universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„New Machine Learning Models for Predicting the Performance of MOFs in Post-Combustion Carbon Capture“

verfasst von / submitted by

Ludwig Schwiedrzik, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2019 / Vienna 2019

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 862
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Chemie
Betreut von / Supervisor:	Univ.-Prof. Dr. Dr. h.c. Leticia González Herrero

Abstract

This master's thesis project was carried out in collaboration with the group of Professor Tom K. Woo, who kindly hosted me and my research during my stay at the University of Ottawa.

Carbon Capture and Storage (CCS) is expected to play a major role in combating global warming by reducing CO_2 emissions, primarily from the energy sector [1]. Adsorption-based post-combustion capture using porous solids is an area of ongoing research, with much interest being focused on Metal-Organic Frameworks (MOFs). MOF structures can be tuned to any set of adsorption requirements [2]. Due to the vast number of possible MOF structures, computational high-throughput (HT) screening is being increasingly used to accelerate the discovery of new high-performing MOFs. Screening large databases using computational methods such as Grand Canonical Monte Carlo (GCMC) is associated with high computational costs [3]. Therefore, prescreening methods that efficiently select promising MOFs for further GCMC screening are a key area of research; such methods are the focus of a long-running research project in the Woo group [4, 5, 6, 7].

My Master's thesis project research goal is to develop Support Vector Regression (SVR)-based prescreening tools using a large, highly diverse database of 352k MOFs to accurately predict CO_2 working capacities (wc) and selectivities (Sel) for post-combustion carbon capture, thereby identifying top performers for further GCMC screening. I make use of GCMC-calculated reference data and the nAP-RDF (normalized Atomic Property Weighted Radial Distribution Functions) as well as geometric descriptors. I also investigate the effects of Principal Component Analysis (PCA)-based pruning of descriptors. In this way, I develop four separate SVR models for MOF wc and Sel and extensively test their accuracy and predictive performance.

All four of the SVR models are shown to fit the training data very well. The models also predict the wc and Sel of MOFs in the 352k database with a high degree of fitting accuracy, although some overprediction of low performers and underprediction of high performers is evident. All four models' ranking accuracy is excellent, and the models function very well as prescreening tools to predict the wc and Sel of MOFs in post-combustion carbon capture. The two models incorporating PCA pruning clearly outperform the other two in terms of computational efficiency, demonstrating the advantages of this approach.

Dieses Masterarbeitsprojekt wurde in Zusammenarbeit mit der Arbeitsgruppe von Professor Tom Woo durchgeführt, der mich und meine Forschung während meines Aufenthaltes an der University of Ottawa freundlicherweise bewirtet hat.

Carbon Capture and Storage (CCS) wird voraussichtlich eine wichtige Rolle bei der Bekämpfung der Erderwärmung spielen, indem es zur Verringerung des CO_2-Ausstoßes, besonders im Energiesektor, beiträgt [1]. Dabei stellt die adsorptions-basierte Abtrennung von CO_2 nach Verbrennung von z.B. Kohle mithilfe poröser Feststoffe aktuell ein bedeutendes Forschungsgebiet dar. Insbesondere Metal-Organic Frameworks (MOFs) erregen aufgrund ihrer rekordbrechend großen Oberflächen und Porenvolumina, aber auch weil ihre Struktur an viele verschiedene Adsorptionsbedingungen angepasst werden kann, großes Aufsehen [2]. Aufgrund der schier unendlichen Zahl an möglichen MOFs erfreut sich computergestütztes High-Throughput (HT) Screening als Technik zur Beschleunigung der Entdeckung neuer leistungsstarker MOFs zunehmender Beliebheit. Das Screening großer Datenbanken mithilfe computergestützter Methoden wie Grand Canonical Monte Carlo (GCMC)-Simulationen ist mit hohem Rechenaufwand verbunden [3]. Daher stellen vorgeschaltete Prescreening-Methoden, die vielversprechende MOFs effizient für weiterführendes GCMC-Screening auswählen, ein Forschungsgebiet von entscheidender Wichtigkeit dar; solche Methoden stehen im Mittelpunkt eines lang anhaltenden Forschungsprojektes in der Arbeitsgruppe Woo [4, 5, 6, 7].

Das Ziel meines Masterarbeitsprojektes ist die Entwicklung eines Support Vector Regression (SVR)-basierten Prescreening-Werkzeuges, das unter Verwendung einer großen, hochgradig diversen Datenbank von 352k MOFs die Arbeitskapazitäten (Working Capacity, wc) und Selektivitäten (Selectivity, Sel) von MOFs bei der Abtrennung von CO_2 nach Verbrennung vorhersagt, wodurch besonders leistungsstarke MOFs für weiteres GCMC-Screening ausgewählt werden sollen. Ich nutze dabei mittels GCMC berechnete Referenzdaten und den nAP-RDF (normalized Atomic Property Weighted Radial Distribution Functions)- sowie geometrische Desktriptoren. Des Weiteren untersuche ich, wie sich Principal Component Analysis (PCA)-basiertes "Pruning" der Deskriptoren auswirkt. Auf diese Weise entwickle ich vier separate SVR-Modelle für MOF wc und Sel und prüfe die Genauigkeit und Vorhersageleistung der Modelle ausführlichst.

Alle vier SVR-Modelle sind sehr gut an die Trainigsdaten angepasst. Die Modelle sagen außerdem die wc und Sel der MOFs in der 352k-Datenband mit hoher Genauigkeit voraus, obwohl eine gewisse Überschätzung bei wenig leistungsfähigen sowie eine Unterschätzung bei hochleitungsfähigen MOFs bemerkbar ist. Die Rangkorrelationen aller vier Modelle ist ausgezeichnet, und die Modelle fungieren weiters sehr gut als Prescreening-Werkzeuge zur Vorhersage der wc und Sel von MOFs bei der Abtrennung von CO_2 nach Verbrennung. Die zwei Modelle, die PCA-basiertes "Pruning" verwenden, übertrefen die anderen beiden Modelle deutlich im Bezug auf ihre Recheneffizienz, wodurch die Vorzüge dieses Ansatzes bewiesen sind.

Acknowledgements

The research that is the subject of the following Master's thesis was carried out during a seven-month stay at the University of Ottawa, Ottawa, ON, Canada from January to July 2018. Many people have contributed to the successful completion of this project, and I would like to thank each of them accordingly.

First of all, I would like to thank Prof. Leticia González. I owe her a great debt for first introducing me to the field of theoretical chemistry as well as teaching me most of what I know about it. She also encouraged and helped me in planning my stay in Ottawa. Her willingness to supervise my thesis over a distance of over 6500 km was crucial in making this unusual project a workable reality, and I hope that the resulting thesis meets her high standards of scientific excellence.

I would also like to thank Prof. Tom Woo. He was incredibly welcoming, going so far as to pick me up form the airport and inviting me to my first hot meal upon my arrival in Canada. He allowed me to join his group for the duration of my stay and supervised my research on a day-to-day basis. Through intense discussion and valuable feedback, he pushed me to continuously improve my results. I am intensely grateful for the amazing opportunity he has afforded me, and I hope to have been able to contribute to the outstanding research carried out in Prof. Woo's research group.

Of the members of the Woo group, I would especially like to thank Tom Burns. He was always the first person I turned to with problems that came up during my work, and his ability to unflusteredly explain programming, physical chemistry, and the particularities of Canadian culture has been invaluable in helping me retain my sanity. I am thankful for his constructive criticism, his insightful comments, and his friendship. I would also like to thank Mykhaylo Krykunov, Sean Collins, and Summer Ho for their respective contributions in bringing my research project to fruition.

Finally, I would like to thank my colleagues Veronika Zeindlhofer and Samuel Senoner for their help in proofreading. I would like to thank my parents for their support throughout the time it has taken me to finish my thesis. I would like to thank my girlfriend, Irene, for always standing by me, especially during the seven long months we spent on different continents. And finally, I would like to thank all the wonderful people I met during my stay in Canada, without whom my stay would possibly have been more productive, but certainly less adventurous.

Contents

List of Figures

List of Tables

List of Abbreviations

AP-RDF	Atomic Property Weighted Radial Distribution Functions
CCS	Carbon Capture and Storage
CoRE	Computation-Ready, Experimental Database
CSD	Cambridge Structural Database
DFT	Density Functional Theory
GCMC	Grand Canonical Monte Carlo
HT	High-Throughput
MOF	Metal-Organic Framework
ML	Machine Learning
PCA	Principal Component Analysis
PSA	Pressure Swing Adsorption
QEq	Charge Equilibration
QSPR	Quantitative Structure-Property Relationship
RBF	Radial Basis Function
RDF	Radial Distribution Function
SBU	Structural Building Unit
Sel	Selectivity
SVM	Support Vector Machine
SVR	Support Vector Regression
TSA	Temperature Swing Adsorption
wc	working capacity

Chapter 1

Introduction

1.1 Carbon Capture and Storage

Climate change is counted among the foremost challenges facing humanity today. The atmospheric concentration of CO_2 has increased by 40% since the beginning of the industrial revolution, making carbon dioxide the single greatest contributor to global warming as a greenhouse gas [11]. Furthermore, some 30% of anthropogenic CO_2 is absorbed by the oceans, leading to progressive ocean acidification [11, 12]. According to the International Energy Agency, the energy sector is responsible for 80% of global CO_2 emissions, of which 45% can be attributed to the burning of coal [13]. Clearly, the reduction of CO_2 emissions is of the utmost importance if climate change is to be effectively combated. One particularly promising approach focuses on capturing CO_2 from point sources and permanently storing it underground. This process is known as Carbon Capture and Storage (CCS).

It is expected that CCS will account for up to 55% of the emission reductions required to meet the global warming target of 2^oC set for the year 2100, underlining its critical role in climate change mitigation [14]. CCS technology offers great utility because it can be integrated into existing systems such as coal-fired power plants, complementing power generation from renewable energy sources (solar, wind, etc.) through fossil fuels' greater flexibility and independence from external factors. A small number of industrial-scale CCS projects are already operational, SaskPower's Boundary Dam Power Station near Estevan, SK, Canada and Petra Nova at the WA Parish Generating Station in Thompson, TX, USA being the best-known examples. While CCS has primarily been considered for application in the energy sector, it could equally be used to decarbonize industries such as cement, steel, and paper production or petroleum refining [1]. Finally, CCS can even be utilized for removing excess CO_2 from the atmosphere, being the key component of negative emission technologies such as Direct Air Capture [15] or Bio-Energy with Carbon Capture and Storage [1]. It can therefore be assumed that the importance of CCS will only increase over time.

1.2 Post-Combustion Carbon Capture

Coal-fired power plants have long been the primary target of CCS research and development due to the high percentage of global CO_2 emissions attributed to them. Three main strategies for capturing CO_2 from such power plants can be identified: In *post-combustion* capture, CO_2 is separated from the flue gas that results from simply burning coal in air. In *pre-combustion* capture, CO_2 is captured before burning by gasifying coal at high pressure and temperature, which eventually produces a mixed gas stream of CO_2 and H_2 that can then be separated. Finally, in *oxy-fuel* capture, coal is burned in pure oxygen, producing only H_2O and CO_2, which are separated through condensation of water [1, 16]. My thesis will focus on post-combustion carbon capture because this technology can most easily be retrofitted to existing power plants, while pre-combustion and oxy-fuel capture require major redesigns or even the construction of entirely new power plants [1].

Within the strategy of post-combustion carbon capture, there are numerous techniques that can be used to separate CO_2 from N_2 and H_2O, the principal components of flue gas (see table 1.1). Aqueous amine scrubbing involves mixing flue gas with a solution of organic amines that react reversibly with CO_2, forming water-soluble salts. This is currently the dominant post-combustion capture technique; it is for example used at both the Boundary Dam and Petra Nova industrial-scale CCS projects. The main disadvantage of aqueous amine scrubbing is that the amine solution, once saturated, needs to be regenerated through boiling, a process that can lead to a significant decrease (ca. 30%) in the efficiency of coal-fired power plants, thereby increasing the cost of power generation. Many amines also decompose to form toxic or corrosive substances, creating maintenance, disposal, and environmental contamination problems [1, 16].

Adsorption-based carbon capture may offer a cheaper and more environmentally friendly alternative to aqueous amine scrubbing. Porous solids such as zeolites are already widely used for industrial-scale gas separation applications such as natural gas purification. Zeolites are microporous aluminosilicate minerals whose crystalline structure is characterized by large pores of 4-15 Å diameter and surface areas in the range of 200-500 m^2g^{-1}. They reversibly and selectively bind CO_2 over N_2 through physisorption. Once saturated, zeolites can typically be regenerated through applying either high temperature (temperature swing adsorption, TSA) or low pressure (pressure swing adsorption, PSA). It is important to note that the energy requirements for regenerating a porous solid binding gas molecules through physisorption are much lower than those needed in aqueous amine scrubbing because the latter involves the breaking of chemical bonds [1].

Table 1.1: Composition of post-combustion flue gas for a coal-fired power plant. Table reproduced from [10].

Chemical Species	Volume Fraction
N_2	73-77%
CO_2	15-16%
H_2O	5-7%
O_2	3-4%
SO_2	800 ppm
NO_x	500 ppm
HCl	100 ppm
CO	20 ppm
hydrocarbons	10 ppm
SO_3	10 ppm
Hg	1 ppb

1.3 Metal-Organic Frameworks

Unfortunately, most zeolites are not well suited to post-combustion capture from coal-fired power plants because they adsorb H_2O present in flue gas streams, reducing their CO_2 uptake under humid conditions [1]. It is therefore desirable to develop new porous materials that can reversibly separate CO_2 from N_2 at atmospheric temperature and pressure and high relative humidity. One class of materials that has generated considerable research interest is Metal-Organic Frameworks (MOFs). MOFs are porous crystalline solids created through self-assembly of inorganic and organic structural building units (SBUs): Metal ions or ion clusters and organic ligands, often carboxylate or imidazolate compounds. Each SBU has two or more coordination sites, necessary for the formation of the three-dimensional framework. The sheer number of possible combinations of inorganic and organic SBUs, combined with the option of postsynthetic functionalization of the organic ligands, ensures that the properties of MOFs can be finely tuned to meet any set of adsorption requirements [1, 8, 2].

Figure 1.1 illustrates the SBUs as well as the molecular and μm-scale structure of MOF-5 [17], one of the most studied MOFs. It is clear how the size of pores and channels within the MOF structure seen in b) could be altered by exchanging the organic SBUs of MOF-5 with ones of identical connectivity, but different dimensions, i.e. a longer molecule would create larger channels. One could also replace the inorganic SBUs with an isostructural ion cluster without fundamentally changing the framework. MOFs that share the same connectivity between SBUs in this manner belong to the same structure type, or more generally to the same net topology [2].

MOFs are known for their record-breaking surface areas, pore volumes, and

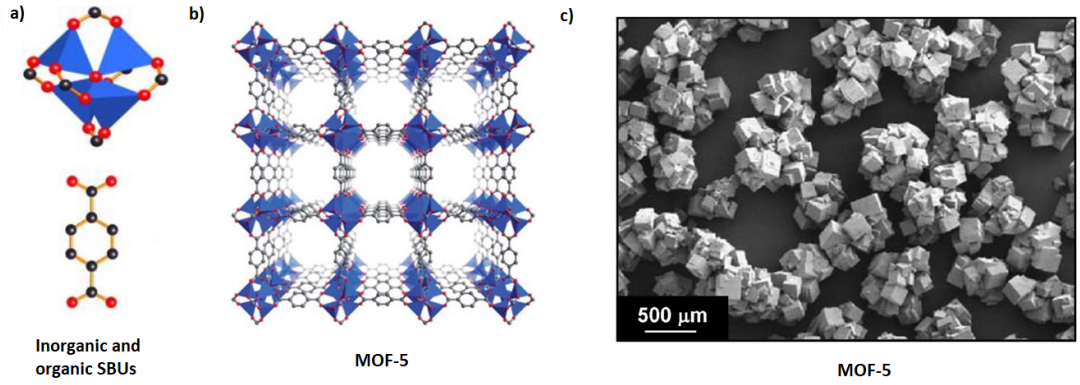

Figure 1.1: a) The inorganic ($Zn_4(O)(CO_2)_6$ cluster) and organic (1,4-benzodicarboxylate) SBUs that constitute MOF-5. b) The structure of MOF-5, as viewed through the (100) plane, showing its primitive cubic (pcu) net topology (see below). Figure adapted from [2]. c) A scanning electron microscope image of MOF-5, showing its morphology on the μm scale. Figure adapted from [8].

void fractions. They can perform gas separation based on two different mechanisms: kinetic and thermodynamic separation. In kinetic separation, a MOF with sufficiently small pores acts as a molecular sieve, allowing molecules with a small kinetic diameter to diffuse into its pores, while larger molecules cannot enter into the framework. As CO_2 and N_2 have very similar kinetic diameters, this separation mechanism is not useful for post-combustion carbon capture. In thermodynamic separation, gas molecules are selectively adsorbed based on their differing affinity to the MOF surface, which is caused by the molecules' physical and chemical properties. For example, table 1.2 shows that CO_2 has a higher polarizability and a much greater quadrupole moment than N_2. CO_2 therefore has a much higher affinity to the pore surfaces of MOFs that are studded with polar or charged groups, such as organic heteroatom functionalities or exposed metal cations, and it is thus selectively adsorbed at such surfaces through physisorption [2].

