EPFL

Thèse n° 7499

Advancing computational and data-driven methods for the design and discovery of nanoporous materials

Présentée le 23 avril 2020

à la Faculté des sciences de base
Laboratoire de simulation moléculaire
Programme doctoral en chimie et génie chimique

pour l'obtention du grade de Docteur ès Sciences

par

Seyedmohamad MOOSAVI

Acceptée sur proposition du jury

Prof. A. Züttel, président du jury
Prof. B. Smit, directeur de thèse
Prof. A. Tkatchenko, rapporteur
Prof. V. Van Speybroeck, rapporteuse
Prof. N. Marzari, rapporteur

École polytechnique fédérale de Lausanne

2020

Acknowledgements

The past years of my life in Switzerland and during the PhD studies have been absolutely influential in my life. These years combine many strong feelings and experiences from missing my family and home to learning how to engage with diverse cultures and yet grow simultaneously. Now that I look back and have a glance at my memories, I see a trajectory of events that bring me to where I am today. I like to take this occasion to acknowledge the people who were with me in this part of my life.

I would start with thanking my supervisor, Prof. Berend Smit, for all the support and mentorship during the past years. The first time, I contacted him in January 2014, I did not expect a response to a mechanical engineer who had not worked with a molecule in his life. He surprised me and arranged a meeting. The first thing I told him at the meeting was: “I don’t know molecular simulation, is that gonna work?”. He responded: “Neither did I, when I was born”. Of course, this was encouraging for me and I decided to work in his group. Now that I know him better, I do not get surprised with this kind of responses, as I know how encouraging Berend is, something that never stopped in the past years. Besides, I learned many other things from him, notably the way to formulate a scientific question, how to sometimes take a step back and look at the problem from a distance, and how to communicate science with other people. Berend’s attitude towards his group, particularly in being and feeling equal, will remain always in my mind.

During my PhD, I had the chance to collaborate with several cutting–edge research groups led by aspiring professors and scientists: Heather J. Kulik (MIT), Lev Sarkisov (Edinburgh), Jeffrey A. Reimer (Berkeley), Laura Gagliardi (Minnesota), Andy Cooper (Liverpool), Wendy Queen (EPFL), and Kyriakos Stylianou (EPFL). I thank all of them and their bright students whom I worked with for all the interesting discussions and guidance. Also, I am sincerely thankful to my PhD jury, Andreas Züttel, Nicola Marzari, Alexandre Tkatchenko, and Veronique van Speybroeck, for dedicating the time to read the thesis and travel to Sion despite all the other commitments they have.

I like to thank the people whom one can name "friend" which is of course a too general and vague term. These "friends" are truly amazing people and have impacted me every day mostly positively. The list will be too long for this short acknowledgment and also some need special thanks which I will do personally. Here, I name only a few: Senja Barthel, Mehrdad Asgari,

Acknowledgements

Saba Gharibzadeh, Sudi Jawahery, Fatemeh Rahimian, Efrem Braun, Kevin Jablonka, Maryam Farzan, Bardiya Valizadeh, Mohammad Tohidi Vahdat, Pete Boyd, Daniele Ongari, Sorour Darvishi, Sadegh Musapour, Soheil Hassanzadeh, Shahab Eghbali, Amin Niayifar, and Yongjin Lee. I like to thank all the people with whom I had a great time working with in the LSMO group and the EPFL Valais.

Lastly, I like to thank my great family, parents “Gholamhossein and Shahnaz”, siblings “Roshanak and Naser”, “Reza and Sahar”, and their amazing kids Arad, Aria, Aneseh, and Borhan. When I was a kid in school, my classmates were using phrases like "the best mom in the world" in their essays. At the time, it always sounded so strange to me to use such a phrase because you do not know other moms, how can you know yours is the best in the world! But now, I have seen many others and I would say I have the best family in the world. I thank them for the encouragement, support, . . . and in one word for being a “family”.

زندگی شوق رسیدن به همان فردایی‌ست ، که نخواهد آمد

Life is to be enthusiastic about a future that might never happen

Sion, 20 January 2020

Abstract

Metal–organic framework (MOFs) and related nanoporous materials have emerged as promising candidates for a variety of applications, such as gas separation and storage, catalysis, sensing, etc. Their building block structure allows us to generate a huge number of distinct materials only by changing the metal nodes and organic linkers. This, in principle, allows the design and discovery of materials that perform optimally for a given application. However, the conventional process of material development, from discovery to synthesis and performance evaluation, is too slow and expensive for exploring this enormous pool of materials. Complimentary methods are therefore needed to accelerate this process. The aim of this thesis is to investigate and expand the computational and data–driven methods for the development of nanoporous materials for gas separation and storage. The prevailing use of these methods is for high–throughput screening of the materials for a target application. However, the capability and success of such a screening approach depends on fast, reliable, and accurate prediction of material properties, as well as on the effective exploration of the chemical space. Therefore, in this thesis, we develop material descriptors for the chemistry and pore geometry of MOFs, which allow us to use machine learning to rapidly evaluate their adsorption properties with high accuracy. We next introduce a methodology to quantify the diversity of material databases to assess how well the chemical space is explored when a given material database is screened. We illustrate the importance of this diversity analysis by showing how the lack of diversity in MOF databases hinders material discovery, leads to chemical insights that are not generalisable, and makes machine learning models not transferable.

The promising materials discovered in a screening study are only of interest if they can be synthesised and are sufficiently stable to withstand the operating conditions of the corresponding application. Therefore, we investigate the applicability and capability of computational and data–driven methods to address some of the challenges in the material synthesis and the mechanical stability of MOFs. Material synthesis still mainly rests on the chemical intuition of synthetic chemists. Here, we introduce a method using machine learning and a genetic algorithm to capture this chemical intuition for MOF synthesis. We demonstrate how this simple approach can be powerful for guiding the synthesis of new materials. Lastly, we study how the mechanical stability of MOFs, which is fundamentally important for most of their practical applications, is affected by its underlying structure, i.e., the framework bonding topology and ligand structure. We show how this understanding can be used to develop strategies to design MOFs with enhanced mechanical stability.

Zusammenfassung

Metallorganische Gerüstverbidungen (engl. Metal–organic frameworks, MOFs) und verwandte nanoporöse Materialien sind vielversprechende Kandidaten für eine Vielzahl von Anwendungen, darunter Trennung und Speicherung von Gasen, Katalyse, Nachweisverfahren, etc. Ihre bausteinhafte Struktur ermöglicht es uns alleine durch das Austauschen der Metallcluster und der organischen Verbindungsstücke eine riesige Anzahl verschiedener Materialien herzustellen. Dies erlaubt prinzipiell das Entwerfen und die Entdeckung von Materialien, die ein ideales Verhalten für eine gewünschte Anwendung aufweisen. Der konventionelle Ablauf der Materialentwicklung, von der Entdeckung bis zur Synthese, ist jedoch zu langsam und zu aufwendig, um diese enorme Menge an Materialien zu erkunden. Daher werden ergänzende Methoden benötigt, um den Ablauf zu beschleunigen. Das Ziel dieser Arbeit ist es, rechnergestützte und datengetriebene (data–driven) Methoden zur Entwicklung von nanoporösen Materialien für die Trennung und Speicherung von Gasen zu untersuchen und weiterzuentwickeln. Diese Methoden werden vorwiegend in Hochdurchsatz Screenings (high–throughput screening) von Materialien mit Hinblick auf eine bestimmte Anwendung verwendet. Jedoch ist die Leistungsfähigkeit und der Erfolg eines solchen Screeningverfahrens von einer schnellen, verlässlichen und genauen Vorhersage der Materialeigenschaften sowie von einer repräsentativen Erfassung aller chemisch möglichen Materialien (chemical space) abhängig. Wir entwickeln daher in dieser Arbeit Größen, die die chemische Zusammensetzung von MOFs und die Geometrie ihrer Poren beschreiben, was es uns erlaubt maschinelles Lernen zur schnellen und hochgenauen Bestimmung ihrer Adsorptionseigenschaften zu verwenden. Im Anschluss führen wir eine Methodologie zur Quantifizierung der Diversität von Materialdatenbanken ein, um beurteilen zu können, wie umfassend die chemischen Möglichkeiten (chemical space) beim Screening einer gegebenen Datenbank erforscht werden. Wir verdeutlichen die Bedeutung dieser Diversitätsanalyse, indem wir zeigen wie ein Mangel an Diversität in einer MOF-Datenbank die Entdeckung von Materialien beeinträchtigt, zu chemischen Einsichten führt, die nicht zu verallgemeinern sind, und nicht übertragbare maschinelle Lernmodelle ergibt.

Die Materialien, die von einem Screening als erfolgversprechend gefunden wurden, sind nur dann wirklich von Interesse, wenn sie sowohl synthetisierbar als auch hinreichend stabil sind, um unter den Betriebsbedingungen der jeweiligen Anwendung bestehen zu können. Wir untersuchen daher die Anwendbarkeit und die Einsatzmöglichkeit von rechnergestützten und datengetriebenen (data–driven) Methoden, um einige der Herausforderungen der Materialsynthese und mechanischen Stabilität anzugehen. Die Synthese von Materialien

beruht noch immer im Wesentlichen auf der Intuition synthetischer Chemiker. Wir stellen hier eine Methode vor, die maschinelles Lernen und einen genetischen Algorithmus nutzt, um diese chemische Intuition zur Synthese von MOFs zu erfassen. Wir demonstrieren wie dieser einfache Ansatz die Synthese von neuen Materialien leistungsstark unterstützen kann. Abschließend untersuchen wir wie die mechanische Stabilität von MOFs, die von grundlegender Bedeutung für die meisten ihrer Anwendungen ist, von der zugrunde liegenden Struktur, d.h. der Topologie ihrer Gerüstverbindung und der Struktur der Verbindungsstücke, beeinflusst wird. Wir zeigen wie diese Einsicht zur Entwicklung von Strategien benutzt werden kann, die es erlauben MOFs mit verbesserter mechanischer Stabilität zu entwickeln.

Contents

Acknowledgements i

Abstract (English/Deutsch) iii

Introduction 1

1 On the importance of structural diversity in metal–organic framework databases 7
1.1 Introduction . . . 8
1.2 Development of descriptors for MOF chemistry . . . 8
1.3 Description of the databases . . . 10
1.4 Predicting adsorption properties of MOFs . . . 10
1.5 Diversity of MOF databases . . . 11
1.6 Applications of diversity analysis . . . 13
1.7 Discussion . . . 15
1.8 Methods . . . 17
1.9 Extended Data Set . . . 20
1.10 Supplementary materials . . . 29

2 Geometric Landscapes for Material Discovery within Energy–Structure–Function Maps 33
2.1 Introduction . . . 34
2.2 Geometric landscapes . . . 36
2.3 Energy-geometry landscapes . . . 40
2.4 Function-geometry landscapes . . . 43
2.5 Conclusions . . . 44
2.6 Methods . . . 45
2.7 Supplementary materials . . . 47

3 Capturing chemical intuition in synthesis of metal-organic frameworks 51
3.1 Introduction . . . 52
3.2 Synthesis and optimisation of the surface area of HKUST-1 . . . 53
3.3 Capturing chemical intuition using machine learning . . . 55
3.4 Application of learned chemical intuition . . . 56
3.5 Outlook . . . 58

3.6 Methods 60
3.7 Supplementary materials 62

4 Improving mechanical stability of MOFs using chemical Caryatids 69
4.1 Introduction 70
4.2 Results and discussion 71
4.3 Conclusions 78
4.4 Methods 80
4.4.1 Hypothetical material generation 80
4.4.2 Structure minimisation procedure 80
4.4.3 Calculation of the mechanical properties 83
4.4.4 Force field 85
4.5 Supplementary materials 85

Conclusions and Future Research 91

Curriculum Vitae 94

List of publications 99

Bibliography 101

Introduction

Computational and data–driven methods are established as complementary to the traditional empirical approach in science [1–5] and engineering [6–8]. These developments are build upon the significant progress in computational sciences and the explosion of the computing power [9], happened in the past decades, which enable us to simulate and optimise complex and multi–scale processes as well as to predict their outcomes by harvesting the generated data in fast, reliable and affordable ways. In this way, computational and data–driven methods make it possible to conceive and link solutions in multiple scales to develop new technologies we need urgently to address the challenges of our century.

These methods have specifically revolutionized the conventional empirical approach in chemical engineering, materials science, chemistry, and related fields. Computational methods have become a practical tool for simulating complex phenomena at atomic scale since the theories of statistical thermodynamics [10–12] and quantum mechanics [13, 14] have led to the emergence of molecular simulation [15, 16] and density functional theory [17–21]. For instance, nowadays, computational modelling is extensively used for *in silico* design and performance evaluation of materials and molecules, e.g., for design of drug molecules [22, 23], catalyst [24–26], adsorbent [27, 28], etc., and to understand and explain experimental observations and data, such as microscopic and spectroscopic data [29]. Data–driven methods are rising tools for finding complex patterns and correlations between materials'/molecules' structure, properties [30–32] and even synthesis [33–36] as we have access to an enormous amount of data provided by the growth of databases and data repositories, e.g., the Cambridge structural database [37] of synthesized compounds, and materials cloud [38] and NOMAD [39] repository for simulated data. Data–driven methods are increasingly used in a wide range of ways, from predictions of molecules and materials function [40–43], to autonomous synthesis in self–driving laboratories [36, 44, 45], and are even used in generative models [31, 46–48] for generating new molecules.

Computational and data–driven methods can be used for the design and discovery of new materials. This process often starts with defining the search space by representing the possible chemical space with a material database. These materials are consequently screened or analysed with respect to a target application. The gained information is then used for the design and discovery of promising candidates that are to be synthesized.

The aim of this thesis is to exploit both computational and data–driven methods for the design and discovery of nanoporous materials for gas separation and storage. Separating gas mixtures is energy intensive but highly demanded industrial process [49, 50]. It is estimated that chemical separation contributes around 10–15% of the world's energy consumption [51]. Therefore, it becomes urgent to improve these processes to address climate change to prevent its drastic consequences. Developing alternative less energy intense solutions can in principle contribute significantly in reforming our global energy landscape towards a sustainable future. Here, we focus mainly on two technologies, namely carbon capture [52] and methane storage [53]. These methodologies and tools are transferable to other separation and storage applications using solid adsorbents.

The prevalent processes that are currently deployed in industry for carbon capture and methane storage are amine scrubbing [54] and liquefied/compressed natural gas (i.e., LNG and CNG) [55], respectively. Using solid adsorbents is a promising alternative for both applications and could reduce the energy consumption considerably [56, 57]. However, the technological advancement for these applications relies on the development of new solid adsorbents with enhanced separation/storage performance and higher stability and processibility. To understand the characteristics of a material with high performance for separation and storage, it is instructive to take a simplified and idealized model for physisorption: In the Langmuir model the guest molecules do not interact with each other and the adsorption sites are all equivalent. In this model, the gas uptake of a porous material at a given pressure is a function of the number of adsorption sites and their strength. Already such a simplified picture unveils that an ideal material should have a high surface area (i.e., many adsorption sites) and tunable chemistry (i.e., tunable strength of sites). Metal–organic framework (MOFs) [58, 59] and related advanced nanoporous materials, including covalent organic frameworks (COFs) [60], zeolitic imidazolate frameworks (ZIFs) [61], and porous molecular crystals (PMCs) [62], are promising materials for this application since they can possess high surface area and are chemically tunable. For example, MOFs with ultra–high surface areas (to date, up to 7800 m^2/g [63]) have been synthesized. In addition, the chemistry of the materials can be tuned by modifying both the metal center and the organic linkers. Therefore, this thesis focuses on these classes of materials, in particular on MOFs. MOFs are formed by self–assembly of inorganic and organic building blocks on a three dimensional topological net (See Figure 1).

The search space for the computational design and discovery of MOFs is intrinsically enormous. The simple building block chemistry allows us to generate millions of possible materials that one would like to explore to find the optimal material for a given application [59, 64]. On this account, the use of computational and data–driven methods for the design and discovery of MOFs is essential as exploring this enormous chemical space experimentally is not feasible. Also, this large number of structures provides the ideal setting for using the data–driven methods to perform optimally. Over the last two decades, over 90'000 MOFs have been synthesized [65]. Yet, this is only a tiny fraction of all possible materials, and therefore, computational methods, which often are referred to as nanoporous material genomics, have been developed to generate hypothetical materials to better represent the possible chemical space [64, 66–68].

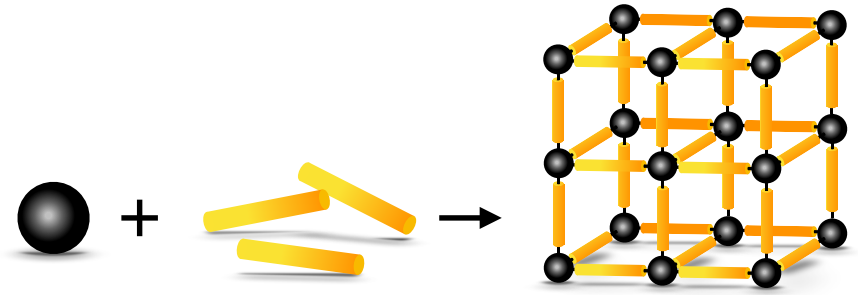


Figure 1 – A cartoon representation of metal–organic framework.

This material genomics approach has resulted in the generation of hundreds of thousands of hypothetical structures and together with the experimental structures, they constitute the search space for our computational and data–driven material discovery workflow. Indeed, many screening studies have been performed on these databases which led to the discovery of many promising materials for a variety of applications [69–71]. Moreover, useful information can be extracted by analysing the data generated from these screening studies [72, 73], for example, the identification of an "adsorbaphore", i.e., an ideal adsorption site for CO_2, by data mining of a large–scale screening study [69].

To screen these ever–growing databases as well as to optimally extract useful information by data mining, it is essential to develop material descriptors to encode and capture the complexity and diversity of the chemical structures. For the specific properties of our interest, i.e., the adsorption properties of nanoporous materials, both material chemistry and pore geometry play a significant role. We therefore start this thesis with presenting the development of descriptors to quantify similarity in material chemistry (Chapter 1) and pore geometry (Chapter 2, also our other publications that are not included in this thesis [74, 75]) of nanoporous materials. Using machine learning on these descriptors we could reach a remarkable accuracy in predicting the adsorption properties of the materials, demonstrating that the descriptors can effectively capture the structural similarity. In these two chapters we also describe how machine learning can be used to prescreen large materials databases and to extract structure–property relationships for our target gas adsorption applications, i.e., carbon capture and methane storage.

In Chapter 1, we show how these descriptors can be used to assess the quality of databases, in particular how well a database represents the chemical space. Since material databases are our representation of the chemical space, it is essential to know whether they truly represent it. Indeed, the success and capability of the screening studies is heavily dependent on how well and diverse the chemical space is explored. The notion of "diversity" in material databases is commonly linked to the fact that one would like to avoid screening a large number of similar structures. However, so far tools to quantify this diversity were missing. In Chapter 1, we introduce an approach for quantifying the diversity that describes how well the chemical space

is explored by a given database. This approach is build upon the success in capturing similarity in material chemistry and pore geometry using our descriptors. Analysing the diversity of several libraries of hypothetical and experimental MOFs, we show that each of the databases suffers from a lack of diversity. We illustrate how this lack of diversity has hindered material discovery, biased the extracted chemical insights and has led to non-transferable machine learning models.

While the discovery of promising structures for gas adsorption is the fundamental step in our computational and data–driven design or discovery workflow, it is evident that even the most promising candidate with respect to any given application will only be of interest if it first can be made, and second is stable enough to withstand the operational conditions. We address problems of these categories in the remaining chapters of the thesis.

In Chapter 3, we introduce a methodology to find the synthesis conditions for the synthesis of a given MOF. The synthesis of MOFs involves the selection of solvents and their composition, temperature, reaction time, etc. Indeed, MOF synthetic chemists use their intuition for selecting these variables since alternative strategies with no prior knowledge, e.g., grid search, would require an unfeasible number of trials due to the combinatorial nature of this search. However, this intuition is developed over years of experience by performing many successful and failed experiments. Specifically, the failed experiments remain unpublished and everyone can only learn from the own failures. We therefore developed a data–driven methodology to capture this intuition for MOF synthesis. This approach closely mimics how synthetic chemists develop their chemical intuition. We first used a genetic algorithm to efficiently explore the synthesis space which produces many failed and a few successful experiments. Consequently, we used machine learning to harvest chemical intuition from all failed and successful experiments. We showed that the chemical intuition that was obtained by our approach is transferable between systems. It furthermore allowed us to find new synthesis conditions for one of the most studied MOFs that resulted in its synthesis with the highest crystal quality reported to date.

In the last Chapter 4, we studied the mechanical stability of MOFs. While performance metrics based on adsorption properties are often used to screen MOFs, a sufficient mechanical stability is also necessary for most practical applications of the materials. However, the mechanical stability of MOFs had received very little attention [76–78]. That motivated us to developed tools to compute the mechanical stability of MOFs and rationalize how the structure of a material influences its mechanical stability. Given the structural tunability of MOFs, it is of particular interest to understand how to tune the mechanical stability to be able to design MOFs with enhanced mechanical stability.

New tools to compute the mechanical stability of MOFs were needed. An evaluation of the mechanical stability of MOFs requires the calculation of the full stiffness tensor as most MOF structures have less than simple cubic symmetry. However, performing quantum calculations for the mechanical properties become prohibitively expensive on a large number of structures,

in particular with MOF crystals typically having large primitive cells. Therefore, we circumvented this obstacle by developing tools to estimate the mechanical property of MOFs based on classical force fields [79]. We benchmarked their accuracy by comparing the mechanical stability of a set of materials with experimental and DFT data [80]. Later, we used these tools for studying structural flexibility [81, 82] and the mechanical stability [83, 84] of several MOFs (these studies are not included in this thesis).

To pinpoint the causes of mechanical instabilities in MOFs and to establish a structure–property relationship for the mechanical stability of MOFs, we present a study in Chapter 4 that revealed how the mechanical stability of a MOF material is related to its underlying structure, i.e., framework bonding topology and ligand structure. We systematically varied these parameters in a library of hypothetical MOFs and computed their mechanical stability. We found that the mechanical stability is primarily determined by the bonding network topology of the material. However, the functional groups on the organic linkers can modify these properties significantly by forming a secondary network. We showed that the optimal mechanical stability is achieved by a synergistic effect of the bonding network and the secondary network. We now are able to both compute the mechanical stability of a MOF and identify its cause.

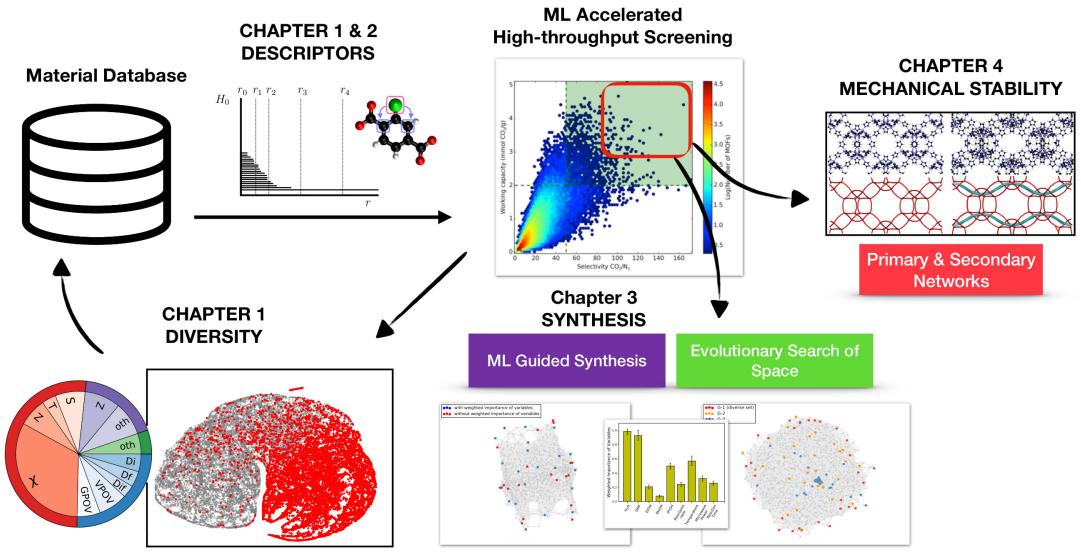
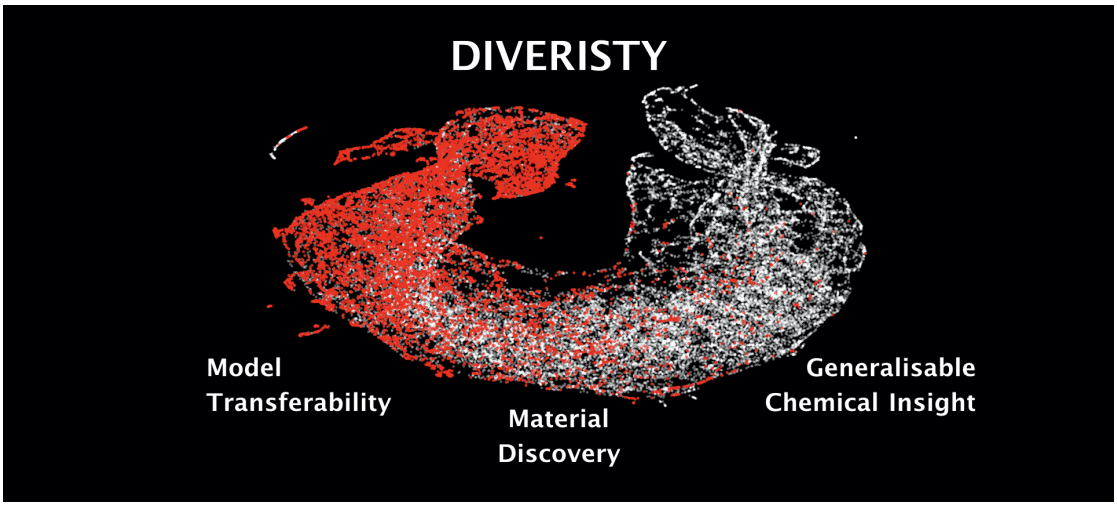


Figure 2 – Schematic of the components of this thesis and how they interact with the computational and data–driven workflow for discovery of nanoporous materials.

The contributions of this thesis to advancing the computational and data–driven methods for design and discovery of nanoporous materials are summarized and depicted in Figure 2. Starting with a material database, using the chemical and geometric descriptors, we can perform machine learning accelerated high–throughput screening of a database for the applications of interest. As a feedback loop, we can quantify the diversity of the database to know whether we

have explored the chemical space sufficiently. The discovered materials are then evaluated for their mechanical stability. Eventually, we can attempt to synthesise the promising candidates. The synergy between these sections can significantly accelerate the discovery of promising nanoporous materials for gas separation and storage applications.

1 On the importance of structural diversity in metal–organic framework databases[1]



[1] This chapter is a preprint version of a manuscript in preparation: Seyed Mohamad Moosavi, Aditya Nandy, Kevin Maik Jablonka, Daniele Ongari, Jon Paul Janet, Peter G. Boyd, Yongjin Lee, Berend Smit,and Heather J. Kulik, in preparation. SMM developed featurisation and machine learning workflows, collected the databases and refined them, performed gas adsorption calculations, and together with BS wrote the manuscript with the help of other coauthors.

Abstract

By combining metal nodes and organic linkers one can make millions of different metal-organic frameworks (MOFs) [59, 85]. At present over 90,000 MOFs have been synthesized [65, 86] and there are databases with over 500,000 predicted structures [64, 71, 87]. This raises the question of whether a new experimental or predicted structure adds new information. For MOF-chemists the chemical design space is a combination of pore geometry, metal nodes, organic linkers, and functional groups, but at present we do not have a formalism to quantify optimal coverage of chemical design space. In this work, we show how machine learning can be used to quantify similarities of MOFs. This quantification allows us to use techniques from ecology to analyse the chemical diversity of these materials in terms of diversity metrics. In particular, we show that this diversity analysis can identify biases in the databases, and how such bias can lead to incorrect conclusions. This formalism provides us with a simple and powerful practical guideline to see whether a set of new structures will have the potential for new insights, or constitute a relatively small variation of existing structures.

1.1 Introduction

The fact that we have an exponentially increasing number of different MOFs ready to be tested for an increasing range of applications opens many avenues for research. However, this rapid increase of data presents concerns over the chemical diversity of these materials. For example, one would like to avoid screening a large number of chemically similar structures. Yet, the way the number of materials evolves is prone to a lack of diversity [88, 89]. For example, one can envision an extremely successful experimental group focusing on the systematic synthesis of a particular class of MOFs for a specific application. Such successes may stimulate other groups to synthesise similar MOFs, which may bias research efforts towards this class of MOFs. In libraries of hypothetical MOFs, biases can be introduced by algorithms that favour the generation of a specific subsets of MOFs. At present, we do not have a theoretical framework to evaluate chemical diversity of MOFs. Such a framework is essential to identify possible biases, quantify the diversity of these libraries, and develop optimal screening strategies. The aim of this work is to introduce a systematic approach to quantify the chemical diversity of the different MOF libraries, and use these insights to remove these biases from the different libraries. The focus of our work is on MOFs as for these materials there has been an exponential growth of the number of studied materials. However, the question on how to correctly sample material design space holds for many classes of materials.

1.2 Development of descriptors for MOF chemistry

One of the aims of this work is to express the diversity of a MOF database in terms of features that can be related to the chemistry that is used in synthesizing MOFs as well as generating the libraries of hypothetical structures. At present, different strategies have been developed

to represent MOFs with feature vectors [74, 90]. However, the global material descriptors [73, 90, 91] that are presently used are not ideal for our purpose. We would like to directly connect to the structural building blocks of MOFs, which closely resemble the chemical intuition of MOF chemists, in which a MOF is a combination of the pore geometry and chemistry (i.e., metal nodes, ligands, and functional groups) [59, 92].

To describe the pore geometry of nanoporous materials we use simple geometric descriptors, such as the pore size and volume [93]. For the MOF chemistry, we adapt the revised autocorrelations (RACs) descriptors [94], which have been successfully applied [94–96] for building structure property relationships in transition metal chemistry [94, 97]. RACs are discrete correlations between heuristic atomic properties (e.g., the Pauling electronegativity, nuclear charge, etc.) of atoms on a graph. We compute RACs using the molecular or crystal graphs derived from the adjacency matrix computed for the primitive cell of the crystal structure (see method section). To describe the MOF chemistry, we extended conventional RACs to include descriptors for all domains of a MOF material, namely metal chemistry, linker chemistry, and functional groups (Figure 1.1 and the method section).

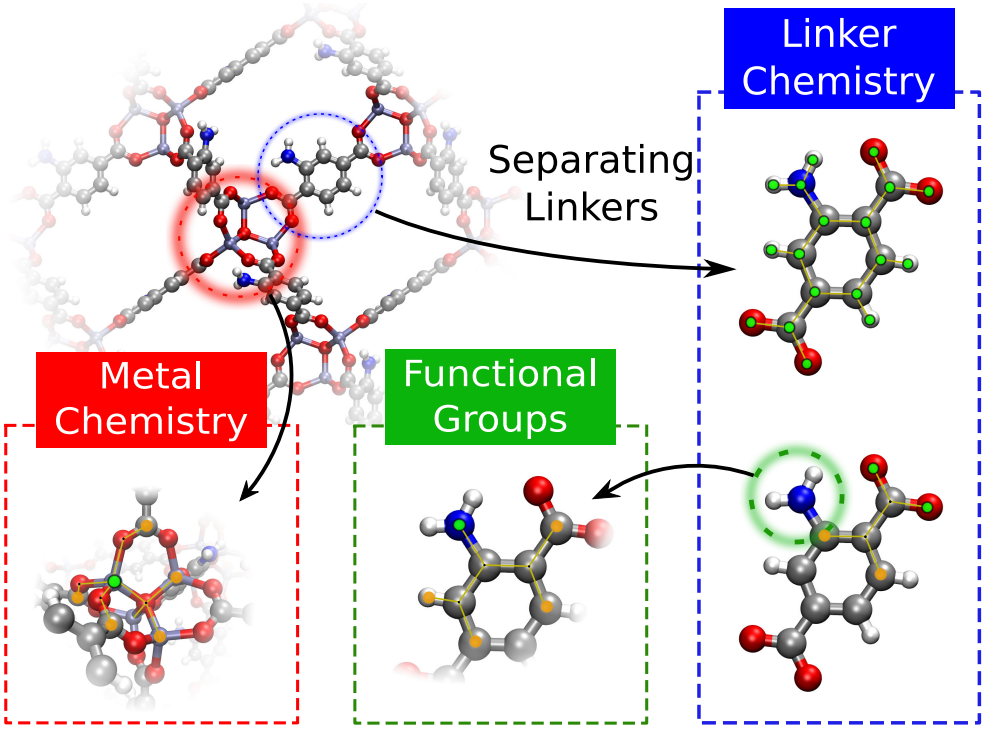


Figure 1.1 – **Description of the three domains of MOF chemistry.** Metal centre RACs are computed on the crystal graph. Linker and functional group RACs are computed on the corresponding linker molecular graph. Linker chemistry includes two types of RACs, namely full linker and linker connecting atoms. The graphs show the start atom (in green) and the nearby atom (in orange) used to define the RACs descriptors (see method section).

1.3 Description of the databases

We consider several MOF databases (see Extended Data Table 1.1): one experimental and five with *in silico* predicted structures. The Computation-Ready, Experimental (CoRE) [37,86,98,99] MOF database represents a selection of synthesised MOFs.

The first *in silico* generated MOF database (**hMOF**) was developed by Wilmer et al. [64] using a "Tinkertoy" algorithm by snapping MOF building blocks to form 130,000 MOF structures. This Tinkertoy algorithm, however, gave only a few underlying nets [100]. An alternative approach, using topology-based algorithms has been applied by Gomez-Gualdron et al. [71] for their **ToBaCCo** database (~13,000 structures), and by Boyd and Woo [69, 87] for their **BW-DB** (over 300,000 structures). A comprehensive review of this topic can be found here [66].

We use **CoRE-2019** and a diverse subset of 20,000 structures from the **BW-DB** (called **BW-20K**) to establish the validity of the material descriptors. In addition, a relatively small database of around 400 structures developed by Anderson et al. [73] (**ARABG-DB**) was included for comparison with their conclusions about importance of structural domains [73]. For this test, we focus on adsorption properties as their accurate prediction requires a meaningful descriptors for both the chemistry and pore geometry. We study the adsorption properties of methane and carbon dioxide. Because of their differences in chemistry (i.e. molecule shape and size, and non-zero quadrupole moment of carbon dioxide), designing porous materials with desired adsorption properties requires different strategies for each gas. To emphasize on these differences, we study the adsorption properties at three different conditions, namely infinite dilution (i.e. Henry regime), low pressure, and high pressure.

1.4 Predicting adsorption properties of MOFs

We will first establish that our descriptors capture the chemical similarity of MOF structures. As a test we show that instance-based machine learning models (kernel ridge regression (KRR)) using these descriptors can accurately predict adsorption properties. Extended Data Table 1.2 shows that the KRR models show good performance in predictions of the adsorption properties of **CoRE-2019** and **BW-20K** databases (see SI for parities and statistics). We observe that for those properties that are less dependent on the chemistry, e.g., the high pressure applications of CH_4 and CO_2, the geometric descriptors are sufficient to describe the materials with the average relative error (RMAE) in the prediction of the gas uptake being bellow 5%. In addition, if we compare the relative ranking of the materials, we also obtain satisfactory agreement as expressed by the Spearman rank correlation coefficient (SRCC) above 0.9. On the other hand, for the applications where chemistry plays a role, e.g., the Henry coefficient of CO_2, the chemical descriptors are essential to accurately predict the materials properties (RMAE ~5% and SRCC ~0.8). The significance of the chemical descriptors is also illustrated by the predictions of the maximum positive charge (MPC) and the minimum negative charge (MNC) of MOF structures (SRCC above 0.9 and 0.7, respectively). The geometric descriptors are nearly

irrelevant for these charges (SRCCs below 0.5 for all cases). This explains the relatively poor performance in prediction of CO_2 adsorption properties using only geometric descriptors as electrostatic interaction plays a crucial role. This analysis shows that our RACs and geometric descriptors are meaningful representations for the chemical space of MOFs for both CH_4 and CO_2 adsorption over the complete range of pressures. As a consequence, if two materials have similar descriptors, their adsorption properties will be similar. Hence, we can now quantify how the different regions of design space are covered by the different databases.

1.5 Diversity of MOF databases

We define the current chemical design space as the combination of all the synthesized materials and the *in silico* predicted structures, i.e., all the materials in the known databases. The real chemical design space, of course, can be much larger, as one can expect that novel classes of MOFs will be discovered. It is instructive to visualize how each MOF database is covering the current design space. This design space, as described by our descriptors, is a high-dimensional space and to visualize this we make a projection on two-dimensions.

The projection of the pore geometry of our current design space is shown in Fig. 1.2a. The color distribution shows a gradient in the pore size of the MOFs, from small to large pores moving on the map from left to the right. Other panels show how the different MOF databases are covering this space. The distributions of the geometric properties of the databases are considerably different from each other. For example, the experimental MOFs (**CoRE-2019**) are mainly in the small pore region of the map. Remarkably, the hypothetical databases also have very different distributions. While **BW-DB** covers the intermediate pore size regions, **ToBaCCo** is biased to the large pore regions of the design space.

The hypothetical structures have been generated to explore the design space of MOFs beyond the experimentally known structures. In Fig. 1.3 we show how these databases are covering the design space (see Extended Data Fig. 1.2 for the distribution of each database and Extended Data Fig. 1.1 for PCA). We use diversity metrics [102] to quantify the coverage of these databases in terms of variety (V), balance (B), and disparity (D). The pore geometry, linker chemistry, and functional groups design spaces are well covered and sampled by the hypothetical databases. However, we observe a serious limitation in diversity, in particular in the variety of the metal chemistry in hypothetical databases (Fig. 1.3b). Compared to the experimental database, the variety of the metal chemistry of MOFs by hypothetical databases is surprisingly low; only a limited number of MOF metal centres are present (18 metal SBUs for all hypothetical databases, see SI).

The choice of the organic linker and the placement of functional groups are readily enumerated; one can take the large databases of organic molecules [103] as a rich source of the possible MOF linkers or functional groups. In contrast, the metal nodes of MOFs are typically only known after a MOF is synthesised. For example, at present we cannot predict that if Zinc atoms during the MOF formation would cluster in a Zinc paddle-wheel (e.g., in

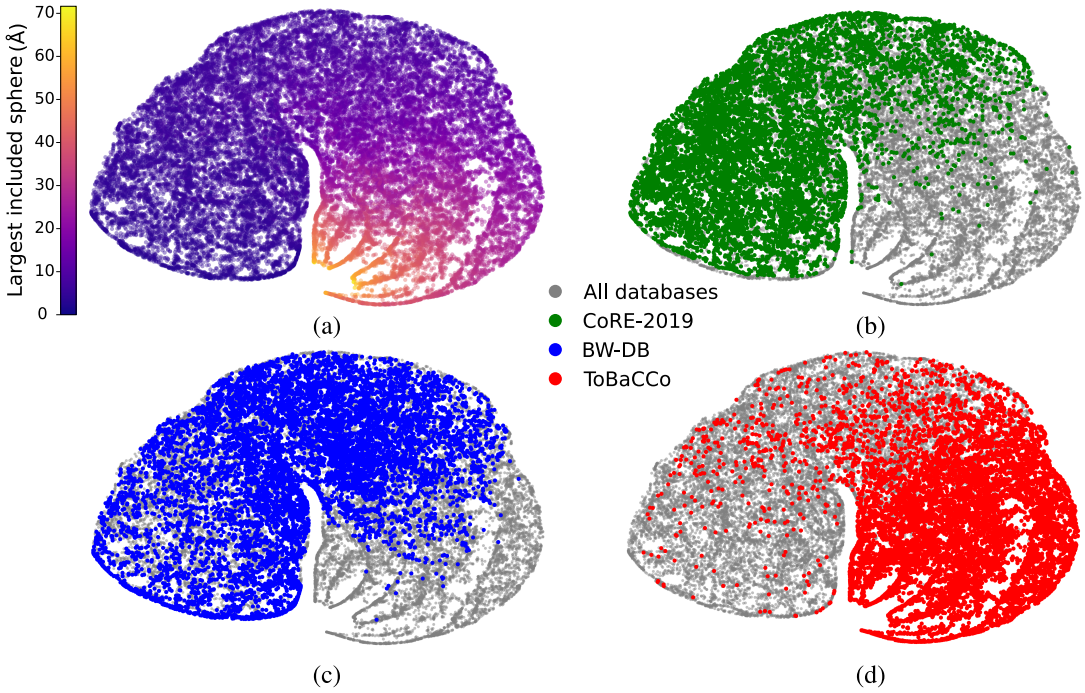


Figure 1.2 – **Map of the pore geometry of MOFs.** To project the geometric descriptor space of MOFs to a 2D map we use the t-Distributed Stochastic Neighbour Embedding (t-SNE) [101] method (see Extended Data Fig. 1.1 for principal component analysis (PCA)). The t-SNE method preserves pairwise distances, ensuring similar structures are mapped close to each other in two dimensions. (a) The current design space colour coded with the largest included sphere. In (b), (c), and (d), the green, blue, and red dots are representing the materials in the **CoRE-2019**, **BW-DB**, and **ToBaCCo** databases, respectively, which are overlaid on the design space represented in grey. See Extended Data Fig. 1.1 for PCA which show a similar distribution of databases.

