An Automated Method for Extracting Adsorption Binding Sites in MOFs from GCMC Simulations with Direct Comparison to Experimentally Determined Binding Sites

Olivier Marchand

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
Master of Science degree in Chemistry

Department of Chemistry and Biomolecular Sciences
Faculty of Science
University of Ottawa

© Olivier Marchand, Ottawa, Canada, 2026

Table of Contents

Abstract

The location and nature of adsorption binding sites in porous materials is central to understanding selective adsorption of guest molecules at the atomic level. While grand canonical Monte Carlo (GCMC) simulations routinely generate three-dimensional adsorbate probability densities, these data are most often interpreted qualitatively. The absence of a standardized and automated framework for extracting binding sites from adsorption probability distributions (APDs) limits quantification and analysis of binding sites in high-throughput screening workflows.

This thesis contains two related parts. The first is the development of a standalone, robust, and semi-automated workflow, termed GALP, for extracting adsorption binding sites from adsorbate probability densities generated by GCMC. GALP applies smoothing, peak identification, clustering, and molecular fitting to transform three-dimensional probability densities into discrete binding-site coordinates, along with binding-site-specific information such as relative occupancy and binding energetics. The development and validation of GALP establishes a general and transferable framework for systematic binding-site identification.

The second part of the thesis quantitatively evaluates the reliability of common classical simulation approximations used in GCMC, specifically the use of generic force field parameters within the framework, standard charge assignment schemes, and the rigid framework approximation. Simulation-predicted binding sites obtained under these assumptions are directly compared with experimentally determined adsorption sites. This comparison assesses the positional accuracy with which classical simulations reproduce known adsorption motifs, thereby identifying the conditions under which these approximations support meaningful mechanistic interpretation and the cases in which limitations become evident.

Together, these results demonstrate that probability-based binding site extraction provides a consistent and reproducible framework for validating classical simulation methodologies against experiment. By enabling direct, spatially resolved comparison of predicted and experimental adsorption sites, the approach moves beyond global adsorption metrics such as isotherms and uptake values, and instead focuses on the underlying structural features that govern adsorption behaviour. As a result, the methods developed here provide a physically grounded basis for scalable high-throughput analysis of adsorption in porous materials.

Table of Figures

Table of Tables

List of Abbreviations

APD Adsorption Probability Distribution

ABS Absorption Binding Sites

Chargemol DDEC Charge Partitioning Package

CIF Crystallographic Information File

COF Covalent Organic Framework

CSD Cambridge Structural Database

CSV Comma-Separated Values

DDEC6 Density Derived Electrostatic and Chemical Charges (version 6)

DFT Density Functional Theory

DL POLY Classical Molecular Dynamics Simulation Package

ESP Electrostatic Potential

GALA Guest Atom Localization Algorithm (legacy name)

GALP Guest Atom Localizer from Probabilities

GCMC Grand Canonical Monte Carlo

LJ Lennard–Jones

MAC MOF/Adsorbate/Conditions Combinations

MAD Mean Absolute Deviation

MAE Mean Absolute Error

MD Molecular Dynamics

ML Machine Learning

MM Molecular Mechanics

MOF Metal–Organic Framework

NPD Neutron Powder Diffraction

OMS Open-Metal sites

PBC Periodic Boundary Condition

PAW Projector Augmented-Wave Method

PBE Perdew–Burke–Ernzerhof Functional

PES Potential Energy Surface

PXRD Powder X-Ray Diffraction

RASPA Molecular Simulation Suite for Adsorption and Diffusion

REPEAT Repeating Electrostatic Potential Extracted Atomic Charges

RMSD Root Mean Square Deviation

RMSE Root Mean Squared Error

SBU Structural Building Unit

SCXRD Single Crystal X-ray Diffraction

SSNMR Solid-State Nuclear Magnetic Resonance

TraPPE Transferable Potentials for Phase Equilibria

UFF Universal Force Field

VASP Vienna Ab initio Simulation Package

VSA Vacuum Swing Adsorption

XRD X-Ray Diffraction

ZIF Zeolitic Imidazolate Framework

Acknowledgements

This thesis would not have been possible without the support and guidance of many people. I would like to first thank my thesis supervisor, Tom Woo, for his guidance, support, and thoughtful feedback throughout my graduate studies. His input shaped both the direction and the rigour of this, and his perspective consistently helped clarify problems and refine ideas. Over the past few years, his support played a central role in my development as a researcher, and I am very grateful for the time and effort he devoted to this project.

I would also like to thank the research group members, in particular Jake Burner and Marco Gibaldi, for sharing their knowledge and experience. They were always available when I had questions and provided clear, practical advice that helped me progress more efficiently. Their support made a significant difference in what I was able to learn and accomplish during this work.

I am grateful for the good friends I made during my studies, especially Rosa Cicciarella, Jun Luo, Dr. Ohmin Kwon, Dr. Hasnain Sajid, and Sari Warshawsky. The many discussions we shared, which often stretched longer than planned, made the long hours of manual validation more engaging and far more enjoyable.

Finally, I would like to thank my family, especially my parents and aunts, as well as my friends outside of the research group, for their continued support and encouragement throughout this journey. Their patience and understanding were constant sources of motivation, and I would not have been able to complete this work without them.

1. Introduction

Porous materials play a central role in adsorption, separations, catalysis, and gas storage because their internal pore structures determine how molecules distribute and interact within the solid phase. Traditional porous adsorbents such as zeolites, activated carbons, and porous silicas have long been used at an industrial scale, although their fixed chemistries and rigid topologies limit the degree of control over adsorption behaviour[1–3]. Metal–organic frameworks (MOFs) extend this landscape by providing modular, reticular architectures whose pore shapes, sizes, and chemical environments can be systematically programmed. As a result, MOFs offer opportunities to tailor adsorption behaviour at the molecular level[4].

Understanding how gases organize, bind, and diffuse within these porous architectures is fundamental for interpreting adsorption mechanisms, predicting global performance, and guiding the design of materials suited for real operating conditions. This chapter introduces the structural principles of MOFs, reviews their major applications, and outlines why adsorption behaviour must ultimately be understood at the level of discrete binding sites. It then discusses the experimental and computational challenges associated with resolving these sites, motivating the methodological developments that comprise this thesis.

1.1 Metal Organic Frameworks

MOFs are crystalline porous materials constructed by linking metal-based nodes with multidentate organic ligands into extended structures through the principles of reticular chemistry[5]. This synthetic strategy allows metal clusters and organic linkers to assemble into defined network topologies, yielding periodic architectures whose pore sizes, shapes, and chemical environments

can be tuned with near-molecular precision[6]. These structural building units include inorganic secondary building units (SBUs) such as metal-oxide clusters and a wide range of functionalized organic linkers, each contributing geometric and chemical features that define the resulting framework. Unlike traditional porous adsorbents, whose chemistries and architectures are largely fixed by their synthesis routes, MOFs can be systematically modified through linker exchange, metal substitution, functional-group incorporation, and topological selection. This flexibility has established MOFs as one of the most versatile families of porous materials developed in the past two decades[7–9]. In 2025, the Nobel Prize in Chemistry was awarded to Susumu Kitagawa, Richard Robson, and Omar M. Yaghi for their foundational contribution to the design and synthesis of MOFs, highlighting the broad impact and tunability of this class of material[10].

A number of benchmark frameworks illustrate the range of achievable structures. MOF-5 demonstrated how Zn_4O clusters and linear linkers can form large, open cubic pores with high surface areas[11]. MIL-101 demonstrated the potential for exceptionally large cages and high surface areas while highlighting the stability of Cr-based clusters[12]. HKUST-1, built from copper paddlewheel SBUs, became a model system for studying strong open-metal-site interactions and guest coordination behaviour[13]. Likewise, MOF-74 and its isostructural variants extend this concept to one-dimensional channels lined with exposed metal ions, producing strongly localized adsorption environments ideally suited for mechanistic studies of gas binding[14].

Other frameworks highlight stability and process compatibility. UiO-66, based on $Zr_6O_4(OH)_4$ clusters, exhibits exceptional chemical and thermal robustness, enabling operation under humid conditions that degrade many other MOFs[15]. ZIF-8 demonstrates that zeolite-like topologies can be recreated through metal-imidazolate coordination, yielding materials with high stability and molecular-sieving apertures[16]. More recently, CALF-20 has emerged as the first MOF

to achieve genuine industrial implementation for CO_2 capture from humid flue gas thanks to its hydrophobic pore surfaces, chemical stability, and selective CO_2 binding, which enable reliable operation under harsh process conditions[17,18].

Together, these examples demonstrate how reticular chemistry enables a systematic exploration of pore geometries and chemical environments, supporting the design of materials that couple structural precision with application-specific performance[5]. This tunability has enabled MOFs to become essential platforms for probing guest confinement, cooperative adsorption, and structure-property relationships that govern uptake, selectivity, and transport[7]. As later sections show, the same structural features that make MOFs synthetically modular also make their adsorption behaviour highly sensitive to local binding environments, emphasizing the need for molecular-level characterization to understand and predict global behaviour.

1.2 Applications of MOFs

The structural precision and chemical tunability outlined above have direct consequences for how MOFs interact with gases under working conditions. Because pore geometry, functional groups, and metal-site environments can be engineered at the molecular scale, MOFs exhibit adsorption behaviour that is far more customizable than that of traditional porous materials. This level of control is central to applications where adsorption strength, selectivity, and deliverable capacity determine performance. As a result, MOFs have become leading candidates in two major adsorption-driven areas: gas capture and gas separation. These applications provide clear examples of how structural features translate into measurable functionality.

In practical adsorption-based technologies, performance is governed not by the absolute uptake at a single pressure, but by the working capacity, defined as the difference in adsorbed

amount between the adsorption and desorption pressures of a given process[19]. Materials that achieve high loadings at low partial pressures while releasing a large fraction of adsorbed gas under modest pressure or temperature swings are therefore the most attractive for real-world operation.

1.2.1 Gas Separation

Gas separation requires not only adsorption but selective discrimination between molecular species, making it one of the most demanding and industrially important applications of MOFs. Separations such as CO_2 over N_2 in flue gas, CO_2 over CH_4 in natural gas, and C_3H_6 over C_3H_8 in petrochemical processing benefit directly from the tunable pore environments and specific interaction sites that MOFs provide.

A benchmark example for CO_2 separation is Mg-MOF-74, whose exposed Mg^{2+} open-metal sites interact strongly with the quadrupole moment of CO_2. At 298 K and 1 bar, Mg-MOF-74 achieves CO_2 uptakes near 8 mmol g^{-1}, significantly outperforming zeolites and activated carbons under comparable conditions[14]. More importantly, the strong binding at low CO_2 partial pressures leads to large working capacities under realistic pressure swing conditions, making Mg-MOF-74 a well-studied platform for post-combustion CO_2 separation.

Among the small number of MOFs that have reached industrial scale testing, CALF-20 represents a particularly important case. CALF-20 is a zinc-based framework designed for post-combustion CO_2 separation under humid, near-ambient conditions. Structured CALF-20 sorbent beds have been deployed in vacuum swing adsorption (VSA) systems applied to cement plant flue gas streams containing 12-15% CO_2, excess N_2, and a high-water content[18]. The material maintains CO_2 loadings of roughly 2 to 3 mmol g^{-1} at 298 to 313 K and 0.15 bar CO_2, even under relative

humidities above 50%[20]. Its exceptional cycling stability, sustained over hundreds of thousands of adsorption-desorption cycles, together with successful full-scale implementation[17], demonstrates that MOF-based separations can achieve true industrial viability.

MOFs are also well suited to kinetic separation, where differences in diffusion rates rather than equilibrium adsorption determine selectivity. ZIF-8 membranes provide a clear example of this behaviour. Their 3.4 Å pore aperture lies between the kinetic diameters of propylene and propane, enabling selective transport of C_3H_6 over C_3H_8. As a result, high-quality ZIF-8 membranes routinely achieve C_3H_6/C_3H_8 separation factors >100 with propylene permeances near 10^{-7} mol m^{-2} s^{-1} Pa^{-1} under near-ambient conditions[21,22]. These performance metrics exceed those of many polymer membranes and underscore the importance of rigid, well defined pore apertures in controlling molecular transport.

1.2.3 Gas Storage

Gas storage applications focus on maximizing the deliverable capacity between realistic charging and discharging pressures. Both gravimetric and volumetric storage metrics are essential, and MOFs are particularly attractive because their high porosity and tunable interaction energies allow both to be optimized simultaneously.

Ultraporous frameworks such as MOF-210 demonstrate how large pore volumes and high surface areas translate into high storage capacities. MOF-210 exhibits ~17 wt% total H_2 uptake at 77 K and 80 bar and delivers close to substantial working capacity for methane storage, delivering near 200 cm^3 (STP) cm^{-3} between 5 and 65 bar at 298 K[23,24]. These values illustrate the potential of MOFs to meet or exceed industrial targets for gas storage.

1.2.4 Other Applications

Although gas-phase adsorption remains the primary focus of MOF research, MOFs exhibit chemical and structural versatility that supports a variety of additional applications. In catalysis, their accessible metal nodes and functional linkers enable site-isolated reactivity, giving rise to activity in oxidation, hydrogenation and other transformations; for example, coordinatively unsaturated frameworks $M_3(btc)_2$ (M = Cr, Fe, Co, Ni, Cu, Zn) have been shown to catalyze CO oxidation using N_2O as the oxidant[25]. Synthetic advances such as modulated self-assembly of catalytically active Zr_6-cluster-based nanosheets further demonstrate the feasibility of fabricating stable, catalytically active MOF-derived materials[26]. As a field, MOF catalysis has matured sufficiently that strategies for improving stability, recyclability, and substrate scope are now regularly discussed in comprehensive reviews[27].

In sensing and detection, MOFs offer tunable pore environments and optical or conductive properties, making them effective chemosensors. Luminescent frameworks have been developed to detect small molecules or gases via changes in fluorescence intensity or wavelength[28]. Conductive MOFs extend this versatility by enabling electrical readout of adsorption-induced changes, broadening the range of detectable analytes[29]. These applications leverage the same host-guest chemistry and structural control that make MOFs valuable for adsorption.

MOFs have also been explored as carriers for therapeutic molecules because their high porosity, tunable surface chemistry, and controllable release environments which align well with the requirements of drug delivery systems. Early work by Horcajada and co-workers established MIL-101 as a benchmark framework in this area, demonstrating that its large pore windows and high surface area enable exceptionally high pharmaceutical loadings, including ibuprofen uptakes exceeding one gram of drug per gram of MOF, while maintaining structural integrity and

exhibiting sustained, pH-dependent release profiles[30]. Subsequent studies broadened the biomedical scope of MIL-101 derivatives. For example, iron-based MIL-101(Fe) and related variants were shown to encapsulate anticancer agents and to release them preferentially under mildly acidic conditions that mimic tumour microenvironments, illustrating how framework chemistry can be tuned to achieve selective, stimulus-responsive delivery[31]. Additional work further refined these concepts by exploiting functionalization and particle-level design to regulate guest-framework interactions, improve dispersion in biological media, and adjust release rates. These studies collectively established MIL-101 as a foundational system for illustrating how pore architecture, linker chemistry, and environmental responsiveness can be engineered to control molecular release. Although MOF-based drug delivery remains at a preclinical stage, the combination of high loading capacities, modular chemistry, and programmable release mechanisms continues to motivate research into MOFs as next-generation therapeutic carriers[32].

While these uses lie outside the scope of this thesis, they illustrate the broad functional potential of MOFs beyond classical gas storage or separation. The diversity of chemical environments, pore architectures, and accessible metal-site chemistries underscores why understanding molecular-level binding behaviour remains central to advancing both adsorption applications and broader functionalities.

1.3 Adsorption in MOFs and the Importance of Binding Sites

Adsorption in MOFs arises from the interplay between pore geometry, surface chemistry, and the underlying energy landscape that governs how guest molecules distribute within a porous structure. Although global metrics such as uptake, selectivity, and working capacity are widely used to assess material performance, these macroscopic quantities ultimately emerge from the

distribution, energetics, and connectivity of discrete binding sites inside the framework. Reviews of CO_2 capture in MOFs consistently emphasize that local adsorption environments, particularly those associated with open metal sites and strongly interacting functional groups, are key determinants of thermodynamic behaviour and separation performance[7].

A binding site corresponds to a local minimum in the adsorption potential and reflects the specific interactions that stabilize an adsorbate at that position. Changes in linker chemistry, metal nodes, pore topology, and functional groups reorganize the number, strength, and spatial arrangement of these sites and can produce substantial differences in adsorption behaviour, even for closely related frameworks. Reticular design enables the introduction of open metal sites, polar functional groups, or size-selective apertures that target specific interactions with guest molecules, which, in turn, shape both the primary adsorption sites and higher-energy secondary sites[5].

The central role of binding environments has been demonstrated explicitly in several detailed studies. Ramsahye and co-workers used periodic density functional theory (DFT) to probe CO_2 adsorption in MIL-53(Al, Cr) and MIL-47(V), showing that a small number of well-defined low-energy pockets near the inorganic chains and linker oxygens dominate CO_2 uptake, and that changes in framework composition shift the relative energies and occupancies of these sites[33]. More recent work has mapped CO_2 adsorption sites across a series of rare earth and zirconium MOFs, revealing that preferred binding locations and their relative populations vary systematically with metal identity and local coordination environment rather than with global surface area or pore volume[34]. In a different context, detailed analysis of C_2H_4/C_2H_6 separation in an Al-PyDC framework showed that high olefin/paraffin selectivity arises from several chemically distinct molecular binding pockets within a single structure, each stabilizing the guests through different interaction motifs that collectively determine the observed separation performance[35].

A review of MOFs for toxic gas capture highlighted that adsorption and selectivity for species such as SO_2, Cl_2, and NH_3 are often controlled less by total porosity and more by the chemistry, accessibility, and distribution of specific binding sites, including open metal centres, basic linkers, and reactive functional groups[36]. Recent work on water-assisted CO_2 adsorption has further shown that the presence of co-adsorbed H_2O can reorganize the adsorption landscape by occupying or modifying primary sites and creating new hydrogen-bonded environments, leading to cooperative enhancement or suppression of CO_2 binding that only emerges when the relevant binding pockets and their shared occupancy are analyzed explicitly[37].

Taken together, these studies make it clear that global adsorption metrics are rooted in a relatively small set of chemically distinct binding environments whose energies and populations depend sensitively on framework structure and operating conditions. A detailed, site-resolved description of adsorption is therefore essential for understanding why a MOF performs well for a given application, for comparing materials on a mechanistic basis, and for building predictive models that connect pore-level interactions to process-level behaviour. This perspective supports the subsequent focus of this thesis on explicit binding-site identification and analysis.

1.4 Experimental Characterization of Adsorption in MOFs

Experimental characterization of adsorption in MOFs is intrinsically challenging because guest molecules are often mobile, weakly interacting, or distributed across several partially occupied positions within the pore network. Although diffraction-based methods remain the most direct approaches for locating adsorbates, unambiguous determination of binding sites is relatively uncommon. More than 100,000 MOF structures are now deposited in the Cambridge Structural Database[38], yet only a small fraction include experimentally resolved guest positions, and these

typically require synchrotron X-ray sources or neutron facilities to obtain sufficient contrast and resolution[39,40]. Even under these optimized conditions, weakly bound or highly dynamic guests yield diffuse or smeared electron density, obscuring individual binding motifs and complicating structural refinement.

Carrington and co-workers highlighted these issues in detail, showing that gas-loaded MOFs often require extensive disorder modelling, partial-occupancy constraints, and careful interpretation of residual density to approximate binding-site geometries[41]. Similar difficulties arise in systems where adsorbates populate several symmetry-related or metastable environments, or in frameworks that exhibit structural flexibility during adsorption. These factors create ensemble-averaged diffraction signatures that cannot readily be decomposed into distinct site contributions.

Spectroscopic methods provide complementary insight but also face inherent limitations. Infrared spectroscopy probes changes in vibrational modes associated with adsorption, and can often identify functional groups involved in binding, but does not resolve three-dimensional guest positions. Solid-state NMR has become a powerful tool for probing local environments, guest mobility, and adsorption-induced structural changes, particularly when diffraction provides insufficient detail[42,43]. However, NMR spectra typically reflect overlapping chemical environments and require computational models to distinguish between plausible adsorption configurations. Recent studies combining NMR with variable-temperature measurements, isotopic labelling, or co-adsorption experiments have demonstrated the value of these techniques, but they still rely on simulations to fully reconstruct the spatial distribution of adsorbates.

Altogether, while experimental methods are indispensable for validating adsorption behaviour and assessing framework stability, their ability to resolve detailed binding-site

distributions is fundamentally limited by disorder, partial occupancy, guest mobility, and instrumental constraints. These limitations motivate the integration of atomistic simulation methods, which can map out three-dimensional adsorption probability distributions and energetics under controlled, idealized conditions, complementing experimental data and enabling a more complete understanding of adsorption in MOFs.

1.5 Atomistic Simulation of Adsorption in MOFs

Atomistic simulations provide a powerful complement to experimental characterization because they can resolve adsorption sites with a level of spatial and energetic detail that is often inaccessible to diffraction or spectroscopy. Several computational strategies have been developed to identify binding sites in MOFs, but each has inherent limitations that become significant when capturing realistic adsorption behaviour under working conditions.

Periodic DFT calculations are among the most widely used tools for studying adsorption at the molecular level. In these approaches, one or a few guest molecules are placed at chemically reasonable positions in the pore and the geometry is relaxed to obtain local minima. DFT has been highly successful at elucidating strongly bound adsorption motifs, such as CO_2 binding at open metal sites in Mg-MOF-74 and related structures[44,45]. However, these optimizations are carried out at effectively 0 K, omitting thermal motion, entropic effects, and guest-guest interactions. They provide discrete stationary points rather than a full occupancy distribution. More importantly, DFT inherently relies on chemically selected starting positions, so unidentified or unexpected sites are easily missed.

On the other hand, force-field-based potential-energy surface (PES) scans offer a complementary strategy. Here, a probe molecule is placed on a real-space grid spanning the pore

volume, and the interaction energy is computed at each point to generate an energy landscape. Local minima are then used to suggest likely binding sites or to place initial configurations for further simulation. Rosen and co-workers used such an approach in their high-throughput DFT screening of catalytic MOFs, where a methane probe was guided by a Lennard-Jones interaction grid[46]. While chemically informative, these PES-based placement schemes are deterministic and guest-specific. They perform well for simple, near-spherical adsorbates such as methane, where orientation plays a minor role. However, they struggle with polyatomic molecules that have multiple chemically distinct interaction sites, where adsorption depends sensitively on molecular orientation and local framework chemistry. In such cases, a limited set of optimized geometries cannot adequately sample the orientational degrees of freedom or capture the existence of multiple competing local minima. These approaches also fail under competitive adsorption or at higher loadings, where guest-guest interactions affect both preferred orientations and binding locations. As with DFT-based geometry optimization, PES-driven methods provide no information on site populations or statistical weighting at finite temperature and pressure. All identified minima are implicitly treated as equally relevant, which is rarely true for adsorption in porous materials.

In contrast, Grand canonical Monte Carlo (GCMC) simulations overcome these limitations by sampling guest configurations according to the Boltzmann distribution at a defined temperature, pressure, and composition. GCMC inherently generates three-dimensional adsorption probability distributions (APDs) that encode the preferred binding locations, their relative statistical weights, and how these distributions evolve with loading. Peaks in the APD correspond to condition-specific adsorption maxima and free energy minima. Their heights reflect the occupancy and stability of each site, allowing primary and secondary binding environments to be distinguished.

Unlike DFT or PES grids, GCMC captures thermal fluctuations, guest mobility, and guest-guest interactions, making it uniquely suited for identifying realistic binding landscapes.

These considerations motivate the use of GCMC-derived probability distributions as the foundation for systematic binding-site extraction. At the same time, they highlight a critical methodological gap: there is currently no standardized, automated, and validated approach for converting APDs into chemically meaningful binding-site coordinates. Most studies still rely on visual inspection or qualitative interpretation, leaving the quantitative structural information contained in probability maps largely underutilized. The overall workflow for transforming noisy APD into discrete binding site configurations is illustrated in Figure 1.1. To clarify how APDs arise from atomistic simulations, the following sections outline the principles of GCMC simulations and describe how sampled configurations are accumulated into three-dimensional probability distributions.

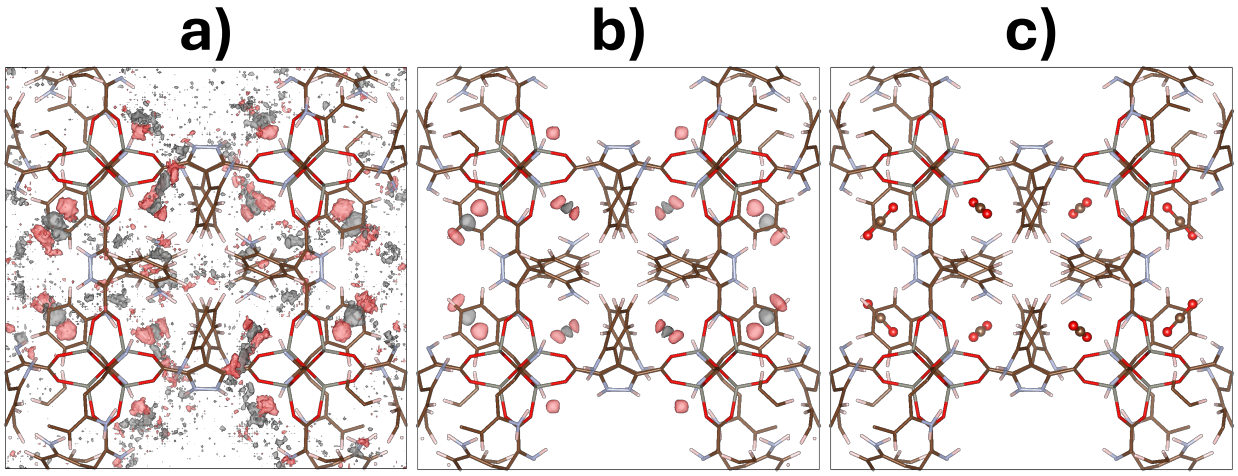

Figure 1.1. Binding site identification from GCMC-derived APD for CO_2 in a representative MOF. a) Raw APD obtained directly from GCMC, showing a noisy and spatially diffuse landscape. b) Smoothed APD highlighting localized regions of high occupancy. c) Final fitted binding site configuration obtained by fitting the CO_2 molecule to the extracted probability maxima, yielding chemically meaningful binding sites. From left to right, the panels illustrate the successive stages of probability smoothing, maxima localization, and guest fitting employed in the GALP workflow. In panels a) and b), red density corresponds to oxygen atoms and gray density corresponds to carbon atoms of CO_2.

1.5.1 Grand Canonical Monte Carlo Simulations

GCMC simulations are widely used to model adsorption in porous materials because they directly sample the equilibrium distribution of guest molecules under experimentally relevant conditions. In the grand canonical ensemble, the temperature, volume, and chemical potential of each adsorbing species are held constant while the number of guest molecules in the simulation cell is allowed to fluctuate. The imposed chemical potential corresponds to a fixed gas pressure, enabling direct comparison with adsorption experiments. In this ensemble, configurations are sampled with a probability proportional to

$$P \propto e^{-\frac{U-\mu N}{kT}} \quad 1.1$$

where U is the total interaction energy of the system, μ is the chemical potential of the guest species, N is the number of guest molecules, k is the Boltzmann constant, and T is the temperature. This expression highlights the defining feature of GCMC simulations, namely that both energetic interactions and fluctuations in guest loading contribute to the equilibrium distribution.

A GCMC simulation proceeds by proposing a sequence of stochastic trial moves that modify the system's configuration. These moves typically include insertions and deletions of guests, as well as translation and rotation of molecules already present within the framework. Each proposed move is accepted or rejected according to the Metropolis criterion, which compares the Boltzmann weighted probability of the new configuration to that of the current one while enforcing detailed balance. This procedure guarantees that, over many iterations, configurations are sampled according to the target grand canonical distribution defined in equation 1.1. The total energy U entering the acceptance criterion is computed as the sum of the framework-guest and guest-guest interaction energies, usually described using classical force fields that include van der Waals and electrostatic contributions. As guest molecules are free to move, reorient, and interact with one

another, GCMC naturally capture thermal motion, configurational disorder, and guest-guest correlations that arise at finite loadings. The resulting ensemble of sampled configurations retains detailed local binding information about the spatial distribution of guest molecules within the framework, providing the basis for constructing APDs that reflect realistic binding behaviours under working conditions.

1.5.2 Generation of Adsorption Probability Distributions

During each accepted GCMC step, the position of guest molecule atoms are mapped onto a three-dimensional grid spanning the simulation cell by assigning each atom to its nearest voxel. At each grid point where an atom is identified, a running count is incremented. Repeating this procedure throughout the simulation generates a raw occupancy histogram for each guest atom types, which is continuously updated as additional configurations are sampled. Upon completion of the simulation, these counts are normalized by the total number of MC steps and by the volume of each voxel to produce three-dimensional APD. The occupancy at a given grid point i, denoted by ρ_i, is calculated as:

$$\rho_i = \frac{number\ of\ steps\ guest\ atom\ binned\ at\ point\ i}{(total\ number\ of\ MC\ steps) \cdot\ (volume\ of\ bin)} \tag{1.2}$$

The volume of each voxel is determined by the unit cell geometry. For a triclinic lattice, the volume of a single bin is given by:

$$Volume\ of\ Bin = \frac{abc\sqrt{1 - \cos^2\alpha - \cos^2\beta - \cos^2\gamma + 2\cos\alpha\cos\beta\cos\gamma}}{N} \tag{1.3}$$

Where a, b, and c are the lattice vector magnitudes in Å, α, β, and γ are the inter-vector angles in degrees, and N is the total number of voxels in the unit cell. The resulting occupancy values, ρ_i,

have units of inverse volume, specifically $Å^{-3}$, representing the probability density of finding a given guest atom at a particular grid point.

Separated APD are generated for each guest atom type, for example, carbon and oxygen in CO_2, enabling independent analysis of the spatial distribution of individual atomic sites. These distributions are stored in the Gaussian cube file format, which preserves the normalized probability data together with the real-space lattice information required for subsequent visualization and post-processing.

1.6 Motivation and Research Problem

The adsorption behaviour of a MOF is controlled by the specific environments within its pores. The locations, geometries, and energetics of individual binding sites determine how guest molecules organize, how strongly they interact with the framework, and how these interactions evolve with temperature, pressure, and loading. Global metrics such as uptake, selectivity, and working capacity are therefore direct consequences of the local adsorption landscape. Understanding these binding sites is essential for interpreting adsorption mechanisms, rationalizing performance differences between materials, and predicting adsorption behaviour under realistic conditions.

GCMC simulations are widely used to model adsorption in MOFs because they capture thermal motion, guest-guest interactions, and finite-loading effects. A key advantage of GCMC is that it produces three-dimensional adsorption probability distributions that directly encode the spatial organization of guest molecules within a framework. These distributions contain far more structural information than global observables, and they reveal which adsorption environments dominate under specific thermodynamic conditions.

Despite their value, no standardized or validated framework exists for converting GCMC-derived probability distributions into explicit binding-site coordinates. Most studies still report only scalar outputs such as uptake or isosteric heat, while the underlying adsorption landscape remains qualitatively inspected or omitted entirely. This lack of a consistent extraction methodology leaves substantial chemically meaningful information unused.

The absence of such a method has two immediate consequences. It limits our ability to systematically compare adsorption behaviour across different guests, materials, or operating conditions. It also prevents meaningful evaluation of the classical approximations used in GCMC simulations, including the choice of force field, charge assignment scheme, framework rigidity, and equation of state. Without a reliable way to extract binding sites, it is not possible to determine whether classical GCMC reproduces known experimental adsorption environments with sufficient accuracy to support mechanistic interpretation. Addressing this unmet need forms the basis of the research problem addressed in this thesis.

1.7 Impact and Future Relevance

A validated and automated method for extracting binding sites from GCMC-derived probability distributions would have a broad impact across molecular simulation, materials design, and adsorption science. At the molecular scale, accurate site identification provides direct insight into the interaction motifs that govern adsorption strength, orientation, cooperativity, and competitive behaviour. This level of detail enables mechanistic interpretation of adsorption trends, allowing researchers to understand why certain frameworks outperform others for CO_2 capture, hydrocarbon separations, or operation under humid conditions.

At the materials-design level, binding-site information supports targeted modification of pore environments through linker functionalization, metal substitution, or the introduction of specific binding motifs. These insights refine how structure-property relationships are established, enabling adsorption performance to be rationally tuned rather than inferred indirectly from global metrics.

At the process scale, a clear description of binding-site structure strengthens the connection between atomistic simulation and engineering models. Site energetics and occupancies influence breakthrough behaviour, working capacity, regeneration energy, and multicomponent co-adsorption. Accurate binding-site maps, therefore, improve predictions for pressure-swing adsorption, vacuum-swing adsorption, and other separation processes that depend sensitively on local adsorption environments.

The long-term relevance extends even further. High-throughput GCMC workflows are already widely used, yet their outputs remain dominated by global properties such as uptake or Henry's coefficients. An automated and validated extraction procedure transforms these workflows by converting raw probability distributions into structured binding-site datasets. Such datasets enable force-field benchmarking, cross-material comparison of adsorption environments, and the development of new descriptors grounded in local interactions. They also provide training data for machine learning models that aim to predict adsorption landscapes, free-energy surfaces, or even full probability distributions at a fraction of the computational cost.

In this sense, the broader impact lies not only in clarifying adsorption mechanisms for individual MOFs but in establishing a foundation for scalable, data-driven approaches to adsorption science.

1.8 Thesis Goals and Chapter Overview

This thesis has two core objectives, both of which shape the organization of the document. The first objective is to develop a standalone, automated workflow that extracts binding sites from APDs generated by GCMC simulations. This workflow performs smoothing, peak identification, clustering, and molecular fitting to convert three-dimensional probability maps into chemically interpretable binding site coordinates. The theoretical foundation and methodological implementation of this procedure are presented in Chapter 2, which defines the full computational pipeline for probability-based binding site identification.

The second objective is to evaluate the reliability of widely used classical GCMC approximations by comparing simulation-predicted binding sites to experimentally determined binding sites. This evaluation assesses the influence of force-field selection, charge-assignment methods, and the rigid framework approximation. The results of this validation effort form the basis of Chapter 3, which examines the conditions under which classical GCMC reproduces known adsorption motifs with sufficient accuracy to support mechanistic interpretation.

Finally, Chapter 4 summarizes the findings from both objectives and discusses their implications for future work. This includes the broader relevance of systematic binding site identification for high-throughput screening and the creation of large adsorption datasets. Although these applications fall outside the scope of this thesis, parallel research in the group has begun exploring them using the methods developed here, demonstrating the adaptability and future value of this approach.

1.9 References

(1) Khan, I.; Altaf, A.; Sadiq, S.; Khan, S.; Khan, A.; Khan, S.; Humayun, M.; Khan, A.; Abumousa, R. A.; Bououdina, M. Towards Sustainable Solutions: Comprehensive Review of Advanced Porous Materials for CO_2 Capture, Hydrogen Generation, Pollutant Degradation, and Energy Application. *Chemical Engineering Journal Advances* **2025**, *21*, 100691. https://doi.org/10.1016/j.ceja.2024.100691.

(2) Broom, D. P.; Thomas, K. M. Gas Adsorption by Nanoporous Materials: Future Applications and Experimental Challenges. *MRS Bull.* **2013**, *38* (5), 412–421. https://doi.org/10.1557/mrs.2013.105.

(3) Li, J.-R.; Kuppler, R. J.; Zhou, H.-C. Selective Gas Adsorption and Separation in Metal–Organic Frameworks. *Chem. Soc. Rev.* **2009**, *38* (5), 1477. https://doi.org/10.1039/b802426j.

(4) Zhou, H.-C.; Long, J. R.; Yaghi, O. M. Introduction to Metal–Organic Frameworks. *Chem. Rev.* **2012**, *112* (2), 673–674. https://doi.org/10.1021/cr300014x.

(5) Furukawa, H.; Cordova, K. E.; O'Keeffe, M.; Yaghi, O. M. The Chemistry and Applications of Metal-Organic Frameworks. *Science (1979).* **2013**, *341* (6149). https://doi.org/10.1126/science.1230444.

(6) Jiang, H.; Alezi, D.; Eddaoudi, M. A Reticular Chemistry Guide for the Design of Periodic Solids. *Nat. Rev. Mater.* **2021**, *6* (6), 466–487. https://doi.org/10.1038/s41578-021-00287-y.

(7) Sumida, K.; Rogow, D. L.; Mason, J. A.; McDonald, T. M.; Bloch, E. D.; Herm, Z. R.; Bae, T.-H.; Long, J. R. Carbon Dioxide Capture in Metal–Organic Frameworks. *Chem. Rev.* **2012**, *112* (2), 724–781. https://doi.org/10.1021/cr2003272.

(8) Wang, C.; Liu, D.; Lin, W. Metal–Organic Frameworks as A Tunable Platform for Designing Functional Molecular Materials. *J. Am. Chem. Soc.* **2013**, *135* (36), 13222–13234. https://doi.org/10.1021/ja308229p.

(9) Rowsell, J. L. C.; Yaghi, O. M. Metal–Organic Frameworks: A New Class of Porous Materials. *Microporous and Mesoporous Materials* **2004**, *73* (1–2), 3–14. https://doi.org/10.1016/j.micromeso.2004.03.034.

(10) Lowe, D. *The 2025 Nobel Prize in Chemistry: Metal-Organic Frameworks*. https://www.science.org/content/blog-post/2025-nobel-prize-chemistry-metal-organic-frameworks.

(11) Li, H.; Eddaoudi, M.; O'Keeffe, M.; Yaghi, O. M. Design and Synthesis of an Exceptionally Stable and Highly Porous Metal-Organic Framework. *Nature* **1999**, *402* (6759), 276–279. https://doi.org/10.1038/46248.

(12) Férey, G.; Mellot-Draznieks, C.; Serre, C.; Millange, F.; Dutour, J.; Surblé, S.; Margiolaki, I. A Chromium Terephthalate-Based Solid with Unusually Large Pore Volumes and Surface Area. *Science (1979).* **2005**, *309* (5743), 2040–2042. https://doi.org/10.1126/science.1116275.

(13) Chui, S. S.-Y.; Lo, S. M.-F.; Charmant, J. P. H.; Orpen, A. G.; Williams, I. D. A Chemically Functionalizable Nanoporous Material [Cu3(TMA)2(H2O)3]n. *Science (1979).* **1999**, *283* (5405), 1148–1150. https://doi.org/10.1126/science.283.5405.1148.

(14) Britt, D.; Furukawa, H.; Wang, B.; Glover, T. G.; Yaghi, O. M. Highly Efficient Separation of Carbon Dioxide by a Metal-Organic Framework Replete with Open Metal Sites. *Proceedings of the National Academy of Sciences* **2009**, *106* (49), 20637–20640. https://doi.org/10.1073/pnas.0909718106.

(15) Cavka, J. H.; Jakobsen, S.; Olsbye, U.; Guillou, N.; Lamberti, C.; Bordiga, S.; Lillerud, K. P. A New Zirconium Inorganic Building Brick Forming Metal Organic Frameworks with Exceptional Stability. *J. Am. Chem. Soc.* **2008**, *130* (42), 13850–13851. https://doi.org/10.1021/ja8057953.

(16) Park, K. S.; Ni, Z.; Côté, A. P.; Choi, J. Y.; Huang, R.; Uribe-Romo, F. J.; Chae, H. K.; O'Keeffe, M.; Yaghi, O. M. Exceptional Chemical and Thermal Stability of Zeolitic Imidazolate Frameworks. *Proceedings of the National Academy of Sciences* **2006**, *103* (27), 10186–10191. https://doi.org/10.1073/pnas.0602439103.

(17) Oktavian, R.; Goeminne, R.; Glasby, L. T.; Song, P.; Huynh, R.; Qazvini, O. T.; Ghaffari-Nik, O.; Masoumifard, N.; Cordiner, J. L.; Hovington, P.; Van Speybroeck, V.; Moghadam, P. Z. Gas Adsorption and Framework Flexibility of CALF-20 Explored via Experiments and Simulations. *Nat. Commun.* **2024**, *15* (1), 3898. https://doi.org/10.1038/s41467-024-48136-0.

