Exploring molecular recognition of small molecule binding, protein-protein interactions, and RNA folding dynamics using enhanced sampling methods

विद्या वाचस्पति की
उपाधि की अपेक्षाओं की आंशिक पूर्ति में प्रस्तुत शोध प्रबंध

A thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy

द्वारा / By

अमल विजय / Amal Vijay

पंजीकरण सं. / Registration No.: 20193628

शोध प्रबंध पर्यवेक्षक / Thesis Supervisor: Prof. Arnab Mukherjee

भारतीय विज्ञान शिक्षा एवं अनुसंधान संस्थान पुणे
INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH PUNE
2024

Dedicated to

My Family Members, Teachers, and Friends

भारतीय बिज्ञान शिक्षा एवं अनुसंधान सँस्थान, पुणे
INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH (IISER), PUNE
(An Autonomous Institution, Ministry of Human Resource Development, Govt. of India)
First floor, Central Tower, Sai Trinity Building, Garware Circle, Sutarwadi,
Pashan, Pune, Maharashtra 411021, India.

CERTIFICATE

Certified that the work incorporated in the thesis entitled, "**Exploring molecular recognition of small molecule binding, protein-protein interactions, and RNA folding dynamics using enhanced sampling methods**" submitted by **Amal Vijay** was carried out by the candidate, under my supervision. The work presented here or any part of it has not been included in any other thesis submitted previously for the award of any degree or diploma from any other University or institution.

Date: 30th December 2023

Pune (MH), India

Prof. Arnab Mukherjee

(Supervisor)

Tel.: +91-20-25908051
Fax: +91-20-25899790
Email: arnab.mukherjee@iiserpune.ac.in
Web:http://www.iiserpune.ac.in/~arnabm

DECLARATION

Name of Student: **Amal Vijay**

Reg. No.: **20193628**

Thesis Supervisor(s): **Prof. Arnab Mukherjee**

Department: **Chemistry**

Date of joining program: **01/01/2019**

Date of Pre-Synopsis Seminar : **03/11/2023**

Title of Thesis: **Exploring molecular recognition of small molecule binding, protein-protein interactions, and RNA folding dynamics using enhanced sampling methods**

I declare that this written submission represents my idea in my own words and where others' ideas have been included; I have adequately cited and referenced the original sources. I declare that I have acknowledged collaborative work and discussions wherever such work has been included. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

The work reported in this thesis is the original work done by me under the guidance of Prof. Arnab Mukherjee.

Date: 25/03/2024

Amal Vijay

ACKNOWLEDGMENTS

I would like to express my heartfelt gratitude to the individuals whose support made this thesis possible. I express my gratitude primarily to my mentor, **Prof. Arnab Mukherjee**, for his invaluable guidance and unwavering support throughout my PhD journey. I am really thankful to him for providing a healthy work environment, necessary freedom, and multiple encouraging and enlightening discussions during my PhD life. His enthusiasm in approaching any research problem is genuinely inspiring and will be warmly remembered.

I extend my gratitude to the members of my Research Advisory Committee (RAC) members, **Dr. Durba Sengupta** and **Dr. Sunish Radhakrishnan**, for monitoring my research progress and providing constructive feedback during my annual evaluations.

I would also like to express my gratitude for the opportunity to collaborate with **Prof. Ioan Andricioaei** at the University of California, Irvine, where I spent approximately nine months as part of the Fulbright fellowship. His endeavor to address complex research challenges has been truly inspiring.

I would like to thank my collaborators, **Prof. Pinaki Talukadar, Dr. Reman, Dr. Javed, Rashmi, Praveen, Rituparno, and Sreyas,** for trusting me in the process of conducting collaborative research work. I have extremely benefited from understanding new concepts and techniques in collaborating with them.

I extend my thanks to my labmates **Sonali, Rahul, Uddipan, Ashish, Bikirna, and Parth** for engaging discussions. Even though my time with **Anant, Harsha, Dheeraj, Sarath, Gopi, Prathith, Mansi, Pratishruti, Atul,** and **Manjul** were brief in the lab, the discussions with them left an indelible impression. I would like to thank my seniors, **Dr. Reman, Dr. Hridya, Dr. Debasish,** and **Dr. Abhijit,** for assisting me in learning the fundamentals of computational

chemistry and providing guidance on the technical aspects of the field. Their advice, support, and encouragement were indelible during the initial days of my Ph.D.

I would like to mention my friends **Arjun, Gokul, Vrinda, Neethu, Naveen, Markose, Cavya, and Sriram** for making my time at IISER the most enjoyable. The trips we took together and tea time discussions will be greatly missed.

I am thankful for the efforts of the administrative staff, especially **Sayalee, Swapnil, and Mayuresh,** for streamlining and easing the paperwork process. I would also like to express my gratitude to IISER Pune for providing a wonderful research experience along the journey.

I am grateful to the IT staff, including **Nisha, Neeta, and Suresh,** who promptly resolved our computer and cluster issues, ensuring smooth operations for our research.

I extend my gratitude to the Param Brahma Computational Facility at IISER Pune for generously providing computational resources for research. I also thank CSIR for the monthly fellowship, DBT-CTEP, and Infosys Fellowship for supporting my participation in international conferences. Special appreciation goes to the Fulbright Visiting Doctoral Researcher Program for facilitating research activities at the University of California, Irvine.

I extend a heartfelt gratitude to my brother, parents, friends, and relatives for their consistent support and encouragement throughout this journey.

Finally, I would also like to express my sincere thanks to those who have directly or indirectly supported me so far in this beautiful journey.

LIST OF PUBLICATIONS

Peer-reviewed

1. **Vijay, A.**; Sreyas Adury, V. S.; Mukherjee, A., Targeting RdRp of SARS-CoV-2 with De Novo Molecule Generation. *ACS Appl Bio Mater* **2023.** *(Accepted Article*)

(Discussed in Chapter 4)

2. **Vijay, A**.; Mukherjee, A., Molecular insights into the stereospecificity of arginine in RNA tetraloop folding. *Phys Chem Chem Phys* **2023,** *25* (16), 11301-11310.

(Discussed in Chapter 6)

3. Sharma, R.; **Vijay, A**.; Chattopadhayay, S.; Mukherjee, A.; Talukdar, P., Self-assembled anion channel formation by bis(1,3-propanediol)-linked meta-dipropynylbenzene-based small molecules. *Chem Commun* **2023,** *59* (24), 3602-3605.

(Discussed in Chapter 8)

4. Sharma, R.; **Vijay, A**.; Mukherjee, A.; Talukdar, P., Bis(cholyl)-based chloride channels with oxalamide and hydrazide selectivity filters. *Org Biomol Chem* **2022,** *20* (10), 2054-2058.

5. Chowdhury, R.; Sai Sreyas Adury, V.; **Vijay, A.**; Singh, R. K.; Mukherjee, A., Atomistic De-novo Inhibitor Generation-Guided Drug Repurposing for SARS-CoV-2 Spike Protein with Free-Energy Validation by Well-Tempered Metadynamics. *Chemistry, an Asian journal* **2021,** *16* (12), 1634-1642.

(Discussed in Chapter 3)

6. Malla, J. A.; Umesh, R. M.; **Vijay, A.**; Mukherjee, A.; Lahiri, M.; Talukdar, P., Apoptosis-inducing activity of a fluorescent barrel-rosette M(+)/Cl(-) channel. *Chem Sci* **2020,** *11* (9), 2420-2428.

Manuscript under preparation

1. **Vijay, A**.; Mukherjee, A., Interaction between β-catenin and its binding partner T-cell factor (TCF)

(Discussed in Chapter 5)

CONTENTS

SYNOPSIS

Title: Exploring molecular recognition of small molecule binding, protein-protein interactions, and RNA folding dynamics using enhanced sampling methods

The specific non-covalent interaction at the molecular level, termed 'molecular recognition,' is vital to biological processes such as small molecule-protein/nucleic acid interactions, folding-unfolding of protein/RNA motifs, etc. Understanding these interactions is crucial in modern chemical and biological research, e.g., in therapeutic applications for the treatment of diseases. These interactions often arise because of the difference in size and shape complementarity, chemical properties, charge differences, etc, leading to a biological response. Especially through experimental techniques, understanding faster intermediates and mechanistic details associated with molecular recognition processes in biological systems can indeed face limitations.

Molecular dynamics technique is found to be a promising tool for understanding the energetics and mechanistic aspects related to such biological responses. However, for the processes, which are usually acoompanied by free energy barriers higher than thermal energy (2.5 kJ/mol) in a multidimensional surface, enhanced sampling methods are used to accelerate the events resulting in a complex free energy landscape. The thesis aims to understand and provide molecular-level aspects of some of such relevant biological processes, especially related to protein and RNA-based systems in various complexity, using enhanced sampling methods.

In this thesis, we utilized different variations of metadynamics technique (a state-of-art enhanced sampling method) along with controlled sampling approaches to gain insights into the molecular events involved in biomolecular recognition processes.

To investigate interactions between proteins and small molecules in the context of SARS-CoV-2 systems, we employed well-tempered metadynamics simulations. The binding free energy stability derived from this method were utilized to establish the binding strength of potential candidates targeting the spike protein and RNA-dependent RNA polymerase of SARS-CoV-2. Moreover, we have utilized a multiple walker approach to comprehensively explore the free energy landscape linked to protein-protein interactions within the β-catenin: T-cell factor (TCF) system. In this instance, we employed the parallel bias version of the metadynamics technique to construct a high-dimensional free energy landscape. The investigation reveals significant changes in the secondary structure of TCF as it undergoes unbinding from the β-

catenin interface. Likewise, the multiple walker version of well-tempered metadynamics is taken forward to understand the phenomena called “RNA breathing” in a protein-RNA recognition process, where we reveal similar energy local base pair transitions in the bound state of the complex. Additionally, we explored the stereospecific impact of a particular amino acid (arginine) on an RNA tetraloop system, aiming to offer molecular insights into an experimentally validated RNA-based theory concerning the origin of homochirality. In this particular instance, we employed a controlled sampling approach using a combination of two enhanced sampling methods (umbrella sampling and parallel bias metadynamics) to achieve efficient sampling and address challenges associated with metadynamics-based simulation protocols. In addition to protein and RNA-based systems, we also modeled and investigated the energetics of the selective ion permeation events through synthetic ion channels in biological membranes to complement the experimental observations from collaborative projects.

This thesis is divided into nine chapters, and details are provided below:

1. ***Chapter 1: Introduction.*** The first chapter provides an overview of the molecular recognition process in biological systems, with relevant examples. This includes discussions about the host-guest interactions in protein-based systems and secondary interactions in RNA-based systems.
2. ***Chapter 2: Methodology.*** This chapter discusses the necessity of advanced molecular dynamics simulation techniques to comprehend the free energy stability associated with diverse biomolecular processes. Moreover, the details regarding various enhanced sampling techniques to understand complex biomolecular processes are also discussed in more detail. This includes metadynamics-based techniques and controlled sampling approaches for the exploration of conformational space. Further, the necessity of collective variables in computational methods to enhance understanding of biomolecular recognition processes is also discussed in more detail.
3. ***Chapter 3: Potential therapeutics for COVID-19 by targeting spike protein of SARS-CoV-2 .*** Targeting the spike protein of the SARS-CoV-2 virus by small molecules was found to be a pivotal strategy in finding therapeutics for the treatment of COVID-19. This chapter focuses on finding potential small molecules that can prevent the interaction between the virus's receptor-binding domain (RBD) and human angiotensin-converting enzyme 2 (hACE2), responsible for the virus's spread. We have utilized an in-house developed drug generation algorithm to generate new small molecules by

targeting the RBD region of the spike protein subjected to free energy validation study using well-tempered metadynamics simulation technique. By rigorous free energy calculations, we showed that some newly generated molecules and some repurposable drugs have shown stronger binding to RBD than hACE2. This work was conducted in collaboration with Rituparno Chowdhury, Venkata Sai Sreyas Adury, and Dr. Reman Kumar Singh.

4. ***Chapter 4: De novo drug generation and free energy validation by targeting RNA Dependent RNA Polymerase (RdRp) of SARS-CoV-2.*** The RNA-dependent RNA polymerase (RdRp) is a crucial enzyme for the replication of RNA viruses, including the SARS-CoV-2 virus responsible for COVID-19. Remdesivir is an FDA-approved antiviral drug that targets the RdRp of SARS-CoV-2. This chapter focuses on finding new molecules that can bind more effectively than Remdesivir in its active form by targeting the active site of RdRp. The active site of RdRp was considered for the successful generation of novel molecules by utilizing the drug generation algorithm developed in-house. Based on free energy calculations using well-tempered metadynamics simulations for all the molecules, we found a few new molecules with better binding affinity than Remdesivir in its active form. We also propose the possible unbinding mechanism & key interactions in Remdesivir in its active form and the best binding novel molecule responsible for free energy stabilization. This work was conducted in collaboration with Venkata Sai Sreyas Adury.

5. ***Chapter 5: Interaction between β-catenin and its binding partner T-cell factor (TCF).*** Similar to protein-small molecule interactions, protein-protein interactions (PPIs) also play a fundamental role in cellular function and regulation. The interaction between beta-catenin and TCF is a key molecular event in the Wnt signaling pathway. This chapter focuses on understanding the molecular events associated with the unbinding of TCF from the β-catenin interface. A multiple-walker version of the parallel bias metadynamics technique was utilized in this study for free energy calculations, as protein-protein interactions are more complex than protein-small molecule interactions. The study reveals that significant secondary structural changes in TCF facilitate the folding of TCF during the unbinding process from the β-catenin interface.
6. ***Chapter 6: Molecular insights into the stereospecificity of arginine in RNA tetraloop folding.*** This chapter aims to understand the secondary interactions in RNA motifs in a

chiral atmosphere of amino acids. RNA-based hypothesis in the context of the origin of homochirality in amino acids suggests that a specific amino acid enantiomer offers enhanced RNA folding stability by acting as a chemical chaperon. However, the molecular details associated with this hypothesis were not entirely known. Here, we study the effect of arginine enantiomers on RNA tetraloop folding-unfolding equilibria using a controlled sampling approach for the construction of a multidimensional landscape. A combination of umbrella sampling and parallel bias metadynamics was utilized for the efficient controlled sampling associated with the process. The study reveals that specific geometric constraint in the RNA motif prefers the D-enantiomer to stack at the terminal bases of RNA, resulting in enhanced folding stability.

7. ***Chapter 7: Scope of "RNA breathing" in Protein: RNA complex recognition.*** Nature uses genetic switches that are composed of RNA-protein complexes (RNPs) to regulate gene expression at the RNA level**.** Targeting specific protein-RNA interactions also holds the potential for developing therapeutic strategies for various diseases. This chapter focuses on understanding local structural changes in the RNA bases in the RNA-bound structure of an L7Ae protein during the specific recognition process. We found that similar to the native bound state, RNA can also exist in a non-native stable confirmation, also defined as RNA breathing phenomena. Further analysis shows that the existence of these states is triggered by changes in the base pairing orientations contributed by edges of interactions (Watson-Crick, Hoogsteen/C-H edge and sugar edges). This study was conducted in collaboration with Prof. Ioan Andricioaei at the University of California Irvine, USA.
8. ***Chapter 8: Selective anion permeation through synthetic ion channel.*** Synthetic ion channels are de novo chemical compounds that insert into lipid bilayers form pores. As a result, the channels can facilitate the ion flow from one side to the other. In addition to the specific recognition processes in protein and RNA-based systems, this chapter aims to provide insights into experimental observations on specific Cl^- ion permeation through synthetic channels made of bis(1,3-propanediol)-linked m-dipropynylbenzene-based molecules. We have modeled the lipid-ion-channel system and investigated the free energy change associated with the permeation of chloride ions through the channel. Multiple walker well-tempered metadynamics simulations were utilized for free energy calculations. We found that the stable interactions between the Cl^- ion and specific groups in the synthetic channel can stabilize the ion channel interactions and further

facilitate the transport through channels. This study was conducted in collaboration with Prof. Pinaki Talukdar's group at IISER Pune.

9. ***Chapter 9: Conclusion***. We found that the choice of good collective variable and specific system-dependent utilization of enhanced sampling methods are essential to reveal the energetics and mechanistic aspects in the biological system of interest. This chapter summarises the outcomes of the studies, importance of system specefic enhanced sampling protocol and future directions with the goal of understanding the molecular recognition process in biomolecules in a better and more efficient way using computational methodologies.

Chapter 1

Introduction

1.1 Molecular recognition in biological systems

In molecular biology, specific non-covalent interaction at the molecular level typically involving biological macromolecules such as proteins, nucleic acids, and carbohydrates is referred to as molecular recognition.[1, 2] The non-covalent interactions include hydrogen bonding, hydrophobic interactions, van der Waals forces, π-π interactions, resonant interactions, etc.[2-4] In instances where such interactions occur within a single molecule, they are termed intra-molecular recognition.[5, 6] Examples of intra-molecular recognition are the protein/RNA folding processes, where a linear sequence of amino acids/nucleotides undergoes a transformation into its three-dimensional, functional structure. Similarly, if the interactions occur most abundantly between two or more molecules, they are referred to as intermolecular recognition. Instances of intermolecular recognition processes include interactions such as protein-small molecule, protein-protein, protein-RNA, and RNA-small molecule interactions.[7-10] Fig. 1 illustrates some of the molecular recognition processes in biological systems. These interactions frequently emerge due to variations in size and shape complementarity, chemical properties, and charge differences, ultimately eliciting a biological response.[7, 11, 12] In this thesis, our aim is to provide insights into questions that arise in molecular recognition processes, primarily focusing on systems based on proteins and RNA.

Computational approaches, which include molecular dynamics simulation techniques, are recognized as promising tools for comprehending biomolecular recognition processes to complement experimental studies.[13-15] Free energy landscape which can be constructed using advanced simulation protocol, can provide valuable insights into the structural changes and conformations adopted by biomolecules during particular recognition processes. Chapter 2 discussed the molecular dynamics simulation techniques and various enhanced sampling techniques to study the free energy change associated with the molecular recognition processes discussed in the thesis.

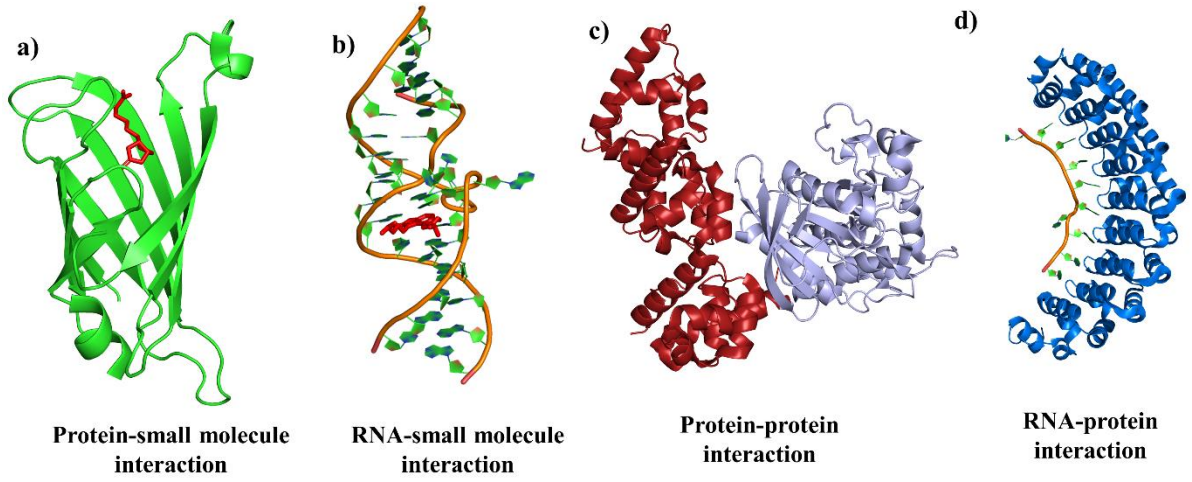

Figure 1. Examples of molecular recognition processes in biological systems. Figures are made using PyMOL[16]. Protein data bank (PDB) ID and name of the complex are provided. a) 3RDO, streptavidin complexed with biotin[17] b) 3TZR, Riboswitch like RNA-ligand complex from the Hepatitis C Virus Internal Ribosome Entry Site[18] c) 6KIP, PTPRD phosphatase domain in complex with Liprin-alpha3 tandem SAM domains[19] d) 3BSX, Crystal Structure of Human Pumilio 1 in complex with Puf5 RNA[20]

1.2 Protein-small molecule interactions

Proteins play a central role in biological processes, facilitating reactions, transporting molecules, orchestrating immune responses against pathogens, and mediating signal transduction between cells.[21] Protein-small interactions play a pivotal role in the field of pharmacology, influencing the efficacy and safety of therapeutic interventions.[22-25] The traditional approach to small molecule drug discovery primarily concentrates on protein–small molecule interactions involving enzymes, ion channels, and receptors, etc. as these proteins commonly possess a clearly defined binding site or, referred to as a hotspot, facilitating interactions with small molecules.[26-28] Exploring protein-small molecule interactions enhances our comprehension of molecular pharmacology and paves the way for uncovering new medications with enhanced selectivity and effectiveness. Anticipating the thermodynamic characteristics of a drug's interaction with its molecular target is crucial for understanding its mechanism of action and advancing the development of novel medications.[14] This endeavor's primary result involves determining the binding mode and computing the free energy associated with small molecule binding. For instance, the COVID-19 pandemic response exemplifies the crucial need to develop innovative molecules aimed at treating SARS-CoV-2, the causative virus of the disease.[29-31] Utilizing computational approaches, which include molecular dynamics simulations and drug discovery algorithms, is recognized as a viable method for identifying and designing novel therapeutics for COVID-19.[32-35] This

comprehensive approach explores vast chemical space, prioritizes the most promising candidates, and refines their chemical structures to enhance antiviral activity.[36, 37] The streamlined process accelerates drug discovery by narrowing down the pool of potential compounds readying them for further experimental validation. In this context, chapters 3 and 4 of this thesis delve into a successful enhanced sampling simulation protocol that utilizes an in-house de novo drug generation algorithm to conduct rigorous free energy validation calculations to identify new small molecules against COVID-19. The specific focus is on targeting the spike protein and RdRp of SARS-CoV-2.[38, 39]

1.3 Protein-Protein interactions

Protein–protein interactions (PPIs) denote the physical associations and binding occurrences involving two or more protein molecules within a biological system.[40-42] PPIs are linked to numerous human diseases, including cancer, infectious diseases, and neurodegenerative disorders.[43-45] As conventional drug targets typically involve protein machinery, the exploration of PPIs has emerged as appealing targets for drug development in recent years.[46, 47]. Viable approaches for discovering hits and leads of PPI modulators include high-throughput screening, fragment-based drug discovery, structure-based design, and virtual screening methods.[27] . A representation of the possible recognition and inhibition in Protein-Protein complex is shown in the Fig. 2. Targeting PPIs with small molecules, however, is commonly perceived as challenging and regarded as “undruggable” due to the following reasons.[48, 49] First, the protein-protein surface area of interaction at the interface (typically extends to 1500–3000 $Å^2$) exceeds the area of interaction observed in protein-ligand interactions (typically extends to 300–1000 $Å^2$). [50, 51]. Second, the PPI interface often contains limited grooves or pockets, posing challenges for the effective binding of designed small molecule compounds.[52-54] Further, as the PPIs are regarded as high-affinity interactions, inhibiting such a strong interaction using small molecules is considered ineffective.[55] Besides, in contrast to conventional drug target protein receptors, reference small molecules considered for drug design towards PPIs is limited.[55]

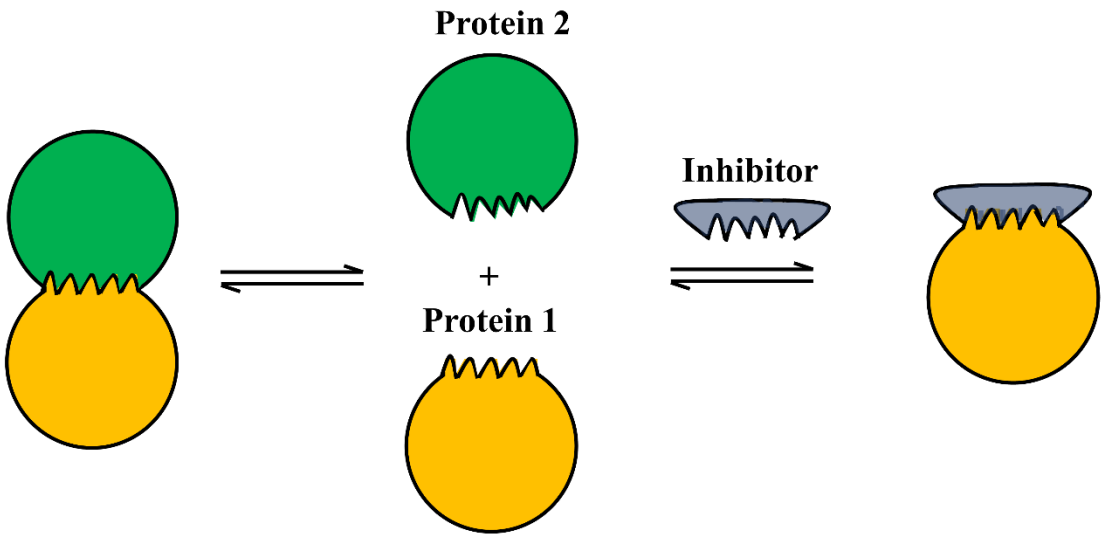

Figure 2. Depiction of the recognition and inhibition in Protein-Protein complex.[56]

Nevertheless, the identification of "hot-spots" facilitates the development of drugs targeting protein-protein interactions, as these regions are expected to contribute to the binding-free energy.[57, 58] The presence of known hotspots in protein-protein complexes provides the advantage of being able to design ligands that directly influence the corresponding protein-protein interaction (PPI). Chapter 5 of the thesis explores the dissociation of a protein-protein complex, specifically focusing on the unbinding process of the β-catenin-TCF (T cell factor) interaction. This investigation holds significant importance in drug discovery and understanding the Wnt signalling pathway.[59]

1.4 Secondary interactions in RNA

Like proteins, RNA motifs also participate in diverse cellular functions, including regulating gene expression, catalysis of various substrates, formation of complex assemblies, and facilitation of molecular chaperoning.[60-62] Secondary interactions in RNA refer to the non-canonical base pairings and structural elements that contribute to the three-dimensional folding and stability of RNA molecules. While the primary structure of RNA involves a linear sequence of nucleotides, the secondary structure arises from interactions such as base pairing between complementary nucleotides, leading to the formation of helixes, hairpin loops, bulge loops, internal loops, and multi-branched loops.[63] A representation of various structural elements in RNA secondary structure is shown in Fig. 3. RNA architectures can adopt diverse structures

featuring A–U and G–C Watson–Crick base pairs, as well as G–U wobble base pairs, collectively referred to as 'canonicals.'[64]

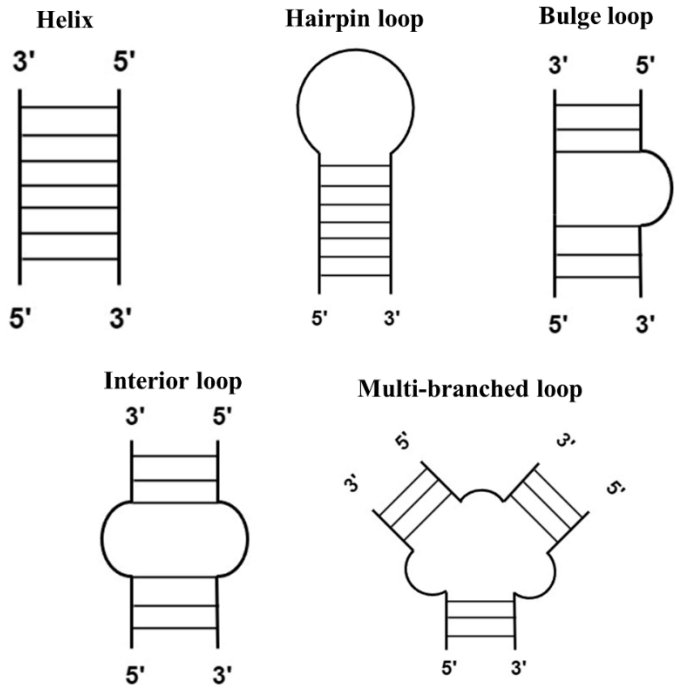

Figure 3. Structural representation of various RNA secondary structure elements: helix, hairpin loop, bulge loop, interior loop and multibranched loop. Horizontal black lines indicate the complimentary base pairing between the strands of RNA.

Additionally, non-canonical base pairings, such as Hoogsteen base pairs and base triples, contribute to the intricate folding patterns of RNA.[65] These non-canonical base pairs, comprising approximately 40% of RNA's base-pairing interactions, have been identified for their involvement in biomolecular recognition events.[66, 67] The Leontis[66] notation is commonly employed to represent canonical and non-canonical base pairs and other features in RNA secondary structures. The foundation of this notation and classification lies in the categorization of edge interactions (Watson–Crick edge, Hoogsteen edge, C-H edge and sugar edge) and the orientation (cis or trans) of glycosidic bonds around RNA residues. This approach helps to describe the local structural changes found in RNA molecules. An illustration of interacting edges and orientation of glycosidic bonds around the sugar backbone around RNA is shown in the Fig. 4.

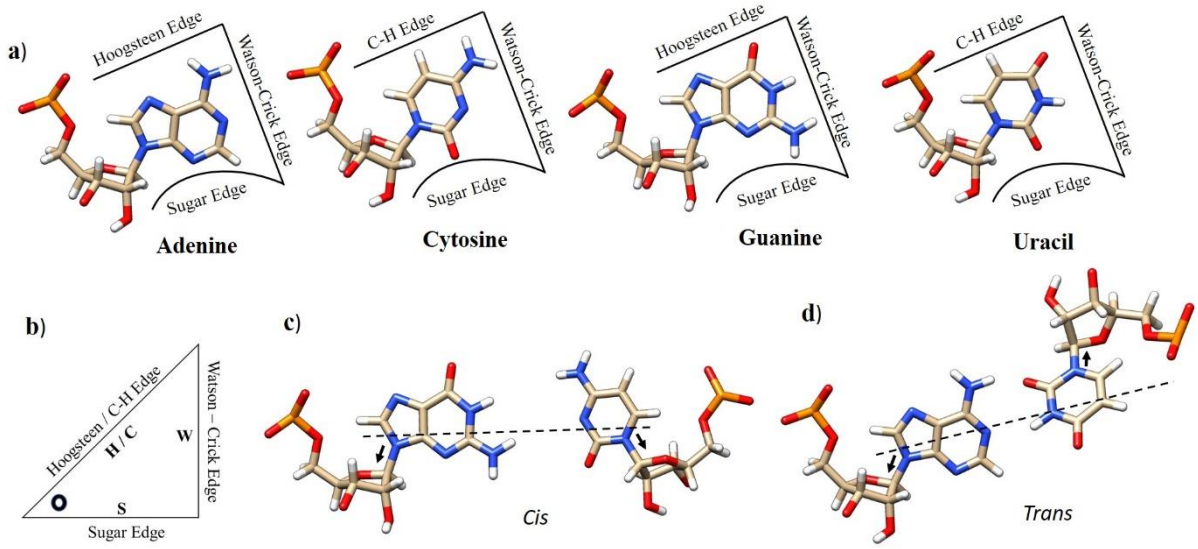

Figure 4. a) Illustration of interacting edges in RNA bases (Adenine, Cytosine, Guanine and Uracil). b) Triangle representation of Watson-Crick, Hoogsteen, Sugar and C-H edges. The black circle indicates the position of ribose unit. c) Illustration of cis geometry formed with respect to the glycosidic bonds formed by complimenatary bases in nucleotide unit. d) Illustration of trans geometry formed with respect to the glycosidic bonds formed by complimentary bases in nucleotide unit.

A three-letter code is assigned to denote each potential complementary interaction between two bases within the RNA structure. In this code, the first letter signifies the likely donating edge for base pairing in a given RNA residue. The second letter specifies the probable acceptor edge of interaction in the complementary residue in RNA. The third letter indicates the potential cis/trans mode of interactions, determined by the glycosidic rotation of the sugar backbone. For instance, the code WHc signifies that the 5'-end of a particular stem residue with a Watson–Crick edge is engaged in interaction with the neighboring Hoogsteen acceptor edge in the cis orientation. The possible notations for the three letter codes for the structural representation of RNA are described in Table 1. Comprehending the discussed secondary interactions is crucial for unraveling the functional and structural aspects of the RNA folding process, as shown in Chapter 6, which delves into the impact of arginine enantiomers on the folding of an RNA tetraloop motif. Experimental approaches, such as RNA structure probing, and computational methods, like RNA folding prediction algorithms, play a crucial role in untangling the intricacies of RNA folding and its regulation.[68, 69] Further, understanding secondary interactions in RNA folding behavior remains a fascinating and complex topic in the context of the origins of life.[70] The goal of the study in Chapter 6 is to offer molecular insights into

the specific folding behavior of RNA within the chiral environment of amino acids, supplementing experimental observations related to RNA-based hypotheses concerning the origin of life.[71]

Table 1. Notation for the three letter codes assigned based on the edges of interactions between the bases and glycosidic bond orientation. The notation is followed based on Lenotis represention of RNA base classification

Sl No	Starting Edge	Neighbouring Edge	Orientation	Representation
1	Hoogsteen	Sugar	*cis*	**WSc**
2	Sugar	Watson-Crick	*cis*	**SWC**
3	Hoogsteen	Hoogsteen	*trans*	**HHt**
4	Sugar	Hoogsteen	*trans*	**SHt**
5	Hoogsteen	Watson-Crick	*cis*	**HWc**
6	Watson-Crick	Sugar	*cis*	**WSc**
7	Hoogsteen	Hoogsteen	*cis*	**HHc**
8	Sugar	Sugar	*cis*	**SSc**
9	Watson-Crick	Hoogsteen	*trans*	**WHt**
10	Watson-Crick	Watson-Crick	*trans*	**WWt**
11	Sugar	Hoogsteen	*cis*	**SHc**
12	Watson-Crick	Hoogsteen	*cis*	**WHc**
13	Watson-Crick	C-H	*cis*	**WCc**
14	Hoogsteen	Watson-Crick	*trans*	**HWt**
15	Sugar	Watson-Crick	*trans*	**SWt**
16	Watson-Crick	Watson-Crick	*cis*	**WWc**
17	Hoogsteen	Sugar	*trans*	**HSt**
18	Watson-Crick	Sugar	*trans*	**WSt**
19	Sugar	Sugar	*trans*	**SSt**

1.5 Protein-RNA interactions

RNA molecules exhibit greater structural versatility compared to DNA, enabling them to interact more dynamically with proteins.[29] Regardless of their structure and function, nearly all RNAs engage with one or more protein partners to fulfill their functions accurately.[72, 73] RNA-binding proteins form a diverse group with various functions, representing a crucial aspect in understanding cellular RNAs.[72] This poses a significant challenge in the post-genomic era. During the protein-RNA recognition process, both RNA and proteins exhibit the capability to undergo structural changes, promoting the establishment of stable complexes and facilitating subsequent biological functions.[74] These changes are brought about through protein-induced RNA conformational changes, RNA-induced protein conformational changes, and co-induced conformational alterations, as illustrated in Fig 5.[74, 75]

Despite the canonical Watson–Crick pairs serving as the fundamental building blocks for constructing three-dimensional frameworks in RNA, non-canonical base pairs have also been recognized for their specific role in hydrogen-bonding interactions with proteins.[76] This is due to the distinct hydrogen bonding patterns and arrangements of bases compared to those found in canonical pairs.[76]. Due to the presence on non-canonical base pairs, induced conformational changes play a significant importance in protein-RNA recognition processes, as in the non-canonical pairs, alternative sets of polar donor and acceptor groups of the bases are accessible for intermolecular contacts with protein. For example, peptide recognition sites in both the HIV-1 Rev response element (RRE) RNA and a Rev-specific aptamer RNA feature non-canonical base pairs. [77, 78] In the crystal structure of the spliceosomal U2B"-U2A' protein complex bound to a fragment of U2 small nuclear RNA, non-canonical base pairs were similarly identified within the recognition site.[79]. Thus, at the interface of RNA and protein molecules, the transition from the native structure in RNA to a non-native base-pairing scheme constitutes an essential aspect of protein-RNA recognition. Our study explores a similar observation, delving into the existence of stable non-native structures similar to the native configuration in RNA-bound proteins. This phenomenon, often challenging to observe in experimental studies, is discussed in chapter 7.

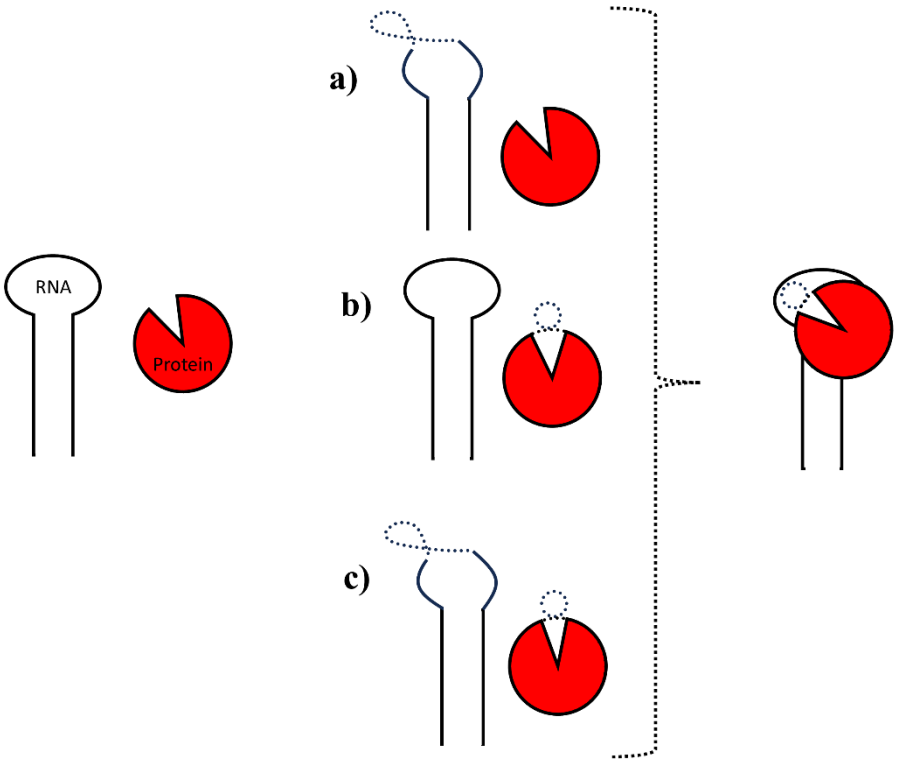

Figure 5. Possible conformational changes during protein-RNA recognition process. a) Protein induced RNA conformational changes b) RNA induced protein conformational changes c) co-induced conformational changes.[74]

1.6 Membrane proteins

Membrane proteins are an important component of cellular membranes and plays a significant role in several biological processes.[80, 81] Membrane proteins are of two kinds, integral membrane proteins and peripheral membrane proteins. Integral membrane proteins are embedded into the membrane and consist of both extracellular and intracellular components, while peripheral proteins are associated with the surface of the membrane.[82-84] Membrane proteins take up several important roles in the physiology of the cell, including signaling, transport, immune recognition, adhesion and catalytic activity.[80, 82] In particular, membranes associated with transport are of three types - pores, channels and carriers.[85] Carrier proteins transport molecules from one side of the membrane to the other by binding to them and undergoing a conformational change, unlike pores, which are always open, allowing molecules to pass through them.[86] Meanwhile, Channel proteins, or ion channels, transport molecules if they are triggered via a stimulus.[87] Representation of an ion transport protein embedded inside a membrane is shown in the Fig. 6.

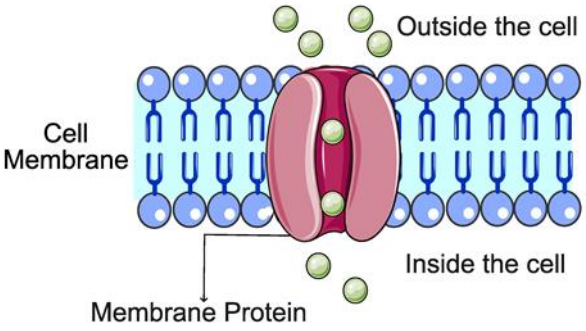

Figure 6. Representation of an ion transport protein embedded inside a membrane. The movement of ions (shown in green) is shown from outside to inside the cell.

Similar to proteins, synthetic ion channels are also recognized for their ability to transport ions across biological membranes.[88] While inspired by the biological functions of natural protein-based channels, they differ in composition. Progress in synthetic chemistry, biology, and nanotechnology has played a crucial role in shaping the development of artificial ion channels with diverse architectures. Chapter 8 of the thesis discusses the driving force and mechanism

governing the selective permeation of Cl- ions through experimentally reported synthetic ion channels, which is based on the bis(1,3-propanediol)-linked m-dipropynylbenzene framework.

Beyond presenting findings and insights on various molecular recognition processes discussed throughout the thesis, Chapter 9 summarises the challenges faced, the strategies used to address them, and potential directions for future research and advancements. The primary emphasis is on deepening the understanding of complex molecular recognition processes within biological systems with the aid of advanced computational techniques.

1.7 References

1. Donato, L., *Molecular Recognition*, in *Encyclopedia of Membranes*, E. Drioli and L. Giorno, Editors. 2015, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 1-2.
2. Pauling, L. and M. Delbruck, *The Nature of the Intermolecular Forces Operative in Biological Processes.* Science, 1940. **92**(2378): p. 77-9.
3. Cosic, I., *Macromolecular bioactivity: is it resonant interaction between macromolecules?--Theory and applications.* IEEE Trans Biomed Eng, 1994. **41**(12): p. 1101-14.
4. Breiten, B., et al., *Water networks contribute to enthalpy/entropy compensation in protein-ligand binding.* J Am Chem Soc, 2013. **135**(41): p. 15579-84.
5. Chen, Y.J. and M. Inouye, *The intramolecular chaperone-mediated protein folding.* Curr Opin Struct Biol, 2008. **18**(6): p. 765-70.
6. Newberry, R.W. and R.T. Raines, *Secondary Forces in Protein Folding.* ACS Chem Biol, 2019. **14**(8): p. 1677-1686.
7. Gilson, M.K. and H.X. Zhou, *Calculation of protein-ligand binding affinities.* Annu Rev Biophys Biomol Struct, 2007. **36**: p. 21-42.
8. Sedov, I.A. and Y.F. Zuev, *Recent Advances in Protein-Protein Interactions.* Int J Mol Sci, 2023. **24**(2).
9. Corley, M., M.C. Burns, and G.W. Yeo, *How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms.* Mol Cell, 2020. **78**(1): p. 9-29.
10. Childs-Disney, J.L., et al., *Targeting RNA structures with small molecules.* Nat Rev Drug Discov, 2022. **21**(10): p. 736-762.
11. Zhang, Q., M. Sanner, and A.J. Olson, *Shape complementarity of protein-protein complexes at multiple resolutions.* Proteins, 2009. **75**(2): p. 453-67.

12. Sulea, T. and E.O. Purisima, *Profiling charge complementarity and selectivity for binding at the protein surface.* Biophys J, 2003. **84**(5): p. 2883-96.
13. Fukunishi, Y., J. Higo, and K. Kasahara, *Computer simulation of molecular recognition in biomolecular system: from in silico screening to generalized ensembles.* Biophys Rev, 2022. **14**(6): p. 1423-1447.
14. Raniolo, S. and V. Limongelli, *Ligand binding free-energy calculations with funnel metadynamics.* Nat Protoc, 2020. **15**(9): p. 2837-2866.
15. Gsponer, J. and A. Caflisch, *Molecular dynamics simulations of protein folding from the transition state.* Proc Natl Acad Sci U S A, 2002. **99**(10): p. 6719-24.
16. Schrodinger, LLC, *The PyMOL Molecular Graphics System, Version 1.8*. 2015.
17. Magalhaes, M.L., et al., *Evolved streptavidin mutants reveal key role of loop residue in high-affinity binding.* Protein Sci, 2011. **20**(7): p. 1145-54.
18. Dibrov, S.M., et al., *Structure of a hepatitis C virus RNA domain in complex with a translation inhibitor reveals a binding mode reminiscent of riboswitches.* Proc Natl Acad Sci U S A, 2012. **109**(14): p. 5223-8.
19. Wakita, M., et al., *Structural insights into selective interaction between type IIa receptor protein tyrosine phosphatases and Liprin-alpha.* Nat Commun, 2020. **11**(1): p. 649.
20. Gupta, Y.K., et al., *Structures of human Pumilio with noncognate RNAs reveal molecular mechanisms for binding promiscuity.* Structure, 2008. **16**(4): p. 549-57.
21. Morris, R., K.A. Black, and E.J. Stollar, *Uncovering protein function: from classification to complexes.* Essays Biochem, 2022. **66**(3): p. 255-285.
22. Zhao, L., et al., *A brief review of protein-ligand interaction prediction.* Comput Struct Biotechnol J, 2022. **20**: p. 2831-2838.
23. Olsson, T.S., et al., *The thermodynamics of protein-ligand interaction and solvation: insights for ligand design.* J Mol Biol, 2008. **384**(4): p. 1002-17.
24. Decherchi, S. and A. Cavalli, *Thermodynamics and Kinetics of Drug-Target Binding by Molecular Simulation.* Chem Rev, 2020. **120**(23): p. 12788-12833.
25. Amaral, M., et al., *Protein conformational flexibility modulates kinetics and thermodynamics of drug binding.* Nat Commun, 2017. **8**(1): p. 2276.
26. Santos, R., et al., *A comprehensive map of molecular drug targets.* Nat Rev Drug Discov, 2017. **16**(1): p. 19-34.

27. Lu, H., et al., *Recent advances in the development of protein-protein interactions modulators: mechanisms and clinical trials.* Signal Transduct Target Ther, 2020. **5**(1): p. 213.
28. Olah, J., et al., *Challenges in Discovering Drugs That Target the Protein-Protein Interactions of Disordered Proteins.* Int J Mol Sci, 2022. **23**(3).
29. Li, G., et al., *Therapeutic strategies for COVID-19: progress and lessons learned.* Nat Rev Drug Discov, 2023. **22**(6): p. 449-475.
30. Robinson, B.W.S., A. Tai, and K. Springer, *Why we still need drugs for COVID-19 and can't just rely on vaccines.* Respirology, 2022. **27**(2): p. 109-111.
31. Zhu, Y., J. Li, and Z. Pang, *Recent insights for the emerging COVID-19: Drug discovery, therapeutic options and vaccine development.* Asian J Pharm Sci, 2021. **16**(1): p. 4-23.
32. Napolitano, F., X. Xu, and X. Gao, *Impact of computational approaches in the fight against COVID-19: an AI guided review of 17 000 studies.* Brief Bioinform, 2022. **23**(1).
33. Muratov, E.N., et al., *A critical overview of computational approaches employed for COVID-19 drug discovery.* Chem Soc Rev, 2021. **50**(16): p. 9121-9151.
34. Liu, C.H., C.H. Lu, and L.T. Lin, *Pandemic strategies with computational and structural biology against COVID-19: A retrospective.* Comput Struct Biotechnol J, 2022. **20**: p. 187-192.
35. Galindez, G., et al., *Lessons from the COVID-19 pandemic for advancing computational drug repurposing strategies.* Nature Computational Science, 2021. **1**(1): p. 33-41.
36. Joshi, G., et al., *Recent efforts for drug identification from phytochemicals against SARS-CoV-2: Exploration of the chemical space to identify druggable leads.* Food Chem Toxicol, 2021. **152**: p. 112160.
37. Kumar, A., et al., *Exploiting cheminformatic and machine learning to navigate the available chemical space of potential small molecule inhibitors of SARS-CoV-2.* Comput Struct Biotechnol J, 2021. **19**: p. 424-438.
38. Chowdhury, R., et al., *Atomistic De-novo Inhibitor Generation-Guided Drug Repurposing for SARS-CoV-2 Spike Protein with Free-Energy Validation by Well-Tempered Metadynamics.* Chem Asian J, 2021. **16**(12): p. 1634-1642.
39. Vijay, A., V.S. Sreyas Adury, and A. Mukherjee, *Targeting RdRp of SARS-CoV-2 with De Novo Molecule Generation.* ACS Appl Bio Mater, 2023.

