DEPARTMENT OF PHYSICS

Doctoral Thesis

gTPS: A machine learning and quantum computer-based algorithm for Transition Path Sampling

Supervisor
Prof. Pietro Faccioli

Candidate
Danial Ghamari

Cycle XXXVI

February 2024

Acknowledgement

First and foremost, I thank my parents and my sister (which will never be enough) for the love and care they have given me throughout my life. They are a huge part of where and who I am today.
Second, I am especially grateful to my supervisor, Pietro Faccioli, for his generosity and skill in guiding me in all the stages of my Ph.D. He has been the best teacher I could have asked for and learned from. I also thank the University of Trento for providing such a fruitful opportunity.
Finally, I thank Camilla for her presence in the last year of my studies which has given me strength and hope.

The following articles have been (or will be) published based on this thesis:

1. **G. D.**, Hauke P., Covino R., Faccioli P. (2022). Sampling rare conformational transitions with a quantum computer. In Scientific Reports (Vol. 12, Issue 1). Springer Science and Business Media LLC. 10.1038/s41598-022-20032-x

2. **G. D.**, Covino R., Faccioli P. (2023). Sampling a rare protein transition using quantum annealing.
   pre-print: arXiv:2311.15891
   The manuscript is submitted and currently under revision in the Journal of Chemical Theory and Computation.

3. **G. D.**, Covino R., Faccioli P. (??) All-atom simulation of protein unfolding on a quantum computer.

   The manuscript is being developed.

Contents

Introduction

Complex systems permeate nature, appearing in everything from microscopic life forms to the unpredictable dynamics of atmospheric phenomena [1, 2]. Characterized by the intricate interplay of numerous components, these systems often exhibit emergent features not easily predicted by the behavior of the individual parts. Among them, the "rare event" phenomenon stands out as a ubiquitous feature occurring with extremely small probability yet in very rapid motion [3]. However, they are not prerogative to systems with many parts. Indeed, one can also observe the same feature in the chemical reactions of a few atoms.

In the vast context of biological macromolecules, rare conformational rearrangements (prototypes of rare events) are responsible for many crucial physiological functions in the body of living organisms [4]. For example, they give rise to hemoglobin's modulation in affinity to Oxygen as the protein goes from the T state to the R state (or R$\rightarrow$T) [5]. In other "not-so-favorable" cases, such transitions lead to the misfolding of α-synuclein protein, instigating the formation of amyloid plaques and ultimately causing Parkinson's disease [6]. Obviously, meticulous and detailed characterization of the mechanism behind these rare transitions is not only theoretically fascinating but of practical importance, especially in advancing drug development and in the wake of the Covid-19 pandemic.

In the quest to characterize the complex structure and dynamics of bio-molecules, an *in vivo* study is arguably the most natural way to approach the problem. However, this approach essentially requires the whole living organism (of which the molecule is only a part) to seize its functioning while we probe different aspects of the molecular system. Conversely, one could resort to *in vitro* methods which provide more control by first isolating the molecule, e.g. using recombinant DNA techniques, and then exhaustively applying tools

such as NMR spectroscopy, to reveal different observables of the system in an equilibrium ensemble [7]. Unfortunately, such an approach is fundamentally bound to introduce artifacts and biases into the system as the cellular regulations and guidance on the molecule are eliminated. Most importantly, neither *in vivo* nor *in vitro* experimental methods provide the resolution to yield a complete characterization of structural changes with an atomic level of detail.

In contrast, *in silico* studies retain virtually unparalleled controllability over the system (the molecule) while they replicate (in principle) the entire dynamic of the cell in full atomistic details [8]. In this realm, the computational "microscope" of Molecular dynamics (MD) has been arguably at the forefront for the past 50 years. Originally developed in the context of condensed matter physics [9, 10], MD soon gained popularity in other areas such as material science and biophysics [11, 12]. During this time, MD has evolved into an indispensable tool in the computational studies of bio-molecular systems, owing to its capacity to track the evolution of every atom at time resolutions hard to match in experiments [13].

The framework of MD comprises a set of algorithms and techniques necessary to integrate the Newtonian equations of motion for each atom of the system. The all-atom forcefields involved, determined by empirical data and *ab initio* quantum mechanical (QM) calculations, ensure realistic and accurate simulation of the dynamics [14, 15], while due to the physical timescales imposed by the chemical bonds, the integration's timesteps are necessarily maintained on the order of $\sim$ 1-fs (femtoseconds). In principle, the latter would be advantageous in allowing us to access the full range of biomolecules' conformational transitions that primarily occur at timescales $\gtrsim$ 1-ps (picoseconds) [16]. However, in practice, these transitions involve a wide range of characteristic timescales, from bond length vibrations at $\sim$ 1-fs to the scale of large structural changes such as protein folding at 1-ms (milliseconds) and beyond. Thus, bridging this massive gap requires the simulations to track the dynamics for $\sim 10^{12} - 10^{15}$ iterations to replicate crucial rare transitions. In addition, the sheer number of calculations (e.g. to evaluate forces) involved in each step of integration, from $3N \sim 3 \times 10^4$ for small proteins to 3×10^6 for large complexes, incur a heavy computational load in terms of memory. This is not to mention statistically the most difficult challenge. In every *in silico* experiment, one needs to retrieve an ensemble of (long) molecular trajectories to characterize a transition/reaction and obtain the thermodynamic observables involved.

To address the computational load of simulating rare events, significant efforts have been dedicated to devising machines and platforms specifically designed to accelerate and also perform MD-based sampling [17]. One of the most influential platforms that has also achieved remarkable results is the Folding@Home project [18]. First developed in 2000, the primary focus of this distributed cloud computing platform has been to investigate the

dynamics of proteins, most importantly in the folding process, and to identify their role in various diseases subsequently. Two of the earliest nominal contributions of Folding@Home were the structural study of the N17 domain of the Huntingtin Protein (Huntington disease) and the investigation of the role of Aβ's 42-residue variant in Alzheimer's disease [19–21].

DE Shaw Research (DESRES), a privately held biochemistry research company, announced in 2008 the development of the first iteration of their supercomputer called Anton [22]. Anton-1 was specifically built to perform large-scale MD simulations for timescales of order milliseconds (ms). While Folding@Home relied on distributed short-lived simulations using their petaFLOPS[1] platform to achieve aggregated 1-ms of simulation time [23], thanks to its highly parallelized (and specialized) network of ASIC[2] systems, Anton was the first to break this crucial barrier using a single continuous simulation [24]. More importantly, Anton achieved this feat using an explicit solvent forcefield, unlike the implicit solvent utilized in Folding@Home's simulations.

In 2010, Anton published their result on the simulation for more than 1 ms of Bovine Pancreatic Trypsin Inhibitor's (BPTI) near-native state dynamics [25]. Additionally, the following year, they provided data on the folding process of 12 "fast folder" molecules, each with a folding time in the range of a few tens of microseconds (μs) or higher [26]. These results were shown to have substantial overlap with the experimental data that had been gathered at the time. Both the Folding@Home and Anton simulations remarkably proved that the existing forcefields are capable of identifying the native structure of proteins (at least in the case of ones with spatially modest size) starting from the unfolded configurations.

With Folding@Home's announcement of exceeding 1.5 exaFLOPs of cumulative computations in 2020, and Anton-3 becoming operational in the following year –reaching an impressive performance of 20μs simulation of 1-million atoms in less than 4 hours–, specialized computing platforms remain central in the computational studies of biomolecules [27]. However, performing ms-long unbiased MD simulations using all-atom forcefields, even with the development of GPU-accelerated MD software, is not yet an everyday routine. This limitation is especially true considering the modest computer clusters available to the average scientific project in this field. Moreover, even supercomputers have great difficulty performing $\sim$ 1-s of consecutive simulation for many atoms. It's worth noting that this is the scale where large proteins fold [28] and even quaternary structures, such as dimers, and oligomers [16], start to appear. As we await the development of more powerful computers, other means must therefore be developed to reduce the computational cost for *in silico* studies and opens the door for more accessible scientific discoveries.

[1] Floating points operations per second
[2] Application-specific integrated circuits

Going back to the 1970s, the advent of MD seemed promising to revolutionize macromolecular modeling, due to its (in principle) capabilities in simulating the entire equilibrium dynamics in fine atomistic details. Such possibilities seemed more tangible as the commercialized computers became more widespread. However, even by 1979 –when McCammon *et al.*, produced a staggering (at the time) 100-ps MD simulation of BPTI's native structure [12]–, it was clear that *in silico* experiments required more efficient alternatives than plain MD. As a consequence, in the past 40 years, a new category of computational methods, aptly named *enhanced sampling methods*, has emerged. These methods extend the utility of MD simulations by employing different levels of theory and approximations – with/out access to large computational resources– in the study of biomolecular systems, and especially in rare conformational transitions.

To give a few examples: Umbrella sampling [29] or Steered MD [30] introduce an unphysical force, either along a Collective Variable (CV) or on specific atoms, to "push" the system to the top of the free energy barrier. Alternatively, in the path-based sampling methods, the focus is to use MD to retrieve primarily trajectories that pertain to overcoming the large free energy regions [31]; thus, avoiding the uninteresting exponential time the system spends thermally fluctuating in the metastable basins. The enhanced sampling methods have greatly improved the applicability of MD in the context of molecular sampling. In some instances, they remain the only practically viable approach to obtaining information on complex conformational transitions. Efforts to map the free energy landscape (FEL) of B3 domain of protein G using experimental data and Metadyanmics[1] [32], and usage of Transition Path Sampling (TPS) to study the binding and unbinding pathways of a base pair in a CGC DNA oligomer [33], are to name a few applications.

Despite all of this, the cost of simulating medium to large-scale biological systems still remains insurmountable. The issue becomes more troublesome when focusing on the transitions between structurally distant metastable states, such as association/dissociation events or protein folding. It is worth mentioning that some of these challenges may be mitigated by utilizing coarse-grained MD [34]. However, this comes at the cost of losing some of the atomistic details that made computational approaches attractive in the first place compared to more physical "wet-lab" alternatives. In addition, defining and computing the effective forces between the degrees of freedom of coarse-grained models is both theoretically and computationally challenging. For example, until quite recently it was believed that the existing protein coarse-grained models are unable to predict native structures or even recognize them among a set of decoy structures [35, 36]. Therefore, the quest remains open to find computationally affordable and accurate methods for studying complex molecules and their transitions.

[1] A method similar to Umbrella sampling but with adaptive biasing force as the sampling progresses.

In the last two decades or so, Artificial Intelligence (AI) has revolutionized the field of computing and all the scientific endeavors that are involved in it [37–40]. Machine Learning (ML) tools have already made a significant impact on how we perform sampling, analyze the data, and even how to make predictions. One famous example in Biophysics and Biomolecular modeling in recent years is the AlphaFold's Deep Learning (DL) model that predicts the 3D folded structure of proteins from their amino acid sequences [41]. In other instances, ML or DL-based approaches are incorporated to: learn the forcefields from QM calculations or molecular modeling [42, 43], expediting the sampling and state detection with dimensionality reduction methods [44], and extracting thermodynamical information [45]. As more practical examples, these methods have also been utilized for drug development and discovery [46, 47]. With the rate that AI is "diffusing" through the scientific community on every level –theory, computational, and experiments– time can only tell what would be the limitations of future frameworks and schemes for studying bio-macromolecules.

On the other side of the computational methods spectrum over the last few years, quantum hardware has grown exponentially both in size and performance [48–50]. Even though the concept of quantum advantage has not yet been fully realized, it is not impossible to see its onset in the near future [51, 52]. This led us to contemplate whether quantum computing (in its current not-optimal state) could provide new opportunities for sampling complex molecular systems. Quantum computers (QCs) deal with quantum bits (qubits) instead of the classical binary variables (bits). The Qubits allow for a quantum superposition of 0 and 1 states in contrast to bits, hence, the amount of information that can be stored in an equal number of qubits would be exponentially larger. This feature, combined with quantum entanglements between the qubits, allows to achieve computations at a much faster rate with a QC than a classical computer (at least in principle). Therefore, perhaps it is timely to address the question: Can MD, ML, and quantum computing be integrated together to tackle outstanding problems in the sampling and simulation of complex biological molecular systems? This question is exactly what we have aimed to answer in this thesis.

Our main interest in this endeavor is to develop a method that can study thermally activated rare molecular transitions, yet, it has to remain agnostic to any *a priori* knowledge of the system under study. Therefore, it can remain as universal as possible without requiring a predefined application-specific CV. To this aim, by focusing our attention on path sampling methods and specifically the Transition Path Sampling (TPS) framework [53], we investigate the potential of introducing QC in this framework. TPS and its variants have traditionally faced difficulty in generating significantly different pathways at an acceptable rate when the associated rare events occur on a characteristic time-scale of $\gtrsim 1\mu$s. Consequently, the resulting MC's Markov chains is left with large auto-correlation values between the sampled transition pathways. By utilizing the inherent properties of QC in conjunction

with the enhancements of ML, we develop a novel TPS framework that directly tackles this auto-correlation problem.

The thesis is organized as follows: In the Chapter 1, we begin by briefly reviewing some of the previous approaches for enhancing molecular sampling. In particular, our focus would be on the advantage of these methods for sampling rare events compared to simple plain MD. The coarse-graining models and approaches are not discussed deliberately (except mentioning of few examples) as they require more extensive discussion and essentially do not fit with the notion of atomistic details we are after. By delving more into the detail of TPS framework, we discuss some of the variants of this approach and their challenges, in the context of studying rare transitions. We then give an overview of quantum computing and in specific quantum annealing [54–58]. This allows us to finally introduce our novel framework, graph Transition Path Sampling (gTPS), which tackles some of the challenges of conventional TPS.

The Chapter 2 focuses on the in-depth explanation of the gTPS framework. First, we explain how the iMapD algorithm [59] rely on ML's manifold learning techniques, guides the MD simulation to rapidly explore the configuration space of the system. We also mention and discuss the relevant challenges of this algorithm that might affect the efficiency of our framework. Then, we delve into the mathematical details of the rigorous approach we have adopted to coarse-grain the dynamics of the system using the iMapD's configurations. This theory is subsequently encoded into a network of transitions where the configurations act as the nodes and the edges carry the thermodynamics cost for the specific transition between them. The discrete ensemble of transition pathways is then implemented in the D-Wave's quantum annealing machine [60, 61] by utilizing a special mathematical formulation called the QUBO optimization [62, 63]. This serves us to finally lay out how the quantum annealers(QA) –integrated into an MC process– can be employed to generate and sample uncorrelated trial pathways representing rare transitions.

After discussing the details of gTPS framework, in Chapter 3, as a proof-of-concept, we apply it to a customary benchmark system called Alanine dipeptide. Even with its modest size, this molecule represents many of the amino acid residues in much larger and more complex proteins. Therefore, it is able to demonstrate, as a benchmark, the capabilities of our framework and validate our claims. In particular, after sampling the transition path ensemble with gTPS, we first illustrate how the pathways in the network, in conjunction with the coarse-grained representation, are able to correctly identify the low-energy regions of FEL. Then, by performing an auto-correlation analysis, we demonstrate how quantum computing is capable of generating pathways with a minimal correlation.

The successful application of Alanine dipeptide led us next to ponder whether current quantum computing machines are capable of tackling larger and biologically more relevant

molecules. To answer this question in Chapter 4, we first modify the iMapD algorithm by adding a new scheme called "Polar Star" shooting. This allows us to stabilize the overall gTPS framework in application to a relatively larger Bovine Pancreatic Trypsin Inhibitor molecule. Subsequently, by comparing the result of the modified version of iMapD and 1-ms plain MD trajectory (provided by Anton supercomputer), we demonstrate the efficiency of this approach in exploring the configuration space with 2 to 3 orders of magnitude lower computational cost. Finally in this chapter, samples of transition pathways are provided –utilizing the D-Wave annealer– which not only follow the regions of low free energy, but are shown to require a lower cost than alternative approaches such as simulated annealing.

The application of Bovine Pancreatic Trypsin Inhibitor elucidates the capabilities of the gTPS framework (and current quantum computing devices) in dealing with rare events that occur on the scale of milliseconds. The next natural course of action is to investigate the capability of our approach in sampling unfolding pathways for biological macromolecules. We expect that the unfolding process possesses less difficulty than folding due to the presence of a large entropic barrier for the latter. To this aim, in the Chapter 5, we exhibit the preliminary results of our ongoing investigations on two molecules: the Headpiece domain of Villin, and the reduced molecule of BPTI.

CHAPTER 1

Sampling Rare Transitions

In the past 40 years, various technologies have been developed to ease the task of sampling the configuration space of biomolecular systems and their rare conformational transitions. Apart from the specialized platforms for MD simulations (Anton and Folding@Home), introduction of GPU-acceleration into MD software was another piece that greatly sped up the computation [17]. Applications such as NAMD [64] (the first to introduce GPU-acceleration) and GROMACS [65, 66] are two famous examples that have become household names in the biophysics and biomolecular simulation community. In addition to this technological growth, the scientific effort has also been dedicated to developing frameworks for more accelerated molecular sampling. Such efforts were revolutionized once the potential of ML-based techniques for sampling gained more attention in the last two decades [40], the idea being that any dataset with complex underlying patterns can be deciphered with ML. Consequently, with the help of the right algorithm, it is believed that it is not only possible to extract these patterns but simultaneously predict certain features for producing new data. Clearly, the concept is very well fitted to the study of molecular systems, both for extracting information efficiently –e.g. correct calculation of thermodynamic observables and reaction rates– and improving the exploration of statistically important regions. Unfortunately, as discussed in the introduction, conformational transitions of medium to large biomolecules are computationally at the edge of what we can achieve with current methods. Therefore, novel ideas are still needed to improve on past technologies and help us break this barrier, hopefully in an accessible way that does not strongly rely on access to large computational power.

We want to dedicate this chapter to first briefly review some of the different enhanced

sampling methods that have been developed throughout the past years. Specifically, we divide these methods into two categories: CV-based and CV-free approaches. Subsequently, we mention methods that combine different enhancing concepts to potentially address the challenges in one another. Some of the advances and improvements that have been achieved using AI are also presented. Next, we turn our attention to a more in-depth discussion of the TPS framework. By presenting the applications of TPS and its variants, we arrive at the challenging issue of producing an ensemble of uncorrelated pathways for particularly complex rare transitions. To tackle this issue, we first set the stage by providing a brief introduction of quantum computing and in specific quantum annealing. Then, we introduce our novel approach for studying rare transitions that combines ML, and quantum annealing in order to tackle the autocorrelation problem of TPS. We postpone the in-detail presentation of this framework to the next chapter.

1.1 Enhanced sampling methods

The theoretical modeling of complex systems undergoing a rare transition is oftentimes illustrated as a point navigating a highly rugged and multidimensional FEL. The ruggedness manifests the large degrees of freedom inherent to these systems. Initially, the representing point is assumed to reside in a metastable state known as the so-called reactant state. Surrounded by free energy barriers larger than the thermal energy induced by the environment, the system continuously bounces around in this state. Occasionally, however, consecutive thermal fluctuations may align such that the system climbs over the lowest energy section of the barrier. At this juncture, there are two possible outcomes: either to turn back into the reactant state or to move along the region of high energy, both with equal probability. If the system decides to move forward, after a small series of fluctuations, it may eventually arrive at a new metastable state, called the Product state [67](see Figure 1.1).

Given this model, the purpose of MD is to replicate the positions and velocities of a complex system of interacting atoms as they evolve on this FEL. If we could somehow ensure that the simulations were ergodic, then statistical mechanics would allow us to replace ensemble averaging with time averaging in order to derive thermodynamic quantities. However, the existence of free energy barriers between two well-separated macrostates greatly hampers the attempts to reach ergodicity in the simulations. As alluded to in the introduction, especially in simulating rare transitions of large complex molecules, ergodicity necessitates computational resources that are almost always not available in *in silico* studies. For example, even with the help of supercomputers like Anton to provide ms-long MD simulations, the chances are that we leave many relevant regions of configuration space unexplored.

Due to the free energy barriers, a gap exists in the temporal scales of the dynamics that separates the thermal oscillations in metastable states from the in-frequent transitions between them. Therefore, when simulating with algorithms such as MD, we have to wait a long time in order for them to sample an ensemble of trajectories corresponding to such complex rare events. However, from a theoretical standpoint, the presence of a temporal gap also implies that a low-dimensional set of CVs exists as functions of the configuration space:

$$\mathbf{Y}(\mathbf{Q}) = f(x_1, x_2, \ldots, x_{3N}) \tag{1.1.1}$$

which are capable of capturing the slow dynamics of the system. Here, $\mathbf{Q}$ is a representative point in the $3N$ dimensional configuration space. These variables act as coordinates for a low-dimensional embedding where we can "perfectly" (at least in principle) map metastable states and the free energy barriers in between. More importantly, they allow us to gauge the system's time evolution for any reaction between these states. As a consequence, if we somehow managed to obtain *a priori* the appropriate "Reaction coordinates" (RCs), then mapping of the free energy in different regions of this low-dimensional hypersurface could be done relatively efficient [68]. Here, we must note that since our primary focus in this thesis is to study rare transitions in (bio)molecular systems, we interchangeably use both terms, CV and RC, as the corresponding coordinates of a reaction.

Due to such utility, it is unsurprising that many methods either incorporate a biasing potential in terms of CVs to accelerate the sampling or use these functions *a posteriori* to extract information from molecular trajectories –e.g. free energy calculations or reaction rates. In fact, for the latter case, it would be significantly hard to characterize the thermodynamics or kinetics of the system without compacting the $3N$-dimension configuration space into a simpler low-dimensional one. In the following, we first focus on the methods specially developed to accelerate the sampling along a chosen CV. Then, we discuss other types of enhancing methods that strictly need no prior knowledge of the system.

1.1.1 CV-based enhanced sampling

One of the earliest approach in the context of enhancing MD-based sampling has been the method of Umbrella Sampling. This method was initially developed in 1977 by Torrie and Valleau in the context of *importance sampling* to study various phases of a Leonard-Jones fluid [29, 69]. In the Umbrella Sampling method, we first divide the region in between the reactant and product states along a selected CV, $\mathbf{Y}(\mathbf{Q})$, into windows. Here, we consider a 1D case as illustrated in the Figure 1.2. Next, for each window i an individual harmonic

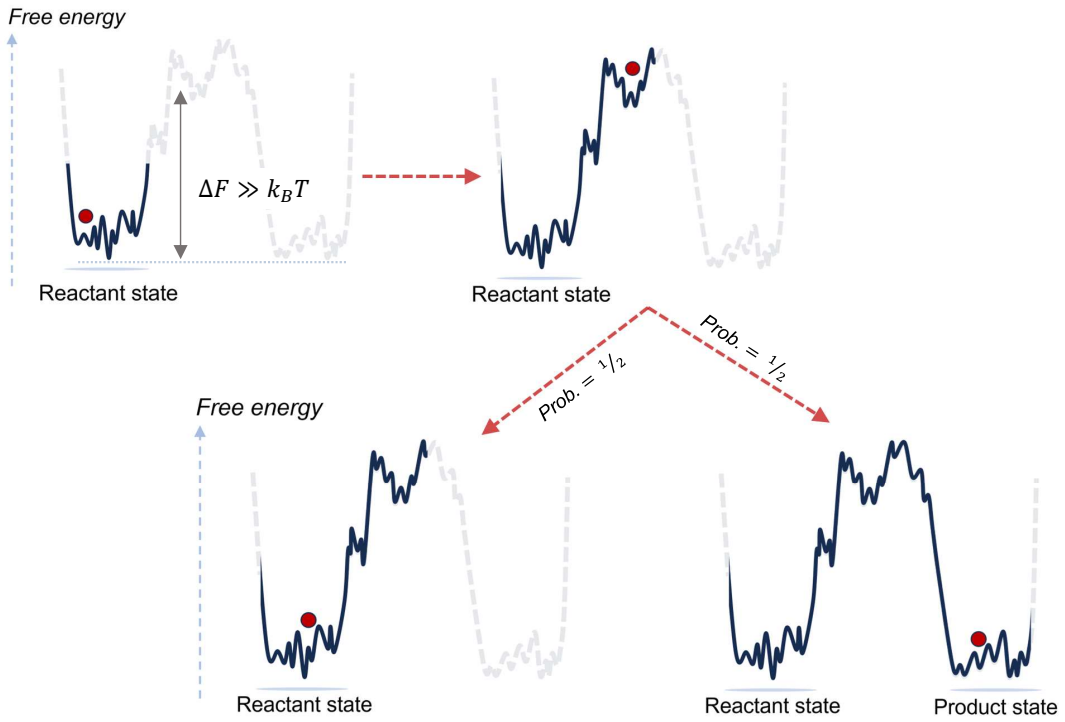


Figure 1.1: Schematic representation of a rare event in a rugged FEL. After spending an exponential time in the reactant state, the thermal fluctuations of the system might lead to overcoming the lowest section of the free energy barrier surrounding it. On the top of the barrier, the system has an equal probability of returning to the reactant state or arriving at a new product state.

potential with the spring constant K is added to the system's energy function:

$$\begin{aligned} V_{\text{new}}(\mathbf{Q}) &= U(\mathbf{Q}) + U^i_{\text{bias}}(\mathbf{Y}(\mathbf{Q})) \\ \textit{where} \quad U^i_{\text{bias}}(\mathbf{Y}(\mathbf{Q})) &= \frac{K}{2}\left(\mathbf{Y}(\mathbf{Q}) - \mathbf{Y}_i\right)^2 \end{aligned} \tag{1.1.2}$$

Here, $\mathbf{Y}_i$ denotes the value of $\mathbf{Y}(\mathbf{Q})$ at the center of i-th window, and $U(\mathbf{Q})$ is the potential energy of the system in equilibrium. We proceed by running an MD simulation for every window using its the corresponding biased potential in Equation (1.1.2). These simulations can be performed in parallel, which is a strength of Umbrella Sampling. In general, the unbiased probability of the system for every value of $\mathbf{Y}$ is given by

$$P(\mathbf{Y}) = \frac{\int d\mathbf{Q}\, e^{-\beta V(\mathbf{Q})}\, \delta(\mathbf{Y}'(\mathbf{Q}) - \mathbf{Y})}{\int d\mathbf{Q}\, e^{-\beta V(\mathbf{Q})}} \tag{1.1.3}$$

where $\beta = 1/k_{\text{B}}T$, $d\mathbf{Q} = dx_1, dx_2, \ldots, dx_{3N}$, and $V(\mathbf{Q}) = U(\mathbf{Q})$ denotes the original energy function of system. If we add a bias energy function to the whole system like in

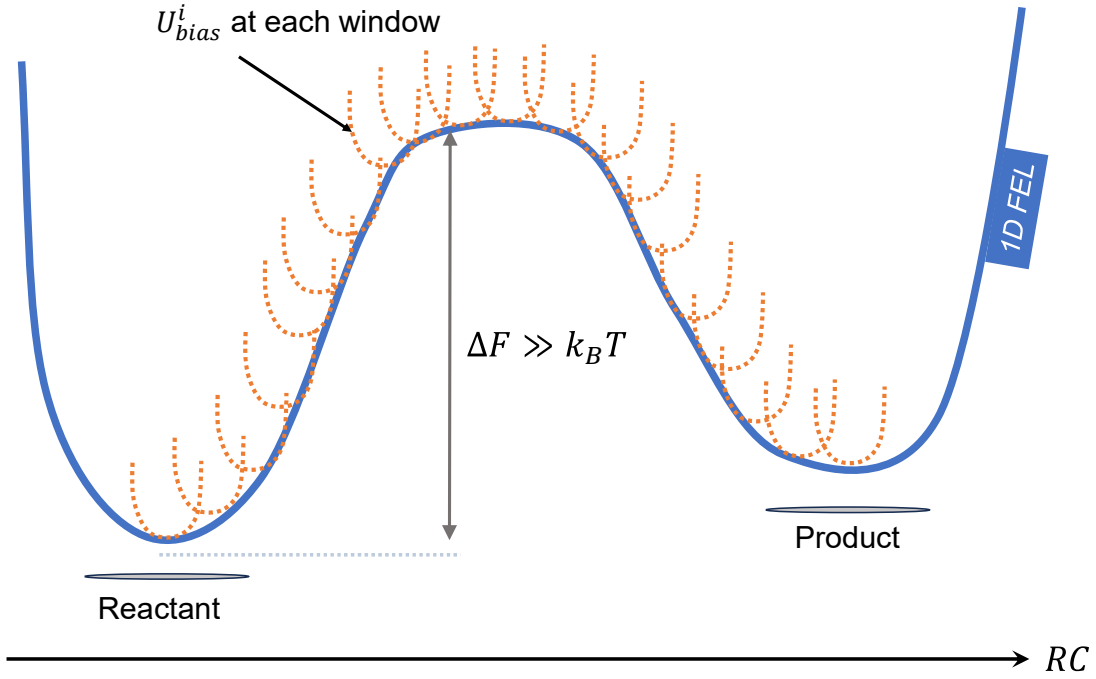


Figure 1.2: Schematic representation of Umbrella sampling algorithm. For every bias U^i_{bias} of window i along the RC $Y(\mathbf{Q})$, the system is simulated according to the Equation (1.1.2). In the end, by using Weighted Histogram Analysis, we calculate the free energy needed for the reaction.

Equation (1.1.2), then the new probability is evaluated as

$$P_{\text{new}}(\mathbf{Y}) = \frac{\int d\mathbf{Q}\, e^{-\beta V_{\text{new}}(\mathbf{Q})}\, \delta(\mathbf{Y}'(\mathbf{Q}) - \mathbf{Y})}{\int d\mathbf{Q}\, e^{-\beta V_{\text{new}}(\mathbf{Q})}} \tag{1.1.4}$$

After a little mathematical manipulation, we arrive at the identity

$$P_{\text{new}}(\mathbf{Y}) = P(\mathbf{Y}) \times e^{-\beta U^i_{\text{bias}}(\mathbf{Y})} \times \langle e^{-\beta U^i_{\text{bias}}(\mathbf{Y})} \rangle \tag{1.1.5}$$

where

$$\langle e^{-\beta U^i_{\text{bias}}(\mathbf{Y})} \rangle = \frac{\int d\mathbf{Q}\, e^{-\beta V(\mathbf{Q})}\, e^{-\beta U^i_{\text{bias}}(\mathbf{Y}(\mathbf{Q}))}}{\int d\mathbf{Q}\, e^{-\beta V(\mathbf{Q})}} \tag{1.1.6}$$

Following the Equation (1.1.5), the free energy, $F = (-1/k_{\text{B}}T) \ln P$, in an unbiased setting is evaluated from the biased dynamics as:

$$F_{\text{unbiased}}(\mathbf{Y}) = F_{\text{bias}}(\mathbf{Y}) - U^i_{\text{bias}}(\mathbf{Y}) - \beta \ln \langle e^{-\beta U^i_{\text{bias}}(\mathbf{Y})} \rangle \tag{1.1.7}$$

The Equation (1.1.7) provides the theoretical ground to approximate the free energy of a transition between two metstable states, once the MD simulations in the Umbrella Sampling

are finished. In practice, the most popular way is the Weighted-Histogram Analysis Method (or its modern variants) [70, 71] to combine the statistics of these independent samplings in evaluating the last term in Equation (1.1.7). The successful applications of Umbrella Sampling can be observed in studies of: the stability of 42-residue variant of Aβ in the fibril phase [72]; base-pair opening in the double helix of B-DNA [73]; protein-ligand binding affinity in the FKBP protein bounding with small molecules 4-Hydroxy-2-Butanone[1] and Tacrolimus (FK-506) [74]; the association/dissociation behavior of Glycophorin A embedded in a lipid membrane using CG simulations [75].

Apart from Umbrella Sampling, two other widely adopted methods that accelerate the sampling by applying a biasing force are SteeredMD and Metadynamics (MetaD). SteeredMD forces the system to evolve away from its initial equilibrium condition using a pulling force in the direction of one or multiple CVs. This inspiration comes from single-molecule pulling experiments in atomic force microscopy [76]. The pulling force in this method is very similar to Umbrella Sampling in being a harmonic function:

$$U_{\text{bias}}^{SMD}(\mathbf{Y}(\mathbf{Q}(t))) = \frac{K}{2}\left(\mathbf{Y}(\mathbf{Q}(t)) - vt - \mathbf{Y}(\mathbf{Q}_i)\right)^2 \tag{1.1.8}$$

where v is the velocity of pulling and $\mathbf{Y}(\mathbf{Q}_i)$ is the value of $\mathbf{Y}$ in the initial state. One interesting feature of SteeredMD is that it can be used as a complementary method for other methods, such as MetaD or Umbrella Sampling, in providing the initial and final conformations. Furthermore, due to the affinity of the method to pulling experiments, Jarzynski or Crooks relations can be utilized to calculate the free energy along any two-state transition [77–79].

SteeredMD has, in particular, provided a reliable method in the calculation of binding energy in *in silico* drug design [80]. To name two successful applications of this method in this regard, we can point to the study of the unbinding affinity of Cyclin-dependent kinase 5 enzyme from inhibitors, an important target for the medicinal chemistry [81], and ligand escape pathways in Acylaminoacyl-peptidase [82].

MetaD, thanks to its adaptive nature, provides a more robust approach for applying the bias in the sampling compared to Umbrella Sampling and SteeredMD. Developed by Laio and Parinello in 2002 [83], during a MetaD simulation, a Gaussian history-dependent potential is applied to the system at a given rate along n-chosen CVs:

$$U_{\text{bias}}^{MetaD}[\mathbf{Y}(\mathbf{Q}(t))] = \sum_{\lambda\tau<t} A(\lambda\tau) e^{\sum_{i=1}^{n} \frac{-\left[\mathbf{Y}_i(\mathbf{Q}(t)) - \mathbf{Y}_i(\mathbf{Q}(\lambda\tau))\right]^2}{2\sigma_i^2}} \tag{1.1.9}$$

In this expression, $\lambda\tau$ is the simulation time of every Gaussian deposited previously at rate τ up to t, $A(\lambda\tau)$ denotes the height of these Gaussians, $\mathbf{Y}_i(\mathbf{Q}(\lambda\tau))$ is the value of i-th

[1] An important intermediate for vitamin A and fragrances

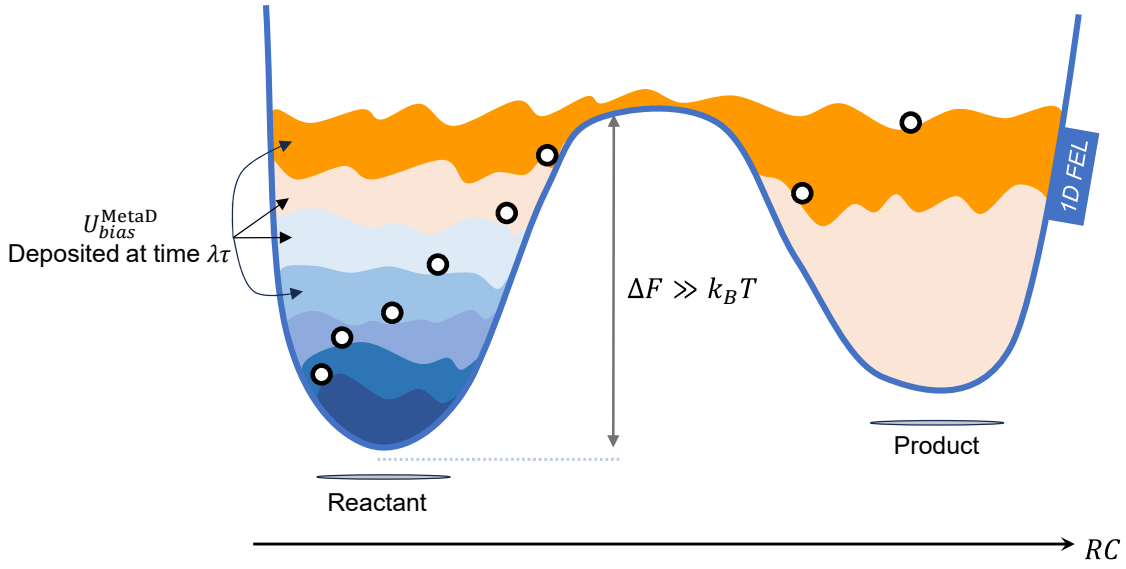


Figure 1.3: Schematic representation of MetaD sampling. At the rate of $\lambda\tau$, Gaussian biases along the RC $Y(\mathbf{Q})$ are deposited to the system. The cumulative sum of biases, Equation (1.1.9), provide the energy for the system (the white circle) to overcome the barrier in a rare reaction.

CV at time $\lambda\tau$, and σ_i is standard deviation of $\mathbf{Y}$. All the τ,λ, and σ_i should be determined before the sampling. The procedure of MetaD follows the illustration depicted in Figure 1.3: Assuming a two-state system located in the reactant state, the deposited Gaussians energetically fill the reactant basin and promote exploration of previously forbidden configurations. Once the height of these Gaussians reaches the free energy barrier at the transition state, the system "rolls" into the product state. The deposition continues until the product basin is also filled. Finally, the difference in potentials added in each state determines the free energy profile of such transition. In the past 20 years, MetaD has been quite influential in the molecular sampling [84, 85]. Applications such as the identification of an intermediate in the folding of the X domain of phosphoprotein of measles virus [86], and the study of ligand unbinding mechanisms in host-guest systems [87] are two (fairly) recent examples.

Even though the CV-based approaches hold a special allure for enhancing the sampling of molecular configurations, the caveat is that a universal and practical definition of such coordinates still eludes us. Often, one relies on physical and chemical intuition to "guess" the optimal CV for the reaction under study. This may be easily achieved for relatively simple systems, such as minimally frustrated proteins, using structural functions like RMSD. However, in more complex systems and reactions, it is highly nontrivial to recognize the correct RC without performing exhaustive sampling using unbiased MD or an enhanced sampling method that does not rely on CV definition. In such cases, the incorrect defini-

tion of the RC may lead to sampling even more inefficient than plain MD or inaccuracy in the *a posteriori* analysis. In addition, while a certain guess of CV may well characterize a specific reaction, another transition could be poorly described with the same coordinate [68, 84, 88].

1.1.2 CV-free enhanced sampling

Recplica Exchange Molecular Dynamics: The class of methods that do not rely on the identification of CV to enhance the sampling contains approaches with different flavors. One such approach is the idea of directly manipulating the temperature to facilitate overcoming free energy barriers and exploration of previously forbidden regions. Arguably, the most well known method that employs this idea is the Replica Exchange MD (RMED) or parallel tempering [89]. In RMED, several instances (or replicas) of the system are simulated in parallel but each at different temperature. Considering two of these simulations (i, j) which are closest in temperature, at regular intervals, we exchange either their temperature or the instantaneous configurations based on a Metropolis selection criteria:

$$P(i \leftrightarrow j) = \min\left\{1, \exp\left(E_i - E_j\right)\left(\frac{1}{k_{\mathrm{B}}T_i} - \frac{1}{k_{\mathrm{B}}T_j}\right)\right\} \tag{1.1.10}$$

where, E is the total energy of the system. RMED has been used to study: The influence of disulfide bridges in the fibril formation of Human Amylin [90, 91], dimer association/dissociation rates [92], and folding of proteins [93–95].

Accelerated Molecular Dynamics: Another enhanced method that directly targets the internal parameters of the simulations for better sampling is the Accelerated MD (AMD). In AMD –developed by McCammon and coworkers [96]– a boost potential is added to the system's dynamics such that it activates every time the system's potential energy falls below a certain threshold $E_{\text{thresh.}}$:

$$\begin{aligned} V_{\text{new}}(\mathbf{Q}) &= U(\mathbf{Q}) + U_{\text{boost}}(\mathbf{Q}) \\ U_{\text{boost}}(\mathbf{Q}) &= \begin{cases} 0 & \text{if } U(\mathbf{Q}) \geq E_{\text{thresh.}} \\ \frac{(E_{\text{thresh.}} - U(\mathbf{Q}))^2}{\alpha + E_{\text{thresh.}} - U(\mathbf{Q})} & \text{if } U(\mathbf{Q}) < E_{\text{thresh.}} \end{cases} \end{aligned} \tag{1.1.11}$$

where α controls the flatness of the boost potential as seen in Figure 1.4. By the end of the AMD simulation, the equilibrium ensemble average of the system can be recovered by reweighting according to the Boltzmann factor $e^{\beta U_{\text{boost}}}$. With this reweighting the average of an observable is written as

$$\langle A \rangle = \frac{\langle A e^{\beta U_{\text{boost}}} \rangle}{\langle e^{\beta U_{\text{boost}}} \rangle} \tag{1.1.12}$$

Applications of AMD has been seen in folding of Trpzip2 (a 13 residue protein) by Yang

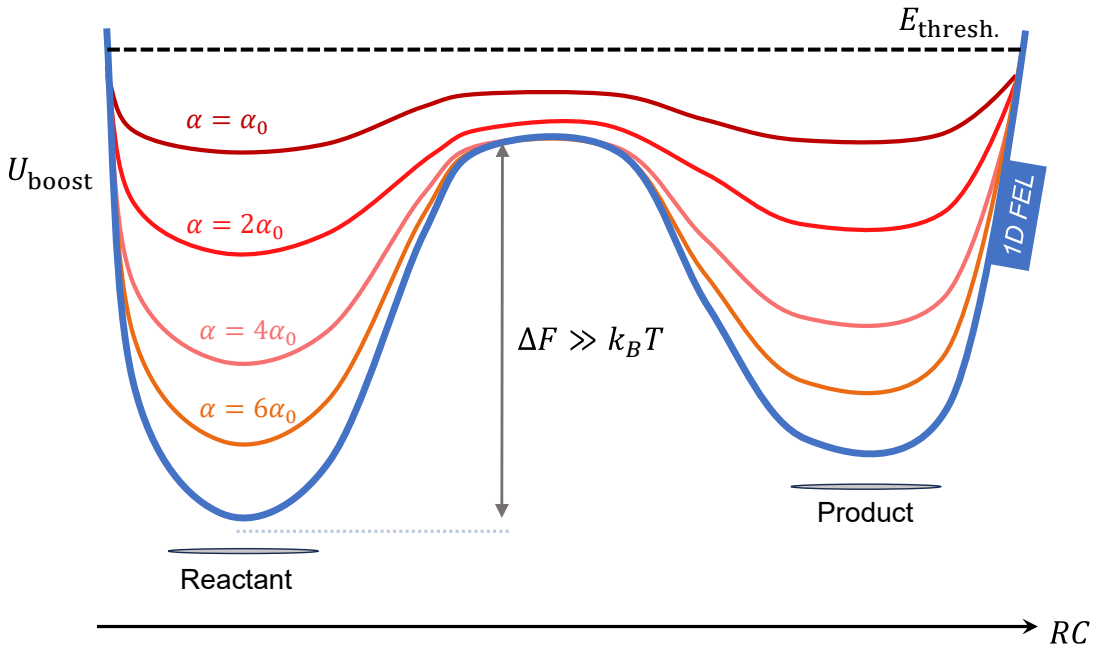


Figure 1.4: Schematic representation of AMD sampling. In this algorithm, a boosting potential is activated every time the energy of the system falls below a threshold, Equation (1.1.11).

et al. [97], and four other fast folding proteins by McCammon and coworkers [98]. Furthermore, it has recently been applied to the study of ligand and peptide binding to proteins [99, 100] and recently one of its newer variants (called Guassian AMD) was used to study conformational changes in DNA complexes [101, 102].

State-based enhancing methods: The presence of a free energy barrier divides the configuration space into two metastable states where the majority of MD-sampled configurations accumulate. In a highly frustrated system, the number of such long-lived states generally scales with the system size, albeit not necessarily monotonically [103]. In such systems, methods that rely on parallel and short simulations in different states have proven more efficient than samplings using a very long simulation. To name some of the well-known approaches that utilize this concept, we can mention Markov State Models (MSMs), Milestoning, and the String method. However, since the last two also require CV definition, we do not discuss these methods and only focus on MSMs here [68].

MSM is a powerful statistical framework that starts from an initial set of molecular configurations, e.g. obtained from a previous long MD simulation. It then builds a set of "microstates" by performing a clustering on the initial dataset. This clustering can be established by first lowering the dimension of the input data using methods such as time-lagged Independent Component Analysis [104, 105] or Principal Component Analysis

(PCA) [106], and then using, for example, KMeans [107]. Once the clustering is obtained, a Markovian transition matrix is built between the microstates, either by running new sets of MD simulations between them or utilizing the original long MD trajectory. One can next coarse-grain the transition matrix by some form of spectral clustering, and obtain a coarser description of states (macrostates), that would correspond to the slow dynamics of the system. Finally, these macrostates can be further utilized to run a new set of simulations that can improve the existing (coarse-grained) transition matrix or lead to the identification of new states.

The most famous applications of MSMs to molecular simulations to date have arguably been reported by the Folding@Home project, which intrinsically relies on the ability to perform many short individual MD simulations for the sampling. Before Folding@Home gained prominence, however, it was Luty and McCammon who first showed that it is possible to accelerate the sampling by studying an enzymatic reaction with a small MSM [108]. Besides the Folding@Home applications of MSMs, of which we only mentioned a few in the introduction, this method has also been recently utilized to study the activation of the G protein-coupled μ-opioid receptor [109], and protein-protein association/dissociation between the Ribonuclease Barnase and its inhibitor Barstar [110].

Path-based enhancing methods: As the final general approach to enhance the sampling of configuration space, the path sampling methods rely on retrieving only the transition pathways in order to increase the computational efficiency while studying rare events. Arguably, the most famous path-based method in the last 20 years has been the TPS framework and its other variants such as *Transition Interface Sampling* or *Forward Flux Sampling* methods. We will discuss in more detail the TPS framework in the next section. Apart from TPS, another notable framework in the context of path-based sampling has been the *Discrete Path Sampling* (DPS) [111, 112], which is akin to both TPS and MSM. In DPS, a double-ended interpolation is performed between any reactant and product states such that it finds stationary points of the potential energy. Then, using the doubly-nudged elastic band framework we find the optimal pathways that connect these points. Finally, utilizing the theory of DPS we can retrieve both the thermodynamic and kinetic of the transition. Using DPS, it is also possible to perform one-ended searches in the configuration space to locate new intermediate states and potentially new metastable states.

1.1.3 Hybrid methods and addition of Machine-Learning

Another common theme in enhanced sampling methods has been to combine the approaches introduced above, in order to potentially address the challenges faced in certain methods. One notable example which has received more attention is the MetaD. Motivated by the idea behind REMD, parallel-tempering MetaD (PT-MetaD) was introduced by Parrinello

and coworkers [113]. PT-MetaD initiates different replicas in order to facilitate easier barrier crossing. In the same paper, PT-MetaD was demonstrated to be more efficient in the folding of β-hairpin than MetaD. Later, both bias potential exchange [114] and CV parallel tempering [115] were introduced to address the challenge of reactions occurring along multiple distinct CVs. In the former, one consecutively exchanges a replica's bias potential that is along one variable, with another potential along a different CV. However, in parallel tempering of CV, each replica simultaneously samples in the direction of multiple variables, and we exchange the replica's potentials. The idea of parallel tempering has also been considered for Umbrella Sampling [116]. In a recent application to protein-protein and lipid-protein interactions in membrane systems [117], this method was shown to have a better convergence rate than normal Umbrella Sampling. Apart from replica exchange, hybrid methods that combined MetaD with SteeredMD [118] or Umbrella Sampling [119–121] have also been introduced. In regard to the latter combination, the more common idea is to utilize MetaD for improving the sampling of different regions in configuration space. Then, use the Umbrella Sampling to evaluate the free energy along the CVs of MetaD.

We now turn our attention to the ML-based additions to enhanced sampling methods. In the study of rare transitions, any physical or computational knowledge available *a priori* can be used to facilitate the sampling. However, once we encounter a new system, even the definition of reactant and product states may be difficult to provide in order to utilize system-agnostic methods e.g. TPS [68]. Considering this challenge, techniques such as dimensionality reduction of ML are particularly useful to extract from an initial sparse dataset the maximum amount of statistical information possible. In this context, we have already mentioned the usage of PCA and tICA in MSMs for facilitating microstate detection. PCA is a linear dimensionality reduction method whose first component characterizes the direction of the original space with the most variance. PCA was first utilized for molecular data to estimate protein configurational entropy by characterizing the anharmonicity of collective motions [68, 122, 123]. Obviously, linearity of PCA is limited when applied to a molecular dataset, and to obtain a better approximation of the CVs more sophisticated approaches must be adopted. In this regard, nonlinear methods like Kernel PCA [124], and Iso Maps [125] have proven useful [126].

In more recent years and with the increase in applicability of GPU-accelerated computations, nonlinear DL-based methods such as autoencoders [127–131] have seen a rise in popularity. Autoencoders are models that, using a Deep Neural Network (DNN), first encode the configurational data into a low-dimensional latent space and then decode through a separate DNN back into molecular ambient space. The autoencoder attempts to minimize the error between the synthetic configuration vector of the second DNN and the original vector. An example of such approaches is the Variational Autoencoder architecture [132, 133]. In the context of using ML/DL methods for sampling configurations, generative DNN mod-

els such as Generative Adversarial Networks [134] and flow-based generative models [135] have received particular attention. Examples of these two approaches can be seen in [136] and [137] (as a Boltzmann Generator) respectively.

1.2 Transition Path Sampling

In rare events, the system attempts to overcome large free energy barriers that can either involve high potential energy or high entropic bottlenecks. A convenient strategy to sample these events is to search for dynamical bottlenecks through which the system passes as it transitions from reactant to product. If the barriers are energetic in nature, then the bottleneck would be the saddle-points of the potential energy. These points, which comprise the ensemble of the transition state, can henceforth be utilized in Transition State Theory (TST) –or its variants– to reveal the reaction coordinate and rate of transition. Conversely, in complex systems with rugged and highly non-linear potential energy landscapes, this view does not hold for a sufficient description of the rare transitions. Here, the entropic barrier involves many points where some may be stationary in potential while most are not. Therefore, by simply searching for the points with zero Laplacian of potential function, we cannot identify the transition macrostate.

Luckily, Transition Path Theory –a natural successor to Transition State Theory– has been developed to provide the correct mathematical description of rare transitions in such cases. To this aim, TPT first introduces the notion of the committor function $q(x)$ for all the points in the transition region. Aptly named, this function signifies the *commitment* probability of any trajectory that has passed through x, to be reactive. Consequently, $q(x)$ is considered to be one of the pivotal concepts in TPT and the optimal CV for any reaction or transition [138]. Furthermore, utilizing the committor, TPT provides the formulation of the probability density and current of reactive pathways in the transition region of the configuration space. Despite the capabilities of TPT, the question still remains: How can one provide these reactive trajectories as the system undergoes a rare event, in order to then utilize TPT?

