

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**EXPLORING GAS STORAGE PROPERTIES OF METAL
ORGANIC FRAMEWORKS BY ARTIFICIAL INTELLIGENCE
COMBINED COMPUTATIONAL METHODS**

ÖMER TAYFUROĞLU
A THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
DEPARTMENT OF CHEMISTRY

GEBZE
2023

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL
AND APPLIED SCIENCES

EXPLORING GAS STORAGE PROPERTIES
OF METAL ORGANIC FRAMEWORKS BY
ARTIFICIAL INTELLIGENCE COMBINED
COMPUTATIONAL METHODS

ÖMER TAYFUROĞLU
A THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
DEPARTMENT OF CHEMISTRY

THESIS SUPERVISOR
PROF. DR. YUNUS ZORLU

GEBZE
2023

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

METAL ORGANİK KAFES YAPILARIN
GAZ DEPOLAMA ÖZELLİKLERİNİN
YAPAY ZEKÂ İLE KOMBİNE EDİLMİŞ
HESAPLAMALI KİMYA YÖNTEMLERİ İLE
ARAŞTIRILMASI

ÖMER TAYFUROĞLU
DOKTORA TEZİ
KİMYA ANABİLİM DALI

DANIŞMANI
PROF. DR. YUNUS ZORLU

GEBZE
2023



DOKTORA JÜRİ ONAY FORMU

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 06/07/2023 tarih ve 2023/37 sayılı kararıyla oluşturulan jüri tarafından 17/07/2023 tarihinde tez savunma sınavı yapılan ÖMER TAYFUROĞLU'nun tez çalışması Kimya Anabilim Dalında DOKTORA tezi olarak kabul edilmiştir.

JÜRİ

ÜYE

(TEZ DANIŞMANI) : Prof. Dr. Yunus ZORLU

ÜYE

: Prof. Dr. Erhan DEMİRBAŞ

ÜYE

: Prof. Dr. Savaş BERBER

ÜYE

: Doç. Dr. İlknur ERUÇAR FINDIKÇI

ÜYE

: Doç. Dr. Ferhat DEMİRAY

ONAY

Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun

...../...../..... tarih ve/..... sayılı kararı.

İMZA/MÜHÜR

SUMMARY

Metal-organic frameworks (MOFs) with high surface areas, pore volumes and adjustable lattice structures are attractive for energy related gas storage, and thus have been a great deal of interest in scientific research. To date, many MOF crystal structures have been reported in literature. Experimental investigation of these materials towards their gas storage properties is cumbersome in terms of time and cost. Computational chemistry tools can be beneficial not only validating experimental results but also explaining atomistic phenomena which otherwise impossible to answer by experiments. Computational methods include but not limited to data driven methods such as text/data mining, quantum mechanical calculations, atomistic simulations and machine learning, which have significant advantages in investigating the structural and gas storage properties of thousands of MOFs with low cost and effort.

In this dissertation, we systematically investigate the surface area and pore volumes of MOFs using either text and data mining (TDM) from published manuscripts or structural property calculations. Combination of both TDM and structural calculation strategies provides a more comprehensive perspective for the investigation of hydrogen storage capacities of MOFs, which elucidates plausibility of IRMOF series as candidates for H₂ storage materials. For these MOFs, we have also developed neural network potential (NNP) at the DFT level accuracy to examine the effect of structural flexibility on gas storage properties. Our developed NNP enables revealing the unusual behaviors of MOFs such as the thermal expansion properties and the effect of mechanical strain on the adsorption of hydrogen and methane molecules for the selected MOFs.

Keywords: Metal-Organic-Framework, Artificial Intelligence, Text&Data Mining, Monte Carlo, Molecular Dynamics, Atomistic Simulation.

ÖZET

Metal Organik Kafes (MOF) yapılar sahip oldukları geniş yüzey alanları ve yoğun gözenekli yapılarının sağladığı birçok üstün özelliklerinden dolayı enerji ile ilgili gazların depolama çözümlerinde araştırmacıların ilgi odağı haline gelerek geniş çapta incelenmektedirler. Bugüne kadar, modüler yapılarının oluşturduğu avantajla çeşitli metal kümeler ve organik ligantlar kullanarak çok sayıda MOF kristal yapısı rapor edilmiştir. Fakat sayıları yüz binleri bulan bu yapıların gaz depolama özelliklerini deneysel çalışmalar ile araştırmak oldukça zahmetlidir. Zaman, maliyet ve birçok açıdan avantajlara sahip hesaplamalı kimya yöntemleri, sadece çok sayıda malzemenin deneysel özelliklerini toplu bir şekilde araştırılmasına olanak tanımakla kalmaz, aynı zamanda deneylerle incelenmesi mümkün olmayan atomik düzeydeki olayları açıklamakta da faydalı olabilmektedir. Yaygın olarak kullanılan kuantum mekanik hesaplamalar ve atomik simülasyonların yanı sıra veri madenciliği ve makine öğrenimi gibi veri odaklı hesaplamalı yöntemler, MOF'ların yapısal ve gaz depolama özelliklerini araştırmak için kullanılan önemli araçlardır.

Bu tezde, öncelikle yayınlanmış makale metinleri üzerinden metin ve veri madenciliği (TDM) tekniği ve üç boyutlu MOF kristal yapıları üzerinden de geometrik hesaplamalar ile literatürdeki MOF'ların yüzey alanlarını (SA) ve gözenek hacimlerini (PV) sistematik olarak araştırdık. Böylece, SA ve PV üzerinden MOF'ların gaz depolama kapasitelerinin geniş bir perspektifte incelenmesi ile, özellikle IRMOF serisinin H₂ depolama malzemeleri için önemli bir aday olduğunu gösterdik. Daha sonra, IRMOF serisi için, yapısal esnekliğin gaz depolama özellikleri üzerindeki etkisini atomik seviyede incelemek amacıyla DFT düzeyinde yapay sinir ağları potansiyelleri (NNP) geliştirdik. Bu şekilde, NNP'lerin atomik simülasyonlarda kullanılmasıyla, bu MOF'ların termal ve mekanik esnekliklerinin H₂ ve CH₄ adsorpsiyonu üzerindeki etkisini inceleyerek MOF'ların önemli gaz depolama davranışlarının ortaya çıkarılmasını sağladık.

Anahtar Kelimeler: Metal-Organic-Framework, Artificial Intelligence, Text&Data Mining, Monte Carlo, Molecular Dynamics, Atomistic Simulation.

ACKNOWLEDGEMENTS

Beş yılı aşkın sürede yapılan bilimsel çalışmalar ve elde edilen sonuçların büyük bir titizlikle bir araya getirilmesinden oluşan bu tezi hazırlarken, beni destekleyen birçok kişiyi hatırlamamak mümkün değildi. Onlara, sadece bilimsel konulardaki yardım ve destekleri için değil, aynı zamanda samimi dostlukları için de minnettarım.

Araştırma grubuna katılma fırsatı veren, kendi bilimsel fikir ve ilgi alanlarımın peşinden gitmem için motive eden ve tez çalışmalarım süresince her konuda desteğini esirgemeyen danışmanım Prof. Dr. Yunus ZORLU'ya teşekkür ederim. Ayrıca değerli tez izleme üyeleri sayın Prof. Dr. Erhan DEMİRBAŞ ve Prof. Dr. Savaş BERBER'e katkılarından ve sabırlarından dolayı teşekkürlerimi sunarım.

Değerli hocam Doç. Dr. Abdulkadir KOÇAK'a teşekkürlerimi bir borç bilirim. Onun olağan üstü desteği ve aydınlatıcı bilimsel tartışmalarda verdiği ilham olmasaydı, bu tezdeki birçok çalışma mümkün olmazdı.

Basel Üniversitesi'ndeki araştırma grubunda altı ay bulunma şansı tanıyan ve engin bilgi ve tecrübesi ile bana yol gösteren Prof. Dr. Stefan Goedecker'e de teşekkürler ederim.

Hesaplamalı bilimlerle tanışmama vesile olduğu için değerli dostum İsmail Erol'a teşekkür ederim. Ayrıca, Gebze Teknik Üniversitesi'nde aynı laboratuvarı paylaştığım ve bu tez üzerinde çalışırken kahve aralarında keyifli sohbetleriyle beni motive eden tüm adaşlarıma teşekkür ederim.

Basel üniversitesinde mutlu kalmamı sağlayan herkese ve özellikle değerli katkılarından dolayı Dr. Jonas Finkler ve Marco Krummenacher'e teşekkür ederim.

Maddi ve manevi desteklerini hayatım boyunca yanı başımda hissettiğim Anneme, Babama, özellikle yurtdışı çalışmalarım sırasında olağan üstü fedakarlığından dolayı sevgili eşim Merve TAYFUROĞLU'na, yeterli zaman ayıramadığım için bir özür niteliğinde kızlarım, Meryem, Sena ve Esma'ya teşekkür ederim.

Bu tez çalışmasını 2211-C Yurt İçi Öncelikli Alanlar Doktora Burs Programı ve 2214-A Yurt Dışı Doktora Sırası Araştırma Burs Programı kapsamında destekleyen Türkiye Bilimsel ve Teknolojik Araştırma Kurumuna (TÜBİTAK) teşekkür ederim. Ayrıca, Yüksek Öğretim Kurumuna (YÖK) 100/2000 Doktora Programına ile doktora eğitimimde sağlamış oldukları destek için teşekkür ederim.

TABLE OF CONTENTS

	<u>Page</u>
SUMMARY	v
ÖZET	vi
ACKNOWLEDGEMENTS	vii
TABLE OF CONTENTS	viii
LIST OF ABBREVIATIONS AND ACRONYMS	xi
LIST OF FIGURES	xiv
LIST OF TABLES	xviii
1. INTRODUCTION.....	19
2. THEORETICAL BACKGROUND	23
2.1. Quantum Mechanics.....	24
2.1.1. The Schrodinger Equation.....	24
2.1.2. The Born-Oppenheimer Approximation	25
2.1.3. Hartree-Fock Theory	26
2.1.4. Density Functional Theory	30
2.2. Atomistic Simulations	34
2.2.1. Molecular Dynamics Simulations	35
2.2.2. Monte Carlo Simulations	36
2.2.3. Ensembles	39
2.2.4. Hybrid MD/GCMC simulation	40
2.2.5. Force Fields	41
2.3. Machine Learning Potentials.....	43
2.3.1. Neural Network Potentials	46
2.4. Structural Calculations	50
2.4.1. Pore Limiting Diameter.....	51
2.4.2. Surface Area	52

2.4.3.	Pore Volume.....	53
3.	IN SILICO INVESTIGATION INTO H ₂ UPTAKE IN MOFS: COMBINED TEXT/DATA MINING AND STRUCTURAL CALCULATIONS ¹	55
3.1.	Introduction	55
3.2.	Text and Data Mining (TDM) Protocol	58
3.3.	SA, PV and H ₂ Storage of MOFs by TDM	62
3.4.	Structural Calculations Protocol	68
3.5.	SA, PV and H ₂ Storage of MOFs by Structural Calculations	70
3.6.	Merging TDM with Structural Calculations	75
3.7.	Conclusions	79
4.	A NEURAL NETWORK POTENTIAL FOR THE IRMOF SERIES AND ITS APPLICATION FOR THERMAL AND MECHANICAL BEHAVIOR ²	81
4.1.	Introduction	81
4.2.	Construction of NNP for IRMOF-n (n=1,4,6,7,10)	82
4.2.1.	Generation of Training Dataset for IRMOF-n (n=1,4,6,7,10) ...	83
4.2.2.	Validation and Test on Fragments	93
4.3.	Performance of NNP on Bulk Properties	98
4.3.1.	Lattice Constants	98
4.3.2.	Energy Profiles on the Organic Linker's Rotation in the Bulk IRMOF-1	100
4.3.3.	Phonon Modes.....	101
4.3.4.	Thermal Behavior.....	105
4.4.	Conclusions	112
5.	THEORETICAL CH ₄ AND H ₂ ADSORPTION ON IRMOF-N ACCOUNTING FLEXIBILITY	113
5.1.	Introduction	113
5.2.	Combined NNP/UFF to Account for Flexibility.....	114

5.3.	Combined MD and GCMC Simulations to Account for Flexibility	116
5.4.	Flexible Snapshot Adsorption Simulations	117
5.5.	Adsorption–Relaxation Simulations	121
5.5.1.	Training NNP for LAMMPs	123
5.5.2.	NNP/UFF based MD/GCMC Simulations for IRMOF-1 and IRMOF-10	126
5.6.	Conclusions	134
6.	CONCLUDING REMARKS AND FUTURE DIRECTIONS	135
REFERENCES.....		138
BIOGRAPHY		156
APPENDICES.....		157

LIST OF ABBREVIATIONS AND ACRONYMS

<u>Abbreviations and Acronyms</u>	<u>Explanations</u>
ACS	: American Chemical Society
ACSF	: Atom Centered Symmetry Function
ADAM	: Adaptive Moment Estimation
AED	: Atomic Environment Descriptors
ANI	: Accurate Neural Network Engine for Molecular Energies
ANN	: Artificial Neural Networks
API	: Application Programming Interface
ASA	: Accessible Surface Area
ASE	: Atomic Simulation Environment
AV	: Accessible Volume
B3LYP	: Becke, 3-Parameter, Lee–Yang–Parr
BDC	: 1,4 Benzenedicarboxylate
BET	: Brunauer–Emmett–Teller
BPDC	: 4,4'-Biphenyldicarboxylate
BPNN	: Behler And Parrinello Neural Network
CCSD	: Coupled Cluster Single-Double-Triple
CG	: Conjugate Gradient
COF	: Covalent Organic Framework
CoRE	: Computation-Ready, Experimental Metal–Organic Framework Database
CSD	: Cambridge Structural Database
DFT	: Density Functional Theory
DOE	: Department Of Energy
DOI	: Digital Object Identifier
DTNN	: Deep Tensor Neural Network
DUT	: Dresden University Of Technology
FF	: Force Field
GAP	: Gaussian Approximation Potentials
GCMC	: Grand Canonical Monte Carlo

GGA	:	Generalized Gradient Approximations
GP	:	Giga Pascal
GPU	:	Graphical Process Unit
HDNN	:	High Dimensional Neural Network
HDNNP	:	High Dimensional Neural Network Potentials
HF	:	Hartree Fock
HIP	:	Hierarchically Interacting Particle Neural Network
HOMC	:	Hybrid Osmotic Monte Carlo
IEEE	:	Institute Of Electrical And Electronics Engineers
IR	:	Infrared
IRMOF	:	Isorecticular Metal Organic Framework
LCAO	:	Linear Combinations Of Atomic Orbitals
LDA	:	Local Density Approximation
LJ	:	Lennard Jones
MAE	:	Mean Absulate Error
MC	:	Monte Carlo
MD	:	Molecular Dynamics
MIL	:	Matériaux De L'Institut Lavoisier
ML	:	Machine Learning
MLP	:	Machine Learning Potentials
MM	:	Molecular Mechanic
MOF	:	Metal Organic Frameworks
MP	:	Møller–Plesset
MPNN	:	Message-Passing Neural Networks
MTP	:	Moment Tensor Potentials
NMR	:	Nuclear Magnetic Resonance
NN	:	Neural Network
NNP	:	Neural Network Potentials
NPD	:	Neutron Powder Diffraction
NPT	:	The Constant-Pressure, Constant-Volume Ensemble
NTE	:	Negative Thermal Expansion
NU	:	Northwestern University
NVT	:	The Constant-Temperature, Constant-Volume Ensemble

PBC	:	Periodic Boundary Conditions
PBE	:	Perdew-Burke-Ernzerhof Parametrization
PC	:	Probe Center
PES	:	Potential Energy Surface
POAV	:	Probe Occupied Pore Volume
PTE	:	Positive Thermal Expansion
PV	:	Pore Volume
QM	:	Quantum Mechanics
RELU	:	Rectified Linear Unit
RMS	:	Root Mean Square
RMSD	:	Root Mean Square Deviation
RMSE	:	Root Mean Square Error
RSC	:	Royal Society of Chemistry
SA	:	Surface Area
SCF	:	Self-Consistent Field
SCXRD	:	Single Crystal X-Ray Diffraction
SNAP	:	Spectral Neighbor Analysis Potential
TDM	:	Text Data Mining
TSA	:	Temperature Swing Adsorption
TZVP	:	Triple Zeta Valence Polarization
UFF	:	Universal Force Field
UM	:	University Of Michigan
ZIF	:	Zeolitic Imidazolate Frameworks

LIST OF FIGURES

<u>Figure No:</u>	<u>Page</u>
2.1: Visualization of a fully-connected artificial neural network.	44
2.2: Two-dimensional representation of the surface area calculation used in the MC algorithm.	53
2.3: Visualization of the a) geometric and b) accessible probe centered volume.	53
3.1: TDM algorithm flowchart.	59
3.2: Flowchart for merging SA and PV values from TDM using Method 2	60
3.3: Flowchart for assigning one SA(PV) values for every single alphanumeric MOF names.	63
3.4: Comparison of TDM-extracted SA (top) and PV (bottom) values to those reported in the literature.	65
3.5: POAV vs ASA calculated by Zeo++ software. The greater the SA, the greater PV values.	68
3.6: Flowchart for structural calculations purely based on single-crystal x-ray diffraction structures.	69
3.7: Cleaning performance of the algorithm for a) SA, b) PV, c) density, and d) H ₂ uptake values.	72
3.8: SA and PV values compared by our calculations (CSD algorithm) and two other algorithms of CoRE and UM (both derived from Siegel's algorithm). The data show that the last two algorithms are more successful in cleaning.	74
3.9: Hydrogen uptakes (>1.0%) calculated by empirical formula and GCMC calculations. The data belonging to GCMC calculations (y-axis) were reprinted with permission from Copyright 2019 ACS. The diagonal dashed line is used to differentiate between the H ₂ uptakes by the two methods.	75
3.10: Comparison of experimental and calculated a) ASA, b) POAV and %H ₂ uptake values with calculated by c) Zeo++ and d) GCMC by method 1 for merge of SA and PV values.	77

3.11:	Probe-occupiable pore volumes (POAV) calculated using a H ₂ probe (radius = 1.20 Å) vs N ₂ probe (radius = 1.86 Å).	78
3.12:	Comparison of probability histograms of a) surface area (SA) and pore volumes (PV) among experimental text and data mining (TDM, solid blue), our calculations (solid red), UM calculations (dashed green), and CoRE calculation (dashed black).	79
4.1:	Systematic Fragmentation of IRMOF-n (n=1,4,6,7,10).	86
4.2:	IRMOF-n (n=1,4,6,7,10) fragments. F1-F5 are generated from IRMOF-1. F6, F7, F8, F9 are from IRMOF-4, 6,7 and 10, respectively. The atoms in the structure were represented by the following colors: white for H, brown for C, red for O, and grey for Zn.	87
4.3:	Construction of NNPs from MOF fragments	90
4.4:	Uncertainty between NNP1 and NNP2 energies according to amount of data.	92
4.5:	Training and validation loss curves	93
4.6:	Errors in energies of fragments for a) training and b) test sets. Color bar shows the number of structures with the corresponding errors. Black dashed lines mark errors of 0.005 eV/atom. DFT energies (x axis) are normalized.	95
4.7:	Errors in forces of fragments for a) training and b) test sets. Color bar shows the number of structures with the corresponding errors. Black dashed lines mark errors of 0.5 eV/Å.	97
4.8:	Potential energy surface in IRMOF-1 by scanning a): dihedral angle between carboxylate and phenylene planes from 0 to 180° and b) distance of aromatic C-H from 1.02Å to 1.16Å. Dashed lines shows experimental values as NMR for torsional barrier and NPD for equilibrium distance.	101
4.9:	Calculated vibrational spectra of IRMOF-n series by NNP. For zoomed sections refer to Figure 4.10.	104
4.10:	IRMOF-1 and IRMOF-7 vibrations zoomed in different regions.	105
4.11:	Representation of SCXRD structure of IRMOF-7 at 100 K. COO planes are orthogonal to each other and angled by 45° to the	109

naphthalene plane to optimize interaction between H atoms of two neighboring B rings; and between H atoms of ring A/B and carboxylate O atoms.

4.12:	NNP-2D scan (with 10° steps) of rotations of two facing aromatic rings in a) IRMOF-1 and b) IRMOF-7. F and Y angles correspond to the angle between carboxylate and aromatic planes on each facing rings. The images were produced the data by fitting 12th order polynomial function. Bottom graphs show 1D scan (a portion in the top images) by fixing F angles to the initial angles (F =0° for IRMOF-1 and Y=45° for IRMOF-7).	110
5.1:	Illustration of the NNP/UFF Combined Force Fields Models	116
5.2:	Changes in density and POAV fraction (top) upon compression along [001] direction for the fragments produced by MD simulations using NNP for a) IRMOF-1 and b) IRMOF-7. MOF structures are shown in cartoon (bottom) before (0%) and after (35%) strain applied.	119
5.3:	Gravimetric (black) and volumetric (blue) gas uptakes upon compression along [001] direction at each fragment obtained by MD simulations using NNP for CH4 a), H2 b)on IRMOF-1 and CH4 c), H2 d) on IRMOF-7.	120
5.4:	Shown is change in unit cell volume, number of molecules throughout the combined MC/MD hybrid simulation. Red arrows indicate properties that change, and blue arrows indicate properties that remain constant.	122
5.5:	Errors in energies of fragments for test sets. Color bar shows the number of structures with the corresponding errors. Black dashed lines mark errors of 0.005 eV/atom. DFT energies (x axis) are normalized for a clear representation.	126
5.6:	Errors in forces of fragments for test sets. Color bar shows the number of structures with the corresponding errors. Black dashed lines mark errors of 0.5 eV/ Å.	126
5.7:	Variations in the types of organic linkers incorporated into IRMOF-1 and IRMOF-10 structures.	127

5.8:	The number of adsorbed H ₂ molecules versus time for the simulations of adsorption of RMOF-1 for hydrogen gas at 100 bar and 77K using UFF4MOF force fields in MD/GCMC hybrid simulations	129
5.9:	The number of adsorbed H ₂ molecules versus time for the simulations of adsorption of RMOF-1 for hydrogen gas at 100 bar and 77K using combined HDNNP/ UFF4MOF force fields in GCMC/MD hybrid simulations	130
5.10:	Adsorption profiles showing the uptake of a) H ₂ and b) on IRMOF-1 using combined HDNNP/UFF4MOF potentials, pure UFF4MOF simulations, and experimental data.	131
5.11:	The ratio of predicted/experimental loadings at different pressures. Predicted values are of pure UFF and HDNNP. HDNNP follows a steady constant line over the pressures while UFF is totally divergent.	132
5.12:	Adsorption profiles showing the uptake of a) H ₂ and b) CH ₄ on IRMOF-10 using combined HDNNP/UFF4MOF potentials and pure UFF4MOF simulations.	133

LIST OF TABLES

<u>Table No:</u>	<u>Page</u>
3.1: Performance tests of the TDM algorithm at various steps	67
3.2: MOF Databases Comparison. UM shows the number of structures with surface area, pore volume, and gravimetric hydrogen uptake percentage greater than 0 or 1. Note that all databases in the table are derived from CSD.	70
4.1: Equilibrium lattice constants and bulk modulus of IRMOF series calculated using NNP and compared with experimental SCXRD and other calculations in literature.	99
4.2: Thermal expansion coefficient values of IRMOF-n series calculated at different temperatures. For the calculation at different temperatures, refer to (Table C.1, Table C.2, Figure C.8 and Figure C.9)	110
4.3: Thermal pressure coefficients calculated from MD simulations using NNP.	112
5.1: Total H ₂ and CH ₄ uptake on IRMOF-1 and IRMOF-10 (wt %) at 100 bar; 77K for H ₂ and 300K for CH ₄	134

1. INTRODUCTION

Over the next generations, due to rising standards of living and growing population, energy consumption is expected to increase enormously [1, 2]. Because of carbon-based fuels' limited supply and environmental issues, increasing demand for clean energy has directed researchers to explore for alternative and renewable energy sources. Among these energy sources, hydrogen (H_2) and methane (CH_4) are considered to be a better alternative due to their high energy density and abundant availability in nature. H_2 is also a very clean energy source in terms of the environment since only water is formed from its oxidation [3]. However, the storage and transportation problems of H_2 and CH_4 constitute an obstacle for the utilization of such sources in particular on-board usages as in the case of light vehicles [4, 5]. Storing these gases under ambient conditions is not suitable, requiring more costly systems such as very low temperatures and/or high pressures [6]. The most promising solution strategies to overcome these obstacles are chemical storage, on-board production, or storage by adsorption on porous materials. Storage of these gases by adsorption on porous materials has been the subject of numerous scientific studies [5]. Metal-Organic Framework (MOFs) have become the center of attention of researchers due to their large surface areas and porous structures and have been extensively studied [7–12].

MOFs are crystalline materials composed of a central group containing a metal atom and organic linkers extending them in one-dimensional (1D), 2D, and 3D periodic lattice structures [13–15]. Owing to their large surface areas, sizable pore volumes and the tremendous structural properties they gain with their modular structure, they are being studied not only for gas storage [16], but also in many fields ranging from drug delivery systems to catalyst applications [17, 18]. The ability of combining different metal sites and organic linkers allows easy synthesis of a variety of MOFs. In addition, alterable metal sites and organic linkers also gain tunability of pore volumes and surface areas in MOFs toward the use as hydrogen and methane storage materials. [19, 20].

Due to the large number of MOF crystal structures reported in the literature, the experimental investigation of these materials and their gas storage properties requires considerable time and cost. In addition, in order to obtain materials with target properties, pioneering studies are needed quickly and less costly. Computational

chemistry tools, which have significant advantages in solving these problems, are generally used either to explain experimental results [21, 22] or to lead experiments [22].

The field of MOFs has witnessed an explosion of experimental results documented in the vast literature. To navigate through the gillions of available data, text-data mining (TDM) strategies prove invaluable in extracting key information. By employing advanced techniques such as natural language processing (NLP), researchers can efficiently uncover trends, correlations, and insights hidden within the extensive MOF literature, accelerating the discovery, and understanding of these remarkable materials. Without proper data extraction, the wealth of knowledge present in these manuscripts remains untapped, rendering it akin to mere garbage. Therefore, employing efficient TDM strategies becomes imperative in order to unlock the treasure trove of insights hidden within the MOF literature and harness the full potential of this valuable scientific resource [23].

Apart from extracting data from published manuscripts for the experimental properties, it is also possible to computationally model MOFs from their single-crystal x-ray diffraction structures for the prediction of these physical properties. There are various computational methods used to investigate the gas storage properties of MOFs. These methods can be basically classified into costly yet the most accurate ab-initio quantum chemical approaches [22] and less costly classical molecular mechanics [24]. These two general approaches have their advantages and disadvantages. The first technique, which uses ab initio techniques and is based on quantum theory, is very precise and produces uncontroversial results in most cases. Unfortunately, the computational cost is very high and can only be applied in very limited systems. On the other hand, classical approaches with a lower level of accuracy allow us to screen thousands of structures. This can sometimes lead to unrealistic results or fail to miss important features at the atomic level [25]. Furthermore, scientists have combined these methods to exploit the advantages of both approaches. In this method, known as "multiscale" approaches, the system can be described either at the quantum mechanical level or classically, depending on the desired properties. [22].

Artificial intelligence (AI) tools have come to play a key role in scientific calculations. Although applications of AI such as machine learning (ML) and neural network (NN) models/algorithms have been known since the 1950s, the techniques have developed very rapidly, especially in the last decade with the exponential increase

in data collection and computational capacity. In particular, its applications in biology [26], physics [27, 28] and chemistry [29] have revolutionized computational science, bringing many important innovations.

Computational studies with thousands of different MOFs obtained by modifying modular building blocks have yielded enormous amounts of diverse data. However, the detailed study of these structures using the most accurate computational techniques is quite expensive in terms of time and computational cost [30, 31]. Considering all these, machine learning techniques can overcome these problems in MOFs and offer the accuracy of ab-initio methods and the speed of classical approaches [31]. Highly accurate and reactive ML potentials can be built to describe MOF systems with both organic and inorganic subunits for simulations. Instead of building electronic structure from the first principles at every time step, their great capabilities for learning of multidimensional potential energy surfaces (PES) allows us to carry out much faster simulations.

By utilizing computational methods, this dissertation methodically explores the physisorption of two significant energy-related gases on extensively porous metal-organic frameworks. It is organized in the following chapters: First, chapter 2 gives a brief description of theoretical methods used.

Next, in chapter 3, we investigate the SA and PV properties of MOFs, which are known to have a great impact on gas storage. Using text-and-data mining (TDM), all reported MOF publications (as many as 60,000) in the literature (until 2020) were examined according to their experimental surface area (SA) and pore volumes (PV). Experimental H₂ uptakes were predicted based on the empirical formula between H₂ uptake and SA/PV values. In this chapter, we also performed structural calculations such as SA, PV and theoretical H₂ uptakes on MOFs available in Cambridge Structural Database (CSD) with x-ray single crystallographic data (as many as 90,000 structures). We finalized this chapter by reporting a list of MOFs with the greatest H₂ storage capacities by combining TDM and structural calculations. Isorecticular metal–organic framework (IRMOF) series in this final list inspired us to further our studies.

Next, in chapter 4, we describe the construction and validation of a generic neural network potential (NNP) for IRMOF-n series (n=1,4,6,7,10) based on DFT reference data and predict bulk properties such as equilibrium lattice constants, phonon density of states and linker orientation. Our developed NNP enables us to reveal

unusual behaviors of MOFs such as thermal expansion properties on the adsorption of hydrogen and methane molecules for the selected MOFs.

Then, in chapter 5, we retrain/transfer our previously constructed NNP for IRMOF-n series so as to use in MD simulations in LAMMPS and investigate the gas adsorption profiles of the MOFs using hybrid NNP-classical potential accounting for extrinsic and intrinsic flexibility with MD/GCMC separate/combined approaches.

Finally, in chapter 6, we discuss the broader impacts of our studies and the plausible future directions/extensions.

2. THEORETICAL BACKGROUND

The adsorption of gas molecules (guest) on MOFs (host) is a dynamic process and requires atomistic simulations, in which both the motions of guest and host molecules need to be described. Atomistic simulations can be classified based on the underlying computational methods used as i) electronic structure based models (also called Quantum Mechanical, QM), which include electronic degrees of freedom and mostly governed by ab-initio molecular dynamics (AIMD), ii) force-field based models (also called Molecular Mechanics, MM), which samples the ensemble of the system of atoms mostly governed by classical Monte Carlo (MC) simulations, or by classical Newtonian mechanics mostly governed by classical molecular dynamics (MD) simulations, in which the physical and thermodynamic properties of the system is retrieved by the ensemble or time averages. In force-field based models, the atoms are treated as point-like objects connected through chemical bonds and affected by associated forces. The neural network potential (NNP) aims to precisely reproduce these forces that are the function of atomic coordinates. The concept of chemical bonding has been established in chemistry for a considerable period, but recent advancements in methodology and computing hardware have enabled a quantitative representation of pairwise atomistic interactions.

For the adsorption of gas molecules into host systems, there have been other types of calculation methods that do not require a simulation of the system but instead predict geometrical (structural) parameters such as surface area (SA), pore volume (PV), void fraction, the density of the MOFs, which are ultimately responsible for the gas adsorption. Thus, there have been empirical coefficients that connect these parameters to gas adsorption properties. We will refer to these types of computations as “structural calculations” or “geometrical calculations” throughout the dissertation.

This chapter provides a concise overview of electronic structure methods, which enable the computation of potential energy, forces, and various other properties of chemical systems and materials. Subsequently, a selection of fundamental simulation techniques utilizing the energies and forces derived from electronic structure methods is also introduced. Finally, important geometrical parameters and their calculations are briefly discussed.

2.1. Quantum Mechanics

2.1.1. The Schrodinger Equation

Quantum mechanics, developed approximately a century ago, provides a description of the interaction between electrons and nuclei, which constitute the fundamental building blocks of the solid matter that surrounds us. A many-body wave function $\Psi(R_I, r_i)$ is used to define the quantum mechanical systems which contains electrons and the nuclei. The variables in the wave function correspond to the positions of the nuclei and electrons, respectively. The solution of the time-independent Schrödinger equation is used to calculate the energy of the ground state of chemical systems.

$$\hat{H}\Psi(R_I, r_i) = E\Psi(R_I, r_i) \quad (2.1)$$

where $\hat{H}$ is the quantum mechanical Hamiltonian operator corresponding the total energy, E , of the system. Equation (2.2) shows the components of the Hamiltonian which includes the following terms: kinetic energies of the nuclei (T_N) and electrons (T_e), as well as the potential energies for the electron-electron interactions (V_{ee}), and nuclei-nuclei interactions (V_{NN}), and electron-nuclei interactions (V_{Ne}).

$$H = - \underbrace{\sum_I \frac{1}{2m_I} \nabla_{R_I}^2}_{T_N} - \underbrace{\sum_i \frac{1}{2} \nabla_{r_i}^2}_{T_e} + \underbrace{\sum_{I < J} \frac{Z_I Z_J}{|R_I - R_J|}}_{V_{NN}} + \underbrace{\sum_{i < j} \frac{1}{|r_i - r_j|}}_{V_{ee}} - \underbrace{\sum_{I,i} \frac{Z_I}{|R_I - r_i|}}_{V_{Ne}} \quad (2.2)$$

Here, Z_I represents to the nuclear charges, m_I donates the masses of the nuclei, and ∇_x denotes the nabla operator pertaining to the three-dimensional coordinate x as follows.

$$\nabla_x = \left(\frac{\partial}{\partial x_x}, \frac{\partial}{\partial x_y}, \frac{\partial}{\partial x_z} \right) \quad (2.3)$$

The Equation (2.2) does not include the electronic masses and charges since it is expressed in atomic units. Atomic units are established in such a way that fundamental quantities such as the Planck constant (reduced) $\hbar = \frac{h}{2\pi}$, elementary charge e , electron mass m_e . Bohr radius $a_0 = \frac{4\pi\epsilon_0\hbar^2}{m_e e^2}$ are all set to value of 1. As such, distances are measured in Bohr (1 Bohr = 0.5292 Å) and energies in Hartree (1 Hartree = 27.2114 eV) atomic units. These units prove highly advantageous as they streamline numerous equations necessary for our calculations. Moreover, employing atomic units often results in computed values, such as energies or forces, being close to unity in magnitude. This does not only enhance readability but also circumvents computational challenges stemming from underflow and overflow problems that arise when dealing with exceedingly small or large numbers. Hence, unless specified else, entirely equations in this dissertation will be presented in atomic units.

2.1.2. The Born-Oppenheimer Approximation

A comparison of the mass of electrons with that of hydrogen, the lightest of all atoms, results in a mass difference of nearly two thousand times heavier than that of electrons. This substantial difference in mass leads to a significant distinction in the timescales of motion between the nuclei and the electrons. As a result, it is often a valid approximation to treat the heavy atomic nuclei as essentially immobile while the light electrons instantaneously adapt to any changes in the nuclear positions. This approximation, known as the Born-Oppenheimer approximation, is commonly employed in quantum mechanics [32]

By decoupling the motion of electrons from that of the nuclei, the Born-Oppenheimer approximation enables us to express the overall wave function of the system as a product of the electronic wave function and the nuclei wave function.

$$\Psi(R_I, r_i) = \psi(r_i)\phi(R_I) \quad (2.4)$$

The electronic wave function is influenced indirectly by the positions of the nuclei through the electrostatic potential they generate. We make an assumption that the electrons rapidly and adiabatically adjust to changes in the nuclear positions, consistently attaining their ground state configuration at each moment in time.

Consequently, the electronic Hamiltonian employed in the Born-Oppenheimer approximation is defined as follows:

$$H = T_e + V_{ee} + V_{Ne} \quad (2.5)$$

As we assume that the electrons instantaneously adjust to the positions of the nuclei, the total potential energy of the system solely depends on the nuclear positions.

$$E = E(R_I) \quad (2.6)$$

The function $E(R)$, referred to as the potential energy surface (PES), constitutes a crucial point of investigation within this thesis. It is of particular interest as various properties, including the geometric ground state, reaction pathways, and thermodynamic quantities, can be derived from it.

2.1.3. Hartree-Fock Theory

The mathematical representation of the electronic wave function relies on the positions of all N electrons present in the system. However, when considering computational approaches to handle such a function, particularly through expansion in a basis set, it becomes evident that the computation space requirements would exponentially increase with the number of electrons. Consequently, attempting to store and operate such wave functions becomes infeasible, even with modern computing hardware, for systems with more than a few electrons. Therefore, the electronic wave function must be simplified. The most straightforward approach is to express the wave function as a product of N individual electron wave functions. This simplification is commonly referred to as the Hartree product.

$$\Phi(r_i) = \phi_1(r_1)\phi_2(r_2)\phi_3(r_3) \dots \phi_N(r_N) \quad (2.7)$$

This approach fails to account for two crucial aspects of interacting electrons: correlation and exchange interactions. Correlation interaction arises from the fact that electrons respond to the instantaneous positions of one another. As a consequence of their electrostatic repulsion, it is highly improbable to find two electrons occupying

the same position or being very close to each other. However, the wave function described above does not reflect this behavior, as the positions of the electrons are entirely independent of one another. Exchange interaction arises due to the Pauli exclusion principle, which dictates that the wave function of electrons, being fermions, need to change sign when the coordinates of two electrons are exchanged.

$$\phi_1(r_1) \dots \phi_i(r_i) \dots \phi_j(r_j) \dots \phi_N(r_N) = -\phi_1(r_1) \dots \phi_i(r_j) \dots \phi_j(r_i) \dots \phi_N(r_N) \quad (2.8)$$

The wave function provided in Equation (2.6) does not inherently satisfy this requirement. To ensure the necessary anti-symmetry of the wave function, it can be expressed by a Slater-determinant of single electron wave functions.

The computation of the total energy for such a wave function is then possible by using the following expression:

$$E_{HF} = \sum \int \phi_i^*(r_i) h_i \phi_i(r_i) d_r + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N ([ij|ij] - [ij|ji]) \quad (2.9)$$

$$h_i(r_i) = \frac{-\nabla_{r_i}^2}{2} - \sum_I \frac{Z_I}{|R_I - r_i|} \quad (2.10)$$

$$[ij|ij] = \int \phi_i^*(r_1) \phi_j^*(r_2) \frac{1}{|r_1 - r_2|} \phi_k(r_1) \phi_l(r_2) d_{r_1} d_{r_2} \quad (2.11)$$

The term $[ij|ji]$ arises due to the implementation of the Slater determinant ansatz, which ensures the required antisymmetry of the wave function. It is commonly referred to as the exchange term. As observed in Equation (2.11), $[ij|ij]$ represents the electrostatic energy associated with the interaction between the electrons.

Minimizing this energy under the constraint that the orbitals ϕ_i are orthogonal reveals that the solution involves selecting the N lowest eigenfunctions of the Fock operator, denoted as $\hat{F}$. The restricted Fock operator ($\hat{F}[\{\phi_i\}](\phi_i)$) for a closed-shell and non-spin-polarized system is provided below. Notably, in this context, only the $N/2$ lowest orbitals are utilized since each orbital is doubly occupied, with one electron having spin-up and the other having spin-down.

$$\hat{F}[\{\phi_i\}]\phi_i(r) = h_i\phi_i(r) + \sum_{j=1}^{N/2} \left(2\hat{J}_j\phi_i(r_1) - \hat{K}_j\phi_i(r_1) \right) \quad (2.12)$$

The Coulomb operator $\hat{J}$ and the exchange operator $\hat{K}$ are expressed as follows:

$$\hat{J}_j\phi_i(r) = \int \frac{\phi_i^*(r')\phi_i(r')}{|r-r'|} d r' \phi_i(r) \quad (2.13)$$

$$\hat{K}_j\phi_i(r) = \int \frac{\phi_j^*(r')\phi_i(r')}{|r-r'|} d r' \phi_j(r) \quad (2.14)$$

The solution to the Hartree-Fock equations is not limited to spin-restricted systems and can also be achieved for non-spin-restricted systems, although it involves a slightly more intricate process.

To numerically solve these equations, an approach is employed that includes expanding the wave functions $\phi_i(r)$ using a basis set $\chi_\mu(r)$.

$$\phi_i(r) = \sum C_{\mu i} \chi_\mu(r) \quad (2.15)$$

This basis set allows for a representation of the wave functions in terms of a set of orthogonal functions χ_μ , which are often chosen to be convenient for the particular system under consideration. By expressing the wave functions in this basis set, the Hartree-Fock equations can be solved iteratively to obtain the desired numerical solution.

By performing appropriate transformations, we arrive at the Roothaan-Hall equation, which can be expressed as a generalized eigenvalue problem involving the coefficient matrix C .

$$FC = SC_\epsilon \quad (2.16)$$

In the given equation, the Fock operator F is expanded in the basis set $\chi_\mu(r)$, while S represents the overlap matrix ($S_{ij} = \int \chi_i(r)\chi_j(r)dr$) and ϵ denotes the orbital

energies. Due to the dependence of the Fock operator F on the molecular orbital coefficients C , the Roothaan-Hall equation requires iterative solving until a self-consistent solution is reached. To initiate the iterative process, an initial guess for the coefficient matrix C is made, which allows for the construction of F . The updated orbital coefficients are then obtained by solving Equation (2.16). This iterative process is repeated until a self-consistent solution for C is achieved.

This solution not only enables us to obtain information about the potential energy surface (PES) in terms of energy and forces but also facilitates the computation of other observables, such as the electron density $\rho(r)$. Additionally, as we will see later, it is important to note that ρ plays a central role in density functional theory (DFT).

$$\rho(r) = \sum_i^{N/2} |\phi_i(r)|^2 \quad (2.17)$$

Various types of basis sets are available, each with its own set of advantages and disadvantages. For instance, plane wave basis sets [33] are particularly suitable for periodic systems. One of their key advantages is that they constitute an orthonormal basis that can be systematically enhanced.

Linear combinations of atomic orbitals (LCAO) have traditionally been among the most widely used basis sets. In this approach, a set of basis functions is placed on each atom to construct the basis set. Each atomic basis function typically comprises a radial part, often represented as a linear combination of Gaussian functions [34], although alternative choices such as numerically defined functions [35] also exist. The angular part of the basis functions is described by spherical harmonics. One advantage of LCAO basis sets is that they generally require a smaller number of basis functions compared to other methods. When Gaussian functions are employed in LCAO basis sets, certain integrals, such as overlap or Coulomb integrals, can be analytically computed. However, achieving systematic convergence in terms of the basis set becomes challenging, making it difficult to accurately estimate the basis-set error in computations utilizing LCAOs.

The Hartree-Fock method is mainly constrained by its approximation of the wave function with a single Slater determinant. Even with a complete basis set, this approximation is insufficient to fully describe the actual wave function due to the

absence of electronic correlation. Consequently, the method cannot adequately account for dispersion interactions, often predicts excessively short bonds, and overestimates bond-breaking energies.

The Hartree-Fock method, while having certain limitations, can still provide reasonable energy differences in many cases. Additionally, it possesses important properties such as size-consistency, where the energy of a system composed of two infinitely separated subsystems is equal to the sum of the energies of the individual systems. The method is also variational, ensuring that the calculated energy serves as an upper bound for the energy of the exact wave function solution.

In terms of computational complexity, a naive implementation of the Hartree-Fock method scales as $\mathcal{O}(n^4)$, where n represents the number of basis functions used, typically proportional to the number of electrons in the system. However, more sophisticated implementations have been developed that significantly reduce the computational complexity to around $\mathcal{O}(n^2)$ [36].

Many methods have been developed to overcome the limitations of the Hartree-Fock method. These methods also rely on the wave function but use correction to effectively calculate electron interaction terms in the Hamiltonian. The correction can be made by applying perturbation theory such as Møller–Plesset (MP2, MP4, etc.) [37], by mixing of different electronic configurations (configuration interaction, CI) [38], or by using the exponential cluster operator (coupled cluster, CC) to account for electron correlation [39, 40]. These methods also have a level of correction, but the calculations become prohibitively expensive even for a few atom systems when the level of the correction is increased. Although these methods are considered as the gold standard for approaching the true solution to the Schrodinger equation, they will not be covered in this dissertation due to their computational cost.

2.1.4. Density Functional Theory

Density functional theory (DFT) provides an alternative approach to solving Schrödinger's equation. Its foundation can be traced back to the influential work of Hohenberg and Kohn [41], who demonstrated that the electronic ground state density $\rho(r)$ uniquely determines the external potential $v(r)$ of the nuclei, up to an additive constant. This means that no two potentials can yield the same electronic density if they differ by more than an additive constant. Furthermore, since the total energy E is

a functional of the electronic wave function, which is uniquely determined by the external potential, it follows that $\rho(r)$ alone is sufficient to uniquely determine E . In other words, the electronic density $\rho(r)$ contains all the necessary information to determine the total energy of the system.

Although the energy of a quantum state is typically a functional of its 3N-dimensional wave function, the aforementioned result reveals that the ground state energy, which often holds the greatest interest, can be expressed solely as a functional of the three-dimensional density $\rho(r)$. In other words, the ground state energy is determined by the spatial distribution of the electrons captured in the density rather than the full complexity of the wave function.

For this reason, there must exist a universal functional $\mathcal{F}[\rho(r)]$ of the electron density $\rho(r)$, which, when minimized, yields the correct ground state density that would be obtained by solving the Schrödinger equation for the external potential $v(r)$. The functional $\mathcal{F}$ captures the essential information about the electronic interactions and allows us to determine the ground state properties solely from the information of the electron density.

$$E = \int v(r)\rho(r)dr + \mathcal{F}[\rho(r)] \quad (2.18)$$

This is a remarkable discovery because it suggests that determining the ground state energy of a system can be achieved by focusing only on the electron density $\rho(r)$. This approach reduces the challenges related with handling the highly complex wave function, which has a large number of dimensions. However, it is important to note that while the theorem confirms the existence of a functional, but it does not tell us anything about how such a functional would look like. Nonetheless, scientists have proposed numerous functionals that have demonstrated great success in predicting various properties of many-body quantum-mechanical systems. As a result, density functional theory has emerged as one of the most extensively utilized tools in the fields of material sciences and chemistry. This powerful approach saves significant computational effort and enables researchers to obtain valuable insights into the behavior and properties of materials [42]

Through a comparison between Equation (2.18) and the Hamiltonian presented in Equation (2.5), a revised representation of the functional $\mathcal{F}[\rho(r)]$ can be obtained.

$$E = \int v(r)\rho(r)dr + \langle \Psi[\rho(r)] | T_e + V_{ee} | \Psi[\rho(r)] \rangle \quad (2.19)$$

In this case, the electron-nuclei interaction V_{Ne} is solely determined by the electron density and is thus incorporated into $v(r)$. However, the kinetic and Coulomb terms of the electrons indirectly rely on the density $\rho(r)$. As a result, Kohn and Sham suggested expanding the electron density using independent electrons ϕ_i .

$$\rho(r) = \sum_i^N |\phi_i(r)|^2 \quad (2.20)$$

Similar to HF theory, the Kohn-Sham wave function is represented by a single Slater determinant. The total energy can be described as the sum of the kinetic and Coulomb energies of the electrons, along with an additional component known as the exchange-correlation energy, $E_{xc}[\rho]$. This energy accounts for the disparities between the Kohn-Sham approach and the exact solution.

$$E = \int v(r)\rho(r)dr + \sum_i^N \int \phi_i^*(r) \frac{\nabla^2}{2} \phi_i(r)dr + \frac{1}{2} \iint \frac{\rho(r)\rho(r')}{|r-r'|} drdr' + E_{xc}[\rho] \quad (2.21)$$

To construct a Hamiltonian, one can define an exchange-correlation potential $V_{xc}[\rho](r)$ as the functional derivative of $E_{xc}[\rho]$. With respect to the density ρ . The electronic orbitals are subsequently determined as the solution that minimizes the energy in the given expression, while satisfying an additional constraint of orthonormality.

$$\left(\frac{\nabla^2}{2} + v(r) + \int \frac{\rho(r')}{|r-r'|} dr' + V_{xc}[\rho](r) \right) \phi_i(r) = \epsilon_i \phi_i(r) \quad (2.22)$$

Since the left side of the equation relies on the charge density ρ , it is necessary to employ a self-consistent approach, similar to the HF method, to ensure consistency and convergence.

The accuracy of DFT is significantly influenced by the quality of the exchange-correlation functional, which encompasses various many-body interactions, including exchange, correlation, self-interaction, and kinetic energy correction. Discovering an appropriate functional is a challenging endeavor, and numerous forms have been proposed. Ongoing research is dedicated to refining and advancing these functionals.

The local density approximation (LDA) is among the most basic forms of the exchange-correlation functional. As suggested by its name, it solely relies on the local electronic density.

$$E_{XC}^{LDA} = \int \varepsilon_{XC}^{LDA}(\rho(r))\rho(r)dr \quad (2.23)$$

Although LDA functionals can yield reasonably accurate geometries, they have a tendency to overestimate binding energies. However, it is important to note that LDA functionals accurately reproduce exact results for homogeneous densities. Therefore, they serve as a foundation for the development of many other functionals.

To obtain a more accurate approximation of the exchange-correlation energy in systems with rapidly varying densities, such as atoms and molecules, it is beneficial to include a dependence on the derivative of the density, denoted as $\nabla\rho(r)$.

$$E_{XC}^{LDA} = \int \varepsilon_{XC}^{GGA}(\rho(r), \nabla\rho(r))\rho(r)dr \quad (2.24)$$

The functionals that incorporate a dependence on the derivative of the density $\nabla\rho(r)$ are referred to as generalized gradient approximations (GGAs). Among the widely used GGAs is the functional developed by Perdew, Burke, and Ernzerhof (PBE) [43]. Unlike some GGAs that are parameterized using experimental data, PBE is derived without the need for empirical information, making it applicable to a broad range of systems.

