Master Thesis

From Potential Energy Surface to Gas Adsorption via Deep Learning

Antonios P. Sarikas

chempp1160@edu.chemistry.uoc.gr

Supervised by
George E. Froudakis

Materials Modeling
&
Design Group

Department of Chemistry

University of Crete

METAL-ORGANIC FRAMEWORKS, or in short MOFs, thanks to their ultra high porosity and surface area, are deemed as prominent candidates for applications involving gas adsorption. However, their intrinsic combinatorial nature translates to a practically infinite material space, rendering the identification of novel materials with traditional methods cumbersome. Over the last years, machine learning approaches based on predictive models have been developed, allowing researchers to rapidly screen large databases of MOFs. The quality of these models is highly dependent on the mathematical representation of a material, thus necessitating the use of informative inputs. In this thesis, we propose a generalized framework for predicting gas adsorption properties, using as one and only input the potential energy surface. We treat the latter as a 3D energy image and then pass it through 3D convolutional neural network, known for its ability to process image-like data. The proposed pipeline is applied in MOFs for predicting CO_2 uptake. The resulting model outperforms both in terms of accuracy and data efficiency a conventional one built upon textual properties. Additionally, we demonstrate the transferability of the approach to other host-guest systems, by examining CH_4 uptake in covalent organic frameworks. The performance and generality of the proposed approach along with the fast input calculation thanks to parallelization, render it suitable for large scale screening. Finally, discussion for improving and extending the suggested scheme is provided.

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 (Trikalitis 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 (S. Ma et al. 2007; Spanopoulos et al. 2016; Tsangarakis et al.

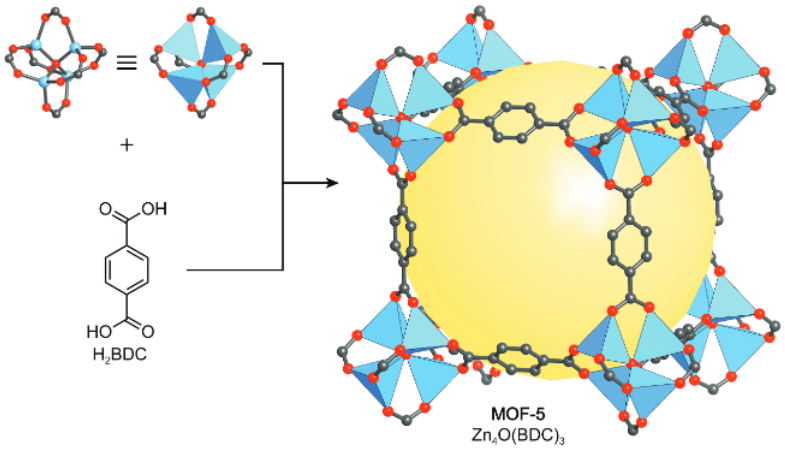


Figure 1.1: Unit cell structure of MOF-5 or IRMOF-1, one of the highest surface area to volume ratio among MOFs, at 2200 m^2 cm^3 and the first MOF studied for hydrogen storage (Rosi et al. 2003).

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 1.2). 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

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

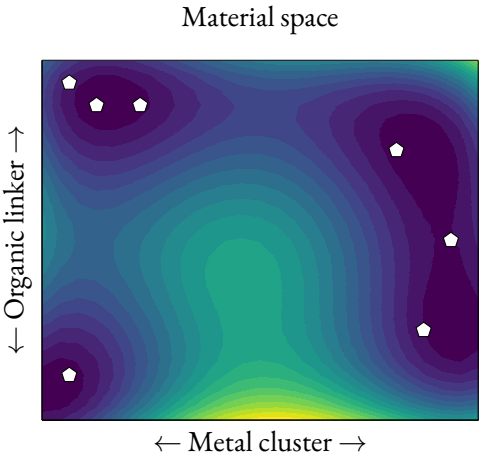


Figure 1.2: Material space of MOFs. Each point in this space corresponds to a unique combination of organic linker and metal cluster, whereas the associated color denotes the "score" of material (point) for a given application. Finding the best material (the pentagons) for a given application, amounts to solving a (probably) non-convex optimization problem.

as the principal tool for the discovery of high-performing MOFs (Simon, J. 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 the screening of large databases. 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.

1.3 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.

[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 2.1.1 for other types of problem tackled by ML.

With regards to gas adsorption in MOFs, one of the first and most commonly used descriptors, are the so called *geometric* ones, which aim to 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, Bobbitt, 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 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, Bobbitt, 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 g L^{-1} and 12.9 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 achieved impressive performance, showing an R^2 (see Section 3.3 for a definition) 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.

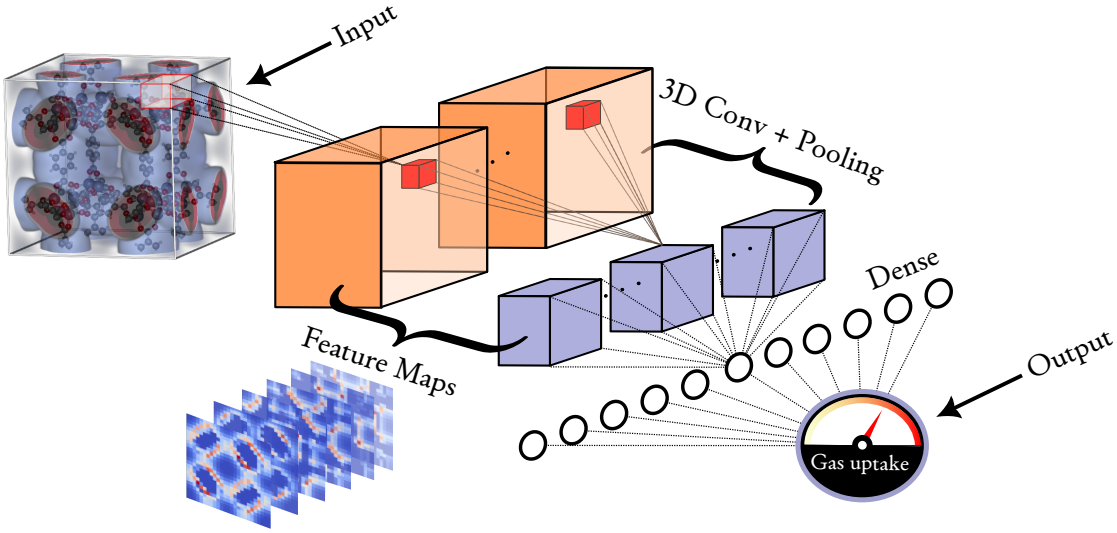


Figure 1.3: Proposed scheme to predict gas adsorption properties, starting from the PES as raw input. A 3D CNN extracts its features from the PES, and then uses them to predict the adsorption property of interest. The IRMOF-1 structure and PES were visualized with the iRASPA software (Dubbeldam et al. 2018).

1.4 Thesis Statement

In the aforedescribed works, 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 its ability to process image-like data. The proposed scheme is schematically presented* in Figure 1.3.

Chapter 2

Theoretical Background

DEEP LEARNING, *a class of ML algorithms based on neural networks (NNs)*, has revolutionized the way we tackle a problem from a ML perspective and is one if not the most important factor for recent ML achievements. Solving complex tasks such image classification or language translation, that for years have bedevilled traditional ML algorithms, constitutes the signature of deep learning (DL). Admittedly, *the advent of a deep CNN*, the AlexNet (Krizhevsky et al. 2012) on September 30 of 2012, signified the "modern birthday" of this field. On this day, AlexNet not only won the ImageNet (Deng et al. 2009) Large Scale Visual Recognition Challenge (ILSVRC), but dominated it, achieving a top-5 accuracy of 85 %, surpassing the runner-up which achieved a top-5 accuracy of 75 %. AlexNet showed that NNs are not merely a pipe-dream, but they can be applied in real-world problems. It is worth to notice that ideas of NNs trace back to 1943, but it was until recently that these ideas got materialized. The reason for this recent breakthrough of DL (and ML) is twofold. First, the availability of large datasets—the era of big data—such as ImageNet. Second, the increase in computational power, mainly of GPUs for DL, accelerating the training of deep NNs and traditional ML algorithms.

2.1 Machine Learning Preliminaries

Since DL is a subfield of ML, it is necessary to familiarize with the later before diving into the former. In this section, the necessary theoretical background and jargon of ML is presented. Machine learning can be defined as *"the science and (art) of programming computers so they can learn from the data"* (Géron 2017). A more technical definition is the following:

Definition 2.1 (Machine learning, Mitchell 1997)**.** *A computer program is said to learn from experience **E** with respect to some class of tasks **T** and some performance measure **P**, if its performance on **T**, as measured by **P**, improves with experience **E**.*

For instance, a computer program that classifies emails into spam and non-spam (the task **T**), can improve its accuracy, i.e. the percentage of correctly classified emails (the performance **P**), through examples of spam and non-spam emails (the experience **E**). But in order to take advantage of the experience aka *data*, it must be written in such a way that *adapts to the patterns in the data*. Certainly, a *traditional spam filter can not learn from experience*, since the latter does not affect the classification rules of the former and as such, its performance. For a traditional spam filter to adapt to new patterns and perform better, it must change its hard-wired rules, but by then it will be a different program. In contrast, a *ML-based filter can adapt to new patterns, simply because it has been programmed to do so*. In other words, *in traditional programming we write rules for solving **T** whereas **in ML we write rules***

to learn the rules *for solving* ***T***. This subtle but essential difference is what gives ML algorithms the ability to take advantage of the data.

2.1.1 Learning paradigms

Depending on the type of experience they are allowed to have during their *training phase* (Goodfellow et al. 2016), ML approaches are divided into three main *learning paradigms*: ***unsupervised learning***, ***supervised learning*** *and* ***reinforcement learning***. The following definitions are not by any means formal, but merely serve as an intuitive description of the different paradigms.

Definition 2.2 (Unsupervised learning)**.** *The experience comes in the form* $\mathcal{D}_{train} = \{\boldsymbol{x}_i\}$*, where* $\boldsymbol{x}_i \sim p(\boldsymbol{x})$ *is the input of the* i*-th training instance aka sample. In this paradigm we are interested in learning useful properties of the underlying structure captured by* $p(\boldsymbol{x})$ *or* $p(\boldsymbol{x})$ *itself.*

For example, suppose we are interested in generating images that look like Picasso paintings. In this case, the input is just the pixel values, i.e. $\boldsymbol{x} \in \mathbb{R}^{W \times H \times 3}$. The latter follow a distribution $p(\boldsymbol{x})$, so all we have to do is to train an unsupervised learning algorithm with many Picasso paintings to get a *model*, that is $\hat{p}(\boldsymbol{x})$. Assuming the estimation of the original distribution is good enough, new realistically looking paintings (with respect to original Picasso paintings) can be "drawn" by just sampling from $\hat{p}(\boldsymbol{x})$. In the ML parlance, this task is known as *generative modeling* while inputs are also called *features, predictors or descriptors*.