To answer this question, Transition Path Sampling (TPS) provides a computationally sound approach that samples directly from the Transition Path Ensemble (TPE) available in a complex rare event. The key enabling feature of TPS is to perform a random walk with the help of the Monte Carlo (MC) algorithm, not in the configuration space but rather in the trajectory space. Then, by focusing the MD simulation on retrieving only the reactive trajectories, TPS samples the TPE ensemble without wasting simulation time for oscillatory dynamics in metastable states. Remarkably, this enhanced sampling approach requires no prior definition of CVs or unphysical forces. Only an unambiguous definition of the reactant

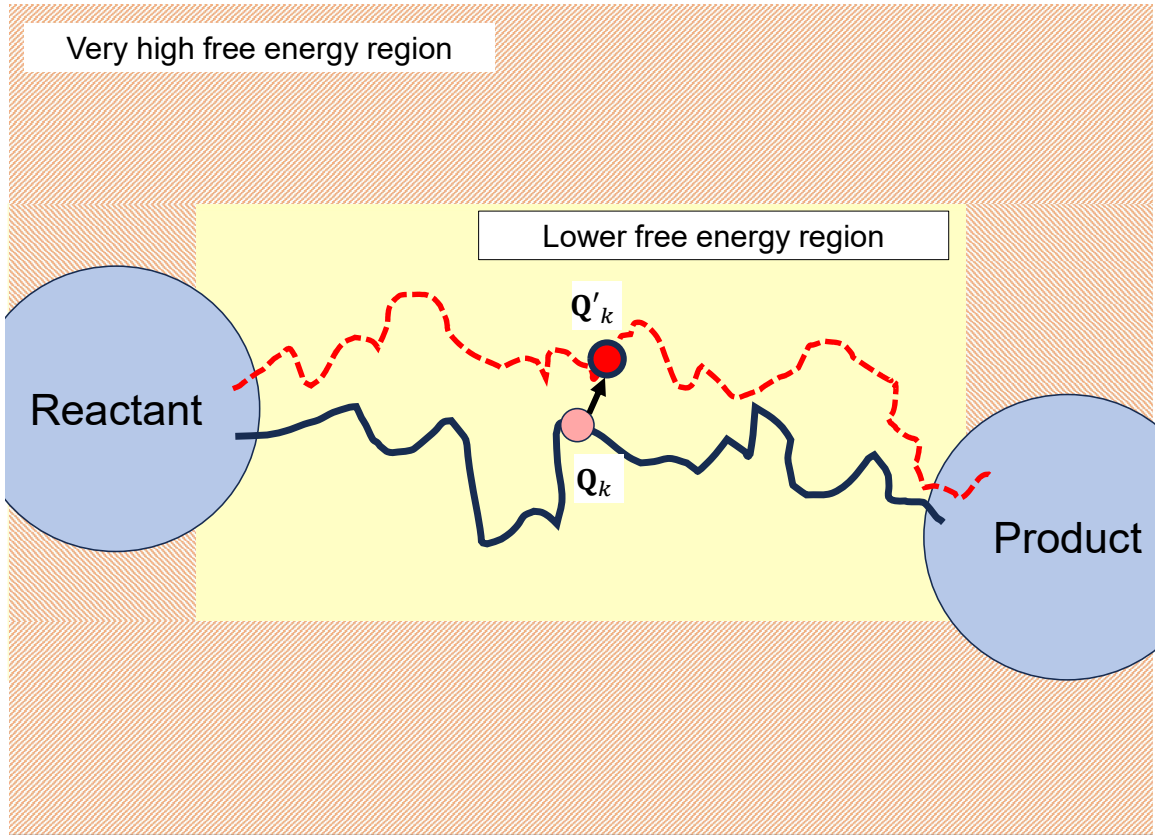


Figure 1.5: Schematic representation of the TPS framework. Starting from a previously retrieved transition path (the black line), we choose a configuration $\mathbf{Q}_k$ randomly in the reactive portion of the path. By perturbing $\mathbf{Q}_k$ in the shooting move and obtaining $\mathbf{Q}'_k$, we initiate two new MDs, one forward and one backward in time (the red dashed line). The combined trajectory of these two is accepted or rejected as new transition path based on the Metropolis criterion in the TPS's MC process.

and product state of the transition in the configuration space is needed. In the following, we briefly present the steps of TPS and some of the kinetic information that it provides. We also provide a (non-comprehensive) review of the advancements that have been achieved in the past 20 years in both the original algorithm and its more recent variants.

Transition Path Ensemble

The Figure 1.5 depicts a possible reactive/transition pathway between schematic reactant and product states: A trajectory that strictly arrives at the product state having been started in the reactant. We denote the equilibrium probability density of any pathway (not only reactive one) in the configuration space that takes the time T, as the $P[\mathcal{X}_T]$. It is convenient to discretize the trajectory into n time slices:

$$\mathcal{X}_T = \{\mathbf{Q}_0, \mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_n\} \tag{1.2.1}$$

where $T = n\delta t$ and, $\mathbf{Q}_k$ is the microstates of the system occupied at time $t = k\delta t$ along the path. Assuming Markovian dynamics, this probability can be decomposed as:

$$P[\mathcal{X}_T] = \left(\prod_{k=1}^{n} p[\mathbf{Q}_{k-1} \to \mathbf{Q}_k] \right) \rho(\mathbf{Q}_0) \tag{1.2.2}$$

where $\rho(\mathbf{Q}_0)$ is the probability of the system initially residing in the microstate $\mathbf{Q}_0$.

To focus only on the reactive trajectories, we follow the expression

$$P_{RP}[\mathcal{X}_T] = h_R[\mathbf{Q}_0] h_P[\mathcal{X}_T] P[\mathcal{X}_T] / \mathcal{Z}_{RP} \tag{1.2.3}$$

where $h_R(\mathbf{Q}_0)$ and $h_P(\mathbf{Q}_T)$ are the characteristic functions of Ω that determine if x is in reactant or product respectively. This is achieved by

$$h_{\mathcal{M}}(x) = \begin{cases} 0 & \text{if } x \notin \mathcal{M} \\ 1 & \text{if } x \in \mathcal{M} \end{cases} \tag{1.2.4}$$

With this choice therefore, the probability $P_{RP}[\mathcal{X}_T]$ is only non-zero for the paths that strictly start from reactant state and end in product state. However, the probability ratio between the reactive pathways remains the same. As we have shrunken the accessible trajectory, to maintain the normalization of probability density, we evaluate the new partition function

$$\mathcal{Z}_{RP} = \int \mathcal{D}[\mathcal{X}_T] h_R(\mathbf{Q}_0) h_P(\mathbf{Q}_T) P[\mathcal{X}_T] \tag{1.2.5}$$

Here, we recall the famous path integral identity

$$\int \mathcal{D}[\mathcal{X}_T] = \int d\mathbf{Q}_0 d\mathbf{Q}_1 d\mathbf{Q}_2 \dots d\mathbf{Q}_n$$

Transition Path Sampling

Having defined the notion of TPE, Dellago *et al.*[53], devised an MC search directly in the trajectory space to sample transition pathways of the system. Starting from an initial guess for a reactive path such as the black line in the Figure 1.5, the TPS algorithm generates a new candidate trajectory (using a method discussed below), which it then decides to either accept or reject depending on a Metropolis criterion. The detailed balance condition associated with this criterion reads:

$$\mathcal{P}(\mathcal{X}_T^{\text{old}} \to \mathcal{X}_T^{\text{new}}) P_{RP}(\mathcal{X}_T^{\text{old}}) = \mathcal{P}(\mathcal{X}_T^{\text{new}} \to \mathcal{X}_T^{\text{old}}) P_{RP}(\mathcal{X}_T^{\text{new}}) \tag{1.2.6}$$

where $\mathcal{P}(\mathcal{X}_T^{\text{old}} \to \mathcal{X}_T^{\text{new}})$ is the probability of generating a new path starting from the old one. The $\mathcal{P}(\mathcal{X}_T^{\text{old}} \to \mathcal{X}_T^{\text{new}})$ can be decomposed as

$$\mathcal{P}(\mathcal{X}_T^{\text{old}} \to \mathcal{X}_T^{\text{new}}) = \mathcal{G}(\mathcal{X}_T^{\text{old}} \to \mathcal{X}_T^{\text{new}}) \times \mathcal{A}(\mathcal{X}_T^{\text{old}} \to \mathcal{X}_T^{\text{new}}) \tag{1.2.7}$$

where $\mathcal{G}(\mathcal{X}_T^{\text{old}} \to \mathcal{X}_T^{\text{new}})$ denotes the probability of generating $\mathcal{X}_T^{\text{new}}$ from $\mathcal{X}_T^{\text{old}}$ and $\mathcal{A}(\mathcal{X}_T^{\text{old}} \to \mathcal{X}_T^{\text{new}})$ is the acceptance probability. Placing back this definition into Equation (1.2.6), we obtain the Metropolis acceptance rule:

$$\mathcal{A}(\mathcal{X}_T^{\text{old}} \to \mathcal{X}_T^{\text{new}}) = \min\left\{1, \frac{\mathcal{G}(\mathcal{X}_T^{\text{new}} \to \mathcal{X}_T^{\text{old}}) P_{RP}(\mathcal{X}_T^{\text{new}})}{\mathcal{G}(\mathcal{X}_T^{\text{old}} \to \mathcal{X}_T^{\text{new}}) P_{RP}(\mathcal{X}_T^{\text{old}})}\right\} \tag{1.2.8}$$

Considering that the old trajectory must always remain reactive, then we obtain

$$\mathcal{A}(\mathcal{X}_T^{\text{old}} \to \mathcal{X}_T^{\text{new}}) = h_R[\mathbf{Q}_0^{\text{new}}] h_P[\mathbf{Q}_T^{\text{new}}] \min\left\{1, \frac{\mathcal{G}(\mathcal{X}_T^{\text{new}} \to \mathcal{X}_T^{\text{old}}) P(\mathcal{X}_T^{\text{new}})}{\mathcal{G}(\mathcal{X}_T^{\text{old}} \to \mathcal{X}_T^{\text{new}}) P(\mathcal{X}_T^{\text{old}})}\right\} \tag{1.2.9}$$

If the proposed trajectory is not reactive, it gets rejected, and the overall step is restarted. In the case of reactivity, however, the acceptance probability $\mathcal{A}$ is evaluated, where for $\mathcal{A} = h_R[\mathbf{Q}_0^{\text{new}}] h_P[\mathbf{Q}_T^{\text{new}}]$ the trajectory is immediately accepted as the new transition path. Otherwise, a random number is generated where the trajectory is accepted only if $\mathcal{A}$ is lower than this number. By iteratively generating a new trajectory using the previously accepted path, TPS obtains a Markov chain ensemble of transition pathways for the reaction.

To generate new paths with significant structural differences in the steps of TPS, a valid approach is called the "shooting move". In the most general form of shooting [139], we randomly perturb one of the configurations $\mathbf{Q}_k$ (either the atoms' velocities or the positions) in the reactive portion of the old trajectory. Here, TPS initiates two trajectories from the modified point $\mathbf{Q}'_k$, one forward in time and one backward, such that they have the same temporal length as the old trajectory. The probability of generating this move can be expressed as:

$$\mathcal{G}(\mathcal{X}_T^{\text{old}} \to \mathcal{X}_T^{\text{new}}) = P_{\text{gen}}(\mathbf{Q}_k \to \mathbf{Q}'_k) \left(\prod_{l=k+1}^{n(=T/\delta t)} p[\mathbf{Q}'_{l-1} \to \mathbf{Q}'_l]\right) \left(\prod_{l=1}^{k} \bar{p}[\mathbf{Q}'_l \to \mathbf{Q}'_{l-1}]\right) \tag{1.2.10}$$

wherethe $\bar{p}[\mathbf{Q}'_l \to \mathbf{Q}'_{l-1}]$ is the probability of Markov jumps in backward dynamics and $P_{\text{gen}}(\mathbf{Q}_k \to \mathbf{Q}'_k)$ denotes the probability of obtaining $\mathbf{Q}'_k$ from $\mathbf{Q}_k$. We assume symmetry for the latter, and thus, its contribution in the acceptance probability is canceled out. By placing the definition of $\mathcal{G}(p \to p')$ into Equation (1.2.9) we obtain

$$\begin{aligned}\mathcal{A}(\mathcal{X}_T^{\text{old}} \to \mathcal{X}_T^{\text{new}}) &= h_R[\mathbf{Q}_0^{\text{new}}] h_P[\mathbf{Q}_T^{\text{new}}] \\ &\min\left\{1, \frac{P(\mathcal{X}_T^{\text{new}}) \left(\prod_{l=k+1}^{n} p[\mathbf{Q}_{l-1} \to \mathbf{Q}_l]\right) \left(\prod_{l=1}^{k} \bar{p}[\mathbf{Q}_l \to \mathbf{Q}_{l-1}]\right)}{P(\mathcal{X}_T^{\text{old}}) \left(\prod_{l=k+1}^{n} p[\mathbf{Q}'_{l-1} \to \mathbf{Q}'_l]\right) \left(\prod_{l=1}^{k} \bar{p}[\mathbf{Q}'_l \to \mathbf{Q}'_{l-1}]\right)}\right\}\end{aligned} \tag{1.2.11}$$

Here, we note that all the Markovian forward jumps are canceled out with the identical terms in the $P[\mathcal{X}_T]$:

$$\mathcal{A}(\mathcal{X}_T^{\text{old}} \to \mathcal{X}_T^{\text{new}}) = \\ h_R[\mathbf{Q}_0^{\text{new}}] h_P[\mathbf{Q}_T^{\text{new}}] \min\left\{1, \frac{\rho(\mathbf{Q}_0')}{\rho(\mathbf{Q}_0)} \prod_{l=1}^{k} \frac{p[\mathbf{Q}_{l-1}' \to \mathbf{Q}_l']\, \bar{p}[\mathbf{Q}_l \to \mathbf{Q}_{l-1}]}{p[\mathbf{Q}_{l-1} \to \mathbf{Q}_l]\, \bar{p}[\mathbf{Q}_l' \to \mathbf{Q}_{l-1}']}\right\} \tag{1.2.12}$$

Upon further assumption of microscopic reversibility

$$\bar{p}[\mathbf{Q}_l \to \mathbf{Q}_{l-1}]\rho(\mathbf{Q}_l) = p[\mathbf{Q}_{l-1} \to \mathbf{Q}_l]\rho(\mathbf{Q}_{l-1})$$

in the dynamics, the expression for acceptance probability can be more simplified

$$\mathcal{A}(\mathcal{X}_T^{\text{old}} \to \mathcal{X}_T^{\text{new}}) = h_R[\mathbf{Q}_0^{\text{new}}] h_P[\mathbf{Q}_T^{\text{new}}] \min\left\{1, \frac{\rho(\mathbf{Q}_k')}{\rho(\mathbf{Q}_k)}\right\} \tag{1.2.13}$$

Therefore, for the cases such as Langevin dynamics, the Metropolis criterion only depends on the ratio of the equilibrium probability of the shooting configurations between the old path and the new. In practice, we can first perturb the configuration $\mathbf{Q}_k$ to obtain $\mathbf{Q}_k'$, and generate a random number from a uniform distribution in the interval $[0, 1]$. If this number is lower than the ratio $\frac{\rho(\mathbf{Q}_k')}{\rho(\mathbf{Q}_k)}$ we accept this modification. Then, we initiate a trajectory from $\mathbf{Q}_k'$ for $n - k$ steps (e.g. using an MD simulation), representative of the forward segment of the dynamics. If this trajectory arrives at product state, then we initiate the backward segment. In case the latter trajectory also arrives at the reactant state, then the combined trajectory is accepted as a new transition path. Otherwise, we start over from a new perturbation of $\mathbf{Q}_k$.

Extracting information from the TPE

In principle, the algorithm described above can generate an ensemble of transition pathways for any complex rare event without requiring any CV, and at a lower cost than using a straightforward equilibrium MD simulation. However, once the ensemble is obtained further calculations are needed to extract useful information such as kinetics. Here, we briefly present two examples of valuable quantities that can be derived using the TPS approach. We then mention few

Let us first discuss the calculation of transition rates. We begin by defining a correlation function [53]

$$C(t) = \frac{\langle h_R(\mathbf{Q}_0) h_P(\mathbf{Q}_t)\rangle}{\langle h_R\rangle} \tag{1.2.14}$$

where the $\langle . \rangle$ denotes an average over the TPE. $C(t)$ reaches its asymptotic value for very large times

$$C(t) \approx \langle h_R(\mathbf{Q}_t) \rangle \left(1 - e^{-t/\tau_r}\right) \tag{1.2.15}$$

where $\tau_r = 1/(k_{RP} + k_{PR})$ denotes the reaction time with k_{RP} being the rate of transition from reactant to product state. If time-scales are well separated for a transition, then there exist a time regime $t \gg 1$ where $C(t)$ scales linearly with the transition rate:

$$C(t) \approx k_{RP} t \tag{1.2.16}$$

Therefore, by monitoring the evolution of $\mathrm{d}C(t)/\mathrm{d}t$, the transition rate k_{RP} can be identified as the value of this quantity in the plateau regime. To evaluate this quantity in practice, one could reformulate $C(t)$ in terms of free energy differences at each time and then utilize Umbrella Sampling to calculate them.

An alternative method to greatly improve the calculation of the rate is provided by a modified version of TPS named the *Transition Interface Sampling* (TIS) [140]. In TIS approach, the transition region is divided into *interfaces* (with the help of an order parameter) such that the boundary of reactant state is the first interface and the last one is the boundary of the product. Then, the rate of transition is calculated as the effective positive fluxes through these interfaces. This way TIS achieves a reduction of the computational time required to evaluate the rates.

Other than the rates, TPS can also be utilized to calculate the committor probability function $q(r)$. A value of $q(r) = 0$ means no commitment at all from the trajectories passing through r to arrive to the product state and $q(r) = 1$ indicates full commitment. Conversely, the configurations with the committor values of $1/2$ denotes the transition state ensemble in the configuration space. Hence, evaluation of committor function provides the optimal CV to characterize the transition.

The quantity $q(r)$ is the stationary state of the time-dependent committor function $q(r,t)$. To calculate the $q(r,t)$ from the ensemble of pathways provided by TPS, one can count the number of trajectories that crossed r and after time t entered the product state. This translates into the expression

$$q(r,t) = \frac{1}{N} \sum_{i=1}^{N} h^(i)_P(\mathbf{Q}_t) \tag{1.2.17}$$

where N denotes the number of the trajectories. Assuming that all the trajectories where independently generated, then the standard deviation of this function can be written as

$$\Delta q(r,t) = \sqrt{\frac{1}{N} q(r,t)(1 - q(r,t))} \tag{1.2.18}$$

By following the evolution of this standard deviation we can terminate the calculation of the committor function once we have reached the desired accuracy.

Applications of TPS and its variants

In the past 25 years since the inception of TPS, a wide range of successful applications have been reported. Besides the initial studies on chemical reactions such as autoionization of liquid water [141], one of the first applications to conformational changes of biomolecules was reported in 2003. In this study by P.G. Bolhuis [142], the folding pathway of the C-terminal of β hairpin of the G-B1 protein was studied with TPS and TIS, where the result showed reasonable agreement with the experimental data. Another notable early application was dedicated to the study of folding/unfolding pathways of the Trp-Cage, a molecule exclusively designed to be a "fast-folder" protein, originally constructed from Exendin-4 protein back in 2002 [143]. In 2006, 5 years before equilibrium simulations provided by Anton [26], Juraszek and Bolhuis [144] applied TPS to a system of Trp-Cage solvated in explicit water and realized two distinct folding pathways for this molecule. Subsequently, by applying TIS and one of its more efficient variants called *Forward Flux Sampling* [145], they calculated the folding's transition rates with good agreements with experiments [146].

As discussed, TPS's efficiency heavily depends on the adopted shooting move when generating a new pathway. In these years, many approaches have been suggested to improve on the original shooting method. *One-way* shooting [147] is one such approach, where one generates only forward or backward segments of a new path and retains the complementary segment from the previous trajectory. This approach, though, increases the possibility of generating a new path with high accepting probability ($\mathcal{A}$ in Equation (1.2.11)) it requires many iterations to de-correlate the generated pathways in the Markov chain. To face this limitation, *Spring* shooting method [139] shifts the time-slice of the shooting in the new trajectory some random frames away from the previous shooting time-slice. Here, depending on the previous shooting to generate a new path in the backward (forward) direction, the new shooting time-slice will be closer to the product (reactant) state, and the newly generated path will accordingly be initiated in the forward (backward) direction. As if in a spring, this method (schematically illustrated in Figure 1.6) pushes the shooting time-slice toward the top of the free-energy barrier and, therefore, retains a high acceptance rate but increases the efficiency in obtaining de-correlated paths. The introduction of spring shooting has remarkably extended the domain of applicability for TPS to large conformational changes, such as association/dissociation of the hydrophilic β-lactoglobulin dimer, which has the experimental rate $k_{\text{off}} < 0.1\text{s}^{-1}$ [148, 149]. The idea of shooting from the top of the energy barrier has also been considered in other approaches, such as the one

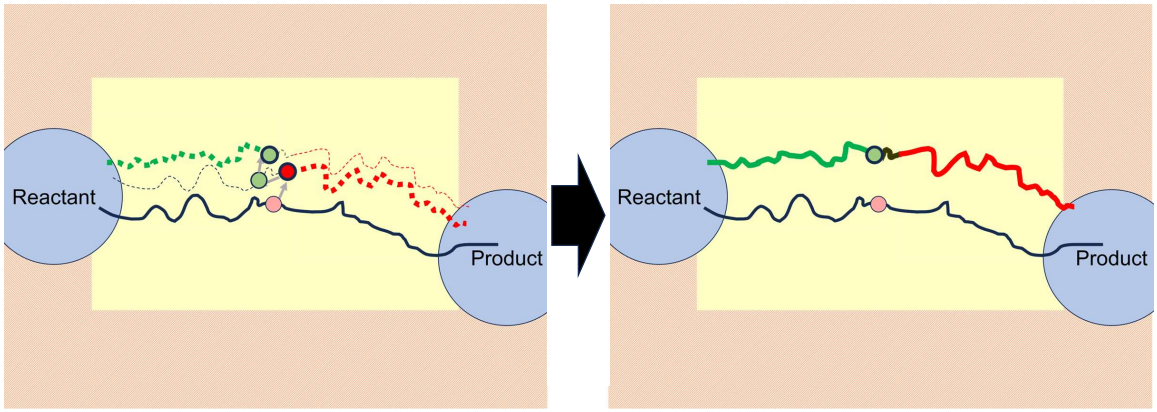


Figure 1.6: The Spring shooting method. (left panel) Similar to one-way shooting, after perturbing a configuration at random (the red circle), we initiate only one MD in the direction for example forward in time (the heavy red line). Then, we join the opposite segment of the previous path to the new trajectory (the thin black dashed line). However, unlike one-way shooting, we perform the next shooting some time steps along the backward direction of the old trajectory (the green circle). Finally, we perform a new MD in the backward direction to get the heavy green line. (right panel) By joining these new segments (red and green lines) to the part of old trajectory that connects the two shooting moves (the small black line) we a get a new trajectory.

suggested by Jung *et al.*[150]. However, this approach requires the definition of an order parameter in order to determine the vicinity of the barrier top.

Other notable challenges that can arise in TPS are the existence of distinct, well-separated transition channels or multiple long-lived intermediate states between the reactant and the product state. In the case of the former, a higher free energy barrier between reaction channels can effectively result in a (meta) rare event in the MC sampling of the path space. To tackle this challenge, other enhanced sampling methods can potentially be integrated into TPS to allow for switching between multiple channels. Borrero *et al.*[151] have devised such an approach based on Metadynamics. Starting from a randomly chosen configuration, here, the shooting move applies a biasing force along a pre-defined CV that identifies the direction toward the top of barrier. Then, a candidate trajectory is generated following the original algorithm of TPS. After multiple successive shootings and generating new paths, the force is built up such that in the next shooting the perturbed configuration would be in the previously not sampled transition channel.

Apart from multiple channels, when intermediate long-lived states exist between the reactant and product basins, the trial trajectories in TPS have a high chance of "getting stuck" in one or more of these states. Consequently, the acceptance rate rapidly decreases in these instances and therefore, sampling reactive pathways becomes rather inefficient. Already in 2008, Rogal and Bolhuis [152] had developed an extension of TPS using the

interfaces in TIS. In their approach, global transition pathways between the initial and final states were formed by combining multiple sample transition trajectories between interfaces that signified the intermediates. This method was then revisited and improved with the addition of Replica Exchange TIS method [153, 154].

Finally, let us briefly mention one of the AI-based improvements that have recently been introduced to accelerating the sampling of pathways in TPS. Jung *et al.*[155] developed a scheme that utilizes a Deep Neural Network (DNN) model to guide the sampling for generating transition paths with high acceptance probability. Starting from a successful reactive pathway, the DNN parameterizes an initial guess for the committor function. Next, a shooting move is initiated around the transition state indicated with the guessed committor. The algorithm then generates a new trajectory according to the steps of originial TPS. Finally, depending on acceptance or rejection of the new trial pathway, the DNN is updated to find a better parameterization of the committor. By studying the transitions of two relatively small systems, Jung *et al.*demonstrated that this reinforcement learning scheme is in fact more efficient in producing the pathways compared to original TPS. Remarkably, Jung and coworkers managed earlier this year to enable this algorithm to even dispose a mathematical expression for the committor function with the help of symbolic regression [156].

1.3 Applying Quantum Computers to tackle the molecular sampling problem

The mentioned challenges of TPS when applied to complex transitions occurring in highly rugged energy landscapes can be summarized into two general categories:

(i) Efficiently generating viable trial trajectories at an acceptable computational cost.

(ii) Simultaneously, reducing the correlation of generated paths.

Once the two metastable states grow in distance (corresponding to a reaction time of $\gtrsim 1\mu$s), the probability of attracting new reactive trajectories decreases rapidly as we shoot away from each state [31]. Consequently, the perturbation must remain appropriately small to ensure a finite acceptance probability in the Metropolis criterion. Such a local stochastic move obstruct the ability of TPS to sample from all the transition channels available to a rare reaction. Therefore, the question is how to perform global moves in the MC sampling of TPE while generating trajectories with high acceptance probability? Before answering, we first introduce another piece of the puzzle, the quantum computing.

Any molecular system is first and foremost comprised of electronic and nuclear degrees of freedom. However, solving the full QM wavefunction of a biomolecule to obtain the

evolution and behavior in time is not practical on any classical computer. In quantum chemistry, the exponential cost of computations in such problems renders the possibility available only for cases with very few atoms and after various approximations (e.g. by using Hartree-Fock method) [157]. Manin in 1980 [158], and Feynmann in 1982 [159] were the first to propose the concept of utilizing a QM system to simulate the behavior of the atoms in a molecule. Such *universal quantum simulators* were envisioned as devices that remedied the curse of dimensionality in classical computation approaches to many-body problems. A quantum simulator maps the Hamiltonian, containing electronic and nuclear interactions in a molecule, into the couplings between its components. Then, it simulates/solves the Schrodinger equation in time (in its most general form)

$$i\hbar\frac{\partial|\Psi(t)\rangle}{t} = \hat{H}(t)|\Psi(t)\rangle$$

, where $H(t)$ is the Hamiltonian of the molecule. This allows us to efficiently replicate and study the behavior of the molecule under study. Efficiency here means, contrary to the classical counterpart, that the required number of qubits (quantum bits) and quantum operations scales utmost polynomially with the system's size. The conjecture that quantum computers can simulate QM systems without an exponential overhead was theoretically proven in 1996 by Llyod [160].

Since 1996, a great effort has been dedicated to realizing such powerful devices and algorithms that can fully wield their power. As an example of quantum algorithms which are advantageous for biomolecular studies, we mention Quantum Phase Estimation that has the potential to evaluate the energy required for ligand-protein and protein-protein interactions [161, 162]. This is obviously not to mention the possibility of simulating the entire molecule dynamic using its quantum Hamiltonian with a QC. However, technical problems such as maintaining coherence time in the presence of environmental noise have not yet allowed the development of fully fault-tolerant and large-scale QCs [157, 163]. Therefore, the application of the quantum algorithms has been limited, which often rely on a large number of entangled qubits to remain coherent under execution of many quantum operations [164].

To measure the performance of quantum computers against each other and older generation devices, "Quantum Volume" was proposed by Moll *et al.*, and later adopted by IBM [165–167]. Quantum Volume evaluates the number of qubits n_{qb} and the number of cycles d_n a random array of two-qubit gates can be performed consecutively on these qubits, such that the output read-out (measurement) can be maintained without error on average 2/3 of the time. This means that a quantum computer that has 6 qubits and can maintain such error only for 6 cycles of two-qubit gates, has the Quantum Volume of $2^6 = 64$. At the time of writing this thesis, the highest performance in terms of Quantum Volume is

associated to "Quantinuum's H1-1" device based on Trapped-Ion qubits that has reached $2^{19} = 524,288$ [168, 169] (19 qubits with 19 cycles of two-qubit gates). The IBM's 127-qubit computer, Eagle, which utilizes superconducting architecture has only reached 2^6 volume[1]. However, to simulate a biologically relevant molecule and even without the presence of a solvent, we most likely require more than 10^3 qubits just to encode the atom's nucleus and they have to remain coherent under a large number of gate cycles. A condition that seems to be out of reach for any near term QC device.

Luckily, in the era of noisy intermediate-scale quantum machines, another class of novel algorithms have emerged that aim for practicality instead of full quantum computer approaches. These methods employ a hybrid combination of classical computing and QC in order to obtain solutions for problems where neither approach is sufficient independently. Algorithms such as Variational Quantum Eigensolver (VQE) and QAOA[2] are two examples from this class [161, 171, 172]. We do not aim to review these algorithms and their respective applications in recent years. Rather, we want to follow their trails to develop a method that can utilize the current NISQ devices for tackling existing problems in molecular sampling with classical computers. In particular, our main concern as mentioned earlier is to sample the TPE of a rare reaction while producing uncorrelated trial pathways at an acceptable computational cost. The hope here is that as the QCs are growing in size the algorithms such as ours will have the edge in the foreseeable future for complex path sampling calculations.

The pioneering applications of QC in the context of quantum chemistry and biology strongly support this potential [171, 173–179]. However, these applications mostly comprised of very small size systems such as finding the ground-state of HeH^+ compound, or searching for folded structure of proteins and/or equilibrium configurations of polymers using lattice models (again only at modest lattice size). In contrast to these cases, our goal is to maintain the atomistic details provided by the current MD forcefields while examining the advantage of current QCs. Considering this reason, we focus on quantum annealing and the DWave annealer since it has the most working qubits available even though it is not technically a universal QC [180]. Nevertheless, our approach has the potential to be reformulated for any other types of QC, given they can provide a useful number of qubits.

1.3.1 Quantum computing: A brief overview

As mentioned QC is made out of qubits, a basic unit of quantum information [181]. A single qubit is a two level quantum system that can take the form of (counter)clockwise

[1] The data on the latest generation of IBM's quantum computer, the 433-qubit computer Osprey, is not yet available [170]

[2] Quantum approximate optimization algorithm

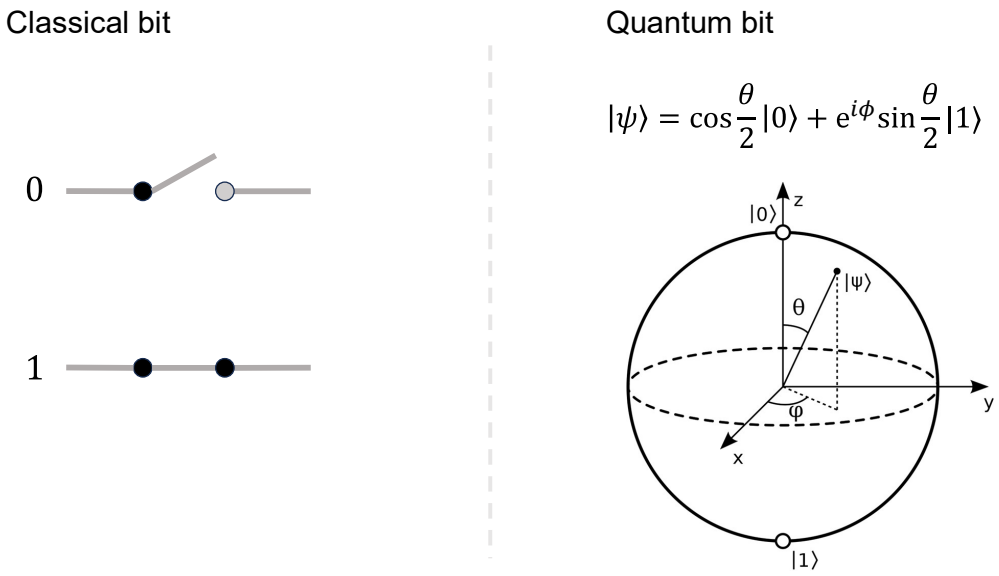


Figure 1.7: The basic unit of information in classical and quantum computation. In classical computation, the bit can be realized schematically as having a circuit either being closed = 1-state or opened = 0. However, qubit is in a quantum superposition of the two states. This in theory allows a QC to have exponential speed up in terms of computation over the classical counterpart.

currents in superconductance, atomic or nuclear spins, or polarization of photons. The pure state of a qubit can be represented as

$$|s\rangle = \alpha|0\rangle + \beta|1\rangle \tag{1.3.1}$$

where $|\alpha|^2$ and $|\beta|^2$ determines the probability of the system being in 0 and 1 states respectively. Here, the computational basis is assumed to be in the basis of Pauli's σ_z matrix such that

$$|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad |1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{1.3.2}$$

The geometric representation of the state in Equation (1.3.1) can be depicted by a vector in Bloch sphere, Figure 1.7. Quantum computing and quantum simulators fall into two categories: the *digital* and *analog* approaches. Digital QCs are the holy grail of quantum computing as they allow for the manipulation and control of every single qubit. They have been theoretically proven to provide the notion of *universal quantum machine* that can formally run any quantum algorithm [182, 183]. In digital QC, the basic quantum operation is presented as a quantum logic gate acting on single or multiple circuits which represent the qubits Figure 1.8.(a). To give an example of the gates, the X-gate is a unitary operator implementing the Pauli's σ_x matrix (Figure 1.8.(b)) that flips the $|0\rangle$ state to $|1\rangle$ or vice versa. Similarly, a two-qubit example would be the quantum CNOT gate (also known

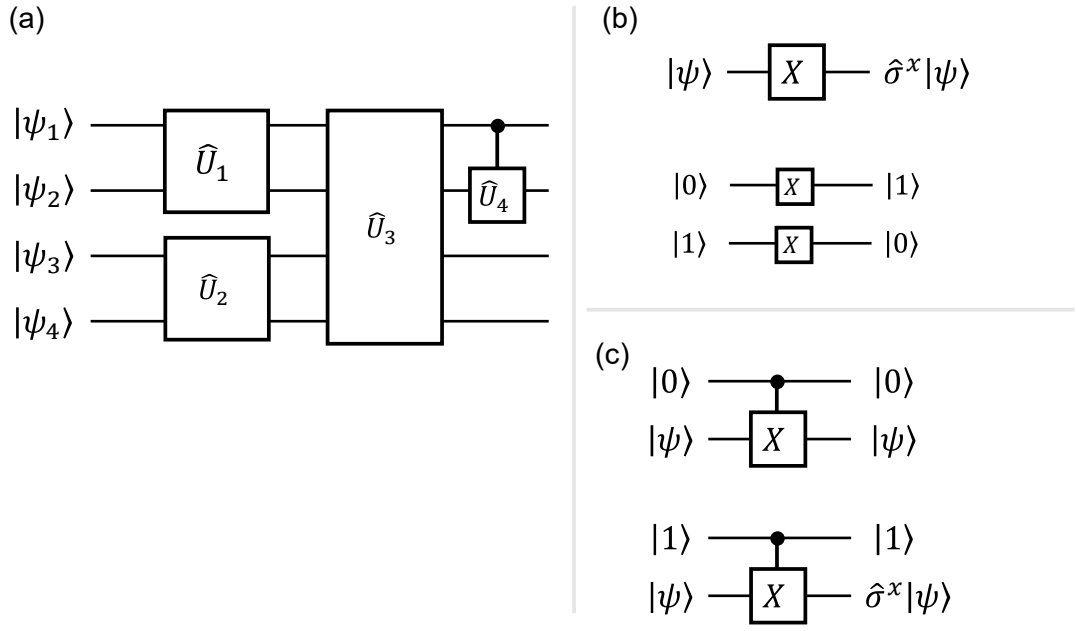


Figure 1.8: Quantum Circuit model of digital quantum computation. (a) In this model, every qubit is depicted as circuit, while the quantum operations in terms of unitary operators $\hat{U}$ are realized through quantum logic gates. (b) The X-gate which is the gate responsible for applying the Pauli's σ^x operator on a qubit and flips the state of that qubit. (c) The CX or quantum CNOT-gate. This gate flips a qubit's state based on whether another qubit is in the state $|0\rangle$ or $|1\rangle$. For more information on circuit model and quantum gates we refer to [181].

as CX gate, Figure 1.8.(c)), that flips the state of one qubit based on the result of another. These gates are formally represented by:

$$\text{X-gate} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \text{CNOT-gate} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{1.3.3}$$

To simulate a system a digital QC, we first reformulate the unitary time evolution $e^{-i\hat{H}t/\hbar}$ of the system in terms of the quantum gates. Here we assume that system is governed by a time-independent Hamiltonian for simplicity. Next, by applying the Trotter decomposition to the $e^{-i\hat{H}t/\hbar}$ we obtain

$$e^{-i\hat{H}t/\hbar} = \prod_{i=1}^{N_t} \left(e^{-i\hat{H}\Delta t/\hbar} \right) \tag{1.3.4}$$

where $t = N_t \times \Delta t$. We then encode the initial configuration of the system $|Psi(0)\rangle$ into the QC's qubits. By iterative application of $e^{-i\hat{H}\Delta t/\hbar}$ (see Figure 1.9) we evolve the qubits until we reach the cumulative time $= t$. The quantum measurement of the qubits after this evolution will give us the final configuration $|\Psi(t)\rangle$ that we were after. One major challenge

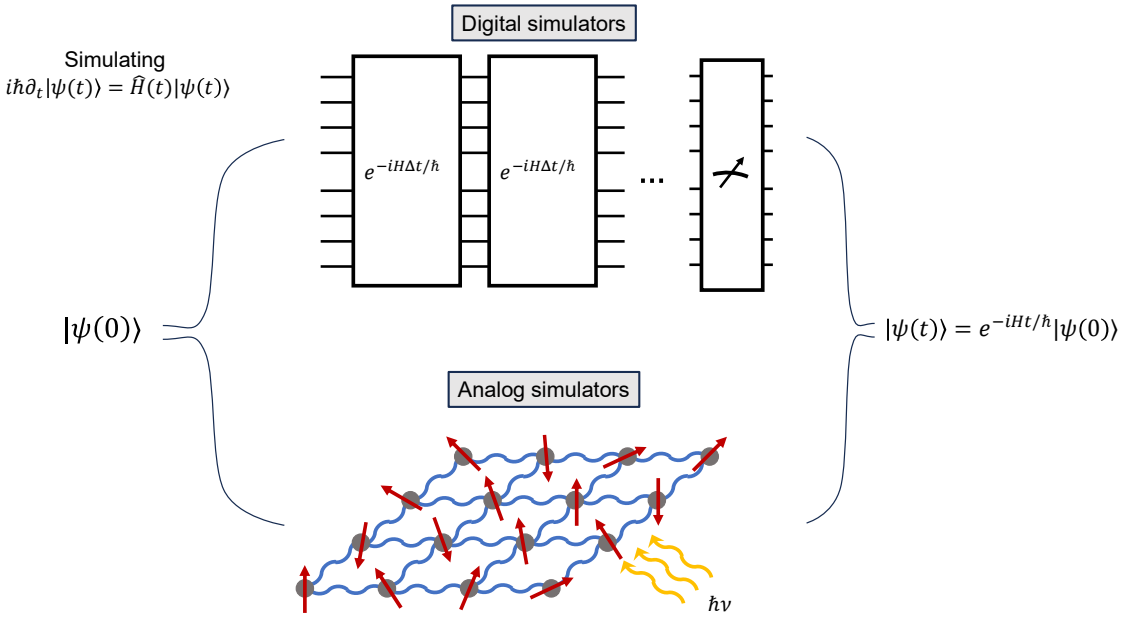


Figure 1.9: A comparison between digital and analog approaches to quantum simulation. First, we prepare the qubits of the QC according to the initial state of the system, for which we are attempting to solve the Schrodinger's equation. In the digital approach, one would express the unitary time operator of the system, $e^{-iHt/\hbar}$, in terms of quantum logic gates. After Trotterization of this operator, then we apply N_t cycles of $e^{-iH\Delta t/\hbar}$ to the qubits, where $t = N_t\Delta t$. In the end, quantities such as $\langle O(t)\rangle$ can be measured by performing $\langle\psi(t)|O(t)|\psi(t)\rangle$ on the final state $|\psi(t)\rangle$ of the qubits. In the analog approach, on the other hand, we tune (e.g. using laser pulses) the coupling between the components/qubits of the quantum device according to Hamiltonian of the system H. In this method, the simulation of Schrodinger's equation is intrinsically achieved by letting the qubits evolve for time t under the influence of H. This schematic representation of the two approach has been adapted from [157].

that has faced all the physical realizations of a fault-tolerant universal QC is: How does one retain the information stored in the qubits in the presence of unwanted noise? In the classical (digital) computers, stability of the bits is guaranteed by following error-correction protocols in computation which are implemented either in the low-level hardware or high-level software [184]. Unfortunately, classical error-correction algorithms cannot be directly migrated into a QC due to a fundamental theorem of QM called *no-cloning theorem*[185, 186]. This theorem forbids the creation of identical copies of an *unknown* quantum state which is essential for classical error-correction. Despite of this limitation, back in 1995, Shor published the first ever quantum error correcting code which allowed to side step the no-cloning theorem by storing information simultaneously into 9 qubits [187]. Ever since, quantum error correction has been an integral part of scientific and industrial effort to build a fault-tolerant QC. However, all the modern approaches introduce an overhead on the number of *physical* qubits required in order to produce a single fault-tolerant *logical*

qubit. However, as the qubits increase, the noise due to the coupling to the environment also increases. This is the reason why it is considerably difficult to realize a true fault-tolerant digital QC that has a useful number of (*logical*) qubits.

Apart from the digital QCs, there is also the analog approach to quantum computing that is more similar to purpose-built machines in classical computation [157]. In analog QCs, the quantum operations are implemented through continuous interactions, such as magnetic field gradients Figure 1.9. In simulations with an analog QC, the coupling between qubits are prepared according to the Hamiltonian of a system $H(0)$ at time zero, whose dynamic we want to replicate (e.g. a molecule). Then, by evolving the interactions in the QC in real time according to $H(t)$, we directly solve/simulate the time-dependent Schrodinger equation. Again, by measuring the state of each qubit at the end, we obtain the final state of the system that we simulated.

The analog QCs offer the advantage of being more easily scalable as they do not require individual manipulation of qubits. However, this also introduces challenges such as more difficult error correction and sensitivity to noise. Moreover, analog quantum computing can be universal only in certain conditions [188]. Nevertheless, analog can be a promising approach for quantum simulation, as physical developments such as DWave's quantum annealer have demonstrated.

1.3.2 Quantum Annealing

Quantum annealing, or adiabatic quantum computation[1], is a category of analog quantum computing with a more specialized purpose. Quantum annealing is built to solve classical combinatorial optimization problems, with applications ranging from computer science problems to many body problems [180, 190]. One often can, in these problems, cast the task of minimizing the cost function onto finding the ground state of a spin-glass Hamiltonian. Such Hamiltonian has many local minima, and stochastic optimization methods implemented on a classical computer, like Simulated Annealing (SA) [191], have a hard time searching for the global minimum. The dynamics in SA algorithm is based on thermal fluctuations, and for ergodic systems, it has an upper bound time complexity of $\mathcal{O}(N)$ with N being the system's size. However, in spin glass's Hamiltonian or traveling salesman's cost function, the heights of the barriers can grow exponentially $\mathcal{O}(2^N)$. Therefore, in these problems, searching for the global minimum with SA effectively becomes a rare event [54]. Quantum annealing was found as an alternative that could potentially perform this task more efficiently. The possibility comes from the fact that through the process of quantum

[1] We should note that quantum annealing is a broader term than adiabatic quantum computation. In fact, quantum annealing allows for nonadiabatic transitions that would be present in a non-ideal experimental setup. However, we have chosen to follow the tradition of considering quantum annealing in the restricted sense of adiabatic quantum computation [58, 189].

annealing, the quantum fluctuations allow for tunneling through the macroscopic barriers. Therefore, by the end of annealing, there is a finite probability that the annealer is in the ground state of the target Hamiltonian.

The quantum adiabatic theorem (QAT) is a central concept in quantum annealing [58, 192]. According to this theorem: "A system remains in its instantaneous eigenstate as it undergoes a sufficiently slow perturbation if there is a gap between its instantaneous eigenvalue and the rest of spectrum [193]". Therefore, the ground state of a target Hamiltonian H_{target} may be found by preparing an Ising system in the ground state of an easy-to-prepare Hamiltonian H_0. H_0 must be chosen such that it does not commute with H_{target}, e.g. having been written in σ^x Pauli's matrices. Then, by performing a very slow quantum switching, we gradually transition the coupling between the spins from H_0 to H_{target}. QAT guarantees to remain in the ground state of H_{target} by the end of this process. Mathematically, this transition can be encoded as a time-dependent Hamiltonian on the system

$$H_{\text{QA}}(t) = A(t)H_0 + B(t)H_{\text{target}} \tag{1.3.5}$$

Here the coefficient $A(t)$ and $B(t)$ encode the switching protocol. The specific form of these functions is not important as long as we ensure that for switching time of $[0, t_f]$, $A(t)$ is slowly reduced from $A(0) = 1$ to $A(t_f) = 0$ and $B(t)$ is accordingly increased from $B(t) = 0$ to $B(t) = 1$.

Using adiabatic approximation one can prove the QAT, for which we refer to [194] for the complete proof and only discuss the results here. The adiabatic approximation is concerned with finding the solution of the time-dependent Schrodinger equation with the Hamiltonian of Equation (1.3.5). Since our concern is to find the ground state of H_{target}, we assume that the initial wave function of annealer $|\psi(0)\rangle$ (or any quantum simulator with quantum annealing capabilities) is the pure ground state $|\phi_0(0)\rangle$ of $H_{\text{QA}}(0)$. Following this assumption, the adiabatic approximation states that:

$$\max_{0<t<t_f} \frac{|\langle\phi_i(t)|\frac{dH(t)}{dt}|\phi_0(0)\rangle|}{[\Delta_{0i}(t)]^2} \ll 1 \quad \text{for} \quad \forall i \geq 1 \tag{1.3.6}$$

is the necessary condition for the instantaneous state of the system to remain near to the instantaneous ground state $\langle\psi(t)|\phi_0(t)\rangle^2 \approx 1$. Here, $\Delta_{0i}(t) = |E_i(t) - E_0(t)|$ is the instantaneous energy difference between ground state and i-th state. In addition to this result, in case the switching protocol is linearly dependent on time ($A(t) = 1 - t/t_f$ and $B(t) = t/t_f$), then adiabatic approximate determines the rate $1/t_f$ of this procedure:

$$t_f \gg \max_{0<t<t_f} \frac{|\langle\phi_1(t)|\frac{dH(t)}{dt}|\phi_0(0)\rangle|}{[\Delta_{01}(t)]^2} \tag{1.3.7}$$

Therefore, based on this result, to perform the transition adiabatically, the switching rate should be much smaller than the jump frequency from the instantaneous ground state to the first excited state.

QUBO encoding: To implement the optimization problem into a QA or a QC that can perform annealing, one must first reformulate the cost function into the Hamiltonian of a quantum Ising model (H_{target} in Equation (1.3.5)). Many classical optimization problems can be formulated as Quadratic Unconstrained Binary Optimization (QUBO), either by natural application or by re-casting alternative formulations [195]. QUBO was shown to be formally equivalent to a classical Ising model. This framework has found applications (as early as the 1960s) in various fields, e.g. industry, economics, and science, and recently has attracted more attention thanks to the advent of commercially available devices like DWave [62].

As the name suggests, in QUBO, one deals with binary variables $\Gamma = \{0, 1\}$. The goal is to find the configuration of variables that would minimize/maximize a 2-order polynomial cost function. The function in question is expressed as

$$H_{\text{QUBO}} = \sum_{i,j} Q_{i,j}\Gamma_i\Gamma_j + \sum_i m_i\Gamma_i \tag{1.3.8}$$

where the optimization problem is fully determined by the values of $Q_{i,j}$ and m_i. Once the QUBO formulation of our problem is established, we can conveniently transform the cost function into an Ising Hamiltonian via $\Gamma = (1 - \sigma^z)/2$:

$$H_{\text{Ising}} = \sum_{i,j} J_{i,j}\sigma_i^z\sigma_j^z + \sum_i h_i\sigma_i^z + \text{const.} \tag{1.3.9}$$

where

$$J_{i,j} = \frac{1}{4}Q_{i,j} \quad \& \quad h_i = \frac{-1}{2}\left(m_i + \sum_j Q_{i,j}\right)$$

. σ^z here is a variable with values $\sigma^z = \{-1, 1\}$. To obtain the quantum Ising model we simply promote the variables σ^z to Pauli matrices $\hat{\sigma}^z$.

1.3.3 graph Transition Path Sampling

We are finally in a position to present our novel framework for sampling transitions pathways. In gTPS framework, ML-driven MD simulations are employed on a classical computer to efficiently perform a preliminary, uncharted exploration of the relevant regions of the molecule's configuration space. Here, each snapshot (of the system) captured during the exploration can identify a finite region of configuration space with a size comparable to

the average nearest-neighbor configuration distance. This realization enables us to derive a coarse-grained (CG) effective theory by assuming the Langevin dynamics as the underlying microscopic theory.

The strength of this representation lies in its ability to evaluate the probability of transitions as the product of probabilities for the system to visit a given ordered sequence of finite-size regions. Thus, it can adapt itself to the non-Boltzmannian distribution of the iMapD-generated dataset. Subsequently, we describe how this CG theory can be encoded into an *undirected* graph, wherein the nodes identify different finite-size regions. Then, the probability of a given reactive path is computed from the sum of the edges that connect the initial and final nodes on the network.

In principle, transition paths in a graph may be sampled with classical computers with a discrete version of the conventional TPS, where the trial trajectories are stochastically generated, e.g., by kinetic MC [196]. However, the effectiveness of these methods falters once the network increases in size and complexity. They would struggle to produce paths that are significantly different than one another in a computationally efficient manner. As a consequence, resulting Markov chains are susceptible to the same auto-correlation problem facing the original TPS.

This challenge can be overcome by resorting to a quantum annealing machine such as DWave to generate the trial paths in the Markov chain. In such an approach, transition paths on the network can be encoded in a QC by assigning qubits to each node and edge in the graph through QUBO formulation. The main point in using a QC approach is that the initial state of the quantum computer can be easily prepared in order to simultaneously encode all the transition paths connecting two given points on the graph. Then, the adiabatic switching process evolves the components of the computer toward a wave function associated with the most statistically significant transition paths. Thanks to quantum fluctuations and quantum measurement, any time the QA is reset in its initial state, the system loses memory of the previously generated path. Therefore, the trajectory obtained at the end of the annealing process is completely uncorrelated to the previous one in the Markov chain.

This way, the QA generates a new uncorrelated trial path at each measurement with a high statistical weight. However, generally in realistic conditions (e.g., for suboptimal choices of the time employed to perform the adiabatic switching procedure), the resulting path distribution may not exactly correspond to that of the underlying Langevin dynamics. Thus, we also incorporate a Metropolis acceptance/rejection criterion implemented on a classical computer to correct for such a deviation. The salient features of gTPS are illustrated in Figure 1.10.