The inclusion of the single-particle kinetic energy density ($\tau^s(r)$) in addition to the electron density $\rho(r)$ characterizes functionals known as meta-generalized gradient approximations (meta-GGAs).

(2.25)

$$E_{XC}^{LDA} = \int \varepsilon_{XC}^{GGA}(\rho(r), \nabla\rho(r), (\tau^s(r)))\rho(r)dr$$

Hybrid functionals, which integrate a portion of exact exchange from HF theory, are employed in various computational approaches. One example is the PBE0 functional [44], where one quarter of the PBE exchange is substituted with exact exchange. Another widely utilized hybrid functional is B3LYP [45], an early meta-GGA that continues to be extensively employed in the field of chemistry.

The computational complexity of DFT calculations typically scales with $\mathcal{O}(n^3)$ in relation to the number of electrons, primarily due to the matrix diagonalization process involved. Therefore, since the metal organic frameworks studied in this dissertation contain numerous atoms in primitive unit cell, they are divided into fragments and const reduced. As we will see in the next chapters, it has become possible to perform fast calculations with ML potentials, which are computational complexity liner scales.

2.2. Atomistic Simulations

In the section 2.1, we explored various methods for obtaining approximate solutions to the electronic ground states of the Schrödinger equation. These solutions yield valuable information such as x-ray diffraction patterns, ionization energies, band gap energies, and more. However, the primary focus of this dissertation is centered around the energy and forces applied on the nuclei, which are obtained through electronic structure calculations.

As discussed earlier, in the Born-Oppenheimer approximation, the nuclei are treated as point-like particles. In this context, the potential energy surface (PES) represents the energy function, $E = E(R_I)$, for a specific configuration of atomic positions R_I . The forces acting on each atom are obtained by taking the negative derivatives of the potential energy function with respect to the atomic coordinates.

$$F_{I\alpha} = -\frac{dE(R_I)}{dR_{I\alpha}}, \quad \alpha = \{x, y, z\} \quad (2.26)$$

The majority of electronic structure codes possess the capability to compute these derivatives analytically and offer them as output. Alongside the forces, the

potential energy surface (PES) offers a multitude of valuable information, including ground state geometries, reaction dynamics, thermodynamic properties, formation energies, and more. However, the PES of a system with N atoms, which is a function of $3N$ coordinates, becomes highly complex when N reaches the hundreds or even thousands. Consequently, exploring PES becomes exceedingly challenging due to the curse of dimensionality. To address these difficulties, numerous simulation methods and algorithms have been developed. In the upcoming sections, we will provide an overview of the most relevant methods.

2.2.1. Molecular Dynamics Simulations

With the knowledge of the forces exerted on each atom and their corresponding masses, we can employ Newton's equation of motion to compute the temporal progression of the system. The trajectories obtained from this process, known as molecular dynamics (MD), can reveal properties that are not easily observable in experiments due to limitations in time and length scales. For instance, MD simulations can provide insights into various properties such as reaction pathways, thermodynamic ensemble averages, and more.

By integrating Newton's equation of motion, we can derive the time evolution of the atomic positions.

$$F_I = m_I \frac{d^2 R_I}{dt^2} \quad (2.27)$$

Where m_I are the atomic masses and in order to perform numerical integration of Newton's equation, we discretize it by employing a finite time step Δt . The Euler method is the simplest approach to propagate the atomic positions, starting from the initial conditions $R(t = 0)$ and velocity $V(t = 0)$. It employs a straightforward finite difference scheme for this purpose.

$$R_I(t + \Delta t) = R_I(t) + V_I(t)\Delta t \quad (2.28)$$

$$V_I(t + \Delta t) = V_I(t) + F_I(R_I(t)) \frac{\Delta t}{m_I} \quad (2.29)$$

However, it is observed that this scheme leads to a drift in the total energy, which ideally should be conserved during a molecular dynamics (MD) simulation.

An alternative algorithm that offers greater stability is the Verlet algorithm [46]. Its enhanced long-term stability arises from its time reversibility property, meaning that applying the algorithm with negative time steps can exactly retrace the forward trajectory taken with positive time steps. Consequently, a time-reversible algorithm avoids introducing systematic energy drifts. While short-term fluctuations may still be observed, they generally oscillate around a mean value in the long term. The velocity Verlet algorithm is employed to propagate the positions according to the following procedure:

$$R_I(t + \Delta t) = R_I(t) + V_I(t)\Delta t + F_I(R_I(t))\frac{\Delta t^2}{2m_I} \quad (2.30)$$

$$V_I(t + \Delta t) = V_I(t) + \left(F_I(R_I(t + \Delta t)) + F_I(R_I(t)) \right) \frac{\Delta t}{2m_I} \quad (2.31)$$

The conservation of total energy in molecular dynamics (MD) simulations allows for sampling the microcanonical ensemble of a system. However, when targeting the canonical ensemble, where atomic configurations are visited according to their Boltzmann probability at a specific temperature, additional adjustments to the MD algorithm are necessary. These modifications, known as thermostats, are employed to control the system's temperature. Commonly used thermostats include the Nosé-Hoover [47, 48], Langevin [49], and Andersen [50] methods.

2.2.2. Monte Carlo Simulations

In the last section, we discussed how MD simulations can be utilized to sample the canonical ensemble of a system. By using this approach, we can evaluate thermodynamic integrals, such as the calculation of expectation values of a quantity X over the Boltzmann distribution. This can be expressed as follows:

$$\langle X \rangle_T = \frac{1}{Z(T)} \int \exp\left(\frac{-E(R)}{k_B T}\right) X(R) d(R) \quad (2.32)$$

Where k_B is the Boltzmann constant, T represents the temperature, and $Z(T)$ denotes the partition function that normalizes the Boltzmann distribution.

$$Z(T) = \int \exp\left(\frac{-E(R)}{k_B T}\right) d(R) \quad (2.33)$$

Instead of employing MD simulations with a thermostat, an alternative approach is to utilize Monte Carlo (MC) simulations. These computational methods, named after the renowned Monte Carlo casino in Monaco, rely on random numbers. Because of their intrinsic randomness, these methods are non-deterministic and provide approximate results with an error that is influenced by the number of iterations. In the area of atomistic simulations, the term "Monte Carlo" typically relates to methods based on the Metropolis-Hastings algorithm [51].

The Monte Carlo method, including the Metropolis-Hastings algorithm, has a wide range of applications beyond the Boltzmann distribution. It serves as a general method for numerically solving high-dimensional integrals. The conventional approach to numerical integration involves subdividing the function into numerous small segments and approximating the area under the curve as the sum of the areas of small rectangles that fit beneath it.

$$\int_a^b f(x) dx \approx \sum_{i=0}^N f\left(\frac{a + i(b-a)}{N}\right) \left(\frac{b-a}{N}\right) \quad (2.34)$$

When a function $f(x)$ depends on multiple variables, it is necessary to employ a nested summation, which involves evaluating the function on a grid that spans all the variables. However, this approach becomes impractical for functions with a large number of variables, such as the potential energy surface (PES), due to the exponential increase in the number of required grid points. Moreover, many of these grid points would likely yield physically unrealistic structures where atoms are excessively close together, leading to high energies and, consequently, low Boltzmann probabilities. As

a result, these configurations would make negligible contributions to the integral being calculated.

By employing importance sampling, it is possible to circumvent this issue. By generating configurations R_i that are already distributed according to the Boltzmann distribution, we can approximate the integral in Equation (2.32) in the following manner.

$$\langle X \rangle_T = \frac{1}{N} \sum_i^N X(R_i) \quad (2.35)$$

The generation of such configurations can be achieved using the Metropolis–Hasting’s algorithm [51], which employs a Markov chain for performing a random walk-in configuration space. To sample a distribution $P(R)$, the algorithm is initialized with an initial configuration R_0 . Subsequently, a new configuration is proposed based on a proposal distribution $g(R_i|R')$ which represents the probability of proposing the new configuration R' given the current configuration R_i . The acceptance of the new configuration is determined by the following probability.

$$\alpha = \min(1, \frac{P(R')g(R_i|R')}{P(R_i)g(R'|R_i)}) \quad (2.36)$$

To determine whether the proposed new configuration is accepted or rejected, a random number u is generated from a uniform distribution between zero and one using a random number generator. This value is then compared to α . If u is less than α , indicating that the new configuration is favorable, it is accepted ($R' = R_{i+1} = R$), added to the list of configurations, and the process continues with the next iteration. On the other hand, if u is greater than or equal to α , the previous configuration is retained ($R_i = R_{i+1}$), and the algorithm proceeds accordingly.

The choice of the magnitude of random displacements in the Metropolis–Hastings algorithm is crucial. If the displacements are too large, the proposed configurations will have high energy, leading to a low acceptance rate. Conversely, if the displacements are too small, the acceptance rate will be high, but there will be a stronger correlation between subsequent samples, requiring more iterations for the

simulation to converge. It is important to select an appropriate displacement magnitude that yields an acceptance rate of approximately 23% [52]. Additionally, it should be noted that since only the fraction of the probabilities $P(R_i)$ and $P(R')$ get involved in the acceptance rate, the normalization of the probability distribution P is not necessary for calculating the acceptance rate. This property is particularly advantageous when sampling the Boltzmann distribution since the partition function is typically unknown.

2.2.3. Ensembles

In the context of a system consisting of N atoms, a total of $6N$ variables ($3N$ positions and $3N$ velocities) are required to fully describe its microscopic state. This $6N$ -dimensional space is referred to as the phase space, and its evolution follows the laws of mechanics. When considering macroscopic systems, variables such as temperature T , volume V , and pressure p are derived as time averages of the microscopic phase space. However, Gibbs introduced the concept of ensembles, where the averaging is performed over a group of microstates that share the same macroscopic properties. This replacement of time averaging with ensemble averaging is based on the ergodic principle, which states that the two types of averaging are equivalent in infinitely long simulations.

By default, MD and MC simulations typically sample the microcanonical ensemble, where the total energy E , number of particles N , and volume V are kept constant. However, in experimental settings, it is more common to maintain constant temperature T , number of particles N , and volume V (canonical ensemble or NVT ensemble). Therefore, it is crucial to ensure that the simulations are designed to closely resemble experimental conditions, accepting meaningful comparisons. Several ensembles are available for both MD and MC simulations such as NVE , NVT and μVT . The NVE (microcanonical) ensemble is employed for isolated systems, where equilibrium can be achieved by adjusting temperature and pressure. In the NVT ensemble, the number of particles, volume, and temperature are maintained, allowing the systems to reach equilibrium by modifying the pressure and total energy. The NpT (Isothermal-Isobaric) ensemble shares similarities with the NVT ensemble, as the pressure is maintained while allowing alterations to the total energy. Both ensembles keep the pressure constant by placing the system in a heat bath, and this is achieved through the thermostat in simulations, where the pressure can be controlled using a

barostat. In the μVT (grand canonical) ensemble, while chemical potential, temperature and volume remain constant, the number of particles changes until the system reaches equilibrium. This ensemble is especially suitable for examining adsorption processes.

2.2.4. Hybrid MD/GCMC simulation

GCMC simulation is commonly used to predict adsorption isotherms by allowing fluctuations in the number of adsorbates, achieving equilibrium under defined chemical potential (μ) and temperature (T) conditions. However, this technique alone cannot directly predict the adsorption isotherm of flexible porous materials. Reparameterization is required, which not only introduces inefficiency due to force field adjustments for each flexible system but also neglects the effects of adsorbent flexibility on adsorption properties. Simulations offer an efficient tool for exploring the microscopic world, and it would be a missed opportunity not to investigate the dynamic interaction between hosts and guests. While MD simulations typically account for host molecule vibrations, the number of guest molecules remains constant throughout the computation.

One of the methods that combines both MD and GCMC algorithms is the hybrid osmotic Monte Carlo (HOMC) simulation [53]. It utilizes the osmotic thermodynamic ensemble, which is suitable for describing the flexibility of an adsorption system ($N_{host}\mu_{ads}\sigma T$) [54]. In this ensemble, the host atom number (N_{host}) and adsorbate chemical potential (μ_{ads}) are kept consistent, while the stress (σ , equivalent to pressure p in an isotropic system) and temperature (T) are maintained at constant values. The simulation volume and number of molecules are adjusted to achieve equilibrium. This approach combines elements from both the NpT and μVT ensembles, resulting in a hybrid statistical ensemble. The probability in this ensemble can be expressed as follows:

$$P(s, V, N_{ads}) \propto \frac{V^N \exp(\beta \mu_{ads} N_{ads})}{\Lambda_{host}^{3N_{host}} N_{host}! \Lambda_{ads}^{3N_{ads}} N_{ads}!} \exp(-\beta(U(s; h) + pV) \quad (2.37)$$

$$\text{Here, } \Lambda_{host} = \frac{h}{\sqrt{2\pi m_{host} k_B T}} \text{ and } \Lambda_{ads} = \frac{h}{\sqrt{2\pi m_{ads} k_B T}}.$$

In this dissertation, we employed the hybrid MD/GCMC simulation approach to investigate the effect of flexibility on adsorbents such as MOF. The MD simulation part was conducted in the isothermal-isobaric (NpT) ensemble. During this simulation, the number of adsorbates (N), pressure (p), and temperature (T) were maintained at constant values, while the volume and instantaneous temperature were allowed to vary, leading to isotropic changes in cell size. In this ensemble, the host has the freedom to vibrate and modify its structure in order to accommodate the adsorption of guest molecules. The resulting configurations were then accepted or rejected based on the Metropolis acceptance criterion.

2.2.5. Force Fields

As mentioned earlier, numerous ab-initio methods are available to explore the PES by solving an approximate model of the Schrödinger equation. However, these methods demand significant computational resources, especially for larger systems, making them suitable only for moderately sized systems and limited calculations. Nevertheless, many applications, including MD and MC simulations, necessitate a vast amount of energy and force evaluations to achieve accurate and dependable results.

In order to facilitate long-time simulations involving many atoms, numerous alternative approaches have been introduced to explore the potential energy surface (PES). These methods primarily fall into the category of classical force fields, where a mathematical functional form is selected based on physical principles and subsequently parameterized using experimental data or ab-initio calculations. Metropolis et al. [55] proposed the utilization of a potential energy function based on pairwise interactions, enabling the calculation of the total energy by considering all pairwise interactions within the system.

$$E = \frac{1}{2} \sum_{i \neq j}^N V(r_{ij}), r_{ij} = |r_i - r_j| \quad (2.38)$$

One of the extensively used pairwise potentials is commonly known as the Lennard-Jones (LJ) potential [56, 57]. This potential incorporates an attractive term a/r^6 to account for Van der Waals interactions, as well as a repulsive term b/r^m . The exponent m of the repulsive term is often chosen as 12, which allows for efficient

computations by simply squaring the attractive term. This choice of LJ potential makes it particularly suitable for modeling interactions among noble gas atoms. Therefore, the 12-6 potential form is widely utilized as the most common representation of the LJ potential.

$$V_{LJ} = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right] \quad (2.39)$$

In pairwise potentials like the LJ potential, the energy contribution from the interaction between every pair of atoms is taken into account. A straightforward approach that considers all pairs of atoms in the system would lead to a computational cost that increases quadratically with the number of atoms. Fortunately, pairwise potentials such as the LJ potential exhibit rapid decay with increasing distance, allowing us to safely disregard the contributions from atom pairs beyond a predefined cutoff radius. Only the interactions between local neighbors need to be considered. To implement this efficiently, lists of neighboring atoms, known as Verlet lists, are necessary. Constructing these Verlet lists requires careful consideration, as comparing all atom pairs would once again result in a quadratic scaling. Instead, methods based on tree-like data structures such as k-D trees or octrees, or methods that bin atoms into a three-dimensional grid, can be used to achieve a scaling of $\mathcal{O}(N_{at} \log(N_{at}))$ or $\mathcal{O}(N_{at})$ respectively.

Pairwise potentials alone often prove insufficient in describing the intricacies of the potential energy surface (PES). Therefore, additional terms, such as angular three-body terms and four-body terms involving dihedral angles, must be incorporated. A wide range of force field implementations [58–63] are now accessible, offering a variety of functional forms for users to select from and enabling a more comprehensive representation of the PES.

Thanks to their simple functional form, force fields can be evaluated with remarkable efficiency. Moreover, the small number of free parameters enables the utilization of only a few high-quality ab-initio reference data points or even experimental data for fitting purposes.

Force fields offer efficient evaluations due to their simple functional form, and their small number of free parameters allows for fitting with a limited set of high-quality ab-initio reference data points or experimental data. A smart choice of

functions used in force fields ensures reasonable behavior in extreme cases not covered by the reference data, such as high repulsive forces at very short bond lengths. However, unchanging functional form and limited number of parameters impose limits on accuracy. To overcome this limitation, machine learned potentials (MLPs) have gained popularity. MLPs utilize machine learning algorithms to represent a larger function space with numerous free parameters. By fitting MLPs to a large amount of reference data, high accuracy can be achieved, often surpassing the accuracy of the methods used to generate the training data.

2.3. Machine Learning Potentials

As discussed in the previous section, force fields provide an efficient approximation of the PES, allowing for significantly faster evaluations compared to ab-initio methods. This enables long-time simulations which involve many atoms. Even with notable achievements, classical force fields have inherent limitations in terms of accuracy since their predefined functional forms for interaction potentials necessarily overlook some of the complexities inherent in real PESs. With the recent advancements in computational resources, it has become feasible to generate large-scale ab-initio reference data sets. This, coupled with algorithmic enhancements, has led to the emergence of machine learned potentials (MLPs). Unlike manually selected functional forms with a limited number of parameters, MLPs leverage machine learning techniques, such as artificial neural networks (ANNs), to represent atomic interactions. These machine learning methods offer highly flexible functional forms with numerous parameters that can be fitted to a vast array of reference data points.

The majority of contemporary MLPs are composed of two key elements. Initially, atomic environments are transformed into encoded representations using atomic environment descriptors (AEDs), which ensure the preservation of crucial symmetries. Subsequently, a machine learning engine is employed to acquire the functional correlation between the environment descriptors and the system's energy. These machine learning engines encompass highly parameterized general functions, with their parameters being determined through the training process. Numerous types of MLPs have been proposed, including neural network potentials [64–67], Gaussian approximation potentials (GAPs) [68], moment tensor potentials (MTPs) [69], spectral neighbor analysis potentials (SNAPs), [70] and various others [71, 72]. Despite the

differences in machine learning algorithms and functional forms among these methods, the fundamental components and challenges, such as the description of atomic environments and long-range interactions, remain consistent. Due to the interatomic potentials constructed in this dissertation relying on ANNs (Artificial Neural Networks), the primary focus of this thesis will be on MLPs based on ANNs.

ANNs, alternatively referred to as multi-layer perceptron in the machine learning community, draw loose inspiration from the neural network structure found in the human brain, composed of interconnected neurons. They are composed of an input layer, multiple hidden layers, and an output layer. Each layer contains a specific number of interconnected nodes, which are linked to both the preceding and succeeding layers.

At its most basic form, the essential components of ANNs are dense (or "fully connected") layers, which act as linear transformations from input vectors x to output vectors y , represented by the equation:

$$y = Wx + b \quad (2.40)$$

Here, the parameters consist of both weights W and biases b .

Figure 2.1 provides an illustration of an ANN, showcasing three input nodes, two fully connected hidden layers with number of nodes each, and an output layer with a single node. The connections between the nodes are depicted as black lines, with each connection having a weight parameter and each node possessing a bias.

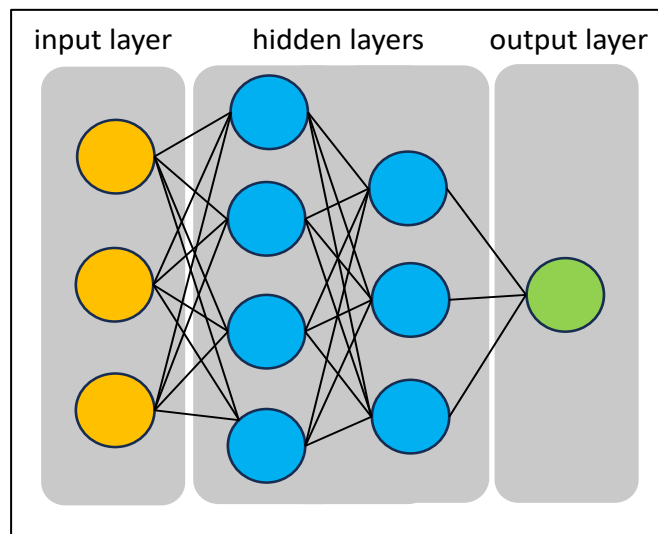


Figure 2.1: Visualization of a fully-connected artificial neural network.

To evaluate the ANN, input values are fed into the input nodes. The values are subsequently propagated through the connections to each node in the initial hidden layer. During this propagation, the values are multiplied by the corresponding connection weight and then aggregated within the node, where the bias is also included. Before passing the values to the next layer, an activation function is applied. This function plays a crucial role in introducing non-linearity to the ANN. In this dissertation, we utilized the hyperbolic tangent function, although other functions like the rectified linear unit (ReLU), sigmoid, or Gaussian function are commonly employed. In the output layer, a linear function is often utilized to avoid constraining the output within a finite range, which can occur with the hyperbolic tangent function. The value of each node can be calculated as follows:

$$v_i^n = f \left(\sum_j w_{ij}^n v_j^{n-1} + b_i^n \right) \quad (2.41)$$

Here, v_i^n represents the i th node, w_{ij}^n donates the weights, b_i^n indicates bias and f represents activation function. This computation can be conveniently written as a matrix-vector multiplication, allowing for a highly efficient evaluation of the ANN on graphics processing units (GPUs). In recent times, various other types of ANN layers, such as convolutional layers [73], pooling layers, or attention-based mechanisms [74], have become predominant.

In training an ANN, the weights and biases are adapted to minimize the error between the represented function $f_{NN}(x)$ by the ANN and the desired target function $y(x)$. This error is measured using a loss function $\mathcal{L}$. The overall loss $\mathcal{L}$ is computed by integrating over all possible inputs x , capturing the extent of the error that needs to be minimized and since this integral cannot be evaluated analytically, we must approximate it using a set of discrete points x_i , the so-called training dataset.

$$\mathcal{L} = \int L(f_{NN}(x) - y(x)) dx \quad (2.42)$$

And now, utilizing any local optimization algorithm, we can minimize the training loss. To achieve this, we require the partial derivatives of the loss concerning

the weights and biases of the ANN. The backpropagation algorithm efficiently computes these derivatives by leveraging the chain rule. During the forward pass (prediction) through the ANN, the values of each node are stored. Subsequently, to obtain the gradients, a backward pass is performed through the ANN, applying the chain rule to derive gradients with respect to the inputs, weights, and biases. It should be noted that in the case of MLPs, the derivatives (forces) might be incorporated into the loss function, slightly complicating the calculation of derivatives with respect to the weights and biases. Fortunately, modern machine learning libraries such as TensorFlow [75] and PyTorch [76] automatically track the calculations performed during ANN prediction, facilitating automatic differentiation without the need for manual implementation.

2.3.1. Neural Network Potentials

In recent years, there has been a significant focus on Neural Network Potentials (NNPs), which are machine learning potentials developed using Artificial Neural Networks. The early implementations of NNPs employed atomic coordinates, such as distances and angles, to represent the structure and model the PES. However, these internal coordinates, despite being roto-translationally invariant, introduced an arbitrary ordering of atoms, which failed to capture the equivalence of permuted inputs [77–79]. Consequently, due to the use of atomic coordinates directly, NNPs could potentially assign distinct energies to structures that are symmetrically equivalent or vice versa. Moreover, the dimensionality of the input, denoted as x , is influenced by the number of atoms, thereby limiting the applicability of the PES to chemical systems of the same size. To overcome these challenges, a possible approach is to decompose the energy prediction using a many-body expansion, which addresses these issues [80, 81]. However, this method scales unfavorably with the size of the system and the number of chemical species, as each term in the many-body expansion necessitates a separate Neural Network model.

Behler and Parrinello [67] introduced the concept of high-dimensional neural network potentials (HDNNPs), which express the total energy of a chemical system as the sum of atomic contributions, $E = \sum E_i$, predicted by a single neural network (or one for each element). This approach is based on the assumption that the energetic contribution, E_i , of each atom primarily depends on its local chemical environment.

By treating all atoms of the same type equally and utilizing the commutative property of summation, the output remains unchanged when the input is permuted. This decomposition into atomic contributions enables the use of the same NNP to predict systems with varying numbers of atoms. In theory, this framework allows for transferability between system sizes, allowing a model trained on small systems to be applied to predict energies and forces for larger systems. However, achieving this requires adequate sampling of local environments to eliminate spurious correlations resulting from the training data distribution, as well as the incorporation of corrections for long-range effects.

The advent of HDNNPs has inspired the development of various neural network architectures, which can be broadly categorized into two types. Descriptor-based NNPs [82–84] rely on predefined rules to encode the atomic environment into a vector x , which is then utilized as input for a conventional feed-forward neural network (as depicted in Equation (2.40)). These architectures include several variants of the original Behler-Parrinello network, such as ANI [65] and TensorMol [85]. On the other hand, end-to-end NNPs [86–88] take nuclear charges and Cartesian coordinates as input and learn a suitable representation directly from the provided data.

2.3.1.1. Descriptor Based NNPs

Behler and Parrinello [67] were the pioneers in introducing the initial descriptor-based NNP, which employed atom-centered symmetry functions (ACSFs) [82]. These ACSFs encompassed two-body terms G_i^2 as well as three-body terms G_i^3 to encode information about the chemical environment of each atom i .

$$G_i^2 = \sum_{j \neq i} \exp(-\eta(r_{ij} - r_s)) f_{cut}(r_{ij}) \quad (2.43)$$

$$G_i^3 = 2^{1-\zeta} \sum_{j \neq i} (1 + \lambda \cos \theta_{ijk})^\zeta \exp(-\eta(r_{ij}^2 + r_{ik}^2 + r_{jk}^2)) f_{cut}(r_{ij}) f_{cut}(r_{ik}) f_{cut}(r_{jk}) \quad (2.44)$$

In this context, r_{ij} represents the distance between atoms i and j , θ_{ijk} denotes the angle formed by atoms i , j and k with atom i at the center, and the summations encompass

all atoms within a specified cutoff distance r_{cut} . Since the values of G_i^2 and G_i^3 are independent of the atom order and are solely derived from internal coordinates, all physical invariants are maintained. By incorporating the cutoff function r_{cut} , the smooth variation of G_i^2 and G_i^3 is guaranteed as atoms enter or exit the cutoff sphere. The parameters η , r_s , ζ , and λ (where λ is equal to ± 1) determine the specific distances or combinations of angles and distances to which the ACSFs are most sensitive.

$$f_{cut}(r) = \begin{cases} \frac{1}{2} \left[\cos\left(\frac{\pi r}{r_{cut}}\right), 1 \right], & r \leq r_{cut} \\ 0, & r > r_{cut} \end{cases} \quad (2.45)$$

By combining a sufficient number of G_i^2 and G_i^3 with diverse parameters and storing them in a vector x_i , a unique "fingerprint" of the local environment surrounding atom i is created. This environment descriptor serves as the input for a neural network, which predicts the energy contributions E_i of the individual atoms. The total energy $E_i = \sum_i E_i$ is then computed by summing up the contributions from all atoms.

Given that the ACSFs rely solely on geometric information, their optimal performance is observed in systems composed only of atoms from a single element, such as silicon atomic system [67]. When consider for a multicomponent system, a common approach involves duplicating the symmetry functions for each combination of elements. Then, separate neural networks are employed to predict the energy contributions for atoms belonging to the same element type [89].

2.3.1.2. End-to-End NNPs

One potential limitation of the ACSFs discussed in the last section is that they require expert selection prior to training the neural network. If the choice of symmetry functions is inadequate, such as when the resulting descriptor is nearly identical for two significantly different structures, it can inherently restrict the expressive power of the neural network and the achievable accuracy. Moreover, as the number of input dimensions grows, computational costs can escalate rapidly, both in terms of calculating the descriptors and evaluating the neural network. This becomes particularly pronounced when modeling multicomponent systems, where assumptions of orthogonality between different elements are commonly made (increasing the input

dimensionality), or when descriptors are weighted by element-dependent factors (which can potentially limit the structural resolution of the descriptor).

End-to-end NNPs directly utilize the atomic types and positions as inputs to obtain suitable representations from the reference data. In a similar manner to descriptor based NNPs, many end-to-end NNPs compute the total energy E by summing the atomic contributions E_i . However, these contributions are estimated based on learned features x_i that encode information regarding the local chemical environment of each atom i . This approach enables the adaptation of features according to the size and distribution of the training set, as well as the specific chemical property of interest, throughout the training process. The idea behind this approach is to acquire knowledge of mapping atomic environments to a high-dimensional feature space, where structurally (and energetically) similar configurations are clustered closely while dissimilar ones are situated farther apart.

Within the deep tensor neural network framework [86], which serves as a prominent example of an end-to-end NNP approach, the refinement of atomic features x_i based on neighboring atoms allows for the mapping to the high-dimensional features through an iterative process. The features are initially initialized as $x_i^{t+1} = e_{z_i}$, where e_z represents learnable element-dependent representations that are subsequently updated for T steps. The approach, which involves iterative refinement of atomic features based on neighboring atoms, has been widely adopted in various end-to-end networks. This general approach can be expressed as follows:

$$x_i^{t+1} = \mathcal{G}^t \left(x_i^t + \sum_{j \neq i} \mathcal{F}^t(x_i^t, x_j^t, g(r_{ij})) f_{cut}(r_{ij}) \right) \quad (2.46)$$

Here, the summation encompasses all atoms located within a distance of r_{cut} , and a cutoff function f_{cut} is employed to ensure the smooth behavior of the system when atoms transition across the cutoff distance. And, to refine the atomic features after incorporating information from neighboring atoms through the interaction function $\mathcal{F}^t$, the "atom-wise" function $\mathcal{G}^t$ is applied. In most cases, the interatomic distance r_{ij} is not directly utilized as input to this function. Instead, it is expanded into a vectorial input $g(r_{ij})$ through a series of uniformly spaced radial basis functions [66, 90] Both the "atom-wise" and interaction functions are implemented as neural networks, although their specific implementations can vary across different end-to-end NNP

architectures. Since only pairwise distances are considered and the order of atoms does not affect the summation's outcome due to its commutative property, the resulting features x_i obtained from Equation (2.46) possess roto-translational and permutational invariance, thereby extending this property to the energy predictions as well.

In the literature, there are numerous end-to-end neural network models which have developed inspiration from Scarselli et al.'s graph neural network [91] and subsequently unified as message-passing neural networks (MPNNs) [88]. Inherently, these models consider molecules as undirected graphs, with atoms represented as nodes and interactions between them as edges. Through message-passing, where information is exchanged between nodes along edges, complex interatomic interactions can be modeled. The Deep Tensor Neural Network (DTNN) stands out as a prominent example [86], and since its introduction, this approach has been subject to continuous refinement, giving rise to new architectures such as SchNet [64, 90], HIP-NN [92] and PhysNet. [66].

In the present dissertation, SchNet was selected as the preferred methodology based on its exceptional track record in accurately representing and modeling interatomic interactions. This choice was made after carefully evaluating various alternatives and considering the extensive success and applicability of SchNet in similar research domains. By leveraging SchNet's capabilities, this study aims to delve deeper into understanding and analyzing interatomic interactions in a more comprehensive and insightful manner.

2.4. Structural Calculations

In order to comprehend the porous structure of MOFs at the mesoscale, conducting a geometrical characterization becomes imperative. Several properties serve as geometric descriptors for MOFs, including density, void fraction, pore limiting diameter (PLD), pore volume (PV) and surface area (SA) since the voids of MOFs are ultimately responsible for allowing guest molecules to adsorb in the host materials. In particular, surface area and pore volumes are of interest in our dissertation. SA and PV can experimentally be determined from the low temperature nitrogen (N_2) gas adsorption. SA and PV can also be calculated from the single-crystal x-ray diffraction structure of MOFs. However, both surface area and pore volumes have different variations that can provide distinct insights into the structure of the

voids. In that sense, MOF structure may have an area/volume that is too small for a guest molecule to occupy, which is called non-accessible area/volume. Alternatively, the surface/volume can be accessible for the guest molecules, called accessible surface area (ASA) or accessible volume. Here, guest molecules are represented as the “*probe*” and for different gas adsorption properties, different probes with different radii can be used in the calculations. On the other hand, the host material may have internal voids that are large enough for guest molecules to trap, but the guest molecules (or the probe) may not find a path to reach that large void due to the small diameter of path (too narrow), so called pore limiting diameter (PLD). Further details of these terms are discussed in the following subsections.

There are several open-source programs available for calculating these properties. In this study, Zeo++ [93, 94] was employed for the geometrical characterization of the MOF. The calculation of structural parameters such as surface area, pore volume, and pore limiting diameter, etc., all start with a probe molecule to visit empty spaces and find the atoms’ arrangements in a periodic domain and then fill the graph representation of the void space with techniques such as Voronoi decomposition. The computation of the Voronoi network (also called tessellation) of a three-dimensional structure serves as the foundation of Zeo++. In this context, the Voronoi network involves partitioning space into regions known as Voronoi cells. These cells possess the property that any points within a given cell are closer to a particular atom than to any other atom. Consequently, the Voronoi network represents the empty space or voids within the material.

2.4.1. Pore Limiting Diameter

By utilizing Voronoi network, it becomes feasible to determine the largest sphere that can fit inside the voids (referred to as the largest cavity diameter or LCD) as well as the largest sphere that remains free within the voids (known as the pore limiting diameter, PLD). In this context, the largest included sphere refers to the maximum distance observed between a node in the network and an adjacent atom. On the other hand, the largest free sphere represents the largest included sphere that can freely move within the empty space or voids. To obtain the largest included sphere, an iterative process is employed, involving all Voronoi nodes within a periodic unit cell to locate the node with the greatest distance to a neighboring atom. Subsequently, the Dijkstra's

optimization algorithm is utilized to determine the travel path with the widest opening, thereby calculating the largest free spheres. By retaining only, the largest among these largest free spheres, the final PLD is derived.

2.4.2. Surface Area

Zeo++ utilizes Monte Carlo (MC) integration methods for the calculation of the accessible surface area (ASA). The algorithm begins by performing a Voronoi tessellation, which divides the space into Voronoi cells around each atom, representing the void space and then checks the intersections between the generated surface points and the atoms in the cell. If a surface point intersects with an atom, it is discarded. The remaining non-intersecting surface points are used to estimate the accessible surface area by summing up the areas of the Voronoi faces exposed to the void space. MC integration is applied by sampling a large number of surface points and calculating the ratio of non-intersecting points to the total generated. This ratio, multiplied by the total surface area of the Voronoi cells, provides an approximation of the accessible surface area. Through this method, Zeo++ accurately estimates the surface area by considering the void space and avoiding intersections with atoms.

Figure 2.2 illustrates the implementation of the Monte Carlo (MC) algorithm in Zeo++. The blue atoms represent the probe, with the green circle representing the accessible surface area. It is defined as a concentric circle with the given atom, having a radius of $r_i + r_p$, where r denotes the radius of atom r_i , and r_p represents the radius of the probe. The MC algorithm calculates the ratio of sampled points f that land on a specific surface area for each atom, ensuring they do not overlap with the von der Waals surface of atoms (grey). The accessible surface area for the designated probe on atom i can be determined as $a = f(4\pi r^2)$. Any sampled points that crash with atoms are represented by the red dashed line. To identify the accessible surface area, the Zeo++ algorithm not only examines the labels of the cubelets and but also performs an additional step that involves assessing the viability of a sampled point based on the convexity of the Voronoi cells [93].

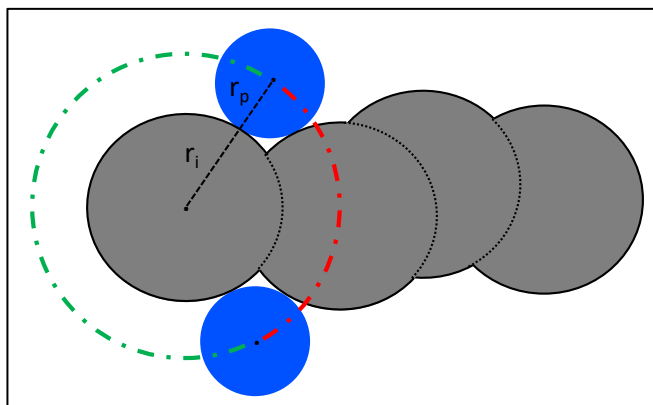


Figure 2.2: Two-dimensional representation of the surface area calculation used in the MC algorithm.

2.4.3. Pore Volume

The method used to calculate the accessible pore volume (probe centered, PC) in Zeo++ is similar to that used for determining the accessible surface area. This involves expanding the sampling area to cover the entire unit cell. It is important to ensure that the distance between each sampled point and any atomic surface is greater than the radius of the probe.

Geometric volume encompasses all the volume within the unit cell that does not overlap with the crystal's atoms whereas probe centered represents the volume where the center of a spherical probe can be located. In Figure 2.3-a, the geometric volume is represented by the light green area while, in Figure 2.3-b, the probe centered volume is depicted as the dark green are.

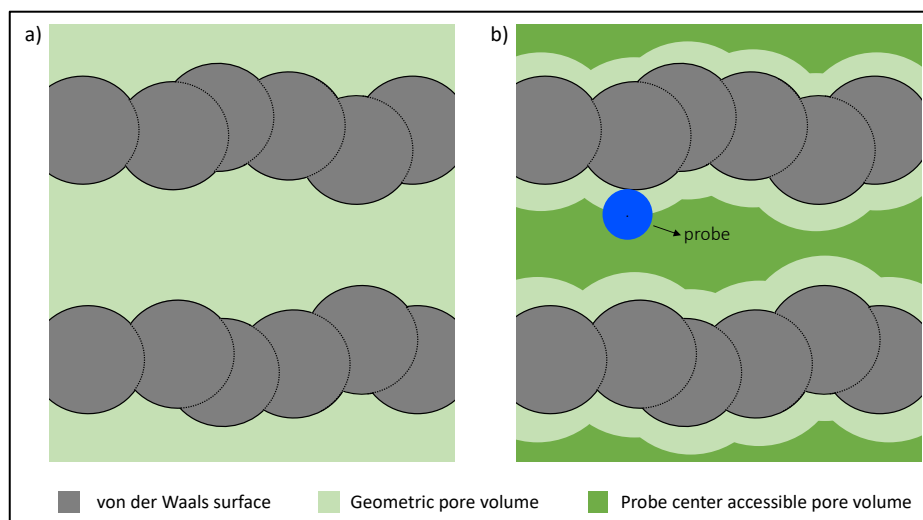


Figure 2.3: Visualization of the a) geometric and b) accessible probe centered volume.

Smit and coworkers have introduced an accurate definition of pore volume known as probe-occupiable pore volume (POAV) [16]. This definition, which is also applied in Zeo++, significantly improves the agreement with experimental results. Briefly, this method aims to ensure that the theoretical pore volume matches the pore volume obtained experimentally from the nitrogen isotherms. The experimental definition assumes that one can take the bulk density of the gas and calculate the volume from the number of adsorbed gas molecules per unit volume. This volume, however, has no notion of atoms and should be defined as the entire volume enclosing all the adsorbed gas atoms. Therefore, in Figure 2.3, this volume has to include the light green (for accessible pores) areas in addition to the dark green. If we have a system with large pores, the difference between the geometric and probe-occupiable volumes is small, but for micropores, however, this difference can be significant. In this dissertation, we employed POAV method to accurately predict the pore volume and successfully compare it with the experimental.

3. IN SILICO INVESTIGATION INTO H₂ UPTAKE IN MOFS: COMBINED TEXT/DATA MINING AND STRUCTURAL CALCULATIONS¹

3.1. Introduction

As mentioned in chapter 1, carbon-based fuels are not sustainable and environment friendly. Therefore, other alternative strategies, hydrogen gas is a good candidate for the long-term solution as energy storage materials [95]. However, hydrogen is a permanent gas and cannot be liquefied at ambient conditions, making it difficult for storage and transportation [96]. Hence, prior to use as fuel, it requires further processes [97]. Some of the solutions include but are not limited to chemical storage, on-board production from metal hydrides, or adsorption on porous materials. Adsorption of hydrogen on porous materials has one of the most attractive strategies and thus has been a topic for a great deal of scientific studies [5]. In particular, metal-organic frameworks (MOFs) have been widely studied due to their greater surface areas and porosities [7, 98–100].

MOFs are crystalline materials composed of a central group containing a metal atom and organic linkers extending them in one-dimensional (1D), 2D, and 3D periodic lattice structures [13, 101]. Owing to their tremendous structural properties such as large surface areas, large pore volumes, high flexibilities, and capabilities of reversible processes, they have been studied not only for hydrogen storage [94] but also for other gas adsorptions in industrial applications [102], drug deliveries [18], separation [101, 103, 104], and catalysis [17]. The ability of combining different metal sites and organic linkers allows easy synthesis of a variety of MOFs. In addition, alterable metal sites and organic linkers also gain tunability of pore volumes [19] and surface areas in MOFs toward the use as hydrogen storage materials [20, 105].

On the other hand, a quite large number of possible MOFs require too much experimental effort and cost for the synthesis and assessment as H₂ storage materials.

¹ This chapter were adapted from O. Tayfuroglu, A. Kocak, and Y. Zorlu, "In Silico Investigation into H₂ Uptake in MOFs: Combined Text/Data Mining and Structural Calculations," *Langmuir*, vol. 36, no. 1, pp. 119–129, 2019 (Copyright 2020 American Chemical Society

In this regard, this problem can be sorted out by the help of computational studies. Grand canonical Monte Carlo (GCMC) simulations [106, 107] have been commonly used in predicting the experimental gas adsorption profiles of MOFs. Apart from GCMC simulations, surface area (SA) and pore volume (PV) values of MOFs can be calculated, which is an indirect indication for gas adsorption profiles. SA and PV values can be calculated from the geometrical voids of known atomic coordinates. The calculations require good initial geometries such as single-crystal x-ray diffraction structures. To date, several databases have been published, including experimental and hypothetical MOF structures. One of the most popular databases is created by Moghadam et al. [108] via applying seven criteria for MOF description and extracting over 90 000 1D, 2D, and 3D MOFs from nearly 1 000 000 available crystal structures (as of January 2019) in the Cambridge Structural Database (CSD). However, it should be noted that most of these structures also include disorders, guest solvent molecules, completing ions, etc., which complicates the prediction of H₂ adsorption capacities via theoretical SA and PV values or the GCMC simulations. The number of non-disordered MOFs in the CSD database is nearly 72 000. Snurr and co-workers reported another database of 5109 non-disordered clean 3D MOFs, so called as “Computation-Ready Experimental MOF Database” (CoRE), by defining bonds between metals and O, N, B, P, S, and C atoms. [109] Both authors also introduced two different algorithms for the removal of guest molecules. Keskin and co-workers [110] compared the two databases, reported 3490 common MOFs, and found that the solvent removal by Moghadam’s to be more successful.

As of 2019, Snurr and co-workers released a newer version of the CoRE database with ~12 000 MOFs [111]. In the most recent study, all available crystallographic and hypothetical MOFs (~500 000) have been screened by Siegel and co-workers for H₂ uptake properties using GCMC simulations. [112]

Apart from theoretical calculations for H₂ uptake using crystallographic coordinates of MOFs, it is also possible to search experimental gas storage capacities using mining strategies in published text data. Most of the experimental data are reported as texts in publications, rather than structural databases [113]. Thus, mining data from text should provide exceptional information on H₂ storage materials. By performing a text and data mining (TDM), it is possible to retrieve experimental SA and PV values of MOFs that can be related to their possible H₂ storage capacities. By TDM, it is possible to search for MOFs that are capable of H₂ storage reported not only

for this purpose but also for other purposes such as catalysis and drug-delivery applications. Recently, Kim and co-workers have introduced a TDM algorithm to extract SA and PV in MOF-related publications [15]. Testing their algorithm on CoRE database, they reported 999 SA and 583 PV values. They kept the data limited to the CoRE database and did not convert these values into hydrogen storage capacities due to the difficulty of merging SA and PV data.

Previous studies on microporous carbon materials showed that excess H_2 adsorption is proportional to surface area ($\bar{n}_{excess} = C \cdot \bar{SA}$). This phenomenon, so called as the “Chahine rule” [114], states that each 1 wt % H_2 uptake requires a 500 m^2 ($C = 0.021 \text{ mg } H_2/m^2$) surface area on porous material at 77 K and 35 bar. As cited by Siegel [109], further studies showed that this relationship also holds for MOFs [115]. Thus, the total H_2 uptake of MOFs can be predicted by adding the amount of gas phase hydrogen ($\bar{n}_{gas}$) present in the MOF’s pore volume. Hence, the total uptake can be described by equation below:

$$H_2 \text{ wt \%} = \bar{n}_{tot} = \bar{n}_{excess} + \bar{n}_{gas} = C \cdot \bar{SA} + \rho_{H_2} \bar{V}_{pore} \quad (3.1)$$

where C is the proportionality constant (0.021 $\text{mg } H_2/m^2$) that connects accessible surface area ($\bar{SA}$) with excess uptake within 1 g of MOF. Here, ρ_{H_2} is the density of hydrogen at 77 K and 35 bar (11.5 kg/m^3) and $\bar{V}_{pore}$ is the volume within the pores of the 1 g of MOF (overbars show the amount per gram of MOF instead of per cell as in the original description)

Recently, Siegel and co-workers [109] have successfully used this formula to estimate total H_2 uptakes. It should be noted that this formula finds the maximum theoretical uptake (upper limit) since the MOFs are freed from any guest molecules, which would be impractical in reality.

In this chapter, we combined two techniques, TDM and structural calculations, to explore broader published data. In the first part, we mined SA and PV values through TDM and estimated experimental H_2 uptakes based on the empirical formula using SA and PV values. We introduced two different approaches for merging SA and PV values. In the second part, we used crystallographic data to calculate the theoretical H_2 adsorption limit from SA and PV values based on structural calculations and compared to experimental data.

3.2. Text and Data Mining (TDM) Protocol

We used a TDM algorithm developed by Park et al [15]. Our mining protocol includes four steps: (1) creating the library of DOI numbers with MOF publications, (2) accessing corresponding full-text manuscripts, (3) extracting SA and PV values and creating SA and PV value tables, and (4) assigning unique SA and PV values for each MOF and merging both tables to estimate H₂ uptake (Figure 3.1).

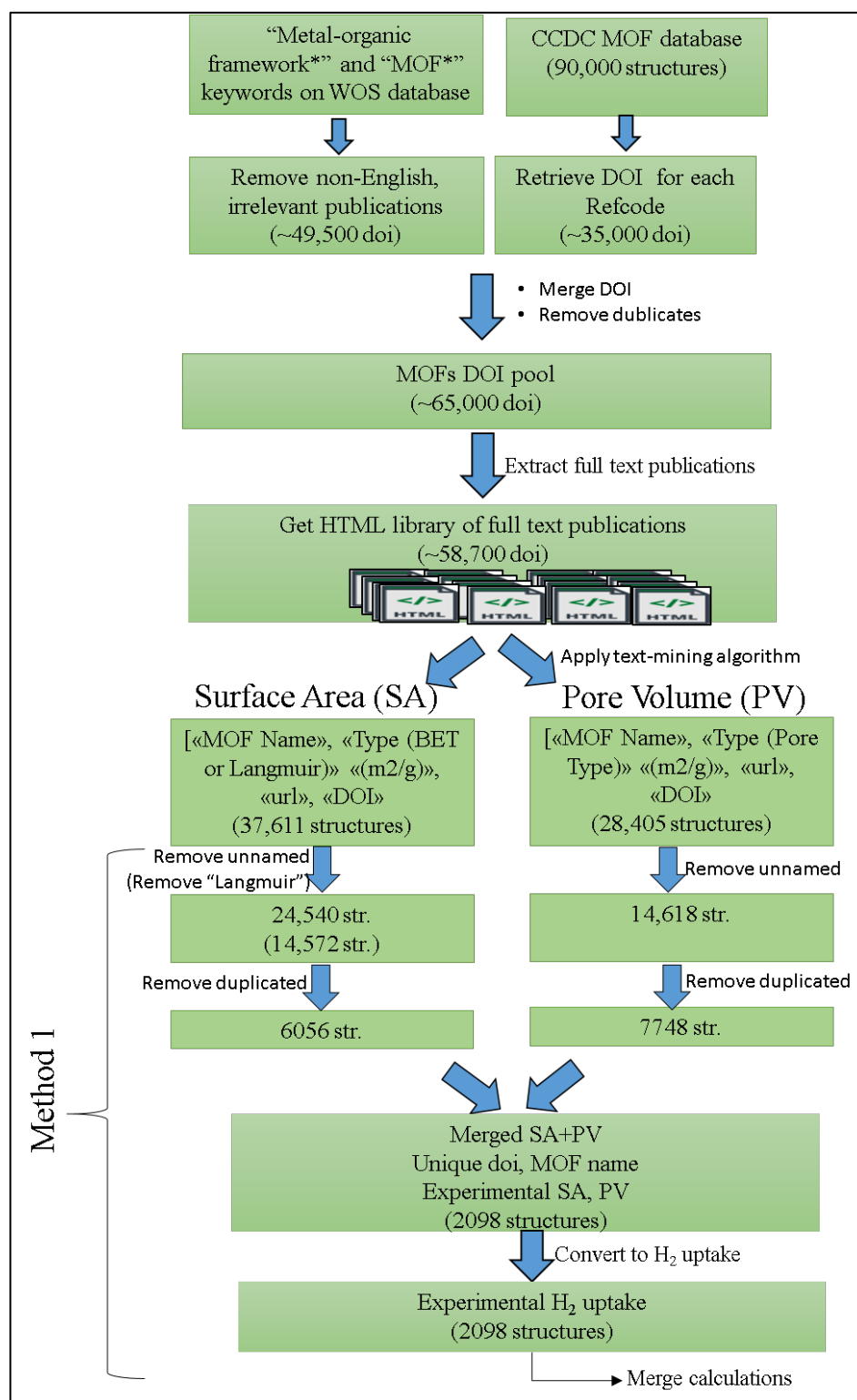


Figure 3.1: TDM algorithm flowchart.