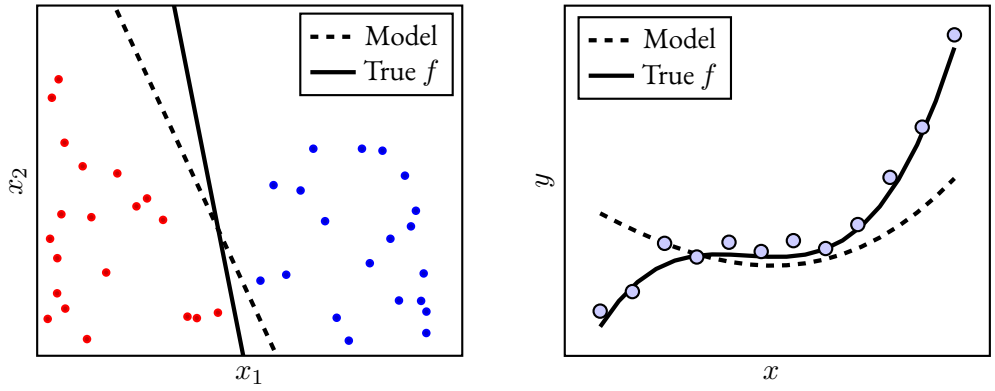
Definition 2.3 (Supervised learning)**.** *The experience comes in the form* $\mathcal{D}_{train} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$*, where* $(\boldsymbol{x}_i, \boldsymbol{y}_i) \sim p(\boldsymbol{x}, \boldsymbol{y})$ *and* $\boldsymbol{y}_i$ *is the output aka label of the* i*-th training instance. In this paradigm we are usually interested in learning a function* $f : \mathcal{X} \to \mathcal{Y}$*.*

This paradigm comes mainly under two flavors: *regression and classification*, which are schematically depicted in Figure 2.1. In regression the interest is in predicting a continuous value given an input, i.e. $y \in \mathbb{R}$, such as a molecular property given a mathematical representation of a molecule. In classification, the interest is to predict in which of k classes an input belongs to, i.e. $y \in \{1, \ldots, k\}$, such as predicting the breed of a dog image given the raw pixel values of the image. The term "supervised" is coined due to the "human supervision" the algorithm receives during its training phase, through the presence of the correct answer (the label) in the experience. In a sense, in this paradigm we "teach" the learning algorithm aka *learner*. It should be emphasized that the label is not constrained to be single-valued, but can also be multi-valued. In this case, one talks about *multi-label regression or classification* (Read et al. 2009).

A more exotic form of supervised learning is *conditional generative modelling*, where the interest is in estimating $p(\boldsymbol{x} \mid \boldsymbol{y})$. For example, one may want to build a model that generates images of a specific category or a *model that designs molecules/materials with tailored properties* (K. Kim et al. 2018; Yao et al. 2021; Gebauer et al. 2022). This is one approach of how ML can tackle the *inverse design problem* in chemistry.

Definition 2.4 (Reinforcement learning)**.** *The experience comes from the interaction of the learner, called agent in this context, with its environment. In other words, there is a feedback loop between the learner and its environment. In this paradigm we are interested in building an agent that can take suitable actions in a given situation.*

The agent observes its *environment*, selects and performs *actions* and gets *rewards or penalties* in return. The goal is to learn an optimal strategy, called a *policy*, that *maximizes the long-term reward*



(A) Classification. (B) Regression.

FIGURE 2.1: Main tasks of supervised learning.

(Géron 2017). A policy simply defines the action that the agent should choose in a given situation. In contrast to supervised learning, where the correct answers are provided to the learner, *in reinforcement learning the learner must find the optimal answers by trial and error* (Bishop 2007). Reinforcement learning techniques find application in fields such as gaming (AlphaGo is a well known example), robotics, autonomous driving and recently chemistry (H. Li et al. 2018; Gow et al. 2022). Since in the present thesis only supervised learning techniques were employed, the remaining of this chapter is devoted to this learning paradigm.

2.1.2 Formulating the problem of supervised learning

The general setting of supervised learning is as follows: we assume that there is some relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon} \tag{2.1}$$

and we want to estimate f from the data. The function f represents the *systematic information* that $\boldsymbol{x}$ gives about $\boldsymbol{y}$ while $\boldsymbol{\epsilon}$ is a random *error term* independent of $\boldsymbol{x}$ and with zero mean. More formally, we have an input space X, an output space Y and we are interested in learning a function $\hat{h}\colon \mathcal{X} \to \mathcal{Y}$, called the *hypothesis*, which produces an output $\boldsymbol{y} \in \mathcal{Y}$ given an input $\boldsymbol{x} \in \mathcal{X}$. At our disposal we have a collection of input-output pairs $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, forming the ***training set*** denoted as $\mathcal{D}_{\text{train}}$, with the pairs drawn i.i.d from $p(\boldsymbol{x}, \boldsymbol{y})$.

Ideally, we would like to learn a hypothesis that minimizes the ***generalization error or loss***:

$$\mathcal{L} \coloneqq \int_{\mathcal{X}\times\mathcal{Y}} \ell(h(\boldsymbol{x}), \boldsymbol{y}) p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} \tag{2.2}$$

that is, the expected value of some *loss function* ℓ over all possible input-output pairs. A loss function just measures the discrepancy of the prediction $h(\boldsymbol{x}) = \hat{\boldsymbol{y}}$ from the true value $\boldsymbol{y}$ and as such, the best hypothesis is the one that minimizes this integral. Obviously, it is impossible to evaluate the integral in EQUATION 2.2, since we don't have access to infinite data.

The idea is to use the *training error or loss*:

$$\mathcal{L}_{\text{train}} \coloneqq \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{i \in \mathcal{D}_{\text{train}}} \ell(h(\boldsymbol{x}_i), \boldsymbol{y}_i) \tag{2.3}$$

as an approximation for the generalization loss, and *choose the hypothesis that minimizes the training loss*, a principle known as *empirical risk minimization*. In other words, to get a hypothesis aka *model* $\mathcal{M}$ from the data, we need to solve the following optimization problem:

$$\hat{h} \leftarrow \arg\min_{h \in \mathcal{H}} \mathcal{L}_{\text{train}} \tag{2.4}$$

which is achieved by just feeding the training data into the learning algorithm $\mathcal{A}$:

$$\mathcal{M} \leftarrow \mathcal{A}(\mathcal{D}_{\text{train}}) \tag{2.5}$$

2.1.3 Components of a learning algorithm

By breaking down Equation 2.4, i.e. the optimization problem the learner needs to solve, the components of a learner can be revealed. The latter is comprised of the following three "orthogonal" components: *a* ***hypothesis space****, a* ***loss function*** *and an* ***optimizer***. We now look into each of them individually and describe the contribution of each one to the solution of the optimization problem. For the ease of notation and clarity, in the remaining of this chapter we will stick to examples from simple (single-valued) regression and binary classification.

Definition 2.5 (Hypothesis space)**.** *The set of hypotheses (functions), denoted as $\mathcal{H}$, from which the learner is allowed to pick the solution of Equation 2.4.*

A simple example of a hypothesis space, is the one used in univariate *linear regression*:

$$\hat{y} = \beta_0 + \beta x \tag{2.6}$$

where $\mathcal{H}$ contains all lines (or hyperplanes in the multivariate case) defined by Equation 2.6. Of course, one can get *a more expressive* hypothesis space, by including polynomial terms, e.g.:

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{2.7}$$

The more expressive the hypothesis space, the larger the ***representational capacity*** of the learning algorithm. For a formal definition of representational capacity, the interested reader can look at *Vapnik-Chervonenkis Dimension* (Hastie et al. 2009).

Definition 2.6 (Loss function)**.** *A function that maps a prediction into a real number, which intuitively represents the quality of a candidate hypothesis.*

For example, a typical loss function used in regression is the *squared loss*:

$$\ell(\hat{y}, y) \coloneqq (\hat{y} - y)^2 \tag{2.8}$$

where $y, \hat{y} \in \mathbb{R}$. A typical loss function for binary classification is the *binary cross entropy loss*:

$$\ell(\hat{y}, y) \coloneqq y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y}) \tag{2.9}$$

where $y \in \{0, 1\}$, indicating the correct class, and $\hat{y} \in [0, 1]$ which corresponds to the predicted probability for class-1. Notice that in both cases the loss is minimum when the prediction is equal to the ground truth. For the cross entropy loss, if $y = 1$ is the correct class, then the model must predict $\hat{y} = 1$ for the loss to be minimized.

Usually, we are not only penalizing a hypothesis for its mispredictions, but also for its *complexity*. This is done in purpose, since a learner with a *very rich hypothesis space* can easily memorize the training set but fail to generalize well to new unseen examples. *Every modification that is made to a learner in order to reduce its generalization loss but not its training loss, is called* ***regularization*** (Goodfellow et al. 2016).

A common—but not the only—way to achieve that, is by including another penalty term called *regularization term or* ***regularizer***, denoted as $\mathcal{R}$, in Equation 2.3:

$$\mathcal{L}_{\text{train}} \coloneqq \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{i \in \mathcal{D}_{\text{train}}} \ell(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_i) + \lambda \mathcal{R} \tag{2.10}$$

The λ factor controls the strength of regularization and it is an ***hyperparameter***, i.e. a parameter that is not learned during training but *whose value is used to control the training phase.* In order to see how λ penalizes model complexity, assume we perform univariate polynomial regression of degree k:

$$\hat{y} = \beta_0 + \sum_{i=1}^{k} \beta_i x^i \tag{2.11}$$

combining mean squared loss (MSL) and *lasso regularization* as training loss:

$$\mathcal{L}_{\text{train}} = \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{i \in \mathcal{D}_{\text{train}}} \ell(\hat{y}_i - y_i)^2 + \lambda \sum_{i=1}^{k} |\beta_i| \tag{2.12}$$

Let's apply a very strong regularization by setting $\lambda \to \infty$ (in practice we set λ to a very large value) and observe what happens to the *weights* β_i. By setting $\lambda \to \infty$, the regularization term dominates the MSL and as such, the only way to minimize the training loss is by setting $\beta_i = 0$. This leave us with a very simple model—only the *bias* β_0 survives—which always predicts the mean value of y in the training set.

Applying a regularizer, is also useful when we need to select between two (or more) competing hypotheses that are equally good. For example, assuming two hypotheses achieve the same (unregularized) training loss, the inclusion of a regularization term help us decide between the two, *by favoring the simplest one*. This is reminiscent of the ***Occam's razor*** *aka* ***principle of parsimony***, which advocates that between two competing theories with equal explanatory power, one should prefer the one with the fewest assumptions.

