Extending the Applications of Boxed Molecular Dynamics: from Simulating Atomic Force Microscopy Experiments to Sampling Trajectories from Virtual Reality

Sarah Jane Mapplebeck

The University of Leeds

School of Chemistry

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

May 2022

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 4 which features in the jointly authored publication:

Mapplebeck, S.; Booth, J.; Shalashilin, D. Simulation of Protein Pulling Dynamics on Second Time Scale with Boxed Molecular Dynamics. *J. Chem. Phys.* **2021**, *155* (8), 085101. https://doi.org/10.1063/5.0059321.

The work, figures and writing in this publication was done by myself, with the exception of some of the theory written by D. Shalashilin. Both D. Shalashilin and J. Booth assisted with the overall editing of the publication.

Acknowledgements

Many thanks go to Dmitrii Shalashilin for his supervision and guidance throughout my PhD. Special thanks go to Robin Shannon for his help and patience when taking on the unofficial role of a postdoctoral mentor. Additional thanks go to David Glowacki for the advice and feedback on manuscripts as well as affording me the opportunity of spending time within his group.

None of this work would have been possible without the use of university facilities, especially ARC HPC systems, nor without the generosity of Dr Simon Waterworth Scholarship Fund. The studentship provided has allowed me to peruse my goal of completing a PhD without the additional worries that come from self-funding, something for which I am extremely grateful.

A massive thank you to all my friends and family for the support they have given me throughout this process. In particular, I would like to thank my mum, dad, grandma, grandad, Karen and Robert who have sat thorough my endless rants and joined in my occasional celebrations.

Abstract

This thesis presents Boxed Molecular Dynamics (BXD) as a useful method of accelerated sampling. It can be used to circumnavigate the rare event problem conduct simulations on extremely long timescales inaccessible to other forms of molecular dynamics, as well as to tackle complex problems for which a simple reaction coordinate cannot be defined and must be described in multidimensional collective variable space.

The BXD method is discussed in both its most primitive one-dimensional form as well as after extension to multidimensional collective variable space. This is followed by a presentation of two new developments to the BXD method, which advance the scope of BXD simulations.

1) Using a one-dimensional reaction coordinate, protein unfolding Atomic Force Microscopy experiments are simulated over a range of pulling velocities. Modifications to the results of unbiased BXD simulations combined with solution of the kinetic master equation allows Atomic Force Microscopy to be modelled at pulling speeds inaccessible to other forms of simulations, helping bridge the gap between experimental and computational methods.

2) A new simulation pipeline for generating free energy surfaces is introduced in which trajectories from virtual reality are used to both define a set of collective variables for the system and as a path for the dynamics to follow. Results for three test systems, each presenting their own unique challenges are reported as a proof of concept for the method. Lastly, as a final validity check a free energy profile is generated for the unfolding of I27 and compared to a previously published version taken from BXD simulations in CHARMM.

Abbreviations

AFM	Atomic Force Microscopy
AXD	Accelerated Molecular Dynamics
BXD	Boxed Molecular Dynamics
CF	Cystic Fibrosis
CV	Collective Variable
CVMD	Constant-Velocity Molecular Dynamics
DR	Dimensionality Reduction
FC	Force Clamp
FPT	First Mean Passage Time
HS-FS	High-Speed Force Spectrometry
LEMS	Lambert–Eaton Myasthenic Syndrome
MD	Molecular Dynamics
MFPT	Mean First Mean Passage Time
MM	Molecular Mechanics
NMR	Nuclear Magnetic Resonance
PES	Potential Energy Surface
PMF	Potential of Mean Force
PCA	Principal Component Analysis
PCs	Principal Coordinates
REMD	Replica Exchange Molecular Dynamics
TSEs	Transmissible Spongiform Encephalopathies
TS	Transition State
TST	Transition State Theory
VC	Velocity Clamp
VGCC	Voltage-Gated Calcium Channels
WHAM	Weighted Histogram Analysis Method
WLC	Worm Like Chain

Table of Contents

List of Figures

Chapter 1: Molecular Dynamics

1.1 Introduction

Molecular Dynamics (MD) is a computer simulation method used to study the evolution of atomic and molecular positions with time, subject to relevant interatomic forces. Before computers, researchers were forced to calculate the trajectories of chemical reactions by hand. [1] This is possible for very small systems such as two interacting particles as a solution can be found analytically, but even a small extension in system size increases the difficulty of such methods enormously. The desire for fast and accurate numerical computation gave rise to the use of computers in scientific research.[2,3]

In 1959 Alder and Wainwright published work detailing how MD can be used to simulate perfectly elastic collisions between hard spheres, which was later followed by Rahman's 1964 study of liquid argon using a Lennard-Jones potential.[2,3] Due to the power of computers growing, it was not long after this until the first MD simulation of a protein was conducted in a landmark study by Martin Karplus.[4] Although the simulation was short and used potentials considered inaccurate when compared to those of the modern day, the significance of this study cannot be overestimated. Demonstrating a protein as a dynamic object whose structure fluctuates shone a light on the role motion plays in the structure and function of biological molecules. Since then, MD has become increasingly helpful in understanding biological processes including molecular transport[5], enzyme catalysis[6] and conformational changes[7].

1.2 Theory

In a classical MD simulation a system comprised of N bodies, which are the nuclei of N atoms, is propagated forward in time from its initial state using Newton's equations of motion. The interactions between the bodies are modelled using a potential energy

function, *V*, which for a conservative system is related to the force acting on the system of atoms through:

$$\vec{F}(t) = -\nabla V\left(\vec{r}(t)\right) \tag{1.1}$$

Where $\vec{\mathrm{r}}(t) \in \mathbb{R}^{3\mathrm{N}}$ is the vector of atomic positions at time *t* and $\vec{\mathrm{F}}(t) \in \mathbb{R}^{3\mathrm{N}}$ is the vector containing the corresponding forces acting on each atom.

Using Newton's second law the acceleration of the atoms in the system at time *t* can be expressed as:

$$\vec{a}(t) = \mathbf{M}^{-1}\vec{F}(t) \tag{1.2}$$

where $\mathbf{M}^{-1} \in \mathbb{R}^{3\mathrm{N}\,\mathrm{x}\,3\mathrm{N}}$ is the inverse of the diagonal matrix containing the masses of each atom and $\vec{a}(t) \in \mathbb{R}^{3\mathrm{N}}$ is the resulting vector of the atomic accelerations. For *N*-bodied systems the above equations cannot be solved analytically and are instead solved numerically over small iterations of time. [2,8] Perhaps the most widely used method of propagation in MD takes the form of the velocity Verlet integration.[9] This is different expression of the original Verlet algorithm[10] in that it explicitly includes velocity into the propagation equations, offering the advantage of being a self-starting method.

In this method, an MD trajectory is considered as a series of 'frames' separated by a small time step, δt, each of which represents a different molecular configuration. For each time step the atomic positions, forces and accelerations are used to propagate the dynamics forward to the time step $\mathrm{t} + \delta\mathrm{t}$, using the following equations:

$$\vec{r}(t+\delta t) = \vec{r}(t) + \delta t\vec{v}(t) + \frac{1}{2}\,\delta t^2\vec{a}(t) \tag{1.3}$$

$$\vec{v}(t+\delta t) = \vec{v}(t) + \frac{1}{2}\,\delta t\left(\vec{a}(t) + \vec{a}(t+\delta t)\right) \tag{1.4}$$

Appendix 1 shows how equations (1.3) and (1.4) are obtained from Taylor expansions around the atomic coordinates and velocities at time *t*.

Iterating over equations (1.1)-(1.4) gives the standard implementation scheme of an MD simulation. A system is propagated according to the underlying potential energy function using the velocity Verlet algorithm as follows:

1. Starting at an initial time *t*, apply potential energy function, *V(r(t))* to the system to return the potential energy as function of atomic coordinates.

2. Use equation (1.1) to get the associated forces acting each atom in the system.

3. Apply Newton's second law (equation (1.2)) to get the atomic accelerations at time *t.*

4. Use the atomic velocities and accelerations to propagate the dynamics forward to time $t + \delta t$.

5. Repeat steps 1- 4 until the simulation reaches an end.

Another common method of integration is the leapfrog algorithm. Simulations follow the same method of iteratively solving Newton's equations of motion, but with velocity calculated at half time steps:

$$\vec{v}\left(t + \frac{\delta t}{2}\right) = \vec{v}\left(t - \frac{\delta t}{2}\right) + \vec{a}(t)\delta t \tag{1.5}$$

$$\vec{r}(t + \delta t) = \vec{r}(t) + \vec{v}\left(t + \frac{\delta t}{2}\right)\delta t \tag{1.6}$$

This scheme requires the specification of $\vec{v}\left(t - \frac{\delta t}{2}\right)$ for its initiation. This can be done using the Euler method of integration, where the velocities at time $t - \frac{\delta t}{2}$ are given by:

$$\vec{v}\left(t - \frac{\delta t}{2}\right) = v(0) - \frac{\delta t}{2}\vec{a}(0) \tag{1.7}$$

Where $v(0)$ and $\vec{a}(0)$ are the initial velocities and accelerations respectively.

To conduct an MD simulation, two things must be in place: a set of initial conditions and a suitable potential energy function. The initial conditions required for starting a simulation are the starting coordinates, usually taken from experimental data such as that found on the Protein Data Bank, and the initial velocities calculated from a Boltzmann distribution.

There exists a wide variety of models for representing the way in which the particles of a system interact, but the choice of one over the other comes from weighing up the need for accuracy in a simulation against the computational expense of running it. The most accurate, and the most computationally expensive of these are the *ab initio* methods, which compute the forces acting on the nuclei from electronic structure calculations 'on the fly' as the trajectory is generated.[11,12] Although these methods are extremely accurate, the huge cost of running these simulations means they are limited to small systems and short timescales.

On the other end of the scale lies the molecular mechanics (MM) method of generating potential energy functions. These potentials, or force fields, do not model electronic structure but rather approximate the quantum mechanical energy surface using classical mechanics, reducing the cost of simulations enormously.[13] The sacrifice of accuracy for savings in computational expense means MM is unsuitable for studying processes such as bond breaking and formation, but excellent for modelling the dynamics of large systems like proteins over longer timescales. The initial model of the system is obtained from experimentally determined structures or comparative modelling data, which for

large biologicals with unfixed structures such as proteins, will be the atomic positions averaged over time.[14]

From here on in all MD simulations are classical, using MM force fields.

1.3 Molecular Mechanics

Once an initial model is in place, the forces acting on each atom are obtained by differentiating the force field equations, which relate the potential energy of the system to the molecular structure.[14,15] But before this can be done the equations of the force field must be derived. To do this, atom types are defined by their bonding and hybridisation[16] (Figure **1.1**(a)), before parameterising interactions with other atoms depending on their interconnectivity (Figure **1.1**(b)).

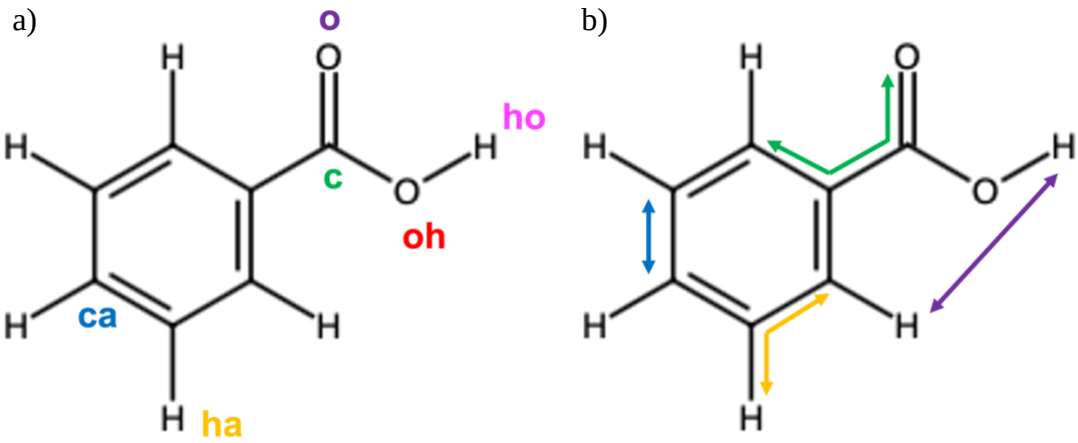


Figure 1.1: (a) Common atom types found in force fields as defined in reference [16]. They include sp^2 hybridised carbons of aromatics (blue) and carbonyls (green), hydrogens bonded to aromatics (gold) and hydroxyl oxygens (pink), as well as oxygens bonded to one atom (purple), or found as part of a hydroxyl group (red). (b) Bonded, angle, dihedral and non-bonded interaction types found in MM force field shown in blue, gold, green and purple respectively.

For cost and efficiency reasons, the force fields treat intramolecular interactions classically.[13,15,17] Covalently bonded atoms are usually parameterised in terms of the sum of three components: the bond, angle, and dihedral terms. The bonded and angle interactions are modelled using harmonic, spring like potentials which each depend on a force constant k_b or k_θ describing the strength of the interaction between the two atoms at a distance *r* or an angle θ:

$$V_{bonds} = \sum_{bonds} k_b(r - r_0)^2 \tag{1.8}$$

$$V_{angles} = \sum_{angles} k_\theta(\theta - \theta_0)^2 \tag{1.9}$$

In the above equations r_0 and θ_0 are the equilibrium distance and angle between the atoms respectively.

Although the dihedral interaction can take a few forms[18], the most common representation uses the cosine of the dihedral angel:

$$V_{dihedrals} = \sum_{diheadrals} k_d(1 + cos(n\phi - \phi_d)) \tag{1.10}$$

Here, k_d is the force constant representing the strength of the interaction, *n* is the periodicity, ϕ_d is the phase and ϕ is the dihedral angle between the four atoms as defined by two planes.

Comparatively, atom pairs at a distance of four or more bonds apart are considered as non-bonded and are considered through their van der Waals interactions. Such interactions are modelled by Lennard-Jones potential, which is repulsive at short distances, attractive at intermediate distances and zero over long ranges [13,15,17] :

$$V_{VDW} = \sum_{i>j} 4\varepsilon \left[\left(\frac{\sigma}{R_{ij}} \right)^{12} - \left(\frac{\sigma}{R_{ij}} \right)^{6} \right] \quad (1.11)$$

Here, ε is the well depth of the potential, R_{ij} is the distance between atoms *i* and *j* and σ is the value of R_{ij} at which the potential is 0. In addition to van der Waals interactions, non-bonded atoms are also considered through their electrostatic interactions, as given by Coulomb's law:

$$V_{elec} = \sum_{i>j} \frac{1}{4\pi\varepsilon} \frac{q_i q_j}{R_{ij}} \quad (1.12)$$

where q_i and q_j are the partial charges of atoms *i* and *j* and here ε is a dielectric constant describing the reduction in the electrostatic force between the atoms arising from surrounding dielectric materials.

The total energy of the system under a MM force field, V_{tot}, is written as the sum of all these potentials:

$$V_{tot} = V_{VDW} + V_{elec} + V_{bonds} + V_{angles} + V_{dihedrals} \quad (1.13)$$

Figure **1.2** shows the different atomic interactions and their corresponding potentials which contribute to V_{tot}.

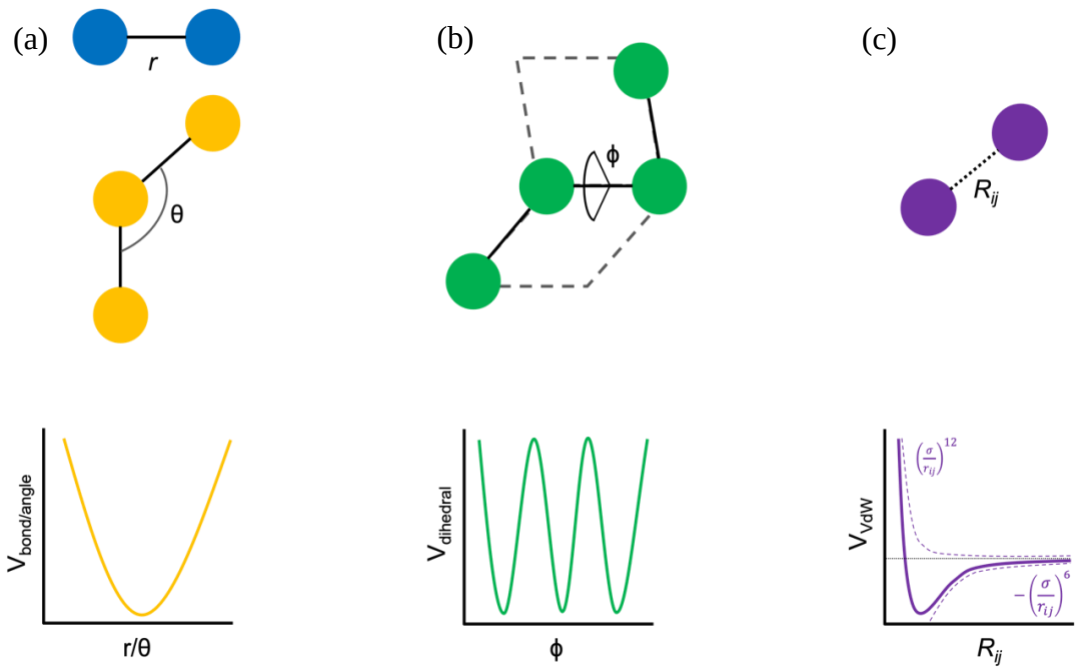


Figure 1.2: Atomic interactions and corresponding potentials used in a MM force field. (a) The harmonic potential used to model V_{bond} and V_{angle} and the corresponding interactions of two atoms at a distance of r (blue) and an angle of θ (gold). (b) the dihedral potential used to represent the interaction between 4 atoms at an angle ϕ defined between the two planes shown by dashed lines. (c) the van der Waals potential comprising of the repulsive and attractive (dashed) potentials of two atoms at a distance of R.

1.4 Reaction Coordinates

The progression of an MD simulation along a reaction pathway can be quantified in terms of its position in collective variable (CV) space. A CV is a low-dimensional degree of freedom, which can be used to describe the molecular structure of system and monitor the state of the simulation. A single CV is referred to as a reaction coordinate, ρ, and is sometimes all that is required to describe a molecular process.

Lots (but not all, see section 1.6) of enhanced sampling MD methods require a predefined CV system to accelerate sampling. Examples of such methods include umbrella sampling [19] and milestoning. [20–22] For low-dimensional systems, chemical intuition alone may be enough to determine the degrees of freedom crucial in reaction

process sufficiently. For example, the change in end-to-end distance of a protein may be a good reaction coordinate for describing protein unfolding (see Chapter 4).

However, for more complex systems (discussed in Chapter 5) identification of the multiple CVs needed to describe the molecular state of the system is more convoluted and less intuitive. In this case, a set of CVs can be returned from a principal component analysis as a linear combination of interatomic distances, multiplied by a coefficient representative of the degree to which the change in each distance is important in describing the molecular process. More details of this can be found in section 5.2.3.

The BXD method central to this thesis requires the identification of a suitable reaction coordinate or set of CVs for the molecular process under investigation. The work in the following chapters shows how BXD can be used to accelerate molecular trajectories along a one-dimensional reaction coordinate enabling extremely long timescales to be reached within a simulation. Additionally, through the development of a new simulation pipeline the need to derive complex sets of CVs by hand is avoided as their acquisition becomes 'blackbox' in nature. Rather, the user need only analyse the results of the principal component analysis to determine if sufficient CVs have been used to capture the main structural variance along a reaction pathway, before BXD is used to accelerate the dynamics through this new CV space.

1.5 The Rare Event Problem

MD propagation from a set of initial conditions results in a trajectory which is a sample of all possible ones, each of which is sensitive to minute changes in the starting conditions.[23] When running an MD simulation, it is easy to assume that with only a good set of initial conditions and a suitable force field the process expected to happen, will be that which is observed. In reality another problem must be overcome: the timescales on which interesting biological processes occur are usually longer than the length of MD simulations.

Events such as protein folding and protein-ligand binding take place on timescales inaccessible to atomistic MD simulations. They occur over milliseconds or longer. Reaching such lengths using typical simulation time steps of around one femtosecond

would require too much computational power to simulate sufficient time steps.[24,25] In fact, Kolinski *et.al* noted that even with a purpose built supercomputer dedicated to atomistic MD it is still only possible to simulate the folding of small, relatively fast folding proteins.[26,27] Additionally, force fields parameterised using data from short timescale simulations have been shown to decrease in validity when their use is extended to the timescales seen in protein folding.[28]

The rate of reaction for these long timescale processes is controlled by the shape of their free energy landscape, which often contain many barriers between meta-stable states. Given the common representation of an initial and product state separated by an activation energy barrier, $\Delta G^{\ddagger}$, used to describe a reaction process, transition state theory (TST)[29] describes the rate of reaction at a given temperature as:

$$\mathrm{k(T)} \propto \left(\frac{-\Delta G^{\ddagger}}{\mathrm{RT}}\right) \tag{1.14}$$

Here, k is the rate of reaction at a given temperature T and R is the universal gas constant. Importantly, equation (1.14) shows that the rate of reaction decreases as the height of the energy barrier increases.

The energy landscape of a real system is unlikely to be as simple as a single reactant and product state separated by one energy barrier. Protein folding as an example, is generally accepted to occur on a globally 'funnelled' energy landscape.[30] Figure **1.3** shows an example of such an energy landscape.

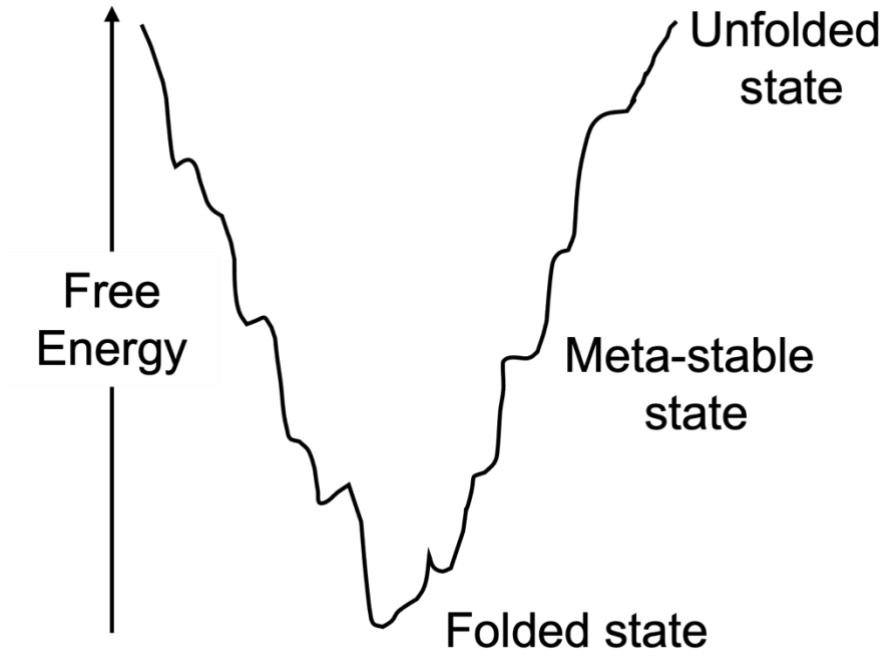


Figure 1.3: Protein folding is generally accepted to happen down a funnel shaped potential energy surface, along which exists meta-stable conformations exist in local minima.

These landscapes are largely directed downwards towards the native state of the protein but exhibit many local minima along the path. This explains the tendency of proteins to eventually return to their native folded state, but why the process occurs over long timescales as there are many pathways through local meta-stable states in which the protein can get trapped in along the way.

Since propagation of an MD trajectory is controlled by the underlying shape of the potential energy surface (PES), if such an energy well is entered throughout the course of a simulation, it can be difficult to get out. If the gradient surrounding the minima is steep then the longer it takes to escape the meta-state, just as described by TST above. This is how the rare event problem in MD arises.

1.6 Addressing the Long Timescale Problem

Addressing the long timescale problem can be done using in various types of accelerated sampling methods which work by controlling a variable within the simulation so that it becomes biased towards sampling the rare event. Some of these methods have the determination of one or more CVs as a prerequisite to simulation.

What follows is a series of examples of accelerated sampling methods. For reasons of simplicity, those requiring the predetermination of any CVs will be discussed in one dimension, with acceleration along a single reaction coordinate.

Accelerated sampling methods can be categorised into three types as follows:

1. **Temperature based methods.** In these methods increasing the temperature is used to overcome energy barriers which would otherwise prevent sampling certain regions of the PES.

2. **Potential energy biasing methods**. In such methods biasing potentials are applied along the reaction coordinate such that energy barriers are easier to cross.

3. **Reactive Flux Methods.** These are based on TST and divide the reaction coordinate of the system into sections, before rate constants are calculated for transitioning between them.

1.6.1 Temperature Based Methods

Increasing the temperature of simulations is a good way to increase the rate at which phase space is explored, as there is more kinetic energy available to overcome barriers along the PES. Replica exchange molecular dynamics (REMD) is one such method that uses this technique to overcome the rare event problem inherent to conventional MD simulations.

1.6.1.1 Replica Exchange

In this method several copies of the same system, known as replicas, are simulated in parallel at different temperatures. The individual replicas cannot interact with one another, but attempts to swap neighbouring configurations are made periodically and accepted with a probability based on the Metropolis criterion.[31] Configurations sampled at high temperature can therefore transition to simulations running at lower temperature (and vice versa if the probability of changing states is accepted). Consequently, simulations can run with sufficient kinetic energy to cross barriers before cooling into minima that would have previously gone unsampled, thus enabling the rare events to be sampled. A schematic representation of the replica exchange method is shown in Figure **1.4**

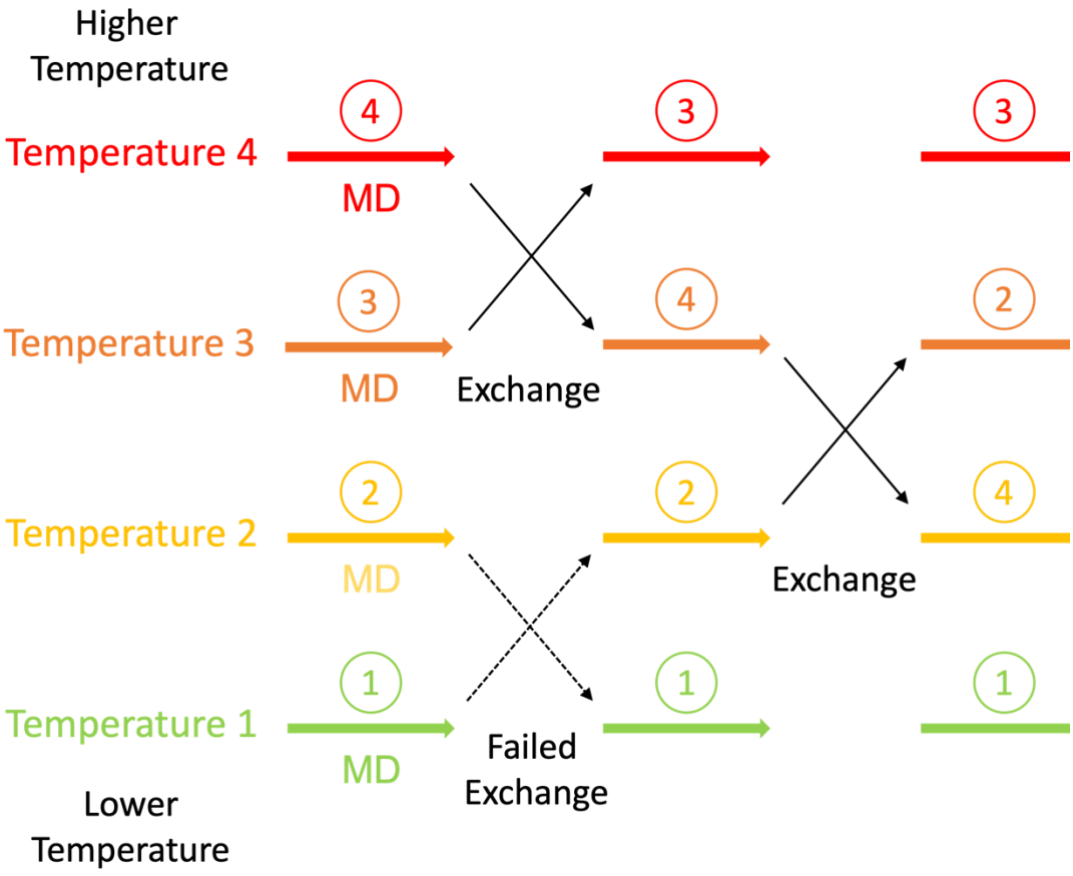


Figure 1.4: Schematic of the replica exchange method. Configurations from high temperature runs can switch with those from lower temperatures allowing energy barriers to be overcome before the system cools into previously unreached minima.

Replica exchange offers the advantage of sampling phase space without portioning it into sections or adding biasing potentials and therefore determination of the reaction coordinate is not a prerequisite for such simulations. The method has been proven useful in the study of protein unfolding and self-assembly.[32–34] However, the need to run many simulations in parallel makes the method computationally expensive, requiring careful consideration of the number of temperatures and runs to use .[35]

1.6.2 Potential Energy Biasing methods

These methods of accelerating MD involve modifying the potential energy of the system to make it easier for the trajectory to cross over energy barriers. Prior knowledge of the CVs or reaction coordinate for the system, ρ, is needed in these methods, an example of which is umbrella sampling.[36,37]

1.6.2.1 Umbrella Sampling

In this method of accelerated sampling, the potential energy function is modified by the addition of biasing potentials called umbrellas, which work to push the trajectory over energy barriers. For a one-dimensional system, the overall modified potential can be expressed as a function of the position along the reaction coordinate:

$$V'(\rho) = V(\rho) + W(\rho) \tag{1.15}$$

where $V'(\rho)$ is the newly modified potential energy function, $V(\rho)$ is the original potential and $W(\rho)$ is the biasing umbrella, usually expressed as a harmonic potential:

$$W(\rho) = k(\rho - \rho_0)^2 \tag{1.16}$$

Here k is a spring constant and ρ_0 is the centre of the umbrella. In this type of sampling many independent simulations are run, with umbrellas centred at different values of ρ_0.

These umbrellas work to confine each trajectory to a given 'window' of the reaction coordinate, although neighbouring windows overlap to ensure the entire system is sampled. Figure **1.5** shows how umbrellas placed along the reaction coordinate help to drive a trajectory over any energy barriers encountered.

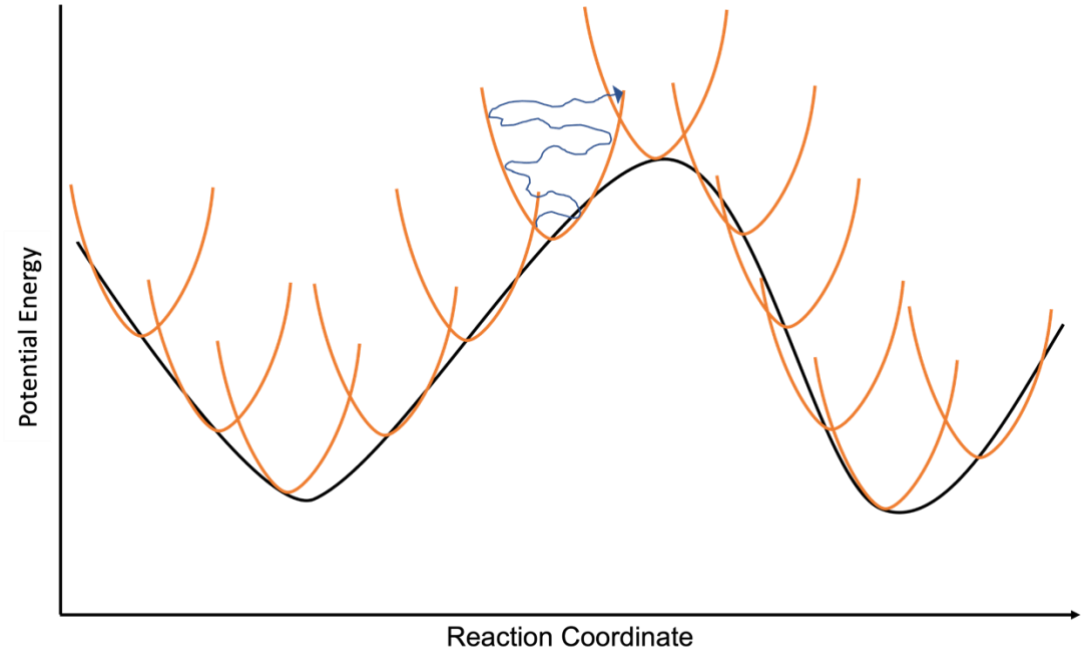


Figure 1.5: A series of overlapping harmonic potentials are added to the underlying potential energy surface in umbrella sampling. Independent MD simulations are run with the dynamics constrained by an umbrella such that higher energy regions of phase are easier to access and the sampling is accelerated, whilst overlapping of umbrellas ensures the entire system is explored.

After the simulations have been run, the statistics gathered must be unbiased if the actual free energy of surface is to be recovered. This is done using the weighted histogram analysis method (WHAM) [38,39], the details of which are to be followed.

The biased probability, $P'_i(\rho)$, of observing a state in the *ith* simulation in which the value of the reaction coordinate is ρ is:

$$P_i'(\rho) = \frac{P(\rho)e^{\frac{-W_i(\rho)}{k_B T}}}{\sum_i e^{-\frac{V(\rho)+W_i(\rho)}{k_B T}}} \quad (1.17)$$

As before $V(\rho)$ is the unbiased potential and $W_i(\rho)$ is the biasing potential from the *ith* simulation. $P(\rho)$ is the probability of the reaction coordinate having a particular value at an unbiased free energy as given by the equation[40]:

$$P(\rho) = e^{\frac{-G(\rho)}{k_B T}} \quad (1.18)$$

Obtaining $P(\rho)$ from equation (1.17), allows the unbiased free energy along the reaction coordinate, $G(\rho)$, to be found from a rearrangement of equation (1.18).

$$G(\rho) = -k_B T \ln[P(\rho)] \quad (1.19)$$

Although umbrella sampling is effective at accelerating MD sampling such that the rare event problem can be overcome, the recovery of $P(\rho)$ from equation (1.17) in the unbiasing process can be convoluted. Additionally, finding a suitable choice of biasing potential and window placement can require significant trial and error. However, both these issues can be addressed by using Adaptive Umbrella Sampling.[41]

This is a more advanced version of umbrella sampling which expands the biasing potential to cover the entire reaction coordinate, with variations being made to it on the fly. This is continued until all states in the system are equally populated, at which point the combination of the free energy and biasing potential is a flat surface. Consequently, the free energy of the system can be recovered as the negative of the biasing potential without the need for complicated unbiasing methods.

1.6.3 Reactive Flux Methods

Reactive flux methods are types of accelerated MD which are based on transition state theory.[29,42] Such methods involve splitting the phase space of a system into two, where states A and B are separated by a dividing surface. The flux through this surface defines the rate constant for transition from A to B:

$$k_{AB}^{TST} = \kappa \frac{k_B T}{h} \frac{e^{-\frac{W(\rho^*)}{k_B T}}}{\int_{-\infty}^{\rho^*} e^{-\frac{W(\rho^*)}{k_B T}} dr} \tag{1.20}$$

Where ρ is the reaction coordinate for the process being studied, and ρ^* is the value of ρ at which the dividing surface is located, i.e., the transition state. $W(\rho^*)$ is the work required to move from state A to ρ^* and is reversible, whilst κ represents the fraction of the trajectories which, upon reaching ρ^* go on to reach state B. This is multiplied by the Boltzmann constant, k_B, and temperature of the simulation, T, over Plank's constant, h.

The fraction in equation (1.20) represents the probability of the system reaching ρ^* over the probability of being in state A anywhere before the transition state. Therefore, the rate constant for transition from state A to state B is given by the probability of finding the trajectory at the transition state multiplied by the chance of it crossing into state B and remaining there. This equation is the base of reactive flux methods.

It should be noted that although in equation (1.20) ρ^* is the formal transition state separating states A and B, the same equation can be applied to for arbitrary states even if they are not stable enough to be isolated. It is in this way that reactive flux methods can accelerate MD sampling, by partitioning the phase space of reactions with multiple dividing surfaces and generating rate constants for transition from one region of phase space to another.

1.6.3.1 Milestoning

Milestoning is a type of reactive flux accelerated sampling method which involves dividing up the reaction coordinate into milestones.[20–22] A milestone can be defined as a hypersurface orientated orthogonal to the CV of the system. In the case of a one-dimensional reaction coordinate, a milestone would be a point along the reaction coordinate, in two dimensions a line, in three a plane and so on.

When using milestoning, an ensemble of conformations is first generated at each plane by running conventional MD simulations within each plane. Then, each conformation is run as a separate trajectory initiated from its original milestone and terminating once it reaches a neighbouring milestone as shown in Figure **1.6**.

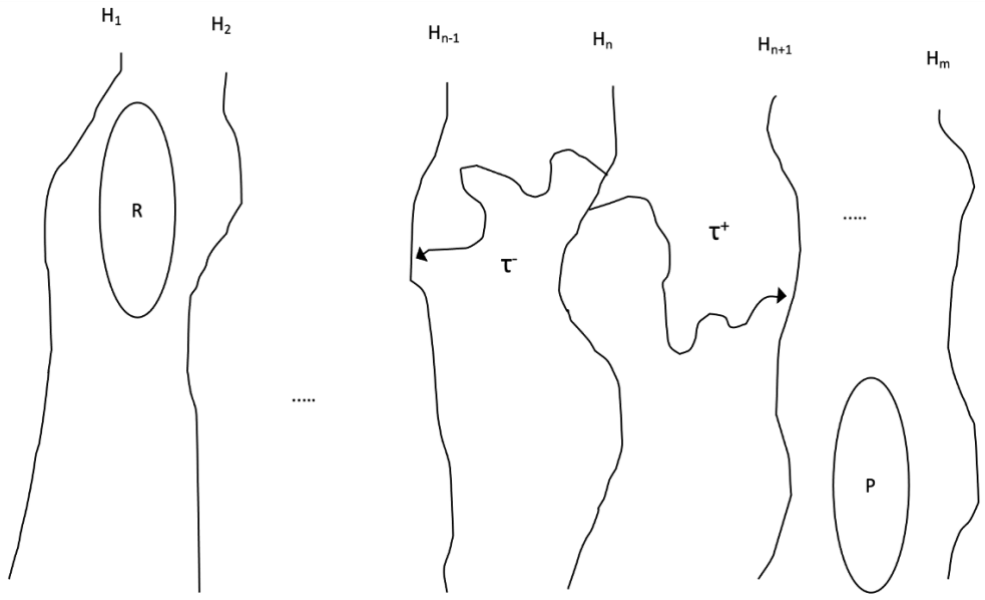


Figure 1.6: A series of planes or milestones separating the reactant, R, and product, P, states of a reaction. Trajectories are set off from one milestone H_n and run until they hit H_{n-1} or H_{n+1} with the time taken to reach the respective milestones recorded as τ^- or τ^+

A trajectory starting at milestone H_n is said to have lifetime of τ^+ if it terminates at H_{n+1} and τ^- if it ends at H_{n-1}. The distribution of lifetimes for going from H_n to $H_{n\pm1}$ is recorded as either $K_n^+(\tau)$ for going to H_{n+1} or $K_n^-(\tau)$ for transition to H_{n-1}. For a system

comprising of *m* planes, K_m^+ and K_0^- are both 0 as above plane H_m and below plane H_0 there are no more planes to reach.

The equilibrium probability of finding a trajectory at a milestone *n*, *Peq (n)*, represents the probability of finding a trajectory at that point along the reaction coordinate. As such, if this was known, the free energy along the reaction coordinate could be found as in equation (1.19).

