Conformational dynamics of DEAH-box helicases from molecular dynamics simulations

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

von
Robert A. Becker

Saarbrücken
2023

Tag des Kolloquiums: .... 22. November 2023
Dekan: .... Prof. Dr. Ludger Santen
Berichterstatter: .... Prof. Dr. Jochen Hub
.... Prof. Dr. Gregor Jung
Vorsitz: .... Prof. Dr. Ludger Santen
Akad. Mitarbeiter: .... Dr. Thomas John

Danksagung

Mein Dank gilt zuerst Prof. Dr. Jochen Hub für die spannende Forschungsidee und -gelegenheit, die umfangreiche Betreuung, die freundlichen und inspirierenden Gespräche, welche immer wieder eine große Unterstützung über die gesamte Laufzeit der Promotion waren. Seine Tür stand sowohl für fachliche als auch persönliche Gespräche immer offen, was ich sehr zu schätzen weiß. Die geführten Dialoge waren für mich immer lehrreich, ermutigend und motivierend.

Ich danke außerdem Prof. Gregor Jung für die wissenschaftliche Betreuung und die Anfertigung des Zweitgutachtens.

Ferner danke ich Prof. Dr. Ralf Ficner und seiner Arbeitsgruppe für die gute Zusammenarbeit, die konstruktiven und aufschlussreichen Gespräche und die Bereitstellung der Kristallstrukturen.

Ein besonderer Dank geht zudem an Leonie für die kritische Begutachtung meiner Arbeit und die sehr wertvollen Kommentare und Verbesserungsvorschläge. Außerdem möchte ich Jeremy, Leonhard, Gari, Johanna, Chetan, Tobias, Milos, Alejandro, Katharina, Massimiliano und auch Leonie für die fortlaufende Unterstützung und die großartige und unvergessliche Zeit danken. Es war mir eine Ehre.

Ein spezieller Dank geht an Alex und Danjano, die mich während meines gesamten Studiums begleitet und tatkräftig unterstützt haben. Auch ohne sie wäre diese Doktorarbeit niemals Zustande gekommen. Vor allem während der Doktorandenzeit möchte ich an dieser Stelle auch meinem Cousin Felix, Benedikt und Jens danken, die diese Zeit um einiges spaßiger gemacht haben.

Ich möchte meiner Freundin Teresa von Herzen für ihre emotionale Unterstützung in allen Lebenslagen bedanken. Ihr offenes Ohr und ihre Ermutigungen haben mich stets dazu motiviert mein Besten zu geben.

Ich widme diese Doktorarbeit meiner lieben Mutter, die mir nicht nur während meiner akademischen Laufbahn, sondern auch mein ganzes Leben lang bedingungslose Untersützung geschenkt hat. Sie hat mir den Weg zum Erreichen dieses Meilensteins geebnet und ich bin ihr unendlich dankbar für ihre liebevolle Begleitung meines gesamten Lebens.

Abstract

Helicases are special kind of ATPases, which are vital to all living organisms. They facilitate nucleic acid strand separation and translocation. DEAH-helicases are involved in the splicing pathway, where they are a part of the spliceosome and facilitate various functions, such as release of mRNA, recycling of spliceosome complexes and proofreading of RNA substrates. The mechanistic function is carried out by the translocation of a single stranded RNA (ssRNA) through the RNA cleft of the helicase facilitated by ATP hydrolysis. This thesis employs molecular dynamics simulations to explore DEAH-helicase conformations and transitions. We present a novel approach by combining Simulated Tempering and Adaptive Sampling, which overcomes the sampling challenge of such a complex system. The combination of these techniques reveals the atomic-level sampling of a complete Prp43 translocation cycle by one nucleotide. Key findings include the role of molecular switches in driving large conformational changes, such as helix-to-loop transitions or small-scale hydrogen bond rearrangements, and the RNA's impact on helicase rigidity. In general, we gained important insights into the impact of the ligands ATP, ADP, and RNA on the dynamics of helicases, which are essential for further investigations and hypotheses. Furthermore, the study validates the hypothesized exit tunnel for phosphate and examines the influence of the G-patch cofactor on protein dynamics, particularly on the modulattion of the enzyme's ATPase activity.

Zusammenfassung

Helikasen sind eine besondere Art von ATPasen, die für alle lebenden Organismen lebenswichtig sind. Sie erleichtern die Trennung und Translokation von Nukleinsäuresträngen. DEAH-Helikasen sind am Spleissweg beteiligt, wo sie Teil des Spleissosoms sind und verschiedene Funktionen wie die Freisetzung von mRNA, das Recycling von Spleissosom-Komplexen und das Korrekturlesen von RNA-Substraten ermöglichen. Die mechanistische Funktion wird durch die Translokation einer einzelsträngigen RNA (ssRNA) durch den RNA-Spalt der Helikase ausgeführt, die durch ATP-Hydrolyse ermöglicht wird. In dieser Arbeit werden Molekulardynamiksimulationen eingesetzt, um Konformationen und Übergänge der DEAH-Helikase zu untersuchen. Ein neuartiger Ansatz, der Simulated Tempering und Adaptive Sampling kombiniert, überwindet die Herausforderung des Samplings in einem komplexen System. Die Kombination dieser Techniken ermöglicht es, einen kompletten Prp43-Translokationszyklus auf atomarer Ebene um ein Nukleotid zu simulieren. Zu den wichtigsten Erkenntnissen gehört die Rolle molekularer Schalter, wie Helix-zu-Schleife-Übergänge oder kleineren Umlagerungen von Wasserstoffbrückenbindungen, die grosse Konformationsänderungen bewirken. Generell haben wir wichtige Erkenntnisse über die Auswirkungen der Liganden ATP, ADP und RNA auf die Dynamik von Helikasen gewonnen, die für weitere Untersuchungen und Hypothesen wichtig sind. Darüber hinaus bestätigt die Studie die Hypothese eines Ausgangstunnels für Phosphat und untersucht den Einfluss des G-Patch-Cofaktors auf die Proteindynamik, insbesondere durch Modulation der ATPase-Aktivität des Enzyms.

Contents

Associated work

This dissertation is mainly based on the content of the following three papers:

1. Florian Hamann, Lars C Zimmerningkat, Robert A Becker, Tim B Garbers, Piotr Neumann, Jochen S Hub, and Ralf Ficner. The structure of Prp2 bound to RNA and ADP-BEF3- reveals structural features important for RNA unwinding by DEAH-box Atpases. *Acta Crystallographica Section D: Structural Biology*, 77(4):496-509, 2021.

2. Robert A Becker and Jochen S Hub. Continuous millisecond conformational cycle of a DEAH-box helicase reveals control of domain motions by atomic-scale transitions. *Communications Biology*, 6(1):379, 2023.

3. Robert A Becker and Jochen S Hub. Molecular simulations of DEAH-box helicases reveal control of domain flexibility by ligands: RNA, ATP, ADP, and G-patch proteins. *Biological Chemistry*, (0), 2023.

Introduction

Nucleotides, such as ribonucleic acid (RNA), deoxyribonucleic acid (DNA), and adenosine nucleoside triphosphates (NTPs), are unique molecules that are essential for life. NTPs, for example, are the monomeric building units of RNA and DNA, but they also facilitate a wide variety of protein mechanisms within each cell. The most prominent member of the NTPs, adenosine triphosphates (ATP), is particularly notable for providing energy to living cells. As a result, many enzymes are ATPases, meaning they lower the activation energy required for ATP hydrolysis and make use of the released energy by the hydrolysis. Conclusively, the decomposition of ATP into ADP and phosphate can release enough energy to drive a wide variety of cellular processes.

RNA and DNA, on the other hand, contain genetic information, which is essential for the function, reproduction, and development of all known organisms, as well as for evolution itself. Crick postulated a fundamental principle in genetics, known as "the central dogma of molecular biology" [1, 2]. In its most general form, this principle states that the flow of genetic information proceeds in only one direction: DNA is transcribed into RNA, and RNA is then translated into proteins. Once a protein has been produced, it cannot be translated back into genetic information. As a result, the process of transcription and translation must be maintained in a highly controlled fashion to avoid errors during protein synthesis, since the errors cannot be reversed.

To accomplish mostly immaculate transcription and translation, organisms have developed large and complex machineries to perform and control these tasks. For example, the transcription of DNA to RNA is carried out by the RNA polymerase, an enzyme with a mass of over 400 kDa. Then, special molecular machines known as ribosomes translate the RNA into proteins, using the RNA sequence as a template for the primary structure of the synthesized proteins. However, before RNA can be processed, it must be prepared by the so-called spliceosomes in a pre-translation step. The spliceosomes facilitate the production of mature messanger RNA (mature mRNA) from pre-mRNA. This process is highly sensitive to errors, which can lead to disease-causing mutations. Therefore, the multiple steps of splicing, including assembly, activation, catalysis, and disassembly, must be supervised by specialized enzymes.

One group of these supervising enzymes are the helicases. Helicases have maintenance and proofreading function. They assist other proteins in binding to the RNA substrate, check for suboptimal substrates and suboptimal substrate binding. If they encounter potential errors, they have the ability to discard the substrate or the binding factor.

Since the stability of the genome depends on helicases, they may be involved in the process of aging, cancerogenic cell alteration, and other genetic diseases and disorders. As a result, helicases have gained increasing attention in scientific research over the last several decades and are considered potential drug targets for preventing cancer formation or other age-related diseases [3, 4].

In summary, helicases play a fundamental role in the functionality and reproduction of all living organisms. These proteins are essential for the flow of genetic information and the production of other proteins, and their proper functionality is vital to the health and survival of cells. The machinery responsible for the splicing of RNA and DNA is complex and multi-faceted and requires the action of a wide variety of enzymes and other molecules to ensure accuracy and efficiency. The study of helicases and their associated machinery offers exciting opportunities for understanding the basic processes of life and for developing potential treatments for a wide range of diseases. Hence, we present a broad investigation of RNA helicases - more specifically DEAH-helicases - via molecular dynamics.

1.1 History

Over 40 years of helicase research. Helicases were first discovered in *E. Coli* by Hoffmann-Berling et al. in 1976 [5,6]. Back then, the only known properties of this newly discovered enzyme were the abilities to hydrolyze ATP and to unwind DNA. Thus, the enzyme was simply called “DNA unwinding enzyme”. Independently in 1976, Mackay and Linn et al. pointed out that the so-called RecBC enzyme in complex with a binding protein is able to unwind DNA in an ATP-depended manner [7]. Two years later, in two follow-up publications of Hoffmann-Berling et al., the enzymes were explicitly renamed to “DNA helicases” [8,9]. This date was the official birthday of the term “helicase”.

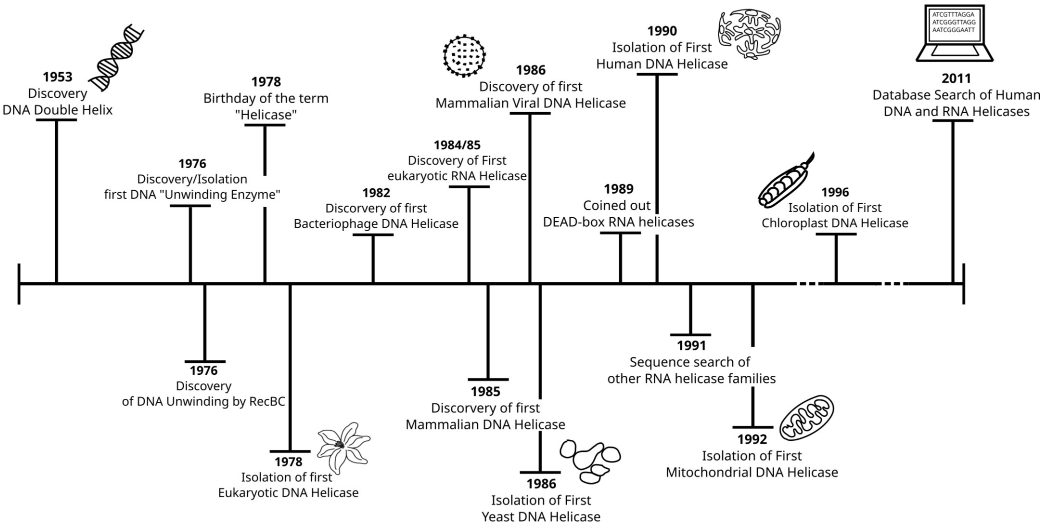


Figure 1.1: *A simplified timeline of helicase discoveries and isolations in various specimen. Starting in the 70s until the early 2000s, helicases were isolated from bacteria, phages, plants, mammals, and other organisms. In 2011, a big database search was carried out, resulting in identification of 95 helicases. Picture was highly inspired by the work of Brosh and Watson [19].*

In the following 20 years, helicases were found in all kinds of procaryotic and eucaryotic cells. The first isolation of an eucaryotic helicase from Lily plants was achieved by Hotta and Stern in 1978 [10], the first bacteriophage helicase was isolated by Nossal et al. in 1982 [11], followed by the first mammal helicase from calf thymus by Stalder and Hubscher in 1985 [12]. Subsequently, in 1986 the first yeast helicase [13], in 1990 the first human helicase [14] and in 1996 the first chloroplast helicase from pea [15] were discovered. The first RNA helicase was found 1984 and 1985 when Grifo et al. [16] and Ray et al. [17], respectively, identified a eukaryotic translation initiation factor, which is named elF4A. Today, this factor is known to be the "godfather" of the DEAD-box proteins [18], which represent one of the most important helicase families by this date. The closely related DEAH-box helicases are often combined to a more general DExD/H-box family because of the high structural similarities. A simplified timeline of helicases is shown in Fig. 1.1. For further and more detailed information about the uncovering of helicase in various species, we refer to the review of Brosh and Watson [19].

1.2 Classification of RNA helicases

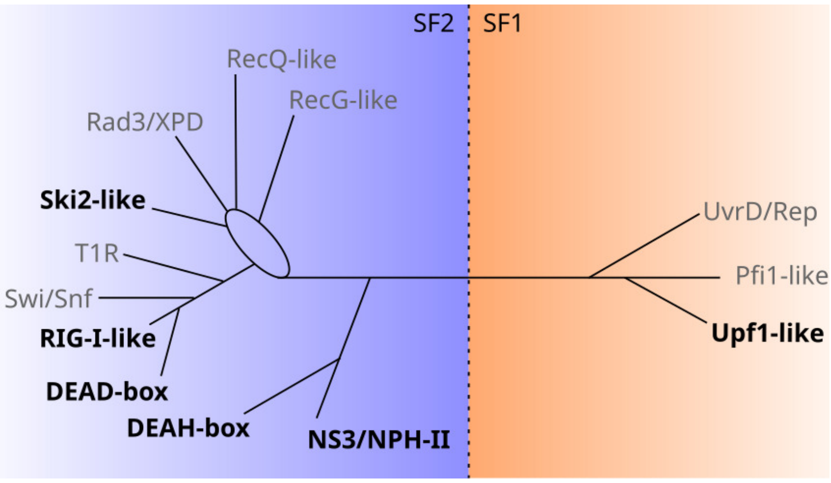


Figure 1.2: *A simplified classification tree of the two super families 1 and 2. The investigated subfamily for this work is the DEAH-box helicases, which are closely related to the DEAD-box helicases. The length of lines in this plot are not to scale. Bold text indicating families which consists of RNA helicases. The picture is adopted from Fairman-Williams et al. [20].*

In 2011, Umate et al. performed a genome wide search and comparison of DNA and RNA helicases to provide a foundation for the organization and characterization of helicases. They identified 95 helicases in the human genome, 31 of them being DNA helicases and 64 being RNA helicases [21].

In general, RNA helicases are characterized by the presence of seven to eight highly conserved motifs, which play a crucial part in the NTP (mostly ATP) hydrolysis [22]. All helicases can roughly be categorized by their quaternary structure in two distinct groups: The ring-shaped hexameric helicases and the monomeric helicases. Sequence analysis by Gorbalenya and Koonin et al. [23] and further structure and function-based analyses by Wigley et al. revealed that helicases can be further divided by their conserved motifs into six superfamilies (SF) [24, 25]. Thereby, the ring-shaped hexameric helicases are members of the SF 3-6 and all other helicases are either members of the SF 1 or SF 2. The SF 2 forms the largest SF of helicases, and it includes, among others, the DExD/H-box, RecQ-like, and Ski-like helicases [20, 22, 26, 27]. A cladogram of the SF 1 and 2 is shown in Fig. 1.2. In addition, to distinguish helicases even further, they can be grouped by their directionality into up- and downstream helicases, denoted as type A and B, respectively, and additionally into helicases working on single- or double-stranded ligands, denoted as α and β types, respectively [28].

Two such translocating SF2Aα DEAH-box helicases are Prp22 and Prp43, which move directionally along the RNA strand by binding to the RNA phosphodiester backbone. X-ray crystallographic studies revealed the structures of Prp22 and Prp43 from *C. thermophilum*, providing insight into the ATP binding and hydrolysis mechanism as well as into certain snapshots along the RNA translocation cycle and give an idea of how they might work in detail [29, 30].

In this work, we want to focus on DEAH-box helicases. Like the name of the DEAD-box helicases, the name of the DEAH-box helicases is derived from a highly conserved motif inside of the helicases core, which consists of aspartic acid. Like the DEAD-box and Ski-like helicases, the DEAH-helicases are key players in the splicing pathway. Their detailed function is discussed in the following section.

1.3 Function

Helicases play crucial roles in any processes which involve DNA or RNA such as transcription, translation, recombination, repair, proof-reading, ribosome biogenesis, RNA transport, splicing, degradation, assembly, disassembly and many more. In most cases, helicases use the energy released by nucleoside triphosphate (NTP) hydrolysis either to unwind double-stranded RNA/DNA or to drive forward the translocation of RNA/DNA [31–36]. Thus, the name helicase is not always appropriate, since a lot of helicases neither cleave double stranded DNA/RNA nor are involved in dsDNA/dsRNA in any way, but rather translocate ssRNA. In fact, unwinding can be achieved with the help of translocation [24, 37, 38], but there exists unwinding without translocation [39–42] and translocation without unwinding [20, 28, 43–46]. In the latter case, the so-called translocases couple the hydrolysis of ATP to the directional movement of single-stranded or double-stranded nucleic acid.

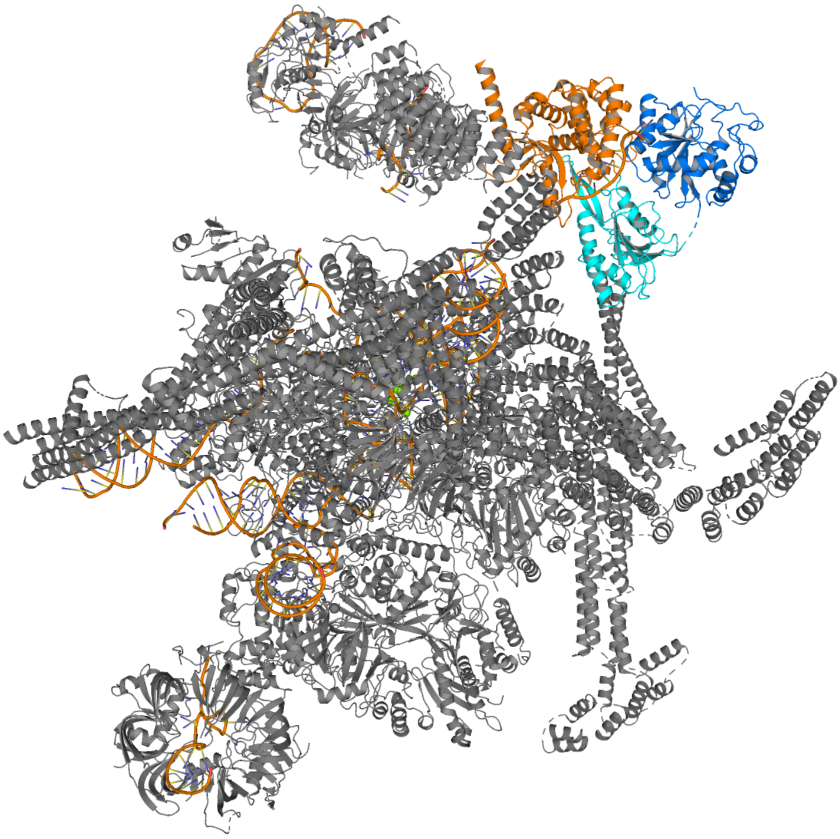


Figure 1.3: *Prp43 connected to the intron lariat spliceosome. Prp43 domains are colored in cyan (RecA2), blue (RecA1) and orange (CTD). The structure was resolved by Wan et al. Wan et al. [47] and is available in the PDB database with the PDB ID 5Y88.*

DEAH-box helicases keep the genome stable. In general, DEAH-box helicases are RNA translocases, which are involved in the activation, catalysis and disassembly of the spliceosome complex during the splicing pathway. Therefore, the helicases are tightly bound to the spliceosome(s) during these processes, thus being an important part of the complete spliceosome complex. For example, Prp43 is bound to the Syf1 domain of the intron lariat spliceosome (ILS) complex as shown in Fig. 1.3 [47]. The major function of DEAH-box helicases in the splicing pathway is the translocation of their substrate RNA by facilitating the conformational rearrangements of the spliceosome. RNA helicases also facilitate the activation of the spliceosome, modulating the connection and release of various splicing factors and disassemble spliceosome complexes. All changes in the spliceosome conformations and interaction with the co-factors is supposedly done for proofreading during the whole mechanism of splicing. Hence, helicases protect the spliceosome to interact with suboptimal substrates [48–50]. Such suboptimal substrates can be pre-mRNA splicing sites, which are altered in their normal sequence, disturbed RNA secondary structures, RNA with wrongly attached binding proteins etc. [50–53]. It has been shown that the proofreading mechanism is established in two different ways [53,54]. The so-called timer model connects the time it takes for a helicase to interact with a substrate [50]. More precisely, the model states that the helicase has a limited time window to act on the substrate during the splicing pathway. Thus, an optimal substrate will interact with the helicase for a small amount of time, because the helicase is built to interact with it in an efficient manner. But, if the substrate is not an optimal interaction partner of the helicase, it will proceed more slowly. Hence, the helicase has more time to act, and the substrate will be discarded. The sensor model on the other hand, suggests that the helicase can reject a non-optimal substrate faster than an optimal substrate, because of the different stability of the resulting substrate-spliceosome complexes [50, 54].

Translocation rate of RNA helicases. The ATPase activity and its turn-over rate are major factors for the speed of a helicase along DNA/RNA. While some RNA helicases translocate just a few base pairs per second, others are able to translocate up to thousands base pairs per second. The variation in the rates is depending on two factors: 1. the helicase itself, most importantly on its structure and 2. the regulation by co-factors. The regulation is downright essential for all mechanisms involving DNA or RNA, because the cells must keep the DNA/RNA strands most of the time in its more stable duplex form than in the single-stranded form [28]. The number of base pairs which are translocated during one cycle of an ATP hydrolysis is called step size. The step size may not be confused with the term rate. The step size is another important quantity to characterize helicases and is essential for the understanding of the mechanism itself. Like the rate, the step size can vary among helicases from 1 bp over to a few bps up to over 20 bps per ATP hydrolysis [55, 56].

Although kinetic quantities can be measured via various experimental approaches, the structural cause and connection is often unknown, because many (intermediate) conformations along the cycle are not resolvable by experiments. For example, due to a lack in RNA-bound crystal structures, the understanding of the relevant domain motions and cascades of the molecular switches remain limited. In the literature, there are two discussed hypotheses for the translocation mechanism derived from the available structures.

The first proposed mechanism is called the hand-over mechanism [57]. The hand-over mechanism is more applicable for the translocation of helicases from the SF 3 to 6. During the hand-over mechanism, enzymes need to be at least dimers or in the case of SF 3 to 6 helicases, hexamers. Since DEAH-helicases are usually monomeric, we will not discuss this mechanism.

The second model is the famous and widely discussed inchworm model, which describes RNA translocation as a stepping process [22,28,58]. Accordingly, while a first RecA-like domain is tightly bound to the RNA, a second RecA-like domain moves along the RNA driven by a power stroke from NTP hydrolysis until it finds a new tightly-binding contact position on the substrate. Next, by another change of the NTP binding state, the first RecA-like domain is now only weakly bound, enabling it to follow the previously moved RecA-like domain along the RNA. Thus, according to the inchworm model, at least one RecA-like domain is tightly bound to the RNA at any time while the other RecA-like domain changes its affinity to the RNA depending on the NTP binding state.

1.4 Structure

As mentioned above, the SF 2 helicases do not appear as ring-shaped hexametric structures. Instead, they are monomeric, with a highly conserved core, which are composed of two ATPase recombination protein A (RecA) in close proximity to each other. The domains sandwich an NTP molecule in order to carry out their ATPase activity [28, 59, 60]. The two RecA domains are referred as RecA1 and RecA2.

In 1982 Walker et al. already pointed out that the RecA protein shares some amino acid regions with other known ATPases such as myosin, kinases and more [61]. The findings led to the conclusion that there might be conserved NTP binding motifs in the group of ATPases. Six years later, similar amino acid motifs could be identified in helicases, highlighting the similarity of helicases with other known ATPases [23]. The highly conserved motifs are known as the Walker A and B motifs or P-Loop and consist of a beta-strand and a glycine-rich loop, which is followed by an α-helix. In the Walker A motif, a lysin sidechain and some of the backbone nitrogen atoms are interacting with the β- and γ-phosphate of the NTP in the pocket, and therefor are important nucleotide binding partners. Additionally, seven other conserved motifs located at the interface between the two RecA domains were identified. Thus, defining them as the nine conserved structural motifs Q, I, Ia, Ib, II, III, IV, V and VI [21, 62]. Motif VI being the mentioned P-loop. Motif Ia, Ib and IV are crucial for the binding and interaction with

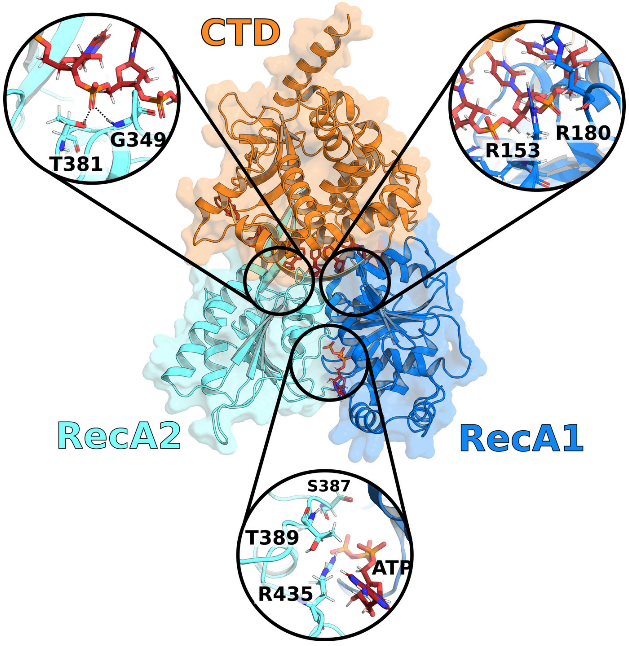


Figure 1.4: *Prp43 in the closed state (PDB ID: 5LTA).*
C-terminal domain (CTD, orange), RecA2 domain (cyan) and RecA1 domain (blue). Boxes show close-up views on the hook-loop (top left), the hook-turn (top right), and on the ATP-pocket (bottom).

the DNA or RNA ligand. In addition, motif III has been proposed to be an important key player, which connects the unwinding process with the ATP hydrolysis [63,64]. Furthermore, Prabu et al. and Ficner et al. identified the distinct beta-hairpin motif and the hook-loop and hook-turn motifs, respectively. Also, in motif V a serine is identified, which is hypothesized to play a crucial role in sensing the catalytic state of the enzyme. In the absence of ATP, the corresponding motif of the serine tends to form a helical structure with the short helix in close proximity by flipping away from the ATP binding pocket. In the presence of ATP, the serine is in a loop structure, bent towards to ATP. A comparison of the two different states is shown in Fig. 1.5. We will refer to this serine as "sensor serine“ and the conformational helix-to-loop (or loop-to-helix) transition as "serine flip“ in this work. The four last mentioned unique structural segments are believed to be involved in the translocation and unwinding mechanism [29,30,65].

Besides the two RecA domains, which build the core of the enzymes, SF 1 and 2 helicases often consists of C-terminal (CTD) and N-terminal (NTD) domains, which can make up most of the enzymes entire mass. For example, the DEAH-box helicases have a large CTD which is further di-

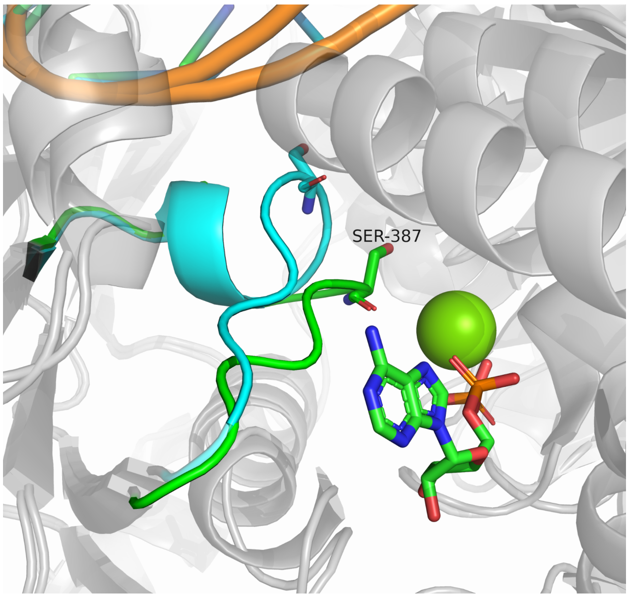


Figure 1.5: *Behavior of the sensor serine in Motif V. When ATP is present in the binding pocket (green), the sensor serine bends towards the ATP-water complex. In the absence of ATP (cyan), the serine flips upwards and points to the RNA strand, which leads eventually to interactions between the phosphate backbone and the sensor serine.*

vided into the winged-helix (WH) domain, a ratchet-like domain, and an oligosaccharide-binding fold (OB) [66,67]. These CTD and NTD are often of structural and functional importance. For example, the RNA tunnel of the helicase Prp43 is defined as the space between the CTD and the two RecA domains. Additionally, the CTD and NTD can promote oligomerization, facilitate protein-protein interactions, and control the enzymes function by, for example, influencing the ligand specificity or recognition of specific nucleic acid regions [20,28,68,69]. Usually, CTD and NTD are not conserved within an SF but in some cases within subfamilies such as the DEAH-box helicases or Ski-like helicases [66, 70].

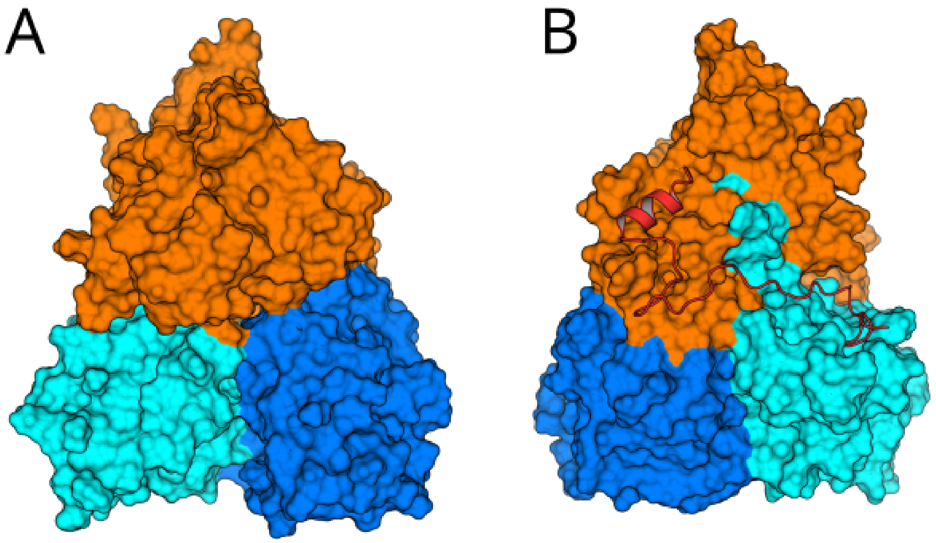


Figure 1.6: *DHX15 in contact with the G-patch NKRF (red). The G-patch is located on the "back" of the enzyme. Here it is connected with a more defined structure on two positions. One is a helix structure on top of the CTD domain on the WH region and the other is a loop-like structure close to the β-hairpin and the RecA2 domain.*

1.5 Regulation by G-patches

Regulation of proteins is essential for most biological processes, because a dysregulation can lead to dramatic failures in the fine-tuned apparatuses in cells, which leads to diseases and/or disorders. In case of RNA helicases, the dysregulation leads to cancer and other age-related diseases and disorders [71]. Fortunately, helicases are regulated in a various ways. This includes substrate-depended auto-inhibition [72, 73], post-translation modifications which renders the catalytic activity depended on specific conditions [74–76], recognition of specific RNA features [77, 78] and most importantly cofactor-depended regulation [79, 80]. The ladder alters the catalytic activity of a helicase by either creating an electrostatic environment which assists the binding to RNA or by mediating direct RNA-protein interactions by conformational changes. There is a large variety of different co-factors for RNA helicases, which can enhance their otherwise poor helicase activity [79,81,82]. One family of those direct co-factors are the so-called G-patch proteins. Their name is derived by their characteristic and highly conserved glycine-rich sequence [83]. The overall motif consists of around 50 amino acids with the conserved sequence $Gx_2hhx_3Gax_2GxGlGx_3pxux_3sx_{10-16}GhG$, where a is an aromatic, h an hydrophobic, l an aliphatic, s a small, u a tiny and x a variable amino acid [80]. The helicases investigated in this study have several different G-patch regulation proteins as binding partners. For example, in yeast the G-patch co-factors Cmg1, Pxr1, Sqs1 and Ntr1 (also called Spp382) associate with Prp43, the cofactor Spp2 binds to Prp2 and the Prp43 human analogous helicase DHX15 has six other G-patch co-factors including NKRF. Every G-patch protein has a unique role in different kind of processes, like the G-patch Pfa1 of Prp43, which regulates the ATPase's unwinding and hydrolysis activity during the ribosome biogenesis or the G-patch Ntr1 and Ntr2, which enhances the ATPase's unwinding activity of Prp43 during the pre-mRNA splicing.

In this study, for the investigation of G-patches, we will focus on the human DEAH-box helicase DHX15, which is an analogous to the yeast and thermophilum helicase Prp43. Two structures of the DHX15 with an attached NKRF was recently resolved by Studer et al. [84] (Fig. 1.6), which makes an investigation via MD simulations possible.

1.6 Malfunction and disease

80% of all viruses are RNA viruses and carry at least one gene encoding a helicases in their genome. The other 20% usually hijack eucaryotic helicases to maintain their reproduction cycle [71]. Thus, RNA helicases have been associated with playing a essential role in bacterial and fungal infections. For example, studies showed that the deactivation of RNA helicases in *Borrelia burgdorferi* (Lyme disease) [85] and in *Cryptococcus neoformans* (Cryptococcosis) [86] drastically reduce their virulency in mice models [71].

In contrast, RNA helicases also play a crucial role in the antiviral defense mechanism of our immune system. For further reading on the impact of RNA helicases on viral infection we refer to the work of Ranji and Boris-Lawrie et al. [87]. Due to their critical role in the genome maintenance, the malfunction of helicases is involved in genetic diseases such as cancer or aging-related disorders such as Bloom syndrome [88,89], Werner syndrome [90], and others [91–95].

Studies showed that helicases are upregulated in tumor cells, which leads to a faster growth rate of cancer cells than of regular cells, which further underlines the impact of helicases in cancer formation. Hence, helicases might be potential drug targets for cancer therapy [71].

1.7 Experiments

Structural biology experiments, such as X-ray crystallography [96,97], NMR [98,99], cryo-EM [100,101], fluorescence single molecular methods [102], and others, play a pivotal role in elucidating the spatial arrangement of atoms in proteins [103]. For instance, X-ray crystallography allows for the determination of protein structures in various conformations by crystallizing proteins under distinct conditions. Meanwhile, cryo-EM enables the visualization of protein complexes and their conformational changes in their native/solvated state. Despite their widespread use and significant impact on understanding proteins and enzymes, these structure determination experiments come with inherent limitations.

One drawback of X-ray crystallography is its dependency on the successful crystallization of the protein of interest, which can be a non-trivial and challenging task, particularly for highly dynamic proteins with flexible domains. Furthermore, the structures captured through crystallization might differ from the proteins' native state in solution, as the frozen unit cell may favor a different conformation.