While resorting to the intrinsic properties of the QA overcomes the autocorrelation

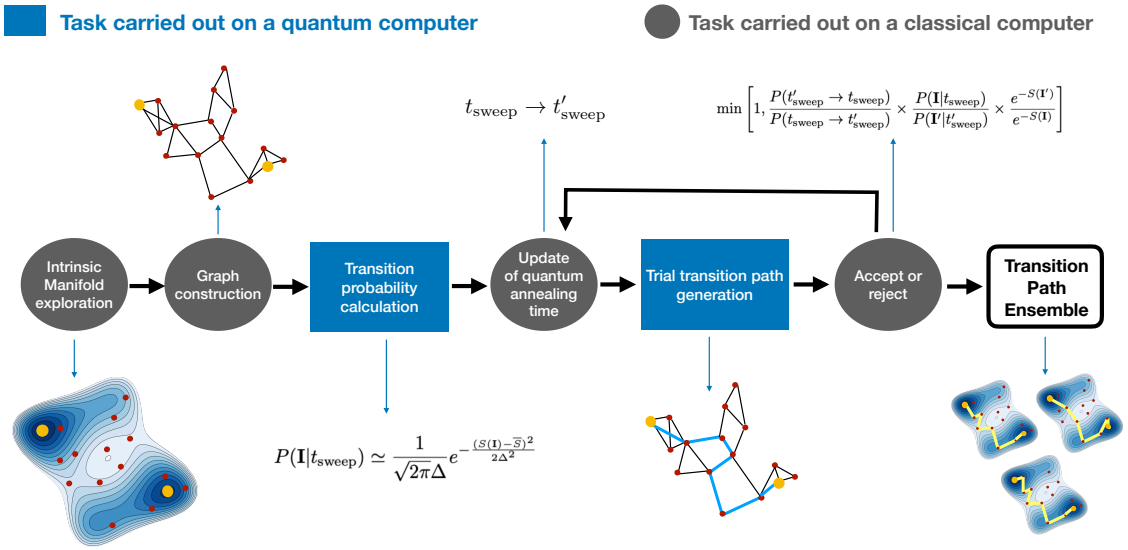


Figure 1.10: Schematic representation of gTPS scheme introduced in this work. This framework combines ML and MD performed on a classical computer and QC on a quantum annealing machine. The interplay between these steps allows our scheme to sample the full transition path ensemble without any use of unphysical biases. First, the ML-guided MD explores rapidly the relevant regions of configuration space. Then, using our CG representation of the dynamic, we encode the transitions of the system into a undirected network between the configurations retrieved in the first step. Finally, using D-Wave's QA in a MC sampling process, we generate uncorrelated trial pathways that have a high probability of being accepted in the Metropolis step. In the next chapter we delve more in-detail through each step of the gTPS framework depicted here.

problem, our adopted ML approach promotes the discovery of multiple transition channels by rapidly exploring the relevant regions of configuration space. Furthermore, the transition network built by the CG effective theory allows for the path sampling algorithm to have a global view of all the possible transition channels. Therefore, the hybrid approach of gTPS can directly tackle outstanding problems of traditional TPS. Moreover, gTPS does not require any simplification of the atomistic details provided by the MD, which is in contrast to most of the previous applications of QC to molecular sampling.

CHAPTER 2

Graph Transition Path Sampling (gTPS): in-depth discussion

In Chapter 1, we presented a concise yet comprehensive summary of the TPS and other enhanced sampling methods. We acknowledged their strength and the challenges they face in sampling rare events. By recognizing the mounting difficulties encountered by conventional TPS frameworks, QCs were then suggested as a new possible remedy. Finally, we presented our newly developed framework that aims to tackle these challenges by harnessing the power of both ML and quantum computing.

Our goal in this chapter is to delve deeper into the various theoretical aspects of gTPS, dissecting each step and discussing potential efficacies. First, we discuss iMapD, the algorithm that enables gTPS to expedite the exploration of the system's FEL. The generated conformations of iMapD provide a ground for building a mathematically rigorous effective theory based on Langevin dynamics. Utilizing Dominant Reaction Pathway [197, 198] formalism enables us to represent this CG theory and the iMapD's conformation with a network of transition that covers the whole explored region of FEL. After encoding the vertices and edges into the D-Wave machine, we finally present our MC process which generates trial paths on the network with quantum annealing.

The validation of gTPS capabilities is left to the subsequent chapters, where the practical implementation and considerations of the framework are also discussed.

2.1 Uncharted exploration of free energy landscape

It is widely accepted that at thermal equilibrium, statistically relevant conformations of molecular systems accumulate on a low-dimensional embedding within the ambient space of molecular coordinates, so-called the *intrinsic manifold* (IM). To explore this manifold efficiently, Intrinsic Map Dynamics (iMapD) provides an enhanced sampling scheme that utilizes ML's dimensionality reduction techniques. Developed by Chiavazzo *et al.*(2017) [59], iMapD starts by learning the geometry of the observed segment of the IM using an initial sampled set of configurations. It then exploits this information to identify strategic configurations that increase the chance of discovering unidentified territories of the underlying manifold, and through unbiased MD simulations initiated from these configurations, it intelligently guides the sampling process. The data-driven approach of iMapD obviates the need for identification of CVs, applying any unphysical force, or *a priori* information on the system, consequently distinguishing it from methods such as metadynamics.

In this section, we outline the main steps of the algorithm while reserving technical details for the Appendix A to maintain clarity. Additionally, we discuss further the main advantages of this algorithm within the gTPS framework while also acknowledging potential pitfalls that may impact our applications.

2.1.1 iMapD framework

Parameterizing the intrinsic manifold and identifying the boundary of observed regions: As in any ML-based method, we start by sampling an ensemble of unbiased molecular trajectories. We operate under the assumption that representative structures from the meta-stable states of interest are available, a condition often met in practice, and from these configurations, we initialize our MD simulations. The generated k configurations are then assembled into a set $\mathcal{C}_{\text{ini}} = \{\mathbf{Q}_i\}_{i=1,\dots,k}$. Next, iMapD employs a procedure known as Diffusion Map (DMAP) [199] which serves as a powerful tool to attain dimensionality reduction and infer the local structure of the intrinsic manifold.

The DMAP space of a dataset is obtained by first building a transition matrix equivalent to a random walk operator. In our implementations, the RMSD function serves as the similarity measure of this matrix. After solving the associated eigenvalue problem of this matrix, we retain a set of n (with $n \ll 3N$) eigenvectors, $\mathbf{Z} = \{\psi_1, \psi_2, \dots, \psi_n\}$, that represent the DMAP components (DCs). These DCs provide a projection of high dimensional data onto a low dimensional embedding such that each configuration Q_k is mapped onto $z_i = \{\psi_1(k), \psi_2(k), \dots, \psi_n(k)\}$ with $\psi_j(k)$ being the k-th element of component ψ_j. An illustrative application of DMAP to an arbitrary 3D dataset is presented in Figure 2.1.

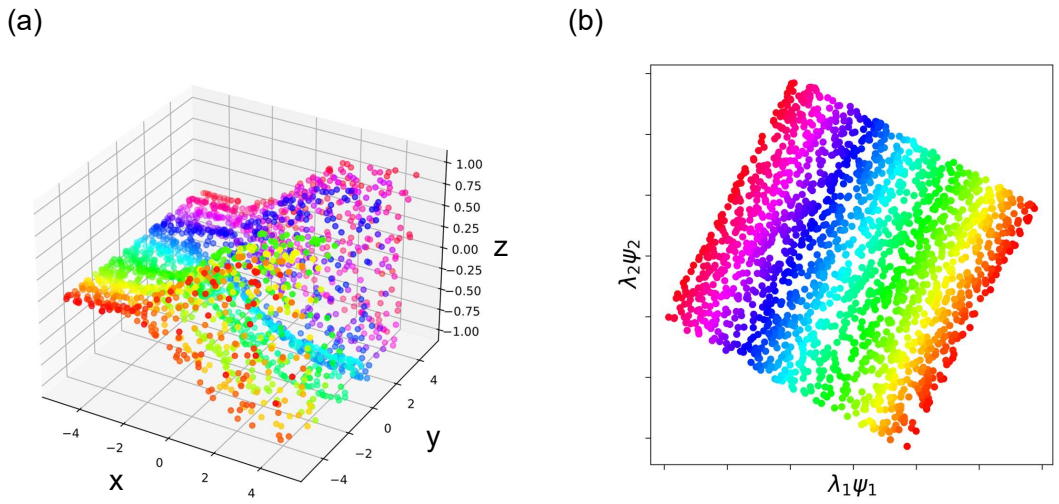


Figure 2.1: (a) A 3D dataset as an illustrative example of identifying the underlying manifold with DMAP. This data was produced by first generating random points on a plane $x, y \in [-5, 5]$. Then the third dimension to each point was added by using the function, $z(x, y) = \cos(y^2 - 4)/(1 + e^{-x})$. (b) The projection onto the first two DCs.

It has been demonstrated [199, 200], owing to the relationship between the eigenvectors of the Fokker-Planck operator and those of DMAP, the Euclidean distance between the points of this embedded space approximate the diffusion distance between members of $\mathcal{C}_{\text{ini}}$. Therefore, by exploiting this connection, a notion of the boundary for the kinetically explored regions can be established by identifying the set of configurations $\mathcal{C}^{\text{B}}$ that form a convex hull [201] in the DMAP projection around the rest of $\mathcal{C}_{\text{ini}}$.

It's noteworthy that we have not specified which eigenvectors in DMAP are the suitable components. While the first non-trivial eigenvector characterizes the principal direction of the IM, the subsequent dominant ones do not necessarily achieve the correct parameterization of other dimensions. They may be harmonic functions of the first eigenvector. According to Chiavazzo *et al.*(2014) [202], it is relatively straightforward to identify the uncorrelated components for 2D DMAPs using a visual test. In such cases, one looks at the scatter plot of $\lambda_1\psi_1$ against $\lambda_i\psi_i$. Here, if the projection appears 1 dimensional (in contrast to having scattered points on a plane like in Figure 2.1) then $\lambda_i\psi_i$ must be a function of $\lambda_1\psi_1$ and thus not independent. However, this test becomes troublesome once we move to the cases that need higher dimensional DMAPs. We refer to [202] for further discussion.

In this thesis, we simply continue the iMapD's original article [59] where they rely on a 2D DMAP space. In our applications, we identified that to this aim the first two dominant eigenvectors remained sufficiently independent.

Extension beyond the boundary: After establishing the set $\mathcal{C}^{\text{B}}$, iMapD proceeds with a "shooting move" whose aim is to generate a new set of configurations beyond the

observed regions of the IM. We first construct for each point $\mathbf{Q}_j^B \in \mathcal{C}^{\mathrm{B}}$ a neighboring set $\mathcal{B}_j$ that contains the nearest configurations based on RMSD. Here, employing PCA enables iMapD to (approximately) navigate the surface in configuration space that is tangent to the local divergence of FEL –hence orthogonal to the kinetic boundary surface (more details on PCA in Appendix A.3). Once the principal components (PCs) for each $\mathcal{B}_j$ are determined, we project both the average conformation

$$\mathbf{Q}_j^{\mathrm{avg}} = \frac{1}{|\mathcal{B}_j|} \sum_{\mathbf{Q}_m \in \mathcal{B}_j} \mathbf{Q}_m$$

and the corresponding boundary point $\mathbf{Q}_j^B$ onto the low dimensional space spanned by the PCs:

$$\begin{aligned} \mathbf{q}_j^{\mathrm{B}} &= \boldsymbol{L}_j \mathbf{Q}_j^{\mathrm{B}} \\ \mathbf{q}_j^{\mathrm{avg}} &= \boldsymbol{L}_j \mathbf{Q}_j^{\mathrm{avg}} \end{aligned} \tag{2.1.1}$$

The $\boldsymbol{L}_j$ is the loading matrix formed by the p-most dominant PCs. Here, the value of p is adaptively determined by using a threshold of how much variance the PCs should retain. This can be mathematically expressed as

$$\frac{\sum_{a=1}^{p} \lambda_a}{\sum_{a=1}^{3N} \lambda_a} \leq \text{threshold} \tag{2.1.2}$$

where λ_a denotes the a-th PCA eigenvalue with $\lambda_{a+1} \leq \lambda_a$. Given the projection in *Equation* (2.1.1), iMapD then "walks" along the direction identified by the PCs to generate a new set of coordinates located beyond the kinetic boundary:

$$\mathbf{q}_j^{\mathrm{new}} = \mathbf{q}_j^{\mathrm{B}} - \mathbf{q}_j^{\mathrm{avg}} + c\,\frac{\mathbf{q}_j^{\mathrm{B}} - \mathbf{q}_j^{\mathrm{avg}}}{|\mathbf{q}_j^{\mathrm{B}} - \mathbf{q}_j^{\mathrm{avg}}|} \tag{2.1.3}$$

In this expression, $c > 0$ is a constant which determines how far from the boundary we are willing to "shoot". Once new coordinates $\mathbf{q}_j^{\mathrm{new}}$ are generated, the Cartesian coordinates can be easily retrieved by the reverse transformation $\mathbf{Q}_j^{\mathrm{new}} = \mathbf{q}_j^{\mathrm{new}} \boldsymbol{L}_j^T + \mathbf{Q}_j^{\mathrm{avg}}$.

The parameter c in Equation (2.1.3) plays a crucial role in iMapD algorithm (as illustrated in Figure 2.2). Large values of this parameter can lead to atomic coordinates $\mathbf{Q}_j^{\mathrm{new}}$ that deviate significantly from chemically viable configurations –violating the molecule's structural constraints. This issue is particularly pronounced in regions where the IM is drastically curved and can lead to systematic errors (see Figure 2.2(b)). To mitigate this risk and potential breakdowns, we are forced to adopt an incremental value of c. However, even with such a conservative choice, obtaining a chemically stable configuration necessitates a corrective step by either running an energy minimization from $\mathbf{Q}_j^{\mathrm{new}}$ or an extra

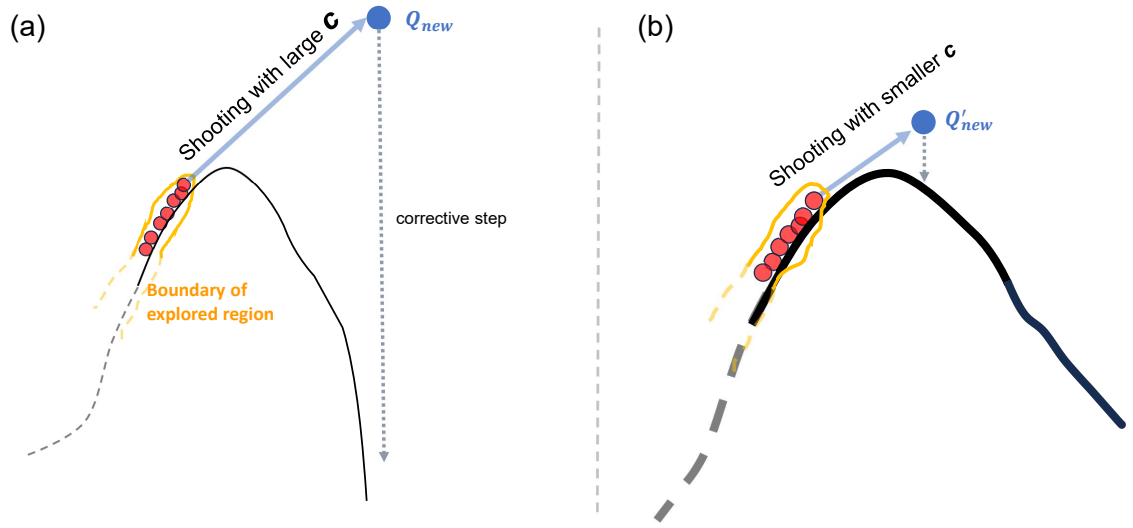


Figure 2.2: Schematic illustration of a 1D manifold depicting iMapD's "shooting move". A large value of c may lead to a set of coordinates significantly distant from the manifold of chemically viable configurations, essentially "breaking" the molecule. This concern is particularly relevant in the regions where the IM experiences a substantial curvature. As a result, even the subsequent corrective step, e.g. in the form of energy minimization, may struggle to restore the correct topology within a finite timeframe.

short burst of unbiased MD. We denoted the resulting chemically stable configurations as $\mathbf{X}_j^{\text{new}}$.

Next iterations of iMapD: Finally, once the $\mathbf{X}_j^{\text{new}}$ are obtained, the exploration is continued by initiating new rounds of short, unbiased MD from these configurations, and merging the data with $\mathcal{C}_{\text{ini}}$. To eliminate any directional bias due to randomness, e.g. when sampling the saddle points of FEL, multiple MDs can be performed with different sets of initial velocities, drawn from the Maxwell-Boltzmann distribution. Next, iMapD iteratively performs the following:

(i) Identifying the boundary of explored regions of IM using the set of configurations sampled in previous iterations.

(ii) Generating new configurations that lie beyond the boundary.

(iii) Restarting the unbiased sampling from these configurations and again merging the new data to the previous set.

The algorithm is terminated once new meta-stable states or transition regions between them have been identified. The overall scheme of iMapD is illustrated in a schematic representation in Figure 2.3.

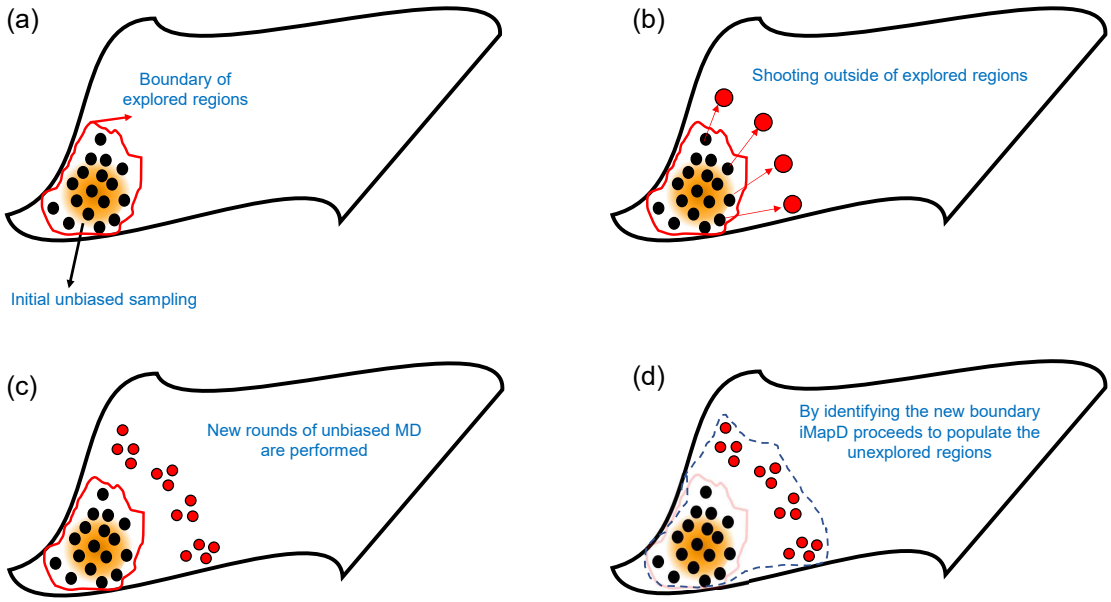


Figure 2.3: A schematic overview of iMapD algorithm. First, utilizing DMAP we parameterize the underlying manifold of previously generated data. By identifying the kinetic boundary in this manifold and subsequently generating configurations beyond that, iMapD guides the MD simulation in exploring new areas of the IM.

2.1.2 Discussion on the application of iMapD

The framework outlined above achieves drastic acceleration in the sampling of statistically relevant molecular conformations. To showcase the power of iMapD, Chiavazzo *et al.*(2017) [59], provide a comparison with the result of a ms-long unbiased MD, investigating the dynamics and configuration space of the transmembrane helices of Mga2 dimers. Mga2 is a protein in the *endoplasmic reticulum* of baker's yeast cell that acts as the bilayer lipid packing sensor [203]. Using a coarse-grained (CG) force field, Covino *et al.* [204] had performed more than 3 milliseconds of MD to characterize the rotational dynamics of the helices in response to changes in saturation. However, the dimers remained in contact form throughout their simulation. In remarkable contrast, iMapD (with the same CG setup) was not only able to recover the data of unbiased sampling but also generate multiple dimer dissociation events, in just 10s of microseconds cumulative simulation time.

In the context of our new framework, iMapD offers an important enhancement: By identifying the kinetic boundary and tracing the geometry of the of FEL in the shooting move, iMapD increases the likelihood of observing (capturing snapshots of the system in) all transition channels with comparable free energy. Moreover, due to its inherent (high) parallelizability, one can initiate multiple shooting moves with different values of c to achieve a more robust sampling as the exploration takes place. This flexibility may facilitate the

exploration of even higher-energy metastable states (e.g. unfolded states of protein) and transition regions. Despite these promising capabilities, it is essential not to overlook the potential challenges and limitations of the algorithm.

Up to this point, we have not discussed how one could infer thermodynamic or kinetic information from the iMapD data. Notably, the deliberate breaking of the detailed balance to accelerate the sampling with iMapD does not allow the final dataset to be distributed according to the Boltzmann weights of each configuration. Consequently, extracting thermodynamic and/or kinetic information by averaging the quantities is not feasible and necessitates further simulations and analysis. Following this challenge, the authors [59] proposed the application of methods such as Umbrella sampling to map out and interpolate the FEL between configurations of $\mathcal{C}_{\text{fin}}$. Moreover, MSMs were suggested for evaluating the kinetic rates between identified metastable states. In the next section, we develop an alternative method based on a rigorous mathematical formulation that enables the sampling of low-free energy transition pathways between the metastable states. This method not only lacks further extensive rate calculations –like in MSM– but it can also be done, in principle, in parallel to iMapDs' samplings thus requiring no additional simulations.

Another aspect that requires our attention is the impact of the parameter c in the shooting move. As previously highlighted, careful considerations and preferably adopting incremental values for c are essential to maintain the stability of the iMapD algorithm. However, such restriction may impede the efficiency of the algorithm, specifically when exploring the transition regions between distant metastable states. To circumvent this issue, we have proposed a modification to the shooting move in Chapter 4. This new scheme while stabilizing the algorithm and allowing us to adopt larger values of c, still retains the original "unsupervised" nature of iMapD.

For the sake of completeness, we revert our attention to the choice of dimension in DMAP. In a general application, the dimension of the underlying manifold is not expected to necessarily remain constant as the exploration progresses. For instance, in the folding of a protein, the dimension of the folded state may be assumed to be $d = 2$ resembling a flat surface. However, as the exploration proceeds it may encounter a transition region with $d = 1$, analogous to a river in a narrow ravine between two mountains. Finally, in the denatured state, the dimensions of the IM might increase ($d \geq 2$) due to an increase in entropy. While one possible approach is to take the largest dimension (the unfolded state in this example) as the definitive dimension for the DMAP application, this assumes retaining such information before applying iMapD, limiting the adaptability of the algorithm.

In this thesis, we did not explicitly address this issue and we refer to [205] for further discussion on determining the dimensionality of the low-dimensional embeddings. However, concerning our application of iMapD to the benchmark system in Chapter 3, it is widely

accepted that the FEL of Alanine dipeptide can be characterized by its 2 CVs, the dihedral angles ϕ and ψ [206]. Therefore, a 2D underlying manifold can be assumed for the application of iMapD in this system. On the other hand for the second application in Chapter 4, we adopted the first 2 dominant DCs, as it is computationally most efficient.

2.2 Building the transition network

The main challenge in iMapD concerning the identification of physical transition pathways was the lack of detailed balance condition between the configurations in the final dataset. In this section, our objective is to construct a stochastic theory that can characterize transitions and dynamics on the IM, while also adhering to the correct form of microscopic reversibility.

To achieve this, we commence by selecting a subset of ν configurations $\mathcal{S} = \{\mathbf{Q}_k^{(s)}\}_{k=1,\dots,\nu}$ out of $\mathcal{C}_{\text{fin}}$. The main criterion here is to maintain a distribution as uniformly as possible with respect to the RMSD distance. The selection can be achieved through simple structural clustering, e.g. KMeans [107], or even more sophisticated techniques. Regardless of the approach, the key is to associate to each configuration $\mathbf{Q}_k^{(s)} \in \mathcal{S}$ a representative region of the configuration space $\boldsymbol{R}_k$ with the size $r = \sigma$. Adopting σ equal to half the average RMSD distance between the nearest-neighbors of $\mathcal{S}$ ensures that the union of all the finite-sized regions $\mathcal{R} = \cup_k \boldsymbol{R}_k$ covers the entire explored portion of the IM. We postpone the exact and in-detail implementation of this step (illustrated in Figure 2.4) to the following chapters.

Having established the partitioning $\mathcal{R}$, we proceed to develop an effective theory by smearing the spatial scale of the microscopic dynamics up to the distance σ (size of individual regions). Here, we employ the powerful formalism of the (stochastic) path integral approach and the renormalization group theory, originally developed in the context of nuclear and subnuclear physics [207]. In this CG representation, the maximum time resolution Δt would accordingly be the average time it takes for the system to diffuse, under the influence of microscopic dynamics, a distance σ in the configuration space.

Therefore our theory has the capacity to characterize the evolution of the system's trajectory as it jumps/transits between the finite space sub-regions of $\mathcal{R}$. Moreover, due to its formal equivalency to the microscopic theory, it inherently follows a detailed balance condition as we will explain.

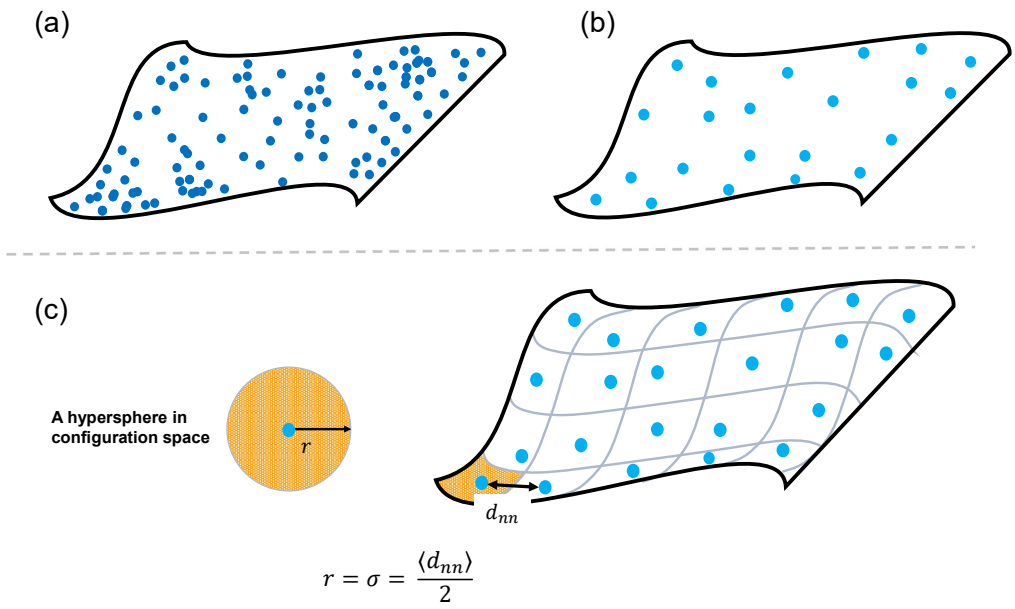


Figure 2.4: Illustration of how to utilize $\mathcal{C}_{\text{fin}}$ to construct regions of the configurations space that partition the IM. (a) The final data set of iMapD, $\mathcal{C}_{\text{fin}}$. (b) By performing a simple clustering, a subset $\mathcal{S}$ is extracted. The configurations of $\mathcal{S}$ are expected to be distributed as uniformly as possible with respect to RMSD. Each configuration is then associated with a region $\boldsymbol{R}_k$ of size $r = \sigma$ in the configuration space. (c) By choosing σ equal to half the average distance between nearest neighbors of $\mathcal{S}$ the union of finite space regions effectively covers the explored portion of the IM.

2.2.1 Coarse-grained dynamics of transitions

Let us begin by assuming that a set of over-damped Langevin equations govern microscopically the structural dynamics of the molecule:

$$\dot{\mathbf{Q}}(t) = \frac{-D}{k_{\text{B}}T} \nabla U(\mathbf{Q}(t)) + \eta(t) \tag{2.2.1}$$

The over-damped limit has been shown to be an appropriate assumption for time resolutions of picoseconds or lower, where the conformational changes of (bio-)macromolecules occur [208]. In this expression $\mathbf{Q} = (\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_N)$ represents a point in the configuration space with $\mathbf{Q}_i = (\mathbf{Q}_{(i)\,x}, \mathbf{Q}_{(i)\,y}, \mathbf{Q}_{(i)\,z})$ and N the number of atoms. The $\dot{\mathbf{Q}}$ denotes the velocity vector of this point, the $U(\mathbf{Q})$ is the underlying potential energy function, and the $D = \frac{k_{\text{B}}T}{m\gamma}$ is the diffusion coefficient, with m as the mass of each atom, γ as the viscosity (assumed for simplicity to be uniform and isotropic), and $k_{\text{B}}T$ as the Boltzmann factor and the temperature respectively. Finally, the $\eta(t)$ is a vector of null average Gaussian noises whose components follow the fluctuation-dissipation theorem:

$$\langle \eta_{(i)a}(t) \eta_{(j)b}(t') \rangle = 2D\delta_{ab}\delta_{ij}\delta(t - t') \tag{2.2.2}$$

with $\eta_{(i)a}$ being the element of vector pertaining to $\mathbf{Q}_{(i)a}$.

The stochastic differential Equation (2.2.1) leads to a probability distribution, which obeys the well-known Fokker-Planck equation (FP):

$$\begin{aligned}\frac{\partial P(\mathbf{Q},t)}{\partial t} &= D\nabla.\left(\nabla + \frac{1}{k_{\mathrm{B}}T}\nabla U\right)P(\mathbf{Q},t)\\ &= -H_{FP}P(\mathbf{Q},t)\end{aligned} \tag{2.2.3}$$

and its Green's function, the conditional probability $P(\mathbf{Q},t|\mathbf{Q}_i)$, to observe the system in configuration $\mathbf{Q}$ at time t subjected to $\mathbf{Q}(0)=\mathbf{Q}_i$, accordingly follows

$$\frac{\partial P(\mathbf{Q},t|\mathbf{Q}_i)}{\partial t} + H_{FP}P(\mathbf{Q},t|\mathbf{Q}_i) = \delta(\mathbf{Q}-\mathbf{Q}_i)\delta(t) \tag{2.2.4}$$

It is convenient to substitute $P(\mathbf{Q},t) = e^{-(U/2k_{\mathrm{B}}T)}\Psi(\mathbf{Q},t)$ in Equation (2.2.3). This recast the expression into a form equivalent to a Schrodinger equation in imaginary time:

$$-\frac{\partial\Psi(\mathbf{Q},t)}{\partial t} = H_{\mathrm{eff}}\Psi(\mathbf{Q},t) \tag{2.2.5}$$

where the effective "Quantum Hamiltonian" operator reads

$$H_{\mathrm{eff}} = -D\nabla^2 + \mathrm{V}_{\mathrm{eff}} \tag{2.2.6}$$

with the effective potential

$$\mathrm{V}_{\mathrm{eff}} = \frac{D}{4(k_{\mathrm{B}}T)^2}\left(|\nabla U|^2 - 2k_{\mathrm{B}}T\nabla^2 U\right) \tag{2.2.7}$$

Given this analogy to quantum mechanics, one can utilize the states $|Q\rangle$ in the $3N$ Hilbert space of the theory –$P(\mathbf{Q},t) = e^-\langle Q|\Psi(t)\rangle$–, to easily express the conditional probability $P(\mathbf{Q},t|\mathbf{Q}_i)$ in the form of a Feynman path integral

$$P(\mathbf{Q}_f,t|\mathbf{Q}_i) = e^{-\frac{1}{2k_{\mathrm{B}}T}\left(U(\mathbf{Q}_f)-U(\mathbf{Q}_i)\right)}\langle\mathbf{Q}_f|e^{-tH_{\mathrm{eff}}}|\mathbf{Q}_i\rangle \tag{2.2.8}$$

where next we perform a Trotter decomposition of the propagator, leading to

$$\begin{aligned}\mathcal{K}(\mathbf{Q}_f,t|\mathbf{Q}_i) &= \langle\mathbf{Q}_f|e^{-tH_{\mathrm{eff}}}|\mathbf{Q}_i\rangle\\ &= \int\left[\prod_{k=0}^{M_t-1} d\mathbf{Q}_K\right]\langle\mathbf{Q}_{k+1}|e^{-H_{\mathrm{eff}}dt}|\mathbf{Q}_k\rangle\,\delta(\mathbf{Q}_0-\mathbf{Q}_i)\,\delta(\mathbf{Q}_{M_t}-\mathbf{Q}_f)\\ &= \mathcal{N}\int\mathcal{D}[\mathbf{Q}]e^{-S_{\mathrm{OM}}}\end{aligned} \tag{2.2.9}$$

Here, we have divided the total time t into M_t differential steps, and

$$S_{\mathrm{OM}} = \int_0^t d\tau \left(\frac{\dot{\mathbf{Q}}^2}{4D} + \mathrm{V}_{\mathrm{eff}}[\mathbf{Q}(\tau)] \right) \tag{2.2.10}$$

denotes the so-called Onsager-Machlup action. The $\mathcal{N}$ is an irrelevant factor that appears due to the change of variables from η_i (Gaussian noises) to $\mathbf{Q}_i$ and essentially ensures the normalization of $\int d\mathbf{Q}\ P(\mathbf{Q}, t|\mathbf{Q}_i) = 1$.

Next, to lower the spatial resolution of this stochastic theory, we define a set of so-called CG states:

$$|Q\} \equiv \int d\mathbf{Q}'\ \phi(\mathbf{Q}' - \mathbf{Q})|Q'\rangle \tag{2.2.11}$$

The $\phi(\mathbf{Q})$ denotes a fast-decaying and smooth smearing function centered around 0, such that $|\phi(\mathbf{Q})| \xrightarrow{|\mathbf{Q}|\gg\sigma} 0$. In the framework of the renormalization group theory, smearing the position eigenstates to the scale σ is equivalent to lowering the spatial resolution by filtering out large momentum states, with $\boldsymbol{K} \gtrsim 1/\sigma$. This procedure which is commonly referred to as "regularization", can be conveniently achieved by adopting

$$\tilde{\phi}(\boldsymbol{K}) = e^{\frac{-\sigma^2 \boldsymbol{K}^2}{4}} \tag{2.2.12}$$

for the Fourier transform of $\phi(\mathbf{Q})$. The factor $1/4$ is adopted for the sake of simple normalization factors in the following equations. Consequently, in the configuration space we obtain

$$\phi(\mathbf{Q} - \mathbf{Q}') = \frac{e^{-\frac{(\mathbf{Q}-\mathbf{Q}')^2}{\sigma^2}}}{(\sigma\sqrt{\pi})^{3N}} \tag{2.2.13}$$

Finally with this choice of $\phi(\mathbf{Q})$, the dot product

$$\begin{aligned} \{Q|Q'\} &= \int d\mathbf{Q}''\phi(\mathbf{Q} - \mathbf{Q}'')\phi(\mathbf{Q}' - \mathbf{Q}'') \\ &= \frac{e^{-\frac{(\mathbf{Q}-\mathbf{Q}')^2}{2\sigma^2}}}{(\sigma\sqrt{2\pi})^{3N}} = \delta_\sigma(\mathbf{Q} - \mathbf{Q}') \end{aligned} \tag{2.2.14}$$

exhibits a non-vanishing overlap between CG states, where we have adopted $\delta_\sigma(\mathbf{Q}-\mathbf{Q}')$ as a modified (smeared) Dirac delta function. In practice, the CG states $|\mathbf{Q}_k\}$ here are associated with the configurations $\mathbf{Q}_k$ in $\mathcal{S}$ and henceforth define the sub-regions $\boldsymbol{R}_k$ surrounding these configurations.

Once the CG states are established, we derive the Feynman propagator of our theory which corresponds to evaluating the Equation (2.2.8) however this time as a path integral performed over the CG states $|Q\}$. This reads

$$\mathcal{K}_{\mathrm{cg}}(\mathbf{Q}_f, t|\mathbf{Q}_i) = \{\mathbf{Q}_f|\hat{U}(t)|\mathbf{Q}_i\}, \tag{2.2.15}$$

where $\hat{U}(t) = e^{-t\hat{H}_{eff}}$. Following the standard derivation in Equation (2.2.9), we proceed by carrying out a new Trotter decomposition:

$$\mathcal{K}_{\rm cg}(\mathbf{Q}_f, t|\mathbf{Q}_i) = \int \mathrm{d}\mathbf{Q}_1 \dots d\mathbf{Q}_{N_t-1} \prod_{n=0}^{N_t-1} \left[\{\mathbf{Q}_{n+1}|\hat{U}(\Delta t)|\mathbf{Q}_n\}\right] \tag{2.2.16}$$

where $\mathbf{Q}_0 = \mathbf{Q}_i$, $\mathbf{Q}_{N_t} = \mathbf{Q}_f$ and the new time steps $\Delta t = \frac{t}{N_t}$ are of the order of the time resolution of the CG theory.

Let us focus on the elementary propagator $\{\mathbf{Q}_{n+1}|\hat{U}(\Delta t)|\mathbf{Q}_n\}$. As usual, by choosing a sufficiently small discretization time, we can ignore commutations between kinetic and potential operators and factorize the evolution operator

$$\hat{U}(\Delta t) \simeq e^{-\hat{T}_{\rm eff}\Delta t} e^{-\hat{V}_{\rm eff}\Delta t}, \tag{2.2.17}$$

where corrections are of higher order in Δt. Using Equation (2.2.11), we obtain

$$\{\mathbf{Q}_{n+1}|U(\Delta t)|\mathbf{Q}_n\} = \int d\boldsymbol{z}\, d\boldsymbol{z}' \, \frac{d\boldsymbol{K}}{(2\pi)^{3N}} \\ e^{-D\Delta t \boldsymbol{K}^2} e^{i\boldsymbol{K}\cdot(\boldsymbol{z}'-\boldsymbol{z})} \phi(\mathbf{Q}_{n+1} - \boldsymbol{z}')\phi(\mathbf{Q}_n - \boldsymbol{z}) e^{-\Delta t \mathrm{V}_{\rm eff}(\boldsymbol{z})}\,. \tag{2.2.18}$$

which after taking the intergal on the momentum $\boldsymbol{K}$ it leads to

$$\{\mathbf{Q}_{n+1}|U(\Delta t)|\mathbf{Q}_n\} \propto \int d\boldsymbol{z}\, d\boldsymbol{z}' e^{-\frac{(z-z')^2}{4D\Delta t} - \frac{1}{\sigma^2}\left((\mathbf{Q}_{n+1}-\boldsymbol{z}')^2 + (\mathbf{Q}_n - \boldsymbol{z})^2\right)} e^{-\Delta t \mathrm{V}_{\rm eff}(\boldsymbol{z})} \tag{2.2.19}$$

Here, we recall that Δt is by definition the smallest time scale of our effective theory. Then we posit that the Gaussian factor $e^{-\frac{m(z-z')^2}{4D\Delta t}}$ remains finite only when the distance between $\boldsymbol{z}$ and $\boldsymbol{z}'$ lies within the spatial resolution of the CG theory, i.e $\sim \sigma$. Consequently, it can be effectively regarded as a Dirac's delta function. Utilizing this assumption, we can carry out the integral in $d\boldsymbol{z}'$ in Equation (2.2.19) and obtain:

$$\{\mathbf{Q}_{n+1}|U(\Delta t)|\mathbf{Q}_n\} \simeq \mathcal{N} e^{-\frac{1}{2\sigma^2}(\mathbf{Q}_{n+1}-\mathbf{Q}_n)^2} \int d\boldsymbol{z}\, e^{-\frac{2}{\sigma^2}\left(\boldsymbol{z} - \frac{\mathbf{Q}_{n+1}+\mathbf{Q}_n}{2}\right)^2} e^{-\Delta t \mathrm{V}_{\rm eff}(\boldsymbol{z})} \tag{2.2.20}$$

where $\mathcal{N}$ is an irrelevant factor. We can cast this expression in a more familiar form by introducing a smeared effective potential defined as the following:

$$e^{-\Delta t V_{\rm cg}(\boldsymbol{X})} \equiv \int d\boldsymbol{z}\, e^{-\frac{2}{\sigma^2}(\boldsymbol{z}-\boldsymbol{X})^2} e^{-\Delta t \mathrm{V}_{\rm eff}(\boldsymbol{z})}. \tag{2.2.21}$$

We emphasize that the average in the right-hand side is dominated by the configurations with the lowest effective potential. From the definition of $\mathrm{V}_{\rm eff}$ in Equation (2.2.7), it follows that these are configurations near mechanical equilibrium points, where the modulus of the

total force $|\nabla U|$ is very small and the Laplacian of the potential energy is positive [209]. Finally, combining all terms, the elementary time propagator becomes

$$\mathcal{K}_{\text{cg}} = \{\mathbf{Q}_{n+1}|\hat{U}(\Delta t)|\mathbf{Q}_n\} \propto e^{-\frac{1}{2\sigma^2}(\mathbf{Q}_{n+1}-\mathbf{Q}_n)^2 - V_{\text{cg}}\left(\frac{\mathbf{Q}_n+\mathbf{Q}_{n+1}}{2}\right)\Delta t}. \tag{2.2.22}$$

Now, we first note that $\mathcal{K}_{\text{cg}}$ vanishes exponentially when the distance between $\mathbf{Q}_n$ and $\mathbf{Q}_{n+1}$ is larger than σ in $e^{-\frac{1}{2\sigma^2}(\mathbf{Q}_{n+1}-\mathbf{Q}_n)^2}$. Furthermore, any structure of $V_{\text{cg}}(Q)$ below the short scale σ is smeared out due to the averaging involved in Equation (2.2.21). Therefore, we may use the approximation $V_{\text{cg}}\left(\frac{\mathbf{Q}_n+\mathbf{Q}_{n+1}}{2}\right) \simeq V_{\text{cg}}\left(\mathbf{Q}_n\right)$, which yields

$$\mathcal{K}_{\text{cg}} \propto e^{-\frac{1}{2\sigma^2}(\mathbf{Q}_{n+1}-\mathbf{Q}_n)^2 - V_{\text{cg}}(\mathbf{Q}_n)\Delta t}. \tag{2.2.23}$$

Equation (2.2.23) qualitatively resembles the structure of the elementary propagator in the microscopic theory in Equation (2.2.9),

$$\langle x_{n+1}|\hat{U}(dt)|x_n\rangle \propto e^{-\frac{1}{4Ddt}(x_{n+1}-x_n)^2 - \text{V}_{\text{eff}}(x_n)dt}. \tag{2.2.24}$$

where for the sake of distinction we have adopted the form $|x\rangle$ as the microscopic states. However, it is important to note that the time discretization step dt in the microscopic theory is generally orders of magnitude smaller than that of the CG theory, Δt.

The CG propagator Equation (2.2.23) can be cast in a form that is completely analog to the microscopic counterpart Equation (2.2.24) by introducing the effective diffusion coefficient of the CG theory D_{cg}:

$$D_{\text{cg}} = \sigma^2/(2\Delta t), \tag{2.2.25}$$

defined in terms of the spatiotemporal resolution scales. After this substitution, we obtain

$$\{\mathbf{Q}_{n+1}|\hat{U}(\Delta t)|\mathbf{Q}_n\} \propto e^{-\frac{1}{4D_{\text{cg}}\Delta t}(\mathbf{Q}_{n+1}-\mathbf{Q}_n)^2 - V_{\text{cg}}(\mathbf{Q}_n)\Delta t} \tag{2.2.26}$$

By plugging back this expression into Equation (2.2.16), we finally obtain

$$\mathcal{K}_{\text{cg}}(\mathbf{Q}_f, t|\mathbf{Q}_i) = \mathcal{N}\int d\mathbf{Q}_1 \dots d\mathbf{Q}_{N_t-1}\, e^{-\sum_{k=1}^{N_t}\frac{(\mathbf{Q}_{k+1}-\mathbf{Q}_k)^2}{4D_{\text{cg}}\Delta t} + V_{\text{cg}}[\mathbf{Q}_k]\Delta t} \tag{2.2.27}$$

We note that the formal connection between the microscopic description and the CG theory also provides a practical approach to evaluate the V_{cg} without exactly calculating the integral Equation (2.2.21). First, it is instructive to compute in microscopic theory, the probability for the system –initially at some given position x– to remain within an

infinitesimal volume after an infinitesimal time interval dt. To remove the dependence on normalization factors, we conveniently compute this probability relative to the same quantity in the purely diffusive limit (i.e. for free Brownian motion). Using the Equation (2.2.24) we obtain:

$$R(x, dt) \equiv \frac{\langle x|e^{-\left(\hat{T}_{\text{eff}}+\hat{\text{V}_{\text{eff}}}\right)dt}|x\rangle}{\langle x|e^{-\hat{T}_{\text{eff}}dt}|x\rangle} = e^{-\text{V}_{\text{eff}}(x)dt}. \tag{2.2.28}$$

Thus, we can identify the $\text{V}_{\text{eff}}(x)$ as the escape rate from an infinitesimal volume centered around x. Performing the same calculation in the CG theory –using Equation (2.2.26)– leads to:

$$R(Q, \Delta t) \equiv \frac{\{Q|e^{-\left(\hat{T}_{\text{eff}}+\hat{V}_{\text{eff}}\Delta t\right)}|Q\}}{\{Q|e^{-\hat{T}_{\text{eff}}\Delta t}|Q\}} = e^{-V_{\text{cg}}(Q)\Delta t}. \tag{2.2.29}$$

Here, the $V_{\text{cg}}(Q)$ can be interpreted as the rate of escape (analog to the microscopic version) from a region $\boldsymbol{R}_k$ around the configuration $\mathbf{Q}_k$, whose radius is the smallest spatial scale of the CG theory. In practice, this rate of escape can be calculated for example by using short MD simulations in iMapD and calculating the average first passage time T^{σ}_{avg} of the trajectory a distance $\geq \sigma$ from $\mathbf{Q}_k$.

2.2.2 Hamilton-Jacobi formulation of the coarse-grained theory

The elementary CG propagator, Equation (2.2.26), is reached by assuming that the spatial scale of the theory adheres to the size of regions in $\mathcal{R}$. Accordingly, we posited that the highest temporal resolution must be the system's average escape time $\Delta t = \langle t_{\text{esc.}} \rangle$ from these regions. Alternatively, we can enforce this limit by directly regularizing the temporal scale. To this aim, we focus on the Laplace transform of the propagator of the microscopic theory Equation (2.2.9)

$$\boldsymbol{G}(\mathbf{Q}_k, s|\mathbf{Q}_j) = \int_0^{\infty} d\tau e^{-s\tau} \boldsymbol{K}(\mathbf{Q}_k, \tau|\mathbf{Q}_j) \tag{2.2.30}$$

This expression essentially describes the probability of transitioning between $\mathbf{Q}_j$ and $\mathbf{Q}_k$ in the frequency space with the characteristic frequency s (not considering the irrelevant prefactor in Equation (2.2.8)). By adopting the correct frequency cut-off and transforming it back into time coordinates, we expect to arrive at the CG propagator in Equation (2.2.29). We now re-direct our attention to the transitions between configurations whose pairwise distance is $\sim 2\sigma$, or equivalently, from one region to its immediate neighbor $\boldsymbol{R}_k, \boldsymbol{R}_j \in \mathcal{R}$. Here, we assume σ to be sufficiently small that the majority of transitions occur along the most dominant pathway $\bar{\mathbf{Q}}$ between $\mathbf{Q}_k$ and $\mathbf{Q}_j$ –associated to the respective regions.

According to Dominant Reaction Pathway (DRP) formalism [197, 210] (briefly reviewed in Appendix A.1), the $\bar{\mathbf{Q}}$ is identified using the saddle-point approximation on the prop-

agator $\boldsymbol{K}(\mathbf{Q}_k, \tau|\mathbf{Q}_j)$. This approximation yields a Newton-type equation of motion along the DRP

$$\frac{\ddot{\bar{\mathbf{Q}}}}{2D} = \nabla \mathrm{V}_{\mathrm{eff}}(\bar{\mathbf{Q}}), \tag{2.2.31}$$

where $\bar{\mathbf{Q}}(\tau)$ obeys the boundary conditions $\mathbf{Q}(0) = \mathbf{Q}_j$ and $\mathbf{Q}(t) = \mathbf{Q}_k$. This implies the conservation of the effective energy

$$E_{\mathrm{eff}} = \frac{\dot{\mathbf{Q}}^2}{4D} - \mathrm{V}_{\mathrm{eff}}(\mathbf{Q}) \tag{2.2.32}$$

along the path $\bar{\mathbf{Q}}(\tau)$. In turn, the conservation of effective energy allows us to equivalently express the DRP with an action that is independent of time. Using the Hamilton-Jacobi (HJ) theory, the most probable path between $\mathbf{Q}_k$ and $\mathbf{Q}_j$ is the one that minimizes the action

$$\boldsymbol{K}(\mathbf{Q}_k, \tau|\mathbf{Q}_j) \sim e^{E_{\mathrm{eff}}\tau - S_{\mathrm{HJ}}[\bar{\mathbf{Q}}]}, \tag{2.2.33}$$

where $S_{\mathrm{HJ}}[\bar{\mathbf{Q}}]$ is the so-called Hamilton–Jacobi (HJ) functional,

$$S_{\mathrm{HJ}}[\bar{\mathbf{Q}}] = \int_{\mathbf{Q}_j}^{\mathbf{Q}_k} dl \sqrt{D\left(E_{\mathrm{eff}} + \mathrm{V}_{\mathrm{eff}}[\bar{\mathbf{Q}}(l)]\right)}, \tag{2.2.34}$$

and dl is the Euclidean distance travelled along the trajectory $\bar{\mathbf{Q}}$. Putting this back into Equation (2.2.30) we get

$$\boldsymbol{G}(\mathbf{Q}_j, s|\mathbf{Q}_j) = \mathcal{N} \int_0^\infty d\tau\, e^{(E_{\mathrm{eff}} - s)\tau}\, e^{-\int_{\mathbf{Q}_j}^{\mathbf{Q}_k} dl \sqrt{D\left(E_{\mathrm{eff}} + \mathrm{V}_{\mathrm{eff}}[\bar{\mathbf{Q}}(l)]\right)}} \tag{2.2.35}$$

where $\mathcal{N}$ is an irrelevant prefactor.

Even though the effective energy along the DRP is conserved, it still depends on the overall time that the path would take, $E_{\mathrm{eff}} = E_{\mathrm{eff}}(\tau)$. Since our focus is on the transitions with spatial distance 2σ, we expect the integral in Equation (2.2.35) to be dominated by the saddle point $\bar{t}$ where $E_{\mathrm{eff}}(\bar{t}) \simeq s$. This approximation leads to

$$\boldsymbol{G}(\mathbf{Q}_k, s|\mathbf{Q}_j) \approx \mathcal{N} e^{-\int_{\mathbf{Q}_j}^{\mathbf{Q}_k} dl \sqrt{D\left(s + \mathrm{V}_{\mathrm{eff}}[\bar{\mathbf{Q}}(l)]\right)}} \tag{2.2.36}$$

Conversely, using this expression in DRP theory, the time taken by the most probable path $\bar{\mathbf{Q}}(l)$ can be written as:

$$t_{j,k} = \int_{\mathbf{Q}_j}^{\mathbf{Q}_j} \frac{dl}{\sqrt{4D\left(\mathrm{V}_{\mathrm{eff}}[\bar{\mathbf{Q}}(l)] + s\right)}}. \tag{2.2.37}$$

In the context of the CG theory, we expect the average transition times $\langle t_{j,k} \rangle$ to be on the order of the temporal scale of the theory Δt. Therefore, one can obtain the time of transition between $\{\boldsymbol{R}_k, \boldsymbol{R}_j\}$ by assuming a maximum cut-off frequency $s_0 \sim 1/\Delta t = 1/\langle t_{j,k} \rangle$ in Equation (2.2.36).