The probability of finding a trajectory at milestone H_n at time *t* is given by:

$$P_n(t) = \int_0^t \left[1 - \int_0^{t-t'} K_n(t-t')\right] Q_n(t')dt' \tag{1.21}$$

Where integrand is the probability of the trajectory arriving at H_n at time *t'* and remaining there until time *t.* In equation (1.21) the probability of leaving H_n between time *t'* and *t* is defined as $K_n(t-t') = K_n^+ + K_n^-$ and $Q_n(t')$ is given by:

$$Q_n(t') = 2\delta(t)\, P_n(0) + \int_0^t Q_{n\pm1}(t'')\, K_{n\pm1}^{\pm}(t-t'')dt'' \tag{1.22}$$

The above equation describes the probability of transition to plane H_n as the initial conditions of the starting trajectories, $P_n(0)$, added to the sum over the probability of transitioning to $H_{n\pm1}$ before moving to H_n.

Obtaining $P_n(t)$ from equations (1.21) and (1.22) allows the equilibrium probability of finding a trajectory at H_n at time *t* to be found as:

$$P^{eq}(n) = \lim_{t\to\infty} P_n(t) \tag{1.23}$$

Where finding $\lim_{t\to\infty} P_n(t)$ is done by running many trajectories until convergence of the distributions of lifetimes is achieved. Now, with $P^{eq}(n)$ in hand, the free energy along the reaction coordinate can be found using equation (1.19).

Reactive flux methods like Milestoning are advantageous in their ability to calculate both kinetic and thermodynamic data from the same simulation. This is not something which is afforded from techniques such as Umbrella sampling, which only provide thermodynamic data. Additionally, the dynamical information obtained remains meaningful as no biasing potentials are required to accelerate the sampling. However, the large number of simulations required for sufficient sampling means that these methods can become relatively expensive.[43]

Chapter 2: Boxed Molecular Dynamics

Boxed Molecular Dynamics (BXD) is an enhanced sampling MD technique which can be used to accelerate the modelling of rare events. Similar to the reactive flux methods discussed in Chapter 1, these simulations can be used to calculate kinetic and thermodynamic properties of slow processes simultaneously.[40,44–46] However, BXD simulations have the added advantage of being very simple to conduct requiring no prior knowledge of the system under investigation or modification of its potential energy surface. Additionally, only a single trajectory is needed in BXD, which makes the process easier to set up than methods such as Milestoning.

This chapter will detail the principles underpinning BXD simulations, including the assumptions upon which it relies and the conditions under which the method works.

2.1 Accelerated Molecular Dynamics

Continuing the likeness to reactive flux methods, BXD has its roots in Transition State Theory.[47] For a system in a BXD simulation, reflective boundaries are placed along a reaction coordinate to separate the phase space of the system into boxes, into which the trajectory can be confined. Figure **2.1** shows how this is done using velocity inversions. The value of the reaction coordinate is continually monitored throughout the simulation and if it exceeds the value at which the extremities of the box lie, then the velocity is inverted so that the trajectory remains within the box. This is done relative to the reaction coordinate, which for a one-dimensional system gives a new velocity of $v' = v - 2v_{parallel,}$ where $v_{parallel}$ is the component of the vector v lying along the reaction coordinate.

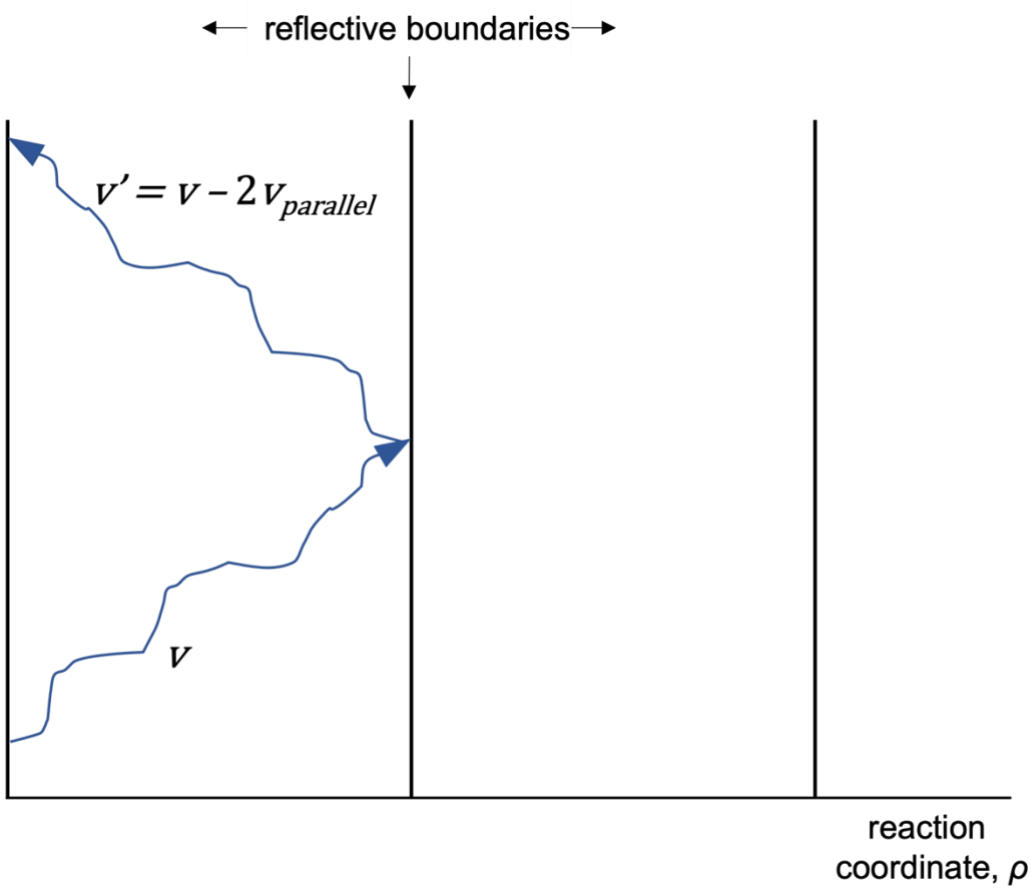


Figure 2.1: Upon collision of the trajectory with a boundary the velocity of each atom is reflected with respect to the reaction coordinate. If the atoms collide with boundary with velocity v and a component along the reaction coordinate of $v_{parallel}$ then the reflected velocity will be $v' = v - 2v_{parallel.}$

The link between BXD and TST is most easily seen when a primitive version of BXD known as Accelerated Molecular Dynamics (AXD) is considered.[46,47] AXD differs from BXD in the number of boxes placed along the reaction coordinate. Figure **2.2** shows the setup of an AXD simulation. Only two reflective boundaries are required: one on the transition state (TS) of a system and one just before. Once the trajectory enters the region of phase space near the transition state, Γ_1 it is prevented from entering Γ_2 or Γ_0 by the reflective boundaries separating these regions.

This differs from TST in the fact that TST assumes the trajectory always crosses the boundary to the product state Γ_0, rather than remaining confined to the region surrounding it. But, provided the boxes are in equilibrium, the reflection of a trajectory

about to collide with a boundary is identical to the reverse of transition across the TS.[48] Thus, if the boxes are in equilibrium, the two approaches become equivalent.

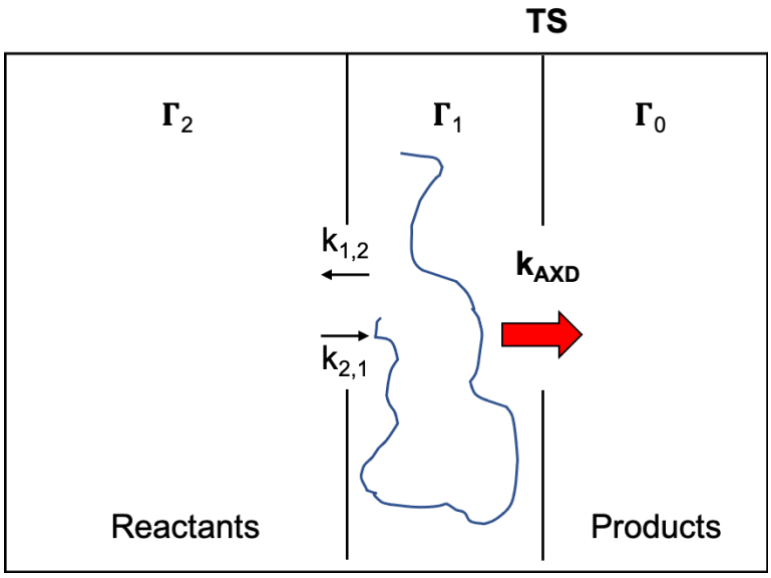


Figure 2.2: The setup of an AXD simulation. Reflective boundaries confine the trajectory to be near the transition state so that this area becomes well sampled and k_{AXD} is converged quickly.

A trajectory which is confined to Γ_1 will sample the area around the TS more often than one which is left to sample the PES freely. Ergo, the reaction rate constant for AXD, $k_{AXD,}$ is accelerated compared to the actual one from TST, $k_{TST.}$ But they are related:

$$k_{TST} = k_{AXD} P_{corr} \tag{2.1}$$

Here, P_{corr} is a correction factor describing the probability of finding the trajectory in $\Gamma_{1.}$ It is equal to the fraction of the phase space found within the region Γ_1 compared to that in both Γ_1 and Γ_2. [46]

$$P_{corr} = \frac{\Gamma_1}{\Gamma_1 + \Gamma_2} \tag{2.2}$$

Describing P_{corr} in terms of rate constants for diffusion across the boundary separating regions Γ_1 and Γ_2 means that equation (2.2) can be written as:

$$P_{corr} = \frac{\Gamma_1}{\Gamma_1 + \Gamma_2} = \frac{1}{1 + \frac{\Gamma_2}{\Gamma_1}} = \frac{1}{1 + \frac{k_{1,2}}{k_{2,1}}} \quad (2.3)$$

Where $k_{1,2}$ is the flux into the region Γ2 and $k_{2,1}$ is that into Γ_1. Equation (2.3) can be derived from TST, in which the rate constant of reaction, k_{TST}, is defined as [29,40]:

$$k_{TST} = \frac{< |\mu| \delta(q,r) \Theta(q,r) >}{\Gamma_R} \quad (2.4)$$

Where $|\mu|$ is the magnitude of the velocity vector normal to the dividing surface in phase space, $\Theta(q,r)$ is a function of the position, r, and momentum, q, of the system which is equal to one when the trajectory is in the reactant region of the system, R, or 0 otherwise and $\delta(q,r)$ is a Dirac delta function equal to one at the dividing surface. It is worth noting that here, k_{TST}, would describe the same transition as k_{AXD} in Figure **2.2**, provided the trajectory hadn't been confine to the region Γ_1.

Recalling that the reactant phase space, Γ_R, is divided into Γ_1 and Γ_2 by a reflective boundary equation (2.4) can be rewritten to reveal k_{AXD}:

$$\begin{aligned} k_{TST} &= \frac{< |\mu| \delta(q,r) \Theta(q,r) >}{\Gamma_R} \\ &= \frac{< |\mu| \delta(q,r) \Theta(q,r) >}{\Gamma_1 + \Gamma_2} \\ &= \frac{< |\mu| \delta(q,r) \Theta(q,r) >}{\Gamma_1} \frac{\Gamma_1}{\Gamma_1 + \Gamma_2} \\ &= k_{AXD} P_{corr} \end{aligned} \quad (2.5)$$

Where $k_{AXD} = \frac{<|\mu|\delta(q,r)\Theta(q,r)>}{\Gamma_1}$ is calculated directly from AXD simulations by confining the trajectory into Γ_1.

The rate constants $k_{1,2}$ and $k_{2,1}$ required for the calculation of P_{corr} can be calculated by molecular dynamics confined to Γ_R split into Γ_1 and Γ_2 with

$$k_{1,2} = \frac{h_{bound}}{t_1} \qquad (2.6)$$

$$k_{2,1} = \frac{h_{bound}}{t_2}$$

Where h_{bound} is the number of hits on the boundary separating Γ_1 and Γ_2 and t_1 and t_2 are the time spent in each respective area of phase space.

AXD is very good at accelerating MD simulations as its much quicker to converge k_{AXD} and P_{corr} separately rather than to converge k_{TST} as one. This is because the trajectory is confined to sampling the area near the TS which, if there was a barrier to reaction, would rarely be visited if the dynamics were allowed to sample the PES freely.

This method is comparable to that suggested by Voter[49], in that the sampling is accelerated by confining the dynamics to regions of phase space near the transition state. The exception being that here they are confined using a reflective boundary placed near the TS, whilst in Voter's method it is additional 'boosting' potentials that work to push the dynamics towards the TS. However, AXD is advantageous in imposing phase space constraints that can be hard to describe by a boosting potential.[47,49]

2.2 The Boxed Molecular Dynamics Algorithm

2.2.1 General method for conducting a BXD simulation

The difference between AXD and BXD is the number of boundaries which are placed along the reaction coordinate, ρ, with BXD using more than two. Placing boundaries at locations other than the formal transition state accelerates the sampling in all areas of phase space whilst remaining a valid approach as TST boundaries do not have to lie only at transition states.[49]

BXD in its simplest form is shown in Figure **2.3**. It is assumed that an atomistic process can be described by a reduced description of the configuration space of the system. In one dimension this is usually referred to as a reaction coordinate and it can be split into multiple boxes into which the dynamics of a trajectory can be locked. To conduct a BXD simulation the trajectory is set off running and is confined to remain within the first box by inverting its velocity as if it had collided with a hard wall upon each collision with a box boundary. After sufficient statistics have been generated for the current box, the trajectory is allowed to diffuse into the subsequent box. Here it becomes confined again until a predetermined number of hits on the boundary to the next box have been recorded and diffusion is permitted once again. A trajectory can only enter the next box along the reaction coordinate in the direction of travel, not the previous one. In this way the boxes push the trajectory along ρ until the final box is reached.

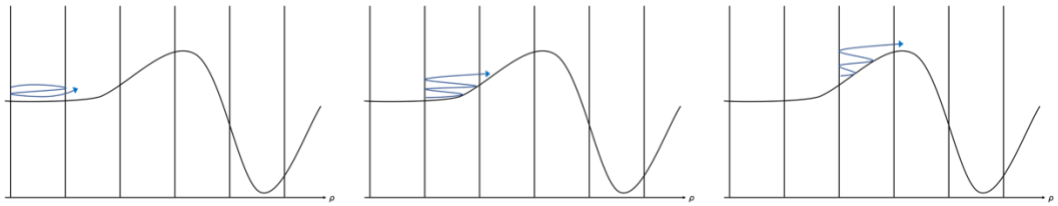


Figure 2.3: Reflective boundaries in BXD help push a trajectory along the reaction coordinate, by only allowing diffusion into the next box in the direction of travel. In this way the BXD boundaries help accelerate trajectories over energy barriers by preventing them from re-entering regions of lower energy.

From here, the course of travel is reversed back down the reaction coordinate until the starting box is reached where the direction is inverted yet again. This is repeated until the entire reaction coordinate has been explored several times in each direction and the sampling converges. Figure **2.4** shows a typical plot of the reaction coordinate vs simulation time step taken from a BXD simulation for the unfolding of a protein – a process in which ρ can be defined as the end-to-end distance of the protein – where sampling is done in both directions multiple times until it converges.

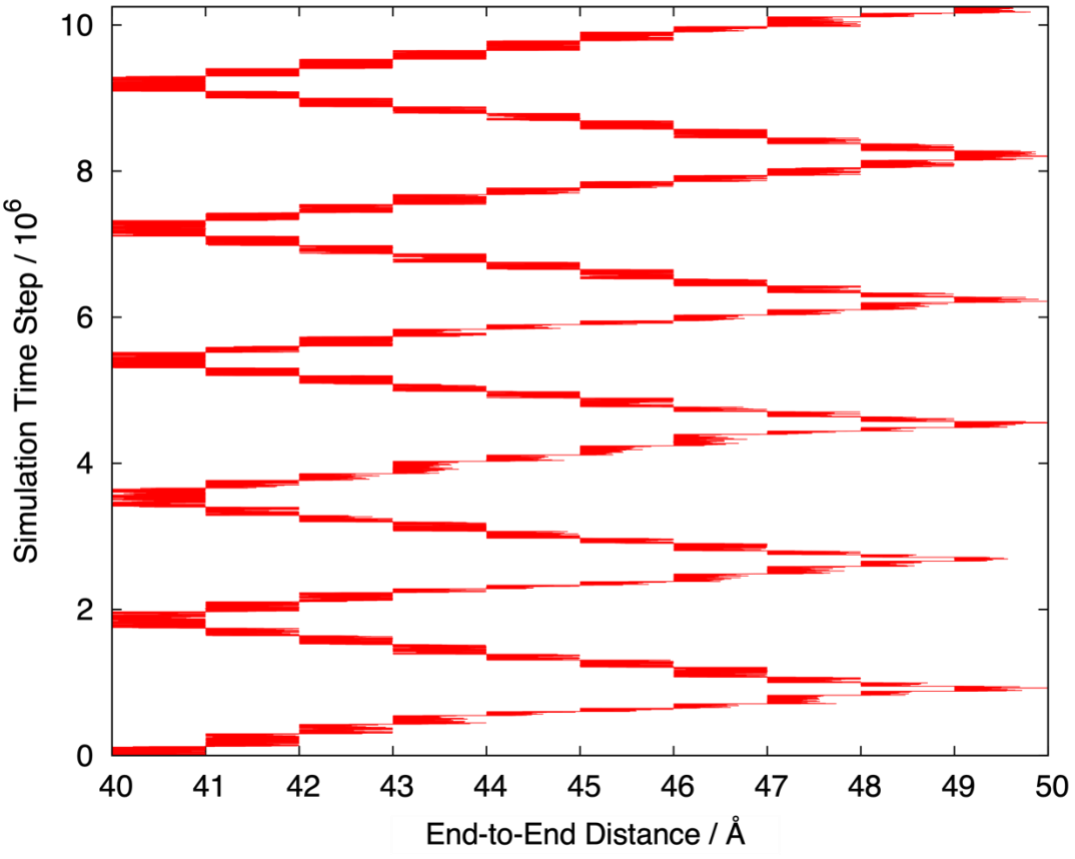


Figure 2.4: A plot of reaction coordinate against the simulation time step demonstrates how a trajectory (red) samples the reaction coordinate multiple times to ensure convergence. Reflective BXD boundaries have been placed at distances of 1 Å along the reaction coordinate.

Division of phase space into boxes not only helps trajectories over energy barriers as highlighted in Figure **2.3**, but also allows calculation of the rate constants and the change in free energy for passing from one box to another, from which the free energy along ρ can be extracted.

Figure **2.5** shows how the BXD method enables calculation of the box-to-box rate constants and subsequent free energies.

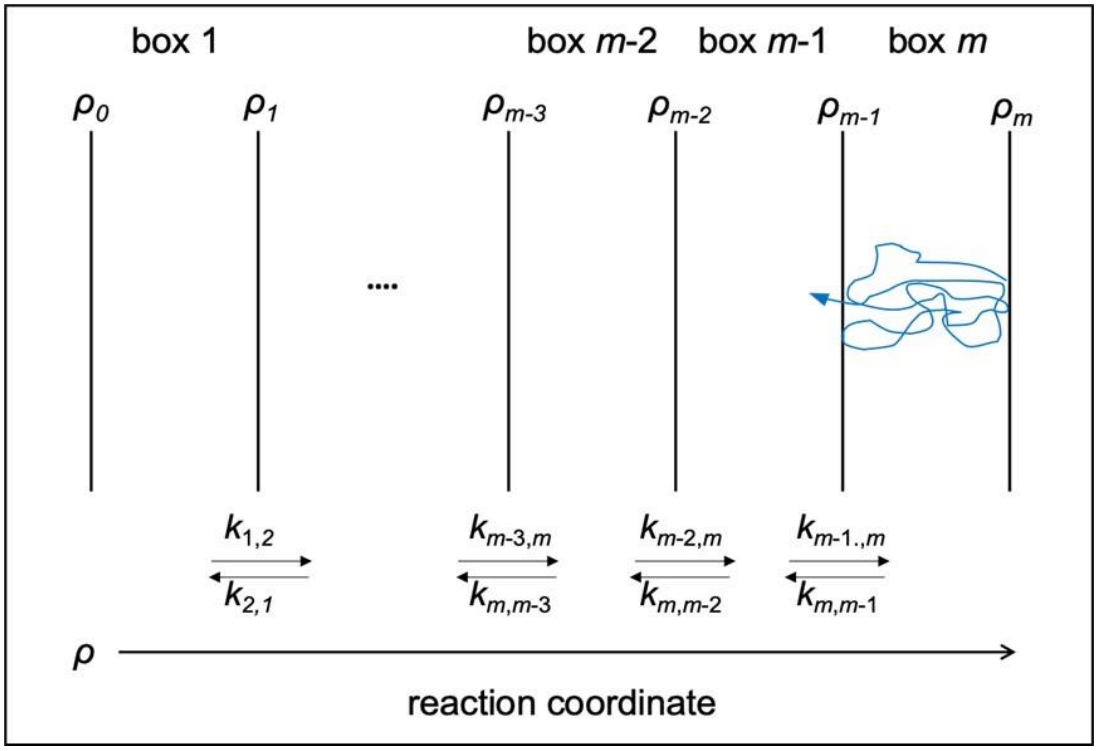


Figure 2.5: Schematic of BXD, where a reaction coordinate, ρ, is split into m boxes into which a trajectory can be confined. After a given number of inversions (two in this case), the trajectory in box m can diffuse across the boundary into box m-1. Dividing the number of hits at boundary $h_{m,m-1}$ by the lifetime of the trajectory in the box gives a rate coefficient for the diffusion into box m-1. This process is repeated until the trajectory has sampled up and down the entire reaction coordinate multiple times generating a set of box-to-box rate coefficients.

For a trajectory locked inside box m the times between successive collisions with boundaries ρ_{m+1} and ρ_{m-1} are referred to as the first passage times (FPTs) for moving up or down a box. Sets of FPTs can be recorded for collisions with both boundaries and used to calculate the rate constants for diffusion up or down a box [40]:

$$k_{m,m+1} = \frac{h_{m,m+1}}{t_m} = \frac{1}{< \tau_{m,m+1} >}$$
$$k_{m,m-1} = \frac{h_{m,m-1}}{t_m} = \frac{1}{< \tau_{m,m-1} >} \qquad (2.7)$$

Here, $k_{m,m+1}$ and $k_{m,m-1}$ are the rate constants for entering box m+1 or m-1 respectively. $h_{m,m+1}$ and $h_{m,m-1}$ represent the number of collisions with either boundary ρ_{m+1} or ρ_{m-1}, and t_m is the time spent inside this box. This is equivalent to the inverse of the mean of

the first passage times (MFPTs) for collisions with the upper and lower boundaries of the *m*th box, $<\tau_{m,\,m+1}>$ and $<\tau_{m,\,m-1}>$.

Having obtained rate constants for diffusion from one box into another, the change in free energy for diffusion between boxes *m*-1 and *m*, $\Delta G_{m-1,m}$, can be calculated as follows:

$$K_{m-1,m} = \frac{k_{m-1,m}}{k_{m,m-1}} = \exp\left(-\frac{\Delta G_{m-1,m}}{RT}\right) \tag{2.8}$$

where $K_{m-1,m}$ is the equilibrium constant for diffusion from box *m*-1 to *m*, *R* is the universal gas constant and *T* is the temperature. Summation of the box-to-box free energy change generates the free energy along the whole reaction coordinate. This can then be differentiated to give the force as a function of the reaction coordinate, $F(\rho)$:

$$F(\rho) = \frac{dG}{d\rho} \tag{2.9}$$

An advantage of the BXD method is that the box-to-box rate constants not only allow for thermodynamic information to be gathered as above, but they also allow the evolution of the box populations to be seen. By obtaining rate constants for diffusion into and out of every box along the reaction coordinate, the dynamics can be reduced to a set of kinetic equations for the time evolution in each box.

$$\frac{dn_1(t)}{dt} = -\big(k_{12}(t) + k_{10}(t)\big)n_1(t) + k_{21}n_2(t)$$

$$\frac{dn_2(t)}{dt} = k_{12}(t)n_1(t) + k_{32}(t)n_3(t) - \big(k_{21}(t) + k_{23}(t)\big)n_2(t) \tag{2.10}$$

……

$$\frac{dn_m(t)}{dt} = k_{m-1,m}(t)n_{m-1}(t) - k_{m,m-1}(t)n_m(t)$$

Equation (2.10) is the kinetic master equation (KME).[5,65–66] $n_1(t)$, $n_2(t)$ and $n_m(t)$ are the populations of the 1st, 2nd and *m*th box as a function of time. The right-hand side of the system of equations is the flux into minus the flux out of each box, where the equation for n_1 assumes diffusion across ρ_0 is irreversible.[48] Equation (2.10) can be written in matrix form:

$$\frac{d\boldsymbol{n}(t)}{dt} = \boldsymbol{M}(t)\boldsymbol{n}(t) \tag{2.11}$$

where $\boldsymbol{M}(t)$ is an N by N sparse matrix of the box-to-box rate constants and $\boldsymbol{n}(t)$ is a vector of length N containing the box populations as a function of time. The solution of equation (2.11) is given by:

$$\boldsymbol{n}(t) = \boldsymbol{U}(t)\Lambda\boldsymbol{U}^{-1}(t)\boldsymbol{n}(0) \tag{2.12}$$

where $\boldsymbol{n}(0)$ contains the initial conditions for each box using a Boltzmann distribution, $\boldsymbol{U}$ is the eigenvector matrix resulting from diagonalisation of $\boldsymbol{M}$, and Λ is a diagonal matrix whose elements, $\Lambda_{ij} = e^{\lambda_j t}$, are determined by λ, the eigenvalue vector corresponding to **M**. The total number of eigenvalue elements in vector $\boldsymbol{\lambda}$ is equal to the number of boxes. Generally, the eigenvalues are all negative and one of them is separated from the rest by orders of magnitude. If the process being studied is an irreversible one then the flux out of the product box will be zero and the smallest eigenvalue can typically be set as the rate constant of the process.[40,52]

2.2.2 Decorrelation and Ergodicity in BXD

2.2.2.1 Decorrelation

BXD assumes the motion of trajectories is stochastic and sequential boundary collisions are uncorrelated. For this to be true, the time between successive boundary hits must be larger than the correlation time such that the trajectory no longer remembers its state

after the previous collision. However, this is often not the case as reflected trajectories can turn back on themselves rapidly causing multiple collisions on the same boundary within such short timescales that the dynamics is still correlated to the previous hit.

To overcome this, FPTs corresponding to short time-correlated events are removed by defining a cut off value, τ_{corr}, below which FPTs are disregarded. This can be done using one of two methods. Firstly, several values of increasing τ_{corr}, can be defined and the free energy of the system calculated for each, with FPTs below τ_{corr} removed from the calculation. This is repeated until the free energies converge, at which point the correct correlation time has been found and any events occurring before this remain ignored.

Alternatively, the survival probability or decay trace, *R(t)*, can be calculated from the lifetime distribution *N(t)* for diffusion across a particular boundary of a given box [48]:

$$R(t) = \int_0^{t_{max}} N(t')dt' - \int_0^{t} N(t')dt' \tag{2.13}$$

Where t_{max} is the maximum value within the lifetime distribution. In other words, the survival probability at time *t* is given by the sum of all the FPTs between 0 and t_{max} minus the sum of those between 0 and time *t*. Inspection of a decay trace such as that seen in Figure **2.6** is used to identify τ_{corr}. Initial steep regions of the trace correspond to dynamically correlated collisions on the given boundary whilst flatter regions of the plot arise from collisions after ergodic exploration of the box. Therefore, τ_{corr} is defined as the FPT at which the steeper region ends and the flatter region begins.

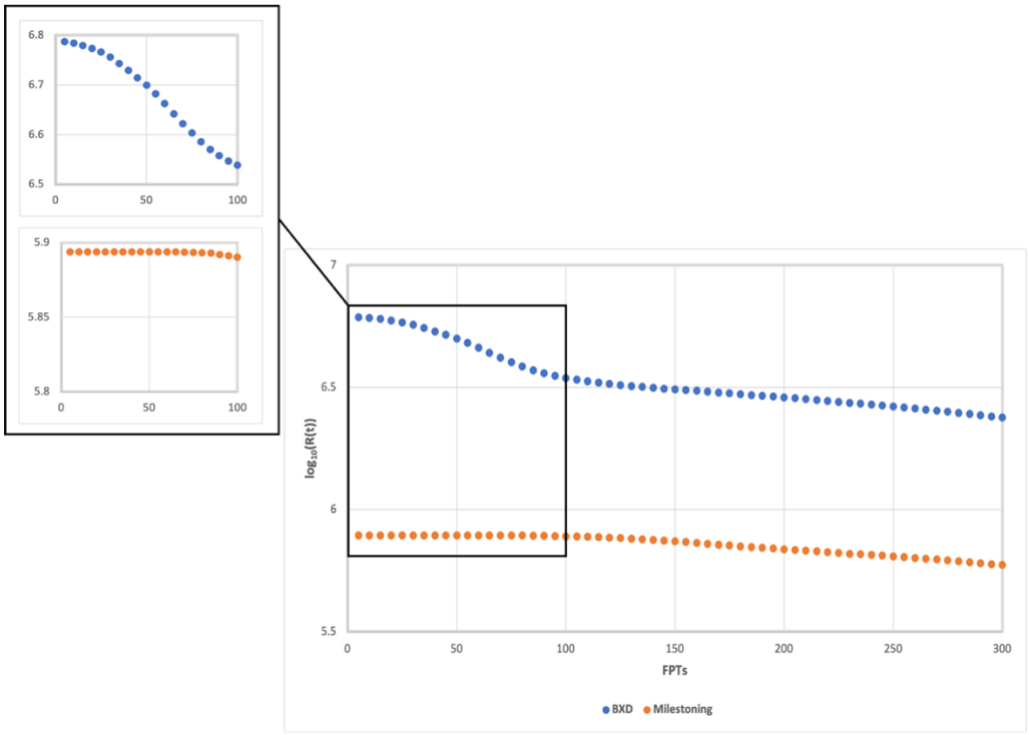


Figure 2.6: Typical decay trace for FPTs of a box boundary in a BXD simulation. Blue points correspond to BXD FPTS whilst orange to Milestoning FPTs. If there is an initial steep region in the decay trace (BXD FPTs) then FPTs in this region are said to be below τ_{corr} and are removed from the free energy calculation. Milestoning FPTs may be used as an alternative BXD FPTs in an effort to avoid need to decorrelate the statistics by hand.

Furthermore, the use of Milestoning FPTs (defined as the number of MD steps between hits on alternate BXD boundaries, see Chapter 2.2.3.5 for more detail) can provide an alternative to decorrelating the statistics by hand after the fact. Providing the boxes are large enough, it is safe to assume that by the time the trajectory reaches the next boundary it will no longer have any memory of the previous collision. The size of a BXD box is important when considering if the system is ergodic, something which is an important assumption in BXD simulations.

2.2.2.2 Ergodicity

An important assumption is made in BXD simulations. Equation (2.2) relies on the assumption that the system under investigation is ergodic and every point in phase space has an equal probability of being explored. But this can only be true if boundary

hits are decorrelated from one another. This requirement imposes a restriction on the size of the box. If the box is too small the time between boundary collisions will be less than τ_{corr} and the trajectory will not have time to relax. Hence, the box length should be larger than the correlation length – the length at which the trajectory loses all memory of its initial conditions.[44] This removes any contribution to the rate coeffects stemming from short time-correlated velocity inversions, improving the quality of the kinetic results.[44]

Inspection of the dynamics within a given BXD box can be used to determine whether or not the box is of an appropriate size. Figure **2.7** shows the differing dynamics between a trajectory confined to a box that is too small (a) and one of an acceptable size (b). If the box is too small, then the trajectory bounces between the boundaries very quickly and remains correlated to the previous hit. Whereas if the box is of an appropriate size, there is ample time to explore the phase space before the next collision. However, it is important to note that the boxes should not be so large as to fail in accelerating the progression of the dynamics along the reaction coordinate.

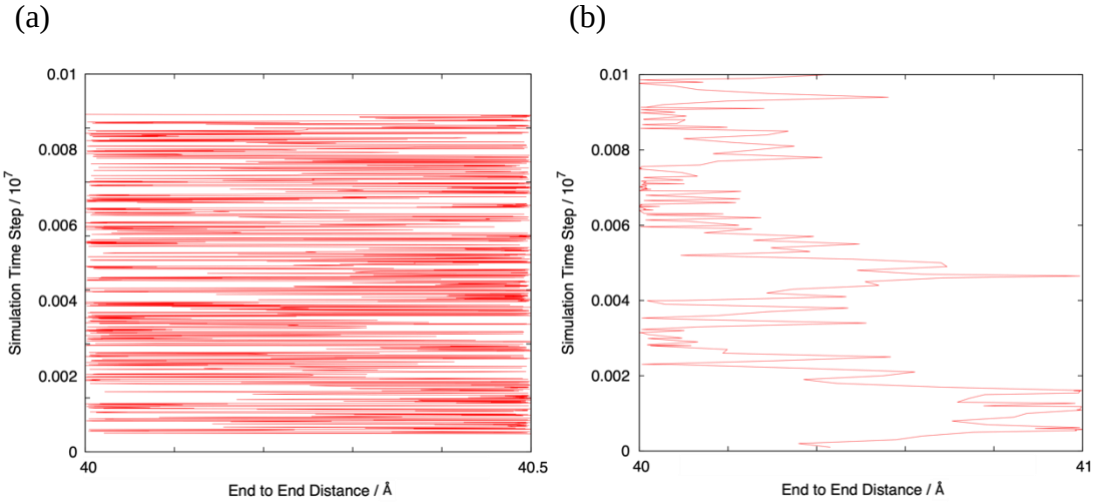


Figure 2.7: (a) If a BXD box is too small the trajectory does not have time to relax between boundary collisions and bounces between boundaries in a ballistic manner rather than exploring all of phase space with equal probability. (b) If the BXD box is larger than the decorrelation length it is possible for the trajectory to come to a state equilibrium with no memory of the previous collision as it explores the box before it's next collision with a boundary.

Consideration of the box sizes in this way is important when using conventional BXD. However, the development of adaptive sampling BXD means such inspections may not

be necessary as boundaries are placed according to the underlying dynamics of the system meaning correlated dynamics from boxes of insufficient size are likely to be avoided.

2.2.3 Adaptive sampling BXD

2.2.3.1 Introduction to Adaptive boundary placing

The BXD method discussed in section 2.2.1 is the simplest form of BXD with boundaries placed at even intervals along a one-dimensional reaction coordinate.

However, more recent work from O'Connor *et. al.* [53] involved developing an algorithm for placing BXD boundaries adaptively based upon the needs of the sampled dynamics. The idea being that the size of the BXD boxes is determined by the shape of the PES. A schematic outlining this idea is shown in
Figure **2.8**. In flat regions, it is easy to explore large regions of configuration space and boundaries can be placed quite far apart from one another. But, as the gradient of the PES increases the volume of configuration space which is readily available for the trajectory to explore decreases resulting in the need for more closely spaced boundaries. Boundaries are placed in this manner until the top of the energy barrier is reached and the trajectory is free to proceed downhill towards the product state without the need to place more boundaries (frame a.) To converge the free energy in the region of space after the energy barrier, boundaries are placed adaptively in the reverse direction using the same methodology (frame b).

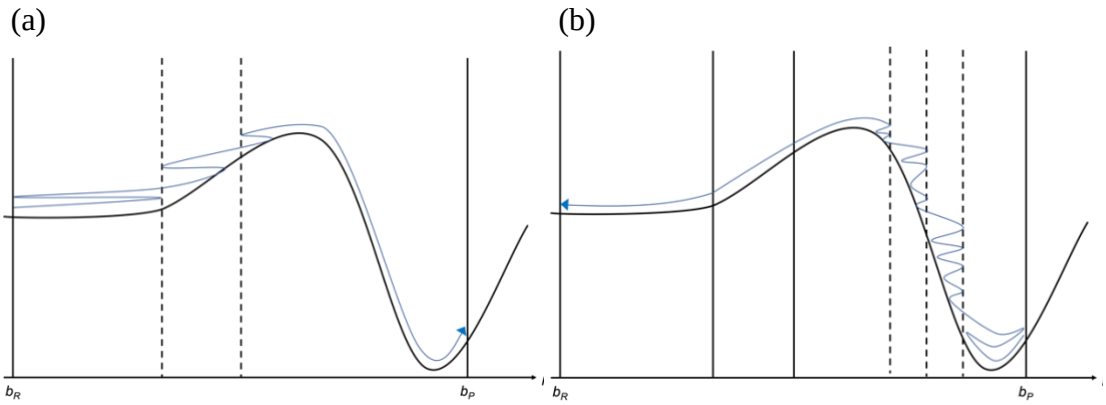


Figure 2.8: Schematics showing BXD boundary placement for a fictitious trajectory along a single dimension, some reaction coordinate ρ. The black curve shows some potential energy barrier which is a function of ρ. The trajectory (blue) progresses along ρ through the various BXD boundaries represented by vertical lines. The panels show the progress of the BXD trajectory when placing adaptive boundaries. For each panel boundaries being placed are shown by dashed lines whilst existing ones are solid. (a) Adaptive boundary placing in the forwards direction. In flat regions of the PES large boxes can be used, whilst in steeper regions smaller boxes are needed to help the trajectory over the potential energy barrier so it can freely proceed to the product state. (b) Adaptive BXD in the reverse direction. Once the product state is reached the direction of the sampling is reversed, with additional boundaries placed when required to get over any potential energy barriers.

The aim of adaptive boundary placement is to achieve boundaries which are placed at optimal distances from one another. That is one which leads to boxes that are narrow enough to ensure the MD trajectory can traverse from one side to another in a relatively small number of MD steps, but wide enough for the dynamics of each box to decorrelate between boundary hits and ensure ergodicity.

The full algorithm for adaptive boundary placement is discussed later in this chapter but first, extending the BXD algorithm to multidimensional space into which the adaptive boundaries will be placed must be discussed.

2.2.3.2 Extending BXD to Multiple Dimensions

The main outcome from the O'Connor *et. al.*[53] paper was to generalise the BXD method to multiple dimensions. To extend BXD to multidimensional collective variable space for a system of *N* atoms, we must first define the Cartesian coordinates and velocities of the atoms as the vectors $\vec{r}(t) \in \mathbb{R}^{3N}$ and $\vec{v}(t) \in \mathbb{R}^{3N}$. A CV at time *t*, $\vec{s}(t)$, is a function of $\vec{r}(t)$ and can be used to describe a system in *M* dimensions as $\vec{s}(t) = [s_1(t),\ s_2(t),\ \ldots, s_M(t)]$.

Generally, for an *M* dimensional CV space the BXD boundaries are of dimension *M*-1. In its most primitive form, as discussed in section 2.2.1, BXD partitions a one-dimensional CV space, referred to as a reaction coordinate, into evenly spaced points of dimension 0. From here on in the term reaction coordinate with the symbol ρ, will be used when discussing a one-dimensional CV and $\vec{s}$ for a CV of two or more dimensions.

For simulations in multidimensional CV space, BXD boundaries are defined as planes in Hessian normal form.[53] That is, for the CV $\vec{s}(t)$ a BXD boundary can be written as:

$$b_j = \left(\sum_{i=1}^{M} n_i s_i\right) + D_j = 0 \tag{2.14}$$

Where $\vec{n} = [n_1,\ n_2,\ \ldots, n_M]$ is a unit norm and D_j is a constant which describes the distance from the origin.