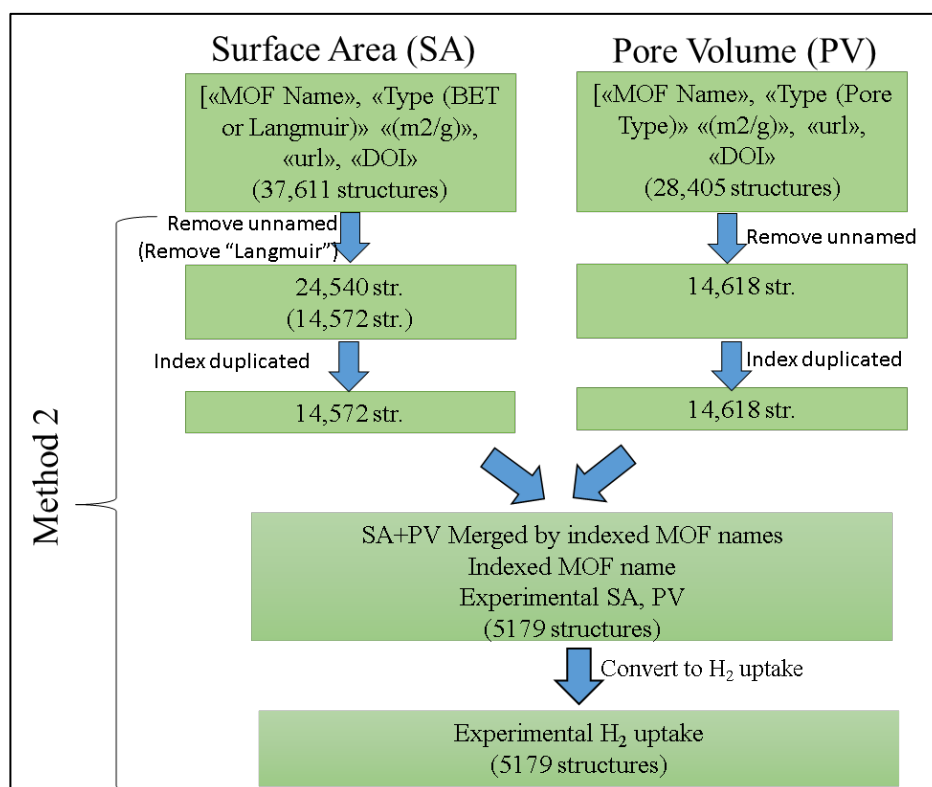


Figure 3.2: Flowchart for merging SA and PV values from TDM using Method 2

In the first step, we have used our own custom database as opposed to the study by Park et al., where they applied the algorithm to the CoRE database. We have created the database by searching all available MOF-related manuscripts using “metal-organic framework*” and “MOF*” keywords on the *Web of Science* database. Then, we have excluded non-English-language manuscripts and eliminated the following publication areas: history, physiology or behavioral sciences, or neurosciences neurology. As a result, we have created a library of total ~49 500 publications’ DOI numbers. In addition, we have also retrieved unique DOI numbers from another source, CSD MOF database using CSD python API and obtained ~35 000 DOI numbers. Finally, we have combined these two libraries removing duplicates and have eventually created a library of ~65 500 unique DOI numbers corresponding to as many MOFs as available in the literature.

In the second step, we accessed all of the full-text manuscripts in html/xml file formats in the custom library created in the first step. To handle large number of files, most publishers have developed the text and data mining application programming interface (TDM-API) with special permissions. For the publications where there is no TDM-API available, we have been limited to access only from the subscribed journals

using our custom python script that uses the CrossRef TDM-API service, a nonprofit organization offering researchers to easily retrieve full-text documents from all participating publishers. The publishers are mainly ACS, Wiley, Springer, RSC, Elsevier, and small contributions from other sources. Among nearly 65 500 DOI numbers, the algorithm by CrossRef has failed to generate the links for the manuscripts of 1863 DOI numbers. In addition, for 4964 DOI numbers, only abstract files could be downloaded. This was mainly due to two reasons: we have no subscription to the requested journal or the full text of manuscript is very old and available only in pdf format (not in html/xml format). Thus, the number of files in DOI pool is dropped to ~58 700.

In the third step, we utilized the TDM algorithm to extract surface area (SA) and pore volume (PV) values from the full texts obtained in the second step. The details of the algorithm have been described elsewhere [15]. Briefly, the algorithm converts an html/xml file into text format. Next, the text is grouped according to sentences by “.” period mark to create sentence-level tokens. After that, each sentence is further divided into words by “empty” spaces to create word-level tokens. The words are stored in a two-dimensional array format for extracting useful information. Finally, two separate tables belonging to SA and PV values are created extracting the data related to a few specific keywords. The table is composed of “MOF name”: common name of the MOF (if exists); “Units”: unit of the surface area value in m^2/g (or unit of pore volume value in cm^3/g); “Value”: the numerical value of SA (or PV); and “Keyword”: the type of the surface area whether a BET or Langmuir. Two tables belonging to SA and PV values of MOFs were created at this step.

The last step is to clean the two tables and assign a unique value for each MOF, and finally combine these two independent tables to be able to use Equation

(3.1) for the prediction of H_2 uptakes. The combined table of unique MOFs with SA, PV, and H_2 uptake values contains 2080 data (method 1 in Figure 3.1). Another way to merge SA and PV values is to create an index and match for each of the sorted value belonging to the same MOF name after unnamed MOF removal step (method 2). In this case, we assumed that a greater SA value of the corresponding MOF will also have a greater PV value. If the number of SA and PV values are uneven, the odd number with the smallest value of SA or PV is simply disregarded.

The text-mined SA and PV values are estimated to be “experimental” in 98% of the cases. Equation

(3.1) has been used to estimate the H₂ uptake in each MOF after the SA and PA values were assigned to each MOF. The tables including experimental SA and PV values of MOFs were compared to the values predicted from structural calculations. Further comparisons of predicted H₂ uptakes from experimental and calculated SA/PV values were carried out.

3.3. SA, PV and H₂ Storage of MOFs by TDM

Among our custom full-text html/xml library of ~58 700 publications, only 10 192 manuscripts comprise 37 612 surface area values and 6103 manuscripts contain 28 520 pore volume values. We have analyzed the distribution of these manuscripts over different publishers (Figure B.1) and the success rates of algorithm's SA/PV value extraction by randomly chosen 100 values from each list. For the first time, we have methodologically applied two different refining approaches to clean the raw data and ultimately report experimental SA and PV values for all MOFs available in the literature. It is also possible to predict H₂ uptakes estimated from experimental SA and PV for MOFs.

In the SA table, MOF name refers to the common name of the structure if a name exists and the algorithm can capture. If it fails to capture any name for the MOF, it assigns the MOF name as "IDK". Our script cannot find a name for 12 542 rows. A primary reason for such large number of unassigned MOF names is the fact that most MOFs available in the literature do not have a common name and thus the algorithm fails to assign a name in the given MOF name description sets. Indeed, a systematic naming method does not exist in MOFs. This is beyond the scope of this study here, but developing a systematic naming procedure would have a great impact on the field. In addition, the algorithm sometimes assigns unrealistic MOF names such as having formed from just numbers, just symbolic initial characters, and too long names (>50 characters). In addition, 10 201 MOFs with either "Langmuir" SA or "unassigned" SA type were removed. Apart from these, some SA values are also problematic, such as having non-numeric and unexpected values (SA > 15 000 m²/g, currently reported greatest value) [116]. The toll number drops to 14 279 (BET only) after this step.

One should note that these data include some repeated MOF names with different SA values. Having multiple SA values for an MOF is common since different researchers can report different values for the same MOF (different conditions, better

solvent removal, etc). In addition, the same MOF can be reported in more than just one manuscript. The number of unique MOF names for 14 279 values in the SA table are only 5975. Therefore, one of the most difficult tasks is to assign a unique SA value for an MOF among the values that are collected by the algorithm. However, this task is a prerequisite only for merging the SA and PV values using method 1. To reduce our data to unique MOF names and SA/PV values, the flowchart in Figure 3.3 was used.

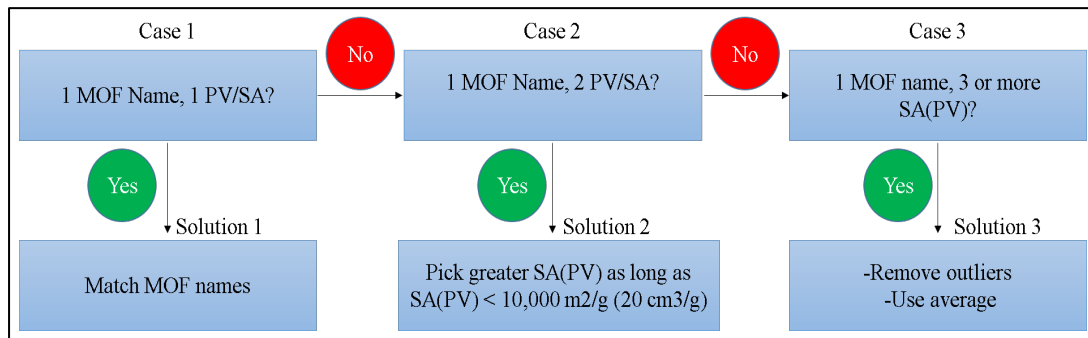


Figure 3.3: Flowchart for assigning one SA(PV) values for every single alphanumeric MOF names.

Among 5975 unique MOF names in the SA table, 4482 MOFs (74%) have already single SA value (case 1 in Figure 3.3). In addition, there are 836 MOFs (14%) with two SA values (case 2). Only 657 of MOFs (11%) are with three or more SA values (case 3). When there is one MOF name with only one SA and PV values in the table, it is straightforward to assign the value for that MOF name (case 1). Among the MOFs, 74% in SA table and 70% in PV table are already with single values. When there exist two values in SA or PV tables for the same MOF name, we assigned the greater value for that name (case 2). When there are at least three values (case 3) for the same MOF name in either table, we chose the most frequently occurred values unless otherwise the maximum value was assigned. Preliminary to the assignment in case 3, we also applied outlier removal protocol according to Tukey's fences [117]. After eliminating outliers and assigning a single SA value for each MOF, the number of unique MOFs with SA values was reduced to 5975. This is the ultimate SA table for method 1.

In the PV table, the MOF name and values were treated the same as SA table. The raw data of 28 520 rows in the PV table was updated to 13 889 after removing 10 451 unnamed and 4180 unrealistic MOFs. The number of unique MOF names in the 13 889 data of PV table is 7481. Among these, 5312 MOFs (71%) have already

single PV values (case 1), while 1347 MOFs (18%) had two PV values (case 2) and rest 822 (11%) data belong to MOFs with three or more PV (case 3). Using the same approach as the SA, a unique value was assigned for each MOF name in the PV table.

In this study, we report 14 279 SA values belonging to 5975 unique MOFs and 13 889 PV values belonging to 7481 unique MOFs. This can be compared to the number of SA and PV values reported in the literature [15], where there are 999 SA values belonging to 490 MOFs and 584 PV values belonging to 407 unique MOFs. To the best of our knowledge, we created the most extensive collection of experimental SA and PV values that can be referred. When comparing common MOFs in our data with those in the literature, both SA and PV values show almost perfect agreement with only a few deviations, mainly attributed to assignment failures in cases with multiple SA/PV values [15]. The Figure 3.4 shows only common MOFs (387 SA and 311 PV values) reported in this study and the literature.

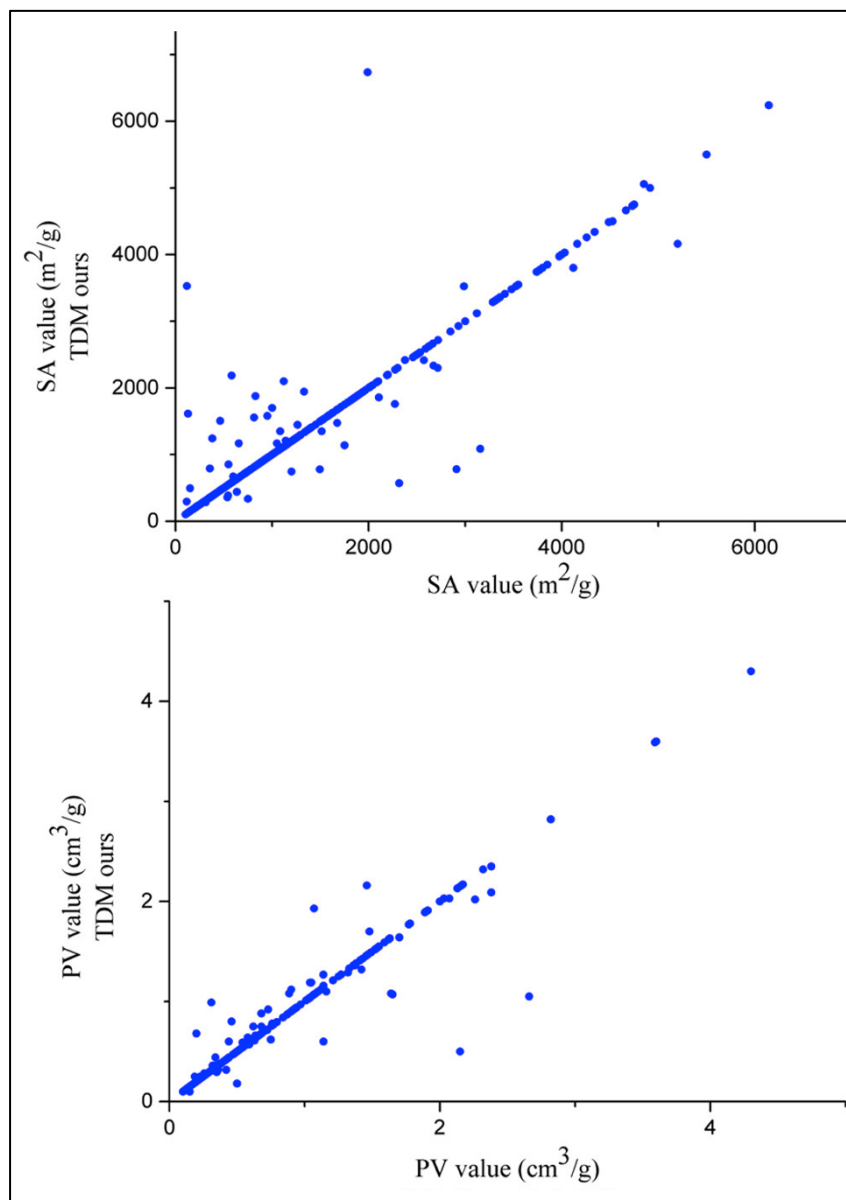


Figure 3.4: Comparison of TDM-extracted SA (top) and PV (bottom) values to those reported in the literature.

In the final part of method 1, two tables belonging SA values of 5975 MOFs and PV values of 7481 MOFs were merged by matching MOF names. There are 2080 common MOF names in the final table that includes unique MOF names, and SA and PV values.

After merging the data, we used Equation

(3.1) to estimate H₂ uptakes. To predict the H₂ uptakes, one more correction on the final merged data needed to be carried out. The merged table has SA and PV values that are predicted to be experimental in 98% of the cases by manual verification of 200 randomly chosen values (100 SA + 100 PV). However, the algorithm does not identify

what experimental conditions, such as pressure range or temperature, were applied in the experiments (we specifically picked BET). Since the equation is valid only at a temperature of 77 K and a pressure of 35 bar, this verification is critical. Manual verification of 200 data over the 2080 merged table also showed that most of the experiments were performed at the temperature of liquid nitrogen (77 K), and the pressure was scanned in the 0–100 bar range using N₂ gas as adsorbate. Therefore, both SA and PV values in the final merged table are of a N₂ probe. Table B.1 shows the first 20 values with top H₂ uptakes along with SA/PV values and experimental conditions.

To assess the performance of the algorithm in retrieving correct values, we randomly picked a subset of data that contains 100 values for each SA and PV. In the SA table, manual confirmation of 100 values over the 24 445 data showed that 71% of SA was retrieved correctly. This is in a complete agreement with the developers of the algorithm [15]. A similar verification for another set of 100 values after reducing the data to 5975 showed that the success rate of SA has improved to 78% by using the flowchart in Figure 3.3. In the PV table, success rates of PV have always been greater than that of SA values. The rate was 77% over the 13 889 data and 82% over the 7481 data when tested with randomly chosen 100 values. We also evaluated the success of the algorithm on the merged data. Another set of 200 randomly picked data over the 2080 data has shown that the algorithms of both SA and PV values are found correctly with 73% success. Table 3.1 summarizes the performance tests carried out at various steps.

Table 3.1: Performance tests of the TDM algorithm at various steps

		# data	# test	% success
SA	Raw data	37611	-	-
	Without unnamed	24540	100	71
	Without duplicates	6056	100	78
PV	Raw data	28405	-	-
	Without unnamed	14618	100	77
	Without duplicates	7748	100	85
SA+PV Merged	Method 1	2098	200	80
	Method 2	5287	100	83

To cover broader data in the table of merged SA and PV values, we also used an alternative method for refining MOFs (Method 2). To match SA and PV data, we have to match each unique MOF name in both tables, and this leads to elimination of considerable data in method 1. Assigning only a single SA/PV value and removing all others may result in noteworthy data loss. Therefore, instead of removing the multiple SA values for an MOF structure, we also made an index of values. For instance, there are seven SA values with MOF name Co-MOF on the raw data of SA table. Sorting values from largest to smallest, we created hypothetical Co-MOF_1, Co-MOF_2, etc., indexed MOF names for these seven data, and assigned SA values to each of these names. The idea of doing this is to avoid removing the multiple SA values predicted by the algorithm for the same MOF. Assuming that an MOF with a greater SA value will also have a greater PV value, the new indexed MOF names may be matched in the SA and PV tables. This assumption can be easily verified by comparing the calculated SA and PV values for MOFs (Figure 3.5)

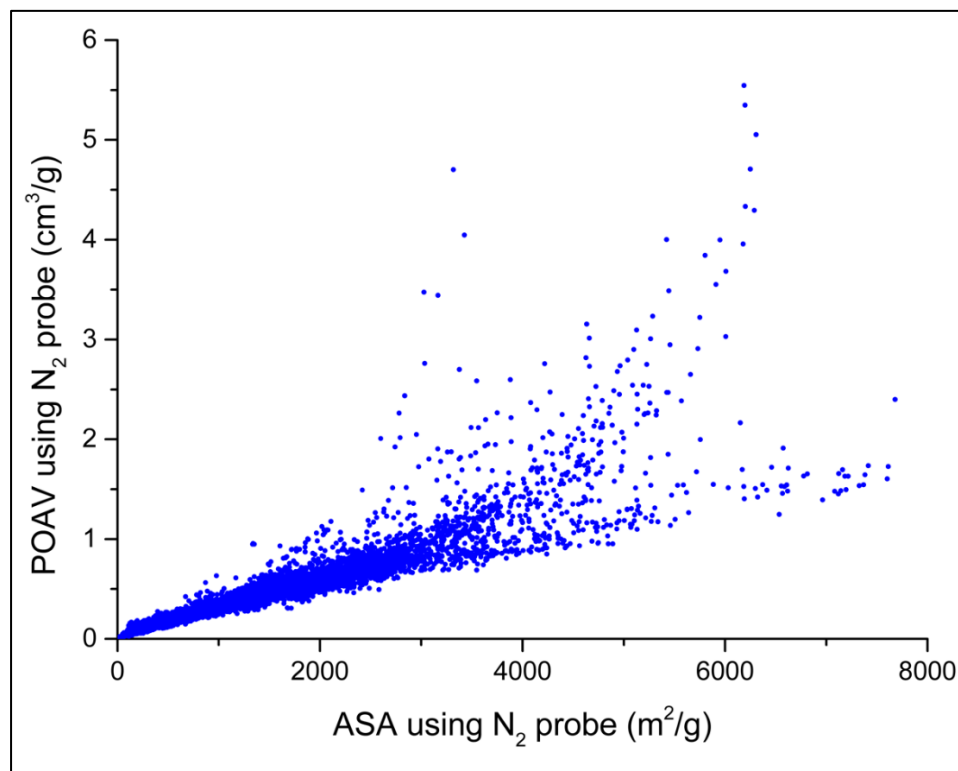


Figure 3.5: POAV vs ASA calculated by Zeo++ software. The greater the SA, the greater PV values.

Using this approach, the 14 279 SA and 13 889 PV values could be merged into a table of 5198 items in which common MOF names are indexed according to the value ranking.

Although it is not as efficient as method 1 and thus less reliable, it is a more expanded MOF table with merged SA and PV values. In this method, the initial numbers of 14 279 and 13 889 data with valid MOF names in the SA and PV tables were dropped to 5198 merged table of indexed MOFs. The success of matching correct experimental SA and PV values for an MOF in this method was found to be 66% by manual verification of randomly picked 100 over 5198 data.

3.4. Structural Calculations Protocol

By “structural calculations”, we refer the data purely based on theoretical calculations using single-crystal x-ray diffraction structures. Our structural calculation procedure includes three steps: (1) retrieving the coordinate files of non-disordered MOFs, (2) cleaning structures by guest molecule removal algorithm, and (3)

calculating H₂ uptake indirectly by using empirical formula that uses calculated SA and PV values (Figure 3.6).

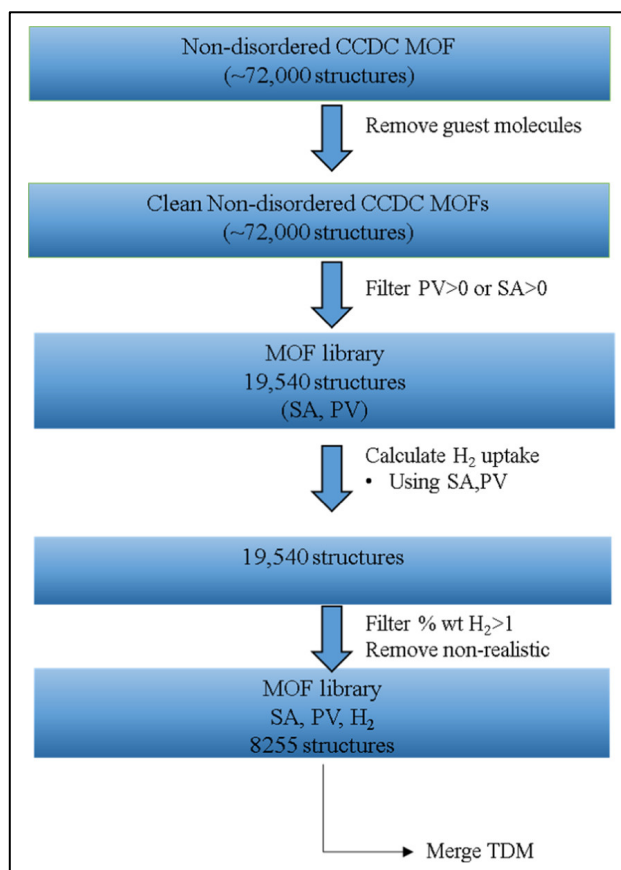


Figure 3.6: Flowchart for structural calculations purely based on single-crystal x-ray diffraction structures.

Since each structure in the CSD database has a specific Refcode, handling a large amount of data is much easier than handling text mining. However, not all structures are computation-ready to perform calculations. A cleaning procedure to remove bound or unbound solvent molecules needs to be applied. Although there are a couple of different methods to remove guest molecules, we used the algorithm by Moghadam et al [108]. Nearly 72 000 non-disordered crystal structures have been retrieved from the CSD database using CSD Python API [118], and solvent removal algorithm has been applied to all structures. We calculated SA and PV values of all 72 000 molecules using a 1.20 Å H₂ probe radius for the pore volume and 1.86 Å N₂ probe radius for SA calculation since the empirical formula is valid only for these parameters [109]. Finally, the structures with H₂_uptake=1 values were eliminated.

As discussed in chapter 2, in the SA and PV calculations, we used the recently developed Zeo++ v0.3 software [93, 94]. The program calculates the accessible surface area (ASA) and probe-occupiable accessible volume (POAV),[16] which was shown to mimic the experimental accessible volume (AV) better than other volumes such as geometric (Gm) and probe center (PC).

3.5. SA, PV and H₂ Storage of MOFs by Structural Calculations

We calculated the H₂ uptake based on structural calculation of MOFs to make connection to our TDM. As of 2019, there are more than 90 000 total and 72 000 non-disordered structures in CSD database that are assigned as MOFs.

Table 3.2: MOF Databases Comparison. UM shows the number of structures with surface area, pore volume, and gravimetric hydrogen uptake percentage greater than 0 or 1. Note that all databases in the table are derived from CSD.

Database	# Str.	#SA>0	#PV>0	##H ₂ >0	##H ₂ >1
CSD	72,609	4830	4840	7345	4005
CSD (cleaned, ours)	72,609	9875	9877	19,540	8255
CoRE (Extended)	14,728	10,103	14,728	-	-
UM	10,881	8003	10,881	10,661	3388

It should be noted that both CoRE and UM databases originally get structures from CSD database but apply different criteria in classification and cleaning procedure for compounds to be considered as MOFs. We used the cleaning algorithm of CSD Python API [118] using all available CSD non-disordered MOF subset of 72 609 (as of 2019, v. 5.40) [108]. The number of MOFs with surface areas and pore volumes that are greater than zero is found to be only 4830 and 4840, respectively. However, upon applying cleaning algorithm, these numbers increased (Table 3.2)

To show the performance of cleaning procedure, we have analyzed the surface areas, pore volumes, densities, and H₂ uptakes based on equation

$$(3.1).$$

Figure 3.7 shows the cleaning performance of the algorithm over 72 609 data. All of the values are tremendously improved upon cleaning. Thus, it provides better estimates for the hydrogen storage by means of maximum theoretical uptake limit. We have further tested the algorithm's cleaning success by randomly choosing 100 structures and manually verifying whether the algorithm cleaned the structures correctly. Over the tested data, 90% of the structures were cleaned perfectly. The possible reasons for the failure of the algorithm in cleaning process for the 10% were discussed in detail by Keskin et al [110]. Briefly, the algorithm has a list of possible guest molecules that sometimes were used as a linker in the MOF architecture. In the cleaning process, these linkers are also removed. This is the major complication of the process that stems from algorithm.

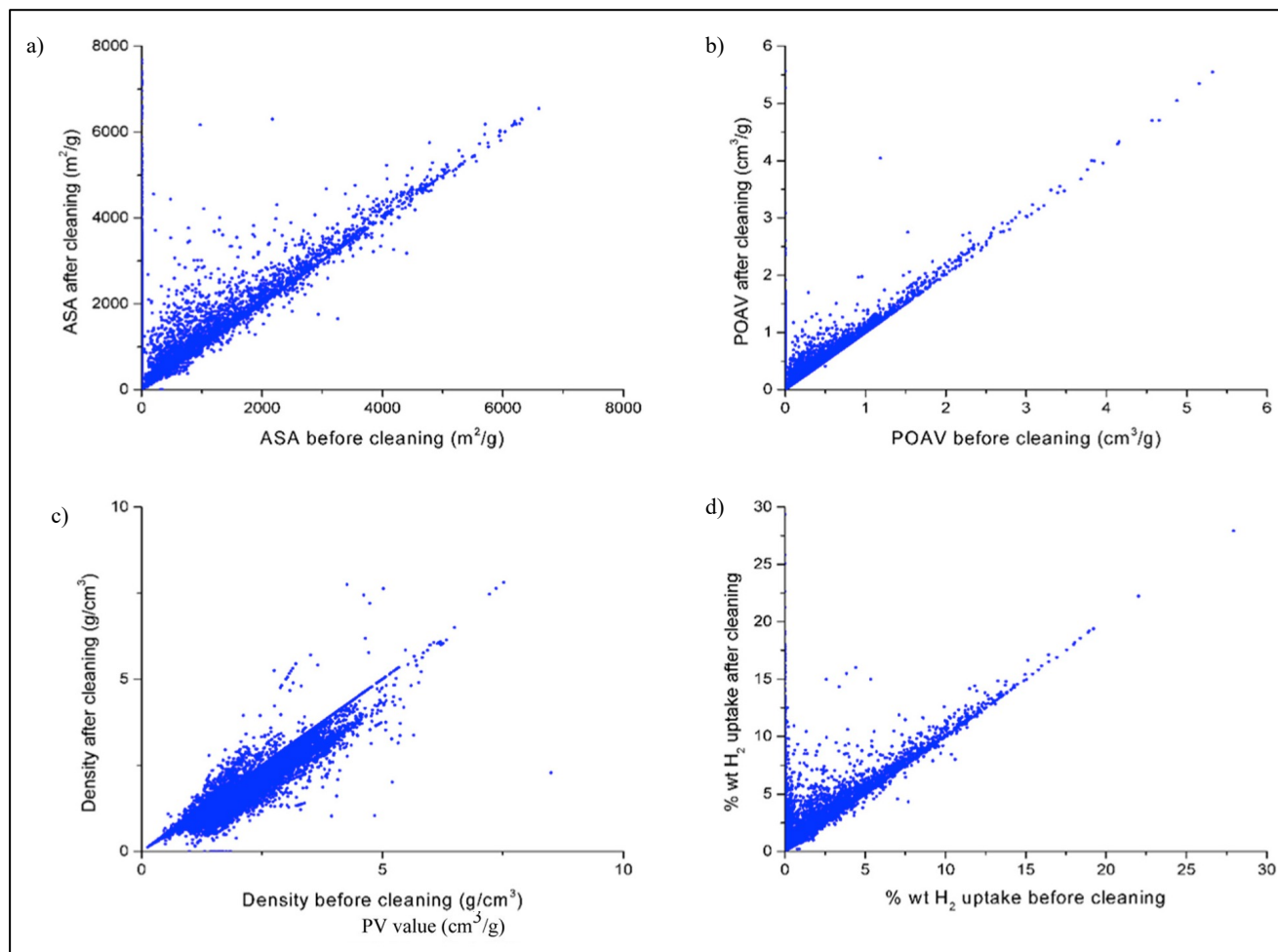


Figure 3.7: Cleaning performance of the algorithm for a) SA, b) PV, c) density, and d) H_2 uptake values.

On the other hand, SA and PV values upon the cleaning process have been recently published using another algorithm by Siegel, from which two new databases of CoRE (extended) and UM databases have been introduced. We also checked our results of SA and PV values of cleaned structures (using CSD cleaning procedure) with the newly released SA and PV data of cleaned structure by the other algorithms.

In Figure 3.8, we compared all three algorithms' surface area and pore volume calculations after cleaning the structures. Although our cleaned SA and PV values have been improved by these two new algorithms, it does not affect overall agreement on the calculations. These new algorithms seem to have less improvement in estimating PV values.

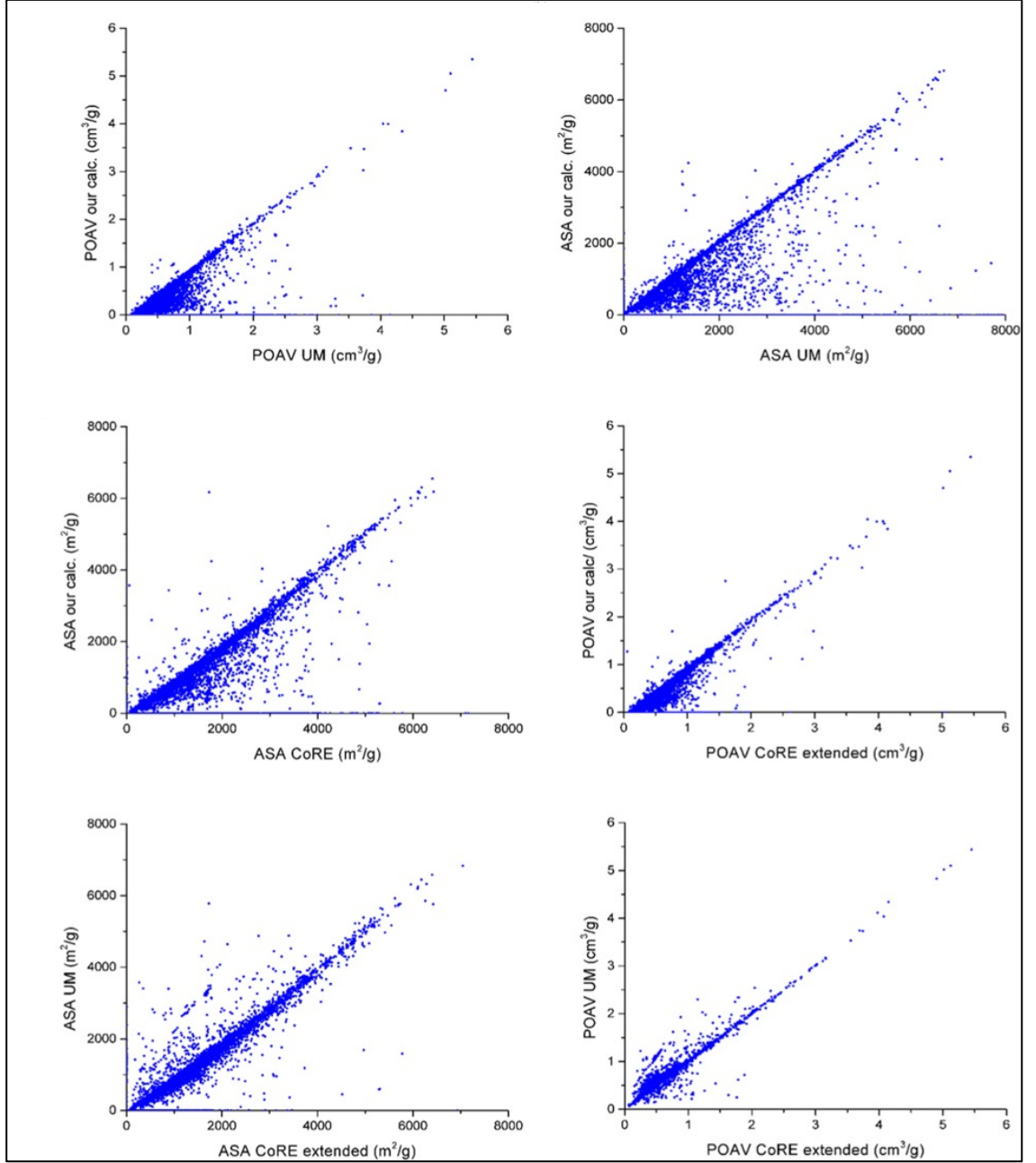


Figure 3.8: SA and PV values compared by our calculations (CSD algorithm) and two other algorithms of CoRE and UM (both derived from Siegel's algorithm). The data show that the last two algorithms are more successful in cleaning.

In addition to SA and PV values, which we used to calculate H_2 uptake via Equation

(3.1), we have also compared our H_2 uptake values with GCMC calculations available in the literature [112]. The H_2 uptakes using our algorithm and GCMC calculations are in a good agreement (Figure 3.9). The major difference in the calculations is due to the cleaning procedure.

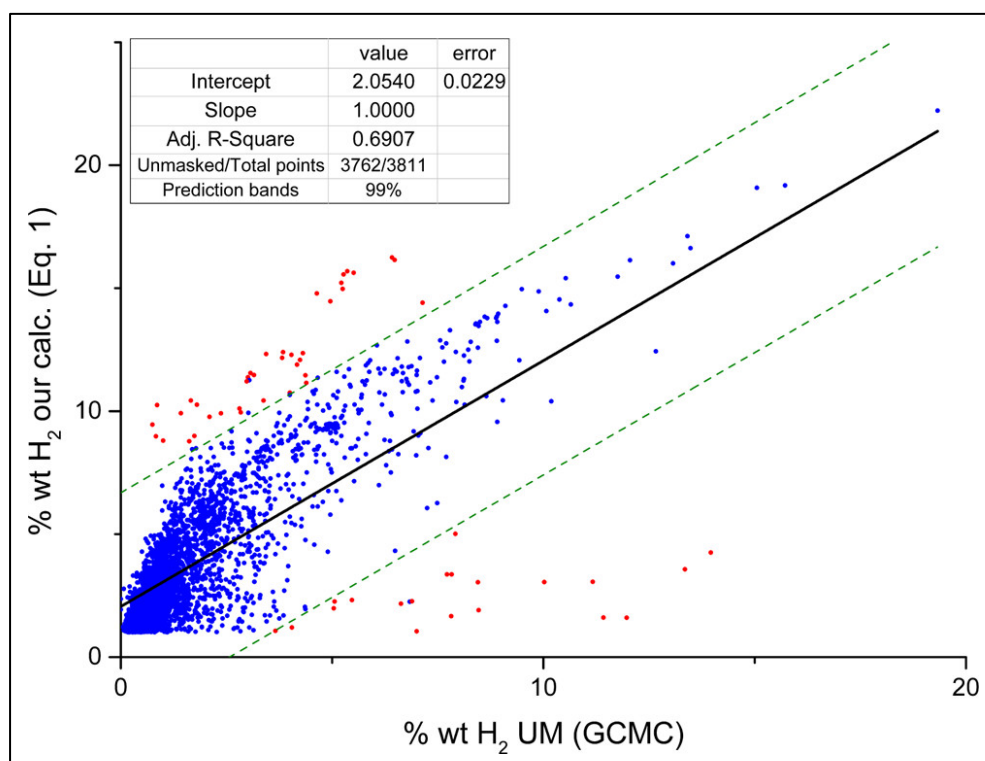


Figure 3.9: Hydrogen uptakes (>1.0%) calculated by empirical formula and GCMC calculations. The data belonging to GCMC calculations (y-axis) were reprinted with permission from Copyright 2019 ACS. The diagonal dashed line is used to differentiate between the H₂ uptakes by the two methods.

3.6. Merging TDM with Structural Calculations

To get insights into the success of the TDM algorithm and success of calculations, we have compared our TDM values (mostly experimental) with structural calculations. Merging the data of TDM and structural calculations is limited to the common MOF names in both the CSD database and TDM-extracted data.

Prior to comparing TDM to structural calculations, we retrieved MOF names (if exists) using CSD API by “synonyms” keyword in the database. Among 72 609 structures, only 3199 have MOF names. Those structures in the database with the same MOF names were treated using the same protocol described in Figure 3.3. After assigning single SA/PV values for each common MOF name in the CSD database, the total number has dropped to 2499. Upon assigning unique SA/PV values for each MOF in the calculated table, it could be possible to compare values to TDM by matching alphanumeric MOF names. The number of matching rows in the two tables is only 296. Figure 3.10 shows the comparison of merged experimental SA + PV values and theoretical ASA/POAV values along with experimental and theoretical H₂ uptakes

using the empirical formula and GCMC calculations. The pore volume and surface area values are both linear between TDM and calculated ones. Although the correlation is poor, the linearity is more obvious in PV values than SA values. One reason to have poor correlation is that the success rates of the TDM algorithm is 80%. In addition, calculations strip the guest molecules that exist inevitably in reality.

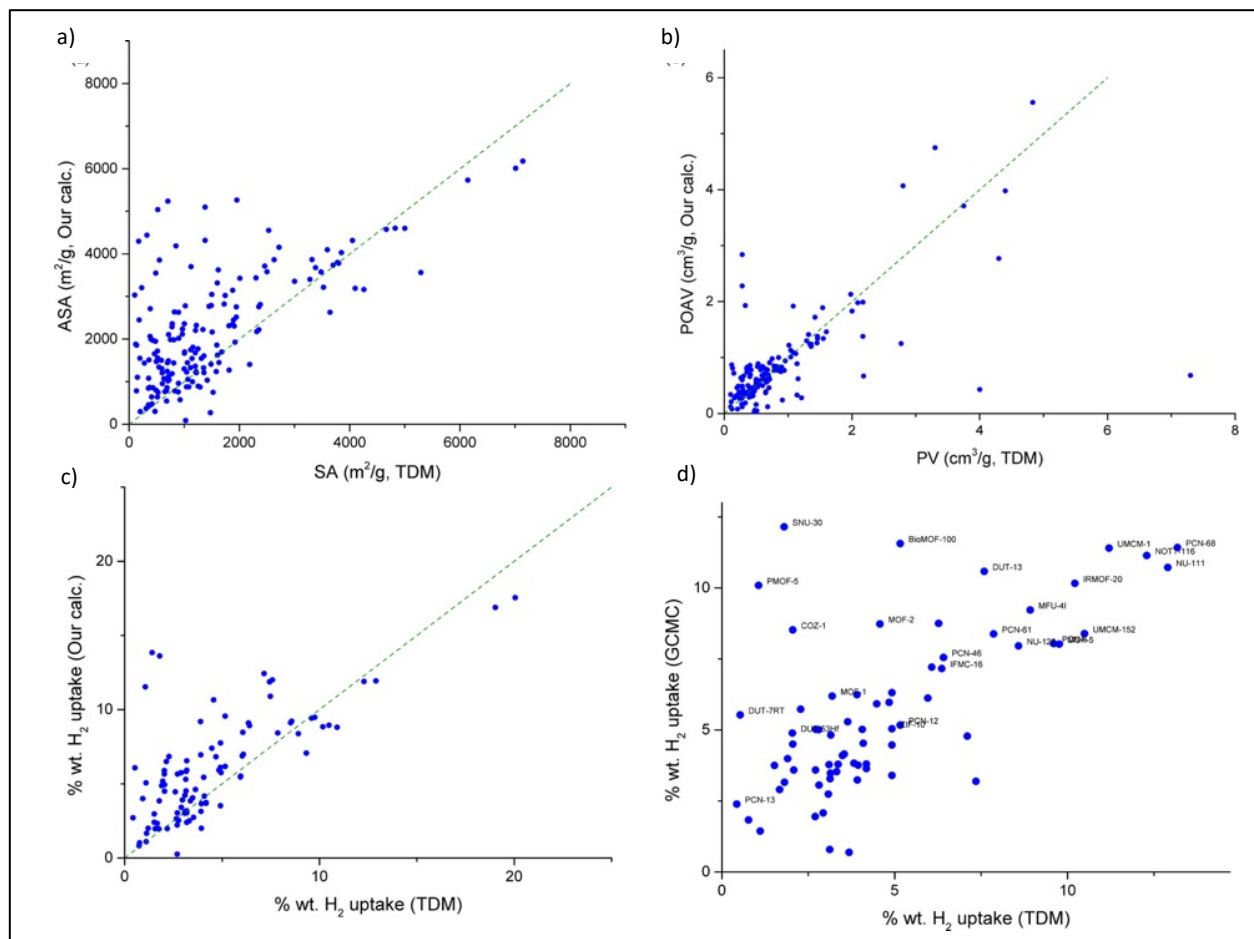


Figure 3.10: Comparison of experimental and calculated a) ASA, b) POAV and %H₂ uptake values with calculated by c) Zeo++ and d) GCMC by method 1 for merge of SA and PV values.

Assuming the validity of linear correlation between calculated and experimental SA/PV values, it is possible to speculate the experimental H₂ uptakes using the empirical formula in Equation

(3.1). It should be noted that experimental PV values are those of a N₂ probe since the experiments have been from mostly N₂ adsorption isotherm, whereas the PV term in the equation is of a H₂ probe. To understand whether PV with both probes different results have, we calculated occupiable pore volumes (POAV) using a N₂ probe radius of 1.85 Å and a H₂ probe radius of 1.20 Å (Figure 3.11). Overall, both values are almost identical with a proportionality constant of 0.9887. It is clear that using a N₂ probe in

(3.1) instead of H₂ should not alter H₂ uptake values.

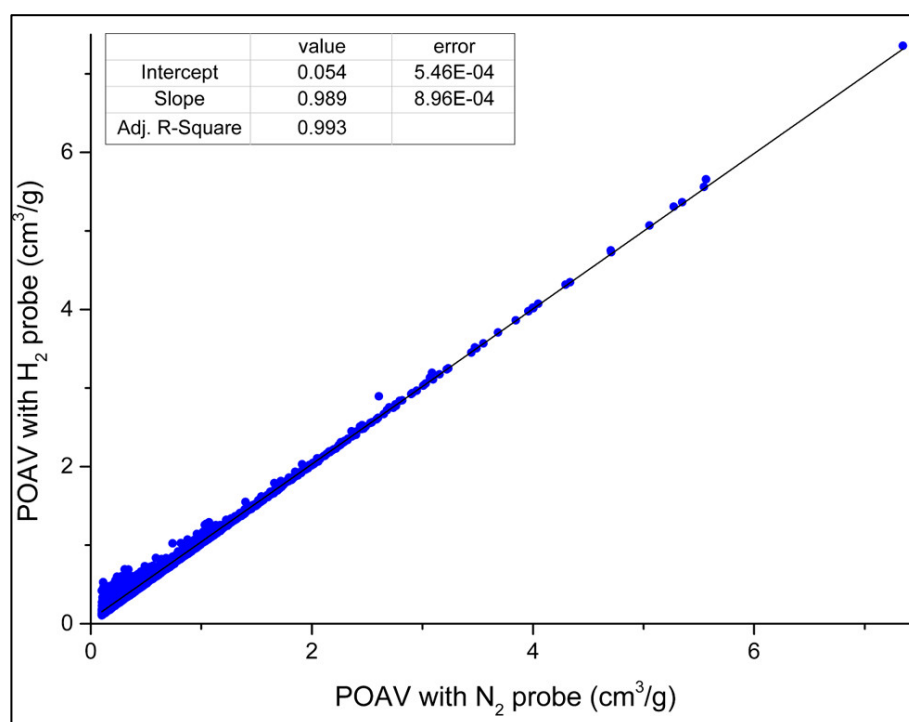


Figure 3.11: Probe-occupiable pore volumes (POAV) calculated using a H₂ probe (radius = 1.20 Å) vs N₂ probe (radius = 1.86 Å).

Similar to SA and PV values, H₂ uptakes also show significant correlation. The GCMC calculations better agree with H₂ uptakes predicted by TDM. Although very limited number of data (~300) could be compared based on MOF names, it still indirectly indicates the success of broader experimental values (2080) using TDM algorithm.

In addition to comparison of TDM with structural calculations using method 1, we analyzed the same comparison using method 2, in which MOF names are indexed rather than reduced to single values. Subsequently, using the same approach, data were merged. A similar linear correlation has been found (Figure B.2).

A better comparison might be to show the distribution of all data without matching MOF names. The advantage of doing this is that it does not require the assignment of unique SA and PV value for each single MOF. Thus, more data can be compared to calculations and/or literature values. Using TDM, it has been possible to retrieve 14 572 SA data (BET only) and 14 618 PV data without assigning unique values to each MOFs. Figure 3.12 shows the histograms of SA/PV values in both experimental values retrieved by the TDM algorithm and calculations carried out by different algorithms (our calculations by CSD algorithm, UM calculations [112], and CoRE extended [111]). All of the calculations are in perfect agreement with experimental data.

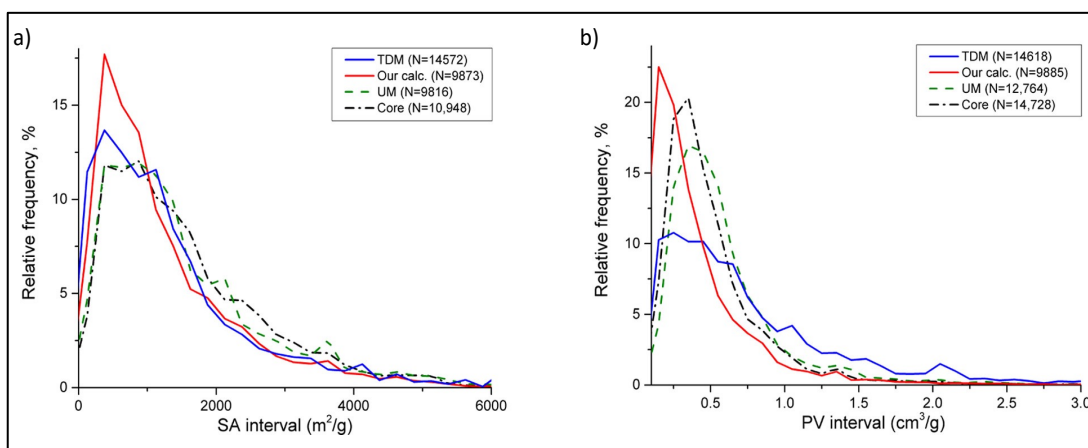


Figure 3.12: Comparison of probability histograms of a) surface area (SA) and pore volumes (PV) among experimental text and data mining (TDM, solid blue), our calculations (solid red), UM calculations (dashed green), and CoRE calculation (dashed black).

3.7. Conclusions

Herein, we have applied a text and data mining (TDM) technique to search for experimental surface area (SA) and pore volume (PV) values of metal-organic frameworks (MOFs). We collected SA and PV values in two separate tables of each nearly 14 600 values. Manual verifications on randomly chosen values showed ~80% success for SA and 85% success for PV values in a complete agreement with literature

values that are limited to ~ 1000 data. In addition, we have introduced two new methods to compare and merge available SA and PV values based on MOF names, ending up with an ultimate table of TDM experimental SA and PV values with nearly 2100 data. Further manual verification of randomly selected test data showed $\sim 80\%$ success in merged data. Using an experimental equation correlating SA and PV values to H_2 uptakes, we predicted H_2 uptakes of nearly 2100 MOFs.

In addition to TDM, we calculated accessible surface area (ASA) and probe-occupiable pore volumes (POAV) purely based on the coordinates of single-crystal x-ray diffraction structure using spherical probes of N_2 and H_2 with 1.86 and 1.20 Å, respectively, as a comparison to and cover of our TDM and available literature data. Our structural calculations based on spherical probes were in a quite good agreement in our TDM. In addition, we compared TDM and structural calculations by means of probability histograms according to SA and PV value intervals with perfect agreement.