Now, we return to the representation of the reactive pathways in the configuration space using the partitions $\mathcal{R}$. In this viewpoint, the transitions correspond to crossing in order a given sequence of finite regions

$$R_{i_0} \to R_{i_2} \to \ldots \to R_{i_N}$$

while going from the region i_0 associated to the reactant to the i_N associated to the product (Figure 2.4). Hence, a transition path in the coarse-grained theory is specified by the integer vector $\mathbf{I} = (i_1, \ldots, i_{N_I})$, where i_k is the region visited at step k. The expression for the conditional probability to perform a transition from i_0 to i_N using the effective theory in Equation (2.2.29) reads:

$$P_{cg}(i_N, t|i_0) \propto \mathcal{K}_{\mathrm{cg}}(\mathbf{Q}_{i_N}, t|\mathbf{Q}_{i_0}), \tag{2.2.38}$$

where the $\mathbf{Q}_{i_k}$ is the configuration representing the region $\boldsymbol{R}_{i_k}$. However, considering the assumption that the transitions between regions occur along the most dominant path, we can evaluate the relative weight of every transition pathway $\mathbf{I}$ with the HJ formulation of CG theory:

$$P(\mathbf{I}) \propto e^{-\sum_{k=0}^{N-1} \Delta l(i_k, i_{k+1}) \sqrt{\frac{1}{D_{\mathrm{cg}}} \left(V_{\mathrm{cg}}(\mathbf{Q}_{i_k}) + s_0 \right)}} \tag{2.2.39}$$

where the $\Delta l(i_k,\, i_{k+1})$ is the Euclidean distance between the configurations $\mathbf{Q}_{i_k}, \mathbf{Q}_{i_{k+1}}$. Accordingly, the time of transition along such a pathway can be calculated as:

$$t \simeq \sum_{k=0}^{N-1} \Delta l(i_k,\, i_{k+1}) \frac{1}{\sqrt{4 D_{\mathrm{cg}} \left(V_{\mathrm{cg}}(\mathbf{Q}_{i_k}) + s_0 \right)}} \tag{2.2.40}$$

It is important to emphasize that this description of the time along a transition path, does not capture the exponential time spent by the system to explore metastable regions. Thus, it can only be used to estimate a lower bound for the transition path time, where jumping from one region to its neighbor involves crossing only a single free-energy barrier.

This feature represents an important difference between the gTPS theory and the MSM approach since the latter yields a complete representation of the relaxation kinetics. On the other hand, defining a MSM is significantly more computationally expensive, since it involves computing the full transition matrix. However, in our approach, we only require

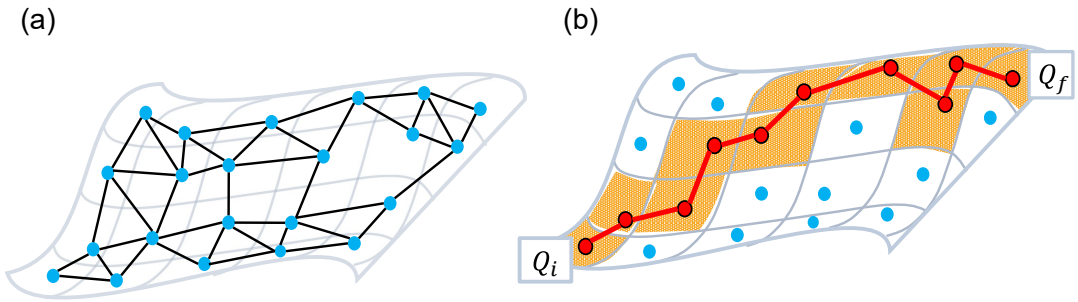


Figure 2.5: (a) A network of transitions connecting the partitions $\mathcal{R}$ on the IM. (b) A possible reactive pathway is depicted that traverses these regions as it transitions from the regions associated with the reactant and product states.

to compute the V_{cg} using the escape time from the finite space regions, which can theoretically be computed simultaneously with the iMapD. We also that corrections to the DRP approximation are only logarithmic in the path's action [211]. Therefore, as long as σ does not exceed atomistic spatial dimensions $\sim 1\mathring{A}$ we do not expect any significant difference between the probability calculated using the HJ action and time-dependent action.

2.2.3 Network of transitions

The HJ formulation above allows us to represent the dynamics of our CG theory using an undirected network that connects the neighboring regions. To this end, each finite region R_{i_k} is assigned to a vertex k, while the edge connecting the vertices i and j is assigned a weight w_{ij} (Figure 2.5):

$$w_{ij} = \frac{|\mathbf{Q}_i - \mathbf{Q}_j|}{2\sqrt{D_{cg}}}(L_i + L_j), \quad L_i = \sqrt{V_{\text{cg}}(\mathbf{Q}_i) + s_0} \tag{2.2.41}$$

This way, the probability of a given coarse-grained path $\mathbf{I}$ is evaluated by the negative exponent of the sum of the weights of all the edges forming the path I:

$$P(\mathbf{I}) \propto e^{-W(\mathbf{I})} = e^{\sum_{\{i,j\}\in\mathbf{I}} w_{i,j}} \tag{2.2.42}$$

In these expressions, we have adopted a trapezoidal rule for the discretization of the integral along the paths.

2.3 Sampling transition pathways

The undirected graph of finite-size regions $\mathcal{R}$ empowered with the CG description of the dynamics, provides an ensemble of discretized transition pathways between the reactant

and product states. To sample this ensemble, we posited in Chapter 1 that a QA could be utilized to generate "uncorrelated" pathways. Our argument was that the existence of inherent quantum fluctuations in these devices, leads to de-correlation in the sampling every time we initiate the annealing procedure. Of course, this is guaranteed only when the number of possible pathways that are also significantly different is sufficiently large.

To demonstrate this in the following, we first utilize QUBO mathematical formulation to encode both the network and its transition pathways into the D-Wave annealer. Then, we integrate the D-Wave sampling step into an MC process implemented on a classical computer. Here, we do not require a fully fair sampling of the space of possible paths, which is one of the challenges in quantum annealer-based sampling [212–214]. Rather, we employ a suitable reweighting procedure to achieve the correct detailed balance condition, while we explore the space of accessible states.

2.3.1 Quantum mechanical encoding of discrete-TPE

Following the procedure for quantum annealing in Section 1.3.1, we need to formulate the problem of sampling from the discrete TPE into an Ising Hamiltonian suitable for D-Wave annealer. Such a function essentially should gauge all the pathways in the ensemble according to their statistical weights, i.e. the more probable paths are represented by the lower energy eigenstates. To this end, we present here a QUBO formulation that exactly performs this ranking by searching for the "optimal path" (equivalently the shortest path) between any two vertices of an undirected network. We note that a similar formulation for the case of a graph with directed edges can also be written, hence, all the following procedures are equally applicable to alternative approaches –e.g. to encode networks produced by MSMs.

Let us begin by denoting the vertices and the edges of the network with two sets of binary variables, Γ_i and $\Gamma_{i,j}$, respectively. Here both i and j run over all the vertices of the graph. In this representation, if $\Gamma_i = 1$ ($\Gamma_i = 0$), then the i-th vertex is (is not) visited by the transition path on the graph. Γ_{ij} is always 0 if the i and j are not adjacent in the graph. If i and j are adjacent, then $\Gamma_{i,j} = 1$ when the path contains the $i \to j$ or $j \to i$ transition. We are specifically interested in configurations of the binary variables in which the set of non-vanishing entries of Γ_i and Γ_{ij} form a topologically connected path, i.e., a continuous line starting from the given initial vertex and terminating in the chosen final vertex.

To generate configurations according to $e^{-W(\mathbf{I})}$ in Equation (2.2.39), we consider the following function:

$$H_{\mathrm{QUBO}} = \alpha H_{\mathrm{C}} + H_{\mathrm{T}}. \tag{2.3.1}$$

H_{C} is the constraint Hamiltonian, a positive semi-definite function that is at its minimum $H_{\mathrm{C}}(\Gamma_i, \Gamma_{i,j}) = 0$ only if the entries of the binary variables satisfy the path topology. What we mean by the correct topology condition can be fulfilled by choosing [215]:

$$H_{\mathrm{C}} = H_{\mathrm{s}} + H_{\mathrm{t}} + H_{\mathrm{r}}, \tag{2.3.2}$$

where

$$\begin{aligned} H_{\mathrm{s}} &= 1 - (\Gamma_s)^2 + \left(\Gamma_s - \sum_k \Gamma_{sk}\right)^2, \\ H_{\mathrm{t}} &= 1 - (\Gamma_t)^2 + \left(\Gamma_t - \sum_k \Gamma_{tk}\right)^2, \\ H_{\mathrm{r}} &= \sum_{j \neq s,t} \left(2\Gamma_j - \sum_i \Gamma_{j,i}\right)^2. \end{aligned} \tag{2.3.3}$$

In this formulation, H_{s} and H_{t} enforce the topology related to the source s and target t vertices. We know that in every path between these two, exactly one edge of the path would be incident on either one of them. Then it is easy to see that H_{s} (or H_{t}) is at its minimum when $\Gamma_s = 1$ ($\Gamma_t = 1$) and only one term in the summation is $\Gamma_{s,k'} = 1$ ($\Gamma_{t,k'} = 1$). Conversely, for the rest of vertices along any given path they are required topologically to have one incoming and one outgoing edges. Therefore, the H_{r}, which controls the flux conservation for all the vertices other than s or y, is at its minimum zero only when for every i, $\Gamma_i = 1$ and exactly two variables Γ_{i,k_1} and Γ_{i,k_2} are equal to 1 in the (second) summation. Finally, in Equation (2.3.1) the $H_{\mathrm{T}} = \sum_{ij} w_{ij}\, \Gamma_{ij}$ is the so-called target-function that H_{QUBO} attempts to minimize. By definition, H_{T} yields the path action $W(\mathbf{I})$ whenever the configuration of the variables Γ_i and $\Gamma_{i,j}$ satisfy a path topology, i.e. $H_{\mathrm{C}} = 0$.

The parameter α in Equation (2.3.1) controls the relative strength of the constraint Hamiltonian, H_{C}, against the H_{T}. In the quantum mechanical sense, this parameter can introduce a gap between the energies of H_{QUBO}. Configurations of binary variables that correspond to real paths occupy the states of H_{QUBO} whose maximum energy is $max(H_{\mathrm{T}}) = C_{max}$. The C_{max} denotes the cost of the largest weighted path possible between the source and target vertices. Immediately above this state, resides another one with eigenvalue $= 2\alpha$ that corresponds to all the binary variables equal to zero, s.t. $H_{\mathrm{T}} = 0$ and $H_{\mathrm{C}} = 2$. In order to discourage any optimization algorithm from accessing "not-path-like" configurations while searching for the shortest weighted path, it is necessary to adopt $2\alpha > C_{\mathrm{max}}$. However, we obviously cannot know the C_{max} beforehand.

Separately, caution must be taken while adjusting parameters such as α in realistic annealing machines, e.g. D-Wave. In particular, apart from energetically distinguishing between the states, this parameter has to be sufficiently small since realistic machines are

not able to arbitrarily access large energy scales. This may not be alarming at first as the pre-processing schemes usually implemented in such machines rescales all the energy couplings to match the physical limitation of the device. However, this can introduce unwarranted biases in the annealing process or, concerning our goal of using D-Wave for sampling, obscure accessing certain states who have slightly higher energy than the minimum. Considering both challenges, we heuristically assign $2\alpha = \sum_{i,j} w_{i,j}$ in both applications of this thesis. This guarantees to satisfy $2\alpha > C_{\max}$ condition while not being too much restrictive concerning the relative scale of energies between H_{C} and H_{T}.

Now, in principle to implement the H_{QUBO} into qubits of D-Wave, we need to switch the binary expression first into a generalized Ising model –following the substitutions $S_i = 2\Gamma_i - 1$, $S_{i,j} = 2\Gamma_{i,j} - 1$– and subsequently promote it to its quantum representation –by replacing the classical Ising variables with Pauli σ^z operators. However, in practice, the programming toolkit provided by D-Wave (OCEAN) is already capable of performing the transformation from QUBO all the way to the quantum Ising model. Thus, we simply rely on OCEAN for encoding the H_{QUBO} into the D-Wave machine.

2.3.2 Sampling pathways with Quantum Annealing

We are now in a position to undertake the task of sampling the TPE by performing the quantum annealing, as implemented in the D-Wave machine. Following the standard procedure in Section 1.3.1, the qubits are initialized in the ground state of a Hamiltonian H_0 that is easy to prepare theoretically and experimentally. The structure of H_0 was not relevant as long as it did not commute with H_{QUBO}, and therefore allowed for quantum fluctuations to penetrate the barriers of the cost function associated with H_{QUBO}. Subsequently, D-Wave gradually evolves the coupling of the qubits from H_0 toward H_{QUBO}, which replaces H_{target} in Equation (1.3.5).

In an ideal closed system and for sufficiently slow annealing, adiabatic theorem Equation (1.3.6) ensures that the system remains in its instantaneous ground state throughout the transition. Naturally, by the end of transition at $t = t_{\mathrm{fin.}}$, the final state $|\Phi_{\mathrm{fin.}}\rangle$ of the system is expected to be the ground state of the QUBO Hamiltonian. Then, performing the quantum measurement in the computational basis of individual qubits (the basis of $\{\sigma_z\}$ operators) reads out the configuration dictated by $|\Phi_{\mathrm{fin.}}\rangle$. However in realistic conditions where annealers like D-Wave operate, the probability of landing onto the ground state remains < 1 even in the limit of very long switching times. Several factors contribute to this. For one, the couplings between the system and its environment often induce dissipative and non-adiabatic corrections in the annealing process. Moreover, hardware limitations such as sub-optimal topological connectivity of the qubits produce unintended biases in the result [216].

Such errors have prompted efforts to investigate the potential of these devices as samplers for classical Boltzmann distributions [217, 218]. Especially, since there exist theoretical and experimental arguments that support this possibility [219–221]. However, if the coupling to the environment is particularly strong, experiments also indicate that the system's relaxation at the end of annealing would not be according to the user-defined parameters, $B(t_{\text{fin.}})H_{\text{QUBO}}$ [218]. Rather, the state of the system may freeze at some time t_{freeze}, between the completion of annealing schedule $t_{\text{fin.}}$ and the time of the critical point t_c where the annealer experiences the minimum energy gap (Δ_{01} in Equation (1.3.7)). We recall that $B(t)$ is the instantaneous coupling of the annealer to the H_{QUBO} (H_{target} in Equation (1.3.5)) s.t. $B(t_{\text{fin.}}) \gg A(t_{\text{fin.}})$. Luckily, in a quasistatic regime, the final states of this frozen annealing coincide with a modified Boltzmann distribution, however, according to $B(t_{\text{freeze}})H_{\text{QUBO}}$ where $B(t_{\text{freeze}}) \gg A(t_{\text{freeze}})$ [218]. Therefore, it is logical to assume that there exists a regime of switching times in which the distribution of the paths generated by multiple hybrid energy minimization has a finite overlap with $e^{-H_{\text{QUBO}}}$. Here, without a lack of generality, we have additionally assumed the physical temperature of the machine and $B(t_{\text{fin.}})$ cancel out each other so that $(k_{\text{B}}T)_{dwave}B(t_{\text{fin.}}) \approx 1$.

Following this argument, to account for the environment-induced fluctuations, we may utilize the machine itself to evaluate the condition probability of generating any path at any given switching time, $P(\mathbf{I}|t_{\text{fin.}})$. With the help of this probability, we then devise a Metropolis criterion in an MC scheme that corrects for biases and errors when exploring the space of transition pathways. Moreover, since the interplay between quantum fluctuations and measurement is fundamentally ever-present in any QA process, the anticipation is for the sampling to "forget" previous iteration every time we initiate a new annealing cycle. Therefore, as long as the space of accessible states is large enough, the resulting Markov chain contains uncorrelated pathways that also obey the correct detail balance condition between one another.

The conditional probability $P(\mathbf{I}|t_{\text{fin.}})$ generally depends on the details of the quantum annealing machine and is challenging to compute using theoretical arguments. This barrier could be overcome by performing a moderate number of annealing sweeps with D-Wave, for each value of $t_{\text{fin.}}$. The spectrum of the target Hamiltonian H_{T} is expected to be non-degenerate since the weights in the graph $w_{i,j}$ are in general all different. In addition, we previously showed that for large values of the parameter α in Equation (2.3.1), all the low-lying states of H_{QUBO} correspond to paths with correct topology. Therefore, each low-lying eigenvalue E of the QUBO Hamiltonian is related to the action of a single path, $E = W(\mathbf{I})$. Then, by performing a frequency histogram of the energies E using the paths that we obtain with multiple annealings at fixed $t_{\text{fin.}}$, $P(\mathbf{I}|t_{\text{fin.}})$ can be directly inferred from $P(E|t_{\text{fin.}})$. In the specific, by calculating the average $\overline{E}$ and the standard deviation ΔE of the energy, we

can estimate $P(E|t_{\text{fin.}})$ by the lowest-order cumulant expansion as

$$P(E|t_{\text{fin.}}) \simeq \frac{1}{\sqrt{2\pi}\Delta} e^{-\frac{(E(\mathbf{I})-\overline{E})^2}{2\Delta^2}} \qquad (\simeq P(\mathbf{I}|t_{\text{fin.}})) \tag{2.3.4}$$

The reason for the lowest-order approximation is the limited time generally available with D-Wave hardware, and therefore it be improved systematically by including higher orders in the cumulant expansion.

Once the conditional probability $P(\mathbf{I}|t_{\text{fin.}})$ is established, the road for sampling is open, so to speak. In a classical computer-based MC, when dealing with systems in equilibrium, one would generally invoke a detail balance condition $T(\mathbf{I}_2|\mathbf{I}_1)e^{-W(\mathbf{I}_1)} = T(\mathbf{I}_1|\mathbf{I}_2)e^{-W(\mathbf{I}_2)}$, where $T(\mathbf{I}_2|\mathbf{I}_1)$ is the transition probability from the path $\mathbf{I}_1$ to the path $\mathbf{I}_2$ in the underlying stochastic process. We choose to generalize this dynamic to include the variation of switching times $t_{\text{fin.}}$ in the Markov chain. We expect that such a modification is necessary due to the limited Quantum Processing Unit (QPU) time of D-Wave. In particular, by performing the majority of annealings at $t_{\text{fin.}} \sim t_0$ where t_0 is a tunable parameter specific to each particular problem, we attempt to find a reasonable compromise between accuracy (slow switching) and efficiency (low consumption of QPU time). The new detailed balance condition reads $\rho(\mathbf{I}_1, t_{\text{fin.}})\, T(t'_{\text{fin.}}, \mathbf{I}_2|t_{\text{fin.}}, \mathbf{I}_1) = \rho(\mathbf{I}_2, t'_{\text{fin.}})\, T(t_{\text{fin.}}, \mathbf{I}_1|(t'_{\text{fin.}}, \mathbf{I}_2)$, where $\rho(\mathbf{I}_1, t_{\text{fin.}})$ is the new equilibrium distribution. Our MC dynamics must be defined in such a way as to ensure that the equilibrium distribution is

$$\rho(t_{\text{fin.}}, p) = p_t(t_{\text{fin.}}) \times e^{-W(\mathbf{I})}, \tag{2.3.5}$$

where $p_t(t_{\text{fin.}})$ is an equilibrium probability centered around t_0, chosen to guarantee the accuracy and efficiency of QPU usage. Following this new condition, to obtain the Metropolis acceptance/rejection, we write the transition probability as a product of a trial move probability $\tau(\mathbf{I}_2, t'_{\text{fin.}}|\mathbf{I}_1, t_{\text{fin.}})$ and a corresponding acceptance probability $A(\mathbf{I}_2, t'_{\text{fin.}}|\mathbf{I}_1, t_{\text{fin.}})$. We posit the following form for the transition probability

$$\tau(\mathbf{I}_2, t'_{\text{fin.}}|\mathbf{I}_1, t_{\text{fin.}}) = \kappa(t'_{\text{fin.}}|t_{\text{fin.}})\ P(\mathbf{I}_2, t'_{\text{fin.}}), \tag{2.3.6}$$

where $\kappa(t'_{\text{fin.}}|t_{\text{fin.}})$ is the probability for the switching time to go from $t_{\text{fin.}}$ to $t'_{\text{fin.}}$ in a Monte Carlo step, while we impose from Equation (2.3.4), $P(p|t_{\text{fin.}}) = P(E|t_{\text{fin.}})$. Combining all terms together, we obtain the acceptance rule:

$$A(\mathbf{I}_2, t'_{\text{fin.}}|\mathbf{I}_1, t_{\text{fin.}}) = \min\left[1, \frac{P(t_{\text{fin.}}|t'_{\text{fin.}})}{P(t'_{\text{fin.}}|t_{\text{fin.}})} \frac{P(\mathbf{I}_1|t_{\text{fin.}})}{P(\mathbf{I}_2|t'_{\text{fin.}})} \frac{p_t(t'_{\text{fin.}})e^{-W(\mathbf{I}_2)}}{p_t(t_{\text{fin.}})e^{-W(\mathbf{I}_1)}}\right] \tag{2.3.7}$$

As the last piece of our MC scheme, in our simulations, we chose to update $t_{\text{fin.}}$ according to Brownian dynamics with a harmonic drift term:

$$t^{i+1}_{\text{fin.}} = t^{i}_{\text{fin.}} - \delta t k(t_{\text{fin.}} - t_0) + \sqrt{2\delta t}\xi^i, \tag{2.3.8}$$

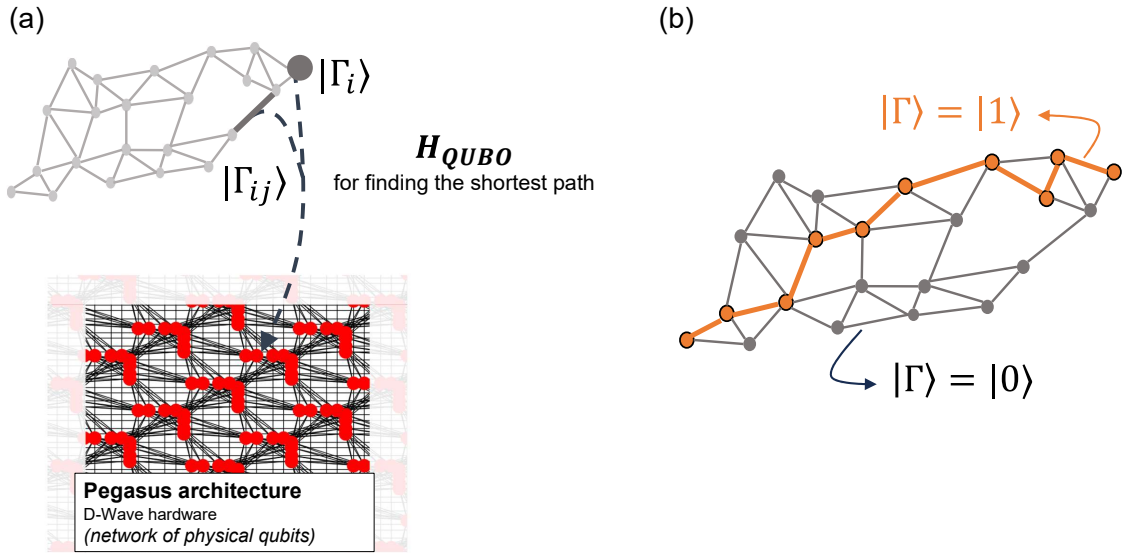


Figure 2.6: (a) Encoding of the network of transition into D-Wave annealing machine using the QUBO formulation. (b) In QUBO, every vertex and edge of the network is represented as a binary variable. If a vertex/edge is present in the path, its associated binary variable takes the value 1, otherwise 0.

where ξ_i is a Gaussian distributed random variable of null mean and unitary variance and δt is an incremental switching time change.

The overall scheme of utilizing QA in the pipeline of gTPS framework is summarized and illustrated in Figure 2.6 and Figure 2.7.

2.4 Discussion

The intricate interplay between the components of our novel framework facilitates the sampling of the transition pathways in a molecular conformational reaction. The unique strength of the gTPS comes from the fusion of iMapD's data-driven approach for rapid uncharted exploration of FEL and the quantum annealing's ability –as implemented in the DWave's machine– to generate minimally correlated pathways. This ultimately yields an ensemble of discrete pathways which by design traverse along the low-energy regions of the FEL, thus maintaining a high acceptance probability in the MC process. As such, we believe gTPS is bound to overcome both of the long outstanding challenges of conventional TPS algorithms mentioned in Section 1.3.1.

Moreover, the incorporation of quantum annealing to generate the reactive paths marks an important difference concerning previous frameworks of TPS. Assuming iMapD is capable of obtaining configurations along every possible channel for a reaction, then the initial

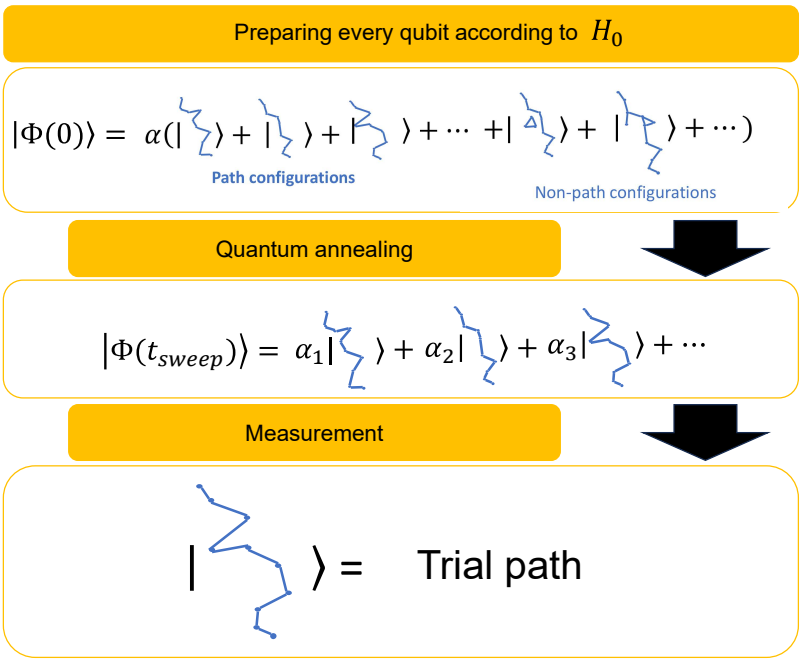


Figure 2.7: Illustration of how a (quasi)- adiabatic switching procedure on a quantum annealing machine yields uncorrelated trial paths on the graph.

quantum state of the annealer holds in superposition the entire ensemble of pathways corresponding to these channels –present in the network of transition. Thus, the gTPS can identify globally the more optimal pathways in the transition network (lower statistical weight) in contrast to classical methods such as kMC which are only able to search locally for these paths using their stochastic approach. This can be considered somewhat as a similar theoretical shift as when the original TPS was introduced as an MC approach which samples directly the trajectory space instead of configuration space [53]. Additionally, gTPS achieves all of these with no simplification or reduction of the system under study in its applications in contrast to other methods involving QCs. To validate and illustrate the capabilities of this approach, we discuss in the next chapter its first application to a benchmark molecule, Alanine dipeptide.

In another significant advancement, in Chapter 4, we introduce a modification designed to enhance the stability of the shooting move in the iMapD algorithm. This additional piece enabled gTPS to seamlessly explore the near-native-state transitions of a much larger molecule with an order of magnitude higher number of atoms than alanine dipeptide. Notably, the same system was previously studied using the special supercomputer, Anton, and by running more than 1 ms of plain MD simulation. Our refined version of iMapD manages to capture conformations in all observed conformational states present in the Anton trajectory, with the exception of the least populated state. Furthermore, using a modest amount of QC time, we obtain transition pathways that closely mimic the same conforma-

tional transitions as in Anton data. Lastly, we have dedicated one chapter to the ongoing applications of gTPS to the unfolding of proteins.

The diagram of iMapD and gTPS algorithms are included in Appendix A.5.

2.4.1 The gTPS's pathways and experiments

Prior to finishing this chapter, we briefly allude to the structural and dynamical observables that one could obtain using the gTPS framework. The kinetic rate of reaction is often one of the first quantities of interest reported in computational studies involving TPS or its variants. As mentioned in Section 2.2, the DRP formalism, while capable of evaluating the thermodynamic cost of transitioning between finite size regions (statistical weights in the transition network), is solely able to estimate a lower bound for the transition path time, i.e the average time it takes to complete a reaction along a productive path. This is because the connections in the network of finite-size regions are established under the assumption that the corresponding transitions (between two connected regions) occur only by overcoming a single free energy barrier and as such the thermodynamic cost of the transition can be evaluated as the average of $S_{\mathrm{HJ}}^{\mathrm{cg}}$ between the two regions. Naturally, this approach does not contain the necessary kinetic information that accounts for multiple-barrier crossing or relaxation between multiple regions. Therefore, gTPS cannot generate real dynamical trajectories where experimental observables such as reaction rates can be evaluated. In contrast, in conventional discrete TPS methods, the information on barrier crossings and relaxation is inherently provided to the kMC process in the form of the transition matrix, to generate the pathways. Alternatively, biophysical experiments can precisely determine the time it takes the system to exit one state and arrive at the next which then can be compared with the prediction of gTPS. Unfortunately, the transition path time is less informative than the reaction rate. Indeed, it depends only logarithmically on the height of the energy barrier. As a result, while transition rates of protein structural rearrangements can vary over many orders of magnitude, the corresponding transition path times are typically within 1 to a few microseconds.

On the other hand, the kinetic rate is not the only experimentally verifiable observable of interest. Capturing certain conformational changes (e.g the order in which the protein's structural motifs form) or identifying intermediate metastable states along a reaction are two other examples that are regularly looked for in computational studies [222–224]. In the folding process of large multidomain proteins (including membrane proteins), the presence of long-lived and partially folded intermediate states is in many cases a rate-limiting factor in the reaction and can make the molecule prone to aggregation and misfolding [224]. Therefore, it is necessary to structurally characterize these metastable states on the folding pathway and understand their role in the kinetics. In particular, the identification of

these states has become a potential new avenue for drug discovery/development as can be observed from companies such as Sibylla Biotech [225].

Meanwhile, the underlying mechanism of the folding process of different proteins cannot be explained in a unifying fashion e.g. by only using the nucleation growth model. This mechanism may very well hold for small single-domain proteins whose folding can be approximate as that of a two-state system (unfolded-folded) [226]. However, in the case of knotted proteins, or the ones with disulfide bridges, or large membrane proteins, the structural evolution of the molecule during folding involves a more complex description [223].

Since the gTPS pathways predominantly favor the low free energy regions, we believe their configurations provide a particularly reliable way to identify metastable states along a reaction. To this aim, one possible approach would be retrieving those highly visited configurations in the obtained TPE. Furthermore, the *transition state* (Section 1.2) of a reaction can be readily identified by finding the configuration in the Dijkstra path from which the probabilities of arriving at product and reactant states are equal (according to Section 2.2.3) [210].

Concerning the conformational evolution along a reaction, we again expect that relying on the low energy pathways of gTPS reveals more easily the essential structural changes of a system such as the order of the α-helix or β-sheet formation in the folding of a protein. We argue that since these pathways are devoid of configurations in high free energy regions, therefore they are less prone to introduce artifacts in the analysis of the structure. We also suspect that as such, these configurations are more valuable for identifying CVs associated with a reaction.

CHAPTER 3

Case study: benchmarking with Alanine dipeptide

In the preceding chapters, we presented the theory behind the framework of gTPS which aims to address the outstanding issues of conventional TPS algorithms. In this chapter, to verify this claim and further demonstrate the capabilities of our framework, we apply gTPS to a benchmark system called Alanine dipeptide (ALA). ALA is a relatively small molecule with 22 atoms. Extensive experimental and computational studies have shown that this molecule shares several chemical features in common with the polypeptide chain of proteins. For example, the structure in metastable states on the ALA's free energy surface –characterized by its torsion angles ϕ and ψ– closely mirror the "allowed/favorable" conformations of amino acids in proteins (excluding Proline and Glycine residues), as the latter form α-helices and β-sheets. This resemblance is why the α_R and β (C_5) states of ALA's landscape have been given these names [206].

Moreover, the existence of a methyl side chain in ALA is representative of all the commonly occurring amino acids except Glycine [227, 228]. The protein-like features and simple structure of ALA make it an excellent prototype for studying the dynamics of backbone atoms. It continues to be a popular choice for benchmarking newly developed algorithms in computational studies of biomolecular systems and is often among the first systems used for tuning/learning force field parameters.

ALA's remarkable characteristics make it an ideal starting point for us to test the capabilities of gTPS framework. Furthermore, its modest size allows it to perform quantum computing calculations on existing D-Wave machines. Consequently, we utilize gTPS to

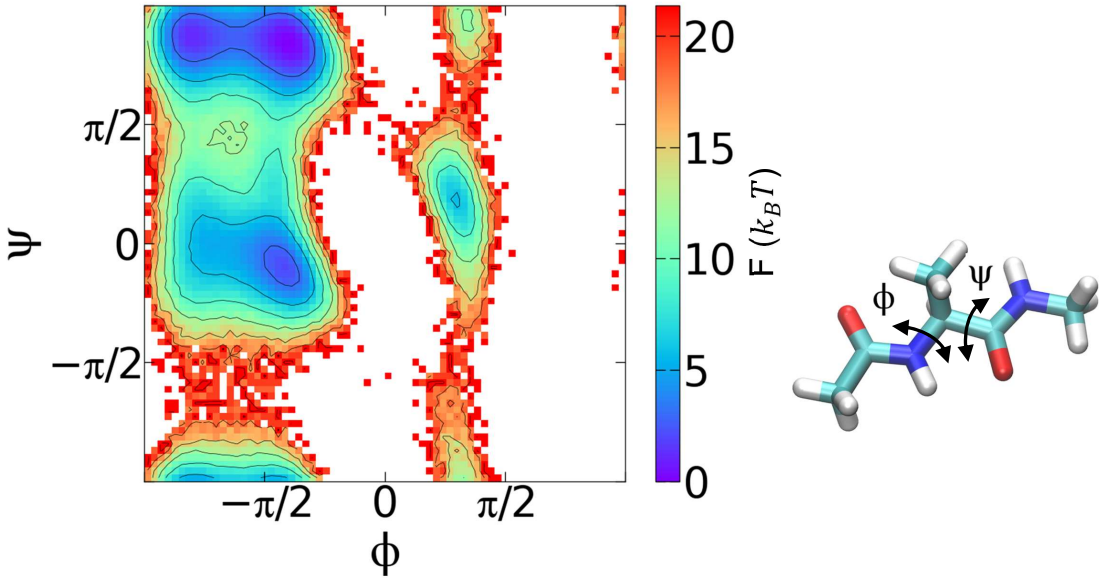


Figure 3.1: Free energy Landscape of alanine dipeptide, projected onto its two main dihedral angles. This figure was generated by simulating alanine dipeptide for 1 μs at T = 300 K, and in explicit TIP3P water. Contour lines are drawn every 3 kJ.

simulate the $C_5 \rightarrow \alpha_R$ transition of ALA.

3.1 Applying gTPS to alanine dipeptide

3.1.1 Exploring ALA's intrinsic manifold

We start by sampling from the two metastable states using OpenMM [229]. One simulation starts from the C_5 region at the top left corner of the Ramachandran plot in Figure 3.1 and another from α_R in the vicinity of $\phi = -75$, and $\psi = -20$. The Figure 3.1 was calculated from a frequency histogram of 1 μs of equilibrium MD at $T = 300K$, generated using OpenMM, in the AMBER99SB force field with explicit TIP3P water. The contour lines of this figure will be used as a reference in all the projections of ALA configurations on the dihedral angles surface (Ramachandran plot). After the initial sampling, we evaluate the DMAP for each set of configurations separately to obtain a low-dimensional representation. We then identify the boundary in this representation to initialize new simulations beyond the region that has already been sampled. We measure the pairwise distance between configurations by calculating RMSD between the backbone atoms after having removed global translations and rotations. This calculation was done using the MDAnalysis package in Python [230, 231]. At every iteration of iMapD, we assigned the scaling parameter $\epsilon = \mu(d_{ij}) - \sigma(d_{ij})$ (see Appendix A.2), where $\mu(d_{ij})$ is the average of RMSD and $\sigma(d_{ij})$ is the standard deviation. The same measure $\mu(d_{ij}) - \sigma(d_{ij})$ was also used in the shooting

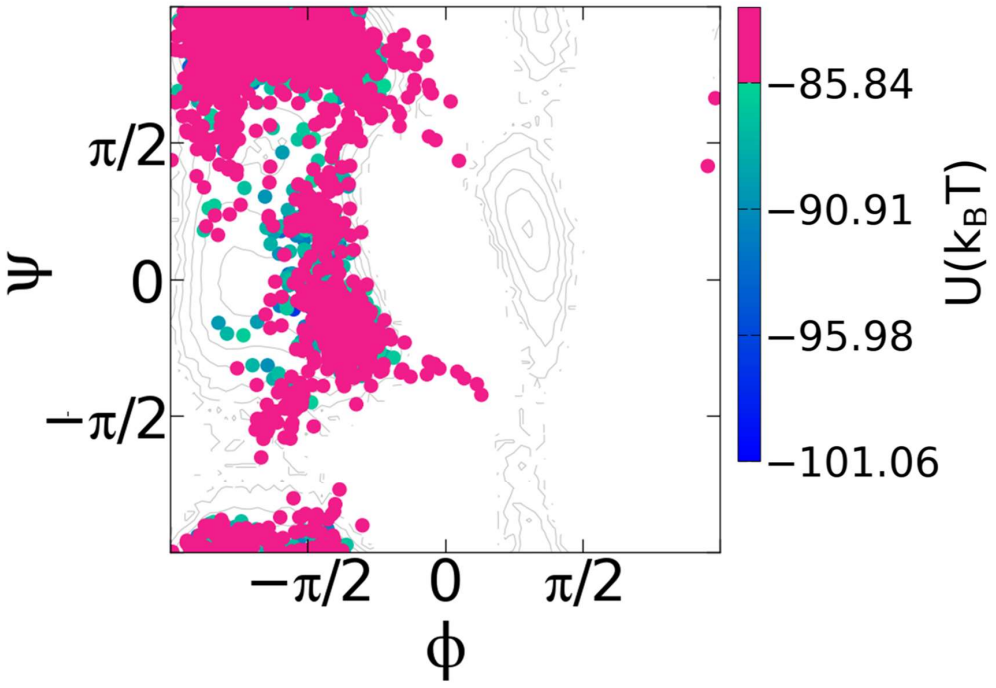


Figure 3.2: Final result of the exploration with iMapD. The neon-pink colored points signify the configurations whose potential energy is higher than the median $85.84 k_{\mathrm{B}}T$

moves to identify the neighboring configurations for every boundary point.

The exploration proceeds according to the steps of iMapD in Section 2.1: Initiating unbiased sampling from each new configuration and then merging all the new data to the previous one. By iterating over DMAP evaluation at every step, finding new configurations, and sampling we populate the transition region between C_5 and α_R states. We eventually terminate the iterations when the configurations explored starting from the two initial metastable state overlap, i.e., when at least two configurations have RMSD closer than $0.3\,\mathring{A}$. The final set of iMapD's configurations form the set, $\mathcal{C}_{\mathrm{fin.}}$, and are depicted in Ramachandran plot in Figure 3.2. In this figure, we have singled out the configurations with potential energy higher than the median.

Numerical details: In all simulations of iMapD, the molecule was placed in a square box with 2.85-nm base vector, solvated with the TIP3P water model using again AMBER99SB forcefield. We energy-minimized the initial configuration using the L-BFGS algorithm implemented in OpenMM with tolerance on the square mean root of all force components at 500 kJ $(\mathrm{nm\ mole})^{-1}$. Simulations were performed at a temperature $T = 300K$, using a Langevin integrator with friction coefficient $\gamma = 91\mathrm{ps}^{-1}$ and timesteps of $\Delta t = 2$ fs. The initial sampling bursts in the two metastable states were 200 ps long starting in C_5 state and 20 ps for α_R. All the subsequent runs were 1 ps long.

3.1.2 Constructing network of transitions

Next, in the gTPS framework, we partition the configuration space between the α_R and $C5$ states to form the set of finite regions $\mathcal{R}$. Then, leveraging the CG theory describing transitions between these regions, we build a network of connectivity over $\mathcal{R}$. Our concern here is to maintain sparse partitioning to ensure compatibility with the D-Wave quantum annealer. In particular, minimizing the required qubits for encoding the network reduces the likelihood of the thermalization and decoherence issues we mentioned in Section 2.3.

The process begins by excluding configurations from $\mathcal{C}_{\text{fin.}}$ with potential energy higher than the median $U(\mathbf{Q}) > 85.84\, k_{\text{B}}T$ (Figure 3.2). Second, we calculated the DMAP of all remaining samples and projected them on the first two DCs, obtaining the set $\mathcal{Z} = \{z_k\}$ (Figure 3.3(b)). We then proceeded by identifying the two configurations in $\mathcal{Z}$ with minimum RMSD from the initial configurations lying in two basins $C5$ and α, denoted as z_i (corresponding to $\mathbf{Q}_i$ in the original space) and z_f ($\mathbf{Q}_f$). Starting from $\mathbf{z}_i$, we kept the nearest point z_{k1} satisfying $D_{\text{diff}}(\mathbf{z}_{k1}, \mathbf{z}_i) > D_{\text{diff}}^{\text{thresh}}$, and removed from $\mathcal{Z}$ all the configurations with lower D_{diff}. Here D_{diff} (diffusion distance) is the Euclidean distance between points in the space spanned by the two DCs (more details in Appendix A.2). We used $D_{\text{diff}}^{\text{thresh}} = 7.5 \times 10^{-4}$ inspecting the histogram of nearest neighbor pairwise diffusion distances among the points in $\mathcal{Z}$, Figure 3.4. This ensures that any point considered kinetically similar to $\mathbf{z}_i$ and $\mathbf{z}_{k1}$ are removed (following the description of DMAP in Appendix A.2). We continued by applying the same procedure between remaining configurations in $\mathcal{Z}$ and $\mathbf{z}_{k1}$, and therefore identifying iteratively the set $\mathcal{Z}_s = \{\mathbf{z}_i, \mathbf{z}_{k1}\mathbf{z}_{k2}, \dots, \}$. We also make sure that we have included z_f (representing $\mathbf{Q}_f$) in the $\mathcal{Z}_s$. By the end of this procedure and lifting back to the original configuration space, we obtained the set $\mathcal{S} = \{\mathbf{Q}_i, \mathbf{Q}_{k1}, \mathbf{Q}_{k2}, \dots, \mathbf{Q}_f\}$ of $\nu = 83$ configurations. This reduced dataset is illustrated both in DMAP space and in the Ramachandran plot in Figure 3.5.

In the next stage, we build a graph having as nodes the sparse set of configurations $\mathcal{S}$. In this graph, we argued that only those configurations that are structurally and kinetically close should be connected. We used two criteria to ensure this condition. Two nodes should be connected if their diffusion distance is smaller than 0.01 and their RMSD closer than 0.8 $\mathring{A}$. Both thresholds were chosen heuristically from the histogram of pairwise diffusion distances and RMSD calculated on the set of configurations generated by iMapD (depicted in Figure 3.6). We should note that these values guarantee to have a single connected network on $\mathcal{S}$ where every node has at least 2 incident edges. Additionally, we point out that in both histograms the existence of two meta-stable basins is evident from the two peaks in the distribution of distances. Next, we calculate the weights $w_{i,j}$ of the edges in this graph, Equation (2.2.39). To evaluate the V_{cg} and consequently the weights for our first illustrative example, we resorted to a simple phenomenological approach instead of

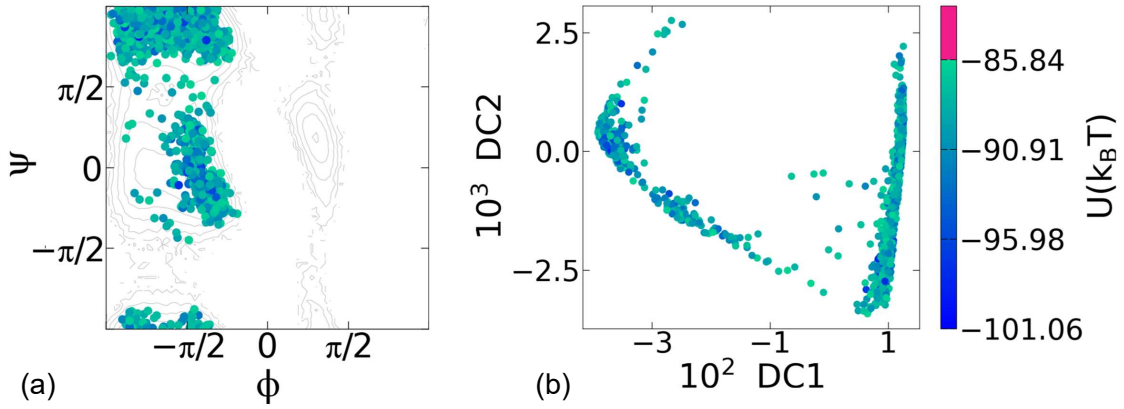


Figure 3.3: (a) Ramachandran plot of iMapD data after removing the configuration with potential energy higher than the median. (b) DMAP embedding of points whose potential is below the median,shown as a function of the first two DCs.

exactly following the procedure provided in Chapter 2. Namely, we first smeared out the short-distance structure of V_{eff} through a window averaging over the groups of points that lay close in the space span by the two DCs in Figure 3.3.(b). Let us denote the value V_{eff} in each window k as $(V_{\text{cg}})_k = \langle V_{\text{eff}} \rangle_k$. Depending on which window the configurations of $\mathcal{S}$ would be located, we assigned to each the value of $(V_{\text{cg}})_k$ respectively. The results can be seen in Figure 3.7. Next, we assigned s_0 in Equation (2.2.39) as the absolute value of $min(V_{\text{cg}})$ and calculated the weights of every edge in the network. To ensure that a single path is not overly represented due to the exponential nature of its probability, we rescaled all the weights by dividing their values by the largest link w_{max}. The network of transition between the configuration in the reduced data set $\mathcal{S}$ can be seen in Figure 3.8.

3.1.3 Sampling transitions paths from C_5 to α_R

Now that we have a fully realized network between the $\mathbf{Q}_i$ and $\mathbf{Q}_f$ we can proceed to sample the paths that represent a transition between C_5 to α_R. To implement our hybrid classical/quantum Monte Carlo scheme, we encode the QUBO Hamiltonian H_{QUBO} that was defined by our graph onto the DWave quantum annealer. As explained in Section 2.3, we utilized the OCEAN programming suite provided by DWave themselves for this step. Our encoding required 578 qubits, given by the sum of the number of nodes and edges of our network. To generate trial paths, we rely on the hybrid solver available with DWave, which combines quantum annealing with classical simulating annealing. We should note that in this case, the $t_{\text{fin.}}$ is now identified with the total hybrid solver's computing time and not the QPU time.

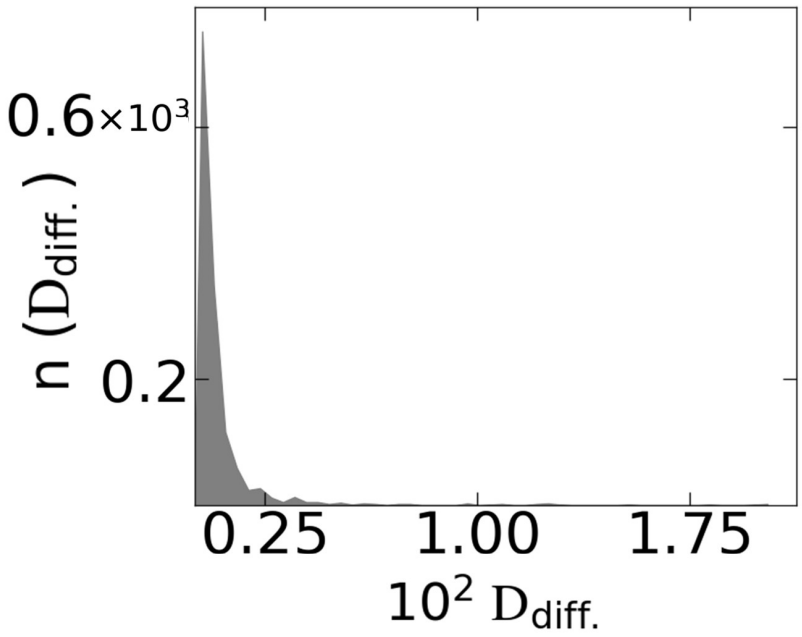


Figure 3.4: D_{diff} histogram for nearest neighbors. $D_{\text{diff}}^{\text{thresh}} = 7.5 \times 10^{-4}$ contains more than 98% of nearest neighbor pairwise distances. In downsampling the iMapD dataset to build the finite space regions $\mathcal{R}$ we adopt this value as the minimum diffusion distance allowed between the configurations.

Num. of attempts	Correct topology	Wrong topology	Success rate
117	69	48	0.59

Table 3.1: Summary of the transition path generation on D-Wave to calculate the histogram reported in Figure 3.9

We proceed by estimating the conditional probability $P(\mathbf{I}|t_{\text{fin.}})$ in Equation (2.3.7) using a direct calculation on D-Wave using Equation (2.3.4). The result of this calculation is summarized in Table 3.1 and can be observed in Figure 3.9. In the latter, we report the average value of the energy $\overline{E}$ and its standard deviation $\Delta(E)$ for every $t_{\text{fin.}}$. Following the MC procedure in Section 2.3, in the next step we initiated three independent Markov chains from arbitrary paths generated by separate runs of hybrid solver at 30 s, $t_{\text{fin.}} = 180$ s, and 240 s, corresponding to $\tilde{1}$ s, $\tilde{6}$ s, and $\tilde{8}$ s of QPU time, respectively. We evolved $t_{\text{fin.}}$ according to Eq. (2.3.8) with $k = 2 \times 10^{-4}$ s^{-1} and $t_0 = 150$s and then accepted or rejected the new paths according to Eq. (2.3.7). We also recall that $\alpha = \sum_{ij} w_{ij}$ in the H_{QUBO}. With this choice, on average, over 60% of the annealing sweeps led to configurations of binary variables Γ with a correct path topology (summarized in Table 3.2). In Figure 3.10 we show the change in path action W_p and the hybrid minimization time $t_{\text{fin.}}$, along our three Markov chains. As these results show, the MC algorithm occasionally accepts trial moves with a higher action. They also show that longer annealing times do not always

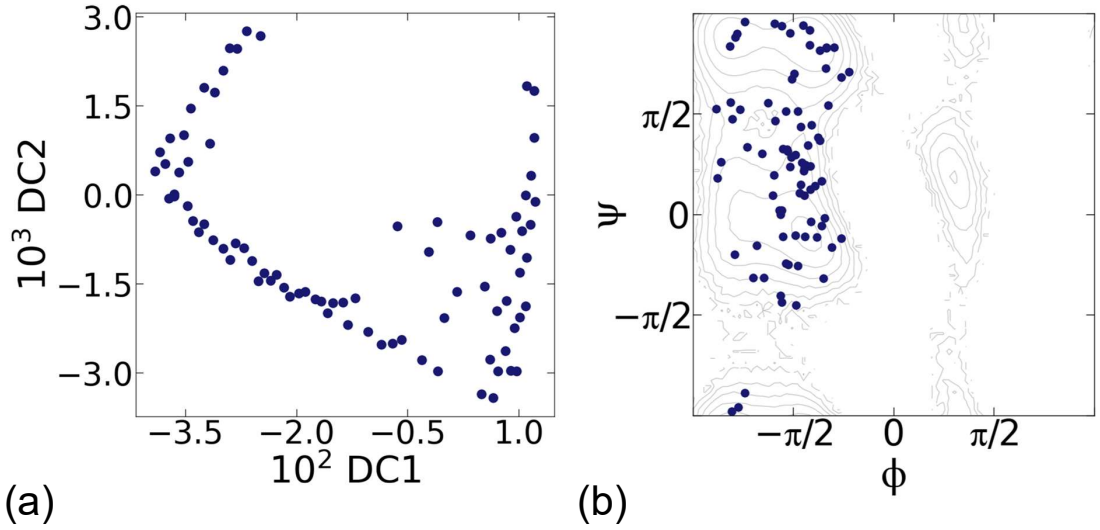


Figure 3.5: Reduced iMapD data projected on (a) the first two DCs and (c) Ramachandran plot.

