Studying DNA opening during transcription by the RNA polymerase II with molecular dynamics simulations, a sampling challenge

Dissertation
zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultät
der Universität des Saarlandes

von
Jeremy Lapierre

Saarbrücken
2023

Tag des Kolloquiums:	5. Mai 2023
Dekan:	Prof. Dr. Ludger Santen
Berichterstatter:	Prof. Dr. Jochen Hub
	Prof. Dr. Bruce Morgan
Akad. Mitglied:	Dr. Thomas John
Vorsitz:	Prof. Dr. Gregor Jung

Acknowledgements

I feel thankful to had the opportunity to grow as a scientist under the supervision of Jochen S. Hub during my PhD, not only because he is objectively a great thinker to learn from, but also because he was always approachable and caring for discussing ideas or concerns. My PhD journey sometimes felt like an expedition through a foggy maze of knowledge, a situation where wise and caring advice are always welcome. Thank you Jochen for your great mentorship!

I would also like to thank all former and present members of the computational biophysics group for our thrilling discussions, everyone has always been friendly to me and contributed to built a great working environment I liked to evolve in. Specifically, I would like to thank Alejandro Martinez Leon for all the headaches he triggered in my head after intense and captivating discussions about common pitfalls we faced during our projects. Many thanks to Leonie Chatzimagas who kindly corrected my Abstract in German. I am grateful to Maciej Wójcik who drastically changed our standard of living from a software support perspective. Finally, I thank Massimiliano Anselmi, Robert Becker and Alejandro Martinez Leon for sharing and building a ready-to-use workflow for running simulated tempering simulations.

I genuinely thank all GROMACS and PLUMED developers for the cutting-edge and convenient softwares they provide, and the community of both ecosystems for helps found in respective Q&As or through direct conversations with members.

Computing resources and related technical supports are crucial to our work, therefore I would like to thank the GWDG HPC team in Göttingen, Dr. Christian Hoffmann and Maciej Brudzynski maintaining our cluster in Saarbrücken University, and the RHRK HPC team in Kaiserslautern.

Funding support from the Deutsche Forschungsgemeinschaft via SFB 860/A16 is gratefully acknowledged.

Last but not least, I probably would not have survived this entire PhD adventure without constant support from my family and friends. I am especially grateful to my wife who has always demonstrated a steadfast support to any tedious projects I chose to tackle, this one was no exception.

Zusammenfassung

Die RNA-Polymerase II (RNAP II) ist ein makromolekularer Komplex, der die RNA aus einer DNA-Matrize synthetisiert. Während des Initiationsschritts der Transkription, öffnet RNAP II die doppelsträngige DNA, um den DNA-Code freizulegen. Da die Bildung der DNA-Transkriptionsblase nur unzureichend verstanden ist, nutzten wir Molekulardynamik-Simulationen (MD), um Erkenntnisse über diesen Prozess zu erlangen.

Da die DNA-Öffnung auf Zeitskalen erfolgt, die für einfache MD Simulationen nicht zugänglich sind, prüften wir verschiedene Enhanced Sampling Methoden, um die MD Simulationen zu beschleunigen und den DNA-Öffnungsprozess zu untersuchen. Wir fanden heraus, dass die vielversprechendste Methode zur Untersuchung der DNA-Öffnung die Steuerung von Simulationen mit einer Kombination aus (i) geführter DNA-Rotation und (ii) Path Collective Variables war. Auf diese Weise erhielten wir kontinuierliche atomare Trajektorien des gesamten DNA-Öffnungsprozesses, welche qualitative Einblicke in die Rolle der Protein–DNA Wechselwirkungen im Allgemeinen ermöglichten.

Mit dem Ziel die DNA-Öffnung quantitativer zu beschreiben, möchten wir weitere Enhanced Sampling Techniken untersuchen, welche wir auf einen einfachen Prozess anwenden: die Permeation von Fosmidomycin durch das OprO Porin. Es zeigte sich, dass das Replica-Exchange Umbrella Sampling in der Lage ist, die Genauigkeit des Profils der freien Energie drastisch zu erhöhen, im Vergleich zu gewöhnlichem Umbrella Sampling.

Abstract

RNA polymerase II (RNAP II) is a macro-molecular complex that synthesizes RNA by reading the DNA code, a process called transcription. During the initiation step of transcription, RNAP II opens double-stranded DNA in order to read the DNA code. Since formation of the DNA transcription bubble remains poorly understood, we used molecular dynamics simulations (MD) to provide atomic-level insights into this process.

Because DNA opening occurs at time-scales that are not accessible to plain MD simulations, we have explored different enhanced sampling methods to accelerate MD simulations enabling to study the DNA opening process. Ultimately, by steering simulations with a combination of (i) guided DNA rotation and (ii) path collective variables, we obtained a continuous atomic trajectories of the complete DNA opening process. The simulations provided qualitative insights into the role of loop dynamics and protein-DNA interactions during DNA opening.

With the aim of obtaining a more quantitative description of DNA opening, we decided to further explore alternative enhanced sampling techniques applied on a simpler process, yet still challenging from a sampling perspective, that is drug permeation through the OprO porin. This study showed that replica-exchange umbrella sampling (REUS) is able to drastically increase precision of free energy profiles compared to standard umbrella sampling.

Contents

List of Figures

List of Tables

1

Introduction

Proteins play major roles in cell functions and structures, thus understanding their origin is of paramount importance to unravel the fundamental mechanism of life. Until 1944, when Avery *et al.* showed that the molecular storage of genetic information was made of DNA (1), proteins were thought to be both the genetic and the functional material of life. Later in 1958, Watson and Crick theorized the central dogma of molecular biology (2, 3) suggesting that proteins are synthesized from RNA through a process called translation, and that RNA is synthesized from DNA through a process called transcription. This proposed model for genetic information transfer is now established, even though some exceptions can break this model, *e.g.* viral RNAs. According to this scheme, DNA is the biomolecule storing the genetic information. From a natural selection standpoint (4, 5), DNA is more suitable than RNA for storing the genetic information because DNA is more stable than RNA (6), and has a higher fidelity during replication (7). The role of RNA is to deliver the genetic message from the nucleus —where DNA and genetic information source-code is protected— to the cytoplasm —where proteins are synthesized—; therefore RNA is involved in genetic information transfer. Moreover, RNA allows genetic expression modulations without impacting the integrity of the source-code held in the DNA, some examples of such modulations are RNA splicing or RNA chemical modifications.

Understanding processes involving DNA, RNA and protein is not only of fundamental interest to comprehend life, but is also of therapeutic significance. Indeed, misregulation of cell cycles in general and of transcription in particular can lead to uncontrolled cell growth and ultimately to cancers (8–10), a disease that has been estimated by the

International Agency for Research on Cancer to rise with 18.1 million new cases and to cause 9.9 million deaths worldwide in 2020 (11). In fact, an active compound targeting proteins involved in transcription, is produced by a plant named *Tripterygium wilfordii* —commonly called "thunder god vine"— and has been used for centuries in traditional Chinese medicine to prevent cell proliferation or inflammation (12, 13). Actinomycin D is another drug targeting the transcriptional machinery used in clinic to treat Ewing's sarcoma, Wilm's tumor and rhabdomyosarcoma (13, 14). Transcription is therefore a cornerstone of molecular biology and comprehending this mechanism paves the way for potential therapeutic applications, motivating the work of the present thesis.

1.1 Deoxyribonucleic acid: DNA

DNA is the biomolecule coding for all proteins of a cell, and therefore contributes to cell structure and function. DNA is a polymer of nucleotides (polynucleotide chain) linked by a phosphodiester bond; each nucleotide monomer is composed of a phosphate group, a deoxyribose and a nitrogenous nucleobase (15–17). The two ends of a nucleotide namely: the 5'-phosphate and 3'-hydroxyl ends, are denoted 5' and 3' ends; this notation is commonly used to refer to a specific DNA reading direction: 3' to 5' or 5' to 3'.

There are four different types of nucleobase in DNA: adenine (A), thymine (T), cytosine (C) and guanine (G) (18); the combination of three of these four nucleobases is called a codon. A succession of codons, framed by so-called start and stop codons, defines a genetic unit called gene. Each codon within a gene codes for a specific amino acid (2). Amino acids are the building block of proteins; hence, the sequence of codons in a gene defines the sequence of amino acids in a protein.

In its canonical 3D structure, DNA is composed of two polynucleotide chains which intertwine and interact with each other mainly by the mean of hydrogen bonds (H-bonds) and base stacking interactions (19, 20) (Fig. 1.1). H-bonds form between complementary base-pairs, *i.e.* between A and T, and between C and G. Stacking interactions result from hydrophobic property of the bases and of attractive London dispersion forces. The two DNA strands form a right-handed B-DNA helical structure; this canonical 3D DNA structure was discovered by James Watson and Francis Crick (21), based on x-ray crystallographic data from Rosalind Franklin and Maurice Wilkins (22), and on works from Erwin Chargaff *et al.* (23, 24) which were key for understanding base-pair

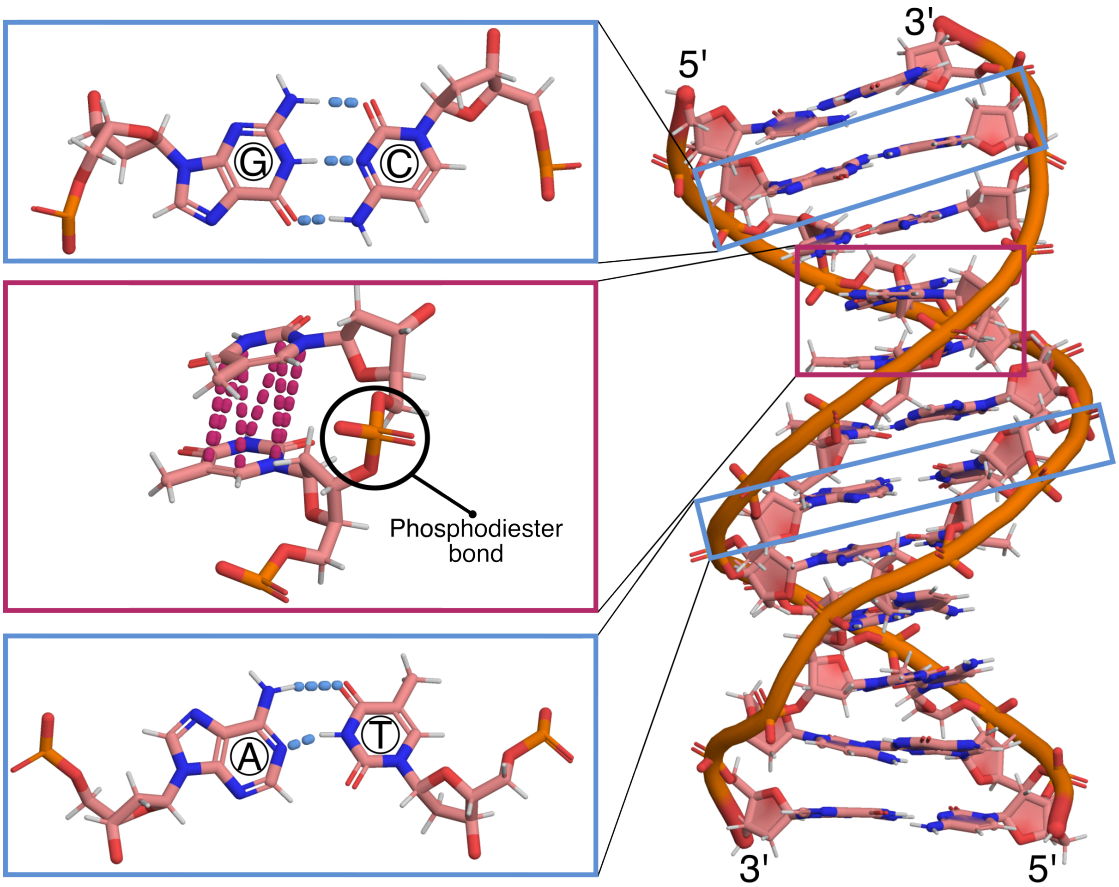

Figure 1.1: B-DNA 3D structure and involved intramolecular interactions - Watson-Crick hydrogen bonds between a G–C and an A–T pairs are shown in blue dashed lines in the blue rectangles. Stacking interactions between two thymine bases are shown in purple dashed lines in the purple rectangle. In this structure, 5'-phosphate ends are missing. The 3D structure of B-DNA shown here can be accessed with the following pdb code: 2BNA (25). This 3D model has been obtained at 16 K to eliminate thermal disorder and to obtain a ground-state for B-DNA.

complementarity principle. Double-stranded structure of DNA ensure minimal exposure of the nucleobases to solvent, and therefore participates to the stability of the genetic code.

1.2 Ribonucleic acid: RNA

RNA is, like DNA, a polymer of nucleotides linked by phosphodiester bonds. However, unlike DNA: (i) RNA nucleotides (called ribonucleotides) are composed of a ribose sugar,

(ii) uracil (U) nucleobase is found in RNA instead of thymine, and (iii) RNA is usually single-stranded, *i.e.* they are constituted of only one polynucleic chain (26). Exceptions exist in viruses for which the genome is constituted of double-stranded RNAs; for this reason, double-stranded RNAs are usually detected as a threat by living organisms and trigger defense mechanism in the host, *e.g.* RNA interference mechanism, in order to degrade double-stranded viral RNAs. In some specific biological processes, RNAs can also be found hybridized to DNA strand, where base-pair complementarity principle stated in the previous section applies, except that U pairs with A in RNA–DNA heteroduplexes. RNA–DNA heteroduplexes are found for example in transcription, where a DNA matrix sequence is used to synthesize a complementary RNA sequence (this process will be detailed in section 1.3). Even if RNA is usually single-stranded, a single RNA strand can fold onto itself an form a particular three-dimensional structure. RNA three-dimensional structures are a result of (i) Watson-Crick interactions within bases of the single-stranded RNA that form specific secondary structures, *e.g.* stem-loop and pseudoknots, and (ii) non-Watson-Crick interactions between secondary structures that form tertiary structures (27).

With respect to their functions, RNAs are classified in two categories: coding RNAs, also called messenger RNAs (mRNAs), and non-coding RNAs (ncRNAs). Messenger RNAs code for proteins while non-coding RNAs are involved in translation (transfer RNAs and ribosomal RNAs), in gene regulations (micro RNAs, small interfering RNAs and long noncoding RNAs) and RNA maturation (small nuclear RNA and small nucleolar RNAs).

1.3 Transcription by RNA polymerase II

The following reviews and articles have been used to write this section: (28–35).

The process of RNA synthesis from DNA is named transcription; such process is catalyzed by a macro-molecular complex named RNA polymerase (RNAP). Five RNAPs are found in eukaryotes: RNAP I through V. In this work we focused on the RNAP II which is involved in synthesis of messenger RNAs, some small nuclear RNAs, small nucleolar RNAs, small interfering RNAs, micro RNAs and long noncoding RNAs (36). RNAP II is constituted of twelve RNA polymerase subunits (RPBs): RPB1 through 12. RPB1 and RPB2 are the largest polymerase subunits and hold the catalytic activity

of the complex (Fig. 1.2). Other proteins are involved in transcription: transcription factors II (TFII) and the mediator complex. They associate to and dissociate from RNAP II at different stages of transcription. The three stages of transcription are: initiation, elongation and termination.

1.3.1 Transcription initiation

The initiation step is the first step of transcription and is itself subdivided into two steps: (i) pre-initiation complex (PIC) assembly —constituted of the twelve RNAP II subunits, TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH—, followed by (ii) DNA opening.

Specific DNA sequences, called promoters, trigger the PIC assembly by recruiting TFIID which binds specifically to this promotor sequence via its TATA-box-binding protein (TBP) subunit (38–44). The first promotor sequence being discovered is called TATA-box (45), hence the name for TBP, and has for consensus sequence: TATAWAWR (W means A or T, and R means A or G). Beside the TATA-box, other TATA-less promotor sequences are found in genes and represent $\sim 85\%$ of the promoters found in coding genes (46–50), nevertheless these promoters are also recognized by TBP. Upon TBP binding to the promotor, TBP induces a $\sim$90°-bending in the DNA (51–53). It has been shown that TFIID–DNA interactions are stabilized by the assembly of TFIIA, however TFIIA is not strictly necessary for transcription in vitro (54, 55). Following TFIID association to the PIC via its TBP subunit, TFIIB binds to DNA at the B recognition elements, and is also involved in early PIC assembly by (i) promoting TBP binding (56) and by (ii) recruiting RNA polymerase II associated to TFIIF (57–59) through contacts with RPB1 and RPB2 (60–62). The complete PIC is finally assembled after bindings of TFIIE and TFIIH.

After PIC assembly, the DNA code has to be accessible to the enzyme in order to transcribe RNAs. Indeed, one DNA strand will serve as a substrate for complementary RNA synthesis, this strand is called the template strand, the other DNA strand is called the non-template strand. To access the template strand, hydrogen bonds between the two DNA strands need to be broken, and the resulting unfolded DNA region is called transcription bubble. The canonical model of DNA opening by the PIC involves ATP hydrolysis by the XPB subunit of TFIIH, providing the energy needed by TFIIH to translocate DNA through the PIC. DNA translocation is a rotational and translation motion of the double-stranded DNA towards the active site. Because of tight interactions

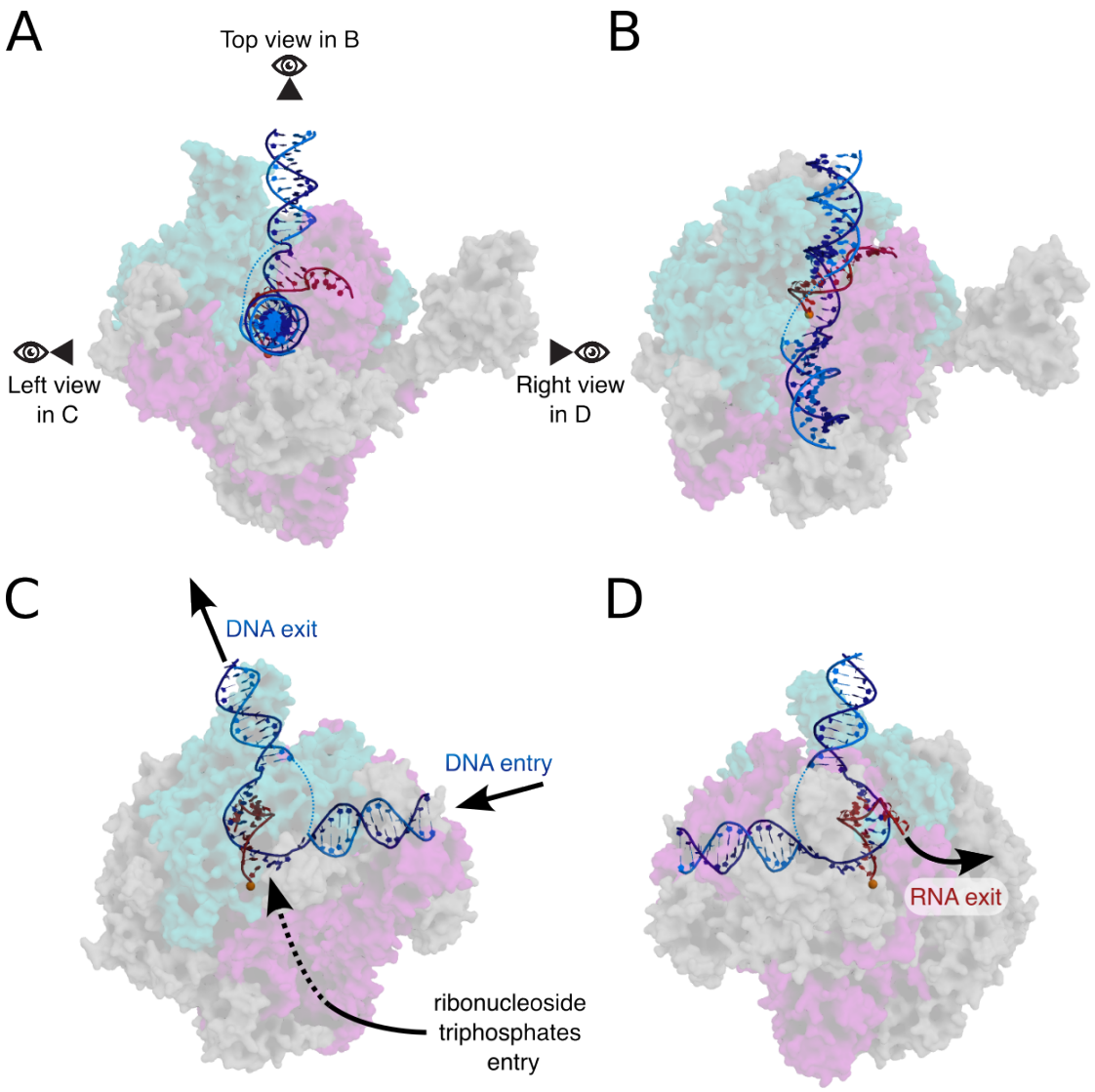

Figure 1.2: Transcription of RNA from a DNA matrix by RNAP II - (A–D) Different point of views of RNAP II during transcription. RNAP II subunits are represented as transparent surfaces, catalytic subunits RPB1 and RPB2 are colored in purple and cyan respectively, all other RPB subunits are depicted in grey. The DNA template strand and non-template strand are colored in dark blue and light blue respectively. Note that the non-template strand has not been resolved at the complex center and has been represented with light blue dashed line. The RNA product is depicted in red and the catalytic magnesium ion is depicted in orange. The RNAP II complex shown in this figure can be found in the pdb with the access code: 5FLM (37).

between the PIC and the promotor region, translocation is only possible up to this DNA region. Combination of DNA translocation towards the active site and restraint on the promotor region due to tight interactions with the PIC mechanically induces DNA unwinding (63–65). A recent study showed that transcription is possible in absence of the XPB subunit, while transcription is hindered by inhibition of the ATPase activity of XPB (66) needed for translocation. These data present XPB as an inhibitor of DNA opening as opposed to the classical model, inhibition released by the presence of ATP allowing translocation of DNA. This alternative model unifies DNA opening mechanism of RNAP I, II and III, as RNAP I and II do not bind to any translocases of helicases during initiation. In line with this alternative model, it has been suggested that TFIIH-independent DNA opening is possible in yeast RNAP II (67) or in human RNAP II under negative supercoiling conditions (68), where DNA melting would be driven by the binding energy released during PIC assembly.

1.3.2 Transcription elongation

After transcription initiation, the PIC is still tightly bound to the promotor region. To enter elongation, the PIC has to leave the promoter region through a process called promotor escape. Promotor escape allows DNA to slide through the PIC, and thus allows RNAPII to read the complete DNA sequence of a gene. Promotor escape is triggered by phosphorylation of Ser-5 in the C-terminal domain (CTD) of RPB1, and this phosphorylation is induced by the CDK7 kinase domain of TFIIH (69). Elongation factor proteins are also needed in order for the elongation to proceed.

Synthesis of RNAs starts at the transcription start site (TSS) which refers to the first deoxyribonucleotide to be read by the polymerase and is numbered: +1. The TSS is positioned ~ 30 base pairs (bp) downstream of the TATA-box. The polymerase synthesizes RNAs by growing the polynucleotide chain from the 5' end to the 3' end, one ribonucleotide at a time. In the canonical model of RNA synthesis, the cycle of ribonucleotide addition is composed of the following steps: (i) a ribonucleoside triphosphate (NTP) binds to the complementary deoxyribonucleotide from the template DNA strand, (ii) specific interactions form between the polymerase complex, the catalytic magnesium ions and the NTP, (iii) the RNA 3'-OH carries out a nucleophilic attack on the NTP γ-phosphate, (iv) RNAP II translocates along DNA to allow the next NTP to bind to the next DNA nucleotide template (33).

If RNAP II introduces by mistake a non-complementary nucleotide in the mRNA sequence, this could lead to proteins with defective functions. To prevent such deleterious effect, natural selection favored evolution of a proofreading mechanism within RNAPs allowing to correct possibly misincorporated nucleotides. By this proofreading mechanism, RNAP II demonstrates very low error rate, reaching up to one misinserted nucleotide every 2×10^5 nucleotides (70).

During elongation, crucial mRNA maturation processes occurred, including notably: 5' terminal capping of mRNA with a methylated guanosine triphosphate (^{m7}Gppp) (71). Addition of the 5' cap is carried out by three enzymes namely: a triphosphatase, a guanylyltransferase, and a methyltransferase. The mRNA capping process follows these steps: (i) the γ-phosphate of the first transcribed RNA residue is cleaved by the triphosphatase, (ii) a GMP is transferred to the remaining diphoshate at the 5' end of the first RNA residue by the guanylyltransferase, this results in a reverse 5' to 5'-triphosphate link between GMP and the first RNA residue, finally (iii) a methyl group is added to the guanine cap on amine N7 (71). Capping of mRNAs has many functions: (i) it protects mRNA from degradation, (ii) it is required to export mRNAs from the nucleus to the cytoplasm in higher eukaryotes (72), (iii) it allows cap-dependent initiation of protein synthesis (73), (iv) it is required for efficient mRNA splicing (74), and it is implicated in mRNA polyadenylation (75). Capping of mRNAs is not the only mRNA maturation process to occur co-transcriptionally. Indeed, splicing also mainly takes place during transcription elongation (76). During splicing, introns —non-coding sequences within mRNAs— are removed and exons —coding sequences within mRNAs— are linked together. Splicing is a regulation process of genetic expression as several orderings and combinations of exons are possible through alternative splicing, leading to different protein sequences derived from a single gene, thus participating to protein diversity (77).

1.3.3 Transcription termination

Two models have been proposed for transcription termination of protein-coding genes: the torpedo model and the allosteric model (35, 78, 79). In the torpedo model, cleavage of the mRNA product produces a new 5' end that enables 5'-3' exoribonuclease 2 to degrade the remaining RNA bound to RNAP II until it reaches the complex and triggers disassembly of RNAP II. In the allosteric model, association of termination transcription factors or disassembly of elongation factors lead to conformational changes of RNAP II

and mRNA release. After termination, mRNAs are polyadenylated, *i.e.* a sequence of adenoside nucleotides is added at the 3' end of mRNAs (80, 81). The poly(A) mRNA's tail is involved in translation initiation (82) and in mRNA stability (83).

1.4 Molecular dynamics simulation

This section has been written based on several books (84–86) and on the gromacs manual (87).

Atomic positions evolve under a potential energy. Molecular dynamics (MD) simulations model movements of atoms using classical mechanics. Each atom is described with its position in space and its velocity, and its position evolve under a potential called forcefield. The forcefield describes the inter-atomic interactions and is often taken as:

$$U(\mathbf{r}^N) = \sum_{i=1}^{N-1} \left(\frac{k_{\mathrm{b},i}}{2}\right)(r_i - r_{i0})^2 + \sum_{i=1}^{N-2} \left(\frac{k_{\theta,i}}{2}\right)(\theta_i - \theta_{i0})^2 + \sum_{i=1}^{N-3} u_i(\Phi_i) + \sum_{i<j} \left\{ \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]\right\} \quad (1.1)$$

where $\mathbf{r}^N$ refers to the whole set of x, y and z coordinates for N atoms. The first term refers to the interaction between pairs of bonded atoms and the second one refers to bond-angle vibration between three bonded atoms *i*-*j*-*k*. Bond stretching and bending angle interactions are defined as harmonic potentials centered in r_{i0} and θ_{i0} respectively and with force constants $k_{\mathrm{b},i}$ and $k_{\theta,i}$ respectively. In the third term, u_i is a function defining the dihedral angle between the *i*-*j*-*k* and the *j*-*k*-*l* planes of four bonded atoms *i*-*j*-*k*-*l*; u_i is generally defined as a sum of cosine functions. The forth term describes electrostatic and van der Waals interactions with a Coulomb and a Lennard-Jones potentials respectively. The first line in equation 1.1 gathers bonded interactions and the second line non-bonded interactions. All forcefield parameters are derived from experimental data or *ab initio* calculations or a combination of both.

In molecular dynamics simulations, atoms evolve according to Newton's second law:

$$m_i \frac{\partial^2 \mathbf{r}_i}{\partial t^2} = \mathbf{F}_i,\ i = 1...N \quad (1.2)$$

where m_i is the mass of atom i, t is the time and $\mathbf{F}_i$ is the force applied on atom i, and defined as the negative derivative of the potential function $U(\mathbf{r}^N)$:

$$\mathbf{F}_i = -\frac{\partial U}{\partial \mathbf{r}_i} \tag{1.3}$$

Several algorithms are available to integrate Newton's equation of motion; within the molecular dynamics package used throughout this thesis, the default integrator algorithm is called leap-frog algorithm(88) and applies the following equations:

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t+\frac{1}{2}\delta t) \tag{1.4}$$

$$\mathbf{v}(t+\frac{1}{2}\delta t) = \mathbf{v}(t-\frac{1}{2}\delta t) + \delta t \mathbf{a}(t) \tag{1.5}$$

where $\mathbf{v}(t)$ and $\mathbf{a}(t)$ are the velocity and the acceleration at time t respectively. The time-step δt has to be chosen carefully to allow accurate integration while keeping simulations computationally affordable (89).

Maintaining physiological temperature and pressure in MD simulations. To model molecular processes occurring in life, one needs to maintain a temperature and a pressure best reproducing those present in the studied organism. For instance, this implies to simulate at $\sim$310 K and $\sim$1 bar for human biomolecules. In order to simulate biomolecules at physiological constant temperature and pressure, *i.e.* to simulate in an NPT ensemble (constant number of particles, constant pressure and constant temperature), several algorithms allowing temperature and pressure coupling have been developed. Because the velocity rescaling algorithm (90) was extensively used to maintain constant temperature in the projects described in this thesis, we will briefly describe this approach. The instantaneous temperature relates to the kinetic energy as follows:

$$T = \frac{2K}{k_B n_{\mathrm{DOF}}} \tag{1.6}$$

with:

$$K = \frac{1}{2}\sum_{i=1}^{N} m_i |\mathbf{v}_i|^2 \tag{1.7}$$

where K is the kinetic energy, k_B the Boltzmann constant, n_{DOF} the number of degrees of freedom, m_i the mass of atom i, and $\mathbf{v}_i$ the velocity of atom i. In its simplest form, the velocity rescaling algorithm consists in rescaling the velocities by a factor λ at a predetermined frequency in order to reach the target temperature T_{target}. Using eq. 1.7 in eq. 1.6, and by scaling the instantaneous velocities by λ we obtain:

$$T_{\text{target}} = \frac{\sum_{i=1}^{N} m_i \lambda^2 |\mathbf{v}_i|^2}{k_B n_{\text{DOF}}} \tag{1.8}$$

solving for λ,

$$\lambda = \sqrt{\frac{k_B T_{\text{target}} n_{\text{DOF}}}{\sum_{i=1}^{N} m_i |\mathbf{v}_i|^2}} \tag{1.9}$$

which can also be written simply as:

$$\lambda = \sqrt{\frac{K_{\text{target}}}{K}} \tag{1.10}$$

where K_{target} and K are the target kinetic energy and the instantaneous kinetic energy respectively. The limitation of this simple implementation is that (i) the kinetic energy does not follow the canonical equilibrium distribution, and (ii) it disturbs considerably the velocities. Instead of choosing K_{target} exactly equal to the kinetic energy corresponding to T_{target}, one could draw K_{target} from the canonical equilibrium distribution, and therefore overcoming the first limitation mentioned above. However, the velocities will still exhibit fast fluctuations. In the implementation of velocity rescaling proposed by Bussi et al. (90) and which was used in this thesis, the change in kinetic energy is rather defined by:

$$\mathrm{d}K = (K_{\text{target}} - K)\frac{\mathrm{d}t}{\tau} + 2\sqrt{\frac{K_{\text{target}} K}{n_{\text{DOF}}}}\frac{\mathrm{d}W}{\sqrt{\tau}} \tag{1.11}$$

where $\mathrm{d}W$ refers to a Wiener noise, τ is the time scale of the thermostat, the first term in the addition corresponds to an exponential decay towards the target kinetic energy, and the last term in the addition corresponds to a stochastic term. This implementation ensures the sampling of proper canonical ensemble of velocities and applies only smooth changes of velocities as the rescaling procedure is distributed over several time steps.

Concerning pressure coupling, one can maintain constant pressure with Berendsen pressure coupling (91). This method consists in modeling the change in volume with an exponential decay towards a reference pressure P_{target} with: $\mathrm{d}P = (P_{\text{target}} - P)\mathrm{d}t/\tau$, where τ is similar to the τ parameter described above for the velocity rescaling thermostat as it controls the slope of the exponential decay. Another approach for pressure coupling is the Parrinello-Rahman barostat (92). It is superior to Berendsen as it generates a true NPT ensemble. The general idea of this method consists in a Lagrangian formulation of the equation of motion where positions of atoms are described with the box vectors and where the target pressure is a constraint. To obtain the target pressure, an external pressure is applied, leading to a change in volume of the simulation box. When the internal and external pressure are in equilibrium, the target pressure is reached.

1.5 Sampling challenge in molecular dynamics simulation

To understand the sampling challenge inherent to molecular dynamics simulations, we first have to introduce some statistical mechanics concepts (84–86).

Total energy of a system depends on the potential and kinetic energies:

$$H(\mathbf{p}^N, \mathbf{r}^N) = U(\mathbf{r}^N) + K(\mathbf{p}^N) \tag{1.12}$$

where H is the Hamiltonian, and equivalent here to the total energy, $\mathbf{r}^N$ has already been referred as the entire set of x, y and z coordinates for N atoms, and $\mathbf{p}^N$ is the set of p_x, p_y, p_z momenta for N atoms. Momentum of atom i is defined as $\mathbf{p}_i = m_i\mathbf{v}_i$.

We now introduce the Boltzmann distribution, which gives the probability of observing a particular microstate in phase space, *i.e.* the probability of observing our simulated atoms in a specific configuration in space and with a specific set of velocities. The Boltzmann distribution is defined as:

$$\begin{aligned} \rho(\mathbf{p}^N, \mathbf{r}^N) &= \frac{\exp\left(-\frac{H(\mathbf{p}^N,\mathbf{r}^N)}{k_B T}\right)}{\int \exp\left(-\frac{H(\mathbf{p}^N,\mathbf{r}^N)}{k_B T}\right) \mathrm{d}\mathbf{p}^N \mathrm{d}\mathbf{r}^N} \\ &= \frac{\exp\left(-\frac{K(\mathbf{p}^N)}{k_B T}\right) \exp\left(-\frac{U(\mathbf{r}^N)}{k_B T}\right)}{\int \exp\left(-\frac{K(\mathbf{p}^N)}{k_B T}\right) \exp\left(-\frac{U(\mathbf{r}^N)}{k_B T}\right) \mathrm{d}\mathbf{p}^N \mathrm{d}\mathbf{r}^N} \end{aligned} \tag{1.13}$$

where the denominator in eq. 1.13 is called partition function and is usually denoted Z. Because momenta and positions of an atom are independent we can factorize the integrals in Z and eq. 1.13 becomes:

$$\begin{aligned}
\rho(\mathbf{p}^N, \mathbf{r}^N) &= \left[\frac{\exp\left(\frac{-K(\mathbf{p}^N)}{k_B T}\right)}{\int \exp\left(\frac{-K(\mathbf{p}^N)}{k_B T}\right) \mathrm{d}\mathbf{p}^N}\right] \left[\frac{\exp\left(\frac{-U(\mathbf{r}^N)}{k_B T}\right)}{\int \exp\left(\frac{-U(\mathbf{r}^N)}{k_B T}\right) \mathrm{d}\mathbf{r}^N}\right] \\
&= \left[\frac{\exp\left(\frac{-K(\mathbf{p}^N)}{k_B T}\right)}{Z_{\text{kinetic}}}\right] \left[\frac{\exp\left(\frac{-U(\mathbf{r}^N)}{k_B T}\right)}{Z_{\text{config}}}\right] \\
&= \rho(\mathbf{p}^N) \times \rho(\mathbf{r}^N) \qquad (1.14)
\end{aligned}$$

where Z_{kinetic} and Z_{config} are the kinetic and the configuration integrals respectively. In this work we are only interested in the distribution of configurations and therefore we will only consider $\rho(\mathbf{r}^N) \propto \exp(-\beta U(\mathbf{r}^N))$, with β defined as the reciprocal of $k_B T$. We are usually not interested in properties of a single configuration, but rather in properties of ensembles of configurations; these ensembles are called macrostates. More specifically, we are typically interested in studying "stable" macrostates, that is to say ensemble of configurations lying in the same energy basin. For instance, from a drug discovery perspective, one could investigate the bound state of a drug to its target and the unbound state of this same drug to understand its binding mechanism. We will now simply refer to configurations belonging to the same energy basin as "state".