Definition 2.7 (Optimizer). *An algorithm that searches through $\mathcal{H}$ for the solution of Equation 2.4.*

Having defined the set of candidate models (the hypothesis space) and a measure that quantifies the quality of a given model (the loss function), all that is remaining is a tool to scan the hypothesis space and pick the model that minimizes the training loss (the optimizer). A naive approach is to check all hypotheses in $\mathcal{H}$ and then pick the one that achieves the lowest training loss. This approach can

work if $\mathcal{H}$ is finite, but obviously doesn't scale in the general case where $\mathcal{H}$ is infinite[1]. More efficient approaches are needed if we are aiming to solve EQUATION 2.4 in finite time.

One optimizer that is frequently used in ML and is the precursor of more refined ones, is ***gradient descent***. With this method, the exploration of hypothesis space[2] involves the following steps:

ALGORITHM 1: Gradient descent

```
θ ← random initialization;
while stopping criterion not met do
    θ ← θ − η∇_θ L_train(θ);
end
```

where η is a small number called the *learning rate*. Gradient descent is based on the idea that if a multivariate function is defined and differentiable at a point $\boldsymbol{x}$, then $f(\boldsymbol{x})$ *decreases fastest if one takes a small step from* $\boldsymbol{x}$ *in the direction of negative gradient at* $\boldsymbol{x}$, $-\nabla f(\boldsymbol{x})$.

The motivation becomes clear if we look at the differential of $f(\boldsymbol{x})$ in direction $\boldsymbol{u}$:

$$f(\boldsymbol{x} + \delta\boldsymbol{u}) - f(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) \cdot \delta\boldsymbol{u} \tag{2.13}$$

EQUATION 2.13 says that this differential is minimized[3] when $\delta\boldsymbol{u}$ is anti-parallel to $\nabla f(\boldsymbol{x})$ and that is why we subtract the gradient in ALGORITHM 1, i.e. move in direction anti-parallel to the gradient. The fact that EQUATION 2.13 holds only locally (magnitude of $\delta\boldsymbol{u}$ must be small) explains why η must be a small number. It should be added that gradient descent can be trapped to (potential) local minima of the training loss and therefore fail to solve EQUATION 2.4. As it will be discussed later, this is not a problem, because *the ultimate purpose is to find a hypothesis that generalizes well, not necessarily the one that minimizes the training loss*[4]. Optimizers are discussed in further detail in SECTION **??**.

Before moving on, it is worth to add that both the regularization and the optimizer have an effect on the "true" or ***effective capacity*** (Goodfellow et al. 2016) of the learner, *which might be less than the representational capacity of the hypothesis space*. For example a regularizer penalizes the complexity of an hypothesis, effectively "shrinking" the representational capacity of the hypothesis space. The effect of the optimizer can be understood by looking on its contribution to the solution of EQUATION 2.4. As described previously, the optimizer searches through the hypothesis space. If this "journey" is not long enough, then this "journey" is practically equivalent to a long "journey" in a shortened version of the original hypothesis space. In the rest of this chapter, by the term ***complexity*** *or capacity of a learner, we mean its effective capacity, which is affected by all its three components*.

2.1.4 Performance, complexity and experience

Suppose that we have trained our learner, and finally we get our model, as stated by EQUATION 2.5. *How can we assess its performance?* Remember, we can't calculate the generalization loss, since we are not given an infinite amount of data. First of all, *we should not report the training loss, because it is*

[1] It is not uncommon for $\mathcal{H}$ to be infinite. Even for simple learners like linear regression $\mathcal{H}$ is infinite, since there infinite lines defined by EQUATION 2.6.

[2] We have implicitly assumed that $\mathcal{H}$ can be parameterized, i.e. $\mathcal{H} = \{h(\boldsymbol{x}; \boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta\}$, where Θ denotes the parameter space, the set of all values the parameter $\boldsymbol{\theta}$ can take. This allows us to write the training loss as function of model parameters and optimize them with gradient descent.

[3] The right hand side of EQUATION 2.13 is a dot product.

[4] Remember we use the training loss (see EQUATION 2.3) as a proxy for the generalization loss (see EQUATION 2.2).

optimistically biased, as it is evaluated on the same data that has been trained on[5]. What we should is to collect new input-output pairs, forming the ***test set*** $\mathcal{D}_{\text{test}}$, and then *estimate the generalization loss* as following:

$$\mathcal{L}_{\text{test}} := \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{i \in \mathcal{D}_{\text{test}}} \ell(h(\boldsymbol{x}_i), \boldsymbol{y}_i) \tag{2.14}$$

The *test error or loss* is evaluated on new—unseen to the learner during the training phase—samples and as such, it is an *unbiased estimate* of the generalization loss. Usually, since many times is not even possible to collect new samples, *we split the initial dataset into training and test sets*.

The general recipe for building and evaluating the performance of a ML model has already been presented. What is missing is how we can improve its performance, or to put it differently, the factors that affect the quality of the returned model. There are two main factors that determine the performance of the model: *complexity and experience*. In general, the larger the experience—the training set— the better the performance, just like we humans perform improve on a task by keep practicing. With regards to the complexity, *learners of low complexity might fail to capture the patterns in the data*, meaning that the resulting model will fail to generalize. In contrast, *learners of higher complexity would be able to capture these patterns*, and as such return models of higher quality. However, *as the complexity of the learner keeps increasing, the latter is more sensitive to noise, i.e. there is a higher chance that the learner will simply memorize its experience* and as such, fail to generalize.

In other words, there is a trade-off between the complexity of the learner and its performance. The learner should be not too simple but also not too complex, in order to generalize well. This in turn implies that we need to find a way to "tune" the complexity. A common way to achieve that is by using another set of instances, known as the ***validation set***. We train learners of different complexity on the training set, evaluate their performance on the validation set, and then choose the learner that performs best on the validation set. The reason we use the validation set instead of the test set for tuning complexity, is to ensure that the performance estimation is unbiased. *The test set should not influence our decisions in any way*. After we have tuned the complexity, we can estimate the performance of the resulting model in the test set. Finally, it is a good practice to retrain the learner on the whole dataset—including validation and test sets—since more data result in models of higher quality.

Theorem 1 (Bias-variance decomposition, Bishop 2006)**.** *From Equation 2.1 and under the assumption that* $\epsilon \sim \mathcal{N}(0, 1)$*, the expected squared loss at* $\boldsymbol{x}^*$ *can be decomposed as following:*

$$\mathbb{E}\left[\left(y^* - \hat{f}(\boldsymbol{x}^*)\right)^2\right] = \left(f(\boldsymbol{x}^*) - \mathbb{E}\left[\hat{f}(\boldsymbol{x}^*)\right]\right)^2 + \mathbb{E}\left[\left(\hat{f}(\boldsymbol{x}^*) - \mathbb{E}\left[\hat{f}(\boldsymbol{x}^*)\right]\right)^2\right] + \sigma_\epsilon^2 \tag{2.15}$$

The expected squared loss refers to the average squared loss we would obtain by repeatedly estimating f *using different training sets, each tested at* $\boldsymbol{x}^*$*. The overall expected squared loss can be computed by averaging the left hand side of Equation 2.15 over all possible values* $\boldsymbol{x}^*$*.*

The trade-off between the complexity of the learner and its performance, is mathematically described in Theorem 1. Equation 2.15 states that the error of the learner can be decomposed into three terms: ***bias***, ***variance*** *and* ***irreducible error***. The bias (squared)—first term of Equation

[5] Intuitively, this is like assessing students' performance based on problems they have already seen before. They can easily achieve zero error, just by recalling their memory.

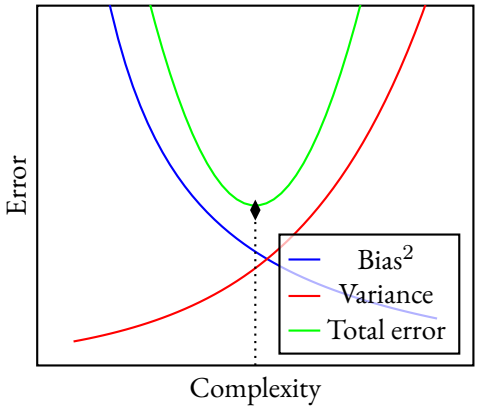


FIGURE 2.2: The bias-variance trade-off. For a given task, there is a "sweetspot" of complexity, that minimizes the total error. Bias^2 and variance correspond to the first and second term of EQUATION 2.15, respectively.

2.15—refers to the error introduced by *approximating a real-world problem, which can be highly complicated, by a much simpler model* (Hastie et al. 2009; James et al. 2014). For instance, if the input-output relationship is highly nonlinearity, using linear regression to approximate f, will undoubtedly introduce some bias in the estimate of f. In contrast, if the input-output relationship is very close to linear, linear regression should be able to produce an accurate estimate of f. In general, more flexible learners, result in less bias (Hastie et al. 2009; James et al. 2014).

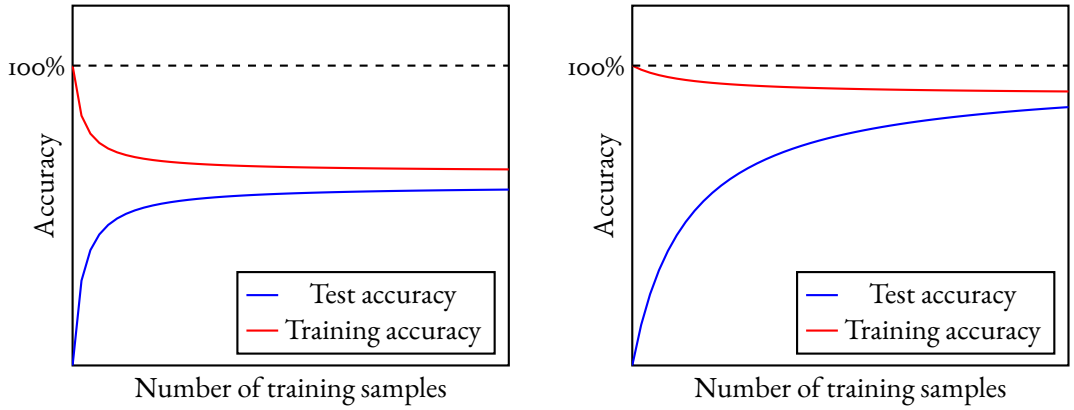
The variance—second term of EQUATION 2.15—refers to the *degree by which $\hat{f}$ would change if it was estimated by different training sets*. Since the training data are used to fit the learning algorithm, different training sets will result in a different estimate of f. Ideally, $\hat{f}$ should not exhibit too much variation between different training sets, since otherwise small changes in the training can result in large changes in $\hat{f}$. In that case, the learner essentially memorizes the training data. Generally, more flexible learners, result in higher variance (Hastie et al. 2009; James et al. 2014).