	Monte Carlo steps	Accepted paths	Wrong topology	Rejected paths
Markov chain 1	9	7	0	2
Markov chain 2	13	8	2	3
Markov chain 3	20	10	4	6

Table 3.2: Summary of the Markov chains sampling process on D-Wave.

yield paths with lower actions. This is expected, since the $P(E|t_{\text{fin.}})$ distributions have significant overlap, which can be inferred from Figure 3.9. By projecting all the paths onto the Ramachandran plot, we can observe in Figure 3.12 that the transition paths generated by our scheme are consistent with the FEL of ALA. In panel (a) we illustrated the first and last accepted transition paths of one of the generated Markov chains. Both paths correctly connect the two meta-stable states, navigate the low-free energy regions of the surface, and cross the barrier at its lowest point. In panel (b) we report how often all the sampled transition paths have passed each node of the network, i.e., the statistical weight of the corresponding configuration in the transition path ensemble. Even though all the paths go through the transition state, due to the presence of fluctuations, a relatively small portion also visits configurations with relatively high free energy. The deterministic least action that we also calculated using the Dijkstra algorithm (Figure 3.11) can only detect the global minimum of the functional W_p. In contrast, our TPS algorithm accounts for fluctuations that lead to the full transition path ensemble.

Auto-correlation between the sampled path: The main strength of our hybrid classical/quantum scheme is that it allows us to efficiently obtain independent transition paths. Following the arguments in Section 1.3 and Section 2.3, the only source of correlation

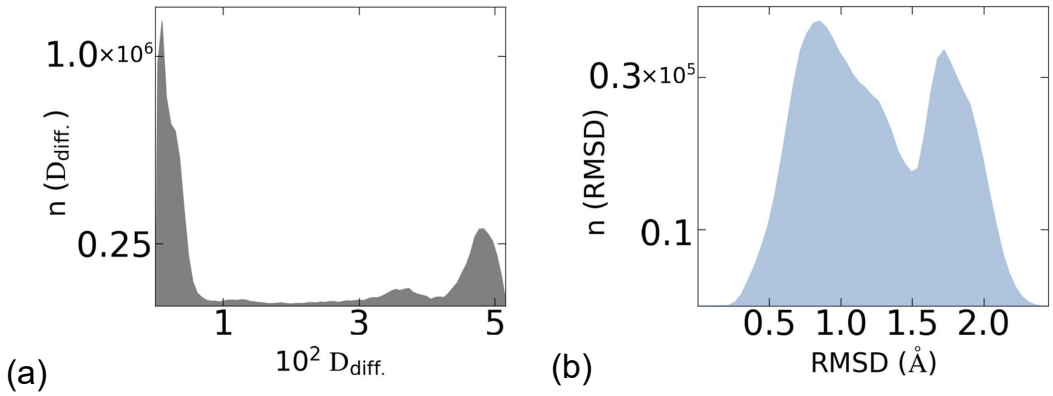


Figure 3.6: (a) Histogram of pairwise D_{diff} (b) and of pairwise RMSD calculated on all configurations sampled by iMapD.

between these paths can be due to the stochastic evolution of the minimization time $t_{\text{fin.}}$ (Eq. (2.3.8)), otherwise gTPS produces uncorrelated paths in the Markov chain. To quantify the degree of correlation in the trajectory of sampled paths, we consider the auto-correlation function $G(N)$:

$$G(N) = \frac{1}{N_{\text{MC}}} \sum_{k=1}^{N_{\text{MC}}} \left[\left\langle \frac{1}{|\mathcal{E}|} \overrightarrow{\Gamma}(k+N).\overrightarrow{\Gamma}(k) \right\rangle - \left\langle \frac{1}{|\mathcal{E}|} \overrightarrow{\Gamma}(k+N)^2 \right\rangle \left\langle \frac{1}{|\mathcal{E}|} \overrightarrow{\Gamma}(k)^2 \right\rangle . \right] \tag{3.1.1}$$

In this equation, the $\overrightarrow{\Gamma}(k)$ represents the vector of all binary variables encoding the edges of the graph (the set of $\{\Gamma_{i,j}\}_{i,j=1,\dots,\nu}$) generated at the k-th Monte Carlo step and we are implicitly assuming periodic boundary conditions, i.e., $\overrightarrow{\Gamma}(N_{\text{MC}}+1) = \overrightarrow{\Gamma}(1)$. Here, N_{MC} is the number of Monte Carlo steps, and $|\mathcal{E}|$ denotes the number of edges present in the graph. Finally, N is the distance in the Monte Carlo chain. The average $\langle f \rangle$ is intended over many Monte Carlo trajectories. In practice, however, the computing time that was available to us on the D-Wave quantum computer was sufficient to generate only 3 Monte Carlo trajectories. Since such limited statistics does not allow us to estimate the averages in Eq. (3.1.1), we chose not to perform the average and directly analyze the behavior of $G(N)$ for each independent trajectory. In Figure 3.13 we plot the behavior of $G(N)$ (evaluated relatively to its initial value $G(0)$) for each independent Markov chain. These results clearly indicate that the correlation of the generated trajectories is suppressed after just a single Monte Carlo step.

Due to the absence of autocorrelation, the cost of generating an ensemble of N independent transition paths scales like $\sim \frac{Ns}{a}$, where s is the cost of a single Monte Carlo step and a is the average acceptance ratio. In our case, s amounts to about 120 seconds of total

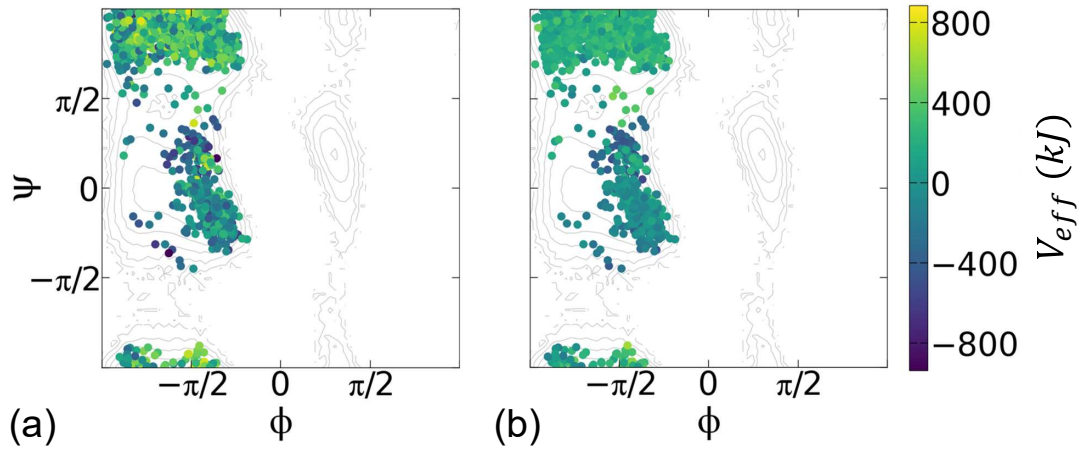


Figure 3.7: The original V_{eff} (a) and the average version which we assign as the V_{cg} (b). To average V_{eff}, we used the embedding of configurations on the first two DCs, and averaged over configurations that were close to each other.

computing (of which about 4 seconds of QPU time) and $a \simeq 60\%$. Therefore, to produce an ensemble of about 100 independent transition paths for this system, our Monte Carlo scheme would require slightly more than 5 hours of total hybrid computing time, including about 10 minutes of QPU time.

3.2 Discussion

Through the application of gTPS to the alanine dipeptide molecule, we have demonstrated its ability to enhance the sampling of the full pathway ensemble during a conformational transition. More notably, our examination of the $C5 \rightarrow \alpha_R$ transition has shown that the gTPS-generated paths correctly follow the low energy regions of FEL while maintaining minimal correlation within the Markov chain. We attribute this crucial feat to gTPS's distinct approach in formulating the problem of sampling TPE. Notably, the TPE is often heterogeneous, displaying multiple transition channels corresponding to alternative molecular mechanisms. gTPS –by utilizing iMapD– rapidly explores all the transition channels that are accessible (concerning their relative free energy) and samples statistically relevant molecular structures in these regions. This approach provides us with a more global and comprehensive view of the IM and the conformational transition.

Subsequently, gTPS builds a discrete representation of TPE by utilizing the effective description of the dynamics directly derived from the iMapD data. This unique formulation enables us to encode the discrete TPE into a quantum computer. Leveraging the inherent features of the device –namely quantum fluctuation and measurement– allows us

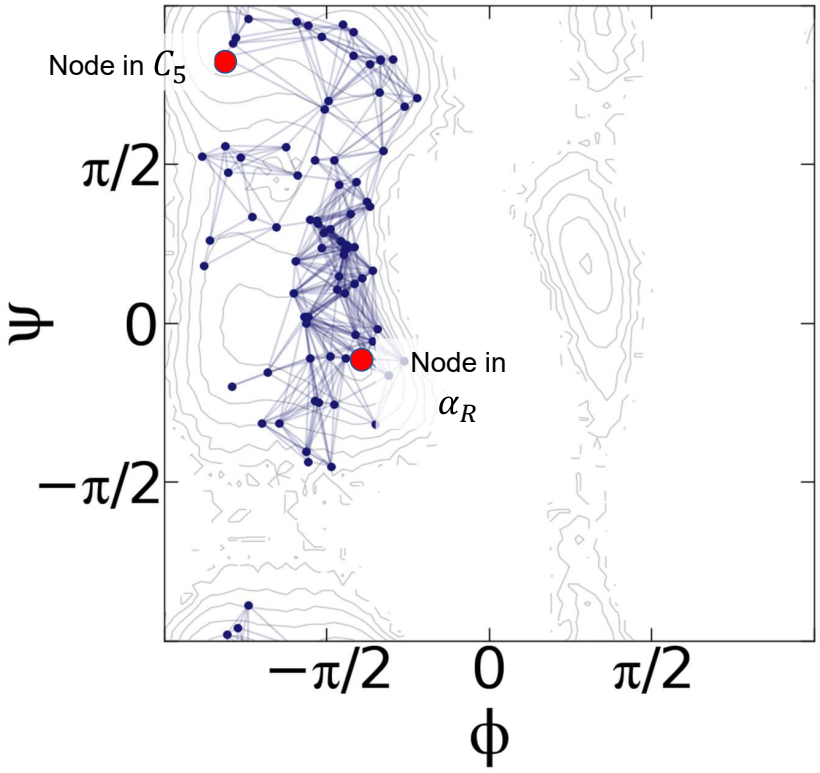


Figure 3.8: Network of transitions plotted on the Ramachandran plot. Nodes correspond to reduced data set $\mathcal{S}$, and edges connect configurations that are kinetically and structurally close. The number of vertices and edges are $|V| = 83$ and $|E| = 495$, respectively.

to generate uncorrelated trial pathways in an MC process. By acknowledging the physical limitations of the device, we extend the metropolis criterion to include the conditional probability $P(E|t_{\text{fin.}})$ calculated from the device itself. This in return facilitates the sampling of transition pathways correctly according to their relative statistical weight on our network.

Finally, we acknowledge that the steps of gTPS were intentionally not strictly followed in the application to the benchmark system of ALA. Specifically, while constructing the network, we deviated from the instruction provided in Section 2.2 for calculating the V_{cg}. This deliberate choice was grounded in the understanding that ALA, being a relatively small molecule, accordingly possesses a geometrically simple effective potential surface. Therefore, a window averaging on the IM is a valid approximation. This was subsequently confirmed as the generated transition paths on the network in Figures 3.11 and 3.12 correctly follow the regions of low free energy.

In the next chapter, we investigate the scalability of gTPS by studying its application to the rare transition of a much larger molecule. There we also perform the gTPS framework in full (calculating V_{cg} according to Equation (2.2.29)) and showcase further the validity of our approach.

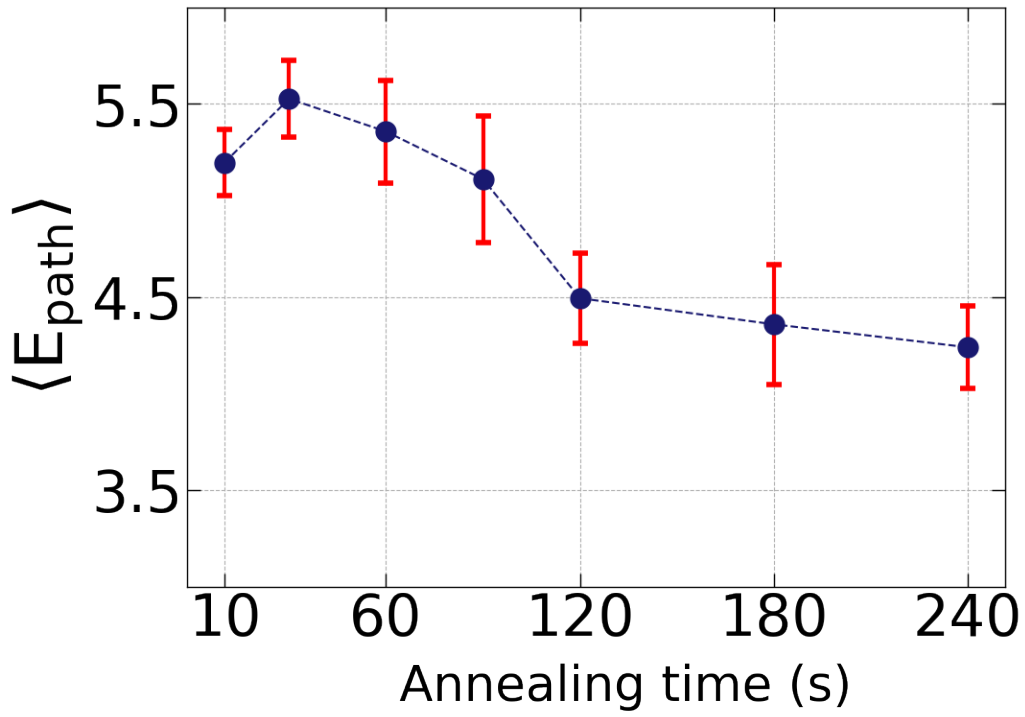


Figure 3.9: Average value and standard error of the mean of the energy E obtained by multiple quantum annealing processes at fixed values of $t_{\text{fin.}}$. These results are used to estimate $P(E|t_{\text{fin.}})$ to the lowest order in the cumulant expansion approximation.

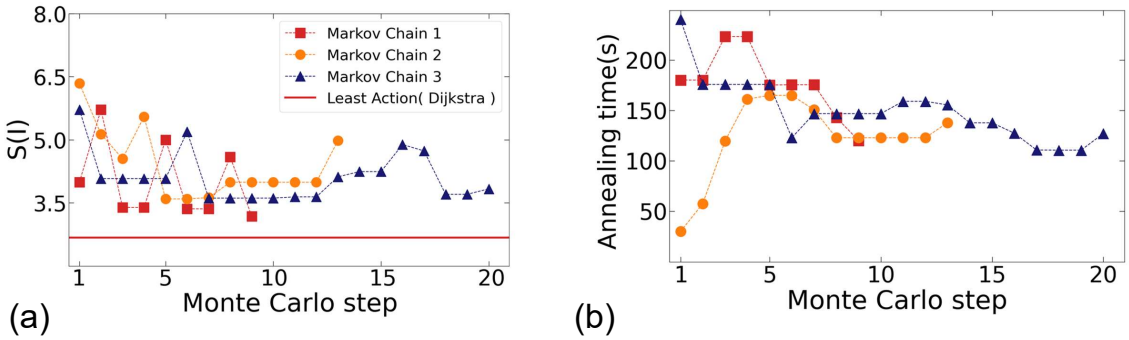


Figure 3.10: (a) Evolution of the path action W_p and (b) annealing time $t_{\text{fin.}}$ along the Monte Carlo paths generated using the hybrid classical/quantum annealing implemented on D-Wave.

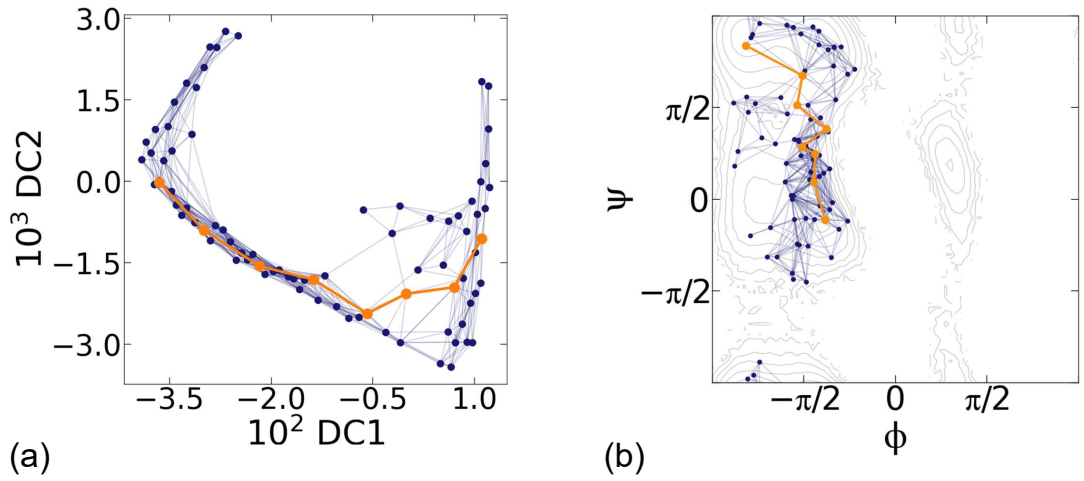


Figure 3.11: The most probable path, calculated via the Dijkstra algorithm, in (a) DMAP embedding and (b) Ramachandran plot.

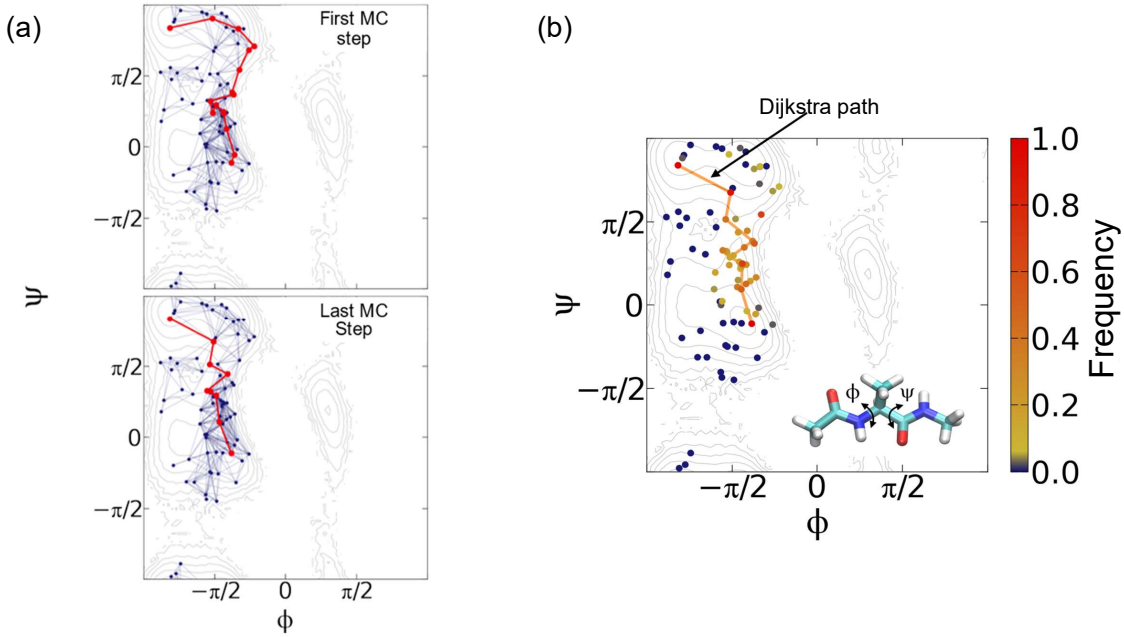


Figure 3.12: (a) Sample of transition pathways for the $C5 \rightarrow \alpha_R$ transition of ALA obtained from D-Wave in an MC process. The red line denotes the first (top) and last (bottom) trajectory in one of the Markov chains. (b) Transition path density, which was evaluated for the ensemble of trajectories calculated from the paths in all the Markov chains of our MC. The solid orange line is the most probable path, Figure 3.11.

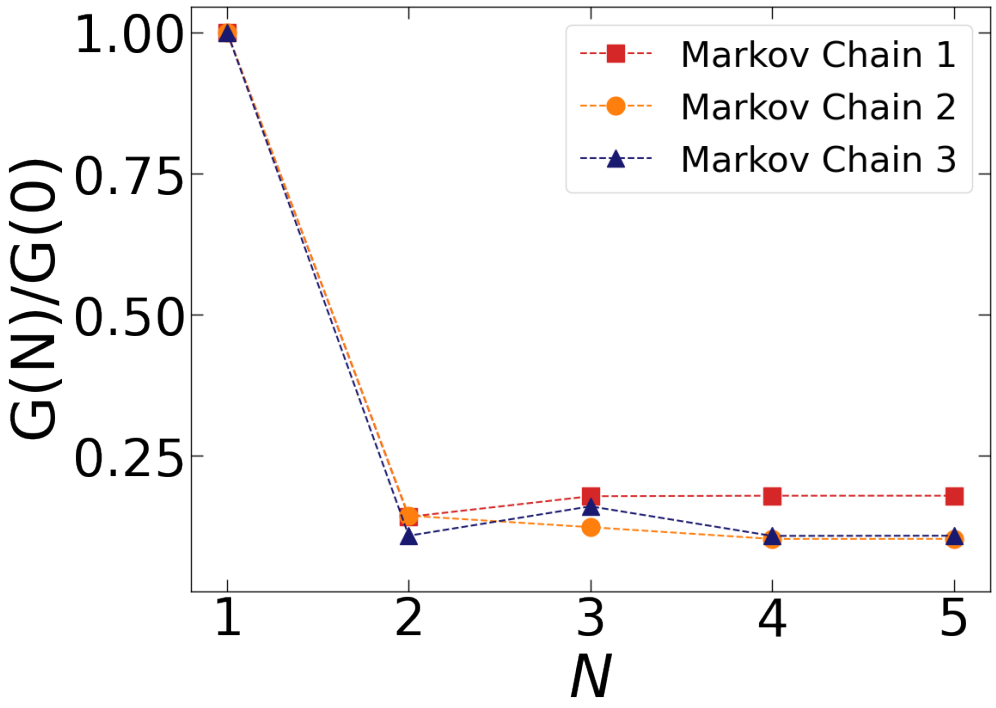


Figure 3.13: The ratio of auto-correlation function $G(N)/G(0)$ plotted as a function of Monte Carlo steps N for three independent Markov chains.

CHAPTER 4

Case study: Bovine Pancreatic Trypsin Inhibitor

In Chapter 3, by resorting to a small GPU desktop and DWave quantum annealer, we successfully demonstrated the ability of the gTPS framework to generate transition paths that explore different regions of configuration space. The critical question is whether our method, implemented on the existing quantum hardware, can accurately simulate transitions too complex to be investigated by plain MD even on large GPU computer clusters. In this chapter, we present the application of gTPS onto Bovine Pancreatic Trypsin Inhibitor (BPTI, PDB code: 5PTI), a substantially larger polypeptide chain than ALA. Our objective is to study the rare transitions of this molecule near its native structure. Previously, the Anton special-purpose supercomputer [25] had fully characterized the same system using more than 1-ms long plain MD simulation. We utilize their result as a benchmark for the gTPS application.

BPTI is a 58-residue globular protein (with 892 atoms), that serves as an inhibitor of proteolytic enzymes such as trypsin, as its name implies. Naturally produced in the bovine's pancreas, its relatively small size has rendered this molecule one of the most studied proteins both computationally and experimentally. BPTI has played a pioneering role in research, being among the earliest proteins to have its crystal structure identified and the first to be simulated using MD [232, 233]. With a research span of 50 years, BPTI offers valuable insights into the study of the folding process, especially for the disulfide-rich proteins due to its 6 Cystine (Cys) residues.

In 2010, D. E. Shaw *et al.* [25], utilized the supercomputer Anton to simulate the

dynamics of BPTI for 1 ms at the temperature of 300K starting from the crystallographic structure. By performing a kinetic clustering on a time auto-correlation function, they analyzed the long-time behavior along their molecular trajectory. Their result revealed 5 structurally distinct states in their data. We present the average configurations in these states in Figure 4.5. Notably, in the two most populated states, the average structure of one corresponds to the crystallographic structure and the other exhibits a left-handed disulfide bridge between the Cystine residues 14 and 38. The existence of the latter had already been confirmed by NMR experiments. Additionally, three states that had not been predicted by any experiments were also identified. Their distinct characteristics include a larger exposed surface area to the water solvent, different rotation rates of aromatic side chains, and the breaking of the hydrogen bonds in the small α-helix near the N-terminus. While iMapD may not allow for such a detailed dynamical analysis, we rely on several structural features to demonstrate that iMapD successfully explores the majority of states observed by Anton, missing only the least populated one.

4.1 Polar Star scheme

Before delving into the application of gTPS, it is crucial to acknowledge the computational challenges of exploring the energy landscape of biological macromolecules with iMapD. Notably, the efficiency of this algorithm, as mentioned in Section 2.1, heavily is tied to the computational cost of generating new viable molecular configurations in the shooting move. One is required to adopt an incremental value for the parameter c (which determines the strength of shooting, Equation (2.1.3)) and perform an energy minimization to obtain a chemically correct configuration, $\mathbf{X}^{\text{new}}$. However, under such constraints, we encountered difficulties in efficiently sampling conformations that were significantly different than the native structure of BPTI. To remedy this issue, we introduce an important improvement to the iMapD algorithm that we refer to as the "Polar Star" scheme. This enhancement is designed to make the algorithm more efficient and robust, enabling its applications to realistic systems, such as protein BPTI.

The key idea is to employ a specific type of biased simulation called Ratchet-and-pawl MD [234–236] (rMD) to drive the MD simulation toward a viable molecular configuration $\mathbf{X}$ starting from the boundary configuration $\mathbf{Q}_j^B$. In an rMD simulation, the equations of motion of the macromolecular system are modified by introducing a history-dependent biasing force defined as:

$$\mathbf{F}_B^i(x, q_m(t)) = -k_{\text{rMD}} \nabla_i q(x)(q(x) - q_{\text{m}}(t))\theta[q(x) - q_m(t)] \tag{4.1.1}$$

Here, the $\theta(x)$ is the Heaviside step function, and $x = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ denotes the set of atomic

coordinates. The $q(x)$ is a CV whose maximum value attained up to time t is denoted by $q_m(t)$.

The CV $q(x)$ is usually calculated via the contact map of instantaneous configuration x during the rMD simulation:

$$\tilde{q}(x) = \frac{\sum_{|i-j|>35}(C_{ij}(x) - C_{ij}^0)^2}{\sum'_{i,j} C_{ij}^{0\,2}}, \tag{4.1.2}$$

In this equation, $C_{ij}(x)$ is a switching function that approaches 1 when atom i and j are in contact and vanishes when they are far apart. In particular, we use:

$$C_{ij}(x) = \frac{1 - \left(\frac{r_{ij}}{r_0}\right)^6}{1 - \left(\frac{r_{ij}}{r_0}\right)^{10}}. \tag{4.1.3}$$

Here, $r_{ij} = |\mathbf{x}_i - \mathbf{x}_j|$ and $r_0 = 4.5$ Å is a threshold reference distance and C_{ij}^0 are entries of the contact map of the target state that we want rMD to reach. We note the restriction $\sum_{|i-j|>35}$ in Equation (4.1.2), which excludes from the summation pairs of atoms that belong to neighboring amino acids. This restriction is introduced to avoid the biasing force acting on atom pairs, which are subject to strong correlations determined by the local chemical structure of the chain. When applying rMD to protein folding, the C_{ij}^0 is often calculated using the native structure of the protein. However, in the application of rMD to Polar Star, the target contact map is defined as:

$$C_{ij}^0 = C_{ij}(\mathbf{Q}_k^{new}) \tag{4.1.4}$$

where $\mathbf{Q}_k^{new}$ is the set of Cartesian coordinates of a point outside the boundary of the explored region, generated with the original shooting move Equation (2.1.3).

Now, in rMD simulation, the system evolves as in plain MD as long as the dynamics spontaneously progress towards configurations with a smaller $q(x)$. Conversely, a harmonic biasing force, defined by Equation (4.1.1), switches on every time the overlap between the instantaneous $C_{ij}(x)$ and the target C_{ij}^0 decreases. Therefore, such simulation initiated from the boundary point $\mathbf{Q}_i^B$, rapidly yields a new viable molecular configuration outside the boundary, with a contact map close to C_{ij}^0.

This "Polar Star" scheme[1], schematically illustrated in Figure 4.1 dramatically improves the efficiency of the iMapD algorithm. Notably, by "dragging" the $\mathbf{Q}_i^B$ along the curvature of FEL, it eliminates the existing restriction on the strength of the "shooting move", i.e.

[1]The name "Polar Star" is based on the analogy with the ancient navigation scheme based on pointing towards a star (outside the Earth's surface manifold) to orient the sailing direction.

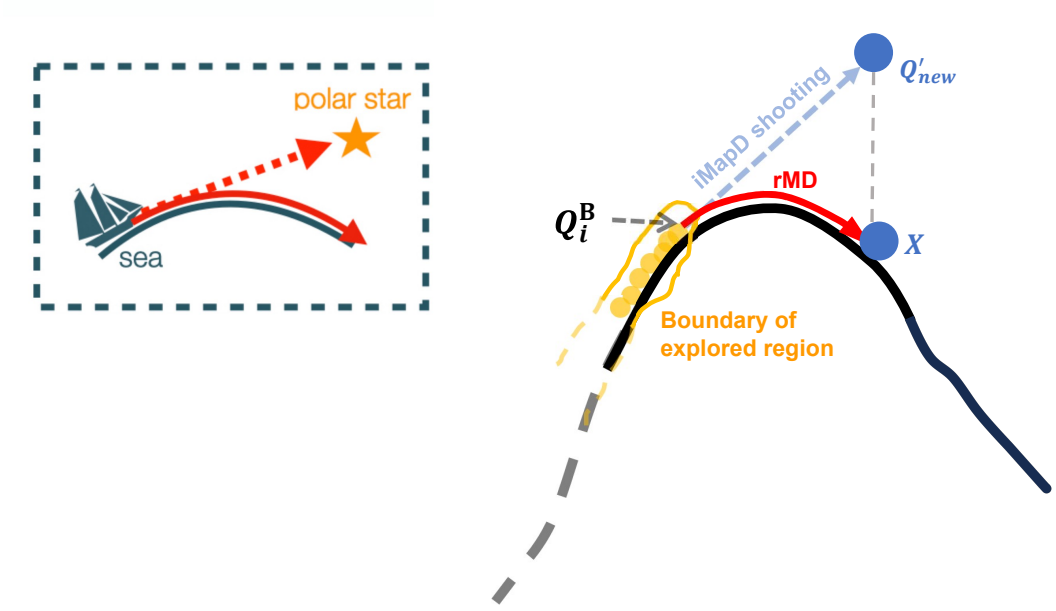


Figure 4.1: Schematic illustration of the Polar Star scheme, which has taken its name from the old marine navigation system. After performing the original shooting move of iMapD, we use rMD to "drag" $\mathbf{Q}_i^B$ toward the $\mathbf{Q}_{\text{new}}$. By the end of rMD, we obtain a new chemically viable configuration $\mathbf{X}^{\text{new}}$ which has a very close contact map to $\mathbf{Q}_{\text{new}}$.

the values of the parameter c. In fact, by performing iMapD with different large values of c, this modification enabled us to cope with the existence of different length scales in the macromolecule's energy landscape.

We also emphasize that the bias for the rMD dynamic is not chosen heuristically by the user. Instead, it is determined once iMapD produces the new set of coordinates outside of the boundary of previously sampled regions. Therefore, the new scheme of exploration remains fully uncharted and faithful to the original algorithm.

4.2 Applying gTPS to BPTI

4.2.1 Exploring BPTI's intrinsic manifold

After implementing the Polar Star modification into iMapD, we initiate the exploration of BPTI's manifold. We start the algorithm by initially solvating the crystal structure of the molecule in a square box of size 6-nm with 6744 molecules of TIP3P water using the forcefield AMBER99SB-ILDN, implemented in GROMACS [66]. The system was neutralized with 6 ions of chloride. Next, we performed energy minimization and equilibrated the system at 300 K in the NVT ensemble for 200-ps with an integration step of 1 fs. We resorted to a stochastic velocity rescaling thermostat, with coupling constant 0.1 ps^{-1}. Fi-

nally, we simulated for 10-ns in the same NVT ensemble. Configurations were saved every 100-ps.

An additional challenge in exploring the protein energy landscape is the co-existence of structures at different length scales. To cope with this issue, we integrated the results of iMapD simulations performed at 3 different values of shooting strengths $c = 0.5, 0.75, 1.0$. The contact map was calculated using the MDTraj package in Python [237]. To perform rMD simulations, we relied on an in-house modified version of GROMACS. Equivalently, this calculation could be performed using the publicly available PLUMED plugin[238].

For each choice of c, we performed 30 iMapD exploration cycles. In each cycle, 5 individual 1-ns long MD simulations were performed starting from configurations outside the boundary generated by the Polar Star. In each exploration cycle, the scaling parameter ϵ (Equation (A.2.1)) that defines the DMAP kernel was set to the average value of pairwise-RMSD between all the configurations sampled by iMapD up to that cycle.

After obtaining the combined dataset of the three separate runs of iMapD $\mathcal{C}_{\text{fin.}}$ (corresponding to $\sim 3\mu$s of cumulative simulation time), our first objective was to assess whether the iMapD algorithm can discover protein conformations close to those observed in the Anton simulations. In particular, using Q (fraction of native contacts) and RMSD, we projected the data of iMapD and Anton onto the space spanned by these two CVs. We note that the energy-minimized structure in the first frame of Anton data was used as a reference (and as analog to the native structure) for the calculation of these CVs. This projection hereafter is called the Q-RMSD plot.

The heatmap of the FEL calculated from Anton molecular trajectory in the Q-RMSD plot is depicted on the Figure 4.2. The contour lines of this figure will be used as the background for all the illustrations in the Q-RMSD plane. The Figure 4.3 illustrates that $\mathcal{C}_{\text{fin.}}$ entirely covers (and even surpasses) the FEL. Moreover, we project the Anton trajectory on one of the mentioned CVs, the RMSD, depicted in Figure 4.4.(a). Namely, the plain MD trajectory can be seen to traverse multiple transitions between the 5 conformational states. In panel (b_1) to (b_3), each iMapD run (projected onto the same CV) appears to visit some of these states as the exploration takes place.

Here, we must note the important addition of the Polar Star moves to iMapD. In particular, based on the result of Figure 4.4(b_1) for shooting strengths of $c \leq 0.5$, it is evident that iMapD even with the new modification has some difficulty in overcoming the energy barrier and necessitates an increase of c. However, in the original shooting move, $c = 0.5$ is already a high value concerning the stability of iMapD algorithm. Therefore, this result can be taken as evidence of the efficiency provided by the Polar Star modification.

Finally, in panel (c) of Figure 4.4, we report the cumulative distribution of RMSD eval-

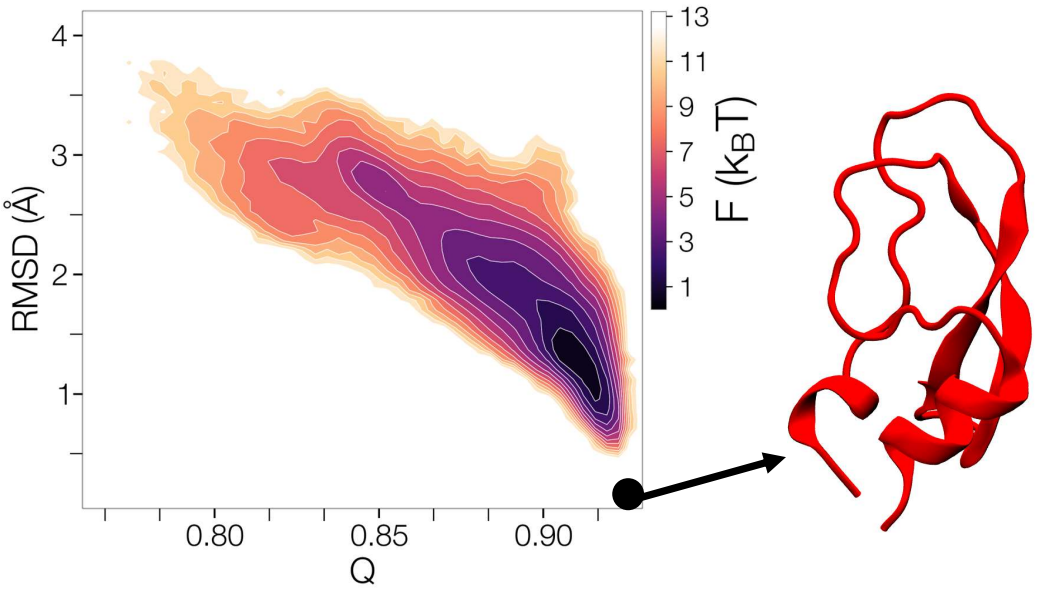


Figure 4.2: The heatmap of FEL near the native structure of BPTI. This was calculated using the frequency histogram of Anton's trajectory.

uated from a frequency histogram of all the configurations in the plain MD and iMapD trajectories. This comparison highlights that iMapD does not yield the Boltzmann distribution at room temperature and thus breaks the detailed balance.

Before continuing the application of gTPS to BPTI, we briefly discuss an analysis based on multiple structural features to demonstrate that iMapD has indeed visited most of the states observed in the Anton simulation. Additionally, we address an interesting question: Could the exploration of iMapD be achieved with comparable accuracy, using plain MD simulations at high temperatures?

Comparison between the structures in iMapD and Anton MD trajectory

As mentioned, the key structural differences between the 4 states other than the native structure include (i) the left-handed chirality of the Cys14-Cys38 disulfide bridge, (ii) the rotation rate of some of the aromatic side chains, (iii) changes in the exposed area of the molecule to the water and (iv) the unfolding of the small α-helix located at the N-terminus. Thankfully, representative configurations from each of these metastable states were provided in the supplementary material of [25] and are reported in panel (a) of Figure 4.5. The color code is consistent with the one adopted in the original paper.

To examine whether our iMapD exploration led to visiting the same states, we first gathered all the configurations generated by iMapD which satisfied Q$>$ 0.9 and RMSD$<$ 2$\mathring{A}$

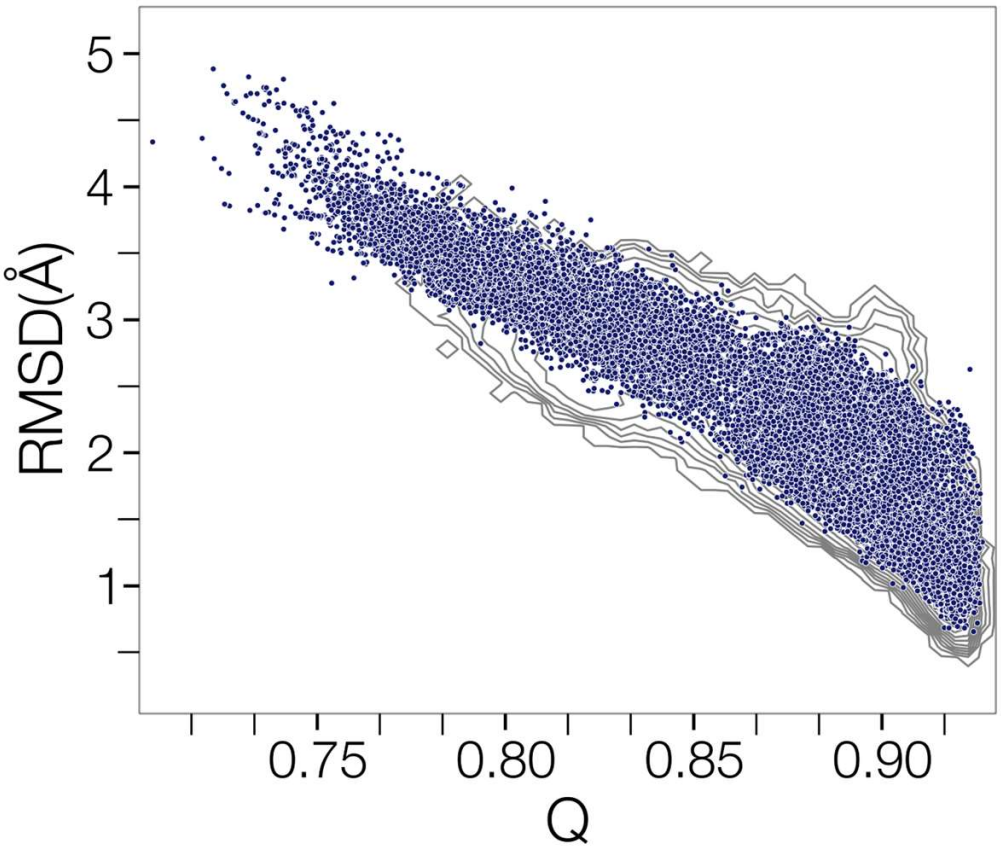


Figure 4.3: The dark blue dots are all the configurations of BPTI generated during the iMapD exploration performed with three different values of c (0.5, 0.75, and 1.0). Each iMapD run was performed for 30 cycles. The contour lines in the background highlight the structure of the free-energy landscape evaluated from a frequency histogram in Figure 4.2

from the structures provided by Anton, representative of each state. The configurations that satisfy the proximity criterion for multiple states were assigned to the one with the smallest RMSD distance. These sets of iMapD configurations are reported in panel (b) of Figure 4.5. According to this criterion, we could not find any iMapD configuration near state-3 (purple dot in Figure 4.5.(a)). We stress that this state is by far the shortest-lived among those detected in plain MD simulations.

Since iMapD breaks microscopical reversibility we could not rely on the dynamical analysis reported in [25] to assess the meta-stability of the configurations explored by iMapD. We, therefore, relied on an analysis of structural properties to check that the iMapD configurations associated with each state, are actually consistent with those found in the metastable states of Anton. In particular, the left-handed chirality of Cys14-Cys38 is one of the main differences between the two largest populated states, state-0 (depicted as the red dot in Figure 4.5(a) , occupied by 56% of Anton trajectory) and state-1 (the crystallographic state,

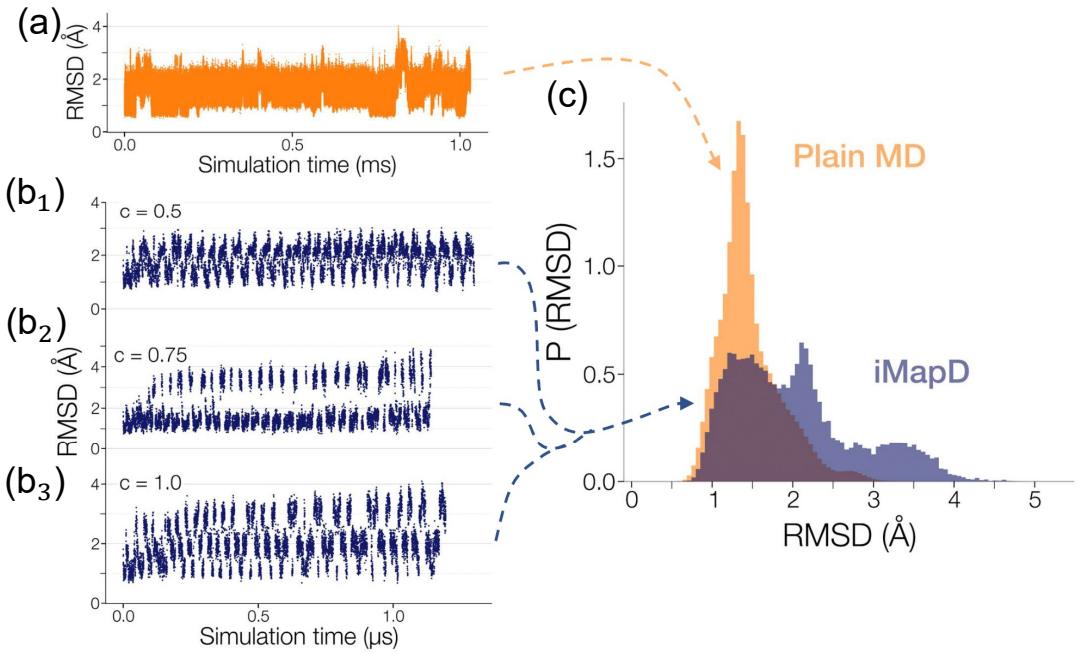


Figure 4.4: (a) The time series of the RMSD from the first frame (evaluated on backbone atoms) for Anton's plain MD trajectory. This frame is taken as a definition of the native state. (b_1)-(b_3) The equivalent time series evaluated along the exploration cycles of the iMapD, with three different values of the parameter c which controls the amplitude of the translation that drives the system outside the explored region. The oscillations visible in the RMSD time series of iMapD are associated with different exploration cycles. (c) The distributions of RMSD generated by plain MD and iMapD. This plot depicts the deviation of iMapD exploration from Boltzmann distribution.

blue dot in Figure 4.5(a), occupied 27% of the time). This property which can be characterized by the dihedral angle $\chi_3 \approx -90$, was evaluated for configurations in all the 4 groups of iMapD configurations, as reported in Figure 4.6. The left-handed disulfide bond in the configurations structurally very close to state-0 provides evidence that iMapD visited this state. Similarly, the characteristic structural properties of state-3 and state-4 observed in [25] were found in the configurations generated by iMapD, as shown in Figure 4.6. Another structural analysis, similar to the one reported in [25], involved the values of χ_2 dihedral angles in the aromatic side chains. We observed the same angles in the iMapD sets of configurations, as reported in Figure 4.7.

Comparing iMapD to high-temperature MD

As a final note on the application iMapD to BPTI, we compare its exploration of the IM with an exploration based on high-temperature plain MD. To this end, in Figure 4.8, we report the results of 300 ns of iMapD (corresponding to the specific choice $c = 1$) with three

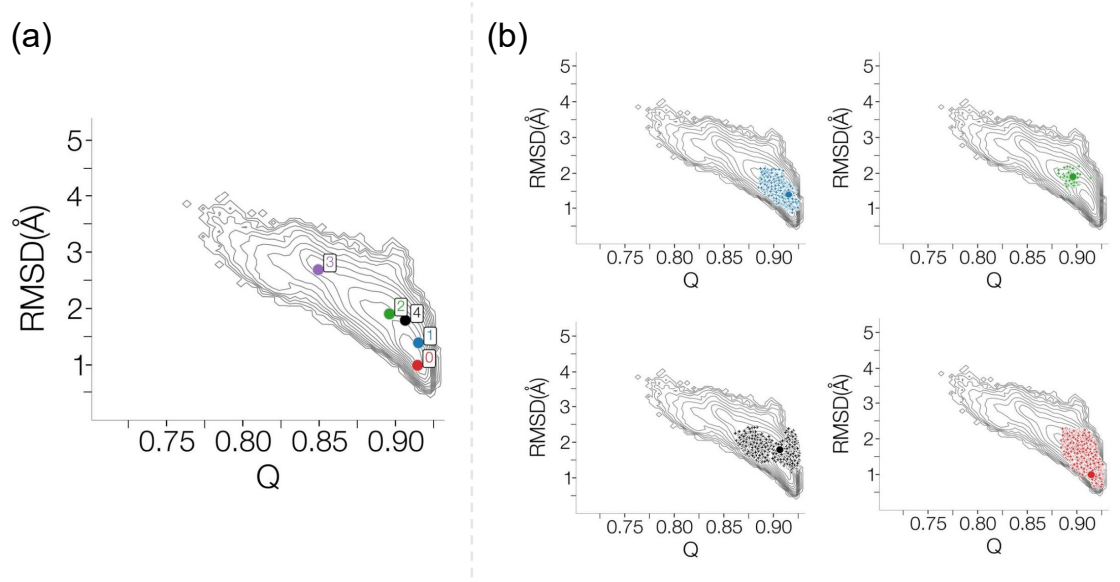


Figure 4.5: (a) The 5 configurations provided as supplementary material of [25]. They represent the 5 states observed in Anton plain MD trajectory. (b) We collected the configurations in iMapD data that satisfied Q$>$ 0.9 and RMSD$<$ 2$\mathring{A}$ from the structures in (a). These represented the configurations in the 5 states of Anton. Based on these criteria, we were not able to identify any configuration in the state-3, which was coincidentally the least populated state in Anton's trajectory.

equally long plain MD simulations, performed at $T = 340$, 360, and 380 K, respectively. For Q$>$ 0.9 and RMSD$<$ 1.5$\mathring{A}$, all of these simulations follow the global shape of FEL in the Q-RMSD plane. However, the configurations generated by high-temperature MD are in general confined in a region with Q$\gtrsim$ 0.85, while the iMapD data reach configurations with a lower fraction of native contacts and larger RMSD to the native state. Furthermore, for Q$<$ 0.9 region, the high-temperature MD configurations drift towards the region with RMSD$>$ 2$\mathring{A}$ (except for a branch of data at $T = 380$K), which is scarcely visited at room temperature. Conversely, the configurations generated by iMapD (which are based on short MD simulations at $T = 300$K) more faithfully profile the low free-energy regions, all the way to Q$<$ 0.8 and RMSD$>$ 2.5$\mathring{A}$.