40. Mangani, S., *Disruption of Protein-Protein Interfaces*. 2013: Springer.
41. Stelzl, U., et al., *A human protein-protein interaction network: a resource for annotating the proteome.* Cell, 2005. **122**(6): p. 957-68.
42. Rual, J.F., et al., *Towards a proteome-scale map of the human protein-protein interaction network.* Nature, 2005. **437**(7062): p. 1173-8.
43. White, A.W., A.D. Westwell, and G. Brahemi, *Protein-protein interactions as targets for small-molecule therapeutics in cancer.* Expert Rev Mol Med, 2008. **10**: p. e8.
44. Blazer, L.L. and R.R. Neubig, *Small molecule protein-protein interaction inhibitors as CNS therapeutic agents: current progress and future hurdles.* Neuropsychopharmacology, 2009. **34**(1): p. 126-41.
45. Rosell, M. and J. Fernandez-Recio, *Hot-spot analysis for drug discovery targeting protein-protein interactions.* Expert Opin Drug Discov, 2018. **13**(4): p. 327-338.
46. Milroy, L.G., et al., *Modulators of protein-protein interactions.* Chem Rev, 2014. **114**(9): p. 4695-748.
47. Nevola, L. and E. Giralt, *Modulating protein-protein interactions: the potential of peptides.* Chem Commun (Camb), 2015. **51**(16): p. 3302-15.
48. Coyne, A.G., D.E. Scott, and C. Abell, *Drugging challenging targets using fragment-based approaches.* Curr Opin Chem Biol, 2010. **14**(3): p. 299-307.
49. Winter, A., et al., *Biophysical and computational fragment-based approaches to targeting protein-protein interactions: applications in structure-guided drug discovery.* Q Rev Biophys, 2012. **45**(4): p. 383-426.
50. Smith, M.C. and J.E. Gestwicki, *Features of protein-protein interactions that translate into potent inhibitors: topology, surface area and affinity.* Expert Rev Mol Med, 2012. **14**: p. e16.
51. Cheng, A.C., et al., *Structure-based maximal affinity model predicts small-molecule druggability.* Nat Biotechnol, 2007. **25**(1): p. 71-5.
52. Buchwald, P., *Small-molecule protein-protein interaction inhibitors: therapeutic potential in light of molecular size, chemical space, and ligand binding efficiency considerations.* IUBMB Life, 2010. **62**(10): p. 724-31.
53. Arkin, M.R. and J.A. Wells, *Small-molecule inhibitors of protein-protein interactions: progressing towards the dream.* Nat Rev Drug Discov, 2004. **3**(4): p. 301-17.

54. Diaz-Eufracio, B.I., J.J. Naveja, and J.L. Medina-Franco, *Protein-Protein Interaction Modulators for Epigenetic Therapies.* Adv Protein Chem Struct Biol, 2018. **110**: p. 65-84.
55. Ivanov, A.A., F.R. Khuri, and H. Fu, *Targeting protein-protein interactions as an anticancer strategy.* Trends Pharmacol Sci, 2013. **34**(7): p. 393-400.
56. Wilson, A.J., *Inhibition of protein-protein interactions using designed molecules.* Chem Soc Rev, 2009. **38**(12): p. 3289-300.
57. Geppert, T., et al., *Context-based identification of protein-protein interfaces and "hot-spot" residues.* Chem Biol, 2011. **18**(3): p. 344-53.
58. Moreira, I.S., P.A. Fernandes, and M.J. Ramos, *Hot spots--a review of the protein-protein interface determinant amino-acid residues.* Proteins, 2007. **68**(4): p. 803-12.
59. Cui, C., et al., *Is beta-Catenin a Druggable Target for Cancer Therapy?* Trends Biochem Sci, 2018. **43**(8): p. 623-634.
60. Choi, S.I., K. Ryu, and B.L. Seong, *RNA-mediated chaperone type for de novo protein folding.* RNA Biol, 2009. **6**(1): p. 21-4.
61. Roundtree, I.A., et al., *Dynamic RNA Modifications in Gene Expression Regulation.* Cell, 2017. **169**(7): p. 1187-1200.
62. Grabow, W.W. and L. Jaeger, *RNA self-assembly and RNA nanotechnology.* Acc Chem Res, 2014. **47**(6): p. 1871-80.
63. Ding, Y. and C.E. Lawrence, *A statistical sampling algorithm for RNA secondary structure prediction.* Nucleic Acids Res, 2003. **31**(24): p. 7280-301.
64. Parisien, M. and F. Major, *The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data.* Nature, 2008. **452**(7183): p. 51-5.
65. Lemieux, S. and F. Major, *RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire.* Nucleic Acids Res, 2002. **30**(19): p. 4250-63.
66. Leontis, N.B. and E. Westhof, *Geometric nomenclature and classification of RNA base pairs.* RNA, 2001. **7**(4): p. 499-512.
67. Sharma, P., et al., *On the role of Hoogsteen:Hoogsteen interactions in RNA: ab initio investigations of structures and energies.* RNA, 2010. **16**(5): p. 942-57.
68. Zhang, H., et al., *A New Method of RNA Secondary Structure Prediction Based on Convolutional Neural Network and Dynamic Programming.* Front Genet, 2019. **10**: p. 467.

69. Wang, X.W., et al., *RNA structure probing uncovers RNA structure-dependent biological functions.* Nat Chem Biol, 2021. **17**(7): p. 755-766.
70. Herschlag, D., *RNA chaperones and the RNA folding problem.* J Biol Chem, 1995. **270**(36): p. 20871-4.
71. Vijay, A. and A. Mukherjee, *Molecular insights into the stereospecificity of arginine in RNA tetraloop folding.* Phys Chem Chem Phys, 2023. **25**(16): p. 11301-11310.
72. Cusack, S., *RNA-protein complexes.* Curr Opin Struct Biol, 1999. **9**(1): p. 66-73.
73. Licatalosi, D.D., et al., *HITS-CLIP yields genome-wide insights into brain alternative RNA processing.* Nature, 2008. **456**(7221): p. 464-9.
74. Gopinath, S.C., *Mapping of RNA-protein interactions.* Anal Chim Acta, 2009. **636**(2): p. 117-28.
75. Majumder, M. and V. Palanisamy, *Compendium of Methods to Uncover RNA-Protein Interactions In Vivo.* Methods Protoc, 2021. **4**(1).
76. Hermann, T. and E. Westhof, *Non-Watson-Crick base pairs in RNA-protein recognition.* Chem Biol, 1999. **6**(12): p. R335-43.
77. Battiste, J.L., et al., *Alpha helix-RNA major groove recognition in an HIV-1 rev peptide-RRE RNA complex.* Science, 1996. **273**(5281): p. 1547-51.
78. Ye, X., et al., *Deep penetration of an alpha-helix into a widened RNA major groove in the HIV-1 rev peptide-RNA aptamer complex.* Nat Struct Biol, 1996. **3**(12): p. 1026-33.
79. Price, S.R., P.R. Evans, and K. Nagai, *Crystal structure of the spliceosomal U2B''-U2A' protein complex bound to a fragment of U2 small nuclear RNA.* Nature, 1998. **394**(6694): p. 645-50.
80. Hedin, L.E., K. Illergard, and A. Elofsson, *An introduction to membrane proteins.* J Proteome Res, 2011. **10**(8): p. 3324-31.
81. Gadsby, D.C., *Ion channels versus ion pumps: the principal difference, in principle.* Nat Rev Mol Cell Biol, 2009. **10**(5): p. 344-52.
82. Jelokhani-Niaraki, M., *Membrane Proteins: Structure, Function and Motion.* Int J Mol Sci, 2022. **24**(1).
83. Boes, D.M., A. Godoy-Hernandez, and D.G.G. McMillan, *Peripheral Membrane Proteins: Promising Therapeutic Targets across Domains of Life.* Membranes (Basel), 2021. **11**(5).
84. Whitelegge, J.P., *Integral membrane proteins and bilayer proteomics.* Anal Chem, 2013. **85**(5): p. 2558-68.

85. Gennis, R.B., *Pores, Channels and Transporters*, in *Biomembranes: Molecular Structure and Function*, R.B. Gennis, Editor. 1989, Springer New York: New York, NY. p. 270-322.
86. Boyd, R., *Chapter 7 - Pumps, Channels, and Carriers: From "Active Patches" to Membrane Transport Proteins*, in *Foundations of Modern Biochemistry*, M.G. Ord and L.A. Stocken, Editors. 1998, JAI. p. 245-265.
87. Abdul Kadir, L., M. Stacey, and R. Barrett-Jolley, *Emerging Roles of the Membrane Potential: Action Beyond the Action Potential.* Front Physiol, 2018. **9**: p. 1661.
88. Fyles, T.M., *Synthetic ion channels in bilayer membranes.* Chem Soc Rev, 2007. **36**(2): p. 335-47.

Chapter 2

Methodology

2.1 Molecular dynamics simulations

The study of a system at a microscopic level often contributes to understanding the system at a macroscopic level. The classical simulations can be used to predict the quantitative and qualitative details of these macroscopic properties based on the structural and dynamical properties of the system at an atomistic level. The molecular dynamics (MD) simulations based on Newtonian dynamics are deterministic in nature, where the atomistic motion is predicted at each instant based on the net force acting on the atoms. MD simulations employ the usage of force fields to define the potential acting on the atoms based on inter-atomic interactions. The general description of a force field is shown below.

$$U = \sum_{bonds} \frac{k_{bond}}{2}(r - r_{eq})^2 + \sum_{angle} \frac{k_{angle}}{2}(\theta - \theta_{eq})^2 + \sum_{torsions} \frac{k_{dihedral}}{2}[1 + \cos(n\emptyset - \gamma)] + \sum_{i>j}\left[\left\{\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}\right\} + \frac{q_i q_j}{\varepsilon R_{ij}}\right],$$

Eq. 2.1

where, k_{bond}, k_{angle}, and $k_{dihedral}$ are the force constants representing the bond, angle, and dihedral potentials, respectively. The equilibrium values for a given bond and angle are represented by r_{eq} and θ_{eq}, respectively. The multiplicity and phase angle for the dihedral angles is indicated by *n* and γ, respectively. Thus, the equation's first three terms contribute to the bonded interactions' potential. The last term in the equation represents the non-bonded interaction between the atoms. The non-bonded interaction is defined as a combination of Lennard-Jones and Coulomb potential. A_{ij} and B_{ij} terms indicate the atom-specific parameters for atoms *i* and *j* for defining the Lennard-Jones potential. q_i, q_j, and ε indicate the partial charges on atoms *i* and *j* and the permittivity of the medium, respectively, for defining the Coulomb potential.

The timescales for various biological processes range from picoseconds (electron transfer process in photosynthesis) to milliseconds (e.g., Protein folding).[1] MD Simulations have become a reliable tool for capturing various biomolecular processes in life sciences. However, predicting and quantifying mechanistic aspects of various rare biomolecular events (e.g.,

protein folding, receptor-ligand interactions) by MD simulations is limited by high computational cost. In order to tackle the timescale constraints associated with these processes, enhanced sampling methods are used to improve the sampling efficiency of biomolecular simulations.

2.2 Enhanced Sampling Techniques

Free energy is crucial for understanding and describing biomolecular recognition processes due to its role in determining the feasibility and spontaneity of these interactions. In principle, the free energy of a state can be determined by assessing the population count associated with a specific state of interest described as follows.

$$G(x) = -K_B T \, ln \; P(x) \qquad \textit{Eq. 2.2}$$

where $P(x)$ is the population of the state at some coordinate x. Enhanced sampling techniques are a specefic class of techniques used in MD simulations to efficiently explore the conformational space of diverse systems, including biomolecular and material systems.[2] These techniques yield valuable insights into the associated free energy landscape, shedding light on the events of interest. Moreover, these techniques can accelerate the sampling and enable the study of rare events, where free energy barriers higher than thermal energy (~2.5 kJ/mol), which are not usually visited by standard molecular dynamics simulations due to limitation in computational resources.

Among various enhanced sampling methods, biasing methods and replica exchange methods are two relevant categories of enhanced sampling techniques. In the biasing methods, an external bias potential is applied to the MD simulation to encourage the system of interest to explore specific regions of the phase space. Some of the popular biasing methods are umbrella sampling (US)[3], metadynamics[4] and its variations[5], adaptive biasing force (ABF)[6] methods, etc. In the case of replica exchange methods, multiple copies or replicas of the system at different temperatures are used and they periodically swap configurations. Popular replica exchange methods include replica exchange molecular dynamics [7] and parallel tempering (PT)[8]. The enhanced methods have led to significant advances and contributions in the scientific community in understanding various recognition processes that were previously limited by timescale problems in computational studies.

A bias potential is commonly applied along the reaction coordinate, often is a collection of coordinates of several atoms and referred to as collective variable (CV), in these accelerated

methods. CVs are descriptors that can identify interesting collective phenomena over long timescales and separate macroscopically different structures.[9] These variables can simplify the complex dynamics of large molecular systems and focus on the essential aspects of interest. By focusing on the defined CVs specific to a system, simulations can capture relevant information and facilitate the study of molecular recognition processes in biological systems.

In this thesis, we have investigated the biomolecular recognition process in protein and RNA-based systems with the aid of enhanced sampling methods such as metadynamics[4], well-tempered metadynamics[10], Multiple walker metadynamics[11], parallel bias metadynamics[11], and a combination of umbrella sampling and parallel bias metadynamics[12] for understanding the mechanistic aspects associated with these processes. We have primarily employed intuitive CVs to gain insights into biomolecular processes. The CVs used in this thesis that are specific to each system will be thoroughly discussed in each chapter. PLUMED[13]patched with GROMACS[14, 15]was utilized for conducting all free energy calculations employing various enhanced sampling techniques. More detailed explanations are provided below about the enhanced sampling techniques used in the thesis.

2.2.1 Metadynamics

Metadynamics[4] is an advanced molecular dynamics simulation technique that can accelerate the rare events in complex molecular systems and provide an estimation of free energy described by a few CVs. In this method, a Gaussian potential is added to the real energy landscape of the system to accelerate the dynamics along the CV space. The final free energy landscape is obtained by summing up the Gaussian potentials added once the system starts fluctuating along the predefined CV space. The sampling efficiency depends on the parameters Gaussian height, Gaussian width, and the time interval between the addition of Gaussian functions that defines the metadynamics potential. During the simulation, the adaptive construction of the metadynamics bias potential aims to explore the regions in the CV space that have not already been visited. Here, the bias deposited to the system in the CV space at a given time is given by,

$$V^b(s,t) = \sum_{\tau < t} w_\tau \; exp\left[- \frac{\{(s - s(\tau)\}^2}{2(\delta s)^2} \right], \qquad Eq.\ 2.3$$

where $V^b(s,t)$ is the metadynamics potential as a function of s (conformational space described by the CV) at time *t*. w_τ and δs represents the Gaussian height and width parameters, respectively to, at a time interval τ. Metadynamics is an adaptive biasing technique by which the bias potential is updated in response to the evolving simulation trajectory, enabling effective sampling of unexplored regions.[4] Hence this technique has the advantage of not requiring a rough estimation of free energy profiles to explore the phase space. Once the simulation has reached convergence, the added bias potential estimates the free energy as a function of the defined CVs. After exploring all the relavent states and filling the wells, the biasing potential is anticipated to serve as a reliable approximation for the underlying free energy surface ($F(s)$), denoted as follows.

$$F(s) = -V_b(s, t \to \infty) \qquad \textit{Eq. 2.4}$$

In standard metadynamics, a fixed hill height is employed for bias deposition, introducing the potential risk of it becoming excessively dominant, leading to over-biasing and impeding the effective exploration of the conformational space.[5] Indeed, when the biasing potential fully compensates for *F(s)*, introducing additional Gaussians introduces errors. As a result, the value of $-V_b(s, t)$ fluctuates around the accurate *F(s)* value.

2.2.2 Well-tempered metadynamics

Well-tempered metadynamics[10] addresses specific limitations of standard metadynamics. This method utilizes scaled Gaussian hill height to add potential in the CV space. This enables a seamless convergence to a precise approximation of *F(s)*, providing enhanced control over the introduced biasing potential and thereby improving the overall efficiency of the simulation. The scaled hill height in the well-tempered metadynamics can be written as follows.

$$w_\tau = \omega_0 \tau_0 \, exp\left[-\frac{V^b(s,t)}{k_b \Delta T}\right], \qquad \textit{Eq. 2.5}$$

where, ω_0 denotes the initial rate of bias deposition, τ_0 represents the time step at which Gaussian potentials are added, and ΔT is a tunable parameter, determining the rate at which hill height diminishes as the wells become filled. The final value of the biasing potential is a proportionally adjusted approximation to *F(s)*, as described below.

$$F(s) = -\frac{T+\Delta T}{\Delta T} V_b(s, t \to \infty) \qquad \textit{Eq. 2.6}$$

2.2.2 Parallel bias metadynamics

When multiple CVs are needed for biasing, the efficiency of exploration in constructing the high-dimensional free energy landscape decreases exponentially with the number of CVs. Parallel bias metadynamics is an enhanced sampling method that facilitates the crossing of high barrier regions associated with high-dimensional free-energy profiles by applying simultaneous multiple low-dimensional bias potentials in the CV space. In this context, the individual low-dimensional free energies are computed directly from the applied bias potentials. Moreover, the entire high-dimensional free energy can be easily reconstructed through the use of standard reweighting techniques. In the parallel bias version of metadynamics (PBMetaD)[16], the bias contribution is defined by several one-dimensional bias potentials, each acting on a different CV s_i. The PBMetaD potential is defined by

$$V^{pb}(s_1, \dots, s_n, t) = -\frac{1}{\beta}\ln \sum_{i}^{n} \exp\left[-\beta V_i^b\ (s_i, t)\right], \qquad Eq.\ 2.7$$

where $\beta = (k_b T)^{-1}$ and V_i^b is given by Eq. 2.3. The Gaussian height introduced in a given dimension is computed using feedback received from other dimensions, as outlined below.

$$w_i(\tau) = w_i(0) \exp\left[-\frac{V_i^b\ (s_i, \tau)}{k_B \Delta T}\right] P_i(s_i), \qquad Eq.\ 2.8$$

where,

$$P_i(s_i) = \frac{exp\left[-\beta V_i^b(s_i, t)\right]}{\sum_j^n exp\ \left[-\beta V_j^b(s_j, t)\right]}\ , and\ i = 1, \dots., n \qquad Eq.\ 2.9$$

$P_i s(i)$ serves as a feedback function that adjusts the Gaussian height according to the bias introduced along other CV.

2.2.3 Multiple walkers approaches

In traditional metadynamics based methods (metadynamics, well-tempered metadynamics, parallel bias metadynamics etc.) as discussed above, a single walker explores the energy landscape by adding a history-dependent bias to the potential energy surface. This bias is typically in the form of Gaussian potentials added at specific points in the coordinate space, preventing the system from getting trapped in local energy minima. Multiple walker approaches for metadynamics involves the use of multiple independent walkers that explore the energy landscape simultaneously.[11] Each walker experiences its own set of biases,

allowing for more efficient exploration of the CV space. This approach can accelerate the sampling, speed up convergence, and improve the overall efficiency of the simulation. It's important to highlight that, in the presented multiple walker metadynamics, the bias experience on all walkers contributes to the reconstruction of the free energy surface (FES). In the multiple walkers approaches, each walker samples a different trajectory, and the walkers communicate key information, such as bias potentials and collective variable distributions, to enhance the exploration of unvisited regions in the CV space.[11, 17] Further, this communication helps in overcoming energy barriers associated with free energy landscape. Multiple walkers approaches can be clubbed with traditional metadynamics based techniques like well-tempered metadynamics, parallel bias metadynamics by taking different starting configurations or copies of the same configuration into consideration.

2.2.4 Combination of umbrella sampling and parallel bias metadynamics

This method is proposed to overcome some of the drawbacks associated with the metadynamics-based methods. Sometimes, due to adaptive sampling, inconsistent sampling, lack of convergence of simulations, an existence of false minima state is observed in metadynamics simulations, especially for the systems associated with high dimensional free energy landscape. These challenges are closely tied to the slow dynamics of the system along multiple CVs, which remain unexplored during biased sampling involving only two or three CVs. Barriers along these orthogonal CVs can constrain the exploration of significant phase space volumes, severely impacting ensemble averages and free energy estimations.[18-20]. The controlled sampling approach takes advantage of two enhanced sampling methods (umbrella sampling and parallel bias metadynamics) to overcome the challenges associated with independent metadynamics simulations by controlled exploration of phase space.[12] This methodology draws inspiration from the WS-MTD[19] and PBTASS[20] methods. In this approach, the CV space is characterized by $s = (s_1, s_2, \dots, s_n)$.[12] The method consist of utilization of two sets of CVs; umbrella sampling coordinate and subsidiary collective variables. The umbrella sampling potential is applied along the primary CV coordinate s_1, which provides the direction of sampling. In a given umbrella sampling window, the harmonic bias is incorporated into the system using the following expression.

$$w_h^b(s) = \frac{1}{2}\,\kappa_h(s - s_h)^2\,, \qquad h = 1, \dots M \qquad Eq.2.10$$

where κ_h and s_h represent the force constant and mean position, respectively, of the added harmonic potential. Additionally, bias contribution from parallel bias metadynamics along the other subsidiary collective variables $(s_2, \dots, s_n)$ is experienced in each umbrella sampling window *h* as described in the Eq. 2.7. Subsidiary CVs are expected to accelerate the sampling along each umbrella sampling window. Thus, the cumulative contribution of bias to the overall system is expressed as:

$$W_h^b(s_1) + V_h^{pb}(s_2, \dots, s_n, t), \qquad h = 1, \dots, m \qquad Eq.2.11$$

where, *h* denotes the number of windows considered for the controlled sampling approach. $W_h^b(s_1)$ and $V_h^{pb}(s_2, \dots, s_n, t)$ are described by the Eq. 2.10 and 2.7, respectively. Thus the modified Hamiltonian can be written as follows.

$$H_h(R,P) = H^0(R,P) + W_h^b(s_1) + V_h^{pb}(s_2, \dots, s_n, t), \qquad h = 1, \dots, m \quad Eq.2.12$$

where $H^0(R,P)$ defines the underlying potential energy of the system without any external bias. To get the unbiased probability distribution, the reweighting procedure is executed in the following manner. Trajectories from individual windows are concatenated to form a global trajectory, allowing the calculation of the bias experienced by each frame within the trajectory. The metadynamics bias potential, contributed by subsidiary collective variables in each of these independent runs, is then reweighted above each frame in the global trajectory using the final bias-based reweighting approach for parallel bias. Subsequently, this reweighted metadynamics potential is combined with the umbrella bias potential and the weight per frame by self-consistently solving the WHAM equations.[21]

2.3 References

1. Shamir, M., et al., *SnapShot: Timescales in Cell Biology.* Cell, 2016. **164**(6): p. 1302-1302 e1.
2. Hénin, J.r.m., et al., *Enhanced Sampling Methods for Molecular Dynamics Simulations [Article v1.0].* Living Journal of Computational Molecular Science, 2022. **4**(1).
3. Torrie, G.M. and J.P. Valleau, *Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling.* Journal of Computational Physics, 1977. **23**(2): p. 187-199.
4. Laio, A. and M. Parrinello, *Escaping free-energy minima.* Proc Natl Acad Sci U S A, 2002. **99**(20): p. 12562-6.

5. Bussi, G. and A. Laio, *Using metadynamics to explore complex free-energy landscapes.* Nature Reviews Physics, 2020. **2**(4): p. 200-212.
6. Darve, E., D. Rodriguez-Gomez, and A. Pohorille, *Adaptive biasing force method for scalar and vector free energy calculations.* J Chem Phys, 2008. **128**(14): p. 144120.
7. Sugita, Y. and Y. Okamoto, *Replica-exchange molecular dynamics method for protein folding.* Chemical Physics Letters, 1999. **314**(1-2): p. 141-151.
8. Earl, D.J. and M.W. Deem, *Parallel tempering: theory, applications, and new perspectives.* Phys Chem Chem Phys, 2005. **7**(23): p. 3910-6.
9. Noe, F. and C. Clementi, *Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods.* Curr Opin Struct Biol, 2017. **43**: p. 141-147.
10. Barducci, A., G. Bussi, and M. Parrinello, *Well-tempered metadynamics: a smoothly converging and tunable free-energy method.* Phys Rev Lett, 2008. **100**(2): p. 020603.
11. Raiteri, P., et al., *Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics.* J Phys Chem B, 2006. **110**(8): p. 3533-9.
12. Vijay, A. and A. Mukherjee, *Molecular insights into the stereospecificity of arginine in RNA tetraloop folding.* Phys Chem Chem Phys, 2023. **25**(16): p. 11301-11310.
13. Tribello, G.A., et al., *PLUMED 2: New feathers for an old bird.* Computer Physics Communications, 2014. **185**(2): p. 604-613.
14. Abraham, M.J., et al., *GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers.* SoftwareX, 2015. **1-2**: p. 19-25.
15. Hess, B., et al., *GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation.* Journal of Chemical Theory and Computation, 2008. **4**(3): p. 435-447.
16. Pfaendtner, J. and M. Bonomi, *Efficient Sampling of High-Dimensional Free-Energy Landscapes with Parallel Bias Metadynamics.* J Chem Theory Comput, 2015. **11**(11): p. 5062-7.
17. Sucur, Z. and V. Spiwok, *Sampling Enhancement and Free Energy Prediction by the Flying Gaussian Method.* J Chem Theory Comput, 2016. **12**(9): p. 4644-50.
18. Acharya, A., J.D. Prajapati, and U. Kleinekathofer, *Improved Sampling and Free Energy Estimates for Antibiotic Permeation through Bacterial Porins.* J Chem Theory Comput, 2021. **17**(7): p. 4564-4577.

19. Awasthi, S., V. Kapil, and N.N. Nair, *Sampling free energy surfaces as slices by combining umbrella sampling and metadynamics.* J Comput Chem, 2016. **37**(16): p. 1413-24.
20. Gupta, A., et al., *Exploration of high dimensional free energy landscapes by a combination of temperature-accelerated sliced sampling and parallel biasing.* J Comput Chem, 2022. **43**(17): p. 1186-1200.
21. Kumar, S., et al., *THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method.* Journal of Computational Chemistry, 1992. **13**(8): p. 1011-1021.

Chapter 3

Potential therapeutics for COVID-19 by targeting spike protein of SARS-CoV-2

3.1 Overview

The urgency and extensive impact of COVID-19 underscored the need for the development of innovative drugs to meet the distinct challenges posed by the SARS-CoV-2 virus, which was responsible for the pandemic. The study aims to identify potential therapeutics for COVID-19, specifically targeting the receptor-binding domain (RBD) of the spike protein of SARS-CoV-2 subjected to rigorous free energy validation studies using enhanced sampling studies. For this purpose, we have employed an in-house made de novo drug generation algorithm that generates molecules from scratch, growing atom by atom through the optimization of their interaction energy with any specified receptor. Moreover, we mapped the generated de novo molecules to present-day drugs from the DrugBank library and chose the best candidates for further analysis based on the docking score. We have validated the binding affinity of all the considered 55 molecules (35 de novo and 20 present-day molecules) by free energy calculations using all-atom explicit-water molecular dynamics simulations aided by well-tempered metadynamics We also showed that the free energy estimate of human angiotensin-converting enzyme 2 (hACE2) and RBD interaction (a protein-protein interaction) is comparable with the experimentally reported values, which validates our simulation protocol. We have identified a minimum of nine molecules exhibiting a higher affinity for binding to the RBD than hACE2. This suggests the possibility of these molecules to disrupt the RBD-hACE2 interaction which is responsible for the spread of viral infection.

3.2 Introduction

SARS-CoV-2, the source of the COVID-19 pandemic, is a beta-coronavirus characterized by a positive-strand RNA genome.[1] These viruses possess a viral envelope comprised of three proteins: the membrane protein, the envelope protein, and the spike protein.[2, 3] The spike proteins bind to the human angiotensin-converting enzyme 2 (hACE2) receptor present on the surface of cells.[4, 5] This interaction facilitates the subsequent entry of the virus into the host cell.[6, 7] Upon entry into the host cell, the virus disrupts the cell's replication machinery, enabling it to propagate and spread throughout the body. Understanding the spike protein-ACE2 interactions is noted as an important step towards treatments and vaccines for COVID-19[8]. While the trimeric spike protein exists in both open and closed forms, the monomeric Receptor Binding Domain (RBD) of the spike protein is found to play a crucial role in the recognition process.[6] The structure of the RBD complexed with hACE2 determined by X-ray diffraction studies was reported[9] and the studies suggests that the RBD region of the spike protein remains a viable target for drug design.[8, 10] Several computational studies have identified diverse small molecule candidates capable of potentially inhibiting the spike protein, pending experimental validation.[11-15] The trimeric structure of the spike protein along with hACE2 is shown in the Fig. 1. The binding constant (K_d) of these RBD-hACE2 interactions is reported to be 34.6 nM by biolayer inferometry. [7] Further, surface plasmon experiments and non-covalent association methods reported the binding affinity for this interaction in the range of 44.2 nM and 2.9 nM, respectively.[6, 16] Thus, the reported studies based on these observations indicate that the free energy stability for these protein-protein complex is in the range between -10.3 kcal/mol and -12.3 kcal/mol. The effectiveness of inhibiting the RBD-hACE2 complex can be achieved if a potential molecule can establish a better binding affinity with the RBD compared to the binding affinity observed between the RBD and hACE2 (Fig. 1). Hence, the primary objective of this study is to employ computational approaches to identify a molecule capable of binding to the RBD region with a binding free energy lower than -10.3 kcal/mol.

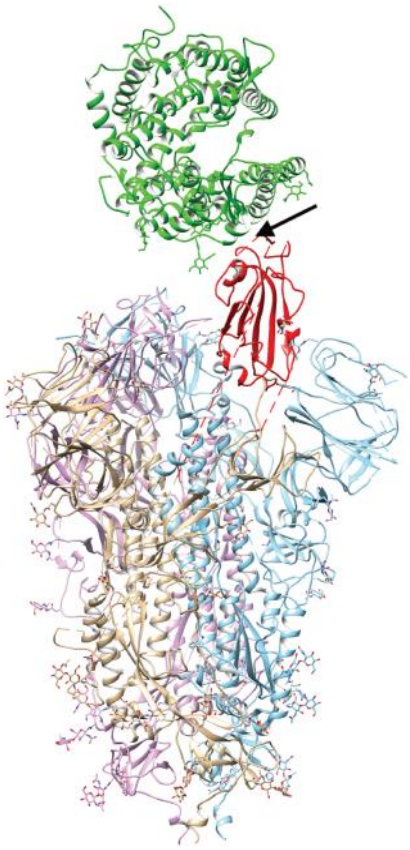

Figure 1. The trimeric structure of the spike protein along with hACE2 (PDB ID: ID 6VSB[7]). The RBD region is shown in red, and hACE2 is in green. The arrow highlights the target site for designing a new inhibitor.

To accomplish this, we have utilized an in-house made de novo molecule generation program that constructs molecules atom by atom within the protein's hotspot.[17] The program focuses on optimizing the interaction energy between the generated molecules and the protein's hotspot.[17] Along with the generated molecules, we also selected additional analogous drugs from the DrugBank[18] library using a similarity-based mapping, aiming for potential repurposing. Further, to understand the binding affinity of all the selected molecules (35 de novo and 20 repurposable drugs all-atom, explicit water well-tempered metadynamics[19] free energy calculations were carried out. Based on our simulation protocol, we observed that nine molecules, comprising four de novo and five repurposable compounds, exhibited lower binding free energy than that observed in the RBD-hACE2 interaction.[17] This indicates the considerable potential of these molecules to efficiently inhibit the viral attack. Furthermore, we also calculated the free energy of binding associated with RBD-hACE2 interactions using well-tempered metadynamics[19] simulations to validate the simulation protocol, and the outcomes were in accordance with experimental observations.

3.3 Computational details

3.3.1 Selection of molecules

We employed an in-house de novo molecule generation program that constructs molecules atom by atom from scratch within a predefined hotspot of the receptor. This approach allowed us to generate molecules around the hotspot of the RBD.[12] We have further shortlisted all de novo-generated molecules that interacted with the protein with energies below our specified cutoff of -61.5 kcal/mol. Following this, a similarity search was conducted using the DrugBank[18], utilizing the Tanimoto algorithm[20]. All approved and investigational drug molecules with a similarity score exceeding 0.4 were considered. Based on the similarity measurements, the selected drug molecules were docked to the hotspots by employing AutoDock Vina version 1.1.2[21]. A box centered around residue 443, was generated with dimensions of 36.7 Å x 26.1 Å x 40.9 Å, using default AutoDock Vina parameter. Drug molecules with docking scores less than -8.0 kcal/mol were subsequently considered for free energy validation studies.

3.3.2 Modeling the systems

All molecules underwent optimization using Hartree Fock theory with a 6-31G* basis set employing Gaussian 09 software[22]. Subsequently, antechamber[23] module was utilized for the generation of force fields for the molecules based on RESP charge calculations. Further, the topology was converted into the GROMACS acceptable format using the a python script acpype.py, available at https://github.com/t-/acpype. The initial configuration of the SARS-CoV-2 spike glycoprotein (PDB ID: 6VSB) was acquired from the Protein Data Bank, and Modeller 9.21[24] was employed to model the missing residues. In investigating the protein-ligand complexes, we have considered the RBD region of the protein, spanning residues 302 to 506. The topology was prepared utilizing the AMBER99SB[25] force field. Every protein-ligand complex system was solvated by approximately 23,000 TIP3P[26] water molecules within a box measuring 70 x 70 x 180 $Å^3$. To maintain physiological conditions, sodium ions (Na^+) and chloride ions (Cl^-) at a concentration of 150 mM were added, along with additional chloride ions to neutralize the system. For carrying out molecular dynamics simulations and subsequent free energy calculations, the docked pose was utilized as the initial point for repurposable molecules. In contrast, the final output complex structure following the execution of the de novo algorithm was employed for de novo-generated molecules

Table 1. The system size and simulation durations for the metadynamics simulations across all systems are provided. Metadynamics simulations were not conducted for the unstable system, and their runtime is indicated as "NA."

System	Runtime (ns)	System Size (number of atoms)	System	Runtime (ns)	System Size (number of atoms)
hACE2 (run 1)	234	236699	hACE2 (run 2)	51	236699
47_68 (run 1)	154	86305	Danoprevir	236	114053
47_68 (run2)	381	86305	Danoprevir	99	114053
37_42 (run 1	41	86337	Solithromycin	149	114129
37_42 (run2)	64	86337	Solithromycin	157	114129
45_23 (run 1)	115	114066	Saquinavir	44	86314
45_23 (run 2)	126	114066	Saquinavir (run 2)	104	86314
41_46 (run 1)	64	114089	Glecaprevir	166	86308
41_46 (run 2)	67	114089	Paritaprevir	106	114087
49_24	48	86318	Ciluprevir	118	86330
50_76	88	86329	Simeprevir	56	114028
50_21	156	114066	Paritaprevir	48	114087
45_24	72	86323	E7107	124	114121
44_18	109	114091	Ergotamine	240	114037
50_23	75	86324	GS-9256	156	86284
50_20	30	86318	Enviomycin	62.9	86351
48_44	92	86318	Nepadutant	53	86294
44_16	20	114068	Vaniprevir (run 1)	97	86309
47_27	63	114028	Vaniprevir (run 2)	151	86309
44_43	37	99557	Telcagepant	88	86316
46_24	16	114066	Pimecrolimus	73	114056
39_32	14	86356	Narlaprevir	10.4	86376
46_13	24	86310	Sacubitril	29	86338
46_12	13	86310	Taurocholic acid	24	86352
29_19	14	114079	Aplaviroc	52	86307
46_37	11	114066	Viomycin	6	86309
47_11	18	86318	49_38	NA	86318
47_26	23	86305	45_74	NA	114074
41_37	16	86283	44_15	NA	86298
50_82	14	86316	44_15	NA	86298
41_31	8	114039	39_57	NA	86330
44_81	3	86293	32_26	NA	86300
46_22	18	114066	40_35	NA	86327
48_59	7	114066			

3.3.3 Simulation details

Firstly, each system underwent energy minimization using the steepest descent method[27] for 10,000 steps. Subsequently, the system was heated to 300 K over 200 ps using the Berendsen thermostat and barostat[28] with a coupling constant of 0.6 ps. During the heating process, position restraints of 25 kcal/mol/Å² were imposed on heavy atoms of the protein and ligand. Thereafter, equilibration was conducted for 2 ns at a constant temperature of 300 K and pressure of 1 bar, without any restraints. The same thermostat and barostat with coupling constants of 0.2 ps each were employed for this process. The final 100 ps of the NPT simulation were utilized to compute the average volume, which was then employed in the subsequent 5 ns unrestrained NVT equilibration using the Nosé-Hoover thermostat[29] with a coupling constant of 0.2 ps. Throughout the simulation, the LINCS algorithm[30] was applied to constrain all bonds, and the Particle Mesh Ewald (PME) method[31] was employed for electrostatics. . The distance cut-offs for the van der Waals (vdW) and electrostatic long-range interaction was kept at 10 Å. The time step for each simulation was taken to be 2 fs.

3.3.4 Free energy calculations

The equilibrated structure of the ligand-bound protein complex underwent final production simulation for 5 ns. If, after this period, the ligand was observed to remain bound, we proceeded to perform free energy calculations on the system. Well-tempered metadynamics[19] simulations were carried out to compute the binding free energy of molecules, incorporating *distvec* and native contacts as collective variables (CVs). In the metadynamics simulation, a hill height of 0.2 kJ/mol, a bias factor of 10, and a hills deposition rate of 2 ps were used. The Gaussian widths for *distvec* and native contacts were set at 0.6 Å and 5, respectively. Additionally, an upper wall restraint was imposed at a 45° angle between two vectors, $\hat{b}$ and $\vec{d}$. The simulation lengths for all systems are detailed in Table 1. For free-energy calculations, PLUMED 2.6[32]patched with GROMACS[33, 34] was used.

3.3.5 Choice of collective variables

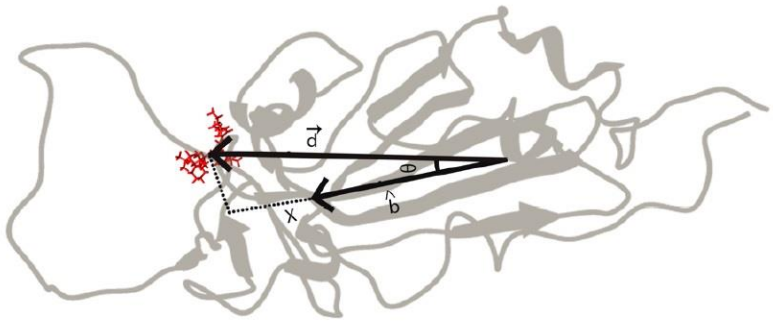

Figure 2. The CV distvec (X) and its components are defined in the picture. The RBD is shown in grey and ligand is shown in red.

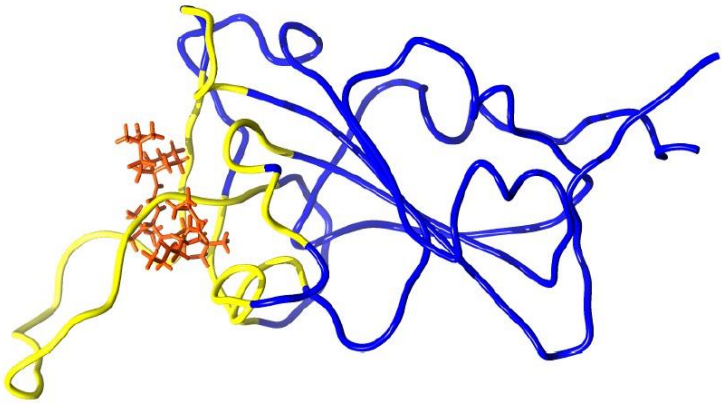

Figure 3. The CV native contact is shown. ***gA*** (ligand) is shown in orange and ***gB*** (part of the RBD) is shown in yellow.

a) collective variable *distvec (X)* : *distvec* is defined by two vectors $\hat{b}$ and $\vec{d}$. The $\hat{b}$ is the unit vector is the unit vector from the centre of mass (COM) of the residues ARG331, VAL369, TYR370, ALA371, LEU487, SER488, GLU490 to the COM of the residues SER323, VAL324, TYR325, VAL375, ILE376, ARG377, GLY378, ASP379, GLU380, VAL381, ARG382, GLN383, GLY390, LYS391, ILE392, TYR395, ASN396, SER412, ASN413, ASN414, LEU415, ASP416, SER417, TYR423, ASN424, TYR425, LEU426, TYR427, ARG428, PRO465, LEU466, GLN467, SER468, TYR469, GLY470, PHE471, GLN472, PRO473, THR474, VAL477, GLY478, TYR479, GLN480, PRO481, TYR482 that directed towards the hotspot region of the RBD. $\vec{d}$ is the vector from the COM of residues ARG331, VAL369, TYR370, ALA371, LEU487, SER488, GLU490 to the COM of selected small molecule. Further $X = \hat{b}.\vec{d}$, and $\theta = \cos^{-1}(\hat{b}.\vec{d}/|d|)$. A representation of the construction of CV is shown in the Fig. 2

b) Native Contacts (N_c): Native contacts (N_c) are defined by the spatial proximity of groups of atoms in the native state which is defined by two groups, *gA* and *gB. gA* consists of heavy atoms of the small molecule taken into consideration. Further, *gB* consists of heavy atoms of the RBD residues 376-377, 379, 382-383,389-392,394-396, 427-430, 440-470, 475, 477, 479.

$$N_c = \sum_{i \epsilon gA} \sum_{i \epsilon gB} s_{ij},$$

where, s_{ij} given by,

$$s_{ij} = \begin{cases} 1 & \forall \ r_{ij} \leq 0 \\ \left(1 - \left(\frac{r_{ij}}{r_0}\right)^n\right) \Big/ \left(1 - \left(\frac{r_{ij}}{r_0}\right)^m\right) & \forall \ r_{ij} > 0 \end{cases}$$

and $r_{ij} = |r_i - r_j| - d_0$. The user-defined parameters were chosen to be n= 6, m= 12, r_0= 5.5 Å and d_0=0 Å. The provided equation guarantees that the variation of s_{ij} remains both continuous and differentiable. A representation of residues of RBD involved in native contact formation is shown in Fig. 3.

3.4 Results & Discussion

3.4.1 RBD-hACE2 interaction

Prior to computing the free energy for all 55 drug molecules, we verified our simulation protocol by calculating the binding free energy associated with the RBD and hACE2 interactions. Well-tempered metadynamics[19] simulations were used for the binding free energy calculations using *distvec* and native contacts as CVs. The CV *distvec* pushes the two proteins apart and native contact aids in unraveling the interactions between two proteins. These CVs were previously successfully utilized for studying DNA-drug intercalation. [35, 36] Fig. 4 depicts the free energy surface associated with RBD and hACE2 interaction. As depicted in Fig. 4, when the proteins detach, there is an observable increase in free energy, accompanied by a significant rise in *distvec*. This also resulted in a consecutive decrease in native contacts. Based on this free energy surface, the estimated binding free energy is found to be -13.3 kcal/mol, a value that closely aligns with the experimental results (between -10.3 kcal/mol and -12.3 kcal/mol).[7, 16]

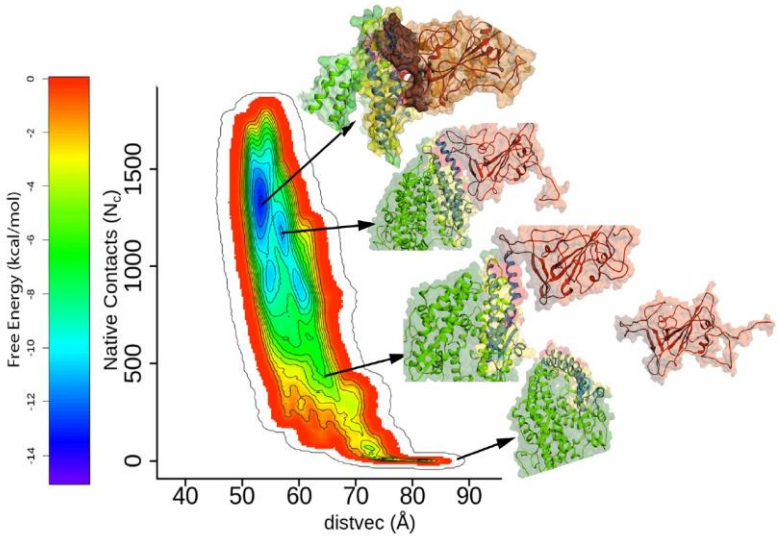

Figure 4. 2D Free Energy surface with respect to distvec and native contacts, illustrating the interactions between hACE2 and RBD. The structures along the unbinding pathway are displayed in the free energy plot.

3.4.2 RBD-small molecule interactions

From a computational perspective, molecules generated de novo exhibit strong binding characteristics, making some of them potential drug candidates. Nevertheless, the synthesis and toxicology studies for these molecules may require a significant time. Hence, the possible repurposable molecules (from DrugBank[18] using Tanimoto (or Jaccard)[37] similarity search) were also considered for the free energy validation study. The chemical structures of all the validated 55 molecules is shown in Figs. 5 and 6. The docking score of the selected repurposable molecule is shown in the Table 2. As we have validated the binding affinity similarly to the experimental results for the RBD-hACE2 interactions, we have followed the same simulation protocol for binding free energy calculations for small molecules towards the RBD region.

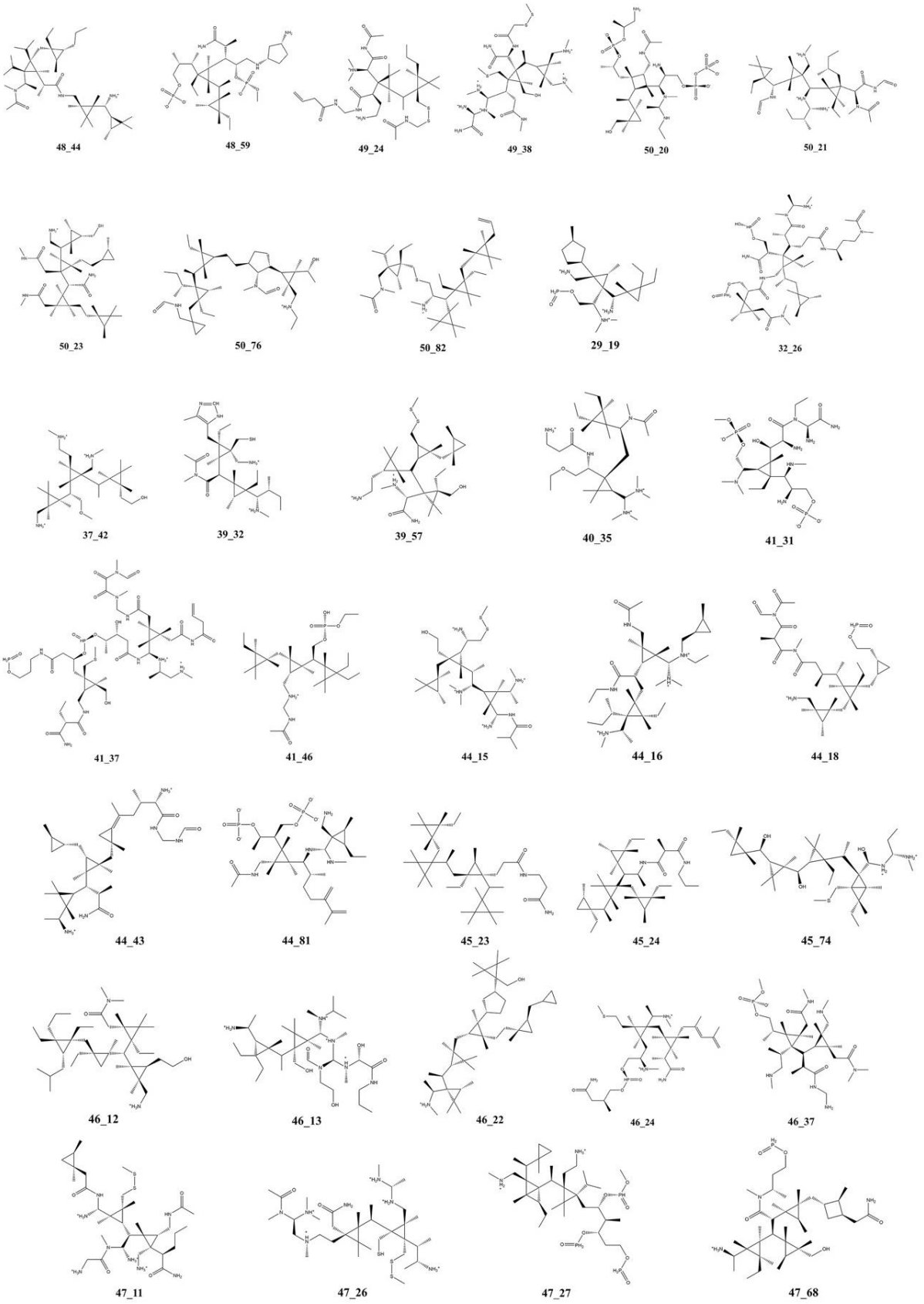

Figure 5. Chemical structures of the 35 de novo molecules considered for free energy validation.