Cryo-EM, on the other hand, does overcome some of the disadvantages of X-ray crystallography by providing native images of proteins in solution. However, it is constrained by capturing snapshots of certain states of the protein and yielding lower resolutions for the dynamic regions of proteins. Despite these limitations, these techniques remain indispensable tools in the field of structural biology, facilitating critical insights into the molecular architecture and functional dynamics of proteins.

MD simulations can address these limitations by efficiently sampling the configurational space of proteins, using structural data from the mentioned experimental methods as a reference or starting point. For instance, MD simulations can perturb a crystal structure by removing a ligand and observe the protein's behavior in response to the perturbation over time. This dynamic sampling allows the protein to explore stable states beyond the initial structure by crossing energy barriers, yielding valuable information about transitions between states stored in a trajectory. Consequently, MD simulations provide more than just rigid structures of specific states; they offer insights to make further predictions and build hypotheses about protein dynamics.

However, MD simulations come with two main disadvantages. Firstly, biological processes, such as large domain motions, can range from micro to milliseconds or even a few seconds. Meaning these processes are relativley slow, considering that the time step of MD simulations lies in the range of a few femtoseconds. Thus, the sampling problem arises especially for large and complex systems, as the undertaking of micro- to millisecond simulations is challenging even with state-of-the-art hardware. Secondly, MD simulations are approximations of the real world, as the interactions are defined by force fields, which rely on a collection of parameters from various sources. Nevertheless, the advancement of high-performance computers, improved MD software, and a vast array of enhanced sampling techniques have made sampling and long time-scale simulations increasingly feasible. Additionally, force fields have become more accurate due to parameters being derived from more precise experimental and quantum calculation data [104, 105].

To perform worthwhile MD simulations of complex or large systems on conventional hardware, the appropriate sampling technique must be selected, and the simulated data must be carefully compared to available experimental data to ensure its biological and physical relevance. Simulations and experiments complement each other effectively, with simulations providing crucial details of protein dynamics and kinetics. The information gained from simulations can make experimental observations more interpretable, particularly as understanding atomistic dynamics and kinetics through experimental methods remains challenging. Conversely, experiments play a paramount role in providing direct data, insights, and visualizations of reality [106]. Hence, MD would not be possible to conduct without the experiments as foundation.

1.8 Sampling problem

Molecular dynamics simulations have been used to follow conformational transitions of enzymes in thousands of studies to this date. However, equilibrium MD simulations of complex conformational cycles involving multiple molecular transitions, including the interaction with one or more ligands, are still challenging because the simulations typically do not cover the functionally relevant time scales. Out-of-equilibrium pulling simulations in principle overcome high energy barriers which resemble transitions on long time scales, but in some cases, they are not applicable because they require the definition of a good reaction coordinate for the process of interest, which is far from trivial for complex, multi-step, non-linear conformational transitions. Therefore, MD simulations face two problems for larger and more complex systems:

1. Simple conventional equilibrium simulations are still not feasible for large proteins which cross their barriers between biological and physical relevant states only once in a few micro- to milliseconds.

2. Out of equilibrium simulations may not be suitable because of the existence of multiple orthogonal degrees of freedom in the protein's dynamics, which are essential to cover the functionally relevant conformational space.

Fortunately, a lot of different approaches were invented to tackle the omnipresent sampling problem. To name a few: meta dynamics, where the user defines a collective variable (CV), which is a function of selected degrees of freedom. Then, a history dependent bias potential is applied along this CV [107]; accelerated MD, which introduces a bias potential for certain areas in the potential energy landscape under a specific threshold, which makes crossing barriers in the landscape much easier [108]; simulated tempering, which lowers the energy barriers between states by using the temperature of the system as a dynamical variable, which results in a more flat energy landscape [109]; adaptive sampling, which samples a specific path in the conformational space by stopping and restarting simulations during selected and desired intermediates [110].

In this study, we obtained a complete conformational cycle of RNA translocation by the helicase Prp43 by combining two enhanced sampling techniques, namely simulated tempering (ST) and adaptive sampling (AS).

ST is able to accelerate the sampling of transitions by approximately one order of magnitude while maintaining the correct Boltzmann distribution and without the need of defining a reaction coordinate [109, 111]. Instead, ST enhances the sampling by switching the temperature of the system along a pre-defined temperature ladder via a Metropolis criterion. Because enthalpic barriers are flattened at higher temperature, the system may carry out transitions in shorter simulation times as compared to a conventional simulation at room temperature.

During AS, multiple rounds of short parallel simulations are carried out [112]. After each round, the most promising simulations are selected as seed for the next round based on their progress along a set of pre-selected structural features such as presence of H-bonds, distances, angles etc. The main advantage of the AS technique is the ability to trivially parallelize simulations and guide the system in a specific direction. Both these factors can lead to a drastic decrease in the overall wall time, i.e. the actual elapsed real time. In addition, AS may enhance the sampling if the rate-limiting transitions are slow and spatially clustered [110].

1.9 Goals of the thesis

In this thesis, we aim to gain a detailed understanding of the overall mechanism of DEAH-helicases, especially for Prp2, Prp22, and Prp43. To study the translocation process of Prp43, we first performed non-equilibrium simulations to find suitable reaction coordinates for the systems. Since this approach does not lead to promising results due to the complexity of the systems, we introduce a combination of two enhance sampling methods, namely simulated tempering, and adaptive sampling (mentioned above). The major achievement of our project is the sampling of a complete translocation cycle by one nucleotide of the helicase Prp43. The found pathway is plausible and in line with earlier insights by experimental studies of Prp43. However, we cannot guarantee that the sampled path is the minimal free energy path of the helicase.

We could refine the current inchworm model by proposing a similar but more detailed version of it. Additionally, we estimated the rates for important transitions during the cycle and constructed a Markov State Model to get insights into the energy landscape of the helicase. We investigated the exit pathway of the phosphate ion, which is produced by the ATP hydrolysis. Here, we could verify an exit tunnel on the back of the enzyme, which was already suggested by crystallographic observations. At last, we investigated the influence of the G-patch proteins on the helicases, providing insigt into the role of G-patch proteins on RNA helicase modulation.

The RNA translocation. The ability of RNA helicases to translocate RNA is essential for the maintenance of the genome because RNA has to reach specific regions and proteins during the splicing pathway in order to be processed. During this translocation process, RNA helicases perform various tasks on RNA. Helicases can recognize errors in the RNA and subsequently initiate a discard routine if they identify an inconsistency in their substrates. This function is crucial for keeping the genome stable and therefore avoid damage of the genetic code and malfunctions of proteins. Despite recent advances in crystal and cryo-EM structures of RNA helicases, an in-depth atomic mechanism is still not defined. Here, MD can help to shed light on the order and behavior of the molecular switches, which control the large domain dynamics and ligand interactions, as well as identifying intermediate states and estimating the respective rates of transitions. It has been reported that there are several important structural features which may drive the mechanism of the RNA helicases forward. These features are, for example, the sensor serine, hook-loop, and hook-turn, just to name the most important three. Additionally, it was proposed that the RNA translocation is a stepwise process in which the RNA is translocated by one to a few nucleotides (depending on the helicase) per ATP hydrolysis in the fashion of the so-called inchworm process. The main part of this thesis concentrates on the investigation of these hypotheses via MD simulations. Here, we present the combination of adaptive sampling and simulated tempering as a strong approach to observe long time-scale transitions in a feasible amount of time. Further, we constructed a MSM from the gained trajectories of RNA translocation to estimate the rates of the most important transitions.

The role of ATP and dissociation of ligands. Since RNA helicases are ATPases, they hydrolyze ATP to ADP and phosphate to facilitate their function. Consequently, ADP and phosphate must leave the enzyme at some point during the mechanism. A potential exit tunnel was identified from crystal structures by Ficner et al. We investigated the possibility for phosphate to exit DEAH-helicases through the proposed tunnel with the random accelerated molecular dynamics method (RAMD).

The impact of G-patches. The influence of cofactors, especially the so-called G-patches, onto the structural dynamics of DEAH-helicases is a major issue of recent helicase research [80, 84, 113]. The G-patches play major roles in the whole interaction of all members of the spliceosome complex and in the modulation and regulation of helicases. Currently, there are only two resolved structures of DEAH-helicases in complex with a G-patch protein, without missing residues at the binding site, available (Apo structure and ADP complex of DHX15 and ADP complex of Prp2) [84, 114], which makes the investigation of the influence of G-patches via MD challenging. Thus, we compared different DHX15 complexes and their dynamics during ST simulations to the influence of the G-patch. In addition, despite the G-patch being an IDP and the lose contacts with the helicase, we observed the necessity of the whole G-patch being present in the simulations to form stable contacts at the bracelet regions on DHX15.

Theory

2.1 Molecular dynamics simulations

Over the last years and decades, MD simulations have been widely established for studying and understanding biological systems in atomic detail. For example, MD simulation help to understand the observations of biological experiments by providing additional information of a protein's dynamics. In addition, MD is widely used in drug discovery [115], material science [116], folding predication [117], and more.

In MD, single atoms or group of atoms are modeled as spherical objects with a distinct charge and mass applied to them. The bonds between said atoms are described by the potential energy as a function of the atoms positions. Hence, the potential energy of the bonds can be simplified as harmonic bonds. In order to make the calculation of the behavior and interactions feasible, the atoms are treated classically. This means MD solves Newton's equation of motion for all atoms. Thus, MD is less computational costly than quantum calculations, which is crucial to make the calculations of the dynamics of a system feasible. If one is interested in the dynamics of a system and not in chemical processes, such as covalent bond creation and breaking for example, the classical approach - instead of a quantum approach -, can be justified by the following three approximations: (i) The Born-Oppenheimer approximation, (ii) classical treatment of nuclei and (iii) empirical force fields.

1. The Born-Oppenheimer approximation. The Born-Oppenheimer approximation is widely used in molecular physics, to separate the motion of the nuclei and electrons in a molecule. This approximation is based on the fact that the nuclei are more inert than the electrons. The Born-Oppenheimer approximation provides enables the description of the potential energy surface of a molecule as a function of the nuclear positions, which is crucial for many applications in chemistry and physics, such as computing molecular behavior, reaction rates, and thermal properties.

In general, the time evolution of a molecular system is described by the time-dependent Schrödinger equation:

$$\mathcal{H}\Psi = i\hbar\frac{\partial\Psi}{\partial t} \tag{2.1}$$

where $\mathcal{H}$ denotes the Hamiltonian of the system, Ψ the wavefunction of the system, $\hbar$ the reduced Plank constant and t the time.

Because the Born-Oppenheimer approximation separates the motion of the nuclei and the electrons, the corresponding total wavefunction of a molecule is split into two parts: a wavefunction for the nuclei, and a wavefunction for the electrons. The total wavefunction is then written as:

$$\Psi(R, r, t) = \Psi_N(R, t)\Psi_e(r; R) \tag{2.2}$$

where R represents the positions of the nuclei, r represents the positions of the electrons, t represents the time, $\Psi_N(R, t)$ is the wavefunction of the nuclei, and $\Psi_e(r; R)$ is the wavefunction of the electrons depending on the nuclei positions only parametrically.

The potential energy surface is calculated using the time-independent Schrödinger equation for the electrons:

$$\mathcal{H}_e(R)\Psi_e(r; R) = E_e(R)\Psi_e(r; R) \tag{2.3}$$

where $\mathcal{H}_e$ is the Hamiltonian operator for the electrons and $E_e(R)$ is the electronic energy. More specific $E_e(R)$ is the potential energy surface of the molecule in the ground state and represent the potential which acts on the nuclei during motion.

The electronic Hamiltonian equals to

$$\mathcal{H}_e(R) = \mathcal{H} - \mathcal{T}_n, \tag{2.4}$$

where $\mathcal{H}$ denotes the Hamiltonian of the complete system and $\mathcal{T}_n$ denotes the kinetic energy operator for the nuclei. Considering equations 2.3 and 2.4, the time-dependent Schrödinger equation for the nuclei is described by

$$(\mathcal{T}_n + E_e(R))\Psi_N(R, t) = i\hbar\frac{\partial\Psi_N(R, t)}{\partial t}. \tag{2.5}$$

The Born-Oppenheimer approximation assumes that the electronic wavefunction is an accurate representation of the electronic structure of the molecule for a given set of nuclear positions. The wavefunction of the electrons is calculated for a fixed set of nuclear positions, R. The electronic energy is an eigenvalue of the electronic wavefunction and is used to get the potential energy surface for the nuclei. The potential energy surface describes the energy of the molecule as a function of the nuclear positions and provides important information about the stability and reactivity of a molecule.

2. The classical treatment Since many biological systems contain usually ten to hundreds of thousands of atoms, solving the given Schrödinger equation for such a system is not possible. Therefore, a classical approach is applied in which the particles are treated as classical particles, which follow Newton's second law:

$$m_i \frac{\partial^2 \mathbf{R}_i}{\partial t^2} = -\nabla_{R_i} V(\mathbf{R}) \tag{2.6}$$

$$m_i \mathbf{a}_i = \mathbf{F}_i \tag{2.7}$$

Here, $V(\mathbf{R})$ is equal to the potential energy surface E_e and m_i, $\mathbf{a}_i$ and $\mathbf{F}_i$ denote the mass, acceleration and the force of and on the atom i, respectively.

This classical treatment is justified by the Ehrenfest theorem. The Ehrenfest theorem describes - as a mathematical prove to the correspondence principle - how the expectation values of quantum mechanical observables evolve over time within a quantum system. Specifically, it states that the rate of change of the expectation value of a quantum observable is related to the commutator of that observable with the system's Hamiltonian. In simpler terms, it states that an expectation value of a physical observable can be described classically. In addition, the phenomenon called decoherence justifies the classical approach as well. Decoherence describes that the phase of the wave function is averaged out when particles are constantly interacting with their environment, which results in the classical behavior of macroscopic systems. Proteins in vitro or in vivo are permanently in contact with their environment, hence decoherence applies and classical treatment is justified.

3. Force fields Force fields are empirically computable potential energy functions $V(\mathbf{R})$ to approximate the potential energy surface E_e. The function $V(\mathbf{R})$ is a sum of multiple different expressions, which describes the interaction type accurate enough to create a satisfying approximation of the true potential energy surface. A common force field takes the form of Eq. 2.8 [104, 105, 118].

$$
\begin{aligned}
V(\mathbf{R}) &= V_b + V_a + V_{dih} + V_{imp.\,dih} + V_{LJ} + V_{Coul} \\
&= \sum_{bonds\ i} \frac{k_i}{2}(l_i - l_{i,0})^2 \\
&+ \sum_{angles\ i} \frac{f_i}{2}(\varphi_i - \varphi_{i,0})^2 \\
&+ \sum_{dihedrals\ i} \frac{V_i}{2}[1 + \cos(n\phi_i - \phi_{i,0})] \\
&+ \sum_{imp.dih.\ i} \kappa_i(\xi_i - \xi_{i,0})^2 \\
&+ \sum_{pairs\ i,j} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \\
&+ \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}}
\end{aligned}
\tag{2.8}
$$

Here, the bond stretching potential V_b, the bond angle potential V_a and the improper dihedral potential $V_{imp.\,dih}$ are represented by harmonic potential functions, which are similar to the description of classical springs. The proper dihedral potential V_{dih} describes the potential of the out-of-plane bending modes modeled by a cosine with periodicity n and potential barriers V_i. These four terms represent the bonded interactions, which are followed by the two non-bonded expressions V_{LJ} and V_{Coul}. The first modeling the Lennard-Jones potential, which describes the short-range repulsive and attractive dispersion interactions as a function of atomic positions to the power of -12 and -6, respectively. The attractive interactions are derived from the London-dispersion. Here, σ_{ij} and ϵ_{ij} describe the bond length and the depth of the potential well, i.e. the strength of the interaction, respectively. The V_{Coul} term describes the electrostatic interactions and is represented by the Coulomb equation, where q_i denotes the partial charge and ϵ_r the relative dielectric constant, which is usually set to 1 for simplicity. Usually, the parameters of all these terms are different depending on the force fields which are used. If the application demands a different representation, the potential energy terms can differ from the ones in Eq. 2.8. The user usually has various choices of force fields to choose from, for example there are AMBER [119–121], CHARMM [122–124], GROMOS [125–127], OPLS [128,129] etc., which all differ from each other in the parameters which they use to describe the interactions. Usually, the parameters come from fits to thermodynamic quantities and quantum calculations. The parameters of different force fields are purposely fitted to different data to make them more similar to certain experiments. Thus, one force field might be more accurate for a specific molecule in a certain environment than another force field.

The Amber and CHARMM force fields are widely used for protein simulations. Also, the CHARMM force fields [122,130] is most commonly the force field of choice, when dealing with membrane simulations, which contain a lot of small (macro-)molecules, e.g., lipids. The choice of the force field does not only depend on the molecules which are simulated, but they may also be used to reduce the computational cost by coarse graining specific parts in the systems. For example, one of the used force fields for coarse-grained simulations is the so-called MARTINI force field [131].

In this thesis, the AMBER14sb force field is used, because it is established as an accurate and valid force field for protein systems containing parameters for DNA or RNA ligands [132].

2.1.1 Time integration

The GROMACS package provides different types of time integration algorithms to calculate iteratively the position and velocities of all atoms in the system. In the present thesis, we want to focus on the *leap-frog* and the *velocity-verlet* integrator. The *leap-frog* algorithm is a special version of the Verlet algorithm. It computes the velocities at each half integration time step and the according coordinates of the atoms are calculated every full integration time step, according to Eq. 2.9 and 2.10.

$$\mathbf{v}(t + \frac{1}{2}\Delta t) = \mathbf{v}(t - \frac{1}{2}\Delta t) + \frac{\Delta t}{m}\mathbf{F}(t) \tag{2.9}$$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \Delta t \mathbf{v}(t + \frac{1}{2}\Delta t) \tag{2.10}$$

Here, $\mathbf{v}$ is the velocity, $\mathbf{r}$ is the spatial coordinate vector and Δt is the integration time step.

We have to use the *velocity Verlet* integrator since the simulated tempering method is not implemented with the conventional *leap-frog* algorithm in GROMACS. In contrast to the *leap-frog* algorithm, the *velocity-verlet* algorithm calculates the velocity and coordinates simultaneously as shown in Eq. 2.11 and 2.12.

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t) + \frac{\Delta t}{2m}\left[\mathbf{F}(t) + \mathbf{F}(t + \Delta t)\right] \tag{2.11}$$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \Delta t \mathbf{v} + \frac{\Delta t^2}{2m}\mathbf{F}(t) \tag{2.12}$$

Velocity Verlet and *leap-frog* algorithm are both symplectic, which means they preserve the total energy of the system, and are time-reversible, which matches in most cases physical processes. The difference between the two algorithms is that the *Velocity Verlet* algorithm provides a more accurate energy conservation with a cost of extra computation. However, with corresponding starting points, both algorithms will produce an identical trajectory.

Since the fastest motions in molecules are usually the hydrogen atom vibration modes, one has to choose the integrator frequency small enough to minimize errors. Thus, in simulations which consider the hydrogen vibrations, the time step must be at least 1 fs. However, the hydrogen vibrations have very little impact on the overall motions of macromolecules. Thus, in MD simulations the vibrations are usually constrained via the LINCS algorithm [133] for all bonds except the water molecules and via the SETTLE algorithm [134] for the water molecules. Those constrains allow the usage of a 2 fs integration time step instead of a 1 fs.

2.1.2 Temperature and pressure coupling

Usually, MD simulations are performed in an isobaric and isothermal ensemble, called NPT, to reproduce experimental conditions.

Isothermal treatment is necessary because of numerical inaccuracies, force cut-offs or dissipative work in non-equilibrium simulations which may lead to nonphysical heating or cooling in different areas of the system. Hence, the simulation box must be coupled to a temperature bath. In this thesis, we used the velocity-rescale scheme (v-rescale) [135], which is closely related to the Berendsen thermostat [136], but introduces an additional stochastic term [135]. The v-rescale method ensures that the temperature of the system decays exponentially over time to a defined temperature [135, 136].

In an NPT ensemble, the system is isobaric and hence it is coupled to a pressure bath to ensure the pressure stays approximately constant over time, similar to the isobaric treatment. In this thesis, the Parrinello-Rahman barostat is used for the conventional (meaning without any enhanced sampling method applied) production simulations, since it guarantees the correct NPT ensemble. For equilibrium simulations and the simulated tempering simulations the Berendsen barostat is used. However, it has been stated by Rizzi et al. [137] that expanded ensemble simulations, such as simulated tempering simulations, in combination with the Berendsen barostat can produce artifacts in the binding free energies. At the beginning of this project, before GROMACS2020, it was only possible to combine the velocity-verlet integrator with the Berendsen barostat for expanded ensemble NTP simulations. Thus, this possible problem could not have been avoided. However, we do not expect that the use of the Berendsen barostat will change the results of this work significantly, because no free energy calculation on basis of the energies during the simulations were performed. Instead we will try to calculate free energies from transition rates, which will be explained later.

2.1.3 Computation of non-bonded interactions

As stated in the force field section, the non-bonded interactions require the calculation of potentials of pairs of all particles and therefor scale quadratically with the number of particles N in the system. Thus, methods are needed to reduce the computational cost of calculating non-bonded interactions can be reduced by defining a spatial cut-off with a certain radius around each atom, to reduce the number of interaction partners. Usually, such a radius is chosen to be 1.0-1.4 nm. The cut-off is justified by the fast r^{12} and r^{6} decay of the LJ potential.

The Coulomb interactions is not simply switched off with a cut-off because that would cause artifacts during the simulation due to missing long-range electrostatic interactions, since the Coulomb potential decays only relatively slow with r. Thus, the prominent particle-mesh Ewald (PME) method was developed, which handles the electrostatics by assigning the charges to a grid [138]. Then, the long-rage potential is handled in reciprocal space as a simple sum. This calculation can be done by a fast Fourier transformation, which leads to a $N \log N$ scaling instead of a $N^{3/2}$ scaling as the original Ewald summation.

2.2 Simulated Tempering

Simulated tempering [109] (ST) is an enhanced sampling technique, which we used to overcome the sampling problem of the DEAH-helicases. In ST, the temperature is a dynamical variable rather than a static parameter. This allows the system to be simulated at higher temperatures, which makes the transition of barriers in conformational space more likely. The on-the-fly changes in temperature render ST a so-called expanded ensemble technique.

In ST, we apply weights to the different temperature states to ensure a uniform distribution of all states. The flat distribution ensures that the ground state is visited frequently enough to ensure the sampling of new minima in conformational space with the correct equilibrium probability. The probability distribution $P(\mathbf{x}, i)$ is

$$P(\mathbf{x}, i) \propto e^{-\beta_i H(\mathbf{x}) + w_i}, \tag{2.13}$$

where $\mathbf{x}$ is a configuration of the system, i is the corresponding temperature state, $\beta_i = 1/k_B T_i$, w_i is the weight of state i and $H(\mathbf{x})$ is the Hamiltonian of the system in the configuration $\mathbf{x}$. Thus, the partition function Z is

$$Z = \sum_i \int e^{-\beta_i H(\mathbf{x}) + w_i} \mathrm{dx} = \sum_i Z_i e^{w_i}, \tag{2.14}$$

where Z_i is the partition function of temperature state i [139].

The transition of the temperature is realized via the Metropolis algorithm, which performs a neighbor random walk on a pre-defined temperature-ladder $T_0 < ... < T_{N-1}$ with N temperature states. The difference between two neighboring temperature states ΔT is chosen small enough to ensure an overlap of the potential energy distributions of the system at both temperatures. After a given time step a transition from temperature T_i to temperature T_j is attempted, where $i = 0, ..., N-1$ and $j = i \pm 1$. Following the Metropolis criterion, if j becomes less than 0 or larger than $N - 1$, the transition is immediately rejected. All other transitions have the acceptance probability given by

$$p_{ij} = \min \left[1{,}\mathrm{exp}^{\Delta_{ij}} \right], \tag{2.15}$$

with

$$\Delta_{ij} = [\beta_i E_i(\mathbf{x}) - w_i] - [\beta_j E_j(\mathbf{x}) - w_j], \tag{2.16}$$

where $E_k(\mathbf{x})$ is the potential energy function of a configuration $\mathbf{x}$ in the temperature T_k.

One disadvantage of ST is the necessity of choosing of the right initial weights. For an optimal choice, the free energies of the respective states are required. However, these energies are not known a priori. To tackle this problem, Pande et al. [140] published an approximation method for the weights. At each temperature T_k a short simulation is performed after which the average potential energies $\langle E \rangle_k$ of the respective temperature state is determined. Then, the weights are estimated by

$$w_j - w_i = (\beta_j - \beta_i) \frac{\langle E_i \rangle + \langle E_j \rangle}{2}, \tag{2.17}$$

where the sum of all w_k are equal to an arbitrary constant. Because the weights of each temperature state is only roughly estimated by this method, there is still a so-called burn-in required in order to reach optimal weights eventually. During simulations with larger conformational transitions, the potential energies of the states can change significantly, which also affects the corresponding weights. Therefore, the weights have to be constantly updated at a fixed frequency to ensure a uniform sampling of states during the whole simulation. An example of the burn-in and the temperature profile during a ST run is shown in Fig. 5.1.

In summary, the used random walk in temperature space enables the system to cross energy barriers between states much faster than in conventional simulations, thus making ST a suited approach for systems with a rough energy landscape. Shaw et al. [111] demonstrated a one order of magnitude reduction in the computation time when using ST in comparison to conventional simulations. A simplified flowchart of the ST procedure is shown in Fig. 2.1.

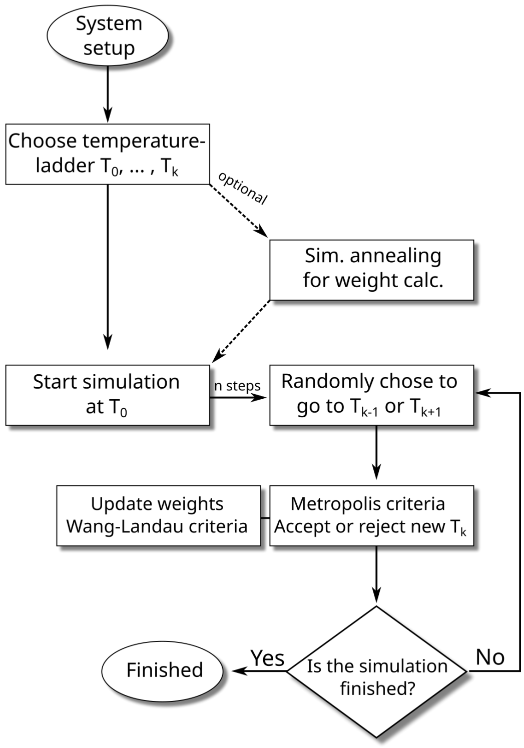

System setup
Choose temperature-ladder T_0, ... , T_k
optional
Sim. annealing for weight calc.
Start simulation at T_0
n steps
Randomly chose to go to T_{k-1} or T_{k+1}
Update weights Wang-Landau criteria
Metropolis criteria Accept or reject new T_k
Is the simulation finished?
Yes
No
Finished

Figure 2.1: *Simplified flow chart for the simulated tempering method. 1. Specify temperature ladder $T_0, ..., T_k$. (1.1 Optional: Initial weights estimation for each temperature state) 2. Start simulation at T_0. 3. Randomly move to T_{k-1} or T_{k+1} in temperature space after n steps. 4. Metropolis criterion. Simultaneously, update the weights of temperature states with the Wang-Landau method to maintain a uniform occupancy over temperature states. 5. Ending criterion.*

2.3 Adaptive Sampling

The Adaptive Sampling method is a computational approach to enhance the sampling of the conformational space. Unlike other methods, such as simulated tempering or replica exchange, this technique does not modify the Hamiltonian of the system to facilitate accelerated sampling. Instead, it employs a parallel execution of many conventional MD simulations, which enhances the likelihood of observing rare events and/or transitions to other local minima. In addition, multiple parallel simulations help to sample underexplored regions in the conformational space.

Due to the short parallel simulations, a stop-and-go mechanism can be exploited. Here, a reference conformation state is pre-defined and the goal is to see a transition to this reference state. If the system is found to be in a preferred conformational state, i.e., a state which is close to the reference, at the end of one or more of the parallel simulation, the user may select this state(s) as starting point for new N parallel simulations. Hence, the vicinity of the conformational landscape of the chosen state is explored further. Repeated iteration of this process may lead to the discovery of multiple domains in a ”divide and conquer” approach or to the pre-defined end state of interest that is inaccessible in a single simulation due to multiple high-energy barriers [110]. A flowchart depicting the adaptive sampling principle is presented in Fig. 2.2.

The underlying principle of the Adaptive Sampling method involves the periodic in-process evaluation of the short trajectories obtained from each batch, enabling the user to guide the system towards desired states without the use of a biasing potential during the simulations. Thus, increasing the likelihood of escaping energy minima. This is possible because rare events, which often occur after prolonged waiting times, are typically accompanied by rapid transitions [141]. Hence, the reference conformational state must be defined by features, such as atomic distances and angles, which are distinctly different from their value in the starting state and have a long enough lifetime to be observable.

The two main advantages of the technique are: 1. the ability to travel along a path in the energy landscape in an unbiased and - if desired- a user-controlled manner and 2. the reduction of the overall wall time and computational time, because of the trivial parallelization of the simulations and the guided exploration of the conformation space [141, 142].

A disadvantage is the requirement a large cluster with many nodes, which is able to parallelize a lot of simulations. Other disadvantages are: 1. the complexity of the data can be hard to analyse because of the amount of data, that is produced, and 2. even though the Hamiltonian is not modified, the system is driven by a certain selection criterion, when choosing a new starting state. Hence, the sampled path in energy landscape is not guaranteed to be the true minimal free energy path.

The method has been widely used in MD studies on proteins trying to identify pathways on long time scales, like ligand binding, protein folding and conformational transitions of proteins [110, 112, 142, 143]. Usually, an additional analyzing technique can be used to take advantage of all the produced data after and during the adaptive sampling method. This way, a lot of additional quantitative information about kinetics and dynamics can be gained. One of the most famous techniques for the analysis of a adaptive sampling outcome is the construction of a Markov State Model (MSM), which is described in section 2.7 Markov State Model in more detail.

Because Prp43 is large and complex system with a high probability of containing a lot of different molecular switches, a single enhanced sampling technique might not be enough to observe a complete translocation cycle of translocation in a feasible amount of time. After a few test simulations, we combine the ST and the AS method to further improve the sampling beyond the ability of the two methods alone. The combination is easily achieved, because both methods operate orthogonally to each other and do not require the definition of complex reaction coordinates.

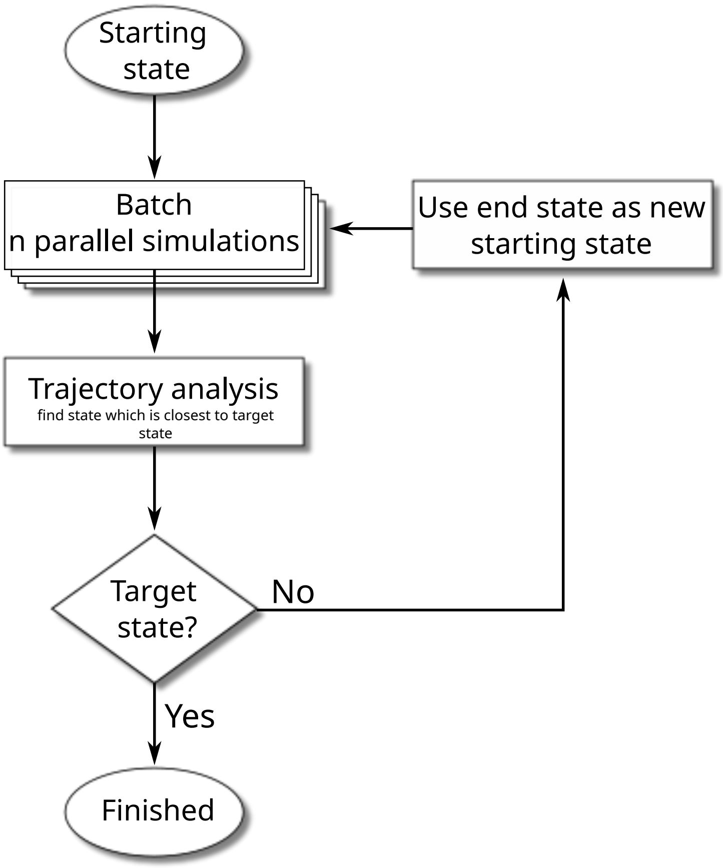


Figure 2.2: *Simplified flowchart for the adaptive sampling routine. A batch of a specified number of separate simulations is initiated using a starting structure. Upon completion of the simulations, the resulting trajectories are evaluated and compared with a pre-determined target structure or, more specifically, the target values of selected features. If a simulation's end state is deemed sufficiently close to the target set, the adaptive sampling process is deemed complete. If not, a new batch of simulations is performed by using the conformation closest to the target state from the previous batch as the starting point.*

2.4 Random acceleration molecular dynamics

The spontaneous entrance and exit of a ligand in and out of a protein usually happens on large timescales and therefore is extremely hard or even impossible to observe during conventional MD simulations. It is shown by experiments and computational studies, that the phosphate release in RNA helicases and other proteins is a rate limiting factor during an ATP hydrolyzation cycle with rate constants of a few releases of phosphate per second [144]. Hence, the observation of a spontaneous unbinding event of phosphate would take seconds of simulation time, which is unfeasible in a limited amount of time. Random acceleration molecular dynamics (RAMD) is able to overcome this drastically high time scales by imposing a force with random orientation on the COM of the ligand of interest in addition to the acting forces introduced by the MD force field [145]. In order to avoid disruption of the secondary structure of the protein by the ligand, which can lead to unreasonable exit tunnels and penetrations into inaccessibly areas of the protein, a fine-tuning of the specific RAMD parameters has to be carried out. A flowchart of the principle of this method is shown in Fig. 2.3.

In the current literature, RAMD has been proven to be an effective method for ligand dissociation simulations, which would be computational too costly to detect with conventional simulations [146]. To name only two, the RAMD method has been used in the entry and exit pathways of carazolol in a β_2-adrenergic G-Protein [147] and in an unbinding study of vitamin D from the vitamin D receptor [148]. Such studies are crucial for a more detailed investigation of pathways for drug dissociation and binding. Hence, the RAMD method can help to find potential drugs, that not only fit into the ligand pocket but can also traverse in and out of the pocket. In addition, RAMD can help to estimate one of the most important properties for drug efficiency, namely the lifetime of the drug-target complex. For example, Wade et al. used an altered version of the RAMD method to estimate the residence time of 70 different ligands of the N-terminal domain of HSP90α [149].

In our study, we use the RAMD method to investigate a potential exit tunnel for the phosphate group after ATP hydrolysis. Hence, we applied RAMD on a modeled structure of 5LTA, in which we altered the ATP into an ADP plus phosphate ion.

