Modeling and Standard Binding Free Energy Calculations of Complex Biological Objects

Marharyta Blazhynska

UNIVERSITÉ DE LORRAINE

École doctorale C2MP

Modeling and Standard Binding Free Energy Calculations of Complex Biological Objects

Thèse

présentée et soutenue publiquement le 16 Octobre 2023

pour l'obtention du

Doctorat de l'Université de Lorraine

(Docteur en Chimie)

par

Marharyta Blazhynska

Composition du jury

Présidente :	Pr. Elise Dumont, Université Côte d'Azur, Nice, France
Rapporteurs :	Pr. Nathalie Reuter, University of Bergen, Norvège Pr. Antonio Monari, ITODYS, Université Paris Cité, France
Examinateurs :	Dr. Hervé Minoux, Sanofi, France Pr. Elise Dumont, Université Côte d'Azur, Nice, France
Encadrants :	Dr. Christophe Chipot, DR-CNRS, HDR, Université de Lorraine Dr. François Dehez, DR-CNRS, HDR, Université de Lorraine

Laboratoire de Physique et Chimie Théoriques - UMR CNRS 7019

"Everything can be taken from a man but one thing: the last of the human freedoms—to choose one's attitude in any given set of circumstances, to choose one's own way."

— Viktor E. Frankl, Man's Search for Meaning

Acknowledgments

I wish to convey my profound appreciation to those who have provided unwavering support and guidance during my doctoral journey. Their invaluable contributions have been instrumental in shaping the final outcome of this thesis, and for that, I am profoundly thankful.

First and foremost, I extend my heartfelt gratitude to my supervisor, Christophe Chipot, and my co-supervisor, François Dehez. Their expertise, guidance, and constant encouragement have been instrumental in my academic and personal growth. I am sincerely thankful for the stimulating research environment they fostered, allowing me to flourish and explore new frontiers of knowledge.

I would also like to sincerely thank the director of the lab, Dragi Karevski, for graciously accepting me here. Additionally, I am grateful to Séverine Bonenberger, the administrative officer of the lab, for her efficiency, professionalism, and willingness to assist with administrative matters associated with my research and studies. Furthermore, I extend my deepest gratitude to Christine Sartori, the pedagogical manager of the Doctoral School. Her exceptional humanity, constant support, and genuine care, which provided me with strength and solace during the challenging years of the war in my homeland, Ukraine, have left an indelible impact on me.

I am indebted to my colleagues, Emma Goulard Coderc de Lacam and Haochuan Chen, for their invaluable companionship and collaborative spirit. Together, we

formed a cohesive team, and I am grateful for their support, intellectual discussions, and camaraderie. Our collective effort and shared enthusiasm have made this journey all the more fulfilling.

Furthermore, I would like to extend my gratitude to the esteemed researchers who have provided invaluable guidance and support throughout my study. I am deeply appreciative of the mentorship and encouragement I received from Benoît Roux, J.C. Gumbart, Jérôme Hénin, and Giacomo Fiorin. Their expertise and willingness to engage in meaningful discussions significantly influenced the development of my work.

I would also like to express my heartfelt appreciation to the group of Emad Tajkhodshid at the Beckman Institute, Urbana, IL. The research stay at their lab was a remarkable experience, and I am grateful for the opportunity to collaborate with such talented colleagues and professors. Special thanks to Alek and Angela Aksiementiev, who provided me with professional support and became dear friends. Our shared Ukrainian background fostered a strong bond, making my time at the institute even more meaningful.

Finally, I would like to express my deepest gratitude to my family members and my partner's family. Their unwavering support, understanding, and sacrifices during challenging times have been crucial in enabling me to pursue this Ph.D. Their belief in me gave me the strength to persevere. I am forever grateful for their boundless love, unwavering encouragement, and for standing by my side every step of the way.

Table of Contents

Publications and Editorial Contributions During Ph.D.

† corresponds to the co-first authorship

List of Publications

- **Blazhynska, M.**, Gumbart, J. C., Chen, H., Tajkhorshid, E., Roux, B., Chipot, C. A rigorous framework for calculating protein-protein binding affinities in a membrane. *J. Chem. Theory Comput.*, **under submission.**

- **Blazhynska, M.**, Goulard Coderc de Lacam, E., Chen, H., Chipot, C. Improving speed and affordability without compromising accuracy: Standard binding free-energy calculations using an enhanced-sampling algorithm, multiple-time stepping, and hydrogen mass repartitioning, *J. Chem. Theory Comput.*, 2023, 19(11), 3091-3101, Doi:10.1021/acs.jctc.3c00141

- Goulard Coderc de Lacam, E., **Blazhynska, M.**,† Chen, H., Gumbart, J.C., Chipot, C. When the dust has settled: Calculation of binding affinities from first principles for SARS-CoV-2 variants with quantitative accuracy, *J. Chem. Theory Comput.*, 2022, 18, 10, 5890–5900, Doi: 10.1021/acs.jctc.2c00604

- **Blazhynska, M.**, Goulard Coderc de Lacam, E., Chen, H., Roux, B. Chipot, C. Hazardous shortcuts in standard binding–free energy calcula-

tions, *J. Phys. Chem. Lett.*, 2022, 13, 27, 6250–6258, Doi: 10.1021/acs.jpclett.2c01490

- Fu, H., Chen, H., **Blazhynska, M.**, Goulard Coderc de Lacam, E., Szczepaniak, F., Pavlova, A., Shao, X., Gumbart, J. C., Dehez, F., Roux, B., Cai, W., & Chipot, C. Accurate determination of protein:ligand standard binding free energies from molecular dynamics simulations., *Nat. Protoc.*, 2022, 17, 1114–1141, Doi.org/10.1038/s41596-021-00676-1.

Public Tutorial Editor Work

- Gumbart, J., Hénin, J., Chipot, C. (2023). **Editor: Blazhynska, M.** In silico alchemy: A tutorial for alchemical free-energy perturbation calculations with NAMD. Retrieved from www.ks.uiuc.edu/Training/Tutorials/

- Hénin, J., Gumbart, J., Chipot, C. (2023). **Editor: Blazhynska, M.** Free energy calculations along a reaction coordinate: A tutorial for adaptive biasing force simulations. Retrieved from www.ks.uiuc.edu/Training/Tutorials/

Table of Abbreviations

ACF Auto-correlation function

CHARMM Chemistry at Harvard molecular mechanics

COM Center of mass

Cryo-EM Cryogenic electron microscopy

CV Collective variable

FF Force field

FEP Free energy perturbation

GPU Graphics processing unit

GUI Graphical user interface

HMR Hydrogen-mass repartitioning

ITC Isothermal titration calorimetry

MD Molecular dynamics

MTS Multi-time stepping

NAMD Not another molecular dynamics [program], University of Illinois

PBC Periodic boundary conditions

PDF Probability distribution function

PME Particle-mesh Ewald

PMF Potential mean force

RDF Radial distribution function

SG Slow growth

SPR Surface plasmon resonance

TI Thermodynamic integration

VMD Visual molecular dynamics, University of Illinois

WTM-eABF Well-tempered metadynamics extended adaptive biasing force

Introduction (English version)

The relentless interest in the intricate dynamics of biomolecular interactions stems from its pivotal role in governing essential processes at the atomic level. These processes include but are not limited to recognition-association phenomena, enzyme-substrate binding, signal transduction, and drug-target interaction.[1–7] Gaining a comprehensive understanding of how proteins establish associations with ligands or other proteins is of utmost significance in multiple research domains, spanning from pharmaceutical sciences to protein engineering.[1,2,8] As a key metric, the standard binding free energy is utilized to measure the strength of the interactions, thus quantifying the reversible association between molecular components.[2,9–13] Remarkably, the theory underlying free energy calculations predates the era of powerful computers capable of conducting large-scale simulations.[2] Pioneering researchers, including Lev Landau,[14] John Kirkwood,[15,16] Robert Zwanzig,[17] William Jorgensen,[18–20] and Peter Kollman[21,22] laid the foundation for the field and set the stage for the developing of the theoretical framework for free energy calculations, which paved the way for the many subsequent advances in the field.[2,23–29]

While experimental techniques, such as isothermal titration calorimetry (ITC)[30,31] or surface plasmon resonance (SPR),[32] are primarily successful in binding affinity determination of both protein-protein and protein-ligand complexes, their application is often costly in terms of time and money spent on synthesis and purification.[2,33,34] To address this challenge, computational methods have been continuously developed to accurately estimate binding free energies through *in*

silico approaches.[2,9–13,35,36] The realm of molecular dynamics (MD) simulations emerged as a groundbreaking tool in this quest, with Berni Julian Alder's pioneering work dating back to 1957.[37] His first "computer experiment" based on the application of Newton's equations of motion allowed him to study phase transitions in liquids, marking the inception of MD. Subsequently, James Andrew McCammon and Martin Karplus advanced the field by designing an MD code and applying it to the bovine pancreatic trypsin inhibitor complex.[38] However, capturing more intricate interactions, such as rare events in complex molecular phenomena, posed a formidable challenge, given the vast timescales involved. In recent years, advancements in computing power coupled with the development of specialized hardware such as graphics processing units (GPUs), have propelled the MD simulations.[36,39] This progress has also played a significant role in elevating approaches for free energy calculation.[8,12,29] During my Ph.D., I had the opportunity to use the advancements in computing power and leverage two MD-based techniques in my research of standard binding free energy calculations, namely, alchemical and geometrical routes, the details of which are discussed in the upcoming chapters.[12,29] By implementing these strategies for specific study cases, I was able to investigate the intricate dynamics of biomolecules at the atomic level and obtain precise estimations of binding free energies.

In this context, my thesis endeavors to contribute to the continuously advancing field of molecular interactions by offering insights into the dynamic behavior of biomolecular complexes and providing accurate and reliable estimations of their binding affinities. Specifically, my work centers on refining and enhanc-

ing the MD-based methodology used for binding free-energy calculations for a range of complexes, including those involving protein-ligand, protein-protein, and transmembrane complexes. Additionally, my research aimed to gain valuable insights into the underlying mechanisms governing these interactions. The ultimate goal of my Ph.D. was to empower researchers by equipping them with the robust methodology to not only compute binding affinities with remarkable accuracy and precision but also to provide them with a profound understanding of the molecular mechanisms at play.

The subsequent chapters cover a range of topics to explore the methodology of the alchemical and geometrical routes and outcomes of my research:

- Chapter 1 serves as a comprehensive survey of the different methodologies presented in the literature to obtain precise standard binding free energy calculations via MD simulations. Additionally, it introduces a succinct yet in-depth discussion of the underlying theoretical principles of the binding affinity approaches used in this study (alchemical and geometrical routes).

- Chapter 2 focuses on providing an overview of the enhanced-sampling algorithms available in the literature. After an extensive evaluation of the strengths and limitations of various enhanced-sampling algorithms, the well-tempered extended metadynamics adaptive biasing force (WTM-eABF) algorithm was chosen for its efficient sampling of the relevant regions in the configuration space and providing accurate standard binding free energies of protein-ligand and protein-protein complexes. A brief overview of the the-

oretical principles underlying this algorithm is also provided to establish a strong theoretical foundation for its use in the present study.

- In Chapter 3, a comprehensive user-friendly software, the Binding Free Energy Estimator 2 (BFEE2) is introduced for automated protein-ligand standard binding free energy calculations, following alchemical and geometrical routes. Specifically, BFEE2 assists in preparing all the necessary input files and performs a post-treatment evaluation toward a final binding affinity value. The chapter provides a comprehensive description of a protocol leveraging BFEE2 with practical application and demonstrates its effectiveness in various scenarios, some of which are presented in the manuscript.

- Chapter 4 elucidates the methodological expansion of the geometrical route for determining protein-protein binding affinity on the straightforward study case of a pig insulin dimer. Its molecular assembly is comprehensively explicated, along with a detailed account of the computational procedures employed in the study.

- Chapter 5 presents a methodological study on the importance of considering restraints applied on collective variables (CVs) in standard binding free energy calculations of protein-ligand and protein-protein complexes. The outcomes of the rigorous geometrical route were compared with those of its shortcut on the basis of two protein-ligand and two protein-protein complexes.

- Following the successful application of the geometrical route for standard

binding free-energy calculations for protein-ligand and protein-protein complexes, Chapter 6 focuses on the binding affinity calculations of SARS-CoV-2 complexes, driven by the urgent need to find effective therapies and develop a better understanding of the molecular interactions involved to fight the global COVID-19 syndemic. Additionally, in this work, the remarkable predictive power of the geometrical route was emphasized.

- In Chapter 7, I extend the application of the geometrical route to explore biological complexes within a membrane-like environment. This chapter focuses on investigating the binding mechanism of a membrane protein-protein complex of glycophorin A (GpA) within a lipid bilayer. To overcome the special challenges posed by the membrane environment and the structural specificity of the complex, the selection of CVs and the evaluation of energetic contributions are approached in a manner distinct from the previous protein-protein complex studies conducted in a water medium. This shift in methodology requires a distinct analytical perspective to accurately assess the binding affinities in membranes.

- Chapter 8 focuses on the optimization of the computational efficiency of standard binding free energy calculations while retaining exceptional accuracy. Applying two techniques, a multiple-time-stepping (MTS) and a hydrogen-mass repartitioning (HMR) trick, and by fine-tuning the extended Lagrangian parameters of WTM-eABF algorithm, the simulations of two complex protein-ligand systems were accelerated by a factor of three without any significant accuracy loss.

Introduction (Version française)

L'intérêt incessant pour la dynamique complexe des interactions biomoléculaires trouve son origine dans son rôle central dans la régulation des processus essentiels au niveau atomique. Ces processus comprennent, sans s'y limiter, les phénomènes de reconnaissance-association, la liaison enzyme-substrat, la transduction de signal et l'interaction médicament-cible.[1–7] Comprendre comment les protéines établissent des associations durables avec des ligands ou d'autres protéines est d'une importance capitale dans de nombreux domaines de recherche, allant des sciences pharmaceutiques à l'ingénierie des protéines.[1,2,8] L'énergie libre de liaison absolue est utilisée pour mesurer la force de ces interactions, servant d'outil pratique pour quantifier l'association réversible entre les composants moléculaires.[2,9–13] La théorie qui sous-tend les calculs d'énergie libre précède l'ère des ordinateurs puissants capables d'effectuer des simulations à grande échelle.[2] Des chercheurs pionniers, dont Lev Landau,[14] John Kirkwood,[15,16] Robert Zwanzig,[17] William Jorgensen,[18–20] et Peter Kollman[21,22] ont établis les bases du domaine et préparé le terrain pour le développement du cadre théorique des calculs de l'énergie libre, ce qui a ouvert la voie aux nombreuses avancées ultérieures dans le domaine.[2,23–29]

Les techniques expérimentales telles que la calorimétrie de titrage isotherme (ITC) ou la résonance plasmonique de surface (SPR), sont principalement utilisées pour déterminer l'affinité de liaison des complexes protéine-protéine et protéine-ligand. Cependant leur application est souvent coûteuse en termes de temps

et d'argent liés aux difficultés de synthèse et la purification des composés.[2,33,34] Pour relever ce défi, des méthodes de calcul ont été continuellement développées pour estimer avec précision les énergies libres de liaison grâce à des approches *in silico*.[2,9–13,35,36] Le domaine des simulations de dynamique moléculaire (MD) a émergé comme un outil révolutionnaire dans cette quête, avec les travaux pionniers de Berni Julian Alder datant de 1957.[37] Sa première *"computer experiment"*, basée sur l'application des équations de mouvement de Newton, lui a permis d'étudier les transitions de phase dans les liquides, marquant le début de la MD . Par la suite, James Andrew McCammon et Martin Karplus ont fait progresser le domaine en concevant un code de MD et en l'appliquant au complexe inhibiteur de la trypsine pancréatique bovine.[38] Cependant, la capture d'interactions plus complexes, telles que les événements rares dans des phénomènes moléculaires complexes, a posé un défi redoutable en raison des vastes échelles de temps impliquées. Au cours des dernières années, les progrès dans la puissance de calcul, conjugués au développement de matériel spécialisé tel que les unités de traitement graphique (GPU), ont propulsé les simulations de MD.[36,39] Les progrès réalisés dans le domaine de la MD ont joué un rôle important dans l'élévation des calculs d'énergie libre au rang de techniques de modélisation fiables et bien établies.[2] Au cours de mon doctorat, j'ai eu l'occasion d'utiliser les progrès de la puissance de calcul et de tirer parti de deux techniques basées sur la MD dans ma recherche de calculs d'énergie libre de liaison absolue, précisément *"alchemical"* et *"geometrical" routes*, dont les détails sont discutés dans les chapitres suivants.[12,29] En mettant en œuvre ces stratégies pour des cas d'étude spécifiques, j'ai pu étudier la dynamique complexe

des biomolécules au niveau atomique et obtenir des estimations précises de ces énergies libres de liaison.

Dans ce contexte, ma thèse s'efforce de contribuer au domaine en constante évolution des interactions moléculaires en offrant des perspectives sur le comportement dynamique des complexes biomoléculaires et en fournissant des estimations précises et fiables de leurs affinités de liaison. Plus précisément, mon travail se concentre sur le raffinement et l'amélioration de la méthodologie basée sur la MD utilisée pour le calcul des énergies libres de liaison pour divers complexes, y compris ceux impliquant des interactions protéine-ligand, protéine-protéine et des complexes transmembranaires. De plus, ma recherche visait à acquérir des informations précieuses sur les mécanismes sous-jacents régissant ces interactions. L'objectif ultime de ma thèse de doctorat était de fournir aux chercheurs une méthodologie robuste leur permettant non seulement de calculer les affinités de liaison avec une précision remarquable, mais également de leur offrir une compréhension profonde des mécanismes moléculaires en jeu.

Les chapitres suivants couvrent une série de sujets détaillant la méthodologie des *alchemical* et *geometrical routes* ainsi que les résultats de mes recherches:

- Le chapitre 1 constitue une présentation complète des différentes méthodologies existantes dans la littérature pour obtenir des calculs précis de l'énergie libre de liaison absolue par le biais de simulations MD. En outre, il inclut une discussion succincte mais approfondie des principes théoriques sous-jacents des approches de l'affinité de liaison utilisées dans cette étude (les *alchemical*

et *geometrical routes*).

- Le chapitre 2 contient une vue d'ensemble des algorithmes d'échantillonnage amélioré disponibles dans la littérature. Après une évaluation approfondie des points forts et des limites des différents algorithmes d'échantillonnage amélioré, l'algorithme WTM-eABF (*well-tempered extended metadynamics adaptive biasing force*) a été choisi pour sa capacité à échantillonner efficacement les régions pertinentes de l'espace de configuration et à calculer avec précision les énergies libres de liaison standard des complexes protéine-ligand et protéine-protéine. Un bref aperçu des principes théoriques qui sous-tendent cet algorithme est également fourni afin d'établir une base théorique solide pour son utilisation dans la présente étude.

- Au chapitre 3, un logiciel complet ergonomique, Binding Free Energy Estimator 2 (BFEE2), est présenté pour effectuer des calculs automatisés de l'énergie libre de liaison protéine-ligand, basé sur les *alchemical* et *geometrical routes*. Plus précisément, BFEE2 aide à préparer tous les fichiers d'entrée nécessaires et effectue une évaluation post-traitement pour obtenir une valeur finale d'affinité de liaison. Le chapitre offre une description complète de l'application pratique d'un protocole pour les calculs standard d'affinité de liaison à l'aide de BFEE2 et démontre son efficacité dans divers scénarios, dont certains sont présentés dans le manuscript.

- Le chapitre 4 dépeint l'expansion méthodologique de la *geometrical route* pour déterminer l'affinité de liaison protéine-protéine sur le cas d'étude simple d'un dimère d'insuline de porc. Son assemblage moléculaire est expliqué

en détail, ainsi que les procédures de calcul employées dans l'étude.

- Le chapitre 5 présente une étude méthodologique sur l'importance de prendre en compte les restrictions appliquées aux variables collectives (CVs) dans les calculs standard de l'énergie libre de liaison des complexes protéine-ligand et protéine-protéine. Les résultats de la *geometrical route* rigoureuse ont été comparés à ceux de son raccourci sur la base de deux complexes protéine-ligand et de deux complexes protéine-protéine.

- Après le succès de l'application de la *geometrical route* pour les calculs d'énergie libre de liaison standard pour les complexes protéine-ligand et protéine-protéine, le chapitre 6 se concentre sur les calculs d'affinité de liaison des complexes du SARS-CoV-2, motivés par le besoin urgent de trouver des thérapies efficaces et de développer une meilleure compréhension des interactions moléculaires impliquées pour lutter contre la syndémie mondiale de COVID-19. De plus, dans cette étude, on a souligné la remarquable puissance prédictive de la *geometrical route*.

- Dans le chapitre 7, j'élargis l'application de la *geometrical route* pour explorer les complexes biologiques dans un environnement de type membranaire. Ce chapitre se concentre sur l'étude du mécanisme de liaison d'un complexe protéine-protéine membranaire,la glycophorine A (GpA), au sein d'une bicouche lipidique. Pour surmonter les défis particuliers posés par l'environnement membranaire et la spécificité structurelle du complexe, la sélection des CVs et l'évaluation des contributions énergétiques sont abordées d'une manière différente des études précédentes sur les complexes protéine-protéine menées

dans un milieu aqueux. Ce changement de méthodologie nécessite une perspective analytique distincte pour évaluer avec précision les affinités de liaison dans les membranes.

- Le chapitre 8 se concentre sur l'optimisation de l'efficacité de calcul de l'énergie libre de liaison standard tout en conservant une précision exceptionnelle. En appliquant deux techniques, *"multiple-time-stepping"* (MTS) et *"hydrogen-mass repartitioning"* (HMR), tout en affinant les paramètres du Lagrangian étendu de l'algorithme WTM-eABF, les simulations de deux systèmes complexes protéine-ligand ont été accélérées d'un facteur de trois sans perte significante de précision.

Chapter 1

Methodology for protein-ligand standard binding free-energy calculations

1.1 Approaches for Estimating Standard Binding Free Energy: Challenges and Advances

With the advent of large-scale MD simulations for studying biological complexes within realistic timescales, the determination of binding free energy has often relied on the use of brute-force MD simulations.[2,36,40] This conventional approach typically involves simulating the separated partners of a complex, anticipating that they will dissociate and reassociate during a finite-length simulation. Using this strategy, Buch et al.[41] employed multiple short and unbiased MD trajectories to statistically reconstruct the binding process of the trypsin-benzamidine enzyme-inhibitor complex by means of a Markov-state model.[42–45] It should be noted that the computational intensity associated with this approach limits its applicability primarily to relatively weak (millimolar) binding interactions, thus limiting its appeal.[13]

Alternatively, the alchemical free energy perturbation (FEP) approach, where

the ligand is gradually removed from its environment via intermediate states, is commonly used for studying ligand binding.[2,17] However, this approach is not suitable for large ligands and cannot be used to treat the binding of two macromolecules due to their size and the extensive configurational rearrangements that occur upon their reversible association, leading to large perturbations, thus, posing challenges in accurate binding affinity estimation.[46] To overcome this limitation, the molecular mechanics/Poisson-Boltzmann and surface area (MM-PBSA) method is often used.[22,47–49] The estimates of the binding free energy of a molecule are obtained as a sum of its gas-phase energy, solvation free energy, and configurational entropy contribution.[47,50] Despite its success, the MM-PBSA method is based on several uncontrolled approximations and has considerable limitations, such as implicit solvation and ignorance of large structural changes upon binding,[51,52] thereby calling into question its significance and general applicability to protein-ligand and protein-protein complexes.[12,53] Another option is steered molecular dynamics (SMD),[54,55] where non-equilibrium pulling simulations are used to separate binding partners along an arbitrary direction.[56–58] The straightforward idea is that the non-equilibrium work to physically separate the complex during a finite-length simulation must reflect the strength of the binding affinity. Theoretically, the binding free energy can be rigorously retrieved by leveraging the Jarzynski identity.[59] However, accurate determination of this quantity requires multiple realizations in a near-equilibrium regime.[60] Furthermore, the convergence of the SMD method is affected by the applied pulling force and may not correspond to the most favorable separation pathway, which can affect the accuracy of

the binding affinity estimation.[61] Consequently, a non-equilibrium SMD strategy requires substantial computational resources to obtain well-converged estimates of the binding free energy.[62]

The accurate estimation of binding free energies using MD simulations poses a significant primary challenge, which stems from the need to precisely capture the intricate changes in configurational entropy that arise during the reversible association of two molecular partners. These changes result from substantial conformational, translational, and orientational movements, underscoring the paramount importance of attaining converged configurational ensemble averages.[12,63] Additionally, MD-based approaches encounter susceptibility to convergence issues due to the partners' size, complexity, and considerable solvation free energy. Consequently, effectively addressing these conceptual and computational issues necessitates being revised from a different perspective.

An effective strategy that ensures the converged configuration ensemble averages, was initially proposed by Hermans and Shankar,[64] and involved two phases. The initial "confine" phase applies a restraining potential to control the movement of the ligand within a specific region of the binding site, whereas the subsequent analytically tractable "release" phase allows the ligand to explore its environment freely, mimicking the ligand's behavior when it dissociates from the binding site.[64–66] Based on a similar idea, Woo and Roux proposed to introduce geometrical restraints[29] acting on a set of collective variables (CVs) to control the ligand's movements.[67] These CVs represent the slow degrees of freedom associated with the relative movements of the two partners. The introduction of the restraints

on the selected CVs permits reducing the configurational space to be sampled and accelerating convergence of the free-energy calculation.[12] The represented restraint-based strategy has served as a direct inspiration to a number of methods, among which the funnel metadynamics,[68] attach-pull-release methods,[69] as well as various restraints-based approaches introduced in alchemical free-energy methodologies.[70–72] From a theoretical standpoint, restraining the ligand means removing certain degrees of freedom and losing orientational, translational, and conformational entropies, contributing to free energy. Hence, the free-energy contribution of these restraints should be evaluated precisely in both states: in the binding pocket and in the bulk.[12,29]

In my work, the deliberate control of sampling was effectuated by utilizing configurational restraints and formulating the calculation in terms of geometric or alchemical transformations, in the so-called "geometrical" and "alchemical" routes.[12,29] Depending on the specific protein-ligand complex case, one has to choose one route or the other. Irrespective of the chosen route, restraints are introduced and rigorously taken into account using a set of template-based and generalized CVs. In the next section of this chapter, I delve into the theoretical background of these two MD-based approaches.

1.2 Theoretical Underpinning for Standard Binding Free Energy Calculations

Theoretical considerations of the binding process require first defining a diluted solution of proteins and ligands that associate in an isotropic environment. Herein, ligands can be considered as substrates, such as drugs, hormones, and analogous compounds. Classically, the association of a ligand to the protein is represented by the equilibrium equation:

$$\text{protein} + \text{ligand} \underset{}{\overset{K_{\text{eq}}}{\rightleftharpoons}} \text{protein:ligand}, \tag{1.1}$$

where K_{eq} is the equilibrium constant of the binding process, defined as:

$$K_{\text{eq}} = \frac{[\text{protein} : \text{ligand}]}{[\text{protein}][\text{ligand}]}, \tag{1.2}$$

with [protein:ligand], [protein], and [ligand] representing the concentration of the protein-ligand complex, protein, and ligand, respectively. Let's assume that the total number of proteins is fixed, $[\text{protein}]_{\text{tot}} = [\text{protein}] + [\text{protein} : \text{ligand}]$. The probability that at least one ligand binds to the protein is equal to p_1, and the probability that none of the ligands binds to the protein is p_0. Thus, $[\text{protein}] = p_0[\text{protein}]_{\text{tot}}$ and $[\text{protein} : \text{ligand}] = p_1[\text{protein}]_{\text{tot}}$, where $[\text{protein}]_{\text{tot}}$ is the total concentration of the protein in the system. By normalization, there can either be zero or one ligand bound to the protein and $p_0 + p_1 = 1$. Therefore, the binding constant can now be expressed as:

$$\begin{aligned} K_{\text{eq}} &= \frac{p_1[\text{protein}]_{\text{tot}}}{[\text{ligand}]\ p_0[\text{protein}]_{\text{tot}}} \\ &= \frac{1}{[\text{ligand}]} \times \frac{p_1}{p_0}. \end{aligned} \tag{1.3}$$

Assuming that the protein concentration is sufficiently low, it is possible to consider a single protein with its center-of-mass held fixed at the origin surrounded by a solution of non-cooperative ligands of the same nature without loss of generality. Then, the logarithm of the ratio p_1/p_0 can be related to the reversible work required to extract one ligand from its bulk environment and to insert it into the binding site of the protein.[12] Therefore, the association constant can be calculated as:[12,73]

$$\begin{aligned} K_{\text{eq}} &= \frac{1}{[\text{ligand}]} \times \frac{N \int_{\text{site}} \mathrm{d}\mathbf{1} \int_{\text{bulk}} \mathrm{d}\mathbf{2} \cdots \int_{\text{bulk}} \mathrm{d}\mathbf{N} \int \mathrm{d}\mathbf{X}\; e^{-\beta U}}{\int_{\text{bulk}} \mathrm{d}\mathbf{1} \int_{\text{bulk}} \mathrm{d}\mathbf{2} \cdots \int_{\text{bulk}} \mathrm{d}\mathbf{N} \int \mathrm{d}\mathbf{X}\; e^{-\beta U}} \\ &= \frac{1}{[\text{ligand}]} \times \frac{N \int_{\text{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{X}\; e^{-\beta U}}{\int_{\text{bulk}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{X}\; e^{-\beta U}}, \end{aligned} \tag{1.4}$$

where U is the total potential energy of the system, $1/\beta = k_{\text{B}}T$ is the Boltzmann constant times temperature, and $\{\mathbf{1}, \mathbf{2}, \cdots \mathbf{N},\ \mathbf{X}\}$ are the degrees of freedom of the N ligands and the remaining molecules (solvent or protein), respectively. The subscripts "site" and "bulk" in the integrals indicate the relevant spatial regions of the configurational space to be included in each integration, representing the bound and unbound states. In eq. (1.4), the ligand molecule has been chosen arbitrarily to occupy the binding site, and the factor N accounts for the fact that any ligand could have been chosen. In the final expression, the integrals over the $(N-1)$ remaining ligands have been omitted for the sake of simplicity, assuming low concentration and absence of ligand-ligand interactions. Since the bulk region

is isotropic and homogeneous, we can assume:[29,71]

$$K_{\text{eq}} = \frac{1}{[\text{ligand}]} \times \frac{N \int_{\text{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta U}}{V_{\text{bulk}} \int_{\text{bulk}} \mathrm{d}\mathbf{1}\; \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X}\; e^{-\beta U}}, \tag{1.5}$$

where $\mathbf{r}_1 \equiv (x_1, y_1, z_1)$ is the position of the center-of-mass of ligand 1 in the 3D bulk region, and $\mathbf{r}_1^* = (x_1^*, y_1^*, z_1^*)$ is some arbitrary (fixed) location in the 3D bulk region, far away from the protein, and V_{bulk} is a volume of the bulk region. Since $[\text{ligand}] = N/V_{\text{bulk}}$, the equilibrium binding constant K_{eq} can be written as:

$$K_{\text{eq}} = \frac{\int_{\text{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta U}}{\int_{\text{bulk}} \mathrm{d}\mathbf{1}\; \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X}\; e^{-\beta U}}. \tag{1.6}$$

The denominator and the numerator of eq. (1.6) each represent the initial and final states of the binding process: the ligand bound to the protein and the ligand with its center-of-mass at $\mathbf{r}_1^*$ in the bulk, respectively. It should be noted that all coordinates are expressed relative to the center of mass of the protein.

However, the evaluation by means of MD simulations of the individual configurational integrals like eq. (1.6) is extremely challenging due to a number of reasons, such as:[2]

- **Sampling and Convergence:** MD simulations entail exploring a wide range of possible configurations of a complex upon binding, which involves sampling the different arrangements it can adopt. To obtain accurate estimates of the equilibrium constant, it is important to achieve convergence and obtain sufficient statistics.

- **System Size and Complexity:** Large systems with intricate interactions require a lot of computational resources and can increase the computational cost of the simulations.

- **Timescale Limitations:** Ligand-binding events may occur on timescales that are beyond the reach of conventional MD simulations.

Depending on the binding mode of the ligand, two general approaches to the calculation of the binding free energy, which consists in inserting intermediate states, in eq. (1.6), can be introduced:[12,17]

1. Geometrical Route[24,26] (based on the PMF simulations)
2. Alchemical Route[14,17] (based on the FEP and thermodynamic integration (TI) simulations)

The corresponding energetic contributions of these states can be determined in separate simulations and then combined to obtain the final binding free energy. Irrespective of the chosen route, the intermediate states are constructed by introducing different restraining potentials on the CVs, that are designed to bias the protein-ligand complex toward the configuration that adopts the bound state. As it was mentioned previously, applying restraints on the set of CVs aims to reduce the configurational space and expedite the convergence of free-energy calculations, facilitating more accurate predictions of protein-ligand binding affinities.[12]

At this end, it is important to establish a local frame of the reference in which the position of the center-of-mass of the ligand relative to the receptor $\mathbf{r}_1$ can be

specified by (r_1, θ_1, ϕ_1) in spherical coordinates, and its orientation can be specified from the three Euler angles $(\Theta_1, \Phi_1, \Psi_1)$ (see Figure 1.1 and Table 1.1). We introduce the "axis" potential $u_{\mathrm{a}}(\theta_1, \phi_1)$, designed to restrain the ligand position along a specific axis as in the bound complex. To restrain the ligand orientation as in the bound complex, we also introduce the potential $u_{\mathrm{o}}(\Theta_1, \Phi_1, \Psi_1)$. The detailed investigation reporting the need for restraining potentials applied on the selected CVs for accurate and efficient estimation of standard binding free energy is presented in Chapter 4.

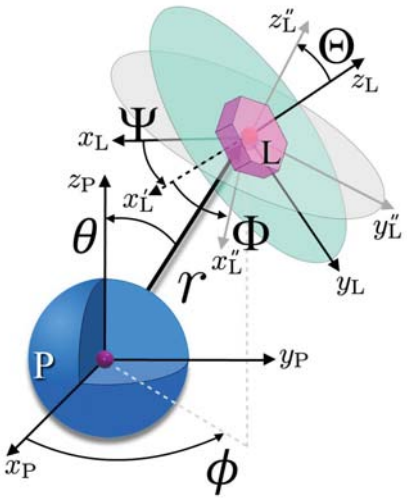

Figure 1.1: Schematic representation of the reference coordinates used to define the orientational and positional restraints, where P and L correspond to protein and ligand, respectively. The P-L center-of-mass distance is represented as r. θ and ϕ relate to the position of L with respect to P. The Euler angles (roll angle Θ, pitch angle Φ, and yaw angle Ψ) determine the relative orientation from the bound state.[74] Reproduced with permission from *J. Phys. Chem. Lett.*, **2022**, 13, 27, 6250–6258, Copyright 2022 American Chemical Society.

Ultimately, it is vital to consider the specific attributes of the protein-ligand complexes when deliberating between the alchemical and geometrical routes. In cases where the ligand is deeply embedded within the protein's binding site, a so-called "buried ligand complexes", the alchemical route emerges as a promising

Table 1.1: Collective variables used in the binding free-energy calculations.

CVs	Description (the ligand with respect to the protein)	Ligand movement mode
RMSD	RMSD on the ligand heavy atoms concerning its bound-state conformation	Conformational
Euler Θ	Roll angle from the bound state orientation	Orientational
Euler Φ	Pitch angle from the bound state orientation	Orientational
Euler Ψ	Yaw angle from the bound state orientation	Orientational
Polar θ	Polar angle in spherical coordinates	Positional
Polar ϕ	Azimuthal angle in spherical coordinates	Positional
r	Center-of-mass distance	Positional

avenue for exploration.[12,75] This methodology facilitates the transformation of the ligand from its bound state to an annihilated state while preserving its structural integrity, thereby enabling a more precise depiction of the intricate interplays and solvation phenomena occurring within the secluded domain. Conversely, the geometrical route, which entails physically displacing the buried ligand from its binding site, presents challenges that may render its implementation impracticable. When the ligand resides in a region that is entirely impermeable to solvent exposure, endeavors to relocate it could engender profound structural rearrangements or even instigate complete dissociation of the complex. In contrast, when the ligand is semi-buried, it means that part of the ligand is exposed to the solvent while the other part is buried within the protein binding site, both routes can be considered for binding free energy calculations, yielding identical binding affinities.[12,75] In certain cases where the ligand is located at the surface of the protein, the geometrical route is often a more practical and feasible approach for studying the binding free energy of the complex, while the alchemical route may pose difficulties due to the complex network of interactions involved. In the following

subsections, each of the routes will be described in greater detail.

1.2.1 Alchemical Route

The alchemical route follows the thermodynamic cycle depicted in Figure 1.2.[76] One of the main challenges that arise during alchemical transformations is known as the "wandering ligand problem," which was first identified by Jan Hermans.[77] This problem arises due to the progressive reduction of interactions between the ligand and the protein during the decoupling phase, allowing the ligand to move at a considerable distance from its original binding site, thereby violating the principle of microreversibility.[12,77] In order to preserve the thermodynamic microreversibility of the binding process, precise geometric restraints are imposed on the ligand within the binding pocket (represented as **ligand*** in Figure 1.2). These restraints, described in detail in Table 1.1, ensure the conservation of the initial conformation, position, and orientation of the ligand with respect to the binding site.[12,17,39]

To compute the free energy associated with the reversible decoupling of the ligand from its surrounding environment, the FEP methodology was employed, in the presence of the restraints mentioned above (see Table 1.1).[12,75] To evaluate the energetic cost of those restraints, TI was employed. In the following subsections, the intricacies of TI and FEP methodologies are presented to shed light on their underlying concepts for the alchemical route.

The equilibrium binding constant in the alchemical route, $K_{\text{eq}}^{\text{AR}}$, is presented in eq. 1.6, can be written as:

$$
\begin{aligned}
K_{\text{eq}}^{\text{AR}} \;=\; & \frac{\int_{\text{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta U_1}}{\int_{\text{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U_1+u_{\text{c}}]}} \\
& \times \frac{\int_{\text{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U_1+u_{\text{c}}]}}{\int_{\text{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U_1+u_{\text{c}}+u_{\text{o}}]}} \\
& \times \frac{\int_{\text{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U_1+u_{\text{c}}+u_{\text{o}}]}}{\int_{\text{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U_1+u_{\text{c}}+u_{\text{o}}+u_{\text{a}}]}} \\
& \times \frac{\int_{\text{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U_1+u_{\text{c}}+u_{\text{o}}+u_{\text{a}}]}}{\int_{\text{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U_1+u_{\text{c}}+u_{\text{o}}+u_{\text{a}}+u_{\text{r}}]}} \\
& \times \frac{\int_{\text{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U_1+u_{\text{c}}+u_{\text{o}}+u_{\text{a}}+u_{\text{r}}]}}{\int_{\text{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U_0+u_{\text{c}}+u_{\text{o}}+u_{\text{a}}+u_{\text{r}}]}} \\
& \times \frac{\int_{\text{bulk}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U_0+u_{\text{c}}+u_{\text{o}}+u_{\text{a}}+u_{\text{r}}]}}{\int_{\text{bulk}} \mathrm{d}\mathbf{1}\; \delta(\mathbf{r}_1-\mathbf{r}_1^*) \int d\mathbf{X}\; e^{-\beta[U_0+u_{\text{c}}+u_{\text{o}}]}} \\
& \times \frac{\int_{\text{bulk}} \mathrm{d}\mathbf{1}\; \delta(\mathbf{r}_1-\mathbf{r}_1^*) \int d\mathbf{X}\; e^{-\beta[U_0+u_{\text{c}}+u_{\text{o}}]}}{\int_{\text{bulk}} \mathrm{d}\mathbf{1}\; \delta(\mathbf{r}_1-\mathbf{r}_1^*) \int d\mathbf{X}\; e^{-\beta[U_0+u_{\text{c}}]}} \\
& \times \frac{\int_{\text{bulk}} \mathrm{d}\mathbf{1}\; \delta(\mathbf{r}_1-\mathbf{r}_1^*) \int d\mathbf{X}\; e^{-\beta[U_0+u_{\text{c}}]}}{\int_{\text{bulk}} \mathrm{d}\mathbf{1}\; \delta(\mathbf{r}_1-\mathbf{r}_1^*) \int d\mathbf{X}\; e^{-\beta[U_1+u_{\text{c}}]}} \\
& \times \frac{\int_{\text{bulk}} \mathrm{d}\mathbf{1}\; \delta(\mathbf{r}_1-\mathbf{r}_1^*) \int d\mathbf{X}\; e^{-\beta[U_1+u_{\text{c}}]}}{\int_{\text{bulk}} \mathrm{d}\mathbf{1}\; \delta(\mathbf{r}_1-\mathbf{r}_1^*) \int d\mathbf{X}\; e^{-\beta U_1}}
\end{aligned}
\tag{1.7}
$$

The potential energy U_0 characterizes the non-interacting state of the lig-

and, while U_1 represents the state in which it is coupled to the environment. The equation mentioned above precisely follows the various steps described in the thermodynamic cycle depicted in Figure 1.2. The first two contributions arise from the conformational, u_{c}, and orientational, u_{o}, restraints imposed on the ligand. The third contribution corresponds to the polar angles term, u_{a}, which represents the positional restraints. The fourth contribution corresponds to the translational term, u_{r}, of the positional restraints. The fifth and eighth contributions, highlighted in magenta, correspond to the alchemical transformations in which the ligand is reversibly decoupled from its environment, respectively, in the bound and unbound states. The sixth and seventh contributions, highlighted in cyan, are analytical ones and account for the reorientation and translation of a rigid body in a homogeneous bulk liquid. The final contribution represents the conformational changes of the ligand in the free, unbound state.[12,78] One may note that the δ function involving $\mathbf{r}_1^*$, when it appears both in the numerator and denominator, does not affect the calculated free energies in the bulk region because it is invariant to translations. In the simplified form, the binding constant can be expressed as follows:[12]

$$K_{\mathrm{eq}}^{\mathrm{AR}} = e^{-\beta(\Delta G_{\mathrm{c}}^{\mathrm{bulk}} - \Delta G_{\mathrm{couple}}^{\mathrm{bulk}} + \Delta G_{\mathrm{o}}^{\mathrm{bulk}} + \Delta G_{\mathrm{r}}^{\mathrm{bulk}} + \Delta G_{\mathrm{a}}^{\mathrm{bulk}} + \Delta G_{\mathrm{decouple}}^{\mathrm{site}} - \Delta G_{\mathrm{a}}^{\mathrm{site}} - \Delta G_{\mathrm{r}}^{\mathrm{site}} - \Delta G_{\mathrm{o}}^{\mathrm{site}} - \Delta G_{\mathrm{c}}^{\mathrm{site}})}, \tag{1.8}$$

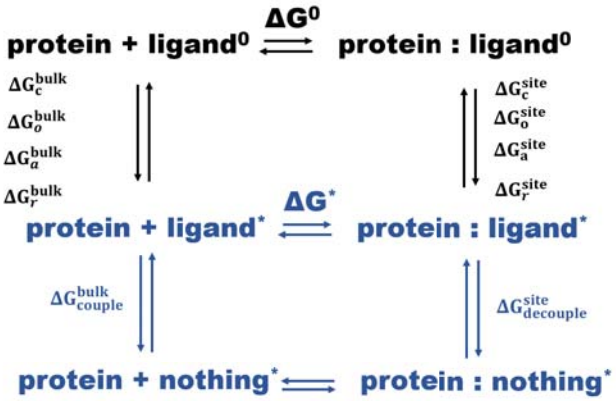

Figure 1.2: The thermodynamic cycle describing the reversible association of a ligand to a protein and the necessary steps to determine the corresponding standard binding free energy, ΔG°. The **ligand**^{0} symbolizes an unrestrained ligand, whereas **ligand*** refers to a ligand restrained in the native conformation, position, and orientation in the protein-ligand complex. The disappearing ligand in the bulk (**nothing***)[12,76] Herein, the energetic contributions of the restraints are indicated by the indices "c", "o", and "a", "r", which refer to conformational, orientational, positional, and center-of-mass distance restraints, respectively, and are evaluated within TI.[15,79] The "couple/decouple" indexes represent the alchemical contribution rising from the FEP calculations, namely, the free energy change of decoupling the ligand from the binding site ($\Delta G^{\text{site}}_{\text{decouple}}$) and from its bulk environment ($\Delta G^{\text{bulk}}_{\text{couple}}$).

