Molecular modelling studies of protein@MOF and the post-translational modification in an archaeal enzyme

A Thesis

Submitted For the Degree of

DOCTOR OF PHILOSOPHY

in the Faculty of Science

by

Oishika Jash

CHEMISTRY AND PHYSICS OF MATERIALS UNIT
JAWAHARLAL NEHRU CENTRE FOR ADVANCED SCIENTIFIC RESEARCH
Bangalore – 560 064, India

NOVEMBER 2024

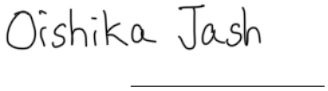
Dedicated to all my pets

♡***Tom, Kalu, Nonte, Fonte, Keltu and Tomtom***♡

DECLARATION

I hereby declare that the matter embodied in the thesis entitled "**Molecular modelling studies of protein@MOF and the post-translational modification in an archaeal enzyme**" is the result of investigations carried out by me at the Chemistry and Physics of Materials Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India under the supervision of Prof. S. Balasubramanian and that it has not been submitted elsewhere for the award of any degree or diploma.

In keeping with the general practice in reporting scientific observations, due acknowledgement has been made whenever the work described is based on the findings of other investigators. Any omission that might have occurred by oversight or error of judgement is regretted.



Oishika Jash

CERTIFICATE

I hereby certify that the matter embodied in this thesis entitled "**Molecular modelling studies of protein@MOF and the post-translational modification in an archaeal enzyme**" has been carried out by Ms. Oishika Jash at the Chemistry and Physics of Materials Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India under my supervision and that it has not been submitted elsewhere for the award of any degree or diploma.

Prof. S. Balasubramanian

(Research Supervisor)

Acknowledgements

The work done in this thesis was supported and encouraged by a group of people to whom I would like to convey my gratitude now.

There are no proper words to express my deepest respect for my research supervisor, Prof. Sundaram Balasubramanian. He has stimulated me to become an independent researcher and made me realize what a creative, hard-working scientist can accomplish. I am much obliged to him.

I am deeply indebted to all my collaborators Prof. Hemalatha Balaram (JNCASR, India), Prof. Padmanabhan Balaram (NCBS, India), Prof. Anand Srivastava (IISc, India) and Anusha Chandrashekarmath (JNCASR, India) for providing me with their expertise and intuition for my scientific problems. I would also like to express my heartfelt gratitude to Prof. Anjukandi Padmesh (IIT Palakkad, India), Nayana Edavan Chathoth (IIT Palakkad, India), Dr. Sudip Das (JNCASR, India), Dr. Abhishek Sharma (JNCASR, India), Dr. Nimish Dwarkanath (JNCASR, India) and Dr. Sudarshan Behera (JNCASR, India), for their valuable suggestions.

I would like to sincerely acknowledge Prof. Anand Srivastava, IISc, for the precious discussions.

I am highly grateful to all my comprehensive examiners, GSAC committee members and colloquium evaluation members Prof. Athinarayanan Sundaresan (JNCASR, India), Prof. Swapan Kumar Pati (JNCASR, India), Prof. Tapas Kumar Maji (JNCASR, India), Prof. Prabal Maiti (IISc, India), Prof. Govardhan Reddy (IISc, India), Prof. Anand Srivastava (IISc, India) and all CPMU, JNCASR faculties for their critical comments.

I would like to thank Prof. Thimmaiah Govindaraju, Prof. Sridhar Rajaram and all my course instructors from NCU, JNCASR, India.

I would like to sincerely thank CCMS at JNCASR, PARAM Yukti supercomputing facility under National Supercomputing Mission (NSM), India, for excellent computational facilities. Also, I would like to thank Suresh Jagannathan, Salman Khan, and all

other NSM technicians, and Complab technicians, JNCASR, for their generous help on different occasions.

I am indeed grateful to CSIR, HRDG, India, for providing me with a research fellowship. I am deeply thankful to JNCASR and NSM for providing me funding during my one year of extended PhD tenure.

I would like to extend my sincere thanks to the open-source communities, specially GROMACS, CP2K, VMD, PLUMED, MDANALYSIS, MOLYWOOD, PyMOL, and PACKMOL, for providing useful software and documentation.

I owe my thanks to all the academic and non-academic staff, particularly security, library, academic office, CPMU office, Tavaragera Basavraju, mess staff, room-cleaning staff and gardeners of JNCASR for making my stay at JNCASR pleasurable.

I am indeed thankful to all the present and past lab members for creating a vibrant learning environment in the lab.

I would like to sincerely thank all my friends and colleagues from Santiniketan, JNCASR, IISc and TIFR, India.

I would like to extend my sincere gratitude to my teachers from my kindergarten (Mrinalini Ananda Pathsala, Santiniketan), my school (Patha-Bhavana, Santiniketan) and my college (Department of Chemistry, Santiniketan, particularly Prof. Bidhan Chandra Bag, Prof. Md Motin Seikh and Prof. Pranab Sarkar) who helped me to bloom.

A journey which "far from being shepherded" started "like a frightened flock hither and thither, helter-skelter" would not have arrived at the harbour without the top-notch support from the hull. Yes, that is my family. My sincerest regards to Maa (my mother, Rita Jash), Baba (my father, Tapan Kumar Jash), Titu (my sister, Ishika Jash), my grandparents (especially Late Kamala Jash and Late Rabindranath Roy Chowdhury), and all my family members. And my pets? They own me!

Preface

The thesis presents results obtained from molecular modelling studies on two types of biomolecular systems. The first is of proteins immobilized in the pores of Metal-Organic Frameworks (MOF). The second system is the post-translational modification in an archaeal enzyme. While the former is important in the biotechnology and pharmaceutical industries, the latter is academically motivated and involves the detailed study of the mechanism of an autocatalyzed reaction.

Chapter 1 provides a generalized description of modelling techniques at the molecular level, along with a short introduction to proteins, MOF, protein@MOF, and post-translational modifications (PTM).

Three work chapters follow the introduction. **Chapter 2** and **Chapter 3** deal with protein@MOF systems. More specifically, **Chapter 2** deals with proteins in channel MOFs, and **Chapter 3** deals with a protein inside a caged MOF. **Chapter 4** studies the reaction mechanism in MjGATase.

In **Chapter 2**, results obtained from equilibrium molecular dynamics simulations of myoglobin and the Green Fluorescent Protein (GFP) in IRMOFs are presented. The current work is motivated by an experimental work that reported the formation of inclusion complexes of these biomolecules with the isoreticular MOF series, i.e., within the channels of IRMOF-74-VII-oeg and IRMOF-74-IX, respectively, where -oeg is the triethylene glocol monomethyl ether group. Employing extensive all-atom equilibrium molecular dynamics simulations, we observe that both these inclusions are mainly governed by van der Waals interactions at the protein-MOF interface. The confinement effect on myoglobin was larger than that of GFP due to the relatively smaller size difference between the former and its MOF host. The primary signature of the confinement was observed in the root mean squared fluctuations of the protein sidechains. Although experiments could not succeed in the inclusion of myoglobin in IRMOF-74-VII-hex (where -hex is the hexyl group), our simulations suggest that it could be easily accommodated in the same, suggesting the possibility of kinetic contributions in the experimental observation. Overall, the tertiary structures and hydration of the surfaces of the proteins were well maintained inside the MOF channels for both proteins.

Chapter 3 examines the mechanism of translocation of proteins through the cavities of the MOF using a model protein-MOF system. The protein chosen is the chicken villin headpiece subdomain, HP35, and the MOF is MIL-101(Cr), as the former's diameter is larger than the size of the window between the cages of the MOF. Equilibrium molecular dynamics simulations demonstrate that the protein is located farther from the center of the cavity and closer to the MOF surface. Molecular interactions with the MOF partially unfolds helix-1 at its N-terminus. The translocation of HP35 through the narrow, hexagonal windows of the mesopores of MIL-101 (Cr), a process that is vital to biomolecular inclusion, was studied using non-equilibrium molecular dynamics simulations. Steered molecular dynamics (SMD), followed by umbrella sampling simulations (US), show that the translocation process across the spherical cavity can be divided into three zones, resulting from both the geometry of the confinement and MOF surface chemistry; in one of them, the protein maintains nearly its native conformational ensemble which encompasses the equilibrium position of the protein. The free energy barrier for the unfolded protein at the cage window, relative to its equilibrium state within the cavity, is estimated to be 16 kcal/mol.

Chapter 4 reports results on the mechanism of an autocatalyzed reaction, the post-translational modification of an enzyme, glutamine amidotransferase (MjGATase), present in a hyperthermophilic archaea. It undergoes the spontaneous formation of succinimide at residue-109 (within an -END- sequence), due to which the enzyme is stable and functional even up to 100°C. Experimental collaborators have discovered through mutation studies of nearby residues that residues D110, Y158, and K151 might have roles in succinimide formation. QM/MM simulations with non-tempered Metadynamics are employed to address how succinimide is formed. The QM region consists of 4 residues (N109, D110, Y158, and K151), including the reactant residue (N109). The reaction proceeds through deprotonation followed by cyclization. Metadynamics simulations yield a free energy barrier for the deprotonation step to be 3.4 kcal/mol; the subsequent cyclization step is likely to be barrierless. However, while intermittent hydrogen bonding between Y158 and the side chains of D110 and N109 was observed, the exact role of Y158 in the product formation was not discernible.

The thesis concludes with **Chapter 5**, which summarizes the work presented and provides a future outlook.

Contents

List of Figures

List of Tables

—the energy surface is one of the fundamental determinants of any reaction, whether a small molecule reaction or protein folding.

Martin Karplus [1]

1 Introduction

This chapter describes the theoretical frameworks and biomolecular systems on which my thesis is based. Specific details used in the projects can be found in Chapters 2 - 4.

The primary method of my thesis is computer simulations. Following is a small narration for that.

1.1 Molecular modelling

Matter is discrete and made up of atoms. Atoms consist of electrons and nuclei. So, the description of any system at the molecular level involves how the nuclei and electrons interact with each other as expressed by the energy operator of an isolated system in Equation 1.1.

$$\mathcal{H} = \sum_i \frac{P_i^2}{2M_i} + \sum_n \frac{P_n^2}{2m} + \frac{1}{2}\sum_{ij} \frac{Z_i Z_j e^2}{|R_i - R_j|} + \frac{1}{2}\sum_{nn'} \frac{e^2}{|r_n - r_n'|} - \sum_{in} \frac{Z_i e^2}{|R_i - r_n|} \tag{1.1}$$

Since nuclei are much heavier than electrons, electronic and nuclear motion can be separated, which came to be known as the Born-Oppenheimer approximation from a 1927 paper by Max Born and Robert Oppenheimer [2, 3] [Equations 1.2, 1.3, and 1.4].

$$\Psi(R_i, r_n) = \Xi(R_i)\, \Phi(r_n; R_i) \tag{1.2}$$

$$H_{el}\Phi(r_n; R_i) = V(R_i)\, \Phi(r_n; R_i) \tag{1.3}$$

$$H = \sum_i \frac{P_i^2}{2M_i} + H_{el} \tag{1.4}$$

Thus, a reduced description of the system becomes how the nuclei are moving in the presence of other nuclei and electrons [Equations 1.5 and 1.6].

$$H\Psi = E\Psi \tag{1.5}$$

$$\left[\sum_i \frac{P_i^2}{2M_i} + V(R_i)\right] \Xi(R_i) = E\,\Xi(R_i) \tag{1.6}$$

In a classical framework, i.e., when the particle-particle-separation is much larger than the de Broglie's wavelength [Equation 1.7a, 1.7b], Equation 1.6 can replaced with Newton's equation of motion [Equation 1.8, 1.9] [4].

$$\Lambda = \sqrt{\frac{2\pi\hbar^2}{Mk_bT}} \tag{1.7a}$$

$$\Lambda << a \tag{1.7b}$$

$$\vec{F_i} = m_i\vec{a_i} \tag{1.8}$$

$$\vec{F_i} = -\nabla_{r_i} V(\vec{r_1}, \vec{r_2}, \vec{r_3}, \ldots, \vec{r_N}) \tag{1.9}$$

Newton's equation depicts that a particle changes its velocities with a force being applied and through the formulation, we can exactly predict the motion at a future time if we know the position and velocities of a particle (or a set of particles) at a present time. Atoms of a molecule experience forces arising from intermolecular interactions. Given an initial set of atom coordinates and velocities, these forces yield a new set of the same. With this new set of positions, a new force is experienced and hence, a new set of velocities generated and again another set of new positions and thus a ceaseless motion [Equations 1.10a, 1.10b] [5].

$$\frac{d\vec{r_i}}{dt} = \frac{\partial H}{\partial \vec{p_i}} = \frac{\vec{p_i}}{m} \tag{1.10a}$$

$$\frac{d\vec{p_i}}{dt} = -\frac{\partial H}{\partial \vec{r_i}} = -\sum_{j(\neq i)} r_{ij}U'_{ij} \tag{1.10b}$$

We approximate this mutual interaction among atoms with an empirical potential energy function (also known as Force field, Figure 1.1). In this thesis, we have used three families of them, namely, CHARMM (Chemistry at HARvard Macromolecular Mechanics) [6], AMBER (Assisted Model Building with Energy Refinement) [7] and UFF (Universal

Force Field) [8] [Equations 1.11, 1.12, and 1.13].

$$U_{CHARMM} = \sum_{bonds} K_b \left(b - b_0\right)^2 + \sum_{UB} K_{UB} \left(s - s_0\right)^2 + \sum_{angles} K_\theta \left(\theta - \theta_0\right)^2 + \sum_{dihedrals} K_\lambda \left(1 + cos(n\chi - \delta)\right) + \sum_{impropers} K_{imp} \left(\varphi - \varphi_0\right)^2 + \sum_{nonbonded} \epsilon \left[\left(\frac{R_{minij}}{r_{ij}}\right)^{12} - \left(\frac{R_{minij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\epsilon_i r_{ij}} \quad (1.11)$$

$$U_{AMBER} = \sum_{bonds} K_r \left(r - r_{eq}\right)^2 + \sum_{angles} K_\theta \left(\theta - \theta_{eq}\right)^2 + \sum_{dihedrals} \frac{V_n}{2} \left[1 + cos\left(n\phi - \gamma\right)\right] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right] \quad (1.12)$$

$$U_{UFF} = \sum_{bonds} \frac{1}{2} K_{ij} \left(r - r_{ij}\right)^2 + \sum_{angles} \frac{1}{2} K_{ijk} \left(\cos\theta - \cos\theta_0\right)^2 + \sum_{dihedrals} \frac{1}{2} K_\phi \left[1 + cos\left(n\phi - \gamma\right)\right] + \sum_{impropers} \frac{1}{2} K_{imp} \left(\varphi - \varphi_0\right)^2 + D_{ij} \left[-2\left[\frac{x_{ij}}{x}\right]^6 + \left[\frac{x_{ij}}{x}\right]^{12}\right] + \frac{Q_i Q_j}{\epsilon R_{ij}} \quad (1.13)$$

Water is modelled as a rigid body, and so no internal degrees of freedom involving

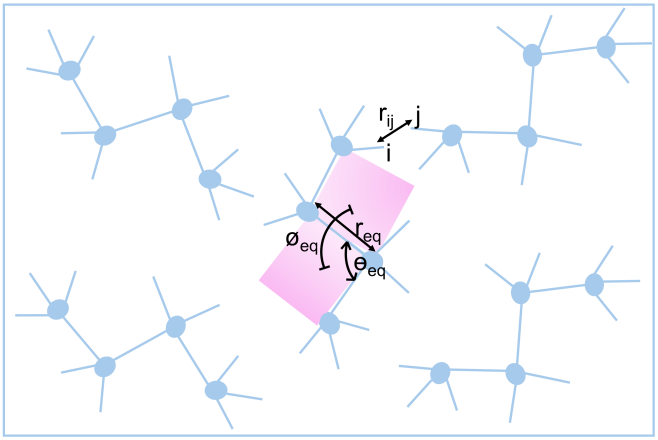


Figure 1.1: The interaction terms in a typical force field. Bonded interactions provide energy contributions from stretching, bending, and rotations (torsions). Nonbonded interactions describe interactions between two atoms separated by more than three bonds and provide energy contributions from short-range dispersion (Lennard-Jones potential) and long-range Coulomb interactions.

bending and stretching are considered. Only non-bonded interactions are included in the force field using the 3-site model (TIP3P [9] and CHARMM TIP3P [6]) and SPC/E [10]. Force field parameters for ions are chosen to be compatible with the water model.

This technique is known as Molecular Dynamics (MD) simulation. As MD is a statistical mechanical method, the trajectory of the particles in phase space (all coordinates and all momenta space) follow a probability distribution function. Since Newton's equation of motion conserves the total energy of the system, this distribution function looks like:

$$\delta(H(\Gamma) - E) \tag{1.14}$$

where Γ denotes all the phase space variables. To simulate the system at constant temperature or pressure, the degrees of freedom have to be coupled to a thermostat or a barostat. In this thesis, we have used the Bussi-Donadio-Parrinello velocity rescaling thermostat [11] [Equations 1.15a, 1.15b]

$$dK = (\bar{K} - K)\frac{dt}{\tau} + 2\sqrt{\frac{K\bar{K}}{N_f}}\frac{dW}{\sqrt{\tau}} \tag{1.15a}$$

$$\bar{P}(K_t)\, dK_t \propto K_t^{\frac{N_f}{2}-1}\, e^{-\beta K_t} dK_t \tag{1.15b}$$

and, Parrinello-Rahman barostat [12] [Equations 1.16a...1.16d, and 1.17a...1.17c].

$$\Omega = \|\mathbf{h}\| = \mathbf{a}.\,(\mathbf{b}\mathbf{x}\mathbf{c}) \tag{1.16a}$$

$$\mathbf{r}_i = \mathbf{h}\mathbf{s}_i = \xi_i\mathbf{a} + \eta_i\mathbf{b} + \zeta_i\mathbf{c} \tag{1.16b}$$

$$\mathbf{G} = \mathbf{h}^{'}\mathbf{h} \tag{1.16c}$$

$$\boldsymbol{\sigma} \equiv \Omega\mathbf{h}^{'-1} \tag{1.16d}$$

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{N} m_i\dot{\mathbf{s}}_i^{'}\mathbf{G}\dot{\mathbf{s}}_i - \sum_{i=1}^{N}\sum_{j>i}^{N}\phi(r_{ij}) + \frac{1}{2}\boldsymbol{W}\,Tr\,\dot{\mathbf{h}}^{'}\dot{\mathbf{h}} - p\Omega \tag{1.17a}$$

$$\mathbf{W}\ddot{\mathbf{h}} = (\pi - p)\,\boldsymbol{\sigma} \tag{1.17b}$$

$$\mathcal{H} = \sum_i \frac{1}{2}m_i\mathbf{v}_i^2 + \sum_{i=1}\sum_{j>i}\phi(r_{ij}) + \frac{1}{2}\boldsymbol{W}\,Tr\,\dot{\mathbf{h}}^{'}\dot{\mathbf{h}} + p\Omega \tag{1.17c}$$

For describing the system, we can have an all-atom description (as the name suggests, considering all the degrees of freedom of all atoms, explicitly) or a bit-coarsened description by lumping together non-polar hydrogens to their covalently attached atom (Carbon atom) using united-atom approach. In this thesis, we have mostly used all-atom modelling.

The challenge for seeing the real motion of particles within computational time is to visit all the possible positions accessible to the system under the thermodynamic condition during simulation (also known as ergodic in simulation timescale). The difficulty comes from two sources: one, any degree of freedom is slow, and two, there are energy barriers in between (Figure 1.2).

To circumvent these, the basic formalism [13] goes to modify the interaction potential energy as:

$$\mathcal{H}(\Gamma, \lambda) = \sum_{i=1}^{N}\frac{\mathbf{p}_i^2}{2m_i} + U(\mathbf{x}_1, \ldots, \mathbf{x}_N; \lambda) \tag{1.18}$$

where λ is the parameter that does the job. This formalism also provides us with an estimate of the free energy changes during any process. In this thesis, we have used three enhanced sampling techniques: Steered Molecular Dynamics (SMD), Umbrella Sampling (US) and Metadynamics. In SMD, we change the potential energy by doing work on the system through:

$$W^F = W^F[\Gamma_t] = \int_0^\tau dt\dot{\lambda}_t^F\frac{\partial U}{\partial\lambda}(\Gamma_t; \lambda_t^F) \tag{1.19}$$

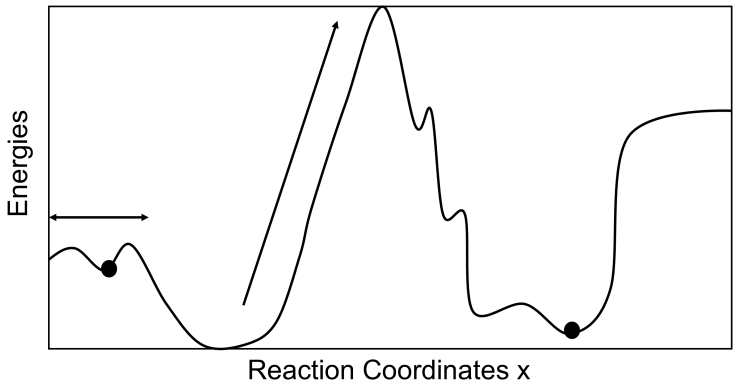


Figure 1.2: The necessity for enhancing the sampling is shown in a representative energy landscape.

In US [14], a harmonic potential is used as:

$$E^b(r) = E^u(r) + w_i(\xi) \tag{1.20a}$$

$$w_i(\xi) = \frac{K}{2}(\xi - \xi_i^{ref})^2 \tag{1.20b}$$

In Metadynamics [15], a repulsive Gaussian potential is used as:

$$V_G(S,t) = \int_0^t dt^{'} w\, e^{-\sum_{i=1}^{d} \frac{(S_i(R) - S_i(R(t^{'})))^2}{2\sigma_i^2}} \tag{1.21a}$$

$$w = \frac{W}{\tau_G} \tag{1.21b}$$

From these biased simulations, we can get back the actual probability distributions (also known as unbiasing). In the case of US, the umbrella potential, used piecewise at different parts (windows) of the path, gives us back the unbiased distribution as:

$$P_i^u(\xi) = P_i^b(\xi) e^{\beta w_i(\xi)} \left\langle e^{-\beta w_i(\xi)} \right\rangle \tag{1.22}$$

and, then, the distributions across all the umbrella windows are stitched together, minimizing the statistical error at every window, using the Weighted Histogram Analysis Method

(WHAM) as:

$$P^u(\xi) = \sum_{i}^{windows} p_i(\xi) P_i^u(\xi) \tag{1.23a}$$

$$\frac{\partial \sigma^2(P^u)}{\partial p_i} = 0 \tag{1.23b}$$

$$\sum p_i = 1 \tag{1.23c}$$

In the case of Metadynamics (non-tempered), the true landscape can be obtained by the negative value of the added Gaussians as:

$$V_G(S, t \to \infty) = -F(S) + C \tag{1.24}$$

Details on the philosophies of these formalisms can be found in [16–21].

For studying chemical reactions, changes in electron densities are accounted explicitly. Since solving the Schrödinger's equation is computationally costly, the common practice is to treat a part of the molecular mechanical system using quantum mechanics (QM), also known as QM/MM simulation (Figure 1.3).

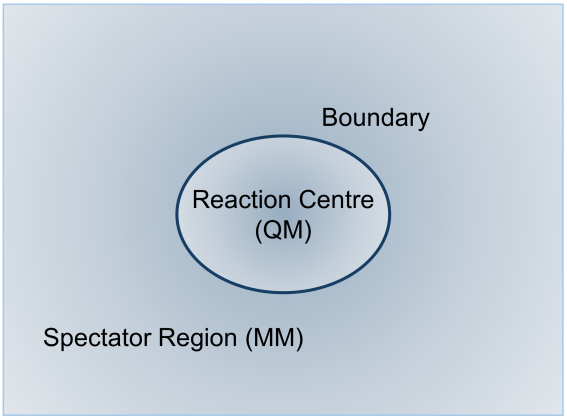


Figure 1.3: Partitioning a complete system into QM region and MM region for QM/MM simulations.

There are two flavours of QM/MM - additive and subtractive. Under the additive scheme, the total energy of the system is additively obtained from three contributions: i) the quantum region, ii) the MM region and iii) the interaction between quantum and MM regions, as [22]:

$$V_{QM/MM} = V_{QM}(QM) + V_{MM}(MM) + V_{QM-MM}(QM + MM) \tag{1.25}$$

The interaction between QM and MM systems can be treated again in three ways - a) at the MM level (1.26a), also known as mechanical embedding [23], b) allowing the influence of MM region electrostatically to the QM region (1.26b), also known as electrostatic embedding, and c) allowing the influence of QM and MM regions electrostatically on each other, also known as polarization embedding.

$$V_{QM-MM}(QM+MM) = V^{el}_{QM-MM} + V^{vdW}_{QM-MM} + V^{b}_{QM-MM} \tag{1.26a}$$

$$h_i^{QM-MM} = h_i^{QM} - \sum_J^M \frac{e^2 Q_J}{4\pi\epsilon_0 \left|\mathbf{r}_i - \mathbf{R}_J\right|} \tag{1.26b}$$

While dividing between QM and MM regions, when we cut across covalent bonds, one of the usual practices is to introduce monovalent link atom (mostly Hydrogen atom) along the bond vector, where the link atoms are included in the QM calculation and the force acting on them is distributed between QM and MM atoms according to the lever rule. Some good reviews on QM/MM simulations can be found in [24–27].

The symbols used in the equations are standard ones. Corresponding papers will give more details.

1.2 Molecular modelling using Computers

We track the motion of the particles in computers with a discrete finite timestep. In this thesis, the machinery used for this purpose is the Leapfrog algorithm [Equations 1.27a, 1.27b]
(https://www.feynmanlectures.caltech.edu/I_09.html).

$$x(t+\epsilon) = x(t) + \epsilon\, v\left(t + \frac{\epsilon}{2}\right) \tag{1.27a}$$

$$v\left(t + \frac{\epsilon}{2}\right) = v\left(t - \frac{\epsilon}{2}\right) + \epsilon\, a(t) \tag{1.27b}$$

Because of this numerical implementation, which comes with finite precision arithmetic, computer generated trajectory does not follow the true trajectory. The divergence is exponential in time if there is a tiny difference in momenta (with exact particle position), also known as Lyapunov instability:

$$\Delta \boldsymbol{r}(t) \sim (\delta \boldsymbol{p})\, e^{\lambda t} \tag{1.28}$$

Periodic boundary conditions (PBC) are usually applied to nullify the surface effects during computer simulations. The "central" simulation box is repeated in space to make an

infinite lattice. If a particle moves in the central box, its periodic images in the neighbouring boxes move as well in exactly the same direction, and if it leaves the central box, one of its images' enters the central box from the opposite direction, eventually making the "central" box boundary-wall-less and surface-less. Thus, the image boxes do not have any degrees of freedom, and the coordinates of the particles from the central box are stored. While deploying PBC, we need to consider the range of intermolecular interaction. If the range becomes longer than the box length, a particle will interact with its own images, resulting in non-realistic motions. To avoid this, the usual practice is to use a cut-off, r_C as the interaction range. With the help of this cut-off, the particle interacts with only its nearmost neighbours (also known as minimum image convention), and accordingly, the condition which should always be followed is:

$$r_C \leq \frac{L}{2} \tag{1.29}$$

where L is the box length of a cubic box (Figure 1.4).

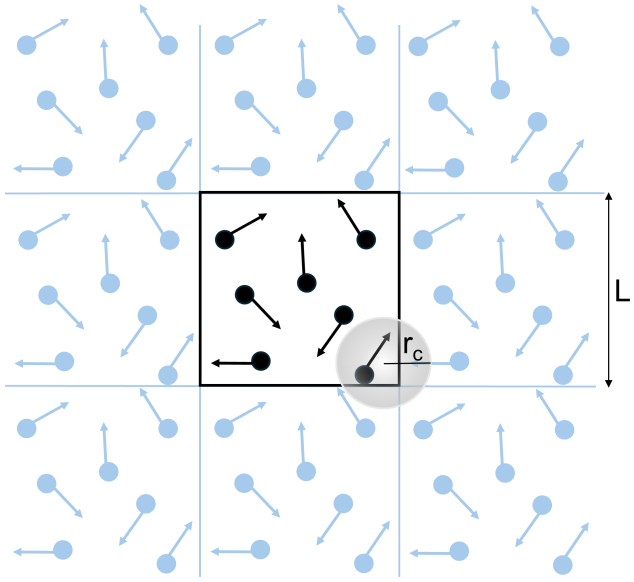


Figure 1.4: A two-dimensional periodic system. The cubic primitive cell (Black colour box) with length L is periodically repeated in two dimensions. The radius r_C of the transparent Grey sphere denotes the cut-off distance according to the minimum image convention.

To determine the neighbours, a further cut-off (larger sphere) is used (proposed by Loup Verlet) to maintain a neighbour list which gets updated at a definite interval. The error introduced by the potential cut-off is corrected through tail correction to the potential for short-range interactions. For long-range interactions as tail correction diverges, the

lattice method (like Ewald sum) is used. Computationally efficient implementation of this method is done through mesh-based charge assignment using fast Fourier transform algorithms (like particle-mesh Ewald or PME [28]). The system should be electroneutral to use Ewald summation, which is achieved by adding counter ions. For speeding up simulations, a common practice is to use constraints (for example, LINCS algorithm [29, 30]) on bond lengths (mainly for bonds containing Hydrogen atoms).

Two good reads on computer simulations by Michael P. Allen and Dominic J. Tildesley [31] and by Daan Frenkel and Berend Smit [32] provide more details.

1.3 Molecular modelling in this Thesis

In this thesis, we have carried out molecular modelling studies on two types of biomolecular systems. The first comprises proteins immobilized in the Metal-Organic Frameworks (MOF) pores. The second system is the post-translational modification in an archaeal enzyme. The following discussion introduces both systems.

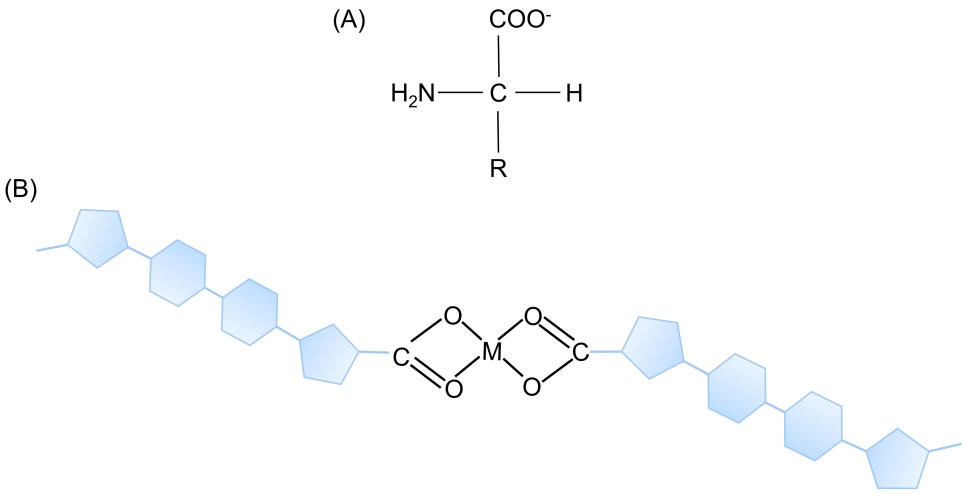


Figure 1.5: (A) Monomer unit (an amino acid, with R as the side chain) of covalent biopolymer, protein. (B) A representative MOF fragment. M stands for metal atoms. Shaded hexagons and pentagons, including the carboxylate group, are part of organic linkers.

1.3.1 Protein structure and organization

Proteins are biological polymers ("Molecular Elephants" as called by John B. Fenn) made up of amino acids. An amino acid contains a carboxyl group, an amino group,

a hydrogen atom and a side-chain. Different amino acids differ in water solubility depending on the chemical nature of side chains. There are different ways to measure the hydrophobicity of amino acids, known as hydrophobicity scales [33]. The *p*Ka values of the side-chain depend on the local environment inside the protein structure and around it. Among the 22 amino acids encoded by genes, glycine is the achiral one. In Nature, proteins contain majorly L-amino acids. But the presence of D-amino acids has also been observed serving specific purposes, and hence, there are enzymes for their digestion as well [34, 35]. Amino acids are represented by one-letter or three-letter codes.

Martin Karplus has written, "Changes in perception are an essential element in the advancement of science" [1]. This is true for the history of protein. From the earlier thinking as a distinct class of molecules towards getting its chemical constituents [36–38], mainly it was Hermann Staudinger's macromolecule hypothesis [39] and Theodor Svedberg's centrifugation experiment on heamoglobin [40] brought proteins into existence. Later, molecular dynamics simulations replaced the static view of proteins with an ever-ceasing dynamical motion. Then comes the protein folding problem. Though Anfinsen [41] and co-workers established that proteins fold reversibly, Cyrus Levinthal [42] posed a "paradox" regarding the folding of the proteins to reach the thermodynamically minimum energy structure in a very fast time scale. In this enterprise [43], a great number of simulations have been done using either phenomenological models (nucleation-growth mechanism [44], diffusion-collision model [45], framework model [46], jigsaw-puzzle model [47]), or with energy landscape models ('golf-course' [48], 'funnel landscape' [49], 'Moat Landscape' [50], 'Champagne Glass Landscape' [51]). Through these studies, it emerged that protein folding problem might not be a NP hard problem. Standing in 2024, it can be said that, at least, we have partly solved the 'first problem' of the 'protein folding problem' i.e. prediction of the three-dimensional structure of a protein with the help of AlphaFold [52]. Partly because AlphaFold2 fails to predict the conformations of IDPs [53–55] and fold-switching proteins [56].

The structure of a protein can be divided hierarchically into primary, secondary, tertiary and quaternary structures. A peptide bond is formed through the condensation reaction between two amino acids. Linear chains of these amino acids tie together through peptide bonds, forming the primary structure of a protein. This amide bond (C-N) acquires a double bond character because of the mesomeric effect and gives atoms around the bond a planar structure, making a dihedral angle (ω) around 180°. Rotation around two single bonds namely, N-C_α, and C_α-C generate a set of dihedral angles known as ϕ and ψ, respectively. Strictly governed by steric hindrance, some sets of ϕ and ψ angles give rise to a "stable" protein structure. G.N. Ramachandran and V. Sasisekharan discovered a set of values for these angles, and the 2D plot of these two angles is known as Ramachandran plot [57]. The primary chain of amino acids connects with itself by hydrogen bonding. The resultant structure is known as the secondary

structure of a protein. Different secondary structural motifs are found in Nature - for example, $alpha$-helix, $beta$-strand, Hairpin Turns and coils. Further, the secondary motifs come together using different intermolecular interactions - i.e. hydrophobic, van der Waals, hydrogen bonding, salt bridges, disulphide bonds, and coordination bonding around cofactors (like heme and FAD) to form the tertiary structure of a protein. Different tertiary structures can glue together to form quarternary structures through hydrophobic and ionic interactions; there, the tertiary structural motifs are called subunits.

Proteins have been classified into class, fold, superfamily and family depending on the similarity they have (a set of proteins without any stable 3D structure, also known as intrinsically disordered proteins (IDP) play many crucial roles from transcriptional regulation to cell-signalling [58]). Some of the well-known protein databases are SCOP (http://scop2.mrc-lmb.cam.ac.uk/), CATH (http://www.cathdb.info), FSSP (http://ekhidna.biocenter.helsinki.fi/dali/), DisProt (https://disprot.org/), RCSB PDB (https://www.rcsb.org/), ExplorEnz (https://www.enzyme-database.org/) and KEGG (https://www.genome.jp/kegg/kegg1a.html). Proteins carry out important functions such as gene regulation, signalling, and metabolism as enzymes, hormones, and antibodies. Textbooks on proteins provide more details on structure and functions of them [59, 60].

1.3.2 MOF structure and organization

Metal-Organic Frameworks (MOF) are composed of two or more secondary building units (SBUs): metal-containing units (single metal atoms or rods and layers like infinite groups) and polytopic organic linkers that might contain metal atoms (for instance, porphyrin). They are crystalline. Crystal nets are simple graphs. The difference between the abstract graph and crystalline MOF is that the latter has an embedding; i.e. the vertices possess coordinates and edges have length [61]. There exist both channel MOFs containing 1-dimensional channels and cage MOFs possessing mesoporous cavities as well as having architectures like Russian dolls, i.e. nested interconnected cages [62, 63]. Following is an abridged note on the topology of MOF [61, 64, 65]. Each MOF contains the topology of a specific periodic net. The tilings covering a 2-D space are polygonal tiles, and a 3-D space (Euclidean) are generalized polyhedra or cages. The tilings are edge-to-edge in 2-D and face-to-face in 3-D. The 1-skeleton, i.e., the set of vertices and edges of the tilings, is called a net. For a given net, there could be more than one and even an infinite number of possible tilings. However, there is a unique special tiling, also known as natural tiling. For 3-periodic nets, natural tiling has to follow certain conditions - 1) tilings should be proper, i.e., tilings should follow the same symmetry as the net 2) tiles cannot have non-face strong rings the same size as or smaller than the smallest face and 3) no tile can have one face bigger than the remaining faces. Conditions 2 and

3 are additional. For most high-symmetry structures, the first condition is enough. A characterizing property for tiling is transitivity. If there are v kinds of vertices, e kinds of edges, f kinds of faces and t kinds of tile, then transitivity is veft. If there is one kind of vertex (or edge or face or tile), the structure is vertex- (or edge-, or face- etc) transitive. Here, one "kind" of vertex means that symmetry operations relate the remaining vertices to this vertex. Sometimes, the net of a crystalline material can be expressed in terms of dual tilings. A dual tiling can be understood by considering a polyhedron obtained by putting a vertex in the middle of each face and joining the new vertices as edges in adjacent faces. The dual of the polyhedron with transitivity pqr has transitivity rqp. Then, the 3-D tiling can be made of dual pairs or self-dual tilings. The octahedron and cube are mutual dual pairs, whereas the tetrahedron is a self-dual polyhedron. Another term known is the coordination figure of a vertex. This figure is formed by the neighbours of a vertex (more specifically by the convex hull of those neighbours) and represents a polygon or polyhedron. In the same line of discussion, another term is augmented structures. These structures are conceptualized by replacing the vertices of the original polyhedra with a coordination figure. The net of augmented polyhedra is made of polygons linked by edges. These concepts differentiate between the abstract graph and embedding (or realization). Nets have an automorphism group. The automorphism group denotes the group of permutations of vertices that leave the connectivity pattern of the net unchanged. This is similar to a crystallographic space group. This space group is the maximum possible symmetry of an embedding (realization) of the net (also known as symmetry of the net). These embeddings of nets have been made available in an online database by Omar Yaghi's group. The database is the Reticular Chemistry Structure Resource (RCSR,http://rcsr.net/).

1.3.3 Protein@MOF

Importance:

Proteins, being the best catalysts, play an important role in various industries (chemical, pharmaceutical, and food). A major concern for the use of enzymes in these applications is the exposure to non-native conditions. Some examples of non-native conditions are poor thermal stability, insolubility in organic solvents (such as THF), poor long-term stability, non-workability beyond optimum pH range conditions, exposure to proteolytic agents (proteases like trypsin) and chaotropic agents (urea) and intolerance for mechanical stresses. Further, product separation and purification create problems [66]. Cryopreservation might be a solution in some cases. But it has two major disadvantages - a) damage to the bio-entity from the repeated freeze-thaw cycle and b) costly cold-chain logistics [67]. On the other hand, many species in Nature survive in extreme conditions by forming shells around their bodies in response to environmental perturbations. For example, endospore coat formation of *Bacillus subtilis*, which is made of about 70 proteins and is nurtured by the mother cell [68], cyst wall formation of *Maryna umbrellata* as a

dormant stage [69, 70], and tun formation of Tardigrades during anhydrobiosis [71, 72]. Taking inspiration from them, many materials have been designed to provide a "shell" to enzymes for their use in the biotechnological industry [73]. For example, bulk materials such as Eupergit® C, Amberlite XAD-7 (acrylic resin) are used to immobilize enzymes by covalent attachment or adsorption, respectively [74] which has made some greatly used enzymes commercially available. Magnetic nanoparticles (Fe_3O_4) and gold-coated iron nanoparticles have been used for the immobilization of cholesterol oxidase and glucose oxidase, among many others [75–77]. Polyethylene glycols (PEG) based hydrogels have been used for lysozyme immobilization [78]. Modifications of graphene sheet with hydroxyl, epoxy and carboxyl groups make them useful for protein immobilization using surface chemistry of the protein (for example, amino groups from proteins' side chains) [79]. Carbon nanotubes provide an immobilization platform either through direct hydrophobic interaction and π-π interactions or through surfactants and polymer coating [80]. Polymers attach to proteins' surfaces through free NH2- or SH- groups [81]. DNA origami-based protein immobilization utilizes different interactions; for example, biotin-streptavidin [82]. Porous solids (like Microporous zeolites networks grow around proteins where Histidines play an important role in entrapping the protein [83], mesoporous silica immobilize enzymes through covalent and non-covalent interactions [84], Metal-Organic Frameworks, Colvalent-Organic Frameworks capsules have been synthesized by etching away MOF layer from @MOF@COF core-shell structure [85], Hydrogen-bonded Organic Frameworks have been used through bottom-up assembly [86].) One of the most celebrated classes of materials in this regard is Metal-Organic Frameworks, or MOF [63, 66, 86–91]. They have applications in the fields of biocatalysis [92], biosensors [93], biodelivery [94] and biobanking [95].

Design strategies:

Following are some examples of design strategies for MOFs or MOF-based protein biocomposites. There are three aspects - a) stability of metal ligation, b) stability of the linker and c) attachment of the protein to the MOF.

a) Stable ligation of metal-linker: One of the characteristics of MOF chemistry is the presence of metal-ligand bonds. Metal ions can have low or high oxidation states that make their radii higher or lower, respectively. High oxidation states with lower ionic radii metal ions (like Zr^{4+}, Cr^{3+}) are hard in nature. So, their coordination bonds with hard bases (like oxygen atoms in carboxylate linkers) will be strong according to the hard and soft acids and bases theory (HSAB). That is why Zr-MOFs and Cr-MOFs are some of the most stable MOFs. Larger ionic radii and lower oxidation states metal ions (Zn^{2+}, Ni^{2+}), being soft in nature, form stable frameworks with soft bases (like pyridine and azolates). Examples of such MOFs are PCN-601, Co(BDP), etc. However, hard acid-soft base or soft acid-hard base pairs usually give rise to less stable MOFs, for example - HKUST-1

and MOF-5 [87].

b) Linker stability: Hydrophobic ligands inside MOF pores can prevent water from approaching the metal-ligand bonds and thus make them water stable. Commonly used hydrophobic functional groups are fluoro, alkyl and aromatic groups. Synthetically, these are done by isoreticular substitution, de novo design or post-synthetic modification. Hydrophobic linkers enhance the hydrophobicity of the framework for targetted applications [87].

c) Protein-MOF biocomposites [66]: As explained earlier, the idea for attaching proteins/enzymes with MOF is to preserve its biocatalytic activity and help in recyclability. There are three ways to attach proteins/enzymes to MOF matrices - i) encapsulation, ii) pore infiltration, and iii) surface binding through covalent or non-covalent interactions. The first two are known as protein@MOF, and the third as protein-on-MOF. Surface immobilization is done by physically grafting enzymes onto the surface through covalent or non-covalent interactions. Encapsulation (in-situ, biomineralization, or de novo synthesis) proceeds via heterogeneous nucleation mechanism. Here, the enzyme serves as the nucleus for MOF to grow. Thus, proteins get physically confined inside the adventitious mesopores. The only condition for encapsulation is that the synthesis conditions should be biocompatible. Encapsulation is a one-step process, whereas infiltration is a two-step process. Pore infiltration requires pores large enough to accommodate enzymes and provide for substrate-product mobility. Stabilizing interactions (van der Waals and electrostatic) hold the protein structure, and specific interactions prevent leaching from the support. For pore diameters just larger than the protein, adsorption becomes reversible, or immobilization is slow (but adsorption is rapid), and loading becomes diffusion-controlled. Immobilization becomes rapid when the pore diameter is approximately three times larger than the protein size. One of the important practical problems with channel MOFs is the blocking of the channels by proteins/enzymes. This is mitigated in cage MOFs where hierarchical pores make it easy for reactant and product movement separately from the enzyme (protein) movement. Pore loading depends on many aspects - mesopore size and accessibility. Retention of loaded protein inside MOF depends on the availability of hydrophobic surface. It has been shown that proteins have an affinity towards hydrophobic surfaces [96]. Compared to surface-bound states, encapsulation and infiltration provide more stability to the biological cargo. Overall, there are different parameters in use for measuring the performance of immobilization [97] (Figure 1.6).

Some representative earlier studies:

Since the very first report [98], enzyme immobilization in MOFs has become a burgeoning research area. Many clinically important proteins like bovine serum albumin

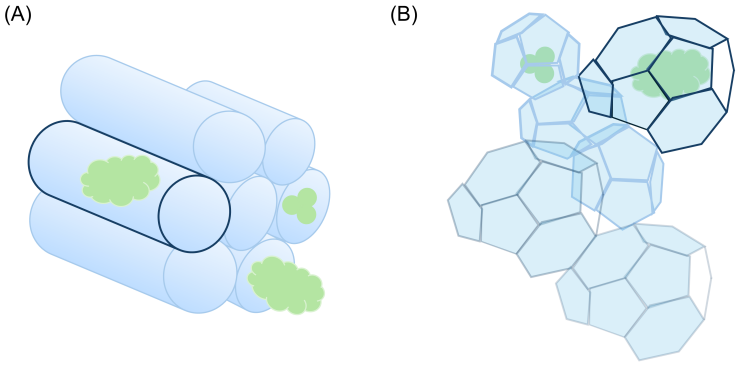


Figure 1.6: (A) Channel MOF and (B) cage MOF; with enzymes and reactant are shown in Green (the larger one being the enzyme). Channel and cage containing protein have been outlined in dark Blue colour.

[99], and cells such as HeLa [100] have been encapsulated using MOF biocomposite. This technique has also been used for the co-delivery of cis-platin for ovarian cancer treatment [101]. Surface binding by covalent and non-covalent means has also had many applications in biocatalysis, including drug delivery. Covalent attachment to the MOF surface exploits different coupling chemistry and click chemistry reactions. On the other hand, through the pore infiltration technique, several MOF families such as Tb-mesoMOF, IRMOF-74 series, PCN, and NU-100x series have been experimentally studied to immobilize enzymes such as MP-11 [102], myoglobin [103, 104], GFP [104], CytC [105], HRP [106], GOx [106], OPAA [107] and insulin [108]. A plethora of experiments have been performed to build MOF-biocomposites and use them for novel applications. Some of them have attempted to understand the protein-MOF interface. For example, electron paramagnetic resonance spectroscopy with site-directed spin labelling has been used to study the orientation of the protein at the MOF interface [109]. Using isothermal titration calorimetry, π-π interaction resulted in enthalpy-driven immobilization and hydrophobic MOF microenvironment makes inclusion an entropically driven process [108]. Ultrasound has been shown to “activate", i.e., open the gate to the heme group of horseradish peroxidase immobilized in ZIF-8 through experiments and MD simulation [110]. In another study using MD simulations and Umbrella Sampling, binding free energies of carbonic anhydrase and myeloperoxidase on the external surfaces of ZIF-8 and MIL-160 were obtained [111]. A handful of studies have employed molecular dynamics simulations to study infiltrated systems. Zhang and coworkers [112] studied the interaction of several individual amino acids within the pores of IRMOF-74. A simulation of a 20-residue miniprotein Trp-cage showed van der Waals (vdW) interaction to be the main driving force for inclusion in the MOF pore, with guest size and polarity balance in the MOF surface playing crucial roles. Hydrogen-bonding, salt-bridge, and π-π interactions between the protein and MOF have

been observed in the case of cutinase@IRMOF-74-VI [113]. Upon inclusion, an enzyme may need to change its shape marginally, as in the case of cutinase@NU-1000 [114] or can change the active site coordination as in the case of CytC@Nu-1000 [115]. All of these studies have been carried out with channel MOFs. A question arises in this context: how does the protein migrate or translocate between neighbouring cages, especially those whose dimensions are larger than the cage apertures? In biological systems, during trafficking between different organelles, proteins get unfolded and translocate through nanometre-sized pores [116–118], the phenomenon being known as co-translocational unfolding. Two experimental reports suggested a similar phenomenon for proteins during entry to the MOF [105, 119]. A part of this thesis has tried to address this process (Figure 1.7).

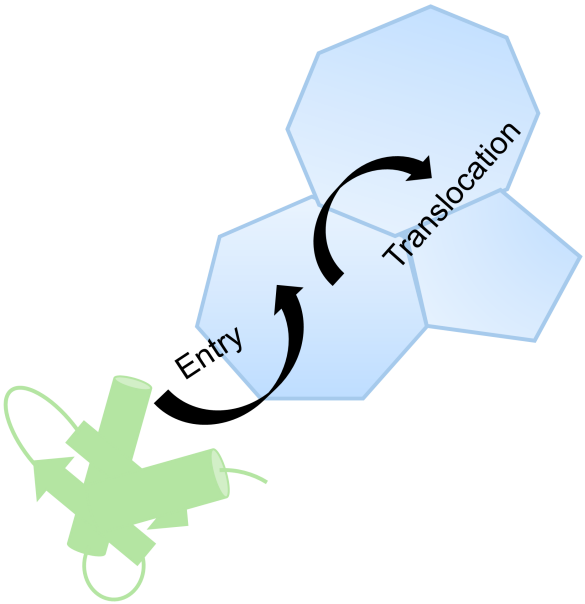


Figure 1.7: A cartoon representation for migration of proteins across the pores of MOF having dimensions larger than the pore sizes.

1.3.4 Post-translational modification in an Archaeal Enzyme

From the earlier idea of a one-sided flow of information, 21st century provides us with a different perspective on Central Dogma [120]. A large inventory of proteins constituting the proteome plays a significant role in information processing. Diversity in proteins stems from two routes: a) alternative m-RNA splicing [121] and b) Post-translational modifications (PTM). Due to PTM, backbone [122] and side chains [123] of proteins get different chemical nature. Some of them are enzyme-mediated, and some are not. These modifications include addition of functional groups and or cyclization of the backbones. One of the cyclized products is the succinimide ring. In most cases, succinimides are prone to hydrolysis, leading to aspartate and iso-aspartate residues. In an archaeal enzyme (glutamine amidotransferase, a subunit of GMP synthetase in *Methanocaldococcus jannaschii*), astonishing stability of the succinimide has been observed even up to 100° C [124]. Not formation of aspartate and iso-aspartate resides in gaining stabilization from the neighbouring residues through electrostatic (n–>π^*) and van der Waals interaction [125]. Mutational studies show the possibility of participation of nearby residues in succinimide formation (Chandrashekarmath et al., unpublished data from our experimental collaborators). A part of the thesis has tried to address how the reaction occurs inside the protein (Figure 1.8).