Lastly, the irreducible error—third term of EQUATION 2.15—refers to the *error caused by stochastic label noise*, as can be seen from EQUATION 2.1. A possible source for this noise, might be ommited features which are useful in predicting the output. It is called irreducible, because no matter how well we estimate f, even if we predict $\hat{y} = f(\boldsymbol{x})$, we can't reduce the error associated to the variability of ϵ. As stated in SECTION 2.1.2, this random error term is indepedent of $\boldsymbol{x}$, and as such, we have no control over it. The *bias-variance trade-off* is schematically depicted in FIGURE 2.2. Interested readers might also appreciate reading about *double descent* (Nakkiran et al. 2019), a phenomenon where increasing further the complexity of the learner, results in a new minimum (hence, the name).

FIGURE 2.3, shows the *learning curves* for learners of different complexity. A learning curve is a plot of the training and test performance[6] of the learner as function of its experience. First, let's look at the learning curve of the low complexity learner. The accuracy[7] starts out high on the training set, since with a small number of samples, the learner can fit them perfectly. However, by adding more training data, learner's training accuracy quickly drops due to learners inflexibility and inexpressivity.

[6] Usually, only the test performance is plotted.

[7] By accuracy we mean any performance metric where higher values are better, not necessarily the classification accuracy.



(A) Learning curve of learner with low complexity.

(B) Learning curve of learner with high complexity.

Figure 2.3: Relation between performance and experience.

That is, it can't fit the patterns in the training data. On the other hand, test accuracy starts out very low since with very few training data, it is unlikely that the training set is a good representation of the underlying distribution $p(\boldsymbol{x}, \boldsymbol{y})$. In other words, it is unlikely that the learner will experience patterns in the training data, that will help it to generalize well. By increasing the training data, test accuracy increases but it never reaches a high value. This happens due to the learners inability to detect and exploit the patterns in the training data. In other words, the learner fails to generalize well not because its experience is low, but because it is biased. That is, it oversimplifies the problem and make strong assumptions that do not capture the complexity of the data.

Now let's consider the learning curve of the high complexity learner. Again, the training accuracy starts out high with a small amount of training data. However, in contrast to the previous case, as the number of training samples increases, the training accuracy remains high since the learner is flexible enough to learn the patterns in the training set, irrespective of its size. At some point, the training data becomes large enough that is a good representation of the underlying distribution $p(\boldsymbol{x}, \boldsymbol{y})$ and since the learner is very flexible, it can capture the true patterns in $p(\boldsymbol{x}, \boldsymbol{y})$, increasing the test accuracy. It is worth pointing out that the learning and complexity curves (see Figure 2.2), are just two slices of the same 3D plot: the plot of performance as function of experience and complexity.

2.2 Fundamentals of Deep Learning

Having covered the basic jargon and concepts of ML, we are now in a position to dive into DL. One might expect that DL is a very complex subfield of ML, given its astonishing results in complex tasks, but quite the opposite holds. Notably, DL shares similar ideas with reticular chemistry: *combining simple computational units, known as* ***neurons****, to achieve intelligent behavior*. And just like we can tune the properties of MOFs by judiciously selecting and combining their building blocks, we can design problem-specific *neural* ***architectures*** by reasonably arranging and connecting the neurons. In other words, both DL and reticular chemistry can be viewed as building with Legos.

Since the term "neuron" is admittedly a neuroscience term, one might wondering what is the re-

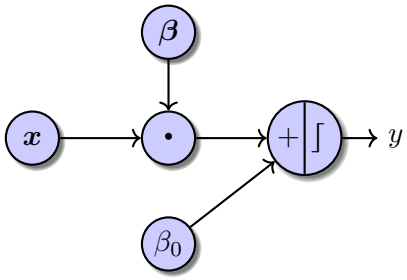


Figure 2.4: The perceptron.

lation between DL and the human brain. The neural perspective on DL is mainly motivated by the following idea: *the brain is a proof by example that intelligent behavior is possible and as such, a straightforward approach to build an intelligent system is by reverse engineering the computational principles behind the brain and duplicating its functionality*. However, the term "deep learning" is not limited to this neuroscientific perspective. It appeals to a more general principle of learning *multiple levels of abstraction*, which is applicable in ML frameworks that are not necessary neurally inspired (Goodfellow et al. 2016).

Definition 2.8 (Deep learning)**.** *Class of machine learning algorithms inspired by brain organization, based on learning multiple levels of representation and abstraction. They achieve great power by learning to represent the world*[8] *as a nested hierarchy of concepts.*

Before exploring NNs we first need to understand how the neuron, the basic building block of NNs, works. *A neuron is nothing more than a device—**a simple computational unit**—that makes decisions by weighing up evidence* (Nielsen 2018). This sounds very similar to the way humans make decisions. For instance, suppose the weekend is coming up and your favorite singer has scheduled a concert near your city. In order to decide whether you should go to the concert or not, you weigh up different factors, such as weather conditions, ease of transportation (you don't own a car) and whether your boyfriend or girlfriend is willing to accompany you. This kind of decision-making can be described mathematically as following:

$$\text{decision} = \begin{cases} 1 & \text{if } \boldsymbol{b}^\top \boldsymbol{x} + \beta_0 > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \text{where} \quad \boldsymbol{b}^\top \boldsymbol{x} := \sum_i \beta_i x_i \tag{2.16}$$

If this weighted sum plus the bias[9]—your willing to to go to the festival irrespective of the evidence—is greater than zero, then your decision is positive, otherwise negative.

The simple decision-making rule specified by Equation 2.16, which is known as the ***perceptron*** (Rosenblatt 1957), is schematically depicted in Figure 2.4. Essentially, the perceptron is a linear binary classifier (see also Figure 2.1a). If we pay a little more attention to Equation 2.16, we can see that that the decision is basically an *application of a linear function*[10] *followed by a nonlinearityity*. As such,

[8] Hierarchy is deeply rooted in our world. Just think the hierarchy from subatomic particles to macroscopic objects.

[9] If you prefer the neuroscientific analogy, you can think of bias as how easy is for a neuron to "fire".

[10] Formally speaking it is an affine function. We can turn it into a linear by "absorbing" the bias term into the weights and adding 1 to the input vector, a procedure known as the *bias trick*.

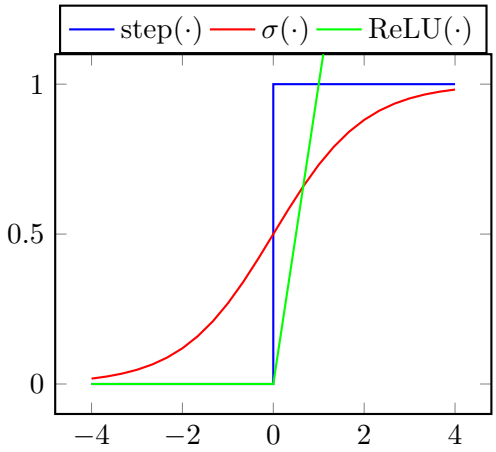


Figure 2.5: Examples of activation functions.

we can rewrite Equation 2.16 as following:

$$y = \phi\left(\boldsymbol{\beta}^\top \boldsymbol{x} + \beta_0\right) \tag{2.17}$$

where $\phi(\cdot)$ is the nonlinear function aka ***activation function***.

In the perceptron, the activation function is the Heavyside step function but in modern NNs it has substituted by functions such as the sigmoid, hyperbolic tangent and currently the rectified linear unit (ReLU) and its variants. Some activation functions are graphically shown in Figure 2.5. The reason that the step function isn't used anymore is that its derivative vanishes everywhere, which is problematic for gradient-based optimization methods that power the training of modern NNs. The ReLU function is defined as:

$$\mathrm{ReLU}(x) \coloneqq \max(0, x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.18}$$

A common variant of ReLU is the leaky rectified linear unit (LeakyReLU) function which is defined as:

$$\mathrm{LeakyReLU}(x) \coloneqq \max(0, x) + a\min(0, x) = \begin{cases} x & \text{if } x > 0 \\ ax & \text{otherwise} \end{cases} \tag{2.19}$$

where a is a small positive constant usually set to 0.01. It is worth to notice how simple the nonlinearities used in NNs are. For instance, ReLU, the most commonly used activation function these days, is just a piecewise linear function. This again highlights the fundamental idea behind DL: *building something complex by combining simple elements*.

2.2.1 Neural networks

Neural networks can be thought as ***collection of neurons*** *organized in layers* and can be represented as *computational graphs*.

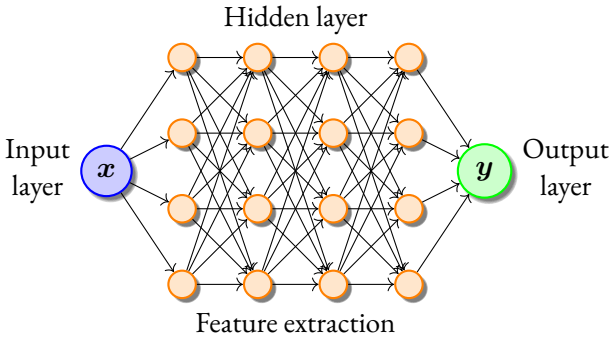
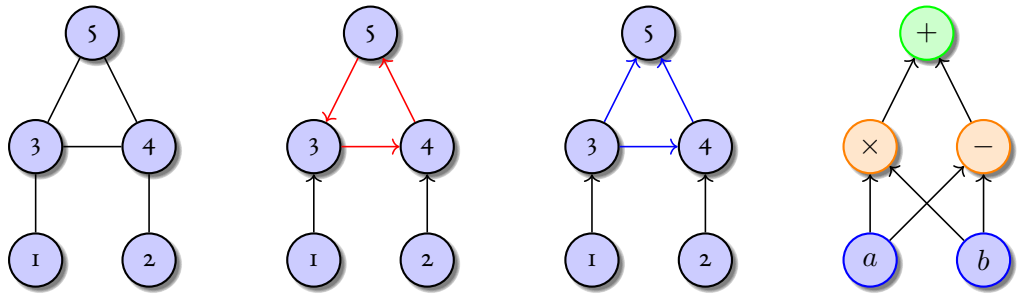


Figure 2.6: The multilayer perceptron. A typical example of a neural network.



(A) Undirected graph. **(B)** Directed cyclic graph. **(C)** Directed acyclic graph. **(D)** Computational graph.

Figure 2.7: Examples of graphs. In a directed graph the edges have direction. If at least one loop is present, they are called cyclic, otherwise acyclic.