(18) Lin, J.-B.; Nguyen, T. T. T.; Vaidhyanathan, R.; Burner, J.; Taylor, J. M.; Durekova, H.; Akhtar, F.; Mah, R. K.; Ghaffari-Nik, O.; Marx, S.; Fylstra, N.; Iremonger, S. S.; Dawson, K. W.; Sarkar, P.; Hovington, P.; Rajendran, A.; Woo, T. K.; Shimizu, G. K. H. A Scalable Metal-Organic Framework as a Durable Physisorbent for Carbon Dioxide Capture. *Science (1979).* **2021**, *374* (6574), 1464–1469. https://doi.org/10.1126/science.abi7281.

(19) Gándara, F.; Furukawa, H.; Lee, S.; Yaghi, O. M. High Methane Storage Capacity in Aluminum Metal–Organic Frameworks. *J. Am. Chem. Soc.* **2014**, *136* (14), 5271–5274. https://doi.org/10.1021/ja501606h.

(20) Gopalsamy, K.; Fan, D.; Naskar, S.; Magnin, Y.; Maurin, G. Engineering of an Isoreticular Series of CALF-20 Metal–Organic Frameworks for CO2 Capture. *ACS Applied Engineering Materials* **2024**, *2* (1), 96–103. https://doi.org/10.1021/acsaenm.3c00622.

(21) Kwon, H. T.; Jeong, H.-K.; Lee, A. S.; An, H. S.; Lee, J. S. Heteroepitaxially Grown Zeolitic Imidazolate Framework Membranes with Unprecedented Propylene/Propane Separation Performances. *J. Am. Chem. Soc.* **2015**, *137* (38), 12304–12311. https://doi.org/10.1021/jacs.5b06730.

(22) Zhao, Y.; Wei, Y.; Lyu, L.; Hou, Q.; Caro, J.; Wang, H. Flexible Polypropylene-Supported ZIF-8 Membranes for Highly Efficient Propene/Propane Separation. *J. Am. Chem. Soc.* **2020**, *142* (50), 20915–20919. https://doi.org/10.1021/jacs.0c07481.

(23) Furukawa, H.; Ko, N.; Go, Y. B.; Aratani, N.; Choi, S. B.; Choi, E.; Yazaydin, A. Ö.; Snurr, R. Q.; O'Keeffe, M.; Kim, J.; Yaghi, O. M. Ultrahigh Porosity in Metal-Organic Frameworks. *Science (1979).* **2010**, *329* (5990), 424–428. https://doi.org/10.1126/science.1192160.

(24) Suh, M. P.; Park, H. J.; Prasad, T. K.; Lim, D.-W. Hydrogen Storage in Metal–Organic Frameworks. *Chem. Rev.* **2012**, *112* (2), 782–835. https://doi.org/10.1021/cr200274s.

(25) Ketrat, S.; Maihom, T.; Wannakao, S.; Probst, M.; Nokbin, S.; Limtrakul, J. Coordinatively Unsaturated Metal–Organic Frameworks M3(Btc)2 (M = Cr, Fe, Co, Ni, Cu, and Zn) Catalyzing the Oxidation of CO by N2O: Insight from DFT Calculations. *Inorg. Chem.* **2017**, *56* (22), 14005–14012. https://doi.org/10.1021/acs.inorgchem.7b02143.

(26) Prasad, R. R. R.; Boyadjieva, S. S.; Zhou, G.; Tan, J.; Firth, F. C. N.; Ling, S.; Huang, Z.; Cliffe, M. J.; Foster, J. A.; Forgan, R. S. Modulated Self-Assembly of Catalytically Active Metal–Organic Nanosheets Containing Zr6 Clusters and Dicarboxylate Ligands. *ACS Appl. Mater. Interfaces* **2024**, *16* (14), 17812–17820. https://doi.org/10.1021/acsami.4c00604.

(27) Yang, D.; Gates, B. C. Catalysis by Metal Organic Frameworks: Perspective and Suggestions for Future Research. *ACS Catal.* **2019**, *9* (3), 1779–1798. https://doi.org/10.1021/acscatal.8b04515.

(28) Hu, Z.; Deibert, B. J.; Li, J. Luminescent Metal–Organic Frameworks for Chemical Sensing and Explosive Detection. *Chem. Soc. Rev.* **2014**, *43* (16), 5815–5840. https://doi.org/10.1039/C4CS00010B.

(29) Ko, M.; Mendecki, L.; Mirica, K. A. Conductive Two-Dimensional Metal–Organic Frameworks as Multifunctional Materials. *Chemical Communications* **2018**, *54* (57), 7873–7891. https://doi.org/10.1039/C8CC02871K.

(30) Horcajada, P.; Chalati, T.; Serre, C.; Gillet, B.; Sebrie, C.; Baati, T.; Eubank, J. F.; Heurtaux, D.; Clayette, P.; Kreuz, C.; Chang, J.-S.; Hwang, Y. K.; Marsaud, V.; Bories, P.-N.; Cynober, L.; Gil, S.; Férey, G.; Couvreur, P.; Gref, R. Porous Metal–Organic-Framework Nanoscale Carriers as a Potential Platform for Drug Delivery and Imaging. *Nat. Mater.* **2010**, *9* (2), 172–178. https://doi.org/10.1038/nmat2608.

(31) Huxford, R. C.; Della Rocca, J.; Lin, W. Metal–Organic Frameworks as Potential Drug Carriers. *Curr. Opin. Chem. Biol.* **2010**, *14* (2), 262–268. https://doi.org/10.1016/j.cbpa.2009.12.012.

(32) Wang, L.; Zheng, M.; Xie, Z. Nanoscale Metal–Organic Frameworks for Drug Delivery: A Conventional Platform with New Promise. *J. Mater. Chem. B* **2018**, *6* (5), 707–717. https://doi.org/10.1039/C7TB02970E.

(33) Ramsahye, N. A.; Maurin, G.; Bourrelly, S.; Llewellyn, P. L.; Serre, C.; Loiseau, T.; Devic, T.; Férey, G. Probing the Adsorption Sites for CO2 in Metal Organic Frameworks Materials MIL-53 (Al, Cr) and MIL-47 (V) by Density Functional Theory. *The Journal of Physical Chemistry C* **2008**, *112* (2), 514–520. https://doi.org/10.1021/jp075782y.

(34) Tassé, D.; Quezada-Novoa, V.; Copeman, C.; Howarth, A. J.; Rochefort, A. Identification of Adsorption Sites for CO2 in a Series of Rare-Earth and Zr-Based Metal-Organic Frameworks. *ChemPhysChem* **2025**, *26* (10). https://doi.org/10.1002/cphc.202401050.

(35) Wu, E.; Gu, X.-W.; Liu, D.; Zhang, X.; Wu, H.; Zhou, W.; Qian, G.; Li, B. Incorporation of Multiple Supramolecular Binding Sites into a Robust MOF for Benchmark One-Step Ethylene Purification. *Nat. Commun.* **2023**, *14* (1), 6146. https://doi.org/10.1038/s41467-023-41692-x.

(36) Wen, M.; Li, G.; Liu, H.; Chen, J.; An, T.; Yamashita, H. Metal–Organic Framework-Based Nanomaterials for Adsorption and Photocatalytic Degradation of Gaseous Pollutants: Recent Progress and Challenges. *Environ. Sci. Nano* **2019**, *6* (4), 1006–1025. https://doi.org/10.1039/C8EN01167B.

(37) Ho, C.-H.; Paesani, F. Elucidating the Competitive Adsorption of H2O and CO2 in CALF-20: New Insights for Enhanced Carbon Capture Metal-Organic Frameworks. September 22, 2023. https://doi.org/10.26434/chemrxiv-2023-qfb1n-v2.

(38) Moghadam, P. Z.; Li, A.; Wiggin, S. B.; Tao, A.; Maloney, A. G. P.; Wood, P. A.; Ward, S. C.; Fairen-Jimenez, D. Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future. *Chemistry of Materials* **2017**, *29* (7), 2618–2625. https://doi.org/10.1021/acs.chemmater.7b00441.

(39) Easun, T. L.; Moreau, F.; Yan, Y.; Yang, S.; Schröder, M. Structural and Dynamic Studies of Substrate Binding in Porous Metal–Organic Frameworks. *Chem. Soc. Rev.* **2017**, *46* (1), 239–274. https://doi.org/10.1039/C6CS00603E.

(40) Han, Z.; Yu, K.-H.; Wang, K.-Y.; Zhou, H.-C. Binding Sites of Automobile Exhaust Gases on Metal–Organic Frameworks: Advances and Perspectives. *Energy & Fuels* **2025**, *39* (13), 6151–6163. https://doi.org/10.1021/acs.energyfuels.5c00552.

(41) Carrington, E. J.; Vitórica-Yrezábal, I. J.; Brammer, L. Crystallographic Studies of Gas Sorption in Metal–Organic Frameworks. *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.* **2014**, *70* (3), 404–422. https://doi.org/10.1107/S2052520614009834.

(42) Lucier, B. E. G.; Chen, S.; Huang, Y. Characterization of Metal–Organic Frameworks: Unlocking the Potential of Solid-State NMR. *Acc. Chem. Res.* **2018**, *51* (2), 319–330. https://doi.org/10.1021/acs.accounts.7b00357.

(43) Martins, V.; Lucier, B. E. G.; Liu, Z.; Liang, H.; Zheng, A.; Terskikh, V. V.; Zhang, W.; Desveaux, B. E.; Huang, Y. Cold, Hot, Dry, and Wet: Locations and Dynamics of CO2 and H2O Co-Adsorbed in an Ultramicroporous MOF. *Inorg. Chem.* **2023**, *62* (28), 11152–11167. https://doi.org/10.1021/acs.inorgchem.3c01251.

(44) Marti, R. M.; Howe, J. D.; Morelock, C. R.; Conradi, M. S.; Walton, K. S.; Sholl, D. S.; Hayes, S. E. CO2 Dynamics in Pure and Mixed-Metal MOFs with Open Metal Sites. *The Journal of Physical Chemistry C* **2017**, *121* (46), 25778–25787. https://doi.org/10.1021/acs.jpcc.7b07179.

(45) Queen, W. L.; Hudson, M. R.; Bloch, E. D.; Mason, J. A.; Gonzalez, M. I.; Lee, J. S.; Gygi, D.; Howe, J. D.; Lee, K.; Darwish, T. A.; James, M.; Peterson, V. K.; Teat, S. J.; Smit, B.; Neaton, J. B.; Long, J. R.; Brown, C. M. Comprehensive Study of Carbon Dioxide Adsorption in the Metal–Organic Frameworks M2(Dobdc) (M = Mg, Mn, Fe, Co, Ni, Cu, Zn). *Chem. Sci.* **2014**, *5* (12), 4569–4581. https://doi.org/10.1039/C4SC02064B.

(46) Rosen, A. S.; Notestein, J. M.; Snurr, R. Q. Identifying Promising Metal–Organic Frameworks for Heterogeneous Catalysis via High-throughput Periodic Density Functional Theory. *J. Comput. Chem.* **2019**, *40* (12), 1305–1318. https://doi.org/10.1002/jcc.25787.

2. Development of a Binding site Identification Tool for MOFs

Enhancing Gas Absorption Analysis in MOFs: Development of an Advanced and Robust Binding Site Identification Tool

2.1 Abstract

GALP (Guest Atom Localizer from Probabilities) is a Python tool that identifies adsorption binding sites directly from three-dimensional absorbate probability distributions generated by grand canonical Monte Carlo (GCMC) simulations. The method addresses a central bottleneck in adsorption studies, namely the lack of a general, automated procedure for converting noisy, grid-based probability distribution data into chemically meaningful guest configurations, at a time when experimentally resolved binding sites are available only for a small set of MOFs. GALP operates on atom-specific adsorbate probability distributions stored on a 3D grid and combines Gaussian smoothing, local maximum detection with occupancy and distance-based filters, and an RMSD minimizing fitting procedure to reconstruct full guest geometries under periodic boundary conditions. Convergence and data quality are quantified with a Tanimoto similarity criterion between symmetry equivalent subcells, complemented by a Shannon entropy metric that reports the degree of spatial delocalization. Parameters controlling smoothing, maxima selection, and molecular fitting are optimized on a development set of 100 diverse MOFs and 8 guests (CO_2, N_2, CH_4, Xe, Kr, C_2H_2, C_3H_6, C_3H_8). With this development set, the number of binding sites determined by GALP is in excellent agreement with those determined manually, where parity plots give a coefficient of determination of ≥ 0.98. A grid spacing of 0.15 Å is shown to reproduce the positions and binding energies of sites obtained from a much finer 0.01 Å grid with average energy deviations of about 0.2 kcal mol^{-1}, while keeping file sizes and runtimes tractable for high-

throughput workflows. Together, these results establish GALP as an accurate and practical framework agnostic tool for constructing binding site datasets, guiding targeted experiments, and enabling automated screening and machine learning models for gas adsorption in porous materials.

2.2 Statement of Work

The work presented in this chapter is based on the development and application of GALP. I was responsible for the full redesign and implementation of the software, as well as for all validation and analysis reported here. Specifically, I rewrote the original Guest Atom Localization Algorithm (GALA) codebase from scratch to improve robustness, readability, and long-term maintainability. I developed a new RMSD-based fitting algorithm for binding site identification and corrected multiple algorithmic and numerical issues in the original implementation. I refactored the code into a fully standalone package by removing proprietary dependencies and replacing them with well-documented and actively maintained libraries, and I extended its compatibility to natively support not only our in-house GCMC code FastMC but also the widely used RASPA3 package. I further designed and executed the validation workflow to assess the method's accuracy and reliability across diverse MOF and guest systems.

2.3 Introduction

Metal-organic frameworks (MOFs) are increasingly recognized as effective materials for gas separation and storage, owing to their highly diverse structures, high porosity, and chemical tunability[47]. The MOF CALF-20 is one such MOF, that has been commercialized for selective CO_2 capture from humid cement making flues[18,48]. Binding sites of the adsorbates are central to

how MOFs selectively capture[35,49] and understanding them is key to the rational design of solid sorbents for specific applications.[50,51]

Experimentally determining the precise locations of adsorbed species within porous materials remains a significant challenge. Techniques such as single-crystal X-ray diffraction (SCXRD), neutron powder diffraction (NPD), powder X-ray diffraction (PXRD), and synchrotron-based methods are indispensable for identifying guest positions within frameworks. However, these methods rely on highly ordered samples with sufficient guest occupancy and minimal disorder. In practice, low scattering contrast for light adsorbates (e.g., H_2, CO_2, CH_4), diffuse scattering from framework flexibility, and site averaging due to dynamic motion often obscure or smear out the electron or nuclear density associated with the guest. Even under optimized conditions, partial occupancy, residual solvent, and framework breathing can further complicate site assignment. Consequently, experimentally resolved adsorption sites are available only for a limited number of systems[51]. Many computational studies have highlighted the same limitation, emphasizing how experimental constraints hinder direct benchmarking of adsorption site predictions[41]. From our analysis, we identified roughly 35 distinct structures where the guest location has been resolved within a MOF, highlighting both the experimental difficulty and the value of computational methods for completing this picture.

Despite the central role that binding sites play in adsorption behaviour, there are currently no general-purpose tools for automatically identifying them in periodic structures. As a result, researchers investigating their synthesized MOFs or analyzing top-performing structures often lack a direct method for locating these sites. Experimental binding site data remain difficult to obtain, while computational alternatives usually rely on manually placing guests and refining them through DFT to locate local minima. A generalizable approach that extracts binding sites from

atomistic simulations would provide a much-needed solution. It would enable not only a more rigorous analysis of adsorption sites but also the construction of large-scale datasets for screening, training machine learning models, and designing new MOFs tailored to specific guest molecules. However, extracting these sites directly from simulation outputs, particularly from the noisy probability distributions generated from atomistic simulations, remains a nontrivial task.

A practical way to explore binding site behaviour and guest uptake in MOFs is through atomistic simulations that model how guest molecules interact with the framework under relevant thermodynamic conditions. GCMC simulations, in particular, are commonly employed to model gas adsorption under constant temperature and volume conditions. In a GCMC simulation, trial moves that insert, delete, translate, and rotate guest molecules are accepted or rejected according to the Metropolis criterion based on changes in the total interaction energy, thereby sampling the equilibrium distribution of guests for a given state point. By averaging over many millions of such configuration, each simulation yields a single point on an adsorption isotherm. Beyond providing uptake data, GCMC simulations also produce three-dimensional probability distributions that represent the likelihood of finding a guest atom at a given position within the framework. These distributions are commonly represented as volumetric grids and visualized as isosurfaces or contour maps. An isosurface corresponds to a surface in three-dimensional space along which the probability density takes a constant value, while contour plots represent the same information on two-dimensional slices through the framework. Regions of high probability in these representations indicate preferred binding locations for the guest, thereby providing direct insight into the spatial organization of the adsorption within the pore structure[18,52].

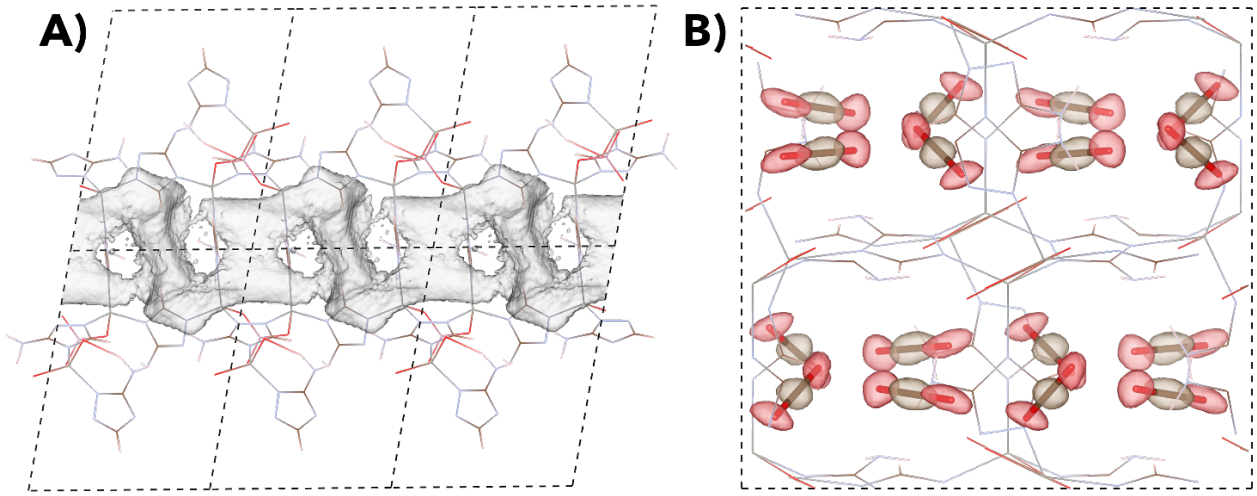

Figure 2.1. A) 3D-isosurfaces of the CO_2 probability distributions (brown – carbon; red – oxygen) in CALF-15 (Zn_2(3-amino-1,2,4-triazole)$_2$(oxalate)), determined from a GCMC simulation. Also shown in the tube representation are the experimental CO_2 binding sites determined from X-ray analysis. B) Centre of mass probability density plots of CO_2 molecules in CALF-16 (Zn_3(3-amino-1,2,4-triazole)$_3$(PO_4)). In both A) and B), the framework of the MOF is shown in line representation for clarity.

For example, Figure 2.1A demonstrates the raw probability plot for the carbon atoms of CO_2. The regions of high occupancy are shown in gray; however, manually identifying the exact maxima is not straightforward. Typically, this visualization provides enough information to determine the general region where the guest resides, but it does not precisely define the binding site. By localizing and extracting these maxima, we obtain clearly defined regions resembling CO_2 molecules, as illustrated in Figure 2.1B. At this stage, a robust algorithm can be applied to fit the guest molecule within the adsorption site accurately.

Identifying binding sites is essential for understanding the atomistic level adsorption behaviour of a MOF. These sites indicate where guest molecules are preferentially located, offering insights into the chemistry and intermolecular interactions that drive binding and selectivity. In binary or multicomponent simulations, binding sites can reveal competitive adsorption, such as displacement driven by stronger framework affinity or cooperative uptake, whereby one molecule enhances the adsorption of another. Examining these sites also allows for the identification of functional groups or local motifs that contribute to strong binding, facilitating

the strategic design of MOFs tailored for specific gases. As larger datasets of guest-specific binding sites are developed, this information can be leveraged to train machine-learning models that assess MOF performance or inform material selection for specific applications[53–55].

In a previous study by Vaidyanathan et al.[56], the authors used a combined experimental and computational approach to elucidate the binding interactions of amine-functionalized MOFs. Starting from the X-ray diffraction determined framework structure, atomistic simulations were used to predict the preferred CO_2 binding locations, which were found to be in direct agreement with the binding sites resolved experimentally. The study revealed cooperative binding effects among CO_2 molecules and showed that both dispersion and electrostatic interactions contribute significantly to the overall binding energy. These results demonstrated how accurate identification of binding sites provides mechanistic insight into adsorption and can be used to guide the design of MOFs with optimized interaction environments for enhanced CO_2 uptake.

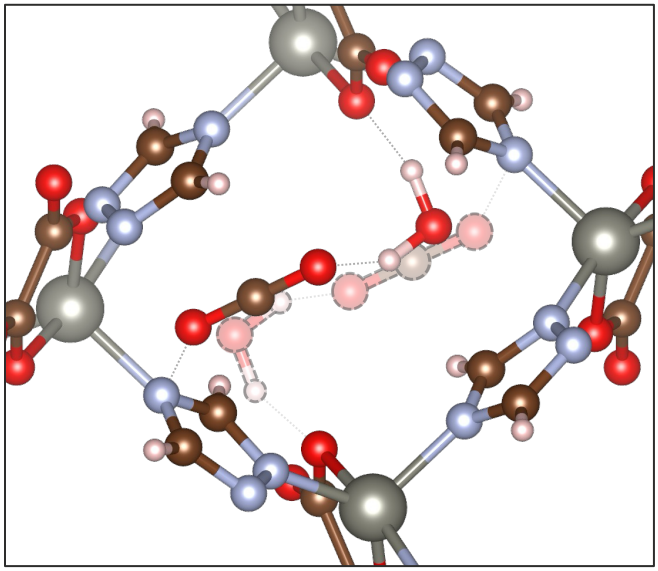

Figure 2.2. Visualization of a competitive binding site in CALF-20 at 5 bar, where both CO_2 and H_2O occupy the same location. The faded (semi-transparent) molecules represent the alternative guest, highlighting the spatial overlap and competition for adsorption at this site.

Similarly, in another study by Ching-hwa Ho et al.[57], the competitive adsorption behaviour of CO_2 and H_2O in CALF-20 was investigated using molecular simulations under varying humidity

conditions. Their results revealed that CO_2 and H_2O molecules compete directly for identical binding sites within the framework, as illustrated in Figure 2.2. The semi-transparent molecules in this figure represent the alternative guest, highlighting the spatial overlap and competitive interaction between CO_2 and H_2O at the same adsorption site. Furthermore, the study found that water significantly influenced CO_2 dynamics, altering both thermodynamic and kinetic properties, thus providing a more straightforward explanation for the anomalies observed in adsorption isotherms. Understanding such competitive interactions is crucial when optimizing MOFs for targeted gas separation processes under realistic conditions.

Extracting binding sites from GCMC-derived probability distributions is a complex task, despite their importance. Raw adsorption probability distributions (APDs) often contain substantial diffuse density associated with weakly interacting or secondary binding sites. Although these regions correspond to lower occupancies, they may still represent chemically relevant adsorption sites, particularly for multi-site guest molecules where one atomic site interacts less strongly with the framework. In such cases, the resulting smeared probability density can overlap with or bias nearby high occupancy regions, complicating the accurate localization of dominant binding sites. Often, the unevenness in the distribution results from both numerical noise and the inherent roughness of the free energy landscape. Even simulations that are well-converged can yield an excessive number of local maxima. For instance, a MOF with fewer than 20 true binding sites can exhibit over 100 distinct maxima in the raw data, resulting in false positives and complicating the interpretation process. In Figure 2.3, the red line reflects the raw output from a stochastic GCMC probability plot (simplified to 1D), whereas the blue line denotes the smoothed distribution achieved through a Gaussian filter. Even converged plots can reveal multiple false maxima due to noise, cluttering the distribution and obscuring true binding sites. Additionally,

grid size and other simulation parameters can affect resolution, leading to artifacts and inconsistencies across different systems. These challenges highlight the necessity for effective post-processing techniques that can smooth and refine the data, facilitating the consistent and accurate identification of chemically significant binding sites. This need is particularly critical for high-throughput workflows or machine learning applications, where reliability and comparability are paramount.

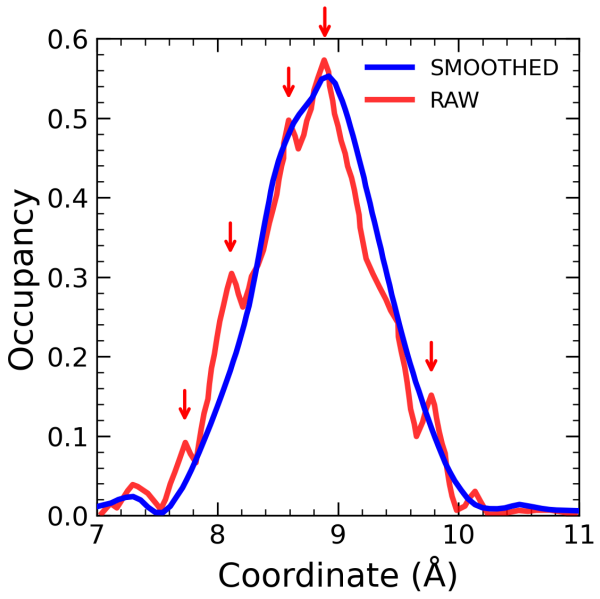

Figure 2.3. 1D probability distribution of the carbon atom derived from GCMC simulation of CO_2 gas adsorption in MOF CALF-15 (red). The probability is plotted along a line which passes through one of the binding sites. The red arrows point to local maxima in the raw probability distribution. The blue line is the result of a "smoothing" of the raw probability distribution with a noise filter (this work).

To address the issues caused by noisy and inconsistent GCMC-derived probability distributions, a reliable algorithm for identifying binding sites must fulfill several key requirements. Firstly, it should consistently differentiate between genuine adsorption sites and false peaks, even in simulations where the raw data reveals many more peaks than the chemically relevant binding locations. This necessitates inherent filtering logic to evaluate peaks based on their statistical significance, allowing users to control the number of binding sites reported by using

a minimal and intuitive set of parameters. For instance, modifying a single parameter should permit users to concentrate on only the most probable areas or widen the search to incorporate larger, less-defined zones. Furthermore, the algorithm needs to be fully automated and compatible with multiple guest molecules to facilitate high-throughput screening with minimal oversight.

Properly addressing periodic boundary conditions is equally crucial. Adsorption sites often form near the edges of the simulation cell, and incorrect treatment can result in split or misidentified regions. Scalability is another essential requirement. With large MOFs surpassing ten million grid points, the algorithm must remain efficient as it scales. Additionally, it should clearly rank or score each predicted site, enabling users to prioritize the most important adsorption areas. Outputs need to be well-organized and compatible with downstream tools, such as periodic DFT codes such as VASP. Lastly, integrated diagnostics should evaluate the sharpness of the probability distribution and identify cases where excessive noise or diffusion undermines the validity of the predictions.

In this study, we present GALP a tool developed to identify adsorption sites directly from probability distributions obtained through GCMC, and to fit complete geometries of guest molecules in those areas. GALP has two operational modes: the case-by-case mode allows users to modify essential parameters to derive chemically precise binding sites suited for specific systems. A second high-throughput mode utilizes default parameters optimized for a wide variety of guest molecules, enabling the quick generation of binding site data with minimal user input. The algorithm has been validated against a benchmark of over 100 MOFs and examined with guest species such as CO_2, N_2, Xe, Kr, C_2H_2, CH_4, C_3H_8, and C_3H_6, for which reference binding sites were manually identified from the corresponding APDs. These adsorbates were selected because they represent widely studied and technologically relevant separation and storage scenarios in the

literature, spanning a broad range of framework-guest and guest-guest interaction systems. These reference sites were then used to assess the algorithm's accuracy and robustness across a wide range of framework chemistries and binding environments.

2.4 Methodology

2.4.1 Adsorbate Probability Distributions

GALP requires the APD for every distinct atomic site of each guest molecule. For example, if CO_2 is the guest, there should be APDs for C and O. The code uses the Gaussian cube format, which stores scalar volumetric data where the simulation cell is divided into N_a x N_b x N_c grid forming voxels, where a, b, and c refer to the lattice vectors and N_x refers to the number of grid divisions along each vector. Each voxel contains the probability of finding the guest atom per unit volume in $Å^{-3}$. The APD is generated by mapping the guest atoms onto voxels at each MC step.

While generating probability distributions from GCMC simulations, we utilized an equitable binning procedure to reduce noise in the resulting distributions. In conventional binning, each atom count is entirely assigned to the voxel it is contained within, as shown in Figure 2.4a. In equitable binning, shares this contribution among the eight closest voxels, with weights based on the atom's proximity to each neighbouring voxel center, as shown in Figure 2.4b. This proportional weighting acts to smooth the distribution, especially in low-occupancy areas. An analysis of the equitable binning in generating the APDs is investigated in the Results and Discussion section. Equitable binning was employed in all simulations throughout this study unless otherwise stated.

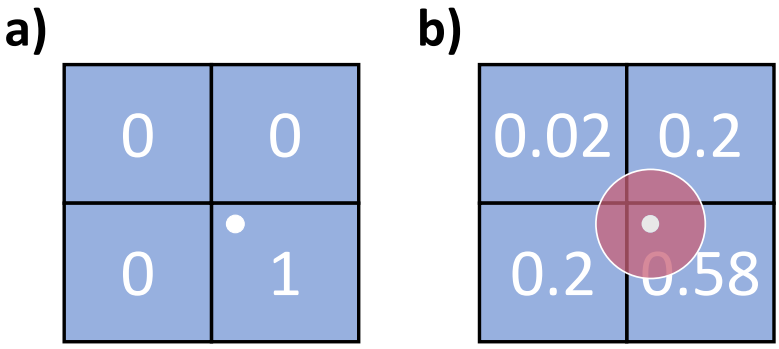

Figure 2.4. A schematic representation of how the position of one atom (the white dot) is distributed to four 2D voxels in a) normal binning procedure compared to b) equitable binning. The numbers in each voxel indicate the contribution of the atom count to the probability distribution in that voxel. The red circular region around the atom position in b) is used to demonstrate the proportional distribution with equitable binning.

2.4.2 Grid Resolution

The resolution of probability distributions is determined by the grid spacing utilized in the GCMC simulations, which reflects the distance between neighbouring grid points along each lattice vector, measured in Å. Smaller grid spacings produce higher-resolution plots, enabling the capture of fine structural details but results in larger data files. Larger spacings result in coarser grids that require less memory to store. While this may result in a loss of local detail, it can also smooth the distributions by averaging guest positions into larger voxels. For all simulations in this study, a default grid spacing of 0.15 Å was employed. The influence and optimization of this parameter are elaborated on in the Results and Discussion section.

2.4.3 Convergence Criteria

The APDs generated from stochastic MC simulations can be noisy unless very long simulations are performed. In GCMC simulations, our experience is that the adsorbate uptake typically converges to acceptable statistical error levels with much fewer MC steps than are required to obtain APDs smooth enough to easily identify the binding sites. While one can simply

run very long GCMC simulations when examining a few materials, in a high-throughput screening workflow, one needs to carefully manage the length of the simulations. Thus, it is critical to define a method to evaluate the convergence of the APDs to stop the simulations when the generated APDs are smooth enough to identify the binding sites. For this, one can evaluate the "smoothness" of an APD to determine convergence. However, the challenge with this approach is to differentiate a noisy APD from one that is intrinsically rugged. Since most simulations involve a supercell, we decided to compare the APD between symmetry-equivalent subcells within the supercell. Symmetry equivalent subcells should ideally have the same distribution and therefore the similarity between APDs of equivalent cells can be used to evaluate the convergence. To evaluate the similarity of two distributions, we use the Tanimoto coefficient as defined in Equation 2.1 where A and B refer two different distributions and $\odot$ is the voxel-wise product. The Tanimoto ranges from 0, indicating two entirely dissimilar distributions, to 1, indicating identical distributions. To provide a more intuitive interpretation, consider a 2 x 2 x 2 supercell where the APD is partitioned into eight symmetry-equivalent subcells. The Tanimoto coefficient is computed between all pairs of these subcells by comparing their voxel grids element by element, which quantifies how consistently the spatial probability density is reproduced across the supercell. High overlap indicates that equivalent regions exhibit the same features, such as peaks occurring at the same positions, while lower overlap reflects discrepancies caused by noise or insufficient sampling. By averaging the Tanimoto over all pairs of equivalent cells, $T_{av,}$ we obtain a single metric that reflects the global consistency of the APD within the supercell. T_{av} is then used as a convergence criterion. We established a minimum T_{av} value of 0.75 as the convergence threshold for all simulations performed in this study. This threshold strikes a practical balance between

simulation duration and the reliability of binding sites, as supported by the validation in the Results and Discussion section.

$$T = \frac{\sum A \odot B}{\sum A \odot A + \sum B \odot B - \sum A \odot B} \quad 2.1$$

2.4.4 Overview of the GALP algorithm

GALP localizes the binding sites from a "converged" set of APDs in the following steps:

i. A Gaussian noise filter is applied to the APDs to further mitigate high-frequency noise in the distributions.
ii. All local maxima are identified from the smoothed distributions. Maxima below a user-defined threshold or within an exclusion radius of a higher occupancy peak are discarded.
iii. For polyatomic guest molecules, maxima (potentially from different atom types) that are roughly in the geometric configuration of the guest molecule are identified and grouped together. The rigid guest molecule is then fit to each group of maxima using a RMSD-minimizing alignment procedure to give the final positions of the binding sites. This process is repeated until no more local maxima can be assigned to binding sites.
iv. The binding energies of the identified binding sites can be optionally evaluated - with or without further geometry optimization.

Upon completion, GALP provides a ranked list of binding site configurations in user-friendly formats, which includes structural coordinates, occupancy values, and binding energies for each guest. In the following, we elaborate on each of the above steps. Any adjustable parameters that are introduced in these steps are highlighted, but the optimization of the parameter values are discussed in a later section.

2.4.4.1 Smoothing of the APD

The use of a Tanimoto convergence test and the equitable binning procedure still generates APDs with noise levels that can impede the identification of binding sites. Thus, we apply a

standard Gaussian noise filter from the SciPy library to isotropically smooth the 3D volumetric data as given in Equation 2.2. σ is a smoothing parameter where larger values lead to more pronounced smoothing effects. A σ value of 0.4 Å is used as a default. The filter handles boundaries using the wrap mode, meaning that probability values at one face of the unit cell are mapped to the opposite face during convolution so that the smoothing respects the periodic boundary conditions of the crystal lattice.

$$G(x, y, z, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x^2+y^2+z^2)}{2\sigma^2}} \quad 2.2$$

2.4.4.2 Local Maxima Identification

After smoothing the APDs, local maxima are identified and filtered. The APD is first rescaled such that the minimum occupancy value is set to 0 and the maximum occupancy value in the APD is set to 1. All maxima in the APD are identified where a maximum is defined as a voxel that has an occupancy greater than all 26 surrounding voxels. Maxima that are lower than a user-defined occupancy cutoff, O_c, are discarded. O_c is set to 10% of the global maximum occupancy or 0.1, by default. Local maxima are also discarded if they are too close to another higher occupancy maximum within the same APD (as opposed to the APD of another atom type). For this we define a user adjustable exclusion radius R_x whose default value is 0.45 Å for most guests. The use of an exclusion radius can potentially eliminate mutually exclusive binding sites. However, we found that an exclusion radius was necessary, particularly in low pressure simulations where high occupancy maxima were often surrounded by many low-occupancy maxima that did not correspond to true binding sites.

2.4.4.3 Fitting of Guest Molecules to Local Maxima

For single site guests such as argon and xenon, the located maxima correspond to binding sites, and no further processing is necessary. However, for multi-site guests, the geometry of the guest must be fit to an appropriate collection of maxima. Since most GCMC simulations of small molecules use a fixed geometry for the guests, GALP utilizes a fixed template structure to fit the guest molecules to the maxima. The first step in the fitting procedure is to determine what collection of maxima are part of the same binding site. For example, for CO_2, one needs to identify a collection of two oxygen maxima and one carbon maximum that are geometrically arranged close to the geometry of a CO_2 molecule. We utilize a procedure that determines what maxima are part of the same binding site in a stepwise fashion. Starting with the first atom in the template structure, the algorithm samples a maximum of the appropriate atom-type, which hasn't been previously assigned to a binding site. It then takes the next atom in the template structure and determines the distance to the previous atom in the template structure. The algorithm then samples maxima of the appropriate atom type until one is found that is within the tolerance of the calculated atom-pair distance. This tolerance is by default set to 0.35 Å. The function continues to "collect" maxima to be part of the same binding site, atom-by-atom until all atoms in the guest are successfully matched to appropriate maxima. During this procedure, if a maximum cannot be found that is within the tolerance of the given atom-pair distance, the binding site is "discarded" and all collected maxima are put back into the pool of available maxima. This process continues until either no more maxima are available to be assigned to a binding site, or no more binding sites can be identified.

Once maxima have been collected into individual binding site groups, the rigid geometry of the template structure is best fit to each binding site group. The first step of the fitting procedure

involves determining the centroid of the binding site maxima and the centroid of the templated guest structure, as shown in Figure 2.5a. The centroid of the template molecule is then placed onto the position of the centroid of maxima as shown in Figure 2.5b. At this point the template structure is not likely to be well aligned with the binding site maxima. The template molecule is then rotated about its centroid to minimize the root mean square distance between the coordinates of the template structure and the positions of the maxima as shown in Figure 2.5c. The rotation matrix that minimizes the root-mean-square distance between the two sets of coordinates is found using the Kabsch algorithm and implemented in the RMSD library[58]. The coordinates of the RMSD fitted template structure of the guest molecule is then output as a single binding site. One can optionally discard poorly fitted binding sites if the RMSD per atom is higher than a user defined threshold, ε.

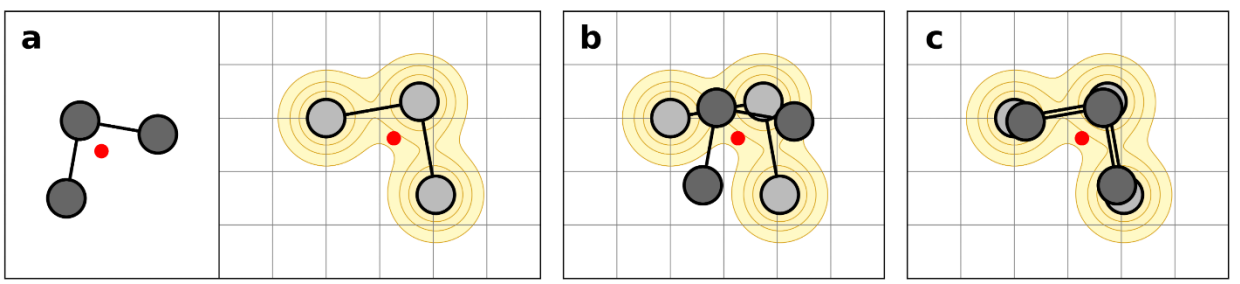

Figure 2.5. Steps of the procedure that fit the template guest molecule structure to the collection of maxima that form a binding site. The grid in a-c represents the probability distribution with the light-yellow contours representing the maxima. The dark grey molecule represents the template guest structure, and the light grey molecule presents the maxima in the APD. a) the centroid (red dot) of the template molecule and the binding site maxima are determined. b) the template guest molecule is placed on the location of the binding site maxima such that the coordinates of the two centroids coincide. c) the template molecule is rotated about the centroid to minimize the RMSD.