The free energy terms $\Delta G^{\text{bulk}}_{\text{r}}$ and $\Delta G^{\text{site}}_{\text{r}}$ correspond to the contribution for changing the position of the ligand along the center-of-mass vector, r, in unbound and bound states, respectively. $G^{\text{site}}_{\text{o}}$ and $\Delta G^{\text{site}}_{\text{a}}$ are the sums of the orientational (i.e., $\Delta G^{\text{site}}_{\Theta}$, $\Delta G^{\text{site}}_{\Phi}$, and $\Delta G^{\text{site}}_{\Psi}$), and positional (i.e., $\Delta G^{\text{site}}_{\theta}$, and $\Delta G^{\text{site}}_{\phi}$) angle contributions in the bound state. $\Delta G^{\text{bulk}}_{\text{o}}$ corresponds to the orientational movement of the unbound partner in an isotropic medium.

The resulting binding free energy can be determined as:

$$\Delta G^\circ_{\text{b}} = -\frac{1}{\beta}\ln(K_{\text{eq}}C^\circ) \tag{1.9}$$

Here, C° denotes the standard concentration of 1 M ($C^\circ = 1/1661$ Å^3) in homogeneous environment.[70]

1.2.1.1 Free Energy Perturbation

Perturbation techniques have a rich historical background in statistical mechanics and were one of the first approaches employed to calculate free energy changes in molecular simulations. Born and Kirkwood laid the groundwork for these techniques in the 1920s and 1930s,[15,16,80] while during the 1950s, Zwanzig further advanced it into the concept of the FEP method.[17] Its original purpose was to analytically calculate equations of state using a high-temperature expansion. Over time, this framework was extended to reversible double annihilation, wherein the ligand is decoupled and recoupled in separate simulations to the protein.[2,12]

To further investigate the FEP concept, let us consider a system characterized by degrees of freedom ($\mathbf{x}$ - atomic coordinate of the system), and having initial and final states described by the potential energy functions $U_{\mathrm{a}}(\mathbf{x})$ and $U_{\mathrm{b}}(\mathbf{x})$, respectively.[1,2] The potential energy function for the final state can be expressed as:

$$U_{\mathrm{b}}(\mathbf{x}) = U_{\mathrm{a}}(\mathbf{x}) + \Delta U(\mathbf{x}), \tag{1.10}$$

where $\Delta U(\mathbf{x})$ corresponds to the perturbation between the two states. It is important to note that in the context of these transformations, we adopt the assumption of mass invariance. This assumption empowers us to work exclusively with potential energies, rather than a complete Hamiltonian.[2] Then, the free energy difference between both states could be represented as:[17]

$$\Delta G_{\mathrm{ab}} = -\frac{1}{\beta} \ln\langle e^{-\beta \Delta U(\mathbf{x})} \rangle_{\mathrm{a}}, \tag{1.11}$$

$$\Delta G_{\mathrm{ba}} = \frac{1}{\beta} \ln \langle e^{\beta \Delta U(\mathbf{x})} \rangle_{\mathrm{b}}, \tag{1.12}$$

where $\langle \cdot \rangle_{\mathrm{a}}$ and $\langle \cdot \rangle_{\mathrm{b}}$ denote ensemble averages over the initial and final states, respectively.

To ensure accurate estimation of the free energy change, where ideally $\Delta G_{\mathrm{ba}} \equiv -\Delta G_{\mathrm{ab}}$, the alchemical transformation simulations should be performed in the forward direction (from state A to state B), and the backward direction (from state B to state A).[81,82] To assess the reliability of the results in line with the coherent sampling and convergence of both directions, it is important to check the discrepancies occurring between the two directions within a hysteresis (Figure 1.3A,C).

It should be noted that eq. 1.11 and 1.12 are the fundamental FEP formulas, which indicate that the computation of free energy differences can be achieved solely by sampling equilibrium configurations from the initial state.[17] It is important to note that these equations hold true in the limit of infinite sampling, assuming that the changes between the two states remain sufficiently small.[1,2] In practice, it is rare for single-step transformations between highly disparate states to meet the requirement of small changes.[2] Therefore, in realistic cases, the reaction pathway connecting states A and B needs to be divided into multiple intermediate states, even though these states may not correspond to physical configurations. The concept of incorporating additional states between the initial and final states, known as "overlap sampling", was first introduced by Bennett.[81] By

incorporating additional intermediate states between the initial and final states, the energy landscape can be more finely sampled, allowing for a smoother transition and better exploration of the perturbation space. A crucial requirement for successful overlap sampling is the overlap between ensembles of these intermediate states.[83,84] To achieve this, the energy is formulated as a function of a coupling parameter, commonly denoted as λ, which governs the transformation between states. Through gradual variation of λ, the system can smoothly traverse from one state to another, enabling accurate estimation of the free energy difference between them.[15] The interval that separates the intermediate states (k) along the transformation between the reference and target systems, corresponding to specific fixed values of λ, is commonly referred to as a "window".[2] The total free energy change over the entire reaction pathway can be expressed as:

$$\Delta G = -\frac{1}{\beta}\sum_{k} \ln\langle e^{-\beta \Delta U_{k,k+1}(\mathbf{x},\lambda)}\rangle_k, \tag{1.13}$$

where $\Delta U_{k,k+1}(\mathbf{x},\lambda)$ corresponds to the perturbation in the energy between adjacent intermediate states k and $k+1$.

Determining the optimal number of windows is a critical consideration in the FEP simulations to achieve reliable results with proper sampling and convergence.[1,2] Insufficient window numbers may result in inadequate exploration of specific energy regions, introducing bias and inaccuracies in the estimation. Conversely, an excessive number of windows can lead to a significant increase in computational costs without substantial improvements in accuracy. Selecting the appropriate number of windows requires careful evaluation of several factors, including the complexity of the transformation, the magnitude of the perturbation,

and the available computational resources.[2,8]

Considering the convergence of the FEP simulations, it can be checked with the probability distribution functions (PDFs) (Figure 1.3B,D). If the PDFs of the forward and backward states of the same window do not completely overlap and exhibit noticeable differences, it indicates a significant variance or mismatch between those states, suggesting that the perturbations occurring in this particular energy region are relatively high.[1] To address this issue and improve convergence, the introduction of extra intermediate states becomes necessary.[81]

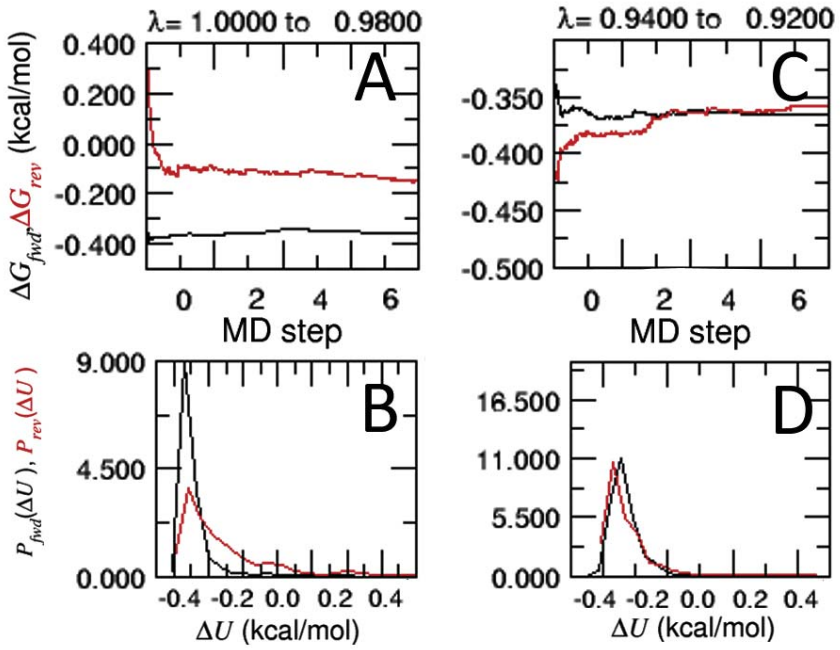

Figure 1.3: Examples of hysteresis and the PDFs necessary to check the accuracy of the calculations following the FEP approach. In this particular case, the $\Delta\lambda$ was equal to 0.02, making in total the use of 50 windows. The graphs are generated automatically using ParseFEP[35,76,81,82] utility included in the VMD package.[85] (A and B) graph corresponds to the hysteresis and the probability function, where there are some discrepancies between the forward and backward estimates, (C and D) signifies reliable and consistent results.

In the FEP calculations, it is also essential to account for the interaction of the ligand with the solvent environment, particularly when the ligand reappears

in regions already occupied by solvent molecules, which poses a challenge known as an "end-point catastrophe".[86–88] This issue is characterized by a substantial increase in the free energy due to collisions between the ligand and solvent. To overcome this issue, the utilization of a soft-core potential has been established as a viable solution,[86,87] enabling a smooth transition between van der Waals (vdW) and electrostatic interactions, thus preventing the occurrence of clashes between molecules. It ensures that two nuclei do not occupy the same spatial region by gradually modulating the strength of intermolecular forces. However, when incorporating the gradual appearance or disappearance of electrostatic interactions using the soft-core potential, it is crucial to ensure that the vdW interactions are appropriately managed to prevent unfavorable clashes or distortions in the molecular system. Overall, in the FEP calculations, it is imperative to establish a robust framework that accounts for the above–mentioned issues to accurately estimate the free energy change.

1.2.1.2 Thermodynamic Integration

The TI is a powerful method that enables the calculation of free energy differences by integrating the derivative of the free energy with respect to a continuous parameter, λ:[2,23,89]

$$\Delta G = \int_0^1 \frac{\partial G_\lambda}{\partial \lambda} \mathrm{d}\lambda = \int_0^1 \left\langle \frac{\partial U(\mathbf{x}, \lambda)}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda \tag{1.14}$$

Unlike FEP, which uses the logarithm of the average of an exponential function (eq. 1.13), the free energy difference is evaluated from the ensemble average directly.[89] In practice, the continuous integral (eq. 1.14) is approximated using techniques

such as the trapezoidal rule, the accuracy of which depends on the smoothness of the integrand and the number of evaluation points.[23] Furthermore, TI offers flexibility in selecting the coupling parameter, λ, allowing for fine-grained control over the transformation between states.[2,89]

In the context of the introduction of restraints on the selected set of CVs in the alchemical route, TI is employed to estimate the energetic cost associated with enforcing these restraints, ensuring reliable and accurate estimation of free energy differences during the alchemical transformations.[12,15,67] Within the framework of TI, the force constant of the harmonic potentials used to apply the restraints is progressively scaled by λ, for both bound and unbound states. In forward simulations, the sequence of discrete λ values evolves linearly from 0 to 1, while in backward simulations, it evolves from 1 to 0.[2] The scaling of the force constant follows a power-law relationship, given by:

$$k_\lambda = k_0 + \lambda^\alpha (k_1 - k_0), \tag{1.15}$$

where k_0, k_λ, and k_1 represent the initial, current, and final values of the force constant, respectively. The exponent α is typically chosen to be greater than 1.0, often set to 4.0, which ensures a smooth distribution of the effects of introducing the restraint over time. This non-linear scaling allows for a gradual and controlled introduction of the restraints on the selected CVs throughout the simulation, enabling accurate estimation of the energetic contributions from these restraints.[67] Additionally, at each step of the λ parameter, the system is pre-equilibrated for a substantial number of steps to ensure that it reaches a stable

state. This pre-equilibration period allows the system to relax and attain a proper equilibrium at the specific λ before data collection begins. The duration of the pre-equilibration phase may vary depending on the specific system and simulation conditions. Following the pre-equilibration, an extensive sampling period is carried out at each λ value. The duration of the sampling period is typically chosen to ensure a thorough exploration of the conformational space and accurate sampling of the relevant thermodynamic states.

In some specific protein-ligand complex cases, the classical TI can be extended to the so-called "slow growth" (SG) technique, which assumes that the increment of λ is small enough to ensure a quasi-static evolution of the system by a finite difference:[23]

$$\left\langle \frac{\partial U(\mathbf{x},\lambda)}{\partial \lambda} \right\rangle_{\lambda} \mathrm{d}\lambda \simeq \frac{\partial U(\mathbf{x},\lambda)}{\partial \lambda} \mathrm{d}\lambda = \Delta U \tag{1.16}$$

In contrast to the classical TI, the SG technique omits the step of pre-equilibration at each λ step. Furthermore, the sampling time in the SG technique is intentionally minimized for each step and focuses only on capturing specific regions or states of interest along the alchemical transformation. This approach is especially valuable when studying complexes with dynamic ligands that undergo conformational changes or when examining interactions involving water molecules that are buried within the protein-ligand interface cite (see Chapter 3, Section Practical Examples for the Protocol Performance Analysis, MUP-I:6-hydroxy-6-methyl-3-heptanone complex).[75] It is important to note that the SG technique should be employed with caution, and its applicability should be evaluated based on the specific complex and research objectives. The decision to minimize sampling time

and omit pre-equilibration steps should be guided by a thorough understanding of the dynamics and properties of the studied complex.[2,75]

1.2.2 Geometrical Route

The geometrical route is built around the concept of the potential of mean force (PMF).[29] By employing the PMF within the geometrical route, it becomes possible to gain insights into the thermodynamics of ligand binding, as well as the underlying molecular mechanisms that govern ligand-protein interactions.[12,46] Herein, it is important to emphasize the notion of the PMF in this context. The PMF (technically, the free energy surface[89]) represents the free energy landscape along a specific CV, such as the separation distance between the ligand and the protein. Historically, the PMF between two particles, denoted as $w(r)$, was originally defined based on the radial distribution function (RDF) $g(r)$.[2,89] In homogeneous environments, the RDF could be derived from the probability density of the interparticle distance, $\rho(r)$, divided by a normalization term proportional to r^2.[89]

$$w(r) = -\frac{1}{\beta}\ln g(r) = -\frac{1}{\beta}\ln\frac{\rho(r)}{r^2} + C = A(r) + \frac{2}{\beta}\ln(r) + C \tag{1.17}$$

where $A(r)$ represents the free energy surface along the interparticle distance, and C is an arbitrary fitting constant.[89] The Jacobian term $\frac{2}{\beta}\ln(r)$, accounts for the contribution of the geometric entropy associated with the CV in an isotropic environment.[67]

Table 1.2: Collective variables and their calculation order for standard binding free-energy calculation of protein-ligand complexes via geometrical route.

Step	CVs	Representations	Restraints
1	RMSD bound	$G_{\mathrm{c}}^{\mathrm{site}}$	
2	Θ	$G_{\Theta}^{\mathrm{site}}$	RMSD bound
3	Φ	G_{Φ}^{site}	RMSD bound, Θ
4	Ψ	G_{Ψ}^{site}	RMSD bound, Θ, Φ
5	θ	$G_{\theta}^{\mathrm{site}}$	RMSD bound, Θ, Φ, Ψ
6	ϕ	G_{ϕ}^{site}	RMSD bound, Θ, Φ, Ψ, θ
7	r	$w(r)$	RMSD bound, Θ, Φ, Ψ, ϕ
8	RMSD unbound	$G_{\mathrm{c}}^{\mathrm{bulk}}$	

In the geometrical route, the restraints are introduced gradually in the following order: the ligand conformational flexibility (RMSD), orientational movements (three Euler angles), and positional polar angle movements relative to the protein, the energetic costs of which are estimated using PMFs (Table 1.2). When all of the contributions from the CVs are determined, the reversible separation PMF as a function of the Euclidean distance, $r = \sqrt{x^2 + y^2 + z^2}$, $w(r)$, can be computed while keeping all other relevant conformational, positional, and orientational components at their equilibrium values through the geometrical restraints by means of suitable harmonic potentials.[12]

Herein, the equilibrium binding constant $K_{\mathrm{eq}}^{\mathrm{GR}}$ in eq. 1.6 can be written as:

$$
\begin{aligned}
K_{\mathrm{eq}}^{\mathrm{GR}} =& \frac{\displaystyle\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta U}}{\displaystyle\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U+u_{\mathrm{c}}]}} \\
&\times \frac{\displaystyle\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U+u_{\mathrm{c}}]}}{\displaystyle\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U+u_{\mathrm{c}}+u_{\mathrm{o}}]}} \\
&\times \frac{\displaystyle\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U+u_{\mathrm{c}}+u_{\mathrm{o}}]}}{\displaystyle\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U+u_{\mathrm{c}}+u_{\mathrm{o}}+u_{\mathrm{a}}]}} \\
&\times \frac{\displaystyle\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U+u_{\mathrm{c}}+u_{\mathrm{o}}+u_{\mathrm{a}}]}}{\displaystyle\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\; \delta(\mathbf{r}_1-\mathbf{r}_1^*) \int d\mathbf{X}\; e^{-\beta[U+u_{\mathrm{c}}+u_{\mathrm{o}}]}} \\
&\times \frac{\displaystyle\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\; \delta(\mathbf{r}_1-\mathbf{r}_1^*) \int d\mathbf{X}\; e^{-\beta[U+u_{\mathrm{c}}+u_{\mathrm{o}}]}}{\displaystyle\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\; \delta(\mathbf{r}_1-\mathbf{r}_1^*) \int d\mathbf{X}\; e^{-\beta[U+u_{\mathrm{c}}]}} \\
&\times \frac{\displaystyle\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\; \delta(\mathbf{r}_1-\mathbf{r}_1^*) \int d\mathbf{X}\; e^{-\beta[U+u_{\mathrm{c}}]}}{\displaystyle\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\; \delta(\mathbf{r}_1-\mathbf{r}_1^*) \int d\mathbf{X}\; e^{-\beta U}}
\end{aligned}
\tag{1.18}
$$

Most of the terms in eq.1.18 are dimensionless ratios of configurational integrals corresponding to free energy differences which can be calculated from a standard application of the PMF simulation:

$$
\begin{aligned}
e^{-\beta G_{\mathrm{c}}^{\mathrm{site}}} &= \frac{\displaystyle\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U+u_{\mathrm{c}}]}}{\displaystyle\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta U}} \\
&= \langle e^{-\beta u_{\mathrm{c}}} \rangle_{(\mathrm{site},U)},
\end{aligned}
\tag{1.19a}
$$

$$
\begin{aligned}
e^{-\beta G_{\mathrm{o}}^{\mathrm{site}}} &= \frac{\displaystyle\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U+u_{\mathrm{c}}+u_{\mathrm{o}}]}}{\displaystyle\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U+u_{\mathrm{c}}]}} \\
&= \langle e^{-\beta u_{\mathrm{o}}} \rangle_{(\mathrm{site},U+u_{\mathrm{c}})},
\end{aligned}
\tag{1.19b}
$$

$$e^{-\beta G_{\mathrm{a}}^{\mathrm{site}}} = \frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta[U+u_{\mathrm{c}}+u_{\mathrm{o}}+u_{\mathrm{a}}]}}{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int d\mathbf{X}\; e^{-\beta(U+u_{\mathrm{c}}+u_{\mathrm{o}})}} = \langle e^{-\beta u_{\mathrm{a}}} \rangle_{(\mathrm{site},U+u_{\mathrm{c}}+u_{\mathrm{o}})}, \tag{1.19c}$$

$$e^{-\beta G_{\mathrm{c}}^{\mathrm{bulk}}} = \frac{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\; \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X}\; e^{-\beta[U+u_{\mathrm{c}}]}}{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\; \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X}\; e^{-\beta U}} = \langle e^{-\beta u_{\mathrm{c}}} \rangle_{(\mathrm{bulk},U)}. \tag{1.19d}$$

Due to the isotropy of the environment,[12] the free energy $G_{\mathrm{o}}^{\mathrm{bulk}}$ shown in cyan in eq. 1.20 can be evaluated analytically:

$$e^{-\beta G_{\mathrm{o}}^{\mathrm{bulk}}} = \frac{1}{8\pi^2} \int_0^{\pi} \sin(\Theta)\, \mathrm{d}\Theta \int_0^{2\pi} \mathrm{d}\Phi \int_0^{2\pi} \mathrm{d}\Psi\; e^{-\beta u_{\mathrm{o}}(\Theta,\Phi,\Psi)}$$

The fourth term shown in magenta in eq. 1.18, which involves a ratio of configurational integrals with the bound ligand (numerator) and the ligand held with its center-of-mass at $\mathbf{r}_1^*$ in the bulk by a delta function, requires special attention because it does not correspond to a free energy difference like the other terms. It can be re-expressed as:

$$\frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{X}\; e^{-\beta[U+u_{\mathrm{c}}+u_{\mathrm{o}}+u_{\mathrm{a}}]}}{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\; \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int \mathrm{d}\mathbf{X}\; e^{-\beta[U+u_{\mathrm{c}}+u_{\mathrm{o}}]}} = S^* I^*, \tag{1.20}$$

where S^* is a surface term, which represents the fraction of a sphere of radius r_1^*, centered around the binding site of the protein accessible to its partner in a homogeneous environment:

$$S^* = (r_1^*)^2 \int_0^{\pi} \mathrm{d}\theta\; \sin\theta \int_0^{2\pi} \mathrm{d}\phi\; \mathrm{e}^{-\beta u_{\mathrm{a}}} \tag{1.21}$$

I^* is a one-dimensional integral over r, defined in terms of the separation PMF:

$$I^* = \int_{\text{site}} \mathrm{d}r_1 \, \mathrm{e}^{-\beta[w(r_1)-w(r_1^*)]} \tag{1.22}$$

where $w(r_1)$ corresponds to the actual separation PMF calculated in the presence of the configurational and orientational, and positional restraints u_{c}, u_{o}, and u_{a}. r_1^* is a reference point that denotes the radial radius between the associated and dissociated states. $w(r_1^*)$ is the PMF at the distance r_1^*, where both partners are located sufficiently far away from each other to no longer interact.[12,46] Then, the binding constant in the isotropic environment for protein-ligand complexes following the geometrical route can be expressed as presented below, and the final binding affinity value can be obtained from eq. 1.9.[29]

$$K_{\text{eq}}^{\text{GR}} = S^* I^* e^{-\beta(G_{\text{c}}^{\text{bulk}} - G_{\text{c}}^{\text{site}} + G_{\text{o}}^{\text{bulk}} - G_{\text{o}}^{\text{site}} - G_{\text{a}}^{\text{site}})} \tag{1.23}$$

Chapter 2

Overcoming free energy barriers: WTM-eABF

2.1 Enhanced sampling algorithms for calculating binding free energy

To efficiently sample the relevant regions of the CV space and accurately calculate the free energy of a complex, it is recommended to use an importance-sampling algorithm.[8,90] The umbrella sampling algorithm (US) is commonly utilized as a straightforward sampling method in this regard.[78,91–95] It involves introducing a biasing potential to the potential energy function, which modifies the distribution along a selected CV. This biasing potential restricts the exploration of specific regions in the configuration space that might be challenging to sample otherwise.[90] Obtaining a complete profile involves a series of simulations that explore specific regions of the reaction coordinate, ξ, (referred to as windows), which must overlap with each other.[91] This biasing approach is based on a priori knowledge of the energy landscape and usually is adjusted through trial and error. The necessity of manually providing a bias for sampling each window can prove

to be a significant obstacle, particularly when the energy landscape is rugged or poorly understood. A variation of the US, known as stratification, partially addresses this issue. This strategy involves simulating a large number of windows with a simple harmonic bias imposed on them, using a relatively strong constant, therefore, forcing sampling along the entire ξ.[2,96,97] In total, the US method has certain limitations. Firstly, it can be computationally expensive due to the requirement of running multiple simulations for each window, which may pose practical challenges, especially when studying complex systems with high-dimensional CV spaces. Additionally, the method is susceptible to slow orthogonal degrees of freedom and can encounter practical issues associated with window design, necessitating careful consideration in advance.[2]

In contrast, metadynamics adds a history-dependent bias potential by depositing small Gaussian potentials that prevent the system from getting trapped in the local minima of the free-energy landscape.[98–100] These Gaussian potentials act as repulsive barriers that discourage the system from revisiting previously explored states and guide it towards unexplored regions of the free-energy surface.[68] The main strength of metadynamics is that it can efficiently sample the free-energy landscape of complex systems and allows for the calculation of free-energy surfaces along multiple CVs. However, an important limitation of metadynamics is that the bias potential can introduce artificial fluctuations in the system, which can affect the accuracy of the results.[12] To overcome this issue, the well-tempered version of metadynamics introduces a tempering factor to control the growth of the bias potential, preventing excessive bias towards higher-energy regions.[101] As a result,

well-tempered metadynamics provides a more efficient and accurate estimation of free-energy differences and enhances the sampling of rare events.[101–103]

ABF methods represent another notable category of advanced sampling algorithms. Within this realm, the standard ABF (stABF) technique[24,104,105] is the classical implementation wherein a biasing force is directly applied to the chosen CV in order to compute the free-energy gradient along it.[105] In stABF, the free energy along the reaction coordinate, ξ, can be considered as a potential resulting from the average force that acts in that direction:[8,106]

$$\frac{\partial A}{\partial \xi} = -\left\langle F_\xi \right\rangle_\xi = \left\langle \frac{\partial U(\mathbf{x})}{\partial \xi} \right\rangle_\xi - \left\langle \frac{1}{\beta} \frac{\partial \ln |\mathbf{J}|}{\partial \xi} \right\rangle_\xi \tag{2.1}$$

$$\mathbf{F}_{\text{bias}} = -\left\langle \mathbf{F}_\xi \right\rangle_\xi \nabla_{\mathbf{x}} \xi \tag{2.2}$$

In equation 2.1, the derivative of the free energy A with respect to ξ provides information about the free energy landscape and guides the sampling along ξ. It combines two terms: the ensemble average of the gradient of the potential energy $U(\mathbf{x})$ with respect to ξ, and the ensemble average of the derivative of the logarithmic Jacobian determinant for transformation from generalized to Cartesian coordinates, $|\mathbf{J}|$, with respect to ξ. The $\mathbf{F}_{\text{bias}}$ in equation 2.2 corresponds to the biasing force equal to the ensemble average of the force $\mathbf{F}_\xi$ acting on ξ, multiplied by the gradient of ξ with respect to the system coordinates $\nabla_{\mathbf{x}}\xi$. The benefit of the ABF-based methods is that the $\mathbf{F}_{\text{bias}}$ is adaptively and locally updated, eliminating the need for prior knowledge of the free energy surface, allowing to efficiently explore the free energy landscape without requiring a predefined biasing

potential.[90,106] However, the stABF imposes strict requirements on CVs, including mutual orthogonality and orthogonality, making the sampling less accessible and more prone to the nonlocal effect of the biasing force.[107–109] To overcome this challenge, the extended-Lagrangian variant of ABF (eABF) was devised.[107,110,111] In this approach, the bias potential is applied to a fictitious particle that is harmonically coupled to the CV. This methodology facilitates more effective sampling and escapes the direct influence of the biasing force exerted on the CV coordinate itself.[112–114] By decoupling the biasing force from the CV, eABF mitigates the issue of being ensnared within high free-energy barriers for an extended duration, thereby enabling more efficient exploration of the free-energy landscape.[107,110,111] Moreover, the merits of eABF are further amplified when it is synergistically combined with other sampling techniques, leading to significant improvements in the convergence rate of simulations.[106,109,115,116]

During my Ph.D., for conducting most of the standard binding free energy calculations, an elevated version of the eABF algorithm, known as a well-tempered meta-extended ABF (WTM-eABF) was employed.[106,109,115,116] The theoretical underpinnings of this algorithm are presented in the next section.

2.2 Theoretical underpinnings of WTM-eABF method

The WTM-eABF algorithm combines two well-established techniques: eABF itself and the well-tempered variant of metadynamics, respectively, allowing for improved sampling of the free energy landscape, without the risk of becoming trapped

in local minima, which can hinder accurate standard binding affinity calculations for both protein-ligand and protein-protein complexes (Figure 2.1).[101,103,106,108,109,115–117]

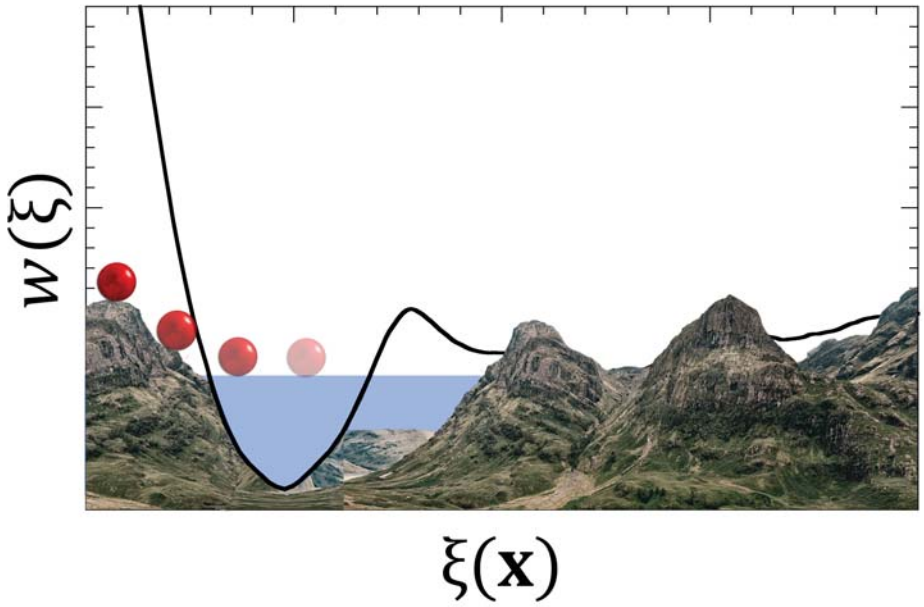

Figure 2.1: Schematic representation of the WTM-eABF algorithm, where eABF shaves the barriers of the free energy landscape (in brown), and well-tempered metadynamics floods its valleys (in blue).[8]

It should be noted that in contrast to the stABF, in the eABF versions, the need for an explicit analytical determination of $|\mathbf{J}|$ is eliminated, making it possible to study the systems with complex entangled movements with multiply defined CVs, where the analytical determination of the Jacobian may be challenging.[106,107]

The WTM-eABF algorithm involves the use of Langevin dynamics to evolve the molecular system at a temperature T. The total energy, denoted by $U(\mathbf{x}, \lambda)$, is not solely dependent on the force-field term, $U_{\mathrm{FF}}(\mathbf{x})$, but also takes into account the spring that links the actual CV to a fictitious particle and additional biasing potentials, $U_{\mathrm{bias}}(\lambda)$. The equations of motion that correspond to this are given as

follows:[116,118]

$$\begin{cases} U(\mathbf{x},\lambda) & = U_{\mathrm{FF}}(\mathbf{x}) + \frac{1}{2}k\big(\xi(\mathbf{x}) - \lambda\big)^2 + U_{\mathrm{bias}}(\lambda) \\ \mathbf{m_x}\ddot{\mathbf{x}} & = -\nabla_{\mathbf{x}}U(\mathbf{x},\lambda) - \gamma_{\mathbf{x}}\mathbf{m_x}\dot{\mathbf{x}} + \sqrt{\frac{2\gamma_{\mathbf{x}}}{\beta}}\ \mathbf{m}_{\mathbf{x}}^{1/2}\ \dot{\mathbf{W}}(t) \\ m_\lambda\ddot{\lambda} & = -\nabla_\lambda U(\mathbf{x},\lambda) - \gamma_\lambda m_\lambda\dot{\lambda} + \sqrt{\frac{2\gamma_\lambda}{\beta}}\ m_\lambda^{1/2}\ \dot{\mathbf{W}}(t) \end{cases} \tag{2.3}$$

where $\xi(\mathbf{x})$ is the real CV, which is a function of the Cartesian coordinates of the system, and the extended degree of freedom, λ, which represents the fictitious particle. The spring connecting the real CV to the fictitious particle is characterized by a force constant k, while the time-dependent Wiener process (also known as Brownian motion), used to describe the random movement of a particle or system, is denoted by $\mathbf{W}(t)$. The masses of the atoms and the extended variable are represented by $\mathbf{m_x}$ and m_λ, respectively, and the friction coefficients of the atoms and the extended variable are represented by $\gamma_{\mathbf{x}}$ and γ_λ, respectively.

The coupling between the real and fictitious particles needs to be sufficiently strong for the applied forces to influence the actual CV. The sampling process is significantly impacted by two critical parameters that regulate the spring, namely, the oscillation period (τ) and the coupling width (σ). These parameters influence the inertial mass of the fictitious particle, as per,

$$\begin{cases} m_\lambda & = & \frac{1}{\beta}\left(\frac{\tau}{2\pi\sigma}\right)^2 \\ k & = & \frac{1}{\beta\sigma^2} \end{cases} \tag{2.4}$$

The extended Langevin dynamics has another parameter known as the damping factor, γ_λ, which can slow down or accelerate the diffusive sampling in the PMF

calculation, depending on its value.[118–122]

The use of the WTM-eABF method resulted in notable improvements in simulation convergence rate, computational efficiency, and stability, compared to previously discussed algorithms, ensuring that the free-energy landscape is adequately sampled without the risk of becoming trapped in local minima, which can hinder accurate standard binding affinity calculations for both protein-ligand and protein-protein complexes.[106,109,115,116] In the following section, it will be explained how the PMF can be reconstructed from the WTM-eABF simulation.

2.3 Reconstruction of the PMF from the MD biased simulations

To obtain the PMF from the WTM-eABF simulation,[106] which uses an extended-Lagrangian formalism, the first step is to extract the average forces acting on the CVs.[75,123] This can be achieved using either the umbrella-integration (UI) estimator[124] or the corrected z-averaged restraint (CZAR) estimator, which both rely on collecting the average forces from the spring connecting the CVs to their fictitious particles. In our work, the CZAR estimator was chosen over the UI estimator, as it has been shown to yield a better estimate of the gradients and PMF, particularly for looser springs.[109]

The CZAR estimator formula[109] for estimating the free-energy gradient, $A'(\xi_i)$, as a function of the i-th collective variable ξ_i, is:

$$A'(\xi_i) = -\frac{1}{\beta}\frac{\mathrm{d}\ln\tilde{\rho}(\boldsymbol{\xi})}{\mathrm{d}\xi_\mathrm{i}} + k_i\langle\lambda_i - \xi_i\rangle_{\boldsymbol{\xi}}, \tag{2.5}$$

where $\tilde{\rho}(\boldsymbol{\xi})$ is the probability density of the system at the CV $\boldsymbol{\xi}$, λ_i is a value of the i-th fictitious particle, k_i is the corresponding spring force constant, and $\langle\rangle_{\boldsymbol{\xi}}$ is the ensemble average. Once the free-energy gradients are estimated by the CZAR estimator, the PMF, $A(\boldsymbol{\xi})$, can be numerically integrated using standard numerical integration techniques such as the trapezoidal rule or Simpson's rule, in the case of only one CV.

Chapter 3

A Comprehensive Protocol for Binding Affinity Calculations

3.1 Effective and Versatile BFEE2 Protocol

Protocol overview. In the initial phase of my doctoral research, our research group undertook the development of a comprehensive protocol aimed at the accurate calculation of protein-ligand standard binding free energies. This protocol was implemented with the software called Binding Free-Energy Estimator 2 (BFEE2),[75,123] and published in a notable journal (H. Fu et al., *Nat. Protoc.*, 2022). The protocol uses advanced simulation techniques that account for the large configurational changes that occur during protein-ligand binding, such as the WTM-eABF algorithm[106,115] and alchemical and geometrical routes,[12] which were discussed in Chapters 1 and 2. The BFEE2 software has been designed with user-friendliness in mind, aiming to minimize the need for manual intervention. It follows a well-defined workflow, comprising the following key sub-steps (also depicted in Figure 3.1):[75]

1. **Modeling.** The initial step involves acquiring a three-dimensional structure

from experimental data, identified by a unique PDB ID.[125,126] Alternatively, if a PDB-formatted file resulting from molecular docking simulations is available, it can be used as well. Additionally, topology files for the protein-ligand complexes in a format compatible with MD engines, such as NAMD[36] and GROMACS,,[127] is essential. This step can be accomplished using modeling tools like CHARMM-GUI[128] and AmberTools.[129]

2. **Input files generation.** During this step, the BFEE2 tool automatically generates all the required configuration files for the multistep free-energy calculation. In addition to selecting the geometrical or alchemical route, the user has the flexibility to fine-tune advanced features. For the alchemical route, the user can regulate the lambda schedule in the TI part of the alchemical route, controlling the gradual transformation of the ligand or specify the number of windows to be used. These advanced features provide greater control and customization of the free-energy calculation process, allowing for optimal sampling and accurate estimation of protein-ligand binding free energies.

3. **Simulation.** This step involves running the MD simulations using the MD engine, which has to be patched with the Colvars module.[67] Additionally, this step may require parameter tuning by the end-user to ensure convergence of the simulations. The trajectory can be visualized using molecular visualization software such as VMD (Visual Molecular Dynamics).[85] Part of the convergence analysis of the geometrical and alchemical routes is avail-

able in the latest version of BFEE2.[123] Regarding the alchemical route, an additional tool, ParseFEP,[76] as a part of the VMD analysis tools,[85] can be used for the automated generation of PDF profiles and hysteresis analysis of the FEP component.

4. **Post-treatment.** BFEE2 can perform all the post-treatments by analyzing the output files generated in the different MD simulations, without requiring any human intervention.

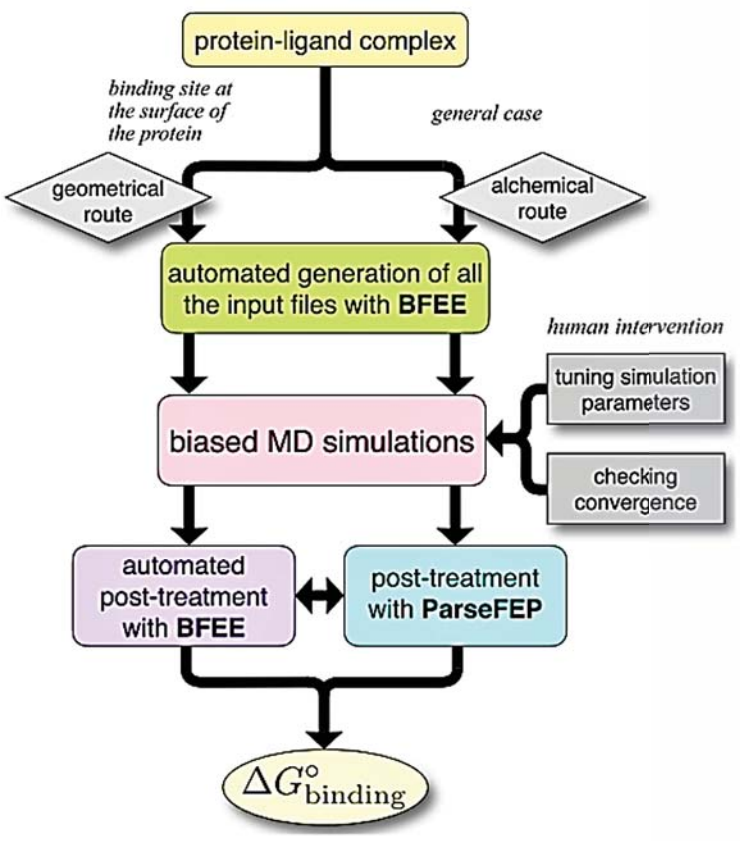

Figure 3.1: The light-yellow box represents the "Modeling" sub-step, the green box, the "Input files generation" sub-step, the pink box, the "Simulation" sub-step, and the purple and cyan boxes, the "Post-treatments" sub-step. Reproduced with permission from *Nat. Protoc.*, **2022**, 17, 1114–1141, Copyright 2022 Springer Nature.