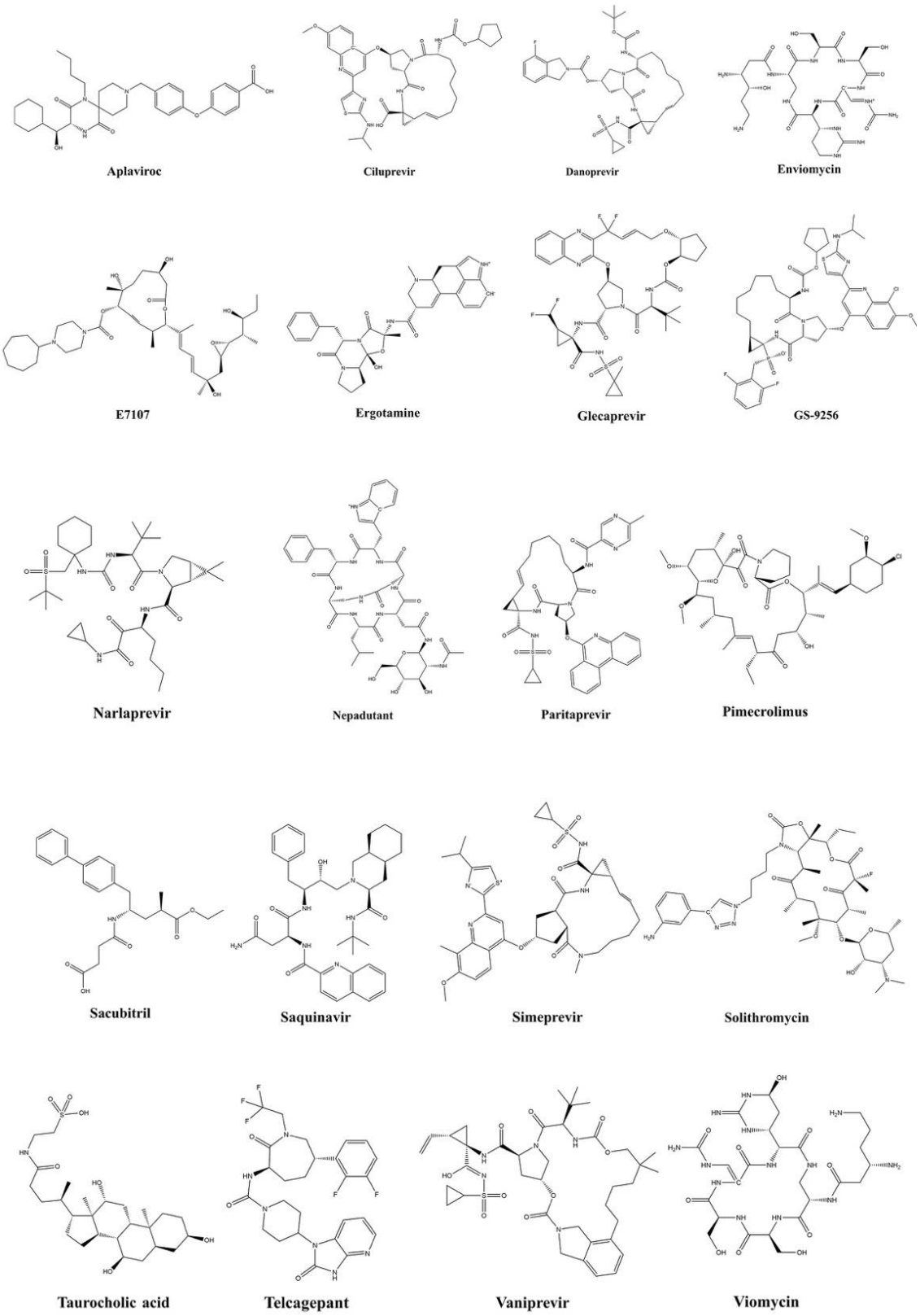

Figure 6. Chemical structures of the 20 drugs (obtained from DrugBank[18]) considered for free energy validation.

Table 2. Docking scores for the selected 20 drugs (obtained from DrugBank[18])

Sl. No	DrugBank ID	Drug Name	Docking Score (kcal/mol)
1	DB09297	Paritaprevir	-9.1
2	DB09308	Solithromycin	-9.0
3	DB11779	Danoprevir	-9.0
4	DB00337	Pimecrolimus	-8.7
5	DB00696	Ergotamine	-8.7
6	DB01232	Saquinavir	-8.7
7	DB12538	Nepadutant	-8.7
8	DB12876	GS-9256	-8.7
9	DB09292	Sacubitril	-8.4
10	DB12508	E7107	-8.4
11	DB14760	Narlaprevir	-8.4
12	DB04348	Taurocholic acid	-8.3
13	DB06290	Simeprevir	-8.3
14	DB12228	Telcagepant	-8.3
15	DB13879	Glecaprevir	-8.3
16	DB06827	Viomycin	-8.2
17	DB08993	Enviomycin	-8.2
18	DB11929	Vaniprevir	-8.2
19	DB05868	Ciluprevir	-8.1
20	DB06497	Aplaviroc	-8.1

Well-tempered metadynamics[19] simualations were conducted for 55 selected small molecules (comprising 35 de novo and 20 known drug molecules) by employing the same set of CVs as used for RBD-hACE2 interactions (*distvec* and native contacts). The free energy stability of all the validated molecules are shown in the Fig. 7. In Fig. 8, the free energy surfaces of the three leading de novo molecules and the top three drugs are presented along with the most stable configuration.

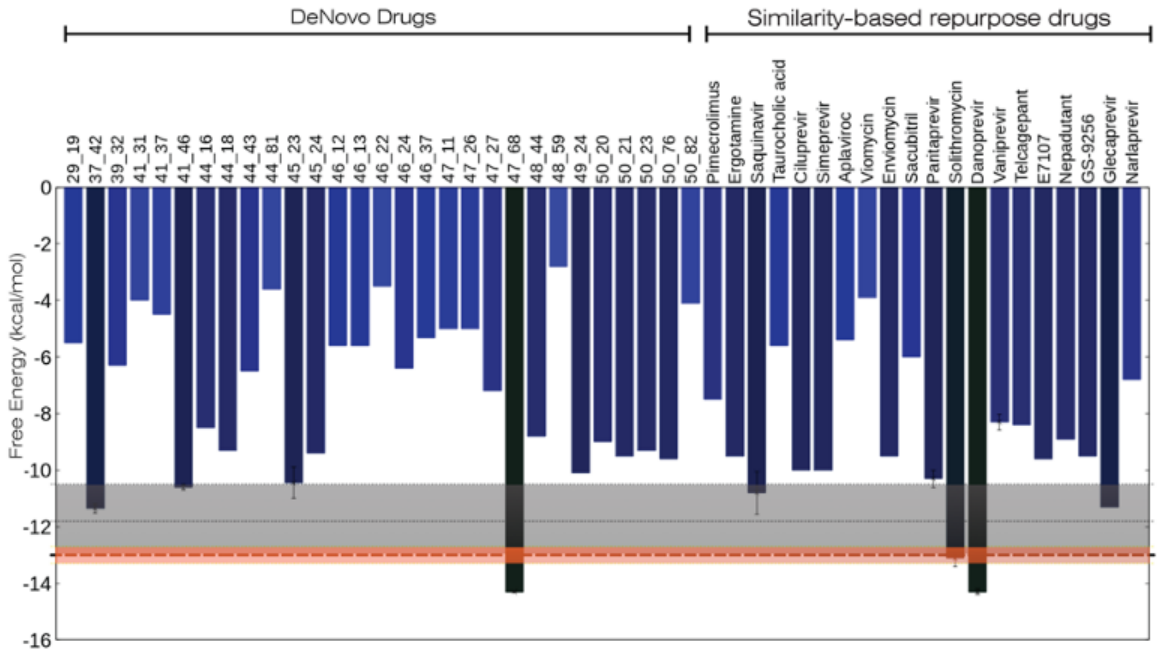

Figure 7. The free energy of binding for the 35 de novo molecules and 20 repurposable drugs to the RBD was determined through well-tempered metadynamics[19] simulations. The horizontal grey bar represents the experimental range of free energy binding between RBD and hACE2. The orange bar indicates the free energy estimate derived from multiple metadynamics simulations of RBD and hACE2. Error bars, obtained from multiple metadynamics simulations, are depicted for some of the strongly binding molecules.

As depicted in Fig. 8, the best binding molecules are well-fitted within the hotspot region by creating a cavity within the protein. The 2D free energy surfaces of all the validated molecules are shown in Fig. 9, 10 and 11 respectively.

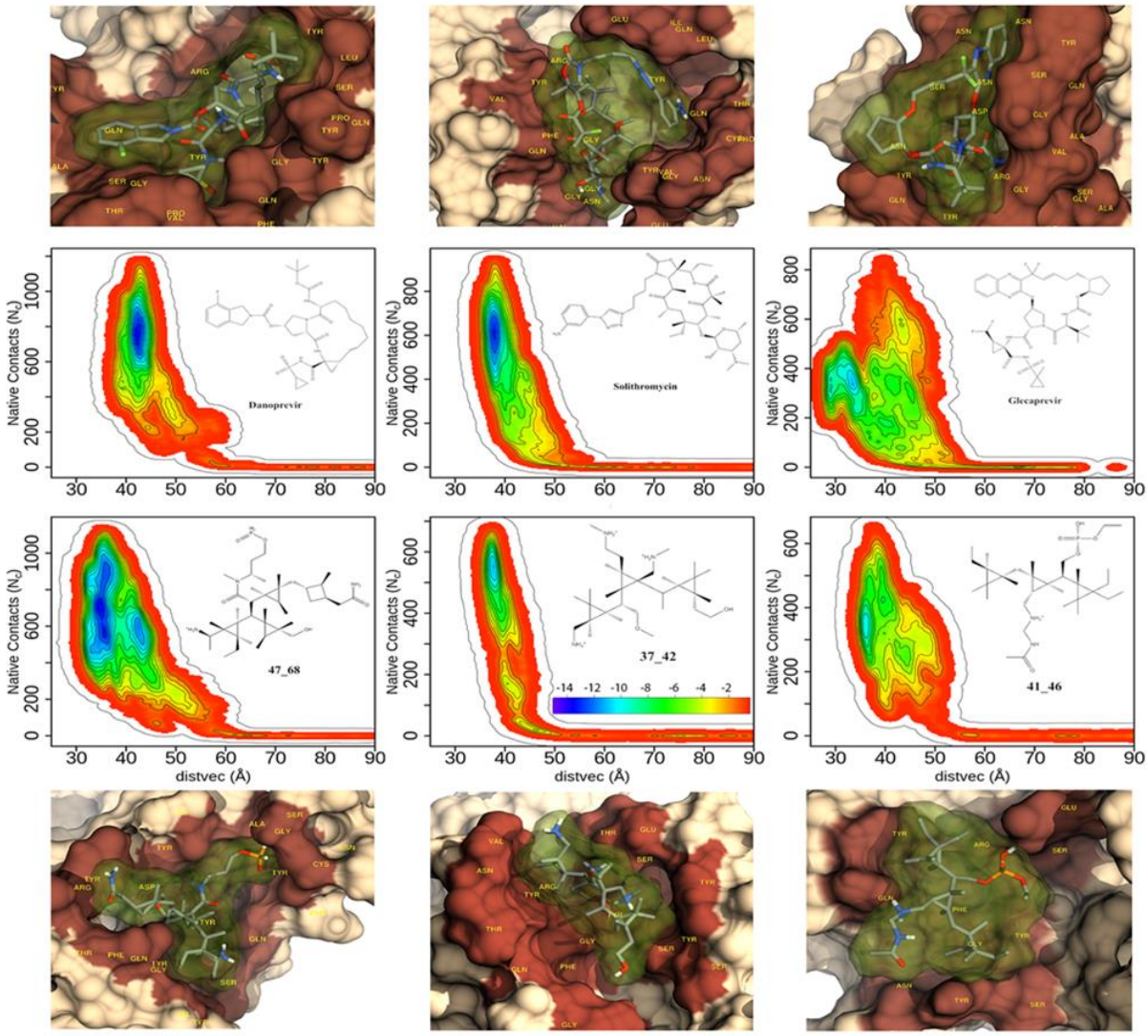

Figure 8. 2D free energy surfaces for the best binding small molecules with each molecule's chemical structure and names (displayed in the inset). The free energy bar used for plotting the free energy surface is shown in the inset of 37_42. Additionally, the free energy minimum structures for each molecule, created using Chimera[38], are presented above or below the corresponding free energy surfaces.

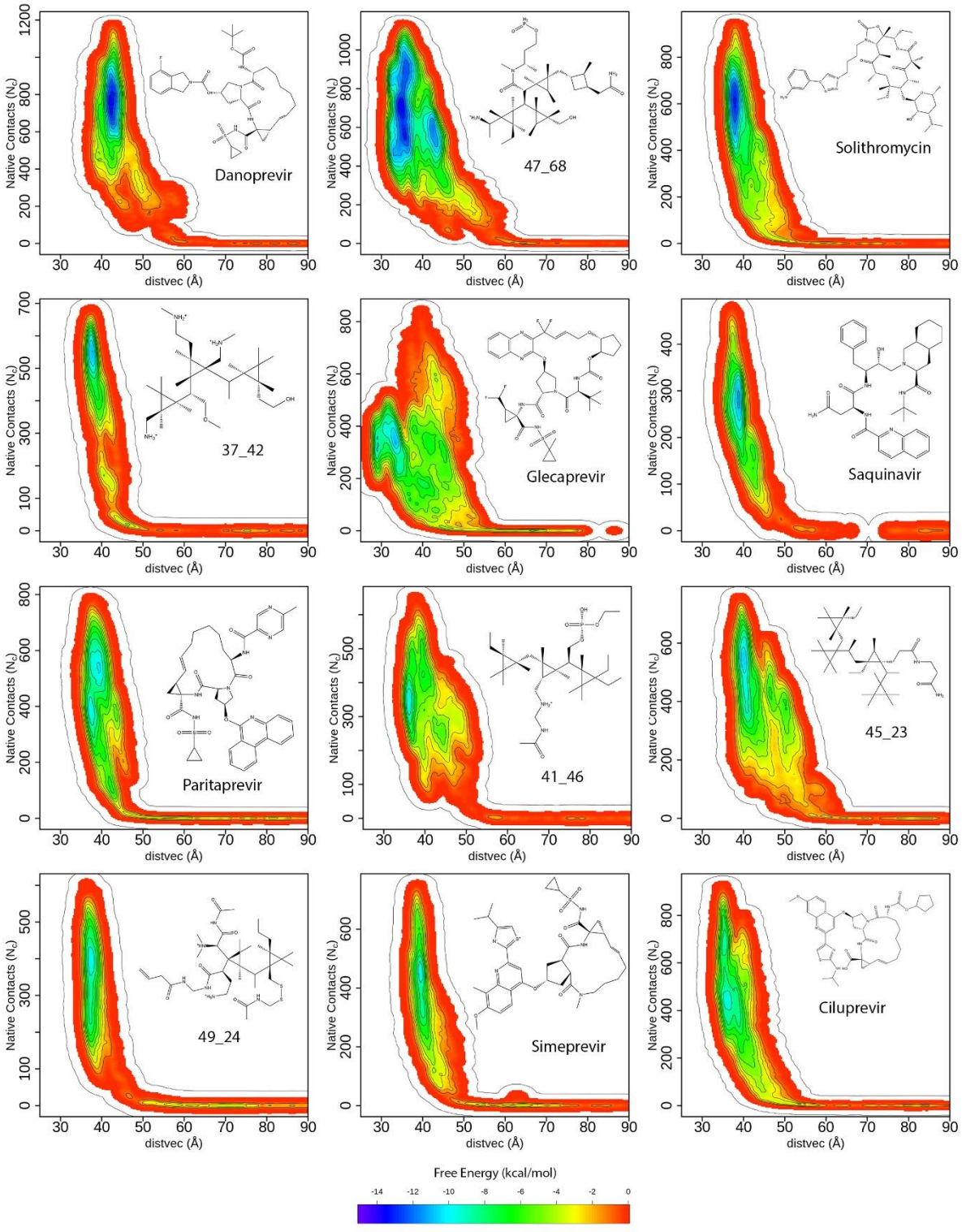

Figure 9. Free energy surfaces of de novo and drug molecules with free energy stability between -14.4 kcal/mol and -9.9 kcal/mol. The chemical structure is shown in the inset of the respective plot.

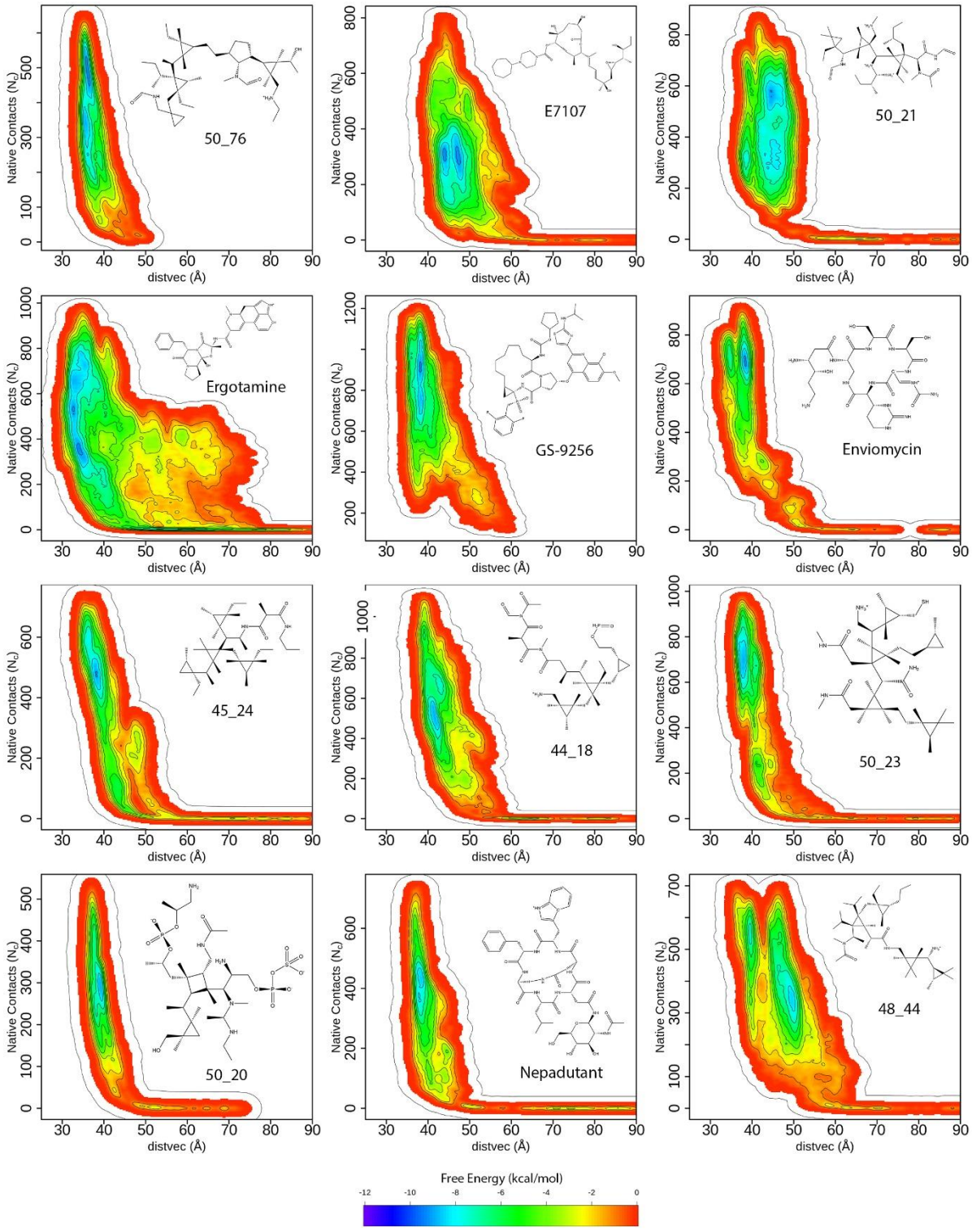

Figure 10. Free Energy surfaces of de novo and drug molecules with free energy stability between -9.8 kcal/mol and -8.7 kcal/mol. The chemical structure is shown in the inset of the respective plot.

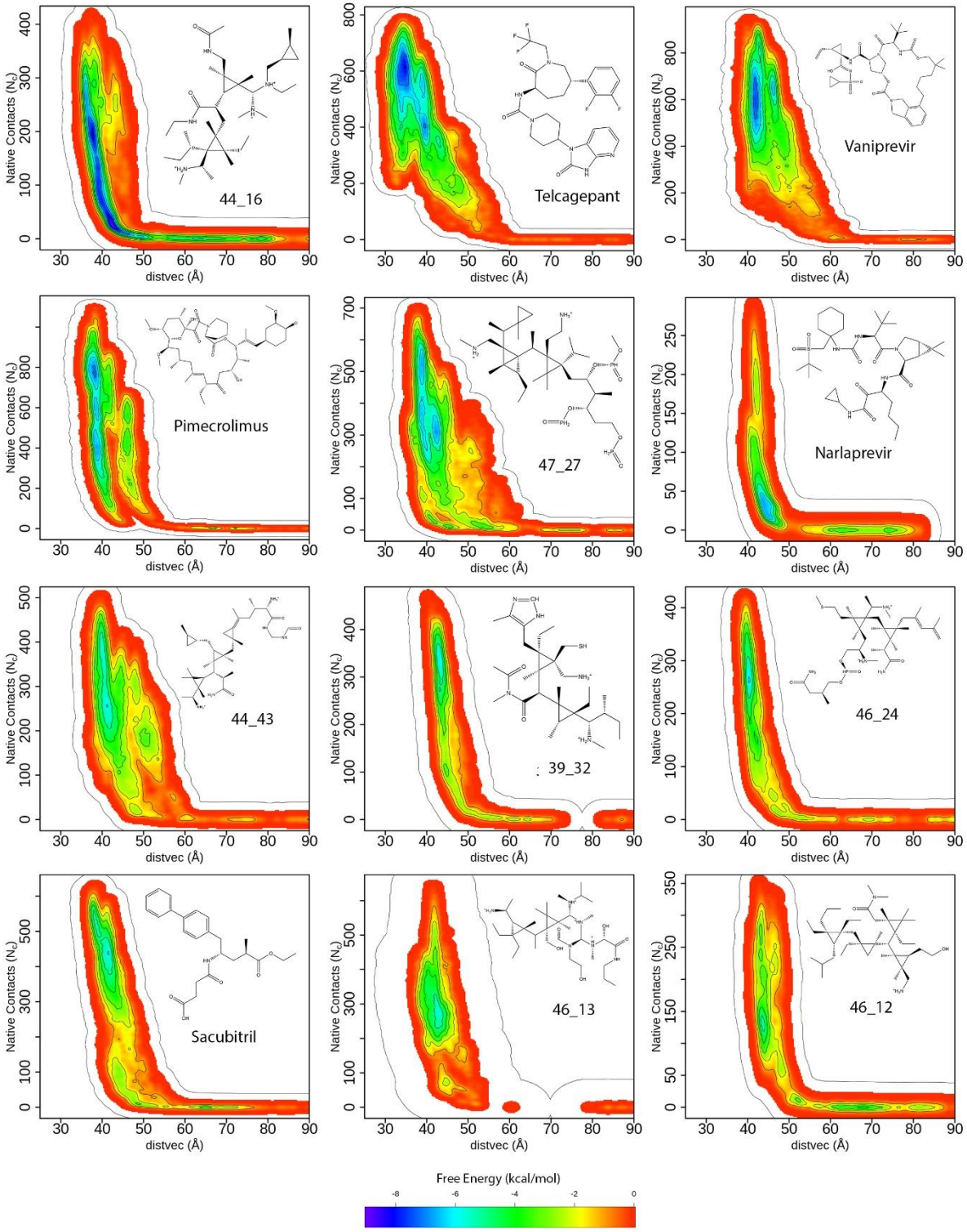

Figure 11. Free Energy surfaces of de novo and drug molecules with free energy stability between -8.6 kcal/mol and -5.6 kcal/mol (both ends are included). The chemical structure is shown in the inset of the respective plot.

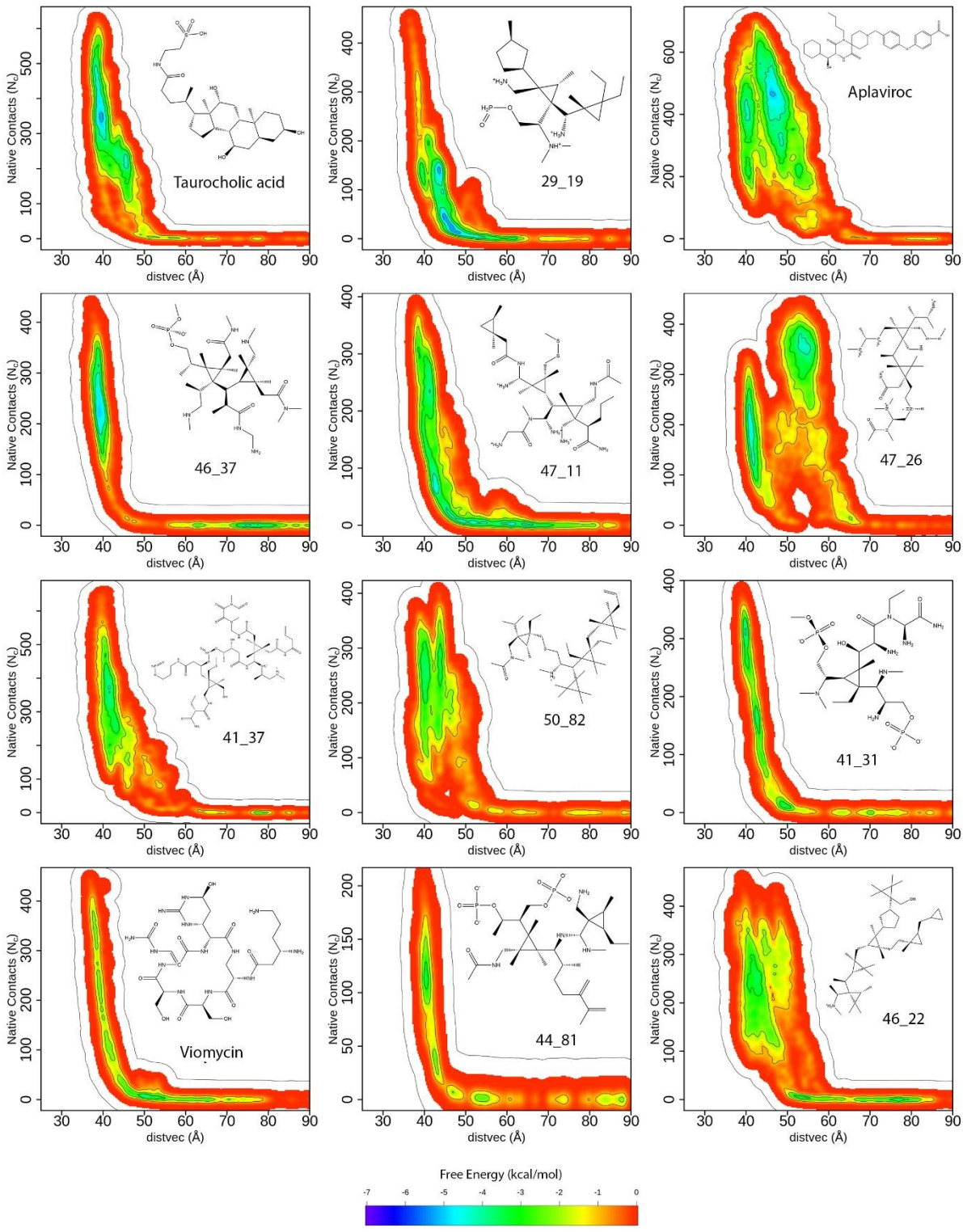

Figure 12. Free Energy surfaces of de novo and drug molecules with free energy stability between -5.5 kcal/mol and -3.4 kcal/mol (both ends are included). The chemical structure is shown in the inset of the respective plot.

As each molecule exhibits distinct behavior, identifying a unifying trend in their binding mechanisms proves challenging. Nevertheless, it is noteworthy that many strongly binding

molecules demonstrate a narrow free energy profile based on native contacts. It is evident from the free energy plots that a minimum of 9 molecules have binding free energy less than the experimental benchmark (-10.3 kcal/mol), suggesting that the RBD would bind to these molecules comparably or even more strongly than hACE2. Significantly, three molecules (47_68, danoprevir, and solithromycin) exceed the binding strength even of the lowest estimated interaction between hACE2 and RBD.

3.5 Conclusion

The strategy employed in this study for identifying COVID-19 therapeutics relies on free energy calculations related to the binding of small molecules to the receptor binding domain of the spike protein. We utilized well-tempered metadynamics, a state-of-the-art enhanced sampling technique to validate the binding affinity of novel molecules generated using an in-house de novo algorithm and a few repurposable molecules. To demonstrate the accuracy of our simulation protocol, we calculated the free energy of RBD and hACE2 to be -13 kcal/mol, a value that aligns closely with the experimental results. The same protocol was applied to all 55 molecules to obtain their binding free energy to RBD. The identical protocol was applied to all 55 molecules to determine their free energy of binding to RBD. Further, we extended the same simulation protocol for 35 de novo molecules 20 repurposable drug molecules (from the similarity mapping). Among these, we identified 21 molecules with free energy values below -9 kcal/mol, signifying the success of our protocol. Notably, at least four de novo molecules and five repurposable drug molecules exhibit binding to RBD at strengths comparable to or higher than that of hACE2. This suggests that these molecules could potentially inhibit the interaction between RBD and hACE2, thereby preventing the viral attack that was responsible for the spread of COVID-19.

3.6 Collaborations

The de novo drug generation and further similarity search for potential repurposable molecules were carried out by Rituparno Chowdhury and Venkata Sai Sreyas Adury. The atomic coordinates of generated de novo molecules and repurposable molecules were taken forward for forcefield generation, modeling, and conducting molecular dynamics simulations. The modeling, molecular dynamics simulations, and free energy calculations were carried out together with Dr. Reman Kumar Singh.

3.7 References

1. Tang, X.C., et al., *Prevalence and genetic diversity of coronaviruses in bats from China.* Journal of Virology, 2006. **80**(15): p. 7481-7490.
2. Boson, B., et al., *The SARS-CoV-2 envelope and membrane proteins modulate maturation and retention of the spike protein, allowing assembly of virus-like particles.* J Biol Chem, 2021. **296**: p. 100111.
3. Jackson, C.B., et al., *Mechanisms of SARS-CoV-2 entry into cells.* Nat Rev Mol Cell Biol, 2022. **23**(1): p. 3-20.
4. Shang, J., et al., *Cell entry mechanisms of SARS-CoV-2.* Proc Natl Acad Sci U S A, 2020. **117**(21): p. 11727-11734.
5. Ge, X.Y., et al., *Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor.* Nature, 2013. **503**(7477): p. 535-+.
6. Walls, A.C., et al., *Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein.* Cell, 2020. **181**(2): p. 281-292 e6.
7. Wrapp, D., et al., *Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation.* Science, 2020. **367**(6483): p. 1260-1263.
8. Du, L.Y., et al., *The spike protein of SARS-CoV - a target for vaccine and therapeutic development.* Nature Reviews Microbiology, 2009. **7**(3): p. 226-236.
9. Yan, R., et al., *Structural basis for the recognition of the SARS-CoV-2 by full-length human ACE2.* Science, 2020.
10. Huang, Y., et al., *Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19.* Acta Pharmacol Sin, 2020. **41**(9): p. 1141-1149.
11. Pandey, P., et al., *Targeting SARS-CoV-2 spike protein of COVID-19 with naturally occurring phytochemicals: an in silico study for drug development.* J Biomol Struct Dyn, 2021. **39**(16): p. 6306-6316.
12. Bansal, P., et al., *In silico molecular docking of SARS-CoV-2 surface proteins with microbial non-ribosomal peptides: identification of potential drugs.* J Proteins Proteom, 2021. **12**(3): p. 177-184.
13. Shamkh, I.M. and D. Pratiwi, *Development of SARS-CoV-2 Inhibitors Using Molecular Docking Study with Different Coronavirus Spike Protein and ACE2.* Journal of Molecular Docking, 2021. **1**(1): p. 1-14.

14. Basu, A., A. Sarkar, and U. Maulik, *Molecular docking study of potential phytochemicals and their effects on the complex of SARS-CoV2 spike protein and human ACE2.* Sci Rep, 2020. **10**(1): p. 17699.
15. Solo, P. and M.A. Doss, *Potential inhibitors of SARS-CoV-2 (COVID 19) spike protein of the delta and delta plus variant: In silico studies of medicinal plants of North-East India.* Curr Res Pharmacol Drug Discov, 2021. **2**: p. 100065.
16. Shang, J., et al., *Structural basis of receptor recognition by SARS-CoV-2.* Nature, 2020.
17. Chowdhury, R., et al., *Atomistic De-novo Inhibitor Generation-Guided Drug Repurposing for SARS-CoV-2 Spike Protein with Free-Energy Validation by Well-Tempered Metadynamics.* Chem Asian J, 2021. **16**(12): p. 1634-1642.
18. Wishart, D.S., et al., *DrugBank: a knowledgebase for drugs, drug actions and drug targets.* Nucleic Acids Res, 2008. **36**(Database issue): p. D901-6.
19. Barducci, A., G. Bussi, and M. Parrinello, *Well-tempered metadynamics: a smoothly converging and tunable free-energy method.* Phys Rev Lett, 2008. **100**(2): p. 020603.
20. Godden, J.W., L. Xue, and J. Bajorath, *Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients.* Journal of Chemical Information and Computer Sciences, 2000. **40**(1): p. 163-166.
21. Goodsell, D.S., G.M. Morris, and A.J. Olson, *Automated docking of flexible ligands: Applications of AutoDock.* Journal of Molecular Recognition, 1996. **9**(1): p. 1-5.
22. Frisch, M.J., et al., *Gaussian 09, Revision A.1*. 2009.
23. Wang, J.M., et al., *Automatic atom type and bond type perception in molecular mechanical calculations.* Journal of Molecular Graphics & Modelling, 2006. **25**(2): p. 247-260.
24. Eswar, N., et al., *Comparative protein structure modeling using Modeller.* Curr Protoc Bioinformatics, 2006. **Chapter 5**: p. Unit-5 6.
25. Hornak, V., et al., *Comparison of multiple amber force fields and development of improved protein backbone parameters.* Proteins-Structure Function and Bioinformatics, 2006. **65**(3): p. 712-725.
26. Jorgensen, W.L., et al., *Comparison of Simple Potential Functions for Simulating Liquid Water*. Journal of Chemical Physics, 1983. **79**(2): p. 926-935.
27. Press, W.H.T., et al., *Numerical Recipes: The Art of Scientific Computing*. 2007: CAMBRIDGE UNIVERSITY PRESS.

28. Berendsen, H.J.C., et al., *Molecular-Dynamics with Coupling to an External Bath.* Journal of Chemical Physics, 1984. **81**(8): p. 3684-3690.
29. Nose, S., *A Molecular-Dynamics Method for Simulations in the Canonical Ensemble.* Molecular Physics, 1984. **52**(2): p. 255-268.
30. Hess, B., et al., *LINCS: A linear constraint solver for molecular simulations.* Journal of Computational Chemistry, 1997. **18**(12): p. 1463-1472.
31. Darden, T., D. York, and L. Pedersen, *Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems.* Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.
32. Tribello, G.A., et al., *PLUMED 2: New feathers for an old bird.* Computer Physics Communications, 2014. **185**(2): p. 604-613.
33. Abraham, M.J., et al., *GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers.* SoftwareX, 2015. **1-2**: p. 19-25.
34. Hess, B., et al., *GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation.* Journal of Chemical Theory and Computation, 2008. **4**(3): p. 435-447.
35. Sasikala, W.D. and A. Mukherjee, *Intercalation and de-intercalation pathway of proflavine through the minor and major grooves of DNA: roles of water and entropy.* Phys Chem Chem Phys, 2013. **15**(17): p. 6446-55.
36. Mukherjee, A. and W.D. Sasikala, *Drug-DNA Intercalation: From Discovery to the Molecular Mechanism.* Dynamics of Proteins and Nucleic Acids, 2013. **92**: p. 1-62.
37. Jaccard, P., *Etude comparative de la distribution florale dans une portion des Alpes et du Jura.* Bull Soc Vaudoise Sci Nat, 1901. **37**: p. 547-589.
38. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.

Chapter 4

De novo drug generation and free energy validation by targeting RNA-dependent RNA Polymerase (RdRp) of SARS-CoV-2

4.1 Overview

Viruses, with their exceptionally high mutation rates, can effectively elude the human immune system and various conventional medical interventions. However, the RNA-dependent RNA polymerase (RdRp) in RNA viruses has predominantly maintained its structure, making substantial mutations in this protein improbable. In this context, the COVID-19 pandemic has shown the necessity for innovative therapeutic approaches. We employed an in-house made de novo drug design algorithm to generate entirely new molecules by targeting the RNA-dependent RNA polymerase (RdRp) of SARS-CoV-2. The generated molecules were subjected to all-atom explicit-water free energy calculations using well-tempered metadynamics simulations. Our studies show that certain newly generated ligands, created de novo algorithm, exhibit a higher affinity for the RNA-dependent RNA polymerase (RdRp) compared to the active form of the recently FDA-approved drug Remdesivir, known as remdesivir triphosphate (RTP). Additionally, we elucidated the unbinding mechanism of RTP and the best binding de novo molecule, aiming to comprehend the fundamental driving forces responsible for stabilization in the bound minima at the active site of RdRp.

4.2 Introduction

The genome of SARS-CoV-2 is composed of single-stranded positive-sense RNA, and its replication occurs through a multi-subunit complex of viral nonstructural proteins (NSPs), facilitating both replication and transcription. [1-4] An essential component of the NSP complex is the RNA-dependent RNA polymerase (RdRp). [1, 5, 6] To carry out its function, RdRp relies on additional factors, including NSP7 and NSP8, for proper activity. [7] The active site of NSP12 is positioned at the substrate domain's center, enabling RNA synthesis by accessing the RNA template through the template input channel. Fig. 1 illustrates the structural representation of NSP7, NSP8, and NSP12. The central RNA-dependent RNA polymerase (RdRp) domain consists of three subdomains: the thumb, palm, and right-handed cup-like fingers.[8, 9] Notably, Fig. 2 highlights the regions of the fingers and thumb that contribute to the open/closed structure of the RdRp protein. Due to its critical role in the viral replication process and its conserved structure, RdRp emerges as a compelling target for drug development aimed at combating SARS-CoV-2 infections.

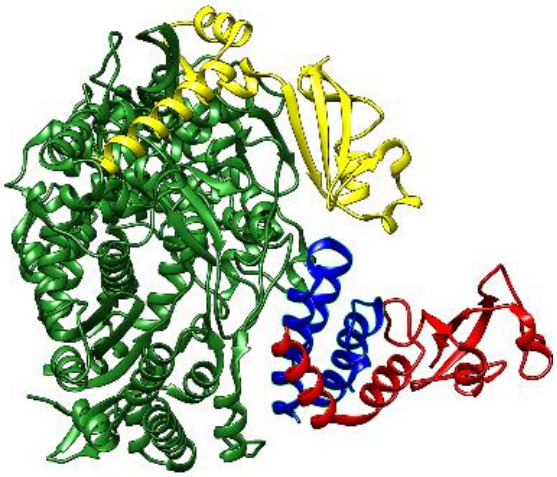

Figure 1. Structure of the RdRp protein (PDB ID: 6NUR), with its key components highlighted: NSP12 is represented in green, NSP7 in blue, chain B of NSP8 in yellow, and chain D of NSP8 in red.

Among various potential targets, such as the spike protein and main protease (Mpro), RdRp stands out as a promising target for drug design. Several antiviral molecules have been identified by specifically targeting crucial residues in the active site of RdRp and other proteins integral to the virus's lifecycle.[10-13] Numerous drug repurposing studies have been reported, employing computational tools like docking to target RdRp.[14-19] These studies have

generated many potential drugs that could be repurposed for COVID-19, pending thorough theoretical or experimental validations. Among the potential drugs targeting RdRp, Remdesivir is notable for having received FDA approval. [11] While the effectiveness of Remdesivir is still debated, numerous studies have been conducted to examine its inhibitory effects. [20-22] Both experimental and computational studies have revealed that the active form of Remdesivir, Remdesivir triphosphate (RTP), is an effective drug that hinders RNA replication at the active site of RdRp.[23-27] Investigations suggest that the binding affinity of RTP falls within the range of approximately ~5 kcal/mol to ~7 kcal/mol, leading to complete inhibition of the polymerization process at the catalytic site. [20, 28-30] In this current study, we employed an atomistic receptor-based de novo generation algorithm[31], along with structural mapping to identify a novel set of molecules and repurpose existing drugs with the potential to bind more effectively than Remdesivir triphosphate (RTP). Initially, we identified the active region of the Remdesivir triphosphate (RTP)-bound 3D structure of RdRp to locate the hotspot region, serving as the target location for generating small molecules. To validate the binding, we conducted molecular dynamics simulations and free energy calculations to compare the binding strength of our de novo molecules with that of RTP. Through analysis of multidimensional free energy surfaces, we identified five molecules, including some known drug molecules, that exhibit stronger binding than RTP at the catalytic site of RdRp.

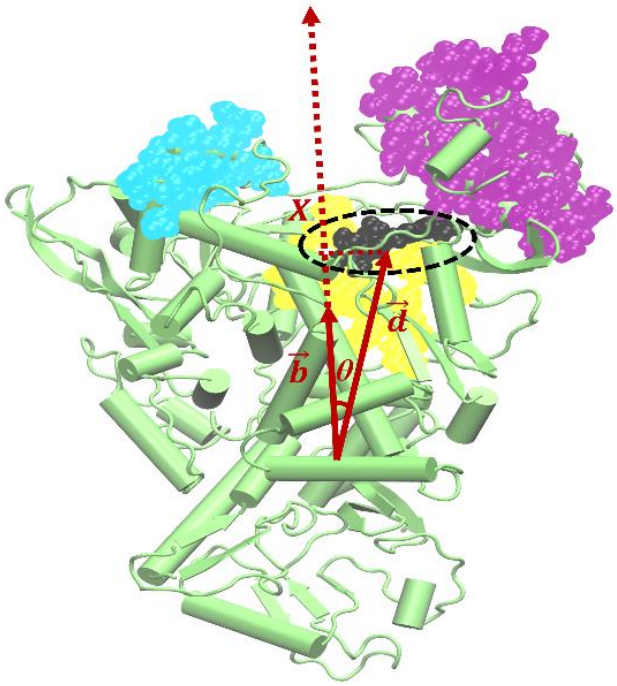

Figure 2. Structure of NSP12 protein. The active site of RdRp is highlighted in yellow. The finger sub-region (residue 500 to 525) and thumbsub–region (residue 850 to 900), forming the ring opening/closed junction, are indicated by cyan and purple, respectively. A sample small molecule is depicted in black, enclosed by a black dotted circle near the protein's active site. The vectorial representation of CVs used for metadynamics simulations is shown by red arrows

4.3 Computational details

4.3.1 Selection of molecules

A total of 5375 molecules were generated by utilizing the in-house made de novo algorithm [31], aiming to select at least 160 molecules with heavy atoms ranging from 11 to 50. The de novo score varied from -36 kcal/mol, which represented the strongest interaction energy, to -5.3 kcal/mol, the weakest. The distribution of interaction energies of the generated molecules to the RdRp active site is shown in the Fig. 3. For a more comprehensive analysis, molecules with interaction scores below three standard deviations from the mean value were taken forward for free energy calculations. Further, we mapped the de novo molecules with approved or investigational (repurposed) drugs from the DrugBank[34] database, employing a Tanimoto similarity score of 0.4 or higher.[35] It was based on the assumption that molecules sharing similar chemical structures are expected to demonstrate comparable functions. For the de novo generated molecules at the predefined hotspot, the bound structure pose (generated by the program) was taken, and the complex was minimized using the steepest descent[36] algorithm to remove the steric clashes, if any. The chemical structures of the de novo generated and repurposable molecules which are used for free energy validation studies, are shown in the Fig. 4.

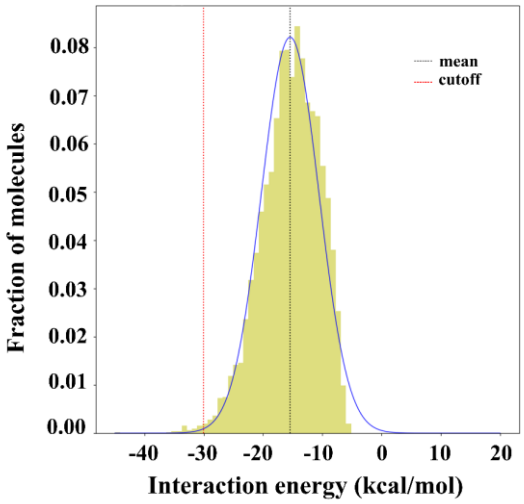

Figure 4. Distribution of de novo interaction energies with the RdRp protein. The mean and cutoff values are shown by black and red dotted lines.

4.3.2 Modeling the systems

Modeller 9.21[32] was used to model missing residues in the RdRp (PDB ID: 6NUR). The RdRp-small molecule complex was constructed according to the following protocols. The atoms in the residues 616-619, 758-762, 798-800, 811-814 were taken into consideration for defining the docking region and small molecule generation (using in-house made de novo algorithm[31]), which includes the active site of RdRp.

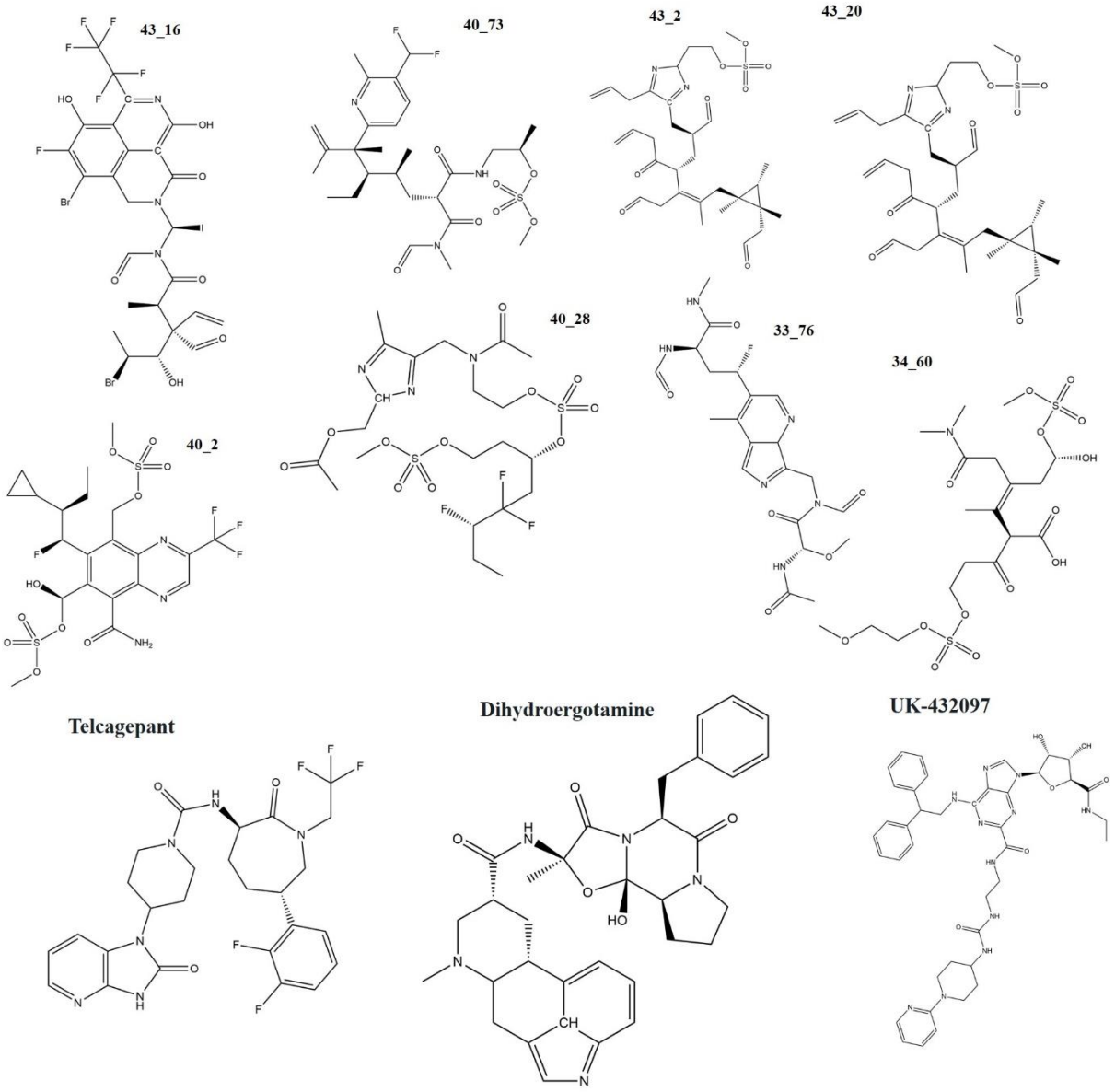

Figure. 3 Chemical structures of the de novo generated and repurposable molecules which were used for free energy validation studies

However, for the repurposable molecules (obtained from the DrugBank[33] database), the structure was docked to the hotspot region according to the docking protocol as follows. A grid box is constructed with dimensions 32.6 Å×33.3 Å×36.5 Å by centering the residues 691 and 622 as midpoints, respectively and incorporating the active site of RdRp protein. The docking protocol was executed using Autodock vina[34]. The best molecules (with docking score > 8 kcal/mol) and RTP were subjected to further free energy calculation. The docking scores for the selected molecules is shown in the Table 1.

