Contents

List of Figures

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 (Y. 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 (Moghadam, A. Li, et al. 2017) while the arrival of in silico designed MOFs (Wilmer et al. 2011; Colón et al. 2017; Boyd et al. 2019; Chung, Haldoupis, 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 (see Figure 1.1. 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 (Simon, Kim, et al. 2015; Banerjee et al. 2016; Gómez-Gualdrón et al. 2016; Jeong et al. 2017; Moghadam, Islamoglu, et al. 2018). 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.

[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 Section **??** for other types of problem tackled by ML.

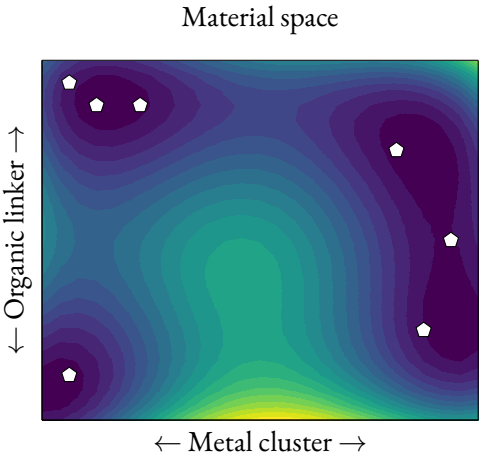


Figure 1.1: Material space of MOFs. Each point in this space corresponds to a unique combination of organic linker and metal cluster, whereas the associated color denotes the "score" of material for a given application. Finding the best material for a given application, amounts to sovling a (probably) non-convex optimization problem.

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, Gkagkas, Tylianakis, Klontzas, 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 his-

tograms, a 3D 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 3D energy grid. The latter is finally 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 et al. (2019) trained Lasso regression models, for predicting: i). H_2 swing capacity between 100 bar and 2 bar at 77 K ii). CH_4 swing capacity between 65 bar and 5.8 bar at 298 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 (Fanourgakis, Gkagkas, Tylianakis, and Froudakis 2020), 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 (Chung, Camp, et al. 2014). The ML models showed impressive performance, showing an R^2 value of: i). 0.874 for H_2 uptake at 77 K and 2 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 Thesis Statement

In the aforedescribed works (Bucior et al. 2019; Fanourgakis, Gkagkas, Tylianakis, and Froudakis 2020), a general pattern can be recognized with regards to the building of the ML models. Starting from the potential energy surface (PES) or an approximation thereof, energetic fingerprints are manually handcrafted based on some heuristic, and these fingerprints are then used to train a ML algorithm. However, a lot of information has been lost during the conversion of the PES into fingerprints, as a 3D object is converted into an 1D object. *Since gas adsorption comes down to the PES, it is reasonable to question whether one can use the PES itself as descriptor*. By doing this: i). The information content that goes into a ML algorithm is increased ii). The computational cost remains the same relative to the previously described works iii). It is no longer necessary to manually handcraft fingerprints.

In this thesis, ***a generalized framework to predict gas adsorption properties is proposed, using the PES as raw input****. In order to be machine understandable, the PES is first voxelized—the voxelized PES is essentially a 3D energy image—and then, it is processed by a 3D convolutional neural network (CNN), known for their ability to process image-like data. The proposed scheme is schematically presented* in Figure 1.2.