4. A NEURAL NETWORK POTENTIAL FOR THE IRMOF SERIES AND ITS APPLICATION FOR THERMAL AND MECHANICAL BEHAVIOR²

4.1. Introduction

As mentioned in Chapter 1, MOFs with their highly ordered microscopic structure have attracted scientific interest due to their large surface areas and pore volumes, high flexibilities, and capabilities of reversible processes. Most of the applications require well defined process conditions such as pressure and temperature. For instance, thermal expansion due to higher temperatures might lead to deformation of plasticity in materials [119, 120]. In particular, adsorption and desorption processes of guest molecules without heat control system might cause drastic temperature changes in MOF structures. On the other hand, adjusting temperature can be beneficial for a controlled adsorption and desorption of guest molecules in some applications such as temperature swing adsorption (TSA) [121, 122]. Thus, thermal properties of MOFs are still interest of investigation. The most interesting thermal property is that most frameworks show negative thermal expansion (NTE), the decrease in the total volume upon heating. This property becomes critical and can be utilized when designing on-board energy related (CH₄ and H₂) storage materials [119]. It also creates thermal pressure that plays important role in the mechanical stability of the frameworks. Adsorption and desorption properties of MOFs have been extensively studied by both experimental and theoretical methods [14]. However, there are few theoretical studies on how adsorption/desorption properties are affected by applied mechanical pressure [123]. Computational studies on adsorption/desorption properties as well as thermal behaviors of MOFs rely on classical force fields. Understanding the atomistic details of host-guest interaction including the reactivity and flexibility of MOFs requires the first-principle accurate methods like density functional theory (DFT).

² This chapter were adapted from O. Tayfuroglu, A. Kocak, and Y. Zorlu, "A neural network potential for the IRMOF series and its application for thermal and mechanical behaviors †," *Phys. Chem. Chem. Phys.*, vol. 24, no. 19, pp. 11882–11897, 2022 (Creative Commons Attribution 4.0 International. Url:<https://creativecommons.org/licenses/by/4.0/>).

On the other hand, MOF structures may include thousands of atoms in unit cells and thus, the first principles definition of such large systems are quite costly in terms of computation [124]. In addition, the existence of multiple transition metals with open shell structures (in inorganic subunits) adds more complications to the calculations. Furthermore, dynamic definition of physical and chemical processes requires extensive sampling of the system with MOFs and/or guest-molecules. Application of ab initio molecular dynamics to bulky MOFs with multiple metal centers is not very practical with the current computational technologies. Hence, new strategies to reduce computational cost without losing reliability are still in high demand.

Over the last decade, Machine Learning (ML) has been greatly applied in many different systems [124, 125]. Highly accurate and reactive ML potentials can be built to describe MOF systems with both organic and inorganic subunits for simulations. Rather than building electronic structure from the first principles at every time step, their great capabilities of learning of multidimensional potential energy surfaces (PES) allows to carry out much faster simulations [126–132].

In the construction of machine learning potentials, several artificial neural network (ANN) architectures have been developed that are mostly based on atomic contribution to the target property (e.g. potential energy surface) such as ANI-torch [133], high dimensional neural network (HDNN) [67, 134], deep tensor neural network (DTNN) [86] and SchNet [64]. Behler and coworkers were the first to construct neural network potentials (NNPs) with an HDNN architecture for MOF-5 (IRMOF1) [135]. The DFT reference energies of nonequilibrium geometries of MOF5 fragments have been used to train NNPs, which could predict bulk properties. This case study showed the applicability of NNPs on MOF like organic-inorganic hybrid systems. However, the pretrained model is not publicly available for performing DFT accurate simulations and the NNP was only limited to MOF-5.

4.2. Construction of NNP for IRMOF-n (n=1,4,6,7,10)

As mentioned in Chapter 2, in the construction NNP process, we used SchNet [132], which is an end-to-end machine learning model (open-source code is available in SchNetPack toolbox [136]). SchNet is a continuous-filter convolutional neural network model designed to predict molecular properties such as energies and atomic forces etc. modified version of deep tensor neural network (DTNN) framework [137].

It is a multilayer neural network consisting of atom embedding, multiple interaction blocks and prediction blocks. At the embedding layer, each atom type Z_i is mapped to feature vectors $x_i^0 = \mathbb{R}^F$, which is collected as a matrix $X^0 = (x_1^0, \dots, x_n^0)$. Here, F denotes the feature dimension and l is the number of layers for i 'th atom. The atoms are similarly represented in other layers and features of the layer, l is given by $X^l = (x_1^l, \dots, x_n^l)$ with $x_i^l = \mathbb{R}^F$. This layer is initialized randomly and refined during training by the interaction blocks. The interaction blocks compute pairwise interactions between all x_i^l within a certain cutoff and then refines atomic representations by a subsequent t step up to T times in total. This procedure is performed by using continuous-filter convolution layers with filter-generating networks. These atom-wise representations are processed to obtain the desired molecular property such as energy, E , by the summation of all atomic contributions, x_i in the prediction block.

$$E = \sum_{i=1}^n e(x_i) \quad (4.1)$$

In addition to calculating energy by reliable neural network (NN) models, SchNet also provides the force models obtained by differentiation of energy with respect to atomic positions.

$$\hat{F}_i(Z_1, \dots, Z_n, \dots, r_1, \dots, r_n) = -\frac{\partial \hat{E}}{\partial r_i}(Z_1, \dots, Z_n, \dots, r_1, \dots, r_n). \quad (4.2)$$

Rotationally and translationally invariant description of energy model allows conservation of both force field and energy. Details of the algorithm can be found in Schutt's works [132, 136, 137].

4.2.1. Generation of Training Dataset for IRMOF-n (n=1,4,6,7,10)

There are many factors that can impact the success of a machine learning (ML) application, including the choice of mathematical descriptors and the ML model itself. The quality of the ML training set is one of the most critical factors in determining the success of the application. Ideally, the training set should include all relevant and representative samples of interest, with minimal redundancy. This ensures that the ML model can learn efficiently and effectively from the data. That is, the training set serves as the foundation for the ML model to generalize to new, unseen data and make accurate predictions. Therefore, careful attention and effort should be given to

generating a high-quality training set, as it is a fundamental and indispensable component of any successful ML application. One of the most successful applications of ML is to train the QM produced potential energy surface for a given set of (equilibrium and non-equilibrium) coordinates, thus allowing system to have energy and force driven evolution by means of MD or MC simulations. This requires performing energy and force calculations using high level of methods (e.g., DFT) for MOFs that have complex structures with hundreds of atoms in a single unit cell in the generation of ML training data. However, this can be computationally challenging. One effective strategy to mitigate this challenge is to use fragmentation [135] and active learning techniques. This approach involves breaking down the large MOF structure into smaller fragments, and performing DFT calculations for the fragments rather than bulk structures. The information retrieved from fragments can be extrapolated for predicting the properties of the overall MOF structure. By using active learning techniques to iteratively select the most informative fragments for inclusion in the model, this approach can significantly reduce the computational cost and improve the accuracy of the predictions. Therefore, to create and refine our training data set, we followed a three-step process. In the first step, the MOF structures, of which the experimental 3D structures are known (from single crystal x-ray diffraction data) were separated into several fragments with different chemical environments. An initial dataset was generated by scaling and randomly displacing these fragments. The second step involves computing the DFT energies and forces for the dataset. Finally, an active learning technique was used to develop this initial dataset and obtain the final dataset. The details of these steps are described in the following subsections.

4.2.1.1. Fragmentation

An important aspect of achieving effective training is to carefully consider the characteristics of the descriptor and neural network (NN) model utilized when fragmenting molecular structures into smaller parts, as this can have a significant impact on the performance of the model. Since the SchNet uses atomic contributions, the environment of atoms must be clearly defined within a certain cutoff. Thus, the MOF structure has been divided into pieces (fragments) by finding chemically inequivalent atoms that form the bulk structure. Similar to Behler's work [135], we generated the different possible fragments so that each fragment includes all different

atom types and their neighbor atoms within a sphere of 6 Å cutoff radius. This cutoff distance was also used by the interaction block of SchNet. In creating MOF fragments, cutting from cyclic structures was avoided. In addition, the missing bonds were saturated with hydrogen to achieve charge balance.

We started with the fragments of IRMOF-1 as it is the base for all other MOFs and thus the fragments of IRMOF-1 are common in all other MOFs studied (Figure 4.1). We defined IRMOF-1 with 5 fragments. Other MOFs differ from IRMOF-1 by only an additional linker. The organic linker in IRMOF-1 is 1,4-benzenedicarboxylate (BDC) while BDC links with n-propoxy, cyclobutyl and fused benzene functional groups, in IRMOF-4, IRMOF-6 and IRMOF-7, respectively. The linker is 4,4'-biphenyldicarboxylate (BPDC) in IRMOF-10. Therefore, an additional fragment for each of the other IRMOF-n (n=4, 6, 7, and 10) was added to the five fragments of IRMOF-1 with a total of 9 fragments that cover all MOFs studied. The common five fragments for all MOFs and additional differing fragments belonging to each MOFs are present in Figure 4.2.

There are various methods available, such as biased or unbiased molecular dynamics (MD) simulations at different temperatures and conformer generation techniques that can be utilized to construct the initial structural dataset [65]. However, in our case, we opted to employ a scaling and random atom displacement technique within a defined cutoff distance, which is better suited for working with fragments. To generate an initial structural data set of 6750 reference fragments, the following process was employed: Firstly, the bonds of each of the 9 fragments were adjusted incrementally from 0.95 to 1.10, using a step size of 0.0002, resulting in a collection of 750 x 9 structures with non-equilibrium geometries. Next, the atomic positions were randomly distorted in the x, y, and z directions, with a maximum displacement of 0.16 Å.

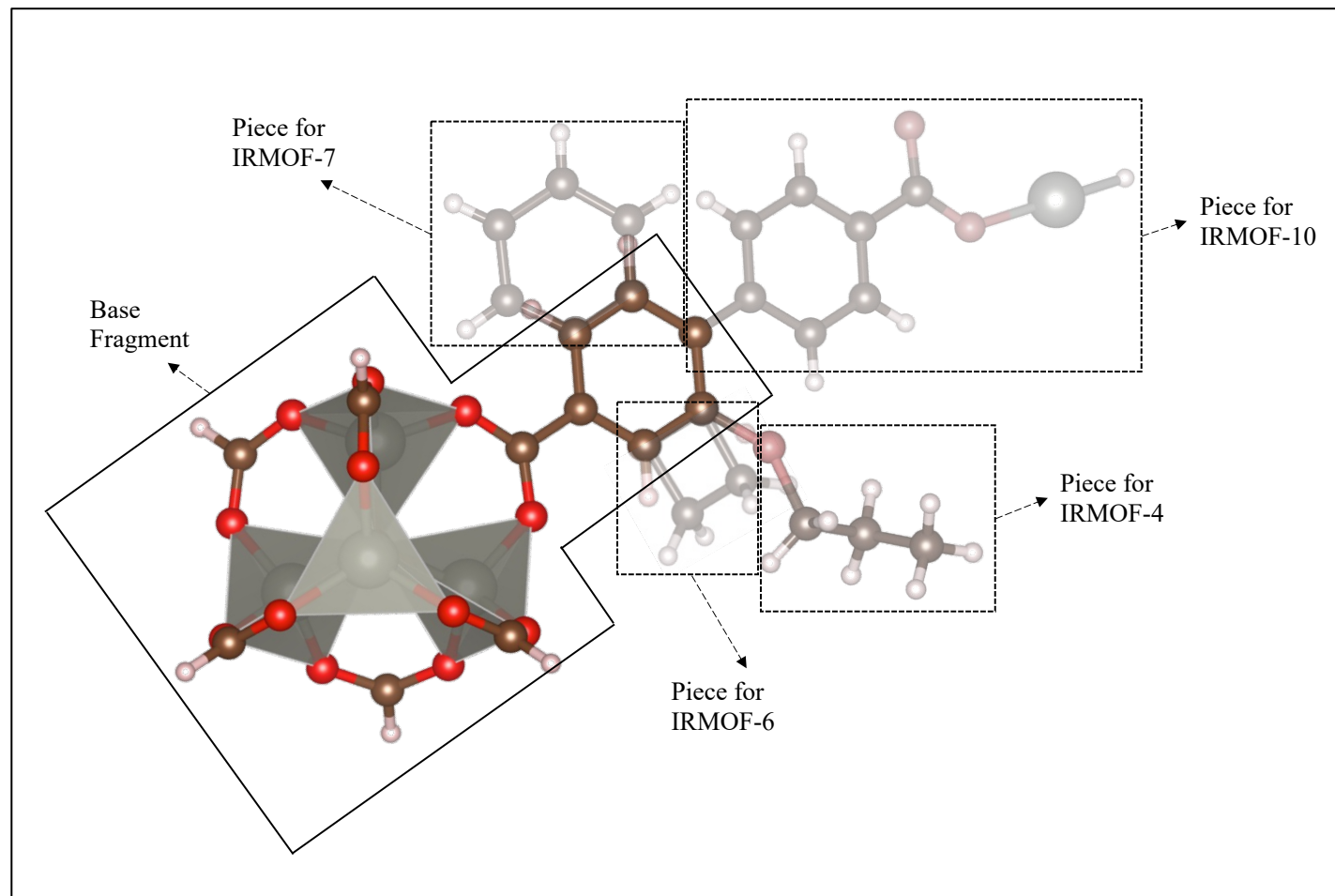


Figure 4.1: Systematic Fragmentation of IRMOF-n (n=1,4,6,7,10).

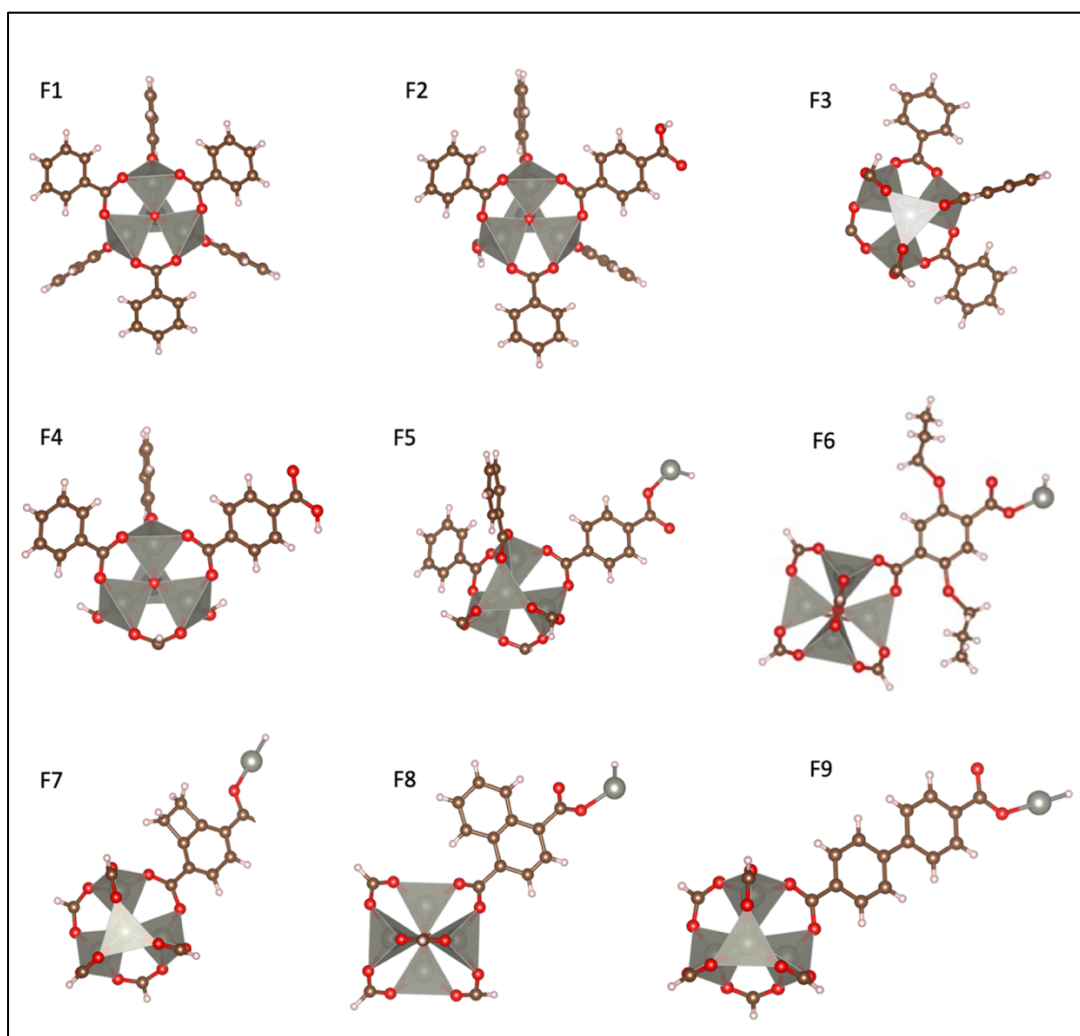


Figure 4.2: IRMOF-n ($n=1,4,6,7,10$) fragments. F1-F5 are generated from IRMOF-1. F6, F7, F8, F9 are from IRMOF-4, 6, 7 and 10, respectively. The atoms in the structure were represented by the following colors: white for H, brown for C, red for O, and grey for Zn.

4.2.1.2. DFT Calculations on the Dataset(s)

Apart from generating and developing the datasets, calculating reference values (i.e., energy and forces) for non-equilibrium geometries (of all datasets) as accurately as possible is another crucial factor in training of neural network potentials (NNPs). As discussed in chapter 2, wave function-based methods such as coupled cluster (CCSD(T), etc.) and perturbation methods (MP2, MP4, etc.) would be the most accurate definition of the systems. However, these methods are still computationally too expensive even for fragments with tens of atoms due to the number of fragment structures (thousands) at the dataset. Instead, density functional theory (DFT)

calculations with much lower computational cost can be utilized. The functional and basis sets employed in DFT calculations should be carefully chosen in order to represent the coordination compounds with transition metal(s) and organic substructures in the unit cell of MOF structures as well as to represent three-dimensional extension of periodic MOF structures [138]. The Perdew-Burke-Ernzerhof parametrization (PBE) has been widely used as a better balance between computational cost and accuracy for such large systems. In particular, this functional has been shown to reflect the most sensitive small changes in equilibrium/non-equilibrium geometry which is crucial for training PES of MOFs [139]. On the other hand, PBE does not account for long range dispersion interactions. The Grimme's functions can be utilized as a dispersion interaction correction to DFT energies [140]. The DFT calculations can also be fine-tuned with different choices for the size of the basis set. The TZVP basis sets are frequently used in computational chemistry to model molecules, providing a balance between computational efficiency and accuracy. Specifically, the def2-TZVP basis set is a popular starting point for constructing consistent and portable basis sets for soft material calculations, within the family of Gaussian basis sets [141, 142].

In this study, all atoms are defined at the PBE/def2-TZVP level in the DFT calculations throughout this dissertation (i.e., we refer this method by DFT). The ORCA software package (version 4.2) was used to perform these calculations. Self-consistent field (SCF) calculations were accelerated with the RIJDX option available in the software which notably increases computational efficiency [143]. Tight SCF criteria were applied for all calculations and interactions were corrected by the use of Grimme's density dependent atom-pairwise dispersion (D4) functions [144]. A custom-built python code was prepared using the Atomic Simulation Environment (ASE) [145] framework to perform calculations and store all geometries and their DFT results in an ASE database format.

4.2.1.3. On-the-fly Data Sampling During MD Simulation and Construction of NNPs

Since the configurational space is not adequately represented with initial training dataset, missing structures from well-represented regions of configurational space must be added in order to develop the NNPs. Thus, wisely chosen new geometries

need to be added to the training set as reference structures. This choice was made by applying two different criteria to the structures from the trajectories sampled from MD simulations. First, if a structure is not inside the training set, the prediction would be purely based on extrapolation and not reliable. Thus, the first criterion is that the structure must be included if the atom centered symmetry function (ACSF) values are outside the training set range. This criterion was carried out by adopting the idea of symmetry function values reported by Behler [146, 147] although we used SchNet rather than HDNN. This is because the symmetry function vectors are translationally and rotationally invariant and it is much easier to compare different structures whether they are inside or outside the training set. Second criterion is that the energy produced by an NNP of different functional form (i.e. NNP2), which is trained by the same dataset, should yield similar result. This is because NNP1 and NNP2 differ only by the number of layers and trained on the same exact dataset until yielding the same loss RMSE [126, 148]. Therefore, if the energy difference for two independent NNPs ($\text{Uncertainty} = \Delta E_{\text{NNP1-NNP2}}$) is above a certain threshold value, that structure must be included in the training set.

As an iterative active learning process (Figure 4.3), starting with primitive neural network potentials which trained initial dataset, are differed by only in functional form, using MD simulations of fragments, new geometry configurations were added to the training datasets to better represent configurational space. This procedure is repeated until the sufficient quality of NNPs were produced (convergence). The simulations of fragments used to train NNP were carried out using SchNetPack software at NVT ensemble with the standard velocity Verlet integrator [136]. The time constant for the simulations was set to 0.5 fs, and Langevin thermostat was used.

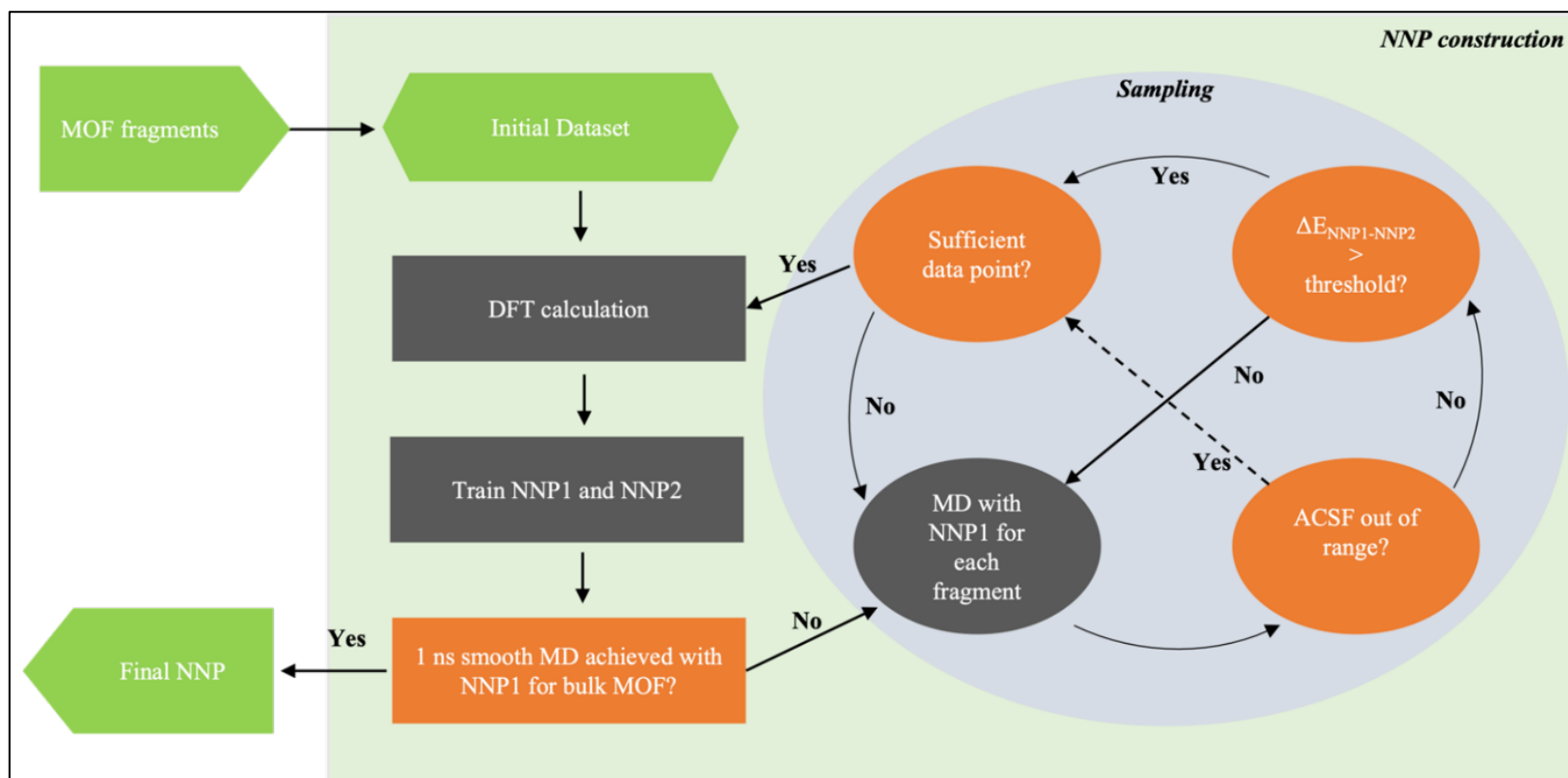


Figure 4.3: Construction of NNPs from MOF fragments

Herein, we construct the primitive NNP1 and NPP2 models were trained with this initial data set using DFT energies and forces based on fragments of the MOFs utilizing SchNet. Hyperparameter optimization was performed using grid search algorithm by Ray Tune software [149]. Optimized hyperparameters are as follows: 96-dimensional feature space, a radial cutoff of 6.0 Å, 64 filters, and T=3 interaction refinements have been used for our metal organic hybrid molecular systems. The prediction block used summation as the aggregation mode and 3 neural network layers. The model was trained with a descending learning rate of 10^{-4} to 10^{-6} , a decay factor of 0.8 and patience of 8 using ADAM optimizer. All other hyperparameters are the default values in SchnetPack [136]. Among the dataset, 80% was used for training while 10% was used for the validation and the remaining 10% was reserved for testing. All training processes have been performed using a Tesla P100 GPU.

Starting with 0.5 eV, this threshold was gradually lowered to 0.05 eV, as the data set develops. The error will be larger in the NNP trained in the early data set, and this error rate decreases as the data set develops. Uncertainty improvement (i.e. RMSE/MAE) between NNP1 and NNP2 by addition of new configurations to the training set shows an exponential decay function (Figure 4.4). When the new structures based on these criteria have reached to a sufficient number (~5000 data points), these new structures were calculated by DFT and added to the reference set, which eventually created the next generation of NNPs. These processes were repeated several times until smooth NVT MD simulations for 1 ns could be run on the periodic bulk MOF structures.

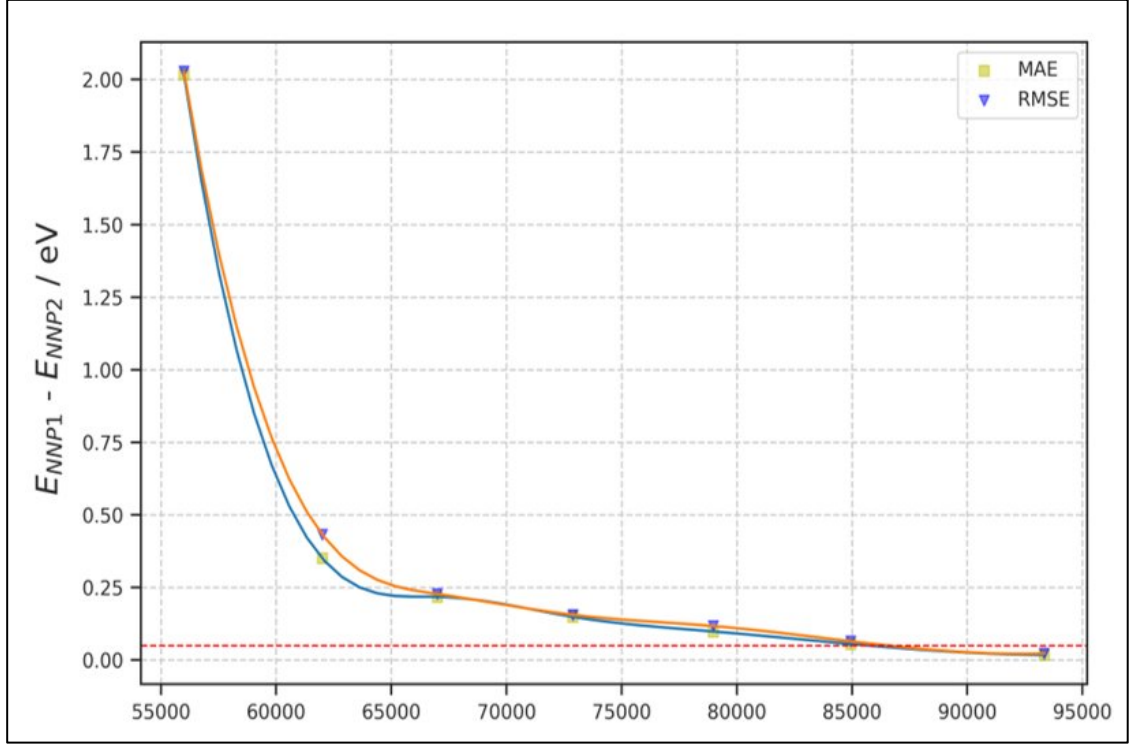


Figure 4.4: Uncertainty between NNP1 and NNP2 energies according to amount of data.

Finally, $\sim 170\text{k}$ reference data points were obtained when the NNPs were converged. The data were split into training ($\sim 126\text{k}$, 80%) and validation ($\sim 17\text{k}$, 10%) sets while the remainder were reserved for tests ($\sim 17\text{k}$, 10%). The RMSE values on energies and forces were selected as metrics for the losses in training and validation. The loss consisted of weighted average RMSE of 99% forces and 1% energies ($\text{RMSE}_{\text{Loss}} = \text{RMSE}_{\text{energy}} \times 0.01 + \text{RMSE}_{\text{force}} \times 0.99$). Early stopping was carried out when the loss of the training and validation curves reached the plateau at 150 epochs after which the energy only changes by 0.001 eV within 107 epochs (~ 152 hours) (Figure 4.5)

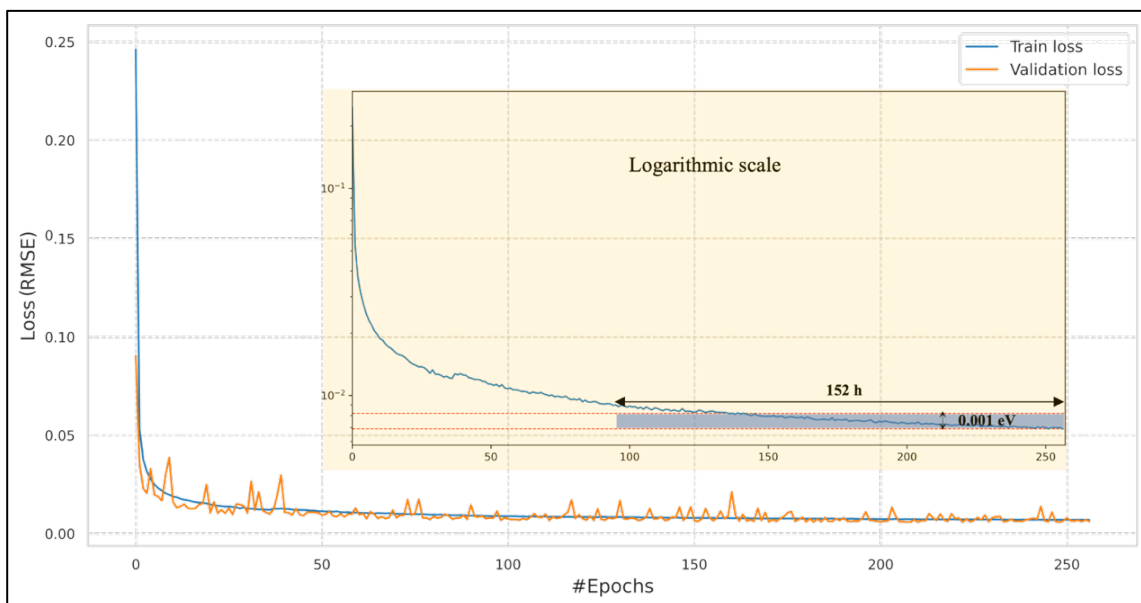


Figure 4.5: Training and validation loss curves

4.2.2. Validation and Test on Fragments

Following the construction of the NNP, the next step involved testing the energies and forces on the individual fragments. This crucial process allowed for the evaluation and optimization of the model's performance and accuracy. By subjecting the fragments to various energy and force tests, it was possible to determine whether the NNP was capable of accurately predicting the behavior and interactions of the constituent atoms and molecules. This step is one of the important tests to ensure that NNP can be used safely in subsequent simulations. Here in, the final version of the NNP was utilized to calculate the energy and force values of the fragments in both the training and test data sets. These values were then compared against the reference DFT values in order to assess the accuracy of the NNP. The maximum errors for the energies are 0.036 eV/atom and 0.026 eV/atom for the training and test sets, respectively.

Figure 4.6 shows the plot of absolute errors between the NNP and DFT energies with respect to DFT energies for the training set and the test sets. The errors on most fragment structures have been estimated to lie within the narrow range of 0.005 eV/atom for both training and test sets. The vertical bar shows the counts of structures for the corresponding error range. Only 1.91% in the training set and 1.58% in the test set is estimated with error greater than 0.005 eV/atom. As the structures approach to the equilibrium geometries, the errors shrink whereas large errors are concentrated in geometries that are too far from the equilibrium, higher-lying energies on PES. The

distribution of the errors is random and fits to a normal Gaussian function (Figure C.1). Reference DFT energies are in a perfect agreement with the those produced by NNP for all 9 fragments ($R^2 \approx 1.000$ for the training set, Figure C.2 and $R^2 \approx 0.999-1.000$ for the test set, Figure C.3). Based on these energy data, the RMSE of the training and test sets is calculated to be 0.0018 and 0.0017 eV/atom, respectively.

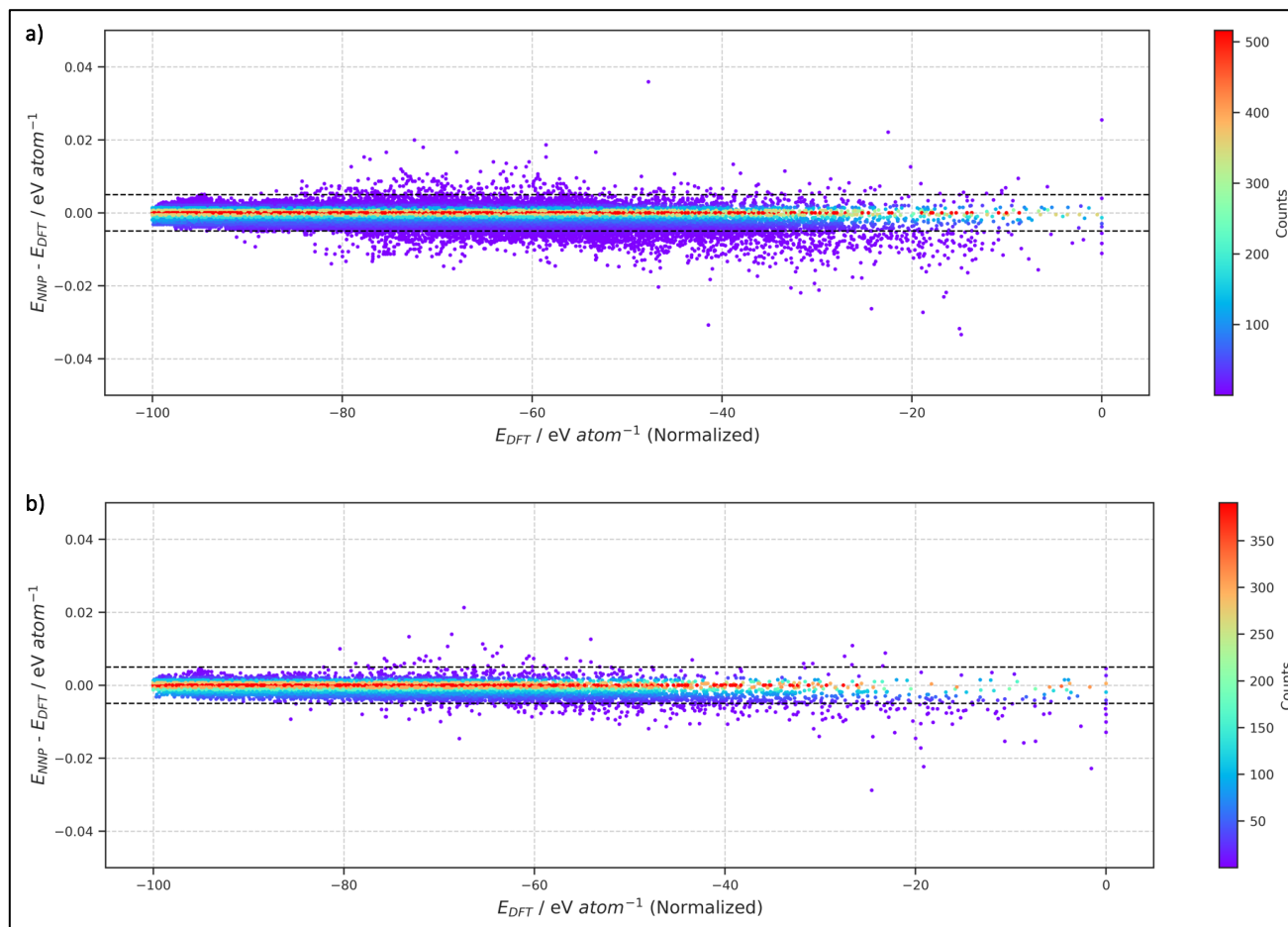
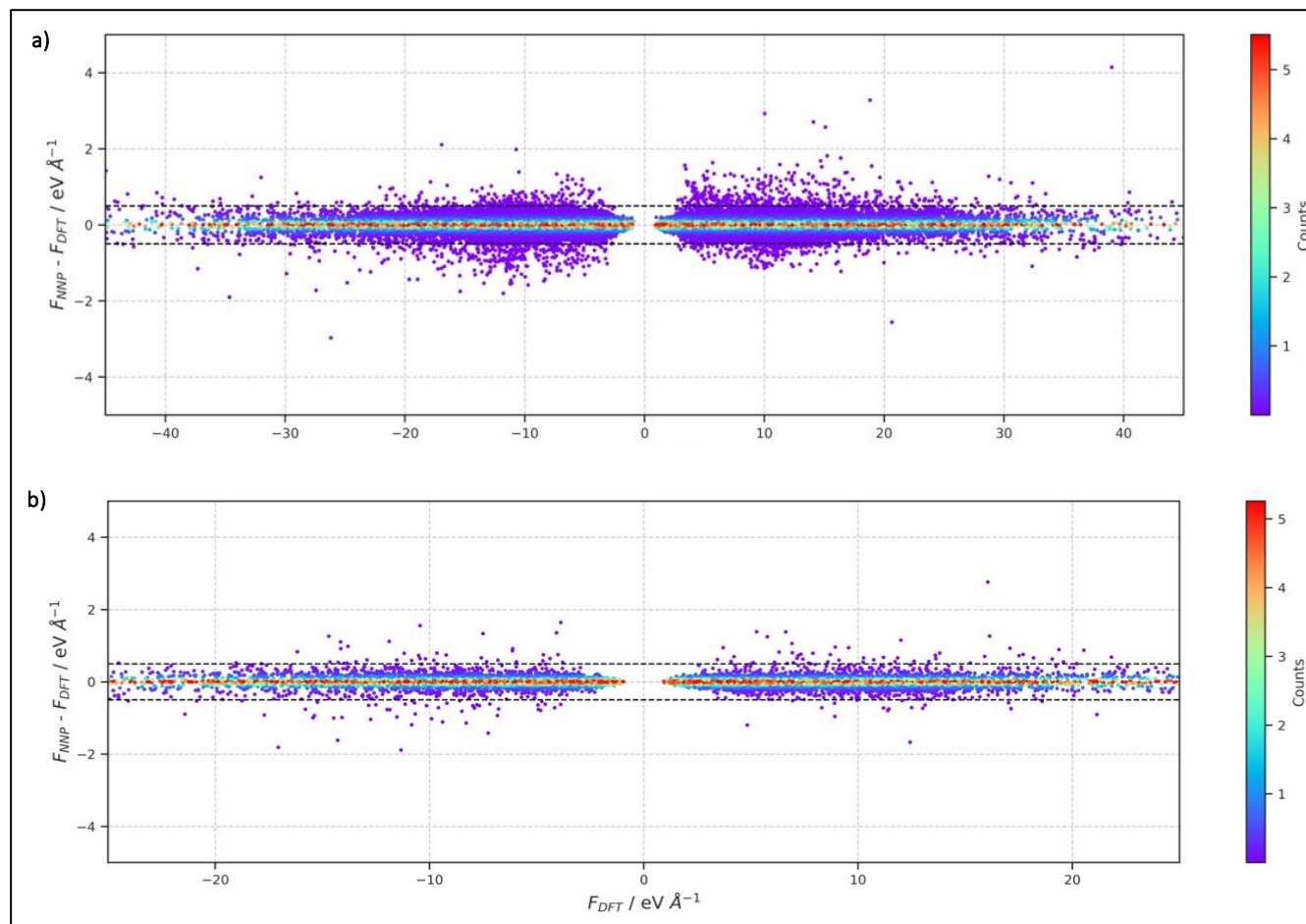


Figure 4.6: Errors in energies of fragments for a) training and b) test sets. Color bar shows the number of structures with the corresponding errors. Black dashed lines mark errors of 0.005 eV/atom. DFT energies (x axis) are normalized.

In addition to the energies, the maximum NNP force components (F_{max}) have been tested by means of deviations from the DFT values. Figure 4.7 also shows the absolute errors between the F_{max} of NNP and DFT with respect to DFT forces for the training and test sets. The RMSE of these deviations is 0.14 eV/Å and 0.15 eV/Å for train and test datasets, respectively. Among the complete data set, only 1.00% of the training and 1.34% of the test sets have errors that are greater than 0.5 eV/Å. (For the distribution of the errors, refer to Figure C.4). Reference DFT forces are in a good agreement with the those produced by NNP for all 9 fragments, $R^2 \approx 1.000$ for the training set, Figure C.5 and $R^2 \approx 0.999-1.000$ for the test set, Figure C.6.).



97 Figure 4.7: Errors in forces of fragments for a) training and b) test sets. Color bar shows the number of structures with the corresponding errors. Black dashed lines mark errors of 0.5 eV/Å.

4.3. Performance of NNP on Bulk Properties

The accuracy of the NNP in modeling bulk structures of MOF from their constituent fragments is of great significance, as it enables the investigation of the properties of IRMOF structures with a high degree of precision. Thus, assessing the performance of the NNP, developed using MOF fragments, on periodic MOF structures is a critical step in ensuring that the model can accurately predict the behavior of bulk structures and facilitate their exploration in various applications.

Apart from testing on fragments, 1 ns NVT MD simulations at 300 K on each periodic bulk IRMOF-1,4,6,7,10 structures were also run following by energy minimization in order to test the stability of NNP. All the configurations visited by the MD simulations were in the well represented regions of NNP. No structural changes apart from subtle thermal fluctuations were observed during the simulations.

4.3.1. Lattice Constants

The equilibrium geometries of fragments in bulk MOFs were not included in the development of NNP in order to avoid memorization of those values. Thus, reproducing the equilibrium lattice constants from the final NNP is one of the most critical test parameters for the success of NNP. Table 4.1 shows the comparison of equilibrium lattice constants produced by NNP of bulk IRMOF structures with experimental single crystal x-ray diffraction (SCXRD) values at 100 K [150] along with literature calculations. Our developed NNP was utilized for carrying out NPT MD simulations at 100 K and 300 K of 500 ps to produce values of equilibrium lattice constants and bulk modulus. The NNP calculated equilibrium lattice constants are in an excellent agreement with experiments and literature calculations, indicative of the reliability of the NNP. The difference between the values of NNP calculated and SCXRD ranges only from 0.04 Å to 0.64 Å for all MOFs trained. It should be stressed that NNP, which has been trained from non-equilibrium geometries of fragments, could successfully predicts the periodic equilibrium parameters.

Table 4.1: Equilibrium lattice constants and bulk modulus of IRMOF series calculated using NNP and compared with experimental SCXRD and other calculations in literature.

IRMOF-n	Equilibrium Lattice Constant (Å)				Bulk Modulus (GPa)		
	SCXR (100 K)	NNP (100 K)	NNP (300 K)	Simulations	NNP (100 K)	NNP (300 K)	Simulations
1	25.89	26.03	25.95	26.05a (0K); 26.08 ^a (0K); 26.88b (0K); 25.26 ^c (100K); 26.09d (100K); 25.21 ^e (300K); 26.08 ^d (300K)	13.17	8.97	16.0a (0K); 19.0 ^a (0K); 11.95f (0K); 15.76 ^d (0K); 8.70b (0K); 19.37 ^c (100K); 16.66c (300K); 13.3 ^d (300K); 12.3 (300K); 14.34 ^e (300K);
4	25.85	25.81	25.82	-	12.55	8.26	-
6	25.84	26.18	26.12	27.03 ^b (0K)	14.59	11.80	12.21b (0K); 12.12 ^c (300K)
7	25.83	25.19	25.20	26.69 ^b (0K)	6.58	2.30	3.70 ^b (0K)
10	34.28	34.73	34.62	35.35b (0K); 34.82 ^d (100K); 34.48d (300K); 34.37 ^e (300K)	6.95	5.23	9.2d (0K); 8.25 ^f (0K); 6.00b (0K); 3.5 ^d (300K); 7.41e (300K); 5 ^g (300K)
a: [124]; b:[151]; c:[152]; d:[153]; e:[154]; f:[155]; g:[119]							

4.3.2. Energy Profiles on the Organic Linker's Rotation in the Bulk IRMOF-1

Orientation of organic linkers is a key feature for the structural stability of IRMOF series and accurate calculation of torsional energies could be an important indicator for the success of the NNP [124]. The phenylene rings are common in all MOFs studied here. Therefore, as a case study, we compared NNP and DFT produced rotational energy barrier of a phenylene ring in IRMOF-1 to assess the reliability of the NNP. In order to reduce cost of the DFT calculation, we truncated the IRMOF-1 structure and saturated it with hydrogens (Figure C.7). NNP and DFT energy profiles were produced by scanning the dihedral angle of one of the phenylene rings. Figure 4.8-a shows the energy profile along a 180° rotation. The DFT and NNP calculated values of 0.58 and 0.65 eV, respectively, diverges only by 0.07 eV. This difference might arise from a single NNP representation of all studied MOFs rather than developing an NNP for each of MOFs. In addition, both methods somewhat overestimate the experimental energy barrier obtained by ²H-NMR studies (0.490±0.087 eV) [156]. The overestimation could be due to the use of truncated structure rather than the periodic one.

Similar to phenylene ring torsions, the same truncated structure was also used to reproduce the aromatic C-H bond distance energy profile of the phenylene ring. Figure 4.8-b shows the relative energies corresponding the C-H distance scan calculated by DFT and NNP. Although NNP was developed using non-equilibrium geometries, it can perfectly predict the DFT calculated equilibrium bond distance of C-H to be 1.09 Å along the scan coordinate. DFT calculated distance was totally reproduced by NNP. Both methods agree reasonably with the experimental equilibrium aromatic C-H distance of 1.10846 Å by neutron powder diffraction (NPD) measurements [157].

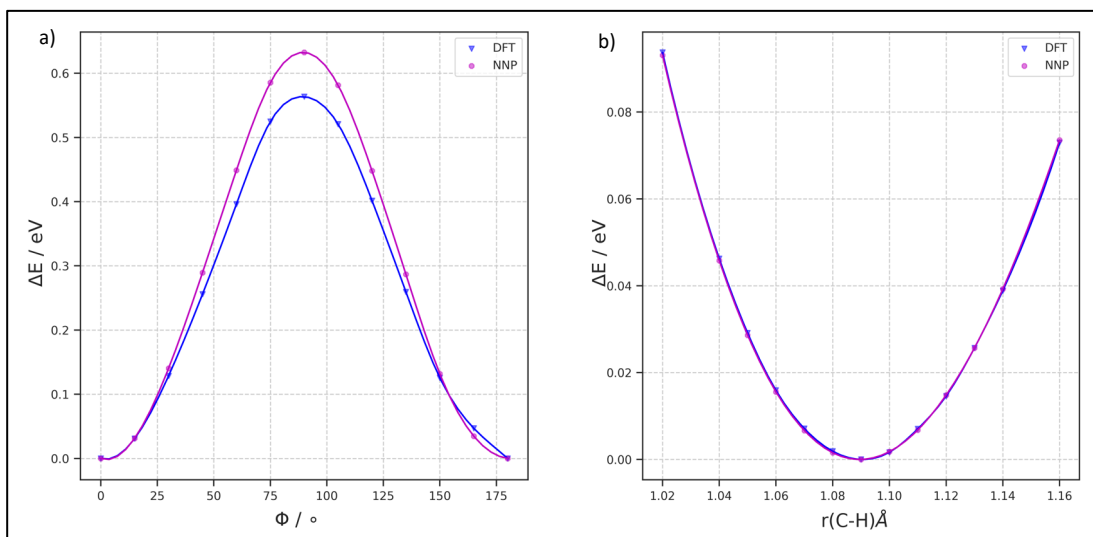


Figure 4.8: Potential energy surface in IRMOF-1 by scanning a): dihedral angle between carboxylate and phenylene planes from 0 to 180° and b) distance of aromatic C-H from 1.02 Å to 1.16 Å. Dashed lines shows experimental values as NMR for torsional barrier and NPD for equilibrium distance.

4.3.3. Phonon Modes

Reproducing phonon modes of a periodic bulk structure from fragment trained NNP is another critical test for the reliability of the potential. Calculating phonon modes directly from first principles can become quite costly for such large MOF systems. However, using the NNP, we utilized the calculation of vibrational modes available in ASE, using finite difference method, with much lower computational cost.