Table 1. Docking scores (calculated using AutoDock Vina) for the repurposable drug molecules

Sl. No.	Drug Name	Docking Score (kcal/mol)
1	Dihydroergotamine	-8.7
2	RTP	-6.7
3	Telcagepent	-8.5
4	UK-432097	-8.9

General Amber Force Field[35] (GAFF) is used for the construction of forcefield parameters for all the small molecules. The molecules with complete valency were optimized based on Hartree-Fock (HF) theory with 6-31G* basis set. The optimization of the molecules was carried out using Gaussian 09 software[36]. Antechamber [38] module of AMBER 18 is used for the RESP charge calculations and corresponding structure parametrization. A Python script acpype.py (available at https://github.com/t-/acpype)[37] was then used to convert the output parameters and the coordinates to GROMACS[38] acceptable format. For the RdRp protein AMBER99SB[39], forcefield is used for parametrization. Once the RdRp-ligand complex was made, the complex was inserted into a box of dimensions 16 X 14 X 11 nm^3. it was further solvated by adding ~76000 water molecules based on TIP3P[40] water model. To maintain the physiological ion concentration (150 mM), Na^+ and Cl^- ions were added to the system, and extra Cl^- ions were added to the system to maintain electrical neutrality.

4.3.3 Simulation details

At first, every system containing the RdRp-ligand complex was subjected to energy minimization using the steepest descent[41] method for 10000 steps. Then, the system was heated by raising the temperature from 0K to 300K in 200 picoseconds using Berendsen thermostat and barostat[42], both having a coupling constant of 0.6 ps. While heating, position restraints of 25 kcal/mol/Å2 were enforced on heavy atoms of the complex. Next, equilibration was carried out for 2 nanoseconds at a constant temperature of 300 K and pressure of 1 bar, without any position restraints. The same thermostat and barostat were used for equilibration, with coupling constants of 0.2 ps for each. The last 100 picoseconds of the NPT simulation were utilized to compute the average volume, which was employed in the final five ns of unconstrained NVT equilibration using the Nosé-Hoover thermostat[43] with a coupling constant of 0.2 ps. Throughout the simulation, all bonds were constrained using LINCS[44] algorithm, and electrostatics were calculated using the Particle Mesh Ewald[45] (PME) method. The van der Waals (vdW) and electrostatic long-range interactions were both cut off at 10 Å. The simulation timestep for all simulations was 2 fs. Finally, the equilibrated structures of the ligand-bound protein systems were subjected to 5 ns production run under NVT conditions. If the ligand was found to be bound after 5 ns of simulation, the system proceeded for free energy calculations. Otherwise, the molecules were rejected due to their weak binding affinity.

4.3.4 Free energy calculations

After equilibration, well-tempered metadynamics[46] simulations were performed to calculate the binding free energy of the molecules. Two collective variables (CVs) were utilized for these simulations: *distvec*(X) and *angvec*(θ). For the metadynamics simulations, an initial hill height of 0.2 kJ/mol, bias factor of 10 were chosen with a hills deposition rate of 2 ps. The Gaussian widths for X and θ was set to be 0.6 Å and 0.002 rad, respectively. These CVs have been successfully applied in previous studies concerning DNA intercalation and protein-small molecule binding studies, producing free energy values in concurrence with experimental outcomes. [31, 47, 48] The MULE[49] program was used to obtain the minimum free energy path from the ligand-bound minimum to the unbound state in the free-energy landscape. The length of the metadynamics simulation for each molecule is shown in Table 2. Using ligplot[50], a 2D representation of small molecules interacting with the protein surface is constructed for the RTP and the best binding molecule.

4.3.5 Choice of collective variables

The CV *distvec*(X) is defined by two vectors, $\hat{b}$ and $\vec{d}$. The unit vector $\hat{b}$ is directed from the center of mass (COM) of residues 127, 130, 131, 134, 139, 140 – 147, 157, 162, 175, and 787 and terminates at the COM of residues 545, 553, 555, 556, 617-619, 622, 623, 680, 682, 687, 691, 759-761, 811, 813, and 814, which are closer to the hotspot region shown by yellow region in Fig. 2. The vector $\vec{d}$ starts at the COM of residues 127, 130, 131, 134, 139, 140 – 147, 157, 162, 175, and 787 and ends at the COM of the heavy atoms of the small molecule considered.

$$X = \hat{b}.\vec{d}.$$

Similarly the CV *angvec*(θ) is defined by

$$\theta = cos^{-1}(\hat{b}.\vec{d}/|d|).$$

Fig. 2 illustrates an intuitive representation of the defined CVs. The unbinding of the molecules is observed anywhere between 50 and 200 ns, ensuring that the molecule moves away from the equilibrated bound structure towards the solution medium with an increase in *X*. The directionality of the unbinding of the molecule is captured by *θ*, which complements *X*.

Table 2. Metadynamics simulation time scale for all the validated molecules

Sl. No.	**Molecule**	**Simulation time (ns)**
1	33_76	54.2
2	34_60	95.5
3	40_2	68
4	40_28	69.3
5	40_73	111
6	43_16	180.4
7	43_2	131.9
8	43_20	91.7
9	Dihydroergotamine	174.4
10	RTP	46.5
11	Telcagepent	169.9
12	UK-432097	127

4.4 Results & Discussion

4.4.1 Binding free energy study of RTP molecule

Initially, the free energy studies have been carried out for the RTP molecule. The 2D-free energy surface for the RTP molecule is shown in the Fig. 5a. The black dotted lines in the free energy profile indicate the minimum free energy path connecting the bound minima state (state I) to the unbound state (state VI) through the intermediate states (states II to 1V). It is to be noted that the unbinding of the RTP molecule from its active site is proceeded by an increase in the value of X and θ, respectively. The free energy stability of the RTP-bound RdRp at its minima is found to be -5.6 kcal/mol, which matches the experimental binding affinity. [30] At the minima, hydrogen bond interaction from the residue 545 is found to stabilize the triphosphate group of the RTP. However, the other terminal end of RTP is found to be stabilized by the hydrogen bond interaction with residue 496 of RdRp. An illustration of the RTP molecule interacting with neighboring residues of the protein is shown in the Fig. 5b. A top view of the minima state of the RTP-RdRp complex is shown in the Fig. 5c. Various intermediates states of RdRp-RTP complex describing the unbinding pathway of RTP from the bound minima is shown in the Fig. 6 in accordance with the minimum free energy path shown in Fig. 5a.

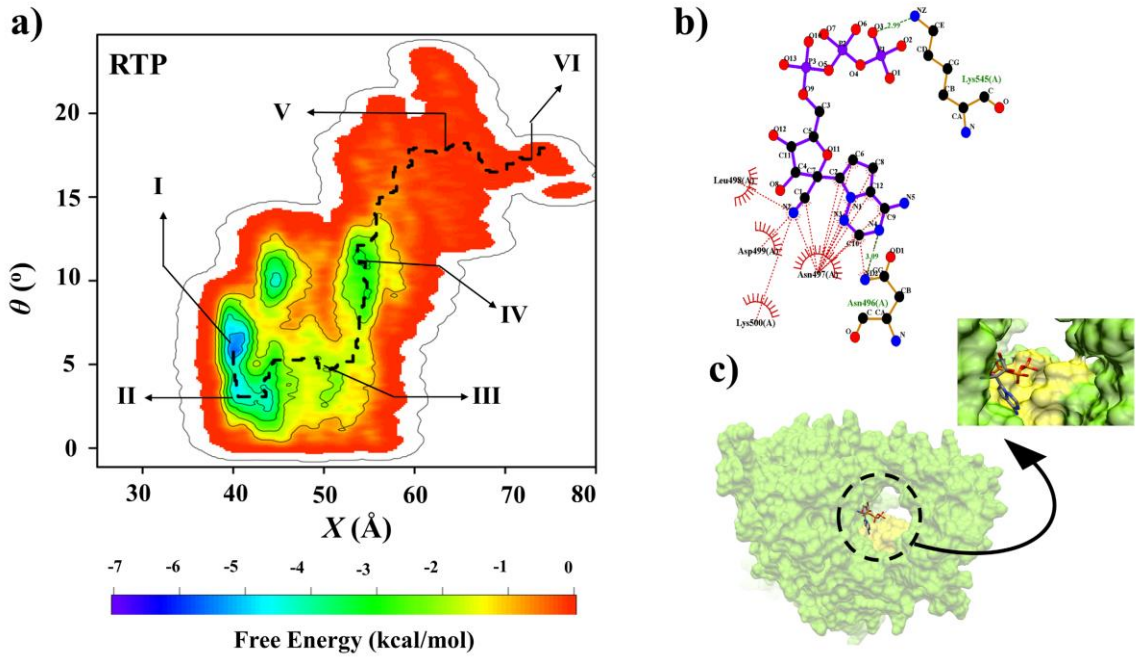

Figure.5 a) 2D free energy surface corresponding to unbinding of RTP molecule. The minimum free energy path connecting the bound state of RTP to the unbound state of RTP is indicated by black dotted lines. b) Illustration of key interactions between RTP molecule and the RdRp. c) Top view of RTP bound RdRp at its minima.

As shown in the representation Fig. 5b the core region of the RTP is also stabilized by residues 497, 498, 499, and 500 through hydrophobic interactions. The core structure of RdRp also includes an architecture composed of finger and palm sub-domains (highlighted in Fig. 2), playing a crucial role in substrate recognition. The minima of RTP are identified at the active site, marked by the yellow-colored region in Fig. 2. As shown in the Fig. 6, after the formation of stabilizing interactions at its active site as shown in states I to III, the RTP move towards finger sub-region (indicated by cyan region and states IV and V). Finally, the unbinding of the molecule towards the bulk water medium occurred from the finger subregion. Fig. 7 shows the 3D structures representing the interaction between the important residues of RdRp and RTP throughout the unbinding pathway (as shown in Fig. 6).

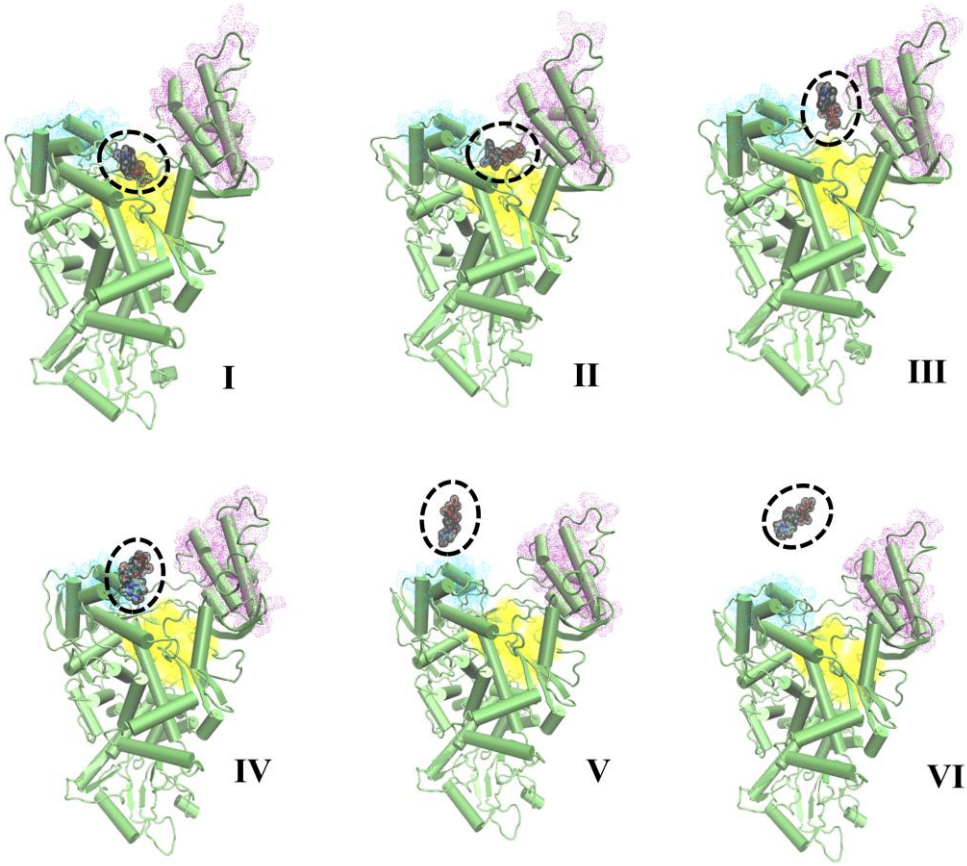

Figure 6. Representation of intermediate states (states I to VI) corresponding to unbinding the RTP molecule from its bound state as per the minimum free energy path shown in Fig. 5a. The RTP molecule is labeled with black dotted circle.

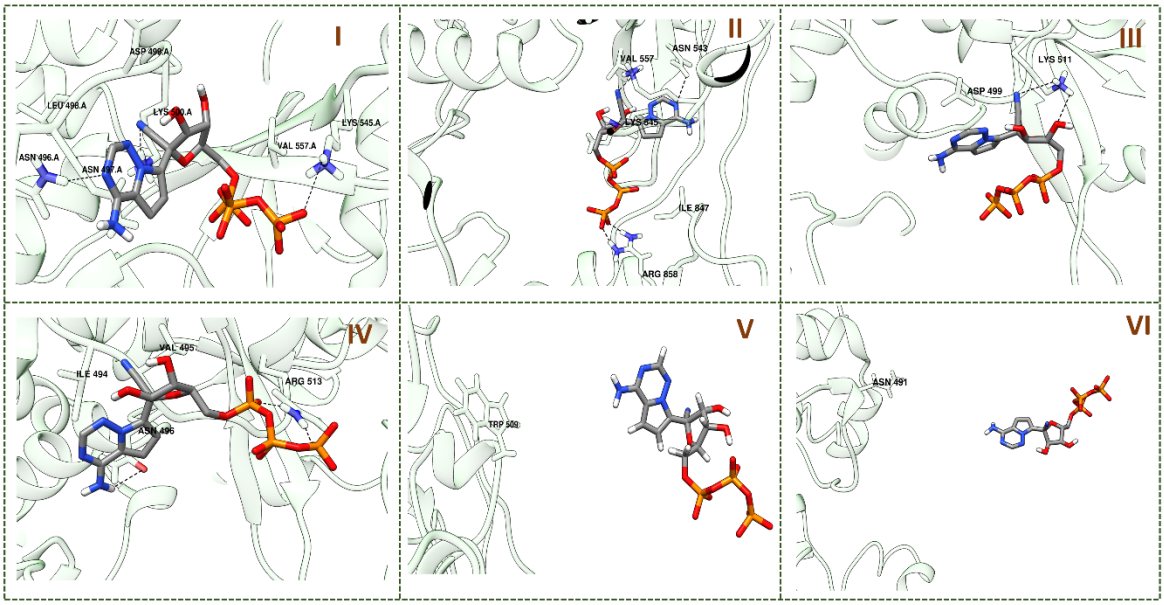

Figure 7. Representative 3D structures depict the interaction between RTP and RdRp residues along the unbinding pathway (I-VI), as defined by the structures presented in Fig. 5a and Fig. 6. Black dotted lines denote hydrogen bonds.

4.4.2 Binding free energy study of de novo and Repurposable molecules

This study aims to identify molecules, both de novo and repurposable, capable of binding more effectively than the RTP molecule. The free energy validation was conducted for all the chosen molecules using the protocol outlined in the methods section. Among the molecules examined in the free energy validation study, ten molecules demonstrated free energy stability less than -5.6 kcal/mol, corresponding to the free energy value of the RTP molecule. A bar diagram showing the free energy stability of all the validated molecules is shown in Fig. 8.

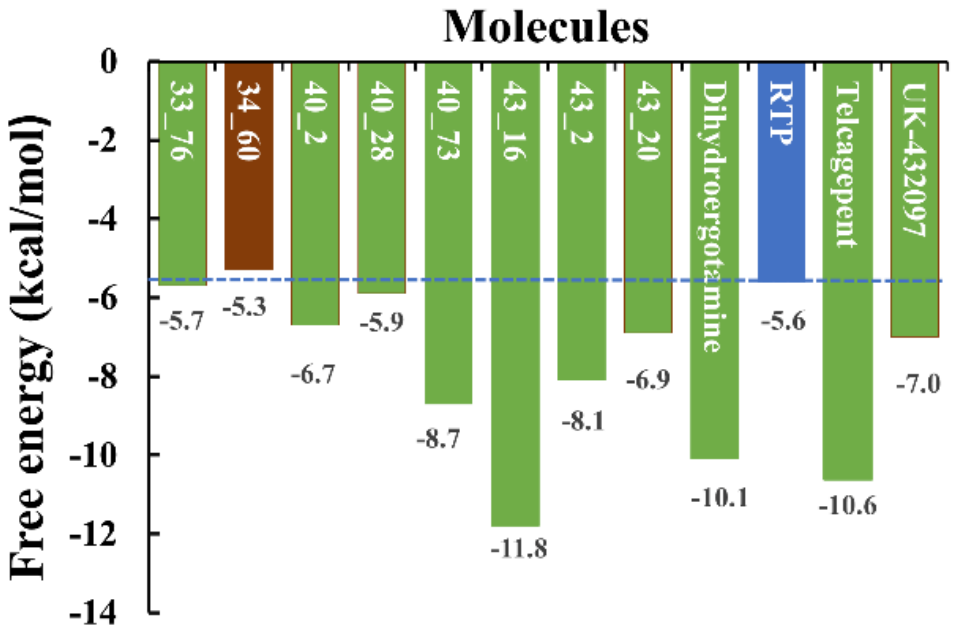

Figure 8. Bar diagram representing the free energy of the molecules obtained by well-tempered metadynamics simulations. The green bars show the molecules with better binding affinity than the RTP (blue bar). The brown bar shows the molecules with weaker binding affinity than RTP.

The leading molecule, designated as 43_16 as per the de novo nomenclature scheme, exhibited a free energy stability of -11.8 kcal/mol. The free energy profile associated with the unbinding of 43_16 is shown in the Fig. 9a.

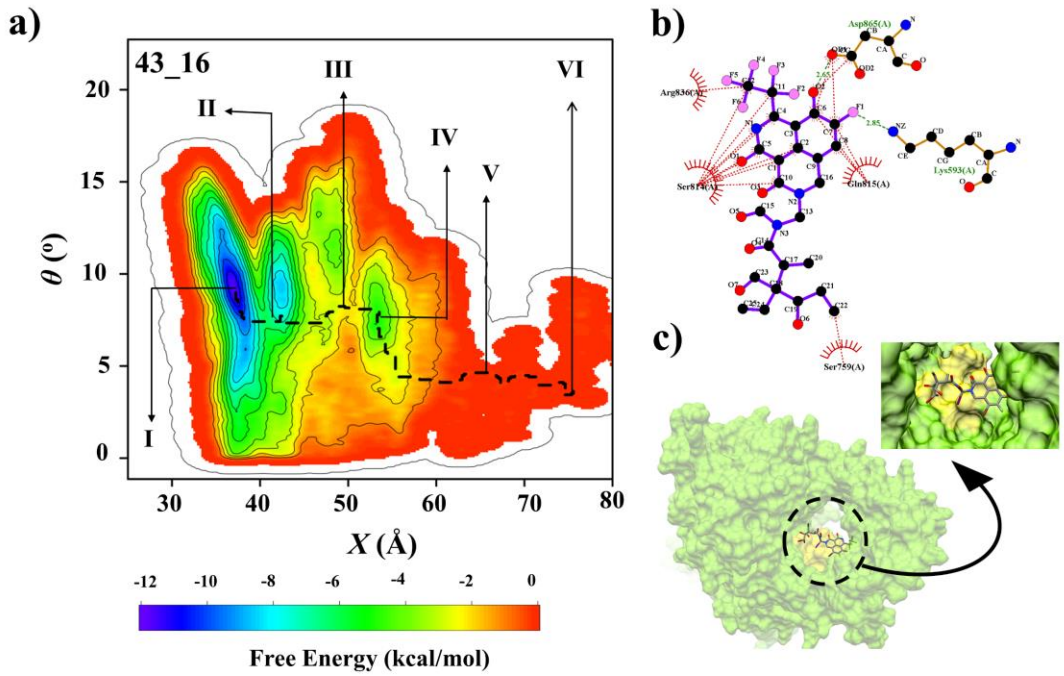

Figure. 9 a) 2D free energy surface corresponding to unbinding of de novo molecule 43-16. The minimum free energy path connecting the bound state of 43_16 to the unbound state of 43_16 is indicated by black dotted lines. b) Illustration of key interactions between 43_16 molecule and the RdRp. c) Top view of 43_16 bound RdRp at its minima state.

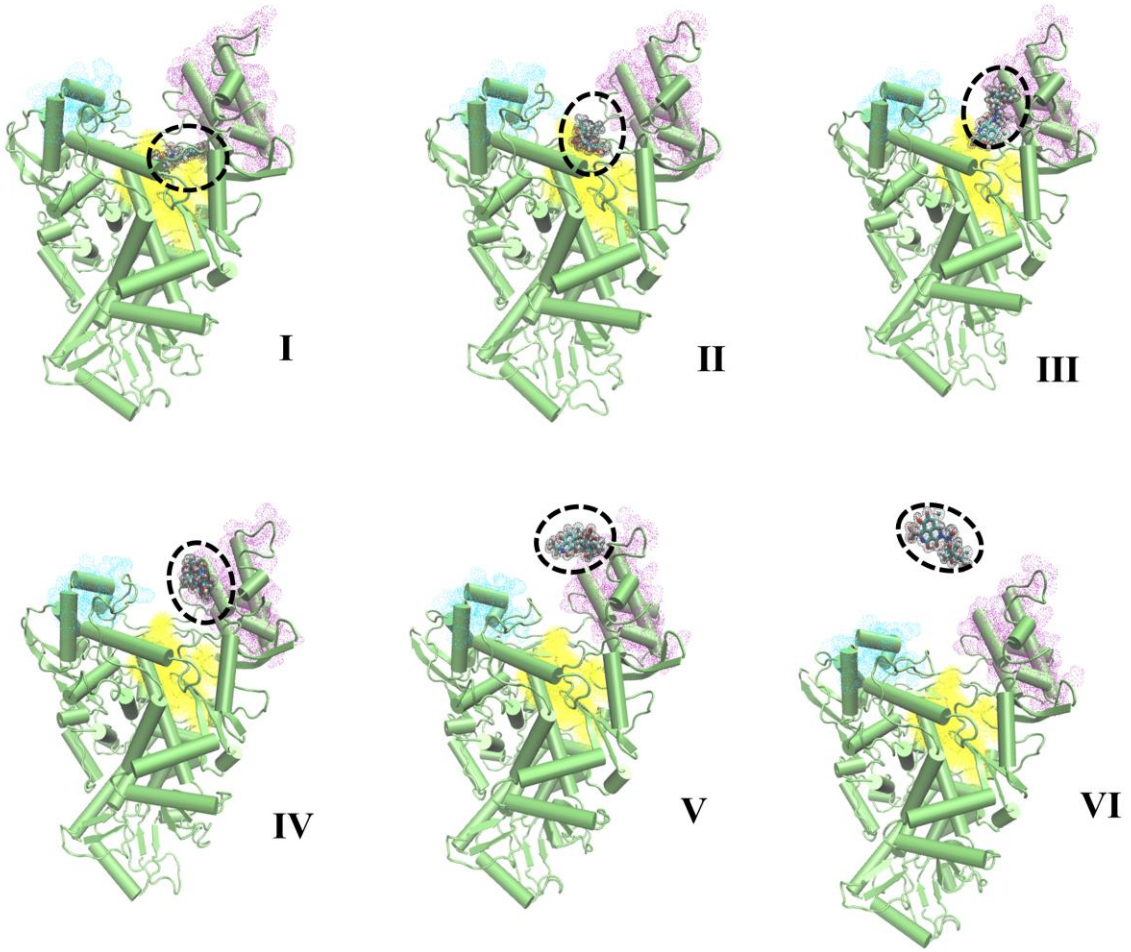

Figure 10. Representation of intermediate states (states I to VI) corresponding to the unbinding the 43_16 molecule from its bound state as per the minimum free energy path shown in Fig. 9a. The 43_16 molecule is labeled with a black dotted circle.

At its minima at the active site of RdRp, 43_16 molecule is stabilized by hydrophobic interactions (initiated by the protein residues 836, 814, 759, and 815) and hydrogen bond interactions (contributed by the protein residues 865 and 593) as shown in Fig. 9b. The major difference in the unbinding pathway of 43_16 compared to that of RTP is that in this case the unbinding is initiated with an increase in value of X and decrease in value of θ, with the detachment of molecule has happened from the thumb subregion (shown by purple color in Fig. 10). Note that the detachment has occurred via finger sub-region in case of RTP. Fig. 11

shows the 3D structures representing the interaction between the important residues of RdRp and 43_16 throughout the unbinding pathway (as shown in Fig. 10).

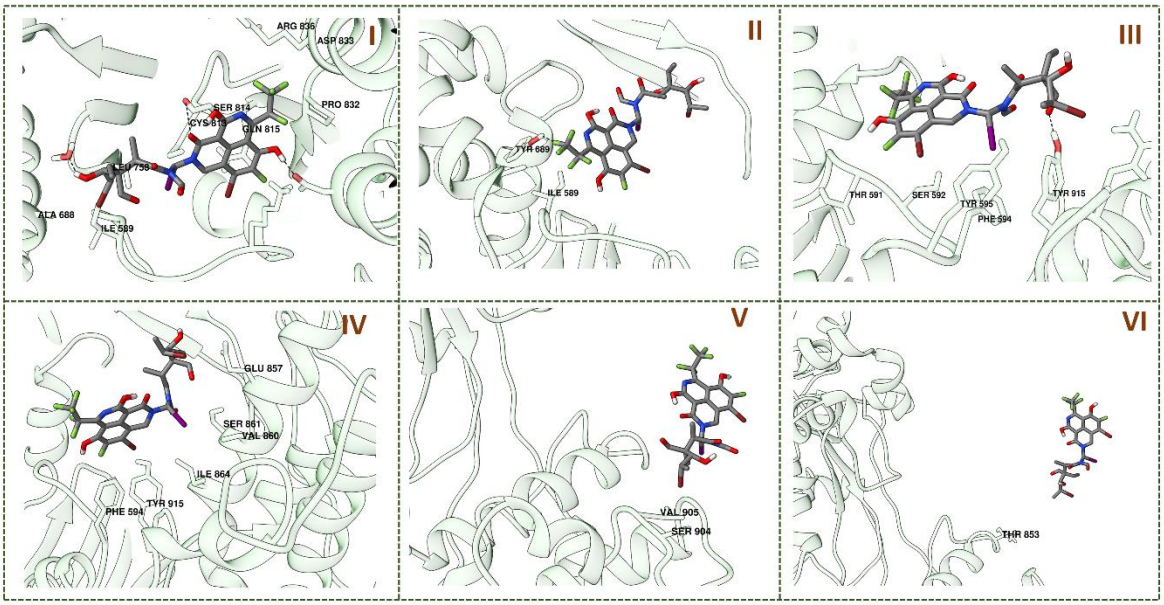

Figure 11. Representative 3D structures depict the interaction between 43_16 and RdRp residues along the unbinding pathway (I-VI), as defined by the structures presented in Fig. 9a and Fig. 10. Black dotted lines denote hydrogen bonds.

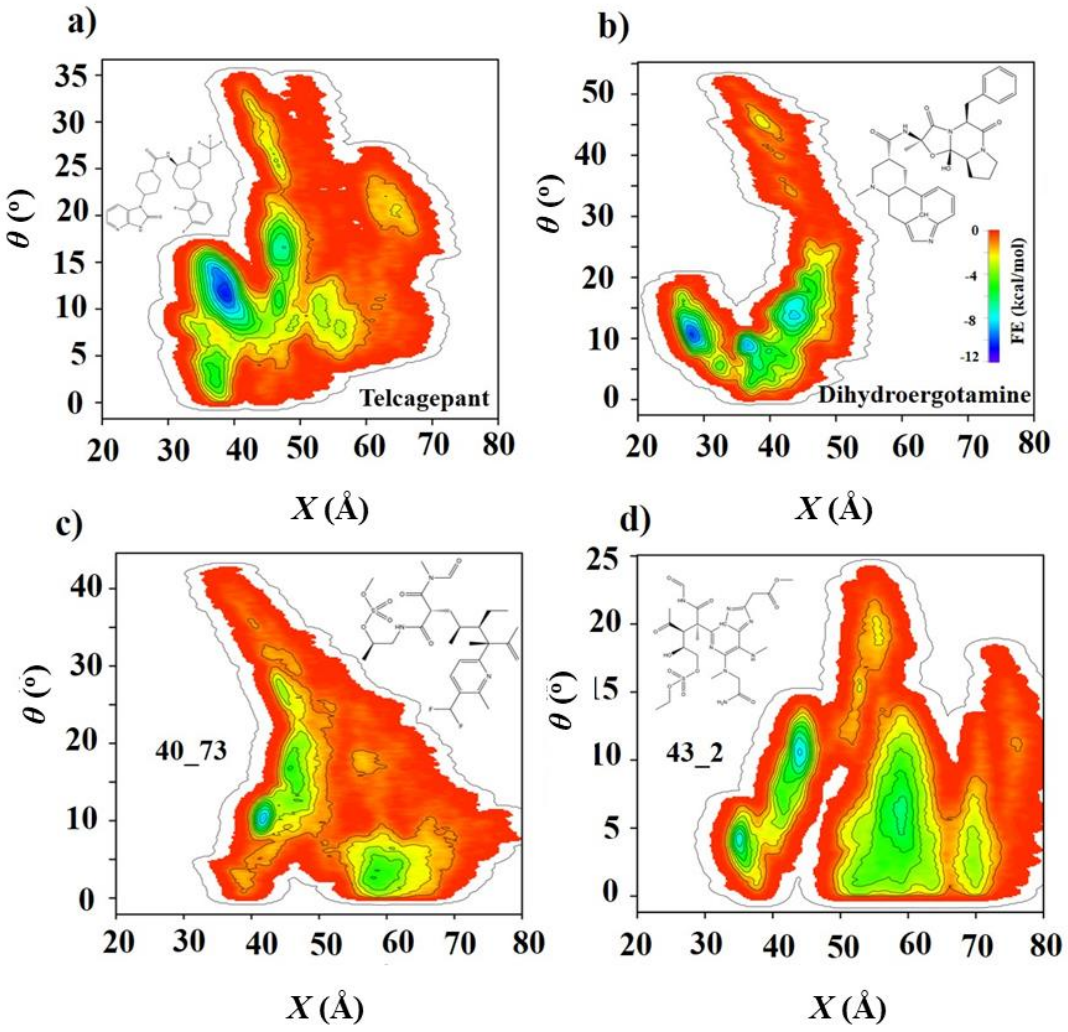

Figure 12. 2D Free energy plots of de novo & repurpose molecules with free energy stability lower than -8 kcal/mol. The chemical structure of the molecules is given in the inset of each free energy surface corresponding to each molecule. a) Telcagepant b) Dihydroergotamine c) 40_73 d) 43_2

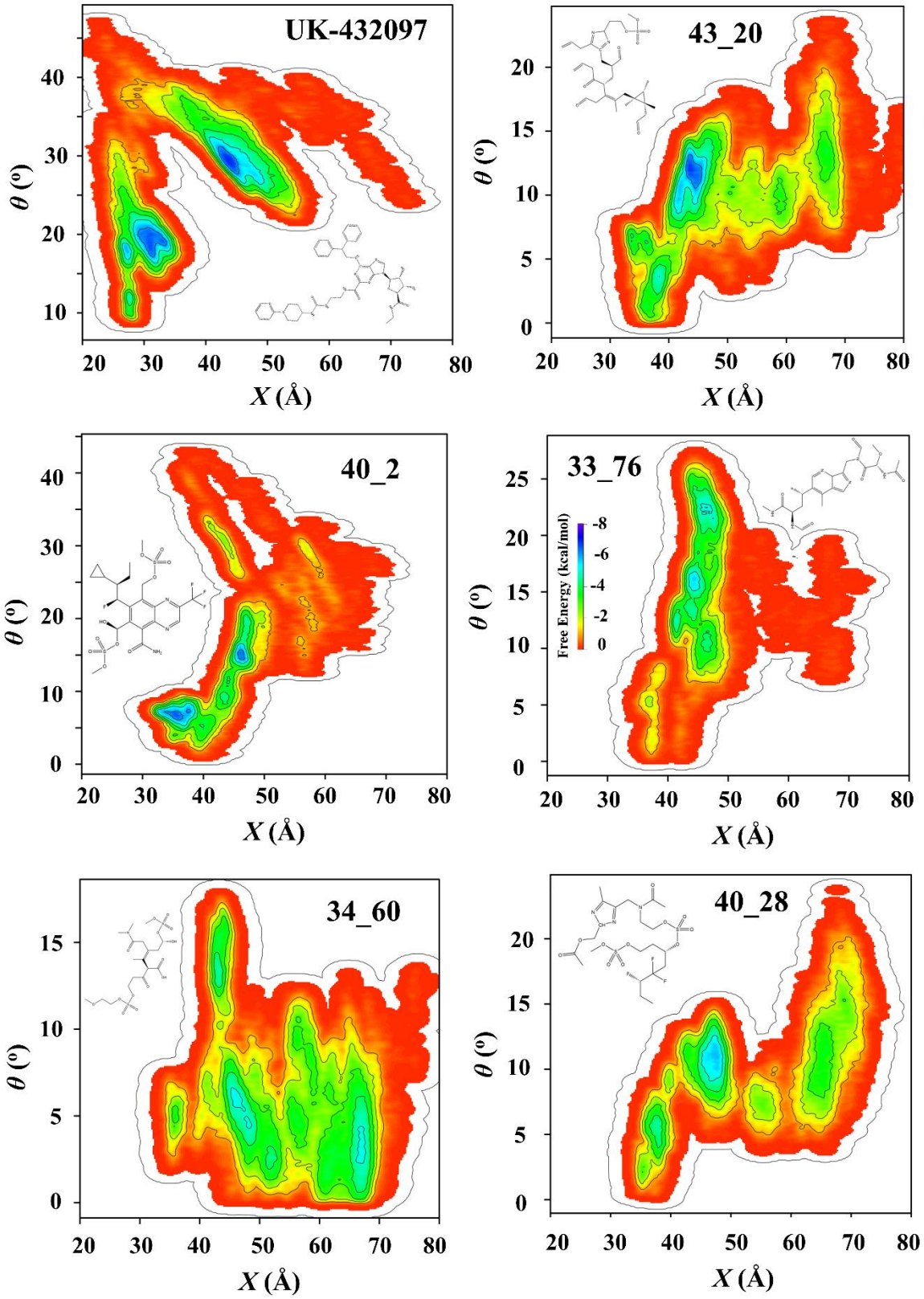

Figure. 13. 2D Free energy plots of de novo & repurposable molecules with free energy stability higher than -8.0 kcal/mol

Along with molecule 43_16, nine other molecules were also found to have a better binding affinity than RTP. Repurposable molecules exhibiting better binding affinity than RTP include Telcagepant, Dihydroergotamine, and UK-432097, with free energy values of -10.6 kcal/mol, -10.1 kcal/mol, and -7.0 kcal/mol, respectively. Similarly, other de novo generated molecules with superior binding affinity than RTP include 43_16, 40_73, 43_2, 43_20, 40_2, 40_28, 33_76, with free energy values of -11.8 kcal/mol, -8.7 kcal/mol, -8.1 kcal/mol, -6.9 kcal/mol, -6.7 kcal/mol, -5.9 kcal/mol, and -5.7 kcal/mol in that order. The free energy surfaces of the molecules with binding free energy lower than -8 kcal/mol are depicted in Fig. 12, along with the chemical structure of the molecules. The free energy surface of all others with binding free energy higher than -8 kcal/mol is shown in Fig. 13. Our approach has proven successful, yielding at least seven de novo molecules and three repurposable molecules that exhibit stronger binding than RTP.

4.5 Conclusion

The active site of the RdRp was utilized to find novel and existing repurposable molecules using an in-house made de novo generation algorithm and rigorous all-atom explicit water free energy validation studies using well-tempered metadynamics simulations. Utilizing CVs X and θ, the free energy landscape offered crucial insights into the directional movement of each candidate during the unbinding process. The protocol found at least ten molecules (including new and existing molecules) that can bind more effectively than reported RTP (free energy stability -5.6 kcal/mol). The best binding molecule exhibits a binding free energy stability of -11.8 kcal/mol. The analysis reveals that the molecules unbind through a narrow opening located at the finger and thumb subregion of the RdRp protein. Moreover, it is observed that the unbinding events of the molecules from the RdRp is facilitated through an opening junction near the finger and thumb subregion of the RdRp protein.

The approach and protocol utilized in the study provides a unique pathway for exploring chemical space in drug design, encompassing the development of novel compounds and the repurposing of readily applicable drugs. The successful validation, transitioning from de novo scoring to achieving free energy stability, provides encouragement to explore the application of this method to other known receptors for lead generation.

4.6 Collaborations

The de novo drug generation and further similarity search for potential repurposable molecules were carried out by Venkata Sai Sreyas Adury. The atomic coordinates of generated de novo molecules and repurposable molecules were taken forward for forcefield generation, modeling, and conducting free energy calculations.

4.7 References

1. McDonald, S.M., *RNA synthetic mechanisms employed by diverse families of RNA viruses.* WIREs RNA, 2013. **4**(4): p. 351-367.
2. Lu, R., et al., *Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding.* Lancet, 2020. **395**(10224): p. 565-574.
3. Ivanov, K.A. and J. Ziebuhr, *Human coronavirus 229E nonstructural protein 13: characterization of duplex-unwinding, nucleoside triphosphatase, and RNA 5'-triphosphatase activities.* J Virol, 2004. **78**(14): p. 7833-8.
4. Ivanov, K.A., et al., *Multiple enzymatic activities associated with severe acute respiratory syndrome coronavirus helicase.* J Virol, 2004. **78**(11): p. 5619-32.
5. Kirchdoerfer, R.N. and A.B. Ward, *Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors.* Nat Commun, 2019. **10**(1): p. 2342.
6. Subissi, L., et al., *One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities.* Proc Natl Acad Sci U S A, 2014. **111**(37): p. E3900-9.
7. Kirchdoerfer, R.N. and A.B. Ward, *Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors.* Nature Communications, 2019. **10**(1): p. 2342.
8. Gao, Y., et al., *Structure of the RNA-dependent RNA polymerase from COVID-19 virus.* Science, 2020. **368**(6492): p. 779-782.
9. Vicenti, I., M. Zazzi, and F. Saladini, *SARS-CoV-2 RNA-dependent RNA polymerase as a therapeutic target for COVID-19.* Expert Opin Ther Pat, 2021. **31**(4): p. 325-337.
10. Elfiky, A.A., *Ribavirin, Remdesivir, Sofosbuvir, Galidesivir, and Tenofovir against SARS-CoV-2 RNA dependent RNA polymerase (RdRp): A molecular docking study.* Life Sciences, 2020. **253**: p. 117592.

11. Gordon, C.J., et al., *The antiviral compound remdesivir potently inhibits RNA-dependent RNA polymerase from Middle East respiratory syndrome coronavirus.* Journal of Biological Chemistry, 2020. **295**(15): p. 4773-4779.
12. Hashemian, S.M.R., et al., *RdRp inhibitors and COVID-19: Is molnupiravir a good option?* Biomed Pharmacother, 2022. **146**: p. 112517.
13. Brunt, D., P.M. Lakernick, and C. Wu, *Discovering new potential inhibitors to SARS-CoV-2 RNA dependent RNA polymerase (RdRp) using high throughput virtual screening and molecular dynamics simulations.* Sci Rep, 2022. **12**(1): p. 19986.
14. Ahmad, J., et al., *SARS-CoV-2 RNA Dependent RNA polymerase (RdRp) - A drug repurposing study.* Heliyon, 2020. **6**(7): p. e04502.
15. Elfiky, A.A., et al., *Molecular dynamics simulations and MM-GBSA reveal novel guanosine derivatives against SARS-CoV-2 RNA dependent RNA polymerase.* RSC Adv, 2022. **12**(5): p. 2741-2750.
16. Behera, S.K., et al., *Computational drug repurposing study elucidating simultaneous inhibition of entry and replication of novel corona virus by Grazoprevir.* Scientific Reports, 2021. **11**(1).
17. Alam, A., et al., *Towards the discovery of potential RdRp inhibitors for the treatment of COVID-19: structure guided virtual screening, computational ADME and molecular dynamics study.* Structural Chemistry, 2022. **33**(5): p. 1569-1583.
18. Celik, I. and T.E. Tallei, *A computational comparative analysis of the binding mechanism of molnupiravir's active metabolite to RNA-dependent RNA polymerase of wild-type and Delta subvariant AY.4 of SARS-CoV-2.* Journal of Cellular Biochemistry, 2022. **123**(4): p. 807-818.
19. Koulgi, S., et al., *Natural plant products as potential inhibitors of RNA dependent RNA polymerase of Severe Acute Respiratory Syndrome Coronavirus-2.* PLoS One, 2021. **16**(5): p. e0251801.
20. Yin, W., et al., *Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir.* Science, 2020. **368**(6498): p. 1499-1504.
21. Frediansyah, A., et al., *Remdesivir and its antiviral activity against COVID-19: A systematic review.* Clin Epidemiol Glob Health, 2021. **9**: p. 123-127.
22. Bravo, J.P.K., et al., *Remdesivir is a delayed translocation inhibitor of SARS-CoV-2 replication.* Mol Cell, 2021. **81**(7): p. 1548-1552 e4.
23. Elfiky, A.A., *SARS-CoV-2 RNA dependent RNA polymerase (RdRp) targeting: an in silico perspective.* Journal of Biomolecular Structure and Dynamics, 2020: p. 1-9.

24. Elfiky, A.A., *Anti-HCV, nucleotide inhibitors, repurposing against COVID-19.* Life Sciences, 2020. **248**: p. 117477.
25. Yu, R., et al., *Computational screening of antagonists against the SARS-CoV-2 (COVID-19) coronavirus by molecular docking.* Int J Antimicrob Agents, 2020. **56**(2): p. 106012.
26. Singh, P.K., S. Pathania, and R.K. Rawal, *Exploring RdRp-remdesivir interactions to screen RdRp inhibitors for the management of novel coronavirus 2019-nCoV.* SAR QSAR Environ Res, 2020. **31**(11): p. 857-867.
27. Wang, J., et al., *Mechanism of Inhibition of the Reproduction of SARS-CoV-2 and Ebola Viruses by Remdesivir.* Biochemistry, 2021. **60**(24): p. 1869-1875.
28. Aranda, J., et al., *Mechanism of reaction of RNA-dependent RNA polymerase from SARS-CoV-2.* Chem Catal, 2022. **2**(5): p. 1084-1099.
29. Zhang, L. and R. Zhou, *Structural Basis of the Potential Binding Mechanism of Remdesivir to SARS-CoV-2 RNA-Dependent RNA Polymerase.* J Phys Chem B, 2020. **124**(32): p. 6955-6962.
30. Zhang, L. and R. Zhou, *Structural Basis of the Potential Binding Mechanism of Remdesivir to SARS-CoV-2 RNA-Dependent RNA Polymerase.* The Journal of Physical Chemistry B, 2020. **124**(32): p. 6955-6962.
31. Chowdhury, R., et al., *Atomistic De-novo Inhibitor Generation-Guided Drug Repurposing for SARS-CoV-2 Spike Protein with Free-Energy Validation by Well-Tempered Metadynamics.* Chem Asian J, 2021. **16**(12): p. 1634-1642.
32. Eswar, N., et al., *Comparative Protein Structure Modeling Using Modeller.* Current Protocols in Bioinformatics, 2006. **15**(1): p. 5.6.1-5.6.30.
33. Wishart, D.S., et al., *DrugBank: a comprehensive resource for in silico drug discovery and exploration.* Nucleic Acids Res, 2006. **34**(Database issue): p. D668-72.
34. Trott, O. and A.J. Olson, *AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.* J Comput Chem, 2010. **31**(2): p. 455-61.
35. Wang, J., et al., *Development and testing of a general amber force field.* J Comput Chem, 2004. **25**(9): p. 1157-74.
36. Frisch, M.J., et al., *Gaussian 09 Rev. A.01*. 2016: Wallingford, CT.
37. Sousa da Silva, A.W. and W.F. Vranken, *ACPYPE - AnteChamber PYthon Parser interfacE.* BMC Research Notes, 2012. **5**(1): p. 367.

38. Van Der Spoel, D., et al., *GROMACS: fast, flexible, and free.* J Comput Chem, 2005. **26**(16): p. 1701-18.
39. Hornak, V., et al., *Comparison of multiple Amber force fields and development of improved protein backbone parameters.* Proteins, 2006. **65**(3): p. 712-25.
40. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water.* The Journal of Chemical Physics, 1983. **79**(2): p. 926-935.
41. Deift, P. and X. Zhou, *A Steepest Descent Method for Oscillatory Riemann--Hilbert Problems. Asymptotics for the MKdV Equation.* The Annals of Mathematics, 1993. **137**(2).
42. Berendsen, H.J.C., et al., *Molecular dynamics with coupling to an external bath.* The Journal of Chemical Physics, 1984. **81**(8): p. 3684-3690.
43. Nosé, S., *A molecular dynamics method for simulations in the canonical ensemble.* Molecular Physics, 1984. **52**(2): p. 255-268.
44. Hess, B., et al., *LINCS: A linear constraint solver for molecular simulations.* Journal of Computational Chemistry, 1997. **18**(12): p. 1463-1472.
45. Darden, T., D. York, and L. Pedersen, *Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems.* The Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.
46. Barducci, A., G. Bussi, and M. Parrinello, *Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method.* Physical Review Letters, 2008. **100**(2): p. 020603.
47. Sasikala, W.D. and A. Mukherjee, *Structure and dynamics of proflavine association around DNA.* Phys Chem Chem Phys, 2016. **18**(15): p. 10383-91.
48. Sasikala, W.D. and A. Mukherjee, *Molecular mechanism of direct proflavine-DNA intercalation: evidence for drug-induced minimum base-stacking penalty pathway.* J Phys Chem B, 2012. **116**(40): p. 12208-12.
49. Fu, H., et al., *Finding an Optimal Pathway on a Multidimensional Free-Energy Landscape.* J Chem Inf Model, 2020. **60**(11): p. 5366-5374.
50. Wallace, A.C., R.A. Laskowski, and J.M. Thornton, *LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions.* Protein Eng, 1995. **8**(2): p. 127-34.

Chapter 5

Interaction between β-catenin and its binding partner T-cell factor (TCF)

5.1 Overview

β-catenin is a crucial protein involved in various biological processes, notably the Wnt Signaling Pathway. During adult tissue homeostasis, β-catenin governs a multitude of cellular activities. The interaction between β-catenin and its binding partner T cell factor (TCF) is pivotal for transducing the signaling pathway. Preventing the interaction between β-catenin and T-cell factor (TCF) emerges as a feasible strategy to inhibit the activation of Wnt target genes. The present study employs the multiple walker parallel bias metadynamics technique, using three collective variables to elucidate the mechanism of TCF unbinding from the interface of β-catenin. Our results correlate with the experimentally reported binding affinity calculations, and the present study reveals the folding-assisted unbinding mechanism governed by significant changes in the free energy landscape associated with the biomolecular events. Our findings align with experimentally reported binding affinity calculations. The multidimensional free energy landscape associated with the process reveals a folding-assisted unbinding pathway characterized by substantial changes in the secondary structure of TCF.