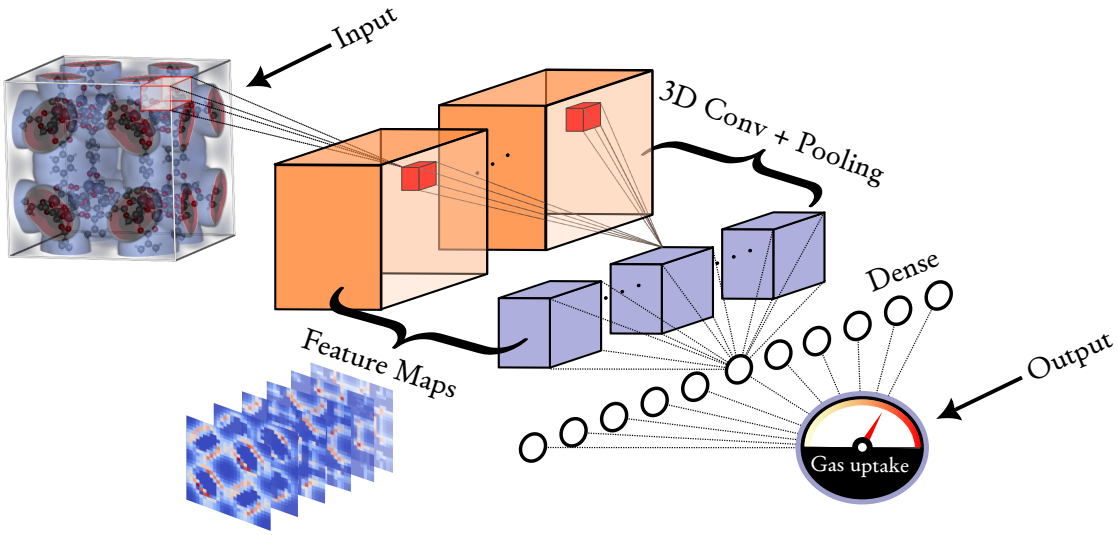


Figure 1.2: Proposed scheme to predict gas adsorption properties, starting from the PES as raw input. A 3D CNN extracts its features from the PES, and then uses them to predict the adsorption property of interest. The IRMOF-1 structure and PES were visualized with the iRASPA software (Dubbeldam et al. 2018).

Chapter 2

Methodology

NEURAL NETWORKS are notorious for being "data hungry", requiring a relatively large amount of training data, in order to unleash their full potential. As such, to get a representative picture of the capabilities of the proposed deep learning (DL) framework, two large datasets are employed to train the 3D CNN. The first one, is a subset of the University of Ottawa (UO) database (Boyd et al. 2019), and is used to verify the applicability of the proposed pipeline, examining CO_2 uptake. The second one, is the COFs database generated by (Mercado et al. 2018), and is employed to demonstrate the transferability of the approach, examining CH_4 uptake. Please note, that these datasets are already labeled, and as such no molecular simulations were performed in this study to generate the labels (gas uptakes) of the materials. Information regarding the Grand Canonical Monte Carlo (GCMC) calculations that were performed to produced the labels, can be found in the original works.

2.1 Datasets

2.1.1 MOFs dataset

The UO database is composed of 324 426 hypothetical MOFs. Randomly selected subsets of size 32 432, 5000 and 27 438, served as the training, validation[1] and test sets, respectively. The absolute CO_2 uptake at 298 K and 0.15 bar was examined and the following eight textual properties were used as input for the conventional models: unit cell's mass and volume, gravimetric surface area, void fraction, void volume, largest free sphere diameter, largest included sphere along free sphere path diameter and largest included sphere diameter. For producing the learning curves shown in Figure 3.3a, the training set size was varied and the following training set sizes were considered:

$$\{100, 500, 1000, 2000, 5000, 10\,000, 15\,000, 20\,000, 32\,432\} \tag{2.1}$$

The energy voxels of MOFs are publicly available in figshare.

2.1.2 COFs dataset

The COFs database contains 69 839 and provides data for five textual properties and CH_4 uptake at different thermodynamic conditions. A randomly selected subset of 55 871 materials served as the training set, whereas the remaining 13 698 correspond to the test set. In this work, CH_4 uptake at

[1] The validation set was used to select the number of epochs, see Section 2.3.2.

298 K and 5.8 bar was examined. The following five textual properties were used to build the conventional models: density, gravimetric surface area, void fraction, pore limiting diameter and largest cavity diameter. For producing the learning curves shown in Figure 3.3b, the training set size was varied and the following training set sizes were considered:

$$\{5000, 10\,000, 15\,000, 20\,000, 35\,000, 55\,871\} \tag{2.2}$$

2.2 Voxelized PES

In order to calculate the voxelized PES, first a 3D grid of size $n \times n \times n$ is overlayed over the unit cell of the material. Second, at each voxel centered at grid point $\boldsymbol{r}_i$, the interaction of the guest molecule with the framework atoms $\mathcal{V}(\boldsymbol{r}_i)$ is calculated, and this energy value "colorizes" the corresponding voxel. The workflow to construct the voxelized PES is schematically depicted in Figure 2.1. The grid size n and the type of the potential control the "trade-off" between information content and computational cost. The greater the grid size n, the greater the resolution of the energy image and as such, the information content. However, this comes at the cost of increased computational cost which is by no means negligible, since voxelization scales up as $\mathcal{O}(n^3)$. Similarly, the more accurate the modeling of interactions, the greater the information content, but again, extra computational burden is required. *The voxelized PES converges to the exact one as $n \to \infty$ and when the voxels are filled with energy values derived from ab-initio calculations.*

