PSL UNIVERSITÉ PARIS

cea

THÈSE DE DOCTORAT

DE L'UNIVERSITÉ PSL

Préparée à Chimie ParisTech
Dans le cadre d'une cotutelle avec le CEA Marcoule

Origines microscopiques de la séparation xénon/krypton dans les matériaux nanoporeux

Screening of the microscopic origins of Xe/Kr separation in nanoporous materials

Présentée par

Emmanuel REN

Soutenance prévue le
26 Septembre 2023

École doctorale n°388

Chimie Physique et Chimie Analytique de Paris Centre

Spécialité

Chimie Physique

Composition du jury :

NAME SURNAME TITLE, PLACE/University		*Présidente du jury*
Sofía CALERO Professeure, Eindhoven University of Technology		*Rapportrice*
Christelle MIQUEU Maître de Conférences, Université de Pau & Pays Adour		*Rapportrice*
Johann WILLIAM Responsable modélisation et IA, Orano		*Membre invité*
Isabelle HABLOT Manager recherche et développement, Orano		*Membre invité*
Philippe GUILBAUD Chef de laboratoire et Directeur de Recherche, CEA		*Encadrant thèse*
François-Xavier COUDERT Directeur de Recherche, Chimie ParisTech		*Directeur de thèse*

Chimie Paris ParisTech | PSL

REMERCIEMENTS

Je tiens tout d'abord à adresser mes plus sincères remerciements aux rapportrices de cette thèse, Sofía Calero et Magali Benoit, pour avoir pris le temps de lire, commenter et évaluer mon manuscrit de thèse. J'adresse également mes plus sincères remerciements à la présidente du jury, NAME SURNAME, pour avoir présidé ma soutenance de thèse malgré ses nombreuses obligations.

J'aimerais également remercier Johann William et Isabelle Hablot pour avoir accepté de faire partie du jury en tant que membres invités et représentants d'Orano. Je suis particulièrement reconnaissant envers Isabelle pour son implication dans l'encadrement de ma thèse, apportant toujours un regard bienveillant et pragmatique à mes travaux.

Je remercie également Philippe Guilbaud pour m'avoir toujours accueilli chalheureusement sur le site du CEA Marcoule, même si les circonstances liées à la situation sanitaire ne m'ont pas permis de m'y rendre aussi souvent que prévu. Je suis également reconnaissant de ton encadrement scientifique éclairé de chimiste théoricien expérimenté, ainsi que pour tes conseils bienveillants dans mes démarches administratives avec le CEA, bien que ce ne soit pas ton rôle.

Je tiens à exprimer ma profonde gratitude envers mon superviseur, François-Xavier Coudert, qui m'a ouvert les portes de son bureau et m'y a accueilli pendant plus de trois ans. Ses conseils bienveillants lors des moments de doute, assortis de remarques constructives, ont été inestimables. Au-delà de son expertise dans l'encadrement de projets scientifiques, il m'a également inspiré par sa passion pour la science et sa grande modestie. Je suis convaincu que ces qualités continueront de me guider tout au long de ma carrière.

Je souhaite également exprimer toute ma sympathie et ma reconnaissance envers les doctorants que j'ai eu la chance de rencontrer et avec qui j'ai partagé de précieux moments, que ce soit au bureau ou en dehors : Nicolas Castel, qui a embarqué sur son aventure doctorale en même temps que moi et avec qui j'ai partagé non seulement un bureau, mais aussi des soirées agréables. Wenke Li, qui terminait sa dernière année de thèse et que je connaissais déjà depuis mon stage, je te remercie pour ta bonne humeur, ta générosité et ton altruisme. Maxime Ducamp, qui entamait ses deux dernières années de thèse avec nous et dont l'habitude de se lever tôt m'a toujours impressionné (8h, c'est vraiment tôt). Merci pour ton aide pour les différentes démarches administratives pour la thèse (ton planning avant la soutenance nous a été bien utile avec Nicolas). Lionel Zoubritzky, qui a rejoint notre grande famille des doctorants au début de ma deuxième année et dont les travaux sur la topologie des matériaux nanoporous m'ont toujours intrigué. Merci pour nos nombreuses discussions sur mes problèmes algorithmiques, tu es peut-être la seule personne avec qui je peux en discuter sans t'ennuyer ! Dune André, qui nous a rejoints pour ma dernière année et avec qui j'aurais aimé passer plus de temps. Merci pour ta bonne humeur et de m'avoir fait découvrir le septième art français, moi qui suis plutôt inculte dans ce domaine.

J'aimerais également remercier les postdoctorants avec qui j'ai pu échanger lors des pauses déjeuner et café : Clément Wespiser, pour le verre que nous a payé à ton pot de départ avant que

tu ne rejoignes le CEA. Ambroise De Izarra, pour nos discussions sur RASPA2. Bravo pour tes articles, ce n'était pas évident de coder dans RASPA2 ! Luca Brugnoli, pour partager ta passion pour les livres et les films de science-fiction. Arthur Hardiagon, pour l'ail des ours que tu m'as ramené de chez toi (le pesto était excellent) et pour les concerts brésiliens de ta fanfare que tu nous proposes à chaque fois (j'ai adoré).

Je tiens à remercier les anciens membres de l'équipe avec qui j'ai pu discuter lors de mes stages et qui m'ont donné envie de revenir pour ma thèse : Elsa Perrin, pour m'avoir initié à la simulation moléculaire et à l'utilisation de bash *sur* Linux. *J'ai de bons souvenirs de ma première conférence de l'*AFA *où tu as fait une superbe présentation. Romain Gaillac, pour avoir pris le temps de discuter de mon projet de thèse autour d'un verre. Guillaume Fraux, pour nos discussions scientifiques et ton aide avec* chemfile *pendant mon stage. Siwar Chibani, pour ta bonne humeur et nos discussions sur les défis du parcours académique en recherche.*

Enfin, je voudrais exprimer ma profonde gratitude envers celle que j'aime, Dabeen Oh, pour son soutien constant. Même si les sciences ne sont pas ton domaine de prédilection, tu as toujours fait des efforts pour t'intéresser à mon travail. Je te remercie infiniment d'avoir pris le temps de comprendre, lire, relire et corriger ma thèse. Cette attention précieuse est une véritable preuve d'amour, et je t'en suis infiniment reconnaissant.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to the rapporteurs of this thesis, Sofía Calero and Magali Benoit, for taking the time to read, comment on, and evaluate my thesis manuscript. I also extend my sincere thanks to the chair of the jury, NAME SURNAME, for presiding over my thesis defense despite their numerous obligations.

I would also like to express my appreciation to Johann William and Isabelle Hablot for accepting to be part of the jury as invited members and representatives of Orano. I am particularly grateful to Isabelle for her involvement in supervising my thesis, always providing a supportive and pragmatic perspective on my work.

I would like to express my thanks to Philippe Guilbaud for warmly welcoming me to the CEA Marcoule site, even though I couldn't visit as often as planned due to the prevailing health situation. I am also grateful for his insightful scientific guidance as a theoretical chemist, and for his kind assistance with administrative matters related to the CEA, despite it not being his role.

I am sincerely grateful to my supervisor, François-Xavier Coudert, who opened the doors of his office to me and has welcomed me for over three years now. He has consistently provided me with benevolent advice during moments of doubt, always offering highly constructive feedback. His supervision of scientific projects is accompanied by his enthusiasm, and his continuous dedication is evident in his personal collaborations and research endeavors. His passion for science and his humility have greatly inspired me and will, I hope, continue to do so throughout my career.

I would also like to extend my gratitude to the fellow doctoral students whom I have had the pleasure of meeting and with whom I have shared enjoyable moments, both in and outside the office: Nicolas Castel, who embarked on his three-year journey of doctoral research alongside me. Thank you for being my co-office mate and also my companion for post-work outings and drinks. Wenke Li, who was in the final year of her Ph.D. when I joined and whom I already knew from my internship. Thank you for your good humor, generosity, and altruism. Maxime Ducamp, , who was starting his last two years of Ph.D. with us and whose early mornings have always amazed me. I try to draw inspiration from it, but I think my average is still around 11 a.m. Thank you for the administrative advice regarding the Ph.D. (your planning for the thesis comitee was very useful). Lionel Zoubritzky, who joined our big family of doctoral students at the beginning of my second year and whose Master's work on the topology of nanoporous materials already intrigued me. Thank you for our numerous discussions on my algorithmic problems, you might be the only person I don't bore with that. Dune André, who joined us for my final year, and whom I would have liked to get to know better. Thank you for your good humor and for introducing me to French cinema during our movie outings, I should confess I am still a bit clueless.

I would also like to thank the postdoctoral researchers with whom I had lengthy discussions during lunch and coffee breaks: Clément Wespiser, for the drinks and dinner we had before you joined CEA. Ambroise De Izarra, for our discussions on RASPA2 (congratulations on your papers; coding

in RASPA2 was not easy). Luca Brugnoli, for sharing your passion for science fiction books and films. Arthur Hardiagon, for the wild garlic you brought from your home (the pesto was excellent) and the Brazilian concerts of your brass band (absolutely loved it).

I extend my gratitude to the former team members with whom I had discussions during my internships and who inspired me to return for my Ph.D.: Elsa Perrin, for introducing me to molecular simulation and bash *scripting on* Linux. *I still have fond memories of my first* AFA *conference where you gave a wonderful presentation. Romain Gaillac for having a drink and discussing my Ph.D. project that I later pursued. Guillaume Fraux, for our scientific discussions and your help with* chemfile *during my internship. Siwar Chibani, for our discussions on the intricacies of an academic career in research.*

Lastly, I would like to express my deepest gratitude to the one I love, Dabeen Oh, for her unwavering support. You have always made efforts to take an interest in what I do, even though science is not your strong suit. Thank you very much for taking the time to understand, read, review, and correct my thesis. It is a true testament of your love and I am immensely grateful.

TABLE OF CONTENTS

GENERAL INTRODUCTION

Industrial gas separation processes are widely used for supplying purified reactants and inert gases for various industries such as chemical, health, agricultural and food industries. They can also be utilized to mitigate the negative environmental impact of certain industrial activities. For instance, in concrete or steel production factories, the highly problematic CO_2 emissions can be separated from other atmospheric gases and captured. Similarly, in nuclear treatment plants, volatile radioactive compounds (e.g., ^{85}Kr) can be captured through an effective separation. Typically, these processes involve the consideration of different small molecules such as nitrogen, oxygen, carbon dioxide, hydrogen, methane, nitrous oxide, and noble gases. The focus of this thesis is on the xenon/krypton separation, which is commonly performed to extract xenon and krypton from the atmosphere,[kerry2007industrial] although the nuclear industry constitutes a more abundant source of noble gases.[Banerjee_2014]

In the industry, Xe/Kr separation is usually based on cryogenic distillation of liquified atmospheric air, which requires significant energy, heavy infrastructures, and a meticulous hazard management. The hazardousness of the process is underscored by the occurrence of recent industrial accidents (1997), which resulted from the reaction between non-filtered hydrocarbons and purified liquid oxygen.[distill_accident, distill_accident2] To address security concerns and reduce installation and operational costs of the gas separation process, researchers are actively exploring a promising technology based on competitive adsorption on nanoporous materials. These materials consist of nanoscale pores that provide a large surface area for molecular interaction and adherence. Industrial adsorption separation commonly utilizes pressure swing adsorption (PSA) — the pores are loaded with a gas mixture at high pressure and the gas is subsequently released by applying lower pressure. If the material preferentially loads a single type of molecules, the composition of the released gas can exhibit a significantly higher content of those molecules, hence achieving gradual separation. In this thesis, the xenon/krypton separation is identified as the most challenging step in the purification of xenon, given their chemical similarities. To address this challenge, some prototypes utilizing beds of nanoporous materials have been developed for xenon/krypton separation.[Banerjee2018]

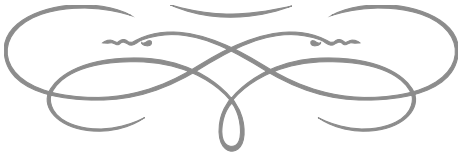
For the process to be viable, materials need to demonstrate improved performance, and numerous studies focus on synthesizing increasingly selective materials by leveraging chemical insights into noble gas adsorption properties.[Chen_2014, Li_2019, Pei_2022] Computational screening plays a crucial role in accelerating the discovery of novel materials with key properties, allowing for the identification of factors that contribute to their performance and the pre-selection of candidates for further experimental studies. The combination of computational discovery and experimental validation, as recently conceptualized by Lyu et al., offers a syner-

gistic workflow to advance material discovery.[Lyu_2020, Jablonka_2022] However, computational chemists face two major challenges in effectively guiding experimental discoveries: generating a greater number of structures reliably and evaluating them using fast and accurate models. This work will primarily focus on the development of tools to address the latter challenge.

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 structures have been digitally constructed.[Wilmer_2012, Boyd_2016, Colon_2017] To efficiently handle this ever-increasing quantity of structures, researchers are developing screening strategies for identifying the best materials, while gaining chemical intuition about the characteristics favorable to a high separation performance. Some studies focus on a multistep screening strategy,[Wilmer_2012, Qiao_2016, Yang_2020] while others use machine learning algorithms to expedite their screening procedures.[Fernandez_2013, Simon_2015, Rosen_2021] Current screening strategies predominantly rely on computational tools that are better suited for single-structure studies rather than high-throughput screenings. Moreover, in the industrial process of PSA introduced earlier and other similar technologies, multiple variables are necessary to fully assess performance, including selectivity, working capacity, and the kinetics and thermodynamics associated with material regeneration (i.e., unloading the pores for another cycle).[Kumar_1994] This thesis aims to address both of these challenges by designing more efficient tools for high-throughput screening, encompassing not only the most commonly studied selectivity performance metric but also transport properties and other relevant factors.

This manuscript begins with a literature review on the screening methodologies applied to various applications of nanoporous materials. The different techniques used in a variety of research fields are explored to inspire the current work.[Ren_2022] For instance, the research focus will be oriented towards the screening of the separation process by breaking down the selectivity metric into thermodynamic quantities such as the enthalpy, the free energy and the entropy. This study, which is based on time-consuming calculations, revealed the effects of pressure, the thermodynamic nature of selectivity and certain structure–property relationships.[Ren_2021] To further improve selectivity screening for Xe/Kr separation, various simulation tools were introduced to evaluate the adsorption performance of a nanoporous material.[Ren_2023] This has opened up new possibilities for developing computationally cheaper and more accurate energy descriptors for the ML prediction of Xe/Kr selectivity at ambient pressure.[Ren_2023_ml] Through this work, faster and more accurate evaluation tools for Xe/Kr separation have been developed, potentially enabling the improvement of the physical description of the system.

As previously mentioned, the gas capacity of the nanoporous materials and the transport properties within them are crucial metrics for evaluating the industrial separation process. The gas capacity can be obtained through GCMC calculations, and alternative methodologies were not thoroughly studied in this thesis. Instead, the fifth chapter of this thesis focuses on determining transport properties and alternative methods of evaluation. Finally, the final chapter provides some perspectives on the physical description of the system (flexibility and polarizability).



1

HIGH-THROUGHPUT COMPUTATIONAL SCREENING OF NANOPOROUS MATERIALS

1.1 Nanoporous materials

Before exploring the screening methodologies of screening, the first section aims to introduce key concepts related to the system of interest, specifically nanoporous materials. These concepts will be referred consistently throughout the text, as the structural characteristics of these materials are intricately linked to their performance in targeted applications.

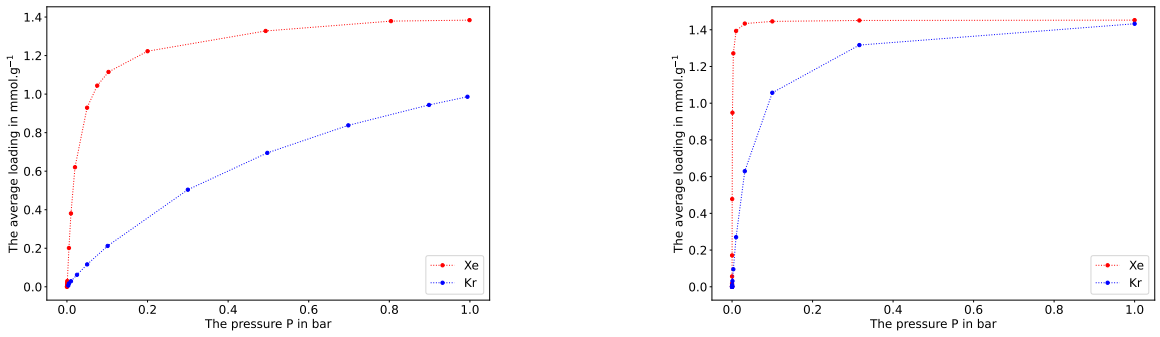
1.1.1 The main characteristics of nanoporous materials

Nanoporous materials are characterized by their nanoscale structure constituted by pores and cavities, some of which are connected by a network of channels. These pores can be empty or filled with a variety of substances called adsorbates. By attaching molecules from either a liquid or a gas phase onto the internal surface of the material, we can use it in a wide range of 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 nature, size, shape and distribution of the pores, we can tailor the physicochemical properties to the targeted application.[Yan_2020]

Adsorption isotherms and geometrical descriptors

The process of particles or molecules adhering on a surface 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. The characterization of adsorption properties of an adsorbate compound typically involves measuring the quantity of adsorbed molecules as a function of its pressure at a given temperature, which is referred to as adsorption isotherm. Figure 1.1 illustrates examples of adsorption isotherms that can be used, along other techniques, to determine the distribution of pore sizes, the accessible surface area and the pore volume.[Rouquerol_1994] These isotherms can also be utilized, through fitting models, to characterize the maximum adsorption uptake, among other adsorption descriptors.[Wang_2020] By using a set of experimental isotherms at similar but distinct 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 determined using calorimetry.[Dunne_1996] Additionally, measurements at infinite dilution enable the establishment of a relationship between the adsorbed quantity and the pressure, as defined by the Henry's law. The Henry adsorption constant, representing the slope of this linear regime,[Finsy2007] serves as another key adsorption descriptor. Although these thermodynamic quantities alone do not provide a complete picture of the adsorption process as an adsorption isotherm would, they are most valuable for comparing experimental data with computational models to rapidly characterize the materials suitable for a target gas adsorption process.



(a) Experimental *(b) Theoretical calculation*

*Figure 1.1: Illustration of monocomponent adsorption isotherms of Xe and Kr obtained experimentally (a) and through GCMC calculations (b). The experimental data are made available[1] by the authors of the Ref. [**Banerjee_2016**].*

[1] Available on Github at https://github.com/CorySimon/XeKrMOFAdsorptionSurv

Most of the materials studied in my thesis have pores with sizes around the nanometer scale, therefore 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, a single terminology – nanopore – will be used to encompass all pores that are a few nanometers or less in size. 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 the structure of the nanoporous material has already been resolved using X-ray diffraction on crystallized porous solids. This method provides the most accurate determination of the PSD, assuming the structure is perfectly rigid and crystalline, allowing for a single set of structural data to 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 be used to evaluate the PSD.**Haynes_1973** Another approach is the Horvath-Kawazoe (HK) method is a semi-empirical analytic model of adsorption isotherm that can extract information about the PSD. Small angle X-ray and neutron scattering methods are non-destructive methods of pore characterization.**Radlinski_2004** My thesis will primarily rely on computationally analyzing experimental structures to deduce pore sizes and other geometric characteristics.

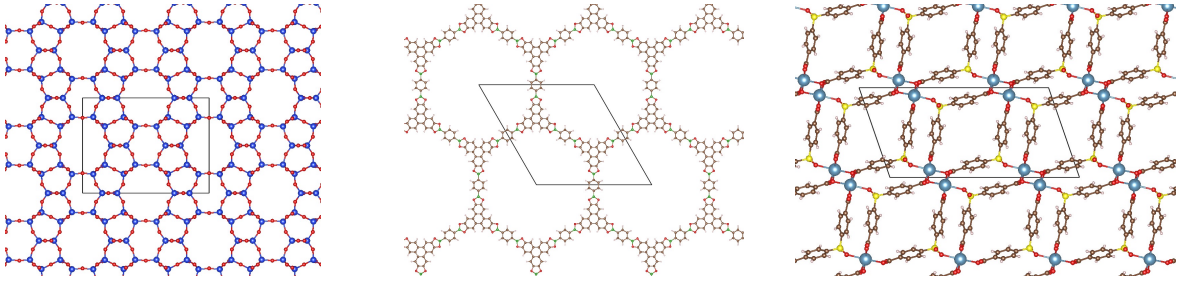
The pore volume of nanoporous materials represents the combined volume of both "closed" and "open" pores. However, the measured quantity can vary depending on the method used. Some pores are not accessible to a specific adsorbate, resulting in different calculated volumes depending on the size of the probe used. Methods that do not rely on adsorption like scattering or stereology techniques can only measure the total pore volume. The porosity or void fraction is defined as the ratio between the pore volume and the apparent framework volume. The specific method used determines whether we can retrieve total porosity, porosity open or closed to a given probe or adsorbate.

The cavities within nanoporous materials provide an exceptionally large adsorbable surface area, which is extremely useful for increasing the number of molecules within a given volume or mass of material, several thousands of square meters can be found in a gram of some nanoporous materials.**Farha_2012** A higher surface area of nanoporous materials allows for a greater number of molecules to be adsorbed for applications such as storage, separation or reactions. Therefore, accurate measurement of surface area using both experimental and computational methods is of utmost importance. The Brunauer–Emmett–Teller (BET) theory is the most widely used method for experimental determination of surface areas based on adsorption isotherms.**Detsi_2011** Most BET areas are calculated using the N_2 isotherm at its boiling temperature (77 K). While alternative probe adsorbates can be considered, they are not standardized.**Tian_2017** However, it is important to note that the definition of surface area is highly dependent on the measurement conditions and the fitting methodology employed. In a statistical experiment involving 61 laboratories, a set of twelve isotherms was provided for BET area calculation, revealing significant disparities in calculation results.**Osterrieth_2022**

Beyond the experimental techniques, Zeo++ and PoreBlazer softwares specialize in computing pore size distributions, surface areas and void fractions based on 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 utilize more comprehensive structural data rather than relying on adsorption models or isotherm data like the BET area calculation. However, the determination of these values heavily relies on a well-designed definition of the volume, the surface and the pore size we aim 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 geometrical descriptors of nanoporous materials.

Classes of nanoporous materials



*(a) Zeolite FER [**FER**]* *(b) COF COF-5 [**Cote_2005**]* *(c) MOF* KAXQIL *[**KAXQIL**]*

*Figure 1.2: Illustration of a zeolite, a covalent organic framework (COF) and a metal–organic framework (MOF). Color code: brown for C, white for H, red for O, blue for Si, cyan for Ca, yellow for S and green for B. The structure visualizations were generated using the VESTA software.***VESTA**

Nanoporous materials exhibit varying degrees of crystallinity, ranging from perfectly crystalline structures to completely amorphous ones. Computational studies primarily focuses on crystalline structures, as atoms are well described within a periodic framework, enabling faster simulations. However, these simulations often neglect the presence of defects, which can explain some discrepancies between simulations and experimental observations. Amorphous materials are described by thousands of atomic positions to capture their inherent non-periodicity.**Thyagarajan_2020** Activated carbons, a well-known class of amorphous material, find extensive industry applications for gas purification purposes. However, characterizing their adsorption properties in a rational manner poses challenges. In general, crystalline nanoporous materials can be categorized into three main classes: inorganic materials such as zeolites (aluminosilicates or aluminophosphates), organic materials like porous polymer networks (PPNs) or covalent organic frameworks (COFs), and 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** They are considered as one of the most mature nanoporous material technologies at our disposal. This class of material offers ample opportunities for innovation as different Al/Si ratios within a specific zeolite type pan out a wide range of structures. Furthermore, zeolite materials have 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 well-established polymer material technology.**Lu_2010**, **Wang_2020**, **Che_2020** However, one of the major drawbacks of this type of material is the formation of irreversible covalent bonds, which make the synthesis kinetically controlled and poses challenges in crystallizing PPNs.**Feng_2012** To overcome this limitation and create crystalline porous materials, Cote et al. developed a strategy using boron-

based organic compounds to form reversible bounds, leading to thermodynamically stable materials COF-1 and COF-5.[Cote_2005] This initiative was led by the group of Yaghi, who has made significant contributions to another very promising and well-known class of materials. A decade earlier, they had 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, resulting in a stable crystalline solid. Although the first synthesis of MOFs dates back to the early 90s,[Abrahams_1991] and brought about a sparking interest in the scientific community a couple of decades later.[Kuppler_2009, Furukawa_2013] The vast number of possible combinations of linkers and metals allows for the theoretical design of an infinite variety of MOFs. Their structure can be tuned to meet our specific requirements and enhance their performance in targeted applications.[Ejsmont_2021] This diversity of nanoporous materials offer a wide range of potential candidates that can be evaluated for any targeted application.

1.1.2 Databases of nanoporous materials

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, nearly a million structures have been considered for applications in separation or storage.[Simon_2015, Simon_2015_EES, Thornton_2017] This extended database can be broken down into synthesized materials and hypothetical ones for all the above-mentioned classes of material.

The International Zeolite Association (IZA) has provided 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 such as the Cambridge Structural Database can be leveraged. However, the raw structures determined experimentally via X-ray cannot be directly used as they are. To obtain a computation-ready dataset, Chung et al. used algorithmic cleaning procedures, resulting in the creation of the publicly available Computation-Ready Experimental MOF (CoRE MOF) database.[Chung_2014, Chung_2019] CoRE MOF 2019 contains about 14,000 MOF structures, making it the largest experimental database available. Similar approach has been applied to organic frameworks leading to the generation of a set of 187 COFs with disorder-free and solvent-free structures.[Tong_2017, Ongari_2019]

These experiment-based databases can provide valuable information on targeted applications using computational screenings, but they have limitations since unknown structures are yet to be discovered. To overcome these limitations and biases of experimental synthesis, the use of artificial methods for generating nanoporous material datasets has proven to be extremely efficient. The first *in silico* generated database of approximately 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] This algorithm served as inspiration for the development of 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 by Colon, Gómez-Gualdrón and Snurr.[Colon_2017] Later, Boyd and Woo introduced another topology-based algorithm using a graph theoretical approach and gener-

ated a 300,000-structure database (BW-DB) based on 46 different network topologies.[Boyd_2016] Similar approaches have been used for other classes of materials. For instance, Deem and co-workers presented a dataset comprising approximately 2.6 million hypothetical zeolite structures.[Earl_2006, Deem_2009, Pophale_2011] However, an important consideration arises regarding the synthesizabiilty and the stability of these hypothetical structures under various 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 with 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. Such predictions can prove very useful for gauging the relative stability of each material and to only consider stable structures. Other types of materials have been explored. For instance, 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 initiated by 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 encompassing all the above-mentioned structures.[kalil2011national, Matgenome, Jain_2013] The fast development of this nanoporous materials genome has motivated Boyd et al. to write a comprehensive review on all the initiatives focused on generating new data for computational analysis.[Boyd_2017]

Yet, simply increasing the size of the databases is not sufficient. It is essential to add diversity to obtain more comprehensive knowledge on the maximum performance and the explanatory features contributing to that performance. Moreover, the diversity of structures ensures the quality of predictions for identifying the best materials for specific applications. To qualitatively or quantitatively assess the diversity of a database, innovative 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. conducted 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 geometric 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 between materials by studying topological patterns in the data.[Lee_2017] These investigations into data structures provide a solid ground for the development of novel materials by objectively defining similarity, diversity and novelty. Based on the analyses conducted thus far, a radical shift in approach is necessary to achieve significant improvements in the diversity of current databases. This requires proposing materials featuring new chemistry, topology or mechanisms (e.g., flexibility).

1.1.3 Exploring the chemical and structural space

With the development of ever-increasing nanoporous material databases, computational chemists have proposed increasingly innovative methods for evaluating and screening thou-

sands of structures. However, new challenges have emerged, including the need to design more efficient screening methods that surpass the brute force screening and effectively analyze big data. Two research groups in Northwestern University, led by R. Snurr and J. Hupp, began to address these challenges by using a "funnel-like" approach to efficiently screen approximately 130,000 hypothetical MOF structures.[Wilmer_2012] To achieve this, they performed a first screening that involved fewer steps of simulation on the entire dataset. They then extracted a subset of top-performing structures to subject them to a second round of simulations, which involved more steps. This procedure was repeated until a few materials were selected through a final round of simulations with reasonable accuracy. Similar "funnel-like" procedures have then been adopted in other fields of applications as described in Figure 1.3. This screening method optimizes computation time by striking a balance between the complexity of calculations and the amount of data to be screened. It ensures that only the most demanding simulations or experiments are applied to the few most promising structures. While this method efficiently identifies top candidates, it has limitations in drawing quantitative structure-property relationships (QSPR) and faces scalability issues beyond a critical dataset size.

To overcome these new challenges, researchers are increasingly turning to transferable models trained by a machine learning (ML) algorithm on diverse and size-limited subsamples. Ideally, such models are 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 design principles for high-performing MOFs in methane storage applications, while also identifying promising structures. In their initial work, they only used geometrical descriptors to describe methane storage.[Fernandez_2013] However, 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 accurately 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]

1.2 Review of screening methodologies

Now that the main concepts around nanoporous materials and their computational representations have been covered, we can introduce the methods used to extract meaningful information from this data. These methods enable us to screen, i.e., evaluate systematically all the structures from a database. Given the scaling issue of such an enterprise, it is essential to employ speedup strategies in these screenings.

The following review aims to provide an exhaustive overview of different screening methodologies used for evaluating nanoporous materials across very different applications. This state-of-the-art review, originally published in Digital Discoveries Ref. [**Ren_2022**], has been adapted for the purpose of this manuscript.

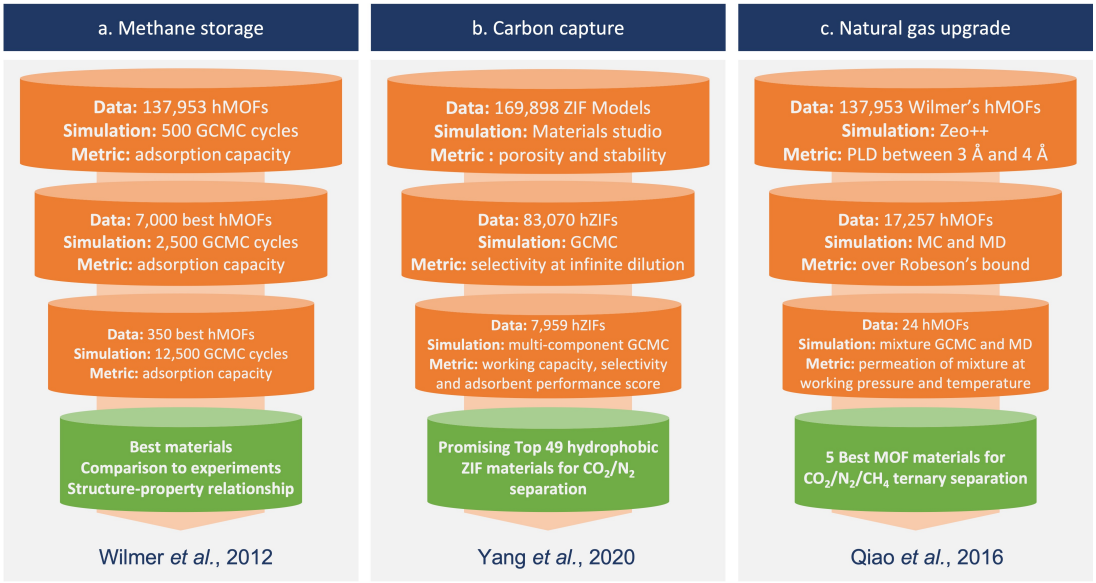


Figure 1.3: Simplified representation of typical funnel-type screening procedures, exemplified on three different applications from the published literature. (a) Wilmer et al.[Wilmer_2012] *used a series of bi-component Grand Canonical Monte Carlo (GCMC) calculations at different levels of complexity 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.*

1.2.1 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 a wide range of other applications. This section focuses on the physical and chemical properties that are not directly related to adsorption processes 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 atomic interactions within the material. To accurately capture these properties, DFT simulations are typically performed. However, the computational cost required for DFT simulations is significantly higher, multiplying the computational time 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 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. While heterogeneous catalysis is generally performed using metallic nonporous structures, the use of nanoporous materials can significantly increase the active surface area and catalytic activity. In this regard, MOFs have demonstrated promising catalytic

properties for several chemical reactions, including hydrogenation, hydrolysis, oxidation among others, as 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 have evolved over the past decade to study larger datasets.

Although the vast majority of computational screenings have been conducted on small series, there have been a few notable attempts to conduct systematic screenings on larger datasets. The scarcity of the latter can be attributed to the high level of computational cost required. Here, we show some examples of such attempts, particularly in the context 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. By screening these descriptors, the authors identified three promising structures for further experimental studies. The best structure 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.

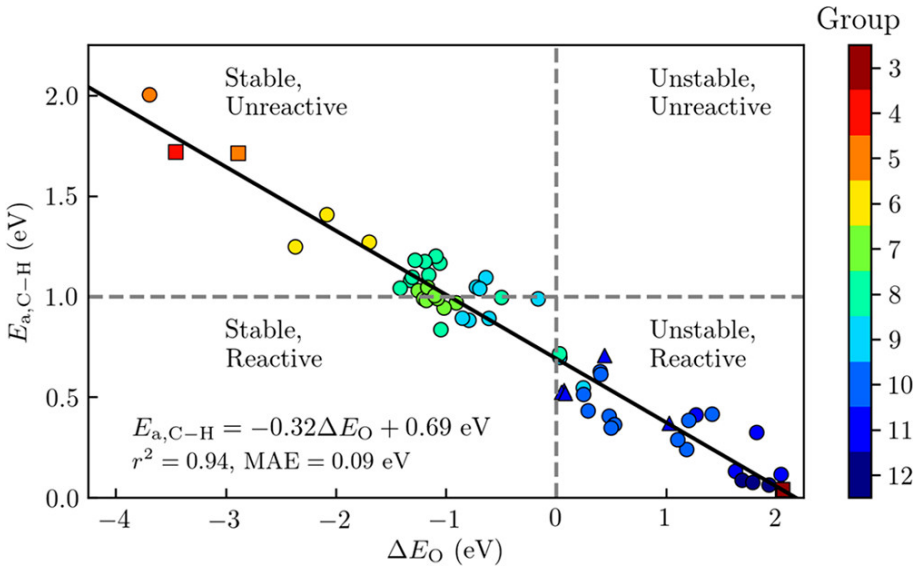


*Figure 1.4: Analysis of a diverse set of experimentally derived metal–organic frameworks (MOFs) with accessible metal sites for the oxidative activation of methane. The graph shows the predicted barrier for the C–H bond activation of methane, E_a, as a function of the metal-oxo formation energy, ΔE_O. 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.*

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 enabling reactant diffusion. They then used a fully automated workflow to place the reactants in the adsorption site and performed periodic DFT calculations to relax the system. As shown in Figure 1.4, they used the bond activation energy $E_{a,C-H}$ and the metal-oxo formation energy ΔE_O as key parameters to classify materials according to their relative stability and reactivity and find the best materials for the application. These energies were then analyzed using physicochemical descriptors such as spin density on oxygen and metal–oxygen distance.

As this brute force screening approach can quickly become cumbersome, researchers in the field are striving to find essential structure-activity relationships to accelerate future computational screenings. Several descriptors have been developed for high-throughput screenings. For instance, Butler et al. used electron removal energies to explain photocatalytic behaviors of MOFs,[Butler_2014] while 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] Fumanal et al. presented a screening protocol based on two energy-based descriptors to predict the photocatalytic properties of MOFs.[Fumanal_descriptor_2020] Rosen et al. recently conducted screenings of thousands of MOF structures to compare different DFT functionals and leveraged the calculated data to train machine learning models capable of rapidly predicting MOF band gaps.[Rosen_2022_high]

The development of ML methods is also critical in the field,[Rosen_2021] although the lack of centralized database with high precision descriptors poses 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. However, these challenges do not undermine the major role of high-throughput screenings in the early design process of nanoporous materials for catalysis. In conclusion, for a more comprehensive understanding of this topic, readers are encouraged to refer to a more exhaustive presentation in Ref..[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 and framework materials. Among these physical properties, mechanical properties have garnered significant interest due to their fundamental importance and relevance to numerous applications. The ability to compute these properties using relatively standard methodologies has further fueled this interest. 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 through 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]

A notable early study that investigated systematically the elastic properties of a specific material family was conducted in 2013 on all-silica zeolites,[Coudert_2013] i.e., crystalline and porous SiO_2 polymorphs. While this study focused on only 121 zeolitic frameworks out of the 244 known

structures, it demonstrated that systematic studies at the DFT level were computationally tractable and provided physical insights into the link between microscopic structure and macroscopic physical properties. This study also showed, 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.

Beyond the specific case of zeolites, other groups have applied DFT calculations to determine elastic constants in a high-throughput manner. For example, de Jong et al. leveraged the structures of the Materials Project[Matgenome, Jain_2013] to chart the diversity of elastic properties across the whole space of inorganic crystalline compounds.[deJong_2015] As shown in Figure 1.5, they provided a database containing the full elastic information of 1,181 inorganic compounds initially. Since then, the database has grown steadily and currently encompasses nearly 14,000 records.[MaterialsProject] This dataset has been used in two different ways by researchers in the field.

Firstly, researchers have employed tensorial analysis to quantitatively investigate the occurrence of certain "anomalous" or rare mechanical behavior within the database of elastic properties. These behaviors include 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 significance, such materials find applications in materials engineering areas, such as energy dissipation (as shock absorbers and for bulletproofing), energy storage, as well as acoustics.[Surjadi_2018] However, quantifying exactly "how rare" these behaviors have been challenging until now. Chibani et al. performed a systematic exploration of available mechanical properties of crystalline materials and demonstrated that the general mechanical trends, which hold for isotropic (noncrystalline) materials at the macroscopic scale, also apply, on average, to crystals. Moreover, they could quantify the presence of materials with rare anomalous mechanical properties: 3% of the crystals exhibited negative linear compressibility, and only 0.3% displayed 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 forcefields 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.

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,

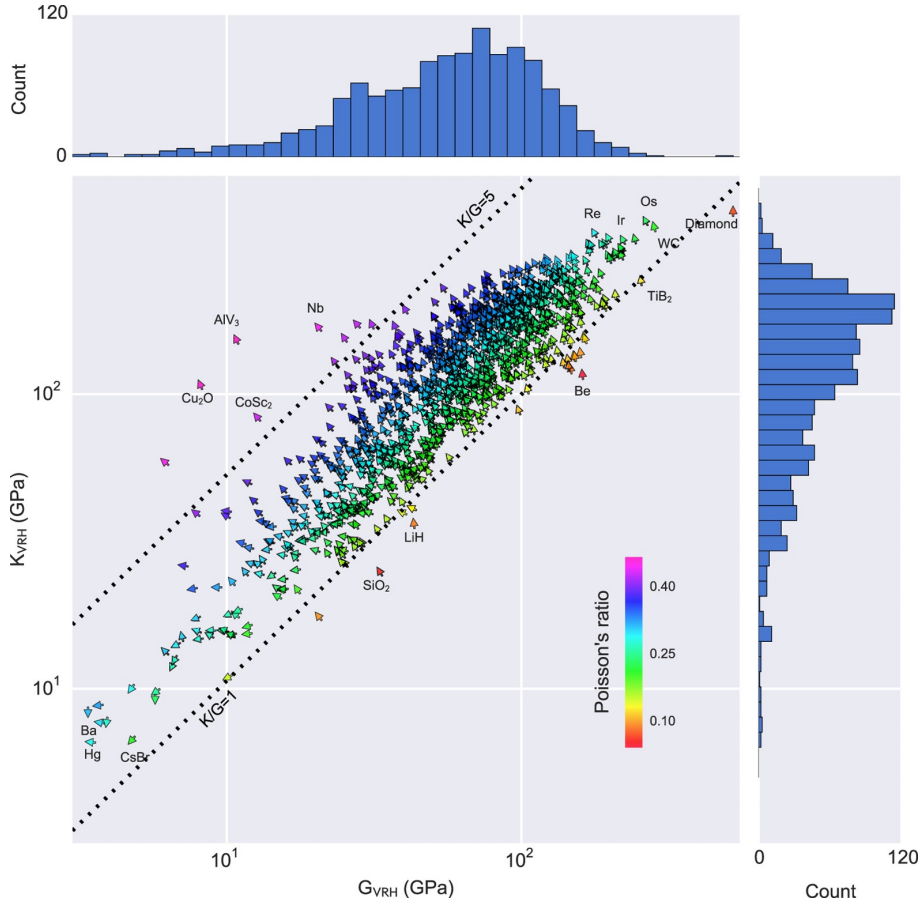


Figure 1.5: Statistical analysis of the calculated volume per atom, Poisson's ratio, bulk modulus K_{VRH} *and shear modulus* G_{VRH} *of 1,181 compounds in the Materials Project database. In the vector field-plot, arrows pointing at 12 o'clock correspond to minimum volume-per-atom and move anti-clockwise 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.

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 1.6, 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.

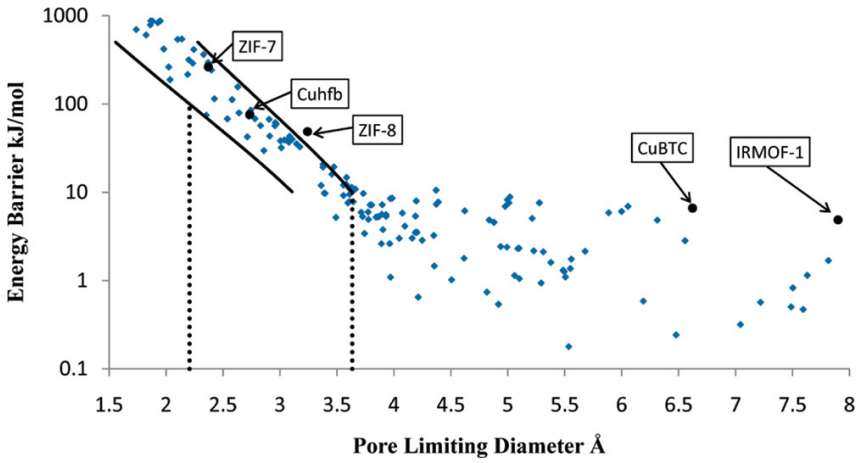


*Figure 1.6: Calculated energy barrier for the diffusion of CH_4 in 216 metal–organic frameworks (MOFs), shown as a function of the pore-limiting diameter. The solid lines represent statistical upper and lower bounds on the energy barrier, in a transition state theory approach. Reprinted with permission 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 1.3 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 Figure 1.7. 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 Figure 1.7, 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]

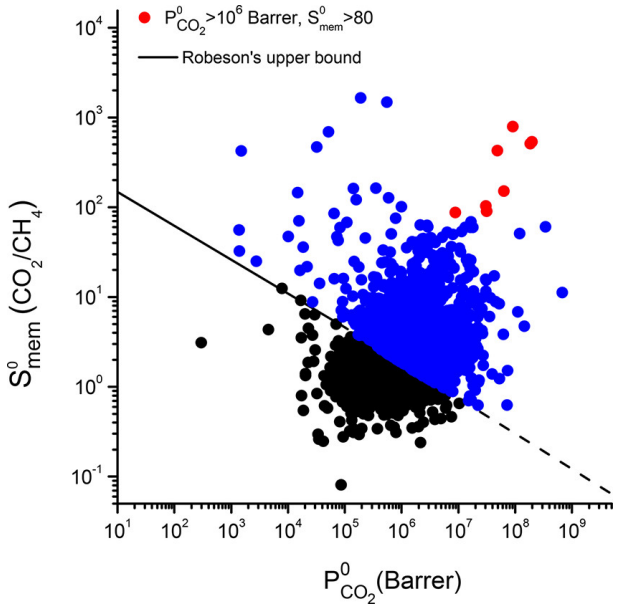


Figure 1.7: Selectivity and permeability of metal–organic framework (MOF) membranes for CO_2/CH_4 separation computed at infinite dilution by combining Grand Canonical Monte Carlo and molecular dynamics simulations.[Altintas_2018] *The black solid line represents the Robeson's upper bound.*[robeson1991correlation, 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 forcefields, 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) and 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 in Figure 1.8. 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

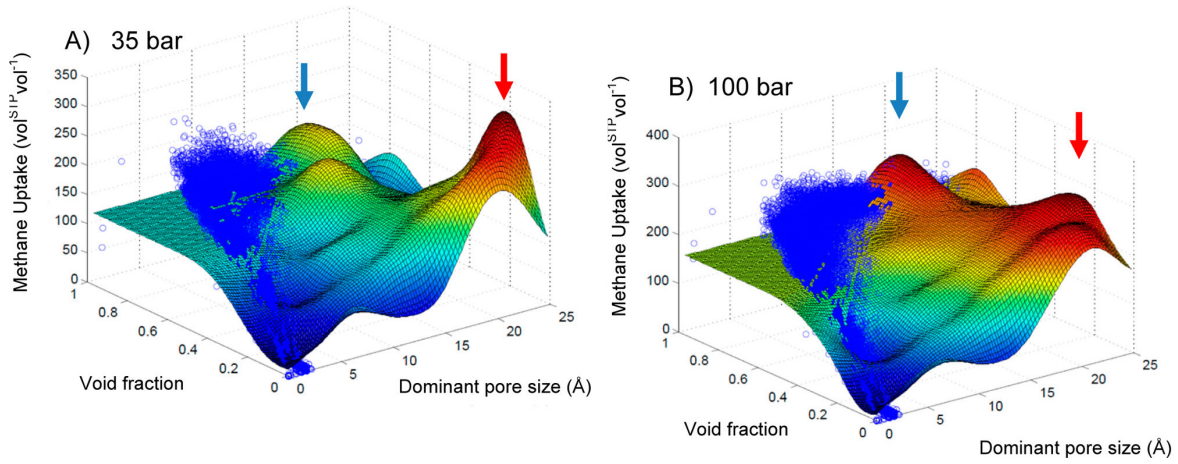


*Figure 1.8: Two-dimensional response surfaces of the support vector machine (SVM) models trained by Fernandez et al. for methane storage at (A) 35 bar and (B) 100 bar using void fraction and dominant pore size. The blue dots represent the GCMC simulated uptake values. The color of the surface represents 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 selective 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_1] 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 in Figure 1.3. 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

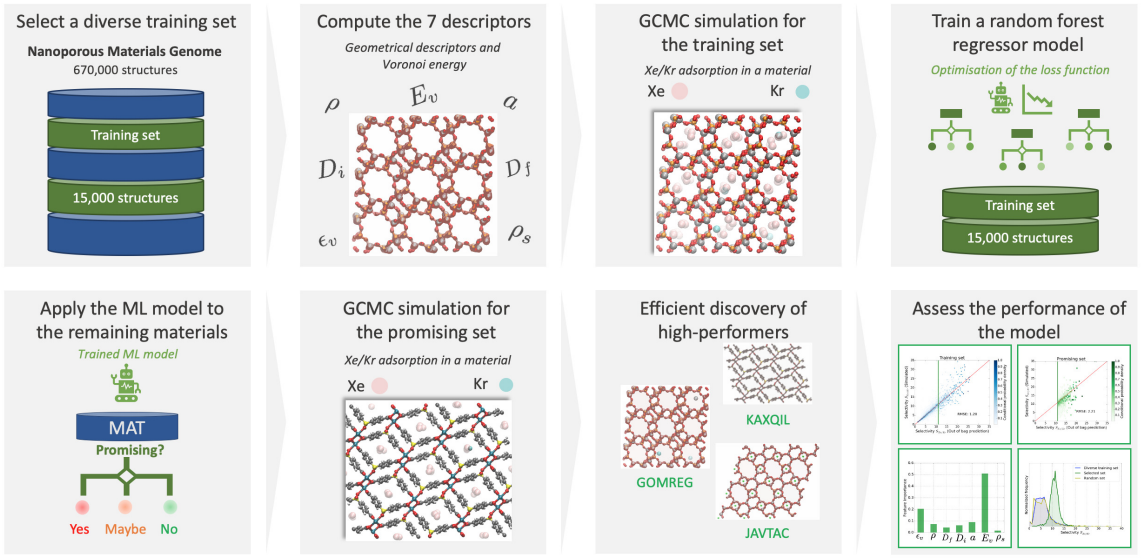


Figure 1.9: Schematic representation of large-scale screening of nanoporous materials for Xe/Kr adsorption-based separation by Simon et al.,[Simon_2015] *based on a combination of Grand Canonical Monte Carlo simulations and machine learning algorithm (Random Forest Regressor). The main goal of this screening 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.*

hypothetical nanoporous material structures for Xe/Kr separation (see Figure 1.9).[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; 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 the further studies needed to better understand the process while discovering new materials.

1.3 Separation of xenon from krypton

This section will be dedicated on how the above-mentioned screening methodologies can 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, holstrater2011intranasal, Mammarappallil_2019] aeronautical,[Patterson_2002, Coxhi] 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., ammonia 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 source	Xe/Kr composition
Extraction from ambient air[kerry2007industrial]	100 ppm/400 ppm
Uranium 235 fission[Blades_1956]	85/15
Spent nuclear fuel[auerbach2003handbook]	90/10
Molten salt reactor[engel1971xenon, Riley_2019]	unknown

*Table 1.1: Composition of the Xe/Kr mixture coming from different industrial sources of production. The Xe/Kr composition can change over time depending on the radioactive decay of Xe and Kr isotopes. The composition out of a molten salt reactor is unknown (no industrial installation exists) and depend on the fissile element, if it is a majority of ^{235}U, then the composition would be very close to the one indicated by the Ref. [**Blades_1956**].*

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 of 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.

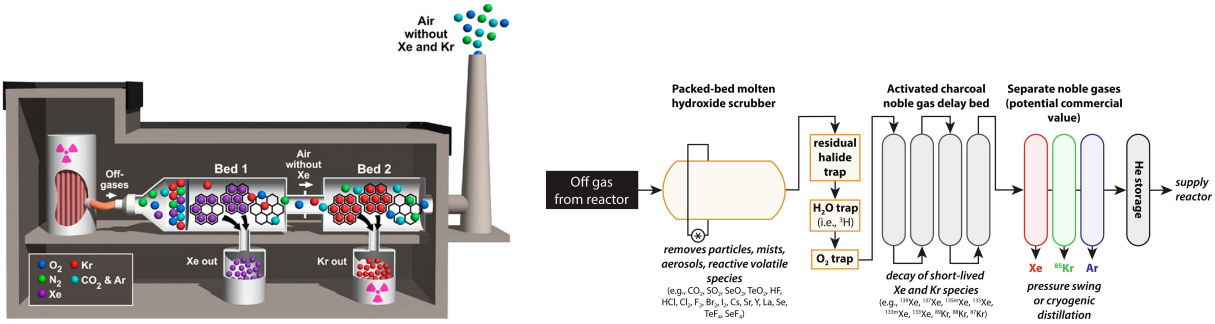


*Figure 1.10: Representation of xenon/krypton separation process using porous materials in a nuclear fuel reprocessing plant (left panel) and in a molten salt reactor (right panel). Reprinted with permission from Ref. [**Banerjee_2014**] copyright © 2014 American Chemical Society and Ref. [**Riley_2019**] copyright © 2019 Elsevier.*

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 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 this work to computational screenings, one of the rare cases of direct contribution of high-throughput screenings to the lab testing of a material will be presented here. 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. 1.9 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 Figure 1.11, 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 immediately imply smaller pore volumes and surface areas since the framework atoms occupy much 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. 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]

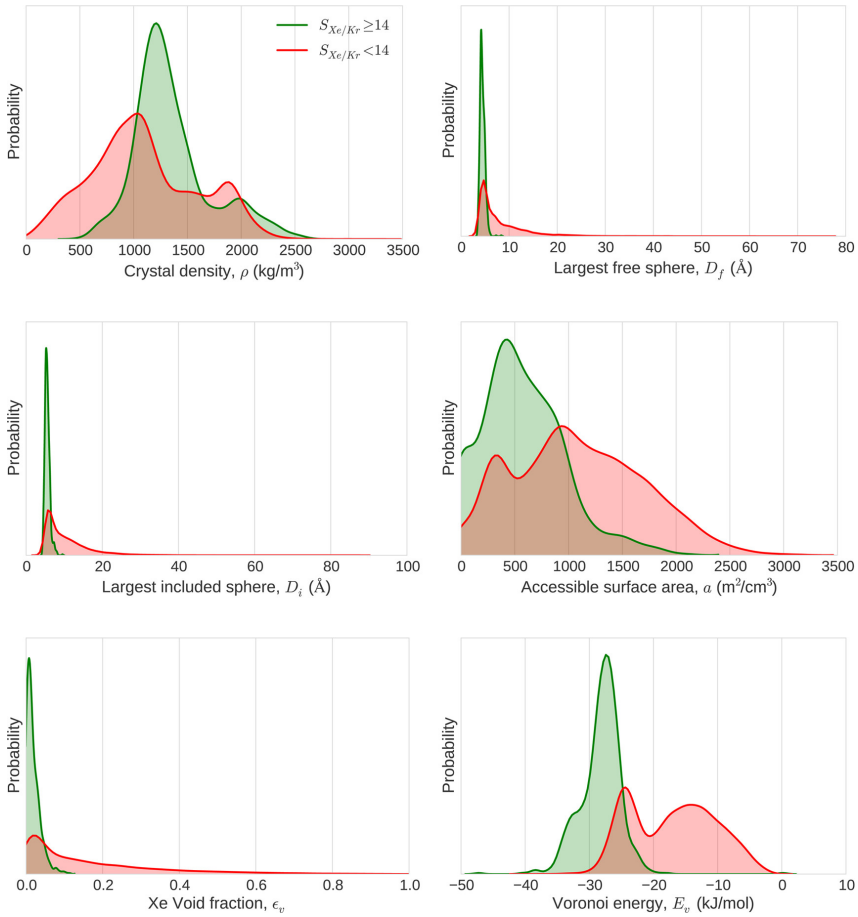


*Figure 1.11: Statistical analysis of the adsorption separation of xenon/krypton mixtures by nanoporous materials. The graphs represent the distributions of structural descriptors explored by highly selective (green) and poorly selective (red) materials separately. Reprinted with permission from Ref. [**Simon_2015**]. Copyright © 2015 American Chemical Society.*

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 (e.g., synthesizability) — a material for industrial xenon/krypton separation we need not only to optimize the selectivity but also the regenerability (how much gas can be retrieved at each cycle) and the capacity and so on, and a key constraint is to know if it is synthesizable or the robustness within the industrial conditions. For instance, although the transport properties of the adsorbates in the material are not key in explaining the separation performance, they are, however, very important in the breakthrough experiments and eventually in the industrial 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 from 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, this thesis focuses 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.

This chapter introduced different studies on the screening of nanoporous material properties coming from a large set of research fields from adsorption to transport properties and exploring the more eclectic mechanical, thermal and catalytic properties. Considering all the different approaches gathered from the literature, the next chapter will aim at better describing the microscopic origins of the xenon/krypton separation in nanoporous materials through a thermodynamics-driven screening. This study will include the study of relationships between the separation performance and geometrical characteristics of the structures, the identification of the thermodynamic nature of the separation. By doing so, some interesting effects were unraveled and we expect to use them at our advantage to accelerate the evaluation of the selectivity at any pressure values.

2

THERMODYNAMIC EXPLORATION OF XENON/KRYPTON SEPARATION

2.1 Characterization of adsorption equilibrium properties

Before exploring the thermodynamic properties of the adsorption-based xenon/krypton separation, this first section aims at defining some key concepts that will be referenced throughout this manuscript. These concepts include the computational definition of the geometrical descriptors mentioned in Chapter 1, as well as the molecular simulations used to assess the separation performance of each material. In addition, the definition of these thermodynamic quantities will be elucidated to ensure a better understanding of the work presented in the following sections.

2.1.1 Geometrical descriptors

Before delving into the details of adsorption properties, it is important to introduce the different simulation techniques used to characterize the internal pore structure of the material. These properties play an essential role in interpreting the adsorption properties obtained using more complex molecular simulations. In this thesis, the Zeo++ software was utilized to calculate all the geometrical descriptors used.[Zeo++] While other tools exist,[First_2013, PoreBlazer] the use of Voronoi decomposition of the volume offers computational efficiency advantages (efficiency gain mainly on volume calculation),[Rycroft_2009] making Zeo++ the preferred tool in this study.

Pore size

Different definitions of pore sizes can be defined depending on the point where we measure it. These diverse pore sizes collectively form what is known as a pore size distribution. However, some pore size values are uniquely defined and can be used to characterize the internal structure. For instance, the diameter of the largest sphere that can freely diffuse in the structure is referred to as 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 2.1 illustrates the difference between these pore sizes. In thermodynamic studies, the term "largest cavity diameter" (LCD) will be frequently employed instead of the largest included sphere D_i. Moreover, the term "pore limiting diameter" will be used instead of D_f, especially when studying the transport effects within the nanopores.

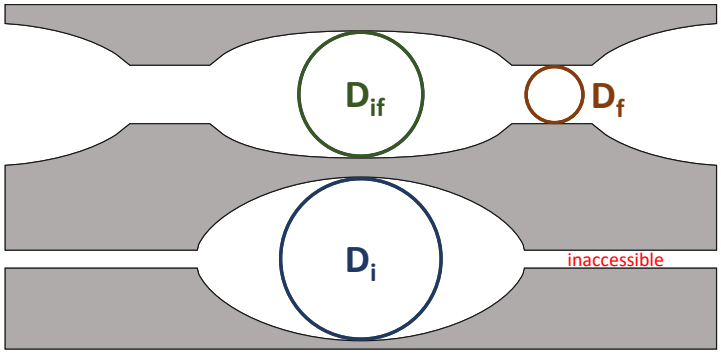


Figure 2.1: Illustration of the different pore sizes D_f, D_i and D_{if}. Note that in some materials D_{if} is equal to D_i, when the largest included sphere is also accessible through a free diffusion path.

To define these pore sizes, it is necessary to first determine the radii of the framework atoms that shape the surrounding pores. These radii can be defined using different methods, with the default mode utilizing the Cambridge Crystallographic Data Centre's (CCDC) radii. This method is widely used in the literature. Additionally, this section introduces another set of radii based on the universal forcefield,[rappe1992] which is used for all types of molecular simulations throughout this thesis. The determination of these radii are inspired by an approach developed by Hung et al.[Hung_2021] The atomic radii correspond to the distance at which the LJ potential reaches $3k_BT/2$, for T = 298 K. This definition enables easier comparison with the quantities obtained from molecular simulations. In case of ambiguity, different indices will be used to differentiate between the two methods. For instance, LCD_{CCDC} corresponds to the standard definition of the LCD that utilizes the CCDC radii when running the Zeo++ software, while the LCD_{UFF} is defined based on the atomic radii that is dependent on the UFF forcefield. In this

chapter, I will mainly use the forcefield-based definition — the largest cavity diameter denoted LCD_{UFF} will be predominantly used, and the studies on void fraction and surface areas will also be defined using this set of radii.

Surface area

The surface areas are calculated using a random sampling technique across the surface of the different atom surfaces. The algorithm counts only the points that do not overlap with another atom. This allows for the calculation of an adsorbable surface for each atom. Ultimately, the surface area can be obtained by summing up all these surfaces. This algorithm, known as the "rolling ball" algorithm, was initially developed by Shrake and Rupley in 1973.**Shrake1973** The Voronoi tessellation determines the accessible and non-accessible areas of the structure by using a probe. Depending on the location of the surfaces, they are categorized as either accessible or non-accessible surface areas. In this chapter, the accessible surface area is defined using a probe of 1.2 Å. This value is computationally equivalent to the experimental N_2 BET surface area.

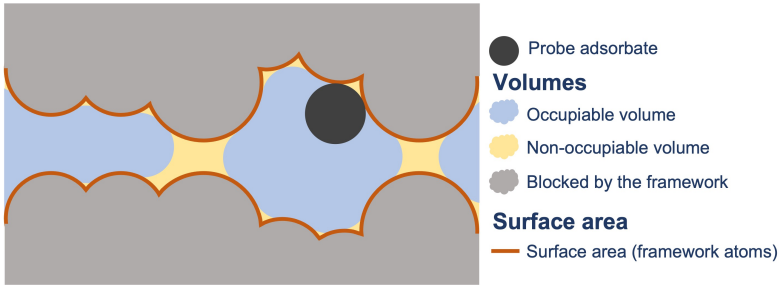


Figure 2.2: Illustration of the pore surface area and volume in a nanoporous material. As illustrated, there are different definitions of the pore volume: we can either consider the whole volume of the pores (occupiable+non-occupiable) or only the volume occupiable by a probe. People usually use the first definition, but the second definition has recently been proposed. Studies have shown that occupiable volume has a better accordance with experimental data.[vol_Ongari2017] *The surface area also changes depending on the definition.*

Pore volume and porosity

The pore volume is calculated by randomly sampling the accessible and inaccessible Voronoi cells. Similarly, other algorithms perform a random sampling over a regular mesh. If the probe used for sampling does not overlap with a framework atom, then it is counted in the number of accessible points N in the volume. The ratio of this number N and the total number of points sampled gives the fraction of the pore volume, also known as the void fraction or porosity. By using the Voronoi decomposition, it is also possible to define the accessible and non-accessible Voronoi cells to reduce the space that needs to be sampled in a Monte Carlo simulation for the surface area and the void fraction calculations.

2.1.2 Intermolecular interactions

In most of the studies in this thesis, rigid structures interacting with guest adsorbates are considered. The intramolecular interactions will not play any significant role in the simulations, as the ionic, chemical, or metallic bonds between the atoms of a molecule are predefined at a specific set of distances and remain unchanged throughout the simulations. As discussed in the final chapter, this approximation can generate discrepancies between the theoretical model and the experimental observations. However, considering the goal of achieving screening

approaches, such as the ones introduced in Chapter 1, adding flexibility in intramolecular interactions would significantly reduce the size of the database that can be screened. For these reasons, the term "interaction energy" will mainly refer to the guest–host and guest–guest intermolecular interactions — host–host interactions would compromise the assumption of framework rigidity.

In classical theory of molecular physics, the intermolecular interactions can be categorized into three different types based on their strength: (i) the ion–dipole and ion–induced dipole interactions (40–600 kJ mol^{-1}), (ii) the hydrogen bonding (10–50 kJ mol^{-1}), and (iii) the van der Waals interactions (1–10 kJ mol^{-1}). It is important to note that these energy values are only indicative and the interaction depends on the nature of the molecules. However, they provide a ranking of the different forces according to their strength. Moreover, the ionic and covalent bonding is always stronger than any intermolecular interactions (over 100 kJ mol^{-1}). The generic term "van der Waals interactions" actually encompasses three different concepts known as the 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** 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, it is possible to consider dipole interactions since they have the most influence 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. However, in fluid phases, the average over all angles is better described by the inverse sixth power, as shown in the equation 2.1 below:

$$U_K = -\frac{\mu_1^2 \mu_2^2}{(4\pi\epsilon_0\epsilon_r)^2 r^6} \times \frac{2}{3k_B T} \tag{2.1}$$

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 2.2.

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

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

$$U_L = -\frac{\alpha_1 \alpha_2}{(4\pi\epsilon_0\epsilon_r)^2 r^6} \times \frac{3}{2} \times \frac{I_1 I_2}{I_1 + I_2} \tag{2.3}$$

where I_1 and I_2 are the first ionization energies. 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 delving into the computational modelization of these long-distance intermolecular forces, it is necessary to specify the repulsive force that

occurs at very short distances. This force can be explained by the Pauli exclusion principle, which states that electrons in both atoms cannot occupy the same quantum space.

For the system of interest, 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. I will use a simple 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 attractive 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, I will use the UFF forcefield (as it performs well compared with *ab initio* forcefields for CO_2 and CH_4 uptake values[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, one usually set a cutoff distance at which the LJ potential can be considered negligible. At this cutoff distance, one can 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. In most of the simulations in my screening studies, I adopted a shifting strategy combined with a cutoff of 12 Å.

In adsorption simulations of other gas molecules with partial charges, it is usually necessary to calculate the Coulomb interaction between the partial charges of the host framework and those of the adsorbate — in periodic systems, an Ewald summation is typically employed to account for this interaction. However, in the case of noble gases, such ion–dipole and dipole–dipole interactions do not exist due to the perfect neutrality of the molecules. Nevertheless, it can be argued that the ion–induced dipole can be adequately described by a simple LJ potential. To provide a comprehensive representation of the intermolecular interactions, the energy induced by the charges of the surrounding framework atoms should be added to the adsorbate. Several approaches have been developed in the literature to enhance the description of the intermolecular interactions by coupling LJ potentials with an induction potential,[Lachet_1998, Becker_2017] which are not used in my work.

To summarize this section on the modelization of the intermolecular interactions in the adsorption simulations, it is essential to highlight the main assumptions in the modelization that may impact the accuracy of the method. Firstly, the framework remains rigid throughout the simulation, which eliminates the need for molecular dynamics simulations of the framework to save time, but it also hides the effects of a known phenomenon.[Witman_2017] Secondly, the polarizability of the adsorbate is only partially considered, as the interaction with the charges of the framework is not taken into account. The difference in polarizability between xenon and krypton can be further exploited to enhance the selectivity, as suggested by experimental studies emphasizing the key role of polar groups and open metal sites.[Li_2019, Pei_2022, Perry_2014]

Lastly, the complex induction and dispersion interactions are described using a two-parameter model. Although this model does not capture all the nuances of differences between the same atom in different environments, it is possible to fine-tune these parameters for very specific cases. However, in a screening strategy, some accuracy on specific cases can be sacrificed to improve the generalization error, as demonstrated by the good performance of the UFF forcefield on a large dataset.[McDaniel_2015] These assumptions have been made to strike a balance between the computational speed and a detailed description of the physical phenomena at stake. Moreover, the focus this thesis is on the development of screening methodologies rather than molecular interaction modeling.

2.1.3 Mixture adsorption: Grand Canonical Monte Carlo

As previously discussed, adsorption can be viewed as a gas–solid or liquid–solid interfacial phenomenon. The adsorbate phase fills the accessible pore volumes depending on the physical conditions of the material. Predicting how adsorbates would interact with the pore surface, the maximum number of molecules that can fit, the most stable configuration, etc., is challenging and cannot be achieved through a simple model. To address these questions, it is necessary to evaluate all possible adsorption configurations each with a different number of adsorbate molecules, and then select the most thermodynamically plausible ones. This evaluation requires these configurations to follow a predefined probability distribution from statistical physics, such as the grand canonical ensemble probability, as it allows for the variation of the number of molecules (adsorbate molecules) and the total energy. By using a Monte Carlo simulation, it is possible to vary the energy and the loading inside the pores so that the distribution of configurations c follows the probability law below:

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

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, I will only focus on the adsorbate molecules. The chemical potential μ and the temperature T ($\beta = 1/k_\mathrm{B}\mathrm{T}$) 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 I will describe now.

Beyond these theoretical considerations, the grand canonical Monte Carlo simulation which refers to a Metroplis-Hastings Monte Carlo algorithm in the context of the grand canonical thermodynamic ensemble, requires several key characteristics to fulfill the above-mentioned probability distribution of the configurations. Monte Carlo (MC) refers to the randomness inherent to gambling games at the eponymous casino on the azure coast of Monaco. In MC simulations, are therefore relying on randomly generating atomic configurations; but it is necessary to remain in the physically possible atomic space to the greatest extent, while exhaustively

exploring all possible chemical configurations. Starting from an initial configuration c_0, the algorithm has different rational moves to change the configuration with a controlled degree of randomness. Some of these moves are illustrated in Figure 2.3. The second key algorithmic step (acceptance or rejection condition), introduced by Metropolis and co-workers, allows the reproduction of any distribution with an unknown multiplicative prefactor.**Metropolis1949** The configuration c_1 resulting from the random move is evaluated by calculating the transition probability (like in a Markov chain) or acceptance rate $acc(c_0 \rightarrow c_1)$:

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

The configuration c_1 is accepted if a number randomly drawn from the $[0, 1]$ interval is higher than the acceptance rate $acc(c_0 \rightarrow c_1)$. For an acceptance rate of 1, when c_1 is more stable than c_0 ($\mathrm{E}_{c_1} - \mu\mathrm{N}_{c_1} \leq \mathrm{E}_{c_0} - \mu\mathrm{N}_{c_0}$), the move is automatically accepted since by construction this randomly drawn number would be lower than 1. On the other hand, if the move is rejected, then we generate another configuration c_1 using another random move, and the acceptance/rejection process restarts until a move is accepted. At the end of the n cycles of the MC simulation, only the accepted configurations $\{c_0, \ldots, c_n\}$ form a Markov chain, and this sequence describes the probability distribution of the grand canonical ensemble described in equation 2.5. The multiplicative prefactor does not influence the algorithm since the acceptance rate corresponds to ratios of probabilities P_c, so that no prior knowledge of the chemical space is needed, which is a valuable simplification.

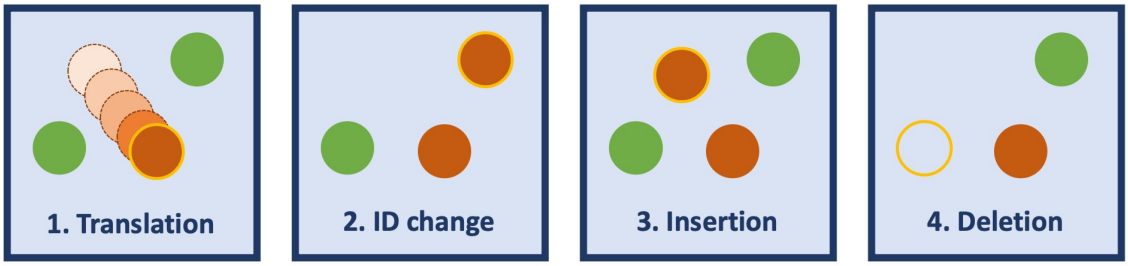


Figure 2.3: MC moves in a system of two types of monoatomic atoms (green and orange). The modification on the first box is highlighted by the yellow circle and the dragging pattern is represented by a set of dashed circles. The boxes 2 to 4 represent the moves going from the initial state represented in box 1, the corresponding move is highlighted by a yellow outer circle. All these moves are used in the GCMC calculations performed using the RASPA2 software.

To complete the description of the grand canonical Monte Carlo (GCMC) simulation, let us now consider the different MC moves used to generate a configuration from another. The probabiities of occurrence of these moves vary depending on the chosen parameterization. For monoatomic molecules, there are only four relevant moves (Figure 2.3): (i) translation of a randomly selected molecule with a displacement randomly chosen within a specific radius, (ii) conversion of the identity of a randomly chosen molecule to another one, (iii) insertion of an adsorbate molecule, and (iv) deletion of an adsorbate molecule. Rotations of the adsorbate are deliberately omitted due to the spherical symmetry of noble gases, and the change of volume is also dismissed since the flexibility of the material framework is not considered. In the GCMC screenings performed in this thesis, the probabilities of translation (i), of identity change (ii), of particle reinsertion ((iii) and (iv)) and of particle swap ((iii) or (iv)) are respectively 1/6, 1/3,

1/6 and 1/3. To clarify the terms used here, for a particle reinsertion, a particle is selected and moved randomly to another location; and for a particle swap, there is the same equal chances to insert a new molecule or to delete one.

By using a GCMC algorithm, it is possible to generate a set of configurations according to their corresponding probability of occurrence. Since the probability law is directly derived from equation 2.5, the series of configurations describe the thermodynamic equilibrium state of a nanoporous material in contact with a reservoir containing a xenon-krypton mixture at a given composition, pressure and temperature. Ensemble averaging enables the derivation of different thermodynamic quantities, such as the averaging loading or uptake at a given pressure (several pressures yield the isotherm) and the isosteric heat of adsorption for 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, it is theoretically sufficient to perform a GCMC calculation at every pressure, temperature and composition conditions. However, such simulations can be very time-consuming due to the need to extensively test insertion/deletion moves to accurately estimate the number of adsorbed molecules and the composition of the mixture. As a result, faster methods (machine learning) have been developed to estimate the selectivity at different physico-chemical conditions.**Simon_2015**, **Kang_2023** For the case of infinite dilution, faster methods are already available. One such method that will be introduced in the following subsection is the Widom insertion, which enables the estimation of adsorption performances at infinite dilution by estimating the Henry adsorption constant.

2.1.4 Infinite dilution adsorption: Widom insertion

In 1963, B. Widom introduced a simple method for calculating thermodynamic properties in materials or fluid mixtures.**Widom1963** This method typically allows accessing to the difference of internal energy before and after the insertion of a test particle while keeping all other particles fixed, thereby comparing the N-particle and (N+1)-particle states. This energy difference $\Delta\Phi$ can then be used to deduce the excess free energy associated with the insertion $\Delta F_{exc} = -k_B T \ln\left(\langle \exp(-\beta\Delta\Phi)\rangle\right)$ by averaging the Boltzmann factors, which corresponds to the excess chemical potential induced by the addition of a particle. More precisely, the average is theoretically performed for all possible positions of the inserted particle; in practice, the tridimensional space is uniformly and randomly sampled until convergence of the value of ΔF_{exc}. In the domain of fluid phase equilibrium, Widom insertion is the most straightforward method to calculate chemical potential values. However, it has limitations in liquid-like phases where the insertable space is very narrow, and no relaxation is implemented to account for the reorganization of surrounding particles.**Nezbeda_1991**

This thesis will only focus on the insertion from 0 to 1 particle, where no issues of overlap between adsorbate particles occur. In this low-loading limit, Widom insertion is simply a random insertion of an adsorbate into an empty nanoporous framework. By randomly sampling

the void space, a distribution of interaction energies $\mathcal{E}_{\text{int}}$ can be obtained. The average of the Boltzmann weights associated with these energies 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, the adsorption enthalpy ΔH_{ads} can also be computed. It should be noted that these quantities remain valid only at infinite dilution, and for higher quantities of adsorbates, the previously described GCMC technique should be used. If the sampling is thorough enough, it is possible to derive the following definitions of ΔG_{ads} (equation 2.9), K_H (equation 2.14) and ΔH_{ads} (equation 2.22) based on a complete sampling of the interaction energies $\mathcal{E}_{\text{int}}$ in all points of the space.