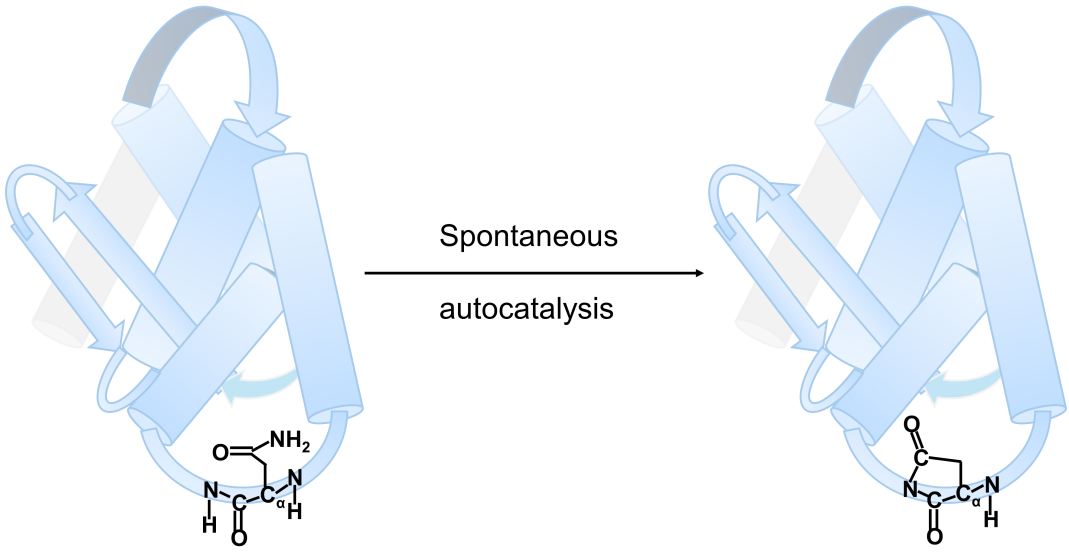


Figure 1.8: A cartoon representation of the post-translational modification in MjGATase - a spontaneous autocatalysis reaction.

1.4 Computer modelling: A quick history and this thesis

The first molecular dynamics simulation was done back in the mid 20^{th} century (64 years ago) [126]. The first simulation of a protein (Bovine pancreatic trypsin inhibitor) without considering solvent by J. Andrew McCammon, Bruce R. Gelin and Martin Karplus using mainframes captured dynamics for about 10^4 steps (9 ps in real-time, 1977) [127]. Since then, because of the very rapid progress in computer hardware, simulation methodologies in terms of force fields, solvation, sampling and workflows, it has now [128] become possible to look into the dynamics with all-atom description (including solvent) over 10^8-10^{10} steps (within hours of real-time). This made Molecular Dynamics reach timescales faster than Moore's law's prediction ([129], Figure 1). For this thesis, for our computer experiments, we have used the Molecular Dynamics library (GROMACS, https://manual.gromacs.org/2022.3/user-guide/index.html) and Quantum Chemistry library (CP2K, https://manual.cp2k.org/trunk/) in mainframe and supercomputer (PARAM Yukti, https://cdac.in/index.aspx?id=print_page&print=moscow,https://nsmindia.in/node/158).

1.5 Scope of the thesis

The research objective of the thesis are presented in brief. As described earlier, over the last decade, research on hybrid biomaterials has grown rapidly. Yet, atomistic modeling of these materials has been lagging behind developments in the synthesis and application domains. MOFs have been used to encapsulate biomolecules, drug molecules, ennzymes, etc., to protect biomatter from harsh operating environments (such as, say, an organic solvent), as drug delivery systems, and to retain the critical secondary structure elements for long-time storage of the biomolecules. Thus, there is a necessity to study these systems from a microscopic perspective using computer simulations. Along with the existing studies in the domain, my effort in understanding these systems are present in two chapters (**Chapter 2** and **Chapter 3**) of the thesis. Through **Chapter 2**, we observed the dynamics of these systems at equilibrium and that gave us some insights for future endeavour with these kinds of systems. In a sense, **Chapter 2** set the stage which helped me to address an almost unaddressed question in the field and that was the mechanism of translocation of larger size proteins across the cavities of the material (**Chapter 3**). I believe the insights from **Chapter 2** and the results from **Chapter 3** would add to the understanding of these systems. On the other hand, understanding reaction mechanism is a all-time quest. As described earlier, our experimental collaborators introduced us to the interesting reaction - a post-translational modification in an archaeal enzyme. Post-translational modifications (PTM) of a protein can occur either be on the side chain of a residue, or on its backbone. In the case of the archaeal enzyme, MjGATase, studied in this thesis, PTM occurs in the backbone and provides it considerable thermostability and function even up to the normal boiling point of water. The final work chapter,

Chapter 4 was able to understand how the intra-protein reaction is taking place with a partial success and shed light on improvement. Followings are short summaries for each chapter.

Chapter 1 provides a generalized description of modelling techniques at the molecular level, along with a short introduction to proteins, MOF, protein@MOF, and post-translational modifications (PTM).

In **Chapter 2**, results obtained from equilibrium molecular dynamics simulations of myoglobin and the Green Fluorescent Protein (GFP) in IRMOFs are presented. The current work is motivated by an experimental work that reported the formation of inclusion complexes of these biomolecules with the isoreticular MOF series, i.e., within the channels of IRMOF-74-VII-oeg and IRMOF-74-IX, respectively, where -oeg is the triethylene glocol monomethyl ether group. Employing extensive all-atom equilibrium molecular dynamics simulations, we observe that both these inclusions are mainly governed by van der Waals interactions at the protein-MOF interface. The confinement effect on myoglobin was larger than that of GFP due to the relatively smaller size difference between the former and its MOF host. The primary signature of the confinement was observed in the root mean squared fluctuations of the protein side chains. Although experiments could not succeed in the inclusion of myoglobin in IRMOF-74-VII-hex (where -hex is the hexyl group), our simulations suggest that it could be easily accommodated in the same, suggesting the possibility of kinetic contributions in the experimental observation. Overall, the tertiary structures and hydration of the surfaces of the proteins were well maintained inside the MOF channels for both proteins.

Chapter 3 examines the mechanism of translocation of proteins through the cavities of the MOF using a model protein-MOF system. The protein chosen is the chicken villin headpiece subdomain, HP35, and the MOF is MIL-101(Cr), as the former's diameter is larger than the size of the window between the cages of the MOF. Equilibrium molecular dynamics simulations demonstrate that the protein is located farther from the center of the cavity and closer to the MOF surface. Molecular interactions with the MOF partially unfold helix-1 at its N-terminus. The translocation of HP35 through the narrow, hexagonal windows of the mesopores of MIL-101 (Cr), a process that is vital to biomolecular inclusion, was studied using non-equilibrium molecular dynamics simulations. Steered molecular dynamics (SMD), followed by umbrella sampling simulations (US), show that the translocation process across the spherical cavity can be divided into three zones, resulting from both the geometry of the confinement and MOF surface chemistry; in one of them, the protein maintains nearly its native conformational ensemble which encompasses the equilibrium position of the protein. The free energy barrier for the unfolded protein at the cage window, relative to its equilibrium state within the cavity, is estimated to be 16 kcal/mol.

Chapter 4 reports results on the mechanism of an autocatalyzed reaction, the post-translational modification of an enzyme, glutamine amidotransferase (MjGATase), present in a hyperthermophilic archaea. It undergoes the spontaneous formation of succinimide at residue-109 (within an -END- sequence), due to which the enzyme is stable and functional even up to 100°C. Experimental collaborators have discovered through mutation studies of nearby residues that residues D110, Y158, and K151 might have roles in succinimide formation. QM/MM simulations with non-tempered Metadynamics are employed to address how succinimide is formed. The QM region consists of 4 residues (N109, D110, Y158, and K151), including the reactant residue (N109). The reaction proceeds through deprotonation followed by cyclization. Metadynamics simulations yield a free energy barrier for the deprotonation step to be 3.4 kcal/mol; the subsequent cyclization step is likely to be barrierless. However, while intermittent hydrogen bonding between Y158 and the side chains of D110 and N109 was observed, the exact role of Y158 in the product formation was not discernible.

The thesis concludes with **Chapter 5**, which summarizes the work presented and provides a future outlook.

2

Nanoconfinement of Myoglobin and Green Fluorescent Protein in IRMOF

2.1 Introduction

As introduced in Chapter 1, enzymes (proteins) are important components for designing smart ultrasmall machines [130] through nanoarchitectonics [131] and lab-on-a-chip devices [132, 133]. In this chapter, we have explored structure and interfacial interaction of channel-confined two large proteins - myoglobin and Green Fluorescent Protein inside IRMOF-74s MOF. This work was like our baby step for computationally studying protein@MOF systems.

The corresponding experimental study was one of the early proof-of-concept experiments in this domain appeared in the literature in 2012. Deng et al. [104] showed that IRMOF-74-VII-oeg and IRMOF-74-IX could incorporate myoglobin and green fluorescent protein (GFP) inside the channels of these MOFs, respectively. The same was reflected in the UV-Vis spectrum corresponding to the absorbance of the Soret band at 409 and 489 nm for the prosthetic HEME group of myoglobin and the chromophore of GFP, respectively. Confocal microscopy of the GFP@IRMOF-74-IX sample confirmed the retention of the protein structure. Two corresponding control systems for myoglobin

The work presented in this chapter has been ChemRxiv-ed (https://doi.org/10.26434/chemrxiv-2024-c8x0p).

and GFP were IRMOF-74-VII-hex and IRMOF-74-V-hex, respectively, and for both of them, the entry of the proteins inside the MOFs was negligible. Both IRMOF-74-VII-hex and IRMOF-74-V-hex possessed hexyl groups, IRMOF-74-VII-oeg contained triethylene glycol mono-methyl ether or -oeg groups and IRMOF-74-IX, too, contained hexyl groups. In the case of GFP, the channels of IRMOF-74-V-hex were too small to accommodate it. For myoglobin, the authors concluded that the hydrophobic nature of the organic linkers of IRMOF-74-VII-hex repelled the hydrophilic surface of myoglobin and hence curtailed its translocation through the MOF channels (Table 2.1). All the MOFs studied here contained one-dimensional channels within which the biomolecule could be included.

Table 2.1: A summary of experimental results from inclusion studies. [104]

MOF	Protein	Inclusion
IRMOF-74-VII-oeg	myoglobin	Yes
IRMOF-74-VII-hex	myoglobin	Negligible
IRMOF-74-IX	GFP	Yes
IRMOF-74-V-hex	GFP	Negligible

Atomistic molecular dynamics simulations have matured to a great extent in the modelling of enzymes and biohybrid materials [134, 135]. They can offer rich insights into microscopic interactions that govern experimental observations such as the inclusion or otherwise of proteins in the pores of structurally similar MOFs [112, 113, 115]. Thus, these MOF-protein complexes are investigated herein through extensive all-atom molecular dynamics (MD) simulations. The systems contained proteins surrounded by channel waters in three MOFs, namely, IRMOF-74-VII-oeg, IRMOF-74-VII-hex and IRMOF-74-IX for myoglobin and GFP inclusions, respectively (Table 2.2). In contrast to experiments, our simulations revealed that myoglobin could be facilely accommodated in the channels of IRMOF-74-VII-hex without any major structural changes of the protein. Its active site was also found to be well preserved. For both the inclusions of myoglobin and GFP, van der Waals interaction was observed to play a major role at the protein-MOF interface.

2.2 Computational details

2.2.1 Myoglobin

The initial structure of myoglobin was taken from the crystal structure of sperm whale myoglobin *Physeter Catodon* (RCSB PDB ID: 1MBC). From the carbonmonoxy myoglobin, deoxymyoglobin was prepared by removing the carbon monoxide ligand; modified parameters for this penta-coordinated heme was employed, following literature [136, 137] (Table A.4). For the prosthetic group (named HEM in PDB), hydrogens were added using GaussView followed by naming them properly to match rtp file names in accordance with the CHARMM force field. Both HEM and HEME were defined as 'Ligand' in the residuetype.dat file. The water molecules (137 in number) present in the crystal structure were retained in the initial configuration of the simulation. A covalent bond between the proximal histidine (resid 93) and HEME Fe was defined through the special bond feature of GROMACS. Protonation states for protein residues were assigned by GROMACS. Missing parameters for proper dihedral angles were assigned force constant and equilibrium values of 0 $kJmol^{-1}$ and 0°, respectively (Figure A.1).

2.2.2 Green Fluorescent Protein

The initial structure of GFP was taken from the crystal structure of jellyfish *Aequorea Victoria* (RCSB PDB ID: 2WUR). There were 2 missing residues near N_{TER}, 6 missing residues from C_{TER} and missing atoms in residue 232. All of these were added using PyMOL. The chromophore GYS was added as 'Protein' in the residuetype.dat file. The protonation states of HIS were taken from the PDB, which also contained the coordinates of hydrogen atoms. 318 water molecules present in the crystal structure were retained in the initial configuration of the protein. Na^+ counterions were added at arbitrary locations by replacing water molecules to neutralize the overall charge of the protein.

The secondary structure for the missing residues at the C_{TER} was determined by Well-Tempered Metadynamics [138]. The radius of gyration of 7 residues (232-238) was taken as the reaction coordinate (collective variable: cv) for the metadynamics run, and residues 1-231 were position restrained. A bias factor of 10.0 and an initial deposition rate of 5 $kJmol^{-1}ps^{-1}$ with a Gaussian potential width of 0.005 nm were used. The integration time step was 1 fs. Simulations were carried out in the NPT ensemble with isotropic Berendsen pressure coupling [139] with a coupling constant of 1 ps. The coupling constant for the Bussi-Donadio-Parrinello velocity-rescaling thermostat [11] was set to 1 ps. All other parameters were the same as in the equilibrium MD runs (Table 2.2). We extracted conformations belonging to the minimum in the free energy profile (Figure A.2, R_g approximately between 0.58 to 0.62 nm). Subsequently, a cluster analysis of these structures were done using the GROMOS method [140] present in GROMACS (Figure A.3). The central structure of the highest populated cluster was taken as the lowest free energy structure (Figure A.4), which was used for MD simulations

downstream.

2.2.3 System preparation

A GitHub repository (https://github.com/scidatasoft/mof, 03.01.2023) provided us with the unit cells of MOFs. This contained MOF structures from the CoRE MOF database [141, 142] along with partial charges on the atoms. We made the metal centers hexa-coordinated by ligating water molecules using GaussView. Using the QUICKSTEP module [143] of CP2K package (version: 7.1) [144] (implementing periodic density functional theory (DFT)), we carried out geometry optimization of the ligated water molecules. This method described Kohn-Sham molecular orbitals through a linear combination of atom-centered Gaussian-type orbitals and the electron density through auxiliary plane-wave basis set along with Goedecker-Teter-Hutter (GTH) pseudopotentials [145]. The convergence criteria for inner and outer loops for the self-consistent field (SCF) was set as 10^{-7}. PBE functional [146] was used with DFT-D3 [147] as dispersion corrections for exchange-correlation interactions.

Each primitive unit cell structure was replicated to a supercell (Table A.1) such that after protein inclusion, periodic images of protein were separated by approximately 30 Å. We took bonds, angles, and dihedrals from the crystal structure geometry and generated MOF topologies using the OBGMX [148] code. Each supercell channel was filled with water molecules. For this purpose, the GROMACS solvation module was used with a scaling factor of 0.47. With a distance tolerance of 1 Å, we kept the proteins in the MOF channels using PACKMOL [149], followed by rotation along the x-,y- and z- directions by arbitrary angles to obtain different initial orientations of the protein with respect to the MOF. Finally, water molecules having hard contact with the proteins within a 2.5 Å radius were removed.

To sample the orientational space of the protein with respect to the MOF, several independent MD simulations, each starting from an arbitrary orientation, were performed. The number of such simulations was: 4 (myoglobin in IRMOF-74-VII-oeg), 4 (myoglobin in IRMOF-74-VII-hex), and 7 (GFP in IRMOF-74-IX). These are displayed in Figure 2.1.

Abbreviations used for referring to different orientations in this chapter is as follows: $\text{Myoglobin}_{\text{W}}$ and GFP_{W} represent myoglobin and GFP in pure water simulations, respectively. VII-oeg, VII-hex, and IX have been used for simulations of supercells of IRMOF-74-VII-oeg, IRMOF-74-VII-hex and IRMOF-74-IX, respectively, where corresponding channels were filled with water molecules. $\text{O1...4}_{\text{myoglobinOEG}}$, $\text{O1...4}_{\text{myoglobinHEX}}$ and $\text{O1...7}_{\text{GFP}}$ correspond to orientations 1 to 4 and 1 to 7 for myoglobin@IRMOF-74-VII-oeg, myoglobin@IRMOF-74-VII-hex and GFP@IRMOF-74-IX, respectively. $\text{O}_{\text{myoglobinOEG}}$, $\text{O}_{\text{myoglobinHEX}}$ and O_{GFP} have been used for

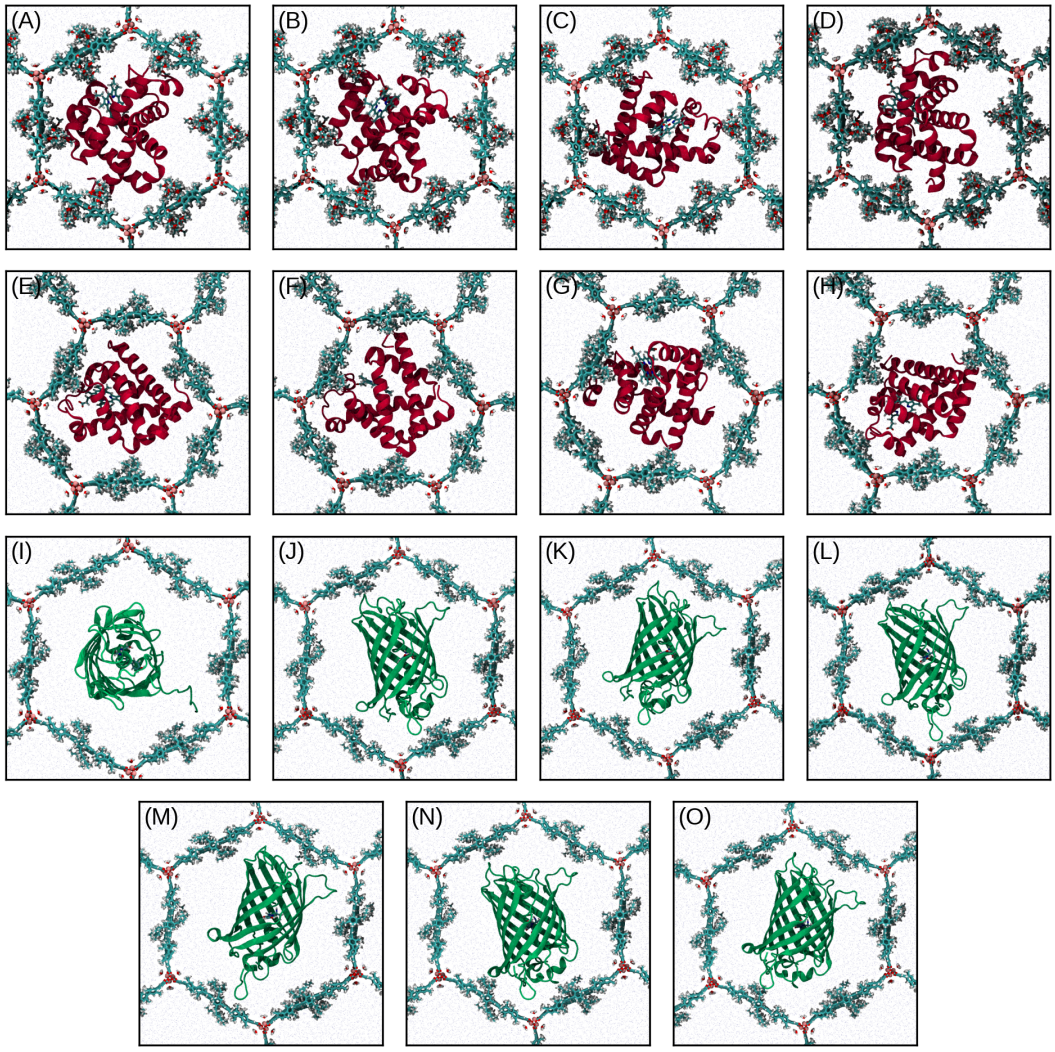


Figure 2.1: Myoglobin and GFP oriented variously within the channels of IRMOF-74. These 15 different orientations were used as starting configurations for protein@MOF MD runs. Panels (A) to (D) are for myoglobin@IRMOF-74-VII-oeg, (E) to (H) are for myoglobin@IRMOF-74-VII-hex, and (I) to (O) are for GFP@IRMOF-74-IX. MOF (except metal sites) is shown in the Licorice representation, and metal sites (Mg atoms) are shown in the vdW representation with a reduced sphere scale. Myoglobin is shown in Red in the New Cartoon representation. HEME of myoglobin is shown as Licorice representation. GFP is shown in Green in the New Cartoon representation, whereas the chromophore inside is highlighted in the Licorice representation. Water molecules filling the channels of the MOF are shown in Iceblue colour in the Lines representation (appear as dots). Ions are not displayed here for clarity.

myoglobin@IRMOF-74-VII-oeg, myoglobin@IRMOF-74-VII-hex and GFP@IRMOF-74-IX, respectively averaged over orientations.

2.2.4 Simulation details

All MD simulations were carried out using GROMACS (version: 2019). The bonded and non-bonded parameters from all-atom CHARMM36m-Jul2020 and UFF (except partial charges) force fields were used to describe the intramolecular interactions of proteins and MOFs, respectively. A rigid CHARMM TIP3P model was used for water molecules. For all our simulations, we have applied periodic boundary conditions (PBC) in all three directions, in addition to defining bonded parameters across PBC for MOF atoms. We used the partial charges for MOF atoms as provided in the same GitHub repository (https://github.com/scidatasoft/mof, 03.01.2023). Metal-ligated water molecules were described through TIP3P charges.

In accordance with the CHARMM family of force fields, the Lennard-Jones potential was smoothly brought down to zero from 10 to 12 Å using the force-switched approach. LJ parameters for two different atom types were obtained using Lorentz-Berthelot mixing rules. 12 Å cut-off was set for Coulomb interaction as well. The switching function for Coulomb interactions was the Potential-shift-Verlet. The particle mesh Ewald (PME) [28] method with an interpolation order of 4 and a relative tolerance of 10^{-5} was used for electrostatics calculation above 12 Å. Scaling factors for 1-4 non-bonded interactions were set in accordance with the CHARMM force field for all simulated systems.

A maximum step size of 0.1 Å and force tolerance of 10 kJmol^{-1}nm^{-1} were used in the steepest-descent algorithm for energy minimization. MD simulations were conducted in isochoric-isothermal (NVT) and isobaric-isothermal (NPT) ensembles. Analysis trajectories were run in the NVT ensemble. For temperature coupling, the Bussi-Donadio-Parrinello velocity-rescaling thermostat [11] with coupling constants 0.5 and 1.0 ps for NVT and NPT, respectively, was employed. Parrinello-Rahman barostat [12] was coupled isotropically for protein in water simulations with a coupling constant of 1.0 ps and anisotropically for MOF-containing systems with a coupling constant of 20 ps. Temperature and pressure were set at 300 K and 1 bar, respectively. Isothermal compressibility for pressure coupling was set to 4.5x10^{-5} bar^{-1}. Covalent bonds containing hydrogen atoms were constrained using the LINCS algorithm with order 4 and warn angle 30°. Equations of motion were integrated using Leap-frog algorithm with a time step of 0.5 and 1.0 fs for NPT and NVT, respectively. Position restraints for non-hydrogen atoms, when made, were done through a harmonic potential with a force constant of 10^3 kJmol^{-1}nm^{-2}.

We had three systems for equilibration: protein in water, solvated MOF supercell, and

protein in solvated MOF supercell. For protein-in-water simulations, the equilibration steps were:

(i) energy minimization of water molecules, position restraining the protein, (ii) energy minimization of protein molecules, position restraining water, (iii) Temperature increase from 0 K to 300 K over 2 ns and doing NVT run at 300 K for 1 ns with no position restraints, (iv) NPT for 5 ns with no restraints.

For the solvated MOF supercell, we carried out the following steps:

(i) energy minimization of water molecules, position restraining MOF, (ii) Temperature increase from 0 K to 300 K over 2 ns and doing NVT run at 300 K for 1 ns, position restraining the MOF (iii) NPT for 5 ns with no restraints.

Constant NPT simulations reproduced the supercell volume of experiments within 5-8 % (Table A.5, and Figure A.5). The last time frame of the NPT trajectory was used to pack proteins.

For the protein in the solvated MOF supercell, the following steps were adopted:

(i) energy minimization of water molecules, position restraining both MOF and the protein, (ii) energy minimization of the protein, position restraining MOF and water, (iii) Temperature increase from 0 K to 300 K over 2 ns and doing NVT run at 300 K for 1 ns, position restraining the MOF (iv) NPT for 5 ns with no restraints.

The production runs were carried out for the protein in water and the protein in solvated MOF supercell systems in the NVT ensemble with no restraints on any atom of the systems. The last 100 ns chunk of production trajectories were analyzed (Figures A.6, A.7, and A.8). In summary, the cumulative simulation lengths for canonical sampling were 1.55 μs, 1.25 μs, and 4.65 μs for myoglobin@IRMOF-74-VII-oeg, myoglobin@IRMOF-74-VII-hex and GFP@IRMOF-74-IX, respectively along with proteins in neat water simulations for 500 ns each (Table 2.2).

2.3 Results and Discussions

2.3.1 Pore size distributions

At first, we calculated the pore size distributions of experimental unit cells (and the corresponding primitive unit cells) using Zeo++ [150–155]. It turned out that the channels of IRMOF-74-VII-hex were smaller than those of IRMOF-74-VII-oeg (Table 2.3), thus disagreeing with the argument provided in the supporting information of the original

experimental paper [104] for the experimental observation of non-inclusion of myoglobin in the former MOF, i.e., that it is the hydrophobic nature of the IRMOF-74-VII-hex that prevented the inclusion of myoglobin. Thus, to investigate further, we carried out MD simulations of myoglobin inclusion in IRMOF-74-VII-hex as well.

2.3.2 Protein location and conformation upon inclusion

A critical analysis of the proteins after their inclusion in the channels of the MOFs through the visualization of their MD simulation trajectories (see movies, links been provided in Appendix A) revealed that they predominantly prefer to move away from the channel axis and towards the MOF surface. This also tallied with our other result [156] of a 35-residue protein, HP35, included in MIL-101(Cr), wherein the protein's equilibrium position was found to be around 8 Å away from the pore center and closer to the internal surface of the MOF pore.

Relative to their values in neat water, the root mean squared fluctuations of the side chains of residues of myoglobin and GFP were attenuated inside MOFs (true for all orientations, see Figures A.9, A.12, and A.10). Figure 2.2 displays the RMSF averaged over all the runs in the MOF initiated from different relative orientations. A few residues show higher fluctuation than in water, but the difference is quite small (within 1 Å). For the overall structure (excluding a couple of residues at the termini), the mean RMSF decreased by 3.8% and 15%, respectively, for myoglobin in IRMOF-74-VII-oeg and IRMOF-74-VII-hex. For GFP, this value increased marginally by 2.7% upon inclusion (Tables 2.4, and 2.5). When considering all the residues (Tables A.6, and A.7) for myoglobin (153 residues), it was found that inside IRMOF-74-VII-hex, the side chain fluctuations decreased by 16.9 %, and in IRMOF-74-VII-oeg, the fluctuation amplitude was comparable to that in neat water. For GFP (238 residues), a decrement by 8% (Tables A.6, and A.7 and Figures A.9, A.12, and A.10) was seen. These observations show that the confinement of the proteins in the MOF channels reduced their conformational flexibility. The effect of confinement is more prominent in the case of IRMOF-74-VII than in IRMOF-74-IX.

Unlike RMSF, which showed a general decrement upon immobilization of the protein in the MOF channel relative to that in neat water, their secondary structure motifs and proportions were largely preserved as seen in Figures 2.3, A.13, A.14, and A.15.

Water access to the side chains of the surface residues, particularly the hydrophilic and polar ones, is important for the function of the enzymes. Thus, we calculated the number of non-hydrogen contacts between the oxygen of the water and the heavy atoms of the protein with a distance cut-off of 3.5 Å. Table 2.6 suggests that the reduction in the extent of hydration is about 8-9% and 4% for myoglobin and GFP, respectively. The same

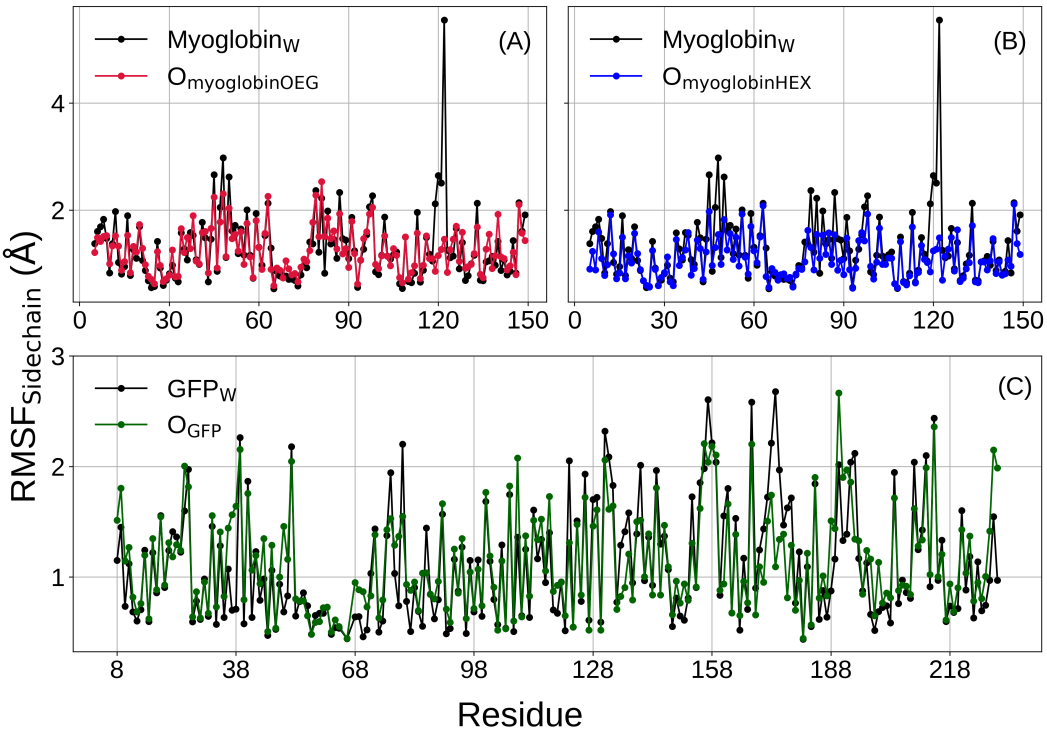


Figure 2.2: Root mean squared fluctuations of side chains of protein residues in pure water and protein@MOF systems. (A) myoglobin@IRMOF-74-VII-oeg, (B) myoglobin@IRMOF-74-VII-hex, and (C) GFP@IRMOF-74-IX. Four residues from each terminus for myoglobin and seven and eight residues from the N- and C-termini for GFP, respectively, were excluded from this calculation.

is reflected in the number of hydrogen bonds between proteins and solvent molecules (Tables A.13, and A.14) and protein-solvent interaction energies as well (Table 2.7). Water being polar, its interaction with the protein has a larger electrostatic component than dispersion. However, the changes in protein-solvent interactions did not discriminate between myoglobin@IRMOF-74-VII-oeg and myoglobin@IRMOF-74-VII-hex. The solvent contribution to the total potential energy was reduced in magnitude for both of these systems.

The inclusion in the MOF did not perturb the active sites of myoglobin and GFP, as shown in Tables A.8, A.9, A.10, and A.11 which provide RMSD and RMSF values for the HEME group of myoglobin and the chromophore of GFP.

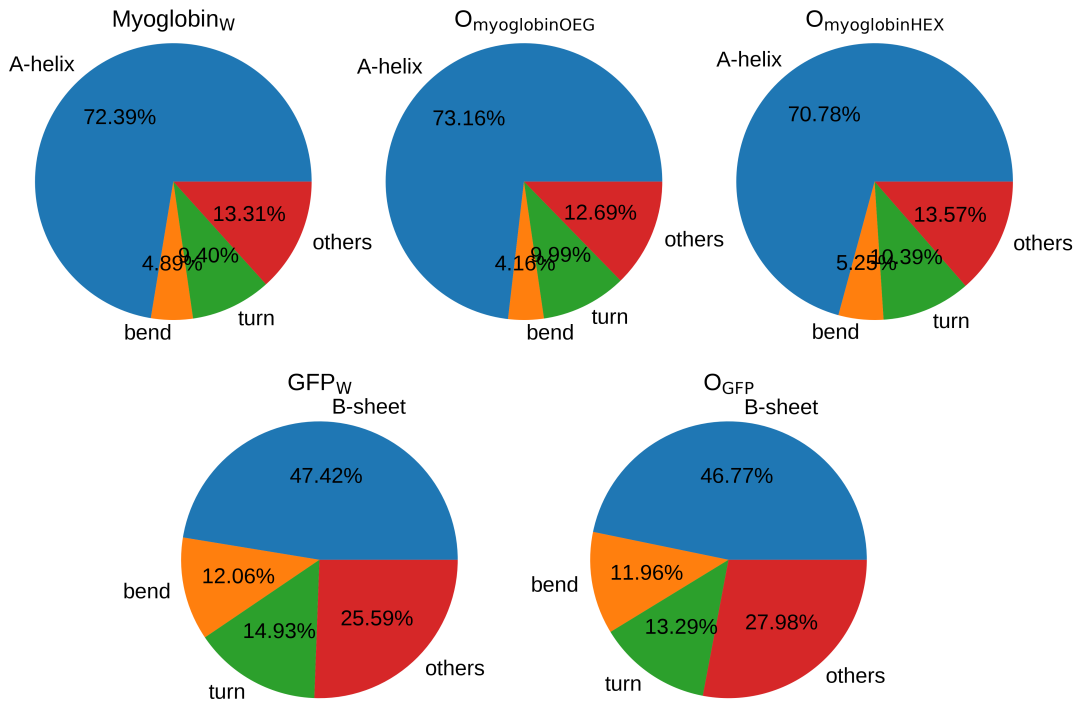


Figure 2.3: Secondary structure of myoglobin and GFP in water and inside the MOF channels.

2.3.3 Interactions at the Protein-MOF interface

The protein-MOF interface is mainly driven by Lennard-Jones interaction as shown in Table 2.8 and Figures A.19, and A.20.

To gauge the contribution of dispersion interactions in stabilizing the protein in the MOF channel, we calculated the number of non-hydrogen contacts between the proteins and the heavy atoms of the MOF with a distance cut-off of 4 Å (Tables A.2 and A.3, Figures A.21, A.22, A.23, A.24, A.25, A.26, A.27, A.28, A.29, A.30, A.31, A.32, A.33, A.34, and A.35). On average, polar residues of myoglobin interact more with the IRMOF-74-VII-oeg surface (Figure 2.4). In the case of myoglobin@IRMOF-74-VII-hex, both polar and nonpolar residues interacted with the MOF surface (Figure 2.5). For GFP@IRMOF-74-IX, the polar residues interacted more (Figure 2.6).

The MOF frameworks themselves can be divided into 'MainChain' and 'SideChain' components (Figure A.16), where the 'SideChain' are either -hex or -oeg herein and the 'MainChain' is the main linker in the framework. We observed that while myoglobin included in IRMOF-74-VII-oeg displayed more number of contacts with 'SideChain' atoms than with the 'MainChain' atoms, in the case of IRMOF-74-VII-hex, both 'MainChain' and 'SideChain' atoms interacted with myoglobin fairly equally. In the

case of IRMOF-74-IX, more number of 'MainChain' atoms were in contact with GFP than 'SideChain' atoms (Table 2.9, Figures A.17, and A.18). These again resonated with the fact that IRMOF-74-VII-oeg was suitably designed (experimentally) for myoglobin inclusion. However, we could not identify any factor which impedes the inclusion of myoglobin in the channels of IRMOF-74-VII-hex, unlike the experimental observation [104].

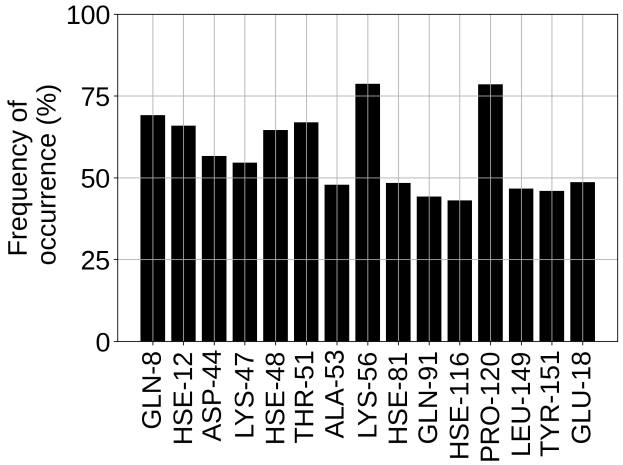


Figure 2.4: Residues of myoglobin whose heavy atoms lie within 4 Å of any of those of IRMOF-74-VII-oeg. Data is averaged over four orientations. Only residues that satisfy this criterion over at least 40% of the simulation frames are shown.

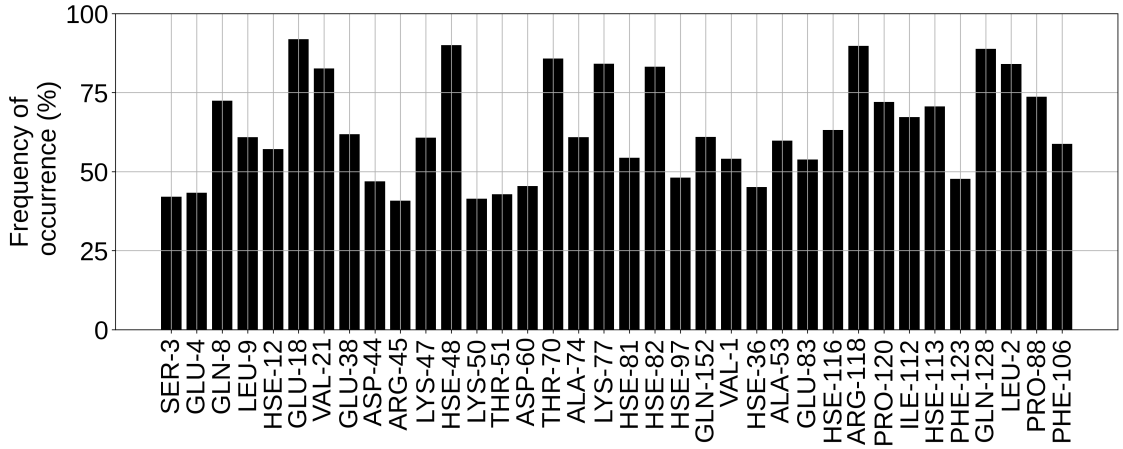


Figure 2.5: Residues of myoglobin whose heavy atoms lie within 4 Å of any of those of IRMOF-74-VII-hex. Data is averaged over four orientations. Only residues that satisfy this criterion over at least 40% of the simulation frames are shown.

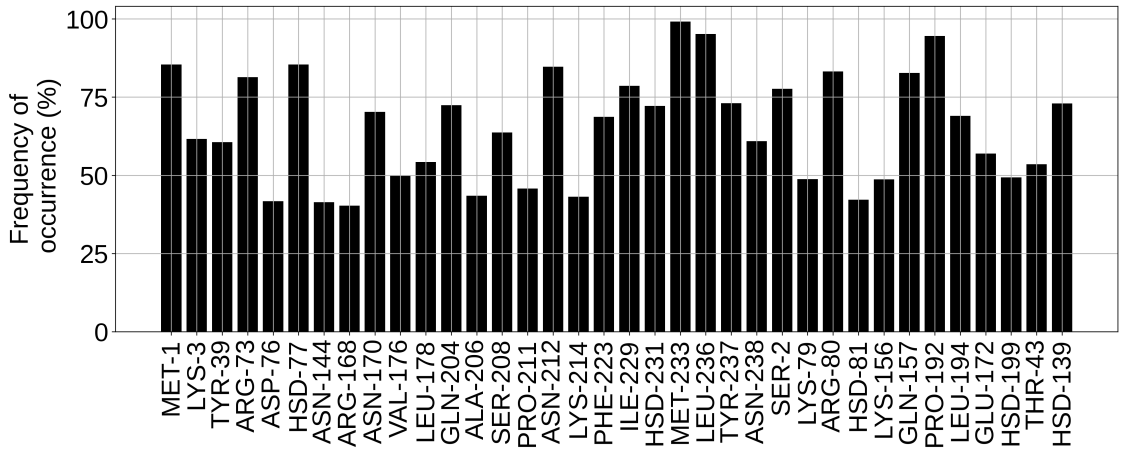


Figure 2.6: Residues of GFP whose heavy atoms lie within 4 Å of any of those of IRMOF-74-IX. Data is averaged over seven orientations. Only those residues that satisfy this criterion over at least 40% of the simulation frames are shown.

Table 2.2: Systems simulated ($\#_{\mathrm{MOF}}$ = Number of MOF atoms, $\#_{\mathrm{Protein}}$ = Number of protein atoms, $\#_{\mathrm{Water}}$ = Number of water atoms, $\#_{\mathrm{Ions}}$ = Number of ions, R_{Length} = Total run length, and A_{Window} = Analysis window).

System	$\#_{\mathrm{MOF}}$	$\#_{\mathrm{Protein}}$	$\#_{\mathrm{Water}}$	$\#_{\mathrm{Ions}}$	R_{Length} (ns)	A_{Window} (ns)
$\mathrm{Myoglobin_W}$	-	2532	32010	-	500	400 - 500
$\mathrm{GFP_W}$	-	3725	61881	7	500	400 - 500
VII-oeg	24648	-	69921	-	-	-
VII-hex	22464	-	70455	-	-	-
IX	23232	-	100596	-	-	-
$\mathrm{O1_{myoglobinOEG}}$	24648	2532	67278	-	200	100 - 200
$\mathrm{O2_{myoglobinOEG}}$	24648	2532	67122	-	500	400 - 500
$\mathrm{O3_{myoglobinOEG}}$	24648	2532	67233	-	350	250 - 350
$\mathrm{O4_{myoglobinOEG}}$	24648	2532	67143	-	500	400 - 500
$\mathrm{O1_{myoglobinHEX}}$	22464	2532	67701	-	450	350 - 450
$\mathrm{O2_{myoglobinHEX}}$	22464	2532	67752	-	200	100 - 200
$\mathrm{O3_{myoglobinHEX}}$	22464	2532	67806	-	200	100 - 200
$\mathrm{O4_{myoglobinHEX}}$	22464	2532	67743	-	400	300 - 400
$\mathrm{O1_{GFP}}$	23232	3725	96543	7	800	700 - 800
$\mathrm{O2_{GFP}}$	23232	3725	96570	7	800	700 - 800
$\mathrm{O3_{GFP}}$	23232	3725	96558	7	300	200 - 300
$\mathrm{O4_{GFP}}$	23232	3725	96600	7	300	200 - 300
$\mathrm{O5_{GFP}}$	23232	3725	96549	7	800	700 - 800
$\mathrm{O6_{GFP}}$	23232	3725	96579	7	800	700 - 800
$\mathrm{O7_{GFP}}$	23232	3725	96576	7	850	750 - 850

Table 2.3: Pore Diameters obtained using Zeo++ software [150] of the MOFs based on their experimentally reported CIFs.

MOF	Diameter (Å)
IRMOF-74-VII-oeg	38
IRMOF-74-VII-hex	35
IRMOF-74-IX	54

Table 2.4: Root mean squared fluctuations averaged over the primary structure. Four residues from each terminal were excluded from the calculation.

Water (Å)	$O_{myoglobinOEG}$ (Å)	$O_{myoglobinHEX}$ (Å)
1.30	1.25	1.10

Table 2.5: Root mean squared fluctuations averaged over the primary sequence. Seven and eight residues from the N- and C-termini, respectively, were excluded from the calculation.

Water (Å)	O_{GFP} (Å)
1.11	1.14

Table 2.6: Average number of water molecules around the protein in neat water and in Protein@MOF systems.

Protein	Protein in water	Protein@MOF
Myoglobin	419	380 ($O_{myoglobinOEG}$)
Myoglobin	-	385 ($O_{myoglobinHEX}$)
GFP	621	596 (O_{GFP})

Table 2.7: Average protein-solvent interaction energies for simulations in pure water and inside MOF channels.

Protein	Coulomb, LJ (in water, kcal/mol)	Coulomb, LJ (in MOF, kcal/mol)
Myoglobin	-3647.9, -219.9	-3435.5, -158.8 ($O_{myoglobinOEG}$)
Myoglobin	-	-3486.2, -161.6 ($O_{myoglobinHEX}$)
GFP	-5562.5, -341.7	-5557.5, -271.8 (O_{GFP})

Table 2.8: Average protein-MOF interaction energies.

Protein	Coulomb, LJ (kcal/mol)
Myoglobin	-43.0, -104.0 ($O_{myoglobinOEG}$)
Myoglobin	-4.5, -135.1 ($O_{myoglobinHEX}$)
GFP	-16.0, -187.7 (O_{GFP})

Table 2.9: Average number of MOF atoms within 4 Å of the protein surface.

Protein	MainChain atoms (MOF)	SideChain atoms (MOF)
Myoglobin	9 ($O_{myoglobinOEG}$)	39 ($O_{myoglobinOEG}$)
Myoglobin	24 ($O_{myoglobinHEX}$)	26 ($O_{myoglobinHEX}$)
GFP	52 (O_{GFP})	26 (O_{GFP})

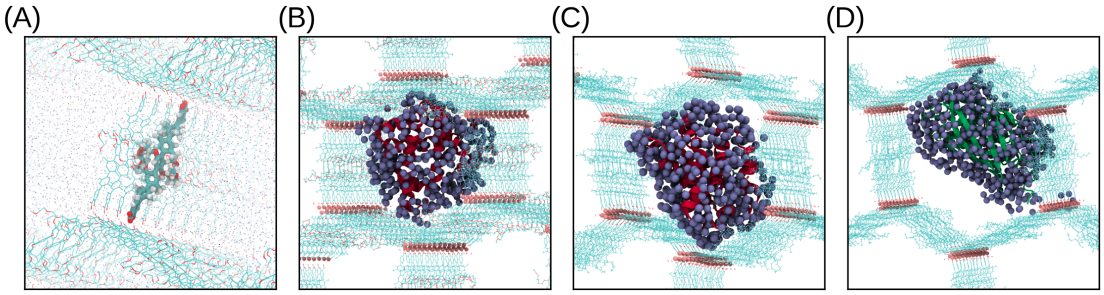


Figure 2.7: (A) Disposition of the organic linker of IRMOF-74-VII-oeg along MOF channel. (B), (C) and (D) show the water coating around myoglobin and GFP in IRMOF-74-VII-oeg, IRMOF-74-VII-hex and IRMOF-74-IX, respectively.

As shown in Figure 2.7 (A), the linker arrangement along the MOF channel was such that there was almost no possibility of π-π interaction between the aromatic side chains of proteins and the linker itself. The possibility of hydrogen bonding existed only in the case of myoglobin inclusion in IRMOF-74-VII-oeg. Across all four orientations, the number of protein-MOF hydrogen bonds was very small (Table A.12). But, the proteins were fairly well solvated, at least in the primary solvation shell; hence, water-bridged hydrogen bonding between the protein and the MOF for myoglobin@IRMOF-74-VII-oeg cannot be ruled out. The average number of OEG groups which interacted with myoglobin was 31.25, and if we consider those remaining with contact for more than 10% of the analysis trajectory, the same value was 18.5 (Figure A.36).

2.4 Conclusions

This work reports large-scale, all-atom molecular dynamics simulations of the encapsulation of two proteins, myoglobin and the green fluorescent protein, in two different metal-organic framework solids. The former was studied in an IRMOF with two kinds of substitutions on its linkers, one which makes the channel surface hydrophilic (the -oeg substitution) and another with a hydrophobic (-hexyl substitution) character. The proteins were studied in varied orientations in these MOFs.

One of our chief observations is the intact secondary structure of the proteins in these immobilized matrices relative to that seen in neat liquid water. However, the flexibility of the protein side chains, as probed through its RMSF, was found to be curtailed for the encapsulated macromolecules as compared to those in bulk water. The key interaction intermediating the protein and the MOF is van der Waals in character. The simulations also suggest that myoglobin inside IRMOF-74-VII-hex is enthalpically stabilized. Thus, the absence of encapsulation of myoglobin in this MOF reported in experiments may have a kinetic origin, which is beyond the capability of the current simulations to probe.

Although specific protein-MOF interactions for the present system has not been reported experimentally, a few reports of other systems have appeared in the literature. Raman spectroscopy revealed [102] $\pi - \pi$ interactions between the heme group of microperoxidase and the organic ligands of a Tb-mesoMOF, which facilitated the retention of the enzyme in this framework solid. Isothermal calorimetry has shown the adsorption of lactoglobulin A in ZIF-8 MOFs to be exothermic, in a similar manner as observed in the current work [157]. Combined ^{1}H, ^{27}Al, and ^{13}C NMR spectroscopy [158] have shown the stabilization offered by van der Waals interactions between lipase and the Al oxo cluster of PCN-333 MOF. And, this work tallies with our observations.

Our effort in this chapter occupies a tiny portion in the realm of proteins' adhesion to surfaces of different chemical natures under the sub-category of 2D confinement. Each protein carries its own 'unique molecular personality', unlike small molecules (gas molecules), which can be treated as rigid. And, because of this heterogeneity, different faces of the protein interact with the same surface differently [159–161] and the mechanism of adsoprtion would also be different. Thus, the 'protein-wettability' of any surface depends both on the surface and the proteins. The post-adsorption conformational changes of a protein depend on how much 'hard' or 'soft' the protein is. Hardness or softness essentially indicates strong or weak internal cohesion, respectively. Hydrophobic effect and electrostatic interactions are the two significant interactions for protein adsorption. And, also pH conditions (changing the charges on protein), ionic strength, and temperature have influences [162, 163]. There are other computational and kinetic modelling studies also to understand this kind of flatland adsorption of proteins [96, 164]. On the other hand, under 2D confinement, the properties of water molecules are quite different from the bulk water [165]. In this context, the mechanism of protein adsorption would be even more complex. Therefore, we believe that improving the modelling methodologies for the various components of the system (protein, surface, and water) would help us get a better real picture.

3

HP35 Protein in the Mesopore of MIL-101(Cr) MOF: A Model to Study Co-translocational Unfolding

3.1 Introduction

As introduced in Chapter 1, this chapter describes protein migration between neighbouring cages of Metal-Organic Frameworks. And the translocation is simultaneously accompanied by unfolding of the protein. We have investigated this in a reconstituted architecture with a model protein.

One of the model systems extensively studied in protein folding-unfolding processes is the chicken villin headpiece (HP35). HP35, an actin-bundling protein, contains two domains: "core" and "headpiece." The headpiece is the F actin-binding domain in the C-TER of super villin [166, 167]. HP35 is the 35-residue subdomain within the headpiece domain. It is one of the smallest monomeric polypeptides, made of naturally occurring amino acids that fold independently and autonomously into a unique and thermostable structure with a melting temperature of 342 K [168–170]. Subdomains that fold independently are important for studying protein folding. As shown by McKnight

The work presented in this chapter was performed in collaboration with Prof. Anand Srivastava, Indian Institute of Science, India. This work is published in *ACS Omega* **2024**, 9, 28, 31185-31194. Reproduced according to the Open Access permission from American Chemical Society, © 2024.

and co-workers, HP35 undergoes a cooperative thermal unfolding transition (unlike molten globules) [168]. Like other small proteins that fold at sub-microsecond timescale, HP35 also shows low folding cooperativity and a low energy barrier [171]. It displays the properties of a fully folded protein with a unique structure (i.e., a clear secondary structure and a well-packed core). It has a three-helix (alpha helices) topology with a closely packed hydrophobic core involving three phenylalanine residues. This structural architecture was revealed by both NMR and XRD experiments [170, 172].