In this work we strived for minimal computational cost, setting $n = 25$ and modeling all interactions with the Lennard-Jones (LJ) potential, using a spherical probe molecule as guest. The interaction energy $\mathcal{V}(\boldsymbol{r}_i)$ between the spherical probe and the framework atoms was calculated as following:

$$\mathcal{V}(\boldsymbol{r}_i) = \sum_{\substack{j=1 \\ r_{ij} \leq r_c}}^{N} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.3}$$

where N is the number of framework atoms, r_c is the cutoff radius which was set to 10 Å, r_{ij} is the distance between the j-th framework atom and the probe molecule and ϵ_{ij} and σ_{ij} combine the ϵ and σ values of the probe molecule and the j-th framework atom using the Lorentz-Berthelot mixing rules:

$$\sigma_{ij} = \frac{\sigma_i + \sigma_j}{2} \quad \wedge \quad \epsilon_{ij} = \sqrt{\epsilon_i + \epsilon_j} \tag{2.4}$$

If there is geometric overlap between a grid point and the position of a framework atom, the interaction energy can be extremely repulsive, leading to very large, even infinite values, which can hamper or not allow the training of a neural network (NN) at all. For this reason, each voxel was filled with $e^{-\beta \mathcal{V}(\boldsymbol{r}_i)}$, which tends to 0 as $\mathcal{V}(\boldsymbol{r}_i) \to \infty$, where $\beta = \frac{1}{k_B T}$ is the Boltzmann constant and T is the temperature, which was set at 298 K. The Python package MOX$\epsilon\lambda$ was introduced to facilitate the calculation of energy voxels. In the remaining of this thesis, the terms "voxelized PES" and "energy voxels", are used interchangeably.

2.3 Machine Learning Details

For the conventional ML models, the RF algorithm as implement in the scikit-learn (Pedregosa et al. 2011) package (version 1.2.2) was used, while the PyTorch (Paszke et al. 2019) framework (version

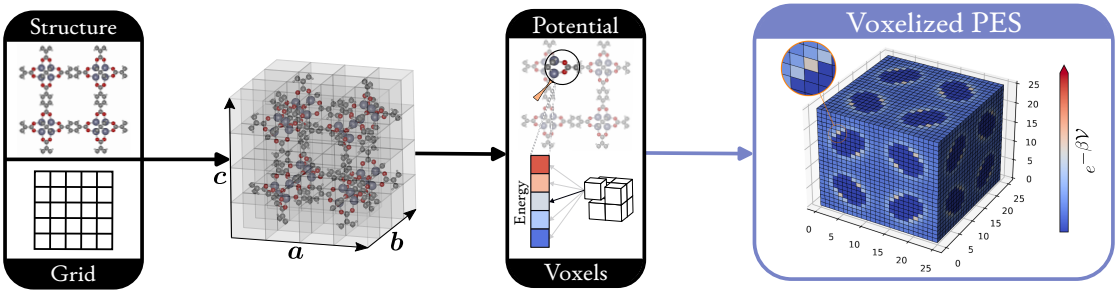


Figure 2.1: Workflow to construct the voxelized PES. The grid size and the type of the potential control the "trade-off" between information content and computational cost. The IRMOF-1 structure was visualized with the iRASPA software (Dubbeldam et al. 2018).

2.0.1+cu118) was employed for the CNN models. The performance, i.e. the generalization ability of the models, was assesed by the coefficient of determination R^2:

$$R^2 := 1 - \frac{\sum_{i=1}^{N_{\text{test}}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N_{\text{test}}} (y_i - \bar{y})^2} \tag{2.5}$$

where N_{test} is the number of samples in test set, $\bar{y}$ is the mean value of y in the test set and y_i, $\hat{y}_i$ are the ground truth and predicted values of the i-the sample, respectively. In all cases where confidence interval (CI) are presented, they were calculated using the percentile bootstrap method (Efron et al. 1994), with 10 000 bootstrapped samples from the test set.