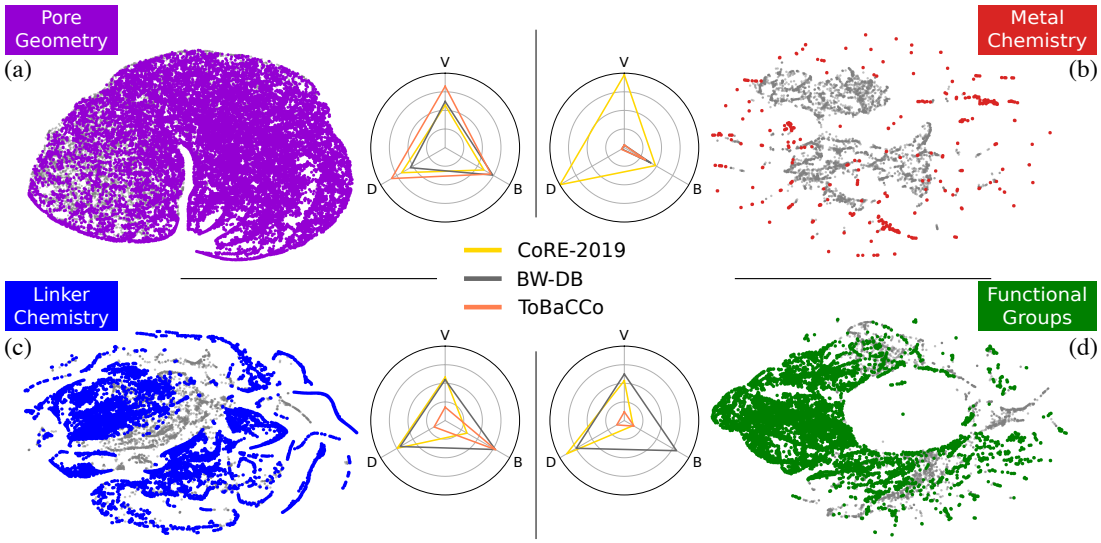


Figure 1.3 – **Diversity metrics and maps of different domains of MOF structures.** The t-SNE method was used to project the (a) pore geometry, (b) metal chemistry, (c) linker chemistry, and (d) functional groups descriptor spaces to 2D maps. Only descriptors up to the second coordination shell were included for metal chemistry to emphasize the local metal chemistry environment. In each panel, the structures from the hypothetical databases are coloured and overlaid on the entire known design space represented in grey. The radar charts show the three diversity metrics: variety (V), balance (B), and disparity (D), for the three databases. For this analysis, first we discretize the space into a fixed number of bins. Variety measures the number of bins that are sampled, balance the evenness of the distribution of materials among the sampled bins, and disparity the spread of the sampled bins.

Zn-HKUST-1), [104] a single node (e.g., in ZIFs) [61], Zn_4O (e.g., in IRMOFs) [59], or to a totally new configuration. Therefore, we expect that there are many missing points on the metal chemistry map in Fig. 1.3b which will be found in the coming years.

1.6 Applications of diversity analysis

We illustrate the importance of quantifying the diversity of the different databases by three examples. The first example illustrates how machine learning can be used to extract insight on how the performance of a material is related to its underlying structure [73, 94, 96]. As our descriptors represent each domain of the MOF architecture, we can quantify the relative importance of these domains on CH_4 and CO_2 adsorption.

Within each database, the importance of variables varies significantly across different gases and different adsorption conditions (see Extended Data Fig. 1.3 and 1.4). These results follow our intuition; the chemistry of the material is more important in the low pressure regime, while at high pressures the pore geometry is the dominant factor. Moreover, we observe that material chemistry is more important for CO_2 than for CH_4 adsorption.

If each of these databases would have covered a representative subset of MOF chemistry, one would expect that each database would give a similar result for the importance of the different variables. However, we observe striking differences when we compare across different databases, which indicates that there are biases in the different databases. An illustrative example is CO_2 adsorption at low pressure. Anderson et al. [73] concluded from their analysis of the (**ARABG-DB**) database that the metal chemistry is not an important variable for CO_2 adsorption. However, Extended Data Fig. 1.5(a) shows that for each of these databases *different* material characteristics are important for the models in predicting CO_2 adsorption. For example, pore geometry is the most important variable in the **BW-20K**, while metal chemistry in **CoRE-2019**, and the functional groups in **ARABG-DB**. The reason why metal chemistry was not identified as an important factor by Andersen et al. was that metal chemistry was not explored sufficiently in their database as only four SBUs were used for structure enumeration.

In our second example, we focus on how our diversity analysis can help us to identify opportunities for the design of new structures. At present, there are over 90,000 MOFs that have been synthesised and one would like to be sure that MOF 90,001 adds relevant information. Similarly for the hypothetical databases one would add new structures to any screening study only if they are complementary to the many that already exist.

For CO_2 capture from flue gases, which corresponds to CO_2 adsorption at low pressure in our study, we have shown that metal chemistry cannot be ignored (Extended Data Fig. 1.5a). Our diversity analysis shows that this domain is not well covered by hypothetical databases (see Fig. 1.3). Therefore, exploring different metal chemistries in these databases would increase the diversity of these databases. For this we have developed a methodology to mine unique MOF building blocks from the experimental MOF databases (see method section). Extended Data Fig. 1.6 show some of these SBUs that have not been used for structure enumeration in these hypothetical databases yet, and including these missing structures in a screening study could lead to the discovery of materials with superior performance.

For methane storage our analysis shows that the single most important factor is the pore geometry (see Extended Data Fig. 1.5b). All databases confirm that pore geometry is the most important factor. For this application, each of the databases have a sufficient diversity in geometric structures and other factors do not matter. This observation provides an important rationale for the provocative conclusion of Simon et al. [105] that there is no point in looking for new structures for methane storage as they are not expected to perform significantly better for this application. Simon et al. arrived at this conclusion from a large screening of 650,000 random selection of structures from many databases of different classes of nanoporous materials. Our study shows that indeed a large selection of structures from different databases will cover the entire geometric space of the current design space. To significantly outperform the best performing materials one would need a completely new chemistry and mechanism, e.g., framework flexibility. [106]

In the final example, we focus on the effect of bias in the databases on the generalisability and

transferability of machine learning predictions. Intuitively, one would expect that if we include structures from all regions of the design space in our training set, our machine learning results should be transferable to any database. We illustrate this point for the two databases **CoRE-2019** and **BW-DB**. We randomly select 2,000 structures that we use as test set. A diverse set of structures based on the chemical and geometric descriptors was obtained from the remaining structures in these two databases [107]. The accuracy of random forest models trained using this diverse set is compared with the models trained using training sets from each database in Extended Data Fig. 1.7. Clearly, the models that were trained on databases which are biased to some regions of the design space result in poor transferability for predictions in unseen regions of the space. In contrast and not surprisingly, the model that is trained with a diverse set performs relatively well for both databases.

1.7 Discussion

An interesting side effect of MOF chemistry is that the enormous number of materials makes this field ideal for big-data science. This development raises all kinds of new, interesting scientific questions. For example, we have now so many experimental and hypothetical materials that brute-force simulations and experiments are only feasible on a subset of materials. Hence, it is essential that this subset covers the relevant chemistry as optimally as possible. In this work, we have developed a theoretical framework on how to arrive at the most diverse set of materials representing the state of the art of MOF chemistry.

Our framework relies on the notion that for chemists the chemical design space of MOFs is a combination of pore geometry, metal nodes, organic linkers, and functional groups. By projecting a material on a set of relevant descriptors characterizing these four domains of MOF chemistry, we can quantify the diversity of databases. Adding structures that increase the diversity metrics, implies that these structures add new information to the database. Given that there are already so many materials and databases, there is a need for a simple and powerful practical guideline to see whether new set of structures will have the potential for new insights, or are relatively small variations of existing structures. Analysis of the diversity can also give us insights in parts of the chemical design space that are not fully explored. An interesting historical perspective is shown in Fig. 1.4, in which we plot as metric of novelty of the discovered materials the distance to the geometry descriptor of the previously discovered materials. The jumps in the graph nicely identifies structures that opened a new direction of MOF research [85, 108–112], where 2012 was an exceptionally good year, which include the discovery of the IRMOF-74 [112] series and the material with the lowest density [110] and highest surface area [108] at their time.

One cannot separate diversity from the application. For example, if one is interested in the optical properties of MOFs, which largely depends on charge transfer between metal and ligand species, diversity in pore geometry might not be that important, and for such a screening study the optimal representative set of materials will be different from say, a

gas adsorption study. Yet, the same procedures to generate such a diverse set can be used provided that the properties depend sufficiently gradual on the relevant descriptors. If one has a property that dramatically changes by a slight change of the structure of the MOF, our method would flag these structures as similar while the properties are in fact very different. Of course, once such property is identified one can re-weight the measure of similarity to ensure that those aspects of the descriptors that can distinguish these materials carry more weight.

MOF chemistry is not a static field, new classes of MOFs will be constantly developed. The protocol that was introduced in this work can be (trivially) extended in the future to include these new MOFs as they get reported, allowing to always generate a set of most diverse structures that is representative of the whole database of known structures.

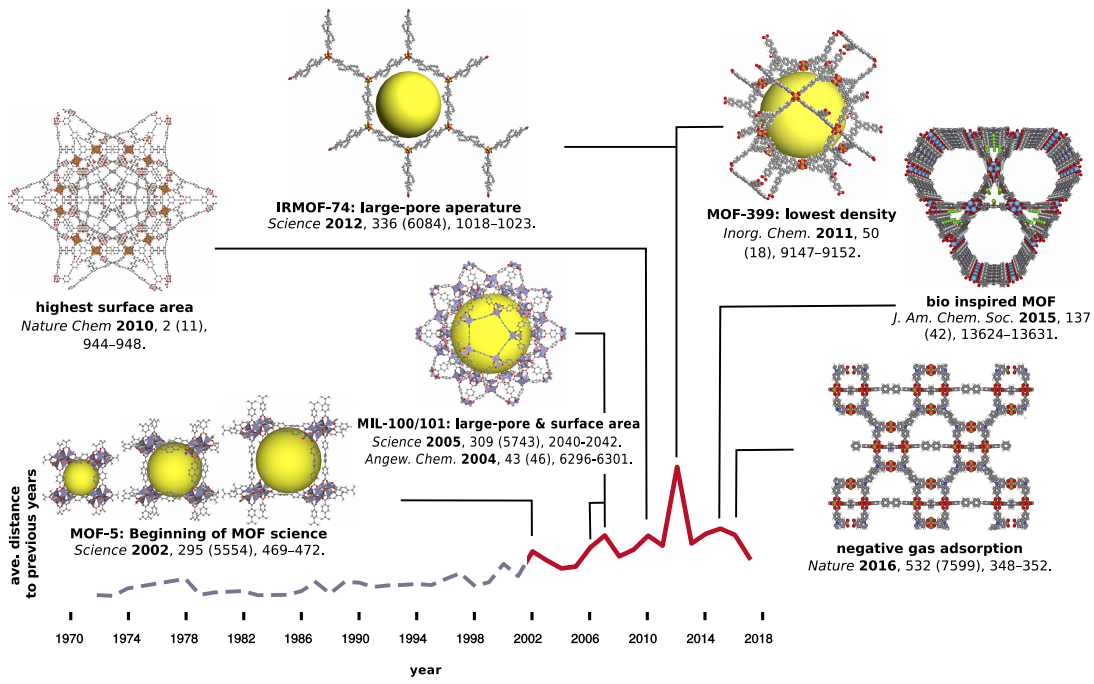


Figure 1.4 – **Timeline of evolution of MOF geometry.** For each year, the average of relative distance in the geometry descriptor space to the MOFs reported in Cambridge structural database (CSD) [37] in the preceding years is shown with red line. The MOFs with largest distance for some of the peaks are shown in the inset [85, 108–114]. The years on the timeline are corresponding to the year that a structure has been deposited in CSD. The gray line shows the coordination polymers reported in CSD before the beginning of the MOF chemistry as a separate field of research, shown in red.

1.8 Methods

RACs for MOFs

RACs [94] are products and differences on the graph of heuristic atomic properties. RACs were first introduced for machine learning open shell transition metal complex properties [94,95,97]. The relative importance of heuristic properties proved valuable for interpreting structure-property relationships and similarity of these transition metal complexes [96]. We have devised an approach to extend RACs to periodic MOF materials by dividing MOFs into their constituent parts. A typical [94] difference-based RAC correlation is computed on the graph representation of the structure using:

$$ {}^{start}_{scope}P_d^{diff} = \sum_i^{start} \sum_j^{scope} (P_i - P_j)\delta(d_{i,j}, d). \tag{1.1} $$

In this equation, atomic property P of atom i selected from *start* atom list is correlated to atom j selected from *scope* atom list when they are separated by d number of bonds. To devise MOF chemistry-specific RACs, we extend the concepts of *start* and *scope* introduced [94] for metal-centered and ligand-centered RACs in transition metal complexes. Two atom lists, namely *start* and *scope*, are needed to compute these RACs (equation 1.1). For the metal centred RACs, we use the crystal graph as the *scope* atom list and the *start* atom list only includes each of the metals in an SBU (see SI for full list). These RACs thus emphasize the metal and SBU contributions to MOF chemistry and property prediction. In describing linkers and functional groups, we use RACs computed on the molecular graph of the corresponding linker. In this approach, we only correlate atoms on the same linker, and therefore, the *scope* atom list includes all the atoms from the same linker of the starting atoms. To construct the molecular graph for each linker, we start by splitting the MOF to the corresponding linker lists. Removing the metals from the crystal graph gives us a set of floating connected components. We remove the atoms that are only bonded to the metals and/or hydrogens, e.g., the bridging oxygen in Zn_4O, and the corresponding hydrogen that are connected to these atoms, leaving us with only the organic linkers and the coordinated organic molecules to the metal centres. By separating the subgraphs of these connected components, we obtain the molecular graph for each linker. Linker chemistry is described with two *start* atom lists, including full linker and linker connecting atoms. Full linker atom list includes all the atoms on the linker. Linker connecting atoms are the atoms that have a chemical bond with a metal center. Lastly, any atom on a linker that is not a carbon or hydrogen atom, and is not linker connecting atom is assigned to be a functional group and is included in the *start* atom list for functional group descriptors. Note that Carbon based functionalisations, e.g., methyl functionalisation, would not be identified as a functional group in this approach.

Similar to applications of RACs on transition metal complexes [94–96], five heuristic atomic properties, including atom identity (I), connectivity (T), Pauling electronegativity (χ), covalent radii (S), and nuclear charge (Z) were used to compute RACs. To this set, we add polarisability

(α) of atoms for the linker descriptors as suggested [73] to be an important factor for gas adsorption properties of MOFs. These properties are used to generate metal centred, linker, and functional group descriptors. Lastly, we take the averages of these descriptors to make a fixed length descriptor. In total, this analysis produces 156 features (see SI for details).

Mining building blocks

The approach explained in the previous section can correctly identifies the organic SBUs. However, rigorously recognising inorganic SBUs is challenging, requires advanced methods, and might be dependent on the crystal graph simplification method [92]. In this study, we leverage a RACs to mine inorganic SBUs specific to our data set. We make an atom list including metal centres and their first and second coordination shells. We extract inorganic SBUs by separating all connected subgraphs after removing all the atoms which are not included in this list from crystal graph. Finally, we identify unique organic and inorganic SBUs by removing all isomorph labelled molecular graphs using Cordella et al.'s [115] approach as implemented in NetworkX [116].

Molecular simulation

The adsorption properties of the materials were computed assuming rigid frameworks. The guest-guest interactions and host-guest interactions were modelled using Lennard-Jones potential truncated and shifted at 12.8 Å and Coulombic electrostatic interactions computed by Ewald summation. The force field parameters of the framework atoms and gas molecules were extracted from UFF and TraPPE force fields, respectively (see full list of parameters in SI), using the Lorentz-Berthelot mixing rule for pairs. Partial atomic charges of framework atoms were generated using EQeq [117]. Grand canonical Monte Carlo and Widom insertion were used to compute gas uptake and Henry coefficient of the materials, respectively. Each calculation consists of 4000 initialisation cycles followed by 6000 equilibrium cycles. All the gas adsorption calculations were performed in RASPA [118]. Adsorption properties were computed at 0.15 bar (5.8 bar) and 16 bar (65 bar) for CO_2 (CH_4) for low and high pressures, respectively. All adsorption calculations were performed for room temperature. The pore geometry was described using eight geometric descriptors, namely largest included sphere (D_i), largest free sphere (D_f), largest included sphere along free path (D_{if}), crystal density ρ, volumetric and gravimetric surface area and pore volume. The geometric descriptors were computed using Zeo++ [93, 119], using a probe radius of 1.86Å.

Machine learning

Random forest regression (RF), gradient boosting regression (GBR), and kernel ridge regression (KRR) models were used in this study. All computations were performed in scikit-learn [120] machine learning toolbox in python.

The hyperparameters for GBR and RF models were chosen by grid search optimisation using 10-fold cross-validation (CV) minimising the mean absolute error (see SI for the range of hyperparameters). For the KRR models, we first perform feature selection. Both recursive feature addition (RFA) and explained variance threshold methods were used to find the the feature subset that minimises the 10-fold CV mean absolute error of the model. For the RFA method, the order of feature addition was done based on the importance of features derived from the random forest mean decrease in impurity importance of variables following the strategy in Ref. [97]. The hyperparameters of the KRR models were chosen by minimising the 10-fold CV score of the model using a mixed optimisation methods, including Tree of Parzen Estimators (TPE), annealing and random search, using the hyperopt [121] package.

The features and labels were centred to zero and scaled using their mean and standard deviation, respectively. Train-test splitting was performed randomly and the size of the train sets are mentioned in the caption of each parity plot or table in the main text and the SI. All the statistics reported were computed by averaging over 10 different random seeds used for train-test splitting except in the figures for transferability of models between databases where fixed test sets were used.

The relative importance of variables were computed for the random forest models. Three different approaches were used to derive the feature importance (see SI for comparison). The first approach is based on the mean decrease in impurity (Gini importance) which is computed while training a random forest regression. The second and third approach are permutation importance and SHapley Additive exPlanations (SHAP) [122], respectively, which were computed for the test or train set.

Diversity metrics

To compute the diversity metrics, we first split the high-dimensional spaces into a fixed number of bins by assigning all the structures to their closest centroid found from k-means clustering. Here, we use the percentage of all the bins sampled by a database as the variety metric. Furthermore, we use Pielou's evenness [123] to measure the balance of a database, i.e., how even the structures are distributed among the sampled bins. Other metrics, including relative entropy and Kullback-Leibler divergence are a transformation of Pielou's evenness and provide the same information (see SI for comparison). Here, we use 1000 bins for these analyses (see sensitivity analysis to the number of bins in SI). Lastly, we compute disparity, a measure of spread of the sampled bins, based on the area of the concave hull of the first two principal components of the structures in a database normalized with the area of the concave hull of the current design space. The areas were computed using Shapely [124] with circumference to area ratio cutoff of 1.

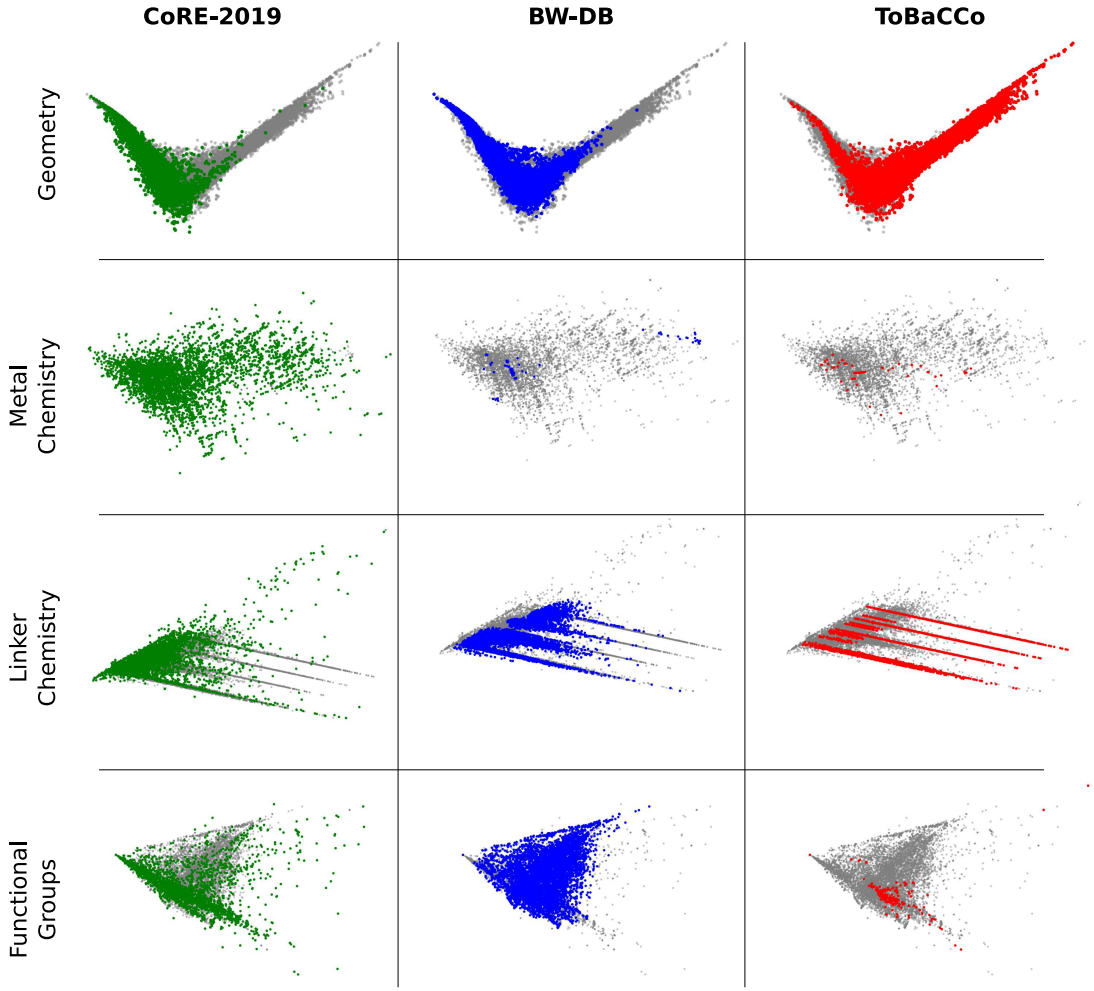
1.9 Extended Data Set

Extended Data Table 1.1 – **Material databases.** The list of the databases investigated in this study. The bold letters are the name that are used to refer to each database throughout the article.

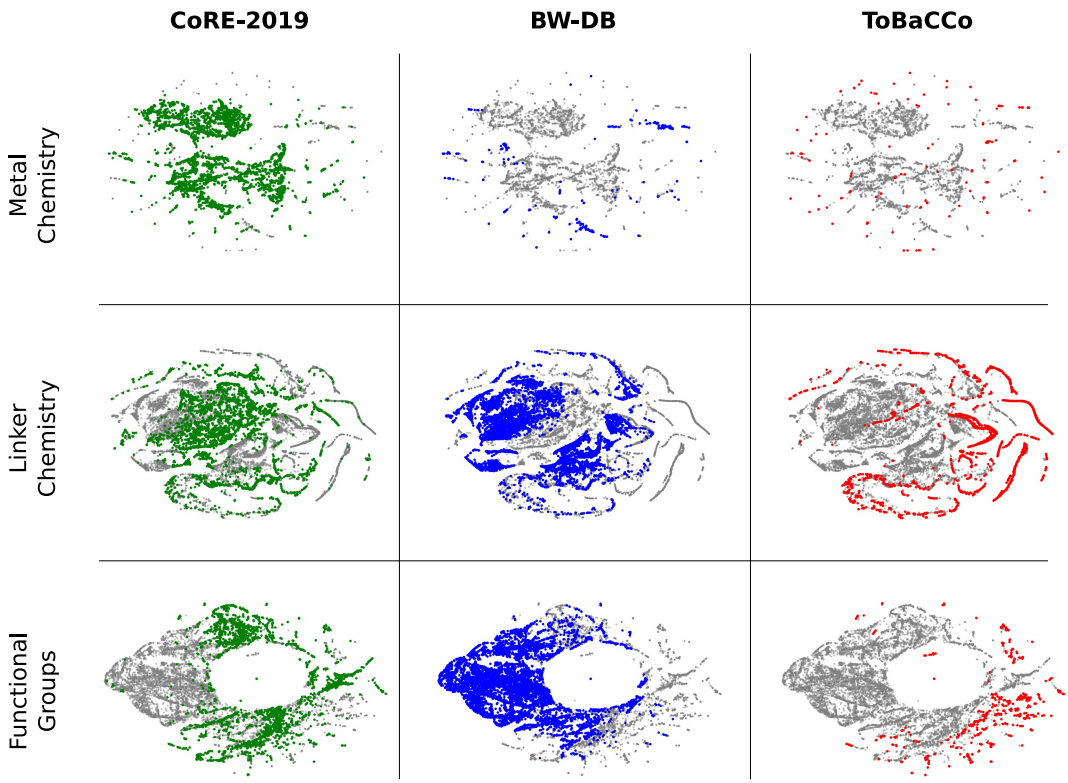
Name	type	number of structures	Notes and References
CoRE-2019	experimental	~12,000	Computational Ready, Experimental (CoRE) MOFs initially developed [98] and later extended by Chung et al. [86]
CoRE-DDEC	experimental	~3,000	The refined subset of CoRE–MOF database [98] with DDEC partial atomic charges developed by Nazarian et al. [99]
hMOF	hypothetical	~130,000	hypothetical MOFs generated by Wilmer et al. [64]
BW-DB	hypothetical	~300,000	hypothetical MOFs generated by Boyd et al. [69, 87]
ToBaCCo	hypothetical	~13,000	hypothetical MOFs generated by Gomez-Gualdron et al. [71]
BW-20K	hypothetical	~20,000	a diverse subset of structures from **BW-DB** database
ARABG-DB	hypothetical	~400	hypothetical MOFs generated by Anderson et al. [73]

Extended Data Table 1.2 – **Accuracy of machine learning predictions of gas adsorption properties of MOFs.** For each database, KRR models were trained using ~7,000 training data points randomly chosen from the database and the remaining structures were used for testing. The numbers are reported for the test set prediction. KRR models use similarity in terms of pairwise distances in feature space for predictions. The statistics are reported as the average over 10 separate train-test splits. Henry coefficient (k_H), gas uptakes and deliverable capacity for CH_4, and gas uptakes for CO_2 are reported in $mol.kg^{-1}.Pa^{-1}$, vSTP/v, and $mol.kg^{-1}$, respectively. MAE: mean absolute error; RMAE: relative mean absolute error (%), and SRCC: Spearman ranking correlation coefficient.

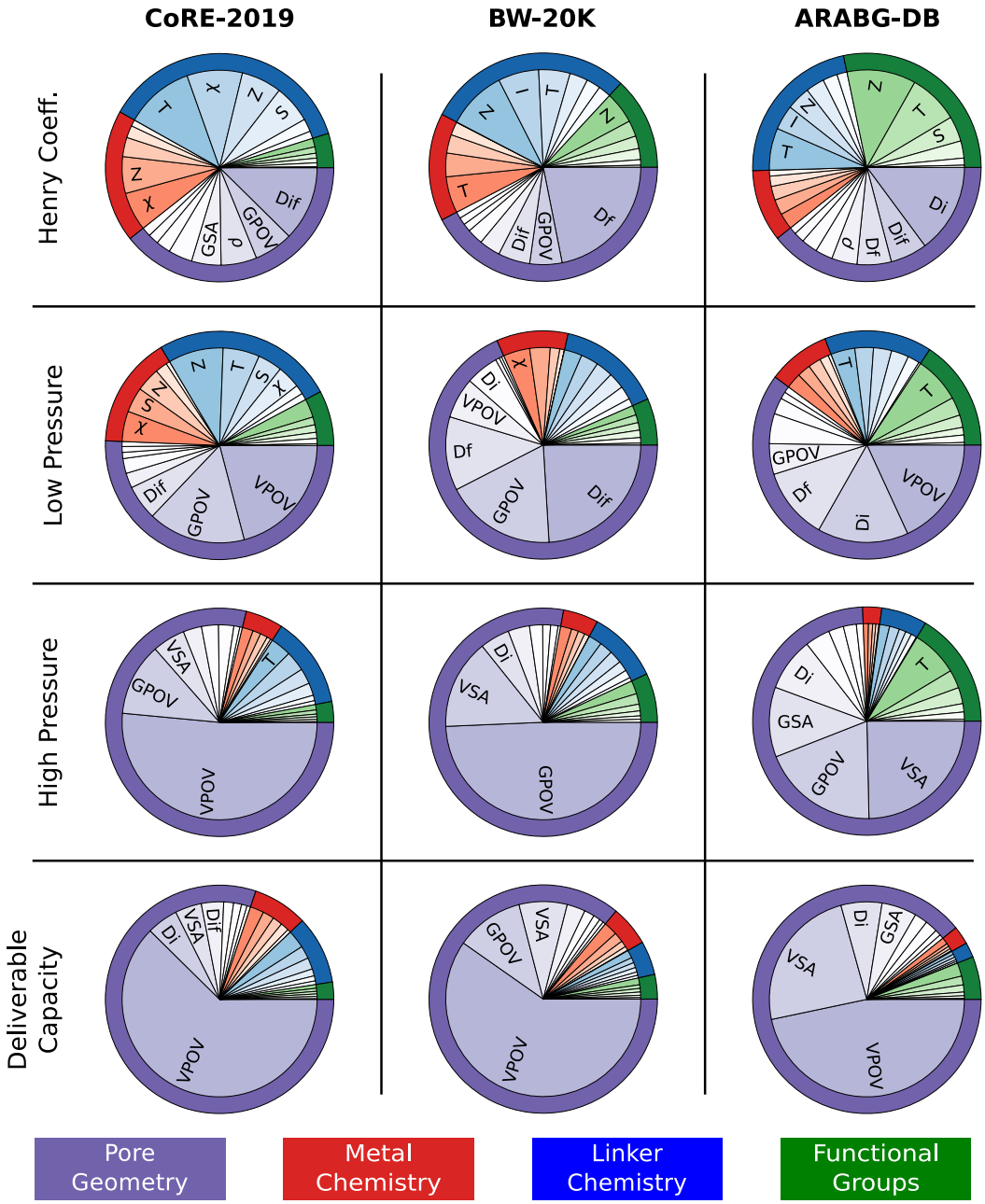
	property	Database	Geo. Descriptors			Geo.&Chem. Descriptors		
			MAE	RMAE	SRCC	MAE	RMAE	SRCC
CH_4	$\log(k_H)$	**CoRE2019**	0.29	4.77	0.67	0.20	3.26	0.84
		BW-20K	0.17	4.24	0.79	0.14	3.35	0.87
	Upt@5.8 bar	**CoRE2019**	19.59	7.34	0.75	12.94	4.85	0.88
		BW-20K	11.54	6.21	0.90	8.80	4.74	0.94
	Upt@65 bar	**CoRE2019**	19.94	5.36	0.92	16.64	4.47	0.94
		BW-20K	14.35	4.90	0.93	10.88	3.72	0.96
	Del. Cap.	**CoRE2019**	14.78	5.15	0.90	13.71	4.78	0.91
		BW-20K	9.90	4.39	0.97	9.90	4.39	0.97
CO_2	$\log(k_H)$	**CoRE2019**	0.74	8.29	0.50	0.51	5.64	0.77
		BW-20K	0.31	4.60	0.82	0.24	3.57	0.89
	Upt@0.15bar	**CoRE2019**	0.92	9.90	0.57	0.57	6.12	0.81
		BW-20K	0.43	5.21	0.83	0.30	3.59	0.92
	Upt@16.0bar	**CoRE2019**	0.83	2.46	0.96	0.65	1.92	0.97
		BW-20K	1.15	3.33	0.98	0.74	2.16	0.99
Charge	MPC	**CoRE2019**	0.28	10.2	0.44	0.07	2.54	0.93
		BW-20K	0.11	4.92	0.63	0.05	2.01	0.90
	MNC	**CoRE2019**	0.17	5.98	0.30	0.11	3.97	0.75
		BW-20K	0.10	3.88	0.35	0.07	2.70	0.71



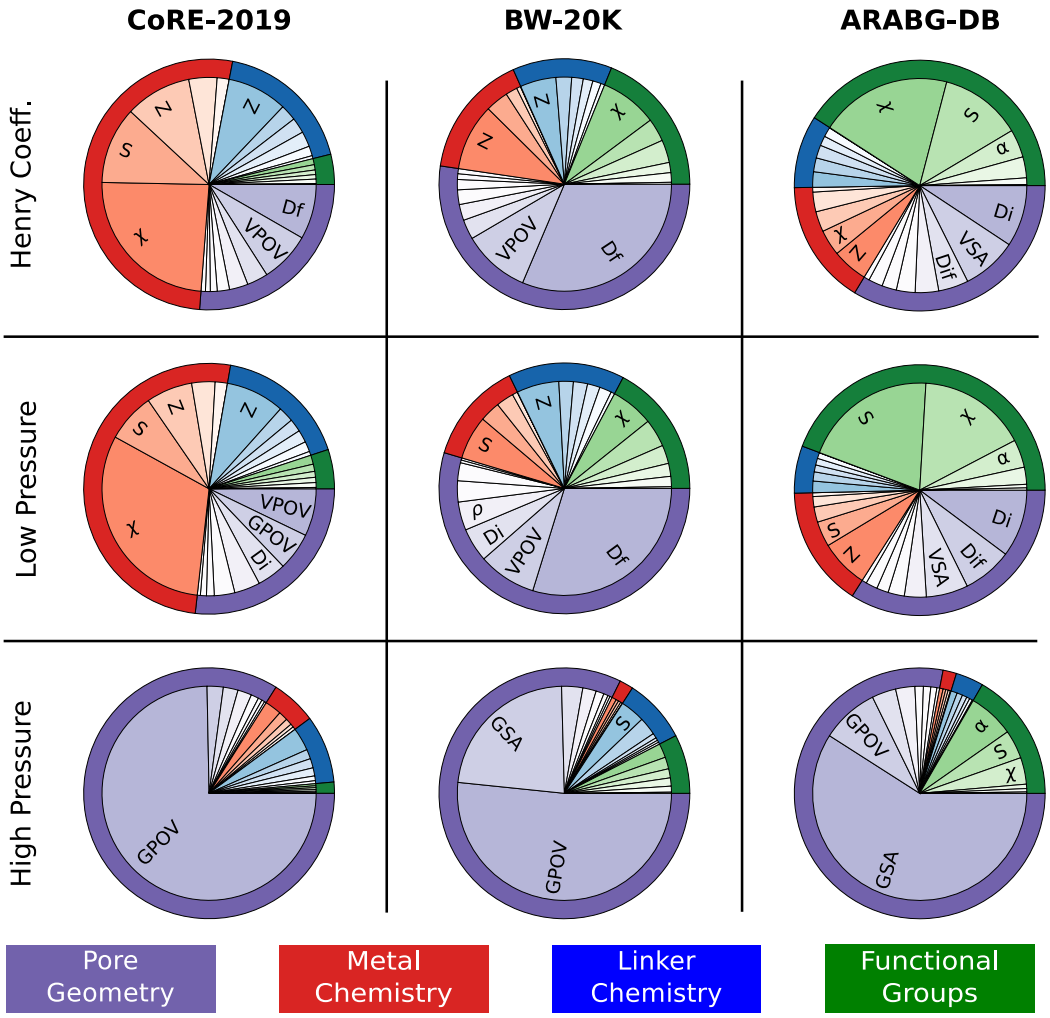
Extended Data Figure 1.1 – **The PCA maps showing the distribution of the materials in each database.** Each database is overlaid using colored dots over the current chemical space that is shown in gray.



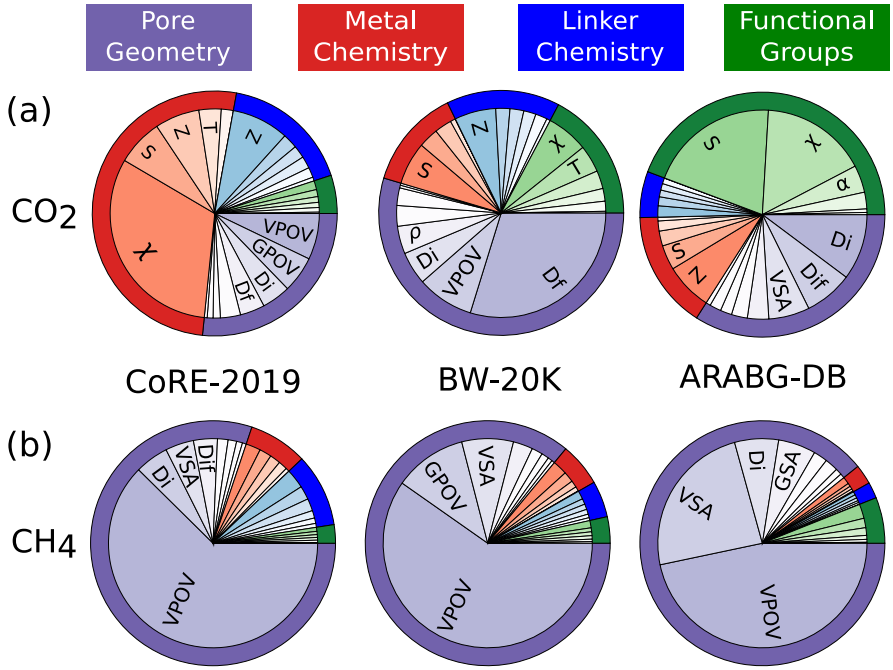
Extended Data Figure 1.2 – **The t-SNE maps showing the distribution of the materials in each database.** Each database is overlaid using colored dots over the current chemical space that is shown in gray.