Definition 2.9 (Graph). *A set of objects in which some pairs of objects are in some sense "related". See Figure 2.7 for some types of graphs.*

Definition 2.10 (Computational graph). *A directed acyclic graph (DAG) where nodes correspond to operations or variables and edges show the data flow between the nodes.*

In general, the architecture of a NN can be broken down into the following three layers: ***input layer***, ***hidden layer*** *and* ***output layer***. Information flow starts from the input layer, passes through the hidden layer(s) and finally ends at the output layer. Neural networks with more than one hidden layer are classified as deep, and shallow otherwise. A typical architecture known as multilayer perceptron (MLP)[11] or fully connected neural network (FCNN) is presented in Figure 2.6. In should be emphasized that all the neurons in the hidden layers aka *hidden units* of the network perform exactly the same operation as that of the perceptron, described in Equation 2.17. As such all the *hidden units at a given layer make decisions based on the decisions of the previous layer*. Unsurprisingly, the kind of decisions made by the neurons depends solely on the problem and the data distribution at hand.

To understand better the purpose of the hidden layers and the functionality of a NN as a whole, let's consider the problem of image classification. This is by no means a trivial task, since we need to learn

[11] The term MLP is kind of a misnomer, since the step function used originally in the perceptron is no longer used in modern NNs.

a mapping from a set of pixels to an object identity. Imagine for a moment you are blindfolded and you need to classify an image. The valid classes are: "person", "car" and "ship". Furthermore, assume that the correct class is "person". *Would you prefer to hear the sequence of pixel values or whether the image contains a face?* In other words, it is a lot easier to classify the content of an image if we know some *high-level features, i.e. a high-level description of the image*. *Neural networks extract such high-level features by exploiting the hierarchical structure of images*. A complex object like a "face" is defined in term of simpler ones, such as "eye" and "nose", which in turn are defined in terms of simpler ones and so on. This hierarchy allows the NN to solve the complex task of image classification by breaking it down into smaller sub-problems. The first layer learns to detect edges. The second layer combines the decisions of the first layer to detect corners. Subsequently, the third layer combines the decisions of the second layer to identify shapes like circles and squares and so on, until we reach the final layer which is able to detect high-level features such as objects or object parts. The deeper we are into the network—i.e. the closer to the output layer—the more abstract and task-specific the detected features become.

In a FCNN with n hidden layers, each hidden layer performs the following operation:

$$\boldsymbol{h}^t = \boldsymbol{\phi}\left(W^t \boldsymbol{h}^{t-1} + \boldsymbol{\beta}_0\right) \quad \text{where} \quad 1 \leq t \leq n \quad \text{and} \quad \boldsymbol{h}^0 \coloneqq \boldsymbol{x} \tag{2.20}$$

which is just a matrix version of Equation 2.17 with the matrix W^t playing the role of the "synapses" between the neurons of the layers $t-1$ and t. Since the "stacking" of many hidden layers is equivalent to a huge composite function:

$$\boldsymbol{\tau}(\boldsymbol{x}) \coloneqq \left(\boldsymbol{h}^t \circ \boldsymbol{h}^{t-1} \cdots \circ \boldsymbol{h}^1\right)(\boldsymbol{x}) \tag{2.21}$$

and the output layer is just a linear function of the last hidden layer, the output of the FCNN can be written as:

$$\hat{y} = \boldsymbol{\beta}^\top \boldsymbol{\tau}(\boldsymbol{x}) + \beta_0 \tag{2.22}$$

or in the general case of multi-valued output:

$$\hat{\boldsymbol{y}} = W \boldsymbol{\tau}(\boldsymbol{x}) + \boldsymbol{\beta}_0 \tag{2.23}$$

In other words, *a linear model on top of the extracted features*. It should be emphasided that Equation 2.23 is not specific to FCNNs, but describes every type of NN used for classification and regression. Moreover, the use of activation function now becomes more clear: *the composition of many linear functions is just another linear function, which implies nonlinearities must be inserted between them, if we aim to learn a nonlinear relationship*. Equation 2.23 can also be understood in the following way: *a problem that is nonlinear—i.e. complex—in the original space, can become linear—i.e. simple—in a transformed space*. Figure 2.8 shows such an example, known as the XOR problem. The solution of Equation 2.23 essentially boils down to finding the right transformation funciton $\boldsymbol{\tau}(\boldsymbol{x})$. Please note that traditional ML algorithms like support vector machines (SVMs), also map the original space to a transformed space. However, they use a *fixed—i.e. not learnable during the training phase—mapping. Deep learning algorithms on the other hand learn this mapping during their training phase, considering both the problem and the data at hand*.

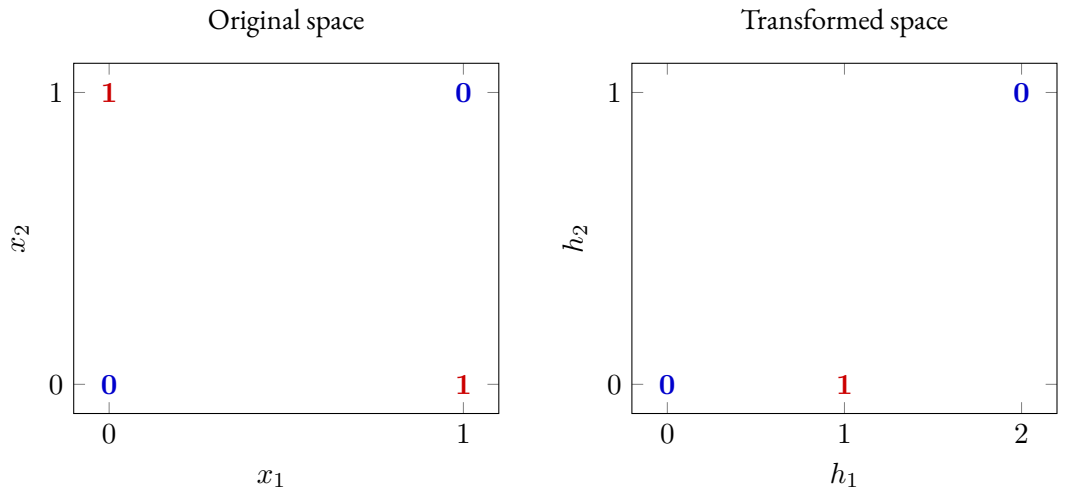


Figure 2.8: Solving the XOR problem. A linear classifier in the original space can't perfectly seperate the "ones" and "zeros". In contrast, if the points are projected into a new space, then they become linearly separable.

Theorem 2 (Universal approximation theorem, Hornik et al. 1989)**.** *A feedforward [12] network with a single hidden layer containing a finite number of neurons can approximate any continuous function.*

One interesting fact about NN is summarized in Theorem 2. An informal proof goes like this. The value of a function f at point x can be viewed as a "spike" or a "bump" at point x with height $f(x)$. If we put a "bump" at every point x we have essentially recovered the function f. The question now boils down to wheter a NN can create "bumps". The answer is affirmative, and a visual proof of the "bump" construction and Theorem 2 is provided by Nielsen 2018.

The universal approximation theorem implies that irrespective of the function we are trying to learn, a network with just one hidden layer and sufficient number of hidden units can represent this function. However, the theorem does not tell us two important things. First, the number of neurons required for the problem we are aiming to solve. Second, *whether we can learn the function at all* (Goodfellow et al. 2016). Learning the true function can fail since the optimizer is not guaranteed to pick the solution that minimizes the generalization loss. Remember it optimizes the training loss as a proxy for the generalization loss.

Theorem 3 (No free lunch theorem, Wolpert 1996)**.** *Averaged over all possible data generating distributions, every learner has the same error rate when predicting previously unobserved points.*

Finally, we close this section with Theorem 3. In essence, it states that *no learner is universally better than any other*. Please note that by "learner" we mean even "dummy" learners, such as random guessing. In other words, the more sophisticated learning algorithm we can conceive has the same

[12] In feedforward newtworks information flows from input to output. In contrast, feedback aka recurrent networks allow the information to travel in both directions by introducing loops. Computations derived from earlier input are fed back into the network, which gives them a kind of memory. This kind of NNs find application in natural language processing (NLP) tasks, such as text generation or classification. Note that although they contain loops, and as such they are not DAGs, we can convert them to DAGs by "unrolling" their computational graph (LeCun et al. 2015).

average performance (over all tasks) as random guessing. While THEOREM 3 seems unintuitive at first glance, it may be easier to understand it with an example. Suppose we see one sheep, and then expect to see another one. We could devise the following strategies (learning algorithms) for predicting the color of the second sheep:

$$\text{strategies} = \begin{Bmatrix} \text{Same color as the first} & \text{(white, white)} \\ \text{Different color than the first} & \text{(black, black)} \\ \text{Always black} & \text{(white, black)} \\ \text{Always white} & \text{(black, white)} \end{Bmatrix} = \text{possible worlds} \tag{2.24}$$

Assuming that all possible worlds (data generating distributions) *are equally likely*, then each strategy has the same expected error: 50%. Fortunately, *in real-world this assumption breaks down*. For example, animals tend to be the same color, so the worlds where the first and the second sheep have different colors are unlikely. In this scenario, guessing the same color as the first is more likely to be correct. Every learning algorithm is equipped with an ***inductive bias***, *that is a set of assumptions about the underlying data distribution*[13]. Whether algorithm $\mathcal{A}_1$ will outperform $\mathcal{A}_2$ on problem $\mathcal{P}$, is just a matter of whose assumptions are better aligned with the structure of $\mathcal{P}$.

2.2.2 Regularizing neural networks

Deep NNs typically have hundreds of thousands, million, or even billion parameters. With such a huge number of parameters, they are capable of memorizing a huge variety of complex training sets. However, memorization is harmful from a generalization point of view. Therefore, it is necessary to employ regularization techniques when training deep NNs, especially when the size of the training set is relatively small. Besides adding penalty terms to the training loss as described in SECTION 2.1.3, common regularization techniques include: ***data augmentation*** *and* ***dropout***.

Data augmentation, as the name implies, is a technique to artificially increase the size of the data set. With this technique, each training instance is replicated many times with *each replicate being randomly distorted in such a way that the corresponding label is left unchanged*[14]. For example, in image classification the original images can undergo *geometric transformations* such as rotations, vertical or horizontal flips, small shifts and random crops. Such kind of transformations force the NN to be more tolerant to variations in orientation, position and size. Another way to augment the original images is by applying *color transformations*, such as changing the brightness or contrast of the image, increasing the NN's tolerance in different lighting conditions. Of course, we can further augment our training set by composing geometric and color transformations. The performance boost thanks to data augmentation can be understood in two ways: i). More data is better. ii). As introducing a useful inductive bias. That is, *we know a priori that the true function ought to be invariant in certain transformations, and the augmented images are a way of imposing this knowledge*.