Adsorption Gibbs free energy

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

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

Henry constant

To derive the Henry constant K_H, let us consider an ideal gas in the bulk phase. The number of adsorbed molecules n_{ads} can then be expressed using the bulk density of the adsorbate molecule $\rho_{\text{ads,bulk}}$ and the volume of the pores V_{pore}:

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

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:

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

Let us apply equation 2.11 and the perfect gas equation of state $P = \rho_{\text{ads,bulk}}RT$ on the bulk gas in equilibrium. The equation 2.10 can be simplified as

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

By considering the gravimetric loading L_{ads} (in mmol g^{-1}) instead of absolute value, we need to divide the equation by the mass density of the framework ρ_f:

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

Since the Henry's law is described by $L_{\text{ads}} = K_H \times P$, we have the final relationship between the Henry adsorption constant and interaction energy distribution. If we consider a mixture, the pressure should be replaced by partial pressures.

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

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 derives from the perfect gas assumption made in equation 2.10, which is a good approximation in the case of noble gas.

Adsorption enthalpy or heat of adsorption

Finally, if we consider an adsorption equilibrium (e.g., $\mathrm{Xe_{(g)}} \rightleftharpoons \mathrm{Xe_{(ads)}}$), we can define an equilibrium constant $\mathrm{K_{ads}}$ based on the thermodynamic activities (partial pressure for a gas and volumetric loading for an adsorption phase) of the adsorbate in the different phases:

$$\mathrm{K_{ads}} = \frac{l_{\mathrm{ads}} \mathrm{P}^o}{y_{\mathrm{gas}} \mathrm{P} c^o} \tag{2.15}$$

where $l_{\mathrm{ads}} = n_{\mathrm{ads}}/\mathrm{V}$ is the volumetric loading in the adsorbed phase (similar to a molar concentration) and y_{gas} the mole fraction in the gas phase for a given compound (e.g., Xe), and P^o is the standard pressure. And, here we assume the gas ideal by taking a fugacity coefficient of 1. For a gas at infinite dilution, the Henry's law can then be applied to derive the following relation:

$$\mathrm{K_{ads}} = \frac{n_{\mathrm{ads}} \mathrm{P}^o}{y_{\mathrm{gas}} \mathrm{P V} c^o} = \frac{\mathrm{K_H} y_{\mathrm{gas}} \mathrm{P} \rho_f \mathrm{V P}^o}{y_{\mathrm{gas}} \mathrm{P V} c^o} = \frac{\mathrm{K_H} \rho_f \mathrm{P}^o}{c^o} = \frac{\langle \exp\left(-\mathcal{E}_{\mathrm{int}}/\mathrm{RT}\right)\rangle}{c^o \mathrm{RT}/\mathrm{P}^o} \tag{2.16}$$

As a sanity check, we can verify that c^o/P^o has a unit homogeneous with a molar energy, which is consistent with $\mathrm{K_{ads}}$ being unitless.

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

$$\Delta \mathrm{H_{ads}} = -\mathrm{R} \frac{\mathrm{d} \ln\left(\mathrm{K_{ads}}(\mathrm{T})\right)}{\mathrm{d}(1/\mathrm{T})} \tag{2.17}$$

Then by decomposing the logarithm on the fraction of equation 2.16,

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

Then, as $c^o \mathrm{R}/\mathrm{P}^o$ is constant for the variable T, the expression can be simplified to two terms, the first one being the logarithmic derivative of itself (1/T) and the second term is the logarithmic derivative of the sum of the exponential terms.

$$\Delta \mathrm{H_{ads}} = 0 - \mathrm{R} \frac{\mathrm{d} \ln\left(\langle \exp\left(-\mathcal{E}_{\mathrm{int}}/\mathrm{RT}\right)\rangle\right)}{\mathrm{d}(1/\mathrm{T})} - \mathrm{RT} \tag{2.19}$$

Using the property that the logarithmic derivative of a function f is obtained by the formula $\frac{\mathrm{d}\ln(f)}{\mathrm{d}x} = f'/f$, we can calculate the derivative of the average of the Boltzmann factors $\langle \exp(-\mathcal{E}_{\mathrm{int}}/\mathrm{RT}) \rangle$:

$$\Delta \mathrm{H_{ads}} = -\mathrm{R} \frac{1}{\frac{1}{\mathrm{N}} \sum e^{-\frac{\mathcal{E}_{\mathrm{int}}}{\mathrm{RT}}}} \frac{1}{\mathrm{N}} \sum \frac{\mathrm{d} \exp(-\mathcal{E}_{\mathrm{int}}/\mathrm{RT})}{\mathrm{d}(1/\mathrm{T})} - \mathrm{RT} \tag{2.20}$$

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 obtain the following expression:

$$\Delta \mathrm{H_{ads}} = -\mathrm{R} \frac{1}{\sum e^{-\frac{\mathcal{E}_{\mathrm{int}}}{\mathrm{RT}}}} \sum -\frac{\mathcal{E}_{\mathrm{int}}}{\mathrm{R}} e^{-\frac{\mathcal{E}_{\mathrm{int}}}{\mathrm{RT}}} - \mathrm{RT} \tag{2.21}$$

With some simplification, the adsorption enthalpy $\Delta \mathrm{H_{ads}}$ can be expressed as the Boltzmann average of the interaction energies minus a term RT that corresponds to the internal energy in the bulk phase under the ideal gas assumption (perfect gas equation of state).

$$\boxed{\Delta \mathrm{H_{ads}} = \frac{\sum \mathcal{E}_{\mathrm{int}} e^{-\frac{\mathcal{E}_{\mathrm{int}}}{\mathrm{RT}}}}{\sum e^{-\frac{\mathcal{E}_{\mathrm{int}}}{\mathrm{RT}}}} - \mathrm{RT}} \tag{2.22}$$

The isosteric heat of adsorption q_{st} is then simply the opposite of the adsorption enthalpy, at infinite dilution.

Adsorption entropy

From the values of the adsorption free energy and enthalpy, we can now deduce the adsorption entropy $\Delta \mathrm{S_{ads}}$ using the definition of the Gibbs free energy ($\mathrm{G} = \mathrm{H} - \mathrm{TS}$):

$$\boxed{\Delta \mathrm{S_{ads}} = \frac{1}{\mathrm{T}} (\Delta \mathrm{H_{ads}} - \Delta \mathrm{G_{ads}})} \tag{2.23}$$

selectivity

In the thesis, the selectivity is defined as the ratio of the proportion of Xe/Kr in the adsorption phase to the proportion in the gas phase in equation 2.8. At infinite dilution, the selectivity can be rewritten using the Henry's law ($q^{\mathrm{i}} = \mathrm{V} \rho_f \mathrm{K_H^i} y^{\mathrm{i}} \mathrm{P}/n_{\mathrm{tot}}$) and simplifying the constant term $\mathrm{PV}\rho_f/n_{\mathrm{tot}}$:

$$\boxed{s = \frac{\mathrm{K_H^{Xe}} y^{\mathrm{Xe}}}{\mathrm{K_H^{Kr}} y^{\mathrm{Kr}}} \times \frac{y^{\mathrm{Kr}}}{y^{\mathrm{Xe}}} = \frac{\mathrm{K_H^{Xe}}}{\mathrm{K_H^{Kr}}}} \tag{2.24}$$

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

This section has simple thermodynamic quantities such as the adsorption Gibbs free energy, enthalpy and entropy from the study of a simple adsorption equilibrium equation. The following section will explore a thermodynamic characterization of the adsorption-based separation process by 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 have been introduced, let us rationalize the separation process by modeling the process within a hypothetical "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:

$$\mathrm{Xe_{(g)}} + \mathrm{Kr_{(ads)}} \rightleftharpoons \mathrm{Xe_{(ads)}} + \mathrm{Kr_{(g)}} \tag{2.25}$$

At any pressure and for a given composition, the equilibrium constant associated with equation 2.25 is simply the selectivity s, defined in equation 2.8, as the gas phase activities of $\mathrm{Xe_{(g)}}$ and $\mathrm{Kr_{(g)}}$ correspond to the partial pressures y^{Xe} and y^{Kr}, while the adsorption phase activities of $\mathrm{Xe_{(ads)}}$ and $\mathrm{Kr_{(ads)}}$ correspond to their mole fractions q^{Xe} and q^{Kr}.

Exchange Gibbs free energy

The Gibbs free energy at equilibrium can be directly defined using the equilibrium constant. Applying this relation to the exchange equilibrium, it is possible to define an exchange Gibbs free energy $\Delta_{\mathrm{exc}}\mathrm{G}$:

$$\boxed{\Delta_{\mathrm{exc}}\mathrm{G} = -\mathrm{RT}\ln(s)} \tag{2.26}$$

Exchange enthalpy

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

$$\boxed{\Delta_{\mathrm{exc}}\mathrm{H}^{\mathrm{Xe/Kr}} = \Delta_{\mathrm{ads}}\mathrm{H}^{\mathrm{Xe}} - \Delta_{\mathrm{ads}}\mathrm{H}^{\mathrm{Kr}}} \tag{2.27}$$

Moreover, the adsorption enthalpy $\Delta_{\mathrm{ads}}\mathrm{H}$ is generally defined by comparing the average energy difference between systems differing by one adsorbate:

$$\Delta_{\mathrm{ads}}\mathrm{H} = \langle \mathrm{E} \rangle \left(\langle \mathrm{N} \rangle + 1 \right) - \langle \mathrm{E} \rangle \left(\langle \mathrm{N} \rangle \right) - \mathrm{RT} \tag{2.28}$$

In a GCMC calculation, we do not use the previous equation as it is, but we use a formula derived from the fluctuation theorem in statistical mechanics (see a complete derivation in this online article[github_simon_gcmc]):

$$\Delta_{\mathrm{ads}}\mathrm{H}^{\mathrm{Xe}} \simeq \frac{\partial \langle \mathrm{E} \rangle}{\partial \langle \mathrm{N} \rangle} - \mathrm{RT} = \frac{\partial \langle \mathrm{E} \rangle}{\partial \langle \beta\mu \rangle} \Big/ \frac{\partial \langle \mathrm{N} \rangle}{\partial \langle \beta\mu \rangle} - \mathrm{RT} = \frac{\langle \mathrm{EN} \rangle - \langle \mathrm{E} \rangle \langle \mathrm{N} \rangle}{\langle \mathrm{N}^2 \rangle - \langle \mathrm{N} \rangle^2} - \mathrm{RT} \tag{2.29}$$

where E corresponds to the total energy of the adsorption system and N the total number of adsorbates at every step of the simulation. Note that this equation remains only valid for $\mathrm{N} \gg 1$, as the first step of the derivation is based on a first order Taylor expansion $\langle \mathrm{E} \rangle \left(\langle \mathrm{N} \rangle + 1 \right) - \langle \mathrm{E} \rangle \left(\langle \mathrm{N} \rangle \right) \simeq \frac{\partial \langle \mathrm{E} \rangle}{\partial \langle \mathrm{N} \rangle}$.

On the other hand, at infinite dilution, we can derive back the equation 2.22 using equation 2.28, where for $\mathrm{N} \to 0$ we now have $\Delta \mathrm{H}_{\mathrm{ads}} = \langle \mathrm{E} \rangle(1) - \langle \mathrm{E} \rangle(0) - \mathrm{RT}$. The average energy with one

adsorbate minus the average energy without adsorbate corresponds to the average over the whole space of the guest–host interaction energies for one adsorbate particle (the host-host energy being encompassed in $\langle E \rangle(0)$). This expression of the adsorption enthalpy echoes with the one derived in equation 2.22.

Exchange entropy

Now that the exchange Gibbs free energy and an exchange enthalpy have been defined at any pressure, the same approach can be applied as in equation 2.23 to derive the exchange entropy:

$$\boxed{\Delta_{exc}S = \frac{1}{T}\left(\Delta_{exc}H - \Delta_{exc}G\right) = \frac{1}{T}\Delta_{exc}H + R\ln(s)} \tag{2.30}$$

Conclusion

Before concluding this methodological section, it is important to note that the thermodynamic quantities associated with the newly defined adsorption exchange equilibrium can be defined at different pressure, temperature and chemical composition conditions. Moreover, various methodologies can be used to calculate them. At infinite dilution, Widom insertions and the adsorption free energies and enthalpies are typically used to deduce the adsorption free energies and enthalpies and the exchange quantities associated with them. At higher pressures, GCMC calculations are necessary to define the free energy (via the loading values) and the isosteric adsorption heat. The following study will focus solely on two characteristic pressures: the ambient pressure (at 1 atm) and the limit of zero loading (infinite dilution). At 1 atm, the previously defined quantities will be denoted with an index 1 to distinguish 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.

As for the simulations details, it is worth mentioning that for the GCMC calculations and Widom insertions, the RASPA2 software, developed by Dubbeldam et al.,[dubbeldam2016] was used. 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 were obtained from the universal forcefield (UFF),[rappe1992] while the LJ parameters for the guest atoms (xenon and krypton) were taken from a previous screening study.[Ryan_2010] All the MOFs described in this study were taken from the CoRE MOF 2019 database.[Chung_2019]

2.2 Preliminary analyses of the separation performance

The previous chapters showed how the computational screening of the nanoporous materials — both existing frameworks and hypothetical structures — for targeted adsorption properties has been a subject of extensive research. Several high-throughput screening studies have particularly focused on noble gas separation, and Xe/Kr separation. In addition to the testing and validation of methodological developments, large-scale studies have generally aimed to achieve one of three main objectives: (i) identify top performing materials for synthesis and/or characterization; (ii) 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 by analyzing correlations between separation performance

and structural properties of the materials, which provides chemical intuitions on designing better-performing materials. In this initial screening study of the thermodynamic quantities, I performed a screening of approximately 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. The focus of this study is to explore different relationships between Xe/Kr selectivity and structural descriptors based on geometrical analyses, as well as different thermodynamic descriptors (free energy, enthalpy, entropy). Some results have already been published in a scientific article [**Ren_2021**].

2.2.1 Structure–selectivity relationships

An adsorption separation process is primarily characterized by a pivotal performance metric, known as the selectivity, as defined in equations 2.8 and 2.24. To characterize materials that are likely to exhibit selectivity for a 20:80 Xe/Kr mixture separation (to compare with most literature screenings on a mixture extracted from the air), this selectivity was compared to geometrical descriptors calculated by the Zeo++ software.[Zeo++] Three structural descriptors have been computed: 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] a list of van der Waals radii was defined to be used in the Zeo++ software.[1] In all Zeo++ calculations, an atomic radius was chosen based on the distance where the LJ potential reaches $3k_BT/2$, for T = 298 K.

Xenon uptake and selectivity

Before delving deeper into the structure–selectivity relationship, the relation between the xenon uptake (the number of adsorbed xenon in the GCMC simulation) and the selectivity at 1 atm will be described. For instance, the xenon uptake is a crucial factor in the separation process, as it defines the working capacity of xenon produced through adsorption/desorption cycles. Figure 2.4 reveals the possibility to have materials with a very high xenon uptake and a moderately high selectivity, or with very high selectivity but associated lower uptakes. Materials can exhibit selectivity values exceeding 100 with Xe uptake around 3 $mmol\,g^{-1}$, whereas an uptake exceeding 6 $mmol\,g^{-1}$ can only be obtained for selectivity values between 10 and 20. It becomes evident that maximizing both uptake and selectivity metrics simultaneously is challenging, and a trade-off must be made when designing nanoporous materials for xenon/krypton separation.[Zhang_2022] Various strategies, such as the adsorbent performance score (APS),[Solanki_2020] have been implemented to optimize both metrics using mixed metrics. This trade-off can be rationalized by using the different structural descriptors (pore size, surface area and volume) introduced earlier.

Furthermore, in the optimization of either xenon uptake or Xe/Kr selectivity, it is important to note that the best materials for each of these metrics are very rare within a given diverse dataset. The histogram shown in Figure 2.4 demonstrates this scarcity, with a very low number of highly selective materials and high-capacity materials. The most frequently observed materials typically have a selectivity ranging from 1 to 10 and an uptake below 3 $mmol\,g^{-1}$. These

[1] A code can be found at https://github.com/eren125/zeopp_radtable

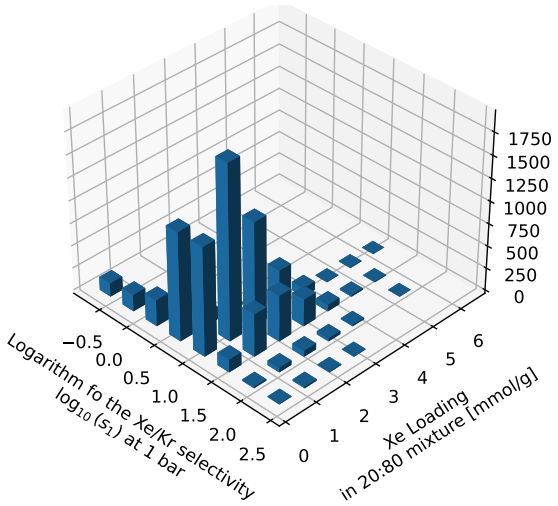


Figure 2.4: 3D histograms of in a bidimensional space formed by the Xe/Kr selectivity and the xenon uptake. The z-axis represents the number of structures with characteristics close to the one specified in x and y-axis. A base-10 logarithm has been applied to the selectivity values.

values can be considered as standard values for nanoporous material used in Xe/Kr separation, serving as reference values for comparing various performance metrics and building a chemical intuition. Therefore, a selectivity exceeding 20 is considered relatively high (even though top-performing materials have a much higher selectivity[Pei_2022]) and a xenon uptake exceeding 4 mmol g^{-1} indicates a significant adsorption capacity. The scarcity of these top-performing materials gives rise to the analogy of searching for a needle in a haystack, prompting some computational studies to design algorithms that focus on identifying the best materials rather than equally describing all materials.[Deshwal_2021, Glasby_2021]

Surface area and selectivity

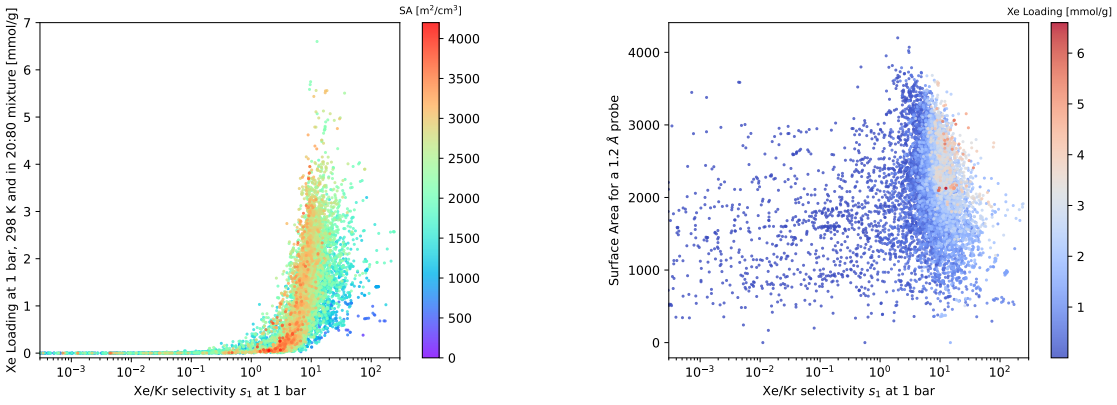


Figure 2.5: On the left: scatterplot of the xenon uptake as a function of the selectivity and labeled by the values of the surface area. On the right: scatterplot of the selectivity and the surface area labeled by the quantity of xenon adsorbed. The selectivity and uptake are calculated by a GCMC simulation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

Studies on methane storage applications, conducted by Wilmer et al.[Wilmer_2012] and by Fernandez et al.,[Fernandez_2013] have shown that methane uptake reaches its maximum within a specific optimal range of surface area values (2500-3000 m^2 cm^3). Increasing the surface area

beyond the range does not lead to higher values of methane uptake. This limitation is also observed for the Xe/Kr selectivity, as shown in the right plot of the Figure 2.5. Materials with a selectivity around 5 tend to have surface areas ranging from 0 to 4000 $m^2\ cm^3$, while those with a selectivity above 40 tend to have a surface area below 2500 $m^2\ cm^3$. On the other hand, the optimal surface area for xenon uptake falls between 2000 and 3000 $m^2\ cm^3$. It is evident that the relationship between selectivity and surface area is highly complex, and a precise range of surface areas does not guarantee high selectivity. Other structural descriptors need to be considered in conjunction with this descriptor to fully characterize selectivity.

The 3D histogram in Figure 2.6 provides a visual representation of the surface area distribution for different selectivity categories. For selectivity values higher than 92, the surface areas are mostly below 2000 $m^2\ cm^3$. In the range of 92 to 35 selectivity, the distribution extends slightly wider, reaching up to 2500 $m^2\ cm^3$. For selectivity values between 35 and 13, the interval spans a larger range, up to 3500 $m^2\ cm^3$, but remains centered predominantly between 1000 and 2500 $m^2\ cm^3$. This split view of the distributions provides a better understanding of the characteristics of the best materials. However, it is important to note that surface area is not a deterministic variable as it is not possible to deduce selectivity based on surface area alone. A surface area between 500 and 1000 $m^2\ cm^3$ may have a relatively high chance of exhibiting selectivity, but it encompasses a large number of materials and is even more likely to have selectivity values between 5 and 35 rather than values higher than that.

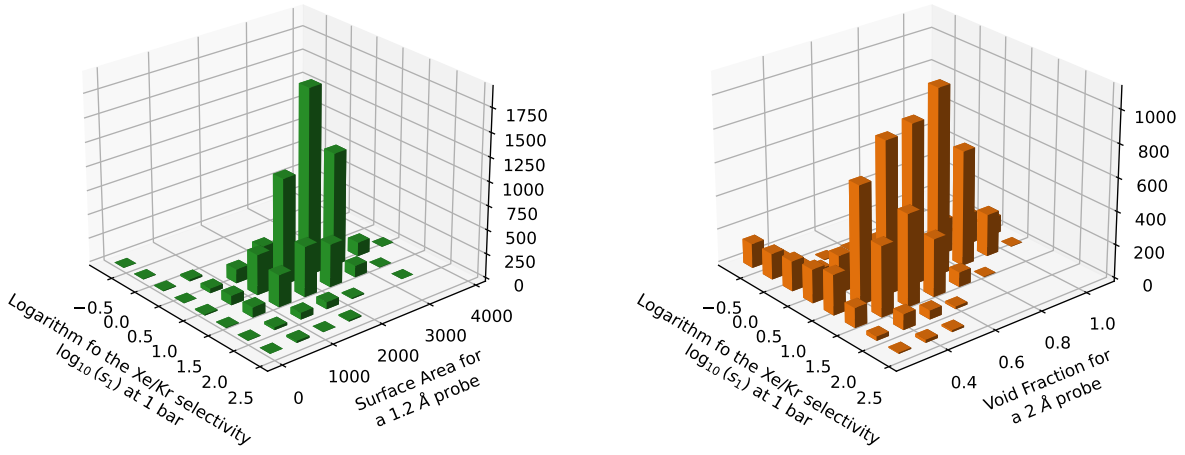


Figure 2.6: 3D histograms of in a bidimensional space formed by the Xe/Kr selectivity and the surface areas (on the left) and formed by the Xe/Kr selectivity and the pore void fractions (on the right). A base 10 logarithm has been applied to the selectivity values. Bin size increased by 2.4 (on log scale)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

Void fraction and selectivity

A similar analysis for void fraction was also conducted by Wilmer et al. for methane storage applications (Figure 5 of Ref. [**Wilmer_2012**]) and they found an optimal void fraction value of approximately 0.8. As shown by the plots in Figure 2.7, materials with the highest value of Xe uptake tend to have void fraction values around 0.5, whereas those with the highest selectivity value exhibit much lower void fractions around 0.1. The optimal range of void fraction for maximizing uptake lies between 0.2 and 0.6, while for selectivity, the optimal range

is completely dissociated and falls below 0.2. Utilizing the void fraction as a descriptor allows for a more refined characterization of selectivity compared to the use of surface area, even though both descriptors yield very similar results. It becomes evident that both descriptors describe a relatively dense material with "microporosity", in accordance with the IUPAC definition,[Sing_1985] indicating materials with medium-low pore volume and surface area.

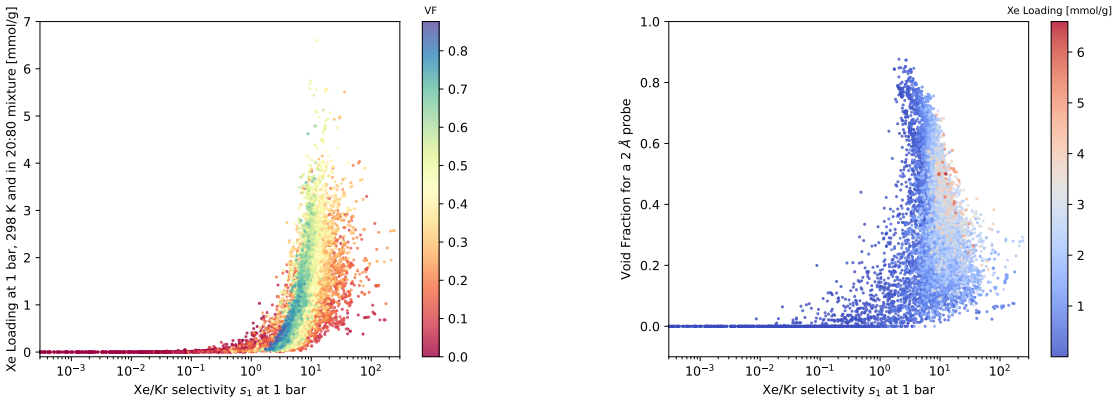


Figure 2.7: On the left: scatterplot of the xenon uptake as a function of the selectivity and labeled by the values of the void fraction. On the right: scatterplot of the selectivity and the void fraction labeled by the quantity of xenon adsorbed. The selectivity and uptake are calculated by a GCMC simulation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

By conducting a similar analysis to that performed for surface areas, but this time focusing on the void fraction using Figure 2.6 (right), it is possible to identify different intervals of void fractions that correspond to highly selective materials. For instance, selectivity values above 92 correspond to materials with a porosity ranging from 0% to 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 ranging from 0% to 50.0% and more frequently concentrated between 12.5% and 37.5%. Selectivity values between 35 and 13 can be observed in materials with a void fraction ranging from 0% to 75.0%, with a bell-shaped distribution centered around 31%. As selectivity values decrease, the peak of the distribution shifts towards higher void fraction values, indicating a preference for lower porosity (below 25%) in terms of selectivity performance. However, similar to surface areas, the void fraction is not a deterministic variable — the void fraction alone does not determine the material's performance. Therefore, it is necessary to investigate whether adding another variable, such as pore size, as a joint variable, can provide a better characterization of the material's performance. As a temporary conclusion, the most selective materials are little porous with a void fraction not exceeding 0.5 and with an internal surface lower than 2500 $m^2\,cm^{-3}$. These materials have pores that are specialized for xenon adsorption, which will be confirmed by the following discussion.

Pore Size and selectivity

The optimal pore size for xenon/krypton separation can be deduced from Figure 2.8. As shown on the right plot, the most selective materials have a pore size close to 5 Å. However, it is challenging to distinguish between materials with very low selectivity that have a similar label color. To better visualize the differences, different colors were used to represent structures with D_i values ranging from 3.6 to 11.6 Å. It becomes evident that the pore size should be around

5 Å, but if it is too small (near 4.5 Å),the selectivity significantly decreases. Therefore, there exists a sweet spot of pore size values that enable the attainment of very high selectivity.

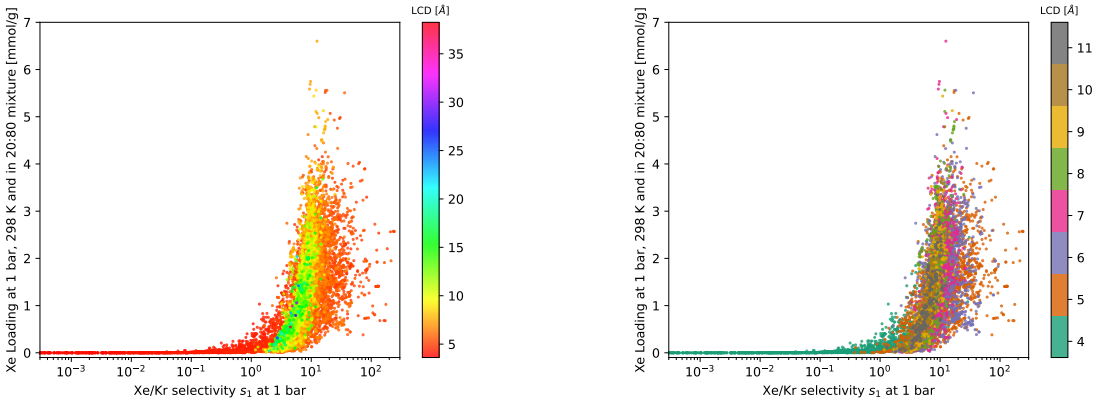


Figure 2.8: Scatterplot of the xenon uptake as a function of the selectivity (20:80) and labeled by the values of LCD_{UFF} (left). The same scatterplot restricted to values of D_i between (3.6 and 11.6 Å) and labeled using a different color code to distinguish the most selective materials from the least 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.

The joint effects of void fraction and largest cavity diameter (D_i) on selectivity reveal a distinct region in the bidimensional descriptor space where the most selective materials are located. On Figure 2.9, structures with a selectivity above 10 are highly likely to have a void fraction below 0.4 with a relatively wider range of D_i. However, as shown on the filtered version of the plot (on the right), the most selective materials (over 40) exist within a very narrow range of D_i values, approximately between 4.8 and 6 Å. This can be attributed to the xenon atom size, which closely matches these D_i values, which allows a maximal stabilization of xenon. On the other hand, krypton, being slightly smaller, exhibits less favorable interaction with the pores, resulting in higher observed selectivity.

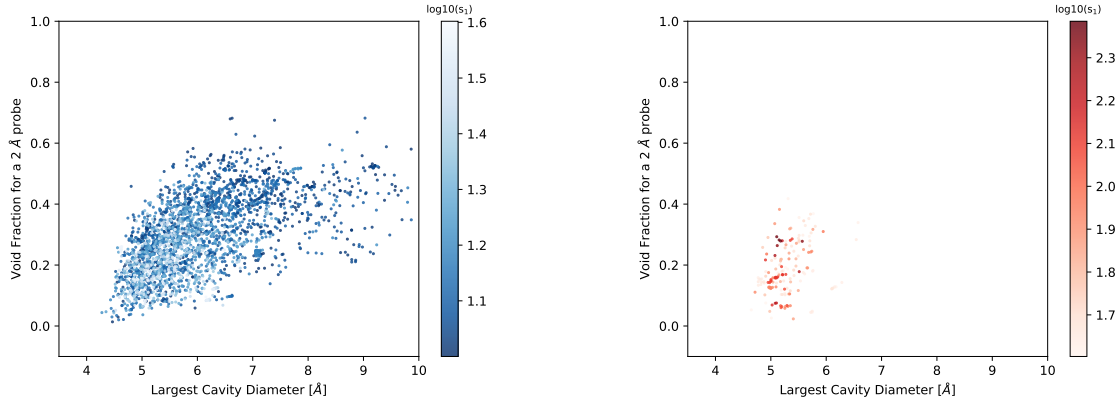


Figure 2.9: Scatterplots of the void fraction as a function of the LCD_{UFF} and labeled by the $\log_{10}$ of the selectivity values. On the left, only the materials with a selectivity between 10 and 40 are considered; and on the right, selectivity values over 40.

As presented in Chapter 1, Simon et al. found that the most selective materials have a pseudo-spherical shape with a size close to the diameter of a xenon. These materials tend to have a dense structure with limited porosity. In this slightly different approach specific intervals of cavity diameter and pore volume have been associated with high selectivity, thus confirming the size requirement already identified by other studies. However, this structure–property relationship serves as a description tool for identifying selective materials, and it does not enable accurate predictions based solely on structural descriptors.

Effect of the composition

The previous analyses focused on a specific type of mixture composition (20:80) associated with the extraction of xenon and krypton from air through cryogenic distillation (see section 1.3.1). In the following section, the effects of composition will be investigated by considering the case of xenon/krypton separation in spent nuclear fuel off-gases. In nuclear applications, the mixture has a higher xenon content than in the previous one, with a typical 90:10 Xe/Kr ratio. For this reason, the quantity of xenon adsorbed in the materials will mechanically be higher compared to the previous scenario. However, the second quotient in the formula of selectivity in equation 2.8 compensates for the inherently higher first quotient. The objective of this analysis is to evaluate these two effects and determine whether they offset each other or if different trends emerge depending on the composition value.

As depicted in Figure 2.10, the selectivity values of both compositions are relatively similar. However, a slight decrease in performance can be observed when increasing the proportion of xenon in the mixture, particularly for materials with moderate selectivity (s between 2 and 50). This decline in performance can be attributed to variations pores displayed by the material, which possess different affinities for xenon. When the xenon proportion is lower, Xe adsorbates preferentially access the most favorable pores, resulting in a concentration of the small quantity xenon in these sites. However, as the xenon content increases, these most favorable sites become saturated, and xenon needs to compete with krypton for less favorable sites, thereby slightly decreasing the selectivity.

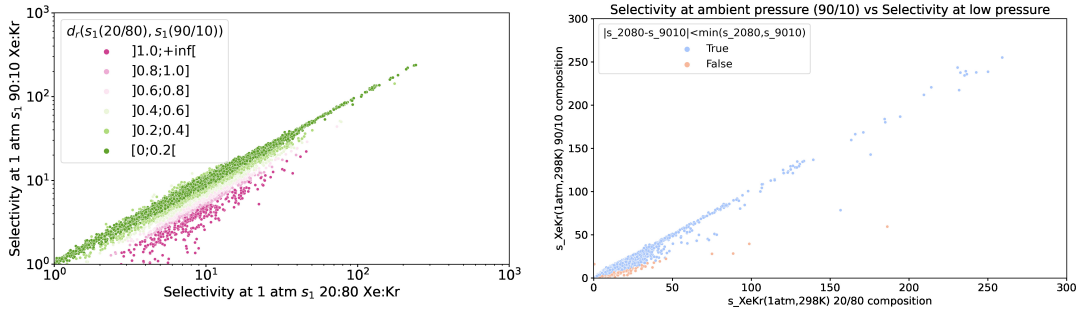


Figure 2.10: Illustration (scatterplot) of the difference of selectivity ($s_1(20:80)$ and $s_1(90:10)$) for two different Xe/Kr mixture compositions 20:80 (x-axis) and 90:10 (y-axis) at 1 atm *and* 298 K. *On the left, the axis is in log scale and the relative difference of selectivity between the two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

The effect of the composition on the different analyses of the different structural descriptors will be discussed here. Notably, when considering a mixture with a higher xenon content, the xenon uptake values experience significant change. The nanopores of selective materials ($1 < s_1 \leq 50$)

are much more saturated with Xe, resulting in a substantially higher maximum xenon uptake. Comparing the Figures 2.7 and 2.11, the maximum uptake increases from 6.6 mmol g^{-1} (for the 20:80 composition) to 11.7 mmol g^{-1}. In the case of moderately selective materials at the 20:80 composition, the xenon competes with krypton primarily in the most selective nanopores. However, with a higher xenon content, xenon has to compete with krypton in much less favorable sites due to saturation of the most preferable sites. It is worth noting the previous conclusion regarding the maximum uptake of xenon for the most selective materials (s_1 > 50) remains valid. The maximum xenon uptake reaches up to 4.0 mmol g^{-1}, and it increases slightly to 4.2 mmol g^{-1} for the composition with a higher xenon content. Despite the change in composition, the nature of the adsorbed state remains unchanged due to the extremely high selectivity, resulting in similar quantities of xenon being present in the pores. Consequently, the higher xenon content does not significantly influence the performance of the most selective materials, but it can alter selectivity and greatly increase the xenon uptake for some moderately selective materials.

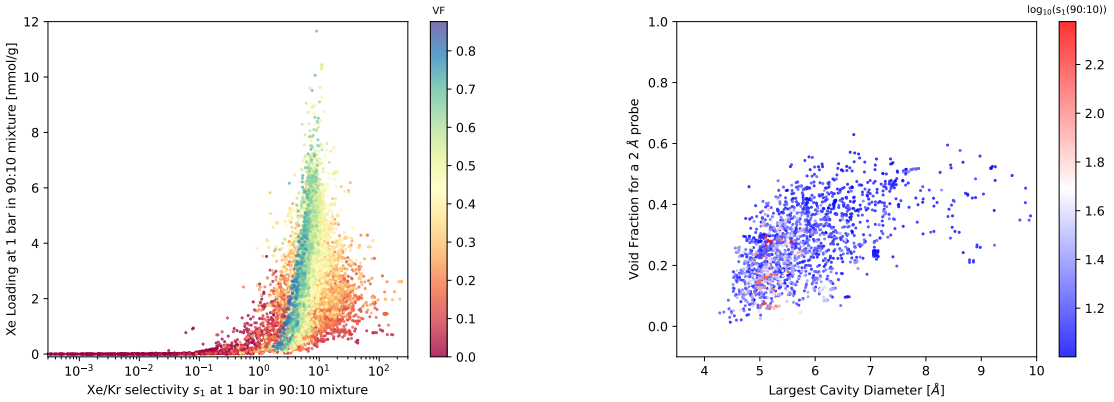


Figure 2.11: Illustration of the effect of the composition by representing the same figures as in 2.7 and 2.9 but for a 90:10 composition. On the left: scatterplot of the xenon uptake as a function of the selectivity ($s_1(90:10)$) and labeled by the values of the void fraction. On the right: scatterplots 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.*

As a conclusion, the composition does not seem to affect the previously determined structural characteristics required for high selectivity. As depicted in the right plot of the Figure 2.11 (right), the most selective materials still exhibit a pore size of approximately 5 Å and a porosity below 40%. This structural domain serves as necessary conditions for achieving selectivity, but they are not sufficient as less selective materials can also display these characteristics. Having established the geometric conditions necessary for obtaining good selectivity, the focus now shifts towards understanding the thermodynamic origins of selectivity by examining energy-based quantities and the various correlations among them.

2.2.2 Thermodynamic quantities correlations at infinite dilution

In this section, my goal is to map out the details of the thermodynamic features of Xe/Kr adsorption and separation in nanoporous materials, rather than to directly address the structure–property relationships. The high-throughput screening methodology was used to map out the space of thermodynamic properties, surpassing the conventional metrics of selectivity and

uptake. The specific emphasis was placed on investigating the role of adsorption enthalpy and entropy, differentiating between Xe and Kr adsorption thermodynamics, and analyzing the variations in selectivity at both low and high pressures. The discussion below is based on a work published[1] in the Faraday Discussions Ref. [**Ren_2021**].

To assess the performance of a given nanoporous material for separation in the low loading (or low pressure) limit, Henry constants are commonly calculated by linearly fitting low-pressure adsorption isotherm data — both experimentally and computationally. In this section, the thermodynamics of Xe and Kr adsorption at low pressure are investigated. Specifically, the low-pressure adsorption properties are obtained using the Widom insertion method**Widom1963**, **frenkel2001widom** on a dataset of 9 668 selected structures. This method provides higher accuracy compared to fitting isotherms, where it can be challenging to determine the extent of the linear adsorption regime. Through these simulations, Henry constants K and adsorption enthalpies $\Delta_{ads}H_0$ (at the zero loading limit) are calculated 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 constants for the two gases. In the following discussion, the statistical relationships among the thermodynamic quantities at low pressure, namely s_0, K^{Xe}, K^{Kr}, $\Delta_{ads}H_0^{Xe}$, $\Delta_{ads}H_0^{Kr}$ and $\Delta_{exc}H_0$, will be examined.

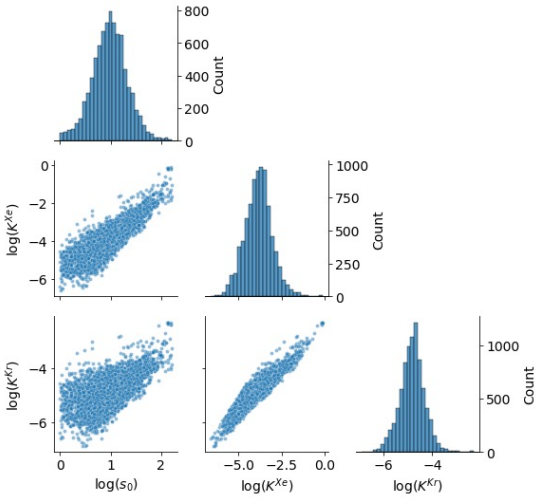


Figure 2.12: For 8 401 MOFs with favorable thermodynamic Xe/Kr selectivity ($s_0 > 1$), pair plots of $\log_{10}(s_0)$, $\log_{10}(K^{Xe})$ and $\log_{10}(K^{Kr})$ (the Henry constants are in $\mathrm{mmol\,g^{-1}\,Pa^{-1}}$) in the off-diagonal subplots (note that the y-axis is displayed on the right 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).

The distribution of thermodynamic properties of materials with favorable thermodynamic Xe/Kr selectivity ($s_0 > 1$) is depicted in Figure 2.12. It is important to note that the plots focus on selectivity values above 1, as these are the materials of interest for separation. This selection eliminates certain outliers with specific geometries or binding sites (without significantly altering the overall conclusions). The plots reveal that while the logarithm of Xe Henry constant K^{Xe} exhibits a weak correlation with the logarithm of selectivity s_0, this correlation is stronger for highly selective materials. Therefore, in a multistep screening study aimed at

[1] The related data can be found at https://github.com/fxcoudert/citable-data/tree/master/132-Ren_FaradayDiscuss_2021

identifying the most selective materials, it could be possible to utilize as a "first filter" criterion based purely on Xe adsorption, by excluding materials below a certain threshold (e.g., the materials with $s_0 \geq 30$ can be found in the subset with $K^{Xe} \geq 2.7\,10^{-1}$ mmol g^{-1} Pa^{-1}). On the other hand, the correlation between K^{Kr} and s_0 is relatively weaker. These results suggest that a high affinity with xenon measured by the Henry constant is a determining factor of high selectivity for the most selective materials.

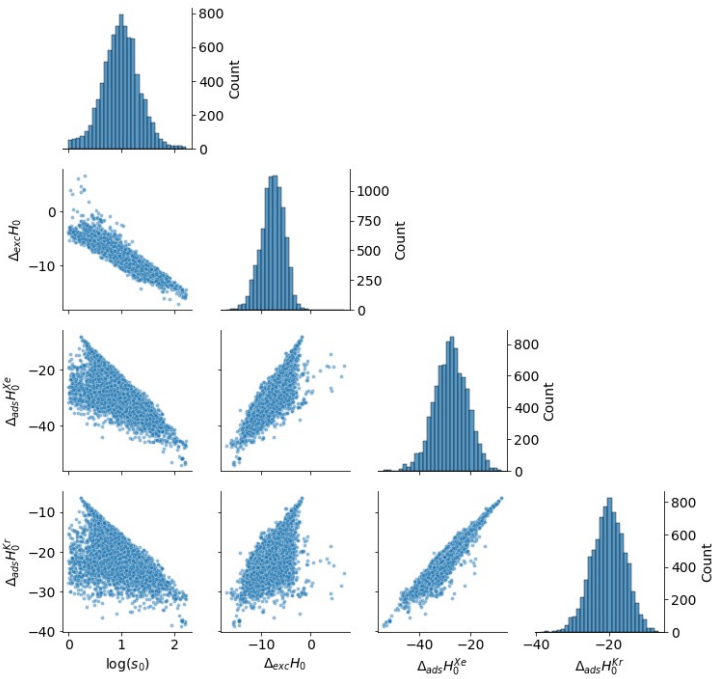


Figure 2.13: For 8 401 MOFs with favorable thermodynamic Xe/Kr selectivity ($s_0 > 1$), pair plots of $\log(s_0)$, $\Delta_{exc}H_0$, $\Delta_{ads}H_0^{Xe}$ *and* $\Delta_{ads}H_0^{Kr}$ *(the enthalpies are in* kJ mol^{-1}*) in the off-diagonal subplots and the distribution of each quantity is on the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

In terms of Henry constants, a diverse range of behaviors is observed, 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}. Additionally, there is a statistical trend indicating that a high affinity for xenon typically translates into a relatively high affinity for krypton. This general trend is observed for noble gases, where the adsorption sites lack strong specificity. To delve deeper into the thermodynamic aspects underlying this wide diversity in behavior, the enthalpies involved were plotted in Figure 2.13.

The low-loading adsorption enthalpy of xenon ($\Delta_{ads}H_0^{Xe}$) is strongly correlated with that of krypton ($\Delta_{ads}H_0^{Kr}$), which aligns with the correlation observed between their respective Henry constants. This suggests the involvement of a rather generic physisorption mechanism in the majority of materials, where the host–adsorbate affinities are primarily determined by the enthalpy. The selectivity of Xe/Kr selectivity is not driven significantly by the xenon or krypton adsorption enthalpy alone (both exhibit weak correlation with selectivity), but rather by their difference, $\Delta_{exc}H_0$, which shows a strong correlation with $\log(s_0)$. This finding is further supported by the lack of correlation between selectivity and adsorption entropies (*cf.* Figure 2.14), indicating that the separation is predominantly enthalpic in nature, and

any dispersion in the correlation between selectivity $\log(s_0)$ and $\Delta_{\text{exc}}H_0$ is influenced by entropy.

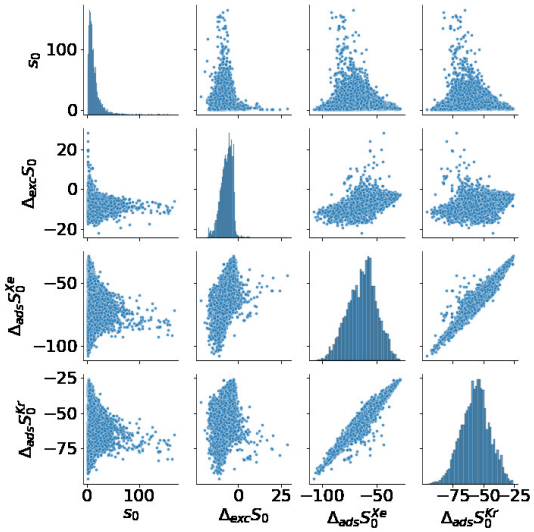


Figure 2.14: For 8 401 MOFs with favorable thermodynamic Xe/Kr selectivity ($s_0 > 1$), pair plots of s_0, $\Delta_{exc}S_0$, $\Delta_{ads}S_0^{Xe}$ and $\Delta_{ads}S_0^{Kr}$ in the off-diagonal subplots and the distribution of each quantity are on the diagonal.

Upon closer analysis of the Figure 2.14, it is observed that the adsorption entropy of xenon and krypton shows a noticeable correlation. However, their difference (the exchange entropy), which represents the exchange entropy, does not exhibit a significant variance value (Figure 2.15) compared to the enthalpy. This suggests that the thermodynamic quantity plays a minor role in the selectivity performance of the materials. However, it is noted that although the most selective materials do not have any exchange entropy values, they are centered around a value of approximately $-10\,\text{kJ}\,\text{mol}^{-1}\,\text{K}^{-1}$. While this correlation is not straightforward, it indicates that possessing an exchange entropy within this range is a necessary attribute for achieving selectivity in materials.

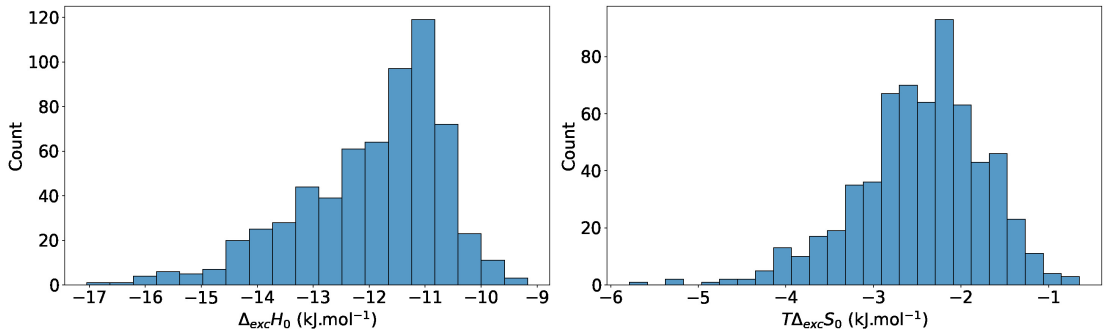


Figure 2.15: Distribution of the enthalpy $\Delta_{exc}H_0$ and entropy $T\Delta_{exc}S_0$ of exchange at low pressure on the 630 most selective structures

To further emphasize the enthalpic nature of the separation process, the base-10 logarithm of the Henry constant (proportional to the adsorption free energy) is compared to the adsorption enthalpy for both xenon and krypton. As shown in Figure 2.16, the free energy can be predominantly explained by the enthalpy, which confirms the secondary role played by entropy in accounting for the dispersion in this relationship. The effect of entropy weakens

the correlation for materials with less favorable adsorption, but as the adsorption enthalpies become increasingly negative, the correlation becomes increasingly stronger. The most selective materials have an almost negligible entropic contribution to the final free energy value (G = H – TS).

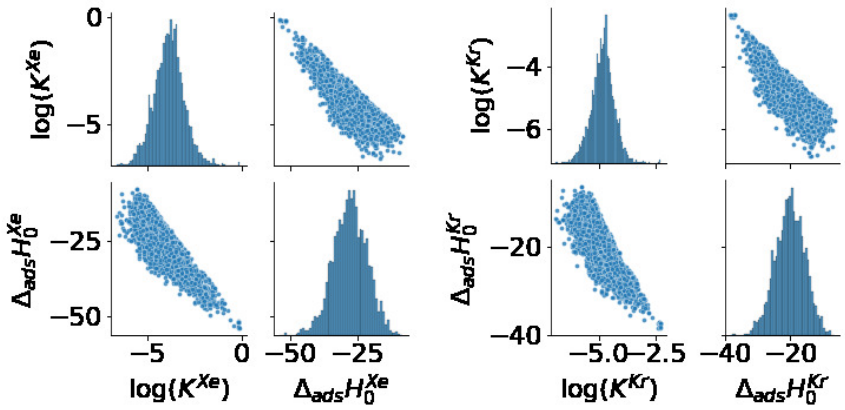


Figure 2.16: For 8 401 MOFs with favorable thermodynamic Xe/Kr selectivity (s_0 > 1), pair plots of $\log(K_H^i)$ *and* $\Delta_{ads}H_0^i$ *in the off-diagonal subplots for both i=Xe and i=Kr and the distribution of each quantity are on the diagonal.*

Upon further analysis of Figure 2.17, it becomes apparent that the entropic effect is influenced by the pore size. Specifically, larger pore sizes tend to yield more positive entropic terms (the entropic term refers to $-T\Delta_{ads}S$). This observation elucidates the weaker correlation observed for less attractive materials throughout the pairplot of Figures 2.13 and 2.12.

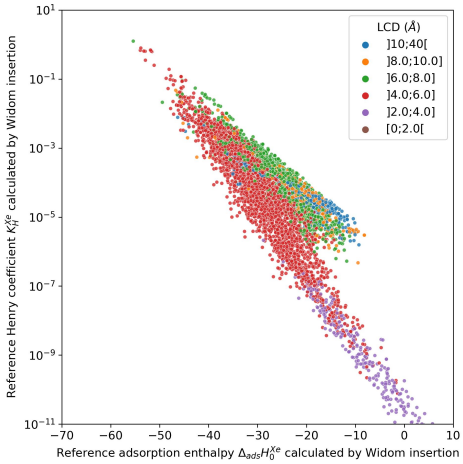


Figure 2.17: Comparison between the Xe Henry constant and Xe adsorption enthalpy labeled by categories of LCD_{UFF} *values for the CoRE MOF structures.*

In the analysis of the influence of the pore size and void fraction on the entropic term $T\Delta_{ads}S_0^{Xe}$ (Figure 2.18), a clear relationship between entropy and pore size, represented by the LCD_{UFF}[1], is observed. Larger pores tend to exhibit higher entropy, likely due to the confinement effect of the pore — a small pore limits the available adsorption positions for xenon, while a larger pore provides more sites for adsorption. A similar trend is observed for pore volume, represented

[1]This corresponds to the diameter of the largest included sphere defined by UFF-based atom radii

by the void fraction here. A weak linear correlation exists between the void fraction (in log-scale) and the adsorption entropic term of xenon. However, it is important to note that these simple geometric descriptors may not capture the entire complexity of the entropic behavior, particularly for larger pore sizes. Other effects that can contribute to the entropic effects include the shape of the channel and cavities (e.g., tortuosity) or the overall distribution of pore sizes that cannot be adequately captured by the LCD_{UFF}.

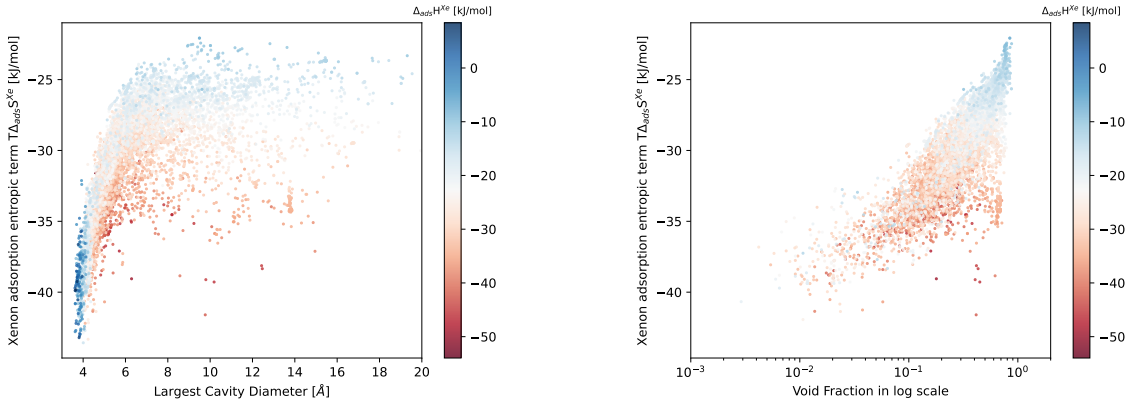


Figure 2.18: Comparison plots of the entropic term $T\Delta_{ads}S_0^{Xe}$ at infinite dilution and two geometric descriptors: the LCD_{UFF} (left) and the void fraction (right).

Cross-referencing these findings with the previous results obtained on the influence of geometric descriptors in the section 2.2.1, it becomes evident that the entropic effect aligns with the enthalpic term in explaining selectivity when the pore size approaches that of xenon. The confinement of xenon within the pores leads to lower entropy in the adsorbed phase compared to the gas phase, especially for pores tailored to xenon's size. Furthermore, the optimal interaction between xenon and the surrounding framework atoms reduces the enthalpic term. Both factors work in concert, elucidating the optimal selectivity observed for this particular pore size value (around 5 Å).

The key takeaways from this section revolve around two relationships. Firstly, there is a correlation between the Henry constant of xenon and selectivity, allowing for rough estimation of Xe/Kr selectivity — the most selective materials exhibit a strong affinity for xenon. Secondly, the selectivity process is primarily driven by enthalpy — the separation process has an enthalpic nature as a first-order approximation, which is particularly true for the most selective materials. Analyzing the energy interactions within the material provides crucial insights into its performance. While the focus here has been on thermodynamic properties at infinite dilution, the subsequent section will delve into the impact of pressure on selectivity, specifically examining a 20:80 Xe/Kr mixture at 1 atm and 298 K.

2.3 Selectivity drop between two pressure regimes

As the previous section has established the relationship between selectivity and some geometrical and thermodynamic descriptors, this section focuses on examining the relationship between selectivity values at infinite dilution and selectivity values at ambient pressure using a thermodynamics-based approach. The aim is to gain a better understanding of the underlying mechanisms driving the observed changes in selectivity as previously discussed in Figure 2.10.

It is worth noting that the findings presented in this section have already been published in Ref. [**Ren_2021**]

2.3.1 Thermodynamic origins

After delving into the thermodynamics of the infinite dilution case, the focus now shifts to examining the impact of changes in working pressure on adsorption selectivity and analyzing its underlying thermodynamic mechanisms. Understanding the impact of pressure on selectivity is crucial for accurately assessing adsorption thermodynamics under different working conditions, particularly in specific industrial processes. Insights into the pressure-dependent selectivity may allow for faster screening of materials limited to specific thermodynamic conditions.

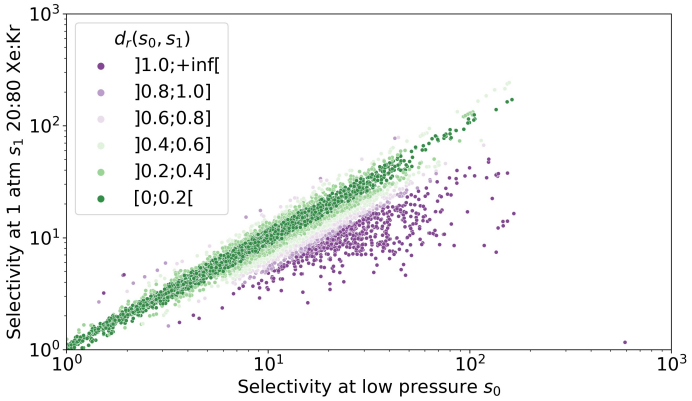


Figure 2.19: Difference of selectivity between low pressure and at a 1013 hPa *pressure for a 20:80 xenon/krypton composition. The relative difference between the low-pressure selectivity and the ambient pressure is particularly high for the points labeled in purple.*

The selectivity s_1 was calculated at a pressure of 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**]). It was observed that for high-selectivity materials, the composition had a minimal impact, as shown in (*cf.* Figure 2.10). In the following analysis, the focus is primarily on the selectivity for the 20:80 mixture, which is the most commonly studied composition in the literature. To quantify the difference in selectivity between low and ambient pressures, a relative difference $d_r(s_0, s_1)$ is considered, as defined in equation 2.31.

$$d_r(s_0, s_1) = \frac{|s_0 - s_1|}{\min(s_0, s_1)} \tag{2.31}$$

Figure 2.19 presents the selectivity at ambient pressure s_1, plotted against its low-pressure counterpart s_0 (for materials where $s_0 > 1$, as before). The points on the plot are color-coded according to the values of $d_r(s_0, s_1)$, which are divided into 6 discrete categories for clarity. A broad correlation is observed, particularly near the diagonal line where approximately 61.5% of the materials exhibit a difference below 20% (close to the $s_0 = s_1$ line). However, it is notable that there are considerably more points, approximately (74.3% of the materials with $d_r(s_0, s_1) \geq 0.2$) below the first bisector ($s_1 < s_0$), indicating that for these materials, the selectivity s_1 at 1 atm is significantly lower than the selectivity s_0 at low pressure.

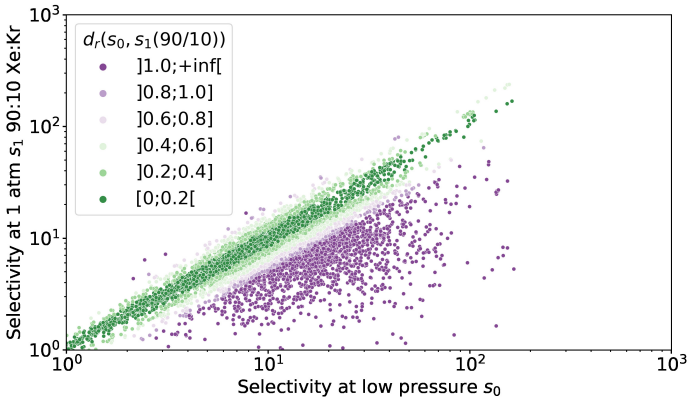


Figure 2.20: Difference of selectivity between low pressure and at a 1013 hPa *pressure for a 90:10 xenon/krypton composition. The relative difference between the low-pressure selectivity and the ambient pressure is particularly high for the points labeled in purple.*

This drop in selectivity primarily affects materials with a relatively high selectivity $s_0 > 10$ (Figure 2.19). It highlights the potential pitfalls of relying solely on pure-component Henry's constant (i.e., zero-pressure selectivity) for materials screening. While calculating low-pressure selectivity is simpler and faster, it can lead to overestimated selectivity by more than 100% in a significant number of materials (646 out of 9,668 in our dataset). To understand the underlying reasons for these shifts in selectivity, a thermodynamic approach will be employed.

Before delving deeper in the analyses of the thermodynamic origins of this pressure-induced selectivity drop, this paragraph will open a small aside on the 90:10 mixture composition. When examining the 90:10 composition, it becomes apparent that the drop in selectivity is even more pronounced. The selectivity for the higher xenon proportion was already found to be higher than that for the 20:80 composition (Figure 2.11). This drop can be attributed to the presence of more or less favorable adsorption sites. In some materials (labeled in purple), at low xenon content composition, xenon and krypton primarily compete for the most favorable sites until these sites become saturated, leaving no xenon to compete in the less selective sites. As the Xe/Kr ratio increases, these less selective nanopores contribute to an overall decrease in selectivity. When combined with the effect of increased pressure, certain materials undergo both phenomena, resulting in a more pronounced drop in selectivity compared to the selectivity at low pressure. These explanations are backed up by the following analyses on the pressure effect, which highlights similar effects of the total pressure (instead of partial pressure) on selectivity — a higher xenon content is actually equivalent to increasing the partial pressure of xenon. Now that the effect of higher xenon content has been characterized, the following analyses will be based on the 20:80 Xe/Kr composition.

To quantitatively assess the thermodynamic effects involved in the competitive adsorption under different regimes (for the 20:80 composition), thermodynamic properties of the "exchange equilibrium" predefined in equation 2.25 are considered. Figure 2.21 displays a scatterplot of the exchange entropy at low pressure, represented as $T\Delta_{exc}S_0$, plotted against the exchange enthalpy $\Delta_{exc}H_0$. The points on the plot are color-coded according to the selectivity s_0 (with discrete categories for clarity), which is related to the enthalpy and entropy through Equation 2.30 — indicating iso-selectivity lines correspond to parallel straight lines in this scatterplot.

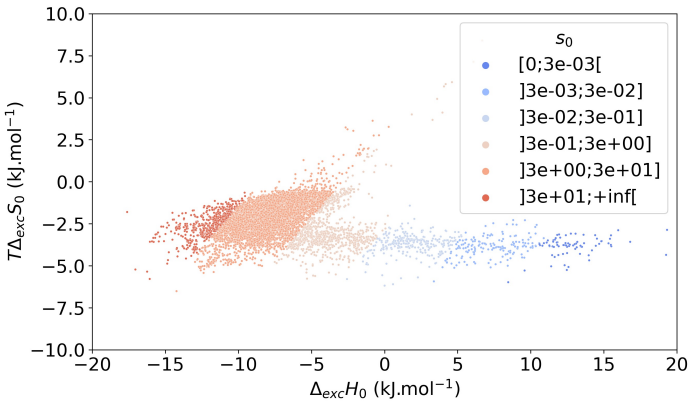


Figure 2.21: The energetic equivalent of exchange entropy $T\Delta_{exc}S_0$ *and enthalpy* $\Delta_{exc}H_0$ *at low pressure labeled using the selectivity* s_0 *at low pressure (for any xenon/krypton composition). The limit between labels follows an affine function of slope* 1/T *and of intercept* $-R\ln(s_0^{lim})$ *where* s_0^{lim} *is the limit selectivity value (*cf. *equation (2.30)). 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 2.15 presents the distributions of the exchange enthalpy and entropy at low pressure. Among the 630 most selective materials ($s_0 > 30$), the distribution of the exchange enthalpy $\Delta_{exc}H_0$ is centered around $-12.0\,kJ\,mol^{-1}$ with a standard deviation of $1.3\,kJ\,mol^{-1}$. On the other hand, the distribution of the exchange entropy (represented as $T\Delta_{exc}S_0$) is centered around $-2.5\,kJ\,mol^{-1}$, with a standard deviation of $0.7\,kJ\,mol^{-1}$. These figures, along with the overall distribution plotted in Figure 2.21, provides further evidence of the relatively modest role of entropy in determining the selectivity at low pressure, which corresponds, on average, to approximately 20% of the exchange enthalpy.

Examining Figure 2.22 for the selectivity at ambient pressure, similar conclusions can be drawn regarding the limited influence of entropy on selectivity values. The distribution of the entropic term $T\Delta_{exc}S_1$ is centered around $-3\,kJ\,mol^{-1}$, which remains relatively small compared to the values of $\Delta_{exc}H_1$. For the most selective materials, the entropic term represents approximately 19% of the exchange enthalpy $\Delta_{exc}H_1$ at ambient pressure.

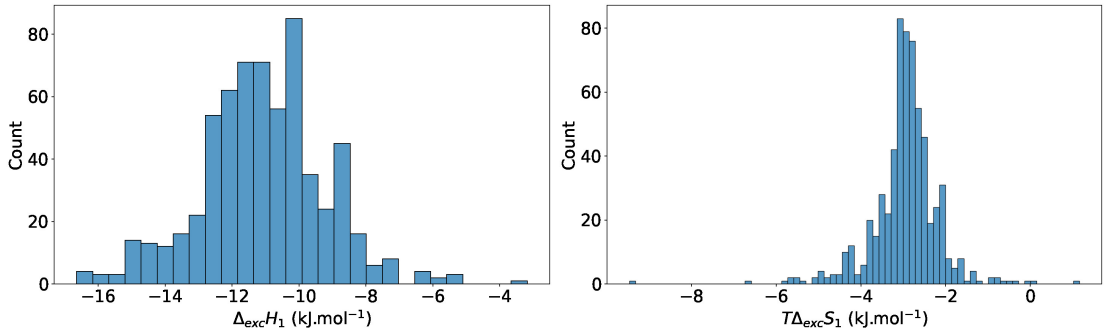


Figure 2.22: Distribution of the enthalpy $\Delta_{exc}H_1$ *and entropic term* $T\Delta_{exc}S_1$ *of exchange at ambient pressure on the 630 most selective structures.*

Figure 2.23 represents a scatterplot of the exchange entropy at P = 1 atm $\Delta_{exc}S_1$ against the exchange enthalpy at ambient pressure $\Delta_{exc}H_1$. The points are color-coded according to the low-pressure selectivity s_0 to compare it with the Fig. 2.21. In comparison to the iso-selectivity

s_1 straight parallel lines (*cf.* Figure 2.24), it can be observed that many materials with high s_0 have lower s_1, as indicated by a migration of points to the right of the plot. This shift is thus mainly due to a higher (less favorable) exchange enthalpy, implying that enthalpy plays a crucial role in determining selectivity at higher pressures.

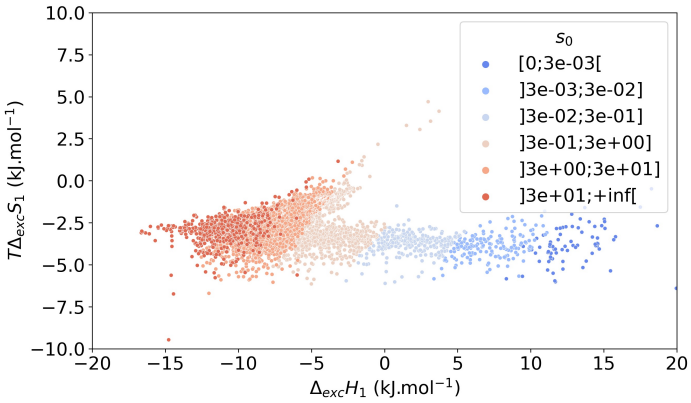


Figure 2.23: The energetic equivalent of exchange entropy $T\Delta_{exc}S_1$ and enthalpy $\Delta_{exc}H_1$ at ambient pressure (for a 20:80 xenon/krypton composition) labeled using the selectivity s_0 at low pressure. The points are layered so that the points with higher s_0 are always above. To see a split version of this plot, please refer to Figure 2.24.

To quantify this change, 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 2.22) are considered. The enthalpy distribution $\Delta_{exc}H_1$ is now centered at $-11.1\,kJ\,mol^{-1}$ with a standard deviation of $1.9\,kJ\,mol^{-1}$. In comparison to the zero-pressure values, the enthalpy distribution exhibits greater dispersion, suggesting significant changes in individual values and an overall increase in average enthalpy. Most materials display lower selectivity at ambient pressure due to enthalpic effects, which can be attributed to the general increase in adsorption enthalpy with increasing gas phase loading, which is linked to the presence of more adsorbed molecules. The correlations shown in Figure 2.12 suggest that highly selective materials have a high affinity for xenon, resulting in substantial uptake at 1 atm. The large Xe loading means the saturation of the most favorable adsorption sites and the subsequent adsorption of weaker host–guest interactions contribute to an overall increase in average adsorption enthalpy at non-zero loading.

The entropic term $T\Delta_{exc}S_1$ is now centered at $-2.9\,kJ\,mol^{-1}$, with a standard deviation of $0.8\,kJ\,mol^{-1}$ (almost unchanged from low-pressure). The average entropy is lower, indicating a less favorable separation overall due to entropic effects. This evolution of the entropic term suggests the possibility of a reorganization of adsorbed molecules within each material. However, the difference in enthalpy distribution has a greater impact on high-pressure selectivity compared to the distribution of entropy. Thus, the overall contribution of enthalpy appears to be more decisive than the role of entropy in determining selectivity changes, even at ambient pressure. This finding is significant for screening studies, as evaluating adsorption enthalpy computationally is generally faster than determining adsorption free energy (or entropy).

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

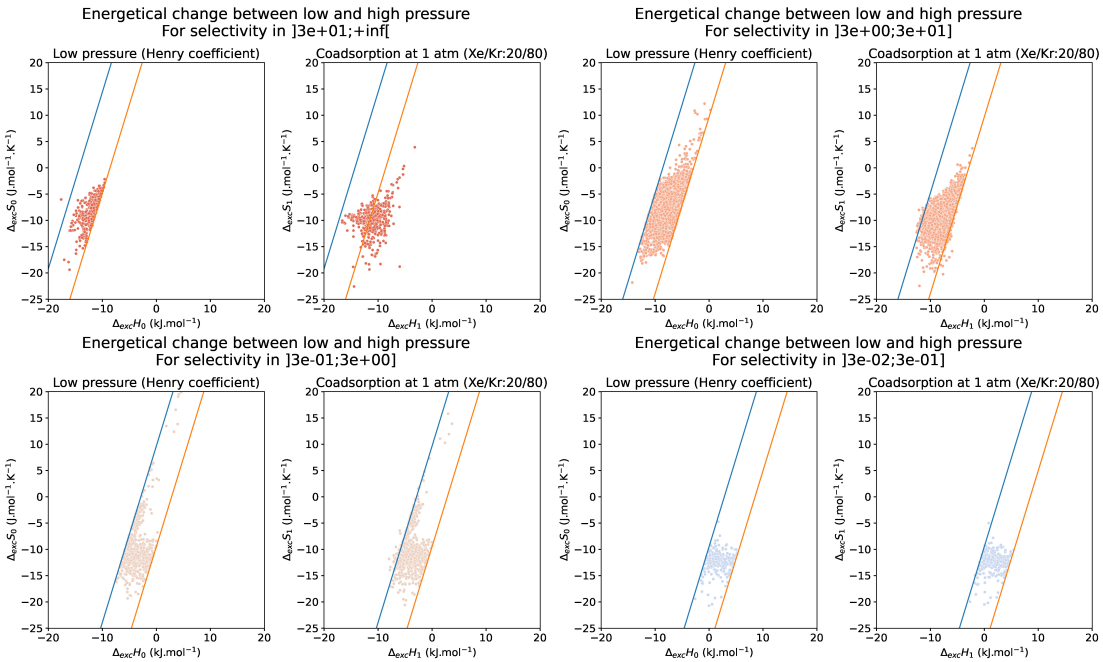


Figure 2.24: Split view of Figure 2.21 and 2.23. The iso-selectivity lines for the limit considered are represented with blue and orange lines. It seems that the shift in exchange enthalpy for the structures with a selectivity higher than 30.