For a boundary described by equation (2.14), the following function gives a measure of how far the system is from that boundary at time *t:*

$$\phi\left(\vec{r}(t)\right) = \vec{s}(t) \cdot \vec{n}_j + D_j \tag{2.15}$$

Changes to the sign in function (2.15), indicate that the boundary b_j has been crossed. Thus, for time steps at which b_j is crossed a single constraint can be enforced on the dynamics such that the trajectory remains on a given side of the boundary

$$\phi(\vec{r}(t)) \geq 0 \tag{2.16}$$

When generating box-to-box rate constants in a BXD run if, at time t $\phi(\vec{r}(t)) \geq 0$ but at the next time step $\phi(\vec{r}(t+\Delta t)) < 0$ b_j has been crossed and the constraint requires enforcing. To do this, the BXD procedure reverts the coordinates back to $\vec{r}(t)$ and inverts the corresponding velocities $\vec{v}(t)$ to generate new ones, $\overrightarrow{v'}(t)$, which when propagated result in a trajectory that satisfies the constraint. Using the chain rule, the derivative of the constraint function with respect to time can be expressed via the projection of the velocities onto the gradient of $\phi(\vec{r}(t))$

$$\frac{d\ \phi(\vec{r}(t))}{dt} = \frac{d\phi(\vec{r}(t))}{d\vec{r}} \cdot \frac{d\ \vec{r}}{dt} = \nabla\phi \cdot \vec{v}(t) \tag{2.17}$$

where $\nabla\phi$ is a row vector of 3 columns representing the x, y and z coordinates of the system.

Only by meeting the equality in equation (2.16) will the constraint remain satisfied at time $t + \Delta t$ whilst ensuring the velocities normal to the boundary b_j have been reflected in a truly elastic manner. Ensuring the constraint will be met at time $t + \Delta t$ can only be achieved if the inverted velocities satisfy the following:

$$\nabla\phi.\vec{v}'(t) + \nabla\phi.\vec{v}(t) = 0 \tag{2.18}$$

The equation of motion for dynamics under a single constraint is [54]:

$$\mathbf{M}\vec{a} = \vec{f} + \vec{g} \tag{2.19}$$

Where $\mathbf{M} \in \mathbb{R}^{3Nx3N}$ is a matrix whose diagonal contains the atomic masses, $\vec{a} \in \mathbb{R}^{3N}$ is the vector of accelerations, $\vec{f}$ is the force vector from the MD simulation and $\vec{g}$ represents the forces due to the constraint, given by:

$$\vec{g} = -\lambda \nabla \phi^T \tag{2.20}$$

where ϕ^T is the transpose of ϕ and λ is a time-dependent Lagrangian multiplier. The constraint is applied to the newly inverted velocities rather than the accelerations, so they become:

$$\vec{v}'(t) = \vec{v}(t) + \lambda \mathbf{M}^{-1} \nabla \phi^T \tag{2.21}$$

For the inverted velocities to be returned from equation (2.21) computation of λ and $\nabla \phi^T$ is required. By subbing equation (2.21) into equation (2.18) and rearranging, the Lagrangian multiplier is found as:

$$\begin{gathered} \nabla \phi \, \vec{v}(t) + \nabla \phi + \lambda \mathbf{M}^{-1} \nabla \phi^T + \nabla \phi . \vec{v}(t) = 0 \\ 2 \nabla \phi \, \vec{v}(t) + \nabla \phi + \lambda \mathbf{M}^{-1} \nabla \phi^T = 0 \\ \mathrm{g} \nabla \phi + \lambda \mathbf{M}^{-1} \nabla \phi^T = -2 \nabla \phi \, \vec{v}(t) \\ \lambda = \frac{-2 \nabla \phi \cdot \vec{v}(t)}{\nabla \phi \mathbf{M}^{-1} \nabla \phi^T} \end{gathered} \tag{2.22}$$

Provided the CV of a system is readily differentiable with respect to Cartesian coordinates, $\nabla \phi^T$ can be evaluated whenever a BXD inversion is required. Details of how to do this can be found in Appendix 2. With $\nabla \phi^T$ in hand, present is everything needed to reflect the velocities according to equation (2.21) and the new velocities can be returned. These new velocities have components normal to the boundary that have been inverted and the constraint is once again satisfied.

2.2.3.3 Boundary Placing in Multidimensional Space

Now that BXD has been generalised to multidimensional space an adaptive scheme for boundary placement within this space can be introduced. Adaptive boundaries are placed according to the sampling dynamics within a given number of MD steps, n_{samp}.

A region of CV space bound only by a single BXD boundary b_j shall be considered. After n_{samp} MD steps have been run the set of the sampled n_{samp} values of $\vec{s}$, $\boldsymbol{S} \in \mathbb{R}^{M\times n}$ can be obtained, where M is the number of dimensions in the system. From this, the set $\vec{R} \in \mathbb{R}^n$ consisting of the distance (r) of each element of $\boldsymbol{S}$ from the lower boundary b_j can be defined. A normalised histogram of $\vec{R}$ gives the cumulative probability distribution function, $P(r)$, representing the likelihood of a given trajectory frame being a certain distance from the lower boundary, which can be used to calculate the optimal region of CV space for placing a new BXD boundary. Then, the distance r_{max} from b_j is calculated, where r_{max} is the centre of bin h_{max} defined by $P(r_{max}) \geq 1-\epsilon$, with ϵ typically taking values of between 0.01 and 0.1. It is at this position that the new "upper" boundary is placed. By defining points $\vec{s}_{min}$corresponding to the mean value of $\vec{s}$ in the first bin of the histogram and $\vec{s}_{max}$ corresponding to the mean value of $\vec{s}$ in the bin h_{max} an approximation can be made for the dynamical path through the box. The new upper boundary is then orientated normal to this trajectory, with the unit norm:

$$\vec{n}_{new} = \frac{\vec{s}_{max} - \vec{s}_{min}}{|\vec{s}_{max} - \vec{s}_{min}|} \tag{2.23}$$

A schematic for placing an adaptive BXD boundary after n_{samp} steps based on the current trajectory through the box is shown in Figure **2.9**.

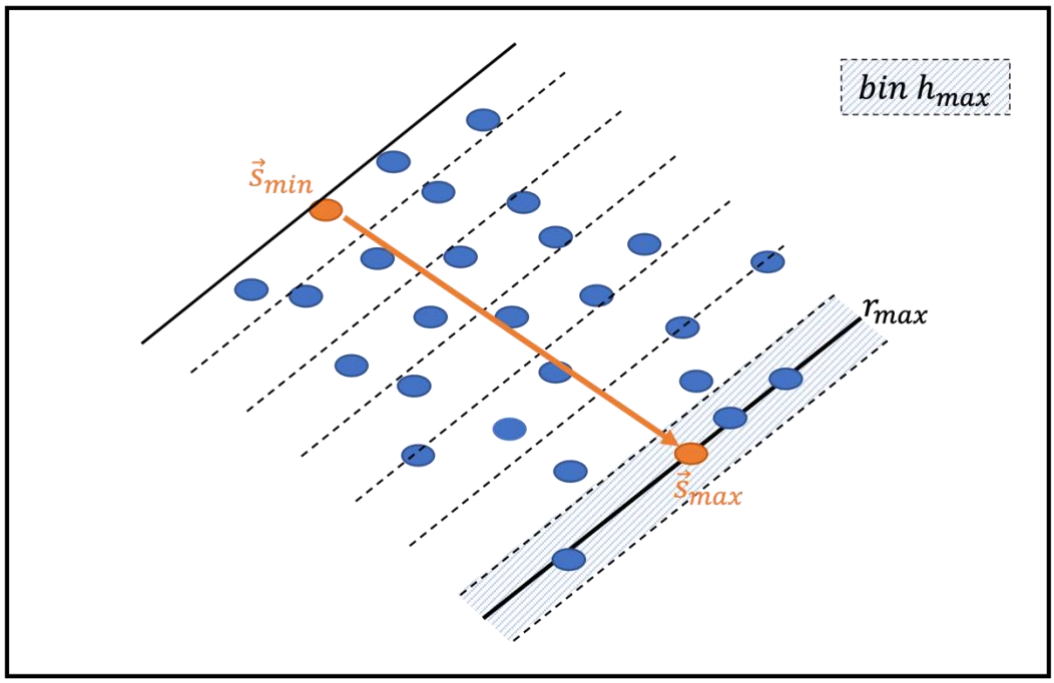


Figure 2.9: After n_{samp} MD steps each sampling a value $\vec{s}$ (blue dots) a new boundary is placed at distance r_{max} from the lower boundary, b_j. The difference between $\vec{s}_{max}$ and $\vec{s}_{min}$, the average value of $\vec{s}$ in the last and first bins gives and approximate path through the box (orange) from which the new upper boundary is orientated normal to.

2.2.3.4 Adaptive BXD runs

With the method of boundary placement dealt with, the exact procedure for conducting an BXD simulation to place these boundaries based on the PES of the system can be outlined. From here on in, such simulations will be termed ‘adaptive runs’.

To start an adaptive run a single box, $B_{0,}$ is generated. However, at this point only the lower boundary $b_{0,lower}$ of the box is defined and is located at starting geometry and orientated with a unit norm along the starting path segment. Then, an MD run is started and left to run for n_{samp} steps. During this time, whenever the lower boundary is hit the BXD inversion procedure described in section 2.2.3.2 is invoked. After n_{samp} steps the data in the box is binned according to the projected distance of each MD frame along the path and an upper boundary $b_{0,upper}$ is placed following the procedure in section 2.2.3.3. A new box $B_{1,}$ with a lower boundary $b_{1,lower}$ corresponding the upper bound of the previous $b_{0,upper}$ is defined and appended to a the list of BXD boxes. The MD run

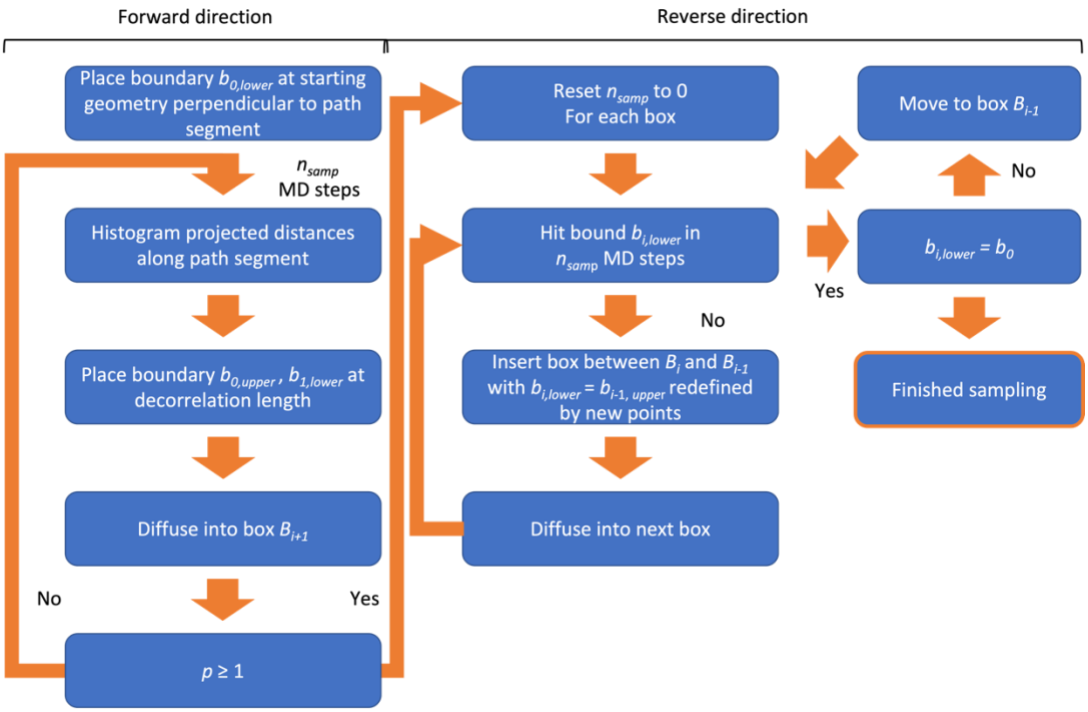
continues and if $b_{0,upper}$ is hit, the dynamics is allowed to enter the new box B_1 where adaptive sampling procedure is repeated.

A BXD box B_i is considered to be "sampling" if the number of data point in the box is less than n_{samp} and "fixed" otherwise. When conducting an adaptive run, a progress metric p is formulated such that it is equal to 0 at the reactant geometry and 1 and the target product geometry. Then, if at any given MD step $p \geq 1$, the process under investigation is considered complete in the forwards direction and the direction of the sampling is reversed as soon as the BXD box changes from "sampling" mode to "fixed".

Reversing the direction of the adaptive procedure is required to fill in extra boxes along the reaction path as required (see Figure **2.8**(b)). Upon entering reverse mode, n_{samp} is reset to zero for each box and sampling is restarted from the current box.

For an adaptive run in the reverse direction, if at any point bound $b_{i,lower}$ is hit, the dynamics is allowed to move into box B_{i-1}. It does not matter if this happens at an MD step much smaller than n_{samp}. But, if there is a BXD box in which n_{samp} data points are recorded, a new box is inserted between B_i and B_{i-1}. Here, boundary $b_{i,\,lower}$ is redefined based upon the sampled dynamics and $b_{i-1,\,upper}$ is identical to $b_{i,lower}$ whilst $b_{i-1,\,lower}$ is set to $b_{i-2,\,upper}$. The process is repeated until the boundary $b_{o,lower}$ is hit, and the adaptive run is complete. In the reverse direction, lower boundaries are always described as transparent, that is, no inversion occurs upon hitting them, whereas at upper boundaries the BXD inversion is always enforced. A flow chart and diagram outlining this procedure can be found in Figure **2.10**.

(a)



(b) (c)

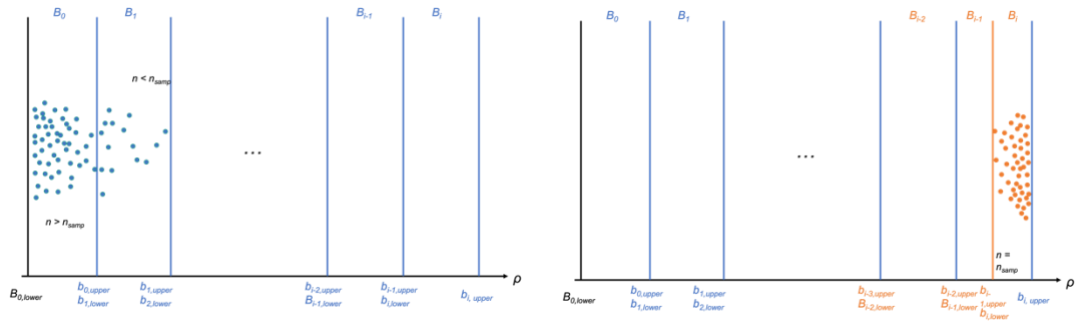


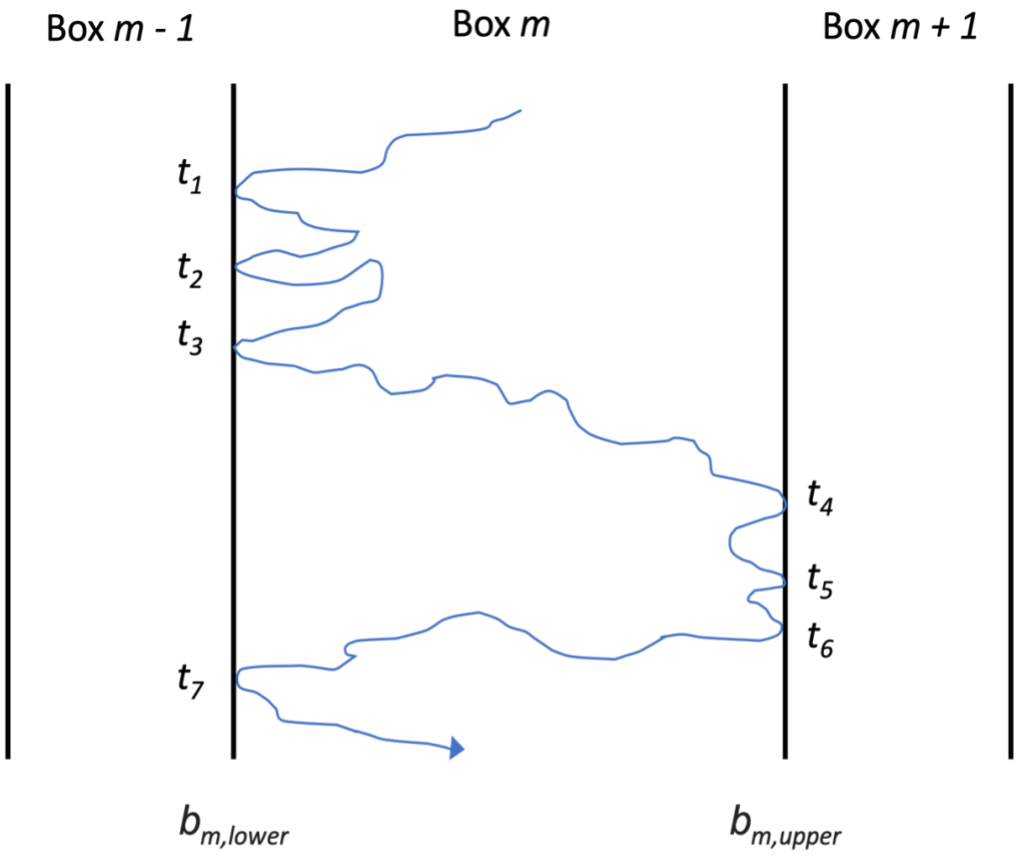
Figure 2.10: (a) A flow chart depicting the workflow for adaptive boundary placing. (b) Adaptive boundary placing in the forward direction. After n_{samp} steps, the data is binned and an upper boundary is placed at a distance r_{max} from the lower boundary, where r_{max} is the centre of the bin b_{max}=$(r_{max}) \geq 1 - \epsilon$ and is orientated noral to the approximate path through the box defined by $\vec{s}_{max} - \vec{s}_{min}$. (c) Adaptive boundary placing in the reverse direction to fill in extra boxes as required. If the trajectory hits the lower boundary of a box at any point it is allowed to diffuse through to the lower box. But if in any box n_{samp} MD steps are reached before hitting the lower boundary a new box is inserted between B_i and B_{i-1}.

2.2.3.5 Converging BXD runs

Once an adaptive run has been performed to generate set of boundaries in CV space a 'converging run' is conducted to obtain box-to-box rate coefficients for diffusion from one box to another.

During a converging run, the box data and number of boundary hits, n_{hits}, are tracked. In the forward direction the lower boundaries of each box are always "fixed" and a velocity inversion is performed each time the trajectory collides with them. The upper boundaries, however, have the ability to change from "fixed" to "transparent" after n_{hits} reaches a user defined limit, thus allowing the trajectory to diffuse into the next box. Upon reaching the final box in the forwards direction, the upper boundaries are set to "fixed" and the lower boundaries are set to become transparent after $n_{hits.}$ and the direction of travel is reversed.

All of the hits on all of the boundaries are recorded and used to generate sets of MFPTs for diffusion of the trajectory from box to box, the same way as in section 2.2.1. Additionally, Milestoning FPT's are defined as the number of MD steps between hits on alternate boundaries. For example, if the lower boundary is hit on time steps t_1,t_2 and t_3, and the upper boundary on step t_4, the BXD FPTs on the lower boundary would be equal to t_2-t_1 and t_3-t_2 whilst the Milestoning FPT on the upper boundary would be equal to t_4-t_1 regardless of the number of intervening hits on the lower boundary. The distinction between Milestoning and normal FPTs is explained in more detail in reference [48] and is highlighted in Figure **2.11**.



BXD FPTS		Milestoning FPTs	
$m, m+1$	$m, m-1$	$m, m+1$	$m, m-1$
$t_5 - t_4$	$t_2 - t_1$	$t_4 - t_1$	$t_7 - t_4$
$t_6 - t_5$	$t_3 - t_2$		
	$t_7 - t_3$		

Figure 2.11: BXD FPTs are calculated as successive hits on the same boundary, whilst Milestoning FPTs are taken from successive hits on alternate boundary of the same box.

Once the MFPTs have been calculated from the FPTs collected for each boundary, be that normal BXD or Milestoning FPT's, a free energy profile for the system can be calculated using equations (2.7) and (2.8), just as in section 2.2.1.

Chapter 3: Atomic Force Microscopy Protein Pulling

3.1 Protein Structure and Function

Proteins are large, complex biopolymers that are vital to many functions of the body such as catalysing biochemical reactions to aiding immune response, as well as enabling movement and providing structural support.[55–57]

The functions of proteins are directly linked to their three-dimensional structure, however this is determined by the order of amino acids within their primary structure.[56] The primary structure of a protein is defined as the linear sequence of amino acid residues contained within the polypeptide chain[58].

There are 20 naturally occurring amino acids, all of which share a general structure of a central carbon atom branched by a carboxyl (-COOH), an amino ($-NH_2$) and an R group, each of which has unique chemical properties.[59] The general structure of an amino acid is given in
Figure **3.1**.

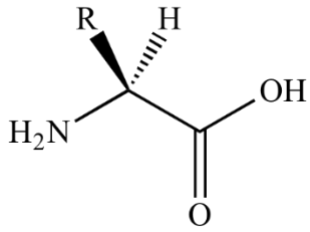


Figure 3.1: General structure the 20 naturally occurring amino acids.

The properties of the different R groups can be used to divide the amino acids into smaller groups categorised by whether they contain non-polar, polar, or ionic side chains.

To avoid water, amino acids containing non-polar hydrophobic side chains aggregate to form the water-insoluble core of proteins. Whilst those with polar and ionic side chains tend to be located on the surface of proteins, where they can interact with water such that they become soluble in aqueous solutions. [60] Additionally, ionic interactions can be found between cationic and anionic side chains, governing the way in which the protein folds itself.[61]

Following this reasoning, the sequence of amino acids in the polypeptide chain, each imposing their own conformational preference, provides a major contribution in determining the ultimate structure and function of proteins.[62]

Peptide bonds formed in condensation reactions link amino acids together one unit at a time to form polypeptide chains. A reaction scheme of a condensation reaction is given in Figure **3.2**. During this process hydrogen and oxygen atoms are lost from the chain as water is formed as a biproduct of condensation reactions, and the amino acids are now termed residues.

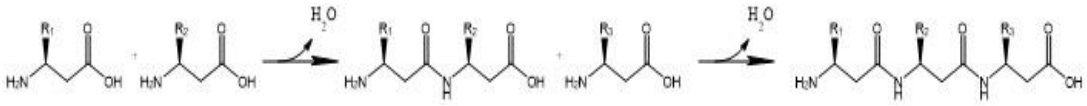


Figure 3.2: The formation of a peptide bond via a condensation reaction.

After the formation of the primary structure, the protein becomes folded into local structural conformations via backbone hydrogen bonding between carbonyl oxygens and amide hydrogens. This is termed the secondary structure and highlights the dependence of three-dimensional protein structure on the original amino acid sequence.[58]

The secondary structure of proteins takes the form of either α-helices or β-sheets. An α-helix is formed when the backbone of a protein becomes folded into a right-handed helical shape, with side chains that radiate outward. Whereas if the protein backbone is

folded into parallel strands with side chains protruding to the side, the structure is that of a β-sheet.

The folding of a protein into its secondary structure brings amino acid residues into close enough proximity for side chain interactions. As a result, the entire polypeptide chain folds to form individual protein domains. This is the tertiary structure of the protein and is stabilised by numerous interactions between amino acid side chains including hydrogen bonding, dipole-dipole interactions, hydrophobic interactions and disulphide bonds between cysteine residues. [60] It is these forces that cause the twisting and folding of α-helices and β-sheets into compact domains in an effort to minimise the energy of the structure. Consequently, both primary and secondary structures control the overall three-dimensional structure of a protein.

The highest level of protein structure is the quaternary structure, which is defined as the arrangement of multiple protein subunits into a multi-subunit complex.[58] Each domain binds to other protein subunits through chemical interactions previously discussed, creating a complex of smaller domains. Figure **3.3** shows the hierarchical structure of proteins.[63]

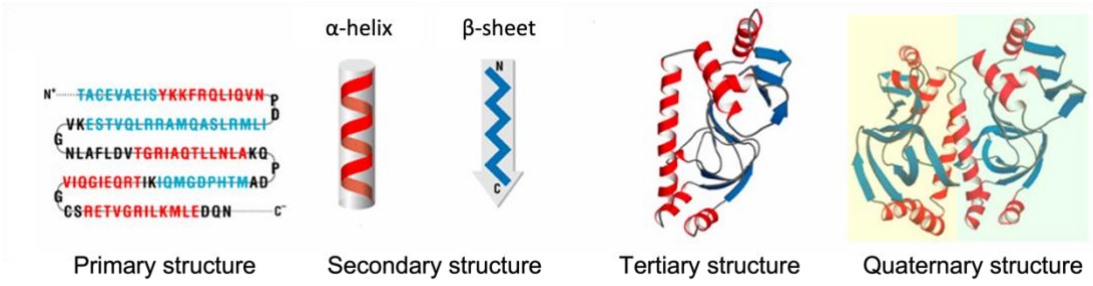


Figure 3.3: The hierarchical nature of protein structure. Picture adapted from reference [63]

The structure and dynamic behaviour of proteins regulates their functions. For example, flexibility is required for haemoglobin to undergo allosteric transition in the presence or absence of oxygen.[64] Whilst the function of the molecular spring titin relies heavily on the protein's mechanical stability. Titin is responsible for the elastic response of sacromeres within skeletal muscle, allowing it to stretch and recoil during muscle movement.[65,66] As it is continually subject to mechanical forces, it is clear flexibility and mechanical robustness are critical to its function.[67]

Understanding the dynamics of conformational changes behind important biological processes carried out by proteins is vital in the prevention and treatment of disease. Investigation into protein structure and dynamics can be carried out by experimental and computational methods, examples of which will be discussed in sections 3.2 and 3.3.

3.2 Experimental Methods

Many experimental methods have been used to monitor protein dynamics. For example, Lewandowski *et. al.*[68] employed multinuclear solid-state nuclear magnetic resonance (NMR) to measure motion of the full hydrated crystalline protein GB1 over various temperatures and time scales. In this experiment NMR observables sensitive to dynamics occurring on different time scales in different regions of the system were monitored over various temperatures, to produce a comprehensible picture of the system dynamics. [68]

Knab *et. al.*[69] monitored the dynamics of egg white lysosome from hens using terahertz time domain spectroscopy; in which the terahertz response was likely due to relaxation response from side chain rotations. Whereas x-ray diffraction was used by Rasmussen *et. al.*[70] to show at 220K ribonuclease A does not bind substrate or inhibitor but at 228K it does so rapidly. This could suggest below 220K the enzyme lacks the flexibility required for the active site atoms to be in the correct position for binding, or even that water molecules bound in the active site are too rigid at low temperatures to be displaced.[70]

There have been many more experimental methods used to examine protein dynamics such as neutron scattering[71] and dielectric spectroscopy.[72] However, the only experimental method relevant to the research in this thesis is atomic force microscopy (AFM) and so this is the only technique that shall be discussed in more depth.

3.3 Atomic Force Microscopy

The atomic force microscope has been used in physical and biological sciences for multiple purposes.[73] It can be used to image the topography of living biological cells[74] as well as to probe the mechanical properties of proteins.[75]

In 1995 the first AFM studies into the mechanical properties of proteins were conducted by Florin and Moy *et. al.* in which the interaction force between two complimentary strands of DNA was investigated.[76,77] One of the strands was attached to a solid bead whilst the other was affixed to the AFM tip. The two strands were brought together for a period of time before being pulled apart, during which the AFM tip measured the force required to separate the two strands.

They found the force required to separate the strands differed depending on the length of time the strands were initially in contact. The dependence of force on initial contact time suggested a dynamical process such as conformational change occurred within the contact time, resulting in interactions of different strength being probed. [28]

These early studies demonstrated that functional groups of large protein molecules could be attached to AFM tips and pulled apart. In turn, a new field of experimental biochemistry concerning AFM force spectroscopy emerged.

In protein pulling AFM experiments the protein is allowed to adsorb from solution onto a flat surface (usually a cleaned cover glass or a gold coated surface to which a C-terminal cysteines can bind),[78,79] before bringing a thin cantilever into contact with the protein for a few seconds so the other end of the protein can bind to it.[80] The cantilever is then pulled away from the surface, causing the protein domains to stretch and unfold.

Deflection of the cantilever from its original position is measured by focusing a laser beam on its rearside and detecting reflected light with a photodiode, which is accurate down to the nanometre scale.[67] The experimental setup of AFM protein pulling is shown in Figure **3.4**.

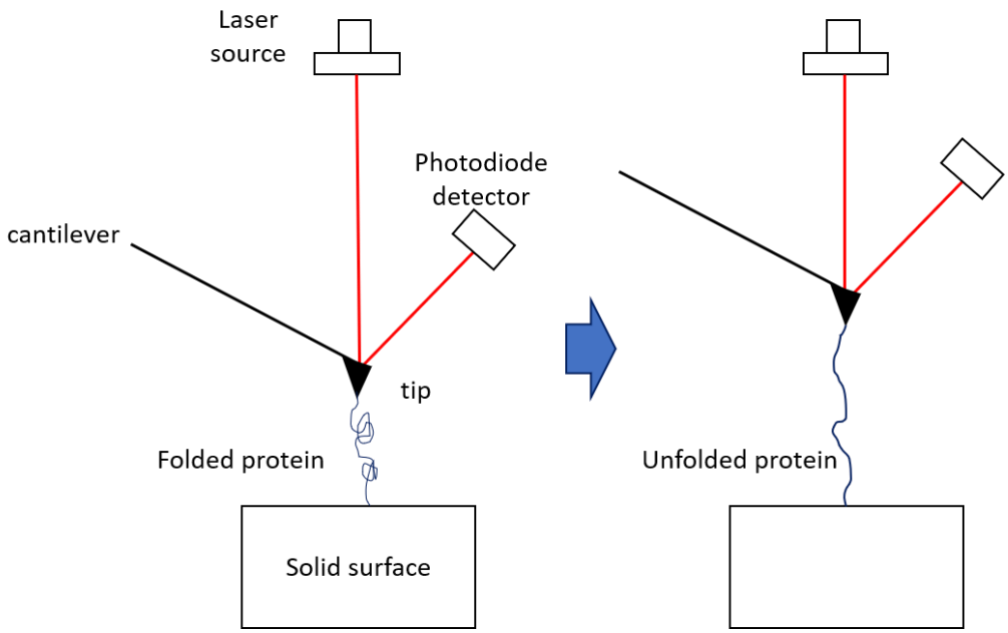


Figure 3.4: Schematic of AFM protein pulling experiment.

Two methods of AFM pulling are in common use: Force Clamp (FC) and Velocity Clamp (VC). The former refers to AFM experiments in which the tip is pulled at a constant force and the latter at a constant velocity.

3.3.1 Force Clamp Atomic Force Microscopy

FC experiments measure domain unfolding as a function of time, producing plots of extension vs time at a range of forces. [81,82] Extension vs time traces from FC experiments demonstrate a staircase pattern in which each step represents the time it takes for each module of the protein chain to unfold, measured from the time the force is applied.[78] An example of one such plot is given in Figure **3.5**

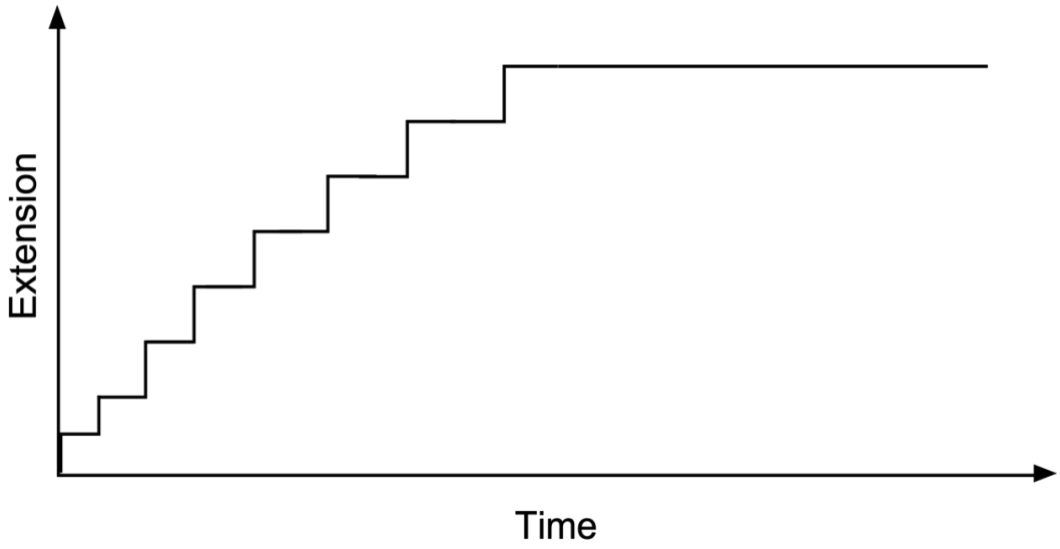


Figure 3.5: Extension vs time plots obtained in FC experiments display a characteristic staircase pattern.

FC experiments have been used to study the unfolding times of many protein domains. For example, Oberhauser *et. al.*[83] stretched engineered $I27_{12}$ under FC conditions and found its elongation occurred in the characteristic time steps. It was reasoned the staircase shape of the plot was due to waiting times for unfolding of its modules being exponentially distributed. The stepwise nature of protein rupture demonstrated by Oberhauser *et. al.*[83] was mirrored in the FC elongation of ubiquitin carried out by Fernandez and Li.[84]

However, perhaps a more insightful use of AFM when it comes to studying protein dynamics is though VC AFM experiments which can be used to measure the mechanical robustness of protein domains which can be related back to their underlying structure.

3.3.2 Velocity Clamp Atomic Force Microscopy

Velocity Clamp experiments yield force vs extension plots. When stretching a protein, the AFM cantilever behaves like a spring, so provided the force constant of the cantilever is known, Hooke's law can be used to calculate the force acting on the AFM tip from its measured displacement.

Stretching of mechanically engineered homopolyproteins (a protein consisting of repeated identical domains) produces a unique saw tooth pattern often with a piecemeal increase in peak size along the reaction coordinate, as more domains are unfolded.[78,80,85,86] Analysis of such profiles suggests the rising phase of the sawtooth reflects the elasticity of the protein and the linker molecule attaching it to the AFM cantilever as they are stretched.[80,86]

Figure **3.6** shows the typical outcome of a single domain unfolding experiment and its interpretation. The domains are connected in sequence. At point 1 element B of the concatemer is ruptured and the cantilever is relaxed. Then, B is extended up to point 2 at which point it is almost straight and stress is put on the next element E, as shown in red. The stress reaches its maximum at point 3 at which point E is ruptured. This coincides with a rapid reduction in force as the cantilever 'snaps back' giving rise to the very steep edge of the sawtooth shape between points 3 and 4. The cantilever further relaxes to reach its equilibrium at point 4 as E continues to unfold E without further resistance. Then the cycle starts again with a new unfolding element (any of A, C, D or F).

Smaller unfolding events are often seen in the sawtooth between points 1 and 3 as weaker structures within the protein are ruptured. Point 3' in Figure **3.6** is an example of one such point. Thus, the region 1-3'-2 of the tooth shown in Figure **3.6** can correspond further extension of domain B, to be followed by the extension and rupture of domain E in the region 2-3-4. Alternatively, peak 3' can correspond to an intermediate unfolding event, such as the partial break of E before its strongest bonds are broken in the main rupture event. The peak forces, i.e. the forces at the points 3 or 3' are of particular interest as they determine the mechanical properties of the protein molecule.

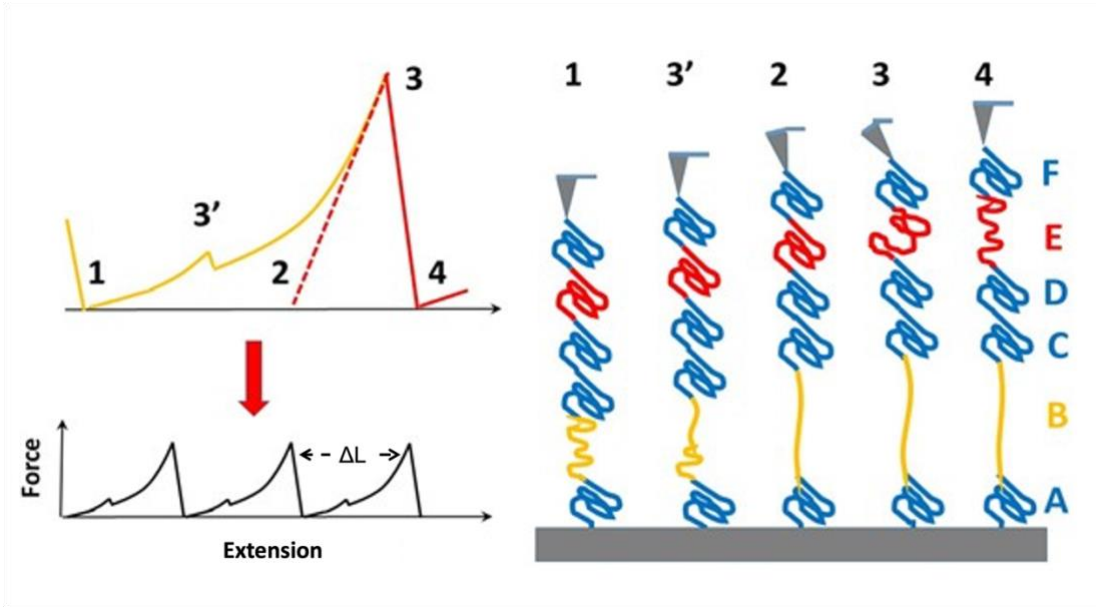


Figure 3.6: In a VC AFM pulling experiment, point 1 corresponds to the concatemer element B (gold) being ruptured but not fully extended. At this point, the cantilever is at equilibrium. Then, the element B is extended, and the cantilever deforms producing Hooke's force. At point 2, B is nearly fully extended, and the next unfolding element E (red) comes under stress. At point 3, the stress reaches its maximum and E ruptures. Then, between the points 3 and 4, the cantilever 'snaps back' and E extends rapidly. After this, the cycle repeats for one of the remaining unfolded domains. On the tooth shaped image, the unfolding events of the domains B and E are indicated by corresponding colours. The extension of a domain can reveal smaller unfolding events, one of which is indicated as 3′. The unfolding forces F_{unfld} that rupture the protein structures are the forces at points 3 and 3′.