2.3.1 CNN architecture

The architecture of the 3D CNN is presented in Figure 3.1, whereas a PyTorch implementation is publicly available in: RetNet. Kernel size is set to 3 for Conv1, Conv2 and 2 for Conv3, Conv4 and Conv5 layers. Stride equals 1 for all Conv layers and only Conv1 layer is padded, with "same" padding and "periodic" mode. For both MaxPool layers, kernel size and stride are both set to 2. For the Dropout layer, the dropout rate p equals 0.3, while the negative slope is set to 0.01 for all LeakyReLU layers.

2.3.2 Preprocessing & CNN training details

Prior to entering the CNN the energy voxels are standardized "on the fly" based on the training set statistics—this transformation is applied both during training and inference—which are computed channel wise[2]. The voxelized PES of a material $\boldsymbol{x}$, enters the CNN as following:

$$\boldsymbol{x}' = \frac{\boldsymbol{x} - \mu_{\text{train}}}{\sigma_{\text{train}}} \tag{2.6}$$

Regarding CNN training, mean squared loss (MSL) is used as loss function and weights are initialized according to the He scheme (He et al. 2015). The Adam optimizer (Kingma et al. 2017) is employed, with $|\mathcal{B}| = 64, \eta = 0.001, \beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 1 \times 10^{-8}$. The CNN training lasts for 50

[2] The voxelized PES is essentialy a single channel, i.e. grayscale, image.

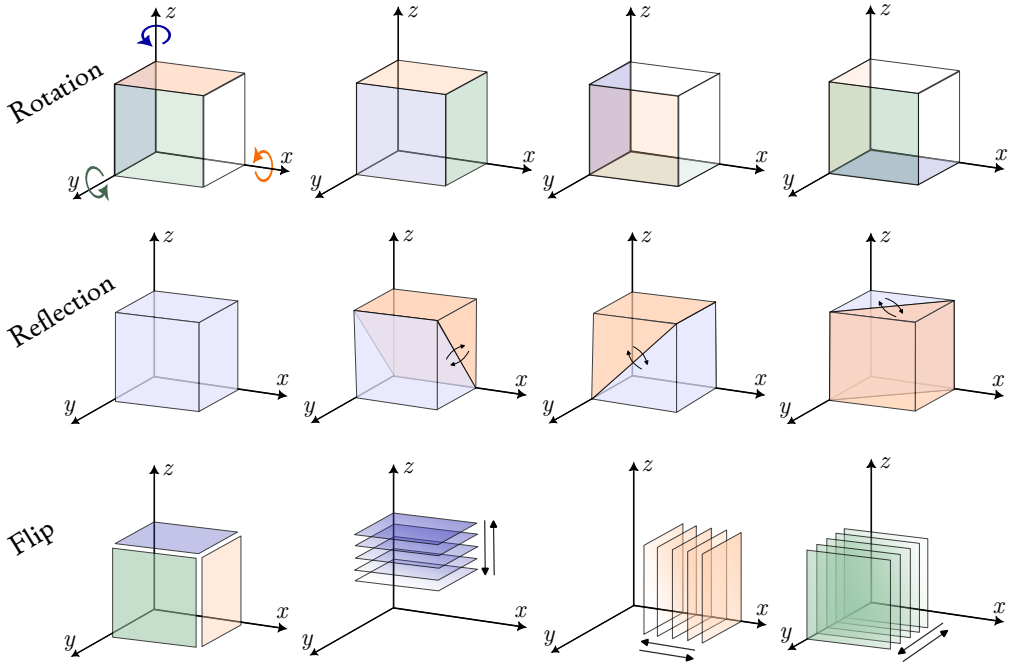


Figure 2.2: Geometric transformations for data augmentation.

epochs with the learning rate being decaybed by 0.5 every 10 epochs. RetNet was trained in the MOFs dataset with the largest training set size (32 432 training samples), for a different number of epochs, namely 10, 20 and 50. The latter value was selected, since it showed the greatest performance in the validation set.

2.3.3 Data augmentation

ADD SECTION FROM THEORY FOR AUGMENTATION

With this technique, the training set is artifically increased, by applying transformations on the input that leave the label unchanged. With regards to gas adsorption, this amounts to applying geometric transformations on the voxelized PES, that leave the gas uptake value of the material unchanged. Data augmentation, helps the CNN to combat overfitting—e.g. memorizing specific orientations of the voxelized PES—and focus on the underlying patterns.