2.4.5 Binding Energy Calculation

Once a list of binding sites and their coordinates has been collected, the user can optionally evaluate the binding energy of all sites either with or without further geometry optimization. One would normally evaluate the binding energy using the same force field that the GCMC simulations

were performed with. For this task GALP is currently only compatible with DL-Poly classic. However, one can easily write extensions to GALP such that it creates inputs for other code packages to perform the energy calculation. The binding energy is calculated such that the guest molecule is placed in an empty framework without any other guest molecules. Thus, the computed binding energy only accounts for the interaction between a single guest molecule and the frozen framework. It is important to realize that the potential energy surface is not necessarily a good surrogate for the free energy distribution that is represented by the APD, particularly at high pressure where guest-guest interactions can be important.

2.4.6 Provided Outputs

GALP provides users with key information on all accepted binding sites. A *.cif* file is created with the framework atoms and the coordinates of all identified binding sites. An .xyz file is also written with each binding site, sorted by the occupancy from high to low. Here, each frame of the file gives the location of a single binding site. In the title line of the *.xyz* file, the binding energy is provided broken down in the electrostatic energy and van der Waals energy.

2.4.7 Implementation

The GALP program is written in the Python 3 language and uses standard, open-source scientific libraries. The main functionality utilizes Pymatgen's[59] molecule and structure objects, while SciPy[60] is used for applying smoothing filters and locating local maxima. Molecular geometries are fitted using routines from the RMSD library[58]. GALP was initially designed to be compatible with APDs generated by our in-house GCMC code, FastMC version 1.4.0[61]. This Fortran-based GCMC code produces separate three-dimensional probability distributions for each

atom type in the guest molecule, allowing GALP to locate and fit distinct atomic binding sites independently. FastMC was modified to support several aspects of this study. Specifically, we implemented an on-the-fly convergence checker that dynamically extended the number of production steps until a user-specified Tanimoto threshold was met. Most of the functions of GALP can easily be extended for use for other GCMC codes, as it can function given only a set of APDs in Gaussian cube format. Compatibility with RASPA3 was subsequently implemented, enabling direct analysis of APDs generated from RASPA3 simulations.

2.5 Computational Details

2.5.1 GCMC Simulations

All GCMC simulations were performed with our in-house GCMC code, FastMC that is derived from the DLPOLY classic MD package. The guest molecules were held rigid, and a fixed framework approximation was used for all simulations. The guest-host interactions are composed of pair-wise point charge electrostatic and Lennard-Jones potentials. The partial atomic charges on the MOF framework atoms are derived from a single-point periodic DFT calculation using the REPEAT method. For the DFT calculations, the VASP package[62] was utilized with a PBE exchange-correlation functional. Further details of the DFT calculations can be found in reference 63. The Lennard-Jones parameters for the framework molecule were taken from the Universal Force Field (UFF)[64]. Guest molecules were assigned partial atomic charge and Lennard-Jones parameters from various literature force fields: CO_2[65], N_2[66], C_2H_2[67], NO[68], Xe[69], Kr[69], CH_4-TraPPE[70,71], H_2O (TIP5P/TIP4P-ew)[72,73], C_3H_6[74] and C_3H_8[70]. Simulation conditions used to optimize the GALP parameters are described in the Results and Discussion section. Supercells

were used such that all atom-atom distances within the cell exceeded 12.5 Å. GCMC simulations were run until the Tanimoto threshold detailed in the Methods section was reached.

2.5.2 Entropy Metric

In addition to the Tanimoto based convergence criterion, we implemented an entropy-based metric to assess the spatial diffusivity of each APD. This indicator reflects whether guest occupancy is concentrated in well-defined regions or dispersed across the framework. To quantify this, we compute the Shannon entropy H_A of the normalized three-dimensional occupancy distribution $p_{i,j,k}$, where i, j, and k index the voxel position in the grid.

$$H_A = -\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\sum_{k=1}^{n_z} p_{i,j,k}\ln\left(p_{i,j,k}\right) \quad 2.3$$

Here, n_x, n_y, and n_z are the grid dimensions, and $p_{i,j,k}$ is the normalized probability associated with each voxel. For a perfectly uniform distribution, where each of the $N = n_x n_y n_z$ voxels have equal probability $p_{i,j,k} = 1/N$, substitution in equation 2.3 yields the maximum entropy:

$$H_{max} = \ln(N) \quad 2.4$$

To enable direct comparison between systems discretized on grids of different sizes, a normalized entropy ratio was defined as:

$$M = \frac{H_A}{H_{max}} = \frac{H_A}{\ln(N)} \quad 2.5$$

This ratio ranges from 0 to 1, where values approaching 1 indicate highly diffuse, delocalized probability distributions and values near 0 correspond to strongly localized binding behaviour. From a practical standpoint, low entropy implies the presence of distinct binding sites suitable for molecular fitting. In contrast, high entropy indicates poor spatial resolution and

potential difficulty in assigning guest configurations. While not used as a convergence criterion, this metric serves as a diagnostic tool. In particular, it complements the Tanimoto score by highlighting cases where simulations appear to converge but fail to yield clearly defined binding regions due to weak or non-specific guest-framework interactions. This metric was used during the manual fitting of binding sites, where it was found that entropy values greater than 0.9 generally corresponded to diffuse distributions that may introduce significant challenges in identifying chemically meaningful binding configurations.

2.6 Results and Discussion

2.6.1 Development set of MOFs

To optimize GALP's ability to identify binding sites and to create a robust set of default adjustable parameters, we curated a development set of 100 MOFs for which the APDs were used to manually validate the GALP binding sites. The structures in the development set were sampled from the ARC-MOF[63] database, which contains about 280K hypothetical and experimentally characterized MOF structures with a diverse range of organic and inorganic SBUs, geometries, topologies and functional groups. Importantly, ARC-MOF has been screened for structural errors that have recently been identified to make up a sizable fraction of many other MOF databases[75]. It is important to note that the purpose of this development set is to optimize the GALP parameters and evaluate the algorithm's robustness across a diverse range of adsorption environments, rather than to validate the method against experimental measurements. The objective is to assess whether GALP can reliably identify maxima and reconstruct binding sites from the underlying APDs produced by GCMC simulations, treating the APD itself as the reference representation of the adsorption landscape. For this reason, the dataset intentionally includes both experimentally

characterized and hypothetical MOFs. Hypothetical structures significantly expand the accessible structural diversity, enabling the exploration of a broader range of pore sizes, topologies, and binding environments than is currently available in purely experimental databases. Throughout this chapter, MOFs are identified using the structure filenames provided in the ARC-MOF database, including their native DB## identifiers where applicable. Experimentally characterized MOFs are typically referenced using crystallographic identifiers associated with their reported structures, whereas hypothetical MOFs are labelled according to their constituent building blocks, topology, or generation scheme within ARC-MOF. This naming convention is used consistently in all figures and tables to uniquely reference each structure. The MOFs in the development set were selected from the ARC-MOF database using a farthest point sampling method[76], considering features such as pore size, limiting diameter, and topology. These descriptors were chosen because they directly control accessible adsorption volume, steric confinement, and connectivity of adsorption environments, which are the primary structural factors governing binding-site location and multiplicity in porous materials. Since the composition of ARC-MOF is dominated by computer generated structures, there are large collections of MOFs that are functionalized variants of a single parent structure. Thus, in our sampling of structures, if a newly sampled structure possessed the same parent structure of a previously selected MOF, the newly sampled structure was discarded. Furthermore, since we wanted our development set to contain ~15% experimental MOFs, and such MOFs only make up 2.59% of ARC-MOF, we first performed farthest point sampling on the experimental MOFs only. Once 15 experimental MOFs were selected, sampling was extended to the full database to fill the remaining 85 structures, again discarding any structures that shared the same parent framework.

The plots shown in Figure 2.6 show a comparison of various geometric features of MOFs in the GALP development set compared to the complete ARC-MOF database. With this development set, we observed that the GCMC computed uptakes for N_2 were very low giving highly delocalized APDs with poorly defined binding sites. Thus, for N_2 we generated a guest specific development set of 10 MOFs from the ARC-MOF database that possessed high N_2 uptakes. For this subset we first identified the top 100 MOFs in ARC-MOF with the highest simulated N_2 uptakes. We then applied farthest point sampling of geometric parameters to select 10 MOFs.

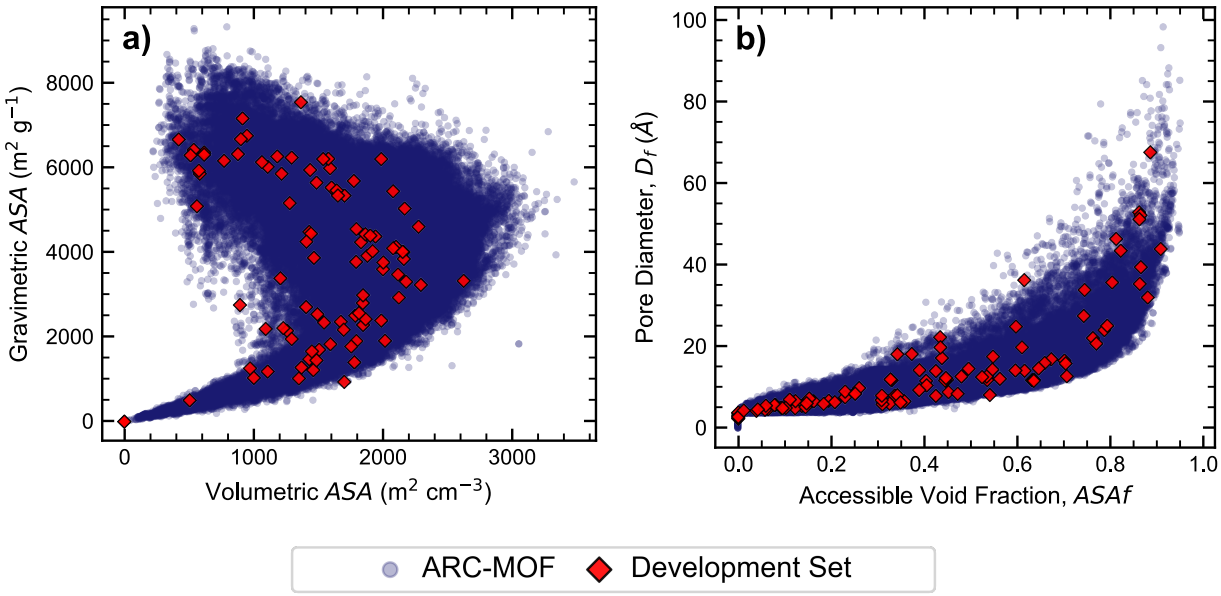

Figure 2.6. Distribution of various geometric parameters of MOFs in the GALP development set compared to the full ARC-MOF database. (a) Plot of the gravimetric versus the volumetric accessible surface areas for the two data sets. (b) Plot of the pore diameter versus the accessible void fraction of the MOFs in the two datasets.

2.6.2 Optimization and Validation of GALP Parameters

2.6.2.1 Optimization of the GALP Parameters

The determination of the GALP parameters followed the workflow shown in Figure 2.7, which outlines the sequential optimization of maxima extraction parameters, including the Gaussian smoothing parameter (σ), occupancy cutoff (O_c), and radius cutoff (R_x), followed by fitting parameters overlap tolerance (OV), and RMSD (ε). At each step, we followed a simple guiding principle, ensuring that all binding-site maxima visible in the unsmoothed, folded APDs were captured while avoiding underfitting that would introduce false sites, or overfitting which would remove true sites. The manually identified maxima from the folded APDs served as the closest available "ground truth" for this optimization. These APDs, generated directly from the underlying GCMC density grids, provided a benchmark against which to judge whether the chosen parameter set successfully reproduced the expected number and location of binding sites. We did not adopt ab initio calculations as a reference because DFT identifies minima on a potential surface corresponding to the zero Kelvin limit. As such, it does not account for finite temperature effects, entropic contributions, or guest-guest interactions that are intrinsic to adsorption under experimental conditions. Moreover, the DFT potential energy surface can differ substantially from the force field potential employed in the GCMC simulations, making direct comparison ambiguous and physically inconsistent. Experimental binding site data were also not used as a primary benchmark. While such measurements are scarce and typically limited to systems with small pores and strong guest-host interactions, the more fundamental limitation is that a direct comparison would primarily assess the accuracy of the underlying simulation approximations rather than the performance of GALP itself. Any disagreement between simulation and experiment binding site

would combine errors arising from force field choice, fixed framework assumptions and charge models with potential limitations of the extraction and fitting algorithm, preventing an unambiguous evaluation of the GALP's ability to identify binding sites from APDs. In contrast, the raw, unsmoothed APDs produced directly by GCMC simulations are available for every system, contain no fitting bias, and reflect the actual sampling of the guest in the framework. For these reasons, the folded APDs provide the most objective and universally applicable internal benchmark for validating GALP, allowing us to ask whether the algorithm can faithfully recover the maxima present in the underlying probability distributions.

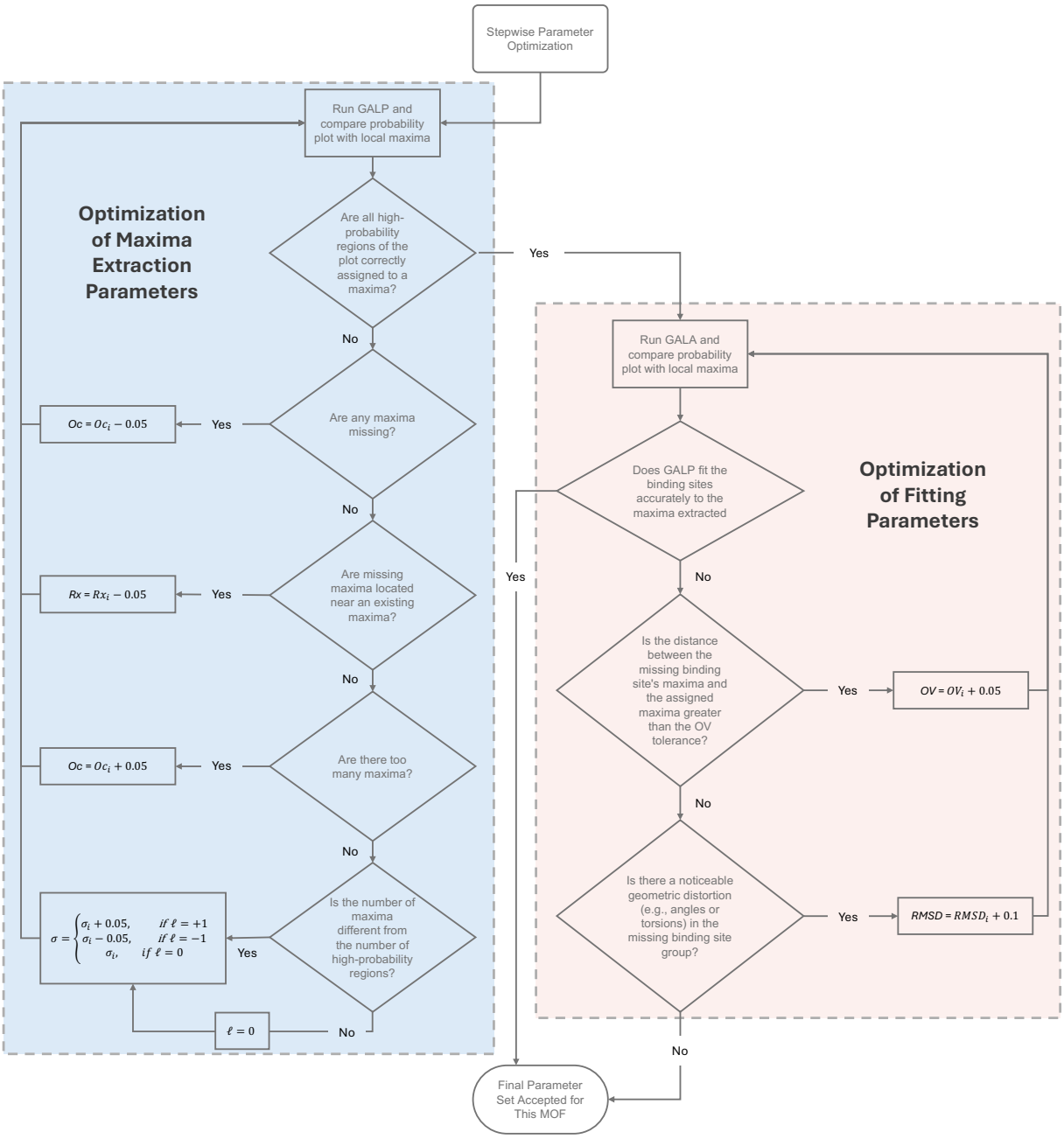

Figure 2.7. Flow chart illustrating the optimization of parameters for each guest molecule in the development set. Five tunable parameters were considered: σ (Gaussian smoothing), R_x (effective radius), O_c (occupancy cutoff), OV (overlap tolerance for molecule building), and RMSD (fitting cutoff). The variable $\boldsymbol{\ell}$ is a diagnostic flag used to evaluate the number of detected maxima relative to the folded probability plot $\boldsymbol{\ell = +1}$ indicates the presence of excess maxima, while $\boldsymbol{\ell = -1}$ indicating that some maxima are missing.

The first stage of parameter optimization focused on maxima extraction, where the Gaussian smoothing width (σ), occupancy cutoff (O_c), and effective radius (R_x) were adjusted. The

goal of this stage was to ensure that all binding-site maxima present in the GCMC-derived APDs were correctly identified by GALP. The key consideration was whether the current parameter set produced the expected number of maxima. Missing peaks indicated underfitting, while excess peaks reflected overfitting. The second stage of parameter optimization addressed the fitting of guest molecules to the extracted maxima, governed by the overlap tolerance (OV) and the RMSD (ε) cutoff. Both parameters were initially set at highly restrictive values to ensure that only chemically reasonable configurations were accepted. In cases where maxima could not be fit under these strict conditions, the cutoffs were gradually relaxed upward, allowing for successful placement without introducing artificial binding sites. This step ensured that the fitted configurations not only reproduced the locations of the probability-density maxima but also maintained geometrical consistency with the underlying molecular models. By combining the maxima-extraction and fitting stages, GALP produced guest-specific parameter sets that faithfully captured the most occupied binding sites across the diverse set of MOFs.

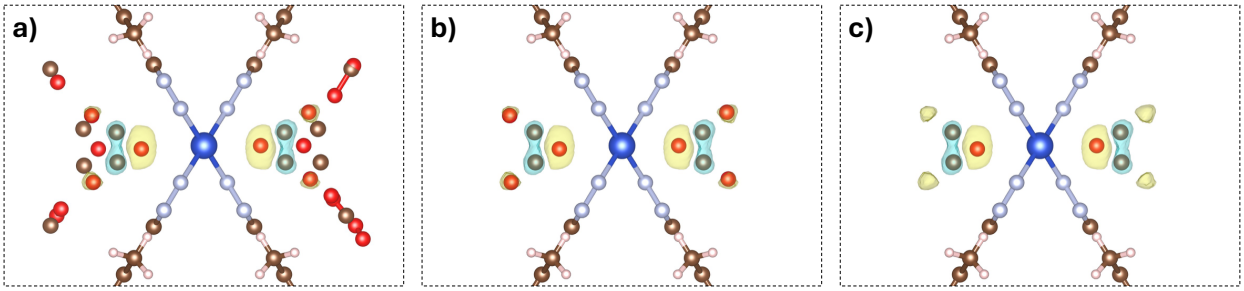

Figure 2.8. Manual fitting of CO_2 adsorption maxima in DB1-Cu2N8-ADC_B-DPAC_B_No1223. Brown and red spheres denote the carbon and oxygen atoms of CO_2, while cyan and yellow isosurfaces represent the corresponding APDs. Panels (a–c) illustrate the effect of parameter choice on the number of extracted maxima.

As illustrated in Figure 2.8, the outcome of maxima extraction depends strongly on the chosen parameter set. In this example, the occupancy cutoff (O_c) was varied, a low value (O_c = 0.01) introduced false maxima in the output (panel a), while a high value (O_c = 0.3) suppressed genuine O_x sites (panel c). Only at an intermediate setting (O_c = 0.2) where all true maxima

recovered (panel b), demonstrating the need to balance sensitivity with noise suppression. This behaviour is not unique to this MOF-guest pair. Because the dataset spans a wide range of frameworks, weak MOF-adsorbate interactions often generate low-intensity noise in the probability plots, which can be misinterpreted as additional maxima. Fine-tuning the parameters for each system yields a distribution of optimized values, from which generalizable parameters can be inferred for high-throughput screening of a given guest.

2.6.2.2 Validation of Algorithm

The reliability of the optimized parameter sets was assessed by comparing the number and locations of binding sites identified by GALP with those determined manually from the same APDs. This manual identification served as the internal benchmark, providing a direct test of whether the algorithm faithfully reproduced the maxima present in the raw APDs. The parity plots in Figure 2.9 demonstrate strong quantitative agreement in the expected number of binding sites across all guest species. Linear regression of GALP-expected versus manually determined binding-site numbers per structure yielded slopes very close to unity, with R^2 values ≥ 0.98. RMSE values ranged from 0.46 to 1.50 binding sites per structure, with most cases close to 0.5 or below, confirming that the optimized parameters recover the expected number of sites with high accuracy. This level of agreement is expected, since the parameters were tuned individually for each MOF-guest system to establish guest-specific defaults and to demonstrate that GALP can, in principle, reproduce the manually identified maxima with near-perfect accuracy. Occasional deviations from perfect parity were observed in systems with very low uptake or highly delocalized densities, but these were rare and confined mainly to challenging guests such as C_3H_8 and C_3H_6.

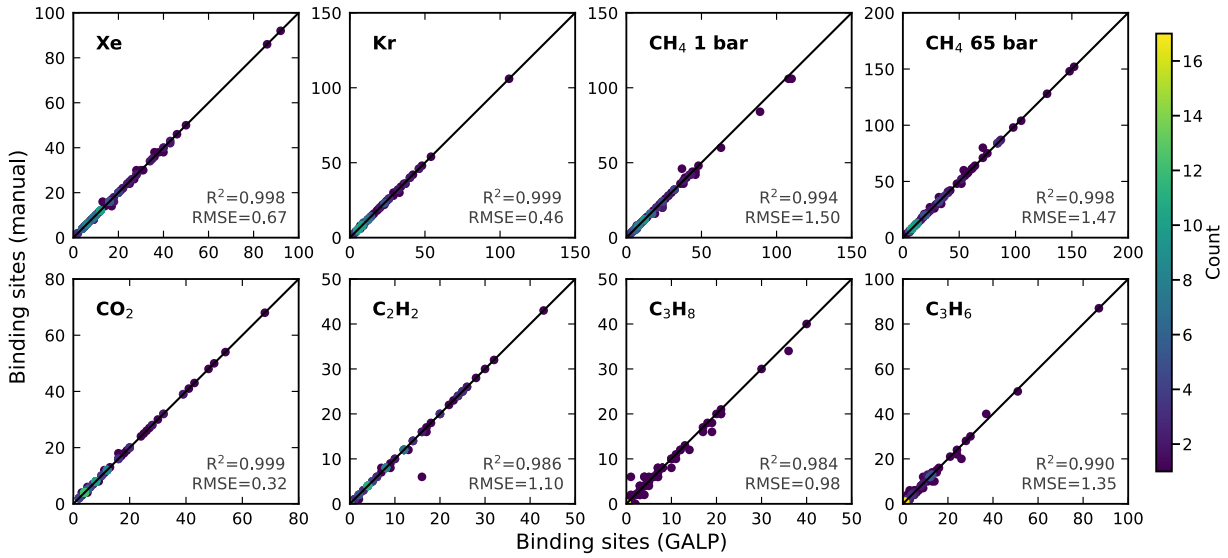

Figure 2.9. Parity plots comparing the number of binding sites identified by GALP and manual labels from folded probability plots. Black lines indicate one-to-one correspondence; regression metrics (R^2 and RMSE) demonstrate strong agreement across all guests, with only minor deviations in low-uptake or delocalized cases.

While the analysis above demonstrates that the algorithm can reliably extract the expected number of binding sites on a case-by-case basis, complete validation also requires assessing the spatial accuracy of the fitted site positions. Figure 2.10, therefore, complements the numerical parity analysis by presenting a structure-resolved comparison between the fitted binding site coordinates and the corresponding folded APDs for a small set of DB12 MOFs and guests. By overlaying the fitted sites directly onto the APDs, this figure enables a direct visual assessment of whether the extracted binding sites' position coincides with the underlying probability maxima sampled by GCMC.

Across most systems, the fitted binding sites are found to align closely with well-localized regions of high probability density for all guest sites, indicating that the algorithm accurately captures both the location and orientation of the preferred adsorption configurations. An illustrative exception is observed for DB12-NEYZAU with CO_2, where the folded APD exhibits carbon density at multiple locations, while the corresponding oxygen density is absent or partially

localized at some of these positions. For multi-atom guests, GALP assigns a binding site only when a complete molecular configuration is supported by coincident probability density for all atoms. As a result, isolated carbon density without accompanying oxygen density is not fitted as a binding site, even when visible in the APD. This behaviour reflects a constraint designed to avoid assigning chemically inconsistent adsorption geometries and to ensure that only physically meaningful binding site positions are fitted.

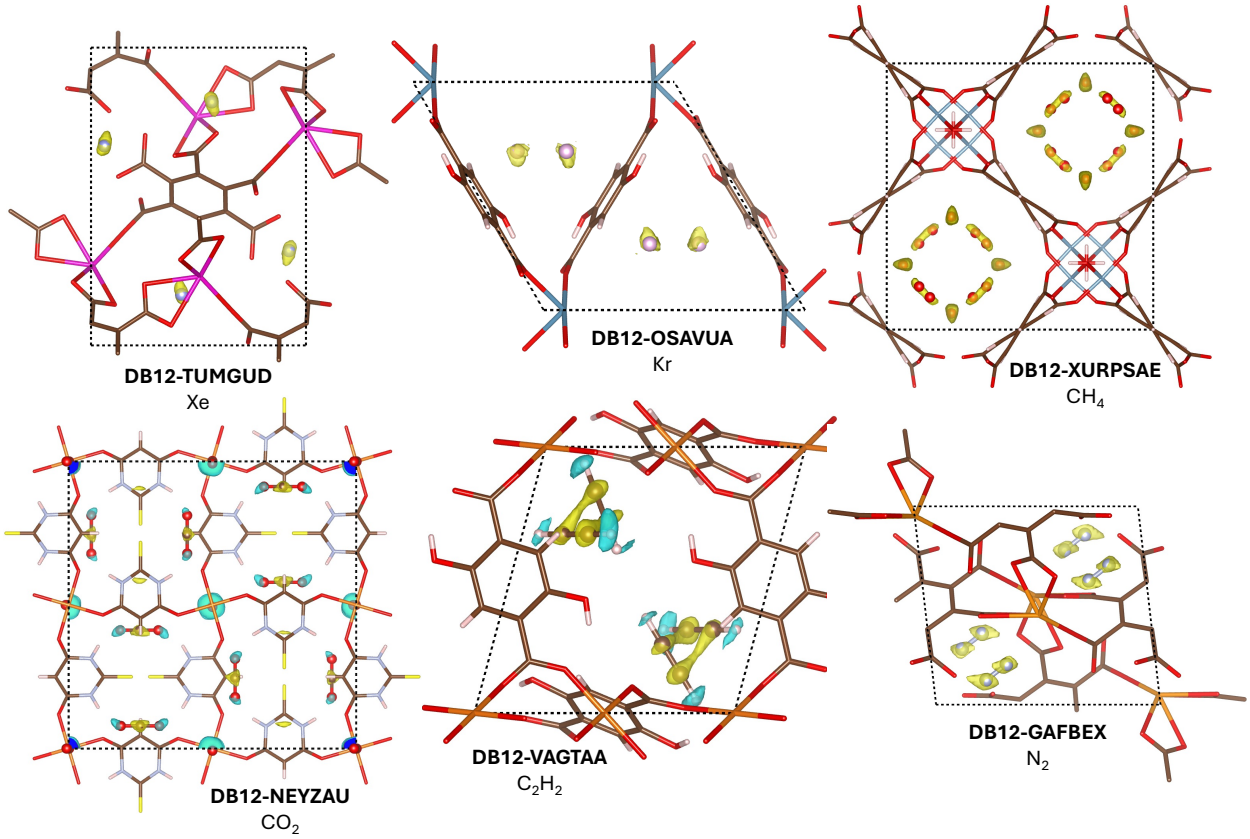

Figure 2.10. Fitted binding sites and corresponding folded adsorption probability distributions for six DB12 MOFs included in the validation of the algorithm. The six guest molecules shown are Xe, Kr, CH_4, CO_2, C_2H_2, and N_2, as labelled. All structures correspond to experimental DB12 entries from the ARC MOF database. All adsorption probability distributions were generated from GCMC simulations performed at 298 K and 1 bar.

2.6.2.3 Optimized Parameters

The final recommended set of optimized parameters for each guest are summarized in Table 2.1. These values represent practical defaults that reproduce the manually identified maxima under the representative temperature and pressure conditions examined in this work, spanning C_2H_2, C_3H_6, C_3H_8, CH_4, CO_2, Kr, Xe and N_2 at temperatures of 298-373K and pressures between 0.75 and 65 bar. To complement the tabulated values, we also examined the distributions of optimized parameters across the development set as represented in Figure 2.11.

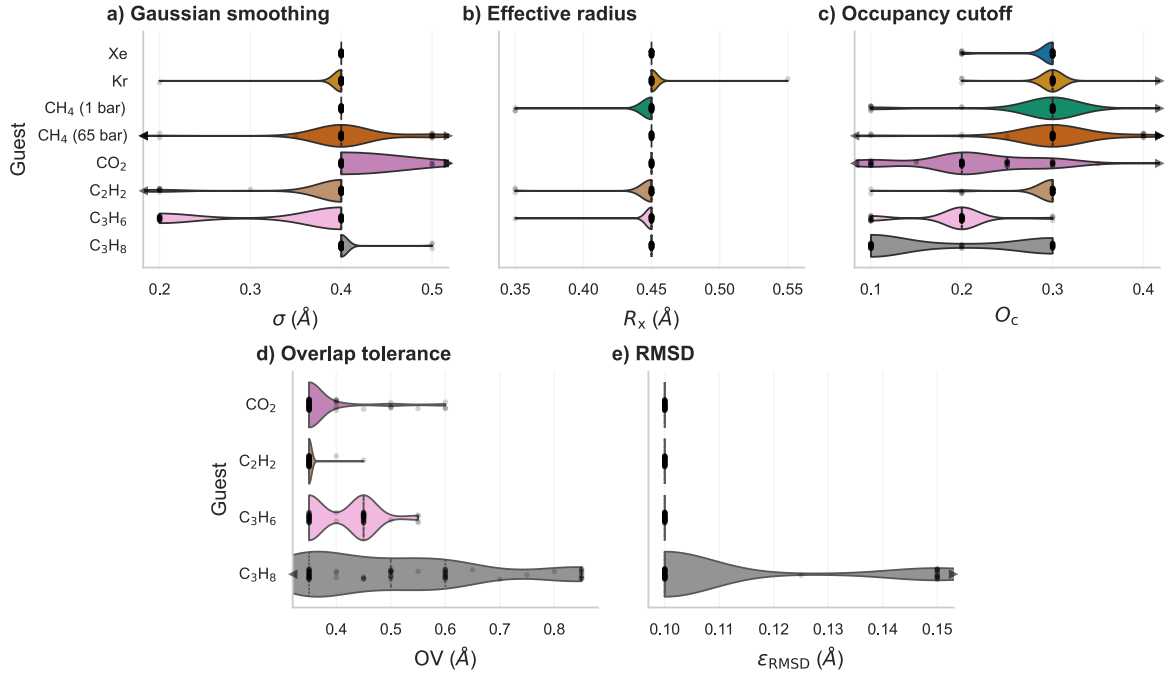

Figure 2.11. Distribution of optimized fitting parameters for each guest molecules in the development set. Panels (a-e) show the fitted values of the Gaussian smoothing parameter (σ), effective radius (R_x), occupancy cutoff (O_c), overlap tolerance (OV), and RMSD (ε), respectively. The distributions reflect the range and consistency of parameter values selected during validation, providing insight into the sensitivity of each parameter to the underlying guest–framework interactions.

The violin plots reveal that σ and R_x were highly consistent across all systems, reflecting the imposed convergence criterion (Tanimoto ≥ 0.75), which intrinsically enforces similarity within the APD and reduces sensitivity to the precise values of these smoothing parameters. In contrast, O_c displayed greater variability because it reflects local occupancy thresholds rather than

global distribution overlap. For single site guests such as CH_4, Xe, and Kr, the optimized occupancy cutoff consistently converged to 0.3. This uniformity reflects the simplicity of these guests, which allows for well-defined maxima to be generate, giving rise to a higher cut-off than with polyatomic guests. By contrast, multi atom guests required lower cutoff values to accommodate more complex spatial distributions. For CO_2, a cutoff of 0.2 was required because rotational freedom of the molecule leads to broader sampling of the terminal atoms, resulting in increased noise in the adsorption probability distribution. In even more complex cases, such as propylene, the lack of molecular symmetry meant that each atomic site was sampled differently, and O_c values as low as 0.1 were required. The overlap tolerance (OV) also showed marked guest dependence. For most guests, OV remained within a narrow range, but significantly larger values were required for propane and propylene. This reflects the more delocalized character of their adsorption density profiles, site maxima for these guests were, on average, spaced farther apart than for others, necessitating a looser tolerance during molecular building. These broader distributions of probability density will be revisited in the limitations section, where we discuss guest-specific limitations in greater detail.

Table 2.1. List of optimized, guest specific GALP parameters and conditions for which GCMC simulations were run on the development set to optimize the parameters.

Guest	T (K)	P (bar)	σ (Å)	O_c	R_x (Å)	Ov (Å)	RMSD (ε) (Å)
Xe	298	1	0.4	0.3	0.45	-	-
Kr	298	1	0.4	0.3	0.45	-	-
CH_4	298	1	0.4	0.3	0.45	-	-
CH_4	298	65	0.4	0.3	0.45	-	-
CO_2	298	1	0.4	0.2	0.45	0.35	0.1
C_2H_2[a]	298	1	0.4	0.3	0.45	0.35	0.1
C_3H_6	373	1	0.2/0.4	0.2	0.45	0.35/0.45	0.1
C_3H_8	373	1	0.4	0.1/0.3	0.45	0.35-0.60	0.1
N_2	298	0.75	0.4	0.3	0.45	0.35	0.1

[a]*The guest molecule was fitted using the GALP option that restricts the fitting procedure to heavy atoms only.*

2.6.2.4 Recovery Accuracy and Guest-Dependent Optimized Parameters

The recovery accuracy obtained using the optimized GALP parameters depends primarily on the adsorption behaviour of each guest rather than on limitations of the fitting algorithm itself. As shown in the parity plots in Figure 2.12, guests that form well-defined adsorption basins are recovered with the highest fidelity, whereas systems characterized by weak or spatially diffuse interactions exhibit greater variability.

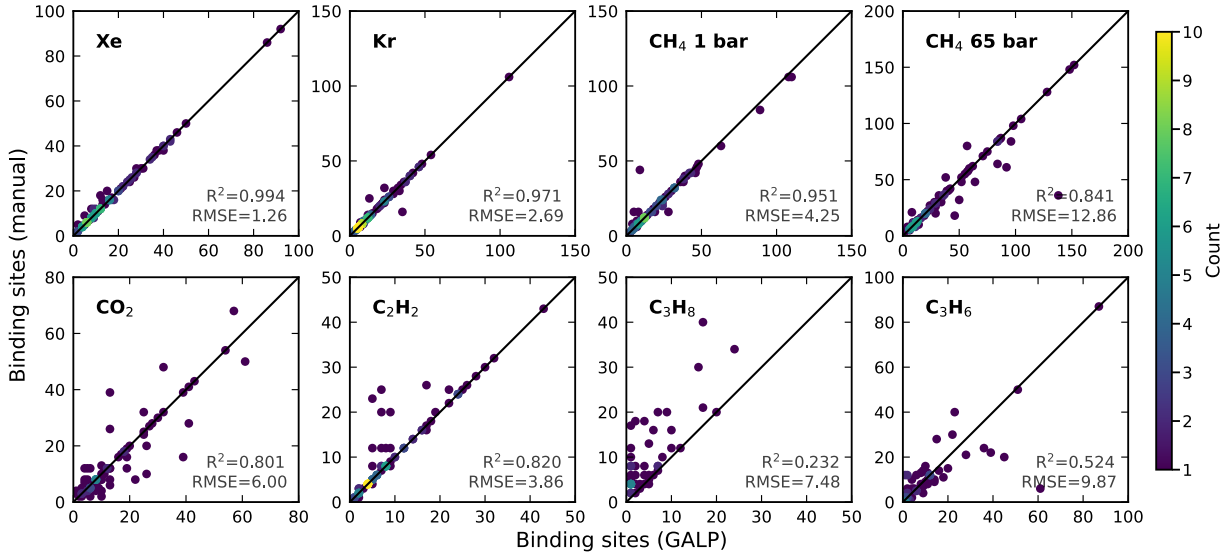

Figure 2.12. Parity plots comparing the number of manually identified binding sites with those recovered by GALP using optimized parameters for each guest. Each point represents a MOF-guest pair, and black lines indicate one-to-one correspondence. Strong agreement is observed for single-site guests and systems with localized adsorption basins, while deviations increase for multi-site guests and weakly interacting systems characterized by diffuse probability distributions. Binding-site counts are reported without accounting for symmetry equivalence.

Single-site guests such as Xe, Kr, and CH_4 at 1 bar display near-ideal parity between manually identified and automatically recovered binding sites (Figure 2.12). These systems involve simple interaction potentials and limited orientational degrees of freedom, making the recovery largely insensitive to small positional variations in the APDs. Multi-site guests, including CO_2 and C_2H_2, require the correct placement and orientation of multiple interaction centres and are therefore more sensitive to local variations in pore geometry and probability density shape.

Although CO_2 often exhibits strong binding in frameworks with open metal sites, this behaviour is not universal across chemically diverse MOFs. As reflected by the increased scatter observed for CO_2 in Figure 2.12, structures with weaker host-guest interactions or broader adsorption landscapes yield higher configurational entropy and reduced spatial confinement, which lowers recovery robustness.

The influence of pressure is further illustrated by CH_4 at elevated loading. At high pressure, CH_4 populates secondary and tertiary pore regions where adsorption free-energy minima are shallow and overlapping, leading to more diffuse APDs and increased deviation from parity relative to the low-pressure case. On the other hand, C_3H_6 and C_3H_8 remain the most challenging guests, exhibiting the largest deviations from parity in Figure 2.12. Their weak and non-specific interactions generate broad APDs with limited structural definition, consistent with their highly entropic adsorption behaviour.

Importantly, all binding-site counts reported in Figure 2.12 include every recovered site without accounting for symmetry equivalence. In highly symmetric frameworks, missing a single symmetry-unique site may therefore lead to an apparent undercount amplified by the number of equivalent sites. Overall, Figure 2.12 demonstrates that the optimized GALP parameters recover binding sites with high fidelity when APD are spatially localized and chemically meaningful. Reduced recovery accuracy arises primarily in systems characterized by weak binding and elevated entropy, which lie outside the conditions where precise binding-site identification is most informative.

2.7 Limitations

2.7.1 Convergence

A key aspect of GALP is that the APDs from GCMC simulations must be sufficiently converged for reliable maximum extraction. We set a Tanimoto threshold of 0.75 as a minimum criterion to ensure convergence of the APDs, but achieving this often required very long simulations. For some guests, such as Kr, C_2H_2, C_3H_6, and C_3H_8, run times sometimes exceeded 15–20 days. Guests that did not reach convergence within this period were excluded from the guest-specific subset.