Additional information regarding the unfolding process of protein domains can be probed from force-extension profiles. The worm-like chain model (WLC) is used for characterising the behaviour of semi-flexible polymers [87,88] and gives an estimate of the length of the domain, termed the contour length. The increment in contour length, ΔL, must therefore reflect the elongation length of a domain; predicted by subtraction of the folded domain length from the unfolded one.[86] Consistent values of ΔL reflect the similar unfolding processes of the identical domains. These incremental contour lengths can be used to fingerprint engineered homopolyproteins based on their mechanical properties. [78,89]

Li and Fernandez[85] used VC AFM to produce a saw-toothed force-extension plot for the engineered polyprotein $(I27\text{-}I1)_4$, in which there were two clear levels of unfolding

force (Figure **3.7**). Comparison of the experimental results to the mechanical unfolding fingerprint of I27 confirmed the higher-level forces corresponded to unfolding of I27 domains. In turn, the lower level forces were from the unfolding of I1 domains, suggesting I1 is less mechanically stable than I27. [85]

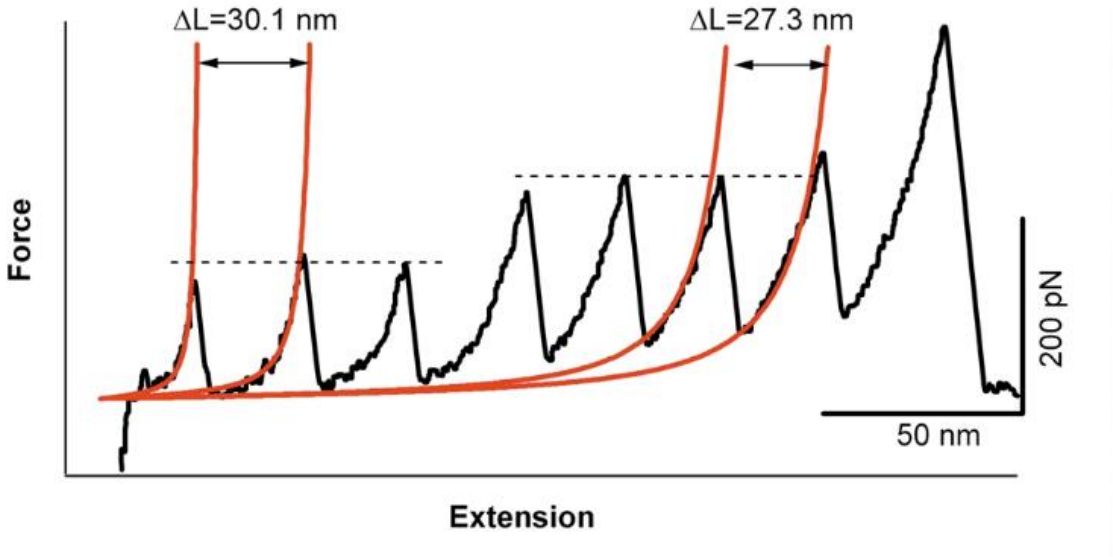


Figure 3.7: Force extension plot of $(I27\text{-}I1)_4$. Comparison of the two levels of peaks to fingerprints of I27 domains can be used to assign the higher-level peaks with ΔL = 27.3 nm as the ones corresponding to the unfolding of I27 domains. Figure adapted from reference [85]

3.3.2.1 Experimental Trends and Bell's Model of Unfolding

The original motivation behind AFM studies was to offer some explanation as to why some domains are more mechanically stable than others [90] by linking the observed unfolding force with the secondary structure of the domains. However, new and interesting phenomena have been observed in constant velocity AFM unfolding experiments which potentially can shed some light not only on the structure of the protein but also on the kinetics of the unfolding. For example, experimentation has shown that rupture force i.e., the force at point 3 or 3' in Figure **3.6** depends on pulling velocity.[79,80,91]

When pulling at intermediate velocities speed (<100 μm/s) the unfolding force often varies linearly with the log of pulling velocity.[80,92] Theoretical explanation of this phenomenon is usually based upon the predictions of Bell's phenomenological model of mechanically assisted biological processes.[93] This is a two-state model in which the folded and unfolded conformations are separated by a single transition state.

In the absence of pulling, the unfolding rate given by transition state theory would be $k_u^o = \kappa V e^{\frac{-\Delta G_{TS-F}}{k_B T}}$ where κ is the transmission coefficient, V is the vibrational frequency at the TS, $\Delta G_{TS\text{-}F}$ is the activation energy for unfolding, k_B is the Boltzmann constant and T is the temperature. Bell's insight was to explicitly account for the impact of mechanical force from the AFM cantilever on such systems. When a constant additional force is applied to pull the protein by the cantilever spring, the work of pulling force, $W = -F \cdot x$, where F is the externally applied force and x is the distance from the folded state to the TS, should be included in the free energy. This decreases the free energy, and the unfolding rate now includes an additional factor $k_u(F) = \kappa V e^{\frac{-(\Delta G_{TS-F} - W)}{k_B T}} = k_u^o e^{\frac{F \cdot x}{k_B T}}$. If $k_u(F)$ is also equal to the loading rate, $r = kv$, (i.e., the rate at which force is loaded onto the bonds in the protein), times by the distance to the transition state, then the logarithmic relationship between unfolding force and the pulling velocity seen in experiments can be given as $F_i = \frac{k_B T}{x} ln\left(\frac{rx}{k_u^o k_B T}\right) = \frac{k_B T}{x} ln\left(\frac{kvx}{k_u^o k_B T}\right)$.[80,86,91,94,95] Where F_i is the unfolding force of the protein domain, k is the cantilever force constant and v is the pulling velocity.

3.3.2.2 More advanced models of unfolding

3.3.2.2.1 Friddle and Noy's model of unfolding

A more sophisticated model by Friddle and Noy[96] accounts for the effect of the pulling more accurately, by assuming that the pulling cantilever is a harmonic spring so that the work of pulling force includes a quadratic term with respect to the extension of the protein. This model predicts that at very slow pulling velocities unfolding force is

independent of the velocity, but when moving to faster pulling velocities typically seen in experiment, a Bell-Evans logarithmic dependence is observed.[80,91,94]

3.3.2.2.2 Hummer and Szabo's microscopic model

Experimentally, differences have been seen in the force-pulling velocity relationship when using intermediate and high pulling speeds. Hummer and Szabo's microscopic model has been used to address the differences in the dynamics for high and low speed pulling and has suggested the presence of three distinct dynamical regions. [79,97] In the first region, when approaching the limit of slow pulling velocities it is suggested the cantilever works to hold back the molecular coordinate, resulting in slower rupture and a negative average unfolding force.[97,98]

Secondly, at intermediate velocities like the ones seen in most AFM experiments (v = 10^{-1} - 10 μm/s) there is a contribution from both the pulling and stochastic motion towards the unfolding force, which averages out to produce an approximately linear relationship with the logarithm of pulling velocity, similar to that predicted by Bell's model. [80,93,97,99] . In their study [97], Hummer and Szabo found a comparison of the exact results from Brownian Dynamics simulations [100,101] of a model system with their microscopic model showed an approximate linear dependency on the logarithm of pulling velocity in this region.

The models' third region occurs at extremely high pulling speeds (> 100 μm/s) at which point stochastic motion becomes irrelevant and the dynamics becomes deterministic due to insufficient time for energy landscape to be explored properly. At this point, TST breaks down as a steady influx into the TS cannot be maintained and Bell's model becomes invalid.

In a recent experiment, Rico *et. al.*[79] used specially designed equipment to pull the I27 domain of titin at velocities extending into this region and observed an upturn in force from the usual linear force vs logarithm of pulling speed dependence. This suggests there may indeed be a more complex relationship between unfolding force and pulling

velocity than originally suggested by Bell; one which bares more similarity to these more advanced models.

3.3.3 Previous computational studies of AFM protein unfolding

In principle, comparison between experiment with MD simulations should be possible. However, AFM experiments usually take place on micro to millisecond timescale[102,103] or even longer, timescales of which are out of reach for conventional unbiased MD simulations.

Instead, most simulations of AFM experiments are conducted using a technique known as Steered Molecular Dynamics (SMD) or Constant-Velocity Molecular Dynamics (CVMD). [104,105] SMD involves attaching a virtual harmonic spring to each end of the protein and moving them apart at a constant velocity, stretching the protein in a similar way to AFM experiments.[104] From this, force vs extension plots are produced similar to those from VC AFM experiments. Alternatively, applying a constant force along the vector between the two ends of the protein during the SMD simulation can be used to reproduce FC AFM experiments.[45]

SMD has been used to investigate mechanical unfolding of several protein domains including the I27 domain of titin. This all β-sheet domain has demonstrated high mechanical robustness in AFM experiments, which SMD simulations by Lu *et. al.* suggest is a result of hydrogen bonding between terminal β-sheets.[106–108] They reported a 10 Å extension of I27 produced a relatively stable unfolding intermediate in which only hydrogen bonds between β-sheets A and B were broken, and it was not until an elongation of 25 Å that all backbone hydrogen bonds were broken and the domain unfolded with less resistance. [106] Figure **3.8** shows the β-sheet hydrogen bonds that are broken in I27 when transitioning from the native to the unfolded structure through an unfolding intermediate.

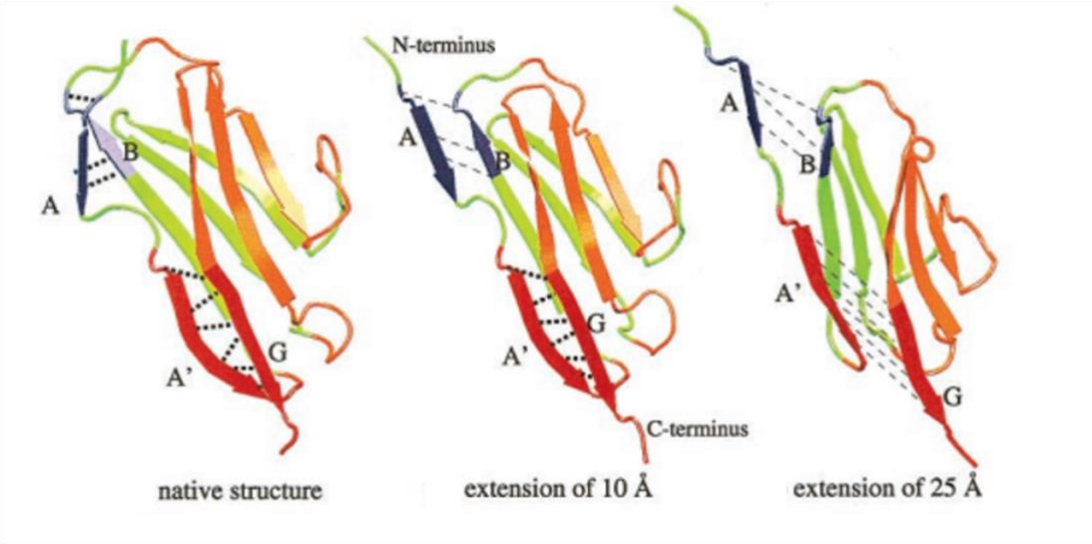


Figure 3.8: As I27 is unfolded in an SMD simulation, β-sheet hydrogen bonds between strands A-B and A'-G seen in the native structure (left) are broken. First, rupture occurs between sheets A-B to move into an intermediate unfolding state at an extension of around 10 Å. This is followed by breaking of the hydrogen bonds between strands A'-G as the domain unfolds at extensions of around 25 Å. Image adapted from reference [106]

Domains containing a mixture of α-helices and β-sheets as well as α-only domains have also been studied using SMD. Brockwell *et.al.*[104] investigated the mechanical unfolding of protein L, a mixed α-helix/β-sheet domain. Similar to I27, a cluster of backbone hydrogen bonding between β-sheets withstood the initial force, but ruptured suddenly causing complete unfolding of the domain. Whilst SMD simulations of α-only domains have shown them to be less mechanically stable and unfold with lower forces. This is thought to be because of a lack of backbone hydrogen bonding between α-helices over which the applied force can be shared.[104,109]

SMD simulations have been useful in highlighting the importance of the secondary structure of proteins to their overall mechanical stability.[102] However, the pulling speeds used in SMD simulations are much faster than in AFM pulling experiment, sometimes by as much as six orders of magnitude. Although AFM studies of I27 have tended to support the results of SMD simulations, the use of such high pulling speeds can spark debate as to the validity of SMD simulations when comparing directly to experimental data.[92,109]

BXD offers an alternative method to SMD for simulating AFM unfolding experiments, without the need to use excessively high pulling speeds. Modelling of AFM at the slowest experimental speeds can be done using BXD simulations without applying extra forces to the system and so can avoid such debate. Moreover, modifications can be made to the results of these BXD simulations to account for the dynamics of the AFM cantilever when pulling at the higher speeds seen in AFM. A discussion of both scenarios will follow in Chapter 4.

Chapter 4: Simulating Atomic Force Microscopy with Boxed Molecular Dynamics

The work in the chapter involves modifying the rate constants obtained from unbiased BXD simulations for the unfolding of I27 conducted in the same way as in refs [40,44,45]. Adjustments were made to these results so that the non-equilibrium kinetics of AFM assisted unfolding could be replicated, allowing a direct comparison of our simulated results to experimental ones taken from refs [79,92,99].

4.1 Pulling at Slow Velocities with BXD

BXD has been used in recent years as a way of circumnavigating the rare event problem inherent to MD simulations thus enabling atomistic simulations to be conducted over very long timescales. [44–46,48,52] Previous work within the Shalashilin group has seen the BXD method described in section 2.2.1 used to generate box to-box rate coefficients for the unfolding of the I27 domain of titin. [40,44,45]

These simulations rely solely on BXD boundaries preventing the trajectory from diffusing down the reaction coordinate back towards the initial state to accelerate the dynamics, and do not have any additional force applied to the system that is representative of an AFM tip. The assumption of equilibrium between boxes made by BXD means that the rate constants from these simulations can only be applied to very slow pulling speeds which do not perturb this equilibrium. Therefore, such simulations are not representative of a typical AFM pulling experiment in which the unfolding force is tested at multiple pulling velocities.

The rate coefficients can, however, be modified after the fact to account for the cantilever dynamics in an AFM experiment. But first, they must be obtained from conventional BXD simulations.

4.1.1 Method of obtaining rate constants

Simulations to obtain these rate constants were done in same way as in refs [40,44,45]. Briefly, this was as follows:

After the initial structure of I27 had been equilibrated and a reaction coordinate for unfolding determined as the end-to-end distance between its two termini, simulations were conducted using the BXD subroutine implemented in CHARMM. This subroutine works by receiving the atomic coordinates and velocities from the CHARMM integrator which are used to update the value of the reaction coordinate. If this value is such that a boundary has been crossed, then a velocity inversion is implemented according to Figure **2.1** and the inverted velocities are used to further propagate the dynamics. After a sufficient number of inversion events (minimum 2000 in this case) have been recorded the diffusion across the boundary is allowed and the process repeated for the next box and all the way along the reaction coordinate.

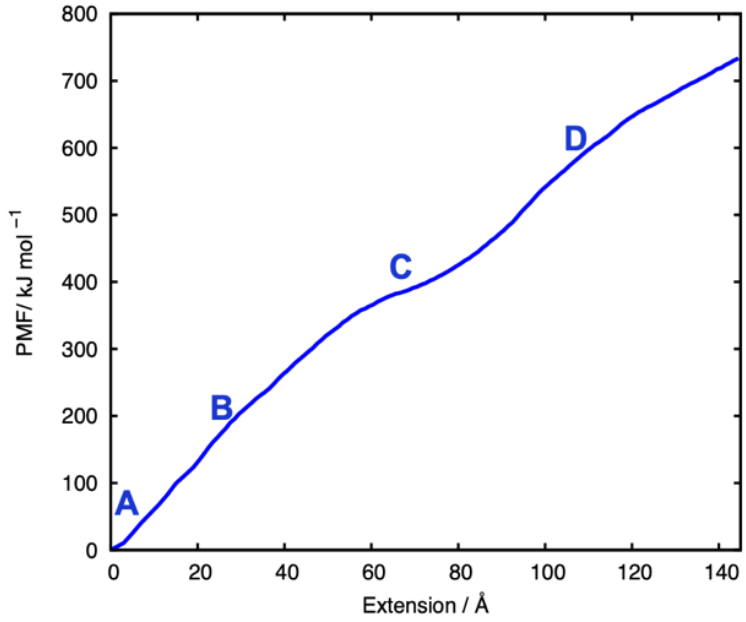
To converge box-to-box rate constants BXD usually scans boxes back and forth along the reaction coordinate several times. But once a large protein is fully extended along its end-to-end reaction coordinate it would not fold back to native state when BXD moves back towards the boxes with a smaller end-to-end distance. For that reason, the protein was stretched from its native state to full extension several times without refolding the protein. In all BXD trajectories the rate coefficients were similar, and their potential of mean force profiles (PMFs) have shown similar features.

The simulations were conducted with the EEF1 implicit solvent model[110] and CHARMM 19 force field with a Langevin thermostat set to 303 K and a friction coefficient of 50 ps^{-1} to replicate bulk water.

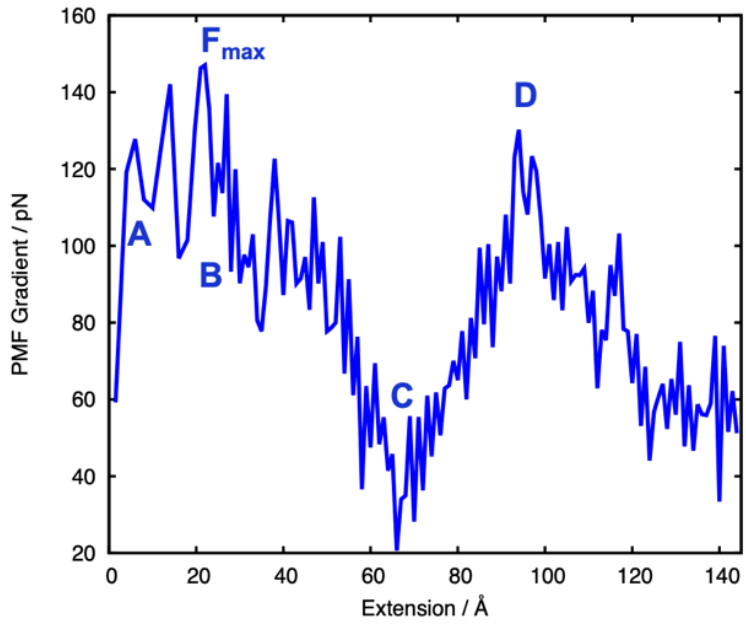
4.1.2 Results using rate constants obtained from the original BXD simulations

Rate constants were obtained using the method above, the same as in references [40,45]. Although, it should be noted this could have also been done as in Chapter 6, in which adaptive BXD sampling was used to generate box-to-box rate constants for unfolding I27. Then, using equation (2.8) a free energy profile for unfolding along the reaction coordinate was generated and subsequently differentiated to give the force as a function of extension. These are shown in Figure 4.1(a) and (b) respectively.

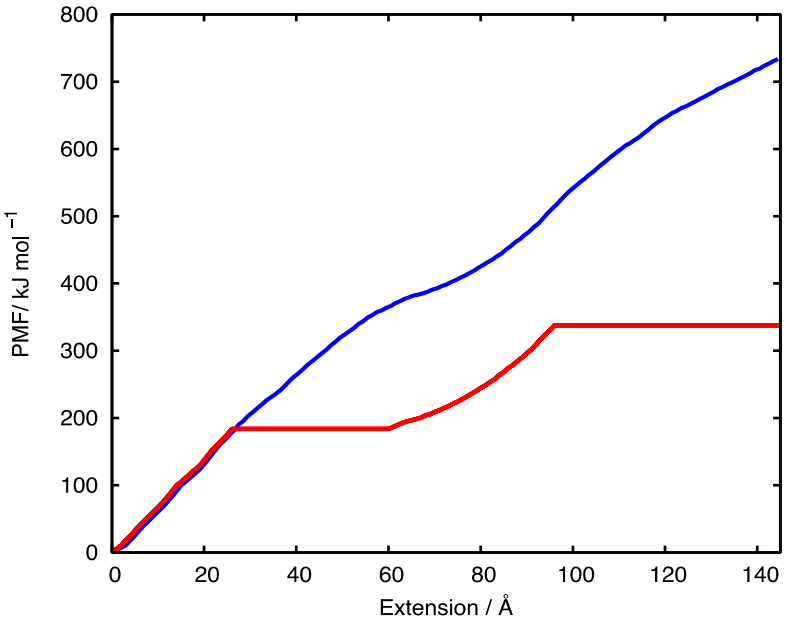
(a)



(b)



(c)



(d)

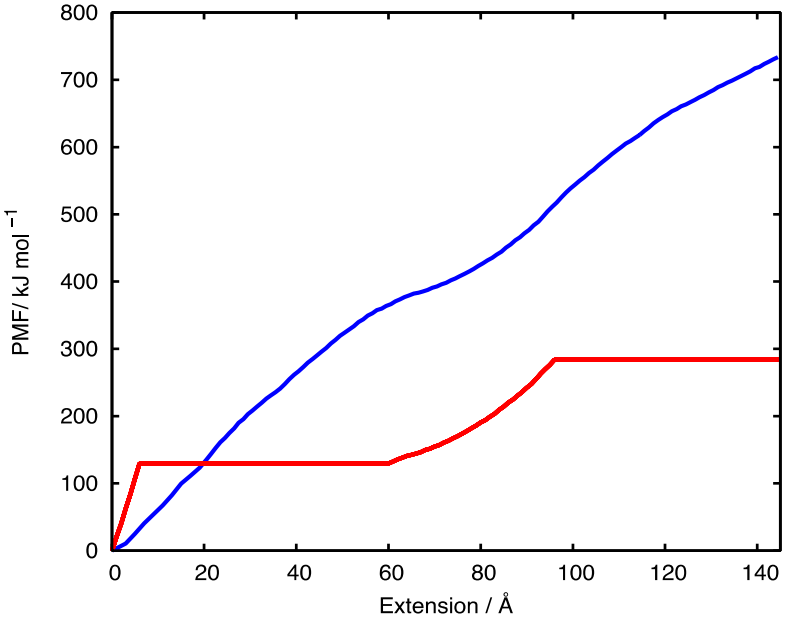


Figure 4.1: Frame (a) shows PMF of I27, i.e. its free energy as a function of extension, calculated with BXD using the EEF1 implicit solvent model and (b) its gradient representing low velocity pulling force. Point A corresponds to the native state of the protein and PMF minimum (not shown). Following this there is a steep increase in PMF to point B without any significant change in the equilibrium structure as the pulling force is spread over hydrogen bonds between I27's β-sheets. After reaching the point B the hydrogen bonds rupture almost simultaneously causing a drop in PMF gradient to point C as the protein slackens and extends. Further pulling increases the gradient up to point D as the next pair of β-sheets connected by hydrogen bonds comes under stress. The hydrogen bond link between these β-sheets is weaker. Fluctuation of the force reflects incomplete convergence of the calculation, however it still qualitatively captures the main features of the PMF. Frame (c) shows a modified PMF1 with flat regions at extensions of 25Å- 60Å and 95Å-145Å to account for

the formation of hydrogen bonds with water and frame (d) shows PMF2 with flat regions positioned at extensions of 5-60 Å and 95-145 Å as well as multiplication of the upwards rate coefficients before 5 Å by 0.0025. This modification provides the best fit to experiment.

In these figures point A corresponds to the native state of the protein, which if I27 were to be compressed as well as extended, would sit in the resulting PMF minimum. Compression of the protein is not shown as only the extension of I27 is relevant to protein pulling experiments. At first, extension of the end-to-end distance from point A causes the PMF increase rapidly whilst there is little change to the equilibrium structure of I27. This is because of hydrogen bonding between the β-sheets of I27. Initially, the force is shared between them, but once it becomes too great to withstand the hydrogen bonds between the A' and G β-sheets of I27 rupture, consistent with the findings of reference [106]. The maximal gradient, marked in frame (b) by F_{max}, corresponds to the steepest region of the PMF curve, just before the β-sheet hydrogen bonds fail. They rupture quickly allowing the I27 domain to slacken and extend in length leading to a reduction in force from the inflection point B to point C. Point D in the figures corresponds to the rupture of another set of hydrogen bonds between β-sheets formed in the middle of the amino-acid sequence. Further extension to approximately 300 Å would see the greatest increase in the PMF gradient as the protein approaches a fully extended linear conformation as was seen in the references [40,44,45]. The overall shape and analysis of the PMF in Figure 4.1(a) matches well with previous explanations of experimental[86] and computational[106] observations.

It should be noted that the unfolding force from protein pulling experiments F_{unfld} experiment is not equal to F_{max}, although there might be a correlation between the two. What follows is how F_{unfld} from experiment can be calculated with the help of BXD.

4.2 Modifications to Better Model AFM

Only in recent years has it been possible to perform AFM experiments at sufficiently high pulling speeds and conduct SMD simulations over long enough timescales for the uppermost and lowermost velocities of the respective methods to overlap.[79] However, the majority of the experimental speeds are still way out of reach of atomistic MD methods and can be described by phenomenological models only. A technique which

bridges the gap between experiment and computational studies in this area such that direct comparison of the results is possible, would clearly be beneficial in furthering the understanding of the mechanical properties of proteins. In this section the results above will be used to describe the unfolding kinetics in AFM experiments. The previously obtained rate coefficients will be used to develop a kinetic model of the pulling process in AFM and model the experimentally observed dependence of unfolding force on pulling velocity.

4.2.1 Modifications to the original PMF

For a good description of experimental AFM to be made, modifications to the above results are needed. Modifications were made to the rate coefficients to both alter the PMF in Figure **4.1**(a) as well as to account for effect the AFM cantilever has on the dynamics of the system. The latter of these is discussed in section 4.2.2 but firstly, the alteration of the PMF in Figure **4.1**(a) will be discussed.

Modified PMF1 and PMF2 are shown in frames (c) and (d) of Figure **4.1**. The implicit solvent model used to generate the rate constants above underestimates the effect of hydrogen bond formation between the newly ruptured protein β-sheets and the surrounding water molecules. Hydrogen bond formation significantly lowers the PMF of the system after the point of rupture (point B in Figure **4.1**(a) and (b)) such that areas of the PMF with small gradients become even flatter. If, for regions of the PMF at which protein-solvent hydrogen bond formation is important, the original rate coefficients are replaced by their geometric mean, the PMF becomes flat in these regions. Figure **4.1**(c) shows the modified PMF1, with flat regions introduced at extensions of 25-60 Å and 95-145 Å, both around the inflection points B and D in Figure **4.1**(a). Whilst Figure **4.1**(d) shows PMF2 with flat regions at extension of 5-60 Å and 95-145 Å which coincide with multiplication of the BXD rate coefficients before 5 Å extension by 0.0025 so that the PMF value at the flat region is similar to that of point B in Figure **4.1**(a). As will be shown in later sections, the modifications shown in frame (d) help achieve unfolding forces in better agreement with those from experiment.

4.2.2 Accounting for cantilever dynamics

BXD assumes equilibrium within each box, allowing box-to-box rate constants to be defined. However, global equilibrium is not required, and as a result BXD is capable of describing nonequilibrium kinetics with the help of the Master Equation. In the case of protein unfolding assisted by AFM, even if initially the protein was in equilibrium the motion of the cantilever distorts the initial equilibrium between boxes, and makes the population move from one box to the next.

Accounting for the effect of the cantilever on the system's dynamics can only be achieved if a term is included within the simulation which represents the interaction between the cantilever and protein. Such a term can be used to modify the rate coefficients from unbiased BXD simulations so that the cantilever dynamics become reflected in the PMF profile.

It is not unusual to assume the total PMF of a system comprising of a protein being pulled by an AFM cantilever can be expressed as the sum of the free energy of unfolding and the mechanical potential energy of cantilever extension.[96,111] Similar to many other works [97,107,108,112,113] the cantilever can be modelled as a harmonic spring with potential energy:

$$V_{harm} = \frac{kx^2}{2} = \frac{k[(\rho - \rho_0(t)])^2}{2} \tag{4.1}$$

where *k* is the cantilever spring constant and *x* is the displacement of the cantilever tip from its initial position, given by the box position along the reaction coordinate, ρ, minus the time dependent position of the tip $\rho_0(t)$.

The tip is moved with velocity *v* such that its position at time *t* is given by:

$$\rho_0(t) = \rho_0(0) + vt \tag{4.2}$$

where $r_0(0)$ is the initial position of the tip. The modified PMF becomes:

$$G_{tot}(\rho) = G_{BXD}(\rho) + V_{harm}(\rho, t) = G_{BXD}(\rho) + \frac{k(\rho - \rho_0(t))^2}{2} \qquad (4.3)$$

The time dependent potential, $V_{harm}(\rho,t)$ creates a new potential difference, $\Delta V_{harm_{m-1,m}}$, for diffusion of the population of one box into the next:

$$\Delta V_{harm_{m-1,m}} = V_{harm}(\rho_{m-1}, t) - V_{harm}(\rho_m, t) \qquad (4.4)$$

The box-to-box rate coefficients are modified such that they reflect the potential difference between boxes imposed by the cantilever tip:

$$k_{m-1,m}(t) = k_{m-1,m}^{BXD}\, e^{-\frac{\Delta V_{harm_{m-1,m}}}{2RT}} \qquad (4.5)$$

$$k_{m,m-1}(t) = k_{m,m-1}^{BXD}\, e^{\frac{\Delta V_{harm_{m-1,m}}}{2RT}}$$

Where $k_{m-1,m}^{BXD}$ and $k_{m,m-1}^{BXD}$ are the original rate coefficients from unbiased BXD simulations for diffusion from box *m*-1 to *m* and *m* to *m*-1 respectively.

By rephrasing equation (2.8) to include the modified rate constants, the change in free energy for diffusion into box *m* from box *m*-1 can now be written as:

$$\begin{aligned} \Delta G_{m-1,m} &= \Delta G_{tot_{m-1,m}} \qquad (4.6) \\ &= -RT\, \ln(K) \\ &= -RT\, \ln\left(\frac{k_{m-1,m}^{BXD}\, e^{-\frac{\Delta V_{harm_{m-1,m}}}{2RT}}}{k_{m,m-1}^{BXD}\, e^{\frac{\Delta V_{harm_{m-1,m}}}{2RT}}}\right) \\ &= \Delta G_{BXD\ m-1,m} + \Delta V_{harm_{m-1,m}} \end{aligned}$$

The modified time-dependent rate constants are calculated at time zero before any pulling takes place and used to generate a starting free energy. An initial equilibrium population in all the boxes is assumed before pulling:

$$n_m(0) = \frac{e^{-\frac{G_m^{BXD}}{RT}}}{\sum_m e^{-\frac{G_m^{BXD}}{RT}}} \tag{4.7}$$

At each time step in a simulation, Δt, the position of the cantilever tip along the reaction coordinate, $\rho_0(t)$, is moved to $\rho_0(t+\Delta t) = \rho_0(t) + v\Delta t$. Here, new box-to-box rate coefficients are generated using equation (4.5) and the KME (equations (2.10) and (2.11)) is solved (equation (2.12)) to get the corresponding box populations $n_m(t+\Delta t)$ after time Δt, calculated from the new initial conditions $n_m(\mathrm{t})$ at time t. This drags the box populations along the reaction coordinate.

Such a kinetic approach to AFM pulling has been outlined in reference [96] albeit with only two states and model kinetic parameters. But our approach uses many boxes with the kinetic rate coefficients between them calculated in atomistic BXD simulations. Therefore, although the evolution of the populations along the reaction coordinate is gathered by solving the KME, our approach remains based on fully atomistic simulations.

The box populations after each time step are used to get the average peptide extension at that time, $\langle\rho(t)\rangle$:

$$\langle\rho(t)\rangle = \sum_m n_m(t)\rho_m \tag{4.8}$$

Which can be used to estimate the experimentally observed force according to the Hooke's law:

$$F_{exp}(\rho_0(t)) = -k(<\rho(t)> - \rho_0(t)) \tag{4.9}$$

where $\rho_0(t)$ is the position of the cantilever at time *t*.

This kinetic approach to modelling AFM experiments allows the time evolution of box populations along the total free energy of the system accounting for new hydrogen bonding to be seen as they are dragged along by the cantilever tip.

Following the above method, a two-state model of unfolding appears. The blue line in Figure **4.2** represents the sum of the harmonic spring (with a cantilever force constant of k=2 pN/Å) and the flattened PMF profile. As pulling begins and the tip is moved to the right (going from frame (a) to (b)) two minima appear. Protein rupture occurs as the evolution of the populations as described by the KME leads to them transitioning from the first minimum to the second. This is shown by the gold arrow in frame (b).

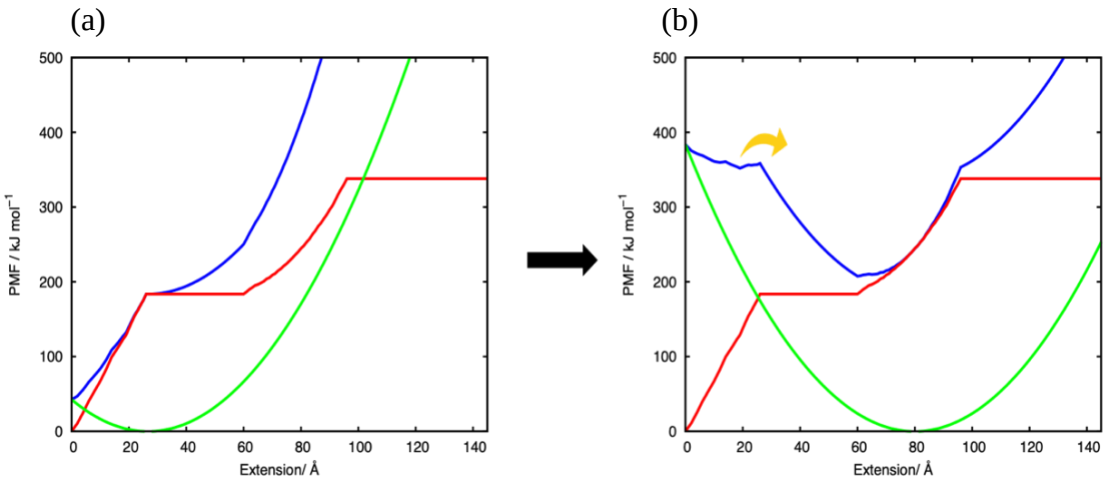


Figure 4.2: The Total PMF (blue) obtained by the addition of a harmonic spring (green) to the new flattened PMF1 profile (red) same as the red line in Figure **4.1**(c). Frames (a) and (b) are for two different positions of the cantilever, 25 Å and 80 Å respectively. Unfolding as shown by the yellow arrow at the frame (b) occurs after the tip is pulled to the right and a second minima which is lower in energy than the first appears in G_{tot}. The figure covers 145 boxes as the box size of 1 Å was used.

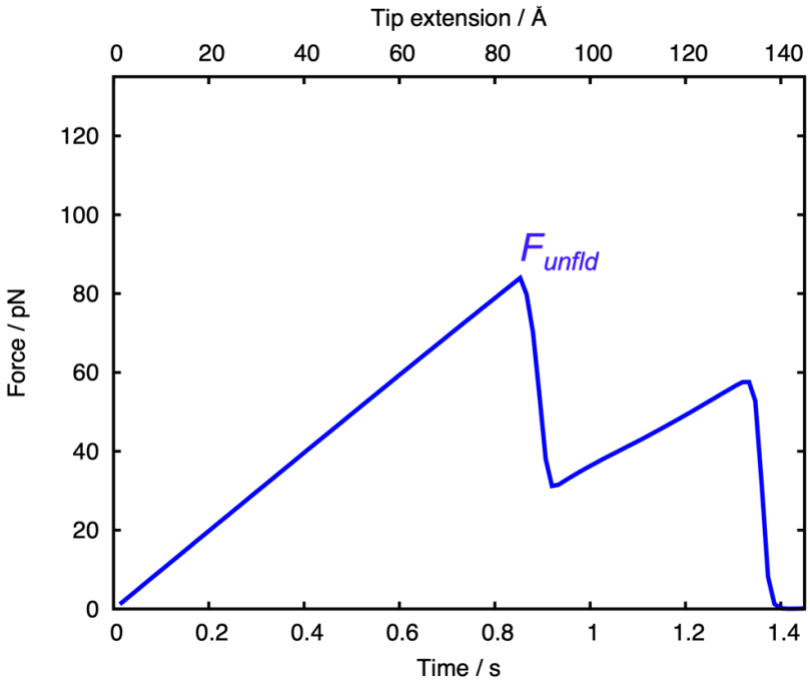
4.3 Results and Discussion

The modifications discussed in sections 4.2.1 and 4.2.2 were used to simulate AFM pulling experiments over a range pulling speeds large enough to cover both conventional AFM and high speed force spectrometry (HS-FS).[79,80,99] The following sections detail the results of these simulations, focusing first on those obtained using PMF1, before discussing the ones generated from PMF2 that better match experiment.

4.3.1 Simulations at all timescales reproduced the characteristic sawtooth shape of AFM force-extension profiles

The characteristic sawtooth shape of the force-extension profiles in AFM was produced for all simulations at each pulling speed. Two examples of these profiles are given in Figure **4.3**. The figure shows force-extension profiles for simulations using a force constant of k=2 pN/Å applied to PMF1 conducted at pulling speeds of v=0.01 and v=10,000 μm/s shown in frames (a) and (b) respectively.

(a)



(b)

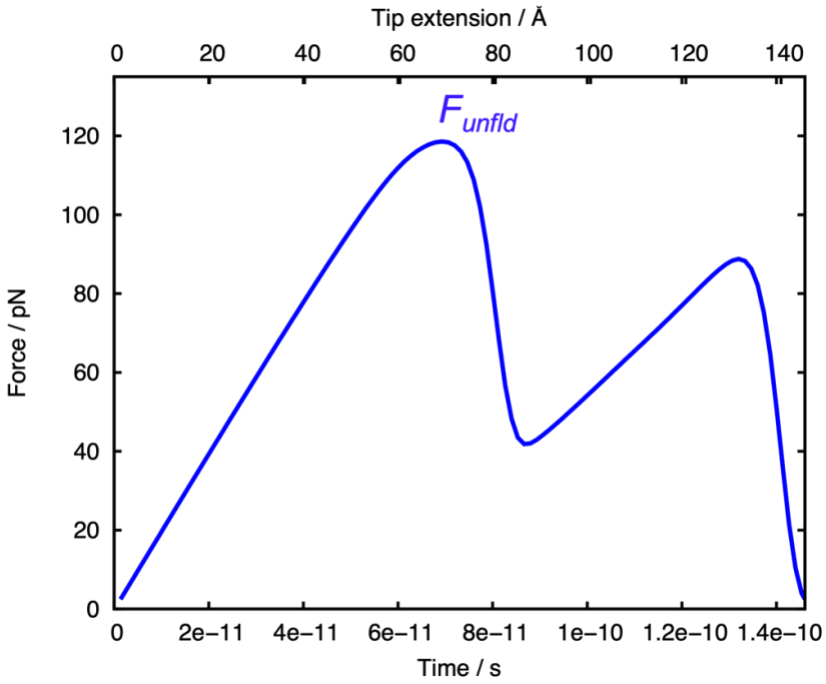


Figure 4.3: The dependence of the Hooke's force on time and cantilever position for v=0.01 (frame (a)) and v=10,000 μm/s (frame (b)) for simulations conducted using a force constant k=2 pN/Å and a flattened PMF1 in the region of 25-60 Å and 95-145 Å. Pulling at higher velocities results in greater unfolding forces occurring on shorter timescales.

Figure **4.3** highlights many important features of pulling experiments that BXD is capable of reproducing, including the shape of the tooth, as well an increase in force with speed. More importantly, it illustrates the main power of BXD. The timescale of the protein pulling process shown frame (a) is in seconds. Timescales of such lengths are usually inaccessible for atomistic MD simulations due to the rare event problem. Nevertheless, combining atomistic MD calculations of the box-to-box rate coefficients with the KME allows such timescales to be reached.

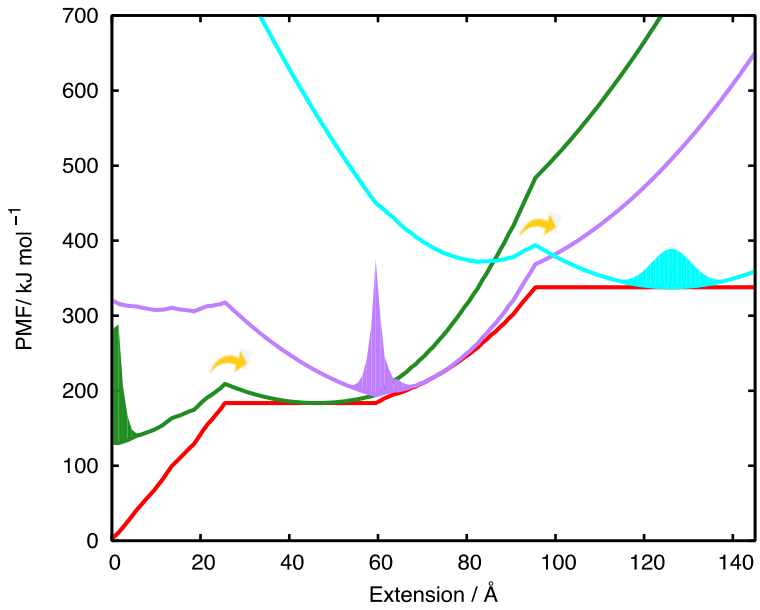
4.3.2 The unfolding kinetics changes with pulling velocity

Figure **4.4** illustrates how the kinetics of unfolding changes with pulling velocity. The green, purple and cyan lines show the total PMF (PMF1, shown in red, with the addition of the harmonic potential energy) when the cantilever is positioned at extensions of 40, 67 and 120 Å along the reaction coordinate which are reached at time

steps of 4 x10^7, 6.7x10^7 and 1.2x10^8 and 4 x10^2, 6.7x10^2 and 1.2x10^3 ns for cantilever speeds of 0.1 and 10000 μm/s respectively.