3.1.1 Standard Error Estimation in the Geometrical Route

In the BFEE2 post-treatment of the geometrical route, in contrast to the alchemical route, the standard error estimation is not automatized. To measure the error, there are three possible ways:

- Parallel simulations of the same energetic contributions. While conceptually straightforward, it can be computationally expensive and time-consuming.

- The estimation of the standard error of the correlated biasing force.[130] By assessing the autocorrelation function of the biasing forces, it is possible to estimate the correlation time, which represents the timescale over which the biasing forces remain correlated. The standard error can then be computed based on this correlation time and the number of uncorrelated samples in the simulation. While this method requires a slightly more complex analysis compared to other error estimation techniques, it provides a robust estimate of the standard error.

- Pseudo–parallel simulations. This approach is particularly useful when dealing with simulations that span a long time period, such as a PMF simulation lasting 100 ns or more. The simulation trajectory is divided into two or more segments, typically of equal length. For example, in the case of a 100 ns PMF simulation, the trajectory can be split into two 50 ns segments, denoted as PMF_1 and PMF_2. The information required for this analysis can be obtained from the **.hist.count** and **.hist.pmf** files generated during the simulation.

This approach provides a way to estimate the variability or standard error between these segments and gives insight into the stability and convergence of the calculated free energies. Pseudo-parallel simulations offer a practical and efficient means to estimate the standard error without the need for extensive parallel simulations or complex statistical analysis. It allows for the assessment of the reproducibility and reliability of the results obtained from different portions of the simulation trajectory.

To calculate the standard errors in the geometrical route simulations, I used the pseudo-parallel approach. In detail, given the PMFs (PMF_1 and PMF_2), firstly, the inverse of their weighted averages has to be found. Here, the weights can be determined based on the number of samples per bin obtained from the biased simulation (***.hist.count**). By taking the weighted averages, the contribution of each sample is appropriately accounted for, giving more weight to bins with a larger number of samples and providing a more accurate estimation of the average values. Let v_i and v_j represent the weights for PMF_1 and PMF_2, respectively. The weighted averages can be calculated as follows:

$$\bar{x}_{v_i} = \frac{\sum x_i v_i}{\sum v_i} \tag{3.1}$$

$$\bar{x}_{v_j} = \frac{\sum x_j v_j}{\sum v_j} \tag{3.2}$$

where x_i and x_j are the values of PMF_1 and PMF_2, respectively.

After calculating the weighted averages, the variances can be computed for

each set. Let n represent the number of values in each set:

$$s_i^2 = \frac{\sum v_i \cdot (x_i - \bar{x_{v_i}})^2}{(n-1) \cdot \sum v_i} \tag{3.3}$$

$$s_j^2 = \frac{\sum v_j \cdot (x_j - \bar{x_{v_j}})^2}{(n-1) \cdot \sum v_j} \tag{3.4}$$

Finally, to compute the total variance, standard deviation (SD), and standard error (SE), let N be the total number of sets:

$$s^2{}_{tot} = s^2{}_i + s^2{}_j \tag{3.5}$$

$$SD = \sqrt{\frac{s^2{}_{tot}}{N}} \tag{3.6}$$

$$SE = \frac{SD}{\sqrt{N}} \tag{3.7}$$

3.1.2 Convergence Analysis of PMF Calculations

The convergence of the PMF calculations of the geometrical route can be obtained using the following steps: (i) generate the free energy gradient along ξ at every time step of the simulation (***.hist.czar.grad** files in outputs generated during the simulations), (ii) calculate the root-mean-square difference (RMSD) of the gradient via:

$$\mathrm{RMSD}(t_i) = \sqrt{(\frac{\sum_{i=0}^{N}(\nabla(\xi, t_N) - \nabla(\xi, t_i))^2)}{N}} \tag{3.8}$$

The time, t_i, varies in the range $[0, t_N]$, where t_N is the final simulation time.[12,105,118] $\mathrm{RMSD}(t_i)$ asymptotically reaches 0 for any PMF obtained in the geometrical route, suggesting convergence of the calculations (Figure 3.2).

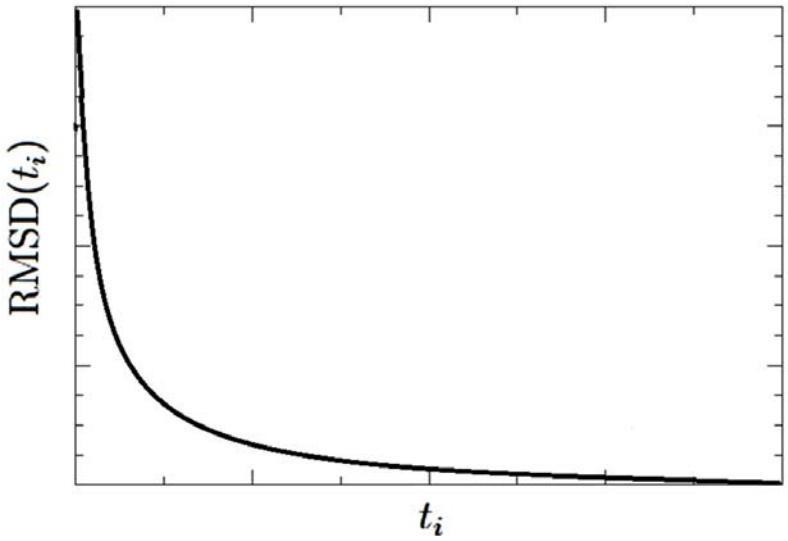

Figure 3.2: An example of the convergence plot observed in the PMF calculations of the geometrical route. The plot shows the RMSD of the gradient, as defined in equation 3.8, as a function of simulation time. As the simulation progresses, the RMSD value gradually decreases and eventually reaches a plateau, indicating convergence of the calculations. The asymptotic behavior of the convergence rate confirms the stability and reliability of the PMF estimation.

3.2 Practical Examples for the Protocol Performance Analysis

To assess the performance of the BFEE2 protocol, a subset of 31 protein-ligand complexes was studied by our research group.[75] Among these complexes, I personally conducted standard binding free energy calculations on three selected systems (MDM2-p53:NVP-CGM097 in geometrical route and MUP-I:2-methoxy-3-isopropylpyrazine and MUP-I:6-hydroxy-6-methyl-3-heptanone in alchemical route). Below, I provide their description, the chosen parameters for the binding free-energy calculation, and the obtained results for each of the studied complexes.

MDM2-p53:NVP-CGM097: Description of the molecular assembly. The first investigated molecular complex consists of a protein complex called MDM2-p53 which has a key role in the degradation of p53 protein, and a ligand, NVP-CGM097, which plays a crucial role in the degradation of the tumor suppressor protein, p53 (Figure 3.3).[131–133] The NVP-CGM097 is a dihydroisoquinolinone derivative that is partially buried inside the protein and occupies the central part of the binding site, reaching three important binding pockets consisting of the amino acid residues Leu26, Trp23, and Phe19 of the p53 protein. The binding affinity of the ligand to the protein was quantified using isothermal calorimetry (ITC), which showed a binding free energy of -11.8 kcal/mol.[131] The initial coordinates for the simulation were obtained from PDB entry 4ZYF of the resolution of 1.80 Å obtained with X-ray crystallography.[131]

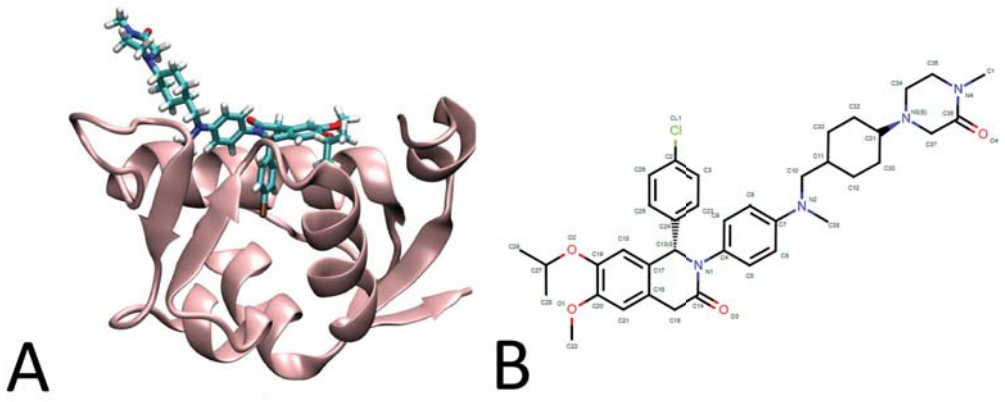

Figure 3.3: (A) Structure of MDM2-p53: NVP-CGM097 complex. (B) The chemical structure of the ligand of the studied complex: (S)-1-(4-chlorophenyl)-7-isopropoxy-6-methoxy-2-(4-(methyl(((1r,4S)-4-(4-methyl-3-oxopiperazin-1-yl)cyclohexyl)methyl)amino)phenyl)-1,2-dihydroisoquinolin-3(4H)-one.

Computational assays. The protein, ligand, and water were modeled using the all-atom CHARMM36m force field,[134] CHARMM general force field[135] and the TIP3P water model,[18] respectively. The ligand parameters were generated through CGenFF program.[136] The lone pair was not included in the simulation for the chlorine atom. The final system consisted of 97,557 atoms atoms in total, among which the ligand has 95 atoms and the protein, 1,570 atoms. The rest of the system was represented by counterions of sodium, chloride, and water molecules. The dimensions of the periodic cells for bound and unbound states were 109 × 95 × 102 $Å^3$.

Prior to using the BFEE2 pre-treatment tool, the system was pre-equilibrated for 100 ns in NVT ensemble with NAMD 3.0 program,[36] at a temperature of 298.15 K (in accordance with the experimental conditions of the ITC experiment for the studied complex) and pressure of 1 atm. Damped Langevin dynamics and Langevin piston were employed for temperature and pressure control,[137,138] respectively, with a time step of 2 fs for the integration of the equation of motion. A smoothed 12 Å spherical cutoff was used for van der Waals and short-range electrostatic interactions, and the particle-mesh Ewald (PME) algorithm was used for long-range electrostatic interactions.[139] The distance between pairs for inclusion in pair lists was 14 Å. The final equilibrated coordinates were saved in a new PDB file and used for further analysis using the BFEE2 tool.

The BFEE2 pre-treatment tool was used to generate the necessary inputs for standard binding free-energy calculations using the geometrical route. The procedure described above was followed for the standard binding free energy cal-

culations. The PMF curves (Figure 3.4) were observed to exhibit continuous and asymptotic behavior, indicating sufficient sampling of the system. To ensure reliable results, the distribution of sampling along the collective variable was checked using the ***.hist.count file**. Additionally, we controlled the convergence rates of the simulations (Figure 3.5). The results of standard binding free energy calculations for this complex are presented in Table 3.1 and are in great agreement with the experiment.[131]

Table 3.1: Contribution to standard binding free energy calculation. The number of nanoseconds enough for the reasonable convergence of the components and computer time used to perform the simulation on 32 CPU cores and 2 GPUs (GeForce RTX 2080 Ti).

Contribution	Free energy (kcal/mol)	Simulation time (ns)	Speed (ns/day)
$\Delta G_{\mathrm{c}}^{\mathrm{site}}$	-7.2 ± 0.1	100	60
$\Delta G_{\Theta}^{\mathrm{site}}$	-0.3 ± 0.0	20	61
$\Delta G_{\Phi}^{\mathrm{site}}$	-0.5 ± 0.0	20	61
$\Delta G_{\Psi}^{\mathrm{site}}$	-0.2 ± 0.0	20	61
$\Delta G_{\theta}^{\mathrm{site}}$	-0.2 ± 0.0	20	44
$\Delta G_{\varphi}^{\mathrm{site}}$	-0.2 ± 0.0	20	42
$(1/\beta)\ln(S^*I^*C^\circ)$	-17.9 ± 0.6	180	36
$\Delta G_{\mathrm{c}}^{\mathrm{bulk}}$	8.6 ± 0.3	100	96
$\Delta G_{\mathrm{o}}^{\mathrm{bulk}}$	6.6		
$\Delta G_{\mathrm{b}}^{\circ}$	-11.3 ± 1.0 (calculation) -11.8 (experiment)[131]		

MUP-I:2-methoxy-3-isopropylpyrazine: Description of the molecular assembly. The MUP-I protein is a pheromone-binding protein that is commonly found in male mouse urine. It has a typical lipocalin fold structure comprising of an eight-stranded β-barrel and a single α-helix, with the interior of the barrel forming a hydrophobic cavity (Figure 3.6). The ligand of this complex is 2-methoxy-3-isopropylpyrazine, which is a small hydrophobic molecule situated inside the cavity of the MUP-I protein.[140]

Table 3.2: Results for each contribution to the binding free energy of MUP-I:2-methoxy-3-isopropylpyrazine. The number of nanoseconds enough for the reasonable convergence of the components and computer time used to perform the simulation on 32 CPU cores and 2 GPUs (GeForce RTX 2080 Ti).

Contribution	Free energy (kcal/mol)	Simulation time (ns)	Speed (ns/day)
$\Delta G^{\text{site}}_{\text{decouple}}$	97.6 ± 0.9	100	22.0
$\Delta G^{\text{site}}_{\text{c}}$	-0.3 ± 0.1	64	25.8
$\Delta G^{\text{site}}_{\Theta}$	-0.6 ± 0.1	64	25.8
$\Delta G^{\text{site}}_{\Phi}$	-0.9 ± 0.2	64	25.8
$\Delta G^{\text{site}}_{\Psi}$	-0.7 ± 0.2	64	25.8
$\Delta G^{\text{site}}_{\theta}$	-0.4 ± 0.1	64	25.8
$\Delta G^{\text{site}}_{\varphi}$	-0.2 ± 0.0	64	25.8
$\Delta G^{\text{site}}_{r}$	-0.3 ± 0.1	64	25.8
$\Delta G^{\text{bulk}}_{\text{decouple}}$	-114.9 ± 0.1	100	54.4
$\Delta G^{\text{bulk}}_{\text{c}}$	0.6 ± 0.0	64	95.1
$\Delta G^{\text{bulk}}_{\text{o+r+a}}$	12.3		
$\Delta G^{\circ}_{\text{b}}$	-7.8 ± 1.0 (calculation) -7.8 (experiment)[140]		

2-methoxy-3-isopropylpyrazine is located inside a hydrophobic cavity formed by the side chains of Phe38, Leu40, Leu42, Ile45, Leu54, Phe56, Met69, Val82, Tyr84, Phe90, Ala103, Leu105, Leu116, and Tyr120, which are part of the eight-stranded β-barrel and a single α-helix of the MUP-I protein. The hydroxyl group of Tyr120 is involved in direct hydrogen bonding with one of the nitrogen atoms of the pyrazine ring. The experimental value of the binding free energy (-7.8 kcal/mol) was determined using ITC, and the experimental snapshot was obtained by X-ray diffraction at a resolution of 1.75 Å under the PDB ID 1QY2.[140]

Computational details. The molecular assembly for this example was modeled using the same procedure as in the case of the previous example (4ZYF). It consists of a total of 55,837 atoms, and the water box dimensions were $87 \times 85 \times 84$ Å^3. Standard binding free-energy calculations were performed employing the alchemical route within the BFEE2 pre-treatment tool. In the advanced settings,

a lambda schedule of 50 windows for the FEP calculations and 15 for TI calculations were chosen for both the bound and unbound states, helping to improve the accuracy and computational time. For the FEP calculations, a sampling time of 1 ns per window and equilibration prior to data collection of 0.2 ns per window was used (Figure 3.7 for bound and Figure 3.8 for unbound states). For TI calculations, a total sampling time of 2 ns per window was chosen, while other parameters were kept at their default values (Figure 3.9). The results of binding affinity calculations are presented in Table 3.2 and are in agreement with the experimental data.[140]

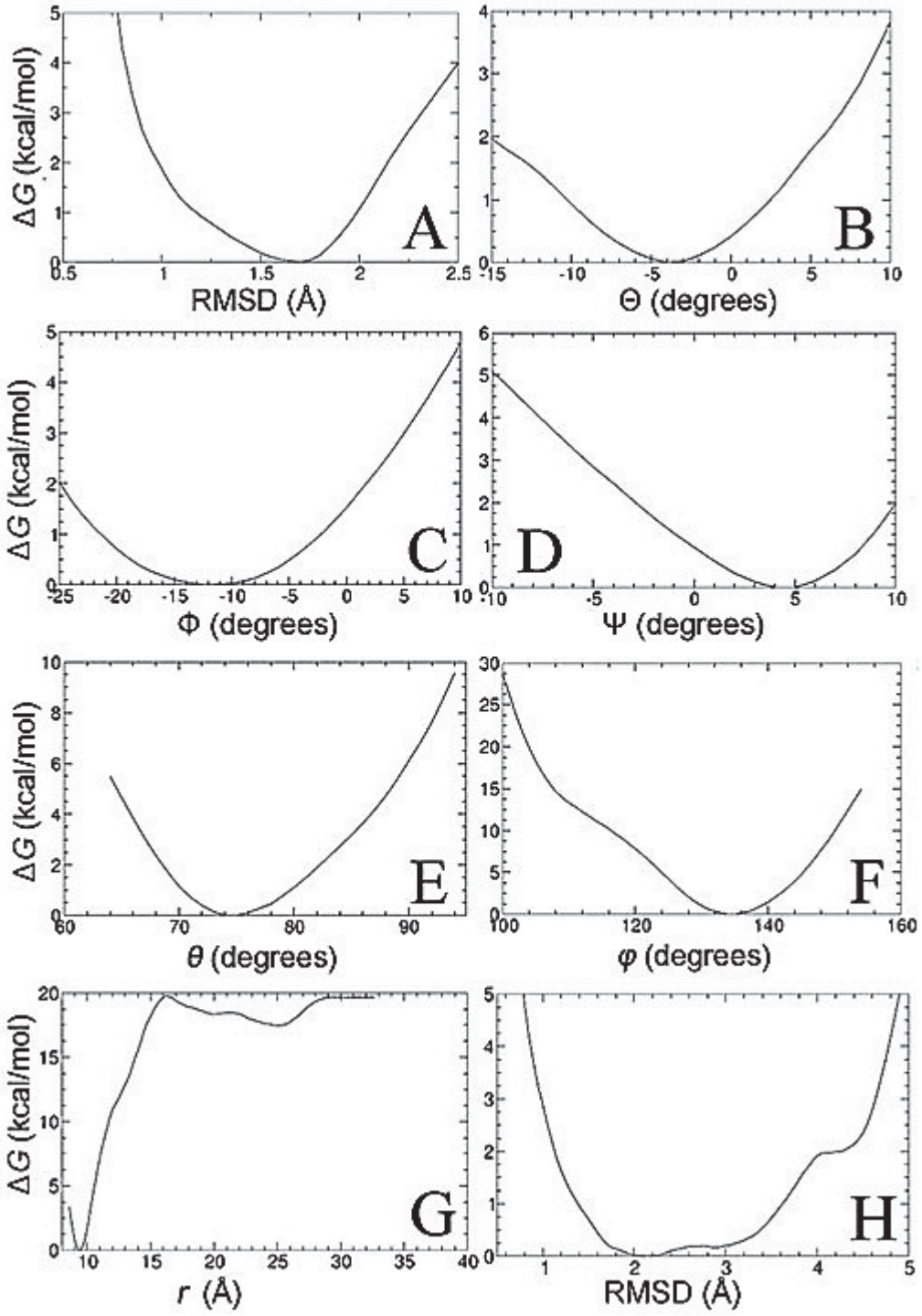

Figure 3.4: Individual PMFs for all components. The PMF calculations using RMSD of the ligand with respect to its bound-state conformation (A), Θ (B), Φ (C), Ψ (D), θ (E), ϕ (F) and centers-of-mass distance between the ligand and protein, with the ligand in the bound state, and RMSD of the ligand with respect to its bound-state conformation with the ligand in the unbound state (H), as the collective variable, respectively. Reproduced with permission from *Nat. Protoc.*, **2022**, 17, 1114–1141, Copyright 2022 Springer Nature.

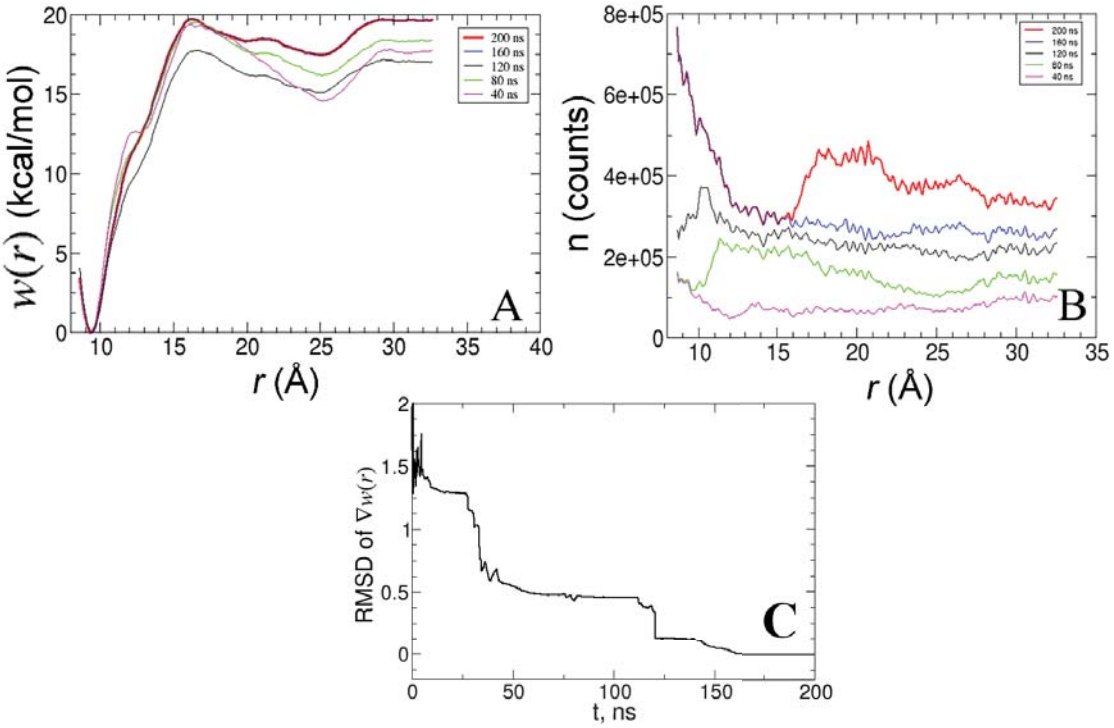

Figure 3.5: Convergence of PMF calculation characterizing the separation of the ligand and the protein. Time-evolution of PMF (A), the distribution of sampling along center-of-mass CV (B), and the progress of the convergence of the simulation (C). Reproduced with permission from *Nat. Protoc.*, **2022**, 17, 1114–1141, Copyright 2022 Springer Nature.

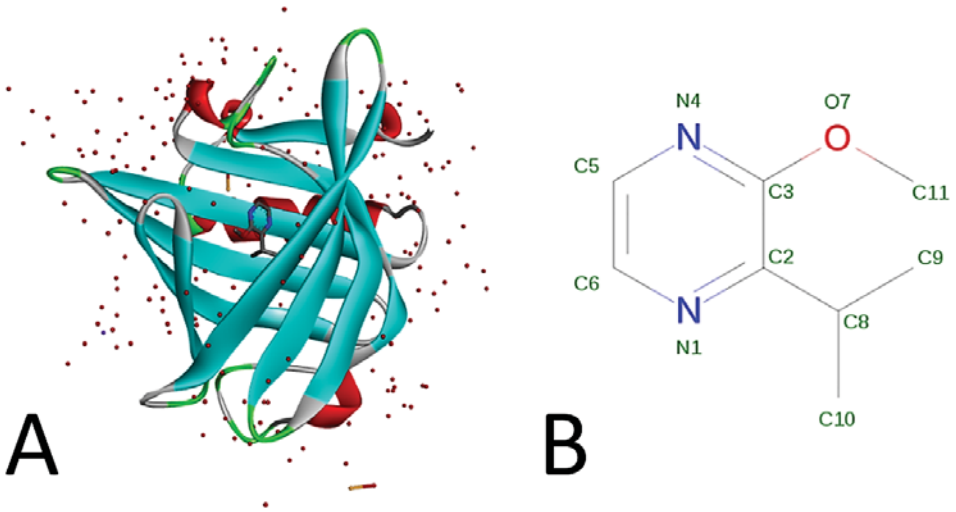

Figure 3.6: (A) Structure of MUP-I:2-methoxy-3-isopropylpyrazine complex. (B) The chemical structure of the ligand of the studied complex.

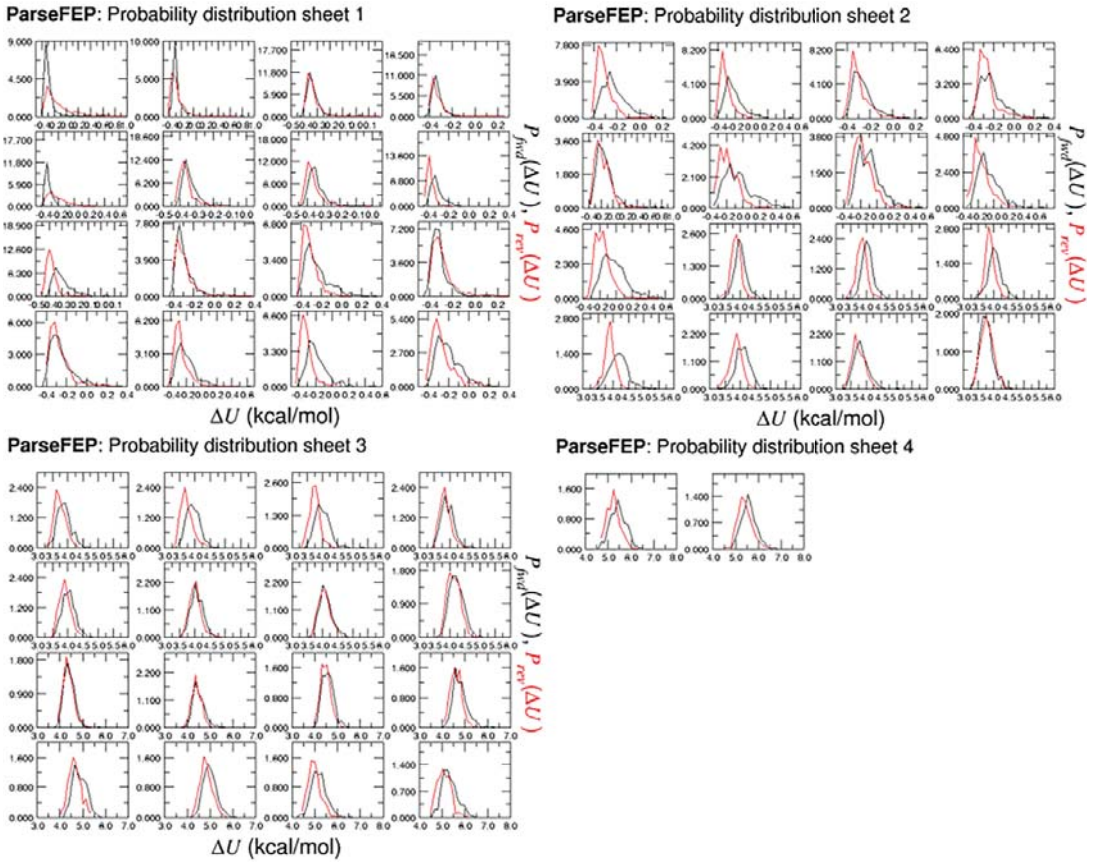

Figure 3.7: Distribution of sampling in each window for reversible decoupling the ligand in the bound state for MUP-I:2-methoxy-3-isopropylpyrazine. Reproduced with permission from *Nat. Protoc.*, **2022**, 17, 1114–1141, Copyright 2022 Springer Nature.

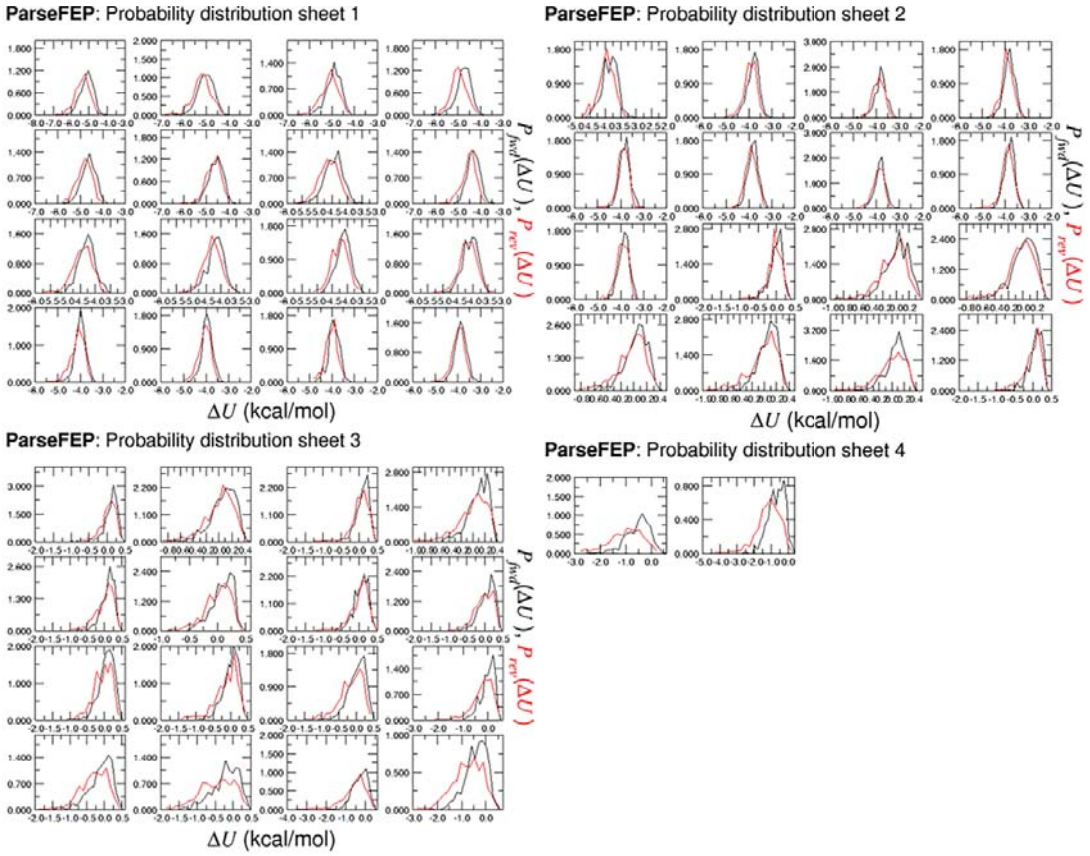

Figure 3.8: Distribution of sampling in each window for reversible decoupling the ligand in the unbound state for MUP-I:2-methoxy-3-isopropylpyrazine.

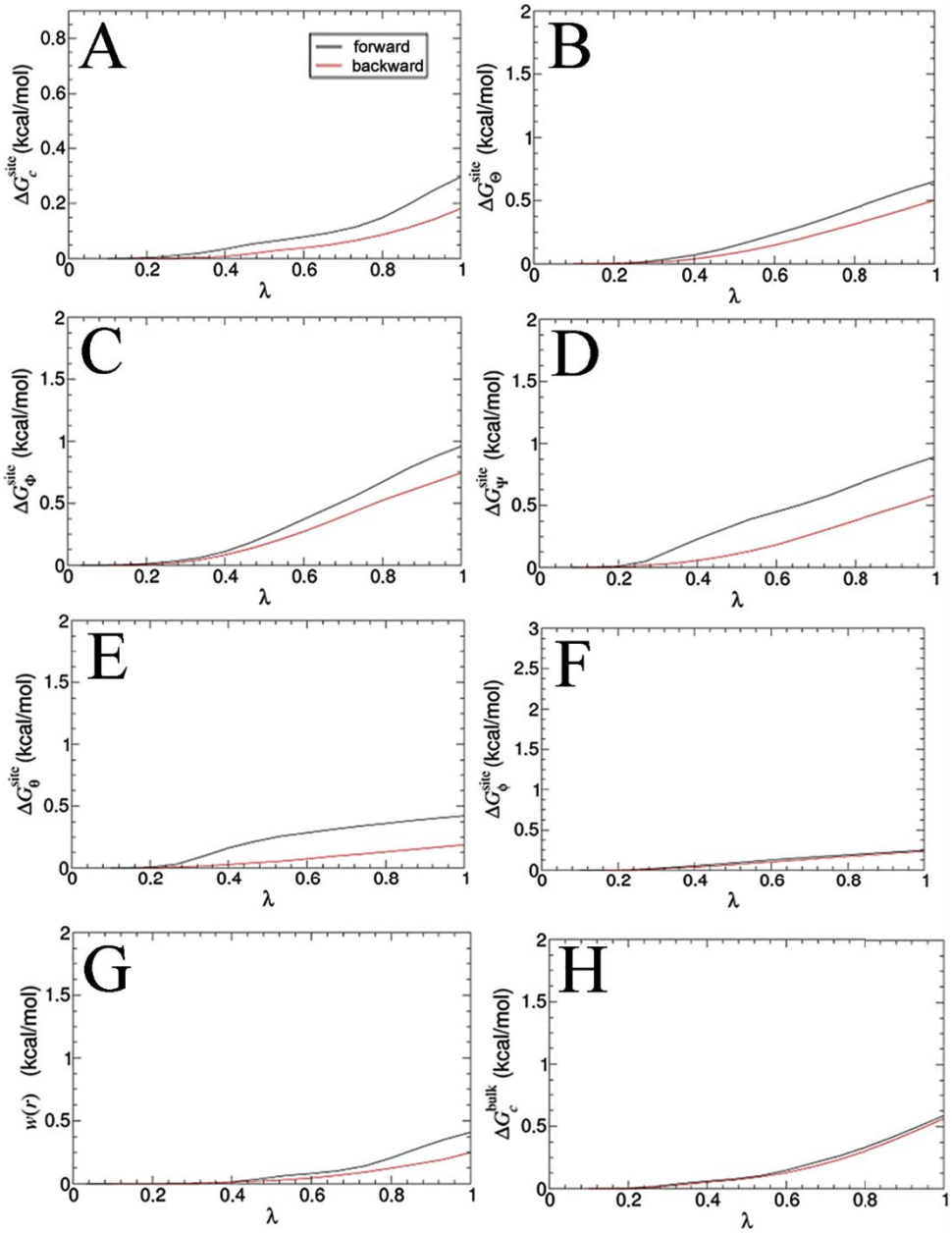

Figure 3.9: Free-energy changes with respect to λ schedule in the TI calculations of the alchemical route. Independent contribution to the free-energy change of each CV is shown, whereas energetic contributions of the reversible decoupling of the ligand from the protein (A-G) and contributions of the unbound ligand (H), respectively. Reproduced with permission from *Nat. Protoc.*, **2022**, 17, 1114–1141, Copyright 2022 Springer Nature.

MUP-I:6-hydroxy-6-methyl-3-heptanone: Description of the molecular assembly. The MUP-I protein is bound by a small hydrophobic molecule known as 6-hydroxy-6-methyl-3-heptanone under PDB ID 1I05.[141] This molecule binds to the hydrophobic environment located at one end of the β-barrel, which is formed by specific side chains such as Phe56, Leu58, Leu60, Ile63, Leu72, Phe74, Met87, Val100, Tyr102, Phe108, Ala121, Leu123, Leu134, and Tyr138.

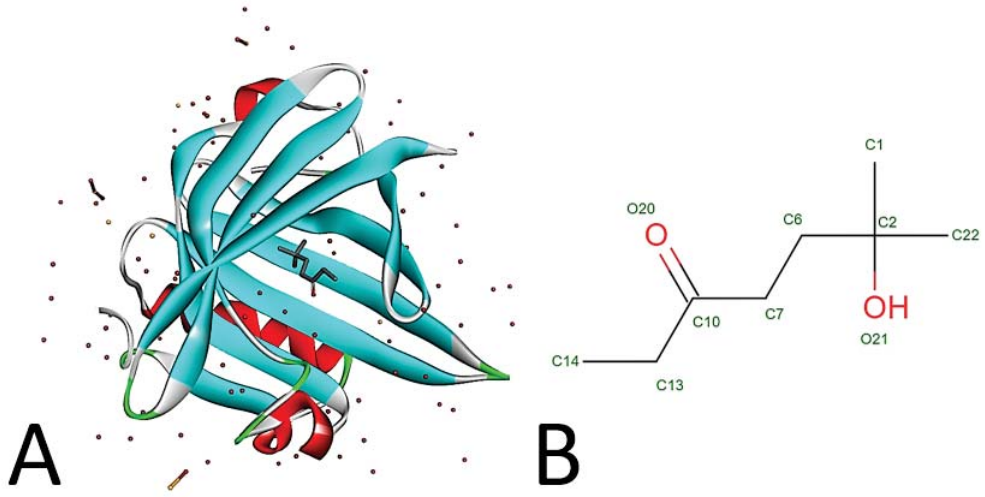

Figure 3.10: (A) Structure of MUP-I:6-hydroxy-6-methyl-3-heptanone complex. (B) The chemical structure of the ligand of the studied complex.

Figure 3.10 displays a snapshot of the complex between the globular protein and the buried ligand. The outermost side chains of the ligand binding site, namely Met87, Phe56, and Leu58, are in contact with the side chain of Tyr102, which serves as a cap for the entrance to the interior of the β-barrel. In addition, two water molecules can be found in the ligand binding site, where they interact with the hydroxyl group of Tyr138 as well as the carbonyl oxygen atoms of Phe56 and Leu58.

The value of the binding free energy was determined via an ITC experiment

at 30°C and was found to be -6 ± 1kcal/mol, with ΔH being -13 kcal/mol and -TΔS being +7 kcal/mol.[141] The enthalpy change implies that the binding process releases energy, indicating an exothermic interaction between the ligand and the binding site. The PDB structure was obtained using X-ray diffraction, with a resolution of 2.00 Å, which provides a satisfactory representation of the ligand binding pose.

Computational details. Herein, the crystal structure of MUP-I:6-hydroxy-6-methyl-3-heptanone is used as the starting point for the standard binding free-energy calculations without any pre-equilibration, as the binding is driven mostly by enthalpy factor and high flexibility of the ligand. As the ligand interacts with the protein and water molecules in the binding site, it undergoes structural adjustments to optimize these interactions. To ensure the stability of the protein-ligand complex, two water molecules are restrained to their hydrogen-bonding partners using a force constant of 0.3 kcal/(mol·Å^2). Additionally, the RMSD of the ligand with respect to its bound-state conformation is restrained using a force constant of 100 kcal/(mol·Å^2) to prevent isomerization of the ligand within the binding pocket. The simulations involve four steps with varying numbers of strata and simulation times and include (i) characterizing the reversible decoupling of the ligand from the protein, (ii) adding restraints on the bound-state ligand, (iii) decoupling the ligand from bulk water, and (iv) adding restraints on the unbound-state ligand. The simulation times per window are 2, 1, 1, and 1 ns for the respective steps, meaning that the TI part of the alchemical route was maintained by applying the SG technique (see Chapter 1). Other parameters remain the same as in the pre-

vious example using the alchemical route. The results of this study are presented in Figures 3.12 3.11, and 3.13, and in Table 3.3.