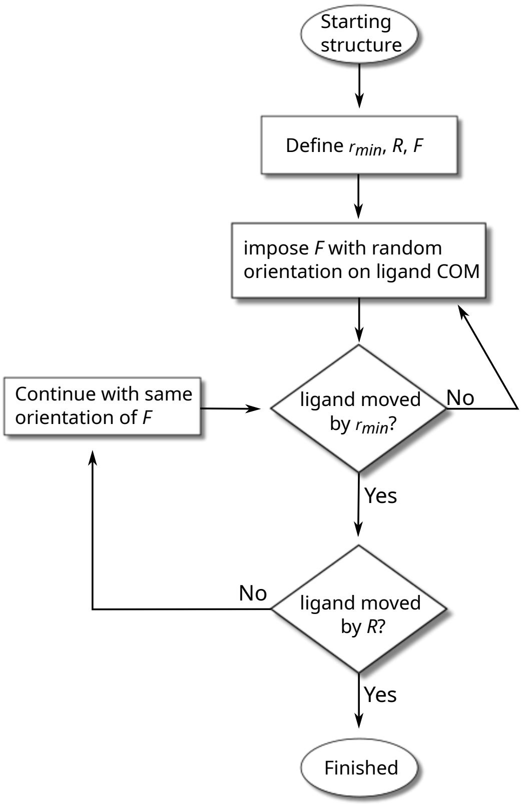

Starting structure
Define r_{min}, R, F
impose F with random orientation on ligand COM
Continue with same orientation of F
ligand moved by r_{min}?
No
Yes
No
ligand moved by R?
Yes
Finished

Figure 2.3: *Flowchart for the RAMD method. After defining minimal distance r_{min}, maximal distance R and force F, the force is applied with random orientation on the center of mass of the moveable ligand. After N steps, the algorithm checks if the ligand moved by at least the minimal distance. If the minimal distance is exceeded, the algorithm checks if the ligand moved more than the maximal distance from its initial position. In this case, the dissociation is successful.*

2.5 Principal component analysis

The Principal Component Analysis (PCA) is an established approach for dimensionality reduction in various fields, from economics to biophysics. It can unravel low dimensional patterns from a statistical distribution of a high-dimensional set of data [150–152]. In the field of MD, covariance matrix C is constructed from the Cartesian coordinates of N atoms of a biological system:

$$C = \langle (\mathbf{r}_i - \langle \mathbf{r}_i \rangle)^T (\mathbf{r}_j - \langle \mathbf{r}_j \rangle) \rangle, \tag{2.18}$$

where $\mathbf{r}_i$ are the mass-weighted Cartesian coordinates and $\langle ... \rangle$ are the averages over all sampled structures. Solving the eigenvalue problem of C leads to the eigenvectors $\mathbf{v_i}$ and the eigenvalues λ_i, which describe the collective modes and their magnitudes, respectively. Projecting the original data onto the eigenvectors leads to the projected data $\mathbf{z}$:

$$\mathbf{z}^T = \mathbf{r}^T \mathbf{V}, \tag{2.19}$$

where $\mathbf{V}$ is the matrix with the eigenvectors $\mathbf{v_i}$ as columns, sorted in descending order by their corresponding eigenvalues λ_i and $\mathbf{z}^T$ is the transposed projected data on the eigenvectors. Thus, the Principal Component with the largest corresponding eigenvalue represents the reduced representation of the motion with the largest contribution to the variance of atomic fluctuations [153, 154].

2.6 Time-structure independent components analysis

The time-structure independent components analysis (tiCA) is closely related to the PCA, but instead of finding the motions, which maximize the variance of the degrees of freedom, the tiCA method searches for the slowest motions in the protein dynamics by maximize the auto-correlation time of these motions. Here, a covariance matrix C of a n-dimensional time series is needed:

$$C = \langle (\mathbf{r}(t) - \langle \mathbf{r}(t) \rangle)^T (\mathbf{r}(t) - \langle \mathbf{r}(t) \rangle) \rangle, \tag{2.20}$$

where $\mathbf{r}(t)$ being the mass-weighted Cartesian coordinates. In addition, a time-lagged covariance matrix $\overline{C}$ needs to be constructed as follows

$$\overline{C} = \langle (\mathbf{r}(t) - \langle \mathbf{r}(t) \rangle)^T (\mathbf{r}(t + \Delta t) - \langle \mathbf{r}(t) \rangle) \rangle, \tag{2.21}$$

where Δt is the respective lag time. Then, the following generalized eigenvalue problem is solved:

$$\overline{C} V = C V \lambda, \tag{2.22}$$

where V and Λ are the eigenvector and eigenvalue matrices respectively. Again, by projecting the eigenvectors which correspond to the slowest motions, e.g., the ones with the highest corresponding eigenvalues (highest autocorrelation), onto the data, we yield a reduced dimensionality representation of that data [155, 156].

2.7 Markov State Model

Markov State Models (MSMs) are stochastic models widely used to describe systems that transition between different states over time. They play a pivotal role in modeling the long-timescale dynamics of molecular systems within the field of Molecular Dynamics (MD). These models provide valuable insights into the conformational landscape of complex biomolecular systems, shedding light on transitions between various states and the kinetics of these processes. Markov Chains are defined by two fundamental properties: the Markov property and the principle of detailed balance. The Markov property states that the future state of the system depends solely on its current state and is independent of its past states, given the current state. In other words, the system's future behavior is ”memoryless,” and the transition probabilities between states are constant over time. The detailed balance property is a crucial concept in the context of Markov chains and is associated with systems in thermodynamic equilibrium. It refers to a specific balance between transition rates, meaning the net probability flux between any two states in the chain becomes zero, meaning that the system is in equilibrium and not biased towards any particular direction. However, in MSMs, only the Markov property must hold true, since the transitions between states in an MSM are usually coming from non-equilibrium simulation and their kinetic clustering, which leads to transitions that may not fulfill the detailed balance requirement. In this work, we present the construction and the application of an MSM on the use case of DEAH-helicases. While we won't delve into all the mathematical intricacies and theoretical underpinnings, interested readers can refer to the comprehensive book by Pande, Bowman, and Noe [157] and to the PyEmma [158] documentation, which was used in the construction and analysis of the MSM.

To create an MSM, one typically combines data from multiple short MD simulations, which may result from adaptive sampling procedures. The first step in constructing an MSM involves clustering the sampled conformations based on a geometric criterion. These clusters are termed microstates and represent conformations that are either similar or closely related in the conformational space. Various clustering algorithms can be employed for this task, such as k-Means clustering [159], k-Center clustering [160], and k-Medoids clustering [161].

Analyzing a large number microstates can be challenging, which can be addressed by further aggregating them into larger entities called macrostates. Kinetic clustering techniques, such as Perron-Cluster Cluster Analysis (PCCA) [162,163] and Robust PCCA (PCCA+) [164,165], are commonly employed to form these macrostates. This simplifies the phase space and makes it more amenable to analysis and interpretation. These clustering approaches are not performed on the raw spatial trajectories given by MD simulations, but rather by a dimensional-reduced subspace of the raw data. High dimensional data can lead to the so-called "curse of dimensionality" and lead to difficulties in the interpretation and representation of the results. The curse of dimensionality describes the phenomenon of an exponential increase of computational cost with the number of dimensions. To solve this issue, tiCA [166,167] can be used to extract motions with the highest autocorrelation time and plot the slowest motions on to each other in a 2D plot to have a more interpretable representation of the spatial data. Then, the microstate and macrostate clustering can be performed and projected onto these so-called tiCs to gain additional insight into the connection and kinetics of the different states. When constructing an MSM, several parameters need to be carefully chosen, including the lag time, features for dimension reduction, the number of clusters, and the number of macrostates. Moreover, a critical evaluation and validation of the constructed MSM are crucial. This validation involves checking implied timescales, verifying Markovian behavior, ensuring detailed balance (if applied), and confirming ergodicity. These checks are essential to ensure that the resulting MSM is a physically meaningful and interpretable model of the system. Once a valid MSM is constructed, it can provide a wealth of additional information about the system and the relationships between different states. For example, one can estimate the free energy of each state using the expression $G_i = -kT\ln(P_i)$, where P_i represents the population of the state i.

In summary, Markov State Models are powerful tools used in MD to explore the long-timescale dynamics of molecular systems. By capturing the essential conformational states and their kinetics, MSMs offer valuable insights into complex biological processes and hold significant promise for advancing our understanding of molecular behavior.

Effects of ligands on the conformations of DEAH helicases

Helicases are essential for life since they play major roles in the genome stability. Thus, helicases were the subject of a lot of studies in the last decades. Only recently, more and more structures of the monomeric DEAH helicases were resolved by different groups via different methods, such as crystallography and cryo-EM. These structures yield crucial insights into the structure of helicases, which allow to construct hypotheses about their dynamics. However, because such methods are only able to resolve single pictures of an ensemble average or only states in a crystal unit cell, the real dynamic of the enzymes remains hidden. Molecular dynamics simulations are able to close this gap by producing possible trajectories of protein structures which capture the motion of enzymes. Here, MD can corroborate the found structures by simulating the enzyme's native state and check the stability of the complex, enforce a change in the enzyme's structure by removal or insertion of a ligand into or onto the protein or simply by applying an artificial force on a part of the protein.

The following results are published in the journal *Biological Chemistry* [168]. We investigated the influence of the removal or/and the replacement of ligands from different crystal structures from various RNA helicases. This approach is useful to get an overall understanding of the effects of the ligands on the three main domains CTD, RecA1 and RecA2 of the corresponding enzymes. Hence, we analyzed the RecA1–RecA2 distance and the CTD–RecA2 distance. For the RecA1–RecA2 distance, we measured the distance between the COM of five beta sheets in RecA1 and RecA2, respectively, since they represent COMs which are stable against inner fluctuations of loop regions and residue sidechains. For the CTD–RecA2 distance, we have chosen to calculate the minimal distance between a LYS in the ratchet-like domain (for example LYS605 in PDB ID 5LTA) and an ASP in the RecA2 domain (for example ASP321 in PDB ID 5LTA), since these amino acids tend to form hydrogen bonds if the RNA cleft gets disrupted.

3.1 Methods

MD simulations of the conventional simulations concerning the investigation and analysis of the various crystal structures were set up as follows with GROMACS 2018.1 [169]. The initial structures of Prp43, Prp22 and Prp2 were taken from the protein data bank (5LTA [29], 5D0U [170], 5LTK [170], 5LTJ [170], 6I3P [30]) representing Prp43 interacting with U_7 RNA and ATP (5LTA), with ADP (5D0U) and ATP (5LTK/5LTJ). Any missing residues were added via Modeller [171]. The ligands were removed according to the desired investigation as stated in the corresponding Results section. The structures were placed into a simulation box of a dodecahedron with a minimum distance of 1 nm between the protein and the box borders. The systems were solvated with water and neutralized with K^+ counter ions. Interactions of protein and RNA were described with the Amber14SB force field [132]. The ATP, ADP, Mg^{2+} and water were positioned to match the positions of the ATP and ADP analogues and the crystal water in the corresponding structures. Parameters of the ATP and ADP were taken from Carlson et al. [172], translated into GROMACS format with the ACPYPE software [173]. Water was modeled with the TIP3P model [174], and parameters for K^+ were taken from Joung et al. [175]. The energy of the system was minimized with the steepest descent algorithm. Then, the system was equilibrated for 100 ps with position restraints acting on the heavy atoms including RNA and Mg ($k = 1000\,\mathrm{kJ\,mol^{-1}nm^{-2}}$). Electrostatic interactions were described with the particle-mesh Ewald method [138]. Dispersion interactions and short-range repulsion were described together with a Lennard-Jones potential with a cut-off at 1 nm. The temperature was controlled at 300 K using velocity-scaling [135], thereby coupling protein, RNA, Mg^{2+}, and ATP (if present) to one heat bath while coupling water and K^+ to a second heat bath ($\tau = 0.5\,\mathrm{ps}$). The pressure was controlled at 1 bar with the Parrinello-Rahman barostat ($\tau = 5\,\mathrm{ps}$) [176]. The md-vv integrator was used for simulated tempering (ST) simulations and the md integrator was used for all other simulations, both with an integration time step of 2 fs. The geometry of water molecules was constrained with SETTLE [134]. All other bonds were constrained with P-LINCS [133].

3.2 Results

3.2.1 Prp43

5LTA

The crystal structure with the PDB ID 5LTA resembles the Prp43 helicase in its RNA and ATP loaded state. In this structure, the RNA cleft has a defined structure which forms a tunnel for the ssRNA. The two RecA domains, RecA1 and RecA2, are in proximity to each other and therefore form a closed interface with each other. Here, ATP acts as a bridge between the two domains.

We performed several different simulations to identify the behavior of Prp43 after removing or replacing its ligands, all starting from the 5LTA crystal structure. More precisely, we carried out four 200 ns of the Prp43·RNA·ATP complex, ten 80 ns simulations of the Prp43·RNA complex, six 300 ns simulations of the Prp43·ATP complex, ten 800 ns simulations of the Prp43·ADP and ten 500 ns simulations of the Prp43 apo structure. The analysis of the domain movements is shown in Fig. 3.2.

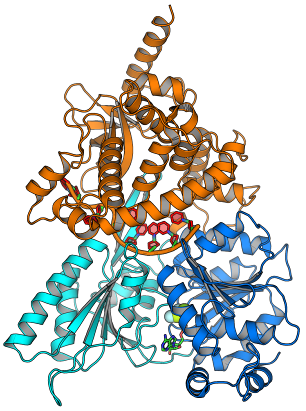


Figure 3.1: *The crystal structure 5LTA from Ficner et al. [30]. The structure represents the Prp43·RNA·ATP complex*

The Prp43·RNA complex has a stable RecA distance and CTD–RecA2 distance, which is shown in the two distance plots of Fig. 3.2A. The slight CTD–RecA2 distance increase might occur due to the poor model of RNA inside the RNA cleft of 5LTA. The other possibility might be that the removal of ATP from the ATP pocket has a slight impact on the RNA tunnel. We can conclude that the RNA strand has a strong impact in the overall conformation, because the RNA itself defines the RNA tunnel and has strong interaction points with both RecA domains. However, we would expect a complete opening of the RecA domains after ATP removal. This conformational transition is probably not observed, because it occurs on way longer time scales than 80 ps.

The Prp43·RNA·ATP complex is the native state of the crystal structure. The RecA domain interface and the RNA cleft are overall stable over the 200 ns simulations time. Higher fluctuations in the domain distances than in the Prp43·RNA complex are observed.

The Prp43·ATP complex possesses a stable ATP pocket even after the removal of RNA as shown in the RecA distance plot in Fig. 3.2C. However, as soon as RNA is removed from the complex, the RNA tunnel gets disturbed, which is noticeable in five out of six simulations as a decrease in the CTD–RecA2 distance of about 0.4–0.7 nm. In two simulations this leads to a H-bond between the measured LYS605 and ASP321 and therefor to a left tilt of the CTD domain towards the RecA2.

The Prp43·ADP complex experiences a small perturbation of the ATP binding pocket which results in a small increase of the RecA distance. This might happen because of the missing interaction link of the γ-phosphate from the original ATP with the RecA2 domain. Similar to the Prp43·ATP complex, the CTD–RecA2 distance is decreasing during the first 100 ns and by the end of six out of ten simulations an H-bond form between LYS605 and ASP321 as seen in Fig. 3.2D.

The Prp43 apo structure carries out the most noticeable movement out of all complex simulations of 5LTA. The CTD–RecA2 distance is much more moving than in the Prp43·ATP and Prp43·ADP complexes. As shown in Fig. 3.2E, the distance between the domains can increase, which indicates an opening of the RNA tunnel. A similar trend in observed in the ATP pocket, as in two out of ten simulations a large increase between the RecA domains occurs. This increase of more than 0.6 nm can be seen as a full rupture of the RecA interface. However, we did not observe a sensor serine loop-to-helix transition in either simulation. Therefore, the two simulations of the Prp43 apo structure, which undergo a rupture of the RecA interface, cannot be seen as a true open state in the sense of the Prp43·RNA complex open state during the translocation.

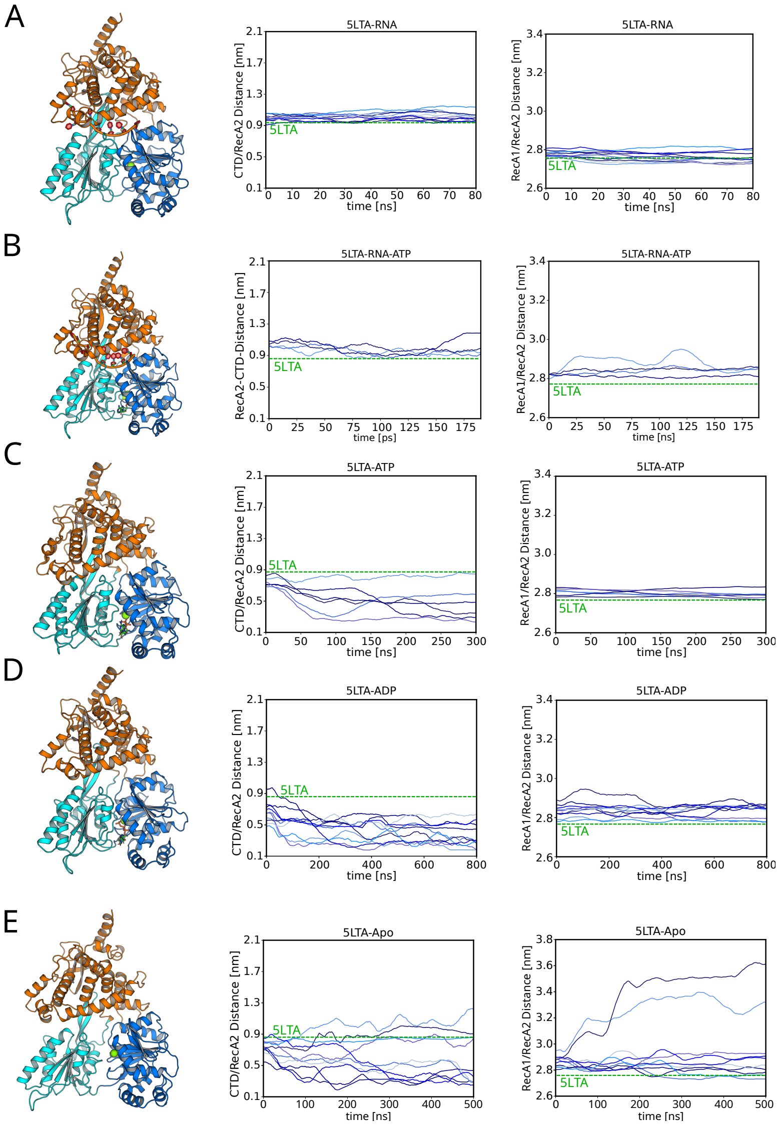


Figure 3.2: *The analysis of all 5LTA simulations. (A) The Prp43·RNA complex. The helicase is overall rigid, and no large conformational change is observed. (B) The Prp43·RNA·ATP complex. The helicase is in its native state as in the crystal structure. The RecA distance and RNA-cleft opening are constant besides some small fluctuations during the 200 ns. (C) The Prp43·ATP complex. The distance between the CTD and RecA2 is decreasing while the RecA domains stay close to each other. (D) The Prp43·ADP complex. The distance between the CTD and RecA2 is decreasing and the RecA domains are more flexible, which results in a small change in the RecA distance. (E) The Prp43 apo structure. The distance between the CTD and the RecA2 is varying a lot. Also, the RecA distance is changing drastically in two simulations*

5D0U

The crystal structure 5D0U resolved by Ficner et al. [170] resembles the Prp43·ADP complex. The complex is in a closed-like state and the sensor serine sequence is already in the helix conformation. Thus, according to this crystal structure, the loop-to-helix transition occurs during or immediately after the ATP hydrolysis. Additionally, the structure shows a clear open RNA cleft at the proximity of the RecA2 domain. In our study, starting from 5D0U, we performed six 300 ns simulations of the native Prp43·ADP complex and ten 500 ns simulations of the Prp43 apo-state by removing ADP from the complex. For analysis, we tracked the change of the RecA domain distance and the size of the entrance of the RNA cleft. In case of the RecA domain distance, we measured the distance between the COM of four β-sheets of the RecA2 domain and the COM of five β-sheets of the RecA1 domain. In case of the RNA cleft opening, we captured the distance between two crucial residues (E320 and S614) close to the entrance of the RNA cleft which is formed by the CTD and the RecA2 (Fig. 3.4).

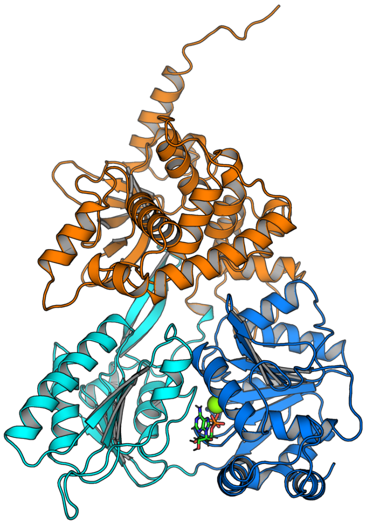


Figure 3.3: *The crystal structure 5D0U from Ficner et al. [170]. The structure represents the Prp43·ADP complex*

The 5D0U Prp43·ADP complex shows a similar behavior like the Prp43·ATP and Prp43·ADP complexes of 5LTA. (Fig. 3.4A) The RecA is stable, but although, the complex is in its native crystal conformation, the CTD and RecA2 domain come closer to each other by 0.2–0.7 nm. In contrast to the Prp43·ADP complex of PDB ID 5LTA, only one simulation reaches a CTD–RecA2 distance of ~0.2 nm after 300 ns. The observation indicates that the RNA tunnel formed between RecA2 and CTD in 5D0U is more stable than in the 5LTA crystal structure after RNA removal. In 5LTA, the RNA tunnel will break down after RNA removal, since the RNA gives it a defined conformation, which is already near RecA2.

The 5D0U Prp43 apo-structure undergoes similar changes like the 5LTA apo structure. (Fig. 3.4B) The CTD–RecA2 distance domain show high fluctuation in the end conformations of the 10 simulations, which indicates a flexible CTD domain. The RecA interface tends to open due to the loss of the ADP contacts between the two domains. In one of the ten simulations, the RecA domains drift apart to a distance of ∼3.15 nm, which is similar to the RecA domain distance of the open conformation of Prp22·RNA complex (PDB ID: 6I3P). The opening of the RecA domains might be easier, because the 5D0U crystal structure resembles a conformation after the loop-to-helix transition of the sensor serine. Thus, probably one of the rate-limiting steps (ATP hydrolysis and/or serine flip) has already occurred during or right after ATP hydrolysis.

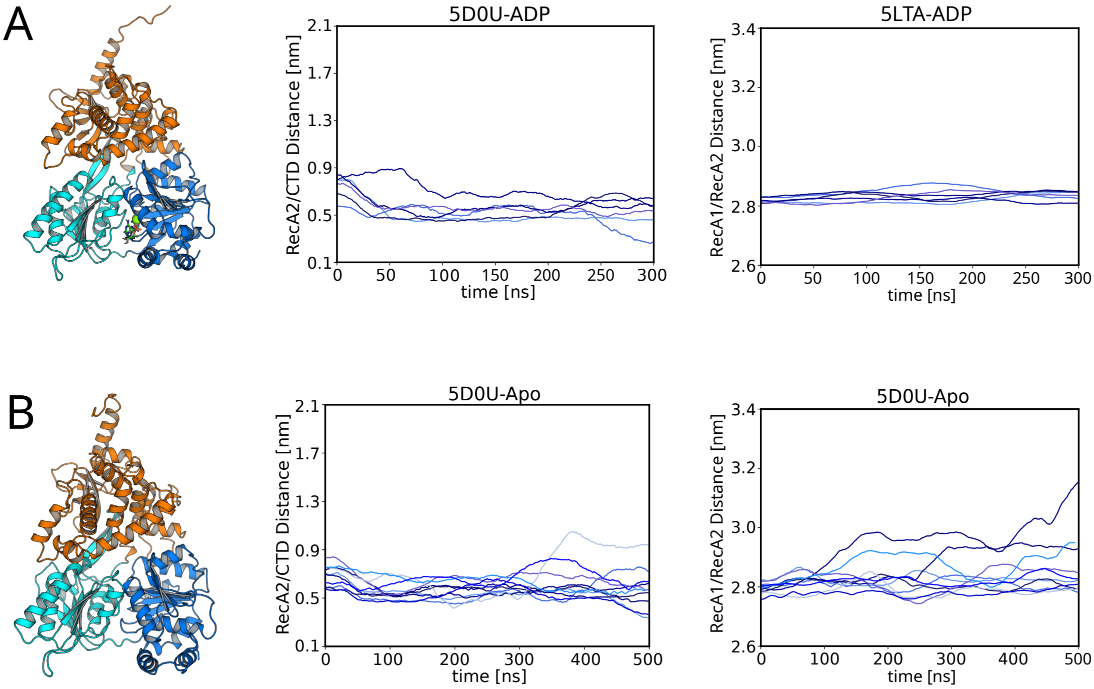


Figure 3.4: *The analysis of the 5D0U simulations.*
A) The Prp43·ADP complex of 5D0U. The helicase is overall rigid, and no large conformational change is observed. B) The Prp43 apo-structure of 5D0U. The distance between the CTD and RecA2 is decreasing while the RecA domains stay close to each other.

5LTJ & 5LTK

The 5LTJ and 5LTK crystal structures resolved by Tauchert et al. [29] represent the helicase Prp43 as ATP-complex. According to Tauchert et al. [29], the enzyme's CTD moved away from RecA2 and RecA1, which results in an open RNA cleft with no defined RNA tunnel (Fig. 3.5). Thus, the conformations resemble the pre-catalytic state of Prp43 before RNA binding. To study the stability of the conformations of 5LTJ and 5LTK, we performed four 250 ns 5LTJ·ATP complex simulations, six 100 ns 5LTJ·ADP complex simulations, four 300 ns 5LTJ·ATP complex simulations, and four 300 ns 5LTJ·ADP complex simulations. As mentioned above, we analyzed the RecA domain distance and the opening of the RNA cleft (Fig. 3.6).

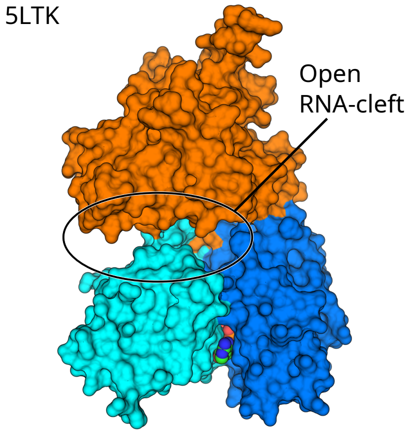


Figure 3.5: *The crystal structure 5LTK from Ficner et al. [29] in surface representation. The structure represents the Prp43·ATP complex with an open RNA cleft.*

Generally, the RecA distance did not vary significantly between the four different setups. This indicates that the native ATP complexes of the two crystal structures are stable at the RecA domain interface.

The 5TLJ and 5LTK Prp43·ADP complexes show a decrease of the CTD–RecA2 distance of ~0.3-0.9 nm. The large difference between the distances suggests, either the CTD or RecA is highly flexible, or the conformations have not reached equilibrium yet. After investigation of the trajectories, we can conclude, that even a 0.3 nm decrease in the distance between CTD and RecA2 is enough for a collapse of the open RNA tunnel as shown in Fig. 3.5. Thus, the RNA tunnel opening is not stable in our simulations (Fig. 3.6A/C).

The 5LTJ and 5LTK Prp43·ATP complexes undergo a decrease in the CTD–RecA2 distance of ~0.3-0.5 nm. In contrast to the Prp43·ADP complexes, the CTD–RecA2 distance equilibrates after roughly 150 ns at an average distance of ~1.5 nm. Thus, although we observe a collapse of the RNA opening in the ATP complexes, the conformation is equilibrating faster than the ADP complex counterparts. Conclusively, the change from ATP to ADP has a slight increase in the stability of the complex and/or the flexibility of the CTD (Fig. 3.6B/D). However, a stable open RNA tunnel as proposed in the crystal structures 5LTJ and 5LTK was not observed.

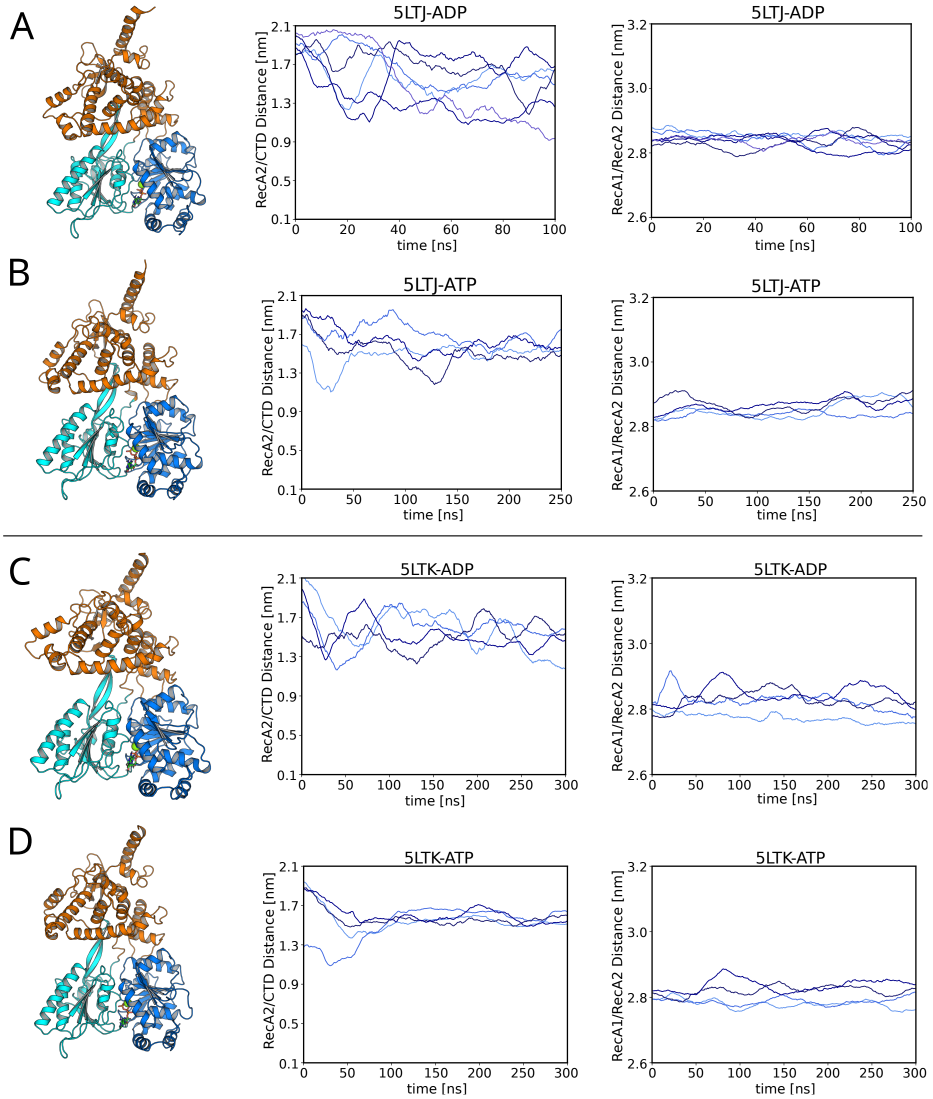


Figure 3.6: *The analysis of the 5LTJ and 5LTK simulations. The RecA interface is stable and the open RNA cleft collapses in all simulations.*
A) The Prp43·ADP complex of 5LTJ and the trend of the RecA domain distance and of the stability of the RNA cleft open state. B) The Prp43·ATP complex of 5LTJ and the trend of the RecA domain distance and of the stability of the RNA cleft open state. C) The Prp43·ADP complex of 5LTK and the trend of the RecA domain distance and of the stability of the RNA cleft open state. D) The Prp43·ATP complex of 5LTK and the trend of the RecA domain distance and of the stability of the RNA cleft open state.

3.2.2 Prp22

6I3P

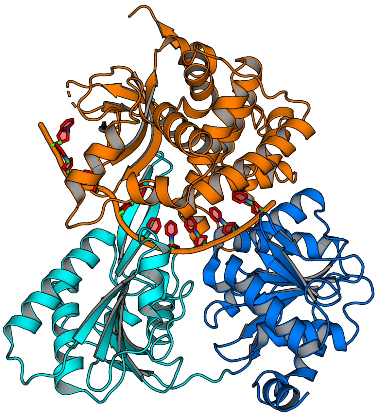


Figure 3.7: *The crystal structure 6I3P from Hamann et al. [30]. The structure represents the Prp22·RNA complex with an open RecA interface.*

The 6I3P crystal structure resolved by Hamann et al. [30] represent the Prp22·RNA complex. The structure has a defined RNA tunnel in which ssRNA is bound. The structure forms an open RecA interface which renders the enzyme in the open conformation. In this conformation the sensor serine is in contact with RNA. Additionally, there is one more nucleotide inside the RNA tunnel than in the closed conformation of the Prp43·RNA·ATP 5LTA. Since the 6I3P crystal structure has missing residues, we modeled the residues with modeller by A. Sali and T. Blundell [171] (more information in Methods section). We performed four 200 ns simulations of Prp22·RNA·ATP complex by inserting ATP inside the ATP binding pocket of 6I3P, four 200 ns simulations of the Prp22·RNA complex starting from the 6I3P crystal structure, four 300 ns simulations of the Prp22·ATP complex by removing RNA and placing ATP in the ATP binding pocket, four 1µs simulations of the Prp22·ADP complex by removing RNA and placing ADP in the ATP binding pocket, and four 300 ns simulations of the Prp22 apo structure by removing RNA from the crystal structure.

The Prp22·RNA complex shows a constant RecA1–RecA2 and CTD–RecA2 distance, which is shown in in Fig. 3.8A. The protein behaves completely stable without any conformational changes, which is in-line with the expectation since it is the original complex from the crystal structure.

The Prp22·RNA·ATP complex has an overall stable RecA interface with only small fluctuations. The CTD–RecA2 distance is not changing significantly during the 200 ns simulations time (Fig. 3.8B). Despite being not the native state of the crystal structure, the insertion of ATP inside the ATP pocket did not have an impact on the protein.

The Prp22·ATP complex shows in one of four simulations a decrease in the RecA distance. All simulations carry out an increase in the CTD–RecA2 distance (Fig. 3.8C). The change in the CTD–RecA distance indicates that the RNA tunnel becomes unstable after around 70 ns after removal of RNA. Both, the disruption of the RecA interface in one simulation and the CTD–RecA2 distance increase in four simulations indicate that the removal of RNA has an impact on the ATP pocket and the RNA tunnel. After removing RNA, the RecA2 domain becomes more flexible and can approach the RecA1 domain, which might enable the formation/closing of the RecA interface. However, in the relative short 300 ns simulation time, no such interface formation can be observed in the case of Prp22.

The Prp22·ADP complex behaves similar to the Prp22·ATP complex. After around 70 ns the CTD–RecA2 distance increases and after around 100 ns the RecA domain distance decreases significantly (Fig. 3.8D). Eventually, both distances converge to a similar value after 1 µs. Here, an average RecA distance of 3.0 nm and an average CTD–RecA2 distance of ∼0.7 nm is achieved. However, we do not observe a full closing of the RecA domains, probably because no sensor serine helix-to-loop transition is occurred, which seems to be essential for the closing process.

The Prp22 apo structure also shows a similar trend like the Prp22·ATP complex. The RecA domain distance is decreasing in two out of four simulations after 50 ns. The final conformations of these two simulations resemble a semi closed state without a helix-to-loop transition of the sensor serine. The CTD–RecA2 distance is increasing in two out of four simulations, which shows the instability of the RNA tunnel after removal of RNA (Fig. 3.8E).

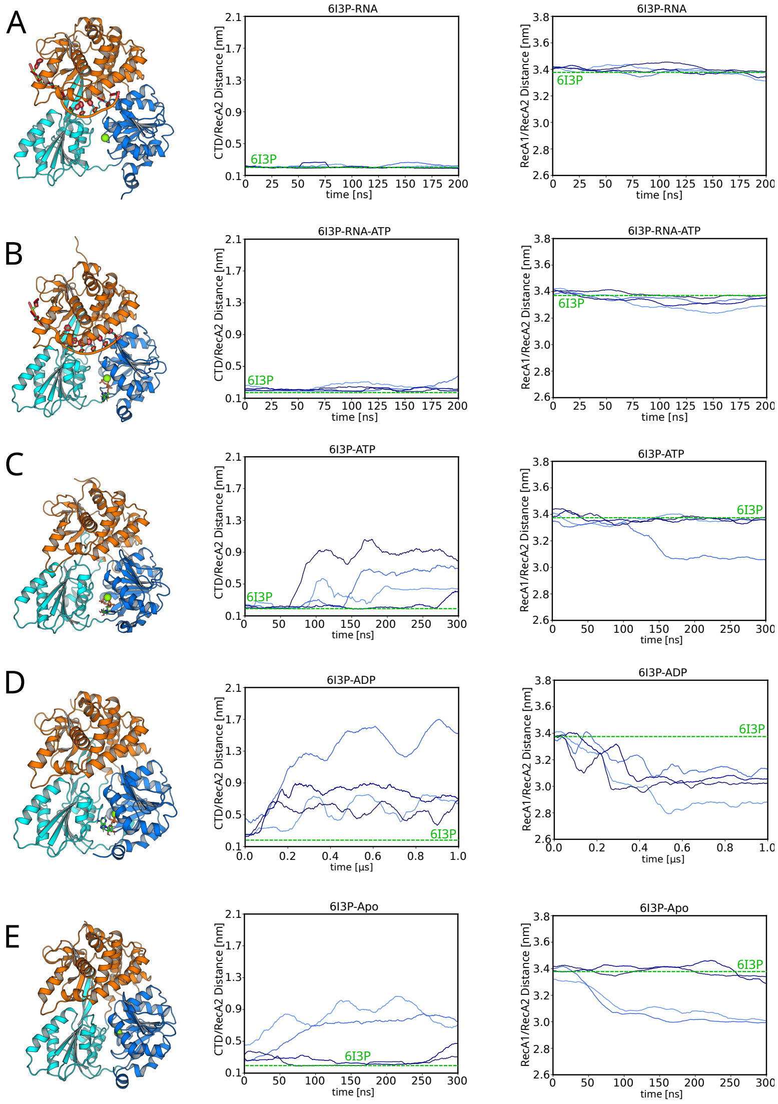


Figure 3.8: *The analysis of all 6I3P simulations. (A) The Prp22·RNA complex. The helicase is overall rigid and no large conformational change is observed. (B) The Prp22·RNA·ATP complex. The helicase is in its native state as in the crystal structure. The RecA distance and RNA-cleft opening are constant besides some small fluctuations during the 200 ns. (C) The Prp22·ATP complex. The distance between the CTD and RecA2 is decreasing while the RecA domains stay close to each other. (D) The Prp22·ADP complex. The distance between the CTD and RecA2 is decreasing and the RecA domains are more flexible, which results in a small change in the RecA distance. (E) The Prp22 apo structure. The distance between the CTD and the RecA2 is varying a lot. Also, the RecA distance is changing drastically in two simulations.*

3.3 Discussion

The simulations of the different complexes of various crystal structures revealed the role of the ligands on the overall protein dynamics. The native complexes from the crystal structures, except for 5LTK and 5LTJ, are stable in their RecA domain distance and in the CTD–RecA2 distance. Hence, we can validate that those crystal structures are not in a state induced by crystallographic artefacts for these complexes.