[13] For example, the *composition of layers* in NNs provides a type of relational inductive bias: *hierarchical processing*. Another example of inductive bias is the linear relationship assumed in linear regression.

[14] We should be careful on how we augment the training set. For example, in character classification tasks there is difference between the characters "b" ↔ "d" and "6" ↔ "9". As such, horizontal flips and 180° rotations are not advisable for this task.

Dropout (Hinton et al. 2012; Srivastava et al. 2014) is a powerful and computational inexpensive method to regularize neural networks, which has proven to be extremely successful. Even the state-of-the-art architectures improved by 1–2 % when dropout was added to them. This may not sound like a lot, but if we consider a model that has already 95 % accuracy, a performance boost of 2 % amounts to 40 % drop in the error rate (going from 5 % to 3 %). Let's see how it works: at every training step the neurons of a hidden layer can be temporarily "dropped out" with probability p aka *dropout rate*, meaning that they will entirely ignored during this training step, but may become active again in the next training step. And that's it except for one technical detail. Dropout is applied only during training and as such, all neurons are active during testing. This means that during this phase a neuron will receive a different amount of input signal compared to its (average) input signal during training. For example, if the dropout rate is set to 0.5, then during testing a neuron will be connected to twice as many input neurons as it was during training (on average). To compensate for that, the neuron's input connections must be multiplied by 0.5 during testing. Otherwise, each neuron in the network will receive a total input signal roughly twice as large as what it was trained on, meaning that each neuron and the network as whole won't perform well. In general, during testing we need to multiply each input connection by the *keep probability* $(1 - p)$. Another way[15] to retain the same amount of input signal for both training and testing phases, is by just dividing each neuron's output with the keep probability, a technique known as *inverted dropout*.

It is quite surprising that this random "resurrection" of neurons improves the performance of the network. Dropout works because it *forces neurons to pay attention to all of their inputs, rather than relying exclusively on just a few of them, making them robust to the loss of any individual piece of evidence*. Or to put it differently, it strengthens them by forcing them to "live" in a stochastic handicap environment. Another way to understand dropout is the following. When different sets of neurons are "dropped out", it is like we are training different NNs. That is, the dropout procedure acts as an average of these different networks. The latter will overfit in different ways and so the net effect of dropout will hopefully mitigate this effect.

2.2.3 Convolutional neural networks

Convolutional neural networks are specialized NN architectures to process image-like data and in general, data which exhibit spatio-temporal relationships. They find application in tasks such as text, audio, video and image classification, semantic segmentation and object detection, just to name a few. As their name implies, CNNs make use of the ***convolution*** operation, so to understand the former we first need to understand the latter.

The convolution between the functions f and g, denoted as $f * g$, is defined as following:

$$(f * g)(t) \coloneqq \int_{-\infty}^{\infty} f(\tau)g(t - \tau)d\tau \tag{2.25}$$

which is the integral of the product of f and g when g is reflected about the y-axis and shifted. In CNN terminology the first argument of the convolution operation is referred to as the *input* and the second one as the ***kernel*** *or* ***filter***. What is the interpretation? It is just a *moving inner product between*

[15] Note that these alternative are not perfectly equivalent, but in practice they work equally well.

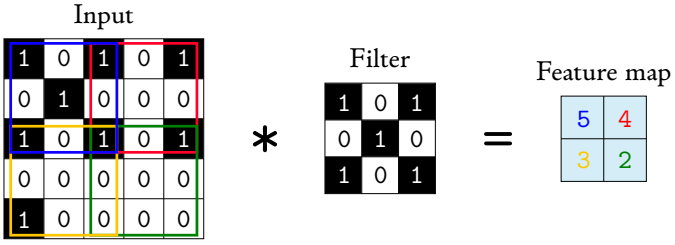


FIGURE 2.9: Convolution operation. Convolving a filter with an image can be seen as template matching. When a local image patch matches the filter—template to be matched—the output in the feature map is highly positive. Sliding of the filter over the image and recording the output of this template-patch similarity, produces the feature map.

the two functions. Recall, that the inner product between two vectors $\boldsymbol{f}$ and $\boldsymbol{g}$ is defined as:

$$\boldsymbol{f} \cdot \boldsymbol{g} := \sum_i f_i g_i \tag{2.26}$$

and can be viewed as *a measure of similarity* between $\boldsymbol{f}$ and $\boldsymbol{g}$. Conceptualizing functions as infinite-dimensional vectors and ignoring the reflection part of convolution[16] let us understand the latter as the inner products between f and shifted versions of g by t. If g represents a pattern—usually local—we are interested in to detect in the signal f, then the output of convolution known as ***feature map*** in the DL jargon, essentially tell us where (hence the term "map") the pattern described by g is located in the signal f.

Usually when we work with data on a computer, the index t will be discretized, meaning that it can take only integer values. Assuming that f and g are defined only for integer t, the discrete convolution is defined as following:

$$(f * g)(t) := \sum_{\tau=-\infty}^{\infty} f(\tau) g(t-\tau) \tag{2.27}$$

In ML applications, the input and the filter are usually multidimensional arrays aka *tensors*. Because these two tensors must be explicitly stored separately, we treat the input and the kernel as functions that are zero everywhere except the finite set of points for which their values are stored (Goodfellow et al. 2016).

Please note that we can and often use convolutions over more than one axis at a time. A typical example is when the input is a 2D image. In that case, the convolution between the image I and the filter K is defined as:

$$(I * K)(i,j) = \sum_m \sum_n I(m,n) K(i-m, j-n) \tag{2.28}$$

and is schematically shown in FIGURE 2.9. The reason that the flipping of the filter is not necessary in ML applications, is simply because the *values of the filter are adapted, i.e. learned, during the training*

[16] As it will be latter discussed, whether the kernel is reflected or not, doesn't make a difference for ML applications.

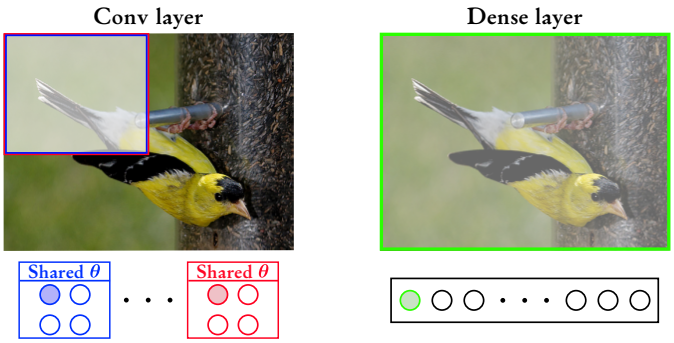


FIGURE 2.10: Neuron's arrangement in convolutional and fully connected aka dense layers. In convolutional layers, neurons are organized into groups and share the same parameters. The receptive field of each member is a small region of the input. In contrast, the receptive field of neurons in a dense layer is the whole input and there is no parameters sharing. Note that the receptive field of neurons increases as we transition to convolutional layers deeper in the network.

phase. Many DL frameworks instead of the convolution implement a related operation called *cross-correlation*, which is the same as convolution but without flipping the filter. We will stick to this convention for the rest of this section.

Now that the convolution operation has been presented, we can appreciate its conbtribution to the mechanics of a CNN. *The convolution operation introduces a beneficial inductive bias* to the network, namely ***sparse connections*** *and* ***parameter sharing***, as shown in FIGURE 2.10, and that's why CNNs outperform FCNNs in object recognition tasks. Sparse connections express the prior knowledge that closely placed pixels are related to each other or to put it differently, *local features such as edges are useful to understand images*. On the other hand, parameter sharing encodes the idea that *a feature detector that is useful in one part of the image is also useful in another part of the image*. For example, we would like to detect a "face" irrespective of its position in the image.

The basic building block of a CNN is the *convolutional layer* which contains many *learnable convolutional filters*, each of which is a template that determines whether a particular local pattern is present in an image. It should be emphasized that a feature map of a given layer combines all the feature maps of the previous layer or just the raw image (in the case of the 1-st convolutional layer). This means that if the layers $t-1$ and t contain n and m feature maps, respectively, then the layer t must learn $m \times n$ filters. By stacking many such layers a CNN extract features hierarchically, with the level of (feature) abstraction increasing the deeper we go into the network.

It is worth to mention that CNNs are essentially regularized FCNNs, meaning that the latter can learn to behave like the former. The catch is that they will probably need a great amount of training data. To understand why this is the case, let's examine the horizontal edge detection problem. The CNN can learn to detect horizontal edges *at any position* by adjusting the values for one of its filters to

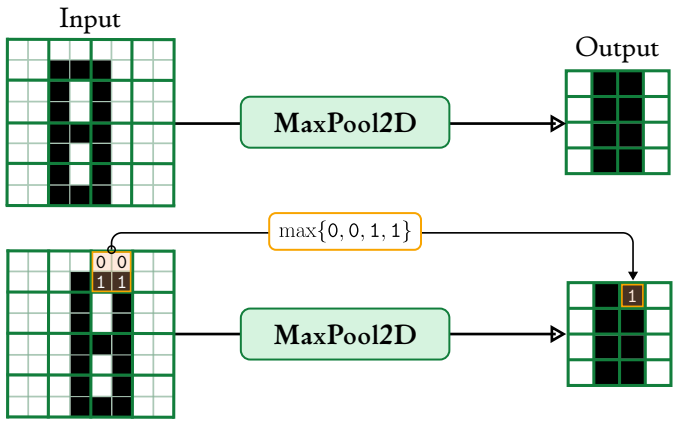


FIGURE 2.11: Max pooling operation. Small translations to the input (input B is just a shifted version of input A by 1 pixel to the right) produce the same output when passed through the max pooling layer, meaning that the latter introduces into the network some level of invariance to small translations.

the following ones[17]:

$$K_x^{\text{edge}} = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} \tag{2.29}$$

What about the FCNN? FIGURE 2.9 gives the answer. A neuron in a dense layer can also learn to detect edges *by just zeroing the weights for its inputs except the small region where the edge needs to be detected*[18]. For this region, it must learn the weights specified by EQUATION 2.29. The problem is that the aforementioned zeroing of weights and the non-zero weights themselves, must be learned by many other neurons for horizontal edges to be detected at different positions, which is of course not guaranteed with limited amount of training data.

Besides convolutional layers, another typical of building block of CNNs are the ***pooling layers***. Their role is to downsample (reduce the resolution) in a parameter-free way the feature maps produced by convolutional layers. By downsampling in this manner, they reduce the memory-computational footprint of the CNN and also the number of parameters, thereby reducing the risk of overfitting (Géron 2017). A pooling layer takes as inputs the feature maps of the preceding convolutional layer and subsamples them by substituting the outputs in a small neighborhood of the feature map with a summary (Goodfellow et al. 2016). FIGURE 2.11 illustrates a common type of pooling, known as ***max pooling***, which uses the max function to compute the summary statistic. Another type of pooling is *average pooling*, which computes the summary statistic through averaging. Just like convolution, the idea of pooling generalizes to more than two dimensions.