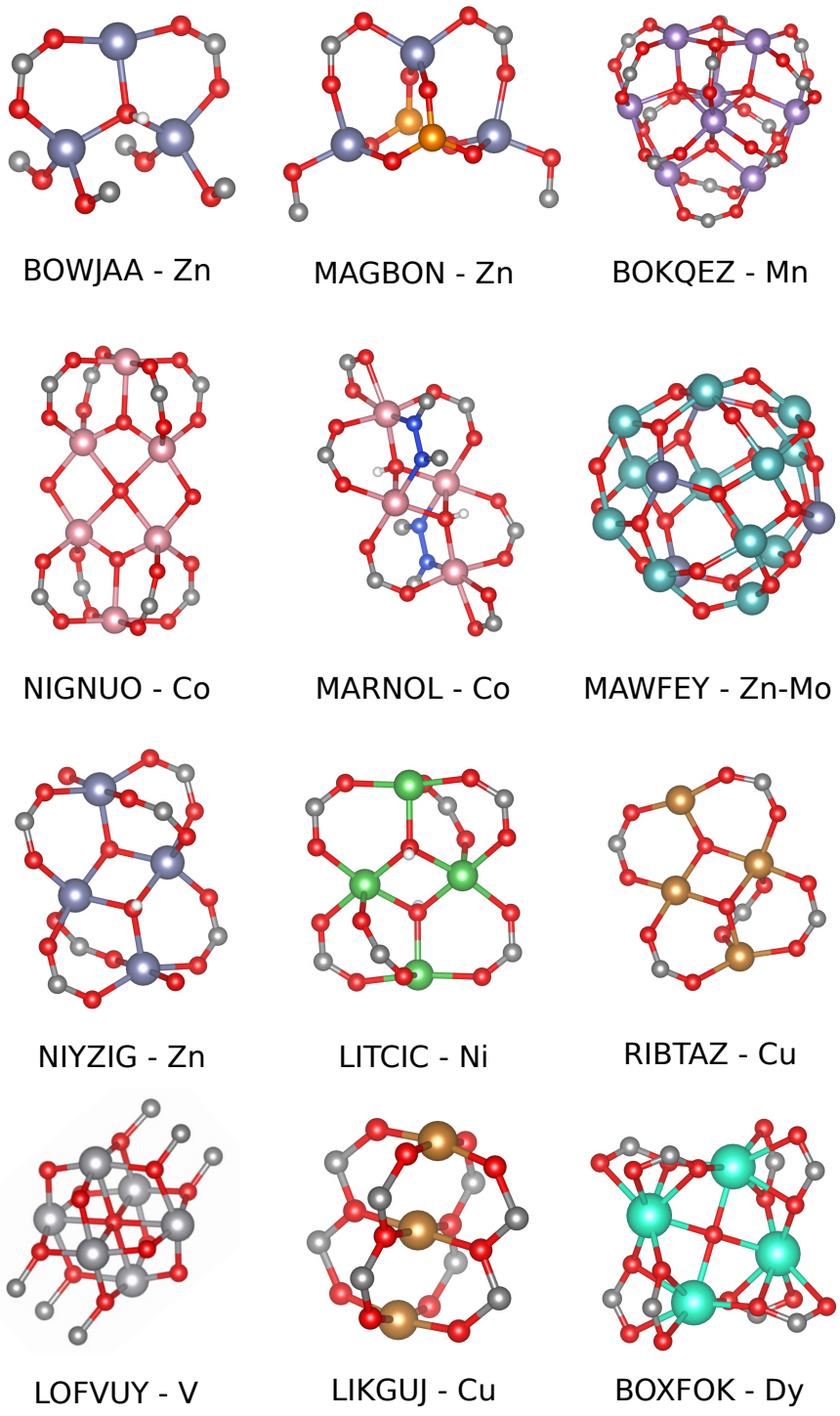
Extended Data Figure 1.3 – **Feature values for CH_4 adsorption properties.** Pie charts showing the SHAP values (importance of variables). SHAP values were computed for the random forest regression models using a training set of **CoRE-2019** and **BW-20K**, and all structures in **ARABG-DB**. For the chemical features, the importance of variables was summed over all RAC depths for each of the heuristic atomic properties. See method section for the meaning of the labels.



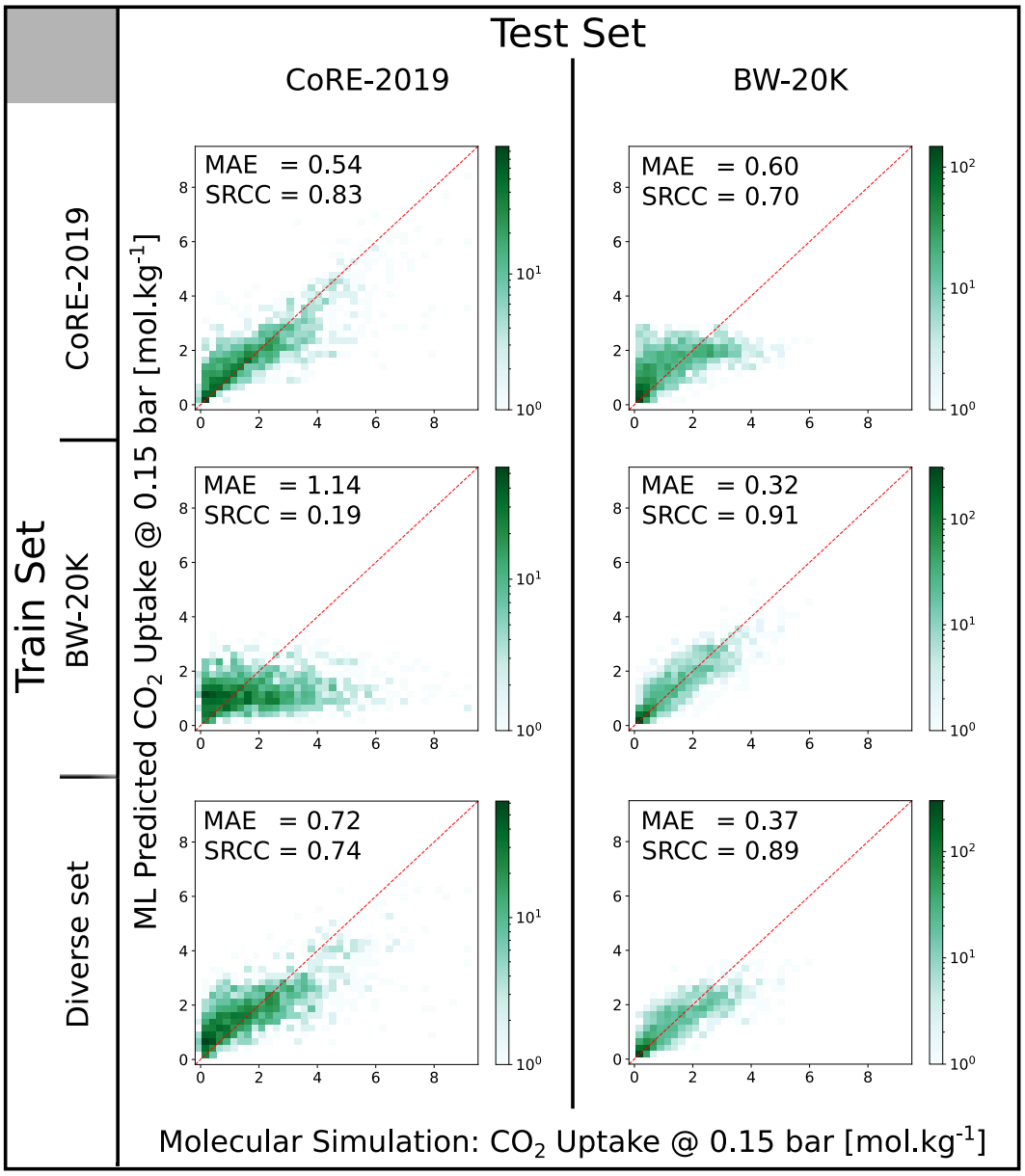
Extended Data Figure 1.4 – **Feature importance for CO_2 adsorption properties.** Pie charts showing the SHAP values (importance of variables). SHAP values were computed for the random forest regression models using a training set of **CoRE-2019** and **BW-20K**, and all structures in **ARABG-DB**. For the chemical features, the importance of variables was summed over all RAC depths for each of the heuristic atomic properties. See method section for the meaning of the labels.



Extended Data Figure 1.5 – **Database dependence of the identified important material characteristics for adsorption properties.** Pie charts showing the SHapley Additive exPlanations (SHAP) values (importance of variables) for (a) the low pressure CO_2 adsorption and (b) CH_4 deliverable capacity. SHAP values were computed for the random forest regression models using a training set of **CoRE-2019** and **BW-20K**, and all structures in **ARABG-DB**. For the chemical features, the importance of variables was summed over all RAC depths for each of the heuristic atomic properties. See method section for the meaning of the labels. Similar values for importance of variables were obtained using other techniques (see SI).



Extended Data Figure 1.6 – **Inorganic SBUs mined from CoRE-2019.** Examples of inorganic SBUs that are missing in hypothetical MOF databases. The CSD names and metal types are shown below each SBU.



Extended Data Figure 1.7 – **Diversity of training data and its impact on transferability of machine learning model.** The parity plots of random forest models using full features; rows and columns correspond to the training and test sets, respectively. The dashed red lines represent the parity. The size of training sets are equal in all cases. The same structures were used as test sets in each column. The diverse set was selected using the MaxMin [107] algorithm using all geometric and chemical descriptors.

1.10 Supplementary materials

Structure refining steps and for featurization and gas adsorption calculations

To prepare the data for featurization and gas adsorption calculations, we carried out a series of steps for cleaning the databases we studied (Figure 1.5). As a first step, we check if the occupancies of the cif file is correct while parsing the structures. We exclude all the cif files that are large, or they do not contain any metals. Then, We compute the periodic pairwise distance matrix between all atoms of the framework and identified cases with atomic overlap when the pairwise distance between two atoms is less than the covalent radii of each atom. After assigning the adjacency matrix, we check each of the connected components of this matrix, and the structures with a connected component that does not contain a metal are identified with having floating atoms (e.g., a solvent molecule) and excluded. If a structure passes all these steps, we perform geometric and RACs featurization for it. The next step is to filter materials for gas adsorption calculation. All the structures that are non-porous to a probe radius of 1.86Åwere excluded for the gas adsorption calculations. We perform partial atomic charges assignment in this step. The structures that take framework maximum positive charge bigger than 3 or minimum positive charge smaller than −3 are recognized to be unrealistic and were excluded.

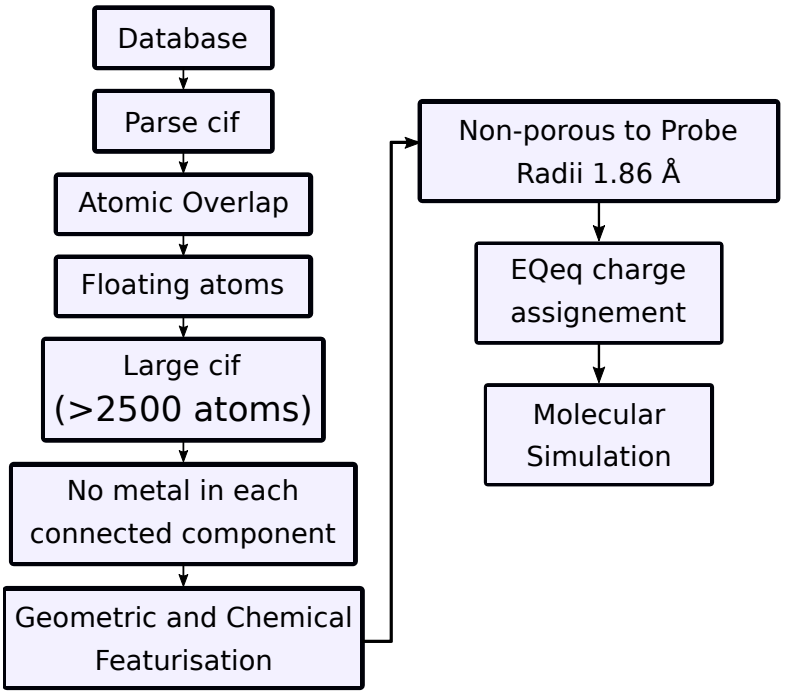


Figure 1.5 – A flowchart representation of the database refinement carried out in this study.

Partial atomic charges

The partial atomic charges for the **CoRE-2019** were derived using the extended charge equilibration method. We use random forest regression models to predict the maximum positive charge (MPC) and minimum negative charge (MNC) of the frameworks using only geometric, and geometric and chemical descriptors. We observe the chemical descriptors are able to learn and predict these attributes of the MOF structures in the **CoRE-2019** with high accuracies (Figure 1.6).

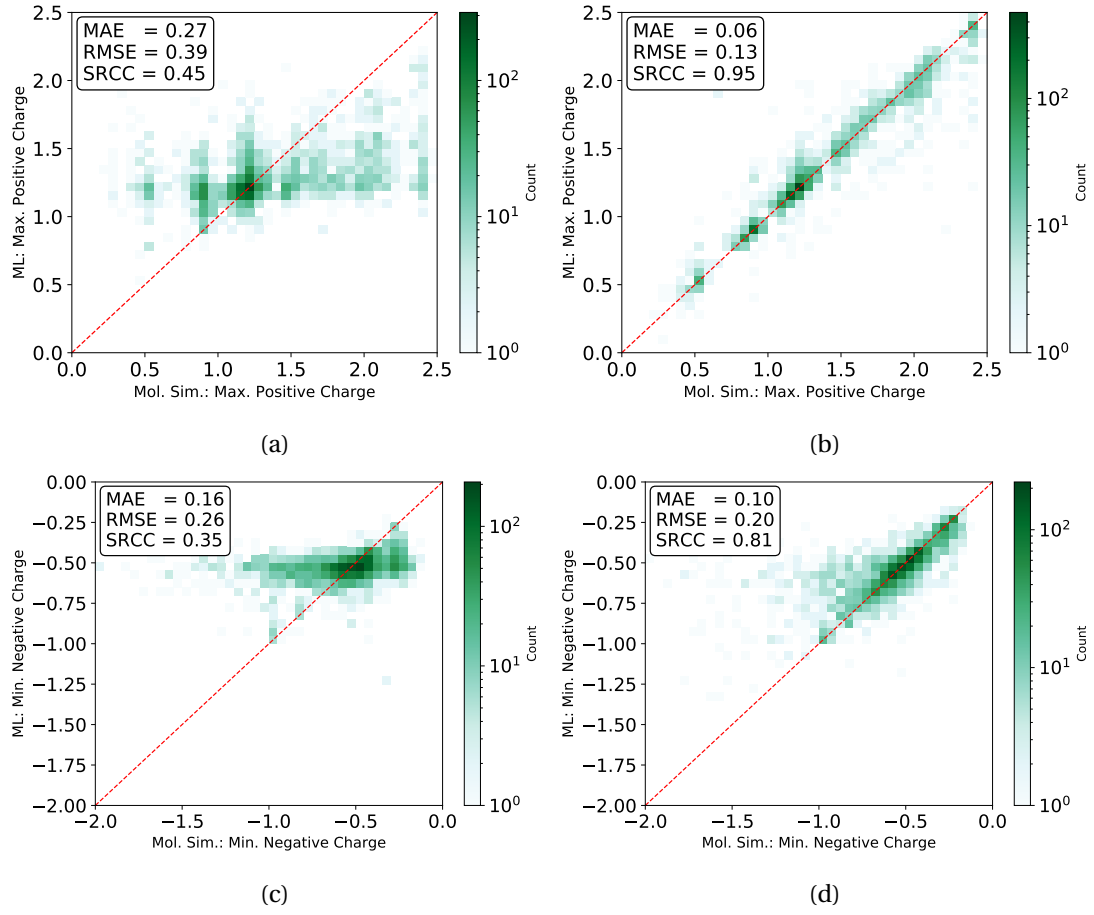
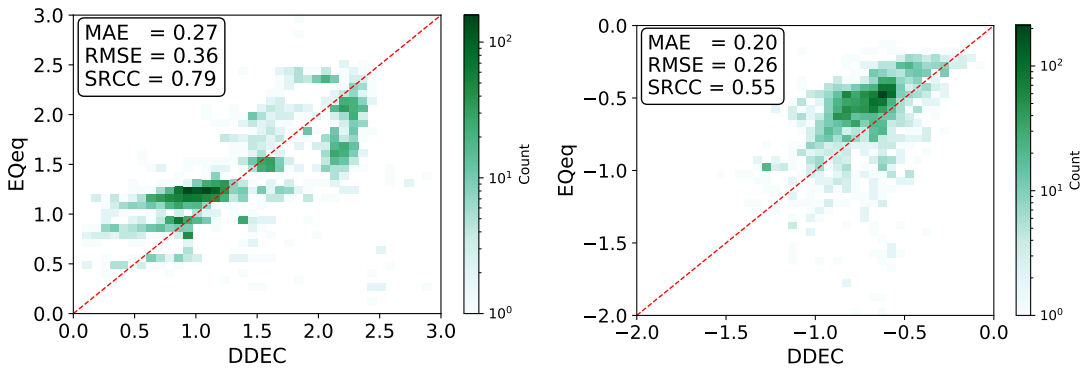


Figure 1.6 – Two–dimensional histogram parity plots and statistics of the accuracy in machine learning predictions of, (a) and (b) the framework maximum positive charge (MPC), and (c) and (d) minimum negative charge (MNC) using only geometric descriptors in (a) and (c), and geometric and chemical descriptors in (b) and (d), for test set from **CoRE-2019**. Partial atomic charges were derived using EQeq method for this database. Random forest regressions were trained using ~7,000 structures and the remaining structures (~2,500 structures) were used as test set. Statistics were reported as average over 10 separate random seeds for train-test splitting. Color-bar shows number of structures in each cell of the histogram.

For a subset of the structures in the CoRE MOF database (2900 structures - **CoRE-DDEC**),

Nazarian et al. performed DFT calculations and derived DDEC charges. Comparing DDEC charges with EQeq is instructive. The correlation between these charges are poor which shows the intrinsic problem with unique featurization of MOF structures using method dependent features (Figure 1.7). We see in figure 1.8 that our chemical descriptors are able to learn the charges derived with both EQeq and DDEC approaches. We note that our prediction accuracies are higher for DDEC charges. This might be due to more smooth behaving of the DFT derived charges which ease the learning process.



(a) (b)

Figure 1.7 – Two-dimensional histogram parity plots and statistics of correlations between two methods for deriving partial atomic charges, namely extended charge equilibration (EQeq) and density derived electrostatics and chemical (DDEC) methods. (a) maximum positive charge of the framework and (b) minimum negative charge of the framework.

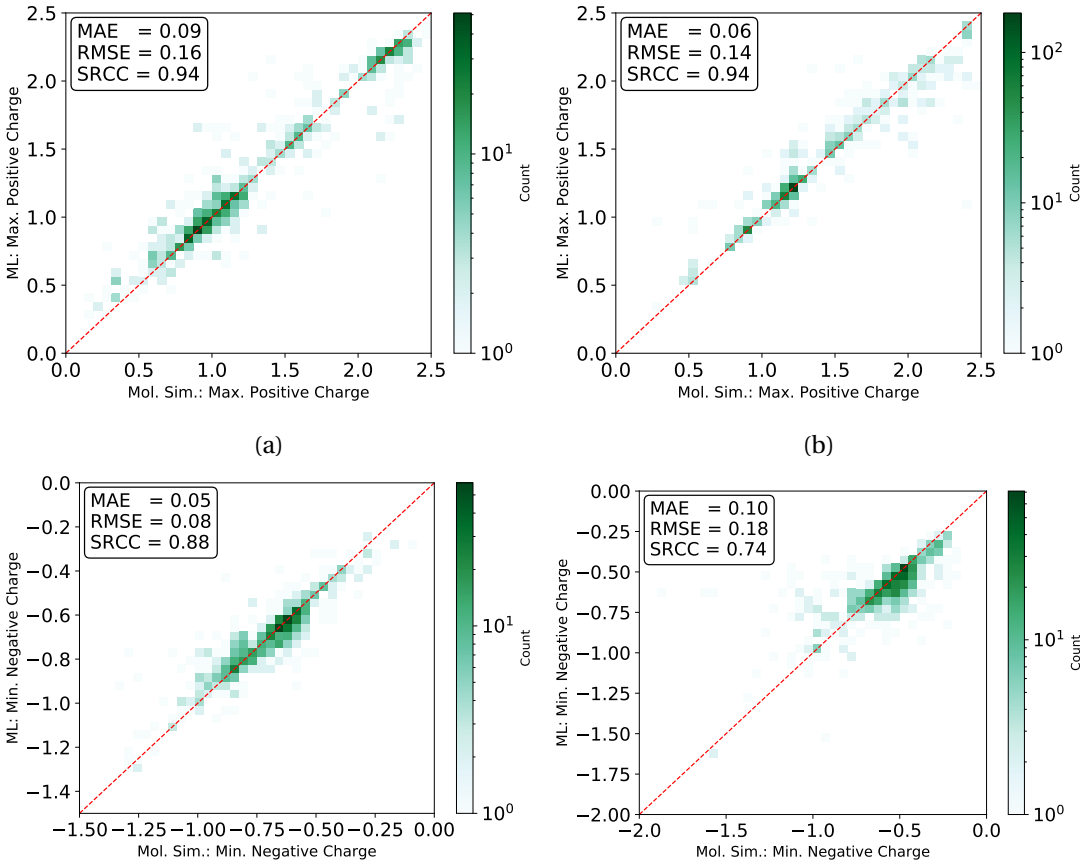
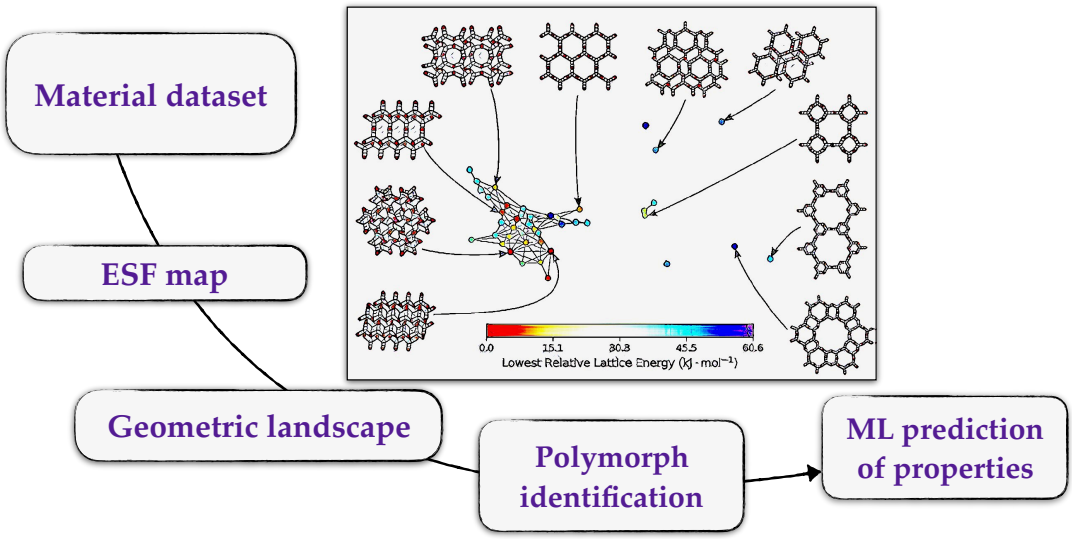


Figure 1.8 – Two–dimensional histogram parity plots and statistics of the accuracy in machine learning predictions of, (a) and (b) the framework maximum positive charge (MPC), and (c) and (d) minimum negative charge (MNC) derived from DDEC method in (a) and (c), and EQeq method in (b) and (d), for **CoRE-DDEC**. Random forest regressions were trained using ~2,000 structures and the remaining structures (~800 structures) were used as test set. Statistics were reported as average over 10 separate random seeds for train-test splitting. Color-bar shows number of structures in each cell of the histogram.

2 Geometric Landscapes for Material Discovery within Energy–Structure–Function Maps [1]



[1] Preprint version of the article Seyed Mohamad Moosavi, Henglu Xu, Linjiang Chen, Andrew I. Cooper, and Berend Smit, 2020, submitted. S.M.M. developed the code for persistent homology, developed the machine learning methodology, performed the calculations, and wrote the manuscript.

Abstract

Porous molecular crystals are an emerging class of porous materials formed by crystallisation of molecules with weak intermolecular interactions, which distinguishes them from extended nanoporous materials like metal–organic frameworks (MOFs). To aid discovery of porous molecular crystals for desired applications, energy–structure–function (ESF) maps were developed that combine *a priori* prediction of both the crystal structure and its functional properties. However, it is a challenge to represent the high–dimensional structural and functional landscapes of an ESF map and to identify energetically favourable and functionally interesting polymorphs among the 1,000s–10,000s of structures typically on a single ESF map. Here, we introduce geometric landscapes, a representation for ESF maps based on geometric similarity, quantified by persistent homology. We show that this representation allows the exploration of complex ESF maps, automatically pinpointing interesting crystalline phases available to the molecule. Furthermore, we show that geometric landscapes can serve as an accountable descriptor for porous materials to predict their performance for gas adsorption applications. A machine learning model trained using this geometric similarity could reach a remarkable accuracy in predicting the materials' performance for methane storage applications.

2.1 Introduction

Design and discovery of porous materials with tailor-made pore sizes, pore shapes, and chemical functionalities is central to a variety of industrial and technological applications, such as gas separation and storage, catalysis, and electronics. [125, 126] Porous molecular crystals are a class of porous materials formed by crystallisation of molecules with shapes that frustrate close packing and/or that have internal, molecular pores. [62, 127] Their discrete molecular building block structures give them certain advantages over other extended framework-type or polymeric materials, such as ease of synthesis and applicability where solubility and amorphous porous phases are desired. [128, 129] Porous molecular crystal materials with high surface areas have been synthesised (to date, up to 3758 $m^2 \cdot g^{-1}$) [130], some of which show promising performance in applications, including hydrogen isotope separation, [131] Xe/Kr separation, [132] and molecular separation. [133]

With the significant progress made in fast and accurate *in silico* prediction of properties and performance of materials, [134, 135] particularly of porous materials, [26, 66, 136] computational modelling plays a significant role in material design and discovery. Using computational techniques, one could generate hypothetical materials to explore the potential chemical space beyond the experimentally realised materials, and then perform *in silico* high-throughput screening of their performance to find the optimal materials for a given application. [71, 137, 138] Unlike framework-type porous materials, such as metal-organic frameworks (MOFs) and covalent organic frameworks (COFs), which are formed by strong coordination or covalent bonds, porous molecular crystals are formed by the balance of many

weak intermolecular interactions, *e.g.*, $\pi - \pi$ stacking and hydrogen bonding. As a result, small changes in the molecular structure can drastically change the landscape of possible crystalline packing, leading to different degrees of propensity for polymorphism and materials properties thereby. [139] Hypothetical material generation techniques that are widely used for framework materials are not generally applicable to porous molecular crystals. To account for this challenge in design and discovery of porous molecular crystals, Pulido *et al.* [140] proposed the concept of energy–structure–function (ESF) maps, combining crystal structure prediction (CSP) with material property prediction, which represents the possible material properties associated with the molecule. For a known molecule, [141] ESF maps revealed new stable polymorphs that were predicted to be promising for different applications before they were targeted for synthesis and measurement in the lab. In this technique, the relative lattice energies of the *in silico* generated structures are projected on a representation of the structural landscape to make a crystal energy landscape, [142] which is used to guide the search of stable packing of the molecule. For molecules showing a simple structural landscape (*e.g.*, with a pronounced minimum well separated from the bulk of the landscape), a 1–dimensional representation of the landscape, often based on the crystal density, [142, 143] is sufficient to reveal the stable packing arrangements for the molecule. [144] However, porous molecules having an internal cavity or a shape that prevents close packing often give rise to a rich, high-dimensional structural landscapes, with multiple local minima. Some of these minima can be easily hidden in a simple 1–dimensional representation. Hence, it is desirable to project an ESF map onto a more complete representation of the CSP landscape, which closely respects the high-dimensional nature of the ESF map, thus improving its predictive ability in pinpointing crystalline packings for desired materials functions.

Ideally, one would construct a crystal energy landscape by representing the free energy surface of the crystals as a function of thermodynamic variables. [142, 145, 146] However, this becomes challenging and infeasible for large molecules or complex energetics of the systems in presence of solvent molecules. [143, 144] Therefore, descriptors able to distinguish different crystalline phases are desired for constructing a good representation of the structural landscape. A robust structural descriptor for crystals should be invariant with respect to the choice of crystal lattice vectors, the permutation of atoms in the crystal structure, and rigid motions of the structure such as translation and rotation. [41] For the purpose of studying porous molecular crystals, a good descriptor should also be invariant to subtle perturbations to the local arrangements of the molecules at their lattice positions. Assuming similar packing leads to similar pore geometries, one can use geometric descriptors to distinguish different molecular packings. Examples of conventional geometric descriptors include crystal density, pore volume, surface area, and pore diameter, all of which satisfy the requirements mentioned above and are cheap to compute; they have been used for representation of structural landscapes. [140] However, each of these conventional descriptors describes partial geometric features only and fails to encode the full picture of the pore shapes of porous materials. [74, 147] Alternatively, one can use persistent homology from mathematics to compute the topological features of shapes. [148] Persistent homology is an algebraic tool which describes these topological

features with a set of persistent barcodes. [149] Persistent homology barcodes provide a quantitative description of the pore shapes, and notably, satisfy the requirements for a geometric representation. While persistent homology was traditionally developed for topological data analysis (TDA), [150–152] it has now been extended to a variety of other disciplines, including material sciences. [153–156]

In this work, we developed a geometric representation based on persistent homology, which allowed us to compute a robust representation of the structural landscapes based on geometric similarity. We show that this representation can be used to automatically explore large databases of porous molecular crystals. This representation has advantages over representations based on a single geometric descriptor in identifying stable crystalline phases, because of its power in encoding the high–dimensional information of an ESF map so as to distinguish structures with unique geometric features not captured by any single geometric descriptor. Moreover, we show that the method offers an explicit structure–function relationship between pore geometries and gas adsorption properties of porous molecular crystals, making ESF maps machine learnable with high accuracies.

2.2 Geometric landscapes

We start with conceiving a representation for the structural landscapes based on geometric similarity. In such a representation, the structures with similar pore geometry should be mapped close to each other. To formulate this representation, we need a metric to assign similarity between pore shapes. Quantifying this geometric similarity is not trivial as, for example, structures with the same crystal density or largest included sphere could be envisioned with totally different pore shapes. [157] Persistent homology, however, allows us to quantify this geometric similarity. Persistent homology can capture the overall similarity of the pore shapes; in contrast to the conventional descriptors, which are more limited. We call such representation of the structural landscape a geometric landscape. The relative lattice energies of the crystals will be projected on this representation to form a crystal energy landscape based on the geometric similarity.

To construct the geometric landscapes, we start with identifying the pore structure of the materials. Here, we use a point cloud sampled on the surface of the accessible pores of the material to a probe atom with a van der Waals radius of 1.5 Å. [93] The persistent homology barcodes then were computed over filtering topological objects to the size of 8 Å of the constructed Vietoris-Rips complexes [158] up to the second dimension for the sampled point cloud (See Figure 2.1, method section, and our previous works [74, 75] for more details). Each dimension of the barcode captures part of the topological features of the pore shape. The zeroth dimension, which gives the number of connected components, is discarded as it does not contain useful information for our analysis. The first and second dimensions of the barcodes capture the features related to the surface and volume of the pore, respectively. Each geometric barcode records the birth and death of these topological objects, which correspond

to the size these features have in space.

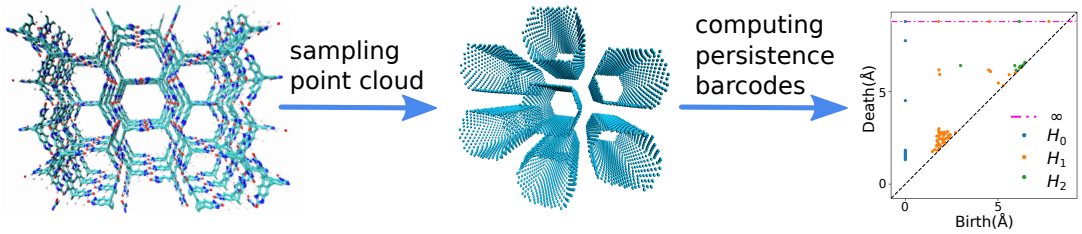


Figure 2.1 – Calculation steps for computing the persistent homology barcodes for a porous molecular crystal. First, a point cloud is sampled on the pore surface of the material. Then, the persistent homology barcodes are computed for this point cloud. The figure on the right is the persistent diagram of the barcodes of the material computed up to the second dimension. This diagram plots the birth and death time of the barcodes.

The persistent homology calculations map each structure in the database to a high-dimensional topological space. In this space, the pairwise distance between each pair of structures is defined by the distance between their persistent homology barcodes. This pairwise distance corresponds to the geometric similarity between the structures in the high-dimensional space where structures with a large distance are geometrically dissimilar while the structures with a small distance are geometrically similar. The L^2 persistence landscape distance [159–161] is used to determine the persistent homology barcode distances because of our previous successful experience in assigning pore geometry similarity using this metric. [74] To make the final representation, instead of including the entire dataset, consisting of 1,000s – 10,000s crystal structures, in the final representation of the geometric landscape, we first classify the dataset to find unique pore-geometry classes. From each class, we use only a landmark structure as a representative structure, to be included on the final geometric landscape. This method allows applying this analysis to extra-large databases (*e.g.*, for datasets that consider multiple conformers) as instead of representing all data points, only representative, low–energy structures are shown on the geometric landscape while still encompassing all the unique classes of pore shapes. Also, it simplifies the representation of the high-dimensional space to avoid over sampling and representing of populated classes with many structures, yet, very similar geometries. This approach is similar to landmark multidimensional scaling, a widely–used dimensionality reduction methodology in computer science and data analysis. [162] To find these representative landmark structures, we perform a Voronoi decomposition of the topological space using the pairwise distances between the barcodes of the materials. To perform this Voronoi decomposition, we select a set of landmark structures covering the topological space with minimum pairwise distance smaller than 10% of the size of the topological space using MaxMin algorithm, [163] which ensures the landmarks were distributed homogeneously in the entire topological space (See method section for details). We assign the remaining structures in the Voronoi cell to their representative landmark structures.

The next step is to apply this technique to generate geometric landscapes for three datasets of

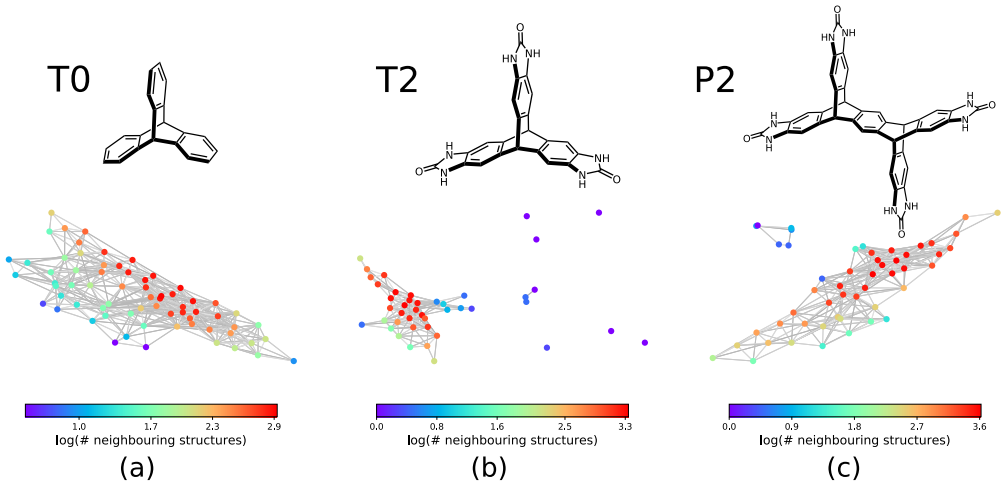


Figure 2.2 – The geometric landscapes of the three triptycene based molecules studied in this work, (a) **T0**, (b) **T2**, and (c) **P2**, with the chemical formula of $C_{20}H_{14}$, $C_{23}H_{14}N_6O_3$, and $C_{38}H_{22}O_4N_8$, respectively. The colour coding shows the number of similar structures to the landmark structure of each node of the geometric landscape. The structures that are contained in a high-dimensional sphere in the topological space centred on the landmark structure with the radius of 15% of the size of the space are counted as similar structures.

crystal structure prediction (CSP) for **T0**, **T2**, and **P2** molecules (Figure 2.2). These molecules possess different directional intermolecular interactions and rigid shapes that promote porosity, and it was shown that they construct multi–minima and complex structural landscapes. [140] Our analysis identified 67, 43, and 51 landmark structures for 2072, 3893, and 7860 porous structures in **T0**, **T2**, and **P2** datasets, respectively. To visualise these geometric landscapes, we use multidimensional scaling (MDS) projection [164] of the relative positions of these unique pore geometry classes using the pairwise distance between the landmark structures in the topological space. MDS representations visualise similarity between individuals in a dataset so that points with relatively small pairwise distances in the high dimensional space are mapped close to each other. The MDS representation of the geometric landscapes of the three databases are shown in Figure 2.2. In these geometric landscapes, each node, *i.e.*, a Voronoi cell of the topological space, represents a set of geometrically similar materials. Nodes with similar barcodes are mapped close to each other and connected when their pairwise distance in the topological space is below 20% of the size of space. The colour coding indicates the number of structures that are similar to each of the landmark points with a cut–off distance of 15% of the size of the topological space. We observe different landscapes for the molecules in Figure 2.2. On the geometric landscape of **T0**, all the landmark structures are closely located to one another, forming one big cluster, which is in line with its featureless, monotonic energy–density distribution reported previously. [140] Similarly, the geometric landscape of **P2** shows one cluster of most of the landmark structures, with a smaller cluster located nearby. By contrast, the **T2** molecule yields a much more interesting geometric landscape, in which the landmark structures are scattered to a larger extent in the spacing, indicating that

these structures have more distinct pore geometries. A proportion of these scattered points corresponds to “spikes” observed in the energy–density landscape for **T2**, [140] though we point out that clusters of similar structures do not have to form such visible “spikes” to be well separated in these geometric projections.

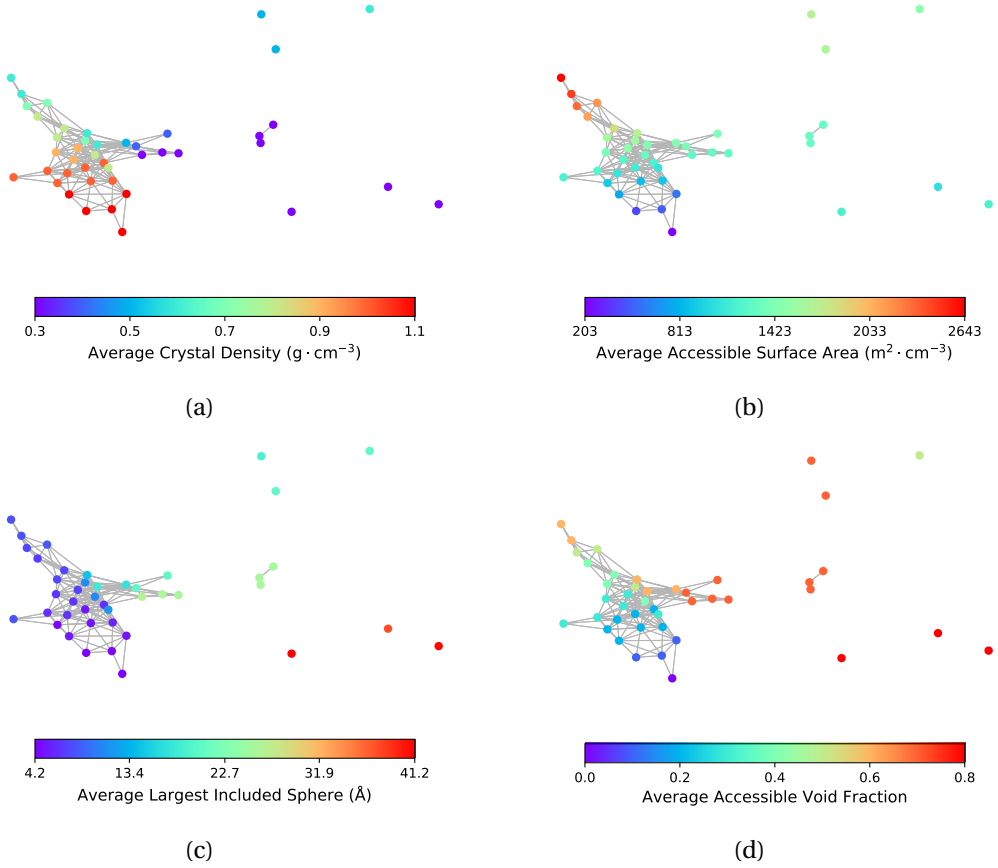


Figure 2.3 – The geometric landscape of the **T2** molecule, colour coded with conventional geometric descriptors, namely (a) crystal density, (b) accessible surface area, (c) largest included sphere, and (d) void fraction.

As a first step, we show that geometric landscapes can capture the expected geometric similarity based on the conventional geometric descriptors. Figure 2.3 shows that the nodes close to each other have similar values of the conventional descriptors, including crystal density, accessible surface area, largest included sphere and accessible void fraction. In other words, the materials that are measured to be similar in the topological space, indeed have similar conventional descriptors. Furthermore, we can see in Figure 2.3 that, for example, there are several landmarks with similar crystal densities (Figure 2.3a) but different cavity sizes (Figure 2.3c). This shows that the geometric landscapes capture information beyond the conventional descriptors used separately, as these landmarks are distinguished and classified in different geometric classes. Capturing multiple geometrical features by one representation

allows for better classification of structures with respect to their pore geometry to represent the full picture of diversity in the pore shapes and geometry of the pores of molecular crystals. If we drew lines from the lowest to the highest value for each conventional descriptor, we would have obtained the 1–D representation of the landscape with respect to the conventional descriptor. In such a 1–D representation, many classes of unique pore geometry will overlay and hence it is difficult to identify. In the geometric landscapes, however, these 1–D representations are embedded into a high-dimensional topological space where all of these unique geometric classes are distinguished from each other.