The selectivity of MOFs for CO_2 can be greatly enhanced by functionalizing MOFs with Lewis bases (e.g. organic amines) that are sufficiently strong to carry out a nucleophilic attack on CO_2, leading to bond formation - a chemisorptive

Table 1.2: Polarizability and quadrupole moment of CO_2 and N_2 [2].

Molecule	Polarizability [cm^{-3}]	Quadrupole Moment [$C{\cdot}m^2$]
CO_2	$29,1 \cdot 10^{-25}$	$13,4 \cdot 10^{-40}$
N_2	$17,4 \cdot 10^{-25}$	$4,7 \cdot 10^{-40}$

process. MOFs with a high selectivity capture CO_2 of higher purity while requiring a smaller adsorbent bed. However, an enhancement in selectivity comes at the cost of increasing the enthalpy of adsorption, which in turn leads to greater energy requirements for adsorbent regeneration. In order for a MOF to perform well in post-combustion carbon capture, its selectivity must be finely balanced against its adsorption enthalpy [2].

1.4 Evaluating Carbon Capture Performance

In addition to balancing selectivity and enthalpy of adsorption, a MOF must fulfill a number of criteria to compete with existing adsorbents in CCS. The MOF must be chemically and thermally stable under operating conditions, and its mechanical stability must be sufficient for dense packing in an adsorbent bed. Its pores must be large enough to allow for rapid diffusion of gas through the framework. Finally, the MOF must exhibit a large absorptive capacity, which is measured both gravimetrically and volumetrically. The gravimetric uptake dictates the mass of MOF needed in an adsorbent bed, while the volumetric uptake influences the total volume of the adsorbent bed. While each of these qualities is important in its own right, the selectivity and capacity of a MOF are typically seen as the two foremost indicators of its usefulness in CCS applications [2].

Having thus identified the qualities that make a useful adsorbent for post-combustion carbon capture, one must now attempt to find a MOF that can outper-

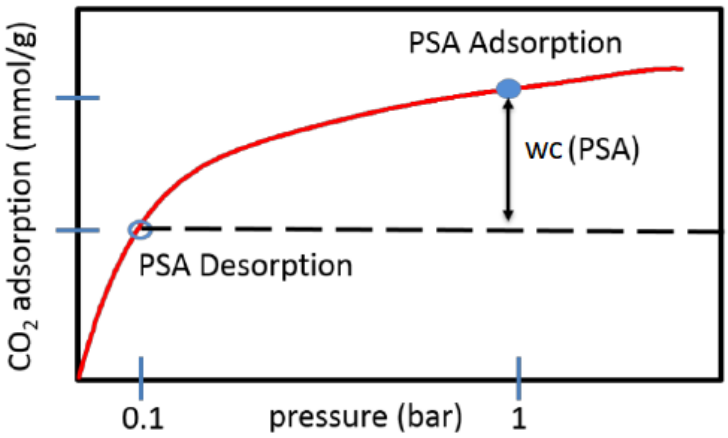

Figure 1.2: An example of an adsorption isotherm, showing the amount of CO_2 adsorbed in mmol/g of adsorbent over the total pressure in bar. The closed circle denotes the CO_2 uptake at adsorption pressure for PSA, q_{ads}. The open circle denotes the CO_2 uptake at desorption pressure for PSA, q_{des}. wc is the working capacity of the material for a full PSA cycle (see equation 1.1). Figure adapted from [7].

form existing materials. To this end, single-component adsorption isotherms for CO_2 and N_2 are measured, and the sample MOF's working capacity (wc) and selectivity (Sel) are calculated from them. Adsorption isotherms record the amount of gas adsorbed over a pressure range at a given constant temperature; an example of an adsorption isotherm is shown in figure 1.2. The wc is defined as

$$\text{wc} = q_{ads} - q_{des}, \tag{1.1}$$

the uptake at adsorption pressure q_{ads} minus the uptake at desorption pressure q_{des}. The Sel is expressed through the selectivity factor,

$$\text{Sel} = \frac{q_1 / q_2}{p_1 / p_2}, \tag{1.2}$$

the uptake of component 1 (CO_2) divided by the uptake of component 2 (N_2), normalized by the ratio of partial pressures of the two components within the flue gas mixture. It should be noted that a selectivity factor calculated on the basis of single-component adsorption isotherms does not directly represent the Sel of the MOF toward CO_2 in the gas mixture. Rather, it gives an estimate of how well a MOF might perform, so that it can be compared to other adsorbents. As experimental multi-component adsorption isotherms are not widely available, this is seen as an acceptable compromise [2].

1.5 High-Throughput Screening

The discovery of new MOFs for adsorption-based carbon capture is a complex chemical challenge due to the practically infinite number of possible MOF structures as well as the variety of performance criteria a candidate MOF has to fulfill to be considered for eventual industrial-scale deployment. While new structures are constantly being synthesized and analyzed, some researchers have come to embrace computational methods as a supplement or even alternative to extensive experimental studies. For example, Density Functional Theory (DFT) has played an important role in determining the arrangement of adsorbed CO_2 molecules in MOF pores and measuring the effects of ligand functionalization, thereby furthering our understanding of the interactions taking place between adsorbent and adsorbate. DFT has even been used to virtually screen dozens of MOFs to select the most promising candidate for synthesis and further characterization. However, to take full advantage of the possibilities offered by modern computational methods, much larger numbers of MOFs must be analyzed using a technique known as High-Throughput (HT) screening [3, 18].

In HT screening, tens or even hundreds of thousands of structures are analyzed to identify promising candidates and uncover previously unknown structure-property relationships. To apply this technique to MOFs in the context of CCS, three things are needed: a large, diverse database of MOF structures

on which HT screening can be carried out; a small, select set of performance parameters; and a method that can calculate these parameters for a given MOF structure in a computationally efficient manner [3, 18].

1.5.1 MOF Databases

Computational simulations require a description of the MOF structure as input, typically in the form of three-dimensional atomic coordinates. For HT screening purposes, these input structures can come from two sources: experimental crystal structures determined using X-ray diffraction; or hypothetical structures generated by computer algorithms [3]. Experimental crystal structures are usually deposited in the Cambridge Structural Database (CSD) when new MOFs are synthesized [19]. However, these experimentally determined structures may contain defects that need to be corrected before such structures can be used in HT screening [3]. Chung et al. have published the Computation-Ready, Experimental (CoRE) database containing 5109 experimental MOF structures (including some duplicates) from the CSD that have been prepared for computational simulations [20].

Hypothetical structures can be generated using either a *bottom-up* or a *top-down* approach. The bottom-up approach is based on sequentially connecting SBUs until a periodic structure is formed [3]. Wilmer et al. developed a method wherein new hypothetical MOFs are constructed in a step-wise manner from SBUs extracted from existing frameworks [21]. They thus created a large database of 137953 MOFs; however, their database exhibits poor structural diversity, containing just six different net topologies [22]. In the top-down approach, a net topology is selected and appropriate SBUs are mapped onto it [3]. Boyd and Woo demonstrated a graph theory-based method (ToBasCCo) that uses a labelled quotient graph of the desired net topology and a list of SBUs with defined connection points to generate large, structurally diverse databases of hundreds of thousands of MOFs [23].

The diversity of a given MOF database can be quantified by computing the textural properties of the MOF structures, such as pore and channel diameters, accessible void volume, void fraction, and internal surface area [3]. These properties can be determined using program packages such as zeo++ [24] or Poreblazer [25]. The importance of determining the structural diversity of a given materials databases is twofold: Firstly, when using the database for HT screening-based materials discovery, researchers want to cover as wide of a variety of structures as possible, as this increases the likelihood of finding high-performing materials. Secondly, computing the textural properties of structures makes it possible to establish previously unknown structure-property relationships, facilitating the design of ever higher-performing MOFs [3].

1.5.2 Grand Canonical Monte Carlo Simulations

As discussed above, the Sel and wc are regarded among the most important indicators of adsorption performance. Both parameters can be determined from single-component adsorption isotherms, which Grand Canonical Monte Carlo (GCMC) simulations are able to calculate in good agreement with experimental data [26]. In the following, I will provide a short explanation of GCMC simulations. For a more thorough discussion, I would refer the reader to the excellent review of Dubbeldam et al. [27].

GCMC uses statistical mechanics to simulate a chemical system under constant chemical potential μ, volume V, and temperature T. From a given starting configuration, particles are randomly perturbed using a selection of predefined moves such as translation, rotation, and conformational changes within a molecule, or even "unphysical" moves like swapping particles or creating and deleting (guest) particles. The energies of the original configuration and of the new configuration (created through perturbation of the original) are evaluated using a force field and compared to one another. If the energy of the new configuration is lower than the energy of the old configuration, the new configuration is added to the ensemble average. If, on the other hand, the energy of the new configuration is higher than that of the old configuration, the new configuration is added to the ensemble average with a probability based on the Boltzmann distribution of the system. This process is repeated until the ensemble average has converged, allowing macroscopic properties of the system to be extracted from the simulation. Because the chemical potential of adsorbed gas molecules is equal to that of non-adsorbed gas molecules at equilibrium, the adsorption isotherm can be derived via the ideal gas equation at low pressures [27].

1.5.3 HT Computational Studies of MOFs

In the following, I will give a short overview of HT computational studies of MOFs, discussing the MOF databases analyzed in each study, what parameters were used to estimate MOF performance, and what methods were utilized for screening. Particular attention will be paid to how the authors dealt with the large computational costs of HT screening. For a more thorough discussion of computational studies of MOFs, I would direct the reader to the exhaustive reviews of Colón and Snurr [3] as well as of Yang et al. [28].

Wilmer et al. used their bottom-up method to generate hypothetical MOF structures (described above), which they screened for their CH_4 storage capabilities using GCMC. They used a three-tiered accuracy approach: All 137k MOFs were subjected to short GCMC runs, and the top 5% in terms of storage performance were selected for longer (more accurate) GCMC runs. The top 5% from this longer run were re-examined one final time with even longer GCMC runs. In this way, the CH_4 storage properties of the top-performing MOFs were computed with a high degree of accuracy, and the amount of computation time

spent on low-performing MOFs was reduced. Over 300 promising new MOF structures were identified, and several structure-property relationships relevant to CH_4 storage were uncovered [21].

Van Heest et al. studied MOFs for noble gas separation applications. They started with some 3000 MOFs from the CSD, which they screened by calculating textural, adsorptive, and transport properties and estimating the Sel of the MOFs based on those properties. In this way, they managed to identify 70 high-Sel structures, for which they generated single- and multi-component adsorption isotherms using GCMC simulations and Ideal Adsorbed Solution Theory [29].

As Coulombic interactions play an important role in the adsorption of CO_2 on MOF surfaces, partial charges must be assigned to the atoms of MOF structures so that GCMC simulations can produce accurate results. This task is usually carried out using computationally expensive ab-initio calculations, but that approach is hardly feasible considering the large number of structures involved in HT screening studies. Several groups have therefore studied alternative charge assignment schemes designed to speed up this necessary but time-consuming simulation step, thereby increasing the computational efficiency of HT screening of materials [3].

Haldoupis et al. developed a new charge equilibration (QEq) method for periodic systems. In QEq, charges are typically assigned based on experimental ionization potentials and electronegativity values. Haldoupis et al. used their new PQEq method to assign charges to 500 MOF structures from the CSD to efficiently determine their adsorption properties and carry out further, more detailed simulations on the most promising MOFs [30].

Kandatsev et al. of the Woo group have published the specialized MOF electrostatic-potential-optimized charge equilibration scheme (MEPO-QEq). In this approach, the atomic partial charges are fit using a genetic algorithm to reproduce the electrostatic potential derived from a periodic DFT calculation. Results were validated by comparing the adsorptive capabilities of over 1200 MOFs, calculated using MEPO-QEq, to values computed for the same structures using charges derived directly from DFT; the two sets of results were found to be in excellent agreement [31].

1.5.4 Predicting Adsorption Performance

A common theme in the computational HT screening studies described above is the need to make efficient use of computational resources. Whether by controlling the length of GCMC simulations or by developing new algorithms for atomic partial charge derivation, researchers are constantly looking for ways to minimize the amount of computational time spent on screening tasks. This makes perfect sense when one considers that, relatively speaking, the majority of any given set of MOFs does not exhibit high performance in terms of CO_2 adsorption (or any other metric) by definition. Therefore, any time spent producing highly accurate results for low-performing MOFs is essentially wasted. The ideal solution to this

problem would be to identify high-performing structures before GCMC-based screening is carried out, so that accurate and expensive simulations are reserved for those high performers.

This approach is more generally referred to as *pre*screening: making some basic selection of structures that are most likely to perform well before the actual screening process. While prescreening can be carried out in a variety of ways, the large number of structures involved in HT screening suggests that Machine Learning (ML) methods from the field of Data Science might be particularly well suited for this task. This idea is the foundation of a long-running research project within the Woo group, which focuses on developing ML-based prescreening tools to facilitate the search for high-performing MOFs for carbon capture.

Based on the observation that Quantitative Structure-Property Relationship (QSPR) models could be developed using the textural properties of MOFs as geometric descriptors to accurately predict the CH_4 uptake of the MOFs [4], Fernandez, Trefiak, and Woo [5] attempted to use the same approach to predict the CO_2 adsorption capabilities of MOFs. However, they found that a more sophisticated descriptor was needed. They therefore developed the Atomic Property Weighted Radial Distribution Functions (AP-RDF) descriptor, which they used in combination with multilinear regression and Support Vector Machine (SVM) ML techniques to predict CO_2 uptakes of the 137k MOF structures from the database of Wilmer et al. [21] While the new descriptor performed well in predicting the performance of MOFs for high-pressure applications, the results obtained in the low-pressure regime relevant to post-combustion carbon capture were less satisfactory [5].

Fernandez et al. [6] went on to develop an SVM-based classifier that allowed them to rapidly sort MOFs into high-performing and low-performing bins, an essential prescreening task. This new classification tool proved to be highly accurate even at low pressures, and allowed the authors to eliminate 90% of a new database of 324500 MOFs in prescreening, thereby reducing screening time by one order of magnitude. However, the new algorithm could not be used to predict numerical values for the CO_2 uptake of MOF structures [6].

Most recently, Dureckova showed in her Master's thesis [7] that SVM-based Support Vector Regression (SVR) algorithms could be utilized to predict the wc and Sel of MOFs for *pre-combustion* carbon capture with high accuracy. She investigated a large and highly topologically diverse database of 358400 MOFs generated partly by an algorithm similar to that of Wilmer et al. [21] and partly by the Woo group's ToBasCCo algorithm [23], making use of both geometric descriptors based on MOF textural properties and a novel normalized variant of the AP-RDF descriptor (nAP-RDF). The accuracy of the SVR-predicted wc and Sel values was validated by comparing them to GCMC results [7].