The general scheme for unfolding along this total PMF is as follows. At first, although the total PMF is substantially distorted by the cantilever the populations are still located near the original native state (green line). Then, as the cantilever is moved to the right, and the second well becomes lower in energy than the first (purple line) the populations transition into it, corresponding to the breaking of the first set of hydrogen bonds. As the cantilever continues to move to the right another well is formed and further unfolding occurs upon transition to the next well, as the next set of hydrogen bonds is ruptured (cyan).

(a)



(b)

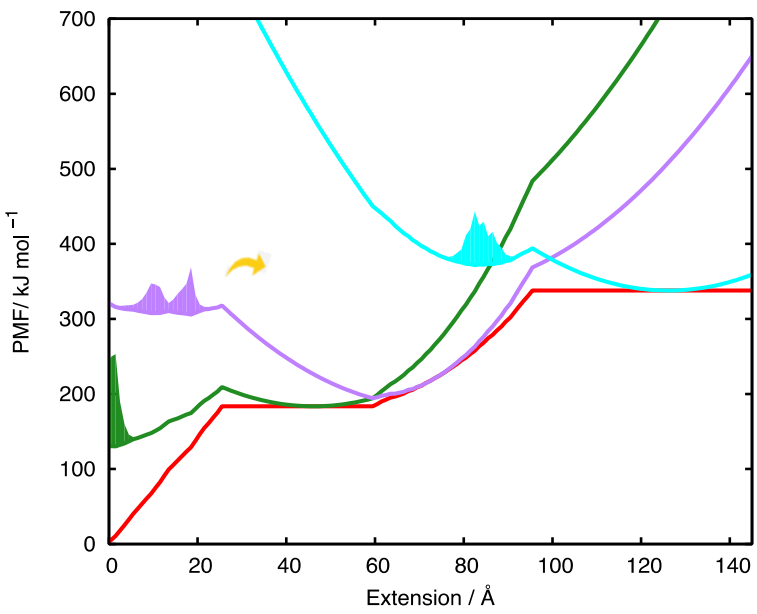


Figure 4.4: Population dynamics taken at three time points, corresponding to the cantilever extended by 40, 67 and 120 Å in an AFM pulling simulation using a force constant of k=2 pN/Å for low (fame (a), v=0.1 μm/s) and high (fame (b), v=10,000 μm/s) speeds. In the figure the leftmost well corresponds to a folded protein domain (green line) and the right wells (purple and cyan) to unfolded protein domains as the protein is stretched. At higher pulling speeds there is less time to transition into the next well and so populations remain in the well for longer. The red line is the PMF1 curve with flat regions at extensions of 25-60 Å and 95-145 Å, whilst the populations at an early, intermediate and later time step have been superimposed onto their corresponding modified PMF1+V_{harm} (equation (4.3)) curve (shown in green, purple and cyan).

However, at high pulling speeds this process is interrupted. Comparing frames (a) and (b) of Figure 4.4 shows that at high pulling speeds (v=10000 μm/s) kinetic inertia leads to populations which lag behind. This results in a in smaller average length of the protein $< r >$ and therefore a larger force calculated from equation (4.9).

At lower pulling velocities the kinetics drives the populations along the unfolding coordinate and over the barrier[95] as soon as the unfolded state becomes thermodynamically lower or equal to the native folded state. Whereas at higher speeds a combination of kinetic inertia and less time available for transition into the next available well lead to a delay in population transfer. As pulling velocity is increased the population density fails to overcome the barrier and follows the cantilever with significant delay. This provides more time for the cantilever to shift to the right during

the transition period, resulting in a larger $r_0(t)$ and larger force (equation (4.9)). Alternatively, one can think of this as faster pulling reducing the time available to escape the first well, therefore either a lower barrier or greater force is required to help population transfer. An important advantage of this technique is that although similar pictures have been suggested [96], BXD combined with the KME enables the population dynamics of a system to be visualised.

4.3.3 At the slowest pulling velocities unfolding force depends only on the cantilever stiffness before transitioning to a linear dependence on velocity as higher ones are used

Figure **4.5** shows the value of F_{unfld} over a range of pulling velocities using different force constants of the cantilever. For all force constants there is no relationship between pulling speed and unfolding when pulling at slower velocities. This could be because unfolding occurs after the next well becomes thermodynamically equal to (or lower than) the previous one and if the pulling speed is very slow the cantilever does not move much during the time interval at which population transfer takes place. Therefore, when concentrated to the lower end of the range of pulling speeds, changes in velocity do not result in large changes to the force when calculated according to equation (4.9). But with faster pulling speeds, a greater increase in $\rho_0(t)$ is seen when increasing the pulling velocity and following the reasoning above, the force increases linearly with the logarithm of the pulling speed. Additionally, there is a clear 'kink' between the flat region of the force spectrum, and the region displaying linear growth in force with the log of pulling speed. Similar behaviour has been seen in a model approach[96], but BXD yields this picture based on atomistic simulations.

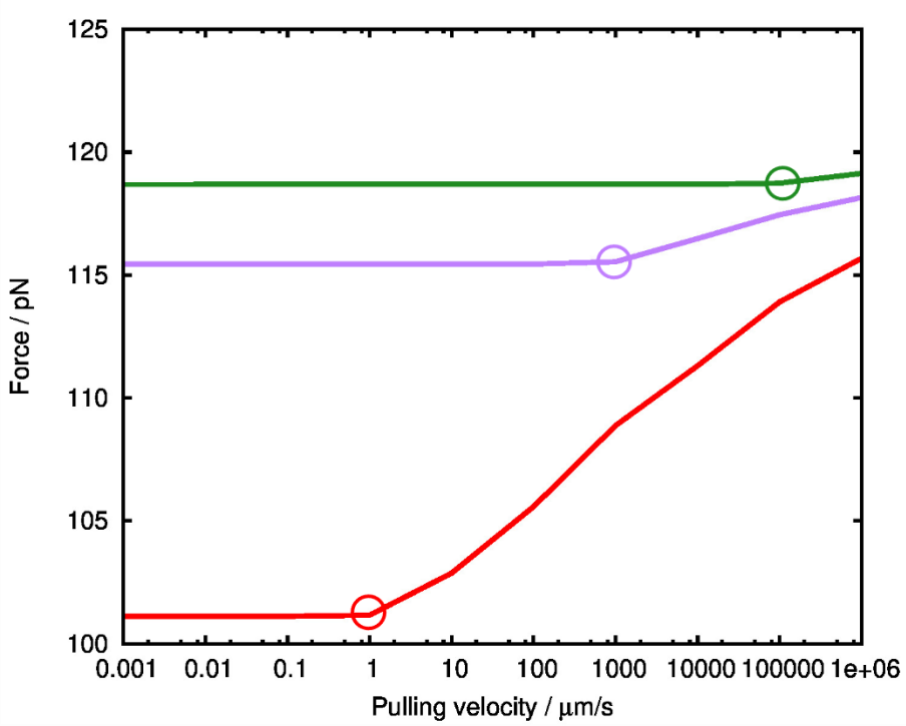


Figure 4.5: The dependence of the unfolding force on the pulling speed. At the lowest pulling speeds the force is independent of *v*. With increased pulling velocity populations have less time to escape the first well and cross the transition state to unfolding, resulting in a higher unfolding force. Increasing the cantilever force constant increases the overall unfolding force and shifts the max force – pulling velocity curve to the right. All lines on the graph are for simulations done with PMF1 shown in Figure **4.1**(c). The red line uses a cantilever with 2 pN/Å, purple with 3 pN/Å and green k= 4 pN/Å. Circles mark the velocity for each curve at which the 'kink' in force spectrum appears as the force shifts from being independent of speed, to increasing linearly with it.

BXD reveals a minimum peak unfolding force that is dependent on and increases with cantilever stiffness. This is similar to the work of Friddle and Noy [96] who developed a model of protein unfolding very similar to ours, with the exception of describing the process qualitatively rather than quantitatively through the use of a real PMF as is the case in our work. They too used the sum of a moving parabolic potential and a PES to model bond rupture in AFM, which when moved to the right produced a second minima into which population transition defined protein rupture.

According to their model [96] there is a minimum force required to rupture protein bonds in an AFM experiment which varies with the cantilever force constant. They suggest any bonds in a protein that rupture from thermal fluctuations alone are held in place

long enough by the rest of the structure to reform, and so an external force is required to destabilise all of the bonds long enough for complete rupture. This implies the existence of a minimum externally applied force required for a protein to unfold in a force spectrometry experiment. In such two-state models of unfolding, the steepness of the barrier separating the bound and unbound states is controlled by the additional potential added to the PES, which varies with cantilever stiffness. Therefore, the minimum amount of externally applied force needed to rupture a protein is dependent on the force constant of the AFM cantilever

This applies only at very slow loading rates near where global equilibrium can be assumed. At higher pulling speeds the observed unfolding force follows Bell's model in which the changing external force exponentially amplifies the unfolding rate, leading to non-equilibrium unfolding kinetics and a logarithmic force-pulling speed relationship. [93,96] At this point, kinetic parameters including the unfolding rate and the distance to the transition state begin to control the unfolding force (see section 3.3.2.1) and the cantilever stiffness becomes less relevant.

The different regions should be easily identifiable on a force spectrum which results from an AFM experiment covering a large enough range of pulling velocities and tests multiple cantilever force constants. According to this model[96] if such an experiment were done, the force spectrum would show several flat lines in the near equilibrium range corresponding to the minimum unfolding force for each cantilever force constant, merging into a single line displaying the linear increase in force associated with Bell's model at higher pulling speeds. [93,96]

Figure **4.5** shows how the results of simulations using BXD and the KME support this model, albeit with the trend shifted to the right of where Friddle and Noy's model predicts.[96] However if the same methodology is applied to PMF2 (Figure **4.1**(d)), then the different regions come much more into line with the velocities they are expected to occur at (see
Figure **4.6**(b)). The modifications used to create PMF2 and the reasoning behind them will be discussed in more detail shortly. It is valid to use PMF2 over PMF1 for this as the modifications to the rate constants to create PMF2 ensure the value of the PMF at which the flat region begins is similar in both PMF1 and 2.

Systematic investigation of the effect of the cantilever force constant on the unfolding force of proteins by experimentalists is something for which we were unable to find results, but is something which could provide future validation of this theory. Not only that, but in combination with the results of this study and others similar, it could provide further insight into the PMFs surrounding these processes.

4.3.4 The use of PMF2 allows for a better fit to experiment

A recent study by Rico *et. al.* [79] used specially designed high-speed force spectrometry (HS-FS) equipment to unfold I27 at many different pulling velocities. The range of velocities covered those found in conventional AFM experiments as well unusually high ones, reaching the lower limits of SMD smulations.[79,109] They observed a linear increase in unfolding force with pulling speed, like that of Bell's model, when pulling at conventional speeds (≤ 100 μm/s), but with faster pulling speeds found a much more rapid increase in unfolding force. The microscopic model developed by Hummer and Szabo [97] (discussed in section 3.3.2.2.2) was used to fit their data and explain the non-linear rise in rupture force with pulling velocity. Consequently, the results of the HS-FS study also suggest the observed upturn in unfolding force at the highest pulling velocities was due to insufficient time for exploration of the energy landscape. [79,97,114]

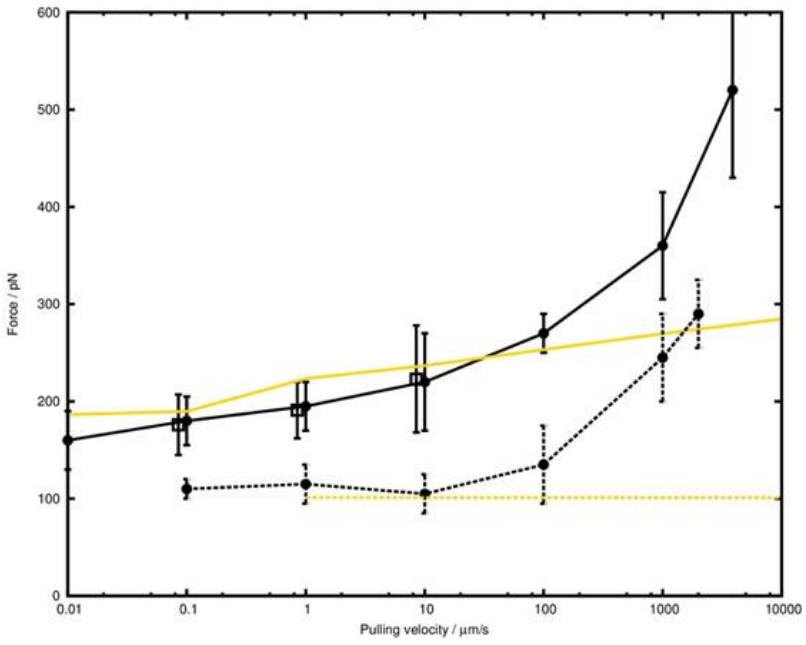
The method in section 4.2.2 can be used to simulate protein unfolding using a wide range of pulling velocities and so in theory should be capable of reproducing the results of the HS-FS experiment.[79] The results presented thus far have used the rate coefficients from previous BXD molecular dynamics simulations [40,45] with only some corrections to account for the interaction of protein with the solvent. However, further modifications were required to better fit the force-pulling speed dependence of the HS-FS experiment. Modifications were made such that the simulations were conducted on PMF2 (Figure **4.1**(d)) rather than PMF1(Figure **4.1**(c)) as above.

To match experiment several changes were made. Firstly, a new force constant of 10 pN/Å was used, the same as in the HS-FS experiment.[79] Increasing the cantilever force

constant increases the steepness of the additional potential and so without any additional modifications to the PMF profile, the maximum separating the folded and unfolded states in the G_{tot} curve disappears. Multiplying the upwards box-to-box rate coefficients before the flat region by 0.0025 steepens the initial region of the PMF and so compensates for the increase in force constant by ensuring the maximum between the two states is maintained. This is atoned for by moving the flat region to begin at 5 Å so that the flat regions on PMFs 1 and 2 lie at similar energies despite the other alterations to the rate coefficients.

Although shortening and steepening the initial rise in PMF2 ensures the beginning of the flat regions in PMF1 and PMF2 are at a similar energy, such changes cannot be made without confidence in the freedom to do so. As with all MD studies BXD simulations rely on a force field, which is based on approximations and is therefore not entirely accurate. Variations between force fields can significantly alter the outcome of the calculations. Given these uncertainties, combined with converging errors and the impact they can have on BXD rate coefficients, there is the freedom to make such changes.

(a)



(b)

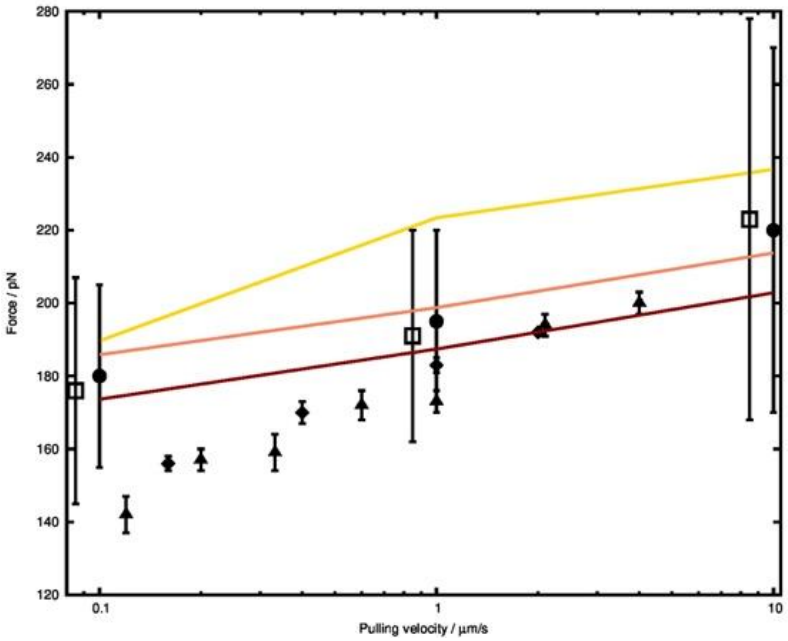


Figure 4.6: (a) Fit of BXD pulling calculations using a spring constant of 10 pN/Å and PMF2 to the experimental HS-FS data using different parameters. The black lines are taken from the dynamic force spectrums for I27 (solid line, square points from conventional AFM and circular from HS-FS) and its unfolding intermediate (black dashed line, circular points from HS-FS) in reference [79]. The gold lines are for the flattened PMF in Figure **4.1**(d) and show the overall maximum unfolding force as a function of pulling speed (solid line) and our second maxima for each pulling speed (dashed line), corresponding to the intermediate unfolding species in [79] (b) BXD calculations match experiment at conventional AFM speeds. The top gold, middle orange and bottom maroon lines are for simulations on PMF2 with k=10,4 and 2 pN/Å respectively. Experimental data taken from [79] is shown by black circles and squares as in frame (a), whilst that taken from [92] and [99] are shown by black diamonds and triangles.

Figure **4.6**(a) shows the relationship between unfolding force and pulling velocity for the overall maximum unfolding force, F_{unfld}, and the intermediate unfolding force of I27 from our simulations, as shown by the gold solid and dashed lines respectively. Examples of the peaks corresponding to F_{unfld} are shown by the first and largest peaks in Figure **4.3** whilst the smaller peaks show give rise to the intermediate unfolding force. In the HS-FS experiment [79] rupture of the hydrogen bonds between I27's A' and G β-sheets was interpreted as the main unfolding event responsible for the peak force, consistent with our findings and previous SMD studies.[106] Whilst breaking of the hydrogen bonds between the A and B β-sheets was said to be the weaker unfolding event in this study, prior to the main A'-G rupture event. However, using BXD the weaker unfolding of two other β-sheets is observed after the main event. It is not easy

to interpret smaller peaks and humps hidden within an AFM sawtooth, like peak 3' in Figure **3.6**. As suggested in reference [79], they can come from intermediate unfolding events before the rupture of the main set of hydrogen bonds, but they can also arise from events after the main unfolding event such as the "unzipping" of another weaker set of hydrogen bonds. BXD calculations suggest the latter of these options. Different studies have reported more than one secondary unfolding event as possible causes of 'humps' in AFM spectra when unfolding I27 and as such their interpretation remains ambiguous.[79,102,115]

All simulations with results shown in frame (a) used a cantilever force constant of 10 pN/Å, just as in the HS-FS experiment. [79] The calculated forces have been compared to the overall and intermediate unfolding forces of I27 from this experiment as shown by the black solid and dashed lines.[79] Agreement with experiment within error limits has been achieved at conventional pulling speeds (v=0.1-100 μm/s), but the BXD method was unable to reproduce the steep upturn in force expected in the high-speed region of the experiment. Nevertheless, these results do show BXD to be capable of modelling AFM at usual pulling velocities to within experimental limits, reproducing the predicted linear increase in force with pulling speed.

Frame (b) of Figure **4.6** shows F_{unfld} compared to experimental results taken from several references [79,92,99] for velocities typically seen in AFM experiments. The experimental data in frame (b) is taken from different studies which used various cantilever force constants of k=10 [79], 5 [99] and 4 [92] pN/Å shown by black circles, squares, triangles and diamonds respectively. Whilst the simulated results shown by the gold orange and light-yellow line use force constants of k=10, 4, and 2 pN/Å applied to PMF2, ensuring the full experimental range of cantilevers was covered. The results from BXD simulations show slightly higher unfolding forces than experimentally determined ones, but quantitative deviation between different studies is normal and can happen for a number of reasons. For example, sample preparation can vary from one study to another and may lead to variation of the observed unfolding force.

Qualitatively however, the results in Figure **4.6**(b) are consistent. Both simulation and experiment show the linear relationship between unfolding force and pulling velocity predicted by Bell's model as well as an overall increase in the force when using stiffer

cantilevers. The figure also further highlights how a systematic investigation into the dependence of the unfolding force on cantilever stiffness may be useful in expanding the breadth of knowledge surrounding protein unfolding.

Overall Figure **4.6**(a) and (b) demonstrate the ability of BXD to reproduce the linear increase in unfolding force with pulling velocity expected at intermediate pulling speeds. This is shown qualitatively for all experimental results [79,92,99] and quantitively for the HS-FS study[79] when pulling using the same cantilever force constant (k=10 pN/Å). However, similar to the predictions of Friddle and Noy[96], BXD failed to reproduce the rapid upturn in unfolding force with speed precited by the microscopic model[97] and observed in HS-FS [79] at velocities higher than 100 μm/s. This may be because at such high speeds the kinetic description of pulling from BXD fails because the MD is faster that the rate of protein-environment equilibration within each box.

Another explanation can also be considered. If a plot is made showing the difference between the observations of Rico *et. al.* [79] and the linear extrapolation of the unfolding force vs logarithm of pulling speed curve from BXD simulations for low speeds displaying a linear dependence, then the extra force is proportional to speed. This is shown in Figure **4.7**.

In the figure, the Stokes formula $F=6\pi\eta vR$ for the friction in water of a spherical object of size R moving at speed v is plotted for several values of R, each representing one dimension of the AFM tip, the length, width and depth as used in experiment as shown by the purple gold and green lines resoectively.[79] If the extra upturn in force was to be explained by viscous drag, then the extrapolated experimental results would present themselves as a linear trend between unfolding force and pulling velocity laying somewhere between the three calculations of Stokes force. This is indeed observed in Figure **4.7**. This raises suspicion that the extra force is simply the Stokes friction acting on the moving cantilever due to the viscosity of water described by its viscosity coefficient $\eta = 8.90 \times 10^{-4}$ Pa s. However, Rico *et. al.* [79] report adjusting their results to account for the additional contribution to the pulling force from viscous drag. Further experiment to investigate the effect of Stokes force in pulling experiments would be a good way to clear up any ambiguity surrounding this matter.

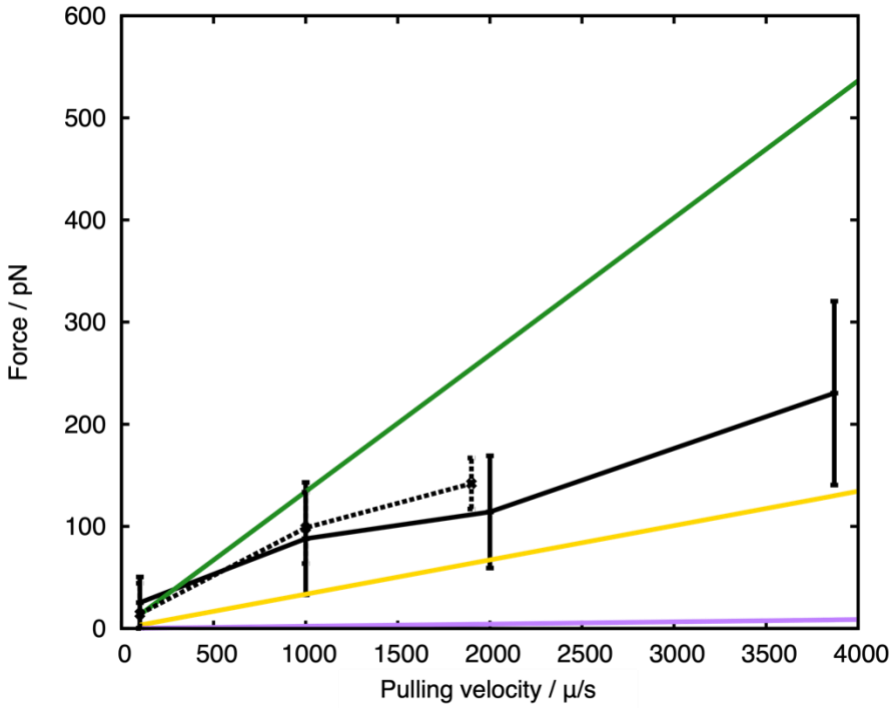


Figure 4.7: Stokes drag force for a moving object with radius R = 8, 2 and 0.13 μm, the dimensions of the cantilever used in [79] shown by the green, yellow and purple lines respectively. When this is compared to the difference between the linear extrapolation of unfolding force vs logarithm of the pulling velocity for $v \leq 10$ μm/s from BXD simulations and the results of [79] for both the total unfolding force (solid black line) and the intermediate species (dashed black line) then suspicion is cast that the extra force observed may just be a result of drag acting on the cantilever.

4.4 Conclusions

The main points to be taken away from the above discussion are:

- BXD can be used to directly simulate protein unfolding at very slow pulling velocities at which equilibrium can be assumed between the boxes, but to simulate AFM over a range of pulling velocities modifications need to be made to the original BXD rate coefficients.

- By accounting for hydrogen bond formation between the ruptured protein and the solvent underestimated by the implicit solvent model used in the original simulations, as well as the potential imposed by the AFM tip, the kinetic equilibrium between BXD boxes is distorted and the populations move from box-to-box.

- This enables the predictions of Friddle and Noy [96] to be reproduced, with the unfolding force depending only on the cantilever force constant at the slowest pulling speeds, before following a Bell-Evans linear increase at convectional pulling speeds.

- But obtaining quantitative agreement with experiment required further adjustments to the rate coefficients such that the V_{harm} was added to PMF2 rather than PMF1.

- With the extra modifications, BXD combined with the KME was able to reproduce the predicted unfolding force vs pulling speed dependence at slow and intermediate velocities as seen in the HS-FS experiment.[79]

- For problems such as protein unfolding chemical intuition would suggest that a one-dimensional reaction coordinate, the change in end-to-end distance of the protein, would be sufficient in describing the reaction progress. However, better fit with experimental data after modifications made to account for the impact of ruptured protein-solvent hydrogen bonding would suggest that a more complex CV inclusive of some water coordinate may be more appropriate.

- However, the rapid increase in unfolding force seen at the very highest pulling speeds was not reproduced with this method. This could be because at such speeds the MD is faster than the rate of equilibration between the protein and the environment in each box, or because of an underestimation of Stokes' force in experiment.

- Combining BXD with the KME allows the kinetic effects which lead to an increase in unfolding force with pulling speed to be rationalised by visualising the box populations moving along the PMF throughout the simulations.

- Using this method AFM unfolding has been modelled over a large range of pulling velocities. These include slow velocities inaccessible to other forms of

MD, with simulations reaching timescales as long as seconds, all the way through to the higher speeds usually seen in experiment.

- Therefore, BXD combined with the KME can be used to bridge the gap between atomistic simulations and protein pulling experiments and help to make a quantitative connection between experimental results and protein structure.

4.5 Future work

Future work to better understand the findings of this project would include:

- Further BXD simulations for the unfolding of I27, done in explicit water to try and understand more about the importance of hydrogen bonding between the ruptured protein and the solvent. These simulations may also be done using a more complex CV which includes a contribution from protein-water hydrogen bond formation after rupture.

- Conducting an experiment to systematically test the dependence of the unfolding force on the force constant of the AFM cantilever.

- Further investigation of the effect of Stokes' force on the observed unfolding force of protein domains through experiment.

Chapter 5: Sampling trajectories from virtual reality

The following chapter contains details of a project in which molecules are manipulated in virtual reality, for which the trajectories are recorded and fed into a BXD code to be used as a guide such that processes that challenging-to-model processes can be simulated. The project builds on the work of O'Connor [53,116] which introduced virtual reality as a tool for MD simulations as well as extending BXD so that boundaries could be placed adaptively in multidimensional CV space and utilises the ChemDyME code for BXD simulations written by Robin Shannon (available from https://github.com/RobinShannon/ChemDyME).

5.1 Introduction and Motivation

Technologies typically associated with gaming are being more frequently used in scientific research[117] due to the enhanced performance of video game processors compared to that of a personal computer's CPU. Interactive molecular dynamics (iMD)[118–121] is an emerging field in computational chemistry in which the immersive environment afforded by virtual reality (VR) technologies can be used for both the visualisation of molecules [122–124] and the study of their interaction.[125–127] Such an iMD-VR approach is implemented in the Narupa code [128], and a number of recent studies have shown this method to be a useful way of intuitively sampling both chemical reaction and conformation space. [116,129–133] The ease with which users can manipulate and guide a simulation in VR goes someway to alleviating the rare event problem inherent to MD. However, the large forces which may be imparted by the user in an iMD-VR simulation must be accounted for if one wishes to extract the free energy of the system directly, and a method for doing so is not immediately obvious.

If an efficient pipeline is created which allows a path to be sampled in VR and then used to guide the BXD process such that a free energy surface can be produced, then the above is no longer a problem and instead emerges a method for simulating processes otherwise hindered by the rare event problem. Previous studies[134–137] have utilised guess paths to define collective variables and associated methods for optimising the

guess reaction path have been developed.[138,139] However, the use of guess paths from iMD-VR as a guide for BXD boundary placement and the restrainment of BXD trajectories to be within regions of CV space close to them, could provide a new way of simulating systems otherwise challenging to the BXD method.

Through combining iMD-VR and BXD, the aim is to generate free energy surfaces for prototypical examples of three particularly challenging problems; the permeation of a nanotube membrane, changing screw sense of helicine and knot tying in the long protein chain 40 Alanine. If this new workflow is capable of producing scientifically reasonable free energy profiles for the three systems, it would be an indication of the effectiveness and robustness of the method.

What follows is a presentation of workflow for integrating VR trajectories into the BXD procedure through use of the ChemDyME code, followed by results for each of the test systems and some final conclusions. The results for each section will be split into three parts: some background information surrounding this class of molecular problem, the specific BXD implementation details and finally the BXD paths and corresponding free energy surfaces.

5.2 Simulation method

5.2.1 The iMD-VR to BXD pipeline seen in ChemDyME

All systems were studied by first creating a guess trajectory using Narupa[128] which was then followed by BXD . The BXD functionality is implemented though ChemDyME, a fully open source BXD code obtainable from: https://github.com/RobinShannon/ChemDyME. What follows in this section is a general description of the key parts involved in an adaptive, path based BXD simulation. Figure **5.1** provides a schematic of the iMD-VR to free energy surface pipeline utilised in ChemDyME.

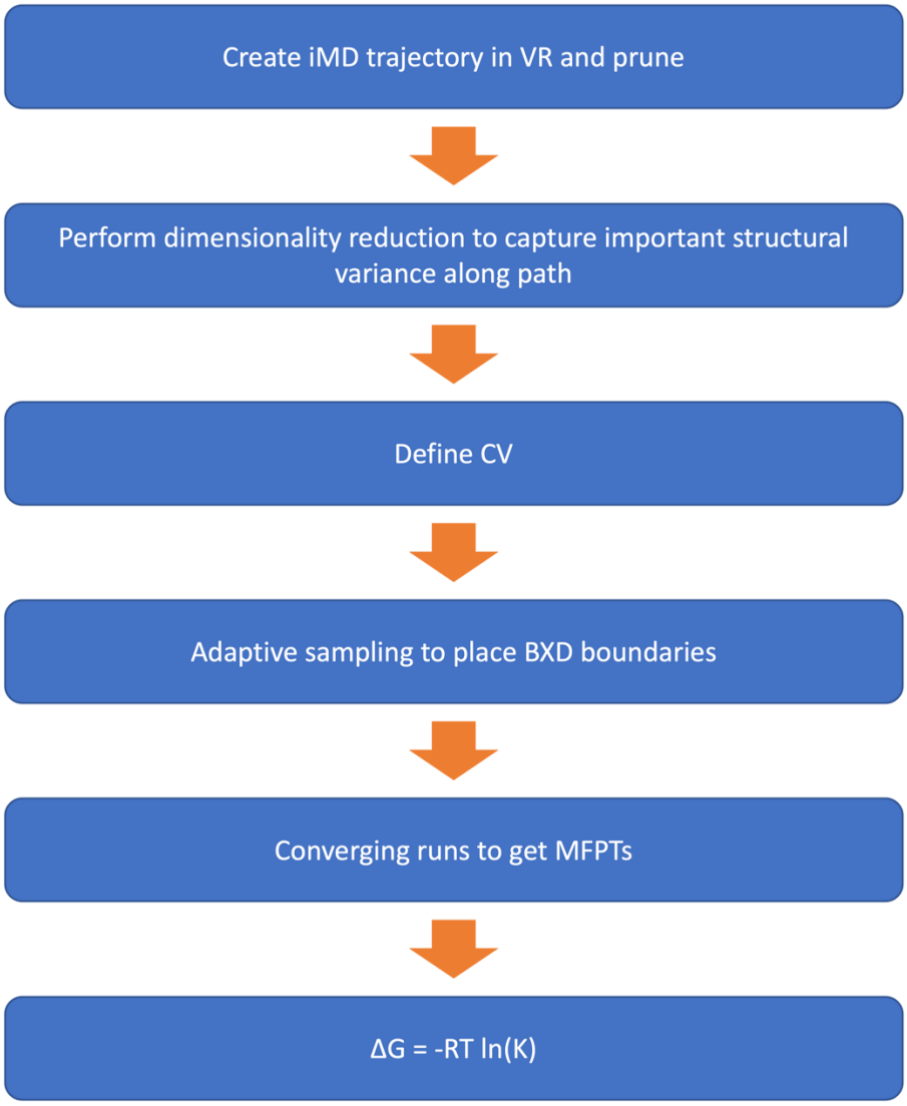


Figure 5.1: Workflow to get from an iMD-VR trajectory to free energy profile using ChemDyME.

5.2.2 The iMD-VR trajectory

The BXD pipeline in ChemDyME does not necessitate a path. In fact, the simplest way of conducting a BXD simulation only requires the specification of starting and target structures, with the path assumed to be a straight line connecting the two points in CV space. However, the focus of this project is to incorporate guess paths into ChemDyME to guide BXD such that challenging systems can be modelled.

Guess paths are easy to obtain by guiding the desired process by hand in Narupa. The Narupa framework allows for the manipulation of rigorous real-time molecular simulations.[116,128] This is shown in Figure 5.2. The figure shows how tracked participants, through the use of head-mounted displays, can be immersed in VR such that they can use wireless hand controllers as atomic 'tweezers' to manipulate molecular systems, in this case a C_{60} molecule.

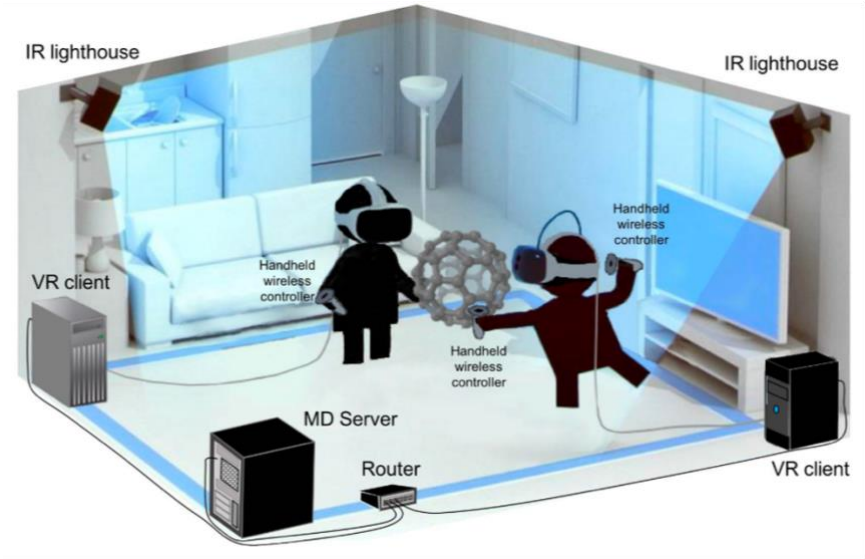


Figure 5.2: The physical set-up of creating an iMD-VR trajectory. Narupa allows participants in VR to manipulate real-time MD simulations of molecular systems and record the resulting trajectory as an xyz file which can be read into ChemDyME as a guess path for BXD. Image taken from reference [128]

The experience of the users in the real world is the same of that within the simulation. In other words, the interaction site between the 'tweezers' and the molecular system is exactly the same in 3D physical space as in 3D simulation space. [128] Consequently, users can intuitively 'lock onto' individual atoms within the system and manipulate the real-time dynamics of the system. Figure **5.3** gives three examples of molecular manipulation in VR via Narupa which shall function as the test systems for the new workflow presented in this chapter. They are: pulling methane through a nanotube (top), reversing the 'screw sense' of helicine (middle) and knot tying in 40 Alanine (bottom). Narupa can record the trajectory of such manipulations as xyz files, which can be read into ChemDyME as list of structures. These guess paths can then be pruned in ChemDyME to remove any unwanted trajectory frames and create a smoother path for BXD to follow.

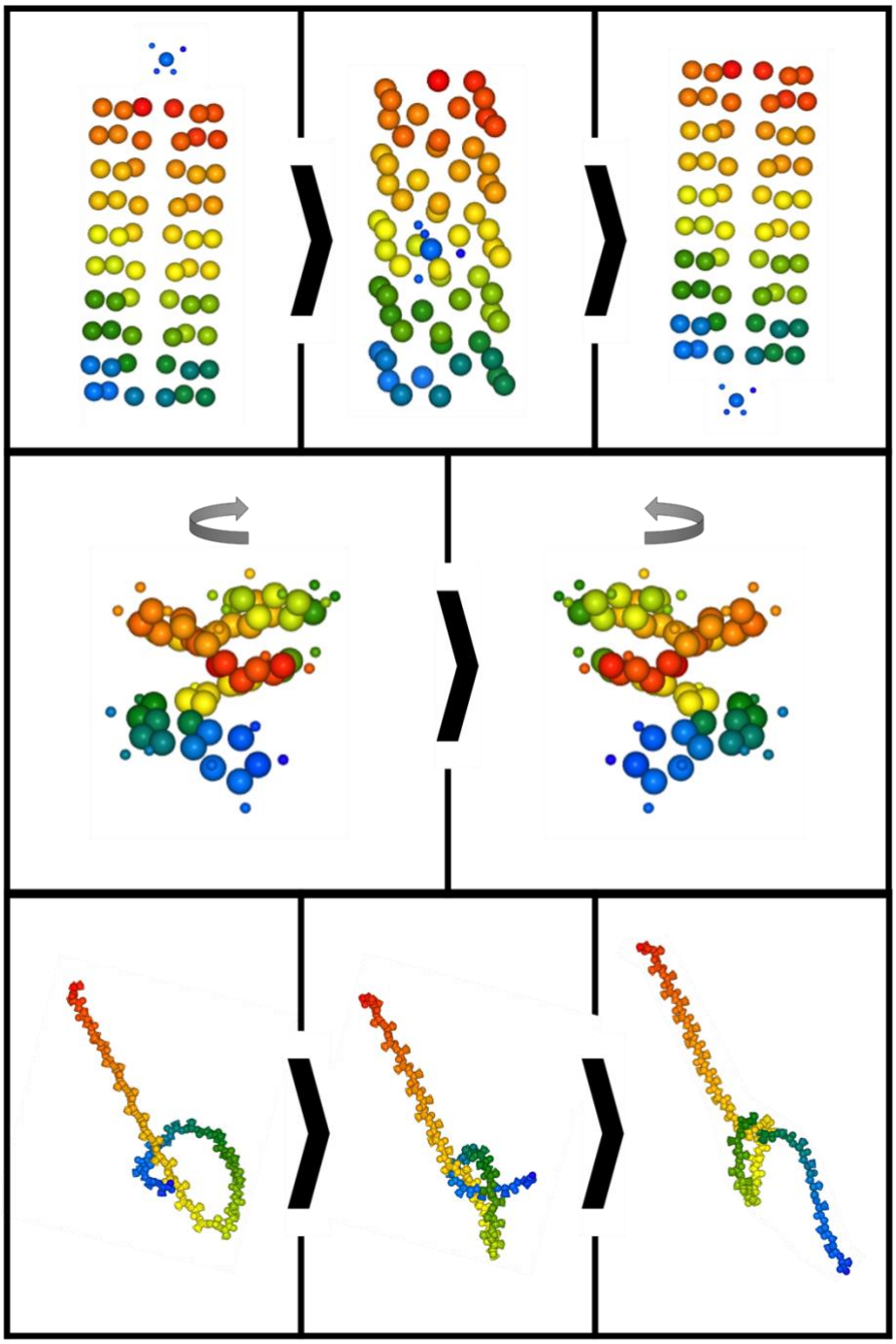


Figure 5.3: Manipulation of iMD-VR trajectories to create guess paths for BXD in ChemDyME. (top) a methane molecule is guided through a carbon nanotube (middle) the helicity of a helicine molecule is reverse and (bottom) a knot is tied in the long protein chain 40 Alanine

5.2.3 Dimensionality Reduction / Collective Variable

Once a guess path has been fed into ChemDyME, the next step in the BXD procedure is to define the CV space to work in. In theory, BXD can be performed in an any number of dimensions/collective variables. But for reasons of efficiency, it is best to use the smallest number possible whilst still describing the most important aspects of any structural change. For example, protein unfolding as discussed in Chapter 4 uses a one-dimensional reaction coordinate, whilst the structural variance of other reactions may be more convoluted requiring multiple CVs to adequately describe the system. When considering complex geometric rearrangements of molecules however, it is not always obvious how to describe the process with a manageable number of CVs, or even to determine which CVs (interatomic distances, angles etc) are most important to the process.