5LTK and 5LTJ take the form of a closed RecA interface and an open RNA cleft, which seems to be a pre-RNA-loading state. In all our simulations, the open RNA cleft always collapsed into a closed RNA cleft, in which an RNA loading is unlikely. Tauchert et al. [29] highly emphasized on the fact that these crystal structures are not crystallization artefacts, because two open structures are resolved under two different experimental conditions. Hence, the simulations may not be able to capture the state with an open RNA cleft, due to one or more of the following reasons: 1. the force field might be slightly inaccurate for some residues and therefore unable to simulate the state correctly, 2. the open state is connected to a closed state with a similar free energy and a small energy barrier between them or 3. the open state is not stable due to a missing buffer solution. The ladder issue is reasoned by the identification of the electrostatic surface of Prp43 by Tauchert et al. [170], which shows negative charges in the ratchet-like domain which encounters the positively charged part of the RecA2 domain. A buffer solution may shield the two domains enough to hinder a favored interaction between the domains.

Overall, RNA is a key player for the stability of the RNA tunnel and the RecA interface. The RNA strand is defining the RNA tunnel by its mere presence, because in our simulations no direct interactions are observed between RNA and the CTD domain. All interactions between RNA and protein are formed via the phosphate backbone of RNA and the RecA domains. Hence, also the RecA interface is influenced by RNA, because the H-bonds holding the RecA domains in place. On the other hand, ATP and ADP only seem to play a larger role for the RecA interface, because it bridges the RecA domains together via strong coulomb interactions between positively charged residues of the protein domains and negatively charged phosphate groups of ATP/ADP.

In conclusion, our extensive MD simulations revealed that RNA plays a significant role in determining the relative arrangements of CTD, RecA1, and RecA2 in Prp43 and Prp22, leading to reduced conformational fluctuations. While ATP and ADP primarily influence the stability of the RecA1-RecA2 interface, their impact on the larger-scale domain fluctuations is relatively minor.

However, we were not able to observe any conformational changes, which would render the resulting conformation in a known state resolved by experiments. Also, we did not detect a helix-to-loop or loop-to-helix transition of the sensor serine sequence. This may indicate that the simulation time was too short to sample rare conformational transitions. Hence, we must consider other methods, which may be able to increase sampling enough to make the observations of such rare events possible in a feasible amount of time.

RecA1–RecA2 interface

Since the conventional MD simulations lacked larger domain motions and critical conformational changes, e.g., the sensor serine flip, we had to try different approaches to enforce these kinds of motions. The first approach was the usage of non-equilibrium pulling. For non-equilibrium pulling a good pulling/reaction coordinate is needed to reveal desired conformational changes which resemble important and physical relevant transitions. Fortunately, a rough mechanistic hypothesis was already proposed by Ficner et al. [29,30,170,177] based on various crystal structures. First, we implemented the RecA1–RecA2 distance as a reaction coordinate, which is the most direct approach to initiate a RecA interface rupture. To check the stability of the resulting conformation, we continue the pulling simulations by starting new conventional simulations without a pull force from the last frames of the pulling simulations as new start states.

According to the revealed structures, the serine flip is essential for the opening process. In the second approach, we tried to enforce the opening of the RecA domains by a pulling on the sensor serine. The resulting reaction coordinate mimics a helix-to-loop transition by pulling the S318 residue towards the RNA U5. To investigate the stability of the resulting conformations, we continued the pulling simulations without a pull force afterwards.

4.1 Methods

Non-equilibrium pulling simulations were set up as follows with GROMACS 2019.1. The initial coordinates were taken from the equilibrated crystal structure 5LTA with removed ATP. Interactions of protein and RNA were described with the Amber14SB force field [132].

Water was modeled with the TIP3P model [174], and parameters for K^+ were taken from [175]. The energy of the system was minimized with the steepest descent algorithm. Then, the system was equilibrated for 100 ps with position restraints acting on the heavy atoms including RNA and Mg ($k = 1000\,\mathrm{kJ\,mol^{-1}nm^{-2}}$). Electrostatic interactions were described with the particle-mesh Ewald method [138]. Dispersion interactions and short-range repulsion were described together with a Lennard-Jones potential with a cut-off at 1 nm. The temperature was controlled at 300 K using velocity-scaling [135], thereby coupling protein, RNA, $\mathrm{Mg^{2+}}$, and ATP (if present) to one heat bath while coupling water and $\mathrm{K^+}$ to a second heat bath ($\tau = 0.5\,\mathrm{ps}$). The pressure was controlled at 1 bar with the Parrinello-Rahman barostat ($\tau = 5\,\mathrm{ps}$) [176]. The geometry of water molecules was constrained with SETTLE [134]. All other bonds were constrained with P-LINCS [133].

For the first four pulling simulations, the COM distance between RecA1 and RecA2 was chosen as reaction coordinate. A force constant of 10000 kJ/mol·nm and pull rate of 0.005 nm/ns was applied for 100 ns to increase the RecA1–RecA2 distance. After 100 ns, each simulation was continued without restraints for an additional 100 ns. For the second four pulling simulations, the COM distance between S387 and U5 O1P atoms was chosen as reaction coordinate. A force constant of 5000 kJ/mol·nm and a pull rate of -0.012 nm/ns was applied for 100 ns to decrease the distance between S387 and RNA-U5 and, thereby, drive the loop-to-helix transition of the sensor loop. After these two 100 ns simulations, each was continued by four independent 400 ns without restraints.

4.2 Results

Starting from the closed Prp43-RNA complex (PDB ID: 5LTA; removed ATP), a complete opening to 3.1 nm was achieved with the RecA1–RecA2 pull coordinate. The domains opened in a linear fashion and seemed to be in a reasonable agreement with the open structure known from crystal structure (PDB ID 6I3P). However, when continuing the simulations without the pulling force, i.e., in conventional free simulations, the yielded open structures fell back to an intermediate semi-open conformation, which shows an average RecA distance of around 2.8 nm to 2.9 nm. Thus, we assume a memory effect due to an incomplete opening. During closer investigation, we spotted a difference between the secondary structure of the sensor serine motif in the open states of the pulling simulations and the crystal structure of Prp22 which resembles the native open state. In the crystal structure of Prp22, the motif containing the sensor serine is in an alpha-helical state is in an alpha-helix state, which is not the case after our pulling simulations. Conclusively, a loop-to-helix transition must occur during the opening process to form a stable open conformation as suggested by Tauchert et al. [29].

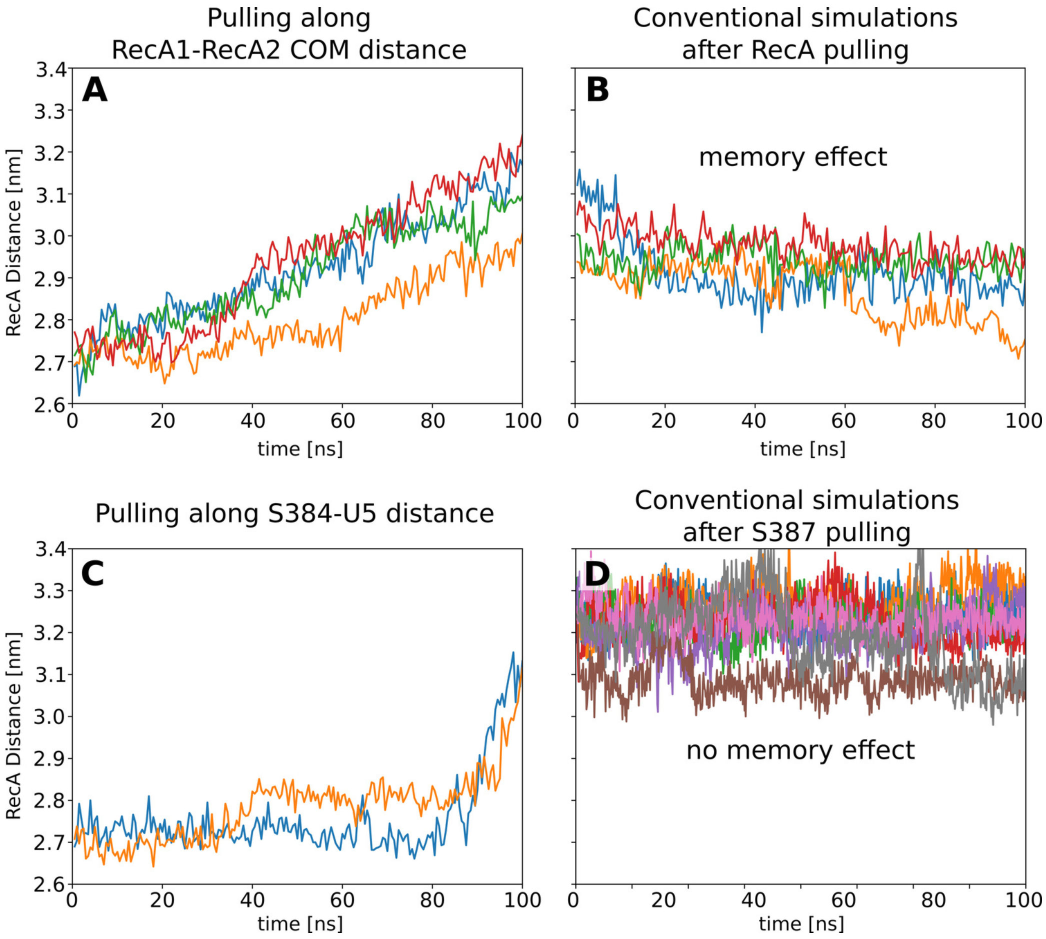


Figure 4.1: *On the correlation between Prp43 opening and loop-to-helix transition of the S387–G392 sensor loop.*
RecA1–RecA2 distance versus simulation time during (A/C) pulling simulations and (B/D) after releasing the pulling force. (A) Pulling along the RecA1–RecA2 center-of-mass (COM) distance leads to major memory effects, as shown by (B) the partial re-closure of the RecA1/RecA2 interface after release of the force. (C) Pulling along the S387–U5 distance, thereby driving the loop-to-helix transition of the sensor loop, leads to opening of the RecA1–RecA2 interface and (D) does not lead to memory effects. The absence of memory effects after the loop-to-helix transition suggests that the sensor loop transitions are critical for Prp43 opening and closing.

This hypothesis was tested by changing the pull force such that the sensor serine S387 is pulled towards its interaction partner, the oxygen atom of the phosphate group of the RNA backbone (O1P-U5). Here, the pulling enforced a loop-to-helix transition of the serine motif, which led to a spontaneous opening of the RecA domains suggested by a 3.1 nm RecA1–RecA2 end distance. Continuing the simulations from the resulted open structures revealed the formation of stable open conformations, which do not fall back to intermediate states. Hence, the memory effect has been eliminated. These results confirm the importance of the sensor serine flip during the opening process.

For the closing process, multiple pulling groups were tested to achieve a closing transition including a shift in the RNA contacts with the protein. The used pull coordinates on the Prp22 open crystal structure were the following: 1. COM of both RecA domains pulled towards each other, 2. the sensor serine pulled towards the magnesium ion and 3. The end of the RNA strand pulled away from the enzyme. None of the listed pulling codes yielded a successful closing or a successful shift of RNA contact by one nucleotide.

4.3 Discussion

The mechanism of the helicase's opening is complex and hence cannot be modelled with a simple one-dimensional reaction coordinate. Either a more sophisticated reaction coordinate is needed, or other enhanced sampling techniques might be more suitable for the project. Since we assumed that a lot of different orthogonal molecular switches control the conformational changes, we sticked with the latter idea, i.e., choosing other enhanced sampling methods to overcome the issues of poor sampling.

Simulated Tempering efficiency

DEAH helicases are complex proteins for which sampling is a fundamental problem as proven by the previous two chapters. Hence, we will use other enhanced sampling techniques which will not require the definition of a reaction coordinate. First, we try to make use of the so-called Simulated Tempering (ST) method. In ST, the temperature of the system is a dynamic variable which can change over time in the boundaries of a pre-defined temperature ladder. Transition from one temperature to another is performed with the Metropolis algorithm. ST is able to accelerate the conformational sampling of a system with a rough free energy landscape by one order of magnitude [109, 111]. Shaw et al. showed that Simulated tempering (ST) can reduce the computational time by one magnitude [111]. For more information see the Theory chapter.

5.1 Methods

We used the ST implementation of GROMACS using a minimal temperature of 300 K and a maximum temperature of 348 K. The temperature difference between neighboring states was set to 4 K resulting in 23 states. Attempts for temperature transitions were carried out every 500 integration steps and accepted or rejected with the Metropolis algorithm. The initial weights were calculated using a preliminary simulated annealing simulations with the routine described by Park et al. [140]. The weights of the states were updated every 500 steps throughout the simulations using the Wang-Landau algorithm [178]. A representative example for the convergence of the weights and for the transitions among temperature states over time is shown in Fig. 5.1.

Before using the ST method in the system of interest, we tested if ST reduces the simulation time enough to capture the desired transitions. Thus, we performed simulations of a spontaneous opening by removing ADP from the crystal structure Prp43·ADP (PDB ID:5D0U), since here a serine flip

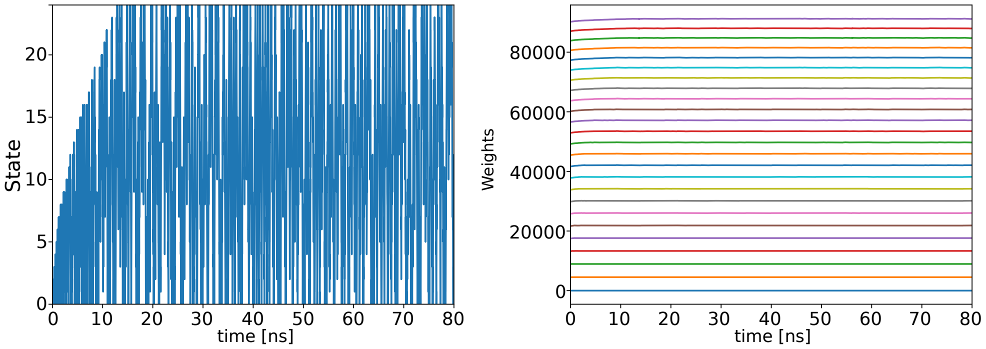


Figure 5.1: *Coverage of temperature states during simulated tempering. Left: temperature state versus simulation time during an example simulated tempering (ST) simulation. During the burn-in phase within the first 17 ns, states with increasingly higher temperatures were visited, reflecting the gradual adaptation of the weights. Right: weights of temperature states versus simulation time. Only the higher weights were slightly adapted at the beginning of the simulation, reflecting decent initial weights. After 17 ns, the weights were converged, and all states are frequently visited.*

already occurred. Ten conventional simulations and ten simulated tempering simulations, each 300 ns run time, were carried out and analyzed. The simulations were performed with GROMACS 2020.2. The initial coordinates for these simulations were taken from PDB data bank (PDB ID 5D0U [29]), representing the Prp43/ADP complex. A non-RNA-loaded structure was chosen to ensure a higher flexibility of the RecA domains. ADP was removed from the system to trigger the opening process, and the system was solvated with 45800 water molecules and neutralized with 1 potassium ions. All other parameters were chosen as described in the previous chapter. Then, the progression of the RecA domain distances were analyzed after 300 ns. As shown in Fig. 5.2, the partial opening of the RecA1/RecA2 interface is greatly accelerated in ST simulations as compared to conventional simulations.

5.2 Results

Among 10 conventional simulations, only two simulations reached a semi-open state within 300 ns, as indicated by the RecA1/RecA2 distance of over 3.0 nm. In contrast, among 10 ST simulations, four reached the semi-open state within only 100 ns and six reached the semi-open state within 300 ns. In addition, four ST reached a fully open state indicated by a distance larger

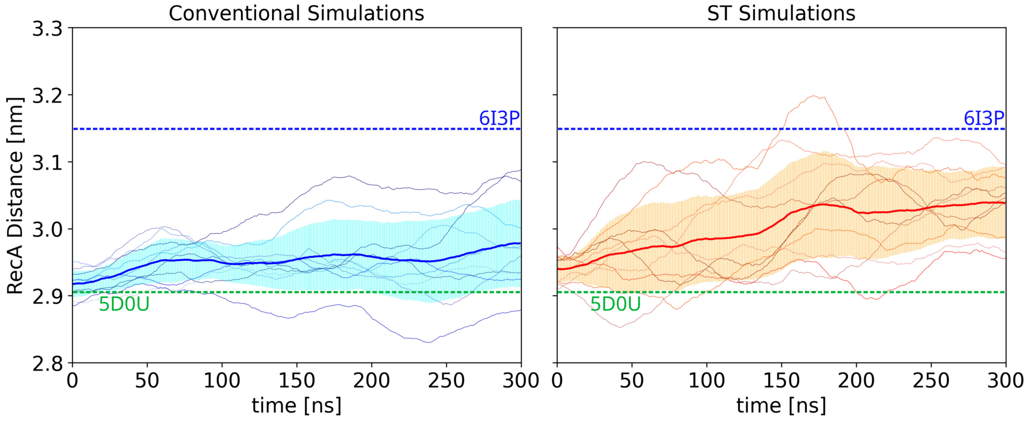


Figure 5.2: *Accelerated conformational sampling with simulated tempering (ST). RecA1–RecA2 distance of Prp43 after removal of ADP in conventional simulations (left) or ST simulations (right).*

3.1 nm. Additionally, as shown in Fig. 5.2, the ST simulations sampled a much broader RecA1–RecA2 distance range, indicated by larger standard deviations than in the conventional simulations. The more rapid increase of the average (thick lines) and of standard deviations (shaded areas) in ST compared to conventional simulations indicate accelerated sampling of the conformational space during a shorter simulation time.

The broad sampling and large-scale conformational fluctuations of the RecA1 and RecA2 domains could be an indicator for the existence of a flexible conformation, which is on average in a semi-open state. The high flexibility may be a reason for the difficulties in the crystallization of a non-ATP/ADP bound state of the proteins. This behavior was suggested by the apo crystal structure of Prp22, because of a low resolution in the RecA2 domain [30].

5.3 Discussion

The ST method yields promising results during the sampling of the DEAH helicase. However, the time of large domain dynamics and even more importantly, ligand translocation, ranges from hundreds of microseconds to milliseconds. Thus, even a one order of magnitude acceleration might still not be sufficient to make the motions of interest visible in the ligand-bound helicases by MD simulations. Therefore, we introduce the combination of ST with the Adaptive Sampling approach in the next chapter.

The RNA translocation cycle of Prp43

As we demonstrated, a full opening or closing of the DEAH-box helicases is not achieved by simple conventional simulations. Additionally, the system's behavior has proven to be too complex to be suited for simulations with non-equilibrium techniques involving pre-defined reaction coordinates. Although, we have shown that pulling enforces an open conformation and the rare transition of the sensor serine loop-to-helix transition, the observation of a complete translocation cycle of RNA is still challenging. The translocation is too complex, because it involves not only a single transition, but rather a series of transitions in a specific order. The ST approach, on the other hand, showed promising results to enhance the sampling of the system by implementing the temperature as a dynamic variable.

Since protein dynamics occur on the timescales of microseconds to milliseconds, such as helix formation and ligand interactions [179], we would need several months to a few years to simulate a full conformational domain transition of a DEAH-box helicase with a single simulation on conventional hardware. Although, ST can enhance the sampling by one magnitude, it might not be sufficient for the observation of such dynamics in a feasible amount of time. Therefore, we combined ST with the Adaptive Sampling (AS) method to further enhance the sampling of the system.

Indeed, the combination approach yielded a complete translocation cycle of RNA by one nucleotide of the DEAH-box helicase Prp43 via MD simulations. In the following sections, we present the general outcome and the analysis of the procedure in detail. This section and the included subsections are based on our own publication [180].

6.1 Methods

6.1.1 Simulation setup of the opening process

Molecular dynamics (MD) simulations of the opening process of the RNA translocation cycle were set up as follows with GROMACS 2019.5 [169]. The initial structure of Prp43 from *C. thermophilum* was taken from the protein data bank (PDB ID 5LTA [29]), representing the complex of Prp43 with U_7-RNA and with the ATP analogue ADP-BeF_3. The ATP analogue and other inorganic molecules were removed from the structure, thereby modeling Prp43•U_7 after the dissociation of the hydrolyzed ATP. The structure was placed into a simulation box of a dodecahedron. The box was solvated with 35350 water molecules and neutralized with 9 potassium ions. Interactions of protein and RNA were described with the Amber14SB force field [132]. Water was modeled with the TIP3P model [174], and parameters for K^+ were taken from [175]. The energy of the system was minimized with the steepest descent algorithm. Then, the system was equilibrated for 100 ps with position restraints acting on the heavy atoms including RNA and Mg ($k = 1000\,\mathrm{kJ\,mol^{-1}nm^{-2}}$).

The following parameters were used in all simulations. Electrostatic interactions were described with the particle-mesh Ewald method [138]. Dispersion interactions and short-range repulsion were described together with a Lennard-Jones potential with a cut-off at 1 nm. The temperature was controlled at 300 K using velocity-scaling [135], thereby coupling protein, RNA, Mg^{2+}, and ATP (if present) to one heat bath while coupling water and K^+ to a second heat bath ($\tau = 0.5\,\mathrm{ps}$). The pressure was controlled at 1 bar with the Parrinello-Rahman barostat ($\tau = 5\,\mathrm{ps}$) [176]. The md-vv integrator was used for simulated tempering (ST) simulations and the md integrator was used for all other simulations, both with an integration time step of 2 fs. The geometry of water molecules was constrained with SETTLE [134]. All other bonds were constrained with P-LINCS [133]. To accelerate the conformational sampling of the Prp43 cycle, we used adaptive sampling (AS) in combination with ST, as described in the following.

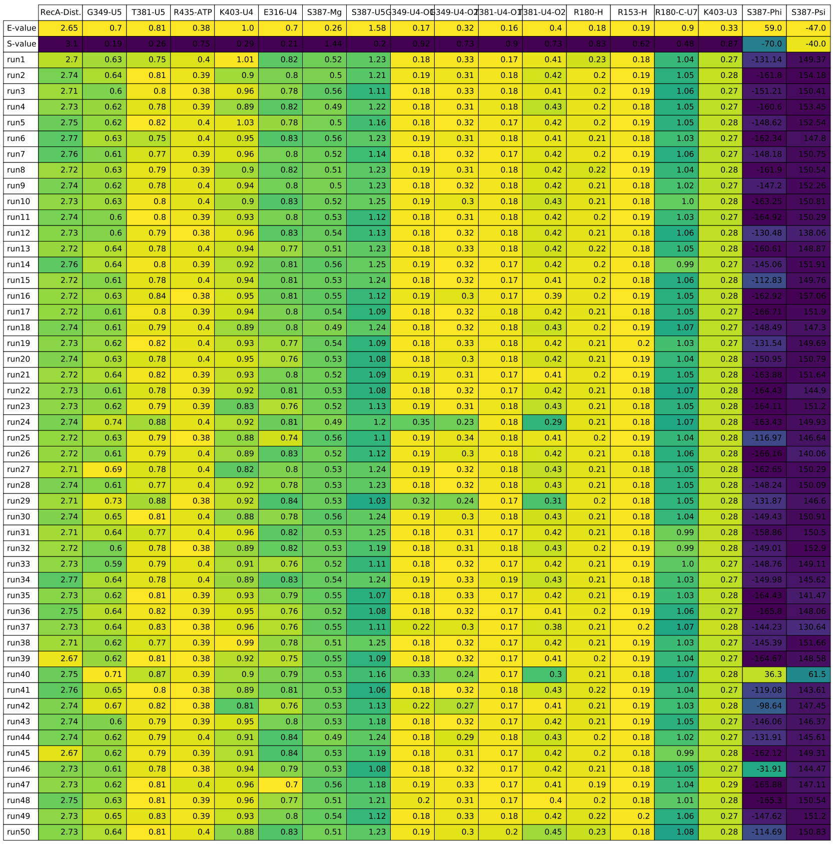
6.1.2 Simulation setup of the closing process

MD simulations of the closing process of the RNA translocation cycle were set up as follow with GROMACS 2020.2. The initial coordinates of Prp43 were taken from the last successful AS simulation of the opening process, representing the protein-RNA complex in the open configuration (Fig. 6.4D, colored representation). The ATP–Mg^{2+}–water complex was inserted by first superimposing the RecA1 domain of the open complex onto the RecA1 domain of the crystal structure using a root mean-square deviation (RMSD) fit (PDB ID 5LTA [29]). Then, the ATP, Mg^{2+} and water were positioned to match at the position of the ATP analogue and of the crystal water in the superimposed 5LTA structure. Parameters of the ATP were taken from Carlson et al. [172], translated into GROMACS format with the ACPYPE software [173]. All other parameters and equilibration steps were identical to the opening process described above.

6.1.3 Adaptive sampling protocol

The Prp43 opening process required 9 rounds of AS, whereas the closing process required 11 rounds to complete the cycle of RNA translocation by one nucleotide. In each round, between 20 and 500 parallel simulations were carried out, which were started from the final conformation of the most successful simulation of the previous round. The parallel simulations were seeded with new random velocities taken from a Maxwell-Boltzmann distribution, thereby obtaining independent trajectories. The individual simulations were carried out between 10 and 100 ns (Table 2). Thereby, the accumulated simulation time was 56.5 µs for the opening process and 40.0 µs for the closing process.

A simulation of an AS round was taken as most successful, if the final frame exhibited the highest similarity to a set of selected features of the features of the reference state (target values). The selected features are listed in Table 1. In this study, the analysis of each round was carried out by plotting the structural features in a heat table as shown in Fig. 6.1. Here, the first and second line in the heat table list the feature values of the starting and the reference structure, respectively. Other lines show the feature values of the independent simulations. The colors indicate the similarity of a feature either to the starting structure (purple) or to the reference structure (yellow). Since it was difficult to weight different features in an automated manner, we selected the most successful simulation by human supervision. This way, from AS round to round, the simulations gradually approached the conformation of the reference state.



	RecA-Dist.	G349-U5	T381-U5	R435-ATP	K403-U4	E316-U4	S387-Mg	S387-U5	G349-U4-O1	G349-U4-O2	T381-U4-O1	T381-U4-O2	R180-H	R153-H	R180-C-U7	K403-U3	S387-Phi	S387-Psi
E-value	2.65	0.7	0.81	0.38	1.0	0.7	0.26	1.58	0.17	0.32	0.16	0.4	0.18	0.19	0.9	0.33	59.0	-47.0
S-value	3.1	0.19	0.26	0.75	0.29	0.21	1.44	0.2	0.92	0.73	0.9	0.73	0.83	0.62	0.48	0.87	-70.0	-40.0
run1	2.7	0.63	0.75	0.4	1.01	0.82	0.52	1.23	0.18	0.33	0.17	0.41	0.23	0.18	1.04	0.27	-131.14	149.37
run2	2.74	0.64	0.81	0.39	0.9	0.8	0.5	1.21	0.19	0.31	0.18	0.42	0.2	0.19	1.05	0.28	-161.8	154.18
run3	2.71	0.6	0.8	0.38	0.96	0.78	0.56	1.11	0.18	0.33	0.18	0.41	0.2	0.19	1.06	0.27	-151.21	150.41
run4	2.73	0.62	0.78	0.39	0.89	0.82	0.49	1.22	0.18	0.31	0.18	0.43	0.2	0.18	1.05	0.27	-160.6	153.45
run5	2.75	0.62	0.82	0.4	1.03	0.78	0.5	1.16	0.18	0.32	0.17	0.42	0.2	0.19	1.05	0.28	-148.62	152.54
run6	2.77	0.63	0.75	0.4	0.95	0.83	0.56	1.23	0.19	0.31	0.18	0.41	0.21	0.18	1.03	0.27	-162.34	147.8
run7	2.76	0.61	0.77	0.39	0.96	0.8	0.52	1.14	0.18	0.32	0.17	0.42	0.2	0.19	1.06	0.27	-148.18	150.75
run8	2.72	0.63	0.79	0.39	0.9	0.82	0.51	1.23	0.19	0.31	0.18	0.42	0.22	0.19	1.04	0.28	-161.9	150.54
run9	2.74	0.62	0.78	0.4	0.94	0.8	0.5	1.23	0.18	0.32	0.18	0.42	0.21	0.18	1.02	0.27	-147.2	152.26
run10	2.73	0.63	0.8	0.4	0.9	0.83	0.52	1.25	0.19	0.3	0.18	0.43	0.21	0.18	1.0	0.28	-163.25	150.81
run11	2.74	0.6	0.8	0.39	0.93	0.8	0.53	1.12	0.18	0.31	0.18	0.42	0.2	0.19	1.03	0.27	-164.92	150.29
run12	2.73	0.6	0.79	0.38	0.96	0.83	0.54	1.13	0.18	0.32	0.18	0.42	0.21	0.18	1.06	0.28	-130.48	138.06
run13	2.72	0.64	0.78	0.4	0.94	0.77	0.51	1.23	0.18	0.33	0.18	0.42	0.22	0.18	1.05	0.28	-160.61	148.87
run14	2.76	0.64	0.8	0.39	0.92	0.81	0.56	1.25	0.19	0.32	0.17	0.42	0.2	0.18	0.99	0.27	-145.06	151.91
run15	2.72	0.61	0.78	0.4	0.94	0.81	0.53	1.24	0.18	0.32	0.17	0.41	0.2	0.18	1.06	0.28	-112.83	149.76
run16	2.72	0.63	0.84	0.38	0.95	0.81	0.55	1.12	0.19	0.3	0.17	0.39	0.2	0.19	1.05	0.28	-162.92	157.06
run17	2.72	0.61	0.8	0.39	0.94	0.8	0.54	1.09	0.18	0.32	0.18	0.42	0.21	0.18	1.05	0.27	-166.71	151.9
run18	2.74	0.61	0.79	0.4	0.89	0.8	0.49	1.24	0.18	0.32	0.18	0.43	0.2	0.19	1.07	0.27	-148.49	147.3
run19	2.73	0.62	0.82	0.4	0.93	0.77	0.54	1.09	0.18	0.33	0.18	0.42	0.21	0.2	1.03	0.27	-131.54	149.69
run20	2.74	0.63	0.78	0.4	0.95	0.76	0.53	1.08	0.18	0.3	0.18	0.42	0.21	0.19	1.04	0.28	-150.95	150.79
run21	2.72	0.64	0.82	0.39	0.93	0.8	0.52	1.09	0.19	0.31	0.17	0.41	0.2	0.19	1.05	0.28	-163.88	151.64
run22	2.73	0.61	0.78	0.39	0.92	0.81	0.53	1.08	0.18	0.32	0.17	0.42	0.21	0.18	1.07	0.28	-164.43	144.9
run23	2.73	0.62	0.79	0.39	0.83	0.76	0.52	1.13	0.19	0.31	0.18	0.43	0.21	0.18	1.05	0.28	-164.11	151.2
run24	2.74	0.74	0.88	0.4	0.92	0.81	0.49	1.2	0.35	0.23	0.18	0.29	0.21	0.18	1.07	0.28	-163.43	149.93
run25	2.72	0.63	0.79	0.38	0.88	0.74	0.56	1.1	0.19	0.34	0.18	0.41	0.2	0.19	1.04	0.28	-116.97	146.64
run26	2.72	0.61	0.79	0.4	0.89	0.83	0.52	1.12	0.19	0.3	0.18	0.42	0.21	0.18	1.06	0.28	-166.16	140.06
run27	2.71	0.69	0.78	0.4	0.82	0.8	0.53	1.24	0.19	0.32	0.18	0.43	0.21	0.18	1.05	0.28	-162.65	150.29
run28	2.74	0.61	0.77	0.4	0.92	0.78	0.53	1.23	0.18	0.32	0.18	0.43	0.21	0.18	1.05	0.28	-148.24	150.09
run29	2.71	0.73	0.88	0.38	0.92	0.84	0.53	1.03	0.32	0.24	0.17	0.31	0.2	0.18	1.05	0.28	-131.87	146.6
run30	2.74	0.65	0.81	0.4	0.88	0.78	0.56	1.24	0.19	0.3	0.18	0.43	0.21	0.18	1.04	0.28	-149.43	150.91
run31	2.71	0.64	0.77	0.4	0.96	0.82	0.53	1.25	0.18	0.31	0.17	0.42	0.21	0.18	0.99	0.28	-158.86	150.5
run32	2.72	0.6	0.78	0.38	0.89	0.82	0.53	1.19	0.18	0.31	0.18	0.43	0.2	0.19	0.99	0.28	-149.01	152.9
run33	2.73	0.59	0.79	0.4	0.91	0.76	0.52	1.11	0.18	0.32	0.17	0.42	0.21	0.19	1.0	0.27	-148.76	149.11
run34	2.77	0.64	0.78	0.4	0.89	0.83	0.54	1.24	0.19	0.33	0.19	0.43	0.21	0.18	1.03	0.27	-149.98	145.62
run35	2.73	0.62	0.81	0.39	0.93	0.79	0.55	1.07	0.18	0.33	0.17	0.42	0.21	0.19	1.03	0.28	-164.43	141.47
run36	2.75	0.64	0.82	0.39	0.95	0.76	0.52	1.08	0.18	0.32	0.17	0.41	0.2	0.19	1.06	0.27	-165.8	148.06
run37	2.73	0.64	0.83	0.38	0.96	0.76	0.55	1.11	0.22	0.3	0.17	0.38	0.21	0.2	1.07	0.28	-144.23	130.64
run38	2.71	0.62	0.77	0.39	0.99	0.78	0.51	1.25	0.18	0.32	0.17	0.42	0.21	0.19	1.03	0.27	-145.39	151.66
run39	2.67	0.62	0.81	0.38	0.92	0.75	0.55	1.09	0.18	0.32	0.17	0.41	0.2	0.19	1.04	0.27	-164.67	148.58
run40	2.75	0.71	0.87	0.39	0.9	0.79	0.53	1.16	0.33	0.24	0.17	0.3	0.21	0.18	1.07	0.28	36.3	61.5
run41	2.76	0.65	0.8	0.38	0.89	0.81	0.53	1.06	0.18	0.32	0.18	0.43	0.22	0.19	1.04	0.27	-119.08	143.61
run42	2.74	0.67	0.82	0.38	0.81	0.76	0.53	1.13	0.22	0.27	0.17	0.41	0.21	0.19	1.03	0.28	-98.64	147.45
run43	2.74	0.6	0.79	0.39	0.95	0.8	0.53	1.18	0.18	0.32	0.17	0.42	0.21	0.19	1.05	0.27	-146.06	146.37
run44	2.74	0.62	0.79	0.4	0.91	0.84	0.49	1.24	0.18	0.29	0.18	0.43	0.2	0.18	1.02	0.27	-131.91	145.61
run45	2.67	0.62	0.79	0.39	0.91	0.84	0.53	1.19	0.18	0.31	0.17	0.42	0.2	0.18	0.99	0.28	-162.12	149.31
run46	2.73	0.61	0.78	0.38	0.94	0.79	0.53	1.08	0.18	0.32	0.17	0.42	0.21	0.18	1.05	0.27	-31.91	144.47
run47	2.73	0.62	0.81	0.4	0.96	0.7	0.56	1.18	0.19	0.33	0.17	0.41	0.19	0.19	1.04	0.29	-165.88	147.11
run48	2.75	0.63	0.81	0.39	0.96	0.77	0.51	1.21	0.2	0.31	0.17	0.4	0.2	0.18	1.01	0.28	-165.3	150.54
run49	2.73	0.65	0.83	0.39	0.93	0.8	0.54	1.12	0.18	0.33	0.18	0.42	0.22	0.2	1.06	0.27	-147.62	151.2
run50	2.73	0.64	0.81	0.4	0.88	0.83	0.51	1.23	0.19	0.3	0.2	0.45	0.23	0.18	1.08	0.28	-114.69	150.83

Figure 6.1: *Example heat table from the closing process, visualizing the progression of one round of AS simulations towards the target structure.*
Rows correspond to 50 individual simulations of one round of AS. Top row: list of structural features including distances, angles, and ϕ/ψ angles. Second row: reference (target) values of the features, here taken from the 6I3P structure of Prp43•U_7•ATP. Third row: starting values of the features, taken from the open simulation frame. Columns show the feature values at the end of this AS round. The color indicates the similarity with the starting feature (purple) or with the reference/target feature (yellow). Such tables have been used extensively to monitor the progression of the AS simulations and to select the most successful simulation to be used a seed for the next round of AS.