Among the several experimental reports of protein@MOF systems, two merit specific attention. Chen et al. [105] demonstrated the immobilization of cytochrome c (Cyt c) in the mesopore of Tb-mesoMOF. While the dimensions of the enzyme in its native state are 2.6x3.2x3.3nm, that of the window to the pore is just 1.7nm. Thus, they argued that the enzyme should unfold during its translocation across the MOF pores, which was confirmed through changes in time-dependent fluorescence spectra upon the uptake of the enzyme by the MOF from an aqueous solution. Gkaniatsou et al. [119] were able to immobilize a mini-enzyme, microperoxidase-8 (MP8) whose native structure has dimensions 3.3x1.1x1.7nm in MIL-101(Cr) MOF. Here again, the enzyme is larger than the window between the pores; thus, the process of its incorporation in the MOF pores must be accompanied by conformational changes; however, in both platforms of (Cyt c)@Tb-mesoMOF and the MP8@MIL-101(Cr), the enzymes are shown to be functional. Thus, they should have folded back to their native states within a pore. Herein, we seek a microscopic understanding of the above experimental findings through a well-constructed model system of HP35@MIL-101(Cr), which shares the size-related characteristics of these two experimentally studied platforms.

In this chapter, where we model the co-translocational unfolding of HP35 in MIL-101(Cr), we have explored (I) the equilibrium location of the protein within the mesopore of the MOF and the energy components that stabilize it, (ii) how MOF confinement affects the structure of HP35 while it moves from one cavity to the neighbouring one, (iii) how confined waters access the protein surface, (iv) and the free energy barrier for protein translocation from the center of the cavity towards the window that separates two neighbouring cavities of the MOF.

Notations that have been used in the following sections are as follows - center of mass of the protein = P_{COM}, center of mass of the cavity = C_{COM}, and reference NMR structure = S_{NMR} and the corresponding definitions are explained in Appendix B.

3.2 Materials and Methods

3.2.1 A hierarchically porous MOF: MIL-101(Cr)

MIL-101(Cr) [173] possesses an MTN zeotype architecture formed by corner-shared super tetrahedra (ST) comprising 1,4-benzenedicarboxylate (BDC) linkers and Cr atoms; each Cr atom has an octahedral environment consisting of four oxygen atoms from BDC, one μ_3-O atom, and one water molecule (Figure B.1). The STs are microporous with an 8.6 Å window aperture. Through corner-sharing of STs, two types of mesoporous cages are formed with internal diameters of 29 Å and 34 Å, respectively. The former has pentagonal windows with a free aperture of 12 Å diameter, while the latter cages are accessible through either pentagonal or hexagonal windows (14.5 x 16 Å free aperture). MIL-101 has been used for the infiltration of microperoxidase-8 (MP-8) [119] and *Aspergillus saitoi* proteinase [174] enzymes. Experimentally determined powder X-ray diffraction patterns of the MOF with and without the enzyme are identical [119]; thus, MIL-101(Cr) is expected not to have structural transformation upon the inclusion of the HP35 protein.

3.2.2 HP35 as a model system for co-translocational unfolding

HP-35 has radii of gyration of 6.2 Å x 7.4 Å x 8.3 Å along its principal axes and cannot translocate across the cavities of MIL-101(Cr) without unfolding due to the smaller dimensions of the intervening window (Figure B.2). Hence, it is a suitable model system for studying the process of co-translocational unfolding. However, the dimensions of the protein are much smaller than the diameter of the mesopore of the MOF; thus, it can be well incorporated in the pore.

3.2.3 Preparation of protein containing MOF supercell

The initial structure of the protein for the simulations was the one with the lowest energy among those solved via NMR experiments (S_{NMR}) from RCSB (PDB ID: 1UNC) [175]. The primitive unit cell of the MOF crystal structure (OCUNAC_manual; CCDC 605510) was taken from a GitHub repository (`https://github.com/scidatasoft/mof,03.01.2023`), where partial charges of cleaned MOF structures from the CoRE MOF database [141, 142] are provided. As the sixth coordination of the metal (Cr) was missing in the primitive unit cell, we added the water molecules manually to the secondary building unit (SBU) of the MOF using Gaussview (version: 5.0.9) and subsequently performed geometry optimization using periodic density functional theory (DFT) implemented in the QUICKSTEP module [143] of CP2K package (version: 7.1) [144]. In this method, a linear combination of atom-centered Gaussian-type orbitals are used to describe the Kohn-Sham orbitals, and the electron density is described in an auxiliary plane-wave basis set in conjunction with

Goedecker-Teter-Hutter (GTH) pseudopotentials [145]. An accuracy of 10^{-7} was used for both inner and outer loops' self-consistent field (SCF) convergence. PBE exchange-correlation functional [146] was used, and dispersion corrections were incorporated using the DFT-D3 approach [147]. The geometry-optimized distance between the metal and the water was 2.30 Å (Figure B.1). We added the water molecules at this distance to the primitive unit cell and replicated it to a supercell of size 2x2x2. The OBGMX code [148] was used to generate the topology of the MOF supercell. Subsequently, solvent water molecules were added using the GROMACS solvation module to fill the supercell using a scaling factor of 0.467. The number of water molecules required to fill the supercell (hence the scaling factor) was approximated from the void volume of the supercell calculated using Biovia Materials Studio 2020 [176]. The protein was packed at the center of a central big cavity in the equilibrated MOF supercell structure using PACKMOL (version: 20.2.2) [149] with a distance tolerance of 1 Å. This was followed by the removal of water molecules within 2.5 Å of the protein (Figure 3.1).

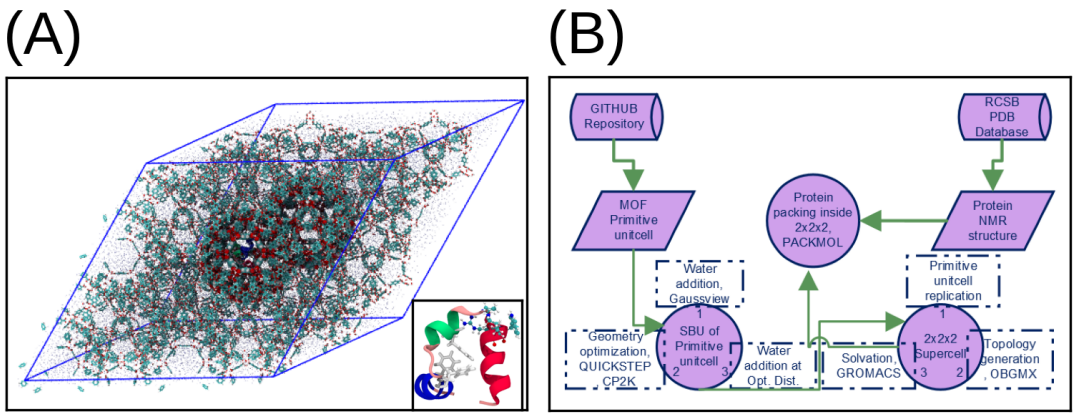


Figure 3.1: (A) Simulation box (supercell) of MIL-101(Cr) with HP35 present in one of the large cavities (highlighted in vdW representation along with the neighbouring cavity) having dimensions 12.568 nm x 12.568 nm x 12.568 nm and 60°cell angles. Water molecules fill all the pores of the MOF and are shown in CPK representation with reduced scale in Iceblue colour. The counterion is not shown. The inset shows geometric motifs of HP35. The hydrophobic core (residues 6,10,17) is highlighted in the White Licorice representation. The PXWK motif (residues 21-24) is highlighted in CPK representation. Helix-1 (residue 3-10), 2 (residue 15-19), and 3 (residue 22-33) are in Blue, Green, and Red, respectively. (B) Flowchart for system preparation.

3.2.4 General protocol for the simulations

All molecular dynamics simulations were performed using GROMACS 2022.3 [177–181]. We employed the all-atom AMBER99sb-star-ildn [182–184] bonded and non-bonded

parameters for the protein, UFF [8, 185, 186] bonded and non-bonded parameters (except partial charges) for the MOF and the rigid TIP3P model [9] for water molecules. Periodic boundary conditions (PBC) were applied in all the three directions. PBC for the MOF included intramolecular potential terms (bond, angle, torsion) across minimum images. Partial charges for MOF atoms (except metal-ligated water) were taken from a repository (https://github.com/scidatasoft/mof, 03.01.2023), wherein they were assigned using a machine learning model [187] which had combined the high accuracy of density-derived electrostatic and chemical charge (DDEC) method and the scalability of the charge equilibration (Qeq) method. TIP3P charges were assigned to the metal-ligated water. Real space cut-offs for the Lennard-Jones (LJ) and Coulomb interactions were set to 10 Å. Lorentz-Berthelot mixing rules were used to obtain the LJ parameters between two atom types. The particle Mesh Ewald (PME) [28] method with an interpolation order of 4 and a relative tolerance of 10^{-5} was used to calculate the electrostatic interactions for distances above 10 Å. Scaling factors for 1-4 non-bonded interactions were set following the AMBER force field. Long-range dispersion corrections were applied to calculate energy and pressure.

Energy minimization was done using the steepest-descent algorithm with an initial step size of 0.01 nm and force tolerance of 10 kJmol^{-1}nm^{-1}. Simulations were performed in the isochoric-isothermal ensemble (NVT). For temperature coupling, Bussi-Donadio-Parrinello velocity-rescaling thermostat [11] with a coupling constant of 0.5 ps was used. The temperature was set to 298 K. Covalent bonds to hydrogen atoms were constrained using the LINCS algorithm [29, 30] with order 4 and warn angle 30°. During equilibration, heavy atoms were position-restrained with a force constant of 10^3 kJmol^{-1}nm^{-2}. Equations of motion were integrated using the leap-frog algorithm with a time step of 1.0 fs.

For equilibrium MD simulations, the following procedures were followed. Before the insertion of the protein into the MOF cavity, the MOF supercell (cavities filled with water molecules) was equilibrated. The steps followed were (i) energy minimization for solvent water, (ii) NVT annealing from 0 K to 298 K over 2 ns followed by equilibration at 298K for 1 ns. (iii) A short production run for 5 ns under constant NVT conditions followed. For the equilibration of the protein@MOF system (the complete system under study), an intermediate step of energy minimization of the protein alone was carried out. During energy minimization of the protein, water was position restrained and vice versa. Non-hydrogen atoms of the MOF were position restrained in all equilibration steps. So, the general protocol was: (i) energy minimization of water, (ii) energy minimization of protein, (iii) and NVT equilibration.

Table 3.1: Details of Simulation of HP35 in MIL-101(Cr) water system. $\#_{\mathrm{MOF}}$, $\#_{\mathrm{Protein}}$, $\#_{\mathrm{Water}}$ and $\#_{\mathrm{Ions}}$ stand for number of MOF atoms , protein atoms, water molecules and ions. $\mathrm{Total}_{\mathrm{atoms}}$ and Prod. stand for total number of atoms and production run.

System name	Run type	$\#_{\mathrm{MOF}}$	$\#_{\mathrm{Protein}}$	$\#_{\mathrm{Water}}$	$\#_{\mathrm{Ions}}$	$\mathrm{Total}_{\mathrm{atoms}}$	Prod.
MOF	NVT	33184	0	36935	0	143989	5 ns
Protein-MOF	$\mathrm{Equil.}_{\mathrm{NVT}}$	33184	574	36629	1	143646	680 ns
Protein-MOF	US	33184	574	36629	1	143646	Table B.1

3.2.5 Prescription for performing translocation experiment in MOF

Initial configurations of the protein for translocation through the hexagonal window of the MOF were generated using Steered Molecular Dynamics (SMD) simulations. As a starting point, HP35 was placed at the center of one of the big cavities of the supercell. During equilibration, a harmonic potential was applied between the $\mathrm{P_{COM}}$ and $\mathrm{C_{COM}}$ along with a restraining potential for 'MOF and protein,' 'MOF and water,' and MOF alone, respectively, in the first three stages of the general protocol (described earlier). This was followed by an additional step of NVT run at 298 K, removing position restraints on the MOF. During the SMD runs, the non-hydrogen atoms of MOF were position-restrained. In these runs, a harmonic spring with a force constant of 10^5 $\mathrm{kJmol^{-1}nm^{-2}}$ attached to the $\mathrm{P_{COM}}$ was pulled with a speed of 0.02 nm/ns along a defined (direction) unit vector towards the hexagonal window. This direction vector was obtained as the cross-product of the hexagonal window's two edges (vectors). 15 SMD runs, each initiated with a different seed for generating initial random velocities of the atoms, were generated. The work profiles from each were examined, and the one displaying the lowest work value was chosen as the putative 'path' for the Umbrella Sampling runs.

Taking the initial positions and coordinates of the protein from the chosen SMD run, 54 umbrella windows that covered the distance from the center of the cavity to the hexagonal window (that connects two adjacent cavities of the MOF) were identified. A harmonic potential was applied to the distance between the $\mathrm{C_{COM}}$ and $\mathrm{P_{COM}}$, with a force constant of 10^5 $\mathrm{kJmol^{-1}nm^{-2}}$ at each of these positions. The non-hydrogen atoms of the MOF were position restrained for all the windows during umbrella sampling simulations with a force constant of 10^3 $\mathrm{kJmol^{-1}nm^{-2}}$. The run lengths in different umbrella windows are presented in Table B.1. The cumulative MD run length over all the Umbrella Sampling windows amounts to 6.54 μs.

At each window, the run length was chosen such that the amplitude of fluctuation of the backbone RMSD of the protein lay between 0.5 to 1.0 Å around a mean value for a duration of at least 5 ns. Thus, the last 5 ns of the molecular dynamics trajectory for each umbrella window was used for further analysis. This procedure ensured that the PMF value at each window was obtained from a converged orientation and conformation of the protein. The free energy surface was reconstructed by reweighting configurations from umbrella windows using the weighted histogram analysis method (WHAM) [21] implemented within gmx_mpi wham code. The number of bins for the histogram was 200 and for estimating error in the potential of mean force (PMF) or the free energy profile, Bayesian bootstrapping was carried out with 200 bootstraps. Table 3.1 presents a summary of the simulations performed.

3.3 Results and Discussion

3.3.1 HP35 is located near the surface of the MOF cavity and not at its center

Although this remains to be verified, an enzyme within the MOF is invariably assumed to be located at the center of the MOF cavity [188, 189]. To understand the behaviour of HP35@MIL-101(Cr), we first performed equilibrium MD simulations at 298 K. The initial location of the protein was with P_{COM} at the center of one of the larger cavities (Figure 3.2A). Even during the equilibration stage (Figure 3.2B inset), the protein moved around 4 Å from this location. Subsequently, this distance, $d_{MOF-Protein}$, reached a value of 7.5 Å in about 400 ns (Figure 3.2B), and maintained the same for a further duration of 280 ns. This observation demonstrates that the protein's equilibrium position is not the cavity's center but closer to the cavity surface of the MOF, as seen from Figure 3.2C.

Our analyses show that HP35 broadly maintains its native structure upon confinement. The root mean squared deviation of the positions of backbone atoms (N, CA, C) of the protein relative to S_{NMR} remains largely (77% of the complete trajectory) within 3 Å over 300 ns (see the 350-650 ns segment of Figure 3.2D). Marginal differences in the conformation of helix-1 of HP35@MOF with that of the reference structure (S_{NMR}) are seen (Figure 3.2F). Inter-molecular interactions between the MOF sites and HP35 were examined to identify the origin of the differences.

One of the ways to quantify the extent of van der Waals interactions between the protein and the MOF is to count the number of heavy atom-heavy atom contacts (i.e., non-hydrogen) that lie within a distance of 4 Å. This quantity, too, increased over the length of the equilibrium trajectory and saturated to a value of around 40 (Figure 3.2E inset). Furthermore, as seen from Figure 3.2E, residues of helix-1 and helix-2 interacted more often with the MOF than those of helix-3, as also those in the

non-helical region (N-TER, residues joining helix1 and helix2, and residues joining helix-2 and helix-3). Within helix-1 and helix-2, no difference in interaction with the MOF between polar and non-polar residues was observed. An overlay of the backbone atoms (N, CA, and C) of the protein in the last time frame of the production run with S_{NMR} is presented in Figure 3.2F. Along with a marginal reduction in the alignment of helix-1, a different orientation for the three F residues of the hydrophobic core is seen; however, the overall secondary structures and, in particular, helix-2 and helix-3 align well.

3.3.2 Existence of a "constriction region" in the protein translocation pathway

To understand the conformational heterogeneity of the protein inside the MOF cavity during translocation, we examined various geometrical parameters of the protein along its migration path from one cavity to the other. The progress of the protein on this path is captured herein from MD trajectories through the sequence of Umbrella Sampling windows. Figure B.5 provides milestones on this path, which are used in the present discussion. The window index increases as the protein moves from the center of the cavity towards the hexagonal aperture that connects two neighbouring MOF cavities. Here, we describe the conformations adopted by HP35 during its translocation.

For a few central umbrella windows (windows 18-29), the root mean square deviation of protein backbone from S_{NMR} (Figure 3.3) lies between 2-3.5 Å. Here, the protein explores its conformational space around its native state. The same is reflected in the plateau region of the helicity plot (Figure 3.3) wherein the alpha-helical content of the protein was the highest and has nearly the same value as in its native state, S_{NMR} (Blue horizontal dotted line). The medians of the solvent accessible surface area of the protein (Figure 3.4) remain close to the two specific SASA values (3056 and 3109 $Å^2$) corresponding to two folded conformations of HP35 in neat water reported in an earlier study [190]. The latter was described as a "dry molten globule" [191]. These two values are labelled in the SASA plot (Figure 3.4) with dashdotted and dotted lines in Blue, respectively. This central zone is termed as the "constriction region" in our study.

Before the constriction region, i.e., closer to the center of the cavity (windows 1-17), the helicity of the protein is reduced (Figure 3.3). The median values of the protein SASA display an undulated profile in this zone (Figure 3.4). Further, the deviation of the backbone atoms (RMSD) from S_{NMR} is within 2-5 Å (Figure 3.3). On the other hand, the number of non-hydrogen contacts between the protein and the MOF (Figure 3.4) displays a steady increase. As the protein is farther from the MOF surface in this zone, the former is more accessible to water molecules. Hence, both the facts – (i) the spontaneous tendency of the protein to have van der Waals contact with the cavity surface of the MOF and (ii) interaction of the confined water molecules with the surface of the protein results

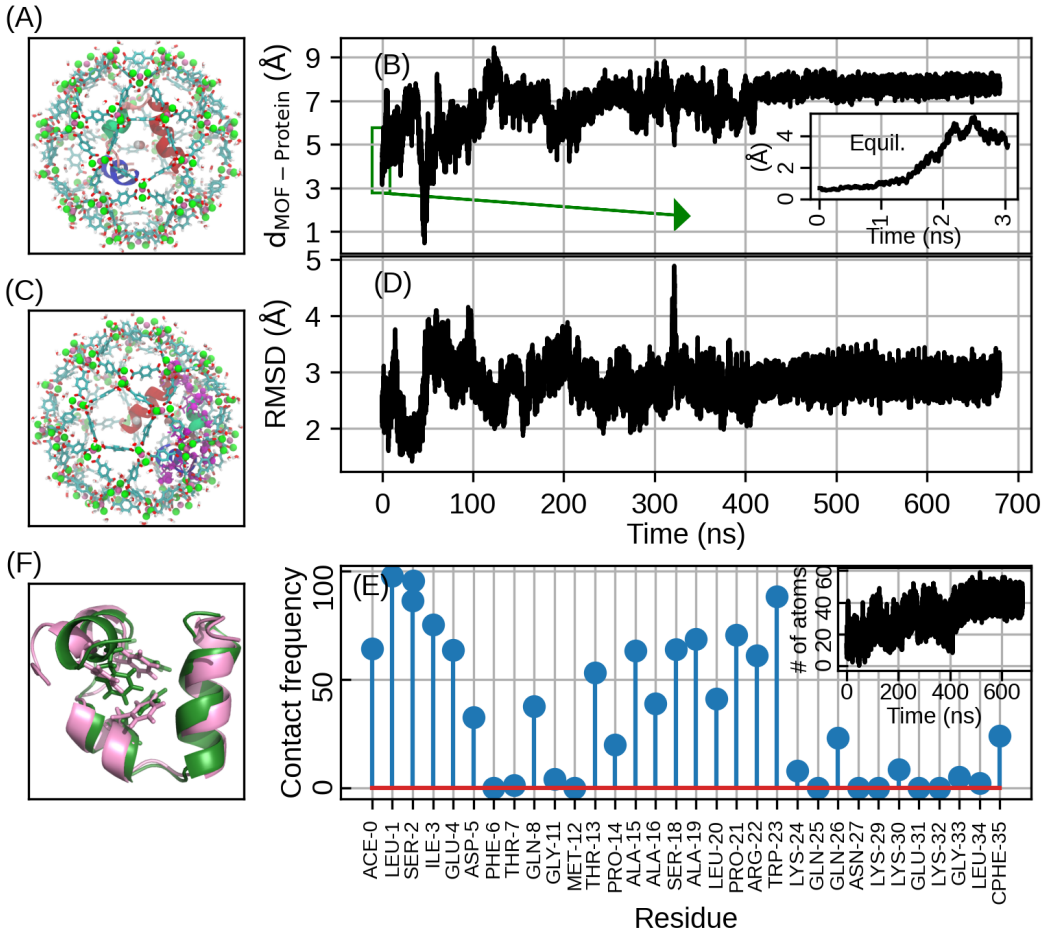


Figure 3.2: Results from the equilibrium MD simulation. (A) HP35 in its initial configuration inside one of the MOF cavities. HP35 is shown in the New Cartoon representation with helix-1, helix-2, and helix-3 in Blue, Green and Red, respectively. MOF: Cr atoms and μ3-Oxygens in Green and Mauve with vdW representation with reduced scale, metal-ligated waters, and organic ligands in Licorice representation. Water molecules filling the cavity are not shown for clarity. (B) Distance between the protein center of mass (COM) and the cavity center of the MOF. The same, but during the equilibration stage, is in inset. (C) Same as in (A) but for the last time frame of the equilibrium MD run. (D) Fraction of time spent by a residue (non-hydrogen atoms) within 4 Å of any MOF non-hydrogen atom. Inset: Total number of protein-MOF atom contacts vs simulation time. (E) Backbone RMSD with respect to solution NMR structure. (F) Overlay of NMR structure of the protein (Green) and that of the last time frame of the protein@MOF run. (Backbone atoms N, CA, C, and O have been used for alignment). The hydrophobic core is highlighted in the Licorice representation.

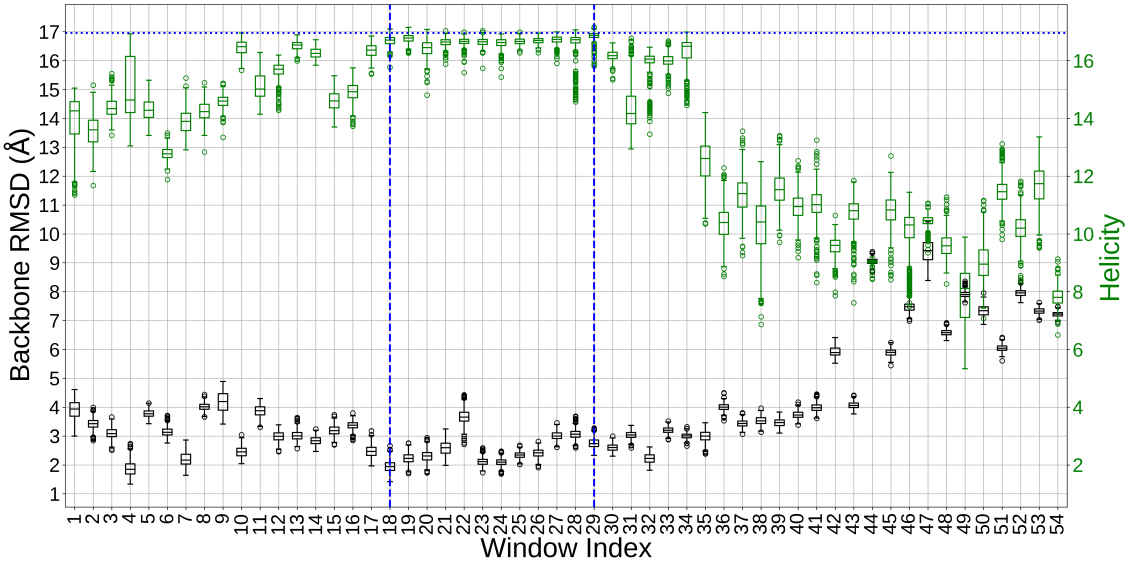


Figure 3.3: Black: RMSD of protein backbone with respect to S_{NMR} across Umbrella Sampling windows. Green: Alpha Helical content across windows. The horizontal dotted line in Blue is the value of helicity of S_{NMR}. The vertical dotted lines in Blue enclose the constriction region. The five number summary statistic, namely Box Plot [192–195] has been used to represent the distribution of different quantities across umbrella sampling windows.

in a zig-zag traversal path of the protein (Figure 3.5(A)). However, the protein did remain intact in its native state, as the medians of SASA of the hydrophobic core (Figure B.10) were close to that of the S_{NMR}.

Beyond the constriction region, i.e., towards the hexagonal window (windows 30-54), the helicity of the protein decreased drastically, while the RMSD of the backbone atoms and the SASA of the protein increased. The non-hydrogen contacts between the protein and the MOF atoms reached the highest value at the hexagonal window. In this region, there existed a sub-zone (approximately, windows 30-41) where the RMSD of the backbone atoms lay within 4 Å; the number of protein-MOF contacts was steady, while the helicity reduced and SASA increased. This region marks the initiation of protein unfolding due to its translocation. The unfolding exposes the hydrophobic core, which interacts with the non-polar linker groups of the MOF. Beyond this sub-zone, the protein accesses more extended structures, driven by the favourable van der Waals contacts with the MOF surface (SASA of the protein accessing values greater than 3700 $Å^2$).

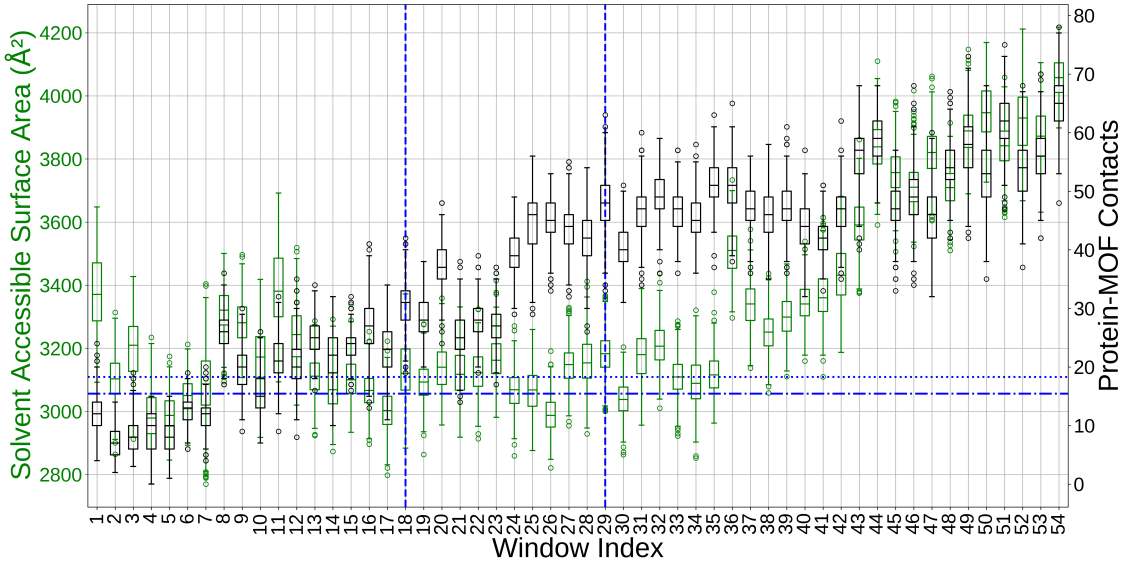


Figure 3.4: Black: Number of Protein-MOF non-hydrogen atom contacts (with a cut-off of 4 Å) across Umbrella Sampling windows. Solvent accessible surface area of the protein across windows is shown in Green (for the horizontal dashdotted and dotted lines in Blue, see text). The vertical dotted lines in Blue enclose the constriction region. The five-number summary statistic, namely Box Plot [192–195] has been used to represent the distribution of different quantities across umbrella sampling windows.

3.3.3 Potential of mean force reveals that unfolding of HP35 during translocation is regulated by both cage geometry and confined waters

The free energy profile of the system during the translocation process obtained through umbrella sampling simulations shows interesting behaviour. Here, the reaction coordinate is the distance between P_{COM} and C_{COM}. The potential of mean force (PMF) displays a minimum at a distance of around 9.5 Å, which lies within the constriction region (approximately 24^{th} window in umbrella sampling). In this region, the protein's conformations are influenced by solvent water molecules and by the atoms of the MOF framework. An optimal arrangement leads to the minimum in the PMF, where the free energy value with respect to the protein at the cavity center is around -9 kcal/mol. Notably, the distance from the cavity center where the PMF minimum is observed matches that obtained from the equilibrium MD simulations and thus serves as an internal consistency check between these two categories of MD runs. The PMF attains its highest value at the hexagonal window where HP35 is in the most extended conformation. With respect to its equilibrium position, the free energy barrier for the co-translocational unfolding of the protein is estimated to be 16 kcal/mol (Figure 3.5).

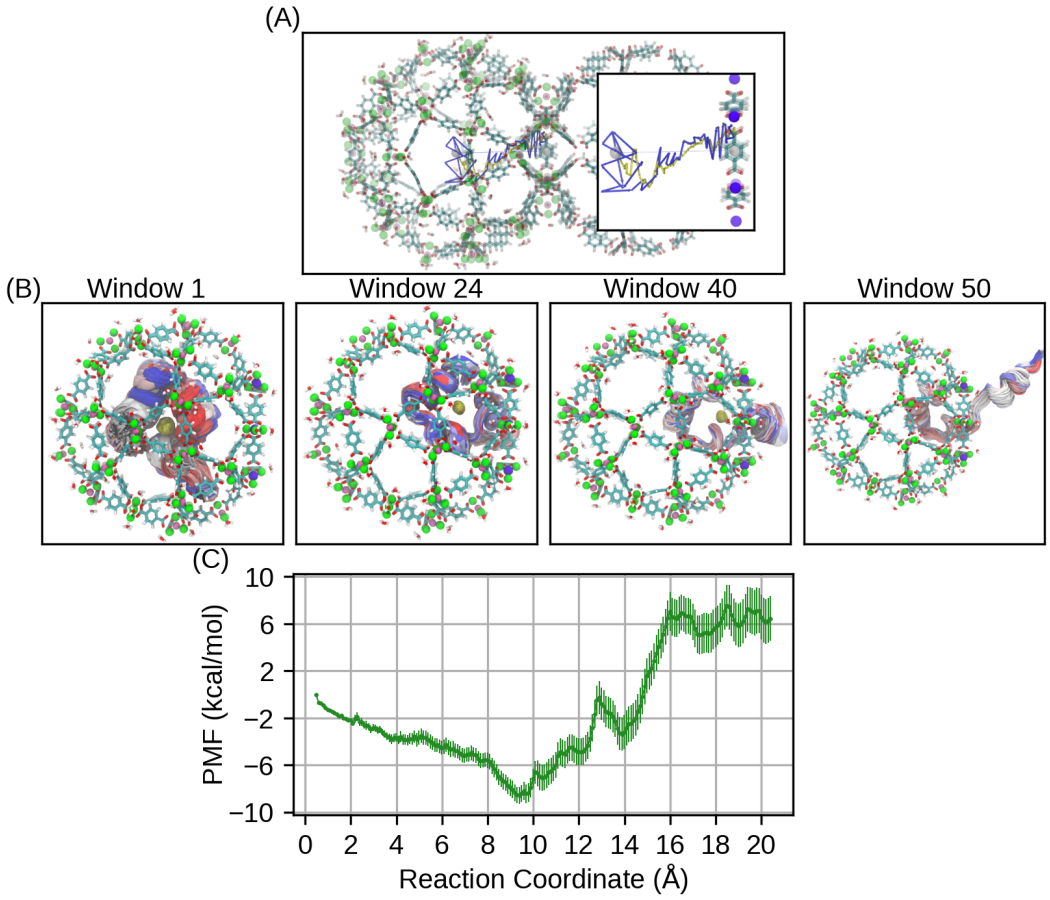


Figure 3.5: (A) Zig-zag excursions of the protein shown as Blue arrows obtained by concatenating the last time frame of all Umbrella Sampling windows. It provides a glimpse of the reaction coordinate. The path traversed during the SMD run (one-way trip) is shown with Yellow arrows. Inset: Zoom-in image of the path. Water molecules and ions are not shown for clarity [196]. (B) New Cartoon representation of the secondary structure of HP35 in a few Umbrella Sampling windows during its translocation. Red to Blue show structures with increasing time. (C) Free Energy profile for the translocation of HP35 from the center of the cavity of MIL-101(Cr) to the hexagonal aperture connecting the neighbouring cavity.

3.4 Conclusions

While the translocations of ions, molecules, and macromolecules across lipid membranes have been studied using MD simulations [197], that of a protein (or an enzyme) through a porous inorganic host such as a metal-organic framework has not been examined so far through computational methods, despite the vast amount of experimental reports on enzyme@MOF as functional biocatalytic platforms. This work addresses this problem and

provides considerable insights into the changes in the secondary structure of the protein as well as that in the free energy along the transport path. This ensemble of structures of the HP35 protein and their interactions with the surface of the MIL-101(Cr) MOF cavity allows us to draw a possible general mechanism for co-translocational unfolding. As shown in the schematic (Figure 3.6), the protein (enzyme) closely resembles its native state when it is present somewhere between the cavity center and the hexagonal window of the MOF (label "2" in Figure 3.6). This key result observed in our equilibrium MD simulations is further confirmed by a free energy minimum away from the cavity center, as seen in the umbrella sampling MD runs. At least concerning HP35@MIL-101(Cr), the center of the MOF cavity is *not* the equilibrium position for the protein. While the general applicability of this observation needs to be verified, it is likely to be followed by proteins that are smaller than the size of the cavity; after all, the macromolecule would prefer to interact with the atoms of the MOF, if possible, without unfolding, via dispersion interactions.

When restrained to stay at the center of the cavity, HP35 undergoes marginal changes in its conformation towards partially unfolded structures in such a manner as to interact with the MOF surface. While translocating through the aperture connecting two cavities, the protein is present in an extended form stabilized by van der Waals contacts with the surfaces of both the MOF cavities. Since the initial configurations for the umbrella sampling simulations performed here were chosen from one of the SMD runs in which helix-1 of the protein is first translocated through the hexagonal window, the same is seen in the US runs as well. However, in other SMD trajectories (not reported here), we observed no preference for any specific helix to cross through the hexagonal window of the MOF first.

The confinement of HP35 in the MOF enabled conformations of the protein that were different from those observed in neat liquid water. This was reflected in the SASA (Figure 3.4). Thus, the solvent water confined inside the MOF could access the protein surface to a greater extent, which was, in turn, made possible through increased protein-MOF direct contact. MIL-101(Cr) lacks groups that can hydrogen bond with the polar side chains of the HP35; thus, van der Waals is the dominant interaction type between the MOF and the protein. This interaction energy displayed a monotonic increase in magnitude (Figure B.12) as the protein approached the hexagonal window during its translocation, unlike the Coulombic interaction (Figure B.11). Our observation that van der Waals is the major interaction in the HP35@MIL-101(Cr) system aligns with earlier studies of biomolecules confined in MOF channels [112, 113].

The free energy barrier for the co-translocational unfolding of HP35 across the hexagonal window of two neighbouring MIL-101(Cr) MOF cavities is estimated to be 16 kcal/mol at ambient conditions. The study also enabled us to examine the partially unfolded

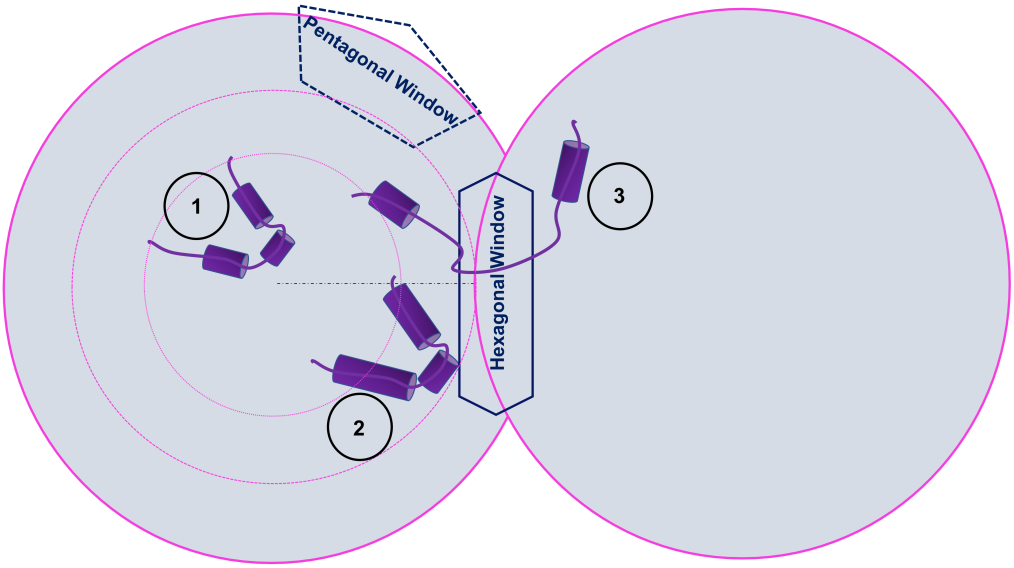


Figure 3.6: A schematic for co-translocational unfolding of protein inside MOF pores. In the context of HP35@MIL-101(Cr), when the protein is near the center of the cavity ("1" in the figure), its native structure is marginally affected due to the competing interactions with the internal surface of the cavity. Near the aperture joining neighbouring cavities, the protein accesses extended structures being partitioned between the two cavities as represented by label "3". In between these two regions (labelled "2" in the figure), the protein maintains its native ensemble of structures fairly well in what we have defined as the "constriction region."

structures of the protein. Simulations of other proteins with different topologies in such porous hybrid materials could further our understanding of unfolded intermediates present under confinement, mimicking living cell milieu.

4

Mechanism of Post-Translational Modification of Asp-109 to Succinimide in an archaeal enzyme, MjGATase

4.1 Introduction

In Chapter 1, we have seen that one of the backbone post-translational modifications (PTM) is the formation of succinimide. The spontaneous nonenzymatic deamidation of asparagine (Asn) and glutamine (Gln) residues to aspartic acid (Asp) and glutamic acid (Glu) residues, respectively, occur under physiological conditions [198–203] and during storage [204]. This is important in developing Alzheimer's disease [205, 206] and cataracts [207–209]. This phenomenon has also been associated with protein aging based on the half-times of deamidation reaction [210, 211]. Noah E. Robinson and Arthur B. Robinson hypothesized this deamidation of asparaginyl and glutaminyl residues as "molecular clocks" [212]. The deamidation of asparaginyl residues is suggested to proceed through succinimide formation [213, 214], which has been studied through experiments [213–221] and computations - through structure-based prediction [222, 223],

The work presented in this chapter was performed in collaboration with Prof. Hemalatha Balaram's group, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India and Prof. Padmanabhan Balaram, National Centre for Biological Sciences, Bangalore, India and valuable discussions with Prof. A Padmesh's group, Indian Institute of Technology, Palakkad, India, and Dr. Sudip Das, Postdoc, Italian Institute of Technology, Genoa, Italy.

using classical molecular dynamics through different classical descriptors - like mean Solvent Accessible Surface Area (SASA) of Asn, C_{γ}-N_{n+1} distance, root mean square fluctuations (RMSF), dihedral angles of backbone and side chain (χ and ψ) and hydrogen bond analysis [224–229] and using quantum mechanical calculations [230–239]. In short, (i) the cyclization happens in mainly two stages - a) deprotonation of n+1 backbone hydrogen and b) cyclization through a covalent bond formation between n+1 backbone nitrogen and side chain carbonyl C of Asn residue at n-th position, (ii) it was proved that the deamidation mechanism was the same for a protein (ribonuclease A) and two model peptides [214], and (iii) the activation barrier for succinimide formation was around 22 kcal/mol (at pH 5-7.5, using Arrhenius equation, based on studies on hexapeptides [240]) from experiment and 20-21 kcal/mol (referring to either the deprotonation or cyclization step) from computations (peptide level studies [239]).

Because of the spontaneity in hydrolysis, stable succinimides are rare. There are two earlier studies reporting stable succinimides in hyperthermophilic archaea - one from *Pyrococcus horikoshii* (https://www.rcsb.org/structure/1WL8) and another from *Methanocaldococcus jannaschii* [124] and in lasso peptides [241]. Through mass spectrometry, our experimental collaborators (Prof. Hemalatha Balaram's group, JNCASR, India) established the presence of the stable succinimide at position 109 in glutamine amidotransferase, a subunit of GMP synthetase in *Methanocaldococcus jannaschii*. This stable succinimide induced remarkable thermal stability on the wild type (WT) protein (resisted unfolding even up to 100° C, mutants melted at appreciably lower temperatures) [124]. Further, the X-ray crystal structure of MjGATase at a resolution of 1.67 Å showed a clear electron density for succinimide at 109^{th} residue. MD simulations coupled with enhanced sampling techniques for the WT crystal structure and two single-point mutants gave more molecular-level insights on the stability induction by succinimide ring. It has been found that the residues preceding and succeeding succinimide played significant roles in protecting succinimide from hydrolysis, even at high temperatures. A set of hydrogen bonds imposed a local "conformational lock" in the succinimide-containing loop of the enzyme and van der Waals interactions involving succinimide was reasoned out for imparting the high stability [125].

Now, the question that arises is how Asn-109 alone turns into the stable succinimide, leaving nine other Asn residues in this enzyme intact. Do the nearby residues play any role? To understand this aspect, our experimental collaborators expressed mutated structures (8 single mutants and two double mutants) of the WT. It was found out from two double mutants (D110V-K151L and Y158F-K151L) and two single mutants (D110P and Y158F) that the 110^{th} residue (n+1, Asp), 158^{th} residue (Tyr) and 151^{th} residue (Lys) might have roles in the formation of succinimide at 109^{th} residue (summary in Table 4.1). This chapter investigates the reaction through QM/MM MD simulations with

non-tempered metadynamics (MTD). Herein, we examine a concerted deprotonation-cyclization mechanism, considering residues that are 'likely' to participate in the PTM within the quantum region, while the remaining residues and water molecules are present in the MM region. We estimated the activation barrier for deprotonation to be about 3.4 kcal/mol followed by a barrierless cyclization.

Table 4.1: Experimental results from Prof. Hemalatha Balaram's group, JNCASR (Chandrashekarmath et al. unpublished, reproduced with permission.)

Mutant	N109 Intact (%)	SNN Formed (%)	D/isoD Formed
D110P	100	-	-
Y158F	76	24	-
D110V-K151L	87	13	-
Y158F-K151L	90	10	-

4.2 Computational details

The Wild Type (WT) enzyme contains the product of post-translational modification, i.e., succinimide (SNN) in the 109^{th} position (PDB ID: 7d40 [125]). However, since we are interested in the mechanism of the PTM from the reactant state (wherein Asn is in the 109^{th} position), we had to generate the initial configuration through in-silico mutation (using PyMOL [242], from Chain A,7d40).

We used GROMACS software [243] for classical molecular dynamics (CMD) simulations and CP2K software [144] patched with PLUMED [244, 245] for QM/MM MD simulations. The protein was placed in a periodic cubic box with 20 Å periodic image distance (box length 72 Å) and solvated using the GROMACS solvation module. Protonation states of ionizable residues were set through the GROMACS default program. Three sodium ions were added to neutralize the complete system. Periodic boundary conditions were applied in all three directions. For maintaining temperature and pressure, we used the Bussi-Donadio-Parrinello thermostat [11] and the Parrinello-Rahman barostat [12], respectively. Time constants for temperature coupling at 298 K were 0.5 ps and 100 fs, respectively, for classical MD and QM/MM MD simulations. Pressure coupling was used only for classical MD runs during equilibration. The pressure was maintained at 1 bar isotropically (isothermal compressibility: $4.5\text{x}10^{-5}$ bar^{-1}) with a 1.0 ps time constant. Real space cut-offs for the Lennard-Jones (LJ) and Coulomb interactions were set to 10

Å, and 14 Å for AMBER and GROMOS, respectively. Lorentz-Berthelot mixing rules were used to obtain the LJ parameters between two atom types. The particle Mesh Ewald (PME) [28] method with an interpolation order of 4 and a relative tolerance of 10^{-5} was used to calculate the electrostatic interactions for distances above 10 Å for classical MD calculations in GROMACS. A Smooth particle Mesh Ewald (SPME) [246] method with an interpolation order of 6 and a relative tolerance of 10^{-6} was used to calculate the electrostatic interactions for distances above 10 Å for MM calculations in CP2K. Time steps for integrating the equations of motion for classical molecular dynamics simulations and QM/MM MD simulations were 1 and 0.5 fs, respectively.

Position restraining was done using a harmonic potential with a force constant of 10^3 kJmol^{-1}nm^{-2} for the non-hydrogen atoms. All bonds were constrained using LINCS with order 4 and warn angle 30°. During analysis, the criteria for hydrogen bonding was < 3.5 Å as donor-acceptor distance and $> 140°$ as donor-H-acceptor angle.

4.2.1 Classical Molecular Dynamics

We carried out classical MD simulations for the system in two steps: at first using GROMOS54A7 [247] and later with AMBER99sb-star-ILDN [182, 183, 248] force fields. Two different force fields were used because using AMBER force field alone, we could not obtain the right set of backbone dihedral angles for residue 109 (SNN being located in the fourth quadrant of the Ramachandran plot). After employing GROMOS54A7 united atom force field (with SPC/E water model [10], as SNN had earlier been treated with united atom parameters [125]), the backbone dihedral angles of residue 109 were properly sampled (in the second quadrant of the Ramachandran plot). The non-hydrogen protein coordinates from the 100$^{\text{th}}$ ns frame of the GROMOS trajectory was used as the initial conformation for subsequent simulations with the all-atom AMBER family of forcefield. AMBER99sb-star-ILDN and TIP3P [9] parameters were used for protein and water molecules, respectively. Leap-frog algorithm was used for solving Newton's equations of motion. The complete system is shown in Figure 4.1. The general protocol for equilibration using both forcefields consisted of four steps: a) Energy minimization of water molecules (position restraining the protein), b) Energy minimization of protein atoms (position restraining water molecules), c) NVT equilibration by raising the temperature from 0 to 298 K over 500 ps and keeping at constant 298 K for the next 500 ps, and d) NPT equilibration at 298 K and 1 atm pressure over 1 ns. Following this, a production run was done in the NVT ensemble at 298 K for 100 ns. We extracted three arbitrary protein conformations by looking at the hydrogen bonding network around 109$^{\text{th}}$ residue from the AMBER production trajectory. These snapshots were used to carry out QM/MM MD simulations (Figure 4.2).

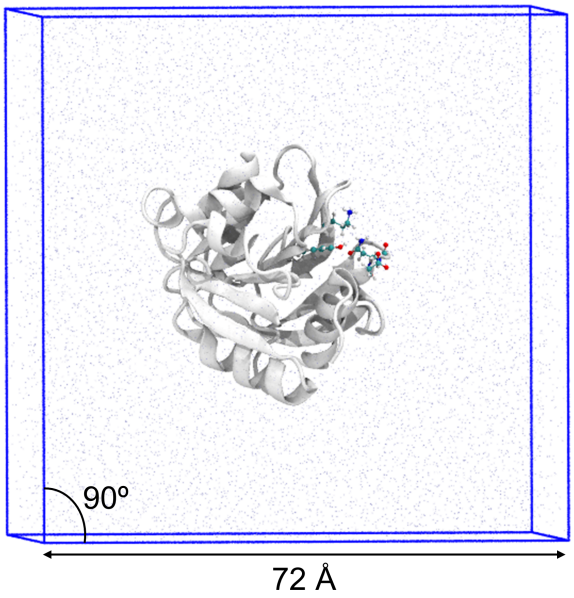


Figure 4.1: The complete system under study. Protein is shown in the New Cartoon representation in White. QM region is highlighted in CPK representation. Water molecules are shown in Iceblue dots. Ions are not shown clarity.

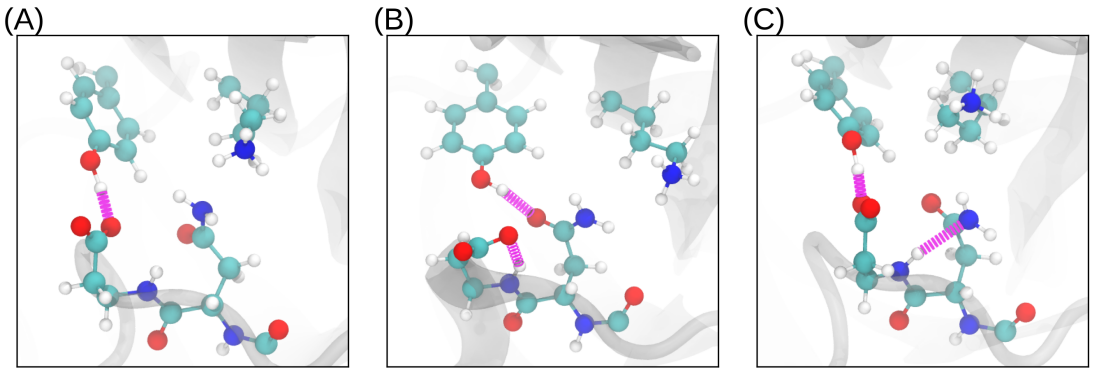


Figure 4.2: (A) to (C) represent the initial protein conformations used for QM/MM MD simulations. Atoms within the quantum region are shown in CPK representation. The remaining protein structure is (partly) shown in White Transparent mode with New Cartoon representation. Water molecules and ions are not shown for clarity. Hydrogen bonds are shown as springs in Magenta.

4.2.2 QM/MM Molecular Dynamics

As mentioned in Chapter 1, treating the complete system with QM description would have been highly computationally expensive for this large system and hence, only residues relevant to the reaction (suggested from experiments) were considered at the

QM level and others at the MM level. Thus, the QM level contained residues 109 (Asn), 110 (Asp), 151 (Lys), and 158 (Tyr). To make sure that we were not cutting through any polarizable bonds, part of residue 108 (backbone carbonyl- of Glu) were included in the QM region. The total number of quantum atoms was thus 61 (Figure 4.3) that included four link atoms (Hydrogen atoms) capping the QM region. The quantum box spanned 15 Å in three directions. For running the QM/MM MD simulations, QUICKSTEP [143] and FIST modules from CP2K were used. The quantum region was described at the DFT level with Perdew-Burke-Ernzerhof (PBE) functional. A double-ζvalence-polarized (DZVP) basis set was used along with Goedecker-Teter-Hutter pseudopotentials (DZVP-GTH-PBE) [145]. The plane wave cut-off was set at 300 Ry, and DFT-D3 dispersion correction [147] was applied. The MM region was described using the same force field used for classical MD simulations. The system was equilibrated through mechanical and electrostatic embedding schemes. The reaction was studied using an enhanced sampling technique (non-tempered metadynamics) with two reaction coordinates and the electrostatic embedding framework to account for the electrostatic interactions between QM and MM regions.