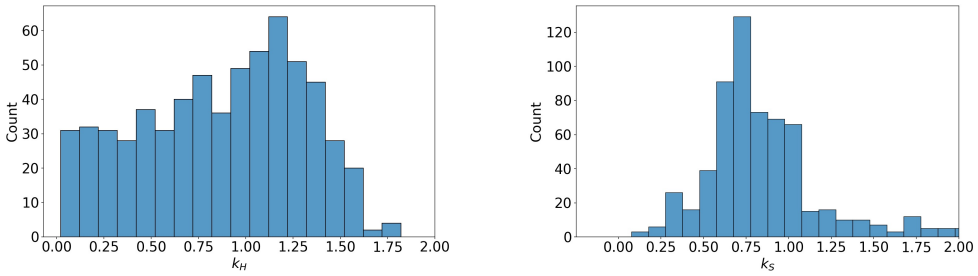


Figure 2.25: Distribution of the enthalpic k_H and entropic k_S contributions to the change of selectivity from low to ambient pressure for the 630 materials with $s_0 > 30$. k_H has a rather broad and uniform distribution, whereas k_S has a bell-like distribution.

where k_{H} and k_{S} are the enthalpic and entropic contributions to the selectivity change defined as:

$$k_{\mathrm{H}} = \exp\left(-\frac{\Delta_{\mathrm{exc}}\mathrm{H}_1 - \Delta_{\mathrm{exc}}\mathrm{H}_0}{\mathrm{RT}}\right)$$
$$k_{\mathrm{S}} = \exp\left(\frac{\Delta_{\mathrm{exc}}\mathrm{S}_1 - \Delta_{\mathrm{exc}}\mathrm{S}_0}{\mathrm{R}}\right) \tag{2.32}$$

As depicted in Figure 2.25, the entropic contribution k_{S} has a bell-shaped distribution, with a mean of 0.9 and a standard deviation of 0.6. This confirms that k_{S} is close to 1, indicating that it has only a marginal effect on the selectivity change. In contrast, the enthalpic contribution k_{H} has a more uniform distribution ranging from 0.1 to 1.5, which means that enthalpy plays a crucial role in the observed selectivity change. Notably, there exists a significant number of materials with k_{H} close to zero, corresponding to the same materials highlighted in Figure 2.19.

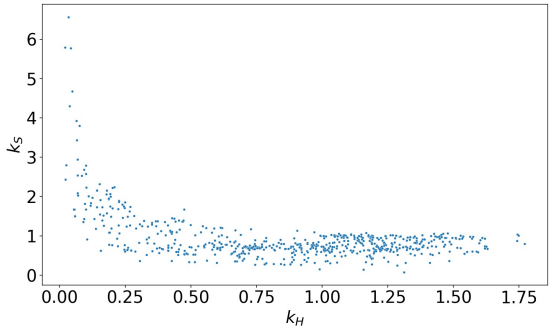


Figure 2.26: Scatterplot of the enthalpic contribution k_H and entropic contribution k_S for the 630 materials with $s_0 > 30$. The entropic compensation occurs when the enthalpic contribution is around 0.1, else its value is around 1 and has little effect on the selectivity change.

Furthermore, the scatterplot of k_H and k_S (Figure 2.26) confirms the relatively moderate effect of entropy. For most materials with $0.25 \leq k_\mathrm{H} \leq 1.75$, k_S is close to 1. The most significant entropic contributions are observed for materials where k_H is close to zero (typically below 0.25). Examining the 29 materials with $k_\mathrm{S} > 2$ in more detail, it appears that the entropic contribution k_S moderately compensates for the enthalpic contribution, resulting in an average ratio of s_1/s_0 around 0.25. In such cases, entropy is non-negligible and can partially offset the enthalpic contribution to the selectivity change. However, the overall trend is still dictated by enthalpy, as the overall selectivity decreases as a result.

2.3.2 Detailed investigation

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

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.

This section will examine the most selective materials identified at low pressure, as listed in Table 2.1, and provide a detailed investigation of the thermodynamic effects underlying their behavior. These materials can be divided into three main categories: materials exhibiting 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, the origins of these different behaviors will be investigated, with reference to the CSD refcodes of the materials.

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

Table 2.2: Thermodynamic quantities associated for a few archetypal structures. Henry 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* $\mathrm{kJ\,mol^{-1}}$

Before delving into the different archetypal structures that undergo different selectivity changes, it is necessary to introduce the fundamental concept of adsorption isotherms. The latter can be understood as a plot of the adsorbed quantity as a function of pressure for different components at a given temperature. The following discussion will only focus on the case of pure-component isotherms at 298 K. Various models have been developed to interpret these plots,**Al_Ghouti_2020** but for the purpose of this study, the Langmuir model will be used exclusively. The Langmuir model is the most well-established local adsorption model that describes the filling of a monolayer by non-interacting adsorbates. Depending on the pore distribution and shape, the isotherms can be effectively modeled by either a 1-site Langmuir or a 2-site Langmuir model, for the simplest cases. At given temperature, certain single site materials' isotherm can accurately be described by the following equation:

$$q(\mathrm{P}) = \mathrm{N_{max}}\frac{\mathrm{KP}}{1+\mathrm{KP}} \tag{2.33}$$

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 adsorption sites, the isotherm can be described by the following equation:

$$q(\mathrm{P}) = \mathrm{N_{max}}\left((1-\alpha_2)\frac{\mathrm{K_1P}}{1+\mathrm{K_1P}} + \alpha_2\frac{\mathrm{K_2P}}{1+\mathrm{K_2P}}\right) \tag{2.34}$$

where q is the total 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.

This section will 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 the material VOKJIQ,**VOKJIQ**

an open-framework aluminophosphate, $[HAl_3P_3O_{13}]\cdot C_3NH_{10}$, the selectivity increases by a factor of 1.5 between low and ambient pressure. Upon closer examination, it is observed that the adsorption enthalpy of xenon $\Delta_{ads}H^{Xe}$ decreases from $-53.9\,kJ\,mol^{-1}$ to $-61.1\,kJ\,mol^{-1}$, whereas the adsorption enthalpy for krypton $\Delta_{ads}H^{Kr}$ decreases from $-38.2\,kJ\,mol^{-1}$ to $-44.5\,kJ\,mol^{-1}$ (*cf.* Table 2.2).

This phenomenon of increased stability of the adsorption sites upon loading is not commonly observed in nanoporous materials for rare gas adsorption. It can be attributed to a cooperative effect between the adsorbed molecules, where the interaction between the adsorbed xenon molecules is more favorable than that between the adsorbed krypton molecules. This preference stems from the stabilization due to the interatomic distance within the pores, which closely matches the energy well for favorable Lennard-Jones potential for xenon-xenon interactions, unlike the case for krypton-krypton interactions (Figure 2.27, where the distance exceeds 4.2 Å).

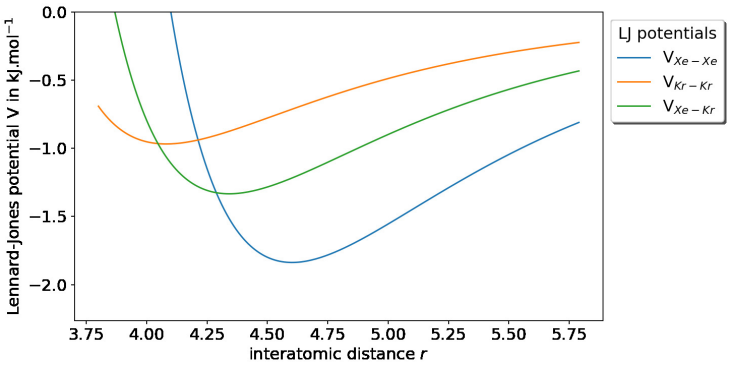


Figure 2.27: The LJ potentials for xenon and krypton interactions. The xenon-xenon interaction is more stabilizing than the krypton-krypton interaction for interatomic distance higher than 4.2 Å.

In the case of KAXQIL, the material features one-dimensional tubes as channels (Figure 2.28), with a distance between two adsorption sites that is approximately the unit cell parameter along the tube direction (5.6 Å). The selectivity of this material increases with pore filling, primarily driven by enthalpic considerations, which can be explained through a relatively straightforward rationale. By estimating the Lennard-Jones potentials U^{LJ} for all species at a distance of 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 simplified model where all adsorbed molecules are assumed to be 5.6 Å apart, the cooperative effect between two xenon molecules is more significant, which accounts for the increased selectivity at high uptake. Further analysis of the adsorption enthalpies of xenon and krypton (*cf.* Table 2.2) reveals that both values increase. This can be attributed to the guest molecules deviating from the "ideal" adsorption sites, resulting in guest-guest interactions that do not fully compensate. Consequently, the selectivity change observed in this material is a consequence of the rearrangement of adsorbate positions within the nanopores, driven by guest-guest interactions.

To further validate the role of the guest–guest interactions, another material with one-dimensional tubelike channels is considered: JUFBIX, a cobalt(II) coordination polymer based on carboxylic acid linkers (Figure 2.29).**JUFBIX** The periodicity along the tube direction is significantly larger at 7.2 Å. The pair interaction energies corresponding to the LJ potentials at this distance are determined as $U^{LJ}_{Xe\text{-}Xe} = -0.24\,kJ\,mol^{-1}$, $U^{LJ}_{Kr\text{-}Kr} = -0.06\,kJ\,mol^{-1}$ and $U^{LJ}_{Xe\text{-}Kr} = -0.13\,kJ\,mol^{-1}$.

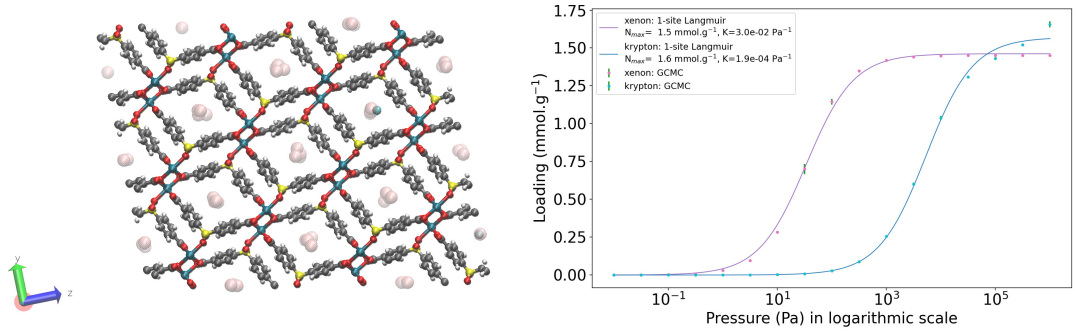


Figure 2.28: KAXQIL*: On the left side, an illustration of a clean version (all solvent removed) of the calcium coordination framework [Ca(SDB)]·H_2O, where SDB = 4,4′-sulfonyldibenzoate loaded with xenon and krypton obtained by GCMC calculations. Color code: Ca in dark cyan, C in gray, O in red, H in 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 2.33) for both xenon and krypton at* 298 K *is represented on the right side.*

Upon analyzing the adsorption enthalpies (Table 2.1), it is observed that these values are too small to affect the positioning of the adsorbed molecules. When the loading is high, resulting in a significant distance between the adsorbed molecules, each adsorption site becomes independent of others. Consequently, the ambient-pressure selectivity s_1 remains equal to the low-pressure selectivity s_0 since the guest–guest interactions become negligible. This finding substantiates the critical role played by cooperative effects among guest molecules when considering a saturated material.

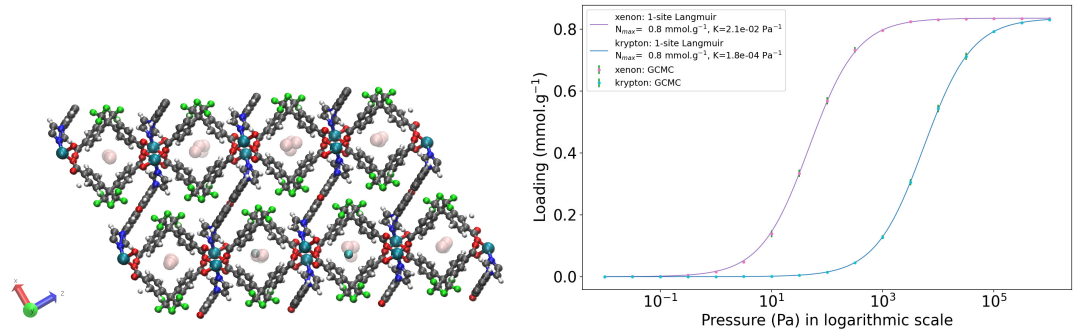


Figure 2.29: JUFBIX*: Representation of a clean version (all solvent removed) of the cobalt(II) coordination framework [Co_2(L)(ppda)$_2$]$_2$·H_2O, where the ligand L is 2,8-di(1H-imidazol-1-yl)dibenzofuran and the carboxylic acid ligand H_2ppda is 4,4′-(perfluoropropane-2,2-diyl)dibenzoic acid loaded 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 2.33) for both xenon and krypton at* 298 K *is represented on the right side.*

GOMREG and JAVTAC are two frameworks categorized as materials with a moderate decrease in selectivity from low to ambient pressure. In the case of GOMREG, the channels consist of one-dimensional tubes that are larger compared to KAXQIL or JUFBIX (Figure 2.30 and Table 2.1). The adsorption sites alternate from left to right inside the channels, resulting in an organized "zigzag" pattern of adsorbed molecules. Analyzing the adsorption enthalpies, it is observed that both xenon and krypton exhibit lower enthalpies by a similar margin, indicating an equivalent

stabilization for both atoms. Consequently, the enthalpic contribution to the selectivity change is close to 1. Due to its smaller size and weaker interaction with the adsorption site, Krypton has more available space within the pore structure. This leads to an entropic advantage for Kr, as reflected by the entropic contribution k_S of 0.64 in Table 2.1. These findings suggest that while enthalpic considerations primarily account for the observed changes at a statistical level, as discussed in previous sections, entropic considerations can play a significant role in pressure-dependent selectivity for specific cases.

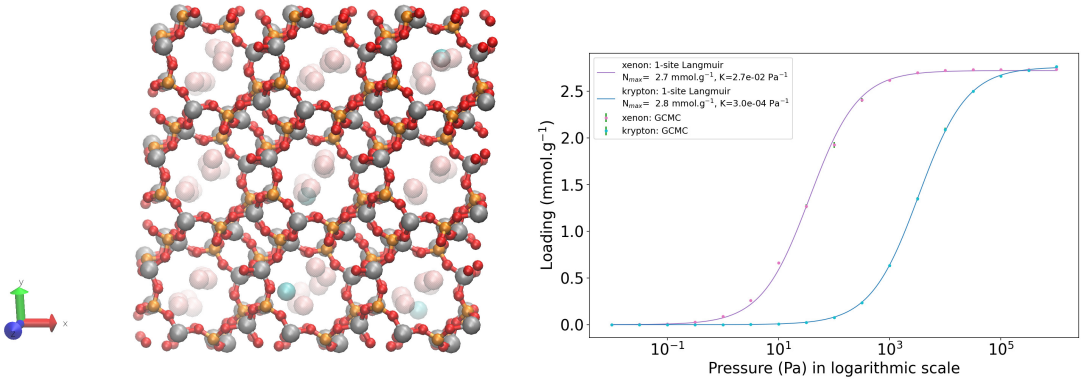


Figure 2.30: GOMREG*: Representation of a clean version (all solvent removed) of this aluminophosphate $AlPO_4$-n that has a zeotype LAU topology with one-dimensional 10-ring channels loaded with xenon and krypton obtained by GCMC calculations. Color code: Al in silver, P in orange, O in red ; Xe in transparent pink and Kr in cyan for the adsorbates. The mono-component isotherms fitted with a 1-site Langmuir model (equation 2.33) for both xenon and krypton at* 298 K *is represented on the right side.*

The remaining materials discussed in this third category exhibit a significant decrease in selectivity from low to ambient pressure. To investigate the factors contributing to this decrease, several phenomena that are relevant for screening studies have been examined, as they can impose limitations on the working performance of materials that initially appear to be "top performer" based on zero-pressure screening.

For example, GOMRAC has a similar structure compared to GOMREG (Figure 2.31), with the distinction of having smaller pores and channels are smaller (see the values of the D_i, and the D_f, in Table 2.1). Consequently, the distances between adsorbed molecules— in their ideal sites — are smaller. At such close distances, it is reasonable to assume that the interactions between adsorbates favor krypton over xenon molecules in GOMRAC (see LJ potentials at distance lower than 4.2 Å in Figure 2.27). This enhanced stabilization of krypton relative to xenon results in an enthalpic contribution k_H of 0.58. Moreover, this finding is consistent with the equivalent guest–guest interactions observed in GOMREG, as discussed earlier. It explains why the difference in adsorption enthalpies becomes smaller for GOMRAC, while it remains unchanged for GOMREG (between low and ambient pressure). This further validates the critical role of interactions between adsorbed molecules and their dependence on guest-guest distances, particularly under high loading conditions.

In the case of MISQIQ, the pure-component Xe isotherm depicted in Figure 2.32 oes not conform to a single-site Langmuir isotherm, but rather aligns well with a two-site Langmuir model (Figure 2.32). Upon visual examination of the adsorbed density at various loadings, it becomes

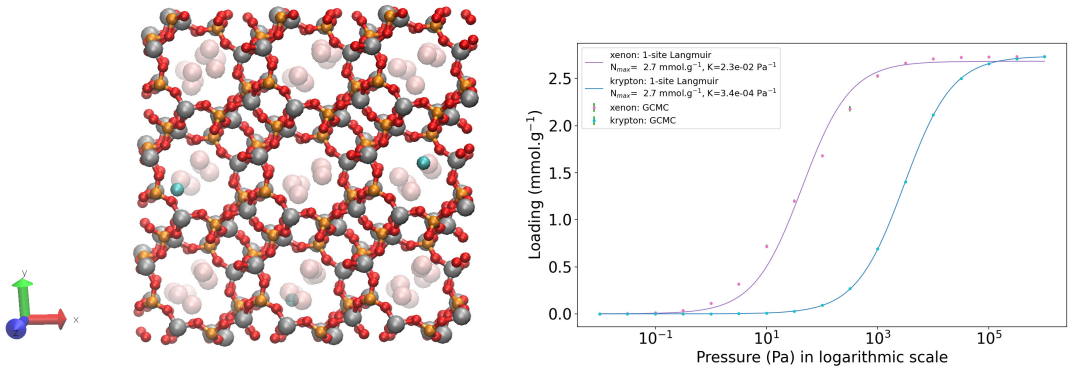


Figure 2.31: GOMRAC*: Representation of a clean version (all solvent removed) of this aluminophosphate $AlPO_4$-n that has a zeotype LAU topology with one-dimensional 10-ring channels loaded with xenon and krypton obtained by GCMC calculations. Color code: Al in silver, P in orange, O in red; Xe in transparent pink and Kr in cyan for the adsorbates. The mono-component isotherms fitted with a 1-site Langmuir model (equation 2.33) 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*.*

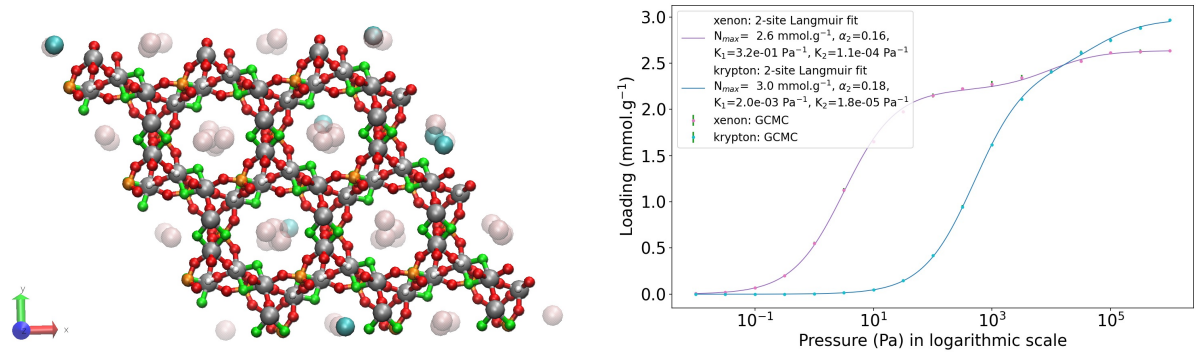


Figure 2.32: Representation of a chiral open-framework fluoroaluminophosphate $[C_4N_3H_{16}]\cdot[Al_6P_3O_{12}F_6(OH)_6]$ denoted AlPO-JU89 (referenced MISQIQ *in the Cambridge structural database), which has been loaded with xenon and krypton in a GCMC simulation, on the left side.*[MISQIQ] *Color code: Al in silver, P in orange, O in red, H in white and F in green for the framework; Xe in transparent pink and Kr in cyan for the adsorbates. The pure-component isotherms fitted with a 2-site Langmuir model (equation 2.34) for both xenon and krypton at* 298 K *on the right side.*

evident that the second step in the isotherm (representing about 20% of the uptake at full loading) corresponds to a reorganization of the adsorbate molecules accompanied by a contraction of interatomic distances. It is important to note that this reorganization does not involve the occupation of a distinct and separate adsorption site at high loading. In this case, the change in selectivity can be attributed to the potential for adsorbate reorganization within the nanopores of the material. This reorganization, which can be detected through the xenon isotherm alone, plays a significant role in determining the material's selectivity at ambient pressure. The repacking of the adsorbed phase during this reorganization process is associated with a strong entropic effect and also influences the enthalpic contribution to selectivity.

The materials BAEDTA01, VIWMOF, LUDLAZ, WOJJOV, and VAPBIZ fall into the category of having more than one available adsorption site, resulting in a significant drop in selectivity from low

to ambient pressure. 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 preferential filling of the most selective sites (i.e., the most favorable for Xe) occurs at low loading, while the less selective sites are populated as the pressure increases. Consequently, a net decrease in selectivity at ambient pressure is observed for these materials. The existence of different types of adsorption sites and their impact on Xe/Kr selectivity (at non-zero pressure) suggests the inclusion of this factor in the screening of pure-component isotherms without the need for explicit multi-component GCMC simulations.

2.4 Towards 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 establishing structure/property relationships, determining theoretical performance limits, and identifying top-performing materials, for both existing experimental structures and novel hypothetical structures yet to be synthesized. To provide a better understanding of the thermodynamics underlying Xe/Kr separation and the microscopic origins of selectivity at low and ambient pressure, a high-throughput screening of Xe, Kr was conducted as well as Xe/Kr mixtures in 12 020 experimental open-framework materials. In addition to structural descriptors such as pore sizes, volume, and surface area, thermodynamic quantities were considered to gain insights into the key factors yielding a high selectivity.

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, it can be concluded that a direct screening of Kr or xenon adsorption free energy may not be essential for a coarse-grained evaluation of the selectivity of nanoporous frameworks. This finding could facilitate the development of more efficient screening methodologies. For instance, a multistage approach could be employed, starting with a preliminary selection on Henry's constant, which is computationally inexpensive. Subsequently, more computationally intensive grand canonical Monte Carlo (GCMC) simulations can be performed 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. It is possible 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 the context of xenon-krypton separation using nanoporous materials, pressure swing adsorption (PSA) processes have been widely used, making pressure a crucial thermodynamic variable in the separation cycle. This study has focused on the selectivity difference between a system under very low pressure (at the zero loading limit, which is calculated at relatively low computational cost) and ambient pressure (closer to working conditions but requiring higher simulation cost). The results demonstrated that selectivity can be highly dependent on pressure, with certain materials maintaining high selectivity at both low and ambient pressures, while

others experience a significant drop in selectivity. It was found that high ambient-pressure selectivity requires high low-pressure selectivity, but the reverse is not necessarily true.

By using a thermodynamic approach to describe the separation selectivity, the differences in selectivity were elucidated between different pressures (and therefore different loading regimes of the frameworks), primarily attributed to the variations in adsorption enthalpies for Xe and Kr.By delving into specific examples, the microscopic origins of these selectivity changes were uncovered and linked to the relative contributions of host–guest and guest–guest interactions. The population of different adsorption sites or repacking of the adsorbed phase at higher loadings can lead to significant alterations in overall selectivity. The underlying mechanisms of 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, this proposed classification of the interactions at play can guide the future design of more efficient high-throughput screening procedures.

For instance, the essentially enthalpic nature of the xenon/krypton separation process underscores the importance of developing more efficient methods for sampling interaction energies and utilizing them as cost-effective descriptors for analyzing an increasing number of structures. In the subsequent chapter, different approaches for evaluating adsorption enthalpy will be explored, considering the computation time required and the accuracy of each method. Additionally, the influence of partial pressure, manifested through changes in composition or pressure, raises questions about the potential use of infinite dilution thermodynamic quantities to predict selectivity at any pressure (GCMC). Numerous studies have focused on predicting GCMC simulations..**Simon_2015**, **Shi_2023**, **Kang_2023**, **Li_2023** The thermodynamics-based approach combined with characterizing pore diversity holds the potential to yield improved results in predicting GCMC values of selectivity, contributing to a more comprehensive understanding of adsorption processes in nanoporous materials.

3

ADSORPTION ENERGIES SAMPLING

This chapter will introduce three distinct energy sampling techniques that can be used to determine an adsorption enthalpy, and, in some cases, deduce Henry constant values.

3.1 Voronoi sampling

The first technique to be discussed has been previously studied for calculating geometric descriptors and, more recently, for deriving energy-based descriptors.**Simon_2015** This sampling method is relatively biased as it relies on a sparse sampling approach based on Voronoi decomposition, which is limited in terms of incorporating prior physical knowledge. A more detailed explanation of this method will be provided in the subsequent discussion. A detailed explanation of this method will follow.

3.1.1 Theoretical considerations

In mathematics, a tessellation of a given space refers to the partitioning of this space into non-overlapping subspaces. In the Voronoi tessellation, named after Georgy Feodosevich Voronoy, a set of points (seeds) corresponds to a tessellation of regions (Voronoi cells). These cells are designed such that each seed possesses a cell wherein all points are closer to that seed than any other seeds.[Rycroft_2009] When applied to materials science, the Voronoi cells attributed to each atom of the framework can be leveraged to determine key geometric 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, 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 each $k \in \{1,\dots,n\}$, a subspace S_k (also called Voronoi cell) can be defined around the atom k, encompassing all points $\boldsymbol{x}$ within this subspace that are 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 termed the Voronoi partition of the space B. The edges and vertices of these polyhedra offer valuable information regarding the void space between adjacent the Voronoi cells associated with them. By leveraging this information, it becomes feasible to determine the accessible and inaccessible points within 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 point closest to the atomic positions $\boldsymbol{x}_{i_1},\dots,\boldsymbol{x}_{i_p}$ — this can be easily demonstrated by combining different conditions outlined in the equation 3.1. The same assessment can be performed for any point located on an edge adjacent to certain subspaces, as it will be closer to the atoms associated with these subspaces than to any other atoms.

This regular Voronoi tessellation is suitable only for separating space among equally sized atoms, as it sets the boundaries at equidistance from all the surrounding atoms, as shown in Figure 3.1. For atoms with unequal sizes, this type of definition may not be desirable, as the boundary could be closer to the surface of an atom compared to another. The initial rationale behind using a Voronoi decomposition is to delimit a region for each atom that is closer to that specific atom than any other. The ambiguity of this definition arises from the definition of "closeness". In this regular Voronoi decomposition, closeness is determined based on the distance between the center of mass of different atoms, which poses a challenge for unequally distributed radii.

Unequal radii

To address this limitation, an alternative approach called the Apollonian Voronoi diagram can be implemented to model the atomic radii. The definition of the Voronoi decomposition, as previously discussed, is limited to equal-sized atoms, as the closest region to an atom is also the closest to its center of mass. This limitation does not apply to the complex atomic structures found in nanoporous frameworks. To overcome this, the Apollonian Voronoi diagram[voronoi_apollonian]

Figure 3.1: Bidimensional illustrations of a Voronoi decomposition using three types of algorithm: (i) for equally sized circles using the equation 3.1 (www.shadertoy.com/view/MslGD8) (ii) for unequally sized circles using the Apollonian Voronoi decomposition condition 3.2 (www.shadertoy.com/view/4sd3D7) and (iii) another algorithm for unequally sized circles using the radical Voronoi condition 3.3 (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.

can be utilized to model the atomic radii $r_1, \ldots, r_n$ of points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ within the same box B. For every $k \in \{1, \ldots, n\}$, the new subspaces A_k are defined as follows:

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

This new definition of the Voronoi diagram takes into account the intuitive property of closeness to the atom's surface rather than its center of mass. This adjustment allows for an unequal distribution of atomic radii, as the diagram now depends on these radii. However, as illustrated in Figure 3.1, the initial implementation presents a convenient definition at the cost of curved edges, which introduces computational challenges.

To overcome these challenges and enhance computational efficiency, a less intuitive but more commonly used implementation known as the radical Voronoi tessellation, power diagram or Laguerre-Voronoi diagram[aurenhammer_1987] is preferred. As depicted in Figure 3.1, this method yields subspaces that are convex polygons with straight edges. Although the condition defining these subspaces is less intuitive, it avoids reliance on a simple definition. The subspaces V_k are now defined based on the following condition:

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

In addition to the polyhedral form of the Voronoi cells, this new implementation presents interesting properties for porosity calculations in frameworks composed of unequal spheres, such as MOFs or zeolites.[voronoi_radical] First, the boundary between two overlapping spheres corresponds simply to the intersection between the spheres themselves. Secondly, the boundary between non-overlapping spheres always lies within the void space separating them. This assertion can easily be demonstrated by considering a point $\boldsymbol{x}$ in V_k and $\|\boldsymbol{x} - \boldsymbol{x}_k\|_2 \geq r_k$ outside

the sphere, which implies $\forall\ l \neq k,\ \|\boldsymbol{x} - \boldsymbol{x}_k\|_2 \geq r_k$. The point $\boldsymbol{x}$ does not overlap with any other atom and resides within the framework's void space.

When considering a point $\boldsymbol{v}$ on a boundary between p Voronoi cells, denoted as $\left\{\mathrm{V}_{i_1}, \ldots, \mathrm{V}_{i_p}\right\}$, this point satisfies the conditions $\|\boldsymbol{x} - \boldsymbol{x}_{i_1}\|_2^2 - r_{i_1}^2 = \cdots = \|\boldsymbol{x} - \boldsymbol{x}_{i_p}\|_2^2 - r_{i_p}^2 = \mathrm{C}$. It is possible to find the minimum distance to the center of mass of neighboring atoms and test possible overlapping. Specifically, in the Zeo++ software,[Zeo++] the Voronoi diagram is characterized by storing the minimum distance to the closest atoms and the corresponding atom indices for every vertices and edges (in cases where edges connect two different periodic images, a periodic displacement vector is also stored). Leveraging this information enables the acceleration of void fraction calculations by bypassing volume calculations in non-adsorbable Voronoi cells. Additionally, it provides a swift approach to determine accessible and non-accessible surface areas and volumes.[Zeo++] It is important to note that when employing a probe with a radius r_{probe}, the sphere radii considered are adjusted accordingly as $r_k = r_{\text{atom}} + r_{\text{probe}}$.

3.1.2 Implementation in a screening

The Voronoi decomposition of geometric characterization of the pore space in materials has become widespread in computational studies over the past decade.[Willems_2012] Its popularity increased notably after its incorporation into the Zeo++ software package.[Pinheiro2013] Recently, this technique was further extended to implement a novel sampling scheme in a study focused on ML-assisted screening of nanoporous materials for xenon/krypton separation. In their work, 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 to each atom in the structure. These vertices (or *Voronoi nodes*) are the points equidistant to at least four atoms of the structure, and can be associated with adsorption sites due to their closeness to the center of the pores.

The Zeo++ software definition of accessibility was used in a screening process aimed at identifying optimal materials for Xe/Kr separation.[Simon_2015] The interaction energies of xenon were calculated exclusively at the accessible nodes, as illustrated in Figure 3.2. The average of the energies at these accessible Voronoi nodes provided an estimation of the adsorption enthalpy. However, this sampling approach assumes that the nodes are close to the actual and most favorable adsorption sites, which implies that the adsorption sites are located at the center of the pores. This assumption holds true only for structures with pore sizes similar to that of the adsorbate. The newly defined adsorption energy descriptor was identified as one of the most influential descriptors in the ML learning model developed by Simon et al. to predict ambient-pressure selectivity.

In light of the initial Voronoi sampling methodology, it is worth questioning the relevance of directly averaging the interaction energies instead of employing Boltzmann averaging to describe the adsorption enthalpy. To gain a deeper understanding of the strengths and weaknesses of this methodology, different methods for approximating adsorption enthalpies have been compared through the Voronoi sampling approach 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.

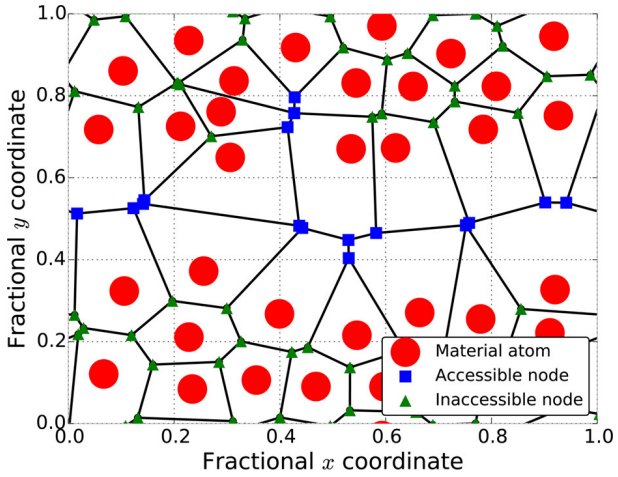


*Figure 3.2: Voronoi network model of void space (2D representation). The unit cell of a toy material is shown. Red circles represent atoms of the material; accessible and inaccessible Voronoi nodes are blue squares and green triangles, respectively. The black lines are the edges in the periodic Voronoi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

3.1.3 Comparative study of the Voronoi sampling

In the previous chapter, the definition of the xenon adsorption enthalpy at infinite dilution (Widom insertion section 2.1.4) and at ambient pressure (GCMC sections 2.1.3 and 2.1.5) was introduced. These methods are widely acknowledged for their accuracy in calculating adsorption enthalpies, which have been established as strongly correlated with the logarithm of selectivity in a previous study on the thermodynamic exploration of xenon/krypton separation using high-throughput screening.

Introduction of the main concepts

The Voronoi energy, as initially conceptualized by Simon et al., is obtained by averaging the xenon interaction energies at the accessible Voronoi nodes. However, for the purpose of comparing with thermodynamic simulations without blocking pockets, the focus is shifted from the accessible Voronoi nodes to the adsorbable Voronoi nodes as they provide a closer approximation to the desired simulation. To simplify the analysis, the set of the adsorbable Voronoi nodes A is defined as the Voronoi nodes with a negative energy value among the ones with a minimum distance to the nearest atom higher than 2 Å. This distance is chosen so that a xenon would be distant from the surface of the neighboring framework atoms (in the Apollonian definition of the Voronoi nodes). Since the xenon diameter is about 4 Å, the set of Voronoi nodes considered are the ones where a xenon particle can be inserted. Additionally, this condition of the distance reduces the computation time required. This average on the adsorbable Voronoi nodes $\mathrm{E}^{Xe}_{\text{voro-A}}$ can be expressed as follows:

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

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 higher than 2 Å. This condition on the distance also reduces the computational time required to find this minimum,

and usually the minimum is always among these adsorbable points. This minimum Voronoi energy $E^{Xe}_{\text{voro-M}}$ can be expressed as follows:

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

Finally, to align with the definition of the heat of adsorption presented in the previous chapter, an energy descriptor can be built using Boltzmann averaging on the same set of nodes V with the same condition on the distance. This condition also reduces the computational cost, while being accurate since the high energy points would be negligible in the averaging. This Boltzmann average of the xenon interaction energies at the Voronoi nodes V is denoted as $E^{Xe}_{\text{voro-B}}$ and can be expressed as follows:

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

It should be noted that the −RT term is necessary to make the expression comparable to the one of adsorption enthalpy (equation 2.22).

Intuitively, since Boltzmann averaging is closer to the definition of the adsorption enthalpy, it would be a more suitable candidate as an energy descriptor, and it can potentially be used to improve the current screening methodology. To test these different methodologies, various energy descriptors will be compared with more accurate evaluations of the adsorption heat.

Low-pressure comparison

The Widom insertion is typically used to calculate the infinite dilution adsorption properties, such as adsorption enthalpy, Henry constant and selectivity. The evaluation of xenon interaction energies at different Voronoi nodes corresponds to a low-pressure averaging and is comparable to the Widom insertion method. However, it is biased by the inhomogeneous sampling of the space, which can account for some of the observed discrepancies.

It is important to note that in this chapter, the standard pore size definition commonly used in the literature, based on atom radii provided by the Cambridge Crystallographic Data Centre (CCDC), will be predominantly used. This pore size will only serve a labeling purpose aid in classifying structures based on their relative size. It predominantly plays a qualitative role, which justifies the omission of a more precise definition based on the forcefield, as employed in the previous chapter.

As illustrated in Figure 3.3, the average of energies (left panel) exhibits suboptimal performance and demonstrates weaker correlation with the adsorption enthalpy compared to the minimum interaction energy (center panel) or the Boltzmann average of interaction energies (right panel). This discrepancy occurs because, in a normal average, high-energy values carry a disproportionately higher weight than in a Boltzmann average, resulting in the average being more significant than expected. The Voronoi average descriptor $E^{Xe}_{\text{voro-A}}$ consistently yields higher values than the infinite dilution adsorption enthalpy $\Delta_{\text{ads}}H^{Xe}_0$.

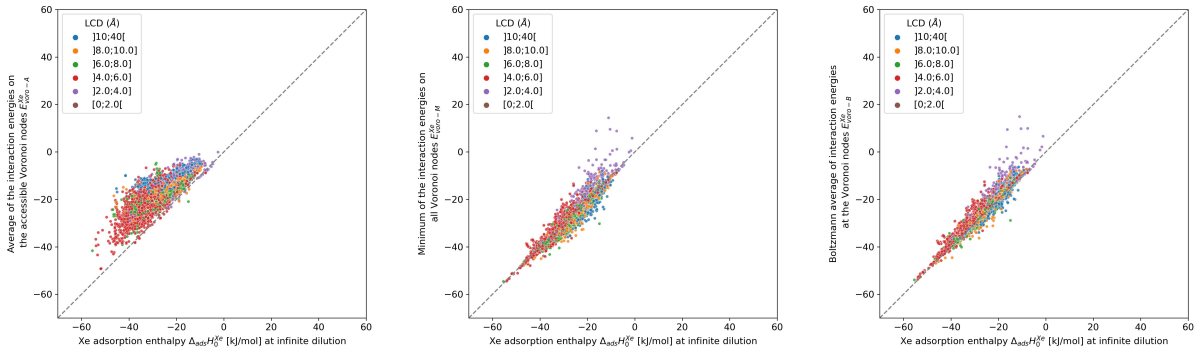


Figure 3.3: Scatterplots of the energy descriptors $E^{Xe}_{voro\text{-}A}$, $E^{Xe}_{voro\text{-}M}$ *and* $E^{Xe}_{voro\text{-}B}$ *calculated by a Voronoi sampling compared to the enthalpies calculated by a 100k-step Widom insertion simulation of xenon in structures of CoRE MOF 2019. The points are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

The Pearson correlation coefficients corroborate the initial observation made in this thesis. The correlation coefficient between $E^{Xe}_{voro\text{-}A}$ and $\Delta_{ads}H^{Xe}_0$ is 0.81, whereas for the minimum $E^{Xe}_{voro\text{-}M}$ and for the Boltzmann average $E^{Xe}_{voro\text{-}B}$, it is respectively equal to 0.95 and 0.97. Based on these coefficients, it is evident that the Boltzmann average is more suitable to evaluate the relevance of a Voronoi energy sampling. As shown in the previous chapter, selectivity is correlated with the difference of adsorption enthalpies between xenon and krypton. Improving the description of enthalpy is a first key step towards a better description of selectivity. As the previous analysis only focused on selectivity values at low pressure, it is essential to explore the behavior of selectivity at higher pressures.

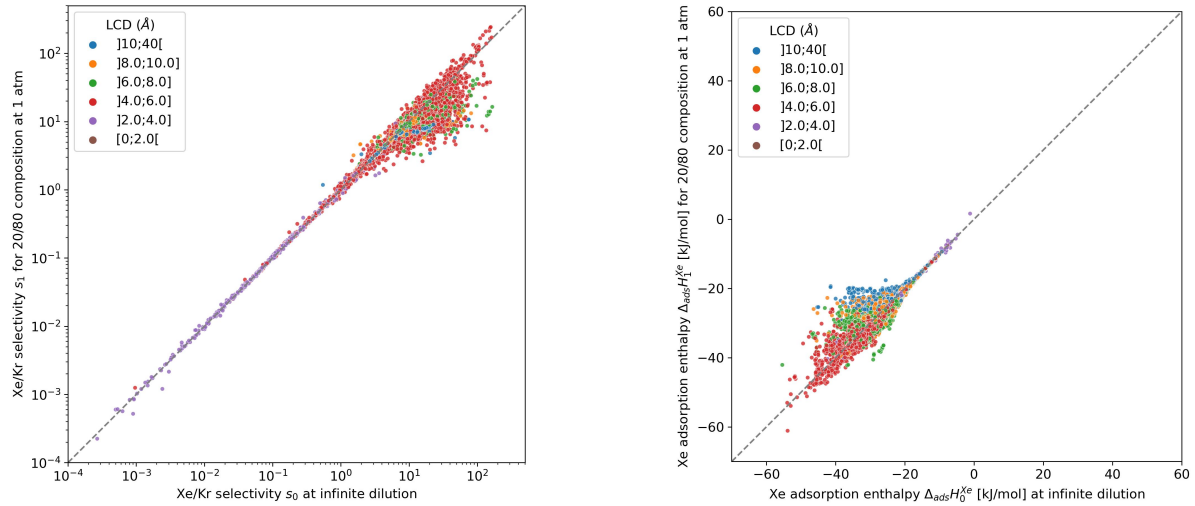


Figure 3.4: Comparison of the ambient-pressure and low-pressure case through two thermodynamic quantities: the Xe/Kr selectivity (left) and the xenon adsorption enthalpy (right).

Figure 3.4 illustrates that the selectivity drop between the low-pressure and ambient-pressure cases has an impact on the enthalpy values of xenon. The xenon affinity decreases as the pressure increases. While the study conducted by Simon et al. primarily focused on predicting

ambient-pressure selectivity, it is worth investigating whether the energy descriptor they developed can also describe the adsorption enthalpy at high pressure.

Ambient-pressure xenon/krypton separation

Upon observing the Figure 3.5, it is not clear which descriptor best performs in describing the enthalpy at ambient pressure. The scatterplots indicate similarly modest correlations for all descriptors, suggesting that the use a regular average may suffice instead of a Boltzmann average. The correlation coefficient for the average $E^{Xe}_{\text{voro-A}}$ is now 0.86, which is equivalent to both the minimum descriptor $E^{Xe}_{\text{voro-M}}$ and slightly lower than the 0.87 for the Boltzmann average $E^{Xe}_{\text{voro-B}}$.

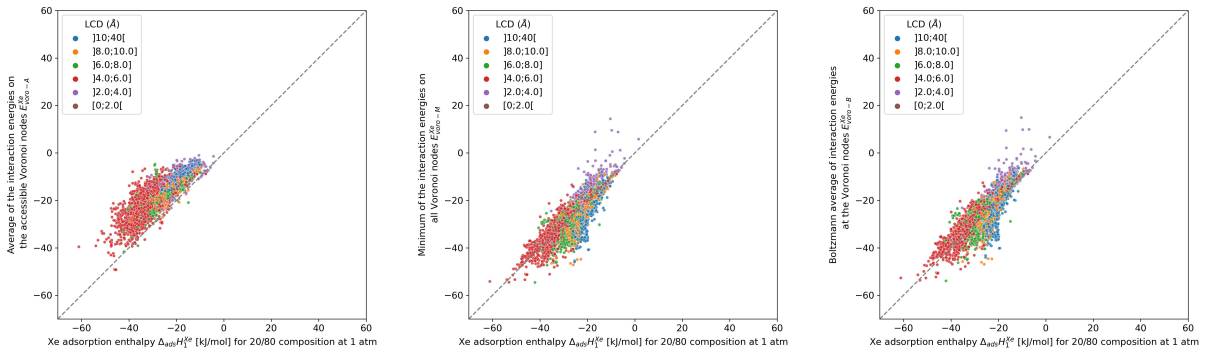


Figure 3.5: Scatterplots of the energy descriptors $E^{Xe}_{voro\text{-}A}$, $E^{Xe}_{voro\text{-}M}$ and $E^{Xe}_{voro\text{-}B}$ calculated by a Voronoi sampling compared to the enthalpies calculated by a 100k-step GCMC simulation of xenon in structures of CoRE MOF 2019. The points are labeled according 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, resulting in a diminished correlation between the Boltzmann average and the minimum of the interaction energies calculated at the Voronoi nodes. The regular averaging approach tends to overestimate the energy values, bringing them closer to the values observed at higher pressures. To address this issue, an alternative averaging method that assigns greater weights to the higher energy values has been developed. This new approach resembles a Boltzmann average with a higher temperature value. The next chapter will focus on testing and evaluating the performance of this alternative method. The overestimation of energy values in the averaging process allows these values to align more closely with the ones observed at higher pressures. Drawing inspiration from this idea, an alternative averaging method that assigns greater weights on the higher values has been developed in this thesis. This modified Boltzmann average with increased weights for higher energy values will also be tested in the next chapter.

3.1.4 Performance of a Voronoi energy sampling

This section will focus on some performance metrics associated with the Boltzmann average at the Voronoi nodes and comparison with the reference sampling method, the Widom insertion with 100,000 cycles. The right plot of the Figure 3.3 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 found to be strong for only a limited number of structures with enthalpy around $-50\,\text{kJ}\,\text{mol}^{-1}$. For structures with higher enthalpy, the correlation diminishes,

particularly for structures with small-pore sizes. For the points in purple in Figure 3.3, the largest cavity diameter is lower than the kinetic diameter of a xenon, and the Voronoi node sampling is clearly deemed insufficient. In addition, the loss of accuracy observed for other points (larger pores) can be explained by the fact that the pores are slightly larger and the center of the pore is no longer an accurate approximation of the adsorption site position, as the adsorption sites are closer to the pore surface than the center of the pore. Consequently, these findings have motivated the proposal of a new sampling scheme based on the molecular surface of the pore space, which will be elaborated in subsequent sections.

Evaluating the performance metrics, the root mean squared error (RMSE) and the mean absolute error (MAE) for Voronoi sampling are determined to be 6.78 kJ mol^{-1} and 2.01 kJ mol^{-1} respectively, when considering all structures in the set. These values appear to be too high to be used for screening purposes. However, non-porous materials would be screened out *a priori* in any high-throughput workflow due to their lack of relevance. Therefore, the focus can be placed on structures with cavities larger than 3.7 Å (slightly lower than 3.96 Å Xe kinetic diameter). By restricting the analysis to such structures, the RMSE and MAE decrease to 2.11 kJ mol^{-1} and 1.55 kJ mol^{-1} respectively. These values can be considered acceptable for a rapid estimation of the guest–host affinity, although they are not suitable for accurate adsorption enthalpy calculations.

The low computational cost of the method further supports its feasibility. The Voronoi tessellation performed by the Zeo++ software is extremely fast, generating the positions of the Voronoi nodes in approximately 0.28 s (on average across all the structures of the CoRE MOF 2019 database), using a typical workstation (a single Intel Xeon Platinum 8168 core at 2.7 GHz). In comparison, a simple Python prototype code for energy calculation required around 27 s per structure, whereas an optimized C++ implementation benchmarked in this thesis achieved Voronoi sampling in approximately 0.4 s. Consequently, the Voronoi sampling method requires only a few hundred milliseconds per structure, whereas a Widom insertion method necessitates approximately hundreds of seconds per structure. Thus, Voronoi sampling exhibits computational efficiency that is 2 to 3 orders of magnitude faster than that of a full sampling of the pore space.

This preliminary study has identified a fast method for adsorption enthalpy calculation that can be widely used in screening procedures. However, its accuracy for quantitative prediction is limited — this sampling technique assumes that the nodes are close to the real, most favorable adsorption sites, which may not always hold true. Specifically, the assumption that the adsorption sites must be located at the center of the pores is only valid for structures with pore sizes close to the size of the adsorbate. This observation raises key questions regarding the importance of selecting appropriate sampling points within the pore space of materials. Consequently, an intermediate technique was developed and optimized to address these limitations and provide a sampling approach that is both fast and accurate for predicting adsorption enthalpy. This new technique focuses the sampling on the surface of the material, aiming to compensate for the primary flaws encountered in the Voronoi sampling approach.

3.2 Rapid Adsorption Enthalpy Surface Sampling (RAESS)

In this section, the development of a new surface sampling algorithm will be described, aiming to achieve higher accuracy than Voronoi sampling and greater efficiency than Widom insertion. My initial idea is based on a series of theoretical considerations: (i) strong adsorption sites are located near the surface of the material; (ii) by changing the problem from 3D to 2D sampling, the complexity can be reduced; and (iii) the algorithm can scale with the number of unique atoms in the structure (rather than the size of the unit cell), which is efficient as many porous frameworks exhibit high symmetry. The first physical intuition ensures that the proposed method will yield more accurate results compared to Voronoi sampling, while the latter two considerations suggest that a well-optimized code can be much faster than a standard sampling (e.g., Widom insertion). To validate these hypotheses, an analysis of both the accuracy and speed of the new algorithm will be conducted and compared against existing methods. It is worth noting that the study described in this section has already been published in the Chemical Science journal [**Ren_2022**].[1]

3.2.1 Initial implementation

The initial implementation of the surface sampling algorithm and its principles are presented in this section. Although this initial implementation is relatively basic, it already demonstrates good performance compared to other methods. In subsequent sections, further refinements will be made through the introduction two additional features that will improve both the accuracy and speed of the algorithm.

This initial implementation accelerates the calculation of adsorption enthalpy in nanoporous materials by sampling interaction energies exclusively near the surface. Figure 3.6 provides an illustration of this approach. To achieve this, a loop is performed over all unique atoms (as defined by crystalline symmetry). For each atom, a sphere is sampled around its position using a uniform distribution, and the number of these sampled points can be adjusted. The default radius for the sampling spheres is set to the distance $r_{\text{min}} = 2^{1/6}\sigma_{ij}$, which corresponds to the minimum of the LJ potential between atoms of type i (belonging to the framework) and j (the guest). This choice represents the strongest possible pair interaction (although the neighboring atoms will also have a second order influence). After calculating the interaction energy $\mathcal{E}_{\text{i}}$ at each sampled point, a Boltzmann average of these energies is obtained. This average corresponds to a biased adsorption enthalpy, as described by the equation 3.7.

$$\Delta \text{H}_{\text{ads}} = \frac{\sum_i \mathcal{E}_{\text{i}} e^{-\frac{\mathcal{E}_{\text{i}}}{\text{RT}}}}{\sum_i e^{-\frac{\mathcal{E}_{\text{i}}}{\text{RT}}}} - \text{RT} \tag{3.7}$$

To validate the accuracy of the approximation made using this sampling, the initial RAESS algorithm was applied with 300,000 sampling points per unique atom. The results, as illustrated by Figure 3.12, demonstrate a good numerical agreement with the reference calculations, the RMSE and MAE being found to be only around 0.90 kJ mol^{-1} and 0.66 kJ mol^{-1} respectively,

[1] The data associated with this work can be found at https://github.com/fxcoudert/citable-data/tree/master/154-Ren_ChemSci_2023

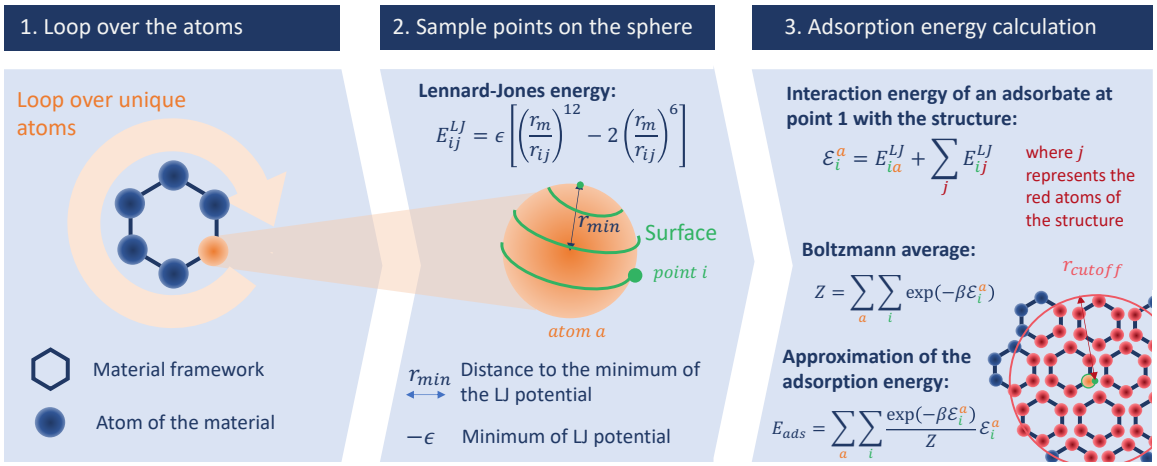


Figure 3.6: Schematic description of the surface sampling based on the three main steps of the algorithm: the loop over the unique atoms, the spiral sampling around each atom, and the energy averaging. The adsorbate is represented by the point i and is moved across all the points around the unique atoms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

considering all the structures from the database. Moreover, no noticeable difference in RMSE was observed when considering the structures with a pore size above 3.7 Å (as determined by the LCD_{CCDC}). Unlike Voronoi sampling, this method provides consistent accuracy across all the structures of the database, with a lower error. The validation of an RMSE value below 1 kJ mol^{-1} supports the intuition that this new sampling technique achieves a balance between the accuracy and efficiency of the previously introduced methods (Voronoi and Widom).

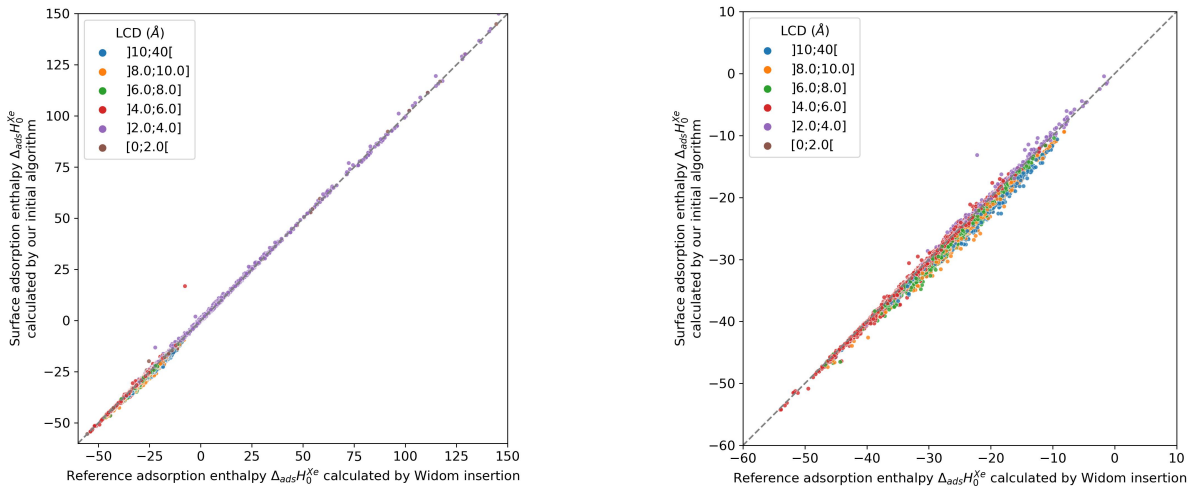


Figure 3.7: Scatterplots of the xenon surface adsorption enthalpy calculated by an initial implementation of the RAESS algorithm as a function of the xenon adsorption enthalpy calculated by a 100k-step Widom insertion simulation using two value windows. The second plot zooms on the negative values corresponding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

After proving its good accuracy, the computation time required for the method was analyzed as a function of the sampling size. Figure 3.8 illustrates that the method reaches an RMSE below 1.0 kJ mol^{-1} promptly, with an average CPU time of 1.2 s, corresponding to 2,000 sampling

points per atom. This is significantly shorter than the 150 s required for a Widom insertion to approach its plateau value, with an RMSE of 0.10 kJ mol^{-1} and 12,000 cycles. It is important to note that the comparison is done with Widom insertion with 100,000 cycles, which explains the convergence of the error of towards a quasi zero value for this method. Additionally, the Widom insertion takes approximately 14 s to achieve a similar RMSE of 1.0 kJ mol^{-1}, which is still slower than the surface sampling. Therefore, this initial implementation of surface sampling exhibits faster computation time than a standard Widom insertion, while maintaining good accuracy.

The observed convergence speed and limit values of the error can be explained by the nature of each sampling method. In surface sampling, the sampled points are biased towards the most attractive adsorption points for xenon, leading to a rapid convergence since the most influential terms of the Boltzmann average are quickly gathered. On the other hand, in a Widom insertion, every point in space has an equal chance of being sampled, which closely aligns with the definition of enthalpy but requires much more time to randomly sample highly attractive adsorption sites. However, due to its biased nature, surface sampling is inherently less accurate, as not all points are considered equally, potentially missing the most optimal adsorption site in some cases, especially if it is located further from the sampled surface.

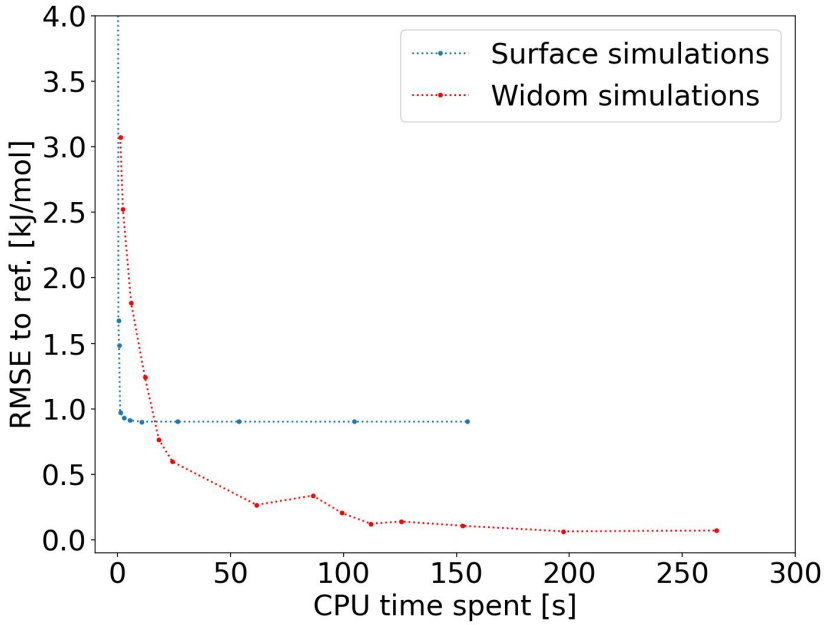


Figure 3.8: Convergence plot of the RMSE on the adsorption enthalpy for the RAESS algorithm (blue) compared to a 100k-step Widom insertion simulation (red) for xenon adsorption in all structures of the CoRE MOF 2019 database.

However, this initial implementation of the method is slower than a Voronoi sampling, which only requires an average of around 1,600 sampled points, as opposed to approximately 13,000 sampled points on average for this implementation — the number of sampled points in RAESS is calculated by multiplying the 2,000 points per atom sphere surface by the average number of symmetrically unique atoms. The sampling process would take approximately 0.15 s, while the generation of Voronoi nodes would take about 0.28 s, resulting in the surface sampling algorithm(1.2 s) being 2 to 3 times slower (both methods implemented in an identical compiled

language, in this case C++). To improve both the accuracy and performance, further adjustments were made to the surface sampling method. The size of the sampling sphere was adjusted, and a fast rejection criterion was implemented. The rejection of high-energy points with little contribution to the final enthalpy value helps reduce simulation time, while the size of the sampling sphere can improve accuracy. As the initially chosen sphere size considered only the interaction with the closest atom, the size was set at the minimum of the Lennard-Jones potential. However, taking into account the interaction with neighboring atoms can further stabilize the adsorbate, and sampling beyond this minimum can potentially increase the accuracy of the 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 most favorable adsorption site corresponds to the minimum of the Lennard-Jones potential. This assumption holds true when the closest atom contributes significantly to the overall interaction. However, in real frameworks, other neighboring atoms also contribute to the host/guest interaction, and in most materials, the adsorption sites are found to be often located farther apart than the LJ potential minimum to maximize the contribution of all atoms — the dissymmetry of the interaction potential well further supports this observation. To explore the possibility of incorporating this insight into the RAESS algorithm, a parameter λ was introduced, and the sampling sphere radius was defined by $R_\lambda = \lambda\sigma$, where σ represents the distance at which the LJ potential is zero. If $\lambda = 2^{1/6}$, the algorithm reverts to the initial definition of the sampling sphere, where the adsorbent is situated at the minimum of the LJ potential for the atom. For $\lambda = 1$, the sampling sphere is centered at the zero of the LJ potential. By varying this parameter, this intuition regarding the optimal positioning of the sampling sphere can be examined and validated.

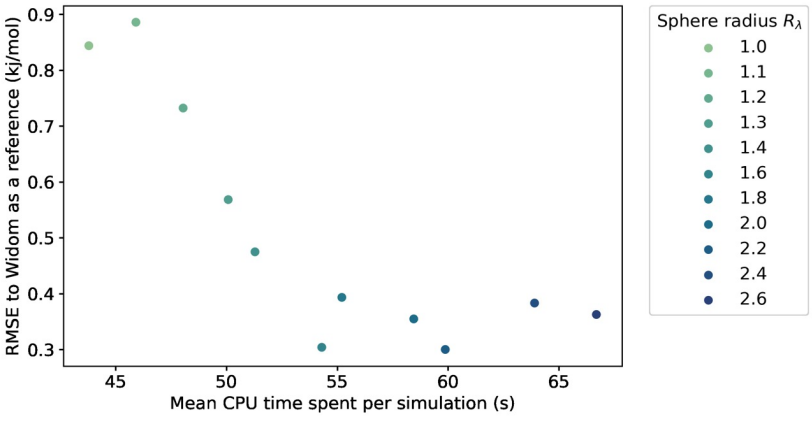


Figure 3.9: Influence of the sampling sphere radius R_λ on the average CPU time required for a simulation of 100k sampling points and the RMSE, compared to the reference adsorption enthalpy. The averaging is done only on the structures with the largest cavity diameter (LCD_{CCDC}) higher than 3.7 Å.

As no analytical model could determine the optimal value for the sampling sphere, a statistical approach was adopted to study the influence of the λ parameter on both the accuracy and computation time. The results are presented in Figure 3.9. It was observed that the RMSE is

relatively high, around 0.90 kJ mol^{-1}, for radius sphere lower than the r_{min}, and then decreases to reach a plateau around 0.35 kJ mol^{-1}, as the radius increases. This confirms that increasing the sampling sphere radius can enhance the accuracy of the algorithm, and it was found that values of λ higher than 1.6 lead to the accuracy stabilized accuracy. This study also found that increasing the sphere radius negatively impacts computational efficiency, as it involves considering a larger number of neighboring atoms in the energy calculation.

By choosing an optimal sampling sphere, it is possible to reduce the error by more than half while increasing the computation time by approximately 20 percent when comparing λ = 1.6 with λ = 1.1 (close to r_{min}). In most cases, this trade-off is acceptable. However, in scenarios where computation time is crucial, such as rapid screening, the optimal choice might not be to increase the sampling sphere at λ = 1.6 but to choose a lower sampling sphere radius at λ = 1.4 or λ = 1.2, resulting in an RMSE around 0.5 kJ mol^{-1} — still considered quite acceptable. The introduction of the new scale parameter in this section allows users to tailor the algorithm according to their specific purposes, prioritizing either accuracy or computation speed. If the method is applied to a completely different database under different conditions, users can choose a default value that works well, such as (*e.g.* λ = 1.4), or optimize the parameter based on a small diverse sample of the unseen data.

Rejection condition

As demonstrated above, the RAESS algorithm exhibits improved accuracy compared to Voronoi sampling. However, its initial implementation was significantly slower, which could hinder its applicability in high-throughput screening workflows involving large numbers of structures, potentially exceeding one million. To address this computational expense, this thesis implemented a mechanism to reject points with minimal contribution to the final enthalpy i.e., the exclusion of sampling points that yield largely positive interaction energies, as these would have negligible impact when exponentiated in the Boltzmann average calculation.

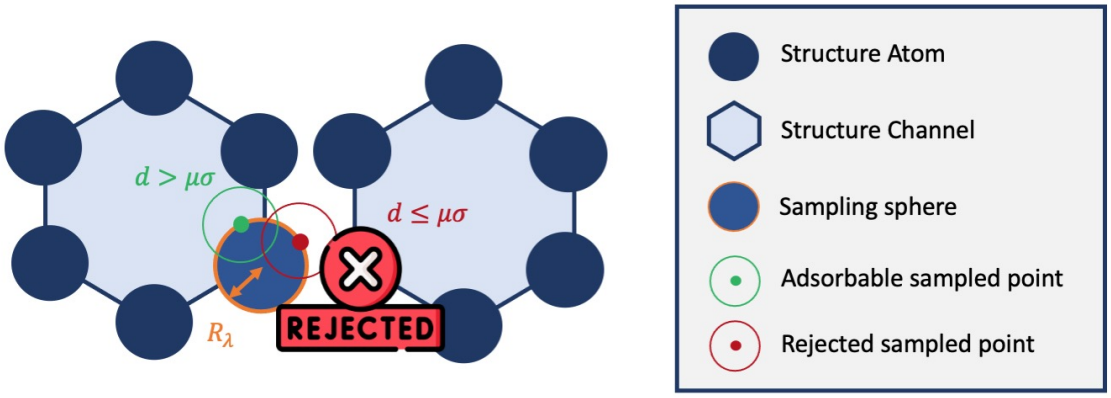


Figure 3.10: Simplified representation of the principle of rejection condition and the concept of sampling sphere inside 2D channels of a nanoporous material.

Inspired by conventional methods for calculating accessible surface, a hard sphere rejection condition based on the distance to neighbors was implemented. 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). The distance threshold is based on the σ_{ij} parameter of the Lennard-Jones potential. To determine the optimal threshold, a factor μ with real values between 0 and 1 was introduced, which modifies the size of the hard sphere rejection condition. If the

guest–host distance is lower than $d_\mu = \mu \times \sigma$, the point is rejected. The absence of a rejection condition occurs when $\mu = 0$, while a value of $\mu = 1$ leads to the rejection of all points with a positive energy interaction with at least one atom of the structure. However, this condition may be overly stringent, resulting in the rejection of points with non-negligible contributions. The rejection condition is schematically illustrated in Figure 3.10.

This rejection condition is expected to speed up the calculation process by avoiding energy computations for rejected sampling points. The energy calculation represents the largest proportion of the CPU time allocated to surface sampling. In the case of the KAXQIL,[Banerjee_2012] as an example, the Lennard-Jones potential calculation represents up to 90% of the calculation time for 100,000 sampling points per sphere (using the initial algorithm). The number of rejections increases with higher values of the factor μ. However, excessive rejections can adversely decrease the accuracy of the results. To strike a balance, a statistical analysis was performed to determine the optimal value of μ, thereby enabling faster sampling without compromising the accuracy of the enthalpy calculation. The results of this analysis are depicted in Figure 3.11.

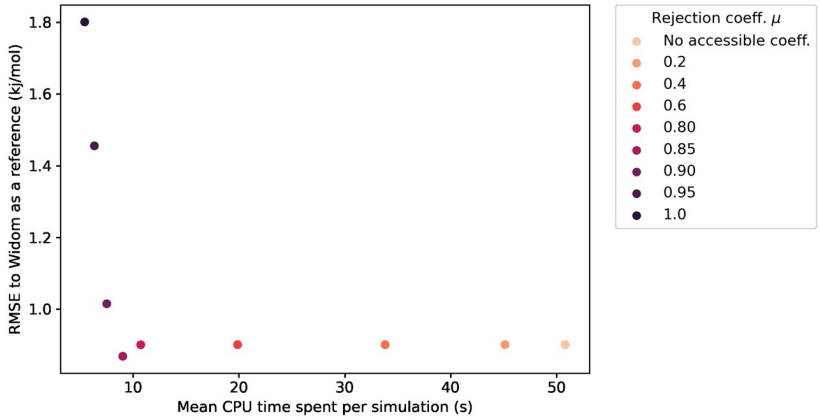


Figure 3.11: Influence of the rejection coefficient μ on the average CPU time required for a simulation of 100k sampling points and the RMSE compared to the reference adsorption enthalpy. The averaging is done only on the structures with the largest cavity diameter (LCD_{CCDC}) superior to 3.7 Å.

The values of RMSE and time presented in Figure 3.11 are averaged only for a subset of the most relevant structures regarding xenon adsorption ($LCD_{CCDC} \geq 3.7$ Å). For $\mu \leq 0.85$, an increase in the value of μ improves the computational speed without affecting the RMSE.[1] For high values of μ, the rejection condition becomes overly stringent, leading to the rejection of points with non-negligible contribution to the overall enthalpy. The RMSE increases as a result. To maintain the same level of accuracy, the optimal value should be $\mu \simeq 0.85$, as it provides the lowest computation time with a good RMSE. However, in specific cases, it may be feasible to explore higher values of μ that trade a slightly reduced accuracy for further gains in speed.

[1]It should be noted that a decrease in accuracy is observed for structures with small pores due to the high probability of rejection within confined spaces, where all sampled points are ultimately rejected. However, these points are not considered when applying a filter on the cavity size ($LCD_{CCDC} \geq 3.7$ Å).

In the simulations depicted in Figure 3.11, the use of a rejection condition $\mu = 0.85$ results in a four-fold acceleration of the simulation compared to the standard algorithm. In the following section, the combination of optimal values for the λ and μ parameters generates an algorithm with highly favorable performance in comparison to Voronoi sampling or Widom insertion methods.

3.2.3 Final surface sampling implementation

Performance comparison

For the calculation of adsorption enthalpy, the proposed surface sampling method strikes a balance between the accuracy of Widom insertion (full sampling of the porous space), and the speed of less accurate methods such as Voronoi sampling. The performance of the algorithm, incorporating the two new features (sampling sphere scaling and rejection criterion), is illustrated in Figure 3.13. This figure showcases the improvements brought about by each feature and provides a comparison with reference simulations. All CPU times are calculated using the possible minimum number of sampling points required for the respective algorithms to achieve convergence. With the implementation of the rejection condition, the surface sampling method is found to outperform Voronoi sampling in terms of computational speed. Moreover, increasing the size of the sampling sphere significantly enhances the accuracy of surface sampling, resulting in an RMSE of 0.33 kJ mol^{-1} and an MAE of 0.21 kJ mol^{-1}. For porous materials from the CoRE MOF 2019 database, the statistically determined set of parameters, ($\lambda = 1.6, \mu = 0.85$), combines the lowest error and the smallest computational cost. By incorporating both of these new features into the algorithm, the final surface sampling method achieves an RMSE of only 0.33 kJ mol^{-1} and an average computation time of 0.34 s per structure. According to the data represented in Figure 3.13, this method is approximately 6 times more accurate and 26% faster than Voronoi sampling, and about 430 times faster than a Widom insertion with 12k cycles.

Finally, 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 the sampling algorithm is to calculate adsorption enthalpy in the zero-loading limit. The method can also calculate the Henry constant and surface area of the materials simultaneously, without incurring significant additional computational cost. The Henry constant serves as a key metric for assessing the affinity of an adsorbate to a nanoporous

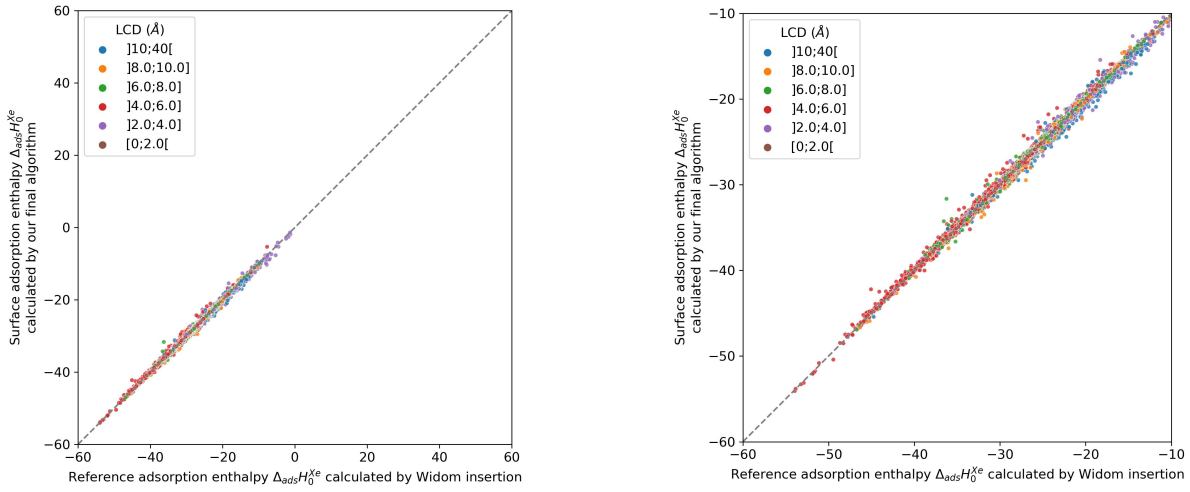


Figure 3.12: Scatterplots of the xenon surface adsorption enthalpy calculated by the final RAESS algorithm (λ = 1.6 and μ = 0.85) as a function of the xenon adsorption enthalpy calculated by a 100k-step Widom insertion simulation using two value windows, in structures of CoRE MOF 2019 with $LCD_{CCDC}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

structure. The Xe/Kr gas selectivity at low pressure is defined as the ratio of the Henry constants of Xe and Kr. This important property can be determined using Equation 2.16 in a Widom insertion calculation. Instead of utilizing the interaction energies at the Widom inserted points, an approximate value for the Henry constant can now be obtained using the surface sampled points.