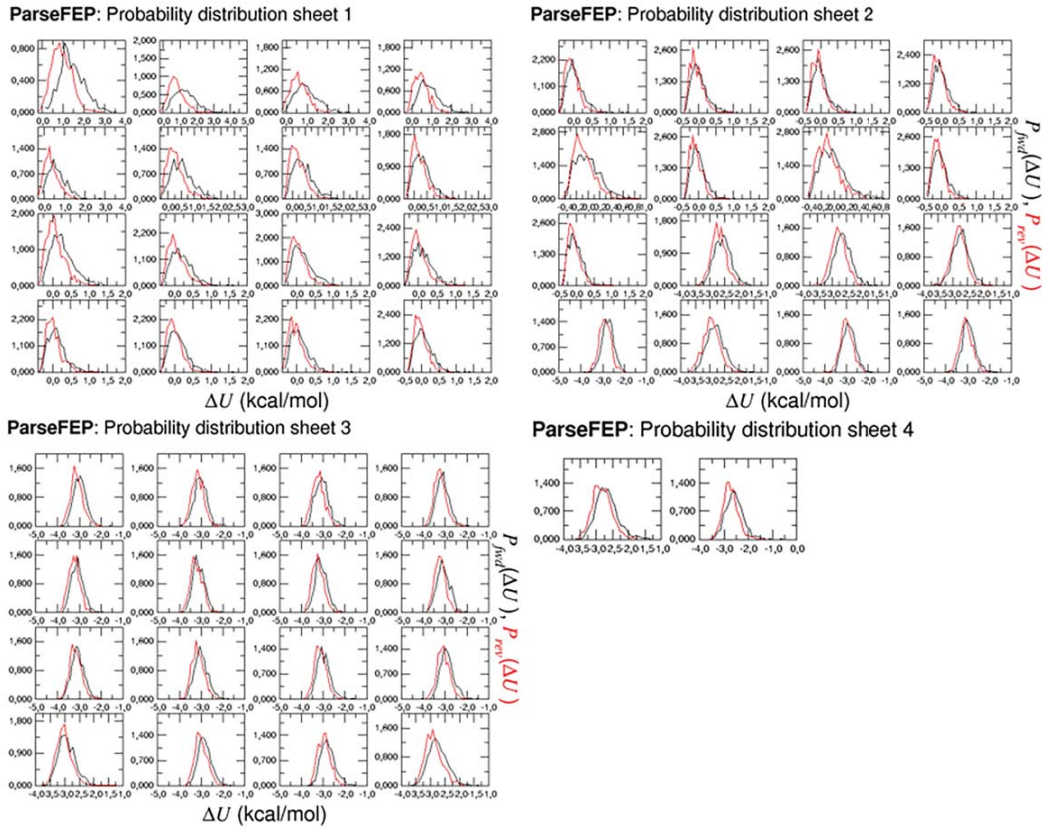

Figure 3.11: Distribution of sampling in each window for reversible decoupling of the ligand in the unbound state for MUP-I:6-hydroxy-6-methyl-3-heptanone. Reproduced with permission from *Nat. Protoc.*, **2022**, 17, 1114–1141, Copyright 2022 Springer Nature.

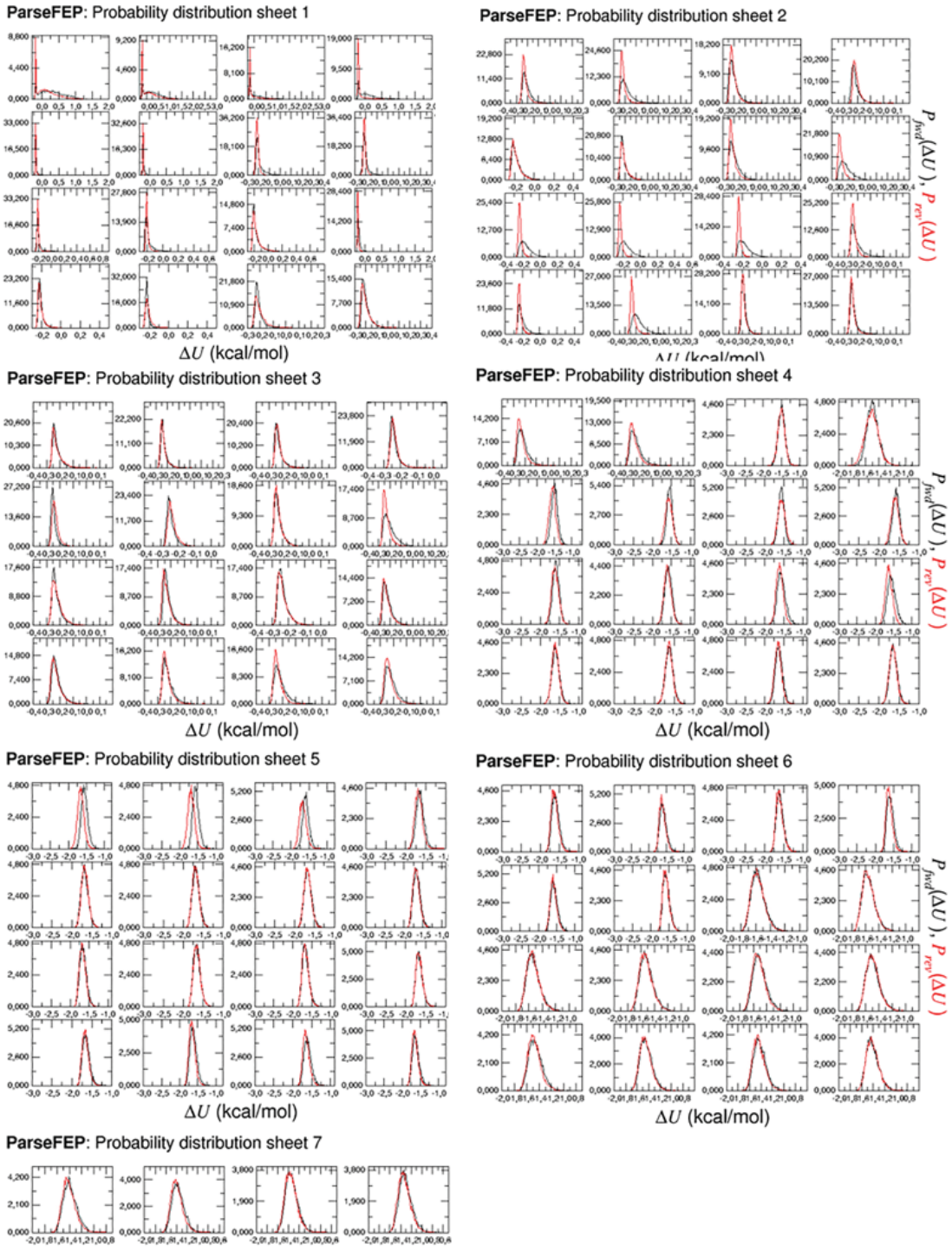

Figure 3.12: Distribution of sampling in each window for reversible decoupling the ligand in the bound state for MUP-I:6-hydroxy-6-methyl-3-heptanone. Reproduced with permission from *Nat. Protoc.*, **2022**, 17, 1114–1141, Copyright 2022 Springer Nature.

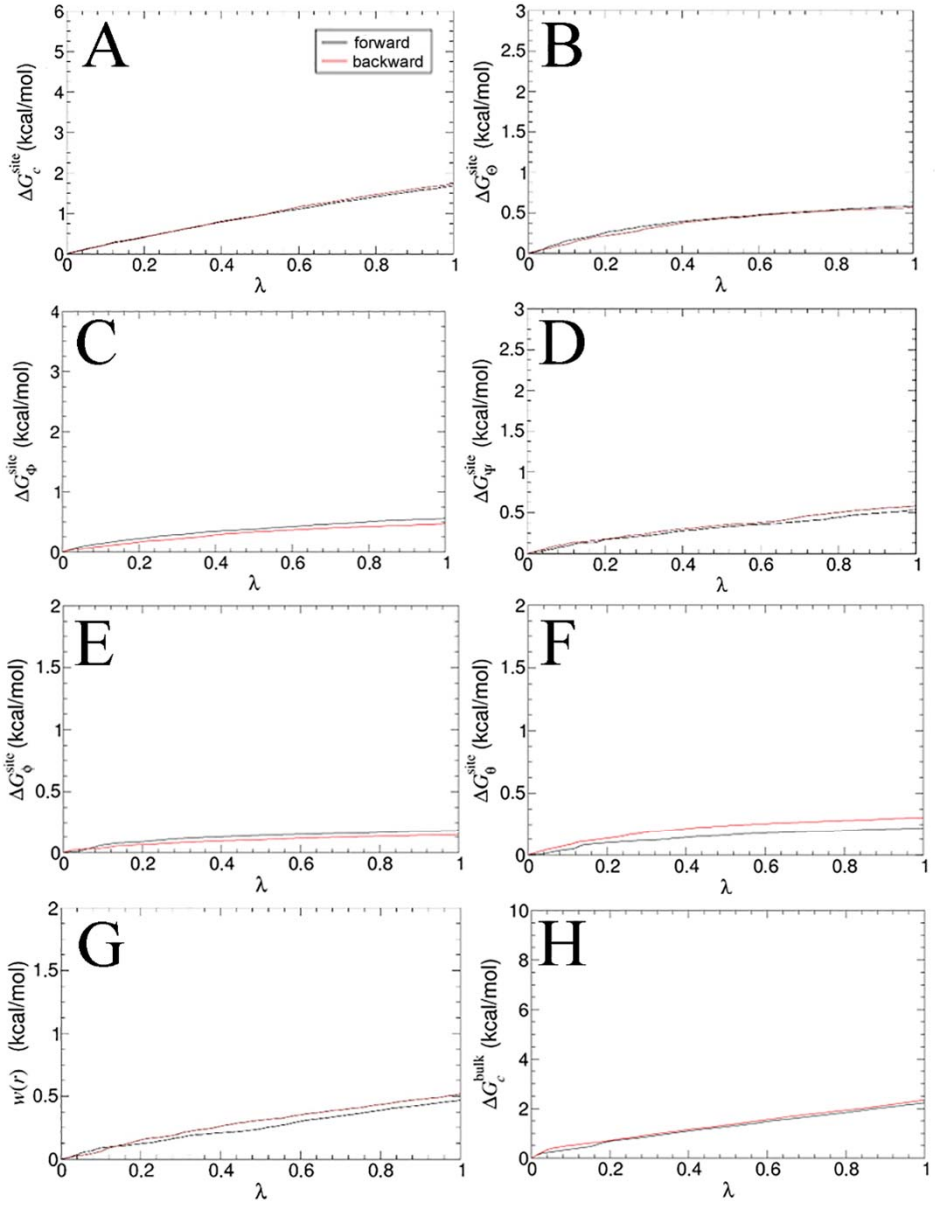

Figure 3.13: Free-energy changes with respect to λ schedule in the TI calculations via the alchemical route. Independent contribution to the free-energy change of each CV is shown, whereas energetic contributions of the reversible decoupling of the ligand from the protein (A-G) and contributions of the unbound ligand (H), respectively. Reproduced with permission from *Nat. Protoc.*, **2022**, 17, 1114–1141, Copyright 2022 Springer Nature.

Table 3.3: Results for each contribution to the binding free energy of MUP-I:6-hydroxy-6-methyl-3-heptanone. The number of nanoseconds enough for the reasonable convergence of the components and computer time used to perform the simulation on 32 CPU cores and 2 GPUs (GeForce RTX 2080 Ti).

Contribution	Free energy (kcal/mol)	Simulation time (ns)	Speed (ns/day)
$\Delta G^{\text{site}}_{\text{decouple}}$	88.7 ± 0.7	200	27.9
$\Delta G^{\text{site}}_{\text{c}}$	-1.8 ± 0.0	0.1	24.0
$\Delta G^{\text{site}}_{\Theta}$	-0.5 ± 0.0	0.1	24.0
$\Delta G^{\text{site}}_{\Phi}$	-0.5 ± 0.0	0.1	24.0
$\Delta G^{\text{site}}_{\Psi}$	-0.5 ± 0.0	0.1	24.0
$\Delta G^{\text{site}}_{\theta}$	-0.3 ± 0.0	0.1	24.0
$\Delta G^{\text{site}}_{\varphi}$	-0.2 ± 0.0	0.1	24.0
$\Delta G^{\text{site}}_{r}$	-0.5 ± 0.0	0.1	24.0
$\Delta G^{\text{bulk}}_{\text{decouple}}$	72.8 ± 0.1	100	49.0
$\Delta G^{\text{bulk}}_{\text{c}}$	2.3 ± 0.0	0.1	40
$\Delta G^{\text{bulk}}_{\text{o+r+a}}$	12.4		
$\Delta G^{\circ}_{\text{b}}$	-5.5 ± 0.7 (calculation) -6.1 (experiment)[141]		

3.2.1 Advantages and Limitations of the Protocol

While the results of the BFEE2 application for the studied systems demonstrate its robustness and accuracy in predicting standard binding free energies, it is also important to consider the pros and cons of this protocol.

Advantages

- **Theoretically rigor**: The protocol is based on a formally rigorous theoretical framework,[29] and free-energy calculations following this protocol are expected to converge within chemical accuracy.

- **Configurational space sampling**: The methodology employs one-dimensional PMF calculations (geometrical route) or alchemical transformations (al-

chemical route) with reduced configurational space sampling by virtue of the introduction of geometric restraints. This strategy obviates the need to capture the large conformational changes that often occur when a protein interacts with a ligand.

- **Minimal human intervention**: The BFEE2 software automates the entire standard binding free-energy calculation process, from the definition of the CV set to the post-treatment of the simulations, minimizing human intervention.

- **Wide applicability**: The BFEE2 tool is suitable for a wide range of protein-ligand and host-guest complexes, including globular and membrane proteins, buried and semi-buried ligands, and both rigid and flexible ligands. It can be easily adapted to specific study cases, allowing users to modify the set of CVs, harmonic force constants, and extended Lagrangian parameters of the WTM-eABF enhanced sampling algorithm. Additionally, the protocol's flexibility and versatility enable it to be extended to investigate protein-protein binding, with only minor modifications, which will be presented in the next Chapter.

- **Easy convergence assessment**: Graphical user interface (GUI)-based tools are available to directly assess the convergence of the PMF calculations or alchemical transformations.

- **Robustness and reproducibility**: The standardized definition of the CV[67] and free-energy calculations[142] ensures that the protocol yields the

same binding affinity for a given protein-ligand complex in replicated simulations.

Limitations

- **Force-field dependence**: The overall accuracy of MD-based binding free-energy calculation strategies, depends on the quality of the force field used. However, the methodology itself used in our protocol is rigorously formulated independently of the force field, and users can improve the reliability of the calculations by turning to more accurate models if necessary.

- **Extremely deeply buried ligands**: For deeply buried ligands, capturing solvent reorganization requires extensive simulation times, which can induce convergence issues in free-energy calculations. In some cases, this issue can be overcome by treating some water molecules as part of the protein or the ligand and by tuning of the simulation parameters.

- **Computational cost**: The methodology is computationally expensive, but this is necessary to guarantee the reliability of the free-energy estimates. Microsecond simulations can now be routinely performed with GPU-based architectures and GPU-accelerated MD engines, and subprocesses in the methodology can be advantageously performed in parallel to reduce the total computational time.

- **Requirement of the native binding motif**: The methodology relies on an assumed initial binding pose, and the outcome is more uncertain

if the pose is inaccurate. It should be noted that alternative simulation strategies that do not rely on prior knowledge of the native binding pose require significant computational effort to discover it from scratch.

Chapter 4

Extension for Protein-Protein Complexes

4.1 Geometrical Route for Protein-Protein Binding Affinity Calculation

When dealing with protein-protein binding affinity, employing the alchemical route for protein-protein complexes is no longer feasible due to considerable solvation-free energies that hinder convergence.[46] Instead, the geometrical route can be employed, which gradually separates the two molecular objects while maintaining conformational, orientational, and positional restraints. The description of these restraints can be found in Table 1.1, Chapter 1. Since any drastic conformational changes occurring in proteins can significantly affect the accuracy and reliability of the binding affinity calculations, the conformational restraints should be applied on both protein backbones during dissociation to ensure structural stability. Optional conformational restraints can also be needed on the interfacing side-chain residues to prevent damage to the binding interface during the separation caused by solvent exposure.[46]

4.2 Case Study of Protein-Protein Binding Affinity Calculation

As a first case study of protein-protein binding affinity calculations following the geometrical route, a pig insulin dimer was investigated.[143] The pig insulin dimer serves as an ideal system for this investigation due to its well-characterized binding properties and the availability of experimental data. Herein, the dimerization is primarily driven by hydrophobic interactions occurring at the interfaces of both monomers. The experimental binding affinity of this dimer was previously reported in the literature,[144] enabling a comparison between the calculated and documented estimates.

Description of the molecular assembly. The dissociation of an insulin dimer to two monomers plays an important role in regulating glucose levels in the blood.[144] Insulin is a double-chain hormone consisting of 21 and 30 residues, making the dimer asymmetric.[145] The first monomer comprises chains A and B, and the second monomer contains chains C and D (Figure 4.1). In the A-B chains, the binding is primarily driven by the ionic interaction between the C-terminus of the B chain and the N-terminus of the A chain, which is absent in the second monomer. The residues Phe24, Phe25, and Tyr26 from chains B and D contribute most significantly to the stability of the dimer by forming four hydrogen bonds and an anti-parallel β-sheet, and also by hydrophobic interactions between the phenylalanine and tyrosine side-chains.[146,147] Other residues at the

interface, including Tyr16, Val12, Pro28, and Gly23 (B and D), make smaller but still favorable contributions to dimerization, while none of the residues from chains A and C are involved in the process.[144,147]

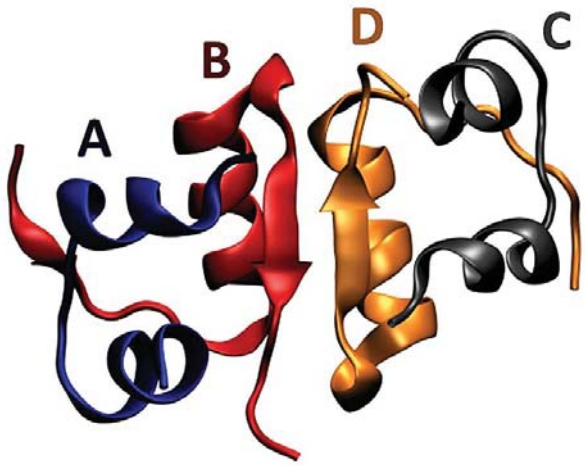

Figure 4.1: The pig insulin dimer structure (PDB: 4INS). Chain A-B and C-D compose two monomers. Reproduced with permission from *J. Phys. Chem. Lett.*, **2022**, 13, 27, 6250–6258, Copyright 2022 American Chemical Society.

Computational details. The starting coordinates were taken from the X-ray diffraction structure of the native pig insulin resolved at 1.5 Å (PDB entry 4INS).[143] The system was described using the all-atom CHARMM36m force field[134] for the protein complex and water, with the TIP3P water model.[18] The system was neutralized with counterions of sodium. The total system comprised 92,711 atoms, with dimensions of the periodic cell measuring 107 × 92 × 104 Å^3. In BFEE2 pre-treatment for protein-protein complexes, the generated configurational files were manually adapted to obtain an additional backbone RMSD restraint of both proteins.

Results. The results of the binding affinity calculations using the geometrical route are represented in Table 4.1. The PMFs of each contribution and the

convergence plots are presented in Figures 4.2 and 4.3.

Table 4.1: Results for each contribution to the binding free energy of the pig insulin dimer in the geometrical route.

Contribution	PMF (kcal/mol)	PMF (ns)
$G^{\text{site}}_{c(A-B)}$	-20.4 ± 0.0	120
$G^{\text{site}}_{c(C-D)}$	-17.5 ± 0.1	120
G^{site}_{Θ}	-0.2 ± 0.1	20
G^{site}_{Φ}	-0.3 ± 0.0	20
G^{site}_{Ψ}	-0.3 ± 0.0	20
G^{site}_{θ}	-0.6 ± 0.0	40
G^{site}_{ϕ}	-0.5 ± 0.0	40
$(1/\beta)\ln(S^*I^*C^\circ)$	-27.4 ± 0.1	100
$G^{\text{bulk}}_{c(A-B)}$	29.0 ± 0.0	200
$G^{\text{bulk}}_{c(C-D)}$	24.6 ± 0.0	200
G^{bulk}_{o}	6.6	
$\Delta G^\circ_{\text{b}}$	-7.0 ± 0.2 (calculation) -7.2 (experiment)[144]	Total: 880 ns

The excellent agreement between the calculated and experimental values of the standard binding free energy for the pig insulin dimer suggests that the geometrical route is a reliable method for predicting protein-protein binding affinities, making it a useful tool for investigating protein-protein interactions.

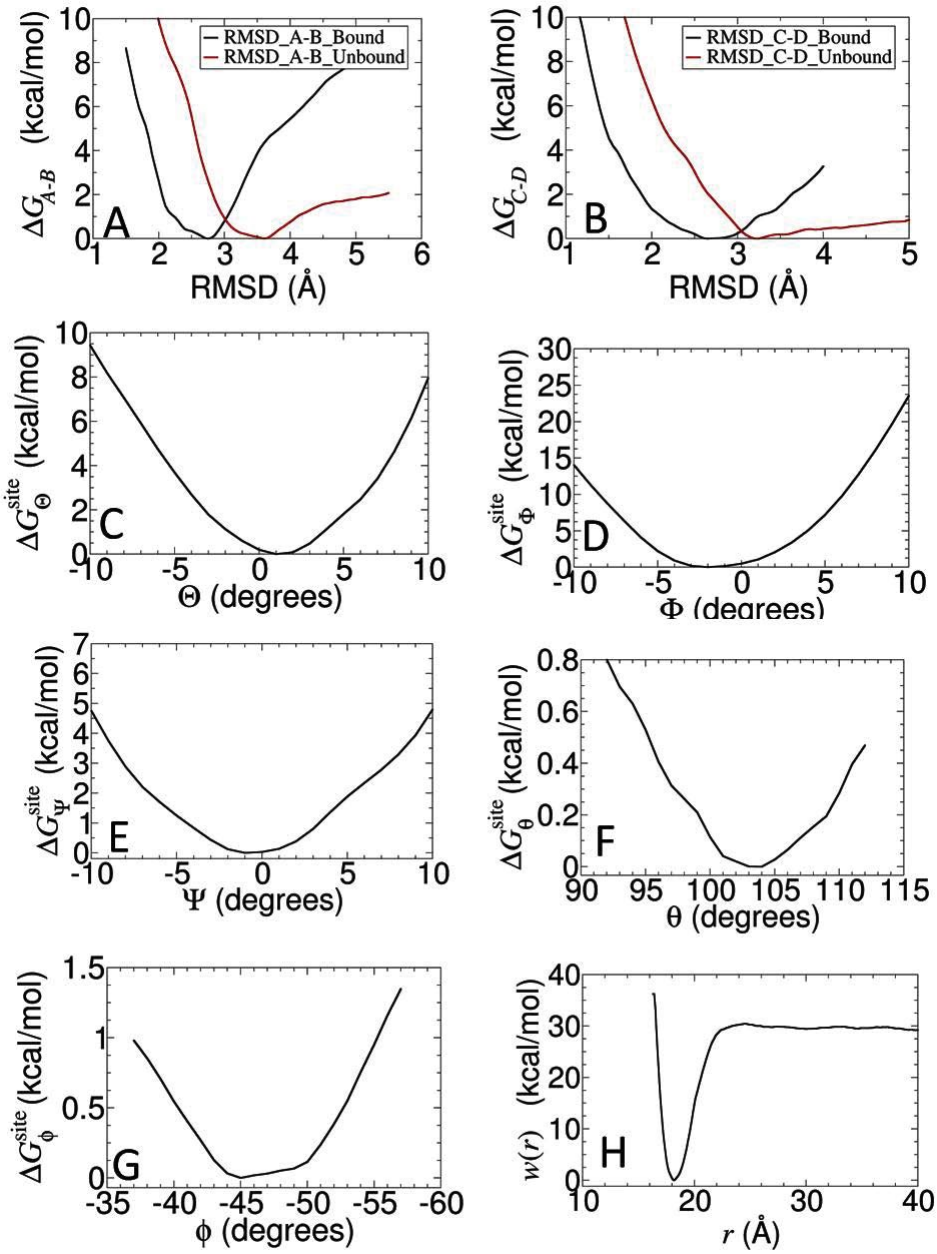

Figure 4.2: Individual PMFs for all components. The PMF calculations using RMSDs of the A-B chains in the bound and unbound state (A), the RMSDs of the C-D chains in the bound and unbound states (B), Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and centers-of-mass distance between two molecular entities (H), as the collective variable, respectively. Reproduced with permission from *J. Phys. Chem. Lett.*, **2022**, 13, 27, 6250–6258, Copyright 2022 American Chemical Society.

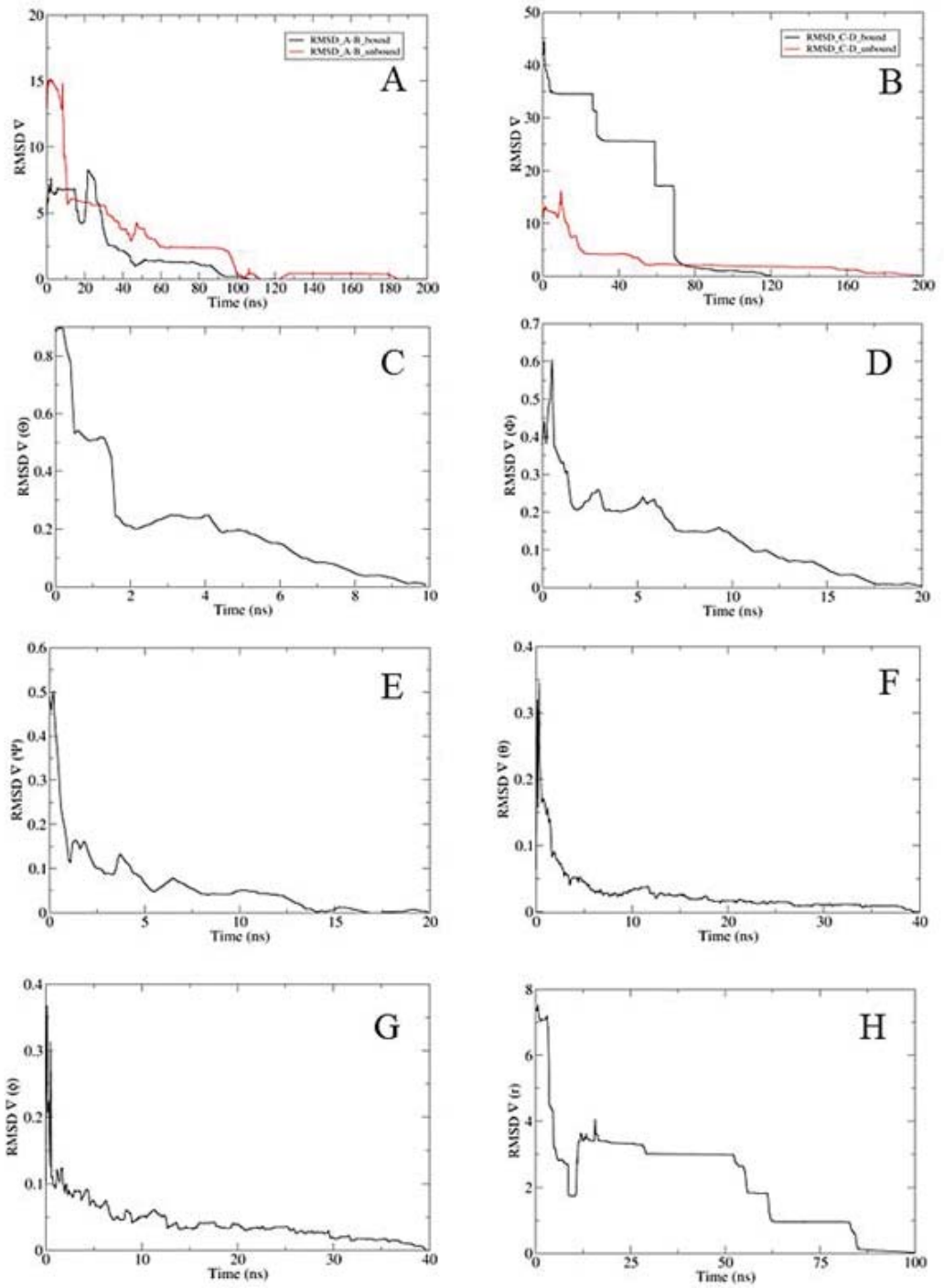

Figure 4.3: Convergence curves for individual PMFs for all components using RMSDs of the pig insulin dimer in the bound and unbound state of monomer A-B (A) and of monomer C-D (B), Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively. Reproduced with permission from *J. Phys. Chem. Lett.*, **2022**, 13, 27, 6250–6258, Copyright 2022 American Chemical Society.

Chapter 5

Hazardous Shortcuts for Binding Affinity Calculations

5.1 Necessity of Restraints Applied on CVs in Binding Affinity Calculations

In order to provide valuable insights into the effectiveness of the restraint-based approaches used for accurate standard binding free-energy calculations for both protein-ligand and protein-protein complexes, a methodological investigation was conducted. Specifically, this part of my research work was focused on comparing the rigorous theoretical framework of the geometrical route with its computational shortcuts, which involve only the physical separation PMF calculations, monitoring all the other CVs described in Table 1.1, yet applying no harmonic potential onto them. In the following, this strategy is referred to as a shortcut of the geometrical route. The results of this investigation ended up with a publication: M. Blazhynska et al. *J. Phys. Chem. Lett.*, 2022.[74]

In cases where the entire configurational space can be adequately sampled within finite-length simulations, such as rapidly relaxing and small binding part-

ners like a benzene dimer,[23] it is possible to calculate a radial physical separation PMF without the need for additional geometrical restraints.[148–151] Assuming that all the degrees of freedom other than the physical separation are suitably averaged, an equilibrium binding constant can be written as follows:[73]

$$K_{\mathrm{eq}} = 4\pi \int_{\mathrm{site}} \mathrm{d}r \; r^2 \; \mathrm{e}^{-\beta[w(r)-w(\mathrm{r}^*)]}, \tag{5.1}$$

where $\beta = (k_{\mathrm{B}}T)^{-1}$, r is the distance between the centers of mass of the two partners, $w(r)$ is a radial separation free-energy profile or PMF, and $w(r^*)$ is the offset at large separation r^* of the complex.

However, when dealing with more complex binders such as protein-ligand and protein-protein complexes, the inclusion of restraints on the chosen CVs, which represent slow degrees of freedom, becomes indispensable. The use of geometrical restraints applied on the CVs prevents the two molecular objects from randomly tumbling while controlling their separation and can help accelerate the convergence of the separation PMF calculation by reducing the conformational space that needs to be sampled. In practice, sampling along all "unrestrained" degrees of freedom other than the distance separating their centers of mass is markedly slowed down by the numerous free-energy barriers to overcome, requiring extensively long simulations, depending upon the height of the free-energy barriers to overcome in order to achieve proper convergence at each separation distance.[12,29,46]

Nevertheless, in the literature, there have been attempts to determine standard binding free energies following elements of the geometrical route with various computational shortcuts, not all of which were formally justified.[152–161] Therefore,

the aim of this work was to demonstrate the necessity of following the complete geometrical route involving PMF calculations along all CVs that that capture the slow degrees of freedom upon binding. This approach ensures both accurate and reproducible standard binding free-energy estimates for a variety of protein-ligand and protein-protein complexes.

5.1.1 Comparing the Geometrical Route with Its Hazardous Shortcuts

In order to illustrate the applicability and effectiveness of the geometrical route compared to its shortcut, two protein-ligand complexes, namely AblSH3:p41[162] and MDM2-p53:NVP-CGM097,[131] along with two protein-protein complexes, the pig insulin dimer and SARS-CoV-2 spike RBD:ACE2, were selected. These representative cases were diligently examined and analyzed as part of my Ph.D. thesis, and are listed in Table 5.1. The study of a protein-protein complex, CheA kinase-P2:CheY, included in the original article, was carried out by my co-author, therefore, will not be shown here. In both strategies, the reference protein was tethered to the origin located at the center of the water box, preventing it from tumbling and drifting. The dimensions of the simulation cell for all the complexes examined herein were large enough to guarantee that in the course of the reversible separation, the binding partners would not interact with their images in the adjacent, periodic cells.

The standard binding free energies for the selected protein–ligand complexes

were already estimated by our group following the geometrical route.[12,75,142] One of the selected protein-ligand complexes is MDM2-p53:NVP-CGM097, which is presented in Chapter 3. In the case of protein–protein complexes, the selection was based on the relatively simple test–cases such as a pig insulin dimer (discussed in Chapter 4) and on the more challenging COVID–19 related complex of SARS-CoV-2 spike RBD:ACE2, for which a 1–μs separation PMF calculated via the shortcut of the geometrical route appeared in the literature.[155,156] The binding affinities via the shortcut of the geometrical route were determined from several replicas (i.e., four for the protein-ligand complexes and two for the protein-protein complexes), each 1–μs long, showing the divergence of the standard binding free–energy estimates (Table 5.1).

The use of geometrical restraints on CVs in finite-length simulations is known to promote convergence of PMF calculations following the geometrical route. This protocol typically yields similar estimates in replicated simulations of protein-ligand and protein-protein complexes.[12,75,142] Therefore, only a single value and its total simulation time for the restrained $\Delta G^{\circ}_{\mathrm{b}}$ are reported in Table 5.1. The following discussion is primarily focused on two protein-protein complexes. Detail of the complexes examined herein, i.e., the structures and computational assays, and analysis, can be found in the next subsection.

The estimates obtained following the geometrical route match the experimental value within chemical accuracy, as shown in Table 5.1. Conversely, the estimates derived from the shortcut method deviate significantly from the experimental data. For instance, for the pig insulin dimer, the values of -4.3 and -8.5

Table 5.1: Standard binding free energies in kcal/mol and association constants for each complex obtained in the shortcut of the geometrical route (with no restraints), the geometrical route (with restraints), and in the experiments

Systems	ΔG_b°		
	(shortcut)	(geometrical route)	(experiment)
Abl kinase-SH3:p41[162]			
1	−5.4		
2	−4.6	−7.9± 0.2 (0.19 μs)[12,142]	−8.0 ± 0.1[162]
3	−4.4		Fluorescence spectroscopy
4	−3.7		
MDM2-p53:NVP-CGM097[131]			
1	−5.4		
2	−5.6	−11.3 ± 0.9 (0.48 μs)[75]	−11.8 ± 1.0[131]
3	−6.0		TR–FRET
4	−6.1		
Pig Insulin Dimer[143]			
1	−4.3	−7.0 ± 0.2 (0.88 μs)	−7.2 ± 0.8[163]
2	−8.5		Spectrophotometry
SARS-CoV-2 spike RBD:ACE2[164]			
1	−4.2	−11.5 ± 0.3 (1.07 μs)	−11.4[164]
2	−7.9		Surface plasmon resonance

kcal/mol for the unrestrained ΔG_b° obtained through the shortcut indicate that it is not a reliable strategy for predicting standard binding free energies and does not provide trustworthy and reproducible results. The substantial deviation observed for nearly all complexes between the unrestrained ΔG_b° estimates and their experimental counterpart further confirms the unreliability of this approach. Although the second replica of the separation PMF calculation for the pig insulin dimer is reasonably close to the experimental value (−8.5 kcal/mol), we consider this result to be an exception that reinforces the general trend.

Figure 5.1 shows the PMFs and time-evolution of the unrestrained ΔG_b° for

the pig insulin dimer in the geometrical route and its shortcut, illustrating that both replicas of the shortcut route exhibit fluctuations in the first half of the free-energy calculation. It is worth noting that one of the replicas even displays an unrestrained $\Delta G_{\mathrm{b}}^{\circ}$ value that reaches as low as -16 kcal/mol before eventually stabilizing at a value approximately $2k_{\mathrm{B}}T$ away from the experimental value. These sudden and pronounced variations observed in the PMF profiles further emphasize the inherent unreliability of employing the shortcut approach within the geometrical route methodology.

Furthermore, the shortcut of the geometrical route imposes sampling the entire configurational space available to the dimer, which makes it challenging to evaluate the contribution of the configurational entropy to the binding affinity at large separations, namely the Jacobian term. Prediction of the Jacobian term requires sufficient information about all possible relative movements within the complex, which, as a matter of principle, is not amenable to finite-length unbiased MD simulations, as shown in Figure 5.1A. This limitation makes the geometrical route a more reliable and accurate approach for predicting standard binding free energies.

In addition, a rough simulation-time estimation necessary to achieve quasi-ergodic sampling along orientational and positional CVs for the pig insulin dimer was estimated. For this purpose, the autocorrelation functions (ACFs), $C(t)$, of angular CVs were fitted, with a simple exponential function,[165] $C(t) = \mathrm{e}^{-t/\tau}$, where t is the simulation time, and τ is the sought quantity, also known as a characteristic relaxation time constant. The estimated times necessary to achieve quasi-ergodic

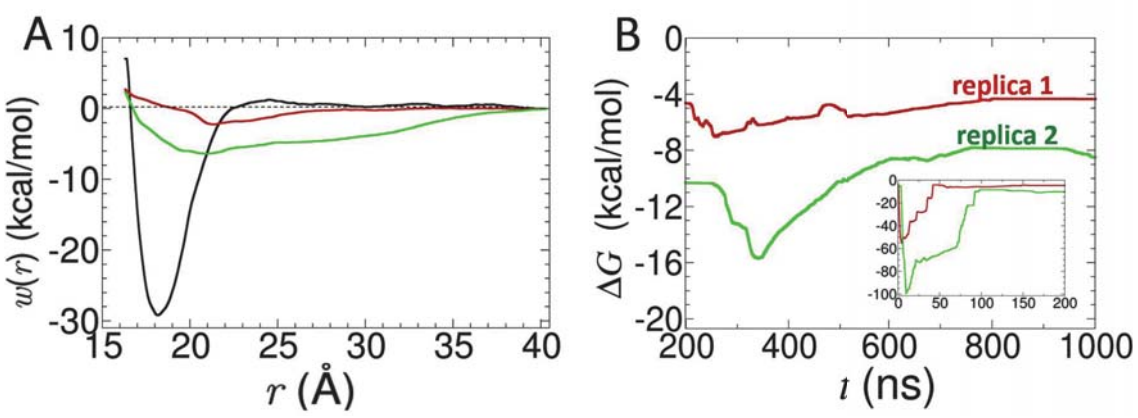

Figure 5.1: (A) Normalized separation PMFs for insulin dimer calculated in the shortcut of the geometrical route for two replicas (in red and green, respectively) in comparison to the separation PMF of the geometrical route (black). All the PMFs were obtained within the separation distance range of [16.3; 40.3] Å. (B) 1–μs evolution of the binding free energy values of insulin dimer obtained in the shortcut calculations for two replicas (in red and green, respectively). The inset provides a closeup of the first 200 ns of the 1–μs trajectories. Reproduced with permission from *J. Phys. Chem. Lett.*, **2022**, 13, 27, 6250–6258, Copyright 2022 American Chemical Society.

sampling along Θ, Φ, Ψ, θ, and ϕ for the simple protein-protein complex of the pig insulin dimer are 0.22, 0.28, 0.28, 35.09, and 13.74 μs, respectively. It is estimated that the value of 35 μs obtained for the polar, θ, likely constitutes an upper bound of the simulation time required to achieve suitable convergence of the strategy that reduces the geometrical route to a mere separation PMF. In addition, the use of fully unrestrained PMF calculations, where the reference protein is allowed to tumble and drift freely, was investigated (Figure 5.3). For the pig insulin dimer, two independent 1-μs-long simulations were conducted, which resulted in final binding free-energy estimates of -6.9 and -3.7 kcal/mol, respectively. These results highlight the inconsistency of the obtained estimates unless the geometrical route is used. The relaxation times required for achieving quasi-ergodic sampling along Θ, Φ, Ψ, θ, and ϕ for the pig insulin dimer are 0.50, 1.70, 1.16, 4.10, and 1.55

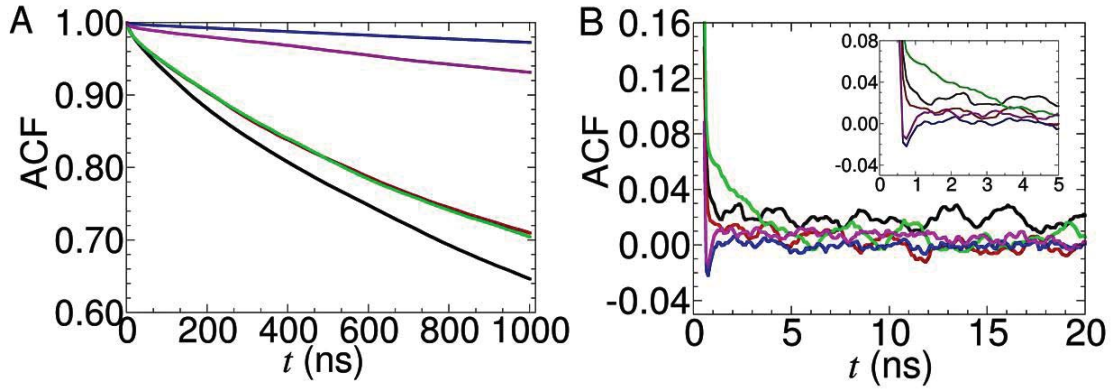

Figure 5.2: (A) ACFs of unrestrained orientational and polar angles CVs of the pig insulin dimer (unrestrained $\Delta G^{\circ}_{\mathrm{b}}$ equal to -8.5 kcal/mol). (B) Variation of the ACF of restrained orientational and polar angles CVs in the 20–ns separation PMF calculation following the geometrical route (restrained $\Delta G^{\circ}_{\mathrm{b}}$ equal to -7.0 kcal/mol). Θ, Φ, Ψ, θ, ϕ are colored in black, red, green, blue, and magenta, respectively. The first 5–ns variation of ACF is shown in the inset. Reproduced with permission from *J. Phys. Chem. Lett.*, **2022**, 13, 27, 6250–6258, Copyright 2022 American Chemical Society.

μs, respectively, when using fully unrestrained PMF calculations. Although this approach can yield a reasonably converged free-energy profile, reflecting adequate orientational and positional averaging, it is much less efficient than the geometrical route, which provides the correct answer for the same protein-protein complex in less than a microsecond timescale (880 ns). Hence, the shortcuts of the geometrical route presented in this study are not competitive with the full geometrical route.

5.1.2 Additional Details about Studied Complexes

MD simulations. All the simulations were performed under the same protocol as it was discussed in Chapter 3. In all the presented study cases of the shortcut of the geometrical route, the protein was pinned to the origin of the box, preventing it from drifting and tumbling. The data collecting step in the shortcut of the geometrical route was 2 ps for all complexes.

