Master Thesis

Antonios P. Sarikas

January 14, 2024

Publisher

Contents

Chapter 1

Introduction

RETICULAR CHEMISTRY, a field that bridges inorganic and organic chemistry (Yaghi 2020), has emerged from a simple albeit powerful idea: *combining molecular building blocks to form extended crystalline structures* (Yaghi 2019). It all started in 1990s, with the advent of metal-organic frameworks (MOFs), the first "offspring" of reticular chemistry. MOFs, a class of nanoporous materials *composed of metal ions or clusters coordinated to organic ligands aka organic linkers*, possess extraordinary properties, such as ultrahigh porosity and huge surface areas (Farha et al. 2012). To get a sense of how extraordinary these materials are, it is suffice to say that *one gram of such a material can have a surface area as large as a soccer field*. The fact that reticular materials are "brought to life" by combining simple building blocks, allows chemists and material scientists to design materials in a judicious manner. The epitome of design in reticular chemistry is found in the synthesis of a zirconium-based MOF (Alezi et al. 2016), incorporating the polybenzene network or "cubic graphite" structure, predicted about 70 years ago.

1.1 Applications of Reticular Chemistry

Owing to their aforedescribed properties along with their extremely tunable and modular nature, MOFs have been considered prominent solutions for gas-adsorption related problems (Li et al. 2007; Jiang et al. 2022). MOFs find application in fields such as gas storage and separation, catalysis and drug delivery, just to name a few.

Carbon capture is a prime example (An et al. 2009; Sumida et al. 2011; Qazvini et al. 2021), where MOF-based sorbents have been deemed as green, low-cost and energy-efficient solutions. These materials provide versatile solutions to carbon capture, spanning various phases of the capture process, with direct air capture (DAC) being a noteworthy example. DAC includes chemical or physical methods for extracting carbon dioxide directly from the ambient air, with MOF-powered DAC showing great potential as a green and sustainable strategy for reducing carbon dioxide levels, contributing to the combating of climate change (Bose et al. 2023).

Hydrogen storage is one of the greatest challenges of hydrogen economy, currently inhibiting the transition from fossil fuels to hydrogen. Fortunately, characteristics of MOF adsorbents such as fast adsorption/desorption kinetics, low operating pressures and high hydrogen capacities, render them as promising answers to the aforementioned challenge (Suh et al. 2011; Suresh et al. 2021).

Methane is an attractive fuel for vehicular applications, being a relatively clean-burning fuel compared to gasoline. ***Methane storage*** in sorbents known as adsorbed natural gas (ANG) exhibit advantages over compressed natural gas (CNG) and liquefied natural gas (LNG), both in terms of energy-efficiency and vehicular safety. MOFs (Ma et al. 2007; Spanopoulos et al. 2016; Tsangarakis et al.

2023) and their “reticular siblings” covalent organic frameworks (COFs)—composed only of light elements—show great promise as ANG solutions (Furukawa et al. 2009; Mendoza-Cortes et al. 2011; Martin et al. 2014; Tong et al. 2018).

1.2 The Problem

The intrinsic combinatorial character of reticular chemistry, translates to practically an infinite number of realizable structures. Currently, the Cambridge Structural Database (CSD) contains more than 100 000 experimentally synthesized MOFs (P. Z. Moghadam et al. 2017) while the arrival of in silico designed MOFs (Wilmer et al. 2011; Colón et al. 2017; Boyd et al. 2019; Chung et al. 2019; Lee et al. 2021; Rosen et al. 2021; De Vos et al. 2023) has immensely expanded the available material pool. The huge size of current and future MOF databases (Lee et al. 2021) is both a blessing and a curse for the identification of novel materials. Blessing, since a large number of candidate structures doesn't limit our choices and as such, the chances to find the right material for a given problem. Curse, since the enormous size of MOFs space makes it harder for researchers to efficiently explore it, complicating the tracing of materials with the desired properties. It is therefore crucial to find a way that allows us to efficiently explore such a huge material space. Another way to rephrase our problem is the following: ***Given a large catalog of MOFs, is there a way to efficiently filter out the most promising ones for the application of interest?***

As a first approach to deal with this challenge, one could, in principle experimentally synthesize and characterize each one of the materials listed in the given catalog. Although *experimental synthesis and characterization* is the ultimate way to assess the performance of a material[1], the fact that a single laboratory study can take days or even months, renders experimental techniques impractical. A more efficient approach is computational screening based on *molecular simulations*, which for years has served as the principal tool for the discovery of high-performing MOFs (**Simon2015**; **Moghadam2018**; Banerjee et al. 2016; Gómez-Gualdrón et al. 2016; Jeong et al. 2017). Although computational screening dramatically accelerates the assessment of a single material compared to experimental techniques, brute-force screening of current and upcoming databases is considered suboptimal, given the size of the latter.