Experimental studies on IRMOF-1 vibrations showed that there are several spectral regions [158]. The dominant vibrations are C-H stretching (due to aromatic C-H on phenylene ring) at 2900-3100 cm^{-1} ; C=O stretching (due to carboxylate linker groups), C=C stretching (due to aromatic C-C on phenylene ring) at 1300-1600 cm^{-1} ; in-plane and out-plane deformation of aromatic ring (due to the phenylene ring) at 600-1200 cm^{-1} ; the Zn involved bands (due to inorganic subunit Zn-O stretches and bends) at 200-600 cm^{-1} ; and collective vibrations and lattice modes below 200 cm^{-1} [158]. The latter (also reported as below 23 meV) are typical for most MOFs and play important role in negative thermal expansion properties.

IRMOF-1 shows aromatic C-H symmetric and antisymmetric stretching bands at the 3126.4 cm^{-1} (23) and 3138.9 cm^{-1} (46), respectively (Figure 4.9 and Figure 4.10). These values are very consistent with previous DFT studies. Phonon modes of simplified model system of IRMOF-1 were reported in literature using two different

functionals, B3LYP [159] and PBEsol-D3 [158] as 3230/3243 cm^{-1} and 3110/3122 cm^{-1} . Since our NNP was trained with PBE-D4, our values are in excellent agreement with PBEsol-D3 values. Not only the frequency values, but the intensities and splitting of 12 cm^{-1} between the symmetric and antisymmetric stretches were also totally reproduced by our NNP. The C-H stretches were also reported by HDNNP [124], which was also trained by PBE-D3, as 3104/3148 cm^{-1} . Our NNP calculated C-H stretching frequencies are also in agreement with previously reported experimental values.

IRMOF-10 shows very similar C-H stretching frequencies to IRMOF-1 since all C-H bonds are aromatic (Figure S8). However, due to unequal chemical environments, C-H bonds at the interface of the rings slightly differ from those that are towards carboxylate groups. The C-H stretches in IRMOF-4 are more complicated due to the existence of aliphatic C-H bonds of n-propoxy groups linked in BDC. Similarly, additional aliphatic C-H stretches are observed in IRMOF-6 due to cyclobutyl group linked in BDC. In IRMOF-7, the C-H stretches of fused benzene ring are observed at slightly different frequencies than those on BDC. This difference can be rationalized by the fact that H atoms on the fused benzene are closer to the carboxylate oxygen atoms whereas the H atoms on BDC are more distant from carboxylate groups. Indeed, two isomeric forms as “twisted” and “planar” can conceivably exist in IRMOF-1 and IRMOF-7 and the stability of one of these forms were dominated by the steric repulsion of the fused benzene ring [160].

The second region of IRMOF-1 consists of the spectrum lying from 1514 cm^{-1} to 1625 cm^{-1} with several peaks including the one at 1612 cm^{-1} and 1595 cm^{-1} , corresponding to symmetric and antisymmetric stretches of aromatic carbon double bonds and carboxylate groups. This is in complete agreement with previously calculated DFT values of 1649 cm^{-1} and 1549 cm^{-1} and experimental IR peaks at 1612 cm^{-1} [150] and Raman peak at 1600 cm^{-1} . In IRMOF-7, these bands are calculated to be red shifted. For IRMOF-7, fused benzene rings’ hydrogens that are proximal to the carboxylate oxygen atoms show more intense C-H bending vibrations.

Experimentally observed Zn-O (the carboxylate oxygen) stretch in IRMOF-1 at 578 cm^{-1} were also reproduced by the peaks in the range 582-612 cm^{-1} . The same mode was reported as 579-606 cm^{-1} in DFT calculations previously. In addition, the Zn-O (inorganic oxygen coordinated by four Zn atoms) is calculated to be 518-523 cm^{-1} . Experimentally, this peak is assigned to be 524 cm^{-1} while the DFT calculations

reported the value of 512 cm^{-1} . Finally, Zn involved collective motions were reported to be below 287 cm^{-1} . Our calculations show 293 cm^{-1} and below are the corresponding motions including Zn-Zn vibrations. Since the additional fragments on the BDC linker increase the degree of freedom, more accessible vibrational states exist in lower lying spectral regions. Overall, our calculated vibrational spectra illustrate a good agreement with those reported experimentally and those predicted by DFT density of state calculations.

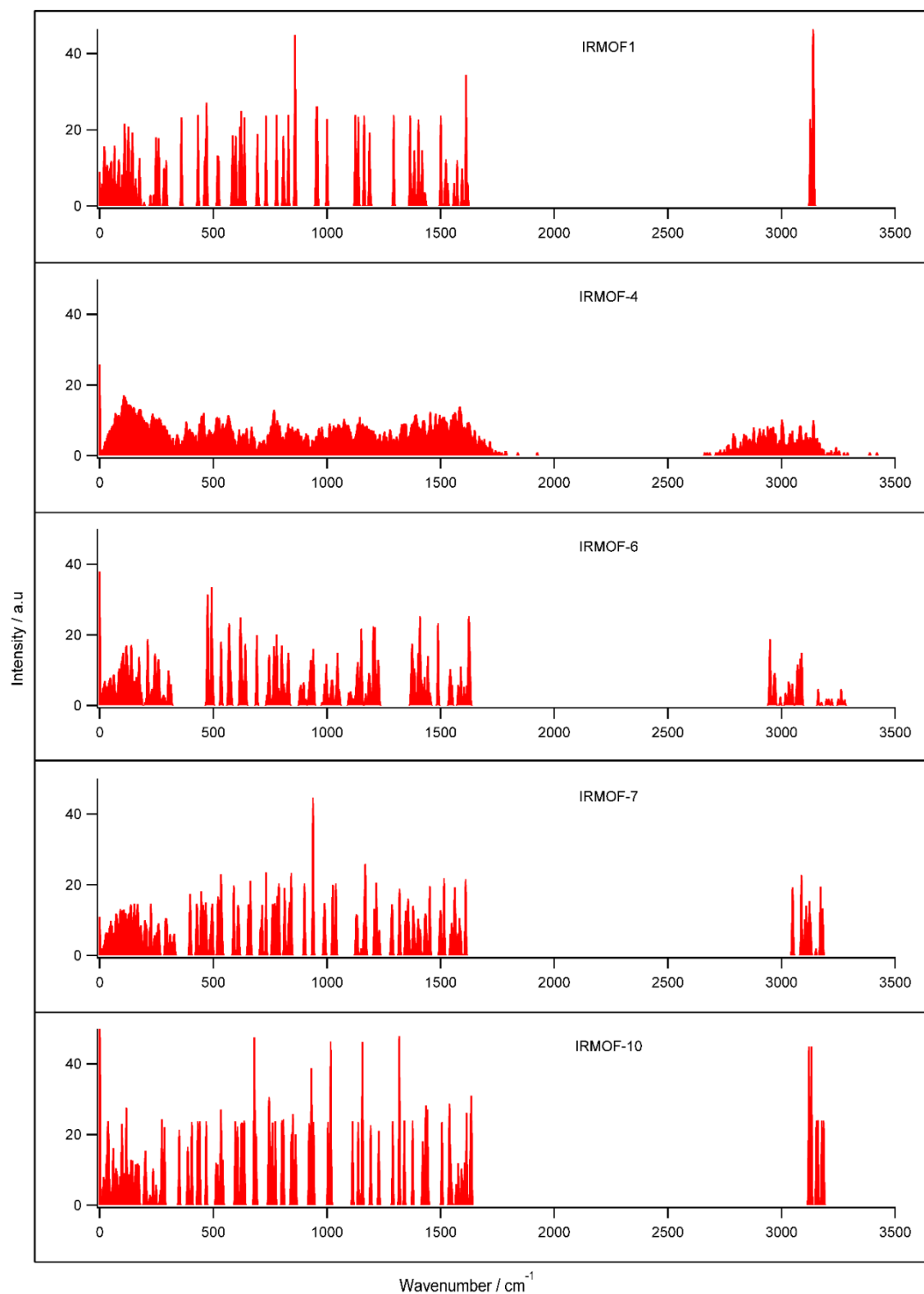


Figure 4.9: Calculated vibrational spectra of IRMOF-n series by NNP. For zoomed sections refer to Figure 4.10.

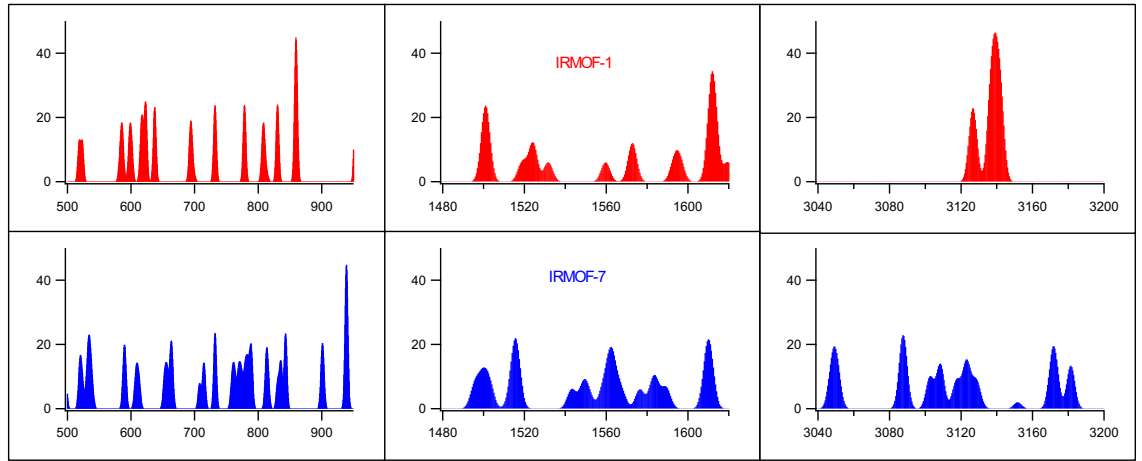


Figure 4.10: IRMOF-1 and IRMOF-7 vibrations zoomed in different regions.

4.3.4. Thermal Behavior

In terms of thermal behaviors of pure MOF crystals, we have calculated thermal expansion coefficient, thermal pressure coefficient and bulk modulus. Using our NNP for energy and force calculations, several MD simulations with temperature ranging from 200 K to 400 K were performed at the NPT ensemble.

The NNP based simulations were carried out using SchNetPack software with the standard velocity Verlet integrator [136]. The time constant for the simulations was set to 0.5 fs, and Langevin thermostat was used. The final simulations of periodic MOF structures were performed to investigate bulk properties for phonon modes and thermal mismatch using SchNetPack and ASE interchangeably. For thermal studies, MD simulations at different temperatures were performed ranging from 200 K to 400 K by 50 K increments. For each temperature value, the energy of the periodic bulk system was minimized. After minimization, MD simulations of 500 ps with 0.5 fs time steps (total of 1 million steps) were carried out using NPT ensemble at 1 atm reference pressure. Langevin thermostat along with Nosé-Hoover barostats were used. Damping constants for both thermostats were set to 0.1 ps. Last 300 ps of the simulations was used for the analysis.

Thermal expansion coefficient (α_0), bulk modulus (β_0) and thermal pressure coefficients (γ) are the main parameters that can be calculated for the thermal behavior [119].

Thermal expansion coefficient (α_0) is defined as:

$$\alpha_0 = \frac{1}{V} \left(\frac{\partial V}{\partial T} \right)_P \quad (4.3)$$

This can be re-written by integrating for volume and temperature;

$$\alpha_0 dT = \frac{1}{V} dV \quad (4.4)$$

$$\int_0^T \alpha_0 dT = \left[\int_{V_0}^V \frac{1}{V} dV \right]_P \quad (4.5)$$

$$\langle \ln(V) \rangle_P \approx \alpha_0 T + V_0 \quad (4.6)$$

where V_0 is the equilibrium volume at 0 K. By running several MD simulations with different temperatures, α_0 can be calculated from the linear correlation between temperature and natural logarithm of volume expansion. The term $\langle \ln(V) \rangle_P$ can be obtained from the ensemble average of an MD simulation at the NPT ensemble.

Thermal pressure is given by:

$$\Delta P = \gamma \cdot \Delta T \quad (4.7)$$

where γ is the thermal pressure coefficient with:

$$\gamma = \alpha_0 \cdot \beta_0 \quad (4.8)$$

where β_0 is the bulk modulus and defined as:

$$\beta_0 = k_B \langle T \rangle \frac{\langle V \rangle}{\langle V^2 \rangle - \langle V \rangle^2} \quad (4.9)$$

where k_B is the Boltzmann coefficient. The volumetric term in the equation corresponds to the relative variance, the fluctuation of the volume throughout the MD simulation at $T=300$ K.

The mechanical pressure was set to 1 atm, and the data of MD simulations with five different temperature values from 200 to 400 K in steps of 50 K were fitted the Equation (4.7) to find thermal expansion coefficient values. Similarly, thermal pressure coefficients and coefficient factors have been calculated from Equations (4.8) and (4.9). Using NNP, we find the thermal expansion coefficient values of IRMOF-1 and IRMOF-10 to be 47 MK^{-1} and 54 MK^{-1} , respectively. In literature, there is no consensus about these values. Experiments suggest two values to be -48 [161] and -39 MK^{-1} [162] while predicted values from simulations in literature range from -9 to -79 MK^{-1} [119, 152, 154, 155, 163–166]. Our result for IRMOF-1 is very close to the one of the two experimentally reported data of -48 and -39 MK^{-1} . This might suggest the superiority of NNP over the classical simulations. Theoretical calculations on IRMOF-10 have also suggested different values ranging between -16 and -55 MK^{-1} [154, 155, 163, 167], which is consistent with our prediction. To the best of our knowledge, the values of IRMOF-4, IRMOF-6 and IRMOF-7 have not been reported in literature. We predicted the value of IRMOF-6 as -35 MK^{-1} . However, we could not get a linear correlation for IRMOF-4 and IRMOF-7.

The origin of negative thermal expansion properties of MOFs is attributed to a few different scenarios. Shortening the bond lengths and phase changes, electronic effects, magnetostriction and vibrational modes of bridging atoms. Bridging atoms can lead to high NTE with more than one contribution. First, the transverse vibrational modes on the bridging phenylene rings cause contraction of the Zn_4O nodes. As temperature increases, the contraction gets larger, resulting with volume shrinking in MOF [124]. Second, the flexible carboxylate groups' vibrational modes lead shortening of the distance between the carboxylate groups and rigid O atoms of Zn_4O [161].

The torsional motion of the phenylene ring may also contribute higher NTE. In particular, IRMOF-7 has fused benzene and IRMOF-4 has n-proxy linked in BDE. Independent torsion of BDC linker proximate these functional groups in IRMOF-4 and IRMOF-7, resulting with greater energy barrier when compared to IRMOF-1, IRMOF-6, and IRMOF-10. This might explain unusual thermal behavior of the former two MOFs. At low temperatures, due to the greater torsional barrier, IRMOF-4 and

IRMOF-7 first expand with temperature, i.e. positive thermal expansion (PTE). After reaching to a thermal energy which can compensate this barrier, the MOFs start contracting with increasing temperature, i.e., negative thermal expansion (NTE). In order to justify this hypothesis, we have scanned the rotation of two of the phenylene rings that are facing to each other in IRMOF-1 and IRMOF-7.

IRMOF-1 and IRMOF-7 structures are similar with some distinct differences. In IRMOF-1 structure, carboxylate groups form two planes that are coplanar to each other (0°). Although these carboxylate planes can be twisted (90°) in principle, they cannot be interconverted without breaking bonds of the network. In addition, the phenylene ring is also coplanar (0°) with these carboxylate groups due to the conjugation between the aromatic system and the carboxylate groups. In addition, the attraction between the proximal H and carboxylate O atoms is dominant in the coplanar carboxylate and phenylene groups since the minimum distance of 2.4 Å is reached. [160]. On the other hand, IRMOF-7 has fused benzene ring attached (ring B) on the phenylene ring (ring A) of the BDC linker (Figure 4.11). The two carboxylate groups were experimentally observed in the twisted (90°) form in IRMOF-7 [150]. For the naphthalene being coplanar with either one of the carboxylate planes would cause steric clashes between the proximal H atoms on the fused ring (ring B) with the O atoms of carboxylate groups. That is, the distance between proximal H atom on ring B and carboxylate O atom can be as low as 1.7 Å, in which repulsions dominates. Thus, the equilibrium geometry of IRMOF-7 has the angle between naphthalene plane and carboxylate groups of 45° in which the distance between proximal H atom of ring B and carboxylate O atom is ~ 2.3 Å, very close to the distance between H atom of the phenylene ring (ring A) and carboxylate O atom in the IRMOF-1 with 2.4 Å. In addition, similar to IRMOF-1, H atom of ring A in IRMOF-7 has a distance of 2.4 Å with carboxylate O atom, which also contributes the stabilization. Steric hindrance is not limited only to the distance between the proximal H atom of ring B and carboxylate O atom but also exists between two neighboring rings. The distance between proximal H atoms between two neighboring B rings can be as low as 1.0 Å ($\Psi=135^\circ$) in the case of a torsion of the naphthalene group (Figure 4.11).

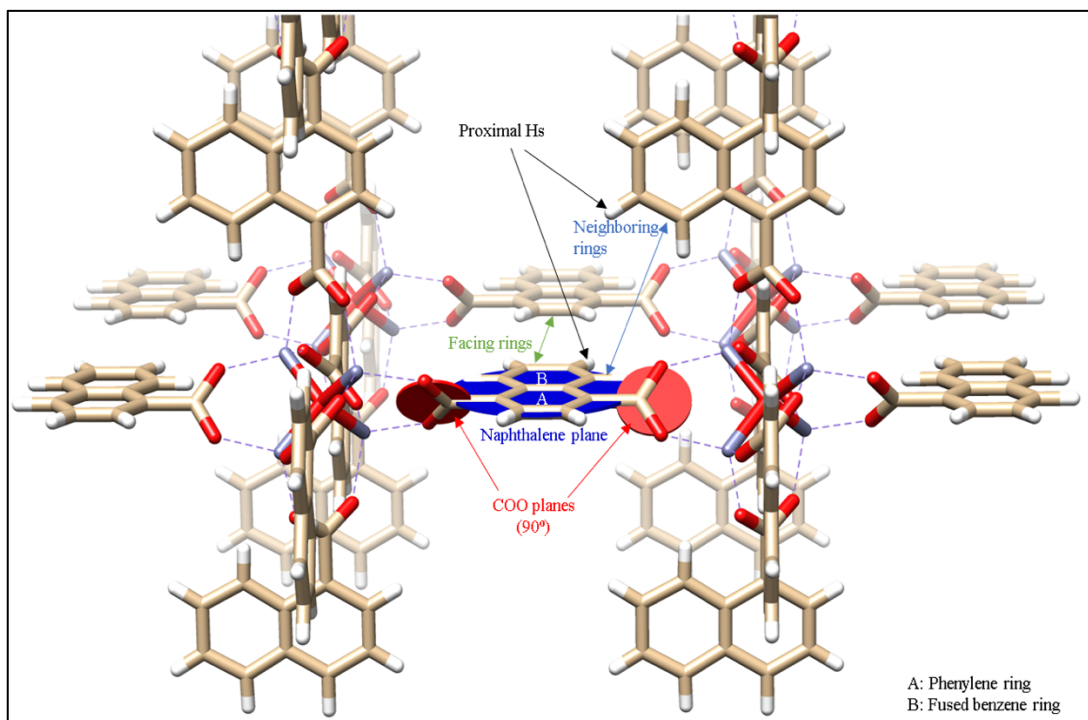


Figure 4.11: Representation of SCXRD structure of IRMOF-7 at 100 K. COO planes are orthogonal to each other and angled by 45° to the naphthalene plane to optimize interaction between H atoms of two neighboring B rings; and between H atoms of ring A/B and carboxylate O atoms.

In order to monitor the effect of steric hindrance from these two scenarios, a rigid scan has been carried out by changing the angle between the phenylene ring and carboxylate planes on a single unit cell of the MOF using NNP in a two-dimensional (2D) scan of two facing rings. Figure 4.12 shows the two-dimensional potential energy surface (2D-PES) of IRMOF-1 and IRMOF-7.

In the case of IRMOF-1, an independent rotation around two facing phenylene rings could bring atoms to a distance that is no less than $2.4 \text{ \AA}$ ($\Phi=\Psi=90^\circ$) with an energy barrier up to $\sim 1.2 \text{ eV}$. For a single dimensional scan, this barrier is almost halved to $\sim 0.6 \text{ eV}$ ($\Phi=0^\circ$, $\Psi=90^\circ$ or vice versa), which is in perfect agreement with experimental and DFT calculated values. On the other hand, due to the neighboring fused benzene rings, the atoms in IRMOF-7 can be as proximate as $1.0 \text{ \AA}$, which gives rise to the energy barrier as much as $\sim 6 \text{ eV}$ ($\Phi=\Psi=135^\circ$). For a single dimensional scan, this barrier is almost halved to $\sim 3 \text{ eV}$ ($\Phi=45^\circ$, $\Psi=135^\circ$ or vice versa). Additionally, another energy barrier (with double shoulders) of $\sim 2.5 \text{ eV}$ (translation to 1D is $\sim 0.6 \text{ eV}$) has been observed from the fused benzene H and carboxylate repulsions with $1.7 \text{ \AA}$ distance ($\Phi=45^\circ$, $\Psi=0^\circ$ and 90°).

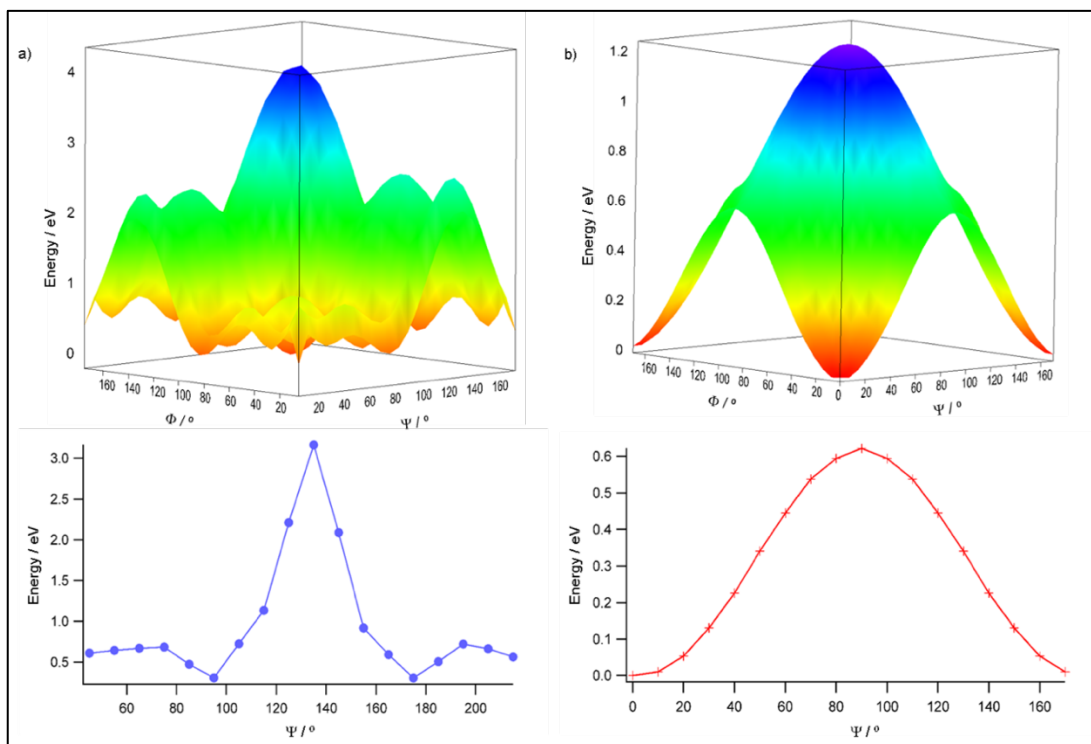


Figure 4.12: NNP-2D scan (with 10^6 steps) of rotations of two facing aromatic rings in a) IRMOF-1 and b) IRMOF-7. Φ and Ψ angles correspond to the angle between carboxylate and aromatic planes on each facing rings. The images were produced the data by fitting 12th order polynomial function. Bottom graphs show 1D scan (a portion in the top images) by fixing Φ angles to the initial angles ($\Phi = 0^\circ$ for IRMOF-1 and $\Psi = 45^\circ$ for IRMOF-7).

Switching from PTE to NTE were first reported in literature and this “switching temperature” were determined as 230 K for DUT-49 [120]. Although experimental observation of this phenomenon required solvated systems, underlying molecular basis can still be investigated for unsolvated systems by theoretical calculations. It should be noted that the thermal expansion coefficients of MOFs with dual thermal expansion characteristics cannot be simply produced by linear $\ln(\langle V \rangle)$ vs. T plot in Equation (4.8) since α_0 is not constant and thus should be reported as a function of temperature.

Table 4.2 lists thermal expansion coefficient values of IRMOF-n series calculated at different temperatures in comparison with other experimental and computational studies.

Table 4.2: Thermal expansion coefficient values of IRMOF-n series calculated at different temperatures. For the calculation at different temperatures, refer to (Table C.1, Table C.2, Figure C.8 and Figure C.9)

IRMOF-n	$\alpha_0 \times 10^{-6}$ (T=200 K)	$\alpha_0 \times 10^{-6}$ (T=300 K)	$\alpha_0 \times 10^{-6}$ (Constant T)	Experimental	Simulations
1	-51.392	-46.876	-46.900	-48[161]; -39 [168]	(-9) – (-79) [119, 154, 155, 163–166]
4	159.92	-91.500	-91.506		
6	-31.623	-34.566	-34.566		
7	433.75	-79.030	-79.032		
10	-40.195	-50.233	-50.233		(-16) – (-55) [154, 155, 163]

IRMOF-1,6 and 10 show only NTE behavior while IRMOF-4 and 7 shows PTE to NTE transition. Although the magnitude of the NTE coefficient of IRMOF-10 was not reported, it was shown to be larger than that of IRMOF-1 in literature both experimentally [164] and from simulations [119]. Our calculations agree well with experiments and computations in terms of the magnitude and the decrease of α_0 from IRMOF-10 to IRMOF-1.

After reaching to the NTE region, IRMOF-4 and IRMOF-7 contract more than IRMOF-1 with increasing temperature. These higher NTE values can be explained by greater repulsion of the phenylene rings due to greater energy barrier discussed earlier.

We also calculated thermal pressure coefficients from MD simulations using NNP (Table 4.3). Thermal pressure is an indicator for the rigidity and stability of MOFs. Thermal pressure coefficient, on the other hand, indicates how the pressure would change when a material in a closed system is subjected to temperature increase. Among the IRMOF series studied, IRMOF-10 shows the lower thermal pressure coefficient than IRMOF-1, which is qualitatively in agreement with simulations in literature [119].

Table 4.3: Thermal pressure coefficients calculated from MD simulations using NNP.

IRMOF-n	Thermal Pressure Coefficient, γ (MPaK ⁻¹)		
	NNP (100K)	NNP (300K)	Literature (300K)
1	-0.6768	-0.4205	-0.408 [169]
4	2.0070	-0.7558	-
6	-0.4614	-0.4079	-
7	2.8541	-0.1818	-
10	-0.2794	-0.2627	-0.233 [169]

4.4. Conclusions

In this Chapter, we have developed neural network potential to describe rotationally and translationally invariant energies and forces of IRMOFs at the DFT level accuracy using fragmentation technique. We have verified the success of our developed NNP on electronic structure properties of fragments and bulk MOFs. Results showed that our NNP can successfully predict equilibrium bulk properties of all IRMOF-n (n=1, 4, 6, 7 and 10) although only five fragments of IRMOF-1 and additional one fragment from each of the other MOFs were used in training. NNP allows us to study bulk properties of MOFs such as lattice constant, bulk modulus and thermal expansion at the DFT accuracy with greatly reduced computational cost. Thermal expansion studies showed that IRMOF-4 and IRMOF-7 have unusual thermal expansion behavior (as both PTE and NTE) whereas the other MOFs have only negative thermal expansion (NTE) in 200 K – 400 K temperature range.

5. THEORETICAL CH₄ AND H₂ ADSORPTION ON IRMOF-N ACCOUNTING FLEXIBILITY

5.1. Introduction

As noted in Chapter 1, MOFs have gathered a great deal of scientific research focus due to its variety of applications that range from gas storage and separation to drug delivery, catalysis and sensors. These and many more diverse applications are mainly because of the MOF's highly tunable physical and chemical properties arising from huge number of possible structures by the combination of a variety of available inorganic and organic building blocks. One of the aspects of MOFs is to utilize their flexibility, distinguishing them from rigid porous materials like zeolites [170], for performance improvements in many applications such as gas storage and separations. The flexibility, which can be extrinsic or intrinsic [171], causes reversible arrangements of organic and inorganic subunits such as expanding, contracting [119], gate-opening and breathing effects [172] under mechanical/thermal stress or host-guest interaction [173].

On the one hand, experimental characterization of materials must be carried out and will always be the vital in understanding the true physical and chemical properties of materials. On the other hand, quite large numbers of possible MOFs require too much experimental effort and cost to the characterization of such materials. In addition, the atomistic details of the host-guest interaction may not even be possible with existing experimental techniques. In this regard, this problem can be sorted out by the help of computational studies. Atomistic simulations, in particular, not only provide complementary information but also explain experiments.

Computational studies on gas adsorption properties of MOFs mostly rely on classical force fields. Classical force fields suffer from the coarse definition (MM based) of MOFs. In addition, the classical force fields use predefined parameters and these parameters do not change throughout the simulation. Indeed, there is an active research area for developing classical force fields to consider the flexibility of MOF systems [174]. The QuickFF [175, 176], MOF-FF [177, 178], BTW-FF [179] ZIF-FF [180] and UFF4MOF [178] are some examples of classical force fields with the former two designed for specific MOFs while the rest being generalized to all MOFs. All these methods still lack the electron involved phenomena such as bond formation or

breaking and the bonded potentials are limited to classical harmonic behavior. On the other hand, DFT methods with inclusion of periodic boundary conditions are much better to mimic bulk properties of MOFs since short-range local environments of atoms are defined more accurately. There have been numerous DFT studies on MOFs with different aspects such as electronic and vibrational properties [22, 181], catalytic behavior [182], and gas adsorption profiles [183, 184]. However, first-principles definition of MOFs, which include thousands of atoms in unit cells, is still a challenging task in terms of computational cost. Long simulation times are required since extensive sampling of the system is a necessity for accounting the dynamic definition of physical and chemical processes such as adsorption, desorption and diffusion of gas molecules, which altogether make *ab initio* simulations prohibitively expensive. A tradeoff between the accuracy and cost is to use machine learning (ML) potentials that are trained by DFT methods.

As mentioned in Chapter 1 and 2, for the last decade, machine learning (ML) has been emerged in a variety of disciplines to cover the gap between the most accurate *ab-initio* and fast classical simulations [68, 124, 125, 132, 185, 186]. MOFs can be described by construction of highly accurate and reactive ML potentials to represent multidimensional potential energy surfaces (PES) in order to perform much more accurate simulations with enormous computational speed-ups [126–132]. In chapter 4, we have reported construction and thermal application of a generic neural network potential (NNP) for the IRMOF-*n* series (*n* = 1, 4, 6, 7, 10) trained by PBE-D4/def2-TZVP reference data of MOF fragments. Our constructed NNP allow us to study much more different aspects of IRMOF series utilizing atomistic simulations such as grand canonical Monte Carlo (GCMC) and molecular dynamics (MD) simulations. In particular, these simulations provide more insights into more accurate gas uptakes as well as how MOFs behave during the adsorption process (i.e., accounting MOFs' flexibility and phase changes). The most recently, Johnson and co-workers developed an ML potential for UiO-66 and applied these potentials for modeling gas diffusion behaviors via molecular dynamics simulations [174].

5.2. Combined NNP/UFF to Account for Flexibility

As discussed in Chapter 1, IRMOFs are composed of Zn_4O coordination centers interconnected by dicarboxylate linkers, leading to the formation of three-dimensional

cubic networks. These materials exhibit various pore sizes while maintaining the same framework topology through the utilization of linkers with different lengths. It is known that the linker size has an impact on the flexibility of MOFs. Flexibility of MOFs have a direct effect on adsorption/desorption behavior of gases as well as their structural changes under different mechanical and thermal conditions (both extrinsic and/or intrinsic). Therefore, it is desired to the most accurately model MOFs in order to study such properties. Ideally, one can do this by performing ab-initio MC/MD simulations to see such effects. Unfortunately, with the current computational technology, it is far beyond the feasibility to model bulk MOF-gas systems. On the other hand, ML potentials (i.e., NNP) allow flexible description of the system at the ab-initio level, allowing us to sample sufficiently in an MD simulation.

It should be noted that none of the NNP developed accounts for the MOF-adsorbate interactions that allow direct use of NNP to study adsorption or diffusion properties of gases on MOFs. Although DFT calculations can in principle be performed in the presence of guest molecules in MOFs to train MOF-adsorbate and adsorbate-adsorbate interactions, this can be computationally too expensive as it requires a massive training set that includes different gases and their mixtures or a new NNP for each MOF-gas pair. Alternatively, adsorbate-adsorbate interactions, which consist of mostly van der Waals interactions, can be defined using classical force fields, which is known to perform very well in simple fluids and mixtures while non-bonded framework-adsorbate interactions (atom-atom or atom-united atom) are defined by combining rules [174, 187]. Thus, hybrid potentials as NNP for MOF and classical UFF for adsorbate could be incorporated in flexible gas adsorption simulations (Figure 5.1).

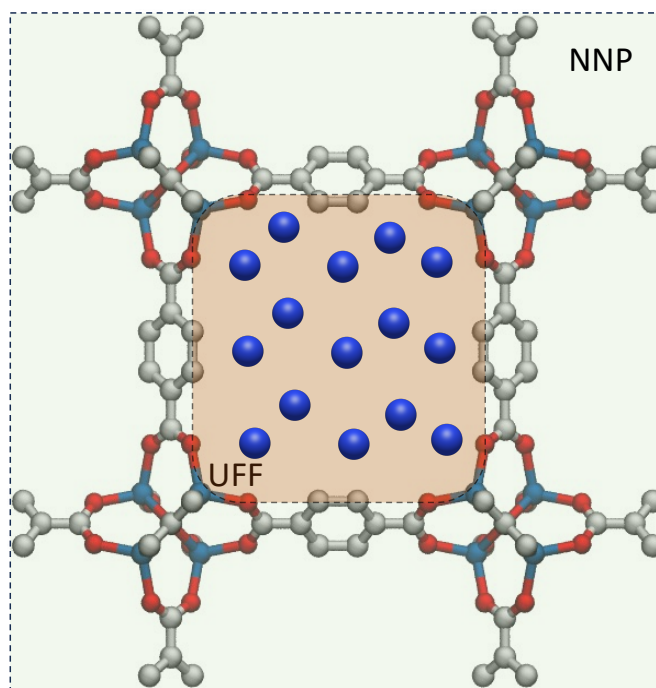


Figure 5.1: Illustration of the NNP/UFF Combined Force Fields Models

5.3. Combined MD and GCMC Simulations to Account for Flexibility

Current methods to study gas uptakes of porous materials rely on the grand canonical Monte Carlo (GCMC) simulations since they allow both the translational/rotational motions and insertion/deletion of the guest molecules. However, the flexibility of MOFs limits the GCMC simulations since the adsorbent is considered to stay rigid during the GCMC steps (“*rigid adsorption*”) [188]. In order to extensively sample the flexibility of MOFs, MD simulations could be performed instead. However, MD does not allow the number of adsorbed particles to change during MD steps [189–191]. Therefore, a single MD simulation cannot be sufficient in studying adsorption isotherms. As discussed in chapter 2, for modeling the dynamic behavior of MOFs so as to sample the flexibility of the host along with adsorption of guest molecules, MD simulations can be incorporated with GCMC simulations. Combination of MD and GCMC simulations not only reflects the extrinsic flexibility, triggered by external stimuli such as mechanical stress and gas adsorption which causes swelling, breathing, etc., but it also allows to observe intrinsic flexibility, due to thermal vibrations even in the absence of external stimuli during the gas adsorption on MOFs [171]. A few attempts have been made to effectively select the simulation

protocol in which GCMC and MD are combined [192, 193]. Using classical force fields, Ghofi and Maurin have studied phase transitions of MIL-53 as an extrinsic flexibility using combined MD/MC methods and experiment for CO₂ adsorption [194].

There are two specific approaches to combine MD simulations with GCMC simulations as “*flexible snapshot adsorption*” and “*adsorption-relaxation*”. The following subsections provide the details of these approaches and discusses the CH₄ and H₂ adsorption on IRMOF series using both approaches.

5.4. Flexible Snapshot Adsorption Simulations

One way of coupling MD and MC simulations is to run MD and MC simulations separately. In this approach, an MD simulation is performed for a MOF with the absence of guest molecules under some external conditions such as mechanical stress. After the MD simulations completed, snapshots are taken from the trajectory of the MD simulations and GCMC simulations are performed for gas adsorption. Since the atoms of MOFs are allowed to evolve throughout the MD simulation under some external condition, the GCMC simulations at the MD frames reflect the flexibility of MOFs at some level. This approach is called “*flexible snapshot adsorption*”.

An advantage of using the flexible snapshot adsorption method is the ease of the computation and a balance between computational cost and simulation accuracy [195]. It is comparatively less costly than *adsorption-relaxation* and more accurate than rigid adsorption. It allows to understand the MOFs’ structural behavior (i.e., accounting the flexibility at some level) under mechanical stress applied externally. The classical FFs might fail to accurately monitor these physical changes during mechanical stress. On the contrary, NNPs enable one to get more insights into these physical transformations during gas adsorption under external mechanical stress. In this context, pressure can also be utilized to control gas storage and release from MOFs. Indeed, the CH₄ uptakes of IRMOF-1 under mechanical compression has been reported by Zheng et al [123] using classical FFs.

Using similar protocol and NNP instead of classical FF for MOF definition, we have explored methane and hydrogen gas storage properties of IRMOF-1 and IRMOF-7 under applied external pressure. We have computed the CH₄ and H₂ uptakes of both compressed and uncompressed structures that are retrieved from strain applied MD

simulations using NNP. Overall, our calculations for loading of CH₄ on IRMOF-1 is in a good agreement with literature.

We have performed MD simulations using our NNP of uniaxial compression by reducing one of the direction vectors in several steps. The uniaxial compression of the systems was introduced by assigning the [001] plane to a vector and its gradual decrease. In line with literature studies, the $-3 \times 10^9 \text{ s}^{-1}$ strain rate was engineered as follows: During the MD simulations at the NVT ensemble, the box volume was reduced by decreasing the [001] plane by 0.03% after each ps simulation. A total 45% volume decrease in 150 ps total simulation time was achieved.

Adsorption properties of methane and hydrogen gases were studied using Grand Canonical Monte Carlo (GCMC) simulations with RASPA software code [196] using deformed/undeformed structures of IRMOF-1 and IRMOF-7 after extracting MD trajectories. At each step, deformed/undeformed MD frames from the trajectory were extracted for the calculation of density, accessible surface area (ASA), probe occupied pore volume (POAV) and POAV fraction using 1.86 Å probe (N₂) by Zeo++ [16, 93].

For both H₂ and CH₄ uptakes, GCMC simulations for all deformed/undeformed structures were run for 1000 cycles with the first 500 being equilibration and the last 500 for the ensemble average. In order to model guest molecules, single-site spherical Lennard-Jones (LJ) 12–6 potential taken from the TraPPE [197] force field was used in the GCMC simulations. The positions of deformed/undeformed MOF structures were kept fixed during GCMC simulations at the universal force field (UFF) [62] mixing with the Lorentz–Berthelot to assess the LJ cross-term guest/host parameters. Periodic boundary conditions (PBC) in all three directions were applied and a 12 Å interaction cutoff distance was used. The methane adsorption calculations were carried out at 298 K and 35 bar while for hydrogen we used 77 K and 35 bar.

As expected, the stress applied increases the density of both MOFs smoothly while reducing POAV fraction. Reducing volume also causes the compression of the organic linkers (Figure 5.2). Due to the steric hindrance of fused benzene rings, the decrease of POAV fraction in IRMOF-7 is predicted to be less than IRMOF1.

Figure 5.3 shows the CH₄ and H₂ uptakes of IRMOF-1 and IRMOF-7 against the strain applied on MOFs. Gravimetric methane uptake starts decreasing once the MOF starts deforming. Since the accessible surface area (ASA) and probe occupied accessible volume (POAV) decreases by the bending of organic linkers upon the stress applied, the total gravimetric uptake decreases.

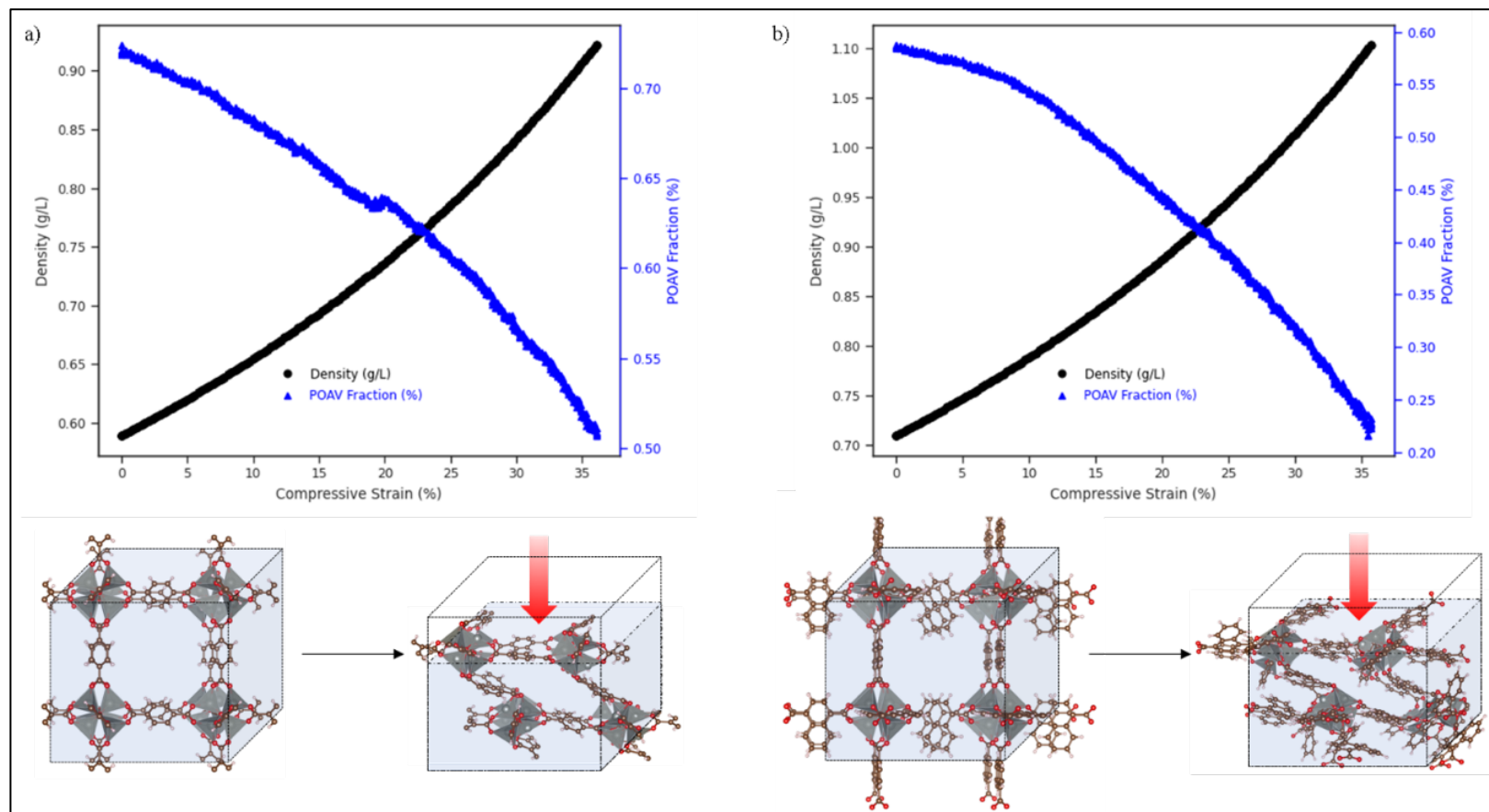


Figure 5.2: Changes in density and POAV fraction (top) upon compression along [001] direction for the fragments produced by MD simulations using NNP for a) IRMOF-1 and b) IRMOF-7. MOF structures are shown in cartoon (bottom) before (0%) and after (35%) strain applied.

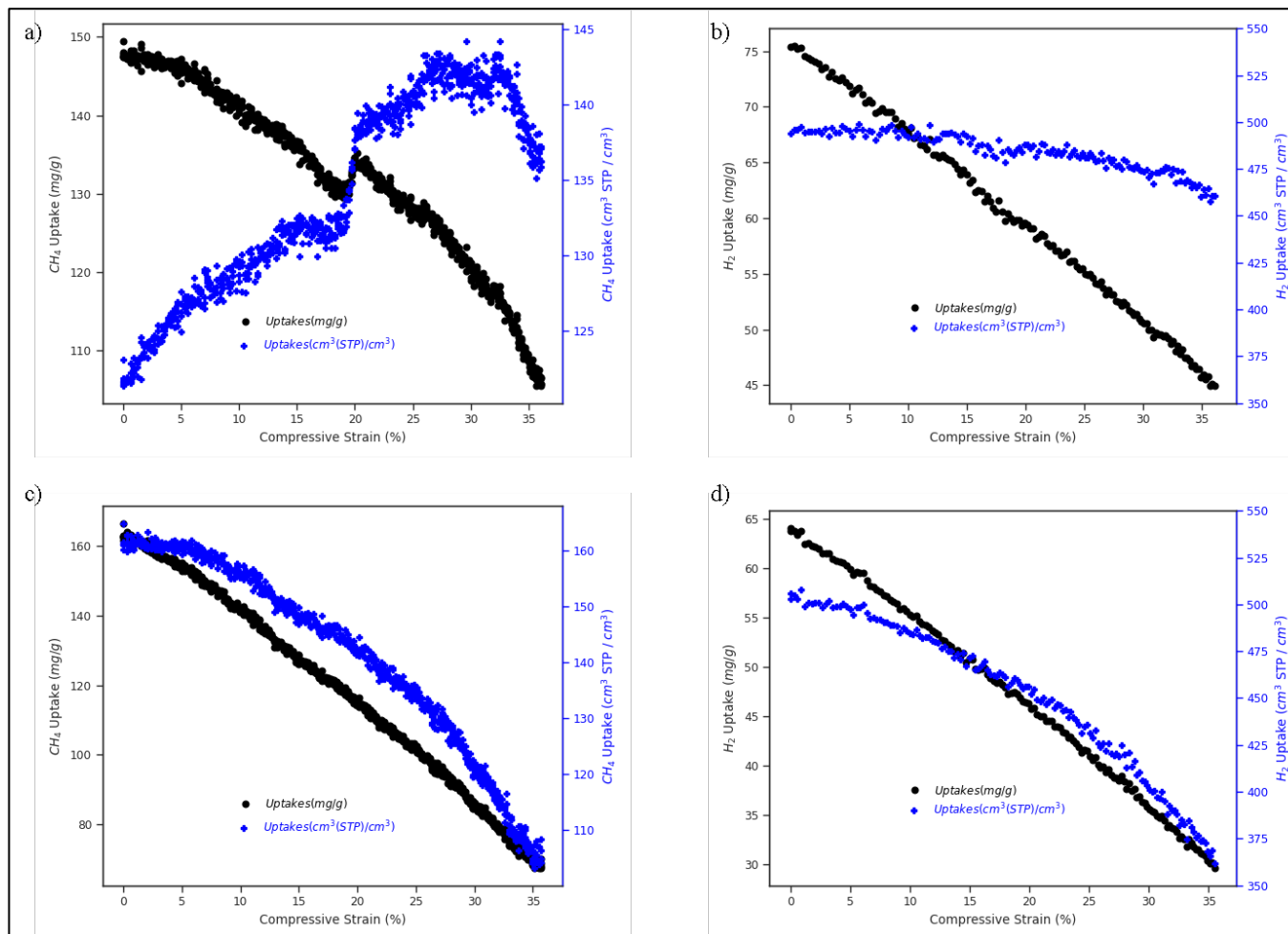


Figure 5.3: Gravimetric (black) and volumetric (blue) gas uptakes upon compression along [001] direction at each fragment obtained by MD simulations using NNP for CH₄ a), H₂ b) on IRMOF-1 and CH₄ c), H₂ d) on IRMOF-7.

5.5. Adsorption–Relaxation Simulations

To the best of our knowledge, accounting the flexibility of MOFs is best represented by *adsorption-relaxation* model. Although in the *flexible snapshot adsorption* model, MD simulations reflect the flexibility and beneficial in relation to computational cost, this method is not quite realistic since it does not allow dynamic interaction of adsorbates with the framework. Gas adsorption is calculated at rigid MOF snapshots instead of allowing the MOF to respond to the guest molecules as they traverse along the window. Furthermore, pressure is formed by forcibly reducing volume rather than inherent gas adsorption.

The *adsorption-relaxation* model can be summarized as follows: First, it starts with gas adsorption with grand canonical Monte Carlo (GCMC) on a periodic (bulk) rigid structure that has been stabilized thermally and energetically. Then, MD simulations are performed with the NPT barostat adjusted to the adsorption pressure to reflect the structural changes caused by the adsorbed gas and to see the effect of the structural changes on the adsorbed sites. Thus, gases and periodic structure are released at the same time with MD simulation, and both the gases and the periodic structure are rearranged according to the conditions in this part of the process. Implemented as an iterative process (Figure 5.4), the MD and GCMC procedures are sequentially applied, and this repetition continues until the structure achieves relaxation and the gas content between two iterations reaches a state of saturation.