4.2.2 Constructing network of transitions

Obtaining the vertices of the network

Next in gTPS, to harvest from $\mathcal{C}_{\text{fin}}$ a subset of configurations representative of finite regions in $\mathcal{R}$, we followed a different approach than the one discussed for ALA. Namely, we applied and compared 3 different methods of structural clustering [107, 239]: KMeans, Hierarchical clustering (HC) with unweighted average linkage, and HC with Centroid linkage. We placed two criteria to identify the best clustering for this step: (i) the distribution of

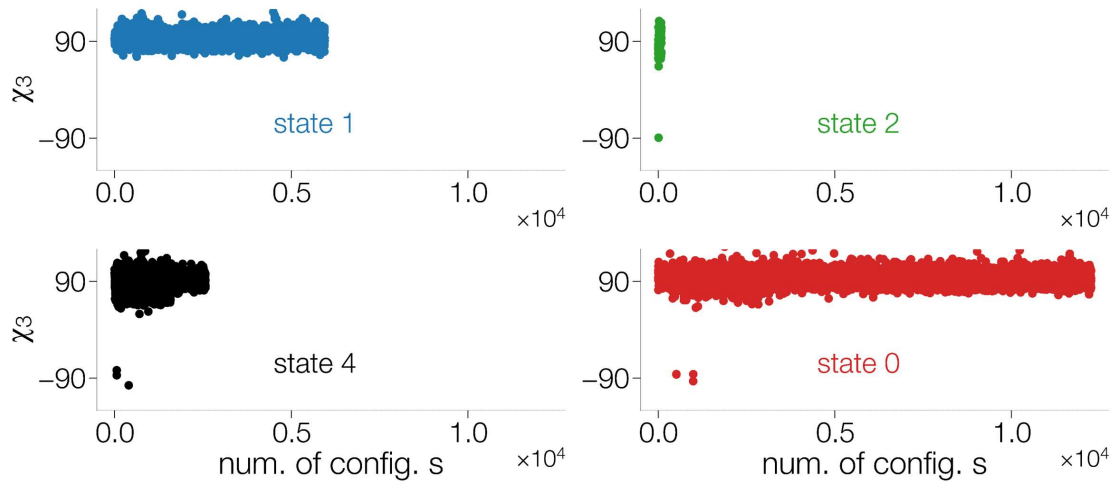


Figure 4.6: The Cys14-Cys38 bridge dihedral angles (χ_3) for the configurations of iMapD data that correspond to each state of Anton. Consistent with [25], we verify that all the states except state-1(crystallographic) contain configurations with left-handed bridges.

nearest-neighbor RMSD calculated over the centroids of the clusters should be narrowest, denoting a uniform spacing of the centroids, and (ii) The centroids of the clusters should be uniformly distributed in the Q-RMSD plane. However, we still have to consider the limitation of DWave hardware that the coherence and henceforth performance decreases as the number of required qubits increases. Due to this reason, we maintained the number of clusters to $\nu = 80$ which is similar to the case of ALA.

The KMeans approach aims to cluster the data while minimizing the average intra-cluster Euclidean distance between members of each cluster. Formally, the objective function that we try to minimize is:

$$\arg\min_{\zeta} \sum_{k=1}^{m} \sum_{i \in \zeta_k} |\mathbf{Q}_i - \mu_i|^2 \tag{4.2.1}$$

where ζ is the set of m-clusters, and μ_i is the centroid of each cluster defined as:

$$\mu_i = \frac{1}{|\zeta_j|} \sum_{j \in \zeta_i} \mathbf{Q}_j \tag{4.2.2}$$

In our application, KMeans was performed using the scikit-learn [240] package with C_α-atoms' coordinates taken as the main initial features of the data set. We set the algorithm to find the optimum clustering after 20 runs of randomly initializing the centroid of the clusters using the "*kmeans++*" built-in method. The algorithm was then terminated once, in successive iterations, the distance between the consecutive cluster centers fell below $0.1\mathring{A}$. The centroid of the KMeans' clusters can be observed in Figure 4.9(a) where each panel

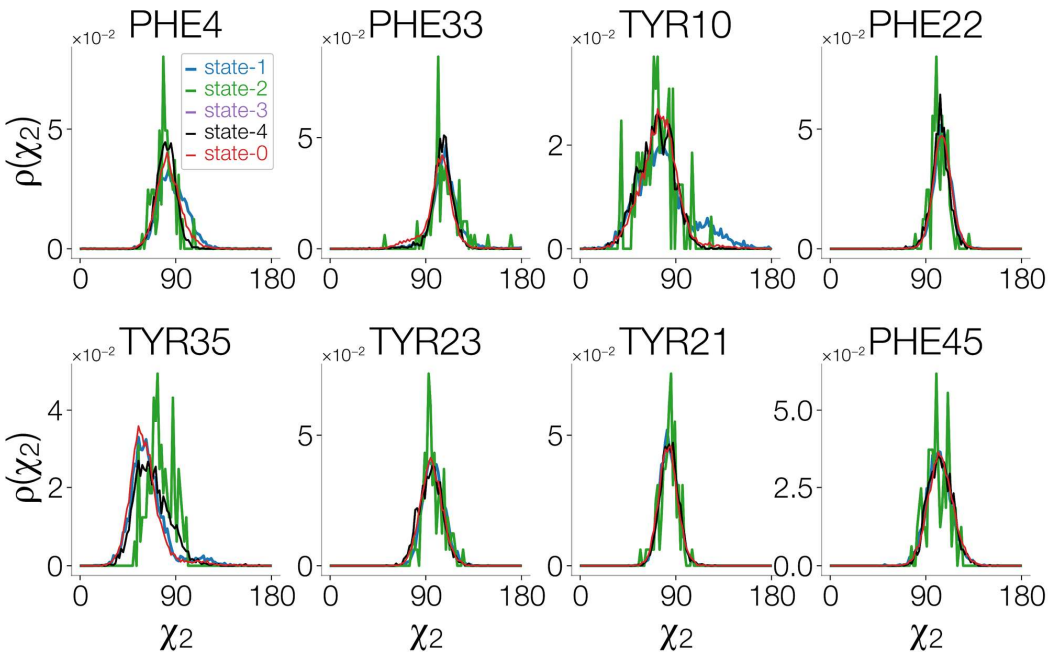


Figure 4.7: The distribution of χ_2 dihedral angles in certain aromatic side chains of BPTI. This figure corresponds to the same analysis done in [25].

shows the representative configurations in the Q-RMSD plane (top) and the distribution of nearest neighbors (bottom).

Alternatively, in HC, we start by taking every point as an isolated cluster. Then, at each step, we form new clusters by connecting those that previously had the minimum "linkage" distance. For HC with unweighted average linkage –UPGMA method–, this distance is defined as

$$D(A,B) = \frac{1}{|A||B|} \sum_{X_a \in A} \sum_{X_b \in B} d(X_a, X_b) \tag{4.2.3}$$

where (in our application) the $d(X_a, X_b)$ is the RMSD of the backbone atoms. Conversely, in the Centroid version of HC –UPGMC method–, the dissimilarity/distance between clusters is taken as the Euclidean distance between the centroid of each cluster

$$D(A,B) = |\mu_A - \mu_B| \tag{4.2.4}$$

where $\mu_K = (1/n_K) \sum_{X_k \in K} X_k$. This linkage distance can be alternatively expressed starting from the RMSD distances between configurations:

$$D(A \cup B, C) = a_1\, D(A,C) + a_2\, D(B,C) - a_3\, D(A,B) \tag{4.2.5}$$

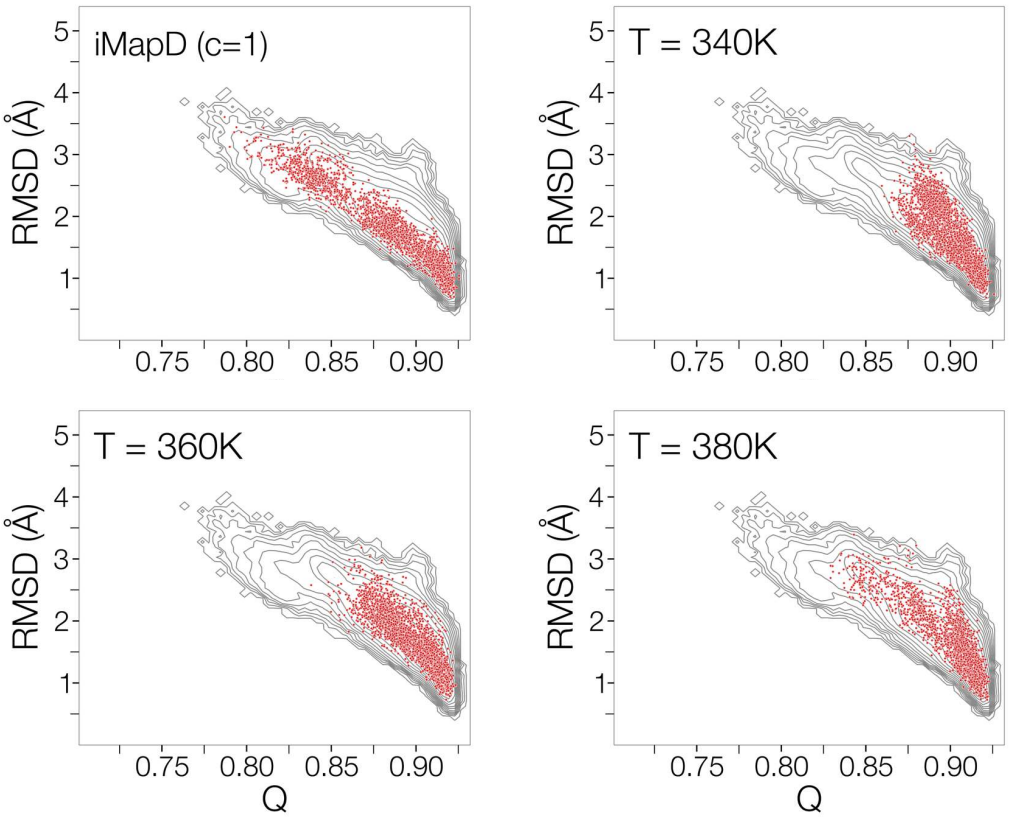


Figure 4.8: Configurations generated by 300 ns of plain MD at three different high temperatures and 300 ns of cumulative iMapD exploration at ($c = 1$), projected onto the Q-RMSD plane. This result demonstrates an advantage of iMapD instead of increasing the temperature for enhancing the molecular configuration. In iMapD, since every short MD is performed at 300K when we follow the gradient of FEL in the shooting step, we remain more faithful to regions of low free energy.

where

$$\begin{aligned} a_1 &= \frac{|A|}{|A|+|B|} \\ a_2 &= \frac{|B|}{|A|+|B|} \\ a_3 &= \frac{|A||B|}{|A|+|B|} \end{aligned} \tag{4.2.6}$$

Such an expression is practically more useful as it allows to utilization pre-computed pairwise RMSD matrices instead of taking Euclidean coordinates of configurations as the argument.

For both HC approaches, we relied on the SciPy package [241] with default values for all the parameters. After obtaining the set of clusters (in each method), we collected those configurations that possessed the lowest intracluster RMSD distance as the centroid of each

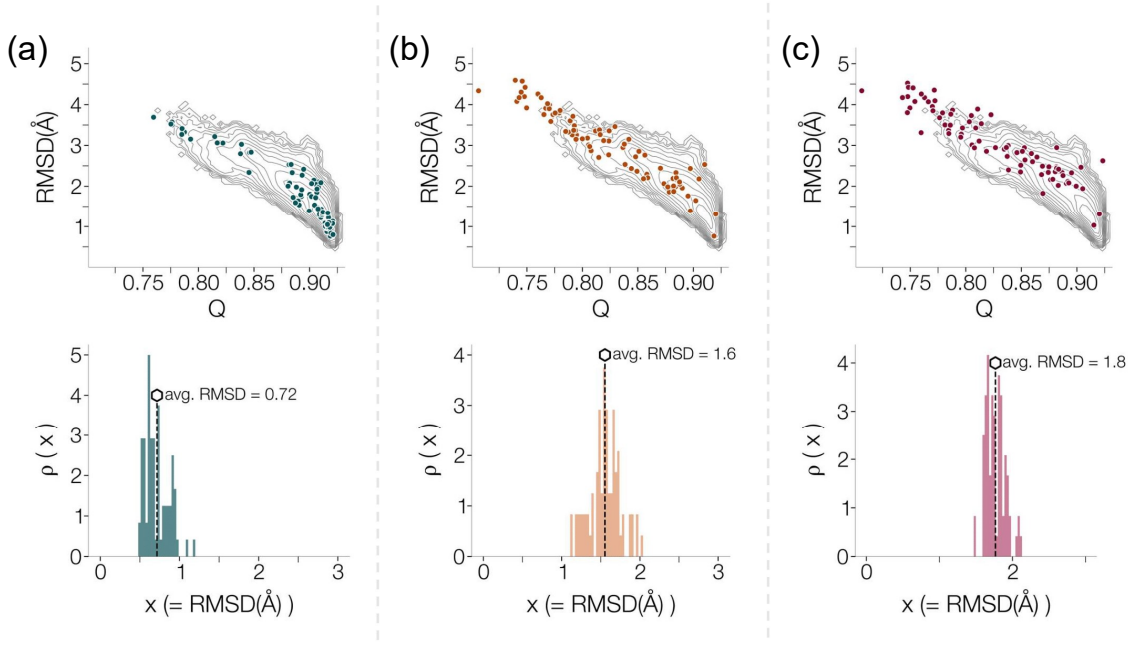


Figure 4.9: Results of clustering iMapD configurations using (a) KMeans, (b) HC with Average linkage, and (c) HC with Centroid linkage. In all panels, Q-RMSD plots on the top depict the representative configurations characterized by the least average intra-cluster RMSD as the centroid of clusters. On the bottom, we have plotted the nearest neighbor RMSD distance between these centroids. Based on these results, we have chosen HC with Average linkage as the representative of $\mathcal{R}$ regions in the gTPS framework.

cluster. The results are reported in Figure 4.9 (b-c).

Based on a comparison between all the panels of Figure 4.9, we identify the clustering of the UPGMA method as the one that better meets the two criteria we were looking for. We denote $\mathcal{S}$ as the set of configurations representing the centroids of the resulting clusters in UPGMA. Next, we continue by utilizing $\mathcal{S}$ to build the network of transitions for BPTI. The Figure 4.10 depicts the centroid configurations of the clustering test overlap onto the whole iMapD dataset.

Connecting the vertices

In the next step of the gTPS framework, we adopted the procedure indicated by Equation (2.2.29) in Section 2.2 to calculate the V_{cg}. Namely, we first established σ –the radius of finite space regions in $\mathcal{R}$– as the half RMSD between the nearest neighbors configurations in $\mathcal{S}$, i.e. $\sigma = 0.8\mathring{A}$. Next, by running 100 individual plain MD simulations ($t = 500$-ps), we evaluated the average time T^{σ}_{avg} it takes for the MD to appear σ RMSD distance from every $\mathbf{Q} \in \mathcal{S}$. As explained in Section 2.2, the inverse of this quantity, which signifies the

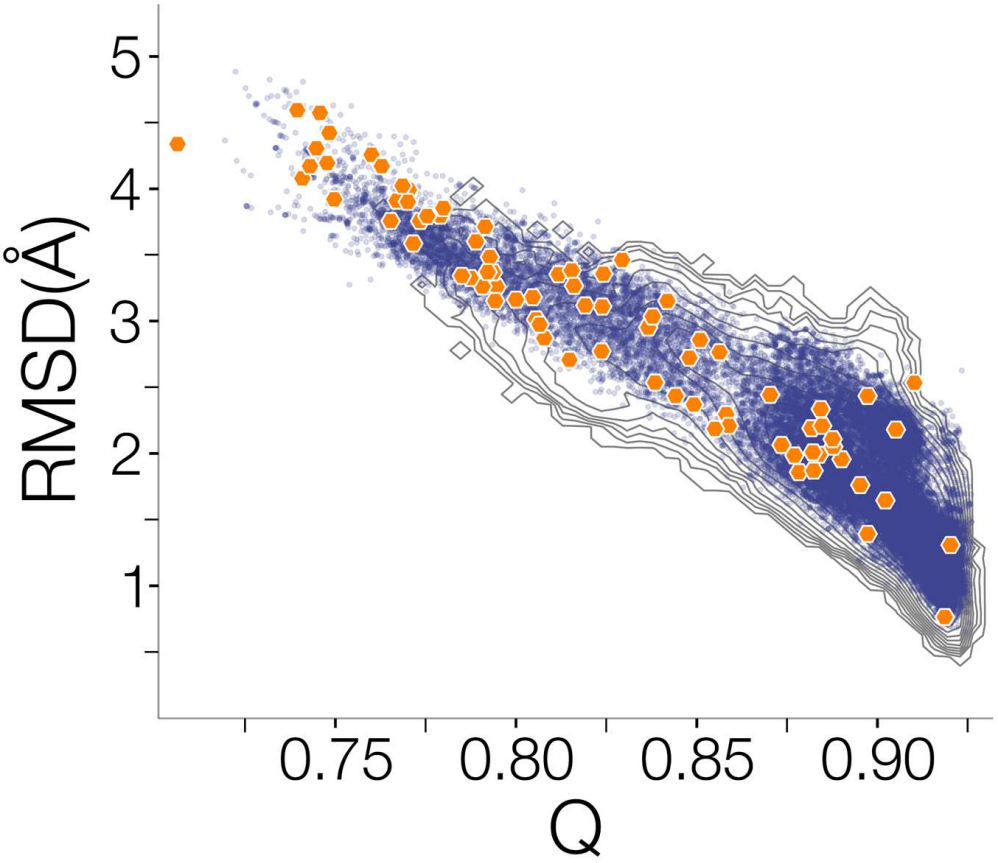


Figure 4.10: The dark blue dots are all the configurations of iMapD and the orange dots are the $\nu = 80$ representative centroids obtained after clustering in Figure 4.9.

rate of escaping the regions in $\mathcal{R}$, also determines the $V_{\text{cg}}(\mathbf{Q})$:

$$V_{\text{cg}}(\mathbf{Q}) = \frac{1}{T^{\sigma}_{avg}} \tag{4.2.7}$$

The result of this calculation can be observed in Figure 4.11, wherein panel (a) illustrates the average escape time in ps, and panel (b) provides the heatmap of V_{cg} in Q-RMSD plane for configurations in $\mathcal{S}$ After obtaining the V_{cg}, we build the connections between the centroids or vertices in $\mathcal{S}$ (representative of finite size regions $\mathcal{R}$) by first establishing edges between those that satisfy relative RMSD $\leq \sigma$. Similar to the application of ALA, we argued that every vertex in this network must retain at least 2 edges (except the source and target vertices) connecting it to the rest. This is an essential criterion for visiting every vertex (and henceforth every region in $\mathcal{R}$) while sampling pathways. Therefore, to guarantee this, we next identified every vertex with degree < 2 and established connection(s) from it to the utmost 2 of its neighbors –albeit with RMSD $> \sigma$. Finally, to ensure the topological connectivity of the graph, we linked each disconnected component to its nearest neighbor

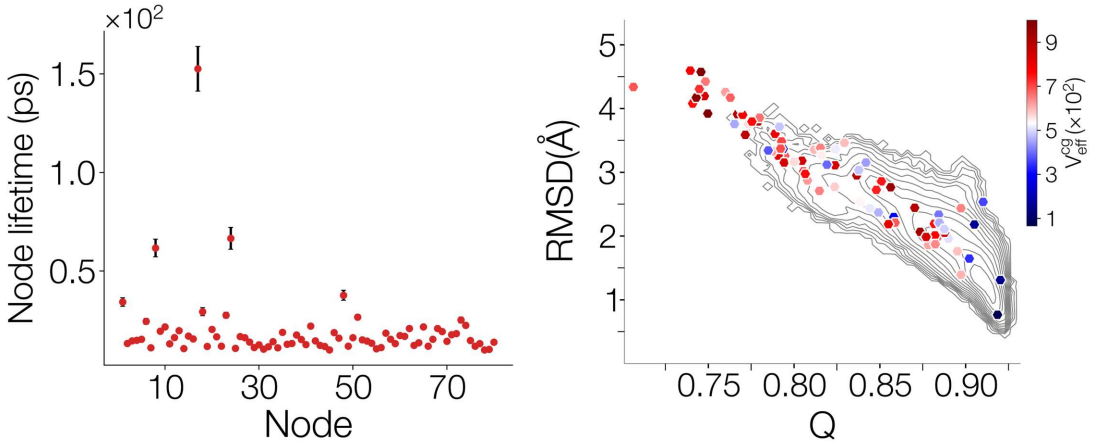


Figure 4.11: The results of the calculation of the CG effective potential $V_{\text{eff}}^{\text{cg}}$ associated with each node of the graph (i.e. the clustered data shown in Figure 4.10). (a) V_{cg} was calculated as the inverse of node lifetime. (b) The value of V_{cg} in the Q-RMSD plane.

using their vertices with minimum RMSD.

Once the network was established, we calculated the weights of each edge using V_{cg} and the Section 2.2.3.

4.2.3 Sampling transitions paths in the basin of BPTI's native structure

To sample pathways from the transition network in the basin of BPTI's native state, we again resort to the DWave annealing machine. For this task, we first identified the two configurations in $\mathcal{S}$ that in the Q-RMSD plane had the largest distance yet overlapped with the FEL contour lines. Then, after assigning them as the source and target nodes for the paths, we implemented the corresponding H_{QUBO} into the DWave machine according to gTPS. The required number of qubits for encoding our network was 207 (equal to the number of edges and vertices). This was a significant reduction compared to the case of ALA.

Before initiating the MC sampling in gTPS, it is necessary to evaluate the conditional probability $P(\mathbf{I}|t_{\text{fin.}})$ (Equation (2.3.7)). Following the application of gTPS to ALA, this step requires a substantial amount of QPU time (about 4 minutes cumulatively). Unfortunately, for the application of gTPS to the BPTI molecule, we only possessed the commonly available time for computation with the DWave machine (20 minutes of hybrid solver usage or equivalently about 40 sec of QPU time). Therefore, as such we could not fully apply the gTPS framework. Nevertheless, we utilized the DWave hybrid solver to generate multiple high-probability transition paths to assess the validity of the overall framework of gTPS in the present application.

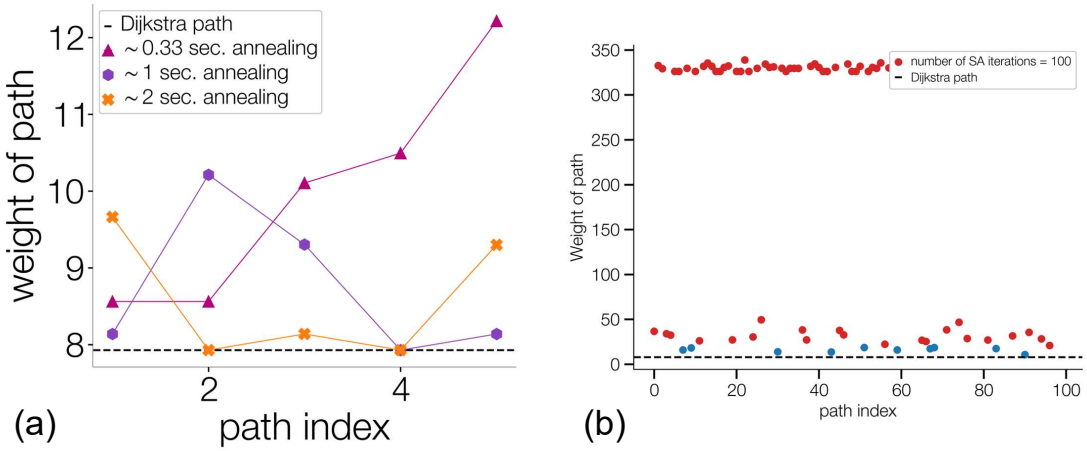


Figure 4.12: (a) Comparison between the statistical weight of the paths generated with D-Wave annealing, and the most probable path obtained by Dijkstra. All the generated paths with annealing only take $\mathcal{O}(1)$-s of QPU time and have weights within factor 2 of the Dijkstra path. (b) In contrast, in our experiments with classical optimization algorithms of SA, only $\sim 10\%$ of the paths had low weights/action (the blue dots), with the majority being $1-2$ orders of magnitude higher than the Dijkstra path. However, it requires $O(10^3)$-s of computation on a single-core classical computer. It is worth noting that this result also demonstrates the gap provided by H_{QUBO} formulation in Section 2.3.

In our annealing cycles, the generation of a single transition path required $\mathcal{O}(10^1)$ s of hybrid solver time (corresponding to $\mathcal{O}(1)$ s of QPU time). For comparison, we have also performed several analog calculations resorting to fully classical simulated annealing (as implemented in OCEAN). Even after increasing the sweeping number to 10^6, which took $\mathcal{O}(10^3)$ s of computational time on a single core, most of the obtained paths have relatively low statistical weight: indeed, their action W_p is typically two orders of magnitude larger than that of the least-action path. However, a small portion of these paths (roughly 10%) has an action within a factor two of that of the least action path. In contrast, the actions of all the paths generated with the hybrid quantum-classical are within a factor 2 from that of the least action path (Figure 4.12) In Figure 4.13.(a), we assess the most probable pathway in our network obtained by the Dijkstra algorithm against the corresponding transition observed in Anton's trajectory. Panel (b) reports some of the unique stochastic trajectories obtained by QA in DWave. In Figure 4.14, we have depicted all the D-Wave paths (whose weights are shown in Figure 4.12) and compared them against the Dijkstra path. As expected, Anton's MD trajectory, the most probable path, and the stochastic trajectories evaluated using DWAVE, all reach the final destination by traveling along the low-free energy region. However, upon closer inspection, a significant portion of the MD trajectory can be seen to accumulate in the vicinity of one of the frames in the gTPS network, indicating the presence of a meta-stable state. This observation highlights the

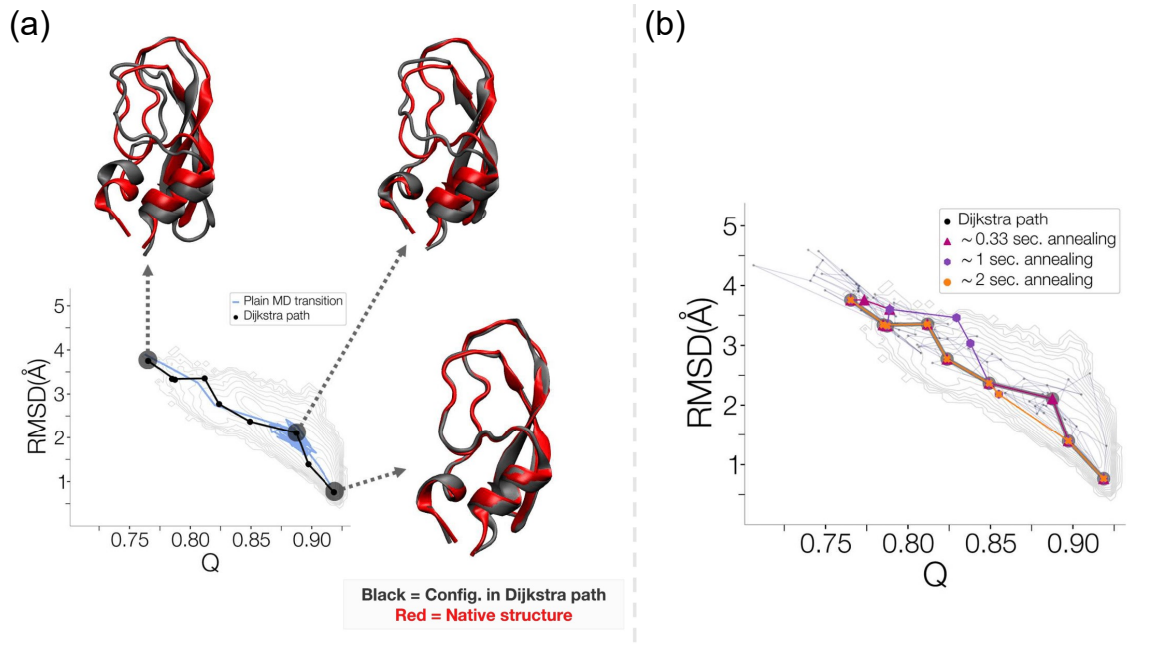


Figure 4.13: (a) The blue line follows the configurations in the transition pathway connecting two distant points in the plane selected by the RMSD to the native state and the fraction of native contacts, obtained by plain MD using the Anton supercomputer. Except for the beginning and the end of the line, the rest depicts the result of an average of over 50 ns-long windows. The solid black line is the most probable path computed in the gTPS scheme, calculated using the Dijkstra algorithm. Selected representative conformations along this path are also included. To illustrate the evolution, these conformations (Black) are superposed onto the native structure (Red). (b) Some representative gTPS transition pathways were sampled using the DWAVE quantum computers and compared with the Dijkstra path.

potential limitation of the gTPS framework, discussed in Section 2.2, as it solely considers overcoming a single free-energy barrier in transitions and may overlook the exploration of metastable states. Meanwhile, in Section 2.2, we also discussed how gTPS yields the lower bound of the transition path time for the stochastic trajectories on the network of transitions. For the least action path shown in panel (a) of Figure 4.13, we obtain a lower-bound estimate $t \sim 500$ ps using Equation (2.2.40). This time scale cannot be directly compared with the transition path time observed in the corresponding plain MD trajectory since the latter involves overcoming more than one energy barrier. However, we analyze the MD transition pathway by projecting it onto the RMSD to the native state and isolating the sections involving a transition between metastable basins. Based on this simple analysis we found the strict transition time of the blue line in Figure 4.13.(a) (excluding the time in a metastable state) to be $t \sim 1.2$ ns. Hence, the transition time observed in MD is about a factor of 2 longer than the lower bound estimate provided by gTPS. This is illustrated in Figure 4.15.

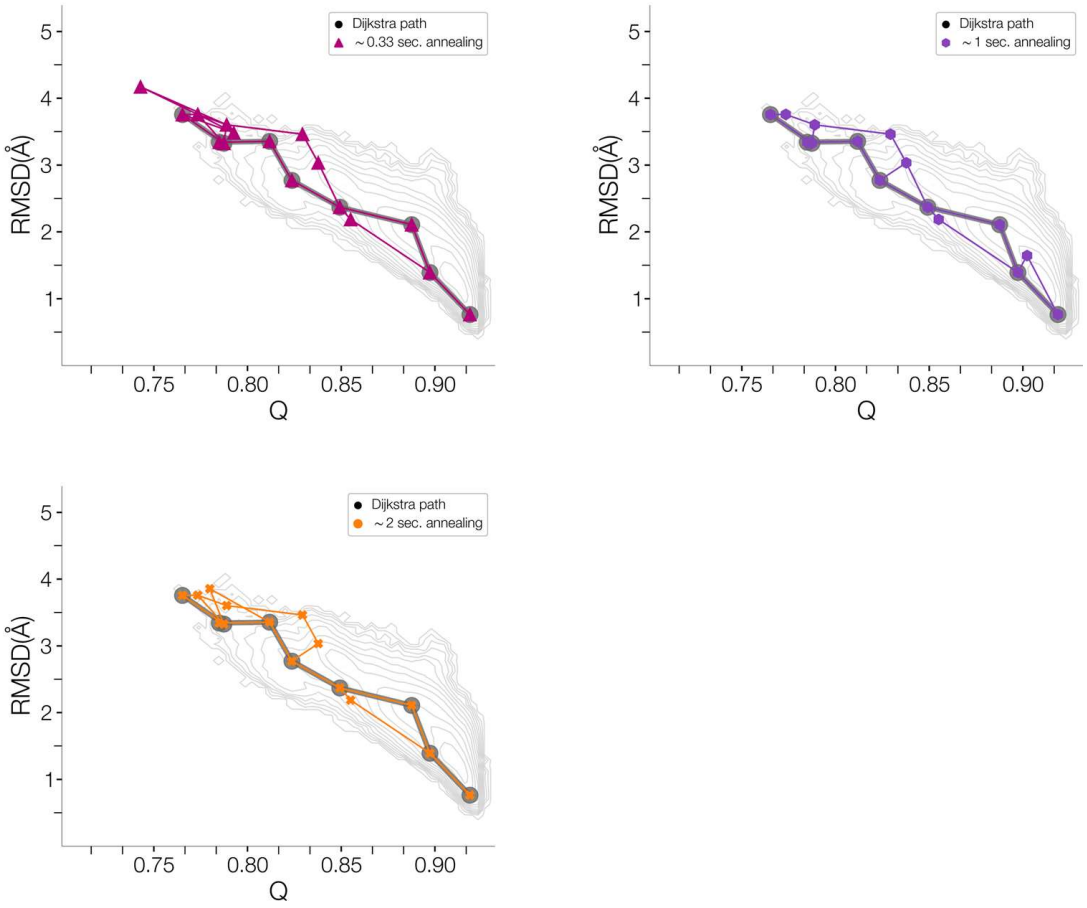


Figure 4.14: Additional transition paths computed using the DWAVE quantum computer and compared with the most probable path obtained with the Dijkstra algorithm.

4.3 Discussion

This chapter was dedicated to the first application of the gTPS framework to a macromolecular system of biological relevance. In particular, we investigated rare structural rearrangements near the native structure of BPTI that spontaneously occur on a millisecond time scale. Our main objective was to demonstrate the capability of current quantum computers and, overall the gTPS framework in investigating challenging protein transitions with full atomic resolution.

By utilizing Anton's trajectory and a modified version of the iMapD algorithm, we demonstrated that gTPS can retrieve conformations of BPTI observed in the plain MD simulation, however, with orders of magnitude lower computational cost. Subsequently, we encoded our effective description of the dynamics into a small network of transition. We note that the constraint on the size of the network is mainly due to the physical limitations present in current QC hardware. Then, we performed several QA cycles with

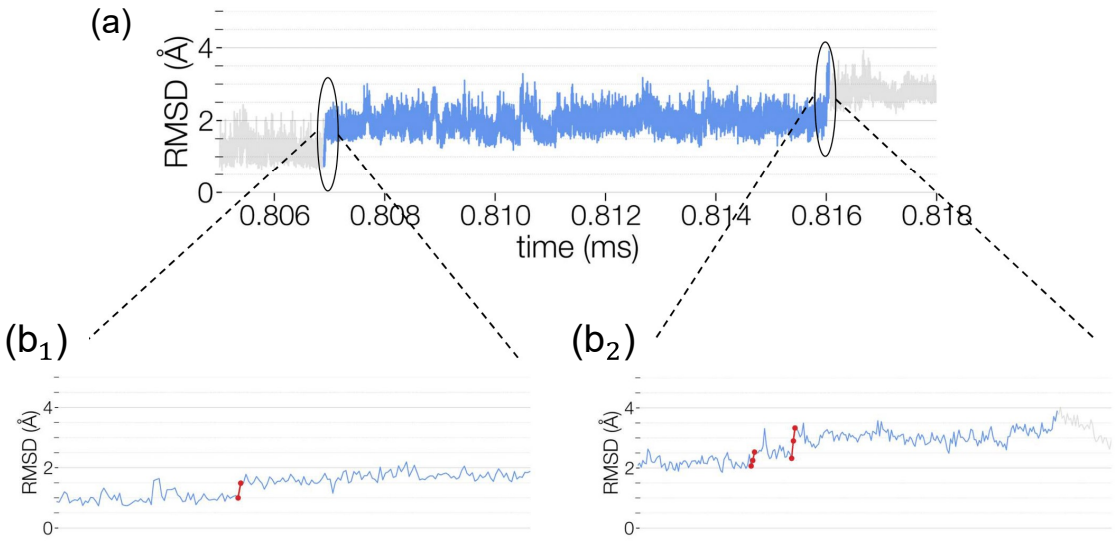


Figure 4.15: (a) The plain MD transition path for the conformational transition of BPTI shown in Figure 4.13. (b) projected on the RMSD from the first frame of Anton's trajectory. The 8 red points have been used to estimate the barrier-crossing time scale $tt \sim 1.2$ ns.

DWave to generate trial transition pathways on the network. The result demonstrated that the DWAVE quantum computer can indeed generate viable paths at an affordable computational cost (with a QPU calculation time of a few seconds per path). Moreover, the results in Figure 4.14 and 4.13 depict how these paths correctly predict the regions of low free energy.

Amplifying the iMapD's shooting move with the Polar Star scheme enabled us to stabilize the overall application of gTPS. We were able to demonstrate the efficiency provided by this addition as we gradually increased the parameter c in Figure 4.4. This result, in combination with the validity of our network of transition in correctly identifying the regions of low free energy, motivated the following question: "Can we utilize the gTPS framework to sample transition pathways along the unfolding of a protein?". We should point out that between folding and unfolding mechanisms, we intuitively expect that the latter would pose a lower level of difficulty to investigate. Since starting from a denatured state, a protein system has to overcome a large "entropic" barrier to reach the folded state, we reason that iMapD would probably be less effective in studying the folding mechanism. To answer this question in the next chapter, we discuss the preliminary results of applying gTPS to the unfolding of two separate molecules.

CHAPTER 5

Ongoing investigations

Understanding the folding/unfolding of proteins is one of the central goals of studies on biological macromolecules. Starting from the ribosomes in what is often referred to as a coiled/denatured/unfolded conformation, proteins have to rearrange their structure in a unique 3D shape to become functional in the body of living organisms. Many numerical and computational approaches (of which we gave just a few examples in Chapter 1) have been mainly developed to address the task of characterizing the underlying mechanism of this rare event.

Following the successful applications of gTPS in previous chapters, we are currently investigating the unfolding of the 35-residue subdomain of the chicken Villin headpiece (HP35, PDB code: 2f4K). This molecule is one of the smallest natural polypeptide chain that autonomously folds into a globular structure with an experimental folding rate of $\sim 4\,\mu$s [242]. The reversible folding/unfolding process of this system has been studied with Anton using $128\,\mu$s of plain MD [26]. In the last two decades, several other studies have also investigated the folding of this molecule which has become a standard benchmark system. Notably, it has been shown that HP35 folds through 3 general channels: (Channel-I) Characterized by the helix-2 and helix-3 segments forming their native structure first (Figure 5.1.(a)), then the helix-1 is formed. This is the mostly observed pathway in the folding of HP35 [26, 243]. (Channel-II) Conversely, the helix-1 and helix-2 can be formed first and then helix-3 [26, 244]. (Channel-III) Most recently a new possible pathway was reported where all three helices cooperatively form their native bond structure [245]. We note that, out of the three, this pathway seldom occurs in the folding of HP35 and in fact, it was not observed in the data of Anton.

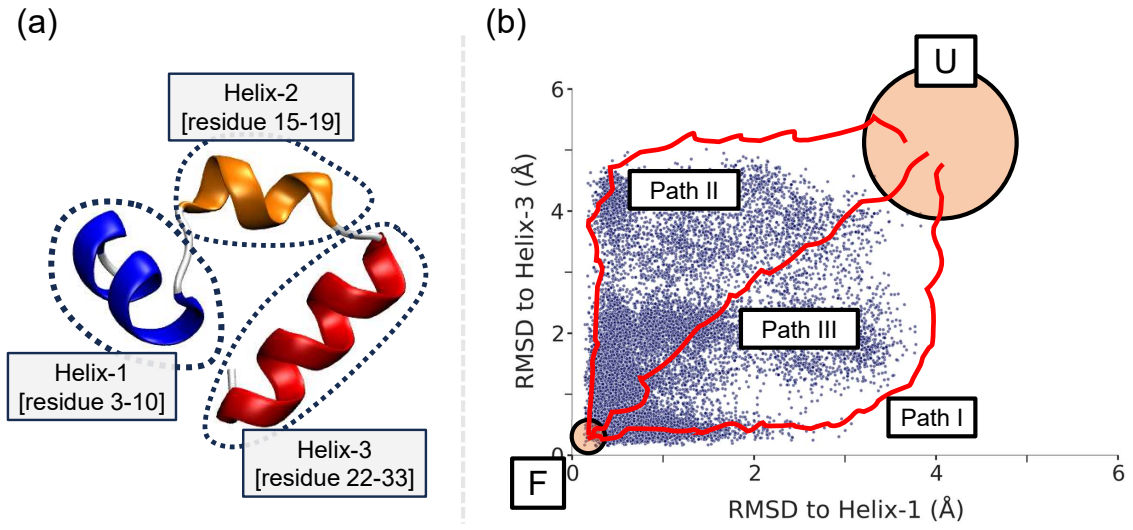


Figure 5.1: (a) The HP35 subdomain of Villin protein. The three helices whose order of formation determines the folding pathway are indicated. (b) The initial sampling of HP35 unfolding with iMapD, at $c = 1.0, 5.0$, for a cumulative simulation time of $\sim 2\mu$s. The data is projected onto the RMSD with respect to the Helix-1 and Helix-3's C_α atoms. Here, F denotes the folded state and U denotes the unfolded state. Previous computational studies revealed that 3 main folding channels exist for this molecule. The 3 red lines in this plot show a schematic illustration of possible pathways in each channel. The path (Path-X) notation is according to in which channel (Channel-X) they can appear. In Channel-I, which is the most observed in previous studies, Helix-3 and Helix-2 form their native contacts first and then Helix-1 is formed. Alternatively in Channel-II, Helix 1 and 2 are the ones to form first. Channel-III is the only one that does not follow a hierarchical formation of helices' native contacts.

We employ all these results as a suitable ground to examine whether gTPS can in fact sample transition pathways for a reaction that occurs through multiple distinct channels. Namely, in both of previous applications, our focus was to investigate whether gTPS is capable of correctly characterizing the low free energy regions for a reaction. Therefore, we were not particularly concerned that in those cases the transition effectively occurred through one channel. However, after observing success of gTPS, folding/unfolding of HP35 with 3 transition channels is arguably a suitable next application in which to test our framework. In Figure 5.1.(b), we have shown our initial application of iMapD to HP35 and a schematic representation of 3 previously identified folding pathways. In Figure 5.2, samples of iMapD configurations in 3 important folding intermediates along each pathway are depicted.

Meanwhile, we have also initiated a separate numerical experiment concerning the unfolding pathways of BPTI protein. As mentioned in the previous chapter, this molecule, due to its size and its chemical structure that contains 3 pairs of Cystine residues, has been extensively studied as a folding model for many disulfide-rich proteins. Due to the

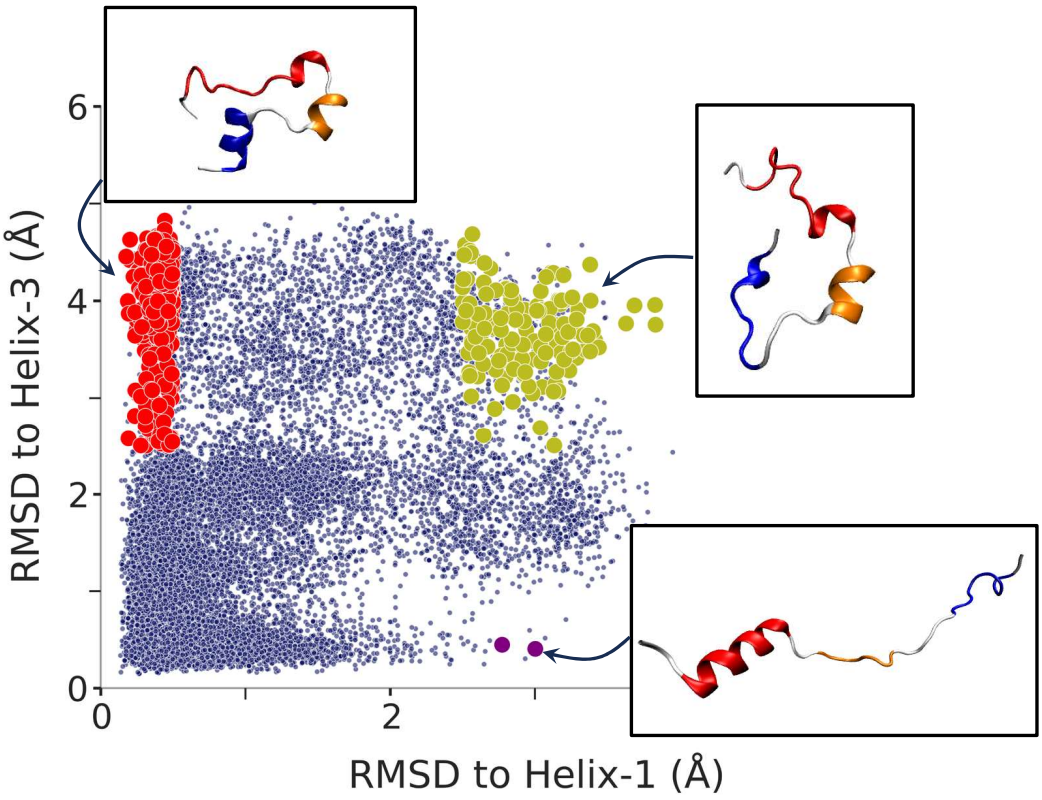


Figure 5.2: Sample iMapD configurations that reside in 3 important intermediates along the HP35 folding pathways. These intermediates are adapted from [245].

present of these bridges which play a stabilizing role for the folded structure of BPTI [246, 247], the folding (unfolding) can only be achieved experimentally through an oxidative (reduced) buffer which allows for forming and breaking of the bonds. In particular, starting from native state N in the unfolding pathway, the solvent exposed disulfide bridge Cys14-Cys38 first has to be reduced to form an intermediate known as N^{SH}_{SH}. Then, remarkably the kinetically preferred mechanism for the unfolding involves intramolecular disulfide rearrangements rather than direct sequential reduction of the remaining disulfide bonds [248]. This means that in the transition pathway, we first observe intermediates such as N', where Cys14-Cys38 is reformed and Cys5-Cys55 bridge is reduced, before observing conformations with only one disulfide bridge. We have depicted the widely accepted pathways for oxidative folding and reductive unfolding of BPTI in Figure 5.3. In the application of gTPS to the unfolding pathway of BPTI, we replicate the reductive environment by manually removing the bonds from the molecule and adding a hydrogen atom to the sulfur atom of the Cystine residues before performing the iMapD simulation. In such an approach, the distances between the disulfide bridges signify the breaking and forming of the disulfide bridges. To the best of our knowledge, a computational study on this system using unbiased sampling

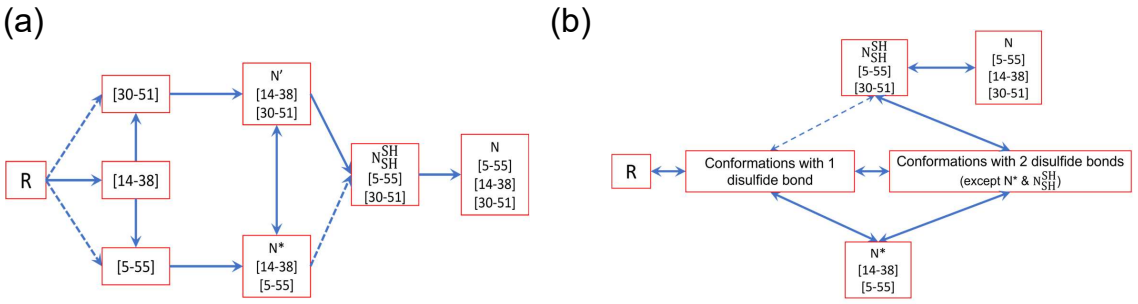


Figure 5.3: The folding/unfolding pathways based on experimental studies. (a) The unfolding pathways in an oxidative environment. The representation here is a combination of the results of Weissmann and Kim [249] and Mousa *et al.* [246]. The [.] indicates which disulfide bridges are formed in the conformations of a state. The R indicated the reduced BPTI conformation and N the native (b) The unfolding pathway as suggested by Mendoza *et al.* [248]. In both folding and unfolding pathways, the N^{SH}_{SH} is a rate-limiting step. Furthermore, the dashed arrows indicate a very slow kinetic rate along the respective transition. Therefore, for folding, the majority of the transition takes place through N′ state.

does not exist and often biased methods are utilized to recover the experimental behavior of bond rearrangements described above. It would be both fascinating and illuminating to test if gTPS can recover unfolding/folding pathways that involve the same behavior. We believe success, in this case, could mark an advancement in validating the utility of current NISQ devices for computational studies in biomolecular sampling. The preliminary results of gTPS (only the application of iMapD) on this system can be observed in Figure 5.4 and Figure 5.5.

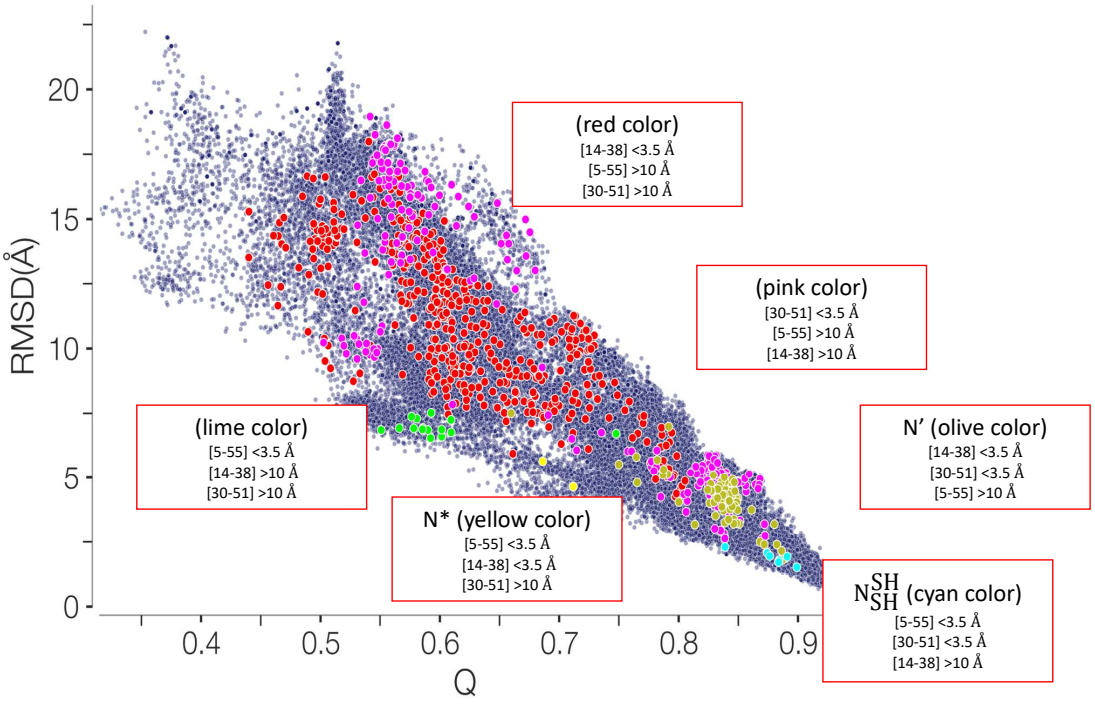


Figure 5.4: The application of iMapD to the BPTI molecule with all the disulfide bonds removed. The dark blue dots are the result of 4 separate runs of iMapD with $c = 1.0\,, 2.0\,, 5.0\,, 10.0$ which was made possible using our Polar Star shooting method. The cumulative time needed to obtain this data was $\sim 7.5\mu$s. In our simulations, the minimum distance between sulfur atoms was measured to be $3.0\mathring{A}$. Therefore, we consider $3.5\mathring{A}$ as the distance that corresponds to an existing bond between the sulfurs. With this in mind and also considering $10\mathring{A}$ as the threshold for a broken disulfide bridge, we have colored the iMapD configurations that correspond to the states in Figure 5.3.(a). We have indicated the respective criteria for each state in the box corresponding to it. We note that the hierarchical expansion of the projected configurations on the Q-RMSD space seems to follow the folding pattern indicated by experiments. In particular, immediately after the basin of the native state, the cyan-colored dots appear which indicate the N_{SH}^{SH} state. Furthermore, we can observe that iMapD has spent more time in the most dominant pathway for the folding, as both the N^* state and the state with only [Cys5-Cys55] bridge have remained rather unobserved.