2.3 Energy-geometry landscapes

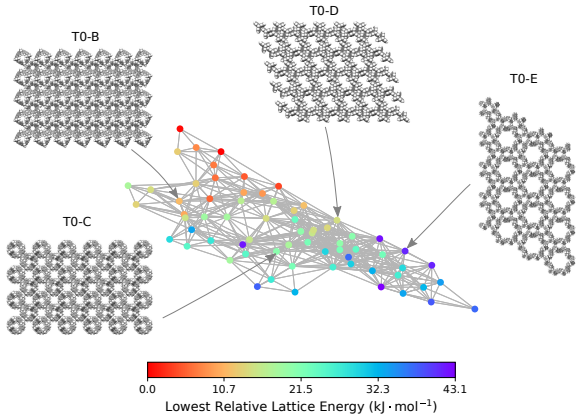
Each node of the geometric landscape represents a unique class of pore geometry, and therefore this representation could be used for identifying unique packing classes of the porous molecular crystals. To find these unique packings, we select the structure with the lowest lattice energy for each node in the geometric landscapes as the stable structure of the corresponding geometry class. Using the geometric landscapes, we could identify many unique classes of packing of the three molecules where some of these structures with ordered packing are shown in Figure 2.4. These landmark structures exhibit a wide range of pore sizes and shapes, immediately revealing potential targets for experimental efforts.

The stability of these polymorphs could be assessed based on their relative lattice energy compared to the global minimum of the landscape. The energetic differences between the polymorphs originate in different ratios of hydrogen bonding network, $\pi-\pi$ stacking, and van der Waals interactions for each packing. We use the **T2** molecule to evaluate the potential of geometric landscapes for exploring crystal structure prediction (CSP) databases to find stable polymorphs because of prior experimental realisation of the molecule. [140, 141] In Figure 2.4, we can see that **T2**-A, **T2**-δ, **T2**-γ, **T2**-α, and **T2**-β have relatively low lattice energy and, hence, one predicts them to be experimentally accessible. Indeed, four of these materials are among the known experimental polymorphs of the **T2** molecule. Therefore, the geometric landscapes could be used to search for stable structures in large CSP databases in one shot.

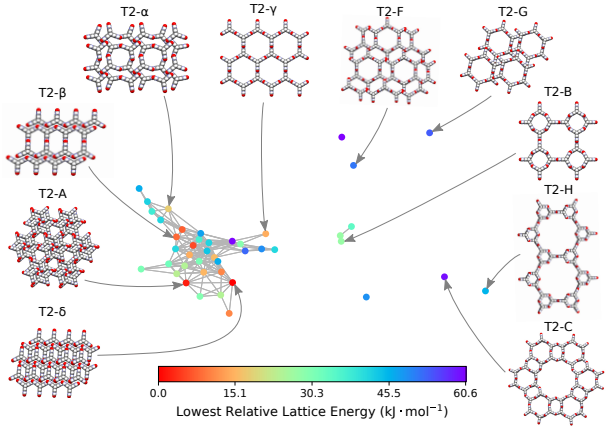
The other materials with higher relative lattice energies in Figure 2.4, yet with unique packings and pore geometries, are potentially interesting because the lattice energy of the porous molecular crystals can be stabilised with proper choice of solvents. Also, previous studies have shown that the lattice energies could vary drastically with dynamics [140] and/or presence of solvents, [165, 166] and therefore one could envision experimental realisation of those materials by solvent stabilisation. However, finding all the experimentally known structures of **T2** molecules in the mainly populated cluster can be a sign of difficulty in synthesising the structures in the smaller or isolated clusters (See 2.2). For those smaller clusters, as the number of neighbouring structures is very low (See 2.2b), the potential well of the landscape is very narrow, and it is unlikely for structures to be trapped in those area of the landscape. This can be explained by the complex architecture of those structures in the small or outlier

clusters, *e.g.*, **T2**-C and **T2**-H in Figure 2.4b, which are more complex assemblies where the **T2** molecules assemble to create a hierarchy of pore sizes.

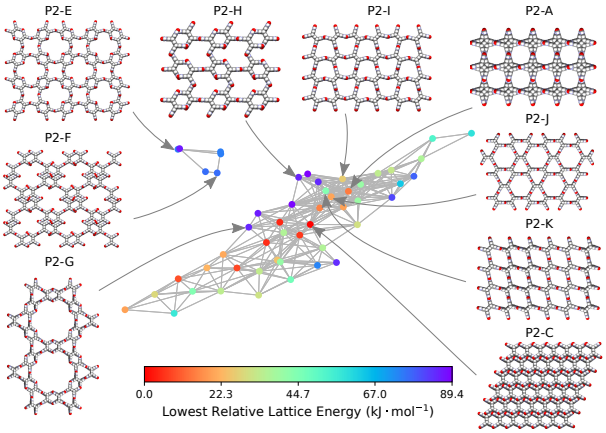
Notably, we see a smaller number of unique ordered packings spotted for the **T0** molecule in comparison to **T2** and **P2** molecules, which implies a comparably simpler landscape of the **T0** molecule. This simplicity can be denoted to the lack of hydrogen bonding motifs in **T0** molecule. Notably, the only experimentally observed structure for **T0** is a densely packed and non-porous structure, where van der Waals interactions are maximised.



(a)



(b)



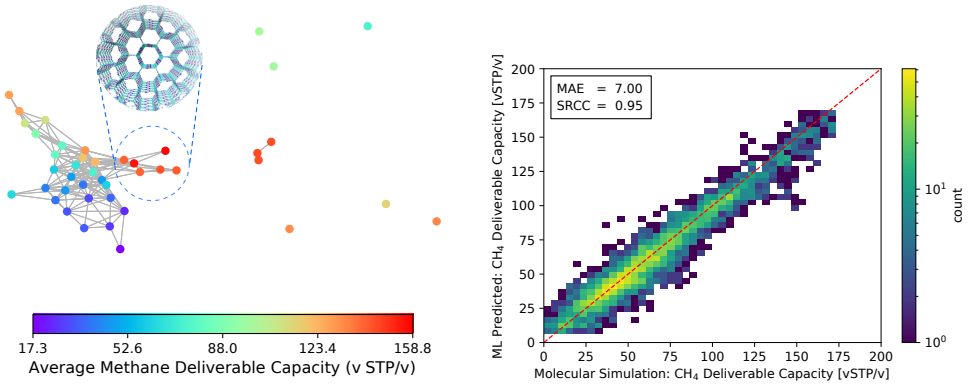
(c)

Figure 2.4 – The energy–geometry landscapes of (a) **T0**, (b) **T2**, and (c) **P2** molecules. The structures with Greek letters are already synthesised in previous works. [140, 141] The letters used for the other structures are chosen in the basis of their relative lattice energy and names used in the previous works. [140] Space–filling representation is used for visualisation of the structures. Carbon, Hydrogen, Oxygen, and Nitrogen atoms are coloured grey, white, red, and blue, respectively.

2.4 Function-geometry landscapes

The pore geometry of porous materials can be optimised for a given adsorption application. Here, we show that geometric landscapes can be used for such optimisation. We show this approach for methane storage application, which is an important application of nanoporous materials. The material's performance for this application is assessed by the deliverable capacity, the difference in the gas uptake in a pressure swing adsorption process reported in standard volumetric units (v STP/v). The adsorption and desorption pressure for this process was set to 5.8 bar and 65 bar, respectively, by Advanced Research Project Agency-Energy (ARPA-E). [53, 167]

Figure 2.5a shows the average methane deliverable capacity of materials in each node of the geometric landscape of the **T2** molecule. A good correlation between geometry and performance is observed as materials mapped close to each other have similar deliverable capacity. This analysis shows that the **T2**-γ structure and the corresponding geometrically similar structures have almost optimal pore shape and size for the methane storage application (Figure 2.5a). These materials have one-dimensional channels with a moderate gravimetric surface area but large volumetric pore volume (Figure 2.3).



(a) (b)

Figure 2.5 – (a) Function-geometry landscape for the methane deliverable capacities of the **T2** molecules. The color coding represents the average methane deliverable capacity of materials in each node of the geometric landscape. (b) Two–dimensional histogram parity plot of the machine learning prediction of the methane deliverable capacity for the materials in the test set. The colour coding shows the number of structures in each cell of the histogram. MAE: mean absolute error. SRCC: Spearman Rank Correlation Coefficient.

The narrow variation of the materials' performance within each node of the geometric landscape shows a clear correlation between the materials' performance and the geometry of the pores (see supplementary material Figure S3 for the standard deviation of the materials' performance in each node). This suggests that the geometric landscapes can be used to explore large databases of porous molecular crystals for finding good performing materials. A possible strategy is to combine them with machine learning to filter out the low–performing materials

from a large database. In such a scenario, [168] instead of performing brute force calculations on the entire database, one carries out calculations only on a subset of structures to obtain enough data, which are used to train the machine learning model. This machine learning model is then used to identify potentially good performing materials where the expensive calculations are worth performing on them. Since persistent homology analysis gives us a metric of similarity, the natural choice for the machine learning model is a kernel based model. [169, 170] In such a machine learning model, the predictions rely on the similarity or dissimilarity (distance) of a data point to all the training data in the feature space, in our case the topological space. [171] Therefore, the prediction accuracy is higher compared to a method relying only on the nearest neighbor, *e.g.*, the landmarks in Figure 2.5a. Here, we use Kernel Ridge Regression (KRR) with combined conventional descriptors and persistent landscape distances (see method section for details). The machine learning predicted deliverable capacities for 3,293 materials in test set are shown and compared to the molecular simulation values in Figure 2.5b. The model accuracy for the out of train samples is remarkable with Mean Absolute Error (MAE) of 7.0 (v STP/v) and Spearman Rank Correlation Coefficient (SRCC) of 0.95. This high accuracy of the machine learning model in predicting material properties and their ranking is promising in comparison to the previous studies where much larger training sets were used. [168, 172, 173] The high SRCC suggest that one can safely use the machine learning model to rank materials and do more expensive calculations on the top performing structures. This will drop the computational costs enormously as only 600 datapoints were used for training the model. The high accuracy of the machine learning model is denoted to the importance of pore geometry in the materials' function. Basically, the adsorption properties of porous materials are a function of their chemistry and pore geometry, [174, 175] and since the chemistry of the molecule is fixed in each of the CSP databases, the geometric similarity could sort out materials with respect to their function nicely.

2.5 Conclusions

We introduced a new representation of the structural landscapes for crystal structure prediction (CSP) datasets and energy–structure–function (ESF) maps of porous molecular crystals based on geometric similarity. We showed this technique has advantage over the typical 1–dimensional representation of the landscapes since it captures both local and global geometric similarity of the pore shapes of the materials. The structures that were identified manually in previous works due to their similar conventional descriptors are classified in different geometric classes in the new representation, allowing automatic identification of unique packing of molecules. Moreover, since the chemistry of the building molecules is fixed in a CSP database, this technique could reveal structure–function relationship for gas the adsorption applications of porous molecular crystals.

We envision the geometric landscapes to be used to automatically explore CSP databases for finding materials with two features, namely unique packing and high performance. This technique allows exploring large CSP databases to find unique packings which could be

subsequently tried to be synthesized experimentally. Besides, instead of performing brute force calculations of a large database of porous materials for a given adsorption application, one can prescreen the database to spot the good performing geometric classes and then do calculations only on those structures that are in an identified good performing geometric class. In this respect, we showed that machine learning could accelerate this procedure even further as geometric landscapes are physically meaningful and machine-understandable [138, 176] material representation for porous materials.

2.6 Methods

Materials

The crystal structure prediction databases of the molecules and the corresponding adsorption properties of the materials were extracted from previous study. [140, 177]

Persistent barcodes and Voronoi decomposition of the space

We retrieved information of pore accessibility for each structure using Zeo++ [93] for a probe radius of 1.5 Å and then sampled accessible pores with a fixed number of points per unit accessible surface area. We constructed the Vietoris-Rips complex and generated zero-dimensional (0D), one-dimensional (1D) and two-dimensional (2D) persistence barcodes, up to a cut-off length of 8.0 Å using Ripser. [158] To quantise pore shape similarity between two structures in the barcode space, we measured the pairwise distance, by a weighted combination of L^2-landscape distance [159, 160] based on their persistence barcodes (Eq. 2.1). $\Lambda_{d=1}$ and $\Lambda_{d=2}$ are the L^2-landscape distances for the first and second dimension of persistent barcodes, respectively.

$$d = \sqrt{0.1 \times |\Delta \text{ ASA}| + 0.45 \times \Lambda_{d=1}^2 + 0.1 \times \Lambda_{d=2}^2}. \tag{2.1}$$

$|\Delta$ ASA$|$ is the differences between accessible surface areas per volume of the two structures. All the conventional descriptors were computed using Zeo++. [93, 119]

To perform Voronoi decomposition, we selected a set of landmark structures using MaxMin algorithm, [162, 178] which ensured all landmarks were distributed homogeneously in the entire barcode space. Then we assigned the remaining structures to their closest landmark structures. When applying MaxMin algorithm, we chose the first landmark structure at random, then for selecting a new landmark structure, we took the following steps:

1. For each structure, calculate its distances to all present landmarks, find the maximal distance, recording as d_i^{Max}, and the minimum distance, recording as d_i^{Min} (i for the ith structure);

2. The new landmark is the structure with the maximal value of d_i^{Min}. We record the maximal value among all d_i^{Max} and assigne the size of the barcode space as the Max(d^{Max}) observed in all steps;

3. Repeat the above steps until Max(d^{Min}) is less than 10% of Max(d^{Max}) to ensure the maximum distance between a structure to its corresponding landmark structure is less than 10% of the maximum pairwise distance in the barcode space (a representative for the size of the barcode space).

Visualising the pore geometry landscape

Multidimensional scaling (MDS) is a visualisation method based on the pairwise distances, similarity or dissimilarity in a set of objects in a high-dimensional space. [164, 179] Here, we used metric MDS using the pairwise distances between landmark structures computed using equation 2.1. The MDS algorithm aims to preserve the relative distances between data points in the high dimensional space when the points are projected on a 2D plane. The metric for evaluating the consistency between the low dimensional representation and the high dimensional distances is called the stress function Eq. 2.2. This function returns the residual sum of squares of the distances in the HD space to the LD space. The stress function was optimised by the stress majorisation algorithm, which is implemented in scikit-learn, a python machine learning package. [120]

$$S = \left(\sum_{i,j=1\cdots N} d_{i,j} - \overline{d}_{i,j} \right)^{\frac{1}{2}} \tag{2.2}$$

Machine learning

Kernel Ridge Regression (KRR), a regression model with l2-norm regularisation and kernel trick, was adapted from scikit-learn. [120] The kernel distances between structures were determined using a combination of their distance in topological space (TS) and conventional geometric space (CS). The distances in TS were computed using persistent homology and equation 2.1. The euclidean distances between the conventional geometric descriptors were used to compute the pairwise distances between structures in CS, using normalized values of largest included sphere, crystal density, void fraction, and accessible surface area. Two radial basis functions (RBF), Gaussian kernel, were used with two independent Gaussian width for the TS and CS. The pairwise distance between data points computed with:

$$K = \lambda K_{\mathrm{TS}}(d_{\mathrm{TS}}, \sigma_{\mathrm{TS}}) + (1-\lambda) K_{\mathrm{CS}}(d_{\mathrm{CS}}, \sigma_{\mathrm{CS}}), \tag{2.3}$$

where

$$K_{\mathrm{TS}\,or\,\mathrm{CS}}(d, \sigma) = \exp\left(-\sigma d^2\right). \tag{2.4}$$

The model was trained using 600 training data using 10-fold cross validation and grid search to find the optimal Gaussian width for each kernel and the regularisation factor. The accuracy of model was found to be highest for λ equal to 0.5.

2.7 Supplementary materials

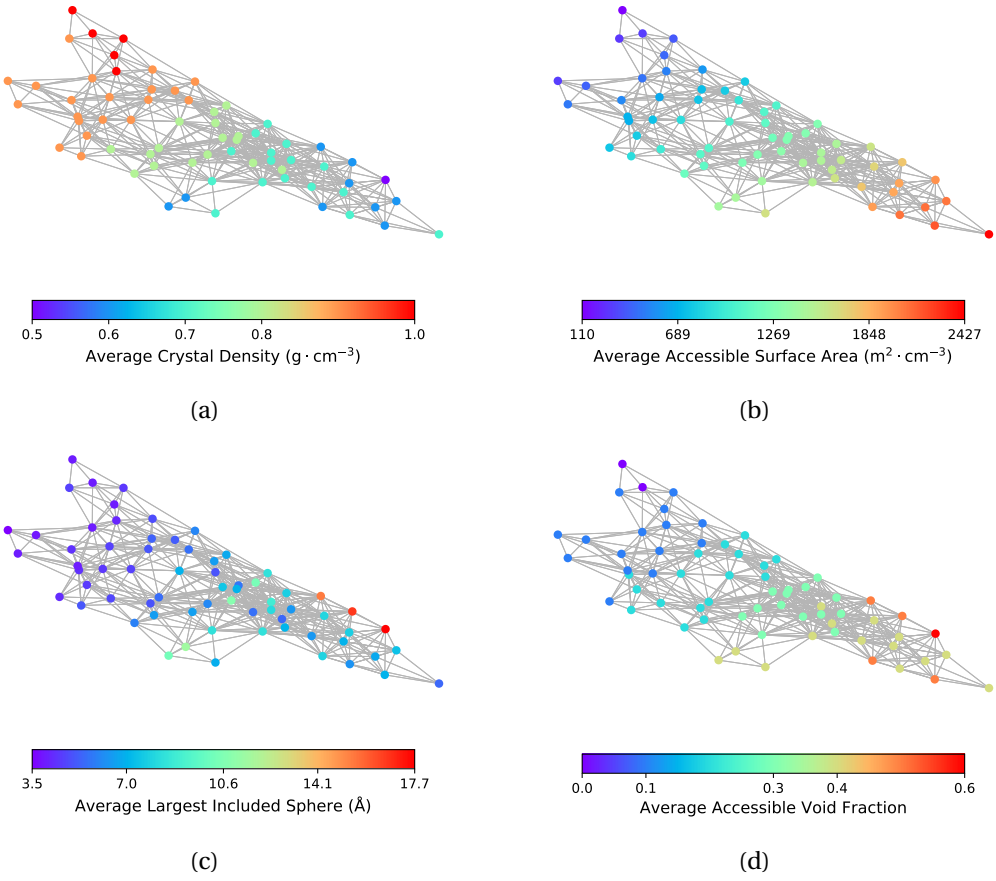


Figure 2.6 – The geometric landscape of **T0** molecule. The color coding shows the average conventional geometric descriptors, (a) crystal density, (b) accessible surface area, (c) largest included sphere, and (d) void fraction, respectively.

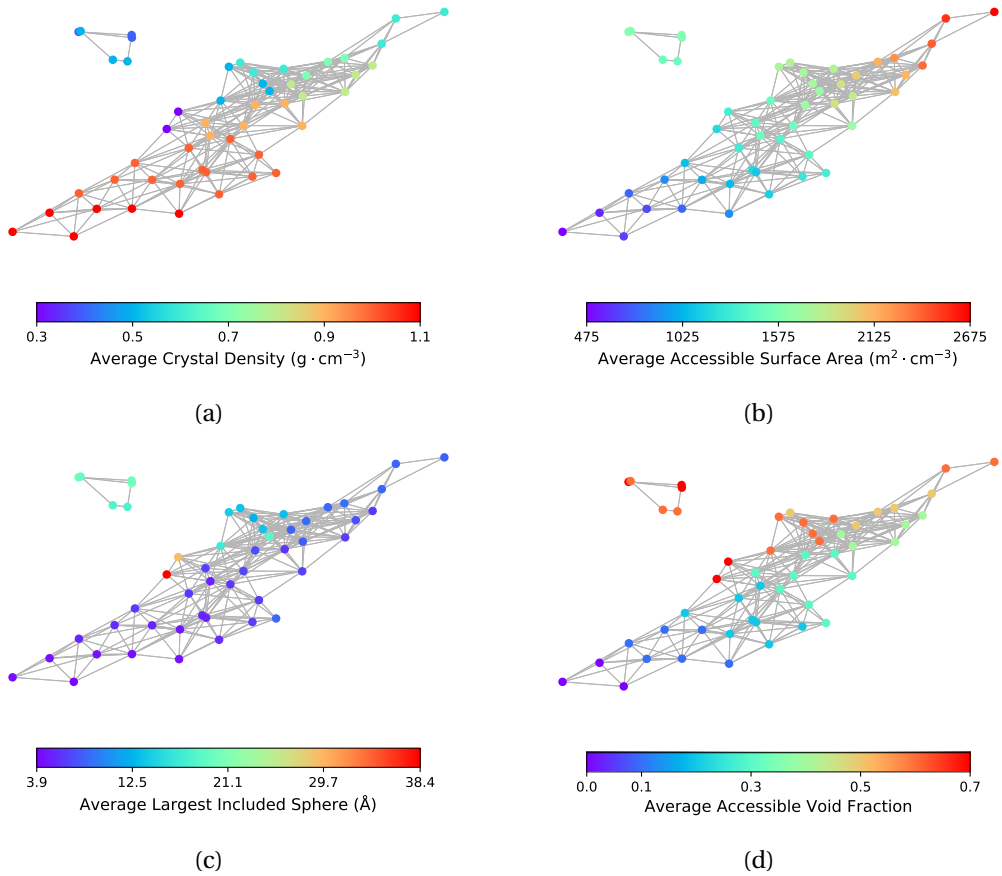


Figure 2.7 – The geometric landscape of **P2** molecule. The color coding shows the average conventional geometric descriptors, (a) crystal density, (b) accessible surface area, (c) largest included sphere, and (d) void fraction, respectively.

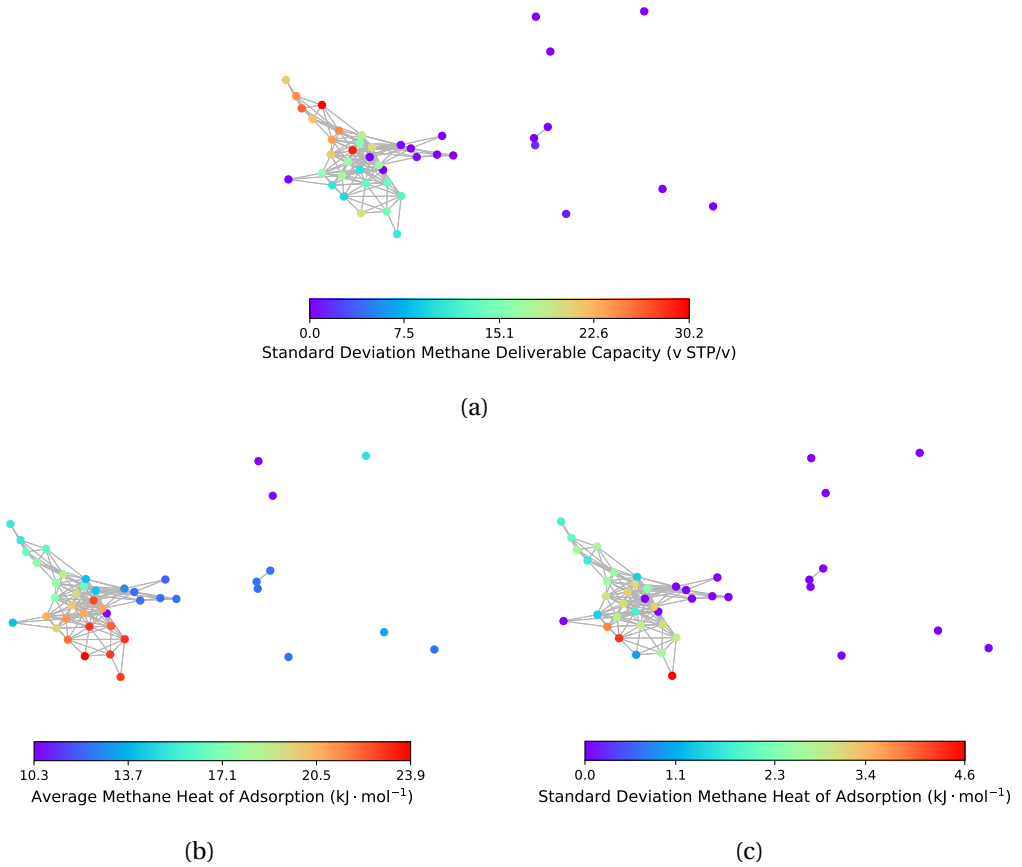


Figure 2.8 – The correlation between geometry and function for methane storage application for **T2** molecule. Low standard deviation in each bin of the geometric landscape shows the extend of importance of pore geometry for this application.

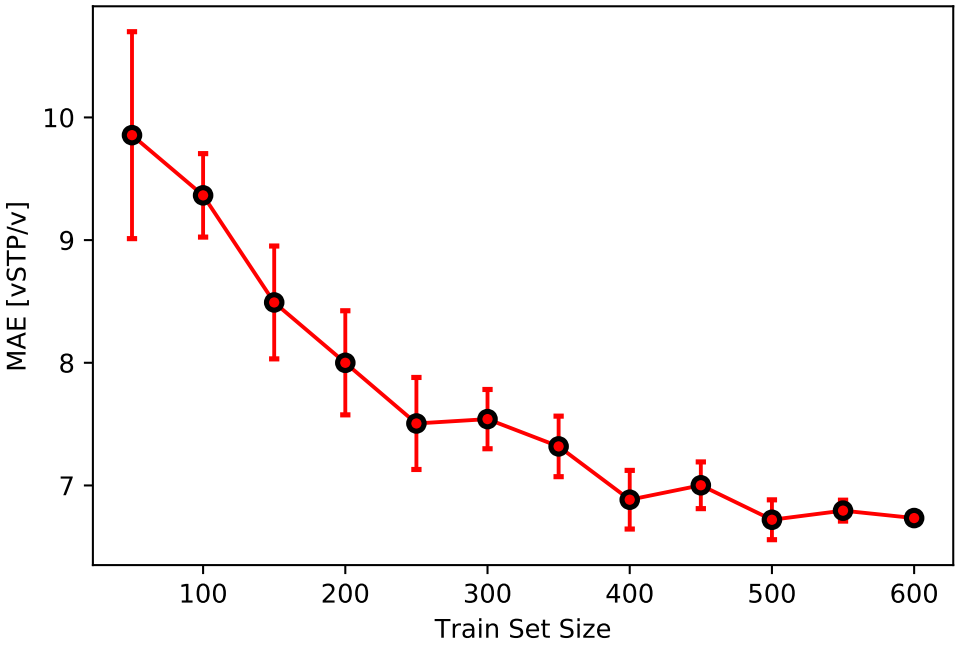
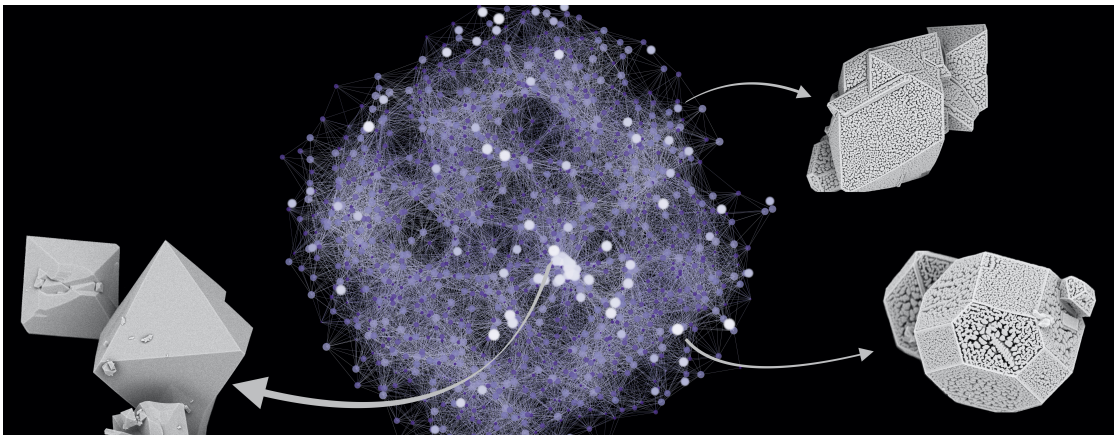


Figure 2.9 – Learning curve of the machine learning model. The mean absolute errors (MAE) were computed 10 times each with a unique random seed for each train set size. Error bars show the standard deviations.

3 Capturing chemical intuition in synthesis of metal-organic frameworks[1]



[1] Postprint version of the article published as: Seyed Mohamad Moosavi, Arunraj Chidambaram, Leopold Talirz, Maciej Haranczyk, Kyriakos C. Stylianou, and Berend Smit, *Nature Communications*, 2019, 10, 539, https://doi.org/10.1038/s41467-019-08483-9. S.M.M. developed the optimisation and machine learning protocols. S.M.M. and L.T. prepared the web application. S.M.M. and B.S. wrote the manuscript with contributions from all authors.

Abstract

We report a methodology using machine learning to capture chemical intuition from a set of (partially) failed attempts to synthesize a metal-organic framework. We define chemical intuition as the collection of unwritten guidelines used by synthetic chemists to find the right synthesis conditions. As (partially) failed experiments usually remain unreported, we have reconstructed a typical track of failed experiments in a successful search for finding the optimal synthesis conditions that yields HKUST-1 with the highest surface area reported to date. We illustrate the importance of quantifying this chemical intuition for the synthesis of novel materials.

3.1 Introduction

Since two decades ago, when metal-organic frameworks (MOFs) emerged as a versatile class of materials for variety of applications, the chemistry and applications of MOFs have been the subject of a large body of research across several disciplines [85, 180]. MOFs were described by the concept of reticular chemistry as materials composed of structural building blocks assembled on a net [59]. The scientific excitement about MOFs originates in the fact that by modifying the building blocks, i.e. changing the metal nodes or organic ligands, MOFs can be tuned for a given application. Therefore, in principle, the number of possible materials is infinitely large; however, since synthesis and optimisation of these materials can be time consuming and laborious [181], only a fraction of them have ever been synthesised.

The synthesis of MOFs involves the self-assembly of the structural building blocks (known as secondary building blocks (SBUs)) in a 3D periodic network. However, our understanding of the self-assembly procedure, i.e. the kinetics and energetics of framework bond formation, nucleation, and crystal growth, has remained too limited to guide the synthesis of these materials. Specifically, since diverse and numerous chemistries exist in MOFs, even the known synthesis conditions for one MOF are typically not transferable to new MOFs, and accordingly, this has prevented chemists to draw a general synthetic route for these materials. The parameters for a typical MOF synthesis include the selection of solvents and their composition, temperature, and reaction time, etc. Considering each parameter as a variable, one needs to probe the high-dimensional chemical space constructed by these variables to find sets of synthesis conditions leading to formation and crystallization of the desired MOF. Without any prior knowledge, one could envision a brute force approach and perform, say, a large grid search of the chemical space using robotic synthesizers. The cost of this approach increases exponentially with the number of variables, e.g. testing only ten choices for a space of nine variables requires a billion experiments. With such poor statistics, one may wonder how so many MOFs could have been synthesized? Clearly, the fact that thousands of MOFs have been synthesized [65] indicates that chemists have been able to beat brute force statistics by orders of magnitude. Given that at present there are at best some empirical guidelines, one can argue that their selection of experimental conditions must have been positively biased by the chemical intuition that

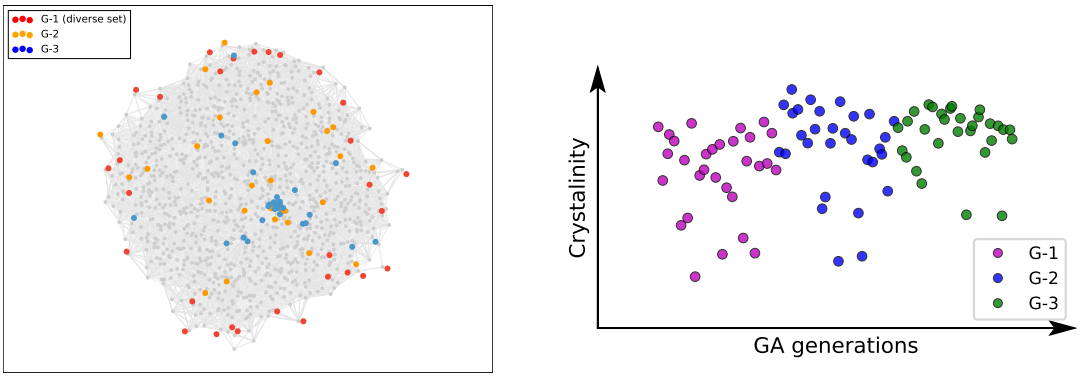
synthetic groups have acquired. While publications typically report only the most successful synthesis conditions, the chemical intuition is built from all experiments, in particular, the substantial number of partially successful and failed experiments. The aim of this work is to develop a systematic approach to capture this chemical intuition.

Recently, machine learning is starting to be applied to chemical synthesis [182–188]. Most of these efforts focus on predicting the outcome of a specific reaction. For instance, Raccuglia et al. proposed and tested successfully the synthesis of a material by machine learning failed experiments using decades of old notebooks of chemical synthesis [186]. Ahneman et al. trained a random forest to predict the performance of the Pd-catalyzed Buchwald-Hartwig reaction [184]. For MOF synthesis, the ligands and metal nodes are in most cases sufficiently simple or even commercially available that their synthesis is often not the bottleneck. Most time and effort are spent in finding the optimal conditions for the ligands and metal nodes to self-assemble into crystals. In this work, we show how machine learning can be used to capture and quantify the chemical intuition that researchers develop in their search for these optimal conditions.

3.2 Synthesis and optimisation of the surface area of HKUST-1

To illustrate our methodology, we focus on a real-life example of MOF synthesis. HKUST-1 is a well-studied MOF that has been synthesized by a large number of different groups [189–192]. Although all groups report high quality powder X-ray diffraction patterns, the different samples show Brunauer–Emmett–Teller (BET) surface area ranging from ~300 to ~2000 m^2g^{-1} (See Figure 3.9 for BET history) [189]. The comparison of the different synthesis conditions shows that they differ mainly in solvent composition (e.g., mixtures of DMF, water, different alcohols, and others), temperature (25°-180°C), and methods (e.g., conventional heating, microwave, electrochemistry, mechanochemistry, ultrasonic, etc.). At present, we lack the knowledge to explain why there are such differences in the BET surface areas, yet from a practical point of view it is important to obtain this material with the highest surface area [193].

One can safely state that this body of work on HKUST-1 involves hundreds if not thousands of experiments, of which only the successful conditions have been published. In this work, we aim to make the case that important and useful information can be obtained, if these groups would also have published their (partially) failed experiments. We use a robotic synthesis procedure to efficiently regenerate part of the failed and partially successful experiments that have been performed in the course to synthesize this material. Using a robotic synthesis platform improves the reproducibility of the generated data. Our robotic synthesizer uses microwave heating and the synthetic procedure involves selecting the setting of 9 different parameters that fully specify the synthesis conditions. Hence, a particular experimental condition can be described as a point in a 9-dimensional (chemical) space (see Table 3.2). We have selected the ranges of synthesis conditions such that they include those solvents and temperatures that have been reported as successful in the literature, but not necessarily



(a) (b)

Figure 3.1 – **Optimisation of synthesis condition of HKUST-1.** (a) Multidimensional scaling projection of the 9-D space of parameters onto a 2-D plane. In this representation, similar conditions are plotted close to each other, and connected if they have normalized pairwise distance below 0.1. Grey dots visualize the extent of the entire bounded (chemical) space, represented by mapping the set of 1000 most diverse synthesis conditions obtained from the MaxMin method. The red dots are the first 30 of this set which are used for the first experiments (G-1), the orange and blue dots mark the second (G-2) and third (G-3) generations obtained from the first via the genetic algorithm (GA). (b) Progress in crystallinity during GA optimization. The color of dots indicates the generation in the GA.

using microwave heating. Our robot can carry out 30 reactions per cycle, where a cycle is completed typically within one day. A simple grid search to explore all possible experimental conditions would require of the order of 10^9 robot cycles, which illustrates the need of this chemical intuition, or in our case, in which we impose a lack of intuition, enhanced sampling techniques.

In the case of HKUST-1, several quite different successful synthesis conditions have been reported. Since the location of these sets of conditions are not known a priori, and for instance, might be clustered in relatively small islands in the high-dimensional space, pinpointing them is genuinely non-trivial. Simple gradient-based algorithms are discarded here due to the high probability of winding up in a local optimum. Genetic algorithms (GAs) have proven to be a robust global optimization algorithm for searching such a complex space [194, 195]. The optimisation strategy in a GA is inspired by natural selection, nature's optimisation strategy. The 9-dimensional synthesis vector takes the role of the chromosome, carrying the synthesis variables as its genes, which are evolved via selection, crossover, and mutation (see section 3.7). Only the mutated genes of successful parents are transferred to the next generation, thus optimizing the synthesis conditions generation by generation.

We start the search for the optimal synthesis conditions without any chemical intuition, i.e. all components of the 9-dimensional synthesis vector are considered equally important. The first run aims to cover the experimental space as widely as possible, using the MaxMin method [196], to obtain the set of 30 most diverse synthesis conditions. Figure 3.1a shows these

conditions in a multidimensional scaling (MDS) projection. MDS-plots visualize the similarity between individuals in a dataset [197]. In this study, the Euclidian distance of normalized variables measures the similarity between synthesis trials. In Figure 3.1a, similar synthesis trials are mapped close to each other while dissimilar experiments are far from each other on the map (see method section for details). As expected, but not intuitively obvious, in such a high dimensional space the most diverse set is located at the edges. The synthesis is attempted for each of the conditions, and the samples are analyzed for crystallinity and phase purity. Using those metrics for the objective function, we evolve the second generation and perform synthesis for all 30 new conditions. We measure crystallinity, phase purity, and BET surface area, and combine those metrics for the objective function for the third generation.

Figure 3.1b shows the progress in crystallinity over the three generations of experiments. The GA generations contain several different synthesis conditions that yielded samples with ideal powder X-ray diffraction pattern and phase purity. For highly crystalline samples in each generation, we determined the BET surface area (see Table 3.1 and Figure3.10 for powder X-ray patterns of the samples), and, not surprisingly, find a wide range of BETs, including the largest reported BET to date. Figure 3.2 illustrates that the optimal conditions for the synthesis of HKUST-1 yielded large crystals, while the samples with a lower BET showed intergrowth and other deviations that are not captured by powder diffraction analysis. Since the BET of $2045\,m^2g^{-1}$ close to the theoretical maximum of $2153\,m^2g^{-1}$ [198], there was no need to further continue our GA using the BET as objective function.

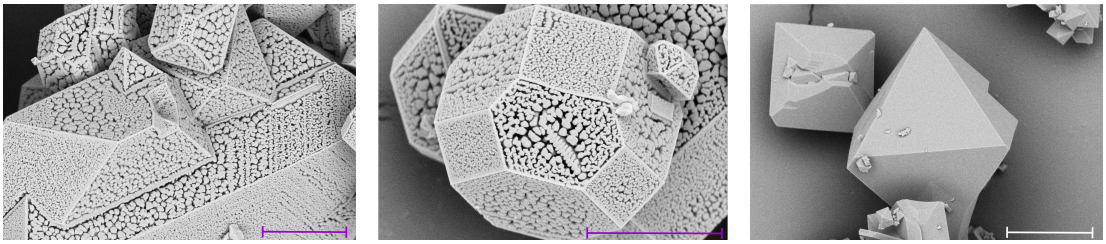


Figure 3.2 – **Scanning electron micrograph of several Cu-HKUST-1 samples.** All these samples have high crystallinity but show a wide range of surface areas (see Table 3.1 for surface areas and Supplementary Figure 11 for more images). Scale bars for sample 1, sample 3 and sample 5 show 5, 4, and 10 μm, respectively.