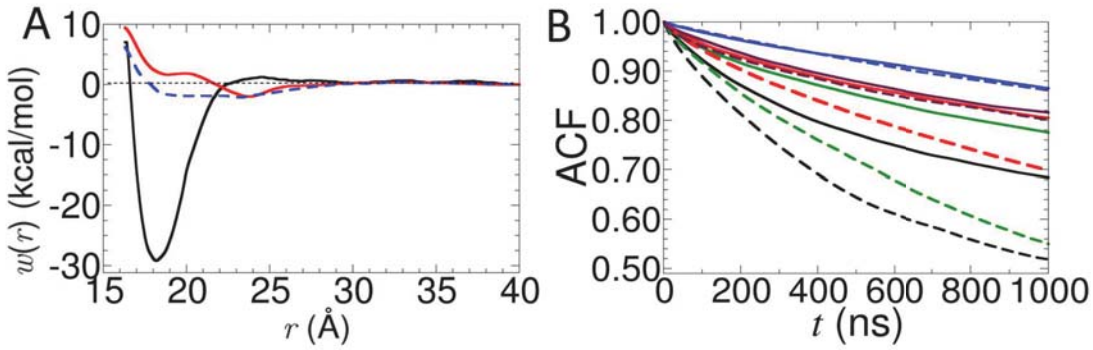

Figure 5.3: (A) Normalized 1–μs separation PMFs for insulin dimer calculated in the totally unrestrained simulation for two replicas (replica 1 in solid red and replica 2 in dashed blue) in comparison to the separation PMF of the geometrical route (black). (B) ACFs of unrestrained orientational and polar angles CVs of both replicas of the pig insulin dimer. Solid and dashed lines correspond to red and blue replicas in (A), respectively. Θ, Φ, Ψ, θ, ϕ are colored in black, red, green, blue, and maroon, accordingly. Reproduced with permission from *J. Phys. Chem. Lett.*, **2022**, 13, 27, 6250–6258, Copyright 2022 American Chemical Society.

Abl kinase-SH3:p41: Description of the Molecular Assembly. The Abl-Src homology 3 domain (SH3) is a crucial component of intracellular signaling pathways. When bound to its ligand, a proline-rich peptide called p41, the SH3 domain forms a stable complex with a binding free energy of −8.0 kcal/mol as determined by experimental measurements.[162] The SH3 domain adopts a β-barrel structure consisting of two anti-parallel three-stranded sheets, and the p41 peptide is partially buried within the Abl kinase-SH3 protein. The complex is stabilized by an intramolecular hydrogen bond between the side-chain of SER5 in the protein and the main-chain carbonyl group of PRO6 in the p41 peptide.[162,166,167] The structure of the complex has been resolved experimentally by X-ray diffraction to a resolution of 1.65 Å (Figure 5.4).

Computational details. The starting coordinates are from the PDB entry 1BBZ.[162] The obtained system consists of 73,054 atoms in total. The dimensions

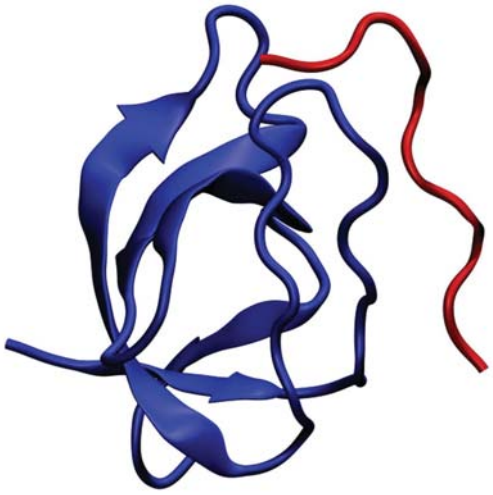

Figure 5.4: The Abl kinase-SH3:p41 structure (PDB: 1BBZ). In blue Abl-SH3 tyrosine kinase and in red peptide p41 are shown. Reproduced with permission from *Phys. Chem. Lett.*, **2022**, 13, 27, 6250–6258, Copyright 2022 American Chemical Society.

of the periodic cells were $94 \times 88 \times 96$ Å^3.

Results. The results of each contribution to the binding free energy of the Abl kinase-SH3:p41 complex in the geometrical route are published elsewhere and will not be presented here.[12,75,142] The physical separation PMFs obtained from the shortcut simulations do not reach a clear-cut shape (e.g., Figure 5.5A), and the evolution of the absolute binding free energy estimates is irregular and non-reproducible (Figure 5.5B). In contrast, the geometrical route provides a converged free-energy profile that mirrors adequate orientational and positional averaging, which makes it a more reliable method to estimate the binding free energy of this complex.

The ACF of the orientational Θ, Φ, and Ψ angles, as well as the polar θ and ϕ angles, for one of the replicas of the shortcut of the geometrical route for the Abl kinase-SH3:p41 complex is shown in Figure 5.6. The orientational Θ angle has an

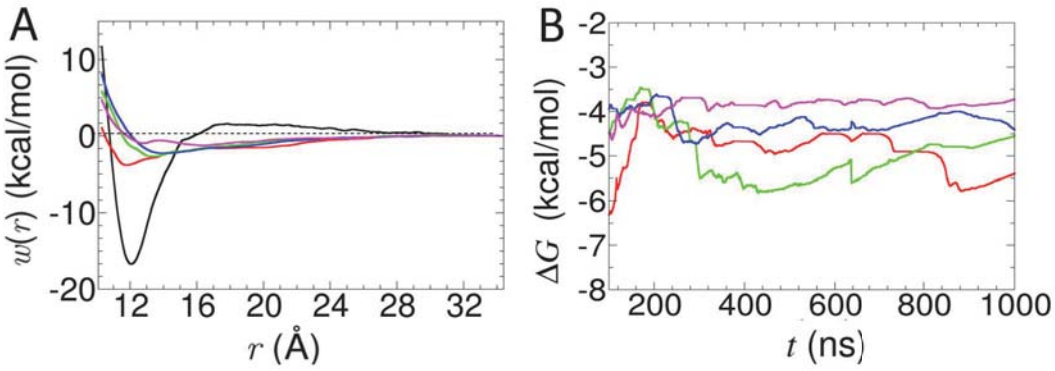

Figure 5.5: (A) Normalized separation PMFs of Abl kinase-SH3:p41 calculated in the shortcut of the geometrical route for four replicas (shown in red, green, blue, magenta) in comparison to the separation PMF obtained in the geometrical route (black). All the PMFs were obtained within the separation distance range of [10.3; 34.3] Å. (B) 1–μs evolution of binding free energy values of Abl kinase-SH3:p41 obtained in unrestrained computations for four replicas. Reproduced with permission from *J. Phys. Chem. Lett.*, **2022**, 13, 27, 6250–6258, Copyright 2022 American Chemical Society.

exponential decay of 0.35 units, whereas the positional θ angle has a slower decay,

reaching only 0.88 units at the end of the 1 μs–simulation.

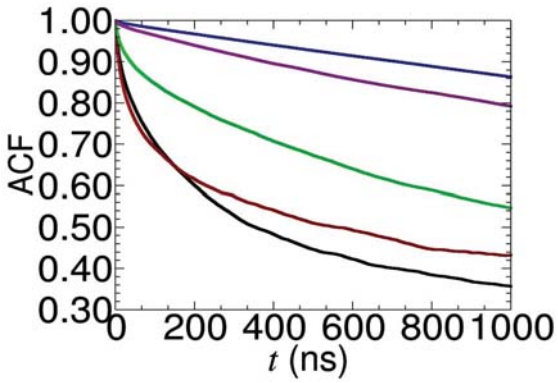

Figure 5.6: ACF of unrestrained orientational and polar angles CVs Θ, Φ, Ψ, θ, ϕ are colored in black, red, green, blue, and magenta, respectively for the red replica of Figure 5.5A. Reproduced with permission from *J. Phys. Chem. Lett.*, **2022**, 13, 27, 6250–6258, Copyright 2022 American Chemical Society.

MDM2-p53:NVP-CGM097: Results. The molecular assembly, the computational details, and the results of the standard binding free energy calculations within the geometrical route of this complex were discussed in Chapter 3.

The separation PMFs of the complex obtained in the geometrical route and its shortcut and the dynamics of ΔG along the shortcut simulation are compared in Figure 5.7. The ACF of Θ, Φ, and Ψ, and polar θ and ϕ angles for one of the replicas of the shortcut of the geometrical route is shown in Figure 5.8.

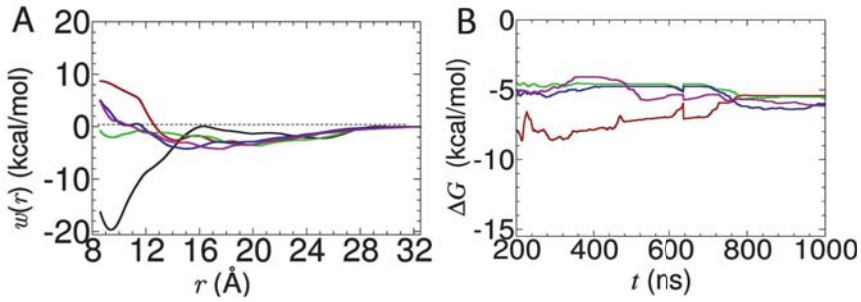

Figure 5.7: (A) Normalized separation PMFs of MDM2-p53:NVP-CGM097 calculated in the shortcut of the geometrical route for four replicas (shown in red, green, blue, magenta) in comparison to the separation PMF obtained in the geometrical route (black). All the PMFs were obtained within the separation distance range of [8.6; 32.6] Å. (B) 1–μs evolution of binding free energy values of MDM2-p53:NVP-CGM097 obtained in unrestrained computations for four replicas. Reproduced with permission from *J. Phys. Chem. Lett.*, **2022**, 13, 27, 6250–6258, Copyright 2022 American Chemical Society.

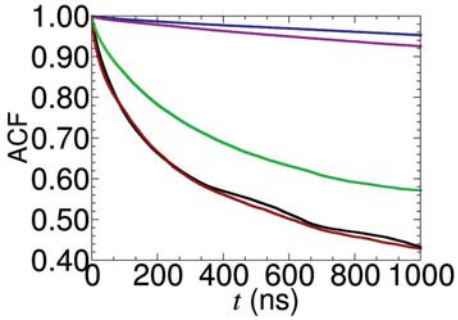

Figure 5.8: ACF of unrestrained orientational and polar angles CVs Θ, Φ, Ψ, θ, ϕ are colored in black, red, green, blue and magenta, respectively for the red replica of Figure 5.7A. Reproduced with permission from *J. Phys. Chem. Lett.*, **2022**, 13, 27, 6250–6258, Copyright 2022 American Chemical Society.

SARS-CoV-2 spike RBD:ACE2: Description of the molecular assembly. The SARS-CoV-2 virus uses its receptor-binding domain (RBD) located on the S1 spike protein to bind to the Angiotensin-Converting Enzyme 2 (ACE2) receptor in human cells. This binding is enabled by a network of hydrophilic interactions, including 13 hydrogen bonds and 2 salt bridges (between the oxygens of $LYS31_{RBD}$-$GLU484_{ACE2}$ and $Lys417_{RBD}$-$Asp30_{ACE2}$). The binding may also be influenced by N-glycans.[164,168–170] Inside the ACE2 protein, a zinc ion plays an important role in catalytic activity and coordinates with water molecules (Figure 5.9).[164] No restraints were used for the glycans and zinc ion in both the geometrical route and its shortcut for the separation of the SARS-CoV-2 spike RBD:ACE2 complex.

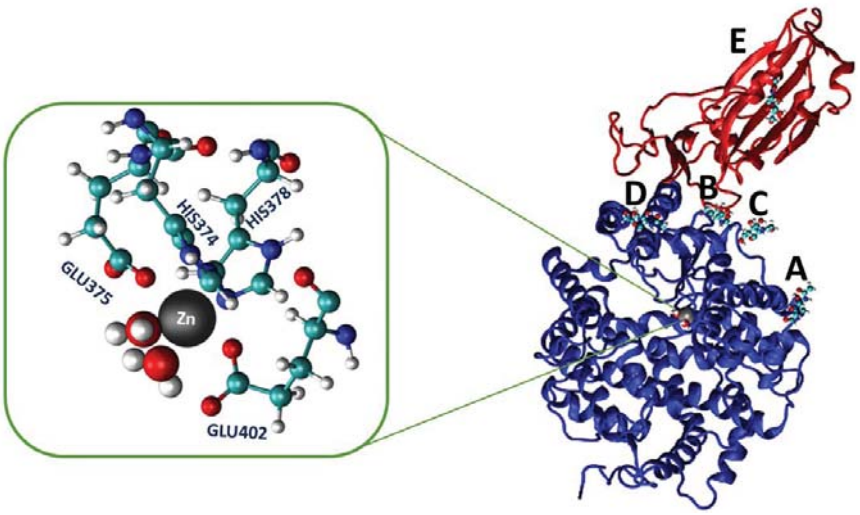

Figure 5.9: SARS-CoV-2 spike RBD in red in a complex with human ACE2 in (blue) (PDB: 6M0J). The N-glycans (2-acetamido-2-deoxy-β-D-glucopyranose) linked to the ACE2 are (A) N546, (B) N90, (C) N322, (D) N53 and (E) N-glycan N343 is linked to the RBD.[164] The Zn ion is shown in gray and coordinates with two water molecules and GLU402, HIS374, HIS378, and GLU375. In the green box, the enlarged coordination of the Zn ion is shown (two water molecules are shown in the VDW representation style). Reproduced with permission from *J. Phys. Chem. Lett.*, **2022**, 13, 27, 6250–6258, Copyright 2022 American Chemical Society.

Table 5.2: Results for each contribution to the binding free energy of the SARS-CoV-2 spike RBD:ACE2 in the geometrical route.

Contribution	PMF (kcal/mol)	PMF (ns)
$G^{\text{site}}_{c(RBD)}$	-7.0 ± 0.1	130
$G^{\text{site}}_{c(ACE2)}$	-6.8 ± 0.1	120
G^{site}_{Θ}	-0.3 ± 0.0	80
G^{site}_{Φ}	-0.1 ± 0.0	40
G^{site}_{Ψ}	-0.1 ± 0.0	40
G^{site}_{θ}	-0.3 ± 0.0	90
G^{site}_{ϕ}	-0.3 ± 0.0	50
$(1/\beta)\ln(S^*I^*C^\circ)$	-21.7 ± 0.0	120
$G^{\text{bulk}}_{c(RBD)}$	5.7 ± 0.1	200
$G^{\text{bulk}}_{c(ACE2)}$	12.8 ± 0.0	200
G^{bulk}_{o}	6.6	
$\Delta G^{\circ}_{\text{bind}}$	-11.5 ± 0.3 (calculation) -11.4 (experiment)[164]	Total: 1070 ns

Computational details. The starting coordinates were taken from the X-ray diffraction structure of the SARS-CoV-2 spike RBD:ACE2 resolved at 2.5 Å (PDB entry 6M0J).[164] All the procedures of the input file creation were similar to the pig insulin dimer discussed in the previous Chapter. The starting coordinates were obtained from the X-ray diffraction structure of the complex, and the procedures for creating the input files were similar to those for the pig insulin dimer. The system contained a total of 421,310 atoms, and the dimensions of the periodic cell were 144 × 156 × 198 Å^3.

Results. The results of the geometrical route are represented in Table 5.2, the PMFs of each contribution, and the convergence plots are presented in Figures 5.10 and 5.11.

The separation PMFs of the complex obtained in the geometrical route and its shortcut and the dynamics of $\Delta G^{\circ}_{\text{bind}}$ along the shortcut simulation are shown

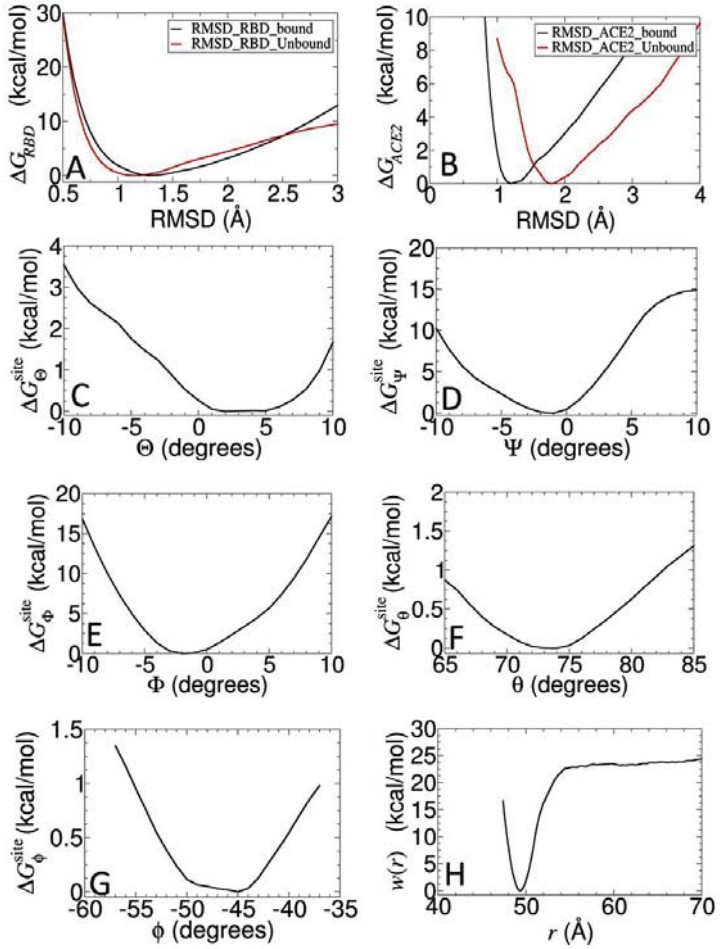

Figure 5.10: Individual PMFs for all components. The PMF calculations using RMSDs of the RBD domain in the bound and unbound state (A), the RMSDs of the ACE2 in the bound and unbound states (B), Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and centers-of-mass distance between two molecular entities (H), as the CV, respectively. Reproduced with permission from *J. Phys. Chem. Lett.*, **2022**, 13, 27, 6250–6258, Copyright 2022 American Chemical Society.

in Figure 5.12. The ACF of Θ, Φ, and Ψ, and polar θ and ϕ angles for one of the replicas of the shortcut of the geometrical route is shown in Figure 5.13.

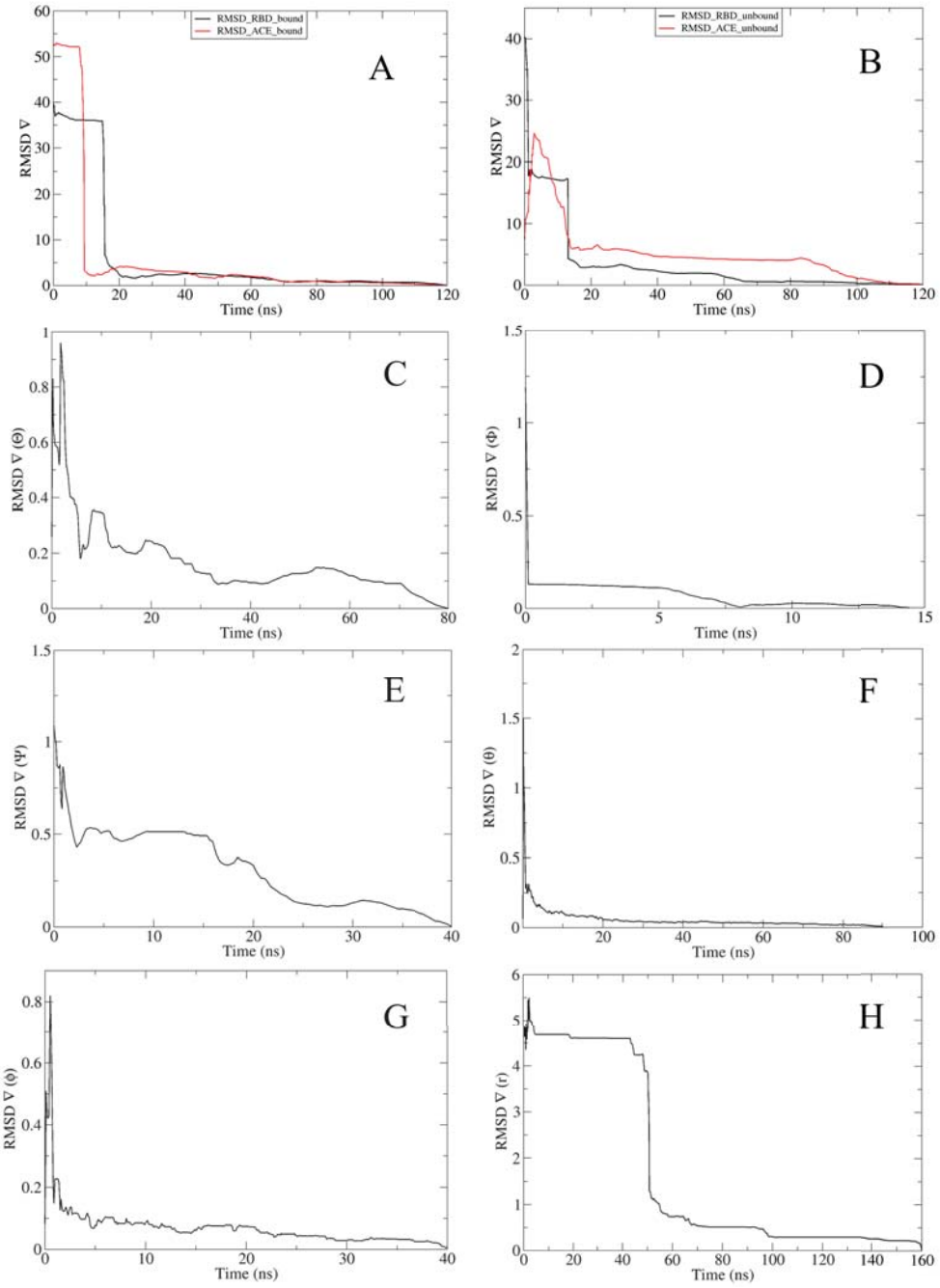

Figure 5.11: Convergence curves for individual PMFs for all components using RMSDs of the WT:RBD and ACE2 proteins of WT:RBD and ACE2 in bound (A) and unbound (B), Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the CV, respectively. Reproduced with permission from *J. Phys. Chem. Lett.*, **2022**, 13, 27, 6250–6258, Copyright 2022 American Chemical Society.

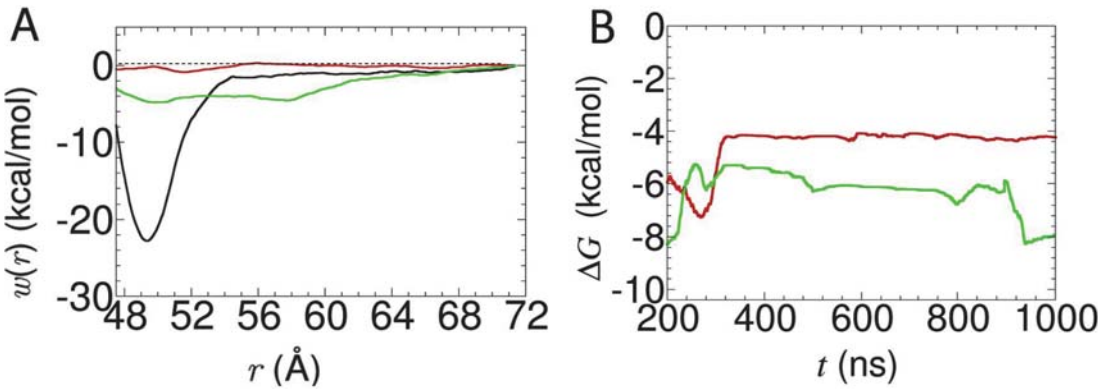

Figure 5.12: (A) Normalized separation PMFs for SARS-CoV-2 RBD:ACE2 complex calculated in the shortcut of the geometrical route for two replicas (the first replica is colored in red and the second one is in green) in comparison to the separation PMF obtained in the geometrical route (black). All the PMFs were obtained within the separation distance range of [47.4; 71.4] Å, (B) 1 μs evolution of binding free energy values of SARS-CoV-2 RBD:ACE2 complex obtained in unrestrained computations for two replicas (in red and in green, respectively). Reproduced with permission from *J. Phys. Chem. Lett.*, **2022**, 13, 27, 6250–6258, Copyright 2022 American Chemical Society.

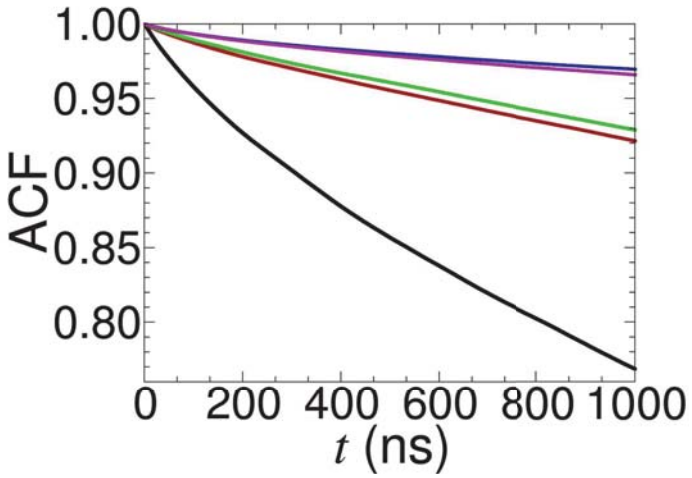

Figure 5.13: ACF of unrestrained orientational and polar angles CVs Θ, Φ, Ψ, θ, ϕ are colored in black, red, green, blue and magenta, respectively for the green replica of Figure 5.12A. Reproduced with permission from *J. Phys. Chem. Lett.*, **2022**, 13, 27, 6250–6258, Copyright 2022 American Chemical Society.

Chapter 6

Binding Affinity Calculations in the Context of COVID-19 Syndemic

6.1 Joining the COVID-19 Research Efforts

After demonstrating the robustness of the geometrical route for standard binding free-energy calculations for both protein-ligand and protein-protein complexes, my work delved into the investigation concerned with the worldwide syndemic of COVID-19,[171] the impact of which has been significant and resulted in numerous human lives lost and disrupted healthcare systems.[172–175]

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus responsible for COVID-19, has a complex structure comprising various proteins. Of particular interest is the spike (S) protein, which plays a pivotal role in viral entry into host cells.[176–178] The S protein can be divided into two subunits, S1 and S2. The S1 subunit encompasses the RBD that specifically binds to the human receptor ACE2, mediating viral attachment to host cells.[179,180] Conversely, the S2 subunit contains the necessary machinery for cell fusion and functions as an anchor to the host cell membrane.[181,182]

The perpetual mutational landscape of SARS-CoV-2 has given rise to a continuous emergence of novel worrisome variants, categorized as variants of concern (VOCs) by the World Health Organization (WHO), thereby augmenting the multifarious challenges encountered in the global battle against COVID-19. These notable variants have exhibited an increased transmissibility potential along with a concomitant reduction in their susceptibility to neutralization by antibodies. This confluence of disconcerting attributes has engendered justified apprehension regarding the efficacy of existent therapeutic modalities, vaccines, and diagnostic methodologies.[183–185] Mutations can potentiate the binding affinity between the RBD and the ACE2, thus facilitating viral entry into host cells.[176,179,180,186] Concordantly, these mutations can confer the ability to elude the immune response, impeding recognition by neutralizing antibodies or T-cell-mediated defenses.[169,187] Among the numerous of mutations, the D614G mutation stands out as a prominent example. This particular mutation has been observed to potentiate the proteolytic cleavage of the S protein, causing a more exposed RBD configuration and facilitating the substantive binding to the ACE2 receptor.[188,189] However, the ubiquity of this mutation across multiple variants, suggests that it alone cannot singularly account for the discernible disparities in fitness among these variants.

Due to the dynamic nature of the SARS-CoV-2 virus, employing standard binding free energy calculations is crucial to provide the quantitative assessment of the thermodynamics that underlies molecular association, stability, and specificity of interactions between SARS-CoV-2 variants and its target molecules. However, numerous early attempts at calculating the binding affinity of COVID-19-related

complexes were hindered by limitations in accuracy, primarily stemming from the reliance on incomplete structural data and the absence of sufficient sampling in oversimplified binding affinity calculation strategies.[155,190–194]

In our group, when we embarked on the binding affinity calculations following the geometrical route, a subset of experimental data and structural information for RBD (WT, Alpha, Beta, Delta, Omicron BA.2) in complexes with either ACE2 or antibodies (H11-D4, S2E12) were already available.[164,195] Our findings highlight the efficacy and potential of this approach in predicting and understanding the molecular interactions involved in COVID-19, offering valuable insights for further investigations and therapeutic strategies. This research work ended up with the following publication: Goulard Coderc de Lacam, E. et al., *J. Chem. Theory Comput.*, 2022.

6.2 Impact of Structural Data on Binding Free Energy Estimates

Results of the binding affinity calculations following geometrical route. In our study, we meticulously investigated a variety of protein complex structures both modeled (WT_{model}:ACE2, $Alpha_{model}$:ACE2, $Beta_{model}$:ACE2, $Delta_{model}$:ACE2, S2E12:$Delta_{model}$) and experimentally derived structures ($Beta_{Cryo-EM}$:ACE2, $WT_{crystal}$:ACE2, H11-D4:WT and Omicron BA.2:ACE2). Initially, we relied on modeled complexes due to the absence of corresponding experimental structures at the start of our research. However, subsequent studies

by Mannar et al.,[196] McCalum et al.,[197] and Yang et al.[198] provided the requisite experimental data and we compared our models with these newly available structures to verify the consistency of the altered interfaces across all the experimental data. In the case of Delta and Alpha variants, the interacting patterns were identical to modeled ones. However, in the case of $\text{Beta}_{\text{model}}$ and WT_{model}, we observed local rearrangements in the interface of the experimental structures, which were absent in our models.[164,196] To demonstrate the impact of the starting structure on the binding affinity using both models (WT_{model}, $\text{Beta}_{\text{model}}$) and experimental structures ($\text{WT}_{\text{crystal}}$, $\text{Beta}_{\text{Cryo-EM}}$), we decided to report the corresponding free-energy estimates (Table 6.1). Our investigation highlights the importance of appropriate structural data, which permits carefully recovering the binding free-energy estimates. It is noteworthy that accurate structure choice is not only an essential precondition to the geometrical route but, in general, to all MD-based binding free-energy strategies.

Besides, we also aimed to investigate the role of glycosylation in the binding of SARS-CoV-2 spike protein to the ACE2 receptor. The crystal structure of the $\text{WT}_{\text{crystal}}$:ACE2 complex used in our study was minimally glycosylated, with only one glycan present at four glycosylation sites on ACE2 and one site on the RBD.[164] Additional glycans were not resolved, likely due to their flexibility. To assess the effect of full-length glycans on binding, we repeated the binding free energy calculations using a fully glycosylated model of the complex as described by Acharya et al.[203] This model contained 8-10 glycans at each site. Our results showed a marginally smaller binding free energy of -10.8 kcal/mol compared to the

Table 6.1: Computed binding free energy of all studied complexes against experimental values obtained via surface plasmon resonance.

Complexes	ΔG°_{b} (kcal/mol)	ΔG°_{exp} (kcal/mol)
$WT_{crystal}$:ACE2	-11.5 ± 0.3	-11.4[164]
WT:ACE2 (full-length glycans)	-10.8 ± 0.3	-11.4[164]
WT_{model}:ACE2	-6.7 ± 2.3	-11.4[164]
$Alpha_{model}$:ACE2	-12.3 ± 1.2	-11.6 ± 0.1[199]
$Beta_{model}$:ACE2	-10.0 ± 1.2	-11.10 ± 0.06[199]
$Beta_{Cryo\text{-}EM}$:ACE2	-11.0 ± 1.6	-11.10 ± 0.06[199]
$Delta_{model}$:ACE2	-9.6 ± 0.5	-9.9[197]
Omicron BA.2:ACE2	-11.4 ± 1.3	-11.5[195]
S2E12:$Delta_{model}$	-12.5 ± 0.3	-12.0*[200]
H11-D4:WT	-9.4 ± 0.5	-9.9 ± 1.5[201]

My contribution to this work is marked in bold. Details about these complexes, except for $WT_{crystal}$:ACE2 discussed in Chapter 5, are shown in the next section.

*The experimental binding free energy was inferred from a neutralization curve reported by Mlcochova et al.,[202] where the Delta variant exhibits a behavior similar to the WT, thus justifying the use of the WT experimental value for the Delta variant in complex with S2E12.

minimally glycosylated model (-11.5 kcal/mol), which is in slight contrast to recent experiments, where glycans were found to contribute about +1 kcal/mol to the binding (-10.3 kcal/mol for the fully glycosylated complex, and -9.7 kcal/mol for that devoid of glycans).[204] Regardless, the calculated binding free energy for the fully glycosylated complex falls within the range observed experimentally, which spans $4k_BT$ (-9.0 kcal/mol,[205] -10.3 kcal/mol,[204] and -11.4 kcal/mol[164]).

Comparing the ΔG°_{b} of all VOCs in complex with ACE2, the Delta variant carries the lowest binding affinity to the receptor (-9.6 kcal/mol), being almost 2 kcal/mol weaker than the WT (see Table 6.1). Notably, experimental studies conducted by Mlcochova et al.[202] demonstrated that the Delta variant does not exhibit a stronger affinity for human ACE2 than either the Alpha variant (i.e., -11.6 kcal/mol)[199] or the WT, corroborating our findings. Therefore, drawing the conclusion that the Delta variant improves its fitness relative to other VOCs

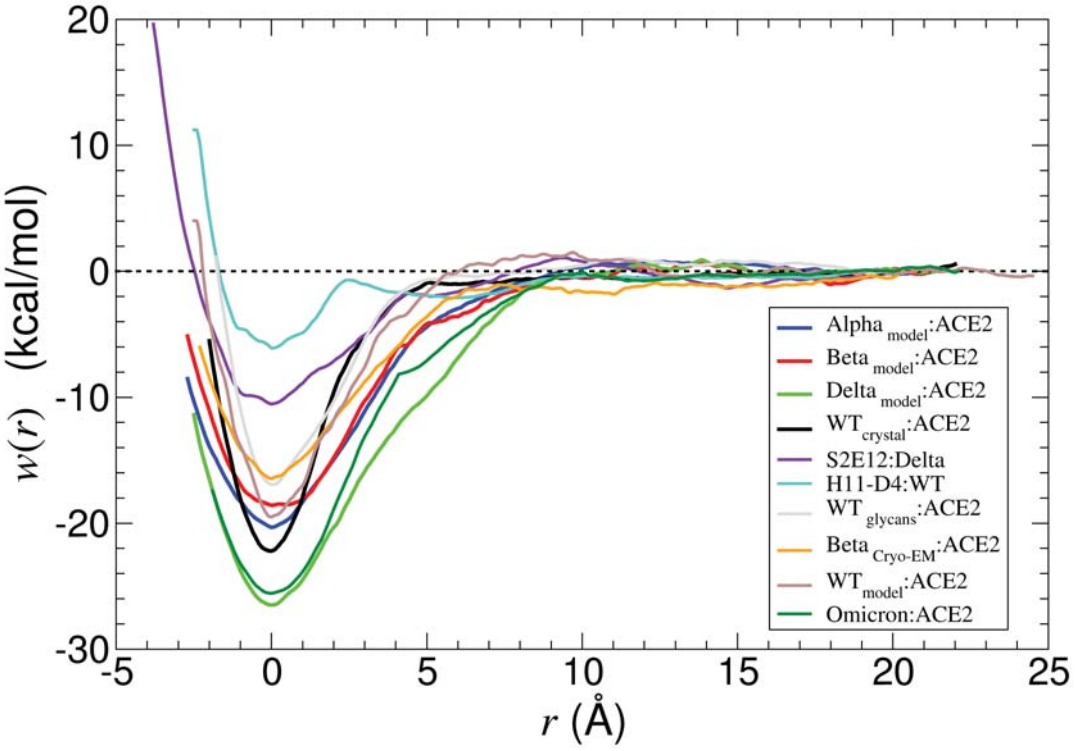

Figure 6.1: PMFs obtained during the reversible separation of ACE2 and the RBD of the WT (with minimal glycans in black, fully glycosylated in grey and the early model in taupe) and the different variants ($\text{Alpha}_{\text{model}}$: blue, $\text{Beta}_{\text{model}}$: red, $\text{Beta}_{\text{Cryo-EM}}$: orange, $\text{Delta}_{\text{model}}$: clear green, Omicron BA.2: dark green) or the RBD and antibodies (S2E12:Delta: violet, H11-D4:WT: cyan). All PMFs have been shifted so that the bound state is set to $r = 0$. Reproduced with permission from *J. Chem. Theory Comput.*, **2022**, 18, 10, 5890–5900, Copyright 2022 American Chemical Society.

by relying more on immune escape may explain the quick dominance of this variant over earlier VOCs, despite the higher vaccination rate amid the population.[206,207]

The separation PMFs for all the studied complexes obtained within the geometrical route are shown in Figure 6.1. Among all the VOCs:ACE2 PMFs, the Delta exhibited the deepest potential well, exceeding −26 kcal/mol, which is about 2.5 times the absolute value of the final $\Delta G_{\text{b}}^{\circ}$ estimate (i.e., −9.6 kcal/mol), underscore the significance of using all degrees of freedom other than the physical separation in the binding free-energy calculations. In the case of Delta, the PMFs

along the RMSDs of both proteins were found to contribute significantly to the binding affinity. Interestingly, the Alpha and Beta variants showed alike PMFs, having a difference of less than 2 kcal/mol in their depths. This finding could be attributed to the utilization of the same initial template and the N501Y mutation that is shared between the two variants. This mutation has been reported to significantly increase the binding affinity to ACE2 receptor in the literature.[186,208,209]

As a result of the binding affinity calculations of the complexes with antibodies, it was found that the S2E12 antibody has a strong affinity for the Delta variant of SARS-CoV-2, suggesting that S2E12 could be, in principle, a promising candidate for COVID-19 therapies. This observation aligns with the findings of Starr et al.,[210] who indicated that S2E12 exhibited binding affinity towards a broad spectrum of variants, highlighting its effectiveness. Specifically, in their study, it was found that the S2E12 is present in the sera of individuals who have been exposed to SARS-CoV-2 or vaccinated against it. Additionally, it was observed that the antibody does not exert strong selective pressure on the virus to mutate and evade its binding, which may contribute to the sustained effectiveness of S2E12 against a wide range of SARS-CoV-2 variants. Therefore, S2E12 represents a promising therapeutic candidate that could potentially maintain its efficacy against the virus, even with the emergence of new variants.[211] A recent study by Huang et al. demonstrated that among 50 monoclonal antibodies tested, S2E12 was one of only three antibodies that retained sufficient neutralizing properties against Omicron sub-variants (IC50 < 1 μg/mL),[212] confirming our statement.

In contrast, the binding affinity for the nanobody H11-D4:WT complex was

found to be −9.9 kcal/mol, which is lower than that of the $WT_{crystal}$:ACE2 complex (i.e., −11.5 kcal/mol), indicating an insufficient neutralizing activity of H11-D4 against the WT strain of SARS-CoV-2, unless used in significantly higher concentrations than ACE2. A similar conclusion was reached by Huo et al.,[201] who recommended the use of H11-D4 in combination with other antibodies that bind different regions of SARS-CoV-2 to enhance the treatment's effectiveness.

Protein-protein interaction networks. Since the geometrical route is a MD-based approach, we were also able to delve into the intricate details of the binding interactions between the various variants and ACE2 to gain a deeper understanding of the molecular underpinnings of the SARS-CoV-2 virus. Through a comprehensive analysis of the physical separation PMF trajectories, we discerned the occupancies of key hydrogen bonds and salt bridges crucial to the binding process for all studied complexes. Our findings reveal pronounced disparities across all cases, as highlighted by the bar plot (Figure 6.2). We observed that the binding between the RBD of the WT and ACE2 is mainly facilitated by the formation of salt bridges between D30:K417 and K31:E484, which is consistent with previous findings by Bhattarai et al.[213] When compared to the WT-strain, the Alpha variant showed a higher occupancy of these salt bridges than the Delta variant and the WT, which may explain its higher binding affinity to ACE2.[213] However, these salt bridges are absent in the Beta variant due to the K417N and E484K mutations, which may have been compensated for by the formation of novel interactions, such as a salt bridge between K484 and E75 (Figure 6.2B). This new interaction could explain why the loss of D30:K417 and K31:E484 did not lead to

a significant decrease in binding affinity for the Beta variant.[199,214]

Analyzing the hydrogen-bond interactions between ACE2 and the different VOCs in detail, and found strong discrepancies in occupancy when the partners are in intimate contact. For example, the occupancy of the E24:A475 hydrogen bond is strongly reduced in the different variants, with the WT having a population of 58% compared to 13%, 9%, 4%, and 16% for the Alpha, $\text{Beta}_{\text{model}}$, $\text{Beta}_{\text{Cryo-EM}}$, and Delta variants, respectively. The $\text{Beta}_{\text{model}}$ exhibits a hydrogen-bond pattern similar to that of its Alpha counterpart for the first three bonds, which is another argument in favor of the similarity of the PMFs and the use of the same initial template. However, the discrepant hydrogen-bond occupancy between the two Beta structures implies that misplaced side chains in the $\text{Beta}_{\text{model}}$ may result in a destabilized interface and explain the difference between their binding free energies.