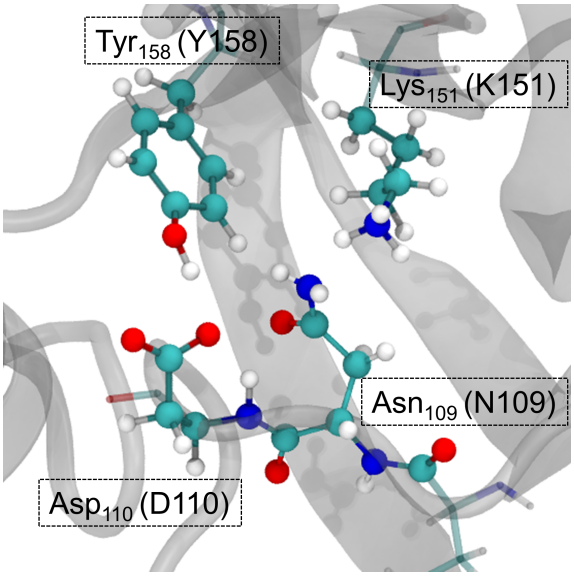


Figure 4.3: The atoms shown in CPK representation are part of the QM region. The complete residues (of which they are part of) are shown in Licorice representation. The remaining protein is (partly) shown in White Transparent mode.

A summary of simulations is tabulated in Tables 4.2, and 4.3. Each of our two collective variables was a combination (using PLUMED function COMBINE, https://www.plumed.org/doc-v2.9/user-doc/html/_c_

o_m_b_i_n_e.html) of two coordination numbers (using PLUMED function COORDINATION, https://www.plumed.org/doc-v2.9/user-doc/html/_c_o_o_r_d_i_n_a_t_i_o_n.html) (Figure 4.4). CV_1 is the combination of the number of contacts between "atom 1 and atom 2" and "atom 2 and atom 3". CV_2 is the combination of the number of contacts between "atom 3 and atom 4" and "atom 1 and atom 4". R_0 value used for the two sets of coordination numbers defining CV_1, i.e., (1, 2) and (2, 3) was 1.34 Å each, respectively. And 1.00 Å each was used as R_0 for the next set of coordination numbers defining CV_2, i.e., (3, 4) and (1, 4). Default values were used for other parameters. So, as per the definition, the values for the combined coordination numbers or collective variables of the reactant and product states were (-0.5, -0.5) and (+0.5, +0.5), respectively (Figure 4.4). Only for the last QM/MM run (Run 7), we used an upper wall bias (https://www.plumed.org/doc-v2.9/user-doc/html/_u_p_p_e_r__w_a_l_l_s.html) between the two nitrogen atoms (side chain amide-N of Asn_{109} and side chain amino-N of Lys_{151}, labelled as atom 3 and atom 5) at 5.5 Å with a force constant of 200 kcal/mol (remaining parameters with default values) as shown in Figure 4.4.

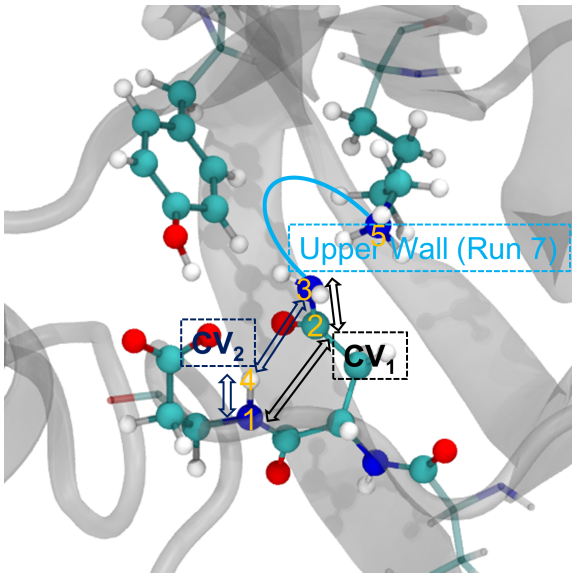


Figure 4.4: Definition of the collective variables used for non-tempered metadynamics. Five atoms are highlighted in Yellow. Atoms 1, 2, and 3 are part of CV_1 and atoms 1, 3 and 4 are part of CV_2. An upper wall (only for Run 7, Table 4.3) is applied between atoms 3 and 5. More explanation is provided in the text.

Table 4.2: Simulation details for classical molecular dynamics (CMD) and QM/MM MD runs. ($\#_{\mathrm{Pro-MM}}$, $\#_{\mathrm{Pro-QM}}$, $\#_{\mathrm{W}}$, $\#_{\mathrm{Ions}}$, $\#_{\mathrm{Total}}$ and Prod. stand for number of protein atoms in MM region, number of protein atoms in QM region, number of water molecules, number of ions, total number of atoms and production run length, respectively. System names, Asn and SNN mean protein structure containing asparagine and succinimide at 109^{th} residue, respectively.)

System	Simulation	$\#_{\mathrm{Pro-MM}}$	$\#_{\mathrm{Pro-QM}}$	$\#_{\mathrm{W}}$	$\#_{\mathrm{Ions}}$	$\#_{\mathrm{Total}}$	Prod. (ns)
Asn	CMD	2973	-	11345	3	37011	100
SNN	CMD	1880	-	12777	3	40214	100
Asn	QM/MM	2916	57	11345	3	37011	-

Table 4.3: Run details for QM/MM metadynamics MD simulations.

Run	σ (CV_1)	σ (CV_2)	Height (kcal/mol)	Pace (fs)
1	0.15	0.15	2	25
2	0.15	0.15	2	25
3	0.15	0.15	2	25
4	0.15	0.15	0.59	25
5	0.15	0.15	0.59	50
6	0.15	0.15	0.59	75
7	0.005	0.006	0.59	100

4.3 Results and Discussions

At first, we focus on runs 1 to 6 from Table 4.3. In all the simulations, we could observe the transformation only once from the reactant state to the product state (Figure 4.5). CV_2 advanced in time than CV_1 i.e., the backbone hydrogen of 110^{th} residue was abstracted by the side chain -NH_2 of 109^{th} residue prior to the cyclization step. In the reactant basin, there were many instances where CV_2 attained the value zero. From the definition, this implies that the backbone hydrogen is neither attached to backbone nitrogen of 110^{th} residue nor to the side chain -NH_2 of 109^{th} residue.

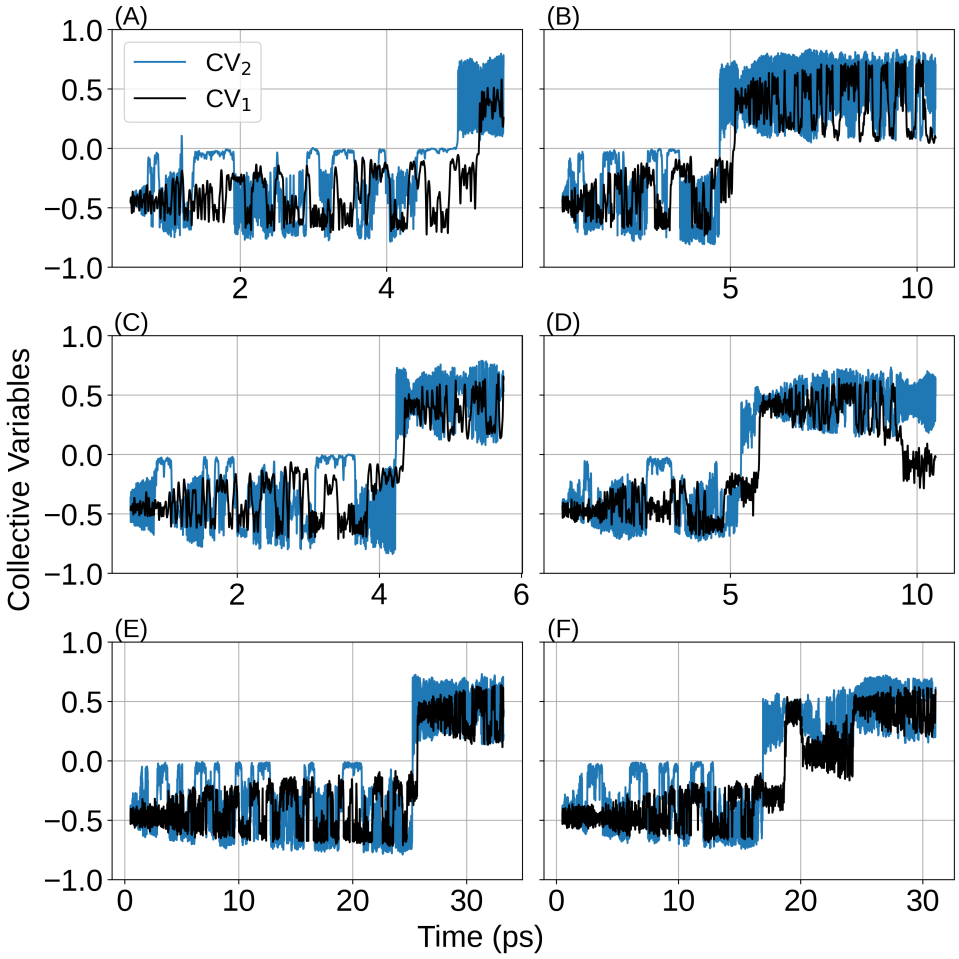


Figure 4.5: Time evolution of the two collective variables used to drive the reaction. The lengths of the simulations are different for different runs. (A) to (F) represent first six QM/MM MD runs (Table 4.3) sequentially, (A) being 1 and (F) being 6.

Hence we further explored four possibilities for this backbone hydrogen abstraction by nearby electron-rich centers. These are: (i) side chain carboxylate- of 110^{th} residue, (ii) side chain hydroxyl- of 158^{th} residue, (iii) side chain amide-O of 109^{th} residue, and (iv) side chain amide-N of 109^{th} residue. As the fourth possibility denotes conformations related to the product basin (as per the definition of CV_2), the value zero in the reactant basin signifies proton abstraction by any of the remaining three possibilities.

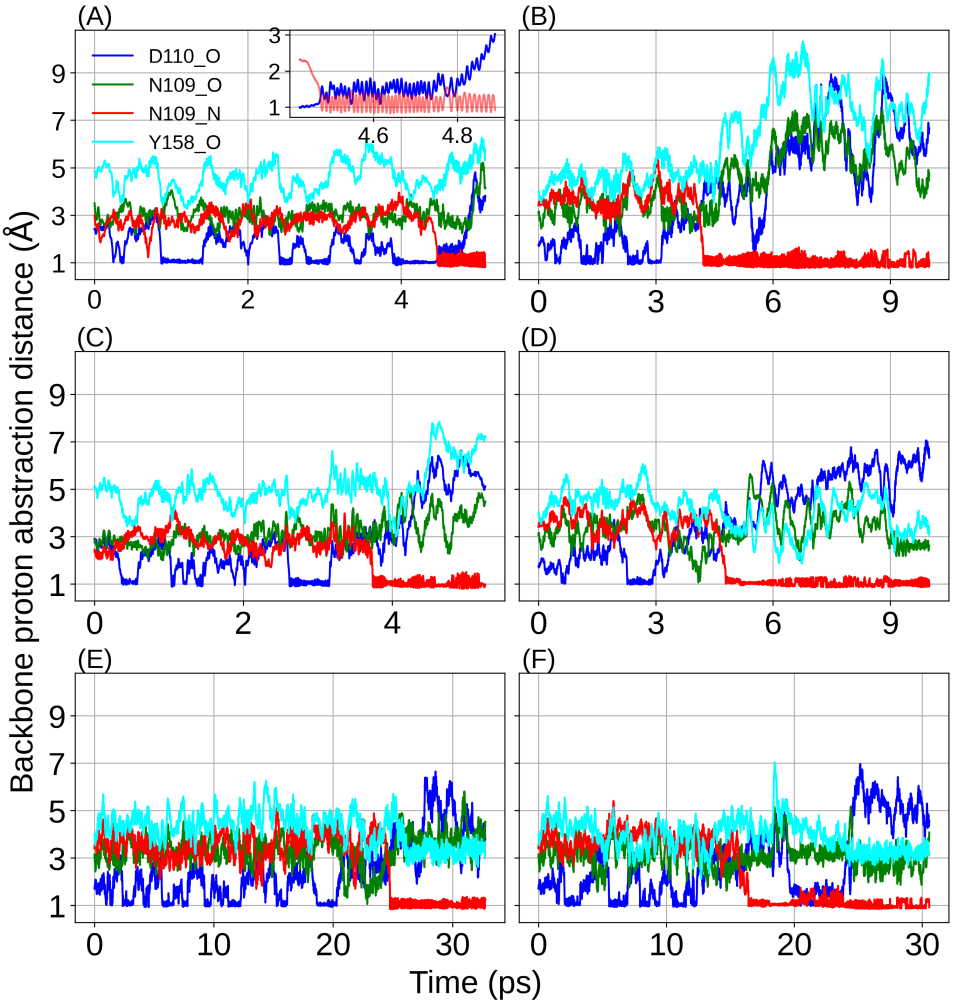


Figure 4.6: Backbone hydrogen abstraction by nearby nucleophilic centers with time. The Y-axis is the acceptor-H distance. (A) to (F) denote data from first six runs (Table 4.3). (A) has an inset figure which zooms into the proton abstraction by the side chain carboxylate-of 110^{th} residue and the side chain primary amino group of 109^{th} residue around 4.5 ps, denoting an exchange. Herein, the backbone proton was first abstracted by carboxylate-of 110^{th} residue and then delivered to the primary amino group of 109^{th} residue. Inset shares the same units for axes as the main graph.

In all the runs, possibility (i) and (iv) outnumbered the other two. In one case, there was an exchange of the proton between them (Panel (A) in Figure 4.6). In two of the runs ((D) and (E) in Figure 4.6), the amide-O of 109^{th} residue grabbed the proton for a couple of instances. Hence, the zero values of CV_2 in Figure 4.5 corresponds to the proton abstraction by side chain COO^- of 110^{th} residue.

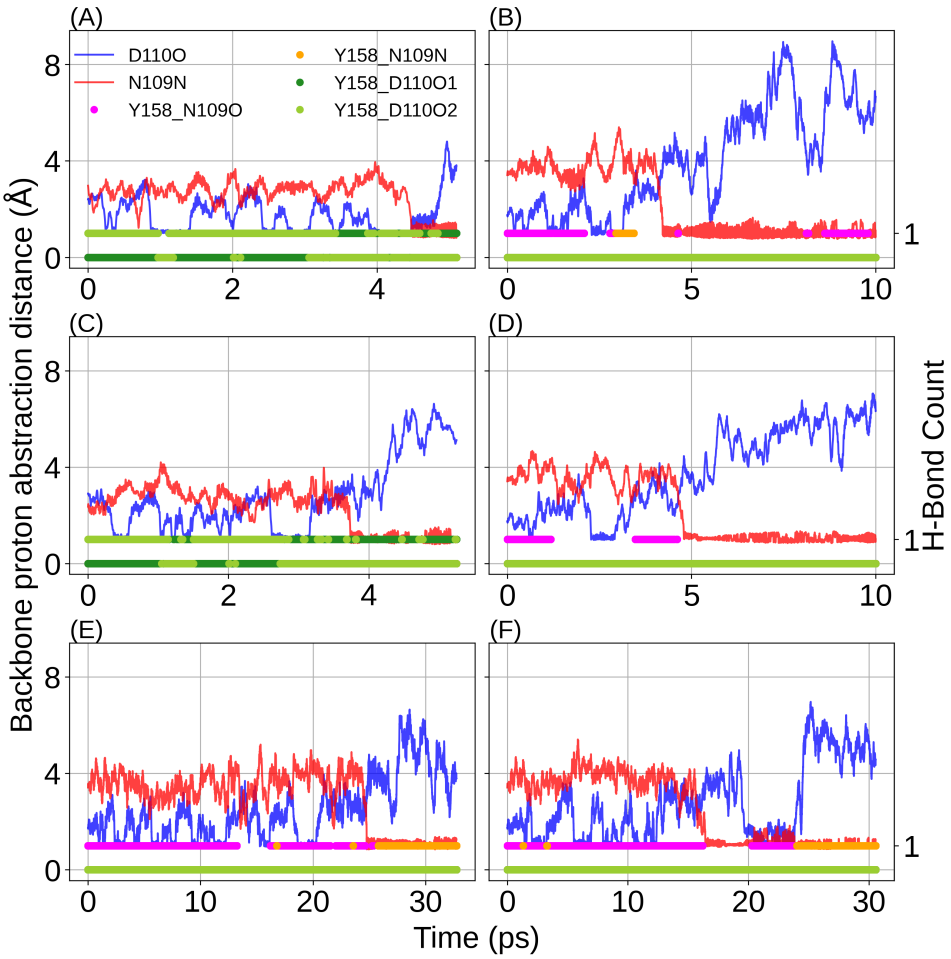


Figure 4.7: Hydrogen bonds with Y158 are shown in dots ((A) to (F) denote data from first six runs (Table 4.3). If the geometric criteria for hydrogen bonding are satisfied, the value is shown as 1; otherwise, it is shown as 0. Hydrogen bonding to either of the oxygens of side chain carboxylate- of residue 110 are denoted as Y158_D110O1 and Y158_D110O2, respectively. Hydrogen bonding to side chain nitrogen or oxygen of residue 109 are denoted by Y158_N109N and Y158_N109O, respectively. The backbone hydrogen abstraction by either the side chain carboxylate- of 110^{th} residue or the side chain primary amino group of 109^{th} residue are also shown in Blue or Red lines, respectively for all the six runs (D110O and N109N).

Experiments by our collaborators suggested that there could be a role of residue 158 for the reaction (Table 4.1). Hence, we examined the possibility of hydrogen bonding of 158^{th} residue as a donor with nearby nucleophilic acceptor sites (Figure 4.7). The hydrogen bond formation (denoted by R.H.S. Y-axis value 1) is not correlated to proton abstraction by the side chains (denoted by L.H.S. Y-axis value 1 Å). As the simultaneous

occurrence of value 1 for a finite stretch of time was missing, we could not able to decipher any direct role of the side chain hydroxyl- of Tyrosine-158 towards succinimide formation.

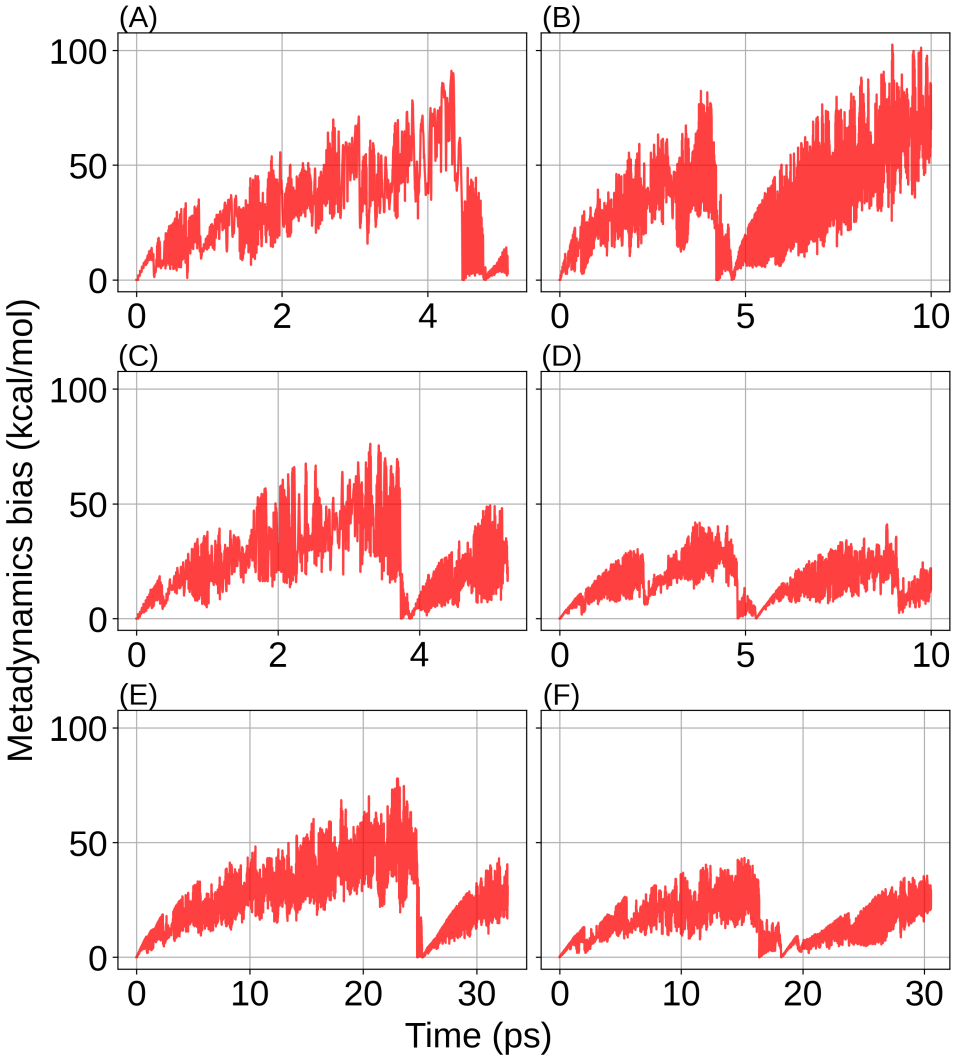


Figure 4.8: (A) to (F) represent time evolution of the deposited Gaussian hills during the first six QM/MM MD runs (Table 4.3).

As introduced in Chapter 1, the philosophy of metadynamics is to deposit repulsive Gaussians on the free energy landscape defined by the reaction coordinates. Once a basin is filled up, the walker fills the next basin. This is reflected in the bias plots through the build-up of the metadynamics bias, followed by a fall-off to zero value (Figure 4.8). As none of the simulations displayed the backward transition (i.e., from product state to

reactant state), we decided to obtain an estimate for the depth of the reactant basin or the activation barrier for the reaction from the reactant state (since that is completely filled up) instead. We have two observations from all the six runs - (i) the activation barrier heights are rather high (for a spontaneous autocatalyzed reaction), and (ii) the range of the values are also large, i.e., 40-75 kcal/mol (Figure 4.8).

Hence, we reexamined the metadynamics parameters and chose more conservative values for a seventh QM/MM-MD-MTD run. We added the Gaussians hills more finely, i.e., with a much smaller width and a larger time lapse between deposition, in the same CV space. In addition, we had to introduce an upper wall to avoid sampling non-interested regions of the reaction space. Here too, we could observe the reactant to product transition once (Figure 4.9); however, the cumulative energy of the deposited hills to fill up the reactant basin was much lower than those observed for earlier runs (Figure 4.10).

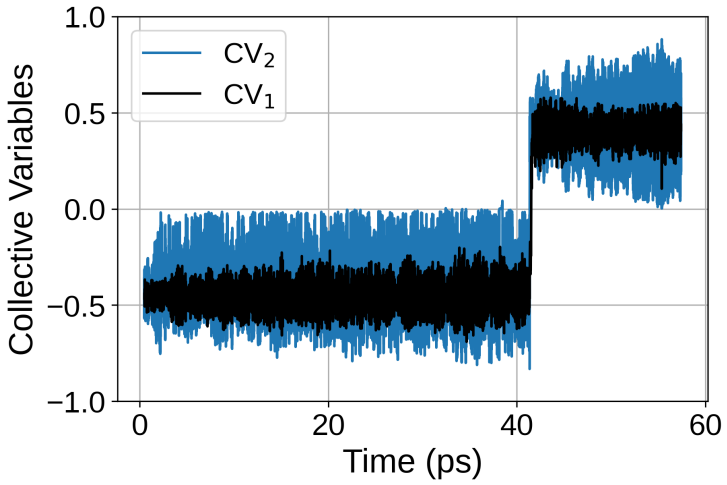


Figure 4.9: Time evolution of the two collective variables studied with conservative MTD bias parameters (Run 7, Table 4.3).

The earlier observations were found to be intact for this seventh run as well; for instance, the abstraction of the backbone proton of the 110^{th} residue occurred largely by the side chain carboxylate- of residue 110 and the side chain amide-N of residue 109. Further, there was no discernible assistance from 158^{th} residue through hydrogen bonding (Figure 4.11), which too aligned with observations from the earlier six runs.

To estimate the activation barrier from the reactant basin, at first, we summed up all the deposited hills (with the correction for the upper wall bias). This process presented a partial free energy surface. Then, we extracted a minimum free energy path joining the reactant state (represented by, coordinate (-0.4,-0.4) in the CV space) to an arbitrary

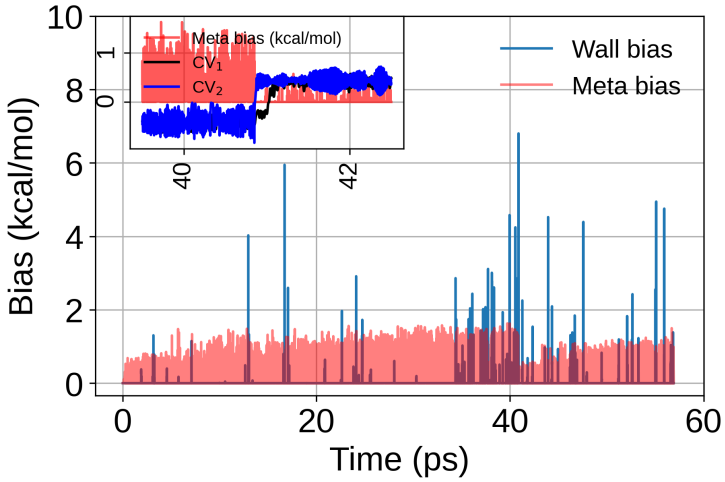


Figure 4.10: Metadynamics bias and upper wall bias for the QM/MM-MTD-MD run with conservative parameters for the repulsive Gaussians (Run 7, Table 4.3). Inset shows a zoomed portion near 40 ps along with the two collective variables. Inset axes share the same units as the main graph.

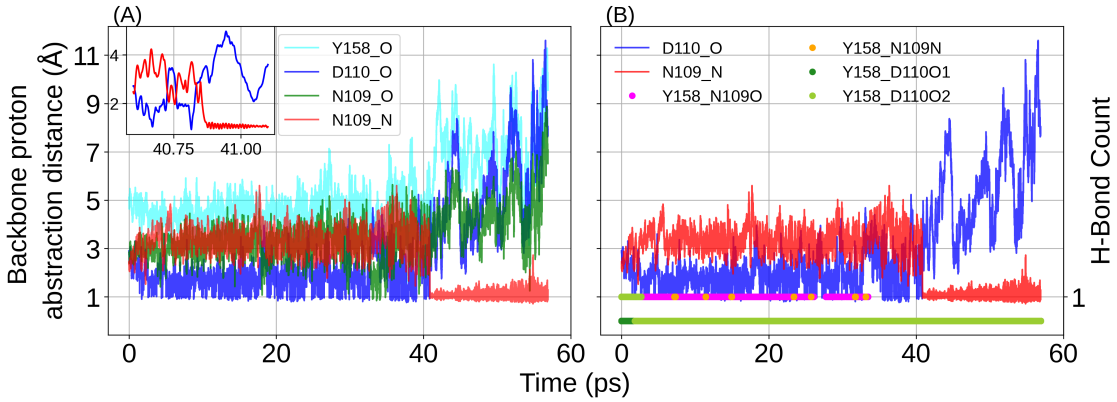


Figure 4.11: (A) Results from Run 7 (Table 4.3) of QM/MM-MTD-MD. Backbone hydrogen abstraction by different nucleophilic sites. The inset shows the zoomed region around 40 ps for proton abstraction by side chain carboxylate- of residue 110 and side chain primary amino group of residue 109. Inset shares the same units for axes as the main graph. (B) Hydrogen bonding through 158$^{\text{th}}$ residue (acting as a donor) along with proton abstraction by side chains of either residue 110 or residue 109.

point (represented by, coordinate (0.0,+0.4) in the CV space) on the free energy surface (Figure 4.12). On the other hand, from the time evolution of the metadynamics bias, we zoomed into those time points where the bias values reached zero and got the corresponding values for the collective variables (Figure 4.13). Uniting these two pictures

(Figure 4.12 (A) and Figure 4.13), we found that the estimated surface resembled the true free energy surface up to Point 2. Point 1 provided us with an estimate for the activation barrier from the reactant basin. The activation barrier was thus estimated to be about 3.4 kcal/mol. Further, we observed that for going from Point 2 to Point 4 (product basin), the addition of the bias was lesser than 0.5 kcal/mol.

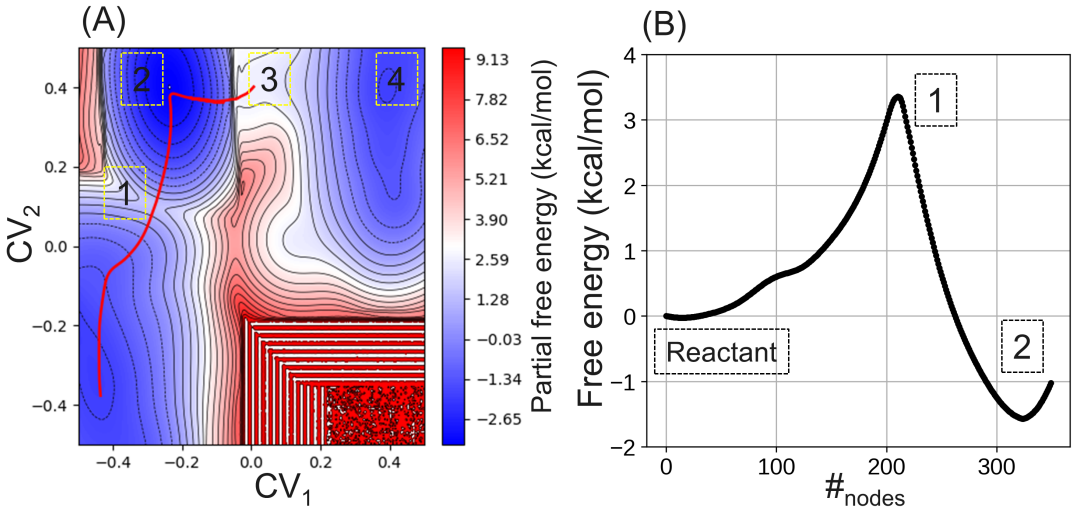


Figure 4.12: (A) Partial free energy surface for the succinimide formation reaction. For regions having free energy values greater than 9.5 kcal/mol, we have used the colour of 9.5 kcal/mol. The minimum free energy path starting from the reactant basin is shown in Red. Four states on the free energy surface are highlighted in Yellow boxes (numbered 1 to 4). (B) The free energy profile along the minimum free energy path (only the relevant fragment). Reactant state, Point 1, and Point 2, have been highlighted in dashed Black boxes.

Finally, we extracted the conformations corresponding to Points 1 - 4 of Figure 4.12 or Figure 4.13 (Figure 4.14). Point 1 corresponds to the structure where the backbone hydrogen (of 110^{th} residue) was detached from the backbone nitrogen (of 110^{th} residue) and near the nitrogen nuclei (amide-N of the side chain of 109^{th} residue). Point 2 corresponds to the structure where ammonia has formed but is still attached to the amide-C of the 109^{th} residue. Point 3 corresponds to the partial cyclization and finally deamination leads towards reaction completion at Point 4. Thus, the 3.4 kcal/mol activation energy corresponds to the deprotonation barrier. Furthermore, from the addition of bias (Figure 4.13), we posit that the cyclization step is likely to be barrierless.

From the handful of quantum mechanics/molecular mechanics (QM/MM) studies and Perturbed Matrix Method (PMM) with peptide sequences (dipeptides, hexapeptides,

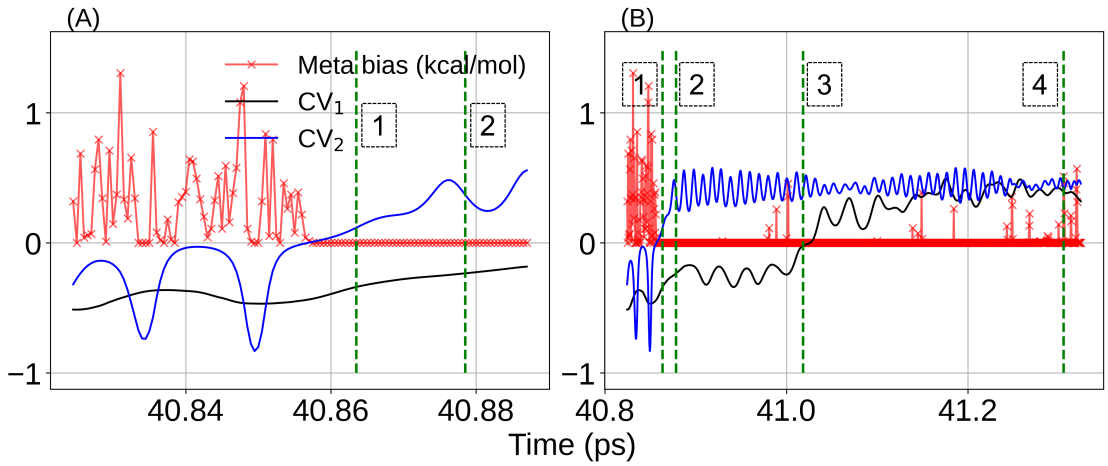


Figure 4.13: (A) Time evolution of the metadynamics bias and the two collective variables (zoomed into a time zone where the bias fills up the reactant basin. (B) Time advancement of the same as in (A) towards the product basin.

etc.), it has been found that the energy barrier for the succinimide formation reaction (Asn –> SNN) was higher than 30-40 kcal/mol [225, 226, 235, 238, 249] or about 20-25 kcal/mol [239, 250] (in agreement with the experimental estimate from peptide sequences [201, 221]) without and with treating water explicitly, respectively. In contrast to the peptide level studies, in the MjGATase enzyme, our experimental collaborators observe the reaction happening very fast at laboratory temperature. Thus, the activation barrier is accessible with the available thermal energy (k_BT, T=298K). So we believe that the predicted value (3.4 kcal/mol) would be a reasonable one.

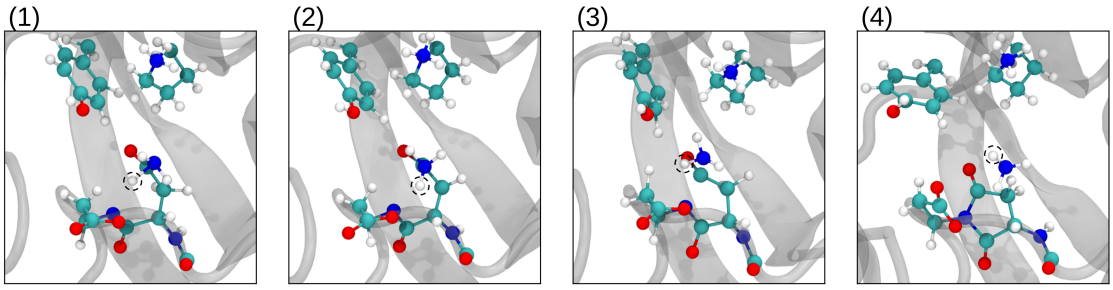


Figure 4.14: Observing succinimide formation. (1) to (4) corresponds to Point 1 to Point 4 of the free energy surface shown in Figure 4.12. The backbone hydrogen of D110, which migrated to -NH2 of the side chain of N109, is encircled with a dashed Black line.

4.4 Conclusions

Before we summarize our results, a brief perspective on the limitations of QM/MM formalism can perhaps be made here. Within the QM/MM simulations, there are two branches - without and with dynamics at the MM region, known as QM/MM, and QM/MM MD, respectively. In the present chapter, we have used QM/MM MD simulations. While there are some common limitations across both formulations, additional ones are present in QM/MM MD from the MD part. According to this methodology, the complete system is described into three parts - reaction centre, boundary, and spectator region, as has been mentioned in Chapter 1. Limitations (below, pointwise) accumulate from all three regions.

Reaction centre:

1. The first assumption is the clamped-nuclei approximation, also known as the Born-Oppenheimer approximation, regarding the motion of electrons in the field of static nuclei and moving electrons. This assumption has its limitations.
2. Under Born-Oppenheimer approximation, the many-body electronic wavefunction or more particularly, mutual interaction among electrons is approximated. Following the variational principle, the ground state wavefunction is searched. Since searching for all eligible functions is impossible, only subsets of acceptable functions are looked for so that a mathematical formulation can be applied for minimization. In this perspective, different approximations are used, like Hartree-Fock (HF) method, semi-empirical methods, and Density Functional theory (DFT). In all of them, the complete many-body wavefunction is replaced with an equivalent system where electrons do not see each other directly but through a mean field produced by other electrons. For example, in the HF method, the many-body effect of a multielectron system is approximated through a single determinantal form (Slater determinant), and electronic correlations are not considered. On the other hand, in DFT the so-called system-independent term, solely the behavior of electrons, is approximated by the Hohenberg-Kohn Functional, where the exchange-correlation functional term is aprroximated. We compromise between simplicity and accuracy by climbing up or down Jacob's Ladder. Core electrons are approximated with pseudopotentials. Interaction from one electron with itself does not get canceled in DFT unlike HF; leading towards self-interaction error (SIE). Since DFT cannot accomodate dispersion satisfactorily, the addition of correction term is a common practice [251]. Earlier works on single point energy (SPE) calculations (QM/MM) reflect the limitations arising from different quantum mechanical methods (SCS-MP2, LCCSD(T), DFT) for the reaction centre [252]. Under DFT calculations, B3LYP sometimes has been shown to underestimate reaction barriers [253]. Also, it has been found that the addition of dispersion decreases relative energies by 5-10 kcal/mol [254, 255]. Initial structures (e.g. different docking modes of the compounds at the active site) also contribute to determining reaction pathways [255]. SIE needs to corrected for large QM systems while determining activation and reaction energies [256].

Boundary and across:
3. Both the sub-systems across the boundary interact with each other. The effect of the environment outside the reaction centre is approximated by two schemes/embeddings - subtractive, and additive, as has been mentioned in Chapter 1. In the subtrative scheme, no communcation takes place between the QM and MM regions, and that leads to two limitations - a) a force field description becomes necessary for the rection centre, and b) modelling reactions like charge transfer becomes a problem because of the absence of possibility of polarization of QM electron density by the MM region. On the other hand, under an additive scheme, an explicit interaction term in present in the energy expression. Depending on how this interaction is being taken care of limits the scheme's applicability (mechanical, electrostatic, and polarization embedding); especially the non-bonded interactions, van der Waals, and electrostatics energy terms across the boundary can introduce errors. van der Waals interaction is described by Lennard-Jones potential where care needs to be taken in using the same parameter set for the QM atoms as that from the MM parametrization, which may lead towards overestimation of electrostatic interaction [257, 258]. As introduced in Chapter 1, mechanical embedding does not allow the polarization effect on the reaction centre by the environment. Polarization embedding on the other hand includes polarization effect of both the subsystems on each other across boundary but the computation becomes quite demanding.
4. Because of the necessary interactions across the boundary, where the boundary should be drawn is a critical matter. Earlier studies from SPE calculations showed that energy values depend on the size of the quantum region [254]. Charge deletion analysis was proposed to determine the optimum size of the QM region [259].
5. Valences of the quantum region is satisfied through different ways (link atoms, LSCF orbitals, GHO orbitals); which may introduce errors to the calculation.

Spectator region:
6. The molecular dynamics of the environment is described with the classical force fields potential energy functions. Borrowing words from Anatol Brodsky it can be said that 'in real molecular dynamics there are no other universal potentials but Coloumbic potentials with quantum electrons.' [260] There exists no one-to-one correspondence between classical energy terms (of the force fields) with the accurate quantum mechanical calculations. Force field accuracy is usually measured through error cancellation between different terms. Even polarizable force fields fail to capture phenomenon like hyperconjugation, charge-penetration, charge transfer, and anisotropy in exchange-repulsion.
7. Water is described with different rigid non-polarizable models. And, none of these models reproduce experimental observations completely but only to a varying degree [261]. Hence any error from water description has to be compensated by the protein (or other biomolecules)-water interaction [262].

Now for our study, the quantum system was of a small size; so hopefully SIE does not have a big role to play. And, thus we believe the partial free energy surface with deprotonation barrier of 3.4 kcal/mol and barrierless cyclization is trustable. Having said this, we also think a systematic evaluation with different functionals would be a good future option. From our calculations we could not yield any signatures of a prominent role of 158^{th} residue in the post-translational modification, unlike results from experimental single-site mutation study and the role of water molecules in the reaction could not be nullified also. Alongwith this, influence of additional residues in the quantum region could possibly be a good check.

Studying reaction mechanism of post-translational modifications stand on its own right because of their importance in evolution. As introduced in the beginning of this chapter that spontaneous deamidation of neurtral asparagine and glutamine residues to their acidic counterparts in proteins lead to neurodegenerative diseases and cataracts. And, this kind of stable succinimide is not so ubiquitous ([241], https://www.rcsb.org/structure/1WL8) in nature. Hence, to further our understanding, studying mechanisms of such reaction is of fundamental interest.

5

Summary and future outlook

An atomic scale understanding of biophysical and biochemical processes is a grand challenge. In this thesis, I describe how molecular dynamics simulations can be used to study proteins under confinement and autocatalyzed reactions in protein.

The confinement here was an organic-inorganic composite - metal-organic frameworks (MOF). Proteins or enzymes encapsulated inside MOF being extensively used in biotechnological industry, is a big enterprise for our civilization. A causal understanding of these systems is thus important. Chapter 2 and Chapter 3 describe protein@MOF systems.

In Chapter 2, we studied two important proteins, myoglobin and GFP in IRMOFs through extensive all atom equilibrium molecular dynamics simulations. This showed us that the main driving force for this kind of confinement was van der Waals interaction. Hence the overall structure of these two big proteins (matching relative sizes of the channels) (myoglobin : 138 residues and GFP: 238 residues) were well maintained including the respective active sites.

This helped us to further ask the question how the proteins enter the MOF or traverse through their pores, where the relative size of the protein is larger than the pore entrance of MOF. Chapter 3 has tried to answer this question with a model system. Intuitively, the protein has to undergo unfolding in order to pass (as has been suggested through experiments as well). We have devised a combination of two enhanced sampling techniques - steered molecular dynamics (SMD) and umbrella sampling (US) for investigating the mechanism and estimating a free energy barrier for the process. Through unfolding

of HP35 while passing through the hexagonal pore window of the hierarchical MOF, MIL-101(Cr), we observed how the geometry of the MOF controlled unfolding of the protein. The translocation path (from center of the cavity to the pore window) could be divided into three major zones. A central zone which mostly preserves the native ensemble of the protein along with partially-unfolding and major-unfolding zones. For this model system, the free energy barrier is estimated to be 16 kcal/mol at 298 K. A near future extension of this would be generating the conformational ensemble of the protein under MOF confinement from the available folding-unfolding pure water trajectory without using any enhanced sampling. These conformations will be the collection of states for the translocation phenomenon as well. This will also help us in bringing up a comparison between the confined ensemble and the water conformations. A far future extension of this would be kinetic modelling of the process which would give us more insights into the mechanism.

In the final work chapter (Chapter 4), we implemented QM/MM with electrostatic embedding to study the autocatalyzed post-translational modification (PTM), succinimide formation in MjGATase. QM region embraced four residues of the protein (including the succinimide forming residue, N_{109}, suggested from mutational experiments by experimental collaborators). We devised non-tempered metadynamics with one pair of collective variable for the succinimide formation reaction consisting of two major steps - deprotonation and deamination. We could observe only one transition from the reactant to the product basin. Though Tyr_{158} engaged in hydrogen bond formation with nearby residues as a donor, we could not conclude any direct participation of Tyr_{158} in the reaction in contrast to experiments. We could not able see any role for Lys_{151}, as well. We estimated a deprotonation activation barrier of 3.4 kcal/mol, followed by barrierless cyclization for the concerted mechanism of succinimide formation. A near future extension would be exploration of other collective variables along with different level of quantum description to observe a complete reaction with forward and backward transitions. A far future extension would be to implement other reaction discovery tools.

Appendices

A

Supplementary Information for Chapter 2

A.1 Abbreviations

For secondary structural motifs of GFP, we have used the following short forms in the plots. Helix-1,2,3 have been denoted as H1, H2, H3, respectively. βstrand-1 to βstrand-11 have been denoted as βS1, βS2, βS3, βS4, βS5, βS6, βS7, βS8, βS9, βS10, βS11, respectively.

A.2 MOF details:

Table A.1: Parameters for MOFs worked with. Unitcell and supercell parameters are given in the format (a x b x c, α, β, γ). Cell lengths are in units of Å.

MOF	CCDC Number	Refcode	Primitive unitcell parameters	Supercell parameters
VII-oeg	841649	RAVXET	6.73 x 54.55 x 54.55, 60.17,87.64,92.36	87.51 x 109.09 x 109.09, 60.17,87.64,92.36
VII-hex	841648	RAVXAP	6.99 x 53.14 x 53.14, 60.19,87.49,92.51	90.85 x 106.29 x 106.29, 60.19,87.49,92.51
IX	841650	RAVXIX	6.94 x 65.73 x 65.73, 60.12,87.98,92.02	76.39 x 131.46 x 131.46, 60.12,87.98,92.02

A.3 Secondary structure motifs

Myoglobin [263]

Table A.2: Definition of secondary structure of myoglobin used for analysis

Structural motif	Residue ID
Loop before A-Helix	1-2
A-Helix	3-18
AB Loop	19
B-Helix	20-35
C-Helix	36-42
Loop between C and D	43-50
D-Helix	51-57
E-Helix	58-77
EF Loop	78-85
F-Helix	86-94
FG Loop	95-99
G-Helix	100-118
GH Loop	119-124
H-Helix	125-148
Loop beyond H-Helix	149-153

End of Table

GFP [264, 265]

Table A.3: Definition of secondary structure of GFP used for analysis

Structural motif	Residue ID
Loop before Helix-1	1-2
Helix-1	3-9
Loop between Helix-1 and βStrand-1	10
βStrand-1	11-23
Loop between βStrand-1 and βStrand-2	24
βStrand-2	25-38
Loop between βStrand-2 and βStrand-3	39
βStrand-3	40-49
Loop between βStrand-3 and Helix-2	50-75
Helix-2	76-81
Loop between Helix-2 and Helix-3	82
Helix-3	83-88
βStrand-4	89-101
Loop between βStrand-4 and βStrand-5	102
βStrand-5	103-114
Loop between βStrand-5 and βStrand-6	115-117
βStrand-6	118-128
Loop between βStrand-6 and βStrand-7	129-142
βStrand-7	143-155
Loop between βStrand-7 and βStrand-8	156-159
βStrand-8	160-171
Loop between βStrand-8 and βStrand-9	172-174
βStrand-9	175-188
Loop between βStrand-9 and βStrand-10	189-196
βStrand-10	197-210
Loop between βStrand-10 and βStrand-11	211-214
βStrand-11	215-228
Loop beyond βStrand-11	229-238

End of Table

A.4 Parameters used for five-coordinated HEME

Table A.4: Modified parameters of the penta-coordinated HEME used in simulation following literature. [136, 137]

Parameter	r (nm) or Θ°	k ($kJmol^{-1}nm^{-2}$) or k ($kJmol^{-1}rad^{-2}$)
Bonds	-	-
NPH-FE	0.21	227776.96
NE2-FE	0.21	54392.00
Angles	-	-
NPH-FE-NPH	90	669.44
NE2-FE-NPH	107	418.40

A.5 Missing Dihedrals around proximal histidine

Four dihedral angles were missing from the CHARMM parameter set and, hence, have been added manually.

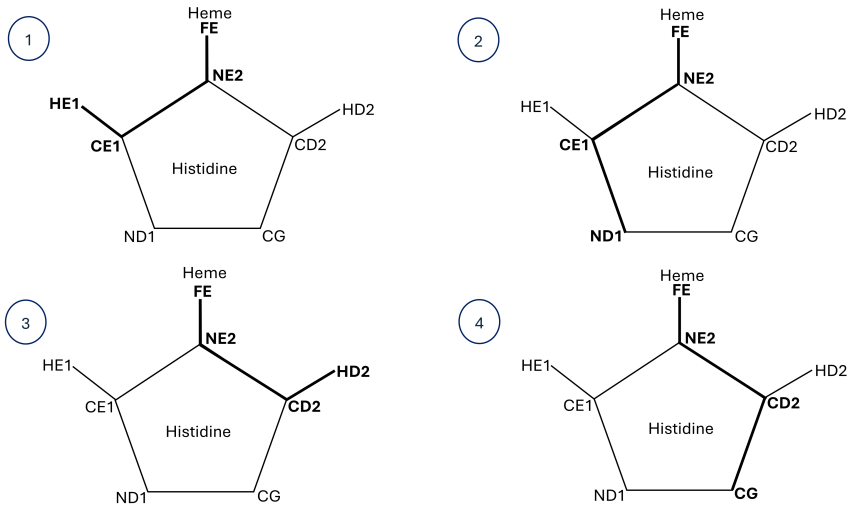


Fig. A.1: Missing dihedrals around Fe (of HEME) ligated HIS-93. Atoms corresponding to the dihedral and the dihedral angles are highlighted in bold.

A.6 Addtional simulation details:

Well-tempered metadynamics simulation for GFP was carried out using PLUMED (verion:2.5.4) that was patched with GROMACS (version:2018.3).

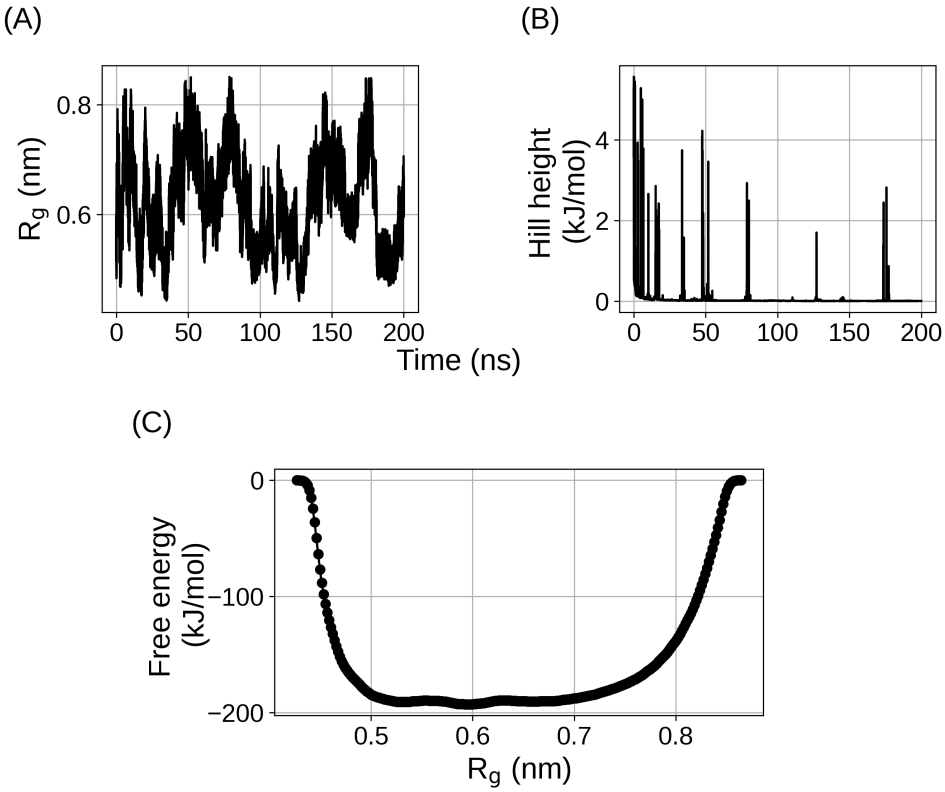


Fig. A.2: Results from well-tempered Metadynamics simulation. (A)Time evolution of radius of gyration. (B)Time evolution of the deposited Gaussian hills. (C)Free energy surface for the conformational landscape of the C-ter segment containing seven residues.