5.2 Introduction

Protein–protein interactions (PPIs) regulate diverse physiological processes in living organisms.[1] A specific protein–protein interface undergoes multiple stabilizations induced by the hydrogen bonding network, secondary structural changes, hydrophobic interactions, etc.[2] Nevertheless, disrupting these robust interactions at a given protein-protein interface using small molecules poses a considerable challenge. Several efforts have been undertaken to disrupt the protein-protein interactions by targeting the hotspots in the complex using small molecules.[3, 4]

The significance of the protein β-catenin lies in its crucial role within the Wnt signaling pathway, a pathway that governs the fate determination of cells.[5-7] A biologically significant interaction of β-catenin involves its binding to T-cell factor (TCF) family members, a process implicated in the onset of various cancer types, such as colorectal cancer.[8, 9] The predicted structure of the β-catenin:TCF complex suggests a tightly bound interface with a relatively high binding affinity, typically in the low nanomolar range. This is corroborated by free energy stability measurements ranging between -10.5 kcal/mol to -12.5 kcal/mol, as determined through isothermal titration calorimetry (ITC) experiments.[10-13] Despite the limited understanding of the unbinding mechanism, the disruption of this robustprotein–protein complex is considered a potential anticancer strategy.[6, 14-16] The structure of equilibrated TCF bound β-catenin is shown in Fig. 1a. β-catenin structure contains three regions, namely the central armadillo (ARM) domain, amino terminus domain (NTD), and carboxyl terminus domain (CTD).[17, 18] Nevertheless, the amino terminus domain (NTD) and carboxyl terminus domain (CTD) have been recognized for their flexibility and lack of defined structure, making them conventionally less favorable as targeting regions for studies.[19] The data from crystallization reveals that the resolved structure of β-catenin lacks the terminal domains.[20]

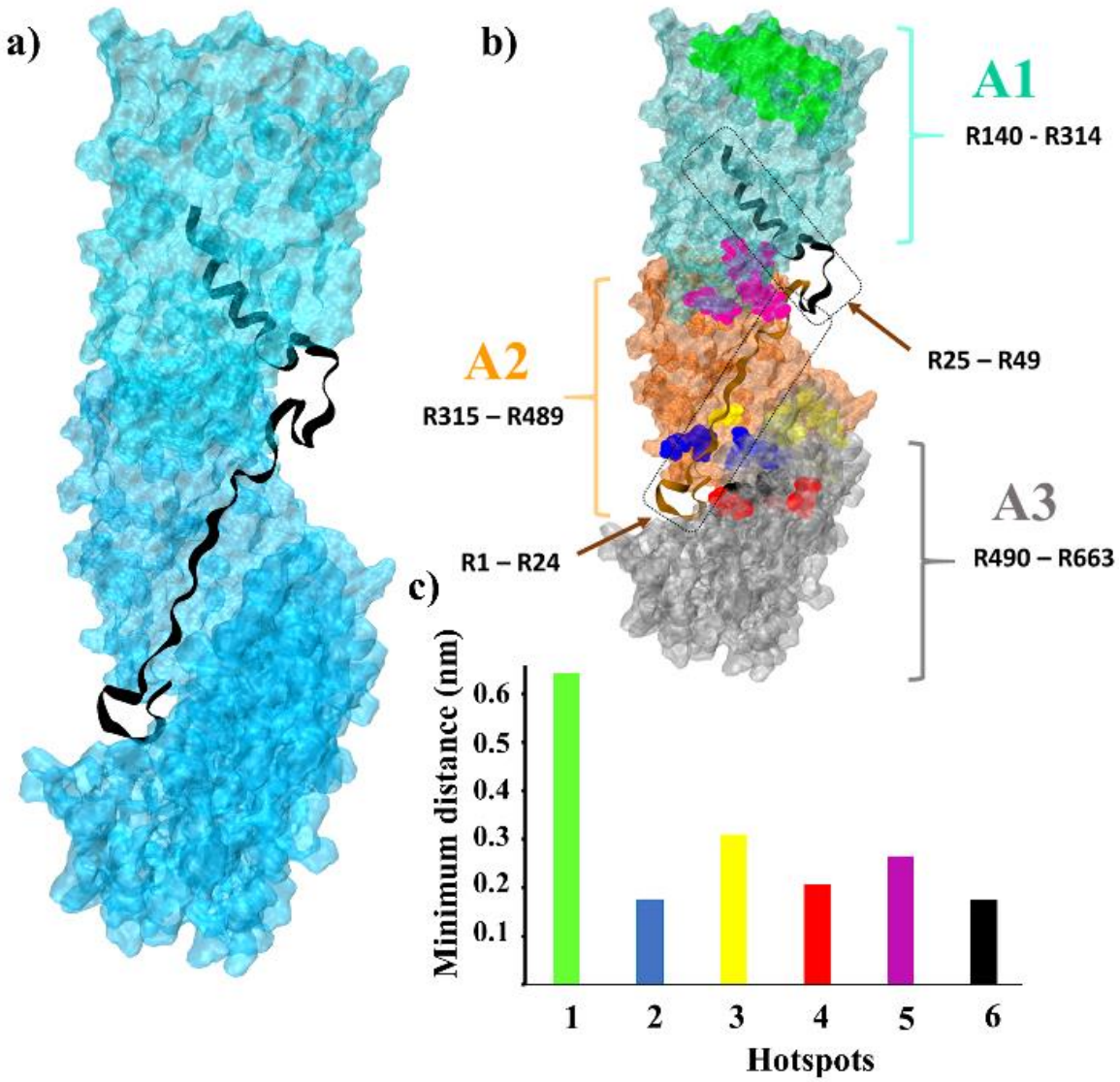

Figure 1. a) Representation of β-catenin: TCF complex. β-catenin and TCF are shown in cyan and black color, respectively. b) Illustration of hotspots and sub-regions in the β-catenin. The highlighted sub-regions, identified as A1 (R140-R314), A2 (R315–R489), and A3 (R490–R663), are depicted using light blue, orange, and grey colors, respectively. c) The minimum distance between the atoms in the hotspots of β-catenin and TCF, with hotspots 1-6 denoted by green, blue, yellow, red, violet, and black colors, respectively.

As a result, much of the research has been directed towards targeting the ARM domain, situated between the flexible amino-terminus domain (NTD) and carboxyl terminus domain (CTD). Moreover, an intriguing characteristic of β-catenin, encompassing the ARM domain, is the absence of flat surface regions and prominent deep binding pockets in its structure, making it conventionally categorized as undruggable.[21, 22] Consequently, due to its conventionally undruggable nature, the ARM domain of β-catenin, spanning from residue numbers 140 to 663

(R140-R663), is infrequently targeted in large-scale screenings. The binding partner TCF is a 49-residue long amino acid chain that tightly binds and covers the interface of β-catenin.

Biochemical assays and the crystal structure of the β-catenin:TCF complex based on experimental results unveil six crucial hotspots within the structure of the ARM domain of β-catenin.[15, 22-26] In the current study, to gain a more comprehensive understanding of the hotspots in the ARM region of β-catenin, the domain is subdivided into three distinct regions with specific residue ranges: A1 (R140-R314), A2 (R315–R489), and A3 (R490–R663). Further, the TCF (R1-R49) envelops the ARM region of β-catenin, forming the complex structure depicted in Fig. 1a and 1b. Table 1 presents the potential molecules reported in the literature, contributing to the likely hotspot regions within the ARM domain of β-catenin.[15]

As shown in the Fig 1b &c, TCF establishes contact with the majority of the hotspot regions within the ARM domains (A1, A2, and A3) of β-catenin. Therefore, gaining insights into the unbinding dynamics of TCF from the β-catenin interface will be pivotal for designing and developing novel drugs, particularly by targeting the identified hotspots.

Table 1. Details of the hotspots in the ARM domain of the β-catenin and corresponding targeted molecules[15] in the literature.

Hotspot	Molecule name	Residue numbers of the interacting residues
1	4FNPC	145, 148, 152, 155, 156, 159, 167, 171, 174, 177, 178, 181
2	UU-T02	430, 435, 469, 508
3	HI-B1	425, 458, 460, 462, 469, 505
4	PNU-74654	435, 469, 511, 515, 519
5	UU-T01	303, 304, 306, 312
6	MSAB	469, 508, 514, 546, 571

The current study aims to investigate the mechanistic pathways and energetics associated with the unbinding of TCF from the β-catenin interface using multiple walker variants of parallel bias metadynamics[27].

5.3 Computational details

5.3.1 System preparation

The initial structure of the TCF bound β-catenin was taken from the Protein Data Bank (PDB ID: 1JPW). Modeller 9.21[28] was utilized to model the missing residues in the β-catenin:TCF

complex. Subsequently, the complex was solvated in a 20 x 20 x 16 nm^3 box using approximately 206,045 TIP3P water molecules. The system was maintained at physiological ion concentration, with 150 mM of Na^+ and Cl^- ions. Additionally, extra Cl^- ions were added to neutralize the system. The topology of the system was constructed using the AMBER99SB[29] forcefield. The molecular dynamics simulations were conducted using GROMACS 2019.6[30, 31] patched with plumed 2.6[32, 33].

5.3.2 Simulation details

The modelled system underwent energy minimization using steepest descent[34] method for 10000 steps. Additionally, the system is heated to 300K in 200 ps using Berendsen thermostat and barostat[35] with a coupling constant of 0.6 ps. During the heating process, position restraints of 25 kcal/mol/$Å^2$ were imposed on heavy atoms. Subsequently, the system underwent equilibration at a constant temperature of 300 K and pressure of 1 bar for 2 ns, without any restraints. The equilibration process employed the same thermostat and barostat as used in the heating process, with a coupling constant of 0.2 ps for each. Further 5 ns NVT equilibration was carried out without any restraints. Nosé-Hoover thermostat[36] with a coupling constant of 0.2 ps was used for this process. The simulation employed the LINCS algorithm[37] to constrain all bonds throughout the duration of the process. Electrostatic interactions during the simulation were computed using the Particle Mesh Ewald (PME) method[38]. A cutoff distance of 10 Å was applied for both van der Waals (vdW) and electrostatic long-range interactions. A time step of 2 fs was used for all the simulations.

5.3.3 Free energy calculations

For free energy calculations, we have used parallel bias metadynamics[27]using three collective variables (CVs) using multiple walker[39] strategy. We employed parallel bias metadynamics (27) by utilizing multiple walker[39] version for free energy calculations. In the parallel bias metadynamics setup, the parameters were configured as follows: A Gaussian bias was applied to the system with a deposition rate of 1 ps. An initial Gaussian hill height of 1.0 kJ/mol was utilized, and a bias factor of 10 was set. The Gaussian

width for the CVs X, Contacts, and R_g (definition of CVs are discussed later) was specified as 0.15 nm, 2.0, and 0.1 nm, respectively. We employed a total of five walkers, each initiated at different starting points within the CV space. The starting geometry and corresponding values of the CVs for each walker are shown in Fig. 2.

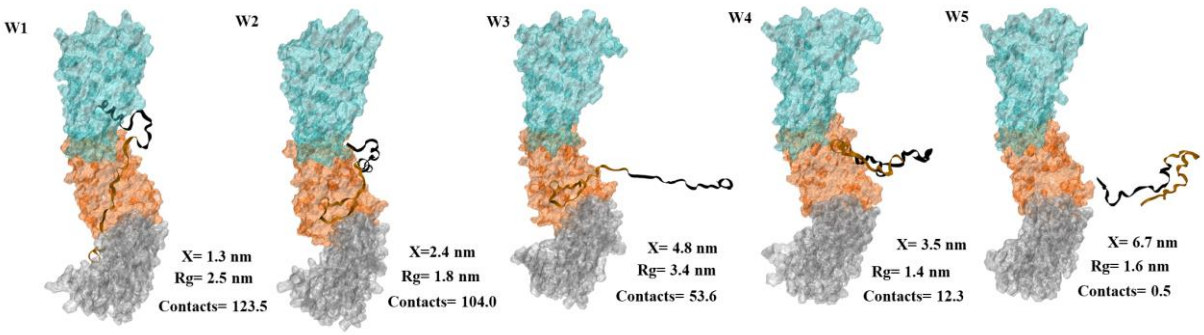

Figure 2. Starting structures of β-catenin:TCF complex used for multiple walker parallel bias metadynamics simulations along with the CV values associated with the structures.

Each walker underwent a simulation of 100 ns, leading to a cumulative simulation time of 500 ns (5 walkers x 100 ns = 500 ns). The reweighting of metadynamics potential was performed using the final bias-based reweighting[32, 40] approach. Bias contribution from each of the walker in the CV space was utilized for reweighting and construction of multidimensional free energy landscape. The MULE[41] program was utilized to construct the Minimum Free Energy Path (MFEP), connecting bound states to unbound states.

5.3.4 Choice of collective variables

In the study, three CVs were employed to elucidate the free energy change and mechanistic aspects associated with the unbinding of TCF from the β-catenin interface. These variables include DISTVEC (displacement along a body-fixed vector), contacts, and radius of gyration. The construction and details of each of these CVs are discussed below.

a) DISTVEC (*X*): This CV is defined as a projection of $\vec{d}$ on a body-fix unit vector $\hat{b}$ and it is defined as follows;

$$X = \hat{b}.\vec{d}$$

Where, $\vec{d}$ is a vector defined from the center of mass (COM) of heavy atoms in residues 312, 313, 315, 316, 321, 347, 348, 350-358 to the COM of heavy atoms in the residues 219, 253, 254, 256, 257, 260, 261, 264, 265, 290, 292, 293, 295, 296, 299, 302, 303, 306, 307, 312, 333, 335, 338, 339, 342, 345, 346, 349, 354, 376, 383, 386, 387, 389, 390, 393, 422, 425, 426, 428-430, 435, 462, 463, 466, 469, 470, 473, 474, 508, 568, 647, 649. Similarly, $\hat{b}$ is a vector defined from the COM of heavy atoms in the residues 312, 313, 315, 316, 321, 347, 348, 350-358 to the COM of the Cα atoms present in the TCF. The angle θ is defined as $\theta = \cos^{-1}(\hat{b}.\vec{d}/|d|)$.

b) Contacts: The contacts CV is defined based on the spatial proximity of two groups, denoted as the Cα atoms in β-catenin (*gA*) and TCF (*gB*), within a specified cutoff distance. It is expressed as follows;

$$Contacts = \sum_{i \in gA} \sum_{i \in gB} s_{ij}$$

Where s_{ij} = 1 if any contact is established between the atoms *i* and *j,* else s_{ij} = 0. In order to make sure that the CV is continuous, s_{ij} is defined by the switching function as follows

$$s_{ij} = \sum_{i} \frac{1 - \left({}^{r_i}/_{r_0}\right)^n}{1 - \left({}^{r_i}/_{r_0}\right)^m}$$

The values for n, m, and r_0 (cutoff distance) were set to 6, 12, and 7 Å, respectively.

c) Radius of gyration (R_g): R_g is defined as follows;

$$R_g = \left(\frac{\sum_i^n m_i |r_i - r_{COM}|^2}{\sum_i^n m_i}\right)^{1/2}$$

where r_{COM} is defined by

$$r_{COM} = \frac{\sum_i^n r_i \, m_i}{\sum_i^n m_i}$$

The Cα atoms present in the TCF was considered for the R_g calculations.

5.4 Results & Discussion

5.4.1 3D free energy landscape

The 3D free energy landscape associated with the unbinding of TCF from the β-catenin interface is shown in the Fig. 3. The projected 3D free energy landscape is constructed with respect to the biased CVs. Six states, namely stable bound minima (B1, B2, B3), transition state (T), and unbound states (U1 and U2), were identified based on the biased CV ranges (Fig.4).

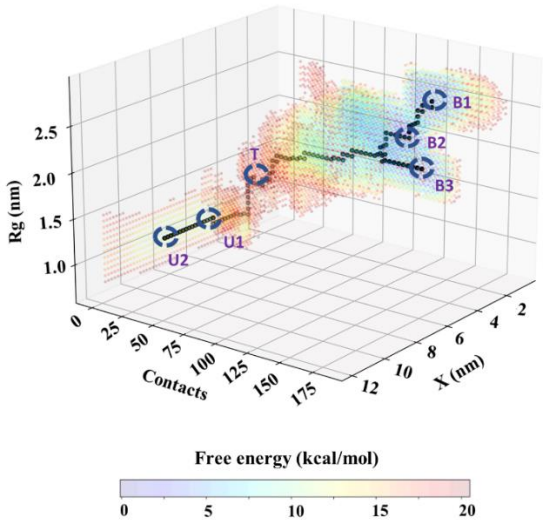

Figure 3. 3D free energy surface with respect to Rg, Contacts, and *X*, depicting the unbinding process of TCF from β-catenin. B1, B2, and B3 denote distinct states of TCF-bound β-catenin. The pathway with the lowest free energy, connecting the bound states (B1, B2, and B3) to unbound states (U1 and U2) through the transition state (T), is illustrated in black dotted line.

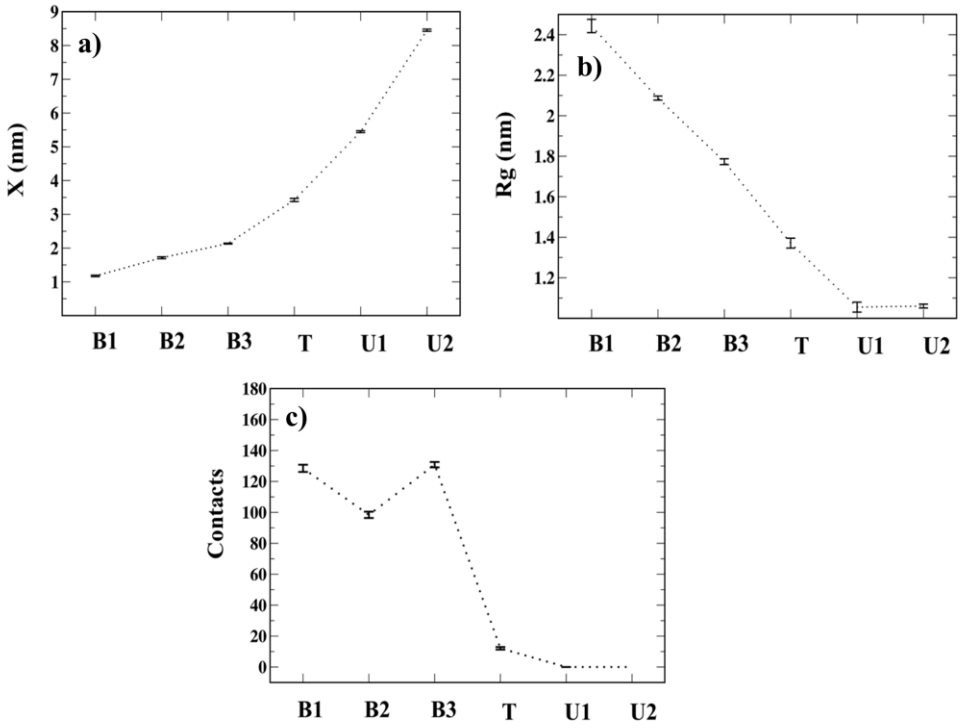

Figure 4. Biased CV values in the states B1, B2, B3, T, U1 and U2. a) *X* b) R_g c) contacts

The rise in the *X* value, progressing from B1 (initial state) to U2 (terminal state), confirms the movement of TCF away from β-catenin toward a fully solvated state. The extent of compactness in TCF throughout the sampling is anticipated to be disclosed by the R_g parameter.

The CV 'contacts' also captures the local structural changes within the bound states. The criteria for defining the bound states are X < 3 nm, R_g > 1.6 nm, and contacts > 90, indicating the presence of proximity between β-catenin and TCF. In contrast, the unbound states are characterized by the absence of any contact between β-catenin and TCF, achieved when TCF is entirely separated from the β-catenin interface, i.e., contacts = 0. The black dotted lines in Fig. 3 depict the minimum free energy path connecting the bound states (B1, B2, and B3) to the unbound states (U1 and U2) through the transition state (T). The variation of free energy along this minimum path is illustrated in Fig. 5

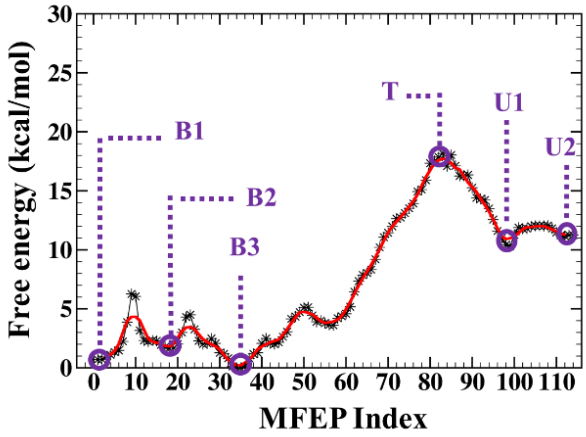

Figure. 5 The relevant states (B1, B2, B3, T, U1, and U2) are labeled according to the 3D free energy surface presented in Fig. 3.

Table 1 displays the free energy changes between the bound states (B1, B2, and B3) and the unbound state U2. Notably, the free energy change between the most stable bound state (B3) and the unbound state (U2), denoted as $\Delta G_{B3\text{-}U2}$ is calculated to be -11.1 kcal/mol. This value aligns well with experimental findings[12, 13], falling within the range of -10.5 kcal/mol to -12.5 kcal/mol. Fig. 6a displays the 1D free energy profiles at various time instances projected with respect to each biased CV indicating the convergence of simulations. Mean free energy and the error estimate associated with the free energy profile with respect to *X*, R_g and contacts during the last 100 ns of simulation is shown in the Fig 6b.

Table 2. Relative free energy stability values associated with the bound states (B1, B2 and B3) and unbound state U2.

Sl. No	Free energy change	Free energy stability (kcal/mol)
1	$\Delta G_{B1\text{-}U2}$	-10.2
2	$\Delta G_{B2\text{-}U2}$	-9.2
3	$\Delta G_{B3\text{-}U2}$	-11.1

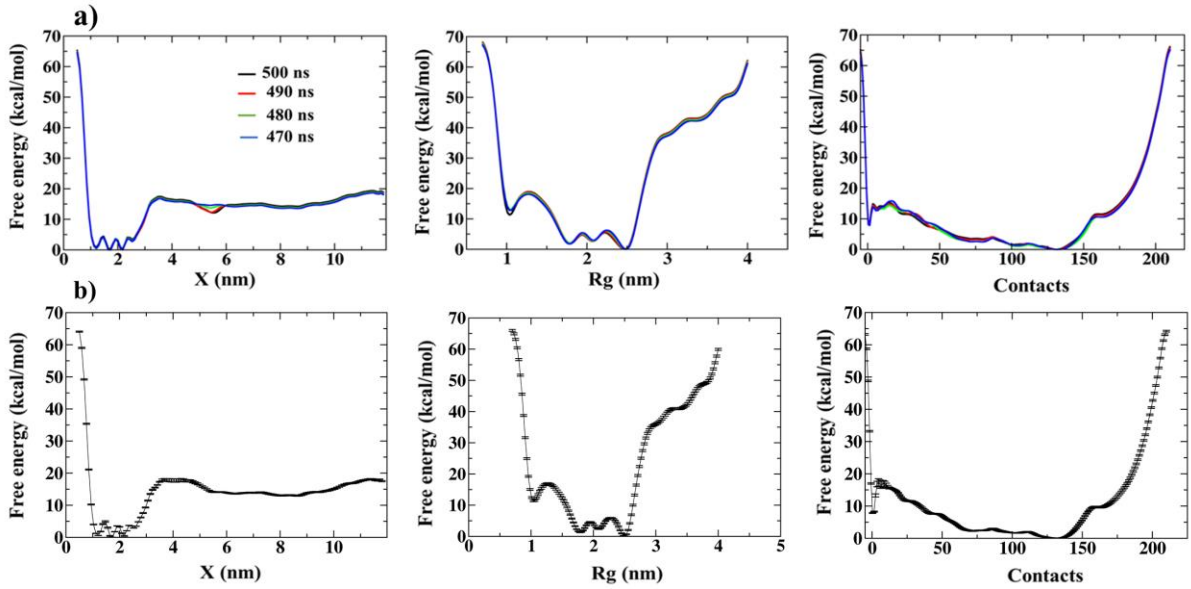

Figure. 6 a) 1D Free energy profiles with respect to *X*, R_g and contacts at different time instances. b) Mean free energy and the error estimate associated with the free energy profile with respect to *X*, R_g and contacts during the last 100 ns of simulation.

5.4.2 Unbinding mechanism

In order to comprehend the process of TCF unbinding at the β-catenin interface, we calculated the minimum distance of atoms from sub-regions A1, A2, and A3 (as shown in Fig. 1b) within the structural clusters associated with the bound, unbound states, and transition state. The minimum distance profile is shown in Fig. 7. In the stable bound states B1 and B2, it was observed that TCF is distributed across the sub-regions A1, A2, and A3 of β-catenin, as indicated by the lowest minimum distance values from A1, A2, and A3. However, upon transitioning to the B3 state, it was observed that TCF exhibits a higher minimum distance from the A1 sub-region compared to the values in the B1 and B2 states. Additionally, the minimum

distance values remain unchanged with respect to the A2 and A3 sub-regions in the B3 state, relative to the B1 and B2 states. This suggests that the initiation of TCF unbinding occurs from the A1 sub-region.

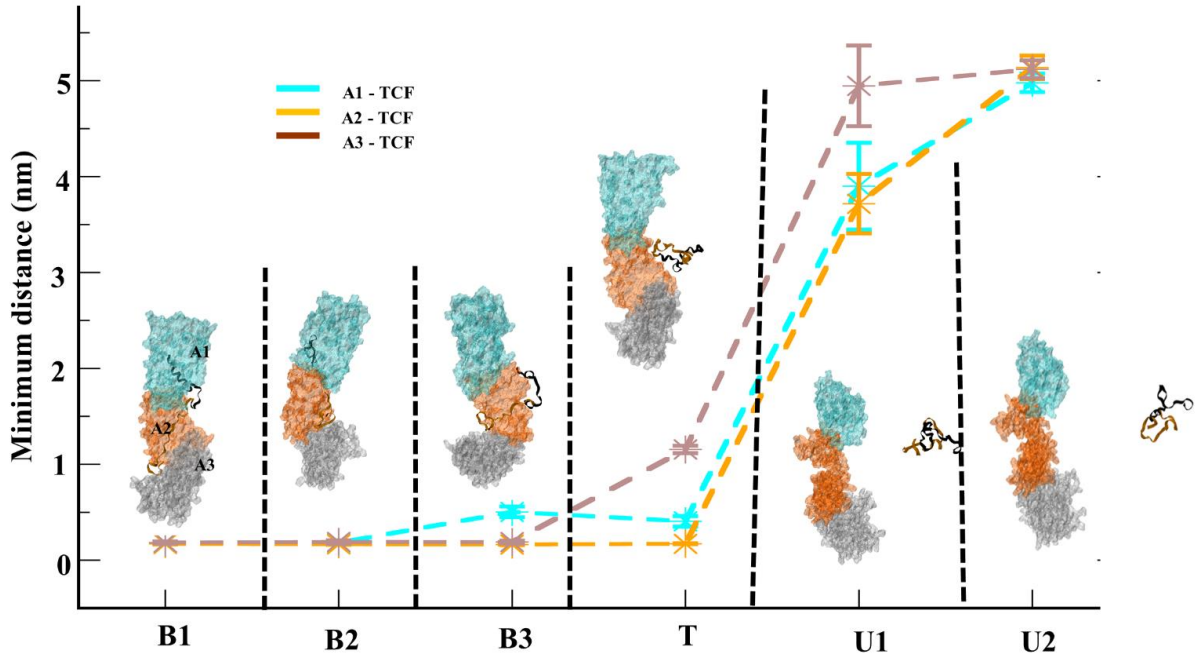

Figure 7. Minimum distance profiles in the states B1, B2, B3, T, U1, and U2. The minimum distance between the TCF and A1, A2 and A3 regions are shown by cyan, orange and brown color respectively.

In the transition state, the minimum distance of TCF towards the A2 sub-region remains unchanged compared to the bound states (B1, B2, and B3). However, notable changes, characterized by an increase in minimum distance, have been observed towards the A1 and A3 sub-regions in the transition state. This suggests that TCF is completely detached from the end sub-regions A1 and A3, now maintaining contact only with the central A2 region. Ultimately, the initiation of TCF detachment occurs from the central A2 region, resulting in no contact with TCF in the unbound states.

5.4.3 Structural changes in TCF during unbinding

The unbinding of TCF from β-catenin is facilitated by notable structural changes in TCF. As depicted in Fig. 4, there is a consecutive decrease in TCF's radius of gyration (Rg) from the B1 state to the U1 state. This suggests that TCF becomes more compact in the unbound state (U1). No notable changes in the radius of gyration (Rg) are observed from the U1 state to the U2 state. This lack of variation is anticipated since these states are indistinguishable, given the absence of contact between TCF and β-catenin. The sole distinction lies in the increase in the

X value, indicating a transition to a deeper solvated state. The only difference lies in the increase in the *X* value, indicating a deeper solvated state. The resemblance in structural features leads to similar free energy values in the U1 and U2 states, as evidenced in Figs. 3 and 5, illustrating a relatively flat landscape. The successive decrease in the radius of gyration (Rg) of TCF during unbinding correspondingly led to alterations in the secondary structure of TCF. The variations in secondary structural elements in these states are depicted in Fig. 8.

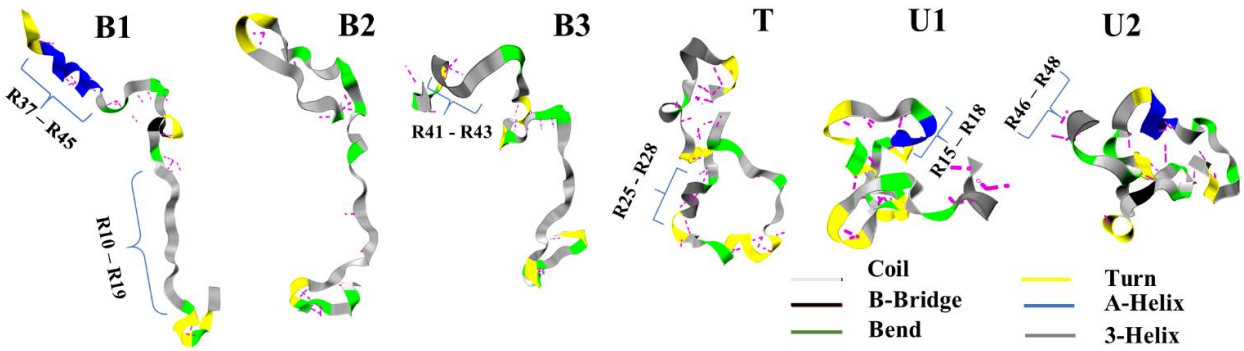

Figure 8. Structures of TCF in the states B1, B2, B3, T, U1 and U2. The secondary structural elements are shown by color codes as indicated in the inset. The important residue junctions are highlighted wherever necessary.

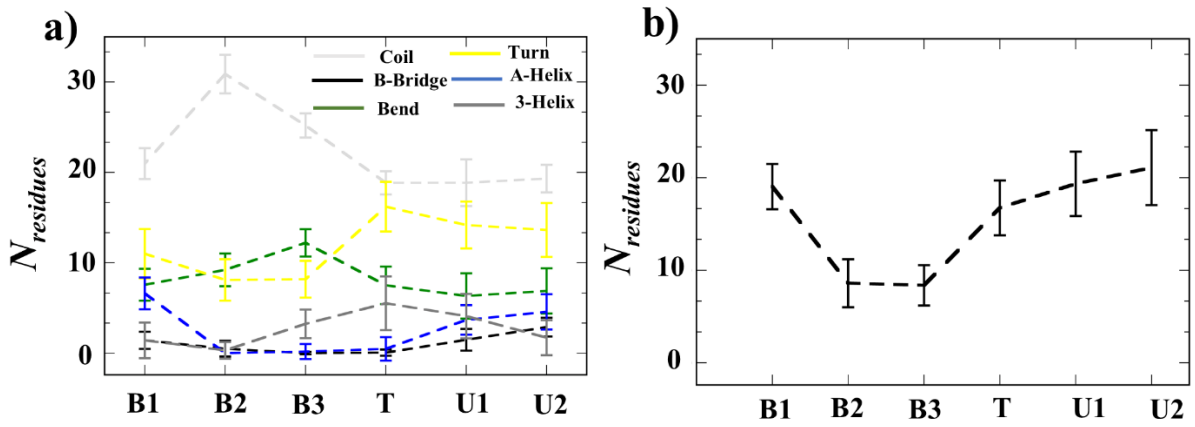

Figure 9. a) The fluctuation in the number of residues participating in the formation of various structural elements (Coil, B-Bridge, Bend, Turn, A-Helix, 3-Helix) in TCF is depicted along the unbinding pathway b) The fraction of secondary structures in states B1, B2, B3, T, U1, and U2 as defined by the equation: *Secondary structure = A-Helix + B-Bridge + Turn*

As depicted in Fig. 9a, the bound states exhibit a higher prevalence of random coil structures compared to the unbound states. Notably, the B2 state displays a greater fraction of random

coil structures among the bound states. This observation might contribute to the lower free energy stability of the B2 state when compared to the B1 and B3 states, as indicated in Fig. 5.

An additional intriguing observation is that the transition state and the unfolded states exhibit a slight advantage in forming turns compared to the bound structures, contributing to a more compact conformation of TCF. The presence of multiple turns in these states may have facilitated the folding of TCF, leading to a lower radius of gyration (Rg) value along the unbinding pathway. This suggests the possibility of an assisted unbinding mechanism. The secondary structure content contribution of TCF, based on the DSSP algorithm, can be calculated using the formula: *Secondary Structure = A-Helix + B-Bridge +Turn.* This equation represents the combined contribution of alpha-helices, beta-bridges, and turns to the overall secondary structure content of TCF as determined by the DSSP[42] algorithm. The secondary structure content according to the above equation along the unbinding pathway is shown in the Fig. 9b.

The B1 state, characterized by greater stability, demonstrates a higher level of structural content in comparison to the other bound states. Additionally, as TCF reaches the transition state and unfolded states, there is an increased fraction of residues contributing to the secondary structure content. The variation in secondary structural forms in TCF on a residue-wise basis during the unbinding pathway is illustrated in Fig. 10.

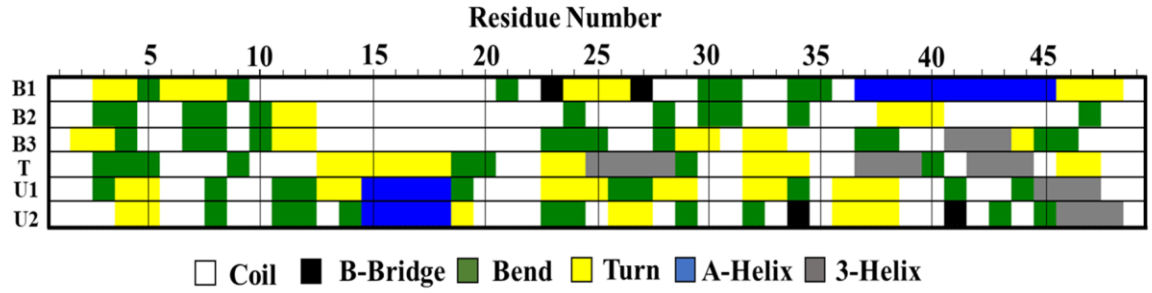

Figure 10. residue-wise variation of secondary structural content in TCF is depicted for the states B1, B2, B3, T, U1, and U2, with the corresponding structural elements shown in the inset.

5.4.4 Hydrogen bond stabilization

The fluctuations in intermolecular hydrogen bonds between β-catenin and TCF, as well as the establishment of intramolecular hydrogen bonds within TCF during the unbinding pathway, are illustrated in Fig. 11. Notably, the intramolecular hydrogen bonds formed in TCF, as depicted in Fig. 11, align with the secondary structural content presented in Fig. 9b.

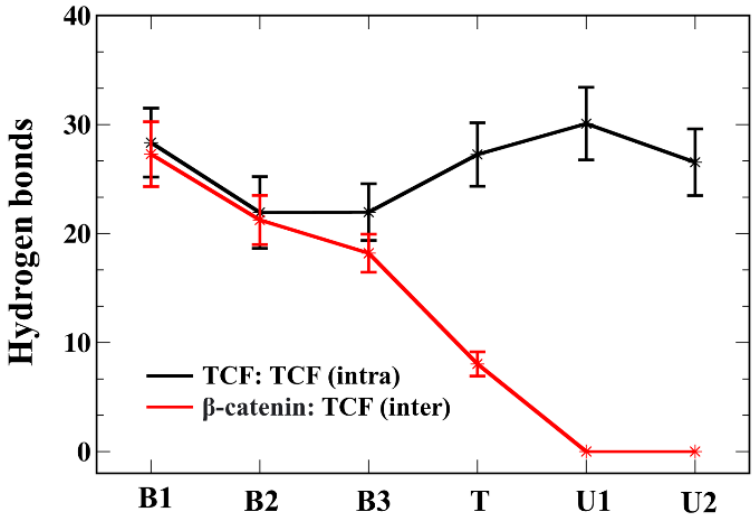

Figure 11. Variation of intramolecular hydrogen bonds in TCF (depicted in black) and intermolecular hydrogen bonds (depicted in red) between β-catenin and TCF across different states during the unbinding pathway.

These results indicate that hydrogen bonds play a role in facilitating the folding-assisted unbinding mechanism. The transition from B3 to T is characterized by an increase in intramolecular hydrogen bonds, supporting the folding process (as evidenced by the low R_g value). Concurrently, there is a decrease in intermolecular hydrogen bonds between β-catenin and TCF, initiating the unbinding events.

5.5 Conclusion

In this study, we explored the molecular events associated with unbinding TCF from the β-catenin interface by employing parallel bias metadynamics, an enhanced sampling technique. Our approach involved the utilization of a multiple walker version of metadynamics for effective sampling, incorporating CVs such as X, R_g, and contacts. The observed free energy change associated with the transition from the bound state to the unbound state is determined to be -11.1 kcal/mol, aligning well with experimental results. Our findings indicate that the unbinding process was facilitated by various folding of TCF, characterized by substantial secondary structural changes. These observations also validate the significance of

intermolecular and intramolecular hydrogen bond interactions triggering the unbinding of TCF from the β-catenin interface.

5.6 References

1. Wells, J.A. and C.L. McClendon, *Reaching for high-hanging fruit in drug discovery at protein–protein interfaces.* Nature, 2007. **450**(7172): p. 1001-1009.
2. Jones, S. and J.M. Thornton, *Principles of protein-protein interactions.* Proc Natl Acad Sci U S A, 1996. **93**(1): p. 13-20.
3. Arkin, M.R. and J.A. Wells, *Small-molecule inhibitors of protein-protein interactions: progressing towards the dream.* Nat Rev Drug Discov, 2004. **3**(4): p. 301-17.
4. Clackson, T. and J.A. Wells, *A Hot Spot of Binding Energy in a Hormone-Receptor Interface.* Science, 1995. **267**(5196): p. 383-386.
5. Huelsken, J. and J. Behrens, *The Wnt signalling pathway.* J Cell Sci, 2002. **115**(Pt 21): p. 3977-8.
6. Liu, J., et al., *Wnt/beta-catenin signalling: function, biological mechanisms, and therapeutic opportunities.* Signal Transduct Target Ther, 2022. **7**(1): p. 3.
7. Valenta, T., G. Hausmann, and K. Basler, *The many faces and functions of beta-catenin.* EMBO J, 2012. **31**(12): p. 2714-36.
8. Omer, C.A., et al., *Identification of Tcf4 residues involved in high-affinity beta-catenin binding.* Biochem Biophys Res Commun, 1999. **256**(3): p. 584-90.
9. Knapp, S., et al., *Thermodynamics of the high-affinity interaction of TCF4 with beta-catenin.* J Mol Biol, 2001. **306**(5): p. 1179-89.
10. Graham, T.A., et al., *Crystal Structure of a β-Catenin/Tcf Complex.* Cell, 2000. **103**(6): p. 885-896.
11. Poy, F., et al., *Structure of a human Tcf4–β-catenin complex.* Nature Structural Biology, 2001. **8**(12): p. 1053-1057.
12. Sun, J. and W.I. Weis, *Biochemical and structural characterization of beta-catenin interactions with nonphosphorylated and CK2-phosphorylated Lef-1*. J Mol Biol, 2011. **405**(2): p. 519-30.
13. Fasolini, M., et al., *Hot spots in Tcf4 for the interaction with beta-catenin.* J Biol Chem, 2003. **278**(23): p. 21092-8.
14. MacDonald, B.T., K. Tamai, and X. He, *Wnt/beta-catenin signaling: components, mechanisms, and diseases.* Dev Cell, 2009. **17**(1): p. 9-26.

15. Cui, C., et al., *Is beta-Catenin a Druggable Target for Cancer Therapy?* Trends Biochem Sci, 2018. **43**(8): p. 623-634.
16. Morin, P.J., *β-catenin signaling and cancer.* BioEssays, 1999. **21**(12): p. 1021-1030.
17. Huber, A.H., W.J. Nelson, and W.I. Weis, *Three-Dimensional Structure of the Armadillo Repeat Region of β-Catenin.* Cell, 1997. **90**(5): p. 871-882.
18. Dar, M.S., et al., *Terminal regions of beta-catenin are critical for regulating its adhesion and transcription functions.* Biochim Biophys Acta, 2016. **1863**(9): p. 2345-57.
19. Xing, Y., et al., *Crystal structure of a full-length beta-catenin.* Structure, 2008. **16**(3): p. 478-87.
20. Xing, Y., et al., *Crystal Structure of a Full-Length β-Catenin.* Structure, 2008. **16**(3): p. 478-487.
21. Lazo, J.S. and E.R. Sharlow, *Drugging Undruggable Molecular Cancer Targets.* Annual Review of Pharmacology and Toxicology, 2016. **56**(1): p. 23-40.
22. Dang, C.V., et al., *Drugging the 'undruggable' cancer targets.* Nature Reviews Cancer, 2017. **17**(8): p. 502-508.
23. Graham, T.A., et al., *Tcf4 can specifically recognize β-catenin using alternative conformations.* Nature Structural Biology, 2001. **8**(12): p. 1048-1052.
24. Knapp, S., et al., *Thermodynamics of the high-affinity interaction of TCF4 with β-catenin1 1Edited by A. R. Fersht.* Journal of Molecular Biology, 2001. **306**(5): p. 1179-1189.
25. von Kries, J.P., et al., *Hot spots in β-catenin for interactions with LEF-1, conductin and APC.* Nature Structural Biology, 2000. **7**(9): p. 800-807.
26. Gail, R., R. Frank, and A. Wittinghofer, *Systematic Peptide Array-based Delineation of the Differential β-Catenin Interaction with Tcf4, E-Cadherin, and Adenomatous Polyposis Coli*.* Journal of Biological Chemistry, 2005. **280**(8): p. 7107-7117.
27. Pfaendtner, J. and M. Bonomi, *Efficient Sampling of High-Dimensional Free-Energy Landscapes with Parallel Bias Metadynamics.* J Chem Theory Comput, 2015. **11**(11): p. 5062-7.
28. Eswar, N., et al., *Comparative Protein Structure Modeling Using Modeller.* Current Protocols in Bioinformatics, 2006. **15**(1): p. 5.6.1-5.6.30.
29. Hornak, V., et al., *Comparison of multiple Amber force fields and development of improved protein backbone parameters.* Proteins, 2006. **65**(3): p. 712-25.

30. Van Der Spoel, D., et al., *GROMACS: fast, flexible, and free.* J Comput Chem, 2005. **26**(16): p. 1701-18.
31. Abraham, M.J., et al., *GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers.* SoftwareX, 2015. **1-2**: p. 19-25.
32. Tribello, G.A., et al., *PLUMED 2: New feathers for an old bird.* Computer Physics Communications, 2014. **185**(2): p. 604-613.
33. consortium, P., *Promoting transparency and reproducibility in enhanced molecular simulations.* Nat Methods, 2019. **16**(8): p. 670-673.
34. Press, W.H., et al., *Numerical Recipes in FORTRAN; The Art of Scientific Computing*. 1993: Cambridge University Press.
35. Berendsen, H.J.C., et al., *Molecular dynamics with coupling to an external bath.* The Journal of Chemical Physics, 1984. **81**(8): p. 3684-3690.
36. Nosé, S., *A molecular dynamics method for simulations in the canonical ensemble.* Molecular Physics, 1984. **52**(2): p. 255-268.
37. Hess, B., et al., *LINCS: A linear constraint solver for molecular simulations.* Journal of Computational Chemistry, 1997. **18**(12): p. 1463-1472.
38. Darden, T., D. York, and L. Pedersen, *Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems.* The Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.
39. Raiteri, P., et al., *Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics.* J Phys Chem B, 2006. **110**(8): p. 3533-9.
40. Bonomi, M., A. Barducci, and M. Parrinello, *Reconstructing the equilibrium Boltzmann distribution from well-tempered metadynamics.* J Comput Chem, 2009. **30**(11): p. 1615-21.
41. Fu, H., et al., *Finding an Optimal Pathway on a Multidimensional Free-Energy Landscape.* J Chem Inf Model, 2020. **60**(11): p. 5366-5374.
42. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-637.

Chapter 6

Molecular insights into the stereospecificity of arginine in RNA tetraloop folding

6.1 Overview

The RNA-based theory on the origin of homochirality suggests that only one stereoisomer of amino acids is responsible for providing extra stability to RNA folding, acting as a chaperone. However, the underlying mechanism for this observation is not well understood in molecular detail. The study focuses on understanding the molecular-level aspects associated with this hypothesis using enhanced sampling methods. A controlled sampling approach is utilized in the present study by a combination of umbrella sampling and parallel bias metadynamics simulations to construct a multidimensional free energy landscape associated with the folding-unfolding equilibrium of a small GAAA RNA tetraloop motif in the presence of enantiomers of arginine. The results show that D-arginine stabilizes the native folded state of RNA, while L-arginine destabilizes it. Moreover, the study identifies a specific geometric constraint that allows D-arginine to stack with the terminal base pairs of RNA; on the other hand, L-arginine tends to bind in the groove.

6.2 Introduction

The fundamental components of biomolecules related to life are present in a definite chiral form. The proteins are known to be made of L-enantiomers of amino acids; on the other hand, sugars in nucleic acids carry D-enantiomers.[1, 2] The theories suggest that in the RNA world preceding the current protein machinery we are seeing today, stereoisomers of amino acids functioned as regulator motifs in a specific manner, assisting the folding of RNA motifs.[3-5] Among various RNA motifs known, tetraloop structures are a unique three-dimensional motif in which most RNA bases exist as pairs, forming the stem portion through Watson-Crick base pairing[6]. Meanwhile, the loop portion of the RNA is formed by an unpaired region with a specific sequence, connecting the stem regions. The dynamics of these motifs are well studied to understand the conformational changes, stability, and receptor recognition using spectroscopic studies.[7-10]. These types of motifs are characterized by a notable trait in which they undergo spontaneous transitions between their native folded states, resulting in a variety of misfolded and extended single-stranded structures.[11-14].

It is known that the kinetics and thermodynamics associated with the folding-unfolding equilibria of tetraloop structures are significantly influenced by the various complexities of amino acids, as revealed by recent experimental FRET studies.[15, 16] These results partially validate arguments based on theories related to the chiral amino acid-based origin of life by showing the enantiospecific effects of amino acids on RNA motifs in the tertiary level of interactions.[15] Computational studies on RNA tetraloop structure were also well established to understand the mechanistic pathways of folding-unfolding events.[12-14, 17-25] However, computational studies exploring the impact of chirality-dependent folding-unfolding are not extensively explored, and molecular details & driving force causing the specific effect of enantiomers on folding–unfolding events is ill-understood in the literature.

Our study aims to understand the chirality-dependence of a given GAAA tetraloop motif of RNA (sequence 5’-gggcGAAAgccc3') in the presence of enantiomers of arginine. Arginine is unique compared to other natural amino acids because of its diverse binding capabilities due to the presence of polar groups at the two ends and the long hydrophobic tail. It is known to facilitate stacking interactions as well as groove binding towards nucleic acids, including RNA motifs. [26-28]. It is to be noted that the stability & specificity of RNA motifs towards amino acids will be highly dependent on the orientation of side groups, the surface of interaction, and the complexity of RNA motifs.[28] We conducted separate folding and

binding studies with the aid of enhanced sampling techniques to gain a molecular-level understanding of the stereospecificity of arginine in the RNA tetraloop dynamics. For the construction of a multidimensional free-energy landscape associated with folding-unfolding equilibria, we have utilized a controlled sampling approach by combining umbrella sampling and parallel bias metadynamics. However, for binding studies, well-tempered metadynamics simulations were utilized.

6.3 Computational details

6.3.1 System preparation

The initial structure of the 12-mer GAAA RNA tetraloop is obtained from the protein data bank (PDB ID: 1ZIF) and further modified to get the target sequence of interest 5'-gggcGAAAgccc3'. The structure of the GAAA RNA tetraloop is shown in Fig. 1a. We have used the amber forcefield-based parameter set for the arginine molecules, which is reported in the literature.[29]. For the RNA, we have used ff99bsc0χOL3 forcefield for parameterization.[19, 30, 31] The study consists of two systems specifically for D-arginine and L-arginine, respectively. The arginine structure and its characteristic groups (guanidinium, alkyl, and Cα groups) are shown in the Fig. 1b.

For folding studies in the presence of arginine, the RNA motif was inserted into a 9 X 9 X 9 nm^3 cubic box, and 132 arginine molecules were added to maintain the concentration of 300 mM. Further, the system is solvated using ~22,500 TIP3P[32] water molecules, and 121 Cl^- was added to neutralize the overall charge of the system. However, in the system without arginine, 150 mM concentration of Na^+ and Cl^- (physiological ion concentration) was maintained along with extra Na^+ ions to neutralize the system, and for single molecule binding studies, the most probable arginine-bound RNA tetraloop in the same physiological concentration environment was considered for the study.

6.3.2 Simulation details

The constructed systems were undergone energy minimizations using the steepest descent method[33] for 10000 steps. This was followed by heating the system to 300K using a Berendsen thermostat and barostat[34] for 200 ps. Position restraints of 25 kcal/mol/$Å^2$ were applied to heavy atoms during the process, and a coupling constant of 0.6 ps was used during the heating process. Further, Equilibration was carried out at a constant temperature of 300 K and 1 bar pressure using Berendsen thermostat and barostat[34] for 2 ns without any position

restraints with a coupling constant of 0.2 ps. Finally, the system underwent 10 ns NVT simulation using Nosé-Hoover thermostat[35] with a coupling constant of 0.2 ps. A time step of 2 fs was assigned to each simulation. Throughout the simulation, the Particle Mesh Ewald (PME) method[36] was utilized for handling electrostatics, and the LINCS algorithm[37] was utilized to constrain the bonds. The van der Waals (vdW) and long-range electrostatic interactions were subjected to distance cutoffs of 10 Å. GROMACS 2019.6[38, 39], patched with plumed 2.6[40, 41], was employed for conducting all molecular dynamics simulations and further free energy calculations (discussed later).