The probability ρ_{A} of observing a specific state A is linked to the free energy of this same state as follow:

$$\begin{aligned}
G_{\text{A}} &= -k_B T \ln\left(\frac{\int_{V_{\text{A}}} \mathrm{d}\mathbf{r}^N \exp(-\beta U(\mathbf{r}^N))}{Z_{\text{config}}}\right) \\
&= -k_B T \ln(\rho_{\text{A}}) \qquad (1.15)
\end{aligned}$$

where V_{A} is the "configurational volume" corresponding to state A. Because biomolecules are high dimensional systems, it is in practice impossible to compute Z_{config} and consequently it is also not possible to compute the absolute free energy of a state. However, computing the free energy difference between two states A and B is more accessible as the configuration integrals from the respective probability distributions cancel each other:

$$
\begin{aligned}
G_{\mathrm{A}} - G_{\mathrm{B}} &= -k_B T \ln\left(\frac{\int_{V_{\mathrm{A}}} \mathrm{d}\mathbf{r}^N \exp(-\beta U(\mathbf{r}^N))}{Z_{\mathrm{config}}}\right) + k_B T \ln\left(\frac{\int_{V_{\mathrm{B}}} \mathrm{d}\mathbf{r}^N \exp(-\beta U(\mathbf{r}^N))}{Z_{\mathrm{config}}}\right) \\
&= -k_B T \ln\left(\frac{\int_{V_{\mathrm{A}}} \mathrm{d}\mathbf{r}^N \exp(-\beta U(\mathbf{r}^N))}{\int_{V_{\mathrm{B}}} \mathrm{d}\mathbf{r}^N \exp(-\beta U(\mathbf{r}^N))}\right) \qquad (1.16)
\end{aligned}
$$

where V_{B} is the "configurational volume" corresponding to state B.

Because it is hard to visualize and understand biological processes in $3N$ dimensions, a projection of this high dimensional space on a lower dimensional space is usually used to describe these processes. The function mapping configurations from the $3N$–space to the lower dimensional space is commonly called collective variable (CV) and will be denoted as: $\hat{\xi}(\mathbf{r}^N)$. The value that this function yields will be denoted ξ. The definitions we gave earlier for the free energy in the full $3N$–space is conceptually similar to the definition of free energy in CV–space: $G(\xi) = -k_B T \ln(\rho(\xi))$; the free energy as a function of a collective variable is called potential of mean force (PMF).

Figure 1.3 shows an example of a two-state model —that is to say, two low energy states separated by a transition state located at the energy barrier— with associated thermodynamics and kinetics. For a real biological process, the potential of mean force would be rougher with multiple minima separated by multiple transition states, however this two-state model still holds in certain cases and is rigorous enough to explain basic thermodynamic and kinetic concepts we are interested in here. Obtaining a potential of mean force of the biological process of interest helps to understand (i) how probable two stable states are relative to each other, measured with free energy difference: ΔG, and (ii) how many transitions from one state to another are to be expected per time unit, measured with reaction rates: $k_{A\to B}$ and $k_{B\to A}$. Determining a collective variable that clearly describes the transition state is of foremost importance to compute accurate rates and to understand at the molecular level what triggers the transition from one stable state to another (more on this in section 1.5.1).

Now that we have introduced rates we can understand what is the sampling problem in atomistic molecular dynamics simulations. The mean first passage time, that is the reciprocal of a rate and which has dimension of a time, tells us on average how long one would have to observe a process until the first transition occurs. In molecular biology these timescales can span from the order of the microsecond for small molecule diffusing

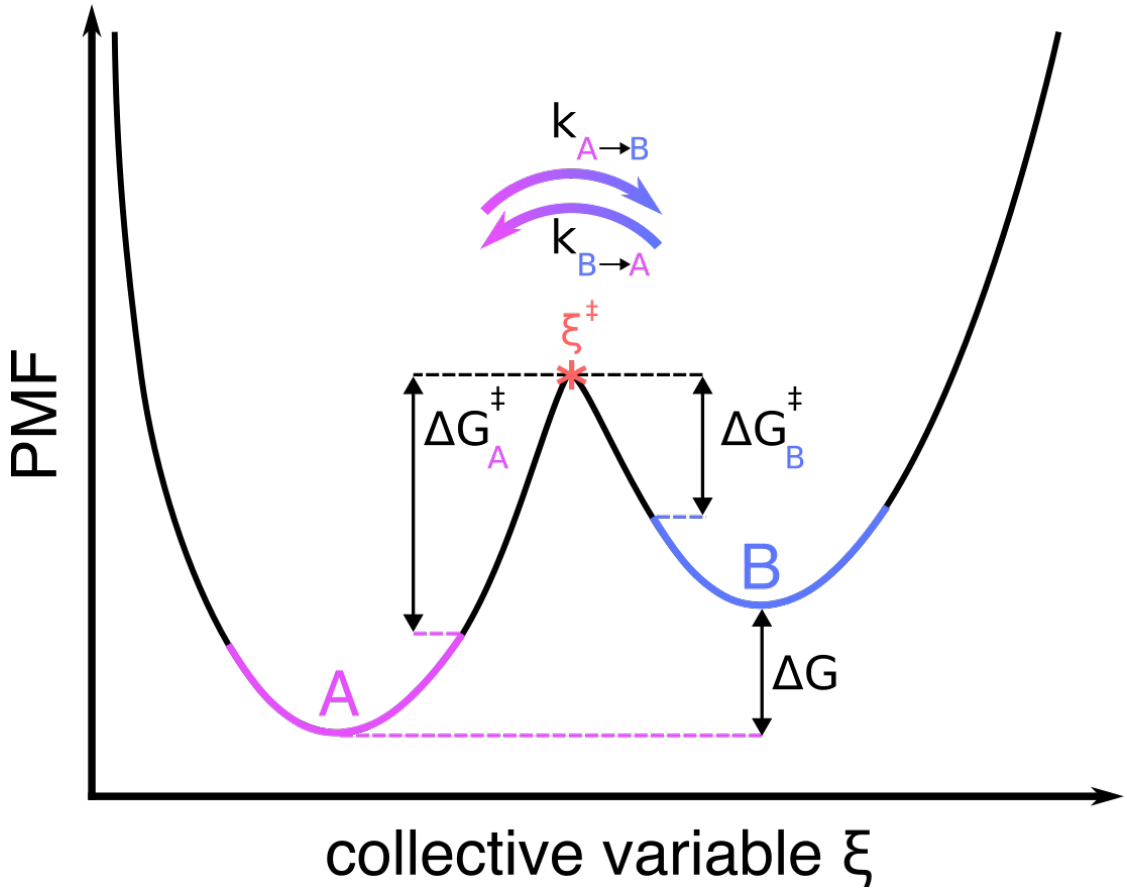

Figure 1.3: Potential of mean force of a two-state model with two minima in state A and B, as a function of a collective variable ξ - The red star represents the transition state and is denoted: $\xi^{\ddagger}$. Thermodynamic parameters of this two-state system: ΔG, $\Delta G_A^{\ddagger}$, and $\Delta G_B^{\ddagger}$ refers to the free energy difference between states A and B, the free energy difference between state A and the transition state, and the free energy difference between state B and the transition state respectively. Kinetic parameters: $k_{A\to B}$ and $k_{B\to A}$ are rates from A to B and from B to A respectively. Rates are related to $\Delta G_A^{\ddagger}$ and $\Delta G_B^{\ddagger}$ as follow: $k_{A\to B} \propto \exp(-\beta \Delta G_A^{\ddagger})$ and $k_{B\to A} \propto \exp(-\beta \Delta G_B^{\ddagger})$.

through a membrane channel to the order of the minute for transcription of a complete gene (93). Considering that with a 2 fs time-step and with current computational power (beside supercomputer) we can reach ∼33 ns/day for an OprO membrane system (∼94410 atoms) or ∼13 ns/day for a RNAP II system (∼832078 atoms), one would need several years to several million years of atomistic simulations to gather a statistically relevant number of transitions to state about thermodynamics or kinetics of the afore-

mentioned processes. This is the reason why several methods have been developed to speed up atomistic simulations, they will be discussed in next sections.

1.5.1 Overcoming the sampling challenge by adding an external potential along a collective variable

By adding a bias potential along one or few relevant collective variables to the forcefield defined previously in eq. 1.1, it is possible to guide simulations through the free energy landscape and therefore to enhanced configurational sampling. A correction has to be applied to the obtained configuration distribution in order to recover the true unbiased distribution. The most crucial aspect of enhanced sampling methods based on collective variable is to determine which function of atomic coordinates, *i.e.* which collective variable, would best characterize the studied process and, therefore, would be the most relevant to use for biasing the simulation. To emphasize how crucial the choice of the collective variable is, we have depicted in Fig. 1.4 two potential of mean forces of fictious two-state models: (i) one where the chosen collective variable (ξ_1) is sufficient to describe the A-B transition (Fig. 1.4A), and (ii) one potential of mean force where a slowly varying degree of freedom (ξ_2) is integrated over, leading to incorrect identification of the transition state (Fig. 1.4B).

When a collective variable rigorously describes the minima and the transition states (as ξ_1 in Fig. 1.4A), the term "reaction coordinate" is preferred (94, 95). A collective variable is rather called order parameter when it properly separates stable states but cannot identify transition states (94, 95) (as ξ_1 in Fig. 1.4B). The theoretical perfect reaction coordinate is called the committor function and gives the probability of a configuration to relax into a stable state before reaching any other stable states. In our two-state model in Fig. 1.4, commitment probability of B for a configuration is the probability for this configuration to relax into stable state B before reaching state A. Furthermore, commitment probability of B should be equal to 0.5 at the transition state —the transition state is the isocommittor point—. Hence, by launching simulations at the transition state identified by using a collective variable, and by computing commitment probabilities of several trajectories, one can check the validity of a putative reaction coordinate (94, 95). This test is illustrated in Fig. 1.4, where the black dotted arrows depicted in the first row are trial trajectories initiated at the transition states

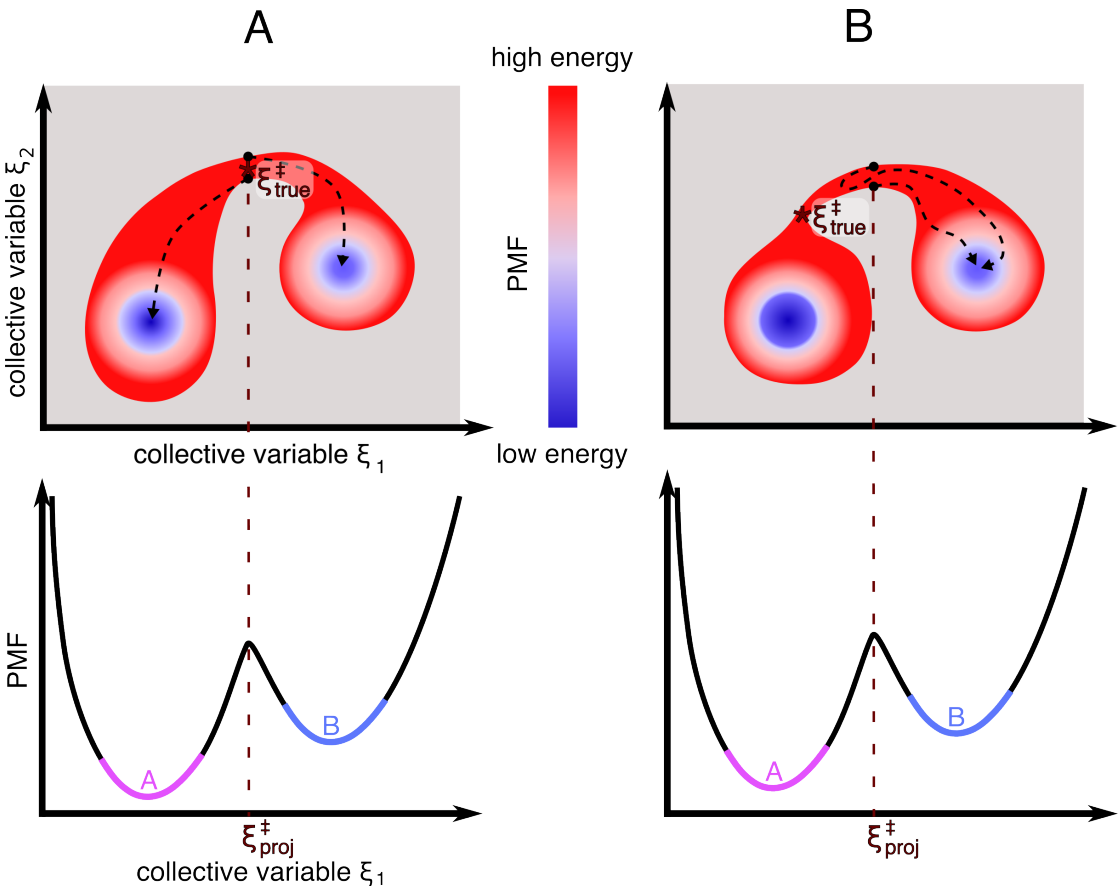

Figure 1.4: Effect of neglecting orthogonal slowly varying degree of freedom when computing PMFs - First row of A and B shows PMFs as a function of two collective variables ξ_1 and ξ_2. The second row corresponds to PMFs depicted in the first row but projected on ξ_1. The true transition states are denoted $\xi^{\ddagger}_{\text{true}}$, the transition states identified by projected PMFs on ξ_1 are denoted $\xi^{\ddagger}_{\text{proj}}$. Because computing a PMF as a function of a CV implies to integrate over other orthogonal degrees of freedom, the CV chosen (here ξ_1) to compute a PMF must contain all important degrees of freedom describing the A−B transition (as in column A); otherwise the transition state will be misidentified and will not correspond to the highest free energy region between the two stable states A and B (as in B). Black dotted arrows are trajectories of simulations started at the transition state $\xi^{\ddagger}_{\text{proj}}$ identified in PMFs projected on ξ_1. This figure has been inspired by Fig. 8 in (94).

$\xi^{\ddagger}_{\text{proj}}$ identified with PMFs in the second row. Because ξ_1 is indeed a reaction coordinate in Fig. 1.4A, trajectories started from $\xi^{\ddagger}_{\text{proj}}$ will have a probability of 0.5 to fall in state A before reaching B, and a probability of 0.5 to fall in state B before reaching A. However, because ξ_1 is not a reaction coordinate in Fig. 1.4B, trajectories started from

$\xi^{\ddagger}_{\mathrm{proj}}$ will have a higher probability to fall first in state B than to fall first in state A. In practice, finding the transition state and, therefore, describing kinetics of a process is challenging; thus, finding a good approximation of the reaction coordinate which is able to describe at least the thermodynamics of a process is already an achievement. A myriad of bias-based enhanced sampling techniques have been developed in the field, but we will detail in the next sections the two most popular ones that were also used in this work: umbrella sampling (US) (96) and metadynamics (97).

Umbrella sampling (US). The first step to apply umbrella sampling is to obtain configurations along a chosen collective variable. To this aim, steered molecular dynamics simulation (SMD), also called constant velocity pulling simulation, is often used (98, 99). Steered molecular dynamics simulation consists in adding a time-dependent harmonic restraint to the system's Hamiltonian and is defined as:

$$\begin{aligned} U_{\mathrm{smd}}(\xi, t) &= \frac{1}{2}k(\xi - \xi_0(t))^2 \\ &= \frac{1}{2}k(\xi - (\xi_0(t=0) + vt))^2 \end{aligned} \tag{1.17}$$

where k is the force constant, ξ defines the instantaneous position of the system in CV–space, $\xi_0(t)$ is the target CV value at time t, and v is a velocity defining how $\xi_0(t)$ changes through time. The next step in applying umbrella sampling is to run i independent simulations restrained with a bias harmonic potential $U_{\mathrm{bias},\,i}(\xi)$, these independent simulations are often called umbrella windows. For each umbrella window i, $U_{\mathrm{bias,\,i}}(\xi)$ is defined as: $U_{\mathrm{bias},\,i}(\xi) = \frac{1}{2}k(\xi - \xi_{0,\,i})^2$, where $\xi_{0,\,i}$ is the reference position along ξ for umbrella window i. The only difference between $U_{\mathrm{bias},\,i}(\xi)$ and $U_{\mathrm{smd}}(\xi, t)$ in eq. 1.17 is that the reference position of the harmonic potential is constant through time for $U_{\mathrm{bias,\,i}}(\xi)$. With these multiple umbrella windows spanning the CV-space, it is now possible to compute probabilities of our system to visit each region of the sampled CV-space. However, to recover unbiased probabilities and ultimately the unbiased PMF, one has to reweight the obtained biased probabilities. A popular method to compute PMF from biased simulations is called Weighted Histogram Analysis Method (WHAM) (100) and has been extensively used in this thesis; this method will therefore be briefly described in the following paragraph.

We first recall from section 1.5 that $\hat{\xi}(\mathbf{r}^N)$ is the function that maps the coordinates $\mathbf{r}^N$ to the collective variable ξ. The aim of WHAM is to recover the unbiased free energy profile, but as we have mentioned in section 1.5, only free energy differences are accessible as the partition function cannot be computed for high dimensional systems. Therefore, the free energies of any points ξ: $G(\xi)$, are in practice expressed relative to the free energy of a chosen reference point ξ^*: $G(\xi^*)$. By defining $G(\xi^*)$ equal to zero, eq. 1.16 becomes:

$$G(\xi) = -\beta^{-1}\ln\left(\frac{\rho(\xi)}{\rho(\xi^*)}\right) \tag{1.18}$$

where $\rho(\xi)$ and $\rho(\xi^*)$ are unbiased probability distributions. For each umbrella window i biased with $U_{\mathrm{bias,\,i}}(\xi)$, it is possible to express the obtained biased probability distribution $\rho_i^{\mathrm{biased}}(\xi)$ as:

$$\begin{aligned}
\rho_i^{\mathrm{biased}}(\xi) &= \frac{\int \mathrm{d}\mathbf{r}^N \delta\left(\hat{\xi}(\mathbf{r}^N)-\xi\right) \mathrm{e}^{-\beta\left(U(\mathbf{r}^N)+U_{\mathrm{bias,\,i}}[\hat{\xi}(\mathbf{r}^N)]\right)}}{\int \mathrm{d}\mathbf{r}^N \mathrm{e}^{-\beta\left(U(\mathbf{r}^N)+U_{\mathrm{bias,\,i}}[\hat{\xi}(\mathbf{r}^N)]\right)}} \\
&= \mathrm{e}^{-\beta U_{bias,\,i}(\xi)} \frac{Z_{\mathrm{config}}^{-1}\int \mathrm{d}\mathbf{r}^N \delta\left(\hat{\xi}(\mathbf{r}^N)-\xi\right) \mathrm{e}^{-\beta U(\mathbf{r}^N)}}{Z_{\mathrm{config}}^{-1}\int \mathrm{d}\mathbf{r}^N \,\mathrm{e}^{-\beta U_{\mathrm{bias,\,i}}[\hat{\xi}(\mathbf{r}^N)]}\mathrm{e}^{-\beta U(\mathbf{r}^N)}} \qquad (1.19) \\
&= \mathrm{e}^{-\beta U_{bias,\,i}(\xi)} \frac{\rho_i(\xi)}{\left\langle \mathrm{e}^{-\beta U_{\mathrm{bias,\,i}}[\hat{\xi}(\mathbf{r}^N)]}\right\rangle} \qquad (1.20)
\end{aligned}$$

where δ is the Dirac delta function, and $\rho_i(\xi)$ is the estimate of the unbiased distribution obtained from umbrella window i. Equation 1.19 is obtained by applying the definition of the Dirac delta function, moreover multiplying the numerator and the denominator by Z_{config}^{-1} makes the presence of $\rho_i(\xi)$ explicit in the numerator. To obtain 1.20 from 1.19 we simply applied the definition of an expectation value in the denominator. Using eq. 1.20 in 1.18 yields:

$$G_i(\xi) = -\beta^{-1}\ln\left(\frac{\rho_i^{\mathrm{biased}}(\xi)}{\rho(\xi^*)}\right) - U_{\mathrm{bias},i}(\xi) + F_i \tag{1.21}$$

where F_i is an undetermined constant defined via:

$$\mathrm{e}^{-\beta F_i} = \langle \mathrm{e}^{-\beta U_{\mathrm{bias,\,i}}[\hat{\xi}(\mathbf{r}^N)]} \rangle. \tag{1.22}$$

The difficulty to obtain $G(\xi)$ from the $G_i(\xi)$ resides in determining the F_i constants. To obtain $G(\xi)$ in WHAM, the overall unbiased distribution is expressed as a weighted sum over each unbiased distribution i:

$$\rho(\xi) = \sum_{i=1}^{N} \omega_i \rho_i(\xi) \tag{1.23}$$

where N is the number of umbrella windows, ω_i is the weight associated to the estimated unbiased probability distribution obtained from umbrella window i, the weights are constrained under: $\sum_{i=1}^{N} \omega_i = 1$. Then the best estimate for $\rho(\xi)$ is obtained by using eq. 1.20 and by minimizing the statistical errors of $\rho(\xi)$ leading to the following two WHAM equations that need to be solved self-consistently until convergence of $\rho(\xi)$:

$$\rho(\xi) = \frac{\sum_{i=1}^{N} n_i \rho_i^{\mathrm{biased}}(\xi)}{\sum_{i}^{N} n_i \exp[-\beta(U_{\mathrm{bias},\,i}(\xi) - F_i)]} \tag{1.24}$$

$$\exp(-\beta F_i) = \int \mathrm{d}\xi \exp[-\beta(U_{\mathrm{bias},\,i}(\xi)]\, \rho(\xi) \tag{1.25}$$

where n_i is the number of data points in the umbrella histogram i.

A critical test to check if important orthogonal degrees of freedom have been omitted when designing collective variables for US is to carry out SMD in forward and reverse directions along the collective variable, and further use these simulations to compute potential of mean forces with umbrella sampling. If these PMFs are different, it means that simulations follow different pathways because they are not guided through important degrees of freedom orthogonal to the designed collective variables. In some other cases, detecting ill-designed collective variables is immediately obvious from the computed PMF as we observe constant increase of energy along the collective variable, meaning that the system is following nonphysical pathways, *i.e.* pathways with very large associated free energy that would not be observed in nature; for this reason the system is always pushed back to the CV-region that was already explored and of lower

free energy. The fact that different pathways are followed depending on the initial condition or that a system keeps on backtracking along CV–space relative to the pulling direction in individual umbrella windows is often called "memory effect" or "hysteresis effect".

Metadynamics. Metadynamics adds an history-dependent bias potential on one or few collective variables, therefore discouraging the simulations to stay in states corresponding to CV values that are often explored, *i.e.* discouraging sampling of low energy regions. The metadynamics bias potential applied at time t is defined as:

$$U_G(\xi, t) = \int_0^t dt' \frac{w}{\tau_G} \exp\left(-\frac{[\xi(\mathbf{r^N}) - \xi(\mathbf{r^N}(t'))]^2}{2\sigma^2}\right) \tag{1.26}$$

where $\xi(\mathbf{r^N})$ is the instantaneous position of the system along the collective variable, $\xi(\mathbf{r^N}(t'))$ is the center of the Gaussian at time t' along the CV, w is the Gaussian height and τ_G is the deposition rate of Gaussians. Several flavors of metadynamics have been developed since its first formulation, among them we find notably (i) well-tempered metadynamics which uses decreasing Gaussian height through time to avoid exploring irrelevant high free energy space (101), and (ii) multiple walkers metadynamics which consists in several metadynamics simulations running in parallel and sharing the same history-dependent biased potential allowing to explore the free energy landscape much faster (102). The basic assumption of standard metadynamics is that after a sufficiently large time t, the bias potential compensates the unbiased PMF and, therefore, the latter can be estimated by $-U_G(\xi, t)$. For multiple walkers metadynamics, a WHAM approach similar to what has been described for umbrella sampling is a valid approach in order to recover the unbiased PMF (102, 103). Several other alternative methods have been proposed to unbias standard metadynamics simulation and its variants, details and comparison of some of these methods can be found in ref. (104). As we have emphasized already, the definition of the collective variable is critical to obtain a valid PMF; but unlike umbrella sampling for which poorly designed collective variable can be easily detected, detecting hysteresis effect in metadynamics might be tedious. Therefore, one has to carefully check the convergence of PMFs obtained with metadynamics before stating about thermodynamics, or *a fortiori* about kinetics.

1.5.2 Overcoming the sampling challenge with generalized-ensemble methods

Replica exchange. As it has been introduced in section 1.5, inefficient sampling is due to low reaction rates between states separated by high energy barriers (Fig. 1.3). Reaction rates can be expressed as a function of free energy differences between stable states and transition state:

$$\begin{aligned} k_{A\to B} &\propto \exp(-\beta\Delta G_A^{\ddagger}) \\ k_{B\to A} &\propto \exp(-\beta\Delta G_B^{\ddagger}) \end{aligned} \tag{1.27}$$

where $k_{A\to B}$ and $k_{B\to A}$ are reaction rates from state A to state B and from state B to state A respectively (Fig. 1.3), and $\Delta G_A^{\ddagger}$ and $\Delta G_B^{\ddagger}$ are free energy differences between the transition state and state A, the transition state and state B, respectively. To further rationalize how we can increase transition rates, we need to introduce the following thermodynamic relationships:

$$\Delta G = \Delta H - T\Delta S \tag{1.28}$$

$$\Delta H = \Delta U - P\Delta V \tag{1.29}$$

where H is the enthalpy, S the entropy, and U the internal energy (*i.e.* the potential energy). By using eq. 1.28 in eq. 1.27 we have:

$$\begin{aligned} k_{A\to B} &\propto \exp\left(-\beta(\Delta H_A^{\ddagger} - T\Delta S_A^{\ddagger})\right) \\ k_{A\to B} &\propto \exp\left(-\frac{\Delta H_A^{\ddagger}}{k_B T} + \frac{\Delta S_A^{\ddagger}}{k_B}\right) \end{aligned} \tag{1.30}$$

where $\Delta H_A^{\ddagger}$ and $\Delta S_A^{\ddagger}$ are respectively enthalpy and entropy differences between transition state and state A. From eq. 1.30 it is clear that, if the barrier is of enthalpic nature, one solution to increase $k_{A\to B}$ is to increase the temperature T.

Replica exchange molecular dynamics (REMD), also called parallel tempering, is one method using simulations at high temperature to overcome free energy barriers (105–108). In parallel tempering, several simulations (called replicas) of the same system

kept at different temperatures are launched in parallel, and at regular time interval, exchange between configurations from neighboring replicas is attempted according to the Metropolis criterion (109). Because exchanges are ruled by the Metropolis criterion, simulations fulfill detailed balance. The probability of exchange acceptance α between neighboring replicas i and j is taken as:

$$\alpha = \min\left[1, \exp\left(\left(\frac{1}{k_B T_j} - \frac{1}{k_B T_i}\right)\left(U(\mathbf{r_j}) - U(\mathbf{r_i})\right)\right)\right] \tag{1.31}$$

where $\mathbf{r_j}$ and $\mathbf{r_i}$ are configurations in replicas j and i respectively, and T_j and T_i are temperatures of replicas j and i respectively. If $\alpha = 1$, then configurations are exchanged between replicas, else if $\alpha \leq 1$ a random number R is drawn between 0 and 1 from a uniform distribution and exchange of configurations is accepted if $R \leq \alpha$. With this method, reactive transitions are favored in replicas at higher temperature, and the replica at the target temperature is also able to sample states explored at higher temperatures with correct Boltzmann distribution due to configuration exchanges with replicas at higher temperature. Equation 1.31 shows that if the temperature difference between replicas is too high, the probability of exchange acceptance will be very low, however if the temperature difference is too low then the probability of crossing a free energy barrier will be lower, defeating the purpose of parallel tempering. It has been shown empirically through several studies that an optimal acceptance probability for exchange resides between 10% and 40% (110–118). One issue with parallel tempering is that temperature spacing between replicas needed to obtain acceptance probability in the [10%, 40%]-range decreases with system size (87), limiting the upper bound of temperature range used for REMD for large systems, hence the difficulty to apply REMD in those cases.

To overcome limitations found in parallel tempering when studying large systems, Hamiltonian replica exchange (HREX) techniques have been developed, and more specifically the REST2 variant of replica exchange solute tempering (REST) (119, 120). In HREX, not only temperature can be scaled between replicas, but also the Hamiltonian, yielding the following probability of exchange acceptance:

$$\alpha = \min\left[1, \exp\left(\frac{-U_i(\mathbf{r_j}) + U_i(\mathbf{r_i})}{k_B T_i} + \frac{-U_j(\mathbf{r_i}) + U_j(\mathbf{r_j})}{k_B T_j}\right)\right] \tag{1.32}$$

where U_i and U_j are scaled forcefield potentials applied in replicas i and j respectively. By scaling down forcefield parameters, one aims to decrease $\Delta H_A^{\ddagger}$ (eq. 1.29) and ultimately to increase transition rates (eq. 1.30). Indeed decreasing enthalpy difference between the transition state and the stable state is equivalent to increasing the temperature (eq. 1.30). Moreover, as the potential energy is an extensive property, it is possible to scale specific parts of the Hamiltonian that are thought to be relevant to decrease free energy barriers, which is not possible with an intensive property like temperature. In practice, one could for example scale down non-bonded interactions and dihedral potentials of the solute of interest, and keeping default forcefield parameters of solvent (120). In this regard, REST2 requires a rough prior knowledge of important degrees of freedom of the studied transitions, as opposed to parallel tempering where no prior knowledge is needed. However the computational cost of REST2 is lower than for parallel tempering, especially for large systems.

Another method called replica-exchange umbrella sampling (REUS, also referred to as windows-exchange umbrella sampling or bias-exchange umbrella sampling) belongs to HREX methods and is used to alleviate potentially ignored slowly varying degrees of freedom orthogonal to chosen collective variables (108). In REUS, exchange of configurations between neighboring umbrella windows is attempted at regular time interval, each umbrella window being harmonically restrained at different position along chosen CVs. Temperature is usually kept constant and therefore the probability of exchange acceptance can be written simply as:

$$\alpha = \min\left[1, \exp\left(\frac{U_i(\mathbf{r_i}) - U_i(\mathbf{r_j}) + U_j(\mathbf{r_j}) - U_j(\mathbf{r_i})}{k_B T}\right)\right] \tag{1.33}$$

where U_i and U_j are forcefield potentials and bias potentials applied in windows i and j respectively, $\mathbf{r_i}$ and $\mathbf{r_j}$ are configurations from windows i and j respectively. With this method, windows exploring degrees of freedom orthogonal to the biased CV are allowed to diffuse along the biased CV-space, and these orthogonal degrees of freedom are then better sampled all along our CV of interest.

Simulated tempering. In simulated tempering (121), temperature is treated as a variable in addition to configuration. The probability distribution of a configuration $\mathbf{r_i}$ at temperature T_k is given by:

$$\rho(\mathbf{r_i}, T_k) \propto \exp\left(-\frac{U(\mathbf{r_i})}{k_B T_k} + g_k\right) \tag{1.34}$$

where g_k is a temperature dependent weight factor. Simulated tempering is a serialization of parallel tempering as configurations are sampled under different temperatures within a single simulation. Transition between temperature states, *i.e.* transition between simulation at T_j and T_i, is ruled by a Metropolis criterion, and the probability of exchange acceptance is given by:

$$\alpha = \min\left[1, \exp\left(-\frac{U(\mathbf{r_j})}{k_B T_j} + \frac{U(\mathbf{r_i})}{k_B T_i} - (g_i - g_j)\right)\right] \tag{1.35}$$

where weight factors g_i and g_j have to be adjusted in order to sample equally all temperature states.

1.5.3 Boosting configurational sampling by increasing the time-step

As already stated when describing integration of Newton's equation of motion in eq. 1.4, choosing a right time-step is crucial to find good balance between calculation time and calculation accuracy. Indeed, if the time-step is too small trajectories will sample a small part of the configurational space; and if the time-step is too large instabilities will arise in the integration of the equation of motion due to some atoms visiting region of potential energy with steep slope. Therefore, the simulation time-step is bound to the fastest simulated atomic motions.

In atomistic molecular dynamics simulations, bond-stretching vibrations involving hydrogens are the fastest atomic motions with a period of about 10 fs (89). In addition, these types of motion are rather of quantum mechanical character than of classical mechanical character, and therefore cannot be rigorously described with classical dynamics. One solution to better describe bond-stretching involving hydrogen and to suppress dependency of time-step on these fast motions is to constrain bond lengths to a fixed distance. Indeed, a quantum oscillator in its ground state resembles a constrained bond more closely than a classical oscillator (87). Three algorithms are available in Gromacs to carry out constraint dynamics (122–124), the main idea of these algorithms is to solve

the equation of motion with additional Lagrange multipliers. By constraining bond-stretching involving hydrogens, it is possible to run molecular dynamics simulations with a 2 fs time-step instead of a 1 fs time-step.

The next fastest atomic motions are angle bendings involving hydrogen atoms, with a period of about 13 fs (89). Two methods can be used to increase this period, and thus to increase further the time-step: virtual interaction sites and hydrogen mass repartitioning (HMR) (89, 125, 126). By modeling hydrogens as virtual interaction sites, their positions are calculated from positions of three nearby heavy atoms, therefore removing oscillation of angle bendings involving these atoms. With hydrogen mass repartitioning, hydrogen masses are scaled up by a factor f_{H} and heavy atom masses connected to hydrogens are scaled down; by this mean, the oscillation period of bond angles involving hydrogen atoms are increased while keeping overall masses of chemical moieties constant. With these two methods, one can use a 4 fs time-step during simulations, doubling the performance of atomistic molecular dynamics simulations.

By using hydrogen mass repartitioning, we artificially change atom masses and concerns about the relevance of the simulations to model our physical world can be raised. To answer these concerns, we have to look where atom masses are involved in atomic models used in MD and what are observables we are interested in. The average of an observable A in the canonical ensemble is defined as:

$$\begin{aligned} \langle A \rangle &= \frac{\iint A \exp(-\beta H) \mathrm{d}\mathbf{p}^N \, \mathrm{d}\mathbf{r}^N}{\iint \exp(-\beta H) \mathrm{d}\mathbf{p}^N \, \mathrm{d}\mathbf{r}^N} \\ &= \frac{\iint A \exp(-\beta U(\mathbf{r}^N)) \exp(-\beta K(\mathbf{p}^N)) \mathrm{d}\mathbf{p}^N \, \mathrm{d}\mathbf{r}^N}{\iint \exp(-\beta U(\mathbf{r}^N)) \exp(-\beta K(\mathbf{p}^N)) \mathrm{d}\mathbf{p}^N \, \mathrm{d}\mathbf{r}^N} \end{aligned} \tag{1.36}$$

where H is the Hamiltonian. Now if our observable is only dependent on positions, *i.e.* if $A = A(\mathbf{r}^N)$, eq. 1.36 becomes:

$$\begin{aligned} \langle A \rangle &= \frac{\int A(\mathbf{r}^N) \exp(-\beta U(\mathbf{r}^N)) \, \mathrm{d}\mathbf{r}^N \int \exp(-\beta K(\mathbf{p}^N)) \mathrm{d}\mathbf{p}^N}{\int \exp(-\beta U(\mathbf{r}^N)) \, \mathrm{d}\mathbf{r}^N \int \exp(-\beta K(\mathbf{p}^N)) \mathrm{d}\mathbf{p}^N} \\ &= \frac{\int A(\mathbf{r}^N) \exp(-\beta U(\mathbf{r}^N)) \, \mathrm{d}\mathbf{r}^N}{\int \exp(-\beta U(\mathbf{r}^N)) \, \mathrm{d}\mathbf{r}^N} \end{aligned} \tag{1.37}$$

The only term in eq. 1.37 that depends on masses is the kinetic energy, however this term cancels out if our observable is only dependent on positions. Therefore in

theory, when observables of interest are only position-dependent, like for example the free energy, results obtained with and without hydrogen mass repartitioning should be comparable.

2

Aims of the project

The ultimate goal of this project is to unravel atomistic details of DNA opening within the pre-initiation complex during transcription initiation. Because molecular dynamics (MD) simulation is able to provide atomic mechanistic insights of biomolecular processes, MD simulations were the method of choice in this thesis to study DNA opening during transcription initiation. However, as this has been introduced, DNA opening is a complex conformational change involving large roto-translational motions at time scales that are not accessible to brute-force MD simulations with current compute power. Thus, a major part of this project has been dedicated to develop and test methods allowing to overcome the sampling challenge.

3

Finding an RMSD–based collective variable to drive large-scale conformational changes

3.1 Proposed RMSD-based collective variable: ξ_{prop}

When the two end states of a conformational transition are known (*e.g.* from cryo-EM or x-ray crystallography experiments), one can use the known 3D structure of these states to drive conformational transition between states by comparing instantaneous conformations with reference states. A widely used metric for protein structure comparison is the root-mean-squared distance (RMSD). After aligning structures we want to compare (127), the RMSD is defined as:

$$\Delta(X,Y) = \sqrt{\frac{\sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{y}_i)^2}{N}} \tag{3.1}$$

where X denotes the set of 3D coordinates for a selection of atoms, *e.g.* C_α atoms, of a configuration to compare with the set of 3D coordinates of another configuration Y. N is the number of atoms to compare, $\mathbf{x}_i$ and $\mathbf{y}_i$ are coordinates of atom i in configuration X and Y respectively.

A collective variable based on RMSDs relative to two end states is commonly used in the literature (128) and has been defined as:

$$\xi_{\text{com}}(X) = \Delta(X, X_B) - \Delta(X, X_A) \tag{3.2}$$

where $\Delta(X, X_A)$ and $\Delta(X, X_B)$ are RMSDs between an instantaneous configuration X and reference states A and B respectively. In order to better compare ξ_{com} with our proposed collective variable which will be introduced hereafter, we will use a normalized form of eq. 3.2:

$$\xi_{\text{com}}(X) = \frac{\Delta(X, X_B) - \Delta(X, X_A)}{\Delta(X_A, X_B)} \tag{3.3}$$

where $\Delta(X_A, X_B)$ is RMSD between reference state A and B. As shown in Fig. 3.1A, ξ_{com} is sub-optimal for defining the two end states A and B. Therefore, we proposed the following functional form to better describe the two end states:

$$\xi_{\text{prop}}(X) = \left(\frac{\Delta_B}{\Delta(X, X_B)}\right)^{\alpha} - \left(\frac{\Delta_A}{\Delta(X, X_A)}\right)^{\alpha} \tag{3.4}$$

where Δ_A and Δ_B are RMSD fluctuations of configurations within state A and B respectively, and α is a tuning parameter controlling curvature of isolines in Fig. 3.1B. Notably, we have found a very close functional form of our proposed CV in the literature (129).