By employing the optimized set of parameters for surface sampling, the algorithm's performance in estimating the Henry constant was assessed by comparing it to the ground truth obtained through 100,000 cycles of Widom insertion. Since the Henry constant corresponds to the exponential of an adsorption free energy and the focus of this study lies on the precision of the free energy, a log-scale evaluation metric is used. For surface sampling, the log-RMSE of K_H is equal to 0.2, indicating that the values are accurately predicted in terms of order of magnitude, as depicted in Figure 3.15. If the derived free energy $\Delta F_{ads} = -RT \log(\rho_f RTK_H)$ is considered, the RMSE is approximately 1.1 kJ mol^{-1}, and this level of error is achieved within a similar amount time of approximately 1 s (Figure 3.15). In contrast, for Widom insertion, a similar level of error is attained within a similar time frame, and an RMSE of approximately 0.1 kJ mol^{-1} is achieved within 86 s (Figure 3.15). For free energy calculation, surface sampling converges 86 times faster. If the main focus is on adsorption enthalpy, the Henry constant can be computed with minimal additional computational cost and reasonable accuracy, thereby obtaining two thermodynamic properties of interest for the cost of one.

Similarly, the algorithm can be adapted to determine the surface area of the material by counting the number of points within the sampling spheres that possess negative energy and represent the points where guest molecules can interact favorably. By dividing this count by the total number of sampled points, the proportion of adsorbable area is obtained for each sphere.

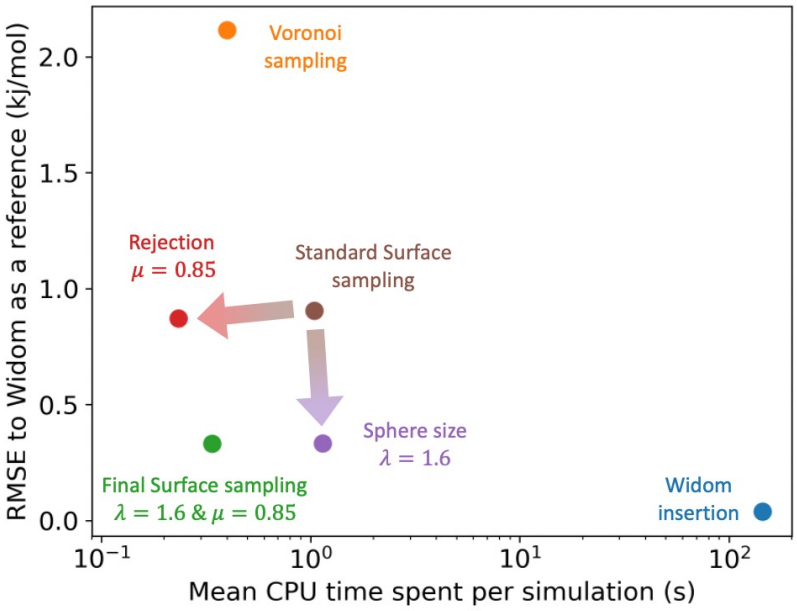


Figure 3.13: Comparison of the RMSE to the reference Widom insertion and the average computation time for different types of enthalpy calculation methods. The surface sampling calculation was all done with 2k sampling points on each sphere and the Widom simulations were done using 12k cycles. These values correspond to the value at the convergence identified using Figure 3.8.

Summing these proportions over all atoms yields the total surface area. This implementation is summarized in equation 3.8:

$$\mathrm{SA} = \frac{1}{\mathrm{V}} \sum_{a \in \mathrm{cell}} \frac{\mathrm{N}_{\mathrm{accessible}}(a)}{\mathrm{N}_{\mathrm{total}}} 4\pi r(a)^2 \tag{3.8}$$

where V is the volume of the cell a atoms of the cell; $\mathrm{N}_{accessible}(a)$ is the number of accessible points around the atom a; N_{total} is the number of sampling points; $r(a)$ is the radius of the sampling sphere around the atom a. When $\lambda = 1$, spheres with a radius σ are sampled, which is equivalent to considering hard spheres defined by σ (a convention used by RASPA2 to calculate surface areas). When comparing simulations with $\lambda = 1$ to those obtained by RASPA2, the surface areas are found to be very close (Figure 3.16). However, when considering $\lambda = 1.6$, the previously observed perfect agreement is lost, and the points show weak correlation in log-scale (Figure 3.16). This difference can be attributed to the larger sphere size, which also alters the proportion of adsorbable points. The relationship between these two adsorption surface areas is far from trivial. Due to the relatively low computational cost of surface area calculation, this implementation would not be highly useful, except for obtaining a rough estimate of the surface area.

3.2.4 Surface sampling application use cases

After introducing the performance of the surface energy sampling algorithm for xenon and specific materials from CoRE MOF 2019 at 298 K, further investigations on other conditions will be conducted to test the transferability of the methodology. First, the algorithm will be used to assess the xenon/krypton selectivity at infinite dilution, in comparison to the standard

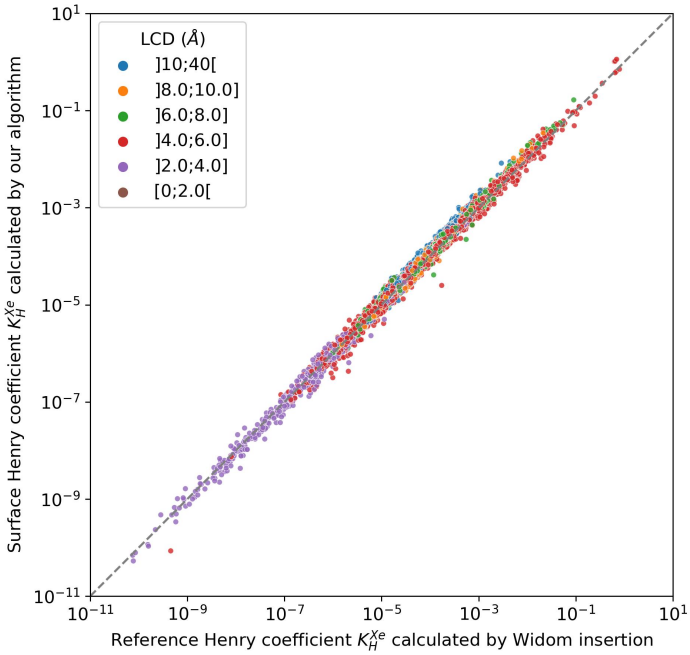


Figure 3.14: Scatterplots of the xenon Henry constants calculated by the RAESS algorithm compared to the ones calculated by a 100k-step Widom insertion simulation using two value windows.

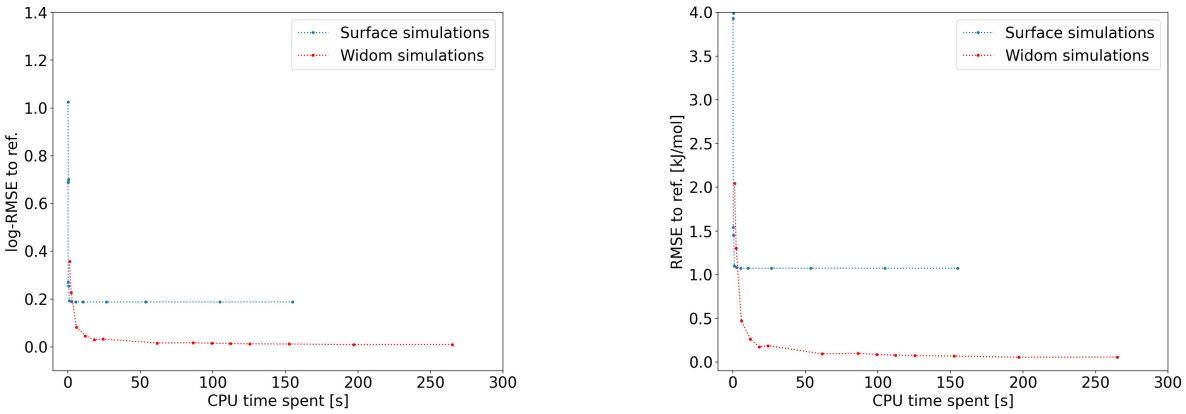


Figure 3.15: Left: convergence plot of the log-RMSE on the xenon Henry constants for both the surface sampling and the Widom insertion. Right: convergence plot of the RMSE on the xenon adsorption Gibbs free energy for the final implementation of the surface sampling and the Widom insertion.

Widom insertion. Secondly, the influence of temperature on the algorithm's performance will be compared, as the performance may be less optimal due to the less concentrated Boltzmann weights on the less attractive points. Lastly, the RAESS algorithm will be tested on databases containing diverse materials.

Selectivity Calculations

The selectivity value, which is the most important metric in evaluating the Xe/Kr separation performance of a nanoporous material, is examined in this study to assess whether a sur-

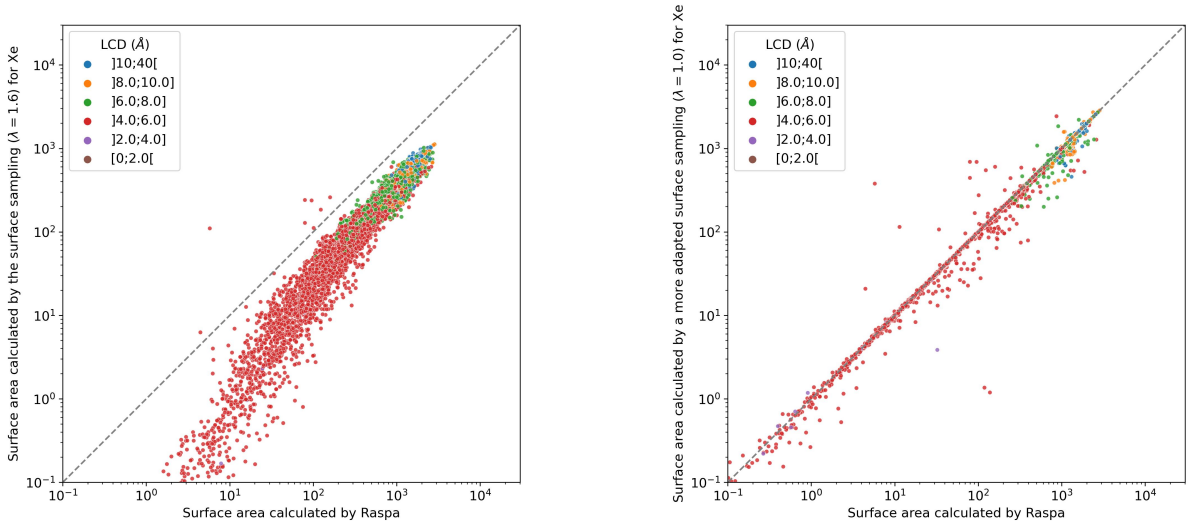


Figure 3.16: Scatterplots of the surface areas calculated by the RAESS algorithm with two different parameterizations compared to the surface area given by a RASPA2 surface area calculation. The left plot corresponds to the surface sampling described in the section 3.2.3 with $\lambda = 1.6$ and $\mu = 0.85$, while the right plot uses a sampling sphere near σ with $\lambda = 1.0$. The second parameterization is much closer to what a RASPA2 sampling based on the σ parameter of a LJ potential does, hence explaining the much better accordance.

face sampling technique can accurately evaluate this metric although it is limited by all the approximations inherent to the technique.

A few precautions should be considered before blindly using the algorithm for selectivity prediction. During the investigation of selectivity calculation, it was observed that the rejection condition on xenon can be high, as the focus if this study is on identifying the most favorable materials for xenon adsorption. However, for krypton, it is necessary to accurately describe very low Henry constants, as a selective material would also exhibit unfavorable characteristics for krypton. Therefore, the parameter μ needs to be chosen wisely, ensuring that it is low enough to obtain accurate Kr Henry constant and selectivity values.

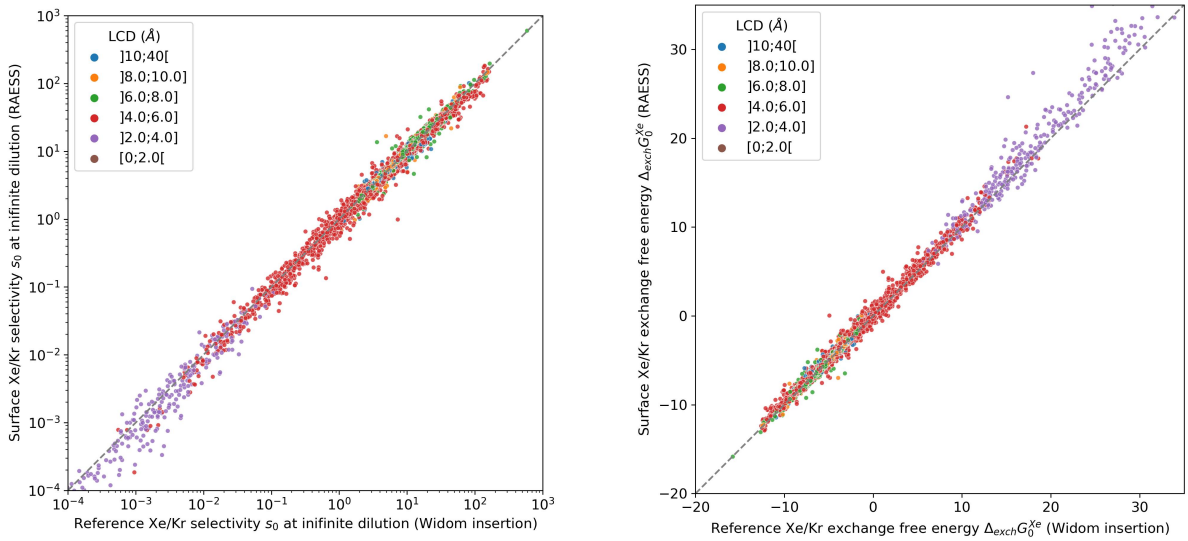
As shown in Table 3.1, the error in selectivity highly depends on the μ value, which determines the exclusion of points at $\mu\sigma$ from a framework atom center. Intuitively, a lower μ enables the sampling of higher energy values that contribute to the Boltzmann averaging. Additionally, dividing by smaller values can amplify any errors in the values, and this effect can be mitigated by increasing the number of sampled points.

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 this initial study, the optimal value is $\mu = 0.5$, as it provides the best accuracy with minimal CPU time. This value will be utilized for krypton in order to conduct a comprehensive study on the performance on the Xe/Kr selectivity for materials from CoRE MOF 2019. The following study will thus 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 metric. By applying the log10 to the selectivity values, the resulting RMSE and MAE are about 0.064 and 0.04 respectively. This implies that the error in comparing the orders of magnitude of the selectivity is around 0.06. For instance, if a selectivity value is predicted to be $s = 10^{-7}$, the actual value s would statistically fall within the range $[10^{-7.06}, 10^{-6.94}]$.



(a) Xe/Kr selectivity (log-scale)

(b) Xe/Kr exchange free energy

Figure 3.17: (a) Scatterplot comparison of the Xe/Kr selectivity calculated by RAESS algorithm and the one calculated by the Widom insertion (in log scale). (b) 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}). Both graphs are color-coded by the cavity size (LCD).

To provide a thermodynamic interpretation, the exchange Gibbs free energy associated $\Delta_{\text{exch}}G_0^{\text{Xe/Kr}}$ with the selectivity defined in the previous chapter (equation 2.26) can be utilized. By using this exchange Gibbs free energy, the assessment of the approach's performance becomes much more straightforward. The resulting RMSE is about 0.36 kJ mol^{-1}. It is not possible to directly compare this error with the errors associated with adsorption enthalpy, as the ranges and interpretations differ significantly. In this case, selective materials exhibit a negative value of $\Delta_{\text{exch}}G_0^{\text{Xe/Kr}}$, ranging up to a maximum of approximately −12.7 kJ mol^{-1}. The relative error is naturally higher for the Gibbs free energy, which can be attributed to the increased uncertainty in the Henry constant and the denominator term introduced by krypton.

To assess the performance of the RAESS algorithm in real screening scenarios, the top 100 most selective materials identified by RAESS and a Widom simulation (RASPA2) were compared in this study. It was observed that 83 structures out of the top 100 materials identified by RAESS

are also included in the top 100 materials obtained through Widom insertion. Although the correlation is not perfect, there will inevitably be some variation in the ordering of the top 100 materials provided by these two methods. The fact that 83% of the materials overlap indicates a relatively narrow difference. Expanding the comparison to the top 150 materials from the Widom simulation, it was found that 94 of them are present in the top 100 materials identified by the surface simulation. This suggests that the RAESS algorithm successfully identifies a large majority of the top candidates obtained through the Widom insertion simulation.

A Higher Temperature

The RAESS method relies on the higher weight of the strong sites close to the surface of the pores. With an increase in temperature, the role of less attractive sites would become more significant, resulting in an expected decrease in the method's accuracy. To understand this limitation of the RAESS algorithm at higher temperatures, a comparison at 600 K and 1 atm was made using the CoRE MOF 2019 database.

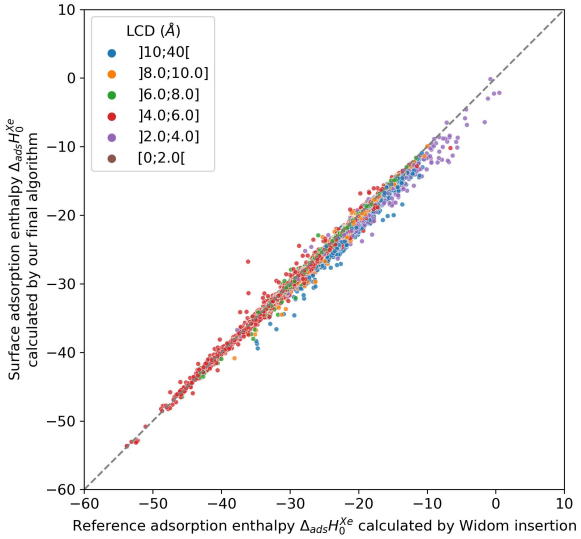


Figure 3.18: Scatterplot of the enthalpies calculated by our final algorithm (λ = 1.6 and μ = 0.85) compared to the enthalpies calculated by a 12k step Widom insertion simulation of xenon in structures of CoRE MOF 2019 with $LCD_{CCDC} \geq 3.7$ Å at 600 K.

As expected, the surface sampling method exhibits lower accuracy when subjected to Boltzmann averaging at 600 K than at ambient temperature. Nevertheless, it still demonstrates an acceptable correlation in adsorption performance, yielding an RMSE 0.70 kJ mol^{-1} and a MAE of 0.41 kJ mol^{-1}. The errors nearly doubled when the temperature was increased from 298 K to 600 K. However, these limitations of the method are not debilitating, as adsorption processes are typically not conducted at very high temperatures. High temperatures are commonly employed in temperature swing adsorption (TSA) for desorbing the adsorbates rather than adsorbing them.

Other databases

ToBaCCo: In this study, a total of 1,000 structures were randomly selected from the 13,511 porous frameworks within the ToBaCCo database[1] to assess the robustness of the RAESS method on a database other than CoRE MOF. Due to the presence of larger pores in the ToBaCCo structures than in other databases, as indicated in a recent study,**Moosavi_2020** these materials exhibit a higher degree of unfavorability towards the adsorption of small molecules (such as Xe). The correlation observed in Figure 3.19 is relatively weak compared to the CoRE MOF 2019 database. It is important to consider this reduced accuracy in light of the unsuitability of these materials for Xe/Kr separation. Moreover, it should be noted that points displaying weaker correlations correspond to those with an LCD_{CCDC} greater than 10 Å, which is suboptimal for Xe-Kr separation.

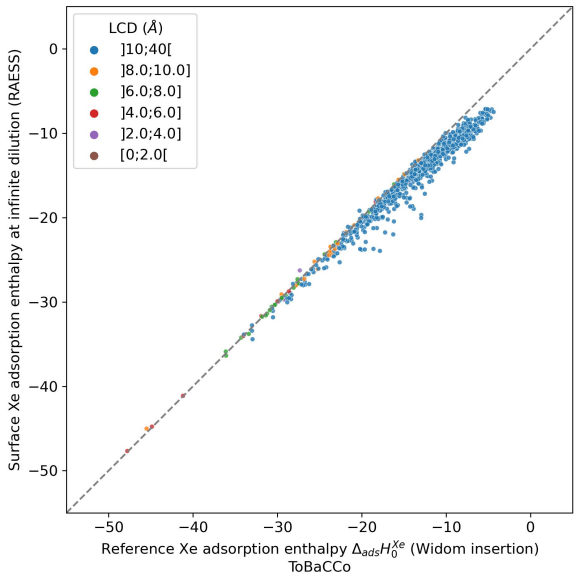


Figure 3.19: Scatterplot comparison of the xenon adsorption enthalpy calculated by the RAESS algorithm and the Widom insertion (RASPA2) on the ToBaCCo database. RMSE = 1.79 kJ mol^{-1} *and MAE* = 1.48 kJ mol^{-1}. *It can be noted that 915 structures have the* LCD_{CCDC} *greater than* 10 Å.

The algorithm, however, demonstrates excellent performance when applied to highly adsorptive materials with xenon adsorption enthalpy values lower than −30 kJ mol^{-1}. This result is primarily due to the proximity of the adsorption sites to the material's surface. For broader pore sizes, some limitations of the methodology become apparent, and it is crucial to acknowledge them. These limitations do not significantly affect the final results when determining the most attractive materials. Moreover, it should be emphasized that this limitation does not have a significant detrimental effect, as the correlation, although weakened, remains intact and does not disappear.

[1]The Topology-Based Crystal Constructor or ToBaCCo corresponds to a topology-based computationally constructed MOFs database.

Amorphous materials: To further extend the potential use cases of the RAESS algorithm, an amorphous material database[1] was subjected to testing with the RAESS algorithm that found results for 176 structures out of 205 — the rejection condition of RAESS does not calculate the adsorption enthalpy of materials with pore sized that cannot fit xenon (there are 20 such structures in the database for μ = 0.85); the remaining 9 structures have been aborted probably due to memory issues considering the high number of atoms inside the structures. The RASPA2 software could not be executed on these amorphous structures with the computers used in this study that ran out of memory due to the large system size. Therefore, no comparison with a Widom simulation could be made. However, an alternative simulation method, which utilizes a homogeneously distributed grid sampled by an optimized algorithm presented in the next section, was employed. This grid sampling approach successfully computed the adsorption energies of 175 structures.

Table 3.2 presents the values of the adsorption enthalpies and Henry constants of selected amorphous materials, along with the corresponding computation times. The substantial number of atoms in each structure significantly increases the required CPU time compared to the crystalline structures of CoRE MOF 2019. Nonetheless, the time requirements remain manageable within a hypothetical screening procedure. Considering all 175 structures that were computable using our methods, the average time required per structure is approximately 75 s.

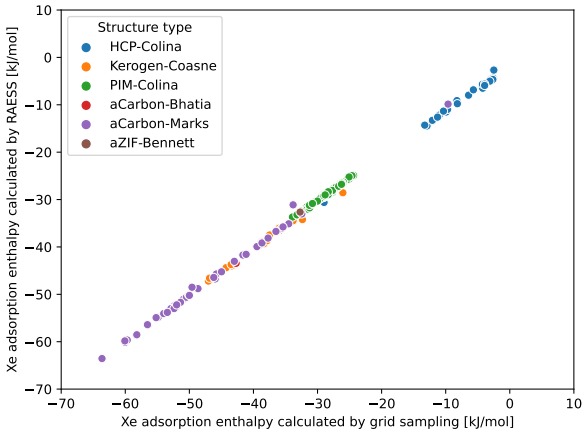


Figure 3.20: Scatterplot comparison of the xenon adsorption enthalpy calculated by the RAESS algorithm and the one calculated by a grid sampling (presented in the next section) on a database of porous rigid amorphous materials.[Thyagarajan_2020] *RASPA2 simulation could not be run on this database. Only the 175 structures computed by both methods are presented here.*

As shown in Figure 3.20, the accuracy of the surface sampling is demonstrated to be high, as evidenced by the highly similar results obtained through unbiased grid-based sampling. The RMSE is about 0.83 kJ mol^{-1}, which is higher than the one for CoRE MOF structures. This method has the potential to serve as a rapid screening tool for evaluating amorphous materials, especially considering the computational time required by the optimized grid sampling is about

[1]This database compiles 205 amorphous nanoporous materials from different classes (polymer of intrinsic microporosity, amorphous carbon, hyper-cross-linked polymer, kerogen, amorphous ZIF, cement) for computational simulation studies[Thyagarajan_2020]

623 s. The dimension reduction inherent to surface sampling makes it one order of magnitude faster than conventional techniques, even for disordered phases.

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`.

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

3.2.5 Perspectives of surface sampling

Here, a novel algorithm for the high-speed calculation of adsorption enthalpy in nanoporous materials has been described, employing a unique approach that significantly reduces the required sampling. Based on the core principle of dimensional reduction from a volume problem to a surface one, this algorithm outperforms the reference Widom insertion method (random sampling of porous space) in terms of both computational speed and accuracy, with an error on the order of 0.4 kJ mol^{-1} observed across the entire CoRE MOF 2019 database for xenon adsorption enthalpy. Furthermore, compared to existing fast sampling techniques such as Voronoi sampling, the surface sampling technique achieves similar CPU time requirements while offering better accuracy.

Based on these results, this algorithm has considerable potential for applications within current computational analysis workflows for material databases, particularly in high-throughput screening studies. For instance, it can be used to rapidly approximate the low-loading adsorption enthalpy of a molecule in nanoporous materials, allowing for the screening of structures with limited affinity for the target adsorbate molecule. It can also serve as a thermodynamic descriptor for selectivity prediction in machine learning models, as demonstrated by Simon et al.[Simon_2015] The computational speedup achieved by this novel methodology also enables the screening of larger-scale materials databases in the future.

It should be noted that the speed of this method primarily lies in the sampling technique itself, rather than the actual energy calculation. While the benchmarking in this work focused on a simple Lennard-Jones interaction potential, the surface sampling technique can equally be applied to accelerate samplings coupled with more computationally expensive modeling strategies, such as polarizable forcefields or density functional theory (DFT) calculations. In the literature, the need for affordable *ab initio* grade thermodynamic properties is typically addressed by employing an importance sampling method based on a classical force.[Vandenbrande2018] In this new method, the description of surface sampling remains independent of any forcefield, and the sampling spheres can be defined based on kinetic radius, van der Waals radius, or any other physically relevant distance. As a result, given a definition of atomic radii, it is possible to define a surface on which other types of simulations, such as neural network potentials, DFT, or other forcefields, can be conducted. While the accuracy and relevance of such sampling methods remain open questions, the approach undeniably accelerates simulations. This acceleration could also be applied to the calculation of adsorption enthalpies while considering intrinsic

structural flexibility,[Witman_2017] a task that is computationally demanding. As surface sampling is hundreds of times faster than standard methodologies, it becomes feasible to utilize hundreds of snapshots in flexibility-aware calculations.

Finally, although the algorithm in its present form can already be applied in a wide range of applications, there is potential for additional development work to generalize it to polyatomic adsorbates. For instance, a definition of the molecular radius for non-spherical adsorbates and consideration of the orientation conformation of the adsorbent would need to be addressed. The distance to the surface could potentially depend on the orientation of the adsorbate or involve sampling a band volume on the surface. Although determining the best implementation of surface sampling for polyatomic adsorbates remains an open question, in theory, it should be feasible to apply it to more complex adsorbates than spherical noble gases. This would add more complexity to the algorithm without altering the fundamental speedup achieved through surface sampling, as similar orientation moves are performed in other standard methodologies. To further improve accuracy, hybrid samplings with multiple sampling spheres or a combination of Voronoi nodes and sampling spheres could be tested. Another possibility is to incorporate fractions of spheres oriented towards the center of the pore defined by the Voronoi node. In theory, a wider variety of sampling points can only enhance the sampling process. Thus, there are multiple potential sampling techniques that could be developed based on 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.

3.3 Grid Adsorption Energies Descriptors (GrAED)

To conclude this overview of novel energy sampling methods, a revised version of the standard grid sampling will be presented. Grid sampling is the most accurate approach as it directly relies on the averaging definitions in equations 3.7 and 2.14. In this section, the inherent symmetry operations of most material structures and the removal of framework occupied space will be leveraged to accelerate this typically slow method. This exhaustive approach allows for the calculation of energy distributions that are less biased compared to other methods. These energy distributions serve as fundamental building blocks for the prediction of ambient-pressure selectivity, which will be discussed in the next section.

3.3.1 Implementation of an efficient grid algorithm

To build more relevant energy descriptors, it is necessary to return to the definitions of adsorption enthalpy and Henry constant (equation 2.22 and 2.14), as the latter require a homogeneous sampling of the adsorption space. The simplest way to achieve this consists in laying a grid in the 3D space. However, this method is known to be time-consuming in theory, as it relies on an exhaustive sampling of all space; random sampling or biased sampling, on the other hand, usually reduces the number of sampled point. Inspired by the work on surface sampling, an approach based on a symmetry-respecting grid was designed by leveraging algorithms from the Gemmi Project.[Wojdyr_2022] In this grid adsorption energy descriptor (GrAED) calculation algorithm, these new features, combined with grid sampling, significantly reduce the computational time required for adsorption energy calculations while maintaining high accuracy.

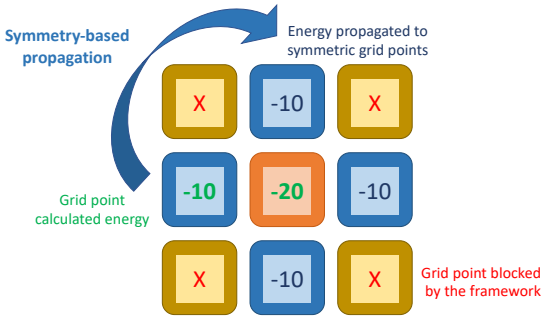


Figure 3.21: Principle of the energy sampling on a symmetry-based grid. On the 9 grid points, 4 points are blocked because they are too close to the framework atoms, 2 points are really calculated using the LJ potential and 3 points are propagated using the inner symmetry of the framework.

The core structure of this novel algorithm encompasses a grid algorithm, where the evaluation of the interaction energy at each point of the preset grid over the structure's unit cell is required. A naive approach would demand an expensive energy calculation at each grid point. To improve this approach, two main simplifications are incorporated into the algorithm — 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. Contributions of high-energy values to the thermodynamic quantities presented in the section 2.1.5 are negligible. Hence, by employing a rejection parameter similar to the one developed for surface sampling as shown in the section 3.2.2, the interaction energy of the grid points within the sphere of radius $\mu \times \sigma_{g-h}$ can be precalculated. If the interaction energy value is higher than a preset energy threshold E_{th}, the corresponding grid point adopts these values as the interaction energy, and no further calculation is performed for that point. The grid's symmetry is determined based on the structure's symmetry using the grid definition of the Gemmi Project. Through the utilization of symmetry operations on a grid point value, it can be propagated to other symmetry-equivalent grid points, as illustrated in Figure 3.21. Since MOF structures are usually highly symmetric, this approach reduces the computation time required to calculate the interaction energy of a guest molecule at a given grid node with all the surrounding framework atoms within a specified cutoff. Having presented the primary components of our optimized grid calculation, the integration of this calculation in the algorithm's implementation will now be demonstrated.

1. A loop performed 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 this preliminary filtering will be referred to as "blocking" in Table 3.3.

2. A second loop over the "unvisited" grid points is performed — at each increment, if the point is "unvisited", the interaction energy is calculated 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 Table 3.3.

A "fast" version of the grid calculation algorithm was built by combining both the "blocking" of the high energy grid points and the "symmetry" based calculation of the interaction energies. This algorithm, which can compete with the previously developed rapid surface sampling method (RAESS), was built. The spacing between the grid points can be adjusted to control the trade-off between accuracy and computation time, with the computation time theoretically inversely proportional to the cube of the spacing. Interestingly, for certain spacing values, this algorithm can even outperform surface sampling on the CoRE MOF database, where symmetry plays a significant role (see Table 3.3). 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

When considering the performance of this new grid sampling algorithm in comparison to previously introduced sampling algorithms, the utilization of this new sampling technique on the CoRE MOF 2019 database proves to be highly advantageous due to its accuracy and speed. The efficient time performance of the grid sampling on the structures within the CoRE MOF 2019 database can be attributed to the relatively small porosity of the materials and their high degree of symmetry. For example, the average void fraction for a 1.2 Å probe radius is equal to 0.16, while the average number of symmetric images is 5.8 (most MOFs present symmetry operations). As a result of the "blocking" procedure, only approximately ∼16% of the grid points necessitate actual calculation on average. Additionally, the "symmetry" procedure ensures that only around ∼17% of points need to be considered. The combination of both procedures significantly reduces the number of relevant points to merely 2.7% of the grid. This reduction substantially decreases the CPU time required for the calculation while maintaining a satisfactory level of accuracy (with a low error on the Xe adsorption enthalpy of 0.014 kJ mol^{-1}) compared to the naive grid approach, as shown in Table 3.3. With the blocking procedure in the grid simulation, the time required is reduced by ∼70.6% when 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 fast grid sampling technique achieves a time reduction of nearly ∼91.6% for a grid spacing of 0.12 Å, aligning with the aforementioned decreased number of sampled points.

As shown in Figure 3.23, the accuracy of the adsorption enthalpy and the Henry constant is not compromised by the approach. An almost perfect agreement between the Widom insertion method and the grid-based approach can be observed when utilizing a finely meshed grid (0.12 Å spacing). This alignment was expected since both methods involve unbiased sampling of adsorption energies. The figure reveals minimal error in 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 extremely low, at 0.01. This method adheres to the initial definition of these quantities at infinite dilution, explaining the unsurprising nature of this perfect correspondence.

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 the 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 spacing values*(0.12, 0.20, 0.30*).*

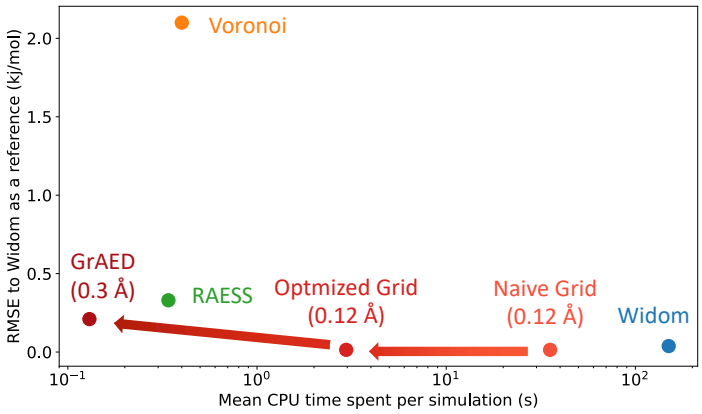


Figure 3.22: Comparison of the RMSE on Xe adsorption enthalpy and the average CPU time required to run a simulation on a structure of CoRE MOF 2019 ($LCD_{CCDC} \geq$ 3.7 Å) for different sampling techniques (Widom, Voronoi, RAESS and GrAED). The values are reported in Table 3.3.

The little computation time required to achieve such an accuracy, however, is much more interesting. When examining the Table 3.3, it can be observed that the highly accurate grid sampling approach attains a similar level of accuracy as a 12k-cycle Widom insertion calculated using the RASPA2 software, but it is 50 times faster. On the CoRE MOF 2019 database, by utilizing a less stringent grid spacing of 0.3 Å, the GrAED algorithm may even be more interesting than the RAESS algorithm, as it reduces the computation time by half while maintaining slightly higher accuracy. This highly comparable performance to a dimensionally reduced sampling technique can be attributed to two factors of the CoRE MOF database. Firstly, the structures have smaller pores, resulting in a higher surface-to-volume ratio, which increases the computation time for RAESS. Secondly, the highly symmetric nature of CoRE MOF structures significantly reduces

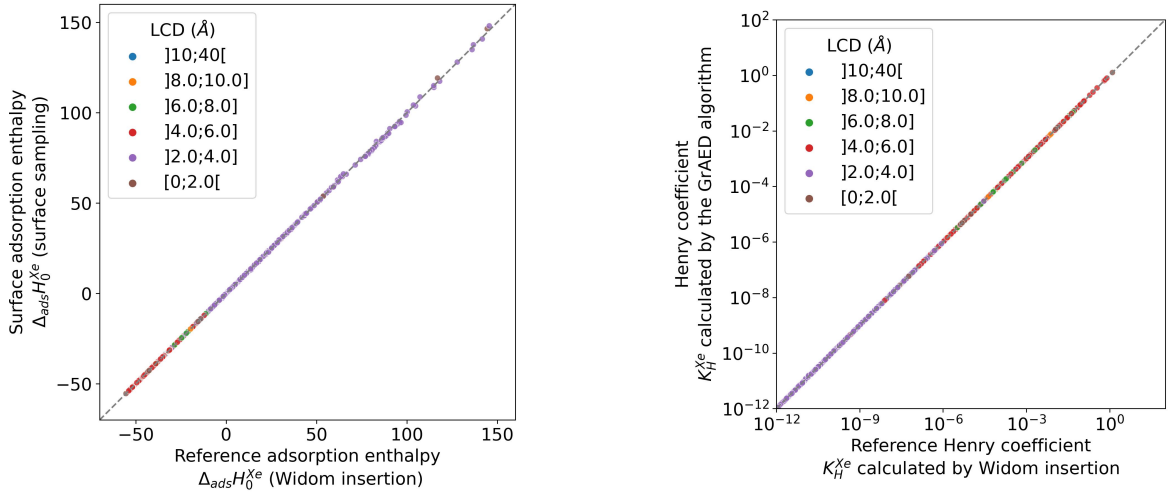


Figure 3.23: Comparison of the xenon adsorption enthalpies (left) and the Henry constants (right) calculated by the optimized grid energy sampling (for a 0.12 Å *spacing, a rejection parameter* $\mu = 0.8$ *and an energy threshold* E_{th} *of* 100 kJ mol^{-1}*) and by the Widom insertion of RASPA2 with 100,000 cycles on the CoRE MOF 2019 structures (*$LCD_{CCDC} \geq 3.7$ Å*).*

the computation time required for GrAED, which is not the case for other databases such as ToBaCCo or the amorphous database previously examined in the section 3.2.4.

For instance, on the amorphous database (see section 3.2.4), the computation time for grid sampling is found to be 750 times longer compared to surface sampling, with an RMSE of only 0.83 kJ mol^{-1}. In the case of amorphous databases, surface sampling outperforms exhaustive grid sampling due to the minimal reduction in the number of sampled points caused by symmetry and overlap considerations, thereby showcasing the greater impact of dimensionality reduction achieved through surface sampling. In the ToBaCCo database,**Colon_2017** where symmetry no longer plays a significant role and the pores are larger, resulting in fewer points obstructed by the framework, the performance of grid sampling is directly affected when compared to the RAESS algorithm. The average time required for the thousand structures in ToBaCCo, as considered in the section 3.2.4, is now 735 s, in contrast to less than 2 s for surface sampling. By increasing the grid spacing to 0.3, a computational time reduction to approximately 47 s can be expected (deduced using a rule of three). However, the accuracy is significantly higher than that of surface sampling (Figure 3.24), reaching an extremely low RMSE of 0.02 kJ mol^{-1}. Depending on the number of structures and their nature (symmetry, porosity), a choice between the more efficient yet less accurate RAESS and the GrAED software must be made.

From the energy values of this grid, a multitude of valuable descriptors for the adsorption process can now be calculated. The performance has been assessed for Xe adsorption enthalpy and Xe Henry constant, as discussed in the section 2.1.5. Additionally, the Xe adsorption Gibbs free energy and Xe adsorption entropy can be derived. With the inclusion of krypton alongside xenon, the thermodynamic quantities for Kr adsorption can be naturally evaluated. Furthermore, the exchange thermodynamic quantities, particularly the Xe/Kr selectivity (the key metric for assessing the separation process of interest), can also be determined.

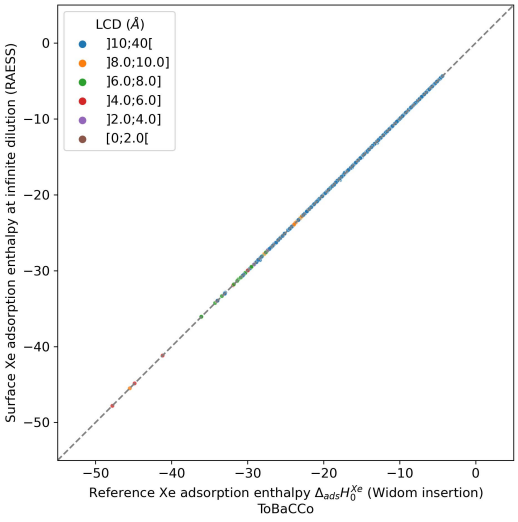


Figure 3.24: Comparison of the xenon adsorption enthalpies (left) and the Henry constants (right) calculated by the optimized grid energy sampling (for a 0.12 Å *spacing, a rejection parameter* $\mu = 0.8$ *and an energy threshold* E_{th} *of* $100\,\mathrm{kJ\,mol^{-1}}$*) and by the Widom insertion of RASPA2 with 100,000 cycles on 1000 randomly selected structure of the ToBaCCo.*[Colon_2017]

3.3.3 Performance on the exchange equilibrium

The Xe/Kr selectivity is commonly used to characterize the competitive adsorption of a binary mixture of xenon and krypton. Unlike a single-component metric such as the Henry constant, the relative uncertainty in the selectivity inherently increases since it involves the quotient of the Henry constants of the competitive adsorbates. In this section, the objective is to quantify this error and determine its relevance in characterizing the separation using the optimized grid sampling method.

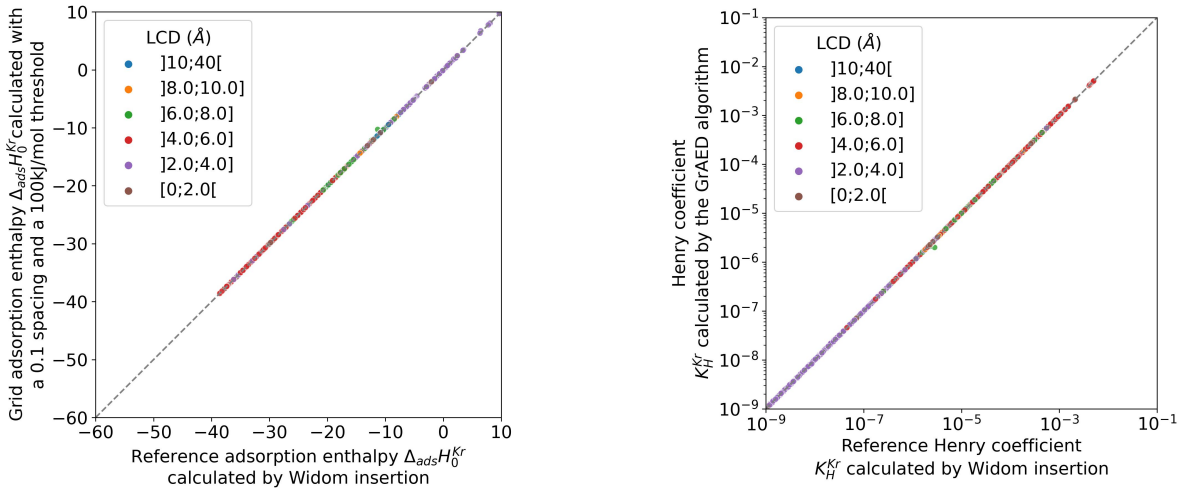


Figure 3.25: Comparison of the krypton adsorption enthalpies (left) and the Henry constants (right) calculated by the optimized grid energy sampling (for a 0.12 Å *spacing, a rejection parameter* $\mu = 0.8$ *and an energy threshold* E_{th} *of* $100\,\mathrm{kJ\,mol^{-1}}$*) and by the Widom insertion of RASPA2 with 100,000 cycles.*

First, the adsorption properties of krypton were also calculated using the same grid spacing of 0.12 Å. The accuracy achieved is approximately equivalent, with an RMSE and MAE on the krypton adsorption enthalpy of around 0.02 kJ mol^{-1} and 0.01 kJ mol^{-1}. As shown in Figure 3.25, there is a strong correlation observed for both the adsorption enthalpy (on a linear scale) and the Henry constant (on a logscale). The RMSE for the base 10 logarithm of the Henry constant (in mmol g^{-1} Pa^{-1}) is typically 0.002, which is similar to the accuracy obtained for xenon. The relative error in the adsorption enthalpies of xenon and krypton does not exceed 0.1% (values of the enthalpy have order of magnitude of dozens of kJ mol^{-1}), and thus, the error in the xenon/krypton exchange enthalpy is expected to be very close to this value. Consequently, there is no significant impact on the exchange enthalpy. To evaluate the selectivity, it is necessary to consider the relative error in the adsorption free energy, which is a logarithmic transformation of the Henry constant. This relative error can be estimated to be around 0.2% (for a Henry constant of 10–4 mmol g^{-1} Pa^{-1}), which is also approximately the expected relative error in the exchange Gibbs free energy or the logarithm of the selectivity.

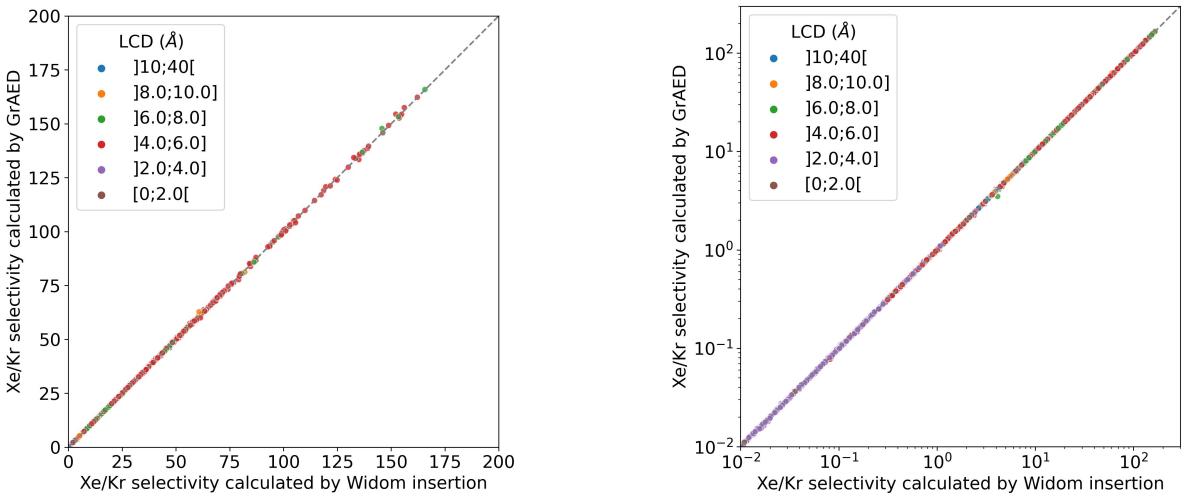


Figure 3.26: Comparison of the Xe/Kr selectivity calculated by the optimized grid energy sampling (for a 0.12 Å *spacing, a rejection parameter* $\mu = 0.8$ *and an energy threshold* E_{th} *of* 100 kJ mol^{-1}*) and by the Widom insertion of RASPA2 with 100,000 cycles. On the left, the axes are in linear scale, whereas the log scale has been used on the right.*

Figure 3.26 demonstrates that the selectivity is accurately represented by the new grid sampling, particularly when considering the logarithmic transformation. The RMSE and MAE for the selectivity values are approximately 0.097 and 0.035, respectively, which are quite low compared to the selectivity values of interest (above 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, indicating a precise understanding of the order of magnitude of the selectivity. If the selectivity were expressed in powers of ten, the exponent would be known with a precision of ± 0.014.

The average computation time required to calculate the selectivity value for a structure in the CoRE MOF 2019 database is approximately 6.5 s, with the krypton component taking about 3.5 s to compute. If an algorithm computes both selectivity values simultaneously, it is possible to save the time required for initializing the software, which can marginally improve this

overall time. This computation time is still much lower than the time required for two Widom insertions.

Having demonstrated the high accuracy and efficiency of the GrAED algorithm for evaluating selectivity at low pressures, the next step is to investigate relationships between descriptors obtained using the grid-based algorithm and the selectivity at ambient pressure.

3.3.4 Description of the ambient-pressure selectivity

Upon initial observation of the left plot in Figure 3.27, the selectivity at ambient pressure shows no correlation with the selectivity at infinite dilution. This suggests that the sampling performed may be ineffective in determining the selectivity values at higher pressures. However, the right plot suggests the existence of a correlation between the logarithm of the selectivity values. The absence of correlation observed in the linear scale plot is actually a phenomenon specific to highly selective materials, which is discussed in detail in Chapter 2. This phenomenon corresponds to a selectivity decrease exhibited by certain highly selective materials (at infinite dilution). In simpler terms, the saturation of the most selective sites diminishes the selectivity of the remaining sites for xenon/krypton separation in these materials.

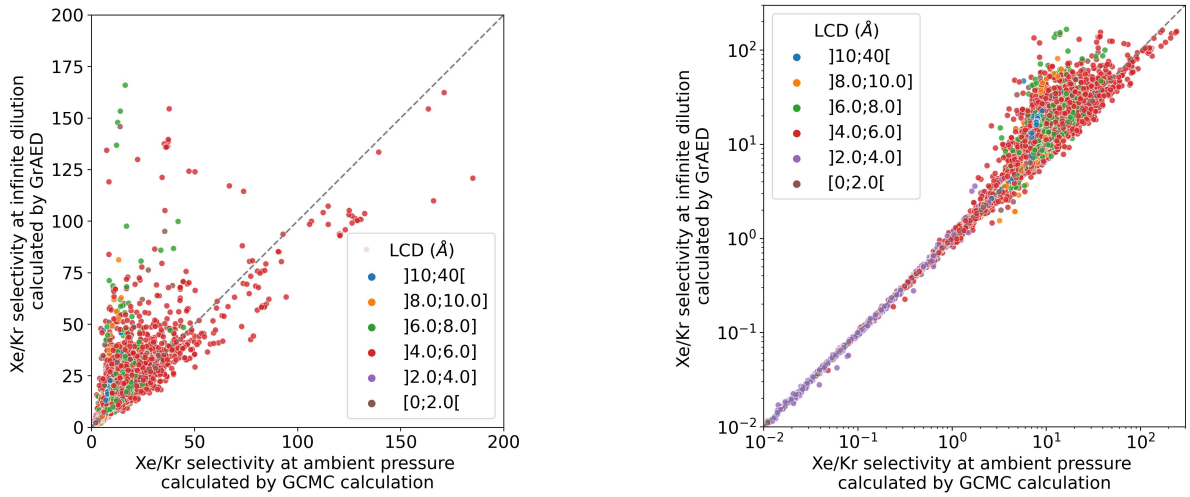


Figure 3.27: Comparison of the low-pressure Xe/Kr selectivity calculated by the GrAED algorithm (same parameters) and the ambient-pressure selectivity calculated by GCMC simulations of RASPA2 with 100,000 cycles. On the left, the axes are in linear scale, whereas the log scale has been used on the right.

The aim is to design descriptors that can help distinguish materials exhibiting a drop in selectivity at higher pressure from those maintaining high selectivity at higher pressure. Three concepts are proposed to gain a better understanding of the origin of this selectivity drop: (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, it is important to note that this method cannot capture guest-guest interactions occurring at higher pressures.

Thermodynamic quantities

In the previous chapter, various thermodynamic quantities were calculated at infinite dilution, including adsorption and exchange Gibbs free energies, enthalpies, and entropies. For the separation of xenon and krypton, a total of nine different descriptors can be generated. The relationship between these quantities and the selectivity values at high pressure will be examined. In the introduction, the relationship between the exchange Gibbs free energy and the selectivity at infinite dilution was already discussed, as shown in Figure 3.27 that plotted the logarithmic transform of the infinite dilution selectivity. This descriptor holds significant importance as it establishes an initial reference value for understanding the problem. The selectivity at high pressure can be viewed as the selectivity at infinite dilution with an additional shift, accounting for the specific adsorption behavior at higher pressures in a given material.

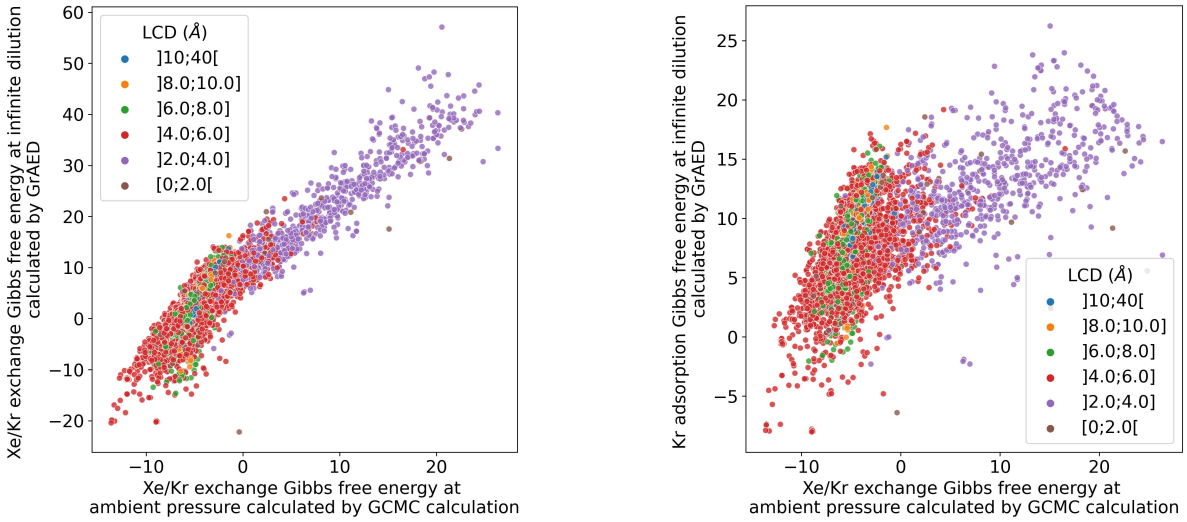


Figure 3.28: Comparison of the ambient-pressure Xe/Kr exchange Gibbs free energy calculated by GCMC simulations of RASPA2 with 100,000 cycles and the low-pressure adsorption free energies of xenon (left) and krypton (right) in kJ mol^{-1} *calculated by the GrAED algorithm (same parameters).*

It comes as no surprise that a good adsorption of xenon is a good indication for the efficiency of the separation from krypton, as shown in Figure3.28, since there is a very strong correlation between the Xe/Kr exchange Gibbs free energy and the xenon adsorption Gibbs free energy. A very weak but positive correlation with the xenon adsorption Gibbs free energy is observed, indicating that a material suitable for efficient Xe/Kr separation would not exhibit very poor krypton adsorption, but rather an average performance. In other words, it is not possible to find a material that is highly effective for xenon adsorption and highly ineffective for krypton adsorption, which explains the theoretical limitation on selectivity, capped under 200 (Figures 3.26 and 3.27, for a UFF level of theory on CoRE MOF 2019). Experimentally, no material has achieved a selectivity value exceeding 100.

The same statement on the importance of the adsorption attractiveness of xenon holds true when looking at the adsorption enthalpies from Figure 3.29. 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 natural solution would, of course, be to include the krypton adsorption performance. The difference of both adsorption enthalpies gives the xenon/krypton exchange enthalpy which

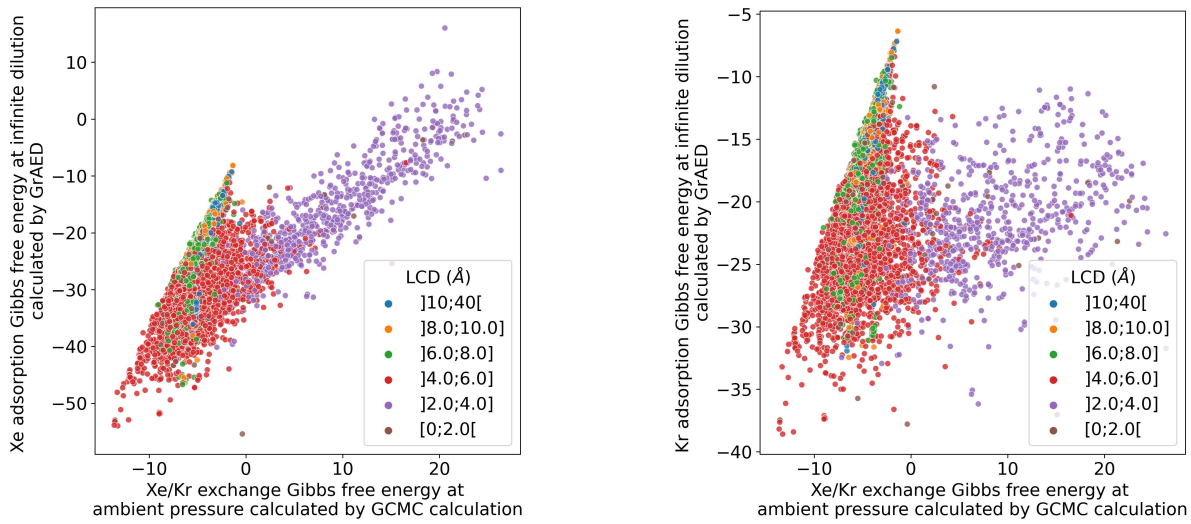


Figure 3.29: Comparison of the ambient-pressure Xe/Kr exchange Gibbs free energy calculated by GCMC simulations of RASPA2 with 100,000 cycles and the low-pressure adsorption enthalpies of xenon (left) and krypton (right) in $\mathrm{kJ\,mol^{-1}}$ *calculated by the GrAED algorithm (same parameters).*

can be used as a separation evaluation metric. The comparison to the krypton adsorption enthalpy alone is not adequate either, the very loose correlation suggests that it is not the main explanatory factor in the separation process.

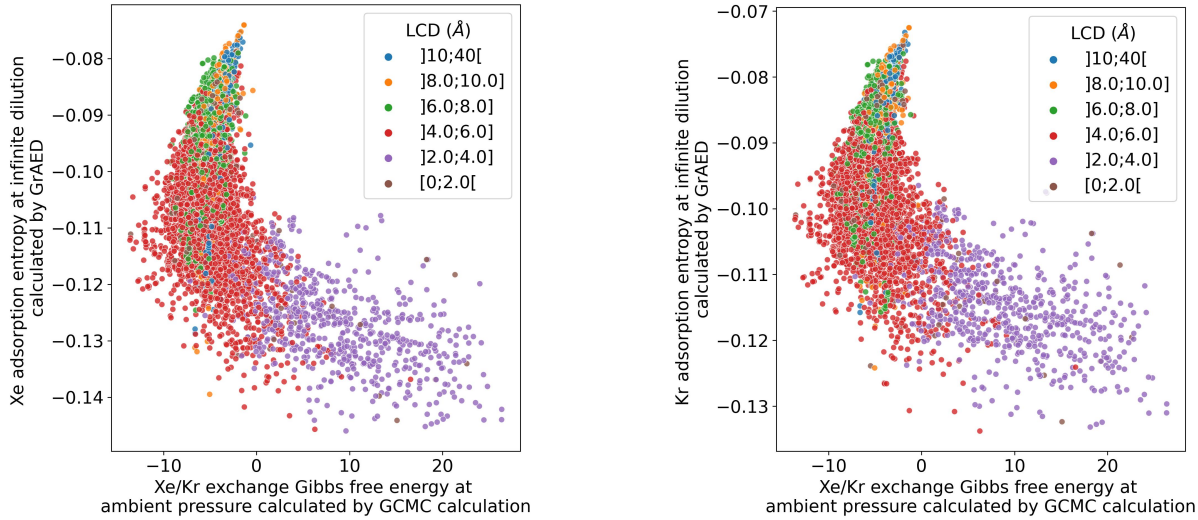


Figure 3.30: Comparison of the ambient-pressure Xe/Kr exchange Gibbs free energy calculated by GCMC simulations of RASPA2 with 100,000 cycles and the low-pressure adsorption entropies of xenon (left) and krypton (right) in $\mathrm{kJ\,mol^{-1}\,K^{-1}}$ *calculated by the GrAED algorithm (same parameters).*

The correlation between the adsorption free energy of xenon and the Xe/Kr exchange free energy at ambient pressure, as well as the weak correlation with xenon's adsorption enthalpy, can be explained by the entropy values. The entropic term (G = H – TS), which represents the difference between enthalpy and free energy, has a minor influence on the correlation, as shown in Figure 3.30. The values of the entropy are relatively stable (ranging from −0.15

to −0.11 kJ mol^{-1} K^{-1}). However, for some structures with ambient-pressure exchange free energy between −10 and 0 kJ mol^{-1}, there is a variation in entropy values ranging from −0.11 to −0.07 kJ mol^{-1} K^{-1} despite having very similar enthalpy values. This discrepancy results in a difference between Gibbs free energy and enthalpy, with a potential span of 12 kJ mol^{-1}, which explains the points deviating from the diagonal in the left plot of Figure 3.29.

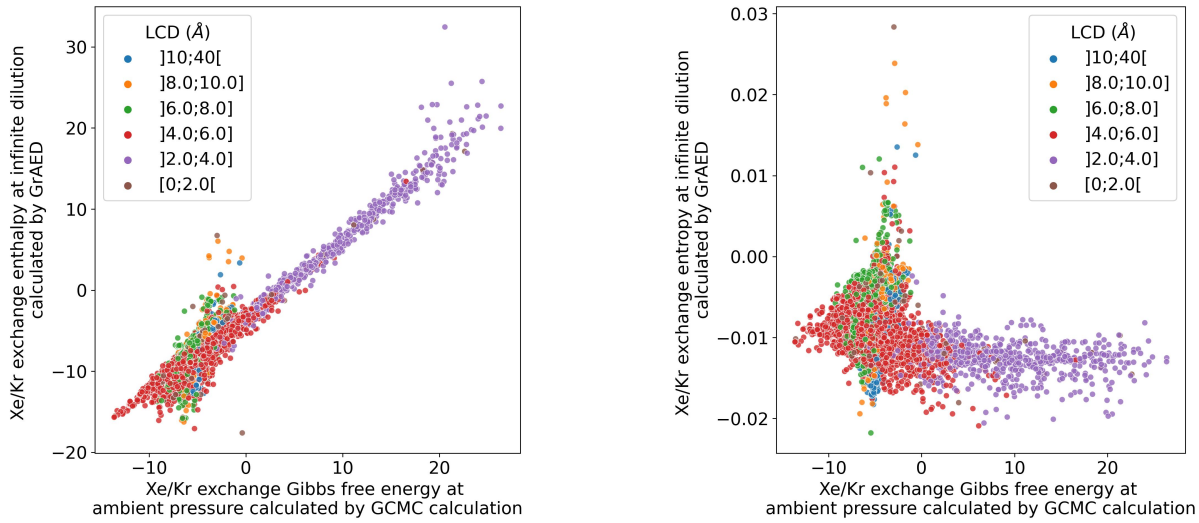


Figure 3.31: Comparison of the ambient-pressure Xe/Kr exchange Gibbs free energy calculated by GCMC simulations of RASPA2 with 100,000 cycles and the low-pressure exchange enthalpy (left, in kJ mol^{-1}*) and entropy (right, in* kJ mol^{-1} K^{-1}*) calculated by the GrAED algorithm (same parameters).*

Upon revisiting the exchange thermodynamic quantities that hold greater relevance to the specific context of this thesis, a notable correlation is observed between the exchange enthalpy at low pressure, calculated using GrAED, and the exchange Gibbs free energy at ambient pressure, calculated by GCMC, as illustrated in Figure 3.31. However, some discrepancies can be detected around the range of −10 and 0 kJ mol^{-1} for the ambient-pressure exchange free energy. These discrepancies can be attributed to the exchange entropy, which remains relatively stable at around −0.01 kJ mol^{-1} K^{-1}, but exhibits a peak for structures within the aforementioned range of ambient-pressure exchange Gibbs free energy. The strong overall correlation can be explained by the enthalpic nature of the separation process of xenon from krypton (Chapter 2). Furthermore, the problematic range, where the correlation weakens, corresponds to the range associated with a drop in selectivity, as illustrated in Figure 3.33. These exchange thermodynamic quantities provide valuable insights for distinguishing materials and improving the modeling of the selectivity drop phenomenon. A more quantitative approach will be developed in the subsequent chapter.

High-temperature quantities

Although the previous quantities provide valuable insights into modeling the adsorption at ambient pressure, they are still insufficient as they pertain to a state where the atoms are adsorbed only on the most attractive sites. The ambient-pressure state, however, is characterized by adsorption on a more diverse set of sites and the increasing significance of the guest-host interaction, which are the main factors contributing to the selectivity difference between the two pressure conditions identified in the previous chapter.

In this section, a descriptor is introduced that provides a better representation of the energy distribution in the ambient-pressure case by assigning greater weight to the more energetic adsorption sites. The simplest approach was to increase the temperature in the Boltzmann averaging for both the Gibbs free energy and the enthalpy, as defined in equations 2.9 and 2.22. Multiple temperatures were tested, and the temperature yielding the higher correlation coefficient between the adsorption enthalpies (at infinite dilution and ambient pressure) was selected.

A temperature of 900 K was found to be the optimal temperature or describing the ambient-pressure adsorption enthalpy of xenon across the structures of CoRE MOF 2019. This choice resulted in a reduced error (RMSE) of 1.76 kJ mol^{-1} compared to 2.87 kJ mol^{-1} for the 298 K case. This improvement has implications for the metrics of exchange free energy and adsorption enthalpy associated with the separation of xenon from krypton. The exchange Gibbs free energy and xenon adsorption enthalpy at ambient pressure exhibit a stronger correlation with their counterparts at lower pressure and higher temperature (900 K) rather than at 298 K. These observations support the use of higher temperature averaging for describing ambient-pressure selectivity.

This new type of descriptor is particularly promising as it performs better in the high selectivity region, where the standard Boltzmann average at very interesting since it performs better around the high selectivity region, where the standard Boltzmann average at 298 K loses its accuracy (Figure 3.27). As shown in Figures 3.32 and 3.33, using averaging at higher temperature yields improved performance in describing the behavior of the most selective materials, while compromising the accuracy of descriptions for less selective materials.

In Figure 3.32, the high-temperature averaging provides a more accurate description of the xenon adsorption enthalpy, with the data points being more centered around the $y = x$ axis, although the correlation is not perfect. Notably, there is greater uncertainty for materials that were initially well predicted as poorly performing materials. The high dispersion around the correlation is likely due to the guest–guest interactions, which are not described in the high temperature averaging but play a non-negligible role in the ambient pressure case.

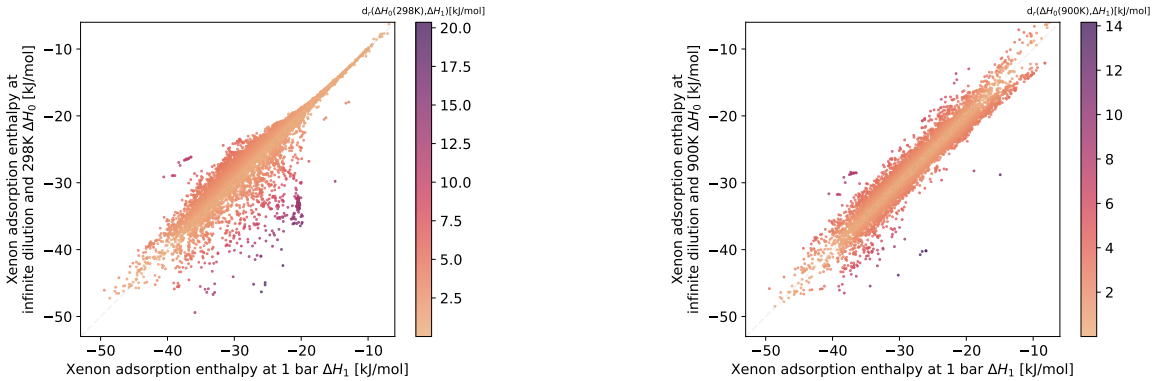


Figure 3.32: Scatterplots of the low-pressure xenon adsorption enthalpy at 298 K *(left) and at* 900 K *(right) calculated by the GrAED algorithm against the ambient-pressure xenon adsorption enthalpy at* 298 K*. Using a higher temperature Boltzmann averaging, the correlation with the ambient-pressure case of interest significantly improves. For instance, the R2 coefficient improves from* 0.80 *to* 0.92*. The RMSE also decreases from* 2.87 kJ mol^{-1} *to* 1.76 kJ mol^{-1}*.*

Figure 3.33 shows that the improvement in xenon adsorption enthalpy does not directly translate into improved performance in the exchange Gibbs free energy. The overall correlation is better between the exchange free energy at 298 K and infinite dilution, and the one at ambient pressure (298 K). However, it can be argued that the exchange free energy at 900 K slightly better describes the materials that experience a selectivity drop, as depicted in Figure 3.27.

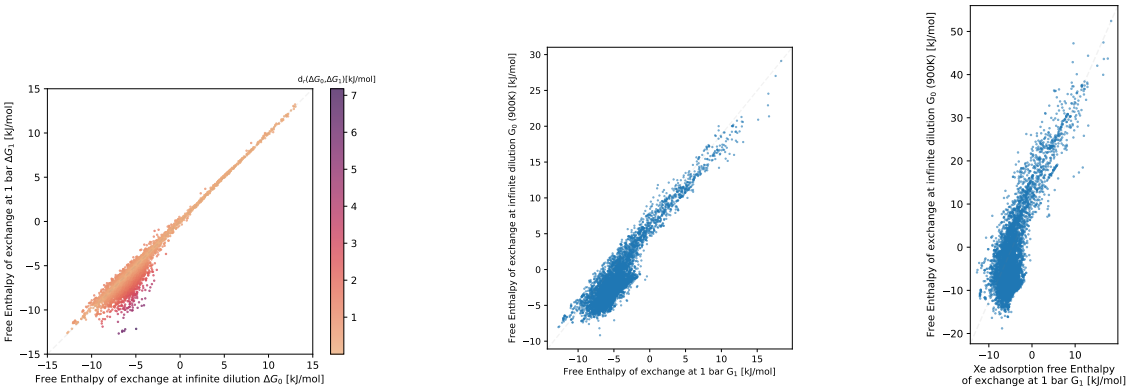


Figure 3.33: Comparison plot between the low-pressure exchange free energy at 298 K *(left) and* 900 K *(right) calculated by the GrAED algorithm and the ambient-pressure exchange free energy at* 298 K *calculated by Widom insertion.*

The utilization of high-temperature averaging enables improved modeling of selectivity in the selective materials that experience a loss of selectivity between low and ambient pressure. These descriptors can quantitatively predict the selectivity at ambient pressure, as demonstrated in the subsequent chapter. Furthermore, these descriptors can provide a qualitative description of the structures that present challenges. It is expected that by leveraging the values obtained through high-temperature averaging, the identification of these problematic structures can be achieved.

Statistical characterization of the energy distributions

To quantify the change of selectivity more accurately, it could be interesting to provide statistical information on the distribution of interaction energies for xenon and krypton, calculated using the grid algorithm. By conducting a statistical analysis, the complexity of the pore adsorption process at higher pressure can be explored through the diversity and distribution of energy values (the quantity of the higher energies in comparison to the lower ones, for example). The grid sampling method presented here utilizes all energy values from the sampled points to construct a histogram representing the energy distribution that can be studied to extract meaningful statistical insights.

These statistical measures encompass moments of different orders (up to 4) of the energy distribution, which provide information on the adsorbate–adsorbent interaction energies in the nanopores at higher loading. The shape of the energy distribution enables quantitative assessment of the changes in selectivity. This approach can be considered a means of summarizing the entire energy distribution in a few statistical values, which is a conventional method employed in the field of data science for analyzing distribution data. Two methods are explored in this context: uniform weighting and Boltzmann weighting of the energy distribution. It is

worth noting that this subsection does not delve into the average of the Boltzmann-weighted distribution, which typically represents the adsorption enthalpy.

Boltzmann weighted distribution

The Boltzmann weighted distribution consists in assigning a weight $\exp(-\beta E)$ to each sampled point according to the corresponding energy E calculated by the GrAED algorithm. This weighting scheme puts a significantly higher weight on the most negative energy values (corresponding to the most favorable adsorption sites) compared to other points. The unfavorable adsorption sites can be considered negligible due to the exponential scaling, which diminishes the importance of these points in the Boltzmann weighted distribution. This distribution has previously been employed to compute the adsorption enthalpy (the first-order moment or average) and indirectly the Henry constant (sum of the weights used for normalization of the distribution). However, this section will focus on other statistical quantities derived from the distribution, which are not commonly used in describing the thermodynamics of the system.