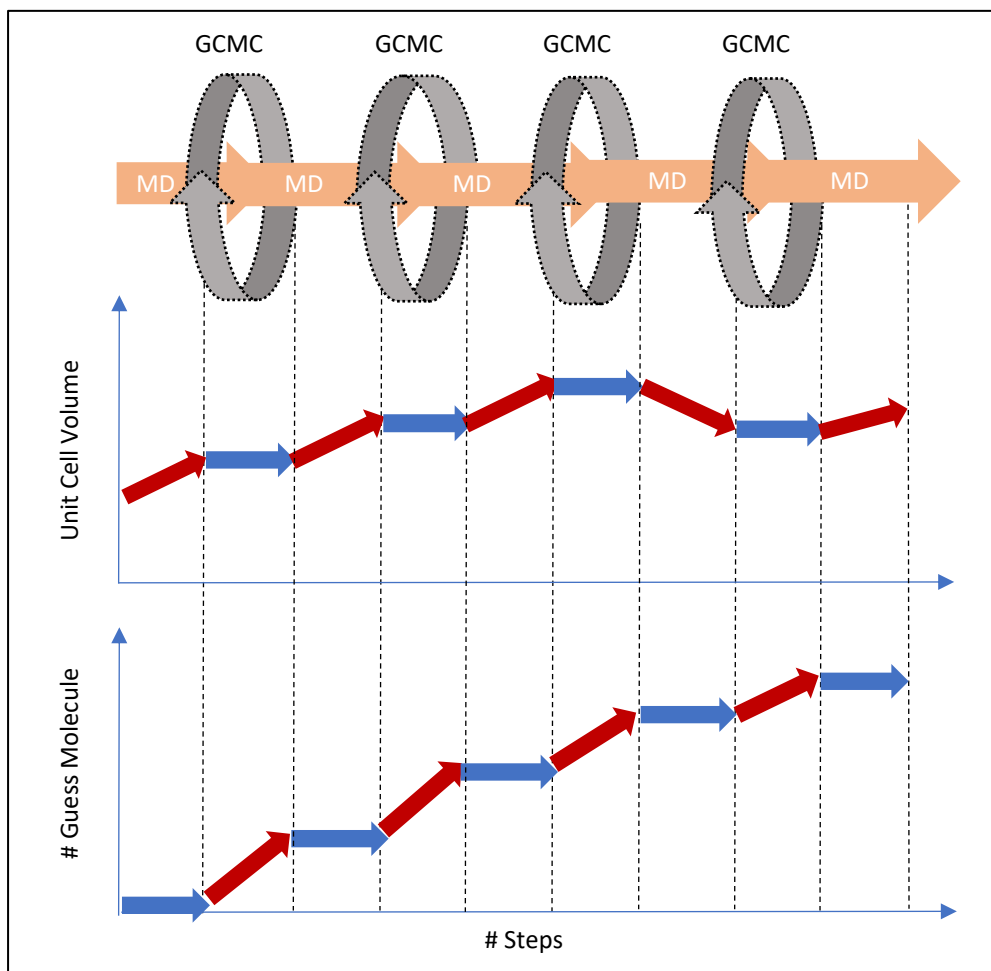


Figure 5.4: Shown is change in unit cell volume, number of molecules throughout the combined MC/MD hybrid simulation. Red arrows indicate properties that change, and blue arrows indicate properties that remain constant.

Recently, Sholl and co-workers have reported the screening of MOFs for the gas uptakes using combined MD/GCMC simulations using *adsorption-relaxation* model, which accounted for both extrinsic and intrinsic flexibility [195]. However, they have used classical force fields, which limited their simulations to classical harmonic vibrations for accounting the intrinsic flexibility. Among their studied MOFs, MIL-53, which has phase transition to possess either large pore or small pore sizes, was reproduced very well with MD/GCMC simulations whereas the rigid defined MIL-53 in either form was too off from the experimental results. In addition, IRMOF-1, which has very low extrinsic flexibility behaved just like the rigid definition. On the other hand, MD/GCMC on UiO-66 did not show significant difference from the rigid defined calculations. This might be due to the limitation of the force field. The same MOF has been also studied by the aforementioned work by Johnson and co-workers using NNP as the force field of the framework in the MD simulations by means of

diffusion of adsorbates. This showed superiority of using NNP with near-DFT accuracy over classical force fields in describing MOF structures. They also showed the feasibility of using hybrid potential as NNP combination with classical force fields in simulations. They successfully applied NNP for interatomic interactions of bare MOF while classical force field for adsorbate-adsorbate interactions along with combining rules to account for adsorbent-adsorbate interactions.

For the *adsorption-relaxation* flexible adsorption simulations on MOFs, we used this hybrid MD/GCMC protocol in which GCMC and MD are performed in LAMMPS software [198, 199]. Prior to gas adsorption simulations, we relaxed bare MOF structures by means of energy minimization using conjugate gradient (CG) algorithm with energy and force thresholds of 10^{-8} eV and 10^{-4} eV/Å, respectively. Following that, an NVT equilibration for 150 ps was performed at different temperatures of interest using The Nose/Hoover thermostat (i.e. 77, 200 and 300 K). Final structures were used in NPT-MD simulations which is combined with GCMC simulations. In the NPT simulations, The Nosé-Hoover thermostat and barostats were defined with 0.5 fs time steps. Damping constants were set to 0.1 ps for both thermostat and barostats. One step of GCMC simulation is performed in every 250 fs time of MD simulations. During the GCMC step, 100 random exchanges of adsorbate are attempted. Monte Carlo trial attempts were made by allowing translations, insertions and deletions all with equal probabilities. The rotation moves were turned off since the adsorbent molecules were assumed spherical. This MD/GCMC altering process is repeated until a total of 2 ns MD simulation time (A total of 8000 GCMC steps altered during 2 ns MD), of which the last 500 ps was used as production run.

5.5.1. Training NNP for LAMMPS

As discussed earlier, instead of using classical FFs in simulations, we rather intended to use NNPs, which should better monitor the structural changes and true adsorption profile. The interatomic interactions of MOFs are defined by our NNP, which captures the dynamic properties of the bulk periodic system. In chapter 4, we reported the construction of an NNP using SchNet via fragmentation of IRMOF series. In the previous subsections of chapter 4, we showed the capability of using NNPs to capture the flexibility of MOFs under mechanical stress using *flexible snapshot adsorption* model. Since this model does not require simultaneous MD and GCMC

simulations, we could successfully utilize ASE and RASPA. However, simultaneous MD/GCMC simulations are not available in ASE and thus, it is necessary to utilize LAMMPS for *adsorption-relaxation* simulations. Unfortunately, the original NNP we developed was not suitable for LAMMPS simulation package. Thus, in order to make it suitable, it is necessary to reconstruct NNP. Due to memory limitation, the data size needed to be reduced. Thus, we have developed LAMMPS suitable NNP using the same database after reducing data points from ~170K to ~98K data point via k-means clustering algorithm according to RMS distance.

In the previous chapter, we have discussed the NNP to model IRMOF-n series (n=1,4,6,7,10). Briefly, we used Behler's idea of fragmentation to reduce the computational cost of DFT calculations so as to train potential energy surfaces along with forces of non-equilibrium geometries. The initial dataset consisting of non-equilibrium geometries of unique fragments that define the MOF structure have been created by scaling interatomic distances and random displacements of atomic positions. This premature dataset has been developed further in iterative MD process using trial NNP as the calculator for the forces and energies.

Herein, we have reconstructed the NNP by n2p2 library using the same dataset in order to be make suitable for LAMMPS simulation package. Due to cost and memory limitations, a reduced version (~98k) of the original dataset (~173k) has been used in training. The elimination of almost half of the data was performed by clustering dataset by k-means and selecting representative from each of the clusters generated. K-means clustering was based on root mean square deviation (RMSD) of the 3D geometries.

Instead of using SchNet in our previous work, we used LAMMPS suitable n2p2 C++ library [200, 201], in which neural network potential formalism of Behler and Parrinello (BPNN) [67] is implemented, to train interatomic potentials. BPNN uses symmetry functions to describe atomic environments, and the total energies/forces are calculated by the summation of pairwise interatomic interactions. We generated symmetry functions with the cutoff radius for atomic environments of 6.0 Å. For the training, multi-stream Kalman filter parameters were used as implemented in n2p2 library as recommended by Singraber et al [200]. The only exception to these parameters was that we used the ratio of energy to force updates fixed at 1:15.

The dataset was split into training (~89k, 90%) and test (~9k, 10%) sets. Hyper parameters were used as default values except that we used 1:15 ratio of energy to

force updates. We stopped the training after 60 epochs (although it was converged within 30) with $E_{\text{trainrmse}} \leq 2.3$ meV/atom ($E_{\text{testrmse}} \leq 1.3$ meV) and $F_{\text{trainrmse}} \leq 45$ meV/Å ($F_{\text{testrmse}} \leq 53$ meV/Å). The loss of the training and validation curves are given in Figure D.1.

After constructing the NNP, we have tested on both train and test data sets. The RMSE of energy and forces are 0.0017 eV/atom and 0.23 eV/Å, respectively. The distribution of the errors is random and fits to a normal Gaussian function (Figure D.2). For all the fragments, NNP produced energies and forces are in a complete agreement with DFT calculated values ($R^2 \approx 0.999-1.000$). In addition, we analyzed the absolute errors produced by NNP and DFT calculations for the fragments. We observed that the absolute errors on almost all of the fragments lie within the narrow range of 0.005 eV/atom and 0.5 eV/Å for energy (Figure 5.5) and force, respectively (Figure 5.1Figure 5.6).

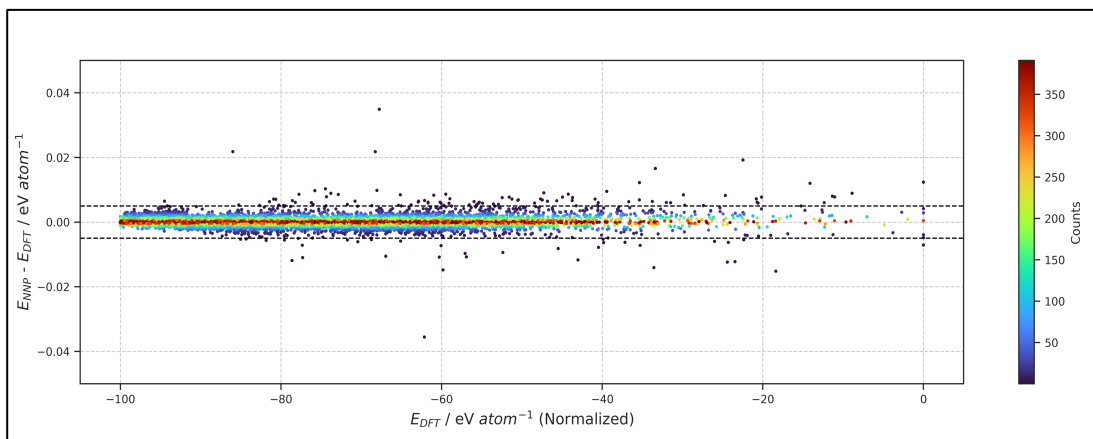


Figure 5.5: Errors in energies of fragments for test sets. Color bar shows the number of structures with the corresponding errors. Black dashed lines mark errors of 0.005 eV/atom. DFT energies (x axis) are normalized for a clear representation.

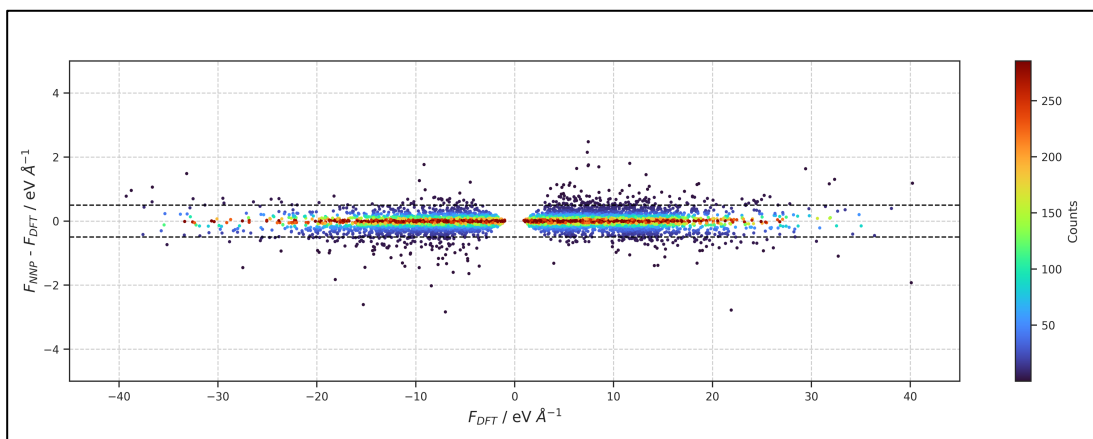


Figure 5.6: Errors in forces of fragments for test sets. Color bar shows the number of structures with the corresponding errors. Black dashed lines mark errors of 0.5 eV/Å.

5.5.2. NNP/UFF based MD/GCMC Simulations for IRMOF-1 and IRMOF-10

In our investigation, we have specifically selected two materials, IRMOF-1, and IRMOF-10, extending via benzene-dicarboxylate and biphenyl-dicarboxylate linkers, respectively. Due to the difference in linker length, IRMOF-1 is rather rigid whereas IRMOF-10 should possess relatively more flexibility due to the longer linker. The choice of these chemically similar linkers, depicted in Figure 5.7, enables a systematic exploration of the impact of pore size. In addition, we aimed to compare the effect of structural flexibility more clearly on gas adsorption including the effect of diffusion

rate by using gases such as H₂ and CH₄ in quite different sizes from each other in our simulations.

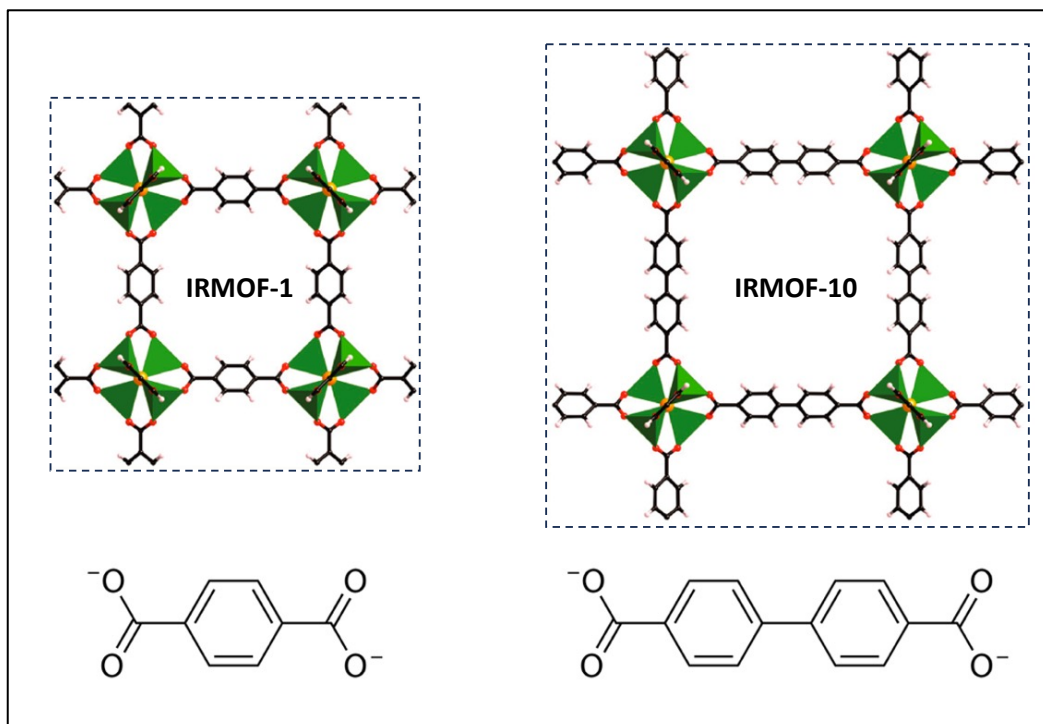


Figure 5.7: Variations in the types of organic linkers incorporated into IRMOF-1 and IRMOF-10 structures.

In order to demonstrate the superiority of our NNP for accurately depicting the structural characteristics of MOFs in comparison to classical force fields (FFs), we investigated the gas adsorption profiles of IRMOF-1 and IRMOF-10 structures with different pore sizes but identical overall topology. Our approach involved calculations utilizing a combination of NNP/FF potentials and GCMC/MD hybrid simulations lead to adsorption-relaxation flexibility, following the same methodology as described previously.

After transferring the NNP to LAMMPS, we have readily studied combined MD/GCMC simulations at different pressures (up to 1-100 bar at regular intervals) to load H₂ and CH₄ gases in each IRMOF-1 and IRMOF-10 using *adsorption-relaxation* model. Here, MOFs modelled by our NNP whereas remaining interaction such as gas-gas and gas-MOF modeled by UFF. The adsorbent gas molecules were assumed spherical Lennard-Jones fluids with no charges. The adsorbent-adsorbent interactions were modeled with the LJ potential.

$$u_{ij} = 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \quad (5.1)$$

where u_{ij} the potential energy, ϵ_{ij} is the well depth, σ_{ij} is the cutoff diameter and r_{ij} is the distance between two spherical gas molecules. For computing MOF-adsorbate interactions, first the LJ parameters for each atom in MOF were assigned by UFF, and then using Lorentz–Berthelot combining rules. ($\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$ and $\sigma_{ij} = \frac{1}{2}(\sigma_i + \sigma_j)$). The cutoff distance for all LJ interactions was set to 12 Å with tail energy corrections applied.

In the case of classical FF simulations, the dynamic properties of MOFs were modeled by UFF4MOF, whose accuracy has been demonstrated by comparing to ab-initio derived force fields by Moghadam et al [178]. Since UFF4MOF does not include partial charges on MOF atoms, the preassigned atomic charges of MOFs were retrieved from reference [202]. The non-bonding interactions among MOF atoms were defined as the coulombic and Lennard-Jones terms whereas MOF-gas and gas-gas interactions use only LJ potentials. For the LJ interactions the same parameters (i.e. 12 Å cutoff, Lorentz-Berthelot combining rules) were used.

To begin, 1 ns NVT MD simulations at 300 K on periodic bulk IRMOF-1 and IRMOF-10 structure were run followed by energy minimization in order to test the stability of the NNP. During the MD simulations, all configurations visited were located within the extensively sampled regions of the NNP, and no significant alterations in structure were observed except for minor thermal fluctuations. MC/MD hybrid simulations were performed at 100 bar and 77K. Each simulation was performed for 2 ns and we determined the amount of gas adsorption using the last 1 ns for analysis.

Figure 5.8 shows H₂ gas uptakes of IRMOF-1 at 100 bar and 77K using a adsorption-relaxation model based on the pure UFF force field. Number of H₂ molecules loaded on IRMOF-1 reach to a plateau at around 250 ps. After that, it maintains its saturated concentration of ~2200 molecules. The figure also shows running average of adsorption profile, which indicates the smoothed adsorption reaching to the same plateau. This figure shows the MD simulation at one specific pressure (100 bar). The adsorption isotherm is created by the collection of different

MD simulations at all pressure values (1-100 bar). Although the simulations are converged within 250 ps, we used the average loading of the last 1 ns in order to ensure the converged loadings. Since the process is involved in both MD and GCMC simulations, at each insertion of gas molecules, the system is equilibrated at the preset pressure. Allowing the system to relax during adsorption causes MOF structures to realign themselves with this new condition, in which adsorbate molecules are inserted or deleted.

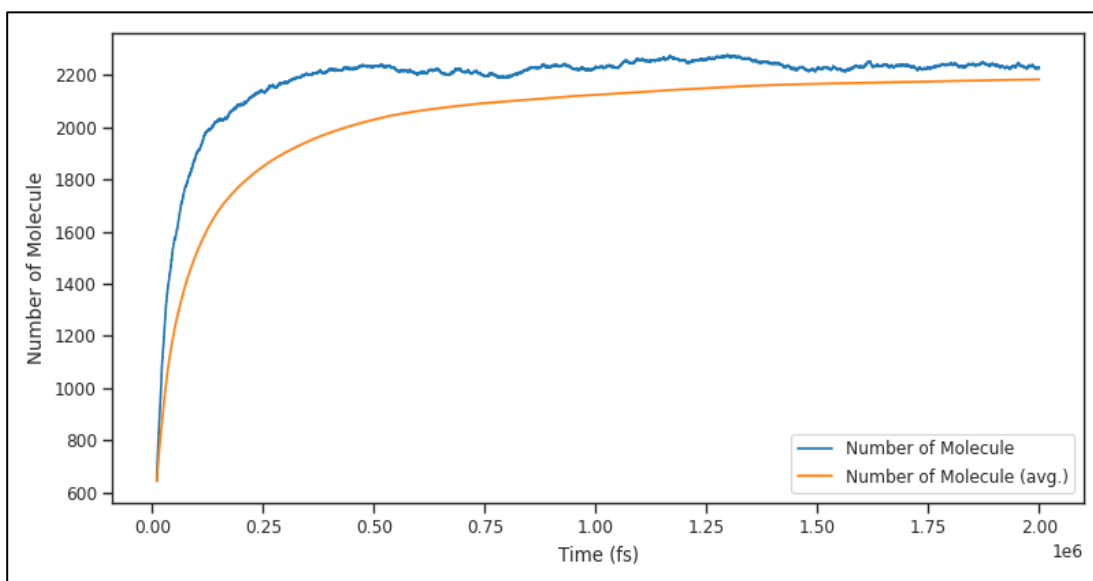


Figure 5.8: The number of adsorbed H₂ molecules versus time for the simulations of adsorption of RMOF-1 for hydrogen gas at 100 bar and 77K using UFF4MOF force fields in MD/GCMC hybrid simulations

Defining the MOF structure with NNP instead of UFF, the same H₂ loading was monitored throughout the MD/GCMC simulation at 100 bar and 77 K (Figure 5.9). Notably, saturation achieved by loading more H₂ molecules (~2600) than with pure UFF (~2200) suggests that the NNP/UFF approach captures adsorption-relaxation processes in a distinct manner. In the following section, we will assess the effectiveness of NNP potentials in capturing the flexibility of MOFs by comparing pure UFF simulations to NNP/UFF simulations. Similar to UFF, the same plateau is reached, and steady loadings are maintained with NNP at around 250 ps, showing the convergence of the MD/GCMC simulations. Once again, the adsorption isotherm is created by the collection of different MD simulations at all pressure values (1-100 bar).

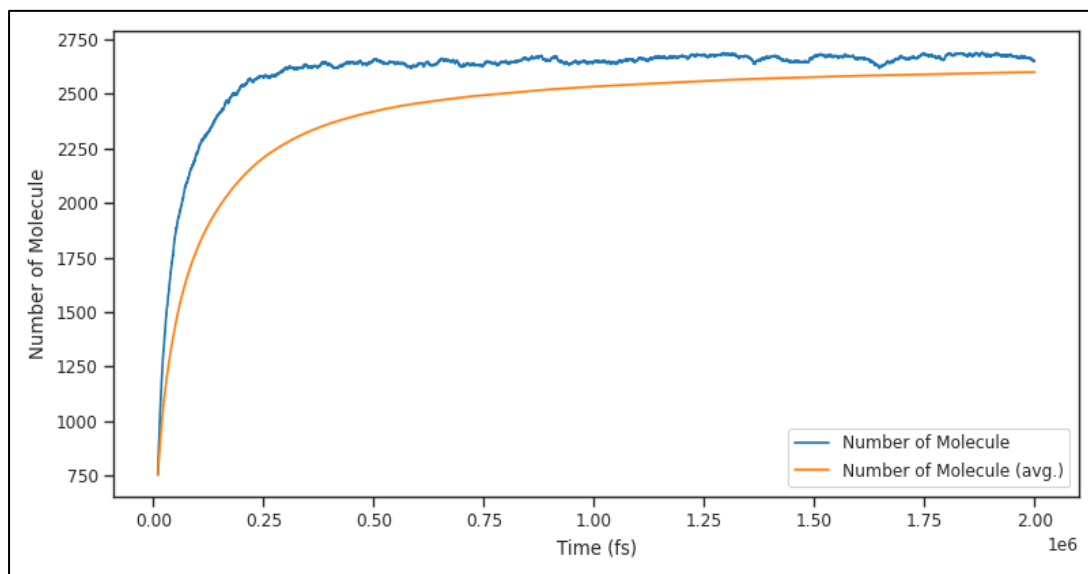


Figure 5.9: The number of adsorbed H₂ molecules versus time for the simulations of adsorption of RMOF-1 for hydrogen gas at 100 bar and 77K using combined HDNNP/UFF4MOF force fields in GCMC/MD hybrid simulations

Figure 5.10 illustrates H₂ a) and CH₄ b) adsorption isotherm on IRMOF-1 obtained from both the HDNNP/UFF and pure UFF as well as experimental data. At low pressure, the general behavior of both methods (HDNNP vs. pure UFF) agree with each other and experimental values. However, when the pressure is increased, HDNNP tend to slightly overestimate the experimental values. This is typical, since simulations employ impurity-free and perfect structures as opposed to experiments, which are affected by impurities. However, at higher pressures, UFF underestimates the H₂ uptake, which should not normally occur. This might indicate the limitation of the use of classical force field at higher pressure values. It should be noted that H₂ is very small, and IRMOF-1 is rather rigid, and thus even classical FF can model such a system accurately. Therefore, the adsorption isotherm produced by HDNNP vs. UFF follow a similar trend even at higher pressures. On the other hand, a larger molecule, CH₄ makes a significant difference in the adsorption behavior predicted by the simulations using HDNNP vs. UFF for IRMOF-1. For this adsorbate, both HDNNP and UFF correctly predict the experimental values at low pressures. The trend of CH₄ uptakes over different pressures is totally diverged from the experiment in UFF whereas for HDNNP, it is still in good agreement with experimental values. A small deviation as in the case of H₂ adsorption can easily be fixed by setting an offset, which should be due to impurities. For a better illustration of how the UFF based calculation diverges from experiment while HDNNP matches with the experiment, we calculated the ratio

of predicted values (from both method) to experimental values (Figure 5.11). The ratio of HDNNP/Experimental follows a rather steady constant line at 1.1 at all pressures while UFF/Experimental does not have a constant value over the pressures and it is changing at every different pressure. From these observations, it can be concluded that the high-pressure gas uptakes of MOFs are best modeled by HDNNPs, demonstrating improved correlation with experimental values.

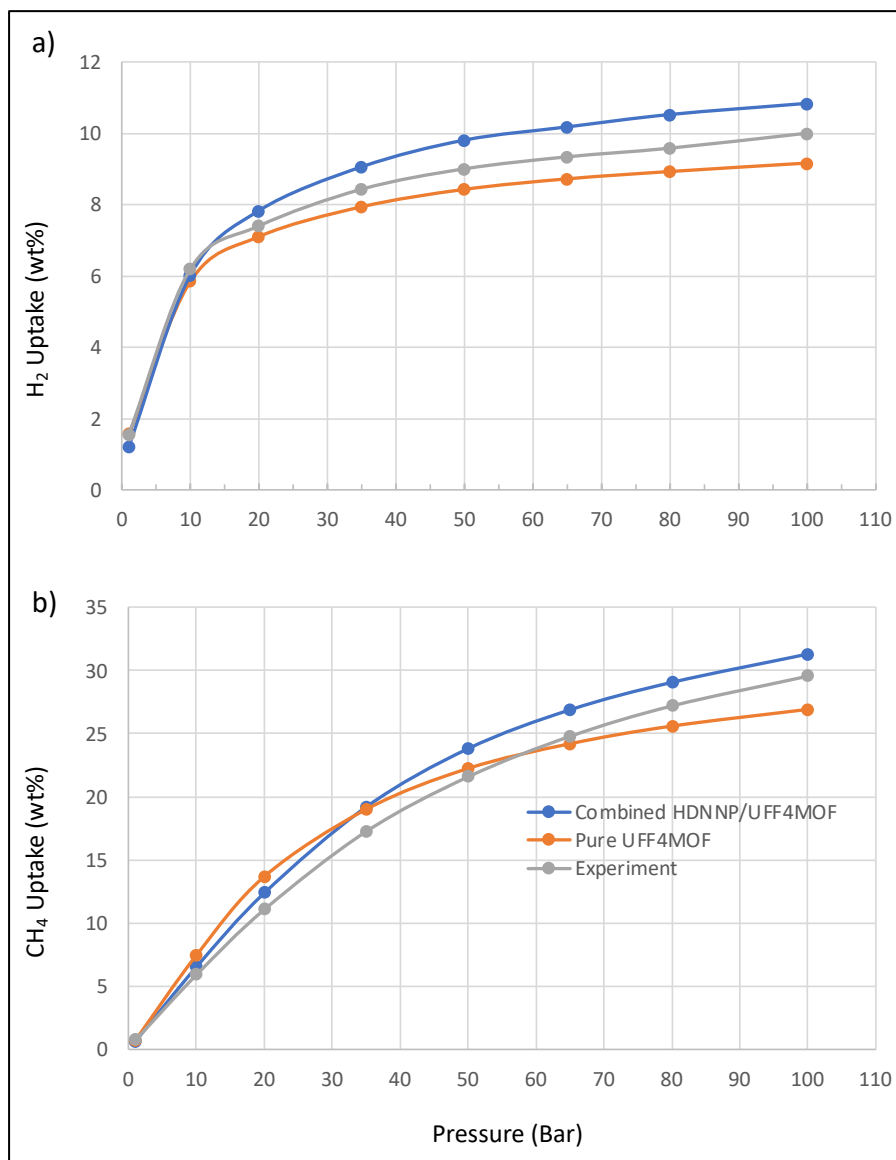


Figure 5.10: Adsorption profiles showing the uptake of a) H₂ and b) CH₄ on IRMOF-1 using combined HDNNP/UFF4MOF potentials, pure UFF4MOF simulations, and experimental data.

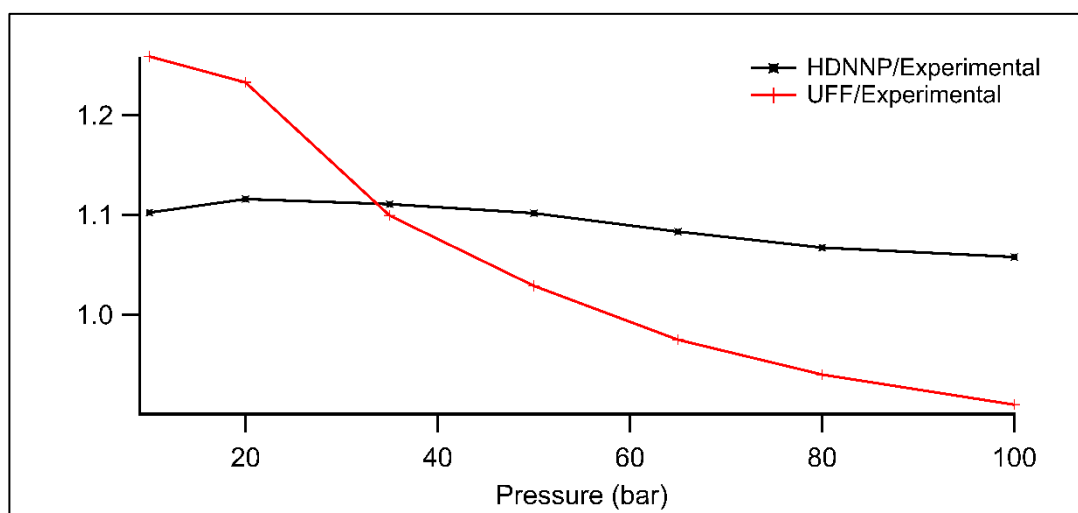


Figure 5.11: The ratio of predicted/experimental loadings at different pressures. Predicted values are of pure UFF and HDNNP. HDNNP follows a steady constant line over the pressures while UFF is totally divergent.

As mentioned earlier, IRMOF-10 should possess more flexibility than IRMOF-1, owing to its longer linker. Due to this flexibility in IRMOF-10, the difference between the adsorption curves of H_2 and CH_4 gas produced by HDNNP and UFF changes significantly at high pressures (Figure 5.12). Unfortunately, there is no experimental data on the adsorption of these gases in IRMOF-10 in the literature, but the fact that this difference is quite large might indicate better modeling of the flexibility by the NNP.

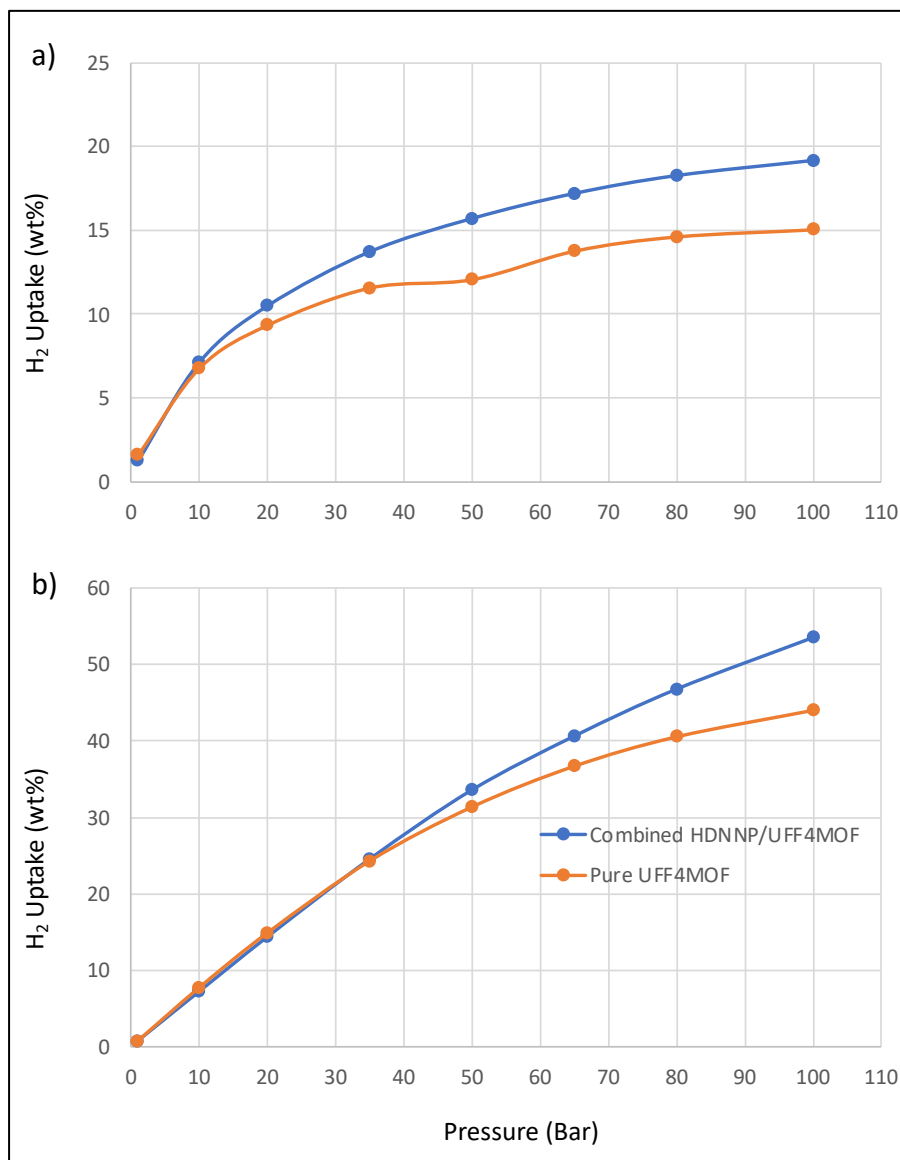


Figure 5.12: Adsorption profiles showing the uptake of a) H₂ and b) CH₄ on IRMOF-10 using combined HDNNP/UFF4MOF potentials and pure UFF4MOF simulations.

Table 5.1 lists the absolute values of uptakes (wt%) of H₂ and CH₄ gases in IRMOF-1 and IRMOF-10 at 100 bars. The absolute value of CH₄ uptake predicted by pure UFF deviates by 2.67 from the experimental value while that of HDNNP is only 1.71. Although there is no experimental data for CH₄ uptake on IRMOF-10, it is believed to have a much closer value to 53.59%.

Table 5.1: Total H₂ and CH₄ uptake on IRMOF-1 and IRMOF-10 (wt %) at 100 bar; 77K for H₂ and 300K for CH₄

	IRMOF-1		IRMOF-10	
Uptake(wt%)	H ₂	CH ₄	H ₂	CH ₄
Pure UFF4MOF	9.16	26.9	15.02	44.06
HDNNP+UFF4MOF	10.8	31.2	19.14	53.59
Experiment	10.0	29.5	-	-

5.6. Conclusions

In this chapter, we have transferred our neural network potential, which was developed in chapter 4, to describe rotationally and translationally invariant energies and forces of IRMOFs at the DFT level of accuracy using the fragmentation technique so as to study H₂ and CH₄ adsorption isotherms considering flexibility. First, mechanical strain was applied on IRMOFs during MD simulations with our developed NNP and snapshots from trajectory was used in GCMC simulations (i.e., *flexible snapshot adsorption*). The results suggested an increase in volumetric CH₄ uptake in IRMOF-1 and a decrease in IRMOF-7. Second, our NNP was successfully utilized to study the flexibility of MOFs during gas adsorption by means of “*adsorption relaxation*” model in which MD and GCMC simulations were performed simultaneously. Our results showed that classical FF may diverge from experiments when accounting for flexibility in MOFs whereas our NNP follows a much better trend to experimental values. In addition, with the absence of experimental data showed that the real number of gas uptake values can be more than what classical FF predicted.

6. CONCLUDING REMARKS AND FUTURE DIRECTIONS

This dissertation systematically investigates the physisorption of two important energy related gases on highly porous metal-organic frameworks by computational methods.

As introduced in chapter 3, the dissertation starts with screening published manuscripts to retrieve the highest hydrogen gas storing MOF via extracting experimental surface area (SA) and pore volume (PV) using a text and data mining (TDM) technique. The TDM method, in which the full text of nearly sixty thousand manuscripts were accessed, has been the first extensive study in the field. In particular, American Chemical Society (ACS) has created solutions for TDM in their journals using our publication as a lead and tutor. Our manuscript is still available on the main page of ACS web site as a case study. In this study, we suggested NU-n and IRMOF-n series to be the most promising H₂ storage materials. After our publication, these MOFs have extensively been the subject of further experimental studies. Our methodology in TDM studies can easily be modified to screen other gases of interest on these and other porous structures. Moreover, our algorithm can be modified to screen other properties than SA and PV for diverse aspects of porous materials.

Due to lack of experimental studies on NU-n series at the time, we have used IRMOF-series in our next studies. Another reason why we chose IRMOFs is that for NNP construction, we followed the procedure by Behler and coworkers, who reported HDNNP for IRMOF-1 [135]. Since the studies discussed in chapter 4 are the state-of-the art in the field, we needed benchmarking with available data in literature both experimentally and computationally. In this chapter, the concept work by Behler and coworkers has been for the first time applied systematically for the features of IRMOFs. Our constructed NNP required only one additional fragment for each of the MOFs derived from IRMOF-1. In this study, we have shown that a single neural network potential can represent all the IRMOF series. In addition, we have for the first time reported very unusual behaviors of MOFs under thermal stress.

Herein, we have applied a text and data mining (TDM) technique to search for experimental surface area (SA) and pore volume (PV) values of metal-organic frameworks (MOFs). We collected SA and PV values in two separate tables of each nearly 14 600 values. Manual verifications on randomly chosen values showed ~80%

success for SA and 85% success for PV values in a complete agreement with literature values that are limited to ~ 1000 data. In addition, we have introduced two new methods to compare and merge available SA and PV values based on MOF names, ending up with an ultimate table of TDM experimental SA and PV values with nearly 2100 data. Further manual verification of randomly selected test data showed $\sim 80\%$ success in merged data. Using an experimental equation correlating SA and PV values to H_2 uptakes, we predicted H_2 uptakes of nearly 2100 MOFs.

In addition to TDM, we calculated accessible surface area (ASA) and probe-occupiable pore volumes (POAV) purely based on the coordinates of single-crystal x-ray diffraction structure using spherical probes of N_2 and H_2 with 1.86 and 1.20 Å, respectively, as a comparison to and cover of our TDM and available literature data. Our structural calculations based on spherical probes were in quite good agreement with our TDM. In addition, we compared TDM and structural calculations by means of probability histograms according to SA and PV value intervals with perfect agreement.

In chapter 5, we successfully retrained and transferred the neural network potential (NNP) to make it suitable for combined MD and GCMC simulations under mechanical stress applied externally or arisen inherently from gas adsorption. We have predicted H_2 and CH_4 adsorption isotherms using two different approaches, which accounted for flexibility of MOFs. Our approach has demonstrated higher agreement with experimental values compared to classical force fields (FF). The study highlighted the importance of considering MOF flexibility in gas adsorption simulations and indicated the potential for higher gas uptake values than predicted by traditional FF models. In particular, other MOFs with more flexibility involved will benefit from the findings of the current studies. The studies will open gates for the studies about the MOF behaviors such as expanding, contracting, breathing effects as well as phase transitions, diffusion rates of different gases. In addition to its findings among MOFs studied, since these studies are among the first studies in the field, it sets a step-by-step guide to the reader to carry out such ultra-modern simulations.

Throughout the dissertation we have developed NNPs and verified the performance of NNP using the available (but limited) experimental and theoretical data in literature. Unfortunately, there is not plenty of experimental data to see broader impact of the studies discussed herein. For instance, the current studies would greatly benefit from experimental analysis such as x-ray single crystal spectroscopy of MOFs

at varying temperatures. The interesting conversion of thermal expansions from negative to positive for IRMOFs discussed in chapter 4 (in particular IRMOF-4 and IRMOF-7) would be experimentally verified. Similarly, H_2 and CH_4 adsorption isotherm at high pressures (up to 100 bar) experiments for other IRMOFs (in particular IRMOF-7 and IRMOF-10) would support the findings of chapter 5.

REFERENCES

- [1] Abe J. O., Popoola A. P. I., Ajenifuja E., Popoola O. M., (2019), “Hydrogen energy, economy and storage: Review and recommendation”, *International Journal of Hydrogen Energy*, 44 (29), 15072–15086, doi: 10.1016/j.ijhydene.2019.04.068.
- [2] Oldenbroek V., Wijtzes S., Blok K., van Wijk A. J. M., (2021), “Fuel cell electric vehicles and hydrogen balancing 100 percent renewable and integrated national transportation and energy systems”, *Energy Conversion and Management: X*, 9 100077, doi: 10.1016/j.ecmx.2021.100077.
- [3] Bičáková O., Straka P., (2012), “Production of hydrogen from renewable resources and its effectiveness”, *International Journal of Hydrogen Energy*, 37 (16), 11563–11578, doi: 10.1016/j.ijhydene.2012.05.047.
- [4] Bosch M., Zhang M., Zhou H.-C., (2014), “Increasing the Stability of Metal-Organic Frameworks”, *Advances in Chemistry*, 2014 1–8, doi: 10.1155/2014/182327.
- [5] Yang J., Sudik A., Wolverton C., Siegel D. J., (2010), “High capacity hydrogenstorage materials: attributes for automotive applications and techniques for materials discovery”, *Chemical Society Reviews*, 39 (2), 656–675, doi: 10.1039/B802882F.
- [6] Xiong Z. *et al.*, (2008), “Interaction of lithium hydride and ammonia borane in THF”, *Chemical Communications*, (43), 5595, doi: 10.1039/b812576g.
- [7] Férey G., (2008), “Hybrid porous solids: past, present, future”, *Chemical Society Reviews*, 37 (1), 191–214, doi: 10.1039/B618320B.
- [8] Horike S., Shimomura S., Kitagawa S., (2009), “Soft porous crystals”, *Nature Chemistry*, 1 (9), 695–704, doi: 10.1038/nchem.444.
- [9] Rowsell J. L. C., Yaghi O. M., (2005), “Strategies for Hydrogen Storage in Metal-Organic Frameworks”, *Angewandte Chemie International Edition*, 44 (30), 4670–4679, doi: 10.1002/anie.200462786.
- [10] Zou L., Zhou H. C., (2017), “Hydrogen storage in metal-organic frameworks”, *Nanostructured Materials for Next-Generation Energy Storage and Conversion: Hydrogen Production, Storage, and Utilization*, 143–170, doi: 10.1007/978-3-662-53514-1_5.
- [11] Gygi D. *et al.*, (2016), “Hydrogen Storage in the Expanded Pore Metal–Organic Frameworks $M_2(\text{dobpdc})$ ($M = \text{Mg, Mn, Fe, Co, Ni, Zn}$)”, *Chemistry of Materials*, 28 (4), 1128–1138, doi: 10.1021/acs.chemmater.5b04538.
- [12] Hu Y. H., Zhang L., (2010), “Hydrogen Storage in Metal-Organic Frameworks”, *Advanced Materials*, 22 (20), E117–E130, doi: 10.1002/adma.200902096.

- [13] Zhou H. C., Long J. R., Yaghi O. M., (2012), “Introduction to Metal–Organic Frameworks”, *Chemical Reviews*, 112 (2), 673–674, doi: 10.1021/cr300014x.
- [14] Li H., Eddaoudi M., O’Keeffe M., Yaghi O. M., (1999), “Design and synthesis of an exceptionally stable and highly porous metal-organic framework”, *Nature*, 402 (6759), 276–279, doi: 10.1038/46248.
- [15] Park S., Kim B., Choi S., Boyd P. G., Smit B., Kim J., (2018), “Text Mining Metal-Organic Framework Papers”, *Journal of Chemical Information and Modeling*, 58 (2), 244–251, doi: 10.1021/acs.jcim.7b00608.
- [16] Ongari D., Boyd P. G., Barthel S., Witman M., Haranczyk M., Smit B., (2017), “Accurate Characterization of the Pore Volume in Microporous Crystalline Materials”, *Langmuir*, 33 (51), 14529–14538, doi: 10.1021/acs.langmuir.7b01682.
- [17] Corma A., García H., Llabrés i Xamena F. X., (2010), “Engineering Metal Organic Frameworks for Heterogeneous Catalysis”, *Chemical Reviews*, 110 (8), 4606–4655, doi: 10.1021/cr9003924.
- [18] Horcajada P. *et al.*, (2010), “Porous metal–organic-framework nanoscale carriers as a potential platform for drug delivery and imaging”, *Nature Materials*, 9 (2), 172–178, doi: 10.1038/nmat2608.
- [19] Yaghi O. M., O’Keeffe M., Ockwig N. W., Chae H. K., Eddaoudi M., Kim J., (2003), “Reticular synthesis and the design of new materials”, *Nature*, 423 (6941), 705–714, doi: 10.1038/nature01650.
- [20] Xiang S.-C. *et al.*, (2011), “Rationally tuned micropores within enantiopure metal-organic frameworks for highly selective separation of acetylene and ethylene”, *Nature Communications*, 2 (1), 204, doi: 10.1038/ncomms1206.
- [21] Lee J. S., Vlasisavljevich B., Britt D. K., Brown C. M., Haranczyk M., Neaton J. B., Smit B., Long J. R., Queen W. L., (2015), “Understanding Small-Molecule Interactions in Metal-Organic Frameworks: Coupling Experiment with Theory”, *Advanced Materials*, 27 (38), 5785–5796, doi: 10.1002/adma.201500966.
- [22] Odoh S. O., Cramer C. J., Truhlar D. G., Gagliardi L., (2015), “Quantum-Chemical Characterization of the Properties and Reactivities of Metal–Organic Frameworks”, *Chemical Reviews*, 115 (12), 6051–6111, doi: 10.1021/cr500551h.
- [23] Tayfuroglu O., Kocak A., Zorlu Y., (2019), “In Silico Investigation into H₂ Uptake in MOFs: Combined Text/Data Mining and Structural Calculations”, *Langmuir*, 36 (1), 119–129, doi: 10.1021/acs.langmuir.9b03618.
- [24] Evans J. D., Fraux G., Gaillac R., Kohen D., Trouselet F., Vanson J.-M., Coudert F.-X., (2017), “Computational Chemistry Methods for Nanoporous Materials”, *Chemistry of Materials*, 29 (1), 199–212, doi: 10.1021/acs.chemmater.6b02994.