3.2 Testing ξ_{prop} on alanine dipeptide toy model

In order to validate our proposed CV, we decided to test it on the alanine dipeptide (CH3-CONH-CHCH3-CONH-CH3) toy model in vacuum for driving isomerization between two stable states namely: c_{7eq} and c_{ax} (Fig. 3.2). For alanine dipeptide isomerization eq. 3.4 becomes:

$$\xi_{\text{prop}}(X) = \left(\frac{\Delta_{c_{7eq}}}{\Delta(X, X_{c_{7eq}})}\right)^{\alpha} - \left(\frac{\Delta_{c_{ax}}}{\Delta(X, X_{c_{ax}})}\right)^{\alpha} \tag{3.5}$$

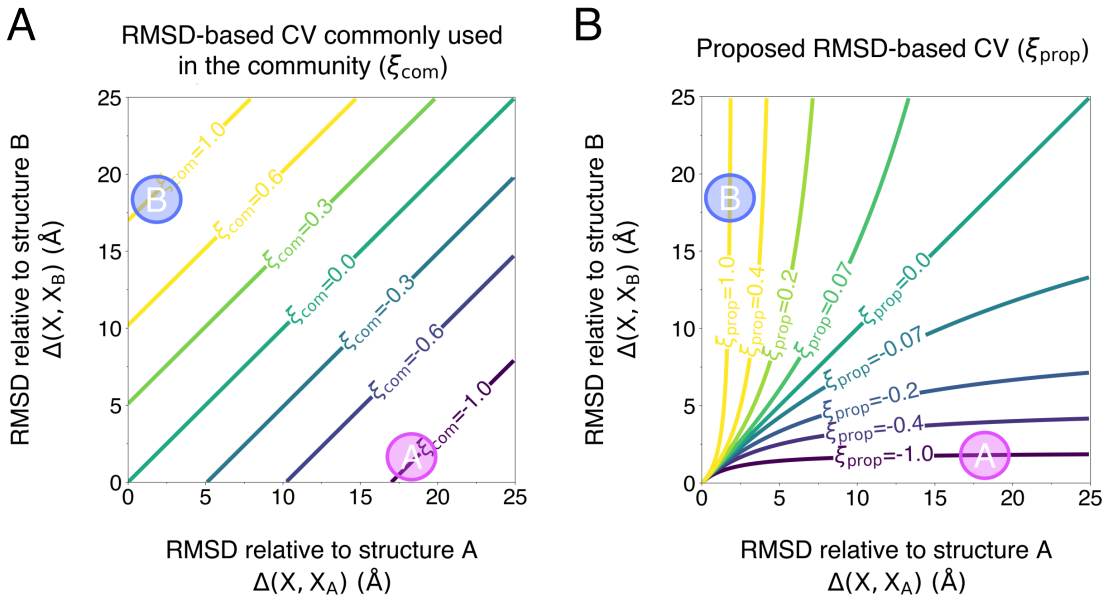

Figure 3.1: Projection of RMSD–based CVs ξ_{com} and ξ_{prop} onto the $\Delta(X, X_A)$ $\Delta(X, X_B)$-plane (see equations 3.3 and 3.4) - (A) Restraining the system to $\xi_{\text{com}} = +1$ or $\xi_{\text{com}} = -1$ is not optimal to sample stable states A and B. Indeed, these configurational-space regions also include configurations up to ∼7.5 Å–RMSD away from reference states. (B) Restraining the system to $\xi_{\text{prop}} = -1$ or $\xi_{\text{prop}} = +1$ allows to specifically sample reference states A and B respectively.

Figure 3.2B shows that RMSD fluctuation within both states is of ∼0.05 nm, thus we can consider that: $\Delta_{c_{7eq}} \approx \Delta_{c_{ax}} \approx 0.05$ nm. Under this approximation eq. 3.5 becomes:

$$\xi_{\text{prop}}(X) = \left(\frac{0.05}{\Delta(X, X_{c_{7eq}})}\right)^{\alpha} - \left(\frac{0.05}{\Delta(X, X_{c_{ax}})}\right)^{\alpha} \tag{3.6}$$

Relevant degrees of freedom describing isomerization of alanine dipeptide are the two dihedral angles: ϕ and ψ; therefore, we have first computed a PMF as a function of ϕ and ψ to have a reference PMF for comparison (Fig. 3.2A).

Pulling simulations along ξ_{prop} exhibit hysteresis effect. After carrying constant velocity pulling simulations along ξ_{prop} to drive the system from c_{ax} to c_{7eq} and from c_{7eq} to c_{ax}, we observed that different paths were sampled depending on the pulling direction (Figs. 3.3A–B); the paths taken when starting pulling simulations from c_{7eq} and c_{ax} will be denoted Path 1 and Path 2 respectively (Figs. 3.3A-B). As explained in the

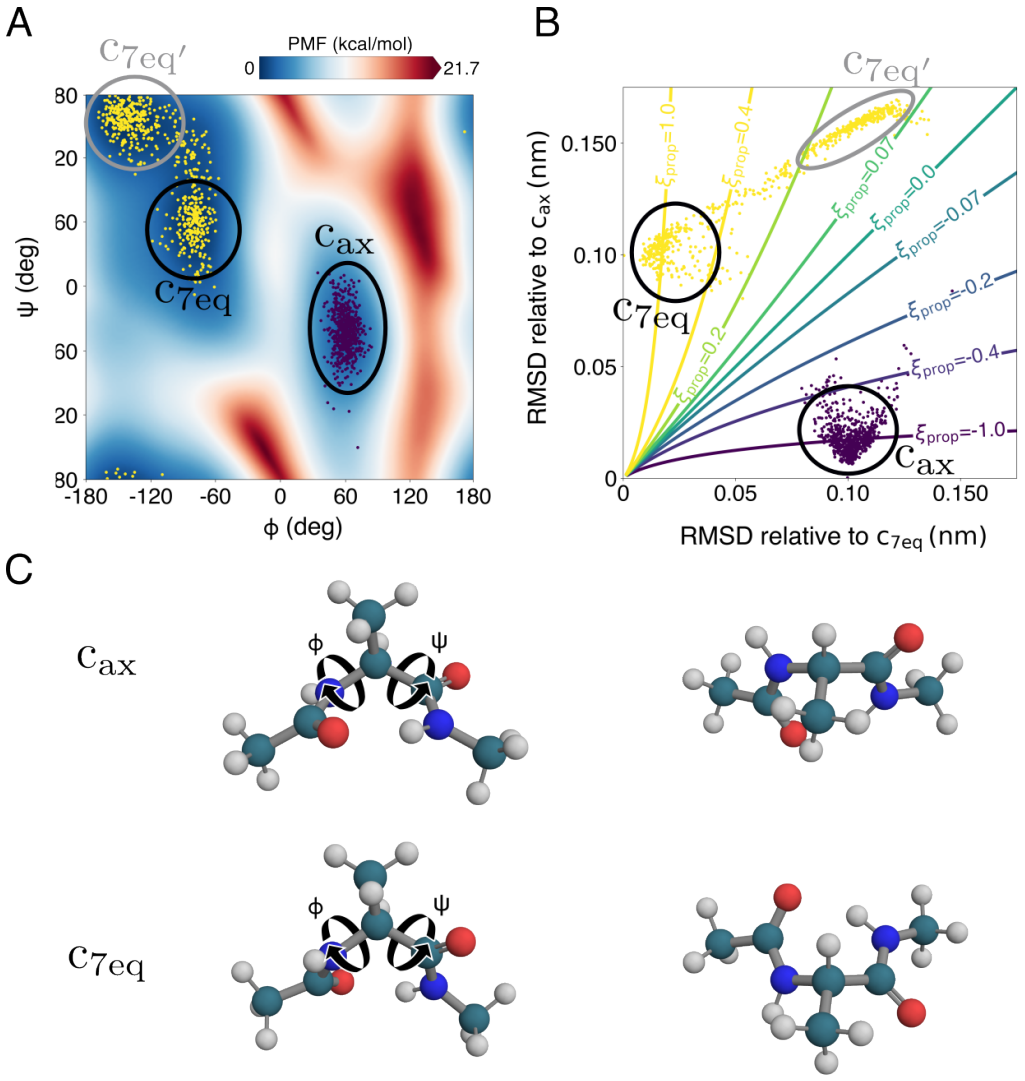

Figure 3.2: Alanine dipeptide overview - (A) Reference alanine dipeptide PMF computed with well-tempered metadynamics. Purple and yellow points represent configurations obtained from 1 ns of free simulation started from c_{ax} and c_{7eq} respectively. When simulating from c_{7eq}, another stable state $c_{7eq'}$ appears. (B) Purple and yellow points are defined as in (A) and are represented on the $\Delta(X, X_{c7eq})$ $\Delta(X, X_{cax})$-plane. Isosurfaces of $\xi_{prop}(X)$ are projected on the RMSD–space, here α =0.5 (see eq. 3.6). (C) Representative configuration of c_{7eq} and c_{ax} states with two points of view.

introduction, this behavior suggests that our proposed collective variable is not a good approximation of the reaction coordinate for isomerization of alanine dipeptide. Indeed,

as shown in Figs. 3.3A–C, the same value of ξ_{prop} maps configurations with different free energies (in Fig. 3.3 we took two configurations at $\xi_{\text{prop}} = -0.21$ as examples). The main issue with our proposed collective variable is that RMSD relative to a reference configuration is a degenarated metric, meaning that a lot of configurations are mapped to large values of RMSD. Therefore, ξ_{prop} gathers configurations with different thermodynamic properties for large RMSD relative to c_{7eq} and c_{ax}, *i.e.* close to $\xi_{\text{prop}} = 0$. Accordingly ξ_{prop} is not a suitable reaction coordinate for conformational change. Nevertheless, ξ_{prop} is a useful order parameter as it can clearly identifies configurations close to our reference states, *i.e.* close to $\xi_{\text{prop}} = 1$ and close to $\xi_{\text{prop}} = -1$.

Combining ξ_{prop} with a second RMSD-based CV alleviates hysteresis effect. In order to overcome degeneracy of our proposed collective variable for values close to $\xi_{\text{prop}} = 0$, we introduced a second collective variable that is RMSD of instantaneous configuration relative to a configuration close to $\xi_{\text{prop}} = 0$. This configuration has been extracted from trajectories following either Path 1 or Path 2, in order to specifically drive the simulation through one of these two paths. We defined: $\xi_{\text{mid1}} = \Delta(X, X_{\text{mid1}})$ and $\xi_{\text{mid2}} = \Delta(X, X_{\text{mid2}})$ as the RMSD relative to the configuration extracted from Path 1 (X_{mid1}) and Path 2 (X_{mid2}) respectively. As shown in Figs. 3.4A and 3.4C, pulling simulations along ξ_{prop} and ξ_{mid1} followed Path 1, independently of the starting configuration. Similarly, pulling simulations along ξ_{prop} and ξ_{mid2} were effectively restrained along Path 2 whether the simulation where started from c_{7eq} or c_{ax}.

Pulling simulations along Path 1 and Path 2 were then used to perform 2D umbrella sampling. From these US simulations, we obtained PMFs shown in Figs. 3.4B and 3.4D. From these PMFs, we computed a free energy difference between c_{ax} and c_{7eq} of: $2.11 \pm 0.20\,\text{kcal.mol}^{-1}$ for Path 1 and of $1.87 \pm 0.05\,\text{kcal.mol}^{-1}$ for Path 2, values which are comparable to free energy difference computed from our reference metadynamics simulation using ϕ and ψ angles as reaction coordinates: $1.71 \pm 0.03\,\text{kcal.mol}^{-1}$.

3.3 Discussion

With this study, we showed that it is not possible to drive conformational change with ξ_{prop} alone because a large number of configurations with different thermodynamic properties correspond to regions close to $\xi_{\text{prop}} = 0$. By choosing a reference configuration

observed in a specific pathway and close to $\xi_{\text{prop}} = 0$, we defined an additional CV that allows to single out configurations within the $\xi_{\text{prop}} = 0$ isosurface that belong to a specific pathway. By this mean, we were able to compute PMFs displaying free energy difference between reference states comparable to free energy difference computed with the reference PMF obtained with ϕ and ψ dihedral angles. Despite these encouraging results, isomerization of alanine dipeptide in vacuum is a very simplistic model compare to protein conformational changes we would like to tackle, *i.e.* DNA opening during transcription initiation. Indeed, with increasing system size we expect to encounter additional RMSD degeneracy problems. Further testing of our combination of CVs on conformational change displayed in lysozyme L99A T4, a system often used for method validations, confirmed that a path defined with three reference configurations is not sufficient to obtain PMFs without hysteresis effects.

Defining collective variables permitting to drive a system through a particular pathway has been subject of several studies in the literature. Among them, the string method (130–134) and path collective variable (PCV) (135) have proven to be efficient methods for enhancing conformational changes (136–141). For our next project: studying DNA opening during transcription initiation, we have chosen to use PCV because it is implemented in PLUMED(142), a plugin for MD packages, facilitating its application.

3.4 Materials and Methods

Simulation setup. Molecular dynamics simulations were carried out with Gromacs version 2018.6 patched with Plumed 2.5.1. Initial atomic coordinates of c_{7eq} and c_{ax} were obtained during the CECAM School "Open source software for enhanced-sampling simulations". Amber99sb-ildn forcefield was used to parameterize alanine dipeptide (143). Stochastic dynamics (SD) integrator was used to integrate the equation of motion. Simulations were carried out in NVT at a temperature of 300 K. The SD integrator was used as a thermostat with a friction constant of $0.5\,\text{ps}^{-1}$. All bonds were constrained with the Lincs algorithm (144).

Obtaining intermediate configurations to define ξ_{mid1} and ξ_{mid2}. Constant-velocity pulling simulations of 20 ns were carried out from $\xi_{\text{prop}} = +1$ to $\xi_{\text{prop}} = -1$ and from $\xi_{\text{prop}} = -1$ to $\xi_{\text{prop}} = +1$ with a force constant of $20{,}000\,\text{kJ}\,\text{mol}^{-1}$. From

these simulations, configurations corresponding to $\xi_{\text{prop}} = 0$ that satisfied equal RMSD–distance between c_{7eq} and c_{ax} were selected to define ξ_{mid1} and ξ_{mid2}. For these pulling simulations and all simulations described bellow, we have chosen $\alpha = 0.5$ in eq. 3.6.

Generating configurations along specific pathway for umbrella sampling. In order to generate starting configurations for US, we have first carried out constant-velocity pulling simulations along ξ_{prop} and ξ_{mid1} for sampling Path 1, and along ξ_{prop} and ξ_{mid2} for sampling Path 2. To check that similar pathways were followed regardless of the pulling direction along ξ_{prop}, we performed five steered MD simulations from $\xi_{\text{prop}} = +1$ to $\xi_{\text{prop}} = -1$ and five steered MD simulations from $\xi_{\text{prop}} = -1$ to $\xi_{\text{prop}} = +1$, for a total of ten simulations per pathways. We set force constants to $5000\,\text{kJ mol}^{-1}$ for ξ_{prop} and to $10{,}000\,\text{kJ mol}^{-1}\,\text{nm}^{-2}$ for ξ_{mid1} and ξ_{mid2}. For Path 1, the simulations were steered from $\xi_{\text{mid1}} = 0.142\,\text{nm}$ to $\xi_{\text{mid1}} = 0\,\text{nm}$ during the first 10 ns; then steered from $\xi_{\text{mid1}} = 0\,\text{nm}$ to $\xi_{\text{mid1}} = 0.152\,\text{nm}$ during the last 10 ns. For Path 2, the simulations were steered from $\xi_{\text{mid1}} = 0.071\,\text{nm}$ to $\xi_{\text{mid1}} = 0\,\text{nm}$ during the first 10 ns; then steered from $\xi_{\text{mid1}} = 0\,\text{nm}$ to $\xi_{\text{mid1}} = 0.071\,\text{nm}$ during the last 10 ns.

Umbrella sampling. We then performed US with 100 umbrella windows, with reference positions between $\xi_{\text{prop}} = -1$ and $\xi_{\text{prop}} = +1$ in steps of 0.02, using a force constant of 8000 kJ mol^{-1} and with 80 ns of simulation per window. Reference positions of harmonic potential along ξ_{mid1} and ξ_{mid2} were chosen equal to the reference position from which the frame were extracted in the initial constant velocity-pulling. Force constants were set to $10{,}000\,\text{kJ mol}^{-1}\,\text{nm}^{-2}$ for ξ_{mid1} and to $35{,}000\,\text{kJ mol}^{-1}\,\text{nm}^{-2}$ for ξ_{mid2}. PMFs were computed with the weighted histogram analysis method (100) implemented for 2D PMF (Grossfield, Alan, "WHAM: the weighted histogram analysis method", version 2.0.11, `http://membrane.urmc.rochester.edu/wordpress/?page_id=126`). Errors were computed with block averaging, using eight time blocks of 10 ns.

Reference well-tempered metadynamics simulation. Well-tempered metadynamics simulation of 80 ns was carried out on ϕ and ψ dihedral angles. Simulation was set-up with a 1 ps time-step Gaussian deposition rate, a Gaussian width of 0.35 rad and a Gaussian height of $1.2\,\text{kJ mol}^{-1}$. After reweighting the histograms accumulated on a ϕ-ψ grid,

the relationship: $G(\xi) = -k_B T \ln\rho(\xi)$ was used in order to compute the reference PMF. Block averaging with eight time blocks were used to compute error bars.

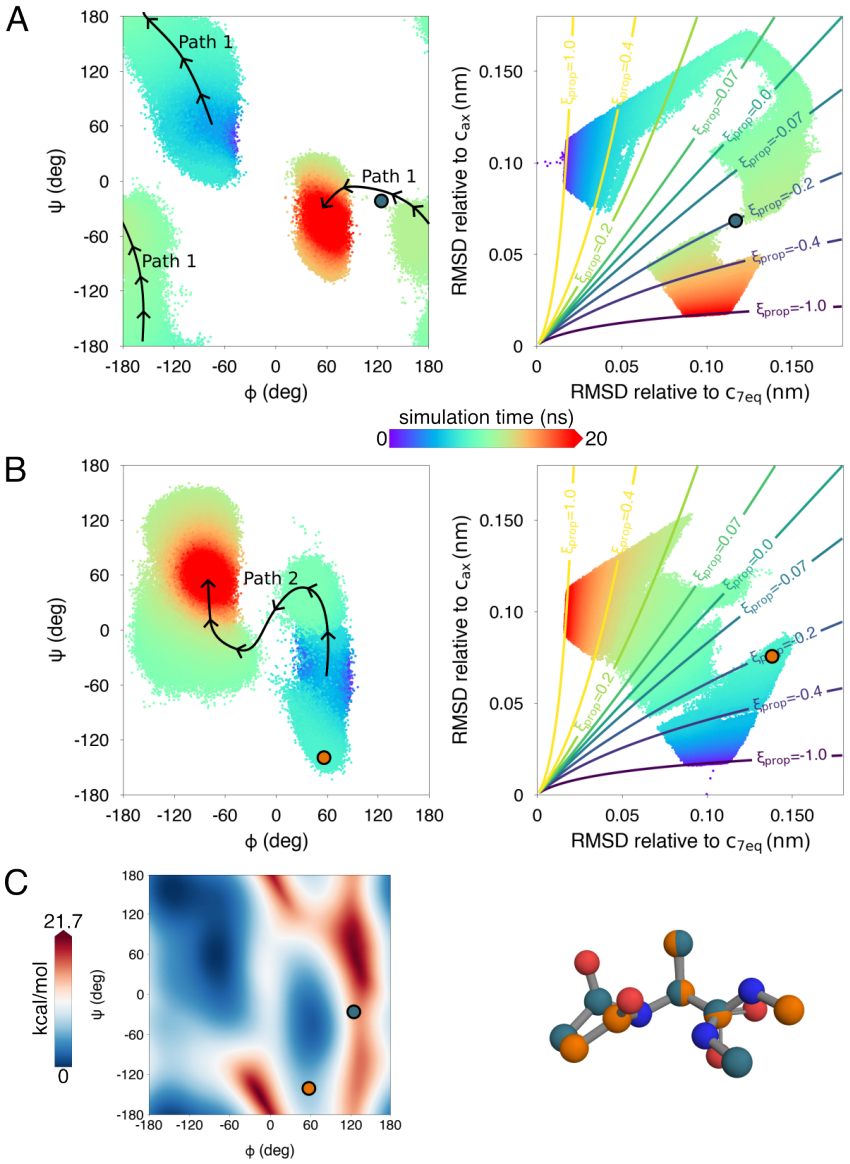

Figure 3.3: Pulling simulations along our proposed collective show memory effect - (A) Evolution of a pulling simulation from c_{7eq} (ξ_{prop} =1) to c_{ax} ($\xi_{prop} = -1$) in ϕ-ψ–space (left plot) and $\Delta(X, X_{c7eq})$-$\Delta(X, X_{cax})$–space (right plot). The blue point corresponds to configuration in (C) with blue carbons. (B) Evolution of a pulling simulation from c_{ax} ($\xi_{prop} = -1$) to c_{7eq} (ξ_{prop} =1) in ϕ-ψ–space (left plot) and $\Delta(X, X_{c7eq})$- $\Delta(X, X_{cax})$–space (right plot). The orange point corresponds to configuration in (C) with orange carbons. Different path are explored in A and B, there are denoted path 1 and path 2 for A and B respectively. (C) Two configurations corresponding to $\xi_{prop} = -0.21$ are mapped on our reference PMF, blue and orange points correspond to configuration on the right side with blue and orange carbons respectively.

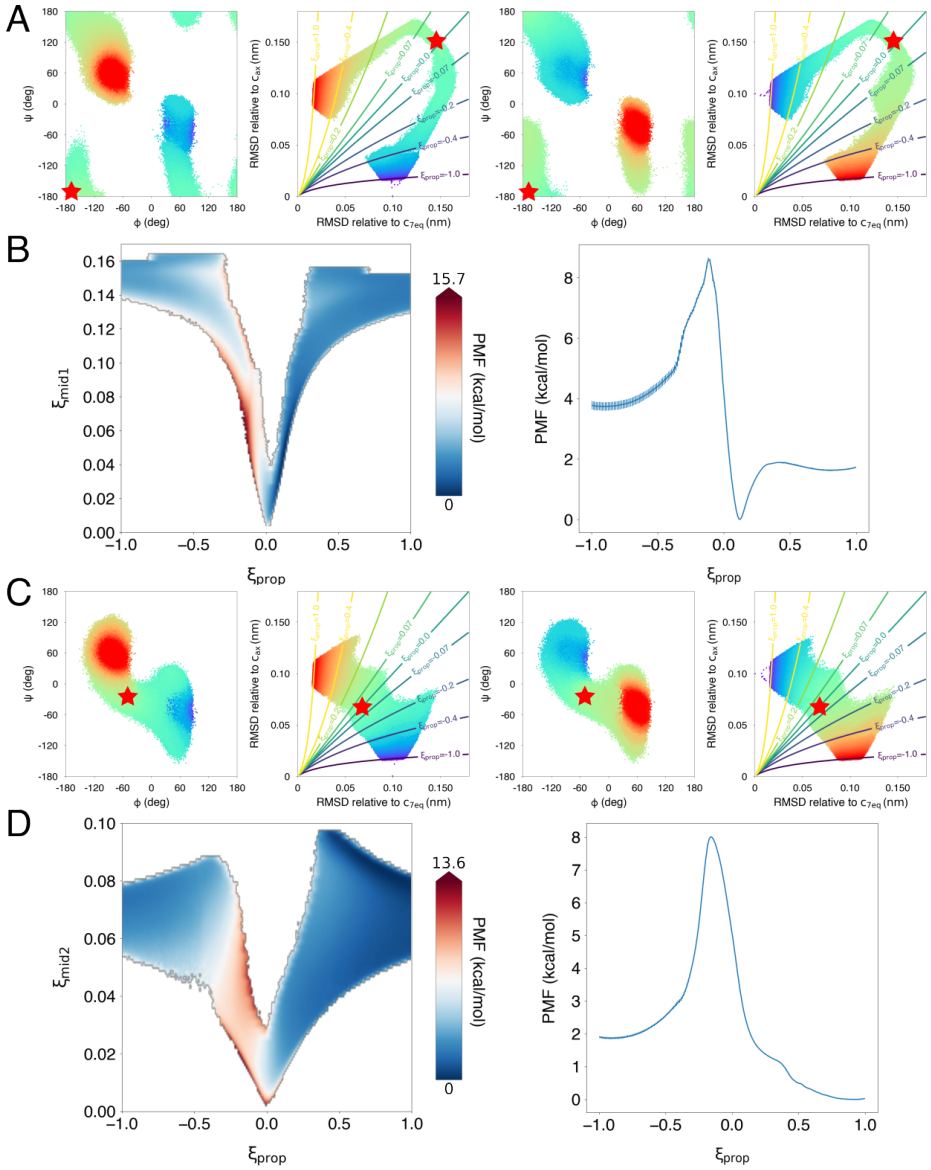

Figure 3.4: Pulling simulations along ξ_{prop} and ξ_{mid} does not exhibit hysteresis effect - (A) Pulling simulations along ξ_{prop} and ξ_{mid1} follow Path 1, regardless of the initial configuration. Blue and red color points depict early and late simulation time points respectively. Red stars represent the configuration used to define ξ_{mid1}. (B) 2D PMF along ξ_{prop} and ξ_{mid1}. This 2D PMF is projected onto ξ_{prop} on the right. (C) Pulling simulations along ξ_{prop} and ξ_{mid2} follow Path 2, regardless of the initial configuration. Blue points and red points correspond to frames close to $t = 0\,\text{ns}$ or $t = 20\,\text{ns}$ respectively. Red stars represent the configuration used to define ξ_{mid2}. (D) 2D PMF along ξ_{prop} and ξ_{mid2}. This 2D PMF is projected onto ξ_{prop} on the right.

4

Driving DNA opening during transcription initiation by RNA polymerase II with atomistic MD simulations

Content of this chapter has been published (145). References to videos have been removed as they are not crucial for understanding our work, but they can be accessed online.

4.1 Introduction

Transcription of DNA to RNA is catalyzed by RNA polymerases (RNAPs), a cornerstone of the central dogma of molecular biology (3). In eukaryotes, RNAP II carries out the synthesis of coding RNAs and of many non-coding RNAs. Transcription involves three main steps: initiation, elongation and termination. To trigger initiation, the 12-subunits RNAP II first assembles with general transcription factors to form the pre-initiation complex (PIC) (63). Within the 12 RNAP II subunits, RNA polymerase subunits 1 and 2 (RPB1 and RPB2, respectively, Fig. 4.1A) form the cleft and the active site. Several loops protrude from the two large subunits (Fig. 4.1A), which are well conserved among eukaryotes, including the rudder (in RPB1), fork loop 1 (FL1, in RPB2), and fork loop 2 (FL2, in RPB2) (146, 147). During initiation, these loops are in proximity with the DNA

as the transcription bubble forms. The architecture of RNAP II and the mechanism of transcription initiation have been described in several excellent reviews (33, 148).

Structural studies provided snapshots of the two end states of the PIC during transcription initiation in eukaryotes (63, 67, 149–156): snapshots of the closed complex (CC), in which DNA is double-stranded and located on top of the RNAP II cleft, and of the open complex (OC), in which the transcription bubble has formed and is loaded into the active site (Fig. 4.1A and 4.1B). While a cryo-EM structural study of the bacterial RNAP also revealed intermediate states of DNA opening (157), atomic details of the DNA opening pathway during transcription initiation in eukaryotes are missing. Consequently, the roles of conserved amino acid motifs of the rudder and of FL1 and FL2 during transcription initiation are largely unclear.

Previous molecular dynamics (MD) simulations focused on the elongation step of transcription (158–166) and on the clamp dynamics during initiation in bacterial RNAP (167). A recent coarse-grained MD study addressed DNA melting by inserting DNA base mismatches (168). However, DNA opening has not been simulated with atomistic models or without DNA base mismatches.

In this work, we used MD simulations to obtain a continuous opening transition from the CC to the OC in atomic detail. Because the CC-to-OC transition involves conformational rearrangements on the scale of several nanometers, obtaining such transition by brute-force MD simulations is computationally prohibitive. Therefore, we used steered MD simulations (98, 99) along a set of collective variables (CVs) to drive DNA opening and to enhance the sampling along the DNA opening pathway. Our CC-to-OC simulation provides insight into the spatial rearrangements of the DNA and of the protein loops during initiation, and they reveal extensive polar interactions of the DNA with the rudder, FL1, and FL2. These observed interactions suggest roles of the protein loops in supporting DNA strand separation and in stabilizing the transcription bubble in the OC.

4.2 Results

Steering a 55 Å-conformational transition with a combination of collective variables. Upon forming the transcription bubble, DNA carries out a transition involving a rotation of the DNA double strand by ∼370° as well as a translation of the

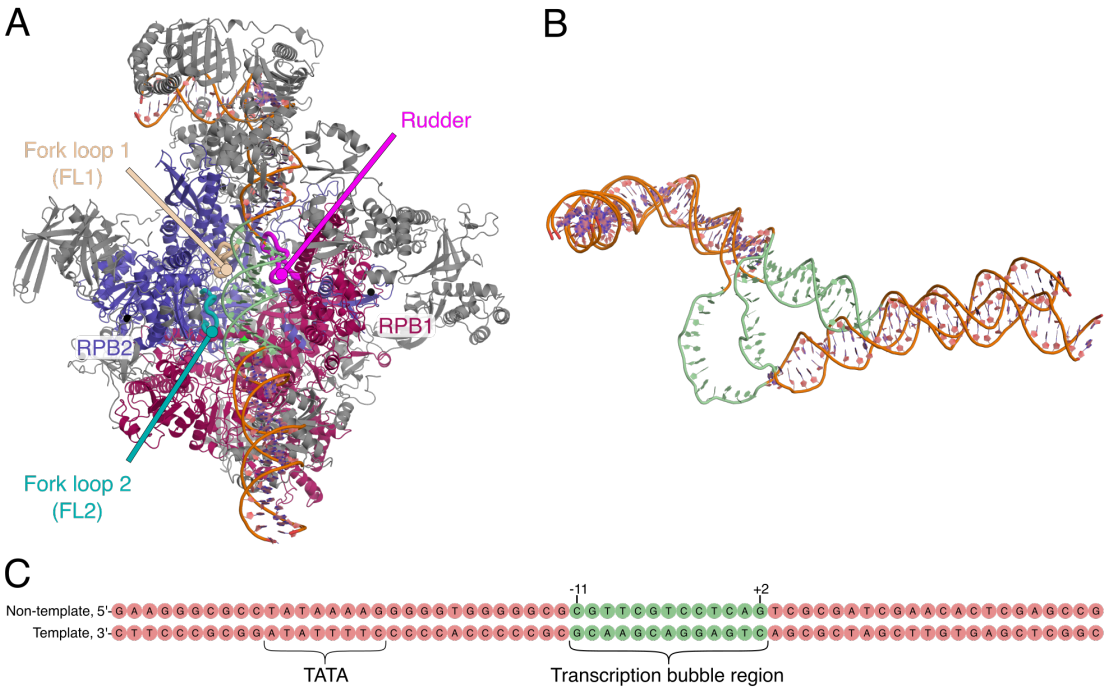

Figure 4.1: PIC complex in CC and overlap of DNA in CC and OC - (A) Cryo-EM structure of the CC without TFIIH and TFIIS (pdb code 5IY6 (151)). Zinc ions shown as black spheres. (B) Overlay of the DNA in CC and OC, taken from structures 5IY6 and 5IYB respectively (151). The DNA region involved in the DNA bubble formation is highlighted in green. (C) DNA sequence simulated in this work, corresponding to the DNA sequence found in 5IYB. DNA numbering according to Ref. (151), where +1 refers to the transcription start site in the OC structure.

DNA strands by up to 55 Å relative to the protein (151). Simulating such large-scale, nonlinear conformational transitions in atomic detail imposes considerable challenges. One possible strategy for favoring these large-scale motions is to introduce base mismatches between the two DNA strands, as used for obtaining the OC cryo-EM structure by He *et al.* (151) or used for favoring DNA melting in coarse-grained MD simulations (168). In contrast to these previous studies, we simulated DNA opening without base mismatches, according to the biologically relevant state of the CC (Fig. 4.1C). We obtained a relaxed pathway of DNA opening with a combination of two methods. First, we obtained an initial pathway using steered MD simulations along a combination of three collective variables (CVs); second, the initial pathway was relaxed using the path collective variable (PCV) method (169).

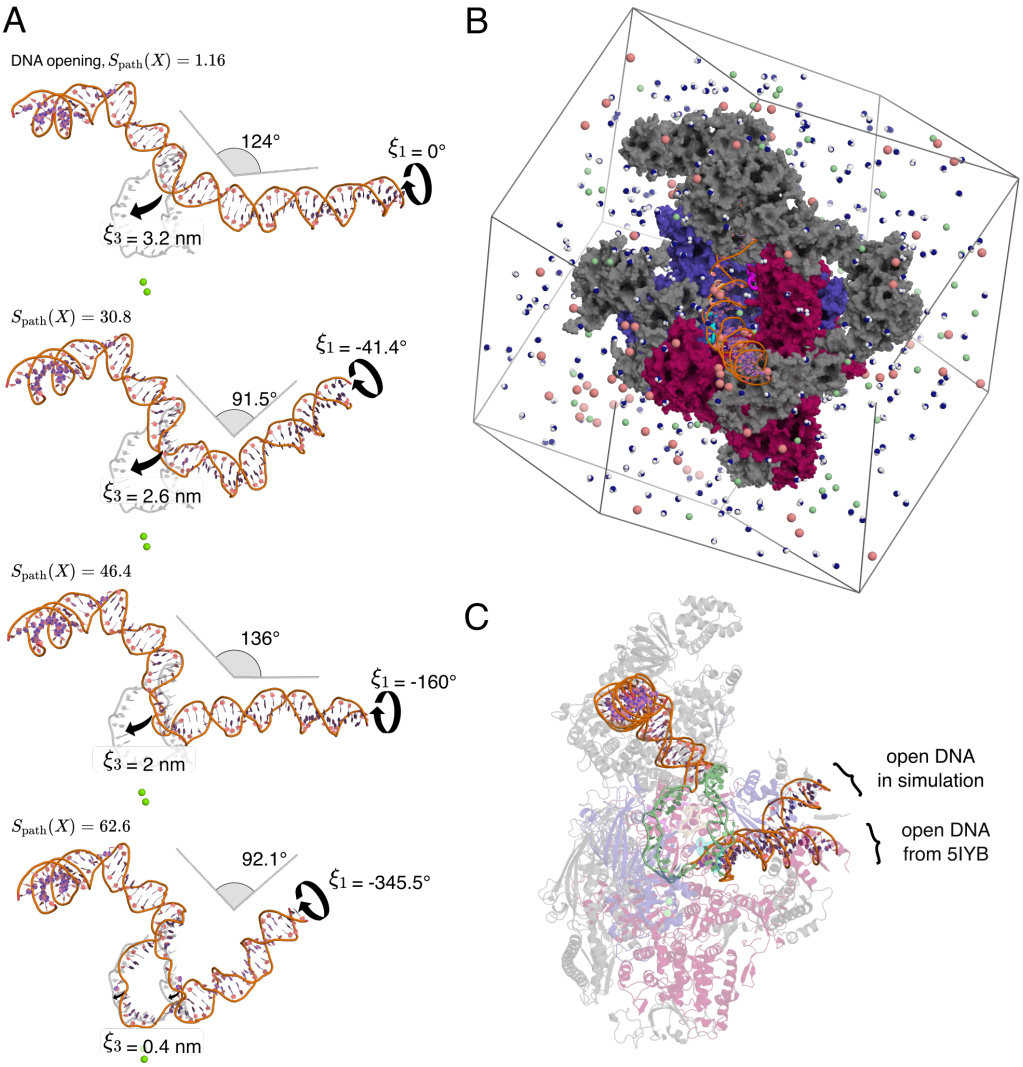

Figure 4.2: Transition from closed to open DNA in atomic detail - (A) Opening transition snapshots with corresponding S_{path} value (progression along the DNA opening), rotation of the downstream DNA helix (ξ_1), RMSD relative to the open bubble (ξ_3), and DNA helix bending angle. The target open bubble conformation is depicted in light gray. For reference, two catalytic magnesium ions are shown as green spheres. (B) Simulation box. Colors for the PIC in CC are consistent with Fig. 4.1. Water molecules, sodium ions, and chloride ions are colored in blue and white, pale pink, and pale green respectively. Most of water molecules and ions have been removed for clarity. (C) Section of the PIC in OC obtained from steered MD simulations. Open DNA from 5IYB structure is shown for reference. The transcription bubble is colored in pale green.

To guide the opening pathway, steered MD simulations were carried out along a combination of the following three CVs: (i) a rotational CV applied to the downstream DNA helix, thereby driving the melting of the DNA strand (Fig. 4.2A, ξ_1); (ii) two CVs given by the root mean-square distance (RMSD) of the sugar-phosphate backbone of the DNA relative to the conformation in the OC, taken from the 5IYB structure (ξ_2 and ξ_3) (151). Figure 4.2A illustrates the evolution of the rotational CV ξ_1 and of the RMSD-based CV ξ_3. By pulling along these CVs we obtained an initial path of DNA opening.