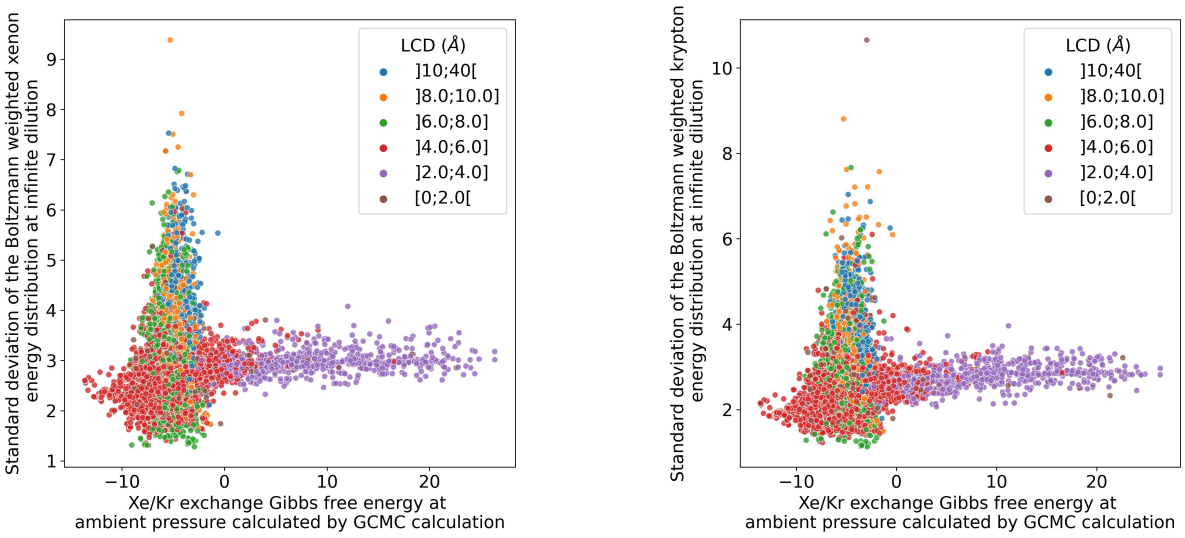


Figure 3.34: Comparison of the ambient-pressure Xe/Kr exchange Gibbs free energy calculated by GCMC simulations of RASPA2 with 100,000 cycles and the standard deviations of the Boltzmann weighted energy distribution of xenon (left) and krypton (right) calculated by the GrAED algorithm at 298 K.

For instance, the standard deviation of the Boltzmann weighted energy distribution is a relevant statistical quantity for evaluating the decline in selectivity. In the previous chapter, the diversity of site attractiveness was identified as a key factor that could explain the drop in selectivity. Therefore, the standard deviation of the energies serves as a useful characterization of the diversity of nature among different adsorption sites. Figure 3.34 presents the calculation of this standard deviation for both xenon and krypton. The values exhibit a higher variation concentrated within a specific range, which aligns with the range of entropy change identified and, more importantly, corresponds to the range where the selectivity drop is observed in Figure 3.33 (between −10 and 0 kJ mol^{-1}). Qualitatively, the standard deviation provides insights into the diversity of pores, which can aid in characterizing the underlying causes of selectivity drop. A higher diversity generally implies a greater probability of experiencing a selectivity

drop. However, quantifying the probability of selectivity drop poses a challenge that cannot be adequately addressed by a simple theoretical model alone. Therefore, the next chapter will focus on exploring machine learning models as a means to overcome this limitation.

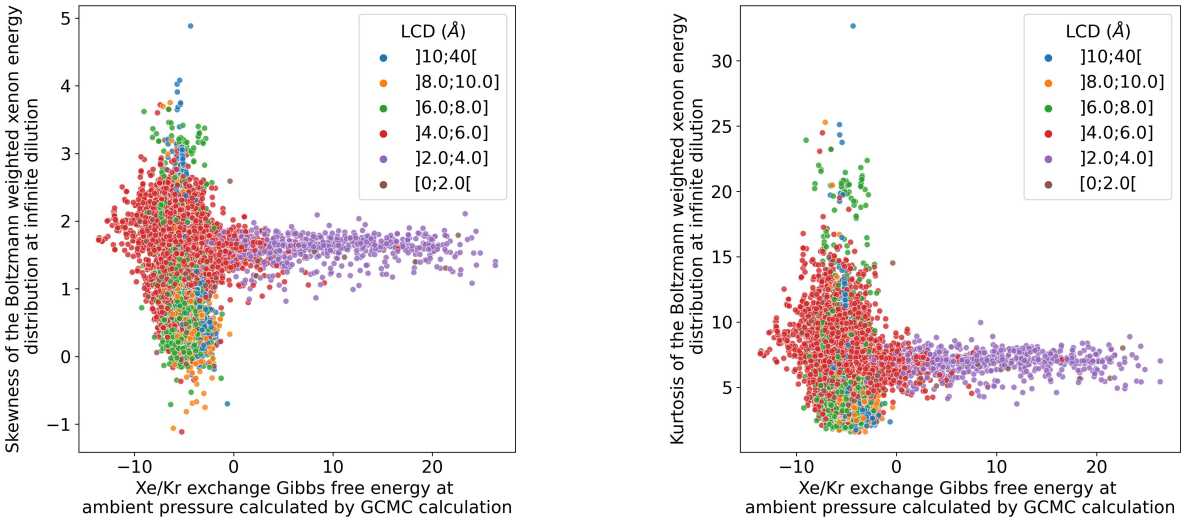


Figure 3.35: Comparison of the ambient-pressure Xe/Kr exchange Gibbs free energy calculated by GCMC simulations of RASPA2 with 100,000 cycles and the skewness (left) and the kurtosis (right) of the Boltzmann weighted energy distribution of xenon calculated by the GrAED algorithm at 298 K.

Two additional statistical quantities, namely skewness and kurtosis, have been introduced to describe the distribution of energy values. Skewness measures the asymmetry of the distribution, while kurtosis quantifies the "tailedness" (the number of values in the tail of the distribution). Skewness is the standardized third moment, and kurtosis is the standardized fourth moment of the distribution. For a random variable X with a mean value of m and a standard deviation σ, the n-th order moment $\mathrm{M}_n(\mathrm{X})$ is defined as:

$$\mathrm{M}_n(\mathrm{X}) = \mathbb{E}\left[\left(\frac{\mathrm{X} - m}{\sigma}\right)^n\right] \tag{3.9}$$

These statistical quantities provide additional insights on the distribution. For instance, if the distribution is skewed towards the most negative pore energies, it indicates a preference for adsorption at higher pressure. Conversely, the opposite skewness would explain a larger drop in selectivity. The overall shape of the distribution requires more than just the standard deviation and mean value to capture the reasons behind the selectivity drop. While having the complete information on the distribution would be ideal, visually comparing structures based on this multidimensional descriptor would be overly complex. The statistical quantities effectively compress the complex energy distribution data.

Figure 3.35 further illustrate the range of selectivity values discussed throughout this section. The different statistical quantities introduced in this discussion can be used to sort materials within this range. The skewness and kurtosis values can potentially establish a theoretical link with the previously identified selectivity drop. However, without a model or framework, it is impossible to find the accurate relationship solely by visual observation.

Uniformly weighted distribution

To conclude this overview of the thermodynamic/energetic descriptors derived from the newly developed grid sampling, a more uniformly weighted energy distribution will be examined. The significantly higher energy values corresponding to the overlap with a framework atom are naturally excluded by the sampling process. In this case, a threshold value of 100 kJ mol^{-1} has been used for the grid sampling, defining a very large overlap range. Energy points below this threshold are considered in the distribution, representing the adsorbable sites (no overlap) that are weighted based on their occupancy of the void volume.

The mean value and standard deviation of this distribution have been analyzed. The mean value shows a weak correlation with the exchange Gibbs energy at ambient pressure. However, the correlation disappears for materials with larger pores, where a more diverse range of energy values is present. The lower mean value in these cases can be attributed to the larger void fraction, resulting in increased weight on the more negative values, but this does not indicate the presence of highly attractive sites, as no Boltzmann weight is applied — the exchange free energy does not follow the same trend and returns to more positive values. It is worth mentioning that the values are generally very high due to the use of an energy threshold value of 100 kJ mol^{-1}, resulting in many points falling within zero to this threshold range, thereby shifting the mean towards these values. Consequently, the statistical analysis of this distribution was not extended beyond the standard deviation. A more refined distribution design should be considered to focus on the negative values, such as lowering the threshold or using a less skewed Boltzmann averaging method (e.g., averaging at higher temperature).

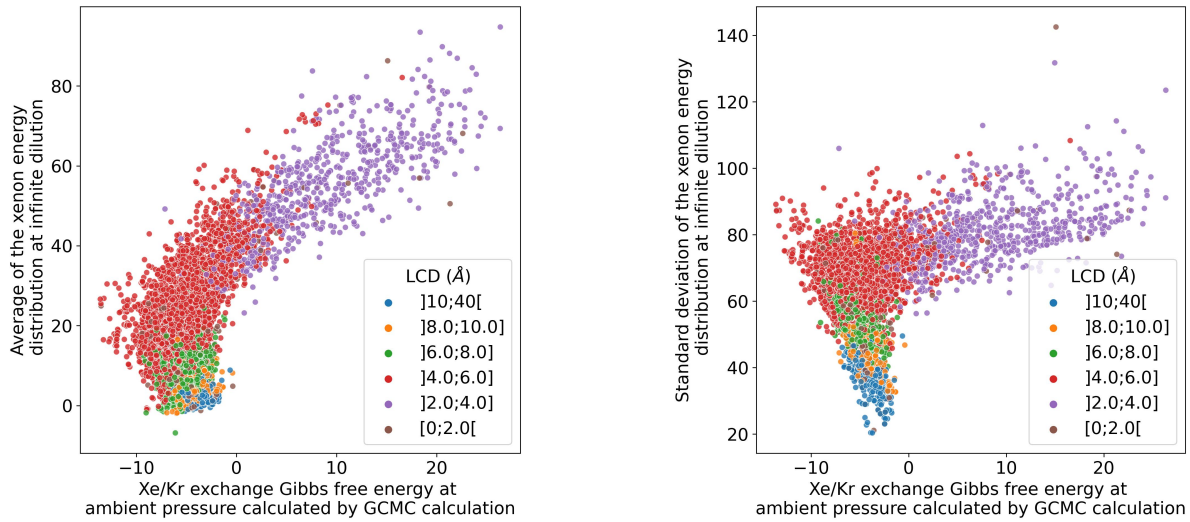


Figure 3.36: Comparison of the ambient-pressure Xe/Kr exchange Gibbs free energy calculated by GCMC simulations of RASPA2 with 100,000 cycles and the mean values (left) and the standard deviation (right) of the uniformly weighted energy distribution of xenon calculated by the GrAED algorithm at 298 K.

The standard deviation of this distribution has a lower value for materials with larger pores, as depicted in Figure 3.36. This observation can be attributed to a higher concentration of points near the average value of the energy (around 0 and 10 kJ mol^{-1}), as depicted by the left plot in the figure. For materials with a standard deviation between 20 and 40 kJ mol^{-1}, it seems clear that their exchange free energy does not exceed −8 kJ mol^{-1}, indicating that they are not among the top selective materials. Therefore, the presented standard deviation aids in

identifying materials that could be promising at low pressure but do not exhibit promising behavior in practice.

To improve this approach, the utilization of a higher temperature Boltzmann average for the distribution weights was explored, without necessarily employing a temperature of 900 K. Additionally, a uniformly weighted average was tested on an energy distribution using different energy thresholds (zero or the mean kinetic energy of a gas $1.5k_BT$). The underlying concept behind these explorations is to characterize a higher energy state close to the state at ambient pressure. Furthermore, higher-order moments can be tested to provide a more picture of the distribution.

This GrAED algorithm proves to be particularly efficient in sampling energies for structures characterized by high symmetry and a large occupied volume. The highly accurate description provided by the grid enables the calculation of additional descriptive metrics, which can be valuable not only for describing adsorption at infinite dilution but also for investigating correlations with selectivity values at higher pressures. Finally, detailed computations on the GrAED sampling technique can be accessed online at github.com/coudertlab/GrAED.

3.4 From statistical description to prediction

In this chapter, a comprehensive overview of fast sampling techniques for evaluating adsorption performance at infinite dilution was provided, excluding the standard Widom insertion method discussed in the previous chapter. The effectiveness of these quantities in describing the thermodynamics of adsorption in nanoporous materials was demonstrated by comparing them to conventional methods. The next step is to examine their predictive value in determining Xe/Kr selectivity under physical conditions closer to industrial settings.

For instance, in the previous section, each thermodynamic descriptor derived from grid sampling was individually examined and compared to ambient-pressure conditions. However, this approach has limitations, as it only offers a descriptive understanding and cannot provide predictions. Moreover, the limited dimensions in visualization restrict the breadth of correlations. While certain key features of the most selective materials (pores adapted to the kinetic diameter, pore shape that maximizes the interactions, etc.) were identified, the correlation-based approach also revealed inherent weaknesses. Understanding high-pressure selectivity using quantities solely based on a host–guest interaction energy grid proves challenging. To overcome these challenges, modern approaches employ statistical learning to capture this relationship using a sufficiently large set of structures with some computed properties. In the upcoming chapter, the potential of these novel descriptors for predicting selectivity values beyond the infinitely diluted case will be explored, all while significantly reducing computational costs.

4

STATISTICAL LEARNING OF ADSORPTION PROPERTIES

4.1 Machine learning models

In the field of nanoporous material study, machine learning (ML) models have been widely used to characterize various properties such as adsorption, transport, catalytic or mechanical properties. These models offer a means to replace time-consuming simulations with simpler calculations of key descriptors, thereby aiding in the prediction of desired properties. In other cases, they are used to describe the structure-property relationships learned by the ML model. However, it should be noted that machine learning cannot be considered as a silver bullet since its application requires a comprehensive understanding of the key variables that improves prediction accuracy. In this study, a machine learning model will be built to characterize the separation of xenon from krypton at ambient pressure, utilizing the work on thermodynamic descriptors and knowledge on the effect of pressure on selectivity from the previous chapters.

4.1.1 From algorithm to machine learning

To understand the learning process of machines, it is necessary to understand how computers perform tasks. The human operator plays a key role in this process by designing the solution based on theoretical considerations and creating a list of instructions, known as an algorithm, which outlines the required actions for the computer to achieve the desired outcome under specific circumstances. In the context of physical or chemical science, these algorithms typically articulate the different components of a theoretical model, such as solving equations without analytical solutions or expressions, and addressing probabilistic problems. The previous chapters presented such algorithms used for simulating adsorption processes. For instance, GCMC simulations are based on the statistical physics of phase equilibrium between a gas phase and an adsorption phase within a nanoporous material, and Monte Carlo models are used to replicate the statistics associated with the grand canonical ensemble. Energy sampling algorithms, along with the Widom insertion are additional examples illustrating how computers assist theoreticians in modeling systems under specific chemical and physical conditions.

Machine learning models are also based on algorithms, but their objective differs significantly from that of the above-mentioned examples — they don't aim at providing comprehensive computational details based on established theoretical principles. As their name suggests, machine learning algorithms aim to learn underlying relationships within input data, enabling them to perform tasks autonomously. The machine learning (ML) algorithm serves as a set of instructions guiding the machine learning process. For instance, clustering algorithms can distinguish different classes of elements within a disordered dataset, leading to the emergence of new concepts. This type of machine learning algorithm is referred to as unsupervised learning, as the data is not pre-labeled, and the machine assists in uncovering the underlying structure. As unsupervised learning extends beyond the scope of this thesis, further details on this algorithm type will not be provided. It is worth mentioning that supervised learning models are the focus of study, which learns the relationship between labeled data and the characteristics (features or descriptors) of a given dataset. Subsequently, these models can predict the label of unlabeled data based on their characteristics.

The focus of this thesis lies on the supervised learning model, which learns the relationship between labels and their characteristics (referred to as features or descriptors) from a given set of labeled data points, enabling the prediction of labels for unlabeled data based on their characteristics. As an example, predicting tomorrow's weather could involve using past weather data from similar dates to infer whether it will rain. The ML model's features comprise the weather history, while the target variable or label of the data corresponds to the future weather.

The distinctions between a standard algorithm and an ML algorithm can be illustrated by a fascinating board game called Go. This game is traditionally played by two players on a 19×19 board, where each player places black/white pieces to gain control over the maximum number of boxes. Based on these simple rules, different algorithms have been developed to make computers 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 released. These programs were based on simple alpha-beta search algorithms that sought to test every possible move (while pruning the less promising

ones). While they worked well in other games like chess (IBM's Deep(er) Blue beat the world chess champion in 1995), these types of programs in Go were only at the level of a novice player. The difference in performance lies in the combinatorics of both games. Chess has a number of legal positions lower than 10^{47},**website_labelle** while Go has approximately 10^{171} legal positions.**Tromp_2007**, **github_tromp_go** The state space to explore in Go is incomparably greater, and an increase in computing power that improved the performance of chess-playing computers would not make a significant difference for Go. A drastic reduction in the space to be explored is required for a computer program to work. The biggest improvement came in 2007 when a Monte Carlo tree search was introduced by Couloms.**Coulom_2007** This algorithm uses heuristics to distinguish between good and bad moves based on human perception of the game. A probability of selection is assigned to the moves according to their potential (policy), and potential moves are randomly selected based on this probability. The average outcomes associated with a parent move provide the value of the move. The computer Go is now more efficient in evaluating moves using a Monte Carlo sampling, and it can now play with average amateur players although it is nowhere near surpassing them. Up until now, the algorithms have been based on human knowledge that the programmer implements directly in the computer using machine instructions. Statistics and randomness are used to guide the machine towards the best moves and reduce their predictability, but the statistics that identify the moves are based on human heuristics that are usually not generalizable. The revolutionary aspect brought about by machine learning in the field aims to improve the evaluation of these statistics using data from previously played games. By using a dataset of 30 million moves, Alpha Go is based on the same Monte Carlo tree search framework, but the formulas behind the probability of searching a move are replaced by a machine learning model called the "policy network" and the evaluation of confidence in winning a position is done by a value "network".**Silver_2016** Alpha Go became the first computer program to beat a world champion in 2016. One year later, an improved version called Alpha Go Zero generated its own data by playing games against itself to train a similar machine learning structure as the one presented before. This new version beat the former version 100 times out of 100,**Silver_2017** marking a new era of computer dominance in Go over the world's best players, with the defeat of another top player further confirming the advent of this new era.

This example showed how the value of each move was learned by the machine through the compilation of knowledge from large 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 objective is not to instruct the computer on how to play using player knowledge implemented in formulas and explicit instructions. Instead, an explicit framework with flexible parameters is provided to the model, which needs to learn these parameters using a database. In other words, the model's parameters are adjusted to match the values of a database while having the ability to generalize to situations outside of the database (further discussion on the notion of generalizability will be presented in the following sections). The purpose of this section is not to provide a comprehensive overview of all existing models, but rather to introduce the main concepts of ML through the example of the model used in this thesis for the prediction of selectivity performance.

4.1.2 Introduction to supervised learning

In this thesis, the focus will be on the most common way to statistically learn from data, which is known as supervised learning. As previously introduced, supervised learning entails 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 referred to as the model or the predictor and is expected to generalize to similar but unseen data. In this section, the goal of the learning algorithm will be formalized when provided with a set of labeled data, to introduce more complex notions in machine learning, such as the bias–variance trade-off, as well as more specific models used in this chapter like the tree-based models. Various books have been consulted to develop this section, primarily the Elements of Statistical 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 denoted as $\mathcal{D}_n = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ 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 represents the label of the data point i that belongs to a set $\mathcal{Y}$ (numerical, categorical or vectorial). The observed characteristics can be modeled by a random variable X, while the label is represented by another random variable Y. The dataset provides only a partial view of the joint probability (see equation 4.1), and the objective is to generalize the relationship to unseen data. (X, Y) represents all possible combinations of seen and unseen data points.

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

The challenge of supervised learning is that a complete picture of the probability law is not provided by the available data. The objective is to determine the most probable label y for a given data point characterized by $\mathbf{x}$, which involves determining the conditional expectation $\mathbb{E}[\mathrm{Y}|\mathrm{X} = \mathbf{x}]$ of Y given the observable $\mathbf{x}$. This determination relies on the conditional probabilities $\mathbb{P}(\mathrm{Y} = y_i|\mathrm{X} = \mathbf{x}_j)$ observed across all data points $i, j \in \{1, \ldots, n\}$.

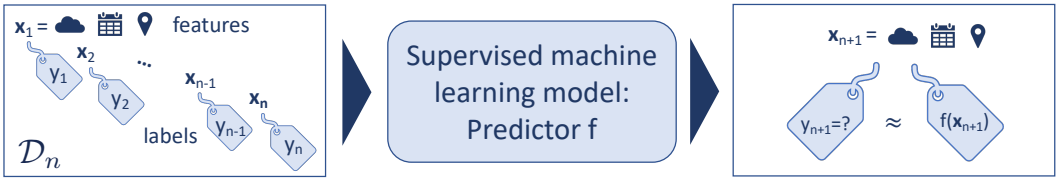


Figure 4.1: Illustration of the core principle of supervised learning. A data point of $\mathbb{D}_n$ corresponds to a set of features $_i$ labeled by y_i. The supervised machine learning model trains a predictor f on the dataset $\mathbb{D}_n$ to predict unknown data y_{n+1} using the features $_{n+1}$ so that $f(y_{n+1}) \simeq y_{n+1}$ (approximate prediction).

To achieve this, the learning algorithm uses a "predictor" f, which can be defined as the function that associates values (features) from $\mathcal{X} = \mathbb{R}^p$ with values from $\mathcal{Y}$. By changing the learning model (subsection 4.1.3) or the feature space $\mathcal{X}$, different domains $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ where the prediction function f is sought, can be defined. The domain $\mathcal{F}$ can be either too restrictive, resulting in the found optimal function being far from the theoretical one, or too large, making

the optimization problem nearly impossible to solve or leading to a solution that is too close to the data. These issues raise questions regarding fitting, which will be discussed later.

This predictor can be interpreted as the function that provides the most probable outcome y for a given input $\mathbf{x}$. To assess the quality of the predictor, a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^p$ is introduced to compare the predicted value $f(\mathbf{x})$ with the true value y on available dataset $\mathcal{D}_n$. The loss function should increase when $f(\mathbf{x})$ deviates from y. To extend the definition of the loss to the entire possible space, the theoretical risk $\mathcal{R}$ of a predictor h is introduced using the random variables X and Y, such that $\mathcal{R}(h) = \mathbb{E}[\mathcal{L}(h(\mathrm{X}), \mathrm{Y})]$. However, since the exact mapping of the random variables is unknown, the empirical risk $\mathcal{R}_n$ on the known dataset $\mathcal{D}_n$ is evaluated instead:

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

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

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

The risk function can utilize various loss functions, with an increasing emphasis on large errors depending on their definitions. For instance, a quadratic cost function highly penalizes outliers, thus prioritizing a few medium errors over a single large error. Conversely, an absolute cost function does not exhibit this behavior. Since regression models were exclusively utilized in this thesis work, the details of classification loss functions will not be discussed extensively. Instead, the focus will be on regression loss functions. The quadratic loss or squared error loss $\mathcal{L}_{\mathrm{SE}}(f(\mathbf{x}), y) = 0.5(y - f(\mathbf{x}))^2$ 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 included to simplify the derivatives. This loss is similar to the mean squared error (MSE) used to compare two quantities across a dataset, where the risk function corresponds to half of the MSE on the predictions $\mathcal{D}_n$:

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

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

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

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

Lastly, a Huber loss can be used to combine the less outlier-sensitive absolute loss with the smoothness of the quadratic loss near the minimal error domain. For a given δ, the Huber loss is defined as:

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

A risk function $\mathcal{R}_\delta$ can also be determined using this loss function. The Huber loss is considered a robust loss function since it is less sensitive to the outliers (high values of error) and has a very smooth gradient near low error values like the squared error. It can be viewed as a combination of the advantages of both the absolute and squared errors as illustrated on Figure 4.2.

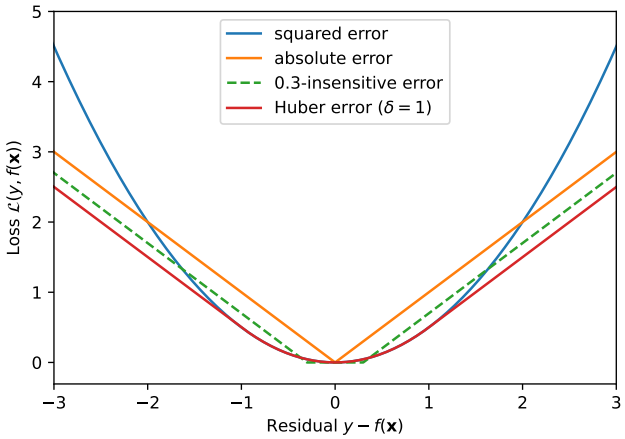


Figure 4.2: Comparison of different loss functions (quadratic loss, absolute loss, ϵ-insensitive and the Huber loss).

Through these theoretical considerations, the process of machine learning from data can hopefully be demystified by formulating this learning process as the optimization of a cost function, which is a common tool in any scientific field. However, this optimization problem poses challenges in the sense that the variable is a function that exists in a high-dimensional space, necessitating approximations to reduce the space. This is why most engineering breakthroughs occur in the conception of the architecture of the ML model, which defines the form of the prediction function f. Another difficulty in machine learning is dealing with an ill-posed problem, where one of the three conditions of Hadamard is not satisfied. These conditions pertain to the existence and unicity of a solution and its continuity with respect to the initial conditions. Typically, this issue is addressed through regularization techniques, such as the one introduced by Tikhonov in the second half of the 20_{th} century. Furthermore, the minimization of the empirical risk does not always align with the minimization of the more global risk (considering all possible observations). In other words, minimizing $\mathcal{R}_n$ does not always yield the same solution as the minimization of $\mathcal{R}$. Therefore, the complexity of the risk optimization problem depends on the chosen loss function and the domain $\mathcal{F}$ defined by the model. Different techniques can be used to construct a solution without any guarantees of its optimality. One of the biggest challenges in ML is overcoming 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 there is no guarantee the model will work on other data points as n goes to infinite. The generalizability of model consists

of ensuring the predictability of unseen data, where the solution does not only 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, the p explanatory variables considered are sometimes not sufficient to model the target phenomenon. To train a generalizable model, it is necessary to ensure sufficient learning to capture the inner relation between X and Y while avoiding fitting the data too closely and capturing the noise along the way. Otherwise, it is said that the model overfits the data. If the model is highly inaccurate even on the training data, it is said to underfit, generally indicating 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 trade-off 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 sensitivity to small variations in the input values. A high bias error corresponds to underfitting, indicating that not enough is learned from the data. A high variance error corresponds to overfitting, indicating too much is learned, even including superfluous relations. To formalize these errors, reference can be made to the empiric risk function $\mathcal{R}_n(f)$ that models the error of the predictor $f \in \mathcal{F}$. To ascertain whether the ideal optimum has been achieved, a comparison with the minimal risk attainable by a predictor possessing infinite knowledge is necessary. This minimal risk is denoted as $\mathcal{R}^* = \min_{h \in \mathcal{Y}^{\mathcal{X}}} \mathcal{R}(h)$. This excess error $\mathcal{R}_n(f) - \mathcal{R}^*$ can then be decomposed into two errors, which can 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, which measures the deviation between the current predictor f and the minimum risk predictor f_n^* (there can be multiple solutions in the case of an ill-posed problem) 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 $\mathcal{F}$ limited availability of data for the prediction model. In the presence of an infinite amount of data, the model f^* associated with the risk $\mathcal{R}^*$ would not be influenced by the noise, as several data points with similar features but with minor noises would yield similar predictions. The difference in loss between this ideal function f^* and the tested current function f would correspond to an overfitting of the noise, which could not be distinguished in the finite case when the considering the domain $\mathcal{F}$ defined by the suitable model. Conversely, if there is a model issue, this error also measures the approximation error resulting from the selection of specific features and a particular model architecture.

In general, if the model is overly complex compared to the available data, it would result in a close fit to the data and a high risk of overfitting. Conversely, if the model is too simplistic, it would lead to a significant bias error and underfitting. This principle is depicted in Figure 4.3 and provides guidance for designing a new ML model. The complex art of achieving an optimal fit between the model and the dataset involves finding the right balance between bias and

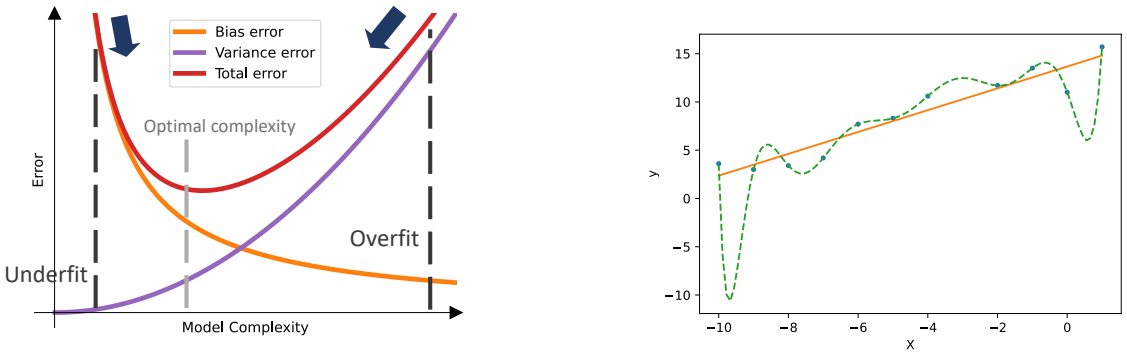


Figure 4.3: On the left, theoretical relation between the bias, variance and total errors and the model complexity. The overfit case is illustrated on the right plot when considering polynomial fits. The lower degree linear function is more generalizable than the biased high degree polynomial that fits perfectly the data.

variance. Fortunately, there are optimization techniques available that can help reduce the variance error by modifying the loss function itself. These tools will be discussed in the following section.

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 solution. This criterion of simplicity is crucial for achieving generalization of the problem. A higher degree polynomial is typically unnecessary when a linear function is a more suitable solution, as demonstrated in Figure 4.3.

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

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

The influence of regularization on the optimization problem varies depending on the expression of the regularization term, denoted as $\Omega_n(f)$.

Since the regularization is a model-specific function (depends on f), the definition of a model is necessary to explore more specific regularization expressions. Consideration is given to a multilinear model $f(\mathbf{x}) = \boldsymbol{\beta}\mathbf{x}^{\mathrm{T}}$, 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 the case of standard multilinear regression with a quadratic loss, the risk function to be minimized can be expressed as follows:

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

Here, y_i represents a scalar quantity in a regression problem ($\mathcal{Y} = \mathbb{R}$). One of the earliest regularization tools introduced by Tikhonov to address ill-posed optimization problem is the L2 regularization. When used in linear regression, this novel model type, known as ridge

regression, consists simply in adding a L2-norm penalty to the model weights within the risk function, as expressed by 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 that controls the importance of the regularization term in the optimization process. By adjusting this parameter, the complexity of the model can be regulated, aiming to find an optimal balance between accuracy and generalizability, as depicted in Figure 4.3. In the case of considering a polynomial function, where the vector $\mathbf{x}_i$ represents various exponentiations of a scalar x_i, such that $\mathbf{x}_i = \left(x_i^0, \ldots, x_i^{n-1}\right)$, the coefficients $\boldsymbol{\beta}$ correspond to the polynomial coefficients of the function f. This serves as a clear illustration of how the regularization terms penalize the complexity of the model by directly penalizing the number of terms utilized and influence impact on the fitting process. It is worth noting that this regularization technique can be adapted to other types of models, provided that a suitable L2-norm of the prediction function f.

Another commonly used regularization term is based on the L1-norm of the prediction function. L1-regularized least square linear regression, known as LASSO (Least Absolute Shrinkage and Selection Operator) regression, allows for a sparser selection of the model weights by permitting certain weights to be zero in the model, unlike L2-regularization. The risk function associated with this regression model can be expressed as follows:

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

where λ_1 is the L1-regularization parameter that controls its importance. The L1-norm can be defined in various ways depending on the model, but its fundamental concept revolves around being a function of the absolute values of the model weights.

Lastly, when both L1 and L2-regularization are combined, linear regression transforms into elastic net regression, and the risk function can be expressed as follows:

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

where $\alpha \in [0,\ 1]$ defines the relative weight of L1 and L2 regularization terms, and $\lambda_{1,2}$ governs the importance of the combined regularization term. This regularization technique consists in simply combining both L1 and L2 regularization, and the different regularization parameters can be adjusted to find the optimal bias–variance trade-off for the final model. These parameters are also commonly referred to as hyperparameters in machine learning, as they influence the parameters at a higher level in the model.

Finally, implicit regularization encompasses alternative forms of controlling the complexity of the model. For instance, it can involve implementing early stopping during the learning process to avoid complete convergence to minimal error with the data. It can also include the removal of outliers that prevent the model from learning properly on relevant data. Furthermore, the architecture of the model itself can contribute to implicit regularization. For instance, random forests are based on an ensemble approach that aims at reducing overfitting, which will be

discussed in the next section. The learning rate in gradient boosting is another regularization parameter that smooths the learning process and will be addressed in the dedicated section. Implicit regularization is related to the construction of the model and will therefore be elucidated in greater detail in the section on machine learning models.

Learning strategies

The theory behind the bias–variance trade-off has been previously introduced, emphasizing the generalization of a model that has a partial glimpse of the available data. However, in practical applications, it is necessary to evaluate the generalization error $\mathcal{R}_n(f) - \mathcal{R}^*$. This evaluation is achieved through a common strategy that consists in randomly splitting the available data into two sets: a training set $\mathcal{D}^{\text{train}} = \left\{(\mathbf{x}_{i_1}, y_{i_1}), \ldots, (\mathbf{x}_{i_N}, y_{i_N})\right\}$ and a test set $\mathcal{D}^{\text{test}} = \left\{(\mathbf{x}_{j_1}, y_{j_1}), \ldots, (\mathbf{x}_{j_{n-N}}, y_{j_{n-N}})\right\}$, such that $\mathcal{D}_n = \mathcal{D}^{\text{train}} \cap \mathcal{D}^{\text{test}}$. The training set is used to solve the optimization problem as defined in equations 4.2 and 4.3, while the test set is employed to assess the generalization error, as it comprises unseen data for the model. In practical applications, a ratio of test data $n - N/n$ (e.g., 20%) that defines the size of the test set from an initial dataset is chosen. The randomness of the split ensures that the data from both sets are similar, yet not identical. However, in some cases, it is important to acknowledge that outliers may be present in the test set in certain cases, thereby resulting in poorer performance than expected. In other cases where the dataset is small and individual data points exhibit significant dissimilarities, the test set may differ significantly from the training set, rendering it impossible for the model to make predictions based on the piecemeal information provided by the training set. Therefore, the percentage of the train/test split should be thoughtfully chosen to maintain representativeness of the training set.

The main property of the test set is that it is composed of an entirely unseen dataset, which means the training of the model should be independent of this set, except for the final evaluation. However, in some cases, different models need to be compared or a "hyperparameter" such as the regularization parameter within the same model architecture needs to be altered. To evaluate these models, the generalization error on the test set cannot be computed for each model, as it would compromise the independence of the test set from the training process. Therefore, validation sets are introduced within the initial training set. A simple training/validation split similar to the train/test split could be performed. Nevertheless, this approach would further weaken the model due to the reduced amount of available data. Moreover, it does not fully utilize the potential of the training set. One widely used technique to test the performance of a model on a training set is cross-validation. This method aims to test the model in multiple configurations by employing different training/validation splits, thereby providing a more comprehensive evaluation of the model's performance through averaging different performances.

The most commonly used method is k-fold cross-validation, which consists in partitioning the training set $\mathcal{D}^{\text{train}}$ into k equal-sized 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 the validation set for all $m \in \{1, \ldots, k\}$. The principle of the k-fold approach is illustrated in Figure 4.4. The approximate generalization error of the model is then computed as the average of the losses calculated on the validation subsets. This tool provides a method for comparing different models without using the test set, which is extremely useful, especially in the parameterization of the ML model.

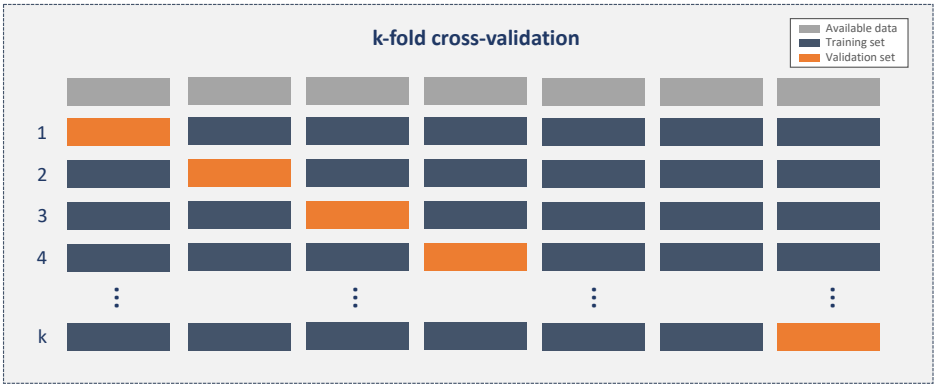


Figure 4.4: Illustration of a k-fold cross-validation. At each step, the machine learning model learns from the training set and is tested on the validation set. The average performance on all validation sets gives an approximation of the generalization error.

Other cross-validation techniques exist and are used in specific cases. For instance, stratification cross-validation ensures the same distribution of labels y_i in each subset, which is particularly useful for classification problems. Increasing the value of k in k-fold validation can make the validation process even more exhaustive. However, this approach requires training the models k times, resulting in an increased computation time. When k reaches the maximum value equal to the size of the training set, the method is referred to as leave-one-out cross-validation. In the case of time series data, the cross-validation technique typically involves sorting the data based on the time history, ensuring that the training set always precedes the validation set. This introduces a completely novel approach to cross-validation. The fundamental concept behind cross-validation is to find multiple training/validation splits to evaluate the model from various points of view. Different strategies exist depending on the specific training problem at hand.

4.1.3 Machine learning models

In this chapter, the transition will be made from the basic components of the model (decision tree) to the more complex ensemble model (e.g., random forest), ultimately concluding with the final stochastic gradient boosting model used in this work. The focus of the discussion will primarily revolve around regression problems rather than classification problems, as the objective is to predict a continuous variable (the xenon/krypton selectivity).

Regression tree

Tree-based models are commonly used in classification problems where the tree classifies the data points into different predefined categories based on a set of binary questions "yes" or "no". These questions are essentially associated with threshold values of the p features or characteristics $C_1, \ldots, C_p$. For instance, a tree node might ask, "Is C_1 higher than 3?", which splits the space into two categories: "yes" and "no". A decision tree can, therefore, be perceived as a splitting of the space into rectangles (in 2D) or their higher-dimensional equivalents in a p-dimensional feature space. To adapt this type of model to regression problems, the label values y can be grouped together into categories represented by the average label value. To summarize, in the context of regression, a decision tree splits the feature space into a set of pseudo-rectangles (volumes separated by finite hyper-surfaces) defined by the tree nodes.

Within each of these subspaces, the average of the different points present in that subspace is assigned. It is worth clarifying that a splitting node corresponds to a boundary between regions, while a terminal node or leaf corresponds to the region itself.

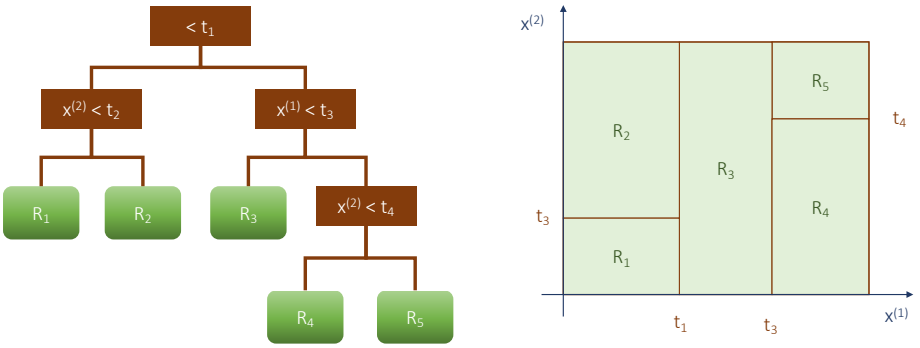


Figure 4.5: Illustration of the decision tree and the region splitting performed by a CART[Breiman_2017] *algorithm. Adapted from an illustration of the book "Elements of Statistical Learning" [***Hastie_2009***].*

The CART[Breiman_2017] algorithm, developed by Breiman et al., is commonly presented as the archetype of a decision tree model. The algorithm follows a straightforward three-step process: (i) Examine every possible split 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] While it is possible to split the decision tree indefinitely, assigning each data point to its own region, this would inevitably lead to a textbook case of overfitting, rendering the model incapable of accurately predicting new data points. To prevent this, the decision tree incorporates a regularization parameter known as the minimum samples split, denoted as $n_{\min}$, which restricts further splitting if a node contains fewer samples than $n_{\min}$, hence treating the latter as terminal nodes. As decision trees are very prone to overfitting, an additional useful regularization parameter, maximum depth of the tree, can be used to halt the iterative tree growth. Finally, a process known as tree pruning can be employed to further regularize decision. Tree pruning simplifies the tree structure and outputs of the final model. A comprehensive discussion of tree pruning is beyond the scope of this work (see Ref. [**Hastie_2009**] for further details). The final tree f_{tree} can be expressed as a function of the different regions $\mathrm{R}_1, \ldots, \mathrm{R}_\mathrm{M}$ created through 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 represents 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 equivalent to the average of the labeling values in the dataset $\mathcal{D}_n$ in the region R_m, i.e. $c_m = \underset{i \in \mathcal{D}_n}{\text{ave}}\,(y_i | \mathbf{x}_i \in \mathrm{R}_m)$. In simple terms, the tree function returns the average value of y (from the dataset) in the region where x (could be new data) is located.

The main advantage of the decision tree lies in its interpretability, as defined in the book by C. Molnar [**molnar2020interpretable**]. This interpretability is derived from the binary

decision made at the root of the decision tree — the explanatory characteristics (R_m) of a predicted value, are easily discernible, and different predictions can be imagined based on the value of $\mathbf{x}$. For smaller trees, the model can even be mentally executed. However, the decision tree model has a reputation for being highly inefficient in identifying simple linear relationships, resulting in a step-like function. It exhibits limited smoothness, with minor changes in the input x potentially causing significant impacts on the predicted value (typically near the boundaries between two regions). Some changes (noise) in the training data can also profoundly alter the tree's structure. The instability of a single tree poses challenges in generalizing over unseen data.**molnar2020interpretable** To address the limitations of a single decision tree, Breiman introduced bagging predictors in 1996. This approach aims to improve the accuracy of models that exhibit instability when subjected to minor changes in the learning set.**Breiman_1996** The concept of bagging predictors has laid the foundations of random forests, which will be discussed in greater detail in the subsequent subsection.

Random Forest

The random forest is built upon the notion that a collection of weak learners, known as an ensemble model, surpasses a single strong learner. This assumption relies on a proven theorem stating 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 a model depends both on the amount of information fed into it and its level of complexity. To achieve a diverse forest consisting of weaker decision trees, two concepts are introduced: bootstrap aggregating (bagging) and random column subsampling. Both methods ensure diversity in the generated trees through random selections and promote relative weakness of the trees by reducing the amount of information accessible to each tree.

The bagging method consists in generating a set $\left\{\phi_b\right\}_{b\in\{1,\dots,\mathrm{B}\}}$ of B weaker learners from different bootstrap datasets $\left\{\mathcal{D}_b^{\text{train}}\right\}_{b\in\{1,\dots,\mathrm{B}\}}$. 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 — It is noteworthy that each bootstrap sample has the same number of elements as $\mathcal{D}^{\text{train}}$, but data points may appear multiple times within it. The frequency with which a data point $(\mathbf{x}_i, y_i)$ appears represents its weight in the bootstrap learning set. In simple terms, each tree model ϕ_b is trained on the $\mathcal{D}_b^{\text{train}}$ dataset, which assigns random weights to the data points. Therefore, each model focuses on different parts of the training data. The generalization error of the model can be evaluated, as certain trees have never seen some data points. The generalization error on unseen data for each tree (similar to cross-validation), known as the out-of-bag error, can thus be assessed.

The second technique consists in randomly choosing a subsample of features for determining the best split (second part of the CART tree growing algorithm). This technique draws inspiration from the work of Ho in 1998, where each tree of a forest is trained exclusively on a randomly chosen subspace of feature.**Tin_Kam_Ho_1998** The only difference in the procedure lies in the fact that the feature space changes at each iteration of the tree growing process, rather than between each tree generation. This method also improves the generalizability of the approach by reducing the likelihood of overfitting in each tree, while the overall accuracy is achieved through 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 proceeds by looping over the number of trees B in the forest. For each tree, denoted as b, a bootstrap sample $\mathcal{D}_b^{\text{train}}$ is randomly drawn (with replacement), and this dataset is utilized to grow the tree (training procedure). During training, a modified CART algorithm is applied to expand the tree by recursively splitting each node: (i) instead of testing all features for the best split, only a random selection of m variables is considered among the p features, (ii) the best split point is determined among the m variables, and (iii) the node is split into two until the minimum leaf size, denoted as n_{min}, is reached. The size of the column subsample defines the number of features randomly considered at each split, serving as another implicit regularization parameter associated with the random forest along with the previously identified regularization parameters of the decision tree, such as the minimal leaf size n_{min} or the maximal depth of a tree. Finally, a set of B trees, denoted as $\{\phi_b\}$, is obtained, which can be used to build an ensemble model Φ, such that:

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

It should be noted that each tree has an equal influence on the prediction, and they are trained on different random samples of the initial training data. Random forest is less prone to overfitting due to the implementation of a cross-validation process known as bootstrapping. However, the algorithm itself does not significantly improve the model's accuracy (bias error). Instead, it relies on the idea that trees mutually compensate for their individual weaknesses in the final ensemble model. In the subsequent section, an alternative algorithm will be introduced, which focuses on leveraging prior knowledge from previous trees to enhance the performance of each tree. This novel technique is referred to as boosting.

From boosting to gradient boosting

In the previous approach, the bootstrap dataset consists of a random selection of samples from the training set $\mathcal{D}^{\text{train}}$ and each tree has an equal voting weight in the final ensemble decision. However, in a boosting algorithm,[drucker1997improving] the paradigm shifts. The data samples are (i) selected based on their predictions by the previous trees, focusing on poorly predicted sample points, and (ii) the tree ϕ_b, trained on this weighted dataset $\mathcal{D}_b^{\text{train}} = \left\{ \left(w_i^{(b)}, \mathbf{x}_i, y_i \right) \right\}$, is evaluated using a confidence α_b, which is determined by the error made (the higher the error is, the lower the confidence is). This confidence measure 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, the CART algorithm (described in the previous sections) is utilized, but with a focus on minimizing a weighted risk function rather than 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 associated with 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 $b = 1$, when there is no previous model, the weights are equidistributed across the samples, i.e., $\forall i,\ w_i^{(1)} = 1/\mathrm{N}$. In practical applications,

to simulate the weighting process, a random selection is performed on each sample with a probability $w_i^{(b)}$ to draw an equally sized N training dataset $\mathcal{D}_b^{\text{train}}$ for ϕ_b.

The specific details of the confidence rate α_b have not been elaborated upon intentionally, as various implementations exist. Generally, it is a decreasing function of the total error of the tree on the weighted dataset. The AdaBoost algorithm typically uses the half of the opposite of the logit transform function $\alpha_b = 0.5 \log\left((1 - \mathcal{R}(\phi_b))/\mathcal{R}(\phi_b)\right)$, which approaches $+\infty$ for very small errors and $-\infty$ for very large ones.**Freund_1997**, **schapire2013explaining** Gentle AdaBoost, on the other hand, assigns an equal say to each tree independent of its performance, which, in some cases, yields better generalization performance compared to regular AdaBoost. However, very high values of α_b can lead to overfitting of the model in some cases, as a very good performance on the weighted dataset may indicate a good fit on noisy data points.**schapire1998improved** To prevent overfitting, an early stopping procedure with a cross-validation (typically k-fold) training procedure is performed to determine the optimal number of trees required to remain generalizable while reducing bias error. As is often the case in machine learning, this involves a trade-off between bias and variance.

In its original implementation, AdaBoost uses stamps, which are trees composed of a single splitting node and two leaves. However, boosting algorithms can be applied to trees of any depth. The tree-depth hyperparameter plays a crucial role in tree-based models as it defines the complexity/strength of each learner tree. Smaller trees generally exhibit less overfitting (see the relationship between complexity and variance in Figure 4.3), and the AdaBoost algorithm uses the smallest possible tree to compensate for its highly aggressive learning procedure. The key takeaway from this study is that boosting focuses on the training trees that compensate for the errors of previous trees, and it can manipulate tree-based hyperparameters (e.g., tree depth, number of trees) to control variance error.

In fact, boosting can be reformulated as a gradient descent problem, as demonstrated by Mason et al.**mason1999boosting** AdaBoost can be viewed as a gradient boosting algorithm with an exponential loss function (same loss and derivative) and follows 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, leading to the minimization of an objective function $\mathcal{R}(\Phi_b)$. The weight $w_i^{(b)}$, which measures the prediction error for each sample i, can be expressed as a derivative of a differentiable loss function $\mathcal{L}$ since the minimum is reached when the derivative is zero.

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

where $\hat{y}_i$ is a derivation variable describing the ensemble tree prediction and is evaluated at Φ_{b-1}. Instead of predicting the y_i values, the weight or pseudo-residual $w_i^{(b)}$ can be predicted, which measures the deviation of the previous model Φ_{b-1} from the ideal Φ (zero weights everywhere in an ideal world). This weight is compensated using a tree ϕ_b. In other words, the CART framework is employed to grow a tree ϕ_b that predicts the gradients $w_i^{(b)}$ from the features $\mathbf{x}_i$, iteratively improving 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, as introduced by Friedman in stochastic gradient boosting, to slow down the learning process and mitigate overfitting.[Friedman2002] In a steepest descent step, the values of this learning rate η minimize the risk function $\mathcal{R}(\Phi_{b-1} + \eta\phi_b)$ associated with 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 $\Phi_1(\mathbf{x}) = \arg\min_{c\in\mathbb{R}} \sum_{i=1}^{N} \mathcal{L}(y_i, c)$. 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 $\mathcal{L}_{SE} = \frac{1}{2}(y_i - f(\mathbf{x}_i))^2$ 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:
   $\Phi_1(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} y_i$
2. For $b = 2$ to B:
   (a) Compute the pseudo-residuals, which are equivalent to real residuals in the case of a quadratic loss $\quad \forall i \in \{1, \ldots, N\},\ w_i^{(b)} = y_i - \Phi_{b-1}(\mathbf{x}_i)$
   (b) Train the weak tree ϕ_b on the dataset $\{(\mathbf{x}_i, w_i)\}_{i\in\{1,\ldots,N\}}$.
   (c) Update the model using a fixed learning rate $\eta \in [0, 1]$ instead of finding $\eta = \arg\min_{c\in\mathbb{R}} \sum_{i=1}^{N} \mathcal{L}\left(y_i, \Phi_{b-1}(\mathbf{x}_i) + c\phi_b(\mathbf{x}_i)\right)$ 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, the different ways of utilizing decision trees to perform predictions on a training dataset $\mathcal{D}^{\text{train}}$ have been demonstrated, with a specific focus on two ensemble models: random forest and gradient boosted trees. The aim of exploring these models is to introduce a prediction model that combines techniques from both ensemble models. This model, known as eXtreme Gradient Boost or XGBoost, was introduced by Chen et al. and offers improved scalability compared to similar methodologies. The implementation improvements will not be discussed in detail here (see Ref. [**chen2016xgboost**] for more details). Instead, the focus will be on the fundamental framework employed by XGBoost, which will enable a better understanding of its core components.

XGBoost model parameterization

The XGBoost model essentially constitutes a gradient boosting model, as discussed in the previous section, with several regularization parameters that can be fine-tuned to enhance its generalizability. In a learning problem involving N learning examples and p features/descriptors, the predictor Φ can be expressed as the sum of weaker tree learners ϕ_b:

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

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 $M = 2^{\max_{\text{depth}}}$, and B is the maximum number of estimators in the ensemble model. The number of estimators is typically determined using early stopping in k-fold cross-validation.

A quadratic loss function, with L1 and L2-regularization terms, was applied to the M leaf weights c_m of a model *phi*, resulting in the following expression for the loss function $\mathcal{L}$:

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

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 be expressed as follows:

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

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 4.1.3) to find the best tree that will be used to predict the pseudo-residuals. As explained earlier, the pseudo-residual is defined as 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 for updating the ensemble model is also a key component of the final model that needs to be adjusted to maximize the generalizability of the model. This parameter slows down and smoothens the convergence to the solution, thereby improving the bias–variance trade-off. Small values below 0.1 are typically used.

To add randomness in the gradient descent procedure, three other parameters that are very similar to the ones implemented in a random forest were utilized. With the integration of these techniques, the model can be referred to as stochastic gradient boosting, as described in the Ref. [**Friedman2002**]. In each iteration, a random subsample of the training data is drawn (without replacement) based on a parameter $\mathrm{N}_{\mathrm{sample}}/\mathrm{N}$. This parameter has a similar effect as the bagging procedure of the random forest, restricting the focus of each weak learner to a portion of the learning set. This approach reduces overfitting, akin to a cross-validation procedure. The different trees learn from distinct segments of the training set, preventing the ensemble model from overfitting the entire dataset. This provides an effective solution to the well-known issue of overfitting encountered in standard gradient boosting. Another procedure involves randomly selecting feature columns, inspired by the concept introduced in Ref. [**Tin_Kam_Ho_1998**]. It entails randomly extracting a subsample of the features for the training of each tree. A parameter is required to determine the size of the portion of features p_{tree}/p used for training each tree. Similarly, column sampling can be performed at each level rather than for each tree, where a proportion p_{level}/p is defined accordingly. Alternatively, a random feature sampling at the node level is also a plausible option, although this parameter was not utilized in this context.

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.

Finally, the parameters used in the construction of the final model are compiled in Table 4.1. The table encompasses a tree-specific parameter "max_depth", an ensemble-specific parameter, "n_estimators", general regularization parameters inspired by linear models "alpha" and "lambda", along with a 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 blending of various concepts drawn from diverse domains within the field of data science. By using this machine learning model, the selectivity drop problem discussed in Chapter 3 will be addressed.

4.2 Prediction of the ambient-pressure selectivity

Before delving into the model of this work, the different literature contributions to xenon/krypton separation screenings will be briefly reviewed. 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 heavily relied on the Voronoi energy, which represents an average of the interaction energies of a xenon atom at each Voronoi node.**Rycroft_2009** To rationalize this increase in performance, the Voronoi energy can be considered as a faster proxy for the adsorption enthalpy. A comparison with the standard Widom insertion revealed that, although faster, the Voronoi energy is less accurate. To address this, a more effective alternative called surface sampling (RAESS) was developed, utilizing symmetry and non-accessible volumes blocking (see section 3.2). Recently, Shi et al. used an energy grid to generate energy histograms as a descriptor for their ML model. This approach provides an exhaustive description of the infinitely diluted adsorption energies**Shi_2023** but can be computationally expensive.

While the aforementioned approaches demonstrate good accuracy in predicting low-pressure adsorption (i.e., in the limit of zero loading), they are not suitable for predicting adsorption in the high-pressure regime when the material is near saturation uptake. Grand Canonical Monte Carlo (GCMC) simulations are commonly employed for this task, but there is a lack of methods for decreasing computational costs for high-throughput screening. The challenge this thesis aims to address is predicting selectivity in the nanopores of a material at high pressure, where adsorbates interact with each other, while having access to information only on the interaction at infinite dilution. Comparing the low- and high-pressure cases provides key information on the origin of selectivity differences. Previous studies have shown that selectivity can decrease between the low and ambient pressure cases in Xe/Kr separation applications (see chapters 2 and 3), primarily due to the presence of different pore sizes and potential reorganizations caused by adsorbate–adsorbate interactions.

By combining grid-based descriptors described in Chapter 3 (section 3.3) with statistical characterizations of pore size, a set of ML descriptors suitable for rapid and accurate ambient-pressure selectivity prediction was proposed. These descriptors were used in an optimized XGBoost model, showcasing its performance in the case of xenon/krypton separation in the CoRE MOF 2019 database.**Chung_2019** The study presented in the following is in the process of being published, and a preprint is available on *Chem. rXiv* in Ref. [**Ren_2023_ml**].[1]

4.2.1 Data Preparation

Target variable

This study aims at building an ML model to predict the Xe/Kr ambient-pressure selectivity faster than standard techniques. To obtain reference values (ground truth), I used the RASPA2 software**dubbeldam2016** to run GCMC calculations (introduced in section 2.1.3) of 20–80 Xe/Kr mixtures at 298 K and 1 atm on our cleaned database. The van der Waals interactions are described by a Lennard-Jones (LJ) potential with a cutoff distance of 12 Å. The LJ parameters of the framework atoms are given by the universal forcefield (UFF),**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 towards predicting other mixtures at different physical conditions (*e.g.* Xe/Kr mixtures out of nuclear off-gases).

To achieve this, a logarithmic transform of the selectivity will be considered instead of the raw value because the goal is rather to predict the order of magnitude of the selectivity values than to predict the higher values of selectivity — an ML model that focuses its prediction on raw selectivity values can reach lower errors by simply focusing on the higher values than the lower ones. By focusing on the logarithmic transform of the selectivity, the different selectivity categories can be separated through the different orders of magnitude of the selectivity values. This approach distributes more evenly the efforts on all the whole spectrum of selectivity values. Moreover, this logarithmic transformation is effectively an exchange Gibbs free energy that was introduced in Chapter 2 and redefined in equation 4.23; it can therefore be easily compared with the energy descriptors I introduced in Chapter 3.

$$\Delta_{exc}G = -RT\ln(s) \tag{4.23}$$

[1] The corresponding data and scripts can be found at: https://github.com/eren125/xe_kr_selectivity_xgb

Database and data generation

This methodology is tested on a set of realistic MOFs by considering 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, the database is reduced to a set of 9,748 structures. Then, with the string information given by the Zeo++ software[zeopp_Willems2012] this number is reduced to 9 177 by removing the structures that are not tridimensional, where solvents are still detected (wrongly classified in "all solvent removed"), or where the metal is radioactive or fissile (e.g., Pu-MOF TAGCIP,[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, a condition on the largest cavity diameter (LCD) is added to keep only the structures with pore sizes allowing the adsorption of xenon: this is the case for 8 523 structures with an LCD higher than 4 Å (approximately the size of a xenon molecule). This is equivalent to removing the structures with very unfavorable adsorption enthalpies, that are not promising for our adsorption-based separation.

Then, the descriptors summarized below (and fully detailed in Supporting Information) were calculated on this restrained dataset. At this stage, 146 structures failed to be calculated in GCMC (the CPU used are limited to 3.75 GB, and for some materials the RASPA2 grid exceeds this memory limitation) and 83 have a zero value for the standard deviation of the pore distribution (skewness and kurtosis go to infinite and cannot be defined). A final dataset of 8,300 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

Examining a number of research papers on supervised ML for the prediction of adsorption properties,[Fernandez_2013, Simon_2015, Fanourgakis_2020, Anderson_2020, Pardakhti_2020] recurrent descriptors are identified: (i) geometrical descriptors obtained using software like Zeo++[zeopp_Willems2012] including surface area (SA), void fraction (VF), largest cavity diameter (LCD) and pore limiting diameter (PLD); and (ii) physical and chemical descriptors such as framework density, framework molar mass, percentage of carbon (C%), nitrogen (N%), oxygen (O%), hydrogen, as well as halogen, nonmetals, metalloids and metals, and degree of unsaturation. Although these descriptors are versatile and widely used in ML models, they fail to provide specific information for the ML task of this study. As demonstrated by Simon et al., energy descriptors greatly influence ML models for selectivity prediction.

The geometric analysis of crystalline porous materials is typically based on the predefined van der Waals (vdW) radii from the Cambridge Crystallographic Data Centre (CCDC). This forcefield-independent choice can create a gap between geometrical descriptors and thermodynamic values obtained through molecular simulations. Inspired by a recent work comparing PLDs and self-diffusion coefficients,[Hung_2021] a list of vdW radii was defined to be read by

the Zeo++ software (more details can be found at github.com/eren125/zeopp_radtable). In this study, all Zeo++ calculations utilize an atomic radius that corresponds to the distance at which the LJ potential reaches $3k_BT/2$ at T = 298 K.

The SA exposed to different probe sizes (1.2 Å, 1.8 Å and 2.0 Å) was tested. The probe occupiable volume was chosen to measure the void fraction (VF) for different adsorbent using probe sizes of 1.8 Å (close to the radius of krypton) and 2.0 Å (close to that of xenon). This definition of pore volume demonstrated better agreement with experimental nitrogen isotherms.[vol_Ongari2017]

Given the objective of predicting the difference between low-pressure selectivity and ambient-pressure selectivity (for a specific gas mixture composition), some descriptors hold little importance, and the key factor lies in the difference in accessible volume and affinity of the remaining pore volume with xenon compared to krypton. The intuition developed in Chapter 2 outlines the role of a diverse distribution of pores with different xenon affinities. Therefore, from the "standard" descriptors mentioned in the literature, the following 7 descriptors were retained: C%, N%, O%, LCD ("D_i_vdw_uff298"), PLD ("D_f_vdw_uff298"), SA for a 1.2 Å probe ("ASA_m2/cm3_1.2") and VF for a 2.0 Å probe ("PO_VF_2.0"). Additionally, a new descriptor ΔVF was created to represent the difference in void fraction values, specifically the difference in volumes occupiable by xenon (2.0 Å) and krypton (1.8 Å). A comprehensive presentation of all these descriptors, including other geometrical descriptors based on pore size distribution, can be found in Table 4.2 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 utilized to generate descriptors based on statistical parameters that describe the overall location, dispersion, shape, and modality of the distribution. In addition to the mean and standard deviation of the distribution, two additional moments were introduced: the skewness (γ), which 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 distribution's tails. Recognizing the importance of characterizing the number of different pore sizes that are likely to have contributed to the observed selectivity drop, this study tried find a simple descriptor for measuring the number of modes in the distribution. The Sarle's bimodality coefficient, $BC = (\gamma^2 + 1)/k$, provides a simple quantification of the extent to which the distribution deviates from unimodality by considering only skewness and kurtosis.[Tarba_2022]

Finally, to assess the diversity of pores, an effective number $n_{eff} = N^2/\sum n_i^2$ of pore sizes was introduced, where N represents the total number of points in the histogram and n_i the number of points associated with the i^{th} bin. This number bears resemblance to a statistical measure widely used in other scientific domains. In political science, it is employed to measure the effective number of political parties,[neffposci_Laakso1979] while in ecology, the inverse Simpson's index evaluates the species diversity in an ecosystem,.[neffbio_Simpson1949] Similarly, in quantum physics, the inverse participation number measures the degree of localization of a wave-function.[neffphys_Kramer1993] This effective number of pore sizes provides an indication of the diversity of pore sizes (considering a binning of 0.1 Å). A highly effective number suggests that multiple pore sizes are well represented in the structure. Thus, this descriptor provides

insight into the scattering of pore sizes within the distribution. All these descriptors contain valuable information regarding the form of the PSD required to understand the loading and selectivity situation in the framework near saturation uptake, which is crucial to predict the ambient-pressure selectivity.

Grid-based and geometrical descriptors

The low-pressure selectivity provides a first intuition about the selectivity at higher pressure, as demonstrated in the previous work, where a correlation between selectivity at both pressures (section 2.3.1) was observed. If the Gibbs free energy formalism is adopted(Equation 2.26), which correspond to a logarithmic transform of the selectivity values, this correlation is confirmed and illustrated in Figure 4.6. It is worth noting that while the selectivity of the majority of structures remains similar under both pressure conditions, a few structures experience a drop in selectivity at higher pressure. The grid sampling approach consistently yields zero-loading selectivity values that are higher or comparable to the ambient-pressure selectivity, providing a reliable foundation for the development of an efficient prediction model. The development of explanatory descriptors related to this selectivity drop phenomenon is the second requirement for a good prediction model. The presence of larger pores, which are less attractive to xenon, is one of the main causes of the selectivity drop. Therefore, additional information on the pore size distributions or the energy landscape would be helpful for this task.

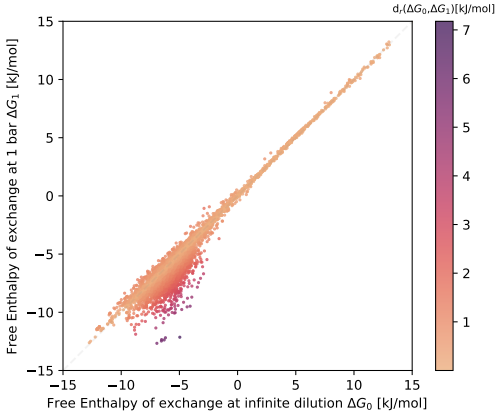


Figure 4.6: Comparison between the Gibbs free energy of exchange at low pressure ΔG_0 and ambient pressure ΔG_1 labeled by the relative distance between them. This plot is equivalent to a logarithmic plot of the selectivity at these two pressure conditions.

To incorporate information on the pore size diversity of the materials, statistical measurements were carried out on the PSD. Explanatory factors at the origin of the observed selectivity drop were detected through analysis. A high degree of multi-modality in the distribution would indicate a diverse set of pores, which can result in a selectivity drop if the pores significantly differ from each other. The chance of observing a selectivity drop increases as the average pore size deviates further from the largest cavity diameter, as a substantial difference in pore sizes leads to lower selectivity. These statistics aim to provide extensive knowledge about a hypothetical selectivity drop and quantitatively estimate its magnitude.

To better quantify the change of selectivity, it could be interesting to present statistics on the distribution of interaction energies for xenon and krypton calculated by the grid algorithm. These statistics include moments of different orders (up to 4) of the energy distribution, which

inform on the adsorbate–adsorbent interaction energies in the nanopores at higher loading. Analyzing the shape of the energy distribution facilitates a quantitative assessment of the selectivity change. This approach compresses the entire energy distribution into a few statistical values, a common method employed in data science to tackle distribution data. The same methodology has been applied to Boltzmann weighted distributions to generate temperature-specific descriptors for the energy distributions. All these quantities have been calculated and compared to ambient-pressure selectivity in Chapter 3 (section 3.3).

As explained in the previous chapter, Boltzmann averaging at higher temperature provided better results in describing ambient-pressure selectivity. This novel type of descriptor proves particularly effective in the high selectivity region, where the standard Boltzmann average at 298 K loses its accuracy (Figure 3.32). This descriptor was utilized to build several descriptors presented in Table 4.3. As illustrated in Figure 4.11, the exchange Gibbs free energy at 900 K and the excess of free energy compared to the 298 K case rank as the second and third most influential descriptors in this ML model. They complement the exchange Gibbs free energy at 298 K in predicting selectivity at higher pressures.

By combining the above-mentioned features with more standard geometrical descriptors, an ML model was trained for ambient pressure selectivity. The model identifies the origins of the selectivity drop and yields 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.

[1] 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: $\Delta_T\left(-T\Delta_{exc}S^{Xe/Kr}\right) = \Delta_T\left(\Delta_{exc}G^{Xe/Kr} - \Delta_{exc}H^{Xe/Kr}\right)$
"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 as 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

The eXtreme Gradient Boosting (XGBoost) algorithm was chosen as the machine learning framework for the predictive model due to its accuracy, efficiency, and simplicity of use. Its performance has been extensively demonstrated, as evidenced by 17 out of 29 winning solutions in Kaggle Challenges being based on this algorithm in 2015. The XGBoost system is highly scalable and parallelized, resulting in fast model training.[chen2016xgboost] Compared to more conventional tree-based algorithms like random forest (commonly used in the field[Simon_2015]), the boosting component of the algorithm enables learning from previous mistakes and allocating greater effort to problematic data points, thereby improving the accuracy of the final ML model.

In the following sections, new descriptors for nanoporous materials will be introduced, along with novel concepts of feature engineering based on energy and pore size histograms. The ML features have been selected through progressive filtering, eliminating less influential features on the accuracy of the final model. A complete list of the feature used can be found in Tables 4.2 and 4.3. The influence or importance of these features will be defined in a subsequent section dedicated to model interpretation. The hyperparameters of the model were fine-tuned through random searches to design the best-performing final model. Lastly, a unified approach will be employed to interpret the influence of the preselected descriptors on the final model.

Hyperparameter optimization

The search for hyperparameters involves finding the best model to optimize the generalization error, as defined in equation 4.7. The most common strategy is to perform cross-validations to evaluate different model configurations, known as hyperparameter search or optimization. In this case, the randomized search algorithm was utilized to find the best parameters within a predefined reasonable range. Through a random search of 30 000 iterations on the parameter space, a set of optimal hyperparameters (refer to Table 4.1 for the meaning of the parameters) for the final ML model was obtained using the following Python dictionary:

```
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),
          }
```

At each iteration, the hyperparameters of the model were evaluated using a 5-fold cross-validation on the training data. The final set of optimal hyperparameters, which corresponded to the model with the lowest RMSE of 0.36 kJ mol^{-1} is provided below. This final set of parameters was then used to train the 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, another 5-fold cross-validation was performed using the set of parameters mentioned above. The convergence plot of the XGBoost model with this parameter set is shown in Figure 4.7. With this configuration, the model was tested on the predefined test set, and interpretation tools were employed to gain a better understanding of the structure-property relationships at play.

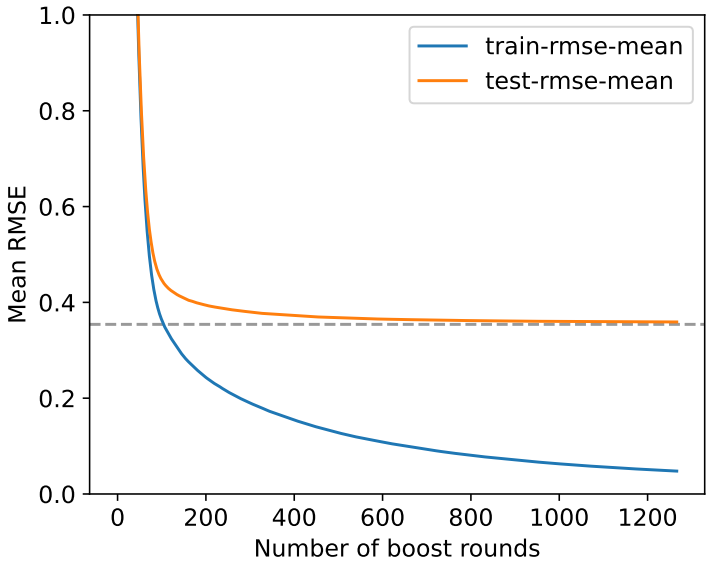
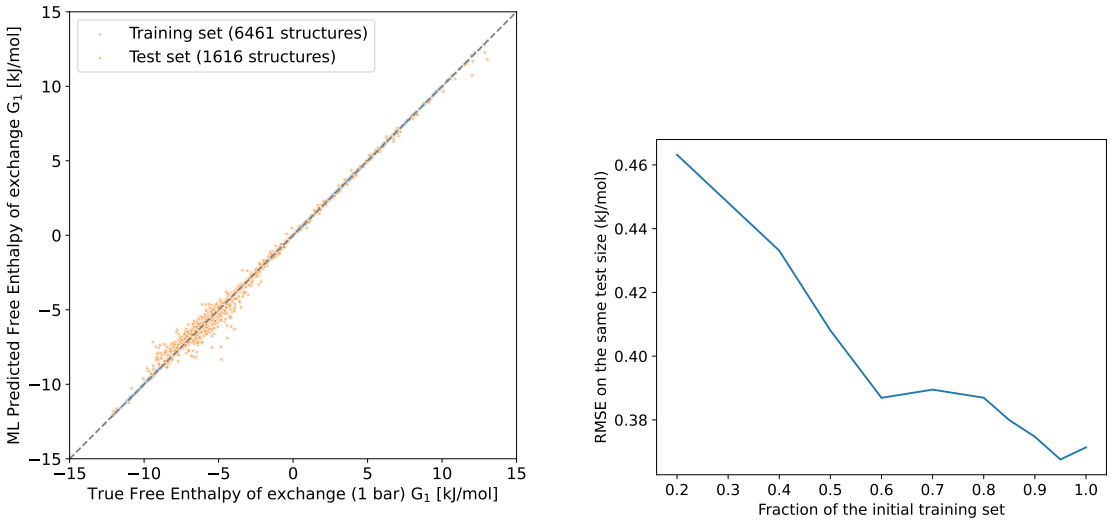


Figure 4.7: Convergence plot of the cross-validation training of our ML model. With the training set considered, the generalization error on the test set converges to 0.36 kJ mol^{-1}.

4.2.4 ML model performance

In this section, the performance of the ML model, which learned the joint effects of all the newly introduced descriptors, will be presented 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 required an average of 2 400 s per structure on the CoRE MOF 2019 database, while the grid-based adsorption calculation only took about 5 s per structure (on a single Intel Xeon Platinum 8168 core at 2.7 GHz). Computing all the necessary features for the prediction would take less than a minute per structure, significantly faster than the 40 minutes required for a GCMC calculation. The ML-based approach clearly demonstrates a speed advantage over standard molecular simulations. However, it needs to maintain a high level of accuracy on an unseen set of structures to be a good substitute.