1.6 Project Summary

To sum up, CCS is expected to play a major role in combating global warming by curbing CO_2 emissions from the energy sector and other industries. Among CCS technologies, adsorption-based post-combustion capture using porous solids is seen as a particularly promising approach [1]. In this context, MOFs have generated great research interest partly because of their record-breaking surface areas and pore volumes. More importantly, MOF structures can be finely tuned to meet any set of adsorption requirements. The wc and Sel of MOFs are seen as the foremost indicators of high carbon capture performance [2]. Due to the sheer endless number of possible MOF structures, HT screening using computational methods has become increasingly popular as a tool to speed up the discovery of new high-performing MOFs and uncover previously unknown QSPRs. HT screening requires large, diverse databases of computation-ready MOF structures, a well-defined set of performance parameters by which to rank MOFs, and a screening method that can efficiently compute those parameters for all the MOFs in a given database. Computer-generated hypothetical databases contain up to hundreds of thousands of MOFs, but their diversity must be verified through computing the structures' textural properties, a process known as database characterization. The wc and Sel of MOFs can be accurately computed by means of GCMC simulations, but the associated computational costs render this approach inefficient for truly large databases [3]. Hence, prescreening methods that efficiently select promising MOFs for further, more accurate screening are an important area of research that is the focus of a long-running research project in the Woo group [4, 5, 6, 7].

Based on the SVR models developed by Dureckova, my Master's thesis project research goal was as follows: To develop SVR-based prescreening tools using the nAP-RDF descriptor on a large, highly diverse database of 352k MOFs to accurately predict CO_2 wc and Sel for post-combustion PSA carbon capture, thereby identifying top performers for further GCMC screening. To this end, I used GCMC simulations to calculate wc and Sel for all 352k MOFs. I computed nAP-RDF and geometric descriptor values and investigated the effects of Principal Component Analysis (PCA)-based pruning of descriptors. In this way, I have developed four separate SVR models for MOF wc and Sel and extensively tested their accuracy and predictive performance.

In the following, I will discuss the methods I utilized, focusing on the database and how reference data was obtained from it, the descriptors and the pruning method applied to them, as well as the basics of SVM and SVR, the ML methods used in this thesis. I will then go on to present the results of my research, characterizing the database utilized, then discussing the accuracy and predictive performance of the SVR models developed, and finally comparing the prescreening tools proposed in the present work to previous iterations of similar prescreening tools developed in the Woo group. I will then summarize my findings and provide a short outlook.

Chapter 2

Methods

2.1 MOF Database and Reference Data

Calculations were carried out on a database of 352177 hypothetical MOFs. The 352k database consists of MOFs constructed from a total of 23 inorganic and 175 organic SBUs, arranged in 1166 net topologies and functionalized with 50 different functional groups. Both a bottom-up algorithm based on that of Wilmer et al. [21] and the top-down algorithm ToBasCCo [23] were used in generating the MOF structures, which were optimized with Molecular Dynamics using the Universal Force Field [32, 33] before being added to the database.

The Woo group's GCMC program package, FA^{3}PS (Fully Automated Adsorption Analysis in Porous Solids), was used to determine the MOFs' adsorption properties. FA^{3}PS simulated PSA cycles using the set of parameters listed in table 2.1. None of the constituents of post-combustion flue gas (cf. table 1.1) beside CO_2 and N_2 were included in the simulations, as this would have gone beyond the scope of the present work. Instead, partial pressures for CO_2 and N_2 were scaled to represent the volume fraction of CO_2 in flue gas (ca. 15%) and to add up to 1 bar for adsorption. Desorption takes place at 0,03 bar total pressure in this PSA cycle, and it is assumed that MOFs retain 99% CO_2 and 1% N_2 for comparison purposes. GCMC point calculations were run at the adsorption and desorption points for CO_2 as well as for N_2, with 20000 steps for equilibration and 20000 steps for production each. Atomic partial charges were assigned using MEPO-QEq [31]. wc were calculated according to equation 1.1, and Sel according to equation 1.2. These wc and Sel values are referred to as 'calculated' values, and they constitute the reference data for the SVR-based prediction used for prescreening.

Textural properties of the MOFs in the 352k database were calculated using zeo++ [24] and used to characterize the database, particularly in terms of its structural diversity. Six textural properties were investigated: dominant pore size, largest accessible pore size, void fraction, density, volumetric surface area, and gravimetric surface area. These properties were chosen because they have been

Table 2.1: GCMC simulation parameters for PSA reference data computation.

	Adsorption	Desorption
T	298 K	298 K
p_{total}	1,00 bar	0,0300 bar
p_{CO_2}	0,15 bar	0,0297 bar
p_{N_2}	0,85 bar	0,0003 bar

shown to be important indicators of the adsorptive performance of MOFs [4, 7]. Furthermore, the textural properties of 4377 charge-neutral experimental MOF structures from the CoRE database[20] were calculated by another member of the Woo group in the same way. The textural properties of MOFs from both the 352k and the CoRE database are reported together to give some experimentally-founded context for the diversity of the 352k database.

2.2 Descriptors

The six textural properties utilized for database characterization were additionally used as geometric descriptors, based on Dureckova's observation that the inclusion of these descriptors increased the accuracy of her SVR models [7].

The Radial Distribution Function (RDF) represents the weighted probability of finding a pair of atoms separated by a given distance in a chemical system [34]. The RDF can be used in conjunction with the minimum image convention to describe the structure of a periodic system. The nAP-RDF descriptor is a development of the AP-RDF descriptor published by Fernandez et al. [5]. It is defined as

$$RDF^P(R) = \frac{1}{N} \sum_{i,j}^{atompairs} P_i P_j e^{-B(r_{ij}-R)^2}, \tag{2.1}$$

wherein P is an atomic property defined for atoms i and j, R is the radius of a spherical volume for which the nAP-RDF descriptor is defined, N is the number of framework atoms in the unit cell, B is a smoothing parameter (set at 10), and r_{ij} is the minimum image convention distance between atoms i and j. The atomic properties utilized in the present work were the respective elements' electronegativity, hardness, and van der Waals volume; their inclusion ensures that chemical as well as structural information is encoded in the nAP-RDF descriptor values.

The nAP-RDF descriptor values were calculated for radii R from a range of 2,0 to 30,0 Å in 113 steps of 0,25 Å for each atomic property, meaning 339 descriptor values were computed for each MOF. The six geometric descriptor values were

added to give a total of 345 descriptor values per MOF. The effects of reducing the number of descriptors through PCA, termed pruning, were investigated in an attempt to increase computational efficiency.

2.3 PCA Pruning

To carry out pruning, 345 descriptor vectors were defined, each containing the descriptor values of all MOFs for a given radius R, centered to their mean and scaled to their variance. These were then subjected to PCA, wherein the 345 original descriptor vectors were replaced by a basis set of 40 orthogonal vectors, the principal components. Each principal component is a linear combination of the original vectors exhibiting the maximum variance permissible under the condition that it be orthogonal with the other principal components [35].

Pruning was carried out in the hope that greater computational efficiency could be achieved by using a smaller set of descriptor values with greater information density while maintaining a high level of accuracy. For comparison, SVR models with and without pruning were generated.

2.4 SVM Regression Models

SVMs are a class of ML algorithm that can be used for both classification and regression. They are well suited to handling high-dimensional input data and are considered to be memory efficient [36]. Like many other ML methods, they excel at interpolating between data points, but show poor extrapolation behaviour. In practice, SVM models are trained by fitting to a set of data points with known target (e.g. wc) values, the training set. They can then be used to predict the wc values of data points outside the training set. In the present work, the predictive performance of SVR models is tested by predicting the wc values of a large number of MOF structures, collectively termed the test set, and comparing the predictions to the GCMC-calculated target values. An important aspect of training an SVM or SVR model is to avoid overfitting: When the model fits the training data too closely, general trends in the data may be lost in favour of modeling mere noise; this can adversely affect predictive performance. Overfitting may be avoided by choosing an appropriate training set and kernel (see below).

Figure 2.1 illustrates the basic principle of SVM on a set of linearly separable 2-class objects. SVM is used to find the line (or more generally hyperplane) that separates the two classes with the greatest possible margin. This is achieved by identifying points in the data (termed support vectors) that define the border of a class: In figure 2.1, the hyperplane H_1 defines the border of class +1, while H_2 defines the border of class -1. They are separated by the margin δ. The solution to the classification problem shown in figure 2.1 is represented by the support

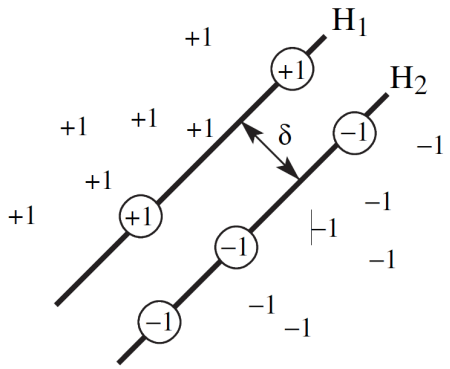

Figure 2.1: A set of objects belonging to two classes (+1 and -1). The borders of the two classes are the hyperplanes H_1 and H_2, separated by the margin δ. The support vectors used to define H_1 and H_2 are circled. Figure adapted from [9].

vectors (circled) that are used to define the maximum margin hyperplane. The remaining data points play no role in defining the hyperplane [9].

SVM can also be used to separate classes of objects that cannot be separated linearly. To this end, the objects are mapped into a higher-dimensional feature space, typically using a kernel, where they can be separated by a hyperplane. This is illustrated in figure 2.2, where two classes of objects are separated by an ellipsis in the input space, which corresponds to the hyperplane H in the feature space. A variety of kernels can be used to enable SVM to model complicated separation hyperplanes; however, overfitting is more likely to occur when using highly complex kernels. The goal of SVM classification should therefore always be to find the simplest function that separates two classes with the lowest possible number of support vectors [9].

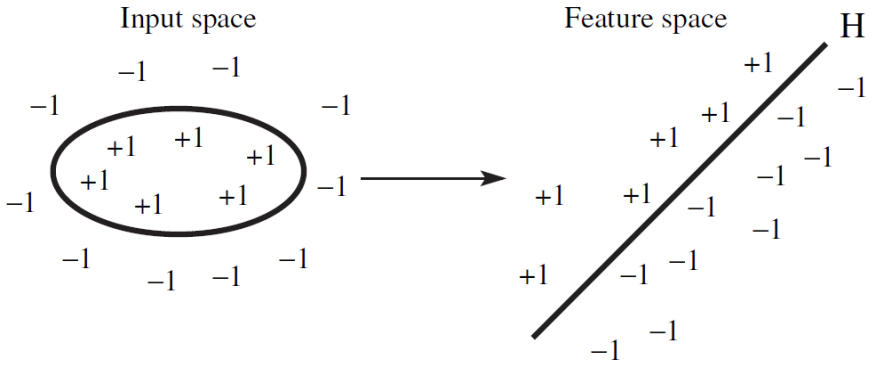

Figure 2.2: A set of objects belonging to two classes (+1 and -1). The objects cannot be separated linearly in the input space, but using a kernel to map them to a higher-dimensional feature space allows separation by the hyperplane H, which corresponds to an ellipsis in the input space. Figure adapted from [9].

For SVR, the definition of the hyperplane is altered: Instead of separating objects into classes, it attempts to identify a trend in a dataset and obtain a fit. A common approach is to use an ε-insensitive loss function: the regression hyperplane is surrounded by a tube with radius ε (see figure 2.3), and all data points within the tube are considered to have an error of zero, while data points outside the tube have an error that increases with distance from the outer margin of the tube. When ε is too small, most data points are outside the tube and will therefore become support vectors, increasing the complexity of the regression model and possibly leading to overfitting. A larger ε value decreases the number of support vectors required and decreases the likelihood that overfitting will occur at the cost of accuracy to the fit [9]. All calculations were carried out using a standard value of $\varepsilon = 0,1$.

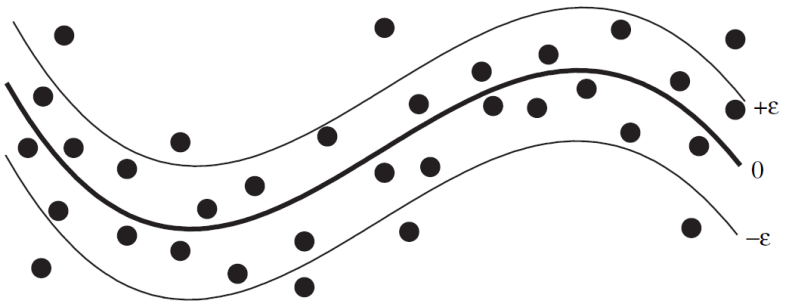

Figure 2.3: An example of an SVR model. The majority of data points are within the margin of $[+\varepsilon, -\varepsilon]$ around the regression hyperplane 0 and therefore have an error of zero. Data points outside the tube are assigned an error that increases with distance from the margin. Figure adapted from [9].

In total, four separate SVR models were generated (see table 2.2) using the SciKit-Learn [36] toolbox for Python. The Radial Basis Function (RBF) kernel was used throughout. The RBF kernel contains the hyperparameter γ, which determines how much influence a single data point has on the overall fit. γ needs to be set by the user, as does the capacity parameter C, which controls the smoothness of the fit [36]. To find the optimum values of C and γ, a grid search approach was used: 10% of the 352k database were randomly chosen to train SVR models using C and γ value pairs corresponding to all 49 possible combinations from the set $\{0{,}001; 0{,}010; 0{,}100; 1{,}000; 10{,}000; 100{,}000; 1000{,}000\}$. A four-fold cross-validation approach was used, meaning the training set was split into four equally sized subsets; each subset was then used once for prediction, while the other three were used for training. Thus, four equal models were created, each with a distinct R^2 value for prediction. The optimum pair of hyperparameter values for each of the four models was chosen based on the mean R^2 values obtained from 4-fold cross-validation, R^2_{Xval}, where

$$R^2 = 1 - \frac{\sum^M (x_i - t_i)^2}{\sum^M (x_i - \bar{t})^2}, \tag{2.2}$$

with M being the number of MOFs, x_i being the SVR-predicted value for MOF i, t_i being the GCMC-calculated target value for MOF i, and $\bar{t}$ being the average target value. R^2_{Xval} was also used to determine the goodness of fit obtained in training each of the models.

After determining the optimum hyperparameters, all four models were retrained using the same training sets selected previously. The wc and Sel values of the remaining 90% of the 352k database (the test set) were then predicted. Two standard metrics were used to judge the models: The R^2 (see equation 2.2) of prediction, R^2_{Pred}, indicates the accuracy of the fit obtained in training relative to the prediction set. The Spearman coefficient ρ (ρ_{Pred} for prediction) is a measure of ranking accuracy. It is defined as

$$\rho = 1 - \frac{6 \sum d_i^2}{M(M^2 - 1)}, \tag{2.3}$$

where d_i is the difference in rank of MOF i between the SVR prediction and the GCMC calculation and M is the number of MOFs in the prediction set. ρ_{Pred} measures the statistical dependence between the rankings of the SVR-predicted values and the calculated target values.

Finally, the prescreening performance of each of the SVR models was examined. To this end, the top 1000 MOFs, ranked by GCMC-calculated wc or Sel, were defined as the Top Performer set. The models were then evaluated by counting how many of the Top Performers (N/1000) were among the top 1%, 2%, etc. of predicted results. Thus, the prescreening efficiency of each of the models could be gauged by measuring how many Top Performers are recovered for further screening when the bottom 99%, 98%, etc. of MOFs according to the SVR model are discarded.

Table 2.2: SVR models generated as part of this research project. wc stands for working capacity, Sel stands for selectivity. Only models utilizing 40 descriptors were subjected to PCA pruning.

Model	Target	Descriptors
wc_BASIC	wc	345
wc_PCA40	wc	40
Sel_BASIC	Sel	345
Sel_PCA40	Sel	40

2.5 Overview

In summary, a database of 352k hypothetical MOF structures was generated and the MOFs' wc and Sel values were determined using GCMC. The 352k database was characterized by calculating the MOFs' dominant pore size, largest accessible pore size, void fraction, density, volumetric surface area, and gravimetric surface area, which were also used as geometric descriptors. The database was further compared to the CoRE database of experimental MOF structures. Additionally, 339 nAP-RDF descriptor values were computed for each MOF. PCA-based pruning was employed to increase computational efficiency. Thus, four SVR prediction models were generated, two each for wc and Sel, one of each of which included pruning: wc_BASIC, wc_PCA40, Sel_BASIC, Sel_PCA40. The hyperparameters C and γ were determined for all models using a grid-based approach. Finally, all models were trained and evaluated according to their training and prediction fitting accuracy, prediction ranking accuracy, and prescreening performance.