A path collective variable (PCV) for steering and relaxing the DNA opening path. Because our steered MD simulations were carried out on much shorter time scales compared to experimental time scales, it is reasonable to believe that the initial path is still biased by non-equilibrium effects. To relax the conformations along the opening pathway and, thereby, to mitigate such non-equilibrium effects, we applied the PCV method (169). Generally, PCVs are defined using two CVs: the position S_{path} along the initial path and the distance Z_{path} from the path, where the path is defined along N intermediate conformations (see Materials and Methods for details). In this study, the initial PCV was defined with 72 intermediate conformations taken form the steered MD simulation. Then, we carried out two rounds of constant-velocity pulling along S_{path}. Within each round, the path was allowed to relax, providing us with an updated set of increasingly relaxed intermediate conformation and, thereby, an updated PCV. The final PCV along 63 relaxed intermediate conformations, allows convenient opening simulations by pulling along the single S_{path}, instead of pulling along the three CVs used for obtaining the initial path (see above). In addition, projection onto the final S_{path} provides a convenient measure for the progress of the opening pathway, as used below in our figures and analysis.

Atomistic transition from the closed to a stable open DNA. By pulling along the aforementioned S_{path}, we obtained all-atom continuous trajectories of DNA opening from the CC to the OC (Fig. 4.2A/C). To test whether we have reached a state with a stable open DNA bubble, we simulated the final state without any biasing potential for 200 ns. In this simulation, the distances between the disrupted base pairs were reasonably stable (Fig. 4.3A, E), demonstrating that the strands did not re-anneal, as

expected for a stable OC. Three additional free simulations of 300 ns each corroborated the stability of the OC (Fig. 4.3B–D, F). The open DNA bubble exhibited a length of 12 base pairs (bp), in reasonable agreement with the length of 13 bp in the reference structure by He *et al.* (151). We quantified the spatial extension of the bubble with the average distance d_{bb} between the 12 disrupted base pairs. In the unbiased 200 ns simulation of the OC, we obtained $d_{\mathrm{bb}} = 2.36$ nm (Fig. 4.3), in reasonable agreement with the value of 2.67 nm in the reference structure. Minor structural differences relative to the reference OC are expected because (i) the DNA bubble is flexible and (ii) RNAP II in the OC accommodates various DNA bubble lengths and widths during initiation and, more generally, during the entire transcription process (170). Overall, the stability of the DNA transcription bubble in our free simulation implies that the previous pulling simulation represents a complete DNA opening pathway.

Figure 4.2A provides an atomic view on the large DNA rearrangements. First, due to the clockwise rotational motion carried out by the downstream DNA, the DNA became underwound in the transcription bubble region. Second, due to the translational motions induced on the transcription bubble towards the active site, the DNA bent at the transcription bubble region. These two topological changes of DNA led ultimately to the disruption of 12 bp. DNA rotational angles, bending angles and RMSD relative to the reference OC are depicted in Fig. 4.2A for four snapshots of our DNA opening trajectory. This interplay between negative supercoiling (clockwise rotational motion of DNA), DNA bending, and base pair disruptions have been reported previously in DNA minicircles (171–174). In addition, negative supercoiling has been shown to promote DNA opening during transcription with minimal transcription factors (68), further corroborating that our simulations reflect experimentally relevant conditions. Together, we obtained a simulation protocol that provides continuous atomistic transitions from the CC to a stable OC with an open DNA transcription bubble within computationally accessible simulation times. The protein–DNA contacts and interaction energies obtained from the simulations are discussed in the following sections.

Fork loop 2 tilts during DNA opening. Because DNA opening occurs inside the PIC, the DNA extensively interacts with protein domains and, in particular, with the protein loops. While DNA was loaded into the active site in the simulations, FL2 tilted into the transcription bubble, in-between the two DNA strands (Fig. 4.4A-D).

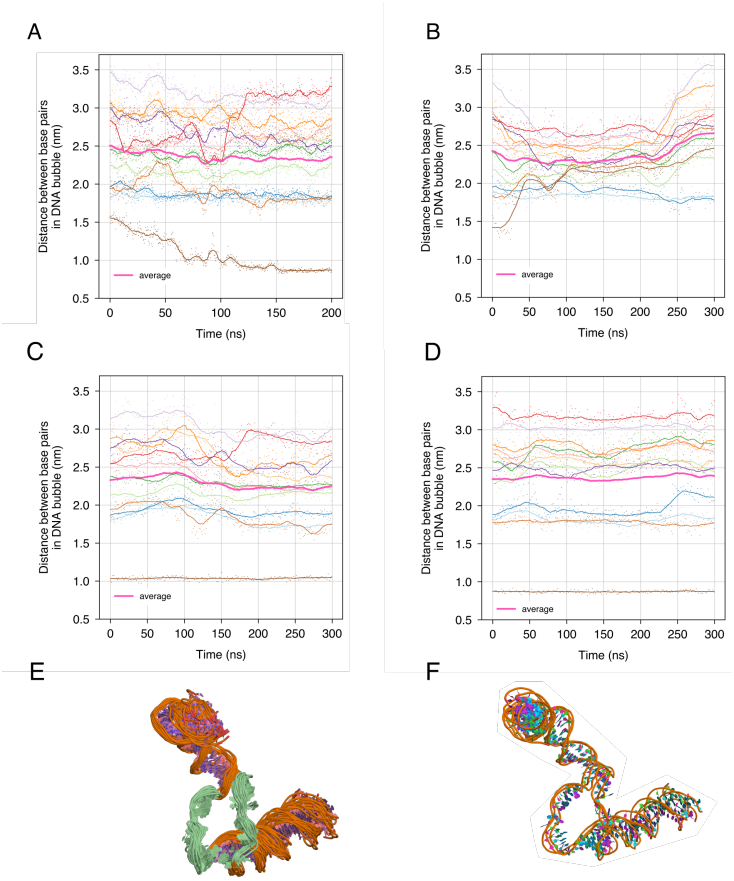

Figure 4.3: Analysis of the stability of the open DNA bubble in free simulation - (A) To test the stability of the DNA transcription bubble, a free simulation of 200 ns was carried out, starting from the final configuration of the pulling simulation along the final S_{path}. To quantify the stability of the DNA bubble, we monitored the DNA backbone COM distance between disrupted base pairs of the transcription bubble during the free simulation: base pair -11, -10, -9, -8, -7, -6, -5, -4, -3, -2, -1 and $+1$ are depicted in light blue, dark blue, light green, dark green, light pink, red, light orange, dark orange, light purple, dark purple, and maroon, respectively. The average distance between the 12 base pairs is depicted in dark pink. (B-D) Similar analysis as in panel (A) for three additional free simulations of the OC, started at $t = 0$ ns, $t = 100$ ns and $t = 190$ ns of trajectory obtained in (A). (E) DNA snapshots taken every 10 ns from the free simulation in (A). (F) Comparison of the last configurations obtained from (B), (C) and (D). Overall, distances between disrupted base pair and visual inspection of the trajectories show that the two strands do not re-anneal, suggesting that a DNA bubble was obtained that is stable at least on the microsecond time scale.

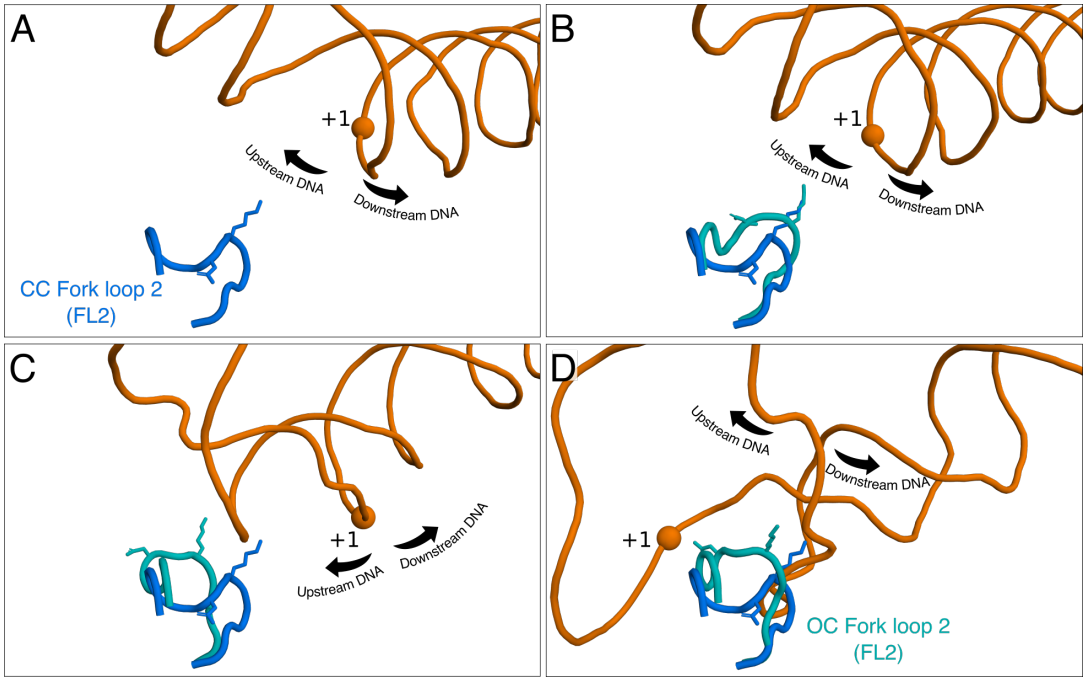

Figure 4.4: Tilting of sensor fork loop 2 (FL2) into the transcription bubble during DNA opening - (A-D) Snapshots of FL2 (cyan) during DNA opening corresponding to $S_{\text{path}} = 1.16$, 31.8, 51.5 and 62.6 respectively. The final tilted state of FL2 is depicted in panel D. For reference, the starting position of FL2 (marine blue), Asp-492 and Lys-494 are shown.

Whereas solvent-exposed protein loops are often flexible, the FL2 conformation pointing into the open bubble was remarkably stable, locked by electrostatic protein–DNA interactions, as observed in the free 200 ns simulation following the opening transition described above. The FL2 tilting in our simulations is compatible with a hypothesized role of FL2 as a sensor for the open transcription bubble (151). A recent study revealed a similar conformational change of FL2 during the transition from the CC to the OC, further supporting that FL2 is acting as a sensor for DNA opening (156).

Fork loops 1 and 2 support DNA opening by hydrogen bond attack on Watson-Crick pairs. The simulations revealed how DNA opening is supported by the rearrangement of hydrogen bonds (H-bonds) between DNA, protein, and water, as shown in Fig. 4.5. Namely, the loss of 31 Watson-Crick (WC) DNA–DNA H-bonds (Fig. 4.5A, orange curve) was predominantly compensated by the formation of approximately

40 DNA–Water H-Bonds (Fig. 4.5A, blue curve). The open DNA was further stabilized by the formation of approx. 12 DNA–protein H-bonds (Fig. 4.5A, green curve), among which ∼50% formed with the Watson-Crick edge (Fig. 4.5A, red curve), and the other ∼50% formed with other edges or with the DNA backbone.

The progression of WC H-bonds (Fig. 4.5A, orange curve), together with visual inspection of the MD trajectories, revealed three key protein residues involved in destabilizing the double-stranded DNA by attacking the Watson-Crick H-bonds. These events are reflected by marked decreases of the number of WC H-bonds at $S_{\mathrm{path}} = 45.9$, 49.6, and 55.4 (Fig. 4.5A, vertical lines) and are visualized in the molecular representations of Figs. 4.5B–D. First, the side chain of Gln-461 of FL1 interacted with the base pairs at −6, thereby competing with the WC base pairing (Fig. 4.5B, $S_{\mathrm{path}} = 45.9$). Second, the side chain of Lys-458 of FL1 interacted via H-bonds with base pairs at position −9 and −10 (Fig. 4.5C, $S_{\mathrm{path}} = 49.6$). Third, at a later stage of the opening process and after FL2 tilted into the open bubble, Arg-491 of FL2 destabilizes WC H-bonds, thereby promoting the unzipping of the double-stranded DNA (Fig. 4.5D, $S_{\mathrm{path}} = 55.4$). Hence, DNA–protein interactions do not merely serve as a compensation for the loss of DNA–DNA interactions in order to energetically stabilize the final OC state, but they might also catalyze the rupture of the WC base pairing.

The DNA dynamics in RNAP II driven by FL1 and FL2 are not unique but instead resemble dynamics observed in other DNA-interacting enzymes. For instance, base flipping has been suggested as an early mechanistic stage for DNA opening in a bacterial promoter (176). Likewise, H-bond attack to WC base pairs has been proposed for the cytosine 5-methyltransferase, where the enzyme infiltrates the DNA helix by forming H-bonds with nucleic bases, consequently destabilizing WC H-bonds and inducing base flipping (177, 178). Similarly, a base flipping event at position −6 of the template strand occurred during DNA opening in our simulation. Here, base flipping was promoted by the aforementioned Gln-461, via disruption of the WC H-bonds during the DNA opening (Fig. 4.5B, black arrow).

To get additional insights into the role of DNA–Protein interactions during DNA opening, and to identify selection pressure on the three key residues mentioned above, we analyzed the residue conservation of FL1 and FL2 among six eukaryotic organisms by means of multiple sequence alignments. Overall, FL1 and FL2 are strongly conserved among eukaryotes demonstrating their critical biological roles (Fig. 4.5E). However,

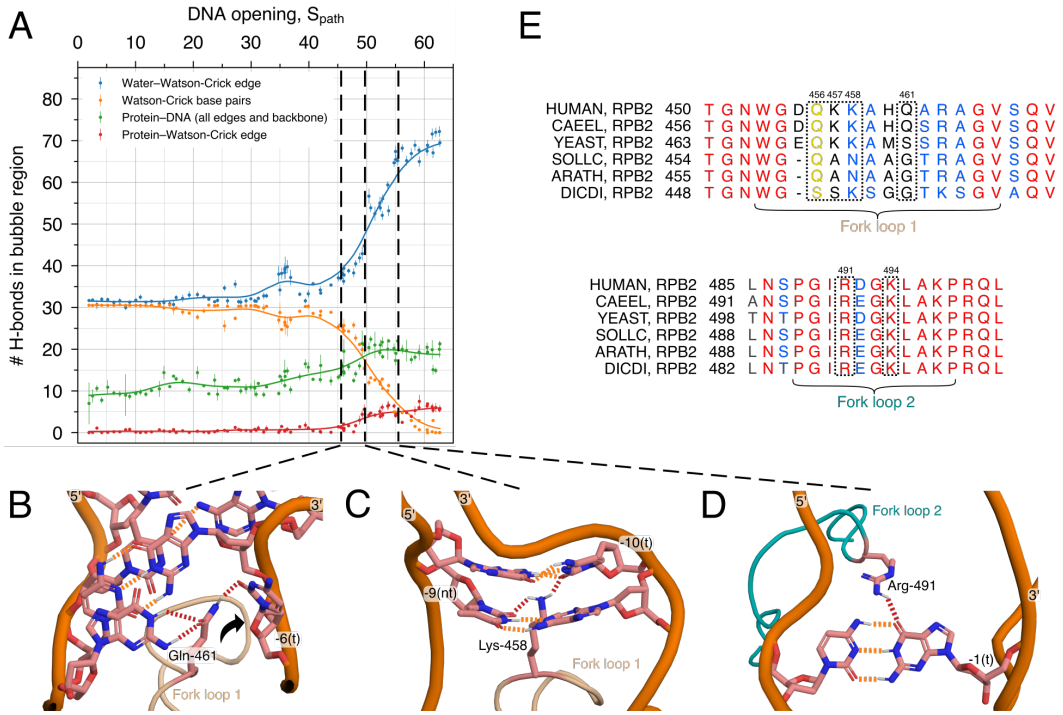

Figure 4.5: Rupture of Watson-Crick H-bonds in the transcription bubble and formation of DNA–protein and DNA–water H-bonds during the DNA opening process - (A) Development of the H-bonds of the transcription bubble region: number of H-bonds between WC edge and water (blue), between base pairs (orange), between DNA and protein (green), and between DNA WC edge and protein (red). Smooth lines are shown to guide the eye. (B) WC H-bond disruption driven by Gln-461 and base flipping (black arrow) of DNA residue −6 in the template strand. (C–D) Attack of WC H-bond by fork loops 1 and 2, respectively. (E) Sequence alignment (CLUSTAL W (175)) of fork loop 1 and 2 from six different eukaryotes: *Homo sapiens*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Solanum lycopersicum*, *Arabidopsis thaliana* and *Dictyostelium discoideum*. The residue color highlights the degree of conservation: invariant residues (red), residues with similar properties (blue), or with weakly similar properties (yellow). Residues in dotted boxes are discussed in the text.

Gln-461 is not conserved among eukaryotes, suggesting that the DNA–Gln-461 interactions observed in our simulations is either not critical for DNA opening or may be replaced with other interactions. In contrast, Lys-458 is well conserved among the analyzed eukaryotes (Fig. 4.5E); we hypothesize that the substitutions with Asn in *Solanum lycopersicum* and *Arabidopsis thaliana* may interact with DNA similar to Lys, thus supporting the role of residue 458 in destabilizing the double-stranded DNA. Arg-491 is

invariant among all the eukaryotic organisms chosen here (Fig. 4.5E), underlining its biological relevance in DNA strand separation. Taken together, our data suggest that residues of FL1 and FL2 catalyze DNA opening by attacking WC H-bonds between double-stranded DNA, providing a rationale for the marked sequence conservation of FL1 and FL2. According to the simulations and the sequence alignment, the conserved residues Lys-458 and Arg-491 are involved in H-bond attack; however, we cannot exclude the possibility that other conserved residues play similar roles.

Fork loops–DNA and rudder–DNA electrostatic interactions stabilize the open DNA conformation. To rationalize the energetic driving forces for DNA opening, we monitored the potential energy from DNA–DNA, DNA–Protein, and DNA–Water interactions (Fig. 4.6A). Here, potential energies were taken as the sum of Lennard-Jones and short-range Coulomb interactions, averaged over 50 ns of simulation and normalized relative to the state of the CC. The loss of interactions between the DNA strands is primarily compensated by a large gain of DNA–Protein interactions, as evident from the large negative DNA–Protein potential energies (Fig. 4.6A, orange and red). Although DNA opening leads to an increase of DNA–water interactions, as expected from the formation of H-bonds between water and the WC edge (Fig. 4.5A, blue), DNA–water interactions (Fig. 4.6A, blue) play a much smaller role as compared to DNA–Protein interactions (Fig. 4.6A, red).

To quantify which type of DNA–Protein interactions drive DNA opening, we further analyzed the number of contacts of the DNA bubble region with different groups of amino acids of common physicochemical properties (Fig. 4.6B). Evidently, the DNA forms ∼50 new contacts with basic protein residues, far more compared to contacts with polar or acidic residues. This finding reflects that RNAP II cleft is highly positively charged which helps to attract the negatively charged DNA backbone deeper into the cleft and, in particular, into the active site. This finding demonstrates, not surprisingly, that electrostatic interactions between the DNA and RNAP II are the key energetic driver for transcription bubble formation.

Visual inspections of the simulations revealed reoccurring salt bridges and hydrogen bonds between the protein and the open DNA. Gln-456 and Lys-457 (in FL1) form H-bonds with the template strand of DNA (Fig. 4.7A), suggesting that FL1 stabilizes the open bubble by compensating the loss of H-bonds between the two DNA strands and by

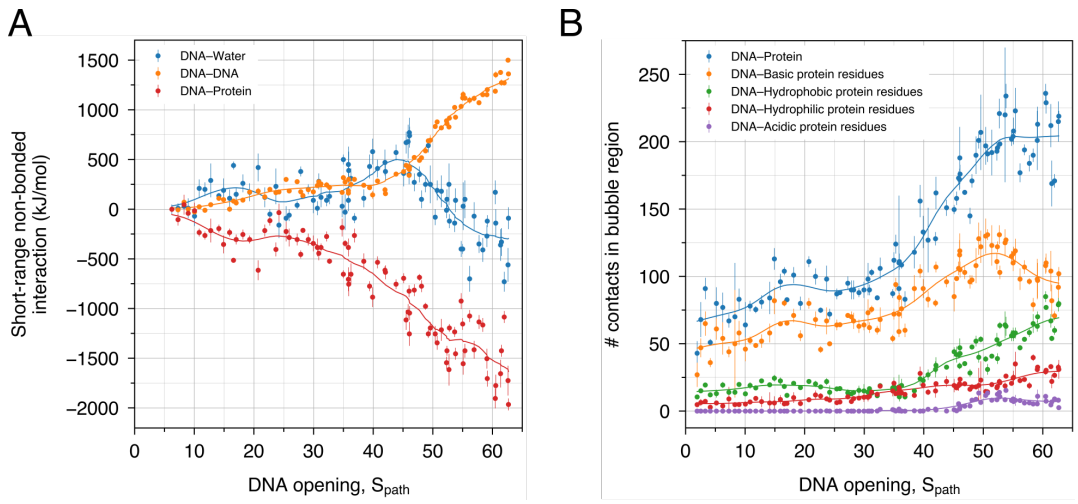

Figure 4.6: Electrostatic interactions support DNA opening - (A) Coulomb and Lennard-Jones short-ranged interactions between DNA and water (blue), between DNA and DNA (orange) and between DNA and protein (red). (B) Number of contacts within the DNA bubble region with all protein residue types (blue), with basic protein residues (orange), with hydrophobic protein residues (green), with hydrophilic protein residues (red) and with acidic protein residues (purple). Smooth lines are shown to guide the eye.

imposing a steric obstacle against strands re-annealing. In close proximity to FL1, Arg-327 (in the rudder) forms a salt bridge with the template strand (Fig. 4.7B). Likewise, Lys-494 (in FL2) and Arg-222 (in RPB2) interact with the non-template strand via salt bridges and H-bonds (Fig. 4.7C). Arg-222, Lys-413 and Arg-416 are examples of residues outside the fork loops or the rudder that form electrostatic interactions with the non-template strand of the open DNA conformation (Fig. 4.7D).

To corroborate the relevance of the DNA–protein interactions observed in our simulations for stabilizing the transcription bubble, we inspected the conservation of the residues mentioned above with sequence alignments. Accordingly, Lys-494 (Fig. 4.5E, FL2) and Arg-222 (Fig. 4.7E, RPB2) are invariant while Lys-413 (Fig. 4.7E, RPB2) is well conserved, supporting their biological relevance. Gln-456 (Fig. 4.5E, FL1) and Arg-416 (Fig. 4.7E, RPB2) are largely conserved except in *Dictyostelium discoideum* and in *Saccharomyces cerevisiae* respectively, underlining their putative role in stabilizing the transcription bubble. In contrast Arg-327 (Fig. 4.7E, rudder) is not conserved but may be replaced with Thr or Gln; however, all those residues are capable of forming H-bonds with the DNA backbone and, thereby may all stabilize the open bubble. Finally, Lys-457

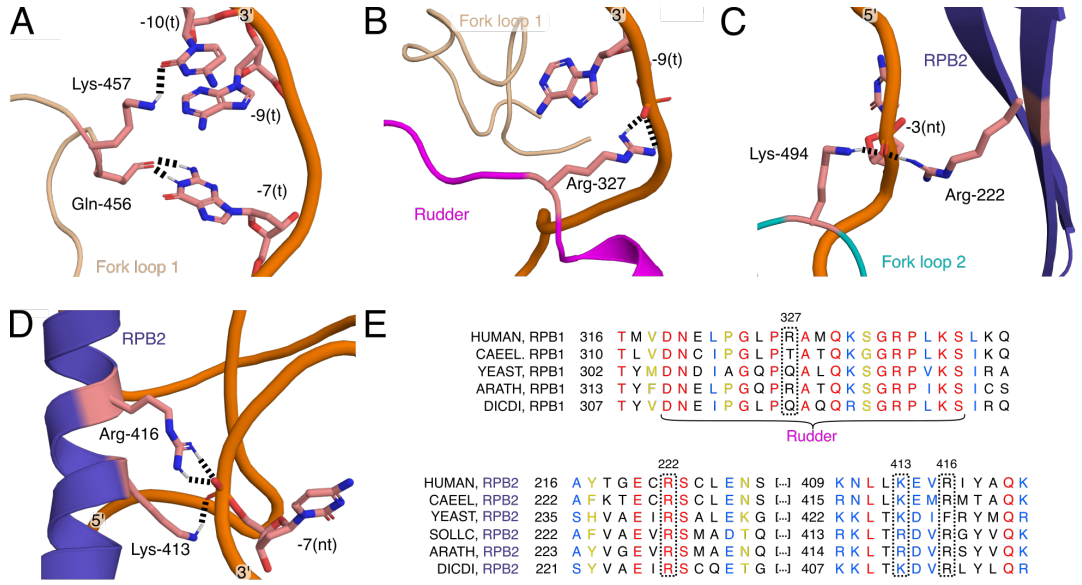

Figure 4.7: Electrostatic interactions between DNA and PIC stabilize open DNA in the OC - (A-D) H-bonds between protein and DNA. (B-D) Salt bridges between cationic residues with the anionic DNA backbone. (E) Sequence alignment (CLUSTAL W) of the rudder from different eukaryote organisms. Red residues are invariant, blue are from groups of strongly similar property and yellow from groups of weakly similar properties.

(Fig. 4.5E, FL1) is not conserved suggesting that this residue is less critical for stabilizing the transcription bubble. Together, our simulations show extensive interactions between the PIC and DNA that stabilize the DNA bubble. In the light of the sequence alignments, many of these interactions are critical, whereas some may be replaced by other interaction.

4.3 Discussion

We have presented the first all-atom simulation of a continuous DNA opening transition within human RNAP II. The simulations revealed extensive electrostatic and polar interactions of the DNA with the protein, predominantly with the two fork loops and with the rudder. Closer inspection of these interactions suggested that the rudder and the two fork loops are involved in (i) the separation of the two DNA strands by means of H-bond attack to WC base pairs and (ii) in maintaining the open DNA conformation by a combination of steric hindrance and electrostatic interactions. The biological relevance

of the protein residues involved in the observed interactions was further scrutinized by analyzing their conservation among eukaryotic amino acid sequences. Finally, we observed a base flipping event as well as the flipping of FL2 into the transcription bubble, in line with previous experiments (151, 156, 176–178).

Mutagenesis experiments targeting the fork loops and the rudder have been carried out in archaeal RNAP II and have revealed that the rudder helps in stabilizing the melted DNA in the OC (179). In addition, these experiments suggested that FL2 and, in particular Arg-451 —the archaeal equivalent of Arg-491 mentioned in this work—, plays a role in unwinding downstream DNA during elongation. However, mutagenesis of FL1 did not impact DNA opening in archaeal RNAP II in this study. Whereas archaeal and human RNAP II exhibit high sequence conservation, the physiological temperature in which DNA opening occurs may strongly differ, which might influence DNA melting. Indeed, a temperature of 70°C was used in the permanganate footprinting experiment by Naji *et al.* compared to a temperature of 37°C expected for human physiological conditions (179). The same kind of mutagenesis experiments in eukaryotic RNAP II system, ideally for human RNAP II, would be highly interesting to confirm that FL1, FL2, and the rudder are essential for DNA opening.

Simulating and relaxing such a large-scale roto-translational conformational transition with atomic MD force fields is computationally challenging. A recent coarse-grained MD study introduced base pair mismatches in the transcription bubble region to favor DNA opening (168). However, if DNA melting occurs without simultaneous rotation of the downstream DNA, high DNA strains in the upstream or downstream DNA region emerge, which is incompatible with the open DNA conformation from experimental structures (150, 151, 180). Therefore, in this work, we used steered MD simulations along a combination of three CVs to guide the large-scale displacement of DNA by up to 55 Å simultaneously with DNA rotation by $\sim 346°$. Finding a suitable set of CVs for obtaining a stable OC without undesired DNA melting outside the transcription bubble required extensive optimization and human supervision; hence, future studies may aim towards more automated protocols for findings suitable sets of CVs. Having obtained an initial DNA opening simulation from steered MD, we used the PCV framework to relax the initial opening pathway. Notably, we tried to compute the potential of mean force along the PCV with umbrella sampling (96) or metadynamics (97), with the aim

to obtain an estimate for the free energy of DNA opening; however, we observed considerable hysteresis problems, suggesting that it is difficult to sample all the degrees of freedom orthogonal to the PCV such as all alternative protein–DNA interaction motifs. This observation further implies that our simulations provided a plausible pathway for DNA opening, but not necessarily the minimum free energy pathway. For instance, alternative pathways may involve sets of protein–DNA pair interactions in addition to the interactions presented in Figs. 4.5 and 4.6. To enable exhaustive conformational sampling and free energy calculations, future simulations may investigate the use of additional enhanced sampling techniques such as bias-exchange umbrella sampling (181) or the use of extensive compute power (182).

In eukaryotes, the transcription factor TFIIH catalyzes, both, DNA translocation and DNA opening by using the energy released by ATP hydrolysis (183–186). DNA opening can also be triggered by torsional stresses generated by negative supercoiling produced mostly by remote transcription processes; therefore, ATP is also the indirect energy source for DNA opening under negative supercoiling conditions (68, 187–192). However, it has been suggested that translocase activity is not necessary for RNA transcription and, thus, that ATP-independent DNA opening is achievable by RNAP II (67) with the use of binding energy generated from PIC assembly (66). In this study, we modeled DNA opening in absence of TFIIH with the use of rotational and translational CVs. However, since TFIIH also produces torsional stress to downstream DNA, our current protocol for DNA opening will be useful to study DNA opening in presence of TFIIH. Simulations with TFIIH will be particularly relevant to understand the role of its XBP subunit, the TFIIH subunit containing the translocase activity and the motor for DNA unwinding (185).

The transcription factor TFIIB contains the B-reader and B-linker elements, which also help DNA opening (152, 193). Because refined atomic models of the B-reader and B-linker were not resolved in the CC structure by He *et al.* (151), we simulated the CC-to-OC transition in the absence of these TFIIB elements. Further atomistic simulation will be needed to investigate whether the B-reader and B-linker support DNA opening using similar interaction motifs as observed here for FL1, FL2, and for the rudder.

To conclude, we obtained an atomic model for a continuous DNA opening event in the human PIC. The simulations revealed extensive interactions of the DNA with the protein, in particular with loops protruding into the polymerase cleft: FL1, FL2, and the

rudder. According to the simulations, the loops play multiple roles for DNA opening: (i) by attacking WC H-bonds, they may catalyze the melting of the DNA; (ii) extensive polar interactions via salt-bridges with the DNA backbone and, to a lower degree, via H-bonds with the DNA backbone and bases stabilize the open DNA conformation; (iii) FL2 tilted into the DNA bubble during opening, a conformational transition that is compatible with a function of FL2 as a sensor for an open transcription bubble.

4.4 Materials and Methods

Simulation setup. MD simulations were carried out with Gromacs (194) version 2020.2 patched with Plumed (142, 195) version 2.6.1, and with Gromacs version 2021 patched with Plumed version 2.7.0. The initial atomic coordinates for the CC were obtained from the Protein Data Bank (accession code 5IY6 (151)) from which we removed TFIIH and TFIIS. We used YASARA version 20.8.23 to add acetyl and N-methyl amide capping groups at the ends of the missing protein regions and at the C and N termini. We also used YASARA to add missing atoms (196). The system was solvated with TIP3P water molecules and Na/Cl counter ions were added to neutralize the system with a salt concentration of 100 mM (197). In total, the system contained 832078 atoms. The OL15 force field was used for the DNA (198). The ff14sb force field was used for the protein (199) except for the zinc(II)-coordinating Cis and His residues, for which the improved parameters by Macchiagodena *et al.* were used (200).

Electrostatic interactions were computed with the particle-mesh Ewald method (201), using a real-space cutoff at 1 nm and a Fourier spacing of 0.16 nm. Dispersion interactions and short-range repulsion were described with a Lennard-Jones potential with a cutoff at 1 nm. Bonds and angles of water were constrained with the SETTLE algorithm (123) and bonds involving other hydrogen atoms were constrained with LINCS (124). To remove atomic clashes, the system was energy minimized with the steepest-descent algorithm. We next equilibrated the system under NVT conditions for 100 ps at 300 K using the velocity-rescale thermostat with one heat bath for the coordinated ions, DNA and protein and another heat bath for water and counter-ions (90). Then, we equilibrated the system at 1 bar for 10 ns under NPT conditions using Parrinello-Rahman pressure coupling and using the same thermostat as in the NVT equilibration (92). During both

equilibration steps, all heavy atoms were position restrained with a force constant of $1000\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$.

To enable the use a 4 fs time-step for further pulling simulations we used hydrogen mass repartitioning (HMR) (89). Accordingly, to increase the oscillation period of the bond angles involving hydrogen atoms, the hydrogen masses were scaled up by a factor of f_{H} and the heavy atom masses connected to hydrogens were scaled down, while keeping the overall masses of chemical moieties constant. To choose a scaling factor f_{H} that yields stable simulations at a 4 fs time step, we tested scaling factors from 2 to 3 in steps of 0.2, where three simulations were carried out for each scaling factor. Each simulation was carried out for 20 ns in *NPT* conditions. None of the simulations with a scaling factor of 2.8 or 3 were stable, whereas all other simulations were stable. For production simulations, we decided to use $f_{\mathrm{H}} = 2.5$.

To exclude that HMR leads to excessive energy drift, we carried out three *NVE* simulations with $f_{\mathrm{H}} = 2.5$ using 4 fs time step for 500 ps and, for reference, three *NVE* simulation without HMR using a 2 fs time step. On average, we obtained an energy drift of $0.06\%\,\mathrm{ns^{-1}}$ with HMR, which was even smaller than the average value of $-0.13\%\,\mathrm{ns^{-1}}$ without HMR (4.1). Hence, integrating Newton's equations of motion with HMR models was numerically stable and exhibited only a marginal energy drift.

Total energy drift with HMR ($\%\,\mathrm{ns^{-1}}$)	Total energy drift without HMR ($\%\,\mathrm{ns^{-1}}$)
0.06	-0.12
0.06	-0.14
0.06	-0.12

Table 4.1: Drifts of the total simulation energy during three independent *NVE* simulations of 500 ps with HMR using a 4 fs time step (left column) or without HMR using a 2 fs time step (right column). Energy drifts are shown relative to the total energy of the simulation per nanosecond. Using HMR does not increase the energy drift.

Simulation of initial DNA opening pathway. We generated an initial path from the CC to the OC with a steered MD simulation of 175 ns using a combination of one rotational CV and two RMSD-based CVs.

As the first CV, we used a rotational CV defined as $\xi_1 = 1/4 \sum_{i=1}^{4} \gamma_i(X; X_0)$, where each dihedral angle γ_i was defined as:

$$\gamma_i(X; X_0) = \mathrm{dih}(\mathbf{v}_i(X_0), \mathbf{u}_{axis}, \mathbf{v}_i(X)) \tag{4.1}$$

Here, $\mathbf{u}_{\mathrm{axis}}$ denotes the helix axis of the DNA in region +3 to +23 (Fig. 4.8A). X_0 is the configuration at $t = 0\,\mathrm{ns}$, and X is the configuration at a later simulation time t. Further, $\mathbf{v}_i(X_0)$ denotes the vector connecting the two centers of mass (COMs) $\mathbf{c}_{i,1}(X_0)$ and $\mathbf{c}_{i,2}(X_0)$ at $t = 0\,\mathrm{ns}$ (Figs. 4.8B and 4.8D-F), and $\mathbf{v}_i(X)$ is the instantaneous vector connecting the same COMs at simulation time t (Figs. 4.8C and 4.8H). The two groups of atoms used to define $\mathbf{c}_{i,1}$ and $\mathbf{c}_{i,2}$, respectively, were constructed by splitting the helix DNA region +3 to +23 along the axis, as illustrated in Figs. 4.8B and 4.8D-F. The four $\mathbf{v}_i(X_0)$ defining ξ_1 are depicted in Fig. 4.8G.

As a second CV, we used $\xi_2 = \Delta(X, X_{\mathrm{OC1}})$. Here, X_{OC1} denotes the DNA backbone atoms of the OC in the region −17 to −5, taken from the 5IYB structure (151), and $\Delta(X, X_{\mathrm{OC1}})$ denotes the RMSD of the instantaneous structure X relative to the reference structure X_{OC1}. As a third CV, we used $\xi_3 = \Delta(X, X_{\mathrm{OC2}})$. Here, X_{OC2} denotes the backbone atoms of the OC in the region −17 to +2, again taken from the 5IYB structure.

During the 175 ns of steered MD simulation, we applied different forces on the three CVs described above:

1. The rotational CV (ξ_1) was pulled from 6.27 rad (close to 2π) to 0.01 rad over the first 100 ns using a force constant of 7000 $\mathrm{kJ\,mol^{-1}\,rad^{-2}}$. Over the next 4 ns, the force applied on ξ_1 was turned off by linearly decreasing the force constant from 7000 $\mathrm{kJ\,mol^{-1}\,rad^{-2}}$ to 0 $\mathrm{kJ\,mol^{-1}\,rad^{-2}}$.

2. The RMSD relative to X_{OC1} (ξ_2) was pulled from 2.35 nm to 0.4 nm over the first 50 ns using a force constant of 10,000 $\mathrm{kJ\,mol^{-1}\,nm^{-2}}$. Over the next 50 ns, ξ_2 was pulled from 0.4 nm to 0 nm using a force constant decreasing linearly from 10,000 $\mathrm{kJ\,mol^{-1}\,nm^{-2}}$ to 0 $\mathrm{kJ\,mol^{-1}\,nm^{-2}}$.