Table 6.1: Key properties of the adaptive sampling rounds.
For each round of adaptive sampling, sum of simulation times t_{sim}, number of simulations N_{sim}, number of observed transitions N_{trans}, estimated transition rate k for the opening (left) and closing process (right).

	Opening				Closing			
AS round	t_{sim} (µs)	N_{sim}	N_{trans}	k (1/µs)	t_{sim} (µs)	$\mathrm{N}_{\mathrm{sim}}$	N_{trans}	k (1/µs)
1	10.0	100	1	0.10	0.8	10	3	3.69
2	10.0	100	1	0.10	0.7	10	0	–
3	10.0	100	9	0.90	0.7	10	0	–
4	10.0	100	0	–	2.0	30	1	0.29
5	5.0	500	2	0.13	3.0	30	0	–
6	8.3	100	20	2.41	2.0	30	2	0.40
7	0.4	40	0	–	2.0	20	2	1.00
8	0.3	40	0	–	4.0	40	0	–
9	1.6	40	1	0.43	5.0	50	5	0.56
10	–	–	–	–	5.0	50	0	–
11	–	–	–	–	4.0	40	0	–
12	–	–	–	–	3.5	40	0	–
13	–	–	–	–	2.7	40	1	0.06

6.1.4 Selected features for adaptive sampling

The progression of the conformational transitions was monitored using the following structural features, where the atom names follow the PDB names: Distances between 1. the center of mass (COM) of RecA1 and RecA2, distances between the following pairs of atoms 2. G349-H and U5-O1P, 3. T381-OG1 and U5-O1P, 4. N382-OD1 and U4-2HO, 5. K403-NZ and U4-O1P, 6. E316-H and U4-O2P, 7. S387-HG and Mg, 8. S387-HG and U5-O1P, 9. G349-H and U4-O1P, 10. G349-H and U4-O2P, 11. T381-HG1 and U4-O1P, 12. T381-HG1 and U4-O2P, 13. N382-OD1 and U3-2HO, 14. K403-NZ and U3-O1P, 15./16. the ψ and ϕ angles of S387. More details are shown in Table 1.

The helicase Prp22 is homologous to Prp43 studied here, as demonstrated by sequence identity and similarity of 47% and 63%, respectively, according to a FASTA sequence alignment [181]. Furthermore, residues of the selected features are conserved among Prp22 and Prp43, except for G349 in Prp43 that is replaced with serine in Prp22. However, since G349 interacts with the RNA mostly via the protein backbone, this replacement has only a minor effect on the characterization of the opening transition. Hence, target values for the selected features were taken from the open Prp22 structure (PDB ID 6I3P [30]).

6.1.5 Estimation of mean first passage times (MFPTs)

In this study, we focused on achieving successful opening and closing transitions of Prp43, rather than exhaustively sampling transitions between all long-living intermediate states. Consequently, we estimated only the order of magnitude of the MFPTs but do not aim towards a comprehensive kinetic network of Prp43 dynamics. Since we applied ST during AS simulation runs, we assume that the transition rates have been accelerated by one order of magnitude [111].

The MFPTs were estimated based on the formalism by Pande and Singhal [182], which involves the calculation of the transition rate matrix between long-living intermediate states. We constructed the transition matrix using the following assumptions: (i) Based on the AS simulations, we modeled both the opening and the closing process as seven transitions between $N = 7$ states. The forward rates $k_{n,n+1}$ were taken from the number of successful forward transitions per total simulation time of the AS round (Table 6.1). (ii) Since the transitions of molecular switches were found to be highly interdependent, we assumed that the opening and closing transitions occurs predominantly via the linear sequence of transitions described in the Results section, suggesting that the rate matrix is tridiagonal ($k_{n,m} = 0$ if $|n-m| > 1$). (iii) Because both the opening transition (in absence of ATP) and the closing transition (in presence of ATP) occur down the free energy landscape ($\Delta G < 0$), we assume that the backward rates $k_{n+1,n}$ are smaller than the respective forward rates $k_{n,n+1}$. To reveal the range of possible MFPTs, we considered the limiting cases where (a) the backward rates equal the forward rates ($k_{n,n+1} = k_{n+1,n}$), corresponding to $\Delta G = 0$, or (b) the backward rates equal 0.01 times the forward rates ($k_{n,n+1} = 100 k_{n+1,n}$), corresponding to a marked downhill process with $\Delta G \ll 0$. These assumptions lead to the following matrix equation:

$$\begin{bmatrix} p_{11}-1 & p_{12} & 0 & \dots & \dots & 0 \\ p_{21} & p_{22}-1 & p_{23} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \dots & 0 \\ \vdots & & & & & \vdots \\ 0 & \dots & 0 & p_{65} & p_{66}-1 & p_{67} \\ 0 & \dots & \dots & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ \vdots \\ x_8 \end{bmatrix} = \begin{bmatrix} -\Delta t \\ -\Delta t \\ \vdots \\ \vdots \\ -\Delta t \\ 0 \end{bmatrix} \tag{6.1}$$

Here, $p_{n,m}$ denotes the probability of transitioning from state n to state m ($n, m = 1, \dots, 7$) within the time delay $\Delta t = 1\,\text{ns}$. Using other time delays between 0.1 ns and 20 ns led to nearly identical estimates for the MFTPs. The symbols x_i denote the MFPTs from state i to the final state $n = 7$. The probabilities obey $p_{n,n} + p_{n,n-1} + p_{n,n+1} = 1$. For the two limiting cases described above, we have $p_{n,n-1} = c \cdot p_{n,n+1}$ with $c = 1$ or $c = 0.01$, respectively.

The simulation data used to compute the transition rates are summarized in Table 6.1 for the opening and the closing process. The table lists, for each round of AS, the overall simulation time of the round, the number of parallel simulations of the round, the number of successful forward transitions taken from all parallel simulations, and the computed forward rate $p_{n,n+1}$. According to Table 6.1 the forward rates for the opening are $k_{n,n+1} = (0.10, 0.10, 0.90, 0.13, 2.41, 0.45)\,\mu s^{-1}$ and the forward rates for the closing are $k_{n,n+1} = (3.69, 0.29, 0.40, 1.00, 0.56, 0.06)\,\mu s^{-1}$. The transitions were identified via the progression of structural features and validated by extensive visual inspection of the simulations. In case that no successful forward transition occurred within an initial set of parallel simulations, a new set was simulated, thereby adding to the total simulation time of the AS round. The probabilities in Eq. 6.1 can be estimated for small rates via

$$p_{n,n+1} = \frac{N^{\text{trans}}_{n,n+1}}{T_n}\,\Delta t = k_{n,n+1}\,\Delta t, \tag{6.2}$$

where $N^{\text{trans}}_{n,n+1}$ is the number of forward transitions in AS round n, and T_n is the overall simulation time of AS round n.

Upper/lower bounds and order of magnitude estimates of MFPTs

For the two limiting cases of (i) a nearly flat free energy landscape ($k_{n,n+1} = k_{n+1,n}$) or (ii) for a marked downhill process with $\Delta G \ll 0$ we obtained the following MFPTs. For case (i), we obtained for the opening $k_{\text{opening}} = 141\,\mu\text{s}$ and closing process $k_{\text{closing}} = 52\,\mu\text{s}$, representing upper bounds of the opening and closing rates. For the marked downhill process, we obtain $k_{\text{opening}} = 32\,\mu\text{s}$ and $k_{\text{closing}} = 28\,\mu\text{s}$. As a more realistic intermediate case of $k_{n,n+1} = 2.5k_{n+1,n}$, which translate into $\Delta G \approx -18\,\text{kJ/mol}$, we obtain $k_{\text{opening}} = 50\,\mu\text{s}$ and $k_{\text{closing}} = 31\,\mu\text{s}$. These values imply orders of magnitudes for the opening and closing rates of 100 µs and 50 µs (see Results).

Increased computational efficiency due to adaptive sampling

The cumulative invested simulation time from all AS runs for the opening and closing simulations were 56 µs and 40 µs, which are both in the order of magnitude of the MFPTs. Hence, main benefit of AS was not to largely reduce the cumulative computational cost for obtaining the cycle. Instead, these values imply that the benefit of AS was the ability to trivially parallelize and to monitor the progression of the simulations towards the target state. As a numerical example, for the Prp43 system, we obtained a simulation performance of 70 ns/day on a compute node with a six-core CPU and a Nvidia RTX 2080Ti GPU. Hence, simulating 150 µs for the conformational cycle (sum of MFPTs) on a single node would have required ∼6 years. Without ST, this value would increase by another order of magnitude. By combing AS with ST, these unacceptable wall clock times were dramatically reduced using only commodity hardware, enabling the study of complex enzymatic cycles as described here.

6.1.6 Construction of the MSM

The construction of the MSM and its analysis were performed with PyEMMA 2.5.7 [183] on all trajectories of the open and closing simulations obtained by the AS procedure. A sub-group of the features from the AS was chosen for the first step of the MSM construction, called the "featurization". We have chosen the following features: RecA1–RecA2 distance, T381–U5 distance, T381–U4 distance, G349–U5 distance, G349–U4 distance, K403–U4 distance, K403–U3 distance, E316–U4 distance, S387–U5 distance, S387–I383 distance, G349–U4 distance, and RMSD towards starting structure. The featurization coarse-grains the feature space of the protein dynamics and, thereby, simplifies the further processing by using only a reduced dimensionality, while keeping key dynamics of the feature space. Next, the time-lagged independent component analysis was performed to further reduce the dimensionality of the system [156, 167, 184], thus breaking the system down to the eight slowest collective motions as a linear combination of all other features. The lag-time used for our tICA decomposition was 5 ns. The resulting first two independent components were chosen for the projection of the kinetically based model obtained by the MSM, because those two components describe the systems dynamics in a simplified fashion (Fig. 6.2A).

The conformational microstates were generated from the tiCA components by the K-means clustering algorithm [185, 186]. We have chosen the value of 100 for the number of cluster centers K. For an optimal number of cluster centers, we calculated the VAMP-2 score as a function of K and checked for convergence as shown in Fig. 6.2B. Another important parameter for an MSM is the lag-time τ of the Markov Model. Here, the resulting implied timescales (ITS) should converge as a function of lag-time to ensure Markovianity, as found for lag times larger than 20 ns (Fig. 6.2C). Once the appropriate parameters were chosen, we computed the free energy landscape by re-weighting the trajectory frames with stationary probabilities from the MSM and projected the resulting free energies (Fig. 6.6) on the first two tiCA components as suggested by the PyEMMA workflow. The PCCA+ algorithm was used to assign each microstate to a corresponding macrostate. In this study, we have chosen to describe the system with five macrostates. The corresponding Chapman-Kolmogorov test for validation is shown in Fig. 6.3. The MFPTs between the macrostates were obtained using the PyEMMA functionality.

6.1.7 Simulations and PCA of isolated RecA1 and RecA2 Domains

MD simulations of $1\,\mu$s for the two isolated RecA-like domains were set up as follows with GROMACS 2019.5. The initial coordinates of RecA1 (residues 97–273) and RecA2 (residues 274–458) were taken from the 5LTA structure [29]. Each domain was placed into a dodecahedral simulation box. The RecA1 box and the RecA2 box were filled with 6308 or 16592 waters molecules and neutralized with one chloride or three sodium ions, respectively. All other parameters and equilibration steps were chosen as described above. PCA was performed to reveal large-scale motions in the isolated

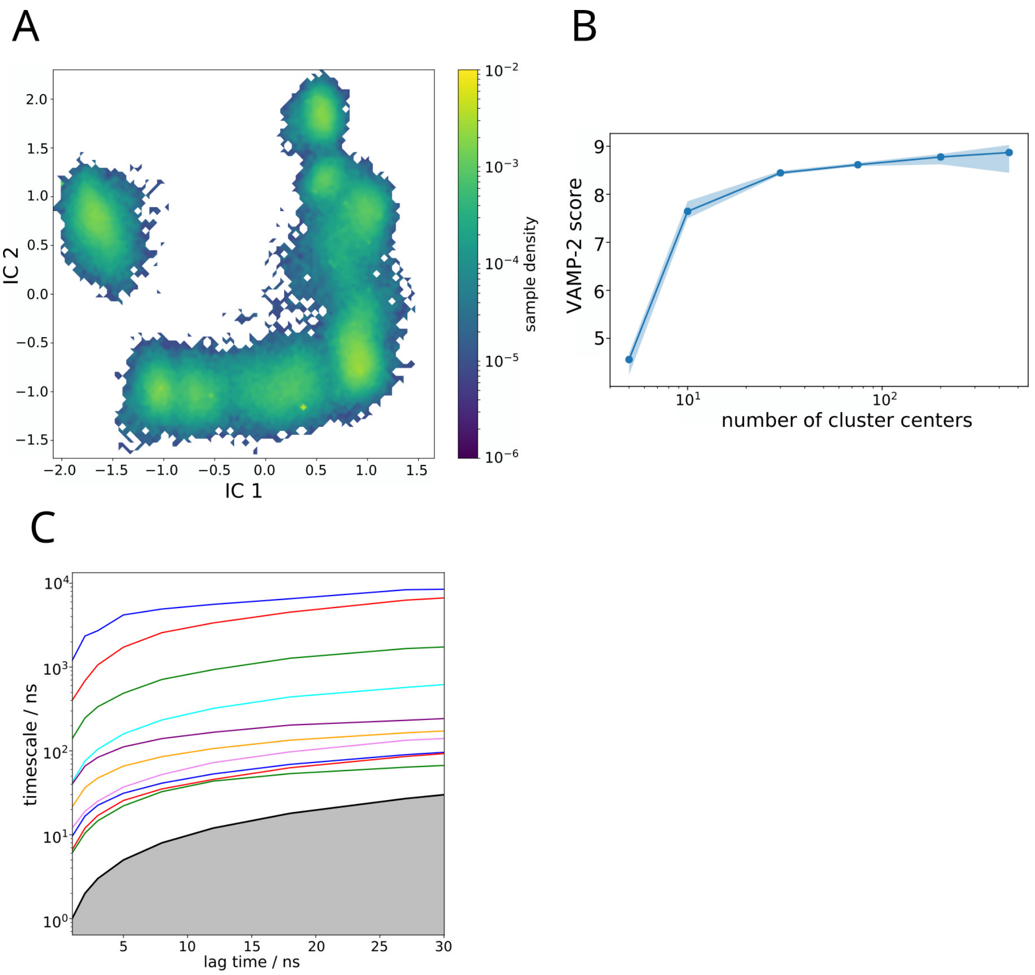


Figure 6.2: *Further analysis and validation of MSM. (A) tiCA plot of the first two independent components (ICs). (B) VAMP2 score versus the number of cluster centers for the microstate assignment. The VAMP2 score converges for more than 80 cluster centers. (C) Implied timescales versus lag time. The time scales converge for lag times greater than approximately 20 ns.*

RecA-like domains. The GROMACS module *gmx covar* was used to calculate and to diagonalize to the covariance matrix. Here, the PCA was applied to the backbone atoms excluding the heavy fluctuating residues T252-N264 of the RecA1 domain and the residues (including the β-hairpin) T401-I421 of the RecA2 domain. The interpolation between the extreme projections of the free simulations onto the first PCA vector of RecA1 and RecA2 are shown in Fig. 6.9.

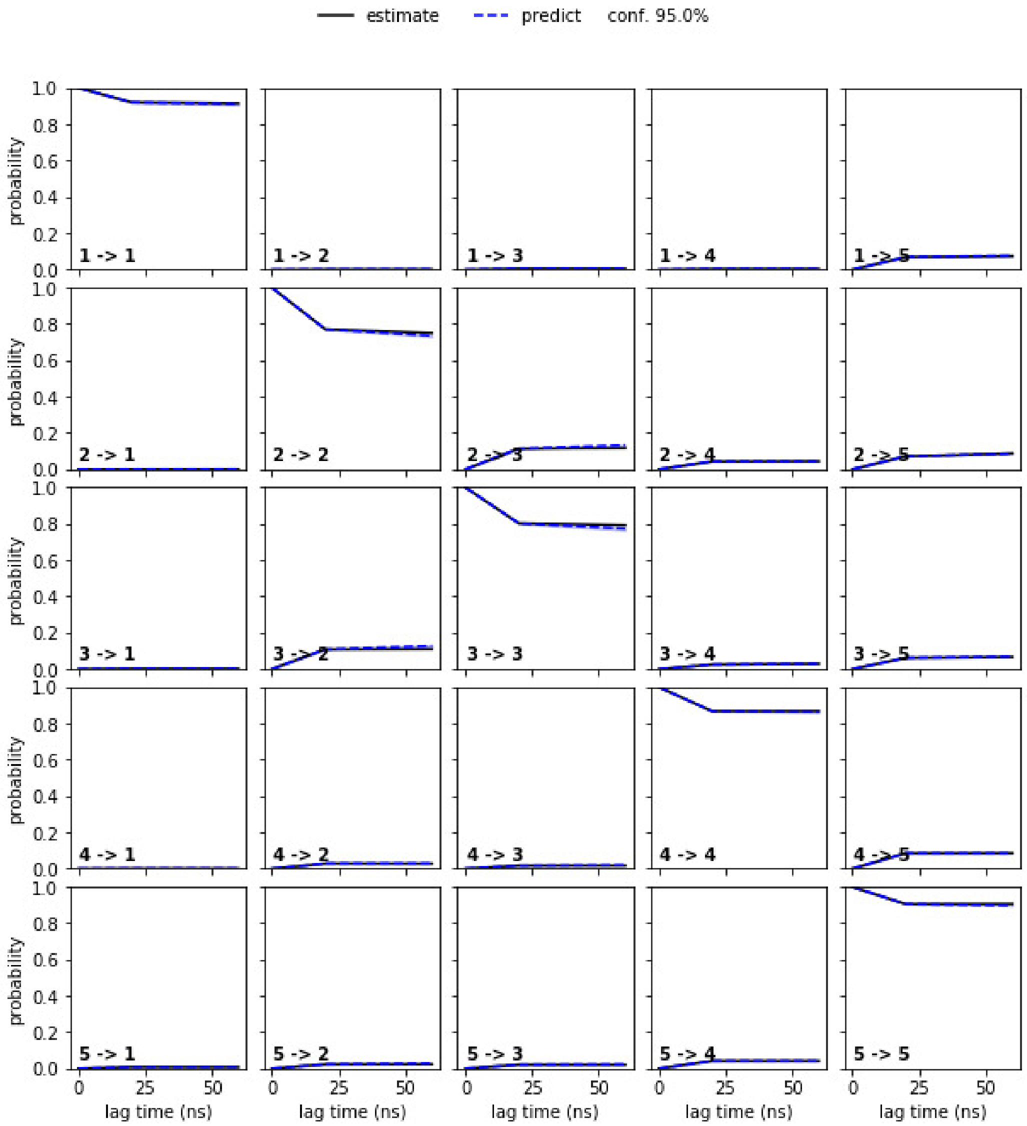


Figure 6.3: *Chapman-Kolmogorov test for the validation of macrostates of the MSM.*

6.2 Results

6.2.1 Full translocation cycle of Prp43

We decomposed the translocation cycle into two major transitions: (i) the "opening transition" involving the opening of the RecA1–RecA2 interface and the sliding of RecA2 along the RNA by one base pair; and (ii) the "closing transition", characterized by the closure of the RecA1–RecA2 interface and the sliding of the RNA along RecA1. AS of the opening process started from the Prp43•U_7•ATP structure (PDB code 5LTA [29], Fig. 1.4) and was triggered by the removal of ATP, thereby modeling Prp43•U_7 after the dissociation of the hydrolyzed ATP. Since the presence of ATP stabilized cationic moieties at the RecA1/RecA2 interface, removal of ATP led to a electrostatic repulsion between the two RecA-like domains. To monitor the opening transition, we selected a set of 18 structural features whose target values were taken from an open structure of the homologous Prp22 (Tab. 6.2, PDB ID 6I3P [30]). The opening process was completed after nine rounds of AS, as evident from a reasonable agreement of the structural features with their target values (Tab. 6.2, middle columns). Here, each round was composed of 40 to 500 simulated tempering simulations of 10 to 100 ns. The concatenated successful simulations summed up to a simulation time of 580 ns, whereas the invested overall simulation time was 56.5 µs.

Table 6.2: **Structural features used to monitor the progress of opening and closing transitions.** Opening: Starting values taken from the closed crystal structure of Prp43 (PDB ID 5LTA), target values taken from open structure of Prp22 (PDB ID 6I3P), and final values of simulated opening process (FoO). Closing: Starting values from the final opening simulation frame (FoO), target values from the closed crystal structure of Prp43 with the RNA shifted by one nucleotide, and final values of the simulated closing process (FoC). Overall, the features of the final simulation frames of the opening and closing are in good agreement with the reference values for the corresponding features.

		Opening			Closing		
Feature	[unit]	5LTA	6I3P	FoO	FoO	s5LTA*	FoC
RecA1 COM – COM RecA2	[nm]	2.65	3.15	3.12	3.12	2.65	2.67
G349 H – O^{1A} U5	[nm]	0.19	0.71	0.70	0.70	0.70	0.74
T381 H$^{\gamma}$ – O^{1A} U5	[nm]	0.26	0.75	0.87	0.87	0.81	1.03
R435 H$^{\eta}$ – ATP	[nm]	–	–	–	0.75	0.38	0.40
K403 H$^{\zeta}$ – O^{1A} U4	[nm]	0.29	0.96	0.77	0.77	0.91	0.93
E316 H – O^{1A} U4	[nm]	0.21	0.72	0.71	0.71	0.70	0.73
S387 H$^{\gamma}$ – Mg	[nm]	–	–	–	1.44	0.26	0.49
S387 H$^{\gamma}$ – O^{1A} U5	[nm]	1.58	0.19	0.17	0.17	1.42	1.18
G349 H – O^{1A} U4	[nm]	0.92	0.24	0.21	0.21	0.17	0.20
G349 H – O^{2A} U4	[nm]	0.73	0.31	0.30	0.30	0.32	0.32
T381 H$^{\gamma}$ – O^{1A} U4	[nm]	0.90	0.17	0.17	0.17	0.16	0.18
T381 H$^{\gamma}$ – O^{2A} U4	[nm]	0.73	0.38	0.42	0.42	0.40	0.42
R180 H – O^{1A} U5	[nm]	–	–	–	0.83	0.18	0.20
R153 H – O^{1A} U5	[nm]	–	–	–	0.62	0.19	0.18
R180 C – O^{1A} U7	[nm]	–	–	–	0.48	0.90	1.04
K403 H$^{\zeta}$ – O^{1A} U3	[nm]	0.87	0.27	0.28	0.28	0.33	0.28
S387 Phi	[°]	59	-65	-68	-68	59	51
S387 Psi	[°]	-47	-47	-33	-33	-47	43

The closing simulation started from the final frame of the opening simulation and was triggered by inserting ATP into the binding pocket of the RecA1 domain. We monitored the progression of the closing transitions by comparing the selected structural features with their values in the closed Prp43•U$_7$•ATP complex (PDB code 5LTA). The closing transition was completed after 13 rounds of AS, again revealed by a reasonable agreement of the structural features with their respective target values (Tab. 6.2, last columns). A typical table used to monitor the progression of the structural features is shown in Fig. 6.1 in the corresponding Methods and Material section. Here, each round involved 10 to 50 individual simulations of 10 to 100 ns each, summing up to an total simulation time of 40 µs, while the concatenated successful simulations summed up to 1.2 µs. The increased simulation time of the concatenated trajectory as compared to the opening simulation likely reflects that the formation of a well-defined protein–protein interface is more challenging than the rupture of an interface.

In summary, by using AS augmented with ST, we obtained a complete conformational cycle of the motor enzyme Prp43. As described in the following, the concatenated successful simulations provided an unprecedented atomic view on the function of a helicase, involving large-scale domain motions as well as the atomic-level rearrangements that were required for RNA translocation.

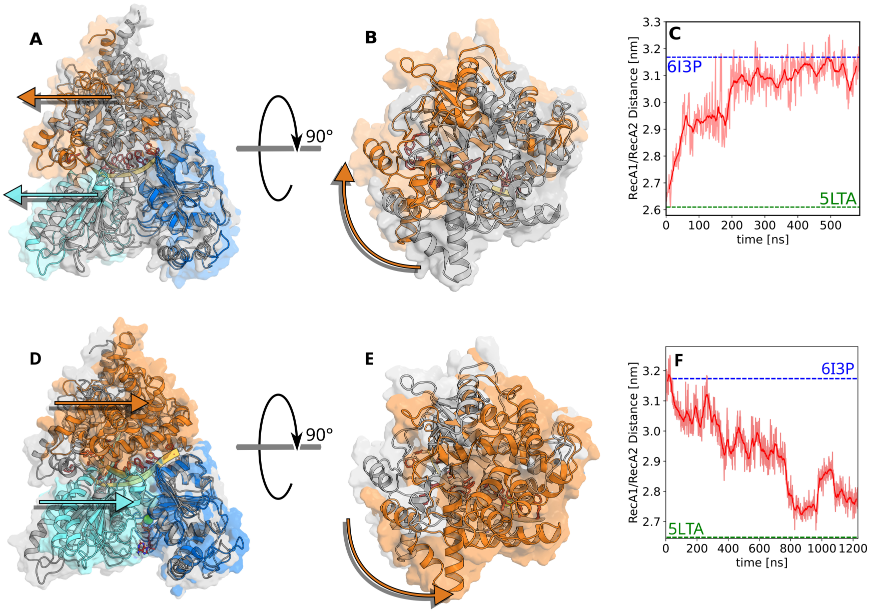


Figure 6.4: *Domain movements during the conformational cycle.*
(A–C) Opening transition and (D–F) closing transition of Prp43. (A/D) Front view and (B/E) top view at the beginning (grey) and end (multi-colored) of the respective process taken from the initial and final frames of the opening or closing trajectory, respectively. Arrows highlight the motions of RecA2 domain (cyan) and CTD domain (orange) relative to RecA1 domain (blue). (C) Center-of-mass distance between RecA1 and RecA2 domains during opening and (F) during closing. Dashed lines indicate the RecA1–RecA2 distances in the closed Prp43 structure (green, pdb code 5LTA) or in the open structure of the homologous Prp22 (blue, pdb code 6I3P).

6.2.2 Large-scale domain motions

Fig. 6.4A–C shows the motions of the RecA2 and the C-terminal domain relative to the RecA1 domain during the opening transition. After release of the ATP, the center-of-mass (COM) distance between RecA1 and RecA2 increased rapidly by 0.3 nm within the first 80 ns of the cumulative simulation time of successful AS simulations. Here, the cumulative simulation time of successful AS runs should not be confused with the by far longer physical time that would be required to observe such transition by a single conventional MD simulation (Fig. 6.4C). The RecA1–RecA2 COM distance exhibited a second sudden increase at 200 ns, followed by a gradual relaxation towards the value of the open conformation of the homologous Prp22 structure. The ongoing fluctuations of the open state are in line with the crystallographic data by Ficner et al. [30], who observed high B-factors of the RecA2 domains in the open Prp22 structure. The sudden changes of the RecA1–RecA2 distance correlate with transitions of molecular switches discussed below, suggesting that large-scale domain motions are controlled by atomic-scale transitions of molecular switches.

Previous crystallographic data revealed that the C-terminal domain (CTD) may carry out large-scale motions relative to RecA1 and RecA2, which are likely required for loading of the RNA into the RNA tunnel along the interface between CTD and RecA1/RecA2 (proposed by Tauchert et al. [29], see also Fig. 6.4A/D). By visual inspection of the simulations and by analyzing the center-of-mass motions of the CTD relative to RecA1 and RecA2, we found that CTD and RecA2 translate concertedly along the RNA during the opening transition while RecA1 remains bound to the RNA (Fig. 6.4A, arrows; Fig. 6.5). In addition to the center-of-mass displacement of the CTD relative to RecA1, the CTD carried out a rotation around a hinge located at the backside of the enzyme (Fig. 6.4B, arrows). This rotation is compatible with the presence of different CTD arrangements observed by X-ray crystallography in different ligand states of Prp43 [29].

During the closing process, the overall domain motions were reversed relative to the opening transition, characterized by a concerted motion of CTD and RecA2 relative to RecA1 (Fig. 6.5). However, in contrast to the opening process, the RNA was tightly bound to RecA2 in the closing process, thereby translocating by one nucleotide along RecA1. The RecA1–RecA2 distance decreased in three major steps at 300 ns, 750 ns, and at 880 ns (Fig. 6.4F), which again correlated with transitions of molecular switches discussed below. The final RecA1–RecA2 distance was in excellent agreement with the Prp43•U_7•ATP structure, suggesting that the RNA translocation cycle was completed.

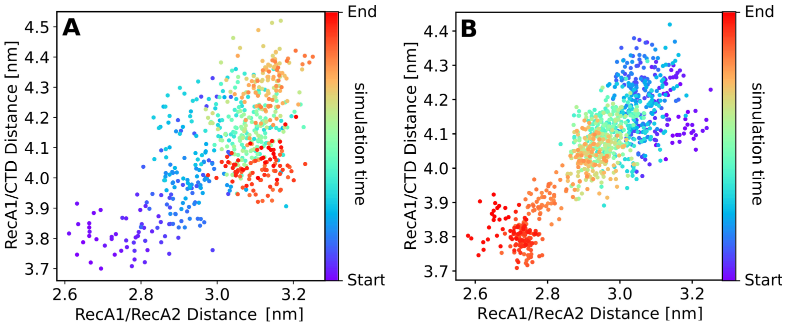


Figure 6.5: *Center-of-mass distances between C-terminal domain relative to RecA-like domains.*
(A) RecA1–CTD distance versus RecA1–RecA2 distance during opening and (B) during closing. Here, the center of mass of CTD was defined with the helices S544–S555, D583–R600, and Y615–K629. The color indicates the cumulative simulation time from start (purple) to end (red). During both the opening (A) and (B) closing process, the RecA1–CTD distance was correlated with the RecA1–RecA2 distance, illustrating that the CTD domain moved concertedly with RecA2 along the RNA, as shown in Figure 2 in the paper.

6.2.3 Kinetic models of opening and closing processes

We used two complementary models to obtain the approximate kinetics of the opening and closing process. First, based on the formalism by Pande and Singhal [182], we modeled the Prp43 dynamics by a linear sequence of transitions, providing an intuitive, simple, and numerically robust kinetic model of the opening and closing processes. Second, we constructed a Markov state model (MSM) which provides, in addition to the approximate kinetics, a view on the underlying free energy landscape and on the conformations of metastable states [156, 157, 187–189]. While MSMs rely on an elaborate theory for dimensionality reduction and kinetic modeling, MSM have been shown to be sensitive with respect to limited sampling, as common when simulating complex protein dynamics [190].

Based on the formalism by Pande and Singhal [182], we modeled both the opening and closing process as seven-step processes, and we estimated the transition state matrix from the successful AS simulations (see Methods and Materials). For the opening and closing processes, we obtained MFPTs in the order of ~100 μs and ~50 μs respectively. Assuming ST accelerates the kinetics by one order of magnitude [111], these values translate into physical MFPTs in the order of 1 ms and 0.5 ms, respectively. This implies a maximum translocation speed of ~600 base pairs per second (bp/s). Owing to contributions from ATP binding, hydrolysis, and release, which may even be rate-limiting for the overall cycle, the translocation speed of Prp43 is likely lower than the maximum value of ∼600 bp/s estimated from our simulations. Notably, assuming some reduction of the translocation speed from ATP binding, hydrolysis and release, our value is in reasonable agreement with translocation speeds of 100 to 300 bp/s observed for other helicases. [27, 191, 192]

Complementary, we derived a MSM of the overall conformational cycle by combining all AS trajectories. The free energy landscape projected onto two independent components revealed several well-separated metastable states, as visualized in Fig. 6.6. The MSM suggested a MFPT of the closing process of ∼40 μs, in reasonable agreement with the value of 50 μs obtained by the linear kinetic model. Furthermore, our MSM passed widely used quality controls, indicating reasonably Markovian dynamics (Figs. 6.2 and 6.3 in Material section). However, the MSM suggested a MFPT of only ∼10 μs for the opening process, which is tenfold lower than the value of 100 μs obtained by the linear kinetic model. This bias in the MSM is likely explained by insufficient sampling of the rate-liming loop-to-helix transition of the sensor serine described below. Hence, whereas the MSM and the free energy landscape presented here should be interpreted with care, the linear kinetic model from successful AS simulations provided a simple and numerically robust approximation to the Prp43 kinetics.

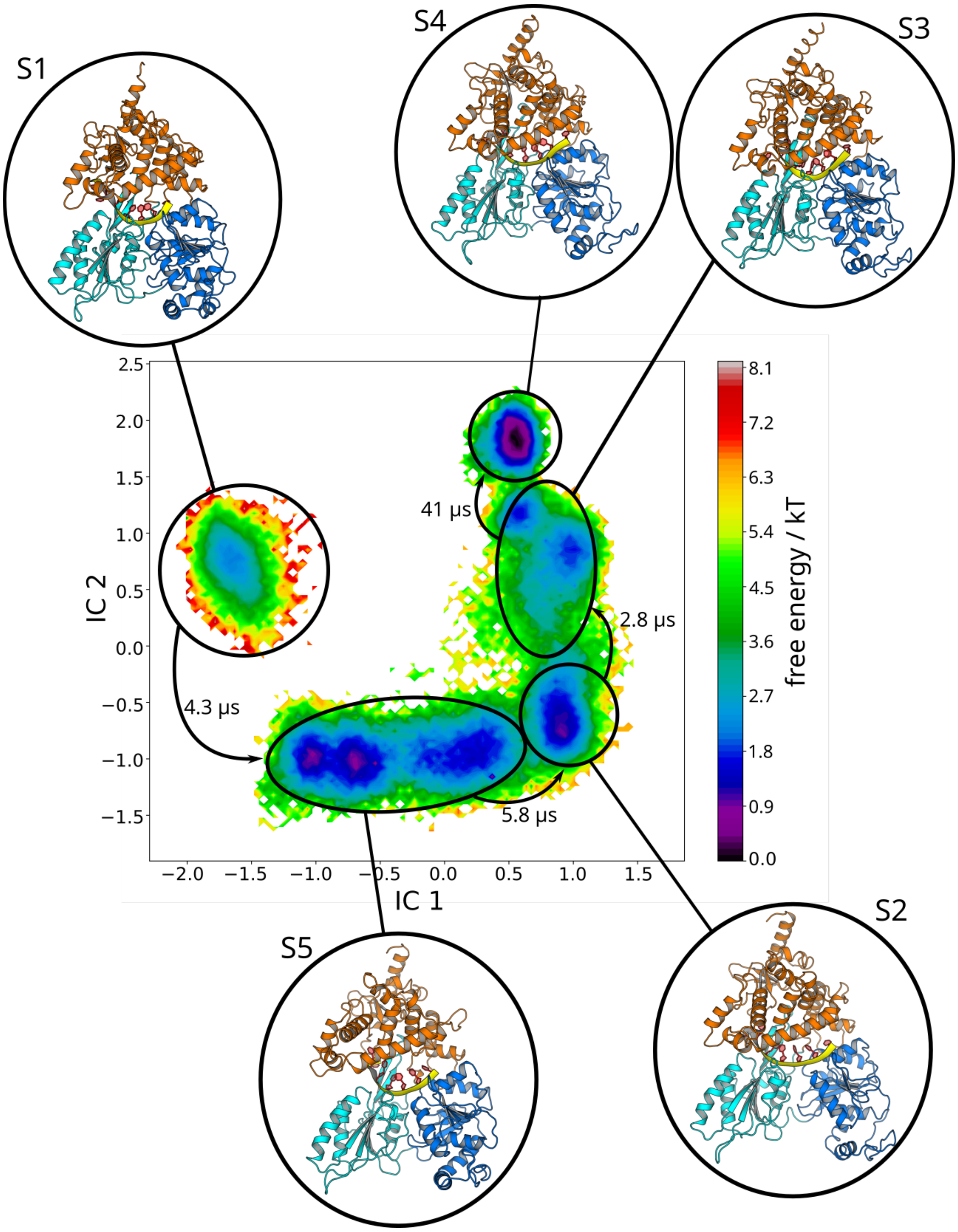


Figure 6.6: *Estimate of the free energy landscape calculated from the stationary distribution during the construction of the MSM. Five macrostates are encircled with black lines, each connected with a representative Prp43 conformation. The translocation path is indicated by arrows with the corresponding MFPT from state to state as suggested by the MSM. The metastable states corresponding to the free energy basins are clearly revealed by the MSM.*

6.2.4 Molecular switches of the opening transition

Transition of the RecA2 β-hairpin and RNA backbone rotation.