Chapter 3

Results & Discussion

3.1 Database Characterization

One of the elementary steps of any HT screening study is the selection of an appropriate database. Screening a large, diverse database increases the likelihood that high-performing materials may be found. However, the goal of this research project is not the immediate search for new MOFs, but rather the development of a ML-based prescreening tool. This prescreening tool should ideally be able to predict the wc and Sel of any given MOF structure; however, to be able to do this, the ML tool must be trained using a training set that is representative of a huge variety of MOFs, as SVR models are vastly better at interpolation than at extrapolation. It is thus necessary to demonstrate that the 352k database used for training and prediction in the present work contains a large variety of MOF structures. To this end, the 352k database is characterized by illustrating the density distributions of several textural properties (dominant pore size, largest accessible pore size, void fraction, density, volumetric surface area, and gravimetric surface area) and compared to the CoRE database of experimental MOFs, thereby demonstrating and contextualizing its high structural diversity.

Figure 3.1 shows the density distribution of the dominant pore size of hypothetical MOFs in the 352k as well as of experimental MOFs in the CoRE database. The 385k distribution has a broad, low peak at 10 Å and a smaller peak at 15 Å; it is distended toward higher values, with the maximum at 107,9 Å. A narrow peak at 5 Å dominates the CoRE distribution, which quickly drops off through a series of minor peaks above 10 Å. The maximum value is at 42,4 Å. The wider distribution of the 352k database indicates that it is more structurally diverse than the CoRE database. However, the 352k database also contains a significant number of MOFs with somewhat unrealistically large pores; there are 4573 MOFs in the 352k database with dominant pore diameters above 40 Å alone.

The largest accessible pore size distribution of the 352k and CoRE databases is depicted in figure 3.2. The 352k distribution is practically identical to that of the dominant pore size (figure 3.1), exhibiting the same characteristic peaks and

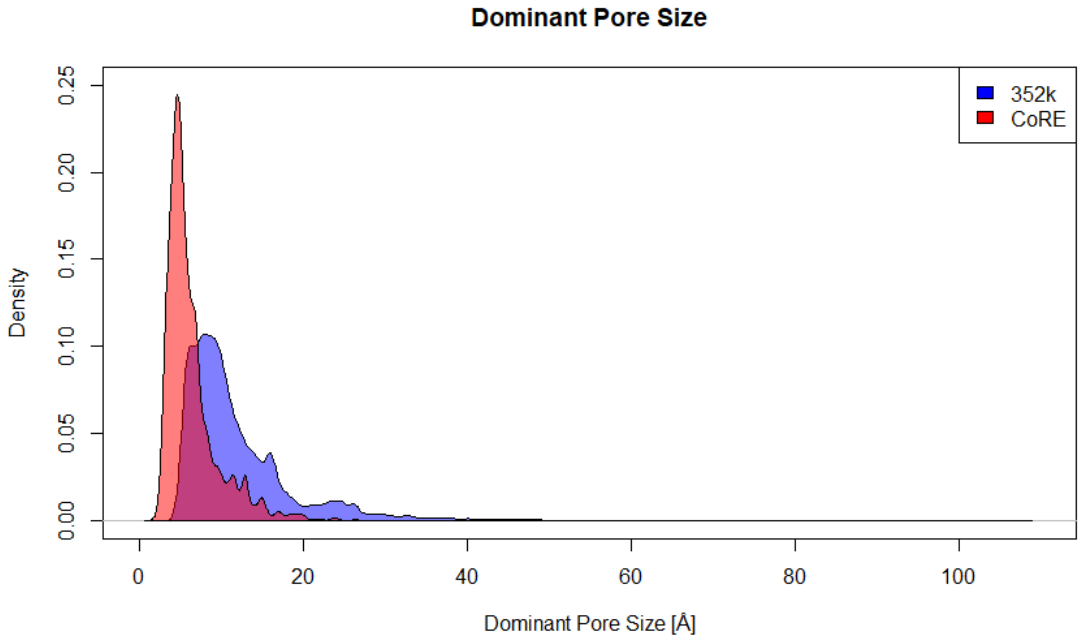

Figure 3.1: The density distribution of the dominant pore size in Å for all MOFs in the 352k (blue) and CoRE (red) databases. The 352k database exhibits a wider distribution of values, indicating that it is more structurally diverse.

maximum value. The same is true for the CoRE distribution. The dominant and largest accessible pore sizes appear to be strongly correlated with one another. While the 352k distribution is again wider than the CoRE distribution, this does not necessarily reinforce the conclusion that the 352k database is structurally more diverse due to the strong apparent correlation between the two pore sizes. However, figure 3.2 again illustrates the large number of MOFs with extremely high pore diameters found in the 352k database, with 4443 MOFs having a largest accessible pore size above 40 Å.

Figure 3.3 illustrates the void fraction distribution of the two databases. The 352k distribution is flat and wide, covering almost the entire range of possible values between 0 and 1, with a maximum value of 0,95. Conversely, the CoRE distribution is dominated by an extremely high peak at 0,0, quickly falling off toward a much smaller peak above 0,4; the maximum value is 0,83. Again, the visibly wider distribution of the 352k database is indicative of greater structural diversity compared to the CoRE database, which is dominated by MOFs of extremely low void fraction. On the other hand, the 352k database contains a large number of MOFs with extremely high void fractions; there are 5847 MOFs with void fractions over 0,8. This observation can be explained by the similar number of MOFs with unrealistically large pore sizes observed in figures 3.1 and 3.2, as

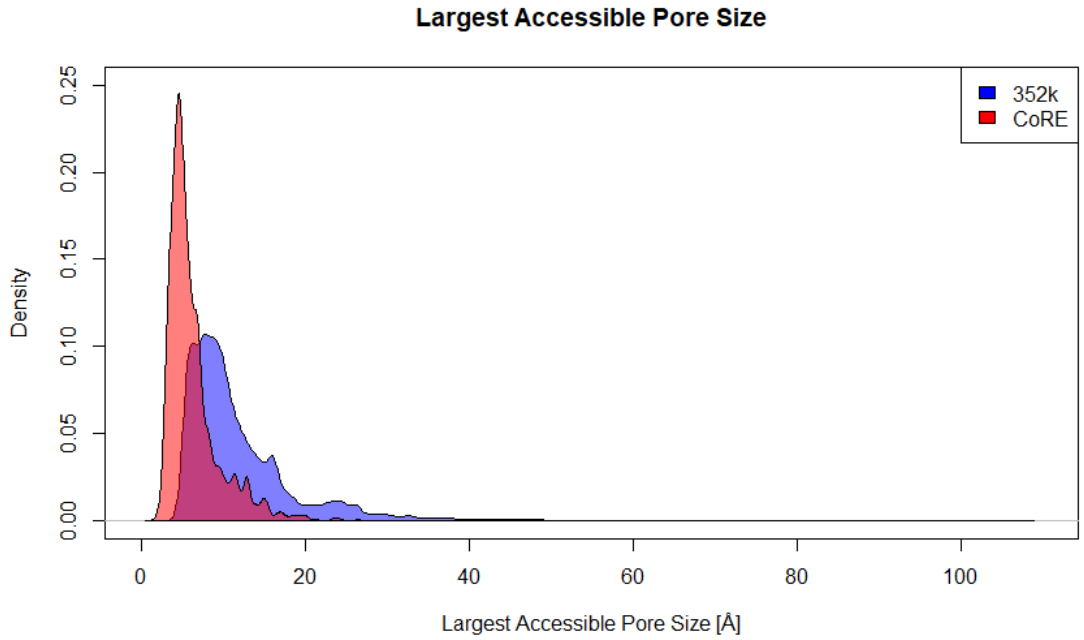

Figure 3.2: The density distribution of the largest accessible pore size in Å for all MOFs in the 352k (blue) and CoRE (red) databases. Both distributions are virtually identical to those found in figure 3.1 for the dominant pore size.

larger pore diameters lead to higher void fractions.

In figure 3.4, the volumetric mass density distribution of the 352k and CoRE databases is shown. The 352k distribution features a narrow peak at 800 kg m^{-3}, with shoulders around 400 and 200 kg m^{-3}; it falls off quickly toward higher values, until a maximum value of 4057 kg m^{-3}. The CoRE distribution has the appearance of a wide, fairly symmetric peak centered around 1200 kg m^{-3}, with a maximum value of 5180 kg m^{-3}. Unlike the other textural properties examined thus far, the 352k distribution in this case is narrower and shifted toward lower values than the CoRE distribution, suggesting that the CoRE database might be more structurally diverse than the 352k database in some ways. The low volumetric mass density of MOFs in the 352k database is closely connected to the large pore sizes and high void fractions that have already been discussed.

The distribution of volumetric surface area values across the 352k and CoRE databases can be seen in figure 3.5. The 352k distribution is marked by a high peak at 2000 m^2cm^{-3}, as well as being distended toward lower values by a shoulder at 1500 m^2cm^{-3}. Its maximum value is 3046 m^2cm^{-3}. The CoRE distribution is flat and even; it is marked by four shallow peaks of progressively declining height near 0, 500, 1500, and 2000 m^2cm^{-3}, with a maximum value of 3495 m^2cm^{-3}. While the 352k is clearly more peaked than the CoRE distri-

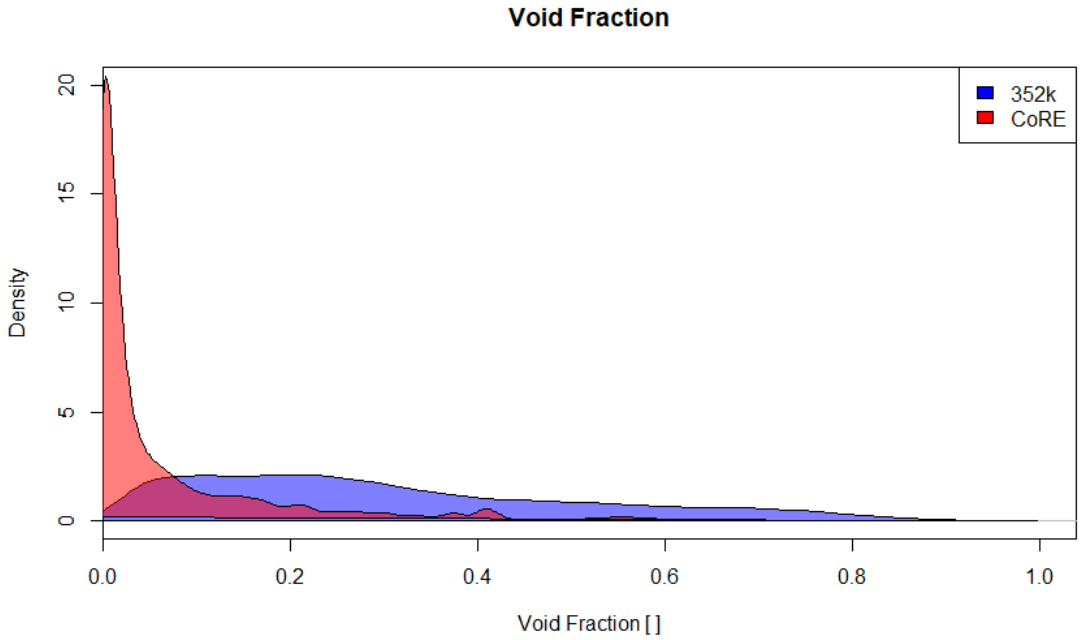

Figure 3.3: The density distribution of the void fraction for all MOFs in the 352k (blue) and CoRE (red) databases. The 352k database exhibits a much wider and flatter distribution of values, indicating that it is more structurally diverse.

bution, the 352k still covers almost the entire value range of 0 to 3000 m^2cm^{-3} also covered by CoRE, while being shifted toward high values. This shift toward higher values can be explained by the previously observed large pore sizes and high void fractions: larger pores lead to greater accessible surface areas per unit of volume as well as lower densities. The two databases exhibit similarly wide distributions, suggesting similar structural diversity, though the CoRE distribution is certainly more even.

Finally, figure 3.6 shows the gravimetric surface area distribution of both databases. The 352k distribution is flat and wide, being centered above 2000 m^2g^{-1} and distended toward even higher values by a shoulder at about 5500 m^2g^{-1}. Its maximum value is 9501 m^2g^{-1}. The CoRE distribution exhibits a large peak with two prominent shoulders at low surface area values (below 2500 m^2g^{-1}) and is much narrower than the 352k overall. Its maximum value is 6603 m^2g^{-1}. The wider distribution of the 352k database indicates that it is more structurally diverse than the CoRE database. In this case, there is no obvious correlation with the previously observed large pore sizes. This is due to the fact that, unlike volumetric textural properties such as the void fraction, density, and volumetric surface area, the *gravimetric* surface area is independent of volume.

In summary, the 352k database of hypothetical MOF structures used for HT

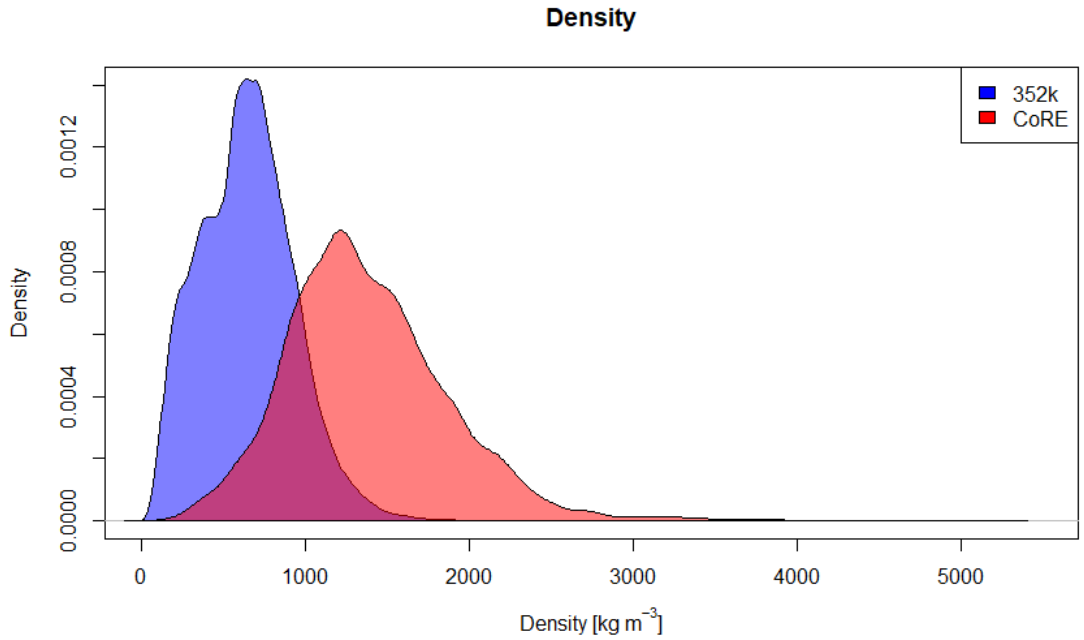

Figure 3.4: The density distribution of the volumetric mass density in kg m^{-3} for all MOFs in the 352k (blue) and CoRE (red) databases. Here, the CoRE database exhibits a wider distribution of values, suggesting that it may be more structurally diverse in some ways.

screening in the present work exhibits large structural diversity. The database contains 352177 MOFs constructed from 23 inorganic and 175 organic SBUs, arranged in 1166 net topologies and functionalized with 50 different functional groups. Furthermore, the 352k database shows wide density distributions for gravimetric surface area, void fraction, as well as dominant and largest accessible pore size (although the two pore sizes may be strongly correlated). However, the distributions of the volumetric surface area and the volumetric mass density for MOFs in the 352k database are narrower. This is most likely due to the large number of MOFs exhibiting somewhat unrealistically large pore diameters and void fractions found in the 352k database. Overall, the density distributions discussed above compare favourably to the distributions of values exhibited by the CoRE database of experimental MOF structures, underlining that the 352k database contains a large variety of MOFs.