3.3 Capturing chemical intuition using machine learning

The common practice is to claim victory and publish the synthesis conditions that yielded the highest experimentally measured BET value. Instead, we would like to focus on the observation that to achieve this high BET surface area, we have over 120 failed and partly successful experiments. In the following, we analyze this data to quantify the relative importance of the experimental variables on the outcome of the synthesis. We use the embedded technique in random decision forest, a machine learning regression model. The result is shown in Figure 3.3a and provides the relative impact of the probed experimental parameters on

Sample	BET [m^2g^{-1}]	Synthesis conditions								
		H_2O [ml]	DMF [ml]	EtOH [ml]	MeOH [ml]	iPrOH [ml]	Rectants Ratio	Temperature [°C]	Microwave Power [W]	Reaction Time [min]
1	367	0.5	0.0	5.0	0.0	1.0	0.9	120	174	58
2	526	0.5	1.0	0.0	4.0	0.0	1.8	176	246	44
3	935	0.0	4.5	0.0	0.0	0.0	1.8	123	200	7
4	1596	0.0	4.0	0.0	0.0	2.0	0.8	200	240	60
5	2045	0.5	2.5	2.0	0.0	0.0	1.5	140	200	20

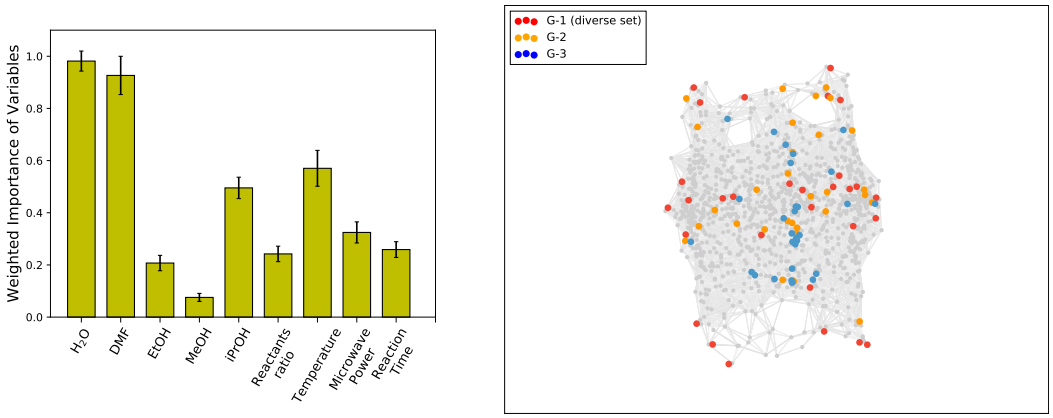
Table 3.1 – BET surfaces and the corresponding synthesis conditions of the five samples with the highest crystallinity.

crystallinity and phase purity. For example, changing the temperature has three times more impact than changes in the reactant ratio. It is this type of information that a synthetic chemist will typically transfer to the next experiments; knowingly, as rules of thumb, or, subconsciously, in the form of "chemical intuition." Machine learning of the recorded data allows us to quantify this intuition, and to use it for subsequent experiments.

Without prior knowledge the difference between synthesis conditions was quantified as the Euclidian distance in 9D space using an equal weight of all parameters. Building on the chemical intuition extracted from our machine-learned model, we now compute the distance in 9D space using the chemical intuition to weight each dimension in the distance measure. If we normalize these weights such that the most important variable has a value of 1, we obtain a chemical space shown in Figure 3.3b. This figure shows how the chemical space for HKUST-1 shrinks in the new metric (the Euclidian distance, weighted by the importance of variables), illustrating that less samples can be placed along less important dimensions without loss of sampling accuracy. Therefore, since the chemical space can be sampled much more efficiently, the chance of success is larger for the same number of trials.

3.4 Application of learned chemical intuition

We now illustrate transferring the quantified chemical intuition to a new synthesis. Most studies on HKUST-1 is focused on the Cu(II) version, but HKUST-1 can also be synthesized with Zn(II) [199]. We can now take three approaches to synthesize Zn-HKUST-1: First, we could assume that the synthesis of Zn-HKUST-1 to be similar to Cu-HKUST-1 and simply reuse the successful conditions of Cu-HKUST-1. For our case, the equivalent of a literature search of successful synthesis conditions for Cu-HKUST-1 is simply testing those optimal synthesis conditions we found for Cu(II). None of the top ten synthesis conditions for Cu(II) yield crystals for Zn(II). Without chemical intuition, this would put us back to square one, and we would have to restart the procedure, i.e., we use the same set of most diverse conditions as used for Cu-HKUST-1. Using our chemical intuition, however, we can sample the space more intelligently by assigning the previously determined importance of variables, resulting in denser sampling of more important experimental parameters. For this weighted set of 20



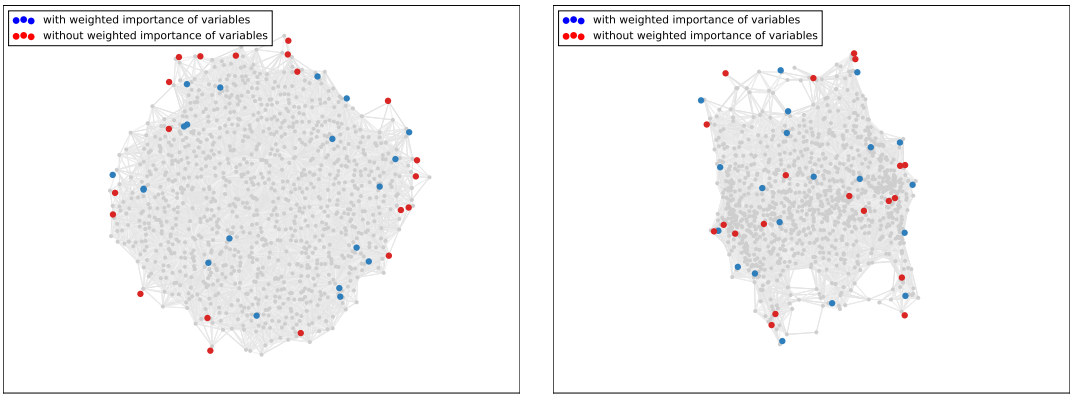
(a) (b)

Figure 3.3 – **Captured chemical intuition and the chemical space in the new metric.** (a) Relative impact of the 9 parameters on Cu–HKUST–1 synthesis, as obtained from machine learning. Maximum impact is normalized to one. The error bars show the standard deviation of the relative importance of variables over 1000 retraining of the random forest with different unique random seeds. (b) Multidimensional scaling projection of the experimental conditions, in which the distance is weighted by the relative importance of the variables. The colour of dots indicates the generation in the GA. The grey dots represent the chemical space in the new metric. Grey dots are the 1000 most diverse conditions obtained using MaxMin method without weighting distances.

diverse conditions, two conditions yielded Zn-HKUST-1 crystals.

The difference in weighted and unweighted synthesis conditions is illustrated in Figure 3.4. As we are sampling a high dimensional space with a low number of points, the most diverse conditions lie at the boundaries of each dimension, and only start populating the interior with sufficient sample points. In the weighted space representation (Figure 3.4b), the set generated without prior knowledge includes several points that are so close to each other that they are not expected to yield additional information. Having determined the (lack of) variation of the sample fitness for the different variables, the variables of lesser importance may be sampled less frequently without loss of accuracy. In fact, the reweighted set samples the most important parameters roughly 10 times more frequently than the least important ones. We note that our 20 intuition-based samples would need to be replaced by order of four to five thousand samples without intuition in order to maintain the same sampling accuracy, illustrating a dramatically increased chance of successful synthesis for a chemist who leverages chemical intuition.

The example of Cu-HKUST-1 and Zn-HKUST-1 illustrates how quantifying and reusing chemical intuition can be beneficial in a case, where the chemistry is too specific for the synthesis conditions themselves to be transferable. In this work we selected HKUST-1 as a case study to illustrate the methodology.



(a) (b)

Figure 3.4 – **Distribution of diverse sets in the chemical space of Zn–HKUST–1.** Multidimensional scaling projection of the set of 20 most diverse synthesis conditions with (blue) and without (red) taking the relative impact of synthesis variables into account. Both sets are shown in the unweighted space (a) and in the weighted space (b).

3.5 Outlook

The main aim of this work was to develop a simple, yet powerful framework that allows to use failed and partially successful experiments to systematically improve synthesis strategies. This framework does not rely on a detailed understanding of how the different synthesis conditions impact the outcome. Rather, it relies on the notion that, over the course of many experiments, chemists develop an intuition, over the course of many experiments, on how to approach the problem of finding the right synthesis conditions. Here, we have developed a simple way of capturing this chemical intuition using machine learning.

Our case study of HKUST-1 was intended as a proof of principle that we can capture and quantify chemical intuition, and effectively use it to develop more efficient synthesis strategies. We note that the data produced in this work are ideal from a machine learning point of view. Using a robotic platform provides precise control over the synthesis variables which results in less noise in the outcome of reactions and improved reproducibility. Furthermore, we are using only one synthesis technique. This allows to obtain an accurate estimate of the chemical intuition using a relatively small set of experiments. If all groups that have worked on the synthesis of HKUST-1 would have published also their failed and partially successful experiments, the data would be significantly less homogenous because of other influencing variables, e.g. size of reactor, purity of reactants, etc., but the much larger data set would also make it easier for machine learning to filter out these inhomogeneities.

Figure 3.5 summarizes how we envision the three components of our framework, synthesis, optimization, and machine learning, to interact. For example, one can use the genetic algorithms to optimize the synthesis conditions while, in parallel, machine learning the relative importance of the experimental variables, leading to more rational experiments. This is the

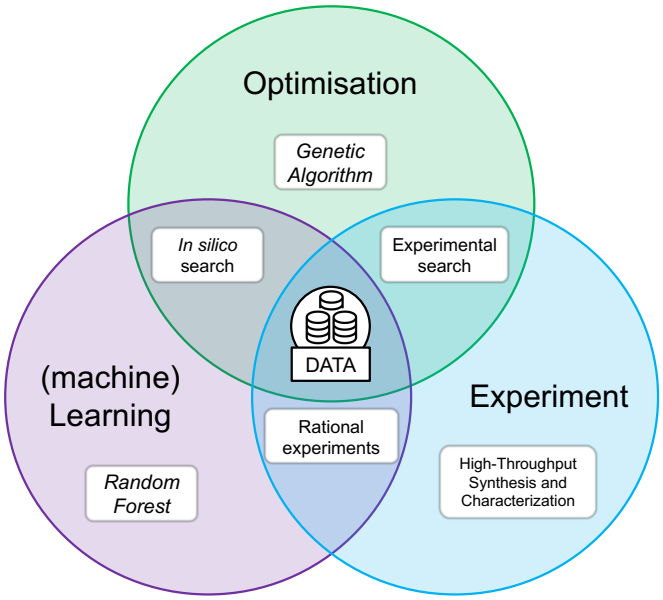


Figure 3.5 – **Schematic of the components of the framework used for MOF synthesis.**

approach we have used for HKUST-1. For more complex synthesis, however, one can take this approach one step further by leveraging the machine learning model in a second way: to score the next generations of the genetic algorithm *in silico*, going back to experiment only once convergence is reached. Appropriately fine-tuned, this has the potential to significantly reduce the number of experiments required (See section 3.7 in for details).

An important practical question is how we envision our approach can be used by other groups. The screening strategy we used can be easily adopted to other synthesis problems. Define the chemical space, generate the most diverse set of conditions, and use a combination of genetic algorithms and machine learning to find the optimal target. Of course, one can only take advantage of the "chemical intuition" in generating the set of most diverse conditions if we have a sufficient number of failed or partially successful experiments using a similar synthesis technique and similar chemical space. A key component here is that the more groups share their failed and partially successful experiments, the more versatile the model's chemical intuition will become. In this respect, each MOF synthesis group has a similar challenge, once the ligands and metal nodes are synthesized: how to find the right synthesis conditions that crystals will form? The quantified "intuition" by machine learning is by no means different from the intuition developed by chemist in the lab; it is useful in many cases, but one always need to keep in mind that in some cases the chemistry can be surprisingly different. The software we have developed for this study is available as a web application on the Materials Cloud [200]. together with the "chemical intuition" which we will be continuously updating and adopting to the needs of the community. If a large number of groups involved in MOF synthesis agree on a systematic reporting of failed or partially successful experiments, this can be an extremely powerful tool that has the potential to change the way our research

community approach synthetic chemistry.

3.6 Methods

To reconstruct the not reported (partially) failed and successful data in the literature, we simulate the steps that are taken by someone with no chemical intuition for synthesis of a MOF by a genetic algorithm (GA) optimization procedure. We start with the set of most diverse synthesis conditions based on a simple algorithm for the MaxMin diversity problem. Chemical intuition can be incorporated by assigning appropriate weights to different variables. The diverse set constitutes the first generation of the optimization cycle. A robotic synthesis and characterization approach is used for synthesis of MOFs, and measurement of X-ray diffraction patterns. We rank the experiments based on their crystallinity and BET surface area. This ranking is fed to the genetic algorithm to generate a new generation of synthesis conditions. Afterwards, the new generation is synthesized and characterized. This procedure continues until it satisfies the objective function of the synthesis. All the data generated in the synthesis procedure is used to train a machine learning model to assess the importance of synthesis variables. Below we summarize the main steps for each part of this procedure. A more detailed description can be found in section 3.7.

Genetic Algorithm

The genetic algorithm (GA) was used as it is implemented in the global optimization toolbox of MATLAB [201]. The population of each generation was fixed to thirty. At each step, the GA was initialized with the last generation and its individuals' fitness. Migration, crossover and mutation genetic functions were applied. The ranking of the individuals was used as the fitness function which determines the chance of each parent in generating children in new generation. The optimization starts with the set of most diverse individuals to ensure exploration of the chemical space with no bias.

Robotic Synthesis and Characterization

The synthesis was carried out in a microwave synthesis reactor (Biotage, Uppsala, Sweden) affixed on a HT robotic platform (Chemspeed technologies, Füllinsdorf, Basel, Switzerland). The synthesis steps inclusive of handling and dispensing of the reactants (metal, ligand, solvents) in to the microwave reaction vials, stirring of the dispensed reactant mixture, capping, crimping, and the transportation of the microwave reaction vials to the microwave reactor cavity was completely automated and executed using the Chemspeed autosuite software. All the chemicals were purchased from commercial sources and used without further purification. Powder X-ray diffraction (PXRD) patterns were collected using the powder diffractometer Bruker D8 Advance with TWIN/TWIN optics and LYNXEYE XE-T detector equipped with high throughput sample changer. The samples were loaded on a silicon (no background) sample

holder and the PXRD pattern was collected in a 2θ range between 2–20° using a monochromatic copper (Cu) X–ray source (λ = 1.54056 Å). The sample holders were rotated about their central axis during data collection, minimizing potential effects from preferred orientation. The diffractometer was controlled using the Bruker's EVA software. All measurements were performed at room temperature. Crystallinity and phase purity of samples were assessed by the full-width at half maximum (FWHM) of the diffraction peaks of the samples' powder X-ray diffraction patterns, and with a penalty in fitness for extra peaks compared to simulated pattern. N2 isotherms (77 K) were recorded to apply the Brunauer-Emmett-Teller (BET) model in the relative pressure range of 0.05-0.30 to determine the surface area of the HKUST-1 MOFs. The isotherms were collected by using an IGA system (Intelligent Gravimetric Analyzer, Hiden Isochema Ltd., Warrington, UK) and the BELSORP mini system (MicrotracBEL Corp., Osaka, Japan). Prior to isotherm collection, the HKUST-1 samples were activated at 220°C under dynamic vacuum for 6 hours to get the desolvated HKUST-1 (dark blue).

Machine Learning

The random forest ensemble learner was used for assessing the importance of variables [202]. Random forest is a supervised learning algorithm for classification and regression problems. A bootstrapped aggregated forest of 200 decision trees with maximum depth of three was trained to predict the outcome of the synthesis based on the synthesis variables. The mean absolute error (MAE) of the predictions was smaller than 9% and 14% for cross-validation and not seen data points, respectively. The importance of variables was estimated by permuting out-of-bag observations. The machine learning algorithm was implemented first using the statistics and machine learning toolbox of MATLAB, and then ported to python (using the scikit-learn package [120]) for the web application.

Multidimensional scaling plots

Multidimensional scaling (MDS) provides a visual representation of data based on the pairwise distances, similarity or dissimilarity within a set of points in a high-dimensional space. Here, we choose metric MDS using the weighted Euclidean pairwise distances between points in both high-dimensional (HD) and low-dimensional (LD) spaces. The algorithm aims to preserve the HD distances between objects in the LD representation. The metric for evaluation of how accurate the LD representation is compared to the high-dimensional distances is called the stress function: $S = (\sum_{i,j=1,\ldots,N} d_{i,j} - \overline{d}_{i,j})^{1/2}$.

This function returns the residual sum of squares of the distances in the HD space (d) to the LD space ($\overline{d}$). We use stress majorization algorithm to minimize the stress function as implemented in scikit-learn python package. The weights in the weighted Euclidian distance function, $d_{a,b} = \sqrt{\sum_i^n w_i(a_i - b_i)^2}$, are set to 1 for all variables in Figures 3.1a and 3.4a (no chemical intuition), and equal to the weighted importance of variables in Figure 3.3b and 3.4b (using chemical intuition).

3.7 Supplementary materials

Genetic Algorithm

Genetic algorithms (GA) are global search methods and the aim of a GA is to search the phase space constructed by the optimisation variables to find the global optimum of the objective function [203]. In a GA optimisation, the value of the objective function is evaluated by a population of individual explorers (chromosomes) distributed in the phase space. Each individual explorer is uniquely defined by its genes, which are the values of the optimisation variables. After evaluation of the objective function, GA randomly selects the good performing individuals to reproduce and create children in the form of a new generation of explorers by the crossover operation that combines parents to generate new children. Furthermore, to explore the not seen parts of the phase space, mutation happens. Hence, a new generation proposed by a GA is a combination of samples with good genes, with a controlled number of new genes. The quality of genes of explorers evolve toward an optimal solution over successive generations.

Here, we have implemented an adaptive genetic algorithm for synthesis of MOFs (Figure 3.6). All the codes are adapted from MATLAB global optimisation toolbox [201]. The GA probes the constrained chemical phase space constructed by the nine synthesis variables listed in Table 3.2. For the synthesis of HKUST-1 we have ensured that our range includes the successful synthesis conditions that are reported in the literature. The volume of solvent was allowed to vary between 1 to 6 ml, as an implicit function of the solvent composition, where 6ml is the maximum volume allowed by the robot. The population size of the GA was fixed to 30 chromosomes for each generation.

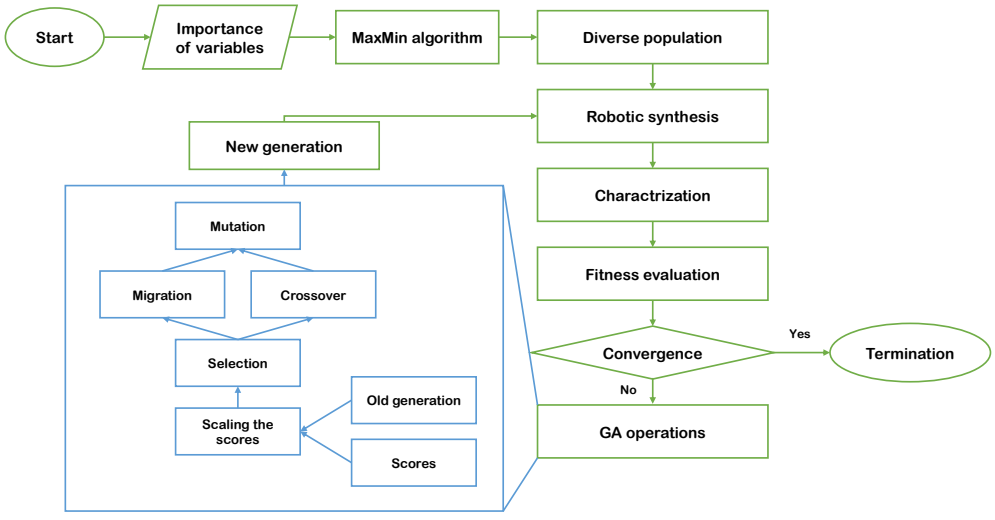


Figure 3.6 – Flowchart representation of the procedure used for synthesis of MOFs.

Synthesis variable	Optimisation constraints	Notes
Water (H_2O)	0-6 ml	The total solvent volume is constrained to be between 1ml to 6 ml.
Dimethylformamide (DMF)	0-6 ml	
Ethanol (EtOH)	0-6 ml	
Methanol (MeOH)	0-6 ml	
Isopropyl alcohol (iPrOH)	0-6 ml	
Reactants ratio	0.8-1.8	
Temperature	100-200 °C	
Microwave power	150-250 W	
Reaction time	2-60 min	

Table 3.2 – The synthesis variables constructing the chemical phase space, and their corresponding optimisation range.

To generate a new generation, the GA takes the chromosomes of the past generation and their corresponding fitness score based on an objective function. The fitness functions in current study are crystallinity or crystallinity and BET surface area of the first and the second generations, respectively. We use the full width at half maximum of powder X-ray diffraction patters as a measure of crystallinity of samples. The algorithm scales the scores with their ranking using $1/\sqrt{r}$, where r is the rank of each chromosome using tied rank for similar performing chromosomes. After scaling the scores, the GA produces 30 new children using the migration and crossover, constituting 10 percent and 90 percent of population, respectively. In migration operation, the genes of the top performing samples from old generation are transformed to the new generation. We use intermediate crossover function to respect the linear constrained of the optimisation. In the intermediate crossover scheme, the child is created using the weighted average of the parents, i.e. *child=parent1+random number*(parent1-parent2)*. The chance of being selected as a parent by the algorithm for crossover operation is proportional to the scaled score, i.e. the rank of chromosomes. The final step in the GA is to mutate the genes of the new chromosomes. This step is crucial to explore the not-seen part of the chemical space. We use gaussian mutation function where the mutated genes are chosen within a Gaussian distribution around the unmutated gene with a shrinking standard deviation of

$$\sigma = S\sigma_0, \tag{3.1}$$

where $S = 1/(Generation\ number)$ is the shrink factor and the initial standard deviation σ_0 was set to 0.2 of ranges of variables. An illustrative example of the genetic operation is shown in Figure 3.7.

The optimisation is initialized with the set of most diverse synthesis conditions based on the MaxMin algorithm. Starting with the set of most diverse genes and keeping a decent mutation rate are essential for efficient exploration of the chemical space and finding global optimum of the objective function.

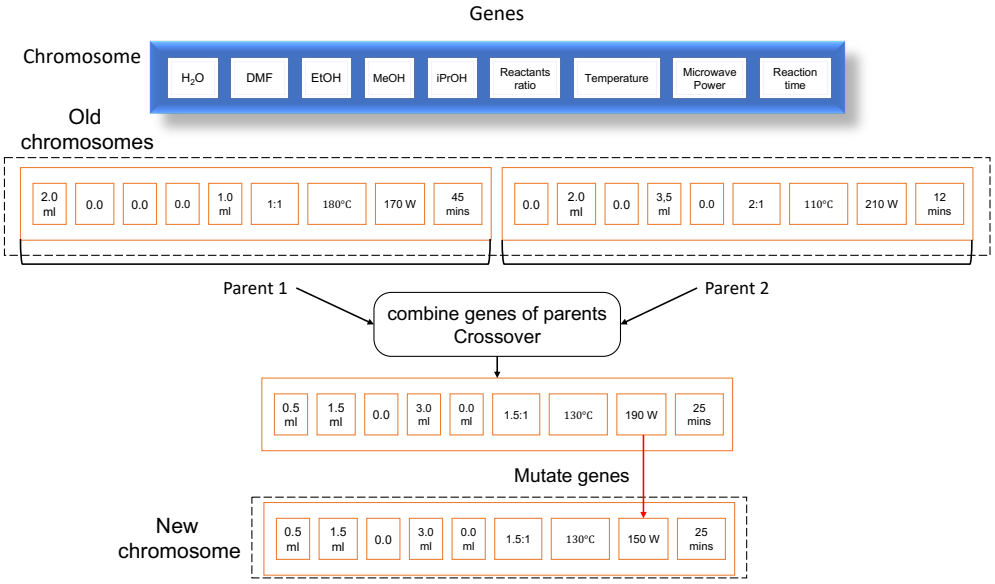


Figure 3.7 – A representative experimental condition and its transformation to the consecutive generation. Each experimental variable is a gene in a chromosome which is an experimental trial. The genetic algorithm operations generate new children using the genes of parents.

The crystallinity of each sample was assessed by the full width at half maximum (FWHM) of powder X-ray diffraction patterns (PXRD) [204, 205]. We start with separating peaks in the PXRD. Afterwards, a Gaussian function is fitted to the peak which give us the FWHM with the following set of equations:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\left(\frac{(x-x_0)^2}{2\sigma^2}\right)\right\}, \tag{3.2}$$

$$\text{FWHM} = 2\sqrt{2\log 2}\sigma, \tag{3.3}$$

where variable x is the 2θ of the diffraction angel. The average FWHM of all the peaks of the PXRD is taken as the measure of crystallinity. Lorentzian, Pearson, and combined Lorentzian, Pearson and Gaussian distributions were also considered, and no considerable differences were observed in the ranking.

For the selection step, GA only takes the ranking of the performance of individuals in the current population, and therefore, its objective function can easily be adapted for more optimization's goals, e.g. reaction yield, crystal morphology, etc. Moreover, adding or removing synthesis variables and conditions is straightforward.

In silico prediction of synthesis outcome

The trained model can be used for prediction of the outcome of synthesis without performing the experiments. For data we collected for H-KUST1, the mean absolute error (MAE) of the trained model for training size of 90 is 9% and 14% for cross-validation and new data points, respectively, which are indeed sufficient and satisfactory for estimating the outcome of a synthesis. Particularly, this prediction is useful to eliminate many chemical hypothesis (synthesis conditions to be tried) that would not yield any favourable outcome.

Moreover, this predictive model can be used to boost the convergence of the GA optimisation for synthesis of difficult MOFs. Since GA takes only the last generation into consideration to propose new synthesis conditions, by construction, after some generations, the algorithm might visit the already seen regions of space. This known weakness is associated to the deficient learning of the GA from the previous failed experiments.

Here, we propose a combined GA with machine learning to address this weakness. After each generation, we train the machine learning model, e.g. RF, and monitor the MAE of the prediction. We can rely on the prediction of the machine learning model after some generations when the MAE decreases to a satisfactory rate. Afterwards, we perform an *in silico* evaluation of the outcome of chemical hypothesis proposed by the GA for several generations until they converge to a set of optimal synthesis conditions based on the machine learning model predictions. Then, we experimentally synthesize the *in silico* predicted optimal conditions. The new experimental data is used to further refine the regression model, and the process is repeated until the objective function is satisfied (Figure 3.8). This procedure can save many experiments in the intermediate steps and help the GA to converge faster. Indeed, this procedure only required if one needs many GA steps to find satisfactory synthesis conditions.

The timeline of HKUST-1 synthesis

Several synthetic routes for HKUST-1 synthesis have been established and reported. Room temperature (RT) synthesis, conventional electric heating (CEH), microwave (MW), electrochemistry (EC), mechanochemistry (MC), and ultrasonication (US) methods are the commonly employed methods for the synthesis of HKUST-1. The prime objective is to identify the optimum synthetic conditions, which would enable the isolation of crystals or a micro crystalline (long range order), phase pure and porous HKUST-1 framework. Using these different synthesis methods, crystalline phase pure HKUST-1 could be obtained. A summary of several trials from literature using these approaches and the corresponding reported BET surface areas are shown in Figure 3.9.

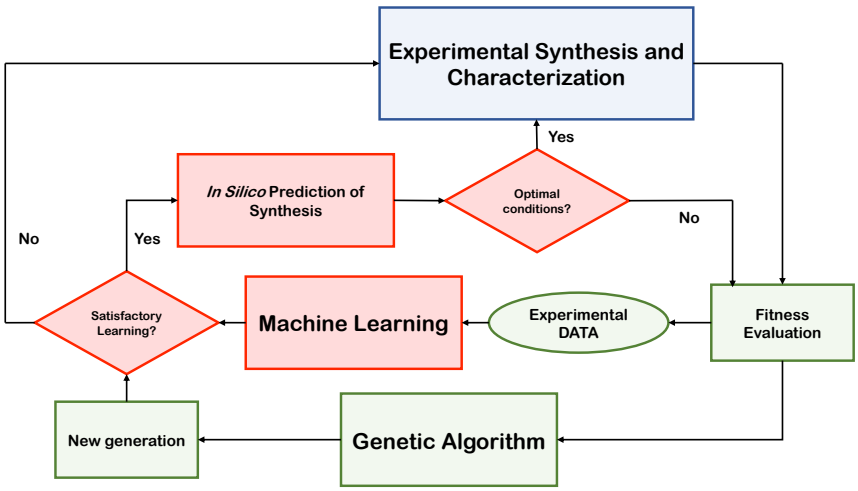


Figure 3.8 – Flowchart of the MOF synthesis accelerated by machine learning.

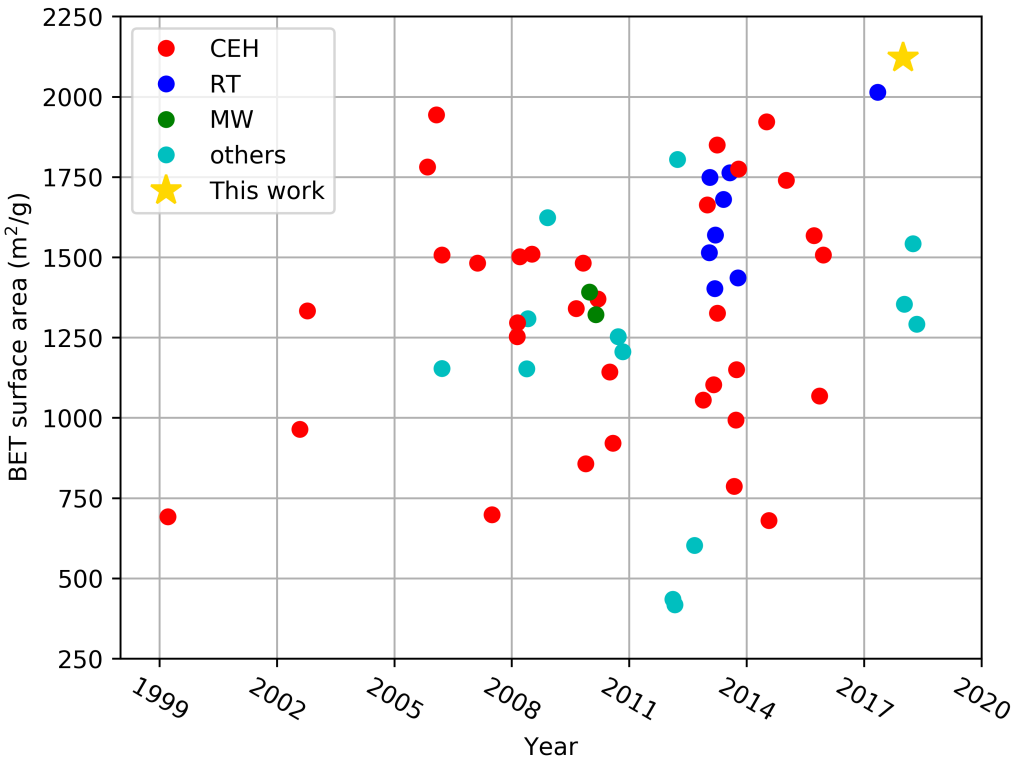


Figure 3.9 – Considerable amount of data lost in the literature as only successful synthesis conditions were reported for these data points (~50 data points). The reported BET surface area by many synthesis groups shows a wide range of variations. The horizontal axis (year) of the data points slightly shifted to avoid overlap.

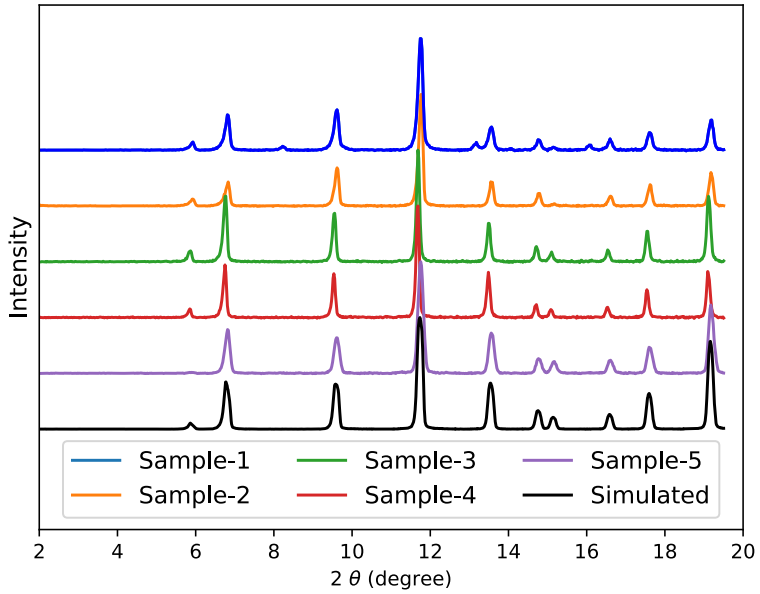
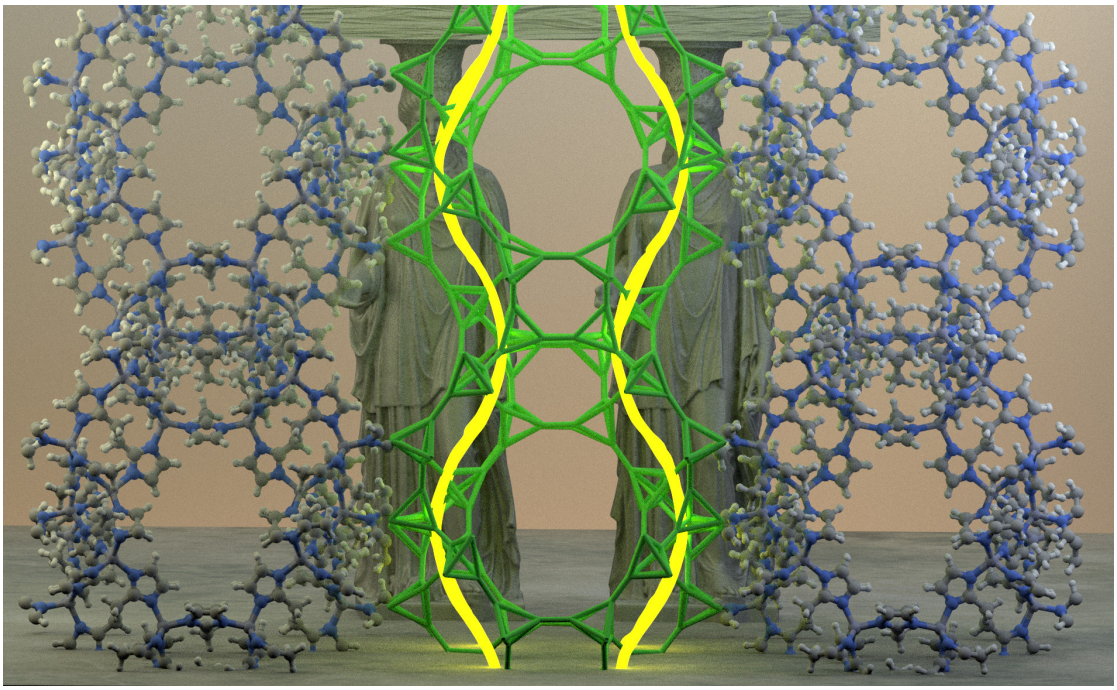


Figure 3.10 – Powder X-ray diffraction pattern of the five samples of Cu-HKUST-1 with high crystallinity and wide range of BET surface area discussed in the manuscript.

4 Improving mechanical stability of metal–organic frameworks using chemical Caryatids [1]



[1] postprint version of the article published as: Seyed Mohamad Moosavi, Peter G. Boyd, Lev Sarkisov, Berend Smit, *ACS Central Science*, 2018, 4, 7, 832-839, https://doi.org/10.1021/acscentsci.8b00157. S.M.M. implemeted the force fields in LAMMPS–interface and developed the code for computing mechanical properties. S.M.M. analysed data and together with B.S. wrote the manuscript with contributions from all authors.

Abstract

Metal-organic frameworks (MOFs) have emerged as versatile materials for applications ranging from gas separation and storage, catalysis, and sensing. The attractive feature of MOFs is that by changing the ligand and/or metal, they can be chemically tuned to perform optimally for a given application. In most, if not all, of these applications one also needs a material that has a sufficient mechanical stability, but our understanding of how changes in the chemical structure influence mechanical stability is limited. In this work, we rationalize how the mechanical properties of MOFs are related to framework bonding topology and ligand structure. We illustrate that the functional groups on the organic ligands can either enhance the mechanical stability through formation of a secondary network of non-bonded interactions or soften the material by destabilizing the bonded network of a MOF. In addition, we show that synergistic effect of the bonding network of the material and the secondary network is required to achieve optimal mechanical stability of a MOF. The developed molecular insights in this work can be used for systematic improvement of the mechanical stability of the materials by careful selection of the functional groups.

4.1 Introduction

Like any other material, metal-organic frameworks (MOFs), as an important class of porous materials with large diversity of pore shapes and sizes, and rich chemical functionalities must pass the stability criteria to be used in most practical applications [180, 206, 207]. Despite having superior performance for many applications, MOFs are vulnerable with respect to stability compared to the competing materials. For instance, due to the relatively weak metal-ligand coordination bonds, many MOFs are chemically unstable and have low endurance in different types of chemicals environments, e.g. acidic or basic environment [207]. Significant progress has been made in developing MOFs that are chemically stable, e.g. Zirconium based MOFs [208]. Since applications of MOFs often involve repetitive, cyclic temperature and pressure variations and capillary forces exerted by guest molecules, sufficient mechanical stability is of equal importance [77,78]. The mechanical stability for porous materials measures the stiffness of a material to withstand its pore size and structure under mechanical load. Clearly, deformations due to external pressure will disrupt pore shape and size, resulting in significantly reduced performance. In this study, we focus on strategies to improve the mechanical stability of a particular MOF.

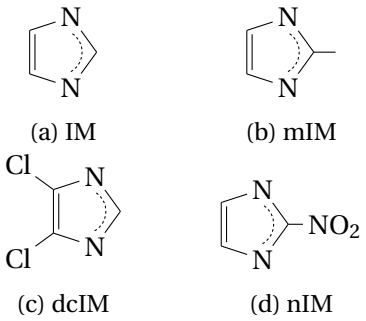
The mechanical properties of materials vary by several orders of magnitude with changing atomic composition and/or crystal structure [209–211]. As the mechanical stiffness, i.e. modulus of elasticity, typically scales quadratically with the density [212], mechanical stability is of particular importance for applications of low-density materials, such as MOFs [78, 213, 214]. For these materials special strategies are often required to improve their mechanical stability. Often these strategies are inspired by nature (e.g., wood and bones [215, 216]) and involve fractal and hierarchical design to make highly connected materials over multiple length

scales [217–219]. Indeed, improving the mechanical stability of MOFs by tuning the chemistry has become an important focus of attention [207, 220–222]. In analogy to the concept of high connectivity of the hierarchical design of materials, it has been shown that the MOFs with high degrees of framework interconnectivity, i.e. high coordination number of metal nodes, have improved the mechanical stability [220, 223]. However, not for all applications a particular MOF can be easily replaced, and therefore, Kapustin *et al.* developed a strategy to retrofit a particular MOF by adding additional ligands to the framework [222]. This strategy is robust but limited to the MOFs that permit ligand installation [224, 225]. In both cases, the mechanical stability is improved by increasing the connectivity of the bonding topology.

In this work, we explore the option of decorating the organic ligands of a MOF with functional groups. The significant progress in computational material science in *in silico* generation of MOFs [64, 66] and reliably prediction of their mechanical properties [213, 226] permits studying a large and diverse set of materials to extract structure-property relationships to design materials with enhanced mechanical stability. We show that the non-bonded interactions play an important role in the stiffness of the materials, and therefore, strategically placed functional groups can introduce extra framework connectivity via non-bonded interactions. This secondary network of non-bonded interactions can enhance the mechanical stability of the framework considerably. We use the term "chemical Caryatids" for those functional groups that are contributing in carrying the mechanical load applied to the material. In addition, we show that the optimum mechanical stability of a MOF framework is obtained by the cooperative effect of the primary network, determined by the bonding topology, and the secondary network, which is governed by the non-bonded interactions.