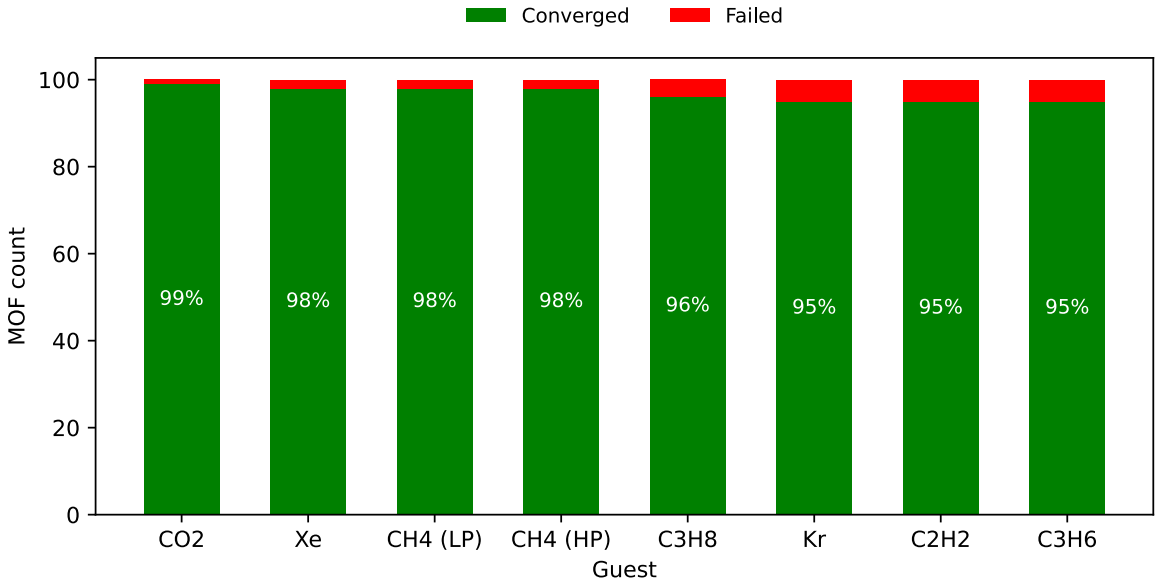

Figure 2.13. Convergence rates for all guest simulations across the full MOF set. The bar chart reports the percentage of structures that reached the Tanimoto convergence criterion for each guest, along with the fraction of remaining failures. The results show consistently high convergence, with most guests exceeding ninety five percent across the dataset.

Overall, as shown in Figure 2.13, 774 of 800 guest–MOF pairs (96.75%) converged within 20 days, with 26 (3.25%) not converging. Convergence failures depended on the guest: 1% for CO_2, 2% for CH_4 and Xe, 4% for C_3H_8, 5% for propylene and Kr. For C_3H_6 and C_3H_8, the weak interactions with the framework result in broad, shallow adsorption landscapes, making it harder

for simulations converge. These guests are over-represented among the failures. For Xe and Kr, behaviour differs according to their Lennard-Jones parameters. Xe, with larger σ and a deeper ε well, creates a more attractive, confined adsorption landscape. Kr, with smaller σ and weaker ε, interacts less strongly and tends to have more diffuse distributions slowing convergence, especially in low-uptake MOFs. This explains why some MOFs converged for Xe but not Kr. C_2H_2 also shows a similar effect due to its short-range interactions; hydrogen sites lack Lennard-Jones potential, so only carbon atoms provide meaningful dispersion forces. Hydrogens only interact electrostatically, leading to more rotational configurations and a broader sampling space. In larger, low-interaction MOFs, this limited contact points forces the simulation to explore a very diffuse landscape, requiring more sampling for convergence.

2.7.3 Delocalization

One of the main challenges in fitting APDs is when the distribution is highly delocalized or diffuse. Delocalization is characterized by a smeared probability distribution that spans a continuous region of the framework, providing little distinction between potential binding sites. Figure 2.14 illustrates varying degrees of delocalization, from diffuse to well-localized distributions. Each contour plot corresponds to a slice along the z-axis taken at the position of the highest identified maxima. No smoothing was applied to the APDs shown. Figure 2.14a represents a strongly delocalized case, where four distinct regions of elevated probability density merge into a continuous distribution. Although certain areas exhibit slightly higher intensity, the lack of well-defined maxima prevents the algorithm from identifying a unique adsorption site. This behaviour directly affects the fitting process, as the algorithm must operate over broad, low-gradient regions rather than sharply peaked densities. In contrast, Figure 2.14b shows a moderately localized

example with two symmetry-equivalent maxima highlighted in yellow. Here, the high-probability regions are clearly separated and easily distinguishable from the surrounding density, simplifying the fitting procedure and improving identification of meaningful adsorption sites. A secondary, weaker maximum is also visible in a lower-density region, corresponding to a less occupied binding site. Figure 2.14c depicts a well-localized case characterized by sharply defined maxima that correspond to strongly interacting binding sites. Two site types are evident, one slightly more populated than the other, both easily identifiable by the algorithm. In systems with multiple sites per guest, delocalization becomes a particularly severe challenge. Extracting several broad or overlapping distributions often leads to inaccurate site identification. The fitting algorithm may miss physically relevant sites altogether or generate redundant results, producing artificially duplicated results. This limitation is unavoidable for systems with weak binding or near-degenerate adsorption sites, where delocalization naturally arises from the underlying free-energy landscape. The relative entropy values for these examples, 0.94 for (a), 0.82 for (b), and 0.89 for (c), quantify this trend. Lower relative entropy (< 0.90) corresponds to more localized distributions, confirming that the metric effectively captures delocalization within the adsorption probability maps.

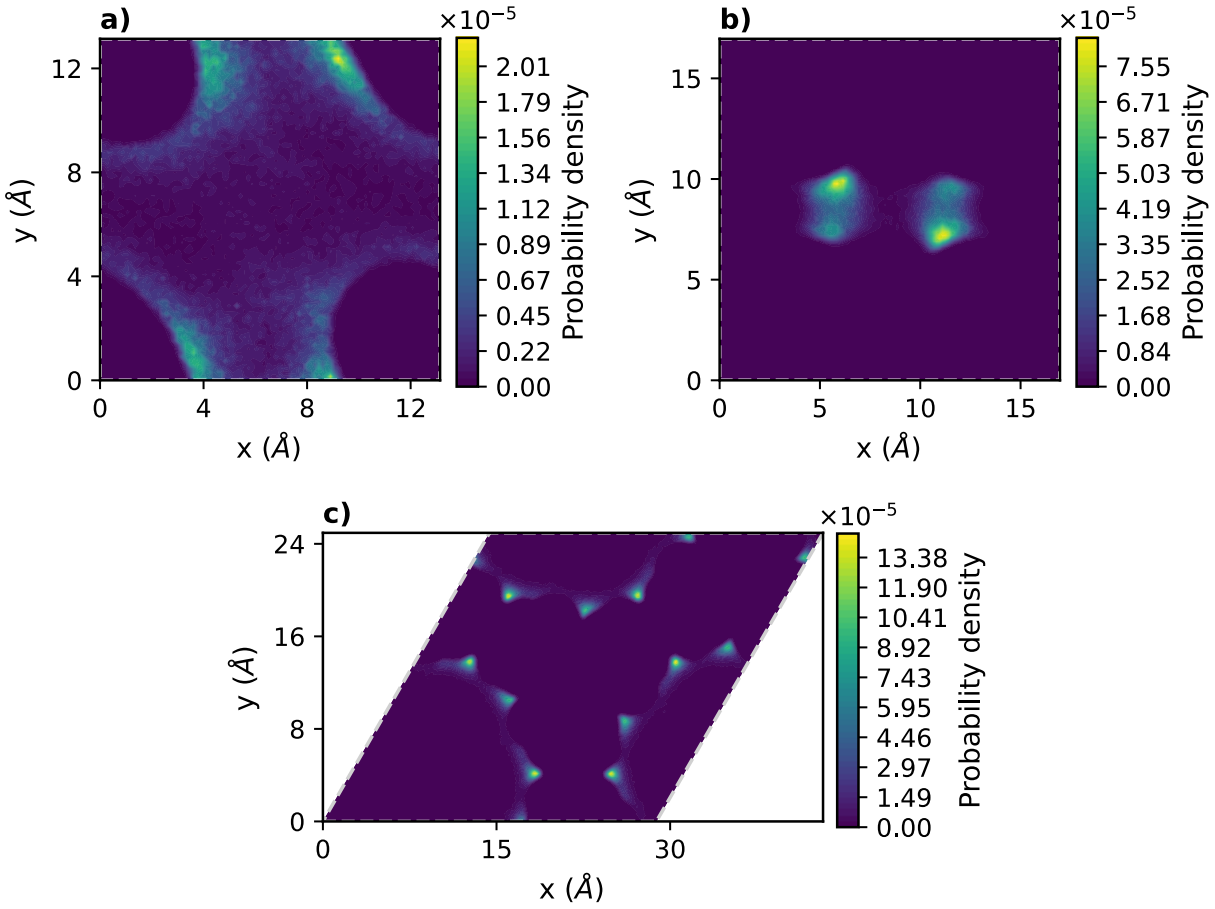

Figure 2.14. Contour plots of (a) DB13-cds-Syn027206 (delocalized), (b) DB12-NEYZAU_clean (localized), and (c) DB1-AlO6-DPAC_A_No7 (localized) showing probability density slices along the z-axis at the respective maxima. All examples correspond to the CM site in the CH_4 guest model at 65 bar and 298 K. The relative entropy values are 0.94, 0.82, and 0.89 for (a), (b), and (c), respectively. The unit cell is outlined with a light gray dotted line.

The delocalization typically occurs in two scenarios. Figure 2.15 demonstrates a clear correlation between a MOF's affinity for a specific guest and the spatial confinement of the adsorbate within the framework. The plot compares the heat of adsorption (HOA) with the mean relative entropy (H/H_{max}), where each point represents a distinct MOF-guest pair. The negative correlation (Spearman $\rho = -0.671$) shows that systems with more localized adsorption probability distributions (lower entropy) exhibit stronger interactions with guest molecules and therefore higher HOAs. In contrast, frameworks that allow broader, more delocalized guest distributions, which correspond to higher entropy, tend to show weaker adsorption energies. The 95%

confidence interval of [-0.718, -0.622] demonstrates that this correlation is both statistically significant and consistent across the dataset. The relatively narrow range indicates that the observed relationship is not driven by a few outliers but instead reflects a general physical trend linking configurational entropy and adsorption strength. The color of the data points in Figure 2.15 represents the accessible surface area (ASA), which further reinforces this relationship. MOFs with large pore volumes and high ASA values generally allow greater guest mobility, leading to higher entropy and weaker binding, whereas smaller or more confined structures restrict guest movement, resulting in lower entropy and stronger host-guest interactions. This trend highlights that both framework topology and pore accessibility jointly determine the balance between entropic freedom and enthalpic stabilization of adsorbed species.

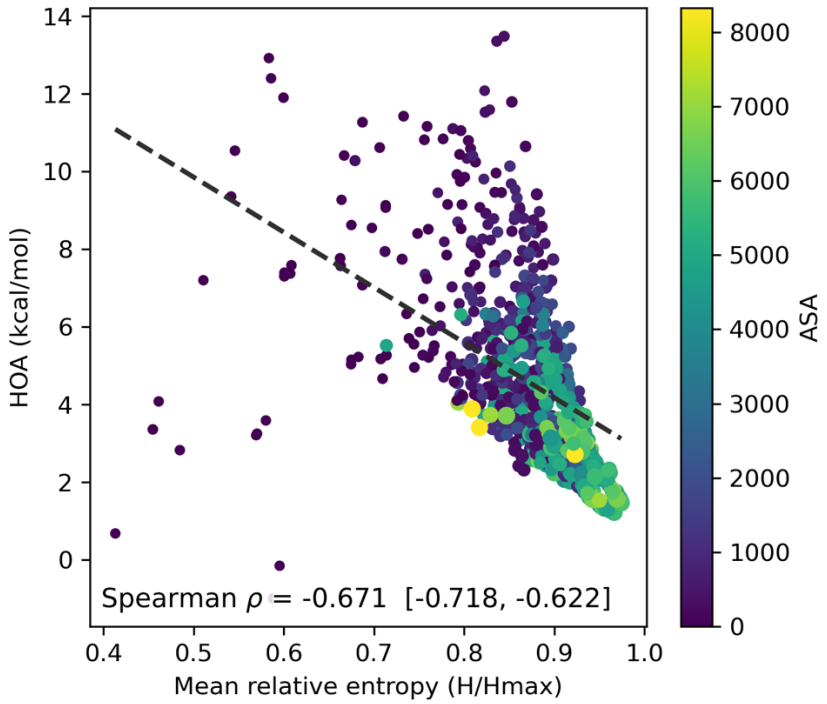

Figure 2.15. Correlation between the mean relative entropy (average of each guest's adsorption density profiles) and the heat of adsorption (HOA), with colours indicating the accessible surface area (ASA). The size of each marker reflects the diameter of the largest included sphere (Di).

2.8 Evaluation of GCMC Specific Parameters

2.8.1 APD Resolution and Grid Spacing

The grid spacing parameter is crucial in determining the resolution of GCMC-generated APDs, which in turn controls GALP's ability to identify binding sites. A finer grid spacing offers higher resolution and more precise localization of binding sites, but at the cost of larger data files and higher compute requirements for post-processing. In this study, we employed a grid spacing of 0.15 Å for all simulations, initially chosen to balance accuracy and computational cost. We evaluate this choice by comparing the binding sites and their energies obtained with grid spacings of 0.50, 0.15, and 0.01 Å for three systems: CO_2 adsorption in CALF-15 and MAF-2, and CH_4 adsorption in Sc_2-bdc_3, under the same conditions used in the corresponding experiments. (MAF-2: 195 K and 20.3 bar; CALF-15: 173 K and 0.85 bar; Sc_2-bdc_3: 230 K and 9 bar, see Table 2).

We first compare the locations of the binding sites obtained with the different grids. Across the three MOFs, a total of 50 distinct binding sites were obtained with the finest grid of 0.01 Å, in agreement with the experimentally determined binding sites (15 in CALF-15, 18 in MAF-2, and 16 in Sc_2-bdc_3). For this comparison, we do not treat the experimental binding sites as the ground truth, because that would implicitly assume that the GCMC simulations employ an exact potential energy surface. Instead, we regard the APD binding sites obtained with the 0.01 Å grid as the most accurate within our test cases and use those as the reference for all comparisons. With a grid spacing of 0.15 Å, GALP recovers all 50 binding sites. The overall average RMSD of all atoms over all binding sites, relative to the 0.01 Å reference, is 0.11 Å. With the coarser grid of 0.50 Å, GALP recovers 48 of the 50 binding sites, missing two equivalent binding sites in MAF-2 (16 out of 18). The overall average RMSD increases to 0.43 Å, roughly four times larger than for the 0.15

Å grid. Even so, an RMSD of 0.11-0.43 Å remains smaller than the uncertainty in many experimentally determined binding site positions.

To further assess the fidelity of the binding sites obtained with the 0.50 and 0.15 Å grids, we compare their binding energies to those computed with the 0.01 Å APD grid. For the 0.15 Å grid, the overall mean absolute deviation (MAD) in binding energy across all 50 sites is 0.14 kcal mol^{-1}, with per MOF MADs of 0.15 kcal mol^{-1} for CALF-15, 0.10 kcal mol^{-1} for MAF-2, and 0.16 kcal mol^{-1} for Sc_2-bdc_3. The maximum deviation remains below about 0.7 kcal mol^{-1}. We regard this as very strong agreement in terms of binding energetics. In contrast, for the 0.50 Å grid the overall MAD increases to about 2.0 kcal mol^{-1}, with per MOF MADs of 2.79 kcal mol^{-1} for CALF-15, 1.00 kcal mol^{-1} for MAF-2, and 2.18 kcal mol^{-1} for Sc_2-bdc_3, and maximum deviations exceeding 5.5 kcal mol^{-1}. These results show that a grid spacing of 0.15 Å yields binding sites that are both geometrically and energetically very close to those obtained with a grid that is 15 times finer, while a coarser grid that is only about 3.3 times larger (0.50 Å) leads to binding energies that are likely to be judged unacceptably different from the reference. A visual representation is shown in Figure 2.16.

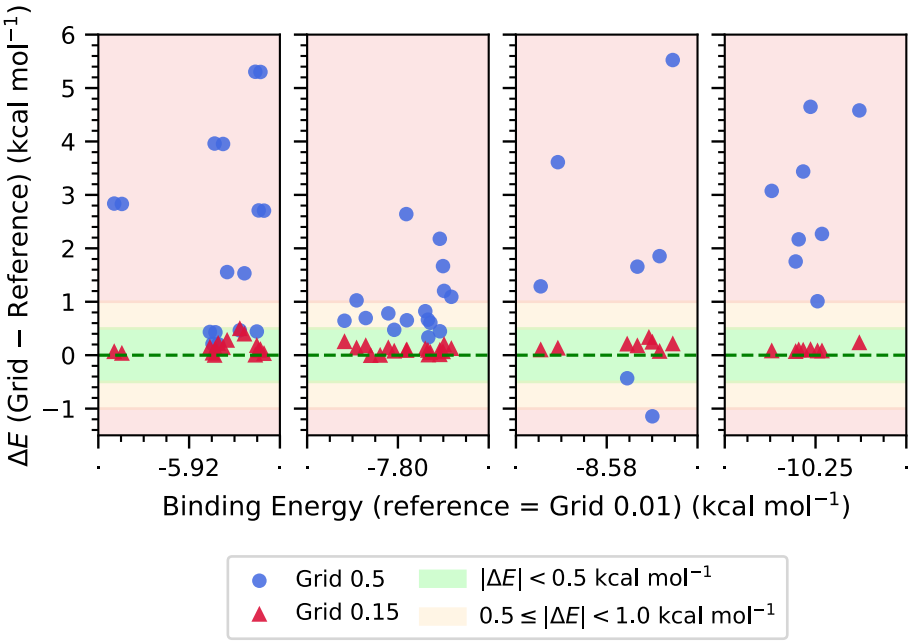

Figure 2.16. Residual binding energies relative to the reference grid spacing of 0.01 Å are shown for coarser grid spacings of 0.5 Å and 0.15 Å. Each subpanel corresponds to a cluster of binding sites extracted from the overall distribution, which includes CALF-15, MAF-2, and Sc_2-BDC_3, representing distinct binding environments for CO_2 and CH_4 guests. The shaded regions indicate energy deviation thresholds: green for $|\Delta E| < 0.5$ kcal mol^{-1} and beige for $0.5 \leq |\Delta E| < 1.0$ kcal mol^{-1}. Values outside these regions (pink) correspond to deviations large enough to potentially alter binding-site ranking or occupancy behaviour.

As expected, reducing the grid spacing significantly increased storage and processing requirements. The probability plot files for the 0.01 Å grid were approximately 2500% larger than those for the 0.15 Å grid. Although we did not expect the GCMC run times to differ significantly, using a grid of 0.01 Å did result in a GCMC simulation that was on average 7% slower for the three MOFs tested than for the 0.15 Å grid. However, as expected, the run time of GALP almost doubled with the 0.01 Å grid as Pymatgen's cube file reader, struggled with the much denser APD grid. These findings suggest that a 0.01 Å grid is costly for high-throughput applications.

These results show that an APD grid of 0.15 Å is balanced, allowing accurate localization of binding sites without being overly demanding computationally. For a small test set of binding sites, this grid size reproduced the binding sites produced by a grid 15 times finer, as quantified on

spatial and energetic terms. In this way, this grid size is a reasonable default or starting point if further refinement is required for a specific guest or host.

2.8.2 Effect of Equitable Binning

In this work, we used equitable binning to generate the probability distributions. Unlike standard binning, which assigns the full "count" to a single voxel, equitable binning distributes the "count" across multiple voxels based on proximity to the center of each voxel as shown in Figure 2.4. While straightforward to implement, not all MC packages may use equitable binning to generate the APDs. Thus, we examine the impact of equitable binning to smooth APDs and reduce the number of MC steps required to reach convergence. To ensure a direct comparison, we modified our GCMC code to generate both normal and equitably binned APDs within the same GCMC simulation.

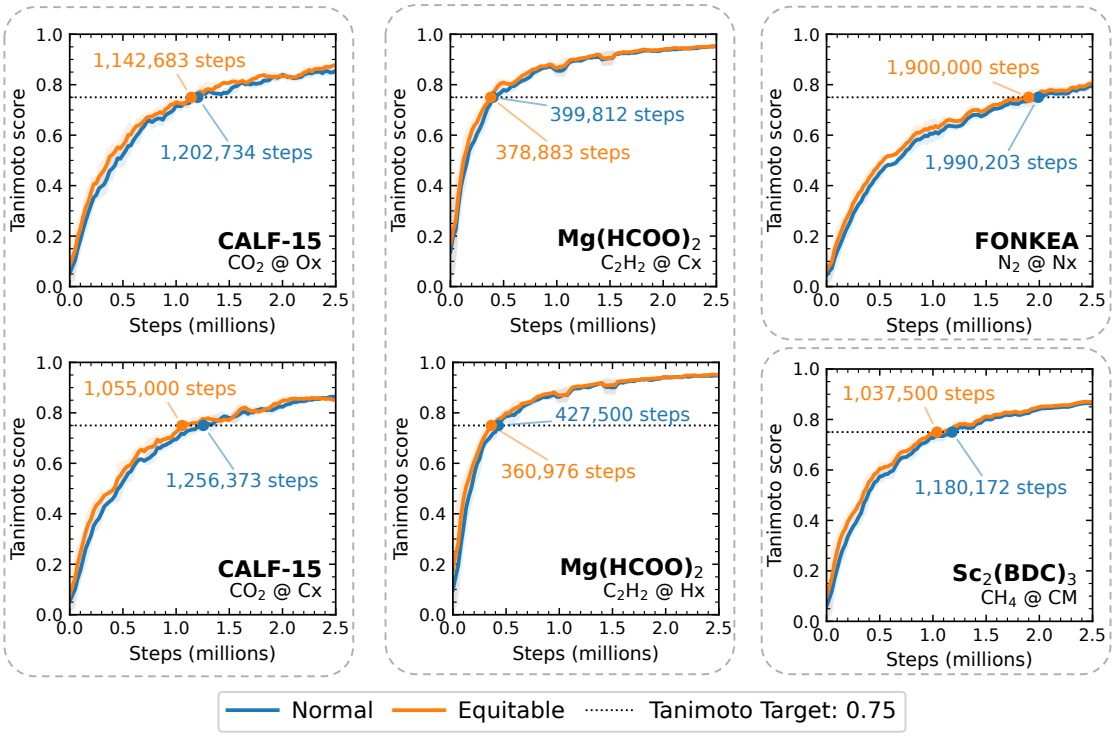

Figure 2.17. Comparison of Tanimoto score convergence for six representative MOF-guest systems using normal and equitable binning. Each panel reports the Tanimoto score as a function of production steps and

highlights the total number of steps required to reach a Tanimoto score of 0.75. Normal binning is shown in blue and equitable binning in orange.

The plots in Figure 2.17 show that equitable binning gives a modest improvement in the convergence of the APDs. Fewer GCMC steps are required in most systems. The effect is usually small, so the absence of equitable binning is not a major limitation. In a few cases where the Tanimoto score begins to plateau before the convergence threshold, the improvement becomes more noticeable. This behaviour appeared in one of the acetylene examples. At this stage, it is not obvious which features of the MOF, the guest, or the adsorption landscape control early or late convergence. Equitable binning introduces no penalty because it does not increase the APD's memory usage or the simulation's computational cost, and while it offers only limited advantages, it can be used when available. However, GALP does not require equitable binning to function correctly and remains fully applicable to APDs generated by codes that do not implement this feature.

2.9 Convergence and Diffusivity Metrics in GALP

GALP utilizes two primary metrics, the Tanimoto criterion and entropy, to assess the quality of adsorption probability distributions. The Tanimoto criterion, described in Section 2.3.3, quantifies the similarity of occupancy distributions across equivalent cells within the supercell. High Tanimoto values indicate that the probability distribution is consistently sampled and that the GCMC simulation has reached convergence. Importantly, however, the Tanimoto criterion only measures sampling convergence and does not guarantee that localized or chemically meaningful binding sites can be identified. In cases where guest–framework interactions are weak, adsorption probability distributions may be intrinsically diffuse, leading to broad regions of occupancy despite high Tanimoto values. To address this limitation, GALP employs a second

metric, entropy (see Section 2.4.2), which quantifies the degree of spatial delocalization in the probability distribution by evaluating the uniformity of voxel occupancies. High entropy values correspond to diffuse distributions, whereas low entropy values indicate well-localized binding regions. Entropy, therefore, provides complementary information that enables assessment of binding site localization beyond convergence alone. By combining the Tanimoto criterion to evaluate sampling convergence with entropy to assess spatial localization, GALP offers a more complete and physically meaningful evaluation of binding site quality.

2.9.1 Evaluation of the Tanimoto Convergence Criterion

In high-throughput workflows, it is important to have a convergence criterion to stop simulations once the generated APD is sufficiently converged. If one stops the simulation too early, the resulting APD may exhibit many false maxima due to noise. On the other hand, allowing the simulation to proceed unnecessarily long may make a high-throughput screening impractical. To determine if the APDs have converged enough to reliably localize binding sites, we have opted to use the Tanimoto similarity between the APDs of equivalent subcells of the simulation cell. The closer the Tanimoto is to unity, the more similar the distributions are and the more converged we consider the APD to be. For this work, we have used an average Tanimoto between all equivalent regions of the APD(s) to be 0.75 as the convergence criteria. To evaluate this criterion, we examine the binding sites obtained from a range of Tanimoto thresholds from 0.10-0.90 and how they differ from the binding sites obtained from a simulation where a threshold of 0.95 is used (considered the ground truth for this analysis). For this test, we have examined the CO_2 adsorption in 8 MOFs, 4 with relatively localized maxima in the APDs, and 4 that have delocalized APDs. Figure 2.18 reinforces this distinction by showing how the number of identified binding sites varies with the

Tanimoto threshold. Localized systems remain stable above 0.5, while delocalized systems fluctuate until much higher thresholds, highlighting their sensitivity to incomplete convergence.

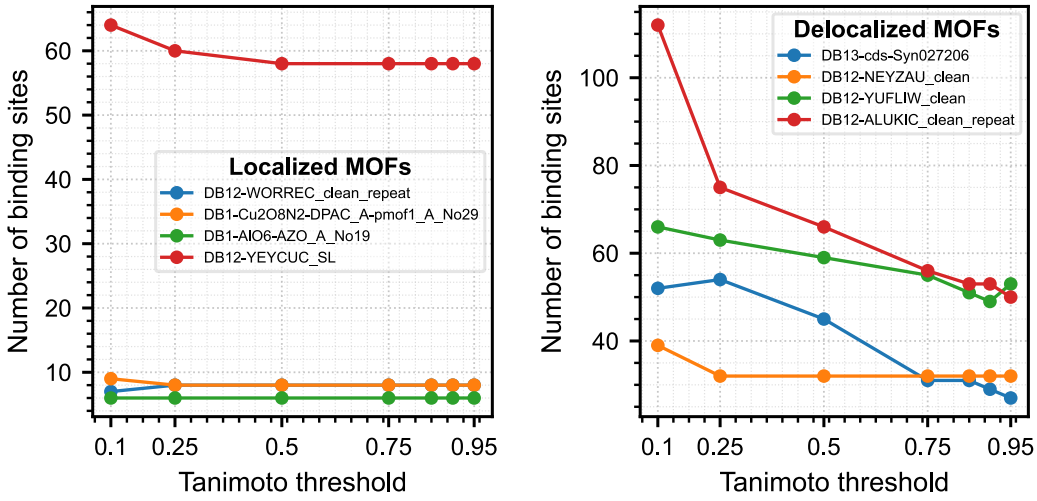

Figure 2.18. Number of binding sites identified by GALP as a function of the Tanimoto threshold for delocalized and localized adsorption distributions. Localized systems show stable behaviour, with the number of sites remaining effectively constant once the threshold exceeds 0.5. Delocalized systems display large variations at low thresholds and begin to stabilize only around 0.75, although minor deviations persist even at high thresholds between 0.9 and 0.95. These trends illustrate the inherent sensitivity of delocalized probability distributions and reinforce the limitations discussed earlier.

We adjusted the FastMC code to progressively increase the number of simulation steps until the desired Tanimoto value was achieved. This approach enabled us to create a controlled set of probability plots at various Tanimoto thresholds: 0.10, 0.25, 0.50, 0.75, 0.85, 0.90, and 0.95. We chose eight MOFs for our analysis: four with localized binding, characterized by distinct maxima, and four with delocalized behaviour, showcasing broad, diffuse occupancy. Using the same GALP settings for each MOF at every Tanimoto level, we extracted binding sites and compared their predicted binding energies, occupancies, and site counts.

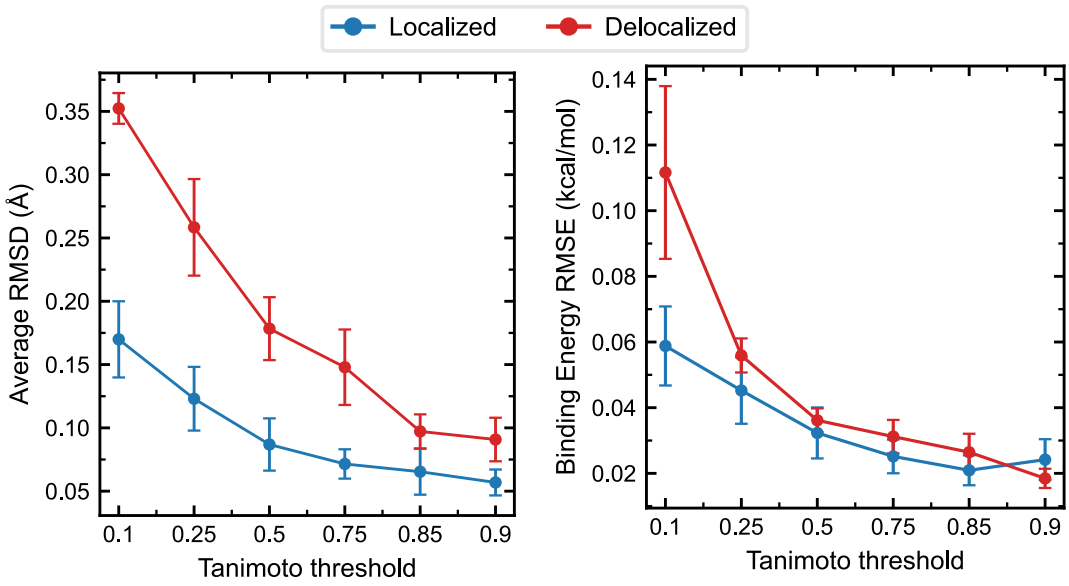

Figure 2.19. Average RMSD and binding energy RMSE as a function of Tanimoto threshold for localized and delocalized probability maps. Error bars denote the standard error of the mean.

Figure 2.19 shows that localized MOFs exhibit minimal changes in binding energy when the Tanimoto threshold exceeds 0.50. This suggests that shorter simulations can yield consistent and converged binding sites for these systems. The reproducibility of these findings reinforces the notion that highly localized systems, characterized by minimal rotational variation and strong guest-framework interactions, create distinct and reliable occupancy profiles even with moderate simulation lengths. In comparison, delocalized MOFs exhibit significantly greater variability in binding energy at lower Tanimoto thresholds (0.10-0.50). Though there is some enhancement between 0.75 and 0.90, considerable fluctuations persist until very high Tanimoto values are achieved. This suggests that delocalized systems necessitate more comprehensive sampling to identify meaningful binding configurations and are more responsive to the length of simulations and convergence criteria.

Users should strive for the highest practical Tanimoto threshold to ensure dependable convergence. Nonetheless, the best value ultimately hinges on the specific application. In cases where basic binding-site identification suffices, such as rapid screening, a Tanimoto value around

0.80 usually balances accuracy and computational efficiency. Conversely, for applications requiring higher precision, particularly those involving detailed structural or energetic assessments, a stricter threshold in the range of 0.85 to 0.90 is advantageous. This range improves the quality and accuracy of the extracted binding sites, particularly in systems with diffuse binding characteristics, by enabling more reliable identification of subtle variations in guest orientation and site location.

2.9.2 Evaluation of the Entropy Metric

The entropy metric should not be interpreted as a standalone criterion for assessing feasibility of binding site fitting. Instead, it functions as a diagnostic indicator that provides complementary information to the Tanimoto similarity score. While the Tanimoto metric quantified the degree of overlap and convergence between subcells of the APD, it does not capture how spatially diffuse the underlying APD is. In contrast, the entropy metric directly reflects the degree of delocalization of the sampled guest configurations within the framework.

High entropy values indicate that APD is distributed broadly across the accessible pore space, rather than concentrated in a small number of well-defined regions. In such cases, the difficulties in fitting binding sites do not arise from poor convergence of the GCMC simulation, but from the absence of chemically meaningful localization. Conversely, low entropy values correspond to specially confined APD that typically give rise to distinct adsorption maxima and are therefore well suited for molecule fitting. When fitting remains problematic despite low entropy, the underlying cause is more likely related to limitations of the force field, charge assignment, or the simulation setup itself.

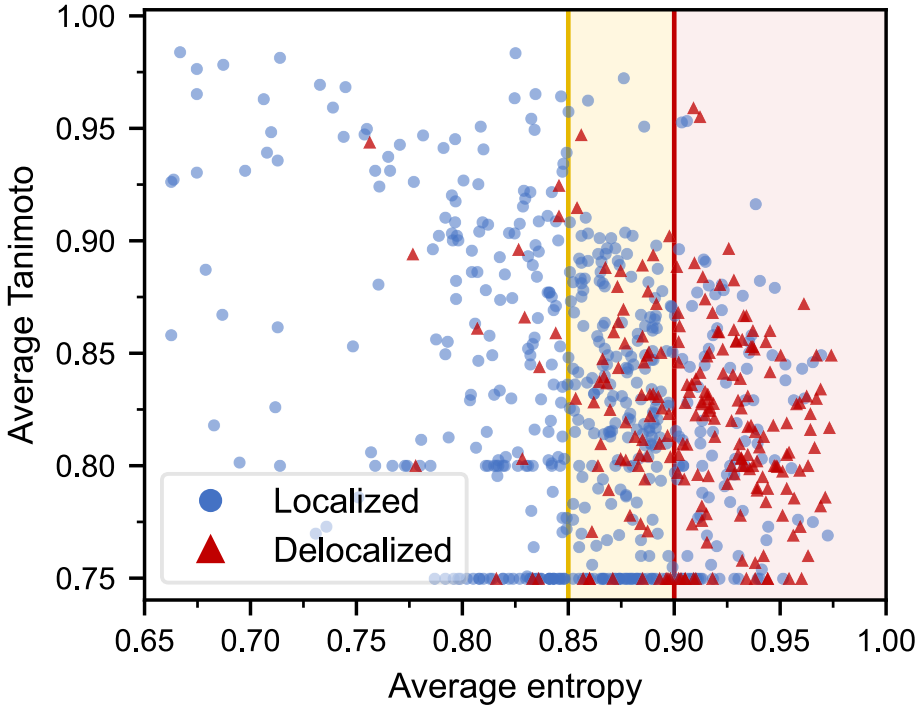

Figure 2.20. Relationship between average entropy and average Tanimoto similarity for all MOF-guest pairs considered in this work. Blue circles represent localized systems with reliably converged fitting, while red triangles indicate delocalized cased identified through manual classification due to unreliable fitting. Shaded regions denote the caution (yellow) and danger (red) entropy regimes, with vertical lines indicating the corresponding threshold.

To assess the practical implication of entropy for binding site identification, Figure 2.20 presents the relationship between the average normalized entropy and the average Tanimoto similarity for all MOF-guest pairs considered in this study. The data are partitioned into three regions: a low entropy "good" region (< 0.85), and intermediate "caution" region (0.85 ≤ and < 0.90), and a high entropy "danger" region (≥ 0.90). These thresholds were not chosen arbitrarily but were motivated by systematic trends observed across the dataset and by empirical experience during manual fitting.

In the low-entropy region, delocalized cases are rare. Only 14 out of 247 MOF-guest pairs are classified as delocalized, corresponding to a delocalized-to-localized ratio of 0.06. This indicates that entropy values below 0.85 are strongly associated with well-localized adsorption

probability distributions and reliable binding-site fitting. In the intermediate region, localized behaviour remains dominant, with a delocalized-to-localized ratio of 0.29. This region represents a transitional area in which adsorption distributions begin to broaden, but fitting remains feasible in most cases.

A qualitative change in behaviour is observed for entropy values above 0.90. In this high-entropy region, delocalized cases become prevalent, with 140 delocalized and 105 localized MOF-guest pairs, corresponding to a delocalized-to-localized ratio of 1.33. Importantly, however, a substantial fraction of localized cases persists even at high entropy. This demonstrates that elevated entropy does not inevitably prevent successful fitting. Rather, it identifies a region in which delocalization becomes common, and the robustness of automated binding-site extraction is reduced.

Taken together, these results show that entropy should not be interpreted as a strict cutoff for determining fit feasibility. Instead, it serves as an early warning indicator, provided by GALP, that flags systems in which weak or non-specific host-guest interactions lead to diffuse adsorption behaviour. In such cases, the binding sites extracted by GALP may be poorly defined and therefore less reliable, even when apparent convergence metrics appear satisfactory. In this sense, entropy complements the Tanimoto score by distinguishing between genuinely converged, well-localized probability distributions from cases where high apparent convergence masks an underlying lack of spatial definition. The combined use of both metrics provides a more physically grounded framework for assessing the reliability of binding-site identification from GCMC-derived APDs.

2.10 Conclusion

This study validates the GALP algorithm as a reliable and flexible framework for extracting and fitting adsorption binding sites from GCMC-generated probability distributions in porous crystalline materials. Through systematic parameter optimization and validation, GALP was assessed across more than 100 MOFs spanning diverse topologies and pore sizes, and 9 chemically distinct guest molecules. This extensiveness demonstrates that GALP is not limited to narrowly defined systems but can be applied robustly across a wide range of adsorption environments.

By optimizing GALP fitting parameters for each guest molecule, the methodology was shown to reliably identify chemically meaningful adsorption sites, including systems involving weak or challenging interactions such as propane and propylene, which are known to exhibit diffuse probability distributions and subtle binding preferences. These results confirm that appropriate parameterization is essential for accurate binding site extraction, particularly for nonpolar or weakly interacting adsorbates.

In addition to algorithmic validation, this work systematically examined the influence of key simulation and analysis choices on binding site localization. Grid resolution was shown to significantly affect both convergence behaviour and localization accuracy, with a grid spacing of 0.15 Å identified as an effective compromise between spatial resolution and computational efficiency. Although equitable binning was found to minimally improve APD convergence, it incurs no extra computational cost. At the same time, one does not need to use equitable binding to use GALP. The use of convergence and localization metrics, including Tanimoto similarity and entropy-based measures, provided quantitative diagnostics for assessing simulation quality beyond simple visual inspection.

Collectively, these results establish GALP as a reliable and transferable tool for identifying binding sites in porous materials. By combining optimized guest-specific parameterization with robust analysis of simulation convergence and localization quality, GALP enables reliable extraction of adsorption sites from GCMC simulations. This capability supports large-scale screening studies and provides a practical link between computational predictions and experimental characterization of adsorption phenomena in crystalline porous frameworks

2.11 References

(18) Lin, J.-B.; Nguyen, T. T. T.; Vaidhyanathan, R.; Burner, J.; Taylor, J. M.; Durekova, H.; Akhtar, F.; Mah, R. K.; Ghaffari-Nik, O.; Marx, S.; Fylstra, N.; Iremonger, S. S.; Dawson, K. W.; Sarkar, P.; Hovington, P.; Rajendran, A.; Woo, T. K.; Shimizu, G. K. H. A Scalable Metal-Organic Framework as a Durable Physisorbent for Carbon Dioxide Capture. *Science (1979).* **2021**, *374* (6574), 1464–1469. https://doi.org/10.1126/science.abi7281.

(35) Wu, E.; Gu, X.-W.; Liu, D.; Zhang, X.; Wu, H.; Zhou, W.; Qian, G.; Li, B. Incorporation of Multiple Supramolecular Binding Sites into a Robust MOF for Benchmark One-Step Ethylene Purification. *Nat. Commun.* **2023**, *14* (1), 6146. https://doi.org/10.1038/s41467-023-41692-x.

(41) Carrington, E. J.; Vitórica-Yrezábal, I. J.; Brammer, L. Crystallographic Studies of Gas Sorption in Metal–Organic Frameworks. *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.* **2014**, *70* (3), 404–422. https://doi.org/10.1107/S2052520614009834.