3. The RMSD relative to X_{OC2} (ξ_3) was pulled from 1.99 nm to 0 nm between simulation times of 50 ns and 100 ns, using a linearly increasing force constant between 20,000 $\mathrm{kJ\,mol^{-1}\,nm^{-2}}$ and 30,000 $\mathrm{kJ\,mol^{-1}\,nm^{-2}}$. Over the next 75 ns, ξ_3 was restrained at 0 nm using a force constant of 30,000 $\mathrm{kJ\,mol^{-1}\,nm^{-2}}$.

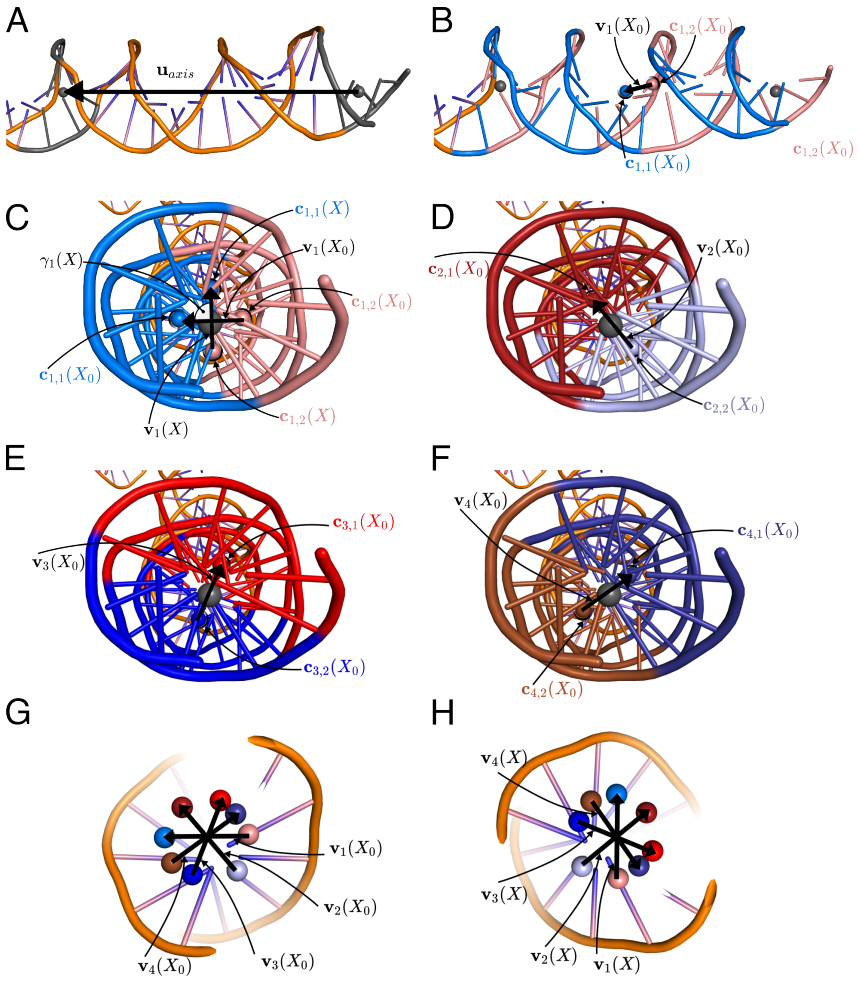

Figure 4.8: Illustration of the definition of the rotational CV ξ_1 - (A) Axis of the DNA helix used to define one vector of the dihedral angles in region +3 to +23. Residues colored in grey were used to compute the centers of geometry defining the axis (grey beads). (B-C) Blue and pink DNA segments highlight two horizontal halves of the DNA, where the blue bead at $\mathbf{c}_{1,1}(X_0)$ and the pink bead at $\mathbf{c}_{1,2}(X_0)$ are located at the center of geometry of the pink and blue halves, respectively. In panel (C), bead positions $\mathbf{c}_{1,1}(X)$ and $\mathbf{c}_{1,2}(X)$ that would lead to a rotation angle of $\gamma_1(X) = -90°$ are illustrated as well. (D-F) Definition of the three other reference vectors $\mathbf{v}_2(X_0)$, $\mathbf{v}_2(X_0)$, and $\mathbf{v}_4(X_0)$. Residues used for computing the COMs are colored accordingly to their associated COMs. (G) All COMs and their associated vectors used to define the rotational CV ξ_1 at $t = 0\,\text{ns}$. (H) Example of a $-90°$ rotation of the DNA helix compared to the reference configuration at $t = 0\,\text{ns}$ in panel (G).

To exclude that independent steered MD simulations would lead to qualitatively different opening pathways, five independent 175 ns simulations were carried out and analyzed in terms of number of hydrogen bonds, contacts, and interaction energies (Fig. 4.9). Whereas independent simulations exhibit different fluctuations of in the downstream DNA helix, the trends of the interactions and conformations of the DNA bubble are largely preserved.

Relaxation of the initial DNA opening pathway. In order to relax and sample intermediate states along the opening pathway, we first used constant-velocity pulling from the CC to the OC with a Path Collective Variable (PCV) (169). The two components of a PCV are S_{path} and Z_{path}, and are defined as:

$$S_{\text{path}}(X) = \frac{\sum_{i=1}^{N} i e^{-\lambda M_i(X)}}{\sum_{i=1}^{N} e^{-\lambda M_i(X)}} \tag{4.2}$$

$$Z_{\text{path}}(X) = -\frac{1}{\lambda} \ln \sum_{i=1}^{N} e^{-\lambda M_i(X)} \tag{4.3}$$

where the unitless S_{path} describes the progression along the path and Z_{path} describes the deviation from the path. N denotes to the number of reference configurations defining the DNA opening path. The distance metric $M_i(X) = \Delta^2(X, X_i)$ is the mean squared deviation (MSD) of the instantaneous configuration X relative to the reference configuration X_i. The choice of the N reference configurations was optimized to obtain similar MSDs between neighboring configurations and a good flatness of the surface spanned by the $N \times N$ MSD matrix (169). The symbol λ is the smoothing parameter, proportional to the inverse of the MSD between adjacent reference configurations.

To relax the initial path along Z_{path}, we performed two rounds of constant-velocity pulling along S_{path}, while applying a wall potential on Z_{path} acting above $Z_{\text{path}}(X) = 0.035\,\text{nm}^{-2}$. For the first relaxation round, we took $N = 72$ reference configurations from the initial path, set $\lambda = 21.7\,\text{nm}^{-2}$, and we carried out 100 ns of constant-velocity pulling along S_{path} from 1.1 to 71.3, corresponding to configurations close to the first or last reference configuration. We used a force constant of $5000\,\text{kJ mol}^{-1}$ for S_{path} and a force constant of $2.8 \times 10^6\,\text{kJ mol}^{-1}$ for Z_{path}. To build a new path for the following relaxation round satisfying the two criteria for reference selection mentioned above without discarding configurations in the OC, another 20 ns simulation was required with an

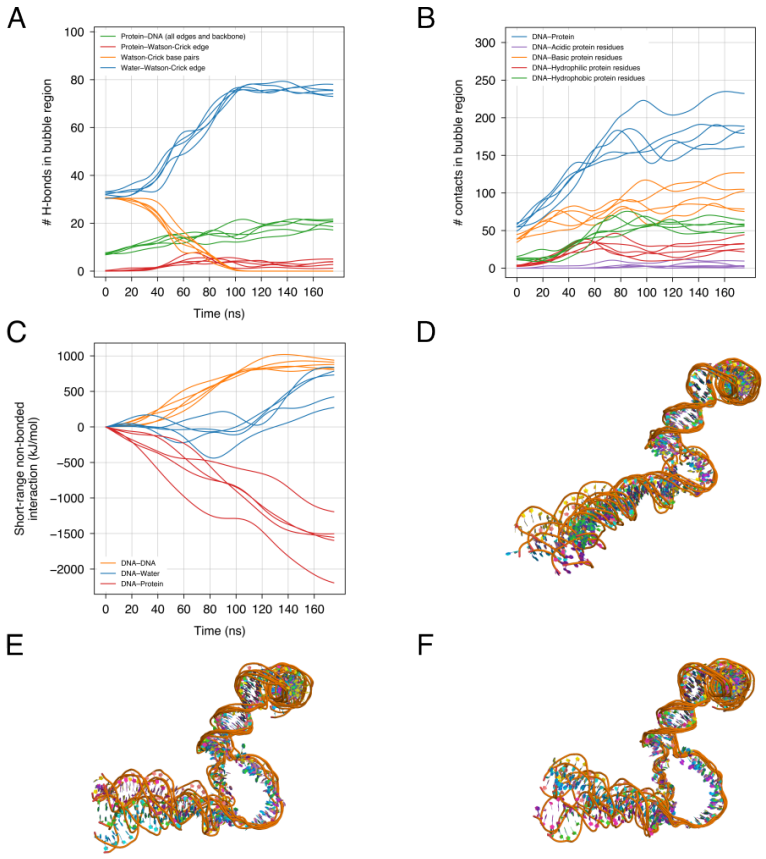

Figure 4.9: Comparison of five independent initial DNA opening simulations obtained by pulling along our combination of RMSD-based and rotational CVs - (A) Development of number of H-bonds in the transcription bubble region: number of H-bonds between WC edge and water (blue), between base pairs (orange), between DNA and protein (green), and between DNA WC edge and protein (red). (B) Number of contacts within the DNA bubble region with all protein residue types (blue), with basic protein residues (orange), with hydrophobic protein residues (green), with hydrophilic protein residues (red) and with acidic protein residues (purple). (C) Coulomb plus Lennard-Jones short-ranged interactions between DNA and water (blue), between DNA and DNA (orange) and between DNA and protein (red). To guide the eye, curves in panels (A–C) were smoothed with Gaussian filters with widths of 0.9 ns, 7 ns, and 14 ns, respectively. (D-F) Overlay of DNA configurations from five independent pulling simulations at time points (D) $t = 57$ ns, (E) $t = 95$ ns and (F) $t = 171$ ns. Protein and solvent is not shown for clarity. Overall, the five independent opening trajectories reveal similar trends in terms of interactions and conformations, albeit the downstream DNA helix exhibits substantial conformational flexibility.

harmonic restraint centered on $S_{\text{path}}(X) = 71.3$, with a force constant of $5000\,\text{kJ}\,\text{mol}^{-1}$ and keeping the wall potential acting above $Z_{\text{path}}(X) = 0.035\,\text{nm}^2$ with an offset of $0.005\,\text{nm}^2$ and a force constant of $2.8 \times 10^6\,\text{kJ}\,\text{mol}^{-1}\,\text{nm}^{-4}$. For the second relaxation round, we used $N = 91$ reference configurations from the first relaxation round, set $\lambda = 45.43\,\text{nm}^{-2}$, and we carried out 100 ns of constant-velocity pulling along S_{path} from 1.15 to 90.85 using a force constant of $4000\,\text{kJ}\,\text{mol}^{-1}$ for S_{path} and a force constant of $4 \times 10^6\,\text{kJ}\,\text{mol}^{-1}$ for Z_{path}. For the same reasons as for the first relaxation round, we extended the second relaxation simulation for another 20 ns with an harmonic restraint centered on $S_{\text{path}}(X) = 90.85$, with a force constant of $4000\,\text{kJ}\,\text{mol}^{-1}$ and keeping the wall potential acting above $Z_{\text{path}}(X) = 0.035\,\text{nm}^2$ with an offset of $0.005\,\text{nm}^2$ and a force constant of $4 \times 10^6\,\text{kJ}\,\text{mol}^{-1}\,\text{nm}^{-4}$.

Sampling the final DNA opening path. From the second relaxation round of constant-velocity pulling described above, we selected $N = 63$ reference configurations to define our final PCV. This final PCV was set up with $\lambda = 36\,\text{nm}^{-2}$. To sample DNA and loop conformations along the DNA opening path, we extracted 90 configurations from the second round of constant-velocity pulling with S_{path} ranging from 1.250 to 62.9. Each of those 90 configurations was used to start a simulation of 50 ns restrained on a particular value of S_{path} with an harmonic potential.

The reference positions and the force constants of the harmonic potentials of these 90 simulations are shown in Figure 4.10. These 90 simulations were used to characterize the opening path in terms of H-bonds, potential energies, and atomic contacts (Figs. 4.5A and 4.6).

Simulation analysis. Hydrogen bonds were defined with a cutoff distance of 0.35 nm between the hydrogen atom and the H-bond acceptor, and with a cutoff angle hydrogen–donor–acceptor of 30°. The base pairs +2 and +3 exhibited disrupted hydrogen bonds but were mismatched with other bases in the downstream DNA fork; hence, during the analysis, we did not consider these bases as part of the open DNA bubble. Contacts were defined with a cutoff distance of 0.3 nm. The potential energies were computed as the average of the sum of Lennard-Jones and short-range Coulomb interactions with a cutoff at 1 nm. Simulation trajectories were visualized with PyMOL (202) and VMD

S_{path}	κ ($kJ\,mol^{-1}$)	S_{path}	κ ($kJ\,mol^{-1}$)	S_{path}	κ ($kJ\,mol^{-1}$)	S_{path}	κ ($kJ\,mol^{-1}$)	S_{path}	κ ($kJ\,mol^{-1}$)
1.250	10	19.148	10	34.100	20	46.000	20	54.945	10
2.244	10	20.143	10	34.600	20	46.840	20	55.250	20
3.239	10	21.137	10	35.025	20	46.990	10	55.940	10
4.233	10	22.131	10	35.058	10	47.680	20	56.000	20
5.227	10	23.126	10	35.950	20	47.985	10	56.750	20
6.222	10	24.120	10	36.052	10	48.520	20	56.934	10
7.216	10	25.115	10	36.875	20	48.979	10	57.500	20
8.210	10	26.109	10	37.047	10	49.360	20	57.928	10
9.205	10	27.103	10	37.800	20	49.973	10	58.923	10
10.199	10	27.200	20	38.041	10	50.200	20	59.917	10
11.194	10	27.900	20	39.035	10	50.550	20	60.000	20
12.188	10	28.098	10	40.030	10	50.968	10	60.911	10
13.182	10	29.092	10	41.024	10	51.962	10	60.933	20
14.177	10	30.086	10	42.019	10	52.956	10	61.867	20
15.171	10	31.081	10	43.013	10	53.000	20	61.906	10
16.165	10	32.075	10	44.007	10	53.750	20	62.300	8
17.160	10	33.069	10	45.002	10	53.951	10	62.800	20
18.154	10	34.064	10	45.996	10	54.500	20	62.900	10

Figure 4.10: Tables of harmonic potential centers S_{path} and force constants κ used to sample the DNA opening path

(203). Images were generated with PyMOL.

Conformational stability of the OC. To test the stability of the OC obtained from the second relaxation round, we performed a free simulation of the OC, i.e., without any biasing potential. To this end, a 200 ns simulation was started from the final configuration of the second relaxation round. We quantified the stability of the OC by monitoring the distances between the 12 disrupted base pairs of the transcription bubble (Fig. 4.3). The 12 distances were computed with the center of geometries of the heavy backbone atoms of the two complementary nucleotides. The 12 distances were then averaged in each time frame. For reference, the same protocol was used to compute the average distances between the 13 dissociated base pairs in the reference OC (5IYB (151)).

Multiple sequence alignments. RPB1 and RPB2 protein sequences were taken from the UniProtKB/Swiss-Prot database (204). The organisms were chosen to cover different kingdoms of the eukaryotic domain. The alignments were carried out with CLUSTAL W version 2.1 (175).

5

Comparing umbrella sampling methods applied on the permeation of fosmidomycin through bacterial outer membrane porin OprO

In order to find enhanced sampling techniques that could help solving hysteresis effects experienced when computing PMFs for DNA opening, we explored (205) how Hamiltonian-replica exchange (119, 120) umbrella sampling, simulated tempering-enhanced umbrella sampling (STeUS) (121, 206), and replica-exchange umbrella sampling (108) perform in computing PMFs compared to standalone umbrella sampling (96).

5.1 Introduction

Pseudomonas aeruginosa bacteria are Gram-negative bacteria, which differentiate from Gram-positive bacteria by the presence of an outer membrane and a thinner peptidoglycan layer 5.1. *Pseudomonas aeruginosa* is an opportunistic pathogen, meaning that it is usually not harmful to healthy individuals, but can infect and cause disease in hosts with defective immune system or already weakened by other diseases. For this reason, infections by *Pseudomonas aeruginosa* are frequent in hospitals, accordingly they are

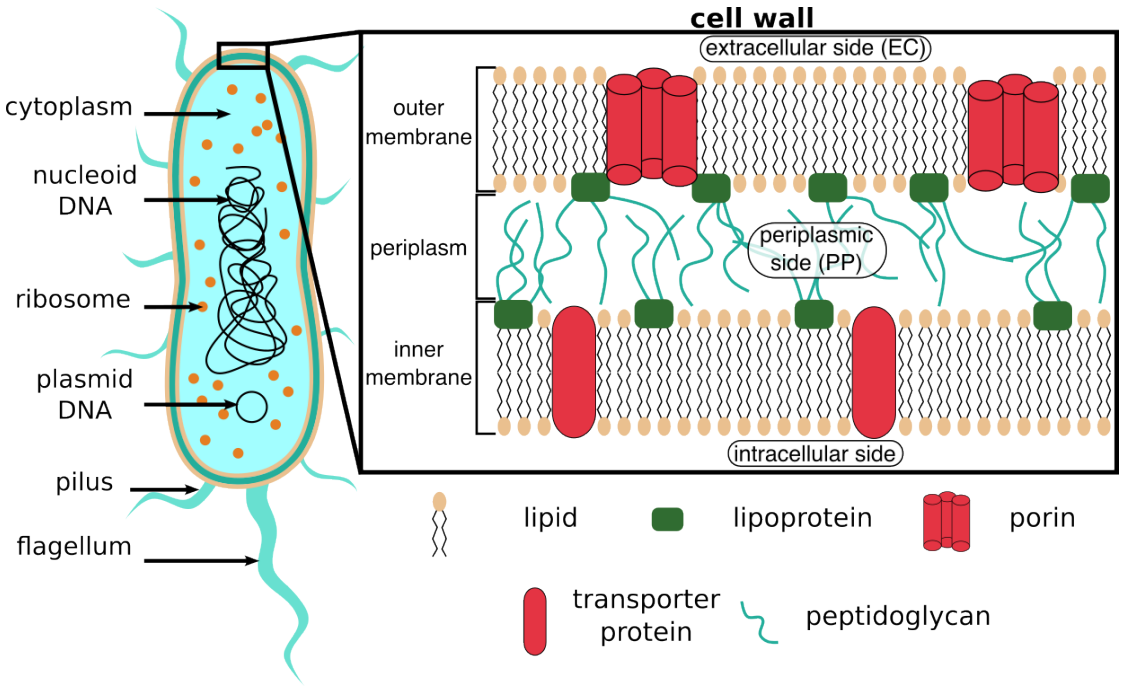

Figure 5.1: *Pseudomonas aeruginosa* bacterium and focus on its cell wall - Overview of the principal molecular and macromolecular components of *Pseudomonas aeruginosa*, a gram-negative bacterium. Only one flagellum is represented here, but in reality *Pseudomonas aeruginosa* would have many more flagella allowing to propel the bacterium in its environment. Pili have adhesive and sexual functions (219).

classified as nosocomial infections (207–209). Pathogens causing nosocomial infections are often multi-drug resistant, this is specifically true for *Pseudomonas aeruginosa* and, therefore, the World Health Organization stated for a critical need of new antibiotics against this group of pathogens in order to protect hospitalized patients (210).

Absorption of nutrients by *Pseudomonas aeruginosa* from the extracellular medium is mediated by outer membrane porin proteins (Fig. 5.1). Porins select incoming molecules by their charge and their size (211–213), thus it is of paramount importance to take into account drug permeability through porins in order to design effective drugs. Notably, *Pseudomonas aeruginosa* lacks many porins found in other Gram-negative bacteria, explaining partly its particular drug-resistance (213–216). Detailed descriptions of bacterial outer membrane porins as well as permeation mechanisms of small-molecules through these molecular gateways are available in recent excellent reviews (217, 218).

In this study we focused on the permeation of fosmidomycin through the OprO porin,

a polyphosphate-specific homotrimeric transmembrane protein. Fosmidomycin is an inhibitor of the 1-deoxy-D-xylulose 5-phosphate reductoisomerase, an enzyme involved in a biosynthesis pathway for isoprenoids, but specific to bacteria and protozoa. Isoprenoids play major roles in all organisms including electron transport and cell signaling (220), therefore blocking the bacterial specific pathway of isoprenoid biosynthesis is an effective strategy to impede proliferation of the pathogen.

Even though understanding the mechanism of fosmidomycin permeation through the OprO porin is of certain interest for drug design, the main motivation for undertaking this study was of methodological interest. Indeed, studying such process with MD simulations represent a sampling challenge, and recent work from Golla *et al.* showed that standalone umbrella sampling (US) (96) using the drug position along the pore as CV was leading to hysteresis effects (221). However, well-tempered metadynamics with multiple walkers along this same CV was also tested and yielded more accurate PMFs (221, 222). In line with the work from Golla *et al.* (221), well-tempered metadynamics with multiple walkers have also been successfully applied for studying permeation processes of several small-molecules through the OmpF porin from *E. coli* (223). Although standalone US has failed to provide meaningful PMFs for the permeation of fosmidomycin through OprO (221), improved flavors of US have been developed, namely: Hamiltonian replica-exchange umbrella sampling (108, 119, 120), temperature-accelerated sliced sampling (TASS) (224), and simulated tempering-enhanced umbrella sampling (STeUS) (206). In fact, replica-exchange US —a specific application of Hamiltonian replica-exchange US— has been successfully applied to obtain quantitative insights into permeation of antibiotics through OmpF (225, 226); and Acharya *et al.* applied TASS to rationalize permeation of ciprofloxacin through OmpF (227). Despite these recent achievements, qualitative comparison of augmented-US approaches described above to study permeation of antibiotics through OprO is lacking.

In this study, we have compared three methods for computing PMFs of the permeation of fosmidomycin through the OprO porin: (i) standalone umbrella sampling (US) (96), (ii) Hamiltonian replica-exchange, and more specifically the related REST2 method (119, 120), in combination with umbrella sampling (US-HREX), (iii) simulated tempering-enhanced umbrella sampling (STeUS) (121, 206), and (iv) replica-exchange umbrella sampling (REUS, also called BEUS) (108). To facilitate implementation of the different methods we simulated only one monomer of the native trimeric OprO

porin (Fig. 5.2A). Although we acknowledge that conformational ensemble of an OprO monomer alone might differs from that of a native OprO monomer in contact with its two monomer partners, we are mainly interested here in finding an efficient sampling method to study drug permeation rather than obtaining a model for the permeation process best predicting that of a native permeation process.

5.2 Results

Permeation of fosmidomycin in orientation 1 with standalone US. In the work from Golla *et al.* (221), PMFs of the permeation of fosmidomycin through OprO computed with standalone US exhibited hysteresis effects. To later compare how REUS, STeUS, and US-HREX improve PMFs obtained from standalone US, we have first computed reference PMFs with standalone US. To drive the permeation process, we used as a collective variable the z-component of the distance vector between the center of mass of OprO porin alpha carbons close to the porin's lumen (referred as pCOM) and the center of mass of fosmidomycin's phosphoryl group and amine group (referred as fCOM) (Fig. 5.2B). We will later simply refer to this collective variable as z. Moreover, in order to better understand what are relevant degrees of freedom orthogonal to z, we reduced the configurational space by (i) restraining the orientation of the antibiotic relative to the z-axis within $\pm 45°$, and (ii) applying a flat-bottom potential restraining the projected xy-distance between the pCOM and the fCOM bellow 1 nm. With orientation restraint applied on fosmidomycin, the antibiotic can enter the porin from the extracellular side either by presenting its amine group first (orientation 1), or its phosphoryl group first (orientation 2); we first investigated the orientation 1 (Fig. 5.2C).

Because the free energy is a state function, PMFs computed with umbrella sampling should not depend on the path taken by the pulling simulations (98, 99) used to generate the initial configurations. Therefore, a robust test to check the validity of PMFs computed with umbrella sampling is to compare PMFs obtained with initial configurations generated from "forward" and "reverse" pulling simulations. Here, this corresponds to initial configurations generated from EC-to-PP and PP-to-EC pulling simulations. Comparison of PMFs obtained from EC-to-PP and PP-to-EC pulling simulations is shown in Fig. 5.3. On this figure we observe that even though confidence intervals mostly overlap for the forward and reverse PMFs, these intervals are covering up to 7.5 $\mathrm{kcal\,mol^{-1}}$ for

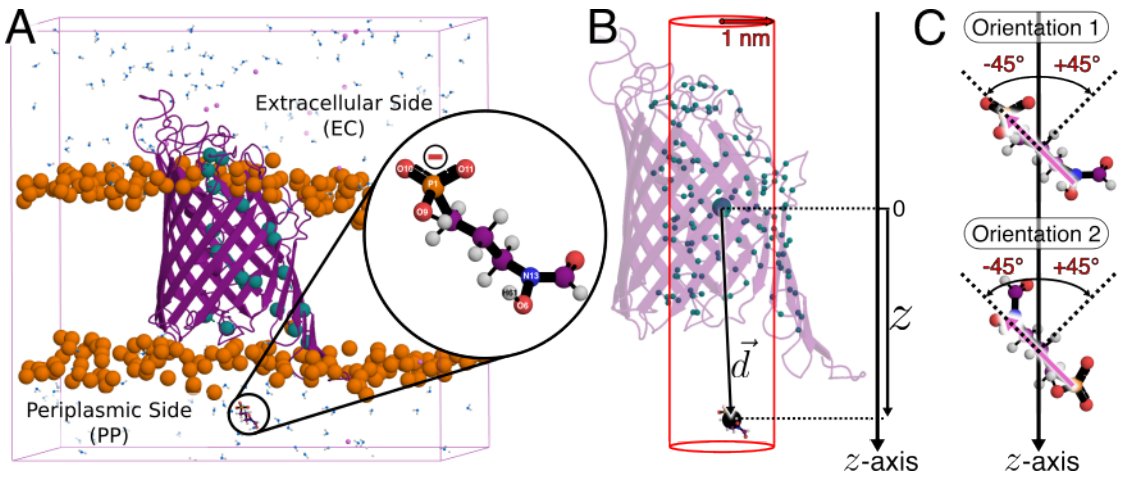

Figure 5.2: Setup for studying permeation of fosmidomycin through the OprO porin - (A) System overview. Orange spheres are phosphate groups of POPE lipids, cyan spheres are main lysine and arginine residues along the porin contributing to its anionic selectivity, fosmidomycin and waters are represented as balls and sticks, and pink spheres are potassium cations. Most of water molecules have been removed for clarity. The focus on fosmidomycin shows key atoms that have been used for defining the CV. (B) The large cyan bead depicts center of mass of all alpha carbons close to the porin's lumen (pCOM) represented as small cyan beads. The large black sphere represents center of mass of the following atoms in fosmidomycin: P1, O9-11, N13, H61 and O6 (fCOM). The z-component of the vector connecting the pCOM and the fCOM has been used as CV for driving the permeation process. The red cylinder represents the cylindrical flat-bottom potential restraint. (C) The two fosmidomycin's orientations studied in this work. The pink arrow represents the vector connecting COMs of phosphoryl and amine group from fosmidomycin; angle between this vector and the z-axis has been restrained with a flat-bottom potential such that this angle does not go above $+45°$or bellow $-45°$.

the EC-to-PP transition, and up to 5.0 kcal mol^{-1} for the PP-to-EC. Therefore, there is a huge uncertainty for the computed PMFs. Moreover, the free energy difference between the PP and EC edges from the EC-to-PP and PP-to-EC PMFs are significantly different. These results suggest that these PMFs suffer from hysteresis effects. To understand the cause of hysteresis, we looked at the bias potential energy over time in different umbrella windows to detect high variation of the bias potential. Through this process, we have notably observed a bias potential peak in umbrella window centered at $z = -0.007$ nm (Fig. 5.4A); this peak correlates with a water molecule getting trapped between fosmidomycin and the protein (Fig. 5.4B). Thus, solvent degrees of freedom are

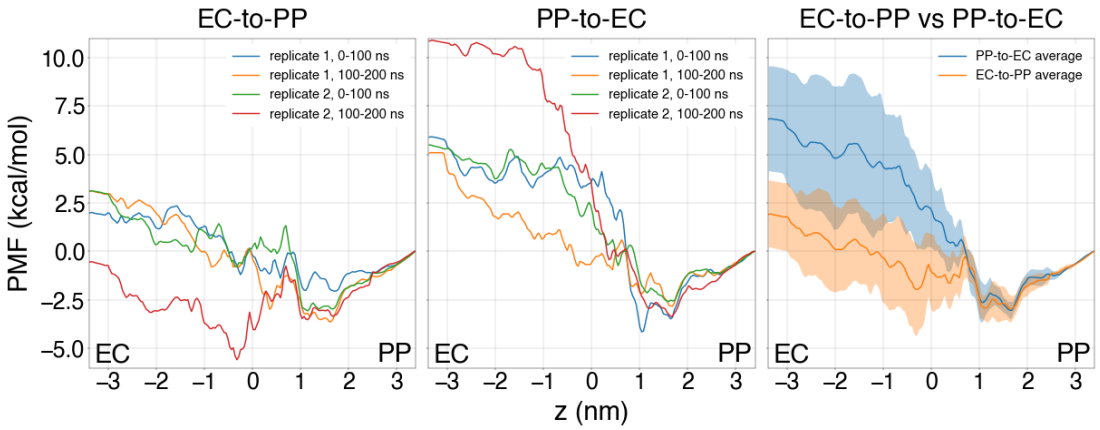

Figure 5.3: Permeation of fosmidomycin in orientation 1 with standalone US - Four independent pulling simulations have been carried out: two for the EC-to-PP transition and two for the PP-to-EC transition; initial configurations from these pulling simulations have been used to perform US. Umbrella windows in each US protocol have been split into two time-blocks: 0-100 ns and 100-200 ns, yielding a total of four PMFs for each transition. The average for each transition is shown on the right graph, confidence intervals represent two standard errors.

likely contributing to the observed hysteresis effects.

Overall, in line with the work from Golla *et al.*, our results confirm that using z as a collective variable with plain umbrella sampling is not sufficient to sample all relevant degrees of freedom involved in the permeation process of fosmidomycin through OprO.

Comparing PMFs of fosmidomycin permeation in orientation 1 obtained with REUS, STeUS and US-HREX. Three improved flavors of umbrella sampling have been tested to overcome the hysteresis effects observed in PMFs computed above with standalone umbrella sampling, namely: REUS, STeUS and US-HREX.

Replica-exchange umbrella sampling exploits the fact that neighboring umbrella windows can explore different regions of phase space and, therefore, by exchanging configurations between windows according to a Metropolis criterion (109), one expect to improve sampling of relevant degrees of freedom orthogonal to the chosen collective variable. In REUS, it is common practice to permit configuration exchanges along the whole CV-space, however in this study we permitted exchanges only between windows within subsets of z to reduce the amount of computational resources needed simultaneously (see

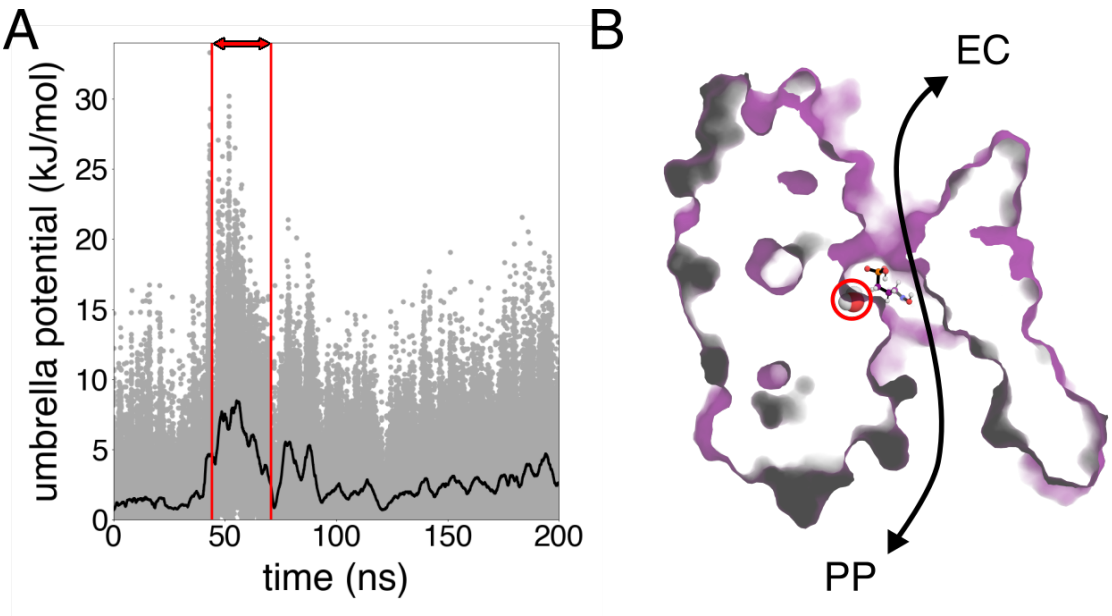

Figure 5.4: Variation of the bias potential in umbrella window corresponding to $z = -0.007$ nm correlates with a water molecule being trapped - (A) The bias potential energy or umbrella window centered at $z = -0.007$ nm as a function of time. We used the uniform_filter1d tool from scipy (228) with a filter size of 5000 points to compute the moving average depicted with the black line. (B) A representative snapshot of the simulation corresponding to simulation time: $45\,\text{ns} \leq t \leq 71\,\text{ns}$, highlighted with the red arrow in (A). The water molecule trapped between fosmidomycin (represented as balls and sticks) and the porin (represented as a surface) is highlighted with a red circle. The double-headed black arrow sketches the OprO lumen.

Materials and Methods for details).

Increasing the temperature is another mean to improve sampling of high energy states. This principle is exploited in the simulated tempering framework, where a Metropolis criterion is used to attempt exchanges between temperature states within a single simulation. The high temperature states have a higher probability to cross energy barriers and, ultimately, exchanges with low temperature states will also improve configurational sampling in the base temperature state. Combining simulated tempering with umbrella sampling in STeUS is therefore an appealing method to improve sampling along relevant degrees of freedom orthogonal to the chosen CV. In this study, simulated tempering was applied in each umbrella window with temperatures ranging from 300 K to 348 K with 4 K-steps, and only data acquired at the base temperature were analyzed

(see Materials and Methods for details).

Umbrella sampling combined with Hamiltonian replica-exchange (US-HREX) is the third approach we have tested to compute PMFs of the fosmidomycin permeation through the OprO porin. To apply US-HREX, one needs a rough idea of the interactions contributing to low energy states. Indeed, the goal of HREX is to increase the potential energies corresponding to the long living states to favor transitions. Here, many positively charged residues inside the OprO porin's lumen contribute favorably to interactions with fosmidomycin (221). Therefore, we chose to scale positive charges within the porin, and to reduce negative charges within the phosphoryl group of fosmidomycin. Down-scaling of positive charges was achieved by multiplying by a factor λ the following charges: (i) positive charges of the guanidine group of arginines, and (ii) positive charges of the amino group of lysine. We used 24 λ-replicas per window, with λ ranging from 0.793 to 1, with a 0.009-λ-step. We then added a positive term to negative charges within the phosphoryl group of fosmidomycin, and to negative charges within the porin (see Materials and Methods for details). With this protocol we weaken electrostatic interactions between fosmidomycin and the porin, while maintaining the system with a neutral net charge. Although it is common practice when applying HREX to allow uniform neutralizing background charge to balance the net charge resulting from charge scaling. Indeed, it has been shown that using background charge could lead to artifacts, especially for non-homogeneous systems like solvated membranes (229). This is the reason why we compensated the overall positive charge loss by decreasing negative charges in the porin until charge balance was achieved.