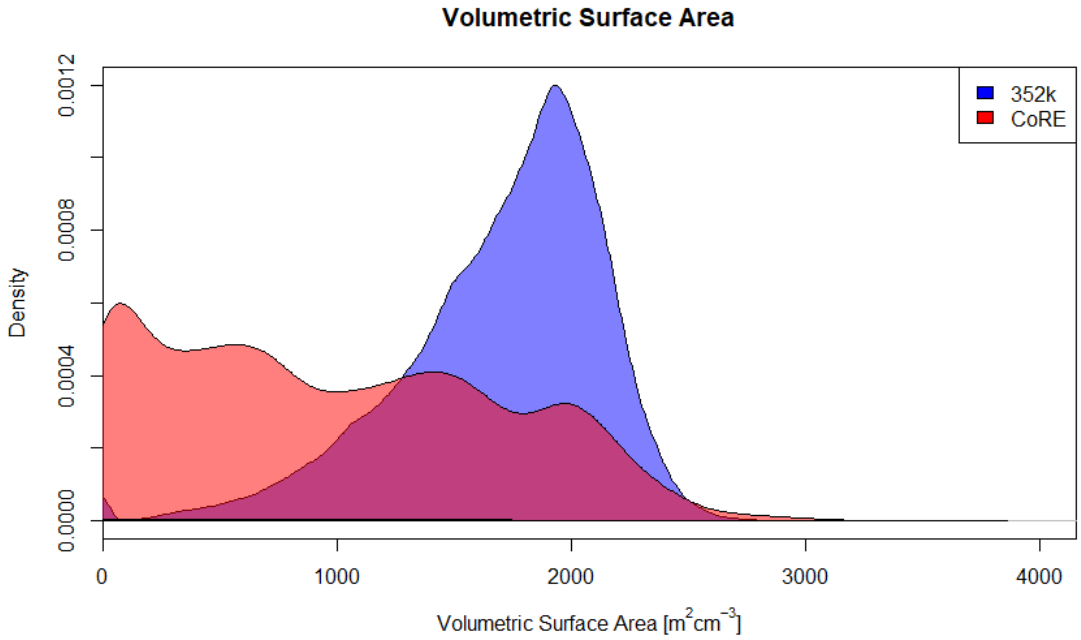

Figure 3.5: The density distribution of the volumetric surface area in m^2cm^{-3} for all MOFs in the 352k (blue) and CoRE (red) databases. Both databases' distributions are similar in width, suggesting similar structural diversity, but the CoRE distribution is more even.

3.2 Working Capacity Prediction

Having thus established that the hypothetical 352k database exhibits sufficient structural diversity when compared to the experimental CoRE database to be considered representative of a large variety of MOFs, the next step is to examine the SVR models constructed for wc prediction. The aim of this section is to answer three basic questions: How well do the proposed wc SVR models fit the training data? How accurately do the wc SVR models predict the wc of MOFs in the test set? And how do the wc SVR models perform as prescreening tools overall? Answering these three questions will permit the evaluation of both wc models on their own, compared to one another, and finally compared to previous work carried out as part of the long-running effort to develop ML-based prescreening tools in the Woo group.

Figure 3.7 illustrates the distribution of wc values across the 352k database. It is readily apparent that the majority of MOFs (276652 MOFs) in the database have low wc values below 1 mmol g^{-1}. The maximum wc value found in the 352k database is 4,731 mmol g^{-1}; there are 39 MOFs whose wc is over 4,0 mmol g^{-1}, 1819 MOFs over 3,0 mmol g^{-1}, and 18154 over 2,0 mmol g^{-1} in the 352k database.

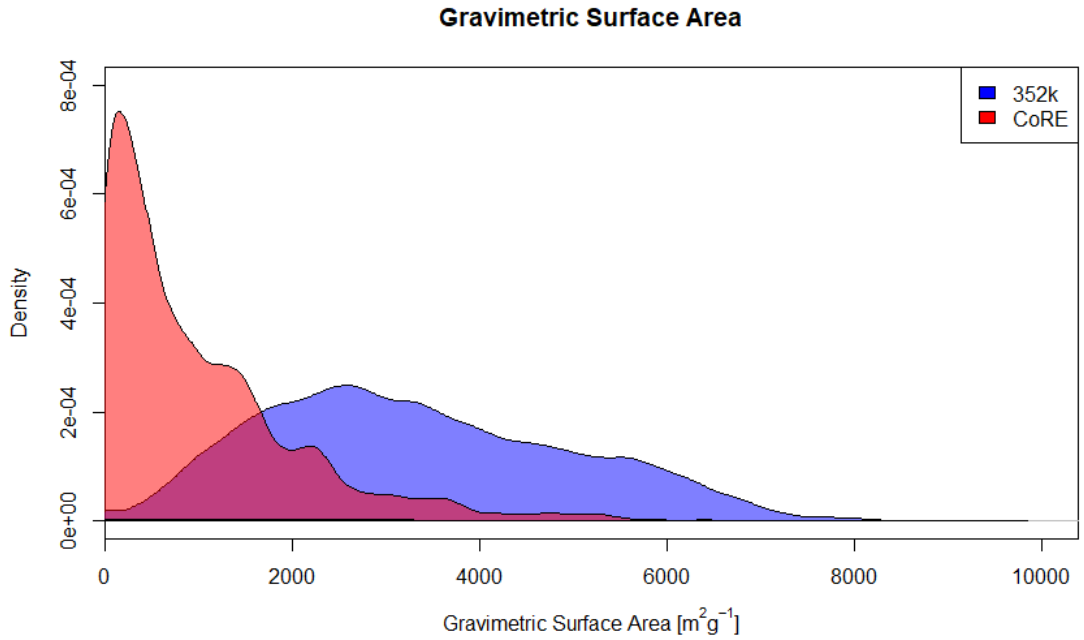

Figure 3.6: The density distribution of the gravimetric surface area in m^2g^{-1} for all MOFs in the 352k (blue) and CoRE (red) databases. The 352k database exhibits a wider distribution of values, indicating that it is more structurally diverse.

The optimum C and γ hyperparameter values for both the wc_BASIC and wc_PCA40 models determined through grid search, as well as their associated 4-fold cross-validation R^2 mean values (R^2_{Xval}) and standard deviations (SD_{Xval}) are listed in table 3.1. In each case, the pair of hyperparameters with the greatest R^2_{Xval} was chosen. R^2_{Xval} for wc_BASIC is 0,897, while R^2_{Xval} of wc_PCA40 is 0,884; both models thus show a good fit to the training data. As the variance between the 4 folds was minimal, it was decided to use the same training set of 10% of the 352k database that had been used for the grid search to train the SVR models for predicting wc on the remaining 90% of the database, i.e. the test set.

Furthermore, table 3.1 shows the R^2 (R^2_{Pred}) and Spearman coefficient (ρ_{Pred}) of prediction for both wc SVR models. These values represent the fitting and ranking accuracy of the models, respectively, and show that both models predict wc with good accuracy. Due to the nature of the prescreening task these models have been developed for, a high ρ_{Pred} is more important than a high R^2_{Pred}: As long as high-performing MOFs are correctly ranked above lower-performing ones, any systematic errors in wc prediction will be corrected by further GCMC screening of the top performers. The difference between wc_BASIC and wc_PCA40 is 0,013 for R^2_{Pred} and 0,003 for ρ_{Pred}, meaning the accuracy of the models is not significantly affected by applying PCA pruning. The computational efficiency, however,

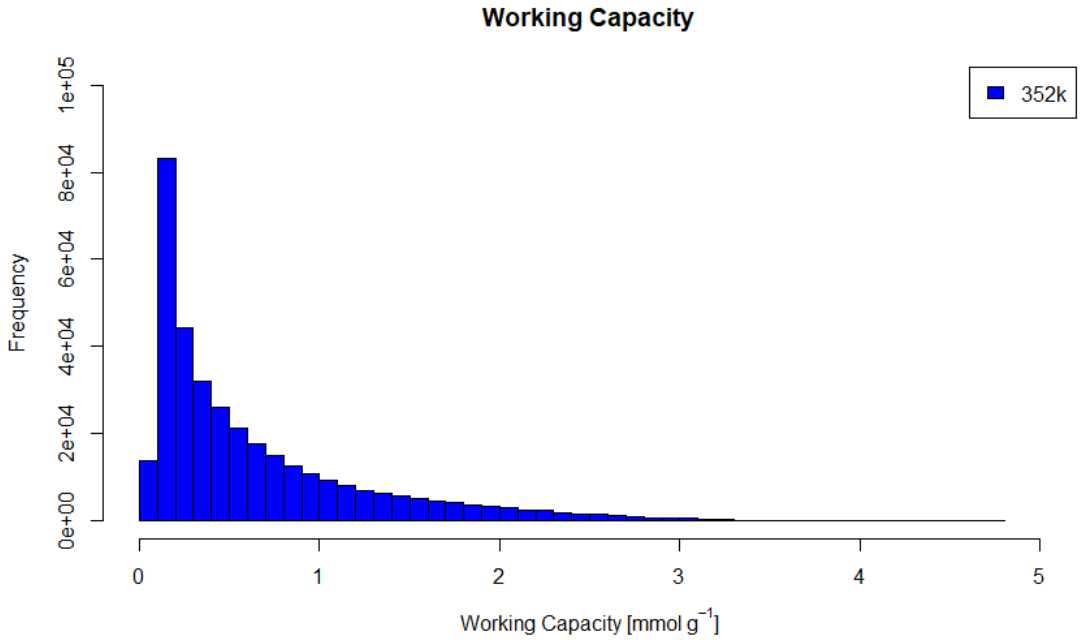

Figure 3.7: The distribution of wc values for all MOFs in the 352k database in mmol g^{-1} (blue).

is much higher for the wc_PCA40 model, which runs training and prediction on the 352k database over five times faster than the wc_BASIC model due to the inclusion of PCA pruning.

Figure 3.8 gives a comprehensive overview of the results of wc prediction by the wc_BASIC model. The heatmap depicts GCMC-calculated vs SVR-predicted wc, with coloured dots representing the MOFs in the test set, corresponding to 90% of the 352k database. The colour gradient is a logarithmic frequency scale, meaning that dark blue dots represent one MOF structure each, while dark red spots represent 10000 MOF structures each. MOFs whose wc was predicted with perfect accuracy are located along the y=x diagonal (grey), as their predicted wc values are equal to the GCMC-calculated target values. The wc of MOFs located above the diagonal was overpredicted, while the wc of MOFs below the diagonal was underpredicted, with distance to the diagonal being proportional to the prediction error.

The heatmap in figure 3.8 has a number of interesting features: Seen as a whole, the distribution of MOFs across the heatmap corresponds to an extremely tall, narrow, and steep frequency peak in the lower left corner (low wc), with an extended ridge sloping down along the diagonal until terminating at approximately (4 mmol g^{-1}; 4 mmol g^{-1}). This peak-and-ridge topography mirrors the distribution of wc values across the 352k database depicted in figure 3.7. The

Table 3.1: The optimum C and γ values and the associated 4-fold cross-validation R^2 mean values (R^2_{Xval}) and standard deviations (SD_{Xval}) for the training set using the wc_BASIC and wc_PCA40 models. Also, R^2 (R^2_{Pred}) and Spearman coefficient (ρ_{Pred}) values for wc prediction on the test set using the wc_BASIC and wc_PCA40 models.

Model	C	γ	R^2_{Xval}	SD_{Xval}	R^2_{Pred}	ρ_{Pred}
wc_BASIC	10,000	0,010	0,897	0,002	0,906	0,950
wc_PCA40	10,000	0,010	0,884	0,003	0,893	0,947

distribution in figure 3.8 also slopes away from the diagonal on both sides, being narrow at low wc values, then widening and remaining at about the same width along most of the diagonal. Keeping in mind the logarithmic nature of the colour scale, it is clear that the vast majority of MOF structures (red to green on the colour scale) are located in the immediate vicinity of the diagonal, meaning their wc values were predicted with a high degree of accuracy. However, this large group of MOFs arranged in an oval shape along the diagonal is characterized by two subtle bulges: Around (1 mmol g^{-1}; 1 mmol g^{-1}), the high-frequency ridge is shifted slightly upwards to overprediction. Meanwhile, above (2 mmol g^{-1}; 2 mmol g^{-1}), the high-frequency ridge is shifted down towards underprediction. The same trend can be observed in the dark blue low-frequency area surrounding the bright high-frequency ridge. This means that the wc_BASIC model has a tendency to slightly overpredict the wc of low-performing MOFs while slightly underpredicting the wc of high-performing MOFs, leading to a somewhat narrower distribution of predicted wc values compared to calculated values. This should be kept in mind when deciding how to select MOFs for further GCMC screening, which will be discussed further below.

Finally, the R^2_{Pred} and ρ_{Pred} values listed in table 3.1 for the wc_BASIC model can be interpreted more easily when compared to the heatmap in figure 3.8. The high R^2_{Pred} of 0,906 is easily explained, considering that most MOFs are located quite close to the diagonal. While a large number of isolated dots on the heatmap are found at a significant distance to the diagonal, these are single outliers, so few in number that their wc prediction errors do not significantly affect R^2_{Pred}. Nevertheless, these outliers serve to underline that the SVR prescreening tools presented in this thesis should not and cannot fully replace screening methods such as GCMC if one wishes to compute wc for MOFs with a high degree of accuracy. These SVR models are only intended to increase the efficiency of HT screening of MOFs. To this end, it is vital that they rank MOFs according to their wc accurately, confirmed by the high ρ_{Pred} of the wc_BASIC model of 0,950. For a given pair of MOFs, this means that the MOF with the higher calculated wc should also have a higher predicted wc, independently of how large the wc difference actually is. Notwithstanding some outliers, this criterion is fulfilled for

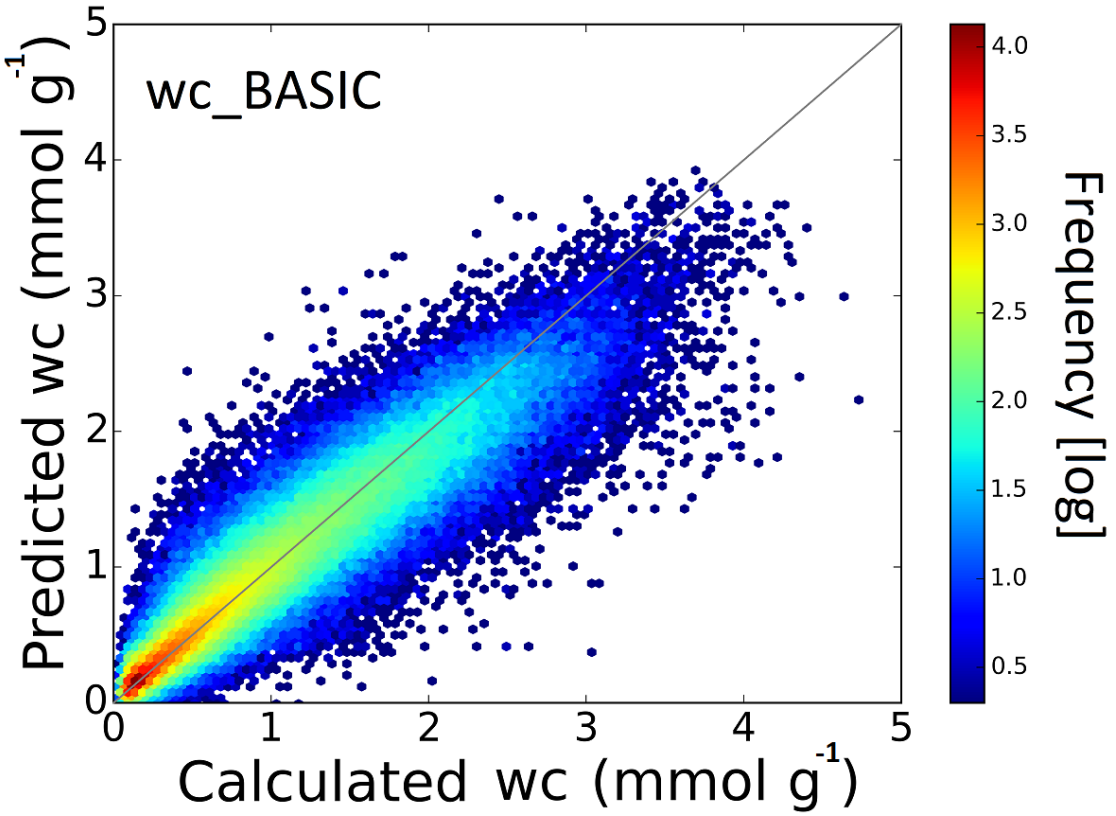

Figure 3.8: A heatmap illustrating predictive performance of the wc_BASIC SVR model. The horizontal axis represents GCMC-calculated wc, the vertical axis represents SVR-predicted wc. MOFs are represented by coloured dots; the color gradient is a logarithmic scale of frequency across the test set. The wc of MOFs located on the y=x diagonal (grey) was predicted accurately; the wc of any MOFs found below the diagonal was underpredicted, while the wc of MOFs above the diagonal was overpredicted.

the vast majority of MOFs in the test set, reflected by the fact that, while not all MOFs in figure 3.8 are located on or immediately next to the diagonal, the overall distribution follows the diagonal in its basic orientation. Thus, a MOF with a higher calculated wc that is further right along the horizontal axis will usually also have a higher predicted wc and be found further up along the vertical axis.