(47) Jiang, C.; Wang, X.; Ouyang, Y.; Lu, K.; Jiang, W.; Xu, H.; Wei, X.; Wang, Z.; Dai, F.; Sun, D. Recent Advances in Metal–Organic Frameworks for Gas Adsorption/Separation. *Nanoscale Adv.* **2022**, *4* (9), 2077–2089. https://doi.org/10.1039/D2NA00061J.

(48) Nguyen, T. T. T.; Lin, J.-B.; Shimizu, G. K. H.; Rajendran, A. Separation of CO2 and N2 on a Hydrophobic Metal Organic Framework CALF-20. *Chemical Engineering Journal* **2022**, *442*, 136263. https://doi.org/10.1016/j.cej.2022.136263.

(49) Pham, T.; Space, B. Insights into the Gas Adsorption Mechanisms in Metal–Organic Frameworks from Classical Molecular Simulations. *Top. Curr. Chem.* **2020**, *378* (1), 14. https://doi.org/10.1007/s41061-019-0276-x.

(50) Main, R. M.; Vornholt, S. M.; Ettlinger, R.; Netzsch, P.; Stanzione, M. G.; Rice, C. M.; Elliott, C.; Russell, S. E.; Warren, M. R.; Ashbrook, S. E.; Morris, R. E. In Situ Single-Crystal X-Ray Diffraction Studies of Physisorption and Chemisorption of SO2 within a Metal–Organic Framework and Its Competitive Adsorption with Water. *J. Am. Chem. Soc.* **2024**, *146* (5), 3270–3278. https://doi.org/10.1021/jacs.3c11847.

(51) Li, S.; Chung, Y. G.; Snurr, R. Q. High-Throughput Screening of Metal–Organic Frameworks for CO2 Capture in the Presence of Water. *Langmuir* **2016**, *32* (40), 10368–10376. https://doi.org/10.1021/acs.langmuir.6b02803.

(52) Nguyen-Thuy, T.; Le-Hoang, P.; Hoang Vu, N.; Le, T. N.-M.; Le Hoang Doan, T.; Kuo, J.-L.; Nguyen, T. T.; Phan, T. B.; Nguyen-Manh, D. Hydrogen Adsorption Mechanism of MOF-74 Metal–Organic Frameworks: An Insight from First Principles Calculations. *RSC Adv.* **2020**, *10* (72), 43940–43949. https://doi.org/10.1039/D0RA08864A.

(53) McDonald, T. M.; Mason, J. A.; Kong, X.; Bloch, E. D.; Gygi, D.; Dani, A.; Crocellà, V.; Giordanino, F.; Odoh, S. O.; Drisdell, W. S.; Vlaisavljevich, B.; Dzubak, A. L.; Poloni, R.; Schnell, S. K.; Planas, N.; Lee, K.; Pascal, T.; Wan, L. F.; Prendergast, D.; Neaton, J. B.; Smit, B.; Kortright, J. B.; Gagliardi, L.; Bordiga, S.; Reimer, J. A.; Long, J. R. Cooperative Insertion of CO2 in Diamine-Appended Metal-Organic Frameworks. *Nature* **2015**, *519* (7543), 303–308. https://doi.org/10.1038/nature14327.

(54) Boyd, P. G.; Chidambaram, A.; García-Díez, E.; Ireland, C. P.; Daff, T. D.; Bounds, R.; Gładysiak, A.; Schouwink, P.; Moosavi, S. M.; Maroto-Valer, M. M.; Reimer, J. A.; Navarro, J. A. R.; Woo, T. K.; Garcia, S.; Stylianou, K. C.; Smit, B. Data-Driven Design of Metal–Organic Frameworks for Wet Flue Gas CO2 Capture. *Nature* **2019**, *576* (7786), 253–256. https://doi.org/10.1038/s41586-019-1798-7.

(55) Vaidhyanathan, R.; Iremonger, S. S.; Shimizu, G. K. H.; Boyd, P. G.; Alavi, S.; Woo, T. K. Direct Observation and Quantification of CO2 Binding Within an Amine-Functionalized Nanoporous Solid. *Science (1979).* **2010**, *330* (6004), 650–653. https://doi.org/10.1126/science.1194237.

(56) Vaidhyanathan, R.; Iremonger, S. S.; Shimizu, G. K. H.; Boyd, P. G.; Alavi, S.; Woo, T. K. Competition and Cooperativity in Carbon Dioxide Sorption by Amine-Functionalized Metal–Organic Frameworks. *Angewandte Chemie* **2012**, *124* (8), 1862–1865. https://doi.org/10.1002/ange.201105109.

(57) Ho, C.-H.; Paesani, F. Elucidating the Competitive Adsorption of H2O and CO2 in CALF-20: New Insights for Enhanced Carbon Capture Metal–Organic Frameworks. *ACS Appl. Mater. Interfaces* **2023**, *15* (41), 48287–48295. https://doi.org/10.1021/acsami.3c11092.

(58) Charnley, A. Rmsd, Version 1.6.4. GitHub January 13, 2025. https://github.com/charnley/rmsd (accessed 2025-07-12).

(59) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (Pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Comput. Mater. Sci.* **2013**, *68*, 314–319. https://doi.org/10.1016/j.commatsci.2012.10.028.

(60) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; Vijaykumar, A.; Bardelli, A. Pietro; Rothberg, A.; Hilboll, A.; Kloeckner, A.; Scopatz, A.; Lee, A.; Rokem, A.; Woods, C. N.; Fulton, C.; Masson, C.; Häggström, C.; Fitzgerald, C.; Nicholson, D. A.; Hagen, D. R.; Pasechnik, D. V.; Olivetti, E.; Martin, E.; Wieser, E.; Silva, F.; Lenders, F.; Wilhelm, F.; Young, G.; Price, G. A.; Ingold, G.-L.; Allen, G. E.; Lee, G. R.; Audren, H.; Probst, I.; Dietrich, J. P.; Silterra, J.; Webber, J. T.; Slavič, J.; Nothman, J.; Buchner, J.; Kulick, J.; Schönberger, J. L.; de Miranda Cardoso, J. V.; Reimer, J.; Harrington, J.; Rodríguez, J. L. C.; Nunez-Iglesias, J.; Kuczynski, J.; Tritz, K.; Thoma, M.; Newville, M.; Kümmerer, M.; Bolingbroke, M.; Tartre, M.; Pak, M.; Smith, N. J.; Nowaczyk, N.; Shebanov, N.; Pavlyk, O.; Brodtkorb, P. A.; Lee, P.; McGibbon, R. T.; Feldbauer, R.; Lewis, S.; Tygier, S.; Sievert, S.; Vigna, S.; Peterson, S.; More, S.; Pudlik, T.; Oshima, T.; Pingel, T. J.; Robitaille, T. P.; Spura, T.; Jones, T. R.; Cera, T.; Leslie, T.; Zito, T.; Krauss, T.; Upadhyay, U.; Halchenko, Y. O.; Vázquez-Baeza, Y. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17* (3), 261–272. https://doi.org/10.1038/s41592-019-0686-2.

(61) Boyd, P.; Marchand, O.; Burner, J. FastMC (v1.4.0). WooLab - Github 2025. https://github.com/uowoolab/FastMC-1.4.0.git (accessed 2025-07-12).

(62) Kresse, G.; Furthmüller, J. Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set. *Comput. Mater. Sci.* **1996**, *6* (1), 15–50. https://doi.org/10.1016/0927-0256(96)00008-0.

(63) Burner, J.; Luo, J.; White, A.; Mirmiran, A.; Kwon, O.; Boyd, P. G.; Maley, S.; Gibaldi, M.; Simrod, S.; Ogden, V.; Woo, T. K. ARC–MOF: A Diverse Database of Metal-Organic Frameworks with DFT-Derived Partial Atomic Charges and Descriptors for Machine Learning. *Chemistry of Materials* **2023**, *35* (3), 900–916. https://doi.org/10.1021/acs.chemmater.2c02485.

(64) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114* (25), 10024–10035. https://doi.org/10.1021/ja00051a040.

(65) García-Sánchez, A.; Ania, C. O.; Parra, J. B.; Dubbeldam, D.; Vlugt, T. J. H.; Krishna, R.; Calero, S. Transferable Force Field for Carbon Dioxide Adsorption in Zeolites. *The Journal of Physical Chemistry C* **2009**, *113* (20), 8814–8820. https://doi.org/10.1021/jp810871f.

(66) Potoff, J. J.; Siepmann, J. I. Vapor–Liquid Equilibria of Mixtures Containing Alkanes, Carbon Dioxide, and Nitrogen. *AIChE Journal* **2001**, *47* (7), 1676–1682. https://doi.org/10.1002/aic.690470719.

(67) Fischer, M.; Hoffmann, F.; Fröba, M. New Microporous Materials for Acetylene Storage and C2H2/CO2 Separation: Insights from Molecular Simulations. *ChemPhysChem* **2010**, *11* (10), 2220–2229. https://doi.org/10.1002/cphc.201000126.

(68) Zhou, Z.; Todd, B. D.; Travis, K. P.; Sadus, R. J. A Molecular Dynamics Study of Nitric Oxide in Water: Diffusion and Structure. *J. Chem. Phys.* **2005**, *123* (5). https://doi.org/10.1063/1.1992482.

(69) Boato, G.; Casanova, G. A Self-Consistent Set of Molecular Parameters for Neon, Argon, Krypton and Xenon. *Physica* **1961**, *27* (6), 571–589. https://doi.org/10.1016/0031-8914(61)90072-6.

(70) Martin, M. G.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 1. United-Atom Description of n-Alkanes. *J. Phys. Chem. B* **1998**. https://doi.org/https://doi.org/10.1021/jp972543+.

(71) Chen, B.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 3. Explicit-Hydrogen Description of Normal Alkanes. *J. Phys. Chem. B* **1999**, *103* (25), 5370–5379. https://doi.org/10.1021/jp990822m.

(72) Mahoney, M. W.; Jorgensen, W. L. A Five-Site Model for Liquid Water and the Reproduction of the Density Anomaly by Rigid, Nonpolarizable Potential Functions. *J. Chem. Phys.* **2000**, *112* (20), 8910–8922. https://doi.org/10.1063/1.481505.

(73) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. Development of an Improved Four-Site Water Model for Biomolecular Simulations: TIP4P-Ew. *J. Chem. Phys.* **2004**, *120* (20), 9665–9678. https://doi.org/10.1063/1.1683075.

(74) Wick, C. D.; Martin, M. G.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 4. United-Atom Description of Linear and Branched Alkenes and Alkylbenzenes. *J. Phys. Chem. B* **2000**, *104* (33), 8008–8016. https://doi.org/10.1021/jp001044x.

(75) White, A.; Gibaldi, M.; Burner, J.; Mayo, R. A.; Woo, T. Alarming Structural Error Rates in MOF Databases Used in Data Driven Workflows Identified via a Novel Metal Oxidation State-Based Method. October 10, 2024. https://doi.org/10.26434/chemrxiv-2024-ftsv3.

(76) Burner, J. Farthest Point Sampling. WooLab - Github 2022. https://github.com/uowoolab/MOF-Diversity-Analysis (accessed 2025-07-12).

3. Accuracy of Atomistic Simulations in Predicting MOF Binding Sites

How well do conventional atomistic simulations predict adsorption binding sites in metal-organic frameworks compared to experiment

3.1 Abstract

Classical force field grand canonical Monte Carlo (GCMC) simulations are widely used to model gas adsorption properties in metal-organic frameworks (MOFs). The accuracy of the approximations made in such simulations has been extensively studied by comparison to global adsorption properties such as uptakes and selectivities, where well-known limitations in force field energetics can lead to significant discrepancies compared to experiment. However, no such investigation has been performed for a local adsorption property such as binding sites, despite its central importance in understanding guest-host interactions and challenges surrounding direct experimental observation. In this work, the literature was searched for MOFs with directly observed binding sites, resolved from diffraction studies. The dataset is composed of 8 adsorbates, and 22 MOFs with varying degrees of accessible open metal sites, flexibility, and dimensionality across various temperature and pressure conditions. Surprisingly, across the dataset classical GCMC recovers the correct number of binding sites and achieves agreement with experiment in binding site positions, with an average RMSD of ~0.6 Å (including adsorbates that bind to open metal sites). Deviations occur most predominantly in cases where experimental uncertainty, framework flexibility, or chemisorption with reactive adsorbates have been reported. In the GCMC simulations, charge assignment was found to be the dominant contributing factor to accuracy, while Lennard-Jones parameter sets have comparatively minor influence. Importantly, it is shown that classical GCMC simulations can be used to accurately determine binding sites, even in cases

where adsorption isotherms are poorly reproduced. This highlights the difference between predicting relative versus absolute adsorption energetics, which dominate binding site position and surface coverage (uptake), respectively. This work provides the first systematic validation of classical GCMC simulations for predicting binding sites of MOFs, establishes a benchmark for future method development, and highlights the surprising robustness and clear limitations of routine computational workflows for resolving guest-host interactions.

3.2 Statement of Work

The work presented in this chapter is derived from a co-first-authored communication submitted to *Materials Advances*. This chapter reproduces the content of the submitted manuscript, with modifications made solely to ensure consistency with the formatting and structure of the thesis. In addition, material that was referenced in the Supporting Information of the communication has been incorporated directly into the main text of this chapter for completeness. Within this context, I was responsible for the data curation, methodological development, formal analysis, investigation, and validation of all results reported here. I designed and executed the complete computational workflow, including the generation, processing, and analysis of the simulation data. I prepared the original draft of the communication, after which the manuscript was revised and finalized in collaboration with the coauthors to prepare it for publication.

3.3 Introduction

Gas separation and storage is one of the most widely studied applications of MOF materials, where it is reaching commercialization. For example, CALF-20[18] is being scaled for industrial CO_2 capture from cement-manufacturing flues[77], and MOFs are being deployed for safe

storage and transport of toxic gases[78,79]. Central to these applications are the adsorption binding sites (ABSs), which govern affinity and selectivity and therefore play a fundamental role in applications from gas storage and separation[54] to catalysis and sensing[80]. Although >100,000 MOFs have been experimentally characterized and deposited in the CSD[38,81], direct observation of adsorbates in these materials using crystallography is relatively rare, often requiring synchrotron radiation sources[39,40]. While other techniques are often used to probe adsorbate interactions in MOFs (e.g., NMR), these usually rely on other experimental or computational methods to elucidate the detailed positions of adsorbates[42,43].

Atomistic GCMC simulations have become a common tool for modelling gas adsorption in MOFs and other porous materials[82,83]. GCMC is particularly popular for simulating gas adsorption isotherms of MOFs. The simulation method can be characterized as a brute-force technique where interaction energies are computed for millions of configurations to generate a single isotherm point using the machinery of statistical mechanics. Therefore, practical use of GCMC simulations generally relies on simplified approximations, which limit their accuracy. For example, generic force field parameters (e.g., UFF[64]) are used to compute pairwise Lennard-Jones potentials to model steric/dispersion interactions, while fixed partial atomic charges are used to model electrostatic interactions (obtained either from empirical models[84], ML models[85], or fit to DFT potentials[86,87]). In addition to these approximations, the MOF is often modelled as a rigid, defect-free crystal.

The accuracy of these approximations and sensitivity to adsorption properties such as uptakes and diffusion of MOFs have been extensively studied[83,88,89]. For example, Cleeton et al.[83] and McCready et al.[88] conducted large systematic evaluations of classical GCMC simulations for computing CO_2 isotherms in MOFs, and both reported significant variability in predicted uptake,

especially when open metal sites (OMSs) are present. Goeminne et al.[90] showed that even small interaction energy errors (~4 kJ/mol) and subtle framework distortions can strongly impact predicted isotherms. Together, these results indicate that common GCMC approximations may be insufficient for reliable isotherm prediction versus experiment. However, while such limitations are known for global observables, their effect on local properties like experimentally resolved ABSs remains much less understood.

In this work, we address this gap by curating a crystallography dataset of directly observed adsorbate positions spanning 23 MOFs with varying degrees of accessible OMSs, flexibility, and dimensionality and eight adsorbates (CO_2, C_2H_2, CH_4, NO, Ar, Xe, Kr, and H_2O). The set was restricted to structures where the adsorbate and MOF atomic positions were fully defined from crystallography. Structures with highly disordered sites or those which relied on computational refinement (e.g., DFT calculations) were excluded since optimized geometries from DFT calculations typically correspond to minima on the single-guest potential energy surface that neglects guest-guest interactions, thermal, and entropic effects. An exhaustive search of the literature resulted in a total of 35 MOF/adsorbate/condition combinations (MACs), for which "conventional" GCMC simulations were performed to obtain atom-specific adsorption probability distributions (APDs), whose maxima correspond to free-energy minima ABSs. We define a simulation as being "conventional" if the following approximations are used: a) the framework is assumed to be rigid and defect-free; b) generic classical force fields are used for the MOF framework to model dispersion interactions; and c) fixed partial charges are used to model electrostatics (i.e., polarization is neglected).The ABSs were fit to the APDs using our in-house code GALP (reference Chapter 2). Each of the 35 MACs were assigned an ID given in column 4 of Table 3.1, column 4. There are many examples of ABSs in 1D M(II) benzoate pyrazines in the

literature, of which we have selected four (**14-16** = Rh(II), **17** = Cu(II), **18-19** = Rh(II) with 2-ethylpyrazine, **1-2** = Rh(II) with dimethylpyrazine) to prevent redundancy (the variants possess similar ABSs). Two examples of M(II) formates (**5** = Mg(II) and **6** = Mn(II)) were also included. CALF-15 (**7**) and CALF-20(**3-4**) are both Zn(II) MOFs composed of oxalate and triazolate linkers, differing only by amine functionalization. Finally, the MOF-74 analogues studied in this work (**11, 26-31**) only differ by metal centre identity and/or degree of OMS capping. This results in 12 classes of structures that exhibit notably unique geometries and chemical compositions.

3.4 Methodology

3.4.1 Experimental Dataset Preparation

As a result of an extensive literature search, MOFs which had experimentally resolved binding sites from XRD or neutron diffraction were collected, and isotherms were also collected when available. The dataset was limited to MOFs where the adsorbate atomic positions were fully defined, such that they were available in a crystallographic information file (CIF) format, or similar. Structures where experimental binding sites exhibited a large degree of disorder, or were refined using computational methods (e.g., DFT) were excluded. This decision reflects the fact that DFT results can have systematic biases depending on the choice of functional, basis set, and dispersion corrections that are difficult to deconvolute. Additionally, DFT structures are often obtained from 0 K geometry optimization, which ignores thermal and entropic effects. These calculations are also usually done with at most one adsorbate in the MOF at a time, which neglects effects of guest-guest interactions.

All obtained crystal structures underwent detailed manual inspection and processing to ensure fidelity of the structures to the original reporting publication for GCMC simulations. The

processing varied on a case-by-case basis, but included correcting structural disorder, adding missing protons, adjusting incorrect protonation states, etc. As a result of this procedure, roughly fifteen MOFs were excluded since the framework or guest disorder was too severe to be corrected. Various combinations of 27 MOFs with seven different adsorbates (CO_2, C_2H_2, CH_4, NO, Ar, Xe, Kr) across various conditions (temperature, pressure) were collected for further use in GCMC simulations, given in Table 1 in the main text. The structures were taken either from the CSD or supplementary information provided in the original reporting publications.

3.4.2 Definition of RMSD Metrics

In this work, several RMSD (Root Mean Square Deviation) metrics are reported to quantify differences between binding site configurations. The atom-centred RMSD, reported in Table 1, is computed from the displacements of all guest atoms across the ensemble of binding sites, excluding the MOF framework, and reflects the overall atomic-level mismatch between two configurations. This quantity corresponds to the average of the RMSDs computed independently for each binding-site pair, such that each binding site contributes equally. In addition, a center-of-mass (COM) RMS Euclidean displacement is included to capture purely translational differences between binding sites, independent of internal molecular distortions or rotations, and is also reported in Table 1. Since all comparisons involve the same guest molecule and the same number of binding sites, the guest-centred RMSD (RMS) is identical to the atom-centred RMSD and is therefore not reported separately.

3.5 Computational Details

3.5.1 DFT Calculations and GCMC Simulations

All MOFs were converted to P1 symmetry and existing guest atoms (from the experimental crystal structure) were removed prior to simulation. All DFT calculations were performed with VASP (version 6.4.2) with the PBE exchange-correlation functional[91] and D3 dispersion corrections[92], and standard VASP pseudopotentials[93] (potpaw.6.4.2), based on the projector augmented wave (PAW) method[94]. For REPEAT[86] calculations and ionic relaxations, k-point sampling was restricted to the gamma point only, an SCF convergence criteria of $1x10^5$ eV was used with the VASP default plane-wave cutoff, and a force cutoff of 0.02 eV/Å for ionic relaxations. The positions of the protons were optimized at the DFT level where both the simulation cell and all other atomic positions were fixed. This choice was made since some structures were missing proton positions, so this protocol allowed for structures to be treated more equally. Moreover, proton positions are often poorly resolved due to weak X-ray scattering. For MOF-74(Ni)@NO and CALF-20@H_2O where adsorbates were considered part of the structure (chemisorbed), the adsorbate atom positions were also optimized in addition to all protons. Structures with disorder were fixed using Mercury. For DDEC-6[87] calculations, a plane-wave cutoff energy of 520 eV, and a Γ-centered k-point mesh was used such that the product of the number of k-points and the lattice vector length in each direction was at least 16 Å, as recommended by the manual in the *chargemol* software package[95]. For MOFs with open metal sites, spin-polarized calculations were performed. All GCMC simulations were performed using our in-house code available on GitHub[61]. The adsorbates and framework were all treated as rigid, such that only non-bonded interactions were required to be computed. Dispersion/steric

interactions were modeled using atom pairwise Lennard-Jones potentials and electrostatic interactions were modelled using fixed atomic partial charges. Lennard-Jones parameters were obtained from the UFF force field[64], and the resulting binding sites were compared to those obtained from using the DREIDING force field[96] for the framework atoms. The Lennard-Jones parameters and partial atomic charges of the guest molecules used in this work correspond to the following force fields: CO_2[65] C_2H_2[67], H_2O[73], NO[68], CH_4[70], Ar[97], Xe[97], and Kr[97]. Atomic partial charges of framework atoms were fit to the electrostatic potential (ESP) obtained from a single-point DFT calculation using the REPEAT methods, as previously described in the DFT section. For comparison, the resulting binding sites were compared to those obtained when using framework atomic partial charges obtained from MEPO-QEq[84], MEPO-ML[85], and the DDEC6[87] methods.

Each GCMC simulation was run until the probability distribution reached a minimum Tanimoto similarity of 0.90, ensuring that adsorption sites were fully sampled. The total number of Monte Carlo steps varied depending on the supercell size and guest-framework interaction strength, ranging approximately from 1.5×10^6 to 4.6×10^8, with a median of about 2.0×10^7 and a mean of roughly 6.6×10^7 steps. The mean value was inflated by a few very long runs; excluding these, the average number of production steps was approximately 2.4×10^7. A cutoff distance of 12.5 Å was applied to all Lennard-Jones non-bonded interactions. Simulation cells were expanded to fully accommodate this cutoff, and no MOF was simulated using a 1x1x1 unit cell. This ensured that replicate subcells were available for Tanimoto-based convergence analysis. For cases where convergence could not be reached due to local maxima (most notably for the MOF-74 series) parallel simulation were performed. Specifically, 25 independent GCMC simulations of 8.0×10^7 production steps each were executed in parallel, and the resulting probability plots were combined

to yield an effective sampling equivalent to approximately $2x10^9$ production steps. This procedure reduced wall time to about two days per structure (given the available cores) while maintaining statistical reliability.

3.5.2 Binding site Identification

To computationally determine binding sites of MOFs, at each GCMC production step, the 3D atomic coordinates of the adsorbates were stored in a histogram to generate 3D APDs, where regions of high probability correspond to free energy minima. The APDs were stored on a uniform grid of maximum spacing 0.15 Å, and were normalized to unity. The convergence of the APDs was evaluated by comparing the probability distribution in replicate unit cells within the simulation cell. If the Tanimoto similarity coefficient between replicate unit cells was ≥ 0.90, the APD was determined to be converged, and the production phase of the simulation was stopped. For most simulations, this required between 10 and 30 million GCMC steps, though certain systems well exceeded 100 million. Once obtained, the APD of the simulation cell was "folded" into a single unit by averaging the APD over all replicate unit cells and renormalized such that the sum of each APD was unity. Binding site positions within the MOF were determined using our GALP method, as described in Chapter 2. In this method, a 3D Gaussian filter is applied to each APD to smooth the distributions. Following this, all local maxima are identified and any maxima with a value below 0.05% of the global maximum were discarded. Only the largest maximum within an exclusion radius of 0.40 Å were retained. Next, maxima are vetted according to whether they align with the geometry of the rigid adsorbate molecule used in the GCMC simulation by applying a RMSD-minimizing alignment procedure. This gives the final positions of the binding sites, for which guest-host interaction energies are determined using the respective force fields/charges.

3.6 Results and Discussion

Figure 3.1 compares experimentally determined ABSs with those obtained from GCMC simulations for representative MACs (full set shown in Figure 3.5-3.39). These figures show that the general positions of the experimental ABSs are well reproduced by the GCMC simulations in almost all cases. The centre-of-mass RMSD between the computed and experimental ABSs (Table 1) is only 0.51 Å averaged over all MACs. This is excellent overall agreement considering this corresponds to only a few grid points on the 0.15 Å resolution APD grids. The standard deviation is ~75% of the mean, highlighting the presence of several outliers, which are discussed in more detail below. In addition to reproducing general ABS positions, the correct number of sites match experiment in all cases except MAC **32** ($[Cu_2(pyrdc)_2(bpp)_2]_n$), where GCMC predicts one additional CO_2 per asymmetric unit that is 25-35% weaker than the others. Experimental and simulated saturation uptakes are internally consistent with full occupation of their respective ABSs (i.e., GCMC predicts higher saturation uptake because it includes an extra site, and vice versa for experiment). The discrepancy therefore likely reflects incomplete activation and/or uncertainty in the refined structure.

For polyatomic adsorbates, the RMSD per atom was computed to incorporate adsorbate orientation in comparing simulated and experimental ABSs. Across all MACs, the mean RMSD/atom is 0.66 Å, well below the ~2 Å threshold commonly considered a good fit in molecular docking studies[98]. This agreement is achieved despite using rigid adsorbates whose fixed internal geometries can differ from crystallography (e.g., NO in MAC 27 is unusually long experimentally at 1.43 Å vs 1.15 Å). Such mismatches contribute an average residual RMSD of 0.06 Å (~10% of the mean).

Classical, rigid GCMC simulations should perform worst when the approximations are most severe, most notably when MOFs present significant framework flexibility and open metal sites (OMSs), since in such cases generalized force fields are insufficient. To investigate this hypothesis, we group the 35 MACs into four categories: (i) no OMSs/minimal flexibility (MACs **1-13**); (ii) no OMSs/significant flexibility (MACs **14-22**); (iii) OMSs/no flexibility (MACs **23-31**); and (iv) OMSs/significant flexibility (MACs **32-35**). As expected, category (i) MACs show the best agreement (mean RMSD/atom = 0.47 Å, vs. 0.66 Å overall), whereas category (iv) is worse (0.72 Å). Category (ii) (**13–22**) is intermediate (0.62 Å), although most "flexibility" here reflects abrupt phase transitions (except NH_2-MIL-53(Al)), limiting broader conclusions. Interestingly, MACs from category (iii) (**23-31**) have the largest mean RMSD/atom of 0.86 Å. While unexpected, this is due to the fact that category (iii) MACs contain two cases in which the adsorbate (NO) binds chemisorptively and the GCMC simulations invert the binding mode (*vide infra*) giving rise to large RMSDs. Furthermore, the OMSs in category (iv) are not nearly as accessible to the adsorbates as MOF-74, which makes up the bulk of MACs in category (iii).

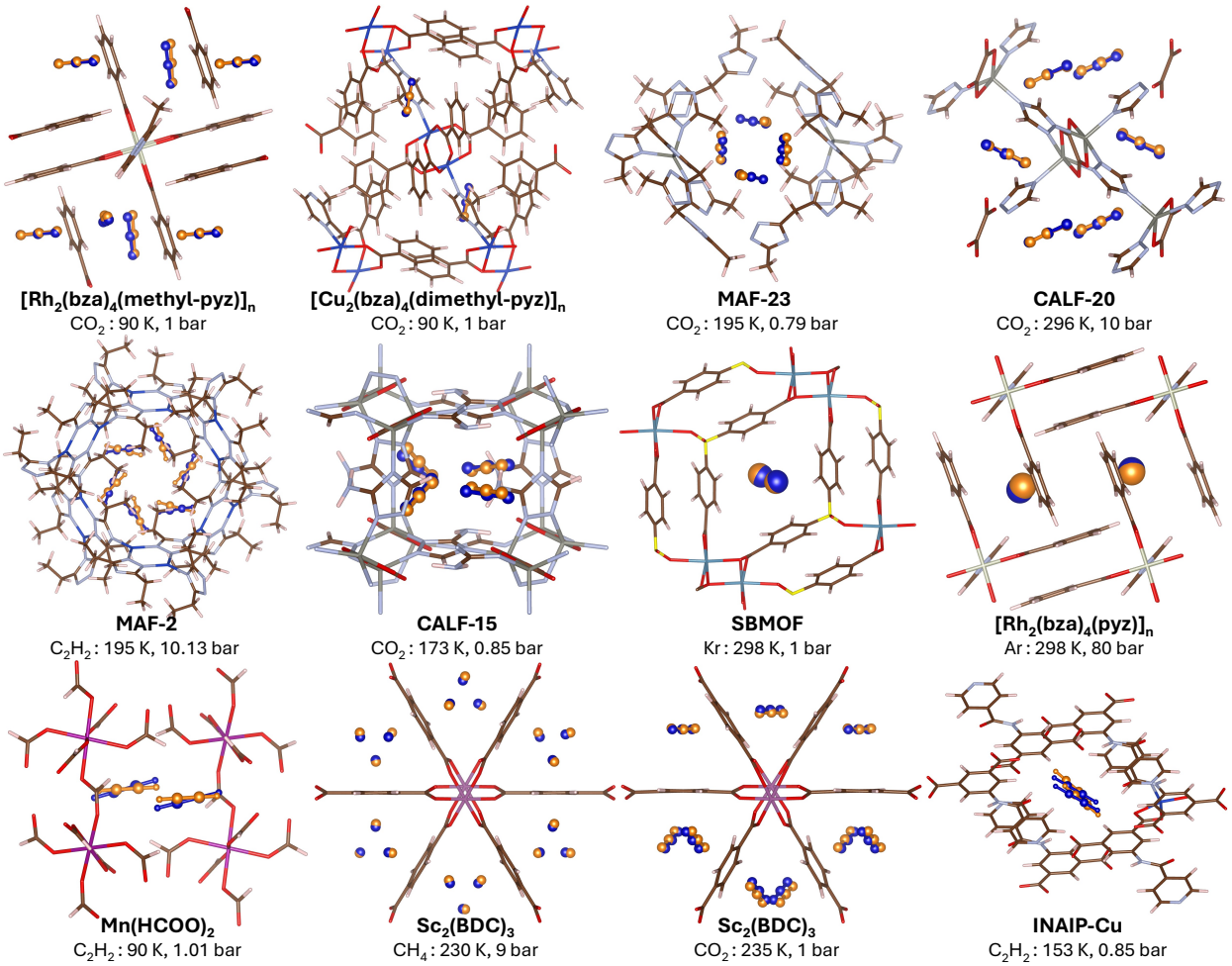

Figure 3.1. Comparison of ABS positions obtained from GCMC simulations (blue balls) and from crystallography experiments (orange balls) for a variety of MOFs and adsorbates selected from Table 1. Each MOF is labelled with the guest, temperature and pressure at which the crystallography and GCMC simulations were performed. Importantly, ABSs correspond to free energy minima instead of potential energy minima.

A perhaps surprising observation is that qualitatively poor agreement in the adsorption isotherms, a global adsorption property, does not necessarily preclude accurate ABS identification, a local property. Figures 3.5-3.39 show the adsorption isotherm comparisons if available—of available isotherms, about half were performed at the same temperature as the crystallography experiments. The experimental and simulated isotherms typically share the same shape, but differ in steepness (e.g., MAF-2, CALF-15, SBMOF-1, the M(II) formates, and the MOF-74 series). For example, MOF-74(Co)@CO_2 (**26**) shows a nearly linear simulated isotherm versus a much steeper

experimental one (Figure 3.2a), yet the ABSs agree well (RMSD = 0.48 Å). This reflects that ABS positions are governed by relative interaction energies (positions of free-energy minima), whereas uptake and isotherm steepness depend on *absolute* adsorption energetics (depth of minima). Considering the exponential dependence of occupancies on energies, even modest energetic errors (~4 kJ/mol[90]) can strongly affect uptake/coverage while leaving site positions largely unchanged. Moreover, MOF-74 ABS positions remain accurate across metals even when isotherms deviate. Corroborating this, the degree of isotherm agreement correlates directly with the agreement in isosteric heats of adsorption (Table 3.1).

Table 3.1. Comparison of isosteric heats of adsorption (Q_{st}) of MOF-74 obtained from simulation and experiment at a coverage of ~0.1 CO_2 per M^{2+}.

MOF	Simulated HOA (kJ/mol)	Experimental HOA (kJ/mol)
MOF-74(Zn)	24.1	26.8
MOF-74(Fe)	25.6	33.2
MOF-74(Co)	24.1	33.6
MOF-74(Mg)	30.1	43.5

*Heat of Adsorption (HOA)

Table 3.2. List of MOF structures where detailed guest positions have been determined experimentally, the conditions they were acquired, and a comparison of the experimental and GCMC determined binding sites.

MOF name	OMS[a]	flexibility[b]	no.	guest	T (K)	P (bar)	cryst. (R-factor)	$RMSD^{c}_{COM}$	RMSD/atom[d]	ref.
$[Rh_2(bza)_4(dimethyl\text{-}pyz)]_n$	no	slight increase in volume at high loading	1	CO_2	90	1	SC-XRD (0.0468)	0.153	0.263	99
			2	CO_2	90	17	SC-XRD (0.1037)	0.199	0.205	99
CALF-20	no	slight phase change with water adsorption	3	CO_2	296	10	SC-XRD (0.0298)	0.156	0.221	100
			4	H_2O	296	–	PXRD (0.0418)	1.026	1.026	101
$Mg(HCOO)_2$	no	–	5	C_2H_2	90	–	SC-XRD (0.0404)	0.770	0.854	102
$Mn(HCOO)_2$	no	–	6	C_2H_2	90	–	SC-XRD (0.0345)	0.788	0.867	102
CALF-15	no	–	7	CO_2	173	0.85	SC-XRD (0.0295)	0.262	0.323	55
$Sc_2(BDC)_3$	no	slight rotational freedom of BDC linkers	8	CH_4	230	9	SC-XRD (0.0401)	0.231	0.231	103
			9	CO_2	235	1	SC-XRD (0.0713)	0.273	0.283	103
MUF-16(Mn)	no	–	10	CO_2	293	1.1	SC-XRD (0.0510)	1.283	1.446	104
MOF-74(Ni) (capped[f])	no	–	11	NO	300	0.08	SC-XRD (0.0490)	–	–	105
SBMOF-1	no	–	12	Xe	100	–	SC-XRD (0.0835)	0.256	0.256	106
			13	Kr	100	–	SC-XRD (0.0518)	0.180	0.180	106
$[Rh_2(bza)_4(pyz)]_n$	no	guest-induced phase change	14	Ar	298	80	SC-XRD (0.1149)	0.243	0.243	107
			15	CO_2	298	35	SC-XRD (0.1047)	0.169	0.389	107
			16	CO_2	93	1.01	SC-XRD (0.1031)	0.208	0.231	108
$[Cu_2(bza)_4(pyz)]_n$	no	guest-induced phase change	17	CO_2	193	1.01	SC-XRD (0.1470)	0.196	0.207	109
$[Rh_2(bza)_4(2\text{-}epyz)]_n$	no	guest-induced phase change	18	CO_2	298	64	SC-XRD (0.0772)	0.145	0.397	110
			19	CO_2	90	36	SC-XRD (0.1681)	0.174	0.192	110
NH2-MIL-53(Al)	no	various phases of pore opening	20	CO_2	253	3.0	PXRD (–)	0.950	1.331	111
			21	CO_2	253	9.5	PXRD (–)	1.395	1.499	111
			22	CO_2	253	18	PXRD (–)	0.607	1.062	111
CPL-1	yes	–	23	C_2H_2	170	0.1	PXRD (0.320)	0.123	0.159	112
INAIP-Cu	yes	–	24	C_2H_2	110	0.18	SC-XRD (0.0409)	0.634	1.026	113
			25	C_2H_2	153	0.85	SC-XRD (0.0357)	0.660	0.708	113

Table 3.2. Continued

MOF name	OMS[a]	flexibility[b]	no.	guest	T (K)	P (bar)	cryst. (R-factor)	$RMSD^{c}_{COM}$	RMSD/atom[d]	ref.
MOF-74(Co)	yes	–	26	CO_2	10	–	NPD (0.0290)	0.393	0.481	45[e]
			27	NO	298	1.01	PXRD (0.0582)	1.148	1.679	114
MOF-74(Fe)	yes	–	28	CO_2	298	1	NPD (0.0176)	0.550	0.738	45[e]
MOF-74(Mg)	yes	–	29	CO_2	10	–	NPD (0.0223)	0.659	0.776	45[e]
MOF-74(Zn)	yes	–	30	CO_2	10	–	NPD (0.0309)	0.456	0.642	45[e]
MOF-74(Ni)	yes	–	31	NO	300	0.4	SC-XRD (0.0551)	1.323	1.495	105
$[Cu_2(pyrdc)_2(bpp)_2]_n$	yes	pillared bilayer; guest-induced phase change	32	CO_2	193	–	PXRD (0.0585)	0.460	0.483	115
MAF-2(Zn)	yes	possible guest-induced flexibility; rotational freedom of ethyl groups	33	C_2H_2	195	10.1	SC-XRD (0.0393)	0.177	0.545	116
			34	CO_2	195	20.3	SC-XRD (0.0458)	0.212	0.316	116
MAF-23(Zn)	yes	guest-induced flexibility at binding site	35	CO_2	195	0.79	SC-XRD (0.0391)	0.998	1.547	117

[a]Open metal site (OMS) identified in structure. [b]Notes on any mention of flexibility in the cited publication(s). [c]COM RMSD is the root mean square deviation between the experimental and simulated center of mass of all binding sites. [d]Root mean square deviation (RMSD) per atom between GCMC binding site coordinates compared to experiment. [e]In the MOF-74 study from Queen et al.[45], both NPD and SC-XRD studies were performed, and were in agreement at similar surface coverages. [f]In the study from Main, et al.[105], a secondary binding site was observed in cases where some OMSs were capped by water, but due to having to optimize chemisorbed NO and water sites, structural differences were too large from the reported structure (disorder in capped sites, and missing protons) to allow for a direct comparison/RMSD determination.

Differences in isotherm shape (e.g., CPL-1, the M(II) benzoate pyrazines, and NH_2-MIL-53(Al)) often arise because these MOFs are flexible and undergo guest-induced phase transitions, so rigid framework simulations are not expected to reproduce the full pressure range; nevertheless, the simulated ABSs are typically still in good agreement with experiment when the correct phase is modelled. A clear example is $[Rh_2(bza)_4(2\text{-epyz})]_n$@$CO_2$ (**18–19**), which shows an abrupt α to β transition that appears near 0.28 bar in the experimental isotherm (Figure 3.2c(i)); using the fixed structure of each phase at the appropriate pressure yields good agreement for both isotherms and ABSs. By contrast, NH_2MIL-53(Al) shows qualitatively different simulated vs experimental CO_2 sites across three phases (Figure 3.22), but here the experimental sites come from *in-situ* PXRD/Rietveld refinement and the original authors report peak evidence for phase coexistence at intermediate pressures. This in conjunction with the fact that such procedures do not necessarily result in a unique solution raises uncertainty in the refined site positions. Therefore, the disagreement is likely partially a result of crystallographic limitations rather than simulation failure alone. The benzoate pyrazines in this work still exhibit low RMSDs despite adsorption-induced transitions, likely because their phase changes occur over a narrow pressure window with limited local flexibility (consistent with the isotherms). We also reiterate that we modelled the phase corresponding to the *in-situ* XRD conditions used to determine the experimental sites. MAF-23 shows analogous flexibility, though at a local rather than global scale. For MAF-23@CO_2 (**35**) (Figure 3.35), the secondary site is notably worse, consistent with the original report that CO_2 binding between chelating triazoles induces a small conformational change, so while the low-pressure regime is reproduced, rigid modelling likely underestimates the saturation uptake.