Machine learning (ML) aka *data-driven techniques* come to the rescue when dealing with *big data* and over the last years have picked up the torch from molecular simulations regarding material characterization. Given a collection of *input-output* pairs, i.e. a mathematical representation of a material and a corresponding property, a ML algorithm[2] seeks to *uncover the underlying structure-property relationship*. To put it in a nutshell, a ML algorithm “eats” *data*—which may come either from experiments, simulations or a combination of the two—and “spits out” a *model*, which can be *used to sort a large catalog of MOFs in just few seconds*. Obviously, for ML approaches to be effective and reliable, it is necessary that the resulting models are of high quality.

- Decide if figures must be added
- Add missing citations

[1] As Richard Feynmann said: *“The test of all knowledge is experiment. Experiment is the sole judge of scientific truth”*.

[2] Note that ML algorithms are not limited to solve only such kind of problems, which fall under the umbrella of supervised learning. See subsection **??** for other types of problem tackled by ML.

1.3 Literature Review

One of, if not the most important factor for the performance of ML models, is the way we select to mathematically represent materials or molecules. In other words, the type and amount of chemical information that is "injected" into these representations commonly known as *descriptors*, can make the difference between a high-performing and a baseline model. As such, it is of uttermost importance to employ descriptors that provide sufficient information for the properties of materials or molecules we are interested in to predict.

With regards to gas adsorption in MOFs, one of the first and most commonly used descriptors, are the so called *geometric* ones, which capture the pore environment of the framework. This type of descriptors includes textual characteristics of MOFs such as void fraction, gravimetric surface area and pore limiting diameter. Although ML models build with these descriptors work particularly well at the high pressure regime (Fernandez et al. 2013; Dureckova et al. 2019; Wu et al. 2020), their performance deteriorates when adsorption takes place at low pressures or the guest molecules exhibit non-negligible electrostatic interactions with the framework atoms. This performance drop should be expected, since geometric descriptors completely ignore the "cornerstone" of adsorption: *host-guest interactions*.

In order to improve the performance of ML models and bypass the limitations of geometric descriptors in the aforementioned conditions, another type of descriptors known as *energy-based* descriptors (Simon, Mercado, et al. 2015; Fanourgakis et al. 2019; Orhan et al. 2023; Shi et al. 2023), has been introduced. This type of descriptors supply ML algorithms with information regarding the energetics of adsorption, and can be used standalone or in combination with geometric descriptors.

In one of the first works to fingerprint the energetic landscape of MOFs (Bucior et al. 2019), energy histograms derived from the interactions of guest molecules with the framework atoms were used to predict hydrogen and methane uptake with remarkable accuracy. Prior to calculating the energy histograms, a three dimensional grid is overlayed on the unit cell of the MOF. Next, at each point of the grid, the interaction between the guest molecule with the framework atoms is calculated, producing a three dimensional energy grid. Finally, this energy grid is converted into a histogram, by partitioning the energy values of the grid into bins of specific energy width. By using solely these histograms as descriptors—without including any textual property—Bucior and his coworkers trained Lasso regression models, for predicting: i). H_2 swing capacity between $100\,\mathrm{bar}$ and $2\,\mathrm{bar}$ at $77\,\mathrm{K}$ ii). CH_4 swing capacity between $65\,\mathrm{bar}$ and $5.8\,\mathrm{bar}$ at $298\,\mathrm{K}$. The resulting models were extremely accurate, achieving a mean absolute error (MAE) of $2.3\,\mathrm{g\,L^{-1}}$ and $12.9\,\mathrm{cm^3\,cm^{-3}}$ for H_2 and CH_4, respectively, tested on the hMOFs database (Wilmer et al. 2011).

In another work (**generic**), a set of descriptors based on the average interaction between fictitious probe particles and the framework atoms was introduced. Two different types of probe particles were proposed: i). Vprobe particles, which account for the van der Waals interactions ii). Dprobe particles, which are neutrally charged electric dipoles and account for the electrostatic interactions. Each of these fictitious probe particles is randomly inserted at different positions of the unit cell, and the interaction energy between the probe and the framework atoms is calculated. The interaction energies at the different positions are averaged out, producing an energetic fingerprint of the material. These fingerprints in combination with six geometric descriptors formed the input for the Random Forest (RF) algorithm, which was trained to predict gas uptake for a plethora of guest molecules and thermodynamic conditions, on the Computation-Ready Experimental (CoRE) MOF database (**chong47**). The ML models showed impressive performance, showing an R^2 value of: i). 0.874 for H_2 uptake at $77\,\mathrm{K}$ and $2\,\mathrm{bar}$

ii). 0.889 for CH_4 uptake at 298 K and 5.8 bar. A highlight of this work was the exceptional performance of the ML model with regards to CO_2 uptake at 300 K and 0.1 bar, achieving an R^2 score of 0.832. At the same conditions, the ML model trained with geometric descriptors only achieved an R^2 score of 0.507. That is, the "injection" of energetic information resulted in 60 % increase in accuracy.

1.4 Objective of Thesis

Index