The cutoff used for GROMOS was 1.8 Å. The population of the first cluster was around 76%. The structure written from each cluster was the middle structure.

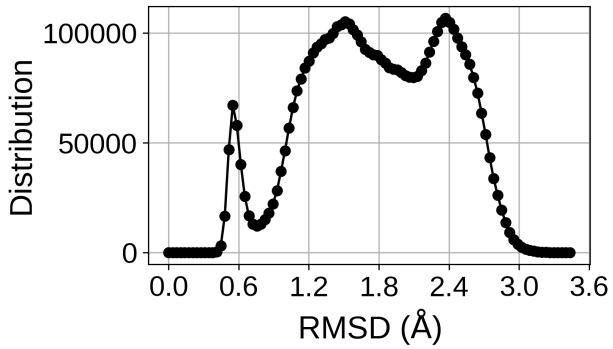


Fig. A.3: RMSD distribution from cluster analysis.

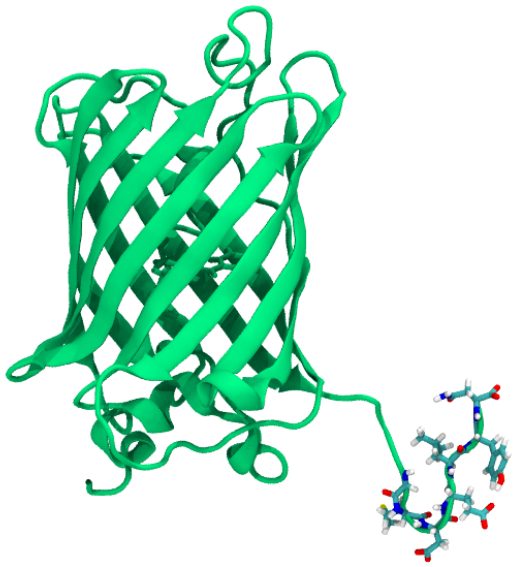


Fig. A.4: Lowest free energy structure from the highest populated cluster. Residues considered for well-tempered Metadynamics are shown in Licorice representation.

A.7 Pore size distribution

Zeo^{++} (version 0.3) was used Voronoi decomposition to get an estimate of the pore sizes in the unit cell and primitive unit cells. The probe radius and the chan radius, were both set to be 1.2 Å. The number of Monte Carlo cycles was 50000.

A.8 NPT simulation: Solvated MOF supercell

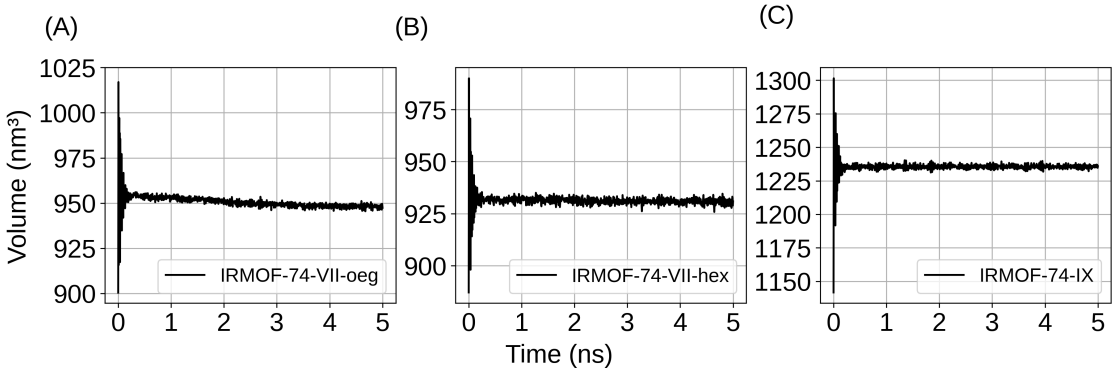


Fig. A.5: Volume change during NPT simulations for the MOFs.

Table A.5: Volume change of MOF supercell containing water during NPT simulations.

MOF	a(Å)	b(Å)	c(Å)	$\alpha(^\circ)$	$\beta(^\circ)$	$\gamma(^\circ)$	Av. Volume(nm^3)
VII-hex	96.27	105.78	105.41	60.56	87.19	92.62	931.25
VII-oeg	93.57	108.40	108.11	60.19	84.67	92.57	950.50
IX	80.81	133.12	132.43	60.42	88.00	92.13	1235.53

A.9 Backbone RMSD for the analyzed trajectory segments.

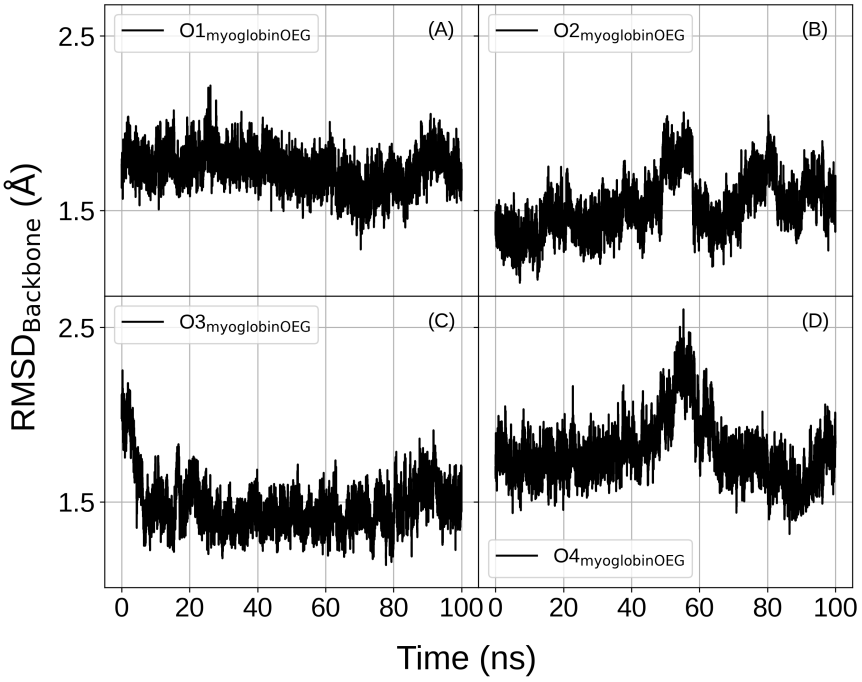


Fig. A.6: Backbone RMSD of myoglobin in IRMOF-74-VII-oeg in MD simulation runs initiated with the enzyme oriented variously with respect to the MOF.

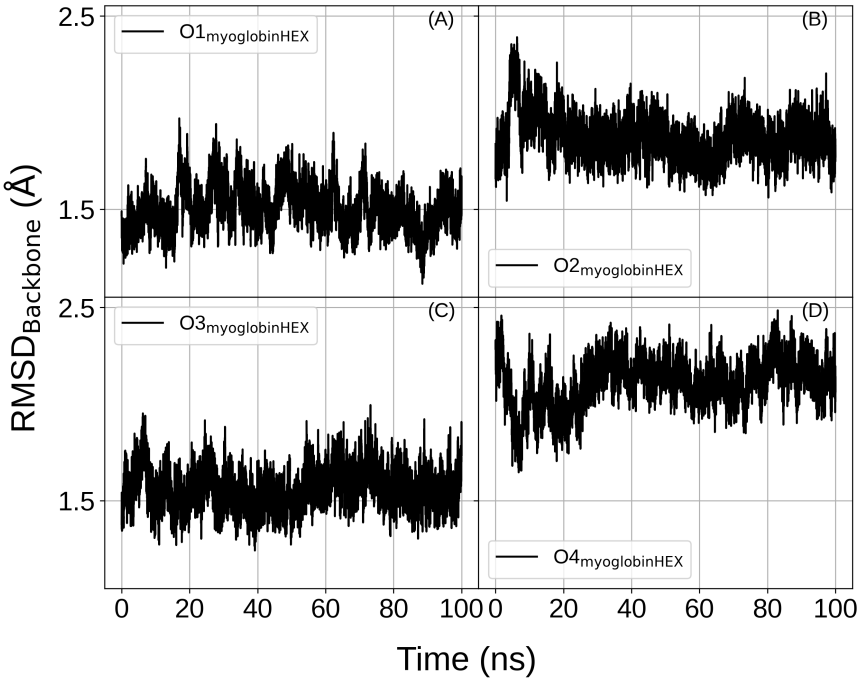


Fig. A.7: Backbone RMSD of myoglobin in IRMOF-74-VII-hex in MD simulation runs initiated with the enzyme oriented variously with respect to the MOF.

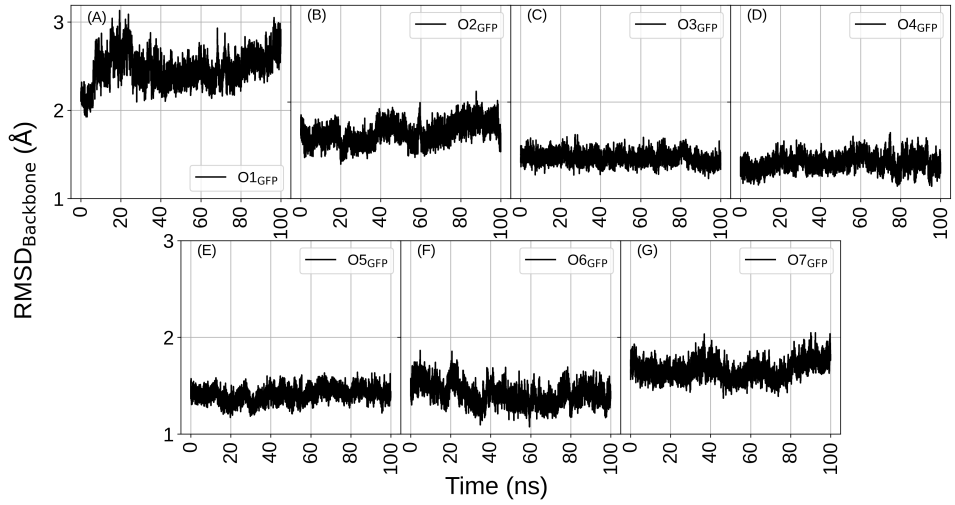


Fig. A.8: Backbone RMSD of GFP in IRMOF-74-IX in MD simulation runs initiated with the protein oriented variously with respect to the MOF.

A.10 Root mean squared fluctuations of the side chains of proteins for all orientations.

Table A.6: Root mean squared fluctuations averaged over the primary sequence for myoglobin@IRMOF-74-VII-oeg and myoglobin@IRMOF-74-VII-hex.

Water (Å)	$O_{myoglobinOEG}$ (Å)	$O_{myoglobinHEX}$ (Å)
1.36	1.36	1.13

Table A.7: Root mean squared fluctuations averaged over the primary sequence for GFP@IRMOF-74-IX.

Water (Å)	O_{GFP} (Å)
1.32	1.21

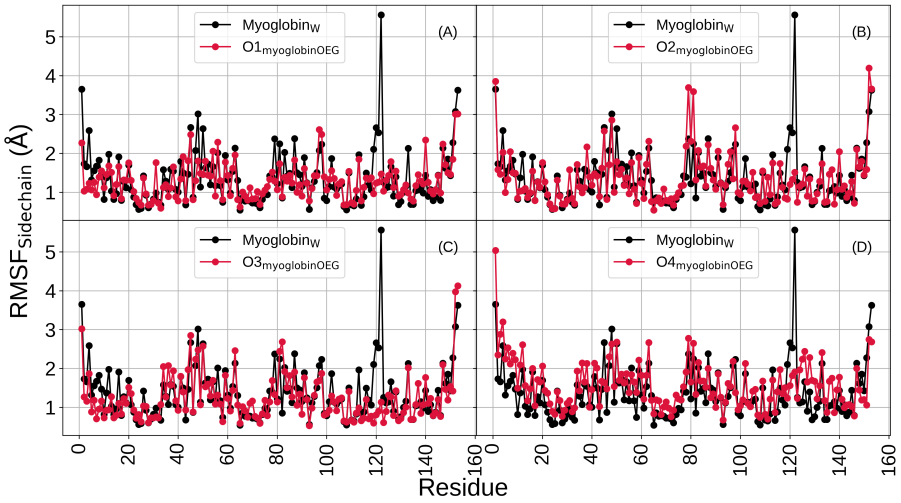


Fig. A.9: RMSF of myoglobin side chains in water and inside IRMOF-74-VII-oeg. (A) to (D) are for orientations 1 to 4, respectively.

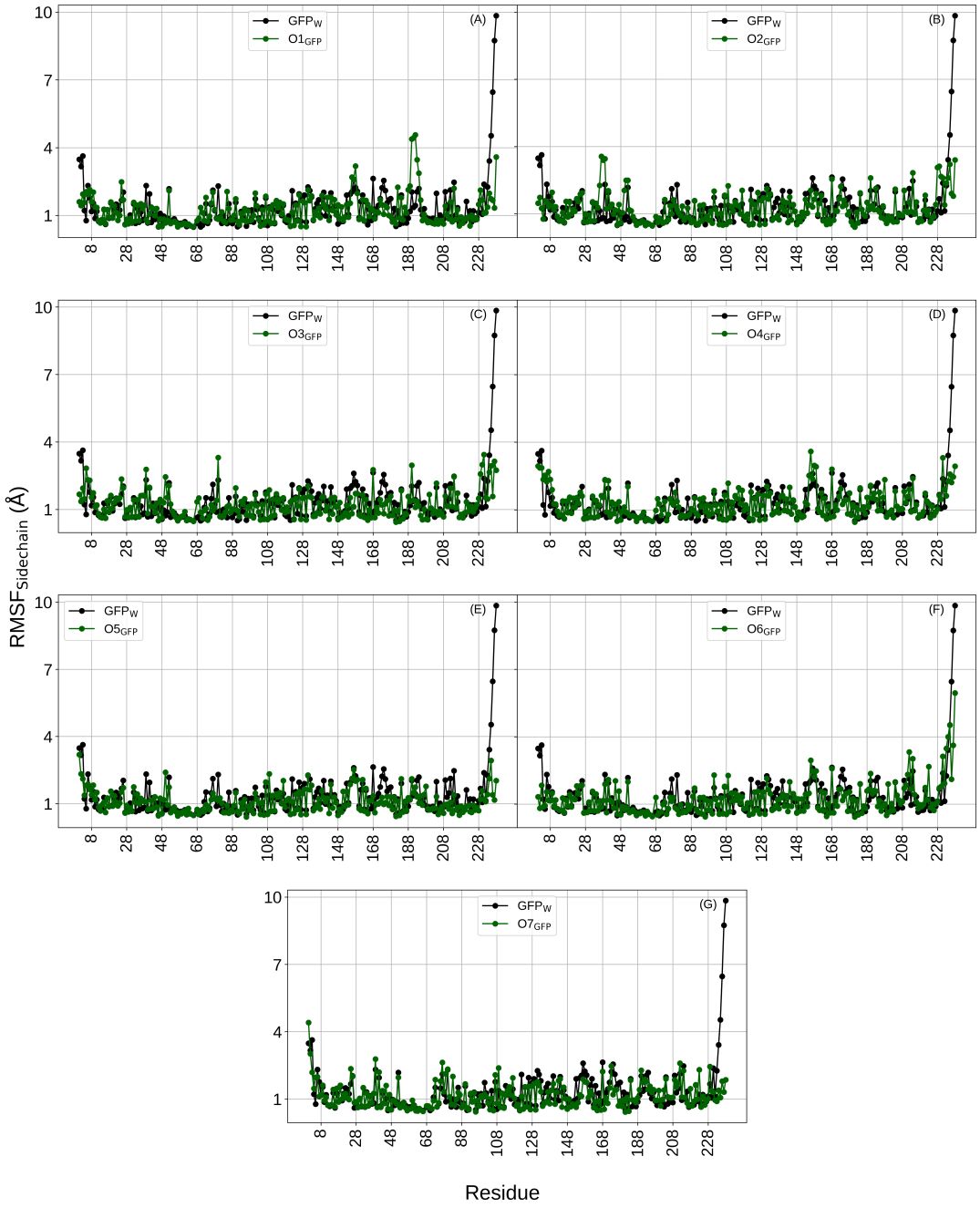


Fig. A.10: RMSF of GFP side chains in water and inside IRMOF-74-IX. (A) to (G) are for orientations 1 to 7, respectively.

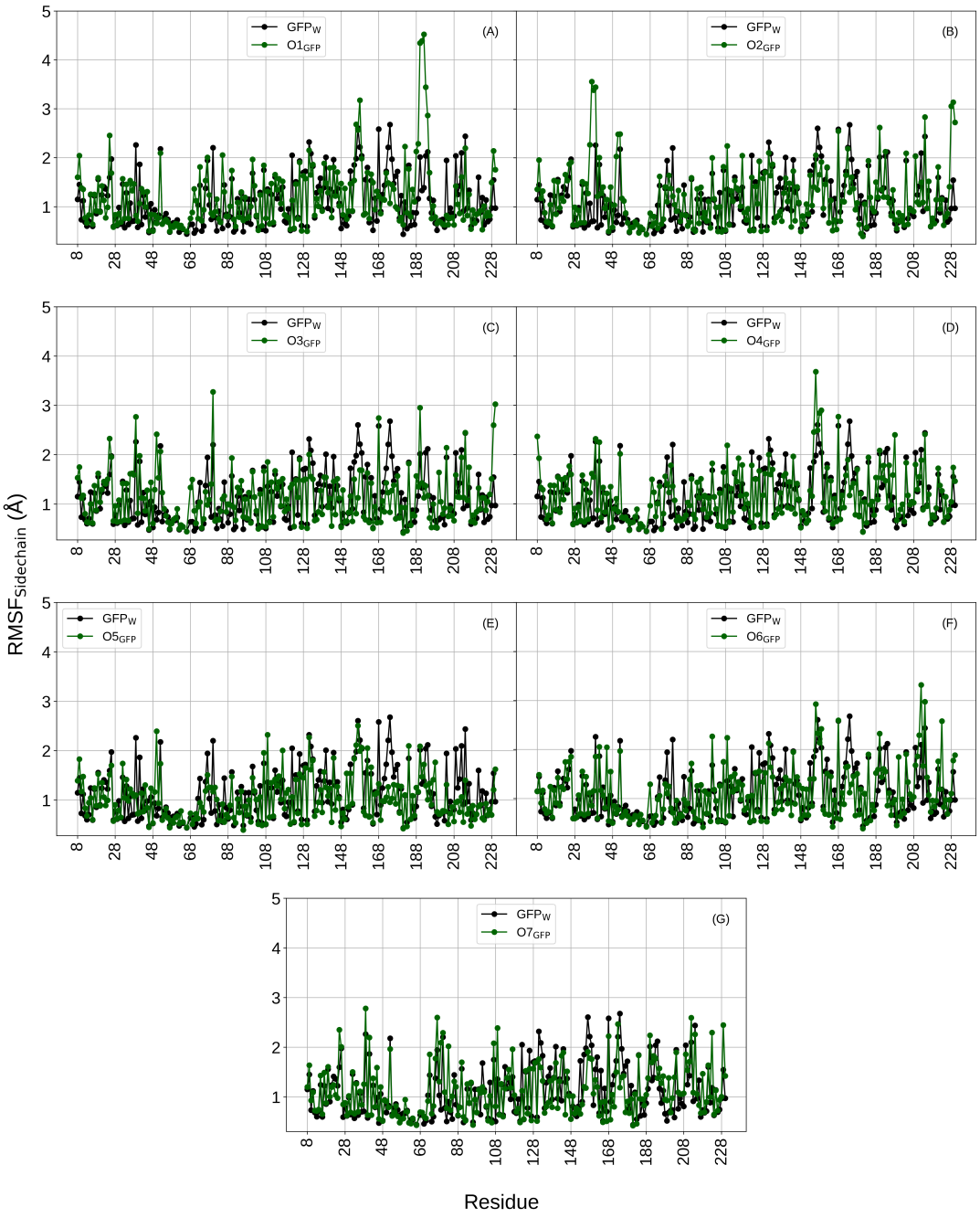


Fig. A.11: RMSF (excluding seven and eight residues from N- and C-TER, respectively) of GFP side chains in water and inside IRMOF-74-IX. (A) to (G) are for orientations 1 to 7, respectively.

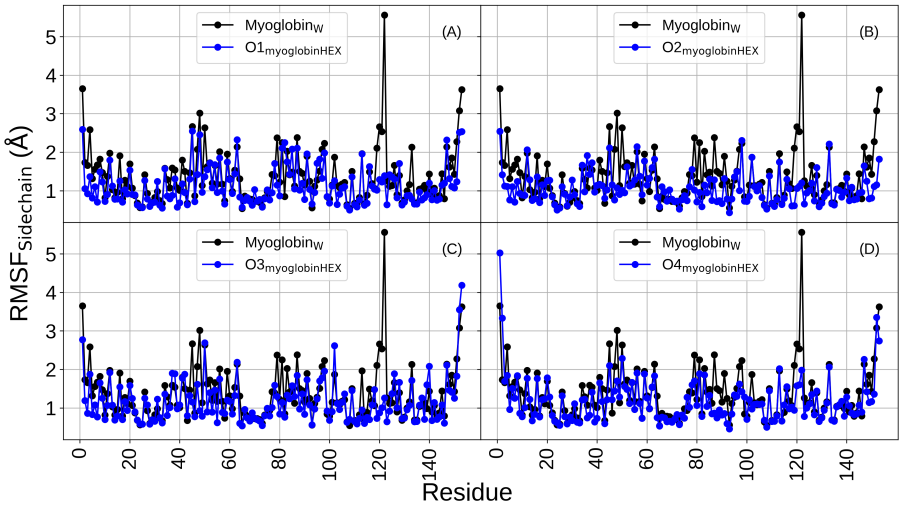


Fig. A.12: RMSF of myoglobin side chains in water and inside IRMOF-74-VII-hex. (A) to (D) are for orientations 1 to 4, respectively.

A.11 Structure of the active site of myoglobin and chromophore of GFP.

Table A.8: RMSF for the prosthetic group of myoglobin (HEME). Av.$_{\text{Ori}}$ stand for average over the orientations.

@MOF	O1 (Å)	O2 (Å)	O3 (Å)	O4 (Å)	Av.$_{\text{Ori}}$
@IRMOF-74-VII-oeg	0.83	0.58	0.65	0.63	0.67
@IRMOF-74-VII-hex	0.71	0.55	0.67	0.58	0.63

Table A.9: RMSF for the chromophore of GFP. Av.$_{\text{Ori}}$ stand for average over the orientations.

@MOF	O1 (Å)	O2 (Å)	O3 (Å)	O4 (Å)	O5 (Å)	O6 (Å)	O7 (Å)	Av.$_{\text{Ori}}$
@IRMOF-74-IX	0.21	0.20	0.20	0.19	0.21	0.18	0.19	0.20

Table A.10: Mean values of the RMSD of the prosthetic group of myoglobin (HEME) with respect to NPT equilibrated structure. $Av._{Ori}$ stand for average over the orientations.

@MOF	O1 (Å)	O2 (Å)	O3 (Å)	O4 (Å)	$Av._{Ori}$
@IRMOF-74-VII-oeg	1.16	1.16	1.02	1.26	1.15
@IRMOF-74-VII-hex	1.38	1.07	0.95	1.80	1.30

Table A.11: Mean values of the RMSD of the chromophore of GFP with respect to the NMR structure of the chromophore (PDB:2WUR). $Av._{Ori}$ stand for average over the orientations.

@MOF	O1 (Å)	O2 (Å)	O3 (Å)	O4 (Å)	O5 (Å)	O6 (Å)	O7 (Å)	$Av._{Ori}$
@IRMOF-74-IX	0.39	0.39	0.37	0.31	0.35	0.33	0.42	0.37

A.12 Secondary structural motifs for myoglobin@IRMOF-74-VII-oeg, myoglobin@IRMOF-74-VII-hex and GFP@IRMOF-74-IX.

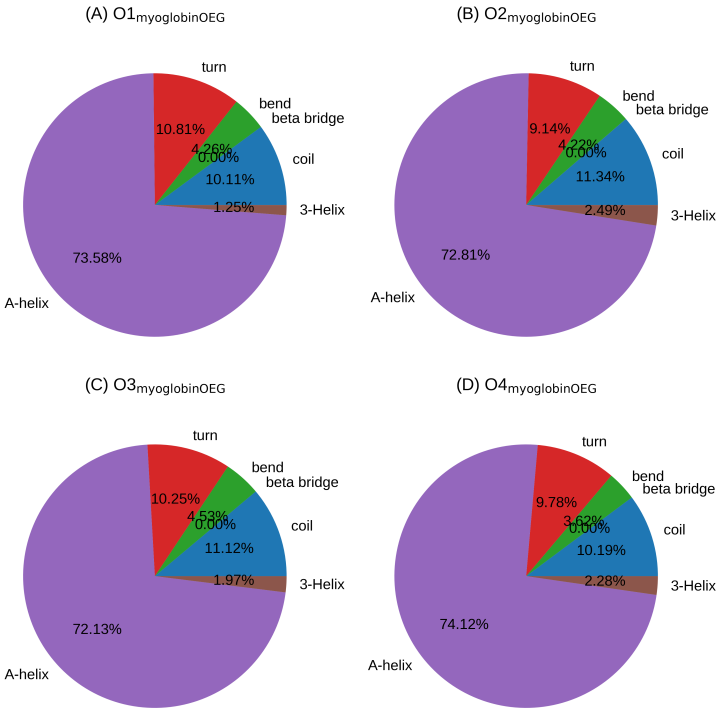


Fig. A.13: Percentage content of secondary structural motifs averaged over the analysis trajectory for myoglobin@IRMOF-74-VII-oeg.(A) to (D) are for four different orientations.

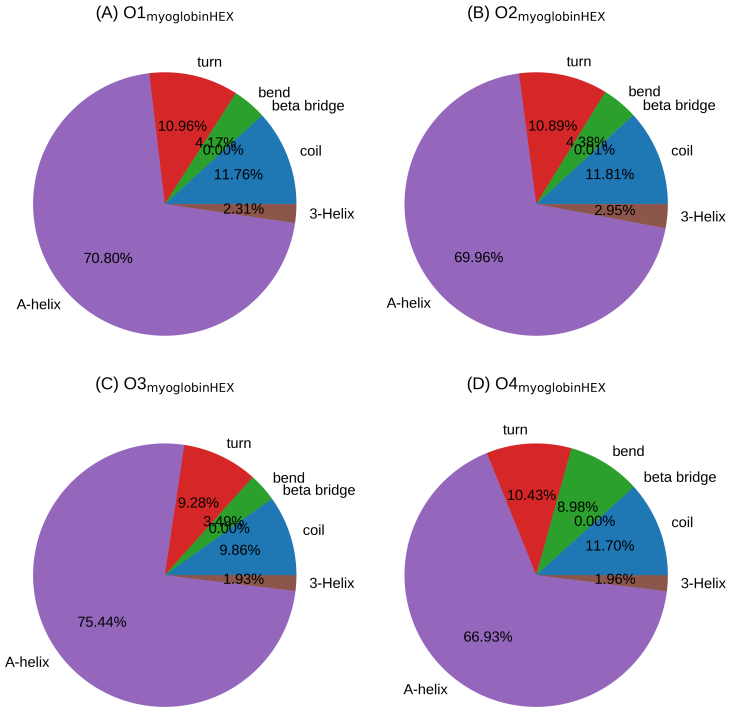


Fig. A.14: Percentage content of secondary structural motifs averaged over the analysis trajectory for myoglobin@IRMOF-74-VII-hex.(A) to (D) are for four different orientations.

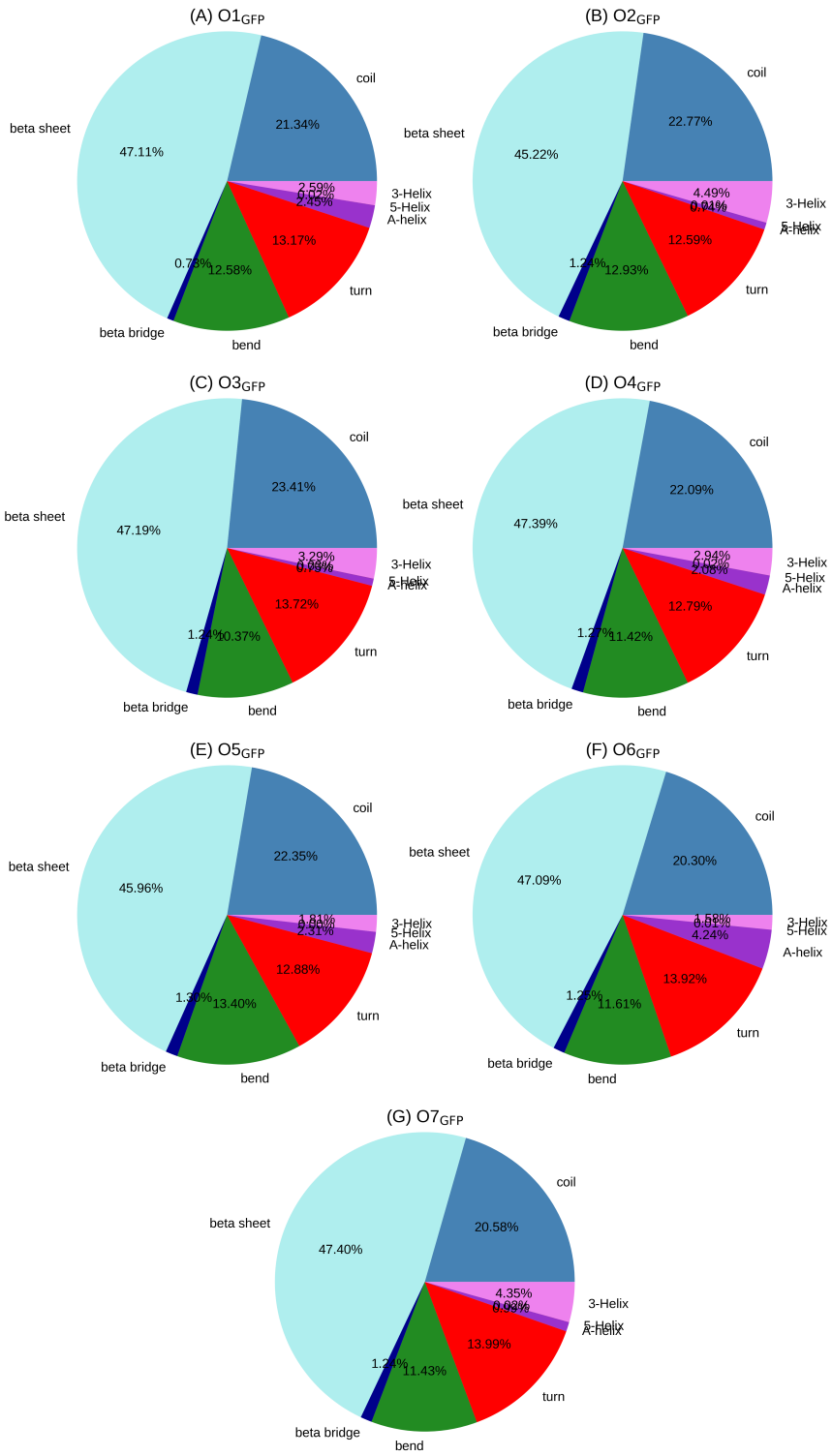


Fig. A.15: Percentage content of secondary structural motifs averaged over the analysis trajectory for GFP@IRMOF-74-IX.(A) to (G) are for seven different orientations.

A.13 Our definition of ‘MainChain’ and ‘SideChain’ of MOFs and contacts between MOF atoms and protein surface.

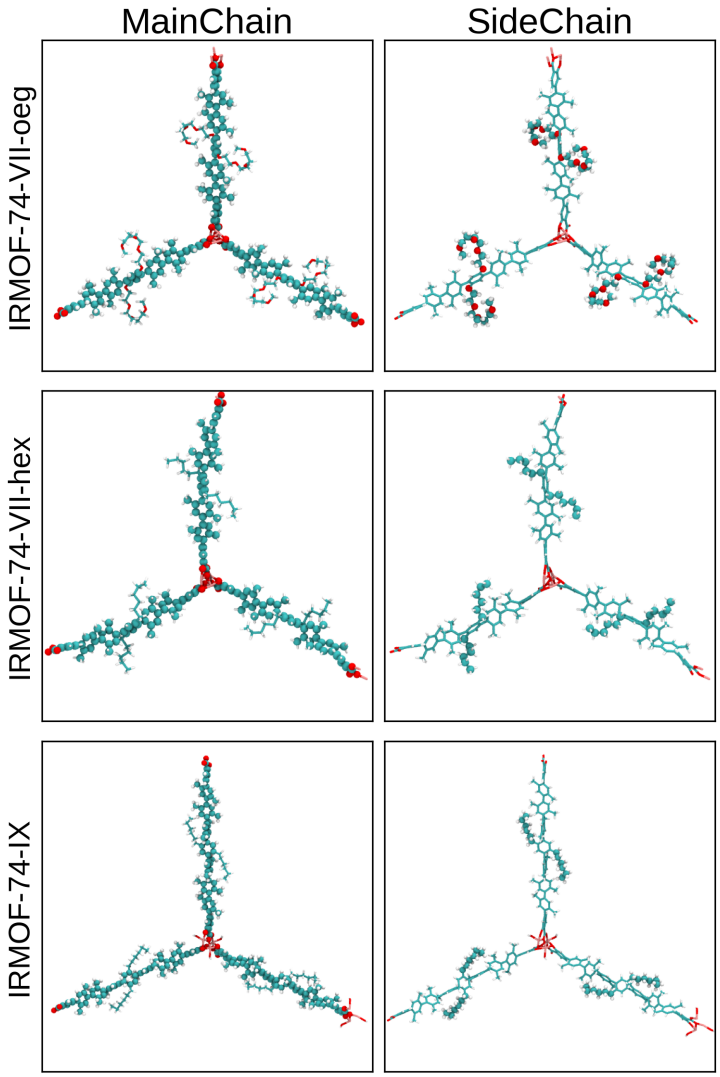


Fig. A.16: Definition of MainChain and SideChain in MOF linker. Main and Side chains are highlighted in vdW representation with reduced sphere scale.

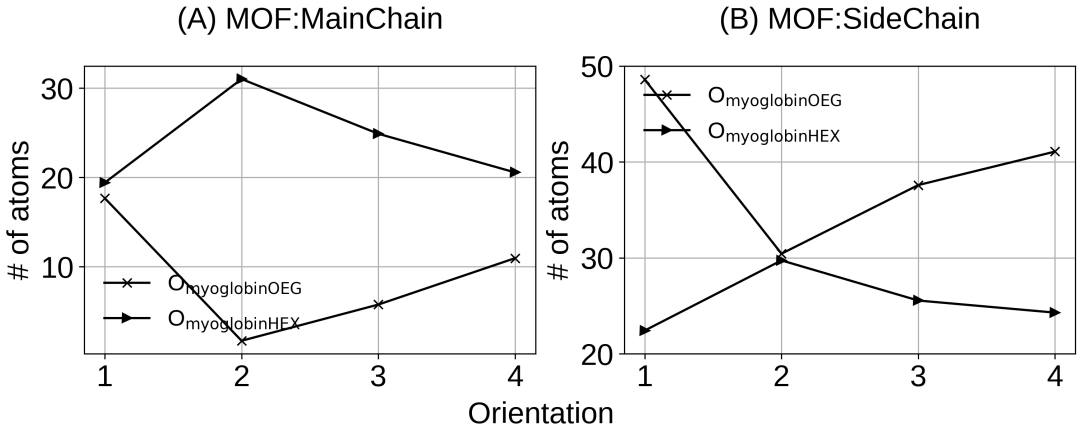


Fig. A.17: Number of (A) Mainchain and (B) SideChain heavy atoms of MOF within 4 Å of any heavy atom of myoglobin.

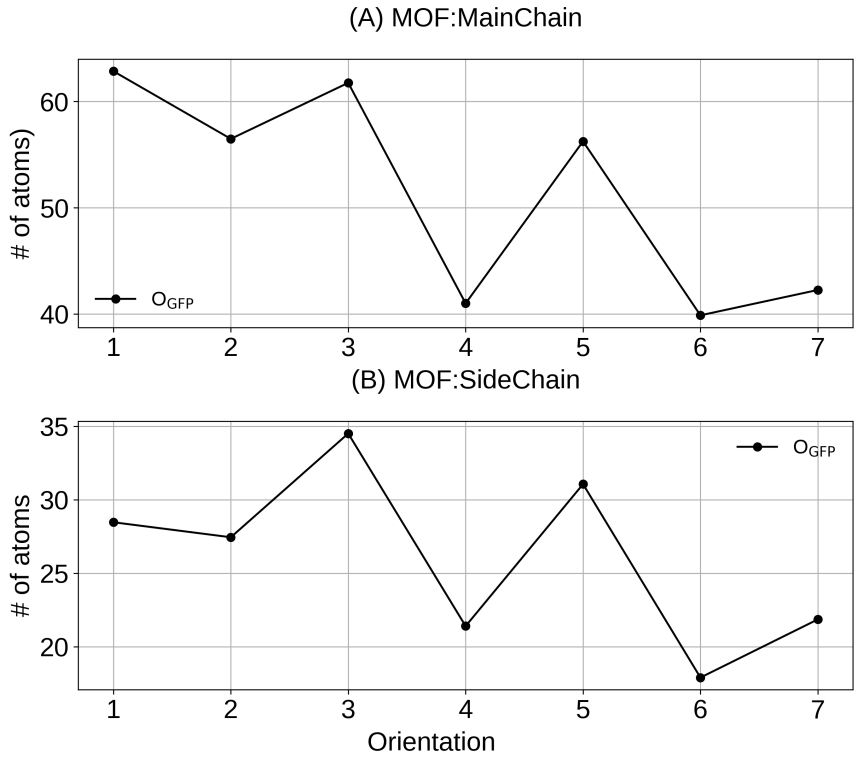


Fig. A.18: Number of (A) Mainchain and (B) SideChain heavy atoms of MOF within 4 Å of any heavy atom of GFP.

A.14 Protein-MOF interaction

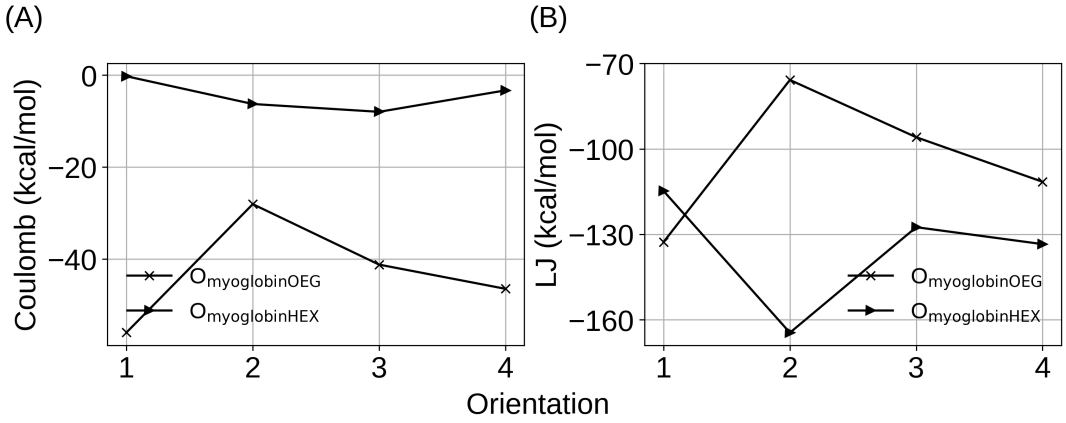


Fig. A.19: Coulomb and LJ interactions between myoglobin and IRMOF-74-VII surface across different orientations for both the inclusions.

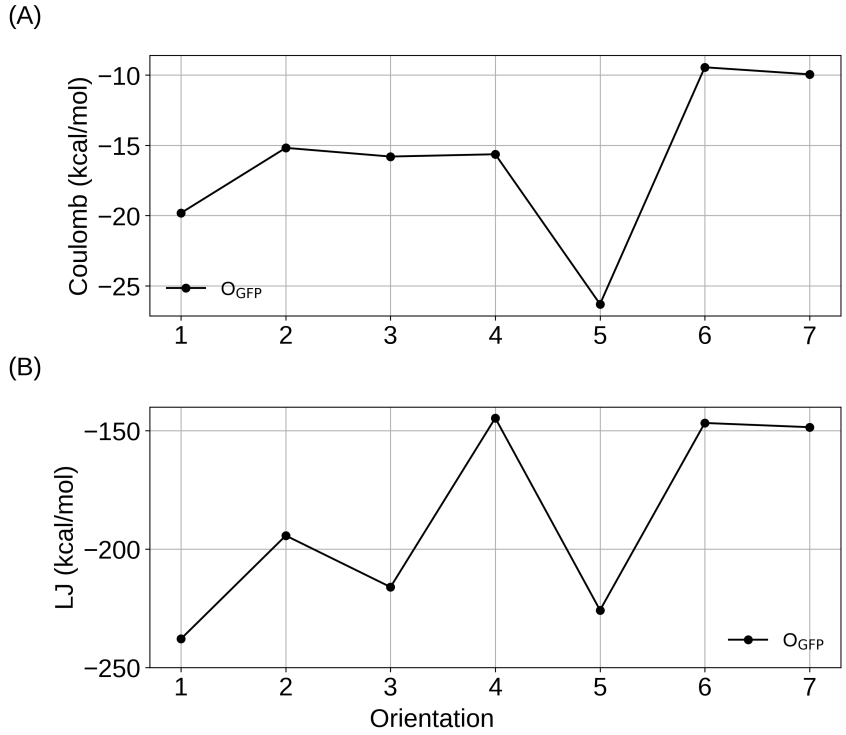


Fig. A.20: Coulomb and LJ interactions between GFP and IRMOF-74-IX surface across different orientations.

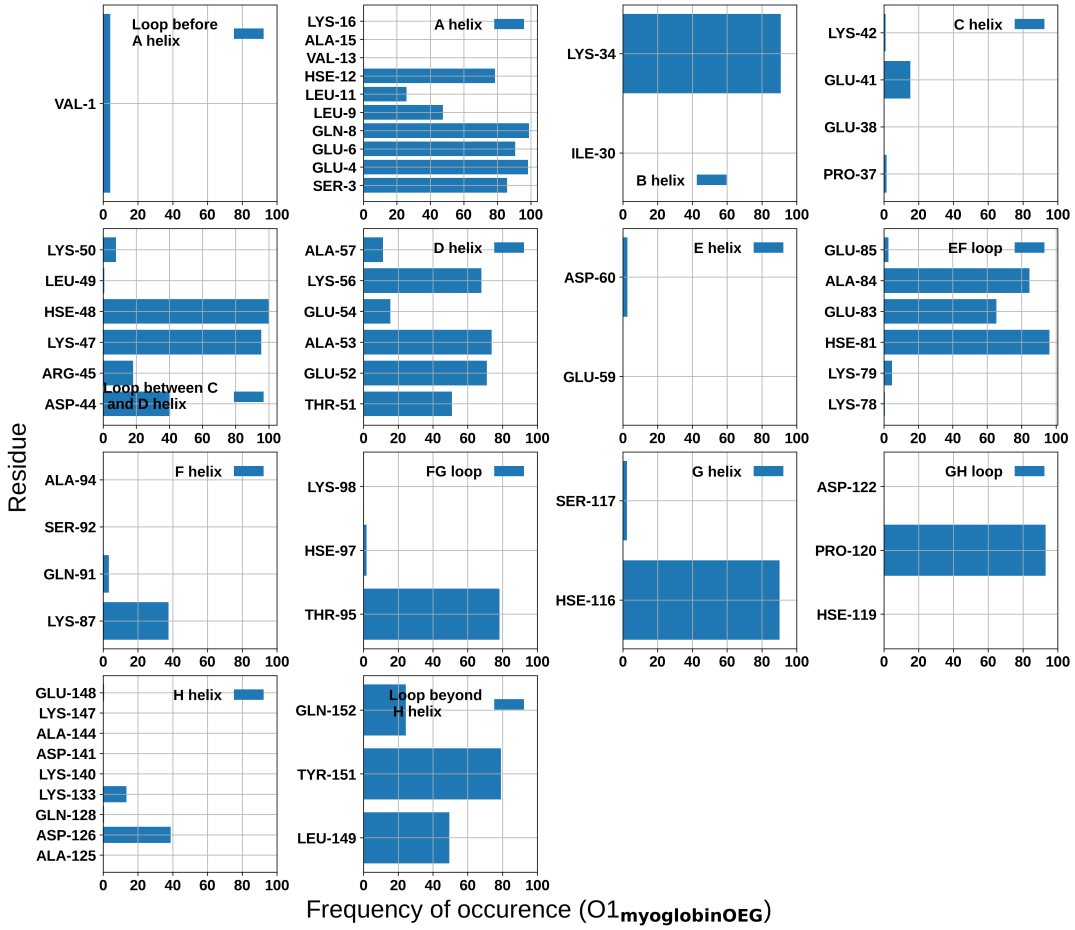


Fig. A.21: Residues for non-hydrogen atom contacts between myoglobin and IRMOF-74-VII-oeg surface with a cutoff of 4 Å for orientation 1.

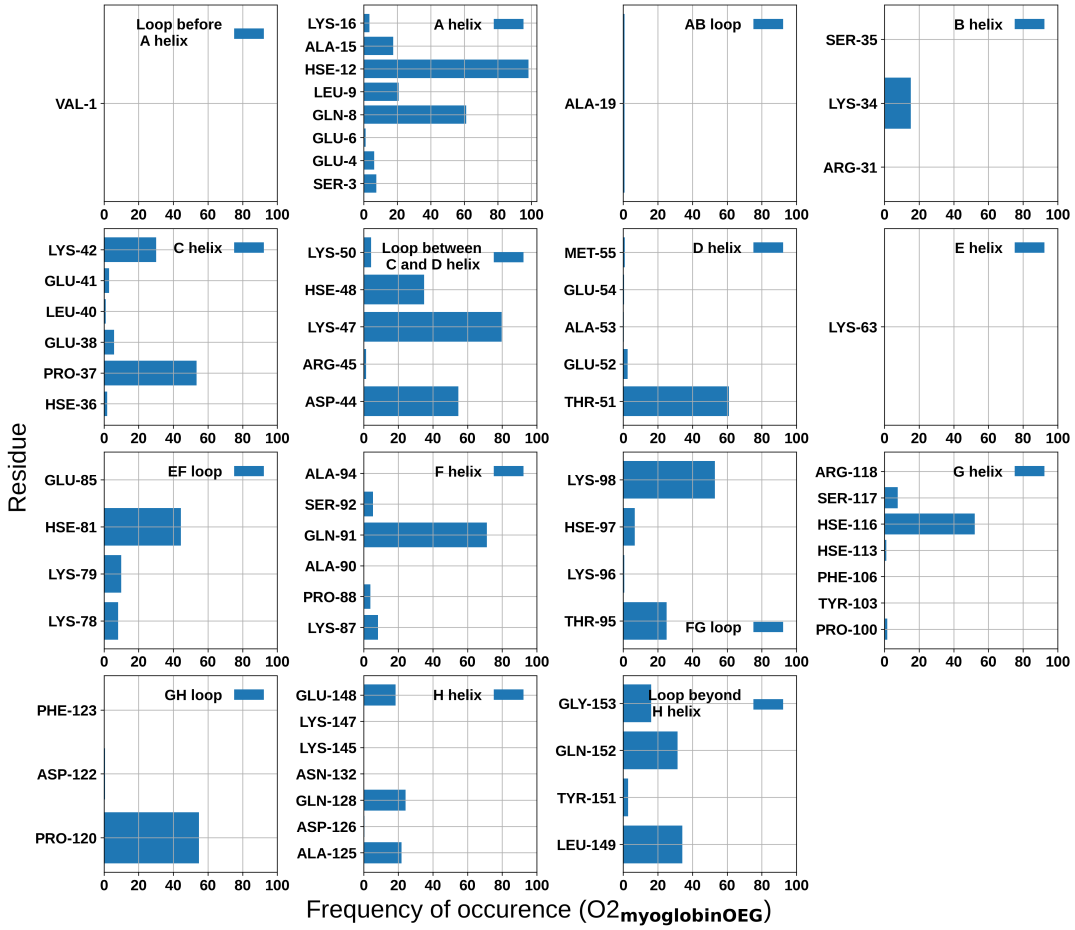


Fig. A.22: Residues for non-hydrogen atom contacts between myoglobin and IRMOF-74-VII-oeg surface with a cutoff of 4 Å for orientation 2.

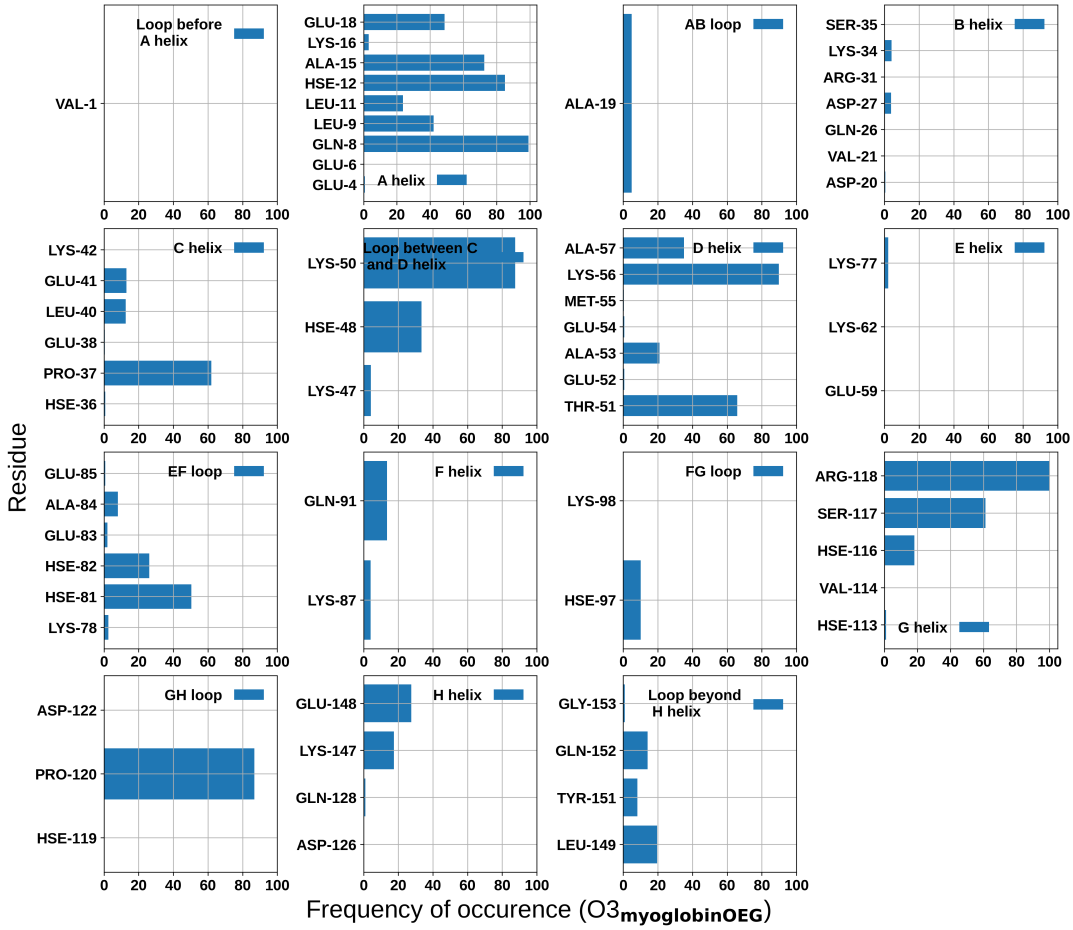


Fig. A.23: Residues for non-hydrogen atom contacts between myoglobin and IRMOF-74-VII-oeg surface with a cutoff of 4 Å for orientation 3.