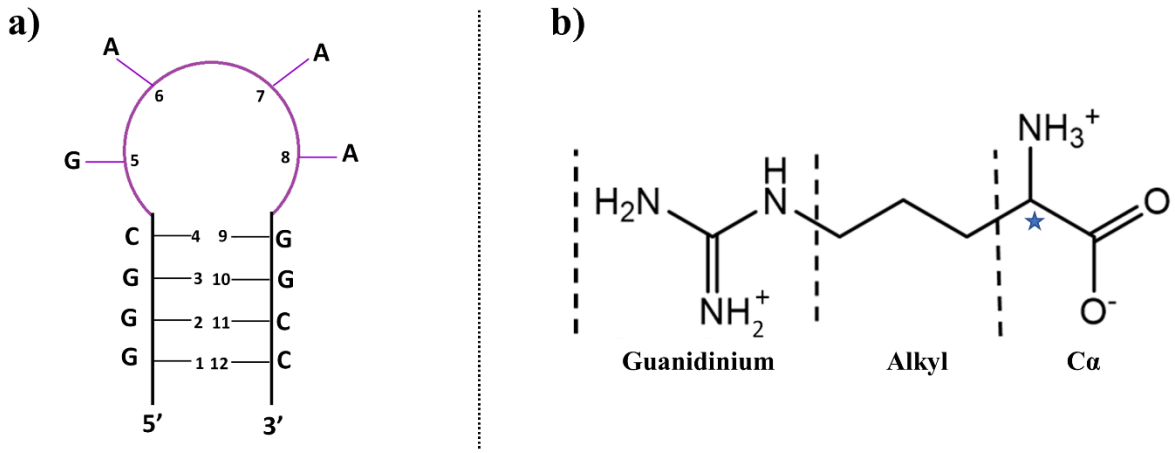

Figure 1. a) Representation of GAAA RNA tetraloop structure. b) Representation of Guanidinium, alkyl, and Cα groups in arginine. The stereocenter is highlighted with a blue asterisk.

6.3.3 Free energy calculations for folding-unfolding studies

The equilibrated system from the unbiased simulation of the final production trajectory was used for free energy calculations associated with folding-unfolding equilibria. Two different enhanced sampling simulation protocols have been utilized for this purpose. They are;

a) well-tempered metadynamics

b) Controlled sampling approach

The details of these approaches are provided below.

a) Well-tempered metadynamics: For optimizing the collective variables (CVs) and understanding the accessible states by RNA tetraloop, well-tempered metadynamics[42] simulations are carried out initially in the absence of amino acids. The CVs utilized for biasing the simulations are hydrogen bonds, end-to-end distance, *ermsd*, and radius of

gyration. We have used various combinations of CVs initially to understand the dynamics. Note that a maximum of two CVs are used for a given simulation.

b) Controlled Sampling Approach: As we have encountered problems with the convergence and existence of false minima states (discussed later), we have used a controlled sampling approach by combining umbrella sampling[43] and parallel bias metadynamics[44, 45] for the construction of multidimensional free energy landscape associated with folding-unfolding equilibria of RNA motif in the presence of arginine. In this approach, two sets of CV sets are defined:

i) Umbrella sampling coordinate

ii) Subsidiary CVs.

Harmonic umbrella sampling potential was applied along the umbrella sampling coordinate. In the current study, stem distance was used as an umbrella sampling coordinate. By biasing along this coordinate, control & directionality of the pathway corresponding to folding-unfolding events are achieved. In this specific case, 20 umbrella windows are used along stem distance with a harmonic potential of 200 kJ/mol/nm^2. The umbrella sampling windows are placed from a stem distance range from 0.8 nm to 4.6 nm, extending from folded state to unfolded state, respectively, with an interval of 0.2 nm between each window.

Further, subsidiary variables are subjected to experience the potential from parallel bias metadynamics across each individual umbrella sampling window. For parallel bias metadynamics, an initial Gaussian hill height of 0.5 kJ/mol with a bias factor of 10 were applied. Gaussian potential was applied to the system with a pace of 1 ps. In addition, the Gaussian widths for *ermsd*, hydrogen bonds, contacts, and radius of gyration are set to be 0.1, 0.4, 1, and 0.05, respectively. For the RNA motif containing D-arginine, simulations were carried out for 950 ns (20 windows x 47.5 ns per window). While for the system containing L-arginine, simulations were carried out for 900 ns (20 windows x 45.0 ns). The MULE program[46] was employed to compute the minimum free energy path, connecting the folded state to the unfolded state through the misfolded state. This calculation involved treating the folded and extended unfolded states as the starting and ending points, respectively.

6.3.4 Single-molecule binding free energy calculations

For single molecule binding free energy calculations, well-tempered metadynamics[42] simulations were carried out using binding distance as a CV. The arginine molecule closest to

the terminal residues of the RNA motif is used as a starting point for the single molecule binding free energy calculations (discussed later). The well-tempered metadynamics parameters were set as follows: an initial Gaussian height of 0.2 kJ/mol, gaussian width (sigma) of 0.1 nm, pace of 2 ps, and bias factor of 10. 2D- free energy surface was constructed by projecting the free energy surface with respect to binding distance and *angvec* CV.

6.3.5 Choice of collective variables

To understand the different states of the RNA motif, CVs were chosen based on already reported studies[13, 47] on similar systems and based on an intuitive understanding of the RNA folding dynamics. The selected CVs chosen for understanding the folding-unfolding events were able to predict the structural changes in the folded states, including the geometric changes in the nuclear bases as well as the changes in the misfolded and extended states. For single molecule binding calculations, CVs were chosen to distinguish the bound and unbound state of arginine enantiomers towards the RNA motif. The details of the CVs are provided below.

a) Radius of gyration (Rg): In this system, the radius of gyration is defined to distinguish the overall shape of RNA motif in folded, misfolded, and unfolded states. The heavy atoms of the RNA motif were considered to calculate the radius of gyration. The radius of gyration is defined by;

$$R_g = \left(\frac{\sum_i^n m_i |r_i - r_{COM}|^2}{\sum_i^n m_i} \right)^{1/2},$$

where r_{COM} is defined by $r_{COM} = \frac{\sum_i^n r_i m_i}{\sum_i^n m_i}$, which defines the center of mass of heavy atoms of the RNA motif.

b) Stem distance: Stem distance is defined as the distance between the center of masses of two stem regions in the RNA motif, which connects the loop region of the structure. The nuclear bases and backbone regions of residues 2, 3, and 4 define the first stem region. Similarly, the nuclear bases and backbone regions of residue 9, 10, and 11 define the second stem region. The stem distance can distinguish the relative motion of defined stem regions during the formation of misfolded and unfolded states during the dynamics, increasing the magnitude of this parameter from 0.8 nm to 4.6 nm from folded state to unfolded state.

c. Hydrogen bonds (HB): The hydrogen bond parameter is defined to understand the local structural changes that happened to the RNA based on the Watson-Crick base pairing scheme between purine and pyrimidine bases in the stem region. The hydrogen bond parameter is defined by;

$$HB = \sum_i \frac{1 - \left({}^{r_i}/_{r_0}\right)^n}{1 - \left({}^{r_i}/_{r_0}\right)^m}$$

where, *i* defines an index value such that a total of 13 hydrogen bonds are formed between the atoms of the RNA motif. The switching function described above ensures that the calculated hydrogen bond values are continuous derivatives. Among them, 12 hydrogen bonds were formed by the purine and pyrimidine base in the stem region. The remaining hydrogen bond was formed in the loop region of the RNA motif, such that the existence of a hydrogen bond is defined between the phosphate and amide group between the bases G (residue position 5) and A (residue position 8), respectively. Note that this hydrogen bond defined in the loop region was not considered for the free energy calculations in the absence of amino acids.

d. End-to-end distance: End-to-end distance is defined by the two terminal atoms connecting the nuclear base and backbone of the RNA (with atom label C1) at residues positioned at 1 and 12, respectively. End-to-end distance gives a brief overview of how the end terminal of the RNA motifs is changed during the folding-unfolding events.

e) *ermsd*: *ermsd*[47] is a metric parameter specifically defined to understand the base-to-base three-dimensional contact map of interactions represented by the residues of RNA motifs. The benefit of using this parameter is that details regarding the orientational preferences can also be understood along with the information related to relative positions. Moreover, this parameter was expected to sample various local base pairing schemes hindered by high energy barriers.

f) Contacts: The contacts are defined by a switching function that incorporates two groups of atoms, which can form possible contacts between the stem region of the RNA motif. Contacts can be described as follows;

$$N_c = \sum_{i \in A} \sum_{i \in B} s_{ij}$$

$$s_{ij} = \frac{1 - \left(\frac{r_{ij} - d_0}{r_0}\right)^n}{1 - \left(\frac{r_{ij} - d_0}{r_0}\right)^m}$$

where, s_{ij} =1, if the contacts are established between the atoms present in the stem regions. The heavy atoms present in the 5' and 3' end of the stem regions are defined using indexes *i* and *j*. The values are set to 3.0 Å, 0 Å, 6, and 12 for the parameters r_0, d_0, n, and m, respectively.

g) Binding distance: Binding distance is defined as the center of mass distance between the heavy atoms of stem regions of the RNA motif and heavy atoms of arginine. Binding distance is used for single molecule binding studies of arginine towards the RNA motif.

h) Collective variable *angvec*: *angvec* CV is defined to understand the orientation of arginine towards the RNA motif. It is defined as the angle established between two vectors $\vec{b}$ and $\vec{d}$, which are defined as follows. $\vec{b}$ was defined by the center of mass distance between the sugar and backbone part of RNA from residues 9, 10, 11, and 12 to residues 1, 2, 3, and 4. A vector that connects the center of mass of guanidinium end to Cα is defined by $\vec{d}$.

$$angvec = cos^{-1}[\vec{b}.\vec{d}/(|\vec{b}||\vec{d}|)].$$

This CV has been utilized before for extensive studies related to DNA intercalation. [48-50]. A representation of $\vec{b}$ and $\vec{d}$ are shown in Fig. 2.

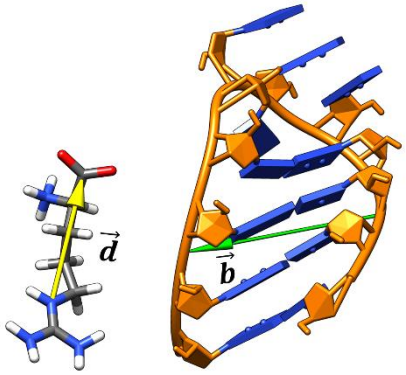

Figure 2. Representation of $\vec{b}$ and $\vec{d}$. The directionality of vectors is shown by green- and yellow-colored arrows for $\vec{b}$ and $\vec{d}$ respectively. The *angvec* CV is defined as the angle between $\vec{b}$ and $\vec{d}$

6.4 Results & Discussion

6.4.1 Folding-unfolding studies in the absence of amino acids

a) Rg and hydrogen bond as CVs

Initially, for free energy calculations based on well-tempered metadynamics, the radius of gyration and hydrogen bonds were considered as CVs to understand the folding-unfolding events of RNA tetraloop. It is to be noted that in this definition of hydrogen bonds, the hydrogen bonds in the loop region of the RNA motif are not considered. In the native folded state, the hydrogen bond is expected to be between 10 and 12, while in the unfolded state, it is expected to be 0. Similarly, Rg is expected to be < 0.9 nm in the folded state and > 1.4 nm in the completely unfolded state. The time evolution of biased CVs and the free energy surface during the 500 ns simulation is shown in Fig. 3. However, we have not observed the existence of the completely unfolded state, indicating the simulation is far from convergence.

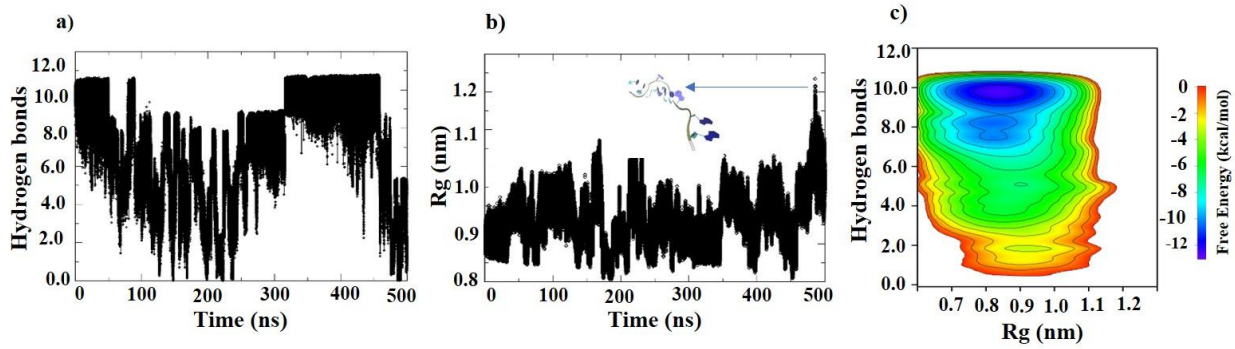

Figure 3. a) Time dependence of hydrogen bonds in the metadynamics simulation using hydrogen bonds and Rg as CVs b) Time dependence of Rg in the metadynamics simulation using hydrogen bonds and Rg as CVs c) Free energy surface with respect to biased variables (Hydrogen bonds and Rg) during 500 ns metadynamics simulation.

Further, to understand specifically the loop motion of the RNA motif, the structures are collected that resemble the equilibrated RNA structure (9.7 <Hydrogen bonds <10 and 0.89 < Rg < 0.93). Based on these CV ranges, four clusters of structures were obtained from the simulation trajectory. The existence of clusters (A, B, C, and D) in various time frames is shown in Fig. 4. To identify the local structural changes in the RNA in the stem region and loop region specifically, the distribution of torsional parameters χ, ζ, ε, and δ (representation of dihedral parameters are shown in Fig. 5) are plotted for clusters A, B, C, and D.

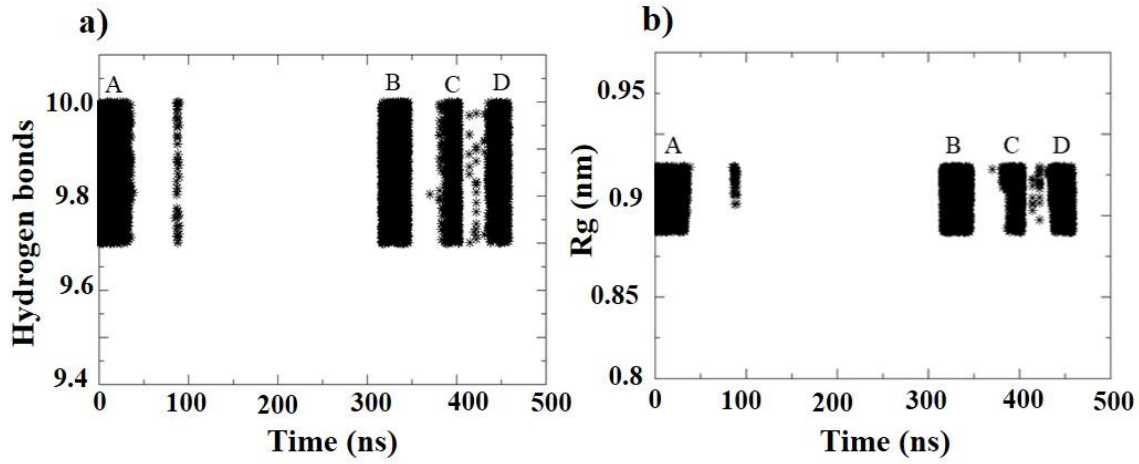

Figure 4. a) Clusters of RNA structures having hydrogen bond values similar to the equilibrated native folded state of RNA during simulation using hydrogen bonds and Rg as C.V.s. b) Clusters of RNA structures having Rg values similar to the equilibrated native folded state of RNA during the simulation using hydrogen bonds and Rg as CVs. The clusters are labeled by B, C, and D, respectively.

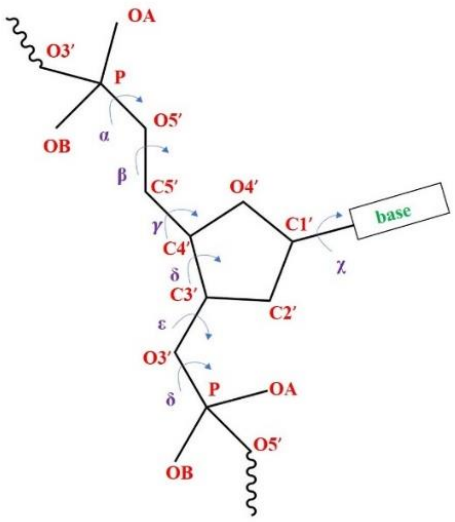

Figure 5. Representation of dihedral parameters in a given nucleotide unit.

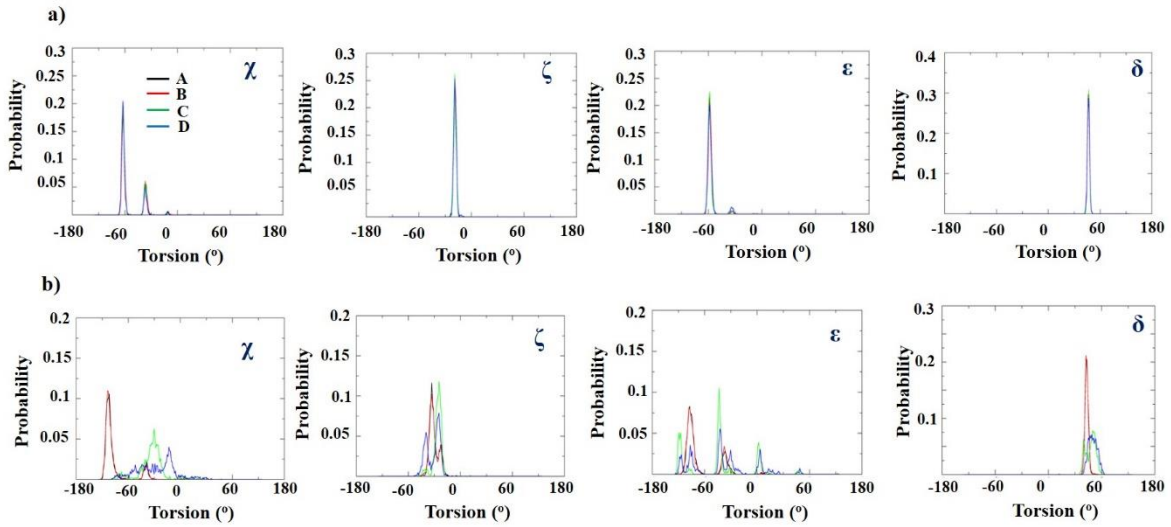

Figure 6. a) Torsional parameters (χ, ζ, ε, and δ) in the clusters A, B, C, and D averaged over the residues in the stem region of RNA. b) Torsional parameters (χ, ζ, ε, and δ) in the clusters A, B, C, and D averaged over the residues in the loop region of RNA.

The distribution of various dihedral parameters averaged over the residues of the stem and loop region in the clusters A, B, C, and D are shown in Fig. 6. It is found that the distribution of dihedral parameters in the stem region is less altered in clusters A, B, C and D. While significant changes are observed in the distribution of the loop region, suggesting modification is required in defining the CVs to understand the loop motions of RNA motifs. Later, we incorporated a given hydrogen bond definition corresponding to phosphate amide interaction in the residues 5 and 8 positions of the loop region along with the Watson-Crick hydrogen bond description defined for the stem part of the RNA. Moreover, we also found that *ermsd* parameter can also distinguish various local structural changes in the RNA. This was reflected in two similar minima states (M1 and M2) in the 2D reweighted free energy profiles with respect to *ermsd*, unlike the broad single minima state obtained in the free energy surface with respect to biased variables (Rg and Hydrogen bonds). The reweighted free energy plots are shown in Fig. 7 with respect to *ermsd* and biased CVs.

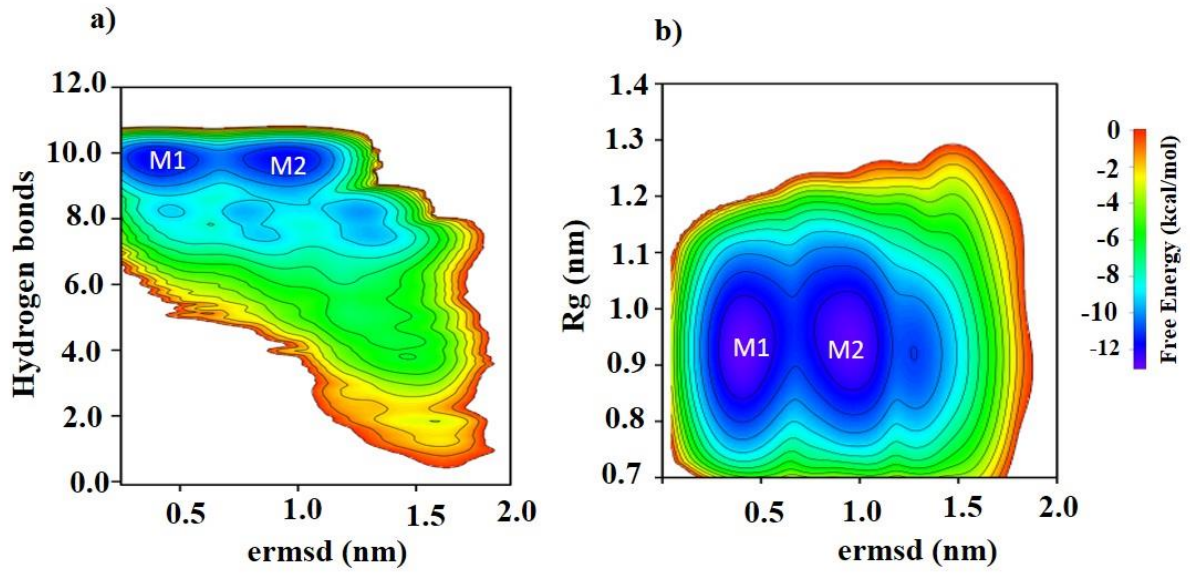

Figure 7. 2D-reweighted free energy surface with respect to hydrogen bonds, Rg, and *ermsd* in the biased simulations using hydrogen bonds and Rg as CVs. M1 and M2 indicate the distinguishable similar minima states, which were previously recognized as the same state with respect to the free energy surface constructed using biased CVs as shown in Fig. 3c.

b) *ermsd* as a CV

ermsd is used as a CV for 1D well-tempered metadynamics simulations to understand the folding-unfolding events of RNA in the CV space. However, it was found that the parameter was unable to push the system to explore the complete unfolded state. Here, the system preferred to stay similar to the native folded state. The evolution of *ermsd*, hydrogen bonds (HB), and Rg in the trajectory is shown in Fig. 8.

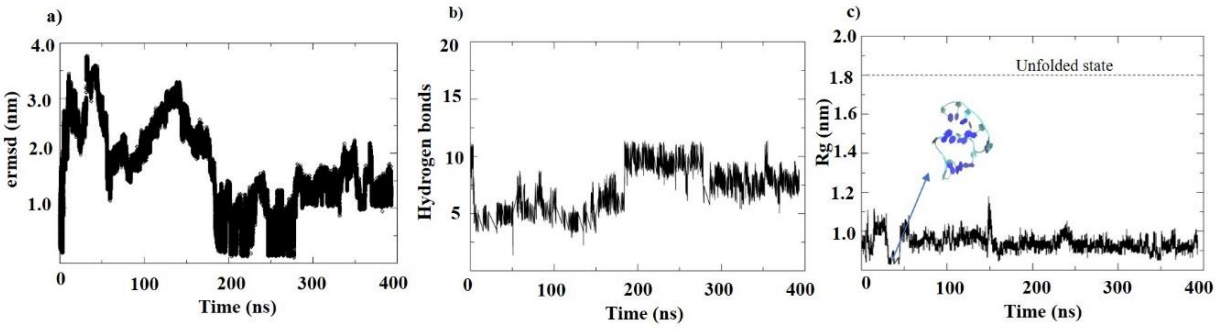

Figure 8. a, b, and c) Time evolution of *ermsd*, Hydrogen bonds, and Rg in the biased simulation using *ermsd* as the CV.

c) End-to-end distance as a CV

As the previous simulations found difficulties in accessing the completely unfolded state, end-to-end distance is used as a CV. By using this CV, the system visited the extended unfolded state. However, once the unfolded states were visited, the system failed to revisit the native folded state. The time evolution of End-to-end distance and hydrogen bonds during the simulation is shown in Fig. 9. Moreover, the formation of complex structures (as shown in Fig 9b at time 50 ns, 80 ns, 175 ns, etc.) was shown by the RNA tetraloop, which prevents visiting the folded state again.

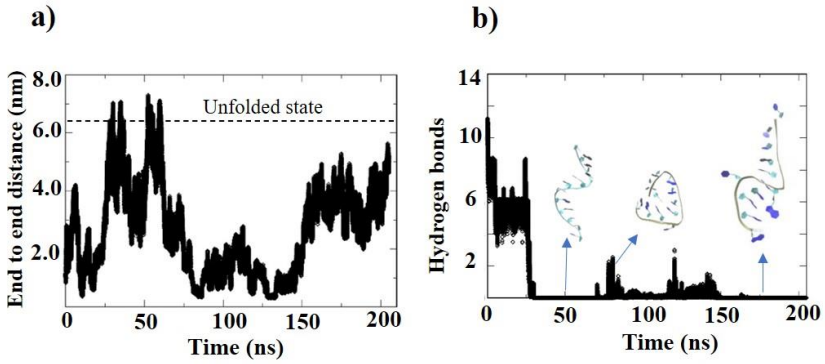

Figure 9. Time evolution of end-to-enddistance and hydrogen bonds in the metadynamics simulations using end to end distance as CV. Some of the complex structures formed by the RNA motif after visiting the unfolded states (after 25 ns) is shown in the inset.

6.4.2 Folding-unfolding studies in the presence of enantiomers of arginine

We have encountered problems related to lack of convergence, the existence of trapped minima states, and CV-dependent folding-unfolding events in the well-tempered metadynamics simulations for the system in the absence of amino acids. To overcome these problems encountered in well-tempered metadynamics simulations, a controlled sampling approach was used to understand the free energy landscape associated with RNA tetraloop folding-unfolding equilibria in the presence of arginine. Five CVs are used for this approach. They are stem distance, *ermsd*, contacts, hydrogen bonds, and radius of gyration. The convergence of simulations was observed at 900 ns and 950 ns for the RNA systems containing L-arginine and D-arginine, respectively. The 1D free energy landscape with respect to all five CVs used in the controlled sampling approach is shown in Fig. 10.

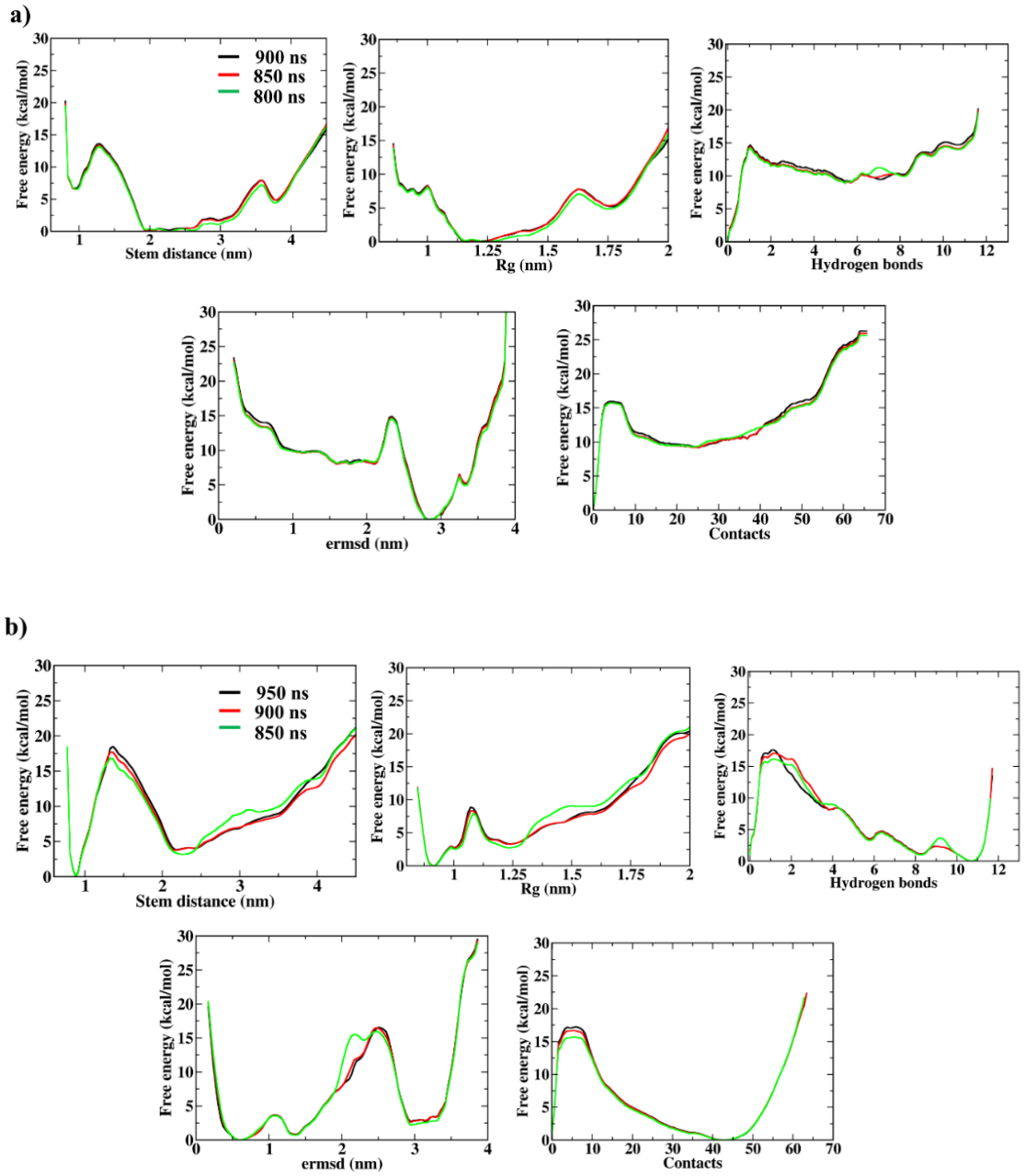

Figure 10. 1D-Free energy profiles at different times instances with respect to all biased collectives (stem distance, Rg, hydrogen bonds, ermsd, and contacts) associated with RNA tetraloop folding-unfolding equilibria a) in the presence of L-arginine, and b) in the presence of D-arginine.

2D-free energy surfaces were constructed with respect to *ermsd* and stem distance to interpret the folding-unfolding equilibrium in more detail. While stem distance measures the extent of unfolding, *ermsd* provides insights about the local base pairing schemes and stacking interactions with respect to reference equilibrated folded structure. The 2D free energy surfaces are shown in the Fig. 11.

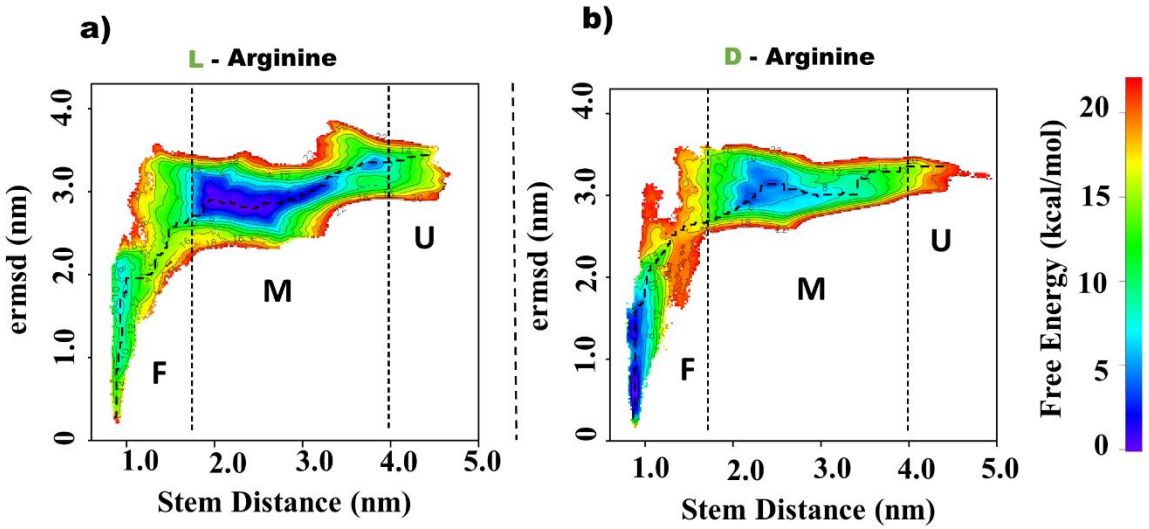

Figure 11. 2D Free energy landscape corresponding to RNA folding-unfolding equilibria with respect to stem distance and *ermsd*. (a) in the presence of L-arginine and (b) in the presence of D-arginine. The minimum free energy path is indicated by black dotted lines connecting the folded state to the unfolded state.

Three broad regions of states are classified based on the ranges of CVs visited by the stem distance and *ermsd*. They are folded state (stem distance < 1.8 nm and *ermsd* < 2.0 nm), extended unfolded states (stem distance > 4.0 nm and ermsd > 2.0 nm), and the misfolded states (1.8 nm ≥ stem distance≤ 4.0 nm and *ermsd* > 2.0 nm) which are denoted by F, U, and M, respectively. In the last 150 ns of the simulation, there were no significant changes in the free energy stability corresponding to the F, U, and M states. The free energy plots with the L^2 error in the last 150 ns simulation are shown in Fig. 12.

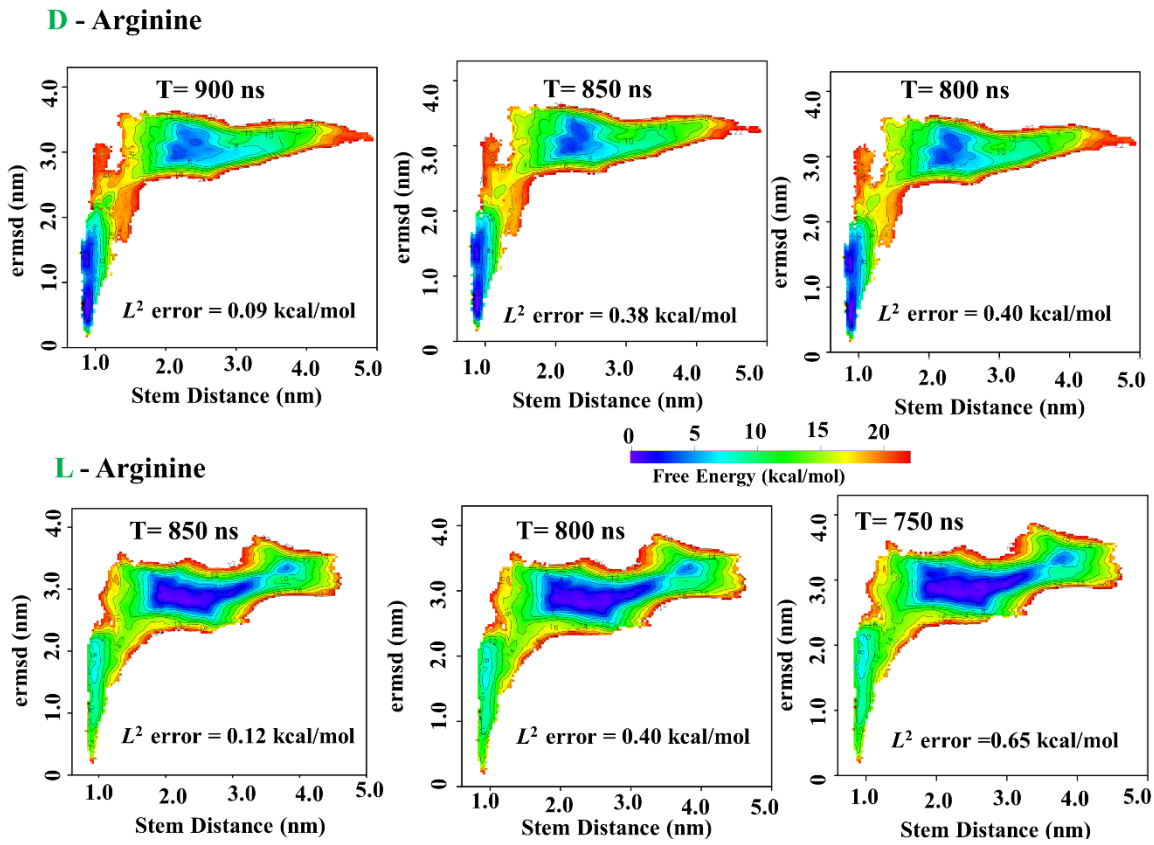

Figure 12. 2D Free energy profiles projected with respect to stem distance and *ermsd* corresponding to folding-unfolding equilibria of RNA motif in the presence of D-arginine and L-arginine. The L^2 error with respect to the final free energy surfaces (950 ns for D-arginine and after 900 ns for L-arginine) is shown in the inset of the plots.

For more details regarding the stability comparisons of RNA in the chiral environment, the 1D free energy profile with respect to stem distance was shown (Fig. 13) relative to the unfolded state. It is evident from Fig. 13 that the folded state of RNA is stable by -22 kcal/mol and -10.9 kcal/mol, respectively, for systems containing D-arginine and L-arginine relative to the unfolded state. However, the free energy stability values for the misfolded state were found to be between -17.0 kcal/mol and 18.0 kcal/mol for both the enantiomers. In the system containing D-arginine, the native state forms the global minima. While in the system containing L-arginine, the misfolded state forms the global minima with a broader region. The results obtained from these plots indicate the chiral-specific interaction of arginine towards RNA, supporting the theories on the origin of homochirality.

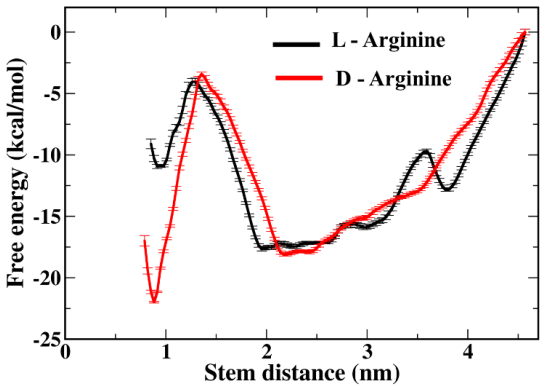

Figure 13. Comparison of 1D -free energy surface projected only on stem distance with respect to unfolded state for both L- and D-arginines.

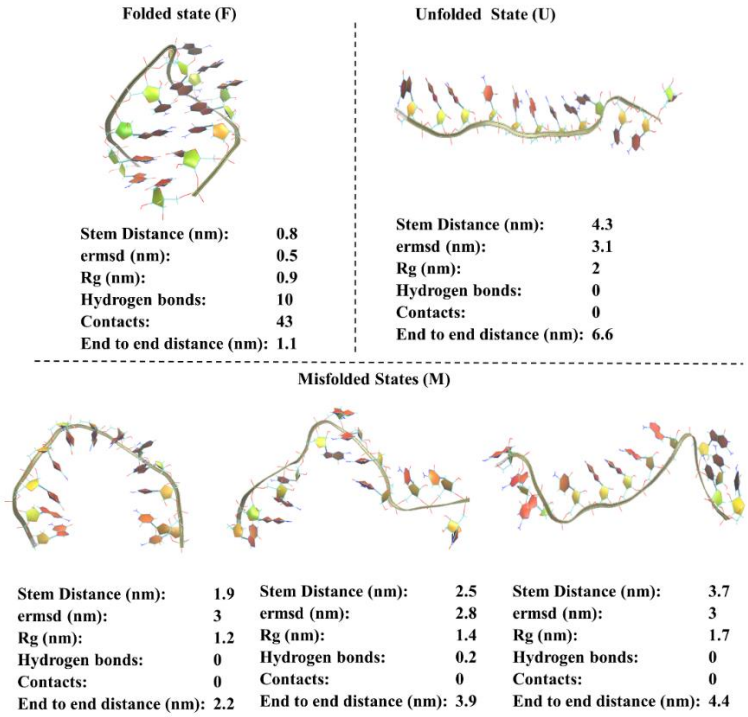

Figure 14. Representative structures of RNA for folded (F), unfolded (U), and misfolded (M) states.

The representation of various structures (folded, unfolded, and misfolded structures) formed by RNA motifs are shown in Fig. 14. While moving from a folded state to a completely extended unfolded state, an increase in stem distance, Rg, and end-to-end distance was observed. However, the misfolded and unfolded states were found to have zero contacts and hydrogen bonds with high *ermsd* values. The 2D reweighted free energy profiles with respect to stem distance and other subsidiary CVs are shown in Fig. 15.

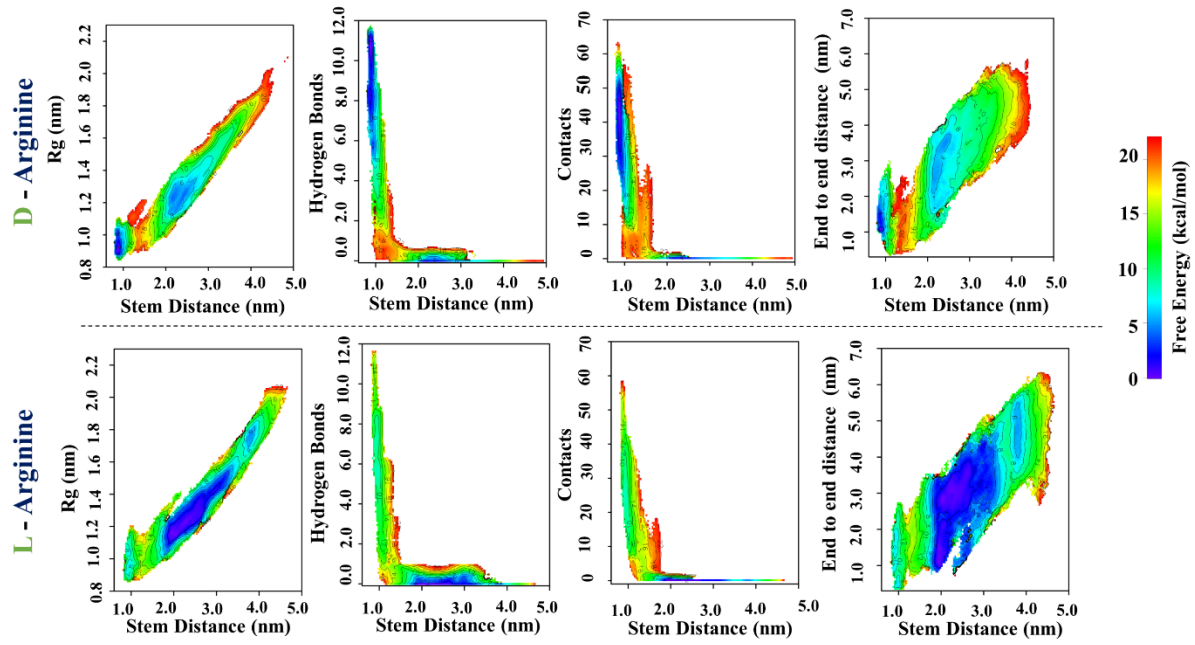

Figure 15. 2D Free energy landscape corresponds to folding-unfolding equilibria in the presence of D-arginine and L-arginine projected with respect to stem distance and other subsidiary CVs.

6.4.3 Influence of enantiomers in the folded state

As D-arginine was found to stabilize the folded state of the RNA motif more effectively than L-arginine, it was initially assumed that interaction energy between the RNA and the enantiomers of arginine would be different. Therefore, the structures corresponding to the folded state (stem distance < 1.8 nm and *ermsd* < 2.0 nm) were collected for further analysis. It is found that the average number of hydrogen bonds and interaction energy between the RNA and the enantiomers of arginine falls into a similar range (Fig. 16). These results rule out the assumption of additional factors that stabilize the folded state of RNA by D-arginine.

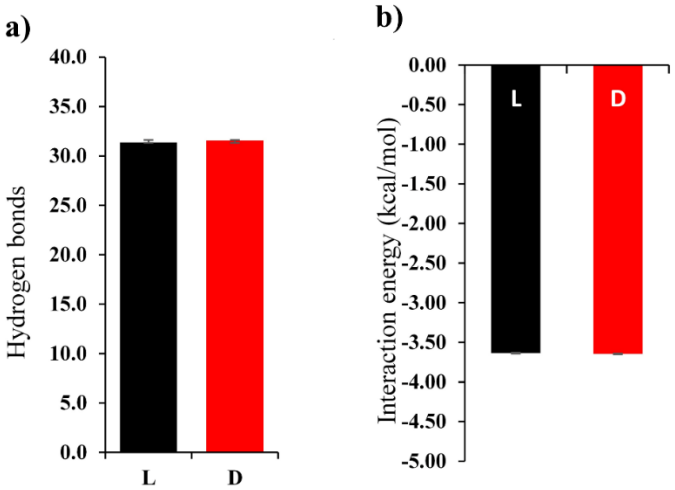

Figure 16. a) Average number of hydrogen bonds between RNA motif and enantiomers of arginine in the folded state of RNA b) Average interaction energy between the RNA motif and the enantiomers of arginine in the folded state of RNA.

The second hypothesis was that the local structural changes that happened in the folded state would be different in the presence of D-arginine and L-arginine. This is attributed to the fact that the bases in the RNA can exist in different geometric forms based on diverse complementary/neighboring base pairing schemes supported by the orientational effects in the backbone of RNA structures. The existence of these states can be described based on the edges of interactions defining the base pairing orientations. The probable edges responsible for these interactions are Watson-Crick, Hoogsteen/C-H edge, and sugar edge. Information regarding these edges of interactions, along with the type of glycosidic rotation (cis/trans orientation) around the sugar backbone, would be able to distinguish the local structural changes in the RNA in different chiral environments. The parameters based on the base pairing schemes and orientational effects are defined based on the way explained by Leontis geometric classification of RNA structures[51]. An illustration of base pairing schemes is shown in the Fig. 4 of Chapter 1.

For a better way of representing the local structures with respect to residues of RNA, the following three-letter code was used. The first letter mentions the donating edge for pairing with the neighboring/complementary base from the 5' end of the RNA motif. The second letter indicates the acceptor edge for the base pairing scheme. Finally, the third letter indicates the orientation of the glycosidic bond facilitating the interactions between the donating edge and acceptor edge. For example, the SHc state demonstrates that the sugar edge forms the base pairing with the Hoogsteen edge of neighboring/complementary residue in *cis* orientation. Similarly, WHc state describes that Watson-Crick is the donating edge and Hoogsteen edge is the acceptor edge such that the bases are interacting in *cis* orientation. Note that the directionality of interaction is from 5' end of the stem region of RNA to 3' end of the stem region of RNA. A representation of SHc state and WCc state is shown in Fig. 17 a and b, respectively. A triangle representation of the possible geometries based on a three-letter code defined as per Leontis representation is provided in Fig. 18.

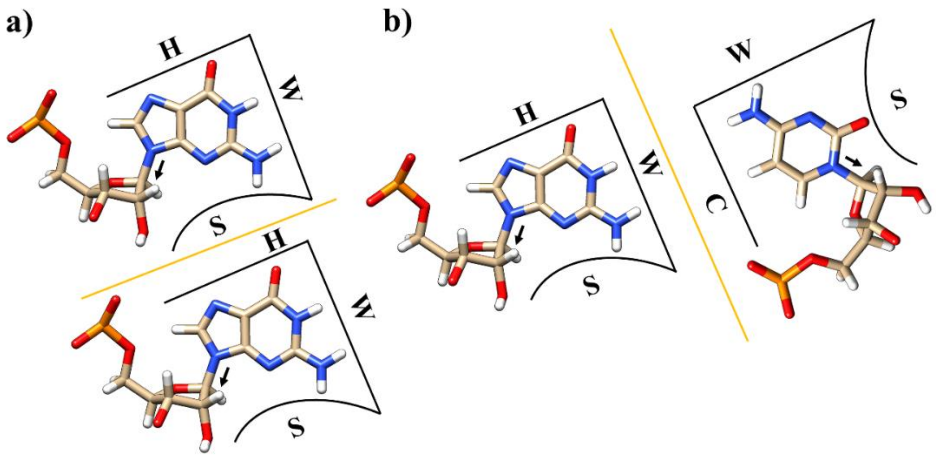

Figure 17. Illustration of base pairing orientations based on three-letter code according to Leontis geometric classification of RNA structures. Black arrows show the directionality of the glycosidic bond. The boundary of interaction between the donating edge and the accepting edge is shown by the solid yellow line. The single letter codes W, H, C, and S indicate Watson-Crick, Hoogsteen, C-H, and sugar edges for the various base pairing schemes. a) SHc state b) WCc state

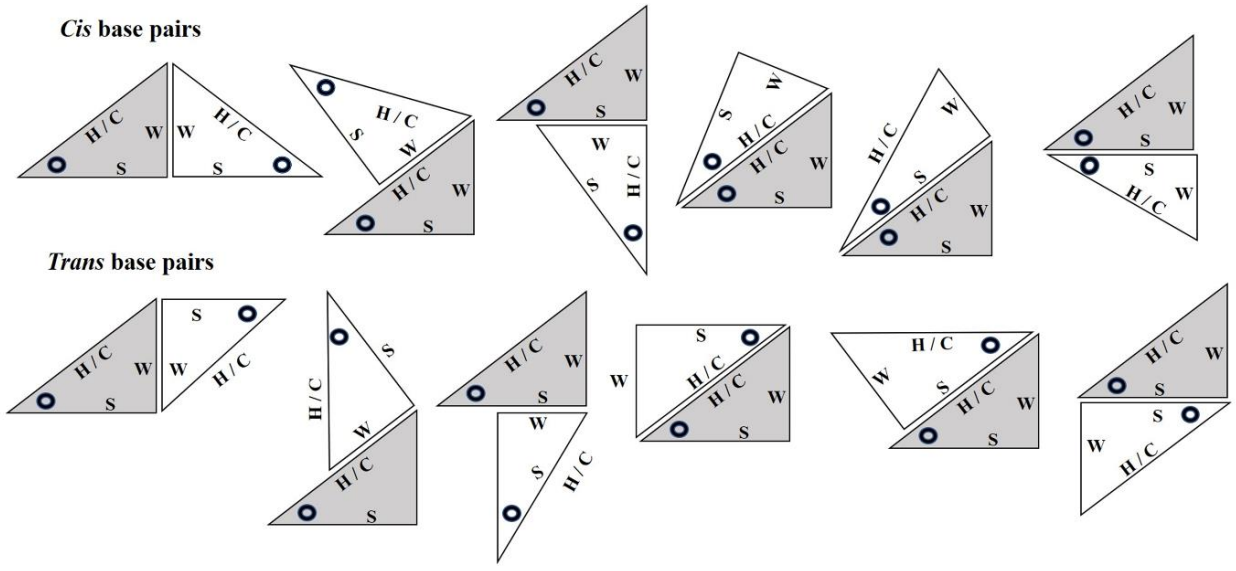

Figure 18. Representation of base pairing schemes in RNA along the edges of interactions with respect to cis/trans geometry.The edges of interactions are identified based on the surface of separation between white and grey triangle which denote the donor and acceptor edges. The black circle indicate the position of the ribose region of the residue in the RNA. The description of three letter notations are also provided in the Table 1 of Chapter 1.