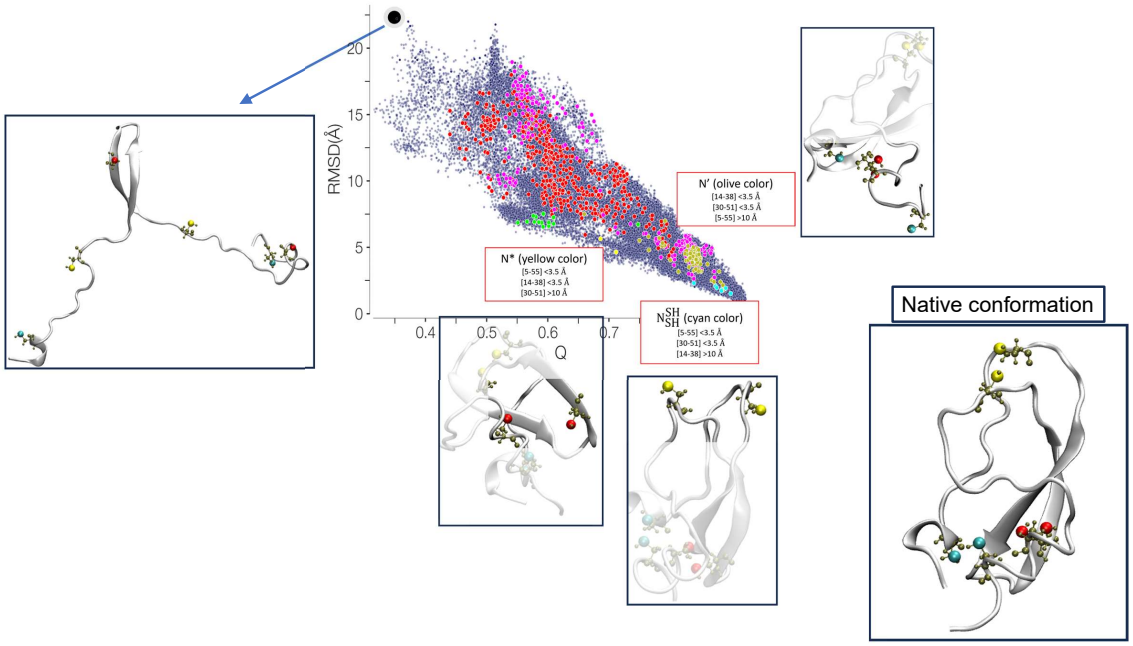


Figure 5.5: Representative configurations of the 3 important states for the folding/unfolding mechanisms. These configurations have been chosen randomly from those who satisfied the criteria for each state (indicated in the corresponding box). The native structure and the conformation with the highest RMSD from the native have also been shown. The Cys residues are indicated by the gray ball and stick representation. The sulfur atoms corresponding to the native disulfide bonds are indicated with the same color. Specifically, sulfur atoms of [5-55] bond are indicated by the color cyan, [14-38] with yellow, and [30-55] with red.

CHAPTER 6

Conlusion

In this thesis, we established a novel computational framework, gTPS, for sampling the full transition path ensemble of rare molecular conformational transitions. gTPS integrates the data-driven method of iMapD for rapid exploration with an MC importance sampling scheme that exploits the potential of quantum computing. This unique feature allows gTPS to tackle outstanding problems in computational methods based on classical computation. In the specific, the dimensionality reduction techniques in iMapD allow us to efficiently navigate the uncharted portions of the configuration space without needing to define a CV or biasing force. However, the original formulation of iMapD is heavily limited by the restrictions on the shooting parameter c. Addressing this limitation in Chapter 4, we introduced our modified version of iMapD with the addition of the Polar Star scheme to the algorithm's shooting move. Polar Star strictly uses the configurations generated during the shooting move of the iMapD to initiate an rMD simulation. Then, in every step of the simulation, it generates an instantaneous configuration that lies outside of previously observed regions of the IM while adhering to the chemical and topological constraints imposed by the structure of the molecule. Therefore, Polar Star method eliminates the restriction of maintaining an incremental value on c and as such greatly enhances the iMapD algorithm. What is important here is that this modification does not obstruct the generality of iMapD (and its unsupervised nature) in application to different systems, specifically in the cases wherein we have no prior insight into the relevant CVs.

By borrowing the renormalization group approach from nuclear and sub-nuclear physics in the next step of gTPS, we proceed to develop a CG representation of the microscopic dynamics following the underlying assumption of the Langevin equation. In this effective

theory, every configuration of iMapD is associated with a finite space region of the configuration space whose union covers the entire portion of IM observed by the uncharted exploration. Subsequently, by encoding our CG theory into the weights of an undirected graph between these regions, we build a transition network where the probability of observing any path is directly determined by the summation of the weights along that path. We emphasize the most important difference between this approach and that of MSMs: the method of gTPS does not require estimating the whole stochastic transition matrix/rate matrix. Instead, it only requires computing the lifetime of each node, a task that can be efficiently carried out even as the iMapD's exploration is taking place. On the other hand, we acknowledge that gTPS can only provide limited information about the kinetics compared to MSMs.

After establishing the network, a QUBO formulation for finding the shortest path on a network allows us to encode the network and the path sampling problem into an actual QA named D-Wave. We integrated this method of utilizing a QC into an MC sampling scheme where the QA generates trial pathways for us. Our main motivation was that –as we argued in Section 1.3.1 and Section 2.3– the intrinsic quantum fluctuation and measurement of the quantum adiabatic switching implemented in D-Wave, would fundamentally decorrelate the sampling every time we perform a trial pathway. Meanwhile, compacting the transition information into a network and subsequently using quantum superposition to encode "all" the pathways into the QA, is the key in allowing MC sampling to perform "global" moves. We recall from Section 1.3, how such an approach was effectively out of the reach of conventional TPS methods. Finally, by utilizing the D-Wave itself to calculate the acceptance/rejection probability of a Metropolis criterion, we account for the physical noise of the machine due to coupling to its environment [217, 218]. Therefore, it guaranteed that we obtained the correct statistical weight in every trajectory we sampled.

In Chapter 3, we presented the first application of gTPS to a molecular system by studying the $C_5 \rightarrow \alpha_R$ transition of ALA. Even though ALA is relatively small compared to other biological macromolecules, it retains many crucial features that resemble those present in the much larger counterpart systems. Conversely, its modest size is perfectly fitted as a benchmark system as the first application of gTPS. We demonstrated how a coherent combination of our CG theory embedded in an undirected network and encoding provided by the QUBO formulation, allows us to successfully sample "uncorrelated transition paths" for ALA's $C_5 \rightarrow \alpha_R$ transition that follow the low energy regions of FEL. Even though in this benchmark, we restricted our sampling to 23 transition paths, in general, the number of trajectories is only limited by the available computational resources. This achievement marked an essential advancing step in tackling outstanding challenges in conventional TPS approaches that are based on classical computers.

Success in the case of ALA prompted us in Chapter 4 to further investigate the potential of our approach and test its scalability. There, we presented the application of gTPS to near-native conformation rearrangements of BPTI which is considerably a larger system than ALA. The transitions in the basin of native conformation of BPTI spontaneously occur in the time scale between micro and milliseconds. Previously, it was made possible to study these computationally only by utilizing the specialized supercomputer Anton [25]. On the other hand, gTPS, by employing a few GPUs on a classical computer cluster and a few hundred qubits on the DWAVE was able to characterize, with the same atomic resolution, these rare transitions.

The algorithm we presented in this thesis is designed to sample the full transition path ensemble. The transition path ensemble is often heterogeneous, displaying several alternative transition channels, corresponding to alternative molecular mechanisms. Therefore, concerning the capabilities of gTPS in this regard, we are currently studying its application to the unfolding transition of the HP35 domain of chicken Villin headpiece and the reduced molecule of BPTI. Both cases contain new unexplored challenges wherein we wish to test gTPS. First, we are seeking to test if gTPS can indeed search through all the transition channels available. Second, in the specific case of BPTI, the folding occurs in the time scales of minutes in oxidative buffer and in the scale of hours for the unfolding in reductive buffers. Therefore, it would be interesting to check if gTPS can successfully retrieve pathways for such a distant and complex transition.

While significant effort has been made towards designing quantum algorithms for quantum many-body problems, only a few applications of quantum computing to classical molecular sampling problems have been reported to date [157, 250]. As we mentioned, most of these attempts assume a simplified molecular representation, among which lattice discretization [178, 251–256]. Our algorithm shows that quantum computers can be successfully applied to investigate challenging protein transitions with full atomic resolution and in explicit solvent. In particular, the DWAVE quantum computer can generate viable transition pathways at an affordable computational cost (with a calculation time of a few seconds per path depending on the size of the network, using the OCEAN's hybrid solver).

A very important quest to be pursued in this phase of the development of quantum technologies is identifying possible new fields of applicability. In the future, heterogeneous platforms for high-performance computing might emerge that fully rely on integrated ML and quantum computing approaches. These new machines will require scientific advancements to be able to fully take advantage of their strengths. As the size of quantum computing hardware continues to grow in size and performance, we may hope that within the foreseeable future, the approach introduced in this work may provide a computationally efficient scheme to perform path sampling calculations for complex macromolecular transi-

tions. Due to the suppression of autocorrelation time, we expect that, for sufficiently large networks, our hybrid scheme may ultimately have an edge over classical stochastic methods.

APPENDIX A

Appendix

A.1 Dominant Reaction Pathways

In the Dominant Reaction Pathway (DRP) approach [197, 210], we are concerned with obtaining the pathways that contribute the most to the propagator $P(\mathbf{Q}_f, t|\mathbf{Q}_i)$ in Equation (2.2.8). We start by assuming that the initial and final configurations

$$\begin{aligned} \mathbf{Q}(t_i) &= \mathbf{Q}_i \\ \mathbf{Q}(t_f) &= \mathbf{Q}_f \end{aligned} \tag{A.1.1}$$

are located in the reactant and product states respectively, hence the $P(\mathbf{Q}_f, t|\mathbf{Q}_i)$ will solely pertains to the reactive pathways. Next, to identify the dominant pathway $\bar{\mathbf{Q}}(t)$ we minimize the S_{OM} through applying the variational principle:

$$\delta S_{\mathrm{OM}}[\mathbf{Q}(t)] = \delta \int_{t_i}^{t_f} \frac{\dot{\mathbf{Q}}(t)^2}{4D} + \mathrm{V}_{\mathrm{eff}}[\mathbf{Q}(t)] = 0 \tag{A.1.2}$$

which leads to the Euler-Lagrange equation

$$\ddot{\bar{\mathbf{Q}}}(t) - 2D\, \nabla \mathrm{V}_{\mathrm{eff}}[\bar{\mathbf{Q}}(t)] = 0. \tag{A.1.3}$$

In principle, numerical integration of Equation (A.1.3), subjected to the boundary conditions in Equation (A.1.1), leads to the path of least action or the DRP. However, in practice due to the separation of time scales, the task of evaluating even a single DRP typically requires integration over $\sim 3N \times \frac{t(\mathrm{react})}{t(\mathrm{min})} = 3N \times 10^{12}$ time-steps. Here the $t(\mathrm{react})$ is the

average time for the reaction to occur –e.g seconds in typical protein folding– and $t(\text{min})$ is the minimum time scale present in the dynamic $\sim$ 1-picosecond [257]. Evidently performing such a task is computationally challenging.

Fortunately, we can "side-step" this issue by reformulating the problem into an alternative description that does not involve integration over time. The key is to realize that the DRP path described by Equation (A.1.3), coincides with the dominant path in a family of pathways along which the "effective energy"

$$E_{\text{eff}} = \frac{\dot{\mathbf{Q}}^2}{4D} - \mathrm{V}_{\text{eff}}[\mathbf{Q}] \tag{A.1.4}$$

remains conserved. We note that this effective energy is different than the total physical energy $\mathbf{E}$. In particular, the friction and noise present in Langevin dynamics does not allow $\mathbf{E}$ to remain conserved.

The E_{eff} allows us to parameterize the same family of pathways using the energy-dependent Hamilton-Jacobi (HJ) action (instead of time-dependent Newtonian dynamics):

$$S_{\text{HJ}} = \int_{\mathbf{Q}_i}^{\mathbf{Q}_f} dl \sqrt{\frac{1}{D}(E_{\text{eff}} + \mathrm{V}_{\text{eff}}[\mathbf{Q}])} \tag{A.1.5}$$

Therefore, once the value of effective energy is established, we then proceed to numerically minimize the target function

$$S_{\text{HJ}} = \sum_j \Delta l_{j,j+1} \sqrt{\frac{1}{D}(E_{\text{eff}} + \mathrm{V}_{\text{eff}}[\mathbf{Q}_j])} - \lambda P \tag{A.1.6}$$

to obtain the DRP. Here $\Delta l_{j,j+1} = \sqrt{(\mathbf{Q}_{j+1} - \mathbf{Q}_j)^2}$ is the Euclidean distance between subsequent steps and $P = \sum_j (\Delta l_{j,j+1} - \langle \Delta l \rangle)$ is a penalty function to keep the $\Delta l_{j,j+1}$ close to their mean value [258, 259].

The advantage of utilizing S_{HJ} over S_{OM} is due to the replacement of differential time steps with incremental (spatial) displacements dl along each trajectory $\mathbf{Q}(t)$. Contrary to the former in the dynamics of macromolecules', the length scales generally do not exhibit any gap. Therefore one may expect that the number of elementary steps required to converge to the path of least action in the discretized S_{HJ} is much lower than in its counterpart S_{OM}. In fact, in the example of protein folding, the transition typically occurs in an overall displacement of order $O(10^2)$ bigger than the smallest spatial scale of the protein, the size of a single atom. Needless to say, identifying the DRP(s) by numerically minimizing the HJ action Equation (A.1.5) is drastically simpler compared to the Newtonian description of Equation (A.1.3).

We finish this section by deriving an expression for the time it takes for the system to traverse the DRP (or any other pathway that conserves the E_{eff}) in the HJ formalism. The velocity of the system along any pathway is given by

$$|\dot{\mathbf{Q}}| = \frac{dl}{dt} \tag{A.1.7}$$

By recalling the alternative expression in Equation (A.1.4), we obtain

$$\frac{dl}{dt} = \sqrt{4D(E_{\text{eff}} + \mathrm{V}_{\text{eff}})} \tag{A.1.8}$$

which will lead to

$$\Delta t_{i\to f} = \int_{\mathbf{Q}_i}^{\mathbf{Q}_f} \frac{dl}{\sqrt{4D(E_{\text{eff}} + \mathrm{V}_{\text{eff}})}} \tag{A.1.9}$$

This is the expression for the traveling time of the system along the DRP pathway.

A.2 How to build the Diffusion maps?

Assuming to have a dataset $\mathcal{C}$ with N members and each with d dimensions, we start by building a Gaussian kernel

$$K(Q_i, Q_j) = e^{-\frac{d(x_i,x_j)^2}{\epsilon}} \tag{A.2.1}$$

where $x_i, x_j \in \mathcal{C}$. The $d(x_i, x_j)$ denotes an affinity function which is supposed to be a local measure of dissimilarity. In all the applications that we have discussed in this thesis, we have adopted RMSD as our affinity function, however in general, it might be advantageous to assign a more case-specific measure. Since the matrix elements in Equation (A.2.1) are analog to the single-step transition probabilities of a random walker that travels between the data points [200], the scaling factor ϵ effectively controls which x_i and x_j are allowed to be connected.

Next, we build the diffusion matrix $\boldsymbol{P}$

$$\boldsymbol{P}_{ij} = \frac{1}{D_i^{0.5}} K(Q_i, Q_j) \frac{1}{D_j^{0.5}} \tag{A.2.2}$$

where $D_{ii} = \sum_j K_{ij}$ (consequently $D_{ii} = \sum_j K_{ij}$). In the next step, we solve the eigenvalue problem $\mathrm{P}\psi = \lambda\psi$ and retrieve the $p = d \times N$ eigenvectors, $\{\psi_1, \psi_2, \ldots, \psi_p\}$, that are referred to as DMAP components. The lowest eigenvalue λ_0 and the lowest eigenvector ψ_0 are trivial and in terms of diffusion are related to the equilibrium condition.

A.3 Principal component analysis

PCA is a dimensionality reduction technique that performs an orthogonal linear transformation on the dataset (contained in $N \times d$ matrix $\boldsymbol{X}$) [260]. The components of PCA are directions in the original space of data and they are ordered according to the amount of variance they entail. The transformation in question is defined by the set of $p \ll N$ vectors of size d, $\{\mathbf{w}_i\}_{i=1,\dots,p}$ where the dot product of projection of dataset on $\mathbf{w}_i$

$$c_i = (\boldsymbol{X}\mathbf{w}_i)^T \boldsymbol{X}\mathbf{w}_i = \mathbf{w}_i^T \boldsymbol{X}^T \boldsymbol{X}\mathbf{w}_i \tag{A.3.1}$$

defines the variance of data in the direction of vector $\mathbf{w}_i$. Therefore, to identify the first component of PCA we must identify the vector that maximizes c_i:

$$\mathbf{w}_1 = \arg\max[\mathbf{w}^T \boldsymbol{X}^T \boldsymbol{X}\mathbf{w}] \tag{A.3.2}$$

The subsequent k component can be found by first removing the variance contained in the previous $k-1$ components:

$$\boldsymbol{X}_k = X - \sum_{i=1}^{k-1} \boldsymbol{X}\mathbf{w}_k\mathbf{w}_k^T \tag{A.3.3}$$

Then, we can apply the same procedure as above to find the new component that contains the maximum variance in $\boldsymbol{X}_k$:

$$\mathbf{w}_k = \arg\max[\mathbf{w}^T \boldsymbol{X}_k^T \boldsymbol{X}_k\mathbf{w}] \tag{A.3.4}$$

Alternatively, we also recognize that $\boldsymbol{X}^T\boldsymbol{X}$ is the covariance matrix of the data [260]. Therefore, solving the eigenvalue problem of this matrix would provide the eigenvectors and eigenvalues which are respectively the $\mathbf{w}_i$ and the variance c_i.

A.4 Dijkstra algorithm

The Dijkstra algorithm is concerned with analytically calculating the shortest path from a source vertex v_s to the target vertex v_t. For simplicity, we assume to have a completely connected network that connects the two vertices (no isolated node). The algorithm goes as follows:

1. All the vertices in the graph except v_s are assigned a *tentative distance* d_t. The tentative distance is going to evaluate the length of shortest path from vertex v_s to

any other vertex. In the initial cycle: Since the only path known is from v_s to itself, its distance is $d_t(v_s) = 0$ and every other vertex has $d_t = \infty$. Initially, we assign all the vertices except v_s as unvisited. The v_s is marked as visited and also the current vertex.

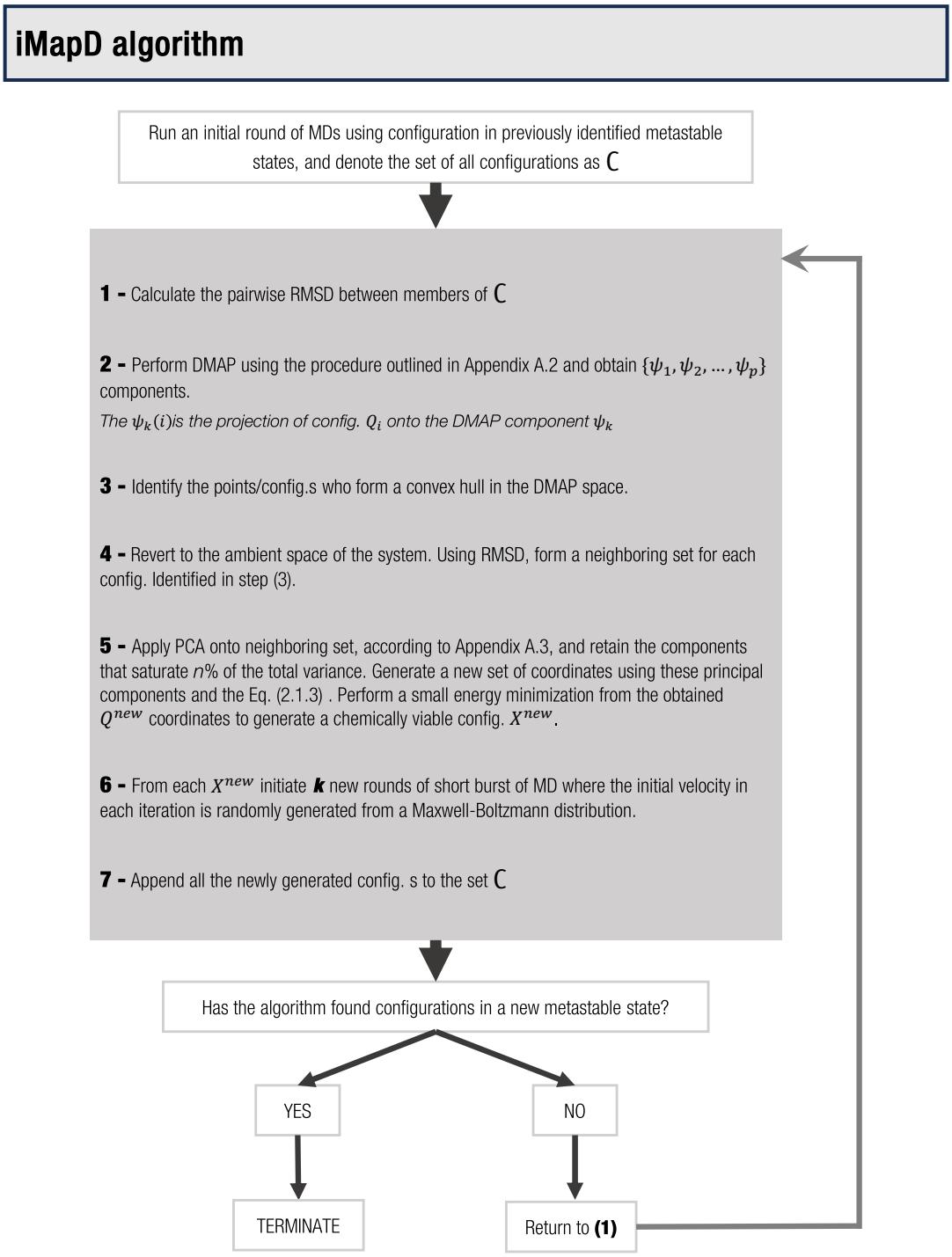
2. (Applies to all the steps) From the currently visiting vertex, we calculate the respective tentative distance of all of its neighbors. Assume we are currently at vertex v_a with tentative distance of $d_t(v_a) = 10$ and v_b is one of its neighbors that is connected to v_a via an edge of weight $w_{ab} = 2$. The tentative distance of v_b (at the current step) is then $d_t(v_b) = d_t(v_a) + w_{ab} = 12$. However, it could happen that in previous steps $d_t(v_b)$ was calculated to be greater than 12. In this case then, we discard the previous tentative distance calculated and update $d_t(v_b) = 12$. Otherwise, we retain the previous distance from before.

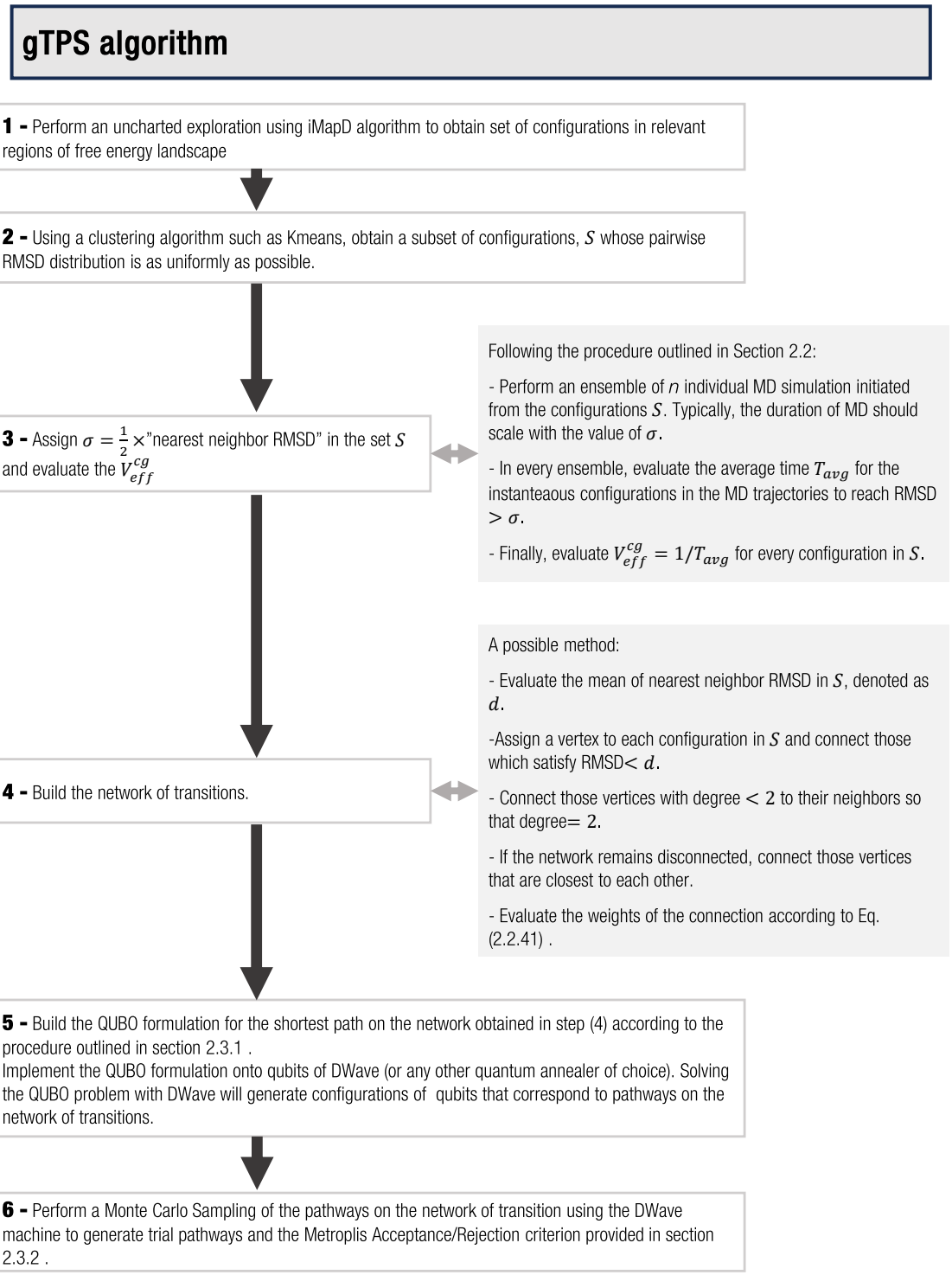
3. After updating all the new tentative distances, we move from v_a to its neighbor that has the lowest tentative distance, e.g. v_c, and has the status of "unvisited". Then, we label v_a as previously visited and update label of v_c to currently visiting. Then we go back to the step II.

4. The algorithm is terminated once the currently visiting vertex is the target vertex v_t.

A.5 Diagram of iMapD and gTPS algorithms

We have diagrammatically illustrated the algorithms of iMapD and gTPS below.


iMapD algorithm
Run an initial round of MDs using configuration in previously identified metastable states, and denote the set of all configurations as C
1 - Calculate the pairwise RMSD between members of C
2 - Perform DMAP using the procedure outlined in Appendix A.2 and obtain $\{\psi_1, \psi_2, \dots, \psi_p\}$ components.
The $\psi_k(i)$is the projection of config. Q_i onto the DMAP component ψ_k
3 - Identify the points/config.s who form a convex hull in the DMAP space.
4 - Revert to the ambient space of the system. Using RMSD, form a neighboring set for each config. Identified in step (3).
5 - Apply PCA onto neighboring set, according to Appendix A.3, and retain the components that saturate n% of the total variance. Generate a new set of coordinates using these principal components and the Eq. (2.1.3) . Perform a small energy minimization from the obtained Q^{new} coordinates to generate a chemically viable config. X^{new}.
6 - From each X^{new} initiate k new rounds of short burst of MD where the initial velocity in each iteration is randomly generated from a Maxwell-Boltzmann distribution.
7 - Append all the newly generated config. s to the set C
Has the algorithm found configurations in a new metastable state?
YES
NO
TERMINATE
Return to (1)


gTPS algorithm
1 - Perform an uncharted exploration using iMapD algorithm to obtain set of configurations in relevant regions of free energy landscape
2 - Using a clustering algorithm such as Kmeans, obtain a subset of configurations, S whose pairwise RMSD distribution is as uniformly as possible.
3 - Assign $\sigma = \frac{1}{2} \times$"nearest neighbor RMSD" in the set S and evaluate the V_{eff}^{cg}
Following the procedure outlined in Section 2.2:
- Perform an ensemble of n individual MD simulation initiated from the configurations S. Typically, the duration of MD should scale with the value of σ.
- In every ensemble, evaluate the average time T_{avg} for the instanteaous configurations in the MD trajectories to reach RMSD $> \sigma$.
- Finally, evaluate $V_{eff}^{cg} = 1/T_{avg}$ for every configuration in S.
4 - Build the network of transitions.
A possible method:
- Evaluate the mean of nearest neighbor RMSD in S, denoted as d.
-Assign a vertex to each configuration in S and connect those which satisfy RMSD$< d$.
- Connect those vertices with degree < 2 to their neighbors so that degree$= 2$.
- If the network remains disconnected, connect those vertices that are closest to each other.
- Evaluate the weights of the connection according to Eq. (2.2.41) .
5 - Build the QUBO formulation for the shortest path on the network obtained in step (4) according to the procedure outlined in section 2.3.1 .
Implement the QUBO formulation onto qubits of DWave (or any other quantum annealer of choice). Solving the QUBO problem with DWave will generate configurations of qubits that correspond to pathways on the network of transitions.
6 - Perform a Monte Carlo Sampling of the pathways on the network of transition using the DWave machine to generate trial pathways and the Metroplis Acceptance/Rejection criterion provided in section 2.3.2 .

Bibliography

[1] N. Malik and U. Ozturk, "Rare events in complex systems: Understanding and prediction," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 30, no. 9, 2020. DOI: 10.1063/5.0024145. [Online]. Available: https://doi.org/10.1063%2F5.0024145.

[2] J. Kwapień and S. Drożdż, "Physical approach to complex systems," *Physics Reports*, vol. 515, no. 3-4, pp. 115–226, 2012. DOI: 10.1016/j.physrep.2012.01.007. [Online]. Available: https://doi.org/10.1016%2Fj.physrep.2012.01.007.

[3] B. Peters, "Introduction," in *Reaction Rate Theory and Rare Events Simulations*, Elsevier, 2017, pp. 1–17.

[4] H. Kuwahara and I. Mura, "An efficient and exact stochastic simulation method to analyze rare events in biochemical systems," *The Journal of Chemical Physics*, vol. 129, no. 16, 2008. DOI: 10.1063/1.2987701. [Online]. Available: https://doi.org/10.1063%2F1.2987701.

[5] M. Cammarata *et al.*, "Unveiling the timescale of the r–t transition in human hemoglobin," *Journal of Molecular Biology*, vol. 400, no. 5, pp. 951–962, 2010. DOI: 10.1016/j.jmb.2010.05.057. [Online]. Available: https://doi.org/10.1016%2Fj.jmb.2010.05.057.

[6] S. Mehra *et al.*, "α-synuclein misfolding and aggregation: Implications in parkinson's disease pathogenesis," *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, vol. 1867, no. 10, pp. 890–908, 2019. DOI: 10.1016/j.bbapap.2019.03.001. [Online]. Available: https://doi.org/10.1016%2Fj.bbapap.2019.03.001.

[7] A. Gupta *et al.*, "Experimental techniques to study protein dynamics and conformations," in *Advances in Protein Molecular and Structural Biology Methods*, Elsevier, 2022, pp. 181–197. DOI: 10.1016/b978-0-323-90264-9.00012-x. [Online]. Available: https://doi.org/10.1016%2Fb978-0-323-90264-9.00012-x.

[8] R. O. Dror *et al.*, "Biomolecular simulation: A computational microscope for molecular biology," *Annual Review of Biophysics*, vol. 41, no. 1, pp. 429–452, 2012. DOI:

10.1146/annurev-biophys-042910-155245. [Online]. Available: https://doi.org/10.1146%2Fannurev-biophys-042910-155245.

[9] B. J. Alder and T. E. Wainwright, "Studies in molecular dynamics. i. general method," *The Journal of Chemical Physics*, vol. 31, no. 2, pp. 459–466, 1959. DOI: 10.1063/1.1730376. [Online]. Available: https://doi.org/10.1063%2F1.1730376.

[10] J. B. Gibson *et al.*, "Dynamics of radiation damage," *Physical Review*, vol. 120, no. 4, pp. 1229–1253, 1960. DOI: 10.1103/physrev.120.1229. [Online]. Available: https://doi.org/10.1103%2Fphysrev.120.1229.

[11] A. Warshel and M. Levitt, "Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme," *Journal of molecular biology*, vol. 103, no. 2, pp. 227–249, 1976.

[12] J. A. McCammon *et al.*, "Dynamics of folded proteins," *nature*, vol. 267, no. 5612, pp. 585–590, 1977.

[13] O. M. Salo-Ahen *et al.*, "Molecular dynamics simulations in drug discovery and pharmaceutical development," *Processes*, vol. 9, no. 1, p. 71, 2020.

[14] S. Stephan *et al.*, "Molmod–an open access database of force fields for molecular simulations of fluids," *Molecular Simulation*, vol. 45, no. 10, pp. 806–814, 2019.

[15] G. A. Khoury *et al.*, "Forcefield_ptm: Ab initio charge and amber forcefield parameters for frequently occurring post-translational modifications," *Journal of chemical theory and computation*, vol. 9, no. 12, pp. 5653–5674, 2013.

[16] E. Lerner *et al.*, "Fret-based dynamic structural biology: Challenges, perspectives and an appeal for open-science practices," *eLife*, vol. 10, Mar. 2021. DOI: 10.7554/eLife.60416.

[17] D. Jones *et al.*, "Accelerators for classical molecular dynamics simulations of biomolecules," *Journal of Chemical Theory and Computation*, vol. 18, no. 7, pp. 4047–4069, 2022. DOI: 10.1021/acs.jctc.1c01214. [Online]. Available: https://doi.org/10.1021%2Facs.jctc.1c01214.

[18] M. Shirts and V. S. Pande, "Screen savers of the world unite!" *Science*, vol. 290, no. 5498, pp. 1903–1904, 2000. DOI: 10.1126/science.290.5498.1903. [Online]. Available: https://doi.org/10.1126%2Fscience.290.5498.1903.

[19] V. A. Voelz *et al.*, "Folding@home: Achievements from over 20 years of citizen science herald the exascale era," *Biophysical Journal*, vol. 122, no. 14, pp. 2852–2863, 2023. DOI: 10.1016/j.bpj.2023.03.028. [Online]. Available: https://doi.org/10.1016%2Fj.bpj.2023.03.028.

[20] N. W. Kelley *et al.*, "The predicted structure of the headpiece of the huntingtin protein and its implications on huntingtin aggregation," *Journal of Molecular Biology*, vol. 388, no. 5, pp. 919–927, 2009. DOI: 10.1016/j.jmb.2009.01.032. [Online]. Available: https://doi.org/10.1016%2Fj.jmb.2009.01.032.

[21] Y.-S. Lin *et al.*, "Investigating how peptide length and a pathogenic mutation modify the structural ensemble of amyloid beta monomer," *Biophysical Journal*, vol. 102, no. 2, pp. 315–324, 2012. DOI: 10.1016/j.bpj.2011.12.002. [Online]. Available: https://doi.org/10.1016%2Fj.bpj.2011.12.002.

[22] D. E. Shaw *et al.*, "Anton, a special-purpose machine for molecular dynamics simulation," *Communications of the ACM*, vol. 51, no. 7, pp. 91–97, 2008. DOI: 10.1145/1364782.1364802. [Online]. Available: https://doi.org/10.1145%2F1364782.1364802.

[23] V. A. Voelz *et al.*, "Molecular simulation of *ab Initio* protein folding for a millisecond folder NTL9(1-39)," *Journal of the American Chemical Society*, vol. 132, no. 5, pp. 1526–1528, 2010. DOI: 10.1021/ja9090353. [Online]. Available: https://doi.org/10.1021%2Fja9090353.

[24] D. E. Shaw *et al.*, "Millisecond-scale molecular dynamics simulations on anton," in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, ACM, 2009. DOI: 10.1145/1654059.1654126. [Online]. Available: https://doi.org/10.1145%2F1654059.1654126.

[25] D. E. Shaw *et al.*, "Atomic-level characterization of the structural dynamics of proteins," *Science*, vol. 330, no. 6002, pp. 341–346, 2010. DOI: 10.1126/science.1187409. [Online]. Available: https://doi.org/10.1126%2Fscience.1187409.

[26] K. Lindorff-Larsen *et al.*, "How fast-folding proteins fold," *Science*, vol. 334, no. 6055, pp. 517–520, 2011. DOI: 10.1126/science.1208351. [Online]. Available: https://doi.org/10.1126%2Fscience.1208351.

[27] D. E. Shaw *et al.*, "Anton 3," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ACM, 2021. DOI: 10.1145/3458817.3487397. [Online]. Available: https://doi.org/10.1145%2F3458817.3487397.

[28] A. N. Naganathan and V. Muñoz, "Scaling of folding times with protein size," *Journal of the American Chemical Society*, vol. 127, no. 2, pp. 480–481, 2004. DOI: 10.1021/ja044449u. [Online]. Available: https://doi.org/10.1021%2Fja044449u.

[29] J. Kästner, "Umbrella sampling," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 1, no. 6, pp. 932–942, 2011. DOI: 10.1002/wcms.66. [Online]. Available: https://doi.org/10.1002%2Fwcms.66.

[30] S. Park and K. Schulten, "Calculating potentials of mean force from steered molecular dynamics simulations," *The Journal of Chemical Physics*, vol. 120, no. 13, pp. 5946–5961, 2004. DOI: 10.1063/1.1651473. [Online]. Available: https://doi.org/10.1063%2F1.1651473.

[31] P. G. Bolhuis and D. W. H. Swenson, "Transition path sampling as markov chain monte carlo of trajectories: Recent algorithms, software, applications, and future outlook," *Advanced Theory and Simulations*, vol. 4, no. 4, 2021. DOI: 10.1002/adts.202000237. [Online]. Available: https://doi.org/10.1002%2Fadts.202000237.

[32] D. Granata *et al.*, "Characterization of the free-energy landscapes of proteins by NMR-guided metadynamics," *Proceedings of the National Academy of Sciences*, vol. 110, no. 17, pp. 6817–6822, 2013. DOI: 10.1073/pnas.1218350110. [Online]. Available: https://doi.org/10.1073%2Fpnas.1218350110.

[33] M. F. Hagan *et al.*, "Atomistic understanding of kinetic pathways for single base-pair binding and unbinding in DNA," *Proceedings of the National Academy of Sciences*, vol. 100, no. 24, pp. 13 922–13 927, 2003. DOI: 10.1073/pnas.2036378100. [Online]. Available: https://doi.org/10.1073%2Fpnas.2036378100.

[34] S. Y. Joshi and S. A. Deshmukh, "A review of advancements in coarse-grained molecular dynamics simulations," *Molecular Simulation*, vol. 47, no. 10-11, pp. 786–803, 2021.

[35] A. P. Latham and B. Zhang, "Consistent force field captures homologue-resolved hp1 phase separation," *Journal of Chemical Theory and Computation*, vol. 17, no. 5, 3134–3144, Apr. 2021, ISSN: 1549-9626. DOI: 10.1021/acs.jctc.0c01220. [Online]. Available: http://dx.doi.org/10.1021/acs.jctc.0c01220.

[36] A. P. Latham and B. Zhang, "Unifying coarse-grained force fields for folded and disordered proteins," *Current Opinion in Structural Biology*, vol. 72, 63–70, Feb. 2022, ISSN: 0959-440X. DOI: 10.1016/j.sbi.2021.08.006. [Online]. Available: http://dx.doi.org/10.1016/j.sbi.2021.08.006.

[37] A. Perrakis and T. K. Sixma, "AI revolutions in biology," *EMBO reports*, vol. 22, no. 11, 2021. DOI: 10.15252/embr.202154046. [Online]. Available: https://doi.org/10.15252%2Fembr.202154046.

[38] M. Ceriotti *et al.*, "Machine learning meets chemical physics," *The Journal of Chemical Physics*, vol. 154, no. 16, 2021. DOI: 10.1063/5.0051418. [Online]. Available: https://doi.org/10.1063%2F5.0051418.

[39] J. Behler, "Perspective: Machine learning potentials for atomistic simulations," *The Journal of Chemical Physics*, vol. 145, no. 17, 2016. DOI: 10.1063/1.4966192. [Online]. Available: https://doi.org/10.1063%2F1.4966192.

[40] F. Noé *et al.*, "Machine learning for molecular simulation," *Annual Review of Physical Chemistry*, vol. 71, no. 1, pp. 361–390, 2020. DOI: 10.1146/annurev-physchem-042018-052331. [Online]. Available: https://doi.org/10.1146%2Fannurev-physchem-042018-052331.

[41] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021. DOI: 10.1038/s41586-021-03819-2. [Online]. Available: https://doi.org/10.1038%2Fs41586-021-03819-2.

[42] O. T. Unke *et al.*, "Machine learning force fields," *Chemical Reviews*, vol. 121, no. 16, pp. 10 142–10 186, 2021. DOI: 10.1021/acs.chemrev.0c01111. [Online]. Available: https://doi.org/10.1021%2Facs.chemrev.0c01111.

[43] P. Gkeka *et al.*, "Machine learning force fields and coarse-grained variables in molecular dynamics: Application to materials and biological systems," *Journal of Chemical Theory and Computation*, vol. 16, no. 8, pp. 4757–4775, 2020. DOI: 10.1021/acs.jctc.0c00355. [Online]. Available: https://doi.org/10.1021%2Facs.jctc.0c00355.

[44] A. Mardt *et al.*, "VAMPnets for deep learning of molecular kinetics," *Nature Communications*, vol. 9, no. 1, 2018. DOI: 10.1038/s41467-017-02388-1. [Online]. Available: https://doi.org/10.1038%2Fs41467-017-02388-1.

[45] H. Sidky and J. K. Whitmer, "Learning free energy landscapes using artificial neural networks," *The Journal of Chemical Physics*, vol. 148, no. 10, 2018. DOI: 10.1063/1.5018708. [Online]. Available: https://doi.org/10.1063%2F1.5018708.

[46] Q. Bai *et al.*, "Application advances of deep learning methods for de novo drug design and molecular dynamics simulation," *WIREs Computational Molecular Science*, vol. 12, no. 3, 2021. DOI: 10.1002/wcms.1581. [Online]. Available: https://doi.org/10.1002%2Fwcms.1581.

[47] M. Pandey *et al.*, "The transformational role of GPU computing and deep learning in drug discovery," *Nature Machine Intelligence*, vol. 4, no. 3, pp. 211–221, 2022. DOI: 10.1038/s42256-022-00463-x. [Online]. Available: https://doi.org/10.1038%2Fs42256-022-00463-x.

[48] P. Scholl *et al.*, "Quantum simulation of 2d antiferromagnets with hundreds of rydberg atoms," *Nature*, vol. 595, no. 7866, pp. 233–238, 2021. DOI: 10.1038/s41586-021-03585-1. [Online]. Available: https://doi.org/10.1038%2Fs41586-021-03585-1.

[49] I. Pogorelov *et al.*, "Compact ion-trap quantum computing demonstrator," *PRX Quantum*, vol. 2, no. 2, 2021. DOI: 10.1103/prxquantum.2.020343. [Online]. Available: https://doi.org/10.1103/prxquantum.2.020343.

[50] P. Ball, "First quantum computer to pack 100 qubits enters crowded race," *Nature*, vol. 599, no. 7886, pp. 542–542, 2021. DOI: 10.1038/d41586-021-03476-5. [Online]. Available: https://doi.org/10.1038%2Fd41586-021-03476-5.

[51] F. Arute *et al.*, "Quantum supremacy using a programmable superconducting processor," *Nature*, vol. 574, no. 7779, pp. 505–510, 2019. DOI: 10.1038/s41586-019-1666-5. [Online]. Available: https://doi.org/10.1038/s41586-019-1666-5.

[52] H.-S. Zhong *et al.*, "Quantum computational advantage using photons," *Science*, vol. 370, no. 6523, pp. 1460–1463, 2020. DOI: 10.1126/science.abe8770. [Online]. Available: https://doi.org/10.1126/science.abe8770.

[53] C. Dellago *et al.*, "On the calculation of reaction rate constants in the transition path ensemble," *The Journal of Chemical Physics*, vol. 110, no. 14, pp. 6617–6625, 1999. DOI: 10.1063/1.478569. [Online]. Available: https://doi.org/10.1063%2F1.478569.

[54] A. Das and B. K. Chakrabarti, Eds., *Quantum annealing and related optimization methods* (Lecture Notes in Physics), 2005th ed. Berlin, Germany: Springer, 2005.

[55] A. Das and B. K. Chakrabarti, "Colloquium: Quantum annealing and analog quantum computation," *Reviews of Modern Physics*, vol. 80, no. 3, pp. 1061–1081, 2008. DOI: 10.1103/revmodphys.80.1061. [Online]. Available: https://doi.org/10.1103%2Frevmodphys.80.1061.

[56] T. Albash and D. A. Lidar, "Adiabatic quantum computation," *Reviews of Modern Physics*, vol. 90, no. 1, 2018. DOI: 10.1103/revmodphys.90.015002. [Online]. Available: https://doi.org/10.1103/revmodphys.90.015002.

[57] S. E. Venegas-Andraca *et al.*, "A cross-disciplinary introduction to quantum annealing-based algorithms," *Contemporary Physics*, vol. 59, no. 2, pp. 174–197, 2018. DOI: 10.1080/00107514.2018.1450720. [Online]. Available: https://doi.org/10.1080/00107514.2018.1450720.

[58] P. Hauke *et al.*, "Perspectives of quantum annealing: Methods and implementations," *Reports on Progress in Physics*, vol. 83, no. 5, p. 054 401, 2020. DOI: 10.1088/1361-6633/ab85b8. [Online]. Available: https://doi.org/10.1088%2F1361-6633%2Fab85b8.

[59] E. Chiavazzo *et al.*, "Intrinsic map dynamics exploration for uncharted effective free-energy landscapes," *Proceedings of the National Academy of Sciences*, vol. 114, no. 28, 2017. DOI: 10.1073/pnas.1621481114. [Online]. Available: https://doi.org/10.1073%2Fpnas.1621481114.

[60] M. W. Johnson *et al.*, "Quantum annealing with manufactured spins," *Nature*, vol. 473, no. 7346, pp. 194–198, 2011. DOI: 10.1038/nature10012. [Online]. Available: https://doi.org/10.1038%2Fnature10012.

[61] D. Willsch *et al.*, "Benchmarking advantage and d-wave 2000q quantum annealers with exact cover problems," *Quantum Information Processing*, vol. 21, no. 4, p. 141, 2022.

[62] G. Kochenberger *et al.*, "The unconstrained binary quadratic programming problem: A survey," *Journal of Combinatorial Optimization*, vol. 28, no. 1, pp. 58–81, 2014. DOI: 10.1007/s10878-014-9734-0. [Online]. Available: https://doi.org/10.1007%2Fs10878-014-9734-0.

[63] A. Lucas, "Ising formulations of many NP problems," *Frontiers in Physics*, vol. 2, 2014. DOI: 10.3389/fphy.2014.00005. [Online]. Available: https://doi.org/10.3389%2Ffphy.2014.00005.

[64] J. C. Phillips *et al.*, "Scalable molecular dynamics with NAMD," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1781–1802, 2005. DOI: 10.1002/jcc.20289. [Online]. Available: https://doi.org/10.1002%2Fjcc.20289.

[65] M. J. Abraham *et al.*, "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, vol. 1-2, pp. 19–25, 2015. DOI: 10.1016/j.softx.2015.06.001. [Online]. Available: https://doi.org/10.1016%2Fj.softx.2015.06.001.

[66] M. Abraham *et al.*, "Gromacs 2023.3 manual," 2023. DOI: 10.5281/ZENODO.10017699. [Online]. Available: https://zenodo.org/doi/10.5281/zenodo.10017699.

[67] B. Peters, "Reaction coordinates and mechanistic hypothesis tests," *Annual Review of Physical Chemistry*, vol. 67, no. 1, pp. 669–690, 2016. DOI: 10.1146/annurev-physchem-040215-112215. [Online]. Available: https://doi.org/10.1146%2Fannurev-physchem-040215-112215.

[68] M. A. Rohrdanz *et al.*, "Discovering mountain passes via torchlight: Methods for the definition of reaction coordinates and pathways in complex macromolecular reactions," *Annual Review of Physical Chemistry*, vol. 64, no. 1, pp. 295–316, 2013. DOI: 10.1146/annurev-physchem-040412-110006. [Online]. Available: https://doi.org/10.1146%2Fannurev-physchem-040412-110006.

[69] G. Torrie and J. Valleau, "Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling," *Journal of Computational Physics*, vol. 23, no. 2, pp. 187–199, 1977. DOI: 10.1016/0021-9991(77)90121-8. [Online]. Available: https://doi.org/10.1016%2F0021-9991%2877%2990121-8.

[70] S. Kumar *et al.*, "THE weighted histogram analysis method for free-energy calculations on biomolecules. i. the method," *Journal of Computational Chemistry*, vol. 13, no. 8, pp. 1011–1021, 1992. DOI: 10.1002/jcc.540130812. [Online]. Available: https://doi.org/10.1002%2Fjcc.540130812.

[71] A. L. Ferguson, "BayesWHAM: A bayesian approach for free energy estimation, reweighting, and uncertainty quantification in the weighted histogram analysis method," *Journal of Computational Chemistry*, vol. 38, no. 18, pp. 1583–1605, 2017. DOI: 10.1002/jcc.24800. [Online]. Available: https://doi.org/10.1002%2Fjcc.24800.

[72] J. A. Lemkul and D. R. Bevan, "Assessing the stability of alzheimer's amyloid protofibrils using molecular dynamics," *The Journal of Physical Chemistry B*, vol. 114, no. 4, pp. 1652–1660, 2010. DOI: 10.1021/jp9110794. [Online]. Available: https://doi.org/10.1021%2Fjp9110794.

[73] E. Giudice, "Base pair opening within b-DNA: Free energy pathways for GC and AT pairs from umbrella sampling simulations," *Nucleic Acids Research*, vol. 31, no. 5, pp. 1434–1443, 2003. DOI: 10.1093/nar/gkg239. [Online]. Available: https://doi.org/10.1093%2Fnar%2Fgkg239.

[74] M. S. Lee and M. A. Olson, "Calculation of absolute protein-ligand binding affinity using path and endpoint approaches," *Biophysical Journal*, vol. 90, no. 3, pp. 864–877, 2006. DOI: 10.1529/biophysj.105.071589. [Online]. Available: https://doi.org/10.1529%2Fbiophysj.105.071589.

[75] D. Sengupta and S. J. Marrink, "Lipid-mediated interactions tune the association of glycophorin a helix and its disruptive mutants in membranes," *Physical Chemistry Chemical Physics*, vol. 12, no. 40, p. 12 987, 2010. DOI: 10.1039/c0cp00101e. [Online]. Available: https://doi.org/10.1039%2Fc0cp00101e.