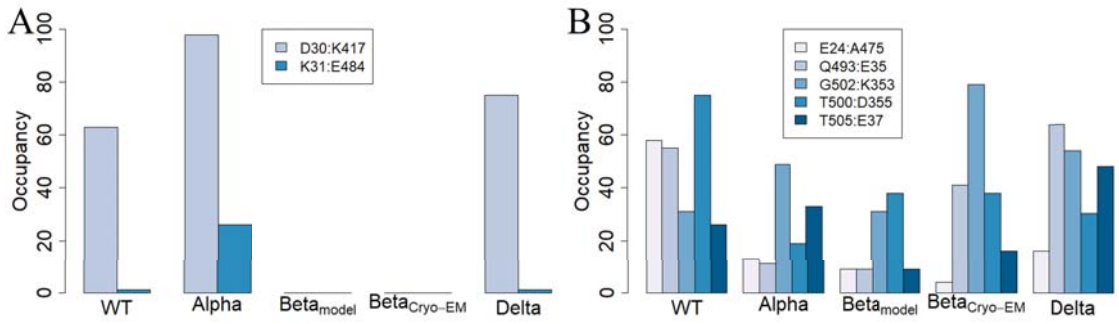

Figure 6.2: (A) Occupancy of the salt bridges in the separation trajectories for the WT ($\text{WT}_{crystal}$: ACE2) and the studied variants computed at a close center-of-mass distance (<50 Å). (B) Occupancy of the hydrogen bonds in the separation trajectories for the WT and the studied variants computed at a close center-of-mass distance (<50 Å). Reproduced with permission from *J. Chem. Theory Comput.*, **2022**, 18, 10, 5890–5900, Copyright 2022 American Chemical Society.

In contrast to all the discussed VOCs, the Omicron BA.2 variant, which has 16 mutations in the RBD alone, exhibits a unique binding mode that differs significantly from the WT (Figure 6.3). Although its standard binding free energy estimate is close to that of the WT (Table 6.1), it lacks the salt bridges observed in the WT and other VOCs. Instead, some hydrogen bonds in the Omicron BA.2 variant show its ability to maintain proper binding to ACE2 despite destabilizing mutations, such as K417N, suggesting a balance between beneficial and detrimental mutations.[196,215] While its numerical value is similar to that of the WT, its separation PMF, exhibits a deeper valley reminiscent of the Delta variant (Figure 5.10).[181,207,216–221] . Recent investigations by Willett et al.[222] and Carabelli et al.[223] further support our findings.

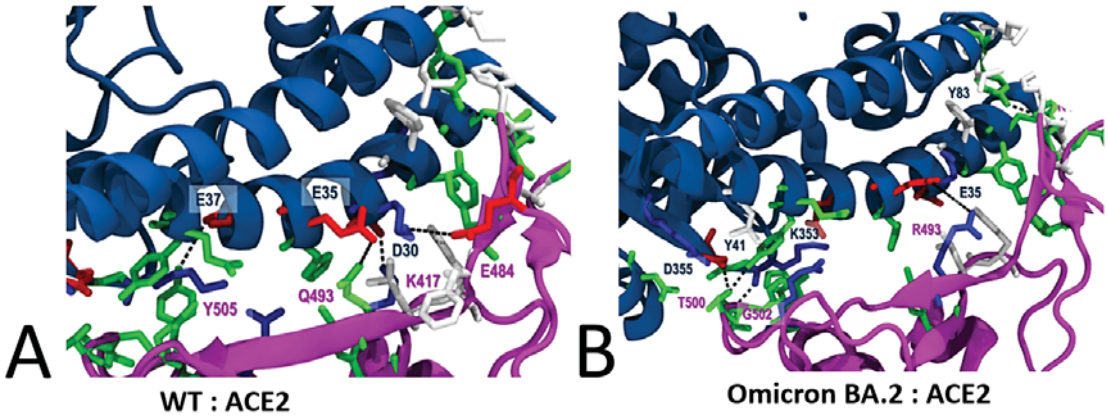

Figure 6.3: Interactions interface of (A) $WT_{crystal}$, (B) Omicron BA.2 variants with salt bridges and hydrogen bonds indicated by the dotted lines. Reproduced with permission from *J. Chem. Theory Comput.*, **2022**, 18, 10, 5890–5900, Copyright 2022 American Chemical Society.

6.3 Additional Details of the Studied Complexes

Omicron BA.2:ACE2: Molecular assembly details. The starting coordinates used to build the molecular assembly were taken from the PDB entry (7ZF7[195]) (Figure 6.4). The structure was then solvated, and an ionic concentration of 0.15 M was added to mimic physiological concentration resulting in a periodic cell of dimension 105 x 110 x 152 $Å^3$ and 167,397 atoms.

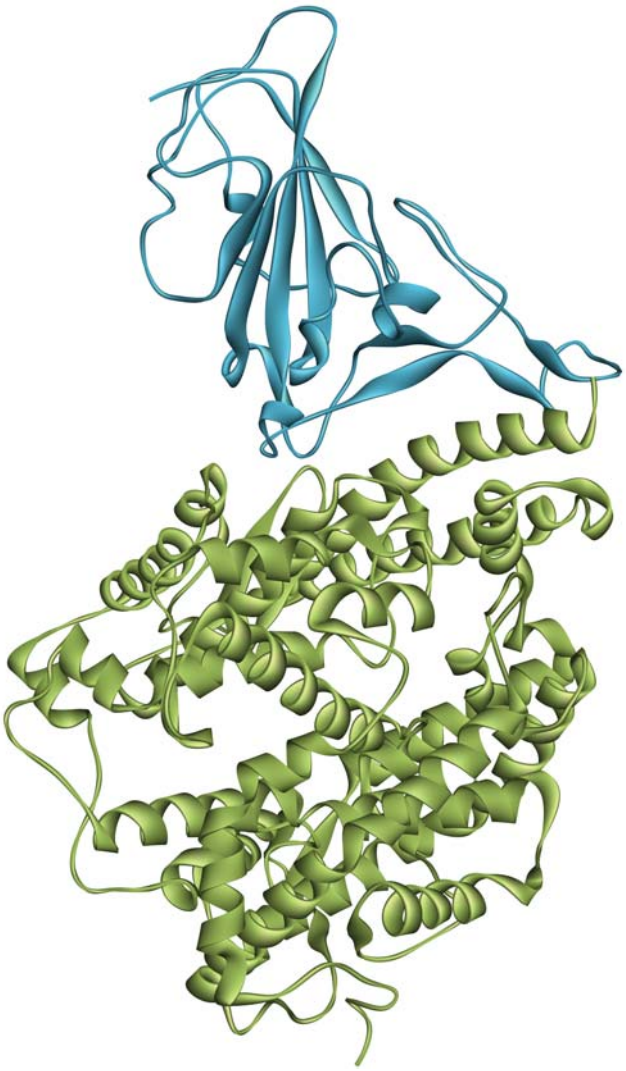

Figure 6.4: The Omicron BA.2:ACE2 structure: the S1 spike RBD and the ACE2 domains are shown in blue and green, respectively. The glycans are not shown but were included in the calculation.

Binding affinity calculation details. The detailed results of the diverse contributions are presented in Table 6.2 and individual components PMFs in Figure 6.5 with the associated convergence in Figure 6.6.

Table 6.2: Results for each contribution to the binding free energy of the Omicron BA.2 SARS-CoV-2 variant spike RBD:ACE2 in the geometrical route.

Contribution	PMF (kcal/mol)	PMF (ns)
$\Delta G^{\text{site}}_{c(\text{RBD})}$	-7.2 ± 0.3	185
$\Delta G^{\text{site}}_{c(\text{ACE2})}$	-8.0 ± 0.1	173
$\Delta G^{\text{site}}_{\Theta}$	-0.1 ± 0.0	30
$\Delta G^{\text{site}}_{\Phi}$	-0.3 ± 0.0	30
$\Delta G^{\text{site}}_{\Psi}$	-0.1 ± 0.0	30
$\Delta G^{\text{site}}_{\theta}$	-0.6 ± 0.0	40
$\Delta G^{\text{site}}_{\phi}$	-0.6 ± 0.0	30
$(1/\beta)\ln(S^* I^* C^\circ)$	-24.0 ± 0.2	280*
$\Delta G^{\text{bulk}}_{c(\text{RBD})}$	15.1 ± 0.3	250
$\Delta G^{\text{bulk}}_{c(\text{ACE2})}$	7.8 ± 0.4	220
$\Delta G^{\text{bulk}}_{o}$	6.6	
$\Delta G^{\circ}_{\text{b}}$	-11.4 ± 1.3 (calculation) $-$ 11.5[195]	1268

*Total simulation time required for the stratification by windows (four equal windows each 60 ns long and merge–simulation of 40 ns).

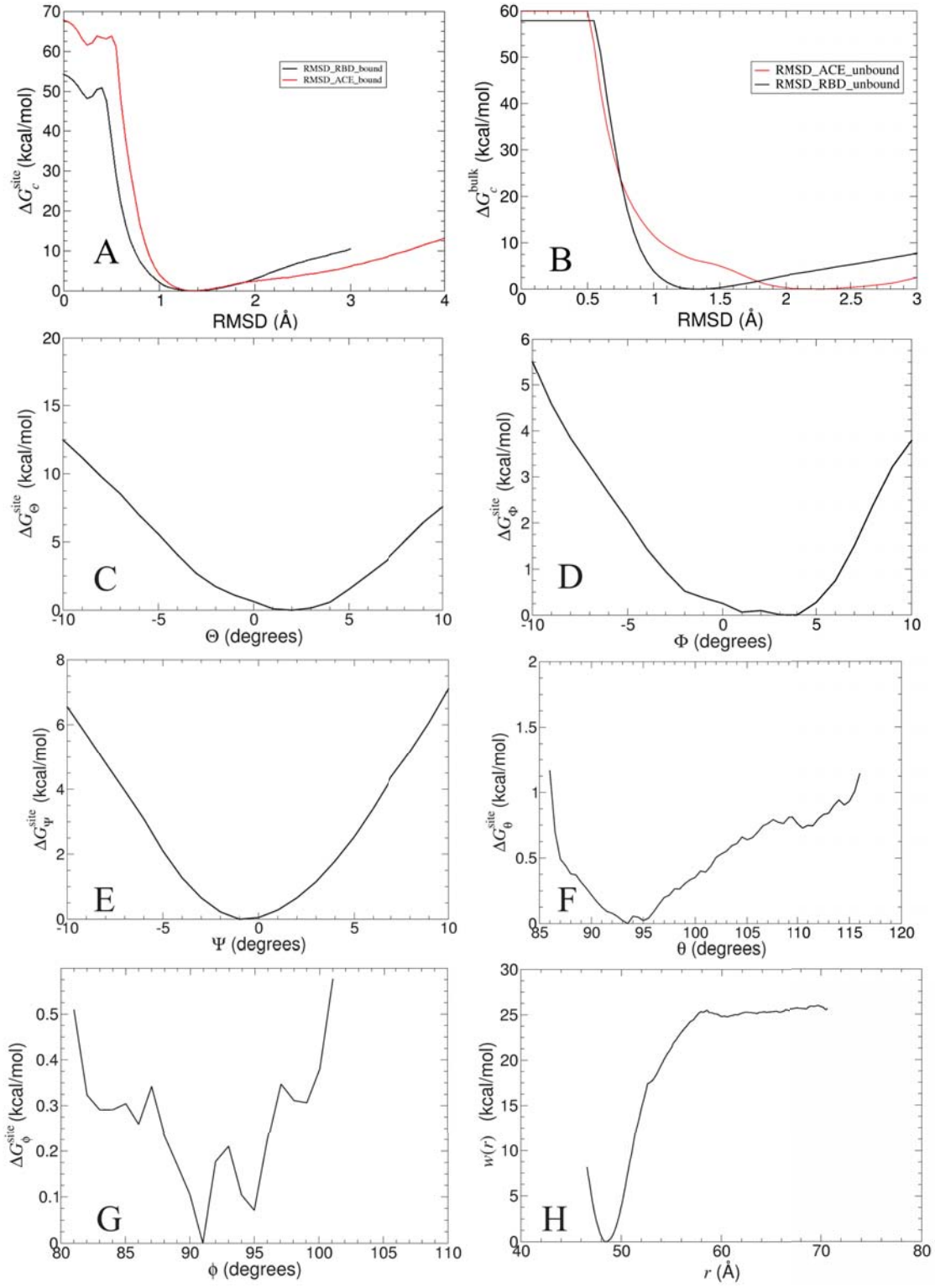

Figure 6.5: Individual PMFs for all components. The PMF calculations using RMSDs of the Omicron RBD and the ACE2 proteins in the bound (A), and unbound state (B), Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the COM distance between two molecular entities (H), as the collective variable, respectively. Reproduced with permission from *J. Chem. Theory Comput.*, **2022**, 18, 10, 5890–5900, Copyright 2022 American Chemical Society.

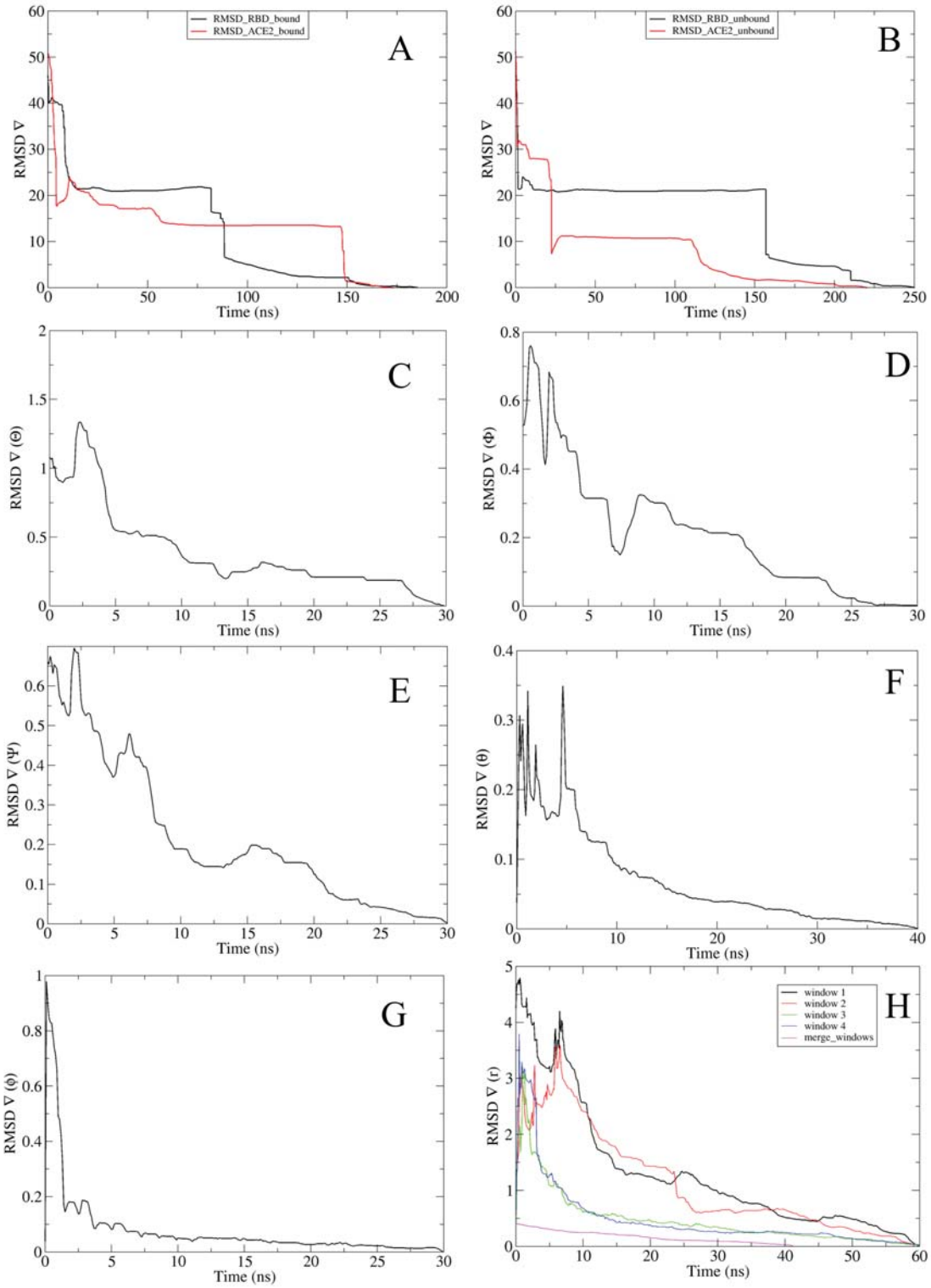

Figure 6.6: Convergence curve for individual PMFs for all components using RMSDs of the Omicron RBD and ACE2 proteins in the bound (A), and unbound state (B), Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the COM distance between two molecular entities (H), as the collective variable, respectively. Reproduced with permission from *J. Chem. Theory Comput.*, **2022**, 18, 10, 5890–5900, Copyright 2022 American Chemical Society.

H11-D4:WT: Molecular assembly details. The starting coordinates were taken from the crystallographic structure 6YZ5 resolved at 1.80 Å (Figure 6.7).[201] Glycans in the crystal structure were retained. The assembly was then solvated in a rectangular box with an ionic force of 0.15 NaCl, resulting in a periodic cell of 101 x 93 x 80 $Å^3$ and a total of 73,977 atoms.

Figure 6.7: The H11-D4:WT structure: the nanobody H11-D4 and the RBD domain are shown in brown and purple, respectively. The glycans are not shown but were included in the calculation.

Binding affinity calculation details. The detailed results of the diverse contributions are presented in Table 6.3 and individual components PMFs in Figure 6.8 with the associated convergence in Figure 6.9.

Table 6.3: Results for each contribution to the binding free energy of the WT SARS-CoV-2 variant spike RBD: H11-D4 in the geometrical route.

Contribution	PMF (kcal/mol)	PMF (ns)
$\Delta G^{\text{site}}_{c(\text{RBD})}$	-11.6 ± 0.1	220
$\Delta G^{\text{site}}_{c(\text{H11-D4})}$	-15.1 ± 0.1	220
$\Delta G^{\text{site}}_{\Theta}$	-0.3 ± 0.0	120
$\Delta G^{\text{site}}_{\Phi}$	-0.2 ± 0.0	120
$\Delta G^{\text{site}}_{\Psi}$	-0.3 ± 0.0	60
$\Delta G^{\text{site}}_{\theta}$	-0.7 ± 0.0	60
$\Delta G^{\text{site}}_{\phi}$	-0.1 ± 0.0	30
$(1/\beta)\ln(S^*I^*C^\circ)$	-4.3 ± 0.0	140
$\Delta G^{\text{bulk}}_{c(\text{RBD})}$	9.3 ± 0.1	160
$\Delta G^{\text{bulk}}_{c(\text{H11-D4})}$	7.3 ± 0.2	160
$\Delta G^{\text{bulk}}_{o}$	6.6	
$\Delta G^{\circ}_{\text{b}}$	-9.4 ± 0.5 (calculation) -9.9 (experiment)[201]	1290

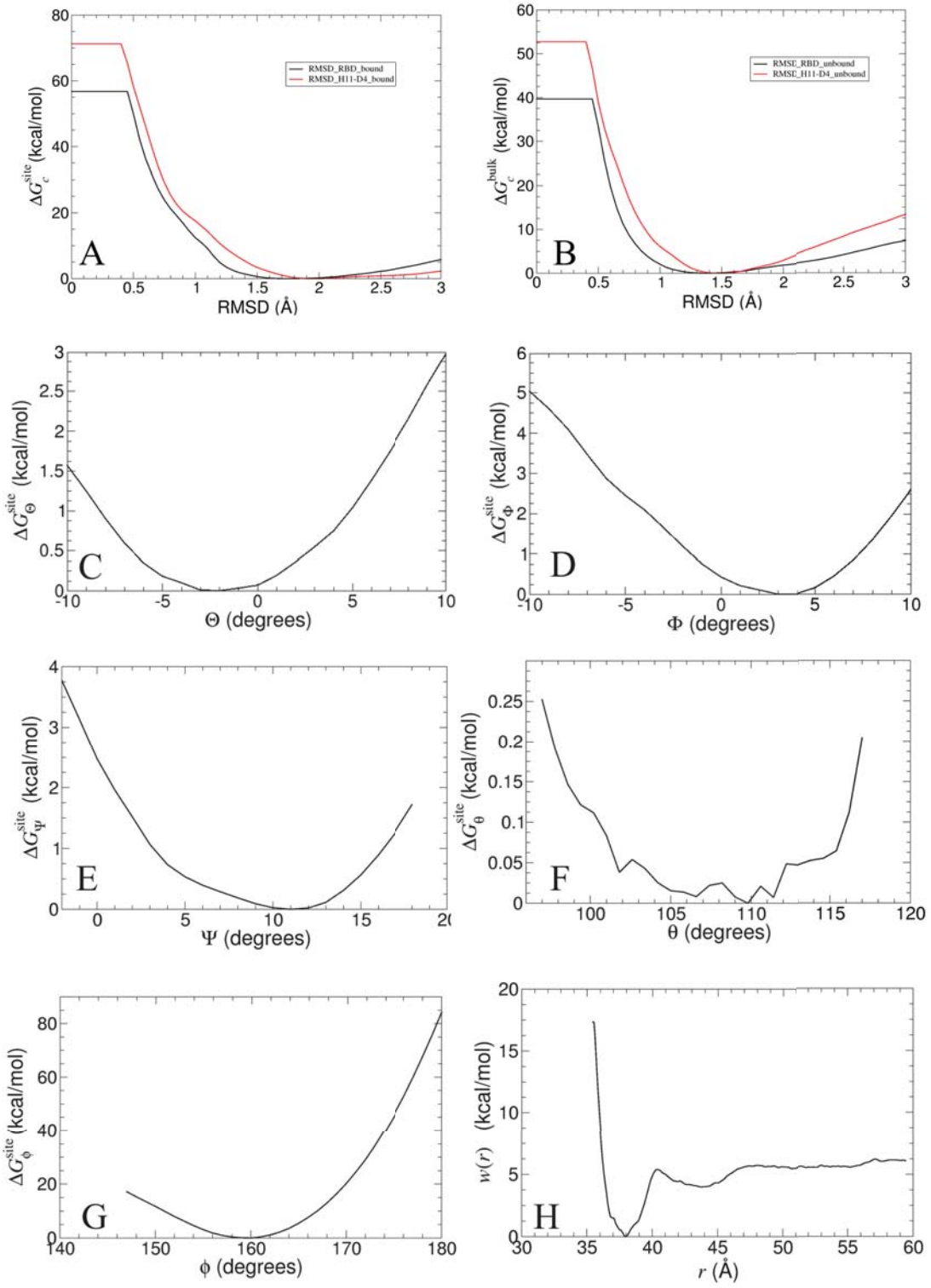

Figure 6.8: Individual PMFs for all components. The PMF calculations using RMSDs of the RBD and the H11-D4 chains in the bound (A), and unbound state (B), Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the COM distance between two molecular entities (H), as the collective variable, respectively. Reproduced with permission from *J. Chem. Theory Comput.*, **2022**, 18, 10, 5890–5900, Copyright 2022 American Chemical Society.

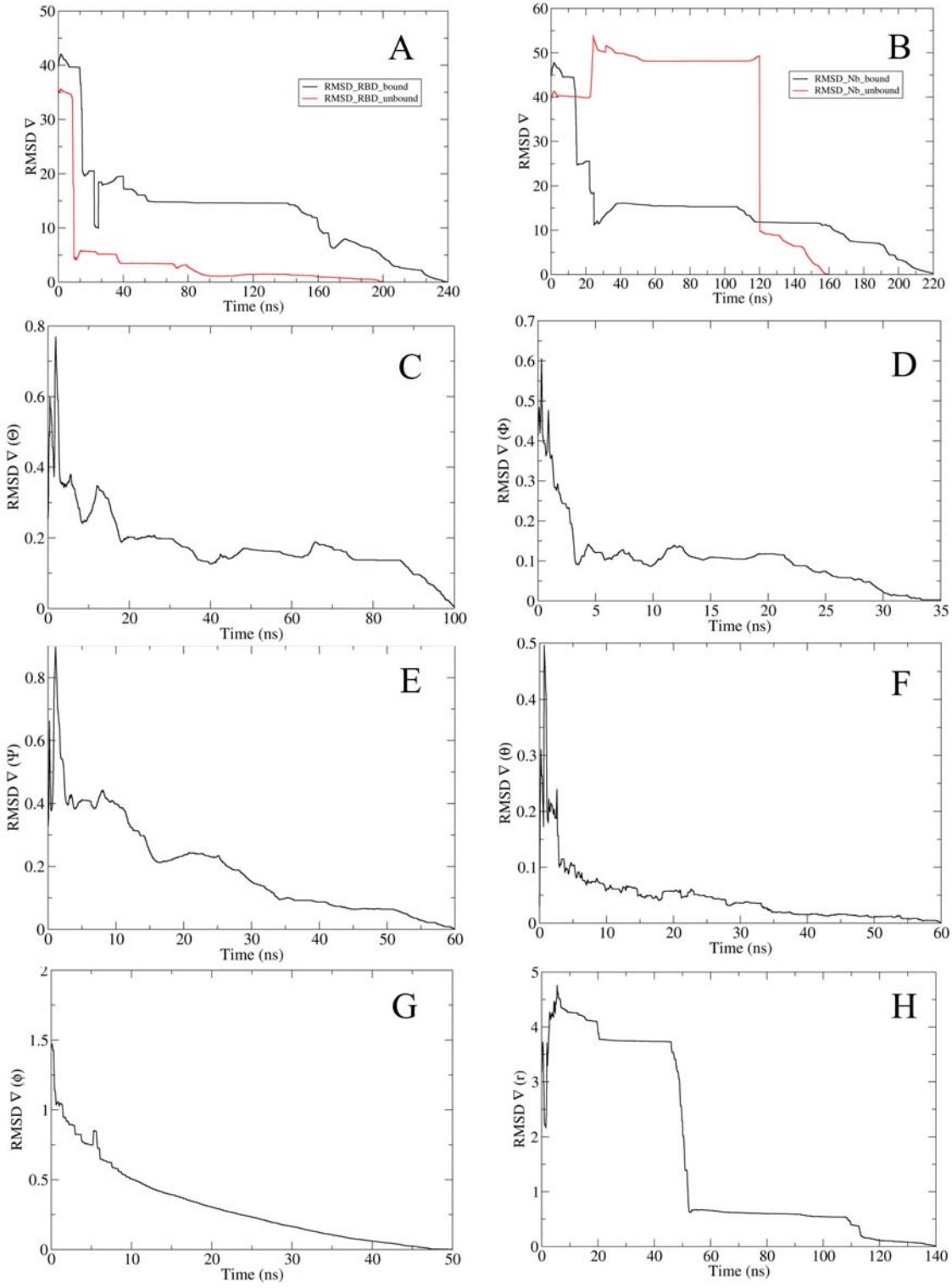

Figure 6.9: Convergence curve for individual PMFs for all components using RMSDs of the RBD and H11-D4 proteins in the bound (A), and unbound state (B), Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the COM distance between two molecular entities (H), as the collective variable, respectively. Reproduced with permission from *J. Chem. Theory Comput.*, **2022**, 18, 10, 5890–5900, Copyright 2022 American Chemical Society.

S2E12:Delta$_{\text{model}}$: Molecular assembly details. The starting coordinates were taken from the PDB structure of the WT RBD in complex with S2E12 (7R6X) resolved at 2.95 Å (Figure 6.10).[200] Point mutations L452R and T478K were introduced in the WT RBD to generate the Delta variant in complex with S2E12 by means of CharmmGUI tool.[128] The assembly was then solvated in a rectangular box with an ionic force of 0.15 NaCl, resulting in a periodic cell of 93 x 89 x 132 Å^3 and a total of 103,029 atoms.

Figure 6.10: The S2E12:Delta$_{\text{model}}$ structure: the antibody Fab S2E12 and the RBD domain are shown in green and purple, respectively. The glycans are not shown but were included in the calculation.

Binding affinity calculation details. The detailed results of the diverse contributions are presented in Table 6.4 and individual components PMFs in Figure 6.11 with the associated convergence in Figure 6.12.

Table 6.4: Results for each contribution to the binding free energy of the Delta SARS-CoV-2 variant spike RBD:S2E12 in the geometrical route.

Contribution	PMF (kcal/mol)	PMF (ns)
$\Delta G^{\text{site}}_{c(\text{RBD})}$	-8.2 ± 0.1	100
$\Delta G^{\text{site}}_{c(\text{S2E12})}$	-21.7 ± 0.1	100
$\Delta G^{\text{site}}_{\Theta}$	-0.4 ± 0.0	100
$\Delta G^{\text{site}}_{\Phi}$	-0.2 ± 0.0	60
$\Delta G^{\text{site}}_{\Psi}$	-0.4 ± 0.0	120
$\Delta G^{\text{site}}_{\theta}$	-0.6 ± 0.0	130
$\Delta G^{\text{site}}_{\phi}$	-0.3 ± 0.0	60
$(1/\beta)\ln(S^* I^* C^\circ)$	-8.0 ± 0.0	130
$\Delta G^{\text{bulk}}_{c(\text{RBD})}$	8.1 ± 0.1	100
$\Delta G^{\text{bulk}}_{c(\text{S2E12})}$	12.6 ± 0.2	100
$\Delta G^{\text{bulk}}_{o}$	6.6	
$\Delta G^{\circ}_{\text{b}}$	-12.5 ± 0.3 (calculation) -12.0 (experiment)[200]	900

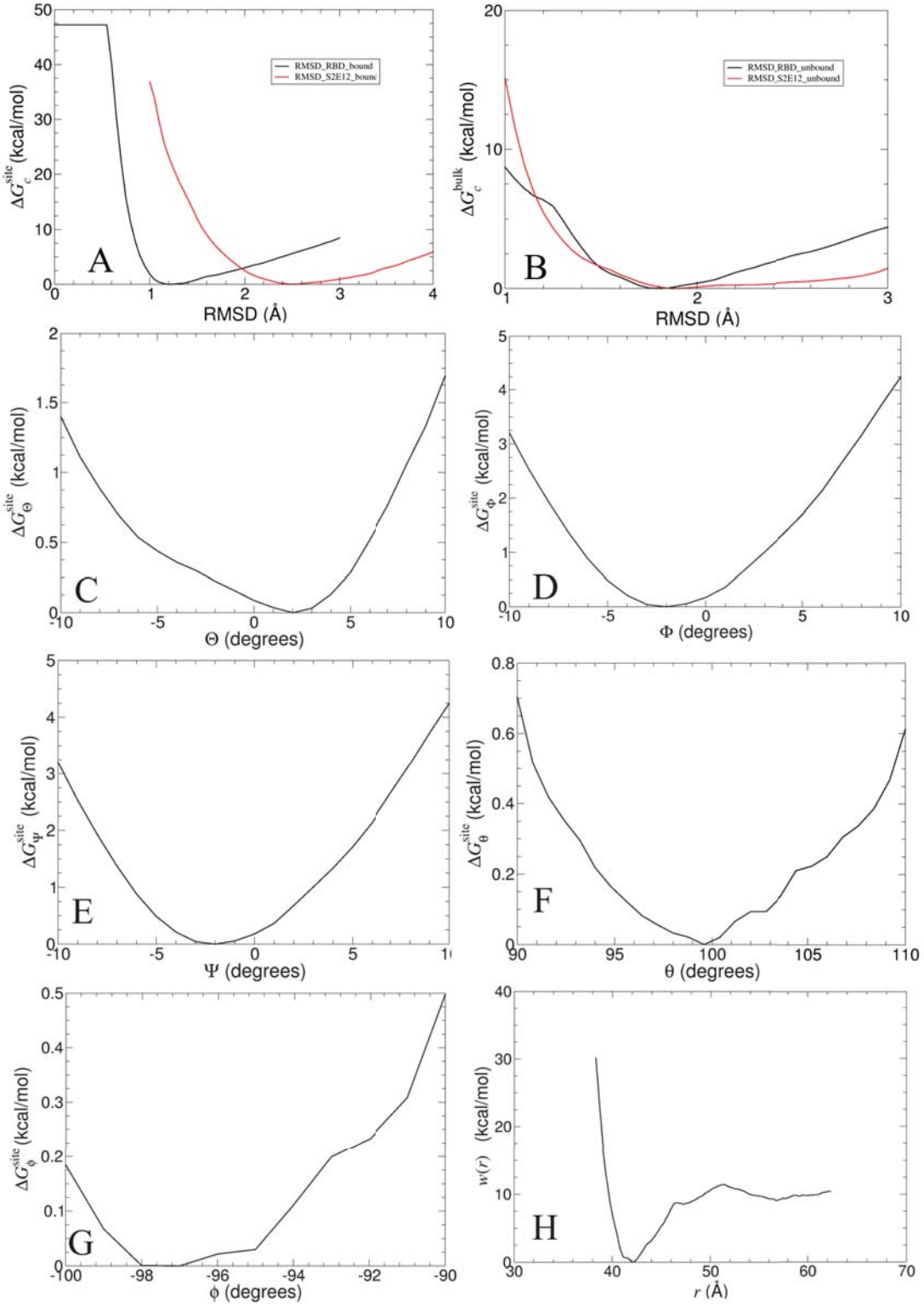

Figure 6.11: Individual PMFs for all components. The PMF calculations using RMSDs of the Delta RBD and the S2E12 chains in the bound (A), and unbound state (B), Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the COM distance between two molecular entities (H), as the collective variable, respectively. Reproduced with permission from *J. Chem. Theory Comput.*, **2022**, 18, 10, 5890–5900, Copyright 2022 American Chemical Society.

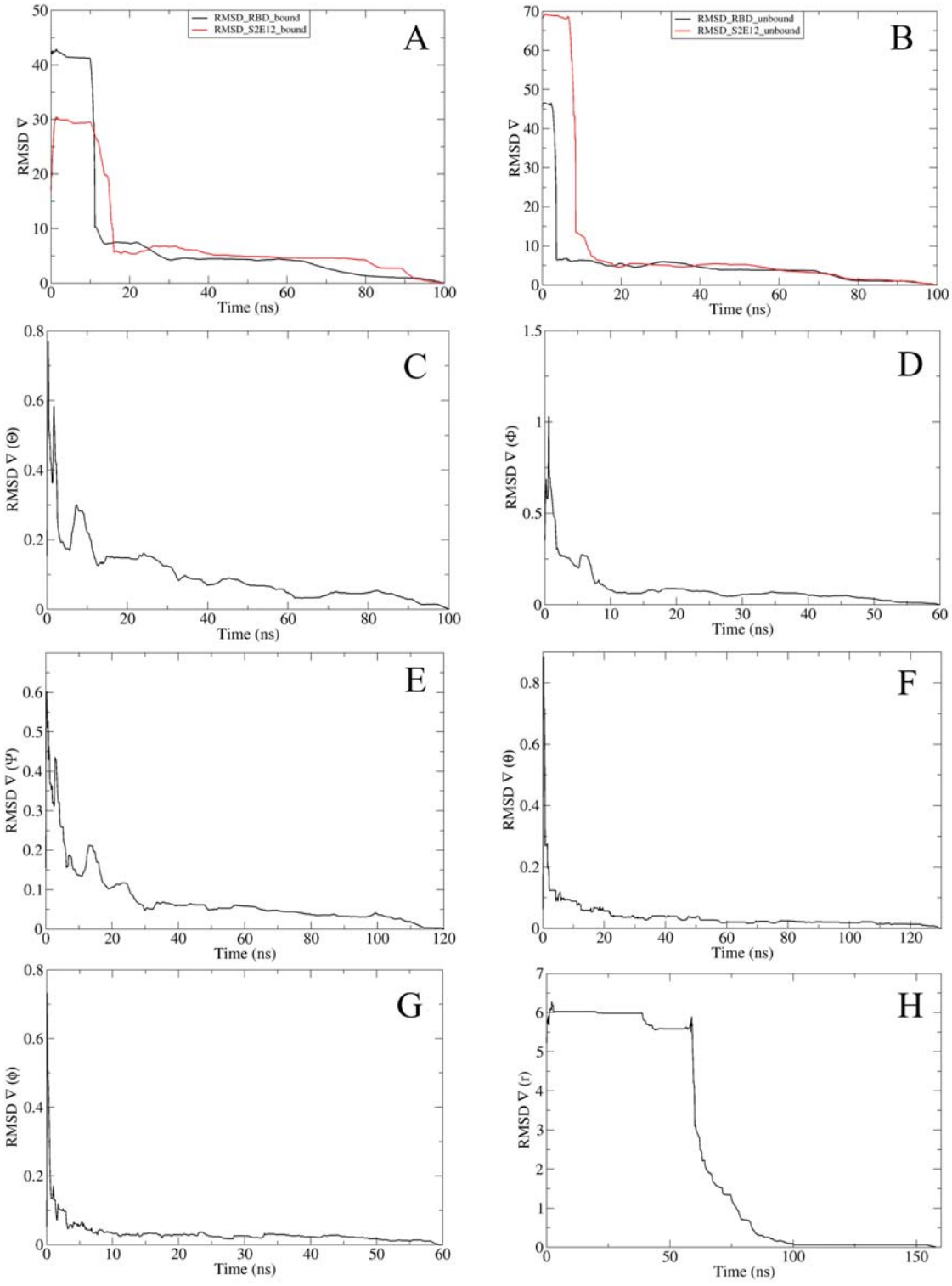

Figure 6.12: Convergence curve for individual PMFs for all components using RMSDs of the Delta RBD and S2E12 chains in the bound (A), and unbound state (B), Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the COM distance between two molecular entities (H), as the collective variable, respectively. Reproduced with permission from *J. Chem. Theory Comput.*, **2022**, 18, 10, 5890–5900, Copyright 2022 American Chemical Society.

Chapter 7

Protein-Protein Binding Affinities in a Membrane

In Chapter 1, I delved into the theoretical underpinnings of the geometric route in the isotropic aqueous environment. In this chapter, I aimed to illustrate how this approach can be adapted for the protein complexes in membranes. For this purpose, I selected the glycophorin A (GpA) homodimer as a model system for studying transmembrane (TM) protein binding. Composed of two α-helical segments, the GpA homodimer has been extensively studied in the literature, yet the fundamental aspects of its recognition and association mechanisms remained elusive, requiring further exploration. To unravel these intricate processes, it was imperative to develop a robust methodological framework that accounted for the symmetry of the GpA homodimer in a lipid bilayer. This study is currently submitted for publication. M. Blazhynska et al. *J. Chem. Theory Comput.*, 2023, and presented in this Chapter in its entirety.

7.1 Submitted Manuscript to J. Chem. Theory Comput.

A Rigorous Framework for Calculating Protein-Protein Binding Affinities in Membrane

Marharyta Blazhynska,[a] James C. Gumbart,[b] Haochuan Chen,[a] Emad Tajkhorshid,[c] Benoît Roux,[d, e] Christophe Chipot[a, c, d, f]

[a] Laboratoire International Associé Centre National de la Recherche Scientifique et University of Illinois at Urbana-Champaign, Unité Mixte de Recherche n°7019, Université de Lorraine, B.P. 70239, 54506 Vandœuvre-lès-Nancy cedex, France

[b] School of Physics, Georgia Institute of Technology, 837 State St., Atlanta, Georgia 30332, USA

[c] Theoretical and Computational Biophysics Group, NIH Center for Macromolecular Modeling and Visualization, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, 405 N. Mathews Ave, Urbana, Illinois 61801, USA

[d] Department of Biochemistry and Molecular Biology, The University of Chicago, 929 E. 57th Street W225, Chicago, Illinois 60637, USA

[e] Department of Chemistry, The University of Chicago, 5735 S Ellis Ave, Chicago, Illinois 60637, USA

[f] Department of Chemistry, The University of Hawai'i at Mānoa, 2545 McCarthy Mall, Honolulu, Hawaii 96822, USA

Abstract

Calculating the binding free energy of integral, transmembrane (TM) proteins is crucial for understanding the mechanisms by which they recognize one another and reversibly associate. The glycophorin A (GpA) homodimer, composed of two α–helical segments, has long served as a model system for studying TM protein reversible association. The present work establishes a methodological framework for calculating the binding affinity of the GpA homodimer in the heterogeneous environment of a membrane. Our investigation carefully considered a variety of protocols, including the appropriate choice of the force field, rigorous standardization reflecting the experimental conditions, sampling algorithm, anisotropic environment, and collective variables to accurately describe GpA dimerization via molecular dynamics-based approaches. Specifically, two strategies were explored: (i) an unrestrained potential mean force (PMF) calculation, which merely enhances sampling along the separation of the two binding partners without any restraint, and (ii) a so-called "geometrical route", whereby the α–helices are progressively separated with imposed restraints on their orientational, positional, and conformational degrees of freedom to accelerate convergence. Our simulations reveal that the simplified, unrestrained PMF approach is inadequate for the description of GpA dimerization. Instead, the geometrical route, tailored specifically to GpA in a membrane environment, yields excellent agreement with experimental data within reasonable computational time. The geometrical route further helps elucidate how environmental forces drive association, before helical interactions

stabilize it. Our simulations also brought to light a distinct, long-lived spatial arrangement, potentially serving as an intermediate state during dimer formation. The methodological advances in the generalized geometrical route provide a powerful tool for accurate and efficient binding-affinity calculations of intricate TM protein complexes in inhomogeneous environments.