PMFs of the fosmidomycin permeation through OprO obtained with the three aforementioned methods are shown in Fig. 5.5. PMFs from US-HREX are difficult to compare at this stage because we have fewer data than for the two other methods. However, by looking at the currently available data, we observe steep rise of free energy around the region z =1 nm suggesting that these PMFs suffer from hysteresis. In addition, the EC-to-PP and PP-to-EC PMFs show big discrepancy within the region z=[-1, 1] nm for PMFs obtained with US-HREX. Concerning PMFs from REUS and STeUS, we remark very few overlap between EC-to-PP and PP-to-EC averages for STeUS, but large overlap for REUS. In addition, we observe that confidence intervals for REUS are smaller than for STeUS. Indeed the sum of two standards errors for the entire EC-to-PP average PMF from REUS equals 76.06 kcal mol^{-1}, whereas the sum of two standards errors for

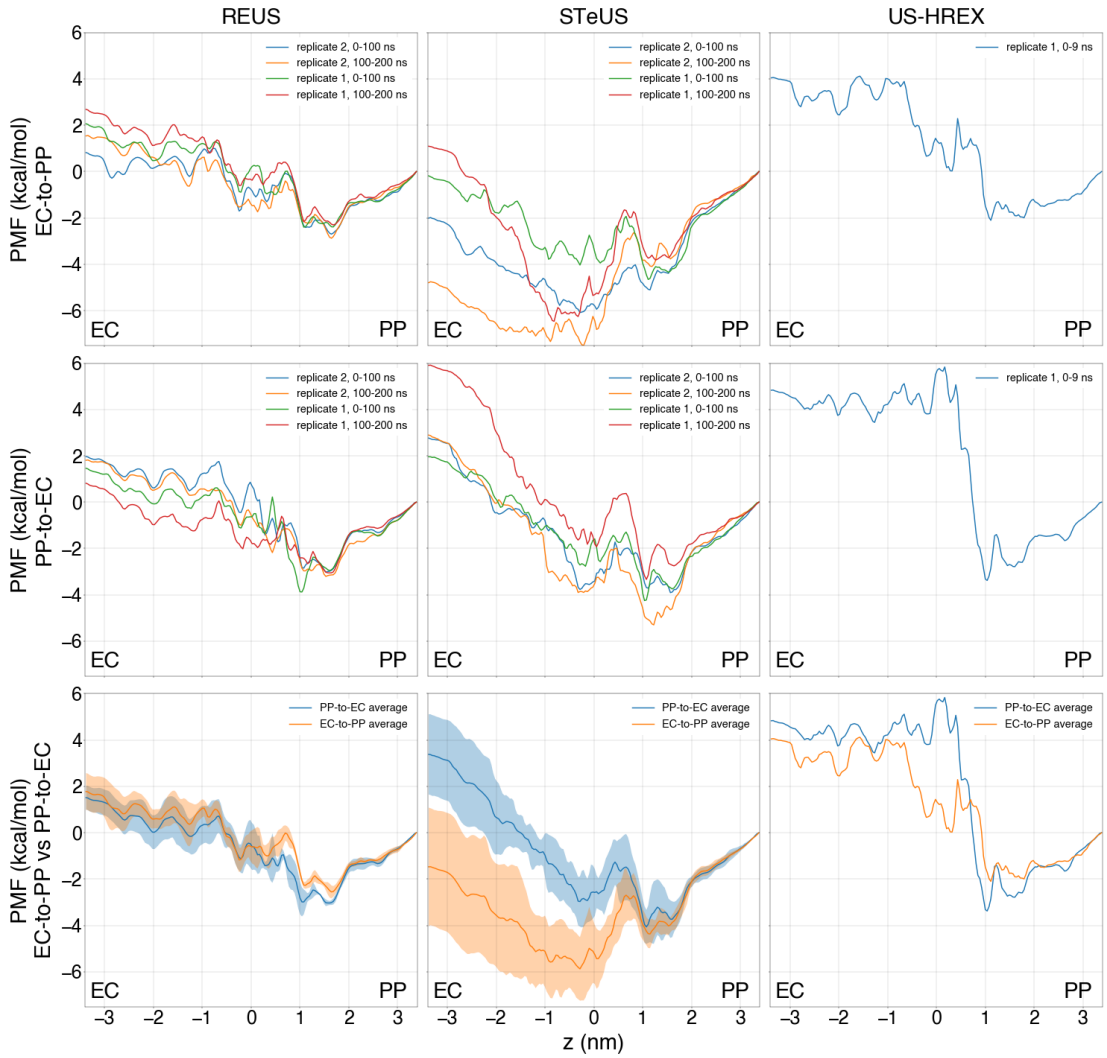

Figure 5.5: Permeation of fosmidomycin in orientation 1 obtained with REUS, STeUS and US-HREX: PMFs comparison - EC-to-PP and PP-to-EC PMFs were computed with initial conformations obtained from EC-to-PP and PP-to-EC pulling simulations respectively. For each EC-to-PP and PP-to-EC setup, two independent pulling simulations were carried out: replicate 1 and 2, except for US-HREX with only one replicate per setup. For each of the two PMFs obtained for each setup, we split the umbrella windows into two time-blocks: 0-100 ns and 100-200 ns, except for US-HREX for which no time-blocking was applied. The average for each EC-to-PP and PP-to-EC setup is shown in the third column, confidence intervals represent two standard errors.

the entire EC-to-PP average PMF from STeUS equals 230.37 kcal mol^{-1}. Similarly, the error sum for the PP-to-EC average PMF from STeUS is ∼1.9 greater than for REUS. Therefore, our results suggest that among the three methods tested here, only REUS is able to provide PMFs of the fosmidomycin permeation in orientation 1 through the OprO porin without hysteresis effect.

To understand why PMFs of the permeation of fosmidomycin through OprO computed with STeUS display hysteresis effects, we took a closer look to the atomic trajectories. Through this process, we have notably noticed that fosmidomycin has been trapped in a pocket at the EC-entrance of OprO in umbrella window centered at $z = -1.63$ nm (Fig. 5.6A–B). This transition occurred at ∼100 ns of simulations time, and fosmidomycin stays in this pocket for the last following 100 ns, meaning that even exchanges across temperature states are not sufficient to free the antibiotic from this pocket. To further test the stability of this unexpected state, we ran three 100 ns free simulations starting from configurations extracted at t =130 ns, t =160 ns and t =200 ns from the base temperature of this specific umbrella window. For each of these simulations, we did not observe fosmidomycin getting out of this pocket. Beside this peculiar state that denotes clearly from other configurations sampled from other replicates, we also noticed that protein loops within the EC and PP entrances were very flexible compared to what is observed in REUS or standard US. In light of these results, we hypothesize that, likewise temperature states way above the physiological protein temperature could sample completely unfolded and irrelevant conformations, even temperature states bellow this upper limit could probably lead to sampling subtle states that are irrelevant when studying a specific process.

By averaging all EC-to-PP and PP-to-EC PMFs computed with REUS in Fig. 5.5, we obtained a PMF with sub-kcal mol^{-1} standard errors shown in Fig. 5.7. Taken together, our results demonstrate that REUS is the best method to study the permeation of fosmidomycin through OprO in orientation 1.

Permeation of fosmidomycin in orientation 2 with REUS. PMFs computed in the above sections correspond to the orientation 1 of fosmidomycin as defined in Fig. 5.2C. To further test if REUS is able to provide accurate PMFs of the permeation of fosmidomycin through OprO, we carried out REUS for the orientation 2.

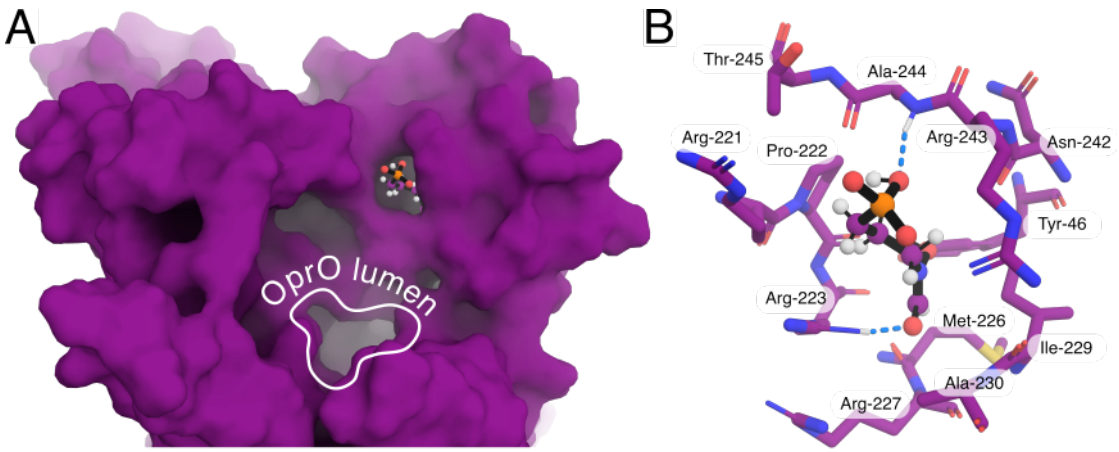

Figure 5.6: Snapshot of the umbrella window centered at $z = -1.63$ nm in STeUS: fosmidomycin gets trapped in a pocket - (A) Overview of a configuration where fosmidomycin is trapped in a pocket at the entrance of the EC. OprO is represented as a purple surface, and fosmidomycin as balls and sticks. (B) Focus on fosmidomycin and protein residues of the pocket. Blue dotted lines are hydrogen bonds between the protein and the antibiotic. In this snapshot, interactions between fosmidomycin and OprO seem to be mostly hydrophobic.

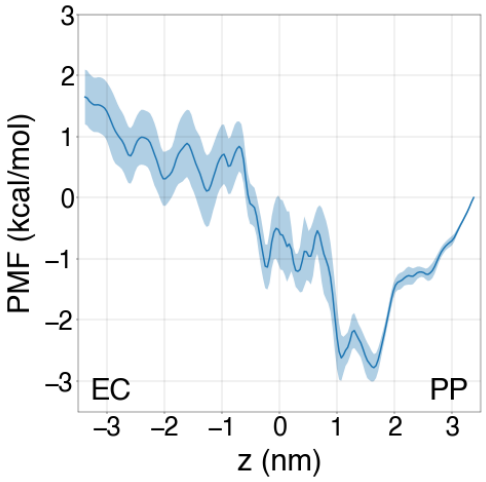

Figure 5.7: Permeation of fosmidomycin in orientation 1 with REUS: final PMF - Average PMF of EC-to-PP and PP-to-EC PMFs computed with REUS in Fig. 5.5, confidence intervals represent two standard errors.

Similarly to the procedure applied for PMFs with fosmidomycin in orientation 1, we checked for hysteresis by comparing EC-to-PP and PP-to-EC PMFs with fosmidomycin in orientation 2, *i.e.* PMFs computed with REUS starting from initial configurations obtained from EC-to-PP or PP-to-EC pulling simulations, respectively (Fig. 5.8A). Unlike with orientation 1 where PMFs were largely overlapping, we observe overlap

between the EC-to-PP and PP-to-EC PMFs only in the region z=[0.5, 3.5] nm; this result suggests that these PMFs suffer from hysteresis. However, in our application of REUS, exchanges are allowed only within subsets of the z-space. Therefore, we hypothesized that sampling could be improved in regions of z-space where we observe large discrepancies between EC-to-PP and PP-to-EC PMFs, by increasing the number of windows that are able to exchange configurations in sub-sampled regions.

To test the aforementioned hypothesis, we first overlaid all PMFs to identify regions with high discrepancies when comparing EC-to-PP and PP-to-EC PMFs (Fig. 5.8A, last column). By doing so, we observe divergence of EC-to-PP and PP-to-EC PMFs in the region z=[0.45, 1.54] nm, while other regions exhibits roughly similar shapes. Thus, we carried out an additional batch of REUS within this specific region for each replicate (red rectangle in Fig. 5.8B, fourth column), and we allowed exchanges between all windows within this batch. We then computed new PMFs by using the data from the new REUS batch in the region z=[0.45, 1.54] nm and the old data elsewhere (Fig. 5.8B). As opposed to what was obtained without this additional REUS batch, the EC-to-PP and PP-to-EC PMFs with data from the new REUS batch show a perfect overlap. Moreover, we observe that confidence intervals are covering only up to 2.9 kcal mol^{-1} and 2.0 kcal mol^{-1} for the PP-to-EC and EC-to-PP PMFs respectively. Therefore, our additional batch of REUS successfully improved the sampling of degrees of freedom orthogonal to z in the region z=[0.45, 1.54] nm.

By averaging all EC-to-PP and PP-to-EC PMFs in Fig. 5.8B, we obtained the PMF with sub-kcal mol^{-1} standard errors shown in Fig. 5.9. Altogether, we showed that REUS is able to provide an accurate PMF of the permeation of fosmidomycin in orientation 2 without hysteresis effect.

5.3 Discussion

In this work we have presented PMFs of the permeation of fosmidomycin through the OprO porin with sub-kcal mol^{-1} accuracy by using REUS. Moreover, by comparing PMFs from multiple replicates computed from simulations initiated with initial configurations generated either from forward or reverse pulling simulations, we have carefully checked that our PMFs are free from hysteresis effects. Our application of REUS is not conventional as we have split our CV-space into several subsets, and only windows

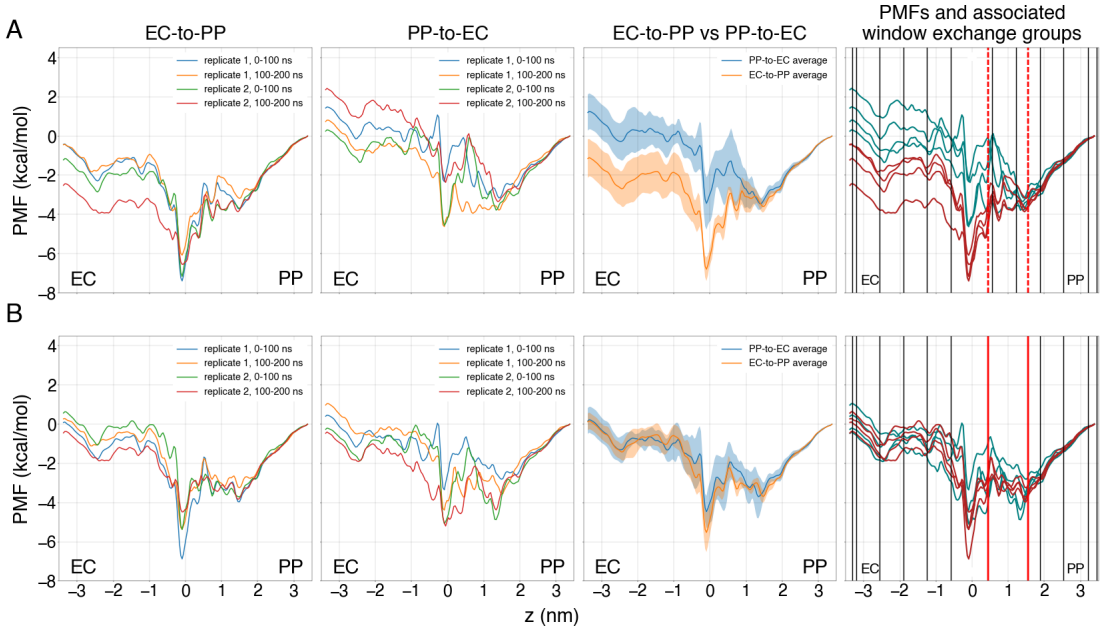

Figure 5.8: Permeation of fosmidomycin in orientation 2 with REUS. - (A) PMFs of the permeation process with sub-optimal choice of window sub-sets within which exchanges are allowed. EC-to-PP and PP-to-EC PMFs refer to REUS setups started with initial conformations obtained from EC-to-PP and PP-to-EC pulling simulations respectively. For each setup, two independent pulling simulations were carried out: replicate 1 and 2. For each of the two PMFs obtained for each setup, we split the umbrella windows into two time-blocks: 0-100 ns and 100-200 ns. The average for each EC-to-PP and PP-to-EC setup is shown in the third column, confidence intervals represent two standard errors. EC-to-PP and PP-to-EC PMFs are respectively depicted in dark red and cyan in the graph of the fourth column. Windows within z-subsets delimited by the dark rectangles are ran in parallel and, therefore, are allowed to exchange configurations. Region within the red-dotted rectangle show high discrepancy between EC-to-PP and PP-to-EC PMFs. (B) PMFs of the permeation process with an optimal choice of window sub-sets within which exchanges are allowed. Descriptions of each column are identical to (A), except that data from a new REUS batch in the region z=[0.5, 3.5] nm (red rectangle region in the fourth column) were used to obtain the PMFs.

within a subset were able to exchange configurations. This allows to reduce the amount of compute resources needed simultaneously, and to dedicate more resources to CV regions that are the most critical from a sampling perspective. Therefore, our work shows that REUS is a robust method to study the permeation process of antibiotics through bacterial porins, in line with previous studies (225, 226).

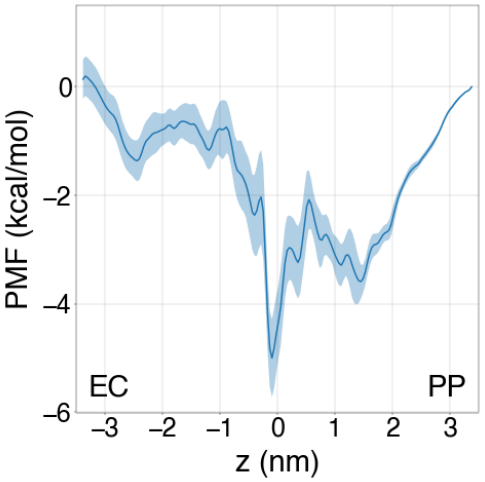

Figure 5.9: Permeation of fosmidomycin in orientation 2 with REUS: final PMF - Average PMF of EC-to-PP and PP-to-EC PMFs computed with REUS in Fig. 5.8B, confidence intervals represent two standard errors.

In this study, we have compared different flavors of umbrella sampling for studying the permeation of fosmidomycin through the OprO porin, namely: REUS, STeUS nad US-HREX. Our data are scarce concerning US-HREX, but the currently available data suggest that our implementation of US-HREX is not optimal to compute PMFs of the permeation process without hysteresis effect. Moreover, the resources needed for US-HREX are prohibitive compared the two other tested methods, which hinders the practical application of US-HREX. Regarding the STeUS approach, hysteresis was hardly improved compared to standalone US. We hypothesize that, for some systems, simulations at high temperatures could sample irrelevant states relative to a specific studied process. Furthermore, in theory, entropic barriers would not benefit from sampling at high temperatures, as opposed to enthalpic barriers. Thus, similarly to what as been done for OmpF (230), studying the nature of the barriers involved in small-molecule permeation through OprO would be needed to understand the nature of the energetic barriers involved in permeation processes through OrpO. Even though STeUS appears to be a poor method for studying the permeation of fosmidomycin through the OprO porin, this does not mean that STeUS is not useful to overcome hysteresis when studying other biological processes. For instance, it has been shown recently that STeUS successfully improves configurational sampling in the case of drug permeation through lipid membranes (206).

Across all the tested methods, we applied several restraints to facilitate configurational sampling: (i) a flat-bottom potential restraining the projected xy-distance between

the pCOM and the fCOM bellow 1 nm, and (ii) a flat-bottom potential restraining the orientation of fosmidomycin relative to the z-axis within $\pm 45°$. Moreover, during the pulling simulations needed to generate initial configurations for the different flavors of umbrella sampling, we applied heavy atom restraints to mitigate non-equilibrium effects. Further studies would have to be undertaken to clearly quantify how much each restraint helps the configurational sampling.

5.4 Materials and Methods

Simulation setup. REUS and US-HREX were carried out with Gromacs (194) version 2020.6 built with Open-MPI and patched with Plumed version 2.7.2 (142, 195). Standalone US and ST-US were carried out with Gromacs version 2020.4 patched with Plumed 2.7.0. Atomic coordinates of the porin trimers (pdb code 4RJW (231)) and fosmidomycin, as well as forcefield parameters were kindly provided by Prof. Ulrich Kleinekathöfer. Protein, lipids, water and ions were parameterized with CHARMM36 forcefield (232) and fosmidomycin with CGenFF-based forcefield generated with ParamChem server (233), validation of fosmidomycin parameters can be found in supplement material of Ref. (221). From atomic coordinates of OprO trimers, two monomers have been removed to obtain a monomeric form of the OprO porin. We next inserted the monomer into a membrane bilayer using 334 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphoethanolamin (POPE) with CHARMM-GUI membrane builder (234). We then solvated the system with TIP3P water molecules (197), and added 14 potassium ions to neutralize the system.

Electrostatic interactions were computed with particle-mesh Ewald method (201), using a Fourier grid spacing of 0.12 nm and a real-space cutoff at 1.12 nm. Short-range repulsion and dispersion interactions were described with a Lennard-Jones potential with a cutoff at 1.2 nm and a force-switch modifier set to 1.0 nm. Angles and bonds of water molecules were constrained with SETTLE (123), and bonds involving other hydrogen atoms were constrained with LINCS (124). Energy minimization was carried out with steepest descent and equilibration was performed following the six-step protocol provided by CHARMM-GUI. In brief, the aforementioned equilibration protocol consisted in two 125 ps-long NVT equilibration simulations, one 125 ps-long NPT equilibration, and three 500 ps-long NPT simulations. During equilibration, position restraints were applied on

lipid phosphate atoms, and protein and fosmidomycin heavy atoms; restraints were slowly released through the six equilibration steps. Pressure at 1 bar was controlled by the Berendsen barostat (τ=5 ps) and temperature at 300 K by the Berendsen thermostat (τ=1 ps) (91).

For production runs, a 4 fs time-step was used for standalone US, US-HREX and REUS, and a 3.5 fs time-step was used for ST-eUS. Using a time-step higher than the commonly used 2 fs time-step was possible by modelling all hydrogens with virtual sites. Parrinello-Rahman barostat (92) was used as barostat and velocity-rescale was used as thermostat (90).

Standalone umbrella sampling. We carried out four 100 ns-long independent pulling simulations with a force constant of 1000 kJ mol^{-1} nm^{-2}, (i) two from PP to EC by pulling along the z-component of the vector connecting the COM of alpha carbons close to the cavity, and the COM of fosmidomycin's phosphoryl group, and (ii) two from PP to EC by pulling along the z-component of the vector connecting the COM of alpha carbons close to the cavity, and the COM of fosmidomycin's amine group. During pulling simulations, the angle between the vector connecting the two ends of the drug and the z-axis was restrained within [-45°, 45°] with a force constant of 7878 kJ mol^{-1} rad^{-2}. During pulling simulations, we also applied a flat-bottom potential to restrain the antibiotic within a cylinder of radius 1 nm centered on the COM of alpha carbons close to the cavity with a force constant of 1000 kJ mol^{-1} nm^{-2}. In addition, to mitigate non-equilibrium effect during pulling simulations, we used restraints on the backbone and side chains heavy atoms of the protein, with a force constant of 1000 kJ mol^{-1} nm^{-2} and 100,kJ mol^{-1} nm^{-2} respectively. Trajectories from pulling simulations were then post-processed to map each frame onto the CV z, and defined as the z-component of the vector connecting the COM of OprO porin alpha carbons close to the cavity, and the center of mass of fosmidomycin. The CV z was not directly used for pulling simulations to mitigate influence of the angle restraint during the pulling simulations.

After mapping frames onto z, we launched 144 umbrella simulations in the range [-3.4, 3.34] nm, with a force constant of 2000 kJ mol^{-1} nm^{-2} for each harmonic potential, for 30 ns. During this equilibration step, force constants of the restraints on the backbone and side chains heavy atoms of the protein were reduce to 100 kJ mol^{-1} nm^{-2} and 10,kJ mol^{-1} nm^{-2} respectively. Final coordinates obtained from the previous step were

then used to launch production simulations of 200 ns without heavy atom restraints. Choice of umbrella windows spacing will be explained when describing the REUS protocol bellow.

All PMFs were computed with wham ("WHAM: the weighted histogram analysis method", version 2.0.11, `http://membrane.urmc.rochester.edu/wordpress/?page_id=126`).

Combining umbrella sampling with Hamiltonian-replica exchange: US-HREX. The CV z chosen to drive the fosmidomycin permeation, number of umbrella windows, spacing between windows, starting initial configurations, restraints, and force constant of harmonic potentials were identical to what has been described in "Standalone umbrella sampling". Scaled charges are depicted in Fig. 5.10A, and the protocol to scale positive and negative charges is depicted in Fig. 5.10B.

To identify the λ-range that maintains stable protein conformation, we ran six simulations for 400 ns with following λ factors: 1, 0.8, 0.6, 0.4, 0.2, 0.1 and 0.05. Simulations with λ =0.1 and 0.05 were not stable and crashed within the first simulation steps. However, simulations with λ =1 through 0.2 were stable, and visual inspection of trajectories as well as root mean square residue fluctuations did not indicate protein unfolding. For production, we ran 24 λ-replicas for each of our 144 umbrella windows. In order to obtain $\sim$ 19% of exchange acceptance between the 24 replicas, we scaled positive charges from $\lambda = 1$ to $\lambda = 0.793$ with 0.009 λ-steps. Each umbrella window was ran for 9 ns.

Combining umbrella sampling with simulated tempering: STeUS. The CV z chosen to drive drug permeation, number of umbrella windows, starting initial configurations, restraints, total simulation time, and force constant of harmonic potentials were identical to what has been described in "Standalone umbrella sampling".

Simulated tempering was carried out with temperatures ranging from 300 K to 348 K with 4 K-steps. To determine initial temperature weights, we followed the procedure described by Park *et. al.* (235) which involves simulated annealing simulation from the lowest to the highest temperature. The weights from this simulation are taken as:

$$g_{n+1} - g_n \approx (\beta_{n+1} - \beta n)\frac{E_n + E_{n+1}}{2} \tag{5.1}$$

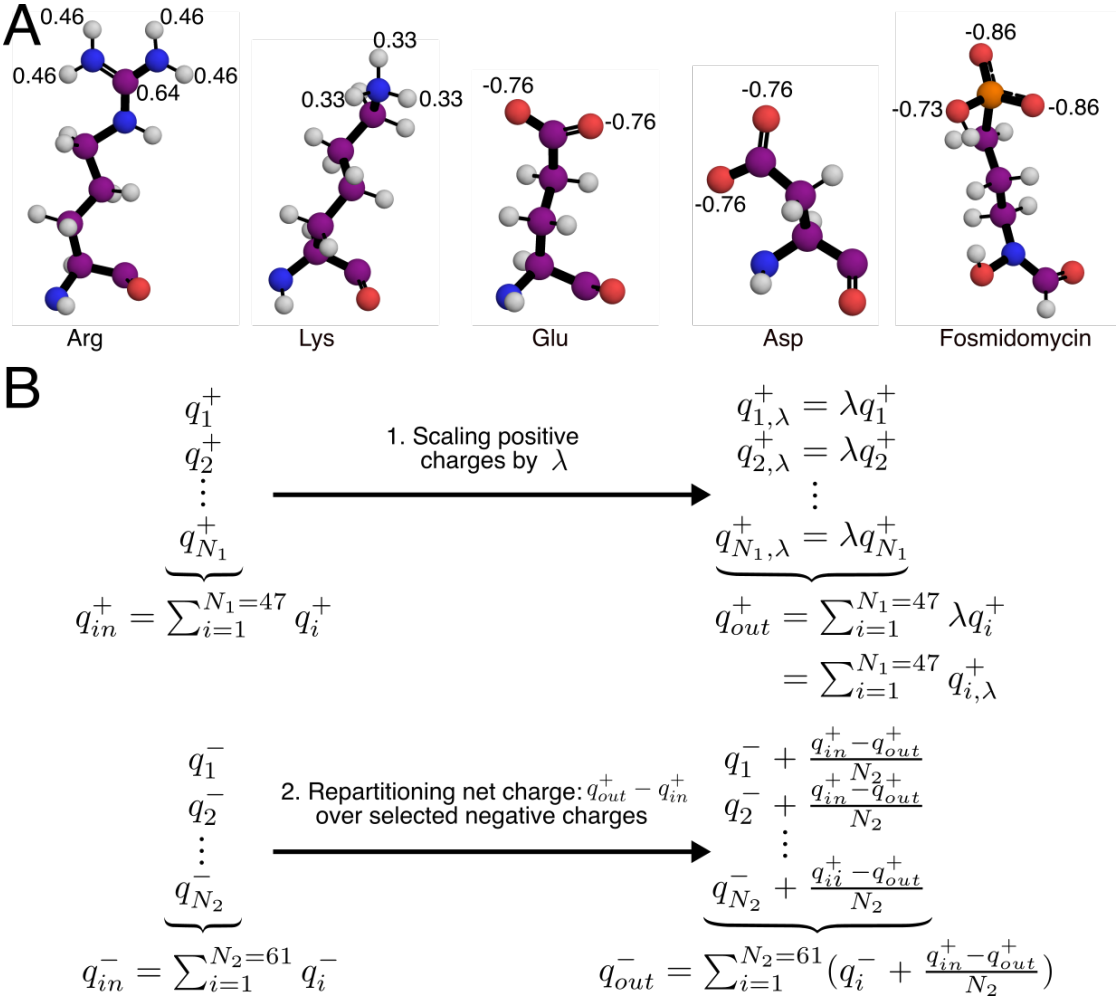

Figure 5.10: Hamiltonian replica-exchange protocol - (A) Residue types implicated in charge scaling are depicted as balls and sticks, charges concerned by scaling are written beside the associated atoms. (B) Scheme of our scaling protocol implemented in python. First, positive charges q_1^+ through $q_{N_1}^+$ are scaled by a factor λ. Scaling leads to a net charge equal to $q_{out}^+ - q_{in}^+$. Therefore, we divide the resulting net charge by the number N_2 of negative charges we want to scale, we then remove this quantity to each of these negative charges q_1^- through $q_{N_2}^+$ to obtain a neutral system.

where g_{n+1} and g_n are weights for temperature states T_{n+1} and T_n respectively ($T_{n+1} > T_n$), β_{n+1} and β_n are equal to $1/k_B T_{n+1}$ and $1/k_B T_n$ respectively, and E_{n+1} and E_n are average potential energies for temperature states T_{n+1} and T_n respectively. From eq. 5.1, each weight is determined from the weight of neighboring lower temperature state, and the weight with the higher index is initialized to 0. Practically, a simulated annealing

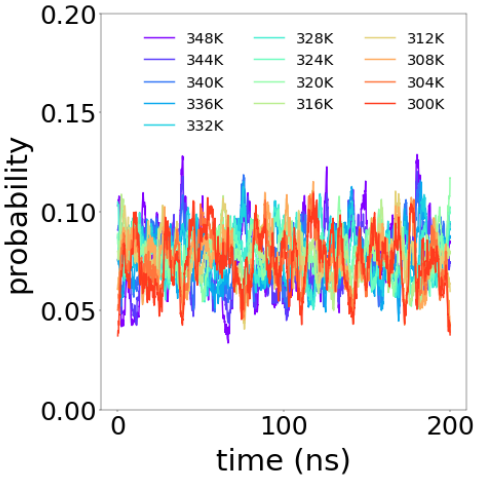

Figure 5.11: Temperature state probabilities through time for window z=0 nm - To check that all temperature states were equally visited during the simulations we looked at the counts for each temperature states through simulation time. Counts for each state have been normalized with the total number of counts and is shown here as a probability. Probabilities curves are smoothen with the uniform_filter1d tool from scipy (228) with a filter size of 500 points.

simulation for umbrella window corresponding to $z = 0$ nm was carried out, for which temperature was increased from 300 K to 348 K, with 2 ns per temperature state and 100 ps for each heating step for a total simulation time of 27.2 ns. With this protocol, we obtained the following weights from g_{12} through g_0: 0, 4370.5, 8600.3, 12694.8, 16659.9, 20500.2, 24220.5, 27826.5, 31322.1, 34711.3, 37999.4, 41189.1, 44284.0.

To further optimize the weights we carried out a simulated tempering simulation for umbrella window corresponding to $z = 0$ nm with the previously determined weights, with exchange attempt every 100 steps, with the wang-landau algorithm, and for a total simulation time of 43 ns. This simulation was used to determine the final weights used for production: 0, 4361.5, 8577.3, 12657.8, 16617.9, 20459.2, 24166.5, 27768.5, 31265.1, 34644.3, 37926.4, 41114.1, 44206.0. Finally, every 144 umbrella windows have been ran with simulated tempering using final weights determined as explained above, with exchange attempt every 100 steps.

We checked that weights were stable through simulations and that all states were sampled equally Fig. 5.11.

Replica-exchange umbrella sampling: REUS. The CV z chosen to drive drug permeation, number of umbrella windows, starting initial configurations, restraints, total simulation time, and force constant of harmonic potentials were identical to what has been described in "Standalone umbrella sampling".

To determine spacing between the 144 umbrella windows along z we ran a series of

60 ps-long simulations and optimize spacing in order to reach an exchange acceptance between neighboring windows between 0.25% and 0.47%. The main idea of the optimization process involves modelling each z-spacing and the corresponding probability of exchange rejection with a linear relationship (Fig. 5.12). From this linear relationship, the algorithm selects z-spacing between windows that correspond to an exchange rejection of 0.66%. Because exchanges between 144 umbrella windows running in parallel would have been computationally prohibitive, we grouped our umbrella windows in eleven subsets of our total z-range, and allowed exchanges only within a subset. To guarantee good overlap between umbrella windows at the edges of one subset and umbrella windows of neighboring subsets, each subset contains two umbrella windows centered at the same z position as the last two windows from the previous subset, and two umbrella windows centered at the same z position as the first two windows from the following subset (Figs. **??** and **??**). Because the constriction region of OprO at [-1, 1] nm is the most difficult region to sample due to extensive contacts between the porin and fosmidomycin, we used more replicas for the subset close to this region. Hence, the subset centered in $z = 0$ nm was composed of 24 windows, all other subsets were composed of 16 windows except for the two subsets including the z-range extrema which were composed of only eight windows. We checked that the average probabilities of exchange of our production REUS were consistent with our optimization procedure, *i.e.* that exchange probabilities were within the [0.25%, 0.47%]-range, and effectively only few acceptance probabilities were outside this range (Figs. 5.13, 5.14, 5.15, and 5.16).

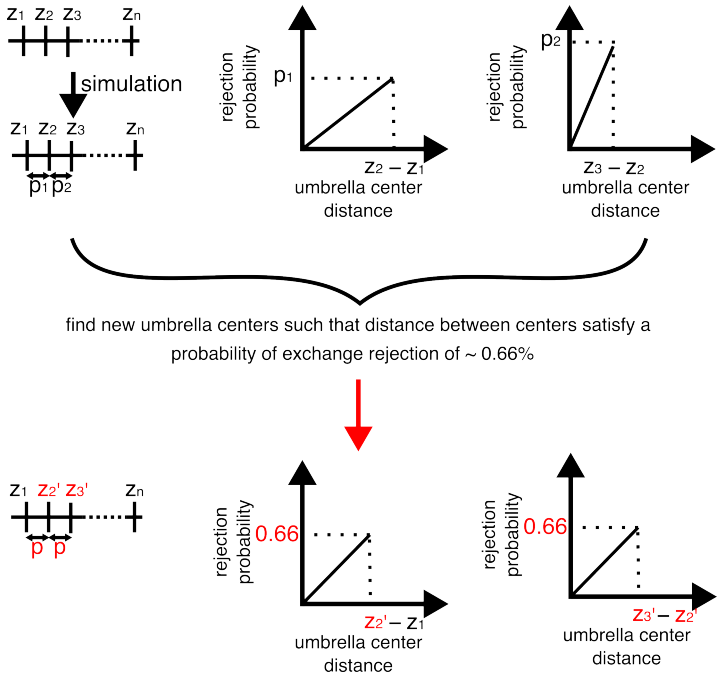

Figure 5.12: General idea behind the algorithm optimizing the distance between centers of bias umbrella potential for REUS. - Series of simulations are ran to compute average probabilities of exchange rejection between windows. From these probabilities, linear relationships between probabilities and distance between centers of bias umbrella potentials are assumed. From this linear relationships, the algorithm find new centers of bias umbrella potential to satisfy rejection probability between windows of about 0.66%, which is equivalent to an acceptance probability of about 0.36%.

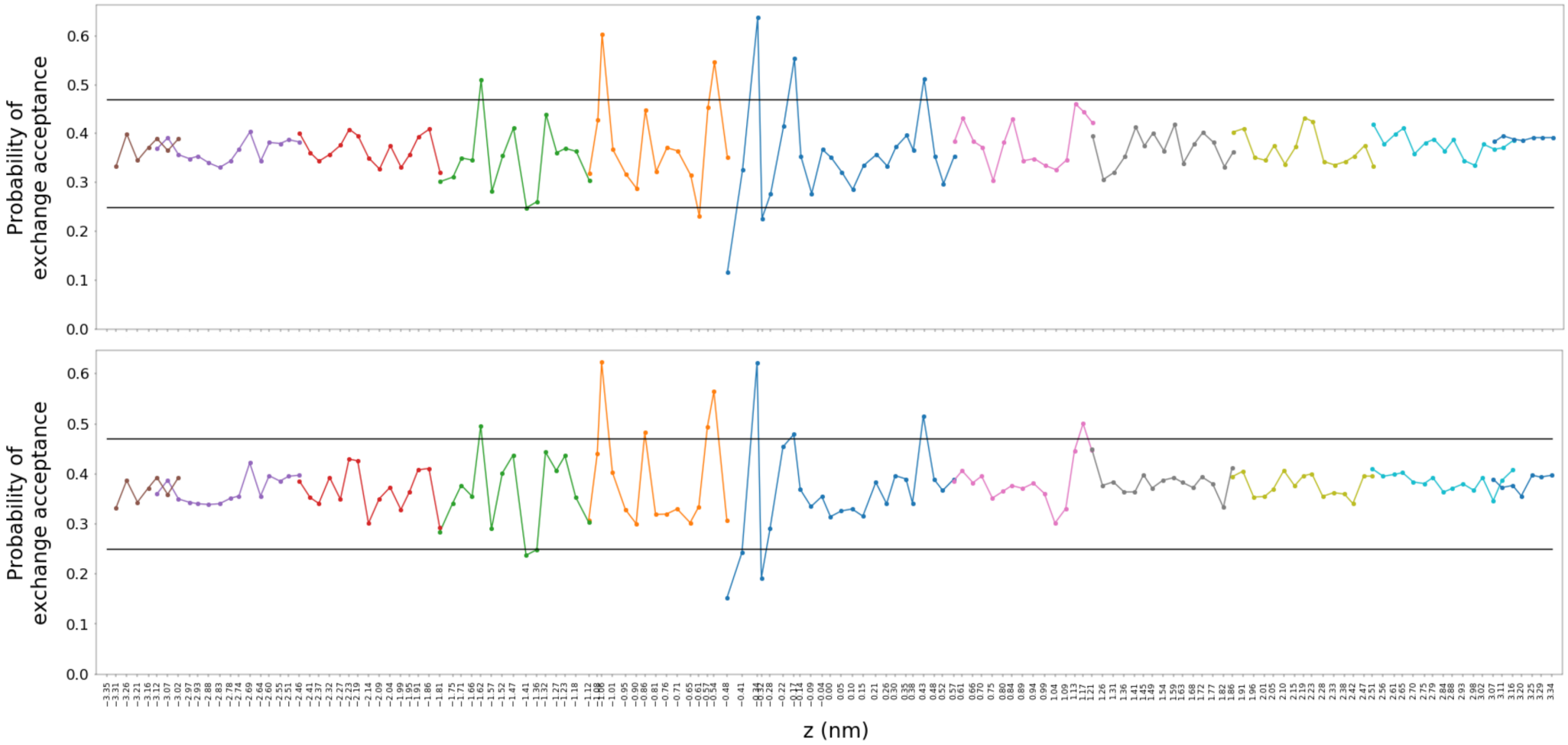

Figure 5.13: Average exchange probabilities between umbrella windows after 200 ns of production simulation for two forward transition replicates, in orientation 1 - Umbrella windows that were allowed to exchange configurations are depicted with the same color. Each point gives the average probability of exchange between the umbrella window centered at the associated z value and the preceding umbrella window on the z-axis; the first point of each exchange subset is not shown here as it does not have proceeding window in the subset. Probability region in-between the two black lines was the target region when optimizing spacing between umbrella windows.