(a) Final model prediction *(b) Generalization error convergence*

Figure 4.8: (a) Scatterplot 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 scatterplot between the logarithm of the ambient-pressure selectivity (Figure 4.9). (b) Root mean squared errors on the same test set (20% of all data) as a function of the fraction of the training set used to train smaller models. The error decreases as the amount of data increases.*

To train and fine-tune the parameters of our model, a set of 80% randomly chosen structures from the final dataset was defined. A randomized search was conducted over a range of maximum depths, learning rates, sizes of feature samples used by tree and by level, sizes of data samples, and alpha regularization parameters. A set of hyperparameters was selected to minimize the average RMSE computed using a 5-fold cross-validation. The ranges used in the randomized search, as well as the final set of hyperparameters, are provided in section 4.2.3. With this parameterization, our XGBoost model achieved 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. Converting these results back to selectivity values, the RMSE on the selectivity values would be 2.5, and the logarithm base 10 of the selectivity would have an RMSE of 0.07, indicating a very accurate estimation of the selectivity order of magnitude. To demonstrate that this performance is not coincidental, a 5-fold cross-validation procedure was conducted on the entire dataset, yielding 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 obtained using a standard train/test split.

To assess the possibility of training a better model with an increased amount of training data, models were trained using different fractions of the training set, as depicted in Figure 4.8b. It is observed that the RMSE decreases predictably as the data amount is increased, with stabilization occurring around a fraction of 95% of the training set. This indicates that the model has sufficient training data to achieve what appears to be the minimum error on this particular test set.

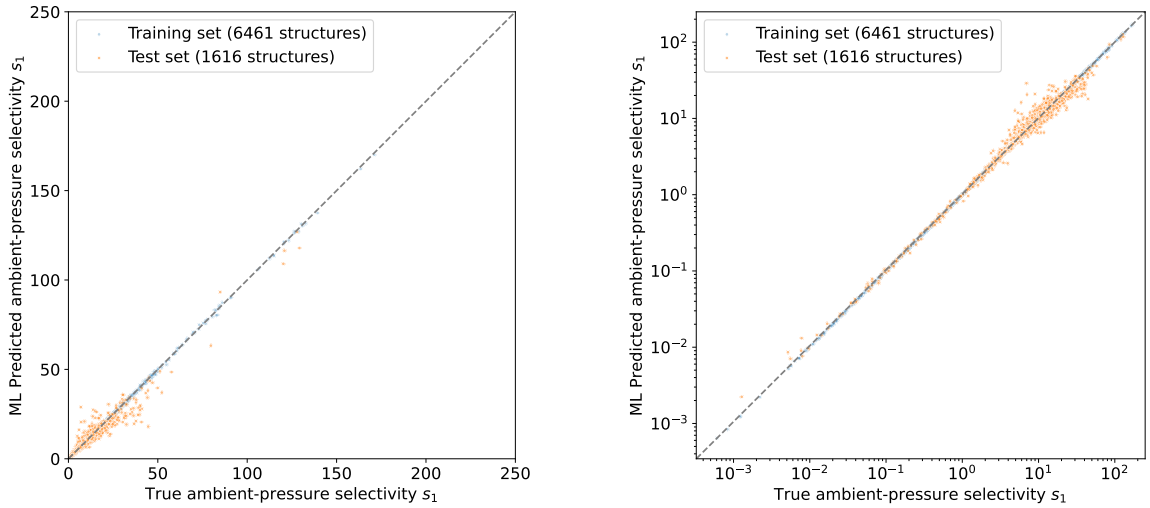


Figure 4.9: Scatterplots of the selectivity predicted by the model of ML compared to the selectivity calculated by GCMC in both log and linear scales. The blue points correspond to the training data while the test data is shown in orange. The focus is on the test data since it shows the generalization of the ML model to unseen data. 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.

This method can later be used in a screening procedure that relies on inexpensive descriptors to filter out obviously undesirable structures, retaining only the promising structures for final evaluation by the ML model. To achieve this, as previously explained in the methods section, only 3D MOF structures with an LCD above 4 Å are retained, as they possess a positive affinity for xenon, which is a necessary condition for a good Xe/Kr selectivity. Given the excellent predictive performance of the model in this thesis regarding ambient pressure selectivity in structures with good xenon affinity, the proposed screening procedure, illustrated in Figure 4.10, would include (i) verifying the nature of the structure to ensure it is a 3D MOF structure, (ii) applying a filter based on the LCD value (above 4 Å), (iii) performing a pre-evaluation of Xe/Kr selectivity at infinite dilution using the grid-based method, and (iv) conducting the ML evaluation to retain only structures above a certain threshold of ambient-pressure selectivity (*e.g.* 30). Additionally, a more comprehensive assessment of the top structures could be conducted using GCMC simulations, *ab initio* calculations, or adsorption experiments.

4.3 Opening the black box

To gain deeper insights into the reasons behind the selectivity drop, the SHAP library of interpretation models[**SHAP**, **molnar2020interpretable**] are utilized to establish relationships between descriptors and predicted ambient-pressure selectivity. Based on Shapley values[**shapley1953value**] which measure the contribution of each descriptor to the prediction, this code facilitates local interpretation of the ML model in this study by disentangling the interdependence between descriptors and extracting individual contributions. To go beyond local interpretation, Shapley values for the entire dataset can be rapidly computed using faster algorithms.[**SHAP**] This allows

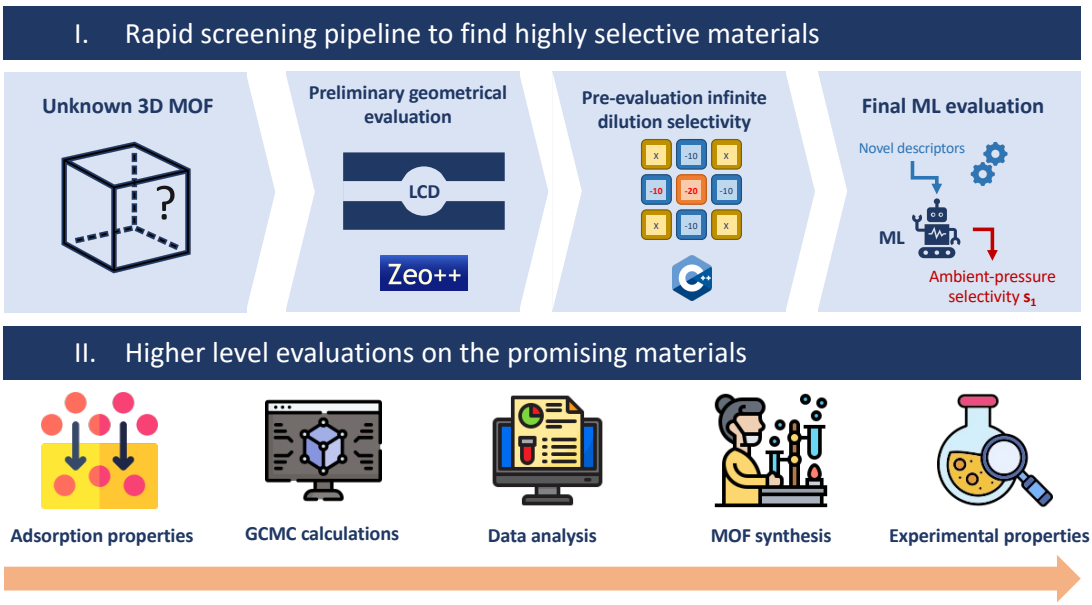


Figure 4.10: An illustration of the screening procedure that could be used to find highly selective materials. The adsorption properties can be rapidly evaluated using structural and energetic conditions on the structure and by confirming it with the ML model. The structures chosen this way can then be tested with higher-level calculations and experiments.

the creation of scatterplots, known as SHAP dependence plots, which depict the contribution as a function of descriptor values and enable a more comprehensive interpretation of our ML model on a global scale. By knowing a descriptor value, it becomes possible to infer, with a certain level of uncertainty, how it influences the final predicted value, thereby shedding light on previously unknown structure–property relationships. Lastly, the mean absolute Shapley values of each feature on the training set were utilized to assess feature importance (Figure 4.11).

Explainable AI

The final model is trained on the predefined training set using XGBoost with fine-tuned hyperparameters. While evaluating its accuracy on the test set provides valuable insights into the performance of the approach, extracting chemical insights into the hidden relationship between the predicted value and the descriptors also help better understand the thermodynamic origins of the performance. In this work, Shapley values,,[shapley1953value] a game theory concept developed by Shapley in 1953, is employed to measure the contribution of each descriptor to the predicted value. This tool is used locally to understand how the characteristics of a given structure contribute to the prediction. To establish structure-property relationships, a global interpretation method such as the SHapley Additive exPlanations (SHAP) method, extensively detailed in Christoph Molnar's online book *Interpretable Machine Learning*.[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. It integrates useful global interpretation modules such as SHAP feature importance and dependence plot.

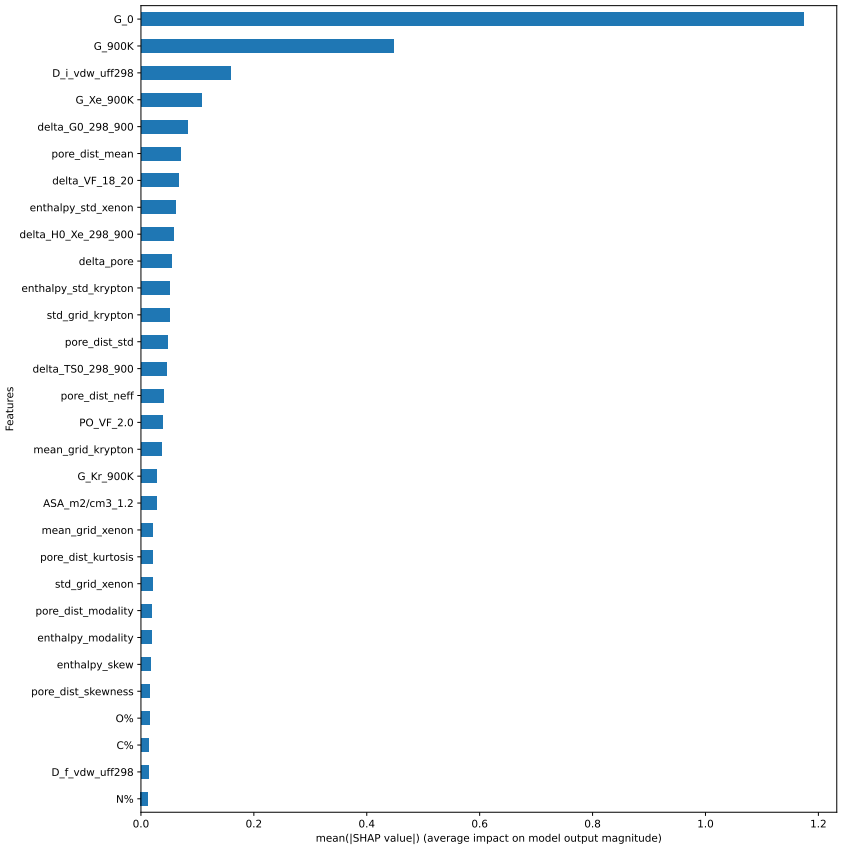


Figure 4.11: Barplot of the feature importance for all the descriptors of our final model. The descriptor labels used in this section are explained in more detail in Tables 4.2 and 4.3.

4.3.1 Global interpretability

The descriptors are ranked based on the mean absolute Shapley values of each descriptor to assess their average impact on the model output's magnitude. The importance plot, associated with these values, is presented in Figure 4.11. Although the low-selectivity exchange Gibbs free energy demonstrates significantly higher SHAP importance compared to other descriptors, it serves as a baseline to achieve a correlation similar to that shown in Figure 4.6. The other descriptors play a major role in moving the outliers of the figure closer to the diagonal line. Energy descriptors significantly influence the model's predictions, while the geometry-based new descriptors play a secondary yet essential role in assessing the differences between the low-pressure and ambient-pressure cases of interest. To gain a deeper understanding of the mechanisms enabling accurate selectivity prediction by the model — the RMSE and MAE on the test set's selectivity being respectively 2.5 and 1.1. The SHAP dependence plots of each interesting descriptor will be examined, depicting the contribution to the predicted value as a function of the actual descriptor value.

The partial dependence module offered by the SHAP library is applied to provide a comprehensive interpretation of the model. Although other methods, such as partial dependence plots, can be used to compute dependence plots (*e.g.* partial dependence plots),,**molnar2020interpretable** maintaining a good level of consistency between global and local interpretations by utilizing the same underlying theory is desirable. The SHAP dependence plots for all descriptors in Figures S9 and S10 exhibit distinct forms, directions, and shapes, which bodes well for the

interpretability of the model. Valuable information regarding how the ML model predicts ambient-pressure selectivity was gleaned from the profile of these dependence plots.

The most important descriptor is the exchange free energy "G_0" associated with low-pressure selectivity. Its contribution displays a very strong positive linear correlation (Figure 4.12). This descriptor establishes a baseline, on top of which other contributions either decrease the free energy (more selective) or increase it (less selective). The model can be interpreted as a combination of a baseline and smaller adjustments estimating the deviation magnitude from the ideal low dilution case. For instance, the next two descriptors "G_900K" (900 K low-pressure exchange free energy) and "G_Xe_900K" (900 K low-pressure xenon adsorption free energy), further contribute to the baseline by providing information on low-pressure selectivity. Moreover, they also offer insights into the deviations necessary to differentiate structures experiencing a drop in selectivity from those maintaining their selectivity. As illustrated in Chapter 3 (Figure 3.32 and 3.33), thermodynamic quantities at high pressure are closer to the 900 K case than to the ambient temperature one. These two descriptors naturally inform the selectivity at higher pressures. In the case of "G_900K" (Figure 4.12), 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 relationship, whereas values between -4 and $5\,\mathrm{kJ\,mol^{-1}}$ constantly yield 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. Further examples highlighting the determination of selectivity contributions using the remaining descriptors will be presented.

The optimal values for the associated descriptors can be highlighted by the U-shape of some SHAP dependence plots. For instance, an optimal value of around 5.1 is observed for "D_i_vdw_uff298" (Figure 4.12), while the optimal average pore size is approximately 5.6. These optimal values align with the physical requirement of having xenon-sized pores to enhance xenon's attraction, as identified in various literature papers. However, it should be noted that these values are slightly higher than those mentioned in the literature due to differences in atom radius definitions.**Hung_2021** Moreover, values of "delta_G0_298_900" between 4 and $6\,\mathrm{kJ\,mol^{-1}}$ (Figure 4.12) have a higher likelihood of contributing negatively, indicating lower ambient-pressure selectivity. These sweet spots provide valuable insights for distinguishing truly selective materials from others. Some SHAP dependence plots have a rather linear domain for the most selective structures (in blue) — a good linear contribution is observed for the difference of pore volumes between Xe and Kr sized probes "delta_VF_18_20" (Figure 4.12). This implies that a lower void fraction difference corresponds to a more selective structure. The same trend is observed for the standard deviations of the PSD, denoted as "pore_dist_std", and the Boltzmann weighted krypton interaction energies distribution, referred to as "enthalpy_std_krypton". Optimal values for these descriptors tend to be zero. As the value approaches zero, the contribution becomes more negative, indicating a more selective structure at ambient pressure.

In some cases, optimal values are not concentrated around well-identified values but are encompassed within larger domains with threshold values separating them. For instance, the difference between the LCD and the average pore size, denoted as "delta_pore", has a threshold value around 0.3 Å. Below this threshold, the contribution for the most selective structures (blue) is negative (Figure 4.12). Although clear correlations are not evident, a threshold value

(of approximately 0.23) indicates a higher probability of achieving high ambient-pressure selectivity. The same type of domain splits can be observed for the average of the 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 ambient temperature "delta_TS0_298_900" (at around 3) and high temperature, and the effective number associated with the PSD "pore_dist_neff" (at around 2.3). These domains serve to distinguish structures with a selectivity drop at ambient pressure from those without, particularly important for identifying selective structures at low pressure.

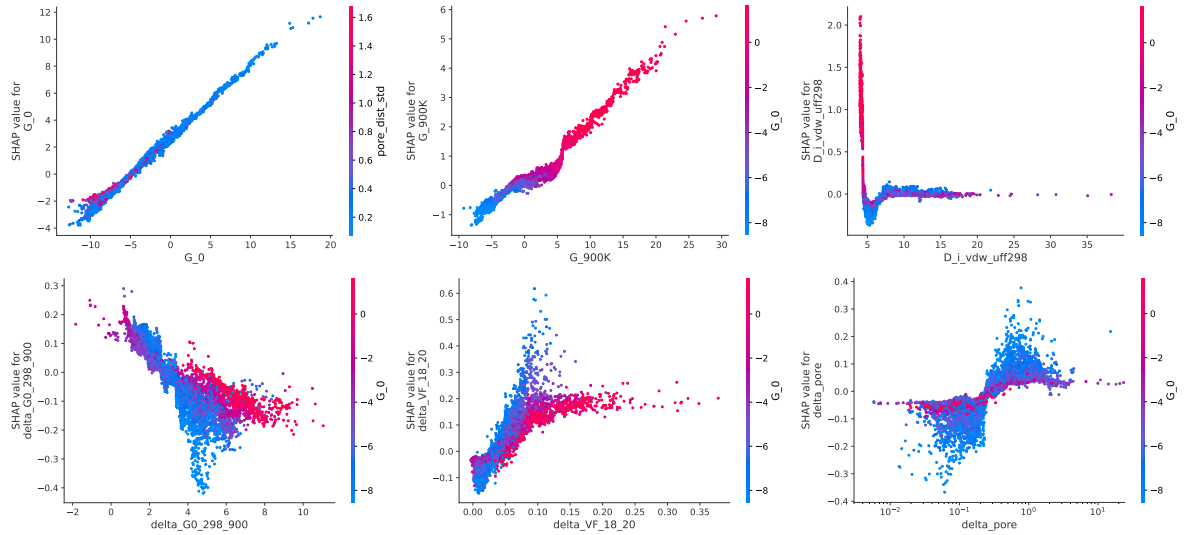


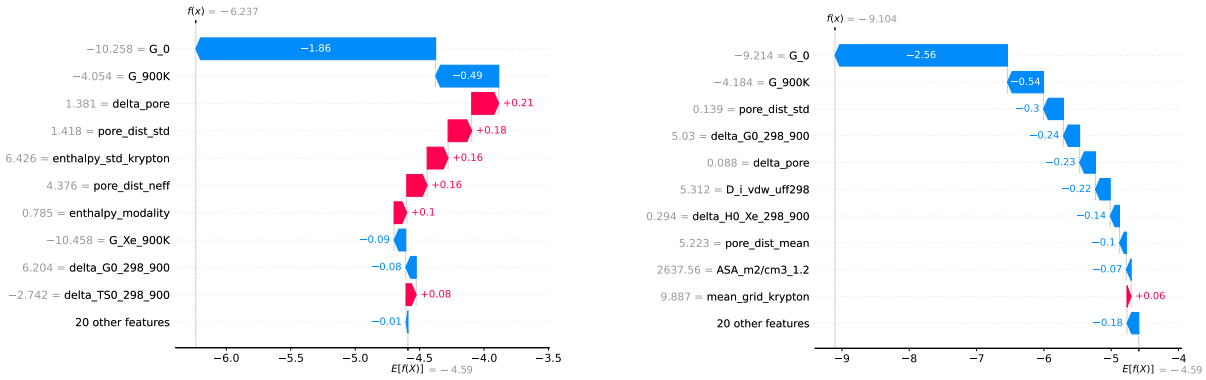
Figure 4.12: Some relevant SHAP dependence plots that are provided here. A SHAP dependence plot corresponds to the Shapley values as a function of the feature values for all structures of the dataset. These SHAP plots show the contribution of the features to the prediction given by the ML model. Each Shapley value depends not only on the value of the feature itself but also on the values of other features. For this reason, the plots are labeled by a relevant second feature.

4.3.2 Local interpretability

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

`VIWMIZ` belongs to the category of highly selective structures that undergo a selectivity drop at ambient pressure. When converting the free energy values to selectivity values, `VIWMIZ` has a selectivity of 62.8 at infinite dilution and 14.5 at ambient pressure. The ML model successfully predicts a close value of 12.0 for the ambient-pressure selectivity based on the given descriptor values. Specifically, the descriptor "G_0" has a highly negative value, which explains its relatively high negative contribution of −1.81. However, the contribution of "G_900K" is relatively low at −0.57 compared to other materials (Figure 4.12), as a value of −4.05 is not the most negative among all structures. Conversely, the remaining descriptors have positive contributions, which lead to the selectivity drop. For instance, the difference in

pore sizes, "delta_pore", has a value of 1.38 Å (above the threshold of 0.23 Å), which contributes +0.25 to the predicted selectivity. This value aligns with the value ranges observed in the associated dependence plot. Similar analyses can be performed on the positive contributions of other descriptors shown in Figure 4.13 by referring to the dependence plots: "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" falls 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 somewhat close to its optimal value, resulting in a negative contribution in this specific prediction. The remaining features have negligible contributions. Analyzing the contributions of each descriptor to the prediction given by the model of this work helps understanding the underlying features of the VIWMIZ structure that explain the selectivity drop at higher pressure. Descriptors such as the shape of the xenon and krypton energy distributions ("enthalpy_std_krypton" and "enthalpy_modality") and the PSD ("pore_dist_std" and "pore_dist_neff") as well as the void fraction difference "delta_pore" play key roles in the lower selectivity at ambient pressure compared to the ideal infinite dilution case. Intuitively, an effective number of pores exceeding 2 suggests the presence of different pore sizes, which aligns with the presence of less attractive pores for xenon, ultimately leading to decreased selectivity. This observation is consistent with a high standard deviation of the PSD or the Boltzmann-weighted krypton interaction energy distribution. Furthermore, a significant difference between the average pore size and the LCD indicates a disparity in pore sizes, resulting in larger pores that become increasingly loaded as pressure rises. However, interpreting the entropic term is more complex and presents unexplored ways of addressing the selectivity drop at higher pressure, as revealed in the previous study (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 Table 4.2 and 4.3.

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 stability of the selectivity is accurately predicted by assigning a value of 40.0. Consequently, the first contribution of "G_0" is one of the most negative contributions, establishing a baseline of −2.4 for subsequent contributions. Although the contributions of "G_900K" and "G_900K", they continue to decrease the predicted selectivity value. The joint contributions of other descriptors will discriminate between the two structures and determine why this particular structure will maintain its selectivity. In contrast to the previously analyzed structure, this structure has a "delta_pore"

value below 0.3 Å, explaining its negative Shapley value in the prediction of this study. The contribution of "delta_G0_298_900", which had only a slightly negative impact on the other structure, now plays a significant role as it falls within the range of 4 to 6 kJ mol^{-1} (Figure 4.13). Additionally, it is observed that "pore_dist_std" is below the threshold, in contrast to the previous structure where it was above the threshold. Furthermore, the other contributions align with the rules suggested by the SHAP dependence plots, and no apparent anomalies are detected. The combined effects of all the descriptors result in a lower free energy value, leading to the conservation of selectivity at higher pressure. The set of descriptor values for this structure significantly differs from the previous one, with many values contributing to opposite domains. This disparity allows the model to differentiate between highly selective structures and identify those that will maintain their selectivity at higher pressure.

These two examples provide a deeper understanding of how the model distinguishes structures that lose selectivity at higher pressure from those that do not. Most dependence plots exhibit a strong association between descriptors and their effects, with outliers being rare enough to comprehend the internal logic of the model. As previously discussed, the first three descriptors establish a baseline for the observed selectivity drop with limited information. Subsequently, the contributions of other descriptors can be positive, negligible, or negative depending on the domain where the values of the descriptor lie. For instance, the average pore size and largest cavity diameter need to be within specific ranges to maximize the likelihood of maintaining selectivity at higher pressure, aligning with previous studies emphasizing the importance of pore sizes similar to xenon for Xe/Kr separation. The difference in entropy between ambient temperature and 900 K is a surprising descriptor that separates selective structures based on whether its value falls within a specified range. Similarly, the difference in void fraction occupied by xenon and krypton is intriguing as it impacts selectivity differently depending on whether the structure is highly selective or not, with the contribution being more or less proportional to its value. Various methods of measuring the disparity of the PSD and interaction energy distribution play a key role in identifying highly selective structures (indicated in blue on the dependence plot in Figure 4.12) that either maintain or decrease in selectivity. These methods include calculating the difference between the average pore size and the LCD, as well as the standard deviation of the PSD or Boltzmann-weighted energy distribution, which exhibit distinct behaviors based on the domain in which the value lies. The SHAP dependence plots provide valuable insights into the mechanisms underlying the ML model presented in this thesis and, more broadly, shed light on the understanding of Xe/Kr separation origins.

4.4 Beyond thermodynamic considerations

To gain a deeper understanding of separation processes within nanoporous materials, a machine learning prediction of Xe/Kr ambient-pressure selectivity was performed, aiming for faster results compared to standard GCMC calculations. The CoRE MOF 2019 database was utilized for MOF structures, enabling the evaluation of xenon/krypton selectivity in less than a minute, whereas an equivalent GCMC calculation typically requires approximately 40 min. Unlike the majority of selectivity predictions in the literature, the decision was made to predict selectivity on a logarithmic scale that focuses on the order of magnitude rather than the exact value of selectivity for highly selective materials. Moreover, converting to an exchange Gibbs free energy allowed for a more thermodynamic approach based on enthalpy, entropy, and free energy values.

The challenge consisted of predicting the free energy equivalent of ambient-pressure selectivity using low-pressure selectivity alongside key energy, geometrical and chemical descriptors. The resulting fully optimized ML model exhibited high performance, yielding an RMSE of 0.36 kJ mol^{-1}, which corresponds to an RMSE of 0.06 on the base-10 log of selectivity.

One specific objective was to uncover the underlying reasons for the observed selectivity drop at high pressure in certain highly selective materials at low pressure. initially studied in Chapter 2. Previous studies (Chapter 2) found that high diversity of pore sizes and channel sizes that favor adsorbate reorganizations could be at the origin of this phenomenon. Through the application of interpretability tools, quantitative factors explaining the conservation or decrease in selectivity for highly selective materials were identified. Depending on energy averaging at 900 K, statistical characterizations of energy or pore size distributions, and differences in occupiable volumes, a structure could exhibit either a selectivity similar to the infinite dilution case or a substantially lower selectivity at higher pressure. The XGBoost model employed in this study utilizes a complex ensemble of decision trees to capture the quantitative rules that can be extracted from the model and used to establish heuristics supporting the intuition about Xe/Kr selectivity in MOF structures.

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

The ambition of this study was to introduce new descriptor ideas that contribute to the development of increasingly efficient screening methodologies for identifying optimal materials for specific applications. However, similar to other studies in this field, the simulations in this study relied on a set of strong assumptions, wherein rigid frameworks and non-polarized classical forcefields were employed. Previous literature suggested that the most selective materials for Xe/Kr separation were designed and synthesized based on the effect of open-metal sites, leveraging the difference in polarizability between the two molecules to achieve efficient separation.[Li_2019, Pei_2022] Moreover, the flexibility of structures could be achieved by employing flexible forcefields with appropriate simulation methodologies[Bousquet2012] or by conducting multiple rigid simulations using snapshots from NPT simulations.[Witman_2017] The simulations could be enhanced at the cost of CPU time by coupling them with a reduction in simulation time, such as the one presented in this chapter. The pursuit of ever-faster evaluation tools enabled the exploration of more complex properties and the discovery of structures with increasingly relevant characteristics. Further discussion on these potential improvements will be presented in greater detail in Chapter 6.

The next chapter will focus on another shortcoming of the simple thermodynamic approach, namely the consideration of kinetic limitations in the screening process. In the problem of adsorbent-material separation, the transport effect does not play a key role in evaluating the separation performance of the material. Instead, it serves as a constraint that could potentially affect the performance of some seemingly top materials. Various methodologies will be examined to incorporate these transport properties, which are computationally demanding and challenging to obtain experimentally.

5

XENON AND KRYPTON TRANSPORT PROPERTIES

In separation processes, transport properties govern the kinetics of the adsorption process. Two distinct use cases for nanoporous materials in separation processes exist: adsorption-based separation, which is primarily a thermodynamic process, and nanoporous separation membranes, which rely on kinetic properties. Depending on the application, diffusion is either the main performance metric or a secondary parameter that is often overlooked. In membrane-based processes, for instance, gases are sieved through a membrane material that selectively blocks certain molecules (e.g., Xe) while allowing other particles to diffuse freely. The performance of the separation is measured through the ratio of diffusion coefficients, rather than the thermodynamic selectivity defined in Chapter 2. However, the thermodynamic selectivity remains the primary performance metric in industrially performed adsorption-based separation processes investigated in my work, such as pressure and/or temperature swing adsorption (PSA, TSA or PTSA). Nevertheless, considering the kinetic performance can enhance the overall industrial process.[Kumar_1994] 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 aim of this chapter is to explore this frequently overlooked diffusion parameter in an adsorption-based Xe/Kr separation process.

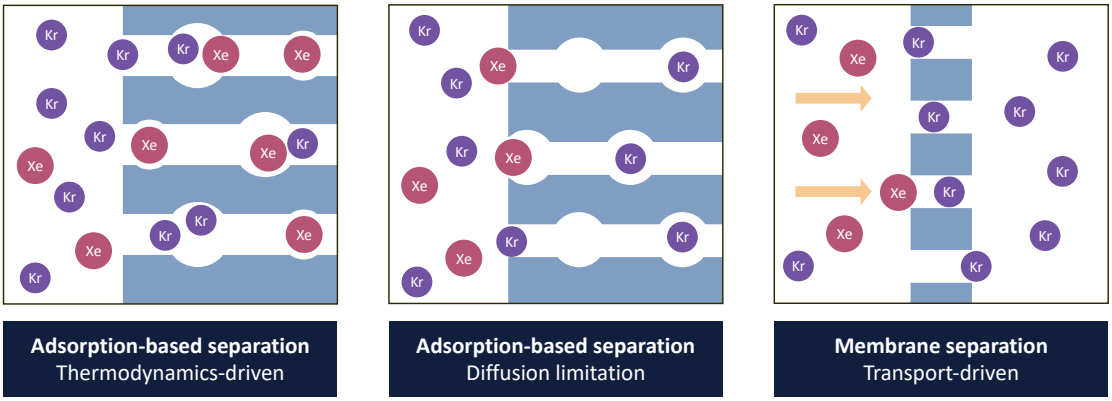


Figure 5.1: Illustration of the comparative role of the thermodynamic and transport properties for Xe/Kr separation in nanoporous materials. From the transport dominated process of membrane separation to the thermodynamically equilibrated separation processes in the nanopores, different more nuanced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

5.1 Modeling the diffusion process

Since the observation of pollen motion by the botanist Brown in 1826, the seemingly erratic movement of particles in a static bulk medium has been thoroughly observed and studied by scientists. Subsequently, Fick proposed a macroscopic model for this phenomenon, known as Brownian motion, by introducing the coefficient D_x in a diffusion equation 5.1 (1D) based on experimental measurements of concentration ϕ..**Fick_1855** According to this law (valid only for independent particles), particles tend to move from regions of higher concentration to regions of lower concentration within the bulk medium.

$$\frac{\partial \phi}{\partial t} = D_x \frac{\partial^2 \phi}{\partial x^2} \tag{5.1}$$

To gain a better understanding of the Brownian motion of suspended particles in a liquid, Einstein derived a microscopic model of diffusion motion based on the molecular-kinetic theory of heat in the momentous year of 1905.**einstein1905motion** To determine the self-diffusion coefficient (referred to as the "diffusion coefficient" hereafter), he observed the motion of an individual particle assumed to be independent of other particles, with time steps large enough to consider two consecutive motions as mutually independent. By considering the particle distribution of N independent diffusing particles, he redefined the diffusion coefficient as a function of the mean squared displacement (MSD) of a particle. In tridimensional space, the following Einstein relation is applicable:

$$\langle r(t)^2 \rangle = 2dD_{\text{diff}}t = 6D_{\text{diff}}t \tag{5.2}$$

where d is the dimension of the space in which the particle diffuses ($d = 3$ in a volume) and $r(t)$ 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 the ease of particle movement within a nanoporous material. A low diffusion coefficient indicates limited access to the structure's pores, as illustrated in Figure 5.1. It is worth mentioning that in porous media, non-Fickian diffusion processes can occur, the MSD has a linear relation to time. For example, when particles are confined in a one-dimensional channel that does not allow a particle to jump over another, the dynamics is described by a single file diffusion equation, and the MSD has square root relation to time.**Levitt_1973**

To model the diffusion coefficient of xenon and krypton inside nanoporous materials, molecular simulations of the adsorbate displacements will be utilized. Although alternative approaches such as the Green-Kubo equation exist, the comparatively less complex Einstein law is preferred for self-diffusion calculations, as demonstrated by this cited comparative study [**Maginn_2020**]. This section will focus on various simulation techniques that can be used for evaluating diffusion in high-throughput screenings. Different methods of assessing the MSD of diffusing particles will be presented, starting from the straightforward approach of molecular dynamics to faster methodologies more suitable for screenings, such as machine-learned surrogate models and kinetic Monte Carlo simulations.

5.1.1 Molecular dynamics

Molecular dynamics (MD) simulations are used to reproduce the microscopic motion of molecules in a given system and to calculate thermodynamic averages by assuming the equivalence between time averaging and ensemble averaging (ergodic hypothesis). For other applications, mechanical properties, thermodynamic properties, or chemical properties can also be determined. However, the main focus here is on calculating diffusion coefficients of monoatomic molecules, with the discussion centered around averaging trajectories to obtain MSD values.

Simulation details

Molecular dynamics aims at describing the motion of particles, subjected to forces exerted by surrounding particles. This process can be viewed as a step-by-step 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 from the cumulated interactions with its surroundings 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 classical modeling, forces are determined using the aptly named forcefield that was previously introduced in Chapter 2. In that context, intermolecular interactions were simplistically modeled by the Lennard-Jones (LJ) interaction potential between atom pairs, which is also employed in this section (although other methods for defining a forcefield exist). By utilizing the LJ potentials U^{LJ} (defined in equation 2.4), the vectorial force $\mathbf{f}_{ij}$ between two atoms i and j can be derived:

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

where ϵ_{ij} and σ_{ij} are the LJ parameters of the atom pair ij. The resulting force is obtained by summing the forces $\mathbf{F}_i = \sum_j \mathbf{f}_{ij}$ exerted by the surrounding atoms j. To reduce the computa-

tion time required, molecular simulations consider only the atoms within a specified cutoff radius.

Now that the force $\mathbf{F}_i$ has been defined, the molecule can be set in motion by numerically integrating equation 5.3 from time t to time $t + \delta t$. Various methods exist to integrate equations of motion, such as the Euler or velocity-Verlet scheme presented in the book by Frenkel et al.[frenkel2001md] In this work, the focus will be on the *leap frog* integration implemented in the RASPA2[dubbeldam2016] software used for the MD simulations. The position $\mathbf{r}_i$ and velocity $\dot{\mathbf{r}}_i$ are updated at 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 $(\mathbf{r}_i(0), \dot{\mathbf{r}}_i(0.5\delta t))$, the center of mass of molecule i can be translated to any position $\mathbf{r}_i(t_n = n * \delta t)$. The rotation step required for polyatomic molecules will be omitted since the study focuses solely on monoatomic noble gases. The different positions $\{(t_n, \mathbf{r}_i(t_n))\}_{n=0,\ldots,\mathrm{N_{tot}}}$ constitute the total trajectory of the MD simulation (velocities are not mentioned for readability, but they are also propagated). This total trajectory can be used to derive the average MSD, which can be further analyzed to calculate the diffusion coefficient.

Diffusivity calculation using an MD trajectory

The MSD sampling technique implemented in RASPA2,[dubbeldam2016] presented in an article [**Dubbeldam_2009**] by several authors of the adsorption simulation software, was employed. The approach is based on a modified version of the order-**n** algorithm described in the book [**frenkel2001msd**] by Frenkel and Smit. The focus in this chapter will be on the multiple-window algorithm used to calculate the diffusion coefficients of xenon and krypton.

To understand this computation, it is necessary to first explain what the window algorithm does and how it can be generalized to the multiple-window algorithm. First, 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. A straightforward approach would be to compute the square displacement $\left\|\mathbf{r}_i(\tau) - \mathbf{r}_i(0)\right\|^2$ for a sub-trajectory $\mathcal{T}(0 \rightarrow \tau)$ of duration τ. However, this alone is insufficient to obtain a statistically meaningful average of the MSD, as described by the Einstein equation 5.2. By assuming independence between two movements of the same particle separated by a time δt, as hypothesized in Einstein's paper [**einstein1905motion**], shifting the origin of time by δt would yield another trajectory. This process can be repeated i times while $\tau + i\delta t \leq t_{\mathrm{tot}}$. Although highly accurate, this approach is highly inefficient when $\tau \gg \delta t$ 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 independent sub-trajectories, a sampling time step of $\delta\tau \lesssim \tau$, on the same order of magnitude as τ, can be employed. To achieve this, the window approach first defines a value $\delta\tau$ and generates $\mathrm{N}_\tau = \lfloor (t_{\mathrm{t\,ot}} - \tau)/\delta\tau \rfloor$ 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 ranging from 1 to K that defines the time window to be sampled. This allows the calculation of the MSD $\langle r(\tau)^2 \rangle$ for duration values of τ equal to $\delta\tau, \ldots, \mathrm{K}\delta\tau$. The resulting MSD $\langle r(\tau)^2 \rangle$ is linear with respect to time, it can then be fitted to equation 5.2 to obtain the diffusion coefficient. The trajectory generation using the window approach is illustrated in Figure 5.2 for a decomposition into sub-trajectories of a duration $\tau = 3\delta\tau$ shifted by $\delta\tau$.

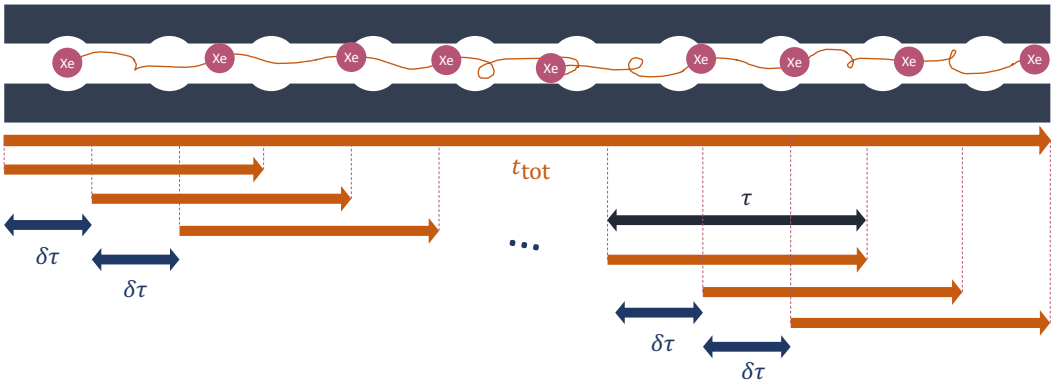


Figure 5.2: Illustration of the generation of trajectories of size τ by shifting the origins of multiple durations $\delta\tau$.

The major drawback of this method is that a timescale $\{\delta\tau, \ldots, K\delta\tau\}$ needs to be defined in advance. To access the different timescales within a single simulation, the multiple-window algorithm developed by Dubbeldam et al. and implemented in the RASPA2 software for calculating mean squared displacements (MSD) in molecular dynamics simulations is employed.

The different time windows are recursively defined using the default parameters of RASPA2. The first time window is defined by $K = 25$ displacements with durations $\delta t, 2\delta t, \ldots, (K-1)\delta t$ and a shift of δt (the default shift value δt of the first window can be modified using the parameter "SampleMSDEvery"). The second window is then based on a sampling time $\delta\tau_1 = K\delta t$, and the sub-trajectories have durations $\tau_1^{(1)} = \delta\tau_1, \ldots, \tau_1^{(K-1)} = (K-1)\delta\tau_1$. This recursive process is repeated, where the i_{th} window has a sampling time of $\delta\tau_i = K^i\delta t$ and sub-trajectories with durations $\tau_i^{(1)} = \delta\tau_i, \ldots, \tau_i^{(K-1)} = (K-1)\delta\tau_i$. A window algorithm similar to the one described above is applied to each window. The algorithm terminates when no further window can be generated, i.e., when $\tau_n^{(k)} > t_{tot}$ where n is the index of the window and k is the index in the single-window algorithm that defines the desired sampling time with respect to $\delta\tau_i$. The timescale $\delta\tau_i = K^i\delta t$ sampled follows a geometric progression, allowing access to very different timescales. This enables the identification of the timescale corresponding to the diffusion regime (linear relationship between the MSD and the duration of the sub-trajectories used in the averaging). The different timescales and the exponent value b from fitting a function of the form $x \mapsto ax^b$ for the different time windows are illustrated in the next section (Figure 5.5) — values of b close to 1 can be associated with a diffusion regime. The determination of the diffusion coefficient is then simplified to a simple fitting problem, which will be further explained in the presentation of the diffusion coefficient screening in section 5.2.

This methodology can then be replicated to thousands of structures to characterize the diffusion properties of these materials. Numerous screenings have already been conducted in the literature, as presented in Chapter 1, specifically in the section dedicated to transport property. The prediction of these quantities using machine learning will now be explored in more detail.

ML MODELING

In a very recent study, Daglar et al. used an ML model to predict the diffusion coefficient of 100,000 hypothetical MOFs by utilizing data on around 5,000 CoRE MOF structures.**Daglar_2022**

Alongside conventional geometrical descriptors, they employed 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 within various MOF materials from CoRE MOF 2019 (training dataset) and hMOF (testing dataset). The combination of kinetic and thermodynamic data for the characterization of MOF materials is an intriguing approach. However, a key limitation of most existing approaches in the literature is the lack of structure–property relationship that elucidates the underlying microscopic origins of diffusion coefficient values.

Following a similar approach to the thermodynamic screening (Chapters 2–4), the present work on transport property screening aims to establish structure-property relationships between diffusion coefficients and geometric descriptors of MOF structures. To gain deeper insights into the diffusion process, the diffusion activation energy will be evaluated using energy grid-based methods described in the literature. These techniques are designed to enhance the prediction of diffusion coefficients, either through direct calculations or by employing ML surrogate models. To this end, kinetic Monte Carlo approach will be introduced, which, although less accurate than the MD approach, offers significantly higher computational efficiency.

5.1.2 Lattice kinetic Monte Carlo

The lattice kinetic Monte Carlo method relies on a predefined lattice of stable points corresponding to adsorption sites. Each site is connected to another if there exists a diffusion path (narrow channel) between them. To calculate the probability of transition from one site to another, a transition state (TS) within the narrow channel corresponding to the highest energy point along the minimal energy diffusion path (the saddle point), must be defined. In the transition state therory, the transition probability is defined based on the energy barrier that must be overcome to traverse the channel. Once the lattice is established, the adsorbate can be propagated from one site to another using the different transition probabilities. Although this approach yields a coarse-grained trajectory compared to the one obtained through MD simulations, it is sufficient for computing the MSD and the diffusion coefficient.

Transition state theory for diffusion

In chemistry, the kinetics of a reaction are often explained using transition state theory. This theory compares the energy of the reactants to that of the transition state, allowing for the calculation of the reaction rate. Along a reaction path, this rate is proportional to the ratio between the Boltzmann factor at the transition state and the integration of Boltzmann factors along the reaction path.

This definition can be directly transposed to the case of a diffusion path instead of a reaction path. Thus, the diffusion rate $\mathcal{R}_{0\to1}$ from site 0 to 1 can be defined as follows:

$$\mathcal{R}_{0\to1} = \kappa\sqrt{\frac{k_\mathrm{B}\mathrm{T}}{2\pi m}}\frac{e^{-\beta \mathrm{E}(\mathbf{r}^{\mathrm{TS}})/k_\mathrm{B}\mathrm{T}}}{\int_{\mathrm{path}} e^{-\beta \mathrm{E}(\mathbf{r})/k_\mathrm{B}\mathrm{T}}\,\mathrm{d}\mathbf{r}} \tag{5.6}$$

κ is the Bennet-Chandler dynamic correction factor (or recrossing probability),[BENNETT1977] and $\kappa = \frac{1}{2}$ if it is equiprobable to reach both sites from the transition state. This requires the determination of the optimal diffusion path before determining the diffusion rate. Wang et al. (2022) adopted this approach in their noble gas separation screening, where they first determined the minimal energy path before calculating diffusion rates.[Wang_2022]

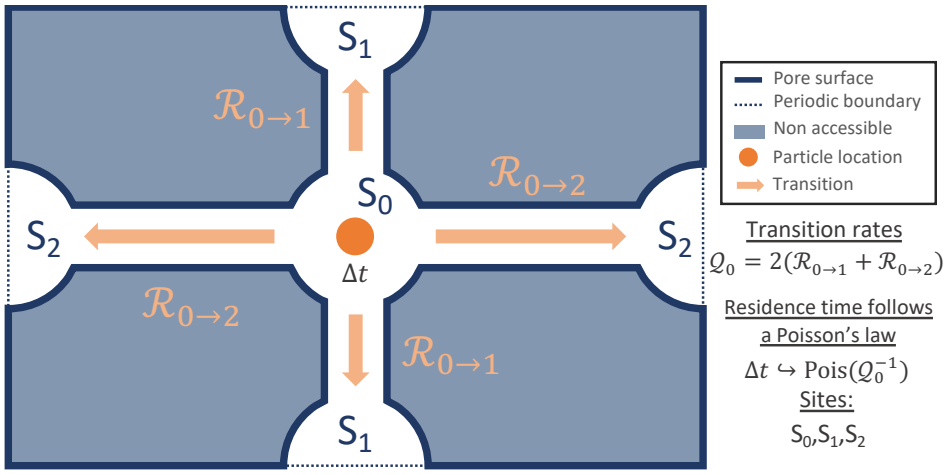


Figure 5.3: Illustration of the core principle of lattice kinetic Monte Carlo in a periodic system. Within the periodic system, the movement of particles is governed by transition probabilities from one site to another. The transition rates are determined using the activation energy needed to move to the transition state between the two stable adsorption sites, as shown in equations 5.6 and 5.7.

Another approach involves determining a transition surface through which the adsorbate passes to diffuse along the channel of the material. In this case, only the determination of the transition only, but it relies on another definition of the transition rate $\mathcal{R}_{0\to1}$:

$$\mathcal{R}_{0\to1} = \kappa \sqrt{\frac{k_\mathrm{B}\mathrm{T}}{2\pi m}} \frac{\int_{\mathcal{S}(\mathrm{TS})} e^{-\beta \mathrm{E}(\mathbf{r})/k_\mathrm{B}\mathrm{T}}\, \mathrm{d}\mathbf{r}}{\int_{\mathcal{V}(\mathrm{S}_0)} e^{-\beta \mathrm{E}(\mathbf{r})/k_\mathrm{B}\mathrm{T}}\, \mathrm{d}\mathbf{r}} \tag{5.7}$$

where the integration for the transition state is performed on a bottleneck surface that a diffusing particle must cross to move from the volume occupied by site 0 to the one occupied by site 1.

TuTraST algorithm for kinetic MC

In this section, the focus shifts to the second approach, which relies on the determination of a transition state as a surface. This approach was developed by Mace et al.[Mace_2019] and involves detecting merging points between basins that represent the adsorption sites. The algorithm developed for this purpose is known as TuTraST, which stands for Tunnels and Transition States. It is a search algorithm that aims to identify the tunnels and transition states that separate different adsorption "basins" within them. Once the adsorption sites, transition states, and connecting tunnels are identified, hopping rates between stable adsorption sites can be calculated using Equation 5.7. Finally, a lattice kinetic Monte Carlo simulation can be employed to move particles within a simplified hopping diffusion system to determine the MSD and, ultimately, the diffusion coefficient.

In practice, the calculation of an energy grid that detects all the different components of the lattice is required. The algorithm iterates over different energy values from $\mathrm{E}_{\min} + \delta\mathrm{E}, \ldots, \mathrm{E}_{\min} + \mathrm{N}\delta\mathrm{E}$, such that the maximum energy is below a predetermined energy cutoff $\mathrm{E}_{\min} + \mathrm{N}\delta\mathrm{E} < \mathrm{E}_{\mathrm{cutoff}}$. At the initial step, the clusters of grid points with an energy below $\mathrm{E}_{\min} + \delta\mathrm{E}$ are naturally formed based on their connectivity. Then, at iteration level L, the clusters found in the previous iteration L – 1 are grown layer by layer (one layer corresponds to the immediate

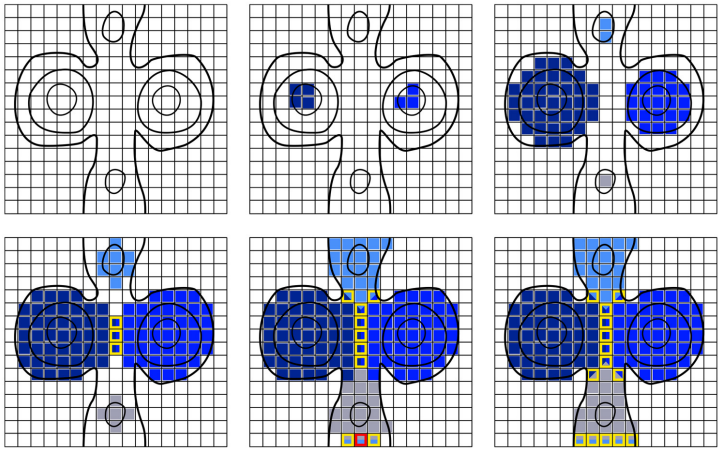


Figure 5.4: Illustration of the cluster growth and the identification of boundary points in the TuTraST algorithm.[Mace_2019] *The clusters are grown from left to right and top to bottom. A tunnel is detected when the points are connected all the way from one periodic boundary to another (red). The boundary points assigned to the TS surface are indicated in yellow. Reprinted with permission from the original paper [***Mace_2019***] copyright © 2019 American Chemical Society.*

neighbors on the grid), as shown in Figure 5.4. If a layer touches another cluster, a boundary point is identified — it is tagged as a point of the TS surface if the energy barrier to go from the basin to this point is sufficiently high, else the energy barrier is negligible, and the basins are merged. At the end of the process, the boundary values are clustered and assigned as the transition surface between different pairs of adsorption basins. If the points of the basins and boundary surfaces are connected all the way through, a tunnel can be defined for diffusion to occur. After a sufficient number of iterations, tunnels with different basins separated by transition state surfaces are established. A kinetic Monte Carlo simulation can then be performed to determine the diffusion coefficient within each tunnel. The diffusion coefficient in the material corresponds to a weighted average of the diffusion coefficients in each tunnel, with the weight determined by the probability of presence in each tunnel, which corresponds to the sum of the Boltzmann factors (proportional to the Henry coefficient in a given tunnel).

This approach is very promising as it is significantly more efficient than MD simulation-based techniques. The code of Mace et al.[Mace_2019] implemented in Matlab (although not computationally very efficient) already outperforms most MD simulations in terms of computation time for diffusion coefficient calculation, with minimal compromise in accuracy, as demonstrated in their diffusion coefficient screening of zeolites. To enhance efficiency further, the algorithm for efficient search of transition states was rewritten in C++. At this stage of development, the focus was on determining the diffusion activation energy, which is independent of transition state detection. Detailed algorithmic information for determining the diffusion activation energy in nanoporous materials using the in-house algorithm is provided in section 5.3, along with the projected development of a faster lattice kinetic Monte Carlo simulation inspired by the TuTraST algorithm.

5.2 Self-diffusion screening

To complement the thermodynamic screenings conducted in chapters 2–4, a screening of transport properties, specifically diffusion coefficients, was also performed. This section presents the screening approach and analyzes the diffusion coefficients in comparison with typical geometric descriptors.

5.2.1 Diffusion in a selective material

Before delving into the details of the screening study, the approach adopted for calculating diffusion coefficients using MSD values is demonstrated through an example: SBMOF-1.[Banerjee_2016] This preliminary study aids in calibrating the time parameters (time step, maximum time) for the final screening study.

First, an MD simulation of 500 million steps (approximately 1–2 days of simulation) was conducted, including a thousand initialization steps and 100 thousand equilibration steps to model a xenon atom diffusing in the KAXQIL[Banerjee2012] MOF at "infinite dilution". To achieve infinite dilution, the box size was adjusted to prevent interactions between different adsorbates. In these conditions, as shown in Figure 5.5, there are different timescales at which distinct transport phenomena occur.

Within the range of 1 fs to 1 ps, a ballistic regime is observed, with the mean squared displacement following a squared dependence. For a particle of mass m, the MSD $\langle r(\tau)^2 \rangle$ in this regime obeys a simple ballistic relation (length equals velocity multiplied by time):

$$\langle r(\tau)^2 \rangle = v_m^2 \tau^2 = \frac{k_\mathrm{B} \mathrm{T}}{2\pi m} \tau^2 \tag{5.8}$$

where v_m is the average velocity of a particle, following the Maxwell-Boltzmann distribution at temperature T. Calculating the squared mean velocity v_m^2 using the standard Maxwell-Boltzmann relation yields a value of $3\,\mathrm{m}^2\,\mathrm{s}^{-2}$, which closely matches the value of $2.8\,\mathrm{m}^2\,\mathrm{s}^{-2}$ obtained by the fit in the right plot of Figure 5.5. This initial regime corresponds to the movement of particles subjected to thermal agitation, but it holds little significance for diffusion.

A transition from the ballistic regime to the pseudo-diffusional regime (where the exponent has not yet reached 1) is observed in the plot in cyan. Between 1 ps and 100 ps, a sub-diffusion regime is observed, characterized by an MSD that follows a power function of time, $\langle r(\tau)^2 \rangle = \mathrm{K}_\alpha \tau^\alpha$, with an exponent less than 1, as illustrated in the left plot of Figure 5.5. This regime corresponds to the confinement of xenon particles within adsorption pores, where only thermal vibration occurs, and no diffusion hopping is observed at this timescale. Diffusion appears to begin at the 10 ns timescale. The MSD between 0.01 ns and 0.4 ns, as shown in Figure 5.6, represents a sub-diffusional regime due to the confinement imposed by the nanopores of KAXQIL. However, at 0.4 ns–9 ns, the MSD starts to exhibit an exponent of 0.7, which is closer to 1, allowing for a linear fit, albeit imperfect (Figure 5.6). Ideally, trajectory sampling on the order of tens of nanoseconds would be desirable for the next timescale. However, with an MD time step of 1 fs, this would multiply the computation time by at least 5 (1–2 weeks for one MD simulation), which becomes prohibitive.

By fitting a linear relation using the right plot of Figure 5.6 and deducing the diffusion coefficient, an underestimated value of the diffusion coefficient of $2.24 \times 10^{-8}\ \mathrm{cm}^2\,\mathrm{s}^{-1}$ is obtained — it

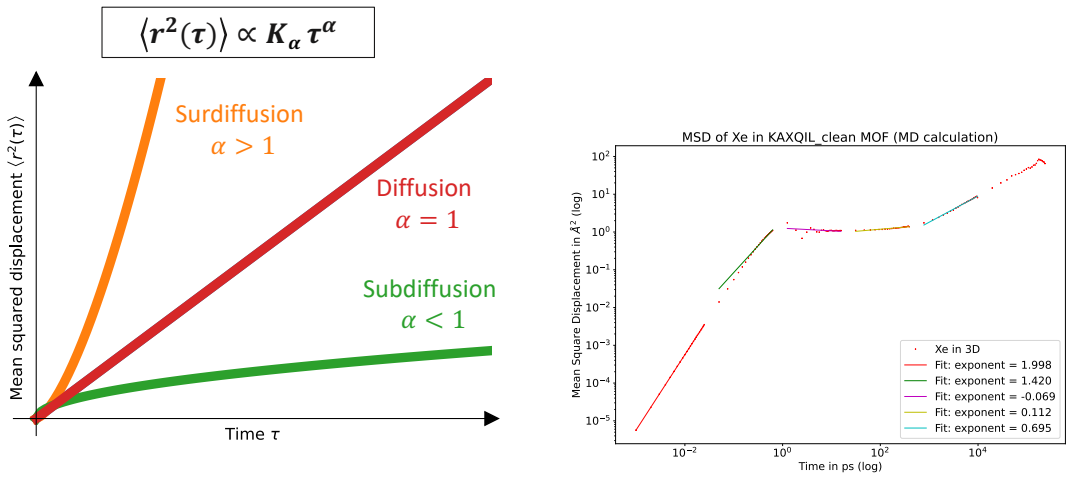


Figure 5.5: Left: Different regimes that could be observed in an MSD plot as a function of time. The ballistic regime can be considered super-diffusional, while the normal diffusion follows a linear relation described by the Einstein equation 5.2. The sub-diffusion regime commonly occurs in obstructed media such as nanoporous materials. The different regimes can be observed in the right plot of the actual MSD, which is calculated using the multiple-window method. Fittings are performed using a generic function $K_\alpha \tau^\alpha$, and the exponents α are provided in the legend.

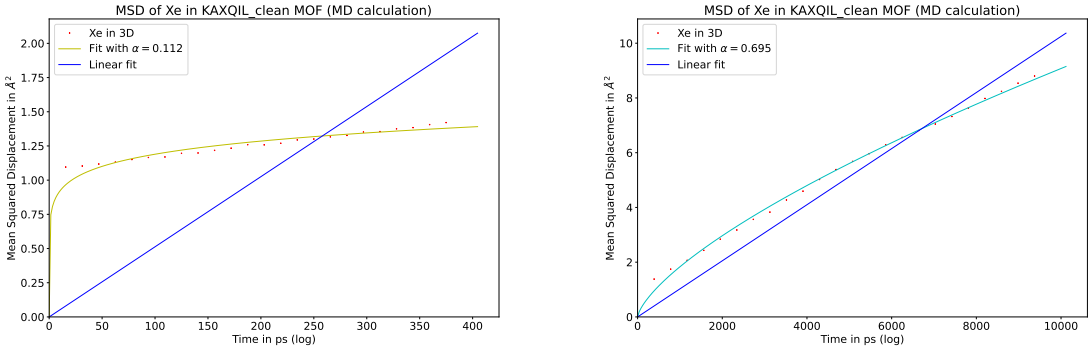


Figure 5.6: Plots of the MSD at the last two timescales considered in Figure 5.5. On the left, the timescale between 0.01 ns *and* 0.4 ns *is considered. The MSD is fitted using a power function with the same previously determined exponent, and a linear fit is provided to demonstrate its incompatibility with the diffusion equation. On the right, a similar approach is employed for the timescale between* 0.4 ns *and* 9 ns.

is an underestimation due to the concave nature of the MSD, which reduces the slope in the fitting process. However, this value is already a good estimation of the diffusion coefficient, considering the relatively high value of the exponent $\alpha = 0.7$ in the fitting equation $K_\alpha \tau^\alpha$.

To account for the randomness in the initial position of xenon (block pockets have been calculated for a 1.5 Å-radius probe), it is necessary to measure the effect of running different MD simulations of the value of the diffusion coefficients. The uncertainty across various MD simulations with different initial positions determined by different random seeds needs to be quantified. In RASPA2, the random seed is equal to the UNIX time upon launching the MD simulation, ensuring that each of the 10 different MD simulations is assigned a different

random seed while using the exact same parameters, as previously mentioned. The diffusion coefficient values were averaged over these simulations, resulting in an average diffusion coefficient of 2.13×10^{-8} cm^2 s^{-1} and a standard deviation of 0.37×10^{-8} cm^2 s^{-1}, representing approximately 17% of the average value. The uncertainty in the diffusion coefficient is estimated to be around 17% for a relatively low coefficient of around 10^{-8} cm²/s, with lower uncertainty expected for less obstructive materials. It is worth mentioning that in the particular case of KAXQIL, where all pores are symmetrically equivalent, the dynamics is more straightforward than for more complex pore architectures that will be tackled later in this chapter.

Increased confidence in the method led to the exploration of higher timescales beyond the reach the previous MD simulation, as the occurrence of the diffusion regime was observed at the 10 ns timescale. An initial calculation with 500 million steps and a timestep of 5 fs was performed to validate the diffusion coefficient value. The time window between 2 ns and 47 ns was considered, and the MSD was calculated from approximately 200 sampled trajectories, resulting in reasonably accurate values. A more accurate diffusion coefficient of 2.6×10^{-8} cm^2 s^{-1} was obtained, which is very close to the value obtained using the previous approach. The value is slightly higher (which was expected), as the previous value was overestimated. Thus, this approach is consistent with the previous one.

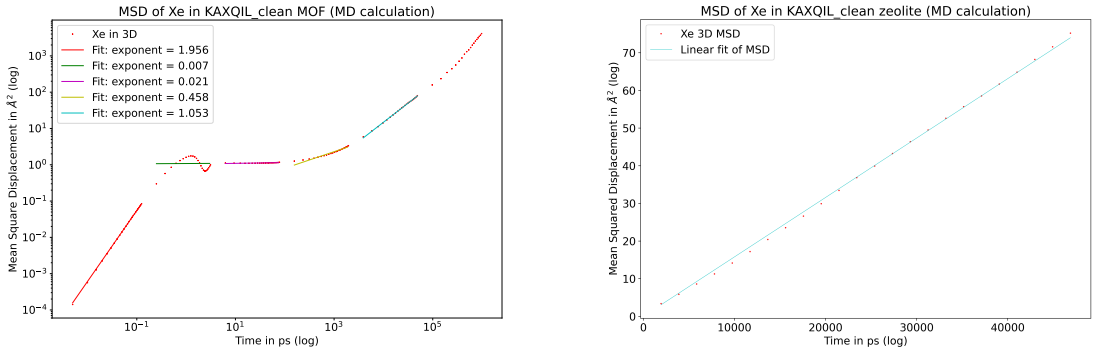


Figure 5.7: On the left panel, the MSD in log-log scale and a fit to the relation MSD(τ) = K$_\alpha \tau^\alpha$ using an MD simulation of a 5 fs-time step and 500 million steps are shown. A better linear fit is obtained with this new configuration of the MD simulation compared to Figure 5.5, as more timescales are explored while utilizing the same computational resources. On the right panel, the MSD on the timescale of the diffusion regime is presented. The linear fit demonstrates improved performance compared to the previous Figure 5.6.

To further support the use of a higher time integration step, the origin of the value of 1 fs needs to be understood. This value is typically justified by the Nyquist-Shannon sampling theorem, which suggests that the integration step should be at most half of the period of the fastest vibration within the system. In the case of a C–H vibration, the maximum time step value is 5 fs, and to be on the safe side, a time step of 1–2 fs is chosen in most diffusion studies in nanoporous materials..**Bukowski_2021** However, in the system of xenon diffusing in a rigid environment, there are no vibrational limitations as previously described. It is hypothesized in this thesis that using higher time steps in this situation can provide easier access to longer timescales. Nevertheless, further studies are required to ensure the validity of the quantities derived from these MD simulations. The value of 5 fs is at the higher end of what is typically observed in MD simulations, but it can be justified by the rigidity of the framework and the

adsorbate being considered. Even higher time steps could be tested, but for more reliable results, a reasonable middle ground of 5 fs was chosen for all high-throughput screening of the transport properties.

5.2.2 High-throughput screening of diffusion coefficients

Screening procedure

To incorporate transport properties into the analysis, MD calculations were performed on 6,525 non-disordered, highly selective materials for xenon or krypton at infinite dilution (no guest–guest interactions). For each material, MD simulations were planned with 500 million steps using the RASPA2 script on the calculation machines (that are restricted to 24-hour runs). To let the simulation run a maximum of steps, 2 to 3 restarts were performed on the slowest simulations, such that every MSD data was obtained after 2–3 days. Out of the planned 500 million steps, only 432 structures completed the simulations at the end of the process. However, this does not imply that the MSD data cannot be utilized for determining diffusion coefficients since the MSD are printed. In fact, 5,125 MSD data are exploitable even if the total 500 millions steps were only partially completed.

To determine the diffusion coefficients, two timescales (2–47ns and 50–950 ns) were analyzed to fit the MSD with a linear function. The linear fit with the highest determination coefficient (within the range of 0 to 1) for both timescales was chosen to obtain a diffusion coefficient value. After this step, structures with a determination coefficient (R^2 that measures the correlation) below 0.9 were removed, leaving 5,125 structures for drawing structure-diffusivity relationships — these structures, for which there is a high degree of confidence in the diffusion coefficient values, will be comparatively studied against different geometrical and thermodynamic quantities in this section. Finally, the final 5,125 diffusion coefficients obtained correspond to the slopes of the best linear fit between the 2–47ns and 50–950 ns timescales.

This approach focuses exclusively on the linear relations between the MSD and time for determining self-diffusion at infinite dilution. The analysis does not explore the nature of the transport property (e.g., single-file diffusion[**Lin_2005**]), by comparing, for example, the exponent of the generalized formula $MSD(\tau) = K_\alpha \tau^\alpha$ with structural descriptors. The more complex diffusion at higher loading values, which may be more accurately described by a collective diffusion coefficient instead of the self-diffusion coefficient, is also not studied. The objective of this study is to identify materials that do not present kinetic limitations, as observed in the case of KAXQIL[**Banerjee2012**] (where the diffusion coefficient of xenon is approximately ten thousand times lower in the material than outside).

Structure–diffusivity relationships

In this section, different relationships between the diffusion coefficient and simple geometrical descriptors will be presented. A forcefield-dependent definition of radii was chosen to ensure better correlation with the results of the MD simulation that utilized the UFF forcefield. The geometrical descriptors were calculated using Zeo++ and these radii to determine the PLD, the largest sphere diameter D_{if} along a free path, the surface area, and the pore volume, as explained in Chapter 2. The use of UFF-based radii for the PLD was further justified by the original paper [**Hung_2021**], which demonstrated a stronger correlation between the PLD and the diffusion constant. This correlation can also be observed in Figure 5.8. The PLD calculated using the standard CCDC-defined atom radii does not fit the diffusion coefficient as well as the

UFF-defined PLD. As shown in a smaller scale in the article [**Haldoupis_2010**] (see Figure 1.6 in Chapter 1), there is a linear relationship between the diffusion activation energy (logarithmic transform of the diffusion coefficient) and the PLD. This linear relationship is much noisier for the PLD defined by the standard CCDC radii than for the UFF-based PLD.

Beyond the practical considerations regarding the calculation of geometrical descriptors, the PLD outlines the variation of diffusion performance within nanoporous materials. First, a linear relationship was observed, as previously highlighted, followed by a plateau. In the first zone, xenon is constrained by channels narrower than its kinetic radius. The diffusion coefficient rises with increasing channel width, and this positive correlation persists until the channel width exceeds approximately 4.6 Å. Beyond this threshold, the diffusion coefficient stabilizes around 3×10^{-4}cm^2 s^{-1}. Variations in this plateau region can only be attributed to other phenomena, such as tortuosity within nanopores or the chemical nature of the surface of nanopores. The diffusion coefficient value can be interpreted as the diffusion coefficient of a "free" xenon, which is less influenced by the surrounding pore surface. For PLD values exceeding 5 Å, the channels are sufficiently wide to allow for xenon toonly undergo a minor slowdown. These findings are consistent with experimental data that measured the diffusion coefficient of xenon dissolved in water at different temperature conditions, where a value of 10^{-5} cm^2 s^{-1} at 303 K was obtained,[**Wise1968**] aligning with the values observed centered around 3×10^{-4} cm^2 s^{-1} at the plateau.

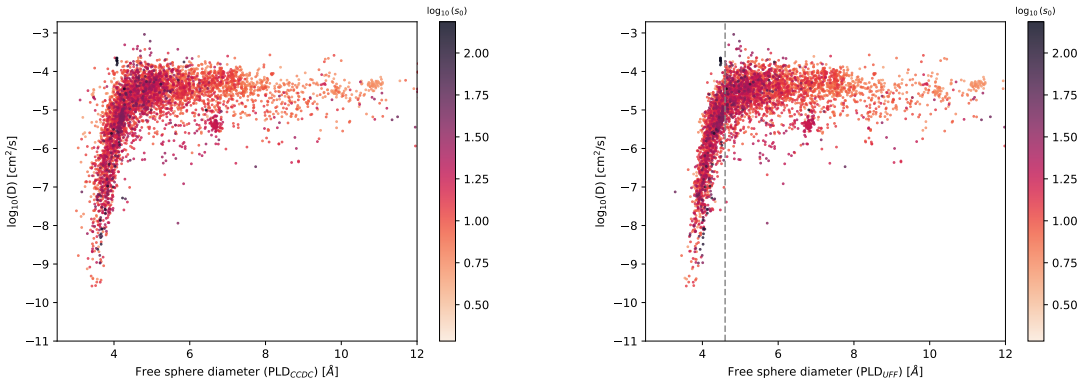


Figure 5.8: Xenon diffusion coefficient at infinite dilution as a function of the pore limiting diameter (PLD). The diameter of the largest free sphere is defined using two different radius systems: the standard CCDC-based PLD (on the left), and the one defined using the UFF forcefield (on the right)[**Hung_2021**] *— as defined in Chapter 2.*

If the channel dimensions (determined using Zeo++) are analyzed, partial information regarding the channel shape can be obtained. The dispersion of diffusion coefficients at the plateau depicted in Figure 5.9 is challenging to characterize visually based on the channel dimension alone.

For this reason, the distribution of diffusion coefficients that depend on the dimensionality of the channels within the framework was plotted in Figure 5.10. The distribution for structures containing 1D structures is characterized by a much heavier tail in terms of low diffusion coefficients. Structures with 1D channels are more likely to have very low diffusion coefficients below 3×10^{-8} cm^2 s^{-1}. The vast majority of structures with tridimensional channels tend to have higher diffusion coefficients between 3×10^{-6} cm^2 s^{-1} and 10^{-4} cm^2 s^{-1}, with almost very

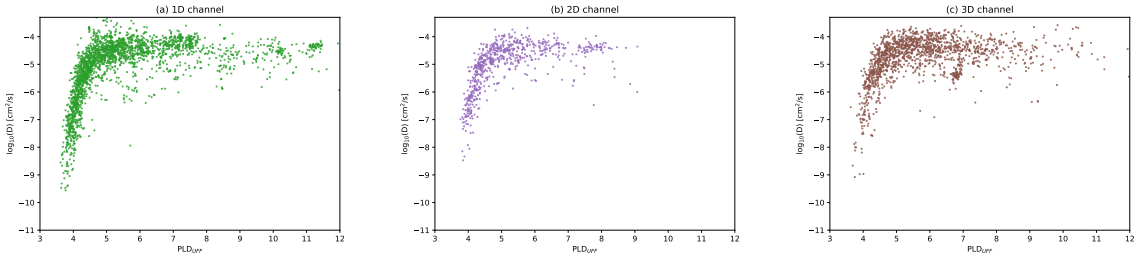


Figure 5.9: Distributions of the base-10 logarithm of the diffusion coefficients of three different subsets of the screened structures. The first one (a) is composed of structures with a unidimensional channel, the second (b) bidimensional channels and the third one (c) tridimensional channels.

few structures having lower diffusion coefficients. The influence of channel dimensionality on diffusion coefficients is not as pronounced for bi- and unidimensional channels. In the case of bi- and unidimensional channels, structures with diffusion coefficients between 3×10^{-8} $cm^2\ s^{-1}$ and 3×10^{-6} $cm^2\ s^{-1}$ are more common, although less frequent in theory. Therefore, the dimensionality of the channels can influence the values of diffusion coefficients, although the relationship is not as clear as for PLD.

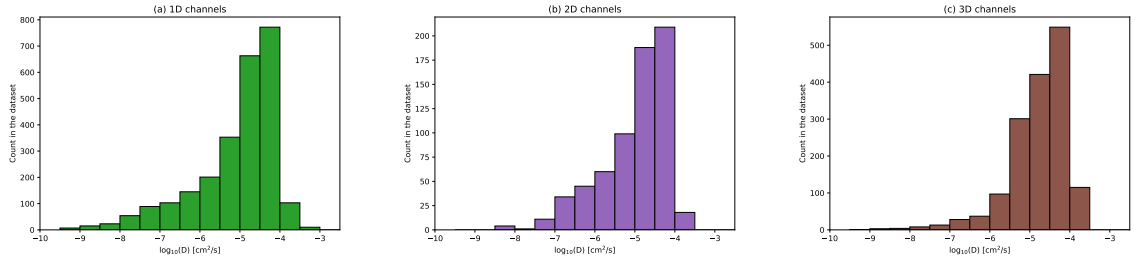


Figure 5.10: Distributions of diffusion coefficient of three different subsets of the screened structures. The first one (a) is composed of structures with a unidimensional channel, the second (b) bidimensional channels and the third one (c) tridimensional channels.

Other geometric properties of the material, such as void fraction and surface area, can also influence diffusion. Low diffusion coefficients are typically observed in materials with small pore volumes below 0.6, as shown in Figure 5.11. However, establishing a direct relationship between void fraction and diffusion coefficient is challenging. The only discernible relationship is that materials with void fractions higher than 0.6 have diffusion coefficients exceeding $3 \times 10^{-6} cm^2\ s^{-1}$. This phenomenon is certainly due to the correlation between PLD and void fraction, as larger PLD values are usually associated with higher void fractions. On the other hand, the accessible surface area for a probe of size 1.2Å does not appear to significantly influence the diffusion coefficient.