As the first critical transition after removal of the ATP, the β-hairpin of RecA2 carried out a rapid rearrangement within the first nanoseconds of the AS simulations (Fig. 6.7A–C). The rearrangement involved a cleavage of the K403–U4 H-bond and a sudden drop of the K403–U3 distance from 0.9 nm to 0.4 nm (Fig. 6.7C). This rapid transition is an indicator of a strong tension in the RNA–RecA2 interface of the closed state, which was likely stabilized by electrostatic interactions of the ATP with RecA1 and RecA2. Hence, the tension was released in an instant after ATP release.

Cleavage of the K403–U4 H-bond allowed the formation of the K403–U3 H-bond after 180 ns of the cumulative simulation time, as required for sliding of RecA2 along the RNA by one base pair (Fig. 6.7C, 180 ns). To enable the formation of the K403–U3 H-bond, a rotation of the RNA backbone around the P–P axis between U3 and U4 was strictly required, thereby pointing the H-bond acceptor O^{1A} of U3 towards K403 (Fig. 6.7C, black arrow). This RNA backbone rotation was, in turn, only enabled by the major opening step during the first 100 ns described above, which rendered RecA2 more flexible and provided the RNA with increased conformational freedom. The successful shift of K403 from U4 to U3 left the U4 phosphate vacant, which enabled the U4 phosphate to rotate and to form a critical connection of U4 with the hook-loop and with T381 described below (Fig. 6.7I).

Arginine finger R435 of RecA2 as anchor for the S387–G392 sensor loop.

Arginine fingers are a reoccurring motif of NTP binding sites, where they are crucial for stabilizing the NTP hydrolysis complex ([193–195]). In the closed conformation of Prp43, the arginine finger R435 of RecA2 is part of the ATP binding pocket and interacts with the β- and γ-phosphate of ATP, thereby bridging the two RecA-like domains (Fig. 1.4). The nearby residue T389 interacts with R435 and with the adenosine moiety of ATP (Fig. 1.4, Fig. 6.7D). After release of the ATP, the arginine finger shifted away from the RecA1 domain and broke the contact with T389 (Fig. 6.7E). Cleavage of the R435–T389 contact allowed increased flexibility of the loop S387–G392, to which we refer as “sensor loop” since it contains the sensor serine S387 [30]. This led to an unfavorable steric clash of the sensor loop with the R153–L167 helix of RecA1, possibly contributing to a repulsion between the two RecA-like domains. At this stage, any further transition of the sensor loop towards its conformation in the open state was obstructed by the current H-bond of

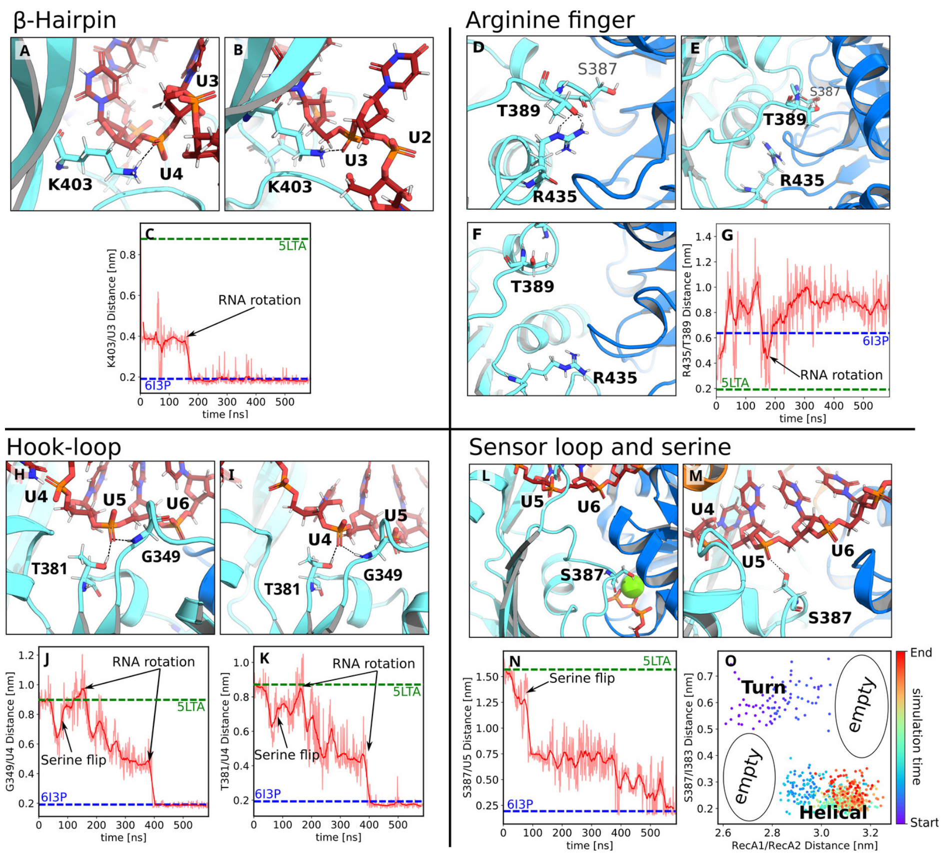


Figure 6.7: *Molecular switches of the opening process. Conformational transitions of (A–C) the β-hairpin, (D–G) the arginine finger, (H–K) the hook-loop, and (L–O) the sensor loop during Prp43 opening. (A/D/H/L) Molecular switches in the initial and (B/F/I/M) final frames of adaptive sampling of the open process. (C/G/J/K/N) Dashed lines indicate the distances in the 5LTA and 6I3P crystal structures. (A) View from the backside of Prp43: K403 of the β-hairpin hydrogen bound to U4 in closed conformation and (B) shifted from U4 to U3 in open conformation. (C) K403–U3 distance versus cumulative simulation time. (D) Closed RecA1–RecA2 interface with the Arginine finger R435 hydrogen bound to T389 and located on top of the P-loop. (E) Broken R435–T389 hydrogen bond after 5 ns. (F) Final conformation with distant R435 and T389 residues and with the arginine finger located underneath the P-loop. (G) R435–T389 distance versus cumulative simulation time. (H) Closed conformation, hook-loop bound to U5 via G349 and T381. (I) Open conformation, G349 and T381 shifted from U5 to U4. S387 formed an H-bond to U5 similar to the conformation in panel M. (J) G349–U4 distance and (K) T381–U4 versus cumulative simulation time. (L) Serine finger S387 bent towards the ATP in closed conformation and (M) pointing towards the RNA forming an H-bond with U5. (N) S387–U5 distance versus cumulative simulation time. (O) S387–I383 distance vs. RecA1–RecA2 distance.*

U5 with the hook-loop and with T381 (Fig. 6.7H).

Upstream motion of the hook-loop and T381 from U5 to U4.

The hook-loop is an important anchor between the RecA2 domain and the RNA. At the beginning of the cycle in the closed state, G349 of the hook-loop and the nearby T381 form H-bonds with the U5 phosphate of the RNA (Fig. 6.7H). During the opening transition, these H-bonds shift by one nucleotide upstream from U5 to U4 (Fig. 6.7I). As illustrated in the G349–U4 and T381–U4 distances in Fig. 6.7J/K, the upstream motion of G349 and T381 occurred in two steps: the first step started at 180 ns and correlated with the RNA backbone rotation between U3 and U4 and the formation of the K403–U3 H-bond described above (Fig. 6.7B/C). The second step at 400 ns, accompanied by a rotation of U4, finalized the upstream motion of G349 and T381 (Fig. 6.7I). This transition left U5 vacant, henceforth allowing the H-bond formation by the sensor serine with U5.

Loop-to-helix transition of sensor loop and binding of sensor serine S387 to U5.

Previous crystallographic data suggested that transitions of the "sensor serine" S387 are critical for RNA translocation [30]. After the cleavage of the H-bond between the arginine finger with T389 described above, the sensor serine S387 (as part of the sensor loop S387–G392) carried out a loop-to-helix transition, decreasing the S387–U5 distance by ≈7 Å (Fig. 6.7N, 100 ns, black arrow). However, only after the upstream motion of the β-hairpin from U4 to U3 (Fig. 6.7A–C, 180 ns) and of the hook-loop from U5 to U4 (Fig. 6.7H–K, 400 ns), the U5 phosphate was vacant, enabling the sensor serine to form a new H-bond with the RNA (Fig. 6.7M/N). The formation of the S387–U5 H-bond finalized the successful opening transition.

Notably, simulating the loop-to-helix transition of S387 with AS was challenging. Among 100 simulation of 100 ns, we observed only a single successful loop-to-helix transition, likely because such changes of secondary structure are slow in sterically tight environments (Fig. 6.8). Hence, the loop-to-helix transition was a rate-limiting step of the opening transition after the release of ATP.

We found that the degree of opening, as given by the RecA1–RecA2 distance, is correlated with the loop-to-helix transition of sensor serine. Namely, a large RecA1–RecA2 distance of ∼3.1 nm could be maintained only if the sensor loop was in the helical state (Fig. 6.7O). As mentioned, we performed an independent test for this correlation in section 4 These simulations confirm

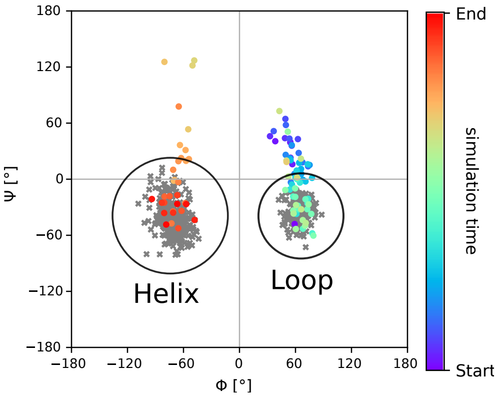


Figure 6.8: *Ramachandran plot of sensor serine S387.*
Grey crosses: ϕ and ψ angles of S387 during free 1 µs simulations starting from the 6I3P structure (left), corresponding to the open conformation, or starting from the 5LTA structure (right), corresponding to the closed conformation. According to the ϕ/ψ angles, S387 remained in the helical or in a loop state in the simulations of the open or closed state, respectively. No transition was observed. Colored dots: ϕ/ψ angles during a AS simulation with a successful loop-to-helix transition of the sensor serine. The color indicates the simulation time of the successful AS run from purple to red.

the findings from AS (Fig. 6.7O), namely that the loop-to-helix transition by the sensor loop is strictly required for a successful opening transition of Prp43.

Hook-loop and serine loop translocations are encoded in the RecA2 dynamics.

As described above, both the hook-loop and serine loop transitions are strictly required for a successful Prp43 opening. Hence, we asked whether these dynamics are guided by the RecA2 interactions with the RNA, or whether they are intrinsically encoded in the dynamics of the RecA2 structure. To this end, we carried out an additional microsecond simulation of the isolated RecA2 domain and analyzed the intrinsic dynamics using principal component analysis (PCA; Fig. 6.9A). The PCA revealed that both the hook-loop and serine loop are highly dynamic in the isolated RecA2 domain, and the first PCA vector describes transitions as observed during the upstream sliding along the RNA. This finding suggests that the RecA2 structure and its intrinsic dynamics have been optimized for enabling the critical H-bonds shifts descried above and, thereby, for RecA2 sliding along the RNA.

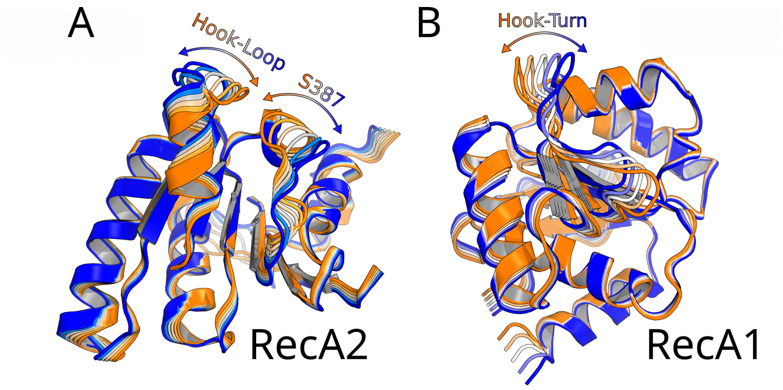


Figure 6.9: *PCA of isolated RecA1 and RecA2.*
A: Motion along the first PCA vector of the isolated RecA2 domain visualized as 6 frames from blue to orange. The hook-loop and sensor serine exhibit the largest contributions to RecA2 fluctuations after excluding the highly mobile β-hairpin from PCA. B: Motion along the first PCA vector of the isolated RecA1 domain. The hook-turn largely contributes to RecA1 fluctuations.

6.2.5 Molecular switches of the closing transition

Anchoring the arginine finger to the ATP.

We triggered the closing process by inserting ATP into the binding pocket of the RecA1 domain of the final frame of the opening simulations. The arginine finger R435 formed a stable contact with the β and γ phosphates of ATP within only 40 ns of a successful AS simulation (Fig. 6.10A–C). This rapid transition, driven by strong electrostatic R435–ATP interactions, anchored the RecA2 domain to the RecA1 domain via the ATP.

Reverse transition of the sensor serine S387 is critical for closing the RecA1/RecA2 interface.

The sensor serine S387 played multiple critical roles during the closing process. S387 rapidly lost contact with U5 (Fig. 6.10D/F), thereby enabling U5 to form an H-bond with R153 of the RecA1 domain, as required for sliding the RNA along the RecA1 domain (Fig. 6.10H–J). In addition, the sensor loop carried out a helix-to-loop transition, thereby extending S387 underneath the R153–L167 helix of RecA1 where S387 interacts with a water molecule of the ATP–Mg^{2+}–water complex to stabilize the closed state. Notably, arrangement of water in the final closed state resembled the structure identified by crystallography [29, 170] (Fig. 6.11). The Mg^{2+} ion was coordinated with three water molecules, thereby bridging interactions with

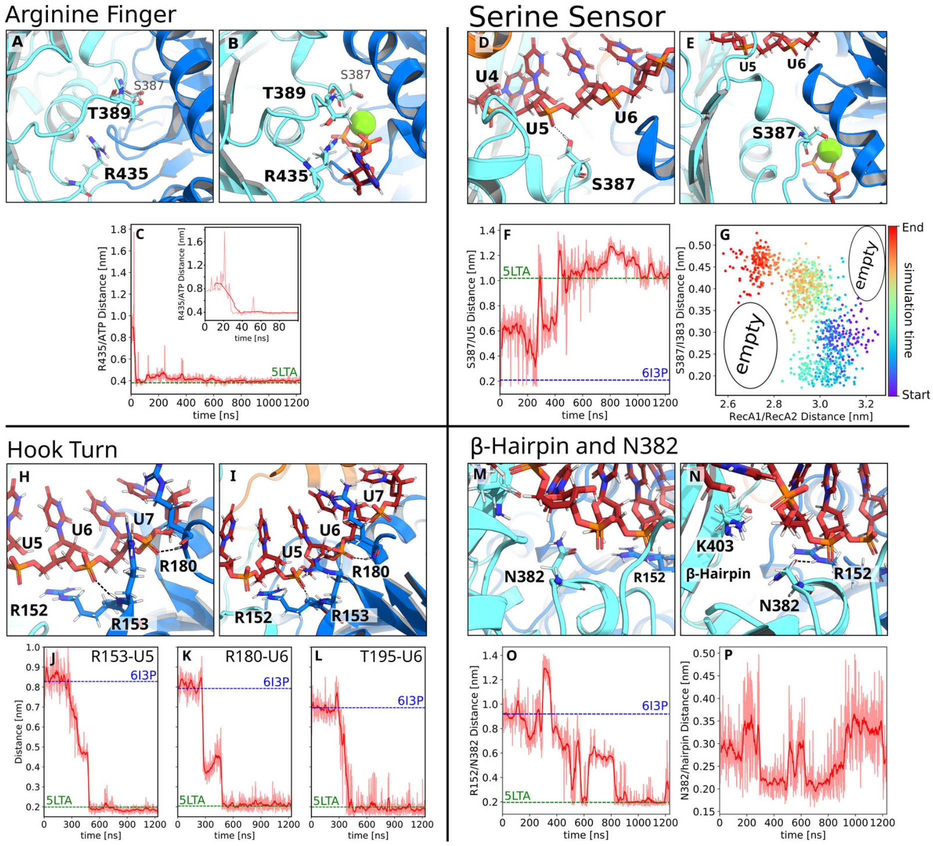


Figure 6.10: *Molecular switches of the closing process.*
Conformational transitions of (A–C) the arginine finger, (D–G) sensor serine, (H–L) RecA1–RNA interactions, and (M–P) β-hairpin during Prp43 closing. (A/D/H/M) Molecular switches in the initial and (B/E/I/N) final frames of adaptive sampling of the closing process. (C/F/J–L/O/P) Critical atomic distances that quantify the progression of the closing transition. Dashed lines indicate the distances in the 5LTA and 6I3P crystal structures. (A) Open RecA1/RecA2 interface with distant R435 and T389 residues. (B) Closed conformation with ATP in the binding pocket, thereby bridging the closed RecA1/RecA2 interface. (C) R435/C_ζ–ATP/O3B distance, revealing an R435–ATP H-bond formation early during the closing process. (D) Open conformation with the sensor serine S387 in the helical state and forming an H-bond with U5 of the RNA. (E) Closed conformation with S387 in the loop conformation, bound to the ATP–Mg^{2+} complex, and reaching underneath the RecA1 R152–L167 helix. (F) R435–U5 distance during the closing process. (G) S387–I383 distance versus RecA1–RecA2 distance. (H) Open conformation with R153 and R180 of RecA2 hydrogen bound to U6 and U7, respectively. (I) Closed conformation with R153 and R180 hydrogen bound to U5 and U6, respectively. (J–L) R153–U5, R180–U6, and T195–U6 distances during the closing process. (M) Open conformation with N382 interacting with the β-hairpin. (N) Closed conformation with N382 interacting with R152. (O/P) R152–N382 distance and N382–β-hairpin distance.

nearby residues D218, E219, and S387. R432 and R435 interacted with water in the ATP pocked and directly with the phosphate moieties of the ATP, as observed by Tauchert *et al.* [29, 170]

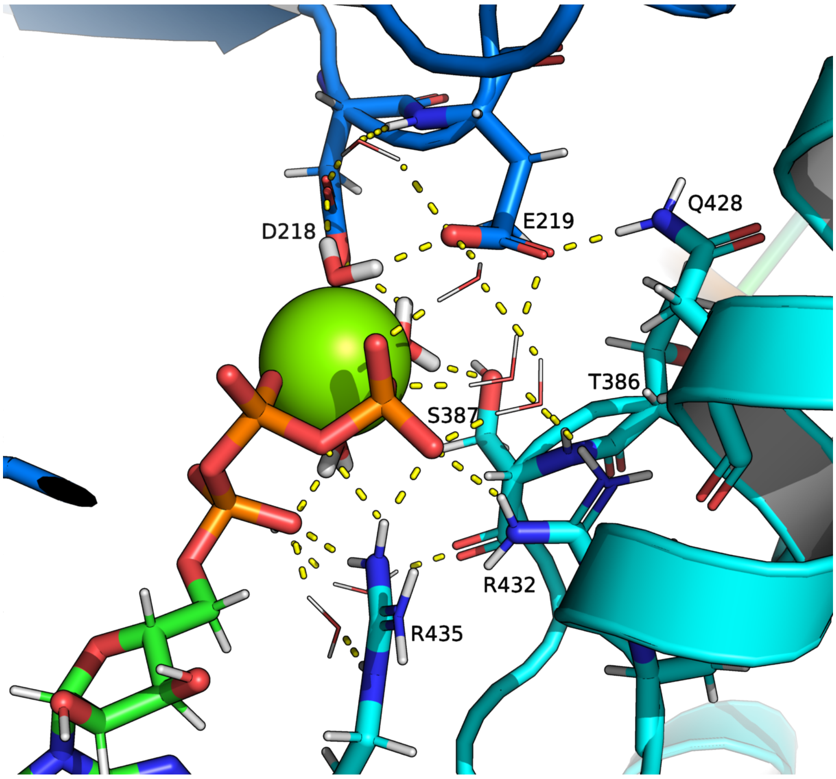


Figure 6.11: *Close-up view on the water complex and network during the end of the closing process. The network is similar to the network proposed by Tauchert et al. for the closed crystal structure 5LTJ. However, the water molecule in position for a S_N2 attack shown by Tauchert et al. was not observed in our simulations. Three water molecules coordinating the Mg^{2+} ions are shown as thick sticks, all other water molecules as thin sticks.*

The helix-to-loop transition of the S387–G392 segment was strictly correlated with the closing of the RecA1–RecA2 interface, in line with the correlation between the loop-to-helix transition and the Prp43 opening described above (Fig. 6.10G and Fig. 6.7O). Namely, only after the transition to the loop state, as evident from an increased S387–I383 distance in Fig. 6.10G (red dots), a tight RecA1–RecA2 interface could form as indicated by a small RecA1–RecA2 distance. Additional support for this correlation was given by visual inspection of the simulations, confirming that the helix-to-loop transition followed by the extension of S387 beneath R152–L167 helix was required to enable tight packing of the RecA1/RecA2 interface without atomic clashes, as suggested previously from crystallographic data [29, 30].

RecA domain sliding along RNA. Sliding of the RecA1 domain along the RNA involves the upstream motion of both the hook-turn and the R152–L167 helix, characterized by the transition of the backbone amine groups of R153 from U6 to U5, of T195 from U7 to U6, and of R180 from U7 to U6 (Fig. 6.10H–L). However, these transitions were initiated only after the sensor loop carried out the helix-to-loop transition, and they were completed only after the sensor serine bound to the ATP-Mg^{2+}-water complex at 400 ns (Fig. 6.10D–F). In addition to the backbone interactions of R153 and R180 with the RNA, the guanidinium moieties of these arginines frequently bound to the RNA backbone. The occasional release of these guanidinium–RNA H-bonds were required for a successful upstream motion. This may indicate that a fine balance in strength and number of RecA1–RNA versus RecA2–RNA H-bonds dictates the upstream motion of Prp43.

To test whether the hook-turn dynamics are intrinsically encoded into the RecA2 domain, we carried out a microsecond simulation of the isolated RecA1 domain, analogous to the simulations of the isolated RecA2 domain discussed above. PCA revealed that fluctuations of the RecA1 domain are dominated by fluctuations of the hook-turn, while the hook-turn transition during RecA1 sliding observed during AS occurred approximately along the first PCA vector (Fig. 6.9B). This analysis complements the PCA of the RecA2 domain described above, together suggesting the largest-scale fluctuation of both RecA-like domains are optimized for enabling RNA sliding.

Visual inspection of the closing trajectory revealed that RecA1–RecA2 interactions are not exclusively established via the ATP. Instead, after sliding of RecA1 along the RNA, R152 of RecA1 frequently interacted with N382 of RecA2, while N382 occasionally interacted with the β-hairpin of RecA2 during the closing transition, which contributes to the tight packing of the RecA1–RecA2 interface and, thereby, to driving the upstream motion along the RNA (Fig. 6.10M–P). In summary, after the formation of the interface of RecA2 with the RecA1/ATP complex and the sliding of RecA1 along the RNA, the closing transition was completed.

Concerning the free energy landscape from the constructed MSM. In the following paragraph, we mainly discuss the free energy landscape obtained by the constructed MSM (6.6). The assigned the corresponding macrostates to the observed energy minima in the plotted free energy landscape. Thus, we add qualitative details to the free energy landscape to shed light on the individual transitions and their relation to each other. Usually,

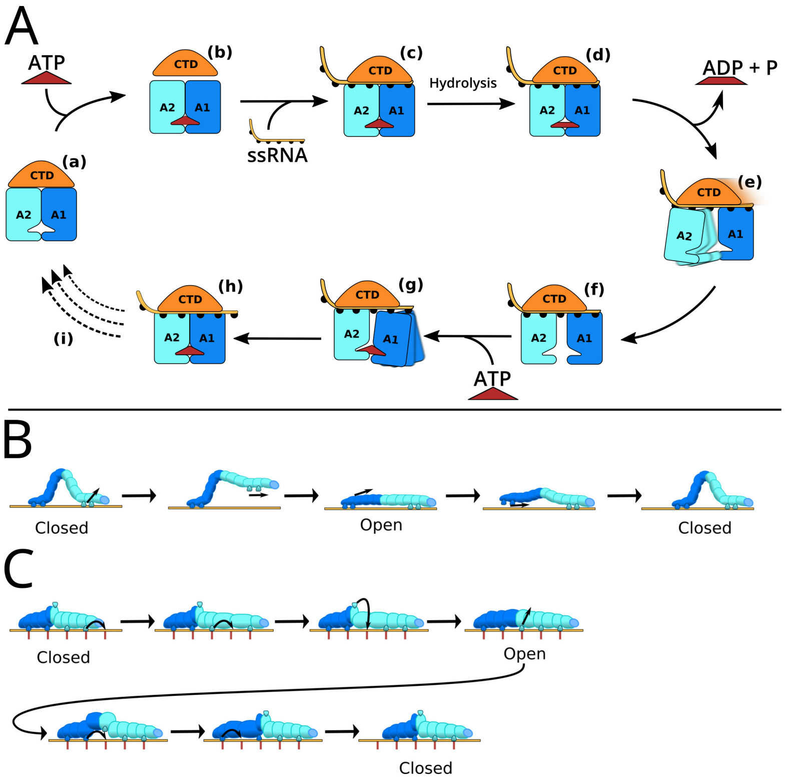


Figure 6.12: *A) Schematic hypothesis of the complete translocation cycle.*
(a) Apo structure. (b) ATP-binding triggers opening of the CTD/RecA interface, allowing (c) binding of ssRNA and formation of the Protein–ATP–RNA complex, represented by PDB ID 5LTA. (d) ADP bound state after ATP hydrolysis phosphate release. (e) RNA-bound state with open RecA interface and weak RecA2–RNA contacts. (f) Stable open structure, represented by PDB ID 6I3P. (g) ATP-binding triggers closure of the RecA interface and sliding of RecA1 along the RNA. (h) Protein–ATP–RNA complex translocated by one nucleotide relative to panel (c). (i) Transition back to the apo-structure after finalizing multiple RNA translocations.
B) The classical inchworm model.
C) Proposed inchworm/caterpillar model to illustrate both the center-of-mass motion of the RecA-like domains (dark blue and cyan) and the crawling of RecA-like along the RNA. Hydrogen bond partners of protein and RNA are sketched as caterpillar legs and as red lines, respectively. Since the RecA-like domains bind the RNA with four or five H-bonds in the closed and open state, respectively, the caterpillar requires five legs. The central leg, modelling the sensor serine S387, carries out a rotation to bind the free nucleotide binding site before reaching the open state.

this intend is difficult to achieve, since in the majority of studies, the curcial features for the dynamics are unknown a priori. In this study, we leverage the identification and sampling of the most crucial features, achieved by the combination of ST and AS, to drastically reduce the dimensionality of the feature space on which the MSM is constructed.

The following construction, analysis and validation were performed with the PyEMMA 2.5.7. [183]. For the construction of the MSM, we used all simulations obtained from the AS of the opening and closing process. Thus, the underlying data of the MSM contains more than 96 µs of simulation time. In the first step, we featurized the trajectories similar to the features of the AS procedure (see Methods section). The resulting dynamics-related data was reduced in dimensionality by using the tiCA and plotted onto the first two resulting independent components (ICs). These two ICs describe the two slowest motions according to the tiCA [156, 167, 184]. The Markovianity was validated with the convergence of the implied timescale estimations and the Chapman-Kolmogorov test [196, 197]. Then, we clustered the featurized data into 70 microstates with the *k-means* clustering algorithm. As suggested by the PyEMMA workflow, we confirmed the VAMP2 convergence of the number of cluster centers (Fig. 6.2B).

For the MSM construction, we used a lag time of 20 ns, because the implied time scales showed convergence beyond this lag time (Fig. 6.2C).

Afterwards, we can calculate the stationary probability distribution over the first two ICs. With this, we can calculate the free energy from the stationary probability p as follows

$$\Delta G = -k_B T \ln \Sigma p_i \tag{6.3}$$

where p_i is the stationary weight of the ith microstate. The resulting free energy plot projected onto the tICA components is shown in Fig. 6.6. The constructed MSM revealed five metastable states - also known as macrostates - after the Perron-cluster cluster analysis (PCCA+). Beforehand, we also checked the validation of the 5 metastable states using the Chapman-Kolmogorov test (Fig. 6.2D). With this result, we can calculate the MFPTs between the five macrostates. The free energy landscape, the macrostates, the conformations belonging to each macrostate and the calculated MFPTs from one state to the next state are shown in Fig. 6.6.

The high energy state S1 can transition into the S5 state with an MFPT of 4.3 µs. The transition region might not be sampled enough, because it contains the rare but fast loop-to-helix transition of the serine loop.

In the S5 region, we have multiple energy minima (Fig. 6.6), which all resemble states which are not yet fully open and not fully closed anymore, meaning they represent states of the transition from the closed conformation to the open conformation. However, the serine loop flip already occurred in all those conformations.

Then, from the S5 state, the helicase transitions with an MFPT of 5.8 µs into the lower energy state S2, which only contains open conformations. Therefore, the S1 to S5 to S2 represents the closed to open pathway, which was simulated in the AS opening process.

The S2 also contains conformations from the beginning of the AS closing process. Here, the S2 will traverse to the S3 state with an MFPT of 2.8 µs, which resembles another transition region from the closed to the open state. All the conformations in S3 represent a mix state of a closed state and a open state before and after the closing-defining Hook-Turn shift.

The last observed transition is observed from S3 to the S4 state with a large MFPT of 41.2µs. S4 is in an energy minimum and only contains fully closed conformations after the Hook-Turn shift.

Conclusively, we see a transition from S1 to S5 to S2 for the opening process and a transition from S2 to S3 to S4 for the closing process. S1 and S4 have a large energy difference, which might be explained by the fact that S1 is in a perturbed state - because of the removal of ATP - and S4 is its native state - bound to ATP. Therefore, the removal of ATP destabilizes the conformation and raised S1 to higher energy. The free energy of S2 might lay between the free energy of S1 and S4 because it is a mix state from simulations with and without ATP, where the simulations without ATP represent its open native state and the simulations with ATP represent the perturbed conformation.

The transition from S1 to S2 has an MFPT of around 10 µs, which is an order of magnitude lower than the MFPT calculated by the approach of Pande et al. However, the MFPT from the MSM of the closing process, i.e., from S2 to S4, is in total 44 µs and thereby in the same range as the MFPT from the linear approach (50 µs). The large difference in the opening MFPTs could result from the lack of sampling in the transition region from S1 to S2.

In summary, from our MSM, we gained interesting insights in the kinetics of the helicase Prp43. In our study, the most impact on the IC 2 of our tiCA decomposition most likely has the RecA1–RecA2 distance since a movement in this coordinate changes the protein from a closed to open conformation and vice versa. The IC 1 on the other hand may describes a combination of the distances of the contacts points between protein and RNA. Although S1 and S4 both describing the "same" closed state, they do not have to overlap in the free energy plot, since our definition of the features, describing the protein–RNA contacts, are dependent on the distance to the next contact point and not on the number of contact points alone. Therefore, a shift of

H-bonds will lead to a separation of the two closed conformations in tiCA space.

6.2.6 Discussion

All-atom simulations of complete conformational cycles of complex processes of enzymes, such as translocations, are still rare in the literature. Here we showed that AS simulations augmented with ST were capable of obtaining a complete RNA translocation cycle of the 80 kDa helicase Prp43. The simulations together with crystallographic studies suggest a translocation mechanism involving the domain rearrangements shown in Fig. 6.12A. Accordingly, ATP binding to apo Prp43 opens the RNA cleft, thereby allowing the binding of RNA (Fig. 6.12A(a–c)) [29]. The hydrolysis of ATP enables the release of ADP and phosphate, while the phosphate ion is predominantly released via a tunnel on the backside of Prp43 [177]. Removal of these negative charges leads to a spring-like conformational change in the ATP binding pocket, driving the movement of RecA2 by one RNA base upstream (Fig. 6.12A(d–f)). Binding of the next ATP triggers the closure of the RecA1–RecA2 interface by moving RecA1 along the RNA (Fig. 6.12A(g–h)). In our simulations, the CTD moved concertedly with the RecA2 domain (Fig. 6.12A(e–h)).

The simulations revealed that the key domain transitions are by no means characterized as diffusive center-of-mass movement of the entire domains. Instead, the large-scale domain dynamics are controlled by atomic-scale molecular switches that occur stochastically along orthogonal degrees of freedom of a highly rugged free energy landscape. Such dynamics have been studied extensively in the context of protein folding, yet much less for conformational cycles of motor enzymes such as Prp43 [198, 199]. Because the transitions of the molecular switches are interdependent, they occurred in a defined temporal order. Hence, our study highlights that large-scale domain motions of enzymes like Prp43 are controlled by atomic-scale molecular switches. This further implies that a mechanistic understanding of the enzyme kinetics, for instance involving the regulation by G-patches [80, 84, 200, 201], requires identification of the molecular switches and understanding of their kinetics.

The overall domain displacement of Prp43 is compatible with an inchworm model [22]. However, an inchworm-like picture may imply that, during upstream motions, the RecA1 and RecA2 domains would fully detach from the RNA and re-bind one nucleotide upstream [30], which is not observed in our simulations (Fig. 6.12B). Instead, RecA1 and RecA2 remain bound to the RNA throughout the cycle and individually crawl along RNA by shifting protein–RNA H-bonds one-by-one, similar caterpillar walking. Hence, we suggest an inchworm/caterpillar model to describe both the relative domain displacements (inchworm) as well as the upstream movements of the individual RecA–RNA contacts (caterpillar; Fig. 6.12C).

In this study, we did not simulate ATP binding or hydrolysis or the release of ADP and phosphate, but instead focused on the domain motions initiated by the insertion and removal of ATP. Hence, the simulations demonstrate that the mere presence or absence of ATP is sufficient to trigger closing or opening of Prp43, respectively, thereby driving RNA translocation on the hundreds of microsecond timescale. This finding suggests that the energy from ATP hydrolysis is not required to generate a "kinetic push" towards Prp43 opening but merely to allow dissociation of the ADP/phosphate products. Together, binding and release of ATP modulates the minima of the rugged free energy landscape, on which the domains move in a stochastic fashion.

Since the overall kinetics of RNA translocation are limited by transitions of molecular switches and not by diffusion of the overall domains, simulating such enzymatic cycles is challenging. Specifically, enhanced sampling techniques that merely enhance the diffusion of the complete domains or steer center-of-mass distances between domains are barely useful for such a system, since all the rate-limiting transitions occur in conformational space orthogonally to the domain center-of-mass distances [202]. Instead, methods such as milestoning [203] or, as used here, AS are suitable for sampling large-scale domain motions of enzymes in rugged energy landscapes as it does not require the definition of reaction coordinates [110]. As a disadvantage, AS may bias the ensemble owing to the selection of the successful simulations, which may lead to over-representation of states with higher free energy; in this study, the enhanced sampling using ST likely reduces such bias as ST accelerates the re-equilibration of the simulations at the beginning of the next AS round.