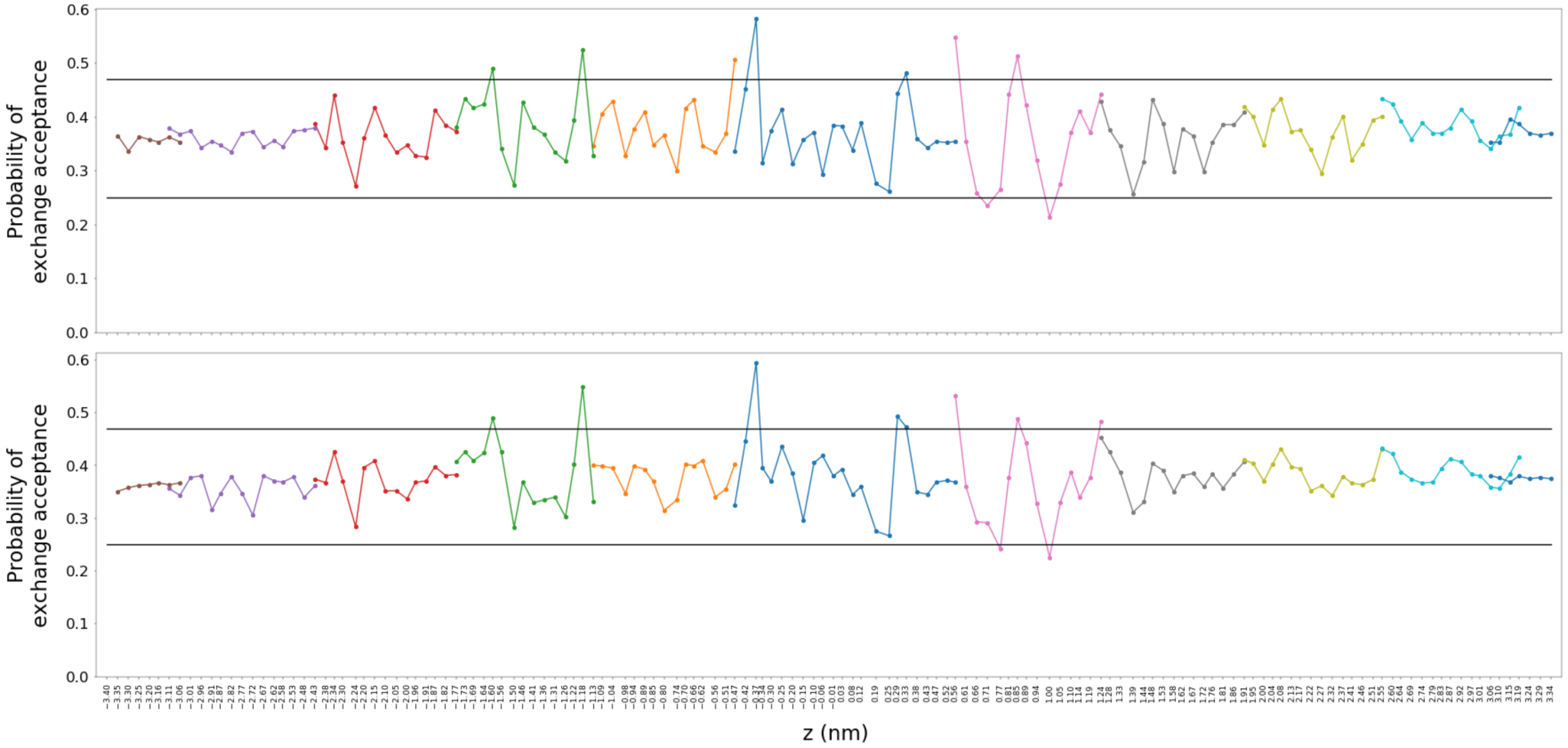

Figure 5.14: Average exchange probabilities between umbrella windows after 200 ns of production simulation for two reverse transition replicates, in orientation 1 - Umbrella windows that were allowed to exchange configurations are depicted with the same color. Each point gives the average probability of exchange between the umbrella window centered at the associated z value and the preceding umbrella window on the z-axis; the first point of each exchange subset is not shown here as it does not have proceeding window in the subset. Probability region in-between the two black lines was the target region when optimizing spacing between umbrella windows.

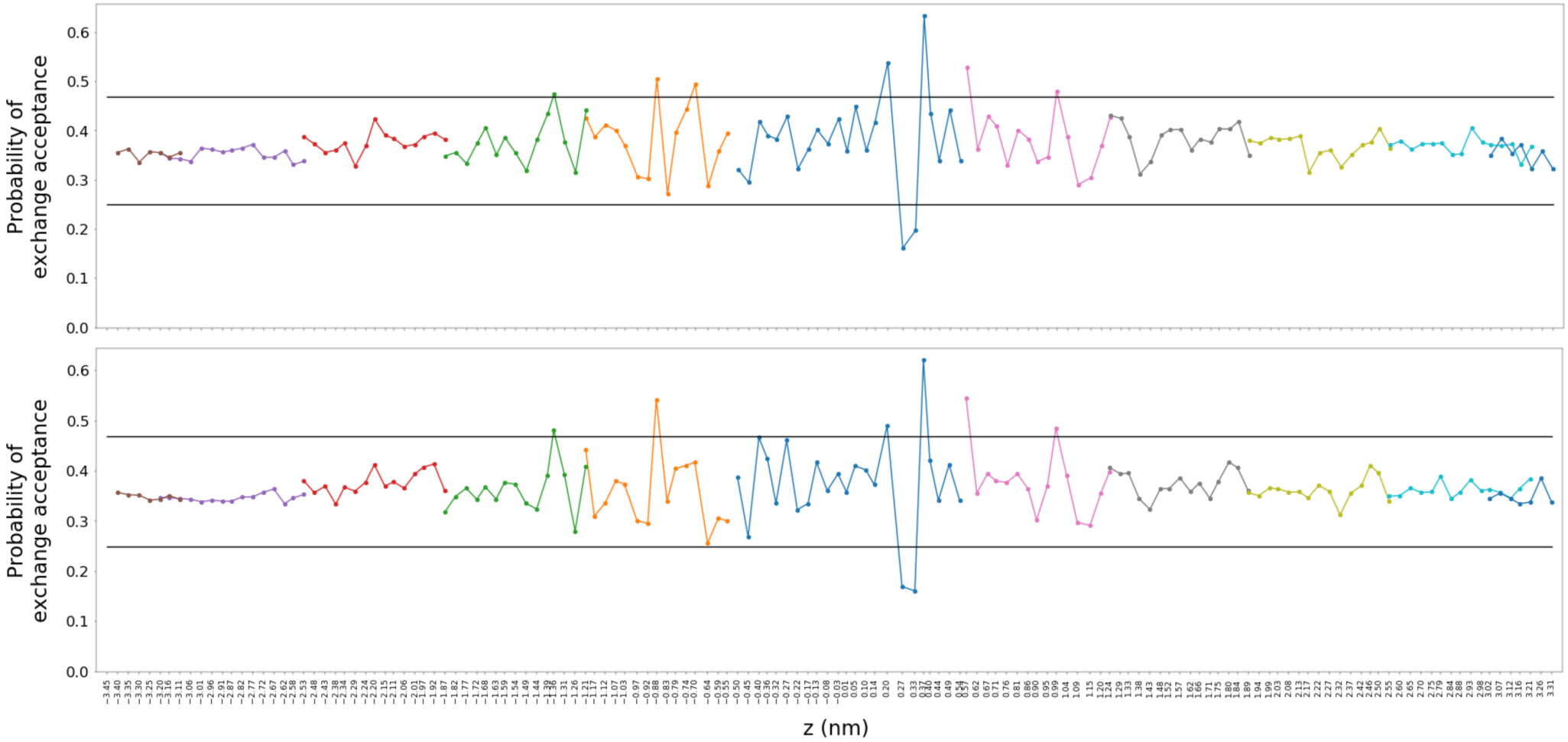

Figure 5.15: Average exchange probabilities between umbrella windows after 200 ns of production simulation for two reverse transition replicates, in orientation 2 - Umbrella windows that were allowed to exchange configurations are depicted with the same color. Each point gives the average probability of exchange between the umbrella window centered at the associated z value and the preceding umbrella window on the z-axis; the first point of each exchange subset is not shown here as it does not have proceeding window in the subset. Probability region in-between the two black lines was the target region when optimizing spacing between umbrella windows.

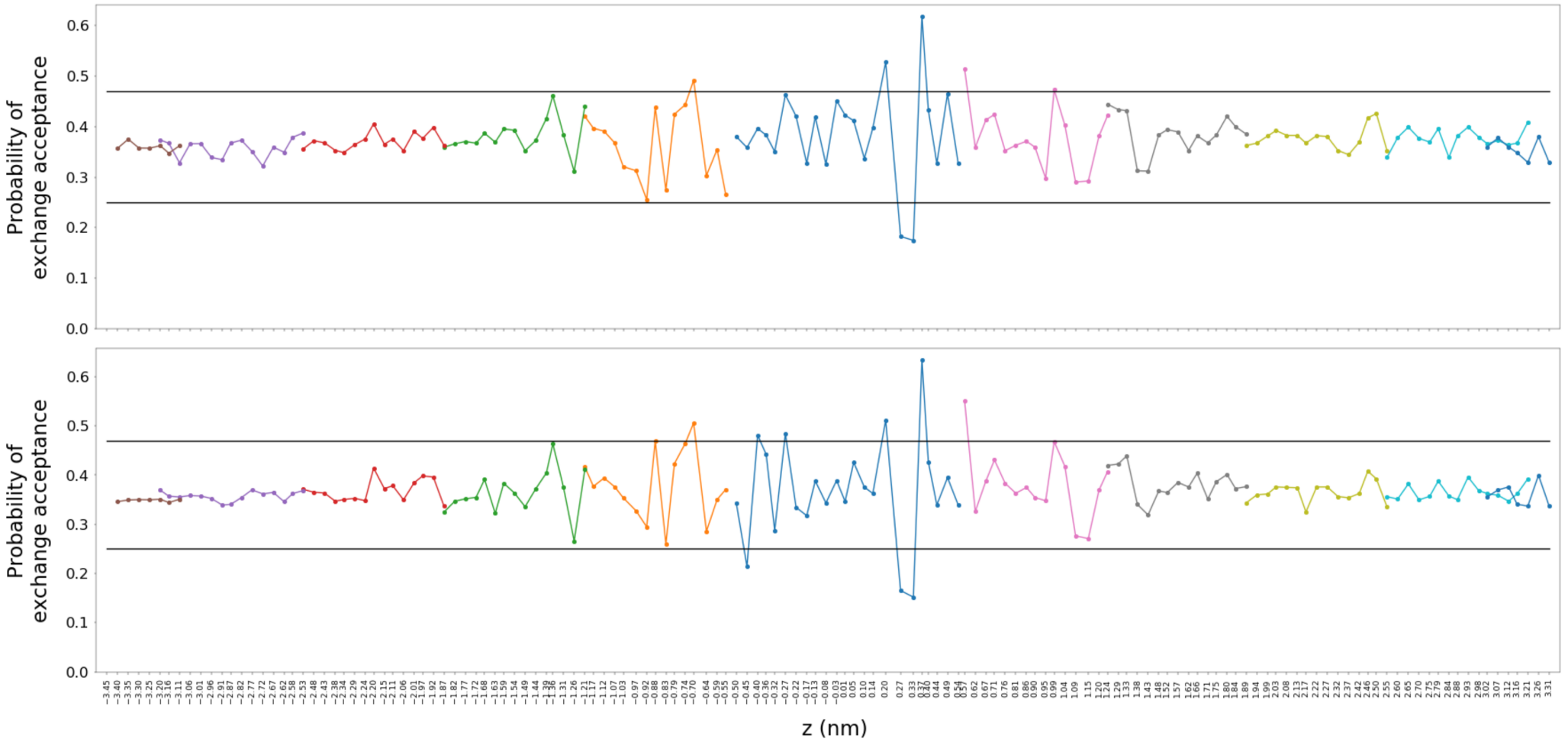

Figure 5.16: Average exchange probabilities between umbrella windows after 200 ns of production simulation for two reverse transition replicates, in orientation 2 - Umbrella windows that were allowed to exchange configurations are depicted with the same color. Each point gives the average probability of exchange between the umbrella window centered at the associated z value and the preceding umbrella window on the z-axis; the first point of each exchange subset is not shown here as it does not have proceeding window in the subset. Probability region in-between the two black lines was the target region when optimizing spacing between umbrella windows.

6

Discussion

Our work on DNA opening during transcription initiation by RNAP II provided the first all-atom simulations of a continuous DNA opening process. From these simulations we revealed qualitative insights into this fundamental process, namely: (i) protein loop–DNA interactions favoring strand separation, and (ii) protein–DNA interactions stabilizing the transcription bubble once formed. Even though the implication of these protein–DNA interactions in RNAP II-dependant eukaryotic transcription initiation lack experimental validations, mutation experiments have shown that some of them are likely involved in archaeal transcription initiation (179). Our work paves the way for future all-atom studies of DNA opening as several aspects of the DNA opening process during transcription remain elusive at the atomic level. For instance, even though TFIIH translocase is known to catalyze DNA opening during transcription (183–186), some studies have suggested that TFIIH-independent DNA opening is achievable (66, 67). Further simulations building upon our proposed approach for studying DNA opening with all-atom simulations could shed light on this particular question.

With the objective to improve our simulations of DNA opening and to obtain quantitative insights of this process, we explored other flavors of umbrella sampling which have been shown to improve configurational sampling, namely: (i) umbrella sampling in combination with Hamiltonian replica-exchange (108, 119, 120) (US-HREX), (ii) simulated tempering-enhanced umbrella sampling (STeUS) (206), and (iii) replica-exchange umbrella sampling (REUS) (108). We have tested these methods on the permeation of fosmidomycin through OprO, a simpler process than DNA opening, yet still challenging from a sampling perspective considering previous work from Golla *et al.* (221). With REUS, we have successfully obtained PMFs of the aforementioned process with sub-kcal mol^{-1} standard errors. This great accomplishment is in line with previous studies demonstrating the efficacy of REUS to obtain PMFs of antibiotics permeation through OmpF (225, 226). Accordingly, REUS seems like a promising approach to support drug discovery endeavors against bacterial infections, some of which have been stated as a priority by the World Health Organization (210).

The three projects presented in this thesis are all connected by a common pitfall:

the sampling challenge, inherent to MD simulations. Until supercomputers like Anton 3 (182) become easily and widely accessible to multi-µs MD projects, overcoming the sampling challenge with ingenious methods inspired by statistical physics will remain central to the field. A horde of enhanced sampling techniques have been developed to this aim, and will continue to flourish in the literature (236). Ultimately, these techniques could be classified according to the amount of prior knowledge that guarantee their success. On one edge of this spectrum we find for instance parallel tempering (105–108) and simulated tempering (121), while on the other edge we find enhanced sampling techniques exclusively based on collective variables like US (96) or metadynamics (97), to name a few. Approaches that do not use any prior knowledge of the studied process might drag simulations to irrelevant part of phase-space; conversely, approaches that solely depend on the definition of collective variables paradoxically require to know, beforehand, the underlying mechanisms of the studied process. An optimal approach might likely lie in the middle of this spectrum, as demonstrated by recent sampling successes in the field (206, 225–227, 237). The burdensome sampling challenge might alternatively be soften with the rise of automated procedures for collective variable design (238), some of which provided great promises for studying the complex world of biomolecules (239–243).

References

[1] Oswald T. Avery, Colin M. Macleod, and Maclyn McCarty. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *Journal of Experimental Medicine*, 79(2):137–158, 2 1944.

[2] F. H.C. Crick, Leslie Barnett, S. Brenner, and R. J. Watts-Tobin. General Nature of the Genetic Code for Proteins. *Nature 1961 192:4809*, 192(4809):1227–1232, 1961.

[3] Crick F. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[4] Charles Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* John Murray, London, 1859.

[5] Richard Dawkins. *The Selfish Gene.* Oxford University Press, New York, 1976.

[6] Tomas Lindahl. Instability and decay of the primary structure of DNA. *Nature*, 362(6422):709–715, 1993.

[7] Kevin Leu, Benedikt Obermayer, Sudha Rajamani, Ulrich Gerland, and Irene A. Chen. The prebiotic evolutionary advantage of transferring genetic information from RNA to DNA. *Nucleic Acids Research*, 39(18):8135–8147, 10 2011.

[8] Tong Ihn Lee and Richard A. Young. Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251, 3 2013.

[9] James E. Bradner, Denes Hnisz, and Richard A. Young. Transcriptional Addiction in Cancer. *Cell*, 168(4):629–643, 2 2017.

[10] Stephin J. Vervoort, Jennifer R. Devlin, Nicholas Kwiatkowski, Mingxing Teng, Nathanael S. Gray, and Ricky W. Johnstone. Targeting transcription cycles in cancer. *Nature Reviews Cancer*, 22(1):5–24, 1 2022.

[11] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 5 2021.

[12] Shao Ru Chen, Yan Dai, Jing Zhao, Ligen Lin, Yitao Wang, and Ying Wang. A mechanistic overview of triptolide and celastrol, natural products from Tripterygium wilfordii Hook F. *Frontiers in Pharmacology*, 9(FEB):104, 2 2018.

[13] Ryan D. Martin, Terence E. Hébert, and Jason C. Tanny. Therapeutic targeting of the general RNA polymerase II transcription machinery. *International Journal of Molecular Sciences*, 21(9), 5 2020.

[14] Olivier Bensaude. Inhibiting eukaryotic transcription: Which compound to choose? How to evaluate its activity? *Transcription*, 2(3):103–108, 2011.

[15] Phoebus Aaron Levene and J. A. Mandel. Über die Konstitution der Thymo-nucleinsäure. *Berichte der Deutschen Chemischen Gesellschaft*, 41:1905–1909, 1908.

[16] Phoebus Aaron Levene and L. W. Bass. *Nucleic acids*, volume 70. The Chemical Catalog Company, Inc., New York, 1931.

[17] P. A. Levene and R. S. Tipson. The ring structure of thymidine. *J. Biol. Chem.*, 109:623–630, 1935.

[18] Albrecht Kossel. *Untersuchungen uber die Nucleine und ihre Spaltungsprodukte.* K.J. Trubner, Strassburg, 1881.

[19] Che-Hung Lee, Hiroshi Mizusawa, and Tsuyoshi Kakefuda. Unwinding of double-stranded DNA helix by dehydration. *Proceedings of the National Academy of Sciences of the United States of America*, 78(5):2838–2842, 1981.

[20] Peter L. Privalov and Colyn Crane-Robinson. Forces maintaining the DNA double helix. *European Biophysics Journal*, 49(5):315–321, 7 2020.

[21] J. D. Watson and F. H.C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature 1953 171:4356*, 171(4356):737–738, 4 1953.

[22] Brenda Maddox. The double helix and the 'wronged heroine'. *Nature 2003 421:6921*, 421(6921):407–408, 1 2003.

[23] E. Chargaff. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. 1950. *Experientia*, 50(4):368–376, 1994.

[24] Erwin Chargaff, Ernst Vischer, Ruth Doniger, Charlotte Green, and Fernanda Misani. The composition of the desoxypentose nucleic acids of thymus and spleen. *The Journal of biological chemistry*, 177(1):405–416, 1 1949.

[25] H. R. Drew, S. Samson, and R. E. Dickerson. Structure of a B-DNA dodecamer at 16 K. *Proceedings of the National Academy of Sciences of the United States of America*, 79(13):4040–4044, 1982.

[26] Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, and Martin Raff. *Molecular Biology of THE CELL.* Number 9. Garland Science, New York, 6th edition, 2014.

[27] Zhichao Miao and Eric Westhof. RNA Structure: Advances and Assessment of 3D Structure Prediction. *Annual Review of Biophysics*, 46:483–503, 5 2017.

[28] Patrick Cramer. Structure and function of RNA polymerase II. *Advances in Protein Chemistry*, 67:1–42, 2004.

[29] Finn Werner and Dina Grohmann. Evolution of multisubunit RNA polymerases in the three domains of life. *Nature Reviews Microbiology*, 9(2):85–98, 2011.

[30] Sarah Sainsbury, Carrie Bernecky, and Patrick Cramer. Structural basis of transcription initiation by RNA polymerase II. *Nature Reviews Molecular Cell Biology*, 16(3):129–143, 2015.

[31] Fei Xavier Chen, Edwin R. Smith, and Ali Shilatifard. Born to run: Control of transcription elongation by RNA polymerase II. *Nature Reviews Molecular Cell Biology*, 19(7):464–478, 7 2018.

[32] Patrick Cramer. Organization and regulation of gene transcription. *Nature*, 573(7772):45–54, 2019.

[33] Sara Osman and Patrick Cramer. Structural Biology of RNA Polymerase II Transcription: 20 Years on. *Annual Review of Cell and Developmental Biology*, 36:1–34, 2020.

[34] Allison C. Schier and Dylan J. Taatjes. Structure and mechanism of the RNA polymerase II transcription machinery. *Genes and Development*, 34(7-8):465–488, 2020.

[35] Mathias Girbig, Agata D Misiaszek, and Christoph W Müller. Structural insights into nuclear transcription by eukaryotic DNA- dependent RNA polymerases. *Nature Reviews Molecular Cell Biology*, 2022.

[36] R. G. Roeder and W. J. Rutter. Specific Nucleolar and Nucleoplasmic RNA Polymerases. *Proceedings of the National Academy of Sciences*, 65(3):675–682, 3 1970.

[37] Carrie Bernecky, Franz Herzog, Wolfgang Baumeister, Jürgen M. Plitzko, and Patrick Cramer. Structure of transcribing mammalian RNA polymerase II. *Nature*, 529(7587):551–554, 1 2016.

[38] F. Gannon, K. O'Hare, F. Perrin, J. P. Lepennec, C. Benoist, M. Cochet, R. Breathnach, A. Royal, A. Garapin, B. Cami, and P. Chambon. Organisation and sequences at the 5 end of a cloned complete ovalbumin gene. *Nature*, 278(5703):428–434, 1979.

[39] W. Chen and K. Struhl. Yeast mRNA initiation sites are determined primarily by specific sequences, not by the distance from the TATA element. *The EMBO Journal*, 4(12):3273–3280, 12 1985.

[40] Thomas W. Burke and James T. Kadonaga. Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes and Development*, 10(6):711–724, 3 1996.

[41] Thomas W. Burke and James T. Kadonaga. The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAF(II)60 of Drosophila. *Genes and Development*, 11(22):3020–3031, 11 1997.

[42] Thierry Lagrange, Achillefs N. Kapanidis, Hong Tang, Danny Reinberg, and Richard H. Ebright. New core promoter element in RNA polymerase II-dependent transcription: Sequence-specific DNA binding by transcription factor IIB. *Genes and Development*, 12(1):34–44, 1 1998.

[43] Wensheng Deng and Stefan G.E. Roberts. A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes and Development*, 19(20):2418–2423, 10 2005.

[44] Tamar Juven-Gershon and James T. Kadonaga. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Developmental Biology*, 339(2):225–229, 3 2010.

[45] Andrew D. Basehoar, Sara J. Zanton, and B. Franklin Pugh. Identification and distinct regulation of yeast TATA box-containing genes. *Cell*, 116(5):699–709, 3 2004.

[46] Piero Carninci, Albin Sandelin, Boris Lenhard, Shintaro Katayama, Kazuro Shimokawa, Jasmina Ponjavic, Colin A.M. Semple, Martin S. Taylor, Pär G. Engström, Martin C. Frith, Alistair R.R. Forrest, Wynand B. Alkema, Sin Lam Tan, Charles Plessy, Rimantas Kodzius, Timothy Ravasi, Takeya Kasukawa, Shiro Fukuda, Mutsumi Kanamori-Katayama, Yayoi Kitazume, Hideya Kawaji, Chikatoshi Kai, Mari Nakamura, Hideaki Konno, Kenji Nakano, Salim Mottagui-Tabar, Peter Arner, Alessandra Chesi, Stefano Gustincich, Francesca Persichetti, Harukazu Suzuki, Sean M. Grimmond, Christine A. Wells, Valerio Orlando, Claes Wahlestedt, Edison T. Liu, Matthias Harbers, Jun Kawai, Vladimir B. Bajic, David A. Hume, and Yoshihide Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, 38(6):626–635, 6 2006.

[47] Long Vo Ngoc, California Jack Cassidy, Cassidy Yunjing Huang, Sascha H.C. Duttke, and James T. Kadonaga. The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. *Genes and Development*, 31(1):6–11, 1 2017.

[48] Long Vo Ngoc, Yuan Liang Wang, George A. Kassavetis, and James T. Kadonaga. The punctilious RNA polymerase II core promoter. *Genes and Development*, 31(13):1289–1301, 7 2017.

[49] Vanja Haberle and Alexander Stark. Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*, 19(10):621–637, 10 2018.

[50] Isaac Fianu, Christian Dienemann, Shintaro Aibara, Sandra Schilbach, and Patrick Cramer. Cryo-EM structure of mammalian RNA polymerase II in complex with human RPAP2. *Communications Biology*, 4(1), 12 2021.

[51] Joseph L. Kim, Dimitar B. Nikolov, and Stephen K. Burley. Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature*, 365(6446):520–527, 1993.

[52] Song Tan, Yvonne Hunziker, David F. Sargent, and Timothy J. Richmond. Crystal structure of a yeast TFIIA/TBP/DNA complex. *Nature*, 381(6578):127–134, 5 1996.

[53] James H. Geiger, Steve Hahn, Sally Lee, and Paul B. Sigler. Crystal structure of the yeast TFIIA/TBP/DNA complex. *Science*, 272(5263):830–836, 1996.

[54] Torill Høiby, Huiqing Zhou, Dimitra J. Mitsiou, and Hendrik G. Stunnenberg. A facelift for the general transcription factor TFIIA. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1769(7-8):429–436, 7 2007.

[55] Kenji Murakami, Guillermo Calero, Christopher R. Brown, Xin Liu, Ralph E. Davis, Hinrich Boeger, and Roger D. Kornberg. Formation and fate of a complete 31-protein RNA polymerase II transcription preinitiation complex. *Journal of Biological Chemistry*, 288(9):6325–6332, 3 2013.

[56] Xuemei Zhao and Winship Herr. A regulated two-step mechanism of TBP binding to DNA: A solvent-exposed surface of TBP inhibits TATA box recognition. *Cell*, 108(5):615–627, 3 2002.

[57] R. G. Roeder. The role of general initiation factors in transcription by RNA polymerase II. *Trends in Biochemical Sciences*, 21(9):327–335, 9 1996.

[58] M T Killeen and J F Greenblatt. The general transcription factor RAP30 binds to RNA polymerase II and prevents it from binding nonspecifically to DNA. *Molecular and Cellular Biology*, 12(1):30–37, 1 1992.

[59] P. Geetha Rani, Jeffrey A. Ranish, and Steven Hahn. RNA Polymerase II (Pol II)-TFIIF and Pol II-Mediator Complexes: the Major Stable Pol II Complexes and Their Activity in Transcription Initiation and Reinitiation. *Molecular and Cellular Biology*, 24(4):1709–1720, 2 2004.

[60] Stephen Buratowski and Hong Zhou. Functional domains of transcription factor TFIIB. *Proceedings of the National Academy of Sciences of the United States of America*, 90(12):5633–5637, 6 1993.

[61] Alcide Barberis, Christoph W. Müller, Stephen C. Harrison, and Mark Ptashne. Delineation of two functional regions of transcription factor TFIIB. *Proceedings of the National Academy of Sciences of the United States of America*, 90(12):5628–5632, 6 1993.

[62] I. Ha, S. Roberts, E. Maldonado, X. Sun, L. U. Kim, M. Green, and D. Reinberg. Multiple functional domains of human transcription factor IIB: Distinct interactions with two general transcription factors and RNA polymerase II. *Genes and Development*, 7(6):1021–1032, 1993.

[63] Yuan He, Jie Fang, Dylan J. Taatjes, and Eva Nogales. Structural visualization of key steps in human transcription initiation. *Nature*, 495(7442):481–486, 3 2013.

[64] Sebastian Grünberg, Linda Warfield, and Steven Hahn. Architecture of the RNA polymerase II preinitiation complex and mechanism of ATP-dependent promoter opening. *Nature Structural and Molecular Biology*, 19(8):788–796, 8 2012.

[65] James Fishburn, Eric Tomko, Eric Galburt, and Steven Hahn. Double-stranded DNA translocase activity of transcription factor TFIIH and the mechanism of RNA polymerase II open complex formation. *Proceedings of the National Academy of Sciences of the United States of America*, 112(13):3961–3966, 3 2015.

[66] Sergey Alekseev, Zita Nagy, Jérémy Sandoz, Amélie Weiss, Jean Marc Egly, Nicolas Le May, and Frederic Coin. Transcription without XPB Establishes a Unified Helicase-Independent Mechanism of Promoter Opening in Eukaryotic Gene Expression. *Molecular Cell*, 65(3):504–514, 2 2017.

[67] Christian Dienemann, Björn Schwalb, Sandra Schilbach, and Patrick Cramer. Promoter Distortion and Opening in the RNA Polymerase II Cleft. *Molecular Cell*, 73(1):97–106, 2019.

[68] Jeffrey D. Parvin and Phillip A. Sharp. DNA topology and a minimal set of basal factors for transcription by RNA polymerase II. *Cell*, 73(3):533–540, 1993.

[69] Young Joon Kim, Stefan Björklund, Yang Li, Michael H. Sayre, and Roger D. Kornberg. A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. *Cell*, 77(4):599–608, 5 1994.

[70] Laure de MERCOYROL, Yves CORDA, Claudette JOB, and Dominique JOB. Accuracy of wheat-germ RNA polymerase II: General enzymatic properties and effect of template conformational transition from right-handed B-DNA to left-handed Z-DNA. *European Journal of Biochemistry*, 206(1):49–58, 1992.

[71] Anand Ramanathan, G. Brett Robb, and Siu Hong Chan. mRNA capping: Biological functions and applications. *Nucleic Acids Research*, 44(16):7511–7526, 9 2016.

[72] Peter Refsing Andersen, Michal Domanski, Maiken S. Kristiansen, Helena Storvall, Evgenia Ntini, Celine Verheggen, Aleks Schein, Jakob Bunkenborg, Ina Poser, Marie Hallais, Rickard Sandberg, Anthony Hyman, John Lacava, Michael P. Rout, Jens S. Andersen, Edouard Bertrand, and Torben Heick Jensen. The human cap-binding complex is functionally connected to the nuclear RNA exosome. *Nature Structural and Molecular Biology*, 20(12):1367–1376, 12 2013.

[73] N. Sonenberg, M. A. Morgan, W. C. Merrick, and A. J. Shatkin. A polypeptide in eukaryotic initiation factors that crosslinks specifically to the 5'-terminal cap in mRNA. *Proceedings of the National Academy of Sciences of the United States of America*, 75(10):4843–4847, 1978.

[74] Maria M. Konarska, Richard A. Padgett, and Phillip A. Sharp. Recognition of cap structure in splicing in vitro of mRNA precursors. *Cell*, 38(3):731–736, 1984.

[75] Xinfu Jiao, Jeong Ho Chang, Turgay Kilic, Liang Tong, and Megerditch Kiledjian. A Mammalian Pre-mRNA 5' End Capping Quality Control Mechanism and an Unexpected Link of Capping to Pre-mRNA Processing. *Molecular Cell*, 50(1):104–115, 4 2013.

[76] Fernando Carrillo Oesterreich, Lydia Herzel, Korinna Straube, Katja Hujer, Jonathon Howard, and Karla M. Neugebauer. Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell*, 165(2):372–381, 4 2016.

[77] Yigong Shi. Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nature Reviews Molecular Cell Biology*, 18(11):655–670, 9 2017.

[78] Odil Porrua and Domenico Libri. Transcription termination and the control of the transcriptome: Why, where and how to stop. *Nature Reviews Molecular Cell Biology*, 16(3):190–202, 3 2015.

[79] Joshua D. Eaton and Steven West. Termination of Transcription by RNA Polymerase II: BOOM! *Trends in Genetics*, 36(9):664–675, 9 2020.

[80] Lori A. Passmore and Jeff Coller. Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression. *Nature Reviews Molecular Cell Biology*, 23(2):93–106, 2 2022.

[81] M. Edmonds and R. Abrams. Polynucleotide biosynthesis: formation of a sequence of adenylate units from adenosine triphosphate by an enzyme from thymus nuclei. *The Journal of biological chemistry*, 235(4):1142–1149, 4 1960.

[82] D. R. Gallie. The cap and poly(A) tail function synergistically to regulate mRNA translational efficiency. *Genes and Development*, 5(11):2108–2116, 1991.

[83] Tim Wilson and Richard Treisman. Removal of poly(A) and consequent degradation of c-fos mRNA facilitated by 3 AU-rich sequences. *Nature*, 336(6197):396–399, 1988.

[84] Andrew R. Leach. *Molecular modelling : principles and applications*. Pearson Prentice Hall, Harlow, 2nd edition, 2001.

[85] Daniel M. Zuckerman. *Statistical Physics of Biomolecules AN INTRODUCTION*. CRC Press, 1st edition, 2010.

[86] Scott M. Shell. *Thermodynamics and Statistical Mechanics*. Number 9. Cambridge University Press, Cambridge, 1st edition, 2015.

[87] GROMACS development team. GROMACS Documentation Release 2022.2, 2022.

[88] R. W. Hockney, S. P. Goel, and J. W. Eastwood. Quiet high-resolution computer models of a plasma. *Journal of Computational Physics*, 14(2):148–158, 2 1974.

[89] K Anton Feenstra, Berk Hess, and Herman J C Berendsen. Improving Efficiency of Large Time-Scale Molecular Dynamics Simulations of Hydrogen-Rich Systems. *Journal of Computational Chemistry*, 20(8), 1999.

[90] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):014101, 1 2007.

[91] H. J.C. Berendsen, J. P.M. Postma, W. F. Van Gunsteren, A. Dinola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, 10 1984.

[92] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182, 8 1998.

[93] Maya Shamir, Yinon Bar-On, Rob Phillips, and Ron Milo. SnapShot: Timescales in Cell Biology. *Cell*, 164(6):1302–1302, 3 2016.

[94] Peter G. Bolhuis, David Chandler, Christoph Dellago, and Phillip L. Geissler. Tansition Path Sampling: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annual Review of Physical Chemistry*, 53(1):291–318, 2002.

[95] Baron Peters. Reaction Coordinates and Mechanistic Hypothesis Tests. *Annual Review of Physical Chemistry*, 67(1):669–690, 5 2016.

[96] Glenn M. Torrie and John P. Valleau. Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid. *Chemical Physics Letters*, 28(4):578–581, 10 1974.

[97] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America*, 2002.

[98] H. Grubmüller, B. Heymann, and P. Tavan. Ligand Binding: Molecular Mechanics Calculation of the Streptavidin-Biotin Rupture Force. *Science*, 271(5251):997–999, 1996.

[99] Jonathan Leech, Jan F. Prins, and Jan Hermans. SMD: Visual steering of molecular dynamics for protein design. *IEEE computational science & engineering*, 3(4):38–45, 1996.

[100] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8):1011–1021, 10 1992.

[101] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Physical Review Letters*, 100(2):020603, 1 2008.

[102] Paolo Raiteri, Alessandro Laio, Francesco Luigi Gervasio, Cristian Micheletti, and Michele Parrinello. Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *Journal of Physical Chemistry B*, 110(8):3533–3539, 2006.

[103] Cristian Micheletti, Alessandro Laio, and Michele Parrinello. Reconstructing the Density of States by History-Dependent Metadynamics. *The American Physical Society*, 92(1), 2004.

[104] Timo M. Schäfer and Giovanni Settanni. Data Reweighting in Metadynamics Simulations. *Journal of Chemical Theory and Computation*, 16(4):2042–2052, 4 2020.

[105] Robert H. Swendsen and Jian Sheng Wang. Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607–2609, 1986.

[106] Koji Hukushima and Koji Nemoto. Exchange Monte Carlo Method and Application to Spin Glass Simulations. *Journal of the Physical Society of Japan*, 65:1604–1608, 1996.

[107] Ulrich H.E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters*, 281(1-3):140–150, 12 1997.

[108] Yuji Sugita, Akio Kitao, and Yuko Okamoto. Multidimensional replica-exchange method for free-energy calculations. *The Journal of Chemical Physics*, 113(15):6042–6051, 10 2000.