In this work, four types of geometric transformations are applied (including the identity one), as shown in Figure 2.2. At each training iteration, the samples in the batch undergo one of these transformations, with all transformations having the same probability to be applied. For instance, at one training iteration, the voxelized PES can be rotated 90° around the x-axis, while at another iteration, it might be flipped along the z-axis. Rotation is performed either clockwise or counterclockwise, around one the three axes. The voxelized PES can also be viewed as a stack of 2D slices. In this view, reflection corresponds to transposing each slice, whereas flip reversed the order of the slices. Reflection takes place along one of the xy, xz, yz planes, whereas flip is performed along one of the three axes. Figure 2.3 illustrates the performance difference when the CNN is trained on the MOFs dataset with and

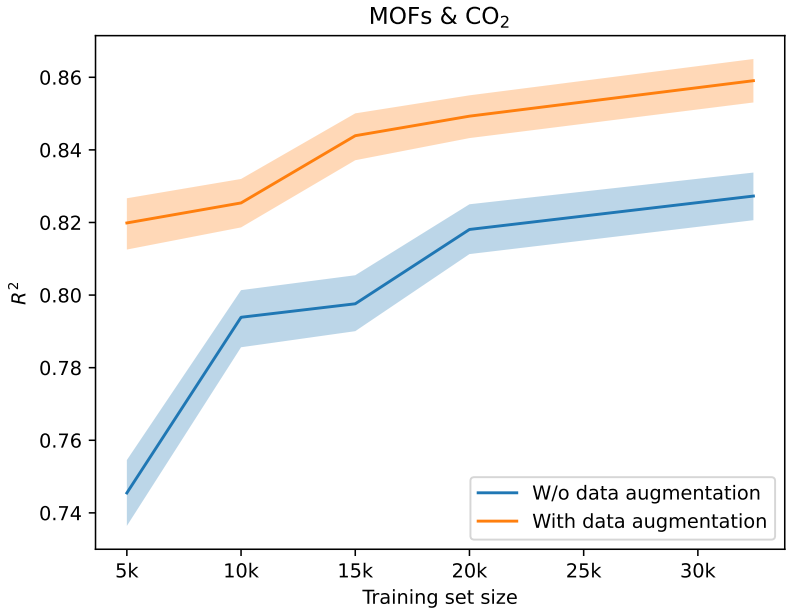


Figure 2.3: CNN performance (R^2 score) on test set with and without data augmentation. Shaded areas correspond to the 95 % CI.

without data augmentation, for training set sizes:

$$\{5000, 10\,000, 15\,000, 20\,000, 32\,432\} \tag{2.7}$$

Chapter 3

Results & Discussion

THE PROPOSED FRAMEWORK is tested on the UO database, for predicting CO_2 uptake in MOFs, the gas that mainly "triggered" the development of energy-based descriptors. In order to evaluate the transferability of the approach, a different host-guest system is also examined. We apply the suggested approach in the database created by Mercado et al. (2018), for predicting CH_4 uptake in COFs. In both cases, the resulting ML models are compared with conventional ones, built upon geometric descriptors. In the rest of this chapter, results from these comparisons are presented, followed by discussion for improvements of the proposed framework. Before delving into the results, we first take a look at RetNet, the 3D CNN under the hood, that takes as input a voxelized PES and outputs a prediction for a gas adsorption property, hereon gas uptake.

3.1 Visualizing RetNet

Figure 3.1 illustrates the processing a voxelized PES undergoes, as it is passing through RetNet. For the purpose of this visualization, we use the model trained on the MOFs dataset with the largest training set size (see Section 2.1). Moreover, for the ease of visualization, only some feature maps of RetNet are visualized. Please note, that each feature map of a given layer, combines all the feature maps of the precedent layer. The only exception are the pooling layers, which just downsample the feature maps from the previous layers.

For example, each feature map of the Conv2 layer takes into account all the twelve feature maps of Conv1 layer. In contrast, the feature maps of the MaxPool1 layer, are just downsampled versions of the corresponding feature maps in Conv2 layer. Although feature maps of CNNs are not meant to be interpreted by humans—especially the ones found deeper in the network—it is worth noticing that early Conv layers (i.e. Conv1 and Conv2) emphasize the texture of the structure. For instance, the third feature map of Conv1 layer delineates the skeleton of the framework.