Framework density and molar mass are immediate characteristics of the structure that do not require complicated simulations to obtain. However, their relation to the diffusion coefficient is not as straightforward, as shown in Figure 5.12. It can be inferred that low-density values favor high diffusion coefficients, which can be explained by the logical relation between low density and high porosity. On the other hand, there does not seem to be any clear relationships between the molar mass of the framework and the diffusion coefficients, and no simple geometric or physical reasoning would justify any.

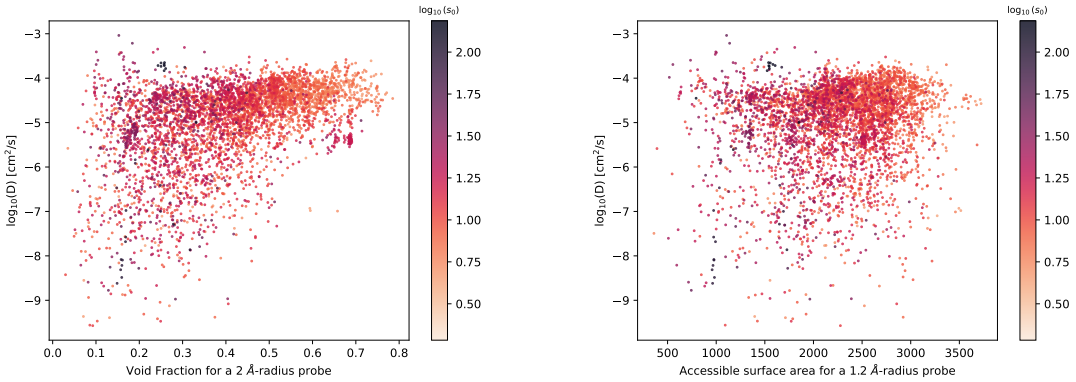


Figure 5.11: Xenon diffusion coefficient at infinite dilution as a function of the accessible surface area (left) and the void fraction (right).

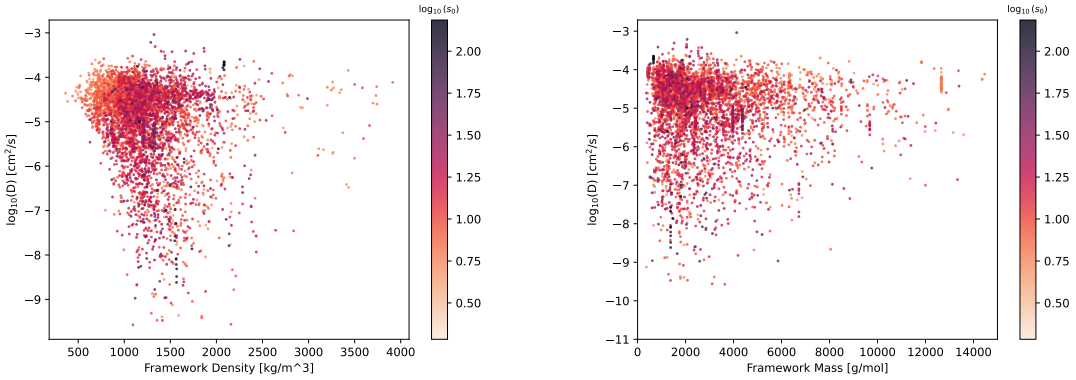


Figure 5.12: Xenon diffusion coefficient at infinite dilution as a function of the density (left) and the mass (right) of the frameworks.

The largest sphere diameter D_{if} along a free diffusion path exhibits a similar relationship to the diffusion coefficient, although the correlations are noisier, as depicted in the left plot of Figure 5.13. This can be explained by the fact that D_{if} is always equal to or greater than the pore limiting diameter D_f by definition. When both diameters are equal, the relationship resembles that shown in Figure 5.8, with a linear correlation and a plateau. However, when it is higher, it creates the observed noise pattern in the left plot of the Figure 5.13.

A final comparison involves a thermodynamic quantity, namely the xenon adsorption enthalpy $\Delta_{ads}^{Xe}H$. There is no relation between diffusion coefficient and the xenon adsorption enthalpy, which is advantageous for screening because it implies the possibility of various configurations. A high diffusion coefficient and a high xenon adsorption affinity (with very negative enthalpy values) can coexist in a material, which represents the ideal configuration for adsorption at infinite dilution. However, it necessitates testing the diffusivity when the material exhibits good affinity to optimize both properties. This approach will constitute the core discussion regarding the optimization of Xe/Kr selectivity and the diffusion coefficients of Xe and Kr.

5.2.3 A trade-off between the selectivity and the diffusion

This section analyzes the screening of diffusion and selectivity properties calculated for xenon and krypton to identify relevant materials that demonstrate both a good Xe/Kr selectivity and

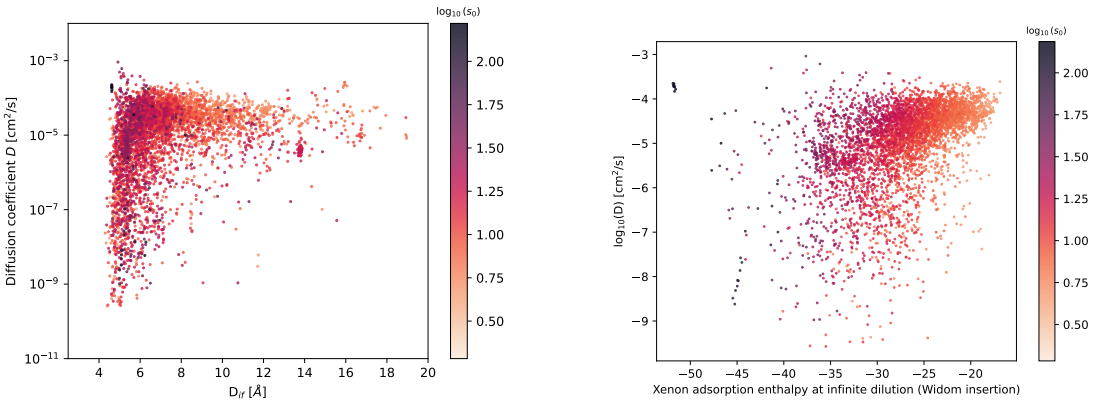


Figure 5.13: Xenon diffusion coefficient at infinite dilution as a function of the largest sphere diameter D_{if} along a free diffusion path (left) and the xenon adsorption enthalpy (right).

a good Xe/Kr diffusion coefficient ratio. To achieve this, a diffusion coefficient screening for krypton was performed, resulting in 4,816 structures with a good determination coefficient R^2 for both linear fits of xenon and krypton MSD. These structures are subsequently tested to find materials that exhibit a balanced combination of thermodynamic and kinetic properties for xenon/krypton separation.

Screening of diffusion selectivity values: a trade-off between adsorption and diffusion

The comparison between xenon/krypton selectivity at infinite dilution and xenon/krypton diffusion coefficients is initiated. Highly selective material can possess a decent diffusion coefficient, indicating that the diffusion limitation observed in the KAXQIL structure is not inevitable, which is encouraging. The left plot of Figure 5.14 clearly demonstrates the possibility of all configurations: high selectivity (above 40) with high diffusion coefficient (over 10^{-6} cm^2 s^{-1}) and high selectivity with low diffusivity. The krypton coefficients exhibit relative stability between 10^{-6} cm^2 s^{-1} and 10^{-3} cm^2 s^{-1}. Consequently, increasing the diffusion selectivity is not the primary leverage for enhancing the diffusion selectivity, as highly selective materials do not display very low krypton diffusion coefficients. To surpass thermodynamic selectivity, a transport-related selectivity metric is examined to identify highly selective materials without significant transport limitations.

The transport properties in a separation process are generally evaluated using the ratio of diffusion coefficients or the diffusion selectivity as performance metrics. For xenon and krypton, the diffusion selectivity can be defined as follows:**Krishna_2010**

$$s_{\text{diff}}^{\text{Xe/Kr}} = \frac{D^{\text{Xe}}}{D^{\text{Kr}}} \tag{5.9}$$

To consider both transport and thermodynamic effects, the thermodynamic adsorption selectivity defined in Chapter 2 (equations 2.24 and 2.24) is combined with the diffusion selectivity to define the membrane selectivity (used to characterize membranes). This membrane selectivity can also be referred to as permselectivity, as it corresponds to the ratio of permeabilities of the

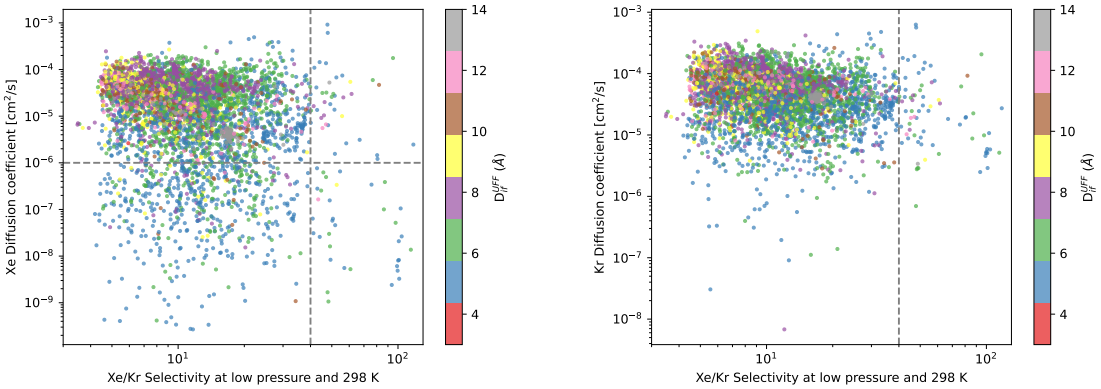


Figure 5.14: On the left panel, the Xe/Kr diffusion selectivity (ratio of the diffusion coefficients) is plotted against the Xe/Kr selectivity values at infinite diluation (calculated by Widom insertion), and the points are color-coded by the largest cavity diameter within a diffusion path D_{if}. On the right panel, the same plot is now color-coded with the average dimension of the channels of the nanoporous structures associated.

components in the binary mixture targeted for separation. The xenon/krypton permselectivity ($s_{\text{perm}}^{\text{Xe/Kr}}$) can thus be defined as follows:

$$s_{\text{perm}}^{\text{Xe/Kr}} = s_{\text{diff}}^{\text{Xe/Kr}} \times s_{\text{ads}}^{\text{Xe/Kr}} \tag{5.10}$$

where $s_{\text{ads}}^{\text{Xe/Kr}}$ corresponds to the adsorption selectivity used throughout the previous chapters (at infinite dilution or higher pressure).

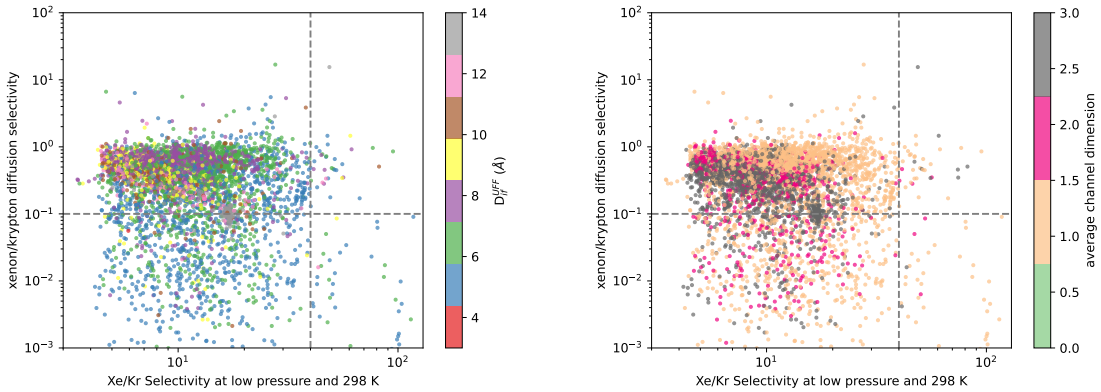


Figure 5.15

Using both $s_{\text{diff}}^{\text{Xe/Kr}}$ and $s_{\text{ads}}^{\text{Xe/Kr}}$ at different pressure conditions, this study will try to find materials that exhibit a relatively high selectivity with high diffusion selectivity. The plots presented in Figure 5.15 demonstrate that a total of 48 structures display a selectivity above 40, along with a good diffusion selectivity over 0.1. Notably, these structures possess rather large pore sizes, represented by the largest included sphere along a free diffusion path D_{if} as depicted in the left plot of Figure 5.15. Additionally, these large pores are associated with structures that exhibit various dimensionalities. Among these structures, one particular standout is characterized by exceptionally high diffusion selectivity (over 15, as indicated by the gray

point on the upper right side of the left plot of Figure 5.15) coupled with a high adsorption selectivity at infinite dilution. It is worth mentioning that this structure, with a CSD code ADOGEH[**Peikert_2012**], features a three-dimensional channel framework with large pores and relatively narrow connecting channels. However, the high adsorption selectivity observed at infinite dilution is not maintained under ambient pressure conditions, as illustrated in Figure 5.16 (refer to Table 5.1 for further details).

Upon closer examination of these 48 structures, it becomes evident that they incorporate a combination of large and small pores, such that the diffusion is not obstructed, while achieving high selectivity within more confined spaces. Materials with varying pore sizes of this nature may experience a decrease in selectivity at higher pressures, as larger pores are less selective and become accessible as the gas pressure increases. This phenomenon is apparent when comparing the plots with those presented in Figure 5.16.

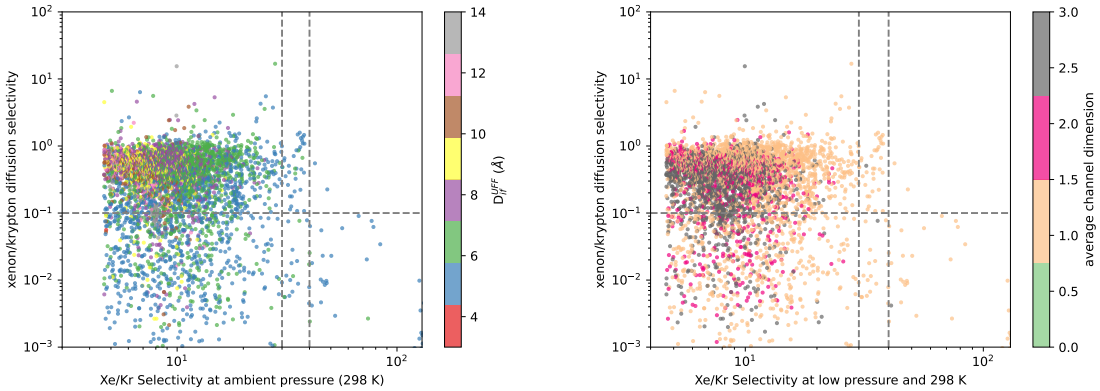


Figure 5.16: The xenon/krypton diffusion selectivity plotted against the ambient-pressure selectivity for a 20:80 Xe/Kr composition and color-coded by the LCD within a diffusion path D_{if} (left panel) and by the average dimension of channels (right panel).

At higher pressure, a shift towards lower selectivity values is observed in some materials. It is noteworthy that only 2 structures exhibit an ambient-pressure selectivity exceeding 40: the MOFs with the CSD codes XUNSOQ[**Abrahams_2014**] and GUMDEZ[**Yin_2014**] (Table 5.1). Most of these materials demonstrate a high cavity size, with only structures having an LCD near 6 Å remaining in this area of the plot. Furthermore, the channel dimension is also equal to one, providing a glimpse into the characteristics of these intriguing materials. They are made of unidimensional channels with small pore sizes, enabling the preservation of selectivity even under higher pressure conditions. Expanding the scope to structures with a selectivity higher than 30 (instead of 40), a total of 38 structures share similar features, including relatively low pore sizes and low channel dimensionality. Some of these structures, like QOZDOY,[**Zhang_2001**] have managed to maintain their selectivity to a certain extent, despite not being detected during the pre-screening based on low-pressure selectivity (Figure 5.15). However, other structures, such as the MOF MISQIQ,[**Tong_2013**] have experienced a significant drop in selectivity values from infinite dilution to ambient pressure (see Table 5.1).

When considering ambient-pressure selectivity, the large majority of highly selective materials actually have relatively low diffusion selectivity values (lower than 0.1), as shown in Figure 5.16 (this was not the case for low-pressure selectivity). This result suggests the necessity of a

trade-off between adsorption selectivity and diffusion selectivity. In the screening approach undertaken in this thesis, the decision was made to lower the adsorption diffusivity to approximately 40 to attain higher diffusion selectivity values. This choice was motivated by the fact that previous literature screenings[Simon_2015, Chung_2019] and the author's own published work [**Ren_2021**] solely focused on maximizing adsorption selectivity — this corresponds to working on the lower right side of the plots in Figure 5.16. To improve upon the previous approach, a kinetic constraint was included in the screening process. An alternative approach consists in optimizing the permselectivity, also known as membrane selectivity (equation 5.10). However, this would address a different application, namely membrane separation, which is extensively studied in the literature.[Anderson_2017, Wang_2022] In the presented screening, the objective was to identify thermodynamically selective materials that are not limited by diffusion. Several interesting materials were identified and will be further examined in the following subsection.

Identification of interesting materials

By cross-referencing the transport data with the thermodynamic data, it is possible to optimize the Xe/Kr adsorption selectivity while imposing a constraint on the diffusion selectivity to ensure it falls within an acceptable range (above 0.1). The structures of the 65 materials exhibiting a low-pressure Xe/Kr selectivity higher than 40 or an ambient-pressure Xe/Kr selectivity higher than 30 were manually visualized and briefly analyzed. Different materials were hand-picked for further analysis based on their unique characteristics. Materials with dissimilar types of channels that can artificially yield high diffusion coefficients were discarded. This phenomenon arises due to the randomness of the initial conditions. For instance, when xenon diffuses in a wider channel while krypton diffuses in a narrower channel, the diffusion selectivity will inevitably be artificially higher. One example is the MOF with a CSD code `OQESAF`,[Xie_2011] which was affected by this phenomenon, as shown in Figure 5.17, where different diffusion coefficient values are observed depending on the channel considered (in this case, a Henry coefficient weighted average needs to be performed). Other materials exhibit moderately high diffusion selectivity values ($\lesssim 1$) and typically consist of unidimensional channels that allow relatively free diffusion of xenon (higher than 4.6 Å) with varying cavity sizes. Multiple factors appear to influence the diffusion coefficients, including the values of channel size and pore size, but notably, the shape of the channel composed of cavities connected by narrower walls also plays a crucial role. The tortuosity of the layout and the relative difference between the cavities and the connecting channels can lead to significant variations in diffusion properties.

In this section, a detailed analysis of the comparative transport and adsorption performances of selected representative structures (Table 5.1) will be conducted to gain a better understanding of the key factors contributing to the observed differences in performance. This work can be used to design more quantitative characteristics that explain better transport performance, similar to the approach employed for the thermodynamic screening, which resulted in the identification of essential thermodynamic descriptors used in the design of an ML model for adsorption selectivity prediction (chapters 3–4). To achieve this, a visualization tool based on the grid calculation principle discussed in the dedicated section 3.3 will be employed, and the corresponding code is available in the same Github repository: `github.com/coudertlab/GrAED`.

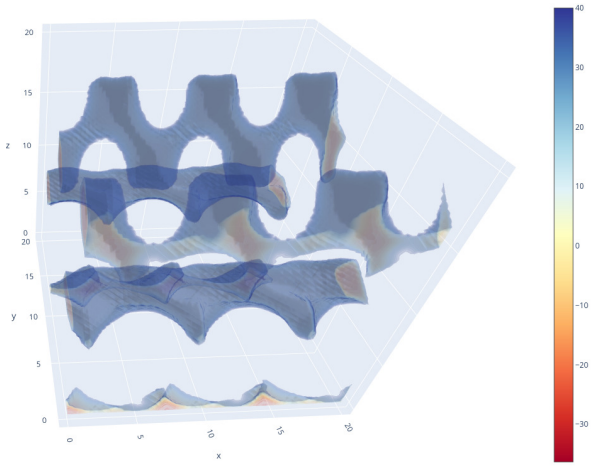


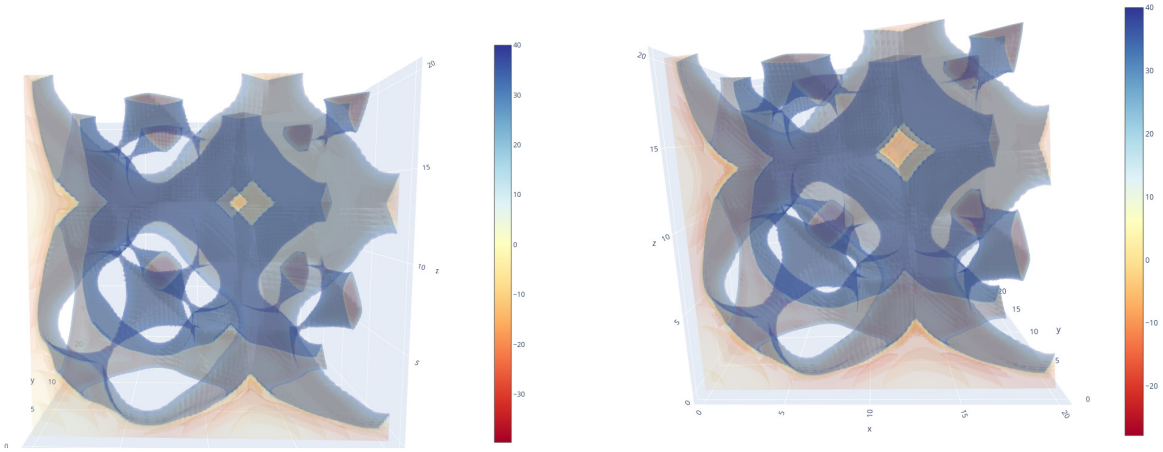
Figure 5.17: Snapshot of a 3D visualization of the xenon interaction energy inside the channels of the OQESAF[Xie_2011] *material. Two distinct unidimensional channels can be observed in the visualization. In an MD simulation of a single xenon per box, in this study, all possible initial positions were not tested out.*

Structure CSD ref. code	$s_0^{Xe/Kr}$	$s_1^{Xe/Kr}$	Pore size D_{if}^{UFF} (Å)	Channel size PLD^{UFF} (Å)	$s_{diff}^{Xe/Kr}$	Diffusion Coeff. D_{diff}^{Xe} ($cm^2\ s^{-1}$)	Xe uptake ($mmol\ g^{-1}$)
OQESAF [**Xie_2011**]	28	28	5.8	5.0	17	4×10^{-5}	3.2
ADOGEH [**Peikert_2012**]	49	10	12.9	5.3	15.5	5×10^{-5}	1.7
KAXQIL [**Banerjee2012**]	104	133	5.2	4.1	0.005	3×10^{-8}	1.4
XUNSOQ [**Abrahams_2014**]	38	48	5.6	4.8	0.23	7×10^{-6}	3.5
BAEDTA01 [**Chen_2010**]	152	38	5.7	4.6	0.4	4×10^{-5}	1.1
TONBII [**Du_2010**]	44	35	5.1	4.8	0.86	1×10^{-4}	1.5
VOHQIS [**Wragg_2001**]	51	48	5.7	3.9	0.01	6×10^{-8}	2.6
QOZDOY [**Zhang_2001**]	52	37	5.6	5.0	0.45	7×10^{-5}	3.7
GUMDEZ [**Yin_2014**]	56	42	5.5	5.1	0.55	7×10^{-5}	3.0
MISQIQ [**Tong_2013**]	140	37	4.6	4.5	1.4	2×10^{-4}	2.3

Table 5.1: Transport and thermodynamic performances of top-performing structures of CoRE MOF 2019 screened out by the approach developed in the section 5.2.3. The thermodynamic properties are determined using xenon uptake at 1 bar *and* 298 K, $s_0^{Xe/Kr}$ *and* $s_1^{Xe/Kr}$ *that correspond to the xenon/krypton adsorption selectivity values respectively at infinite dilution and ambient pressure condition. The pore size is defined as the largest cavity along a free diffusion path* D_{if}^{UFF} *and the channel size is defined using the pore limiting diameter* PLD^{UFF} *using atom radii defined by the UFF. The transport properties are evaluated using the xenon/krypton diffusion selectivity* $s_{diff}^{Xe/Kr}$ *and the xenon diffusion coefficient* D_{diff}^{Xe} *calculated by the MD-based screening presented above.*

The structure ADOGEH,[Peikert_2012] an amino-substituted version of the well-known HKUST-1 or $Cu_3(btc)_2$ (btc = 1,3,5-benzenetricarboxylate), was not found when comparing the transport data with the ambient-pressure selectivity values and the infinite dilution selectivity values, which explains that the selectivity s_1 for this structure is relatively low (10) compared to other materials (over 35). However, ADOGEH was detected when considering the selectivity s_0 at infinite dilution

due to its exceptional diffusion selectivity (around 10). This suggests that as a membrane material, ADOGEH could exhibit a selectivity of approximately 100, one of the highest values observed. Even as an adsorption-based separation material, it demonstrates an outstanding low-pressure selectivity of 49 coupled with its high diffusion selectivity, making it suitable for certain applications involving very low partial pressures of xenon and krypton.



(a) Xe energy grid in ADOGEH *(b) Kr energy grid in* ADOGEH

Figure 5.18: 3D volume plot of the xenon (a) and krypton (b) interaction energy values inside the material ADOGEH$^{\textbf{Peikert_2012}}$ *calculated using an energy grid as described in the section 3.3.*

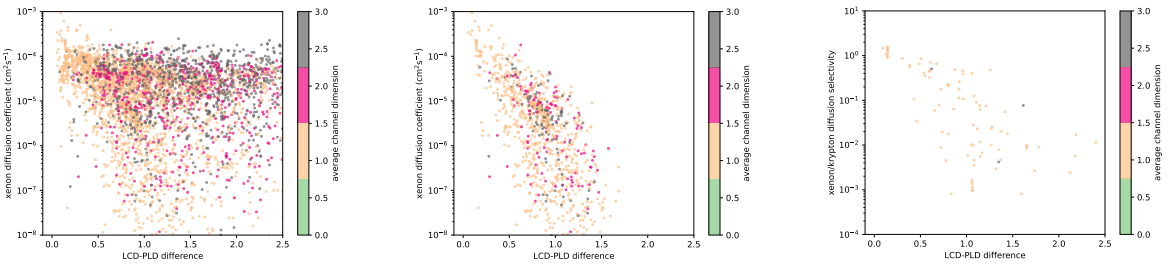
Even when used as an experimental material, the diffusion properties of xenon and krypton in this material are of significant interest in themselves. It was observed that only two materials displayed a diffusion selectivity over 10 in Figure 5.15. However, the other material has an artificially high diffusion selectivity due to the above-mentioned randomness of the initial position in the MD simulation and the presence of two types of channels (refer to Figure 5.17). Among all the screened materials for diffusion performance, ADOGEH stands out as the material with the highest diffusion selectivity. In a unidimensional system, it is more natural to expect a higher or equivalent diffusion coefficient for krypton compared to xenon due to their significant size difference.

This noteworthy behavior of adsorption ADOGEH can be explained by a special mechanism occurring in its tridimensional channel network. As depicted in Figure 5.18, both xenon and krypton have access to all dimensions for diffusion through the channels in the three directions of space. However, when examining the type of "pocket" that connects the channels diagonally, it becomes apparent that the access differs when comparing the two 3D energy grid plots. This pocket can be accessed by a xenon or a krypton atom even if the energy barrier to cross is relatively high. Figure 5.18a shows that the connection is narrower for xenon compared to krypton at the same energy threshold, which implies a higher energy barrier for xenon to access the "pocket" compared to krypton. This discrepancy in energy barrier explains the unusual difference in diffusion coefficients between xenon and krypton since krypton has a greater number of diffusion directions of space available than xenon, increasing the probability of turning around, which slows krypton down on the long run. In other words, xenon can diffuse in the 3D space using only three main directions, while krypton deviates from the main channel axes. Moreover, when krypton takes the small channel towards the "pocket", it experiences a

non-negligible residence time inside, further slowing down its diffusion compared to xenon. These "pockets" can be considered as traps for krypton in the nanoporous material, creating a competition between the two adsorbates.

Beyond the specific cases of OQESAF and ADOGEH, other nanoporous materials exhibit lower diffusion selectivity values. For instance, all the other materials listed in Table 5.1 have diffusion selectivity values ranging between 0.2 and 1.4. The diffusion selectivity and xenon diffusion coefficient vary depending on the shape and size distribution of the porous channels. A weak correlation can be observed between the pore size characteristics (LCD^{UFF}–PLD^{UFF}) and diffusion performance for structures with an LCD^{UFF} value below 6 Å — in these structures, pore size has a higher chance of influencing the transport properties. The correlation arises from the fact that a higher difference between LCD and PLD corresponds to a higher energy barrier for xenon to move within the channels, consequently decreasing the diffusion coefficient. However, when considering all available structures, the correlation disappears since the movement of diffusing xenon is less influenced by the pore walls in materials with higher LCD values. Moreover, for LCD values higher than 7 Å, the diffusion coefficient stabilizes around 10^{-5}cm^2 s^{-1}, as depicted in Figure5.13.

Structures with high ambient-pressure selectivity (Figure 5.19c) exhibit a negative linear relationship between the LCD-PLD difference and xenon/krypton diffusion selectivity. This is due to the fact that highly selective materials have pore sizes close to that of xenon atom, as explained in previous chapters. The effect on the Xe diffusion coefficient can also be extended to the Xe/Kr diffusion selectivity, as suggested by the noisy and stable values of krypton diffusion coefficients around 3×10^{-5}cm^2 s^{-1} (refer to Figure5.14). The weakness of the correlation can be explained by the inherent uncertainty in the MD methodology for diffusion coefficient calculation (estimated to be around 20% for the material KAXQIL), as well as other phenomena not accounted for by the simple arithmetic difference of two pore characteristics. The tortuosity of the channel, for instance, could contribute to the comprehension of diffusion coefficient values, but it is challenging to quantify in a tridimensional space. While various channel shapes will be qualitatively discussed in the following examples, the current work does not make an attempt to quantify these effects.

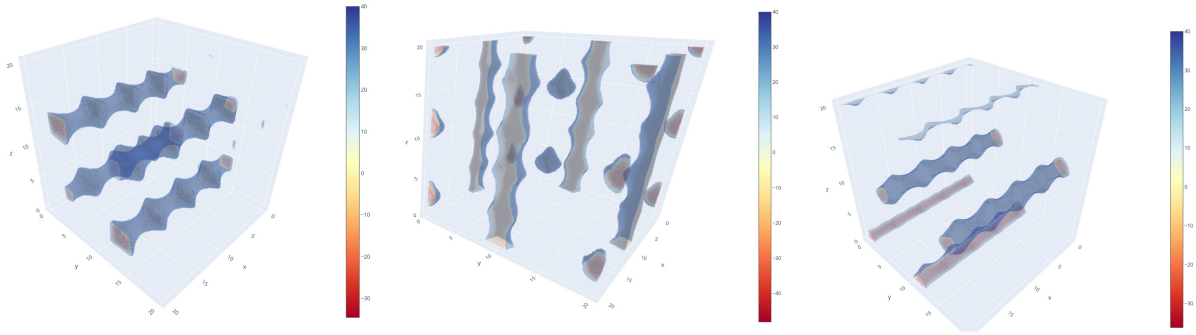


(a) All structures *(b) $LCD^{UFF} < 5.5$* *(c) $s_1^{Xe/Kr} > 30$*

Figure 5.19: Scatterplots of the diffusion coefficient compared to the LCD-PLD difference labeled using the channel dimension for all structures (a) and for structures with an LCD above 5.5 Å (b). On the subfigure (c), the scatterplot of the xenon/krypton diffusion selectivity compared to the LCD-PLD difference for the most selective structures ($s_1^{Xe/Kr} > 30$).

The highly selective materials displayed in Figure 5.19c are presented in Table 5.1. The negative linear relation between the LCD-PLD difference and the Xe/Kr diffusion selectivity or the Xe diffusion coefficient can be reaffirmed by examining the values in Table 5.1 for materials with 1D channels (except ADOGEH). Two categories will be created to explore the tortuosity difference between these materials.

Materials with nanopores composed of pseudo-spherical cavities connected by cylindrical channels following a straight line are the first category, as shown in Figure 5.20. These channels would in fact have very low tortuosity if evaluated. For TONBII, the very small LCD-PLD difference explains the relatively high diffusion selectivity near 1. There is hardly any difference in the diffusion of xenon and krypton in the channels of this material. As the LCD value increases for similar PLD values, the diffusion selectivity of materials like BAEDTA01 and XUNSOQ decreases, as indicated in Table 5.1. This drop can be attributed to a lower xenon diffusion coefficient.



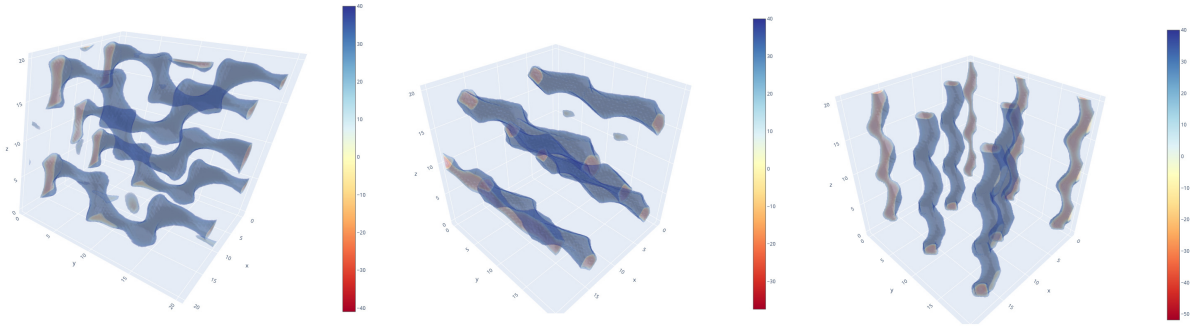
(a) XUNSOQ *[**Abrahams_2014**]* *(b)* BAEDTA01 *[**Chen_2010**]* *(c)* TONBII *[**Du_2010**]*

Figure 5.20: 3D volume plot of the xenon interaction energy values inside materials with non-tortuous unidimensional channels calculated using an energy grid, as previously described.

Beyond the consideration of the pure diffusion properties, the relatively high adsorption selectivity coupled with minimal diffusion limitations make these materials intriguing for further study. The material BAEDTA01, which was previously discussed in relation to the selectivity drop caused by changes in pressure conditions, reveals the microscopic origins of this drop in Figure 5.20b. The two distinct adsorption sites (one narrower than the other) can be clearly observed — with the narrower site contributing to the extremely high selectivity at low pressures. The other materials, TONBII and XUNSOQ, maintain a relatively stable selectivity between low and ambient pressure cases. Compared to the KAXQIL structure identified in a previous high-throughput screening,**Simon_2015** these materials exhibit higher diffusion coefficients and resolve the potential issue of diffusion limitation. However, the selectivity values of these materials are yet to be confirmed. While the value of the PLD is the primary factor explaining the lower diffusion coefficient of KAXQIL, the LCD-PLD difference can serve as a secondary variable to distinguish materials with similar PLD values but different diffusion coefficients.

Other materials, as seen in Figure 5.21, consist of channels that are much more tortuous than the previous type. They exhibit a "zigzag-like" shape. Quantifying the effect of tortuosity on the diffusion coefficient is challenging with the limited data available at this stage — to achieve this, a comparison of highly similar materials (same chemical nature, same pore size) with differing

tortuosity would be necessary. A theoretical perspective suggests that tortuosity typically has a negative impact on diffusion coefficients. For example, VOHQIS displays high degree of tortuosity, as shown in Figure 5.21a, and its diffusion properties are not particularly high. However, it is difficult to disentangle the effects of pore size (significant difference between LCD and PLD) from the impact of tortuosity. A more quantitative approach is required to gain deeper insight into the diffusion process in nanoporous materials.



(a) VOHQIS *[**Wragg_2001**]* *(b)* GUMDEZ *[**Yin_2014**]* *(c)* MISQIQ *[**Tong_2013**]*

Figure 5.21: 3D volume plot of the xenon interaction energy values inside materials with tortuous unidimensional channels calculated using an energy grid as previously described.

A similar analysis to the previous materials with straight channels reveals similar findings for these more tortuous ones, which is not surprising as these materials are included in the correlation plot in Figure 5.19c. GUMDEZ is much less tortuous, as shown in Figure 5.21b, but the difference in pore size is also smaller, resulting in a higher diffusion coefficient, which explains the good diffusion selectivity. This material also has significantly higher xenon uptake compared to more selective materials like KAXQIL, reflecting the uptake–selectivity as illustrated in Figure 2.5. Therefore, GUMDEZ and QOZDOY were previously identified in the literature by Chung et al.[**Chung_2019**] during the xenon/krypton separation screening conducted while introducing the CoRE MOF 2019 database. These materials were noted for their high xenon/krypton selectivity and substantially higher xenon uptake, which is a key metric for industrial separation processes — it typically determines the amount of xenon retrieved per adsorption–desorption cycle. To improve the screening process, optimization of xenon uptake could be added alongside optimization of diffusion and selectivity properties. Fortuitously, this study identified materials such as XUNSOQ, QOZDOY, and GUMDEZ with good adsorption selectivity, diffusion selectivity, and xenon uptake that could be much more versatile than highly specialized materials like KAXQIL.

In summary, a screening of diffusion properties for xenon and krypton was conducted to complement the previous thermodynamic properties screening. Materials with a balanced combination of diffusion and adsorption selectivity values were identified, some of which exhibited very high Xe uptake, potentially enhancing the productivity of xenon separation processes by increasing separative capacity and facilitating rapid gas penetration within the material. This study further justifies the multivariate nature of the optimization problem when searching for suitable materials for xenon/krypton separation — relying solely on a single variable is insufficient. The study provides a more comprehensive approach compared to existing studies (on other systems), highlighting the significance of kinetic effects in adsorption processes.[**Stanton_2022**] Further investigation is required to gain a better understanding of the

relationship between diffusion properties, tortuosity, and each pore size effect. The next section will be dedicated to the development of faster methodologies for screening transport properties to scale up the screening process for larger databases. To achieve faster transport property screening, methods based on the transition state theory and machine learning prediction models were explored.

5.3 Fast diffusion calculation algorithm

To overcome the high computational demands of MD simulations, alternative methods were developed during this thesis to calculate the diffusion coefficient. One such method involves utilizing the transition state theory to generate MSD at larger timescales more efficiently. By applying a similar algorithm as in TuTraST, it becomes possible to address the initialization problem encountered in MD simulations when dealing with different channels in an automated screening process — while it is feasible to manually place a particle in a specific initial state, achieving this within a screening process can be challenging. Although the implementation of the C++ algorithm that directly reproduces diffusion coefficients through a rejection-free lattice kinetic Monte Carlo algorithm is not yet complete, the initial implementation was employed to calculate maximum energy barriers within a material. These energy values provide additional information alongside PLD values for predicting diffusion coefficients, and a ML model was trained to predict diffusion coefficients much faster than the current MD method.

5.3.1 Code based on the TuTraST algorithm

The GrAED algorithm, as presented in section 3.3, was employed to calculate the xenon interaction energy with the material at each non-overlapping point of the symmetry-aware grid. This energy grid enables the identification of different channels, adsorption pores, and the transition surface that separates them.

A slight modification was made to the approach previously described in section 5.1.2 and adopted by Mace et al., to identify the three main components of the lattice Monte Carlo approach. Rather than detecting the channels on the fly, the decision was made to pre-detect the channels to restrict the cluster growth to a specific channel. This approach aims to reduce the computation time required during the clustering step by simulating only one representative channel out of all the potentially equivalent channels — due to the high level of symmetry, multiple equivalent channels typically exist within a single unit cell, as illustrated in Figures 5.20 and 5.21.

To identify the various channels, a breadth-first search algorithm was utilized to find all connected grid points with an energy below a specified threshold value (E_{cutoff}). The determination of connections between rhomboidal grid voxels (same angles as the unit cell) was based on voxel faces, resulting in six nearest neighbors. Additional connections from the eight edges were considered, totaling 14 nearest neighbors. Including the vertices would yield 26 nearest neighbors. However, for simplicity, only the six primary connections were employed. The breadth-first search algorithm for a grid system is relatively straightforward:

All the points of the grid are iterated over:

1. If the point has not been visited and its energy is below the threshold, it is added to the cluster and a queue. A search can be initiated to identify all connected neighbors.
2. Each (face-connected) neighbor of the point is tested and added to the cluster and a queue if it has not been visited and its energy is below the threshold.
3. The process is repeated for all elements in the queue until the queue becomes empty. This yields all grid points connected to the initial point at the end. The main loop then restarts, and the search is only initiated if an unvisited point is encountered.

The breadth-first search ends with clusters of connected points that are below the energy threshold. Each of these clusters can be tested to see if they represent channels (connected all the way through a periodic boundary).

Having identified well-defined channels and pockets within the nanoporous material, symmetrically equivalent channels can be identified using the symmetric grid explained in section 3.3. Typically, only a few unique channels (usually fewer than three) remain, which can be used for basin-cluster growth and the detection of TS surfaces that separate these basins.

Next, the algorithm proceeds by iterating over energy values (with a step size of $1\,\mathrm{kJ\,mol^{-1}}$ used in the original paper) and growing the cluster layer by layer. An improvement is introduced at this stage, wherein the previous search algorithm is employed to efficiently count the number of clusters within a given channel. If this number changes, it indicates that some clusters have merged. When the energy gap is sufficiently large, TS surfaces can be detected using a layer-by-layer growth approach within a small energy range ($[\mathrm{E},\ \mathrm{E} + \delta\mathrm{E}]$) to generate a smoother surface.

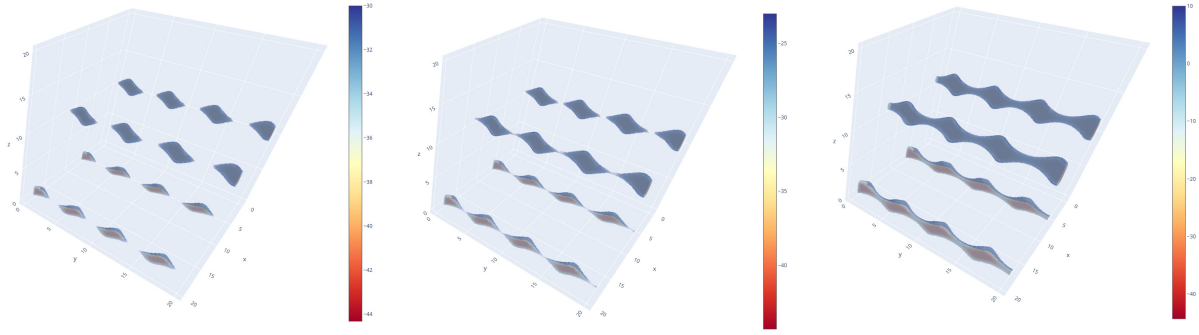
The current development of this detection algorithm is at its final stage, as it was temporarily paused to focus on analyzing the barrier energies calculated from a modified version of the initial implementation, which will be detailed in the following discussions. The detection of transition surfaces will be further explored in the following chapter, with an attempt to implement a more optimized version that avoids the computationally expensive layer-by-layer growth.

5.3.2 Calculation of a diffusion activation energy

As the aim of this study is to only determine the activation energy, the more computationally demanding TS detection and kinetic Monte Carlo steps can be skipped. Instead, the breadth-first search algorithm is employed to label different connected components within a given channel between $\mathrm{E_{min}}$ and $\mathrm{E_{min}} + i\delta\mathrm{E}$ (at the i^{th} iteration). By monitoring changes in the number of connected components between two energy values, the code automatically detects the energy $\mathrm{E_{trans}}$ at which components reconnect and form a channel (allowing diffusion from one boundary to another). The activation energy $\mathrm{E_a}$ is then calculated as the difference between the calculated transition state energy $\mathrm{E_{trans}}$ and the minimum energy $\mathrm{E_{min}}$ within the channel.

In the case of `KAXQIL`, barrier detection was performed using an energy step $\delta\mathrm{E}$ of $0.3\,\mathrm{kJ\,mol^{-1}}$. A single symmetrically unique type of channel was identified in `KAXQIL`, with a minimal energy of $-44.3\,\mathrm{kJ\,mol^{-1}}$ — the various channels shown in Figure 5.22c are all symmetrically equivalent. The code detected a single merge that resulted in a fully connected component within the

channel. This merging occurred at an energy of −25.7 kJ mol^{-1} (as depicted in Figure 5.22b), indicating that the estimated activation energy is 18.6 kJ mol^{-1} with an error of 0.3 kJ mol^{-1} (due to the energy step used).



(a) $E_{max} < -30$ kJ mol^{-1} *(b)* $E_{max} < -23$ kJ mol^{-1} *(c)* $E_{max} < 10$ kJ mol^{-1}

Figure 5.22: 3D visualization of channels within KAXQIL *using different energy thresholds* E_{max}*. Depending on the maximum value of energy allowed, the channel is either composed of unconnected basins (a), or they are fully connected (b) and (c). This illustrates the principle of the energy barrier detection.*

In this case of one unique merge of a unidimensional channel, the method demonstrates strong performance, and it becomes possible to associate the activation energy with a diffusion rate k_{diff} using the Arrhenius equation:

$$k_{\text{diff}} = A \exp\left(-\frac{E_a}{k_B T}\right) \tag{5.11}$$

where A is a prefactor that depends on the temperature and system (adsorbate, adsorbent). This is a simplified version of the equations 5.6 and 5.7 used in transition state theory-based methods. In the case of a unidimensional channel with a single possible transition, the diffusion coefficient is directly associated with the diffusion rate. The problem can be reduced to a unidimensional random walk with a given transition probability, and the diffusion coefficient is given by $D = 0.5 k_{\text{diff}} L^2$ where L is the distance between two basins (in one dimension). In this special case, there exists a direct relationship between the diffusion coefficient and the activation energy such that $\log(D) \propto E_a$. For more complex systems than KAXQIL, these methods may not yield satisfactory results.

Describing the case of multistep diffusion is particularly challenging. For example, a particle can cross a series of lower barriers instead of encountering the highest energy gap as calculated by the method of this work, as illustrated in Figure 5.23. In such cases, the relevant activation energy is the maximum value among these two activation energies. Even when considering the maximum activation energy, if the values are similar, this approximation may not be justified. Both transitions would influence the diffusion process. This approximation holds true only when one of the activation energies is significantly larger than the other ($E_a^1 \gg E_a^2$).

To improve this approach, intermediate transition state energy values can be detected by examining the change in the number of clusters, for instance. However, it remains difficult to determine which combinations of energy differences are relevant, and a more detailed

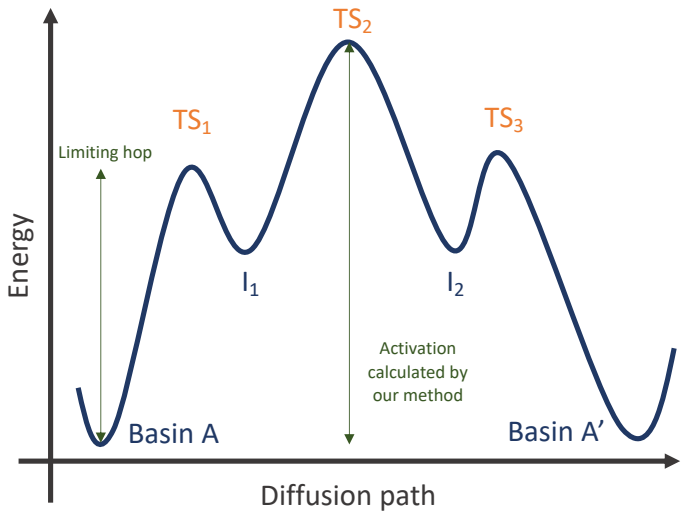


Figure 5.23: Multistep diffusion from a basin A to a basin A'. The diffusion process is modeled by transition states TS_1, TS_2 and TS_3 and intermediate steps I_1 and I_2. In this particular case, there is a difference between the real limiting activation energy and the activation energy calculated by the simplified method.

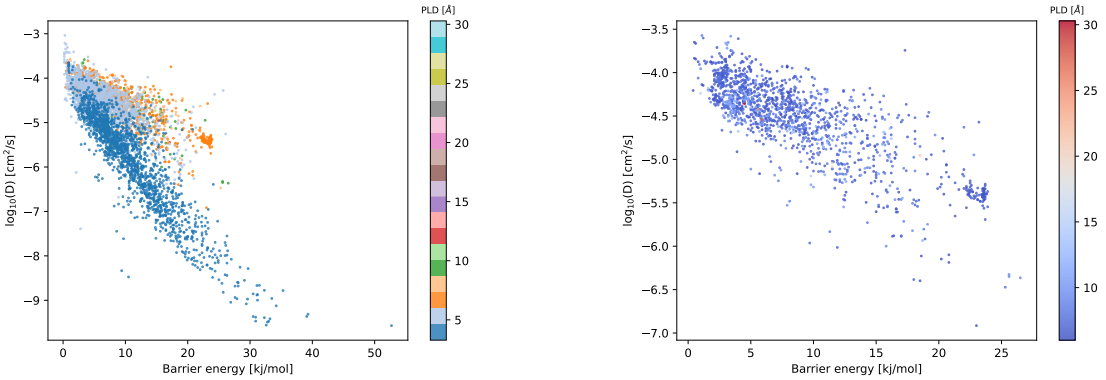
investigation of the location of these transition states is required, which brings the discussion back to the initial issue of TS surface detection that has remained on standby. Despite these limitations, this quickly measurable activation energy can be employed as a proxy for the diffusion coefficient. The subsequent discussion will focus on the relationship between this approximated activation energy value and the diffusion coefficients. This new diffusion descriptor can later be incorporated into prediction models as a complementary feature to the PLD, providing a more comprehensive picture of the diffusion process.

5.3.3 Relation of this activation energy to the diffusion

The xenon diffusion activation energy was calculated for all the 5,125 structures selected for the xenon diffusion coefficient screening presented in section 5.2.2. An energy step of δE of 0.1 kJ mol^{-1} was employed during the energy loop to determine the minimal energy barrier for each unique channel in the material. Subsequently, an activation energy can be derived and compared to the diffusion coefficients. To avoid any potential noise arising from the MD simulation initialization problem, materials with significantly different energy barrier values from one channel to another (standard deviation of energy barrier values higher than 1 kJ mol^{-1}) were excluded. These materials accounted for only 145 out of the total 5,125 materials.

As shown in Figure 5.24a, the activation energy is correlated with the diffusion coefficient for xenon. A stronger correlation is observed for points with a PLD around 4.5 Å, while for PLD values exceeding 6 Å, the correlation appears to be weaker compared to smaller PLD values, as illustrated in Figure 5.24b

This correlation between the energy barrier and the diffusion coefficient is confirmed in Figure 5.25. The points are labeled according to their energy barrier value, and the highest energy barrier points tend to be concentrated among lower diffusion coefficient values. However, a few points with very high energy barriers are also observed for diffusion coefficients that are quite low. There are two such points detectable with the naked eye.



(a) All structures *(b) PLD over 6 Å*

Figure 5.24: Scatterplots of the $\log_{10}$ *of the diffusion coefficient (in* $\mathrm{cm}^2\,\mathrm{s}^{-1}$*) compared to the diffusion activation energy* E_a *in* $\mathrm{kJ\,mol}^{-1}$ *(a) for all structures and (b) for the structures with a PLD above 6 Å. For all structures, the Pearson correlation coefficient is equal to −0.77, whereas for the restriction to structures with a PLD below 6 Å this correlation is stronger with a Pearson coefficient of −0.85. For structures with a PLD above 6 Å, this coefficient decreases to reach −0.74.*

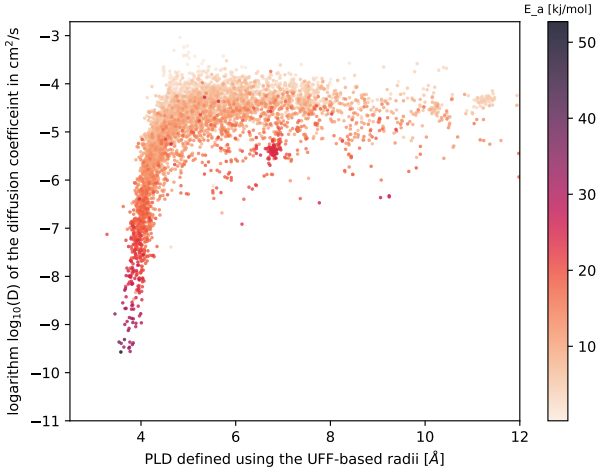


Figure 5.25: Scatterplot of the $\log_{10}$ *of the diffusion coefficient in* $\mathrm{cm}^2\,\mathrm{s}^{-1}$ *as a function of the PLD values and labeled by the barrier activation energy. The higher barriers seem to correspond to lower diffusion coefficients, thus echoing the correlation observed in the previous figure 5.24.*

This barrier activation energy descriptor completes the description of the diffusion coefficient given by PLD values. As discussed in the dedicated section, PLD values cannot distinguish between structures over 6 Å in the "plateau", and the difference in diffusion coefficient values was considered as noise in the previous analysis. However, Figure 5.25 reveals that higher values of barrier energies are associated with lower diffusion coefficients within the plateau range, thereby explaining the variations in diffusion coefficient within the plateau based on the activation barrier values. Although the correlation is not perfect, this barrier descriptor provides better insights into this uncharted area of PLD values above 6 Å, which cannot be explained by simple geometric considerations. The barrier activation energy value sheds light on the chemical nature of the diffusion barrier that needs to be overcome.

In the final section of this chapter, the combination of geometrical descriptors with this energy barrier will be employed to train a machine learning model, following the same approach as in the previous Chapter 4. The energy barrier and PLD values, the highest correlated descriptors, will play a prominent role in the final ML model. This model can then be utilized to evaluate the diffusion coefficient of xenon in other materials, offering a significantly faster alternative to MD simulations.

5.3.4 ML prediction model

Calculating diffusion coefficients is an extremely time-consuming process, complicated by various challenges in the final fitting procedure. Out of the initial 6,525 structures, over a thousand were lost, resulting in a success rate of approximately 79%, mainly due to either insufficient time for obtaining a usable MSD or MSD that describe non-diffusional regimes. By utilizing an unconventionally higher time step, the time required to investigate the diffusion regime could be reduced to just a couple of days per structure. Compared to the 12 seconds required for energy barrier calculations with an energy step of 0.1 kJ mol^{-1}, and the few minutes required to run Zeo++, the MD method is extremely slow. Even under highly optimistic assumptions for MD simulations, the comparison is essentially between 24 hours and at most 10 minutes per structure, which corresponds to an approximate speedup of 150-fold (though, in reality, it is much higher). However, the relationships between energy barrier, PLD, and diffusion coefficient remain unclear — the Arrhenius law generalizes limited consideration of the weak correlation shown in Figure 5.24a. The aim of the ML model is to establish this relationship and achieve accurate predictions while significantly reducing the time required for predicting the diffusion coefficient of future selective materials.

The ML model was trained using 80% of the 4,873 structures that survived all the different imposed filters. A total of 12 descriptors described in Table5.2 were employed to train the model. The hyperparameters of the XGBoost model were determined using a similar approach as in Chapter 4, and the following values were utilized:

```
optimal_params = {
    'objective': 'reg:squarederror',
    'n_estimators': 1500,
    'max_depth': 4,
    'colsample_bytree': 1,
    'colsample_bylevel': 0.75,
    'subsample': 0.75,
    'alpha': 0.6,
    'lambda': 1,
    'learning_rate': 0.04,
  }
```

With this parameterization, this ML model predicts the $\log_{10}$ of the diffusion coefficient ($cm^2\ s^{-1}$) with a root mean square error of 0.26 on the test set and a mean absolute error of 0.18. This implies that the exponent α is known with an error of approximately 0.2 when expressing the diffusion coefficient as $D = 10^{\alpha}$. For comparison, the previous ML model for thermodynamic selectivity predicts the $\log_{10}$ of selectivity with an error of about 0.07. It is important to note that the goal here is not to predict the exact values of the diffusion coefficient due to the

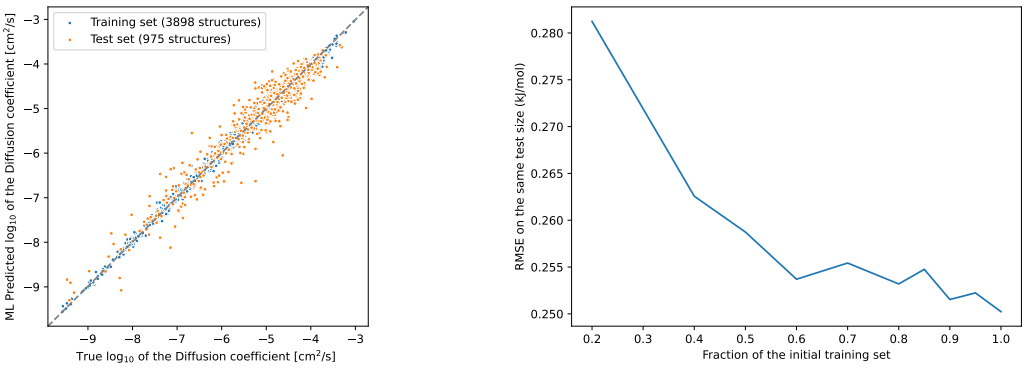
Feature name	Symbol	Description
"Framework Mass (g/mol)"	M_f	Molar mass of the framework material considered
"Framework Density (kg/m3̂)"	ρ_f	Mass density of the framework material considered
"ASA_m2/cm3_1.2"	SA	Surface area accessible to a 1.2 Å radius probe in $m^2\ cm^{-3}$
"PO_VF_2.0"	VF $\frac{V_{pore}}{V_{tot}}$	The void fraction or the ratio of the pore volume occupied by a 2 Å radius probe over the total material volume
"D_f_vdw_uff298"	PLD or D_f	Pore limiting diameter of the largest free sphere diameter calculated using the UFF dependent definition
"D_if_vdw_uff298"	LCD or D_{if}	The largest included free sphere diameter in a free diffusion path calculated using the UFF dependent definition
"Adsorption_enthalpy"	$\Delta_{ads}H_0^{Xe}$(channel)	Xenon adsorption enthalpy within a channel calculated using the barrier algorithm
"barrier_kjmol"	E_a	difference between transition state energy E_{trans} and the minimal energy E_a within a channel
"delta_LCD_PLD"	LCD−PLD	difference between the LCD and PLD values
"1D_chan"	$\mathbb{1}_{1D}$	categorical feature: 1 if there is a unidimensional channel, 0 else
"2D_chan"	$\mathbb{1}_{2D}$	categorical feature: 1 if there is a bidimensional channel, 0 else
"3D_chan"	$\mathbb{1}_{3D}$	categorical feature: 1 if there is a tridimensional channel, 0 else

Table 5.2: Features used in the ML model for diffusion coefficient prediction.

inherent noise the values generated by MD simulation (about 20% relative error for KAXQIL). Instead, the objective is to determine the order of magnitude of the diffusion coefficient. The proposed model achieves this objective effectively, as illustrated in Figure 5.26a, where the predicted diffusion coefficient aligns closely with the true values when represented on a log scale.

The training curve (Figure 5.26b) was examined to assess whether the model had sufficient training data or required additional data. As the amount of training data increased, the error converged to 0.25, indicating that no further data was necessary for training the model. However, it is conceivable to train a similar model using less data (50% instead of 80% of the total data could probably suffice to train a similar model).

The ML model was interpreted using the SHAP algorithms discussed in the previous chapter. As expected, the most important features were found to be the PLD and the barrier activation



(a) Prediction result *(b) Training curve*

Figure 5.26: (a) Comparison of the $\log_{10}$ *of the diffusion coefficient predicted by an ML model and the true values. (b) Root mean squared errors on the same test set (20% of all data) as a function of the fraction of the training set used to train smaller models. The error decreases as the amount of data increases and seems to stabilize near* 0.25.

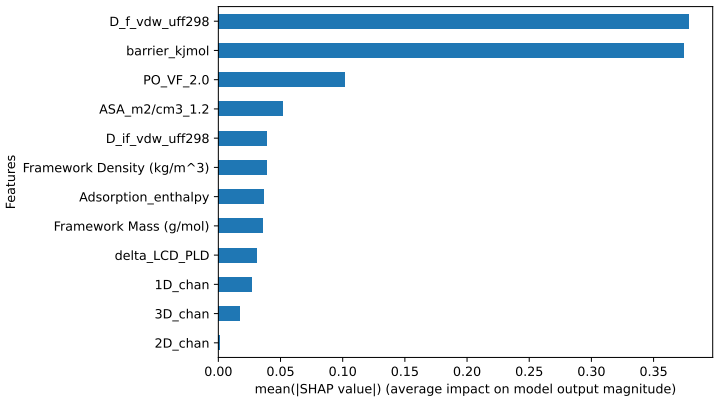


Figure 5.27: Feature importance determined using the average of the absolute Shapley values for each feature based on every training data. An influential feature would have a very high average absolute SHAP value. The features are detailed in Table 5.2.

energy, as demonstrated in the previous section. The void fraction also appeared to play a non-negligible role.

To unravel the relationship between these features and the target diffusion coefficient, partial dependence plots (PDPs) were examined for these features shown in Figure 5.28.

The PLD has a contribution similar to that described in section 5.2.2. A linear contribution was observed when the PLD values were below 6 Å, followed by a constant contribution for PLD values above this threshold. The activation energy showed a negative correlation with the log of the diffusion coefficient, which explained the linear contribution observed in the dependence plot.

The model also revealed less obvious contributions. Figure 5.11 indicates that no clear relationships can be inferred between surface areas or void fractions and the diffusion coefficient. These factors played a more secondary role, slightly adjusting the obtained values with contributions of the order of 0.2. For instance, the model identifies a positive relation between the

void fraction and the contribution to the diffusion coefficient, which aligns with the physical understanding that lower void fractions correspond to lower diffusion rates within the material, assuming other parameters are equal. Conversely, larger surface areas imply more interaction with the pore walls, which slows down the diffusion of particles. Regarding the LCD, the LCD-PLD difference, xenon adsorption enthalpy, framework's mass, and density, no clear contribution patterns were observed. This may be attributed the fact that the previous features account for a substantial portion of the contribution due to the correlation between all these features.

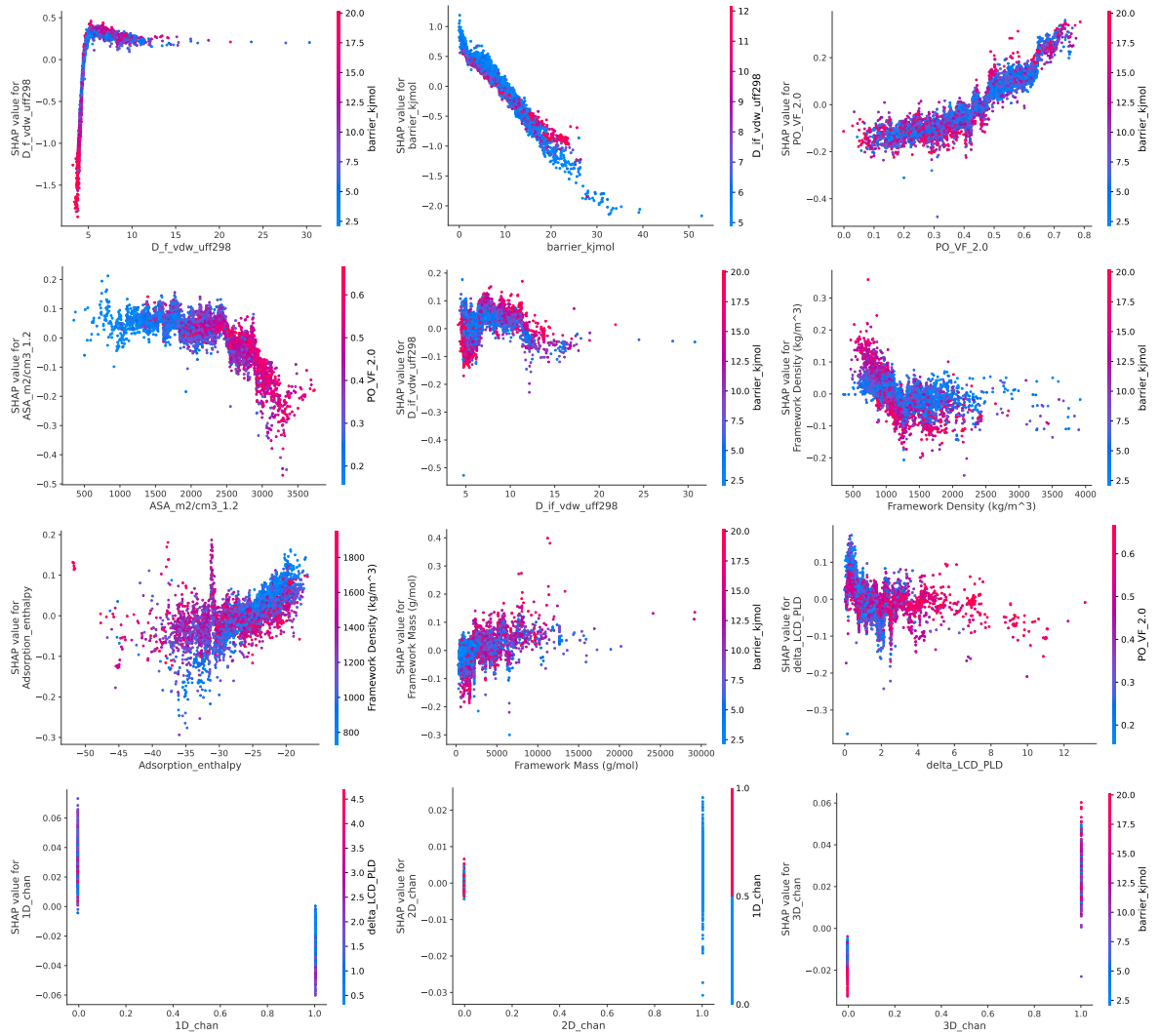


Figure 5.28: A SHAP dependence plot corresponds to the Shapley values as a function of the feature values for every structure. These SHAP plots show the contribution of the features to the prediction given by the ML model. Each Shapley value depends not only on the value of the feature itself but also on the other features. For this reason, the plots are labeled based on a relevant second feature. The partial dependence plots of every feature in the diffusion prediction model are presented here.

The final predicted values are only marginally influenced by the channel dimension, despite its association with a clear physical phenomenon. The behavior of diffusion coefficients varies depending on the dimensionality of the channel. Figure 5.28 illustrates that a 1D channel has a lower diffusion coefficient when all other features are similar. On the other hand, a 2D channel demonstrates a higher contribution, which is further confirmed by the partial dependence

plots. A tridimensional channel exhibits an even higher diffusion coefficient. The model can distinguish between different material types based on their channel dimension.

This section presented an ML-based approach for computing diffusion coefficient values using computationally cheaper energy descriptors combined with geometrical descriptors. This method is much more efficient than the conventional MD simulation, as it requires only one costly training session using MD simulations at the beginning. To further accelerate the process, alternative methods can be employed to generate diffusion coefficients, as demonstrated by Mace et al. in their work.[Mace_2019] This work is currently in progress, and it is expected that the future implementation will yield diffusion values comparable to those derived from MD simulations.

5.4 Beyond self-diffusion screenings

In this chapter, different methods have been introduced to evaluate transport properties of an adsorbate inside a nanoporous material. The most accurate method requires considerable computational time and meticulous attention to achieve optimal accuracy. For instance, careful parameter selection in MD simulations is essential to obtain relevant mean square displacement data for diffusion coefficient calculation. A screening of diffusion coefficient values for xenon and krypton has been performed to identify materials with notable thermodynamic and kinetic separation performance. These values have also served as baseline data for testing other methods such as activation energy detection and an ML model. The final ML model seems to show promising performance, achieving a root mean squared error of only 0.25 on the base-10 logarithm scale of the diffusion coefficient. This indicates the ability to accurately assess the order of magnitude of diffusion properties. Such assessments can help identify potential diffusion limitations in promising materials and optimize this property to expedite equilibration in adsorption-based separations. Furthermore, the techniques developed in this study, as well as future developments, can be applied to membrane separation processes.

The obtained results provide the foundation for various follow-up studies. For instance, the effect of tortuosity on diffusion coefficient values and relevant definitions for tortuosity remain open questions. Unidimensional channels can be particularly examined, where the frequency and magnitude of changes in direction can be analyzed to quantify their occurrence.[Bullitt_2003] Another challenge could consist in measuring different diffusion regimes, such as single-file diffusion characterized by a square root time relation in the mean square displacement (MSD).[Lin_2005] In this study, materials with MSD relations other than linear were excluded since only materials with high determination coefficients in the linear fit were considered.

To expand beyond conventional studies, the diffusion coefficient can be utilized to model breakthrough experiments, is the closest a lab experiment can get from the industrial adsorption process. The recent development of the RUPTURA software[Sharma_2023] opens new perspectives in modeling. For instance, the axial dispersion coefficient used in a breakthrough model can be calculated using transport properties, combined with thermodynamic data on the adsorption process of xenon and krypton. This presents an opportunity for experiment-theory comparison, fostering a virtuous feedback loop to improve modeling and facilitate the discovery of superior materials.

The diffusion coefficients calculated using the aforementioned methodologies solely describe self-diffusion in an infinitely diluted environment. To better describe transport properties in industrial conditions, it is necessary to study diffusion coefficients in a higher loading environment to account for host–host interactions. Furthermore, mixture simulations can be directly conducted to obtain the so-called Onsager diffusion coefficients, which are based on the Maxwell-Stefan diffusion equation rather than Fick's equation.[Krishna_2008] The calculation of such quantities requires significant computational resources, as MD simulations on mixtures at relatively high loading must be run for a sufficiently long duration to capture the diffusion regime. Therefore, applying this approach to large-scale screening is impractical, but some interesting materials can be tested to study the effects of mixtures and loading on transport properties.

This chapter presented a particular aspect that was not considered in standard high-throughput screenings for xenon/krypton separation, namely transport properties. The subsequent chapter will focus on other factors that can contribute to a more comprehensive picture, bringing it closer to experimental systems. The flexibility of the nanoporous framework, for instance, can impact adsorption performance.[Witman_2017] Additionally, the difference in polarizability between xenon and krypton can be better leveraged in the screening process if it can be modeled through the use of higher-level theories than the Lennard-Jones potential. Both the flexibility and polarization aspects are still under investigation, and although some results will be presented, the chapter mainly serves as a compilation of potential research focuses and solutions to enhance the current understanding.

6

TOWARDS THE NEXT GENERATION OF SCREENING

6.1 Limits of the current screening methodologies

As presented in the review of different methodologies for screening materials in Chapter 1, it is a common practice to screen for a specific metric, such as selectivity, permselectivity, or capacity, depending on the targeted application. Attempts to screen for materials that exhibit high selectivity while also possessing a good capacity are increasingly prevalent in current research.[Chung_2019, Zhang_2022, Solanki_2020] For instance, improvements can be made in selectivity screening regarding calculation efficiency and the accuracy of molecular description. The previous chapters primarily focused on enhancing efficiency by exploring various techniques for sampling adsorption energy and comparing their computational time and accuracy. Furthermore, the screening procedure was enhanced by incorporating transport properties. Additionally, alternative calculation strategies were explored and developed to increase the efficiency of screening diffusion coefficients (e.g., transition state detection, machine learning models) compared to computationally expensive conventional methods (MD simulations).

To address the limitations of the current methodologies for adsorption screening, a more accurate physical description of the nanoporous system is required. For example, the rigid nature of structures in most screening procedures can sometimes lead to misleading results, as materials may appear to have high selectivity, but exhibit decreased computed selectivity values

due to their flexible nature. Considering flexibility in the analysis can modify the rankings obtained from previous screenings and potentially identify other top materials. Another physical property that can significantly impact the screening results is polarization. In the case of adsorbates like xenon and krypton, the difference in polarizability plays a crucial role in the separability of these gases using adsorbent materials. A more precise characterization of this property has the potential to completely alter the screening outcomes. Notably, the best experimental materials often feature decorations with polar groups (Ref.[**Li_2019**]) or possess open metal sites (Ref.[**Pei_2022**]), although these criteria were not deemed essential in the current screenings.

This final chapter will explore prospective studies focusing on three main research areas: (i) the efficiency improvement of the calculation of transport properties, (ii) the adsorption calculations in flexible frameworks for screening purposes, and (iii) a more accurate description of polarization interactions in molecular simulations.

6.2 Future developments on transport properties

During the Ph.D. project, the transport properties were thoroughly studied using MD simulations and an ML model primarily based on the PLD and a proxy of the diffusion activation energy. This approach provides very promising results as it significantly accelerates the evaluation of diffusion coefficient values. However, when employing an ML-based approach, the generalizability to other types of systems cannot be guaranteed. To create a simulation that is faster than MD simulations while ensuring higher accuracy and reliability than the ML-based approach, the next steps involve completing the development of the diffusion coefficient calculation code based on the transition state theory (TST) and kinetic Monte Carlo (kMC) simulations. This code is currently in its final stages of development, as explained in the preceding chapter. Once this new approach is developed, it can provide diffusion coefficients that can be utilized in breakthrough modeling software for comparison with experimental data. The subsequent section presents a description of such software, exploring the perspectives offered by RUPTURA.[**Sharma_2023**]

6.2.1 Final development of the optimized version of TuTraST

The diffusion calculation code, based on the TST and kMC, already possesses several capabilities: (i) calculating the energy grid using the GrAED algorithm, (ii) identifying connected components or clusters through a breadth-first search algorithm, (iii) detecting channels using a simple all-direction scanning algorithm on the identified clusters, and (iv) determining the energy barrier by utilizing (ii) and (iii) in a loop over the energy values. The energy barrier of a particular channel is defined as the energy at which the channel reconnects with at least one channel connected through at least one direction.