[109] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[110] Ruhong Zhou, Bruce J. Berne, and Robert Germain. The free energy landscape for β hairpin folding in explicit water. *Proceedings of the National Academy of Sciences of the United States of America*, 98(26):14931–14936, 12 2001.

[111] K. Y. Sanbonmatsu and A. E. García. Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. *Proteins: Structure, Function and Genetics*, 46(2):225–234, 2 2002.

[112] Jed W. Pitera and William Swope. Understanding folding and design: Replica-exchange simulations of "Trp-cage" miniproteins. *Proceedings of the National Academy of Sciences of the United States of America*, 100(13):7587–7592, 6 2003.

[113] Francesco Rao and Amedeo Caflisch. Replica exchange molecular dynamics simulations of reversible folding. *Journal of Chemical Physics*, 119(7):4035–4042, 8 2003.

[114] Angel E. García and José N. Onuchic. Folding a protein in a computer: An atomic description of the folding/unfolding of protein A. *Proceedings of the National Academy of Sciences of the United States of America*, 100(SUPPL. 2):13898–13903, 11 2003.

[115] Ruhong Zhou. Trp-cage: Folding free energy landscape in explicit water. *Proceedings of the National Academy of Sciences*, 100(23):13280–13285, 11 2003.

[116] M. Cecchini, F. Rao, M. Seeber, and A. Caflisch. Replica exchange molecular dynamics simulations of amyloid peptide aggregation. *Journal of Chemical Physics*, 121(21):10748–10756, 12 2004.

[117] Dietmar Paschek and Angel E. García. Reversible temperature and pressure denaturation of a protein fragment: A replica exchange molecular dynamics simulation study. *Physical Review Letters*, 93(23):238105, 12 2004.

[118] Daniel Sindhikara, Yilin Meng, and Adrian E. Roitberg. Exchange frequency in replica exchange molecular dynamics. *Journal of Chemical Physics*, 128(2), 2008.

[119] Lingle Wang, Richard A. Friesner, and B. J. Berne. Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2). *Journal of Physical Chemistry B*, 115(30):9431–9438, 8 2011.

[120] Giovanni Bussi. Hamiltonian replica exchange in GROMACS: A flexible implementation. *Molecular Physics*, 112(3-4):379–384, 2014.

[121] E. Marinari and G. Parisi. Simulated tempering: A New Monte Carlo Scheme. *EPL*, 19(6):451–458, 7 1992.

[122] Jean Paul Ryckaert, Giovanni Ciccotti, and Herman J.C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, 3 1977.

[123] Shuichi Miyamoto and Peter A. Kollman. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry*, 13(8):952–962, 10 1992.

[124] Berk Hess, Henk Bekker, Herman J C Berendsen, and Johannes G E M Fraaije. LINCS: A Linear Constraint Solver for Molecular Simulations. *J Comput Chem*, 18:14631472, 1997.

[125] Chad W. Hopkins, Scott Le Grand, Ross C. Walker, and Adrian E. Roitberg. Long-time-step molecular dynamics through hydrogen mass repartitioning. *Journal of Chemical Theory and Computation*, 11(4):1864–1874, 4 2015.

[126] Curtis Balusek, Hyea Hwang, Chun Hon Lau, Karl Lundquist, Anthony Hazel, Anna Pavlova, Diane L. Lynch, Patricia H. Reggio, Yi Wang, and James C. Gumbart. Accelerating Membrane Simulations with Hydrogen Mass Repartitioning. *Journal of Chemical Theory and Computation*, 15(8):4673–4686, 8 2019.

[127] S. K. Kearsley. On the orthogonal transformation used for structural comparisons. *Acta Crystallographica Section A*, 45(2):208–210, 2 1989.

[128] Nilesh K. Banavali and Benoît Roux. Free energy landscape of A-DNA to B-DNA conversion in aqueous solution. *Journal of the American Chemical Society*, 127(18):6866–6876, 2005.

[129] Pavel I Zhuravlev, Sangwook Wu, Davit A Potoyan, Michael Rubinstein, and Garegin A Papoian. Computing free energies of protein conformations from explicit solvent simulations. *Methods (San Diego, Calif.)*, 52(1):115–21, 9 2010.

[130] Weinan E and Eric Vanden-Eijnden. Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events. *Annual Review of Physical Chemistry*, 61(1):391–420, 2010.

[131] Luca Maragliano, Alexander Fischer, Eric Vanden-Eijnden, and Giovanni Ciccotti. String method in collective variables: Minimum free energy paths and isocommittor surfaces. *Journal of Chemical Physics*, 125(2):1–15, 2006.

[132] Albert C. Pan, Deniz Sezer, and Benoît Roux. Finding transition pathways using the string method with swarms of trajectories. *Journal of Physical Chemistry B*, 112(11):3432–3440, 3 2008.

[133] Luca Maragliano, Benoiît Roux, and Eric Vanden-Eijnden. Comparison between mean forces and swarms-of-trajectories string methods. *Journal of Chemical Theory and Computation*, 10(2):524–533, 2014.

[134] Davide Branduardi and José D. Faraldo-Gómez. String Method for Calculation of Minimum Free-Energy Paths in Cartesian Space in Freely Tumbling Systems. *Journal of Chemical Theory and Computation*, 9(9):4140–4154, 9 2013.

[135] Davide Branduardi, Francesco Luigi Gervasio, and Michele Parrinello. From A to B in free energy space. *The Journal of Chemical Physics*, 126(5):054103, 2 2007.

[136] Wenxun Gan, Sichun Yang, and Benoît Roux. Atomistic view of the conformational activation of Src kinase using the string method with swarms-of-trajectories. *Biophysical Journal*, 97(4):L8, 8 2009.

[137] Anna Berteotti, Andrea Cavalli, Davide Branduardi, Francesco Luigi Gervasio, Maurizio Recanatini, and Michele Parrinello. Protein conformational transitions: The closure mechanism of a kinase explored by atomistic simulations. *Journal of the American Chemical Society*, 131(1):244–250, 2009.

[138] Victor Ovchinnikov, Martin Karplus, and Eric Vanden-Eijnden. Free energy of conformational transition paths in biomolecules: The string method and its application to myosin VI. *The Journal of Chemical Physics*, 134(8):085103, 2 2011.

[139] Yasuhiro Matsunaga, Hiroshi Fujisaki, Tohru Terada, Tadaomi Furuta, Kei Moritsugu, and Akinori Kidera. Minimum free energy path of ligand-induced transition in Adenylate kinase. *PLoS Computational Biology*, 8(6), 6 2012.

[140] Yong Wang, Elena Papaleo, and Kresten Lindorff-Larsen. Mapping transiently formed and sparsely populated conformations on a complex energy landscape. *eLife*, 5(AUGUST):1–35, 2016.

[141] Avisek Das, Huan Rui, Robert Nakamoto, and Benoît Roux. Conformational Transitions and Alternating-Access Mechanism in the Sarcoplasmic Reticulum Calcium Pump. *Journal of Molecular Biology*, 429(5):647–666, 3 2017.

[142] Gareth A. Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New feathers for an old bird. *Computer Physics Communications*, 185(2):604–613, 2014.

[143] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function and Bioinformatics*, 78(8):1950–1958, 6 2010.

[144] Berk Hess, Henk Bekker, Herman J C Berendsen, and Johannes G E M Fraaije. LINCS: A Linear Constraint Solver for Molecular Simulations. *J Comput Chem*, 18:14631472, 1997.

[145] Jeremy Lapierre and Jochen S. Hub. DNA opening during transcription initiation by RNA polymerase II in atomic detail. *Biophysical Journal*, 121:4299–4310, 11 2022.

[146] P. Cramer, D. A. Bushnell, and R. D. Kornberg. Structural basis of transcription: RNA polymerase II at 2.8 ångstrom resolution. *Science*, 292(5523):1863–1876, 2001.

[147] A. L. Gnatt, P. Cramer, J. Fu, D. A. Bushnell, and R. D. Kornberg. Structural basis of transcription: An RNA polymerase II elongation complex at 3.3 Å resolution. *Science*, 292(5523):1876–1882, 6 2001.

[148] Eva Nogales, Robert K. Louder, and Yuan He. Structural Insights into the Eukaryotic Transcription Initiation Machinery. *Annual Review of Biophysics*, 46:59–83, 2017.

[149] Kenji Murakami, Kuang Lei Tsai, Nir Kalisman, David A. Bushnell, Francisco J. Asturias, and Roger D. Kornberg. Structure of an RNA polymerase II preinitiation complex. *Proceedings of the National Academy of Sciences of the United States of America*, 112(44):13543–13548, 11 2015.

[150] C. Plaschka, L. Larivière, L. Wenzeck, M. Seizl, M. Hemann, D. Tegunov, E. V. Petrotchenko, C. H. Borchers, W. Baumeister, F. Herzog, E. Villa, and P. Cramer. Architecture of the RNA polymerase II-Mediator core initiation complex. *Nature*, 518(7539):376–380, 2015.

[151] Yuan He, Chunli Yan, Jie Fang, Carla Inouye, Robert Tjian, Ivaylo Ivanov, and Eva Nogales. Near-atomic resolution visualization of human transcription promoter opening. *Nature*, 533:359–365, 5 2016.

[152] C. Plaschka, M. Hantsche, C. Dienemann, C. Burzinski, J. Plitzko, and P. Cramer. Transcription initiation complex structures elucidate DNA opening. *Nature*, 533:353–358, 5 2016.

[153] S. Schilbach, M. Hantsche, D. Tegunov, C. Dienemann, C. Wigge, H. Urlaub, and P. Cramer. Structures of transcription pre-initiation complex with TFIIH and Mediator. *Nature*, 551(7679):204–209, 2017.

[154] Chunli Yan, Thomas Dodd, Yuan He, John A Tainer, Susan E Tsutakawa, and Ivaylo Ivanov. Transcription preinitiation complex structure and dynamics provide insight into genetic diseases. *Nature Structural & Molecular Biology*, 26:397–406, 2019.

[155] Sandra Schilbach, Shintaro Aibara, Christian Dienemann, Frauke Grabbe, and Patrick Cramer. Structure of RNA polymerase II pre-initiation complex at 2.9 Å defines initial DNA opening. *Cell*, 184(15):4064–4072, 7 2021.

[156] Shintaro Aibara, Sandra Schilbach, and Patrick Cramer. Structures of mammalian RNA polymerase II pre-initiation complexes. *Nature*, 594(7861):124–128, 2021.

[157] James Chen, Courtney Chiu, Saumya Gopalkrishnan, Albert Y. Chen, Paul Dominic B. Olinares, Ruth M. Saecker, Jared T. Winkelman, Michael F. Maloney, Brian T. Chait, Wilma Ross, Richard L. Gourse, Elizabeth A. Campbell, and Seth A. Darst. Stepwise Promoter Melting by Bacterial RNA Polymerase. *Molecular Cell*, 78(2):275–288, 2020.

[158] Xuhui Huang, Dong Wang, Dahlia R. Weiss, David A. Bushnell, Roger D. Kornberg, and Michael Levitt. RNA polymerase II trigger loop residues stabilize and position the incoming nucleotide triphosphate in transcription. *Proceedings of the National Academy of Sciences of the United States of America*, 107(36):15745–15750, 9 2010.

[159] Michael Feig and Zachary F. Burton. RNA polymerase II with open and closed trigger loops: Active site dynamics and nucleic acid translocation. *Biophysical Journal*, 99(8):2577–2586, 10 2010.

[160] Lin-Tai Da, Dong Wang, and Xuhui Huang. Dynamics of Pyrophosphate Ion Release and Its Coupled Trigger Loop Motion from Closed to Open State in RNA Polymerase II. *J. Am. Chem.Soc*, 134:11, 2012.

[161] Beibei Wang, Alexander V. Predeus, Zachary F. Burton, and Michael Feig. Energetic and structural details of the trigger-loop closing transition in RNA polymerase II. *Biophysical Journal*, 105(3):767–775, 8 2013.

[162] Jin Yu, Lu Bai, Michelle D Wang, Yong-Shun Song, Yao-Gen Shu, Xin Zhou, Lin-Tai Da, and Xuhui Huang. Constructing kinetic models to elucidate structural dynamics of a complete RNA polymerase II elongation cycle. *Physical Biology*, 12(1):016004, 12 2014.

[163] Daniel Adriano Silva, Dahlia R. Weiss, Fátima Pardo Avila, Lin Tai Da, Michael Levitt, Dong Wang, and Xuhui Huang. Millisecond dynamics of RNA polymerase II translocation at atomic resolution. *Proceedings of the National Academy of Sciences of the United States of America*, 111(21):7665–7670, 5 2014.

[164] Lu Zhang, Daniel Adriano Silva, Fátima Pardo-Avila, Dong Wang, and Xuhui Huang. Structural Model of RNA Polymerase II Elongation Complex with Complete Transcription Bubble Reveals NTP Entry Routes. *PLoS computational biology*, 11(7), 7 2015.

[165] Ilona Christy Unarta, Lizhe Zhu, Carmen Ka Man Tse, Peter Pak Hang Cheung, Jin Yu, and Xuhui Huang. Molecular mechanisms of RNA polymerase II transcription elongation elucidated by kinetic network models. *Current Opinion in Structural Biology*, 49:54–62, 4 2018.

[166] Carmen Ka Man Tse, Jun Xu, Liang Xu, Fu Kit Sheong, Shenglong Wang, Hoi Yee Chow, Xin Gao, Xuechen Li, Peter Pak Hang Cheung, Dong Wang, Yingkai Zhang, and Xuhui Huang. Intrinsic cleavage of RNA polymerase II adopts a nucleobase-independent mechanism assisted by transcript phosphate. *Nature Catalysis*, 2(3):228–235, 2 2019.

[167] Ilona Christy Unarta, Siqin Cao, Shintaroh Kubo, Wei Wang, Peter Pak-Hang Cheung, Xin Gao, Shoji Takada, and Xuhui Huang. Role of bacterial RNA polymerase gate opening dynamics in DNA loading and antibiotics inhibition elucidated by quasi-Markov State Model. *Proceedings of the National Academy of Sciences of the United States of America*, 2021.

[168] Genki Shino and Shoji Takada. Modeling DNA Opening in the Eukaryotic Transcription Initiation Complexes via Coarse-Grained Models. *Frontiers in Molecular Biosciences*, 8(November):1–12, 2021.

[169] Davide Branduardi, Francesco Luigi Gervasio, and Michele Parrinello. From A to B in free energy space. *The Journal of Chemical Physics*, 126(5):054103, 2 2007.

[170] Mahadeb Pal, Alfred S. Ponticelli, and Donal S. Luse. The Role of the Transcription Bubble and TFIIB in Promoter Clearance by RNA Polymerase II. *Molecular Cell*, 19(1):101–110, 7 2005.

[171] Filip Lankaš, Richard Lavery, and John H. Maddocks. Kinking Occurs during Molecular Dynamics Simulations of Small DNA Minicircles. *Structure*, 14(10):1527–1534, 2006.

[172] Graham L. Randall, Lynn Zechiedrich, and B. Montgomery Pettitt. In the absence of writhe, DNA relieves torsional stress with localized, sequence-dependent structural failure to preserve B-form. *Nucleic Acids Research*, 37(16):5568–5577, 2009.

[173] J. S. Mitchell, C. A. Laughton, and Sarah A. Harris. Atomistic simulations reveal bubbles, kinks and wrinkles in supercoiled DNA. *Nucleic Acids Research*, 39(9):3928–3938, 2011.

[174] Rossitza N. Irobalieva, Jonathan M. Fogg, Daniel J. Catanese, Thana Sutthibutpong, Muyuan Chen, Anna K. Barker, Steven J. Ludtke, Sarah A. Harris, Michael F. Schmid, Wah Chiu, and Lynn Zechiedrich. Structural diversity of supercoiled DNA. *Nature Communications*, 6, 10 2015.

[175] Julie D Thompson, Desmond G Higgins+, and Toby J Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680, 1994.

[176] Andrey Feklistov and Seth A. Darst. Structural basis for promoter -10 element recognition by the bacterial RNA polymerase σ subunit. *Cell*, 147(6):1257–1269, 2011.

[177] Saulius Klimasauskas, Sanjay Kumar, Richard J. Roberts, and Xiaodong Cheng. Hhal methyltransferase flips its target base out of the DNA helix. *Cell*, 76(2):357–369, 1994.

[178] Niu Huang, Nilesh K. Banavali, and Alexander D. MacKerell. Protein-facilitated base flipping in DNA by cytosine-5-methyltransferase. *Proceedings of the National Academy of Sciences of the United States of America*, 100(1):68–73, 2003.

[179] Souad Naji, Michela G. Bertero, Patrizia Spitalny, Patrick Cramer, and Michael Thomm. Structure-function analysis of the RNA polymerase cleft loops elucidates initial transcription, DNA unwinding and RNA displacement. *Nucleic Acids Research*, 36(2):676–687, 2008.

[180] Christopher O. Barnes, Monica Calero, Indranil Malik, Brian W. Graham, Henrik Spahr, Guowu Lin, Aina E. Cohen, Ian S. Brown, Qiangmin Zhang, Filippo Pullara, Michael A. Trakselis, Craig D. Kaplan, and Guillermo Calero. Crystal Structure of a Transcribing RNA Polymerase II Complex Reveals a Complete Transcription Bubble. *Molecular Cell*, 59(2):258–269, 2015.

[181] Hiroaki Fukunishi, Osamu Watanabe, and Shoji Takada. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *The Journal of Chemical Physics*, 116(20):9058, 5 2002.

[182] David E. Shaw, Peter J. Adams, Asaph Azaria, Joseph A. Bank, Brannon Batson, Alistair Bell, Michael Bergdorf, Jhanvi Bhatt, J. Adam Butts, Timothy Correi, Robert M. Dirks, Ron O. Dror, Michael P. Eastwoo, Bruce Edwards, Amos Even, Peter Feldmann, Michael Fenn, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Maria Gorlatova, Brian Greskamp, J. P. Grossman, Justin Gullingsrud, Anissa Harper, William Hasenplaugh, Mark Heily, Benjamin Colin Heshmat, Jeremy Hunt, Douglas J. Ierardi,

Lev Iserovich, Bryan L. Jackson, Nick P. Johnson, Mollie M. Kirk, John L. Klepeis, Jeffrey S. Kuskin, Kenneth M. Mackenzie, Roy J. Mader, Richard McGowen, Adam McLaughlin, Mark A. Moraes, Mohamed H. Nasr, Lawrence J. Nociolo, Lief O'Donnell, Andrew Parker, Jon L. Peticolas, Goran Pocina, Cristian Predescu, Terry Quan, John K. Salmon, Carl Schwink, Keun Sup Shim, Naseer Siddique, Jochen Spengler, Tamas Szalay, Raymond Tabladillo, Reinhard Tartler, Andrew G. Taube, Michael Theobald, Brian Towles, William Vick, Stanley C. Wang, Michael Wazlowski, Madeleine J. Weingarten, John M. Williams, and Kevin A. Yuh. Anton 3: Twenty Microseconds of Molecular Dynamics Simulation before Lunch. *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, 2021.

[183] Frank C.P. Holstege, Ulrike Fiedler, and H. Th Marc Timmers. Three transitions in the RNA polymerase II transcription complex during initiation. *EMBO Journal*, 16(24):7468–7480, 1997.

[184] Tae Kyung Kim, Richard H. Ebright, and Danny Reinberg. Mechanism of ATP-dependent promoter melting by transcription factor IIH. *Science*, 288(5470):1418–1421, 2000.

[185] Sebastian Grünberg, Linda Warfield, and Steven Hahn. Architecture of the RNA polymerase II preinitiation complex and mechanism of ATP-dependent promoter opening. *Nature Structural & Molecular Biology 2012 19:8*, 19(8):788–796, 7 2012.

[186] James Fishburn, Eric Tomko, Eric Galburt, and Steven Hahn. Double-stranded DNA translocase activity of transcription factor TFIIH and the mechanism of RNA polymerase II open complex formation. *Proceedings of the National Academy of Sciences of the United States of America*, 112(13):3961–3966, 2015.

[187] L. F. Liu and J. C. Wang. Supercoiling of the DNA template during transcription. *Proceedings of the National Academy of Sciences of the United States of America*, 84(20):7024–7027, 1987.

[188] Fedor Kouzine, Suzanne Sanford, Zichrini Elisha-Feil, and David Levens. The functional response of upstream DNA to dynamic supercoiling in vivo. *Nature Structural and Molecular Biology*, 15(2):146–154, 2 2008.

[189] Catherine Naughton, Nicolaos Avlonitis, Samuel Corless, James G. Prendergast, Ioulia K. Mati, Paul P. Eijk, Scott L. Cockroft, Mark Bradley, Bauke Ylstra, and Nick Gilbert. Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nature Structural and Molecular Biology*, 20(3):387–395, 3 2013.

[190] Fedor Kouzine, Ashutosh Gupta, Laura Baranello, Damian Wojtowicz, Khadija Ben-Aissa, Juhong Liu, Teresa M. Przytycka, and David Levens. Transcription-dependent dynamic supercoiling is a short-range genomic force. *Nature Structural and Molecular Biology*, 20(3):396–403, 3 2013.

[191] Sheila S. Teves, Christopher M. Weber, and Steven Henikoff. Transcribing through the nucleosome. *Trends in Biochemical Sciences*, 39(12):577–586, 12 2014.

[192] Samuel Corless and Nick Gilbert. Effects of DNA supercoiling on chromatin architecture. *Biophysical Reviews*, 8(3):245–258, 9 2016.

[193] Dirk Kostrewa, Mirijam E. Zeller, Karim Jean Armache, Martin Seizl, Kristin Leike, Michael Thomm, and Patrick Cramer. RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature*, 462(7271):323–330, 2009.

[194] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindah. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, 9 2015.

[195] Massimiliano Bonomi, Giovanni Bussi, Carlo Camilloni, Gareth A. Tribello, Pavel Banáš, Alessandro Barducci, Mattia Bernetti, Peter G. Bolhuis, Sandro Bottaro, Davide Branduardi, Riccardo Capelli, Paolo Carloni, Michele Ceriotti, Andrea Cesari, Haochuan Chen, Wei Chen, Francesco Colizzi, Sandip De, Marco De La Pierre, Davide Donadio, Viktor Drobot, Bernd Ensing, Andrew L. Ferguson, Marta Filizola, James S. Fraser, Haohao Fu, Piero Gasparotto, Francesco Luigi Gervasio, Federico Giberti, Alejandro Gil-Ley, Toni Giorgino, Gabriella T. Heller, Glen M. Hocky, Marcella Iannuzzi, Michele Invernizzi, Kim E. Jelfs, Alexander Jussupow, Evgeny Kirilin, Alessandro Laio, Vittorio Limongelli, Kresten Lindorff-Larsen, Thomas Löhr, Fabrizio Marinelli, Layla Martin-Samos, Matteo Masetti, Ralf Meyer, Angelos Michaelides, Carla Molteni, Tetsuya Morishita, Marco Nava, Cristina Paissoni, Elena Papaleo, Michele Parrinello, Jim Pfaendtner, Pablo Piaggi, Giovanni Maria Piccini, Adriana Pietropaolo, Fabio Pietrucci, Silvio Pipolo, Davide Provasi, David Quigley, Paolo Raiteri, Stefano Raniolo, Jakub Rydzewski, Matteo Salvalaglio, Gabriele Cesare Sosso, Vojtěch Spiwok, Jiří Šponer, David W.H. Swenson, Pratyush Tiwary, Omar Valsson, Michele Vendruscolo, Gregory A. Voth, and Andrew White. Promoting transparency and reproducibility in enhanced molecular simulations. *Nature Methods 2019 16:8*, 16(8):670–673, 7 2019.

[196] Krieger E and Vriend G. YASARA View - molecular graphics for all devices - from smartphones to workstations. *Bioinformatics (Oxford, England)*, 30(20):2981–2982, 10 2014.

[197] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 1983.

[198] Marie Zgarbová, Jiří Šponer, Michal Otyepka, Thomas E. Cheatham, Rodrigo Galindo-Murillo, and Petr Jurečka. Refinement of the Sugar-Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *Journal of Chemical Theory and Computation*, 11(12):5723–5736, 11 2015.

[199] James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, 8 2015.

[200] Marina Macchiagodena, Marco Pagliai, Claudia Andreini, Antonio Rosato, and Piero Procacci. Upgrading and Validation of the AMBER Force Field for Histidine and Cysteine Zinc(II)-Binding Residues in Sites with Four Protein Ligands. *Journal of Chemical Information and Modeling*, 59(9):3803–3816, 2019.

[201] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993.

[202] L L C Schrödinger and Warren DeLano. The PyMOL Molecular Graphics System, Version˜2.3.0.

[203] William Humphrey, Andrew Dalke, and Klaus Schulten. {VMD} – {V}isual {M}olecular {D}ynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[204] Alex Bateman, Maria Jesus Martin, Sandra Orchard, Michele Magrane, Rahat Agivetova, Shadab Ahmad, Emanuele Alpi, Emily H. Bowler-Barnett, Ramona Britto, Borisas Bursteinas, Hema Bye-A-Jee, Ray Coetzee, Austra Cukura, Alan Da Silva, Paul Denny, Tunca Dogan, Thank God Ebenezer, Jun Fan, Leyla Garcia Castro, Penelope Garmiri, George Georghiou, Leonardo Gonzales, Emma Hatton-Ellis, Abdulrahman Hussein, Alexandr Ignatchenko, Giuseppe Insana, Rizwan Ishtiaq, Petteri Jokinen, Vishal Joshi, Dushyanth Jyothi, Antonia Lock, Rodrigo Lopez, Aurelien Luciani, Jie Luo, Yvonne Lussi, Alistair MacDougall, Fabio Madeira, Mahdi Mahmoudy, Manuela Menchi, Alok Mishra, Katie Moulang, Andrew Nightingale, Carla Susana Oliveira, Sangya Pundir, Guoying Qi, Shriya Raj, Daniel Rice, Milagros Rodriguez Lopez, Rabie Saidi, Joseph Sampson, Tony Sawford, Elena Speretta, Edward Turner, Nidhi Tyagi, Preethi Vasudev, Vladimir Volynkin, Kate Warner, Xavier Watkins, Rossana Zaru, Hermann Zellner, Alan Bridge, Sylvain Poux, Nicole Redaschi, Lucila Aimo, Ghislaine Argoud-Puy, Andrea Auchincloss, Kristian Axelsen, Parit Bansal, Delphine Baratin, Marie Claude Blatter, Jerven Bolleman, Emmanuel Boutet, Lionel Breuza, Cristina Casals-Casas, Edouard de Castro, Kamal Chikh Echioukh, Elisabeth Coudert, Beatrice Cuche, Mikael Doche, Dolnide Dornevil, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Guillaume Keller, Arnaud Kerhornou, Vicente Lara, Philippe Le Mercier, Damien Lieberherr, Thierry Lombardot, Xavier Martin, Patrick Masson, Anne Morgat, Teresa Batista Neto, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Lucille Pourcel, Monica Pozzato, Manuela Pruess, Catherine Rivoire, Christian Sigrist, Karin Sonesson, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Laure Verbregue, Cathy H. Wu, Cecilia N. Arighi, Leslie Arminski, Chuming Chen, Yongxing Chen, John S. Garavelli, Hongzhan Huang, Kati Laiho, Peter McGarvey, Darren A. Natale, Karen Ross, C. R. Vinayaka, Qinghua Wang, Yuqi Wang, Lai Su Yeh, and Jian Zhang. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 2021.

[205] Jeremy Lapierre and Jochen S. Hub. Converging pmf calculations of antibiotic permeation across an outer membrane porin with subkilocalorie per mole accuracy. *Journal of Chemical Information and Modeling*, 8 2023.

[206] Carla F. Sousa, Robert A. Becker, Claus-Michael Lehr, Olga V. Kalinina, and Jochen S. Hub. Simulated tempering-enhanced umbrella sampling improves convergence of free energy calculations of drug membrane permeation. *bioRxiv*, page 2022.11.13.516136, 11 2022.

[207] A. Cross, J. R. Allen, J. Burke, G. Ducel, A. Harris, J. John, D. Johnson, M. Lew, B. MacMillan, and P. Meers. Nosocomial infections due to Pseudomonas aeruginosa: review of recent trends. *Reviews of infectious diseases*, 5 Suppl 5, 1983.

[208] Hossein Fazzeli, Reza Akbari, Sharareh Moghim, Tahmineh Narimani, Mohammad Reza Arabestani, and Ali Reza Ghoddousi. Pseudomonas aeruginosa infections in patients, hospital means, and personnel's specimens. *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences*, 17(4):332, 2012.

[209] Shugang Qin, Wen Xiao, Chuanmin Zhou, Qinqin Pu, Xin Deng, Lefu Lan, Haihua Liang, Xiangrong Song, and Min Wu. Pseudomonas aeruginosa: pathogenesis, virulence factors, antibiotic resistance, interaction with host, technology advances and emerging therapeutics. *Signal Transduction and Targeted Therapy 2022 7:1*, 7(1):1–27, 6 2022.

[210] George L. Daikos, Clóvis Arns da Cunha, Gian Maria Rossolini, Gregory G. Stone, Nathalie Baillon-Plot, Margaret Tawadrous, and Paurus Irani. Review of Ceftazidime-Avibactam for the Treatment of Infections Caused by Pseudomonas aeruginosa. *Antibiotics*, 10(9), 9 2021.

[211] Roland Benz. Permeation of hydrophilic solutes through mitochondrial outer membranes: review on mitochondrial porins. *Biochimica et biophysica acta*, 1197(2):167–196, 6 1994.

[212] Hiroshi Nikaido. Molecular basis of bacterial outer membrane permeability revisited. *Microbiology and molecular biology reviews : MMBR*, 67(4):593–656, 12 2003.

[213] Claudio Piselli and Roland Benz. Fosmidomycin transport through the phosphate-specific porins OprO and OprP of Pseudomonas aeruginosa. *Molecular Microbiology*, 116(1):97–108, 7 2021.

[214] P. A. Lambert. Cellular impermeability and uptake of biocides and antibiotics in Gram-positive bacteria and mycobacteria. *Journal of Applied Microbiology*, 92(1):46S–54S, 2002.

[215] Shaan L. Gellatly and Robert E.W. Hancock. Pseudomonas aeruginosa : new insights into pathogenesis and host defenses. *Pathogens and Disease*, 67(3):159–173, 4 2013.

[216] Sylvie Chevalier, Emeline Bouffartigues, Josselin Bodilis, Olivier Maillot, Olivier Lesouhaitier, Marc G.J. Feuilloley, Nicole Orange, Alain Dufour, and Pierre Cornelis. Structure, function and regulation of Pseudomonas aeruginosa porins. *FEMS microbiology reviews*, 41(5):698–722, 9 2017.

[217] Julia Vergalli, Igor V. Bodrenko, Muriel Masi, Lucile Moynié, Silvia Acosta-Gutiérrez, James H. Naismith, Anne Davin-Regli, Matteo Ceccarelli, Bert van den

Berg, Mathias Winterhalter, and Jean Marie Pagès. Porins and small-molecule translocation across the outer membrane of Gram-negative bacteria. *Nature Reviews Microbiology 2019 18:3*, 18(3):164–176, 12 2019.

[218] Jigneshkumar Dahyabhai Prajapati, Ulrich Kleinekathöfer, and Mathias Winterhalter. How to Enter a Bacterium: Bacterial Porins and the Permeation of Antibiotics. *Chem. Rev*, 121:5192, 2021.

[219] Samuel Baron. *Medical Microbiology*. University of Texas Medical Branch, Galveston, 4th edition edition, 1996.

[220] Yosuke Hoshino and Eric A. Gaucher. On the Origin of Isoprenoid Biosynthesis. *Molecular Biology and Evolution*, 35(9):2185–2197, 9 2018.

[221] Vinaya Kumar Golla, Jigneshkumar Dahyabhai Prajapati, Manas Joshi, and Ulrich Kleinekathöfer. Exploration of Free Energy Surfaces across a Membrane Channel Using Metadynamics and Umbrella Sampling. *Journal of Chemical Theory and Computation*, 16(4):2751–2765, 2020.

[222] Vinaya Kumar Golla, Claudio Piselli, Ulrich Kleinekathöfer, and Roland Benz. Permeation of Fosfomycin through the Phosphate-Specific Channels OprP and OprO of Pseudomonas aeruginosa. *J. Phys. Chem. B*, 2022:126–1388, 2022.

[223] Alessandro Pira, Mariano Andrea Scorciapino, Igor V. Bodrenko, Andrea Bosin, Silvia Acosta-Gutiérrez, and Matteo Ceccarelli. Permeation of β-lactamase inhibitors through the general porins of gram-negative bacteria. *Molecules*, 25(23), 12 2020.

[224] Shalini Awasthi and Nisanth N. Nair. Exploring high dimensional free energy landscapes: Temperature accelerated sliced sampling. *Journal of Chemical Physics*, 146(9):094108, 3 2017.

[225] Nandan Haloi, Archit Kumar Vasan, Emily J. Geddes, Arjun Prasanna, Po Chao Wen, William W. Metcalf, Paul J. Hergenrother, and Emad Tajkhorshid. Rationalizing the generation of broad spectrum antibiotics with the addition of a positive charge. *Chemical Science*, 12(45):15028–15044, 11 2021.

[226] Archit Kumar Vasan, Nandan Haloi, Rebecca Joy Ulrich, Mary Elizabeth Metcalf, Po Chao Wen, William W. Metcalf, Paul J. Hergenrother, Diwakar Shukla, and Emad Tajkhorshid. Role of internal loop dynamics in antibiotic permeability of outer membrane porins. *Proceedings of the National Academy of Sciences of the United States of America*, 119(8):e2117009119, 2 2022.

[227] Abhishek Acharya, Jigneshkumar Dahyabhai Prajapati, and Ulrich Kleinekathöfer. Improved sampling and free energy estimates for antibiotic permeation through bacterial porins. *Journal of Chemical Theory and Computation*, 17(7):4564–4577, 7 2021.

[228] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R.J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T. Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods 2020 17:3*, 17(3):261–272, 2 2020.

[229] Jochen S. Hub, Bert L. De Groot, Helmut Grubmüller, and Gerrit Groenhof. Quantifying artifacts in Ewald simulations of inhomogeneous systems with a net charge. *Journal of Chemical Theory and Computation*, 10(1):381–390, 1 2014.

[230] Igor V. Bodrenko, Samuele Salis, Silvia Acosta-Gutierrez, and Matteo Ceccarelli. Diffusion of large particles through small pores: From entropic to enthalpic transport. *Journal of Chemical Physics*, 150(21), 6 2019.

[231] Niraj Modi, Sonalli Ganguly, Iván Bárcena-Uribarri, Roland Benz, Bert Van Den Berg, and Ulrich Kleinekathöfer. Structure, dynamics, and substrate specificity of the OprO porin from Pseudomonas aeruginosa. *Biophysical Journal*, 109(7):1429–1438, 10 2015.

[232] Jeffery B. Klauda, Richard M. Venable, J. Alfredo Freites, Joseph W. O'Connor, Douglas J. Tobias, Carlos Mondragon-Ramirez, Igor Vorobyov, Alexander D. MacKerell, and Richard W. Pastor. Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. *Journal of Physical Chemistry B*, 114(23):7830–7843, 6 2010.

[233] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry*, 31(4):671–690, 3 2010.

[234] Sunhwan Jo, Taehoon Kim, Vidyashankara G. Iyer, and Wonpil Im. CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*, 29(11):1859–1865, 8 2008.

[235] Sanghyun Park and Vijay S. Pande. Choosing weights for simulated tempering. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 76(1), 7 2007.

[236] Jérôme Hénin, Tony Lelièvre, Michael R. Shirts, Omar Valsson, and Lucie Delemotte. Enhanced sampling methods for molecular dynamics simulations. *arXiv*, 2 2022.

[237] Alejandro Gil-Ley and Giovanni Bussi. Enhanced Conformational Sampling Using Replica Exchange with Collective-Variable Tempering. *Journal of chemical theory and computation*, 11(3):1077–85, 3 2015.

[238] Soumendranath Bhakat. Collective variable discovery in the age of machine learning: reality, hype and everything in between. *RSC Advances*, 12(38):25010–25024, 12 2021.

[239] Yihang Wang, João Marcelo Lamim Ribeiro, and Pratyush Tiwary. Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nature Communications*, 10(1):1–8, 12 2019.

[240] Shashank Pant, Zachary Smith, and Yihang Wang. Confronting pitfalls of AI-augmented molecular dynamics using statistical physics COLLECTIONS ARTICLES YOU MAY BE INTERESTED IN Reweighted autoencoded variational Bayes for enhanced sampling. *RAVE) The Journal of Chemical Physics*, 153:72301, 2020.

[241] Zachary Smith, Pavan Ravindra, Yihang Wang, Rory Cooley, and Pratyush Tiwary. Discovering loop conformational flexibility in T4 lysozyme mutants through artificial intelligence aided molecular dynamics. 2020.

[242] Rhys Evans, Ladislav Hovan, Gareth A. Tribello, Benjamin P. Cossins, Carolina Estarellas, and Francesco L. Gervasio. Combining Machine Learning and Enhanced Sampling Techniques for Efficient and Accurate Calculation of Absolute Binding Free Energies. *Journal of Chemical Theory and Computation*, 16(7):4641–4654, 2020.

[243] Shams Mehdi, Dedi Wang, Shashank Pant, and Pratyush Tiwary. Accelerating All-Atom Simulations and Gaining Mechanistic Understanding of Biophysical Systems through State Predictive Information Bottleneck. *Journal of Chemical Theory and Computation*, 18:3231–3238, 2022.