Introduction

Accurate determination of the binding free energy of integral, transmembrane (TM) proteins is of paramount importance for apprehending the folding of individual α–helices, as well as their subsequent assembly into oligomers and folded proteins.[224–229] A conceptual framework for interpreting the formation of α–helical bundles in membranes is provided by the so-called two-stage model originally proposed by Popot and Engelman.[225,230,231] According to this model, TM helices oligomerize in two steps: (i) folding into independently stable TM α–helices, and (ii) formation of the α–helical bundle in the membrane. Glycophorin A (GpA) is a well-characterized model system for studying the association of TM helices, as it was the first clear example of a single-span membrane protein that dimerizes, investigated over thirty years ago by Bormann and Engelman.[232,233]

The first experimental structure of $GpA_{62–101}$ in dodecyl phosphocholine micelles was derived using solution NMR spectroscopy by MacKenzie et al.[234] This specific structure, identified as PDB 1AFO, was utilized in our study. Early investigations revealed that the non-covalent dimer of $GpA_{73–96}$, formed between

the membrane-spanning domains of $GpA_{62–101}$,[235,236] adopts an α–helical conformation.[234,237] The interactions between the protein and the hydrophobic environment of the membrane, mediated by the TM domain of $GpA_{75–98}$, play a critical role in the formation of the dimer.[234,236,238–248]

Another available structure of GpA, which was determined by Smith et al. in dimyristoyl phosphocholine, using solid-state NMR spectroscopy,[237,249] exhibits differences with that of MacKenzie et al.[234] Specifically, the interacting helical faces in the model of Smith et al. are rotated around their axes, resulting in a slightly reduced crossing angle of 35°,[237] compared to the 40° angle reported by MacKenzie et al.[234] For the latter, the experiment was conducted at 40°C, whereas for that of Smith et al. the temperature was −10°C. Furthermore, a more recent X–ray structure of GpA, obtained by Trenker et al. in a monooleic cubic-phase bilayer with a resolution of 2.81 Å,[250] aligns closely with the original findings of MacKenzie et al. Overall, the observed disparities in the experimental structures of the GpA dimer can be attributed to different environments and specific experimental conditions.

While there is a reasonable amount of structural information on GpA, the thermodynamic basis of its dimerization has yet to be determined.[251,252] In order to contextualize our results on binding free-energy calculations presented in the Results and Discussions section, we have compiled previous experimental and theoretical estimation for GpA dimerization in Table 7.1.

In addition to the aforementioned factors influencing the GpA structure,

Table 7.1: Reported experimental and computational values for the GpA dimerization free energy.

Experimental data				
GpA sequence	Medium	Methodology	ΔG_b°, kcal/mol	Standard concentration
TM domain	Pentaoxyethylene octyl ether	Ultracentrifugation	-9.0[253]	1 M
	Pentaoxyethylene octyl ether		-7.0[254]	
	Sodium dodecyl sulfate		-5.5 to -4.5[254]	
	C14 betaine micelle		-5.7[255]	
	Sodium dodecyl sulfate	FRET	-5.7[246]	
	Dodecyl dimethyl ammonium bromide		-3.8[246]	
	Dodecyl maltoside		-7.5[246]	
	Decyl maltoside		-6.6[246]	
	Dodecyl dimethyl amine oxide		-5.9[246]	
	Undecyl dimethyl amine oxide		-3.7[246]	
	Decyl dimethyl amine oxide		-4.1[246]	
	Escherichia coli inner membrane	GALLEX assay[256]	-7.5[257]	
	POPC bilayer	Steric trap	-12.1[258]	1 molecule/nm^2
	Plasma membrane	FRET	-3.9[259]	
Residues 75–98	Plasma membrane		-3.4 to -4.0[260]	
Computational data				
Residues 73–95	Dodecane	All-atom[a]	-11.5[261]	1 M
Residues 69–97	POPC bilayer		-3.0 to -3.8[262]	1 molecule/nm^2
		Coarse-grained[a]	-8.4[263]	
			-5.9[252]	
			-9.1,-9.3,-7.5[264]	
TM domain	DPPC bilayer		-9.1[265]	
Residues 70–96	DPPC bilayer		-7.5[251]	
	DLPC bilayer		-6.7[251]	
	DOPC bilayer		-6.6[251]	

[a] type of force field used in the simulations.

other aspects, including, but not limited to, the sequence length of the protein and the choice of the standard concentration, may have significantly contributed to the variations in the reported experimental binding free energies, $\Delta G_{\mathrm{b}}^{\circ}$.[246,251,253,259,260,266,267] With regard to the standardization, different research groups[253–255,258–260,266–272] have used various concentration units to infer $\Delta G_{\mathrm{b}}^{\circ}$ for protein complexes in anisotropic media.[246,253,258] Due to the gamut of standards employed, the reported free energies of GpA dimerization in similar media may differ markedly, making it challenging to precisely compare and interpret the reported results.[267,268] For instance, Nash et al.[257] and Chen et al.[259] studied experimentally the association of the GpA α–helices in biological membranes, applying different standard concentrations for the $\Delta G_{\mathrm{b}}^{\circ}$ estimations. While Nash et al.[257] used 1 M as the reference concentration for the $\Delta G_{\mathrm{b}}^{\circ}$ estimation, Chen et al.[259] used a standard state of 1 molecule/nm^2. As a result, the reported values for the standard free energy of GpA dimerization from these studies differ by more than 3.5 kcal/mol (-7.5 kcal/mol[257] versus -3.9 kcal/mol[259]). It should be noted that the application of standard volume concentration units, such as 1 M, for the TM complexes becomes questionable owing to the restriction of the membrane to a two–dimensional space.[258–260,266,272] In contrast, standard surface concentration, expressed as molecules per unit area,[258–260] acknowledges its intrinsic dimensionality, while offering a more accurate depiction of the spatial organization and intermolecular interactions within the membrane milieu.

Various experimental techniques, e.g., FRET,[246,259,260] steric trap,[258] GALLEX assays,[257] and analytical ultracentrifugation,[253–255] have been used to determine

ΔG_b° for the GpA dimerization in different anisotropic media, leading to a variety of experimental estimates ranging from $-3.4\,\mathrm{kcal/mol}$ to $-12.1\,\mathrm{kcal/mol}$.[246,253,254,258–260] Furthermore, the choice of the solvent may affect protein insertion into the membrane, helix topology, and protein stability.[273,273–279] Table 7.1 reveals a trend wherein increasing the hydrophobicity of the membrane-like environment correlates with less favorable GpA dimerization, as reflected in the increasing values of ΔG_b° from the POPC bilayer ($-12.1\,\mathrm{kcal/mol}$[258]) to detergent micelles (e.g., $-3.8\,\mathrm{kcal/mol}$ in dodecyl dimethyl ammonium bromide[246]). The impact of the detergent chain length on the strength of GpA dimerization was also investigated by Fisher et al.[246] Using nonionic maltoside and zwitterionic dimethyl amine oxide detergents, they showed that the interactions of the GpA α–helices are weaker in the presence of shorter-chain detergents.[246] In the context of membrane environments, Hong et al. proposed that the discrepancies in the experimental ΔG_b° values obtained from model POPC bilayers and natural membranes might stem from the probability of the GpA monomers to find competitive binding partners in natural membranes, stabilizing the dissociated forms and hindering GpA dimer formation.[258,280]

While experimental studies have provided valuable insights, computational approaches furnish powerful complementary tools for the calculation of the binding constants for GpA helices in a lipid environment.[251,252,261–265] However, computational studies also yield diverse results (see Table 7.1). In addition to the discussed factors influencing GpA binding affinity in experiments, computational approaches require careful consideration of various aspects to ensure accurate binding free-

energy calculations. They include reproducing the membrane-like environment, as well as selecting suitable force fields, sampling algorithms, and methodological approaches for standard binding free-energy calculations.[251,252,261–264,275,281–283] An appropriate force field is particularly crucial, as it directly impacts the energetics and dynamics of GpA–lipid interactions, thus influencing the reliability of the computations.[284,285] Among the GpA dimerization studies based on all-atom force fields,[261,262] the work of Domański et al.[262] calls into question the ability of the CHARMM36 force field[286,287] to reproduce the stability of the native GpA dimer in POPC membranes, and suggests the use of a dispersion correction for protein–protein interactions. The investigation by Balusek et al.,[288] on the other hand, supports the reliability of CHARMM36 for GpA dimerization. However accurate for membrane protein complexes,[135,287,288] all–atom force fields have, nevertheless, seldom been used in GpA dimerization calculations due to their computational cost. While coarse-grained force fields are computationally less expensive,[251,252,263–265] their ability to faithfully capture the molecular interactions involved in GpA dimerization is limited by their simplified representation of the system at hand.[264] For example, there is a potential risk of promoting excessive nonspecific protein aggregations in coarse-grained simulations, which may affect the reliability of the theoretical predictions.

Furthermore, it is essential to recognize that certain computational techniques used for evaluating $\Delta G_{\mathrm{b}}^{\circ}$ values for membrane proteins neglect the anisotropic properties of the surrounding environment.[252,262,263] By overlooking the membrane symmetry, these approaches can lead to inaccurate estimates of binding

free energy,[261] complicating the thermodynamic characterization of GpA dimerization. Overall, both experimental and computational approaches have their own complexities when determining and comparing $\Delta G_{\mathrm{b}}^{\circ}$. These data, therefore, ought to be interpreted with utmost caution.

In the present work, we lay out a methodology for the standard binding free-energy calculations for GpA homodimerization in a membrane-like environment. We combine two strategies: (1) unrestrained physical separation potential-of-mean-force (PMF) calculations along the distance between the centers of mass (COMs), or COM-to-COM distance, one of the most abundantly used approaches for GpA dimerization studies,[252,261,262,265] and (2) the so-called "geometrical route" approach for free-energy calculation, generalized for membrane proteins, which involves physical separation of the two proteins while imposing restraints on their configurational changes.[12,29,46,142] With the application of the tailored geometrical route, we emphasize the importance of adequately describing the slow degrees of freedom involved in the dimerization process and controlling them by applying restraints along suitably chosen collective variables (CVs),[67] whose contributions can be evaluated numerically. The application of such restraints on the CVs suitably reduces the configurational space that needs to be sampled. The introduced CVs handle orientational, positional, and conformational movements. Although the membrane environment naturally restrains the relative orientation of the α–helices upon binding,[289–293] introducing geometric restraints on these CVs is imperative to accelerate the convergence of the binding free-energy calculations within an accessible computational time.

Contrary to the geometrical route, in unrestrained PMF calculations, where the conformational, orientational, and positional (through the polar and azimuthal angles) restraints are absent, the configurational space available to the dimer is no longer reduced. Covering all possible degrees of freedom during the physical separation of the α–helical segments, therefore, requires extensive sampling of all possible configurations, and, hence, substantial simulation times to ensure adequate convergence.[74] It should be noted that even with the application of the most advanced enhanced-sampling algorithms,[90] in the unrestrained PMF approach, the complexity of GpA dimerization in the membrane cannot be fully captured. Apart from its technical advance, the geometrical route offers valuable insights into the mechanism of recognition and association of the GpA TM segments. The separation simulation following this approach reveals that the lipids exert a force that brings the α-helices closer together, prompting inter-helical interactions, which is in agreement with the two–stage model of membrane protein folding and oligomerization.[225,230,231,294] Our results show that GpA dimerization, although appearing simple at first glance, involves (i) a number of binding stages prior to the fully associated state, and (ii) an intermediate state, as was observed during the physical separation simulation, following the geometrical route.

Methods

We consider a dilute system of membrane-bound receptors R and ligands L that can associate in a biomolecular fashion. While both the receptor and the ligand could be proteins or other membrane constituents, we adopt this notation

for the sake of clarity. Classically, the equilibrium constant K_{eq} of the binding process, $\mathrm{L} + \mathrm{R} \rightleftharpoons \mathrm{RL}$, is defined as a function of the concentrations of each species, [RL], [L], and [R], at equilibrium as $K_{\mathrm{eq}} = [\mathrm{RL}]/[\mathrm{L}][\mathrm{R}]$. However, in the case of membrane-confined processes, such as GpA dimerization studied here, the two components evolve within the two-dimensional phase of a membrane, and their concentrations might be best represented as the number of moles per area of the membrane. Without loss of generality, we consider a single receptor in a fixed orientation with its COM held at the origin surrounded by a solution of ligands. We can express the equilibrium binding constant as,

$$K_{\mathrm{eq}} = \frac{\displaystyle\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta U}}{\displaystyle\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\ \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int \mathrm{d}\mathbf{X}\ \mathrm{e}^{-\beta U}}. \tag{7.1}$$

where U is the total potential energy of the system, $1/\beta = k_{\mathrm{B}}T$ is the Boltzmann constant times temperature, and $\mathbf{1}$ and $\mathbf{X}$ represent the degrees of freedom of the ligand $\mathbf{1}$, and the remaining atoms (solvent, lipid, or protein), respectively. The subscripts "site" and "bulk" in the integrals indicate the relevant spatial regions of the configurational space to be included in each integration, representing the bound and unbound states. Here, "bulk" means somewhere within the two-dimensional membrane. Vector $\mathbf{r}_1 \equiv (x_1, y_1)$ is the position of the COM of ligand $\mathbf{1}$ in the two-dimensional bulk region, and $\mathbf{r}_1^* = (x^*, y^*)$ is some arbitrary (fixed) location in this region, far away from the receptor.

The denominator and the numerator of eq. 7.1 represent, respectively, the initial and final states of the binding process, that is the ligand bound to the

receptor, and the ligand with its COM at $\mathbf{r}_1^*$ in the bulk (note that all coordinates are expressed relative to the COM of the receptor). The central idea in molecular-dynamics-based standard binding free-energy calculations is to insert a series of convenient intermediate states between the initial and final states, in order to render the computation feasible.

The main practical challenge in the case of membrane proteins is the selection of appropriate CVs along which the $\Delta G_\mathrm{b}^\circ$ can be accurately estimated.[8] While the evaluation of the free energy along the COM-to-COM distance, just like in the unrestrained physical separation PMF approach, is a straightforward choice,[261,265] it leaves all degrees of freedom orthogonal to the separation distance uncontrolled. As a consequence, the system has access to a vast configurational space, encompassing all angular and conformational degrees of freedom during the physical separation of the two binding partners. Such an unrestricted exploration of the configurational space can dramatically slow down sampling, thus requiring long simulation times to achieve adequate convergence.[2,74]

A more sophisticated approach, compared to the unrestrained methodology, involves the combined use of multiple CVs that project the conformational CV onto the COM-to-COM distance CV.[262–264] Although this strategy appears to be more reasonable, it still may not embrace all the relevant degrees of freedom in the system, potentially leading to inaccurate or erroneous characterization of the binding process.[12,46,63,74] Notably, the reversible association of the two α–helices during GpA dimerization is accompanied by significant conformational, translational, and orientational movements, which are major factors contributing

to changes in configurational entropy, and have to be considered when selecting relevant CVs to obtain reliable results.

In the geometrical route, the significant changes in configurational entropy that accompany binding are reduced through the introducing and restraining conformational, orientational, and positional CVs, in order to control the corresponding modifications in the molecular moiety with respect to the binding pose in both the bound and the unbound states.[46,295] The choice of the CVs for the GpA dimerization process is described in Table 7.2 and visualized in Figure 7.1. Starting from step 2 (see Table 7.2), the restraints at the i–th step keep all CVs from step 1 to step $i-1$ at their equilibrium values, which can be determined from the minima of the one-dimensional PMF calculations in steps 1 through i-1. These restraints gradually limit the conformational flexibility, as well as the orientational and positional movements of the relevant molecular moiety prior to the PMF calculation for the physical separation along the distance between the COMs of the two partners.

The selection of root-mean-square deviation (RMSD) over backbone atoms, applied to both α–helical segments in the membrane (denoted as $G^{\text{site}}_{\text{c(H}_1\text{)}}$, $G^{\text{site}}_{\text{c(H}_2\text{)}}$, $G^{\text{bulk}}_{\text{c}}$ in Table 7.2), holds similar importance as for soluble protein complexes,[46,295] primarily due to the high flexibility exhibited by the proteins, which can result in notable deviations from the bound-state structure. In certain scenarios, additional RMSD restraints may be required to control the isomerization of interfacial side chains during the physical separation of the binding partners owing to solvent exposure, which might result in a loss of interaction and progressive deterioration

Table 7.2: Collective variables and their calculation order for standard binding free-energy calculation of GpA dimerization via geometrical route.

Step	CVs	Partner movement	Representations[a]	Restraints
1	RMSD(H_1)	Conformational	$G^{\text{site}}_{\text{c}(H_1)}$	
2	RMSD(H_2)		$G^{\text{site}}_{\text{c}(H_2)}$	RMSD(H_1)
3	Θ	Orientational	G^{site}_{Θ}	RMSD(H_1), RMSD(H_2)
4	Φ		G^{site}_{Φ}	RMSD(H_1), RMSD(H_2), Θ
5	Ψ		G^{site}_{Ψ}	RMSD(H_1), RMSD(H_2), Θ, Φ
6	ϕ	Positional	G^{site}_{ϕ}	RMSD(H_1), RMSD(H_2), Θ, Φ, Ψ
7	r		$(1/\beta)\ln(L^* \cdot I^* \cdot C^{\circ}_{\text{surf}})$[b]	RMSD(H_1), RMSD(H_2), Θ, Φ, Ψ, ϕ
8	RMSD	Conformational	$G^{\text{bulk}}_{\text{c}}$	
9	Θ	Orientational	G^{bulk}_{Θ}	RMSD
10	Φ		G^{bulk}_{Φ}	RMSD, Θ
11	Ψ		G^{bulk}_{Ψ}	RMSD, Θ, Φ

[a]The superscripts "site" and "bulk" refer to the bound and the unbound states, respectively.
[b]Corresponds to the energetic contribution arising from the evaluation of the physical separation PMF along the distance between the COMs of the two partners. The details of the parameters β, L^*, I^*, C°_{surf} are discussed in this section.

of the standard binding free-energy estimate.[46]

Furthermore, when calculating the binding free energy of the GpA complex, the structural specificity of the latter calls for particular attention. The GpA homodimer comprises two indistinguishable α–helices, H_1 and H_2, which exhibit identical characteristics in the unbound state. It is crucial to note that both monomers contribute equivalently to the bulk free energy (G_c^{bulk}) in the unbound state. However, due to the presence of two α–helical segments, the value obtained for the unbound state should be doubled to correctly reflect the overall energetic landscape and account for the dimeric nature of GpA (see eq. 7.2).

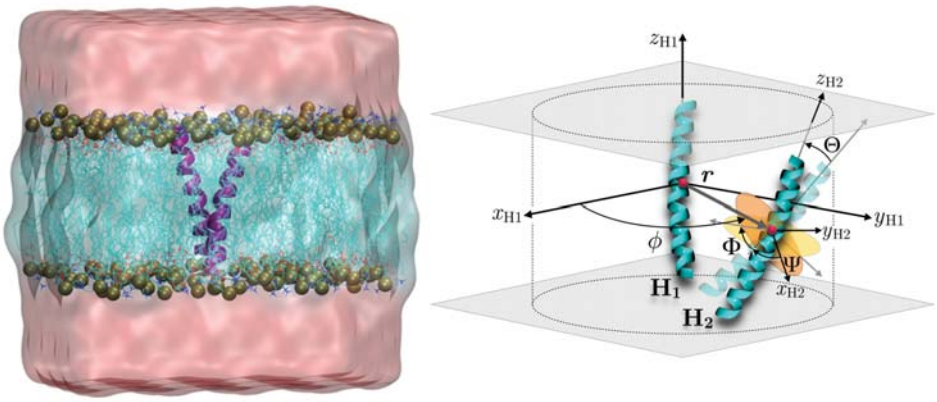

Figure 7.1: GpA_{69-97} dimer in a POPC (1-palmitoyl-2-oleoyl-sn-glycero-3-phosphatidylcholine) lipid bilayer (left). The scheme of the reference coordinates is used to define the orientational and positional restraints, where H_1 and H_2 correspond to the symmetric α–helices. In the geometrical route, the H_1 helix is pinned to the origin of the simulation box and prevented from tumbling. The dashed line between two red balls represents the COM distance in the cylindrical coordinate frame of reference. The azimuthal angle, ϕ, relates to the position of H_2 with respect to H_1. The Euler angles (roll angle Θ, pitch angle Φ, and yaw angle Ψ) determine the relative orientation of H_2 with respect to H_1 (right).

In the context of integral, TM complexes, it is also important to consider the membrane environment when calculating the binding affinity, $\Delta G_{\mathrm{b}}^{\circ}$. Here, the lipid bilayer may naturally impose orientational restraints on the α–helices, preventing their free rotation within its hydrophobic environment.[289–292] Hence, it may be argued that the benefit of the application of orientational restraints introduced on the three Euler angles (Θ, Φ, Ψ) in the geometrical route over the simple unrestrained physical separation PMF calculations is only worthwhile in the case of homogeneous, isotropic environments, such as water.[74] However, our study of GpA dimerization demonstrates that the geometrical route, with a careful selection of CVs, including orientational degrees of freedom, denoted as $G_{\Theta}^{\mathrm{site}}$, G_{Φ}^{site}, and G_{Ψ}^{site} in Table 7.2, yields a more accurate and reliable estimate of the standard binding free energy, compared to the unrestrained PMF approach. This result underscores the importance of the choice of the CVs to capture all

the relevant contributions to the changes in the configurational entropy. Given the inhomogeneous nature of the membrane environment, the contribution of the orientational restraints in the bulk cannot be assessed analytically owing to the dissimilarity in the probability of exploring all the available space, compared to an isotropic medium.[12,46] It is, therefore, essential to properly estimate the unbound-state free-energy contributions (G_{Θ}^{bulk}, G_{Φ}^{bulk}, and G_{Ψ}^{bulk}) for an accurate evaluation of standard binding affinities for membrane proteins (for additional detail, see the Supporting Information (SI)).

An additional issue that ought to be considered to thoroughly characterize TM-protein association is the symmetry of the system, cylindrical in nature. Unlike protein complexes in an aqueous solution, where the relative position of the binding partners can be described using spherical coordinates,[46,74,295] the physical separation of TM proteins in the membrane requires that the binding constant be expressed in cylindrical coordinates.[261] Therefore, only an azimuthal angle CV (ϕ, G_{ϕ}^{site} in Table 7.2) is utilized for standard binding free-energy calculations.

Here, the equilibrium constant, $K_{\text{eq}}^{\text{GR}}$, obeying the geometrical route, can be expressed as:

$$K_{\text{eq}}^{\text{GR}} = L^* I^* \text{e}^{-\beta(G_{\text{c}}^{\text{bulk}} - G_{\text{c(H}_1\text{)}}^{\text{site}} - G_{\text{c(H}_2\text{)}}^{\text{site}} + G_{\Theta}^{\text{bulk}} + G_{\Phi}^{\text{bulk}} + G_{\Psi}^{\text{bulk}} - G_{\Theta}^{\text{site}} - G_{\Phi}^{\text{site}} - G_{\Psi}^{\text{site}} - G_{\phi}^{\text{site}})}, \quad (7.2)$$

where $\beta = (k_{\text{B}}T)^{-1}$, k_{B} is the Boltzmann constant, and T is the temperature. L^* has the dimension of length, as defined below. The theoretical developments leading to the equilibrium constant, $K_{\text{eq}}^{\text{GR}}$, are thoroughly described in the SI. A

key contribution is I^*. It is a one-dimensional integral over r, defined in terms of the separation PMF:[12,46]

$$I^* = \int \mathrm{d}r \; \mathrm{e}^{-\beta[w(r)-w(r^*)]}, \tag{7.3}$$

where $w(r)$ corresponds to the actual separation PMF subject to the imposed constraints on other CVs, and $w(r^*)$ is the PMF at distance r^*, where both partners are located sufficiently far away from each other to no longer interact. r corresponds to the cylindrical COM-to-COM distance between helices H_1 and H_2 in the plane of the membrane, equal to $\left((x_1 - x_2)^2 + (y_1 - y_2)^2\right)^{1/2}$, where x_i and y_i represent the two-dimensional coordinates of the α–helices. r^* denotes a separation at which the two binding partners no longer interact.

In the two-dimensional membrane environment treated with cylindrical coordinates, the length-scale term, L^*, is given as:

$$L^* = r^* \int_0^{2\pi} \mathrm{d}\phi \; \mathrm{e}^{-\beta u(\phi)}, \tag{7.4}$$

where $u(\phi)$ is the harmonic restraint potential acting on the azimuthal angle, ϕ. The complete expression for $K_{\mathrm{eq}}^{\mathrm{GR}}$ following the tailored geometrical route for the GpA dimer can be found in the SI.

Previously, it was demonstrated by Gumbart et al.[12] that for protein-ligand complexes in an aqueous environment, the PMF (technically, the free-energy surface[296]) as calculated includes a Jacobian term, which gradually decreases at large

distances due to increasing entropy. This decrease takes the functional form $w(r^*) = -1/\beta \ \ln\left(\alpha\, r^{*2}\right)$, where α corresponds to a fitting constant,[26,297,298] and it keeps the product of S^* (as denoted there) and I^* independent of the choice of r^*. However, in the case of the membrane protein, the PMF evaluation is restrained to the (x,y)–plane. Consequently, the entropic term takes the form $w(r^*) = -1/\beta \ \ln r^*$ at large separation (see SI).

In the context of membrane proteins, it is important to consider the specific location of the protein within the lipid bilayer, as well as its solubility properties. Depending on these factors, additional restraints along the z–axis, which represents the direction perpendicular to the membrane plane (as illustrated in Figure 7.1), may need to be introduced during the separation process. These restraints, accompanied by the associated free energies ($G_{\mathrm{z}}^{\mathrm{bulk}}$ and $G_{\mathrm{z}}^{\mathrm{site}}$), would allow the relative orientation and position of the binding partners to be controlled precisely along the membrane normal. However, in the case of the GpA dimer, the complex primarily resides within the lipid bilayer, which has a sufficient thickness to accommodate the two α–helices without significant distortion (as depicted in the Results and Discussions section). Consequently, for the specific investigation of the GpA dimer here, we have chosen to not include any additional restraint along the z–axis, since the z–coordinate is already adequately sampled.

It should be noted that in the unrestrained separation PMF approach, the exponential terms in eq. 7.2 and eq. 7.4 are equal to one because the contributions of the degrees of freedom other than the COM-to-COM distance are unknown. Additionally, there is a difference in the evaluation of K_{eq} between the unrestrained

physical separation PMF strategy (i.e., $K_{\mathrm{eq}}^{\mathrm{unrestr}}$) and the geometrical route in terms of the homodimer symmetry. In the geometrical route, H_1 is fixed at the origin and prevented from tumbling and drifting, obviating the need to consider the symmetry number of the homodimer.[261] In contrast, in the unrestrained separation PMF strategy, the reference monomer is uncontrolled during the physical separation, meaning that the final equilibrium constant should be multiplied by the symmetry factor $S = 1/2$.[251,261,265] The binding constant in unrestrained PMF calculations should, therefore, be expressed as:

$$K_{\mathrm{eq}}^{\mathrm{unrestr}} = 2\pi\, S \int \mathrm{d}r\; \mathrm{e}^{-\beta[w(r)-w(r^*)]} \tag{7.5}$$

Henceforth, the binding affinity, $\Delta G_{\mathrm{b}}^{\circ}$, will be evaluated as:

$$\Delta G_{\mathrm{b}}^{\circ} = -\frac{1}{\beta}\,\ln(K_{\mathrm{eq}}\, C_{\mathrm{surf}}^{\circ}) \tag{7.6}$$

In regard to the significance of the choice of the standard state emphasized previously, the standard unit of concentration employed in the present study is expressed as molecules per unit area ($1/\text{Å}^2$).[266,268] We have determined K_{eq} values from the separation simulations, which are presented in units of Å^2. This particular choice allows us to express $\Delta G_{\mathrm{b}}^{\circ}$ in the more customary units of kcal/mol, which are frequently employed in thermodynamic analyses, and facilitate comparison across different investigations. By employing this conversion, we enhance the interpretability of our results and foster compatibility with the existing scientific

literature.

Simulation Details

The selection of POPE (phosphatidylethanolamine) and POPC for bilayers is supported by their frequent occurrence in biological membranes.[299–303] Additionally, we employed widely accepted and recognized all-atom force fields (CHARMM36, CHARMM36m, and CHARMM27 as an extension of CHARMM22 with cross-term (CMAP) corrections for proteins and protein-lipid interactions),[304,305] which have been extensively tested and validated in the literature.[286,300,303] It should be noted that the lipid and protein parameters are vastly improved in CHARMM36 and CHARMM36m compared to CHARMM27.[287,306,307] Moreover, CHARMM36m has been shown to be an effective framework to capture the dynamic behavior and conformational ensembles, even when simulating proteins lacking an ordered three-dimensional structure.[134]

The enhanced-sampling algorithms presented herein can be classified as (i) generalized ensemble methods, such as the bias-replica-exchange MD (B-REMD),[308–310] (ii) CV-based methods that apply biasing forces to overcome free-energy barriers along the CVs, such as the adaptive biasing force (ABF) method and its variants, and (iii) a combination of methods from (i) and (ii).[2,116] The B-REMD algorithm[308–310] involves the exchange of harmonic restraints between adjacent replicas (windows), which are executed in parallel, enabling the system to escape from local minima in the free-energy landscape. However, B-REMD suffers from

slow exchange rates between replicas when the energy barriers are high, which can affect the sampling efficiency and accuracy of the simulations.[311–313] In the classical implementation of ABF, namely standard ABF (stABF),[24,104,105] the bias is applied directly onto the selected CV. A poorly chosen CV can cause the system to be trapped between high free-energy barriers lingering in one valley for an extended period before transitioning to the next. This shortcoming can be overcome with multiple walkers ABF (MW-ABF)[107,314–316] as this algorithm enables the simultaneous sampling of multiple energy valleys along the CV and exchange of information at fixed intervals. However, the walker selection rules must be appropriately applied to prevent the possibility of kinetic trapping in the same valley. In contrast, in the extended Lagrangian variant of ABF (eABF),[107,110,111] the bias is applied to a fictitious particle harmonically coupled to the CV, which permits more effective sampling, and avoids the direct influence of the biasing force on the coordinate.[112–114] When combined with multiple walkers[107,314] and well-tempered metadynamics,[98,99,102] which mollifies the free-energy surface, the simulation convergence rate is significantly improved compared to the aforementioned mentioned algorithms.[106,109,115,116]

Results and Discussions

Unrestrained separation PMF calculation

In an effort to depict the performance of the unrestrained separation PMF approach, using as the CV only the COM-to-COM distance between the two helices,

Table 7.3: GpA reversible association in unrestrained PMF calculation under different computational conditions.

Index	GpA length	Force field[a]	Sampling algorithm	Bilayer	ΔG_b°, kcal/mol	Ω^b, degree	Simulation length, μs
A	69–97	C27	MW-WTM-eABF	POPC	−8.2	36	16.0
B	69–97	C36	stABF	POPC	−6.3	37	6.0
C	69–97	C36	MW-ABF	POPC	−3.0	28	4.8
D	72–96	C36	B-REMD	POPC	−8.3	50	2.6
E	70–97	C36	stABF	POPE	−6.5	41	1.8
F	69–97	C36	MW-WTM-eABF	POPE	−5.4	39	1.6
G	70–97	C27	stABF	POPE	−12.8	45	1.7

[a]CHARMM27 is denoted as C27, C36 corresponds to CHARMM36m/CHARMM36 used for proteins/lipids. [b]refers to the average crossing angle between the two helices corresponding to the minima in their separation PMF.

we explored seven cases with distinct computational conditions (see Table 7.3 and Figure 7.2). Among these conditions are different GpA sequence lengths, monounsaturated bilayers (POPC and POPE), sampling algorithms (standard adaptive bias force, or stABF),[24,104,105] multiple walker ABF (MW-ABF),[107,314,315] multiple walker well–tempered metadynamics extended ABF (MW-WTM-eABF),[106,109,115,116,315] and B-REMD[308,310]). All–atom force fields (CHARMM27,[304,305] CHARMM36,[286,287] and CHARMM-36m[134]) were used in all simulations. The ΔG_b° presented in Table 7.3 were calculated with eq. 7.5. Detail of the MD protocols used for the unrestrained physical separation PMF calculations can be found in the SI.

While all the calculated ΔG_b° values fall within the range of experimentally observed binding free energies for GpA dimerization (−3.4 to −12.1 kcal/mol),[253,254,257–260]

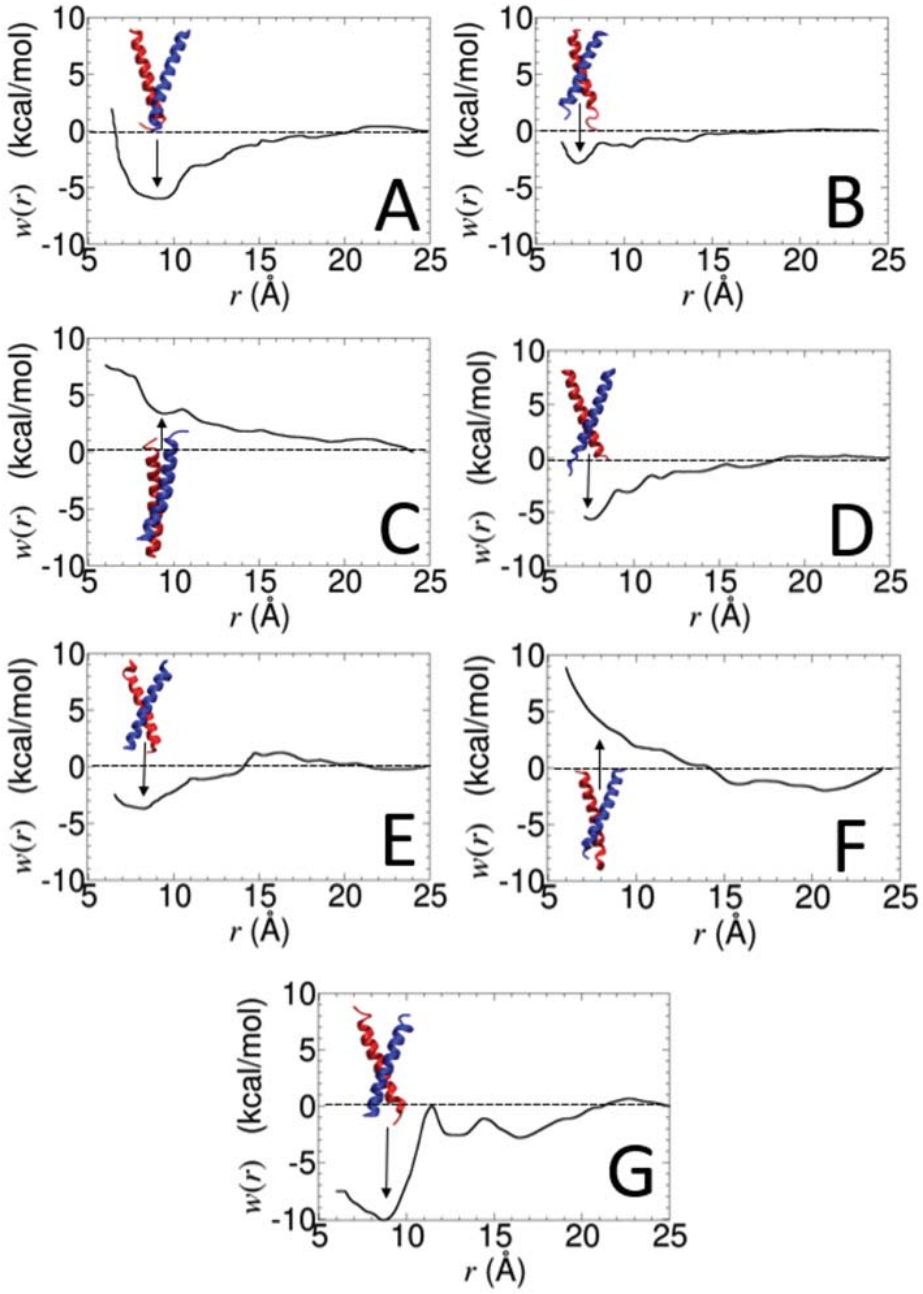

Figure 7.2: GpA reversible association in unrestrained-separation PMF calculations under different computational conditions. The molecular image in each panel depicts a snapshot of the molecular structure corresponding to the minima of the PMF, where H_1 and H_2 helices are shown in blue and red, respectively.

these results should be interpreted with great caution, and cannot be considered as formally reproducible in any of the reported cases (see Figure 7.2A–G). The shape of the unrestrained physical separation PMF and the absence of a global

minimum at small separation, r, (cases C, E, F in Figure 7.2) indicates that the estimated $\Delta G_{\mathrm{b}}^{\circ}$ values for these cases are likely to be fortuitous.

In case A, the quality of the separation PMF suggests that the COM-to-COM distance between the α–helices alone does not capture the complexity and the multidimensionality of the underlying free-energy landscape beyond 14Å. Consequently, even with substantially long simulations, the sampling algorithm, regardless of its sophistication, may still struggle to adequately explore and visit all the relevant conformations in that region.

To gain insight into the stability of the homodimer spatial arrangement, we examined the crossing angle, Ω, formed by the longitudinal axes of the two α–helices at small inter-helical distances in the unrestrained physical-separation PMFs (see Figure 7.2). Specifically, Ω provides information about the relative orientation of the TM helices, and can indicate whether they are favorably positioned to establish stable interactions.[234,267,317,318] For the PMFs that did not exhibit a clear minimum, e.g., cases E and F, we selected an arbitrarily small distance between the two α–helices to evaluate Ω, aiming at capturing the behavior of the binding partners in close proximity. In contrast, for the PMFs that exhibit a clear minimum, we determined Ω within 0.05 Å of this minimum. It is noteworthy that the reported value of Ω for the experimental structure used in our study is 40°.

Since the PMFs for cases E and F do not exhibit clear-cut minima, the close agreement with the experimental value of Ω ought to be interpreted with caution, as it might be fortuitous and stem from insufficient sampling of configurational

space, rather than from the stability of the homodimer. Furthermore, our analysis reveals that the dimeric structures associated to the free-energy minima of cases A, B, C, D, and G correspond to crossing angles that fall outside the expected range (see Table 7.3). These findings suggest that, despite the belief that the lipid bilayer alone sufficiently restrains the relative orientation of the α–helices, imposing retraining potentials on additional degrees of freedom ought to be considered. Furthermore, the observed deviations of Ω are indicative of the ability of the TM α-helices to adopt different orientations, pointing to the coexistence of multiple conformational states and transitions between them.[251,319–321]

The common trait of the reported unrestrained, physical separation PMFs is, in all likelihood, an insufficient sampling along r, which should be viewed as a consequence of the complexity of the GpA dimerization process, rather than a blatant failure of the chosen sampling algorithm, or of the force field. Sampling all possible configurations associated with α–helix binding would probably require far greater simulation times to ensure adequate convergence of the free-energy calculation. In stark contrast, controlling the changes in the internal conformational, relative orientational and positional degrees of freedom that accompany the reversible association of the TM helical segments with suitably chosen restraining potentials—as would be the case in the geometrical route—could markedly accelerate convergence, and, hence, reduce the required computational time, as will be seen hereafter.[74]

Geometrical route

In order to determine as accurately as possible the standard binding free energy associated with GpA dimerization, we employed the geometrical route,[12,46] which allows the energetic contributions arising from the geometric restraints applied sequentially to the different degrees of freedom to be evaluated precisely. Here, use was made of the CHARMM36 and CHARMM36m force fields for lipids and proteins, respectively, due to their ability to capture with suitable accuracy the behavior of TM proteins in lipid bilayers.[134,286,288] The choice of POPC for the lipid bilayer was based on its resemblance to natural lipids found in mammalian membranes.[300,322] Additionally, the extensive use of POPC in previous computational studies of GpA dimerization further warrants its choice for the present theoretical investigation.[252,262–264] To sample efficiently the free-energy landscape, we employed the WTM-eABF advanced sampling algorithm.[106,115,116]

The results of the standard binding free-energy calculations for the GpA homodimer, following the geometrical route, are gathered in Table 7.4. In this table, $\Delta G_{\mathrm{b}}^{\circ}$ is reported for two specific standard states, demonstrating the impact of the units chosen for $C_{\mathrm{surf}}^{\circ}$ on the calculated binding free energy. It is worth noting that the obtained $\Delta G_{\mathrm{b}}^{\circ}$ estimate of -11.1 kcal/mol closely aligns with the documented experimental value reported by Hong et al.[258] (see Table 7.1). In their study, a similar standardization scheme was employed, further lending support of our comparison of theory and experiment. Specifically, Hong et al. reported a $\Delta G_{\mathrm{b}}^{\circ}$ value of -12.1 kcal/mol, being less than $2k_{\mathrm{B}}T$ away from our estimate, underscoring

the reliability and consistency of our approach.

In the geometrical route, the energetic terms arising from the internal conformational changes within the two TM α–helices, as well as their physical separation contribute significantly to the final $\Delta G_{\mathrm{b}}^{\circ}$ estimate (see Table 7.4), emphasizing their relevance in the binding process. Despite the relatively small contribution of the orientational angular restraints ($G_{\mathrm{o}}^{\mathrm{site}} + G_{\mathrm{o}}^{\mathrm{bulk}}$) to the total $\Delta G_{\mathrm{b}}^{\circ}$ (viz. 1.5 kcal/mol), accurate evaluation of the contributions arising from these restraints remains essential.[12,46,74]

Table 7.4: Detailed results of the different contributions to the standard binding free energy of the GpA dimerization.

Representation	Contribution (kcal/mol)	Simulation time (μs)
$G^{\text{site}}_{c(\text{H1})}$	$-$ 11.7 $\pm$ 0.1	0.40
$G^{\text{site}}_{c(\text{H2})}$	$-$ 15.2 $\pm$ 0.0	0.40
G^{site}_{Θ}	$-$ 0.4 $\pm$ 0.0	0.12
G^{site}_{Φ}	$-$ 0.5 $\pm$ 0.0	0.16
G^{site}_{Ψ}	$-$ 0.2 $\pm$ 0.0	0.12
G^{site}_{ϕ}	$-$ 0.5 $\pm$ 0.0	0.12
$(1/\beta)\ln(L^* \cdot I^* \cdot C^{\circ}_{\text{surf}})$	-6.8 ± 0.0	1.30
G^{bulk}_{c}	21.6 $\pm$ 0.1[a]	0.50
G^{bulk}_{Θ}	0.8 $\pm$ 0.0	0.20
G^{bulk}_{Φ}	1.2 $\pm$ 0.0	0.20
G^{bulk}_{Ψ}	0.6 $\pm$ 0.0	0.20
$\Delta G^{\circ}_{\text{b}}$	-11.1 ± 0.2	3.72

[a]is a doubled contribution arising from the identical structures of the helices.