To finalize the implementation of the algorithm, the final step entails detecting the transition state surfaces that separate different clusters. This final mapping, which establishes basins connected by transition surfaces, can then be used in a simple kMC scheme to determine the diffusion coefficient. In the original work by Mace et al.,[**Mace_2019**] the authors achieved the growth of clusters with energy values below $E_{min} + i\delta E$ by incrementally expanding them layer by layer until they reached energy values below $E_{min} + (i + 1)\delta E$. When a point from a layer

of one cluster touches another cluster, that point can be considered a transition point if the energy gap is sufficiently high. Otherwise, the two clusters merge to form a single cluster. The loop over the energy values is employed to restrict the transition points to the range between $E_{min} + (i)\delta E$ and $E_{min} + (i+1)\delta E$. Upon reflection, the layer-by-layer growth method does not identify the highest energy point in a given direction, as conventionally defined for transition states. Instead, it detects points that are equidistant from the surfaces of two previous clusters. Both definitions are equivalent if the value of δE is infinitely small.

To avoid using a computationally expensive layer-by-layer growth, an alternative possibility is to assign labels using a breadth-first search approach. The boundary points between two connected components can be determined as the points that are "equidistant" to the clusters. The definition of equidistant depends on the definition of distance. It can be directly defined based on the grid cells, corresponding to the Manhattan distance. However, this distance metric may be sensitive to the angle values of the unit cell — a tilted cell could introduce bias in the neighbor search towards a particular direction. To overcome this limitation, a Cartesian coordinate grid would be necessary, and a bucket queue prioritized according to Cartesian distances would replace the standard queue based on grid neighbors (as presented in section 5.1.2).

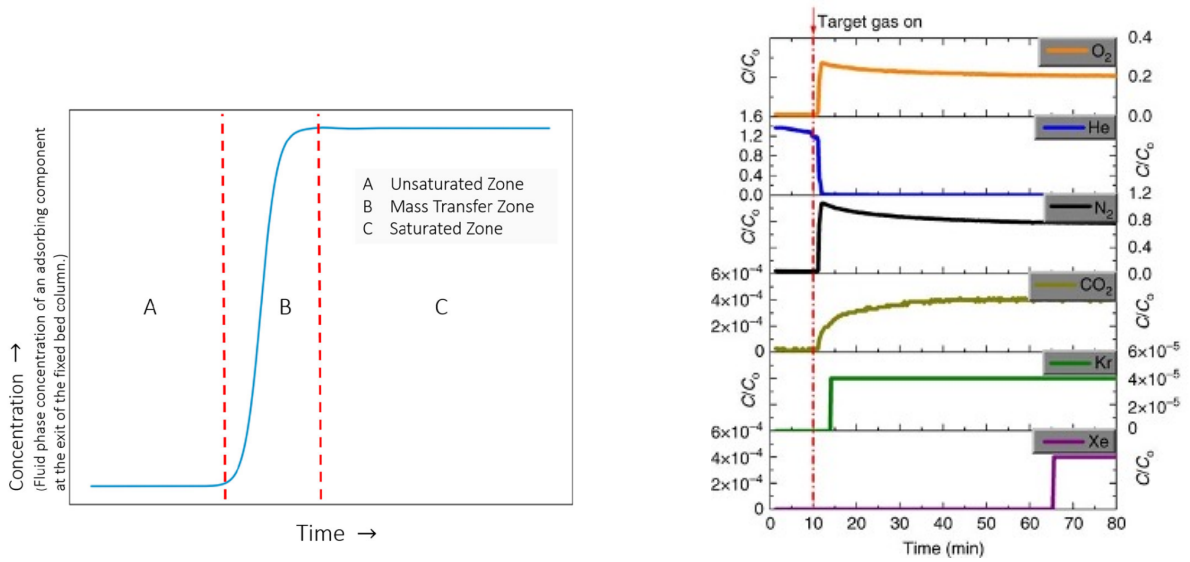
In future research, the focus of the thesis will be on studying different approaches: layer-by-layer growth, breadth-first search with a standard queue, or a bucket queue prioritized using Cartesian distances. The objective is to compare their performance in terms of time and accuracy.

6.2.2 Connection to the breakthrough experiments

The relevance of the diffusion coefficient calculated by these methods can only be reliably validated through a comparison with experimental data. However, accessing the diffusion coefficient inside a nanoporous material can be challenging through experiments. The shape of the experimental breakthrough curves provides a glimpse of the kinetic performance. An indirect comparison can be achieved by generating breakthrough curves from the computed diffusion coefficients and comparing them with experimental data. To separate the kinetics from the thermodynamics of the adsorption process, a breakthrough model based on both the computed transport properties and the experimental isotherm data might be necessary.

A breakthrough curve can be broken down into three different zones: an unsaturated zone, a mass transfer zone, and a saturated zone, as shown in Figure 6.1a. The mass transfer zone can be qualitatively interpreted as a consequence of the different transport properties. For instance, The breakthrough curves are based on quantities obtained from an adsorption isotherm fit, and mass transfer properties such as the self-diffusion coefficient and material surface diffusion (Knudsen diffusion)..**Sharma_2023** This tool has the potential to generate breakthrough curves and compare them to experimental curves. When the isotherm fitting properties are derived from experimental calculations, the only variable remaining is the mass transfer term. In this case, this tool can be used to qualitatively validate a calculated self-diffusion coefficient value.

Using SBMOF-1 as an example (Figure 6.1b), a rather slow mass transfer can be associated with CO_2, while the mass transfer corresponds to a vertical line for the other components,



(a) Breakthrough curve zones *(b) SBMOF-1 experimental breakthrough*

*Figure 6.1: (a) Different zones in a breakthrough curve reprinted from the open-access article [**Sharma_2023**]. (b) Experimental breakthrough curves in SBMOF-1 for a gas mixture with 400 ppm Xe and 40 ppm Kr balanced with dry air. Reprinted with permission from Ref. [**Banerjee_2016**] copyright © 2016 Springer Nature.*

indicating fast diffusion rates. Consequently, there is a diffusion limitation for CO_2 but not for the other adsorbates in the SBMOF-1 material. However, the absence of diffusion limitation for xenon sheds light on an apparent inconsistency with the diffusion coefficient calculated by an MD simulation ($3 \times 10^{-8} cm^2\ s^{-1}$). The next section on the flexible nature of SBMOF-1 offers a reasonable explanation for this discrepancy between the experimental and simulated evaluations of the diffusion.

6.3 Screening of flexible materials

The preference for studying rigid frameworks in computational studies is due to the high complexity associated with the simulation of the dynamics of a flexible framework. Given the considerable cost associated with simulating a grand canonical ensemble using MC methods, the simulation of a flexible framework would be even more computationally expensive, as it would require relaxation of the volume and simulation of an osmotic ensemble (μ,P,T), which necessitates additional MC moves on the volume of the unit cell itself..**Bousquet2012** Although this type of MC simulation describes more accurately every aspect of flexibility, including intrinsic flexibility due to thermal agitation and adsorbate-induced flexibility, it is prohibitively time-consuming in large-scale screening procedures. Therefore, it is more practical to use this type of simulation as a precise method to confirm the properties of a few top materials.

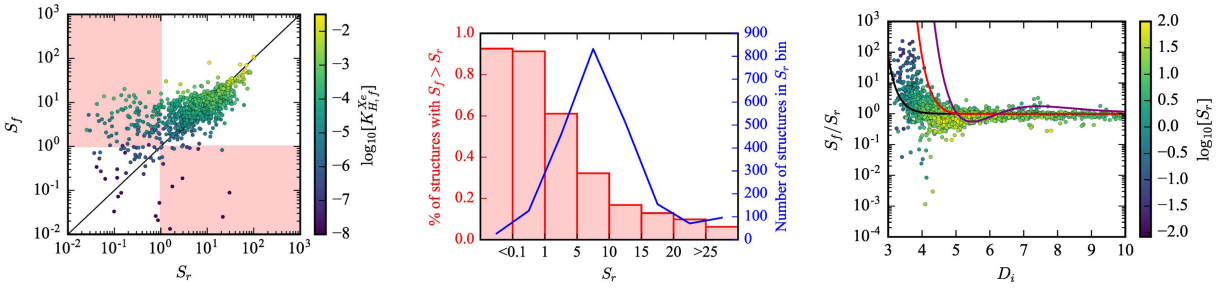
In order to incorporate flexibility effects in the screening procedure at a minimal computational cost, another approach consists in using a set of rigid structures that reflects the structural diversity generated by the thermal agitation of the nanoporous material. A first study on the effect of this intrinsic flexibility on the Xe/Kr selectivity suggests that some materials could lose selectivity due to the less favorable pore size as the structure vibrates.**Witman_2017** For instance,

it was found that the differences between the experimental and theoretical Xe/Kr selectivity of KAXQIL[KAXQIL] are caused by its intrinsic flexibility, which questions the performance ranking obtained through rigid-framework screening. In this section, the study of Witman et al.[Witman_2017] on a screening of intrinsic flexibility will be detailed, with a specific focus on the case of KAXQIL. The subsequent flexibility study will be based on the structural diversity among similar deposited experimental structures. Notably, there exist a dozen different structures with the same chemical nature as KAXQIL, but with very distinct structural characteristics depending on the loaded adsorbate, which suggests an adsorbate-induced flexibility in addition to the previously studied intrinsic flexibility.

6.3.1 Snapshot method

Methodology

To model the dynamics of the framework, Witman et al. used the UFF forcefield to describe the non-electrostatic framework bond potentials, except for the metal bonding. For the bond dynamics around the metal, a harmonic equilibrium potential was fixed around the values extracted from the experimental structure. For this reason, this forcefield definition is referred to as the UFF-fix-metal (UFF-FM). In addition of the Lennard-Jones description, the point charge Coulomb interactions were described using the standard Ewald summation technique based on the charges calculated by the density derived electrostatic and chemical (DDEC) method.[manz2010chemically] Using this forcefield, the authors carried out a systematic snapshot generation of the structures from the CoRE MOF 2014 database with pre-calculated DDEC charges. These snapshots were then utilized to determine the Xe and Kr Henry constant values for the flexible structures, as well as the infinite dilution Xe/Kr selectivity. The study revealed that the flexible selectivity was lower for 95% of the materials with a rigid selectivity over 25 (as shown in Figure 6.2b), which suggests an overestimation of the top performing materials. Furthermore, the effect of flexibility is much more important for materials with smaller pore sizes due to the increased intensity of interactions at shorter distances.



(a) Flexible vs. *Rigid* *(b) selectivity underestimation* *(c) Flexibility effect* vs. *LCD*

*Figure 6.2: (a) A scatter plot of the flexible selectivity against the rigid selectivity labeled by the $\log_10$ of the flexible Xe Henry constants. (b) Barplot of the fraction of the underestimated selectivity ($s_f > s_r$) for different categories of materials going from the least selective ones to the most selective ones ($s_r > 25$). (c) Effect of the flexibility measured using the ratio s_f/s_r as a function of the largest included sphere diameter. The line plots corresponds to analytical modeling of the effect that will not be detailed here. Reprinted with permission from the original paper [**Witman_2017**] copyright © 2017 American Chemical Society.*

Data source	Flexible structure	Xe Henry Constant mmol g^{-1} Pa^{-1}	Kr Henry Constant mmol g^{-1} Pa^{-1}	Xe/Kr selectivity
Experimental data[Banerjee_2016]	maybe	$3.84\,10^{-4}$	$2.37\,10^{-5}$	16
Rigid structure SBMOF-1[Banerjee_2016]	no	$1.45\,10^{-2}$	$2.70\,10^{-4}$	54
PBE+D3 (2,2,1 unit cell)	yes	$6.80\,10^{-3}$	$1.77\,10^{-4}$	38
UFF-FM	yes	$6.24\,10^{-3}$	$1.67\,10^{-4}$	37
UFF-DCM	yes	$3.18\,10^{-3}$	$1.28\,10^{-4}$	25

*Table 6.1: Results of the flexibility analysis carried out by Witman et al., flexibility reduces the values originally calculated in a rigid structure. Reproduced with permission from the original paper [***Witman_2017***] copyright © 2017 American Chemical Society.*

Turning to the issue of flexibility in KAXQIL, the authors used several methods to evaluate its effect on the Xe and Kr Henry constants and the Xe/Kr selectivity. For instance, they leveraged an alternative description of the metal–ligand bond utilizing a cationic dummy model (UFF-CDM) and an *ab initio* MD simulation performed using the PBE DFT function,[Perdew_1996] with a Grimme's D3 van der Waals correction[Grimme_2010] (PBE+D3). Each of these three methods was employed to generate approximately 30 snapshots, which were subsequently used to determine the flexible framework's adsorption properties for KAXQIL.

The authors found that the lower experimental selectivity value of 16, as compared to the UFF-determined value, could be partially attributed to a flexibility effect. As shown in Table 6.1, the selectivity value decreases from 54 to 25 when changing from a rigid to a flexible structure. The selectivity evaluated using the standard UFF forcefield on a rigid SBMOF-1 structure is considerably higher than the selectivity obtained when considering snapshots of a vibrating structure. Although the *ab initio* MD method should provide the closest representation of the actual dynamics, it did not fully capture the phenomenon due to the dependence on system size. Typically, multiple unit cell replications are required to observe crystallographic deformations. Moreover, the UFF forcefield does not provide a perfect picture of the interaction energies at play in the system. Nonetheless, this study establishes an overall trend by attributing the discrepancies between experimental and theoretical data to the rigidity hypothesis.

Although this approach does not fully describe the flexibility effect on the selectivity value, it can rapidly identify a weakness in the rigidity hypothesis, thereby warning of a possible over- or under-estimation of the selectivity. This can lead to the wrongful identification of a material as the best or the missed opportunity of finding a better material. The main advantage of this technique is its relative speed compared to an osmotic ensemble Monte Carlo simulation.[Bousquet2012] However, the imperfect description of the intrinsic flexibility as the only phenomenon at play is its main drawback. For instance, the following discussion will focus on some adsorbate-induced effects that were outlooked but can be retrieved by utilizing multiple works on the same SBMOF-1 material. This approach avoids the issues around simulating the flexible structure, as the reasoning is solely based on experimentally observed structural changes.

6.3.2 Experimental database approach

According to original paper on SBMOF-1,[Banerjee_2016] the theoretical selectivity calculated by UFF is around 70.6. However, the experimental selectivity is significantly lower, around 16.

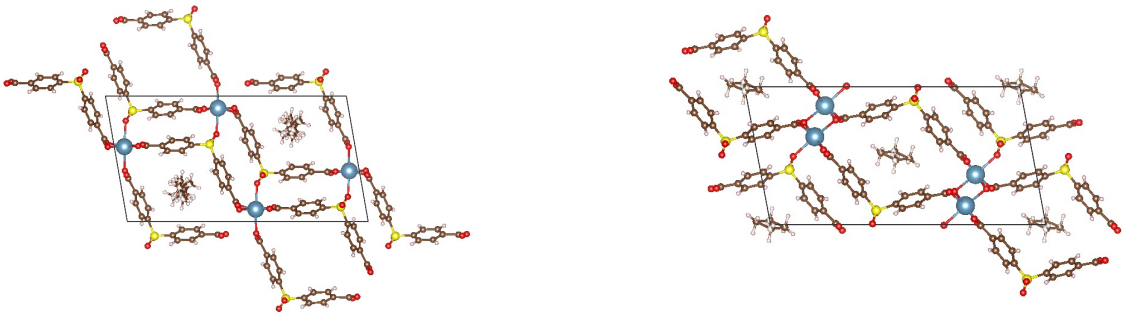
Experimental structure CCSD ref. code	Adsorbate in the structure	Selectivity $s_0^{Xe/Kr}$	K_H^{Xe} (mmol g^{-1} Pa^{-1})	LCD (Å)	PLD (Å)	Xe Diff. Coeff. cm^2 s^{-1}
KAXQOR01[Yeh2012]	Not specified	101	3×10^{-2}	4.99	3.66	3×10^{-09}
KAXQOR[Banerjee2012]	Not specified	22	4×10^{-3}	4.51	4.04	7×10^{-06}
KAXQIL[Banerjee2012]	H_2O	104	3×10^{-2}	5.12	3.77	3×10^{-08}
QUXRIM[Banerjee2016hydro]	hexane	52	1×10^{-2}	4.75	4.31	3×10^{-05}
QUXRUY[Banerjee2016hydro]	hexane	96	3×10^{-2}	4.91	3.57	9×10^{-10}
QUXROS[Banerjee2016hydro]	hexane	99	3×10^{-2}	5.00	3.66	5×10^{-09}
QUXREI[Banerjee2016hydro]	hexane	101	3×10^{-2}	5.02	3.67	7×10^{-09}
QUXRAE[Banerjee2016hydro]	hexane	100	3×10^{-2}	5.03	3.68	7×10^{-09}
QUXQUX[Banerjee2016hydro]	butane	103	3×10^{-2}	5.17	3.83	1×10^{-07}
QUWYEO[Banerjee2016hydro]	butane	100	3×10^{-2}	4.99	3.65	5×10^{-09}
UQEFAZ[Banerjee_2016]	krypton	23	5×10^{-3}	4.53	4.08	5×10^{-06}
UQEFED[Banerjee_2016]	xenon	63	3×10^{-2}	4.89	3.54	1×10^{-11}

Table 6.2: Structural, adsorption and transport properties of structures in the CSD database that are similar to SBMOF-1.[Banerjee_2016] *The last structures actually correspond to the structures resolved in the paper presenting SBMOF-1 in Nature Communications. We can note the structural diversity that induces this diversity of properties. (The pore sizes are calculated using the CCDC radii definition.)*

To solve this difference, Witman et al. used a snapshot-based method to evaluate the effect of selectivity. The intrinsic flexibility lowers the selectivity, which aligns with explaining the difference in selectivity, but it does not appear to capture the whole picture.

For instance, the observed discrepancies could also be explained by the deformation induced by the loading of adsorbate inside the material. For instance, experimentally, a structure is often not empty when resolved by X-ray, and molecules are actually loaded inside. As shown in Table 6.2, the structure that was originally published for its good CO_2/N_2 selectivity[Yeh2012, Banerjee2012] was also tested for water adsorption, and two different structures emerged from this study: KAXQOR and KAXQIL. The first one is loaded with either air or CO_2, and the structure does not seem to be stretched as much (low LCD values around 4.5 Å). The second one, on the other hand, is filled with water that forms big clusters inside the pores and therefore stretches the pore size towards higher values (high LCD values around 5.0 Å). Looking at the structures resolved in the *Nature Communications* study,[Banerjee_2016] depending on the adsorbate (UQEFAZ for krypton or UQEFED for xenon), the LCD and PLD values change in the first order according to the size of the adsorbate as illustrated in Figure 6.4. There are, of course, other effects, like the clustering mentioned for water, but also less expected effects such as the orientation of the adsorbate inside the structure.

As shown in Figure 6.3, the orientation of the hexane molecule inside the material seems to favor either a configuration with a large LCD and a low PLD (QUXRUY), or a slightly lower LCD with a slightly higher PLD (QUXRIM). The material configurations are, however, slightly different from the ones observed with KAXQOR or KAXQIL.



(a) QUXRIM *(b)* QUXRUY

Figure 6.3: An illustration of the effect of the orientation of hexane inside a SBMOF-1-like material. In QUXRIM *(a), the carbon atoms are oriented towards the S atoms, whereas in* QUXRUY *(b) they are oriented towards the Ca atoms. This difference in the orientation could explain the different structural properties of the materials reported in Table 6.2. Color code: brown for C, white for H, red for O, cyan for Ca, yellow for S. The structure visualizations are generated using the VESTA software.*[VESTA]

Now that the adsorbate effects have been fully characterized on a few example configurations, it is easier to understand the thought process that led to the identification of KAXQIL as a candidate for Xe/Kr separation. The KAXQIL structure actually represents the material loaded by water with large pores, which enables a good interaction with a large molecule like xenon. For this reason, it was identified as a top selective material. However, when it was experimentally tested for low-pressure adsorption using the Henry constant, it is most likely that the pores are not stretched, which implies lower Henry constants than expected. The structures UQEFAZ or KAXQOR seem to provide a better description of this low-pressure case since the experimental selectivity values are much more consistent with their theoretical selectivity values.

To confirm this hypothesis, a high-loading Xe/Kr binary mixture adsorption uptake would need to be measured. If xenon is highly represented in the adsorbent material, then the structure would be much more favorable to xenon adsorption, hence increasing the selectivity value closer to the theoretically predicted one. This also highlights a composition effect; if the initial mixture has a low xenon content, the structure would most likely have narrower pores, which could decrease the selectivity. By changing the composition of the binary mixture, this effect could also be measured experimentally if the initial hypothesis on the adsorbate-induced flexibility is correct.

This method could be generalized to other systems by screening for materials with a similar chemical composition and topology, for example. However, finding structures in very different adsorption conditions is not always possible due to biases in research focus. To overcome these limitations, these structures could be either experimentally generated when a material seems interesting to see if flexibility plays a role in the adsorption process, or computationally generated by running structure optimizations on loaded structures. Either way, this new approach to flexibility seems complementary to the ones mentioned previously as it seems to have a similar (or slightly higher due to the adsorbate) computational cost as the Witman approach, while avoiding the computationally prohibitive calculation (in a screening) presented by Bousquet et al.[Bousquet2012]

Diffusion in a flexible environment

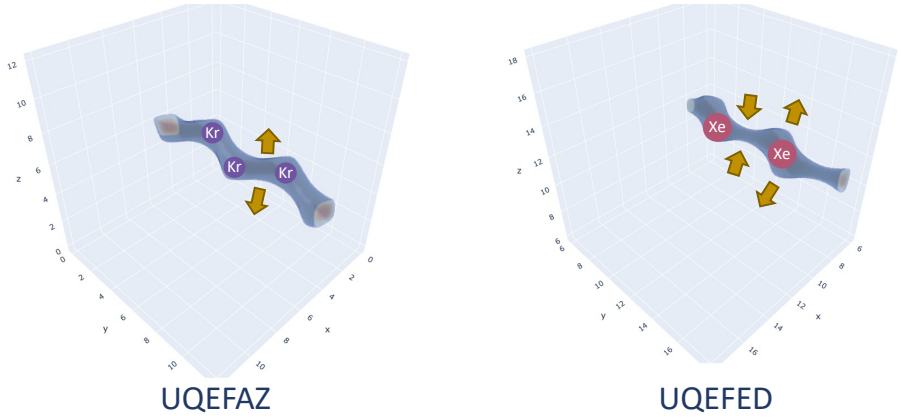


Figure 6.4: Visualization of the pore size stretching effect using the GrAED algorithm. The xenon increases the LCD value while diminishing the PLD value.

Guest transport can also be modulated by the adsorbate-induced flexibility of SBMOF-1. Depending on the structural configuration of the material, the diffusion coefficient becomes limiting only for some configurations of the material: it is equal to 3×10^{-8} cm^2 s^{-1} for KAXQIL and 1×10^{-11} cm^2 s^{-1} for UQEFED (Table 6.2). This lower diffusion coefficient can be explained by the change in PLD value, the channel bottleneck diameter, induced by the stretching illustrated in Figure 6.4. For a material predominantly loaded with krypton molecules, the diffusion channel is significantly broader, which explains the much higher diffusion coefficient of xenon (5×10^{-6} cm^2 s^{-1}). Although the xenon adsorbate does not diffuse freely in the structure, the diffusion is much less obstructed than in the previous material configuration of SBMOF-1 (UQEFED). The Figure 6.4 suggests that the SBMOF-1 material changes its pore configuration according to the shape and size of the atoms loaded inside, and these induced deformations lead to significantly different separation performances.

Now having identified two completely distinct diffusion behaviors, these results can be extrapolated to hypothetical conditions. For instance, if there is a relation between the quantity of xenon inside the pores and the structural similarity towards UQEFED, then the material could kinetically limit the adsorption of xenon at high loading of xenon. In other words, adsorbing xenon at higher xenon loading values may be kinetically more challenging. However, at lower xenon loading, there are no diffusion limitations, as indicated by the steep mass transfer zone observed in the breakthrough curve of xenon in Figure 6.1b. By connecting these results to the influence of flexibility on transport properties and the adsorption process, it can be inferred that xenon adsorption is thermodynamically much more favorable at higher xenon loading. There is a thermodynamics/kinetics trade-off, as articulated in the previous chapter on KAXQIL. Since KAXQIL and UQEFED are structurally similar, the combination of diffusion limitation and high selectivity can be extended to the xenon-loaded structure (UQEFED), as confirmed by the diffusion coefficient and selectivity values reported in Table 6.2. From an industrial perspective, the inclusion of transport effects in the analysis reveals additional costs associated with adsorbing xenon at high loading values, if this theoretical study is validated by experiments.

By employing simple simulation methods (Widom insertion and MD) on rigid structures, the effects of flexibility on both the adsorption and transport properties were probed using experi-

mentally resolved structures under different adsorption conditions. These results shed light on the experiment-theory discrepancies and provides insights into similar problematic systems. In this study, the experimental data published on the SBMOF-1 structure was used, but such resources may not always be available for other systems. In such cases, generating data using experiments or simulations could be necessary. If generalized, this approach would enable its automatic application to a series of structures, paving the way for "flexibility-aware" screenings in the future. Recognizing the importance of both flexibility and transport effects, other studies have attempted to incorporate both in a small-scale screening process.[Stanton_2022] These authors used a flexible forcefield and MD simulations to determine the diffusion coefficient, while adsorption performance was assessed through DFT calculations at the adsorption site. The main issue of this method is its computational cost, but it can serve as an alternative solution to the one introduced here, when no prior knowledge is available on the structure's flexibility.

6.4 Noble gas polarizability

The last effect that could greatly influence adsorption performance is the level of theory behind the modeling of interaction energy. In most screening studies, (our group included), very low-level classical theories are commonly utilized to describe guest–host interactions due to their low computational cost. To improve the accuracy of descriptions, some studies focus on a few specific structures and use higher levels of theory such as DFT calculations. However, the computational cost associated with these methods is prohibitive for high-throughput screenings.

Prior to exploring higher-cost methods, it is essential to first identify the limitations of molecular modeling in current screening methodologies. This work is motivated by recent advancements in experimental design of nanoporous materials for Xe/Kr separation. The most selective materials are based on highly polar groups or exposed open-metal sites.[Li_2019, Pei_2022] The polarization phenomenon is, therefore, central to these materials, but it cannot be adequately described by a simple Lennard-Jones potential, particularly when induced by high partial charge values.

For this reason, it is necessary to develop a polarizable forcefield that incorporates the effect of the surrounding partial charges into the guest–host interactions. The difference in polarizability between xenon and krypton may lead to the emergence of new materials. The ranking of the best materials obtained through this type of screening would differ significantly from the standard ranking. This section will present the problem of current methodologies through an experiment-theory comparison and explore alternative methodologies that can account for polarization in the commonly employed Lennard-Jones potentials.

6.4.1 Problem definition

If the selectivity of good materials for xenon/krypton separation that are often presented in the literature is considered, the materials named $Co_3(HCOO)_6$,[Wang_2014] CC3,[Chen_2014] SBMOF-2,[Chen_2015] CROFOUR-1-Ni,[Mohamed_2016] SBMOF-1,[Banerjee_2016] $Co_3(C_4O_4)_2(OH)_2$[Li_2019] and ZJU-17a[Pei_2022] often appear as top separation materials. When the selectivity values obtained through a Widom insertion with the UFF forcefield are compared to the experimental

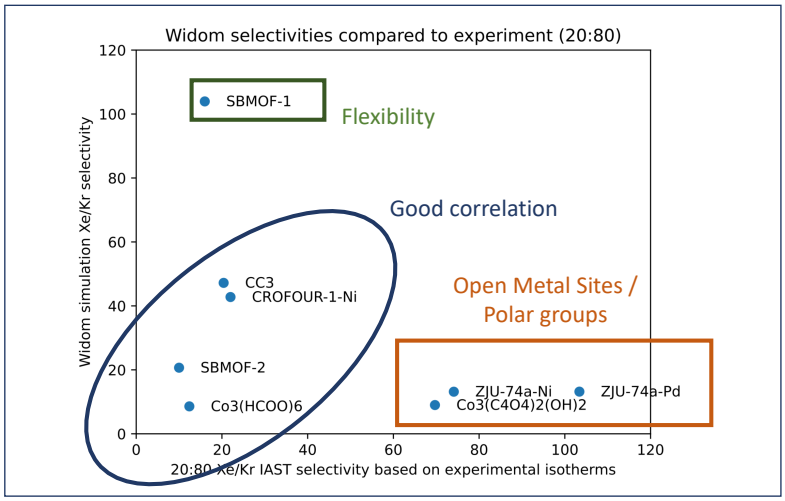
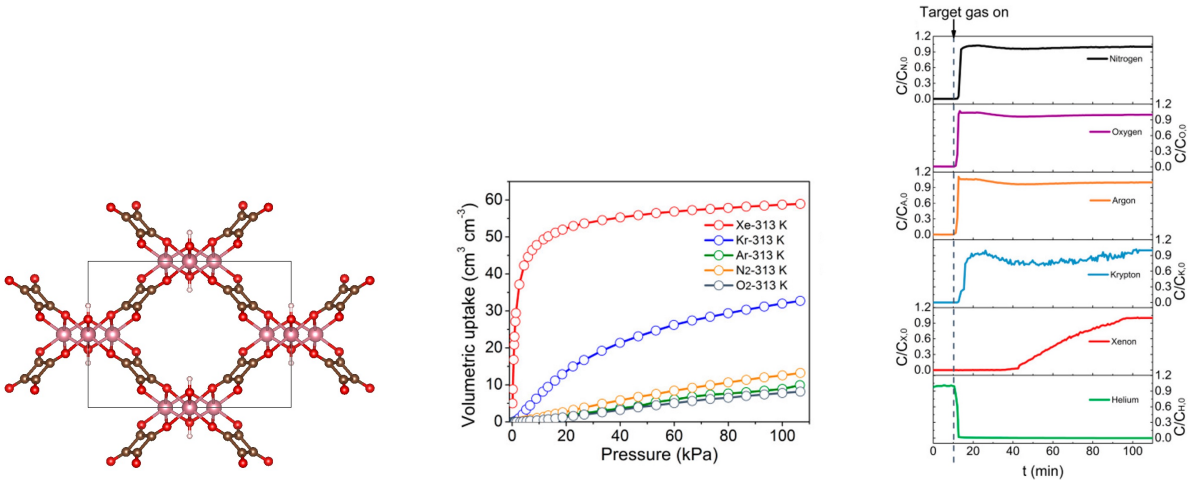


Figure 6.5: Comparison between the selectivity values obtained experimentally and computationally. The structures are split into three categories depending on the difference between experiments and theory. The case of SBMOF-1 can be attributed to its flexibility. The discrepancies between the materials in the lower right correspond to the ones introduced by Li et al. and Pei et al.,[Li_2019, Pei_2022] *and the difference can be explained by the polarization that is not included in the level of theory considered.*

values as shown in Figure 6.5, a good correlation is generally observed. However, in some cases like SBMOF-1, the difference observed could be explained by other effects (see previous section on flexibility). In other cases, the difference could be explained by the polarization effect that was not taken into account. To better understand this phenomenon, this study will focus on two papers that found record-breaking Xe/Kr selectivity values based on polar hydroxyl groups and open metal sites.

The first paper of Li et al.[Li_2019] published in 2019 introduced a squarate-based MOF with a Xe/Kr selectivity of 69.7 for a 20:80 binary mixture estimated by the ideal adsorbed solution theory (IAST).[Cessford_2012] The authors explained this outstanding xenon affinity by two factors: a pore size close to the kinetic diameter of a xenon and the stabilization effect of the hydroxyl group. DFT calculations determined binding energies of the order of 44.1 kJ mol^{-1} for xenon and 33.7 kJ mol^{-1}, which suggests a separation process of enthalpic nature (usually the case for highly selective materials). Due to the high electronegativity of the oxygen atom, the hydroxyl group pointing to the pore center (as illustrated in Figure 6.6a) interacts strongly with the xenon through a permanent dipole–induced dipole interaction (introduced in the section 2.1.2). This high xenon affinity is illustrated by the experimental isotherms in Figure 6.6b. On the other hand, the pore wall creates unidimensional channels that present adsorption pores with a 4.1 Å×4.3 Å size, which is very close to the xenon kinetic diameter of about 4.0 Å.

Finally, the breakthrough experiment data (Figure 6.6c) reveals a relatively slow release of xenon in the mass transfer zone, which suggests a relatively low xenon diffusion coefficient in this material. The diffusion limitation seems to occur even at very low xenon partial pressure (400 ppm) in this material. This was not the case for SBMOF-1, as the xenon breakthrough curve was much steeper (rapid mass transfer) as shown in Figure 6.1b. To have a closer look



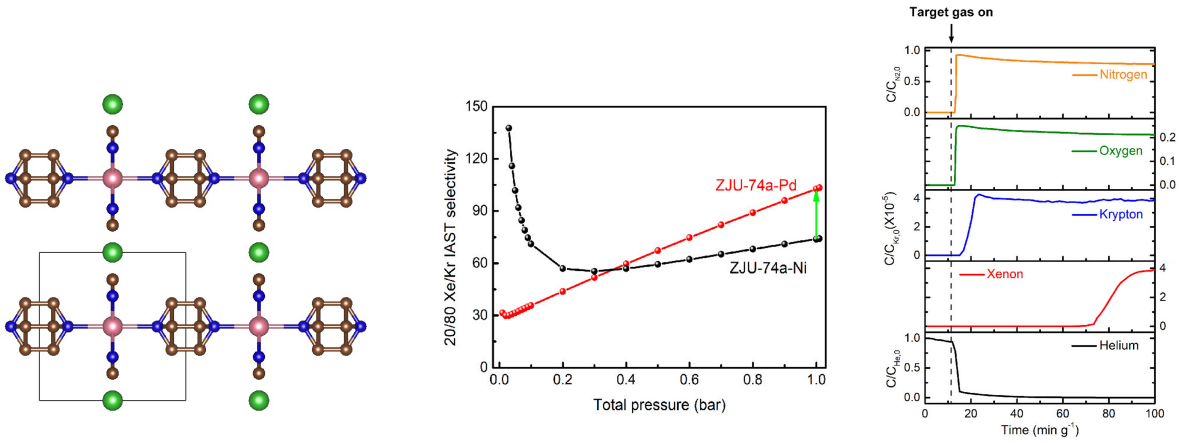
(a) VESTA visualization[VESTA] *(b) Adsorption isotherms* *(c) Breakthrough curves*

*Figure 6.6: (a) Representation of the squarate-MOF $Ce_3(C_4O_4)_2(OH)_2$ structure with the color code: brown for C, white for H, red for O, pink for Co. The hydroxyl group, to which the high Xe/Kr selectivity is attributed, is visible in the structure. (b) Mono-component adsorption isotherm measured experimentally for Xe, Kr, Ar, N_2 and O_2. (c) Experimental breakthrough curves for a gas mixture with 400 ppm Xe and 40 ppm Kr balanced with dry air. Reprinted with permission from Ref. [***Li_2019***] copyright © 2019 American Chemical Society.*

at the transport effect in this squarate-MOF, similar simulations should be performed as for SBMOF-1 with a polarizable forcefield.

In the second work,[Pei_2022] Pei et al. introduced two Hofmann-type MOFs with record-breaking Xe/Kr selectivity values. The first Co/Ni-based MOF, called ZJU-74a-Ni, has an estimated IAST selectivity of 74.1 for a Xe/Kr binary mixture of composition 20:80 at 1 bar and 298 K, while the second Co/Pd-biased MOF, ZJU-74a-Pd, displays a selectivity of 103.4 in the same ambient-pressure conditions. As shown in Figure 6.7b, the IAST selectivity of ZJU-74a-Pd is not always that high and can decrease to 30 at very low-pressure conditions. The authors attribute the record-breaking selectivity values of these materials by a size close to the kinetic diameter of xenon and, above all, the increased interaction with the open metal site, either the nickel or the palladium atoms. The Horvath–Kawazoe method provided a pore size of 4.0 Å and 3.8 Å for the Ni and Pd-based MOFs, respectively. The xenon binding energy was evaluated to be around 38 kJ mol^{-1} for ZJU-74a-Ni using a UFF-based method. It should be noted that this value is much lower than the one obtained for the squarate-based MOF. Further investigations with a DFT method are required to determine the real binding energy for this material due to its higher experimental performance compared to the computational result.

Finally, the breakthrough experiment suggests a relatively slow mass transfer in the material. However, when compared to the squarate-based material under similar conditions, the mass transfer seems to be much faster, as the mass transfer zone is shorter. Therefore, it can be inferred that ZJU-74a-Pd is probably a better material than $Co_3(C_4O_4)_2(OH)_2$ due to its superior adsorption and transport properties for Xe/Kr separation. In comparison to SBMOF-1 with a relatively low xenon partial pressure, a slight diffusion limitation phenomenon seems to be present. The retention of xenon is, on the other hand, longer in ZJU-74a-Pd (around 70 s) than in SBMOF-1 (around 65 s). More information regarding the ambient-pressure selectivity of



(a) VESTA visualization[VESTA] *(b) IAST selectivity* *(c) Breakthrough curves*

*Figure 6.7: (a) Representation of the ZJU-74a-Ni structure with the color code: brown for C, white for H, red for O, pink for Co, green for Ni. We can see the open metal sites or coordinatively unsaturated nickel metals that could interact with an adsorbate in the center of the pore. (b) Selectivity values at different pressure conditions for a 20:80 Xe/Kr binary mixture calculated by the IAST theory. (c) Experimental breakthrough curves of a gas mixture with 400 ppm Xe and 40 ppm Kr balanced with dry air in ZJU-74a-Pd. Reprinted with permission from Ref. [**Pei_2022**] copyright © 2022 American Chemical Society.*

SBMOF-1 is necessary to complete the comparison. This material is also noteworthy for its robustness in different pH, humidity and radiation conditions, making it an ideal choice for capturing xenon produced by nuclear reactions in nuclear installations.

These two studies clearly demonstrate the failure of current screening methodologies in identifying materials whose performance relies on polarization effects. The next and final discussion will introduce some methods for incorporating polarization into Lennard-Jones potentials that could be used in a screening procedure.

6.4.2 Studying the polarization

The physical reason behind the consideration of the polarization effect for xenon/krypton separation is to exploit the difference in polarizability between Xe (4.0 $Å^3$) and Kr (2.5 $Å^3$)[Olney1997] to its full potential. Seen from a broad perspective, the order of magnitude of the induction energy is actually higher than other standard van der Waals energies, as explained in section 2.1.2. For instance, the ion-induced dipole interaction is reported to range between 40–600 kJ mol^{-1}. In the case of ZJU-74a-Ni, the Ni^{2+}—Xe—Ni^{2+} interaction originates from selectivity, which corresponds to this specific type of interaction and predominantly explains the experimental selectivity values. Incorporating polarization into the screening procedure has the potential to completely alter the obtained structures and the types of interactions at play.

Bearing this in mind, Becker et al. carried out an interesting study on a series of MOF materials with a high density of open-metal sites, the M-MOF-74 with M = Co, Cr, Cu, Fe, Mg, Mn, Ni, Ti, V, and Zn.[Becker_2017] By introducing a potential induced by the surrounding partial charges to a modified LJ potential, the authors successfully replicated the experimental isotherm data for CO_2 and CH_4 adsorption on this series of MOFs. They also showed the inadequacy of the

standard UFF force field in describing the adsorption behavior of CO_2 on the open-metal sites, thereby overlooking a highly adsorptive site at infinite dilution.

This novel method is based on the procedure developed by Lachet et al.,[Lachet_1998] which considers the induced dipole method, where the induction energy U_{ind} is expressed as follows:

$$U_{ind} = -\frac{1}{2}\sum_{i=1}^{N} \boldsymbol{\mu}_i \cdot \mathbf{E}_i^0 \tag{6.1}$$

where $\boldsymbol{\mu}_i$ is the induced dipole, and $\mathbf{E}_i^0$ is the electric field created by the surrounding atoms' partial charges on the particle *i*. Since the induced dipole also interacts with the surrounding induced dipoles $j \neq i$, the induced dipole is usually calculated using a back-propagation algorithm as described in the Ref. [**Lachet_1998**]. However, it was found that back-propagation accounts for less than 5% of the total induction energy. For this reason, the equation can be simplified by considering only the interaction between the induced dipole and the surrounding electric field, without taking into account induced-dipole–induced-dipole interactions. Moreover, skipping the back-propagation step saves valuable computation time during screening. The induction energy can then be expressed as follows:

$$U_{ind} = -\frac{1}{2}\sum_{i=1}^{N} \alpha_i \left|\mathbf{E}_i^0\right|^2 \tag{6.2}$$

Since a portion of the induction energy is already incorporated into the Lennard-Jones potential, the authors rescaled the LJ parameters to eliminate the induction part from the LJ energy. This part appears to be system-dependent and may be debatable since it allows for fitting to experimental data without solid theoretical justification. To address this properly, a force field should be designed around this concept to fine-tune the LJ parameters based on specific experimental data, similar to standard force field development practices.

By customizing this method for xenon/krypton separation, it may be possible to conduct screenings that yield materials comparable to ZJU-74a-Pd. Further optimization of the process can be achieved by narrowing down the materials' selection through restrictions on pore size and the presence of open metal sites within a database. Subsequently, the restricted material list can be evaluated using higher-level methods like the ones described in this chapter.

GENERAL CONCLUSIONS

This thesis explored various approaches to find the best nanoporous materials for adsorption-based industrial xenon/krypton separation (e.g., pressure-swing adsorption). As highlighted in the literature review,[Ren_2022] high-throughput screening methods focus on a specific property of nanoporous materials to identify the most suitable material for targeted application. Such screenings face three main challenges: (i) achieving the accuracy of the methods used to characterize the key properties, (ii) reducing the experimental/computation time required to determine those properties, and (iii) incorporating additional properties often overlooked in performance evaluation. Adsorption selectivity, commonly used to assess separation performance, is typically used in this regard.

Chapter 2 focused on the different methodologies used to evaluate selectivity in different physical conditions, and demonstrated how screenings can provide a realistic picture of selective materials.[Ren_2021] The influence of composition, pressure and some structural descriptors were thoroughly examined. It was found that tailoring the pore size to match the size of xenon is key in achieving maximum selectivity. The Xe/Kr separation can be approximately described by the affinity of xenon for the material, which predominantly manifests in its enthalpic nature. When the partial pressure of xenon increases, the most favorable pores for xenon adsorption become saturated, leading to an observed selectivity decrease in certain materials.

Considering the prominent role of the enthalpic term, Chapter 3 introduced faster sampling techniques to evaluate selectivity under infinite-dilution conditions. In addition to the widely used Widom insertion method, various biased sampling techniques such as Voronoi sampling and surface sampling (RAESS) were described. The RAESS algorithm[Ren_2023] demonstrated superior speed compared to Widom insertion and higher accuracy than the previously introduced Voronoi energy[Simon_2015] on the CoRE MOF 2019 database. Finally, an unbiased sampling approach utilizing a symmetric grid (GrAED) was introduced to generate valuable energy descriptors for the design of finely tuned energetic descriptors. The GrAED algorithm provides interesting descriptors that can be further used, for instance, in an ML modeling. These techniques can be incorporated into a multiscale screening to efficiently identify promising materials for more time-consuming calculations or experiments.

Chapter 4 proposed an ML model based on structural, chemical, and energetic descriptors to achieve GCMC-level accuracy combined with a speed comparable to faster low-dilution calculations.[Ren_2023_ml] This ML model demonstrated high accuracy and enabled GCMC-grade evaluations to be obtained with minimal computational resources. Importantly, the interpretation of this ML model offered novel approaches for investigating the structure-property relationship beyond conventional correlation analyses.

To date, extensive research has focused on investigating thermodynamic properties computed using relatively simplistic assumptions through multiple correlation analysis and the development of various performance evaluation tools. To encompass different key properties, the transport properties were studied in Chapter 5. Different methodologies were investigated, including: (i) molecular dynamics, which represents the most physically accurate but also the slowest method, requiring simulations of at least a few tens of nanoseconds to capture the diffusion process accurately; (ii) transition state-based methodologies that approximate the diffusion process by hopping from one site to another; (iii) an ML-based approach that uses descriptors based on activation energies (transition state theory) to predict the diffusion coefficient. By leveraging these calculated transport properties, the screening process successfully identified selective materials without kinetic limitations (that also happen to have high xenon capacity). This outcome validates the multivariate nature of the screening process, as such materials have the potential to significantly enhance productivity, yield a greater output during each pressure-swing cycle and enable faster cycles (in a PSA process).

The final chapter provides prospects for future research studies regarding additional physical properties of materials that have been overlooked throughout this thesis. These properties include the flexibility of the material and the interactions induced by charged atoms or polar groups. By incorporating the polarization effect into the screening process, it becomes possible to identify materials with significantly higher experimental selectivity, as suggested by the characteristics of recently identified top-performing materials for Xe/Kr separation.[Li_2019, Pei_2022] Furthermore, the flexibility of the material can potentially provide insights into theory–experiment discrepancies that would otherwise remain unresolved, thus highlighting the importance of employing more accurate descriptions through flexibility-aware molecular simulations.

This work paves the way for more efficient screening strategies aimed at investigating separation properties under diverse physical and chemical conditions. Moreover, the novel tools developed in this thesis can readily facilitate the integration of transport properties into future screening for gas separation involving nanoporous materials. By combining these tools with the emerging concept of Digital Reticular Chemistry,[Lyu_2020] new possibilities for material design and discovery can be envisioned.

The methodologies developed in this thesis also enable the integration of understudied properties that have a key role in the industrial process of xenon/krypton separation. By integrating the faster methods introduced throughout this thesis, it becomes feasible to consider physical phenomena that are typically overlooked in the screening procedure. The faster sampling of adsorption energies can serve as a foundation for modeling flexible materials using a snapshot approach, as demonstrated by Witman et al..[Witman_2017] The evaluation of induced energy[Lachet_1998] can also be integrated into the evaluation tools used throughout this thesis.

LIST OF PUBLICATIONS

Peer-reviewed papers

Preprint

RÉSUMÉ EN FRANÇAIS

Introduction

Les procédés industriels de séparation des gaz sont utilisés dans diverses industries, telles que la chimie, la santé, l'agriculture et l'alimentation, pour fournir des réactifs purifiés et des gaz inertes. Ces procédés sont également utilisés pour atténuer les effets négatifs de certaines activités industrielles sur l'environnement, comme la capture du dioxyde de carbone dans les usines de production de béton ou d'acier ou encore le piégeage de composés radioactifs volatils des usines de retraitement des combustibles nucléaires. Différentes petites molécules comme le diazote, le dioxygène, le dioxyde de carbone, le dihydrogène, le méthane, le protoxyde d'azote ou les gaz rares sont ainsi séparées, purifiées puis stockées. Cette thèse se concentre sur la séparation xénon/krypton communément utilisée pour extraire ces gaz de l'atmosphère, mais aussi de l'industrie du nucléaire qui constitue une source bien plus abondante de xénon et de krypton.

Les procédés industriels de séparation Xe/Kr sont encore bien souvent basés sur la distillation cryogénique de l'air ambiant, ce qui requiert beaucoup d'énergie, une infrastructure complexe et un contrôle minutieux des risques. On peut par exemple évoquer les récents accidents d'exploitation d'usine de séparation de gaz (1997) qui ont été causés notamment par la réaction d'hydrocarbures de l'environnement avec l'oxygène liquéfié de l'usine. Pour éviter les problèmes de sécurité et de coûts importants, de nombreux chercheurs s'attèlent à développer des méthodes de séparation industrielle basées sur l'adsorption dans des matériaux nanoporeux. Ces matériaux nanoporeux sont constitués de pores à l'échelle nanoscopique qui offrent une large surface aux molécules pour y interagir puis s'adsorber. Des procédés industriels basés sur cette technologie existent déjà, ils utilisent notamment le *pressure swing adsorption* (PSA) qui consiste à remplir les pores d'un mélange de gaz à haute pression, puis de récupérer un gaz ainsi purifié. En effet, les pores du matériau permettent l'adsorption préférentielle d'une molécule par rapport aux autres ce qui permet d'augmenter la teneur en une certaine molécule du mélange sortant. En répétant ce procédé, on peut ainsi séparer les différentes molécules d'un gaz. Dans le cadre de

ma thèse, le xénon étant chimiquement proche du krypton, la purification par ce procédé reste un défi majeur. Certains prototypes industriels ont déjà été imaginés, mais la recherche d'un matériau pour effectuer au mieux cette tâche reste aujourd'hui une question ouverte.

Pour développer un procédé viable, il faut donc choisir avec soin les matériaux que l'on utilise dans ces dispositifs industriels. La recherche se focalise aujourd'hui sur la conception de matériaux toujours plus sélectifs en se basant sur des intuitions chimiques construites au fil des études. Afin d'éviter les expériences coûteuses pour tester tous les matériaux, les criblages computationnels sont de plus en plus utilisés. Ces criblages ou *screenings* en anglais permettent de passer en revue de grandes quantités de structures afin d'en évaluer leur potentielle performance. Tout l'enjeu est donc de former une bonne synergie entre la conception minutieuse de matériaux et la recherche et évaluation rapide des matériaux via des méthodes informatiques. Du côté du traitement informatique des matériaux, les deux défis majeurs sont la génération de données fiables et diverses afin de couvrir le spectre des possibles et le développement de nouveaux outils pour l'évaluation rapidement et avec précision les performances de ces matériaux.

La quantité de matériaux est potentiellement infinie, rien que pour les *metal–organic frameworks* (MOFs) en anglais, plus de 90 000 structures ont été synthétisées et 500 000 ont été construits de manière digitale. Pour pouvoir évaluer tous ces matériaux, différentes stratégies ont été élaborées. Certains utilisent des criblages à plusieurs niveaux qui permettent de réduire au fur et à mesure les matériaux à évaluer avec des méthodes plus coûteuses, d'autres se basent sur des algorithmes d'apprentissage statistiques. Cependant, peu d'études se focalisent sur les outils de calcul, en eux-mêmes, qui sont souvent davantage adaptés à des calculs sur des structures uniques plutôt que pour être déployés sur des centaines de milliers de structures. Cette thèse s'emploie donc à développer des outils pour accélérer les procédés de criblages actuels tout en travaillant sur la précision des évaluations de performance. Outre la sélectivité, d'autres variables revêtent une importance significative : la capacité d'adsorption du matériau, la cinétique et la thermodynamique derrière la régénération du matériau (c'est-à-dire en vider les pores). Pour cette raison, ma thèse étudie également les propriétés de transport du xénon et du krypton dans ces matériaux nanoporeux.

Étude thermodynamique de la séparation Xe/Kr

En premier lieu, mes travaux ont porté sur l'analyse poussée des corrélations qu'il pouvait exister entre les différentes grandeurs thermodynamiques décrivant la séparation xénon/krypton. Pour cela, mes travaux se basent sur la base de données CoRE MOF 2019 pour comparer les différentes grandeurs thermodynamiques grâce à des analyses de corrélation. Différentes conditions de pression et de composition ont été étudiées et des explications physiques à l'échelle microscopique sont proposées pour comprendre l'origine des différences observées.

Pour commencer, j'ai étudié les corrélations entre l'enthalpie et la sélectivité. Sur la figure R1, l'enthalpie d'adsorption du xénon est assez bien corrélée au logarithme de la sélectivité à basse pression suggérant ainsi que l'affinité du xénon avec le matériau peut expliquer la sélectivité. Cette corrélation diminue cependant pour les matériaux moins sélectifs. Les matériaux les plus sélectifs ont en effet des pores dont la taille est très favorable à l'adsorption du xénon comme le suggèrent d'autres études. Pour des gaz nobles, seules les interactions de Van der

Waals jouent un rôle important, ainsi la taille des pores permettent d'expliquer en grande partie l'affinité comparée entre deux molécules de tailles différentes le xénon et le krypton. Ainsi, dans des matériaux avec de petits pores, les phénomènes sont dominés par les interactions entre les pores et l'adsorbat, c'est-à-dire par l'enthalpie. Alors que dans de larges pores, les effets entropiques jouent un rôle plus important.

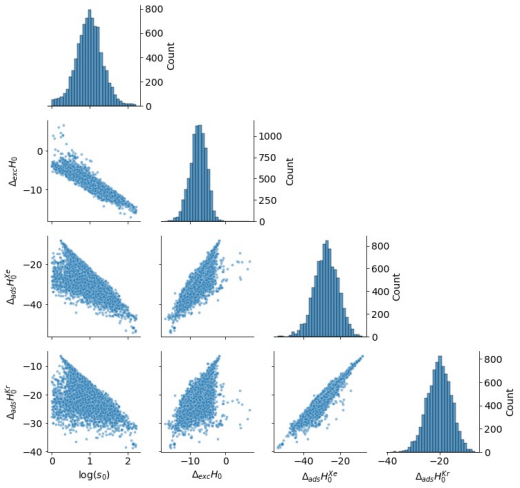


FIGURE R1 : Pour 8 401 MOFs avec une sélectivité Xe/Kr favorable ($s_0 > 1$), pair-plots entre les différentes grandeurs $\log(s_0)$, $\Delta_{exc}H_0$, $\Delta_{ads}H_0^{Xe}$ *et* $\Delta_{ads}H_0^{Kr}$ *(les enthalpies sont en* $kJ\,mol^{-1}$*) en dehors de la diagonale et la distribution de chaque grandeur sur la diagonale.*

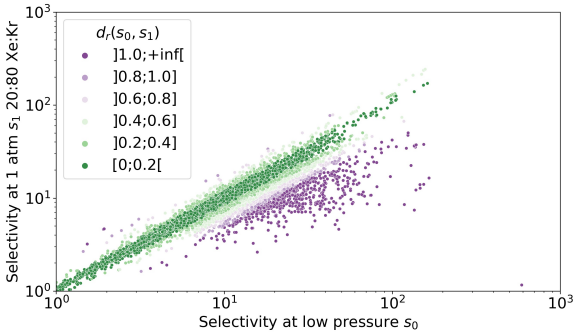


FIGURE R2 : Sélectivité à 1 atm de pression en fonction de la sélectivité à basse pression pour une composition 20:80 Xe/Kr. Les points sont étiquetés selon la différence relative entre les deux sélectivités. Les points violets ont une grande différence relative entre les sélectivités.

D'autre part, nous observons sur la figure R1 que l'enthalpie d'échange est très bien corrélée à la sélectivité. Cela peut s'interpréter à l'aide de l'équation suivante $\Delta_{exc}H = T\Delta_{exc}S - RT \ln s$ dans le cas où $T\Delta_{exc}S$ serait quasi constante. En effet, l'entropie joue le rôle de bruit d'un point de vue statistique ce qui est confirmé par d'autres figures au chapitre 2 de cette thèse, où l'on observe clairement l'absence totale de corrélation avec la sélectivité. Cette première figure nous renseigne ainsi sur le rôle prédominant de l'enthalpie d'échange pour expliquer la sélectivité observée.

La figure R2 quant à elle met en évidence la chute de la sélectivité de certains matériaux lorsque l'on passe de la basse pression à la pression ambiante. Cette différence de sélectivité est étudiée à l'aide de l'enthalpie et l'entropie d'échange. Et nous remarquons à nouveau que ce changement de sélectivité est en grande partie expliqué par une augmentation de l'enthalpie d'échange pour ces structures. L'entropie joue encore un rôle relativement mineur sur ce phénomène. L'étude des données thermodynamiques sur un ensemble de 9 668 structures nous suggère que l'enthalpie d'échange définie précédemment permet d'expliquer en grande partie les tendances des sélectivités thermodynamiques à haute et basse pression. La séparation xénon/krypton est donc dominée par des effets enthalpiques.

Pour mettre en évidence les phénomènes physiques à l'origine de la chute de sélectivité pour certains matériaux, nous allons présenter dans ce résumé une structure problématique en particulier pour illustrer les caractéristiques de la structure. Dans ce matériau, il n'y a pas qu'un seul type de site d'adsorption ni un seul canal unidirectionnel. Le matériau WOJJOV

(figure R3) est un exemple de structure contenant deux types de pores comme on peut le voir sur la représentation graphique et ce qui est confirmé par la validité d'un modèle à 2 sites pour décrire les isothermes corps pur. Le premier type de type est plus petit et a une taille parfaite pour adsorber le xénon. C'est pourquoi, à basse pression la sélectivité calculée est très élevée $s_0 = 146$. Lorsqu'on augmente la pression, les sites plus larges commencent à être occupés. Or ces sites plus larges sont moins sélectifs du fait de leur taille. C'est pourquoi la sélectivité diminue grandement et passe à $s_1 = 14$ à pression ambiante. D'autres structures ayant un système plus complexe de canaux baissent également en sélectivité avec la pression.

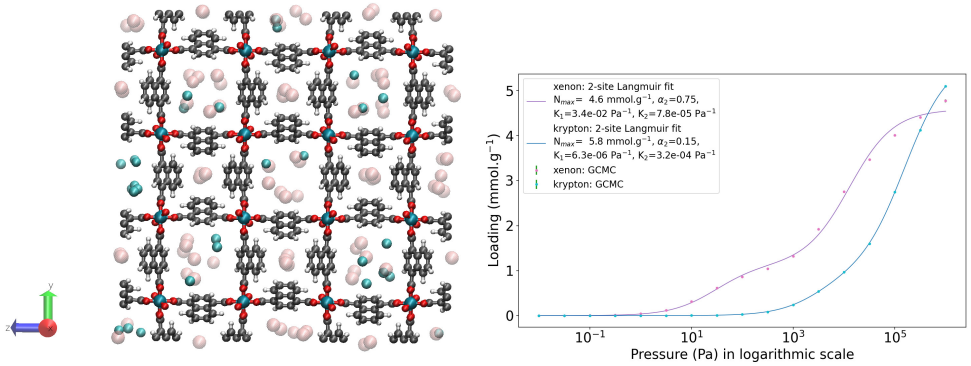


FIGURE R3 : WOJJOV : Représentation d'un MOF [Al(OH)(1,4-NDC)]·2(H_2O) où NDC signifie naphthalenedicarboxylate. Code couleur : Cu en cyan foncé, C en gris, O en rouge, H en blanc ; Xe en rose et Kr en cyan clair. Sur la droite, les isothermes corps pur du xénon et krypton à 298 K ainsi qu'un modèle d'isotherme à deux sites.

Pour conclure, la présence de différents types de site et les réorganisations dues aux interactions du mélange Xe/Kr dans la phase d'adsorption permettent d'expliquer à l'échelle moléculaire la différence de sélectivité à basse et haute pression pour un certain nombre d'exemples. De plus, la séparation Xe/Kr est dominée par les effets enthalpiques, donc une bonne description de l'énergie d'adsorption est primordiale pour décrire la sélectivité d'un matériau.

Développement d'outils de criblage

Dans cette partie on va s'intéresser à différentes méthodes de calcul de l'enthalpie d'adsorption qui joue un rôle central dans la performance d'un matériau. Cette enthalpie à basse pression peut être théoriquement calculée grâce à un échantillonnage des énergies d'interaction pour tous les points accessibles de l'espace, mais cette méthode est coûteuse en temps de calcul. C'est pourquoi les méthodes d'échantillonnage aléatoire des points de l'espace se basant sur les algorithmes Monte Carlo sont plus souvent utilisées (insertion de Widom). Cependant, cet échantillonnage aléatoire ne tient donc pas en compte des informations que l'on a sur les matériaux nanoporeux. En effet, les adsorbats ne se situent pas à des endroits imprévisibles, ils sont souvent aux centres des pores (si la taille est adaptée) ou sur la surface des pores. On a donc exploité ces informations afin de diminuer le temps de calcul nécessaire à la détermination de l'enthalpie d'adsorption.

La première méthode approchée d'échantillonnage consiste à calculer les énergies sur les nœuds de Voronoï. Les nœuds de Voronoï sont des points équidistants à au moins quatre atomes de la structure. Si on considère uniquement les points de Voronoï accessibles, ces points seront situés

au centre des pores. La deuxième méthode quant à elle échantillonne les surfaces des pores. Pour cela l'algorithme RAESS parcourt les points à la surface des atomes de la structure et y calcule l'énergie d'interaction avec le matériau. Et enfin, la dernière méthode développée durant cette thèse se base sur une grille symétrique optimisée pour faire baisser le temps de calcul par rapport à l'approche usuelle par grille. L'algorithme associé GrAED (*Grid Adsorption Energy Descriptors*) est ainsi très intéressant pour des bases de données ayant des petits pores et des structures avec un haut degré de symétrie. La figure R4 compile les différentes performances de précision et de temps pour toutes les méthodes de calcul de l'enthalpie d'adsorption étudiée durant ma thèse.

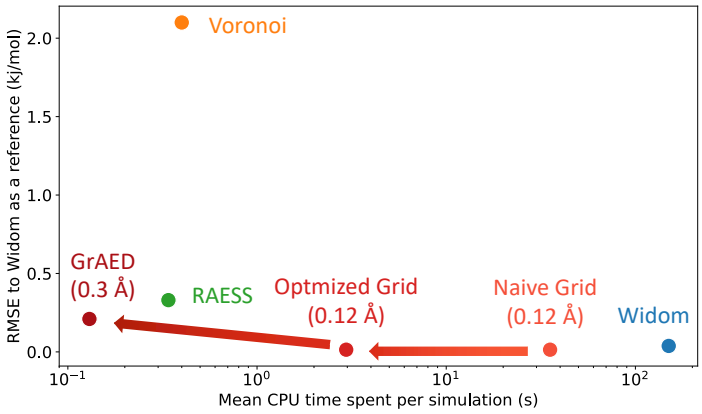


FIGURE R4 : Comparaison de la racine de l'erreur quadratique moyenne sur l'enthalpie d'adsorption du xénon et le temps de simulation par structure pour différentes méthodes d'échantillonnage sur la base de données CoRE MOF 2019 (pour une taille de pore supérieure à 3.7 Å*).*

Finalement, j'ai également développé un modèle de machine learning basé sur les descripteurs de GrAED et des descripteurs structurels plus couramment utilisés. Ce modèle donne de très bons résultats de prédiction tout en étant bien plus rapide que les simulations GCMC plus couramment utilisées. La figure R5 montre l'excellent accord entre les valeurs réelles et celles prédites par le modèle. Quantitativement, l'erreur sur le logarithme base 10 de la sélectivité vaut environ $0,07$ (RMSE).

Ce modèle peut ensuite être utilisé pour accélérer l'évaluation de la sélectivité Xe/Kr à pression ambiante pour trouver plus rapidement les meilleurs matériaux pour la séparation. Cette partie conclut l'étude des effets thermodynamiques de la séparation xénon/krypton.

Propriétés de transport

Enfin, mes derniers travaux portent sur la modélisation des effets de transport du xénon et du krypton dans les structures poreuses de CoRE MOF 2019. Les effets de transport peuvent influencer les performances d'un matériau utilisé comme un adsorbant, comme illustré sur la figure R6. L'accès aux pores pour adsorption peut être plus ou moins rapide selon la vitesse de diffusion dans le matériau. Pour des membranes de séparation, l'effet de transport devient même la mesure principale de performance. Les effets de transport ont été estimés en utilisant des simulations de dynamique moléculaire.

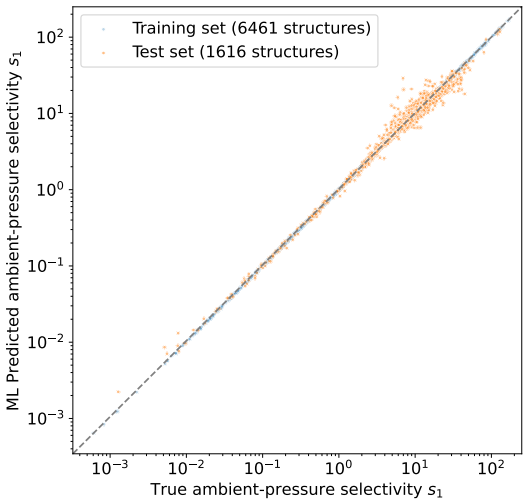


FIGURE R5 : Graphe de comparaison de la sélectivité Xe/Kr (à pression et température ambiantes pour une composition 20:80) prédite par le modèle et celle calculée par GCMC en échelle logarithmique. Les points colorés en bleu correspondent au jeu de données d'entraînement, tandis que ceux en orange correspondent au jeu de test. La superposition des points est faite de tel sort que l'on voit davantage les résultats sur le jeu de test pour évaluer la généralisabilité du modèle.

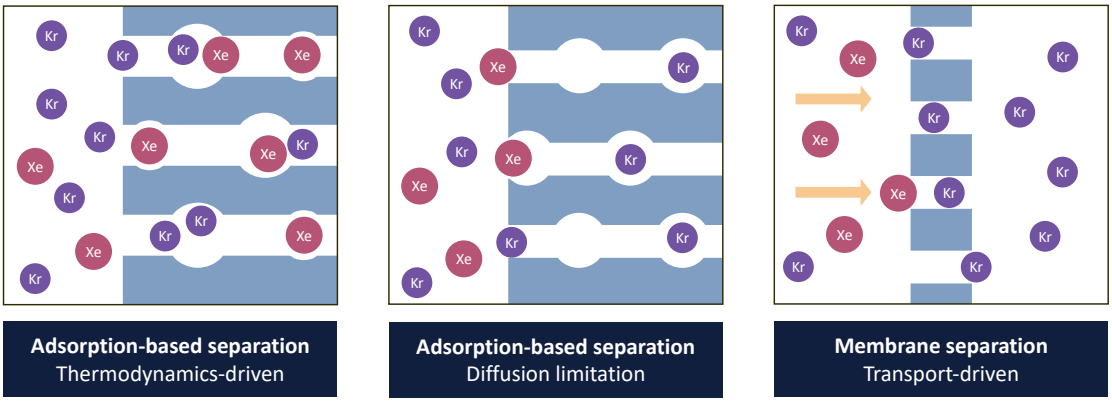
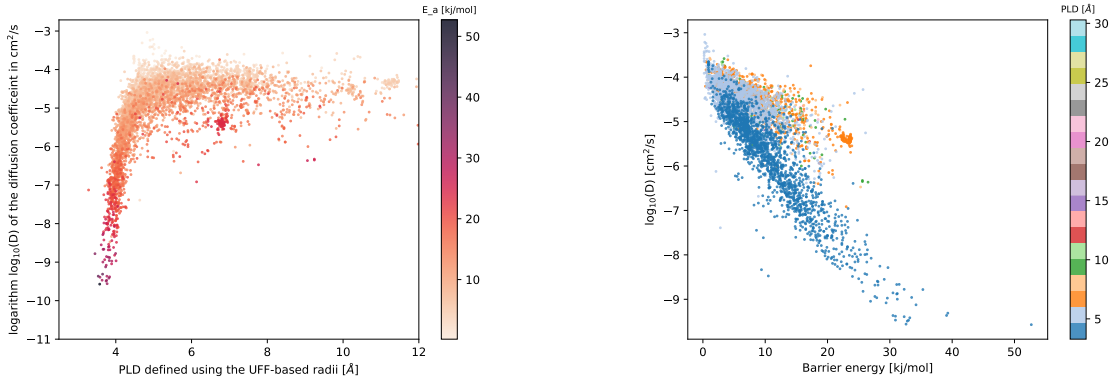


FIGURE R6 : Illustration of the comparative role of the thermodynamic and transport properties for Xe/Kr separation in nanoporous materials. From the transport dominated process of membrane separation to the thermodynamically equilibrated separation processes in the nanopores, different more nuanced cases could emerge where the diffusion imposes kinetic limitations.

Dans cette étude, de nombreuses corrélations ont été analysées et deux descripteurs semblent expliquer les valeurs du coefficient de diffusion. En effet, le diamètre de la plus petite sphère pouvant diffuser librement dans les canaux du matériau (PLD), une caractéristique structurelle facilement calculable, semble corrélé au logarithme du coefficient de diffusion. On peut distinguer deux régimes sur la figure R7a : une relation plutôt linéaire suivie d'un plateau.

Une mesure de la barrière énergétique de diffusion est également proposée en utilisant une grille calculée par l'algorithme GrAED. Cette barrière d'énergie semble inversement proportionnelle au logarithme du coefficient de diffusion comme on peut le voir sur la figure R7b. Cette



(a) Taille caractéristique des canaux *(b) Barrière énergétique de diffusion*

FIGURE R7 : (a) Graphe comparant le logarithme base 10 du coefficient de diffusion du xénon à la taille des canaux mesurée par le diamètre minimal du canal (PLD, en anglais), et les points sont étiquetés par les énergies de barrière. (b) Graphe comparant le logarithme base 10 du coefficient de diffusion du xénon à la barrière énergétique de diffusion, les points sont étiquetés par le diamètre PLD.

corrélation n'est pas très forte avec un coefficient de Pearson de l'ordre de −0.77 si on considère toutes les structures.

En suivant une approche similaire à celle utilisée pour prédire la sélectivité, le logarithme du coefficient de coefficient du xénon a été prédit en utilisant le diamètre PLD, la barrière énergétique et d'autres descripteurs thermodynamiques et structurels. Ce modèle de machine learning permet ainsi de remplacer la méthode coûteuse de dynamique moléculaire par l'utilisation de simulations moins coûteuses (barrière d'énergie et descripteurs structurels).

FIGURE R8 : (a) Comparaison du $\log_{10}$ *du coefficient de diffusion du xénon prédit par le modèle ML et les valeurs simulés par dynamique moléculaire. La racine de l'erreur quadratique moyenne sur le logarithme base 10 vaut* 0.25.

Comme on peut le voir sur la figure R8, le modèle semble bien prédire l'ordre de grandeur du coefficient de diffusion du xénon avec une erreur de l'ordre de 0.25 sur le $\log_{10}$ de ce coefficient. Cela signifie que l'on a une bonne connaissance de l'exposant du coefficient de diffusion exprimé comme une puissance de 10.

D'autres méthodes basées sur la théorie des états de transition ont également été explorées pour calculer le coefficient de diffusion. Ces travaux ont notamment mené à la conception de l'algorithme utilisé pour calculer des barrières d'énergie.

Enfin, en comparant les rapports des coefficients de diffusion du xénon et du krypton aux valeurs de la sélectivité Xe/Kr, il est possible d'identifier de nouveaux matériaux qui ne présentent pas de blocage cinétique tout en ayant une sélectivité très importante.

Conclusion

Cette thèse étudie ainsi les grandeurs thermodynamiques et cinétiques de la séparation xénon/krypton dans les matériaux nanoporeux. On a pu caractériser de manière plus fine les caractéristiques des matériaux les plus sélectifs. En ajoutant des contraintes sur la diffusibilité, différents types de matériaux ont été identifiés.

La prise en compte de phénomènes physiques non inclus dans les études de cette thèse ouvre des perspectives pour de nombreux travaux sur ce sujet. En effet, de nouveaux travaux expérimentaux montrent l'importance de la prise en compte de la polarisation. Le matériau le plus sélectif à ce jour pour la séparation xénon/krypton se base sur l'interaction induite par des métaux non coordinés.**Pei_2022** Enfin, la flexibilité du matériau est également importante à prendre en compte. Dans certains cas, la flexibilité permet même de mieux comprendre l'incohérence entre les résultats expérimentaux et théoriques. De nombreuses pistes peuvent être explorées pour intégrer à l'avenir ces effets dans un criblage.**Lachet_1998**, **Witman_2017**

RÉSUMÉ

Cette thèse se concentre sur l'amélioration de la séparation xénon/krypton en utilisant des matériaux nanoporeux. L'objectif est de développer des outils de description microscopique de ces matériaux en utilisant différents niveaux de modélisation moléculaire. Pour en évaluer rapidement les performances, des approches de criblage à haut débit et d'apprentissage statistique sont déployées en exploitant les bases de données existantes de matériaux nanoporeux. L'étude se concentre principalement sur la sélectivité Xe/Kr en utilisant des grandeurs thermodynamiques pertinents. Outre la sélectivité, d'autres propriétés importantes pour le procédé industriel de séparation de gaz, telles que la capacité d'adsorption et la vitesse de diffusion à l'intérieur des nanopores, sont également étudiées. Ces travaux de recherche contribuent à explorer des solutions plus efficaces et durables pour séparer efficacement des mélanges de gaz dans diverses industries.

MOTS CLÉS

simulation moléculaire, séparation de gaz, adsorption, matériaux nanoporeux, criblage haut-débit, apprentissage statistique

ABSTRACT

This thesis aims to improve the xenon/krypton separation using nanoporous materials. The primary objective is to develop microscopic characterization tools employing diverse levels of molecular modeling. High-throughput screening and statistical learning approaches are utilized, leveraging material databases to quickly assess their performances. Specifically, the Xe/Kr selectivity is investigated through a thermodynamically driven approach. Beyond selectivity, other relevant properties for gas separation, such as adsorption capacity and, more specifically, diffusion rates within nanopores, are studied. These research efforts contribute to the exploration of more efficient and sustainable solutions for gas separation in various industries.

KEYWORDS

molecular simulation, gas separation, adsorption, nanoporous materials, high-throughput screening , machine learning