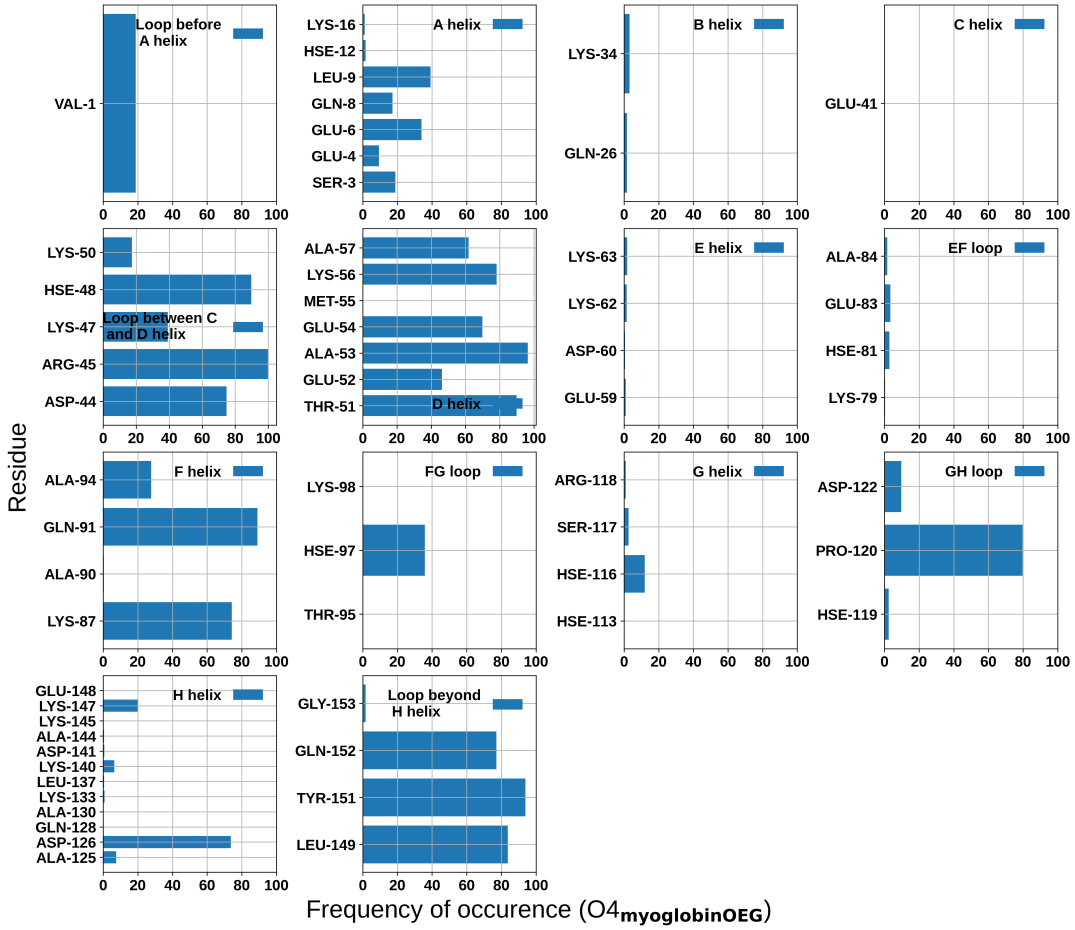


Fig. A.24: Residues for non-hydrogen atom contacts between myoglobin and IRMOF-74-VII-oeg surface with a cutoff of 4 Å for orientation 4.

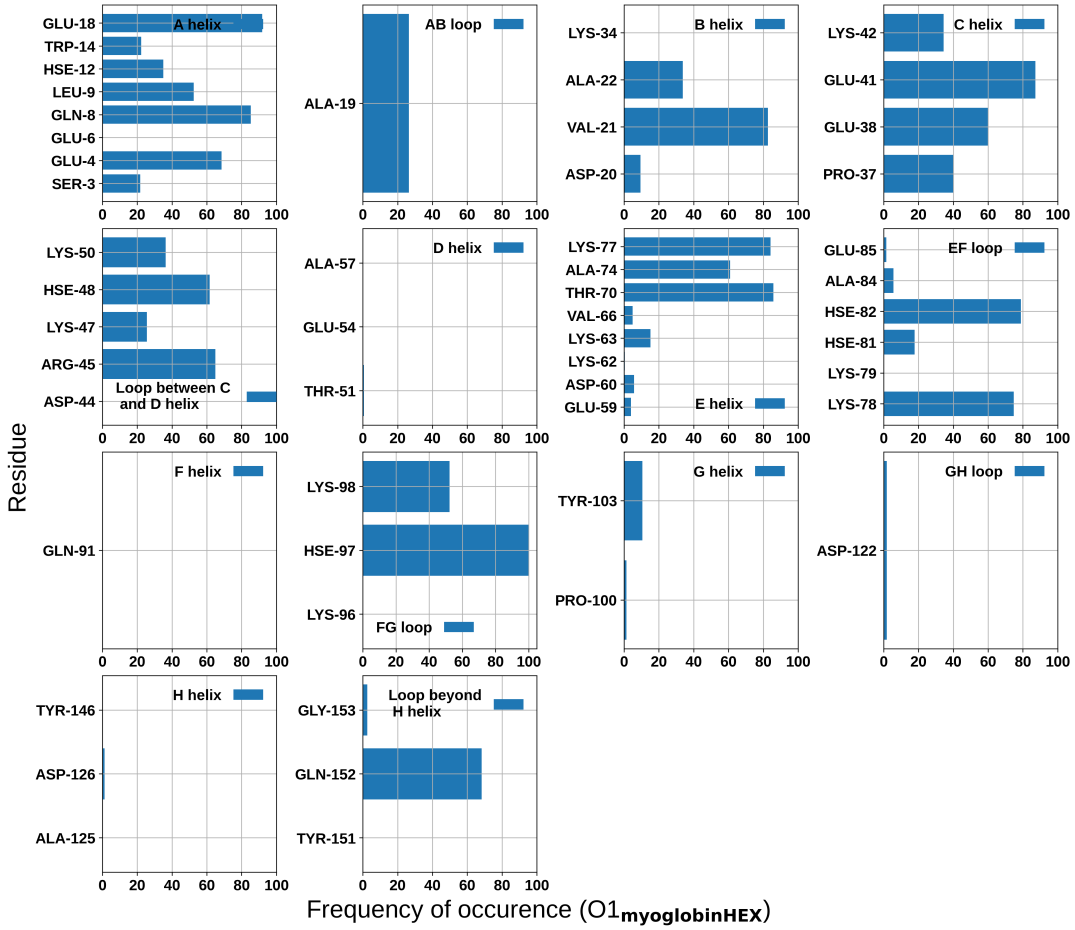


Fig. A.25: Residues for non-hydrogen atom contacts between myoglobin and IRMOF-74-VII-hex surface with a cutoff of 4 Å for orientation 1.

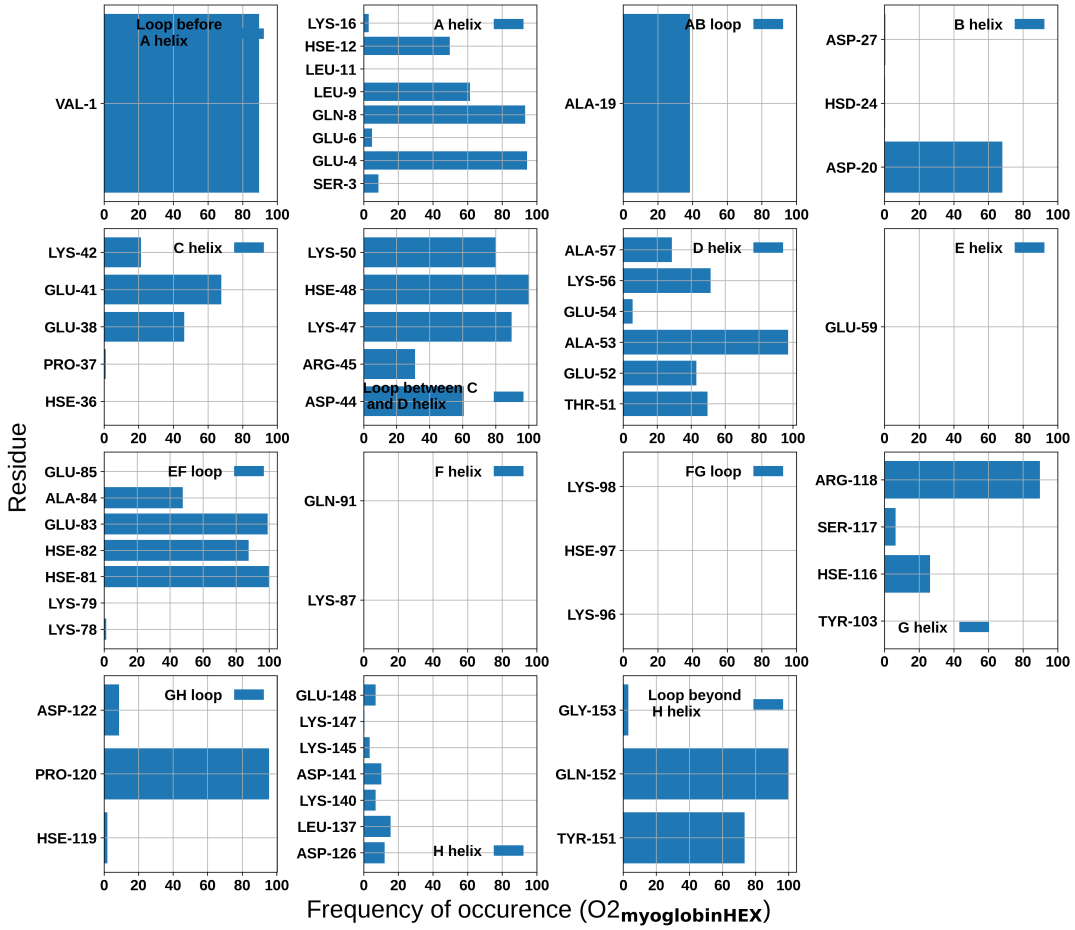


Fig. A.26: Residues for non-hydrogen atom contacts between myoglobin and IRMOF-74-VII-hex surface with a cutoff of 4 Å for orientation 2.

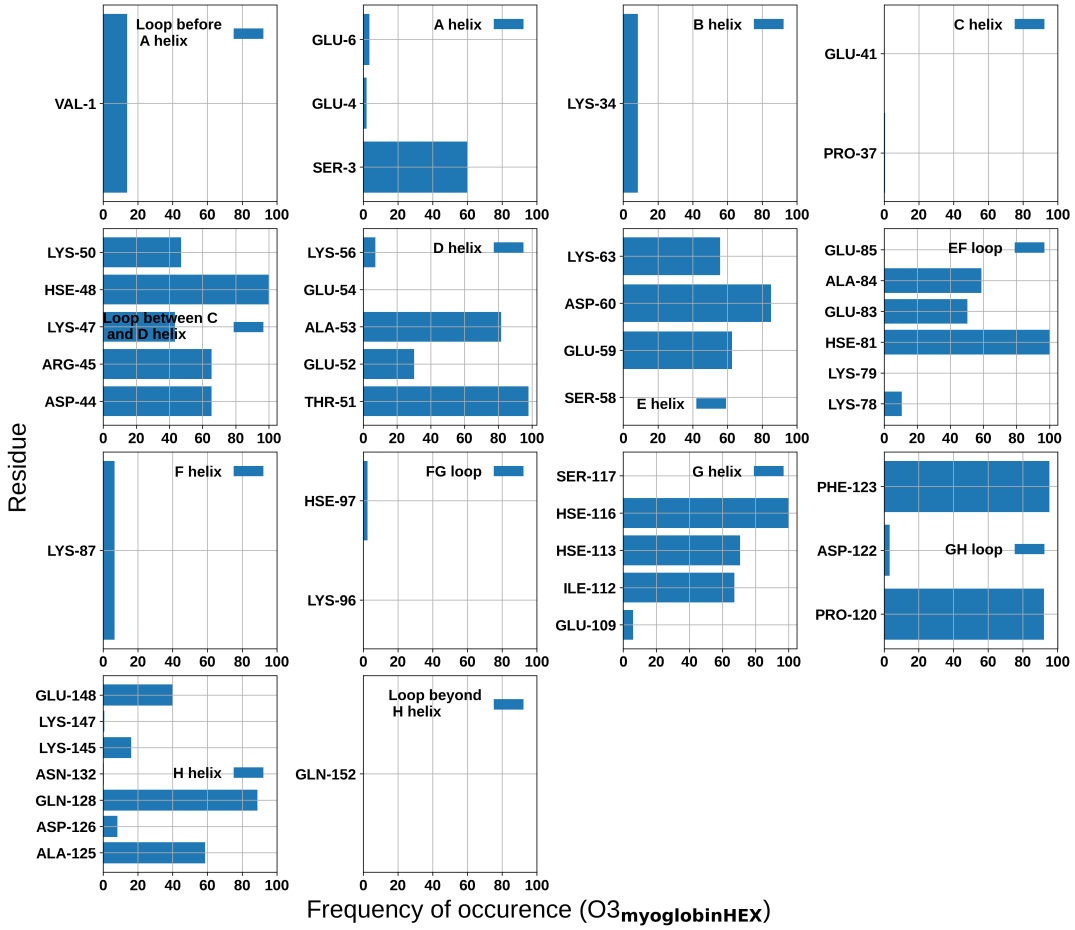


Fig. A.27: Residues for non-hydrogen atom contacts between myoglobin and IRMOF-74-VII-hex surface with a cutoff of 4 Å for orientation 3.

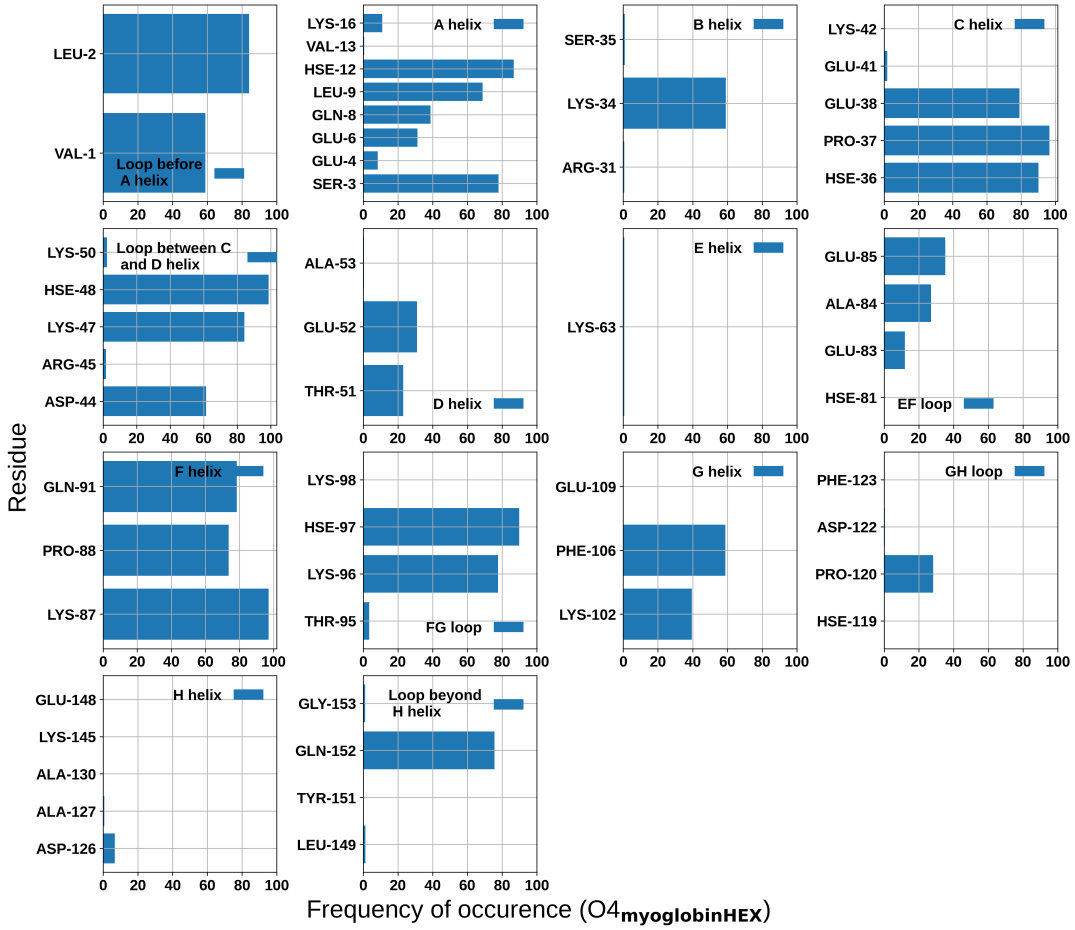


Fig. A.28: Residues for non-hydrogen atom contacts between myoglobin and IRMOF-74-VII-hex surface with a cutoff of 4 Å for orientation 4.

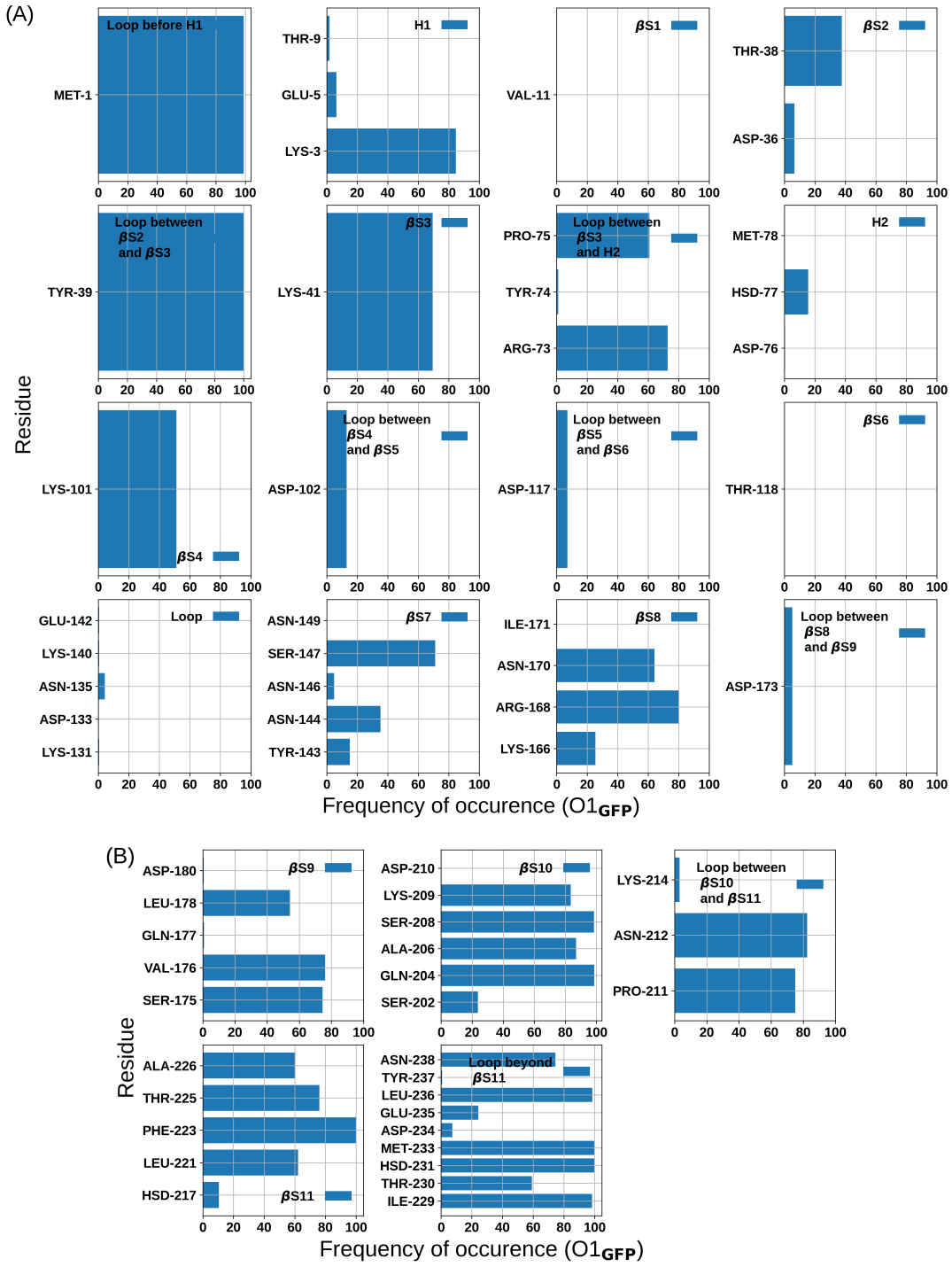


Fig. A.29: Residues for non-hydrogen atom contacts between GFP and IRMOF-74-IX surface with a cutoff of 4 Å for orientation 1. (A) and (B) for different structural elements.

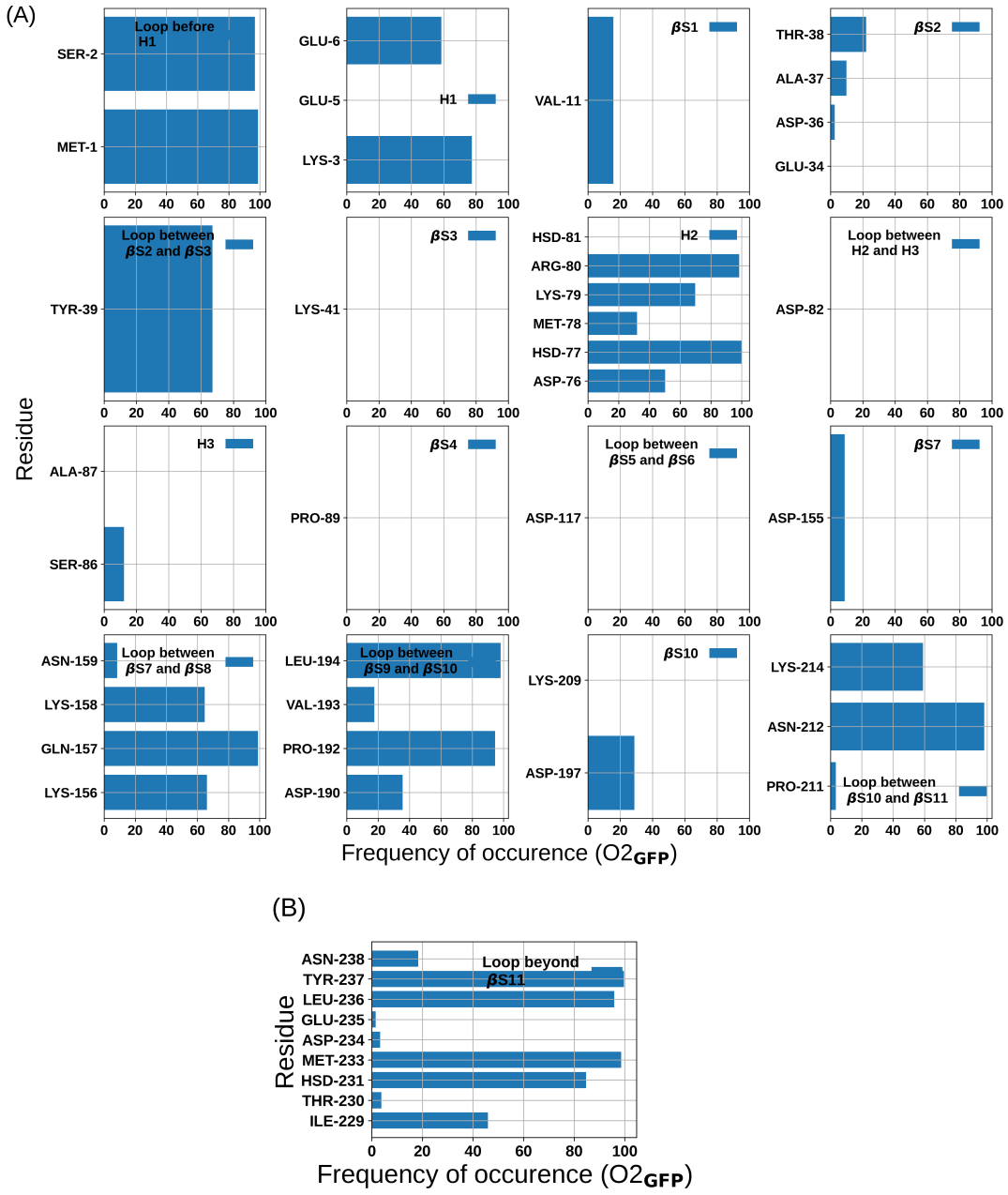


Fig. A.30: Residues for non-hydrogen atom contacts between GFP and IRMOF-74-IX surface with a cutoff of 4 Å for orientation 2. (A) and (B) for different structural elements.

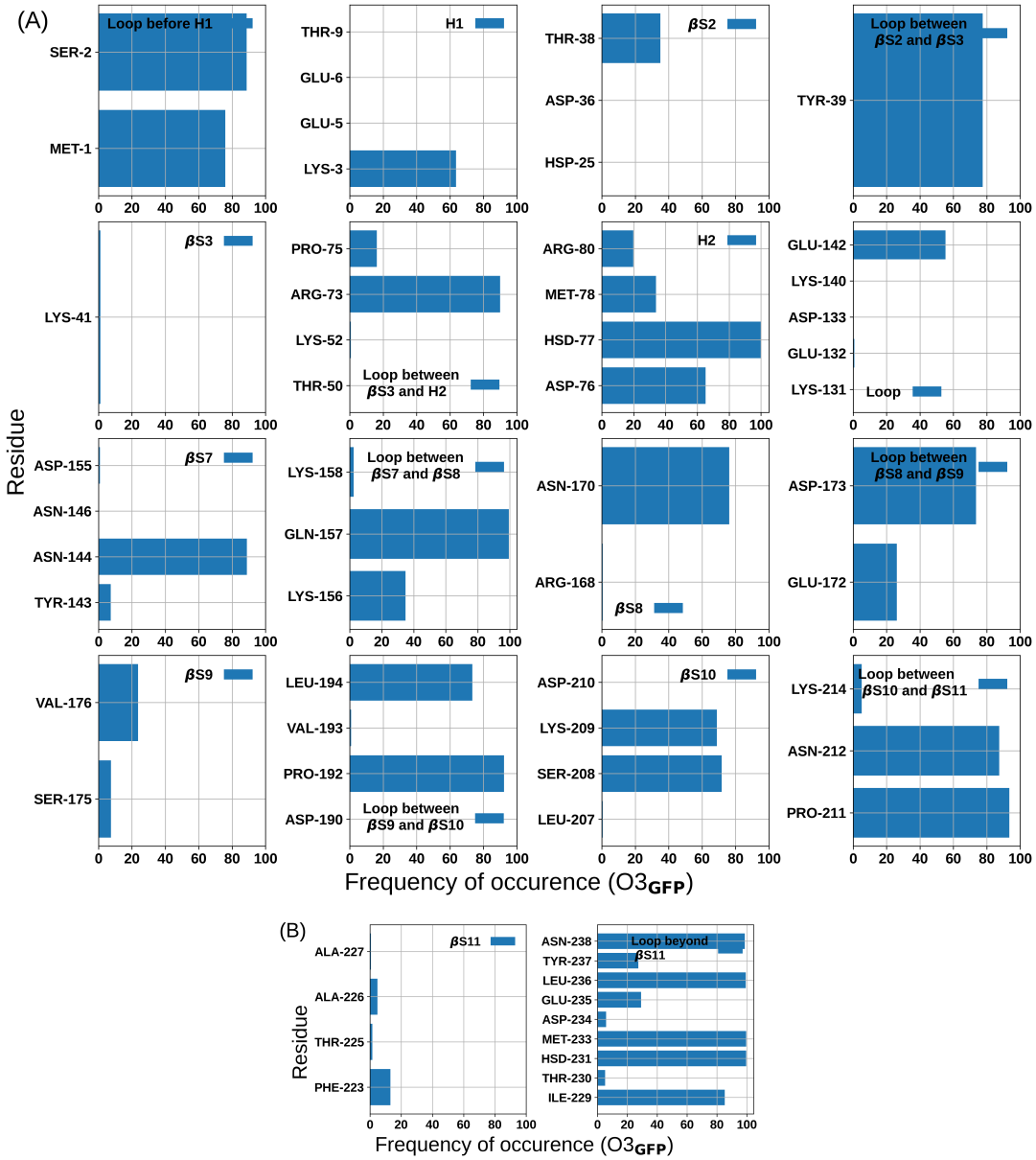


Fig. A.31: Residues for non-hydrogen atom contacts between GFP and IRMOF-74-IX surface with a cutoff of 4 Å for orientation 3. (A) and (B) for different structural elements.

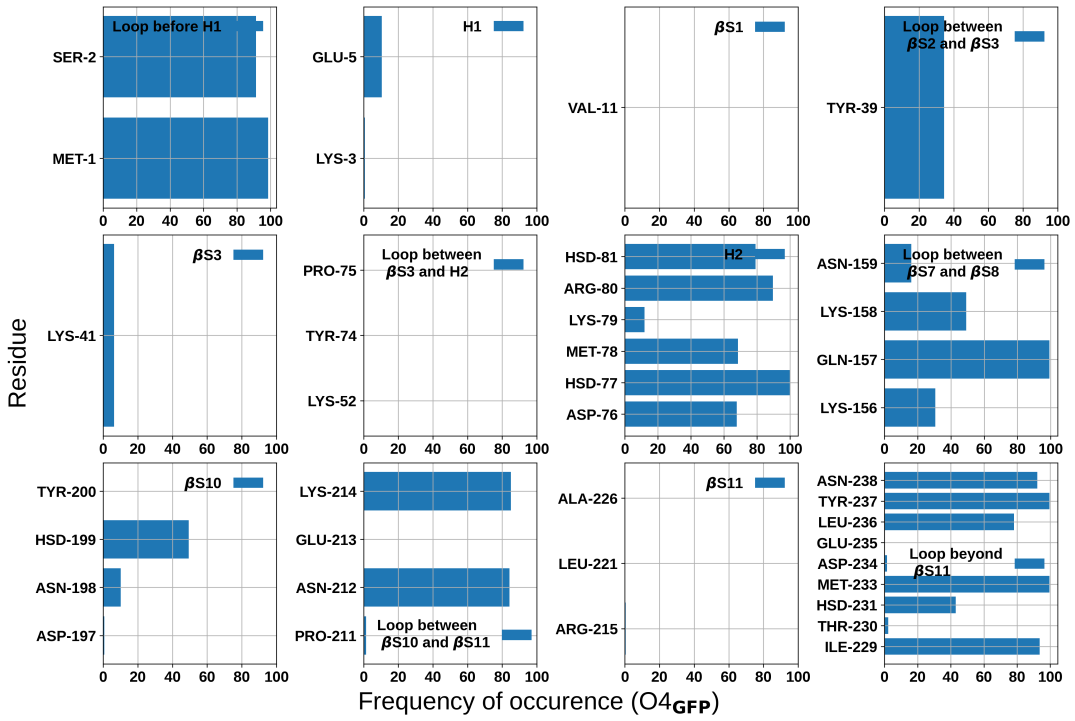


Fig. A.32: Residues for non-hydrogen atom contacts between GFP and IRMOF-74-IX surface with a cutoff of 4 Å for orientation 4.

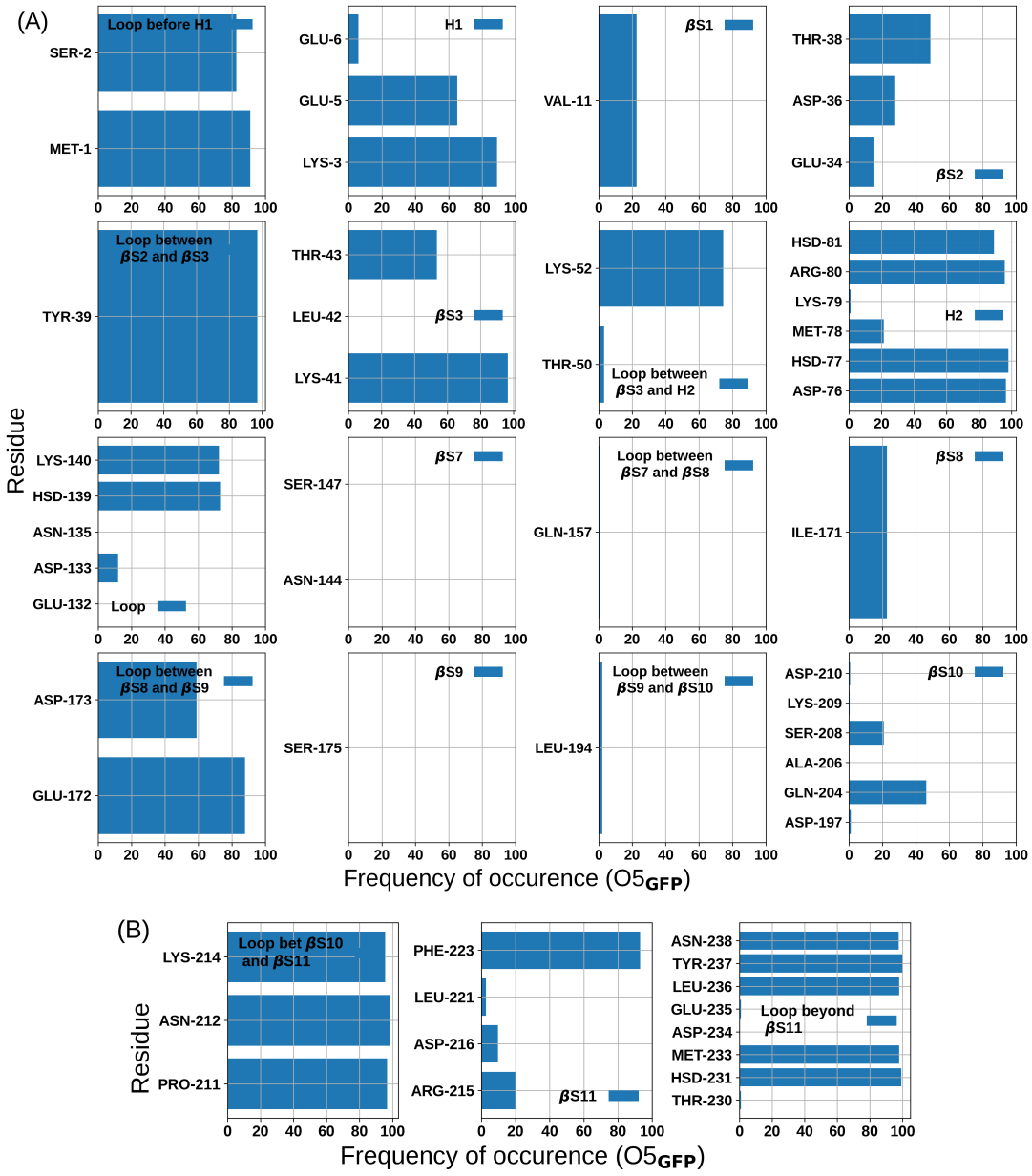


Fig. A.33: Non-hydrogen atom contacts between GFP and IRMOF-74-IX surface with a cutoff of 4 Å for orientation 5. (A) and (B) for different structural elements.

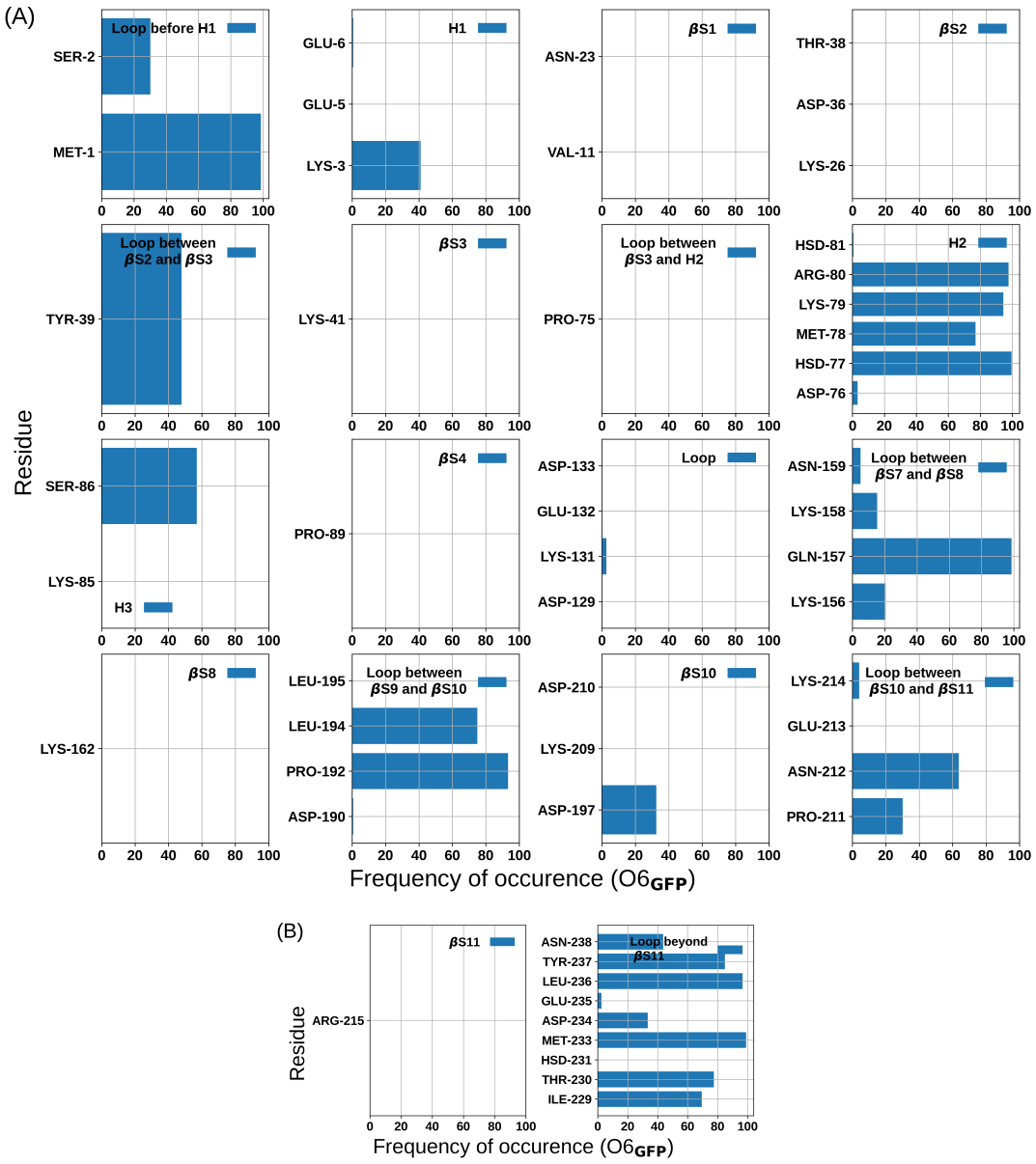


Fig. A.34: Non-hydrogen atom contacts between GFP and IRMOF-74-IX surface with a cutoff of 4 Å for orientation 6. (A) and (B) for different structural elements.

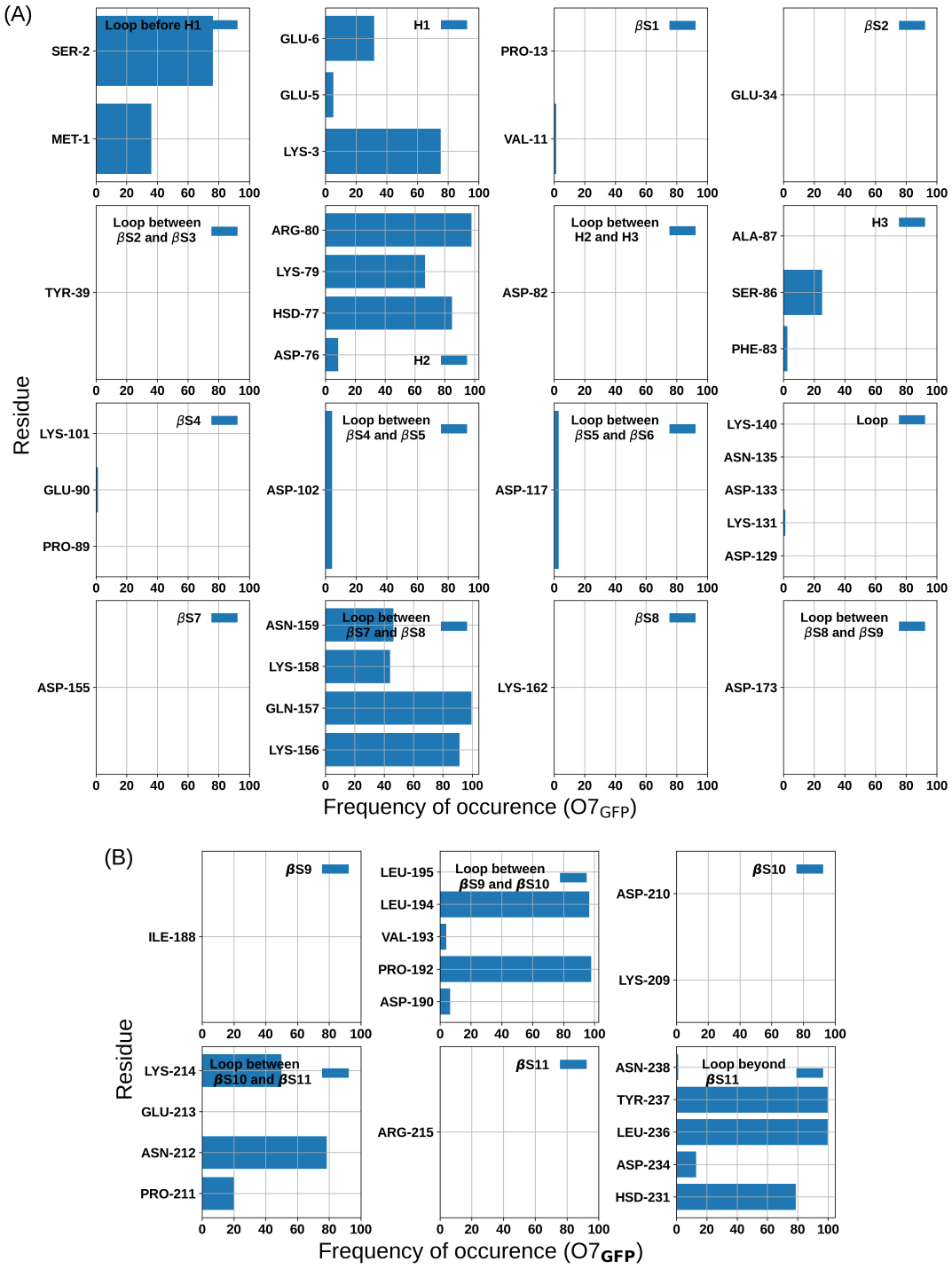


Fig. A.35: Non-hydrogen atom contacts between GFP and IRMOF-74-IX surface with a cutoff of 4 Å for orientation 7. (A) and (B) for different structural elements.

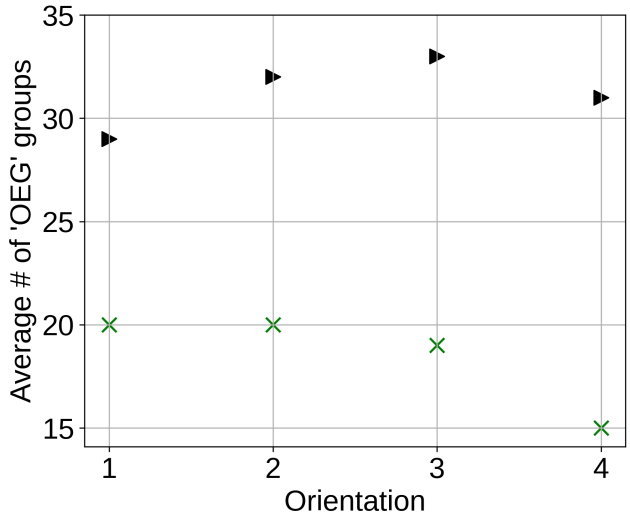


Fig. A.36: 'OEG' groups (SideChains of IRMOF-74-VII-oeg) in contact with myoglobin surface (non-hydrogen atoms within 4 Å) across four orientation for myoglobin@IRMOF-74-VII-oeg. The total number of OEGs is shown in Black and the same with occurrence greater than 10% of the analysis trajectory is shown in Green.

A.15 Hydrogen bonding interaction

Table A.12: Average number of hydrogen bonds between myoglobin and IRMOF-74-VII-oeg.

O1	O2	O3	O4
2.97	2.07	4.18	4.84

Table A.13: Average number of protein-solvent hydrogen bonds in myoglobin encapsulated systems.

System	O1	O2	O3	O4	Average
@water	-	-	-	-	332.45
@IRMOF-74-VII-oeg	311.91	319.52	324.37	312.18	317.00
@IRMOF-74-VII-hex	325.09	320.58	314.63	342.73	325.76

Table A.14: Average number of protein-solvent hydrogen bonds in GFP encapsulated system.

System	O1	O2	O3	O4	O5	O6	O7	Average
@water	-	-	-	-	-	-	-	564.14
@IRMOF-74-IX	559.12	572.31	553.85	543.49	558.99	543.18	562.31	556.18

A.16 Supplementary Movie

For visualization of the systems, a set of .mp4 files are available at https://drive.google.com/drive/folders/1QfVn1lDOlAdUBNaCTWRwBC5KUGzaBjOc?usp=sharing. myoglobinOEG1..4.mp4 and myoglobinHEX1..4.mp4 files are for corresponding analysis windows in four different orientations of myoglobin@IRMOF-74-VII-oeg and myoglobin@IRMOF-74-VII-hex simulations, respectively. Myoglobin is shown in NewCartoon representation in Red. The prosthetic group (HEME) is shown in Licorice representation. Organic linkers of MOF are shown in Line representation. Metal atoms of SBU (MOF) has been shown in vdW representation with reduced sphere scale in Pink. Metal ligated waters are shown in Licorice representation. Number of protein contacts with MOF (non-hydrogen atoms within 4 Å) are shown in Blue and Magenta for atoms and the corresponding residues, respectively. Waters in the systems are shown in Line representation (appear as dots) in Iceblue color.

GFPO1..7.mp4 files are for complete simulations (i.e. from the 0th step of energy minimization until production) in seven different orientations of GFP@IRMOF-74-IX, respectively. GFP is shown in NewCartoon representation in Green. Chromophore is shown in Licorice representation in Green. The remaining descriptions are the same as in myoglobin movies.

While preparing the movie files, the MOF orientation has been kept the same with respect to the symmetry of the pore axis. In myoglobin, MOF have been shown in tilted orientations whereas for GFP, not.

B

Supplementary Information for Chapter 3

B.1 Definitions and abbreviations

P_{COM} or the center of mass of the protein was determined considering all the atoms of the protein, HP35. C_{COM} or the Center of mass of the cavity was determined considering the μ_3-O atoms of that particular cavity of MIL-101(Cr), which contained the HP35. $W_{hexagonal}$ or the center of mass of the hexagonal window of MIL-101(Cr) (chosen for translocation) was determined considering the μ_3-O atoms of the hexagonal window. S_{NMR} structure was the lowest energy NMR structure (the first structure in the PDB file, 1UNC). The helicity (or helical content) of the protein was calculated considering residues 2-34 of HP35.

For most of the VMD representations, only the protein-containing cavity was shown instead of showing the complete supercell for clarity.

B.2 Toolbox for the simulations and data analysis

Superposition of protein structures was done using PyMOL (version: 2.5.0) [266]. Analyses were carried out using GROMACS (version: 2022.3) routines, Tcl (VMD version: 1.9.3, 1.9.4a38), PLUMED (version: 2.9.0-dev) [245, 267–269], Python (version: 3.10.4, 3.11.8, using Numpy (version: 1.22.4, 1.26.4) [270] library) and BASH. Data visualization was done using VMD [271, 272], Matplotlib (version: 3.5.1, 3.8.0) library [273] and OpenCV (version: 4.6.0, 4.7.0) library in Python. To calculate the helical content of the protein, we used the ALPHARMSD function with default parameters implemented in PLUMED.

B.2.1 Geometry optimization of the secondary building unit (SBU) using CP2K to obtain the optimum distance between the Cr atom and the O atom of the ligated water molecule.

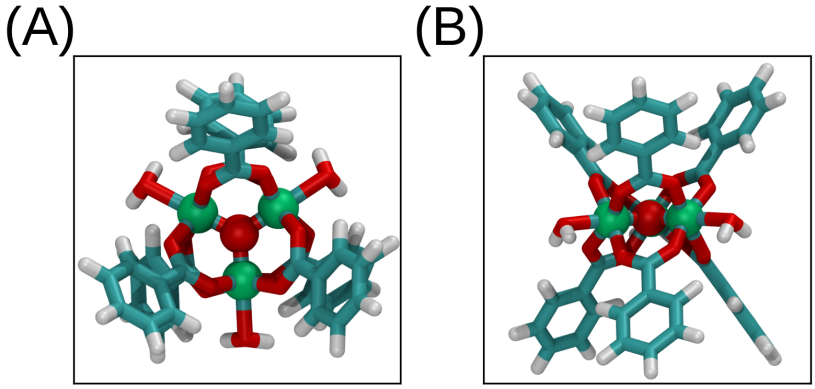


Fig. B.1: Geometry optimized structure of secondary building unit (SBU) of MIL-101 (Cr). Metal atoms (Cr) are shown in Green vdW representation. The metal ligated waters are at a distance of 2.30 Å from the metal center.

B.3 Visual representation illustrating the necessity of unfolding of HP35 at the hexagonal window of MIL-101(Cr)

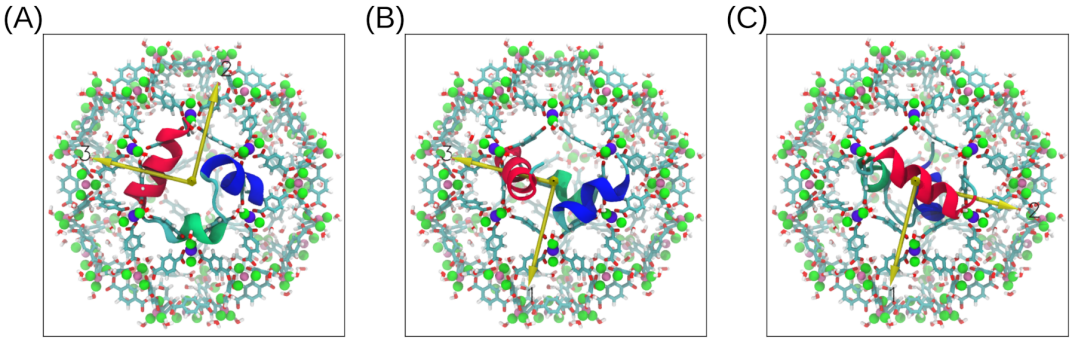


Fig. B.2: HP35 shown at the center of the hexagonal window to display its size relative to that of the hexagonal window and to demonstrate that it has to perforce unfold for translocation to the neighboring cage. Violet spheres are the μ_3-Os of the hexagonal window. (A) Orientation of the first principal axis of HP35 along the collinear vector joining C_{COM} and $W_{hexagonal}$. (B) and (C) are similar alignments with the second and third principal axes of HP35, respectively.