ChemDyME interfaces with the dimensionality reduction code pathReducer [140] to return a user-defined number of principal coordinates (PCs) which aim to capture the most important aspects of structural variance along the path. This code works to perform a principal component analysis[141] (PCA) based dimensionality reduction (DR) on the molecular trajectory data[140,142–144], resulting in a set of PCs formed as a linear combination of either interatomic distances or Cartesian coordinate basis functions.

Briefly, this is done as follows:

Consider some data of n dimensions, which in this case would be the interatomic distances, r, of each unique atom pair considered in the DR at every frame in the molecular trajectory. Each value is first standardised by subtraction of the mean of the data in that dimension, $\bar{r}$. An example of this mean adjusted is shown in Table **5.1**. Adjusting the data in this way is done so that the distribution of the altered values sit around a mean of 0, or the origin of a graph.

	Frame 1	Frame 2	Frame 3	…
r_{C_1,C_2}	r - $\bar{r}$	r - $\bar{r}$	r - $\bar{r}$	…
r_{C_1,C_3}	r - $\bar{r}$	r - $\bar{r}$	r - $\bar{r}$	…
r_{C_1,C_4}	r - $\bar{r}$	r - $\bar{r}$	r - $\bar{r}$	…
…	…	…	…	…

Table 5.1: Example mean adjusted data for conducting a principal component analysis.

From here, the covariance matrix of the data is found. This is a square matrix of the form $C_{i,j} = cov(r_i, r_j)$ where $C \in \mathbb{R}^{n \times n}$ containing $\frac{n(n-1)}{2}$ covariance values. As a simple illustration, if $n = 3$ the covariance matrix would be given by:

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix} \tag{5.1}$$

Where x, y and z represent the three dimensions of the system. Along the main diagonal of the matrix, the covariance value is calculated between one of the n dimensions and itself and is simply the variance along that dimension, whilst the non-diagonal elements give an indication of the relationship between the two variables.

Next, the normalised eigenvectors and corresponding eigenvalues of the covariance matrix are calculated. The eigenvectors are the axes along which there is most variance and thus the most information about the system and the corresponding eigenvalues are coefficients describing the amount of variance along each axis. The eigenvector with the largest eigenvalue is the one along which the data is best characterised and corresponds the first principal coordinate, PC1, this is followed by the one with the second largest eigenvalue which belongs to PC2 and so on.

Following this, the feature vector is created. This is a matrix containing as columns the eigenvalues being kept, ordered so that the first column contains the eigenvalues corresponding to PC1, the second to PC2 etc. This is the first step in reducing the

dimensions of the system, as by choosing to keep only *p* eigenvectors out of the original *n*, the final data set will be only of dimensionality *p*.

Finally, the feature vector is used to reorientate the data from the original axes so that it is expressed solely in terms of the selected axes. This is done by multiplying the transpose of the feature vector by the transpose of the mean adjusted data:

$$\mathrm{PCs} = \mathrm{FeatureVector}^T \times \mathrm{MeanAdjustedData}^T \quad (5.2)$$

Deriving the PCs in this way means they are constructed as linear combinations of the *n* dimensions considered. In turn, the system can be expressed along axes that characterise it, in this case that is in terms of the structural variance of the system. Additionally, the PCs are uncorrelated with most of the information from the initial dimensions compressed into the first PCs. By keeping only the PCs containing the most information and transforming the system data so that it is expressed in terms of those axes, systems of high dimensionality can be reduced to ones of lower dimensionality whilst still being described to a reasonable accuracy.

This method of DR is particularly appealing for the proposed pipeline. Trajectories can be taken straight from Narupa and passed into pathReducer to automatically produce a given number of PCs describing the main structural variations over the course of the trajectory. ChemDyME then prints the percentage of the structural variance captured by each PC to provide users with a measurable estimate of whether or not a sufficient number of PCs have been chosen. Performing the DR in this way streamlines the process as choosing the CVs important in describing the process is made as automated and 'blackbox' as possible. Rather than painstakingly derive bespoke CVs for a given problem by hand, the user need only make two considerations. Firstly, by applying a degree of 'chemical intuition' to which structural changes within the molecule are most likely to be important in describing the process being modelled; a decision can be made as to whether all atoms in the system should be considered in the DR or if a specific subset of atoms will suffice. Secondly, by inspecting the percentage of structural variance along the trajectory that is captured when retaining a given number of PCs

from the PCA, the number which leads to a sufficient description of the system can be determined. Furthermore, the CVs are returned in a form which is convenient for inputting into the BXD algorithm.

The key requirement for a CV to be used in in BXD is that it is readily differentiable with respect to the Cartesian coordinates, such that the constraint matrix, $\nabla\phi$, can be evaluated whenever a BXD inversion is required (see section 2.2.3.2 and Appendix 2). By default, the form of each PC produced by the dimensionality reduction is a linear combination of interatomic distances which can be easily differentiated with respected to the Cartesian coordinates of the system using the chain rule (see Appendix 2).

Having performed the DR, ChemDyME stores the CV as an object containing functions designed to transform any point in Cartesian space into the defined PC space.

5.2.4 Adaptive and Converging runs

Having defined a CV in which to work, the next steps are to conduct the adaptive and converging runs to first create and place the BXD boundaries and then collect the rate coefficients for diffusion of the trajectory from one box to another. The procedures that these runs follow are detailed in sections 2.2.3.4 and 2.2.3.5 respectively. Importantly though, both types of run require some measurement of the progression along the reaction coordinate/collective variable so that the BXD procedure can tell once the product geometry has been reached and the direction of the sampling should be reverse and halted altogether upon returning to the initial conformation.

5.2.5 Progress metric

5.2.5.1 "Path based" modifications to the BXD method

When O'Connor [53] introduced the adaptive scheme for BXD boundary placing in multidimensional CV space, the system being sampled was defined simply by the reactant and target geometries. In such a situation, the progression along the reaction coordinate for a point in CV space is defined only by the distance of the lower boundary

of the current BXD box from the starting geometry. This may be sufficient for simple biomolecular reactions, but when studying more complex reactions or those of a higher dimensionality, the system has the potential to get 'lost' and a method of monitoring its position along the CV is required.
Furthermore, it may be desirable to sample the free energy along the specific reaction pathway which is not possible using the progress metric described above. Thus, a way of projecting any point back onto the guess path is needed to monitor the evolution of trajectory along the specified reaction path.

Through the introduction of a guess path from VR, three modifications can be made to the BXD algorithm which allow a specific reaction pathway to be followed. Firstly, through the addition of extra BXD constraints running parallel to the guess path, the sampled dynamics are confined within a hypercylinder around the path whose radius is defined by the user; secondly, the orientation of the BXD boundaries are calculated from the sampling dynamics within this region and therefore controlled by the guess path; thirdly, a new procedure for projecting any point in CV space onto the guess path can be introduced as metric for determining progress along the reaction coordinate. The details outlining how this final procedure is conducted follow below.

5.2.5.2 The "path" as a progress metric

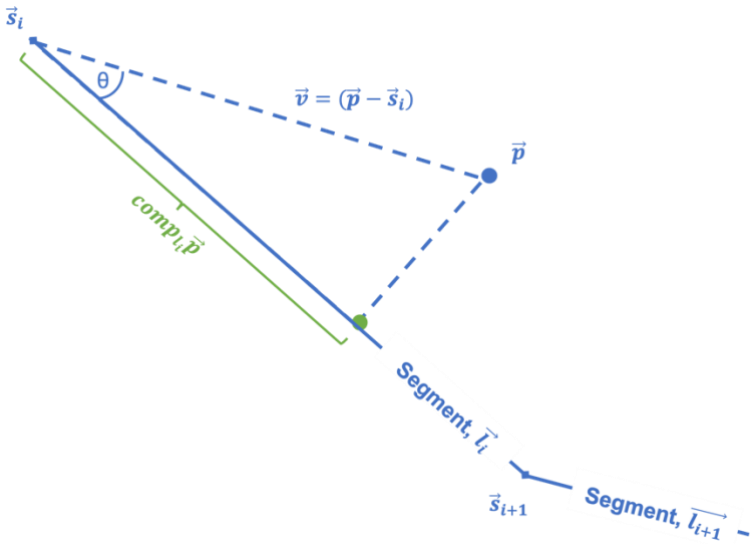
The simplest way to represent a path in an *n* dimensional CV space is through linear interpolation of each individual points along the path. Let $S = (\vec{s}_1, \dots, \vec{s}_M)$, $\vec{s}_i \in \mathbb{R}^n$ be a list of M molecular trajectory frames projected into some CV space. The linear interpolated path then consists of a set of *M*-1 linear segments $L = (\vec{l}_1, \dots, \vec{l}_{M-1})$, $\vec{l}_i \in \mathbb{R}^n$ where $\vec{l}_i = \vec{s}_{i+1} - \vec{s}_i$. It also necessary to define the cumulative distance along the path at each point $\vec{s}_i$ as $D = (d_1, \dots, d_{M-1})$, $d_i \in \mathbb{R}$ where $d_i = \sum_{i=1}^{i} \|l_i\|$.

To project an arbitrary point $\vec{p}$ in CV space, onto this linearly interpolated path the shortest distance between $\vec{p}$ and each path segment $\vec{l}_i$ needs to be determined. To do this, at each segment a vector is created between $\vec{p}$ and the starting point of the segment, ($\vec{v} = \vec{p} - \vec{s}_i$), for which the scalar projection onto the segment, $comp_{l_i}\vec{p} = \frac{\vec{v} \cdot \vec{l}_i}{|\vec{l}_i|}$, is

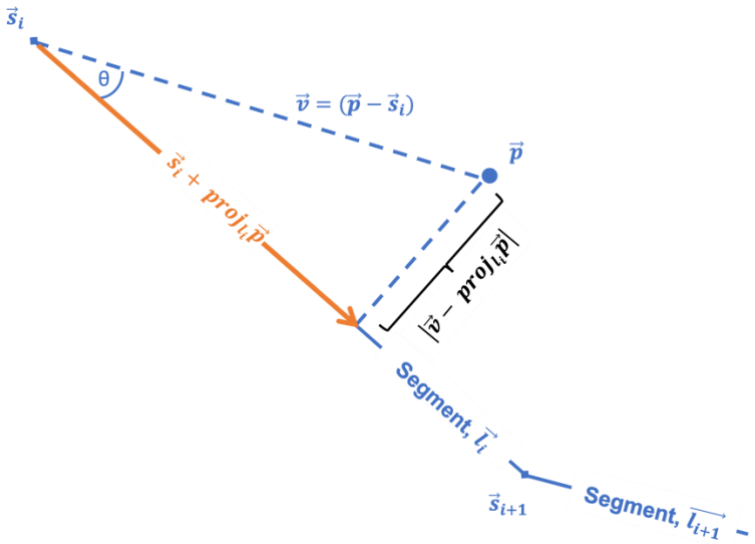
calculated as seen in Figure 5.4(a). Then, the corresponding vector projection, $proj_{l_i}\vec{p} = \vec{s}_i + comp_{l_i}\vec{p}\frac{\vec{l}_i}{|\vec{l}_i|}$, can be used to calculate the distance of the point from the path as the magnitude of $\vec{v} - proj_{l_i}\vec{p}$ (Figure 5.4 (b)).

The distance to the path is calculated for a user-specified number of segments which neighbour the current segment. Then, the cumulative distance along the path for the current MD frame is calculated as the path segment with the smallest distance to the point, d_i, added to the scalar projection of the point onto that segment. When comparing frames (b) and (c), it can be seen $\vec{p}$ lies closer to segment $\vec{l}_i$than to $\vec{l}_{i+1}$ and as such the distance of the point to the path in this case would be given as $d_i + comp_{l_i}\vec{p}$.

(a)



(b)



(c)

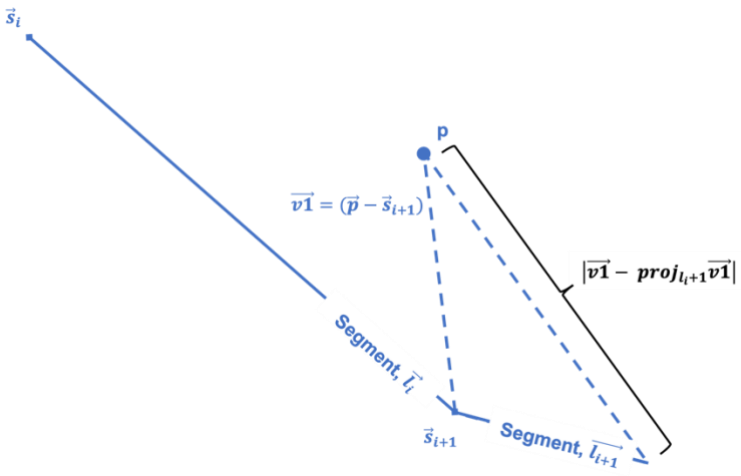


Figure 5.4: Projecting an arbitrary point $\vec{p}$ onto the linearly interpreted path. In frame (a) the scalar projection of $\vec{p}$ onto a path segment is used to obtain the corresponding vector projection. Following this, the magnitude of $\vec{v} - proj_{l_i}\vec{p}$ gives the distance of $\vec{p}$ from the segment $\vec{l}_i$ (frame (b)). This is calculated for several path segments near $\vec{p}$ and the one with the smallest difference defines the segment closest to $\vec{p}$. The cumulative distance along the path up to this segment is calculated and onto which the scalar projection of $\vec{p}$ is added to return the cumaltive distance along the path for $\vec{p}$ at a given MD frame. When comparing frames (b) and (c) it can be seen the closest path segment is that of $\vec{l}_i$, not $\vec{l}_{i+1}$.

ChemDyME stores a python object containing all the details of the progress metric which is used to convert each point in the BXD trajectory into progress along the reaction path and is subsequently used to determine whether or not sampling in a given direction has finished. Then, once both the adaptive and converging runs have been completed and the sampling terminated, the box-to-box rate coefficients can be used as in equation (2.8) to obtain a free energy profile for the process being studied. It should be noted that although beyond the scope of this project, the rate coefficients from the converging runs could also be used in combination with the KME to investigate the kinetics of the process under investigation, in the manner discussed in section 2.2. the difference here being that here, the CV of the system is created through a PCA of a molecular trajectory taken from VR.

5.3 Results

The above workflow was followed to generate reaction paths and free energy profiles for the following test systems: methane travelling through a nanotube, changing the helicity of helicine and tying a knot in 40 Alanine. The results for each system are presented in the following sections.

5.3.1 Nanotube

5.3.1.1 Background and Motivation

Ion channels are specialised proteins embedded within cell membranes whose structures enable them to selectively control the passage of ions through the plasma membrane. There exists a wide variety of ion channels, which can open and close in response to different stimuli including temperature, pH and mechanical force.[145] Excitable cells, so called because of their ability to generate tiny electrical currents which enable cell signalling and muscle constriction within the body, rely on voltage gated ion channels to selectively allow permeation of the membrane. This creates electrochemical gradients between extracellular and intracellular environments, along which ion flow produces electrical signals which are propagated along neurons and used to communicate with other cells.

Disruption to the usual functions of ion channels can cause serious disease such as Cystic Fibrosis (CF), Parkinson's and Lambert–Eaton myasthenic syndrome (LEMS).[146–148] For example, antibodies against P/Q-type voltage-gated calcium channels (VGCC) which block Ca^{2+} influx into nerve endings have been found in 85–90% of patients with LEMS.[148] This reduces the amount of the neurotransmitter ACh released from presynaptic membranes, so less can bind to postsynaptic receptors and induce muscle contraction.

Without doubt, the more is understood about the biological processes that cause such diseases, the better equipped we are for preventing and treating future cases. In fact, greater scientific insight within recent years has led to the emergence of membrane-

based nanoparticles capable of mimicking the surface features of native cells as an approach to targeted drug delivery.[149,150]

The complexity of structures comprising biological systems and the convoluted interactions between them makes achieving an exhaustive understanding of the nano-bio interface from experiment alone challenging. Therefore, the need for computational modelling as a counterpart to experimental studies becomes clear if we hope to gain further insight into the molecular conformations and interactions controlling these processes.[151,152]

Monitoring the progression of a methane molecule through a carbon nanotube acts as a primitive and archetypal model of ion transportation through a nanopore. If this new method of simulation proves to be capable of simulating such a system, one could hope to move on to systems of greater complexity where the real-life implications are more apparent.

5.3.1.2 Method

A system comprising of a carbon nanotube and a methane molecule was parameterised using MM3 forcefield parameters defined in a bespoke openMM xml file. Then, openMM was used to generate forces and energies for the system which were interfaced into NarupaIMD and ChemDyME respectively to propagate the MD. Using Narupa-iMD a methane molecule was guided through a carbon nanotube to create a guess path for BXD to follow. The resulting trajectory was passed into pathReducer to perform a dimensionality reduction, in which the hydrogen atoms were omitted. Two PCs taken from this analysis were found to be sufficient for describing the process, capturing 98% of the structural variance along the trajectory.

All BXD simulations for this system were run at 500 K to accelerate the dynamics and a friction of 0.5 was used in the Langevin integrator. This friction is much higher than that used is in other simulations (by factor of 50 for helicine and 5 for 40 Alanine). This was necessary in reducing the timescale for dynamical decorrelation given the steep potential energy gradient surrounding either end of the nanotube faces as methane

enters or leaves the nanotube. In such regions of the PES, the potential energy dominates over kinetic energy, dragging the dynamics down the slope leading to correlated collisions on the BXD boundary positioned at the point in CV space lower in potential energy. By increasing the friction to simulate more collisions with water, there is greater kinetic contribution to the overall energy of the system and the potential and kinetic energies equilibrate faster so the downwards drag of the PES is felt less.

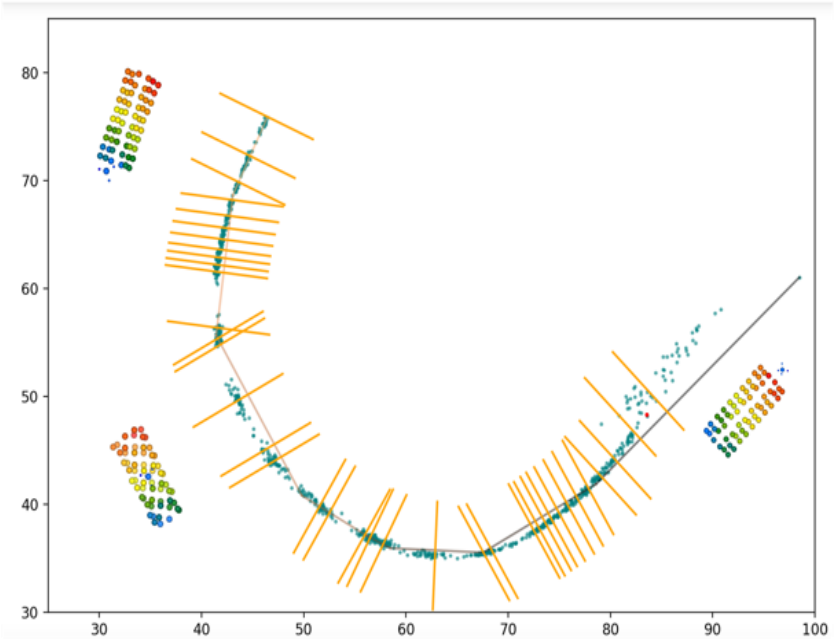
Following the procedure outlined in section 2.2.3.5 ,the adaptive BXD runs sampled each box with n_{samp} = 2500 MD steps before placing a new boundary with an epsilon value of 0.05. After some initial test simulations, it was found that the radius of the hypersphere surrounding the guess was crucial in determining whether or not the methane would pass through the nanotube. Placing path boundaries at a distance of 4 Å from the guess path was sufficient to force the methane through the nanotube, whilst if they were placed at 8 Å methane would travel along the outside of the nanotube to avoid the energetic penalty of entering the nanotube faces. BXD runs were performed for both cases so that the free energy profiles for travelling through and along the outside of the nanotube could be compared. For both cases converging runs were performed to generate milestoning rather than BXD MFPTs for the box-to-box transitions, which were then used to calculate free energy profiles for the trajectories.

5.3.1.3 Results and Discussion

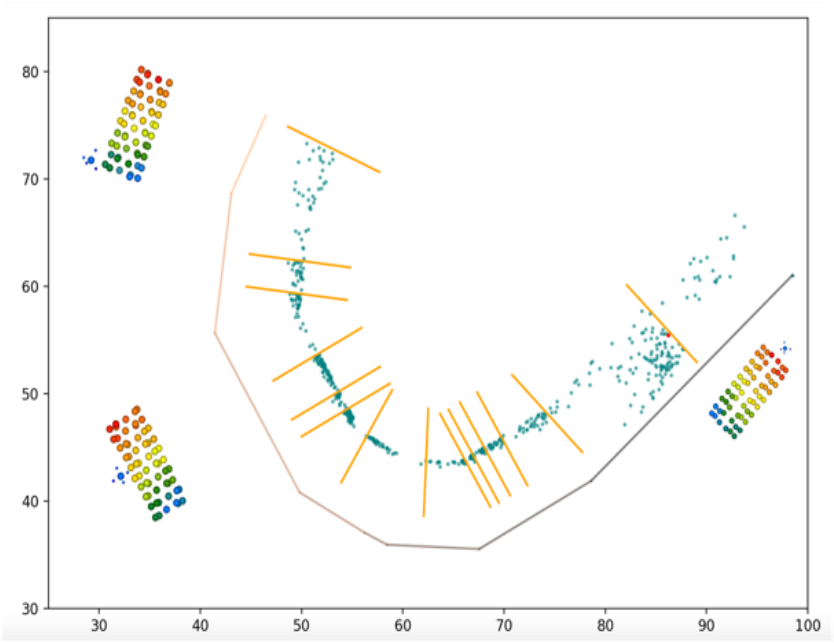
Figure **5.5**(a) and (b) show the adaptive sampling data points obtained from confining BXD to within 4 or 8 Å of the reduced VR path projected into CV space, onto which they are superimposed. Utilising VR in this way avoids the need to derive reaction coordinates by hand. Rather, the user simply pulls the methane through the nanotube and pathReducer returns a set of PCs describing the structural variance occurring throughout the process. All that is required of the user is to decide on how many PCs are sufficient to describe the most important aspects of the structural change. In this case 2 PCs is clearly enough, capturing 98% of the overall process. These PCs make up the x and y axes of Figure **5.5** (a) and (b), with PC1 as the x axis and PC2 as the y axis. The axes are somewhat arbitrary, not simply describing the position of the methane molecule along the nanotube. Rather, they describe multiple degrees of freedom within

the system, in the form of a linear combination of interatomic distances. Methane travelling through a nanotube may not be the most convoluted system to conduct a dimensionality reduction on and an educated guess could be made as to what the 2 PCs represent could be made. For example, one of these PCs may correspond mainly to changes in the interatomic distances between the carbon atom of the methane molecule and carbons along the nanotube, whilst the other to small changes in the diameter of the nanotube as the methane passes through it. However, what these PCs physically represent is irrelevant to this project as the whole iMD-ChemDyME pipeline is designed to circumvent the need for users to define sets of CVs for themselves. There exists other 'less-intuitive' systems for which avoiding such thinking is more advantageous, some of which will be discussed later. Nevertheless, this is still convenient.

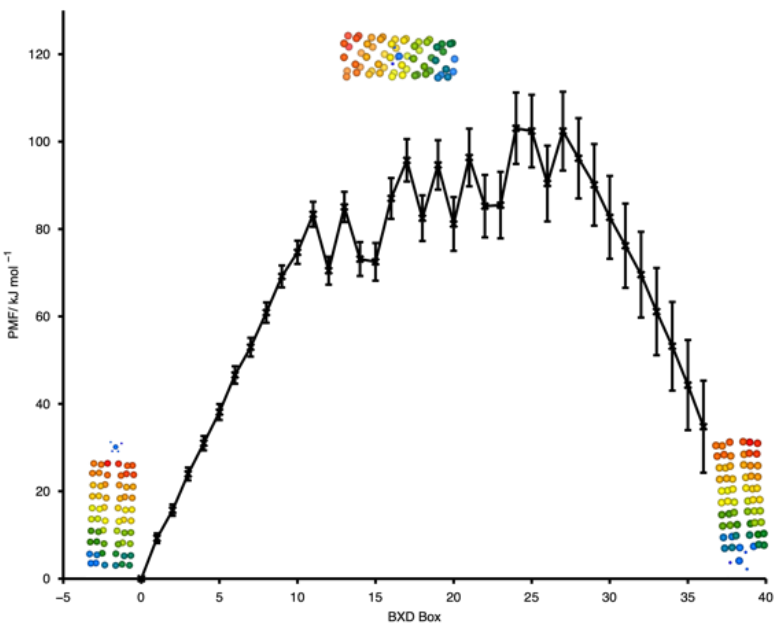
(a)



(b)


80
70
60
50
40
30
30
40
50
60
70
80
90
100

(c)


PMF/ kJ mol⁻¹
120
100
80
60
40
20
0
−5
0
5
10
15
20
25
30
35
40
BXD Box

(d)

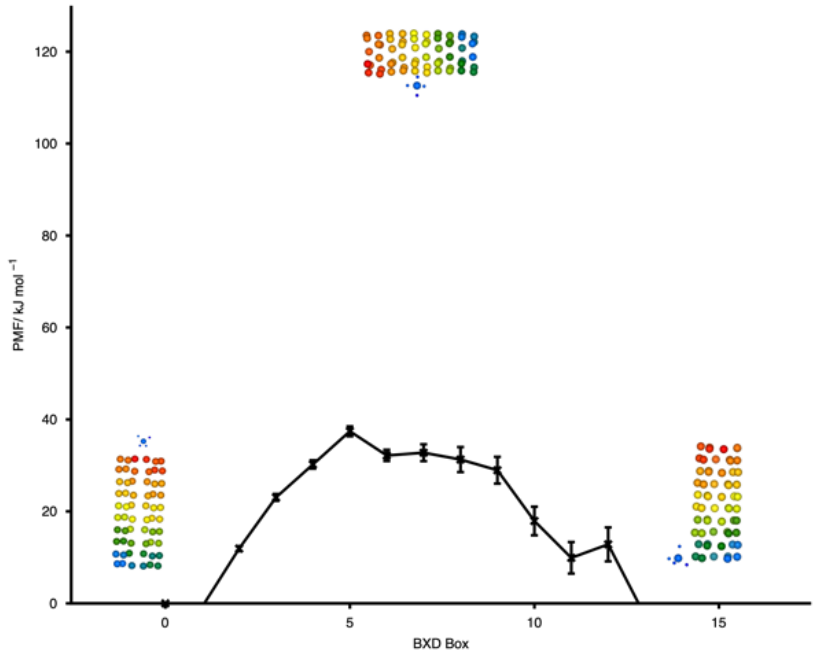


Figure 5.5: Reduced path considering only the carbon atoms in the system when pulling methane through a nanotube projected into CV space for simulations at 500K and 0.5 friction. Superimposed on top are the BXD adaptive sampling points when confining BXD to within 4 and 8 Å of the path shown in frames (a) and (b) respectively. In these frames, the x-axis corresponds to PC1, likely a linear combination of changing interatomic distances (in units of Angstroms) between the carbon atom of the methane and other carbons along the nanotube, and the y-axis to PC2, possibly representative of small changes in the diameter of the nanotube. The free energy profiles from converging runs at the same temperature and fiction, are shown in frames (c) and (d) when simulations are conducted with path boundaries places at 4 and 8 Å respectively. When BXD is allowed to deviate further from the reduced path, it takes the energetically more favourable path alongside the nanotube rather than through it.

Frame (c) and (d) of Figure 35 show the free energy profiles calculated from the results of BXD simulations at 500K and 0.5 friction in which the dynamics were limited to only exploring CV space within 4 or 8 Å from the reduced path. With greater restriction on the phase space available for sampling, BXD is forced to follow the guess path more rigorously (Figure 5.5 (a)), thus it follows the energetically unfavourable path through the nanotube (Figure 5.5 (c)). However, with greater sampling freedom afforded to BXD it is free to pass by the nanotube encountering a much lower free energy penalty as shown in Figure 5.5(b) and (d). It should be noted that one would expect the free energy profile in Figure 5.5(c) to be almost completely flat in the region corresponding to travel through the nanotube, and a reason as to why this is not the case is still under

investigation. Nevertheless, there is a clear distinction between the two free energy profiles resulting from the different levels of dynamical confinement. This combined with the structures taken at relevant time steps in the simulations, as superimposed onto the corresponding free energy profiles, indicates the use of path boundaries to control the path of a trajectory through CV space does indeed work as expected.

Integrating VR guess paths into ChemDyME which BXD can follow to differing degrees of rigour provides an easy method of comparing alternative paths to the same endpoint. This is something which may come in useful when trying to understand the dynamics of biological mechanisms such as transportation through ion channels.

5.3.2 Helicine

5.3.2.1 Background and motivation

Helical structures are commonplace within the human body. In fact, estimates put the percentage of the human proteome made up of alpha-helical membrane proteins to be as high as 27%.[153] Transmissible spongiform encephalopathies (TSEs) are a group of uniquely transmittable neurovegetative prion diseases, in which the formation of tiny holes within the brain lead to it's distinctly 'spongy' degradation and death.

One possible explanation for the onset of TSEs is the 'protein-only hypothesis'. This suggests it is a conformational change of the cellular prion protein from one rich in α-helices (PrP^{C}) to one mainly consisting of β-sheets (PrP^{Sc}) that is responsible for causing such diseases.[154–156] Such a change initiates an autocatalytic reaction leading protein aggregation in the central nervous system and the deterioration of mental and physical abilities of the affected.

Although the initial steps have been taken in investigating the exact mechanism which drives the conformational change of PrP^{C} to PrP^{SC}, there is still much more that needs to be understood.[157] Coupling experimental findings with the insights from computational studies would be a good approach to deepening this understanding along with that of other disease-causing changes in protein conformation. Changing the

screw-sense of helicine works as a test case for proving ChemDyME capable of modelling structural changes in helical type proteins.

5.3.2.2 Method

The helicene system was parameterised using MM3 forcefield parameters defined in a bespoke openMM xml file. Following this, forces and energies for the system were generated using openMM and interfaced with Narupa and ChemDyME such that the MD could be integrated. A guess path was generated in VR in which the helicity of the helicene molecule was reversed. This was then passed into pathReducer to conduct a dimensionality reduction using only every third carbon atom in the system. From this, six PCs consisting of linear combinations of interatomic distances within the helicene molecule were found to capture 98% of the structural variance along the trajectory. The BXD simulations were all run at 500 K to accelerate the BXD process and a friction of 0.01 was used in the Langevin integrator. The adaptive BXD run sampled each box with n_{samp} = 50000 MD steps before placing a new boundary using an epsilon value of 0.025. These adaptive runs were repeated, each time altering the position of the path boundaries so that runs were conducted with them positioned at 0.5, 0.75 and 1.0 Å from the path. For each distance at which the path boundaries were placed, adaptive runs were successfully used to switch the helicene screw-sense before generating milestoning MFPTs with converging runs.

5.3.2.3 Results and discussions

Figure **5.6** shows the reduced path in CV space for reversing the screw-sense of helicine as conducted in Narupa. Superimposed on top are the BXD data points and boundaries taken from an adaptive run following this path to within 0.5 Å. Helicine starts with its screw-sense orientated anticlockwise and BXD follows the guess path until the conformation of helicine is altered so that its helicity is rotating clockwise.

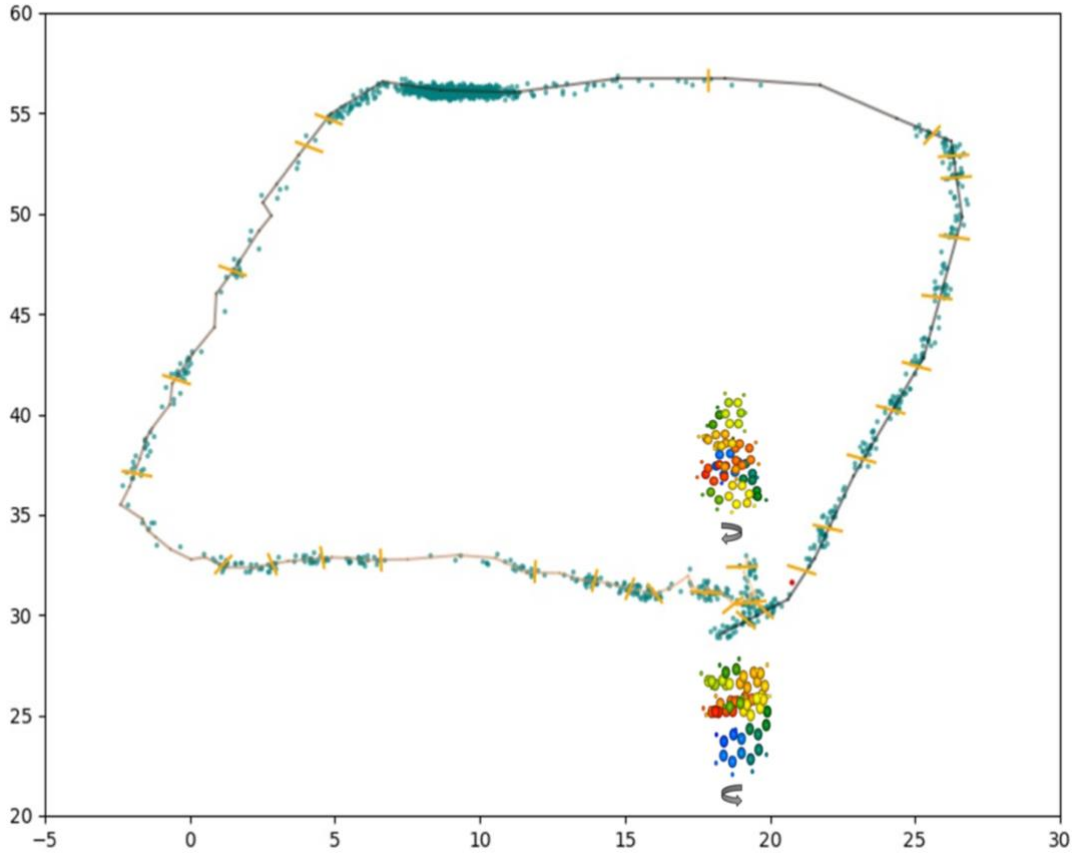


Figure 5.6: Reduced path projected into CV space for changing the screw sense of helicine when considering only every third carbon atom in the system. The data points from adaptive sampling simulations at simulations at 500K using a friction of 0.01 when confining BXD to within 0.5 Å are superimposed on top. Helicine begins with a screw sense orientated in the anticlockwise (bottom) and finishes with a clockwise helicity (top).

For systems such as this, the quality of the guess path is very important and must be as smooth as possible. A comparison of the quality of different VR paths is beyond the scope of this project, designed only to prove the iMD-ChemDyME workflow as an effective method of molecular simulation. However, the ability to control the type and number of atoms considered in the dimensionality reduction in the ChemDyME-pathReducer interface offers an easy and efficient way of optimising the reduced path.

The whole conformational change occurs in a relatively small area of CV space. Therefore, if molecular vibrations are not taken into consideration sections of the path

can become so close that they end up intersecting or containing overlapping BXD boundaries. This can cause problems in converging runs in which the BXD algorithm to either get ‘lost’ or stuck in an infinite loop as another boundary is hit when invoking the velocity inversion procedure upon collision with a boundary. It is for this reason that only every third carbon atom was considered in the dimensionality reduction, resulting in a much smoother path.

However, even with so few atoms considered in the dimensionality reduction, 98% of the important structure variance along the reaction coordinate was captured with 6 PCs. This highlights the true power of utilising iMD for defining reaction coordinates. Simply by constructing the workflow in this manner the need to define a complex reaction coordinate by hand is circumvented. Instead, 6 PCs whose physical meaning may not be immediately obvious are obtained without the need for any ‘chemical intuition’ that may have previously been required.

Figure **5.7** shows the free energy profiles for the helicine system when confining BXD to within 0.5, 0.75 and 1 Å from the guess path, shown by the black, green and purple lines respectively. The shape of these profiles makes sense given the symmetrical nature of the reduced path. There is an increase in free energy as intramolecular bonds break enabling the conformation of helicine to change and energetically unfavourable sterics are encountered along the path; but his is followed by a decrease in free energy as the helicine enters its other enantiomeric form.

In the provision of a good quality reduced path, allowing the dynamics to deviate a small way from the guess path doesn’t have much of an impact on the resulting free energy profile. There isn’t an energetically more favourable path between the anticlockwise and clockwise enantiomers of helicine obvious to the naked eye and so such findings are not surprising.

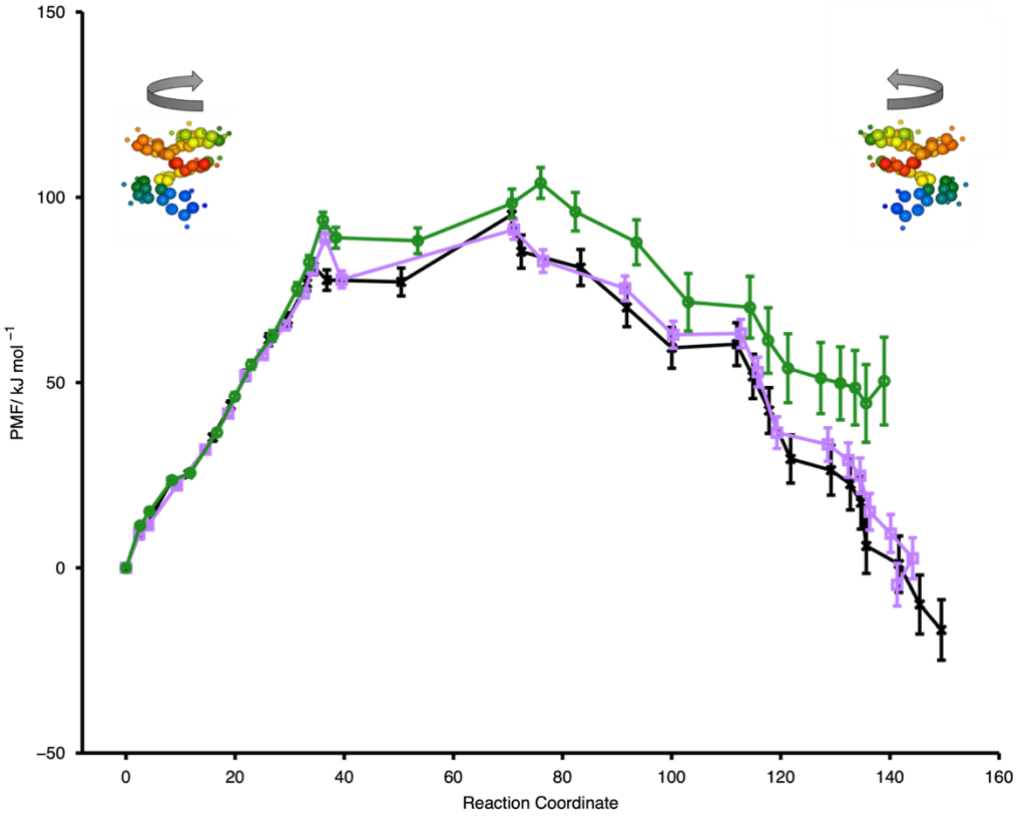


Figure 5.7: Free energy profiles for changing the helicity of helicine, taken from converging runs at 500K, 0.01 friction and using path boundaries placed at 0.5, 0.75 and 1 Å shown by black, purple and green lines respectively. Changing the maximum distance BXD is allowed to stray from path does not change the free energy profiles very much as no other path for changing the screw-sense of helicine that is lower in energy is immediately available.