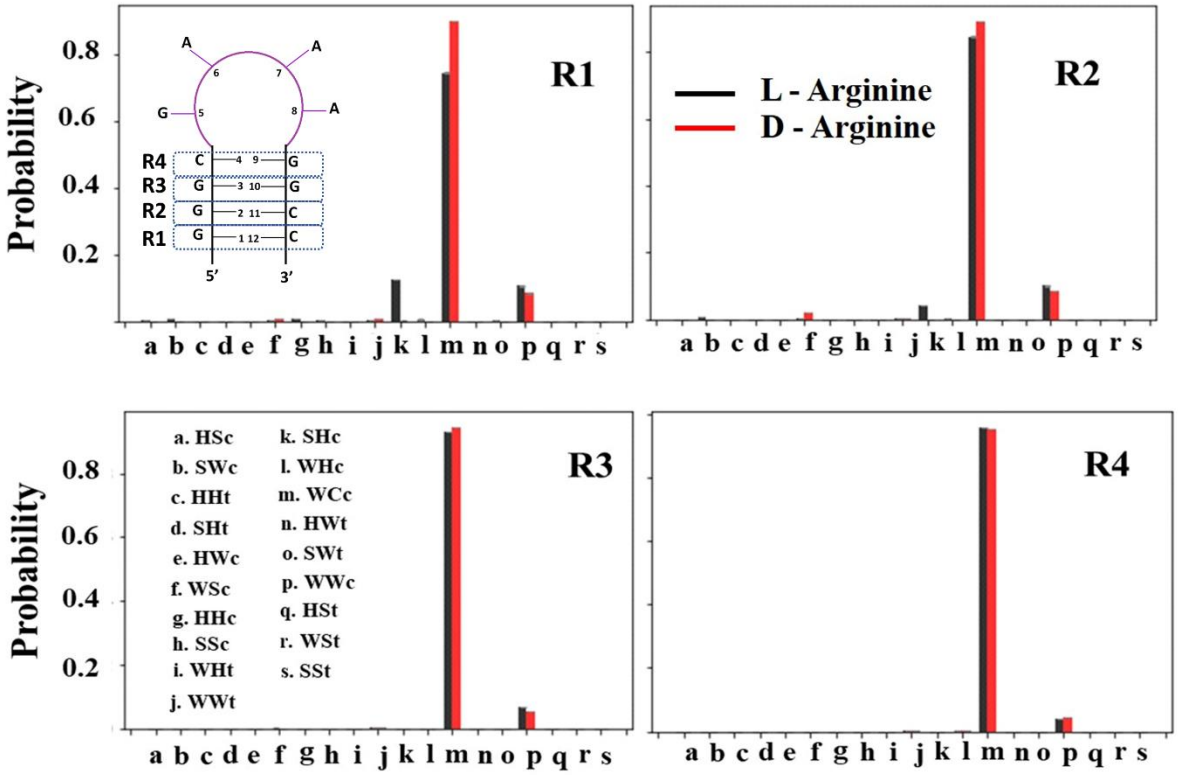

Figure 19. Distribution of base pairing schemes in the native folded state of RNA corresponding to stem region. R1, R2, R3, and R4 indicate the residual junctions in the stem region of RNA. The residue junctions in the RNA tetraloop are shown in the inset of the figure.

As the impact of chirality is predominantly visible in the folded state, the distribution of the local states of RNA according to the Leontis representation is calculated for the residue junctions in the stem region (Fig. 19). The residue junctions are named from R1 to R4, starting from the terminal junction (residues 1 and 12) towards the junction just after the loop (residues 5 and 9). In the majority of the junctions, WCc state is found to exist in higher probability, followed by WWc state. However, the WCc state distribution is altered more in the R1 junction in the presence of D-arginine. Moreover, the SHc state is present only in the R1 junction in the presence of L-arginine. The significant changes in the probability distribution in the R1 junction indicate that the chiral-specific effects are majorly affected only in the terminal base pairs of the RNA motif.

6.4.4 Binding modes of arginine enantiomers on the terminal residues of RNA

The unbinding free energy calculations were carried out to understand the effect of arginine enantiomers on the terminal region of the RNA. From the structures corresponding to the folded state of RNA, the most probable confirmation of arginine (both D-arginine and L-arginine) closest to the terminal residues of RNA was taken for unbinding studies using well-tempered metadynamics simulations (see methods). Binding distance distinguishes the bound state and unbound state, while the *angvec* CV reveals information about the possible position of arginine around the terminal region of RNA. The 2D free energy profile with respect to binding distance and *angvec* is shown in Fig. 20.

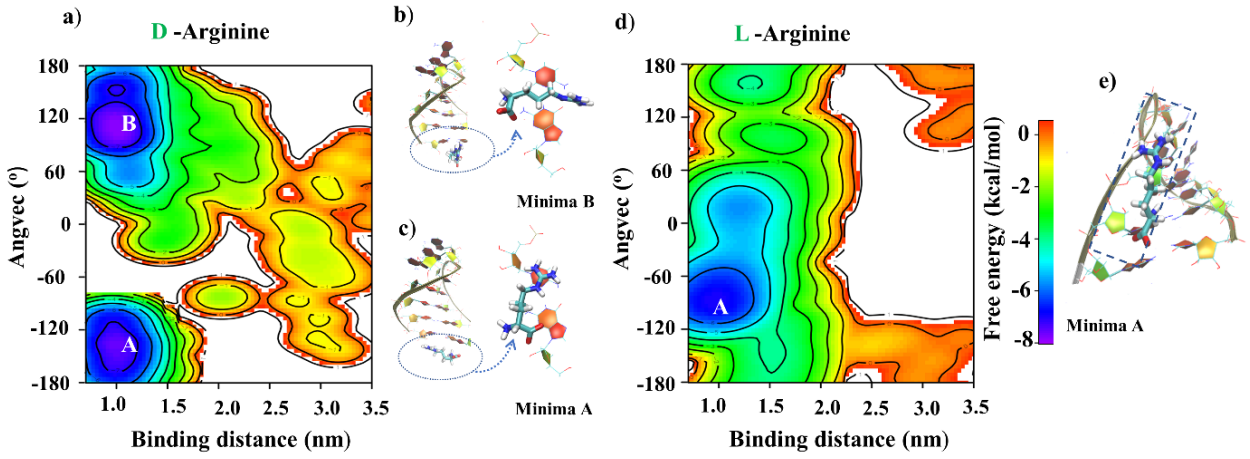

Figure 20. a) 2D free energy plot corresponding to unbinding of a single D-arginine and b &c) Minima structures for the D-arginine bound to RNA. d) 2D free energy plot corresponding to unbinding of a single L-arginine e)) Minima structure for the L-arginine bound to RNA.

In the case of D-arginine, two minima were found (minima A & B) with similar free energy values around -7.8 kcal/mol. These minima correspond to two different stacking orientations at the bottom of the RNA motif, as shown in Fig. 20b & c, respectively. However, in the case of L-arginine, only a single minima was observed with the free energy stability -6.8 kcal/mol, which corresponds to the groove-bound state (Fig. 20e). To understand these specific effects of arginine enantiomer, the architecture of arginine structure is divided into (i) Cα (ii) alkyl (iii) guanidinium regions as shown in Fig. 21a. In the case of the minima structure of the system containing L-arginine, Cα part is found to be close to the sugar-backbone part of the RNA motif. This structural existence can be due to the favorable energetics causing nucleophilic interaction between the hydroxyl group in the sugar region of the RNA and Cα

part of the arginine. A similar existence of favorable interactions was reported in the trans-acylation reaction between the acylated amino acids and specific RNA motifs.[52] A representation of this favorable approach is shown in Fig. 21b. However, the D-arginine cannot form this stabilizing interaction as the steric hindrance caused by the alkyl chain of arginine blocks the attack for nucleophilic attack initiated by the hydroxyl group of RNA (Fig. 21c).

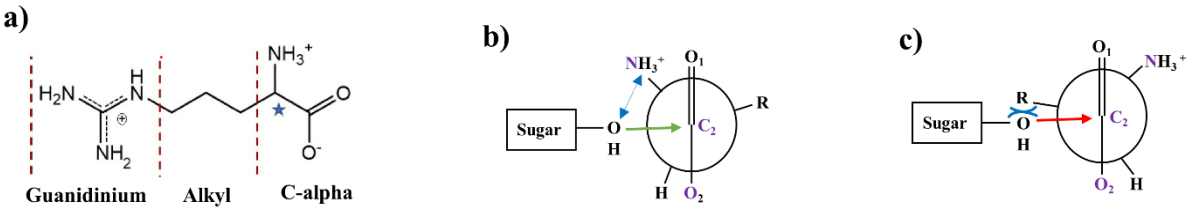

Figure 21. (a) Representation of arginine with interacting regions. The stereocenter is shown by a star label. Schematic representation for (b) Representation of stabilizing nucleophilic interaction between RNA and carbonyl part of L-arginine and (c)) Representation of non-stabilizing nucleophilic interaction between RNA and carbonyl part of D-arginine.

To understand further the suitable binding mode, a dihedral angle is defined θ (O2, C2, C1, N), which specifies the orientational effects caused at the stereocenter of the L-arginine and D-arginine in the presence of RNA motif. θ distribution is found to be different for the systems containing D-arginine and L-arginine as shown in Fig. 22. This indicates that the stereocenters at the respective amino acids play a key role in recognizing the RNA motifs.

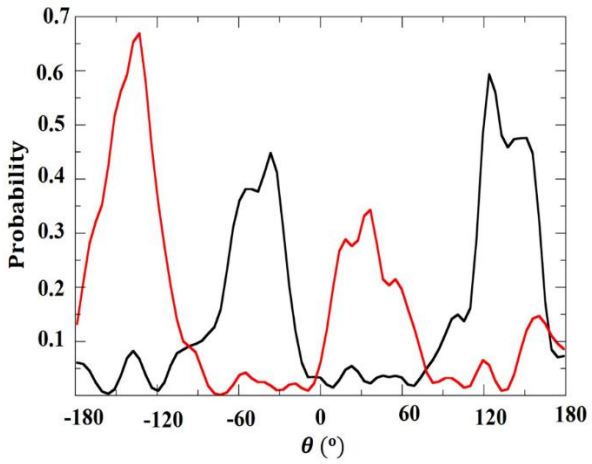

Figure. 22 Dihedral angle (θ (O2, C2, C1, N)) distribution characterizing the binding mode for L-arginine and D-arginine.

Further, to comment on the fact that only D-arginine is found on the terminal regions of the base pairs, the interaction energy is calculated between the side constituents of arginine enantiomers and terminal nuclear bases. Note that the geometry around terminal residues of RNA for the L-arginine is obtained by inverting the stereochemistry for the D-arginine at the stacked minima state (minima A in the system containing D-arginine, Fig. 20c). It was found that for the D-arginine, the carboxylate region in the c-alpha part faces towards the amine part of residue 1 of the RNA, stabilizing the arginine along with stacking interactions at residue 12. However, in the system containing L-arginine, the carboxylate group is found to be opposite to the amine part of residue 1, lacking stabilization at the residue 1 region. An illustration of these interactions at the terminal base pairs is shown in the Fig. 23. These stabilizing interactions are also reflected in the difference in energetics associated with the carboxyl-amino interactions, as shown in Table 1, Even though the magnitude of stacking interactions is found to be similar.

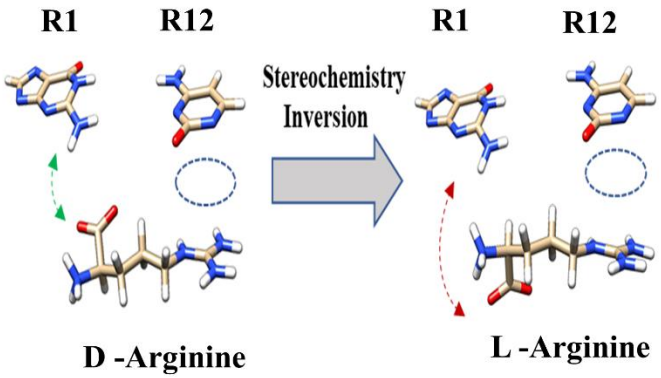

Figure 23. Illustration of interaction between terminal bases (residue 1 and residue 12) of RNA motif and arginine enantiomers. R1 and R2 represent the RNA motif's terminal residues 1 and 12, respectively. Dotted blue circles show stacking interactions initiated by the Guanidinium group of arginine and residue 12. The green arrow shows the preferred geometry and stabilizing interactions between the carboxylate group in D-enantiomer and residue 1. The non-referrable mode of interaction between the carboxylate group in L-enantiomer and residue 1 is indicated by the red arrow.

Table 1. Interaction energy profiles representing stacking and carboxyl-amino interactions between the terminal residues of the RNA and stereoisomers of arginine.

Sl No	System	Interaction energy (kJ/mol)	
		Stacking	**Carboxyl - Amino interaction**
1	D-arginine	-12.0 ± 0.046	-5.2 ± 0.016
2	L-arginine	-12.2 ± 0.037	-0.6 ± 0.016

It is assumed that the specific binding modes between RNA motif and D-arginine at the terminal region of RNA contribute to extra stabilization of the folded state, as observed in the free energy landscape associated with folding-unfolding results. This stable interaction at the terminal junction might have caused the RNA motif to prevent the formation of unfolded states.

6.5 Conclusion

The study, focused on understanding the folding-unfolding events of 12-mer GAAA RNA tetraloop in the presence of enantiomers of arginine using a controlled sampling approach. A combination of umbrella sampling and parallel bias metadynamics is used to construct free energy surfaces efficiently, describing the folding-unfolding equilibria. Five CVs (one umbrella sampling coordinate and four subsidiary CVs) were used for efficient sampling associated with the process.

The results based on converged free energy surfaces indicate that D-arginine stabilizes the folded states of the RNA motif. At the same time, the L-arginine is found to destabilize the folded state of the RNA motif. Interestingly, it was also found that the free energy stability of misfolded states is found to be the same in both chiral environments. However, it has been found that the local structural changes at terminal base pair junctions of RNA caused extra stabilization to the native folded state in the presence of D-arginine.

Upon further investigation, D-arginine stabilization is attributed to two factors: stacking interactions facilitated by the guanidinium group and electrostatic stabilization by the C-alpha part of D-arginine. Conversely, in the case of L-arginine, the absence of favorable electrostatic stabilization is due to the inversion of the chiral center. Consequently, stabilization in L-arginine primarily occurs through the favorable interaction of the sugar backbone with the arginine side groups during groove binding.

As the D-arginine stabilizes the tetraloop in the terminal junction, it prevents the unfolding of the RNA tetraloop. Hence, our findings elucidate the molecular basis of chiral structures and their impact on the thermodynamics of RNA hairpin structures, exemplified by the GAAA motif. These results suggest that analogous specific interactions between RNA and chiral amino acids support the RNA-based hypothesis on the origin of homochirality.

6.6 References

1. Bonner, W.A., *Chirality and life.* Orig Life Evol Biosph, 1995. **25**(1-3): p. 175-90.
2. Mason, S.F., *Origins of biomolecular handedness.* Nature, 1984. **311**(5981): p. 19-23.
3. Morris, K.V. and J.S. Mattick, *The rise of regulatory RNA.* Nat Rev Genet, 2014. **15**(6): p. 423-37.
4. Joyce, G.F., *The antiquity of RNA-based evolution.* Nature, 2002. **418**(6894): p. 214-21.
5. Herschlag, D., *RNA chaperones and the RNA folding problem.* J Biol Chem, 1995. **270**(36): p. 20871-4.
6. Woese, C.R., S. Winker, and R.R. Gutell, *Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops".* Proceedings of the National Academy of Sciences, 1990. **87**(21): p. 8467-8471.
7. Jucker, F.M. and A. Pardi, *Solution structure of the CUUG hairpin loop: a novel RNA tetraloop motif.* Biochemistry, 1995. **34**(44): p. 14416-27.
8. Jaeger, L., F. Michel, and E. Westhof, *Involvement of a GNRA tetraloop in long-range RNA tertiary interactions.* Journal of Molecular Biology, 1994. **236**(5): p. 1271-1276.
9. Molinaro, M. and I. Tinoco, Jr., *Use of ultra stable UNCG tetraloop hairpins to fold RNA structures: thermodynamic and spectroscopic applications.* Nucleic Acids Res, 1995. **23**(15): p. 3056-63.
10. Baumruk, V., et al., *Comparison between CUUG and UUCG tetraloops: thermodynamic stability and structural features analyzed by UV absorption and vibrational spectroscopy.* Nucleic Acids Res, 2001. **29**(19): p. 4089-96.
11. Chen, A.A. and A.E. Garcia, *High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations.* Proc Natl Acad Sci U S A, 2013. **110**(42): p. 16820-5.

12. Ma, H., et al., *Exploring the energy landscape of a small RNA hairpin.* J Am Chem Soc, 2006. **128**(5): p. 1523-30.
13. Kuhrova, P., et al., *Computer Folding of RNA Tetraloops? Are We There Yet?* J Chem Theory Comput, 2013. **9**(4): p. 2115-25.
14. Smith, L.G., et al., *Chemically Accurate Relative Folding Stability of RNA Hairpins from Molecular Simulations.* J Chem Theory Comput, 2018. **14**(12): p. 6598-6612.
15. Sengupta, A., H.L. Sung, and D.J. Nesbitt, *Amino Acid Specific Effects on RNA Tertiary Interactions: Single-Molecule Kinetic and Thermodynamic Studies.* J Phys Chem B, 2016. **120**(41): p. 10615-10627.
16. Nicholson, D.A., et al., *Amino Acid Stabilization of Nucleic Acid Secondary Structure: Kinetic Insights from Single-Molecule Studies.* J Phys Chem B, 2018. **122**(43): p. 9869-9876.
17. Zerze, G.H., P.M. Piaggi, and P.G. Debenedetti, *A Computational Study of RNA Tetraloop Thermodynamics, Including Misfolded States.* J Phys Chem B, 2021. **125**(50): p. 13685-13695.
18. Bottaro, S. and K. Lindorff-Larsen, *Mapping the Universe of RNA Tetraloop Folds.* Biophys J, 2017. **113**(2): p. 257-267.
19. Banas, P., et al., *Performance of Molecular Mechanics Force Fields for RNA Simulations: Stability of UUCG and GNRA Hairpins.* J Chem Theory Comput, 2010. **6**(12): p. 3836-3849.
20. Mohan, S., et al., *RNA tetraloop folding reveals tension between backbone restraints and molecular interactions.* J Am Chem Soc, 2010. **132**(36): p. 12679-89.
21. Proctor, D.J., et al., *Folding thermodynamics and kinetics of YNMG RNA hairpins: specific incorporation of 8-bromoguanosine leads to stabilization by enhancement of the folding rate.* Biochemistry, 2004. **43**(44): p. 14004-14.
22. Haldar, S., et al., *Insights into Stability and Folding of GNRA and UNCG Tetraloops Revealed by Microsecond Molecular Dynamics and Well-Tempered Metadynamics.* J Chem Theory Comput, 2015. **11**(8): p. 3866-77.
23. Zhang, W. and S.J. Chen, *Exploring the complex folding kinetics of RNA hairpins: II. Effect of sequence, length, and misfolded states.* Biophys J, 2006. **90**(3): p. 778-87.
24. Pathak, A.K. and T. Bandyopadhyay, *Water isotope effect on the thermostability of a polio viral RNA hairpin: A metadynamics study.* J Chem Phys, 2017. **146**(16): p. 165104.

25. Halder, A., et al., *Mg(2+) Sensing by an RNA Fragment: Role of Mg(2+)-Coordinated Water Molecules.* J Chem Theory Comput, 2020. **16**(10): p. 6702-6715.
26. Yarus, M., *A specific amino acid binding site composed of RNA.* Science, 1988. **240**(4860): p. 1751-8.
27. Janas, T., et al., *Simple, recurring RNA binding sites for L-arginine.* RNA, 2010. **16**(4): p. 805-16.
28. Yarus, M., J.J. Widmann, and R. Knight, *RNA-amino acid binding: a stereochemical era for the genetic code.* J Mol Evol, 2009. **69**(5): p. 406-29.
29. Horn, A.H., *A consistent force field parameter set for zwitterionic amino acid residues.* J Mol Model, 2014. **20**(11): p. 2478.
30. Zgarbova, M., et al., *Refinement of the Sugar-Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA.* J Chem Theory Comput, 2015. **11**(12): p. 5723-36.
31. Zgarbova, M., et al., *Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles.* J Chem Theory Comput, 2011. **7**(9): p. 2886-2902.
32. Mark, P. and L. Nilsson, *Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K.* The Journal of Physical Chemistry A, 2001. **105**(43): p. 9954-9960.
33. Deift, P. and X. Zhou, *A Steepest Descent Method for Oscillatory Riemann--Hilbert Problems. Asymptotics for the MKdV Equation.* The Annals of Mathematics, 1993. **137**(2).
34. Berendsen, H.J.C., et al., *Molecular dynamics with coupling to an external bath.* The Journal of Chemical Physics, 1984. **81**(8): p. 3684-3690.
35. Nosé, S., *A molecular dynamics method for simulations in the canonical ensemble.* Molecular Physics, 2006. **52**(2): p. 255-268.
36. Darden, T., D. York, and L. Pedersen, *Particle mesh Ewald: AnN·log(N) method for Ewald sums in large systems.* The Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.
37. Hess, B., et al., *LINCS: A linear constraint solver for molecular simulations.* Journal of Computational Chemistry, 1997. **18**(12): p. 1463-1472.
38. Abraham, M.J., et al., *GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers.* SoftwareX, 2015. **1-2**: p. 19-25.

39. Van Der Spoel, D., et al., *GROMACS: fast, flexible, and free.* J Comput Chem, 2005. **26**(16): p. 1701-18.
40. consortium, P., *Promoting transparency and reproducibility in enhanced molecular simulations.* Nat Methods, 2019. **16**(8): p. 670-673.
41. Tribello, G.A., et al., *PLUMED 2: New feathers for an old bird.* Computer Physics Communications, 2014. **185**(2): p. 604-613.
42. Barducci, A., G. Bussi, and M. Parrinello, *Well-tempered metadynamics: a smoothly converging and tunable free-energy method.* Phys Rev Lett, 2008. **100**(2): p. 020603.
43. Torrie, G.M. and J.P. Valleau, *Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling.* Journal of Computational Physics, 1977. **23**(2): p. 187-199.
44. Pfaendtner, J. and M. Bonomi, *Efficient Sampling of High-Dimensional Free-Energy Landscapes with Parallel Bias Metadynamics.* J Chem Theory Comput, 2015. **11**(11): p. 5062-7.
45. Prakash, A., et al., *Biasing Smarter, Not Harder, by Partitioning Collective Variables into Families in Parallel Bias Metadynamics.* J Chem Theory Comput, 2018. **14**(10): p. 4985-4990.
46. Fu, H., et al., *Finding an Optimal Pathway on a Multidimensional Free-Energy Landscape.* J Chem Inf Model, 2020. **60**(11): p. 5366-5374.
47. Bottaro, S., F. Di Palma, and G. Bussi, *The role of nucleobase interactions in RNA structure and dynamics.* Nucleic Acids Res, 2014. **42**(21): p. 13306-14.
48. Sasikala, W.D. and A. Mukherjee, *Molecular mechanism of direct proflavine-DNA intercalation: evidence for drug-induced minimum base-stacking penalty pathway.* J Phys Chem B, 2012. **116**(40): p. 12208-12.
49. Mukherjee, A., et al., *On the molecular mechanism of drug intercalation into DNA: a simulation study of the intercalation pathway, free energy, and DNA structural changes.* J Am Chem Soc, 2008. **130**(30): p. 9747-55.
50. Wilhelm, M., et al., *Multistep drug intercalation: molecular dynamics and free energy studies of the binding of daunomycin to DNA.* J Am Chem Soc, 2012. **134**(20): p. 8588-96.
51. Leontis, N.B. and E. Westhof, *Geometric nomenclature and classification of RNA base pairs.* RNA, 2001. **7**(4): p. 499-512.

52. Ando, T., S. Takahashi, and K. Tamura, *Principles of chemical geometry underlying chiral selectivity in RNA minihelix aminoacylation.* Nucleic Acids Res, 2018. **46**(21): p. 11144-11152.

Chapter 7

Scope of "RNA breathing" in Protein: RNA complex recognition

7.1 Overview

Protein-RNA interactions are pivotal in numerous cellular processes, playing a key role in regulating gene expression, RNA processing, and other essential biological functions. The secondary structure of RNA involves base pairing interactions between complementary nucleotides based on edges of interactions (Watson-Crick, Hoogsteen, etc.). RNA breathing refers to the dynamic fluctuations experienced by these structures, manifesting as an oscillation between open and closed states resembling a breathing motion. The fluctuating motion enables RNA molecules to adapt various stable conformations, a feature crucial for their biological functions.

The study aims to identify the scope of RNA breathing in a given protein-RNA(L7Ae: k-turn RNA complex) recognition process using enhanced sampling techniques. We have utilized the multiple walker well-tempered metadynamics simulations to understand the phenomenon. We found a stable non-native state similar to a native state in the bound state of the protein-RNA complex. The non-native state preserves the overall architecture of RNA as in the native bound state. However, they differ in the orientation of base pair interactions. The intriguing observations shed light on the scope of the phenomenon known as "RNA breathing."

7.2 Introduction

The interaction between protein and RNA motifs plays a significant role in maintaining cellular homeostasis.[1] Any perturbation in such high-affinity interactions may lead to dysfunction of the cells, causing various diseases.[2] L7Ae is one such RNA binding protein that can form complexation with k-turn RNA motifs. A structural representation of L7Ae: k-turn RNA complex is shown in Fig. 1. These interactions were well studied by experimental techniques such as NMR and FRET, indicating that the binding affinity of these interactions in the picomolar range indicates a strong interaction.[3] These complexes are important in understanding RNA splicing and site-specific RNA modification.[4] According to the experimental results, the bound state of the complex is found to have multiple stable states and validates the existence of non-canonical base pairing schemes in RNA structure. As the experimental evidence demonstrates the existence of multiple stable states, it opens the scope of "RNA breathing" in the protein-RNA recognition process. The phenomena can be regarded as "breathing" as the residues in the RNA can show transitions from native Watson-Crick base pairing schemes to non-canonical base pairing schemes due to the presence of sugar edges and Hoogsteen[5] edges in the nuclear bases of RNA.

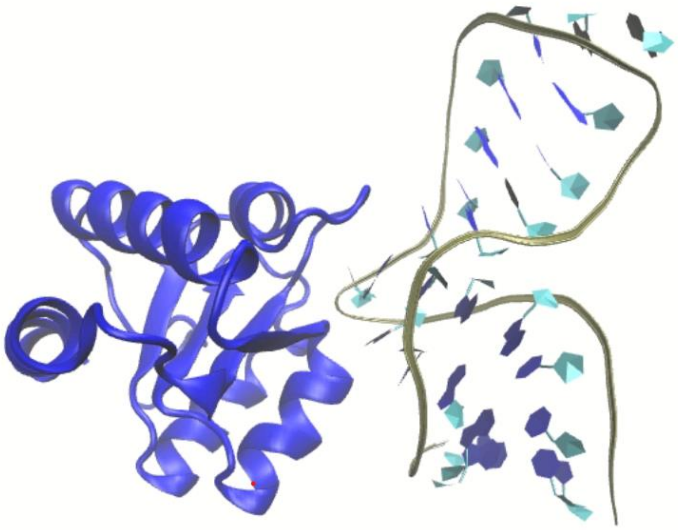

Figure 1. Structural representation of L7Ae: k-turn RNA complex (PDB ID: 4BW0)

Our study aims to understand the scope of "RNA breathing" in the protein-RNA recognition process using advanced molecular dynamics simulation protocols. We have utilized multiple walker[6] version of well-tempered metadynamics[7] using minimum distance and *ermsd[8]* as collective variables (CVs).

7.3 Computational details

7.3.1 Modeling the system

L7Ae protein: k-urn RNA complex structure (PDB ID: 4BW0) was taken from the protein data bank. The complex is inserted into a cubical box with dimensions 10.4 nm X 10.4 nm X 10.4 nm. Further, the system is solvated by adding ~35140 water molecules based on TIP3P[9] water model. To maintain the physiological ion concentration (150 mM), Na^+ and Cl^- ions were added to the system, and extra Cl^- ions were added to the system to maintain the electrical neutrality. We have used amber OL_3 (chiOL_3)[10] force field for RNA, along with amber ff14SB[11] forcefield for protein.

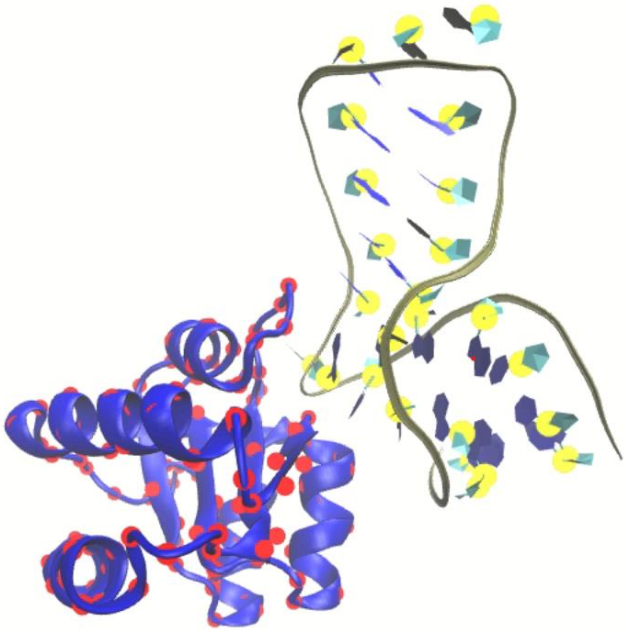

Figure 2. Representation of atoms in protein and RNA described for defining CVs. Cα atoms in protein are shown in red, and C1 carbon atoms of RNA are shown in yellow.

7.3.2 Simulation details

The protein-RNA complex was subjected to energy minimization using the steepest descent[12] method for 10000 steps. Then, annealing of the system was carried out by increasing the temperature of the system from 0K to 300K in 200 picoseconds using Berendsen[13] thermostat and barostat, both having a coupling constant of 0.6 ps. During this process, position restraints of 25 kcal/mol/$Å^2$ were enforced on heavy atoms of the protein and RNA. This was followed by a 2 ns equilibration simulation at a constant temperature of

300 K and pressure of 1 bar, without any position restraints. The same thermostat and barostat were used for equilibration, with coupling constants of 0.2 ps for each. Finally, unrestrained NVT equilibration was carried out using the Nosé-Hoover[14] thermostat with a coupling constant of 0.2 ps. All bonds were constrained during the simulations using the LINCS[15] algorithm, and the electrostatics were measured using the Particle Mesh Ewald (PME)[16] method. The cutoff distance for van der Waals (vdW) and electrostatic long-range interactions were kept at 10 Å. The timestep was set to 2 fs for all the simulations.

7.3.3 Free energy calculations

The equilibrated structure during the unbiased simulation was used for free energy calculations by employing well-tempered metadynamics simulations. The parameters for well-tempered metadynamics simulations were configured as follows: a Gaussian deposition stride of 1 ps, an initial Gaussian hill height of 1.0 kJ/mol, accompanied by a bias factor of 10. The Gaussian widths for the CVs, minimum distance, and *ermsd[8]* were specified as 0.04 nm and 0.1 nm, respectively. Two sets of simulations were carried out using either one or two CV combinations for biasing the simulations. For the simulations using one CV, minimum distance was used as a CV for well-tempered metadynamics simulations. This simulation is carried out for 100 ns. However, for the simulations, both minimum distance and *ermsd[8]* were used as CVs for bising the simulations multiple walker[6] version of well-tempered metadynamics was utilized. For 2D well-tempered metadynamics simulations, all the walkers (a total of four walkers) were started from the same starting point (from the equilibrated structure). For each walker, ~90 ns simulation was carried out. For constructing the final 2D-free energy landscape with respect to minimum distance and *ermsd[8],* bias contribution from each of the walkers was considered. The final bias-based reweighing approach was employed for reweighing purposes. [17, 18]

7.3.4 Choice of collective variables

The CVs used for the study include (i) minimum distance (ii) *ermsd[8].* The minimum distance is anticipated to identify and find the smallest possible separation or distance between the chosen atoms in protein and RNA, thereby differentiating the protein and RNA from the structure observed in the bound state and unbound state. For minimum distance calculations, the Cα atoms of protein and C1 carbon atoms of the RNA were considered as shown in the Fig. 2. The *ermsd* parameter was expected to provide local structural changes

within the bases of RNA. For *ermsd[8]* calculations, all the residues of RNA were considered with default parameters.

7.4 Results & Discussion

In the initial biased metadynamics simulation using minimum distance as a CV, we were able to identify the bound (where there is definite contact between protein and RNA) and unbound (contacts between protein and RNA is nearly zero). The time evolution of minimum distance (biased CV) and contacts during the 1D well-tempered metadynamics simulation is shown in the Fig. 3.

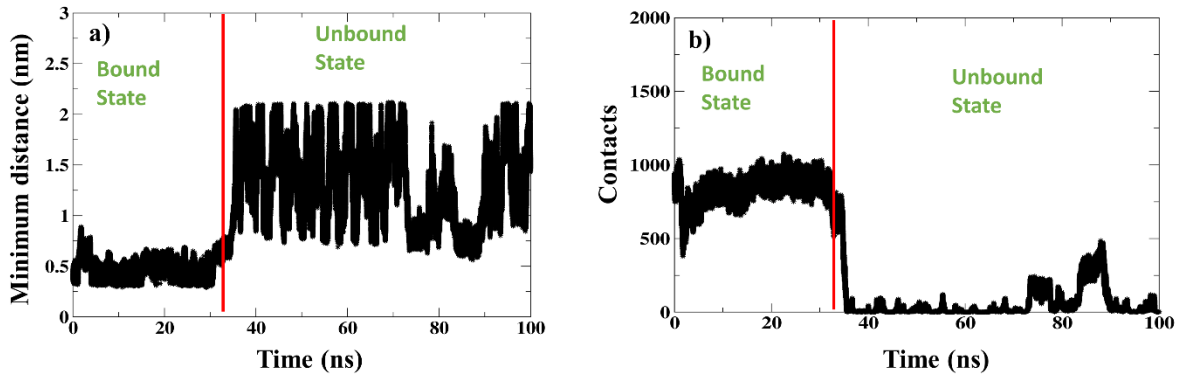

Figure. 3. a) Time evolution of minimum distance (biased variable) during well-tempered metadynamics simulation. b) Time evolution of contacts during well-tempered metadynamics simulation.

Structures corresponding to both the bound state and unbound states were extracted from the simulation trajectory. Further, the probability of the existence of various local conformational changes in the RNA residues was calculated in these clusters according to Leontis[19] geometric classification of RNA structures (Fig. 4). However, no significant changes were monitored in the probability distribution of RNA structures in the bound and unbound states. In both the bound and unbound states, WCc is the predominant classification, followed by SHc and HSc. Other transient classifications were found to be less than 10%.

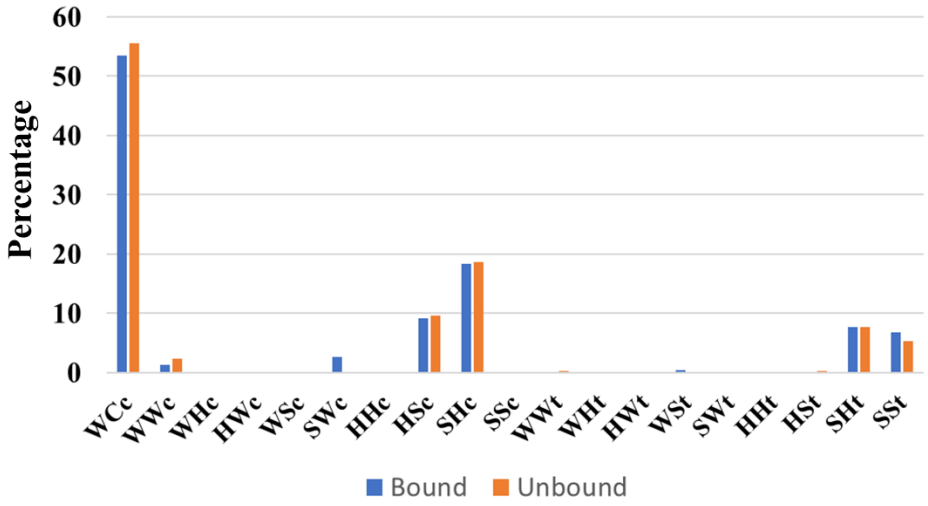

Figure 4. Distribution of various structures in RNA according to Leontis[19] geometric classification.

Further, multiple walker well-tempered metadynamics simulations were carried out using minimum distance and *ermsd* as CVs. In this strategy, *ermsd* is expected to sample the RNA states in more detail, which might have been hindered by high energy barriers in previous simulations using a single CV (minimum distance). Moreover, the multiple version, in this case, ensures better sampling efficiency. The time evolution of minimum distance, *ermsd*, and contacts during these simulations is shown in Fig. 5.

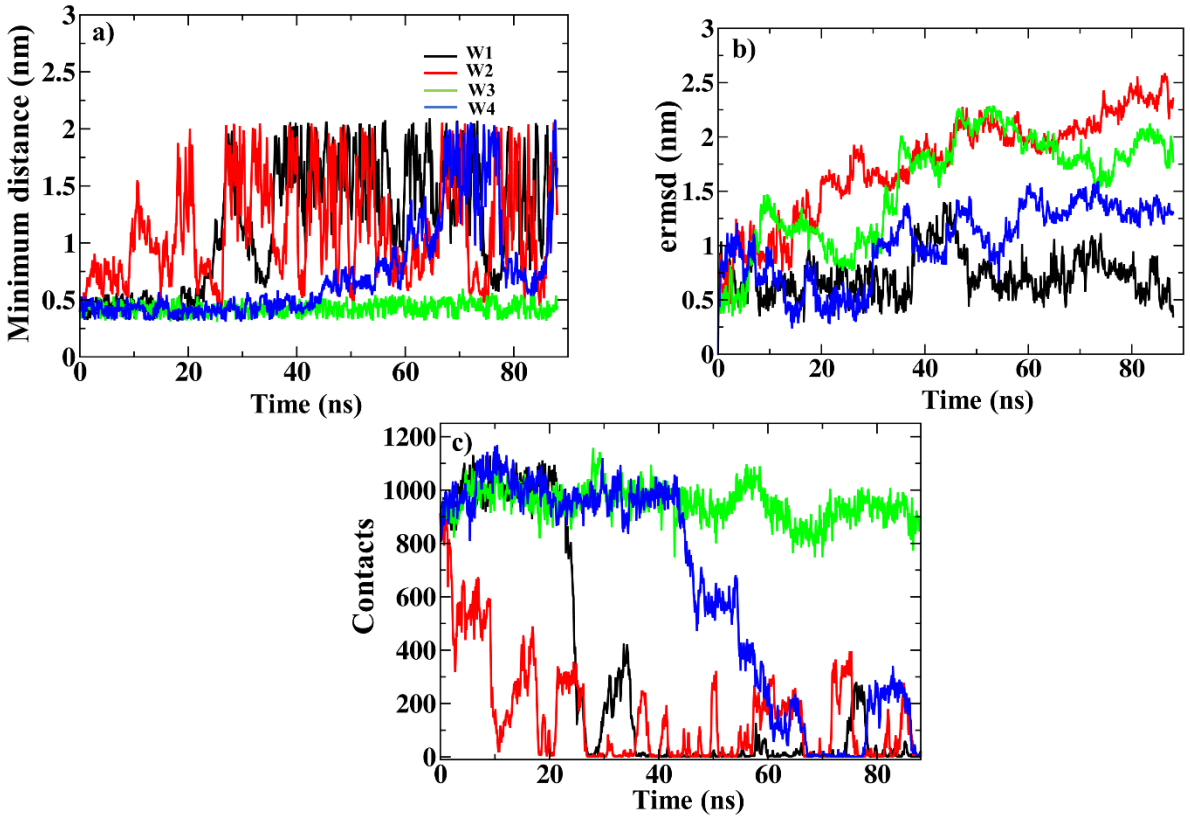

Figure 5. Time evolution of CVs in each of the walkers during multiple walker metadynamics simulations. a) minimum distance b) *ermsd* c) contacts. The color labels for each of the walkers are shown in the inset.

2D Free energy surface is constructed by combining the bias potential experienced by each of the walkers in the CV space. The 2D free energy landscape is shown in the Fig. 6, corresponding to L7Ae: k-turn RNA recognition.

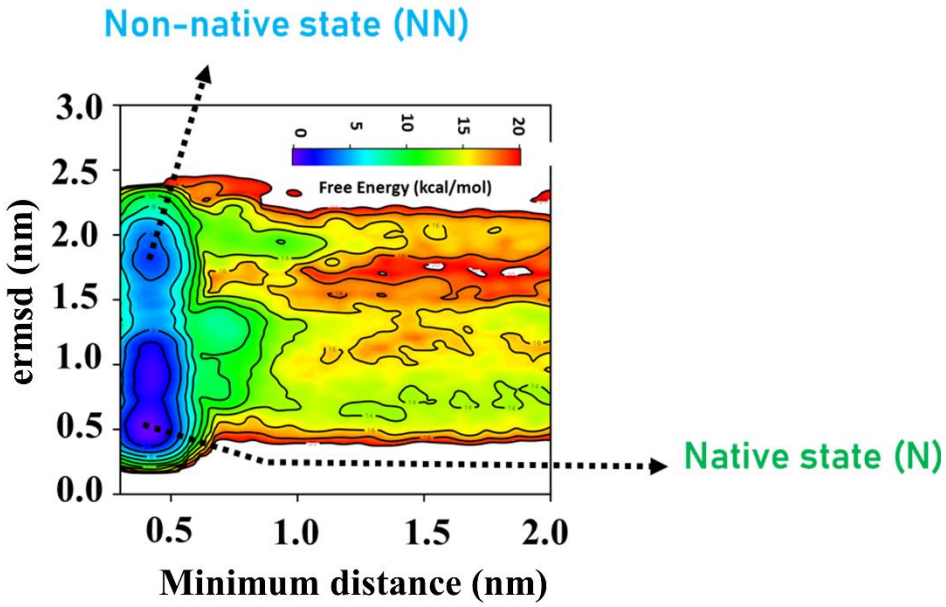

Figure 6. 2D free energy landscape with respect to minimum distance and *ermsd* showing the existence of the native state and non native state in the bound state of protein-RNA complex.

Two relevant minima were observed in the constructed free energy landscape with respect to minimum distance and *ermsd*. These are native state (minimum distance=0.4 nm, *ermsd*=0.5 nm) and non-native state (minimum distance=0.4 nm, *ermsd*=1.8 nm). It is to be noted that both the minima have the same value of minimum distance. The native state is the same as the equilibrated structure used as the starting structure for the multiple walker metadynamics simulations. However, in the non-native state, the *ermsd* is found to have a higher value. To discern the distinctions in bound structures, the contacts between the Cα atoms of protein and C1 atoms of RNA were computed for both the native and non-native states. The distribution of contacts in both native and non-native states are shown in Fig 7.

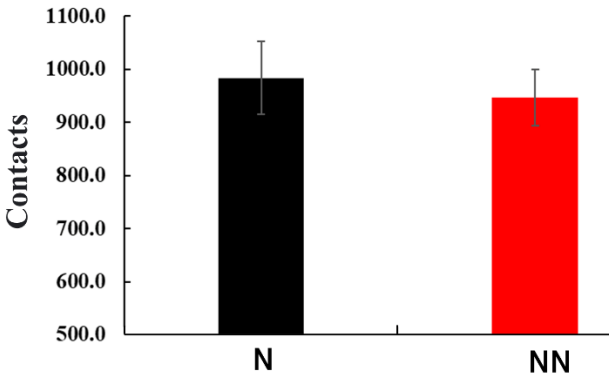

Figure. 7 Distribution of contacts between protein and RNA in the native state (N) and non-native states

In the contacts distribution (between RNA and protein), a drastic change was not observed in the non-native state compared to the native state. The superimposed structure of the non-native structure of RNA with respect to the native state of RNA is shown in the Fig. 8.

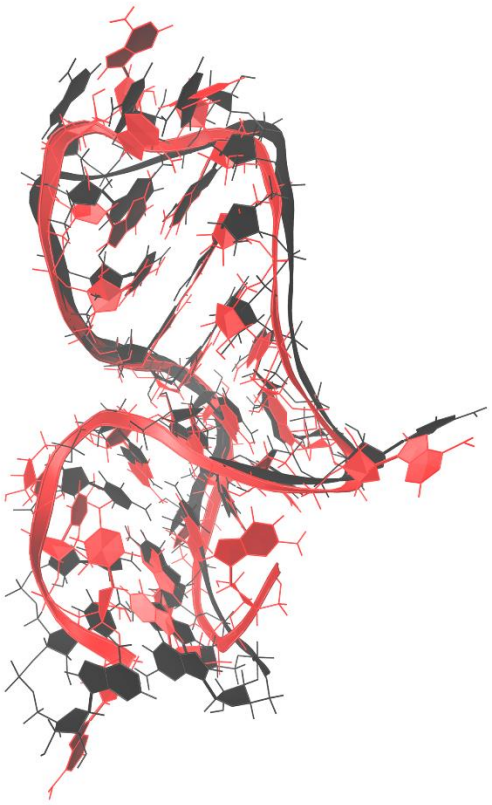

Figure 8. Superimposed structure of RNA in the non-native state (highlighted in red) with respect to the native state (black).

As shown in the Fig. 8, the overall architecture and shape of existence of RNA backbone was found to be similar in both native and non-native states. However, it was found that the internal geometry of nuclear bases was significantly altered. Various structural forms in

RNA, as described by Leontis[19] geometric classification in the native state and non-native state is shown in Fig. 9.

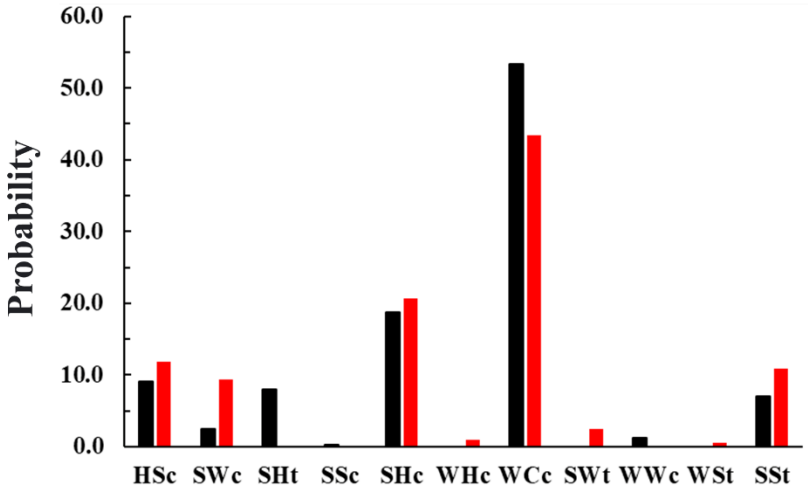

Figure 9. Probability of existence of various structures in RNA according to Leontis[19] geometric classification in the native state (shown in black) and non-native state (shown in red).