B.4 Results from Steered Molecular Dynamics (SMD) runs

We calculated work profiles for the 15 SMD runs, and these are presented in Figure B.3.

The SMD path was defined through a direction vector , which was calculated as the cross product of two edge vectors of the hexagonal window considering that the μ_3-Oxygens formed a hexagon. In Umbrella Sampling runs, the reaction coordinate was defined as the distance between C_{COM} and P_{COM}. This distance was calculated for the SMD run with the lowest work value (see Figure B.3 (A)), as shown in Figure B.4.

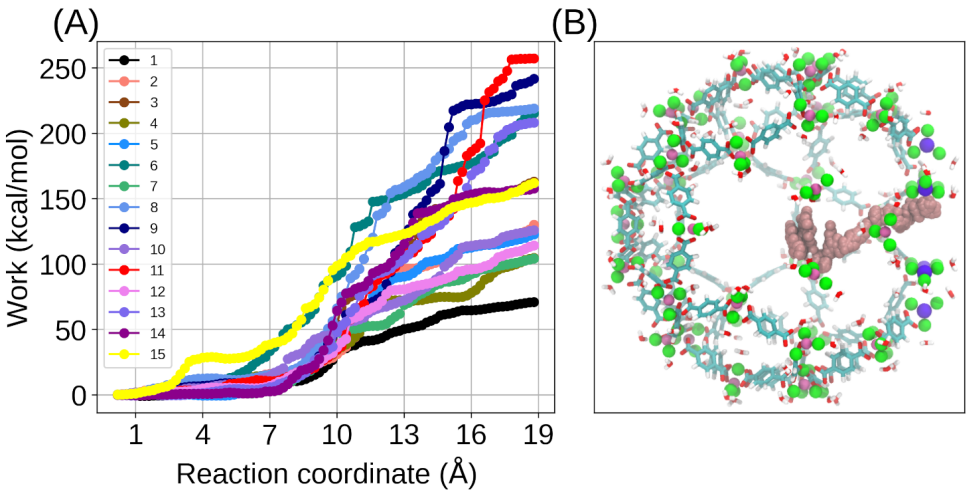


Fig. B.3: (A) Non-equilibrium work profiles calculated for the 15 SMD trajectories. (B) The path traversed by the protein during the SMD run corresponds to the profile with the lowest work shown in panel (A) (Black). P_{COM} is drawn with spheres in Pink from the center of the cavity towards the hexagonal window. Some of the windows of the MOF in the foreground are not shown for clarity.

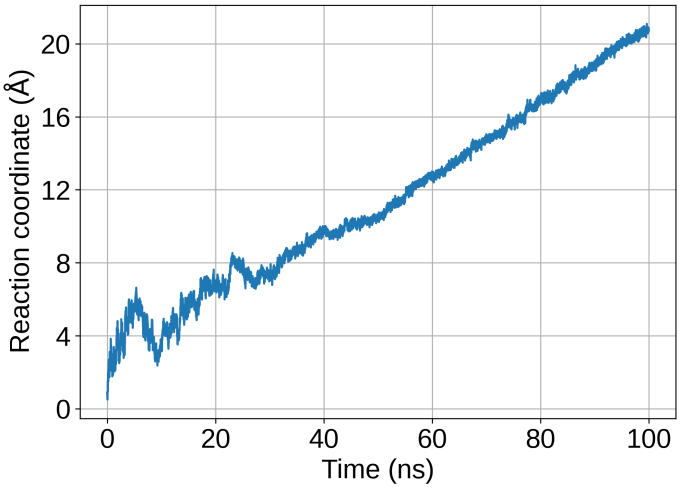


Fig. B.4: Reaction coordinate for SMD run.

B.5 Results from Umbrella Sampling (US)

The umbrella sampling results explained in the main manuscript could be followed through Figure B.5. We explain the constriction region first, followed by the region to its left and later the one on its right, i.e., the region closer to the hexagonal window.

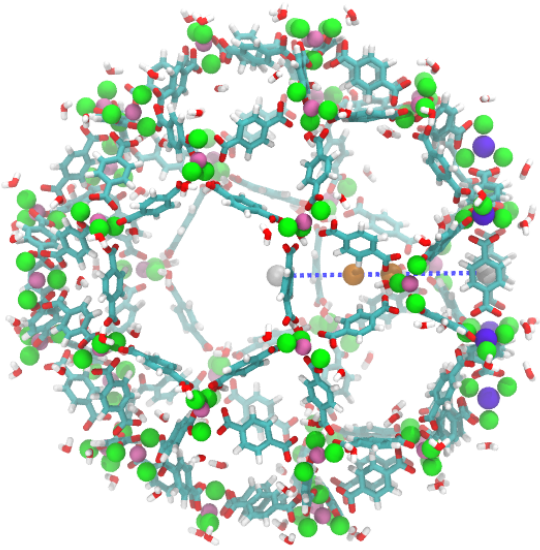


Fig. B.5: The cavity of MIL-101(Cr) containing HP35 is shown (protein has not been highlighted for clarity). Two White spheres join the C_{COM} and $W_{hexagonal}$ with a Blue dotted line. The two Orange spheres along the line bound the constriction region.

The reaction coordinate for the Umbrella Sampling run was the distance joining P_{COM} to P_{COM}, as defined earlier. At each value of the reaction coordinate, a harmonic restraining potential was applied for umbrella sampling simulations. The values of the reaction coordinate across the 54 umbrella windows are shown in Figure B.6.

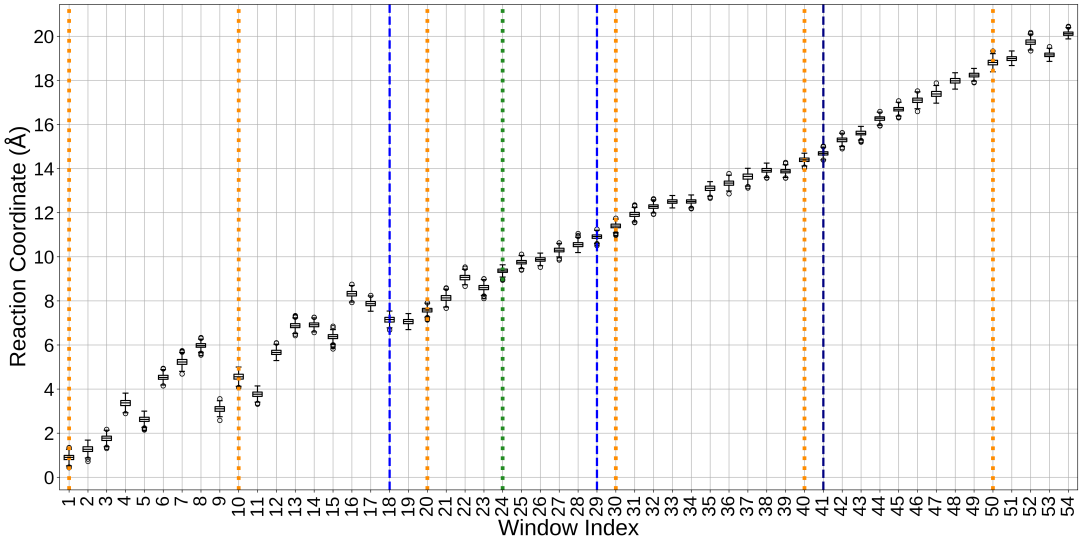


Fig. B.6: Reaction coordinate value vs umbrella window index. Six Orange dotted vertical lines correspond to windows for which protein structures over the last 5ns of umbrella sampling runs are shown separately in Figure B.7. The dashed vertical lines in the Blue box bound the constriction region (windows 18-29). The Green vertical dotted line is the minimum position seen in the PMF. The vertical dashed line in Navy denotes the upper limit of windows for the sub-zone (windows 30-41).

Protein conformations over the last five ns of the trajectory for 6 umbrella windows are shown in Figure B.7; the locations of these windows are indicated through Orange dotted lines in Figure B.6.

The run lengths in different umbrella windows are tabulated in Table B.1.

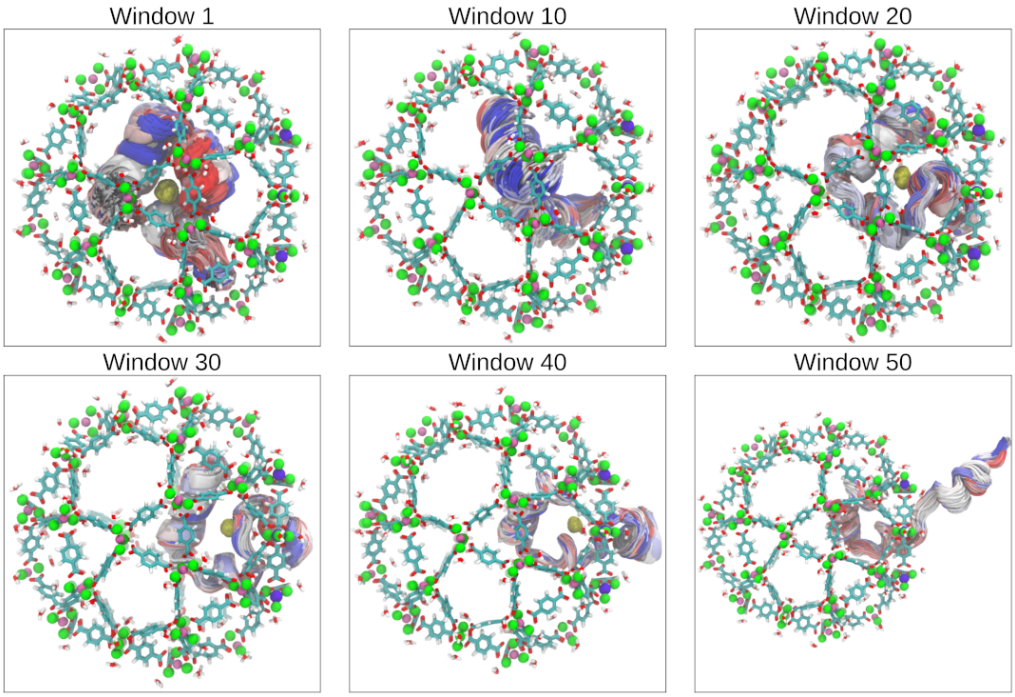


Fig. B.7: Protein conformations inside the MOF cavity at various umbrella window locations. P_{COM} are shown as Yellow spheres.

Table B.1: Details of Umbrella Sampling windows

Window Index	Run Length (ns)	Analysis window range (ns)
1	300.00	295.00-300.00
2	100.00	95.00-100.00
3	100.00	95.00-100.00
4	140.98	135.98-140.98
5	261.76	256.76-261.76
6	275.63	270.63-275.63
7	100.00	95.00-100.00

Continued to the next page

Continued from previous page

Window Index	Run Length (ns)	Analysis window range (ns)
8	225.93	220.93-225.93
9	254.62	249.62-254.62
10	132.51	127.51-132.51
11	100.00	95.00-100.00
12	50.00	45.00-50.00
13	200.00	195.00-200.00
14	50.00	45.00-50.00
15	141.00	136.00-141.00
16	200.00	195.00-200.00
17	100.00	95.00-100.00
18	181.67	176.67-181.67
19	137.46	132.46-137.46
20	100.00	95.00-100.00
21	73.97	68.97-73.97
22	181.92	176.92-181.92
23	100.00	95.00-100.00
24	90.77	85.77-90.77
25	50.00	45.00-50.00
26	90.69	85.69-90.69
27	90.89	85.89-90.89

Continued to the next page

Continued from previous page

Window Index	Run Length (ns)	Analysis window range (ns)
28	50.00	45.00-50.00
29	100.00	95.00-100.00
30	100.00	95.00-100.00
31	100.00	95.00-100.00
32	100.00	95.00-100.00
33	177.86	172.86-177.86
34	100.00	95.00-100.00
35	73.63	68.63-73.63
36	71.08	66.08-71.08
37	50.00	45.00-50.00
38	50.00	45.00-50.00
39	81.98	76.98-81.98
40	90.88	85.88-90.88
41	50.00	45.00-50.00
42	84.19	79.19-84.19
43	170.64	165.64-170.64
44	170.13	165.13-170.13
45	74.03	69.03-74.03
46	100.00	95.00-100.00
47	134.61	129.61-134.61

Continued to the next page

Continued from previous page

Window Index	Run Length (ns)	Analysis window range (ns)
48	75.28	70.28-75.28
49	75.17	70.17-75.17
50	73.89	68.89-73.89
51	100.00	95.00-100.00
52	179.35	174.35-179.35
53	132.56	127.56-132.56
54	145.85	140.85-145.85

End of Table

Umbrella histograms are shown in Figure B.8.

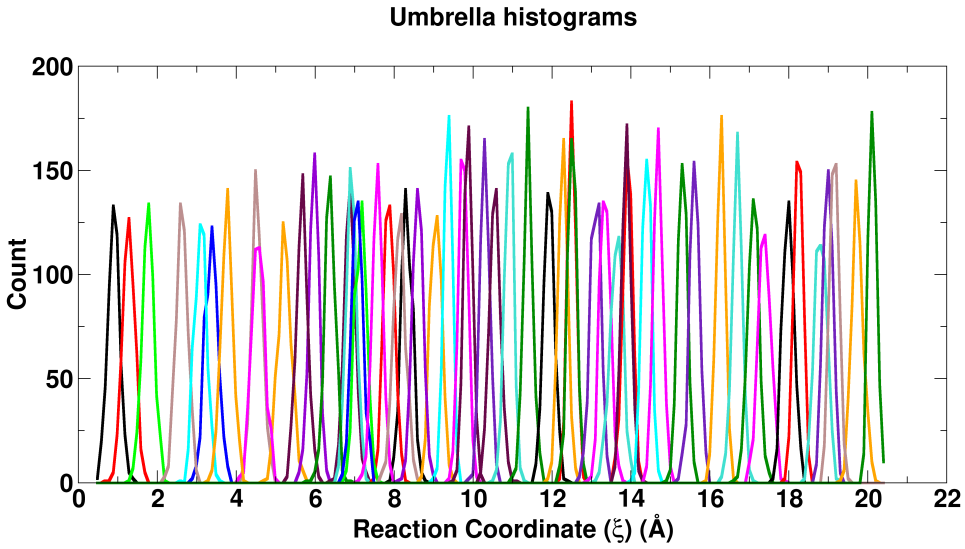


Fig. B.8: Biased probability distribution across umbrella windows.

Backbone RMSD of the protein over the complete trajectory in each umbrella window are shown in Figure B.9.

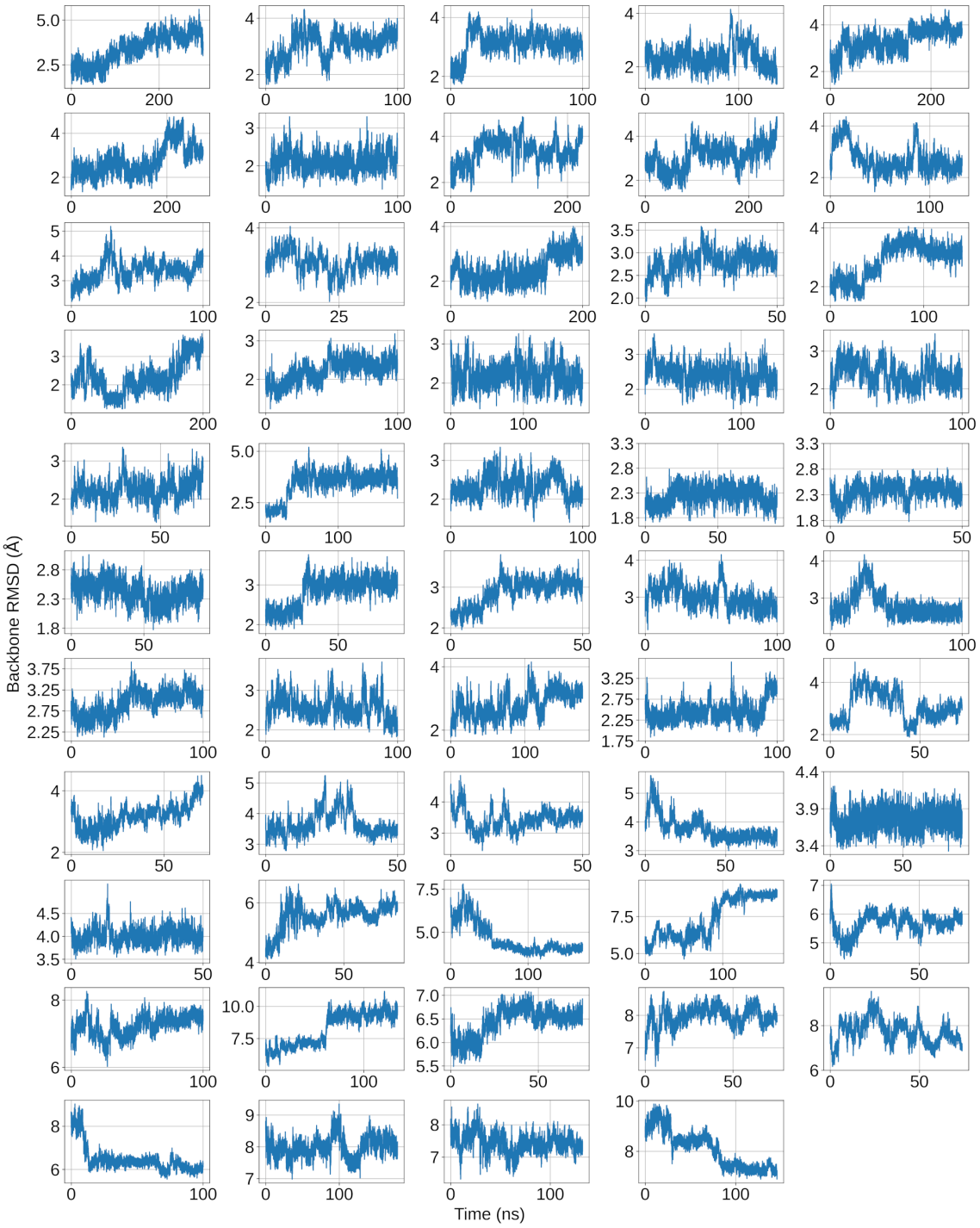


Fig. B.9: Backbone RMSD w.r.t. experimental NMR structure for all the windows over the entire trajectory length.

Solvent accessible surface area of the hydrophobic core of HP35 is shown in Figure B.10.

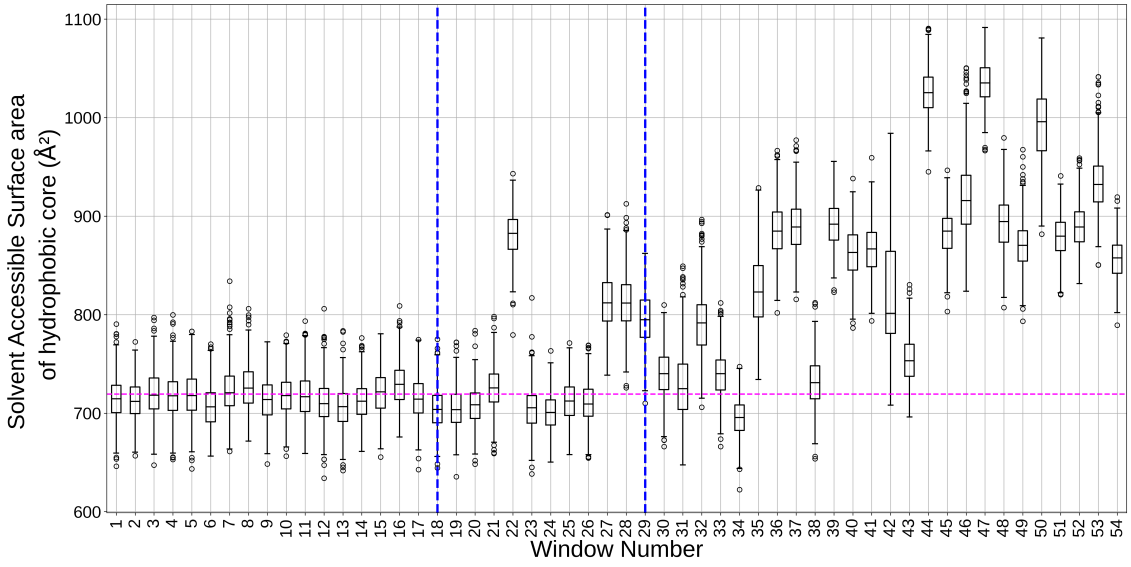


Fig. B.10: Solvent accessible surface area of three F residues of HP35 across umbrella windows. Blue vertical dotted lines box the constriction region. The Magenta horizontal dotted line represents the SASA value of the same hydrophobic core in S_{NMR}.

Protein-MOF interaction energies across umbrella windows are displayed in Figures B.11, and B.12.

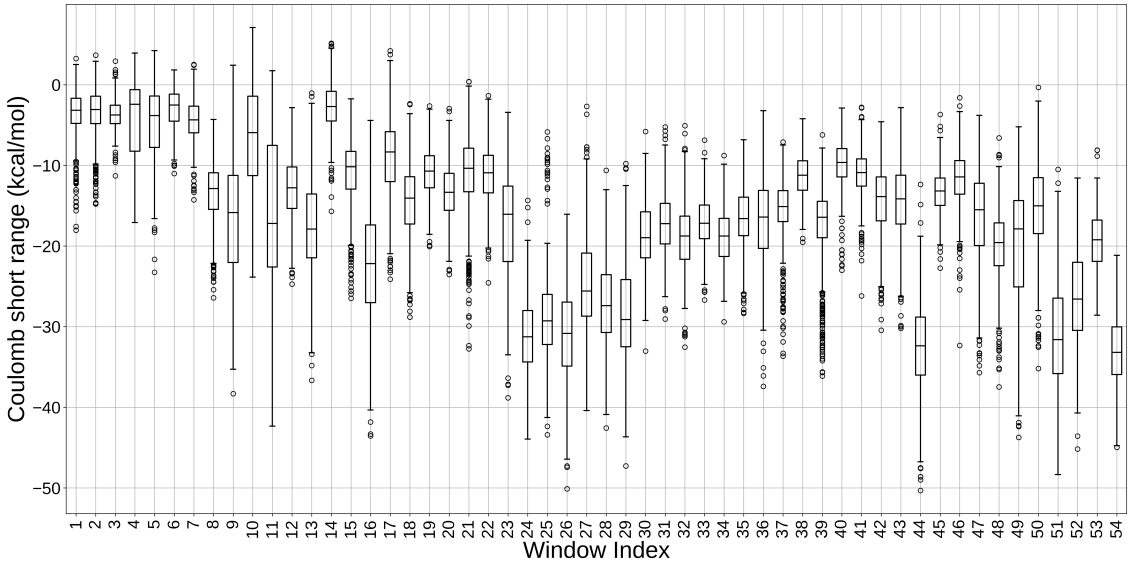


Fig. B.11: Protein-MOF interaction Energy (Coulomb SR) across umbrella windows

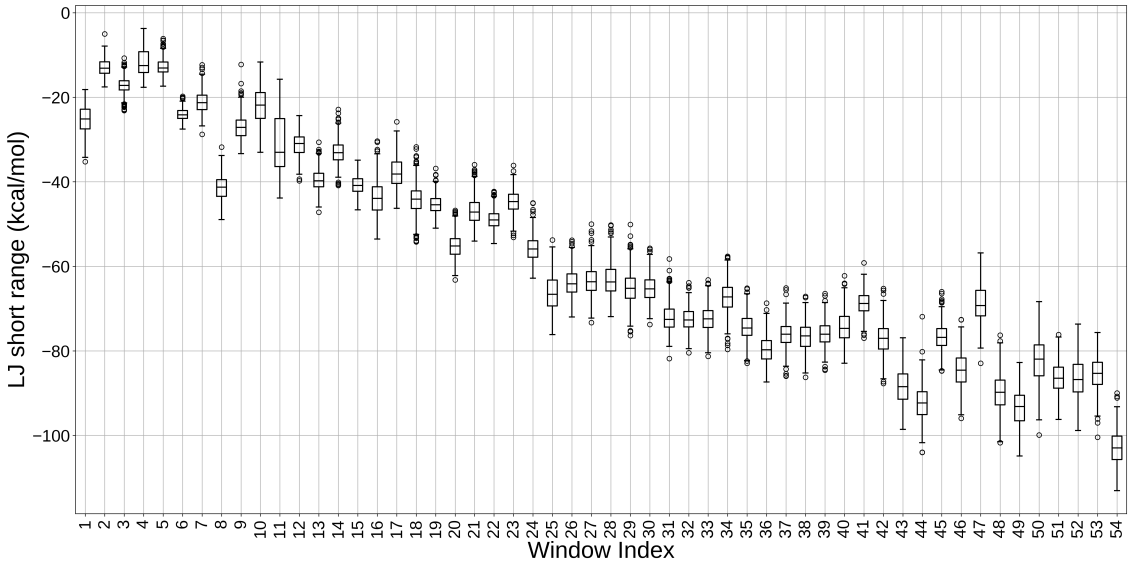


Fig. B.12: Protein-MOF interaction Energy (LJ SR) across umbrella windows

B.6 Supplementary Movie

The movie file is us.mp4 (https://drive.google.com/drive/folders/1maop3-IpFjIbZHz8nCqxJk5R_DjUJUJt?usp=sharing). This is not a "real" movie corresponding to a molecular dynamics simulation. Instead, it was made by concatenating snapshots of the protein from the last time frame of each umbrella sampling window. Helix-1 of the protein is shown in Blue, helix-2 in Green, and

helix-3 in Red. Cavities containing the protein and its neighbour are shown in Licorice representation with the left cavity in full description (Cr atoms in Green, μ_3-O in Pink vdW representation with reduced sphere scale and the metal-ligated waters in Licorice representation.) The μ_3-O of the hexagonal window are shown in Violet vdW representation with reduced sphere scale. Two White fictitious spheres are drawn at the center of the cavity and the hexagonal window; a Blue straight line joins them. The zigzag path from the concatenation of the snapshots from the umbrella windows is shown in Blue, while the path followed by steered MD is shown in Yellow. Non-hydrogen atoms of the protein within four Å of non-hydrogen atoms of MOF are shown in Cyan vdW representation with the complete residue in Cyan Licorice representation. These get updated at every window. Although the MOF structure may appear rigid in this "movie", in reality, in the simulations, they are part of the degrees of freedom. For clarity, the MOF structure is shown as rigid in this movie.

B.7 Code availability

The input files of our simulations are available in the GitHub repository: https://github.com/Oishika-1/CoTranslocationalUnfolding.

C

Supplementary Information for Chapter 4

C.1 QM region definition

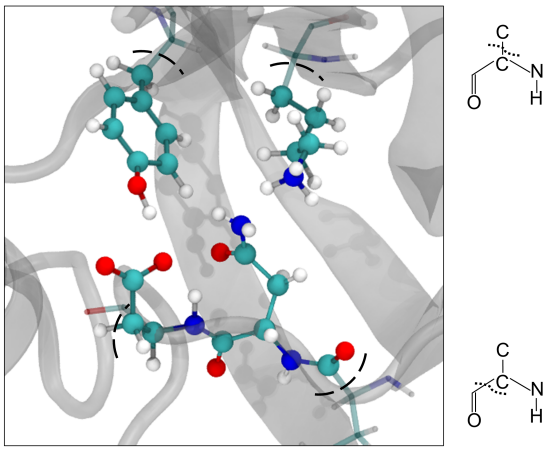


Fig. C.1: Part of the enzyme that was treated at QM level. Four covalent (non-polarizable) bonds were broken at the QM-MM boundary (was described through link atoms). They were C_α-C_β and C_α-C(=O).

C.2 Collective Variable space:

The initial conformations in the CV space (Table C.1).

Table C.1: Values of collective variables for the three initial conformations.

Conformations	CV_1	CV_1
1	-0.40	-0.36
2	-0.49	-0.48
3	-0.50	-0.49

The plots show how the collective variable space was sampled during the reaction for all the runs (Figures C.2, and C.3).

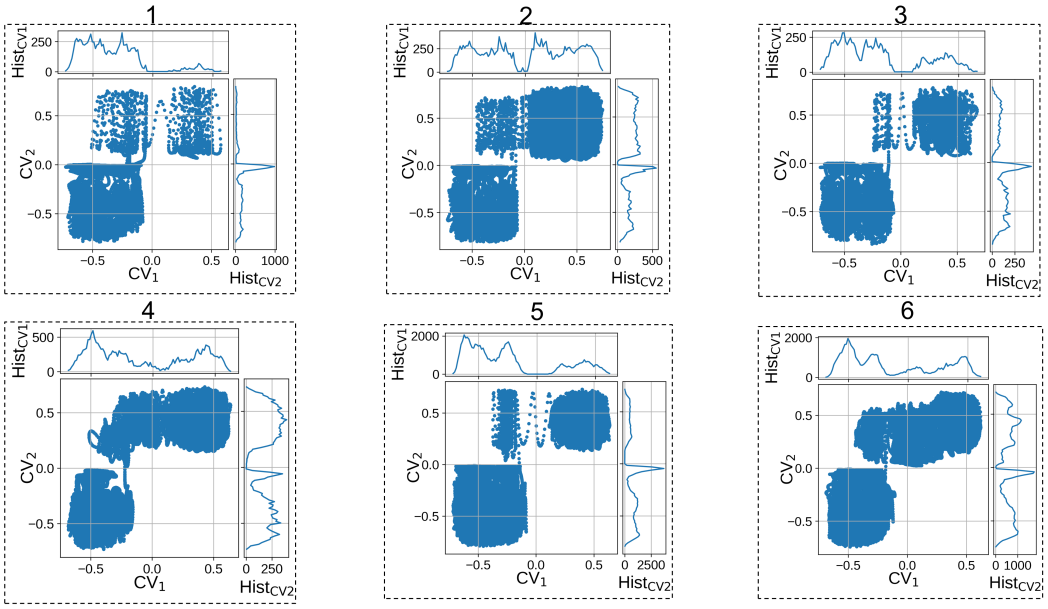


Fig. C.2: Sampling of conformations in the collective variable space. 1 to 6 stand for the first six runs.

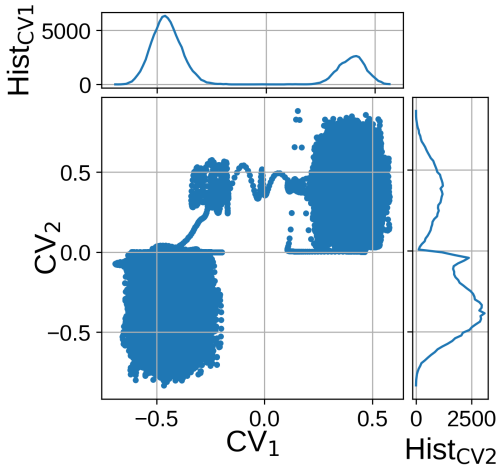


Fig. C.3: Sampling of conformations in the collective variable space for the seventh run with the redefined conservative Gaussian bias parameters.

C.3 Characterising the 109th residue through backbone and side chain dihedrals.

Ramachandran plots and Janin plots for residue 109 are shown in Figures C.4, C.5, C.6, and C.7. 109_{ASN} and 109_{SNN} titled plots are results from the classical molecular dynamics simulations with asparagine and succinimide at the 109th position, respectively. 1 to 6

denoted the six QM/MM runs. Yellow dotted regions highlight product zone. The product zone is defined with a distance cutoff (upper bound) of 1.75 Å between the backbone nitrogen of residue 110 and the side chain carbonyl-carbon of residue 109. In the case of Ramachandran plots, the product states from our simulations are mainly in the third quadrant, unlike the experimental product state in the fourth quadrant (which is reflected from the 109_{SNN} plot). However, our collaborators have observed that succinimide alone (from the PDB database) cluster in the Ramachandran plot's third and fourth quadrants (Chandrashekarmath et al., unpublished). Thus, we think the movement from the third to the fourth quadrant could happen after the product has formed. We could not observe this in our simulations, as the product basin was only partially filled. In the case of the side chain dihedrals (Figure C.6, and C.7), product basins from our QM/MM runs sampled the succinimide state partly.

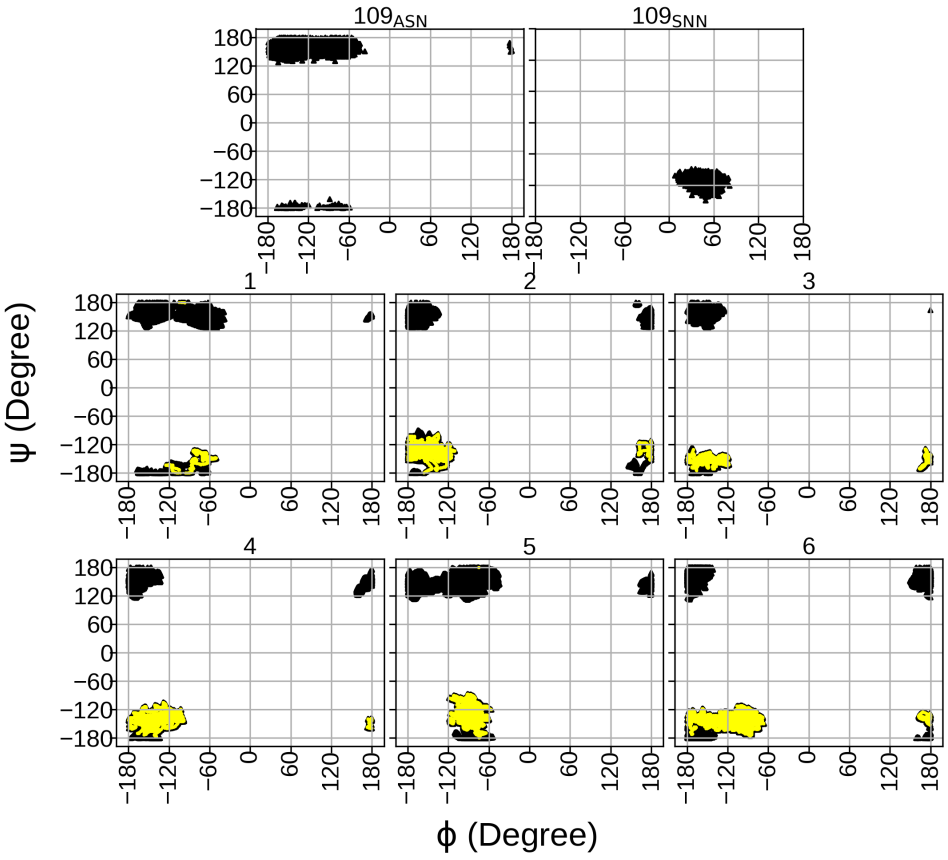


Fig. C.4: Ramachandran plots for the 109$^{\text{th}}$ residue in the classical molecular dynamics and QM/MM runs (1 to 6).

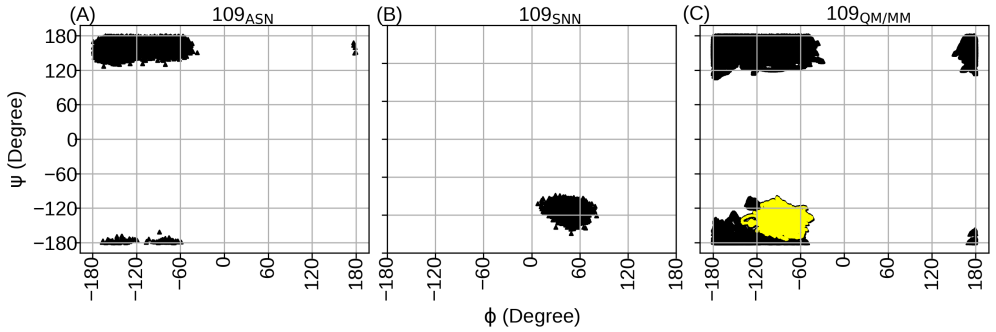


Fig. C.5: Ramachandran plot for 109$^{\text{th}}$ residue for QM/MM simulation with the redefined, conservative parameters (Run 7) along with 109_{ASN} and 109_{SNN}.

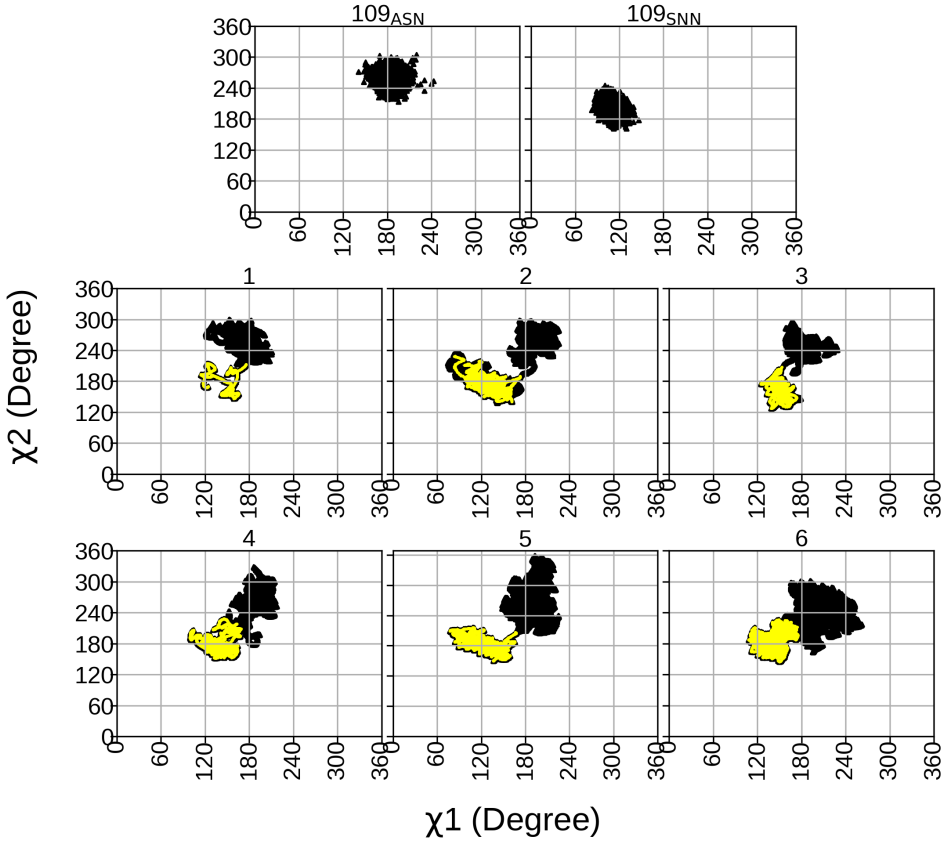


Fig. C.6: Janin plots for 109$^{\text{th}}$ residue in the classical molecular dynamics and QM/MM runs (1 to 6).

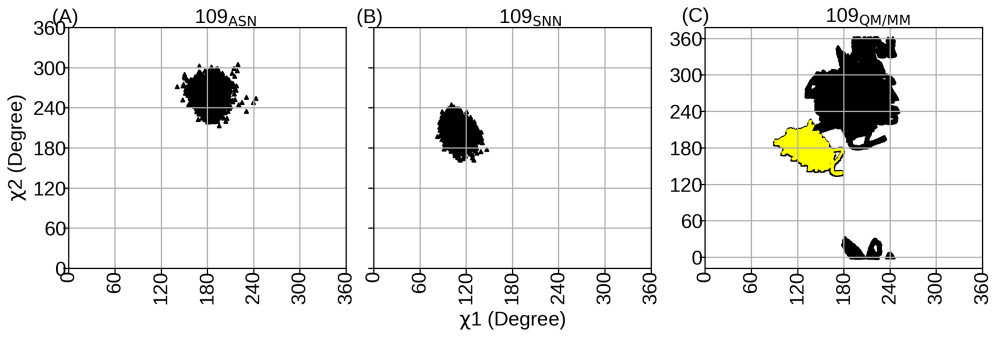


Fig. C.7: Janin plot for 109^{th} residue for QM/MM simulation with the redefined, conservative parameters (Run 7) along with 109_{ASN} and 109_{SNN}.

C.4 Calculation of minimum free energy path

This is carried out with the help of a GUI based software http://bioweb.cbm.uam.es/software/MEPSA/ which uses a similar algorithm [274] as Dijkstra's [275] to produce a minimum value path, in this case, in the free energy landscape.

C.5 Possible reasons for not observing a converged free energy surface

For all the seven QM/MM MD MTD runs, we could not 'see' the backward reaction (from product state to reactant state) as the jobs crashed with QM atoms going out of the QM box. We think there are two reasons for this not filling up of the product basin - (a) plane wave description and (b) use of polarized basis sets for the QM atoms. Both of these are known to cause spilling of electron densities at the boundary out into the MM region. A possible solution could be to use a Gaussian distribution of charges for the MM atoms or to go for a polarization embedding scheme.

C.6 Supplementary Movie

For visualization purposes, the movie file corresponding to QM/MM Run 7 has been provided: https://drive.google.com/drive/folders/1dmEbYcDjfDQnFDsQDvq_IthEq4mt9vIa?usp=sharing. The reaction center is highlighted with CPK representation. Hydrogen bonds (with donor-acceptor distance cutoff 3.5 Å and donor-hydrogen-acceptor angle cutoff 140°) are shown in Magenta Springs. From the spectator region, protein is (partly) shown in a White, Transparent New Cartoon representation. Water molecules and ions are not shown for clarity.

C.7 Code availability

The input files for classical molecular dynamics simulations and QM/MM Run 7 are available in the GitHub repository: https://github.com/Oishika-1/QM-MM_MjGATase_Succinimide.

Bibliography

[1] Karplus, M. The Levinthal paradox: yesterday and today. *Folding and Design* **1997**, *2*, S69–S75.

[2] Born, M.; Oppenheimer, R. Zur Quantentheorie der Molekeln. *Annalen der Physik* **1927**, *389*, 457–484.

[3] Silva-Feaver, M. Born Oppenheimer Approximation. 2018.

[4] Ercolessi, F. A molecular dynamics primer. *Spring College in Computational Physics, ICTP, Trieste* **1997**, *19*.

[5] Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics* **1980**, *72*, 2384–2393.

[6] MacKerell Jr, A. D.; Bashford, D.; Bellott, M.; Dunbrack Jr, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; et al., All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B* **1998**, *102*, 3586–3616.

[7] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society* **1995**, *117*, 5179–5197.

[8] Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard III, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society* **1992**, *114*, 10024–10035.

[9] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79*, 926–935.

[10] Berendsen, H. J.; Grigera, J.-R.; Straatsma, T. P. The missing term in effective pair potentials. *Journal of Physical Chemistry* **1987**, *91*, 6269–6271.

[11] Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **2007**, *126*.

[12] Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **1981**, *52*, 7182–7190.

[13] Pohorille, A.; Jarzynski, C.; Chipot, C. Good practices in free-energy calculations. *The Journal of Physical Chemistry B* **2010**, *114*, 10235–10253.

[14] Kästner, J. Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 932–942.

[15] Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 826–843.

[16] Jarzynski, C. Nonequilibrium equality for free energy differences. *Physical Review Letters* **1997**, *78*, 2690.

[17] Crooks, G. E. Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. *Journal of Statistical Physics* **1998**, *90*, 1481–1487.

[18] Hummer, G.; Szabo, A. Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proceedings of the National Academy of Sciences* **2001**, *98*, 3658–3661.

[19] Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry* **1992**, *13*, 1011–1021.

[20] Souaille, M.; Roux, B. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Computer Physics Communications* **2001**, *135*, 40–57.

[21] Hub, J. S.; De Groot, B. L.; Van Der Spoel, D. g_wham– A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *Journal of Chemical Theory and Computation* **2010**, *6*, 3713–3720.

[22] Groenhof, G. In *Biomolecular Simulations: Methods and Protocols*; Monticelli, L., Salonen, E., Eds.; Humana Press: Totowa, NJ, 2013; pp 43–66.

[23] Senn, H. M.; Thiel, W. In *Atomistic Approaches in Modern Biology: From Quantum Chemistry to Molecular Simulations*; Reiher, M., Ed.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2007; pp 173–290.

[24] Mouvet, F.; Villard, J.; Bolnykh, V.; Rothlisberger, U. Recent advances in first-principles based molecular dynamics. *Accounts of Chemical Research* **2022**, *55*, 221–230.

[25] Carloni, P.; Rothlisberger, U.; Parrinello, M. The role and perspective of ab initio molecular dynamics in the study of biological systems. *Accounts of Chemical Research* **2002**, *35*, 455–464.

[26] Thar, J.; Reckien, W.; Kirchner, B. In *Atomistic Approaches in Modern Biology: From Quantum Chemistry to Molecular Simulations*; Reiher, M., Ed.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2007; pp 133–171.

[27] Senn, H. M.; Thiel, W. QM/MM methods for biomolecular systems. *Angewandte Chemie International Edition* **2009**, *48*, 1198–1229.

[28] Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log (N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **1993**, *98*, 10089–10092.

[29] Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **1997**, *18*, 1463–1472.

[30] Hess, B. P-LINCS: A parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation* **2008**, *4*, 116–122.

[31] Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford university press: Oxford, England, 2017.

[32] Frenkel, D.; Smit, B. *Understanding molecular simulation: From Algorithms to Applications*; Elsevier: Cambridge, Massachusetts, United States, 2023.

[33] Simm, S.; Einloft, J.; Mirus, O.; Schleiff, E. 50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification. *Biological Research* **2016**, *49*, 1–19.

[34] Genchi, G. An overview on D-amino acids. *Amino Acids* **2017**, *49*, 1521–1533.

[35] Martínez-Rodríguez, S.; Martínez-Gómez, A. I.; Rodríguez-Vico, F.; Clemente-Jiménez, J. M.; Las Heras-Vázquez, F. J. Natural occurrence and industrial applications of D-amino acids: An overview. *Chemistry & Biodiversity* **2010**, *7*, 1531–1548.

[36] Vickery, H. B. The origin of the word protein. *The Yale Journal of Biology and Medicine* **1950**, *22*, 387.

[37] Hartley, H. Origin of the Word 'Protein'. *Nature* **1951**, *168*, 244–244.

[38] Calvete, J. J.; Bini, L.; Hochstrasser, D.; Sanchez, J.-C.; Turck, N. The magic of words. *Journal of Proteomics* **2014**, *107*, 1–4, Special Issue: "20 years of Proteomics" in memory of Vitaliano Pallini.

[39] Staudinger, H. In *A Source Book in Chemistry, 1900–1950*; Leicester, H. M., Ed.; Harvard University Press: Cambridge, Massachusetts, United States, 1968; pp 259–264.

[40] Kyle, R. A.; Shampo, M. A. Theodor Svedberg and the Ultracentrifuge. *Mayo Clinic Proceedings* **1997**, *72*, 830.

[41] Anfinsen, C. B.; Haber, E.; Sela, M.; White Jr, F. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences* **1961**, *47*, 1309–1314.

[42] Levinthal, C. Are there pathways for protein folding? *Journal de Chimie Physique* **1968**, *65*, 44–45.

[43] Dill, K. A.; Chan, H. S. From Levinthal to pathways to funnels. *Nature Structural & Molecular Biology* **1997**, *4*, 10–19.

[44] Wetlaufer, D. B. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proceedings of the National Academy of Sciences* **1973**, *70*, 697–701.

[45] Karplus, M.; Weaver, D. L. Protein-folding dynamics. *Nature* **1976**, *260*, 404–406.

[46] Kim, P. S.; Baldwin, R. L. Intermediates in the folding reactions of small proteins. *Annual Review of Biochemistry* **1990**, *59*, 631–660.

[47] Harrison, S. C.; Durbin, R. Is there a single pathway for the folding of a polypeptide chain? *Proceedings of the National Academy of Sciences* **1985**, *82*, 4028–4030.

[48] Bryngelson, J. D.; Wolynes, P. G. Intermediates and barrier crossing in a random energy model (with applications to protein folding). *The Journal of Physical Chemistry* **1989**, *93*, 6902–6915.

[49] Leopold, P. E.; Montal, M.; Onuchic, J. N. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences* **1992**, *89*, 8721–8725.

[50] Chan, H. S.; Dill, K. A. Transition states and folding dynamics of proteins and heteropolymers. *The Journal of Chemical Physics* **1994**, *100*, 9238–9257.

[51] Zhou, H.-X.; Zwanzig, R. A rate process with an entropy barrier. *The Journal of Chemical Physics* **1991**, *94*, 6147–6152.

[52] Jumper, J.; et al., Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

[53] Tunyasuvunakool, K.; et al., Highly accurate protein structure prediction for the human proteome. *Nature* **2021**, *596*, 590–596.

[54] Ruff, K. M.; Pappu, R. V. AlphaFold and implications for intrinsically disordered proteins. *Journal of Molecular Biology* **2021**, *433*, 167208.

[55] David, A.; Islam, S.; Tankhilevich, E.; Sternberg, M. J. The AlphaFold database of protein structures: a biologist's guide. *Journal of Molecular Biology* **2022**, *434*, 167336.

[56] Chakravarty, D.; Porter, L. L. AlphaFold2 fails to predict protein fold switching. *Protein Science* **2022**, *31*, e4353.

[57] Ramachandran, G.; Sasisekharan, V. Conformation of Polypeptides and Proteins. *Advances in Protein Chemistry* **1968**, *23*, 283–437.

[58] Holehouse, A. S.; Kragelund, B. B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nature Reviews Molecular Cell Biology* **2024**, *25*, 187–211.

[59] Buxbaum, E. *Fundamentals of Protein Structure and Function*; Springer Cham: Cham, Switzerland, 2015.

[60] Voet, D.; Voet, J. G. *Biochemistry*; Wiley: New Jersey, United States, 2010.

[61] Schoedel, A.; Li, M.; Li, D.; O'Keeffe, M.; Yaghi, O. M. Structures of metal–organic frameworks with rod secondary building units. *Chemical Reviews* **2016**, *116*, 12466–12535.

[62] Zheng, S.-T.; Wu, T.; Irfanoglu, B.; Zuo, F.; Feng, P.; Bu, X. Multicomponent Self-Assembly of a Nested Co24@ Co48 Metal–Organic Polyhedral Framework. *Angewandte Chemie* **2011**, *123*, 8184–8187.

[63] Furukawa, H.; Cordova, K. E.; O'Keeffe, M.; Yaghi, O. M. The chemistry and applications of metal-organic frameworks. *Science* **2013**, *341*, 1230444.

[64] O'keeffe, M.; Peskov, M. A.; Ramsden, S. J.; Yaghi, O. M. The reticular chemistry structure resource (RCSR) database of, and symbols for, crystal nets. *Accounts of Chemical Research* **2008**, *41*, 1782–1789.

[65] Delgado-Friedrichs, O.; O'Keeffe, M.; Yaghi, O. M. Taxonomy of periodic nets and the design of materials. *Physical Chemistry Chemical Physics* **2007**, *9*, 1035–1043.

[66] Liang, W.; Wied, P.; Carraro, F.; Sumby, C. J.; Nidetzky, B.; Tsung, C.-K.; Falcaro, P.; Doonan, C. J. Metal–organic framework-based enzyme biocomposites. *Chemical Reviews* **2021**, *121*, 1077–1129.

[67] Riccò, R.; Liang, W.; Li, S.; Gassensmith, J. J.; Caruso, F.; Doonan, C.; Falcaro, P. Metal–organic frameworks for cell and virus biology: a perspective. *ACS Nano* **2018**, *12*, 13–23.

[68] McKenney, P. T.; Driks, A.; Eichenberger, P. The Bacillus subtilis endospore: assembly and functions of the multilayered coat. *Nature Reviews Microbiology* **2013**, *11*, 33–44.

[69] Foissner, W. The stunning, glass-covered resting cyst of Maryna umbrellata (Ciliophora, Colpodea). *Acta Protozoologica* **2009**, *48*, 223.