4.2 Results and discussion

In this work, we focus on Zeolitic Imidazolate Frameworks (ZIFs), which are a special class of MOFs comprised of four coordinated metals, typically Zinc, with imidazolate (IM) derivative ligands. ZIFs are an ideal case study for our work because they all have the same coordination environment, but diverse bonding topologies and functional groups [227, 228]. This allows us to focus on the effects of bonding topology and functional groups on the mechanical properties, while keeping coordination environment fixed, i.e. keeping the same metal node. In addition, because of the pioneering work of Cheetham and co-workers, ZIFs are among the very few MOFs for which systematic research has been done on their mechanical stability [77, 78, 214]. To characterize the mechanical properties of ZIFs, Cheetham and co-workers used nano-indentation to measure the Young's modulus, i.e. the resistance of materials to the tensile stress [78]. These and related studies concluded that for these materials the mechanical properties can be described with the low density-stiffness correlation [78, 229–232]. As these experiments require sufficiently large single crystals, the number of studied structures is relatively small compared to the total number of possible ZIFs. In this work, we expand the studied materials to, in addition to the known ZIF structures, a large set of *in silico* constructed materials using fifty different zeolite topologies [233] with four type of ligands. Such a large set



(a) IM (b) mIM

(c) dcIM (d) nIM

Figure 4.1 – The four different ligands used to construct hypothetical materials. (a) IM = imidazolate, (b) mIM = 2-methylimidazolate, (c) dcIM = dichloroimidazolate, (d) nIM = 2-nitroimidazolate.

of materials allows us to cover a representative range of bonding topologies and functional groups. The ligands used for *in silico* construction of materials include the commonly [61,227] used derivatives of IM shown in figure 4.1.

Theoretically, mechanical properties of materials are described by their stiffness matrix [234]. Young's and other moduli of elasticity, including bulk and shear modulus, which characterize material's resistance to hydrostatic pressure and shear stress, respectively, can be extracted from the stiffness matrix. Since the mechanical properties of the materials in our study do not involve the breaking/formation of chemical bonds and other quantum effects, we used an approach based on a classical force field to compute the stiffness matrix for each material. The reliability of our force field is evaluated by comparison with the experimental and *ab initio* calculated values of Young's modulus reported in the literature. Figure 4.2 shows a comparable agreement between the *ab initio* and force field results with the experimental data, supporting the conclusion of our previous work that these classical force fields yield sufficiently reliable data on the mechanical properties of these materials [80].

If we focus on those materials in figure 4.2 for which experimental data are available, we observe the same low density-stiffness correlation as found experimentally [78]. However, if we include all our data, the picture becomes quite different. By expanding the chemistry and topology of ZIF structures, figure 4.2 shows large deviations from the density-stiffness correlation. Changing the underlying network topology and/or ligand can lead to larger variations in mechanical stability than changes in density, and in some cases, even reverse the trend. For instance, many ZIF structures with dcIM ligand have similar or lower stiffness in comparison to the structures in mIM and nIM ligand families, although they have higher density. A molecular-level explanation of these deviations is provided below, the understanding of which will allow us to exploit the chemical and topological features of a material to improve its mechanical stability.

The structures in figure 4.2 differ in their bonding topology and/or functional group of ligand. We introduce a computational approach to disentangle the effects of changes of the topology from changes of the ligand. To distinguish the role of the bonding topology on the mechanical

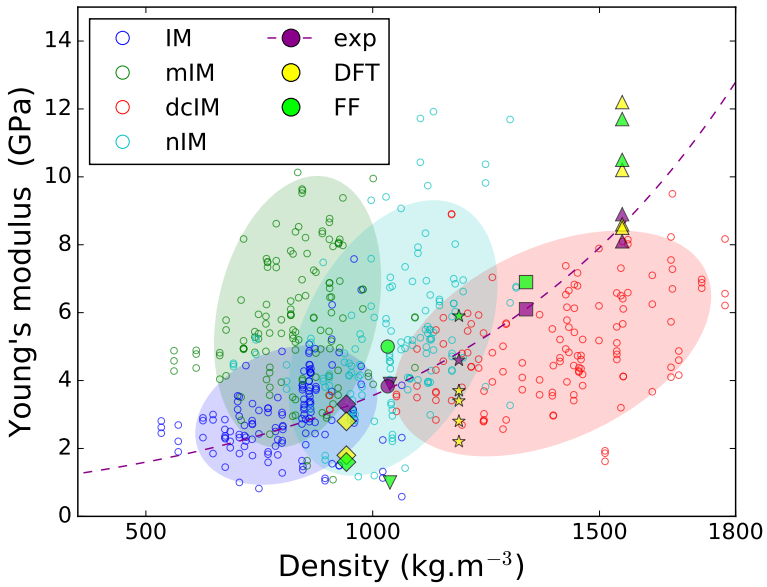


Figure 4.2 – Young's modulus versus density; for each material we plot the value along each of the three lattice principle axes. The filled markers with unique marker for each structure are used for those structures we can compare our force field (FF) with experimental (exp) or *ab initio* density functional theory (DFT) calculations, with the markers representing: ◆: ZIF-8 [78, 214], ●: ZIF-20 [78], ▼: ZIF-68 [78], ★: ZIF-4 [78, 230, 231], ■: ZIF-7 [78] and ▲: ZIF-zni [78, 230, 232]. The color coding is used to indicate the different ligands. If the density-stiffness correlation were perfectly obeyed a principle component analysis would give a narrow cloud around the dashed line. The clouds derived from principle component analysis demonstrate the deviations for the different ligands. The complete set of data can be found on materials cloud [235].

properties, we first look at the mechanical stability of a simplified network of atoms comprised of atomic bonding, and we refer to this network as the primary network. Several approaches have been used to define such a primary network [223, 236]. Here, we define the primary network as the ZIF structure in the absence of non-bonded interactions. Since the ligands in our study only differ in their functional groups, the primary network of the structures with the same underlying network topology but different ligands are nearly identical. Hence, we expect similar mechanical properties for the structures with the same underlying network topology. Indeed, figure 4.3a shows that all ZIFs with the same topology have similar bulk and shear modulus, and hence, superpose on each other.

Figure 4.3b and 4.3c show the effects of switching on the non-bonded interactions where a large effect of functionalization on mechanical properties is observed. As there is no functional group on the IM ligand, it can be seen as bare backbone, and we see that the mechanical properties for this ligand are indeed dominated by the primary network. However, for the other ligands, functionalization can have a large effect on some topologies while on others surprisingly little. Moreover, although mIM, dcIM and nIM exhibit observable contributions to the stiffness of ZIF structures in comparison to IM, depending on the topology one functional

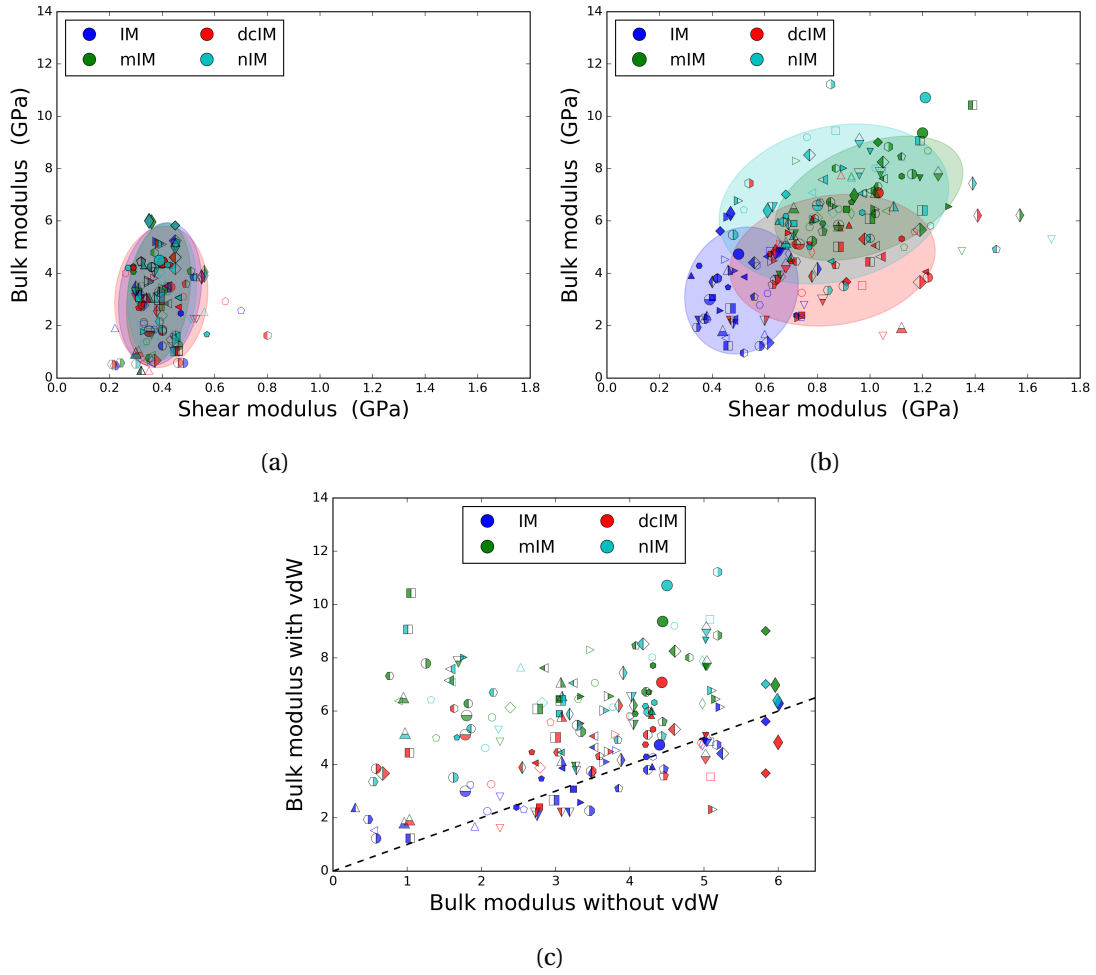


Figure 4.3 – Differentiating the contributions from bonding topology (primary network) and non-bonded interactions (secondary network) in mechanical properties. Considerable contribution from the secondary network is observed in some of the materials with functional groups. (a) and (b) Bulk modulus with respect to shear modulus of the materials computed without and with non-bonded interactions, respectively. (c) Bulk modulus of the materials versus the bulk modulus of them without non-bonded interactions. Dashed line represents identical properties computed with and without non-bonded interactions, i.e. no contribution from the secondary network. In all sub-figures, each marker (open, half-filled and filled) represent a unique underlying network topology while the colours represent the ligand.

group might show greater enhancement. For instance, for LTA topology, mIM gives higher stiffness, while for GIS topology (ZIF-6 [61,237]) dcIM has higher stiffness. Similar changes in the mechanical properties were observed experimentally for ZIFs with the same underlying net but different functional groups and was associated to the ligand-ligand interactions [78].

It is instructive to try to explain these deviations with a simple extension of the density-stiffness model. This model assumes a solid which has only non-bonded interactions, for example, a primitive cubic lattice with only nearest neighbour, (Lennard-Jones type) pairwise interactions. In this simple model, the only variable is the density dependent nearest neighbour distance. The bulk modulus is given by the second derivative of the potential energy of the crystal with respect to isotropic deformations. The second derivative of the Lennard-Jones potential changes sign from positive to negative at $\sim 1.2\sigma$, where σ is the van der Waals radius (see Figure 4.7a). As the second derivative for each pairwise interaction can be positive or negative depending on the nearest neighbour distance, the bulk modulus of this simple solid consists of a sum of positive or negative contributions, giving the well known density-stiffness correlation. In a ZIF structure, however, there is a distribution of inter-atomic distances, some have a positive contribution (i.e. stiffening interactions) and some have a negative contribution (i.e. softening interactions) to bulk modulus. One can argue that this distribution depends on the topology and functional group. If we now assume that the contributions of the non-bonded interactions are independent of the contribution of the primary network, we can obtain a simple correction to the density-stiffness correlation by adding the sum of the contribution of the non-bonded interactions to the bulk modulus resulting from the primary network.

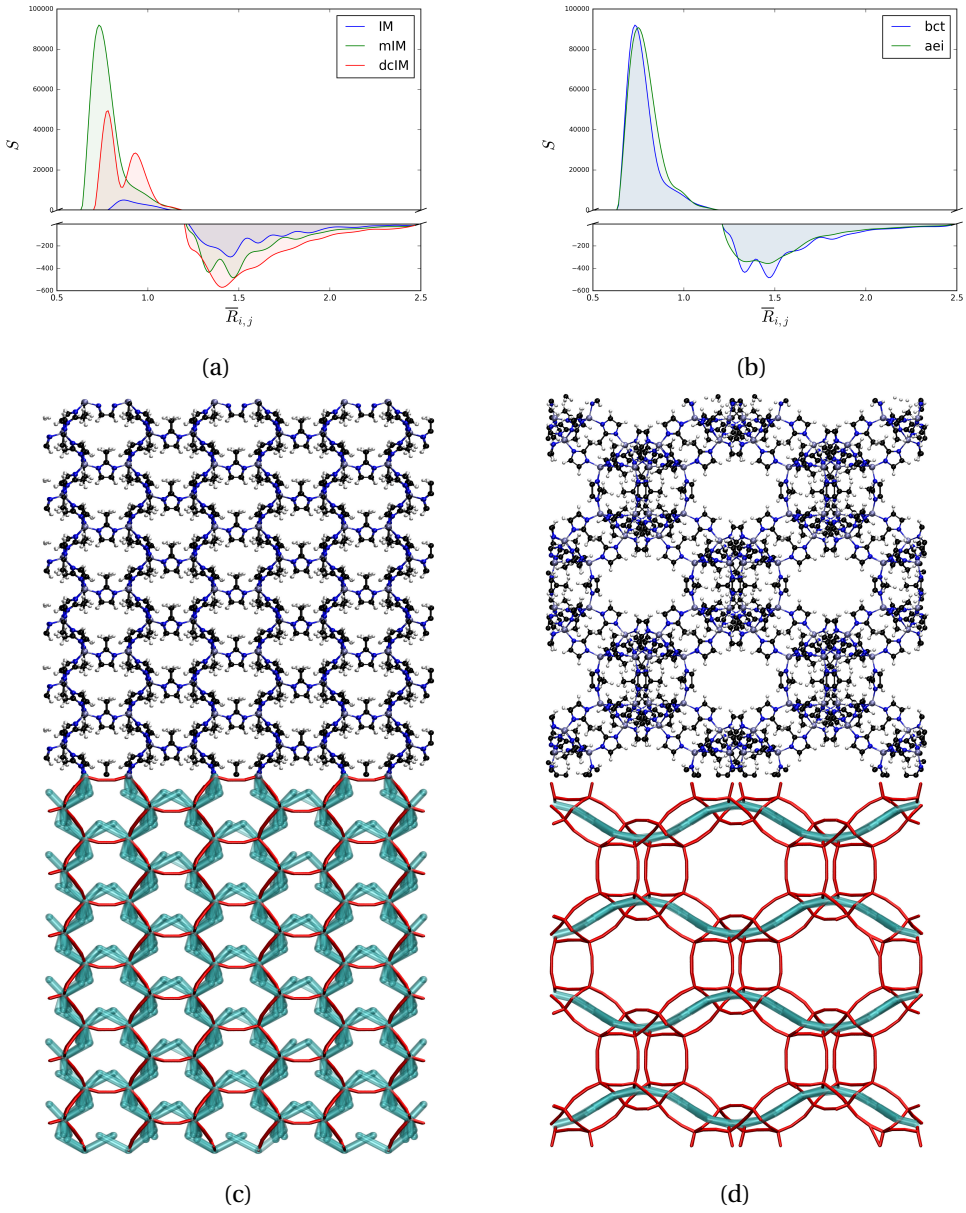


Figure 4.4 – Distribution of stiffening/softening non-bonded contributions for: (a) structures with BCT topology and IM, mIM and dcIM ligands, and (b) structures with mIM ligand and BCT and AEI topologies. The vertical axes represent the sum of second derivative of van der Waals (vdW) energy ($S = \sum \partial^2 E_{vdW} / \partial (r_{i,j}/\sigma_{i,j})^2$) plotted with respect to the inter-atomic distances normalized with vdW radii ($\overline{R}_{i,j} = r_{i,j}/\sigma_{i,j}$). The bulk moduli are 4.7, 9.4 and 7.1 for the BCT structures with IM, mIM and dcIM ligands, and 6.3 and 7.0 for the AEI structures with IM and mIM ligands, respectively (values are in GPa). (c) and (d) The atomic representation and the primary and secondary networks for the mIM ligand structures with BCT and AEI topologies, respectively. Details of ligands and metals were omitted in visualization of the primary and secondary networks for clarity. The primary net is demonstrated with red tubes and secondary net with cyan tubes; white, black, blue, and grey spheres represent H, C, N, and Zn atoms, respectively. The corresponding structures with IM ligand have the same primary net and no secondary net.

In figure 4.4a, we plot the distribution of stiffening/softening contributions for the ZIFs with BCT topology for three different ligands. The BCT zeolite topology include some known ZIFs, e.g. ZIF-1 [61, 237]. As expected, for IM, which has no functional group, this contribution is small. For the mIM and dcIM ligands, figure 4.4a shows higher peaks in the stiffening regime which is consistent with the observed increase in mechanical stability due to functionalization. Figure 4.4b shows an example of two materials in which the distributions of the stiffening and softening contributions are nearly identical. For the BCT topology structure we observe the expected stiffening compared to the primary network. However, for AEI topology we observe only a small effect of the non-bonded interactions on the bulk modulus. This is where our simple correction to the density-stiffness correlation breaks down. This example illustrates that the contributions of non-bonded interactions and the primary network to the stiffness can be highly non-additive. The reason for this non-additive behaviour becomes clear by introducing the concept of a secondary network.

We define the secondary network by connecting pairs of atoms with non-bonded interactions that have a stiffening contribution to the bulk modulus. Figure 4.4c and 4.4d shows the primary (red tubes) and secondary (cyan tubes) networks for the BCT and AEI topologies, respectively. Both materials have a 3D percolating primary network, but the pronounced difference is in the secondary networks. For BCT the secondary network is percolating in all three dimensions, while for AEI topology it percolates only in one dimension, and there are no contributions in the other two dimensions. Inspection of the primary network of AEI topology shows that the weak spots are on the ligands while the backbone is relatively stiff. Figure 4.4d shows that the corresponding secondary network reinforces this stiff backbone, but not the links between the backbones. Hence, the secondary network is only supporting AEI topology in a direction in which the primary network is already strong. As the mechanical properties are dominated by the weakest link, we now understand why we see such a small effect of the secondary network on the mechanical properties. To have an effect, we need to add a functional group that would form a secondary network orthogonal to the current network which it would significantly increase the bulk modulus. This type of synergy between the primary and secondary networks explains why some topologies show a large effect of functionalization, while for others this effect can be small.

It is interesting to apply our concept of primary and secondary networks to MOF-520-BPDC. Kapustin *et al.* [222] retrofitted the mechanically unstable MOF-520 by adding an additional linker to to allow for its use at high pressures. This retrofitting procedure changes the underlying network topology from **fon** net to more connected **skl** net [238]. This improved mechanical stability can be explained in terms of changes of the primary network (see subsection of mechanical properties of MOF-520 and MOF-520-BPDC in supplementary information in section 4.5). This form of topological tunability is very robust. However, it does rely on the ability to add extra linkers to support the weak spots of the primary network, which can be challenging from a chemical point of view for most materials.

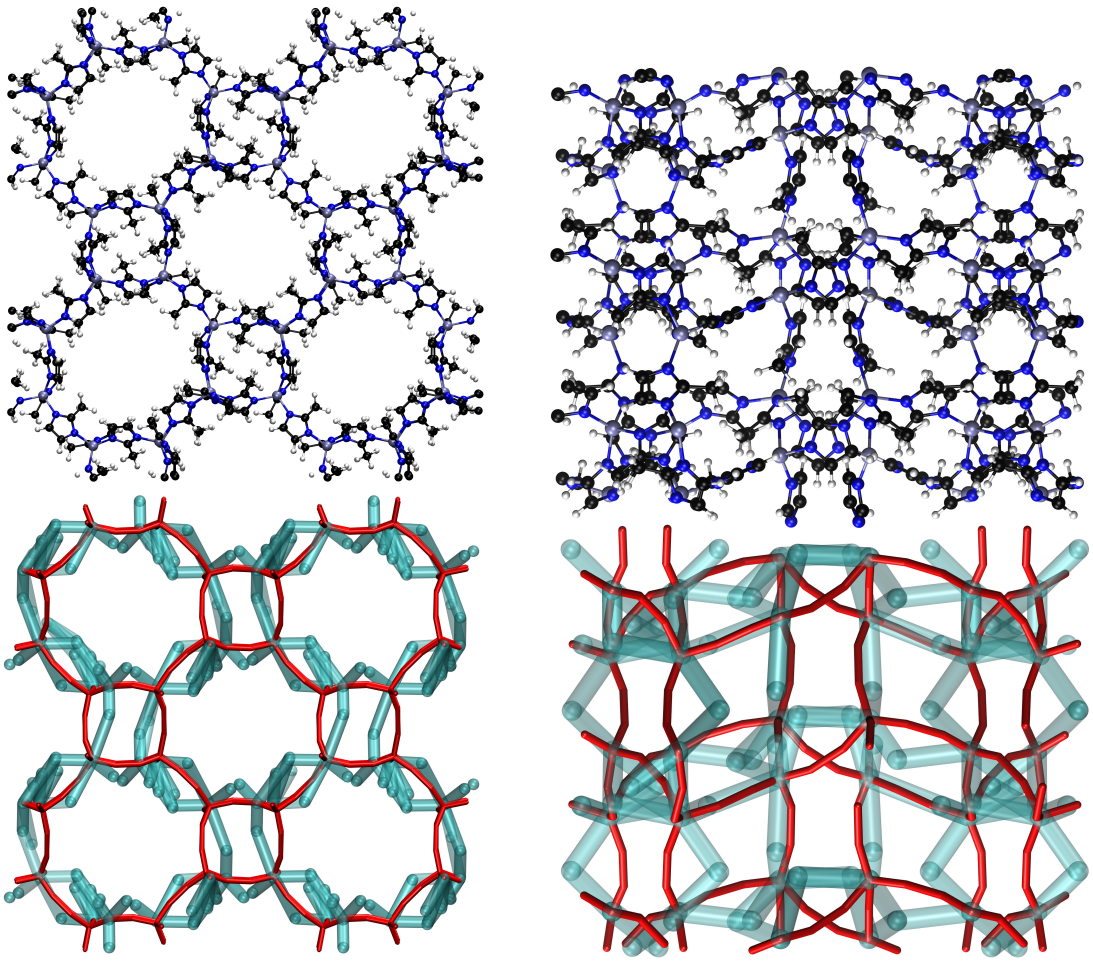
Alternatively, the mechanical properties of MOFs can be tuned by creating a secondary network

via ligand functionalization. The presence of such a secondary network can shed some light on the experimental observation on the amorphization of ZIFs [239]. Amorphization is directly related to the mechanical stability of these materials [240]. Cheetham and co-workers showed that ZIFs with the bare IM ligand amorphize relatively easily under pressure and heating, while the corresponding ZIFs with functionalization ligands required extreme conditions, specifically, they observed thermal amorphization only in ZIFs with the bare IM ligand [239, 241]. These results are consistent with our molecular dynamics simulations (see amorphization of ZIF-3 and ZIF-4 in section 4.5). Our analysis of the mechanical stability shows that "switching on" the secondary network in ZIF-3 and ZIF-4 improve the mechanical stability by as much as ~ 80% in shear modulus of both structures, and 300% and 150% in their bulk modulus, respectively. Figure 4.5 shows that for both ZIFs the functionalized structures have a secondary network that spans the entire unit cell in all three directions. Such increased mechanical stability explains why these materials are stable at conditions where the unsubstituted IM structure amorphize.

4.3 Conclusions

Our study shows that there are two strategies to improve the mechanical stability of a nanoporous material: modifying the primary and/or secondary network. Changing the primary network can be challenging as it requires the addition of extra linkers. In this respect the work of Kapustin *et al* [222] is a remarkable, but exceptional achievement. functionalization of ligands to create or modify the secondary network, much like the Caryatids holding up the porch of the Erechtheion on the Acropolis, might be a more generally applicable route. Our study shows that such a network, however, is only effective if it supports the weak points of the primary network.

It is interesting to envision how these results can be used from an experimental perspective. Suppose we have a particular MOF for a given application, but the mechanical stability needs to be improved. As the tools developed in this work are applicable to any MOF, we can determine the primary and secondary network of this material. If this analysis shows weak spots, a simple screening of different functional groups should give a clear prediction whether the mechanical properties of the material can be improved. As these functional groups may change the details of the pores, other computational tools should be used to ensure that these changes do not influence the performance of the modified material.



(a) (b)

Figure 4.5 – (a) and (b) Atomic representation and the primary and secondary networks of ZIF-3 and ZIF-4 structures with mIM ligands, respectively. The corresponding structures with IM ligand have the same primary net and no secondary net. The bulk and shear moduli for ZIF-3 (ZIF-4) are 2.0 and 0.53 (3.1 and 0.80) for IM and 7.8 and 0.96 (7.6 and 1.49) for mIM structures, respectively (values are in GPa). The functional groups of the ligands form a secondary network which enhance the mechanical stability considerably. The primary net is demonstrated with red tubes and secondary net with cyan tubes; white, black, blue, and grey spheres represent H, C, N, and Zn atoms, respectively.

4.4 Methods

To compute the mechanical properties of a materials we start with the crystal structure either from experimental or from an *in silico* predicted structure. The procedure of computing the mechanical properties relies on the assumption that the structure corresponds to the minimum energy configuration that is consistent with the force field used to describe the potential energy surface of the material. We developed a structural minimisation procedure to efficiently obtain this minimum energy configuration for all materials. All calculations were carried out within the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) molecular simulation package [242]. The VMD–Visualize Molecular Dynamics package was used for the structural figures and visualization of the primary and secondary networks [243]. No unexpected or unusually high safety hazards were encountered. Below we summarise the computational procedures that we have used. A more detail description can be found in the supplementary information.

4.4.1 Hypothetical material generation

Each material was assembled with the ToBasCCo algorithm [87], using a representative set of fifty zeolite topologies. Input into the program were the underlying networks, as obtained from the International Zeolite Association website [233], and two geometric building blocks; a 4-connected tetrahedral (Zn^{2+}) and 2-connected imidazole type ligands. This procedure yielded 200 materials, i.e. fifty structures for each of the four types of ligands, IM, nIM, mIM, and dcIM. All the structures are available through the materials cloud website and supplementary information.

4.4.2 Structure minimisation procedure

Simulated annealing algorithm was used to minimise lattice parameters and atomic sites using DREIDING force field [244] as implemented [80] in LAMMPS for all the structures. To avoid getting trapped in local minima we combined temperature annealing with expansion/relaxation cycles. The details of algorithm and its efficiency are discussed below.

The stiffness matrix represents the multidimensional curvature of the potential energy surface (PES) of a material with respect to infinitesimal lattice shape and size deformations. Indeed, this curvature is dependent on the atomic configuration, including lattice parameters and atomic sites. Consequently, the very first step in calculation of the mechanical properties is to find the minimum energy configuration of the system. This miminum energy configuration is dependent on the force field used to describe the potential energy surface. Depending on the number of degrees of freedom of a system, it can be challenging to find the global minimum of the PES of a strcuture. Conventionally, gradient based minimization methods have been widely used to relax atomic sites and lattice parameters in an iterative way to find the minimum energy configuration [245]. For MOFs with typically large number of atoms in a

method	Framework energy	K_{VHR}	G_{VHR}	E_{xx}	E_{yy}	E_{zz}
simulated annealing	825.7	8.8	1.1	8.0	8.0	8.0
conjugate gradient	831.3	6.4	1.25	5.3	5.0	6.1

Table 4.1 – The properties of ZIFs with mIM ligand and **sod** topology minimised with two methods. The lattice energies are normalized with the number of Zinc atoms and reported in kcal/mol/Zn. The moduli of elasticity are in GPa.

unitcell and large cell parameters, it is non-trivial to minimize both lattice parameters and atomic positions as the degrees of freedom is huge for optimisation. Specially, structures with the same underlying network topology but rotated ligands can introduce more complexity for the optimisation [246, 247]. Thus, the gradient based minimization methods are susceptible to fail and converge to a local minimimum. Simulated annealing minimization is a method to escape these local minima by adding sufficient energy to overcome energy barriers and reducing it slowly, the system evolves eventually to the global minimum [248].

Since the energy barrier of rotation of ligands are significantly higher than thermal energy due to atomic overlap and steric effects, we introduce a combination of thermal and mechanical simulated annealing for minimizing both lattice parameters and atomic coordinates. To provide the required space and energy for rotation of ligands, each structure was expanded 15% in all directions and heated to 550K. This was followed by sequential molecular dynamics at constant volume and temperature with gradual relaxation of the thermal energy and lattice parameters. This procedure was repeated until the potential energy converged to a minimum. The expansion factor, heating temperature and length of simulation were optimised for the purpose of minimizing ZIFs in our study by evaluating the final crystal energy.

ZIF-8, a known and well-studied material, was taken as a system [227] for evaluation of this procedure. ZIF-8 is comprised of Zinc and mIM ligands with Sodalite (**sod**) topology. The *in silico* constructed ZIF with the same topology and chemical composition was minimized with two methods, namely conjugate gradient and annealing. The lattice energy and powder X-ray diffraction patter of the two resulting structures are compared with the experimental structure (Figure 4.6a and 4.6b). As it can be seen, the powder diffraction pattern and crystal energy of the anneal structure have converged to the experimental ZIF-8 structure. However, the gradient based optimisation trapped in a local minimum. Although the structure resulting from the gradient based optimisation has the same chemical composition and topology, the pores of the Sodalite cage are distorted (Figure 4.6c and 4.6d). Crucially for the purpose of our study, the two structure show different mechanical behaviour (See table 4.1). Notably, without careful inspection of the powder diffraction and crystal energy, one cannot distinguish the correct structure, as both structures are mechanically stable, i.e. all the eigen values of their stiffness matrix are positive.

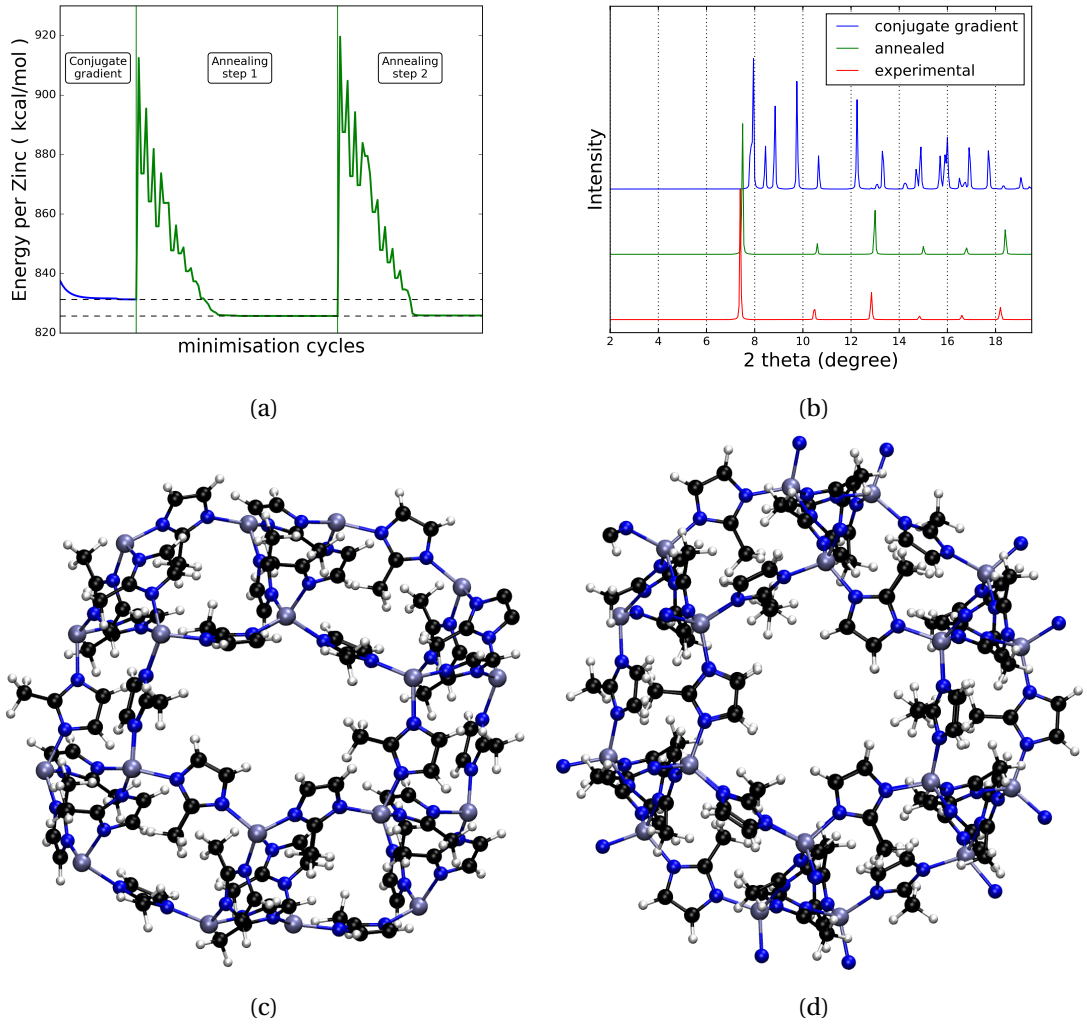


Figure 4.6 – Evaluation and significance of minimisation schemes. (a) Evolution of the crystal energy during minimisation. The conjugate gradient method fails and traps in a local minimum. (b) Powder x-ray diffraction pattern of structures minimised with conjugate gradient and simulated annealing and the corresponding experimental structure [61]. (c) and (d) Sodalite cage of minimised structures with conjugate gradient and simulated annealing minimisation , respectively. While the annealed structure recovers the symmetric sodalite cage, the conjugate gradient minimisation winds up in a structure with distorted cages.

4.4.3 Calculation of the mechanical properties

Calculation of the moduli of elasticity for a crystal with an arbitrary shape requires matrix representation of the mechanical properties. In continuum mechanics, for a material in elastic regime, the normalized structural deformation (strain) is linearly proportional to the applied mechanical load (stress). Stress ($\boldsymbol{\sigma}$) and strain ($\boldsymbol{\epsilon}$) are both second order tensors which are 3×3 matrices defining the directional values of these two quantities for a finite element of the material.

$$\boldsymbol{\sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix} \tag{4.1}$$

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{pmatrix} \tag{4.2}$$

The correlation between these two second order tensors is a property of the material, which is expressed by a forth order tensor (9×9 matrix) called stiffness tensor (**C**) or its inverse, compliance tensor (**S**).

$$\sigma_{ij} = C_{ijkl}\epsilon_{kl} \tag{4.3}$$

$$\epsilon_{ij} = S_{ijkl}\sigma_{kl} \tag{4.4}$$

$$\mathbf{S} = \mathbf{C}^{-1} \tag{4.5}$$

Both stress and strain tensors are symmetric and have six independent elements. In Voigt notation by keeping only these six elements, the stress and strain are represented by 6×1 matrices.

$$\boldsymbol{\sigma} = (\sigma_{11}\ \sigma_{22}\ \sigma_{33}\ \sigma_{23}\ \sigma_{13}\ \sigma_{12})^T \tag{4.6}$$

$$\boldsymbol{\epsilon} = (\epsilon_{11}\ \epsilon_{22}\ \epsilon_{33}\ \epsilon_{23}\ \epsilon_{13}\ \epsilon_{12})^T \tag{4.7}$$

Accordingly, the stiffness and compliance tensors are reduced to 6×6 matrices. It can be shown that, these matrices are symmetric, and hence have 21 independent elements. A detailed explanation and derivation of the elasticity theory and its tensorial representation

can be found in these references [234, 249]. Stiffness and compliance tensors contain all the mechanical properties of a material in elastic regime. Integration of the stress-strain curve, starting from the relaxed structure to an arbitrary strain, gives the energy difference between the two states. Thus, by computing this energy difference for all the possible deformations, one can calculate the corresponding elements of the stiffness tensor. A similar procedure is used for calculation of stiffness matrix by using finite differences of the crystal's energy relative to its ground state energy by applying series of deformation to the lattice parameters. The elements of the stiffness matrix are evaluated by fitting second order polynomials to the energy-strain curves. Previously, this method has been widely used for calculation of stiffness matrix based on force field or *ab initio* energy [250, 251]. All 21 elements of the stiffness matrix evaluated without taking symmetry into account. For each element, 21 different strain rates imposed with the maximum strain of 1%.

Conventionally, moduli of elasticity are extracted from the stiffness tensor based on strain (Voigt average), or from the compliance tensor based on stress (Reuss average). While Voigt averages provide the upper bound of the properties, the Reuss averages provide their lower bound. In our study, Voigt-Reuss-Hill averages, which are the average of these two bounds, are used as the representative properties of the materials. These average properties are computed with the following equations where V, R and H stand for Voigt, Reuss and Hill conventions, and K and G represent bulk and shear modulus, respectively.

$$K_V = \frac{c_{11} + c_{22} + c_{33} + 2(c_{12} + c_{13} + c_{23})}{9} \tag{4.8}$$

$$K_R = \frac{1}{s_{11} + s_{22} + s_{33} + 2(s_{12} + s_{13} + s_{23})} \tag{4.9}$$

$$K_{VRH} = \frac{K_V + K_R}{2} \tag{4.10}$$

$$G_V = \frac{c_{11} + c_{22} + c_{33} - (c_{12} + c_{13} + c_{23}) + 3(c_{44} + c_{55} + c_{66})}{15} \tag{4.11}$$

$$G_R = \frac{15}{4(s_{11} + s_{22} + s_{33}) - 4(s_{12} + s_{13} + s_{23}) + 3(s_{44} + s_{55} + s_{66})} \tag{4.12}$$

$$G_{VRH} = \frac{G_V + G_R}{2} \tag{4.13}$$

Although one can define a single value for the Young's modulus of a crystal, directional evaluation of the Young's modulus is more informative as these materials are highly anisotropic [213]. Young's modulus along an arbitrary unit vector **a** is defined as the inverse of compliance,

$$E_{\mathbf{a}} = (s_{\mathbf{a}}^{'})^{-1}, \tag{4.14}$$

where compliance along the vector is defined as

$$s_{\mathbf{a}}^{'} = \sum_{i}^{3}\sum_{j}^{3}\sum_{k}^{3}\sum_{l}^{3} a_i a_j a_k a_l s_{ijkl}. \tag{4.15}$$

In this equation, s_{ijkl} is an element of the compliance matrix **S** in the original notation, i.e. not represented in Voigt notation.

4.4.4 Force field

The potential energy surfaces of the structures were described by DREIDING force field [244], a classical force field. DREIDING force field is known to be able to model dynamics and structures of organic molecules correctly. The force field is a series of molecular mechanics potentials and parameters, including bonded and non-bonded van der Waals interactions.

$$\mathrm{E_{total}} = \mathrm{E_{bonded}} + \mathrm{E_{non\text{-}bonded}} \tag{4.16}$$

$$\mathrm{E_{bonded}} = \sum_{\mathrm{bonds}} \mathrm{E_{stretch}} + \sum_{\mathrm{angles}} \mathrm{E_{bend}} + \sum_{\mathrm{dihedrals}} \mathrm{E_{torsion}} + \sum_{\mathrm{impropers}} \mathrm{E_{out\text{-}of\text{-}plane}} \tag{4.17}$$

$$\mathrm{E_{non\text{-}bonded}} = \sum_{\mathrm{pairs}} \mathrm{E_{vdW}}. \tag{4.18}$$

The main inadequacy of generic force fields, such as DREIDING, is their relatively poor parameters for modelling the coordination environment of the metallic nodes of MOFs. Hence, we modified the force field parameters, bond stretching and angle bending, of Zinc tetrahedrals to correctly model the coordination environment of Zinc atoms. The parameters were extracted from the experimental vibrational frequencies of far-infrared spectrum of ZIFs and density functional theory calculations in literature [252, 253]. The modified parameters and their functionals are summarized in table 4.2. Bonding topology calculation and force field parameter assignment are discussed in detail in reference [80].