Combining iMD with adaptive BXD for systems in which deriving a mathematical expression for the reaction coordinate may not be trivial, but with the application of some 'chemical intuition' in the creation of trajectories and their subsequent dimensionality reduction, can provide an efficient way of obtaining both kinetic (MFPTs) and thermodynamic (free energy) data simultaneously.

5.3.3 40 Alanine

5.3.3.1 Background and motivation

Although very rare (one analysis of the PDB showing only 0.8%), knotted proteins have been found to exist in nature. [158–160] In most cases the functional reasons for their

existence have proved challenging to decipher. However, some suggestions include helping to shape the binding site of enzymes and even alter their activity. [160–162]

The structure of all proteins, including knotted ones, are determined by the way in which they fold. This is a process which, if it goes wrong, plays a central role in the causation of neurodegenerative diseases.[163,164] For example, the tau hypothesis states that the main cause of cell death and subsequent development of Alzheimer's disease is the existence of neurofibrillary tangles formed of misfolded, hyperphosphorylated tau protein.[164–166]

Without the identification of such disease-causing pathways, the fight against neurodegenerative diseases is all but lost. Thus, it is paramount to explore all avenues towards such and computational studies are increasingly being used as a way of complementing the experimental studies in this area. Knot tying 40 Alanine can be thought of as a more intricate version of protein misfolding. Therefore, by demonstrating this methodology as being robust enough to obtain free energy profiles for a system of such complexity, it can be inferred as being more than capable of simulating simpler, more biologically relevant problems.

5.3.3.2 System setup

The 40 Alanine was parameterised using MM3 forcefield parameters defined in a bespoke openMM xml file. Subsequently, openMM was used to generate forces and energies for the system such that the MD could be propagated. Using Narupa-iMD a knot was tied in 40 Alanine 20 times to create 20 different guess paths, for each of which the change in the x, y and z coordinates of the carbon atoms only were used to generate a PC file. In the following BXD simulations all 20 PCs were used in the CV space, and the guess trajectory was taken to be a combination of the smoothest five VR trajectories.

To accelerate the dynamics the adaptive run was run at 1000K and with friction of 0.01 in the Langevin integrator. Path boundaries were set to be at a distance 0.1 Å from the reduced path whilst CV space was sampled for n_{samp} = 10000 MD steps before placing

a new boundary using an epsilon value of 0.01. Under these conditions, a knot was tied in 40 Alanine and subsequent converging runs were performed at a friction of 0.1 to generate milestoning MFPTs for diffusion from one box into the next.

5.3.3.3 Results and discussion

Figure **5.8** shows the reduced path for knot tying in 40 Alanine projected into Cartesian CV space. The data points and BXD boundaries from the adaptive run when following the guess path to within 0.1 Å have been superimposed on top.

For each of the previous systems the PCs were given by a linear combination of interatomic distances in the form of $PC = c_1 * r_1 \cdots c_n * r_n$. In this expression n represents the number of unique pairs of atoms considered in the reduction, r the interatomic distances between the atoms in each pair, and c a coefficient detailing the degree to which the change in r helps capture the overall structural variance of the system throughout the trajectory. However, a Cartesian coordinate system is a more appropriate CV for knot tying in 40 Alanine. For this study, each PC is simply the change in the x, y and z coordinate of each atom considered when defining the CV, which in this case was the carbon atoms only.

The angles between atoms cannot be ignored when defining a CV for knot tying. The region of space occupied by the loop when attempting to tie the knot is so cramped that two atoms in this region could remain at the same distance to each other but be in a conformation of 40 Alanine in which the knot ties or one in which it ‘slips past’ the loop, depending on the angle with which they’re orientated to one another. By using a Cartesian coordinate system for 40 Alanine as opposed to conducting a PCA, the interatomic angles remain as feature of the CV and BXD can follow the guess path effectively enough to tie the knot.

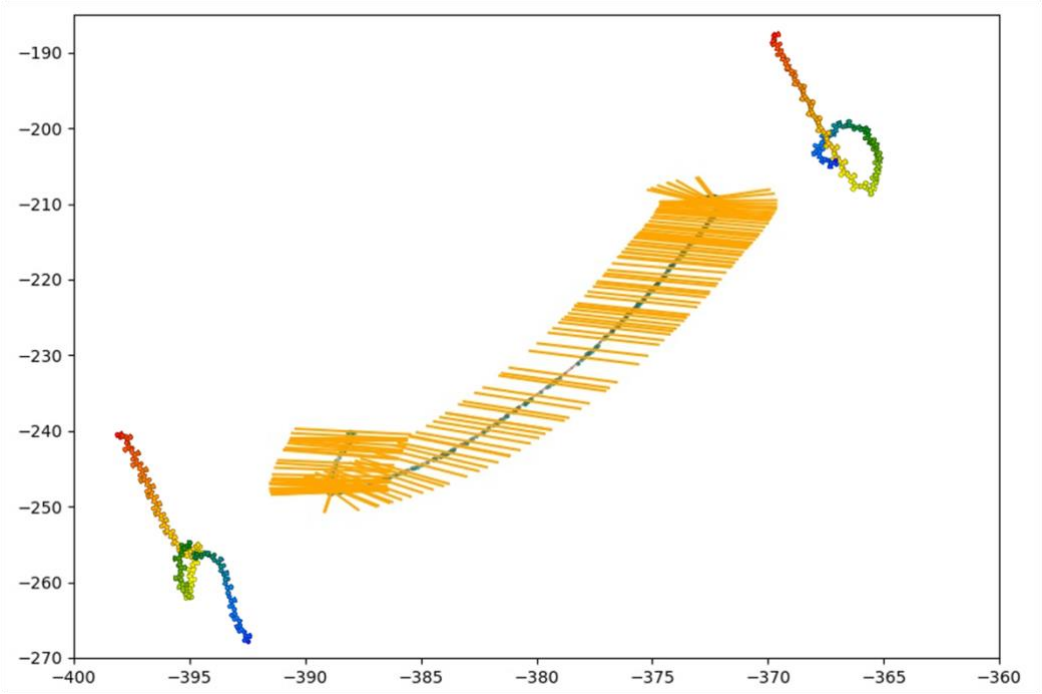


Figure 5.8: Reduced path projected into CV space tying a knot in 40 Alanine considering only the carbons atom in the system. The data points from adaptive sampling simulations at simulations at 1000K using a friction of 0.01 when confining BXD to within 0.1 Å are superimposed onto the path.

Figure **5.9** shows the free energy profile obtained for the 40 Alanine system when using a friction coefficient of 0.1 and confining BXD to within 0.1 Å from the reduced guess path. A higher friction was used in the converging runs than in the adaptive runs after considering the possibility of any correlation effects in steep regions of the potential energy surface for knot tying. Only by placing path boundaries at 0.1 Å from the guess path would the knot tie. If the dynamics were free to stray further from the path it would take the energetically 'easier' path of slipping past the loop, rather than tying the knot. Conversely, restricting the freedom of BXD further so that it can roam only to distances smaller than 0.1 Å from the guess path leaves too little space for the dynamics to move in. This project is formatted to be a proof of concept rather than a set of results for medicinally relevant real-life problems, and as such only attaining results for knot tying under very limited conditions is inconsequential. Rather, it should be thought of that the success of ChemDyME in simulating such an energetically unfavourable and chemically intricate system is a good indicator for success with simpler, more realistic problems.

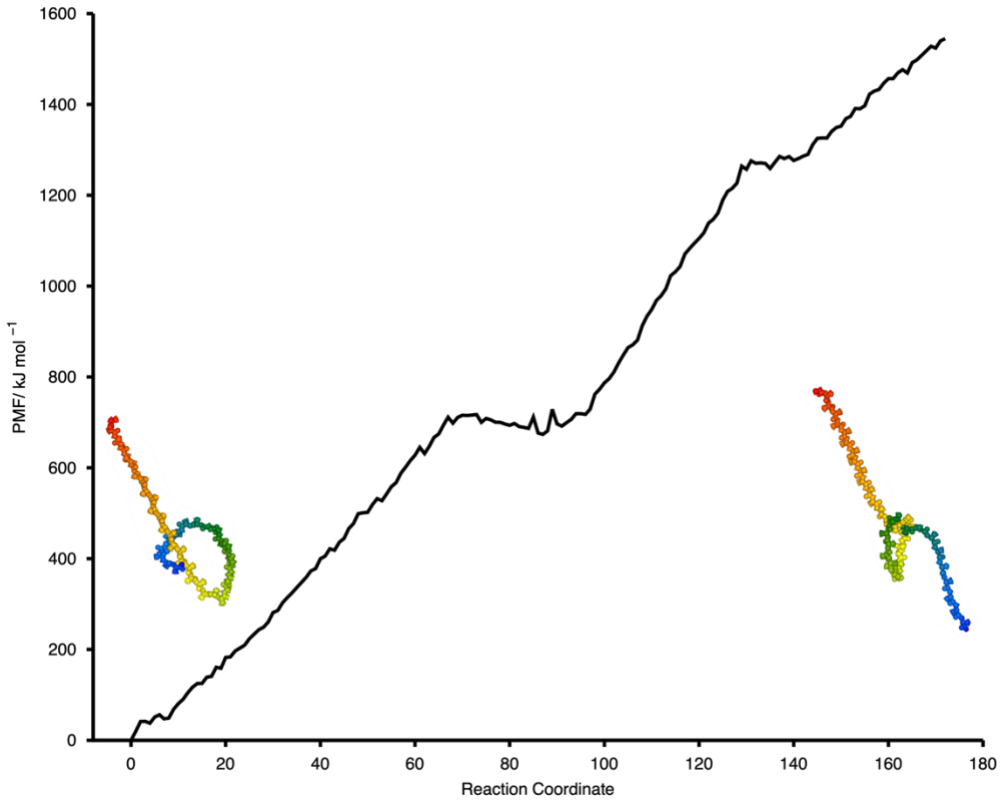


Figure 5.9: Free energy profile for tying a knot in 40 Alanine from BXD converging runs done at 1000K and 0.1 friction with a maximum distance from the path set to 0.1 Å. Only one distance from the path was used as deviation of more than 0.1 Å resulted in no knot tying, whilst confining it more left insufficient room for the dynamics to move in.

5.4 Conclusions

Through studying the three problems above, this workflow has been shown to be suitable for tackling a range of problems, offering advantages specific to the challenges of each:

- Through setting path boundaries at different distances from the guess path in the nanotube system, the ability to contrast different pathways to the same end point just by changing some input parameters in ChemDyME was highlighted. This could be particularly useful if studying a system where more than one route to a given end geometry is suspected. By controlling the path boundaries in such a way that the dynamics must follow the route set out by the guess path, one can

ensure the dynamical path aiming to be studied is indeed that under investigation. For additional piece of mind in such a situation a comparative BXD simulation with much greater freedom to explore alternative pathways could be conducted and the resulting free energy profiles compared.

- Being able to create trajectories in VR which can be interfaced with pathReducer to create a set of PCs for the system automatically, enables the user to elude deriving a CV for the system by hand. For systems like that of helicine, this is a big advantage as the changes in interatomic distance which are the most important in describing the overall structural variance of the system may not be immediately obvious, and so a 'black box' approach to define a CV can have a dramatic increase on efficiency and the ease with which a simulation is done.

- Finally, by tying a knot in 40 Alanine, this workflow has been shown as robust. This system presents a couple of challenges which ChemDyME must overcome for a successful simulation. Firstly, it is such an energetically unfavourable process that BXD will naturally want to work against it. Secondly, small deviations from the guess path can lead to the end of the peptide chain 'slipping past' the loop rather than going through, so that no knot is tied. However, despite these challenges, the iMD-ChemDyME workflow was proves successful in simulating such a challenging problem and is therefore more than likely capable of tackling other tricky systems.

5.5 Future work

Future developments of this project could include:

- Using the above workflow to generate a free energy profile for a well-studied system, so that the results can be compared to previously published work as a final sanity check of the validity of this method.

- Using the iMD-ChemDyME workflow to simulate real-life systems with greater biological significance, for example nucleic acids.

Chapter 6: Further validation of ChemDyME through adaptive sampling of I27

6.1 Introduction

In Chapter 5, a proposed area of future work to further validate the iMD-ChemDyME workflow was to compare the free energy profiles for a well-known system resulting from the interactive BXD method described and an alternative method of simulation. The unfolding of I27 has been studied widely both in experiment [78] and through computational methods, in which recent work has seen free energy profiles for the process emerge. [76,163,164] Therefore, I27 would seem a natural choice for a comparison of free energy profiles generated from the different methodologies. Comparing the free energy profile shown in Figure **4.1**(a) for the unfolding of I27 taken from unbiased BXD simulations done in CHARMM to one produced using the iMD-BXD method would be a good way to further validate the new method whilst simultaneously strengthening the link between the projects in Chapters 4 and 5. As before, all simulations conducted using the iMD-BXD workflow build on the work of O'Connor [128,129] and use the ChemDyME code developed by Robin Shannon.

6.2 Method

Like the work in Chapter 5, MM3 forcefield parameters were defined in a bespoke openMM xml file and used to parameterise the I27 system. Using openMM, the forces and energies for the system were generated and interfaced with Narupa such that the MD could be propagated. Once the system was set up in VR, I27 was pulled apart by its termini to generate a guess path for it's unfolding. This guess path was then minimised before being passed into pathReducer so that a dimensionality reduction could be performed using only the carbon atoms in the system. Three PCs, each representing a linear combination of interatomic distances within I27 were produced from this which accounted for 99.5% of the structural variance along the trajectory.

Due to the large size of the I27 system and the slower nature of python, the language of ChemDyME, as compared to Fortran for which CHARMM is written in, the original guess path was split into 19 trajectories. Each of these were used as a shorter guess path for a separate adaptive BXD run. This resulted in 19 smaller adaptive BXD simulations, corresponding to different chunks of the overall guess path being run in parallel. However, all simulations were run in the same CV space.

Each BXD simulation was run at 298 K and a friction of 0.01 was used in the Langevin integrator. The dynamics in each adaptive BXD run followed the guess path to within 0.5Å and sampled each box with n_{samp} = 25000 MD steps before placing a new boundary with an epsilon value of 0.05. Following this procedure, adaptive runs were conducted for each trajectory, all of which successfully unfolded that section of the overall guess path.

Since the adaptive runs were all completed in the same CV space, it was possible to alter the sets of BXD boundaries for each run to include the last boundary of the *n*th trajectory as the first boundary of the *n*th+1 trajectory. i.e. the uppermost boundary of the first trajectory would be appended to the set of boundaries for the second trajectory as the lowermost boundary, etc. This way, after the converging runs were completed for each of the cropped down trajectories, the resulting free energy profiles could be joined together without leaving any regions of CV space unexplored.

Under the same conditions as before, converging runs were performed for each of the shorter trajectories to generate milestoning MFPTs. These were used to obtain free energy data for each trajectory, which were joined together to create an overall free energy profile for the unfolding of I27.

6.3 Results and Discussion

Figure **6.1** shows the free energy profile obtained using the iMD-ChemDyME workflow and from CHARMM simulations (see Figure **4.1**(a)) for the unfolding of I27. The output from ChemDyME lists columns for each BXD box and the point in CV

space for each value of the free energy. Since both these values are somewhat arbitrary, it was possible to convert the x-axis into the more meaningful quantity of extension by dividing the number of the BXD box by 1.65, such that the free energy profiles from each simulation could be compared along the same axis.

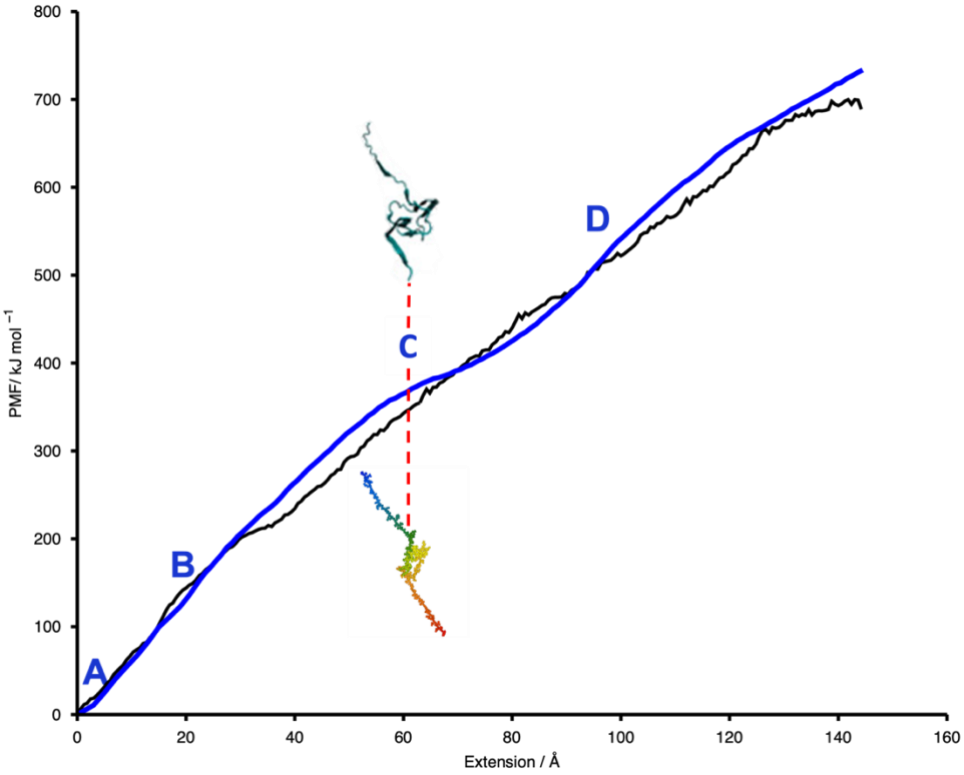


Figure 6.1: Free energy profile for the unfolding of I27 from simulations conducted using CHARMM (blue line) and the iMD-ChemDyME workflow (black line). Compassion of the structures taken at point C taken from reference [45] at the top and from ChemDyME at the bottom, shows them to be similar indicating the conversion from BXD box to extension for the x-axis is sensible. The profiles show very similar free energies for the same extensions and therefore this comparison of data from ChemDyME to the well-established software CHARMM is further evidence of the validity of the iMD-ChemDyME method.

Both profiles in Figure **6.1**, show the free energy for I27 as the hydrogen bonds holding its secondary structure in place are ruptured, but neither reach extensions long enough to change the conformation of I27 to being fully linear as is shown in reference [45] where the rate constants for unfolding simulations in CHARMM were taken from. This is simply because for the work in Chapter 4, only data for extensions up to the point of rupture within I27 was required and the extensions achievable in VR are limited by the

physical reach of the user. However, to double check the check the conversion of BXD box into extension, the structures of I27 at point C in Figure **4.1**(a) (i.e. point C in the corresponding PMF Figure 1 of [45]) and that at the same extension in the ChemDyME simulation were compared and found to be similar indicating the conversion was reasonable. These structures have been superimposed onto the free energy profiles in Figure 6.1, with the all blue structure being taken from reference [45] and the multicoloured one taken from the ChemDyME simulation.

Obtaining a very similar free energy profile from simulations conducted with ChemDyME to those done in a well-established simulation package such as CHARMM further validates the iMD-ChemDyME workflow. Previously, the largest system tested on this workflow was 40 Alanine. Whereas the use of I27 as a test case has proved the effectiveness of the method against larger systems, of sizes more comparable to those found in real-life biological systems. Data on the PBD shows I27 to contain 98 amino acid residues, which when compared to the average length of a protein domain, 100 amino acids [169], would suggest I27 as an excellent test case in relation to potential studies of protein domains.

6.4 Conclusions

A summary of the work in this chapter can be written as:

- Through the use of iMD combined with ChemDyME a free energy profile for the unfolding of I27 was generated and found to closely resemble that of previous work conducted in CHARMM.[76,163,164] Generating such similar results through both methods provides evidence of the validity of the new workflow presented in chapter 5.

- By running the simulations in parallel, the size of the system which can be studied using this procedure can be increased to that of a typical protein domain, indicating the method is suitable for simulating biologically relevant systems.

- Comparing the results of unfolding simulations for I27 from Chapter 4 to those conducted using the iMD-ChemDyME workflow makes for a nice way of connecting the two main projects in this thesis by another means than just the BXD method.

6.5 Future work

The work in this chapter could be expanded on by:

- Increasing the number of atoms in the system under investigation with this method, in search of the limiting system size for this workflow. It would be interesting to know whether the method is more likely to be limited by the number of atoms which can be rendered in VR or the required time to conduct all the simulations in ChemDyME becoming too computationally expensive for it to be worthwhile.

- Taking the free energy profile for unfolding I27 from iMD-ChemDyME and using it as the starting PMF from which to make the modifications required to simulate AFM experiments over a range of velocities as in Chapter 4. This could be done to see if the small differences between the two starting PMFs lead to one set of results with a better match to experiment than the other.

Chapter 7: Conclusions and outlook

The rare event problem in MD sampling is one which still limits the length of simulations to this day and makes producing well converged data for processes which occur over long timescales a challenge to this day. BXD has been shown to be a powerful tool in tackling this problem, by providing a method for enhancing sampling along a reaction coordinate without requiring any modifications to the PES. It is because of this, BXD has proven itself to be a simple yet effective way to simulate long timescale problems such as protein unfolding. But the assumption of equilibrium between BXD boxes means these simulations can only apply to VC AFM experiments at very slow pulling speeds.

However, using the methods introduced in this thesis to modify the results of unbiased BXD simulations it is possible to successfully simulate AFM experiments over the full range of pulling speeds seen in conventional AFM. This is something that other methods of simulation have failed to do. However, for the very highest speeds like the ones seen in HS-FS further work needs to be done to understand why the BXD method fails to reproduce the expected upturn in force with pulling velocity seen in experiment. Nonetheless, the work in this thesis has gone some way to bridging the gap between experimental and computational studies.

Additional work in this thesis has shown the development of a novel method of chemical simulation in which trajectories can be created in VR and then interfaced into a BXD code to guide the dynamics. Several archetypes of biological systems, all presenting their own unique challenges, were simulated using the iMD-ChemDyME workflow and reasonable free energy profiles generated for each. This was taken as an initial ‘proof of concept’ before a comparison of the results from ChemDyME for the well-studied system I27 for which the free energy profile was already known, provided further validation of the method.

Through this work, integrating iMD with BXD has been shown as an effective way to ensure the exact dynamical path between a starting and final geometry is sampled, whilst also eliminating the need to define a reaction coordinate or set of CVs for the

process by hand, increasing the overall ease with which a simulation is conducted. The ability of ChemDyME to interface with the pathReducer code means that BXD simulations can be guided along both a one-dimensional reaction coordinate, as well as through multidimensional CV space with relative ease. Therefore, systems of greater complexity can be described in higher dimensionality CV space without bestowing much extra effort on the user, opening the possibility of using BXD to simulate more convoluted biologically relevant problems.

All work in this thesis is linked through BXD, a technique whose main advantage lies in its simplicity and ability to simultaneously produce kinetic and thermodynamic data for long timescale processes, even reaching the second timescale. The work in this thesis has shown the power of BXD as a base from which to tackle more complicated problems; be that simulating experimental AFM over a wide range of pulling velocities inaccessible to other forms of MD or extending the algorithm so it can follow trajectories from VR in multidimensional CV space. Further work to build on the power of BXD could include:

- Conducting BXD simulations in explicit water to better understand the impact of hydrogen bond formation between ruptured protein domains and the solvent system. MD modelling of protein unfolding in explicit water requires a large simulation box containing many water molecules, which can quickly become prohibitively expensive and so a hybrid hydrodynamics/molecular dynamics approach to such simulations may be considered.

- Extending AFM simulations to concatamers rather than just single protein domains to better match experiment.

- Further study of challenging systems using the iMD-ChemDyME workflow to ensure its suitability in future research.

- Integrating the BXD algorithm into other MD software packages such as LAMMPS or GROMACS for wider usage.

Appendices

Appendix 1

Equations (1.3) and (1.4) for the propagation of an MD trajectory can be derived as follows.

A Taylor expansion around the current atomic coordinates gives an estimate of them at time $t + \delta t$:

$$\vec{r}(t+\delta t) = \vec{r}(t) + \delta t \frac{d\vec{r}(t)}{dt} + \frac{\delta t^2}{2!} \frac{d^2\vec{r}(t)}{dt^2} \cdots \quad \text{(A1 1)}$$

By definition $\frac{d\vec{r}(t)}{dt}$ is $\vec{v}(t)$ and $\frac{d^2\vec{r}(t)}{dt^2}$ is $\vec{a}(t)$, and therefore equation 1.3 is returned

$$\vec{r}(t+\delta t) = \vec{r}(t) + \delta t\vec{v}(t) + \frac{1}{2}\delta t^2\vec{a}(t) \quad (1.3)$$

Similarly, a Taylor expansion around velocities $\vec{v}(t)$ gives:

$$\vec{v}(t+\delta t) = \vec{v}(t) + \delta t \frac{d\vec{v}(t)}{dt} + \frac{\delta t^2}{2!} \frac{d^2v(t)}{dt^2} \cdots \quad \text{(A1 2)}$$

But the third term in this Taylor expansion, $\frac{\delta t^2}{2!} \frac{d^2v(t)}{dt^2}$, is complicated it's simplification not trivial. Instead, a further Taylor expansion can be done around $\frac{d\vec{v}(t)}{dt}$:

$$\frac{\vec{v}(t+\delta t)}{dt} = \frac{d\vec{v}(t)}{dt} + \delta t \frac{d^2v(t)}{dt^2} \cdots \quad \text{(A1 3)}$$

Multiplying A1 3 by $\frac{\delta t}{2}$ and rearranging yields:

$$\frac{\delta t}{2}\frac{\vec{v}(t+\delta t)}{dt} = \frac{\delta t}{2}\frac{d\vec{v}(t)}{dt} + \frac{\delta t^2}{2}\frac{d^2v(t)}{dt^2} \quad \text{(A1 4)}$$

$$\frac{\delta t^2}{2}\frac{d^2v(t)}{dt^2} = \frac{\delta t}{2}\frac{\vec{v}(t+\delta t)}{dt} - \frac{\delta t}{2}\frac{d\vec{v}(t)}{dt} \quad \text{(A1 5)}$$

By substituting A1 5 into A1 2 we get:

$$\vec{v}(t+\delta t) = \vec{v}(t) + \delta t\,\frac{d\vec{v}(t)}{dt} + \frac{\delta t}{2}\frac{\vec{v}(t+\delta t)}{dt} - \frac{\delta t}{2}\frac{d\vec{v}(t)}{dt} \quad \text{(A1 6)}$$

$$= \vec{v}(t) + \frac{\delta t}{2}\frac{\vec{v}(t+\delta t)}{dt} + \frac{\delta t}{2}\frac{d\vec{v}(t)}{dt}$$

Recalling the derivative of velocity with respect to time is acceleration equation A1 6 can be written as in equation 1.4:

$$\vec{v}(t+\delta t) = \vec{v}(t) + \frac{1}{2}\,\delta t\big(\vec{a}(t) + \vec{a}(t+\delta t)\big) \quad \text{(1.4)}$$

Appendix 2

What follows is an illustration of calculating $\nabla\phi^T$ in a BXD velocity inversion.

Equation (2.17) showed how the chain rule is used to obtain the derivative of the constraint function with respect to time in terms of the gradient of ϕ.

This can be calculated as the linear combination of the derivatives of the components of the system's collective variable**:**

$$\frac{d\,\phi}{d\vec{r}} = n_1 \frac{d\,s_1}{d\vec{r}} + n_2 \frac{d\,s_2}{d\vec{r}} \ldots n_M \frac{d\,s_M}{d\vec{r}} \tag{A2 1}$$

Where M is the number of dimensions in the CV.

If, as an illustration we think of a system comprising of atoms A,B and C, in which for reasons of simplicity is restricted to 2 spatial coordinates, the coordinates and corresponding velocities can be defined as $\vec{r} = [a_x, a_y, b_x, b_y, c_x, c_y]$ and $\vec{v} = \left[V_x^a, V_y^a, V_x^b, V_y^b, V_x^c, V_y^c\right]$ respectively. The atomic masses can also be represented in the diagonal matrix:

$$\mathbf{M} = \begin{bmatrix} m_a & 0 & 0 & 0 & 0 & 0 \\ 0 & m_a & 0 & 0 & 0 & 0 \\ 0 & 0 & m_b & 0 & 0 & 0 \\ 0 & 0 & 0 & m_b & 0 & 0 \\ 0 & 0 & 0 & 0 & m_c & 0 \\ 0 & 0 & 0 & 0 & 0 & m_c \end{bmatrix} \tag{A2 2}$$

Then, the CV can be defined in terms of the interatomic distances r_{AB} and r_{BC}:

$$s(\vec{r}) = (r_{AB}, r_{BC}) \tag{A2 3}$$

with:

$$r_{AB} = \sqrt{(a_x - b_x)^2 + \left(a_y - b_y\right)^2} \quad \text{(A2 4)}$$

$$r_{BC} = \sqrt{(b_x - c_x)^2 + \left(b_y - c_y\right)^2}$$

Now there is an expression for $s(\vec{r})$ equation A2 1 can be used to calculate $\nabla\phi$, whose transpose is given by:

$$\nabla\phi^T = \frac{d\phi}{d\vec{r}} = n_1 \frac{dr_{AB}}{d\vec{r}} + n_2 \frac{dr_{BC}}{d\vec{r}} \quad \text{(A2 5)}$$

$$= \begin{bmatrix} n_1(a_x - b_x)/r_{AB} \\ n_1\left(a_y - b_y\right)/r_{AB} \\ n_1(a_x - b_x)/r_{AB} + n_2(b_x - c_x)/r_{AB} \\ n_1\left(a_y - b_y\right)/r_{AB} + n_2\left(b_y - c_y\right)/r_{AB} \\ n_2(b_x - c_x)/r_{AB} \\ n_2\left(b_y - c_y\right)/r_{AB} \end{bmatrix}$$

Where:

$$n_1 \frac{\partial r_{AB}}{\partial \vec{r}} = n_1 \begin{bmatrix} \frac{\partial}{\partial a_x} n_1 \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2} \\ \frac{\partial}{\partial a_y} n_1 \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2} \\ \frac{\partial}{\partial b_x} n_1 \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2} \\ \frac{\partial}{\partial b_y} n_1 \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2} \\ \frac{\partial}{\partial c_x} n_1 \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2} \\ \frac{\partial}{\partial c_y} n_1 \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2} \end{bmatrix} \quad \text{(A2 6)}$$

$$= \begin{bmatrix} n_1(a_x - b_x)/r_{AB} \\ n_1(a_y - b_y)/r_{AB} \\ n_1(a_x - b_x)/r_{AB} \\ n_1(a_y - b_y)/r_{AB} \\ 0 \\ 0 \end{bmatrix}$$

and

$$n_2 \frac{\partial r_{BC}}{\partial \vec{r}} = n_2 \begin{bmatrix} \frac{\partial}{\partial a_x} n_2 \sqrt{(b_x - c_x)^2 + (b_y - c_y)^2} \\ \frac{\partial}{\partial a_y} n_2 \sqrt{(b_x - c_x)^2 + (b_y - c_y)^2} \\ \frac{\partial}{\partial b_x} n_2 \sqrt{(b_x - c_x)^2 + (b_y - c_y)^2} \\ \frac{\partial}{\partial b_y} n_2 \sqrt{(b_x - c_x)^2 + (b_y - c_y)^2} \\ \frac{\partial}{\partial c_x} n_2 \sqrt{(b_x - c_x)^2 + (b_y - c_y)^2} \\ \frac{\partial}{\partial c_y} n_2 \sqrt{(b_x - c_x)^2 + (b_y - c_y)^2} \end{bmatrix} \quad \text{(A2 7)}$$

$$= \begin{bmatrix} 0 \\ 0 \\ n_2(b_x - c_x)/r_{AB} \\ n_2(b_y - c_y)/r_{AB} \\ n_2(b_x - c_x)/r_{AB} \\ n_2(b_y - c_y)/r_{AB} \end{bmatrix}$$

Now we have the atomic coordinates, velocities, and masses and have obtained an expression for $\nabla\phi^T$, along with, we can use equations (2.21)and (2.22) to return the newly inverted velocities.

References

(1) Hirschfelder, J.; Eyring, H.; Topley, B. Reactions Involving Hydrogen Molecules and Atoms. *J. Chem. Phys.* **1936**, *4* (3), 170. https://doi.org/10.1063/1.1749815.

(2) Alder, B. J.; Wainwright, T. E. Studies in Molecular Dynamics. I. General Method. *J. Chem. Phys.* **1959**, *31* (2), 459. https://doi.org/10.1063/1.1730376.

(3) Rahman, A. Correlations in the Motion of Atoms in Liquid Argon. *Phys. Rev.* **1964**, *136* (2A), A405. https://doi.org/10.1103/PhysRev.136.A405.

(4) JA, M.; BR, G.; M, K. Dynamics of Folded Proteins. *Nature* **1977**, *267* (5612), 585–590. https://doi.org/10.1038/267585A0.

(5) Roux, B.; Schulten, K. Computational Studies of Membrane Channels. *Structure* **2004**, *12* (8), 1343–1351. https://doi.org/10.1016/j.str.2004.06.013.

(6) Yang, W.; Gao, Y. Q.; Cui, Q.; Ma, J.; Karplus, M. The Missing Link between Thermodynamics and Structure in F1-ATPase. *Proc. Natl. Acad. Sci.* **2003**, *100* (3), 874–879. https://doi.org/10.1073/pnas.0337432100.

(7) Shan, Y.; Seeliger, M. A.; Eastwood, M. P.; Frank, F.; Xu, H.; Jensen, M. O.; Dror, R. O.; Kuriyan, J.; Shaw, D. E. A Conserved Protonation-Dependent Switch Controls Drug Binding in the Abl Kinase. *Proc. Natl. Acad. Sci.* **2009**, *106* (1), 139–144. https://doi.org/10.1073/pnas.0811223106.

(8) Leach, A. R. *Molecular Modelling Principles and Application*, 2nd ed.; Pearson Education Ltd: Essex, 2001.

(9) Frenkel, D.; Smit, B. *Understanding Molecular Simulation*, 2nd ed.; Academic Press: London, UK, 2002. https://doi.org/10.1016/B978-0-12-267351-1.X5000-7.

(10) Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159* (1), 98. https://doi.org/10.1103/PhysRev.159.98.

(11) Marx, D.; Hutter, J. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods* ; Cambridge University Press: Cambridge, 2009.

(12) Iftimie, R.; Minary, P.; Tuckerman, M. E. Ab Initio Molecular Dynamics: Concepts, Recent Developments, and Future Trends. *Proc. Natl. Acad. Sci.* **2005**, *102* (19), 6654–6659. https://doi.org/10.1073/PNAS.0500193102.

(13) Vanommeslaeghe, K.; Guvench, O.; MacKerell, A. D.; Jr. Molecular Mechanics. *Curr. Pharm. Des.* **2014**, *20* (20), 3281.

(14) Hospital, A.; Goñi, J. R.; Orozco, M.; Gelpí, J. L. Molecular Dynamics Simulations: Advances and Applications. *Adv. Appl. Bioinform. Chem.* **2015**, *19* (8), 37–47. https://doi.org/10.2147/AABC.S70333.

(15) Adcock, S. A.; Mccammon, J. A. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.* **2006**, *106* (5), 1589–1615. https://doi.org/10.1021/cr040426m.

(16) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174. https://doi.org/10.1002/JCC.20035.

(17) Maple, J. R.; Dinurt, U.; Hagler, A. T. *Derivation of Force Fields for Molecular Mechanics and Dynamics from Ab Initio Energy Surfaces*; 1988; Vol. 85.

(18) González, M. A. Force Fields and Molecular Dynamics Simulations. *École thématique la Société Française la Neutron.* **2011**, *12*, 169–200. https://doi.org/10.1051/SFN/201112009.

(19) Torrie, G. M. M.; Valleau, J. P. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23* (2), 187–199. https://doi.org/10.1016/0021-9991(77)90121-8.

(20) West, A. M. A.; Elber, R.; Shalloway, D. Extending Molecular Dynamics Timescales with Milestoning: Example of Complex Kinetics in a Solvated Peptide. *J. Chem. Phys.* **2007**, *126* (14). https://doi.org/10.1063/1.2716389.

(21) Faradjian, A. K.; Elber, R. Computing Time Scales from Reaction Coordinates by Milestoning. *J. Chem. Phys.* **2004**, *120* (23), 10880–10889. https://doi.org/10.1063/1.1738640.

(22) Wei, W.; Elber, R. ScMile: A Script to Investigate Kinetics with Short Time Molecular Dynamics Trajectories and the Milestoning Theory. *J. Chem. Theory Comput.* **2020**, *16* (2), 860. https://doi.org/10.1021/ACS.JCTC.9B01030.

(23) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*; Academic press: California, 2002; Vol. 1.

(24) Dill, K. A.; MacCallum, J. L. The Protein-Folding Problem, 50 Years On. *Science* **2012**, *338* (6110), 1042–1046. https://doi.org/10.1126/SCIENCE.1219021.

(25) Stank, A.; Kokh, D. B.; Fuller, J. C.; Wade, R. C. Protein Binding Pocket

Dynamics. *Acc. Chem. Res.* **2016**, *49* (5), 809–815. https://doi.org/10.1021/ACS.ACCOUNTS.5B00516.

(26) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chem. Rev.* **2016**, *116* (14), 7898–7936. https://doi.org/10.1021/ACS.CHEMREV.6B00163/SUPPL_FILE/CR6B00163_LIVESLIDES.MP4.

(27) SHAW, D. .; GROSSMAN, J. W.; Bank, J. A.; Batson, B.; Butts, J. A.; Chao, J. C.; Deneroff, M. M.; Dror, R. O.; Even, A.; Fenton, C. H.; et al. Anton 2: RAISING the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*; IEEE Press: New Orleans, 2014; pp 41–53.

(28) Piana, S.; Klepeis, J. L.; Shaw, D. E. Assessing the Accuracy of Physical Models Used in Protein-Folding Simulations: Quantitative Evidence from Long Molecular Dynamics Simulations. *Curr. Opin. Struct. Biol.* **2014**, *24* (1), 98–105. https://doi.org/10.1016/J.SBI.2013.12.006.

(29) Truhlar, D. G.; Garrett, B. C.; Klippenstein, S. J. Current Status of Transition-State Theory. *J. Phys. Chem.* **1996**, *100* (31), 12771–12800. https://doi.org/10.1021/JP953748Q.