In addition, AS is suitable for estimating the kinetics of the overall pathway by collecting the rates between neighboring metastable states. For the successful sequence of transitions, using a simple linear kinetic model, we estimated the MFPTs of the opening and closing transitions in the order of 1 ms or 0.5 ms, respectively (assuming tenfold accelerated rates owing to the use of ST). These values are in reasonable agreement with experimental data for other helicases. [27, 191, 192] Complementary, we constructed a MSM for the conformational cycle. While the MSM suggested a closing rate in reasonable agreement with the linear kinetic model, the MSM overestimated the opening rate, which we ascribe to insufficient sampling of the rate-limiting loop-to-helix transition of the sensor serine. Hence, for complex transitions as studied here, for which constructing a converged MSM remains challenging, the linear kinetic model provides a numerically robust and useful alternative.

Notably, by using AS, the computational cost for obtaining a conformational cycle was reduced only by a factor of approximately two compared to using few long simulations, primarily because the opening and closing processes occur down the gradient of the free energy landscape (Materials and Methods for details). Hence, a key advantage of AS for the present study was also the ability to trivially parallelize the simulations with commodity hardware, thereby drastically reducing the elapsed real time (or wall clock time) for completing the first conformational cycle. Another key for obtaining the Prp43 cycle with acceptable computational costs was to augment AS simulations with ST. In line with the enhanced sampling of conformational transitions of small domains [111], we obtained significantly improved sampling of the enzyme dynamics (Fig. 5.2), which rendered the simulations feasible. We expect that the combination of AS with ST will be useful for a wide range of future enzyme simulations.

Impact of G-patch on helicase dynamics

Activator or cofactor proteins are needed to enhance and regulate the activity of helicases. Different cofactor proteins carry out different functions for the corresponding enzymes which they are attached to [79, 81, 82]. For example, DEAH-box helicases have a diverse group of cofactor proteins, the so-called G-patch proteins [83]. The G-patch proteins are mostly intrinsically disordered proteins (IDPs), which are connected to the outer surface of a helicase. Here, they are believed to influence the conformational dynamics of a helicase, which enhances or regulates its activity in specific situations.

In this study, we investigated the DHX15 apo-structure resolved by Studer et al. [84] (PDB ID 6SH7). To get a better understanding of the influence of the NKRF G-patch on the DHX15 helicase structure, we performed ten 1µs ST simulations of the native crystal structure with G-patch attached and ten 1µs ST simulations with the NKRF G-patch removed from the crystal structure. We have chosen the apo-structure of the DHX15 helicase, because the apo structure is the most flexible and therefore has the lowest timescales on domain transitions of all DEAH-box helicase complexes. The reason behind this might be the missing ligands which usually provide stability to the structures. Thus, the chosen structure is a suitable test system to study the influence of the G-patch on the domain movements. The following sections are based on our published work about the influence of the various ligands on the dynamics of helicases [168].

7.1 Methods

MD simulations of the simulations concerning the investigation of the G-patch structure (6SH7) were set up as follows. First, we used Modeller to fix the missing residues of the 6SH7 structure from the PDB database. For the comparable apo-structure system the G-patch was removed from the crystal structure. In GROMACS 2020.2 [169], the structures were placed into a simulation box of a dodecahedron. The energy of the systems was minimized with the steepest descent algorithm. Then, the systems were equilibrated for 100 ps with position restraints acting on the heavy atoms including present ligands ($k = 1000\,\mathrm{kJ\,mol^{-1}nm^{-2}}$). Water was modeled with the TIP3P model [174], and parameters for K^+ were taken from [175]. The energy of the system was minimized with the steepest descent algorithm. Then, the system was equilibrated for 100 ps with position restraints acting on the heavy atoms including RNA and Mg ($k = 1000\,\mathrm{kJ\,mol^{-1}nm^{-2}}$).

Electrostatic interactions were described with the particle-mesh Ewald method [138]. Dispersion interactions and short-range repulsion were described together with a Lennard-Jones potential with a cut-off at 1 nm. The temperature was controlled at 300 K using velocity-scaling [135], thereby coupling protein, RNA, Mg^{2+}, and ATP (if present) to one heat bath while coupling water and K^+ to a second heat bath ($\tau = 0.5\,\mathrm{ps}$). The pressure was controlled at 1 bar with the Parrinello-Rahman barostat ($\tau = 5\,\mathrm{ps}$) [176]. The md-vv integrator was used for simulated tempering (ST) simulations and the md integrator was used for all other simulations, both with an integration time step of 2 fs. The geometry of water molecules was constrained with SETTLE [134]. All other bonds were constrained with P-LINCS [133]. For the ST parameters, we used a temperature ladder from 300 K to 348 K with a temperature difference of 2 K between the states, resulting in 24 different temperatures states. The weights were determined using the SA approach mentioned above. With this setup, we performed ten 1µs ST simulations of the apo-structure of DHX15 and ten 1µs ST simulations of the DHX15 with attached G-patch.

7.2 Results

The simulations of the two different systems showed a significant difference in the domain dynamics. We first plotted the COM distance of the two RecA domains over time as shown in Fig. 7.1A and Fig. 7.1B. They show the time evolution of the RecA domains of DHX15 with G-patch and without G-patch, respectively. At first glance, the time evolution of these distances looks rather chaotic. But, taking a look on the corresponding histograms, two distinct states are observed. One of the states in the protein–G-patch complex represents the native closed state of the crystal structure with a RecA distance of around 2.8-2.9 nm. The other state has a RecA1–RecA2 distance of 3.15-3.25 nm. The data suggests, the enzyme with an attached G-patch prefers both states (Fig. 7.1D), while the other system behaves rather random and diffusive as seen in Fig. 7.1C. The G-patch might stabilize the two captured states by restricting the overall flexibility of the RecA domains. This hypothesis is also proposed by Studer et al. [84] during the analysis of crystal structures. They claim that NKRF acts as a brace on the domains of the enzyme, which ultimately leads to less extreme open and closed states, which results in a higher ATPase activity. Conclusively, for the G-patch bound structure, at any given time, there are more proteins in a conformation with higher ATP affinity than in the more flexible non-G-patch conformations.

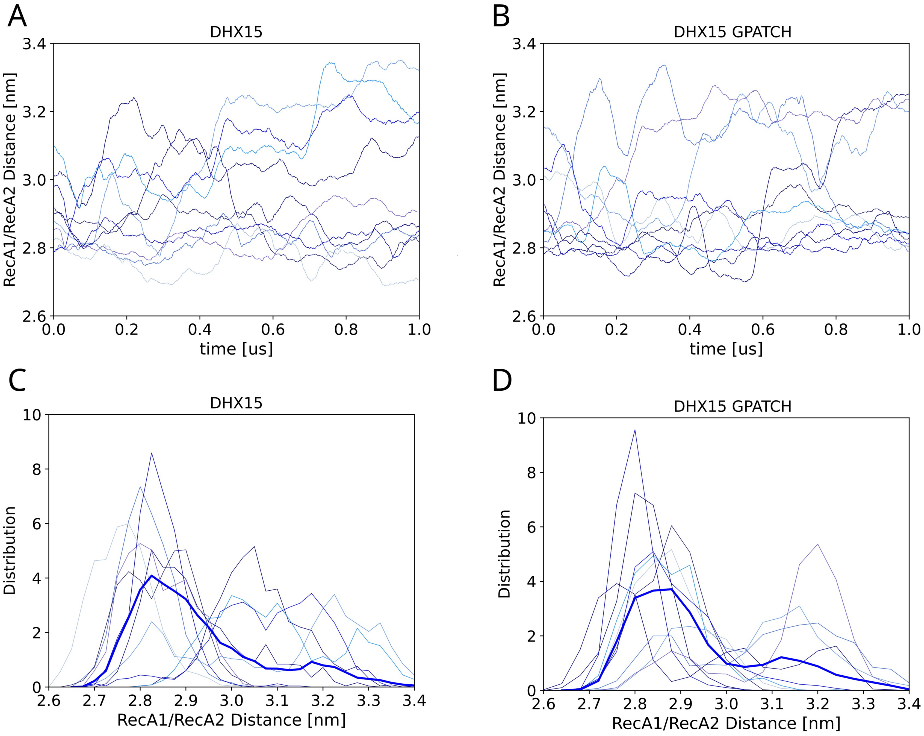


Figure 7.1: *RecA domain distance of DHX15 and DHX15·G-patch complex. (A) Time evolution of the RecA distance of the DHX15 apo-structure. (B) Time evolution of the RecA distance of the DHX15·G-patch complex. (C) Distance distribution of the RecA distances for DHX15 apo-structure. (D) Distance distribution of the RecA distances for DHX15·G-patch complex.*

Additionally, we investigated the influence of the G-patch on the CTD domain. Here, we measured the time evolution of the distance between GLU497 in the ratchet-like domain and two positively charged LYS259 and ARG263 in an outer helix of the RecA2 domain. The helix and the GLU497 are important interaction partner between CTD and RecA2. As shown in Fig.

7.2A/B, the domains show a similar trend like the RecA domains, because in the G-patch bound structure, two distinct states are taken in, while in the non-G-Patch structure a rather diffuse behaviour of the domains is observed. The first state at around 0.2 nm is an H-bond between GLU497 and LYS259 or between GLU497 and ARG263 as shown in Fig. 7.2C. In contrast, the DHX15 apo-structure ends up in a state which represents a disconnection between CTD and RecA2 domain as shown in Fig. 7.2D. The second distinct state is a conformation with a different contact point between RecA2 and ratchet-like domain, e.g., the residue pairs LYS259/TYR537 and ARG544/ASP255 form an stable H-bonds which leads to a long living conformation.

We can further distinguish the G-patch and the apo-structure by comparing the opening of the RNA tunnel formed by the RecA2 and CTD domain. We plot the time evolution of distances between the residues ARG640 and PRO495 in Fig. 7.3A-B. The distance between the two residues describes the nearness of the ratchet-like domain to the inner RNA tunnel and therefor estimates a the tightness of the RNA tunnel. For clarity, the two residues, at specific times during the DHX15 apo-structure and the DHX15·G-patch complex simulations, are shown in Fig. 7.3C-D, respectively. Figs. 7.3A/B show a clear difference between the G-patch structure and the apo-structure. In case of the apo-structure the two residues approach each other right at the beginning of the simulations, which indicates a tight RNA tunnel. In contrast, the G-patch structure follows a more constant evolution of the distance between the residues, which resembles a rather stable RNA tunnel.

7.3 Discussion

Overall, the protein–G-patch complex shows to be more rigid than the apo-protein, because the ensemble adopted by the G-patch-protein complex shows two distinct states while the apo-protein has more diffusive and flexible domain movements. These observations are in line with the results from the structural study of Studer et al. [84] and from the FRET study of the Prp43-pfa1 complex by Ficner et al. [204]. In the latter study, Ficner et al. observed a distinct open state in addition to the closed state in the G-patch bound structure of Prp43. In contrast, the open state is not seen in the apo-structure of Prp43 during the FRET experiments. But, the results of our study are a strong indication that the G-patch modulates the enzymes kinetics in a way that leads to the stabilization of a distinct closed state alongside the open state. This closed state could have a higher ATP affinity than the open state, which would explain the higher ATPase activity of the G-patch bound

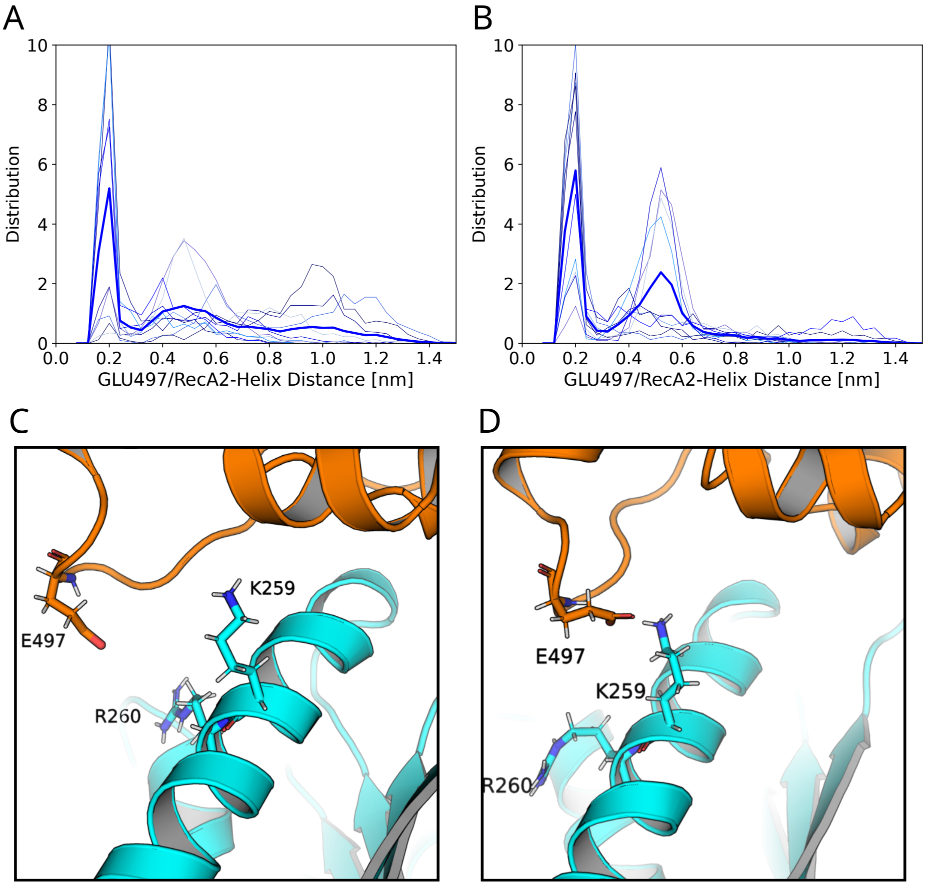


Figure 7.2: *GLU497/RecA2 distance distribution of the DHX15 apo-structure (A) and of the DHX15·G-patch complex (B). (C) Specific chosen snapshot of a DHX15 apo-structure simulation representing a large distance between the RecA2 and the outer residue E497 of the ratchet-like domain. (D) Specific chosen snapshot of a DHX15·G-patch simulation representing a connection between the RecA2 and the outer residue E497 of the ratchet-like domain.*

complex in contrast to the non-bound complex.

Additional simulations and experimental studies are required to fully resolve the effect of G-patch proteins on the conformational dynamics of DEAH-box helicases, especially during ATP binding.

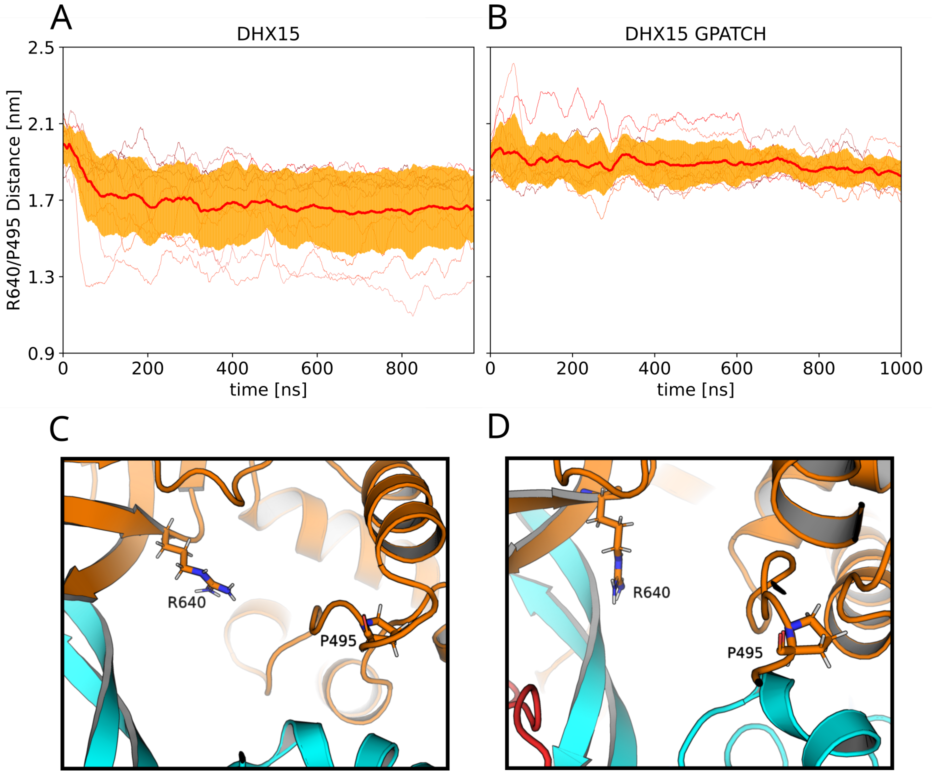


Figure 7.3: *(A) Time evolution of R640/P495 distance of the DHX15 apo-structure simulations. The average over all simulations is shown as red line with the standard deviation represented by the orange area. (B) Time evolution of R640/P495 distance of the DHX15·G-patch simulations. The average over all simulations is shown as red line with the standard deviation represented by the orange area. (C) Specific chosen snapshot of a DHX15 apo-structure simulation representing a low distance between R640 and P495. (D) Specific chosen snapshot of a DHX15 apo-structure simulation representing a large distance between R640 and P495.*

Phosphate exit tunnel of Prp22 and Prp43

Hydrolysis and product release are processes which occur on long timescales and therefore are often the time limiting steps during a complete mechanism. Thus, it is of great importance to gain more knowledge about the exit paths a product might take when leaving the enzyme. Crystallographic structures of Prp22, resolved by Hamann et al., already revealed a possible exit pathway on the back of the protein. To substantiate the hypothesis of the existence of the exit path way, we carried out RAMD simulations of the phosphate which is produced after the hydrolysis of ATP. The following results have been published in a joint experimental and computational study in Ref. [177].

8.1 Methods

MD simulations of the phosphate release from the ATP-binding pocket were set up as follow. In the simulations, the Prp2 and the Prp43 structure were used, both representing the complex of the respective enzyme with U_7 RNA and the ATP analog $\mathrm{ADP} - \mathrm{BeF}_3^-$. The ATP analog was replaced with ADP and dihydrogen phosphate (DHP). The structure was placed into a simulation box with the shape of a dodecahedron. The box was solvated with 24 798 water molecules for Prp2 and 36 616 water molecules for Prp43. Each system was then neutralized by 13 potassium counter-ions. Interactions of protein and RNA were described with the Amber14SB force field ([132]). Water was modeled with the TIP3P model [174]. Parameters for ADP and DHP were taken from Meagher et al. [172] and Kashefolgheta & Vila Verde [205], respectively. The parameters were translated into the GROMACS format via the *ACPYPE* software [206]. The parameters for K^+ ions were taken from Joung & Cheathma [175]. For the interactions between Mg^{2+} and DHP:O (the negative-charged O atom of DHP), the combination rule for Lennard–Jones interactions was overwritten with the nonbonded interactions suggested by Panteva et al. [207]. The energy of the system was minimized with the steepest-descent algorithm. The system was then equilibrated for 100 ps with positional restraints acting on the heavy atoms, including RNA, ADP and DHP ($k = 1000\,\text{kJmol}^{-1}\text{nm}^{-1}$). Electrostatic interactions were described with the particle mesh Ewald. Dispersion interactions and short-range repulsion were described together using a Lennard–Jones potential with a cutoff at 1 nm. The temperature was controlled at 300 K using velocity scaling [135] by coupling protein/RNA/ADP/DHP and water/K^+ to two separate heat baths ($\tau = 0.5ps$). The pressure was controlled at 1 bar with the Parrinello–Rahman barostat ($\tau = 5ps$ [176]). An integration time step of 2 fs was used. The geometry of water molecules was constrained with *SETTLE* [134], while all other bonds were constrained with *P-LINCS* [133]. To accelerate the dissociation of DHP from the complex, we used random-accelerated MD simulations (RAMD; [145]. The *GROMACS* code (version 2020.1) extended for RAMD was taken from https://github.com/HITS-MCM/gromacs-ramd [208]. For the simulations reported in this study, we used the following RAMD settings. Mg^{2+} and DPH were considered as the receptor and the ligand, respectively. An accelerating force of 585.2 $\text{kJmol}^{-1}\text{nm}^{-1}$ was used, and simulations were evaluated every 50 steps. A different random seed was used for each simulation. If the ligand had traveled less than 0.005 nm within 50 steps, the direction of force was changed. The simulation stopped at a ligand–receptor distance of 4 nm. For Prp2 30 RAMD simulations were performed with these parameters, and 15 RAMD simulations were carried out for Prp43. In addition, we tested 5 RAMD simulations with a force of 635 $\text{kJmol}^{-1}\text{nm}^{-1}$ and 5 RAMD simulations with a force of 700 $\text{kJmol}^{-1}\text{nm}^{-1}$ for Prp43. Notably, we tested various alternative RAMD settings; in the case of successful dissociation events, these simulations revealed similar DHP-exit pathways.

8.2 Results

We used the RAMD method to proof the existence of a γ-phosphate exit tunnel on the back of the enzymes Prp2 and Prp43. The usage of RAMD was necessary because dissociation events usually occur on large time scales. In RAMD an additional force in a random direction is applied to a defined ligand. This force is updated in a different direction if the ligand cannot travel a certain distance in a specific time interval. This is done to avoid the unproductive movement against a dead end.

Among 30 RAMD simulations that successfully led to DHP dissociation

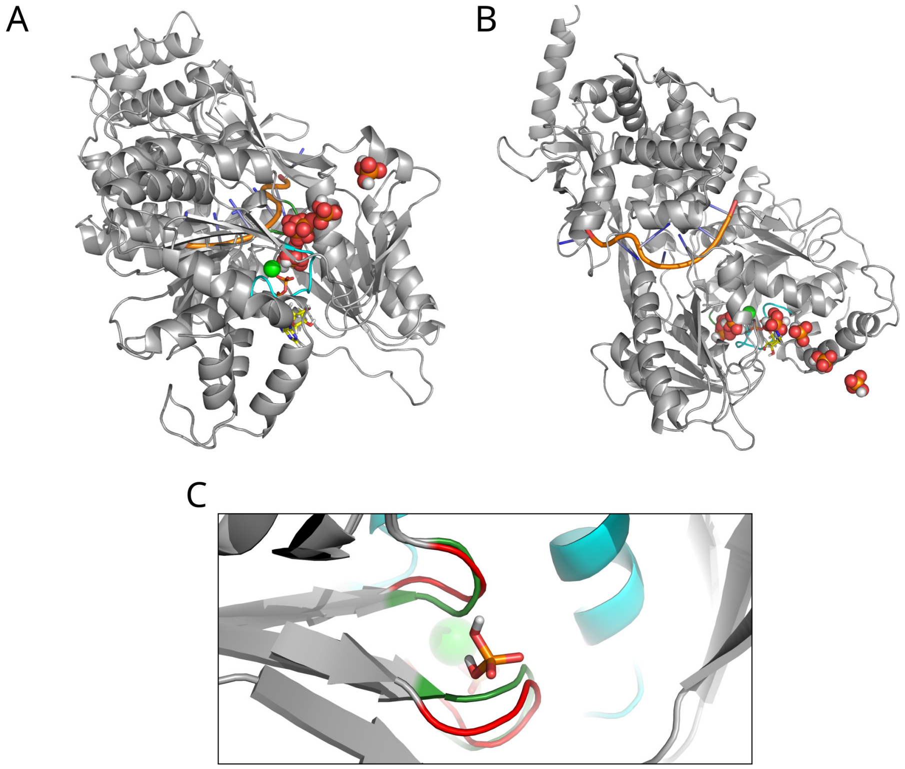


Figure 8.1: *(A) Exit path of the DHP on the back of the enzyme. This path has shown to be the predominant exit tunnel in our simulations. (B) The alternative exit path on the front of the enzyme. Here, the DHP has to find a way to pass the negative charged moiety of the ADP molecule. (C) The small change in the loop location of the motifs I and III are sufficient to enable an exit of DHP.*

from Prp2, 24 of them resulted in exit through the pathway between motifs I and III (Fig. 8.1A). Only 4 out of 30 simulations showed an exit through the ATP-binding site in the opposite direction (Fig. 8.1B). Similar results were observed in Prp43, with DHP exiting through the proposed channel in 18 of 25 simulations and an alternative pathway being present in 4 simulations. These findings suggest that the pathway between motifs I and III has the lowest energy barrier, while alternative pathways would require larger structural rearrangements with higher energetic costs. Our visual inspection of the trajectories showed that minor fluctuations of motifs I and III were sufficient for DHP dissociation (Fig. 8.1C). Our simulations provide strong evidence that the pathway between motifs I and III is the predominant exit pathway for DHP in Prp2 and Prp43 and may be conserved in other members of the family.

Summary and Conclusion

Helicases perform translocation and unwinding mechanisms to control certain key points in the splicing pathway. Therefore, they are crucial for the stability of our genome and conclusively the protein synthesis. One family of RNA helicases are the so-called DEAH-box helicases, named after a highly conserved motif in their structural core. Three prominent members of the DEAH-box family are the Prp43/DHX15, Prp22 and Prp2. Their main function is the proof reading and discarding process during splicing. Structural studies suggested that these DEAH-box helicases translocate ssRNA during the execution of their function. In the last ten years, RNA helicases became a highly investigated field of research, since a variety of different structures were resolved during the last decade. Thus, new hypotheses about their mechanism were proposed actively. Especially in the last few years, studies of the influence of G-patches onto the structure and function of helicases have raised considerable attention. Despite the past efforts, a detailed understanding of the mechanism of such a complex motor enzyme has not yet been obtained.

Simulating complex conformational cycles in atomic detail has been a long-standing goal of computational biophysics. This study provides detailed structural, kinetic, and energetic insights of the function of DEAH-box helicases. These results are made possible by overcoming the sampling problem of our system with the combination of two enhanced sampling techniques. Thereby, we achieved the simulation of a transition pathway which leads to a complete and continuous translocation cycle with a shift of the H-bonds between RNA and protein by one nucleotide upstream in atomistic detail. This complete cycle provided important insights into the dynamics and kinetics of the individual domains and smaller features/residues. Thereby, the thesis is offering a refined inchworm mechanism as an alternative to the classical inchworm model. Our work highlights the significance of small molecular switches in driving large-scale domain dynamics. Hence, the sampling problem arises from the fact that multiple atomic features, rather than large-scale domain motions, are the rate-limiting steps for the overall process. A typical example of such a molecular switch is the loop-to-helix (or helix-to-loop) transition of the sensor serine of RecA2 (Chapter 6).

The success in accomplishing this goal was not without challenges. Conventional pulling simulations along simple reaction coordinates - such as domain-domain center-of-mass pulling - proved insufficient to induce productive conformational changes. As mentioned, we introduced a novel combination of Simulated Tempering (ST) and Adaptive Sampling (AS) techniques as a solution to the sampling problem. This innovative approach yielded a complete conformational cycle of the motor protein Prp43, providing atomic-level details. We anticipate that the ST and AS combination will prove valuable for a wide range of future MD studies aiming to uncover protein conformational transitions, especially when the transitions are too complex for reaction coordinates and appear on very long timescales.

Apart from the conformational cycle of Prp43, we studied several additional aspects of the helicase dynamics and functions. For example, we showed the importance of RNA in the maintenance of the structural integrity of the whole enzyme, the responsibility of the G-patch for the modulation of the DEAH-helicases ATPase activity and the existence of an exit tunnel for phosphate after the ATP hydrolysis.

More precisely, we found that removing ATP from the complex or replacing it with ADP yielded limited changes. However, the removal of RNA from the complex triggered notable movements in the CTD, leading to the collapse of the original RNA tunnel. Remarkably, if ATP was kept in the complex during RNA removal, it had no large impact on the relative positions of the RecA domains. On the other hand, in apo structure simulations, substantial movements were observed in both the RNA cleft and the ATP core, resulting in conformational changes. These observations suggested that RNA played a crucial role in stabilizing all three domains, as any alteration in the RNA tunnel led to major movements. This observation aligned with findings from crystal structures, where RNA defined the RNA cleft and created a defined and rigid protein structure.

For the study on the influence of G-patch on the helicases, we compared the behavior of a DHX15-G-patch complex to the behavior of an apo-structure. Here, we removed the G-patch from the complex. Then, we ran ten 1 µs simulations of the G-patch bound native structure and ten 1 µs simulations of the apo-structure. The results indicate that the G-patch influences the rigidity and fine tunes the conformational movements such that distinct states are taken in. The apo-structure instead behaves more diffusive and flexible. These results are in line with recent studies like the structural study of Studer et al. itself and the FRET analysis of the Prp43-pfa1 performed by Ficner et al. [84, 204]

While the thesis successfully captures a full translocation cycle by one nucleotide and provides additional details on the influence of atomic-scale switches and ligands, some critical steps of the helicase mechanism remain unobserved. Future research should focus on investigating the open RNA tunnel during the RNA-loading step, considering changes in the mechanism due to different RNA compositions, and determining the polarity of helicases. Additionally, further exploration of the influence of G-patches on domain movements and the roles of different G-patches in the spliceosome's mechanism are essential steps toward unraveling the complexity of DEAH helicases in splicing. Here, more structures from experimental studies, especially RNA and/or G-patch bound conformations, are needed.

Further studies on the modulation of helicases, e.g., by investigating more cofactors or mutations of crucial features, may shed light into the emergence of related diseases and disorders. The present study provides information, which could be used in cell line studies by mutating specific features of the helicases. The provided insights could also be used for a high-throughput screening to find potential compounds which might modulate the helicase dynamics.

An additional long-term goal is the study of the spliceosome. Since there are structures of helicases bound to the larger spliceosome complex, it would be a milestone to investigate the impact of the found dynamics on the whole spliceosome complex. Of course, for an observation with computational methods, this would require even more sophisticated techniques or much stronger hardware. However, the possible insights could yield crucial information about the function of the spliceosome in general.

Overall, this work lays the foundation for a deeper understanding of these molecular machines and their crucial roles in RNA translocation and in the spliceosome in general.

Bibliography

[1] Francis Crick. Split genes and RNA splicing. *Science*, 204(4390):264–271, 1979.

[2] Francis H Crick. On protein synthesis. In *Symp Soc Exp Biol*, volume 12, page 8, 1958.

[3] Edmond M Chan, Tsukasa Shibue, James M McFarland, Benjamin Gaeta, Mahmoud Ghandi, Nancy Dumont, Alfredo Gonzalez, Justine S McPartlan, Tianxia Li, Yanxi Zhang, et al. WRN helicase is a synthetic lethal target in microsatellite unstable cancers. *Nature*, 568(7753):551–556, 2019.

[4] Marina K Kukhanova, Inna L Karpenko, and Alexander V Ivanov. DEAD-box RNA helicase DDX3: Functional properties and development of DDX3 inhibitors as antiviral and anticancer drugs. *Molecules*, 25(4):1015, 2020.

[5] Mahmoud Abdel-Monem and Hartmut HOFFMANN-BERLING. Enzymic unwinding of DNA: 1. Purification and characterization of a DNA-Dependent ATPase from escherichia coli. *Eur. J. Biochem.*, 65(2):431–440, 1976.

[6] Mahmoud ABDEL-MONEM, Hildegard DÜRWALD, and Hartmut HOFFMANN-BERLING. Enzymic unwinding of DNA: 2. Chain separation by an ATP-Dependent DNA unwinding enzyme. *Eur. J. Biochem.*, 65(2):441–449, 1976.

[7] Vivian Mackay and Stuart Linn. Selective inhibition of the dnase activity of the recBC enzyme by the DNA binding protein from Escherichia coli. *J. Biol. Chem.*, 251(12):3716–3719, 1976.

[8] B Kuhn, M Abdel-Monem, H Krell, and H Hoffmann-Berling. Evidence for two mechanisms for DNA unwinding catalyzed by DNA helicases. *J. Biol. Chem.*, 254(22):11343–11350, 1979.

[9] Bernd Kuhn, Mahmoud Abdel-Monem, and Hartmut Hoffmann-Berling. DNA helicases. In *Cold Spring Harb. Symp. Quant. Biol.*, volume 43, pages 63–67. Cold Spring Harbor Laboratory Press, 1979.

[10] Yasuo Hotta and Herbert Stern. DNA unwinding protein from meiotic cells of Lilium. *Biochemistry*, 17(10):1872–1880, 1978.

[11] M Venkatesan, LL Silver, and NG Nossal. Bacteriophage T4 gene 41 protein, required for the synthesis of RNA primers, is also a DNA helicase. *J. Biol. Chem.*, 257(20):12426–12434, 1982.

[12] Ulrich Hübscher and Hans-Peter Stalder. Mammalian DNA helicase. *Nucleic Acids Res.*, 13(15):5471–5483, 1985.

[13] A Sugino, B Ho Ryu, T Sugino, L Naumovski, and EC Friedberg. A new DNA-dependent ATPase which stimulates yeast DNA polymerase I and has DNA-unwinding activity. *J. Biol. Chem.*, 261(25):11744–11750, 1986.

[14] Narendra Tuteja, Renu Tuteja, Khalilur Rahman, Li-Ya Kang, and Arturo Falaschi. A DNA helicase from human cells. *Nucleic Acids Res.*, 18(23):6785–6792, 1990.

[15] Narendra Tuteja, Tuan-Nghia Phan, and Krishna K Tewari. Purification and characterization of a DNA helicase from pea chloroplast that translocates in the 3-to-5 direction. *Eur. J. Biochem.*, 238(1):54–63, 1996.

[16] JA Grifo, RD Abramson, CA Satler, and WC Merrick. RNA-stimulated ATPase activity of eukaryotic initiation factors. *J. Biol. Chem.*, 259(13):8648–8654, 1984.

[17] Bimal K Ray, T Glen Lawson, JC Kramer, MH Cladaras, JA Grifo, RD Abramson, WC Merrick, and Robert E Thach. ATP-dependent unwinding of messenger RNA structure by eukaryotic initiation factors. *J. Biol. Chem.*, 260(12):7651–7658, 1985.

[18] Patrick Linder and Frances V Fuller-Pace. Looking back on the birth of DEAD-box RNA helicases. *Biochim. Biophys. Acta BBA-Gene Regul. Mech.*, 1829(8):750–755, 2013.

[19] Robert M Brosh Jr and Steven W Matson. History of DNA helicases. *Genes*, 11(3):255, 2020.

[20] Margaret E Fairman-Williams, Ulf-Peter Guenther, and Eckhard Jankowsky. SF1 and SF2 helicases: Family matters. *Curr. Opin. Struct. Biol.*, 20(3):313–324, 2010.

[21] Pavan Umate, Narendra Tuteja, and Renu Tuteja. Genome-wide comprehensive analysis of human helicases. *Commun. Integr. Biol.*, 4(1):118–137, 2011.

[22] N Kyle Tanner and Patrick Linder. DExD/H box RNA helicases: From generic motors to specific dissociation functions. *Mol. Cell*, 8(2):251–262, 2001.

[23] Alexander E Gorbalenya, Eugene V Koonin, Alexei P Donchenko, and Vladimir M Blinov. A novel superfamily of nucleoside triphosphate-binding motif containing proteins which are probably involved in duplex unwinding in DNA and RNA replication and recombination. *FEBS Lett.*, 235(1-2):16–24, 1988.

[24] Martin R Singleton, Mark S Dillingham, and Dale B Wigley. Structure and mechanism of helicases and nucleic acid translocases. *Annu. Rev. Biochem.*, 76:23–50, 2007.

[25] Alexander E Gorbalenya and Eugene V Koonin. Helicases: Amino acid sequence comparisons and structure-function relationships. *Curr. Opin. Struct. Biol.*, 3(3):419–429, 1993.

[26] Frances V Fuller-Pace. DExD/H box RNA helicases: Multifunctional proteins with important roles in transcriptional regulation. *Nucleic Acids Res.*, 34(15):4206–4215, 2006.

[27] Alicia K Byrd, Dennis L Matlock, Debjani Bagchi, Suja Aarattuthodiyil, David Harrison, Vincent Croquette, and Kevin D Raney. Dda helicase tightly couples translocation on single-stranded DNA to unwinding of duplex DNA: Dda is an optimally active helicase. *J. Mol. Biol.*, 420(3):141–154, 2012.

[28] Martin R Singleton, Mark S Dillingham, and Dale B Wigley. Structure and mechanism of helicases and nucleic acid translocases. *Annu. Rev. Biochem.*, 76:23–50, 2007.