4.5 Supplementary materials

Lennard-Jones lattice model

Bulk modulus is the second derivative of the energy with respect to isotropic deformations. For a primitive cubic lattice solid with only nearest neighbour Lennard-Jones type of interactions, the second derivative of the energy can be expressed analytically as a function of the

$E_{stretch} = K_{IJ}(r - r_0)^2$	$K_{IJ}\,[kcal/mol]$	r_0
Zn - N	90.0	1.97
$E_{bend} = C_{IJK}[\cos(\theta) - \cos(\theta_0)]^2$	$C_{IJK}\,[kcal/mol]$	θ_0
N - Zn - N	19.67	109.47
C - N - Zn	23.3	120

Table 4.2 – The force field parameters used for Zinc tetrahedrals.

interatomic distance of atoms on the lattice. In such a simple model, there is a one by one correspondence between the interatomic distance and density (ρ),

$$\rho = \frac{m}{a^3}, \tag{4.19}$$

where a is lattice parameter and m is atomic mass. Lennard-Jones (LJ) potential is a simple potential which evaluates the non-bonded van der Waals interactions as a function of the interatomic distance between two atoms (r):

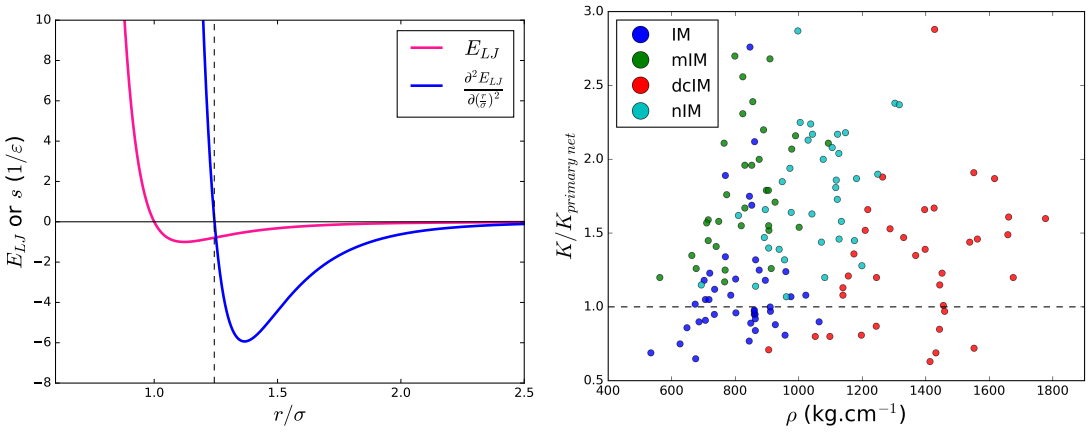
$$E_{LJ} = 4\varepsilon[(\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^{6}] \tag{4.20}$$

In this equation, σ and ε are the van der Waals radii and the potential well depth of the LJ potential, respectively. The second derivative of the Lennard-Jones potential with respect to normalized distance is

$$s = \frac{\partial^2 E_{LJ}}{\partial(\frac{r}{\sigma})^2} = 4\varepsilon[12 \times 13(\frac{\sigma}{r})^{14} - 6 \times 7(\frac{\sigma}{r})^{8}]. \tag{4.21}$$

Figure 4.7a shows there are two stiffening and softening regimes of non-bonded van der Waals interactions which are corresponding to the positive and negative parts of the second derivative of the potential energy, respectively. The second derivative of the potential changes sign at interatomic distance $\sim 1.2\sigma$.

Assuming an additive nature for the contributions from the non-bonded interactions to the primary network, one would expect high density materials gain more from the non-bonded interactions, as higher density corresponds to higher population in stiffening regime of the non-bonded interactions. To examine this hypothesis, we define enhancement ratio as the ratio of the bulk modulus of a structure with respect to the bulk modulus of its primary network ($K/K_{primary\ net}$). In figure 4.7b, density and the enhancement ratio of the structures are represented. Evidently, there is no correlation between density and the enhancement ratio. This demonstrates that the spatial distribution of atoms is the key for formation of a secondary



(a) (b)

Figure 4.7 – (a) Lennard-Jones potential and its second derivative (s) with respect to normalized pairwise distance. (b) Ratio of the bulk modulus of structures computed with non-bonded interactions to turned off non-bonded interactions as a function of density.

structure	K_{VHR}	G_{VHR}	E_{xx}	E_{yy}	E_{zz}
MOF-520	10.6	1.3	7.5	7.5	11.1
MOF-520-BPDC	15.3	2.5	20.2	20.2	17.1

Table 4.3 – Mechanical properties of MOF-520 and its retrofitted version, MOF-520-BPDC. All the properties are in GPa.

network.

Mechanical properties of MOF-520 and MOF-520-BPDC

Kapustin *et al.* proposed a molecular retrofication procedure by adding extra ligands to the weak points of MOF frameworks [222]. They successfully applied this procedure by adding extra ligands to retrofit MOF-520 in its xy plane by modifying the underlying network topology, i.e. the primary network of the structure [254]. The primary network of the structures are shown in figure 4.8. The analysis of underlying network topology of the two structures were performed by ToposPro version 5.3.0.2 [255]. To charactrize the mechanical properties of MOF-520 and its retrofitted version MOF-520-BPDC, we calculated the moduli of elasticity of the structures. The same force field [244] with adapting the geometric parameters of metals from the experimental x-ray data as explained in reference [80] was used to describe the PES of the structures. The computed properties agree with the experimentally observed higher mechanical stability of the retrofitted MOF (See table 4.3). More specifically, as one expects, the effect of extra ligands is more emanate in the Young's modulus of the structure along x and y axis. Unfortunately, due to lack of experimental data on moduli of elasticity, we cannot compare directly the computed values with experimental data.

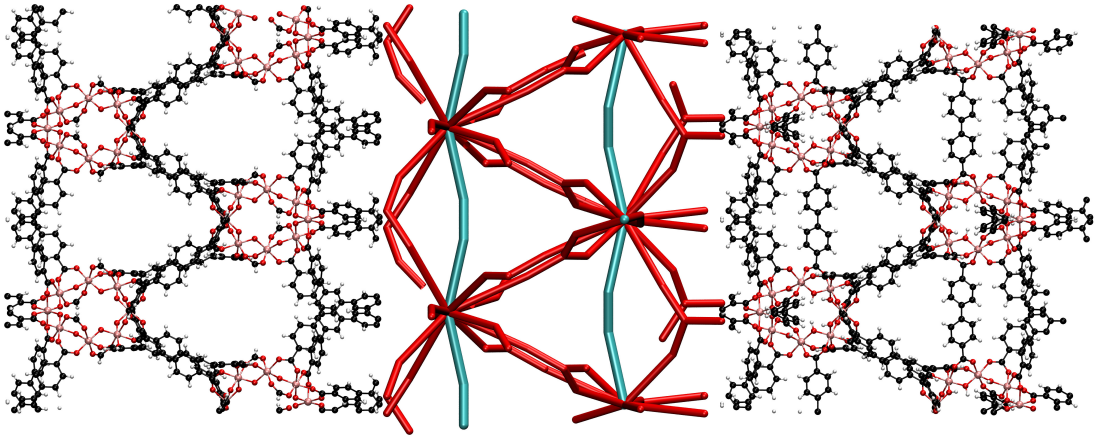


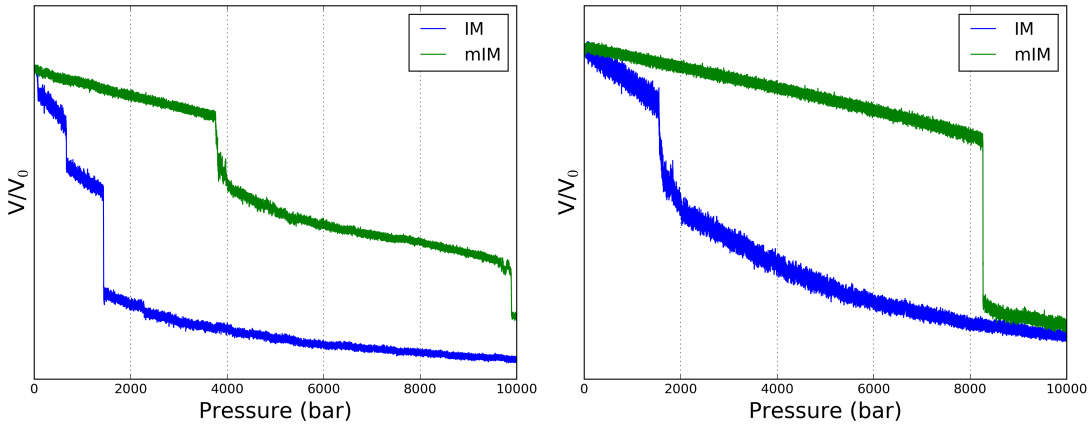
Figure 4.8 – MOF-520 (left) and its retrofitted version MOF-520-BPDC (right) are shown together with their primary networks (middle). By adding extra linkage (cyan tubes) to the primary network of MOF-520 (red tubes), it was adapted to high pressures. White, black, red and pink spheres represent H, C, O, and Al atoms, respectively.

structure	K_{VHR}	G_{VHR}	E_{xx}	E_{yy}	E_{zz}
IM-ZIF-3	2.0	0.5	1.3	1.1	4.6
mIM-ZIF-3	7.8	1.0	3.5	3.8	5.1
IM-ZIF-4	3.1	0.8	5.6	4.9	2.6
mIM-ZIF-4	7.6	1.5	7.6	7.2	8.7

Table 4.4 – Mechanical properties of ZIF-3 and ZIF-4 with two IM and mIM ligands. All the properties are in GPa

Amorphization of ZIF-3 and ZIF-4

The moduli of elasticity of ZIF-3 and ZIF-4 with two IM and mIM ligands are represented in table 4.4. Evidently, formation of secondary networks in the structures with substituted ligands, namely mIM-ZIF-3 and mIM-ZIF-4, has enhanced the mechanical stability of the structures. Knowing that amorphisation occurs due to material softening upon heating and/or pressure, we expect ZIFs with higher mechanical stability, i.e. higher moduli of elasticity, amorphise at more extreme conditions. To examine this hypothesis, we carried out molecular dynamic simulations at constant temperature and pressures (NPT ensemble [256]) using Nosé-Hoover thermostat and barostat as implemented in LAMMPS. Figure 4.9 shows the variation of volume with respect to applied pressure for the structures. Indeed, amorphisation delayed in the modified structures to much extreme pressures. Interestingly, the results in table 4.4 and figre 4.9 show that the materials with lower shear modulus amorphise at lower pressure. This observation confirms the explanation of a previous study on amorphisation of ZIFs, which showed amorphisation happens due to shear mode softening [240].



(a) (b)

Figure 4.9 – Evolution of unit cell volume of structures as a function of pressure for (a) ZIF-3 and (b) ZIF-4 extracted from molecular dynamic simulations. Pressure-induced amorphisation is delayed by incorporation of secondary networks in the mIM structures.

Conclusions and Future Research

In this thesis, we explored some of the emerging applications of computational and data–driven methods for the design and development of porous materials for gas separation and storage. We demonstrated how these methods can be beneficial at different stages of the material development, from database generation and *in silico* screening, to material synthesis and stability assessment. In this respect, this thesis shows some novel promising domains of applicability of the computational and data–driven methods for the development of porous materials.

The motivation of developing descriptors for the chemistry and pore geometry of MOFs was to pave the first steps towards exploiting the capability of machine learning and material informatics for the design and discovery of these materials. The success of these approaches relies on how well the material descriptors preserve structural similarity. In Chapter 1, we showed that the graph–based autocorrelation descriptors are simple, symmetry invariant, cheap, and yet chemically rich representations of the chemistry of MOFs. Furthermore, in Chapter 2, we presented a methodology based on persistent homology to quantify similarity in pore geometry and shape of materials. The remarkable accuracy we achieved in predicting gas adsorption and partial atomic charges of MOFs using machine learning trained using our descriptors is a proof of concept, showing that we can quantify their structural similarity and exploit machine learning to predict their properties in a fast, reliable and cheap way.

A natural future research direction is to use the here developed descriptors for machine learning predictions of other properties of MOFs. Machine learning can in particular be a practical attractive alternative for obtaining properties, for which computational screening is too demanding. For example, large–scale screenings of MOFs for their electronic structure properties are prohibited due to the exceedingly large computational cost. These properties are of great importance for applications, such as (photo-)catalysis, sensing, and electronics. Since some of these properties are near–sighted and mainly depend on the local chemical environment and atomic connectivity, e.g., the catalytically active site, we expect that here the graph–based descriptors will be useful. Noticeably, these graph–based descriptors were initially developed for the electronic properties of transition metal complexes and showed remarkable performance.

In Chapter 1, we introduced a methodology to quantify the diversity of MOF databases and

showed that each of the MOF databases is biased towards specific regions of the MOF chemical space. The primary aim of this study is to aid material discovery with a way to more efficient and exploratory sampling the possible chemical space. Our results strongly demonstrated the importance of this exploratory approach as we showed that the bias in the databases have obstructed material discovery and led to not generalizable structure–property relationships, and not transferable machine learning models. A way to reduce these biases is to generate structures in the less explored or unexplored regions of the chemical space. For example, this can be done by mining inorganic building blocks from the experimental databases to assemble them with many organic linkers that we can retrieve from organic molecule databases.

In addition, it is important to realize that the notion of diversity depends on the context of the problem at hand. For example, if we are interested in gas adsorption or separation any diversity in non-porous MOFs will be irrelevant, while for optical properties the shape and size of pores will be of minor importance. Our diversity analysis was performed in the broad and fundamental context of structural building blocks and pore geometry, to recognise the problem and find the limitations in the available databases independent of the application. However, in future, we aim to perform diversity analysis for a specific given applications. Essential to reach this aim is to be able to identify the key structural characteristics which influence and determine the material properties of interest. Since our chemical descriptors are inspired by the concept of reticular chemistry, and they closely resemble the chemical intuition of MOF chemists, in which a MOF is a combination of metal nodes, organic linkers and functional groups, they allow us to identify the key characteristics and directly use them in the structure generation step. This approach ensures optimal exploration of the design space for a given application.

We all know the importance of chemical synthesis when we discover a new promising material in computer. However, chemical synthesis is a fuzzy procedure that only “expert” synthetic chemists have an “intuition” for how to do it. The aim of our study in Chapter 3 was to develop a data–driven framework to capture this intuition using data from a set of failed and successful experiments. Our study on a MOF showed that such approach in principle can be feasible. Similar to the way chemist develop their chemical intuition over the course of many failed and successful experiments, our framework’s intuition become more versatile when it is provided more data. Hence, such framework becomes a powerful tool only if we can collect data from many experimental groups. However, most of the experimental data, specifically almost all the failed experiments, remained unreported. Therefore, an important direction for future work is to develop infrastructure to collect all experimental data. A possible tractable way is to use electronic lab notebooks (ELNs) coupled with data repositories to collect these data.

In our study on HKUST–1, we used genetic algorithm to optimise objectives based on the crystal quality metrics, namely crystallinity and surface area. However, the aim of synthesis is often to obtain the crystals with optimum desired functionality. For example, HKUST–1 has the highest deliverable capacity for the methane storage application. As the objective function of the genetic algorithm is flexible to any desired property, one can simply use our

tools which are available online to optimise the methane deliverable capacity of HKUST–1. Moreover, it is fundamentally important to find out whether the crystal quality metrics correspond to functionality. This insight has also practical merits as adsorption properties characterisations (e.g., methane adsorption) are much more time consuming than crystal quality characterisations (e.g., X–ray diffraction).

With the discovery of many promising MOFs, other properties of materials which have fundamental importance for the practical use of the materials, e,g. mechanical stability, become relevant. Our study on the mechanical stability of MOFs in Chapter 4 aimed to establish a basic understanding of how the mechanical stability is related to the underlaying molecular structure. We showed that the underlying network topology is the primary factor for the mechanical stability. Additionally, the functional groups of the linker can form a secondary network which can enhance and/or tune the mechanical stability of MOFs. While both underlying network and secondary network can be rationally designed, designing the secondary network is more practical since linker functionalisation is feasible in many ways, e.g., post synthesis modification.

The research on the mechanical stability of MOFs is only at the beginning and many more studies to come. For example, our study was focused on a relatively small set of MOFs. Since the methodology and tools developed for this study can be applied to other MOFs, a possible research project is to extend the set of materials to include more chemical diversity to establish a more general structure–property relationship. In particular, for different isoreticular MOF series, e.g. IRMOF-1, IRMOF-74, etc., it is of great importance to develop specific strategies to enhance their mechanical stability.

Curriculum vitae: Seyed Mohamad Moosavi

Personal information

Address: EPFL Valais Wallis, EPFL SB ISIC LSMO
Rue de l'Industrie 17 Case postale 440
1950 Sion, Switzerland
Mobile: +41787140578
Email: s.moh.moosavi@gmail.com

Born: 27 June 1991
Publication list on Google Scholar

Education

Ph.D. in Chemistry and Chemical Engineering **Sep. 2015 – Feb. 2020**
Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland
Thesis title: *"Advancing computational and data-driven methods for the design and discovery of nanoporous materials"*
Advisor: Prof. Berend Smit

SNSF Visiting Scholar, Chemical Engineering **Jan. 2019 – July 2019**
Massachusetts Institute of Technology (MIT), United States
Project title: *"Machine learning for material properties of metal–organic frameworks"*
Advisor: Prof. Heather J. Kulik

M.Sc. in Mechanical Engineering **2013 – 2015**
Minor in Management of Technology and Entrepreneurship
Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland

B.Sc. in Mechanical Engineering **2009 – 2013**
Sharif University of Technology (SUT), Iran

Honors, awards, and fellowships

- Doctoral mobility fellowship, Swiss National Science Foundation (SNSF) (∼30K$), Switzerland, 2018
- Selected as one of the top 5 Ph.D. thesis at EPFL in class of 2015 to represent EPFL for Schmidt Science Fellows, Switzerland, 2019
- Chemistry travel award of Swiss Chemical Society (1K$), Switzerland, 2018
- Silver medal in the national student olympiad of mechanical engineering, Iran, 2013
- 3nd place in the B.Sc. program, Sharif University of Technology, Iran, 2013
- Ranked 59^{th} among over 300'000 in the nationwide entrance exam for undergraduate admissions in the public universities, Iran, 2009

Teaching Activities

University of Amsterdam (UvA) – Amsterdam Centre for Multi-scale Modelling (ACMM)

- Lecturer, "Basics of machine learning in chemistry and materials science", MolSim, Jan 2020
- Assistant, understanding molecular simulation, CECAM winter school, Jan 2018

Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland

- Assistant, understanding advanced molecular simulation, M.Sc. courses, Spring 2018
- Assistant, advanced diffusional separation processes, M.Sc. course, Fall 2016
- Assistant, analysis 1 and 2, B.Sc. courses, Spring and Fall 2014

Sharif University of Technology (SUT)

- Assistant, mechanical design, B.Sc. course, Fall 2012

Academic Service and memberships

- Scientific reviewing paper for Joule, Journal of Material Chemistry A, ChemSusChem and Journal of Chemical Theory and Computation
- Editorial board member of Sharif Mechanics Magazine, 2010-2011
- Member of the organisation committee of the 5th Sharif Mechanics Industrial Festival, 2010
- Member of the executive committee of the UNESCO conference on technologies development hosted by CODEV, 2015, EPFL

Work Experience – internships

Engineer internship **Spring 2015**
FUELMAT, Yverdon-les-Bains Switzerland
Modeling of Electrochemical Impedance Spectroscopy (EIS) of Solid Oxide Fuel Cell (SOFC) on stack scale using gPROMS

Engineer internship **Summer 2012**
Sazeh Consultants, Stationary Equipment Department, Iran
Pressure vessel design for South Pars / North Dome gas-condensate field project

Contribution to Conferences and Seminars

Oral presentations

- Oral presentation in ML4MS workshop CECAM, Toulouse, October 2019
- Invited talk in Nanoporous Material Genome Center and Chemical Theory Center in university of Minnesota, Minneapolis, US, May 2019
- Invited talk in the Platform for Advanced Scientific Computing (PASC) Conference, Basel, Switzerland, July 2018
- Department of BioChemistry seminar, University of Oxford, November 2018

Poster presentations

- MOF-2018 conference, University of Auckland, December 2018
- Computational Chemistry meets Artificial Intelligence (CC2AI), EPFL, June 2018
- NCCR-MARVEL Junior Retreat, EPFL, July 2018
- NCCR-MARVEL Site Visit, EPFL, April 2018
- Understanding Molecular Simulation (MOLSIM), Amsterdam, January 2016
- Hands-on workshop density functional theory and beyond, Isfahan, May 2016

Languages

English	fluent
French	A2 level of CEFRL
German	A2 level of CEFRL
Arabic	basic
Farsi	mother tongue

List of publications

- Peter G Boyd, Arunraj Chidambaram, Enrique García-Díez, Christopher P Ireland, Thomas D Daff, Richard Bounds, Andrzej Gladysiak, Pascal Schouwink, Seyed Mohamad Moosavi, M Mercedes Maroto-Valer, Jeffrey A Reimer, Jorge AR Navarro, Tom K Woo, Susana Garcia, Kyriakos C Stylianou, Berend Smit, "Data-driven design of metal–organic frameworks for wet flue gas CO 2 capture." *Nature* 576.7786 (2019): 253-256. **10.1038/s41586-019-1798-7**

- Li Peng, Shuliang Yang, Sudi Jawahery, Seyed Mohamad Moosavi, Aron J Huckaba, Mehrdad Asgari, Emad Oveisi, Mohammad Khaja Nazeeruddin, Berend Smit, Wendy L Queen, "Preserving Porosity of Mesoporous Metal–Organic Frameworks through the Introduction of Polymer Guests." *Journal of the American Chemical Society* 141.31 (2019): 12397-12405. **10.1021/jacs.9b05967**

- Andrzej Gladysiak, Seyed Mohamad Moosavi, Lev Sarkisov, Berend Smit, Kyriakos C Stylianou, "Guest-dependent negative thermal expansion in a lanthanide-based metal-organic framework." *CrystEngComm* 21.35 (2019): 5292-5298. **10.1039/C9CE00941H**

- Sudi Jawahery, Nakul Rampal, Seyed Mohamad Moosavi, Matthew Witman, Berend Smit, "Ab Initio Flexible Force Field for Metal-Organic Frameworks Using Dummy Model Coordination Bonds." *Journal of chemical theory and computation* 15.6 (2019): 3666-3677. **10.1021/acs.jctc.9b00135**

- Seyed Mohamad Moosavi, Arunraj Chidambaram, Leopold Talirz, Maciej Haranczyk, Kyriakos C Stylianou, Berend Smit, "Capturing chemical intuition in synthesis of metal-organic frameworks." *Nature communications* 10.1 (2019): 539. **10.1038/s41467-019-08483-9**

- Andrzej Gladysiak, Kathryn S Deeg, Iurii Dovgaliuk, Arunraj Chidambaram, Kaili Ordiz, Peter G Boyd, Seyed Mohamad Moosavi, Daniele Ongari, Jorge AR Navarro, Berend Smit, Kyriakos C Stylianou, "Biporous Metal Organic Framework with Tunable CO2/CH4 Separation Performance Facilitated by Intrinsic Flexibility." *ACS applied materials & interfaces* 10.42 (2018): 36144-36156. **10.1021/acsami.8b13362**

- Efrem Braun, Yongjin Lee, Seyed Mohamad Moosavi, Senja Barthel, Rocio Mercado, Igor A Baburin, Davide M Proserpio, Berend Smit, "Generating carbon schwarzites

via zeolite-templating." *Proceedings of the National Academy of Sciences* 115.35 (2018): E8116-E8124. **10.1073/pnas.1805062115**

- Efrem Braun, Seyed Mohamad Moosavi, Berend Smit, "Anomalous effects of velocity rescaling algorithms: the flying ice cube effect revisited." *Journal of chemical theory and computation* 14.10 (2018): 5262-5272. **10.1021/acs.jctc.8b00446**

- Yongjin Lee, Senja D Barthel, Pawel Dlotko, Seyed Mohamad Moosavi, Kathryn Hess, Berend Smit, "High-throughput screening approach for nanoporous materials genome using topological data analysis: application to zeolites." *Journal of chemical theory and computation* 14.8 (2018): 4427-4437. **10.1021/acs.jctc.8b00253**

- Seyed Mohamad Moosavi, Peter G Boyd, Lev Sarkisov, Berend Smit, "Improving the mechanical stability of metal–organic frameworks using chemical caryatids." *ACS Central Science* 4.7 (2018): 832-839. **10.1021/acscentsci.8b00157**

- Daniel T Sun, Li Peng, Washington S Reeder, Seyed Mohamad Moosavi, Davide Tiana, David K Britt, Emad Oveisi, Wendy L Queen, "Rapid, selective heavy metal removal from water by a metal–organic framework/polydopamine composite." *ACS central science* 4.3 (2018): 349-356. **10.1021/acscentsci.7b00605**

- Seyed Mohamad Moosavi, Mathias Niffeler, Jeff Gostick, Sophia Haussener, "Transport characteristics of saturated gas diffusion layers treated with hydrophobic coatings." *Chemical Engineering Science* 176 (2018): 503-514. **10.1016/j.ces.2017.10.035**

- Velencia J Witherspoon, Lucy M Yu, Sudi Jawahery, Efrem Braun, Seyed Mohamad Moosavi, Sondre K Schnell, Berend Smit, Jeffrey A Reimer, "Translational and Rotational Motion of C8 Aromatics Adsorbed in Isotropic Porous Media (MOF-5): NMR Studies and MD Simulations." *The Journal of Physical Chemistry C* 121.28 (2017): 15456-15462. **10.1021/acs.jpcc.7b03181**

- Yongjin Lee, Senja D Barthel, Pawel Dlotko, Seyed Mohamad Moosavi, Kathryn Hess, Berend Smit, "Quantifying similarity of pore-geometry in nanoporous materials." *Nature communications* 8 (2017): 15396.**10.1038/ncomms15396**

- Peter G Boyd, Seyed Mohamad Moosavi, Matthew Witman, Berend Smit, "Force-field prediction of materials properties in metal-organic frameworks." *The journal of physical chemistry letters* 8.2 (2017): 357-363. **10.1021/acs.jpclett.6b02532**

- Seyed Mohamad Moosavi, Arman Sadeghi, Mohammad Said Saidi, "Electrophoretic velocity of spherical particles in Quemada fluids." *Colloids and Surfaces A: Physicochemical and Engineering Aspects* 436 (2013): 225-230. **10.1016/j.colsurfa.2013.06.028**

Bibliography

[1] Frank Jensen. Introduction to computational chemistry. John wiley & sons, 2017.

[2] Christopher J Cramer. Essentials of computational chemistry: theories and models. John Wiley & Sons, 2013.

[3] Eugene Isaacson and Herbert Bishop Keller. Analysis of numerical methods. Courier Corporation, 2012.

[4] Kenneth L Judd and Kenneth L Judd. Numerical methods in economics. MIT press, 1998.

[5] James E Gentle. Computational statistics, volume 308. Springer, 2009.

[6] Steven C Chapra and Raymond P Canale. Numerical methods for engineers, volume 2. Mcgraw-hill New York, 1998.

[7] John David Anderson and J Wendt. Computational fluid dynamics, volume 206. Springer, 1995.

[8] Leonard Meirovitch. Computational methods in structural dynamics, volume 5. Springer Science & Business Media, 1980.

[9] Gordon E Moore et al. Cramming more components onto integrated circuits, 1965.

[10] David Chandler. Introduction to modern statistical mechanics. Introduction to Modern Statistical Mechanics, by David Chandler, pp. 288. Foreword by David Chandler. Oxford University Press, Sep 1987. ISBN-10: 0195042778. ISBN-13: 9780195042771, page 288, 1987.

[11] Radu Balescu. Equilibrium and nonequilibrium statistical mechanics. NASA STI/Recon Technical Report A, 76, 1975.

[12] Terrell L Hill. An introduction to statistical thermodynamics. Courier Corporation, 1986.

[13] Jun John Sakurai and Eugene D Commins. Modern quantum mechanics. AAPT, 1995.

[14] José E Moyal. Quantum mechanics as a statistical theory. In Mathematical Proceedings of the Cambridge Philosophical Society, volume 45, pages 99–124. Cambridge University Press, 1949.

[15] Daan Frenkel and Berend Smit. Understanding molecular simulation: from algorithms to applications, volume 1. Elsevier, 2001.

[16] Mark Tuckerman. Statistical mechanics: theory and molecular simulation. Oxford university press, 2010.

[17] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. Phys. Rev., 140:A1133–A1138, Nov 1965.

[18] John P Perdew and Alex Zunger. Self-interaction correction to density-functional approximations for many-electron systems. Physical Review B, 23(10):5048, 1981.

[19] Wolfram Koch and Max C Holthausen. A chemist's guide to density functional theory. John Wiley & Sons, 2015.

[20] Robert G Parr. Density functional theory of atoms and molecules. In Horizons of Quantum Chemistry, pages 5–15. Springer, 1980.

[21] David Sholl and Janice A Steckel. Density functional theory: a practical introduction. John Wiley & Sons, 2011.

[22] Theodora Katsila, Georgios A Spyroulias, George P Patrinos, and Minos-Timotheos Matsoukas. Computational approaches in target identification and drug discovery. Computational and structural biotechnology journal, 14:177–184, 2016.

[23] Weilin Zhang, Jianfeng Pei, and Luhua Lai. Computational multitarget drug design. Journal of chemical information and modeling, 57(3):403–412, 2017.

[24] Jens Kehlet Nørskov, Thomas Bligaard, Jan Rossmeisl, and Claus Hviid Christensen. Towards the computational design of solid catalysts. Nature chemistry, 1(1):37, 2009.

[25] Justin B Siegel, Alexandre Zanghellini, Helena M Lovick, Gert Kiss, Abigail R Lambert, Jennifer L St Clair, Jasmine L Gallaher, Donald Hilvert, Michael H Gelb, Barry L Stoddard, et al. Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. Science, 329(5989):309–313, 2010.

[26] Berend Smit and Theo LM Maesen. Molecular simulations of zeolites: adsorption, diffusion, and shape selectivity. Chemical reviews, 108(10):4125–4184, 2008.

[27] Jörg Neugebauer and Matthias Scheffler. Adsorbate-substrate and adsorbate-adsorbate interactions of na and k adlayers on al (111). Physical Review B, 46(24):16067, 1992.

[28] Alain H Fuchs and Anthony K Cheetham. Adsorption of guest molecules in zeolitic materials: computational aspects, 2001.

[29] Vincenzo Barone. Computational strategies for spectroscopy: from small molecules to nano systems. John Wiley & Sons, 2011.

[30] Daniel P Tabor, Loïc M Roch, Semion K Saikin, Christoph Kreisbeck, Dennis Sheberla, Joseph H Montoya, Shyam Dwaraknath, Muratahan Aykol, Carlos Ortiz, Hermann Tribukait, et al. Accelerating the discovery of materials for clean energy in the era of smart automation. Nature Reviews Materials, 3(5):5, 2018.

[31] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. Science, 361(6400):360–365, 2018.

[32] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical review letters, 120(14):145301, 2018.

[33] Connor W Coley, Regina Barzilay, Tommi S Jaakkola, William H Green, and Klavs F Jensen. Prediction of organic reaction outcomes using machine learning. ACS central science, 3(5):434–443, 2017.

[34] Connor W Coley, William H Green, and Klavs F Jensen. Machine learning in computer-aided synthesis planning. Accounts of chemical research, 51(5):1281–1289, 2018.

[35] Jarosław M Granda, Liva Donina, Vincenza Dragone, De-Liang Long, and Leroy Cronin. Controlling an organic synthesis robot with machine learning to search for new reactivity. Nature, 559(7714):377, 2018.

[36] Klavs F Jensen, Connor W Coley, and Natalie S Eyke. Autonomous discovery in the chemical sciences part i: Progress. Angewandte Chemie International Edition, 2019.

[37] Colin R Groom, Ian J Bruno, Matthew P Lightfoot, and Suzanna C Ward. The cambridge structural database. Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials, 72(2):171–179, 2016.

[38] Materials cloud. http://https://www.materialscloud.org/.

[39] Claudia Draxl and Matthias Scheffler. Nomad: The fair concept for big data-driven materials science. MRS Bulletin, 43(9):676–682, 2018.

[40] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. Physical review letters, 108(5):058301, 2012.

[41] Sandip De, Albert P Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. Physical Chemistry Chemical Physics, 18(20):13754–13769, 2016.

[42] Albert P Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R Kermode, Gábor Csányi, and Michele Ceriotti. Machine learning unifies the modeling of materials and molecules. Science advances, 3(12):e1701816, 2017.

[43] Jon Paul Janet and Heather J Kulik. Predicting electronic structure properties of transition metal complexes with neural networks. Chemical science, 8(7):5137–5152, 2017.

[44] Connor W Coley, Natalie S Eyke, and Klavs F Jensen. Autonomous discovery in the chemical sciences part ii: Outlook. Angewandte Chemie International Edition, 2019.

[45] Florian Häse, Loïc M Roch, and Alán Aspuru-Guzik. Next-generation experimentation with self-driving laboratories. Trends in Chemistry, 2019.

[46] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. ACS central science, 4(2):268–276, 2018.

[47] Gabriel Lima Guimaraes, Benjamin Sanchez-Lengeling, Carlos Outeiral, Pedro Luis Cunha Farias, and Alán Aspuru-Guzik. Objective-reinforced generative adversarial networks (organ) for sequence generation models. arXiv preprint arXiv:1705.10843, 2017.

[48] Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs. arXiv preprint arXiv:1805.11973, 2018.

[49] Mai Bui, Claire S Adjiman, André Bardow, Edward J Anthony, Andy Boston, Solomon Brown, Paul S Fennell, Sabine Fuss, Amparo Galindo, Leigh A Hackett, et al. Carbon capture and storage (ccs): the way forward. Energy & Environmental Science, 11(5):1062–1176, 2018.

[50] R Bruce Eldridge. Olefin/paraffin separation technology: a review. Industrial & engineering chemistry research, 32(10):2208–2212, 1993.

[51] David S Sholl and Ryan P Lively. Seven chemical separations to change the world. Nature News, 532(7600):435, 2016.

[52] Smit Berend et al. Introduction to carbon capture and sequestration, volume 1. World Scientific, 2014.

[53] Trevor A Makal, Jian-Rong Li, Weigang Lu, and Hong-Cai Zhou. Methane storage in advanced porous materials. Chemical Society Reviews, 41(23):7761–7779, 2012.

[54] Gary T Rochelle. Amine scrubbing for co2 capture. Science, 325(5948):1652–1654, 2009.

[55] Yabing He, Wei Zhou, Guodong Qian, and Banglin Chen. Methane storage in metal–organic frameworks. Chemical Society Reviews, 43(16):5657–5678, 2014.

[56] Erin Baker, Haewon Chon, and Jeffrey Keisler. Carbon capture and storage: combining economic analysis with expert elicitations to inform climate policy. Climatic Change, 96(3):379–408, 2009.

[57] Michael J Economides et al. The economics of gas to liquids compared to liquefied natural gas. World Energy, 8(1):136–140, 2005.

[58] Omar M Yaghi, Markus J Kalmutzki, and Christian S Diercks. Introduction to Reticular Chemistry: Metal-Organic Frameworks and Covalent Organic Frameworks. John Wiley & Sons, 2019.

[59] Omar M Yaghi, Michael O'Keeffe, Nathan W Ockwig, Hee K Chae, Mohamed Eddaoudi, and Jaheon Kim. Reticular synthesis and the design of new materials. Nature, 423(6941):705, 2003.

[60] Adrien P Cote, Annabelle I Benin, Nathan W Ockwig, Michael O'Keeffe, Adam J Matzger, and Omar M Yaghi. Porous, crystalline, covalent organic frameworks. science, 310(5751):1166–1170, 2005.

[61] Kyo Sung Park, Zheng Ni, Adrien P Côté, Jae Yong Choi, Rudan Huang, Fernando J Uribe-Romo, Hee K Chae, Michael O'Keeffe, and Omar M Yaghi. Exceptional chemical and thermal stability of zeolitic imidazolate frameworks. Proceedings of the National Academy of Sciences, 103(27):10186–10191, 2006.

[62] J.T.A. Jones, T. Hasell, X. Wu, J. Bacsa, K.E. Jelfs, M. Schmidtmann, S.Y. Chong, D.J. Adams, A. Trewin, F. Schiffman, F. Cora, B. Slater, A. Steiner, G.M. Day, and A.I. Cooper. Modular and predictable assembly of porous organic molecular crystals. Nature, 474(7351):367–371, 2011.

[63] Ines M Hönicke, Irena Senkovska, Volodymyr Bon, Igor A Baburin, Nadine Bönisch, Silvia Raschke, Jack D Evans, and Stefan Kaskel. Balancing mechanical stability and ultrahigh porosity in crystalline framework materials. Angewandte Chemie International Edition, 57(42):13780–13783, 2018.

[64] Christopher E Wilmer, Michael Leaf, Chang Yeon Lee, Omar K Farha, Brad G Hauser, Joseph T Hupp, and Randall Q Snurr. Large-scale screening of hypothetical metal–organic frameworks. Nature Chemistry, 4(2):83, 2012.

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

[66] Peter G Boyd, Yongjin Lee, and Berend Smit. Computational development of the nanoporous materials genome. Nature Reviews Materials, 2(8):17037, 2017.

[67] Peter G Boyd and Tom K. Woo. A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. CrystEngComm, 18(21):3777–3792, 2016.

[68] Diego A Gomez-Gualdron, Oleksii V Gutov, Vaiva Krungleviciute, Bhaskarjyoti Borah, Joseph E Mondloch, Joseph T Hupp, Taner Yildirim, Omar K Farha, and Randall Q Snurr. Computational design of metal–organic frameworks based on stable zirconium building units for storage and delivery of methane. Chemistry of Materials, 26(19):5632–5639, 2014.

[69] Peter G Boyd, Arunraj Chidambaram, Enrique García-Díez, Christopher P Ireland, Thomas D Daff, Richard Bounds, Andrzej Gładysiak, Pascal Schouwink, Seyed Mohamad Moosavi, M Mercedes Maroto-Valer, et al. Data-driven design of metal–organic frameworks for wet flue gas co 2 capture. Nature, 576(7786):253–256, 2019.

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

[71] Diego A Gómez-Gualdrón, Yamil J Colón, Xu Zhang, Timothy C Wang, Yu-Sheng Chen, Joseph T Hupp, Taner Yildirim, Omar K Farha, Jian Zhang, and Randall Q Snurr. Evaluating topologically diverse metal–organic frameworks for cryo-adsorbed hydrogen storage. Energy & Environmental Science, 9(10):3279–3289, 2016.

[72] Andrew S Rosen, Justin M Notestein, and Randall Q Snurr. Structure–activity relationships that identify metal–organic framework catalysts for methane activation. ACS Catalysis, 9(4):3576–3587, 2019.

[73] Ryther Anderson, Jacob Rodgers, Edwin Argueta, Achay Biong, and Diego A Gomez-Gualdron. Role of pore chemistry and topology in the co2 capture capabilities of mofs: from molecular simulation to machine learning. Chemistry of Materials, 30(18):6325–6337, 2018.

[74] Yongjin Lee, Senja D Barthel, Paweł Dłotko, S Mohamad Moosavi, Kathryn Hess, and Berend Smit. Quantifying similarity of pore-geometry in nanoporous materials. Nature communications, 8:15396, 2017.

[75] Yongjin Lee, Senja D Barthel, Paweł Dłotko, Seyed Mohamad Moosavi, Kathryn Hess, and Berend Smit. High-throughput screening approach for nanoporous materials genome using topological data analysis: application to zeolites. Journal of chemical theory and computation, 14(8):4427–4437, 2018.

[76] Peyman Z Moghadam, Sven MJ Rogge, Aurelia Li, Chun-Man Chow, Jelle Wieme, Noushin Moharrami, Marta Aragones-Anglada, Gareth Conduit, Diego A Gomez-Gualdron, Veronique Van Speybroeck, et al. Structure-mechanical stability relations of metal-organic frameworks via machine learning. Matter, 1(1):219–234, 2019.

[77] Jin Chong Tan and Anthony K Cheetham. Mechanical properties of hybrid inorganic–organic framework materials: establishing fundamental structure–property relationships. Chemical Society Reviews, 40(2):1059–1080, 2011.

[78] Jin Chong Tan, Thomas D Bennett, and Anthony K Cheetham. Chemical structure, network topology, and porosity effects on the mechanical properties of zeolitic imidazolate frameworks. Proceedings of the National Academy of Sciences, 107(22):9938–9943, 2010.

[79] Seyed Mohamad Moosavi, Peter G Boyd, Lev Sarkisov, and Berend Smit. Improving the mechanical stability of metal–organic frameworks using chemical caryatids. ACS Central Science, 4(7):832–839, 2018.

[80] Peter G Boyd, Seyed Mohamad Moosavi, Matthew Witman, and Berend Smit. Force-field prediction of materials properties in metal-organic frameworks. The journal of physical chemistry letters, 8(2):357–363, 2017.

[81] Andrzej Gładysiak, Seyed Mohamad Moosavi, Lev Sarkisov, Berend Smit, and Kyriakos C Stylianou. Guest-dependent negative thermal expansion in a lanthanide-based metal–organic framework. CrystEngComm, 21(35):5292–5298, 2019.