[70] Foissner, W. Biogeography and dispersal of micro-organisms: a review emphasizing protists. *Acta Protozoologica* **2006**, *45*, 111–136.

[71] Weronika, E.; Łukasz, K. Tardigrades in space research-past and future. *Origins of Life and Evolution of Biospheres* **2017**, *47*, 545–553.

[72] Møbjerg, N.; Neves, R. C. New insights into survival strategies of tardigrades. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology* **2021**, *254*, 110890.

[73] Sun, H.; Li, Y.; Yu, S.; Liu, J. Metal-organic frameworks (MOFs) for biopreservation: From biomacromolecules, living organisms to biological devices. *Nano Today* **2020**, *35*, 100985.

[74] Sheldon, R. A. Enzyme immobilization: the quest for optimum performance. *Advanced Synthesis & Catalysis* **2007**, *349*, 1289–1307.

[75] Kouassi, G. K.; Irudayaraj, J.; McCarty, G. Examination of cholesterol oxidase attachment to magnetic nanoparticles. *Journal of Nanobiotechnology* **2005**, *3*, 1–9.

[76] Zhang, F.; Cho, S.; Yang, S.; Seo, S.; Cha, G.; Nam, H. Gold Nanoparticle-Based Mediatorless Biosensor Prepared on Microporous Electrode. *Electroanalysis* **2006**, *18*, 217–222.

[77] Ansari, S. A.; Husain, Q. Potential applications of enzymes immobilized on/in nano materials: A review. *Biotechnology Advances* **2012**, *30*, 512–523.

[78] Lee, E. S.; Kwon, M. J.; Lee, H.; Kim, J. J. Stabilization of protein encapsulated in poly (lactide-co-glycolide) microspheres by novel viscous S/W/O/W method. *International Journal of Pharmaceutics* **2007**, *331*, 27–37.

[79] Pavlidis, I. V.; Patila, M.; Bornscheuer, U. T.; Gournis, D.; Stamatis, H. Graphene-based nanobiocatalytic systems: recent advances and future prospects. *Trends in Biotechnology* **2014**, *32*, 312–320.

[80] Feng, W.; Ji, P. Enzymes immobilized on carbon nanotubes. *Biotechnology Advances* **2011**, *29*, 889–895.

[81] Veronese, F. M. Peptide and protein PEGylation: a review of problems and solutions. *Biomaterials* **2001**, *22*, 405–417.

[82] Kong, G.; Xiong, M.; Liu, L.; Hu, L.; Meng, H.-M.; Ke, G.; Zhang, X.-B.; Tan, W. DNA origami-based protein networks: From basic construction to emerging applications. *Chemical Society Reviews* **2021**, *50*, 1846–1873.

[83] Wang, H.; Han, L.; Zheng, D.; Yang, M.; Andaloussi, Y. H.; Cheng, P.; Zhang, Z.; Ma, S.; Zaworotko, M. J.; Feng, Y.; et al., Protein-structure-directed metal–organic zeolite-like networks as biomacromolecule carriers. *Angewandte Chemie International Edition* **2020**, *59*, 6263–6267.

[84] Magner, E. Immobilisation of enzymes on mesoporous silicate materials. *Chemical Society Reviews* **2013**, *42*, 6213–6222.

[85] Li, M.; Qiao, S.; Zheng, Y.; Andaloussi, Y. H.; Li, X.; Zhang, Z.; Li, A.; Cheng, P.; Ma, S.; Chen, Y. Fabricating covalent organic framework capsules with commodious microenvironment for enzymes. *Journal of the American Chemical Society* **2020**, *142*, 6675–6681.

[86] Wied, P.; Carraro, F.; Bolivar, J. M.; Doonan, C. J.; Falcaro, P.; Nidetzky, B. Combining a Genetically Engineered Oxidase with Hydrogen-Bonded Organic Frameworks (HOFs) for Highly Efficient Biocomposites. *Angewandte Chemie International Edition* **2022**, *61*, e202117345.

[87] Zhang, X.; Chen, Z.; Liu, X.; Hanna, S. L.; Wang, X.; Taheri-Ledari, R.; Maleki, A.; Li, P.; Farha, O. K. A historical overview of the activation and porosity of metal–organic frameworks. *Chemical Society Reviews* **2020**, *49*, 7406–7427.

[88] Yaghi, O. M. The Reticular Chemist. *Nano Letters* **2020**, *20*, 8432–8434.

[89] Velásquez-Hernández, M. d. J.; Linares-Moreau, M.; Astria, E.; Carraro, F.; Alyami, M. Z.; Khashab, N. M.; Sumby, C. J.; Doonan, C. J.; Falcaro, P. Towards applications of bioentities@ MOFs in biomedicine. *Coordination Chemistry Reviews* **2021**, *429*, 213651.

[90] Chen, Z.; Kirlikovali, K. O.; Li, P.; Farha, O. K. Reticular chemistry for highly porous metal–organic frameworks: The chemistry and applications. *Accounts of Chemical Research* **2022**, *55*, 579–591.

[91] Wang, X.; et al., Spatially confined protein assembly in hierarchical mesoporous metal-organic framework. *Nature Communications* **2023**, *14*, 973.

[92] Cao, S.-L.; Yue, D.-M.; Li, X.-H.; Smith, T. J.; Li, N.; Zong, M.-H.; Wu, H.; Ma, Y.-Z.; Lou, W.-Y. Novel nano-/micro-biocatalyst: soybean epoxide hydrolase immobilized on UiO-66-NH2 MOF for efficient biosynthesis of enantiopure (R)-1, 2-octanediol in deep eutectic solvents. *ACS Sustainable Chemistry & Engineering* **2016**, *4*, 3586–3595.

[93] Patra, S.; Crespo, T. H.; Permyakova, A.; Sicard, C.; Serre, C.; Chaussé, A.; Steunou, N.; Legrand, L. Design of metal organic framework–enzyme based bioelectrodes as a novel and highly sensitive biosensing platform. *Journal of Materials Chemistry B* **2015**, *3*, 8983–8992.

[94] Carné-Sánchez, A.; Carmona, F. J.; Kim, C.; Furukawa, S. Porous materials as carriers of gasotransmitters towards gas biology and therapeutic applications. *Chemical Communications* **2020**, *56*, 9750–9766.

[95] Gascón, V.; Castro-Miguel, E.; Díaz-García, M.; Blanco, R. M.; Sanchez-Sanchez, M. In situ and post-synthesis immobilization of enzymes on nanocrystalline MOF platforms to yield active biocatalysts. *Journal of Chemical Technology & Biotechnology* **2017**, *92*, 2583–2593.

[96] Rabe, M.; Verdes, D.; Seeger, S. Understanding protein adsorption phenomena at solid surfaces. *Advances in Colloid and Interface Science* **2011**, *162*, 87–106.

[97] Illanes, A., Ed. *Enzyme Biocatalysis Principles and Applications*; Springer Dordrecht: Netherlands, 2008.

[98] Pisklak, T. J.; Macías, M.; Coutinho, D. H.; Huang, R. S.; Balkus, K. J. Hybrid materials for immobilization of MP-11 catalyst. *Topics in Catalysis* **2006**, *38*, 269–278.

[99] Liang, K.; et al., Biomimetic mineralization of metal-organic frameworks as protective coatings for biomacromolecules. *Nature Communications* **2015**, *6*, 7240.

[100] Zhu, W.; et al., SupraCells: living mammalian cells protected within functional modular nanoparticle-based exoskeletons. *Advanced Materials* **2019**, *31*, 1900545.

[101] He, C.; Lu, K.; Liu, D.; Lin, W. Nanoscale metal–organic frameworks for the co-delivery of cisplatin and pooled siRNAs to enhance therapeutic efficacy in drug-resistant ovarian cancer cells. *Journal of the American Chemical Society* **2014**, *136*, 5181–5184.

[102] Chen, Y.; Han, S.; Li, X.; Zhang, Z.; Ma, S. Why does enzyme not leach from metal–organic frameworks (MOFs)? Unveiling the interactions between an enzyme molecule and a MOF. *Inorganic Chemistry* **2014**, *53*, 10006–10008.

[103] Chen, Y.; Lykourinou, V.; Hoang, T.; Ming, L.-J.; Ma, S. Size-selective biocatalysis of myoglobin immobilized into a mesoporous metal–organic framework with hierarchical pore sizes. *Inorganic Chemistry* **2012**, *51*, 9156–9158.

[104] Deng, H.; et al., Large-pore apertures in a series of metal-organic frameworks. *Science* **2012**, *336*, 1018–1023.

[105] Chen, Y.; Lykourinou, V.; Vetromile, C.; Hoang, T.; Ming, L.-J.; Larsen, R. W.; Ma, S. How can proteins enter the interior of a MOF? Investigation of cytochrome c translocation into a MOF consisting of mesoporous cages with microporous windows. *Journal of the American Chemical Society* **2012**, *134*, 13188–13191.

[106] Lian, X.; Chen, Y.-P.; Liu, T.-F.; Zhou, H.-C. Coupling two enzymes into a tandem nanoreactor utilizing a hierarchically structured MOF. *Chemical Science* **2016**, *7*, 6969–6973.

[107] Li, P.; Moon, S.-Y.; Guelta, M. A.; Lin, L.; Gómez-Gualdrón, D. A.; Snurr, R. Q.; Harvey, S. P.; Hupp, J. T.; Farha, O. K. Nanosizing a metal–organic framework enzyme carrier for accelerating nerve agent hydrolysis. *ACS Nano* **2016**, *10*, 9174–9182.

[108] Tai, T.-Y.; Sha, F.; Wang, X.; Wang, X.; Ma, K.; Kirlikovali, K. O.; Su, S.; Islamoglu, T.; Kato, S.; Farha, O. K. Leveraging Isothermal Titration Calorimetry to Explore Structure–Property Relationships of Protein Immobilization in Metal–Organic Frameworks. *Angewandte Chemie* **2022**, *134*, e202209110.

[109] Pan, Y.; Li, H.; Li, Q.; Lenertz, M.; Zhu, X.; Chen, B.; Yang, Z. Site-directed spin labeling-electron paramagnetic resonance spectroscopy in biocatalysis: Enzyme orientation and dynamics in nanoscale confinement. *Chem Catalysis* **2021**, *1*, 207–231.

[110] Liang, J.; Bin Zulkifli, M. Y.; Yong, J.; Du, Z.; Ao, Z.; Rawal, A.; Scott, J. A.; Harmer, J. R.; Wang, J.; Liang, K. Locking the ultrasound-induced active conformation of metalloenzymes in metal–organic frameworks. *Journal of the American Chemical Society* **2022**, *144*, 17865–17875.

[111] Chapman, J.; Zoica Dinu, C. Assessment of Enzyme Functionality at Metal–Organic Framework Interfaces Developed through Molecular Simulations. *Langmuir* **2023**, *39*, 1750–1763.

[112] Zhang, H.; Lv, Y.; Tan, T.; van der Spoel, D. Atomistic simulation of protein encapsulation in metal–organic frameworks. *The Journal of Physical Chemistry B* **2016**, *120*, 477–484.

[113] Tuan Kob, T.; Ismail, M.; Abdul Rahman, M.; Cordova, K. E.; Mohammad Latif, M. Unraveling the structural dynamics of an enzyme encapsulated within a metal–organic framework. *The Journal of Physical Chemistry B* **2020**, *124*, 3678–3685.

[114] Li, P.; Modica, J. A.; Howarth, A. J.; Vargas, E.; Moghadam, P. Z.; Snurr, R. Q.; Mrksich, M.; Hupp, J. T.; Farha, O. K. Toward design rules for enzyme immobilization in hierarchical mesoporous metal-organic frameworks. *Chem* **2016**, *1*, 154–169.

[115] Chen, Y.; et al., Insights into the enhanced catalytic activity of cytochrome c when encapsulated in a metal–organic framework. *Journal of the American Chemical Society* **2020**, *142*, 18576–18582.

[116] Sato, T.; Esaki, M.; Fernandez, J. M.; Endo, T. Comparison of the protein-unfolding pathways between mitochondrial protein import and atomic-force microscopy measurements. *Proceedings of the National Academy of Sciences* **2005**, *102*, 17999–18004.

[117] Berko, D.; et al., The direction of protein entry into the proteasome determines the variety of products and depends on the force needed to unfold its two termini. *Molecular Cell* **2012**, *48*, 601–611.

[118] Makarov, D. E. Computer simulations and theory of protein translocation. *Accounts of Chemical Research* **2009**, *42*, 281–289.

[119] Gkaniatsou, E.; Sicard, C.; Ricoux, R.; Benahmed, L.; Bourdreux, F.; Zhang, Q.; Serre, C.; Mahy, J.-P.; Steunou, N. Enzyme encapsulation in mesoporous metal–organic frameworks for selective biodegradation of harmful dye molecules. *Angewandte Chemie* **2018**, *130*, 16373–16378.

[120] Miller Jr, W. B.; Baluška, F.; Reber, A. S. A revised central dogma for the 21st century: All biology is cognitive information processing. *Progress in Biophysics and Molecular Biology* **2023**, *182*, 34–48.

[121] Black, D. L. Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry* **2003**, *72*, 291–336.

[122] Mueller, M. M. Post-Translational Modifications of Protein Backbones: Unique Functions, Mechanisms, and Challenges. *Biochemistry* **2018**, *57*, 177–185.

[123] Walsh, C. T.; Garneau-Tsodikova, S.; Gatto Jr, G. J. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angewandte Chemie International Edition* **2005**, *44*, 7342–7372.

[124] Kumar, S.; Prakash, S.; Gupta, K.; Dongre, A.; Balaram, P.; Balaram, H. Unexpected functional implication of a stable succinimide in the structural stability of Methanocaldococcus jannaschii glutaminase. *Nature Communications* **2016**, *7*, 12798.

[125] Dongre, A. V.; Das, S.; Bellur, A.; Kumar, S.; Chandrashekarmath, A.; Karmakar, T.; Balaram, P.; Balasubramanian, S.; Balaram, H. Structural basis for the hyperthermostability of an archaeal enzyme induced by succinimide formation. *Biophysical Journal* **2021**, *120*, 3732–3746.

[126] Alder, B. J.; Wainwright, T. E. Studies in molecular dynamics. II. Behavior of a small number of elastic spheres. *The Journal of Chemical Physics* **1960**, *33*, 1439–1451.

[127] McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of folded proteins. *Nature* **1977**, *267*, 585–590.

[128] Brini, E.; Simmerling, C.; Dill, K. Protein storytelling through physics. *Science* **2020**, *370*, eaaz3041.

[129] Perez, A.; Morrone, J. A.; Simmerling, C.; Dill, K. A. Advances in free-energy-based simulations of protein folding and ligand binding. *Current Opinion in Structural Biology* **2016**, *36*, 25–31.

[130] Wallace, G. G.; Higgins, M. J.; Moulton, S. E.; Wang, C. Nanobionics: the impact of nanotechnology on implantable medical bionic devices. *Nanoscale* **2012**, *4*, 4327–4347.

[131] Aono, M.; Ariga, K. The way to nanoarchitectonics and the way of nanoarchitectonics. *Advanced Materials* **2016**, *28*, 989–992.

[132] Dusastre, V. MOF patterns. *Nature Materials* **2013**, *12*, 778–778.

[133] Schmid, A.; Dordick, J.; Hauer, B.; Kiener, A.; Wubbolts, M.; Witholt, B. Industrial biocatalysis today and tomorrow. *Nature* **2001**, *409*, 258–268.

[134] Bhattacharjee, N.; Alonso-Cotchico, L.; Lucas, M. F. Enzyme immobilization studied through molecular dynamic simulations. *Frontiers in Bioengineering and Biotechnology* **2023**, *11*, 1200293.

[135] Barati, F.; et al., In-silico approaches to investigate enzyme immobilization: a comprehensive systematic review. *Physical Chemistry Chemical Physics* **2024**, *26*, 5744–5761.

[136] Leu, B. M.; Zhang, Y.; Bu, L.; Straub, J. E.; Zhao, J.; Sturhahn, W.; Alp, E. E.; Sage, J. T. Resilience of the iron environment in heme proteins. *Biophysical Journal* **2008**, *95*, 5874–5889.

[137] Meuwly, M.; Becker, O. M.; Stote, R.; Karplus, M. NO rebinding to myoglobin: a reactive molecular dynamics study. *Biophysical Chemistry* **2002**, *98*, 183–207.

[138] Barducci, A.; Bussi, G.; Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical Review Letters* **2008**, *100*, 020603.

[139] Berendsen, H. J.; Postma, J. v.; Van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **1984**, *81*, 3684–3690.

[140] Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; Van Gunsteren, W. F.; Mark, A. E. Peptide folding: when simulation meets experiment. *Angewandte Chemie International Edition* **1999**, *38*, 236–240.

[141] Chung, Y. G.; Camp, J.; Haranczyk, M.; Sikora, B. J.; Bury, W.; Krungleviciute, V.; Yildirim, T.; Farha, O. K.; Sholl, D. S.; Snurr, R. Q. Computation-ready, experimental metal–organic frameworks: A tool to enable high-throughput screening of nanoporous crystals. *Chemistry of Materials* **2014**, *26*, 6185–6192.

[142] Chung, Y. G.; et al., Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: CoRE MOF 2019. *Journal of Chemical & Engineering Data* **2019**, *64*, 5985–5998.

[143] VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach. *Computer Physics Communications* **2005**, *167*, 103–128.

[144] Kühne, T. D.; et al., CP2K: An electronic structure and molecular dynamics software package-Quickstep: Efficient and accurate electronic structure calculations. *The Journal of Chemical Physics* **2020**, *152*.

[145] Goedecker, S.; Teter, M.; Hutter, J. Separable dual-space Gaussian pseudopotentials. *Physical Review B* **1996**, *54*, 1703.

[146] Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Physical Review Letters* **1996**, *77*, 3865.

[147] 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. *The Journal of Chemical Physics* **2010**, *132*.

[148] Garberoglio, G. OBGMX: A web-based generator of GROMACS topologies for molecular and periodic systems using the universal force field. *Journal of Computational Chemistry* **2012**, *33*, 2204–2208.

[149] Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *Journal of Computational Chemistry* **2009**, *30*, 2157–2164.

[151] Martin, R. L.; Smit, B.; Haranczyk, M. Addressing challenges of identifying geometrically diverse sets of crystalline porous materials. *Journal of Chemical Information and Modeling* **2012**, *52*, 308–318.

[152] Pinheiro, M.; Martin, R. L.; Rycroft, C. H.; Jones, A.; Iglesia, E.; Haranczyk, M. Characterization and comparison of pore landscapes in crystalline porous materials. *Journal of Molecular Graphics and Modelling* **2013**, *44*, 208–219.

[153] Pinheiro, M.; Martin, R. L.; Rycroft, C. H.; Haranczyk, M. High accuracy geometric analysis of crystalline porous materials. *CrystEngComm* **2013**, *15*, 7531–7538.

[154] Ongari, D.; Boyd, P. G.; Barthel, S.; Witman, M.; Haranczyk, M.; Smit, B. Accurate characterization of the pore volume in microporous crystalline materials. *Langmuir* **2017**, *33*, 14529–14538.

[155] Martin, R. L.; Haranczyk, M. Construction and characterization of structure models of crystalline porous polymers. *Crystal Growth & Design* **2014**, *14*, 2431–2440.

[150] Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous and Mesoporous Materials* **2012**, *149*, 134–141.

[156] Jash, O.; Srivastava, A.; Balasubramanian, S. HP35 Protein in the Mesopore of MIL-101 (Cr) MOF: A Model to Study Cotranslocational Unfolding. *ACS Omega* **2024**, *9*, 31185–31194.

[157] Turner, J. G.; Murphy, C. J. How Do Proteins Associate with Nanoscale Metal–Organic Framework Surfaces? *Langmuir* **2021**, *37*, 9910–9919.

[158] Yang, W.; Liang, W.; O'Dell, L. A.; Toop, H. D.; Maddigan, N.; Zhang, X.; Kochubei, A.; Doonan, C. J.; Jiang, Y.; Huang, J. Insights into the interaction between immobilized biocatalysts and metal–organic frameworks: a case study of PCN-333. *JACS Au* **2021**, *1*, 2172–2181.

[159] Andrade, J.; Hlady, V.; Wei, A. Adsorption of complex proteins at interfaces. *Pure and Applied Chemistry* **1992**, *64*, 1777–1781.

[160] Wang, Q.; Wang, M.-h.; Wang, K.-f.; Liu, Y.; Zhang, H.-p.; Lu, X.; Zhang, X.-d. Computer simulation of biomolecule–biomaterial interactions at surfaces and interfaces. *Biomedical Materials* **2015**, *10*, 032001.

[161] Walsh, T. R. Pathways to structure–property relationships of peptide–materials interfaces: Challenges in predicting molecular structures. *Accounts of Chemical Research* **2017**, *50*, 1617–1624.

[162] Norde, W. *Macromolecular Symposia*; 1996; Vol. 103; pp 5–18.

[163] Norde, W. My voyage of discovery to proteins in flatland... and beyond. *Colloids and Surfaces. B, Biointerfaces* **2007**, *61*, 1–9.

[164] Adamczyk, Z. Protein adsorption: A quest for a universal mechanism. *Current Opinion in Colloid & Interface Science* **2019**, *41*, 50–65.

[165] Le Caër, S.; Pin, S.; Esnouf, S.; Raffy, Q.; Renault, J. P.; Brubach, J.-B.; Creff, G.; Roy, P. A trapped water network in nanoporous material: the role of interfaces. *Physical Chemistry Chemical Physics* **2011**, *13*, 17658–17666.

[166] Tang, Y.; Grey, M. J.; McKnight, J.; Palmer III, A. G.; Raleigh, D. P. Multistate folding of the villin headpiece domain. *Journal of Molecular Biology* **2006**, *355*, 1066–1077.

[167] Vardar, D.; Chishti, A.; Frank, B.; Luna, E. J.; Noegel, A.; Oh, S. W.; Schleicher, M.; McKnight, C. Villin-type headpiece domains show a wide range of F-actin-binding affinities. *Cell Motility and the Cytoskeleton* **2002**, *52*, 9–21.

[168] McKnight, J. C.; Doering, D. S.; Matsudaira, P. T.; Kim, P. S. A thermostable 35-residue subdomain within villin headpiece. *Journal of Molecular Biology* **1996**, *260*, 126–134.

[169] Kubelka, J.; Eaton, W. A.; Hofrichter, J. Experimental Tests of Villin Subdomain Folding Simulations. *Journal of Molecular Biology* **2003**, *329*, 625–630.

[170] McKnight, C. J.; Matsudaira, P. T.; Kim, P. S. NMR structure of the 35-residue villin headpiece subdomain. *Nature Structural Biology* **1997**, *4*, 180–184.

[171] Muñoz, V. Conformational dynamics and ensembles in protein folding. *Annual Review of Biophysics and Biomolecular Structure* **2007**, *36*, 395–412.

[172] Chiu, T. K.; Kubelka, J.; Herbst-Irmer, R.; Eaton, W. A.; Hofrichter, J.; Davies, D. R. High-resolution x-ray crystal structures of the villin headpiece subdomain, an ultrafast folding protein. *Proceedings of the National Academy of Sciences* **2005**, *102*, 7517–7522.

[173] 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* **2005**, *309*, 2040–2042.

[174] Navarro-Sánchez, J.; Almora-Barrios, N.; Lerma-Berlanga, B.; Ruiz-Pernía, J. J.; Lorenz-Fonfria, V. A.; Tuñón, I.; Marti-Gastaldo, C. Translocation of enzymes into a mesoporous MOF for enhanced catalytic activity under extreme conditions. *Chemical Science* **2019**, *10*, 4082–4088.

[175] Vermeulen, W.; Vanhaesebrouck, P.; Van Troys, M.; Verschueren, M.; Fant, F.; Goethals, M.; Ampe, C.; Martins, J. C.; Borremans, F. A. Solution structures of the C-terminal headpiece subdomains of human villin and advillin, evaluation of headpiece F-actin-binding requirements. *Protein Science* **2004**, *13*, 1276–1287.

[176] BIOVIA, Dassault Systèmes, Materials Studio, 2020, San Diego: Dassault Systèmes, 2020.

[177] Berendsen, H. J.; van der Spoel, D.; van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications* **1995**, *91*, 43–56.

[178] Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation* **2008**, *4*, 435–447.

[179] Pronk, S.; et al., GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845–854.

[180] Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1*, 19–25.

[181] Páll, S.; Abraham, M. J.; Kutzner, C.; Hess, B.; Lindahl, E. In *Solving Software Challenges for Exascale*; Markidis, S., Laure, E., Eds.; Springer International Publishing: Cham, 2015; pp 3–27.

[182] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics* **2006**, *65*, 712–725.

[183] Best, R. B.; Hummer, G. Optimized molecular dynamics force fields applied to the helix- coil transition of polypeptides. *The Journal of Physical Chemistry B* **2009**, *113*, 9004–9015.

[184] Lindorff-Larsen, K.; Trbovic, N.; Maragakis, P.; Piana, S.; Shaw, D. E. Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *Journal of the American Chemical Society* **2012**, *134*, 3787–3791.

[185] Addicoat, M. A.; Vankova, N.; Akter, I. F.; Heine, T. Extension of the universal force field to metal–organic frameworks. *Journal of Chemical Theory and Computation* **2014**, *10*, 880–891.

[186] Coupry, D. E.; Addicoat, M. A.; Heine, T. Extension of the universal force field for metal–organic frameworks. *Journal of Chemical Theory and Computation* **2016**, *12*, 5215–5225.

[187] Korolev, V. V.; Mitrofanov, A.; Marchenko, E. I.; Eremin, N. N.; Tkachenko, V.; Kalmykov, S. N. Transferable and extensible machine learning-derived atomic charges for modeling hybrid nanoporous materials. *Chemistry of Materials* **2020**, *32*, 7822–7831.

[188] Wang, X.; Lan, P. C.; Ma, S. Metal–organic frameworks for enzyme immobilization: beyond host matrix materials. *ACS Central Science* **2020**, *6*, 1497–1506.

[189] Xia, H.; Li, N.; Zhong, X.; Jiang, Y. Metal-organic frameworks: a potential platform for enzyme immobilization and related applications. *Frontiers in Bioengineering and Biotechnology* **2020**, *8*, 695.

[190] Saladino, G.; Marenchino, M.; Gervasio, F. Bridging the gap between folding simulations and experiments: The case of the villin headpiece. *Journal of Chemical Theory and Computation* **2011**, *7*, 2675–2680.

[191] Reiner, A.; Henklein, P.; Kiefhaber, T. An unlocking/relocking barrier in conformational fluctuations of villin headpiece subdomain. *Proceedings of the National Academy of Sciences* **2010**, *107*, 4955–4960.

[192] Spear, M. *Charting Statistics*; McGraw-Hill Book Company, Inc, 1952.

[193] McGill, R.; Tukey, J. W.; Larsen, W. A. Variations of box plots. *The American Statistician* **1978**, *32*, 12–16.

[194] Wickham, H.; Stryjewski, L. 40 years of boxplots. *Am. Statistician* **2011**,

[195] Tukey, J. W. *Exploratory Data Analysis: Limited Preliminary Edition*; Addison-Wesley Publishing Company Ann Arbor, MI, USA, 1970.

[196] Zhang, Y.; Voth, G. A. Combined metadynamics and umbrella sampling method for the calculation of ion permeation free energy profiles. *Journal of Chemical Theory and Computation* **2011**, *7*, 2277–2283.

[197] Hu, Y.; Liu, X.; Sinha, S. K.; Patel, S. Translocation thermodynamics of linear and cyclic nonaarginine into model DPPC bilayer via coarse-grained molecular dynamics simulation: implications of pore formation and nonadditivity. *The Journal of Physical Chemistry B* **2014**, *118*, 2670–2682.

[198] Robinson, A. B.; Rudd, C. J. Deamidation of glutaminyl and asparaginyl residues in peptides and proteins. *Current Topics in Cellular Regulation* **1974**, *8*, 247–295.

[199] Robinson, A. B.; Scotchler, J. W.; McKerrow, J. H. Rates of nonenzymic deamidation of glutaminyl and asparaginyl residues in pentapeptides. *Journal of the American Chemical Society* **1973**, *95*, 8156–8159.

[200] Bornstein, P.; Balian, G. The specific nonenzymatic cleavage of bovine ribonuclease with hydroxylamine. *Journal of Biological Chemistry* **1970**, *245*, 4854–4856.

[201] Geiger, T.; Clarke, S. Deamidation, isomerization, and racemization at asparaginyl and aspartyl residues in peptides. Succinimide-linked reactions that contribute to protein degradation. *Journal of Biological Chemistry* **1987**, *262*, 785–794.

[202] MEINWALD, Y. C.; STIMSON, E. R.; SCHERAGA, H. A. Deamidation of the asparaginyl-glycyl sequence. *International Journal of Peptide and Protein Research* **1986**, *28*, 79–84.

[203] Reissner, K.; Aswad, D. Deamidation and isoaspartate formation in proteins: unwanted alterations or surreptitious signals? *Cellular and Molecular Life Sciences CMLS* **2003**, *60*, 1281–1295.

[204] Wright, H. T. Nonenzymatic deamidation of asparaginyl and glutaminyl residues in proteins. *Critical Reviews in Biochemistry and Molecular Biology* **1991**, *26*, 1–52.

[205] Payan, I. L.; Chou, S.-J.; Fisher, G. H.; Man, E. H.; Emory, C.; Frey, W. H. Altered aspartate in Alzheimer neurofibrillary tangles. *Neurochemical Research* **1992**, *17*, 187–191.

[206] Hubbard, E. E.; et al., Spontaneous Isomerization of Asp387 in Tau is Diagnostic for Alzheimer's Disease: An Endogenous Indicator of Reduced Autophagic Flux. *bioRxiv* **2021**, 2021–04.

[207] Takata, T.; Oxford, J. T.; Demeler, B.; Lampi, K. J. Deamidation destabilizes and triggers aggregation of a lens protein, βA3-crystallin. *Protein Science* **2008**, *17*, 1565–1575.

[208] Truscott, R. J.; Schey, K. L.; Friedrich, M. G. Old proteins in man: a field in its infancy. *Trends in Biochemical Sciences* **2016**, *41*, 654–664.

[209] Pande, A.; Mokhor, N.; Pande, J. Deamidation of human γS-crystallin increases attractive protein interactions: implications for cataract. *Biochemistry* **2015**, *54*, 4890–4899.

[210] Robinson, N.; Robinson, A. Deamidation of human proteins. *Proceedings of the National Academy of Sciences* **2001**, *98*, 12409–12413.

[211] Weintraub, S. J.; Manson, S. R. Asparagine deamidation: a regulatory hourglass. *Mechanisms of Ageing and Development* **2004**, *125*, 255–257.

[212] Robinson, N. E.; Robinson, A. B. Molecular clocks. *Proceedings of the National Academy of Sciences* **2001**, *98*, 944–949.

[213] Capasso, S.; Mazzarella, L.; Sica, F.; Zagari, A.; Salvadori, S. Kinetics and mechanism of succinimide ring formation in the deamidation process of asparagine residues. *Journal of the Chemical Society, Perkin Transactions 2* **1993**, 679–682.

[214] Capasso, S.; Salvadori, S. Effect of the three-dimensional structure on the deamidation reaction of ribonuclease A. *The Journal of Peptide Research* **1999**, *54*, 377–382.

[215] Capasso, S.; Mazzarella, L.; Sica, F.; Zagari, A. Deamidation via cyclic imide in asparaginyl peptides. *Peptide Research* **1989**, *2*, 195–200.

[216] Tam, J. P.; Riemen, M.; Merrifield, R. Mechanisms of aspartimide formation: the effects of protecting groups, acid, base, temperature and time. *Peptide Research* **1988**, *1*, 6–18.

[217] Robinson, N.; Robinson, A. Prediction of primary structure deamidation rates of asparaginyl and glutaminyl peptides through steric and catalytic effects. *The Journal of Peptide Research* **2004**, *63*, 437–448.

[218] Robinson, N.; Robinson, Z.; Robinson, B.; Robinson, A.; Robinson, J.; Robinson, M.; Robinson, A. B. Structure-dependent nonenzymatic deamidation of glutaminyl and asparaginyl pentapeptides. *The Journal of Peptide Research* **2004**, *63*, 426–436.

[219] Robinson, N. E.; Robinson, A. *Molecular clocks Deamidation of Asparaginyl and Glutaminyl Residues in Peptides and Proteins*; Althouse press: London, ON, 2004.

[220] Robinson, N. E.; Robinson, A. B. Prediction of protein deamidation rates from primary and three-dimensional structure. *Proceedings of the National Academy of Sciences* **2001**, *98*, 4367–4372.

[221] Patel, K.; Borchardt, R. T. Chemical pathways of peptide degradation. III. Effect of primary sequence on the pathways of deamidation of asparaginyl residues in hexapeptides. *Pharmaceutical Research* **1990**, *7*, 787–793.

[222] Yan, Q.; Huang, M.; Lewis, M. J.; Hu, P. Structure based prediction of asparagine deamidation propensity in monoclonal antibodies. *mAbs* **2018**, *10*, 901–912.

[223] Sydow, J. F.; et al., Structure-based prediction of asparagine and aspartate degradation sites in antibody variable regions. *PLOS One* **2014**, *9*, e100736.

[224] Irudayanathan, F. J.; Zarzar, J.; Lin, J.; Izadi, S. Deciphering deamidation and isomerization in therapeutic proteins: Effect of neighboring residue. *mAbs* **2022**, *14*, 2143006.

[225] Ugur, I.; Marion, A.; Aviyente, V.; Monard, G. Why does Asn71 deamidate faster than Asn15 in the enzyme triosephosphate isomerase? Answers from microsecond molecular dynamics simulation and QM/MM free energy calculations. *Biochemistry* **2015**, *54*, 1429–1439.

[226] Plotnikov, N. V.; Singh, S. K.; Rouse, J. C.; Kumar, S. Quantifying the risks of asparagine deamidation and aspartate isomerization in biopharmaceuticals by computing reaction free-energy surfaces. *The Journal of Physical Chemistry B* **2017**, *121*, 719–730.

[227] Kumar, S.; Plotnikov, N. V.; Rouse, J. C.; Singh, S. K. Biopharmaceutical informatics: supporting biologic drug development via molecular modelling and informatics. *Journal of Pharmacy and Pharmacology* **2018**, *70*, 595–608.

[228] Ugur, I.; Aviyente, V.; Monard, G. Initiation of the reaction of deamidation in triosephosphate isomerase: investigations by means of molecular dynamics simulations. *The Journal of Physical Chemistry B* **2012**, *116*, 6288–6301.

[229] Vatsa, S. In silico prediction of post-translational modifications in therapeutic antibodies. *mAbs* **2022**, *14*, 2023938.

[230] Konuklar, F. A.; Aviyente, V.; Sen, T. Z.; Bahar, I. Modeling the deamidation of asparagine residues via succinimide intermediates. *Journal of Molecular Modeling* **2001**, *7*, 147–160.

[231] Konuklar, F. A. S.; Aviyente, V. Modelling the hydrolysis of succinimide: formation of aspartate and reversible isomerization of aspartic acid via succinimide. *Organic & Biomolecular Chemistry* **2003**, *1*, 2290–2297.

[232] Konuklar, F. A.; Aviyente, V.; Ruiz Lopez, M. F. Theoretical study on the alkaline and neutral hydrolysis of succinimide derivatives in deamidation reactions. *The Journal of Physical Chemistry A* **2002**, *106*, 11205–11214.

[233] Radkiewicz, J. L.; Zipse, H.; Clarke, S.; Houk, K. Accelerated racemization of aspartic acid and asparagine residues via succinimide intermediates: an ab initio theoretical exploration of mechanism. *Journal of the American Chemical Society* **1996**, *118*, 9148–9155.

[234] Radkiewicz, J. L.; Zipse, H.; Clarke, S.; Houk, K. Neighboring side chain effects on asparaginyl and aspartyl degradation: an ab initio study of the relationship between peptide conformation and backbone NH acidity. *Journal of the American Chemical Society* **2001**, *123*, 3499–3506.

[235] Peters, B.; Trout, B. L. Asparagine deamidation: pH-dependent mechanism from density functional theory. *Biochemistry* **2006**, *45*, 5384–5392.

[236] Catak, S.; Monard, G.; Aviyente, V.; Ruiz-Lopez, M. F. Reaction mechanism of deamidation of asparaginyl residues in peptides: Effect of solvent molecules. *The Journal of Physical Chemistry A* **2006**, *110*, 8354–8365.

[237] Catak, S.; Monard, G.; Aviyente, V.; Ruiz-López, M. F. Computational study on nonenzymatic peptide bond cleavage at asparagine and aspartic acid. *The Journal of Physical Chemistry A* **2008**, *112*, 8752–8761.

[238] Catak, S.; Monard, G.; Aviyente, V.; Ruiz-López, M. F. Deamidation of asparagine residues: direct hydrolysis versus succinimide-mediated deamidation mechanisms. *The Journal of Physical Chemistry A* **2009**, *113*, 1111–1120.

[239] Kaliman, I.; Nemukhin, A.; Varfolomeev, S. Free energy barriers for the N-terminal asparagine to succinimide conversion: quantum molecular dynamics simulations for the fully solvated model. *Journal of Chemical Theory and Computation* **2010**, *6*, 184–189.

[240] Patel, K.; Borchardt, R. T. Chemical pathways of peptide degradation. II. Kinetics of deamidation of an asparaginyl residue in a model hexapeptide. *Pharmaceutical Research* **1990**, *7*, 703–711.

[241] Cao, L.; Beiser, M.; Koos, J. D.; Orlova, M.; Elashal, H. E.; Schroder, H. V.; Link, A. J. Cellulonodin-2 and Lihuanodin: Lasso peptides with an aspartimide

post-translational modification. *Journal of the American Chemical Society* **2021**, *143*, 11690–11702.

[242] Schrödinger, LLC, The PyMOL Molecular Graphics System, Version 1.8. 2015.

[243] M.J., A.; van der Spoel D.,; E., L.; B., H.; the GROMACS development team, GROMACS User Manual version 2019.4, http://www.gromacs.org.

[244] Colizzi, F.; et al., Promoting transparency and reproducibility in enhanced molecular simulations. *Nature Methods* **2019**, *16*, 670–673.

[245] Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Computer Physics Communications* **2014**, *185*, 604–613.

[246] Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *The Journal of Chemical Physics* **1995**, *103*, 8577–8593.

[247] Schmid, N.; Eichenberger, A. P.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A. E.; Van Gunsteren, W. F. Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *European Biophysics Journal* **2011**, *40*, 843–856.

[248] Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics* **2010**, *78*, 1950–1958.

[249] Lawson, K. E.; Dekle, J. K.; Evans, M. N.; Adamczyk, A. J. Deamidation reaction network mapping of pharmacologic and related proteins: impact of solvation dielectric on the degradation energetics of asparagine dipeptides. *Reaction Chemistry & Engineering* **2022**, *7*, 1525–1543.

[250] De Sciscio, M. L.; Nardi, A. N.; Centola, F.; Rossi, M.; Guarnera, E.; D'Abramo, M. Molecular Modeling of the Deamidation Reaction in Solution: A Theoretical–Computational Study. *The Journal of Physical Chemistry B* **2023**, *127*, 9550–9559.

[251] Johnson, E. R.; Mackie, I. D.; DiLabio, G. A. Dispersion interactions in density-functional theory. *Journal of Physical Organic Chemistry* **2009**, *22*, 1127–1135.

[252] Lawan, N.; Ranaghan, K. E.; Manby, F. R.; Mulholland, A. J. Comparison of DFT and ab initio QM/MM methods for modelling reaction in chorismate synthase. *Chemical Physics Letters* **2014**, *608*, 380–385.

[253] Siegbahn, P. E.; Himo, F. Recent developments of the quantum chemical cluster approach for modeling enzyme reactions. *JBIC Journal of Biological Inorganic Chemistry* **2009**, *14*, 643–651.

[254] Kaiyawet, N.; Lonsdale, R.; Rungrotmongkol, T.; Mulholland, A. J.; Hannongbua, S. High-level QM/MM calculations support the concerted mechanism for Michael addition and covalent complex formation in thymidylate synthase. *Journal of Chemical Theory and Computation* **2015**, *11*, 713–722.

[255] Lonsdale, R.; Houghton, K. T.; Zurek, J.; Bathelt, C. M.; Foloppe, N.; de Groot, M. J.; Harvey, J. N.; Mulholland, A. J. Quantum mechanics/molecular mechanics modeling of regioselectivity of drug metabolism in cytochrome P450 2C9. *Journal of the American Chemical Society* **2013**, *135*, 8001–8015.

[256] Fouda, A.; Ryde, U. Does the DFT self-interaction error affect energies calculated in proteins with large QM systems? *Journal of Chemical Theory and Computation* **2016**, *12*, 5667–5679.

[257] Pentikainen, U.; Shaw, K. E.; Senthilkumar, K.; Woods, C. J.; Mulholland, A. J. Lennard- Jones Parameters for B3LYP/CHARMM27 QM/MM Modeling of Nucleic Acid Bases. *Journal of Chemical Theory and Computation* **2009**, *5*, 396–410.

[258] Riccardi, D.; Li, G.; Cui, Q. Importance of van der Waals Interactions in QM/MM Simulations. *The Journal of Physical Chemistry B* **2004**, *108*, 6467–6478.

[259] Liao, R.-Z.; Thiel, W. Convergence in the QM-only and QM/MM modeling of enzymatic reactions: A case study for acetylene hydratase. *Journal of Computational Chemistry* **2013**, *34*, 2389–2397.

[260] Brodsky, A. Is there predictive value in water computer simulations? *Chemical Physics Letters* **1996**, *261*, 563–568.

[261] Vega, C.; Abascal, J. L. Simulating water with rigid non-polarizable models: a general perspective. *Physical Chemistry Chemical Physics* **2011**, *13*, 19663–19688.

[262] Cui, Q. Perspective: Quantum mechanical methods in biochemistry and biophysics. *The Journal of Chemical Physics* **2016**, *145*.

[263] Kuriyan, J.; Wilz, S.; Karplus, M.; Petsko, G. A. X-ray structure and refinement of carbon-monoxy (Fe II)-myoglobin at 1.5 Å resolution. *Journal of Molecular Biology* **1986**, *192*, 133–154.

[264] Ormö, M.; Cubitt, A. B.; Kallio, K.; Gross, L. A.; Tsien, R. Y.; Remington, S. J. Crystal structure of the Aequorea victoria green fluorescent protein. *Science* **1996**, *273*, 1392–1395.

[265] Shinobu, A.; Agmon, N. The hole in the barrel: water exchange at the GFP chromophore. *The Journal of Physical Chemistry B* **2015**, *119*, 3464–3478.

[266] DeLano, W. L.; et al., Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr* **2002**, *40*, 82–92.

[267] Bonomi, M.; et al., PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications* **2009**, *180*, 1961–1972.

[268] Bonomi, M.; et al., Promoting transparency and reproducibility in enhanced molecular simulations. *Nature Methods* **2019**, *16*, 670–673.

[269] Pietrucci, F.; Laio, A. A collective variable for the efficient exploration of protein beta-sheet structures: application to SH3 and GB1. *Journal of Chemical Theory and Computation* **2009**, *5*, 2197–2201.

[270] Harris, C. R.; et al., Array programming with NumPy. *Nature* **2020**, *585*, 357–362.

[271] Frishman, D.; Argos, P. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics* **1995**, *23*, 566–579.

[272] Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *Journal of Molecular Graphics* **1996**, *14*, 33–38.

[273] Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **2007**, *9*, 90–95.

[274] Marcos-Alcalde, I.; Setoain, J.; Mendieta-Moreno, J. I.; Mendieta, J.; Gomez-Puertas, P. MEPSA: minimum energy pathway analysis for energy landscapes. *Bioinformatics* **2015**, *31*, 3853–3855.

[275] Dijkstra, E. A Note on Two Problems in Connation with Graphs Numerische Mathematik. 1959.

List of publications

- HP35 Protein in the Mesopore of MIL-101 (Cr) MOF: A Model to Study Cotranslocational Unfolding
  Jash, Oishika; Srivastava, Anand; Balasubramanian, S.* *ACS Omega* **2024**, 9, 28, 31185-31194.

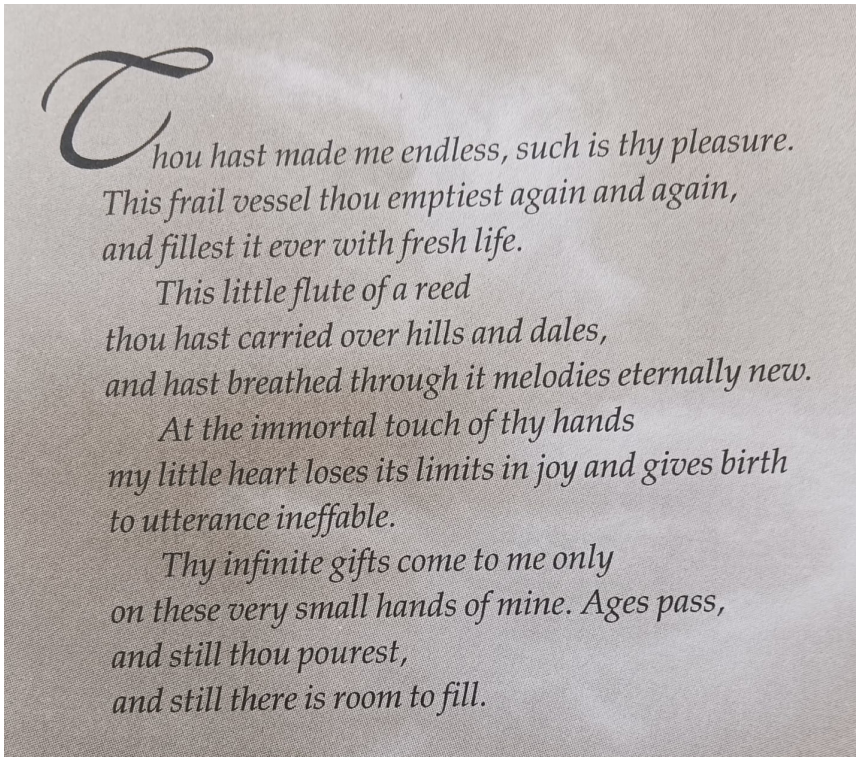
- Insights towards the succinimide formation mechanism in MjGATase
  Chandrashekarmath, Anusha; **Jash, Oishika**; Singh, Karan; Bellur, Asutosh; Sen Roy, Chitralekha; Balaram, Padmanabhan; Balasubramanian, S.; Balaram, Hemalatha* *(Manuscript under preparation)*.

Chem Archived:

- Protein in Metal-Organic Frameworks (MOFs): A Molecular Dynamics Study
  Jash, Oishika; Balasubramanian, S.* *ChemRxiv* **2024**, 10.26434/chemrxiv-2024-c8x0p.

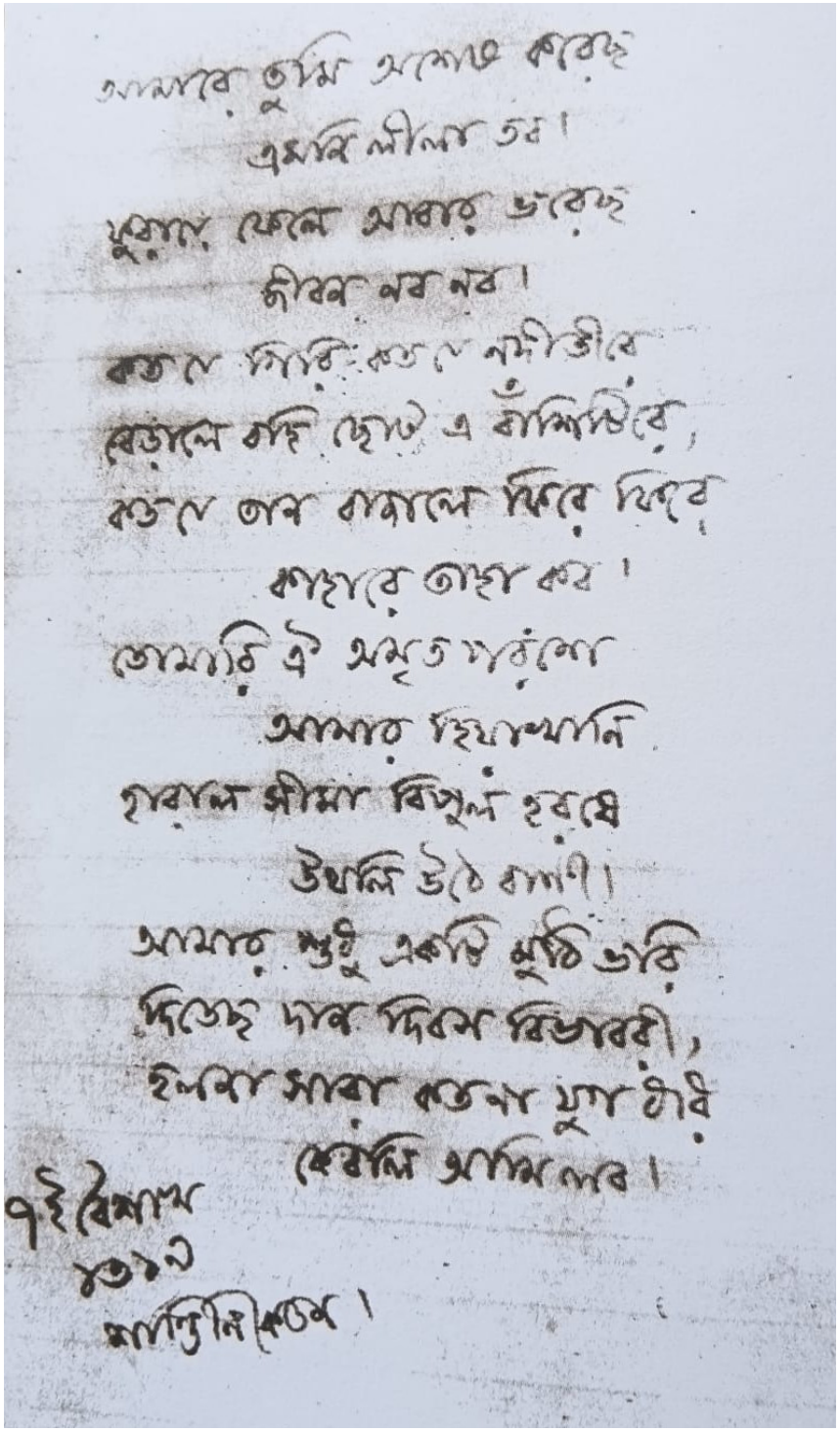
- Co-translocational Unfolding of HP35 in MIL-101(Cr) MOF
  Jash, Oishika; Srivastava, Anand; Balasubramanian, S.* *ChemRxiv* **2024**, 10.26434/chemrxiv-2024-t0zn6.

And, here comes the end of the voyage …



Thou hast made me endless, such is thy pleasure.
This frail vessel thou emptiest again and again,
and fillest it ever with fresh life.
This little flute of a reed
thou hast carried over hills and dales,
and hast breathed through it melodies eternally new.
At the immortal touch of thy hands
my little heart loses its limits in joy and gives birth
to utterance ineffable.
Thy infinite gifts come to me only
on these very small hands of mine. Ages pass,
and still thou pourest,
and still there is room to fill.

From: Tagore, Rabindranath. Gitanjali (Song Offerings). New Delhi, UBS Publishers' Distribution Pvt. Ltd. (UBSPD) in association with Visva-Bharati, Santiniketan, 2014.



From: Tagore, Rabindranath. Gitanjali (Song Offerings). New Delhi, UBS Publishers' Distribution Pvt. Ltd. (UBSPD) in association with Visva-Bharati, Santiniketan, 2014.