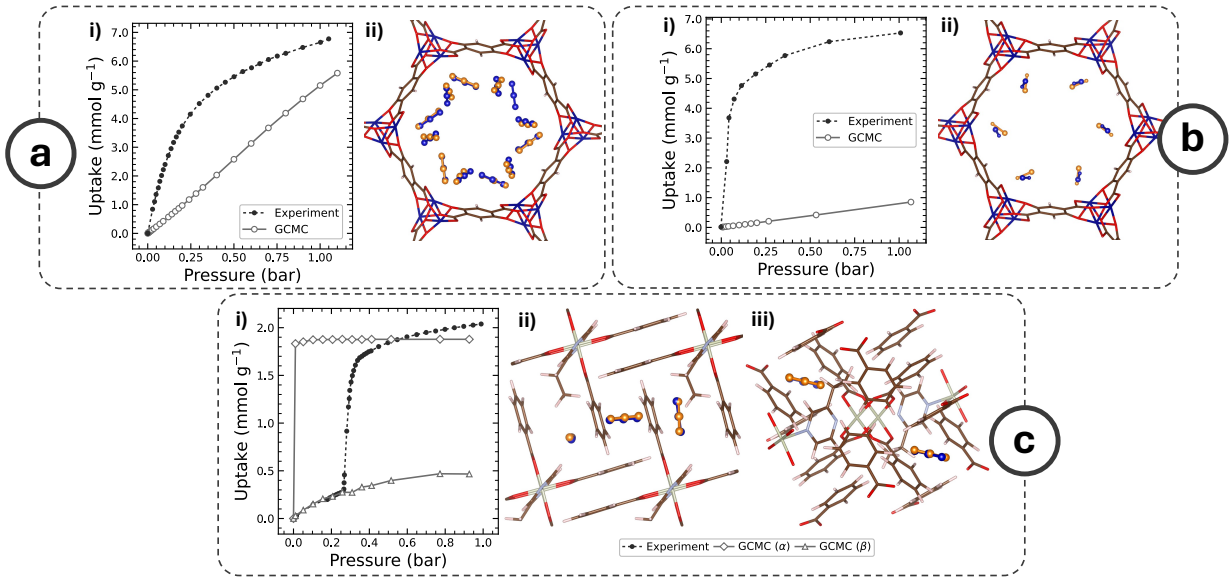

Figure 3.2. Comparisons of binding sites and adsorption isotherms between GCMC and experiment for a) MOF-74(Co)@CO_2 (MAC 26), b) MOF74(Co)@NO (MAC 27), and c) $[Rh_2(bza)_4(2\text{-epyz})]_n$@$CO_2$ (MACs 1-2).The simulated (blue) and experimental (orange) binding sites correspond to conditions specified in Table 1. The atoms of NO with larger radii in b) correspond to oxygen. The isotherms correspond to temperatures of a, b) 304 K, and c) 195 K. Structures in c) correspond to the (ii) α and (iii) β phases.

Another exception with qualitative disagreement in the binding sites arises for MOF-74@NO, where unlike the previous examples there is no notable flexibility, but there is a considerable degree of chemisorption. In MOF-74(Co)@NO (**27**) and MOF74(Ni)@NO (**31**), the simulated primary ABS has NO bound to the OMS via O instead of N as observed experimentally. This is because the interaction is expected to be chemisorptive to a large degree, and the generic classical force field is unable to treat such an interaction properly. This is also reflected in the simulated isotherm, which severely underestimates uptake as shown in the MOF-74(Co)@NO isotherm (Fig 2b(i)). However, the centre of mass of the ABS is nevertheless correctly predicted. We note the NO bond distance from crystallography is unusually long (1.43 Å), whereas the value used in the GCMC simulation is 1.15 Å(much closer to the 1.18 Å length obtained from a full geometry optimization of MOF-74(Co)@NO at the DFT level). To separate chemisorption effects from NO-specific complexity, we also examined a secondary physisorptive NO ABS in MOF-74(Ni)@NO where the OMSs are capped with H_2O (Figure 3.14, MAC **11**). There is still some

disagreement between experimental and simulated ABSs (with experimental disorder noted in the nitrogen position of NO)[105], but the qualitative binding motif is recovered: a dipole "chain" from $H_{H_2O}(\delta^+) \rightarrow O_{NO}(\delta^-) \rightarrow N_{NO}(\delta^+) \rightarrow MOF\ carboxylate(\delta^-)$ (Figure 3.3).

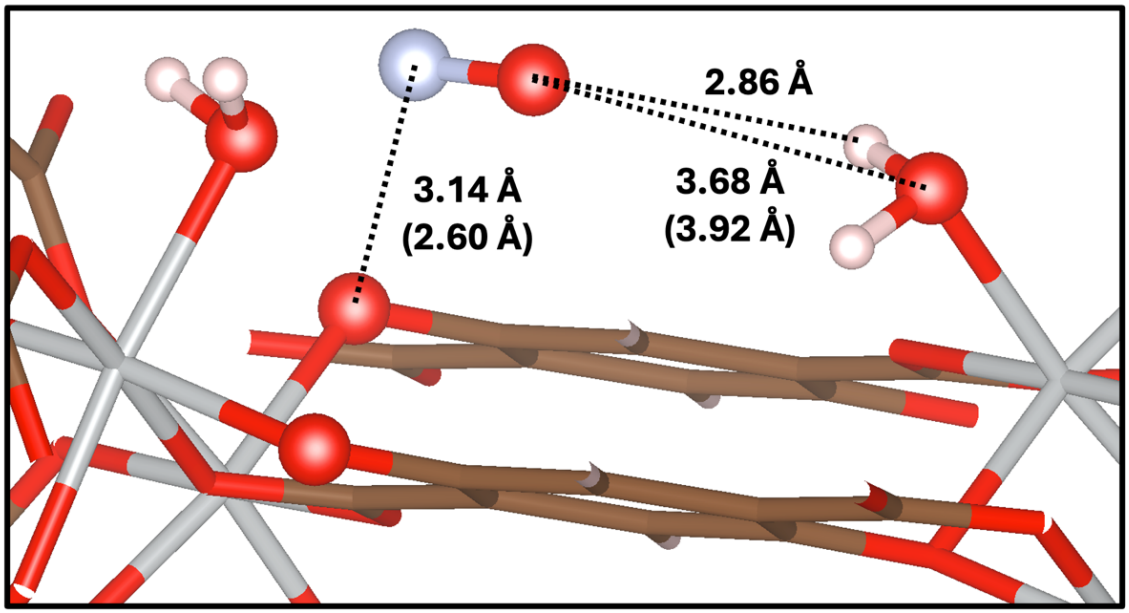

Figure 3.3. Physisorptive NO binding site in MOF-74(Ni) determined by GCMC, where 100% of OMSs are capped by water. Values in parentheses are distances obtained from one of the experimental crystal structures reported in the publication (CCDC refcode UJOCEF)[18]. The simulation conditions were 196 K, 0.40 bar NO. Atomic positions of capping sites and protons were optimized at the PBE level prior to simulation.

Overall, the large errors for MOF-74@NO likely reflect both OMS limitations and the challenges of modelling a reactive, open-shell adsorbate with a classical force field. Despite generally good agreement, an important question is whether other force fields predict the same ABSs. In this work, sensitivity of the ABSs to the force fields used to model adsorbate interactions were not investigated, since the ABSs were found to be largely insensitive to the choice of force field for the framework atoms. However, we note that other available or more specialized adsorbate force fields may exhibit higher accuracy for modelling certain ABSs. Most MOF screening studies model sterics/dispersion with UFF Lennard-Jones potentials and electrostatics with ESP fitted charges (e.g., REPEAT[86], DDEC6[87]) or charge equilibration (MEPO-QEq[84]). Excluding MOF-

74(Zn) and MOF-74(Fe), where metal parameters differ drastically (DREIDING RMSD/atom = 3.4 and 2.6 Å), UFF and DREIDING force fields give the same mean RMSD (0.56 Å). Empirical charges such as QEq often give qualitatively different results with the wrong number of sites, but MEPO-ML[85] (trained to reproduce ESP-fitted charges) provides a low cost alternative to DFT-derived charges with comparable accuracy (RMSD/atom = 0.63 Å). While RMSD analysis assesses the structural fidelity of the predicted ABS locations, a complementary perspective is obtained by examining the consistency of the associated binding energies across force field and charge model choices. To this end, a pairwise RMSE matrix (Figure 3.4) compares the binding energies predicted for each mapped binding site across 34 MOF guest systems, excluding MOF-74(Zn) and MOF-74(Fe) as described previously.

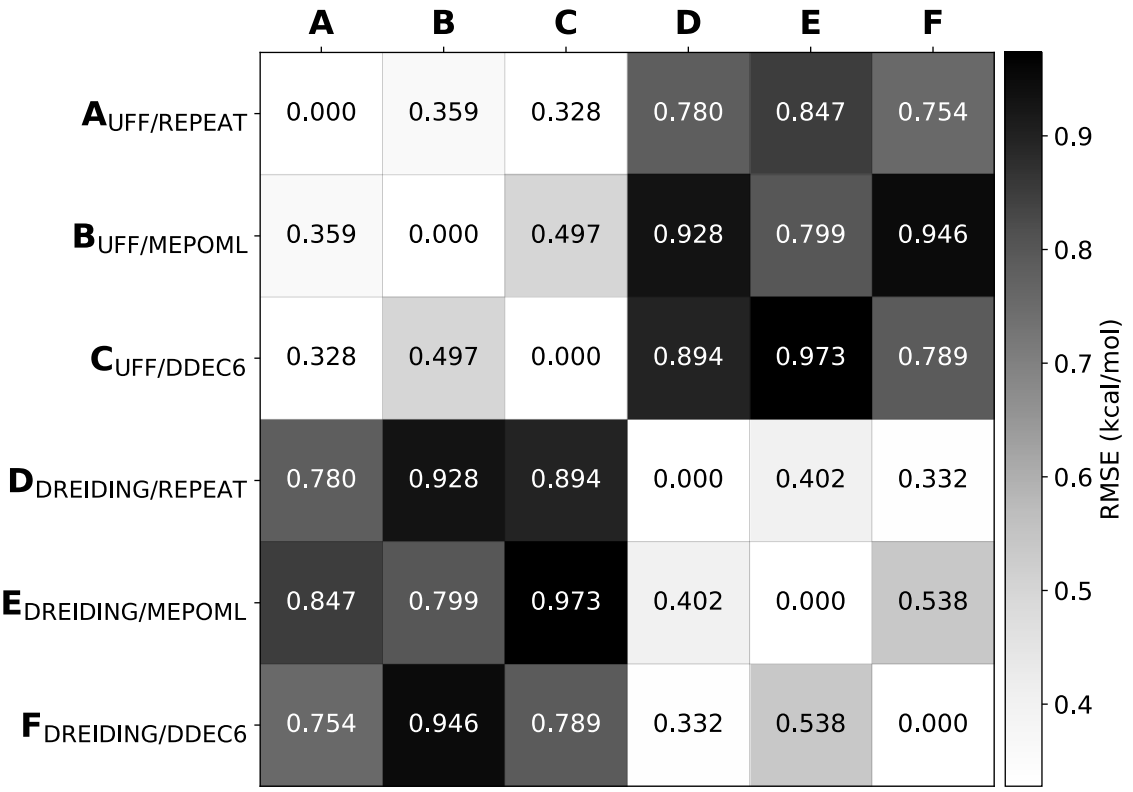

Figure 3.4. Pairwise RMSE (kcal mol^{-1}) of binding energies across different force fields (UFF, DREIDING) and charge methods (REPEAT, MEPO-ML, DDEC6). Lower values along the diagonal blocks indicate consistency with each force field, while higher cross-field values reflect differences in the Lennard-Jones parameters.

Within each forcefield, the three charge methods yield closely aligned energies, with UFF showing intra-set RMSEs of 0.33-0.50 kcal mol^{-1} and DREIDING showing 0.33-0.54 kcal mol^{-1}. In contrast, the deviations between UFF and DREIDING are considerably larger, typically 0.78-0.97 kcal mol^{-1}. This clustering by force field indicates that differences in the Lennard-Jones parameters, rather than the charge model, dominate variation in the computed binding energies. We therefore recommend UFF with DFT-quality charges for reliable ABS prediction. Across diverse MOFs and adsorbates, direct comparison to crystallography shows that classical GCMC reliably predicts physisorptive ABSs when the correct framework phase is used. Notably, predicted ABSs are accurate even when isotherms are poorly reproduced because they depend more on relative than absolute energetics. Accuracy diminishes with chemisorption, strong flexibility, and substantial experimental disorder. Outside these cases, routine simulations offer a practical way to resolve ABSs that are often inaccessible experimentally, providing a benchmark for future methods and a tool for diagnosing experiment-simulation discrepancies in MOF adsorption.

3.7 Complete Set of Validation Figures

This section compiles the complete set of RMSD and isotherm validation figures referenced throughout this chapter. The figures provide visual confirmation of the agreement between GCMC derived and experimentally determined binding sites, as quantified by the RMSD values reported in Table 3.2, together with comparisons between simulated and experimental adsorption isotherms where available. These visualizations support the discussion of binding site accuracy for the UFF with REPEAT charge model, which serves as the reference dataset for the communication, and are grouped here to preserve the flow of the chapter while presenting the full validation results.

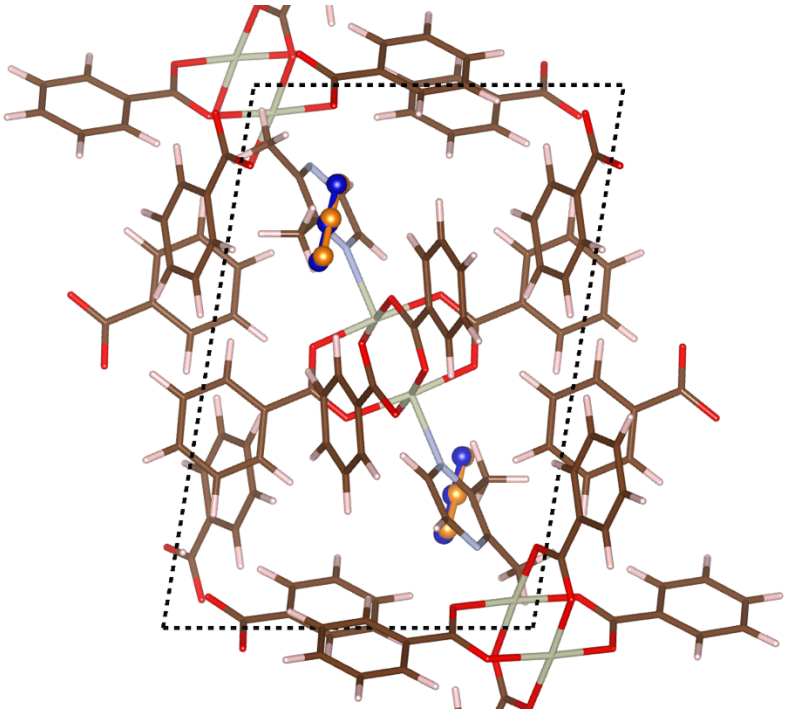

Figure 3.5. Comparison of binding site positions (experimental and simulated at 90 K, 1 bar) of CO_2 in $[Rh_2(bza)_4(dimethyl\text{-}pyz)]_n$. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

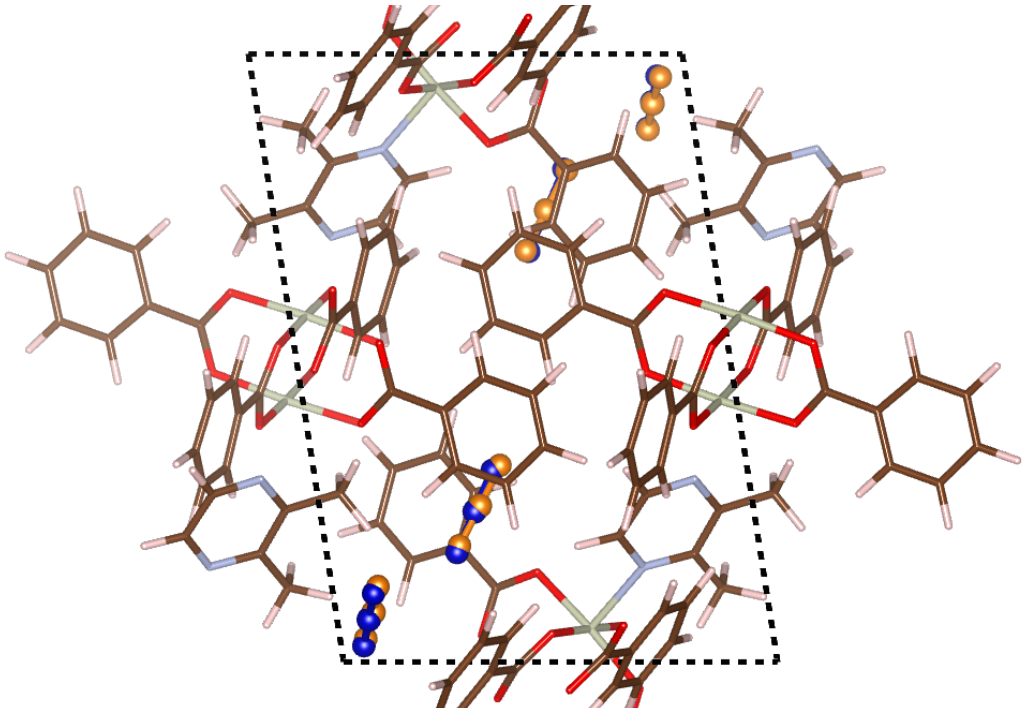

Figure 3.6. Comparison of binding site positions (experimental and simulated at 90 K, 17 bar) of CO_2 in $[Rh_2(bza)_4(dimethyl\text{-}pyz)]_n$. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

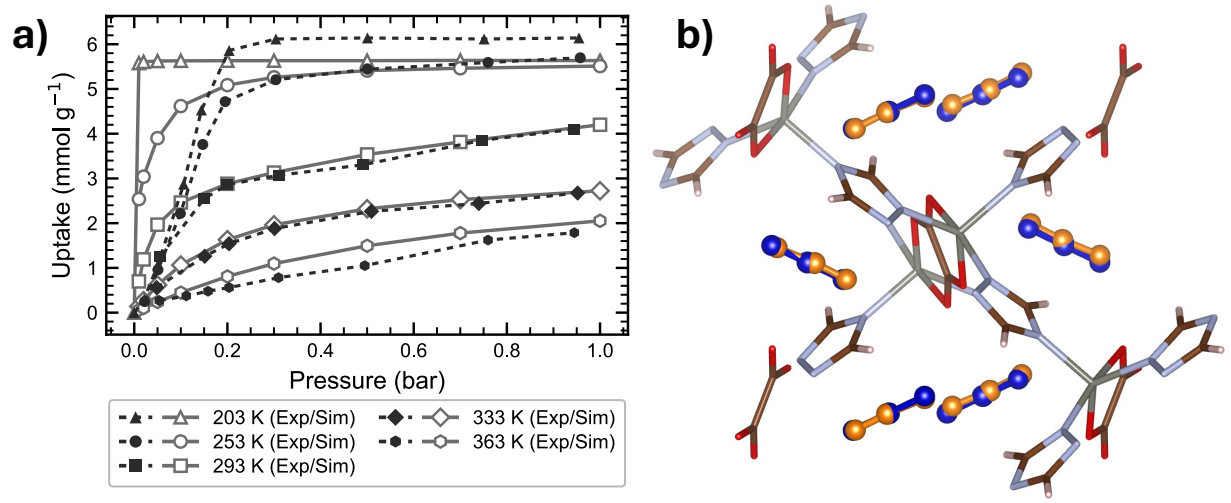

Figure 3.7. Comparison of binding site positions (experimental and simulated at 298 K, 10 bar) with CO_2 adsorption isotherms measured at 203 K, 253 K, 293 K, 333 K and 363 K for CALF-20. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

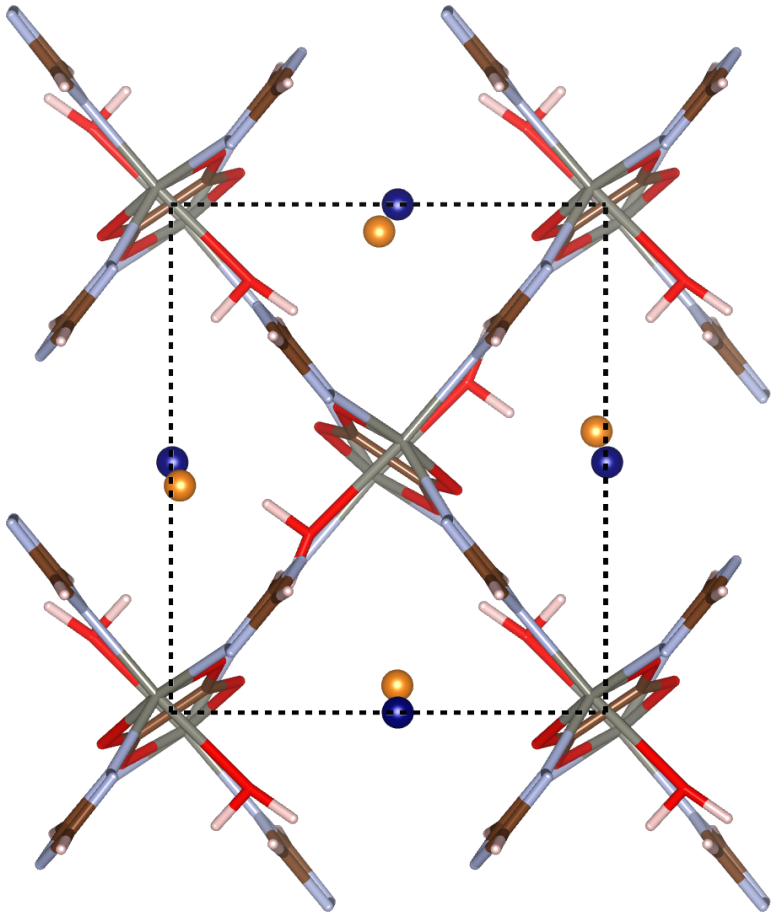

Figure 3.8. Comparison of binding site positions (experimental and simulated at 296 K, 100% relative humidity) of H_2O in calf-20. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres. *Hydrogens are excluded from the comparison since experimental data did not report orientation.*

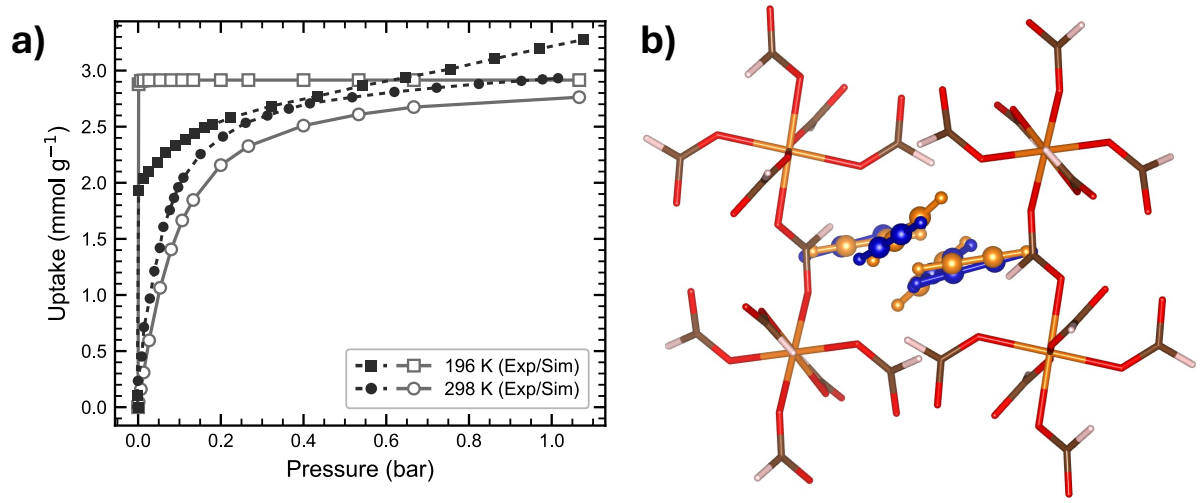

Figure 3.9. Comparison of binding site positions (experimental and simulated at 90 K, 1.01 bar) with C_2H_2 adsorption isotherms measured at 196 K and 298 K for $Mg(HCOO)_2$. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

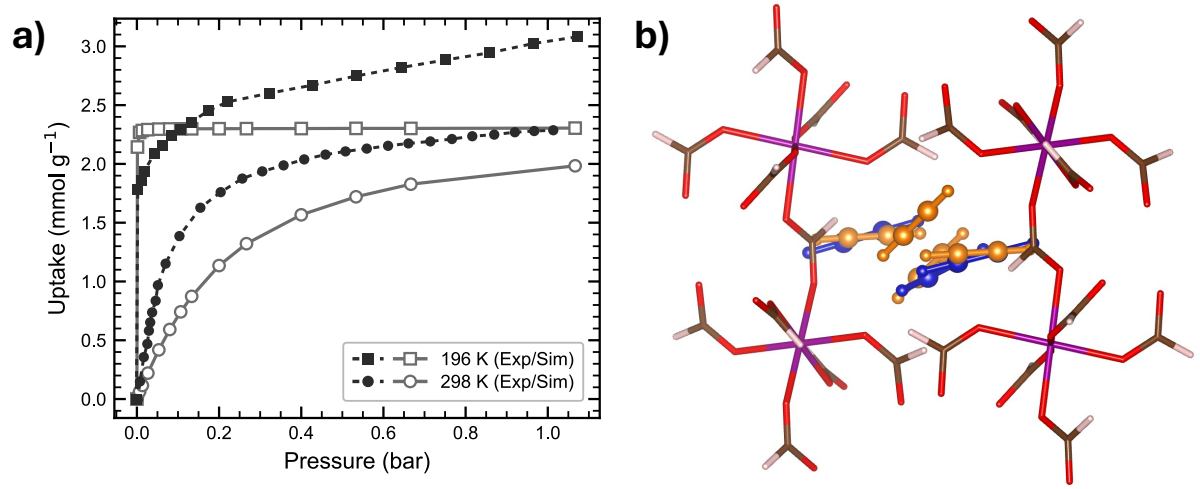

Figure 3.10. Comparison of binding site positions (experimental and simulated at 90 K, 1.01 bar) with C_2H_2 adsorption isotherms measured at 196 K and 298 K for $Mn(HCOO)_2$. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

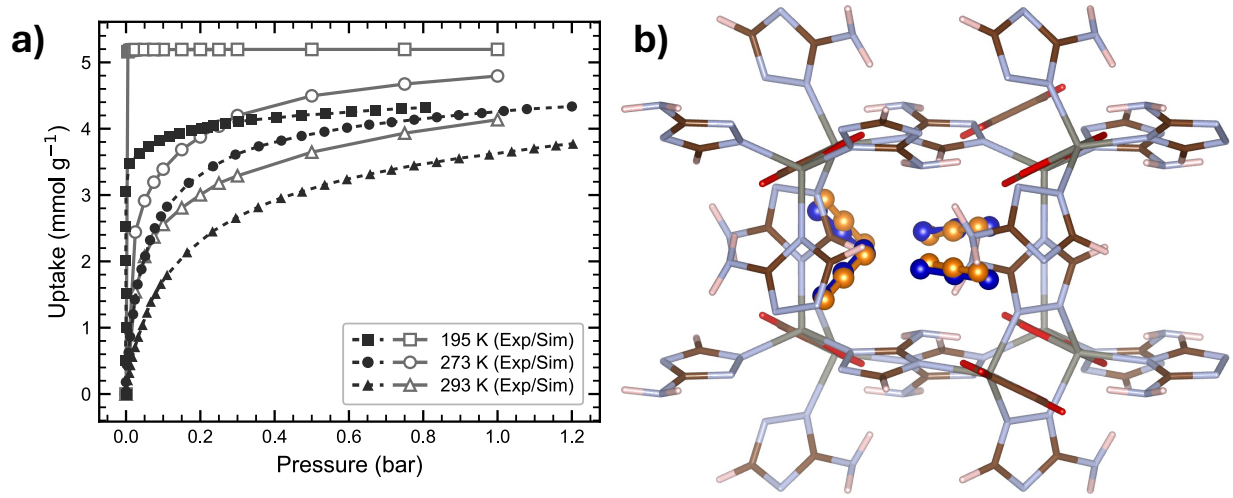

Figure 3.11. Comparison of binding site positions (experimental and simulated at 173 K, 0.85 bar) with CO_2 adsorption isotherms measured at 195 K, 273 K and 293 K for CALF-15. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

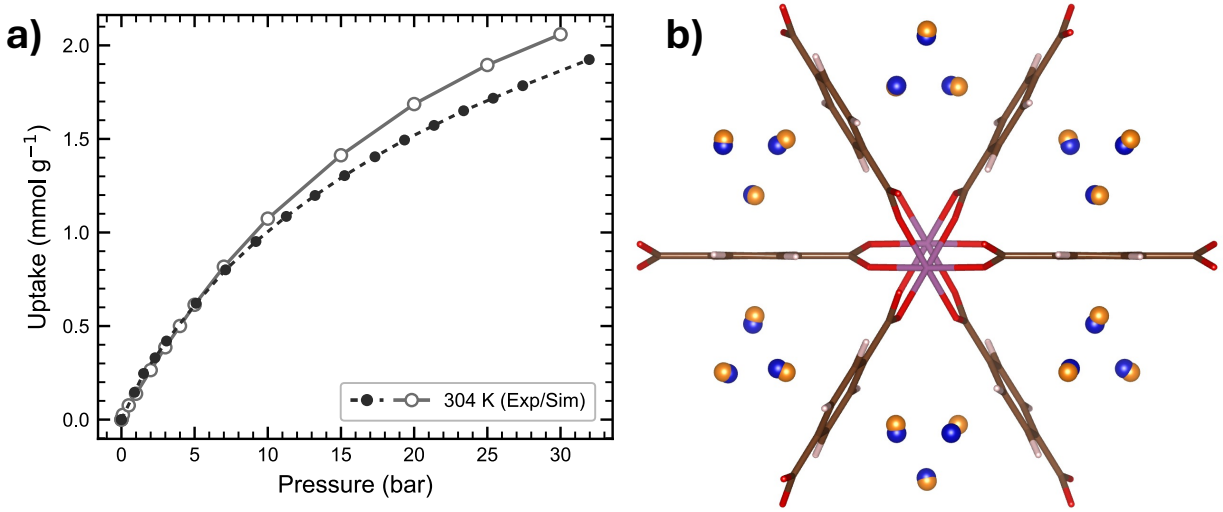

Figure 3.12. Comparison of binding site positions (experimental and simulated at 230 K, 9 bar) with CH_4 adsorption isotherms measured at 304 K for $Sc_2(BDC)_3$. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

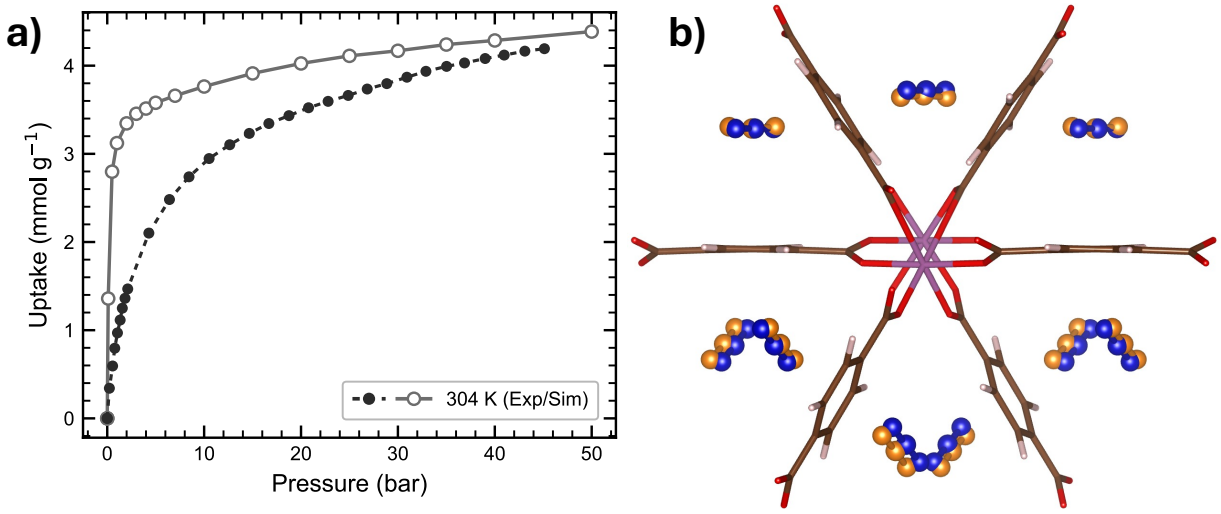

Figure 3.13. Comparison of binding site positions (experimental and simulated at 235 K, 1 bar) with CO_2 adsorption isotherms measured at 304 K for $Sc_2(BDC)_3$. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

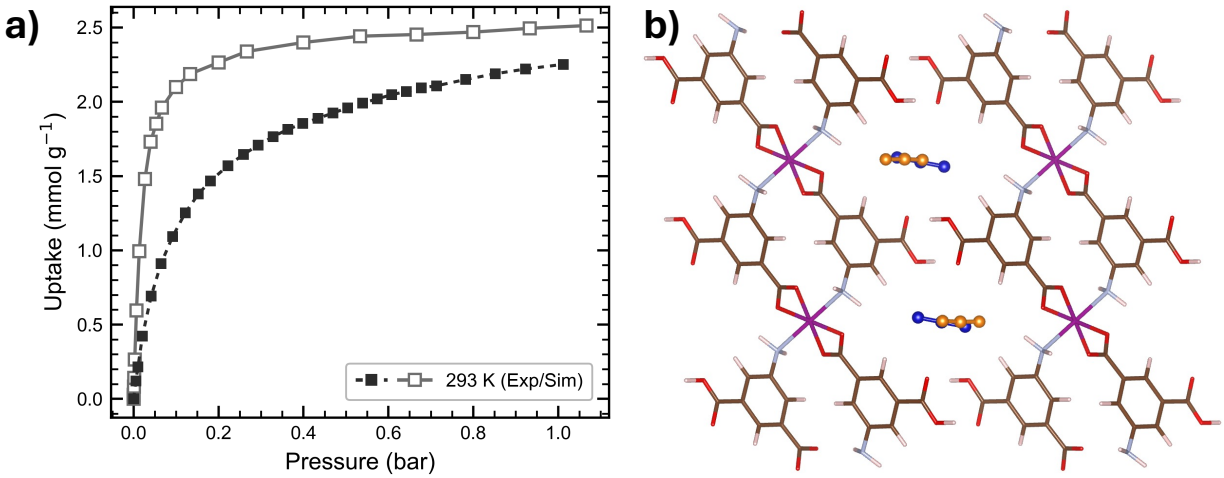

Figure 3.14. Comparison of binding site positions (experimental and simulated at 293 K, 1.1 bar) with CO_2 adsorption isotherms measured at 293 K for MUF-16(Mn). Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

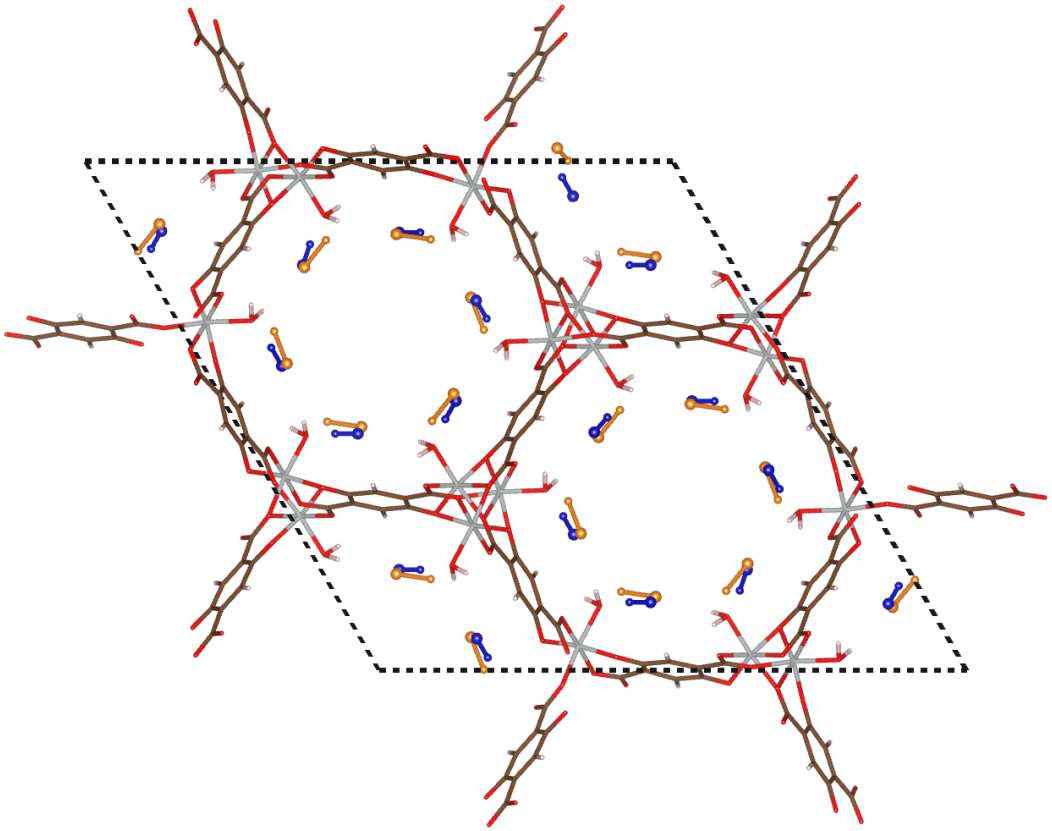

Figure 3.15. Comparison of binding site positions (experimental and simulated at 196 K, 0.4 bar) of NO in MOF-74(Ni) (H_2O capped). Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres. The nitrogen atoms are represented as smaller spheres to distinguish orientation.