- [25] Datta S. S., Ramakrishnan T. S., Weitz D. A., (2014), “Mobilization of a trapped non-wetting fluid from a three-dimensional porous medium”, *Physics of Fluids*, 26 (2), 83–89, doi: 10.1063/1.4866641.
- [26] Tarca A. L., Carey V. J., Chen X., Romero R., Drăghici S., (2007), “Machine Learning and Its Applications to Biology”, *PLoS Computational Biology*, 3 (6), e116, doi: 10.1371/journal.pcbi.0030116.
- [27] Carrasquilla J., Melko R. G., (2017), “Machine learning phases of matter”, *Nature Physics*, 13 (5), 431–434, doi: 10.1038/nphys4035.
- [28] Carleo G., Cirac I., Cranmer K., Daudet L., Schuld M., Tishby N., Vogt-Maranto L., Zdeborová L., (2019), “Machine learning and the physical sciences”, *Reviews of Modern Physics*, 91 (4), 045002, doi: 10.1103/RevModPhys.91.045002.
- [29] Goh G. B., Hodas N. O., Vishnu A., (2017), “Deep learning for computational chemistry”, *Journal of Computational Chemistry*, 38 (16), 1291–1307, doi: 10.1002/jcc.24764.
- [30] Fanourgakis G. S., Gkagkas K., Tylianakis E., Froudakis G., (2020), “A Generic Machine Learning Algorithm for the Prediction of Gas Adsorption in Nanoporous Materials”, *The Journal of Physical Chemistry C*, 124 (13), 7117–7126, doi: 10.1021/acs.jpcc.9b10766.
- [31] Borboudakis G., Stergiannakos T., Frysali M., Klontzas E., Tsamardinos I., Froudakis G. E., (2017), “Chemically intuited, large-scale screening of MOFs by machine learning techniques”, *npj Computational Materials*, 3 (1), 40, doi: 10.1038/s41524-017-0045-8.
- [32] Born M., Oppenheimer R., (1927), “Zur Quantentheorie der Molekeln”, *Annalen der Physik*, 389 (20), 457–484, doi: <https://doi.org/10.1002/andp.19273892002>.
- [33] Kresse G., Furthmüller J., (1996), “Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set”, *Computational Materials Science*, 6 (1), 15–50, doi: 10.1016/0927-0256(96)00008-0.
- [34] Hehre W. J., Stewart R. F., Pople J. A., (1969), “Self-Consistent Molecular-Orbital Methods. I. Use of Gaussian Expansions of Slater-Type Atomic Orbitals”, *The Journal of Chemical Physics*, 51 (6), 2657–2664, doi: 10.1063/1.1672392.
- [35] Blum V., Gehrke R., Hanke F., Havu P., Havu V., Ren X., Reuter K., Scheffler M., (2009), “Ab initio molecular simulations with numeric atom-centered orbitals”, *Computer Physics Communications*, 180 (11), 2175–2196, doi: 10.1016/j.cpc.2009.06.022.
- [36] Strout D. L., Scuseria G. E., (1995), “A quantitative study of the scaling properties of the Hartree–Fock method”, *The Journal of Chemical Physics*, 102 (21), 8448–8452, doi: 10.1063/1.468836.

- [37] Møller Chr., Plesset M. S., (1934), “Note on an Approximation Treatment for Many-Electron Systems”, *Physical Review*, 46 (7), 618–622, doi: 10.1103/PhysRev.46.618.
- [38] David Sherrill C., Schaefer H. F., “The Configuration Interaction Method: Advances in Highly Correlated Approaches,” in *Advances in Quantum Chemistry*, 1999, 143–269. doi: 10.1016/S0065-3276(08)60532-8.
- [39] Sinanoğlu O., (1962), “Many-Electron Theory of Atoms and Molecules. I. Shells, Electron Pairs vs Many-Electron Correlations”, *The Journal of Chemical Physics*, 36 (3), 706–717, doi: 10.1063/1.1732596.
- [40] KÜMMEL H. G., (2003), “A BIOGRAPHY OF THE COUPLED CLUSTER METHOD”, *International Journal of Modern Physics B*, 17 (28), 5311–5325, doi: 10.1142/S0217979203020442.
- [41] Rajagopal A. K., Callaway J., (1973), “Inhomogeneous Electron Gas”, *Physical Review B*, 7 (5), 1912–1919, doi: 10.1103/PhysRevB.7.1912.
- [42] Harker A. H., (2016), “Materials Modelling using Density Functional Theory: Properties and Predictions, by Giustino Feliciano”, *Contemporary Physics*, 57 (1), , doi: 10.1080/00107514.2015.1100218.
- [43] Perdew J. P., Burke K., Ernzerhof M., (1996), “Generalized Gradient Approximation Made Simple”, *Physical Review Letters*, 77 (18), 3865–3868, doi: 10.1103/PhysRevLett.77.3865.
- [44] Perdew J. P., Ernzerhof M., Burke K., (1996), “Rationale for mixing exact exchange with density functional approximations”, *The Journal of Chemical Physics*, 105 (22), 9982–9985, doi: 10.1063/1.472933.
- [45] Stephens P. J., Devlin F. J., Chabalowski C. F., Frisch M. J., (1994), “Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields”, *The Journal of Physical Chemistry*, 98 (45), 11623–11627, doi: 10.1021/j100096a001.
- [46] Verlet L., (1967), “Computer ‘Experiments’ on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules”, *Physical Review*, 159 (1), 98–103, doi: 10.1103/PhysRev.159.98.
- [47] Nosé S., (1984), “A unified formulation of the constant temperature molecular dynamics methods”, *The Journal of Chemical Physics*, 81 (1), 511–519, doi: 10.1063/1.447334.
- [48] Hoover W. G., (1985), “Canonical dynamics: Equilibrium phase-space distributions”, *Physical Review A*, 31 (3), 1695–1697, doi: 10.1103/PhysRevA.31.1695.
- [49] Hoover W. G., Ladd A. J. C., Moran B., (1982), “High-strain-rate plastic flow studied via nonequilibrium molecular dynamics”, *Physical Review Letters*, 48 (26), , doi: 10.1103/PhysRevLett.48.1818.

- [50] Andersen H. C., (1980), “Molecular dynamics simulations at constant pressure and/or temperature”, *The Journal of Chemical Physics*, 72 (4), 2384–2393, doi: 10.1063/1.439486.
- [51] Hastings W. K., (1970), “Monte Carlo sampling methods using Markov chains and their applications”, *Biometrika*, 57 (1), 97–109, doi: 10.1093/biomet/57.1.97.
- [52] Gelman A., Gilks W. R., Roberts G. O., (1997), “Weak convergence and optimal scaling of random walk Metropolis algorithms”, *The Annals of Applied Probability*, 7 (1), , doi: 10.1214/aoap/1034625254.
- [53] Brotin T., Dutasta J.-P., (2009), “Cryptophanes and Their Complexes—Present and Future”, *Chemical Reviews*, 109 (1), 88–130, doi: 10.1021/cr0680437.
- [54] Taratula O., Hill P. A., Khan N. S., Carroll P. J., Dmochowski I. J., (2010), “Crystallographic observation of ‘induced fit’ in a cryptophane host–guest model system”, *Nature Communications*, 1 (1), 148, doi: 10.1038/ncomms1151.
- [55] Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H., Teller E., (1953), “Equation of State Calculations by Fast Computing Machines”, *The Journal of Chemical Physics*, 21 (6), 1087–1092, doi: 10.1063/1.1699114.
- [56] Jones J. E., (1924), “On the determination of molecular fields.—I. From the variation of the viscosity of a gas with temperature”, *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 106 (738), 441–462, doi: 10.1098/rspa.1924.0081.
- [57] Jones J. E., (1924), “On the determination of molecular fields. —II. From the equation of state of a gas”, *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 106 (738), 463–477, doi: 10.1098/rspa.1924.0082.
- [58] van Duin A. C. T., Dasgupta S., Lorant F., Goddard W. A., (2001), “ReaxFF: A Reactive Force Field for Hydrocarbons”, *The Journal of Physical Chemistry A*, 105 (41), 9396–9409, doi: 10.1021/jp004368u.
- [59] Gale J. D., (1996), “Empirical potential derivation for ionic materials”, *Philosophical Magazine B*, 73 (1), 3–19, doi: 10.1080/13642819608239107.
- [60] Brooks B. R., Bruccoleri R. E., Olafson B. D., States D. J., Swaminathan S., Karplus M., (1983), “CHARMM: A program for macromolecular energy, minimization, and dynamics calculations”, *Journal of Computational Chemistry*, 4 (2), 187–217, doi: 10.1002/jcc.540040211.
- [61] Phillips J. C. *et al.*, (2005), “Scalable molecular dynamics with NAMD”, *Journal of Computational Chemistry*, 26 (16), 1781–1802, doi: 10.1002/jcc.20289.
- [62] Rappe A. K., Casewit C. J., Colwell K. S., Goddard W. A., Skiff W. M., (1992), “UFF, a full periodic table force field for molecular mechanics and molecular

- dynamics simulations”, *Journal of the American Chemical Society*, 114 (25), 10024–10035, doi: 10.1021/ja00051a040.
- [63] Harrison J. A., Schall J. D., Maskey S., Mikulski P. T., Knippenberg M. T., Morrow B. H., (2018), “Review of force fields and intermolecular potentials used in atomistic computational materials research”, *Applied Physics Reviews*, 5 (3), 031104, doi: 10.1063/1.5020808.
 - [64] Schütt K. T., Sauceda H. E., Kindermans P.-J., Tkatchenko A., Müller K.-R., (2018), “SchNet – A deep learning architecture for molecules and materials”, *The Journal of Chemical Physics*, 148 (24), 241722, doi: 10.1063/1.5019779.
 - [65] Smith J. S., Isayev O., Roitberg A. E., (2016), “ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost”, *Chemical Science*, 8 (4), 3192–3203, doi: 10.1039/C6SC05720A.
 - [66] Unke O. T., Meuwly M., (2019), “PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges.”, *Journal of chemical theory and computation*, 15 (6), 3678–3693, doi: 10.1021/acs.jctc.9b00181.
 - [67] Behler J., Parrinello M., (2007), “Generalized neural-network representation of high-dimensional potential-energy surfaces”, *Physical Review Letters*, 98 (14), 1–4, doi: 10.1103/PhysRevLett.98.146401.
 - [68] Bartók A. P., Payne M. C., Kondor R., Csányi G., (2010), “Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons”, *Physical Review Letters*, 104 (13), 136403, doi: 10.1103/PhysRevLett.104.136403.
 - [69] Shapeev A. V., (2016), “Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials”, *Multiscale Modeling & Simulation*, 14 (3), 1153–1173, doi: 10.1137/15M1054183.
 - [70] Thompson A. P., Swiler L. P., Trott C. R., Foiles S. M., Tucker G. J., (2015), “Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials”, *Journal of Computational Physics*, 285 316–330, doi: 10.1016/j.jcp.2014.12.018.
 - [71] Drautz R., (2019), “Atomic cluster expansion for accurate and transferable interatomic potentials”, *Physical Review B*, 99 (1), 014104, doi: 10.1103/PhysRevB.99.014104.
 - [72] Balabin R. M., Lomakina E. I., (2011), “Support vector machine regression (LS-SVM)—an alternative to artificial neural networks (ANNs) for the analysis of quantum chemistry data?”, *Physical Chemistry Chemical Physics*, 13 (24), 11710, doi: 10.1039/c1cp00051a.
 - [73] LeCun Y., Boser B., Denker J. S., Henderson D., Howard R. E., Hubbard W., Jackel L. D., (1989), “Backpropagation Applied to Handwritten Zip Code Recognition”, *Neural Computation*, 1 (4), 541–551, doi: 10.1162/neco.1989.1.4.541.

- [74] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I., “Attention is all you need,” in *Advances in Neural Information Processing Systems*, 2017.
- [75] Marcham F., (2016), “Tensorflow: LArge-scale Machine Learning on Heterogeneous Distributed Systems”, *The Library*, s4-X (3), 265–283.
- [76] Paszke A. *et al.*, “Automatic Differentiation in PyTorch,” in *NIPS 2017 Workshop on Autodiff*, 2017. [Online]. Available: <https://openreview.net/forum?id=BJJsrnfCZ>
- [77] Brown D. F. R., Gibbs M. N., Clary D. C., (1996), “Combining *ab initio* computations, neural networks, and diffusion Monte Carlo: An efficient method to treat weakly bound molecules”, *The Journal of Chemical Physics*, 105 (17), 7597–7604, doi: 10.1063/1.472596.
- [78] Tafeit E., Estelberger W., Horejsi R., Moeller R., Oettl K., Vrecko K., Reibnegger G., (1996), “Neural networks as a tool for compact representation of *ab initio* molecular potential energy surfaces”, *Journal of Molecular Graphics*, 14 (1), 12–18, doi: 10.1016/0263-7855(95)00087-9.
- [79] Tai No K., Ha Chang B., Yeon Kim S., Shik Jhon M., Scheraga H. A., (1997), “Description of the potential energy surface of the water dimer with an artificial neural network”, *Chemical Physics Letters*, 271 (1–3), 152–156, doi: 10.1016/S0009-2614(97)00448-X.
- [80] Manzhos S., Carrington T., (2006), “A random-sampling high dimensional model representation neural network for building potential energy surfaces”, *The Journal of Chemical Physics*, 125 (8), 084109, doi: 10.1063/1.2336223.
- [81] Malshe M., Narulkar R., Raff L. M., Hagan M., Bukkapatnam S., Agrawal P. M., Komanduri R., (2009), “Development of generalized potential-energy surfaces using many-body expansions, neural networks, and moiety energy approximations”, *The Journal of Chemical Physics*, 130 (18), 184102, doi: 10.1063/1.3124802.
- [82] Behler J., (2011), “Atom-centered symmetry functions for constructing high-dimensional neural network potentials”, *Journal of Chemical Physics*, 134 (7), 1–13, doi: 10.1063/1.3553717.
- [83] Artrith N., Urban A., Ceder G., (2017), “Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species”, *Physical Review B*, 96 (1), 014112, doi: 10.1103/PhysRevB.96.014112.
- [84] Khorshidi A., Peterson A. A., (2016), “Amp: A modular approach to machine learning in atomistic simulations”, *Computer Physics Communications*, 207 310–324, doi: 10.1016/j.cpc.2016.05.010.
- [85] Yao K., Herr J. E., Toth D. W., McKintyre R., Parkhill J., (2018), “The TensorMol-0.1 model chemistry: A neural network augmented with long-range physics”, *Chemical Science*, 9 (8), 2261–2269, doi: 10.1039/c7sc04934j.

- [86] Schütt K. T., Arbabzadah F., Chmiela S., Müller K. R., Tkatchenko A., (2017), “Quantum-chemical insights from deep tensor neural networks”, *Nature Communications*, 8 6–13, doi: 10.1038/ncomms13890.
- [87] Duvenaud D., Maclaurin D., Aguilera-Iparraguirre J., Gómez-Bombarelli R., Hirzel T., Aspuru-Guzik A., Adams R. P., (2015), “Convolutional networks on graphs for learning molecular fingerprints”, *Advances in Neural Information Processing Systems*, 2015-Janua 2224–2232.
- [88] Gilmer J., Schoenholz S. S., Riley P. F., Vinyals O., Dahl G. E., (2017), “Neural message passing for quantum chemistry”, *34th International Conference on Machine Learning, ICML 2017*, 3 2053–2070.
- [89] Behler J., (2015), “Constructing high-dimensional neural network potentials: A tutorial review”, *International Journal of Quantum Chemistry*, 115 (16), 1032–1050, doi: 10.1002/qua.24890.
- [90] Schütt K. T., Kindermans P. J., Sauceda H. E., Chmiela S., Tkatchenko A., Müller K. R., (2017), “SchNet: A continuous-filter convolutional neural network for modeling quantum interactions”, *Advances in Neural Information Processing Systems*, 2017-Decem (July), 992–1002.
- [91] Scarselli F., Gori M., Tsoi A. C., Hagenbuchner M., Monfardini G., (2009), “The Graph Neural Network Model”, *IEEE Transactions on Neural Networks*, 20 (1), 61–80, doi: 10.1109/TNN.2008.2005605.
- [92] Lubbers N., Smith J. S., Barros K., (2018), “Hierarchical modeling of molecular energies using a deep neural network”, *The Journal of Chemical Physics*, 148 (24), 241715, doi: 10.1063/1.5011181.
- [93] Willems T. F., Rycroft C. H., Kazi M., Meza J. C., Haranczyk M., (2012), “Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials”, *Microporous and Mesoporous Materials*, 149 (1), 134–141, doi: 10.1016/j.micromeso.2011.08.020.
- [94] Ongari D., Boyd P. G., Barthel S., Witman M., Haranczyk M., Smit B., (2017), “Accurate Characterization of the Pore Volume in Microporous Crystalline Materials”, *Langmuir*, 33 (51), 14529–14538, doi: 10.1021/acs.langmuir.7b01682.
- [95] Grochala W., Edwards P. P., (2004), “Thermal Decomposition of the Non-Interstitial Hydrides for the Storage and Production of Hydrogen”, *Chemical Reviews*, 104 (3), 1283–1316, doi: 10.1021/cr030691s.
- [96] Xiong Z. *et al.*, (2008), “High-capacity hydrogen storage in lithium and sodium amidoboranes”, *Nature Materials*, 7 (2), 138–141, doi: 10.1038/nmat2081.
- [97] Ahmed A., Liu Y., Purewal J., Tran L. D., Wong-Foy A. G., Veenstra M., Matzger A. J., Siegel D. J., (2017), “Balancing gravimetric and volumetric hydrogen density in MOFs”, *Energy & Environmental Science*, 10 (11), 2459–2471, doi: 10.1039/C7EE02477K.

- [98] Li H., Eddaoudi M., O’Keeffe M., Yaghi O. M., (1999), “Design and synthesis of an exceptionally stable and highly porous metal-organic framework”, *Nature*, 402 (6759), 276–279, doi: 10.1038/46248.
- [99] Suh M. P., Park H. J., Prasad T. K., Lim D.-W., (2012), “Hydrogen Storage in Metal–Organic Frameworks”, *Chemical Reviews*, 112 (2), 782–835, doi: 10.1021/cr200274s.
- [100] Rosi N. L., Eckert J., Eddaoudi M., Vodak D. T., Kim J., O’Keeffe M., Yaghi O. M., (2003), “Hydrogen Storage in Microporous Metal-Organic Frameworks”, *Science*, 300 1127–1130.
- [101] Li J.-R., Kuppler R. J., Zhou H.-C., (2009), “Selective gas adsorption and separation in metal–organic frameworks”, *Chemical Society Reviews*, 38 (5), 1477, doi: 10.1039/b802426j.
- [102] Gómez-Gualdrón D. A., Moghadam P. Z., Hupp J. T., Farha O. K., Snurr R. Q., (2016), “Application of Consistency Criteria To Calculate BET Areas of Micro- And Mesoporous Metal–Organic Frameworks”, *Journal of the American Chemical Society*, 138 (1), 215–224, doi: 10.1021/jacs.5b10266.
- [103] Gallo M., Glossman-Mitnik D., (2009), “Fuel gas storage and separations by Metal-organic frameworks: Simulated adsorption isotherms for h₂ and ch₄ and their equimolar mixture”, *Journal of Physical Chemistry C*, 113 (16), 6634–6642, doi: 10.1021/jp809539w.
- [104] Yu J., Xie L. H., Li J. R., Ma Y., Seminario J. M., Balbuena P. B., (2017), “CO₂ Capture and Separations Using MOFs: Computational and Experimental Studies”, *Chemical Reviews*, 117 (14), 9674–9754, doi: 10.1021/acs.chemrev.6b00626.
- [105] Dawson R., Adams D. J., Cooper A. I., (2011), “Chemical tuning of CO₂ sorption in robust nanoporous organic polymers”, *Chemical Science*, 2 (6), 1173, doi: 10.1039/c1sc00100k.
- [106] Jiang J., Babarao R., Hu Z., (2011), “Molecular simulations for energy, environmental and pharmaceutical applications of nanoporous materials: from zeolites, metal–organic frameworks to protein crystals”, *Chemical Society Reviews*, 40 (7), 3599, doi: 10.1039/c0cs00128g.
- [107] Jiang J., (2014), “Molecular simulations in metal–organic frameworks for diverse potential applications”, *Molecular Simulation*, 40 (7–9), 516–536, doi: 10.1080/08927022.2013.832247.
- [108] Peyman Z. Moghadam, Aurelia Li, Seth B. Wiggin, Andi Tao, Andrew G. P. Maloney, Peter A. Wood, Suzanna C. Ward, David Fairen-Jimenez*, (2017), “Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future”, *Chemistry of Materials*, 29 (7), 2618–2625, doi: 10.1021/acs.chemmater.7b00441.
- [109] Goldsmith J., Wong-Foy A. G., Cafarella M. J., Siegel D. J., (2013), “Theoretical limits of hydrogen storage in metal-organic frameworks:

- Opportunities and trade-offs”, *Chemistry of Materials*, 25 (16), 3373–3382, doi: 10.1021/cm401978e.
- [110] Altintas C., Avci G., Daglar H., Nemati Vesali Azar A., Erucar I., Velioglu S., Keskin S., (2019), “An extensive comparative analysis of two MOF databases: high-throughput screening of computation-ready MOFs for CH₄ and H₂ adsorption”, *Journal of Materials Chemistry A*, 7 (16), 9593–9608, doi: 10.1039/c9ta01378d.
- [111] Chung Y. G. *et al.*, (2019), “Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019”, *Journal of Chemical & Engineering Data*, 64 (12), 5985–5998, doi: 10.1021/acs.jced.9b00835.
- [112] Ahmed A., Seth S., Purewal J., Wong-Foy A. G., Veenstra M., Matzger A. J., Siegel D. J., (2019), “Exceptional hydrogen storage achieved by screening nearly half a million metal-organic frameworks”, *Nature Communications*, 10 (1), 1568, doi: 10.1038/s41467-019-09365-w.
- [113] Tshitoyan V., Dagdelen J., Weston L., Dunn A., Rong Z., Kononova O., Persson K. A., Ceder G., Jain A., (2019), “Unsupervised word embeddings capture latent knowledge from materials science literature”, *Nature*, 571 (7763), 95–98, doi: 10.1038/s41586-019-1335-8.
- [114] Panella B., Hirscher M., Roth S., (2005), “Hydrogen adsorption in different carbon nanostructures”, *Carbon*, 43 (10), 2209–2214, doi: 10.1016/j.carbon.2005.03.037.
- [115] Kolotilov S. V, Pavlishchuk V. V, (2009), “Effect of structural and thermodynamic factors on the sorption of hydrogen by metal–organic framework compounds”, *Theoretical and Experimental Chemistry*, 45 (2), 75–97, doi: 10.1007/s11237-009-9068-7.
- [116] Farha O. K. *et al.*, (2012), “Metal–Organic Framework Materials with Ultrahigh Surface Areas: Is the Sky the Limit?”, *Journal of the American Chemical Society*, 134 (36), 15016–15021, doi: 10.1021/ja3055639.
- [117] Tukey J. W., *Exploratory Data Analysis*. 1977.
- [118] Groom C. R., Bruno I. J., Lightfoot M. P., Ward S. C., (2016), “The Cambridge Structural Database”, *Acta Crystallographica Section B*, 72 (2), 171–179, doi: 10.1107/S2052520616003954.
- [119] Wieme J., Van Speybroeck V., (2021), “Unravelling thermal stress due to thermal expansion mismatch in metal–organic frameworks for methane storage”, *Journal of Materials Chemistry A*, 9 (8), 4898–4906, doi: 10.1039/D0TA09462E.
- [120] Garai B., Bon V., Efimova A., Gerlach M., Senkovska I., Kaskel S., (2020), “Reversible switching between positive and negative thermal expansion in a metal–organic framework DUT-49”, *Journal of Materials Chemistry A*, 8 (39), 20420–20428, doi: 10.1039/D0TA06830F.

- [121] Schneider C., Bodesheim D., Ehrenreich M. G., Crocellà V., Mink J., Fischer R. A., Butler K. T., Kieslich G., (2019), “Tuning the Negative Thermal Expansion Behavior of the Metal-Organic Framework Cu(3)BTC(2) by Retrofitting.”, *Journal of the American Chemical Society*, 141 (26), 10504–10509, doi: 10.1021/jacs.9b04755.
- [122] Grobler I., Smith V. J., Bhatt P. M., Herbert S. A., Barbour L. J., (2013), “Tunable anisotropic thermal expansion of a porous zinc(II) metal-organic framework.”, *Journal of the American Chemical Society*, 135 (17), 6411–6414, doi: 10.1021/ja401671p.
- [123] Zheng B., Tian D., Zhang L., Zheng X., Liu C., Lu H., Chen J., Maurin G., Shi Q., (2019), “Investigation of Methane Adsorption in Strained IRMOF-1”, *The Journal of Physical Chemistry C*, 123 (40), 24592–24597, doi: 10.1021/acs.jpcc.9b06960.
- [124] Eckhoff M., Behler J., (2019), “From Molecular Fragments to the Bulk: Development of a Neural Network Potential for MOF-5”, *Journal of Chemical Theory and Computation*, 15 (6), 3793–3809, doi: 10.1021/acs.jctc.8b01288.
- [125] Behler J., (2011), “Neural network potential-energy surfaces in chemistry: A tool for large-scale simulations”, *Physical Chemistry Chemical Physics*, 13 (40), 17930–17955, doi: 10.1039/c1cp21668f.
- [126] Behler J., (2017), “First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems”, *Angewandte Chemie - International Edition*, 56 (42), 12828–12840, doi: 10.1002/anie.201703114.
- [127] Behler J., (2014), “Representing potential energy surfaces by high-dimensional neural network potentials”, *Journal of Physics Condensed Matter*, 26 (18), 183001, doi: 10.1088/0953-8984/26/18/183001.
- [128] Smith J. S., Isayev O., Roitberg A. E., (2017), “ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost”, *Chemical Science*, 8 (4), 3192–3203, doi: 10.1039/C6SC05720A.
- [129] Artrith N., Urban A., (2016), “An implementation of artificial neural-network potentials for atomistic materials simulations: Performance for TiO₂”, *Computational Materials Science*, 114 135–150, doi: 10.1016/j.commatsci.2015.11.047.
- [130] Kocer E., Mason J. K., Erturk H., (2019), “A novel approach to describe chemical environments in high-dimensional neural network potentials”, *Journal of Chemical Physics*, 150 (15), 1–9, doi: 10.1063/1.5086167.
- [131] Zuo Y. *et al.*, (2020), “Performance and Cost Assessment of Machine Learning Interatomic Potentials”, *Journal of Physical Chemistry A*, 124 (4), 731–745, doi: 10.1021/acs.jpca.9b08723.

- [132] Schuett K. T. *et al.*, (2018), “SchNet - A deep learning architecture for molecules and materials”, *J. Chem. Phys.*, 148 (24), 241722, doi: 10.1063/1.5019779.
- [133] Gao X., Ramezanghorbani F., Isayev O., Smith J. S., Roitberg A. E., (2020), “TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials”, *Journal of Chemical Information and Modeling*, 60 (7), 3408–3415, doi: 10.1021/acs.jcim.0c00451.
- [134] Behler J., (2014), “Representing potential energy surfaces by high-dimensional neural network potentials”, *Journal of Physics Condensed Matter*, 26 (18), , doi: 10.1088/0953-8984/26/18/183001.
- [135] Eckhoff M., Behler J., (2019), “From Molecular Fragments to the Bulk: Development of a Neural Network Potential for MOF-5”, *Journal of Chemical Theory and Computation*, 15 (6), 3793–3809, doi: 10.1021/acs.jctc.8b01288.
- [136] Schütt K. T., Kessel P., Gastegger M., Nicoli K. A., Tkatchenko A., Müller K. R., (2019), “SchNetPack: A Deep Learning Toolbox for Atomistic Systems”, *Journal of Chemical Theory and Computation*, 15 (1), 448–455, doi: 10.1021/acs.jctc.8b00908.
- [137] Schütt K. T., Arbabzadah F., Chmiela S., Müller K. R., Tkatchenko A., (2017), “Quantum-chemical insights from deep tensor neural networks”, *Nature Communications*, 8 (1), 13890, doi: 10.1038/ncomms13890.
- [138] Zanca F., Glasby L. T., Chong S., Chen S., Kim J., Fairen-Jimenez D., Monserrat B., Moghadam P. Z., “Computational techniques for characterisation of electrically conductive MOFs: quantum calculations and machine learning approaches,” *Journal of Materials Chemistry C*, 9 (39), . 2021. doi: 10.1039/d1tc02543k.
- [139] Ling S., Slater B., (2015), “Unusually Large Band Gap Changes in Breathing Metal–Organic Framework Materials”, *The Journal of Physical Chemistry C*, 119 (29), 16667–16677, doi: 10.1021/acs.jpcc.5b04050.
- [140] Grimme S., (2011), “Density functional theory with London dispersion corrections”, *WIREs Computational Molecular Science*, 1 (2), 211–228, doi: 10.1002/wcms.30.
- [141] St. John P. C., Guan Y., Kim Y., Etz B. D., Kim S., Paton R. S., (2020), “Quantum chemical calculations for over 200,000 organic radical species and 40,000 associated closed-shell molecules”, *Scientific Data*, 7 (1), 244, doi: 10.1038/s41597-020-00588-x.
- [142] Debeve L. M., Pollock C. J., (2021), “Systematic assessment of DFT methods for geometry optimization of mononuclear platinum-containing complexes”, *Physical Chemistry Chemical Physics*, 23 (43), 24780–24788, doi: 10.1039/D1CP01851E.
- [143] Neese F., (2003), “An improvement of the resolution of the identity approximation for the formation of the Coulomb matrix”, *Journal of*

Computational Chemistry, 24 (14), 1740–1747, doi:
<https://doi.org/10.1002/jcc.10318>.

- [144] Caldeweyher E., Ehlert S., Hansen A., Neugebauer H., Spicher S., Bannwarth C., Grimme S., (2019), “A generally applicable atomic-charge dependent London dispersion correction”, *The Journal of Chemical Physics*, 150 (15), 154122, doi: 10.1063/1.5090222.
- [145] Hjorth Larsen A. *et al.*, (2017), “The atomic simulation environment—a Python library for working with atoms”, *Journal of Physics: Condensed Matter*, 29 (27), 273002, doi: 10.1088/1361-648x/aa680e.
- [146] Behler J., (2015), “Constructing high-dimensional neural network potentials: A tutorial review”, *International Journal of Quantum Chemistry*, 115 (16), 1032–1050, doi: 10.1002/qua.24890.
- [147] Behler J., (2011), “Atom-centered symmetry functions for constructing high-dimensional neural network potentials”, *Journal of Chemical Physics*, 134 (7), 1–13, doi: 10.1063/1.3553717.
- [148] Behler J., (2021), “Four Generations of High-Dimensional Neural Network Potentials”, *Chemical Reviews*, 121 (16), 10037–10072, doi: 10.1021/acs.chemrev.0c00868.
- [149] Liaw R., Liang E., Nishihara R., Moritz P., Gonzalez J. E., Stoica I., (2018), “Tune: A Research Platform for Distributed Model Selection and Training”, doi: 10.48550/arxiv.1807.05118.
- [150] Eddaoudi M., Kim J., Rosi N., Vodak D., Wachter J., O’Keeffe M., Yaghi O. M., (2002), “Systematic design of pore size and functionality in isorecticular MOFs and their application in methane storage”, *Science*, 295 (5554), 469–472, doi: 10.1126/science.1067208.
- [151] Kuc A., Enyashin A., Seifert G., (2007), “Metal–Organic Frameworks: Structural, Energetic, Electronic, and Mechanical Properties”, *The Journal of Physical Chemistry B*, 111 (28), 8179–8186, doi: 10.1021/jp072085x.
- [152] Han S. S., Goddard W. A., (2007), “Metal–Organic Frameworks Provide Large Negative Thermal Expansion Behavior”, *The Journal of Physical Chemistry C*, 111 (42), 15185–15191, doi: 10.1021/jp075389s.
- [153] Banlusan K., Strachan A., (2017), “First-principles study of elastic mechanical responses to applied deformation of metal-organic frameworks”, *The Journal of Chemical Physics*, 146 (18), 184705, doi: 10.1063/1.4982356.
- [154] Sun Y., Sun H., (2014), “An all-atom force field developed for Zn₄O(RCO₂)₆ metal organic frameworks”, *Journal of Molecular Modeling*, 20 (3), 2146, doi: 10.1007/s00894-014-2146-3.

- [155] Bristow J. K., Tiana D., Walsh A., (2014), “Transferable Force Field for Metal–Organic Frameworks from First-Principles: BTW-FF”, *Journal of Chemical Theory and Computation*, 10 (10), 4644–4652, doi: 10.1021/ct500515h.
- [156] Gould S. L., Tranchemontagne D., Yaghi O. M., Garcia-Garibay M. A., (2008), “Amphidynamic Character of Crystalline MOF-5: Rotational Dynamics of Terephthalate Phenylenes in a Free-Volume, Sterically Unhindered Environment”, *J. Am. Chem. Soc.*, 130 3246.
- [157] Yildirim T., Hartman M. R., (2005), “Direct Observation of Hydrogen Adsorption Sites and Nanocage Formation in Metal-Organic Frameworks”, *Physical Review Letters*, 95 (21), 215504, doi: 10.1103/PhysRevLett.95.215504.
- [158] Civalleri B., Napoli F., Noel Y., Roetti C., Dovesi R., Noël Y., Roetti C., Dovesi R., (2006), “Ab-initio prediction of materials properties with CRYSTAL: MOF-5 as a case study”, *CrystEngComm*, 8 (5), 364, doi: 10.1039/b603150c.
- [159] Bristow J. K., Skelton J. M., Svane K. L., Walsh A., Gale J. D., (2016), “A general forcefield for accurate phonon properties of metal-organic frameworks”, *Physical Chemistry Chemical Physics*, 18 (42), 29316–29329, doi: 10.1039/c6cp05106e.
- [160] Amirjalayer S., Schmid R., (2008), “Conformational isomerism in the isorecticular metal organic framework family: A force field investigation”, *Journal of Physical Chemistry C*, 112 (38), 14980–14987, doi: 10.1021/jp8061948.
- [161] Zhou W., Wu H., Yildirim T., Simpson J. R., Walker A. R. H., (2008), “Origin of the exceptional negative thermal expansion in metal-organic framework MOF-5”, *Physical Review B*, 78 (5), 054114, doi: 10.1103/PhysRevB.78.054114.
- [162] Lock N., Wu Y., Christensen M., Cameron L. J., Peterson V. K., Bridgeman A. J., Kepert C. J., Iversen B. B., (2010), “Elucidating Negative Thermal Expansion in MOF-5”, *The Journal of Physical Chemistry C*, 114 (39), 16181–16186, doi: 10.1021/jp103212z.
- [163] Wieme J., Vandenbrande S., Lamaire A., Kapil V., Vanduyfhuys L., Van Speybroeck V., (2019), “Thermal Engineering of Metal–Organic Frameworks for Adsorption Applications: A Molecular Simulation Perspective”, *ACS Applied Materials & Interfaces*, 11 (42), 38697–38707, doi: 10.1021/acsami.9b12533.
- [164] Dubbeldam D., Walton K. S., Ellis D. E., Snurr R. Q., (2007), “Exceptional Negative Thermal Expansion in Isorecticular Metal–Organic Frameworks”, *Angewandte Chemie International Edition*, 46 (24), 4496–4499, doi: <https://doi.org/10.1002/anie.200700218>.
- [165] Ryder M. R., Maul J., Civalleri B., Erba A., (2019), “Quasi-Harmonic Lattice Dynamics of a Prototypical Metal–Organic Framework”, *Advanced Theory and Simulations*, 2 (11), 1900093, doi: <https://doi.org/10.1002/adts.201900093>.

- [166] Greathouse J. A., Allendorf M. D., (2008), “Force Field Validation for Molecular Dynamics Simulations of IRMOF-1 and Other Isoreticular Zinc Carboxylate Coordination Polymers”, *The Journal of Physical Chemistry C*, 112 (15), 5795–5802, doi: 10.1021/jp076853w.
- [167] Boyd P. G., Moosavi S. M., Witman M., Smit B., (2017), “Force-Field Prediction of Materials Properties in Metal-Organic Frameworks”, *Journal of Physical Chemistry Letters*, 8 (2), 357–363, doi: 10.1021/acs.jpclett.6b02532.
- [168] Lock N., Wu Y., Christensen M., Cameron L. J., Peterson V. K., Bridgeman A. J., Kepert C. J., Iversen B. B., (2010), “Elucidating Negative Thermal Expansion in MOF-5”, *J. Phys. Chem. C*, 114 16181.
- [169] Wieme J., Van Speybroeck V., (2021), “Unravelling thermal stress due to thermal expansion mismatch in metal–organic frameworks for methane storage”, *Journal of Materials Chemistry A*, 9 (8), 4898–4906, doi: 10.1039/D0TA09462E.
- [170] Lee J. H., Jeoung S., Chung Y. G., Moon H. R., (2019), “Elucidation of flexible metal-organic frameworks: Research progresses and recent developments”, *Coordination Chemistry Reviews*, 389 161–188, doi: 10.1016/j.ccr.2019.03.008.
- [171] Lee S., Kim B., Cho H., Lee H., Lee S. Y., Cho E. S., Kim J., (2021), “Computational Screening of Trillions of Metal-Organic Frameworks for High-Performance Methane Storage.”, *ACS applied materials & interfaces*, 13 (20), 23647–23654, doi: 10.1021/acsami.1c02471.
- [172] Gao M., Huang R.-K., Zheng B., Wang P., Shi Q., Zhang W.-X., Dong J., (2022), “Large breathing effect in ZIF-65(Zn) with expansion and contraction of the SOD cage”, *Nature Communications*, 13 (1), 4569, doi: 10.1038/s41467-022-32332-x.
- [173] Schneemann A., Bon V., Schwedler I., Senkovska I., Kaskel S., Fischer R. A., (2014), “Flexible metal-organic frameworks”, *Chemical Society Reviews*, 43 (16), 6062–6096, doi: 10.1039/c4cs00101j.
- [174] Achar S. K., Wardzala J. J., Bernasconi L., Zhang L., Johnson J. K., (2022), “Combined Deep Learning and Classical Potential Approach for Modeling Diffusion in UiO-66”, *Journal of Chemical Theory and Computation*, 18 (6), 3593–3606, doi: 10.1021/acs.jctc.2c00010.
- [175] Vanduyfhuys L., Vandenbrande S., Verstraelen T., Schmid R., Waroquier M., Van Speybroeck V., (2015), “QuickFF: A program for a quick and easy derivation of force fields for metal-organic frameworks from ab initio input”, *Journal of Computational Chemistry*, 36 (13), 1015–1027, doi: 10.1002/jcc.23877.
- [176] Vanduyfhuys L., Vandenbrande S., Wieme J., Waroquier M., Verstraelen T., Van Speybroeck V., (2018), “Extension of the QuickFF force field protocol for an improved accuracy of structural, vibrational, mechanical and thermal

properties of metal-organic frameworks”, *Journal of Computational Chemistry*, 39 (16), 999–1011, doi: 10.1002/jcc.25173.

- [177] Dürholt J. P., Fraux G., Coudert F. X., Schmid R., (2019), “Ab Initio Derived Force Fields for Zeolitic Imidazolate Frameworks: MOF-FF for ZIFs”, *Journal of Chemical Theory and Computation*, 15 (4), 2420–2432, doi: 10.1021/acs.jctc.8b01041.
- [178] E. Coupry D., A. Addicoat M., Heine T., (2016), “Extension of the Universal Force Field for Metal–Organic Frameworks”, *Journal of Chemical Theory and Computation*, 12 (10), 5215–5225, doi: 10.1021/acs.jctc.6b00664.
- [179] Bristow J. K., Tiana D., Walsh A., (2014), “Transferable force field for metal-organic frameworks from first-principles: BTW-FF”, *Journal of Chemical Theory and Computation*, 10 (10), 4644–4652, doi: 10.1021/ct500515h.
- [180] Weng T., Schmidt J. R., (2019), “Flexible and Transferable ab Initio Force Field for Zeolitic Imidazolate Frameworks: ZIF-FF”, *Journal of Physical Chemistry A*, 123 (13), 3000–3012, doi: 10.1021/acs.jpca.8b12311.
- [181] Ryder M. R., Maul J., Civalleri B., Erba A., (2019), “Quasi-Harmonic Lattice Dynamics of a Prototypical Metal–Organic Framework”, *Advanced Theory and Simulations*, 2 (11), 1900093, doi: 10.1002/adts.201900093.
- [182] Saiz F., Bernasconi L., (2020), “Density-functional theory models of Fe(4)O reactivity in metal–organic frameworks: self-interaction error, spin delocalisation and the role of hybrid exchange”, *Physical Chemistry Chemical Physics*, 22 (22), 12821–12830, doi: 10.1039/D0CP01285H.
- [183] Liu Y., Liu H., Hu Y., Jiang J., (2010), “Density Functional Theory for Adsorption of Gas Mixtures in Metal–Organic Frameworks”, *The Journal of Physical Chemistry B*, 114 (8), 2820–2827, doi: 10.1021/jp9104932.
- [184] Lee K., Howe J. D., Lin L.-C., Smit B., Neaton J. B., (2015), “Small-Molecule Adsorption in Open-Site Metal–Organic Frameworks: A Systematic Density Functional Theory Study for Rational Design”, *Chemistry of Materials*, 27 (3), 668–678, doi: 10.1021/cm502760q.
- [185] Rupp M., Tkatchenko A., Müller K.-R., von Lilienfeld O. A., (2012), “Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning”, *Physical Review Letters*, 108 (5), 058301, doi: 10.1103/PhysRevLett.108.058301.
- [186] Faraji S., Ghasemi S. A., Rostami S., Rasoulkhani R., Schaefer B., Goedecker S., Amsler M., (2017), “High accuracy and transferability of a neural network potential through charge equilibration for calcium fluoride”, *Physical Review B*, 95 (10), 104105, doi: 10.1103/PhysRevB.95.104105.
- [187] Karimi-Varzaneh H. A., Müller-Plathe F., *Multiscale Molecular Methods in Applied Chemistry*, 307 . in *Topics in Current Chemistry*, vol. 307. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. doi: 10.1007/978-3-642-24968-6.

- [188] Rogge S. M. J., Goeminne R., Demuynck R., Gutiérrez-Sevillano J. J., Vandenbrande S., Vanduyfhuys L., Waroquier M., Verstraelen T., Van Speybroeck V., (2019), “Modeling Gas Adsorption in Flexible Metal–Organic Frameworks via Hybrid Monte Carlo/Molecular Dynamics Schemes”, *Advanced Theory and Simulations*, 2 (4), 1–15, doi: 10.1002/adts.201800177.
- [189] Vanduyfhuys L., Rogge S. M. J., Wieme J., Vandenbrande S., Maurin G., Waroquier M., Van Speybroeck V., (2018), “Thermodynamic insight into stimuli-responsive behaviour of soft porous crystals”, *Nature Communications*, 9 (1), 204, doi: 10.1038/s41467-017-02666-y.
- [190] Rogge S. M. J., Vanduyfhuys L., Ghysels A., Waroquier M., Verstraelen T., Maurin G., Van Speybroeck V., (2015), “A Comparison of Barostats for the Mechanical Characterization of Metal–Organic Frameworks”, *Journal of Chemical Theory and Computation*, 11 (12), 5583–5597, doi: 10.1021/acs.jctc.5b00748.
- [191] Demuynck R., Rogge S. M. J., Vanduyfhuys L., Wieme J., Waroquier M., Van Speybroeck V., (2017), “Efficient Construction of Free Energy Profiles of Breathing Metal–Organic Frameworks Using Advanced Molecular Dynamics Simulations”, *Journal of Chemical Theory and Computation*, 13 (12), 5861–5873, doi: 10.1021/acs.jctc.7b01014.
- [192] Mehlig B., Heermann D. W., Forrest B. M., (1992), “Hybrid Monte Carlo method for condensed-matter systems”, *Physical Review B*, 45 (2), 679–685, doi: 10.1103/PhysRevB.45.679.
- [193] Guo J., Haji-Akbari A., Palmer J. C., (2018), “Hybrid Monte Carlo with LAMMPS”, *Journal of Theoretical and Computational Chemistry*, 17 (03), 1840002, doi: 10.1142/S0219633618400023.
- [194] Chanut N., Ghoufi A., Coulet M.-V., Bourrelly S., Kuchta B., Maurin G., Llewellyn P. L., (2020), “Tailoring the separation properties of flexible metal-organic frameworks using mechanical pressure”, *Nature Communications*, 11 (1), 1216, doi: 10.1038/s41467-020-15036-y.
- [195] Yu Z., Anstine D. M., Boulfelfel S. E., Gu C., Colina C. M., Sholl D. S., (2021), “Incorporating Flexibility Effects into Metal–Organic Framework Adsorption Simulations Using Different Models”, *ACS Applied Materials & Interfaces*, 13 (51), 61305–61315, doi: 10.1021/acsami.1c20583.
- [196] Dubbeldam D., Calero S., Ellis D. E., Snurr R. Q., (2016), “RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials”, *Molecular Simulation*, 42 (2), 81–101, doi: 10.1080/08927022.2015.1010082.
- [197] Martin M. G., Siepmann J. I., (1998), “Transferable potentials for phase equilibria. 1. United-atom description of n-alkanes”, *Journal of Physical Chemistry B*, 102 (14), 2569–2577, doi: 10.1021/jp972543+.

- [198] Plimpton S., (1995), “Fast Parallel Algorithms for Short-Range Molecular Dynamics”, *Journal of Computational Physics*, 117 (1), 1–19, doi: 10.1006/jcph.1995.1039.
- [199] Thompson A. P. *et al.*, (2022), “LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales”, *Computer Physics Communications*, 271 108171, doi: 10.1016/j.cpc.2021.108171.
- [200] Singraber A., Morawietz T., Behler J., Dellago C., (2019), “Parallel Multistream Training of High-Dimensional Neural Network Potentials”, *Journal of Chemical Theory and Computation*, 15 (5), 3075–3092, doi: 10.1021/acs.jctc.8b01092.
- [201] Morawietz T., Singraber A., Dellago C., Behler J., (2016), “How van der waals interactions determine the unique properties of water”, *Proceedings of the National Academy of Sciences of the United States of America*, 113 (30), 8368–8373, doi: 10.1073/pnas.1602375113.
- [202] Chung Y. G. *et al.*, (2014), “Computation-Ready, Experimental Metal–Organic Frameworks: A Tool To Enable High-Throughput Screening of Nanoporous Crystals”, *Chemistry of Materials*, 26 (21), 6185–6192, doi: 10.1021/cm502594j.
- [203] S. Kaye S., Dailly A., M. Yaghi O., R. Long J., (2007), “Impact of Preparation and Handling on the Hydrogen Storage Properties of Zn₄O(1,4-benzenedicarboxylate)₃ (MOF-5)”, *Journal of the American Chemical Society*, 129 (46), 14176–14177, doi: 10.1021/ja076877g.

BIOGRAPHY

Ömer Tayfuroğlu completed his bachelor's degree in 2011 and master's degree in 2016 at Yıldız Technical University, majoring in Chemistry. He began his PhD journey at Gebze Technical University in the Department of Chemistry in 2016. As part of his doctoral studies, He spent six productive months conducting research in the working group of Prof. Dr. Stefan Goedecker at the University of Basel in Switzerland. Alongside his academic pursuits, Ömer cherishes his family life; he is married to Merve Tayfuroğlu and takes pride in being a father to three wonderful daughters, Meryem, Sena, and Esma.

APPENDICES

Appendix A. Publications Related to These Works

O. Tayfuroglu, A. Kocak, and Y. Zorlu, “A neural network potential for the IRMOF series and its application for thermal and mechanical behaviors”, *Phys. Chem. Chem. Phys.*, vol. 24, no. 19, pp. 11882–11897, 2022, doi: 10.1039/d1cp05973d.

O. Tayfuroglu, A. Kocak, and Y. Zorlu, “In Silico Investigation into H₂ Uptake in MOFs: Combined Text/Data Mining and Structural Calculations”, *Langmuir*, vol. 36, no. 1, pp. 119–129, 2019, doi: 10.1021/acs.langmuir

Appendix B. Supporting informatin for chapter 3

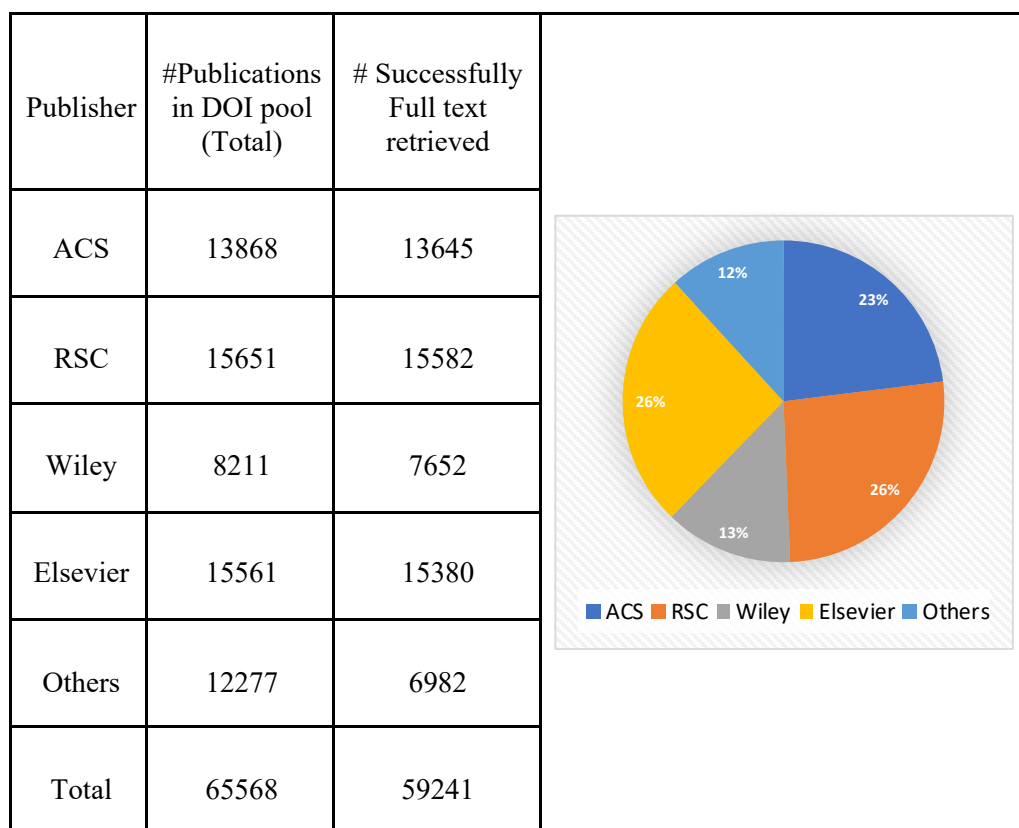


Figure B.1: Manuscript distribution over publishers.

Table B.1: Top 20 MOFs sorted by experimental % wt. H₂ storage capabilities predicted using SA and PV values extracted through TDM algorithm. For the full list, refer to Excel document.