Moving towards the output layer, the alternation of MaxPool and Conv layers continues until the Flatten layer, which just flattens out and concatenates[1] all feature maps from Conv2 layer into a single vector of size 3240. This vector is then processed by a fully connected neural network (FCNN)—i.e. the stack of Dense and Output layers—to give the final prediction. Since the Output layer is really nothing more than a linear layer, all that RetNet does is the following:

[1] Given m feature maps of size $n \times n \times n$, a Flatten layer converts them into a vector of size mn^3.

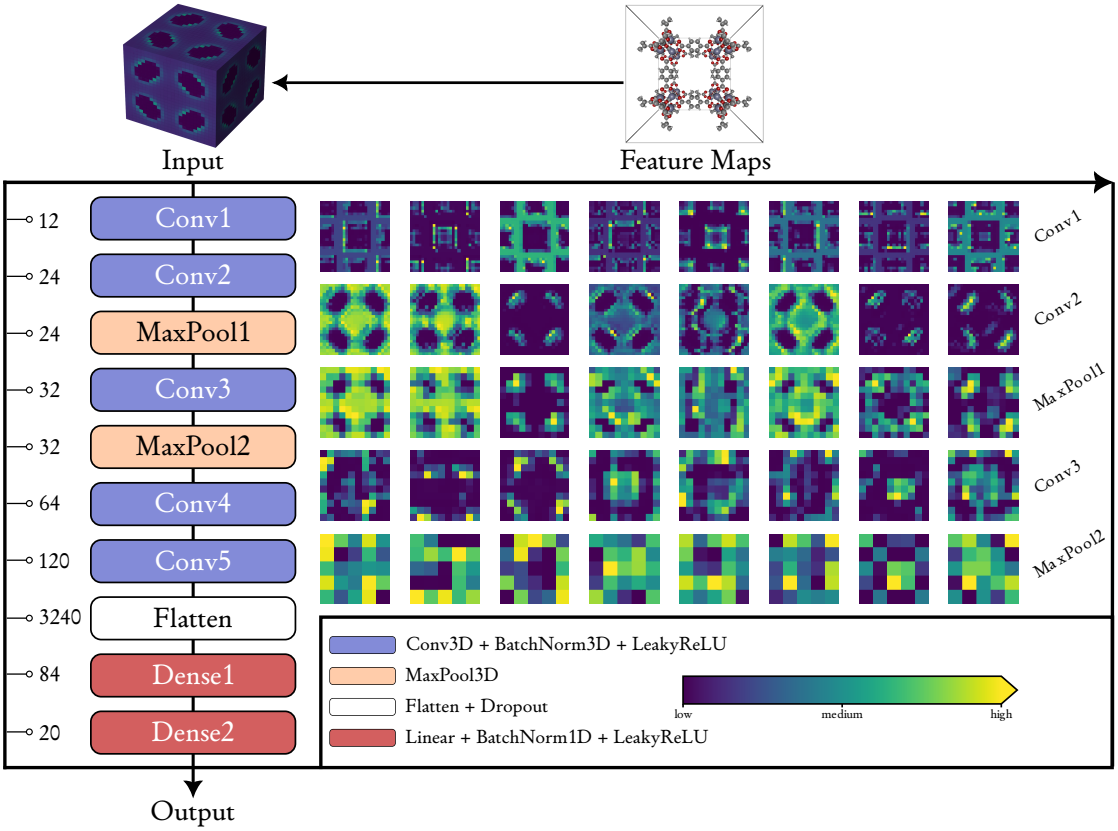


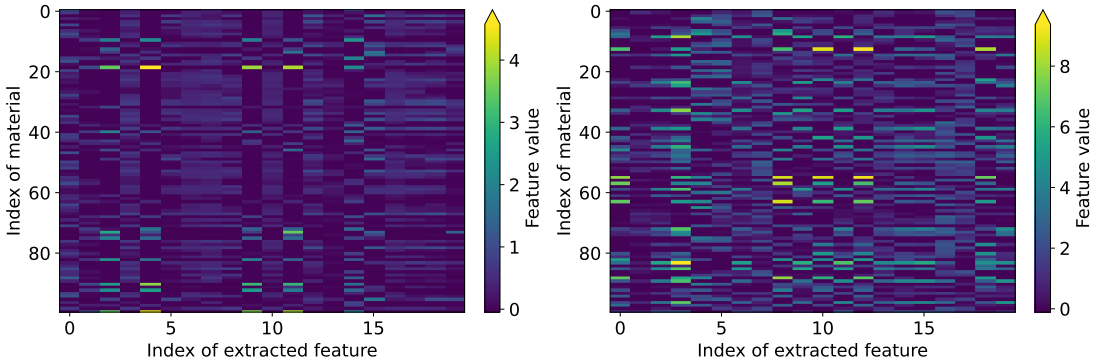
Figure 3.1: Forward pass of IRMOF-1 through RetNet. For the sake of visualization, only slices (feature maps are 3D matrices) of eight feature maps from the first five layers are visualized. For Conv1 layer, the fifth slice is presented, while for the remaining layers, the first slice is presented. The IRMOF-1 structure was visualized with the iRASPA software (Dubbeldam et al. 2018).

$$\underbrace{\boldsymbol{x}}_{\substack{\text{PES}\\ \text{input}}} \longrightarrow \underbrace{\phi(\boldsymbol{x};\boldsymbol{\theta})}_{\substack{\text{fingerprint}\\ \text{feature extraction}}} \longrightarrow \overbrace{\boldsymbol{\beta}^\top \phi(\boldsymbol{x};\boldsymbol{\theta}) + \beta_0}^{\substack{\text{gas uptake}\\ \text{output}}} \tag{3.1}$$