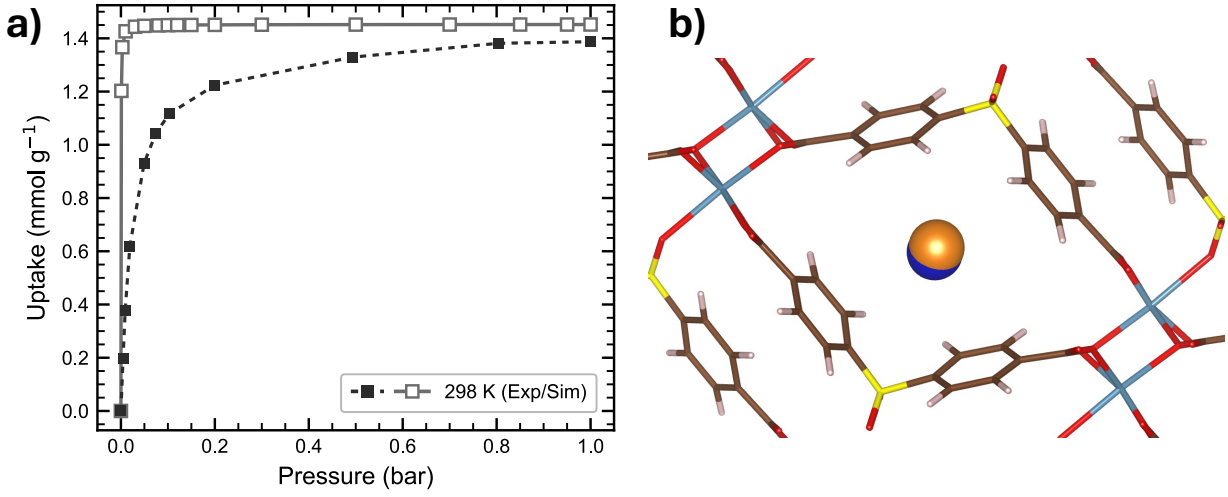

Figure 3.16. Comparison of binding site positions (experimental and simulated at 298 K, 1 bar) with Xe adsorption isotherms measured at 298 K for SBMOF-1. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

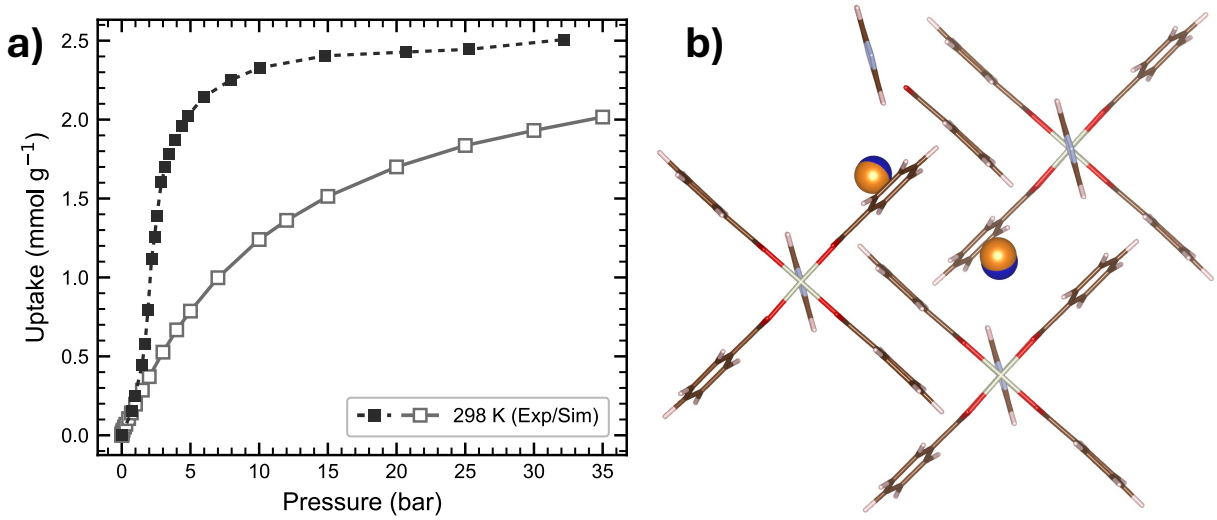

Figure 3.17. Comparison of binding site positions (experimental and simulated at 298 K, 80 bar) with Ar adsorption isotherms measured at 298 K for $[Rh_2(bza)_4(pyz)]_n$. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

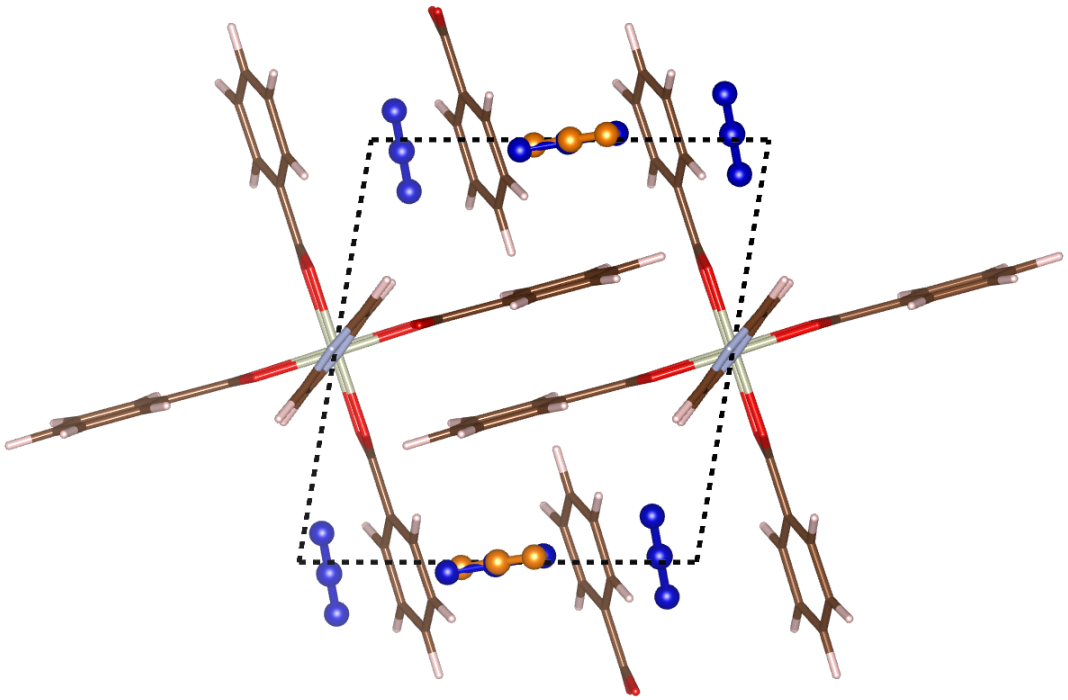

Figure 3.18. Comparison of binding site positions (experimental and simulated at 298 K, 35 bar) of CO_2 in $[Rh_2(bza)_4(pyz)]_n$. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

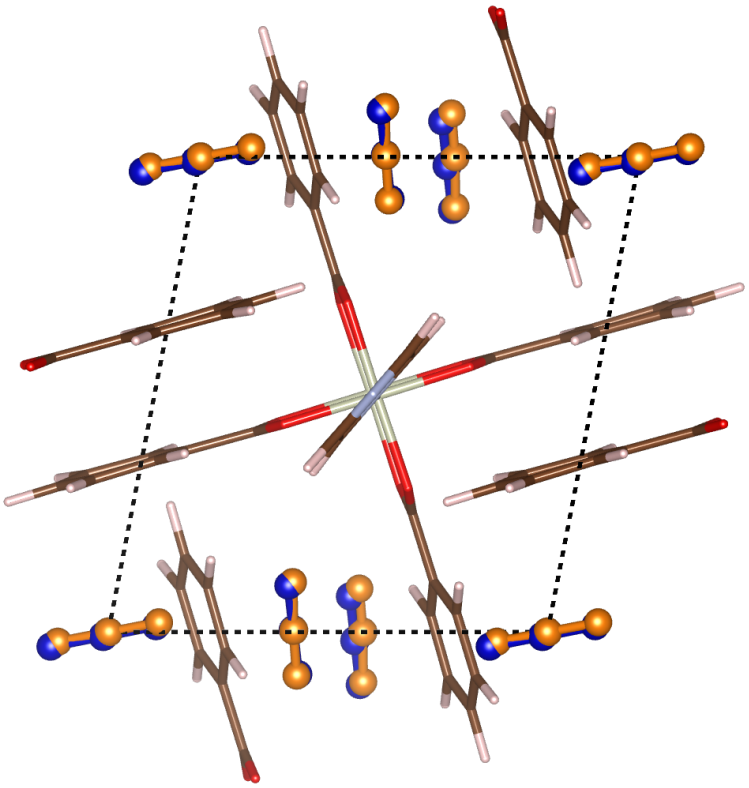

Figure 3.19. Comparison of binding site positions (experimental and simulated at 93 K, 1.01 bar) of CO_2 in ($[Rh_2(O_2CPh)_4(pyz)]_n$). Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

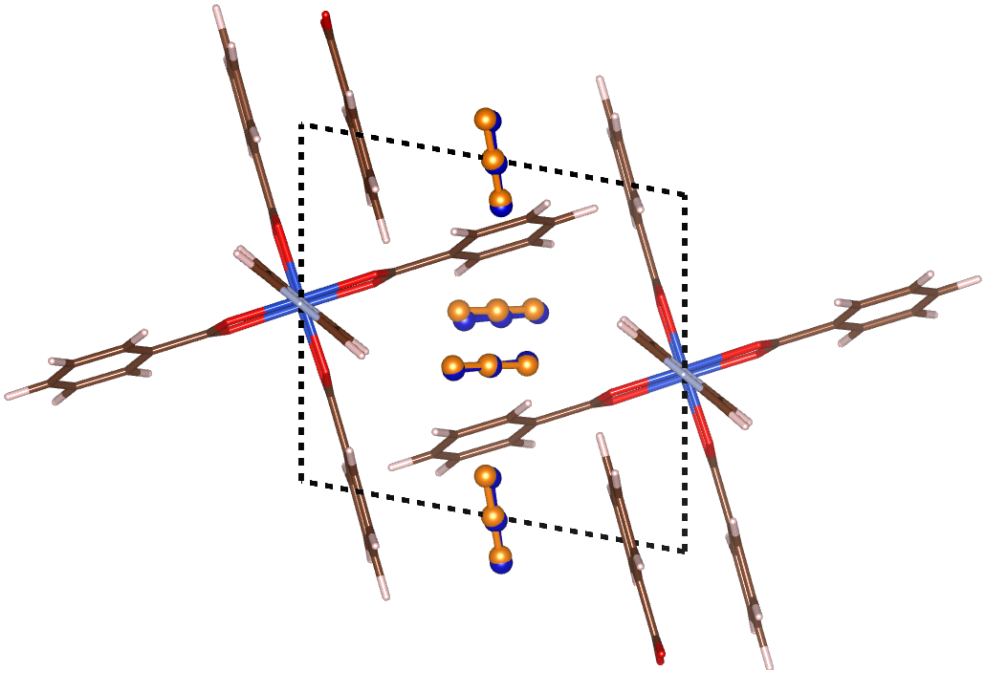

Figure 3.20. Comparison of binding site positions (experimental and simulated at 193 K, 1.01 bar) of CO_2 in $[Cu_2(bza)_4(pyz)]_n$. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

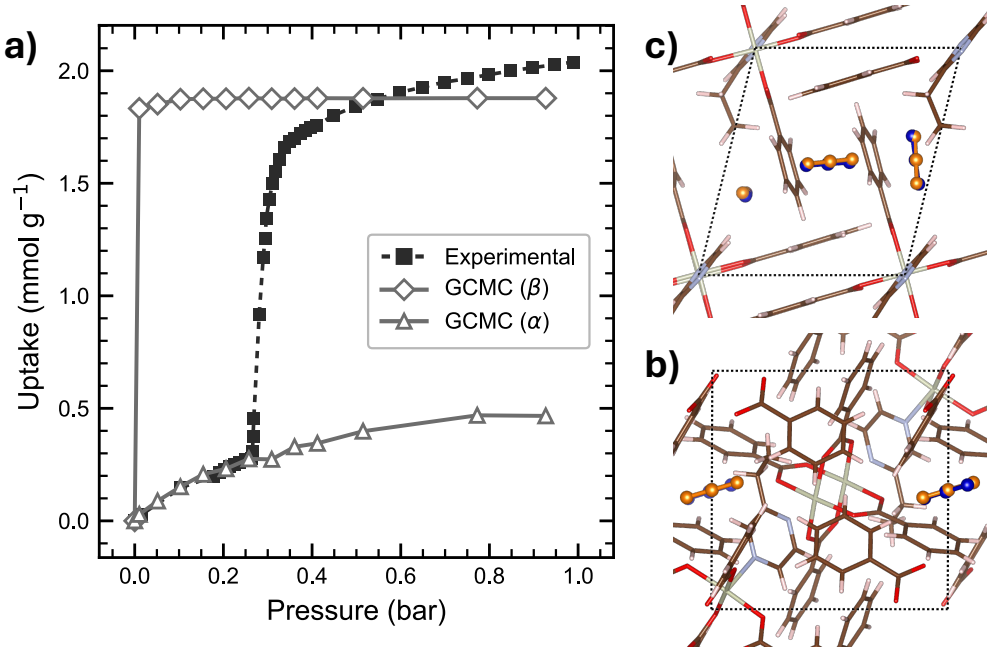

Figure 3.21. Comparison of binding site positions (experimental and simulated at 298 K and 64 bar in subfigure (c), 90 K and 36 bar in subfigure (b)) with CO_2 adsorption isotherms measured at 195 K for the flexible MOF $[Rh_2(bza)_4(2\text{-epyz})]_n$. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

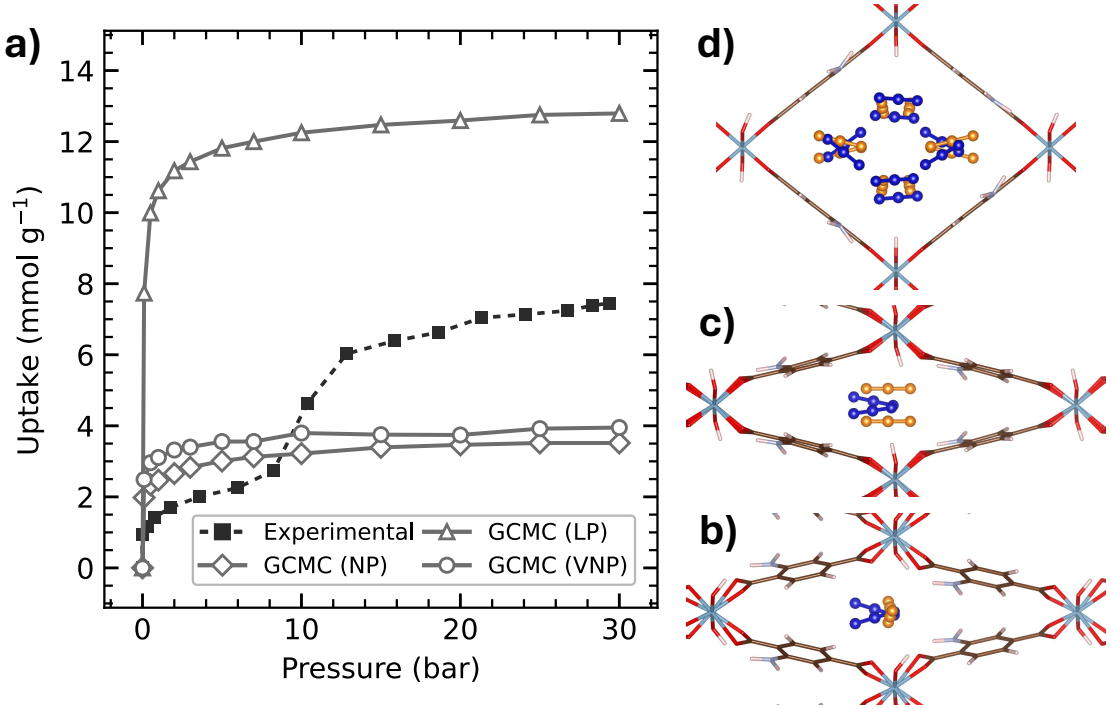

Figure 3.22. Comparison of binding site positions (experimental and simulated at 253 K, 3 bar (b), 9.5 bar (c) and 18 bar (d)) with CO_2 adsorption isotherms measured at 283 K for the flexible MOF MIL-53(Al). Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

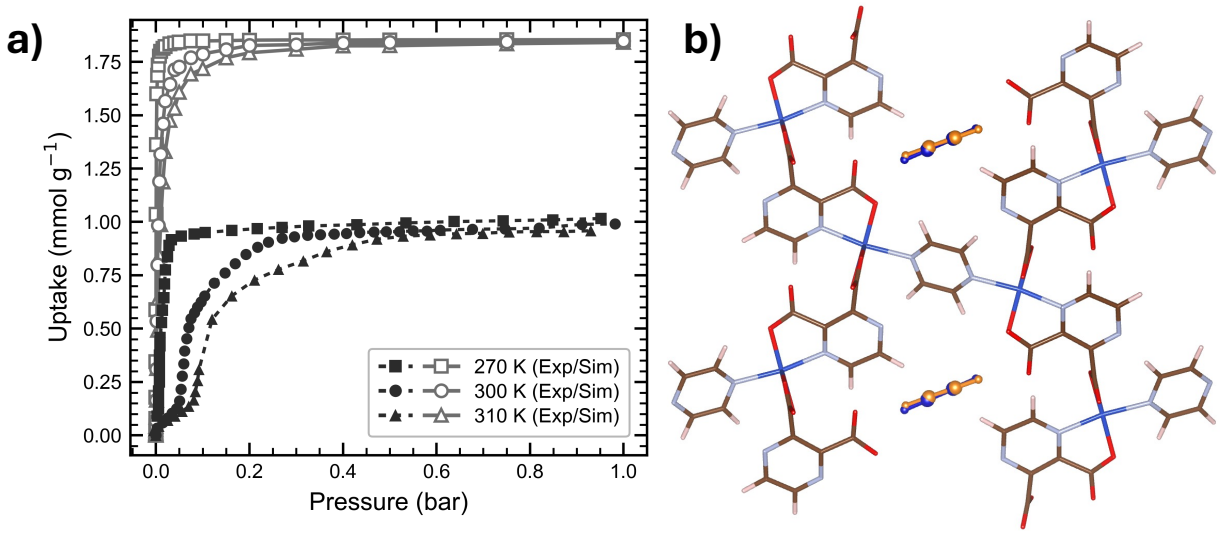

Figure 3.23. Comparison of binding site positions (experimental and simulated at 170 K, 0.1 bar) with C_2H_2 adsorption isotherms measured at 270 K, 300 K and 310 K for CPL-1. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

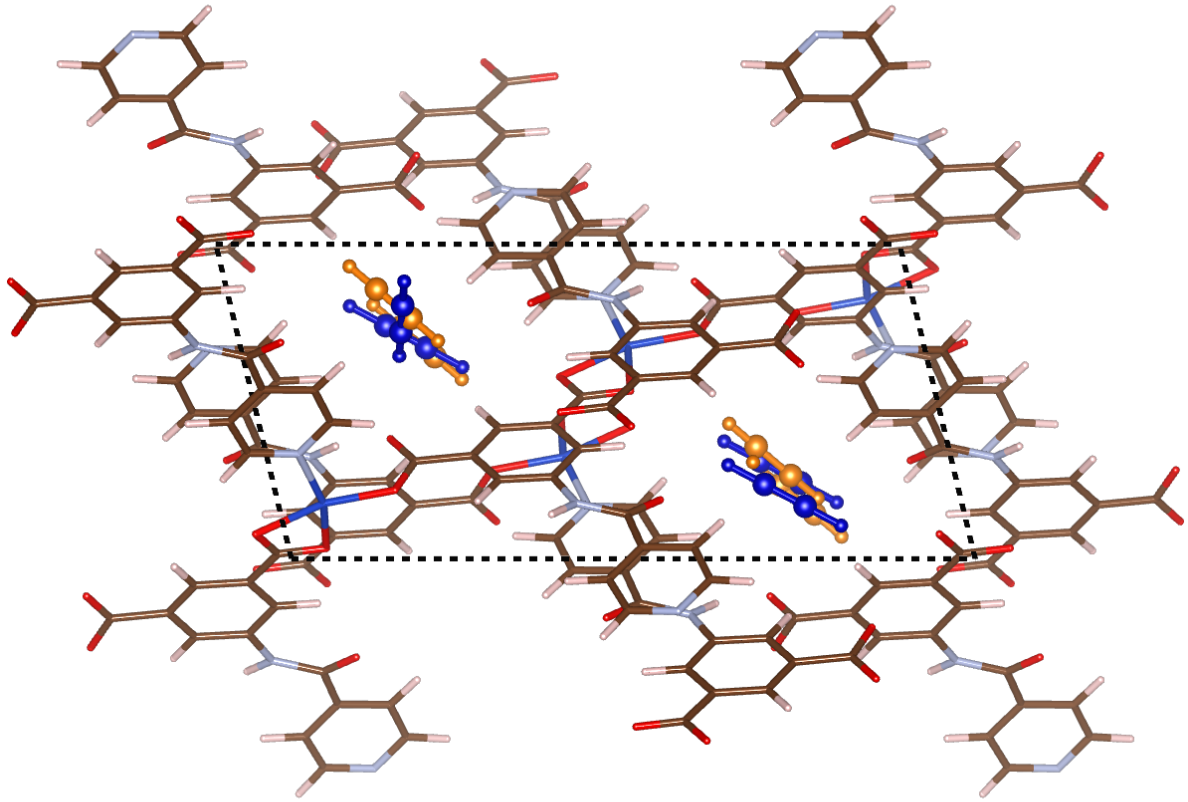

Figure 3.24. Comparison of binding site positions (experimental and simulated at 110 K, 0.18 bar) of C_2H_2 in INAIP-Cu. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

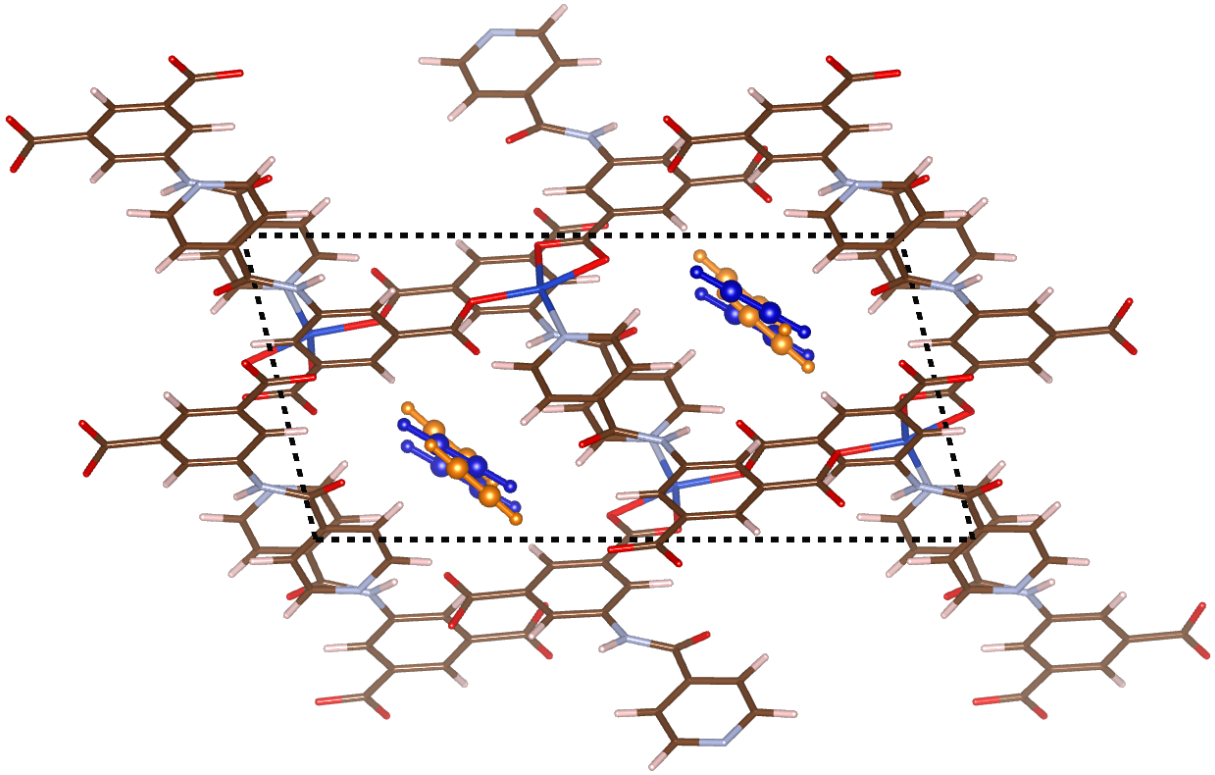

Figure 3.25. Comparison of binding site positions (experimental and simulated at 153 K, 0.85 bar) of C_2H_2 in INAIP-Cu. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

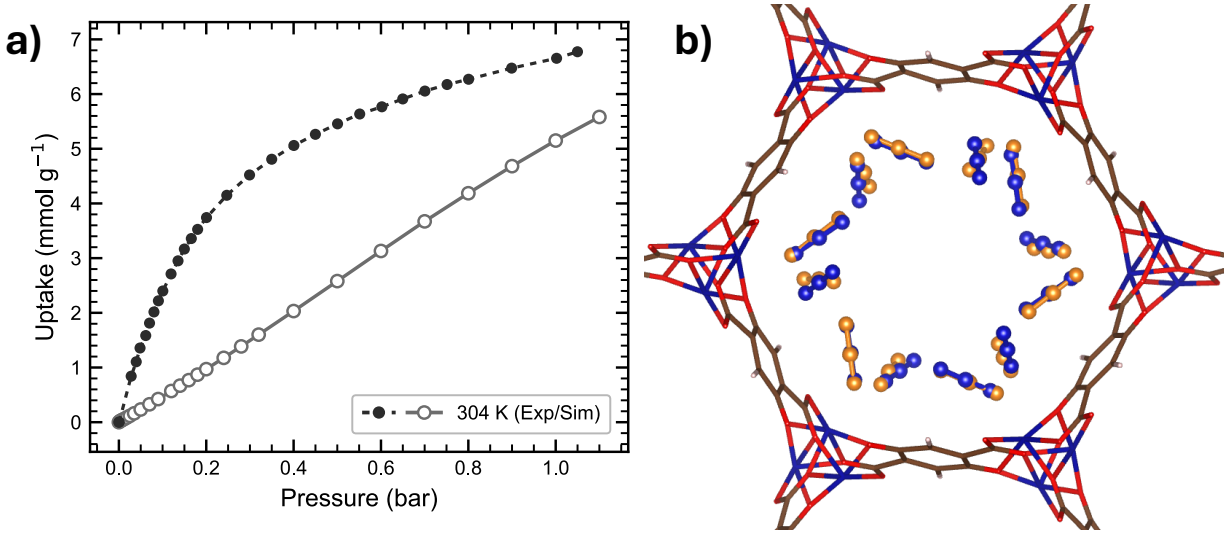

Figure 3.26. Comparison of binding site positions (experimental and simulated at 196 K, 1.06 bar) with CO_2 adsorption isotherms measured at 304 K for MOF-74(Co). Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

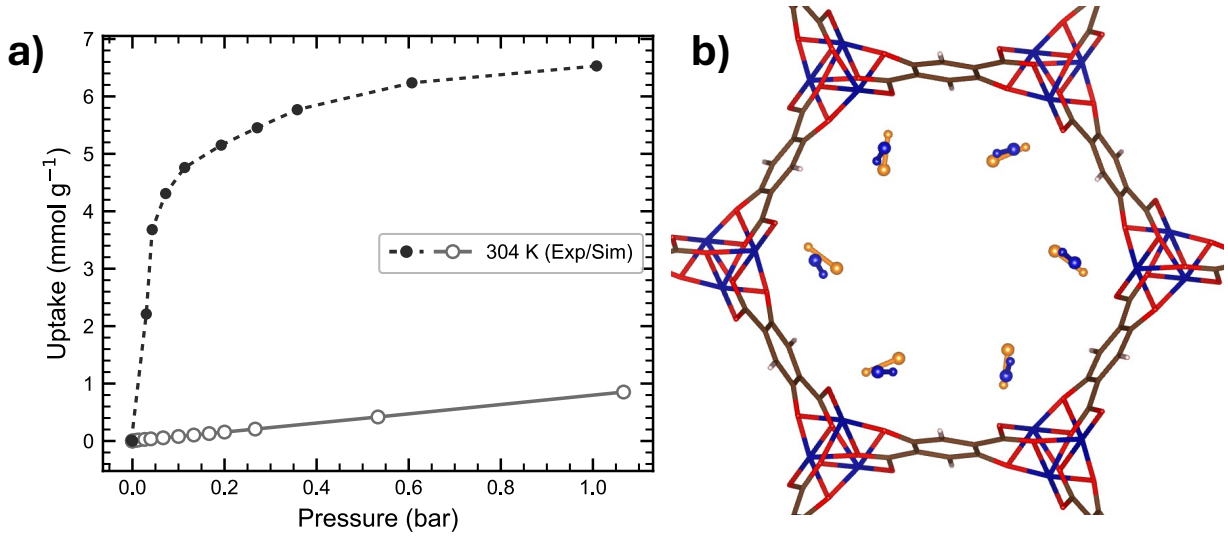

Figure 3.27. Comparison of binding site positions (experimental and simulated at 298 K, 1.01 bar) with NO adsorption isotherms measured at 304 K for MOF-74(Co). Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres. The nitrogen atoms are represented as smaller spheres to distinguish orientation.

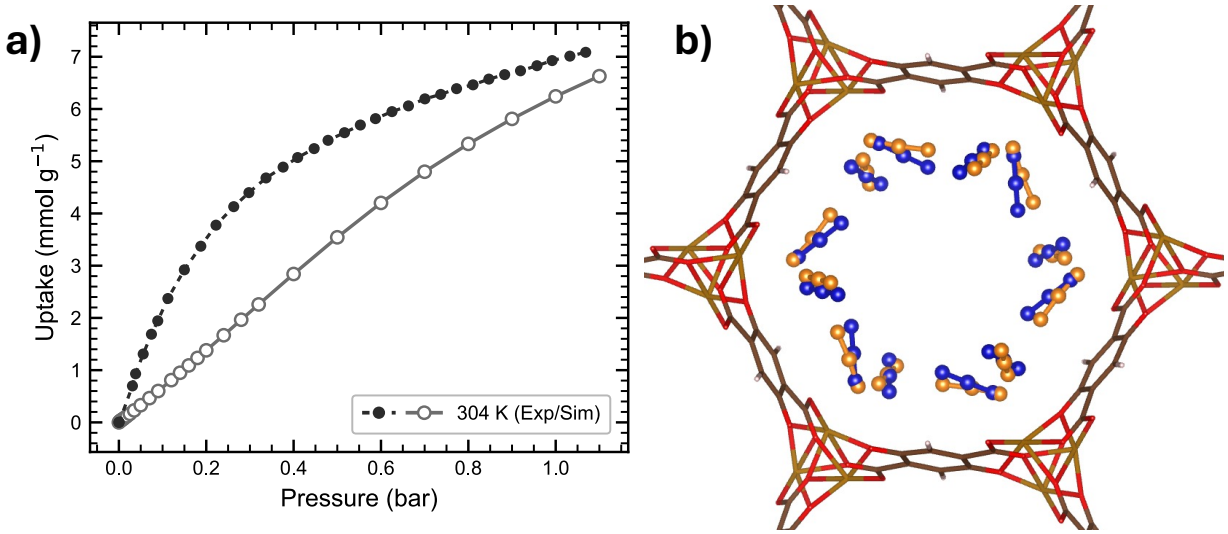

Figure 3.28. Comparison of binding site positions (experimental and simulated at 298 K, 1.01 bar) with CO_2 adsorption isotherms measured at 304 K for MOF-74(Fe). Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

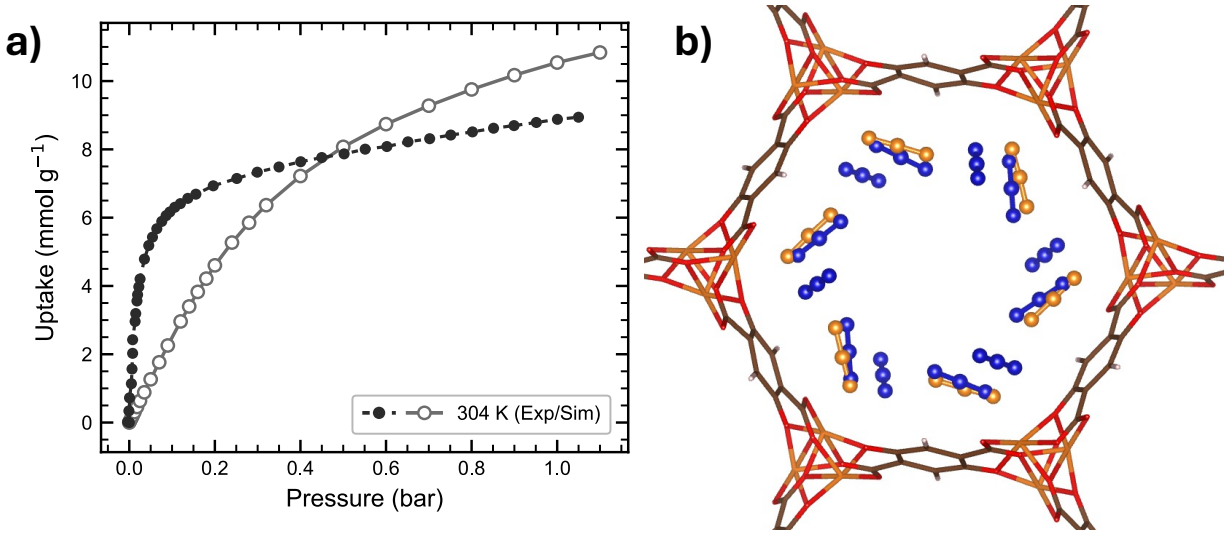

Figure 3.29. Comparison of binding site positions (experimental and simulated at 196 K, 1.06 bar) with CO_2 adsorption isotherms measured at 304 K for MOF-74(Mg). Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

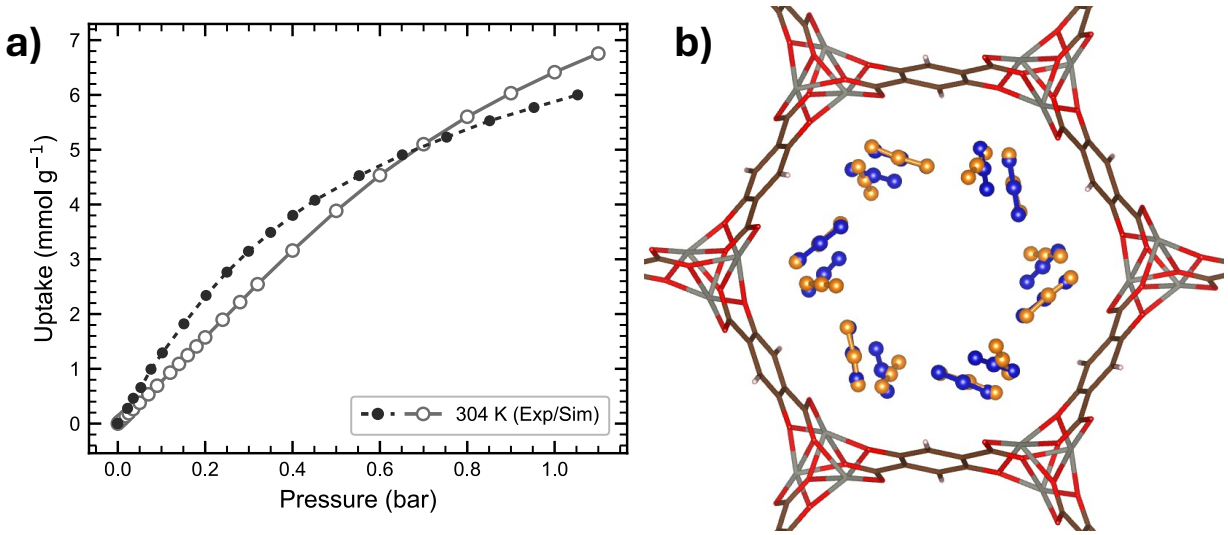

Figure 3.30. Comparison of binding site positions (experimental and simulated at 196 K, 1.06 bar) with CO_2 adsorption isotherms measured at 304 K for MOF-74(Zn). Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

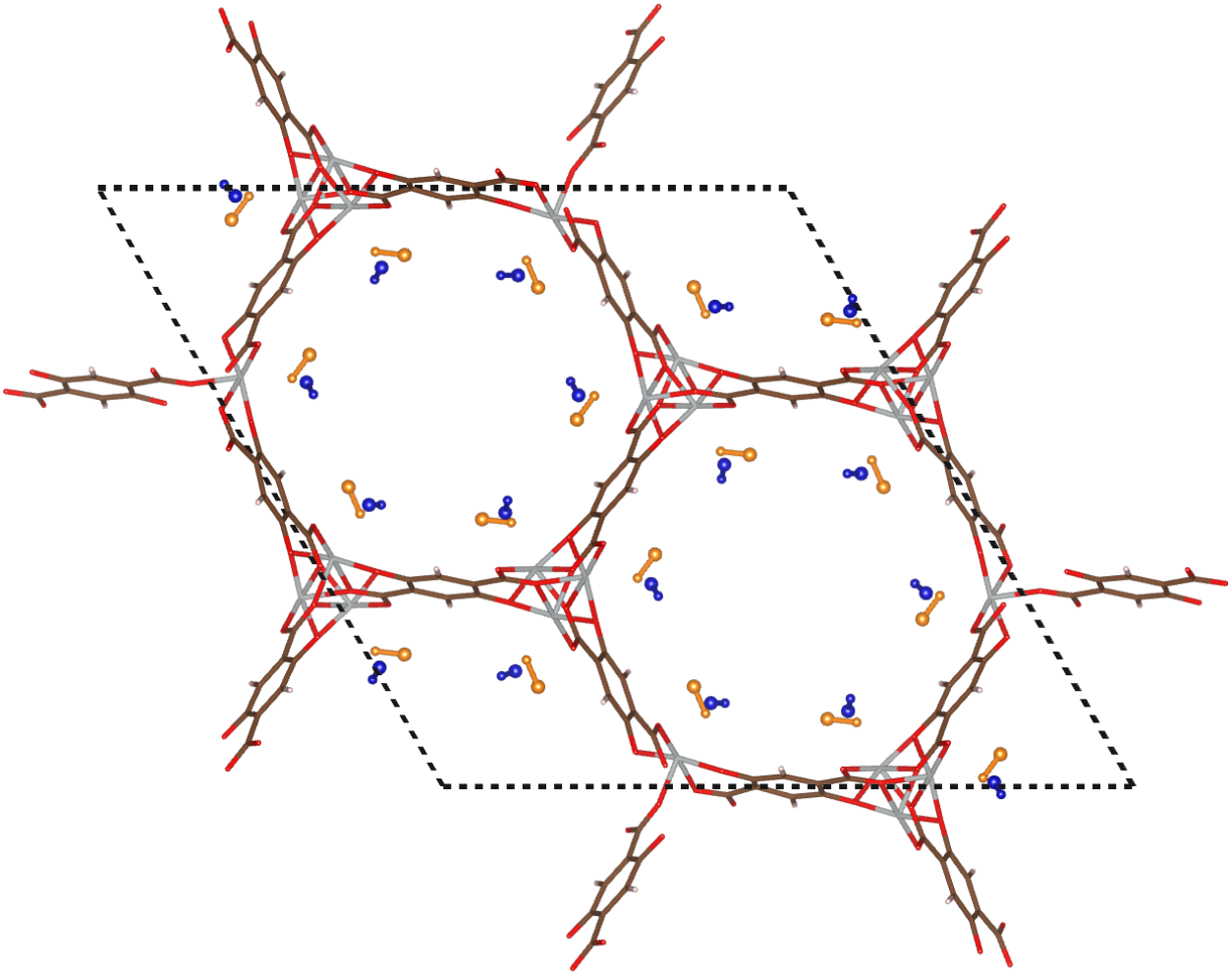

Figure 3.31. Comparison of binding site positions (experimental and simulated at 196 K, 0.4 bar) of NO in MOF-74(Ni) (OMS). Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres. The nitrogen atoms are represented as smaller spheres to distinguish orientation.