[82] Andrzej Gładysiak, Kathryn S Deeg, Iurii Dovgaliuk, Arunraj Chidambaram, Kaili Ordiz, Peter G Boyd, Seyed Mohamad Moosavi, Daniele Ongari, Jorge AR Navarro, Berend Smit, et al. Biporous metal–organic framework with tunable co2/ch4 separation performance facilitated by intrinsic flexibility. ACS applied materials & interfaces, 10(42):36144–36156, 2018.

[83] Li Peng, Shuliang Yang, Sudi Jawahery, Seyed Mohamad Moosavi, Aron J Huckaba, Mehrdad Asgari, Emad Oveisi, Mohammad Khaja Nazeeruddin, Berend Smit, and Wendy L Queen. Preserving porosity of mesoporous metal–organic frameworks through the introduction of polymer guests. Journal of the American Chemical Society, 141(31):12397–12405, 2019.

[84] Sudi Jawahery, Nakul Rampal, Seyed Mohamad Moosavi, Matthew Witman, and Berend Smit. Ab initio flexible force field for metal-organic frameworks using dummy model coordination bonds. Journal of chemical theory and computation, 2019.

[85] Mohamed Eddaoudi, Jaheon Kim, Nathaniel Rosi, David Vodak, Joseph Wachter, Michael O'Keeffe, and Omar M Yaghi. Systematic design of pore size and functionality in isoreticular mofs and their application in methane storage. Science, 295(5554):469–472, 2002.

[86] Yongchul G Chung, Emmanuel Haldoupis, Benjamin J Bucior, Maciej Haranczyk, Seulchan Lee, Hongda Zhang, Konstantinos D Vogiatzis, Marija Milisavljevic, Sanliang Ling, Jeffrey S Camp, Ben Slater, J. Ilja Siepmann, David S. Sholl, and Randall Q. Snurr. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: Core mof 2019. Journal of Chemical & Engineering Data, 2019.

[87] Peter G Boyd and Tom K Woo. A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. CrystEngComm, 18(21):3777–3792, 2016.

[88] Xiwen Jia, Allyson Lynch, Yuheng Huang, Matthew Danielson, Immaculate Lang'at, Alexander Milder, Aaron E Ruby, Hao Wang, Sorelle A Friedler, Alexander J Norquist, and Joshua Schrier. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. Nature, 573(7773):251–255, 2019.

[89] Anang A Shelat and R Kiplin Guy. Scaffold composition and biological relevance of screening libraries. Nature chemical biology, 3(8):442–446, 2007.

[90] Michael Fernandez, Nicholas R Trefiak, and Tom K Woo. Atomic property weighted radial distribution functions descriptors of metal–organic frameworks for the prediction of gas uptake capacity. The Journal of Physical Chemistry C, 117(27):14095–14105, 2013.

[91] Yuping He, Ekin D Cubuk, Mark D Allendorf, and Evan J Reed. Metallic metal–organic frameworks predicted by the combination of machine learning methods and ab initio calculations. The journal of physical chemistry letters, 9(16):4562–4569, 2018.

[92] Benjamin J Bucior, Andrew S Rosen, Maciej Haranczyk, Zhenpeng Yao, Michael E Ziebel, Omar K Farha, Joseph T Hupp, J Ilja Siepmann, Alán Aspuru-Guzik, and Randall Q Snurr. Identification schemes for metal–organic frameworks to enable rapid search and cheminformatics analysis. Crystal Growth & Design, 19(11):6682–6697, 2019.

[93] Thomas F Willems, Chris H Rycroft, Michaeel Kazi, Juan C Meza, and Maciej Haranczyk. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. Microporous and Mesoporous Materials, 149(1):134–141, 2012.

[94] Jon Paul Janet and Heather J Kulik. Resolving transition metal chemical space: Feature selection for machine learning and structure–property relationships. The Journal of Physical Chemistry A, 121(46):8939–8954, 2017.

[95] Aditya Nandy, Jiazhou Zhu, Jon Paul Janet, Chenru Duan, Rachel B Getman, and Heather J Kulik. Machine learning accelerates the discovery of design rules and exceptions in stable metal-oxo intermediate formation. ACS Catalysis, 2019.

[96] Jon Paul Janet, Fang Liu, Aditya Nandy, Chenru Duan, Tzuhsiung Yang, Sean Lin, and Heather J. Kulik. Designing in the face of uncertainty: Exploiting electronic structure and machine learning models for discovery in inorganic chemistry. Inorganic Chemistry, 58:10592–10606, 2019.

[97] Aditya Nandy, Chenru Duan, Jon Paul Janet, Stefan Gugler, and Heather J Kulik. Strategies and software for machine learning accelerated discovery in transition metal chemistry. Industrial & Engineering Chemistry Research, 57(42):13973–13986, 2018.

[98] Yongchul G Chung, Jeffrey Camp, Maciej Haranczyk, Benjamin J Sikora, Wojciech Bury, Vaiva Krungleviciute, Taner Yildirim, Omar K Farha, David S Sholl, and Randall Q Snurr. Computation-ready, experimental metal–organic frameworks: a tool to enable high-throughput screening of nanoporous crystals. Chemistry of Materials, 26(21):6185–6192, 2014.

[99] Dalar Nazarian, Jeffrey S Camp, and David S Sholl. A comprehensive set of high-quality point charges for simulations of metal–organic frameworks. Chemistry of Materials, 28(3):785–793, 2016.

[100] Benjamin J Sikora, Randy Winnegar, Davide M Proserpio, and Randall Q Snurr. Textural properties of a large collection of computationally constructed mofs and zeolites. Microporous and mesoporous materials, 186:207–213, 2014.

[101] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605, 2008.

[102] Andrew Stirling. Diversity and ignorance in electricity supply investment: Addressing the solution rather than the problem. Energy policy, 22(3):195–216, 1994.

[103] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2019 update: improved access to chemical data. Nucleic acids research, 47(D1):D1102–D1109, 2019.

[104] Manas K Bhunia, James T Hughes, James C Fettinger, and Alexandra Navrotsky. Thermochemistry of paddle wheel mofs: Cu-hkust-1 and zn-hkust-1. Langmuir, 29(25):8140–8145, 2013.

[105] Cory M Simon, Jihan Kim, Diego A Gomez-Gualdron, Jeffrey S Camp, Yongchul G Chung, Richard L Martin, Rocio Mercado, Michael W Deem, Dan Gunter, Maciej Haranczyk, et al. The materials genome in action: identifying the performance limits for methane storage. Energy & Environmental Science, 8(4):1190–1199, 2015.

[106] Jarad A. Mason, Julia Oktawiec, Mercedes K. Taylor, Matthew R. Hudson, Julien Rodriguez, Jonathan E. Bachman, Miguel I. Gonzalez, Antonio Cervellino, Antonietta Guagliardi, Craig M. Brown, Philip L. Llewellyn, Norberto Masciocchi, and Jeffrey R. Long. Methane storage in flexible metal-organic frameworks with intrinsic thermal management. Nature, 527(7578):357–361, 2015.

[107] Ronald W Kennard and Larry A Stone. Computer aided design of experiments. Technometrics, 11(1):137–148, 1969.

[108] Omar K Farha, A Özgür Yazaydın, Ibrahim Eryazici, Christos D Malliakas, Brad G Hauser, Mercouri G Kanatzidis, SonBinh T Nguyen, Randall Q Snurr, and Joseph T Hupp. De novo synthesis of a metal–organic framework material featuring ultrahigh surface area and gas storage capacities. Nature chemistry, 2(11):944–948, 2010.

[109] Simon Krause, Volodymyr Bon, Irena Senkovska, Ulrich Stoeck, Dirk Wallacher, Daniel M Többens, Stefan Zander, Renjith S Pillai, Guillaume Maurin, François-Xavier Coudert, et al. A pressure-amplifying framework material with negative gas adsorption transitions. Nature, 532(7599):348–352, 2016.

[110] Hiroyasu Furukawa, Yong Bok Go, Nakeun Ko, Young Kwan Park, Fernando J Uribe-Romo, Jaheon Kim, Michael O'Keeffe, and Omar M Yaghi. Isoreticular expansion of metal–organic frameworks with triangular and square building units and the lowest calculated density for porous crystals. Inorganic chemistry, 50(18):9147–9152, 2011.

[111] M Hassan Beyzavi, Nicolaas A Vermeulen, Ashlee J Howarth, Samat Tussupbayev, Aaron B League, Neil M Schweitzer, James R Gallagher, Ana E Platero-Prats, Nema Hafezi, Amy A Sarjeant, et al. A hafnium-based metal–organic framework as a nature-inspired tandem reaction catalyst. Journal of the American Chemical Society, 137(42):13624–13631, 2015.

[112] H. Deng, S. Grunder, K. E. Cordova, C. Valente, H. Furukawa, M. Hmadeh, F. Gandara, A. C. Whalley, Z. Liu, S. Asahina, H. Kazumori, M. O'Keeffe, O. Terasaki, J. F. Stoddart, and O. M. Yaghi. Large-pore apertures in a series of metal-organic frameworks. Science, 336(6084):1018–1023, May 2012.

[113] Gérard Férey, Christian Serre, Caroline Mellot-Draznieks, Franck Millange, Suzy Surblé, Julien Dutour, and Irène Margiolaki. A hybrid solid with giant pores prepared by a combination of targeted chemistry, simulation, and powder diffraction. Angewandte Chemie International Edition, 43(46):6296–6301, 2004.

[114] Gerard Férey, Caroline Mellot-Draznieks, Christian Serre, Franck Millange, Julien Dutour, Suzy Surblé, and Irena Margiolaki. A chromium terephthalate-based solid with unusually large pore volumes and surface area. Science, 309(5743):2040–2042, 2005.

[115] Luigi Pietro Cordella, Pasquale Foggia, Carlo Sansone, and Mario Vento. An improved algorithm for matching large graphs. In 3rd IAPR-TC15 workshop on graph-based representations in pattern recognition, pages 149–159, 2001.

[116] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[117] Christopher E Wilmer, Ki Chul Kim, and Randall Q Snurr. An extended charge equilibration method. The journal of physical chemistry letters, 3(17):2506–2511, 2012.

[118] David Dubbeldam, Sofía Calero, Donald E Ellis, and Randall Q Snurr. Raspa: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. Molecular Simulation, 42(2):81–101, 2016.

[119] Daniele Ongari, Peter G Boyd, Senja Barthel, Matthew Witman, Maciej Haranczyk, and Berend Smit. Accurate characterization of the pore volume in microporous crystalline materials. Langmuir, 33(51):14529–14538, 2017.

[120] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. Journal of machine learning research, 12(Oct):2825–2830, 2011.

[121] James Bergstra, Daniel Yamins, and David Daniel Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures, 2013.

[122] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. Explainable ai for trees: From local explanations to global understanding. arXiv preprint arXiv:1905.04610, 2019.

[123] Evelyn C Pielou. The measurement of diversity in different types of biological collections. Journal of theoretical biology, 13:131–144, 1966.

[124] Sean Gillies, A Bierbaum, K Lautaportti, and O Tonnhofer. Shapely: manipulation and analysis of geometric objects, 2007.

[125] Anna G Slater and Andrew I Cooper. Porous materials. function-led design of new porous materials. Science (New York, NY), 348(6238):aaa8075–aaa8075, 2015.

[126] Mark E Davis. Ordered porous materials for emerging applications. Nature, 417(6891):813, 2002.

[127] Tom Hasell and Andrew I Cooper. Porous organic cages: soluble, modular and molecular pores. Nature Reviews Materials, 1(9):16053, 2016.

[128] Andrew I Cooper. Porous molecular solids and liquids. ACS central science, 3(6):544–553, 2017.

[129] Jian Tian, Praveen K Thallapally, Scott J Dalgarno, Peter B McGrail, and Jerry L Atwood. Amorphous molecular organic solids for gas adsorption. Angewandte Chemie International Edition, 48(30):5492–5495, 2009.

[130] Gang Zhang, Oliver Presly, Fraser White, Iris M Oppel, and Michael Mastalerz. A permanent mesoporous organic cage with an exceptionally high surface area. Angewandte Chemie International Edition, 53(6):1516–1520, 2014.

[131] Ming Liu, Linda Zhang, Marc A. Little, Venkat Kapil, Michele Ceriotti, Siyuan Yang, Lifeng Ding, Daniel L. Holden, Rafael Balderas-Xicohténcatl, Donglin He, Rob Clowes, Samantha Y. Chong, Gisela Schütz, Linjiang Chen, Michael Hirscher, and Andrew I. Cooper.

Barely porous organic cages for hydrogen isotope separation. Science, 366(6465):613–620, 2019.

[132] L. Chen, P.S. Reiss, S.Y. Chong, D. Holden, K.E. Jelfs, T. Hasell, M.A. Little, A. Kewley, M.E. Briggs, A. Stephenson, K.M. Thomas, J.A. Armstrong, J. Bell, J. Busto, R. Noel, J. Liu, D.M. Strachan, P.K. Thallapally, and A.I. Cooper. Separation of rare gases and chiral molecules by selective binding in porous organic cages. Nature Materials, 13(10):954–960, 2014.

[133] Tamoghna Mitra, Kim E Jelfs, Marc Schmidtmann, Adham Ahmed, Samantha Y Chong, Dave J Adams, and Andrew I Cooper. Molecular shape sorting using molecular organic cages. Nature chemistry, 5(4):276, 2013.

[134] Alan Aspuru-Guzik, Roland Lindh, and Markus Reiher. The matter simulation (r) evolution. ACS central science, 4(2):144–152, 2018.

[135] Kurt Lejaeghere, Gustav Bihlmayer, Torbjörn Björkman, Peter Blaha, Stefan Blügel, Volker Blum, Damien Caliste, Ivano E. Castelli, Stewart J. Clark, Andrea Dal Corso, Stefano de Gironcoli, Thierry Deutsch, John Kay Dewhurst, Igor Di Marco, Claudia Draxl, Marcin Dułak, Olle Eriksson, José A. Flores-Livas, Kevin F. Garrity, Luigi Genovese, Paolo Giannozzi, Matteo Giantomassi, Stefan Goedecker, Xavier Gonze, Oscar Grånäs, E. K. U. Gross, Andris Gulans, François Gygi, D. R. Hamann, Phil J. Hasnip, N. A. W. Holzwarth, Diana Iuşan, Dominik B. Jochym, François Jollet, Daniel Jones, Georg Kresse, Klaus Koepernik, Emine Küçükbenli, Yaroslav O. Kvashnin, Inka L. M. Locht, Sven Lubeck, Martijn Marsman, Nicola Marzari, Ulrike Nitzsche, Lars Nordström, Taisuke Ozaki, Lorenzo Paulatto, Chris J. Pickard, Ward Poelmans, Matt I. J. Probert, Keith Refson, Manuel Richter, Gian-Marco Rignanese, Santanu Saha, Matthias Scheffler, Martin Schlipf, Karlheinz Schwarz, Sangeeta Sharma, Francesca Tavazza, Patrik Thunström, Alexandre Tkatchenko, Marc Torrent, David Vanderbilt, Michiel J. van Setten, Veronique Van Speybroeck, John M. Wills, Jonathan R. Yates, Guo-Xu Zhang, and Stefaan Cottenier. Reproducibility in density functional theory calculations of solids. Science, 351(6280), 2016.

[136] Qingyuan Yang, Dahuan Liu, Chongli Zhong, and Jian-Rong Li. Development of computational methodologies for metal–organic frameworks and their application in gas separations. Chemical Reviews, 113(10):8261–8323, 2013.

[137] Li-Chiang Lin, Adam H. Berger, Richard L. Martin, Jihan Kim, Joseph A. Swisher, Kuldeep Jariwala, Chris H. Rycroft, Abhoyjit S. Bhown, Michael W. Deem, Maciej Haranczyk, and Berend Smit. In silico screening of carbon-capture materials. Nature Materials, 11:633 EP –, 05 2012.

[138] Lukas Turcani, Rebecca L Greenaway, and Kim E Jelfs. Machine learning for organic cage property prediction. Chemistry of Materials, 2018.

[139] Valentina Santolini, Marcin Miklitz, Enrico Berardo, and Kim E Jelfs. Topological landscapes of porous organic cages. Nanoscale, 9(16):5280–5298, 2017.

[140] Angeles Pulido, Linjiang Chen, Tomasz Kaczorowski, Daniel Holden, Marc A. Little, Samantha Y. Chong, Benjamin J. Slater, David P. McMahon, Baltasar Bonillo, Chloe J. Stackhouse, Andrew Stephenson, Christopher M. Kane, Rob Clowes, Tom Hasell, Andrew I. Cooper, and Graeme M. Day. Functional materials discovery using energy–structure–function maps. Nature, 543:657 EP –, 03 2017.

[141] Michael Mastalerz and Iris M Oppel. Rational construction of an extrinsic porous molecular crystal with an extraordinary high specific surface area. Angewandte Chemie International Edition, 51(21):5252–5255, 2012.

[142] Sarah L Price. From crystal structure prediction to polymorph prediction: interpreting the crystal energy landscape. Physical Chemistry Chemical Physics, 10(15):1996–2009, 2008.

[143] Edward O Pyzer-Knapp, Hugh PG Thompson, Florian Schiffmann, Kim E Jelfs, Samantha Y Chong, Marc A Little, Andrew I Cooper, and Graeme M Day. Predicted crystal energy landscapes of porous organic cages. Chemical Science, 5(6):2235–2245, 2014.

[144] Sarah L Price. Predicting crystal structures of organic compounds. Chemical Society Reviews, 43(7):2098–2111, 2014.

[145] Pablo M. Piaggi. Entropy as a tool for crystal discovery. PhD thesis, École Polytechnique Fédérale de Lausanne (EPFL), 2019.

[146] Pablo M Piaggi and Michele Parrinello. Predicting polymorphism in molecular crystals using orientational entropy. Proceedings of the National Academy of Sciences, 115(41):10251–10256, 2018.

[147] Daniel Schwalbe-Koda, Zach Jensen, Elsa Olivetti, and Rafael Gómez-Bombarelli. Graph similarity drives zeolite diffusionless transformations and intergrowth. Nature materials, 18(11):1177–1181, 2019.

[148] Herbert Edelsbrunner and John Harer. Computational topology: an introduction. American Mathematical Soc., 2010.

[149] Herbert Edelsbrunner and John Harer. Persistent homology-a survey. Contemporary mathematics, 453:257–282, 2008.

[150] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. arXiv preprint arXiv:1710.04019, 2017.

[151] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. In Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on, pages 454–463. IEEE, 2000.

[152] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. Discrete & Computational Geometry, 33(2):249–274, 2005.

[153] Mohammad Saadatfar, Hiroshi Takeuchi, Vanessa Robins, Nicolas Francois, and Yasuaki Hiraoka. Pore configuration landscape of granular crystallization. Nature communications, 8:15082, 2017.

[154] Yasuaki Hiraoka, Takenobu Nakamura, Akihiko Hirata, Emerson G Escolar, Kaname Matsue, and Yasumasa Nishiura. Hierarchical structures of amorphous solids characterized by persistent homology. Proceedings of the National Academy of Sciences, page 201520877, 2016.

[155] Wenxiang Xu, Dongyang Zhang, Peng Lan, and Yang Jiao. Multiple-inclusion model for the transport properties of porous composites considering coupled effects of pores and interphase around spheroidal particles. International Journal of Mechanical Sciences, 150:610–616, 2019.

[156] Ludovic Duponchel. When remote sensing meets topological data analysis. Journal of Spectral Imaging, 2018.

[157] Richard Luis Martin, Berend Smit, and Maciej Haranczyk. Addressing challenges of identifying geometrically diverse sets of crystalline porous materials. Journal of chemical information and modeling, 52(2):308–318, 2011.

[158] Ulrich Bauer. Ripser. http://ripser.org. Accessed: 2017-08-01.

[159] Peter Bubenik. Statistical topological data analysis using persistence landscapes. The Journal of Machine Learning Research, 16(1):77–102, 2015.

[160] Peter Bubenik and Paweł Dłotko. A persistence landscapes toolbox for topological statistics. Journal of Symbolic Computation, 78:91–114, 2017.

[161] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. EPJ Data Science, 6(1):17, 2017.

[162] Vin De Silva and Joshua B Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, Technical report, Stanford University, 2004.

[163] Nathan Brown. In Silico Medicinal Chemistry: Computational Methods to Support Drug Design. Royal Society of Chemistry, 2015.

[164] Ingwer Borg and P Groenen. Modern multidimensional scaling: theory and applications. Journal of Educational Measurement, 40(3):277–280, 2003.

[165] Marta K Dudek and Graeme M Day. Explaining crystallization preferences of two polyphenolic diastereoisomers by crystal structure prediction. CrystEngComm, 2019.

[166] David P McMahon, Andrew Stephenson, Samantha Y Chong, Marc Little, James TA Jones, Andrew I Cooper, and Graeme Matthew Day. Computational modelling of solvent effects in a prolific solvatomorphic porous organic cage. Faraday Discussions, 2018.

[167] Robert W Howarth, Renee Santoro, and Anthony Ingraffea. Methane and the greenhouse-gas footprint of natural gas from shale formations. Climatic change, 106(4):679, 2011.

[168] Michael Fernandez, Tom K Woo, Christopher E Wilmer, and Randall Q Snurr. Large-scale quantitative structure–property relationship (qspr) analysis of methane storage in metal–organic frameworks. The Journal of Physical Chemistry C, 117(15):7681–7689, 2013.

[169] Genki Kusano, Kenji Fukumizu, and Yasuaki Hiraoka. Kernel method for persistence diagrams via kernel embedding and weight factor. The Journal of Machine Learning Research, 18(1):6947–6987, 2017.

[170] Chi Seng Pun, Kelin Xia, and Si Xian Lee. Persistent-homology-based machine learning and its applications–a survey. arXiv preprint arXiv:1811.00252, 2018.

[171] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An introduction to statistical learning, volume 112. Springer, 2013.

[172] Maryam Pardakhti, Ehsan Moharreri, David Wanik, Steven L Suib, and Ranjan Srivastava. Machine learning using combined structural and chemical descriptors for prediction of methane adsorption performance of metal organic frameworks (mofs). ACS combinatorial science, 19(10):640–645, 2017.

[173] George S Fanourgakis, Konstantinos Gkagkas, Emmanuel Tylianakis, Emmanuel Klontzas, and George E Froudakis. A robust machine learning algorithm for the prediction of methane adsorption in nanoporous materials. The Journal of Physical Chemistry A, 2019.

[174] Tina Düren, Lev Sarkisov, Omar M Yaghi, and Randall Q Snurr. Design of new materials for methane storage. Langmuir, 20(7):2683–2689, 2004.

[175] Cory M Simon, Jihan Kim, Li-Chiang Lin, Richard L Martin, Maciej Haranczyk, and Berend Smit. Optimizing nanoporous materials for gas storage. Physical Chemistry Chemical Physics, 16(12):5499–5513, 2014.

[176] Arni Sturluson, Melanie T Huynh, Arthur HP York, and Cory M Simon. Eigencages: Learning a latent space of porous cage molecules. ACS central science, 4(12):1663–1676, 2018.

[177] Angeles Pulido, Linjiang Chen, Tomasz Kaczorowski, Daniel Holden, Marc A. Little, Samantha Y. Chong, Benjamin J. Slater, David P. McMahon, Baltasar Bonillo, Chloe J. Stackhouse, Andrew Stephenson, Christopher M. Kane, Rob Clowes, Tom Hasell, Andrew I. Cooper, and Graeme M. Day. Additional computational data (related to "functional materials discovery using energy–structure–function maps" manuscript). http://eprints.soton.ac.uk/404749/. Accessed: 2018-1-1.

[178] Seyed Mohamad Moosavi, Arunraj Chidambaram, Leopold Talirz, Maciej Haranczyk, Kyriakos C Stylianou, and Berend Smit. Capturing chemical intuition in synthesis of metal-organic frameworks. Nature Communications, 10(1):539, 2019.

[179] J Douglas Carroll and Phipps Arabie. Multidimensional scaling. Annual review of psychology, 31(1):607–649, 1980.

[180] Hiroyasu Furukawa, Kyle E Cordova, Michael O'Keeffe, and Omar M Yaghi. The chemistry and applications of metal-organic frameworks. Science, 341(6149):1230444, 2013.

[181] N. Stock and S. Biswas. Synthesis of metal-organic frameworks (mofs): routes to various mof topologies, morphologies, and composites. Chem. Rev., 112, 2012.

[182] V. Duros. Human versus robots in the discovery and crystallization of gigantic polyoxometalates. Angew. Chem.-Int Ed., 56, 2017.

[183] Lin s., et al. mapping the dark space of chemical reactions with extended nanomole synthesis and maldi-tof ms. science361, pii: eaar6236 (2018).

[184] D. T. Ahneman, J. G. Estrada, S. S. Lin, S. D. Dreher, and A. G. Doyle. Predicting reaction performance in c-n cross-coupling using machine learning. Science, 360, 2018.

[185] B. Maryasin, P. Marquetand, and N. Maulide. Machine learning for organic synthesis: are robots replacing chemists? Angew. Chem.-Int Ed., 57, 2018.

[186] P. Raccuglia. Machine-learning-assisted materials discovery using failed experiments. Nature, 533, 2016.

[187] J. N. Wei, D. Duvenaud, and A. Aspuru-Guzik. Neural networks for the prediction of organic chemistry reactions. ACS Cent. Sci., 2, 2016.

[188] Z. P. Zhou, X. C. Li, and R. N. Zare. Optimizing chemical reactions with deep reinforcement learning. ACS Cent. Sci., 3, 2017.

[189] Mu x., chen y., lester e., wu t. optimized synthesis of nano-scale high quality hkust-1 under mild conditions and its application in co2 capture. microporous mesoporous mat. 270, 249-257 (2018).

[190] E. Biemmi, S. Christian, N. Stock, and T. Bein. High-throughput screening of synthesis parameters in the formation of the metal-organic frameworks mof-5 and hkust-1. Microporous Mesoporous Mat., 117, 2009.

[191] K. J. Kim. High-rate synthesis of cu-btc metal-organic frameworks. Chem. Commun., 49, 2013.

[192] S. S. Y. Chui, S. M. F. Lo, J. P. H. Charmant, A. G. Orpen, and I. D. Williams. A chemically functionalizable nanoporous material. Science, 283, 1999.

[193] Y. Peng. Methane storage in metal-organic frameworks: current records, surprise findings, and challenges. J. Am. Chem. Soc., 135, 2013.

[194] T. C. Le and D. A. Winkler. Discovery and optimization of materials using evolutionary approaches. Chem. Rev., 116, 2016.

[195] Henson a. b., gromski p. s., cronin l. designing algorithms to aid discovery by chemical robots. acs cent. sci. 4, 793-804 (2018).

[196] Polinsky a., feinstein r., shi s., kuki a. librain. in molecular diversity and combinatorial chemistry: Libraries and drug discovery vol. 996 (eds chaiken i.m., janda k.d. 219-232 (american chemical society), 1996).

[197] Kruskal j. b., wish m. multidimensional scaling, vol. 11 (sage: Newbury park, 1978).

[198] T. Duren, F. Millange, G. Ferey, K. S. Walton, and R. Q. Snurr. Calculating geometric surface areas as a characterization tool for metal-organic frameworks. J. Phys. Chem. C., 111, 2007.

[199] M. K. Bhunia, J. T. Hughes, J. C. Fettinger, and A. Navrotsky. Thermochemistry of paddle wheel mofs: Cu-hkust-1 and zn-hkust-1. Langmuir, 29, 2013.

[200] Moosavi s. m., et al. synthesis of metal-organic frameworks: capturing chemical intuition. available from: http://dx.doi.org/10.24435/materialscloud:2018.0011/v3 (2018).

[201] Matlab 2018a. global optimisation toolbox and statistics and machine learning toolbox. a natick ed. massachusetts, united states: The mathworks, inc.; 2018.

[202] A. Liaw and M. Wiener. Classification and regression by randomforest. R. News, 2, 2002.

[203] Lawrence Davis. Handbook of genetic algorithms. CUMINCAD, 1991.

[204] Zhou Jian and Wang Hejing. The physical meanings of 5 basic parameters for an x-ray diffraction peak and their application. Chinese journal of geochemistry, 22(1):38–44, 2003.

[205] DL Dorset. X-ray diffraction: a practical approach. Microscopy and microanalysis, 4(5):513–515, 1998.

[206] U Mueller, M Schubert, F Teich, H Puetter, K Schierle-Arndt, and J Pastre. Metal–organic frameworks—prospective industrial applications. Journal of Materials Chemistry, 16(7):626–636, 2006.

[207] Ashlee J Howarth, Yangyang Liu, Peng Li, Zhanyong Li, Timothy C Wang, Joseph T Hupp, and Omar K Farha. Chemical, thermal and mechanical stabilities of metal–organic frameworks. Nature Reviews Materials, 1:15018, 2016.

[208] Jasmina Hafizovic Cavka, Søren Jakobsen, Unni Olsbye, Nathalie Guillou, Carlo Lamberti, Silvia Bordiga, and Karl Petter Lillerud. A new zirconium inorganic building brick forming metal organic frameworks with exceptional stability. Journal of the American Chemical Society, 130(42):13850–13851, 2008.

[209] Rodney Hill. Elastic properties of reinforced solids: some theoretical principles. Journal of the Mechanics and Physics of Solids, 11(5):357–372, 1963.

[210] Ole Sigmund. Tailoring materials with prescribed elastic properties. Mechanics of Materials, 20(4):351–368, 1995.

[211] Maarten De Jong, Wei Chen, Thomas Angsten, Anubhav Jain, Randy Notestine, Anthony Gamst, Marcel Sluiter, Chaitanya Krishna Ande, Sybrand Van Der Zwaag, Jose J Plata, Comer Toher, Stefano Curtarolo, Gerbrand Ceder, Kristin A. Persson, and Mark Asta. Charting the complete elastic properties of inorganic crystalline compounds. Scientific data, 2:150009, 2015.

[212] Hongyou Fan, Christopher Hartshorn, Thomas Buchheit, David Tallant, Roger Assink, Regina Simpson, Dave J Kissel, Daniel J Lacks, Salvatore Torquato, and C Jeffrey Brinker. Modulus–density scaling behaviour and framework architecture of nanoporous self-assembled silicas. Nature materials, 6(6):418–423, 2007.

[213] Aurélie U Ortiz, Anne Boutin, Alain H Fuchs, and François-Xavier Coudert. Anisotropic elastic properties of flexible metal-organic frameworks: how soft are soft porous crystals? Physical review letters, 109(19):195502, 2012.

[214] Jin-Chong Tan, Bartolomeo Civalleri, Chung-Cherng Lin, Loredana Valenzano, Raimondas Galvelis, Po-Fei Chen, Thomas D Bennett, Caroline Mellot-Draznieks, Claudio M Zicovich-Wilson, and Anthony K Cheetham. Exceptionally low shear modulus in a prototypical imidazole-based metal-organic framework. Physical review letters, 108(9):095502, 2012.

[215] Benoit B Mandelbrot and Roberto Pignoni. The fractal geometry of nature, volume 1. WH freeman New York, 1983.

[216] Joanna Aizenberg, James C Weaver, Monica S Thanawala, Vikram C Sundar, Daniel E Morse, and Peter Fratzl. Skeleton of euplectella sp.: structural hierarchy from the nanoscale to the macroscale. Science, 309(5732):275–278, 2005.

[217] Roderic Lakes. Materials with structural hierarchy. Nature, 361(6412):511–515, 1993.

[218] Xiaoyu Zheng, Howon Lee, Todd H Weisgraber, Maxim Shusteff, Joshua DeOtte, Eric B Duoss, Joshua D Kuntz, Monika M Biener, Qi Ge, Julie A Jackson, Sergei O. Kucheyev, Nicholas X. Fang, and Christopher M. Spadaccini. Ultralight, ultrastiff mechanical metamaterials. Science, 344(6190):1373–1377, 2014.

[219] Lucas R Meza, Alex J Zelhofer, Nigel Clarke, Arturo J Mateos, Dennis M Kochmann, and Julia R Greer. Resilient 3d hierarchical architected metamaterials. Proceedings of the National Academy of Sciences, 112(37):11502–11507, 2015.

[220] Hui Wu, Taner Yildirim, and Wei Zhou. Exceptional mechanical stability of highly porous zirconium metal–organic framework uio-66 and its important implications. The journal of physical chemistry letters, 4(6):925–930, 2013.

[221] Lila Bouëssel du Bourg, Aurélie U Ortiz, Anne Boutin, and François-Xavier Coudert. Thermal and mechanical stability of zeolitic imidazolate frameworks polymorphs. APL materials, 2(12):124110, 2014.

[222] Eugene A Kapustin, Seungkyu Lee, Ahmad S Alshammari, and Omar M Yaghi. Molecular retrofitting adapts a metal–organic framework to extreme pressure. ACS Central Science, 3(6):662–667, 2017.

[223] Lev Sarkisov, Richard L Martin, Maciej Haranczyk, and Berend Smit. On the flexibility of metal–organic frameworks. Journal of the American Chemical Society, 136(6):2228–2231, 2014.

[224] Shuai Yuan, Weigang Lu, Ying-Pin Chen, Qiang Zhang, Tian-Fu Liu, Dawei Feng, Xuan Wang, Junsheng Qin, and Hong-Cai Zhou. Sequential linker installation: precise placement of functional groups in multivariate metal–organic frameworks. Journal of the American Chemical Society, 137(9):3177–3180, 2015.

[225] Quan-Guo Zhai, Xianhui Bu, Xiang Zhao, Dong-Sheng Li, and Pingyun Feng. Pore space partition in metal–organic frameworks. Accounts of chemical research, 50(2):407–417, 2017.

[226] Sven MJ Rogge, Michel Waroquier, and Veronique Van Speybroeck. Reliably modeling the mechanical stability of rigid and flexible metal–organic frameworks. Accounts of chemical research, 51:138 – 148, 2017.

[227] Rahul Banerjee, Anh Phan, Bo Wang, Carolyn Knobler, Hiroyasu Furukawa, Michael O'keeffe, and Omar M Yaghi. High-throughput synthesis of zeolitic imidazolate frameworks and application to co2 capture. Science, 319(5865):939–943, 2008.

[228] Jingjing Yang, Yue-Biao Zhang, Qi Liu, Christopher A Trickett, Enrique Gutierrez-Puebla, M Ángeles Monge, Hengjiang Cong, Abdulrahman Aldossary, Hexiang Deng, and Omar M Yaghi. Principles of designing extra-large pore openings and cages in zeolitic imidazolate frameworks. Journal of the American Chemical Society, 139(18):6448 – 6455, 2017.

[229] TD Bennett, J Sotelo, Jin-Chong Tan, and SA Moggach. Mechanical properties of zeolitic metal–organic frameworks: mechanically flexible topologies and stabilization against structural collapse. CrystEngComm, 17(2):286–289, 2015.

[230] Jin-Chong Tan, Bartolomeo Civalleri, Alessandro Erba, and Elisa Albanese. Quantum mechanical predictions to elucidate the anisotropic elastic properties of zeolitic imidazolate frameworks: Zif-4 vs. zif-zni. CrystEngComm, 17(2):375–382, 2015.

[231] Matthew R Ryder and Jin-Chong Tan. Explaining the mechanical mechanisms of zeolitic metal–organic frameworks: revealing auxeticity and anomalous elasticity. Dalton Transactions, 45(10):4154–4161, 2016.

[232] Matthew R Ryder, Thomas Douglas Bennett, Chris Kelley, Mark Frogley, Gianfelice Cinque, and Jin-Chong Tan. Tracking thermal-induced amorphization of a zeolitic imidazolate framework via synchrotron in situ far-infrared spectroscopy. Chemical Communications, 53:7041–7044, 2017.

[233] Ch. Baerlocher and L. B. McCusker. Database of zeolite structures, accessed in October 2016.

[234] John Frederick Nye. Physical properties of crystals: their representation by tensors and matrices. Oxford university press, New York, 1985.

[235] Seyed mohamad moosavi, peter g. boyd, lev sarkisov, berend smit, improving the mechanical stability of metal-organic frameworks using chemical caryatids, materials cloud archive (2018), doi: 10.24435/materialscloud:2018.0004/v4.

[236] Donald J Jacobs, Andrew J Rader, Leslie A Kuhn, and Michael F Thorpe. Protein flexibility predictions using graph theory. Proteins: Structure, Function, and Bioinformatics, 44(2):150–165, 2001.

[237] Anh Phan, Christian J Doonan, Fernando J Uribe-Romo, Carolyn B Knobler, Michael O'keeffe, and Omar M Yaghi. Synthesis, structure, and carbon dioxide capture properties of zeolitic imidazolate frameworks. Acc. Chem. Res, 43(1):58–67, 2010.

[238] Michael O'Keeffe, Maxim A Peskov, Stuart J Ramsden, and Omar M Yaghi. The reticular chemistry structure resource (rcsr) database of, and symbols for, crystal nets. Accounts of chemical research, 41(12):1782–1789, 2008.

[239] Thomas D Bennett and Anthony K Cheetham. Amorphous metal–organic frameworks. Accounts of chemical research, 47(5):1555–1562, 2014.

[240] Aurélie U Ortiz, Anne Boutin, Alain H Fuchs, and François-Xavier Coudert. Investigating the pressure-induced amorphization of zeolitic imidazolate framework zif-8: mechanical instability due to shear mode softening. The journal of physical chemistry letters, 4(11):1861–1865, 2013.

[241] Thomas D Bennett, David A Keen, Jin-Chong Tan, Emma R Barney, Andrew L Goodwin, and Anthony K Cheetham. Thermal amorphization of zeolitic imidazolate frameworks. Angewandte Chemie International Edition, 50(13):3067–3071, 2011.

[242] Steve Plimpton. Fast parallel algorithms for short-range molecular dynamics. Journal of computational physics, 117(1):1–19, 1995.

[243] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: visual molecular dynamics. Journal of molecular graphics, 14(1):33–38, 1996.

[244] Stephen L Mayo, Barry D Olafson, and William A Goddard. Dreiding: a generic force field for molecular simulations. Journal of Physical chemistry, 94(26):8897–8909, 1990.

[245] John D Head and Michael C Zerner. A broyden–fletcher–goldfarb–shanno optimization procedure for molecular geometries. Chemical physics letters, 122(3):264–270, 1985.

[246] IA Baburin, Stefano Leoni, and G Seifert. Enumeration of not-yet-synthesized zeolitic zinc imidazolate mof networks: a topological and dft approach. The Journal of Physical Chemistry B, 112(31):9437–9443, 2008.

[247] Christian A Schröder, Igor A Baburin, Leo van Wüllen, Michael Wiebcke, and Stefano Leoni. Subtle polymorphism of zinc imidazolate frameworks: temperature-dependent ground states in the energy landscape revealed by experiment and theory. CrystEngComm, 15(20):4036–4040, 2013.

[248] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. science, 220(4598):671–680, 1983.

[249] Charles Kittel. Introduction to solid state physics. Wiley, New York, 1996.

[250] Rostam Golesorkhtabar, Pasquale Pavone, Jürgen Spitaler, Peter Puschnig, and Claudia Draxl. Elastic: A tool for calculating second-order elastic constants from first principles. Computer Physics Communications, 184(8):1861–1873, 2013.

[251] Roberto Dovesi, Roberto Orlando, Bartolomeo Civalleri, Carla Roetti, Victor R Saunders, and Claudio M Zicovich-Wilson. Crystal: a computational tool for the ab initio study of the electronic properties of crystals. Zeitschrift für Kristallographie-Crystalline Materials, 220(5/6):571–573, 2005.

[252] Matthew R Ryder, Bartolomeo Civalleri, Thomas D Bennett, Sebastian Henke, Svemir Rudić, Gianfelice Cinque, Felix Fernandez-Alonso, and Jin-Chong Tan. Identifying the role of terahertz vibrations in metal-organic frameworks: from gate-opening phenomenon to shear-driven structural destabilization. Physical review letters, 113(21):215502, 2014.

[253] Bin Zheng, Marco Sant, Pierfranco Demontis, and Giuseppe B Suffritti. Force field for molecular dynamics computations in flexible zif-8 framework. The Journal of Physical Chemistry C, 116(1):933–938, 2012.

[254] Lars Öhrström. Framework chemistry transforming our perception of the solid state, 2017.

[255] Vladislav A Blatov, Alexander P Shevchenko, and Davide M Proserpio. Applied topological analysis of crystal structures with the program package topospro. Crystal Growth & Design, 14(7):3576–3586, 2014.

[256] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. Journal of Applied physics, 52(12):7182–7190, 1981.