[29] Marcel J Tauchert, Jean-Baptiste Fourmann, Reinhard Lührmann, and Ralf Ficner. Structural insights into the mechanism of the DEAH-box RNA helicase Prp43. *Elife*, 6:e21510, 2017.

[30] Florian Hamann, Marieke Enders, and Ralf Ficner. Structural basis for RNA translocation by DEAH-box ATPases. *Nucleic Acids Res.*, 47(8):4349–4362, May 2019.

[31] Peter H von Hippel and Emmanuelle Delagoutte. Macromolecular complexes that unwind nucleic acids. *Bioessays*, 25(12):1168–1177, 2003.

[32] Yong-Joo Jeong, Mikhail K Levin, and Smita S Patel. The DNA-unwinding mechanism of the ring helicase of bacteriophage T7. *Proc. Natl. Acad. Sci.*, 101(19):7264–7269, 2004.

[33] Mikhail K Levin, Madhura Gurjar, and Smita S Patel. A Brownian motor mechanism of translocation and strand separation by hepatitis C virus helicase. *Nat. Struct. Mol. Biol.*, 12(5):429–435, 2005.

[34] Mikhail K Levin, Madhura M Gurjar, and Smita S Patel. ATP binding modulates the nucleic acid affinity of hepatitis C virus helicase. *J. Biol. Chem.*, 278(26):23311–23316, 2003.

[35] Narendra Tuteja and Renu Tuteja. Unraveling DNA helicases: Motif, structure, mechanism and function. *Eur. J. Biochem.*, 271(10):1849–1863, 2004.

[36] Inga Jarmoskaite and Rick Russell. RNA helicase proteins as chaperones and remodelers. *Annu. Rev. Biochem.*, 83:697–725, 2014.

[37] Timothy M Lohman, Eric J Tomko, and Colin G Wu. Non-hexameric DNA helicases and translocases: Mechanisms and regulation. *Nat. Rev. Mol. Cell Biol.*, 9(5):391–401, 2008.

[38] Anna Marie Pyle. Translocation and unwinding mechanisms of RNA and DNA helicases. *Annu. Rev. Biophys.*, 37:317–336, 2008.

[39] Pilar Tijerina, Hari Bhaskaran, and Rick Russell. Nonspecific binding to structured RNA and preferential unwinding of an exposed helix by the CYT-19 protein, a DEAD-box RNA chaperone. *Proc. Natl. Acad. Sci.*, 103(45):16698–16703, 2006.

[40] Thierry Bizebard, Ilaria Ferlenghi, Isabelle Iost, and Marc Dreyfus. Studies on three E. coli DEAD-box helicases point to an unwinding mechanism different from that of model DNA helicases. *Biochemistry*, 43(24):7857–7866, 2004.

[41] Quansheng Yang and Eckhard Jankowsky. The DEAD-box protein Ded1 unwinds RNA duplexes by a mode distinct from translocating helicases. *Nat. Struct. Mol. Biol.*, 13(11):981–986, 2006.

[42] Quansheng Yang, Mark Del Campo, Alan M Lambowitz, and Eckhard Jankowsky. DEAD-box proteins unwind duplexes by local strand separation. *Mol. Cell*, 28(2):253–263, 2007.

[43] Panos Soultanas, Mark S Dillingham, Paul Wiley, Martin R Webb, and Dale B Wigley. Uncoupling DNA translocation and helicase activity in PcrA: Direct evidence for an active mechanism. *EMBO J.*, 19(14):3799–3810, 2000.

[44] Ralf Seidel, Joost GP Bloom, Cees Dekker, and Mark D Szczelkun. Motor step size and ATP coupling efficiency of the dsDNA translocase EcoR124I. *EMBO J.*, 27(9):1388–1398, 2008.

[45] Sua Myong, Sheng Cui, Peter V Cornish, Axel Kirchhofer, Michaela U Gack, Jae U Jung, Karl-Peter Hopfner, and Taekjip Ha. Cytosolic viral sensor RIG-I is a 5'-Triphosphate–dependent translocase on double-stranded RNA. *Science*, 323(5917):1070–1074, 2009.

[46] H Durr, A Flaus, T Owen-Hughes, and K-P Hopfner. Snf2 family ATPases and DExx box helicases: Differences and unifying concepts from high-resolution crystal structures. *Nucleic Acids Res.*, 34(15):4160–4167, 2006.

[47] Ruixue Wan, Chuangye Yan, Rui Bai, Jianlin Lei, and Yigong Shi. Structure of an intron lariat spliceosome from Saccharomyces cerevisiae. *Cell*, 171(1):120–132, 2017.

[48] Jacques Ninio. Kinetic amplification of enzyme discrimination. *Biochimie*, 57(5):587–595, 1975.

[49] John J Hopfield. Kinetic proofreading: A new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Natl. Acad. Sci.*, 71(10):4135–4139, 1974.

[50] Daniel R Semlow and Jonathan P Staley. Staying on message: Ensuring fidelity in pre-mRNA splicing. *Trends Biochem. Sci.*, 37(7):263–273, 2012.

[51] Pavanapuresan P Vaidyanathan, Ishraq AlSadhan, Dawn K Merriman, Hashim M Al-Hashimi, and Daniel Herschlag. Pseudouridine and N6-methyladenosine modifications weaken PUF protein/RNA interactions. *RNA*, 23(5):611–618, 2017.

[52] Nicole M Martinez, Amanda Su, Margaret C Burns, Julia K Nussbacher, Cassandra Schaening, Shashank Sathe, Gene W Yeo, and Wendy V Gilbert. Pseudouridine synthases modify human pre-mRNA co-transcriptionally and affect pre-mRNA processing. *Mol. Cell*, 82(3):645–659, 2022.

[53] Prakash Koodathingal and Jonathan P Staley. Splicing fidelity: DEAD/H-box ATPases as molecular clocks. *RNA Biol.*, 10(7):1073–1079, 2013.

[54] Francesca De Bortoli, Sara Espinosa, and Rui Zhao. DEAH-box RNA helicases in pre-mRNA splicing. *Trends Biochem. Sci.*, 46(3):225–238, 2021.

[55] Jae Young Lee and Wei Yang. UvrD helicase unwinds DNA one base pair at a time by a two-part power stroke. *Cell*, 127(7):1349–1360, 2006.

[56] Piero R Bianco and Stephen C Kowalczykowski. Translocation step size and mechanism of the RecBC DNA helicase. *Nature*, 405(6784):368–372, 2000.

[57] Ahmet Yildiz, Michio Tomishige, Ronald D Vale, and Paul R Selvin. Kinesin walks hand-over-hand. *Science*, 303(5658):676–678, 2004.

[58] Sameer S Velankar, Panos Soultanas, Mark S Dillingham, Hosahalli S Subramanya, and Dale B Wigley. Crystal structures of complexes of PcrA DNA helicase with a DNA substrate indicate an inchworm mechanism. *Cell*, 97(1):75–84, 1999.

[59] Manuel Hilbert, Anne R Karow, and Dagmar Klostermeier. The mechanism of ATP-dependent RNA unwinding by DEAD box proteins. *Biol. Chem.*, 390:1237–1250, 2009.

[60] Patrick Linder and Eckhard Jankowsky. From unwinding to clamping—the DEAD box RNA helicase family. *Nat. Rev. Mol. Cell Biol.*, 12(8):505–516, 2011.

[61] John E Walker, Matti Saraste, Michael J Runswick, and Nicholas J Gay. Distantly related sequences in the alpha-and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.*, 1(8):945–951, 1982.

[62] Eckhard Jankowsky and Margaret E. Fairman. RNA helicases - one fold for many functions. *Curr. Opin. Struct. Biol.*, 17(3):316–324, 2007.

[63] A Pause and N Sonenberg. Mutational analysis of a DEAD box RNA helicase: The mammalian translation initiation factor eIF-4A. *EMBO J.*, 11(7):2643–2654, 1992.

[64] Beate Schwer and Tamar Meszaros. RNA helicase dynamics in pre-mRNA splicing. *EMBO J.*, 19(23):6582–6591, 2000.

[65] J Rajan Prabu, Marisa Müller, Andreas W Thomae, Steffen Schüssler, Fabien Bonneau, Peter B Becker, and Elena Conti. Structure of the RNA helicase MLE reveals the molecular mechanisms for uridine specificity and RNA-ATP coupling. *Mol. Cell*, 60(3):487–499, 2015.

[66] Yangzi He, Gregers R. Andersen, and Klaus H. Nielsen. Structural basis for the function of DEAH helicases. *EMBO Rep.*, 11(3):180–186, 2010.

[67] Hélène Walbott, Saïda Mouffok, Régine Capeyrou, Simon Lebaron, Odile Humbert, Herman Van Tilbeurgh, Yves Henry, and Nicolas Leulliot. Prp43p contains a processive helicase structural architecture with a specific regulatory domain. *EMBO J.*, 29(13):2194–2204, 2010.

[68] Mitsutoshi Yoneyama and Takashi Fujita. Structural mechanism of RNA recognition by the RIG-I-like receptors. *Immunity*, 29(2):178–181, 2008.

[69] Robert D Shereda, Nicholas J Reiter, Samuel E Butcher, and James L Keck. Identification of the SSB binding site on E. coli RecQ reveals a conserved surface for binding SSB's C terminus. *J. Mol. Biol.*, 386(3):612–625, 2009.

[70] Katharina Büttner, Sebastian Nehring, and Karl-Peter Hopfner. Structural basis for DNA duplex separation by a superfamily-2 helicase. *Nat. Struct. Mol. Biol.*, 14(7):647–652, 2007.

[71] Lenz Steimer and Dagmar Klostermeier. RNA helicases in infection and disease. *RNA Biol.*, 9(6):751–771, 2012.

[72] Eva Absmeier, Karine F Santos, and Markus C Wahl. Molecular mechanism underlying inhibition of intrinsic ATPase activity in a Ski2-like RNA helicase. *Structure*, 28(2):236–243, 2020.

[73] Manjeera Gowravaram, Fabien Bonneau, Joanne Kanaan, Vincent D Maciej, Francesca Fiorini, Saurabh Raj, Vincent Croquette, Hervé Le Hir, and Sutapa Chakrabarti. A conserved structural element in the RNA helicase UPF1 regulates its catalytic activity in an isoform-specific manner. *Nucleic Acids Res.*, 46(5):2648–2659, 2018.

[74] Chenlin Song, Agnes Hotz-Wagenblatt, Renate Voit, and Ingrid Grummt. SIRT7 and the DEAD-box helicase DDX21 cooperate to resolve genomic R loops and safeguard genome stability. *Genes Dev.*, 31(13):1370–1381, 2017.

[75] Rebecca Mathew, Klaus Hartmuth, Sina Möhlmann, Henning Urlaub, Ralf Ficner, and Reinhard Lührmann. Phosphorylation of human PRP28 by SRPK2 is required for integration of the U4/U6-U5 tri-snRNP into the spliceosome. *Nat. Struct. Mol. Biol.*, 15(5):435–443, 2008.

[76] A-MF Jacobs, SM Nicol, RG Hislop, EG Jaffray, RT Hay, and FV Fuller-Pace. SUMO modification of the DEAD box protein p68 modulates its transcriptional activity and promotes its interaction with HDAC1. *Oncogene*, 26(40):5866–5876, 2007.

[77] Jens Kretschmer, Harita Rao, Philipp Hackert, Katherine E Sloan, Claudia Höbartner, and Markus T Bohnsack. The m6A reader protein YTHDC2 interacts with the small ribosomal subunit and the 5'–3' exoribonuclease XRN1. *Rna*, 24(10):1339–1350, 2018.

[78] Magdalena Natalia Wojtas, Radha Raman Pandey, Mateusz Mendel, David Homolka, Ravi Sachidanandam, and Ramesh S Pillai. Regulation of m6A transcripts by the 3'-¿5' RNA helicase YTHDC2 is essential for a successful meiotic program in the mammalian germline. *Mol. Cell*, 68(2):374–387, 2017.

[79] Katherine E Sloan and Markus T Bohnsack. Unravelling the mechanisms of RNA helicase regulation. *Trends Biochem. Sci.*, 43(4):237–250, 2018.

[80] Katherine E Bohnsack, Ralf Ficner, Markus T Bohnsack, and Stefanie Jonas. Regulation of DEAH-box RNA helicases by G-patch proteins. *Biol. Chem.*, 402(5):561–579, 2021.

[81] Sevim Ozgur, Gretel Buchwald, Sebastian Falk, Sutapa Chakrabarti, Jesuraj Rajan Prabu, and Elena Conti. The conformational plasticity of eukaryotic RNA-dependent ATP ases. *FEBS J.*, 282(5):850–863, 2015.

[82] Edward Silverman, Gretchen Edwalds-Gilbert, and Ren-Jang Lin. DExD/H-box proteins and their partners: Helping RNA helicases unwind. *Gene*, 312:1–16, 2003.

[83] L Aravind and Eugene V Koonin. G-patch: A new conserved domain in eukaryotic RNA-processing proteins and type D retroviral polyproteins. *Trends Biochem. Sci.*, 24(9):342–344, 1999.

[84] Michael K Studer, Lazar Ivanović, Marco E Weber, Sabrina Marti, and Stefanie Jonas. Structural basis for DEAH-helicase activation by G-patch proteins. *Proc. Natl. Acad. Sci.*, 117(13):7159–7170, 2020.

[85] Aydan Salman-Dilgimen, Pierre-Olivier Hardy, Ashley R Dresser, and George Chaconas. HrpA, a DEAH-box RNA helicase, is involved in global gene regulation in the Lyme disease spirochete. *PloS one*, 6(7):e22168, 2011.

[86] John Panepinto, Lide Liu, Jeanie Ramos, Xudong Zhu, Tibor Valyi-Nagy, Saliha Eksi, Jianmin Fu, H Ari Jaffe, Brian Wickes, Peter R Williamson, et al. The DEAD-box RNA helicase Vad1 regulates multiple virulence-associated genes in Cryptococcus neoformans. *J. Clin. Invest.*, 115(3):632–641, 2005.

[87] Arnaz Ranji and Kathleen Boris-Lawrie. RNA helicases: Emerging roles in viral replication and the host innate response. *RNA Biol.*, 7(6):775–787, 2010.

[88] Nives Selak, Csanad Z Bachrati, Igor Shevelev, Tobias Dietschy, Barbara van Loon, Anette Jacob, Ulrich Hübscher, Joerg D Hoheisel, Ian D Hickson, and Igor Stagljar. The Bloom's syndrome helicase (BLM) interacts physically and functionally with p12, the smallest subunit of human DNA polymerase δ. *Nucleic Acids Res.*, 36(16):5166, 2008.

[89] Nathan A Ellis, Joanna Groden, Tian-Zhang Ye, Joel Straughen, David J Lennon, Susan Ciocci, Maria Proytcheva, and James German. The Bloom's syndrome gene product is homologous to RecQ helicases. *Cell*, 83(4):655–666, 1995.

[90] Matthew D Gray, Jiang-Cheng Shen, Ashwini S Kamath-Loeb, Anne Blank, Bryce L Sopher, George M Martin, Junko Oshima, and Lawrence A Loeb. The werner syndrome protein is a DNA helicase. *Nat. Genet.*, 17(1):100–103, 1997.

[91] Saori Kitao, Akira Shimamoto, Makoto Goto, Robert W Miller, William A Smithson, Noralane M Lindor, and Yasuhiro Furuichi. Mutations in RECQL4 cause a subset of cases of Rothmund-Thomson syndrome. *Nat. Genet.*, 22(1):82–84, 1999.

[92] Johannes N Spelbrink, Fang-Yuan Li, Valeria Tiranti, Kaisu Nikali, Qiu-Ping Yuan, Muhammed Tariq, Sjoerd Wanrooij, Nuria Garrido, Giacomo Comi, Lucia Morandi, et al. Human mitochondrial DNA deletions associated with mutations in the gene encoding Twinkle, a phage T7 gene 4-like protein localized in mitochondria. *Nat. Genet.*, 28(3):223–231, 2001.

[93] Sharon B Cantor, Daphne W Bell, Shridar Ganesan, Elizabeth M Kass, Ronny Drapkin, Steven Grossman, Doke CR Wahrer, Dennis C Sgroi, William S Lane, Daniel A Haber, et al. BACH1, a novel helicase-like protein, interacts directly with BRCA1 and contributes to its DNA repair function. *Cell*, 105(1):149–160, 2001.

[94] Anja J Van Brabant, Rodica Stan, and Nathan A Ellis. DNA helicases, genomic instability, and human genetic disease. *Annu. Rev. Genomics Hum. Genet.*, 1, 2000.

[95] Frances V Fuller-Pace. DEAD box RNA helicase functions in cancer. *RNA Biol.*, 10(1):121–132, 2013.

[96] Yigong Shi. A glimpse of structural biology through X-ray crystallography. *Cell*, 159(5):995–1014, 2014.

[97] MS Smyth and JHJ Martin. X ray crystallography. *Mol. Pathol.*, 53(1):8, 2000.

[98] Kurt Wüthrich. Protein structure determination in solution by NMR spectroscopy. *J. Biol. Chem.*, 265(36):22059–22062, 1990.

[99] Andrea Cavalli, Xavier Salvatella, Christopher M Dobson, and Michele Vendruscolo. Protein structure determination from NMR chemical shifts. *Proc. Natl. Acad. Sci.*, 104(23):9615–9620, 2007.

[100] Xiao-Chen Bai, Greg McMullan, and Sjors HW Scheres. How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci.*, 40(1):49–57, 2015.

[101] Ka Man Yip, Niels Fischer, Elham Paknia, Ashwin Chari, and Holger Stark. Atomic-resolution protein structure determination by cryo-EM. *Nature*, 587(7832):157–161, 2020.

[102] Sviatlana Shashkova and Mark C Leake. Single-molecule fluorescence microscopy review: Shedding new light on old problems. *Biosci. Rep.*, 37(4), 2017.

[103] Dorothee Liebschner, Pavel V Afonine, Matthew L Baker, Gábor Bunkóczi, Vincent B Chen, Tristan I Croll, Bradley Hintze, L-W Hung, Swati Jain, Airlie J McCoy, et al. Macromolecular structure determination using X-rays, neutrons and electrons: Recent developments in Phenix. *Acta Crystallogr. Sect. Struct. Biol.*, 75(10):861–877, 2019.

[104] Alexander D MacKerell Jr. Empirical force fields for biological macromolecules: Overview and issues. *J. Comput. Chem.*, 25(13):1584–1604, 2004.

[105] William L Jorgensen and Julian Tirado-Rives. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci.*, 102(19):6665–6670, 2005.

[106] Martin Karplus and J Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.*, 9(9):646–652, 2002.

[107] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. Metadynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 1(5):826–843, 2011.

[108] Danny Perez, Blas P Uberuaga, Yunsic Shim, Jacques G Amar, and Arthur F Voter. Accelerated molecular dynamics methods: Introduction and recent developments. *Annu. Rep. Comput. Chem.*, 5:79–98, 2009.

[109] E. Marinari and G. Parisi. Simulated tempering: A new monte carlo scheme. *Epl*, 19(6):451–458, 1992.

[110] Eugen Hruska, Jayvee R Abella, Feliks Nüske, Lydia E Kavraki, and Cecilia Clementi. Quantitative comparison of adaptive sampling methods for protein dynamics. *J. Chem. Phys.*, 149(24):244119, 2018.

[111] Albert C Pan, Thomas M Weinreich, Stefano Piana, and David E Shaw. Demonstrating an order-of-magnitude sampling enhancement in molecular dynamics simulations of complex protein systems. *J. Chem. Theory Comput.*, 12(3):1360–1367, 2016.

[112] S Doerr and G De Fabritiis. On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations. *J. Chem. Theory Comput.*, 10(5):2064–2069, 2014.

[113] Julien Robert-Paganin, Stéphane Réty, and Nicolas Leulliot. Regulation of DEAH/RHA helicases by G-patch proteins. *BioMed Res. Int.*, 2015, 2015.

[114] Florian Hamann, Andreas Schmitt, Filippo Favretto, Romina Hofele, Piotr Neumann, ShengQi Xiang, Henning Urlaub, Markus Zweckstetter, and Ralf Ficner. Structural analysis of the intrinsically disordered splicing factor Spp2 and its binding to the DEAH-box ATPase Prp2. *Proc. Natl. Acad. Sci.*, 117(6):2948–2956, 2020.

[115] Inessa De, Jana Schmitzová, and Vladimir Pena. The organization and contribution of helicases to RNA splicing. *Wiley Interdiscip. Rev. RNA*, 7(2):259–274, 2016.

[116] Yue Han and James Elliott. Molecular dynamics simulations of the elastic properties of polymer/carbon nanotube composites. *Comput. Mater. Sci.*, 39(2):315–323, 2007.

[117] Kiersten M Ruff and Rohit V Pappu. AlphaFold and implications for intrinsically disordered proteins. *J. Mol. Biol.*, 433(20):167208, 2021.

[118] Martin Karplus and Gregory A Petsko. Molecular dynamics simulations in biology. *Nature*, 347(6294):631–639, 1990.

[119] Wendy D Cornell, Piotr Cieplak, Christopher I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, David C Spellmeyer, Thomas Fox, James W Caldwell, and Peter A Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117(19):5179–5197, 1995.

[120] Paul K Weiner and Peter A Kollman. AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions. *J. Comput. Chem.*, 2(3):287–303, 1981.

[121] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25(9):1157–1174, 2004.

[122] Alex D MacKerell Jr, Donald Bashford, MLDR Bellott, Roland Leslie Dunbrack Jr, Jeffrey D Evanseck, Martin J Field, Stefan Fischer, Jiali Gao, H Guo, Sookhee Ha, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102(18):3586–3616, 1998.

[123] Alexander D MacKerell Jr, Joanna Wiorkiewicz-Kuczera, and Martin Karplus. An all-atom empirical energy function for the simulation of nucleic acids. *J. Am. Chem. Soc.*, 117(48):11946–11975, 1995.

[124] Kenno Vanommeslaeghe, Elizabeth Hatcher, Chayan Acharya, Sibsankar Kundu, Shijun Zhong, Jihyun Shim, Eva Darian, Olgun Guvench, P Lopes, Igor Vorobyov, et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.*, 31(4):671–690, 2010.

[125] Wilfred F van Gunsteren, SR Billeter, AA Eising, PH Hünenberger, PKHC Krüger, AE Mark, WRP Scott, and IG Tironi. Biomolecular simulation: The GROMOS96 manual and user guide. *Vdf Hochschulverlag AG ETH Zür. Zür.*, 86:1–1044, 1996.

[126] Chris Oostenbrink, Alessandra Villa, Alan E Mark, and Wilfred F Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.*, 25(13):1656–1676, 2004.

[127] Lukas D Schuler, Xavier Daura, and Wilfred F Van Gunsteren. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.*, 22(11):1205–1218, 2001.

[128] William L Jorgensen, David S Maxwell, and Julian Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118(45):11225–11236, 1996.

[129] William L Jorgensen and Julian Tirado-Rives. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, 110(6):1657–1666, 1988.

[130] Alexander D MacKerell Jr and Nilesh K Banavali. All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution. *J. Comput. Chem.*, 21(2):105–120, 2000.

[131] Siewert J Marrink, H Jelger Risselada, Serge Yefimov, D Peter Tieleman, and Alex H De Vries. The MARTINI force field: Coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, 111(27):7812–7824, 2007.

[132] 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. *J. Chem. Theory Comput.*, 11(8):3696–3713, 2015.

[133] Berk Hess. P-LINCS: A parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.*, 4(1):116–122, 2008.

[134] Shuichi Miyamoto and Peter A Kollman. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.*, 13(8):952–962, 1992.

[135] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

[136] Herman JC Berendsen, JPM van Postma, Wilfred F Van Gunsteren, ARHJ DiNola, and Jan R Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984.

[137] Andrea Rizzi, Travis Jensen, David R Slochower, Matteo Aldeghi, Vytautas Gapsys, Dimitris Ntekoumes, Stefano Bosisio, Michail Papadourakis, Niel M Henriksen, Bert L De Groot, et al. The SAMPL6 SAMPLing challenge: Assessing the reliability and efficiency of binding free energy calculations. *J. Comput. Aided Mol. Des.*, 34(5):601–633, 2020.

[138] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An N·log (N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[139] 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. *Journal of Chemical Theory and Computation*, 19(6):1898–1907, 2023.

[140] Sanghyun Park and Vijay S Pande. Choosing weights for simulated tempering. *Phys. Rev. E*, 76(1):016703, 2007.

[141] Gregory R Bowman, Daniel L Ensign, and Vijay S Pande. Enhanced modeling via network theory: Adaptive sampling of Markov state models. *J. Chem. Theory Comput.*, 6(3):787–794, 2010.

[142] Jeffrey K Weber and Vijay S Pande. Characterization and rapid sampling of protein folding Markov state model topologies. *J. Chem. Theory Comput.*, 7(10):3405–3411, 2011.

[143] Robin M Betz and Ron O Dror. How effectively can adaptive sampling methods capture spontaneous ligand binding? *J. Chem. Theory Comput.*, 15(3):2053–2063, 2019.

[144] Qixin Wang, Jamie J Arnold, Akira Uchida, Kevin D Raney, and Craig E Cameron. Phosphate release contributes to the rate-limiting step for unwinding by an RNA helicase. *Nucleic Acids Res.*, 38(4):1312–1324, 2010.

[145] Susanna K Lüdemann, Valère Lounnas, and Rebecca C Wade. How do substrates enter and products exit the buried active site of cytochrome P450cam? 1. Random expulsion molecular dynamics investigation of ligand access channels and mechanisms. *J. Mol. Biol.*, 303(5):797–811, 2000.

[146] Jérémie Mortier, Christin Rakers, Marcel Bermudez, Manuela S Murgueitio, Sereina Riniker, and Gerhard Wolber. The impact of molecular dynamics on drug design: Applications for the characterization of ligand–macromolecule complexes. *Drug Discov. Today*, 20(6):686–702, 2015.

[147] Ting Wang and Yong Duan. Ligand entry and exit pathways in the β2-adrenergic receptor. *J. Mol. Biol.*, 392(4):1102–1115, 2009.

[148] Mikael Peräkylä. Ligand unbinding pathways from the vitamin D receptor studied by molecular dynamics simulations. *Eur. Biophys. J.*, 38:185–198, 2009.

[149] Daria B Kokh, Marta Amaral, Joerg Bomke, Ulrich Grädler, Djordje Musil, Hans-Peter Buchstaller, Matthias K Dreyer, Matthias Frech, Maryse Lowinski, Francois Vallee, et al. Estimation of drug-target residence times by τ-random acceleration molecular dynamics simulations. *J. Chem. Theory Comput.*, 14(7):3859–3869, 2018.

[150] Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.*, 2(11):559–572, 1901.

[151] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, 24(6):417, 1933.

[152] Ian T Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, 374(2065):20150202, 2016.

[153] Manel A Balsera, Willy Wriggers, Yoshitsugu Oono, and Klaus Schulten. Principal component analysis and long time protein dynamics. *J. Phys. Chem.*, 100(7):2567–2572, 1996.

[154] Florian Sittel, Abhinav Jain, and Gerhard Stock. Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates. *J. Chem. Phys.*, 141(1):07B605_1, 2014.

[155] Yusuke Naritomi and Sotaro Fuchigami. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions. *J. Chem. Phys.*, 134(6):02B617, 2011.

[156] Christian R Schwantes and Vijay S Pande. Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9. *J. Chem. Theory Comput.*, 9(4):2000–2009, 2013.

[157] Gregory R Bowman, Vijay S Pande, and Frank Noé. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, volume 797. Springer Science & Business Media, 2013.

[158] Martin K Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Pérez-Hernández, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noé. PyEMMA 2: A software package for estimation, validation, and analysis of Markov models. *J. Chem. Theory Comput.*, 11(11):5525–5542, 2015.

[159] Stuart Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–137, 1982.

[160] Teofilo F Gonzalez. Clustering to minimize the maximum intercluster distance. *Theor. Comput. Sci.*, 38:293–306, 1985.

[161] Leonard Kaufman and Peter J Rousseeuw. Partitioning around medoids (program pam). *Find. Groups Data Introd. Clust. Anal.*, 344:68–125, 1990.

[162] Peter Deuflhard, Wilhelm Huisinga, Alexander Fischer, and Ch Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra Its Appl.*, 315(1-3):39–59, 2000.

[163] Ch Schütte, Alexander Fischer, Wilhelm Huisinga, and Peter Deuflhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys.*, 151(1):146–168, 1999.

[164] Peter Deuflhard and Marcus Weber. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra Its Appl.*, 398:161–184, 2005.

[165] Frank Noé, Illia Horenko, Christof Schütte, and Jeremy C Smith. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J. Chem. Phys.*, 126(15):04B617, 2007.

[166] James V Stone. Independent component analysis: An introduction. *Trends Cogn. Sci.*, 6(2):59–64, 2002.

[167] Lutz Molgedey and Heinz Georg Schuster. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, 72(23):3634, 1994.

[168] Robert A Becker and Jochen S Hub. Molecular simulations of DEAH-box helicases reveal control of domain flexibility by ligands: RNA, ATP, ADP, and G-patch proteins. *Biol. Chem.*, (0), 2023.

[169] Paul Bauer, Berk Hess, and Erik Lindahl. 2019.

[170] Marcel J Tauchert, J-B Fourmann, Henning Christian, Reinhard Lührmann, and Ralf Ficner. Structural and functional analysis of the RNA helicase Prp43 from the thermophilic eukaryote Chaetomium thermophilum. *Acta Crystallogr. Sect. F Struct. Biol. Commun.*, 72(2):112–120, 2016.

[171] Andrej Šali and Tom L Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, 234(3):779–815, 1993.

[172] Kristin L. Meagher, Luke T. Redman, and Heather A. Carlson. Development of polyphosphate parameters for use with the AMBER force field. *J. Comput. Chem.*, 24(9):1016–1025, 2003.

[173] Alan W Sousa Da Silva and Wim F Vranken. ACPYPE-Antechamber python parser interface. *BMC Res. Notes*, 5(1):1–8, 2012.

[174] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[175] In Suk Joung and Thomas E Cheatham III. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B*, 112(30):9020–9041, 2008.

[176] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52(12):7182–7190, 1981.

[177] Florian Hamann, Lars C Zimmerningkat, Robert A Becker, Tim B Garbers, Piotr Neumann, Jochen S Hub, and Ralf Ficner. The structure of Prp2 bound to RNA and ADP-BeF3- reveals structural features important for RNA unwinding by DEAH-box ATPases. *Acta Crystallogr. Sect. Struct. Biol.*, 77(4):496–509, 2021.

[178] Fugao Wang and David P Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.*, 86(10):2050, 2001.

[179] Gira Bhabha, Justin T Biel, and James S Fraser. Keep on moving: Discovering and perturbing the conformational dynamics of enzymes. *Acc. Chem. Res.*, 48(2):423–430, 2015.

[180] Robert A Becker and Jochen S Hub. Continuous millisecond conformational cycle of a DEAH box helicase reveals control of domain motions by atomic-scale transitions. *Commun. Biol.*, 6(1):379, 2023.

[181] William R. Pearson. Rapid and sensitive sequence comparison with FASTP and FASTA. In *Methods in Enzymology*, volume 183, pages 63–98. Elsevier, 1990.

[182] Nina Singhal and Vijay S Pande. Error analysis and efficient sampling in Markovian state models for molecular dynamics. *J. Chem. Phys.*, 123(20):204909, 2005.

[183] Martin K. Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Pérez-Hernández, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noé. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.*, 11:5525–5542, October 2015.

[184] Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.*, 139(1):07B604_1, 2013.

[185] Hugo Steinhaus et al. Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci*, 1(804):801, 1956.

[186] J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp Math Stat. Probab.*, pages 281–297, 1967.

[187] John D Chodera and Frank Noé. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.*, 25:135–144, 2014.

[188] Brooke E Husic and Vijay S Pande. Markov state models: From an art to a science. *J. Am. Chem. Soc.*, 140(7):2386–2396, 2018.

[189] Vijay S Pande, Kyle Beauchamp, and Gregory R Bowman. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods*, 52(1):99–105, 2010.

[190] Ziwei He, Fabian Paul, and Benoît Roux. A critical perspective on Markov state model treatments of protein–protein association using coarse-grained simulations. *J. Chem. Phys.*, 154(8):084101, 2021.

[191] Sean P Carney, Wen Ma, Kevin D Whitley, Haifeng Jia, Timothy M Lohman, Zaida Luthey-Schulten, and Yann R Chemla. Kinetic and structural mechanism for DNA unwinding by a non-hexameric helicase. *Nat. Commun.*, 12(1):1–14, 2021.

[192] Natalie M Stano, Yong-Joo Jeong, Ilker Donmez, Padmaja Tummalapalli, Mikhail K Levin, and Smita S Patel. DNA synthesis provides the driving force to accelerate DNA unwinding by a helicase. *Nature*, 435(7040):370–373, 2005.

[193] Henry R Bourne. The arginine finger strikes again. *Nature*, 389(6652):673–674, 1997.

[194] Daniel Mann, Christian Teuber, Stefan A Tennigkeit, Grit Schröter, Klaus Gerwert, and Carsten Kötting. Mechanism of the intrinsic arginine finger in heterotrimeric G proteins. *Proc. Natl. Acad. Sci.*, 113(50):E8041–E8050, 2016.

[195] Gergely N Nagy, Reynier Suardíaz, Anna Lopata, Olivér Ozohanics, Károly Vé´key, Bernard R Brooks, Ibolya Leveles, Judit Tóth, Beata G Vértessy, and Edina Rosta. Structural characterization of arginine fingers: Identification of an arginine finger for the pyrophosphatase dUTPases. *J. Am. Chem. Soc.*, 138(45):15035–15045, 2016.

[196] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D Chodera, Christof Schütte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.*, 134(17):174105, 2011.

[197] Marco Sarich, Frank Noé, and Christof Schütte. On the approximation quality of Markov state models. *Multiscale Model. Simul.*, 8(4):1154–1177, 2010.

[198] Xuhui Huang, Gregory R Bowman, and Vijay S Pande. Convergence of folding free energy landscapes via application of enhanced sampling methods in a distributed computing environment. *J. Chem. Phys.*, 128(20):05B622, 2008.

[199] Ken A Dill and Justin L MacCallum. The protein-folding problem, 50 years on. *science*, 338(6110):1042–1046, 2012.

[200] Rui Bai, Ruixue Wan, Chuangye Yan, Qi Jia, Jianlin Lei, and Yigong Shi. Mechanism of spliceosome remodeling by the ATPase/helicase Prp2 and its coactivator Spp2. *Science*, 371(6525):eabe8863, 2021.

[201] Andreas Schmitt, Florian Hamann, Piotr Neumann, and Ralf Ficner. Crystal structure of the spliceosomal DEAH-box ATPase Prp2. *Acta Crystallogr. Sect. Struct. Biol.*, 74(7):643–654, 2018.

[202] Roberto A Rodriguez, Lili Yu, and Liao Y Chen. Computing protein–protein association affinity with hybrid steered molecular dynamics. *J. Chem. Theory Comput.*, 11(9):4427–4438, 2015.

[203] Wen Ma and Klaus Schulten. Mechanism of substrate translocation by a ring-shaped ATPase motor at millisecond resolution. *J. Am. Chem. Soc.*, 137(8):3031–3040, 2015.

[204] Marieke Enders, Ralf Ficner, and Sarah Adio. Regulation of the DEAH/RHA helicase Prp43 by the G-patch factor Pfa1. *Proc. Natl. Acad. Sci.*, 119(48):e2203567119, 2022.

[205] Sadra Kashefolgheta and Ana Vila Verde. Developing force fields when experimental data is sparse: AMBER/GAFF-compatible parameters for inorganic and alkyl oxoanions. *Phys. Chem. Chem. Phys.*, 19(31):20593–20607, 2017.

[206] Alan W Sousa da Silva and Wim F Vranken. ACPYPE-Antechamber python parser interface. *BMC Res. Notes*, 5(1):1–8, 2012.

[207] Maria T Panteva, George M Giambasu, and Darrin M York. Force field for Mg2+, Mn2+, Zn2+, and Cd2+ ions that have balanced interactions with nucleic acids. *J. Phys. Chem. B*, 119(50):15460–15470, 2015.

[208] Daria B Kokh, Bernd Doser, Stefan Richter, Fabian Ormersbach, Xingyi Cheng, and Rebecca C Wade. A workflow for exploring ligand dissociation from a macromolecule: Efficient random acceleration molecular dynamics simulation and interaction fingerprint analysis of ligand trajectories. *J. Chem. Phys.*, 153(12):125102, 2020.