Figure 3.9 shows the calculated-vs-predicted heatmap for the wc_PCA40 model. The distribution of MOFs is depicted in the same manner as in figure 3.8 (see above). In fact, figures 3.9 and 3.8 are extremely similar in most respects. The heatmap in figure 3.9 is also marked by a peak-and-ridge topography originating at low wc values and extending out to about (4 mmol g^{-1}; 4 mmol g^{-1}), with vaguely symmetrical slopes to both sides of the diagonal. Again, two bulges

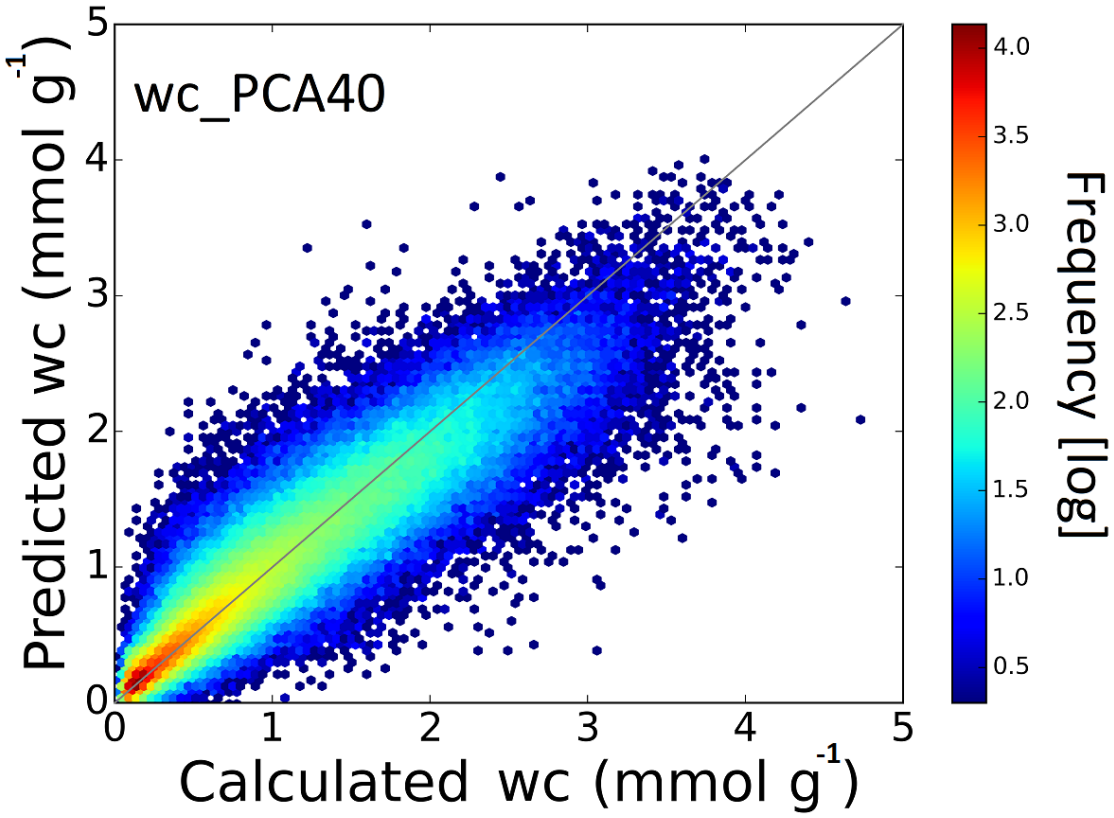

Figure 3.9: A heatmap illustrating predictive performance of the wc_PCA40 SVR model. The horizontal axis represents GCMC-calculated wc, the vertical axis represents SVR-predicted wc. MOFs are represented by coloured dots; the color gradient is a logarithmic scale of frequency across the test set. The wc of MOFs located on the y=x diagonal (grey) was predicted accurately; the wc of any MOFs found below the diagonal was underpredicted, while the wc of MOFs above the diagonal was overpredicted.

indicate some overprediction around (1 mmol g^{-1}; 1 mmol g^{-1}) and underprediction above (2 mmol g^{-1}; 2 mmol g^{-1}). The slightly lower R^2_{Pred} value of 0,893 and the minimally lower ρ_{Pred} value of 0,947 exhibited by the wc_PCA40 model compared to wc_BASIC are reflected by the slightly wider distribution of MOFs on the heatmap in figure 3.9 compared to figure 3.8. It is thus evident that while the fitting accuracy of prediction for the SVR model is lowered marginally by the inclusion of PCA pruning, the ranking accuracy is essentially unaffected.

It is clear from their high R^2_{Pred} and ρ_{Pred} values as well as from the heatmaps in figures 3.8 and 3.9 that both wc SVR models predict the wc of MOFs with a high degree of accuracy. To gauge how well wc_BASIC and wc_PCA40 function as prescreening tools, it was decided to measure the recovery rate of Top Performers

Table 3.2: Having defined the 1000 top MOFs ranked according to their GCMC-calculated wc as Top Performers, this table gives the number N of Top Performers recovered when various sample sizes of top SVR-predicted wc MOFs are considered to have PASSED prescreening, for the wc_BASIC and wc_PCA40 models, respectively. Sample sizes are given in number of MOFs (corresponding percentage of test set).

Sample PASSED (% of test set)	N/1000 Top Performers for wc_BASIC	wc_PCA40
1000 (0,3%)	474	447
3170 (1,0%)	734	702
6340 (2,0%)	847	827
9510 (3,0%)	909	888
12680 (4,0%)	943	933
15850 (5,0%)	962	951

relative to the size of sample PASSED through prescreening. To this end, the MOFs in the test set were ordered according to their GCMC-calculated wc, and the top 1000 MOFs were defined as the Top Performers subset. Then, the MOFs in the test set were reordered according to their SVR-predicted wc, and various sample sizes (e.g. 3170 MOFs, equal to 1% of the test set) of the highest performers from this list were checked for how many of the Top Performers subset (N/1000) had been recovered within these samples. The results of this analysis are listed in table 3.2 and graphically represented in figure 3.10 for both models.

This N/1000 analysis is essentially a simulation of actual prescreening, where a certain number of top results is passed on for further GCMC screening (PASSED) and the rest are discarded. This is one way of defining a prescreening threshold; in practice, it may be preferable to set a cutoff wc value. As shown in table 3.2, 47% of Top Performers are recovered when the bottom 99,7% of MOFs according to wc_BASIC are discarded before further screening. In this case, the percentage of Top Performers among the sample PASSED through prescreening is also 47%. Figure 3.10 illustrates that the percentage of recovered Top Performers quickly rises as more MOFs are passed on to further screening, up to 96% recovery when 5% are PASSED (and 95% discarded). In this case, Top Performers only make up 6% of the PASSED sample, and returns diminish even further beyond this point. Much of the same holds true for wc_PCA40, with just under 45% recovery at 0,3% PASSED and 95% recovery at 5% PASSED. Figure 3.10 also clearly shows that the difference in prescreening performance between the two wc models is negligible, while wc_PCA40 is by far the more computationally efficient of the two.

As a final note, the results of the N/1000 analysis also show that the underprediction of the wc of high-performing MOFs evidenced in the heatmaps of the

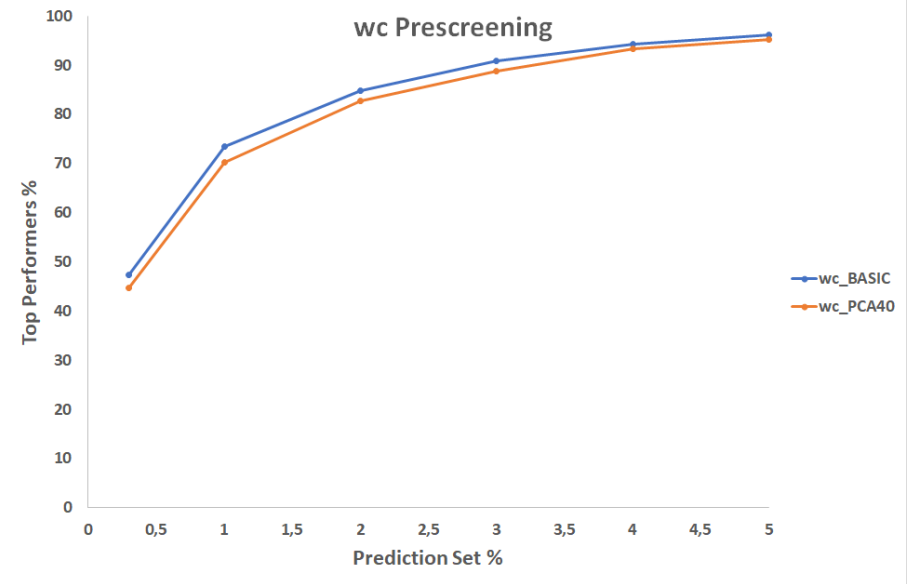

Figure 3.10: A graphic depiction of the results from table 3.2. The horizontal axis represents the top SVR prediction results PASSED through prescreening as a percentage of the test set. The vertical axis represents the number N of Top Performers recovered as a percentage of the total number of Top Performers (1000). Results for the wc_BASIC model in blue, wc_PCA40 in orange. N increases with greater sample sizes PASSED, up to ca. 95% of Top Performers recovered for 5% sample size PASSED.

two SVR models does not limit the prescreening performance of the models in terms of identifying high performers. As discussed above, the accuracy of wc prediction for any given MOF or set of MOFs is much less important than the ranking accuracy of the SVR model as a whole. Thus, the excellent results of the N/1000 analysis are in good agreement with the extremely high ρ_{Pred} of both wc SVR models, indicating both models perform prescreening very well.

To sum up, both wc models fit the training data very well. Both wc models also predict the wc of MOFs in the 352k database with a high degree of fitting accuracy, although some overprediction of low performers and underprediction of high performers is evident. Both models' ranking accuracy is excellent, and both models function well as prescreening tools to predict the wc of MOFs in post-combustion carbon capture. However, wc_PCA40 clearly outperforms wc_BASIC in terms of computational efficiency, running over five times faster due to the inclusion of PCA pruning.

3.3 Selectivity Prediction

In the following section, the SVR models for Sel prediction are examined so as to answer the same three questions posed in the previous section, i.e.: How well do the proposed Sel SVR models fit the training data? How accurately do the Sel SVR models predict the Sel of MOFs in the test set? And how do the Sel SVR models perform as prescreening tools overall? Again, the aim is to evaluate both Sel models according to their own merits as well as in comparison with one another and previous work carried out by the Woo group.

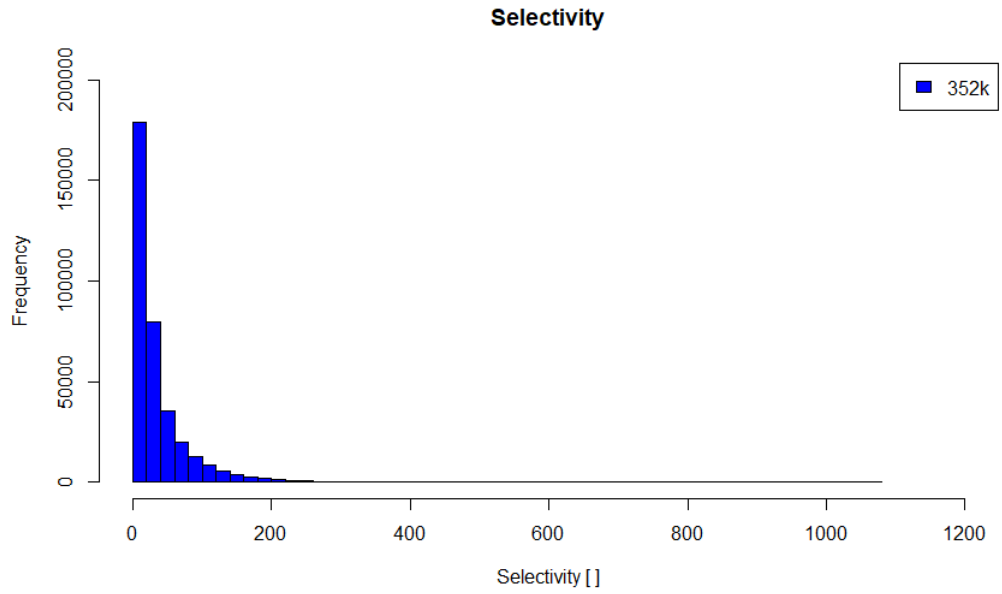

Figure 3.11: The distribution of CO_2/N_2 Sel values for all MOFs in the 352k database (blue).

The distribution of CO_2/N_2 Sel values across the 352k database is shown in figure 3.11. Even more so than was the case for wc, the vast majority of MOFs (326825 MOFs) clearly perform relatively poorly, having Sel values below 100. Nevertheless, there is a significant number of high-performing MOFs, with the max Sel value found in the database being 1069,036. That is the only Sel value over 800, while there are 15 MOFs over 400, and 3291 MOFs over 200 in the 352k database.

Table 3.3 lists the optimum C and γ hyperparameter values for the Sel_BASIC and Sel_PCA40 models along with their associated cross-validation R^2 mean values (R^2_{Xval}) and standard deviations (SD_{Xval}). Optimal hyperparameters were once again chosen by highest R^2_{Xval}. R^2_{Xval} for Sel_BASIC is 0,886, while R^2_{Xval} of Sel_PCA40 is 0,875; both models thus show a good fit to the training data. Due to the minimal variance between cross-validation folds, SVR model train-

Table 3.3: The optimum C and γ values and the associated 4-fold cross-validation R^2 mean values (R^2_{Xval}) and standard deviations (SD_{Xval}) for the training set using the Sel_BASIC and Sel_PCA40 models. Also, R^2 (R^2_{Pred}) and Spearman coefficient (ρ_{Pred}) values for Sel prediction on the test set using the Sel_BASIC and Sel_PCA40 models.

Model	C	γ	R^2_{Xval}	SD_{Xval}	R^2_{Pred}	ρ_{Pred}
Sel_BASIC	10,000	0,010	0,886	0,005	0,895	0,956
Sel_PCA40	10,000	0,010	0,875	0,005	0,885	0,953

ing proceeded with the same 10% training set used for the grid search, while the remaining 90% of the 352k database served as the test set.

The R^2 (R^2_{Pred}) and Spearman coefficient (ρ_{Pred}) of prediction for both Sel models are also reproduced in table 3.3. Both models predict Sel with good fitting and ranking accuracy, though R^2_{Pred} is somewhat lower in both cases than for the wc models, while ρ_{Pred} is marginally higher. The difference in R^2_{Pred} between the two Sel models is 0,010, and the difference in ρ_{Pred} is 0,003, with Sel_BASIC being the more accurate in both cases. Once again, the accuracy is not very much adversely affected by PCA pruning, with Sel_PCA40 offering computation times shorter than those of Sel_BASIC by up to a factor of 8 due to the use of pruning.