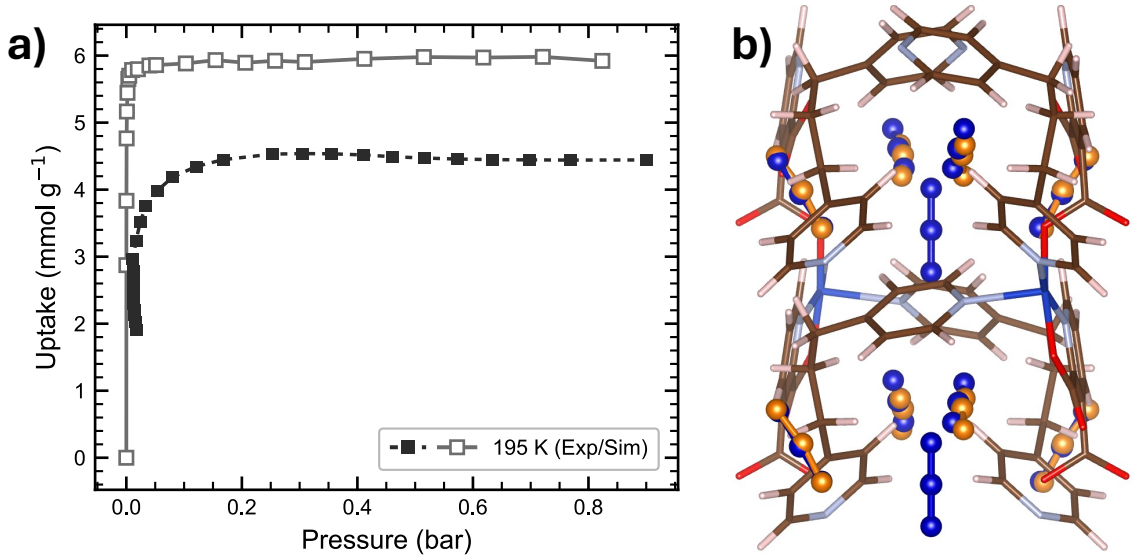

Figure 3.32. Comparison of binding site positions (experimental and simulated at 193 K, 1.01 bar) with CO_2 adsorption isotherms measured at 195 K for $[Cu_2(pyrdc)_2(bpp)_2]_n$. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

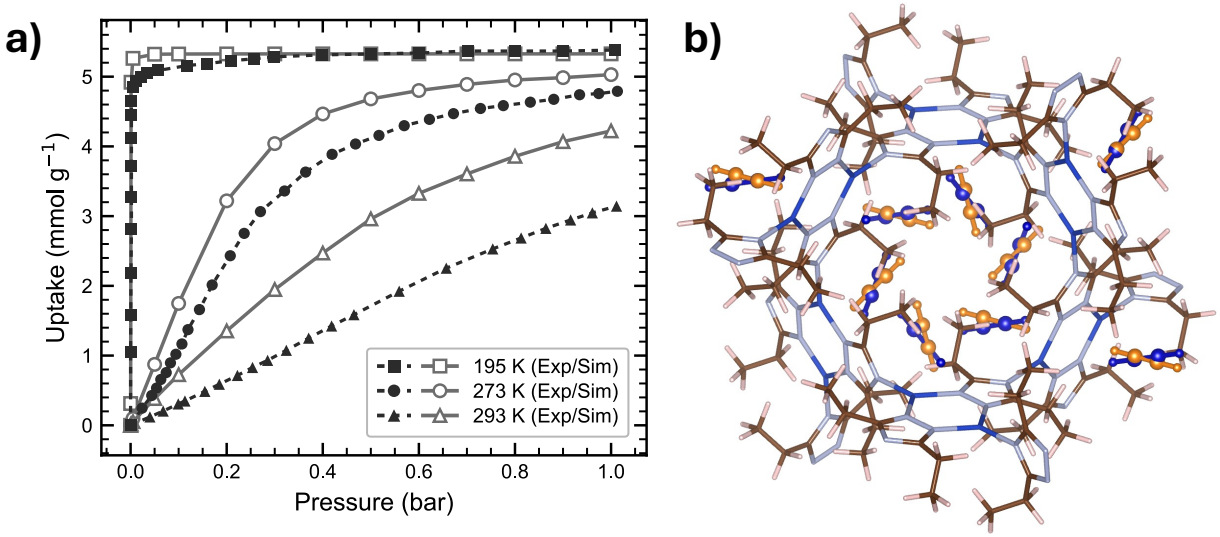

Figure 3.33. Comparison of binding site positions (experimental and simulated at 195 K, 10.13 bar) with C_2H_2 adsorption isotherms measured at 195 K, 273 K and 293 K for MAF-2(Zn). Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

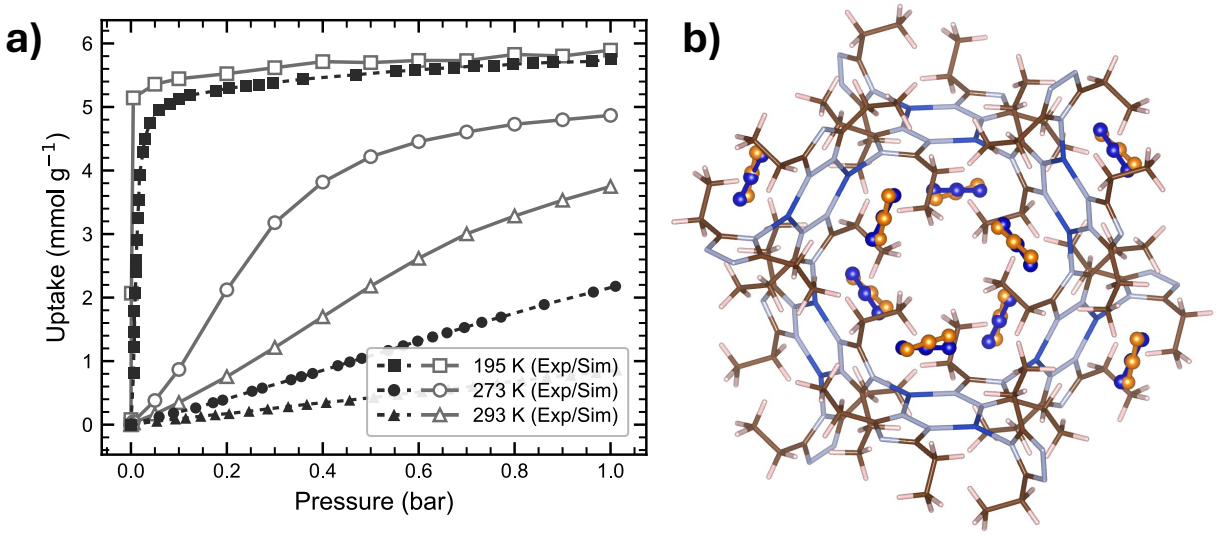

Figure 3.34. Comparison of binding site positions (experimental and simulated at 195 K, 20.27 bar) with CO_2 adsorption isotherms measured at 195 K, 273 K and 293 K for MAF-2(Zn). Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

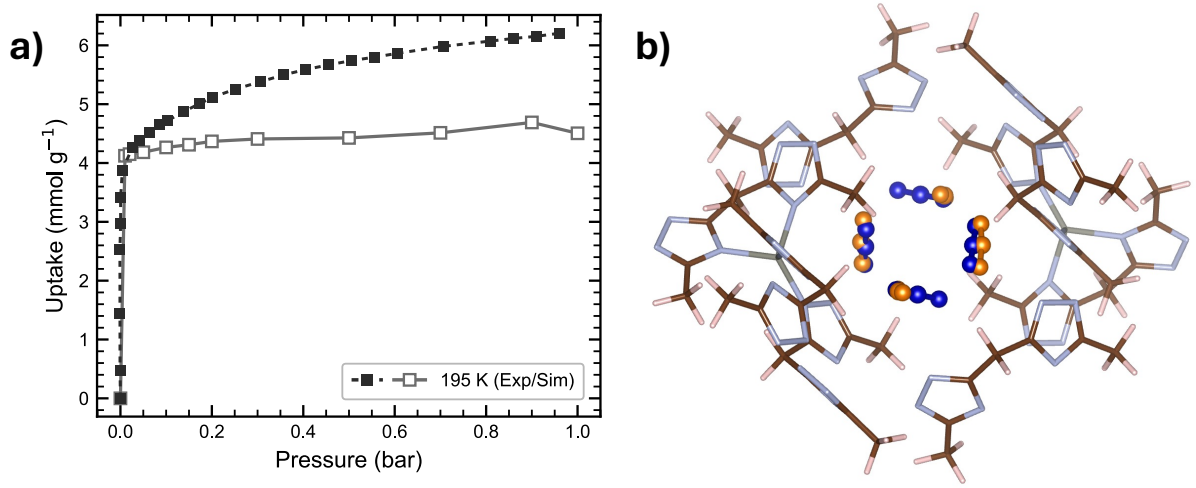

Figure 3.35. Comparison of binding site positions (experimental and simulated at 195 K, 0.79 bar) with CO_2 adsorption isotherms measured at 195 K for MAF-23(Zn). Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

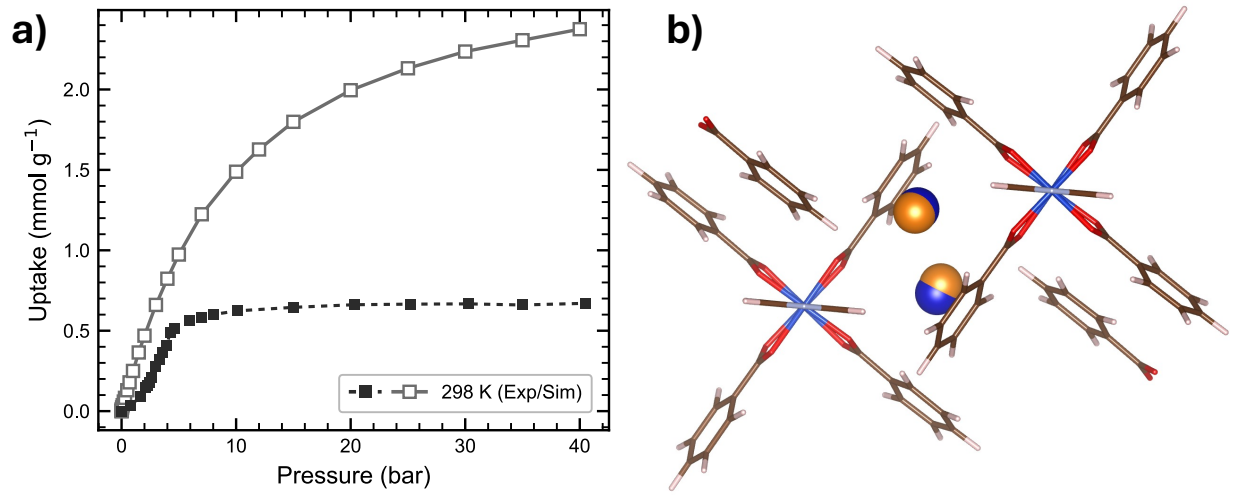

Figure 3.36. Comparison of binding site positions (experimental and simulated at 298 K, 100 bar) with Ar adsorption isotherms measured at 298 K for $[Cu_2(bza)_4(pyz)]_n$. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

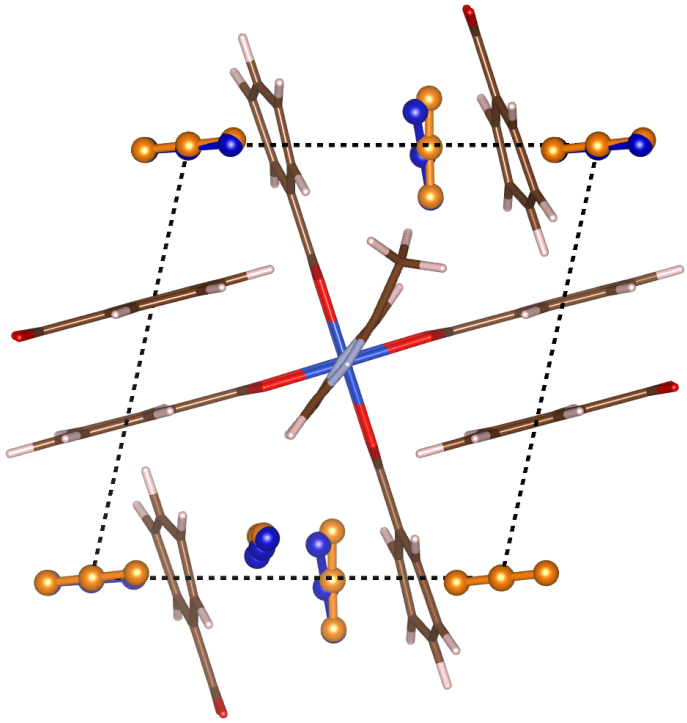

Figure 3.37. Comparison of binding site positions (experimental and simulated at 90 K, 1 bar) of CO_2 in $[Cu_2(bza)_4(methyl\text{-}pyz)]_n$. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

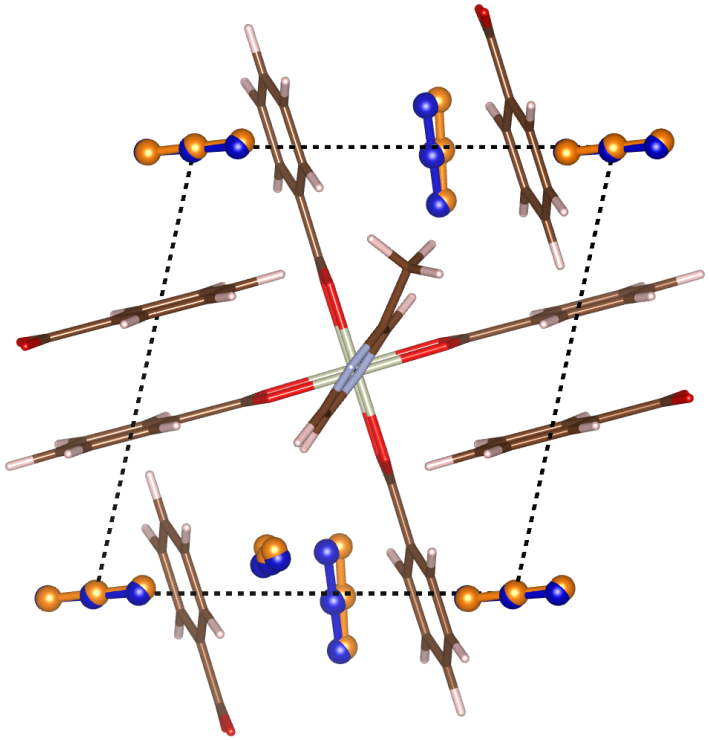

Figure 3.38. Comparison of binding site positions (experimental and simulated at 90 K, 1 bar) of CO_2 in $[Rh_2(bza)_4(methyl\text{-}pyz)]_n$. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

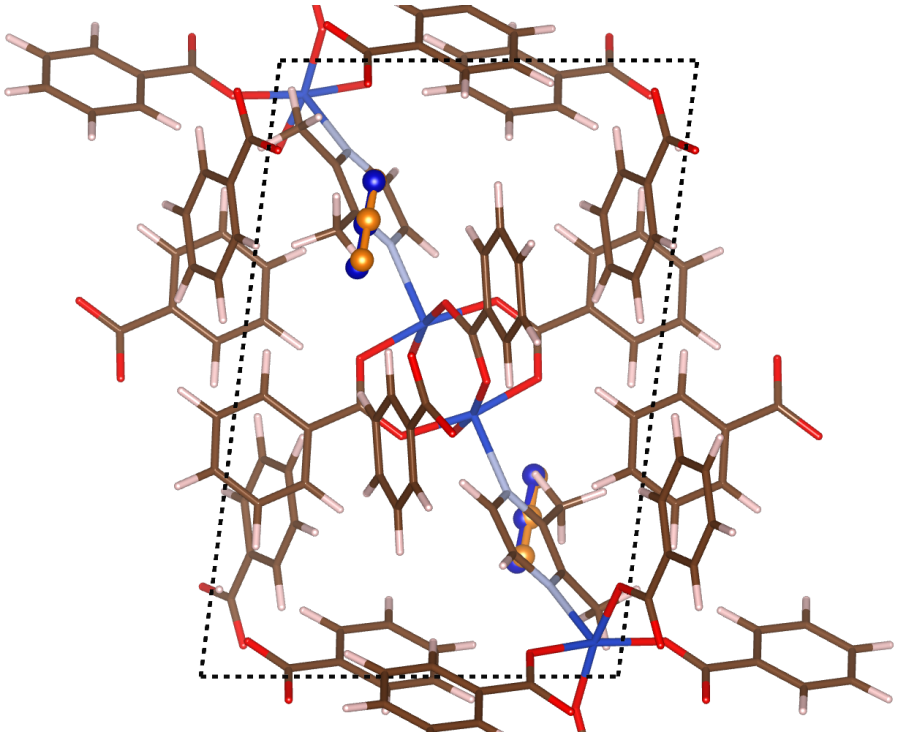

Figure 3.39. Comparison of binding site positions (experimental and simulated at 90 K, 1 bar) of CO_2 in $[Cu_2(bza)_4(dimethyl\text{-}pyz)]_n$. Experimental binding sites are shown as orange spheres, and simulated binding sites are shown as blue spheres.

3.8 Conclusion

This chapter evaluated how well conventional atomistic simulations reproduce experimentally determined adsorption binding sites in MOFs. Direct comparison with crystallographic data across a diverse set of MACs shows that classical GCMC simulation reliably predict the location of physisorptive binding sties with the correct framework phase is used. The agreement between simulated and experimental bindings sites is consistently high, with RMSD values well within the resolution of the APDs.

A key result is that accurate binding site prediction does not require quantitative agreement to adsorption isotherms. Even when simulation uptakes deviate from experiment, binding site locations remain well reproduced because they depend primarily on relative, rather than absolute, interactions energetics. In contrast, accuracy decreases in systems dominated by chemisorption, strong framework flexibility, or significant experimental disorder, where generic force fields are no longer adequate.

Overall, these results demonstrate that routine classical simulations provide a robust and computationally efficient approach for identifying adsorption binding sites in many MOFs, despite their limitations for global adsorption properties. This establishes classical GCMC as a practical tool for interpreting experimental absorption behaviour and for benchmarking more advanced simulation methodologies.

3.9 References

(18) Lin, J.-B.; Nguyen, T. T. T.; Vaidhyanathan, R.; Burner, J.; Taylor, J. M.; Durekova, H.; Akhtar, F.; Mah, R. K.; Ghaffari-Nik, O.; Marx, S.; Fylstra, N.; Iremonger, S. S.; Dawson, K. W.; Sarkar, P.; Hovington, P.; Rajendran, A.; Woo, T. K.; Shimizu, G. K. H. A Scalable Metal-Organic Framework as a Durable Physisorbent for Carbon Dioxide Capture. *Science (1979).* **2021**, *374* (6574), 1464–1469. https://doi.org/10.1126/science.abi7281.

(39) Easun, T. L.; Moreau, F.; Yan, Y.; Yang, S.; Schröder, M. Structural and Dynamic Studies of Substrate Binding in Porous Metal–Organic Frameworks. *Chem. Soc. Rev.* **2017**, *46* (1), 239–274. https://doi.org/10.1039/C6CS00603E.

(40) Han, Z.; Yu, K.-H.; Wang, K.-Y.; Zhou, H.-C. Binding Sites of Automobile Exhaust Gases on Metal–Organic Frameworks: Advances and Perspectives. *Energy & Fuels* **2025**, *39* (13), 6151–6163. https://doi.org/10.1021/acs.energyfuels.5c00552.

(42) Lucier, B. E. G.; Chen, S.; Huang, Y. Characterization of Metal–Organic Frameworks: Unlocking the Potential of Solid-State NMR. *Acc. Chem. Res.* **2018**, *51* (2), 319–330. https://doi.org/10.1021/acs.accounts.7b00357.

(43) Martins, V.; Lucier, B. E. G.; Liu, Z.; Liang, H.; Zheng, A.; Terskikh, V. V.; Zhang, W.; Desveaux, B. E.; Huang, Y. Cold, Hot, Dry, and Wet: Locations and Dynamics of CO2 and H2O Co-Adsorbed in an Ultramicroporous MOF. *Inorg. Chem.* **2023**, *62* (28), 11152–11167. https://doi.org/10.1021/acs.inorgchem.3c01251.

(45) Queen, W. L.; Hudson, M. R.; Bloch, E. D.; Mason, J. A.; Gonzalez, M. I.; Lee, J. S.; Gygi, D.; Howe, J. D.; Lee, K.; Darwish, T. A.; James, M.; Peterson, V. K.; Teat, S. J.; Smit, B.; Neaton, J. B.; Long, J. R.; Brown, C. M. Comprehensive Study of Carbon Dioxide Adsorption in the Metal–Organic Frameworks M2(Dobdc) (M = Mg, Mn, Fe, Co, Ni, Cu, Zn). *Chem. Sci.* **2014**, *5* (12), 4569–4581. https://doi.org/10.1039/C4SC02064B.

(54) Boyd, P. G.; Chidambaram, A.; García-Díez, E.; Ireland, C. P.; Daff, T. D.; Bounds, R.; Gładysiak, A.; Schouwink, P.; Moosavi, S. M.; Maroto-Valer, M. M.; Reimer, J. A.; Navarro, J. A. R.; Woo, T. K.; Garcia, S.; Stylianou, K. C.; Smit, B. Data-Driven Design of Metal–Organic Frameworks for Wet Flue Gas CO2 Capture. *Nature* **2019**, *576* (7786), 253–256. https://doi.org/10.1038/s41586-019-1798-7.

(55) Vaidhyanathan, R.; Iremonger, S. S.; Shimizu, G. K. H.; Boyd, P. G.; Alavi, S.; Woo, T. K. Direct Observation and Quantification of CO2 Binding Within an Amine-Functionalized Nanoporous Solid. *Science (1979).* **2010**, *330* (6004), 650–653. https://doi.org/10.1126/science.1194237.

(64) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114* (25), 10024–10035. https://doi.org/10.1021/ja00051a040.

(67) Fischer, M.; Hoffmann, F.; Fröba, M. New Microporous Materials for Acetylene Storage and C2H2/CO2 Separation: Insights from Molecular Simulations. *ChemPhysChem* **2010**, *11* (10), 2220–2229. https://doi.org/10.1002/cphc.201000126.

(69) Boato, G.; Casanova, G. A Self-Consistent Set of Molecular Parameters for Neon, Argon, Krypton and Xenon. *Physica* **1961**, *27* (6), 571–589. https://doi.org/10.1016/0031-8914(61)90072-6.

(77) Ozin, G. *CALF-20: A carbon capture success story*. https://www.advancedsciencenews.com/calf-20-a-carbon-capture-success-story/.

(78) Craig Bettenhausen. Numat Plans MOF Factory in Chicago. *C&EN Global Enterprise* **2023**, *101* (38), 13–13. https://doi.org/10.1021/cen-10138-buscon14.

(79) Mitch Jacoby. NuMat Technologies. *C&EN Global Enterprise* **2016**, *94* (43), 32–32. https://doi.org/10.1021/cen-09443-cover8.

(80) Chen, Y.; Lu, W.; Schröder, M.; Yang, S. Analysis and Refinement of Host–Guest Interactions in Metal–Organic Frameworks. *Acc. Chem. Res.* **2023**, *56* (19), 2569–2581. https://doi.org/10.1021/acs.accounts.3c00243.

(81) Li, A.; Bueno-Perez, R.; Wiggin, S.; Fairen-Jimenez, D. Enabling Efficient Exploration of Metal–Organic Frameworks in the Cambridge Structural Database. *CrystEngComm* **2020**, *22* (43), 7152–7161. https://doi.org/10.1039/D0CE00299B.

(82) Ran, Y. A.; Sharma, S.; Balestra, S. R. G.; Li, Z.; Calero, S.; Vlugt, T. J. H.; Snurr, R. Q.; Dubbeldam, D. RASPA3: A Monte Carlo Code for Computing Adsorption and Diffusion in Nanoporous Materials and Thermodynamics Properties of Fluids. *J. Chem. Phys.* **2024**, *161* (11). https://doi.org/10.1063/5.0226249.

(83) Cleeton, C.; de Oliveira, F. L.; Neumann, R. F.; Farmahini, A. H.; Luan, B.; Steiner, M.; Sarkisov, L. A Process-Level Perspective of the Impact of Molecular Force Fields on the Computational Screening of MOFs for Carbon Capture. *Energy Environ. Sci.* **2023**, *16* (9), 3899–3918. https://doi.org/10.1039/D3EE00858D.

(84) Kadantsev, E. S.; Boyd, P. G.; Daff, T. D.; Woo, T. K. Fast and Accurate Electrostatics in Metal Organic Frameworks with a Robust Charge Equilibration Parameterization for High-Throughput Virtual Screening of Gas Adsorption. *J. Phys. Chem. Lett.* **2013**, *4* (18), 3056–3061. https://doi.org/10.1021/jz401479k.

(85) Luo, J.; Said, O. Ben; Xie, P.; Gibaldi, M.; Burner, J.; Pereira, C.; Woo, T. K. MEPO-ML: A Robust Graph Attention Network Model for Rapid Generation of Partial Atomic Charges in Metal-Organic Frameworks. *NPJ Comput. Mater.* **2024**, *10* (1), 224. https://doi.org/10.1038/s41524-024-01413-4.

(86) Campañá, C.; Mussard, B.; Woo, T. K. Electrostatic Potential Derived Atomic Charges for Periodic Systems Using a Modified Error Functional. *J. Chem. Theory Comput.* **2009**, *5* (10), 2866–2878. https://doi.org/10.1021/ct9003405.

(87) Manz, T. A.; Limas, N. G. Introducing DDEC6 Atomic Population Analysis: Part 1. Charge Partitioning Theory and Methodology. *RSC Adv.* **2016**, *6* (53), 47771–47801. https://doi.org/10.1039/C6RA04656H.

(88) McCready, C.; Sladekova, K.; Conroy, S.; Gomes, J. R. B.; Fletcher, A. J.; Jorge, M. Quantifying the Uncertainty of Force Field Selection on Adsorption Predictions in MOFs. *J. Chem. Theory Comput.* **2024**, *20* (11), 4869–4884. https://doi.org/10.1021/acs.jctc.4c00287.

(89) Yu, Z.; Anstine, D. M.; Boulfelfel, S. E.; Gu, C.; Colina, C. M.; Sholl, D. S. Incorporating Flexibility Effects into Metal–Organic Framework Adsorption Simulations Using Different Models. *ACS Appl. Mater. Interfaces* **2021**, *13* (51), 61305–61315. https://doi.org/10.1021/acsami.1c20583.

(90) Goeminne, R.; Van Speybroeck, V. Ab Initio Predictions of Adsorption in Flexible Metal–Organic Frameworks for Water Harvesting Applications. *J. Am. Chem. Soc.* **2025**, *147* (4), 3615–3630. https://doi.org/10.1021/jacs.4c15287.

(91) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865–3868. https://doi.org/10.1103/PhysRevLett.77.3865.

(92) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate *Ab Initio* Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15). https://doi.org/10.1063/1.3382344.

(93) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B* **1999**, *59* (3), 1758–1775. https://doi.org/10.1103/PhysRevB.59.1758.

(94) Blöchl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50* (24), 17953–17979. https://doi.org/10.1103/PhysRevB.50.17953.

(95) Berquist, E. Chargemol. https://github.com/berquist/chargemol.

(96) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: A Generic Force Field for Molecular Simulations. *J. Phys. Chem.* **1990**, *94* (26), 8897–8909. https://doi.org/10.1021/j100389a010.

(97) Boato, G.; Casanova, G. A Self-Consistent Set of Molecular Parameters for Neon, Argon, Krypton and Xenon. *Physica* **1961**, *27* (6), 571–589. https://doi.org/10.1016/0031-8914(61)90072-6.

(98) Rao, S. N.; Head, M. S.; Kulkarni, A.; LaLonde, J. M. Validation Studies of the Site-Directed Docking Program LibDock. *J. Chem. Inf. Model.* **2007**, *47* (6), 2159–2171. https://doi.org/10.1021/ci6004299.

(99) Takamizawa, S.; Nataka, E.; Akatsuka, T.; Miyake, R.; Kakizaki, Y.; Takeuchi, H.; Maruta, G.; Takeda, S. Crystal Transformation and Host Molecular Motions in CO2 Adsorption Process of a Metal Benzoate Pyrazine (M(II) = Rh, Cu). *J. Am. Chem. Soc.* **2010**, *132* (11), 3783–3792. https://doi.org/10.1021/ja9091598.

(100) Drwęska, J.; Formalik, F.; Roztocki, K.; Snurr, R. Q.; Barbour, L. J.; Janiak, A. M. Unveiling Temperature-Induced Structural Phase Transformations and CO2 Binding Sites in CALF-20. *Inorg. Chem.* **2024**, *63* (41), 19277–19286. https://doi.org/10.1021/acs.inorgchem.4c02952.

(101) Bette, S.; Sleptsova, A.; Lotsch, B. V.; Dinnebier, R. E.; Marx, S.; Loloei, M.; Adeleke, A. A.; Masoumifard, N.; Vaidhyanathan, R. CO2 and H2O Sorption Induced Bulk-Phase Changes of CALF-20 Captured Using In Situ Laboratory X-Ray Powder Diffraction. *J. Am. Chem. Soc.* **2025**, *147* (29), 25662–25671. https://doi.org/10.1021/jacs.5c06866.

(102) Samsonenko, D. G.; Kim, H.; Sun, Y.; Kim, G.; Lee, H.; Kim, K. Microporous Magnesium and Manganese Formates for Acetylene Storage and Separation. *Chem. Asian J.* **2007**, *2* (4), 484–488. https://doi.org/10.1002/asia.200600390.

(103) Miller, S. R.; Wright, P. A.; Devic, T.; Serre, C.; Férey, G.; Llewellyn, P. L.; Denoyel, R.; Gaberova, L.; Filinchuk, Y. Single Crystal X-Ray Diffraction Studies of Carbon Dioxide and Fuel-Related Gases Adsorbed on the Small Pore Scandium Terephthalate Metal Organic Framework, Sc2(O2CC6H4CO2)3. *Langmuir* **2009**, *25* (6), 3618–3626. https://doi.org/10.1021/la803788u.

(104) Qazvini, O. T.; Babarao, R.; Telfer, S. G. Selective Capture of Carbon Dioxide from Hydrocarbons Using a Metal-Organic Framework. *Nat. Commun.* **2021**, *12* (1), 197. https://doi.org/10.1038/s41467-020-20489-2.

(105) Main, R. M.; Ettlinger, R.; Tajnšek, T. K.; Brako-Amoafo, D. A.; Stanzione, M. G.; Duncan, M. J.; Ettlinger, P.; Lawrence, G. B.; Warren, M. R.; Heard, C. J.; Morris, R. E. In Situ Single-Crystal X-Ray Diffraction Studies of an Anomalous Nitric Oxide Adsorption in a Partially Activated Metal–Organic Framework. *J. Am. Chem. Soc.* **2025**, *147* (34), 31260–31269. https://doi.org/10.1021/jacs.5c10395.

(106) Banerjee, D.; Simon, C. M.; Plonka, A. M.; Motkuri, R. K.; Liu, J.; Chen, X.; Smit, B.; Parise, J. B.; Haranczyk, M.; Thallapally, P. K. Metal–Organic Framework with Optimally Selective Xenon Adsorption and Separation. *Nat. Commun.* **2016**, *7* (1), ncomms11831. https://doi.org/10.1038/ncomms11831.

(107) Takamizawa, S.; Nakata, E.; Saito, T.; Kojima, K. Structural Determination of Physisorbed Sites for CO2 and Ar Gases inside an Organometallic Framework. *CrystEngComm* **2003**, *5* (72), 411. https://doi.org/10.1039/b312553j.

(108) Takamizawa, S.; Nakata, E.; Yokoyama, H.; Mochizuki, K.; Mori, W. Carbon Dioxide Inclusion Phases of a Transformable 1D Coordination Polymer Host [Rh2(O2CPh)4(Pyz)]n. *Angew. Chem. Int. Ed.* **2003**, *42* (36), 4331–4334. https://doi.org/10.1002/anie.200351368.

(109) Takamizawa, S.; Nakata, E.; Saito, T. Structural Determination of Copper(II) Benzoate–Pyrazine Containing Carbon Dioxide Molecules. *Inorg. Chem. Commun.* **2004**, *7* (1), 1–3. https://doi.org/10.1016/j.inoche.2003.09.011.

(110) Takamizawa, S.; Kojima, K.; Akatsuka, T. Channel-Switching Crystal with Guest Stress Drive. *Inorg. Chem.* **2006**, *45* (12), 4580–4582. https://doi.org/10.1021/ic051952e.

(111) Couck, S.; Gobechiya, E.; Kirschhock, C. E. A.; Serra-Crespo, P.; Juan-Alcañiz, J.; Martinez Joaristi, A.; Stavitski, E.; Gascon, J.; Kapteijn, F.; Baron, G. V.; Denayer, J. F. M. Adsorption and Separation of Light Gases on an Amino-Functionalized Metal–Organic Framework: An Adsorption and In Situ XRD Study. *ChemSusChem* **2012**, *5* (4), 740–750. https://doi.org/10.1002/cssc.201100378.

(112) Matsuda, R.; Kitaura, R.; Kitagawa, S.; Kubota, Y.; Belosludov, R. V.; Kobayashi, T. C.; Sakamoto, H.; Chiba, T.; Takata, M.; Kawazoe, Y.; Mita, Y. Highly Controlled Acetylene Accommodation in a Metal–Organic Microporous Material. *Nature* **2005**, *436* (7048), 238–241. https://doi.org/10.1038/nature03852.

(113) Chen, M.; Chen, S.; Chen, W.; Lucier, B. E. G.; Zhang, Y.; Zheng, A.; Huang, Y. Analyzing Gas Adsorption in an Amide-Functionalized Metal Organic Framework: Are the Carbonyl or Amine Groups Responsible? *Chemistry of Materials* **2018**, *30* (11), 3613–3617. https://doi.org/10.1021/acs.chemmater.8b00681.

(114) McKinlay, A. C.; Xiao, B.; Wragg, D. S.; Wheatley, P. S.; Megson, I. L.; Morris, R. E. Exceptional Behavior over the Whole Adsorption−Storage−Delivery Cycle for NO in Porous Metal Organic Frameworks. *J. Am. Chem. Soc.* **2008**, *130* (31), 10440–10444. https://doi.org/10.1021/ja801997r.

(115) Maji, T. K.; Mostafa, G.; Matsuda, R.; Kitagawa, S. Guest-Induced Asymmetry in a Metal−Organic Porous Solid with Reversible Single-Crystal-to-Single-Crystal Structural Transformation. *J. Am. Chem. Soc.* **2005**, *127* (49), 17152–17153. https://doi.org/10.1021/ja0561439.

(116) Zhang, J.-P.; Chen, X.-M. Optimized Acetylene/Carbon Dioxide Sorption in a Dynamic Porous Crystal. *J. Am. Chem. Soc.* **2009**, *131* (15), 5516–5521. https://doi.org/10.1021/ja8089872.

(117) Liao, P.-Q.; Zhou, D.-D.; Zhu, A.-X.; Jiang, L.; Lin, R.-B.; Zhang, J.-P.; Chen, X.-M. Strong and Dynamic CO2 Sorption in a Flexible Porous Framework Possessing Guest Chelating Claws. *J. Am. Chem. Soc.* **2012**, *134* (42), 17380–17383. https://doi.org/10.1021/ja3073512.

4. Conclusions and Future Work

4.1 Conclusions

This thesis addressed two central challenges in adsorption modelling for MOFs. First, the field lacked a validated, fully standalone procedure for extracting binding-site locations from three-dimensional adsorption probability distributions generated by GCMC simulations. Second, the reliability of classical GCMC approximations for predicting adsorption environments had not been systematically assessed against experimentally determined binding sites. The work presented here resolves both issues. It establishes a robust algorithmic framework for converting probabilistic adsorption data into explicit structural information and demonstrates that classical GCMC retains significant predictive value even when used with commonly adopted approximations.

4.1.1 Guest Atom Localizer from Probabilities (GALP)

Chapter 2 introduced GALP, the Guest Atom Localizer from Probabilities, which is the central methodological contribution of this thesis. GALP provides a unified workflow for transforming GCMC-derived probability distributions into chemically meaningful binding-site coordinates through a sequence of smoothing, peak identification, spatial clustering, and RMSD-based molecular placement. The method was validated across 100 experimentally characterized MOFs and 9 representative guest conditions, including C_2H_2 (298 K, 1 bar), C_3H_6 (373 K, 1 bar), C_3H_8 (373 K, 1 bar), CH_4 (298 K, 1 bar), CH_4 (298 K, 65 bar), CO_2 (298 K, 1 bar), Kr (298 K, 1 bar), Xe (298 K, 1 bar), and N_2 (298 K, 0.75 bar). Across this diverse test set, GALP consistently identified the correct probability maxima and accurately placed the guest molecules within their corresponding adsorption environments. These results confirm that GALP reliably extracts both

the geometric positions and molecular orientations associated with physically relevant binding sites, thereby establishing a rigorous and scalable foundation for the structural interpretation of GCMC sampling.

4.1.2 Assessment of GCMC approximations in simulations

The work in Chapter 3 examined whether the approximations commonly used in classical GCMC simulations compromise the accuracy of predicted binding sites. Force-field choice, charge assignment method, rigid framework treatment, and equation-of-state choice were each evaluated by comparing GALP-extracted binding sites to experimentally resolved binding sites. The results show that, despite occasional discrepancies in isotherm predictions, classical GCMC consistently reproduces the correct binding motifs across a broad range of chemistries and interaction strengths. This finding underscores that global deviations in uptake do not necessarily translate into structural anomalies at the binding-site level. The analysis therefore reinforces the reliability of classical GCMC for understanding adsorption mechanisms and for predicting adsorption environments under realistic conditions.

4.1.3 High-throughput & Machine learning application

Although not explored in depth within this thesis, the methods developed here have already been applied in a high-throughput context by a coworker. GALP was integrated into a large-scale probability-distribution generation pipeline and used to produce training data for a graph neural network capable of predicting full adsorption probability distributions at a fraction of the computational cost, as described in the manuscript “Rapid prediction of adsorbate probability distributions in metal–organic frameworks using graph neural networks”[118]. This application

demonstrates that GALP is not only accurate and generalizable but also practical for large-scale automation and machine-learning-driven adsorption modelling.

4.2 Future Work

4.2.1 Extensions of the Binding Site Identification Framework

Future improvements to GALP should focus on expanding its utility rather than increasing complexity. A key direction is to develop binding-environment descriptors that translate the three-dimensional adsorption probability landscape into compact numerical representations suitable for machine-learning workflows. These descriptors would capture features such as the number, geometry, symmetry, and relative strength of binding sites, enabling supervised models to learn structure–probability relationships directly. Such representations would also enable rapid prediction of uptake or selectivity using models that incorporate binding environment information rather than relying solely on framework-level descriptors.

On the algorithmic side, only modest refinements are needed. Improving the robustness of the molecular-building step will allow GALP to handle larger polyatomic guests (e.g., benzene, methanol), including short alcohols and other small polar molecules, once reliable force fields for these adsorbates become available. Extending compatibility to alternative GCMC programs such as RASPA3 will broaden accessibility and facilitate integration with existing simulation pipelines. Beyond these extensions, the core workflow is already mature, and most future developments will come from its application in larger datasets and from the adoption of binding-site descriptors in data-driven adsorption modelling.

4.2.2 Improving the Predictive Accuracy of Atomistic Simulations

The validation performed in Chapter 3 could be extended as additional high-quality experimental adsorption structures become available, enabling a more rigorous and systematic assessment of the limits of current atomistic simulation approximations. On the simulation side, such data would enable targeted evaluation of how specific modelling choices, including rigid framework assumptions, force field parametrization, and sampling protocols, influence the predicted locations and populations of adsorption sites. Frameworks exhibiting backbone flexibility, multiple competing binding pockets, or strong loading-dependent guest-guest correlations would be particularly valuable for identifying when and why classical GCMC descriptions begin to deviate from experiment. Extending the analysis to mixed-gas systems would further test the ability of classical GCMC to capture competitive and cooperative adsorption effects, where site preference and occupancy depend sensitively on inter-guest interactions and composition.

Together, these extensions would not only broaden the scope of validation but also help close existing gaps between simulation and experiment by isolating the specific assumptions responsible for observed discrepancies. Rather than pursuing indiscriminate force field refinement, this approach would enable targeted improvements by distinguishing errors arising from interaction models, structural rigidity, or incomplete sampling. On the experimental side, advances in neutron diffraction and solid-state NMR methodology, including improved spatial resolution, reduced crystallographic disorder, and better control over adsorption conditions, would provide more reliable reference data under realistic operating conditions. Such improvements are essential for meaningful benchmarking, as uncertainty in experimentally determined binding site positions directly limits the ability to assess and refine simulation-based predictions.

4.3 References

(118) Burner, J.; Marchand, O.; Cicciarella, R.; Gibaldi, M.; Woo, T. K. Rapid Prediction of Adsorbate Probability Distributions in Metal-Organic Frameworks Using Graph Neural Networks. September 17, 2025. https://doi.org/10.26434/chemrxiv-2025-tn0rh.