[17] Recall that an "edge" is nothing else than a significant local change in the image intensity. Based on the formula of symmetric derivative: $\partial_x f \propto f(x+h) + 0 \cdot f(x) - f(x-h)$. In essence, by convolving an image with the filter specified in EQUATION 2.29, we calculate the gradient along the x-axis in a discretized fashion. Detecting vertical or adges along any direction follows the same idea. You can play with different filters here.

[18] We can view each neuron in a dense layer as performing convolution with a kernel size as large as the input.

Chapter 3

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 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.

3.1 Datasets

3.1.1 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 (see SECTION 2.1.4), 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 4.3a, the training set size was varied and the following training set sizes were considered:

$$\{100, 500, 1000, 2000, 5000, 10\,000, 15\,000, 20\,000, 32\,432\} \tag{3.1}$$

The energy voxels of MOFs are publicly available in figshare.

3.1.2 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 298 K and 5.8 bar was examined. The following five textual properties were used to build the conventional models:

[1] The validation set was used to select the number of epochs, see SECTION 3.3.2.

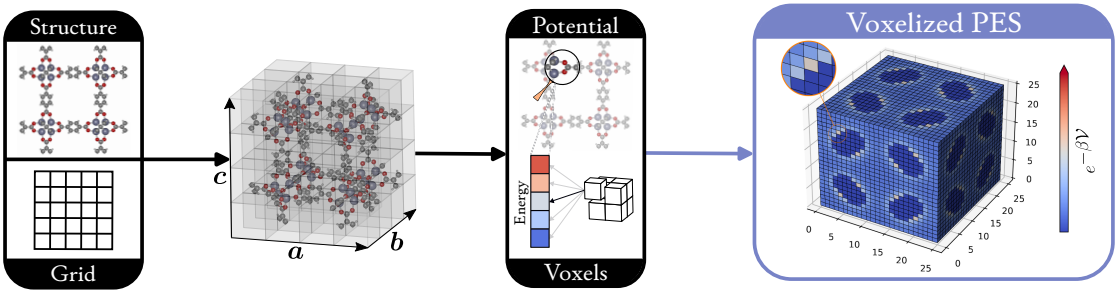


Figure 3.1: Workflow to construct the voxelized PES. The grid size and the type of the potential control the "trade-off" between information content and computational cost. The IRMOF-1 structure was visualized with the iRASPA software (Dubbeldam et al. 2018).

density, gravimetric surface area, void fraction, pore limiting diameter and largest cavity diameter. For producing the learning curves shown in Figure 4.3b, the training set size was varied and the following training set sizes were considered:

$$\{5000, 10\,000, 15\,000, 20\,000, 35\,000, 55\,871\} \tag{3.2}$$

3.2 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 3.1. 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:

$$\mathcal{V}(\boldsymbol{r}_i) = \sum_{\substack{j=1 \\ r_{ij} \leq r_c}}^{N} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{3.3}$$

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:

$$\sigma_{ij} = \frac{\sigma_i + \sigma_j}{2} \quad \wedge \quad \epsilon_{ij} = \sqrt{\epsilon_i + \epsilon_j} \tag{3.4}$$

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 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 and speed through parallelization the calculation of energy voxels. In the remaining of this thesis, the terms "voxelized PES" and "energy voxels", are used interchangeably.

3.3 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 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:

$$R^2 := 1 - \frac{\sum_{i=1}^{N_{\text{test}}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N_{\text{test}}} (y_i - \bar{y})^2} \tag{3.5}$$

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.

3.3.1 CNN architecture

The architecture of the 3D CNN is presented in FIGURE 4.1, 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.

3.3.2 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:

$$\boldsymbol{x}' = \frac{\boldsymbol{x} - \mu_{\text{train}}}{\sigma_{\text{train}}} \tag{3.6}$$

Regarding CNN training, 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 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] The voxelized PES is essentialy a single channel, i.e. grayscale, image.

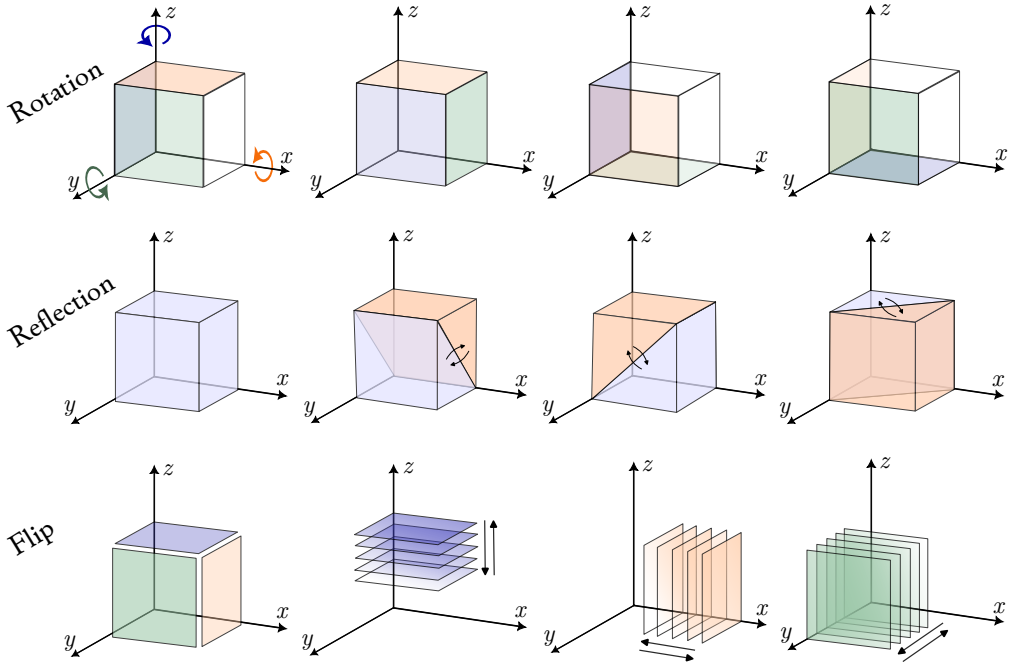


Figure 3.2: Geometric transformations for data augmentation.

3.3.3 Data augmentation

With this technique, the training set is artifically increased, by applying transformations on the input that leave the label unchanged (see also Section 2.2.2). 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 3.2. 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 3.3 illustrates the performance difference when the CNN is trained on the MOFs dataset with and without data augmentation, for training set sizes:

$$\{5000, 10\,000, 15\,000, 20\,000, 32\,432\} \tag{3.7}$$

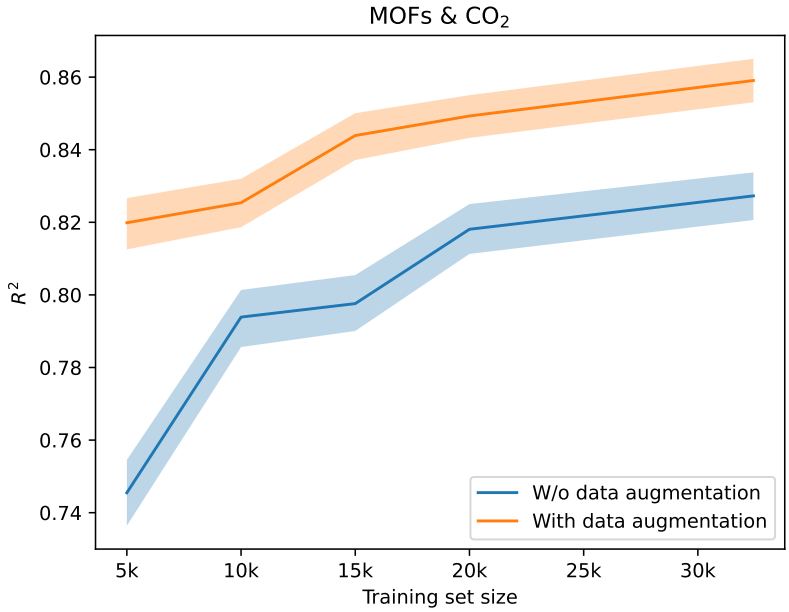


Figure 3.3: CNN performance (R^2 score) on test set with and without data augmentation. Shaded areas correspond to the 95 % CI.

Chapter 4

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.

4.1 Visualizing RetNet

FIGURE 4.1 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 3.1). 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 12 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 3rd 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 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.