[76] H. Lu and K. Schulten, "Steered molecular dynamics simulations of force-induced protein domain unfolding," *Proteins: Structure, Function, and Genetics*, vol. 35, no. 4, pp. 453–463, 1999. DOI: 10.1002/(sici)1097-0134(19990601)35:4<453::aid-prot9>3.0.co;2-m. [Online]. Available: https://doi.org/10.1002%2F%28sici%291097-0134%2819990601%2935%3A4%3C453%3A%3Aaid-prot9%3E3.0.co%3B2-m.

[77] C. Jarzynski, "Nonequilibrium equality for free energy differences," *Physical Review Letters*, vol. 78, no. 14, pp. 2690–2693, 1997. DOI: 10.1103/physrevlett.78.2690. [Online]. Available: https://doi.org/10.1103%2Fphysrevlett.78.2690.

[78] G. E. Crooks, "Path-ensemble averages in systems driven far from equilibrium," *Physical Review E*, vol. 61, no. 3, pp. 2361–2366, 2000. DOI: 10.1103/physreve.61.2361. [Online]. Available: https://doi.org/10.1103%2Fphysreve.61.2361.

[79] J. R. Gullingsrud *et al.*, "Reconstructing potentials of mean force through time series analysis of steered molecular dynamics simulations," *Journal of Computational Physics*, vol. 151, no. 1, pp. 190–211, 1999. DOI: 10.1006/jcph.1999.6218. [Online]. Available: https://doi.org/10.1006%2Fjcph.1999.6218.

[80] P.-C. Do *et al.*, "Steered molecular dynamics simulation in rational drug design," *Journal of Chemical Information and Modeling*, vol. 58, no. 8, pp. 1473–1482, 2018. DOI: 10.1021/acs.jcim.8b00261. [Online]. Available: https://doi.org/10.1021%2Facs.jcim.8b00261.

[81] J. S. Patel *et al.*, "Steered molecular dynamics simulations for studying protein–ligand interaction in cyclin-dependent kinase 5," *Journal of Chemical Information and Modeling*, vol. 54, no. 2, pp. 470–480, 2014. DOI: 10.1021/ci4003574. [Online]. Available: https://doi.org/10.1021%2Fci4003574.

[82] H. Jin *et al.*, "Exploring the different ligand escape pathways in acylaminoacyl peptidase by random acceleration and steered molecular dynamics simulations," *RSC Advances*, vol. 6, no. 13, pp. 10 987–10 996, 2016. DOI: 10.1039/c5ra24952j. [Online]. Available: https://doi.org/10.1039%2Fc5ra24952j.

[83] A. Laio and M. Parrinello, "Escaping free-energy minima," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12 562–12 566, 2002. DOI: 10.1073/pnas.202427399. [Online]. Available: https://doi.org/10.1073%2Fpnas.202427399.

[84] G. Bussi and A. Laio, "Using metadynamics to explore complex free-energy landscapes," *Nature Reviews Physics*, vol. 2, no. 4, pp. 200–212, 2020. DOI: 10.1038/s42254-020-0153-0. [Online]. Available: https://doi.org/10.1038%2Fs42254-020-0153-0.

[85] D. Ray and M. Parrinello, "Kinetics from metadynamics: Principles, applications, and outlook," *Journal of Chemical Theory and Computation*, vol. 19, no. 17, pp. 5649–5670, 2023. DOI: 10.1021/acs.jctc.3c00660. [Online]. Available: https://doi.org/10.1021%2Facs.jctc.3c00660.

[86] D. Bonetti *et al.*, "Identification and structural characterization of an intermediate in the folding of the measles virus x domain," *Journal of Biological Chemistry*, vol. 291, no. 20, pp. 10 886–10 892, 2016. DOI: 10.1074/jbc.m116.721126. [Online]. Available: https://doi.org/10.1074%2Fjbc.m116.721126.

[87] J. Debnath and M. Parrinello, "Computing rates and understanding unbinding mechanisms in host–guest systems," *Journal of Chemical Theory and Computation*, vol. 18, no. 3, pp. 1314–1319, 2022. DOI: 10.1021/acs.jctc.1c01075. [Online]. Available: https://doi.org/10.1021%2Facs.jctc.1c01075.

[88] Q. Liao, "Enhanced sampling and free energy calculations for protein simulations," in *Computational Approaches for Understanding Dynamical Systems: Protein Folding and Assembly*, Elsevier, 2020, pp. 177–213. DOI: 10.1016/bs.pmbts.2020.01.006. [Online]. Available: https://doi.org/10.1016%2Fbs.pmbts.2020.01.006.

[89] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," *Chemical Physics Letters*, vol. 314, no. 1-2, pp. 141–151, 1999. DOI: 10.1016/s0009-2614(99)01123-9. [Online]. Available: https://doi.org/10.1016%2Fs0009-2614%2899%2901123-9.

[90] R. Laghaei *et al.*, "Effect of the disulfide bond on the monomeric structure of human amylin studied by combined hamiltonian and temperature replica exchange molecular dynamics simulations," *The Journal of Physical Chemistry B*, vol. 114, no. 20, pp. 7071–7077, 2010. DOI: 10.1021/jp100205w. [Online]. Available: https://doi.org/10.1021%2Fjp100205w.

[91] R. Laghaei *et al.*, "Structure and thermodynamics of amylin dimer studied by hamiltonian-temperature replica exchange molecular dynamics simulations," *The Journal of Physical Chemistry B*, vol. 115, no. 12, pp. 3146–3154, 2011. DOI: 10.1021/jp108870q. [Online]. Available: https://doi.org/10.1021%2Fjp108870q.

[92] C. T. Leahy *et al.*, "Peptide dimerization-dissociation rates from replica exchange molecular dynamics," *The Journal of Chemical Physics*, vol. 147, no. 15, 2017. DOI: 10.1063/1.5004774. [Online]. Available: https://doi.org/10.1063%2F1.5004774.

[93] F. Ding *et al.*, "Ab initio folding of proteins with all-atom discrete molecular dynamics," *Structure*, vol. 16, no. 7, pp. 1010–1018, 2008. DOI: 10.1016/j.str.2008.03.013. [Online]. Available: https://doi.org/10.1016%2Fj.str.2008.03.013.

[94] D. Paschek *et al.*, "Computing the stability diagram of the trp-cage miniprotein," *Proceedings of the National Academy of Sciences*, vol. 105, no. 46, pp. 17 754–17 759, 2008. DOI: 10.1073/pnas.0804775105. [Online]. Available: https://doi.org/10.1073%2Fpnas.0804775105.

[95] E. K. Peter *et al.*, "How water layers on graphene affect folding and adsorption of TrpZip2," *The Journal of Chemical Physics*, vol. 141, no. 22, 2014. DOI: 10.1063/1.4896984. [Online]. Available: https://doi.org/10.1063%2F1.4896984.

[96] D. Hamelberg *et al.*, "Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules," *The Journal of Chemical Physics*, vol. 120, no. 24, pp. 11 919–11 929, 2004. DOI: 10.1063/1.1755656. [Online]. Available: https://doi.org/10.1063%2F1.1755656.

[97] L. Yang *et al.*, "Thermodynamics and folding pathways of trpzip2: An accelerated molecular dynamics simulation study," *The Journal of Physical Chemistry B*, vol. 113, no. 3, pp. 803–808, 2008. DOI: 10.1021/jp803160f. [Online]. Available: https://doi.org/10.1021%2Fjp803160f.

[98] Y. Miao *et al.*, "Accelerated molecular dynamics simulations of protein folding," *Journal of Computational Chemistry*, vol. 36, no. 20, pp. 1536–1549, 2015. DOI: 10.1002/jcc.23964. [Online]. Available: https://doi.org/10.1002%2Fjcc.23964.

[99] Y. Miao *et al.*, "Ligand binding pathways and conformational transitions of the HIV protease," *Biochemistry*, vol. 57, no. 9, pp. 1533–1541, 2018. DOI: 10.1021/acs.biochem.7b01248. [Online]. Available: https://doi.org/10.1021%2Facs.biochem.7b01248.

[100] J. Wang *et al.*, "Improved modeling of peptide-protein binding through global docking and accelerated molecular dynamics simulations," *Frontiers in Molecular Biosciences*, vol. 6, 2019. DOI: 10.3389/fmolb.2019.00112. [Online]. Available: https://doi.org/10.3389%2Ffmolb.2019.00112.

[101] Y. Miao *et al.*, "Gaussian accelerated molecular dynamics: Unconstrained enhanced sampling and free energy calculation," *Journal of Chemical Theory and Computation*, vol. 11, no. 8, pp. 3584–3595, 2015. DOI: 10.1021/acs.jctc.5b00436. [Online]. Available: https://doi.org/10.1021%2Facs.jctc.5b00436.

[102] G. Palermo, "Structure and dynamics of the CRISPR–cas9 catalytic complex," *Journal of Chemical Information and Modeling*, vol. 59, no. 5, pp. 2394–2406, 2019. DOI: 10.1021/acs.jcim.8b00988. [Online]. Available: https://doi.org/10.1021%2Facs.jcim.8b00988.

[103] V. S. Pande *et al.*, "Everything you wanted to know about markov state models but were afraid to ask," *Methods*, vol. 52, no. 1, pp. 99–105, 2010. DOI: 10.1016/j.ymeth.2010.06.002. [Online]. Available: https://doi.org/10.1016%2Fj.ymeth.2010.06.002.

[104] L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," *Physical Review Letters*, vol. 72, no. 23, pp. 3634–3637, 1994. DOI: 10.1103/physrevlett.72.3634. [Online]. Available: https://doi.org/10.1103%2Fphysrevlett.72.3634.

[105] B. E. Husic and V. S. Pande, "Markov state models: From an art to a science," *Journal of the American Chemical Society*, vol. 140, no. 7, pp. 2386–2396, 2018. DOI: 10.1021/jacs.7b12191. [Online]. Available: https://doi.org/10.1021%2Fjacs.7b12191.

[106] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20 150 202, 2016. DOI: 10.1098/rsta.2015.0202. [Online]. Available: https://doi.org/10.1098%2Frsta.2015.0202.

[107] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, CA, USA, vol. 1, 1967, pp. 281–297.

[108] B. A. Luty and J. A. McCammon, "Simulation of bimolecular reactions," *Molecular Simulation*, vol. 10, no. 1, pp. 61–65, 1993. DOI: 10.1080/08927029308022498. [Online]. Available: https://doi.org/10.1080%2F08927029308022498.

[109] D. Meral *et al.*, "An efficient strategy to estimate thermodynamics and kinetics of g protein-coupled receptor activation using metadynamics and maximum caliber," *The Journal of Chemical Physics*, vol. 149, no. 22, 2018. DOI: 10.1063/1.5060960. [Online]. Available: https://doi.org/10.1063%2F1.5060960.

[110] N. Plattner *et al.*, "Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and markov modelling," *Nature Chemistry*, vol. 9, no. 10, pp. 1005–1011, 2017. DOI: 10.1038/nchem.2785. [Online]. Available: https://doi.org/10.1038%2Fnchem.2785.

[111] D. J. WALES, "Discrete path sampling," *Molecular Physics*, vol. 100, no. 20, pp. 3285–3305, 2002. DOI: 10.1080/00268970210162691. [Online]. Available: https://doi.org/10.1080%2F00268970210162691.

[112] J. A. Joseph *et al.*, "Exploring biomolecular energy landscapes," *Chemical Communications*, vol. 53, no. 52, pp. 6974–6988, 2017. DOI: 10.1039/c7cc02413d. [Online]. Available: https://doi.org/10.1039%2Fc7cc02413d.

[113] G. Bussi *et al.*, "Free-energy landscape for β hairpin folding from combined parallel tempering and metadynamics," *Journal of the American Chemical Society*, vol. 128, no. 41, pp. 13 435–13 441, 2006. DOI: 10.1021/ja062463w. [Online]. Available: https://doi.org/10.1021%2Fja062463w.

[114] S. Piana and A. Laio, "A bias-exchange approach to protein folding," *The Journal of Physical Chemistry B*, vol. 111, no. 17, pp. 4553–4559, 2007. DOI: 10.1021/jp067373l. [Online]. Available: https://doi.org/10.1021%2Fjp067373l.

[115] A. Gil-Ley and G. Bussi, "Enhanced conformational sampling using replica exchange with collective-variable tempering," *Journal of Chemical Theory and Computation*, vol. 11, no. 3, pp. 1077–1085, 2015. DOI: 10.1021/ct5009087. [Online]. Available: https://doi.org/10.1021%2Fct5009087.

[116] Y. Sugita *et al.*, "Multidimensional replica-exchange method for free-energy calculations," *The Journal of Chemical Physics*, vol. 113, no. 15, pp. 6042–6051, 2000. DOI: 10.1063/1.1308516. [Online]. Available: https://doi.org/10.1063%2F1.1308516.

[117] J. Domański *et al.*, "Convergence and sampling in determining free energy landscapes for membrane protein association," *The Journal of Physical Chemistry B*, vol. 121, no. 15, pp. 3364–3375, 2016. DOI: 10.1021/acs.jpcb.6b08445. [Online]. Available: https://doi.org/10.1021%2Facs.jpcb.6b08445.

[118] M. Moradi and E. Tajkhorshid, "Driven metadynamics: Reconstructing equilibrium free energies from driven adaptive-bias simulations," *The Journal of Physical Chemistry Letters*, vol. 4, no. 11, pp. 1882–1887, 2013. DOI: 10.1021/jz400816x. [Online]. Available: https://doi.org/10.1021%2Fjz400816x.

[119] Y. Zhang and G. A. Voth, "Combined metadynamics and umbrella sampling method for the calculation of ion permeation free energy profiles," *Journal of Chemical Theory and Computation*, vol. 7, no. 7, pp. 2277–2283, 2011. DOI: 10.1021/ct200100e. [Online]. Available: https://doi.org/10.1021%2Fct200100e.

[120] J. M. Johnston *et al.*, "Assessing the relative stability of dimer interfaces in g protein-coupled receptors," *PLoS Computational Biology*, vol. 8, no. 8, E. Tajkhorshid, Ed., e1002649, 2012. DOI: 10.1371/journal.pcbi.1002649. [Online]. Available: https://doi.org/10.1371%2Fjournal.pcbi.1002649.

[121] S. Awasthi *et al.*, "Sampling free energy surfaces as slices by combining umbrella sampling and metadynamics," *Journal of Computational Chemistry*, vol. 37, no. 16, pp. 1413–1424, 2016. DOI: 10.1002/jcc.24349. [Online]. Available: https://doi.org/10.1002%2Fjcc.24349.

[122] M. Karplus and J. N. Kushick, "Method for estimating the configurational entropy of macromolecules," *Macromolecules*, vol. 14, no. 2, pp. 325–332, 1981. DOI: 10.1021/ma50003a019. [Online]. Available: https://doi.org/10.1021%2Fma50003a019.

[123] T. Ichiye and M. Karplus, "Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations," *Proteins: Structure, Function, and Bioinformatics*, vol. 11, no. 3, pp. 205–217, 1991. DOI: 10.1002/prot.340110305. [Online]. Available: https://doi.org/10.1002%2Fprot.340110305.

[124] B. Schölkopf *et al.*, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998. DOI: 10.1162/089976698300017467. [Online]. Available: https://doi.org/10.1162%2F089976698300017467.

[125] J. B. Tenenbaum *et al.*, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000. DOI: 10.1126/science.290.5500.2319. [Online]. Available: https://doi.org/10.1126%2Fscience.290.5500.2319.

[126] A. Glielmo *et al.*, "Unsupervised learning methods for molecular simulation data," *Chemical Reviews*, vol. 121, no. 16, pp. 9722–9758, 2021. DOI: 10.1021/acs.chemrev.0c01195. [Online]. Available: https://doi.org/10.1021%2Facs.chemrev.0c01195.

[127] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE Journal*, vol. 37, no. 2, pp. 233–243, 1991. DOI: 10.1002/aic.690370209. [Online]. Available: https://doi.org/10.1002%2Faic.690370209.

[128] W. Chen and A. L. Ferguson, "Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration," *Journal of Computational Chemistry*, vol. 39, no. 25, pp. 2079–2102, 2018. DOI: 10.1002/jcc.25520. [Online]. Available: https://doi.org/10.1002%2Fjcc.25520.

[129] W. Chen *et al.*, "Collective variable discovery and enhanced sampling using autoencoders: Innovations in network architecture and error function design," *The Journal of Chemical Physics*, vol. 149, no. 7, 2018. DOI: 10.1063/1.5023804. [Online]. Available: https://doi.org/10.1063%2F1.5023804.

[130] C. Wehmeyer and F. Noé, "Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics," *The Journal of Chemical Physics*, vol. 148, no. 24, 2018. DOI: 10.1063/1.5011399. [Online]. Available: https://doi.org/10.1063%2F1.5011399.

[131] R. Winter *et al.*, *Auto-encoding molecular conformations*, 2021. arXiv: 2101.01618 [cs.LG].

[132] R. Lopez *et al.*, "Information constraints on auto-encoding variational bayes," in *Neural Information Processing Systems*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:44052631.

[133] C. X. Hernández *et al.*, "Variational encoding of complex dynamics," *Physical Review E*, vol. 97, no. 6, 2018. DOI: 10.1103/physreve.97.062412. [Online]. Available: https://doi.org/10.1103%2Fphysreve.97.062412.

[134] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani *et al.*, Eds., vol. 27, Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[135] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach and D. Blei, Eds., ser. Proceedings of Machine Learning Research, vol. 37, Lille, France: PMLR, 2015, pp. 1530–1538. [Online]. Available: https://proceedings.mlr.press/v37/rezende15.html.

[136] J. Zhang *et al.*, "Targeted adversarial learning optimized sampling," *The Journal of Physical Chemistry Letters*, vol. 10, no. 19, pp. 5791–5797, 2019. DOI: 10.1021/acs.jpclett.9b02173. [Online]. Available: https://doi.org/10.1021%2Facs.jpclett.9b02173.

[137] F. Noé *et al.*, "Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning," *Science*, vol. 365, no. 6457, 2019. DOI: 10.1126/science.aaw1147. [Online]. Available: https://doi.org/10.1126%2Fscience.aaw1147.

[138] R. Elber *et al.*, "Calculating iso-committor surfaces as optimal reaction coordinates with milestoning," *Entropy*, vol. 19, no. 5, p. 219, 2017. DOI: 10.3390/e19050219. [Online]. Available: https://doi.org/10.3390%2Fe19050219.

[139] Z. F. Brotzakis and P. G. Bolhuis, "A one-way shooting algorithm for transition path sampling of asymmetric barriers," *The Journal of Chemical Physics*, vol. 145, no. 16, 2016. DOI: 10.1063/1.4965882. [Online]. Available: https://doi.org/10.1063%2F1.4965882.

[140] T. S. van Erp *et al.*, "A novel path sampling method for the calculation of rate constants," *The Journal of Chemical Physics*, vol. 118, no. 17, pp. 7762–7774, 2003. DOI: 10.1063/1.1562614. [Online]. Available: https://doi.org/10.1063%2F1.1562614.

[141] P. L. Geissler *et al.*, "Autoionization in liquid water," *Science*, vol. 291, no. 5511, pp. 2121–2124, 2001. DOI: 10.1126/science.1056991. [Online]. Available: https://doi.org/10.1126%2Fscience.1056991.

[142] P. G. Bolhuis, "Transition-path sampling of β-hairpin folding," *Proceedings of the National Academy of Sciences*, vol. 100, no. 21, pp. 12 129–12 134, 2003. DOI: 10.1073/pnas.1534924100. [Online]. Available: https://doi.org/10.1073%2Fpnas.1534924100.

[143] X. Ding *et al.*, "Exendin-4, a glucagon-like protein-1 (GLP-1) receptor agonist, reverses hepatic steatosis inob/ob mice," *Hepatology*, vol. 43, no. 1, pp. 173–181, 2005. DOI: 10.1002/hep.21006. [Online]. Available: https://doi.org/10.1002%2Fhep.21006.

[144] J. Juraszek and P. G. Bolhuis, "Sampling the multiple folding mechanisms of trp-cage in explicit solvent," *Proceedings of the National Academy of Sciences*, vol. 103, no. 43, pp. 15 859–15 864, 2006. DOI: 10.1073/pnas.0606692103. [Online]. Available: https://doi.org/10.1073%2Fpnas.0606692103.

[145] R. J. Allen *et al.*, "Sampling rare switching events in biochemical networks," *Physical Review Letters*, vol. 94, no. 1, 2005. DOI: 10.1103/physrevlett.94.018104. [Online]. Available: https://doi.org/10.1103%2Fphysrevlett.94.018104.

[146] J. Juraszek and P. G. Bolhuis, "Rate constant and reaction coordinate of trp-cage folding in explicit water," *Biophysical Journal*, vol. 95, no. 9, pp. 4246–4257, 2008. DOI: 10.1529/biophysj.108.136267. [Online]. Available: https://doi.org/10.1529%2Fbiophysj.108.136267.

[147] K. B. Lipkowitz, Ed., *Reviews in Computational Chemistry* (Reviews in Computational Chemistry), en, 2nd ed. Nashville, TN: John Wiley & Sons, Sep. 2010.

[148] Z. F. Brotzakis and P. G. Bolhuis, "Unbiased atomistic insight into the mechanisms and solvent role for globular protein dimer dissociation," *The Journal of Physical Chemistry B*, vol. 123, no. 9, pp. 1883–1895, 2019. DOI: 10.1021/acs.jpcb.8b10005. [Online]. Available: https://doi.org/10.1021%2Facs.jpcb.8b10005.

[149] D. Mercadante *et al.*, "Bovine β-lactoglobulin is dimeric under imitative physiological conditions: Dissociation equilibrium and rate constants over the pH range of 2.5–7.5," *Biophysical Journal*, vol. 103, no. 2, pp. 303–312, 2012. DOI: 10.1016/j.bpj.2012.05.041. [Online]. Available: https://doi.org/10.1016%2Fj.bpj.2012.05.041.

[150] H. Jung *et al.*, "Transition path sampling of rare events by shooting from the top," *The Journal of Chemical Physics*, vol. 147, no. 15, 2017. DOI: 10.1063/1.4997378. [Online]. Available: https://doi.org/10.1063%2F1.4997378.

[151] E. Borrero and C. Dellago, "Avoiding traps in trajectory space: Metadynamics enhanced transition path sampling," *The European Physical Journal Special Topics*, vol. 225, no. 8-9, pp. 1609–1620, 2016. DOI: 10.1140/epjst/e2016-60106-y. [Online]. Available: https://doi.org/10.1140%2Fepjst%2Fe2016-60106-y.

[152] J. Rogal and P. G. Bolhuis, "Multiple state transition path sampling," *The Journal of Chemical Physics*, vol. 129, no. 22, 2008. DOI: 10.1063/1.3029696. [Online]. Available: https://doi.org/10.1063%2F1.3029696.

[153] T. S. van Erp, "Reaction rate calculation by parallel path swapping," *Physical Review Letters*, vol. 98, no. 26, 2007. DOI: 10.1103/physrevlett.98.268301. [Online]. Available: https://doi.org/10.1103%2Fphysrevlett.98.268301.

[154] D. W. H. Swenson and P. G. Bolhuis, "A replica exchange transition interface sampling method with multiple interface sets for investigating networks of rare events," *The Journal of Chemical Physics*, vol. 141, no. 4, 2014. DOI: 10.1063/1.4890037. [Online]. Available: https://doi.org/10.1063%2F1.4890037.

[155] H. Jung *et al.*, *Artificial intelligence assists discovery of reaction coordinates and mechanisms from molecular dynamics simulations*, 2019. arXiv: 1901.04595 [physics.chem-ph].

[156] H. Jung *et al.*, "Machine-guided path sampling to discover mechanisms of molecular self-organization," *Nature Computational Science*, vol. 3, no. 4, pp. 334–345, 2023. DOI: 10.1038/s43588-023-00428-z. [Online]. Available: https://doi.org/10.1038%2Fs43588-023-00428-z.

[157] A. Baiardi *et al.*, "Quantum computing for molecular biology," *ChemBioChem*, vol. 24, no. 13, 2023. DOI: 10.1002/cbic.202300120. [Online]. Available: https://doi.org/10.1002%2Fcbic.202300120.

[158] Y. Manin, "Computable and uncomputable," *Sovetskoye Radio, Moscow*, vol. 128, p. 28, 1980.

[159] R. P. Feynman, "Simulating physics with computers," *International Journal of Theoretical Physics*, vol. 21, no. 6-7, pp. 467–488, 1982. DOI: 10.1007/bf02650179. [Online]. Available: https://doi.org/10.1007%2Fbf02650179.

[160] S. Lloyd, "Universal quantum simulators," *Science*, vol. 273, no. 5278, pp. 1073–1078, 1996. DOI: 10.1126/science.273.5278.1073. [Online]. Available: https://doi.org/10.1126%2Fscience.273.5278.1073.

[161] Y. Cao *et al.*, "Quantum chemistry in the age of quantum computing," *Chemical Reviews*, vol. 119, no. 19, pp. 10 856–10 915, 2019. DOI: 10.1021/acs.chemrev.8b00803. [Online]. Available: https://doi.org/10.1021%2Facs.chemrev.8b00803.

[162] A. Kitaev *et al.*, *Classical and Quantum Computation*. American Mathematical Society, 2002. DOI: 10.1090/gsm/047. [Online]. Available: https://doi.org/10.1090%2Fgsm%2F047.

[163] J. Preskill, "Quantum computing in the NISQ era and beyond," *Quantum*, vol. 2, p. 79, 2018. DOI: 10.22331/q-2018-08-06-79. [Online]. Available: https://doi.org/10.22331%2Fq-2018-08-06-79.

[164] M. Suchara *et al.*, "QuRE: The quantum resource estimator toolbox," in *2013 IEEE 31st International Conference on Computer Design (ICCD)*, IEEE, 2013. DOI: 10.1109/iccd.2013.6657074. [Online]. Available: https://doi.org/10.1109%2Ficcd.2013.6657074.

[165] N. Moll *et al.*, "Quantum optimization using variational algorithms on near-term quantum devices," *Quantum Science and Technology*, vol. 3, no. 3, p. 030 503, 2018. DOI: 10.1088/2058-9565/aab822. [Online]. Available: https://doi.org/10.1088%2F2058-9565%2Faab822.

[166] C. H. Baldwin *et al.*, "Re-examining the quantum volume test: Ideal distributions, compiler optimizations, confidence intervals, and scalable resource estimations," *Quantum*, vol. 6, p. 707, May 2022, ISSN: 2521-327X. DOI: 10.22331/q-2022-05-09-707. [Online]. Available: https://doi.org/10.22331/q-2022-05-09-707.

[167] K. Miller *et al.*, *An improved volumetric metric for quantum computers via more representative quantum circuit shapes*, 2022. arXiv: 2207.02315 [quant-ph].

[168] *Quantinuum website claiming quantum volume of* 2^{19} *for their h1-1 quantum computer.* https://web.archive.org/web/20230924011854/https://www.quantinuum.com/news/quantinuum-h-series-quantum-computer-accelerates-through-3-more-performance-records-for-quantum-volume-217-218-and-219, Accessed: 2023-09-24.

[169] S. A. Moses *et al.*, *A race track trapped-ion quantum processor*, 2023. arXiv: 2305.03828 [quant-ph].

[170] *Ibm website announcing the 433-qubit quantum computer, osprey.* https://web.archive.org/web/20231002100729/https://research.ibm.com/blog/next-wave-quantum-centric-supercomputing, Accessed: 2023-11-02.

[171] A. Peruzzo *et al.*, "A variational eigenvalue solver on a photonic quantum processor," *Nature Communications*, vol. 5, no. 1, 2014. DOI: 10.1038/ncomms5213. [Online]. Available: https://doi.org/10.1038%2Fncomms5213.

[172] E. Farhi *et al.*, *A quantum approximate optimization algorithm*, 2014. arXiv: 1411.4028 [quant-ph].

[173] C. Hempel *et al.*, "Quantum chemistry calculations on a trapped-ion quantum simulator," *Physical Review X*, vol. 8, no. 3, 2018. DOI: 10.1103/physrevx.8.031022. [Online]. Available: https://doi.org/10.1103/physrevx.8.031022.

[174] Y. Cao *et al.*, "Quantum chemistry in the age of quantum computing," *Chemical Reviews*, vol. 119, no. 19, pp. 10 856–10 915, 2019. DOI: 10.1021/acs.chemrev.8b00803. [Online]. Available: https://doi.org/10.1021/acs.chemrev.8b00803.

[175] S. N. Genin *et al.*, "Quantum chemistry on quantum annealers," 2019. arXiv: 1901.04715 [physics.chem-ph].

[176] C. Outeiral *et al.*, "The prospects of quantum computing in computational molecular biology," *WIREs Computational Molecular Science*, vol. 11, no. 1, May 2020. DOI: 10.1002/wcms.1481. [Online]. Available: https://doi.org/10.1002/wcms.1481.

[177] S. McArdle *et al.*, "Quantum computational chemistry," *Reviews of Modern Physics*, vol. 92, no. 1, 2020. DOI: 10.1103/revmodphys.92.015003. [Online]. Available: https://doi.org/10.1103/revmodphys.92.015003.

[178] C. Micheletti *et al.*, "Polymer physics by quantum computing," *Physical Review Letters*, vol. 127, no. 8, 2021. DOI: 10.1103/physrevlett.127.080501. [Online]. Available: https://doi.org/10.1103/physrevlett.127.080501.

[179] Y. Shen *et al.*, "Quantum implementation of the unitary coupled cluster for simulating molecular electronic structure," *Physical Review A*, vol. 95, no. 2, 2017. DOI: 10.1103/physreva.95.020501. [Online]. Available: https://doi.org/10.1103%2Fphysreva.95.020501.

[180] A. D. King *et al.*, "Quantum critical dynamics in a 5,000-qubit programmable spin glass," *Nature*, vol. 617, no. 7959, pp. 61–66, 2023. DOI: 10.1038/s41586-023-05867-2. [Online]. Available: https://doi.org/10.1038%2Fs41586-023-05867-2.

[181] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*. Cambridge University Press, 2012. DOI: 10.1017/cbo9780511976667. [Online]. Available: https://doi.org/10.1017%2Fcbo9780511976667.

[182] A. C.-C. Yao, "Quantum circuit complexity," in *Proceedings of 1993 IEEE 34th Annual Foundations of Computer Science*, IEEE. DOI: 10.1109/sfcs.1993.366852. [Online]. Available: https://doi.org/10.1109%2Fsfcs.1993.366852.

[183] A. Molina and J. Watrous, "Revisiting the simulation of quantum turing machines by quantum circuits," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 475, no. 2226, p. 20 180 767, 2019. DOI: 10.1098/rspa.2018.0767. [Online]. Available: https://doi.org/10.1098%2Frspa.2018.0767.

[184] N. Glover and T. Dudley, *Practical error correction design for engineers*, en. Data Systems Technology Corporation, 1991.

[185] D. Dieks, "Communication by EPR devices," *Physics Letters A*, vol. 92, no. 6, pp. 271–272, 1982. DOI: 10.1016/0375-9601(82)90084-6. [Online]. Available: https://doi.org/10.1016%2F0375-9601%2882%2990084-6.

[186] W. K. Wootters and W. H. Zurek, "A single quantum cannot be cloned," *Nature*, vol. 299, no. 5886, pp. 802–803, 1982. DOI: 10.1038/299802a0. [Online]. Available: https://doi.org/10.1038%2F299802a0.

[187] P. W. Shor, "Scheme for reducing decoherence in quantum computer memory," *Physical Review A*, vol. 52, no. 4, R2493–R2496, 1995. DOI: 10.1103/physreva.52.r2493. [Online]. Available: https://doi.org/10.1103%2Fphysreva.52.r2493.

[188] T. S. Cubitt *et al.*, "Universal quantum hamiltonians," *Proceedings of the National Academy of Sciences*, vol. 115, no. 38, pp. 9497–9502, 2018. DOI: 10.1073/pnas.1804949115. [Online]. Available: https://doi.org/10.1073%2Fpnas.1804949115.

[189] W. Vinci and D. A. Lidar, "Non-stoquastic hamiltonians in quantum annealing via geometric phases," *npj Quantum Information*, vol. 3, no. 1, 2017. DOI: 10.1038/s41534-017-0037-z. [Online]. Available: https://doi.org/10.1038%2Fs41534-017-0037-z.

[190] E. Farhi *et al.*, "A quantum adiabatic evolution algorithm applied to random instances of an NP-complete problem," *Science*, vol. 292, no. 5516, pp. 472–475, 2001. DOI: 10.1126/science.1057726. [Online]. Available: https://doi.org/10.1126%2Fscience.1057726.

[191] S. Kirkpatrick *et al.*, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983. DOI: 10.1126/science.220.4598.671. [Online]. Available: https://doi.org/10.1126%2Fscience.220.4598.671.

[192] M. H. S. Amin, "Consistency of the adiabatic theorem," *Physical Review Letters*, vol. 102, no. 22, 2009. DOI: 10.1103/physrevlett.102.220401. [Online]. Available: https://doi.org/10.1103%2Fphysrevlett.102.220401.

[193] M. Born and V. Fock, "Beweis des adiabatensatzes," *Zeitschrift für Physik*, vol. 51, no. 3-4, pp. 165–180, 1928. DOI: 10.1007/bf01343193. [Online]. Available: https://doi.org/10.1007%2Fbf01343193.

[194] A. Messiah, *Mechanics V1 P*. Nashville, TN: John Wiley & Sons, Jan. 1976.

[195] F. Glover *et al.*, "Quantum bridge analytics i: A tutorial on formulating and using QUBO models," *Annals of Operations Research*, vol. 314, no. 1, pp. 141–183, 2022. DOI: 10.1007/s10479-022-04634-2. [Online]. Available: https://doi.org/10.1007%2Fs10479-022-04634-2.

[196] N. Eidelson and B. Peters, "Transition path sampling for discrete master equations with absorbing states," *The Journal of Chemical Physics*, vol. 137, no. 9, 2012. DOI: 10.1063/1.4747338. [Online]. Available: https://doi.org/10.1063%2F1.4747338.

[197] P. Faccioli *et al.*, "Dominant pathways in protein folding," *Physical Review Letters*, vol. 97, no. 10, 2006. DOI: 10.1103/physrevlett.97.108101. [Online]. Available: https://doi.org/10.1103%2Fphysrevlett.97.108101.

[198] P. Hauke *et al.*, "Dominant reaction pathways by quantum computing," *Physical Review Letters*, vol. 126, no. 2, 2021. DOI: 10.1103/physrevlett.126.028104. [Online]. Available: https://doi.org/10.1103%2Fphysrevlett.126.028104.

[199] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006. DOI: 10.1016/j.acha.2006.04.006. [Online]. Available: https://doi.org/10.1016%2Fj.acha.2006.04.006.

[200] R. R. Coifman *et al.*, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," *Proceedings of the National Academy of Sciences*, vol. 102, no. 21, pp. 7426–7431, 2005. DOI: 10.1073/pnas.0500334102. [Online]. Available: https://doi.org/10.1073%2Fpnas.0500334102.

[201] F. P. Preparata and M. I. Shamos, *Computational Geometry*. Springer New York, 1985. DOI: 10.1007/978-1-4612-1098-6. [Online]. Available: https://doi.org/10.1007%2F978-1-4612-1098-6.

[202] E. Chiavazzo *et al.*, "Reduced models in chemical kinetics via nonlinear data-mining," *Processes*, vol. 2, no. 1, pp. 112–140, 2014. DOI: 10.3390/pr2010112. [Online]. Available: https://doi.org/10.3390%2Fpr2010112.

[203] S. Ballweg and R. Ernst, "Control of membrane fluidity: The OLE pathway in focus," *Biological Chemistry*, vol. 398, no. 2, pp. 215–228, 2016. DOI: 10.1515/hsz-2016-0277. [Online]. Available: https://doi.org/10.1515%2Fhsz-2016-0277.

[204] R. Covino *et al.*, "A eukaryotic sensor for membrane lipid saturation," *Molecular Cell*, vol. 63, no. 1, pp. 49–59, 2016. DOI: 10.1016/j.molcel.2016.05.015. [Online]. Available: https://doi.org/10.1016%2Fj.molcel.2016.05.015.

[205] M. Belkin *et al.*, *Graph laplacians on singular manifolds: Toward understanding complex spaces: Graph laplacians on manifolds with singularities and boundaries*, 2012. arXiv: 1211.6727 [cs.AI].

[206] J. Hermans, "The amino acid dipeptide: Small but still influential after 50 years," *Proceedings of the National Academy of Sciences*, vol. 108, no. 8, pp. 3095–3096, 2011. DOI: 10.1073/pnas.1019470108. [Online]. Available: https://doi.org/10.1073%2Fpnas.1019470108.

[207] P. Lepage, *How to renormalize the schrodinger equation*, 1997. arXiv: nucl-th/9706029 [nucl-th].

[208] E Pitard and H Orland, "Dynamics of the swelling or collapse of a homopolymer," *Europhysics Letters (EPL)*, vol. 41, no. 4, pp. 467–472, 1998. DOI: 10.1209/epl/i1998-00175-8. [Online]. Available: https://doi.org/10.1209%2Fepl%2Fi1998-00175-8.

[209] M. Sega *et al.*, "Quantitative protein dynamics from dominant folding pathways," *Physical Review Letters*, vol. 99, no. 11, 2007. DOI: 10.1103/physrevlett.99.118102. [Online]. Available: https://doi.org/10.1103%2Fphysrevlett.99.118102.

[210] M. Sega *et al.*, "Quantitative protein dynamics from dominant folding pathways," *Physical Review Letters*, vol. 99, no. 11, 2007. DOI: 10.1103/physrevlett.99.118102. [Online]. Available: https://doi.org/10.1103%2Fphysrevlett.99.118102.

[211] G. Mazzola *et al.*, "Fluctuations in the ensemble of reaction pathways," *The Journal of Chemical Physics*, vol. 134, no. 16, 2011. DOI: 10.1063/1.3581892. [Online]. Available: https://doi.org/10.1063%2F1.3581892.

[212] M. S. Könz *et al.*, "Uncertain fate of fair sampling in quantum annealing," *Physical Review A*, vol. 100, no. 3, 2019. DOI: 10.1103/physreva.100.030303. [Online]. Available: https://doi.org/10.1103/physreva.100.030303.

[213] M. Yamamoto *et al.*, "Fair sampling by simulated annealing on quantum annealer," *Journal of the Physical Society of Japan*, vol. 89, no. 2, p. 025 002, 2020. DOI: 10.7566/jpsj.89.025002. [Online]. Available: https://doi.org/10.7566/jpsj.89.025002.

[214] V. Kumar *et al.*, *Achieving fair sampling in quantum annealing*, 2020. arXiv: 2007.08487 [quant-ph].

[215] T. Krauss and J. McCollum, "Solving the network shortest path problem on a quantum annealer," *IEEE Transactions on Quantum Engineering*, vol. 1, pp. 1–12, 2020. DOI: 10.1109/tqe.2020.3021921. [Online]. Available: https://doi.org/10.1109%2Ftqe.2020.3021921.

[216] M. H. S. Amin *et al.*, "Role of single-qubit decoherence time in adiabatic quantum computation," *Physical Review A*, vol. 80, no. 2, 2009. DOI: 10.1103/physreva.80.022303. [Online]. Available: https://doi.org/10.1103/physreva.80.022303.

[217] M. H. Amin, "Searching for quantum speedup in quasistatic quantum annealers," *Physical Review A*, vol. 92, no. 5, 2015. DOI: 10.1103/physreva.92.052323. [Online]. Available: https://doi.org/10.1103/physreva.92.052323.

[218] M. Benedetti *et al.*, "Estimation of effective temperatures in quantum annealers for sampling applications: A case study with possible applications in deep learning," *Physical Review A*, vol. 94, no. 2, 2016. DOI: 10.1103/physreva.94.022308. [Online]. Available: https://doi.org/10.1103/physreva.94.022308.

[219] Z. Bian *et al.*, "The ising model: Teaching an old problem new tricks," *D-wave systems*, vol. 2, pp. 1–32, 2010.

[220] M. H. Amin, "Searching for quantum speedup in quasistatic quantum annealers," *Physical Review A*, vol. 92, no. 5, Nov. 2015, ISSN: 1094-1622. DOI: 10.1103/physreva.92.052323. [Online]. Available: http://dx.doi.org/10.1103/PhysRevA.92.052323.

[221] A. Perdomo-Ortiz *et al.*, "Determination and correction of persistent biases in quantum annealers," *Scientific Reports*, vol. 6, no. 1, Jan. 2016, ISSN: 2045-2322. DOI: 10.1038/srep18628. [Online]. Available: http://dx.doi.org/10.1038/srep18628.

[222] S. L. Quaytman and S. D. Schwartz, "Reaction coordinate of an enzymatic reaction revealed by transition path sampling," *Proceedings of the National Academy of Sciences*, vol. 104, no. 30, 12253–12258, Jul. 2007, ISSN: 1091-6490. DOI: 10.1073/pnas.0704304104. [Online]. Available: http://dx.doi.org/10.1073/pnas.0704304104.

[223] F. Wang *et al.*, "Folding mechanism of proteins im7 and im9: Insight from all-atom simulations in implicit and explicit solvent," *The Journal of Physical Chemistry B*, vol. 120, no. 35, 9297–9307, 2016, ISSN: 1520-5207. DOI: 10.1021/acs.jpcb.6b05819. [Online]. Available: http://dx.doi.org/10.1021/acs.jpcb.6b05819.

[224] F. Dingfelder *et al.*, "Slow escape from a helical misfolded state of the pore-forming toxin cytolysin a," *JACS Au*, vol. 1, no. 8, 1217–1230, 2021, ISSN: 2691-3704. DOI: 10.1021/jacsau.1c00175. [Online]. Available: http://dx.doi.org/10.1021/jacsau.1c00175.

[225] G. Spagnolli *et al.*, "Pharmacological inactivation of the prion protein by targeting a folding intermediate," *Communications Biology*, vol. 4, no. 1, 2021, ISSN: 2399-3642. DOI: 10.1038/s42003-020-01585-x. [Online]. Available: http://dx.doi.org/10.1038/s42003-020-01585-x.

[226] M. Tsytlonok and L. S. Itzhaki, "The how's and why's of protein folding intermediates," *Archives of Biochemistry and Biophysics*, vol. 531, no. 1–2, 14–23, 2013, ISSN: 0003-9861. DOI: 10.1016/j.abb.2012.10.006. [Online]. Available: http://dx.doi.org/10.1016/j.abb.2012.10.006.

[227] M. Feig, "Is alanine dipeptide a good model for representing the torsional preferences of protein backbones?" *Journal of Chemical Theory and Computation*, vol. 4, no. 9, pp. 1555–1564, 2008. DOI: 10.1021/ct800153n. [Online]. Available: https://doi.org/10.1021%2Fct800153n.

[228] T. Head-Gordon *et al.*, "A theoretical study of alanine dipeptide and analogs," *International Journal of Quantum Chemistry*, vol. 36, no. S16, pp. 311–322, 2009. DOI: 10.1002/qua.560360725. [Online]. Available: https://doi.org/10.1002%2Fqua.560360725.

[229] P. Eastman *et al.*, "OpenMM 7: Rapid development of high performance algorithms for molecular dynamics," *PLOS Computational Biology*, vol. 13, no. 7, R. Gentleman, Ed., e1005659, 2017. DOI: 10.1371/journal.pcbi.1005659. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1005659.

[230] N. Michaud-Agrawal *et al.*, "MDAnalysis: A toolkit for the analysis of molecular dynamics simulations," *Journal of Computational Chemistry*, vol. 32, no. 10, pp. 2319–2327, 2011. DOI: 10.1002/jcc.21787. [Online]. Available: https://doi.org/10.1002%2Fjcc.21787.

[231] R. Gowers *et al.*, "MDAnalysis: A python package for the rapid analysis of molecular dynamics simulations," in *Proceedings of the Python in Science Conference*, SciPy, 2016. DOI: 10.25080/majora-629e541a-00e. [Online]. Available: https://doi.org/10.25080%2Fmajora-629e541a-00e.

[232] M. Karplus and J. A. McCammon, "Molecular dynamics simulations of biomolecules," *Nature Structural Biology*, vol. 9, no. 9, pp. 646–652, 2002. DOI: 10.1038/nsb0902-646. [Online]. Available: https://doi.org/10.1038%2Fnsb0902-646.

[233] J. A. McCammon *et al.*, "Dynamics of folded proteins," *Nature*, vol. 267, no. 5612, pp. 585–590, 1977. DOI: 10.1038/267585a0. [Online]. Available: https://doi.org/10.1038%2F267585a0.

[234] E. Paci and M. Karplus, "Forced unfolding of fibronectin type 3 modules: An analysis by biased molecular dynamics simulations.," *Journal of Molecular Biology*, vol. 288, pp. 441–459, 3 May 1999, ISSN: 0022-2836. DOI: 10.1006/jmbi.1999.2670, ppublish.

[235] C. Camilloni *et al.*, "Hierarchy of folding and unfolding events of protein g, ci2, and acbp from explicit-solvent simulations.," *Journal of Chemical Physics*, vol. 134, p. 045 105, 4 Jan. 2011, ISSN: 1089-7690. DOI: 10.1063/1.3523345, ppublish.

[236] G. Bartolucci *et al.*, "Transition path theory from biased simulations.," *Journal of Chemical Physics*, vol. 149, p. 072 336, 7 Aug. 2018, ISSN: 1089-7690. DOI: 10.1063/1.5027253, ppublish.

[237] R. T. McGibbon *et al.*, "Mdtraj: A modern open library for the analysis of molecular dynamics trajectories," *Biophysical Journal*, vol. 109, no. 8, pp. 1528 –1532, 2015. DOI: 10.1016/j.bpj.2015.08.015.

[238] G. A. Tribello *et al.*, "Plumed 2: New feathers for an old bird.," *Computer Physics Communications*, vol. 185, pp. 604–613, 2014.

[239] F. Murtagh and P. Contreras, *Methods of hierarchical clustering*, 2011. arXiv: 1105.0121 [cs.IR].

[240] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[241] P. Virtanen *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.

[242] J. Kubelka *et al.*, "Sub-microsecond protein folding," *Journal of Molecular Biology*, vol. 359, no. 3, 546–553, Jun. 2006, ISSN: 0022-2836. DOI: 10.1016/j.jmb.2006.03.034. [Online]. Available: http://dx.doi.org/10.1016/j.jmb.2006.03.034.

[243] H. Lei *et al.*, "Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations," *Proceedings of the National Academy of Sciences*, vol. 104, no. 12, 4925–4930, Mar. 2007, ISSN: 1091-6490. DOI: 10.1073/pnas.0608432104. [Online]. Available: http://dx.doi.org/10.1073/pnas.0608432104.

[244] R. Harada and A. Kitao, "The fast-folding mechanism of villin headpiece subdomain studied by multiscale distributed computing," *Journal of Chemical Theory and Computation*, vol. 8, no. 1, 290–299, Dec. 2011, ISSN: 1549-9626. DOI: 10.1021/ct200363h. [Online]. Available: http://dx.doi.org/10.1021/ct200363h.

[245] E. Wang *et al.*, "A novel folding pathway of the villin headpiece subdomain hp35," *Physical Chemistry Chemical Physics*, vol. 21, no. 33, 18219–18226, 2019, ISSN: 1463-9084. DOI: 10.1039/c9cp01703h. [Online]. Available: http://dx.doi.org/10.1039/c9cp01703h.

[246] R. Mousa *et al.*, "Bpti folding revisited: Switching a disulfide into methylene thioacetal reveals a previously hidden path," *Chemical Science*, vol. 9, no. 21, 4814–4820, 2018, ISSN: 2041-6539. DOI: 10.1039/c8sc01110a. [Online]. Available: http://dx.doi.org/10.1039/c8sc01110a.

[247] T. E. Creighton, "Conformational restrictions on the pathway of folding and unfolding of the pancreatic trypsin inhibitor," *Journal of Molecular Biology*, vol. 113, no. 2, pp. 275–293, 1977. DOI: 10.1016/0022-2836(77)90142-5. [Online]. Available: https://doi.org/10.1016%2F0022-2836%2877%2990142-5.

[248] J. A. Mendoza *et al.*, "Effects of amino acid replacements on the reductive unfolding kinetics of pancreatic trypsin inhibitor," *Biochemistry*, vol. 33, no. 5, 1143–1148, Feb. 1994, ISSN: 1520-4995. DOI: 10.1021/bi00171a013. [Online]. Available: http://dx.doi.org/10.1021/bi00171a013.

[249] J. S. Weissman and P. S. Kim, "Reexamination of the folding of bpti: Predominance of native intermediates," *Science*, vol. 253, no. 5026, 1386–1393, Sep. 1991, ISSN: 1095-9203. DOI: 10.1126/science.1716783. [Online]. Available: http://dx.doi.org/10.1126/science.1716783.

[250] G. Mazzola, "Sampling, rates, and reaction currents through reverse stochastic quantization on quantum computers," *Physical Review A*, vol. 104, no. 2, 2021. DOI: 10.1103/physreva.104.022431. [Online]. Available: https://doi.org/10.1103/physreva.104.022431.

[251] A. Perdomo-Ortiz *et al.*, "Finding low-energy conformations of lattice protein models by quantum annealing," *Scientific Reports*, vol. 2, no. 1, 2012. DOI: 10.1038/srep00571. [Online]. Available: https://doi.org/10.1038/srep00571.

[252] L.-H. Lu and Y.-Q. Li, "Quantum approach to fast protein-folding time," *Chinese Physics Letters*, vol. 36, no. 8, p. 080 305, 2019. DOI: 10.1088/0256-307x/36/8/080305. [Online]. Available: https://doi.org/10.1088/0256-307x/36/8/080305.

[253] A. Irbäck *et al.*, "Folding lattice proteins with quantum annealing," *Phys. Rev. Res.*, vol. 4, p. 043 013, 4 2022. DOI: 10.1103/PhysRevResearch.4.043013. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevResearch.4.043013.

[254] A. Robert *et al.*, "Resource-efficient quantum algorithm for protein folding," *npj Quantum Inf*, vol. 7, p. 38, 2021. DOI: 10.1038/s41534-021-00368-4.

[255] S. Boulebnane *et al.*, "Resource-efficient quantum algorithm for protein folding," *npj Quantum Inf*, vol. 9, p. 043 013, 2023. DOI: /10.1038/s41534-023-00733-5.

[256] R. Babbush *et al.*, "Construction of energy functions for lattice heteropolymer models: Efficient encodings for constraint satisfaction programming and quantum annealing," *Advances in Chemical Physics*, vol. 155, 2014. DOI: 10.1002/9781118755815.ch05.

[257] J. Kubelka *et al.*, "The protein folding 'speed limit'," *Current Opinion in Structural Biology*, vol. 14, no. 1, pp. 76–88, 2004. DOI: 10.1016/j.sbi.2004.01.013. [Online]. Available: https://doi.org/10.1016%2Fj.sbi.2004.01.013.

[258] R. Olender and R. Elber, "Yet another look at the steepest descent path," *Journal of Molecular Structure: THEOCHEM*, vol. 398-399, pp. 63–71, 1997. DOI: 10.1016/s0166-1280(97)00038-9. [Online]. Available: https://doi.org/10.1016%2Fs0166-1280%2897%2900038-9.

[259] A. Ghosh *et al.*, "An atomically detailed study of the folding pathways of protein a with the stochastic difference equation," *Proceedings of the National Academy of Sciences*, vol. 99, no. 16, pp. 10 394–10 398, 2002. DOI: 10.1073/pnas.142288099. [Online]. Available: https://doi.org/10.1073%2Fpnas.142288099.

[260] I. T. Jolliffe, *Principal Component Analysis* (Springer Series in Statistics), en, 2nd ed. New York, NY: Springer, Oct. 2002.