The calculated-vs-predicted heatmap of Sel_BASIC is depicted in figure 3.12 in the same manner as the wc models were in figures 3.8 and 3.9. The Sel_BASIC heatmap is characterized by an extreme peak at low Sel values, as well as a short, narrow ridge that quickly slopes down along the y=x diagonal before terminating around (300; 300). This topography reflects the distribution of Sel values shown in figure 3.11, which is strongly focused toward low values. The high-frequency area around the diagonal, coloured red to green, forms an oval that once again shows two bulges: One upward around (50; 50), indicating overprediction, and one downward above (100; 100), indicating underprediction. The blue low-frequency area forms a roughly triangular shape, with the tip at (0; 0) and the base located in the upper right quadrant of the heatmap. There is a significant number of isolated outliers, more so than in the wc models, showing that the Sel of high-performing MOFs is quite likely to be underpredicted by Sel_BASIC.

The lower R^2_{Pred} of Sel_BASIC of 0,895 compared to both of the wc models can be explained by the large number of outliers, though once again the vast majority of MOFs are quite close to the diagonal. However, the $\rho_{Pred}R$ value of 0,956 is slightly higher even than that of wc_BASIC. This excellent ranking accuracy is reflected by the generally diagonal orientation of the MOF distribution on the heatmap.

Figure 3.13 shows the heatmap for the Sel_PCA40 model, which is very similar to the heatmap for Sel_BASIC. It also features the peak-and-ridge topography

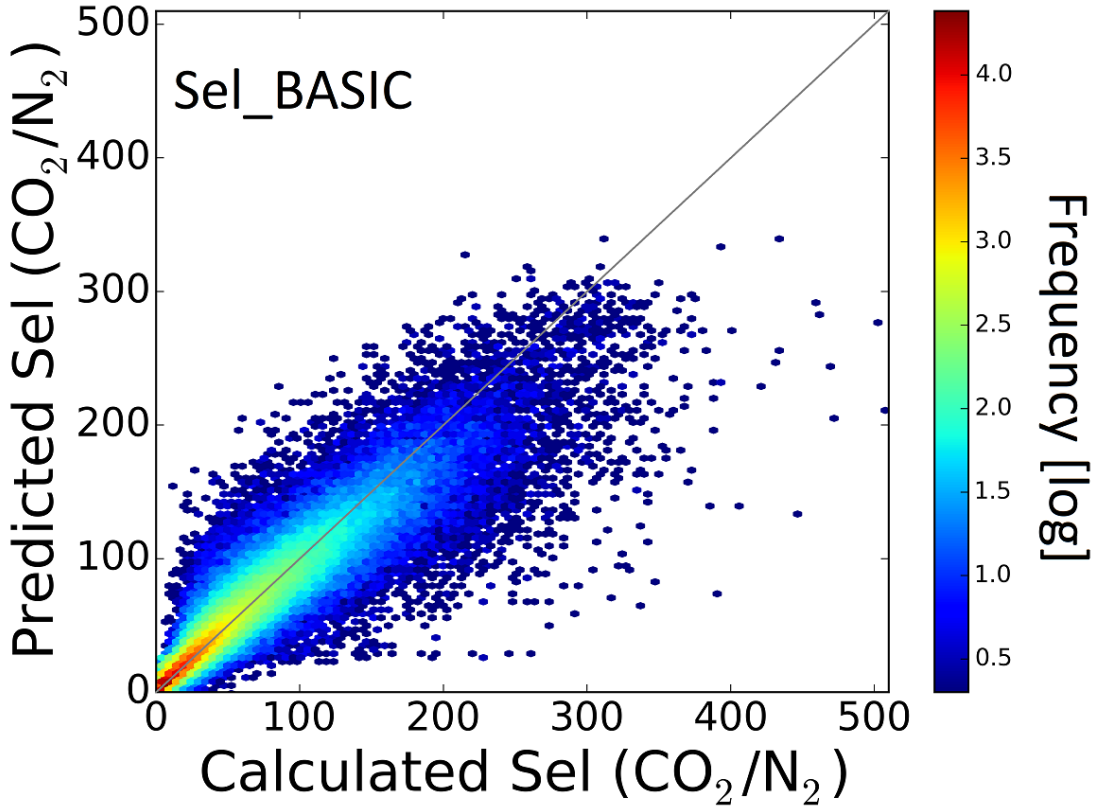

Figure 3.12: A heatmap illustrating predictive performance of the Sel_BASIC SVR model. The horizontal axis represents GCMC-calculated Sel, the vertical axis represents SVR-predicted Sel. MOFs are represented by coloured dots; the color gradient is a logarithmic scale of frequency across the test set. The Sel of MOFs located on the y=x diagonal (grey) was predicted accurately; the Sel of any MOFs found below the diagonal was underpredicted, while the Sel of MOFs above the diagonal was overpredicted.

with slopes to both sides of the diagonal and two bulges, the first around (50; 50) being indicative of over-, the second above (100; 100) of underprediction. The distribution of MOFs is slightly wider than in the Sel_BASIC model, reflected by the slightly lower R^2_{Pred} of 0,885. $\rho_{Pred}R$ is also marginally lower, at 0,953. Again, while the fitting accuracy of prediction is very slightly lowered by the inclusion of PCA pruning, the ranking accuracy remains practically the same.

Judging from the R^2_{Pred} and $\rho_{Pred}R$ values and the distribution of MOFs on the heatmaps in figure 3.12 and 3.13, both Sel SVR models are highly accurate in predicting the Sel of MOFs. For the N/1000 metric, the Top Performers subset was redefined by ranking MOFs according to GCMC-calculated Sel, and the test set was ordered by SVR-predicted Sel for sampling. Table 3.3 shows the results of the

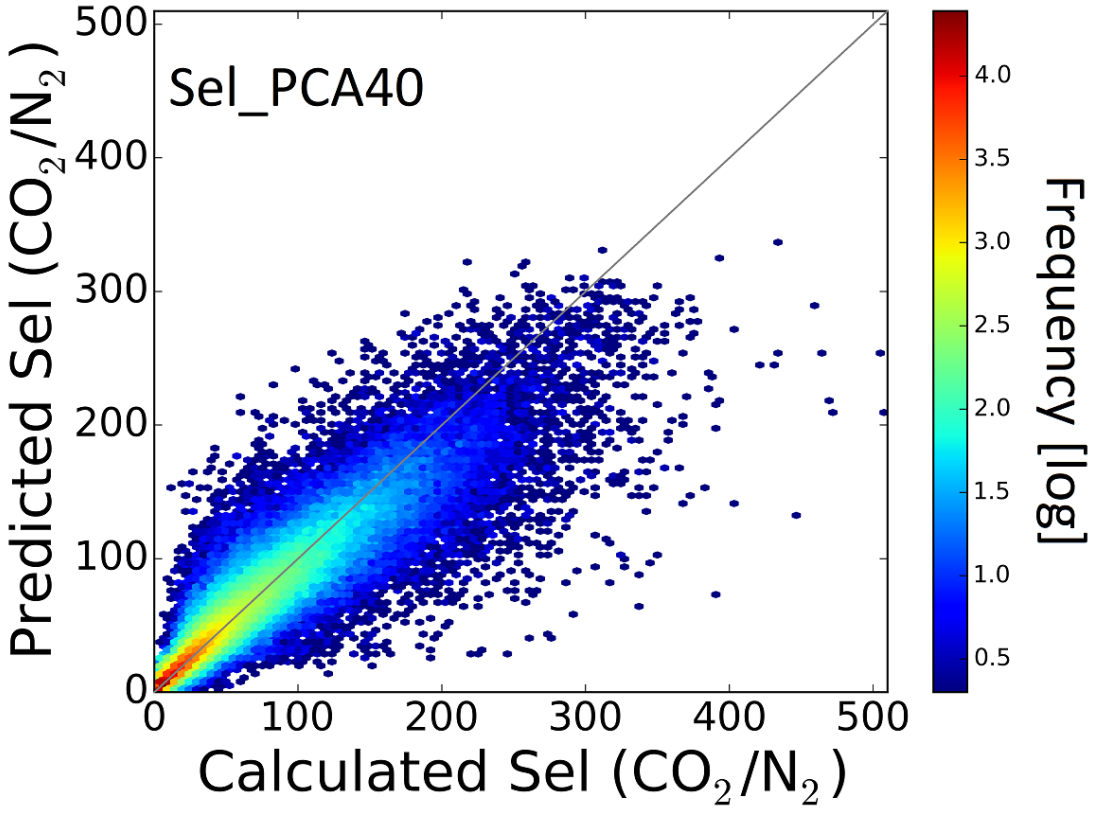

Figure 3.13: A heatmap illustrating predictive performance of the Sel_PCA40 SVR model. The horizontal axis represents GCMC-calculated Sel, the vertical axis represents SVR-predicted Sel. MOFs are represented by coloured dots; the color gradient is a logarithmic scale of frequency across the test set. The Sel of MOFs located on the y=x diagonal (grey) was predicted accurately; the Sel of any MOFs found below the diagonal was underpredicted, while the Sel of MOFs above the diagonal was overpredicted.

prescreening analysis: When 0,3% of the test set are PASSED, Sel_BASIC recovers 52% of the Top Performers, while Sel_PCA40 recovers 50%; the percentage of Top Performers in the PASSED sample corresponds accordingly. This quickly rises to 96% for Sel_BASIC and Sel_PCA40 when 5% are passed. However, in this case, Top Performers again only constitute 6% of the sample PASSED, with diminishing returns from passing more of the test set, as can be seen in figure 3.14. The difference in prescreening performance is even smaller between the two Sel models than it was between the wc models, leaving Sel_PCA40 as the clear favourite due to the model's greater computational efficiency.

Again, the N/1000 analysis demonstrates that the underprediction of the Sel of high-performing MOFs, more prominent in the Sel than in the wc SVR models,

Table 3.4: Having defined the 1000 top MOFs ranked according to their GCMC-calculated Sel as Top Performers, this table gives the number N of Top Performers recovered when various sample sizes of top SVR-predicted Sel MOFs are considered to have PASSED prescreening, for the Sel_BASIC and Sel_PCA40 models, respectively. Sample sizes are given in number of MOFs (corresponding percentage of test set).

Sample PASSED (% of test set)	N/1000 Top Performers for Sel_BASIC	Sel_PCA40
1000 (0,3%)	515	504
3170 (1,0%)	771	764
6340 (2,0%)	884	878
9510 (3,0%)	927	918
12680 (4,0%)	948	943
15850 (5,0%)	960	955

does not have a limiting effect on the prescreening performance of the Sel models when it comes to identifying high performers. Both Sel SVR models function well as prescreening tools, as evidenced by the results of the N/1000 analysis as well as their extremely high ρ_{Pred}.

Summing up, both Sel models fit the training data very well. Both models predict the Sel of MOFs with high fitting accuracy, although some overprediction of low performers and underprediction of high performers can again be observed. The ranking accuracy of both Sel models is outstanding, and both models function very well as prescreening tools to predict the Sel of MOFs in post-combustion carbon capture. Sel_PCA40 greatly outperforms Sel_BASIC in terms of computational efficiency, running over eight times faster due to the use of pruning.

3.4 Improving Prescreening Efficiency

The four SVR models discussed above (wc_BASIC, wc_PCA40, Sel_BASIC, Sel_PCA40) exhibit excellent predictive performance both in terms of fitting and ranking accuracy, and their utility as prescreening tools for HT screening of a large database of hypothetical MOF structures to identify high performers has been demonstrated. Going further, the two models utilizing PCA-based pruning (wc_PCA40, Sel_PCA40) have demonstrated outstanding computational efficiency. It remains to be examined to what extent the present work improves upon and goes beyond previous iterations of prescreening tools for MOFs developed in the Woo group.

In their 2013 paper, Fernandez, Trefiak, and Woo [5] used the AP-RDF descriptor with SVR to predict the CO_2 uptake at 0,1 bar of 83000 MOFs from the

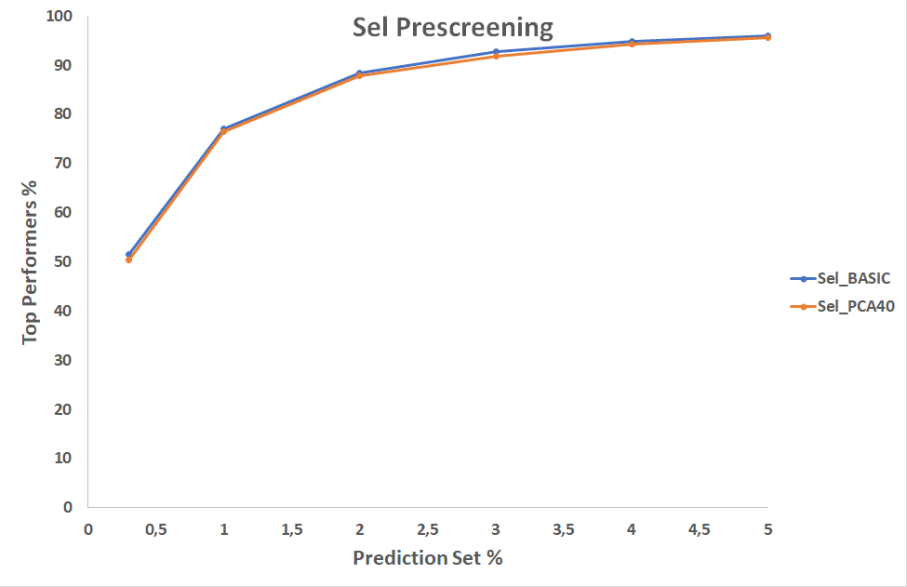

Figure 3.14: A graphic depiction of the results from table 3.4. The horizontal axis represents the top SVR prediction results PASSED through prescreening as a percentage of the test set. The vertical axis represents the number N of TOP Performers recovered as a percentage of the total number of Top Performers (1000). Results for the Sel_BASIC model in blue, Sel_PCA40 in orange. N increases with greater sample sizes PASSED, up to ca. 95% of Top Performers recovered for 5% sample size PASSED.

database of Wilmer et al. They used a grid search approach with 3-fold cross-validation to determine the hyperparameters of their SVR model, attaining an R^2_{Xval} of 0,69. Their R^2_{Pred} was 0,673. While this study is not directly comparable to the present work, mainly because it focuses on predicting CO_2 uptake rather than wc, the difference in both goodness of fit (R^2_{Xval}; wc_BASIC: 0,897; wc_PCA40: 0,884) and prediction accuracy (R^2_{Pred}; wc_BASIC: 0,906; wc_PCA40: 0,893) is very noticeable. While adding a normalization parameter has certainly contributed to making the SVR models more accurate, it is also possible that using a more structurally diverse database (1166 vs 6 net topologies) has improved the training fit. The training set used by Fernandez, Trefiak, and Woo [5] contained 58000 MOFs from the low-diversity database of Wilmer et al. [21], while the training set used in the present work consists of just 35218 MOFs from the much higher-diversity 352k database. While using a larger training set is generally said to improve the predictive capabilities of ML models, this is only true as long as adding additional training data expands the chemical space (e.g. the variety of MOF structures) covered by the model, thereby extending its capability for interpolation. If, however, additional training data comes from a part of chemical space already sufficiently covered by the SVR model, this data will increase the danger of overfitting without providing much added information.

Another point of comparison is the 2014 paper by Fernandez et al. [6], in which the authors presented an SVM-based classifier that sorted MOFs into high and low performers for post-combustion carbon capture based on CO_2 uptake at 0,15 bar. Their reported area under curve (a measure of ranking accuracy for classifiers) was 0,979. Furthermore, the authors used an N/1000 metric to analyze the predictive performance of their classifier, finding that it recovered 790/1000 Top Performers when 5% of the test set were PASSED, 945/1000 at 10% PASSED, and 999/1000 at 28% passed. As this paper represents the Woo group's most recent publication on ML-based prescreening of MOFs for post-combustion carbon capture, the N/1000 analysis in this work was carried out partly to compare the performance of the prescreening tools proposed herein to the existing classifier. While both the 2014 paper and the present work show excellent ranking accuracy, the prescreening efficiency of the four models discussed in this thesis is about double that of the classifier, with slightly higher numbers of Top Performers (wc_BASIC: 962; wc_PCA40: 951; Sel_BASIC: 960; Sel_PCA40: 955) recovered at 5% PASSED than the classifier reportedly recovered at 10% PASSED. Again, the main differences between the two studies are that the 2014 paper used a non-normalized APRDF descriptor and a less structurally diverse database, consisting of 324500 MOFs made from a combination of 66 SBUs and 19 functional groups, while the present work utilizes a database of 352178 MOFs constructed from a combination of 198 SBUs (23 inorganic and 175 organic) and 50 functional groups. It is possible that these changes in descriptor and database improved the fit obtained by the SVR models, perhaps particularly with respect to high-performing MOFs, due to similar reasons as those discussed in the preceding paragraph.