No	MOF_name	SA (m ² /g)	PV (cm ³ /g)	H ₂ wt (%)
1	MOF-L7	7514.00	7.42	24.31
2	NU-110SP	7800.00	4.44	21.49
3	NU-110E	7800.00	4.40	21.44
4	NU-109SP	7560.00	4.12	20.61
5	NU-110SE	7400.00	4.18	20.35
6	NU-109SE	6950.00	3.90	19.08
7	NU-109E	7010.00	3.75	19.03
8	NU-109	7005.00	3.75	19.02
9	[Zn ₄ O(bbc) ₂]	6260.00	3.59	17.27
10	(Zn ₄ O(bte) ₄ /3(bpdc)	6240.00	3.60	17.24
11	NU-110	5778.40	4.35	17.13

Table B.1 continued.

12	DUT-76	6344.00	3.25	17.06
13	PPN-4	6455.92	3.04	17.05
14	COF-108	4885.00	5.45	16.53
15	MOF-L4	5784.00	3.50	16.17
16	MOF-210	5784.45	3.49	16.16
17	Al-soc-MOF-1	5892.50	2.73	15.52
18	DUT-49	5476.00	2.91	14.85
19	NU-1301	4750.00	3.90	14.46
20	COF-102-Ant	6461.00	0.75	14.43

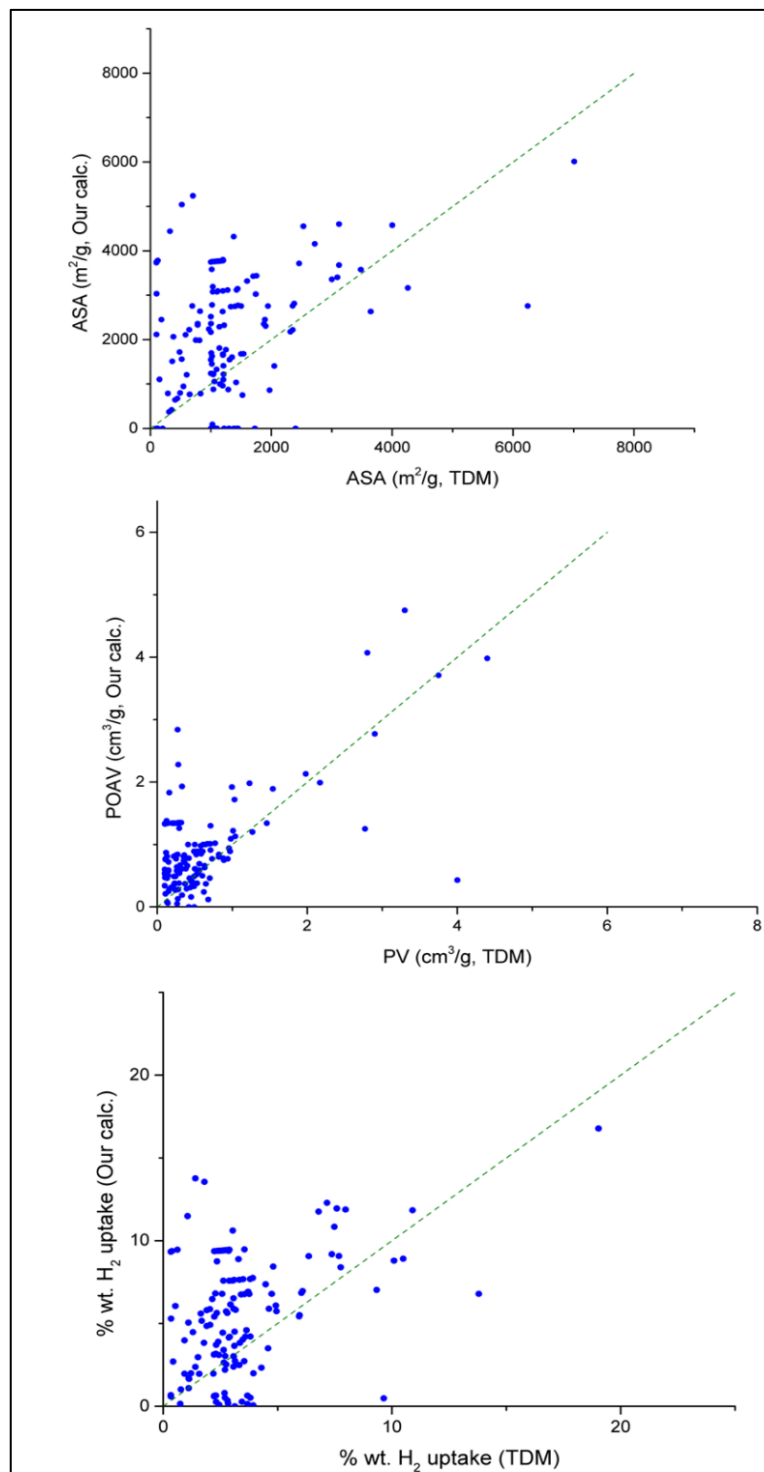


Figure B.2: Comparison of experimental and calculated ASA, POAV and %H₂ uptake values with calculated by Zeo++, GCMC and Method 2 for merge of SA and PV values

Appendix C. Supporting informatin for chapter 4

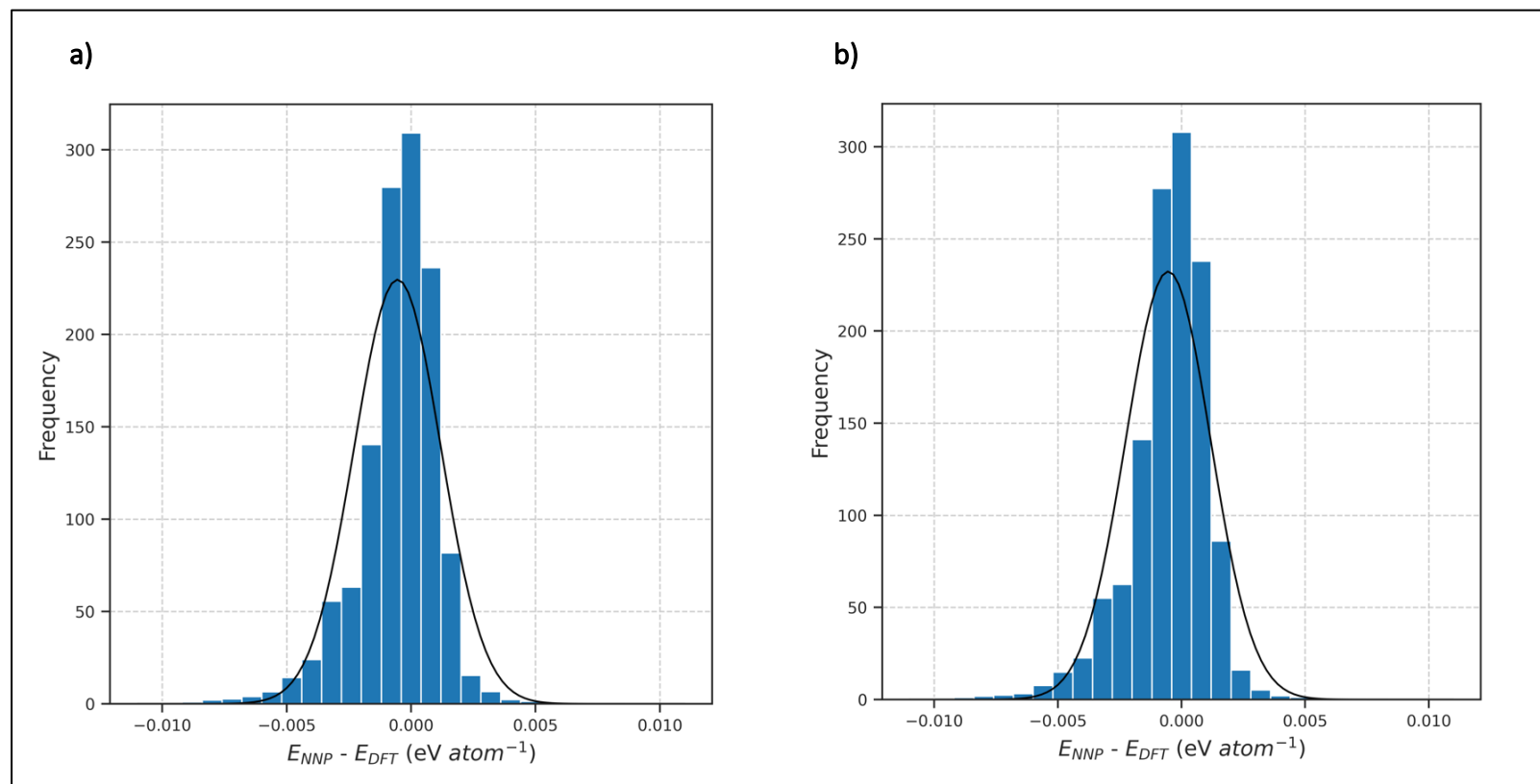


Figure C.1: Histograms of energy errors show normal distribution a-training b-tests.

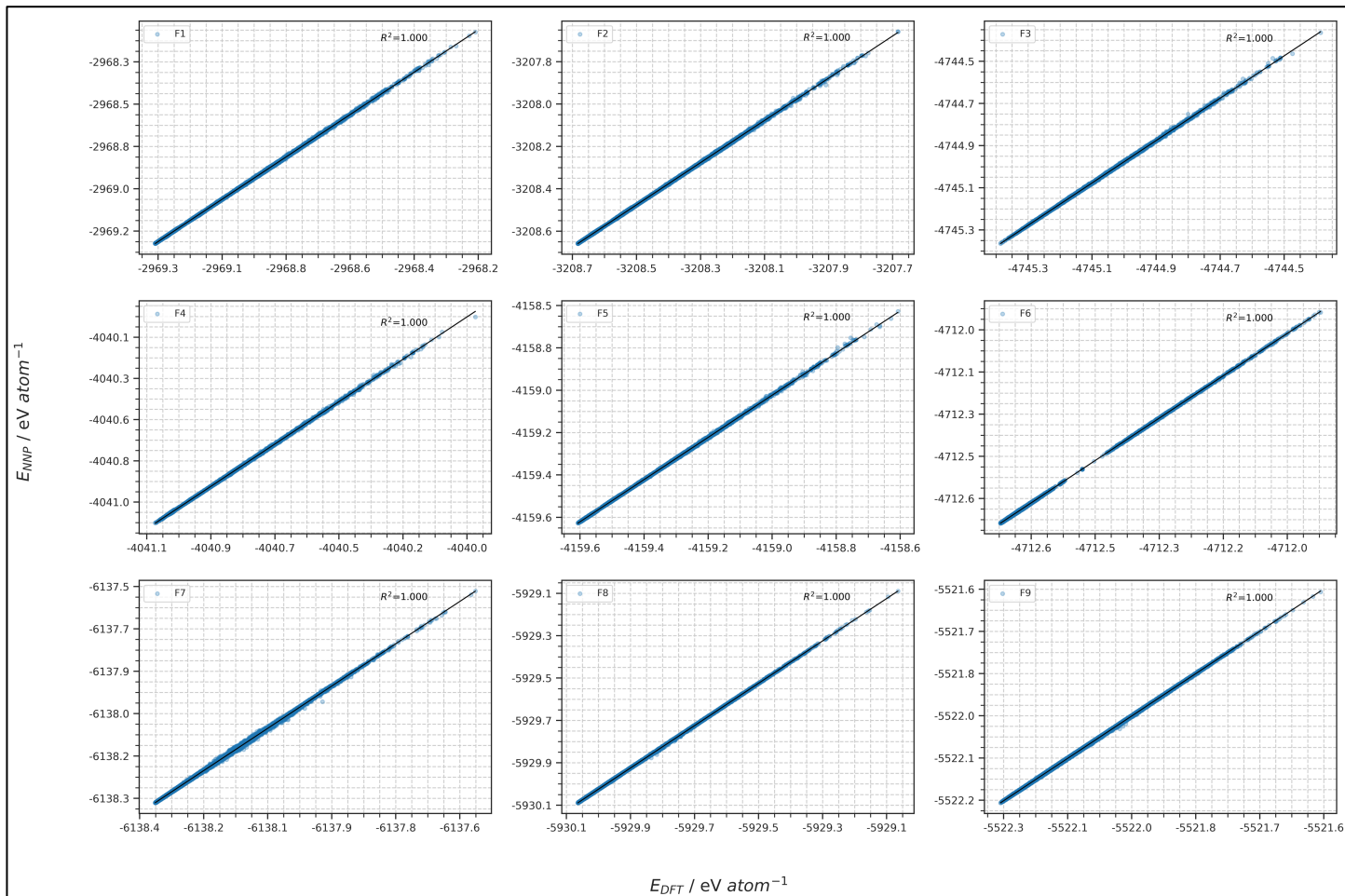


Figure C.2: DFT vs NNP energies of 9 fragments for training set.

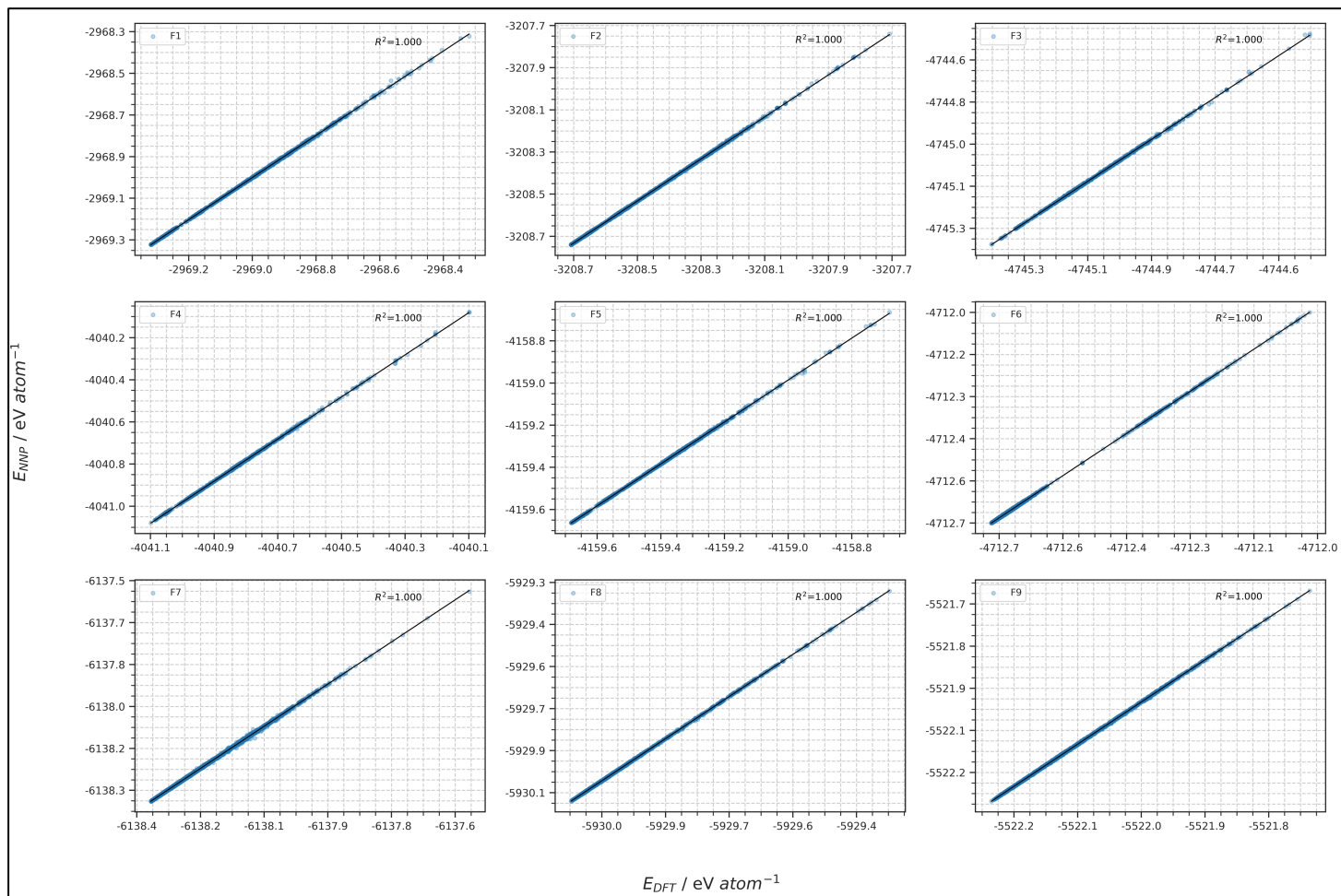


Figure C.3: DFT vs NNP energies of 9 fragments for test set.

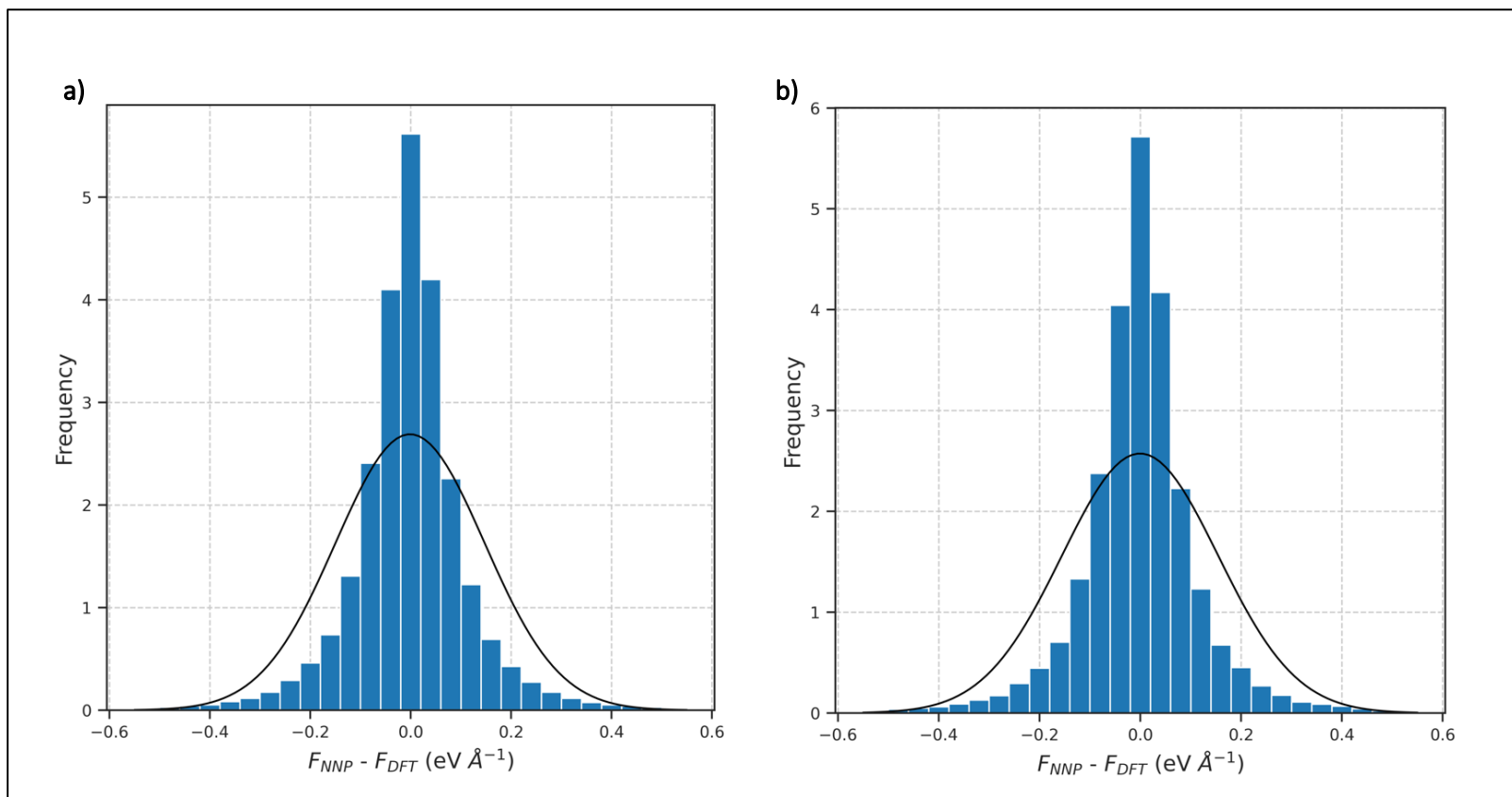


Figure C.4: Histograms of forces errors show normal distribution a) training b) tests.

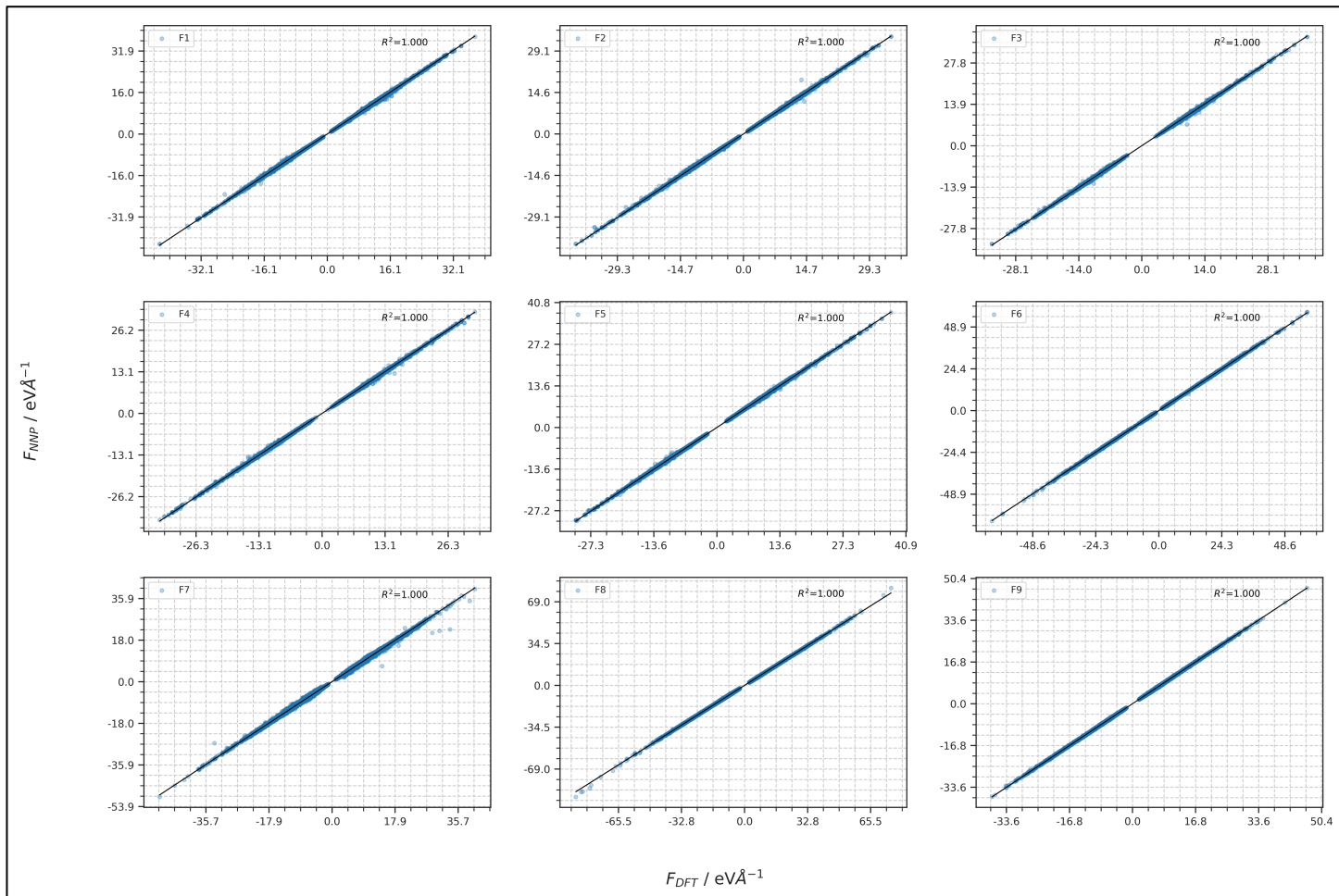


Figure C.5: DFT vs NNP forces of 9 fragments for training set

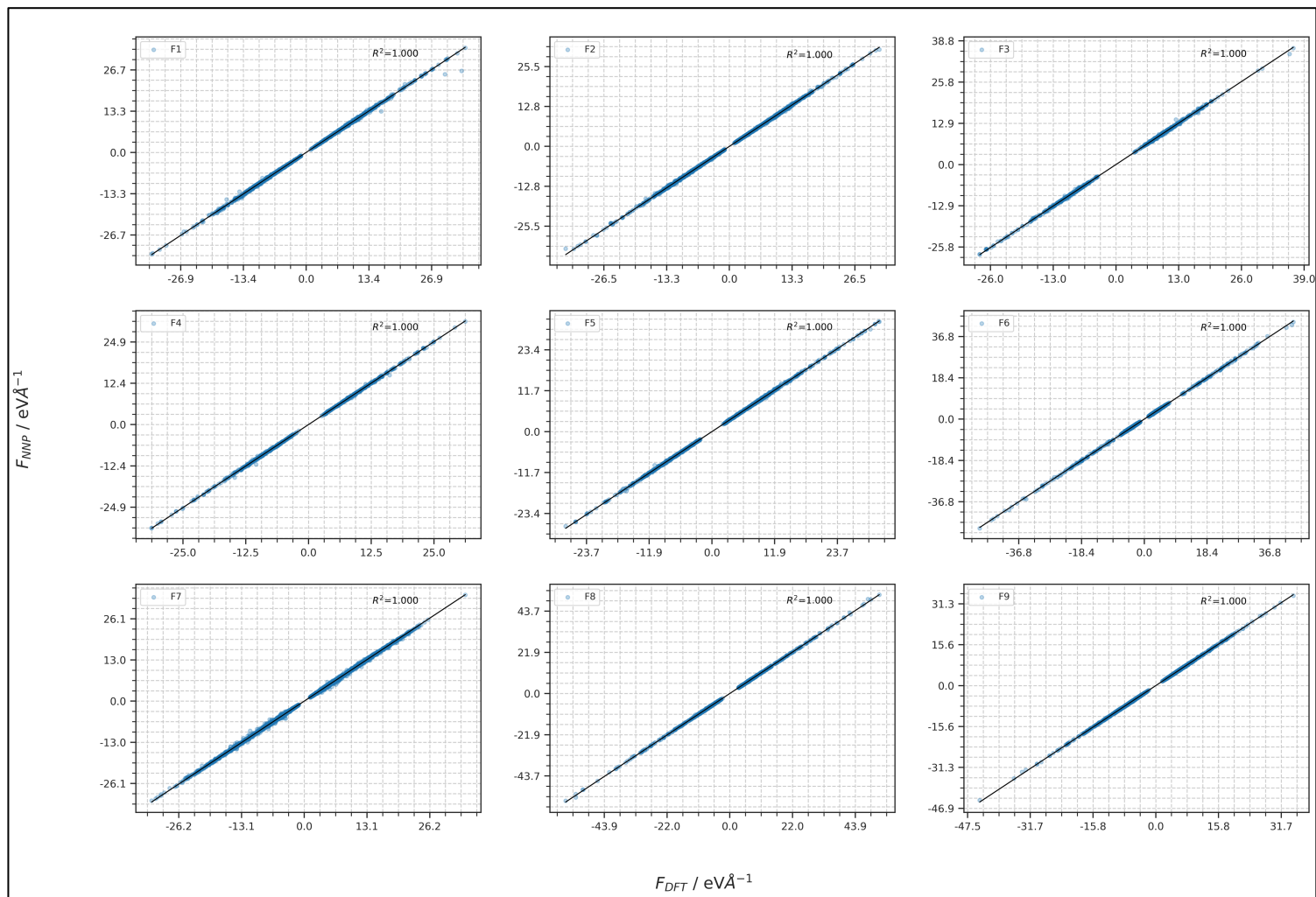


Figure C.6: DFT vs NNP forces of 9 fragments for test set

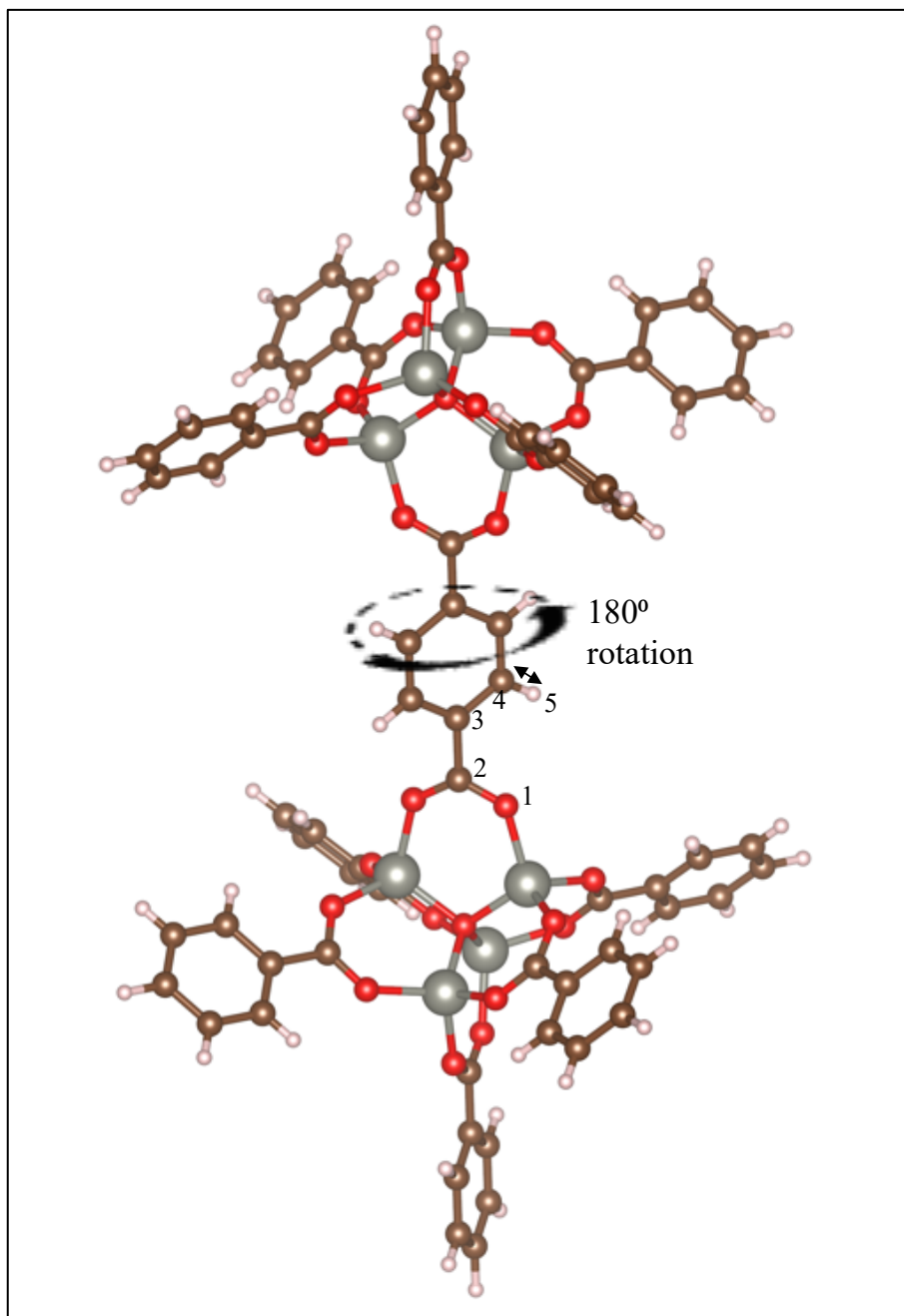


Figure C.7: IRMOF-1 truncated structure for phenylene ring torsion by scan of dihedral angle, $\widehat{\phi_{1234}}$ from 0 to 180° and aromatic C-H distance of $\overline{r_{45}}$ from 1.02Å to 1.16Å.

Table C.1: Thermal expansion coefficient calculations based on quadratic and linear equations. ST shows PTE to NTE switching temperature.

							$\alpha_0=m+nT$		$\alpha_0=\text{constant}$
IRMOF-n	$m \times 10^6$	$n \times 10^8$	Quadratic equation	Linear equation	$ST=-m/n$		200K $\times 10^6$	300K $\times 10^6$	T independent α_0
1	-60.42	4.52	$2.2576E-08T^2 - 6.0422E-05T + 9.7841E+00$	$-4.6876E-05T + 9.7822E+00$		Only NTE	-51.392	-46.876	-46.900
4	662.76	-251.42	$-1.2571E-06T^2 + 6.6276E-04T + 9.6715E+00$	$-9.1506E-05T + 9.7784E+00$	263.606	PTE to NTE	159.920	-91.500	-91.506
6	-25.74	-2.94	$-1.4719E-08T^2 - 2.5735E-05T + 9.7972E+00$	$-3.4566E-05T + 9.7984E+00$		Only NTE	-31.623	-34.566	-34.566
7	1459.31	-512.78	$-2.5639E-06T^2 + 1.4593E-03T + 9.4736E+00$	$-7.9032E-05T + 9.6915E+00$	284.587	PTE to NTE	433.750	-79.030	-79.032
10	-20.12	-10.04	$-5.0190E-08T^2 - 2.0119E-05T + 1.0644E+01$	$-5.0233E-05T + 1.0648E+01$		Only NTE	-40.195	-50.233	-50.233

Table C.2: Equilibrium lattice parameters produced by NNP at different temperatures.

	Lattice constant (a=b=c) in Angstrom					ln(<Volume>)				
T (K)	IRMOF-1	IRMOF-4	IRMOF-6	IRMOF-7	IRMOF-10	IRMOF-1	IRMOF-4	IRMOF-6	IRMOF-7	IRMOF-10
200	25.988	25.820	26.150	25.060	34.674	9.773	9.753	9.792	9.664	10.638
250	25.967	25.881	26.134	25.169	34.645	9.770	9.761	9.790	9.677	10.635
300	25.948	25.824	26.120	25.196	34.617	9.768	9.754	9.788	9.680	10.633
350	25.923	25.808	26.107	25.134	34.596	9.765	9.752	9.787	9.673	10.631
400	25.909	25.660	26.088	24.913	34.554	9.764	9.735	9.784	9.646	10.628

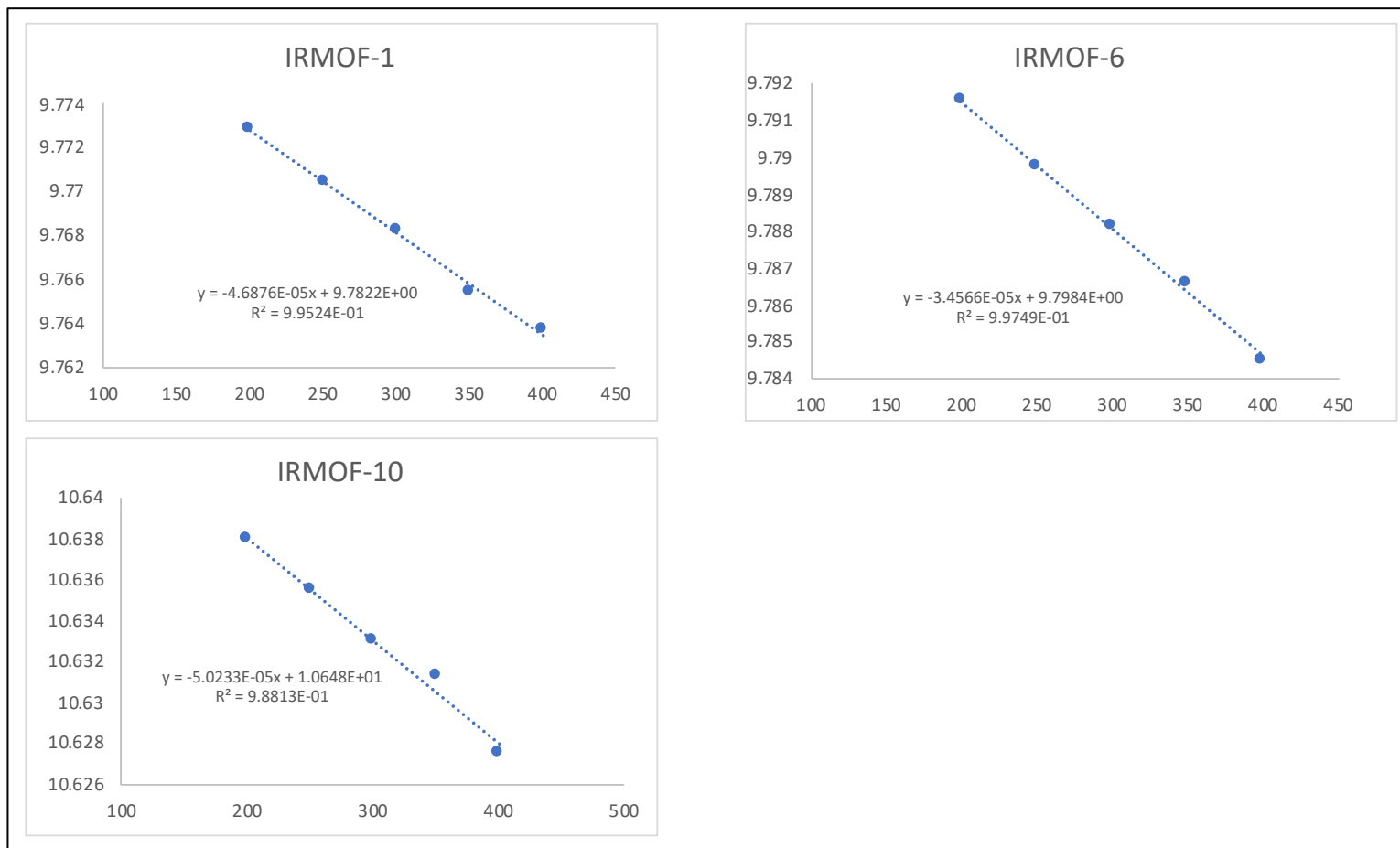


Figure C.8: $\ln(\langle V \rangle)$ vs. T (in Kelvin) plots for thermal expansion coefficients for IRMOF-1, 6 and 10.

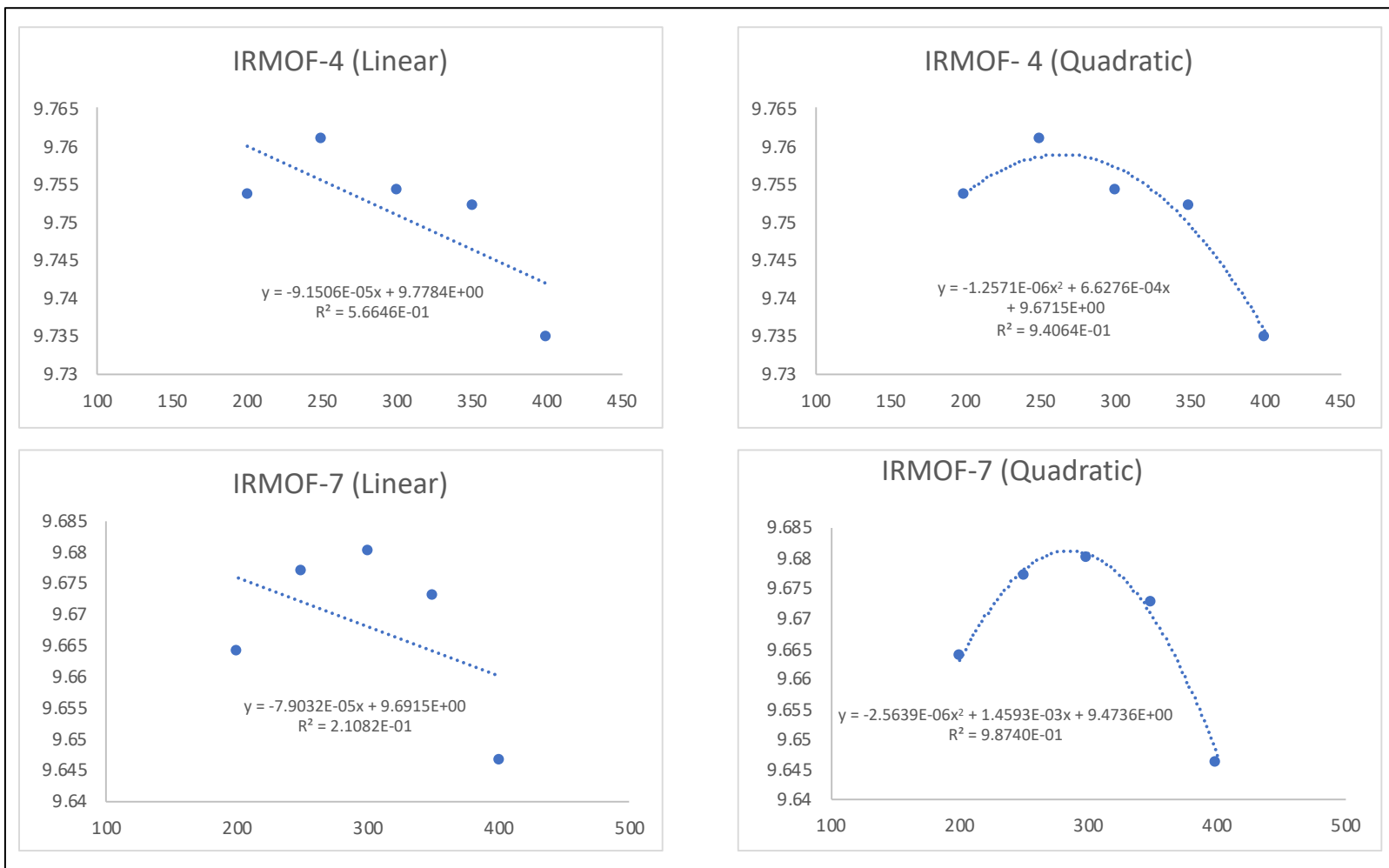


Figure C.9: $\ln(\langle V \rangle)$ vs. T (in Kelvin) plots for thermal expansion coefficients for IRMOF-4 and IRMOF 7.

Appendix D. Supporting informatin for chapter 5

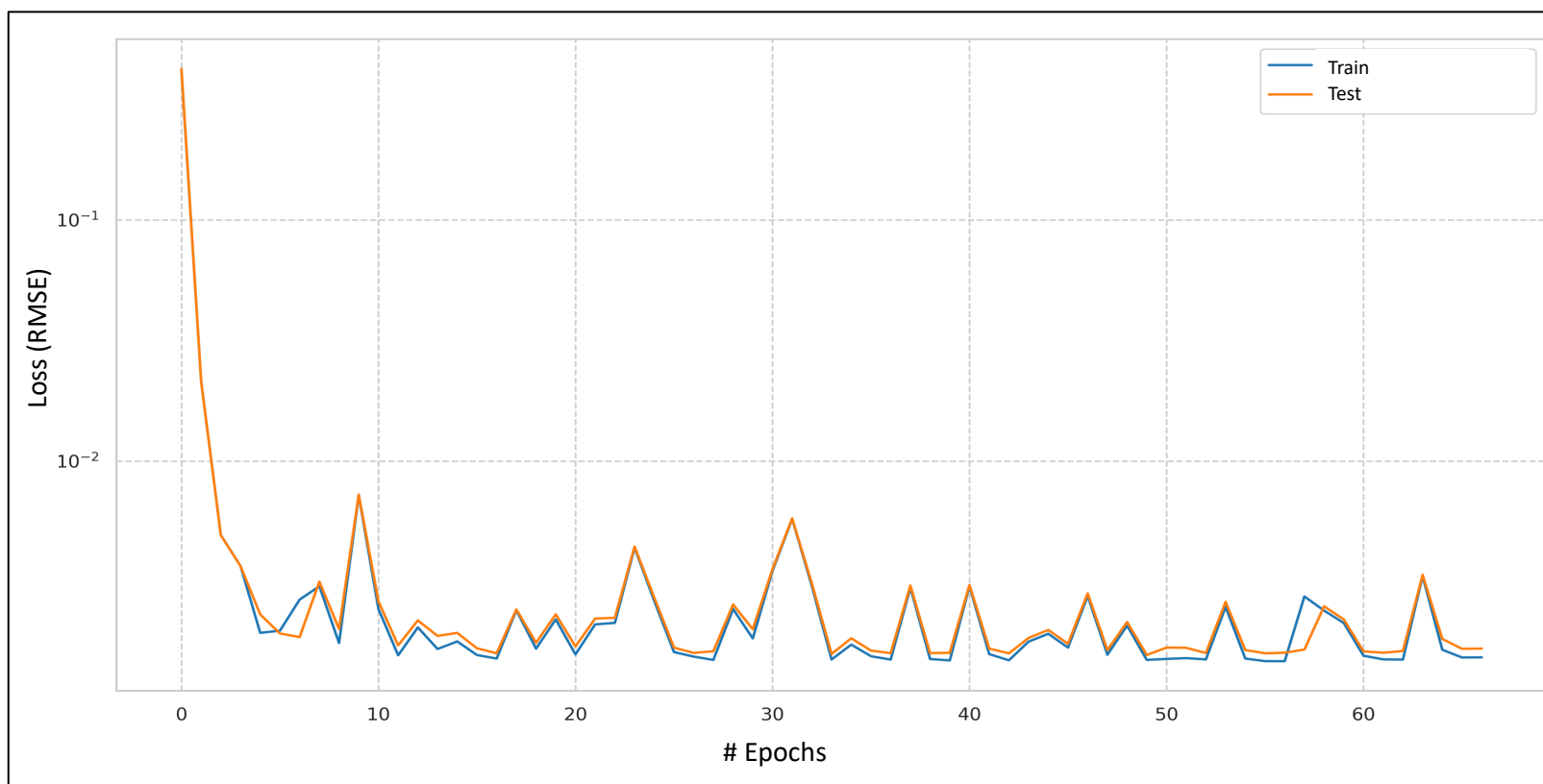


Figure D.1: Training loss curves of single HDNNP for IRMOFs.

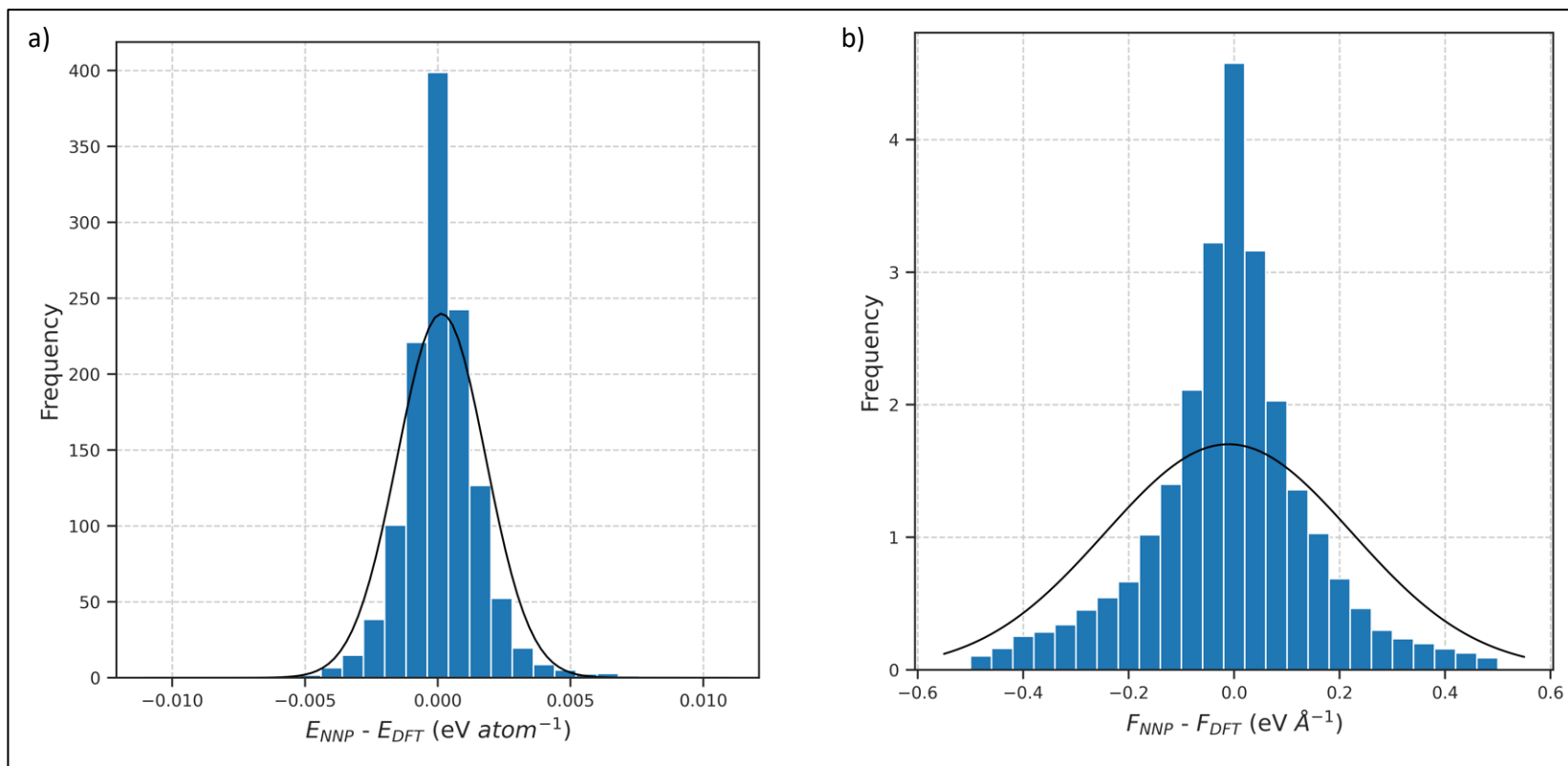


Figure D.2: Histograms of energy errors show normal distribution a) training b) test.