Equation 3.1 says that RetNet, *starting from the PES, it extracts a fingerprint—that is, a high level representation of the PES—and then predicts the gas uptake by using a linear model on top of this fingerprint*. All intermediate layers between Input and Output layer

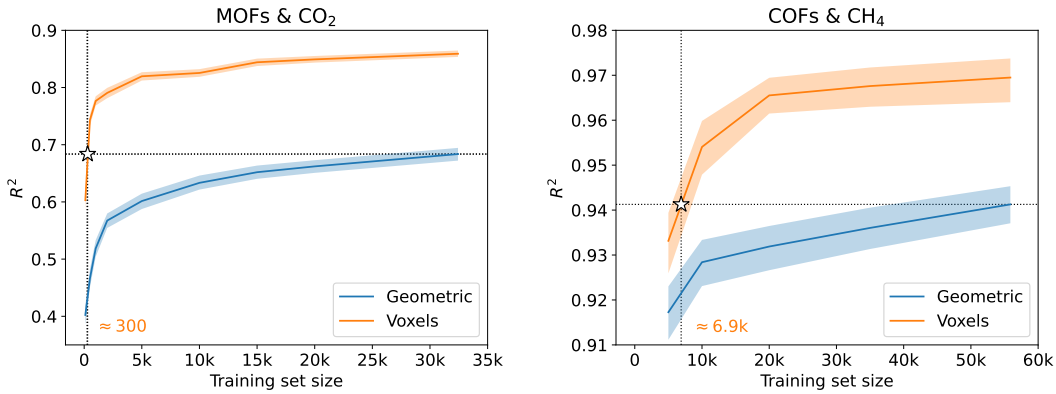
3.2 Learning Curves

For the performance boost, please see Section 2.3.3



(A) Fingerprints extracted from the MOFs dataset. (B) Fingerprints extracted from the COFs dataset.

Figure 3.2: Output of the last LeakyReLU layer of RetNet trained on MOFs (left) and COFs (right) datasets, with the corresponding maximum training set size. The fingerprints of the first 100 materials in the training set are depicted.



(A) Learning curves for MOFs & CO_2. (B) Learning curves for COFs & CH_4.

Figure 3.3: Performance (R^2 score) on test set as function of the training set size for conventional and CNN models. Shaded areas correspond to the 95 % CI. The x-coordinate of the white star denotes the trainig set size where the CNN model reaches the performance of the conventional one, the y-coordinate. "Geometric" stands for geometric descriptors, while "Voxels" stands for energy voxels.

3.3 Discussion

Index

INDEX

Acronyms

ANG adsorbed natural gas.

CNG compressed natural gas.

COF covalent organic framework.

CSD Cambridge Structural Database.

DAC direct air capture.

LNG liquefied natural gas.

MAE mean absolute error.

ML machine learning.

MOF metal-organic framework.