Finally, Dureckova's Master's thesis [7] used nAP-RDF as well as geometric descriptors with SVR to predict the wc and Sel of MOFs for *pre-combustion* carbon capture. For wc, she reported an R^2_{Xval} of 0,935 and an R^2_{Pred} of 0,944, while for Sel, she reported an R^2_{Xval} of 0,862 and an R^2_{Pred} of 0,876. Using a product of the normalized wc and Sel as a unified performance metric, she managed to obtain 309/1000 Top Performers at 0,3% of test set PASSED. While Dureckova's work is aimed at a different application, her methods were very similar to those employed in the present work, and her results are therefore comparable to some extent. For MOF wc, her R^2_{Xval} and R^2_{Pred} are noticeably higher than those of wc_BASIC (0,897; 0,906) or of wc_PCA40 (0,884; 0,893). Meanwhile for Sel, her R^2_{Xval} and R^2_{Pred} are somewhat lower than those of Sel_BASIC (0,886; 0,895) or of Sel_PCA40 (0,875; 0,885). Due to the difference in metric used, Dureckova's N/1000 analysis cannot be compared to that presented here in a meaningful way. It appears therefore that the SVR models proposed by Dureckova and in the present work are similar in both training and prediction accuracy. However, the inclusion of PCA pruning has opened up the possibility of prescreening at much greater speed and comparable accuracy, effectively multiplying the prescreening efficiency of the models presented here compared to previous iterations.

In summary, the SVR models developed in the course of this research project

represent a substantial increase in accuracy and N/1000 performance over previous prescreening tools used in the Woo group for predicting MOF performance in post-combustion carbon capture [5, 6]. This is likely due to the use of the newly adopted nAP-RDF descriptor and the hugely structurally diverse 352k database. Furthermore, the wc_PCA40 and Sel_PCA40 models have demonstrated that the inclusion of PCA-based pruning can vastly increase the computational efficiency of SVR-based prescreening compared to wc_BASIC and Sel_BASIC. It is to be hoped that the same improvements can be achieved by applying PCA pruning to pre-combustion capture.

Chapter 4

Conclusion

CCS technology will play an important role in combating global warming by reducing CO_2 emissions, primarily from the energy sector [1]. Adsorption-based post-combustion capture using porous solids is an important area of ongoing research. Within this field, MOFs are of particular interest because of their record-breaking surface areas and pore volumes, and also because MOF structures can be tuned to meet various sets of adsorption requirements [2]. Due to the vast number of possible MOF structures, computational HT screening is being increasingly used to accelerate the discovery of new high-performing MOFs. HT screening requires three things: firstly, a large, diverse database of computation-ready MOF structures; both experimental and hypothetical MOF databases are used. Secondly, a well-defined set of performance parameters by which to rank MOFs - the most widely accepted being wc and Sel. Finally, a screening method that can efficiently compute wc and Sel for all MOFs in a database, such as GCMC. Screening large databases using GCMC is associated with high computational costs [3]. Therefore, prescreening methods that efficiently select promising MOFs for GCMC screening are a key area of research; such methods are the focus of a long-running research project in the Woo group [4, 5, 6, 7].

In the course of my research project, a database of 352k hypothetical MOF structures was generated from 23 inorganic and 175 organic SBUs, arranged in 1166 net topologies and studded with 50 different functional groups. The MOFs' wc and Sel values were determined using GCMC. The 352k database was characterized by calculating the MOFs' dominant pore size, largest accessible pore size, void fraction, density, volumetric surface area, and gravimetric surface area, which were also used as geometric descriptors. The database was also compared to the CoRE database of experimental MOF structures for the sake of contextualization. Additionally, 339 nAP-RDF descriptor values were computed for each MOF. PCA-based pruning was employed to increase computational efficiency. Thus, four SVR prediction models were generated: wc_BASIC, wc_PCA40, Sel_BASIC, Sel_PCA40. The hyperparameters C and γ were determined for all models using a grid-based approach. Finally, all models were trained and evaluated according to their training and prediction fitting accuracy, prediction rank-

ing accuracy, and prescreening performance.

The results of the database characterization carried out through analysis of the 352k database's textural property distributions show that the 352k database exhibits large structural diversity. The database shows wide density distributions for gravimetric surface area, void fraction, as well as dominant and largest accessible pore size. However, the distributions of the volumetric surface area and mass density are narrower. This is most likely caused by the large number of MOFs with unrealistically large pore diameters and void fractions present in the 352k database. Overall, the density distributions of the 352k database compare favourably to the distributions of the CoRE database of experimental MOF structures, underlining that the 352k database contains a large variety of MOFs.

Both of the wc SVR models, wc_BASIC and wc_PCA40, have been shown to fit the training data very well. Both models also predict the wc of MOFs in the 352k database with a high degree of fitting accuracy, although some overprediction of low performers and underprediction of high performers is evident. Both wc models' ranking accuracy is excellent, and both models function very well as prescreening tools to predict the wc of MOFs in post-combustion carbon capture. However, wc_PCA40 clearly outperforms wc_BASIC in terms of computational efficiency, running over five times faster.

Both Sel SVR models, Sel_BASIC and Sel_PCA40, fit the training data very well. Both models predict the Sel of MOFs with high fitting accuracy, although some overprediction of low performers and underprediction of high performers can again be observed. The ranking accuracy of both Sel models is outstanding, and both models function very well as prescreening tools to predict the Sel of MOFs in post-combustion carbon capture. Sel_PCA40 greatly outperforms Sel_BASIC in terms of computational efficiency, running over eight times faster.

In conclusion, all four SVR models developed in the course of this research project demonstrate a substantial increase in accuracy and N/1000 performance over previous prescreening tools used in the Woo group for predicting MOF performance in post-combustion carbon capture [5, 6]. This is most likely caused by the adoption of the new nAP-RDF descriptor and the hugely structurally diverse 352k database. Finally, the wc_PCA40 and Sel_PCA40 models have shown that PCA-based pruning can vastly increase the computational efficiency of SVR-based prescreening compared to the wc_BASIC and Sel_BASIC models that did not employ pruning.

Having proven the effectiveness of the SVR models as prescreening tools by training and testing them on the 352k database, the next step would be to apply them to an even larger database of MOFs in their intended function. If used in the context of a HT screening materials discovery study focusing on post-combustion carbon capture via PSA, the SVR models would allow the user to eliminate some 95% of a given database of MOFs while retaining a significant proportion of high performers. The wc_PCA40 and Sel_PCA40 models in particular offer high-efficiency, accurate prediction of important adsorptive performance parameters. Thus, GCMC simulations could be limited to just 1/20 of the database, allow-

ing for greater accuracy through longer GCMC runs. The most promising results from GCMC screening could then be investigated further using e.g. DFT, until a small selection of a handful high-performing candidates fulfilling a variety of performance goals is put together. This selection could then be used to guide experimental efforts, hopefully resulting in the successful synthesis of several new MOFs with excellent adsorptive properties.

Bibliography

[1] M. Bui, C. S. Adjiman, A. Bardow, E. J. Anthony, A. Boston, S. Brown, P. S. Fennell, S. Fuss, A. Galindo, L. A. Hackett, J. P. Hallett, H. J. Herzog, G. Jackson, J. Kemper, S. Krevor, G. C. Maitland, M. Matuszewski, I. S. Metcalfe, C. Petit, G. Puxty, J. Reimer, D. M. Reiner, E. S. Rubin, S. A. Scott, N. Shah, B. Smit, J. P. M. Trusler, P. Webley, J. Wilcox, and N. Mac Dowell, "Carbon capture and storage (CCS): the way forward," *Energy Environ. Sci.*, vol. 11, pp. 1062–1176, 2018.

[2] K. Sumida, D. L. Rogow, J. A. Mason, T. M. McDonald, E. D. Bloch, Z. R. Herm, T.-H. Bae, and J. R. Long, "Carbon Dioxide Capture in Metal–Organic Frameworks," *Chem. Rev.*, vol. 112, no. 2, pp. 724–781, 2012.

[3] Y. J. Colón and R. Q. Snurr, "High-throughput computational screening of metal–organic frameworks," *Chem. Soc. Rev.*, vol. 43, pp. 5735–5749, 2014.

[4] M. Fernandez, T. K. Woo, C. E. Wilmer, and R. Q. Snurr, "Large-Scale Quantitative Structure–Property Relationship (QSPR) Analysis of Methane Storage in Metal–Organic Frameworks," *J. Phys. Chem. C*, vol. 117, no. 15, pp. 7681–7689, 2013.

[5] M. Fernandez, N. R. Trefiak, and T. K. Woo, "Atomic Property Weighted Radial Distribution Functions Descriptors of Metal–Organic Frameworks for the Prediction of Gas Uptake Capacity," *J. Phys. Chem. C*, vol. 117, no. 27, pp. 14095–14105, 2013.

[6] M. Fernandez, P. G. Boyd, T. D. Daff, M. Z. Aghaji, and T. K. Woo, "Rapid and Accurate Machine Learning Recognition of High Performing Metal Organic Frameworks for CO2 Capture," *J. Phys. Chem. Lett.*, vol. 5, no. 17, pp. 3056–3060, 2014.

[7] H. Dureckova, *Robust Machine Learning QSPR Models for Recognizing High Performing MOFs for Pre-Combustion Carbon Capture and Using Molecular Simulation to Study Adsorption of Water and Gases in Novel MOFs*. MSc Thesis, University of Ottawa, 2018.

[8] A. U. Czaja, N. Trukhan, and U. Müller, "Industrial applications of metal–organic frameworks," *Chem. Soc. Rev.*, vol. 38, pp. 1284–1293, 2009.

[9] O. Ivanciuc, *Applications of Support Vector Machines in Chemistry*, ch. 6, pp. 291–400. John Wiley & Sons, Ltd, 2007.

[10] E. J. Granite and H. W. Pennline, "Photochemical Removal of Mercury from Flue Gas," *Ind. Eng. Chem. Res.*, vol. 41, no. 22, pp. 5470–5476, 2002.

[11] IPCC, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 2013.

[12] J. C. Orr, V. J. Fabry, O. Andmont, L. Bopp, S. C. Doney, R. A. Feely, A. Gnanadesikan, N. Gruber, A. Ishida, F. Joos, R. M. Key, K. Lindsay, E. Maier-Reimer, R. Matear, P. Monfray, A. Mouchet, R. G. Najjar, G.-K. Plattner, K. B. Rodgers, C. L. Sabine, J. L. Sarmiento, R. Schlitzer, R. D. Slater, I. J. Totterdell, M.-F. Weirig, Y. Yamanaka, and A. Yool, "Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms," *Nature*, vol. 437, p. 681–686, 2005.

[13] International Energy Agency, *CO2 Emissions from Fuel Combustion 2017*. 2017.

[14] B. S. Koelbl, M. A. van den Broek, A. P. C. Faaij, and D. P. van Vuuren, "Uncertainty in Carbon Capture and Storage (CCS) deployment projections: a cross-model comparison exercise," *Clim. Change*, vol. 123, pp. 461–476, Apr 2014.

[15] J. Wilcox, P. C. Psarras, and S. Liguori, "Assessment of reasonable opportunities for direct air capture," *Environ. Res. Lett.*, vol. 12, no. 6, p. 065001, 2017.

[16] N. MacDowell, N. Florin, A. Buchard, J. Hallett, A. Galindo, G. Jackson, C. S. Adjiman, C. K. Williams, N. Shah, and P. Fennell, "An overview of co2 capture technologies," *Energy Environ. Sci.*, vol. 3, pp. 1645–1669, 2010.

[17] H. Li, M. Eddaoudi, M. O'Keeffe, and O. M. Yaghi, "Design and Synthesis of an Exceptionally Stable and Highly Porous Metal-Organic Framework," *Nature*, vol. 402, pp. 276–279, 1999.

[18] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, "The High-Throughput Highway to Computational Materials Design," *Nat. Mater.*, vol. 12, pp. 191–201, 2013.

[19] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward, "The Cambridge Structural Database," *Acta Crystallogr., Sect. B*, vol. 72, pp. 171–179, Apr 2016.

[20] Y. G. Chung, J. Camp, M. Haranczyk, B. J. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O. K. Farha, D. S. Sholl, and R. Q. Snurr, "Computation-Ready, Experimental Metal–Organic Frameworks: A Tool To Enable High-Throughput Screening of Nanoporous Crystals," *Chem. Mater.*, vol. 26, no. 21, pp. 6185–6192, 2014.

[21] C. E. Wilmer, M. Leaf, C. Y. Lee, O. K. Farha, B. G. Hauser, J. T. Hupp, and R. Q. Snurr, "Large-Scale Screening of Hypothetical Metal-Organic Frameworks," *Nat. Chem.*, vol. 4, pp. 83–89, 2011.

[22] B. J. Sikora, R. Winnegar, D. M. Proserpio, and R. Q. Snurr, "Textural properties of a large collection of computationally constructed MOFs and zeolites," *Microporous Mesoporous Mater.*, vol. 186, pp. 207 – 213, 2014.

[23] P. G. Boyd and T. K. Woo, "A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory," *CrystEngComm*, vol. 18, pp. 3777–3792, 2016.

[24] T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza, and M. Haranczyk, "Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials," *Microporous Mesoporous Mater.*, vol. 149, no. 1, pp. 134 – 141, 2012.

[25] L. Sarkisov and A. Harrison, "Computational structure characterisation tools in application to ordered and disordered porous materials," *Mol. Simul.*, vol. 37, no. 15, pp. 1248–1257, 2011.

[26] J. R. Karra and K. S. Walton, "Molecular Simulations and Experimental Studies of CO2, CO, and N2 Adsorption in Metal-Organic Frameworks," *J. Phys. Chem. C*, vol. 114, no. 37, pp. 15735–15740, 2010.

[27] D. Dubbeldam, A. Torres-Knoop, and K. S. Walton, "On the inner workings of Monte Carlo codes," *Mol. Simul.*, vol. 39, no. 14-15, pp. 1253–1292, 2013.

[28] Q. Yang, D. Liu, C. Zhong, and J.-R. Li, "Development of Computational Methodologies for Metal-Organic Frameworks and Their Application in Gas Separations," *Chem. Rev.*, vol. 113, no. 10, pp. 8261–8323, 2013.

[29] T. Van Heest, S. L. Teich-McGoldrick, J. A. Greathouse, M. D. Allendorf, and D. S. Sholl, "Identification of Metal-Organic Framework Materials for Adsorption Separation of Rare Gases: Applicability of Ideal Adsorbed Solution Theory (IAST) and Effects of Inaccessible Framework Regions," *J. Phys. Chem. C*, vol. 116, no. 24, pp. 13183–13195, 2012.

[30] E. Haldoupis, S. Nair, and D. S. Sholl, "Finding MOFs for Highly Selective CO2/N2 Adsorption Using Materials Screening Based on Efficient Assignment of Atomic Point Charges," *J. Am. Chem. Soc.*, vol. 134, no. 9, pp. 4313–4323, 2012.

[31] E. S. Kadantsev, P. G. Boyd, T. D. Daff, and T. K. Woo, "Fast and Accurate Electrostatics in Metal Organic Frameworks with a Robust Charge Equilibration Parameterization for High-Throughput Virtual Screening of Gas Adsorption," *J. Phys. Chem. Lett.*, vol. 4, no. 18, pp. 3056–3061, 2013.

[32] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff, "UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations," *J. Am. Chem. Soc.*, vol. 114, no. 25, pp. 10024–10035, 1992.

[33] D. E. Coupry, M. A. Addicoat, and T. Heine, "Extension of the Universal Force Field for Metal–Organic Frameworks," *J. Chem. Theory Comput.*, vol. 12, no. 10, pp. 5215–5225, 2016.

[34] C. A. Young and A. L. Goodwin, "Applications of pair distribution function methods to contemporary problems in materials chemistry," *J. Mater. Chem.*, vol. 21, pp. 6464–6476, 2011.

[35] R. Bro and A. K. Smilde, "Principal component analysis," *Anal. Methods*, vol. 6, pp. 2812–2831, 2014.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.