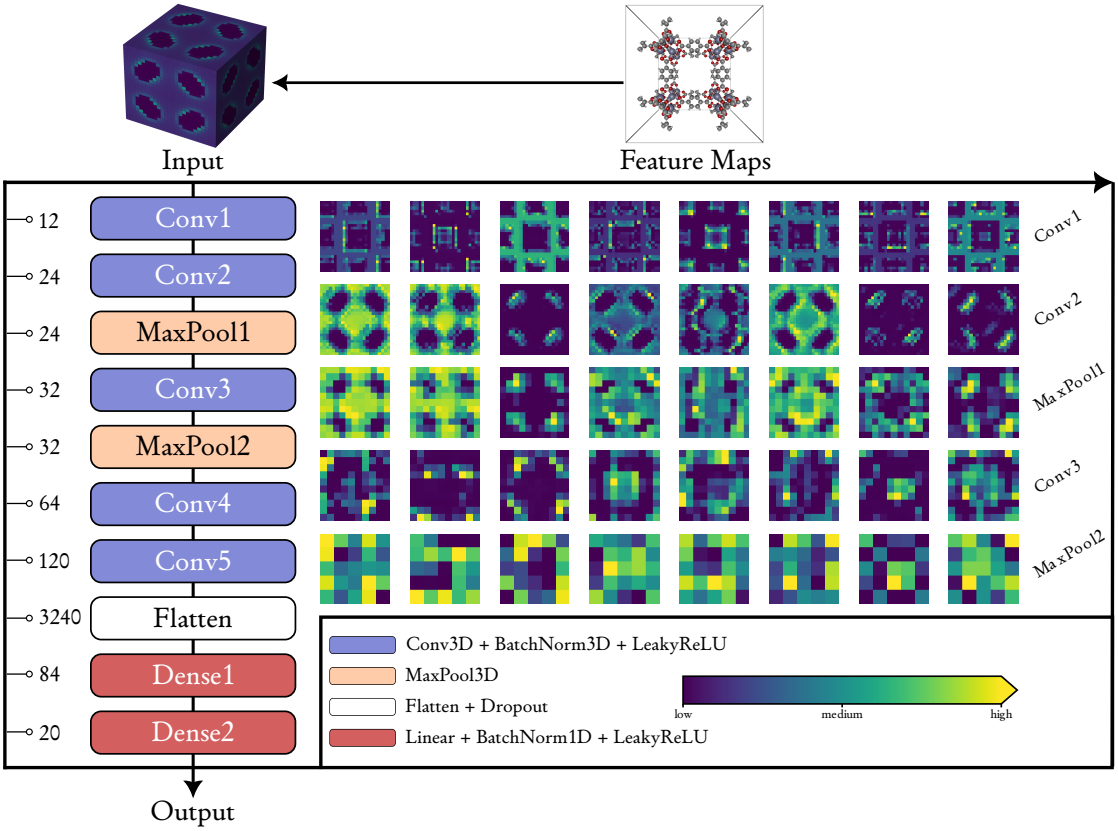
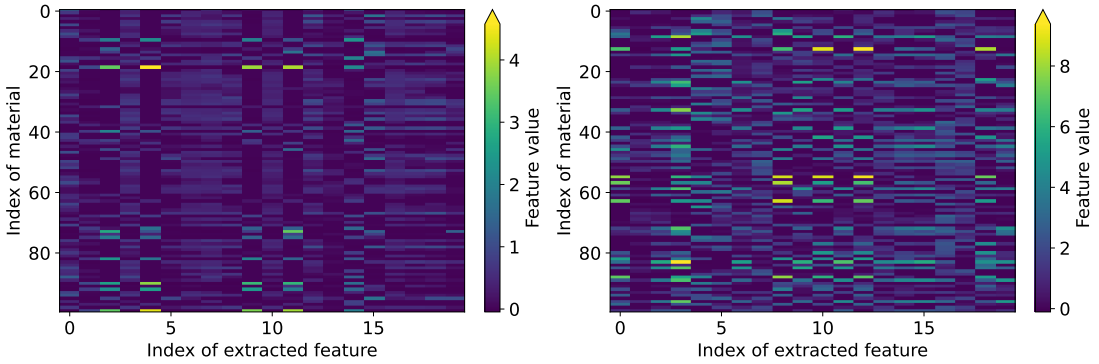


Figure 4.1: Forward pass of IRMOF-1 through RetNet. For the sake of visualization, only slices (feature maps are 3D matrices) of 8 feature maps from the first 5 layers are visualized. For Conv1 layer, the 5-th slice is presented, while for the remaining layers, the 1-st slice is presented. The IRMOF-1 structure was visualized with the iRASPA software (Dubbeldam et al. 2018).

$$\underbrace{\boldsymbol{x}}_{\substack{\text{PES}\\ \text{input}}} \longrightarrow \underbrace{\phi(\boldsymbol{x};\boldsymbol{\theta})}_{\substack{\text{fingerprint}\\ \text{feature extraction}}} \longrightarrow \overbrace{\boldsymbol{\beta}^{\top}\phi(\boldsymbol{x};\boldsymbol{\theta})+\beta_0}^{\substack{\text{gas uptake}\\ \text{output}}} \tag{4.1}$$

Equation 4.1 says that RetNet, *starting from the PES, 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 participate in this feature extraction step, with the Dense2 layer determining the size of the fingerprint, which is a vector of size 20, i.e. $\phi(\boldsymbol{x}) \in \mathbb{R}^{20}$ (see Figure 4.2). The fact that *this fingerprint extraction step is learnable*—the parameters $\boldsymbol{\theta}$ of ϕ are learned during the training phase—*is what fundamentally distinguishes the proposed approach from methods that use hand-crafted fingerprints* (see Section 1.3). *In these methods the fingerprint or extraction step is fixed*, and based on some heuristic, such as energy histograms (Bucior, Bobbitt, et al. 2019) or average interactions (Fanourgakis, Gkagkas, Tylianakis, and Froudakis 2020).



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

FIGURE 4.2: 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.

Hereon, feature extraction from the PES is no longer fixed, but is an essential part of the training phase.

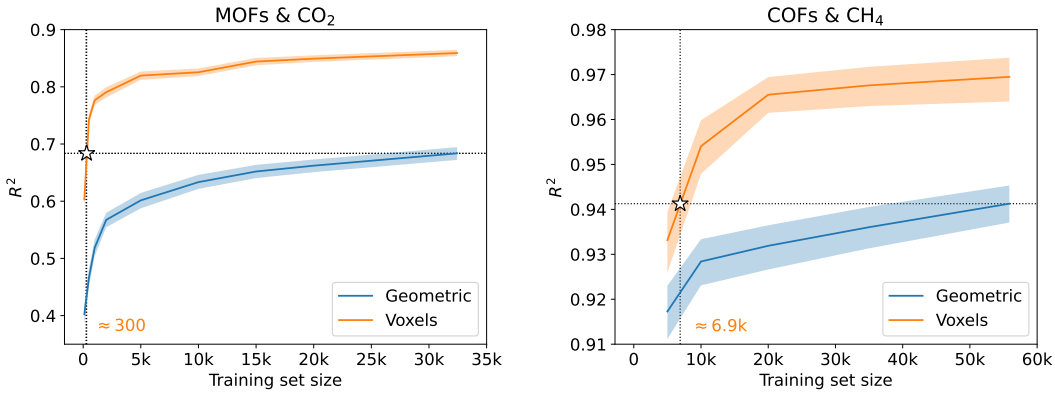
4.2 Learning Curves

The learning curves of the conventional models—built upon geometric descriptors—and the proposed ones–built upon energy voxels—are shown in FIGURE 4.3. As it can be seen from FIGURE 4.3a, in the MOF-CO_2 case, the CNN model achieves an R^2 score of 0.859, outperforming the conventional model, which shows an R^2 score of 0.690. This amounts to around 25 % increase in accuracy, even with such a coarse approximation of the PES[2]. Moreover, from the same figure, one can notice that the proposed model reaches the peak performance of the conventional one—that is, the performance when trained with the maximum training set size—by requiring two orders of magnitude less training data, around 300.

Analogous results are observed when examining the COF-CH_4 case. Again the CNN model performs better, showing an R^2 score 0.969 compared to 0.941 for the conventional one. Similar to the previous case, a substantially smaller amount of training data are required—one order of magnitude less training, around 6900—for the CNN model to match the performance of the conventional model.

The fact that in both cases, the learning curves of the proposed models lie above the corresponding ones of the conventional models, should be credited to the following factors: i). The increased informativeness of the voxelized the voxelized PES—in comparison to geometric descriptors. ii). The ability of CNNs to handle images and image-like data, such as the voxelized PES, which is essentially a single channel 3D image. iii). The data augmentation technique, which was applied during the CNN training (see SECTION 3.3.3).

[2] In this work, all host-guest interactions were modeled with the LJ potential (see SECTION 3.2), which neglects electrostatic interactions.



(A) Learning curves for MOFs & CO_2.

(B) Learning curves for COFs & CH_4.

FIGURE 4.3: 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.

4.3 Discussion

It is worth mentioning the increase in performance, approximately 13 %, of the CNN model in the COFs-CH_4 case ($R^2 = 0.969$) compared to the MOFs-CO_2 case ($R^2 = 0.859$). In contrast to CO_2, which exhibits strong electrostatic interactions with the framework atoms, CH_4 lacks dipole or quadrupole moment. Given that the same resolution—i.e. the same grid size—was used in both cases and that the LJ potential doesn't account for electrostatic interactions, this performance gap should be attributed to the absence of the latter in the voxelized PES. *In other words, the extra "contrast" that such strong interactions add to the energy image of the masterial, is missing from the voxelized PES. As such, a straightforward approach to improve the performance of the proposed approach, especially for adsorbates like CO_2, H_2 and H_2S, is to include this type of interactions into the voxelized PES*. Of course, there is no free lunch, since these refinements require the assignment of partial charges to each framework atom, which is a computationally expensive task. Luckily, ML-based approaches have already been developed (Bleiziffer et al. 2018; Raza et al. 2020; Kancharlapalli et al. 2021), which can assign partial charges rapidly and with high fidelity, enabling the efficient construction of a more accurate voxelized PES.

Improving the input, and as such, the performance of the suggested pipeline is a major concern, but not the only one. *What about the data efficiency of the pipeline?* Imagine that we are asked to predict CH_4 uptake at various thermodynamic conditions. A naive approach would be to collect training data and retrain from scratch the CNN for every thermodynamic condition, which is of course a laborious task. *Can we do something smarter?* Well, the fact that the proposed framework uses a DL algorithm under the hood, opens the door for applying ***transfer learning*** *techniques*. In a nutshell, transfer learning (Zhuang et al. 2019; R. Ma et al. 2020; Kang et al. 2023) is based on the following idea: *a violist can learn to play piano faster than others, since both the piano and the violin are musical instruments, and may share some common knowledge.* Translating this to NNs, a pre-trained NN on an original task—

known as the *source task*—may require less training data to perform well on a new task—known as the *target task*—if there is some *similarity between the tasks*. Coming back to our "imaginary" scenario, all we have to do is to train the CNN once in a specific thermodynamic condition[3] and then fine-tune this pre-trained model on the other conditions.

Throughout this work we focused on gas adsorption, but of course this doesn't mean we are not interested in predicting other properties of reticular materials. *What if we are asked to predict properties such as band gap or bulk modulus?* In that case, quantities such as *electron density* are more informative over host-guest interactions with regards to the aforementioned properties. This entails that the *voxelized electron density* should substitute the voxelized PES, as input to the 3D CNN. Nevertheless, ***wouldn't be great if all properties could be predicted from one and only one input?*** If our aim is to predict *different properties for the same structure, shouldn't the structure itself be used as input?* Currently, the approaches to tackle this challenge are based on *text representations* (Bucior, Rosen, et al. 2019; Cao et al. 2023) *and crystal graphs* (Xie et al. 2018; Chen et al. 2019). The main drawback of these approaches, is their inability to represent exactly the structure, that is the *exact arrangement of the atoms in the 3D space*.

Point clouds (Qi et al. 2016; Bello et al. 2020) are a natural way to solve this problem, since they are just *a set of coordinates and associated features*. In our context, the coordinates are the coordinates of the atoms, and the associated features are the types of the atoms. It should be emphasized, that *a point cloud is not another mathematical representation of a material—in the sense of a descriptor—it is the material itself*[4]. Therefore, an answer to the original question is to *couple point clouds with a neural network that can handle such kind of input*. This approach might have to overcome the current immaturity of DL over point clouds—especially regarding materials and molecules—but from a chemical perspective, it is the only one that truly respects the 3D nature of chemistry and of course, reticular chemistry.

[3] Preferably, the one where we have more labeled training data.
[4] Same ideas apply for molecules and in general, for any chemical system.