(30) Leopold, P. E.; Montal, M.; Onuchic, J. N. Protein Folding Funnels: A Kinetic Approach to the Sequence-Structure Relationship. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89* (18), 8721. https://doi.org/10.1073/PNAS.89.18.8721.

(31) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. Replica Exchange with Solute Tempering: A Method for Sampling Biological Systems in Explicit Water. *Proc. Natl. Acad. Sci.* **2005**, *102* (39), 13749–13754. https://doi.org/10.1073/PNAS.0506346102.

(32) Arefi, H. H.; Yamamoto, T. Communication: Self-Assembly of a Model Supramolecular Polymer Studied by Replica Exchange with Solute Tempering. *J. Chem. Phys.* **2017**, *147* (21), 211102. https://doi.org/10.1063/1.5008275.

(33) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314* (1–2), 141–151. https://doi.org/10.1016/S0009-2614(99)01123-9.

(34) Qi, R.; Wei, G.; Ma, B.; Nussinov, R. Replica Exchange Molecular Dynamics:

A Practical Application Protocol with Solutions to Common Problems and a Peptide Aggregation and Self-Assembly Example. *Methods Mol. Biol.* **2018**, *1777*, 101. https://doi.org/10.1007/978-1-4939-7811-3_5.

(35) Valsson, O.; Tiwary, P.; Parrinello, M. Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annu. Rev. Phys. Chem.* **2016**, *67* (1), 159–184. https://doi.org/10.1146/annurev-physchem-040215-112229.

(36) Mills, M.; Andricioaei, I. An Experimentally Guided Umbrella Sampling Protocol for Biomolecules. *J. Chem. Phys.* **2008**, *129* (11). https://doi.org/10.1063/1.2976440.

(37) You, W.; Tang, Z.; Chang, C. E. A. Potential Mean Force from Umbrella Sampling Simulations: What Can We Learn and What Is Missed? *J. Chem. Theory Comput.* **2019**, *15* (4), 2433–2443. https://doi.org/10.1021/ACS.JCTC.8B01142/SUPPL_FILE/CT8B01142_SI_001.PDF.

(38) Souaille, M.; Roux, B. Extension to the Weighted Histogram Analysis Method: Combining Umbrella Sampling with Free Energy Calculations. *Comput. Phys. Commun.* **2001**, *135* (1), 40–57. https://doi.org/10.1016/S0010-4655(00)00215-0.

(39) Bartels, C.; Karplus, M. Multidimensional Adaptive Umbrella Sampling: Applications to Main Chain and Side Chain Peptide Conformations. *J. Comput. Chem.* **1997**, *18* (12), 1450–1462. https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1450::AID-JCC3>3.0.CO;2-I.

(40) Booth, J. J. New Applications of Boxed Molecular Dynamics: Efficient Simulation of Rare Events, University of Leeds, 2016.

(41) Mezei, M. Adaptive Umbrella Sampling: Self-Consistent Determination of the Non-Boltzmann Bias. *J. Comput. Phys.* **1987**, *68* (1), 237–248. https://doi.org/10.1016/0021-9991(87)90054-4.

(42) Pechukas, P. Transition State Theory. *Annu. Rev. Phys. Chem.* **1981**, *32* (1), 159–177. https://doi.org/10.1146/ANNUREV.PC.32.100181.001111.

(43) Votapka, L. W.; Amaro, R. E. Multiscale Estimation of Binding Kinetics Using Brownian Dynamics, Molecular Dynamics and Milestoning. *PLOS Comput. Biol.* **2015**, *11* (10), e1004381. https://doi.org/10.1371/JOURNAL.PCBI.1004381.

(44) Glowacki, D. R.; Booth, J.; Vazquez, S.; Martinez-Nunez, E.; Marks, A.; Rodgers, J.; Shalashilin, D. V. Recent Applications of Boxed Molecular Dynamics: A Simple Multiscale Technique for Atomistic Simulations. *Philos. Trans. R. Soc. Lond. A.* **2014**, *372*, 20130384. https://doi.org/10.1098/rsta.2013.0384.

(45) Booth, J. J.; Shalashilin, D. V. Fully Atomistic Simulations of Protein Unfolding in Low Speed Atomic Force Microscope and Force Clamp Experiments with the Help of Boxed Molecular Dynamics. *J. Phys. Chem. B* **2016**, *120* (4), 700–708. https://doi.org/10.1021/acs.jpcb.5b11519.

(46) Glowacki, D. R.; Paci, E.; Shalashilin, D. V. Boxed Molecular Dynamics: A Simple and General Technique for Accelerating Rare Event Kinetics and Mapping Free Energy in Large Molecular Systems. *J. Phys. Chem. B* **2009**, *113* (52), 16603–16611. https://doi.org/10.1021/jp9074898.

(47) Martínez-Núñez, E.; Shalashilin, D. V. Acceleration of Classical Mechanics by Phase Space Constraints. *J. Chem. Theory Comput.* **2006**, *2* (4), 912–919. https://doi.org/10.1021/ct060042z.

(48) Glowacki, D. R.; Paci, E.; Shalashilin, D. V. Boxed Molecular Dynamics: Decorrelation Time Scales and the Kinetic Master Equation. *J. Chem. Theory Comput* **2011**, *7*, 1244–1252. https://doi.org/10.1021/ct200011e.

(49) Voter, A. F. A Method for Accelerating the Molecular Dynamics Simulation of Infrequent Events. **1997**, *106* (11), 4665. https://doi.org/10.1063/1.473503.

(50) Buchete, N.-V.; Hummer, G. Coarse Master Equations for Peptide Folding Dynamics†. *J. Phys. Chem. B.* **2008**, *112* (19), 6057–6069. https://doi.org/10.1021/JP0761665.

(51) Wales, D. J. Energy Landscapes: Calculating Pathways and Rates. *Int. Rev. Phys. Chem.* **2006**, *25* (2), 237–282. https://doi.org/10.1080/01442350600676921.

(52) Shalashilin, D. V; Beddard, G. S.; Paci, E.; Glowacki, D. R. Peptide Kinetics from Picoseconds to Microseconds Using Boxed Molecular Dynamics: Power Law Rate Coefficients in Cyclisation Reactions. *J. Chem. Phys.* **2012**, *137*, 165102. https://doi.org/10.1063/1.4759088.

(53) O'Connor, M.; Paci, E.; McIntosh-Smith, S.; Glowacki, D. R. Adaptive Free Energy Sampling in Multidimensional Collective Variable Space Using Boxed Molecular Dynamics. *Faraday Discuss.* **2016**, *195* (0), 395–419. https://doi.org/10.1039/C6FD00138F.

(54) Lotsted, P. Mechanical Systems of Rigid Bodies Subject to Unilateral Constraints. *SIAM J. Appl. Math.* **1982**, *42* (2), 281–296. https://doi.org/10.1137/0142022ï.

(55) O'Connor, C.; Adams, J. *Essentials of Cell Biology*; MA: NPG Education: Cambridge, 2010.

(56) Berg, J. M.; Tymoczko, J. L.; Stryer, L. *Biochemistry. 5th Ed.*, 5th ed.; New York: W H Freeman: New York, 2002.

(57) U.S National Library of Medicine. What are proteins and what do they do? https://ghr.nlm.nih.gov/primer/howgeneswork/protein (accessed Jan 21, 2019).

(58) *Oxford Dictionary of Biochemistry and Molecular Biology*; Cammack, R., Atwood, T., Campbell, P., Parish, H., Smith, A., Vella, F., Stirling, J., Eds.; Oxford University Press, 2006. https://doi.org/10.1093/acref/9780198529170.001.0001.

(59) Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. *Molecular Biology of the Cell. 4th Ed.*, 4th ed.; Garland Science: New York, 2002.

(60) Lodish, H.; Berk, A.; Zipursky, S. L.; Matsudaira, P.; Baltimore, D.; Darnell, J. *Molecular Cell Biology. 4th Ed.*, 4th ed.; W. H. Freeman: new York, 2000.

(61) Zhou, N. E.; Kay, C. M.; Hodges, R. S. The Role of Interhelical Ionic Interactions in Controlling Protein Folding and Stability: De Novo Designed Synthetic Two-Stranded α-Helical Coiled-Coils. *J. Mol. Biol.* **1994**, *237* (4), 500–512. https://doi.org/10.1006/JMBI.1994.1250.

(62) Damodaran, S. Beyond the Hydrophobic Effect: Critical Function of Water at Biological Phase Boundaries — A Hypothesis. *Adv. Colloid Interface Sci.* **2015**, *221*, 22–33. https://doi.org/10.1016/j.cis.2015.03.005.

(63) Volkert, M. High Pressure-Low Temperature Induced Structures in Dairy Foams and Protein Model Systems, Technischen Universität Berlin, 2009.

(64) Janin, J.; Sternberg, M. J. E. Protein Flexibility, Not Disorder, Is Intrinsic to Molecular Recognition. *F1000 Biol. Rep.* **2013**, *5*, 2. https://doi.org/10.3410/B5-2.

(65) Minajeva, A.; Kulke, M.; Fernandez, J. M.; Linke, W. A. Unfolding of Titin Domains Explains the Viscoelastic Behavior of Skeletal Myofibrils. *Biophys. J.* **2001**, *80* (3), 1442–1451. https://doi.org/10.1016/S0006-3495(01)76116-4.

(66) von Castelmur, E.; Marino, M.; Svergun, D. I.; Kreplak, L.; Ucurum-Fotiadis, Z.; Konarev, P. V.; Urzhumtsev, A.; Labeit, D.; Labeit, S.; Mayans, O. A

Regular Pattern of Ig Super-Motifs Defines Segmental Flexibility as the Elastic Mechanism of the Titin Chain. *Proc. Natl. Acad. Sci.* **2008**, *105* (4), 1186–1191. https://doi.org/10.1073/pnas.0707163105.

(67) Rico, F.; Rigato, A.; Picas, L.; Scheuring, S. Mechanics of Proteins with a Focus on Atomic Force Microscopy. *J. Nanobiotechnology* **2013**, *11 Suppl 1* (Suppl 1), S3. https://doi.org/10.1186/1477-3155-11-S1-S3.

(68) Lewandowski, J. R.; Halse, M. E.; Blackledge, M.; Emsley, L. Protein Dynamics. Direct Observation of Hierarchical Protein Dynamics. *Science (80-.).* **2015**, *348* (6234), 578–581. https://doi.org/10.1126/science.aaa6111.

(69) Knab, J. R.; Chen, J.-Y.; He, Y.; Markelz, A. G. Terahertz Measurements of Protein Relaxational Dynamics. *Proc. IEEE* **2007**, *95* (8), 1605–1610. https://doi.org/10.1109/JPROC.2007.898906.

(70) Rasmussen, B. F.; Stock, A. M.; Ringe, D.; Petsko, G. A. Crystalline Ribonuclease A Loses Function below the Dynamical Transition at 220 K. *Nature* **1992**, *357*, 423–424. https://doi.org/10.0.

(71) Doster, W.; Cusack, S.; Petry, W. Dynamical Transition of Myoglobin Revealed by Inelastic Neutron Scattering. *Nature* **1989**, *337* (6209), 754–756. https://doi.org/10.1038/337754a0.

(72) Frauenfelder, H.; Chen, G.; Berendzen, J.; Fenimore, P. W.; Jansson, H. N.; Mcmahon, B. H.; Stroe, I. R.; Swenson, J.; Young, R. D. *A Unified Model of Protein Dynamics*; 2009; Vol. 106.

(73) Last, J. A.; Russell, P.; Nealey, P. F.; Murphy, C. J. The Applications of Atomic Force Microscopy to Vision Science. *Invest. Ophthalmol. Vis. Sci.* **2010**, *51* (12), 6083–6094. https://doi.org/10.1167/iovs.10-5470.

(74) Puricelli, L.; Galluzzi, M.; Schulte, C.; Podestà, A.; Milani, P. Nanomechanical and Topographical Imaging of Living Cells by Atomic Force Microscopy with Colloidal Probes. *Rev. Sci. Instrum.* **2015**, *86* (3). https://doi.org/10.1063/1.4915896.

(75) Carrion-Vazquez, M.; Oberhauser, A. F.; Fowler, S. B.; Marszalek, P. E.; Broedel, S. E.; Clarke, J.; Fernandez, J. M. Mechanical and Chemical Unfolding of a Single Protein: A Comparison. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96* (7), 3694–3699. https://doi.org/10.1073/PNAS.96.7.3694.

(76) Florin, E.-L.; Rief, M.; Lehmann, H.; Ludwig, M.; Dornmair, C.; Moy, V. T. T.; Gaub, H. E. E. *Sensing Specific Molecular Interactions with the Atomic Force*

Microscope; Elsevier, 1995; Vol. 10, pp 895–901. https://doi.org/10.1016/0956-5663(95)99227-C.

(77) Moy, V. T.; Florin, E. L.; Rief, M.; Lehmann, H.; Ludwig, M.; Gaub, H. E.; Dornmair, K. Probing the Forces between Complementary Strands of DNA with the Atomic Force Microscope. *Proc. SPIE - Int. Soc. Opt. Eng.* **1995**, *2384*, 2–12.

(78) Linke, W. A.; Grützner, A. Pulling Single Molecules of Titin by AFM—Recent Advances and Physiological Implications. *Pflügers Arch. - Eur. J. Physiol.* **2008**, *456* (1), 101–115. https://doi.org/10.1007/s00424-007-0389-x.

(79) Rico, F.; Gonzalez, L.; Casuso, I.; Puig-vidal, M.; Scheuring, S. High-Speed Force Spectroscopy Molecular Dynamics Simulations. *Science (80-.).* **2013**, *342* (November), 741–743. https://doi.org/10.1126/science.1239764.

(80) Rief, M.; Gautel, M.; Oesterhelt, F.; Fernandez, J. M.; Gaub, H. E. *Reversible Unfolding of Individual Titin Immunoglobulin Domains by AFM*; 1997; Vol. 276.

(81) Brujić, J.; Hermans Z., R. I.; Walther, K. A.; Fernandez, J. M. Single-Molecule Force Spectroscopy Reveals Signatures of Glassy Dynamics in the Energy Landscape of Ubiquitin. *Nat. Phys.* **2006**, *2*, 282–286. https://doi.org/10.1038/nphys269.

(82) Garcia-Manyes, S.; Brujić, J.; Badilla, C. L.; Fernández, J. M. Force-Clamp Spectroscopy of Single-Protein Monomers Reveals the Individual Unfolding and Folding Pathways of I27 and Ubiquitin. *Biophys. J.* **2007**, *93* (7), 2436–2446. https://doi.org/10.1529/biophysj.107.104422.

(83) Oberhauser, A. F.; Hansma, P. K.; Carrion-Vazquez, M.; Fernandez, J. M. Stepwise Unfolding of Titin under Force-Clamp Atomic Force Microscopy. *Proc. Natl. Acad. Sci.* **2001**, *98* (2), 468–472. https://doi.org/10.1073/pnas.021321798.

(84) Fernandez, J. M. Force-Clamp Spectroscopy Monitors the Folding Trajectory of a Single Protein. *Science (80-.).* **2004**, *303* (5664), 1674–1678. https://doi.org/10.1126/science.1092497.

(85) Li, H.; Fernandez, J. M. Mechanical Design of the First Proximal Ig Domain of Human Cardiac Titin Revealed by Single Molecule Force Spectroscopy. *J. Mol. Biol.* **2003**, *334* (1), 75–86. https://doi.org/10.1016/J.JMB.2003.09.036.

(86) Taniguchi, Y.; Kobayashi, A.; Kawakami, M. Mechanical Unfolding Studies of

Protein Molecules. *Biophys. (Nagoya-shi, Japan)* **2012**, *8*, 51–58. https://doi.org/10.2142/biophysics.8.51.

(87) Bouchiat, C.; Wang, M. D.; Allemand, J.-F.; Strick, T.; Block, S. M.; Croquette, V. Estimating the Persistence Length of a Worm-Like Chain Molecule from Force-Extension Measurements. *Biophys. J.* **1999**, *76* (1), 409–413. https://doi.org/10.1016/S0006-3495(99)77207-3.

(88) Hsu, H.-P.; Paul, W.; Binder, K. Breakdown of the Kratky-Porod Wormlike Chain Model for Semiflexible Polymers in Two Dimensions. *EPL (Europhysics Lett.* **2011**, *95*, 68004. https://doi.org/10.1209/0295-5075/95/68004.

(89) Fernandez, J. M. Fingerprinting Single Molecules In Vivo. *Biophys. J.* **2005**, *89*, 3676–3677. https://doi.org/10.1529/biophysj.105.072223.

(90) Hedberg, C.; Toledo, A. G.; Gustafsson, C. M.; Larson, G.; Oldfors, A.; Macao, B. Hereditary Myopathy with Early Respiratory Failure Is Associated with Misfolding of the Titin Fibronectin III 119 Subdomain. *Neuromuscul. Disord.* **2014**, *24* (5), 373–379. https://doi.org/10.1016/j.nmd.2014.02.003.

(91) Best, R. B.; Clarke, J. What Can Atomic Force Microscopy Tell Us about Protein Folding? *Chem. Commun.* **2002**, No. 3, 183–192. https://doi.org/10.1039/b108159b.

(92) Tych, K. M.; Hughes, M. L.; Bourke, J.; Taniguchi, Y.; Kawakami, M.; Brockwell, D. J.; Dougan, L. Optimizing the Calculation of Energy Landscape Parameters from Single-Molecule Protein Unfolding Experiments. *Phys. Rev. E Stat. Nonlinear, Soft Matter Phys.* **2015**, *91* (1), 012710-012719 (10). https://doi.org/10.1103/PhysRevE.91.012710.

(93) Bell, G. Models for the Specific Adhesion of Cells to Cells. *Science (80-.).* **1978**, *200* (4342), 618–627. https://doi.org/10.1126/science.347575.

(94) Evans, E.; Ritchie, K. Dynamic Strength of Molecular Adhesion Bonds. *Biophys. J.* **1997**, *72* (4), 1541–1555. https://doi.org/10.1016/S0006-3495(97)78802-7.

(95) Yu-Shiu Lo; Ying-Jie Zhu, A.; Thomas P. Beebe, J. . Loading-Rate Dependence of Individual Ligand−Receptor Bond-Rupture Forces Studied by Atomic Force Microscopy. *Langmuir* **2001**, *17*, 3741–3748. https://doi.org/10.1021/LA001569G.

(96) Noy, A.; Friddle, R. W. Practical Single Molecule Force Spectroscopy: How to Determine Fundamental Thermodynamic Parameters of Intermolecular Bonds

with an Atomic Force Microscope. *Methods* **2013**, *60* (2), 142–150. https://doi.org/10.1016/J.YMETH.2013.03.014.

(97) Hummer, G.; Szabo, A. Kinetics from Nonequilibrium Single-Molecule Pulling Experiments. *Biophys. J.* **2003**, *85* (1), 5–15. https://doi.org/10.1016/S0006-3495(03)74449-X.

(98) Heymann, B.; Grubmüller, H. *Dynamic Force Spectroscopy of Molecular Adhesion Bonds*; 2000.

(99) Brockwell, D. J.; Beddard, G. S.; Clarkson, J.; Zinober, R. C.; Blake, A. W.; Trinick, J.; Olmsted, P. D.; Smith, D. A.; Radford, S. E. The Effect of Core Destabilization on the Mechanical Resistance of I27. *Biophys. J.* **2002**, *83* (1), 458–472. https://doi.org/10.1016/S0006-3495(02)75182-5.

(100) Satō, A. *Introduction to Practice of Molecular Simulation : Molecular Dynamics, Monte Carlo, Brownian Dynamics, Lattice Boltzmann, Dissipative Particle Dynamics*; Elsevier: London, 2011.

(101) Rojnuckarin, A.; Kim, S.; Subramaniam, S. Brownian Dynamics Simulations of Protein Folding: Access to Milliseconds Time Scale and Beyond. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95* (8), 4288–4292. https://doi.org/10.1073/pnas.95.8.4288.

(102) Marszalek, P. E.; Lu, H.; Li, H.; Carrion-Vazquez, M.; Oberhauser, A. F.; Schulten, K.; Fernandez, J. M. Mechanical Unfolding Intermediates in Titin Modules. *Nature* **1999**, *402* (6757), 100–103. https://doi.org/10.1038/47083.

(103) Crampton, N.; Brockwell, D. J. Unravelling the Design Principles for Single Protein Mechanical Strength. *Curr. Opin. Struct. Biol.* **2010**, *20* (4), 508–517. https://doi.org/10.1016/j.sbi.2010.05.005.

(104) Brockwell, D. J.; Beddard, G. S.; Paci, E.; West, D. K.; Olmsted, P. D.; Smith, D. A.; Radford, S. E. Mechanically Unfolding the Small, Topologically Simple Protein L. *Biophys. J.* **2005**, *89* (1), 506–519. https://doi.org/10.1529/biophysj.105.061465.

(105) Strzelecki, J.; Mikulska, K.; Lekka, M.; Kulik, A.; Balter, A.; Nowak, W. *AFM Force Spectroscopy and Steered Molecular Dynamics Simulation of Protein Contactin 4*; 2009; Vol. 116.

(106) Lu, H.; Schulten, K. The Key Event in Force-Induced Unfolding of Titin's Immunoglobulin Domains. *Biophys. J.* **2000**, *79*, 51–65.

(107) Lu, H.; Schulten, K. Steered Molecular Dynamics Simulations of Force-Induced

Protein Domain Unfolding. *Proteins Struct. Funct. Genet.* **1999**, *35* (4), 453–463. https://doi.org/10.1002/(SICI)1097-0134(19990601)35:4<453::AID-PROT9>3.0.CO;2-M.

(108) Lu, H.; Isralewitz, B.; Krammer, A.; Vogel, V.; Schulten, K. Unfolding of Titin Immunoglobulin Domains by Steered Molecular Dynamics Simulation. *Biophys. J.* **1998**, *75* (2), 662–671. https://doi.org/10.1016/S0006-3495(98)77556-3.

(109) Lee, E. H.; Hsin, J.; Sotomayor, M.; Comellas, G.; Schulten, K. Discovery Through the Computational Microscope. *Structure* **2009**, *17* (10), 1295–1306. https://doi.org/10.1016/J.STR.2009.09.001.

(110) Lazaridis, T.; Karplus, M. Effective Energy Function for Proteins in Solution. https://doi.org/10.1002/(SICI)1097-0134(19990501)35:2.

(111) Friddle, R. W.; Noy, A.; De Yoreo, J. J. Interpreting the Widespread Nonlinear Force Spectra of Intermolecular Bonds. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (34), 13573–13578. https://doi.org/10.1073/pnas.1202946109.

(112) Gao, M.; Wilmanns, M.; Schulten, K. *Steered Molecular Dynamics Studies of Titin I1 Domain Unfolding*; 2002.

(113) Balsera, M.; Stepaniants, S.; Izrailev, S.; Oono, Y.; Schulten, K. Reconstructing Potential Energy Functions from Simulated Force-Induced Unbinding Processes. *Biophys. J.* **1997**, *73* (3), 1281–1287. https://doi.org/10.1016/S0006-3495(97)78161-X.

(114) Dudko, O. K.; Hummer, G.; Szabo, A. Intrinsic Rates and Activation Free Energies from Single-Molecule Pulling Experiments. **2006**. https://doi.org/10.1103/PhysRevLett.96.108101.

(115) Nunes, J. M.; Hensen, U.; Ge, L.; Lipinsky, M.; Helenius, J.; Grubmüller, H.; Muller, D. J. A "Force Buffer" Protecting Immunoglobulin Titin. *Angew. Chemie - Int. Ed.* **2010**, *49* (20), 3528–3531. https://doi.org/10.1002/ANIE.200906388.

(116) O'Connor, M.; Deeks, H. M.; Dawn, E.; Metatla, O.; Roudaut, A.; Sutton, M.; Thomas, L. M.; Glowacki, B. R.; Sage, R.; Tew, P.; et al. Sampling Molecular Conformations and Dynamics in a Multiuser Virtual Reality Framework. *Sci. Adv.* **2018**, *4* (6), eaat2731.

(117) Narumi, T.; Kameoka, S.; Taiji, M.; Yasuoka, K. Accelerating Molecular Dynamics Simulations on PlayStation 3 Platform Using Virtual-GRAPE

Programming Model. *SIAM J. Sci. Comput.* **2008**, *30* (6), 3108–3125. https://doi.org/10.1137/070692054.

(118) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5.

(119) Schroeder, D. V. Interactive Molecular Dynamics. *J. Phys* **2015**, *83* (3), 210–218. https://doi.org/10.1119/1.4901185.

(120) Grayson, P.; Tajkhorshid, E.; Schulten, K. Mechanisms of Selectivity in Channels and Enzymes Studied with Interactive Molecular Dynamics. *Biophys. J.* **2003**, *85* (1), 36–48. https://doi.org/10.1016/S0006-3495(03)74452-X.

(121) Dreher, M.; Piuzzi, M.; Turki, A.; Chavent, M.; Baaden, M.; Férey, N.; Limet, S.; Raffin, B.; Robert, S. Interactive Molecular Dynamics: Scaling up to Large Systems. In *Procedia Computer Science*; Elsevier B.V., 2013; Vol. 18, pp 20–29. https://doi.org/10.1016/j.procs.2013.05.165.

(122) Cassidy, K. C.; Šefčík, J.; Raghav, Y.; Chang, A.; Durrant, J. D. ProteinVR: Web-Based Molecular Visualization in Virtual Reality. *PLOS Comput. Biol.* **2020**, *16* (3), e1007747. https://doi.org/10.1371/journal.pcbi.1007747.

(123) Goddard, T. D.; Brilliant, A. A.; Skillman, T. L.; Vergenz, S.; Tyrwhitt-Drake, J.; Meng, E. C.; Ferrin, T. E. Molecular Visualization on the Holodeck. *Journal of Molecular Biology*. Academic Press October 2018, pp 3982–3996. https://doi.org/10.1016/j.jmb.2018.06.040.

(124) Anderson, A.; Weng, Z. VRDD: Applying Virtual Reality Visualization to Protein Docking and Design. *J. Mol. Graph. Model.* **1999**, *17* (3–4), 180–186. https://doi.org/10.1016/S1093-3263(99)00029-7.

(125) Juárez-Jiménez, J.; Tew, P.; o'connor, M.; Llabres, S.; Sage, R.; Glowacki, D.; Michel, J. A Virtual Reality Ensemble Molecular Dynamics Workflow to Study Complex Conformational Changes in Proteins. **2020**. https://doi.org/10.26434/CHEMRXIV.11833470.V2.

(126) Férey, N.; Nelson, J.; Martin, C.; Picinali, L.; Bouyer, G.; Tek, A.; Bourdot, P.; Burkhardt, J. M.; Katz, B. F. G.; Ammi, M.; et al. Multisensory VR Interaction for Protein-Docking in the CoRSAIRe Project. *Virtual Real.* **2009**, *13* (4), 273–293. https://doi.org/10.1007/s10055-009-0136-z.

(127) Doblack, B. N.; Allis, T.; Dávila, L. P. Novel 3D/VR Interactive Environment for MD Simulations, Visualization and Analysis. *J. Vis. Exp.* **2014**, No. 94.

https://doi.org/10.3791/51384.

(128) O'connor, M. B.; Bennie, S. J.; Deeks, H. M.; Jamieson-Binnie, A.; Jones, A. J.; Shannon, R. J.; Walters, R.; Mitchell, T. J.; Mulholland, A. J.; Glowacki, D. R. *Interactive Molecular Dynamics in Virtual Reality from Quantum Chemistry to Drug Binding: An Open-Source Multi-Person Framework*; American Institute of Physics Inc., 2019; Vol. 150. https://doi.org/10.1063/1.5092590.

(129) O'Connor, M. B.; Bennie, S. J.; Deeks, H. M.; Jamieson-Binnie, A.; Jones, A. J.; Shannon, R. J.; Walters, R. K.; Mitchell, T. J.; Mulholland, A. J.; Glowacki, D. R.; et al. Interactive Molecular Dynamics in Virtual Reality for Accurate Flexible Protein-Ligand Docking. *J. Chem. Phys.* **2020**, *150* (22). https://doi.org/10.1371/journal.pone.0228461.

(130) Bennie, S. J.; Ranaghan, K. E.; Deeks, H.; Goldsmith, H. E.; O'Connor, M. B.; Mulholland, A. J.; Glowacki, D. R. Teaching Enzyme Catalysis Using Interactive Molecular Dynamics in Virtual Reality. *J. Chem. Educ.* **2019**, *96* (11), 2488–2496. https://doi.org/10.1021/acs.jchemed.9b00181.

(131) Thomas, L. M.; Deeks, H. M.; Jones, A. J.; Metatla, O.; Glowacki, D. R. Somatic Practices for Understanding Real, Imagined, and Virtual Realities. **2019**.

(132) Amabilino, S.; Bratholm, L. A.; Bennie, S. J.; Vaucher, A. C.; Reiher, M.; Glowacki, D. R. Training Neural Nets to Learn Reactive Potential Energy Surfaces Using Interactive Quantum Chemistry in Virtual Reality. *J. Phys. Chem. A* **2019**, *123* (20), 4486–4499. https://doi.org/10.1021/acs.jpca.9b01006.

(133) Arbon, R. E.; Jones, A. J.; Bratholm, L. A.; Mitchell, T.; Glowacki, D. R. Sonifying Stochastic Walks on Biomolecular Energy Landscapes; International Community for Auditory Display, 2018; pp 232–239. https://doi.org/10.21785/icad2018.032.

(134) Zinovjev, K.; Tuñón, I. Reaction Coordinates and Transition States in Enzymatic Catalysis. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. Blackwell Publishing Inc. January 2018, p 1329. https://doi.org/10.1002/wcms.1329.

(135) Pérez de Alba Ortíz, A.; Vreede, J.; Ensing, B. The Adaptive Path Collective Variable: A Versatile Biasing Approach to Compute the Average Transition Path and Free Energy of Molecular Transitions. In *Methods in Molecular Biology*; Humana Press Inc., 2019; Vol. 2022, pp 255–290. https://doi.org/10.1007/978-1-4939-9608-7_11.

(136) Mandelli, D.; Hirshberg, B.; Parrinello, M. Metadynamics of Paths. *Phys. Rev. Lett.* **2020**, *125* (2), 026001. https://doi.org/10.1103/PhysRevLett.125.026001.

(137) Bešker, N.; Gervasio, F. L. Using Metadynamics and Path Collective Variables to Study Ligand Binding and Induced Conformational Transitions. *Methods Mol. Biol.* **2012**, *819*, 501–513. https://doi.org/10.1007/978-1-61779-465-0_29.

(138) Zinovjev, K.; Tuñón, I. Exploring Chemical Reactivity of Complex Systems with Path-Based Coordinates: Role of the Distance Metric. *J. Comput. Chem.* **2014**, *35* (23), 1672–1681. https://doi.org/10.1002/jcc.23673.

(139) Zinovjev, K.; Tuñón, I. Adaptive Finite Temperature String Method in Collective Variables. *J. Phys. Chem. A* **2017**, *121* (51), 9764–9772. https://doi.org/10.1021/acs.jpca.7b10842.

(140) Hare, S. R.; Bratholm, L. A.; Glowacki, D. R.; Carpenter, B. K. Low Dimensional Representations along Intrinsic Reaction Coordinates and Molecular Dynamics Trajectories Using Interatomic Distance Matrices. *Chem. Sci.* **2019**, *10* (43), 9954–9968. https://doi.org/10.1039/c9sc02742d.

(141) Jollife, I. T.; Cadima, J. Principal Component Analysis: A Review and Recent Developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. Royal Society of London April 2016. https://doi.org/10.1098/rsta.2015.0202.

(142) Yoo, C. K.; Shahlaei, M. The Applications of PCA in QSAR Studies: A Case Study on CCR5 Antagonists. *Chem. Biol. Drug Des.* **2018**, *91* (1), 137–152. https://doi.org/10.1111/cbdd.13064.

(143) Hemmateenejad, B.; Miri, R.; Elyasi, M. A Segmented Principal Component Analysis-Regression Approach to QSAR Study of Peptides. *J. Theor. Biol.* **2012**, *305*, 37–44. https://doi.org/10.1016/j.jtbi.2012.03.028.

(144) Liu, C.; Kelley, C. T.; Jakubikova, E. Molecular Dynamics Simulations on Relaxed Reduced-Dimensional Potential Energy Surfaces. *J. Phys. Chem. A* **2019**, *123* (21), 4543–4554. https://doi.org/10.1021/acs.jpca.9b02298.

(145) Hille, B. *Ion Channels of Excitable Membranes.*, 3rd Edition.; Sinauer Associates: Massachusetts , 2001.

(146) Mall, M. A.; Galietta, L. J. V. Targeting Ion Channels in Cystic Fibrosis. *J. Cyst. Fibros.* **2015**, *14* (5), 561–570. https://doi.org/10.1016/J.JCF.2015.06.002.

(147) Chen, X.; Xue, B.; Wang, J.; Liu, H.; Shi, L.; Xie, J. Potassium Channels: A Potential Therapeutic Target for Parkinson's Disease. *Neurosci. Bull.* **2018**, *34*

(2), 341–348. https://doi.org/10.1007/s12264-017-0177-3.

(148) Titulaer, M. J.; G M Verschuuren, J. J.; Titulaer, M. J.; Lang, B.; G M Verschuuren, J. J. *Lambert-Eaton Myasthenic Syndrome: From Clinical Characteristics to Therapeutic Strategies*; 2011; Vol. 10. https://doi.org/10.1016/S1474-4422(11)70245-9.

(149) Yurkin, S. T.; Wang, Z. Cell Membrane-Derived Nanoparticles: Emerging Clinical Opportunities for Targeted Drug Delivery. *Nanomedicine (Lond).* **2017**, *12* (16), 2007–2019. https://doi.org/10.2217/nnm-2017-0100.

(150) Sushnitha, M.; Evangelopoulos, M.; Tasciotti, E.; Taraballi, F. Cell Membrane-Based Biomimetic Nanoparticles and the Immune System: Immunomodulatory Interactions to Therapeutic Applications. *Front. Bioeng. Biotechnol.* **2020**, *8*, 627. https://doi.org/10.3389/fbioe.2020.00627.

(151) Chandana Epa, V.; Burden, F. R.; Tassa, C.; Weissleder, R.; Shaw, S.; Winkler, D. A. Modeling Biological Activities of Nanoparticles. *Nano Lett* **2012**, *12*, 49. https://doi.org/10.1021/nl303144k.

(152) Casalini, T.; Limongelli, V.; Schmutz, M.; Som, C.; Jordan, O.; Wick, P.; Borchard, G.; Perale, G. Molecular Modeling for Nanomaterial-Biology Interactions: Opportunities, Challenges, and Perspectives. *Front. Bioeng. Biotechnol.* **2019**, *7*, 268. https://doi.org/10.3389/fbioe.2019.00268.

(153) Almén, M. S.; Nordström, K. J. V; Fredriksson, R.; Schiöth, H. B. Mapping the Human Membrane Proteome: A Majority of the Human Membrane Proteins Can Be Classified According to Function and Evolutionary Origin. *BMC Biol.* **2009**, *7* (50). https://doi.org/doi: 10.1186/1741-7007-7-50.

(154) Kupfer, L.; Hinrichs, W.; Groschup, M. . Prion Protein Misfolding. *Curr. Mol. Med.* **2009**, *9* (7), 826. https://doi.org/10.2174/156652409789105543.

(155) Westergard, L.; Christensen, H. M.; Harris, D. A. The Cellular Prion Protein (PrPC): Its Physiological Function and Role in Disease. *Biochim. Biophys. Acta* **2007**, *1772* (6), 629. https://doi.org/10.1016/J.BBADIS.2007.02.011.

(156) Wulf, M.-A.; Senatore, A.; Aguzzi, A. The Biological Function of the Cellular Prion Protein: An Update. *BMC Biol. 2017 151* **2017**, *15* (1), 1–13. https://doi.org/10.1186/S12915-017-0375-5.

(157) Ma, J.; Wang, F. Prion Disease and the 'Protein-Only Hypothesis.' *Essays Biochem.* **2014**, *56* (1), 181. https://doi.org/10.1042/BSE0560181.

(158) Lua, R. C.; Grosberg, A. Y. Statistics of Knots, Geometry of Conformations, and

Evolution of Proteins. *PLoS Comput. Biol.* **2006**, *2* (5), e45. https://doi.org/10.1371/journal.pcbi.0020045.

(159) Jamroz, M.; Niemyska, W.; Rawdon, E. J.; Stasiak, A.; Millett, K. C.; Sułkowski, P.; Sulkowska, J. I. KnotProt: A Database of Proteins with Knots and Slipknots. *Nucleic Acids Res.* **2015**, *43* (D1), D306–D314. https://doi.org/10.1093/nar/gku1059.

(160) Faísca, P. F. N. Knotted Proteins: A Tangled Tale of Structural Biology. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 459–468. https://doi.org/10.1016/J.CSBJ.2015.08.003.

(161) Nureki, O.; Watanabe, K.; Fukai, S.; Ishii, R.; Endo, Y.; Hori, H.; Yokoyama, S. Deep Knot Structure for Construction of Active Site and Cofactor Binding Site of TRNA Modification Enzyme. *Structure* **2004**, *12* (4), 593–602. https://doi.org/10.1016/J.STR.2004.03.003.

(162) Alam, M. T.; Yamada, T.; Carlsson, U.; Ikai, A. The Importance of Being Knotted: Effects of the C-Terminal Knot Structure on Enzymatic and Mechanical Properties of Bovine Carbonic Anhydrase II. *FEBS Lett.* **2002**, *519* (1–3), 35–40. https://doi.org/10.1016/S0014-5793(02)02693-5.

(163) Shacham, T.; Sharma, N.; Lederkremer, G. Z. Protein Misfolding and ER Stress in Huntington's Disease. *Front. Mol. Biosci.* **2019**, *6*, 20. https://doi.org/10.3389/fmolb.2019.00020.

(164) Mroczko, B.; Groblewska, M.; Litman-Zawadzka, A. The Role of Protein Misfolding and Tau Oligomers (TauOs) in Alzheimer's Disease (AD). *Int. J. Mol. Sci.* **2019**, *20* (19). https://doi.org/10.3390/ijms20194661.

(165) Hammond, T. C.; Xing, X.; Wang, C.; Ma, D.; Nho, K.; Crane, P. K.; Elahi, F.; Ziegler, D. A.; Liang, G.; Cheng, Q.; et al. β-Amyloid and Tau Drive Early Alzheimer's Disease Decline While Glucose Hypometabolism Drives Late Decline. *Commun. Biol.* **2020**, *3* (1), 352. https://doi.org/10.1038/s42003-020-1079-x.

(166) Hallinan, G. I.; Vargas-Caballero, M.; West, J.; Deinhardt, K. Tau Misfolding Efficiently Propagates between Individual Intact Hippocampal Neurons. *J. Neurosci.* **2019**, *39* (48), 9623–9632. https://doi.org/10.1523/JNEUROSCI.1590-19.2019.

(167) Mapplebeck, S.; Booth, J.; Shalashilin, D. Simulation of Protein Pulling Dynamics on Second Time Scale with Boxed Molecular Dynamics. *J. Chem.*

Phys. **2021**, *155* (8), 085101. https://doi.org/10.1063/5.0059321.

(168) Booth, J.; Vazquez, S.; Martinez-Nunez, E.; Marks, A.; Rodgers, J.; Glowacki, D. R.; Shalashilin, D. V. Recent Applications of Boxed Molecular Dynamics: A Simple Multiscale Technique for Atomistic Simulations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. Royal Society August 2014. https://doi.org/10.1098/rsta.2013.0384.

(169) Lin, M. M.; Zewail, A. H. Hydrophobic Forces and the Length Limit of Foldable Protein Domains. *Proc. Natl. Acad. Sci.* **2012**, *109* (25), 9851–9856. https://doi.org/10.1073/PNAS.1207382109.