As shown in Fig. 9, the distribution is found to have altered significantly in native state and non-native states. The WCc state is the predominant state in both native and non-native states. However, the population of WCc state is relatively less in non-native states. However, in the HSc, SWc, SHc, and SSt states, there is a higher population in non-native structures than in native structures. Moreover, the native state is found to have SHt state, which is absent in the non-native state. The uneven distribution of local base pairing modes at the native and non-native states at the same bound state describes the scope of "breathing" phenomena in the L7Ae: k-turn RNA complex.

7.5 Conclusion

To understand the scope of the "breathing" phenomenon in L7Ae: k-turn RNA complex, well-tempered metadynamics simulation was utilized. During the biased simulation using only minimum distance as a CV, the complex's bound state and separated state were observed. However, it was found that biasing *ermsd* was also essential to capture the local nuclear base level structural changes in the RNA. By utilizing multiple walker metadynamics simulations using minimum distance and *ermsd* as CVs, similar to the native state, non-native states have been identified in the bound state. Even though the architecture and shape of RNA were similar in both the bound states, the local orientation of nuclear bases is found to be

different in these states, opening up the scope of "RNA breathing" . More research is going on to examine these states closely.

7.6 References

1. Lin, J. and A. Amir, *Homeostasis of protein and mRNA concentrations in growing cells.* Nature Communications, 2018. **9**(1): p. 4496.
2. Gebauer, F., et al., *RNA-binding proteins in human genetic disease.* Nature Reviews Genetics, 2021. **22**(3): p. 185-198.
3. Huang, L. and D.M. Lilley, *The molecular recognition of kink-turn structure by the L7Ae class of proteins.* RNA, 2013. **19**(12): p. 1703-10.
4. Ohno, H., et al., *Synthetic RNA–protein complex shaped like an equilateral triangle.* Nature Nanotechnology, 2011. **6**(2): p. 116-120.
5. Hoogsteen, K., *The structure of crystals containing a hydrogen-bonded complex of 1-methylthymine and 9-methyladenine.* Acta Crystallographica, 1959. **12**(10): p. 822-823.
6. Raiteri, P., et al., *Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics.* J Phys Chem B, 2006. **110**(8): p. 3533-9.
7. Barducci, A., G. Bussi, and M. Parrinello, *Well-tempered metadynamics: a smoothly converging and tunable free-energy method.* Phys Rev Lett, 2008. **100**(2): p. 020603.
8. Bottaro, S., F. Di Palma, and G. Bussi, *The role of nucleobase interactions in RNA structure and dynamics.* Nucleic Acids Res, 2014. **42**(21): p. 13306-14.
9. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water.* The Journal of Chemical Physics, 1983. **79**(2): p. 926-935.
10. Zgarbova, M., et al., *Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles.* J Chem Theory Comput, 2011. **7**(9): p. 2886-2902.
11. Maier, J.A., et al., *ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB.* J Chem Theory Comput, 2015. **11**(8): p. 3696-713.
12. Press, W.H., et al., *Numerical Recipes in FORTRAN; The Art of Scientific Computing*. 1993: Cambridge University Press.
13. Berendsen, H.J.C., et al., *Molecular dynamics with coupling to an external bath.* The Journal of Chemical Physics, 1984. **81**(8): p. 3684-3690.

14. Nosé, S., *A molecular dynamics method for simulations in the canonical ensemble.* Molecular Physics, 1984. **52**(2): p. 255-268.
15. Hess, B., et al., *LINCS: A linear constraint solver for molecular simulations.* Journal of Computational Chemistry, 1997. **18**(12): p. 1463-1472.
16. Darden, T., D. York, and L. Pedersen, *Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems.* The Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.
17. Bonomi, M., A. Barducci, and M. Parrinello, *Reconstructing the equilibrium Boltzmann distribution from well-tempered metadynamics.* J Comput Chem, 2009. **30**(11): p. 1615-21.
18. Tribello, G.A., et al., *PLUMED 2: New feathers for an old bird.* Computer Physics Communications, 2014. **185**(2): p. 604-613.
19. Leontis, N.B. and E. Westhof, *Geometric nomenclature and classification of RNA base pairs.* RNA, 2001. **7**(4): p. 499-512.

Chapter 8

Selective anion permeation through synthetic ion channel

8.1 Overview

Ion transport across cell membranes is an essential component for the optimal functioning of cells. Synthetic ion channels have been recognized for their ability to selectively facilitate the permeation of anions through biological membranes, mimicking the properties inherent in naturally occurring protein channels. These synthetic ion channels offer several merits, particularly in the fields of nanotechnology and medical applications. The study aims to provide molecular insights into the experimentally reported observations on the selective Cl^- ion transport through the synthetic ion channels, constructed based on bis(1,3-propanediol)-linked m-dipropynylbenzene. Through molecular dynamics simulations, we identified that selective Cl^- ion permeation is triggered by hydrogen bonds formed by the atoms present in the channels and Cl^- ion. Additionally, we employed multiple walker metadynamics simulations to gain insights into the mechanistic aspects of the permeation of a single Cl- ion through the channel. This approach allows us to uncover the free energy stability associated with the selective permeation events.

8.2 Introduction

The transport of ions across the cell membrane is a vitally important physiological process, ensuring the optimal functioning of cells.[1-5] Ion transport proteins are pivotal in facilitating the movement of ions across cell membranes, a fundamental process for diverse physiological functions.[6] These proteins actively contribute to upholding the essential electrochemical gradients crucial for cell function, signal transduction, and the regulation of osmotic balance.[7, 8]. Any defect or damage in these intricate transport systems, referred to as channelopathies, can have profound and detrimental effects on the physiological functions of the cell.[9] Disruptions in the function of ion channels lead to various channelopathies, including cystic fibrosis, Barter's syndrome, Dent's disease, Long QT syndrome, osteoporosis, epilepsy, Parkinson's disease, etc.[9-11] Nevertheless, mimicking the functions of the nature of natural ion transport channels poses a challenge for synthetic chemists, along with designing simplified systems capable of facilitating the transport of specific ions. Synthetic ion channels are artificially crafted chemical compounds that mimic the function of natural ion channels and enable the transport of ions across cell membranes or lipid bilayers.[12]

1,3-Propanediol is recognized for its involvement in diverse self-assembly processes facilitated by hydrogen bonding interactions.[13-16] This intriguing characteristic expands the possibilities for designing novel supramolecular ion channels that feature single-file ion recognition sites. As per the experimental observations conducted by Pinaki Talukdar's group, IISER Pune, the channel linked by amide has shown significant activity and demonstrated selectivity towards chloride ions within lipid bilayer membranes. The study aims to elucidate the mechanistic events linked to the selective transport of Cl^- ions through the reported bis(1,3-propanediol)-linked m-dipropynylbenzene based supramolecular ion channels (Fig. 1), aiming to complement the experimental observations through computational studies.[17] The molecular dynamics simulations conducted by us have shown that the selectivity is attributed to the favorable interactions provided by the components of the channel and the Cl^- ion.

8.3 Computational details

8.3.1 System preparation

A system consisting of a channel and membrane is made by incorporating the previously optimized synthetic channel into the DPHPC membrane. A simulation box with dimensions 7.0 X 7.4 X 10.0 nm^3 is used to incorporate the channel and membrane in the system. This

system was then solvated with 10586 SPC model l[18] water molecules, along with 63 Na+ and 63 Cl- ions making a target concentration of 200 mM. The pre-equilibrated structure and parameters for DPHPC (1,2-diphytanoyl-sn-glycero-3-phosphocholine) were obtained from Lipidbook[19]. To position the optimized channel in the simulation box, inflategro[20] methodology was utilized, wherein 128 DPHPC molecules surrounded the channel. The automated topology builder[21], developed by Malde et al., was employed to generate the topology parameters for the channel using the GROMOS-53a6 force field[22].

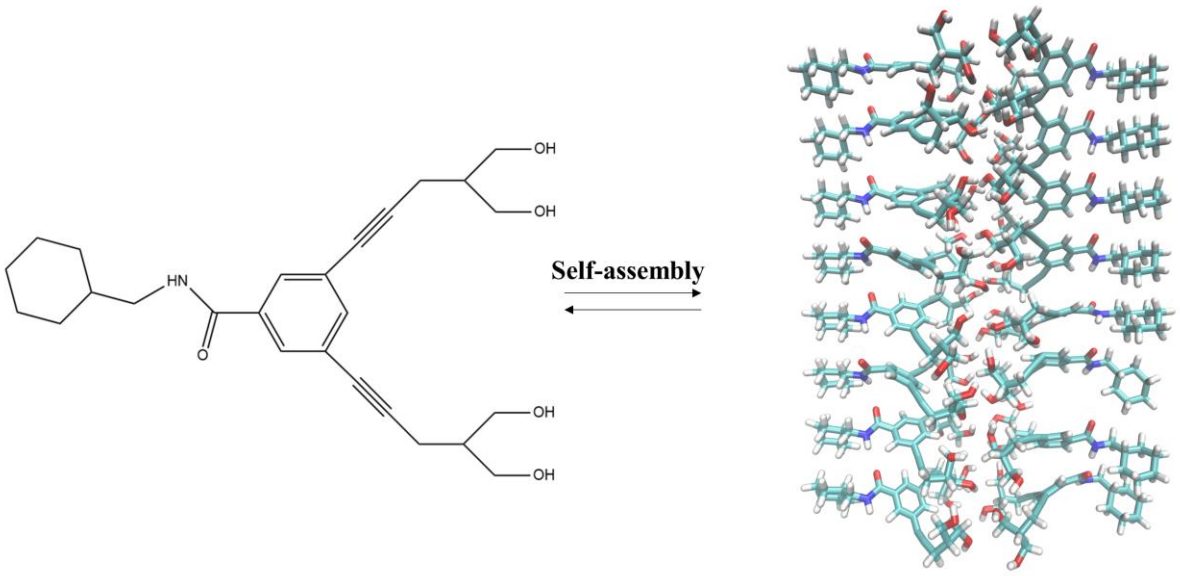

Figure 1. Structural representation of reported monomer of bis(1,3-propanediol)-linked m-dipropynylbenzene based molecule and corresponding supramolecular channel.[17]

8.3.2 Simulation details

The modeled system (Fig. 2) underwent energy minimization for 50000 steps using the steepest descent method[23]. This was followed by Equilibration at a constant temperature of 323 K and pressure of 1 bar for 10 ns, employing the Nosé-Hoover thermostat[24] and Parrinello-Rahman barostat[25] with coupling constants of 0.5 ps and 5 ps, respectively. The simulations were conducted with a time step of 2 fs and utilized PME electrostatics[26] for electrostatic interactions, with a cutoff of 12 Å for both electrostatic and van der Waals (vdW) interactions.

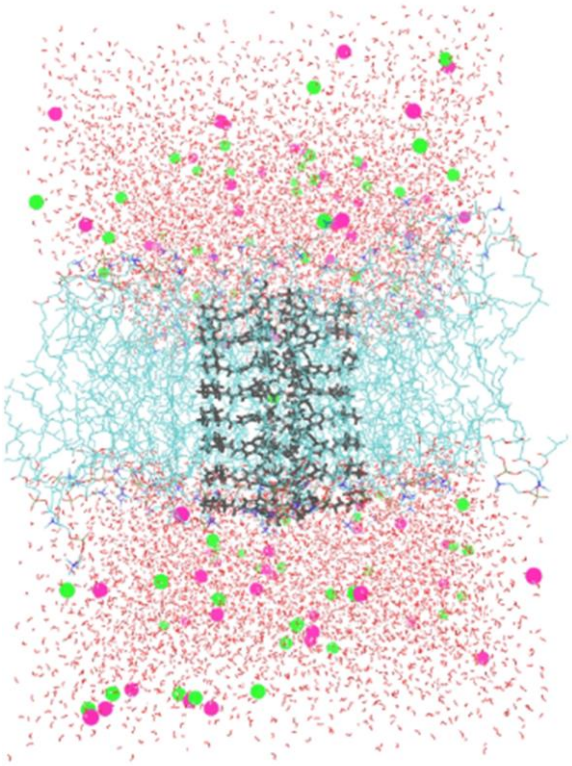

Figure 2. Representation of the modeled system consisting of a channel inserted in the DPHPC membrane surrounded by water (red), sodium ions (purple), chloride ions (green)

Position restraints of 500 kJ/mol nm^2 were applied to the heavy atoms of the channel atoms. Finally, a 50 ns molecular dynamics simulation was performed at a constant temperature of 323 K and pressure of 1 bar, utilizing the Nosé-Hoover thermostat[24] and Parrinello-Rahman barostat[25] with coupling constants of 0.5 ps and 2 ps, respectively. The electrostatics and vdW interactions were kept similar to the equilibration process. All the simulations were performed using the GROMACS 2019.6[27] package. The partial density profiles of the components during the Equilibration are shown in the Fig. 3a. The overall alignment of the membrane with respect to the channel axis is found to be stable during the simulation as the probability distribution corresponding to the angle formed by the channel axis, head group and the tail group of the membrane forms two distinct perks (Fig. 3b) corresponds to alignment in opposite directions with respect to channel center.

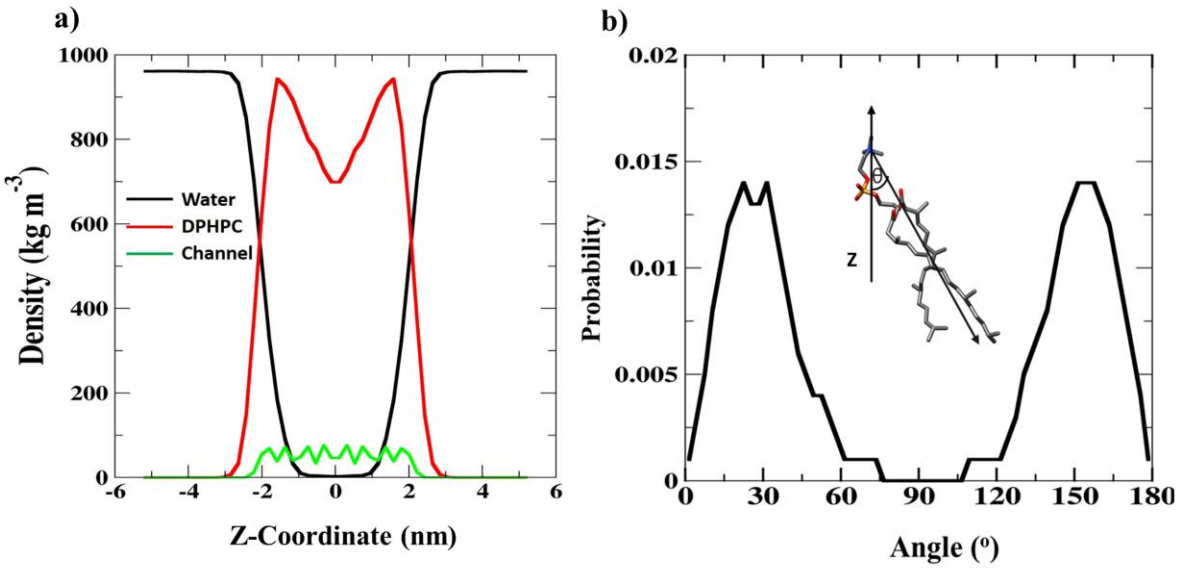

Figure 3. a) Partial density profiles of the components water, DPHPC membrane, and channel) present in the system. b) Probability distribution of the angle established by the channel axis in Z–direction and a vector that connects the head group of the DPHPC to the tail end of the DPHPC membrane. A representation of the angle is shown in the inset.

8.3.3 Free energy calculations

The multiple walker version[28] of well-tempered metadynamics[29] is used for free energy calculations associated with ion permeation through the channel. To construct multiple starting points for each walker, the reference chloride ion was kept at different channel locations. These initial configurations were taken from the permeation event monitored in the unbiased molecular dynamics simulation trajectory. In this multiple walker metadynamics simulation protocol, individual biased simulations were conducted parallelly, and the bias contribution from each walker was taken into account to construct the final free energy surface for the permeation process until the final ion exited the channel.

A total of 8 walkers, each representing a different independent trajectory, were constructed to study permeation events. A total of 160 ns simulations were carried out, each lasting 20 ns. Permeation distance (discussed later) is used as a collective variable (CV) for these biased simulations. Gaussian hill height of 0.2 kJ/mol, a bias factor of 10, a hill deposition rate of 1 ps, and Gaussian widths of 0.05 nm were used as metadynamics simulation parameters.

8.3.4 Choice of collective variables

a) Permeation distance: Permeation distance is defined as the displacement of Cl- ion along the Z-coordinate between the center of mass of the channel and the reference chloride ion. This

CV quantitatively describes the position of an ion inside or outside the channel during the simulation. The permeation distance has a value of 0 nm near the center of the channel, while it exceeds 2.5 nm when the ion reaches the bulk water.

b) Coordination number: Coordination number calculates the possible contacts between the reference chloride ion (group A) and oxygen atoms (group B) of the water molecules within a given cutoff distance. This parameter is defined by a switching function as follows:

$$Coord.Number = \sum_{i \in A} \sum_{i \in B} s_{ij}$$

$$s_{ij} = \frac{1 - \left(\frac{r_{ij} - d_0}{r_0}\right)^n}{1 - \left(\frac{r_{ij} - d_0}{r_0}\right)^m}$$

The distance between the reference chloride ion and the neighboring oxygen atoms of the water molecules is described by r_{ij}. The values of r_0 (cutoff distance) and d_0 are set to 0.35 nm and 0 nm, respectively. Further, the user-defined parameters were chosen to be n= 6, m= 12.

8.4 Results & Discussion

8.4.1 Permeation event during the unbiased simulation

A permeation event of chloride ion was observed during the unbiased production run between the time intervals 19.4 ns and 26.0 ns (Fig. 4). The position of chloride ion with respect to Z-Coordinate during the permeation event is shown in Fig. 5. During the permeation event, it was found that when the chloride ion comes nearer to the channel terminal, it establishes hydrogen bonds with the phenyl hydrogens and hydroxyl hydrogens. Moreover, the hydrogen bonds formed between the ion and water are relatively less than in bulk water during the event. The time evolution of hydrogen bonds established by the channel hydrogen atoms (phenyl hydrogen and hydroxyl hydrogen) towards the chloride ion during the unbiased simulation is shown in Fig. 6a and 6b, respectively. Similarly, the time evolution of the coordination number established between the permeated chloride ion and water molecules is shown in Fig. 6c.

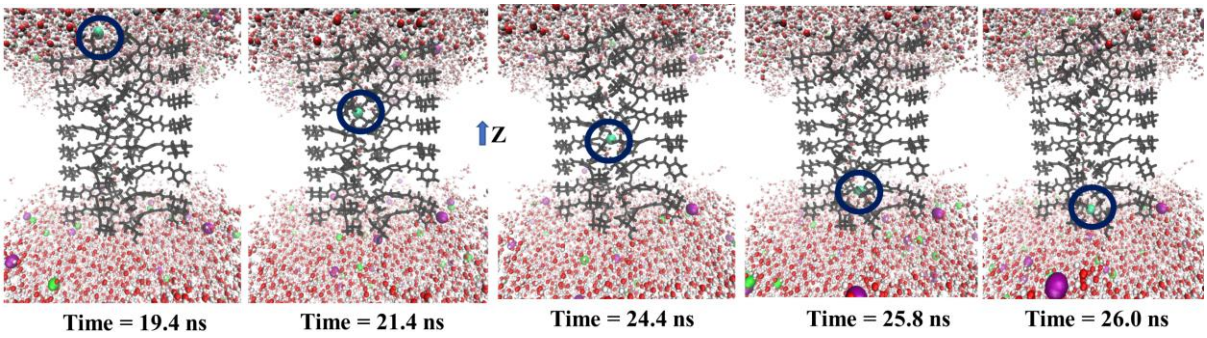

Figure 4. Representation of permeation events during different time instances of the unbiased simulation. Channel, water, sodium, and chloride are shown in black, red, purple, and green, respectively. Atoms of the membrane were omitted for clarity. The permeating ion is encircled in blue at different times.

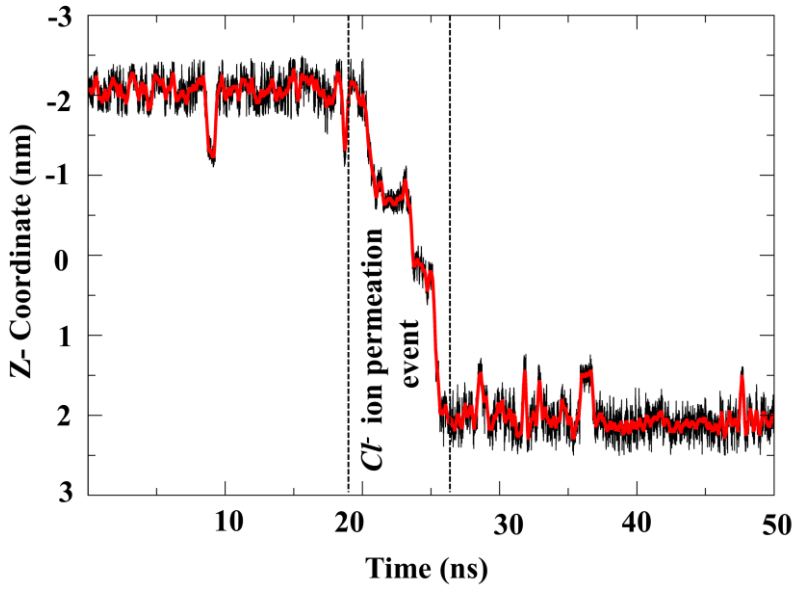

Figure 5. Variation in the Z-coordinate of chloride ion during the permeation event (shown in black) and its running average (shown in red) over 100 frames. Permeation event is observed during the time interval between 19.4 ns and 26.0 ns.

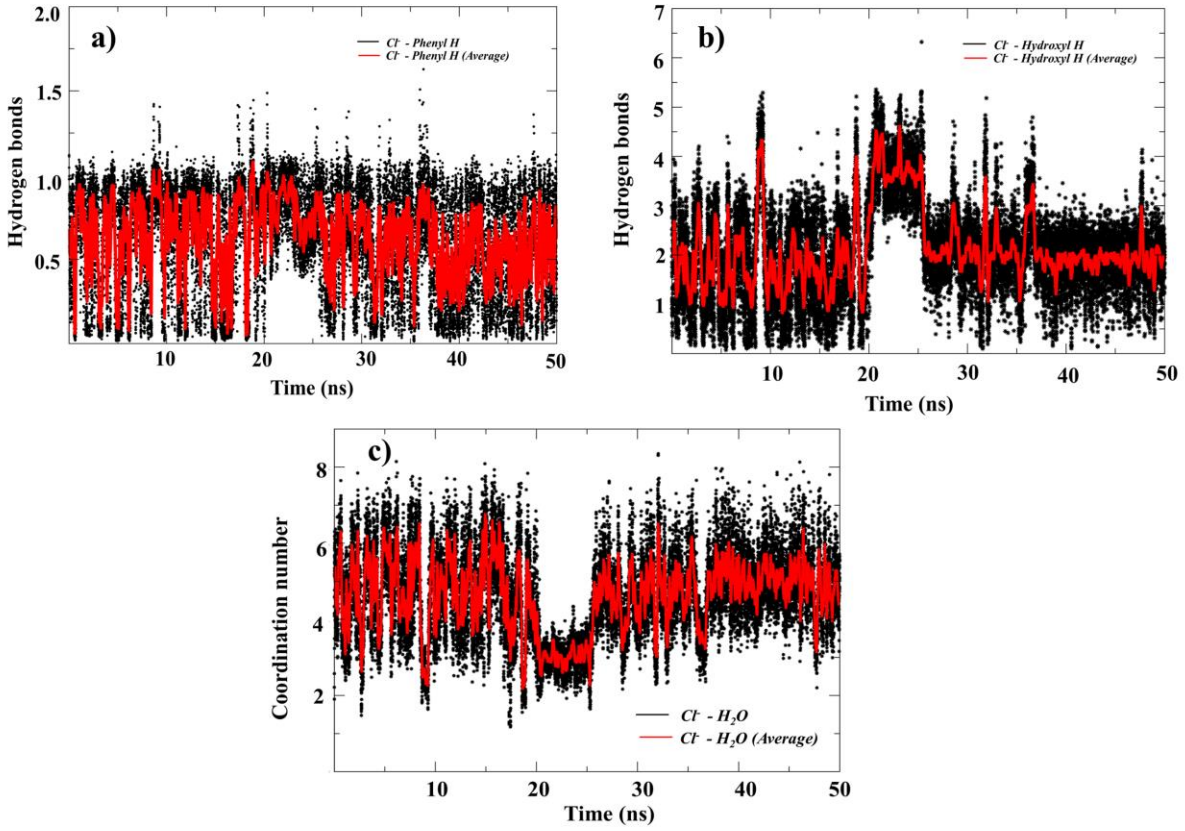

Figure 6. a) Time dependence of hydrogen bond between the phenyl hydrogen and permeated chloride ion b) Time dependence of hydrogen bond between hydroxyl hydrogen and permeated chloride ion c) Time dependence of Coordination number between the permeated chloride ion and water molecules.

These interesting observations show that the interface between channel ends and bulk water acts as a transitory intermediate during the permeation event. The variation of hydrogen bonds and the coordination number with water molecules during the permeation pathway (between 19.4 ns and 26.0 ns) with respect to the Z-coordinate (channel axis) is shown in the Fig. 7.

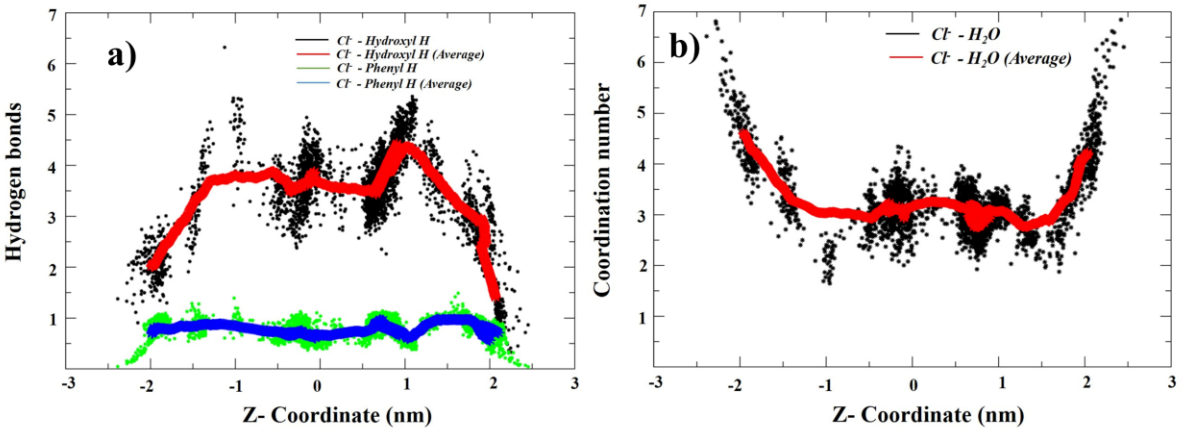

Figure. 7. a) Dependence of hydrogen bonds with respect to Z-coordinate, established between the permeated chloride ions with hydroxyl and phenyl hydrogens present in the channel during the permeation event. The start

and end terminals of the channel residues have the Z-coordinate values -2.2 nm and 2.2 nm, respectively. b) Dependence of coordination number between chloride ion and water molecules during the permeation event.

8.4.2 Free energy calculations

To understand the free energy change associated with the permeation of a single Cl^- ion, multiple walker metadynamics[28] simulation was carried out (see methods) by placing the ion in the various parts of the channel, varying from the center of the channel to end terminals. The starting configurations of channel and ion for constructing multiple walkers are shown in Fig. 8. These starting configurations were taken from the trajectory corresponding to the permeation event during unbiased simulation. 2D free energy surface is constructed by reweighing (final bias-based reweighing algorithm) [30-32]the bias contribution from each of the walkers used in the simulation with respect to permeation distance and coordination number (between water molecules with the Cl^-).

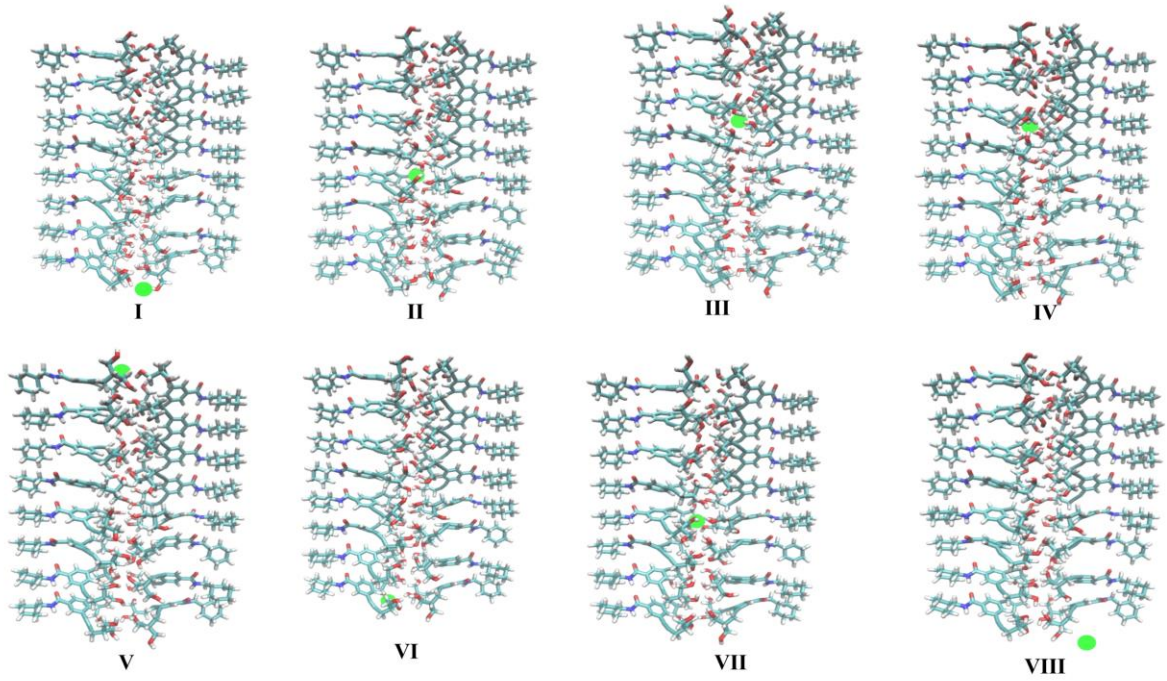

Figure 8. Representation of initial configurations sed for multiple walker metadynamics simulations. The reference chloride ion (green) is positioned at the various parts of the channel (Z-coordinate between -2.2 nm and 2.2 nm).

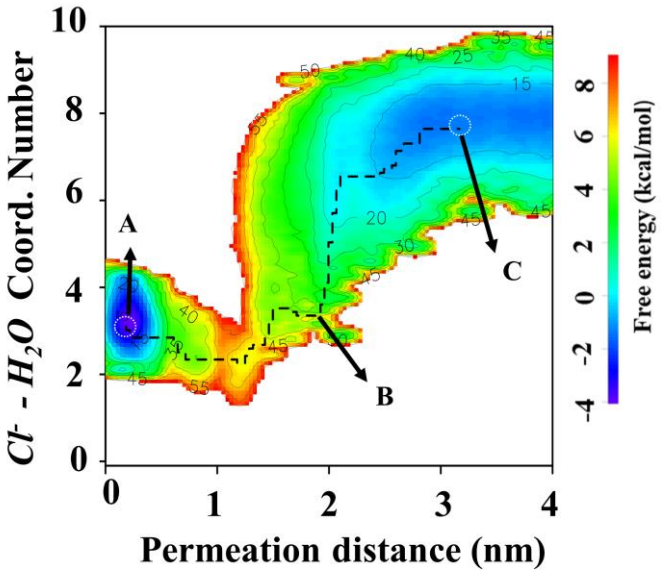

Figure 9. 2D-free energy surface corresponds to the permeation of ion through the channel, constructed based on multiple walker metadynamics simulations. Minima A and C represent the fully permeated state (inside the channel) and free state in the bulk water (outside the channel) respectively. The transition state is indicated by state B. The black dotted line in the free energy surface shows the Minimum free energy path connecting the minima A and C.

The 2D-dimensional free energy plot corresponding to the permeation of ions through the channel is shown in Fig. 9. The minimum free energy path connecting the permeated state and free state (bulk water) is shown by black dotted lines. A stable minima (minima A) is observed to correspond to a completely permeated state at a permeation distance of 0.2 nm and coordination number 3. When the permeated chloride ion reaches the terminal region of the channel, it visits a transition state B. Further, a flat minima state (minima C) is monitored when the chloride ion arrives at the bulk water. The representation channel, ion and sorounding waters around the ion in the states A, B, and C are shown in the Fig. 10. The free energy change between the free state (when the ion is in the bulk water) and when the ion is in the completely permeated state is found to be -2.9 kcal/mol.

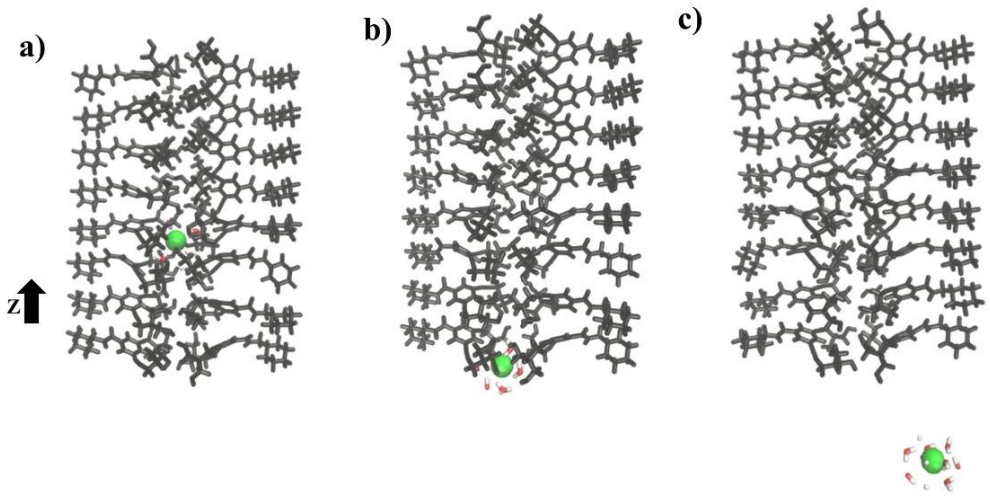

Figure 10. Representation of relevant states associated with the permeation of chloride ion, based on the free energy surface shown in Fig. 9. Green represents the Cl^- ion, and red represents the surrounding water molecules. The channel is shown in black. a) minima state A (permeated state) b) Transition state B c) minima state C (free state in bulk water)

8.5 Conclusion

Using molecular dynamics simulation techniques, we investigated the selective permeation of chloride ions through a self-assembled structure comprising amide-linked bis(1,3-propanediol)-based molecules incorporated across a lipid bilayer membrane. Our results show that the hydrogen bond between the Cl^- ion and phenyl hydrogens in the channel triggers the selective permeation through the channel. Moreover, we also found that the interface between the bulk water and channel acts as a transitory intermediate state, which assists the permeation events. In order to comprehend the free energy change linked with the permeation event, we conducted multiple walker metadynamics simulations. The resulting analysis reveals a free energy change of -2.9 kcal/mol associated with the permeation of Cl- ions from the center of the channel to bulk water. The simulation protocol can be extended to study the permeation vents through synthetic ion channels and also to comment on the mechanistic details associated with similar processes.

8.6 Collaborations

This work was conducted in collaboration with Rashmi Sharma and Prof. Pinaki Talukdar at IISER Pune. The molecule and the initial atomic coordinates of the assembled synthetic ion channel were provided by Pinaki Talukdar's group for us to perform computational calculations. All computational work and subsequent analyses were carried out by me.

8.7 References

1. Jentsch, T.J., C.A. Hubner, and J.C. Fuhrmann, *Ion channels: function unravelled by dysfunction.* Nat Cell Biol, 2004. **6**(11): p. 1039-47.
2. Song, L., et al., *Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore.* Science, 1996. **274**(5294): p. 1859-66.
3. Davis, J.T., O. Okunola, and R. Quesada, *Recent advances in the transmembrane transport of anions.* Chem Soc Rev, 2010. **39**(10): p. 3843-62.
4. Hladky, S.B., *Ion transport and displacement currents with membrane-bound carriers: the theory for voltage-clamp currents, charge-pulse transients and admittance for symmetrical systems.* J Membr Biol, 1979. **46**(3): p. 213-37.
5. Chiu, S.Y. and G.F. Wilson, *The role of potassium channels in Schwann cell proliferation in Wallerian degeneration of explant rabbit sciatic nerves.* J Physiol, 1989. **408**: p. 199-222.
6. Gadsby, D.C., *Ion channels versus ion pumps: the principal difference, in principle.* Nat Rev Mol Cell Biol, 2009. **10**(5): p. 344-52.
7. Diamond, J.M. and E.M. Wright, *Biological membranes: the physical basis of ion and nonelectrolyte selectivity.* Annu Rev Physiol, 1969. **31**: p. 581-646.
8. Cahalan, M.D., et al., *A voltage-gated potassium channel in human T lymphocytes.* J Physiol, 1985. **358**: p. 197-237.
9. Dworakowska, B. and K. Dołowy, *Ion channels-related diseases.* Acta Biochimica Polonica, 2000. **47**(3): p. 685-703.
10. Gale, P.A., J.T. Davis, and R. Quesada, *Anion transport and supramolecular medicinal chemistry.* Chem Soc Rev, 2017. **46**(9): p. 2497-2519.
11. Kasianowicz, J.J., *Introduction to ion channels and disease.* Chem Rev, 2012. **112**(12): p. 6215-7.
12. Fyles, T.M., *Synthetic ion channels in bilayer membranes.* Chem Soc Rev, 2007. **36**(2): p. 335-47.
13. Marshall, J.A. and G.S. Bartley, *Observations Regarding the Ag(I)-Catalyzed Conversion of Allenones to Furans.* The Journal of Organic Chemistry, 2002. **59**(23): p. 7169-7171.
14. Long, R., et al., *Asymmetric total synthesis of (-)-lingzhiol via a Rh-catalysed [3+2] cycloaddition.* Nat Commun, 2014. **5**: p. 5707.

15. Huang, L.B., et al., *Hydroxy Channels-Adaptive Pathways for Selective Water Cluster Permeation.* J Am Chem Soc, 2021. **143**(11): p. 4224-4233.
16. Mondal, T., et al., *Extended supramolecular organization of pi-systems using yet unexplored simultaneous intra- and inter-molecular H-bonding motifs of 1,3-dihydroxy derivatives.* Chem Commun (Camb), 2015. **51**(24): p. 5040-3.
17. Sharma, R., et al., *Self-assembled anion channel formation by bis(1,3-propanediol)-linked meta-dipropynylbenzene-based small molecules.* Chem Commun (Camb), 2023. **59**(24): p. 3602-3605.
18. Berendsen, H.J.C., et al., *Interaction Models for Water in Relation to Protein Hydration.* 1981. **14**: p. 331-342.
19. Domański, J., et al., *Lipidbook: A Public Repository for Force-Field Parameters Used in Membrane Simulations.* The Journal of Membrane Biology, 2010. **236**(3): p. 255-258.
20. Kandt, C., W.L. Ash, and D. Peter Tieleman, *Setting up and running molecular dynamics simulations of membrane proteins.* Methods, 2007. **41**(4): p. 475-488.
21. Malde, A.K., et al., *An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0.* Journal of Chemical Theory and Computation, 2011. **7**(12): p. 4026-4037.
22. Oostenbrink, C., et al., *A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6.* Journal of Computational Chemistry, 2004. **25**(13): p. 1656-1676.
23. Deift, P. and X. Zhou, *A Steepest Descent Method for Oscillatory Riemann--Hilbert Problems. Asymptotics for the MKdV Equation.* The Annals of Mathematics, 1993. **137**(2).
24. Nosé, S., *A molecular dynamics method for simulations in the canonical ensemble.* Molecular Physics, 1984. **52**(2): p. 255-268.
25. Parrinello, M. and A. Rahman, *Polymorphic transitions in single crystals: A new molecular dynamics method.* Journal of Applied Physics, 1981. **52**(12): p. 7182-7190.
26. Darden, T., D. York, and L. Pedersen, *Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems.* The Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.
27. Hess, B., et al., *GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation.* Journal of Chemical Theory and Computation, 2008. **4**(3): p. 435-447.

28. Raiteri, P., et al., *Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics.* J Phys Chem B, 2006. **110**(8): p. 3533-9.
29. Barducci, A., G. Bussi, and M. Parrinello, *Well-tempered metadynamics: a smoothly converging and tunable free-energy method.* Phys Rev Lett, 2008. **100**(2): p. 020603.
30. Bonomi, M., A. Barducci, and M. Parrinello, *Reconstructing the equilibrium Boltzmann distribution from well-tempered metadynamics.* Journal of Computational Chemistry, 2009. **30**(11): p. 1615-1621.
31. Pfaendtner, J. and M. Bonomi, *Efficient Sampling of High-Dimensional Free-Energy Landscapes with Parallel Bias Metadynamics.* J Chem Theory Comput, 2015. **11**(11): p. 5062-7.
32. Tribello, G.A., et al., *PLUMED 2: New feathers for an old bird.* Computer Physics Communications, 2014. **185**(2): p. 604-613.

Chapter 9

Conclusion

9.1 Summary

This thesis uses enhanced sampling techniques to investigate a few molecular recognition processes in biological systems. Free energy – a thermodynamic parameter was used to understand the interactions, stability, and molecular mechanisms associated with the discussed molecular recognition processes. The selection and design of collective variables (CVs) in the studies were determined through a process of trial and error, drawing from reported CVs in similar systems and guided by visual inspection of the molecular dynamics trajectories.

We found that well-tempered metadynamics, an adaptive enhanced sampling method, efficiently accelerated the sampling of rare events and overcame energy barriers in the configuration space associated with the biological processes. The effectiveness of metadynamics in expediting the sampling of rare events and reconstructing free-energy landscapes is highly contingent on the chosen CVs. This method enabled us to accurately replicate the experimental binding free energy stability of the processes of interest, particularly in the context of interactions between proteins and small molecules in SARS-CoV-2-based systems. Through thorough validation studies employing the well-tempered metadynamics simulation protocol as discussed in Chapters 3 and 4, we identified a selection of both existing and novel small molecules that exhibit effective binding at the active sites of the spike protein and RdRp. However, for a few complex systems such as protein-protein/protein-RNA complexes, we found that multiple walker versions of well-tempered metadynamics will be well-suited, as multiple starting configurations can be considered for biasing, increasing the quality of sampling.

In the L7Ae: k-turn RNA-based systems we have extensively used the multiple walker well-tempered metadynamics using two CVs to account for the “breathing” phenomena in the RNA: Protein recognition process. As discussed in Chapter 7, we have used 4 walkers, starting at the same configuration of the equilibrated bound structure. Similarly, we have utilized multiple walker well-tempered metadynamics to understand ion permeation events through synthetic ion channels. Note that here 8 walkers are used each starting from a different configuration, as discussed in Chapter 8.

However, as the dimensionality of the CV space increases, filling a multidimensional Gaussians becomes more expensive. In these cases, where more than two CVs are required to sample the high-dimensional free energy landscape, we found that parallel bias metadynamics will be a good choice, as this method considers simultaneous multiple low-dimensional bias potentials along the CV space. Multiple walker parallel bias metadynamics using three CVs were used to estimate free energy and understand the mechanistic information associated with the unbinding T-cell factor (TCF) from the β-catenin interface (discussed in Chapter 5). Using this simulation protocol using three CVs and multiple configurations as the starting points (five walkers), we matched the experimentally reported binding affinity and suggested relevant secondary structural changes in the TCF during the unbinding pathway.

In the metadynamics-based methods, we observed the convergence-related issues and false minima states, especially in the RNA tetraloop folding studies. We identified that employing controlled exploration methods is well-adapted to address these challenges. We have utilized a combination of umbrella sampling and parallel bias metadynamics by using five CVs to understand the RNA tetraloop folding-unfolding equilibria in the presence of enantiomers of arginine. We were able to construct the converged free energy landscape using this protocol as discussed in Chapter 6. Our results indicate that the molecular origin of chiral specificity in RNA folding behavior is because of differences in different binding modes established by the enantiomers.

Moreover, choosing the wrong choice of collective for biasing may result in exploring false minima during the simulation and may lead to misinterpretation of results, especially convergence-related issues. We found that selecting a suitable intuitive CV for comprehending the targeted biological process is not straightforward. The wrong choice of CVs for biasing can result in result misinterpretation, exploration of false minima, and convergence-related challenges. Furthermore, opting for an enhanced sampling method tailored to the specific system is crucial for obtaining a converged free energy landscape. Advanced metadynamics versions or controlled sampling approaches can be employed to construct the high-dimensional free energy landscape and extract meaningful insights from the free energy landscapes.

To conclude, the multiple walker approach improves the overall sampling efficiency compared to single starting point-based traditional metadynamics techniques (including non-tempered and well-tempered metadynamics). If the convergence of the simulation is not observed in the single starting point-based metadynamics techniques, even after the selection of a good

collective variable, the multiple walker approach will be a better option. Further, if the free energy landscape is high-dimensional, where more than two collective variables are required for biasing, the multiple walker approach can be combined with the parallel bias metadynamics technique to increase the exploration of states. However, if none of the aforementioned strategies prove effective and if trapped minima states and high energy barriers are present, employing controlled sampling approaches (a combination of umbrella sampling and metadynamics-based methods) may be worth considering for the system of interest.

9.2 Future directions

This thesis highlights the necessity for an appropriate enhanced sampling method, aided with relevant intuitive CVs, to gain insights into the molecular recognition processes within a few protein and RNA-based systems. The primary future goal is to identify more effective non-intuitive CV-based approaches that can effectively describe conformational changes, metastable states, and energy barrier regions. This includes employing machine learning-based CV optimization (HLDA, tICA, etc.) and path CV-based studies for the investigation of diverse molecular recognition processes and comment on the free energy stability.[1]

Besides understanding free energy stability, obtaining kinetic information from biomolecular simulations is challenging yet provides valuable insights for comparing experimentally reported observations. The second future plan includes extracting the biomolecular kinetics for relavent protein-drug systems from simulations that rely on both the molecular pathways connecting the stable states and the free energy barrier associated with transition states. This information holds significant value from a computational perspective, as the increasing recognition of ligand-target residence time as a crucial parameter in drug discovery underscores its reliable predictive capacity for in vivo drug efficacy.[2]

The third goal is to utilize quantum mechanics/molecular mechanics (QM/MM) calculations to study relevant biological responses in protein-drug complexes that involve reactions with complex electronic structures. For example, extracting the bond-breaking/formation information from the free energy minima associated with the bound state in a given protein-drug interaction could provide new insights into the role of specific residues in small molecule recognition. [3] Such molecular-level information could account for building/modeling novel/existing protein-ligand complexes and for setting new simulation protocols for gaining new information along with the results from advanced molecular dynamics simulation techniques.

9.3 References

1. Bussi, G. and A. Laio, *Using metadynamics to explore complex free-energy landscapes.* Nature Reviews Physics, 2020. **2**(4): p. 200-212.
2. Mollica, L., et al., *Molecular Dynamics Simulations and Kinetic Measurements to Estimate and Predict Protein-Ligand Residence Times.* J Med Chem, 2016. **59**(15): p. 7167-76.
3. Gleeson, M.P. and D. Gleeson, *QM/MM calculations in drug discovery: a useful method for studying binding phenomena?* J Chem Inf Model, 2009. **49**(3): p. 670-7.

Targeting RdRp of SARS-CoV-2 with De Novo Molecule Generation

Author: Amal Vijay, Venkata Sai Sreyas Adury, Arnab Mukherjee
Publication: ACS Applied Bio Materials
Publisher: American Chemical Society
Date: Aug 1, 2023

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms and Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from {COMPLETE REFERENCE CITATION}. Copyright {YEAR} American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your RightsLink request. No additional uses are granted (such as derivative works or other editions). For any uses, please submit a new request.

If credit is given to another source for the material you requested from RightsLink, permission must be obtained from that source.

BACK CLOSE WINDOW

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

Dec 26, 2023

This Agreement between Mr. Amal Vijay ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	5696370896515
License date	Dec 26, 2023
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Chemistry - An Asian Journal
Licensed Content Title	Atomistic De-novo Inhibitor Generation-Guided Drug Repurposing for SARS-CoV-2 Spike Protein with Free-Energy Validation by Well-Tempered Metadynamics
Licensed Content Author	Arnab Mukherjee, Reman K. Singh, Amal Vijay, et al
Licensed Content Date	May 18, 2021
Licensed Content Volume	16
Licensed Content Issue	12
Licensed Content Pages	9
Type of use	Dissertation/Thesis

Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Title of new work	Exploring molecular recognition of small molecule binding, protein-protein interactions, and RNA folding dynamics using enhanced sampling methods
Institution name	IISER Pune
Expected presentation date	Jan 2024
Requestor Location	Mr. Amal Vijay Main Building IISER Pune Pashan Pune, 411008 India Attn: Mr. Amal Vijay
Publisher Tax ID	EU826007151
Total	0.00 USD

Terms and Conditions

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.

- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order,** is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.**For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts,** You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED

BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.

Further details can be found on Wiley Online Library
http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions:

v1.10 Last updated September 2015

Questions? customercare@copyright.com.