In this section, we focus solely on the analysis of the physical separation PMF simulation through the geometrical route (see Figure 7.3A), and its implications for the GpA binding process. Initially, we performed the separation PMF calculation for the first 0.9 μs in a single window ranging from 4.8 to 25.8 Å. To improve sampling in each bin, we stratified this window into four separate strata of even length, and performed additional sampling, corresponding to an additional 0.1 μs for each stratum. Finally, we combined the individual free-energy profile from

these windows into a single PMF spanning the entire reaction pathway. Details of the computational assays that yielded the PMFs for the different energetic contributions, alongside their convergence, can be found in the SI.

Akin to the analysis of the unrestrained separation PMF approach above, we extracted from the trajectories configurations within 0.05 Å from the minimum of $w(r)$ determined within the geometrical route, and found the average crossing angle formed by the two α–helices to be equal to 41°. This theoretical result is consistent with the experimental measurement for the structure utilized (viz. 40°).[234] Our findings further suggest that following the geometrical route helps render a realistic picture of the GpA dimerization mechanism.

The pair–interaction analysis of the separation PMF was carried out in connection with the underlying hypotheses of the two-stage model put forth by Popot et al.[225,230,231,323] (see Figure 7.3B). The physical separation PMF was partitioned into van der Waals and electrostatic free-energy contributions arising from helix-helix and helix-solvent interactions as a function of the separation coordinate. The results of the pair-interaction analysis differ somewhat from those previously reported by Hénin et al.,[261] which can be attributed to various factors, including the length of the GpA sequence used, the presence of capping termini, and the different nature of the anisotropic environment, namely a genuine lipid bilayer versus a slab of oil mimicking the latter (see Table 7.1). As mentioned previously, the GpA model used in our study consists of residues 69–97, which is longer than that used by Hénin et al. (residues 73–95).[261] Moreover, in our model the C– (e.g. ARG97) and N–termini (e.g., SER69) are not blocked by capping groups, which

contribute to distinct interactions between the α–helices and their environment.

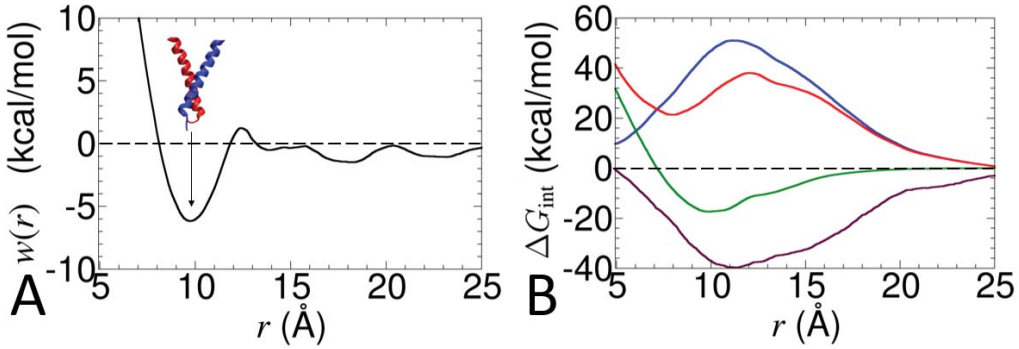

Figure 7.3: (A) Normalized physical separation PMF of the GpA homodimer in POPC obtained within the geometrical route simulation at 1300 ns. (B) ΔG_{int} corresponds to the energy interaction components obtained from the partition of the physical separation PMF into helix–helix electrostatic (blue), helix–helix vdW (green), helix–helix (sum of electrostatic and vdW interactions) (red), and sum of electrostatic and vdW interactions of helix–solvent contributions (maroon). The solvent includes POPC lipid bilayer, water, sodium, and chloride ions.

The global minimum of the van der Waals inter-helical interaction term (the green curve in Figure 7.3B) is located at 9.7 Å, close to the global minimum of the PMF. Disruption of the α–helix dimer occurs around 12 Å, corresponding to a separation at which the inter-helical association becomes unfavorable, and the TM α–helices begin to move apart. The variation of the free energy in this range is accompanied by an increase of the van der Waals term and a decrease of the electrostatic term, indicative that at short range, association is driven primarily by dispersive interactions. At larger separations ($r > 15$ Å), the inter-helical van der Waals free-energy contribution reaches a plateau and remains constant, while the inter-helical electrostatic free energy decreases between 15 and 20 Å, which is related to a thermodynamically unfavorable rupture of ionic interaction at the C– and N–termini. The behavior of the inter-helical electrostatic term at large separations is also consistent with that of two interacting macro-dipoles, obeying an r^{-3}

decay. The helix–solvent energetic contribution reaches its minimum around 11 Å, below which the solvent contribution increases, becoming the primary driving force in the association of the α–helices. This observation is in agreement with the two-stage model,[225,230,231] and suggests that formation of the complex is stabilized by the environment.

In addition, we investigated the recognition mechanism through the pair distribution function of the residues involved in the formation of inter-helical contacts, as reported by Hénin et al.[261] (see Figure 7.4). The initial inter-helical association occurs around 15 Å, mainly due to the formation of an ILE73–THR74 interaction, which is later disrupted at approximately 10 Å. As the distance between the α–helices decreases, intermittent ILE88–ILE88 and ILE88–ILE91 contacts are formed. At smaller separations, i.e., $r < 6$ Å, multiple interactions involving the THR87–ILE88, LEU75–ILE76, and ILE76–ILE76 pairs of residues contribute to the formation of the dimer, persisting in the homodimeric structure.

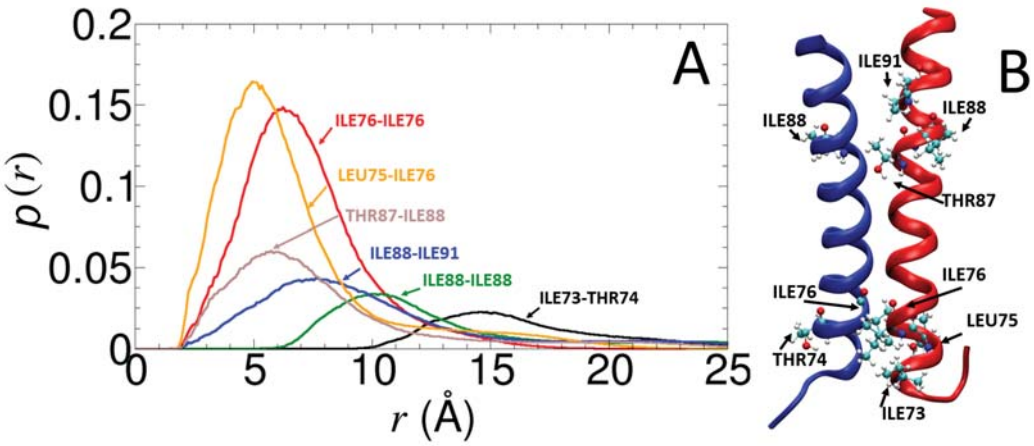

Figure 7.4: (A) Pair distribution function of the residue pairs. (B) The residues that form the interhelical contacts[261] during the association of GpA helices.

To explore the structural changes within the POPC bilayer in the course of

reversible α–helix association, we examined the evolution of its thickness over the first 0.9 μs of the simulation. The bilayer thickness was determined from the distance separating the COM of the upper and lower leaflets. We compared the overall bilayer thickness of the system and the local bilayer thickness at a distance of 16 Å from helix H_2 (see Figure 7.5). The overall bilayer thickness lies within the range of reported values for a POPC bilayer, i.e., typically 30–40 Å.[287,324,325] At the beginning of the simulation, it was about 23 Å, subsequently increasing to about 30 Å due to lipid exchange and deformation as the α–helices separated. The local bilayer thickness is equal to about 20 Å, which is slightly thinner than the overall bilayer thickness.

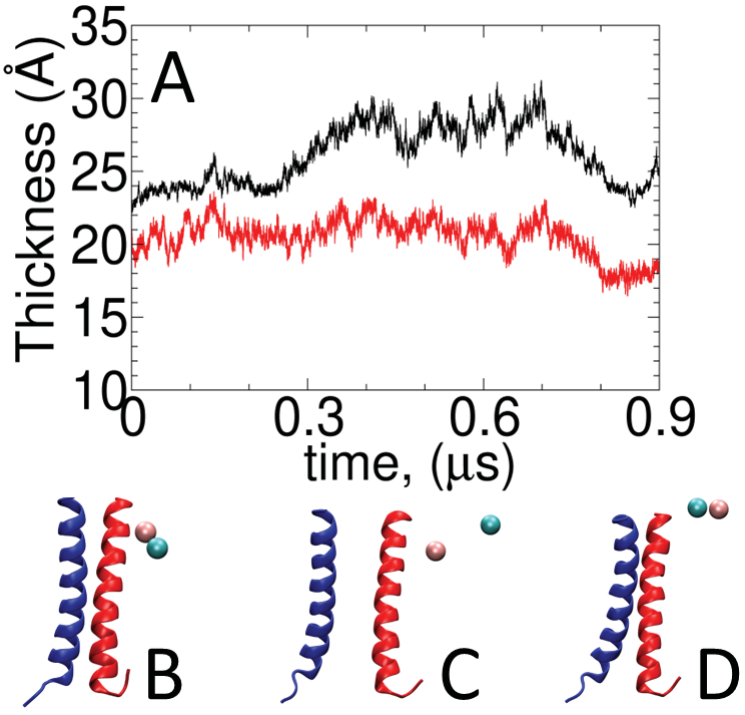

Figure 7.5: (A) Time evolution of the overall bilayer thickness of the system (black) and local bilayer thickness within a distance of 16 Å from helix H_2 (red) during the separation PMF simulations. Snapshots of the rearrangement of the selected phosphorus atoms (shown in pink and cyan balls) (B) at the beginning of the simulation, (C) at the moment of separation, (D) after reversible association.

To illustrate how the lipid bilayer behaves during TM α–helix separation, we provide snapshots of two POPC atoms of the upper leaflet located at 2 Å from the H_2 helix (see Figure 7.5B–D). In the course of the reversible association, the lipid chains adjust, deform, and regroup, indirectly contributing to the binding free energy. Our observations indicate that the local perturbations in the lipid bilayer thickness act as a driving force towards α–helix dimerization, bringing together the two TM helical segments. This behavior is reminiscent of hydrophobic mismatch, which forces the membrane to deform, translating to an energetic penalty. The present result is in line with a host of theoretical and experimental investigations of membrane proteins and their lipid environments.[268,326–337]

Intermediate states in the association pathway

In the detailed study of the physical separation PMF calculations by means of the geometrical route, we observed that the TM α–helices, while reversibly associating, go through a long-lived spatial arrangement, potentially acting as an intermediate state in dimer formation. This arrangement is distinct from the established native configuration of the GpA dimer (see Figure 7.6A), and manifests itself at approximately 6 Å, suggesting that it is part of the recognition and association process. Despite its extended persistence, this particular intermediate state should be considered in the context of the broader dynamics of the GpA complex, and prompts the consideration of other potential intermediate states,[224] some of which are acknowledged and documented in early research works on TM proteins in references 224, 233, 294, 323.

To understand the nature of the observed intermediate state, we conducted an extensive analysis of this structure via a 1–μs–equilibrium simulation, and compared the latter with a 1–μs–equilibrium simulation of the native structure. Analysis of the RMSD over backbone atoms of the dimer shows that the value for the non–native structure is slightly higher ($\sim$ 3.5 Å) than that for the native one ($\sim$ 2.5 Å). The flat RMSD time series for the newly observed state also indicates that the structure does not undergo significant changes in the backbone conformation over the simulation time. Moreover, we characterized the protein-protein interface for both dimers through the analysis of their respective buried surface area (BSA).[338] In the equilibrium simulations with the POPC lipid bilayer, the BSA for both states amounts to $\sim$ 900 Å^2 (see Figure 7.6D), which is consistent with previous studies,[261,339] and indicates that the TM helical segments remained closely associated.

The GpA dimer can also be characterized quantitatively by the crossing and tilt angles, denoted as Ω and θ, respectively (see Figure 7.6E). We have used previously Ω to assess the stability of the complex in the course of the physical separation PMF calculations. Specifically, Ω is defined as the angle between the two helical axes at their closest approach point in the dimer interface, whereas θ is the angle between the helix axis and the normal to the membrane.[261,340,341] The reported tilt angle represents the average of both helices.

Our investigation indicates that the average crossing angle between the TM α–helices for both the native and intermediate structures is approximately 40°, indicating a conserved orientation of the two helical segments with respect to

each other in both states. This finding is consistent with previous experimental studies reporting the same crossing angle.[234] We also determined that the averaged tilt angle for both structures is close to the experimental one[234] and stable (viz., ∼20°), indicating a conserved helix orientation relative to the membrane normal. A supplementary analysis on hydrogen-bond occupancies for both the native and intermediate states can be found in the SI.

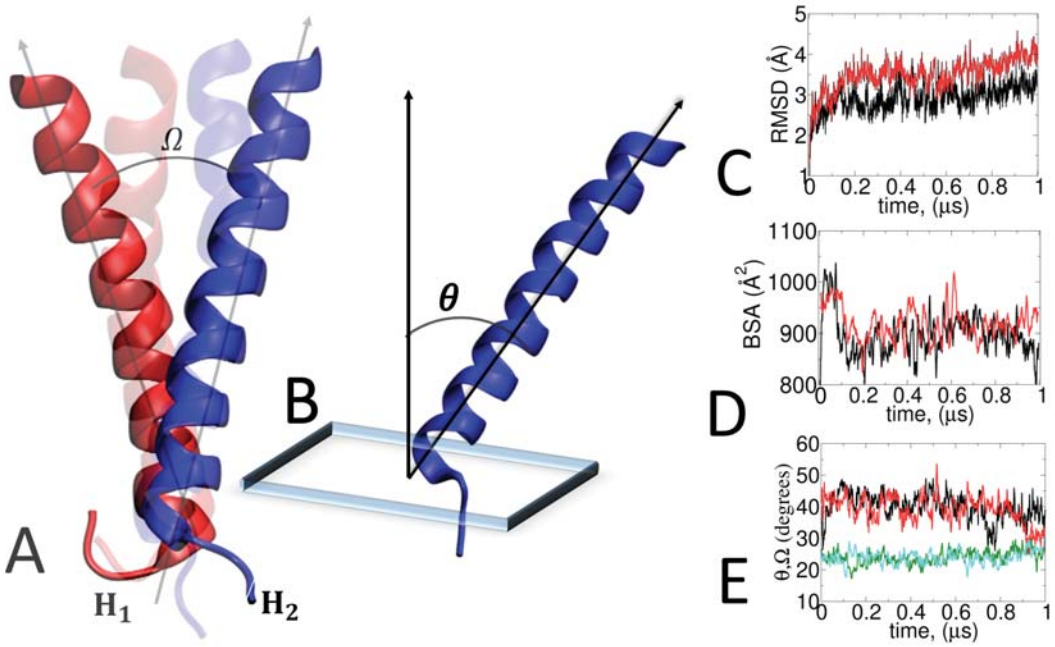

Figure 7.6: (A) The observed intermediate state is shown in transparent mode, and the native state is shown in solid colors. Ω corresponds to the crossing angle between two helices. (B) θ corresponds to the tilt angle. In analyses (C)–(D), the curves of the native state are shown in black and of the intermediate state in red. (C) RMSD of the backbone in 1–μs–equilibrium simulation. (D) Time evolution of their buried molecular surface areas (BSA). (E) Time evolution of the crossing angle Ω (black and red curves correspond to the native and the intermediate states) and averaged tilt angle θ between two helices (green and cyan curves correspond to the native and the intermediate states)[234] in 1–μs–equilibrium simulations.

Conclusions

Recent advances in standard binding free-energy calculations have provided a unique opportunity to gain insights into the mechanisms underlying TM α–helical dimerization, allowing the thermodynamic aspects of the reversible binding process to be explored in a comprehensive fashion, while shedding light on the molecular interactions at play. According to the two-stage model,[225,230,231] the conceptual framework for the analysis of integral membrane protein folding, the TM α–helices remain independently stable prior to oligomerization. Being driven by the solvent, their association is then stabilized by the formation of inter-helical interactions. To probe this model from a theoretical perspective, we have investigated the mechanism of GpA dimerization in lipid bilayers by means of unrestrained separation PMF calculations, as well as a rigorous theoretical framework for standard binding free-energy calculations that combines MD and PMF calculations—the so-called "geometrical route".[12,46]

While the lipid bilayer imposes significant natural restraints on membrane proteins, proper choice of the CVs to investigate the binding process remains critical. In the present work, we underscore the importance of these CVs, upon which sequential application of restraining potentials offers control of the selected degrees of freedom in the reversible association, and substantively minimizes the computational time needed to reach convergence. In the context of anisotropic environments, such as those found in membrane proteins, it is important to approach standard binding free-energy estimation from a distinct perspective compared to

protein-protein complexes in aqueous solutions. Toward this end, we put forth a revised formulation of the geometrical route for reversible dimerization, accounting for the pseudo-cylindrical symmetry of the planar membrane region around the receptor located at the origin. We further show that this approach can be generalized to calculate binding free energies in any inhomogeneous and anisotropic environment. The methodology employed herein provides new insights into the mechanism of TM α–helix dimerization, and holds promise for the theoretical prediction of binding affinity of more complex membrane protein assemblies.

Our study revealed that the unrestrained physical separation PMF approach does not provide a rigorous picture of GpA dimerization, failing to capture all the configurational changes that accompany binding. In contrast, analysis of the physical separation PMF obtained in the geometrical route shows that the α–helices are brought closer together by the lipids at large separation, strengthening inter-helical interactions, which is in agreement with the two-stage model for membrane-protein folding and oligomerization.[225,230,231] Additionally, we observed that the TM α–helical segments, while reversibly associating, go through an intermediate state, whereby they adopt a conformation distinct from that of the native state. Furthermore, insofar as computational approaches for calculating the standard binding free energy of membrane proteins are concerned, several key factors ought to be taken into account to ensure accurate predictions. These key factors include the selection of an appropriate force field and lipid environment, as well as the careful consideration of the native binding motif.

Overall, our study provides valuable insights into the mechanism of TM α–

helix dimerization, and demonstrates the potential of the revised geometrical route approach for predicting binding free energies in inhomogeneous and anisotropic environments. These findings are envisioned to pave the way for the characterization of more complex membrane assemblies, and contribute to the development of more accurate computational approaches for determining the binding free energy of membrane proteins.

7.2 Supplementary Information

Equilibrium Constant in Two Dimensions

The equilibrium binding constant K_{eq} can be written as,

$$K_{\text{eq}} = \frac{\displaystyle\int_{\text{site}} d\mathbf{1} \int d\mathbf{X}\ \mathrm{e}^{-\beta U}}{\displaystyle\int_{\text{bulk}} d\mathbf{1}\ \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X}\ \mathrm{e}^{-\beta U}}. \tag{7.7}$$

The denominator and the numerator of eq. 7.7 each represent the initial and final states of the binding process: the ligand bound to the receptor and the ligand with its center-of-mass at $\mathbf{r}_1^*$ in the bulk, respectively (note that all coordinates are expressed relative to the center of mass of the receptor). The design of a practical computational method consists in inserting intermediate states in eq. 7.7 such that each individual contribution can be calculated easily using PMF simulations. In the following, the intermediate states are constructed by introducing various restraining potentials, which are designed to bias the ligand-protein complex toward

the configuration that adopts the bound state.

It is useful to first establish a local frame of reference in which the position of the center-of-mass of the ligand relative to the receptor $\mathbf{r}_1$ can be specified by (r_1, ϕ_1) in cylindrical coordinates, its z-depth can be specified by z_1, and its orientation can be specified by the three Euler angles $(\Theta_1, \Phi_1, \Psi_1)$. We introduce the "axis" potential $u_{\mathrm{a}}(\phi_1)$, designed to restrain the ligand position along a specific axis as in the bound complex.

To restrain the z-depth of the ligand around its mean position in the bound complex, we can also introduce the z-depth restraining potential $u_{\mathrm{z}}(z_1)$. However, if the free ligand in the membrane spontaneously fluctuates around the same value (as is the case of the glycophorin A dimer), then it may not be necessary to introduce this restraint in the binding calculation. This is equivalent to setting $u_{\mathrm{z}}(z_1)$ to zero in the development. In the following, we carry over the restraining potential u_{z} for the sake of completeness, although it was not used in this work. To restrain the ligand orientation as in the bound complex, we also introduce the potential $u_{\mathrm{o}}(\Theta_1, \Phi_1, \Psi_1)$. Lastly, we introduce the potential u_{c}, designed to restrain the conformation of the receptors and ligand around the average conformations that they adopt in the bound complex. Though other choices are possible, a simple potential can be constructed on the basis of ξ, the root-mean-square-deviation (RMSD) of the ligand relative to its average conformation. With these definitions,

the equilibrium binding constant $K_{\rm eq}$ in eq. 7.7 can be written as:

$$
\begin{aligned}
K_{\rm eq} \;=\; & \frac{\int_{\rm site} d\mathbf{1} \int d\mathbf{X}\; {\rm e}^{-\beta U}}{\int_{\rm site} d\mathbf{1} \int d\mathbf{X}\; {\rm e}^{-\beta[U+u_{\rm c}]}} \\
& \times \frac{\int_{\rm site} d\mathbf{1} \int d\mathbf{X}\; {\rm e}^{-\beta[U+u_{\rm c}]}}{\int_{\rm site} d\mathbf{1} \int d\mathbf{X}\; {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm o}+u_{\rm z}]}} \\
& \times \frac{\int_{\rm site} d\mathbf{1} \int d\mathbf{X}\; {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm o}+u_{\rm z}]}}{\int_{\rm site} d\mathbf{1} \int d\mathbf{X}\; {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm o}+u_{\rm z}+u_{\rm a}]}} \\
& \times \frac{\int_{\rm site} d\mathbf{1} \int d\mathbf{X}\; {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm o}+u_{\rm z}+u_{\rm a}]}}{\int_{\rm bulk} d\mathbf{1}\; \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X}\; {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm o}+u_{\rm z}]}} \\
& \times \frac{\int_{\rm bulk} d\mathbf{1}\; \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X}\; {\rm e}^{-\beta[U+u_{\rm c}+u_{\rm o}+u_{\rm z}]}}{\int_{\rm bulk} d\mathbf{1}\; \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X}\; {\rm e}^{-\beta[U+u_{\rm c}]}} \\
& \times \frac{\int_{\rm bulk} d\mathbf{1}\; \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X}\; {\rm e}^{-\beta[U+u_{\rm c}]}}{\int_{\rm bulk} d\mathbf{1}\; \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X}\; {\rm e}^{-\beta U}}.
\end{aligned}
\tag{7.8}
$$

Most of the terms in eq. 7.8 are dimensionless ratios of configurational integrals corresponding to free energy differences which can be calculated from a standard application of the PMF simulation technique,

$$
\begin{aligned}
{\rm e}^{-\beta G_{\rm c}^{\rm site}} \;&=\; \frac{\int_{\rm site} d\mathbf{1} \int d\mathbf{X}\; {\rm e}^{-\beta[U+u_{\rm c}]}}{\int_{\rm site} d\mathbf{1} \int d\mathbf{X}\; {\rm e}^{-\beta U}} \\
&=\; \langle {\rm e}^{-\beta u_{\rm c}} \rangle_{({\rm site},U)},
\end{aligned}
\tag{7.9a}
$$

$$\begin{aligned} \mathrm{e}^{-\beta G_{\mathrm{o}}^{\mathrm{site}}} &= \frac{\displaystyle\int_{\mathrm{site}} d\mathbf{1} \int d\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c}}+u_{\mathrm{o}}+u_{\mathrm{z}}]}}{\displaystyle\int_{\mathrm{site}} d\mathbf{1} \int d\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c}}]}} \\ &= \left\langle \mathrm{e}^{-\beta[u_{\mathrm{o}}+u_{\mathrm{z}}]} \right\rangle_{(\mathrm{site},U+u_{\mathrm{c}})}, \end{aligned} \tag{7.9b}$$

$$\begin{aligned} \mathrm{e}^{-\beta G_{\mathrm{a}}^{\mathrm{site}}} &= \frac{\displaystyle\int_{\mathrm{site}} d\mathbf{1} \int d\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c}}+u_{\mathrm{o}}+u_{\mathrm{a}}]}}{\displaystyle\int_{\mathrm{site}} d\mathbf{1} \int d\mathbf{X}\ \mathrm{e}^{-\beta(U+u_{\mathrm{c}}+u_{\mathrm{o}})}} \\ &= \left\langle \mathrm{e}^{-\beta u_{\mathrm{a}}} \right\rangle_{(\mathrm{site},U+u_{\mathrm{c}}+u_{\mathrm{o}})}, \end{aligned} \tag{7.9c}$$

$$\begin{aligned} \mathrm{e}^{-\beta G_{\mathrm{o}}^{\mathrm{bulk}}} &= \frac{\displaystyle\int_{\mathrm{bulk}} d\mathbf{1}\ \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c}}+u_{\mathrm{o}}+u_{\mathrm{z}}]}}{\displaystyle\int_{\mathrm{bulk}} d\mathbf{1}\ \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c}}]}} \\ &= \left\langle \mathrm{e}^{-\beta[u_{\mathrm{o}}+u_{\mathrm{z}}]} \right\rangle_{(\mathrm{bulk},U+u_{\mathrm{c}})}, \end{aligned} \tag{7.9d}$$

(for the sake of simplicity, we can lump the z-depth restraint together with the orientational restraint for both the site and bulk calculations).

$$\begin{aligned} \mathrm{e}^{-\beta G_{\mathrm{c}}^{\mathrm{bulk}}} &= \frac{\displaystyle\int_{\mathrm{bulk}} d\mathbf{1}\ \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\mathrm{c}}]}}{\displaystyle\int_{\mathrm{bulk}} d\mathbf{1}\ \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X}\ \mathrm{e}^{-\beta U}} \\ &= \left\langle \mathrm{e}^{-\beta u_{\mathrm{c}}} \right\rangle_{(\mathrm{bulk},U)}. \end{aligned} \tag{7.9e}$$

One may note that the delta function involving $\mathbf{r}_1^*$, when it appears both in the numerator and denominator, does not affect the calculated free energies in the bulk region because it is invariant to translations.

The free energy $G_{\mathrm{o}}^{\mathrm{bulk}}$ in eq. 7.9d requires a calculation of over the Euler angles

restraining the ligand in the anisotropic membrane environment. Furthermore, the free energy contribution associated with the restraint along the z-depth requires the calculation of an additional PMF for the isolated ligand in the membrane. The full orientational restraint based on Euler angles is $u_{\rm o}(\Theta_1,\Phi_1,\Psi_1) = u_\Theta(\Theta_1) + u_\Phi(\Phi_1) + u_\Psi(\Psi_1)$, yielding the total free energy $G_{\rm o}^{\rm bulk} + G_{\rm z}^{\rm bulk} = G_\Theta^{\rm bulk} + G_\Phi^{\rm bulk} + G_\Psi^{\rm bulk} + G_{\rm z}^{\rm bulk}$, where:

$$\mathrm{e}^{-\beta G_\Theta^{\rm bulk}} = \langle \mathrm{e}^{-\beta u_\Theta} \rangle_{({\rm bulk},U+u_{\rm c})} \tag{7.10}$$

$$\mathrm{e}^{-\beta G_\Phi^{\rm bulk}} = \langle \mathrm{e}^{-\beta u_\Phi} \rangle_{({\rm bulk},U+u_{\rm c}+u_\Theta)} \tag{7.11}$$

$$\mathrm{e}^{-\beta G_\Psi^{\rm bulk}} = \langle \mathrm{e}^{-\beta u_\Theta} \rangle_{({\rm bulk},U+u_{\rm c}+u_\Theta+u_\Phi)} \tag{7.12}$$

$$\mathrm{e}^{-\beta G_{\rm z}^{\rm bulk}} = \langle \mathrm{e}^{-\beta u_{\rm z}} \rangle_{({\rm bulk},U+u_{\rm c}+u_\Theta+u_\Phi+u_\Psi)} \tag{7.13}$$

All these expressions in the site or in the bulk have the same form and can be calculated from a conditional potential of mean force (PMF), $w^*(\xi)$, computed with the potential energy function U^*,

$$\begin{aligned} \langle \mathrm{e}^{-\beta u(\xi)} \rangle_{(U^*)} &= \int d\xi' \; \mathrm{e}^{-\beta u(\xi')} \; \langle \delta(\xi - \xi') \rangle_{(U^*)} \\ &= \frac{\displaystyle\int d\xi' \; \mathrm{e}^{-\beta u(\xi')} \mathrm{e}^{-\beta w^*(\xi')}}{\displaystyle\int d\xi'' \; \mathrm{e}^{-\beta w^*(\xi'')}} \end{aligned} \tag{7.14}$$

The fourth term in eq. 7.8, which involves a ratio of configurational integrals with the bound ligand (numerator) and the ligand held with its center-of-mass at $\mathbf{r}_1^*$ in the bulk by a delta function (denominator), requires special attention

because it does not correspond to a free energy difference like the other terms. It can be re-expressed as:

$$\frac{\displaystyle\int_{\text{site}} d\mathbf{1} \int d\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\text{c}}+u_{\text{o}}+u_{\text{z}}+u_{\text{a}}]}}{\displaystyle\int_{\text{bulk}} d\mathbf{1}\ \delta(\mathbf{r}_1-\mathbf{r}_1^*)\int d\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\text{c}}+u_{\text{o}}+u_{\text{z}}]}} = L^* I^*, \tag{7.15}$$

where L^* is a length defined as an integral over the azimuthal angle ϕ_1,

$$L^* = r_1^* \int_0^{2\pi} d\phi_1\ \mathrm{e}^{-\beta u_{\text{a}}(\phi_1)}, \tag{7.16}$$

and I^* is a one-dimensional integral over the distance r_1

$$I^* = \int_{\text{site}} dr_1\ \mathrm{e}^{-\beta[w(r_1)-w(r_1^*)]} \tag{7.17}$$

defined in terms of the PMF $w(r_1)$ calculated in the presence of the configurational and orientational restraints u_{c}, u_{o}, and u_{a} (see eq. 7.20). To obtain these expressions, we consider the left-hand-side of eq. 7.15 which can be re-expressed in a simpler form by first considering the function $\rho(\mathbf{r}_1)$ defined as:

$$\rho(\mathbf{r}_1) \equiv \int d\mathbf{1}\ \delta(\mathbf{r}_1-\mathbf{r}_1')\int d\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\text{c}}+u_{\text{o}}+u_{\text{z}}]}. \tag{7.18}$$

The position of the center-of-mass is expressed in cylindrical coordinates to account for the planar symmetry of the membrane, and at a sufficiently large distance, the ligand and the receptor are well-separated, and the function $\rho(\mathbf{r}_1^*) \equiv \rho(r_1^*, \phi_1^*, z_1^*)$ becomes independent of ϕ_1^* and z_1^*, i.e., $\rho(\mathbf{r}_1^*) = \rho(r_1^*, 0, 0)$.

It follows that eq. 7.15 can be written as,

$$\begin{aligned}
\frac{\displaystyle\int_{\text{site}} d\mathbf{1} \int d\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\text{c}}+u_{\text{o}}+u_{\text{z}}+u_{\text{a}}]}}{\displaystyle\int_{\text{bulk}} d\mathbf{1}\ \delta(\mathbf{r}_1-\mathbf{r}_1^*)\int d\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\text{c}}+u_{\text{o}}+u_{\text{z}}]}} &= \frac{\displaystyle\int_{\text{site}} d\mathbf{r}_1\ \rho(\mathbf{r}_1)\ \mathrm{e}^{-\beta u_{\text{a}}}}{\rho(\mathbf{r}_1^*)} \\
&= \frac{\displaystyle\int_{\text{site}} d\mathbf{r}_1\ \rho(\mathbf{r}_1)\ \mathrm{e}^{-\beta u_{\text{a}}}}{\rho(r_1^*,0,0)} \\
&= \frac{\displaystyle\int_{\text{site}} d\mathbf{r}_1\ \rho(\mathbf{r}_1)\ \mathrm{e}^{-\beta u_{\text{a}}}}{\rho(r_1^*,0,0)\times\dfrac{\int d\mathbf{r}_1\ \delta(r_1-r_1^*)\ \mathrm{e}^{-\beta u_{\text{a}}}}{\int d\mathbf{r}_1\ \delta(r_1-r_1^*)\ \mathrm{e}^{-\beta u_{\text{a}}}}} \\
&= L^*\frac{\displaystyle\int_{\text{site}} d\mathbf{r}_1\ \rho(\mathbf{r}_1)\ \mathrm{e}^{-\beta u_{\text{a}}}}{\rho(r_1^*,0,0)\displaystyle\int d\mathbf{r}_1\ \delta(r_1-r_1^*)\ \mathrm{e}^{-\beta u_{\text{a}}}} \\
&= L^*\frac{\displaystyle\int_{\text{site}} d\mathbf{r}_1\ \rho(\mathbf{r}_1)\ \mathrm{e}^{-\beta u_{\text{a}}}}{\displaystyle\int d\mathbf{r}_1\ \delta(r_1-r_1^*)\ \rho(\mathbf{r}_1)\ \mathrm{e}^{-\beta u_{\text{a}}}} \\
&= L^*\int_{\text{site}} dr_1'\left[\frac{\int d\mathbf{r}_1\ \delta(r_1-r_1')\rho(\mathbf{r}_1)\ \mathrm{e}^{-\beta u_{\text{a}}}}{\int d\mathbf{r}_1\ \delta(r_1-r_1^*)\rho(\mathbf{r}_1)\ \mathrm{e}^{-\beta u_{\text{a}}}}\right] \\
&= L^*\int_{\text{site}} dr_1'\ \mathrm{e}^{-\beta[w(r_1')-w(r_1^*)]}, \qquad (7.19)
\end{aligned}$$

where $w(r_1)$ is a PMF defined as a function of the cylindrical distance $r_1 = \sqrt{(x_1)^2+(y_1)^2}$ projected onto the 2D membrane plane calculated in the presence of the configurational and oriental restraints u_{c} and u_{o} (and also, if so desired, the z-depth restraint u_{z}),

$$\mathrm{e}^{-\beta[w(r_1')-w(r_1^*)]} = \frac{\displaystyle\int d\mathbf{1}\ \delta(r_1-r_1')\int d\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\text{c}}+u_{\text{o}}+u_{\text{a}}+u_{\text{z}}]}}{\displaystyle\int d\mathbf{1}\ \delta(r_1-r_1^*)\int d\mathbf{X}\ \mathrm{e}^{-\beta[U+u_{\text{c}}+u_{\text{o}}+u_{\text{a}}+u_{\text{z}}]}} \qquad (7.20)$$

and L^* is a length integral given by:

$$\begin{aligned} L^* &= \int d\mathbf{r}_1 \; \delta(r_1 - r_1^*) \; \mathrm{e}^{-\beta u_\mathrm{a}(\phi_1)} \\ &= r_1^* \int_0^{2\pi} d\phi_1 \; \mathrm{e}^{-\beta u_\mathrm{a}(\phi_1)}. \end{aligned} \tag{7.21}$$

It follows that the binding constant can be expressed as:

$$K_\mathrm{eq} = L^* \, I^* \, \mathrm{e}^{-\beta[G_\mathrm{c}^\mathrm{bulk} + G_\mathrm{o}^\mathrm{bulk} + G_\mathrm{z}^\mathrm{bulk} - G_\mathrm{a}^\mathrm{site} - G_\mathrm{o}^\mathrm{site} - G_\mathrm{c}^\mathrm{site} - G_\mathrm{z}^\mathrm{site}]} \tag{7.22}$$

By definition, K_eq calculated from MD in the standard units has the dimension of surface area (e.g., Å^2).

MD simulation protocols

The simulations in this study were conducted on servers equipped with 32 CPU cores and 2 GPUs (GeForce RTX 2080 Ti), using the NAMD3 MD engine for PMF calculations and NAMD 2.14 for pairwise force collection.[36] The visualization and snapshots were taken via VMD software.[85]

Geometrical route and equilibrium simulations. The temperature and the pressure were maintained at 303 K and 1 atm, respectively, using overdamped Langevin dynamics and the Langevin piston, respectively.[137,138] The equations of motion were integrated using a time step of 2 fs for the geometrical route free-energy calculations, with the exception of the physical separation PMF calculation, which used a time step of 4 fs to improve the computation time through

hydrogen-mass repartitioning (HMR).[288] Short-range interactions were smoothly turned to zero between 10 and 12 Å. The pair–list distance was 14 Å. For the long-range electrostatic interactions, the PME algorithm[139] was used. The coordinates and CHARMM36 force-field parameters for the lipids,[134] CHARMM36m for protein,[286,287] and the TIP3P water model[18] were used as inputs in the binding free-energy estimator 2 (BFEE2),[75] a tool based on the Colvars module[67] for streamlining and automating the setup of binding free-energy calculations, originally designed to tackle protein–ligand complexes.

To further broaden the scope of the BFEE2 tool[75] for protein-protein complexes within membranes, several refinements were implemented. Specifically, RMSD of the backbone calculations were conducted for each protein in the bound state. In the original automated setup of the BFEE2 tool, the inclusion of the polar Theta calculations was considered. This feature was designed to accommodate complexes in a homogeneous environment. However, in our specific case of a 2D membrane environment, the PMF calculation of the polar Theta angle was no longer necessary. Therefore, we opted to exclude the polar Theta calculations from our setup. Moreover, to account for the identical structures of the helices, only one RMSD of the unbound state is needed.

Additionally, to capture the impact of a non-homogeneous membrane environment, three additional calculations of the Euler angles (Θ, Φ, Ψ) in the unbound state were incorporated. To perform PMF calculations for the Euler angles in the bulk, the setup involved several steps. Firstly, a simulation box containing only one helix underwent re-solvation and re-neutralization. Subsequently, a thorough

pre-equilibration was carried out for a duration of 10 ns for each angle. Following the pre-equilibration, consecutive PMF simulations (for Θ, Φ, and Ψ) were conducted. Herein, for calculation of the energetic contributions rising from the orientational angles and conformational changes in the bulk restraints were introduced one by one:

1. RMSD unbound,
2. RMSD unbound + Θ,
3. RMSD unbound + Θ + Φ,
4. RMSD unbound + Θ + Φ + Ψ, where

1. for (1) the PMF as a function of the RMSD of the unbound state calculated with no other restraint,
2. for (2) the PMF as a function of Θ, calculated with the RMSD of the unbound state restraints,
3. for (3) the PMF as a function of Φ calculated with the RMSD of the unbound state and Θ restraints,
4. for (4) the PMF as a function of Ψ calculated with the RMSD of the unbound state, Θ, and Φ restraints.

It is important to note that as the reference, the configuration of the helix taken from the bound state was employed. In total, 11 consecutive calculation steps were applied to evaluate the binding free energy of the GpA complex.

Table 7.5: MD protocols used for unrestrained separation PMF calculations.

Index	Time step, fs	Temperature, K	Windows number	ns per window
A	4	308	4	4000
B	4	308	4	1500
C	4	303	6	800
D	2	310	26	100
E	4	308	6	300
F	4	303	4	400
G	4	308	5	340

Unrestrained physical separation PMF simulations. As in the case of the geometrical route, the unrestrained PMF simulations were carried out in the NPT ensemble with 1 atm for maintaining the constant pressure. However, the temperature conditions, as well as conditions for Langevin dynamics, differ. Table 7.5 summarizes the MD protocols used for unrestrained separation PMF calculations. The time step of 4 fs corresponds to the application of the HMR technique.[288]

Computational details

The starting coordinates for the simulation were obtained from the PDB entry 1AFO.[234] The resulting system comprised a total of 52,938 atoms, including 926 atoms of the $GpA_{69\text{-}97}$ homodimer, 23,985 atoms of the POPC lipid bilayer, and 27,978 atoms of water molecules. The remaining atoms consisted of sodium and chloride ions, which were added to ensure electric neutrality and a 150 mM

salt concentration, mimicking physiological conditions.[342] The dimensions of the periodic cells were 93 × 90 × 91 Å^3.

The convergence plots of the PMF calculations obtained in the geometrical route were obtained using the following steps: (i) generate the free energy gradient along ξ at every time step of the simulation, (ii) calculate the root-mean-square difference (k) of the gradient via:

$$k(t_i) = \sqrt{(\frac{\sum_{i=0}^{N}(\nabla(\xi, t_N) - \nabla(\xi, t_i))^2)}{N}},$$

where the time, t_i, varies in the range [0,t_N], and t_N is the final simulation time. More details about the gradient RMSD calculations can be found in reference.[105] k(t_i) asymptotically reaches 0 for every PMF, suggesting convergence of the calculations.

Results of the geometrical route

The results of the PMFs of each contribution for the geometrical route and the convergence plots are presented in Figures 7.7 and 7.8.

Jacobian in cylindrical coordinates

In cylindrical coordinates, a point is specified by its distance from the origin (r), its azimuthal angle (ϕ), and its height or displacement along the axis of the cylinder (z). The transformation equations from cylindrical coordinates (r, ϕ, z)

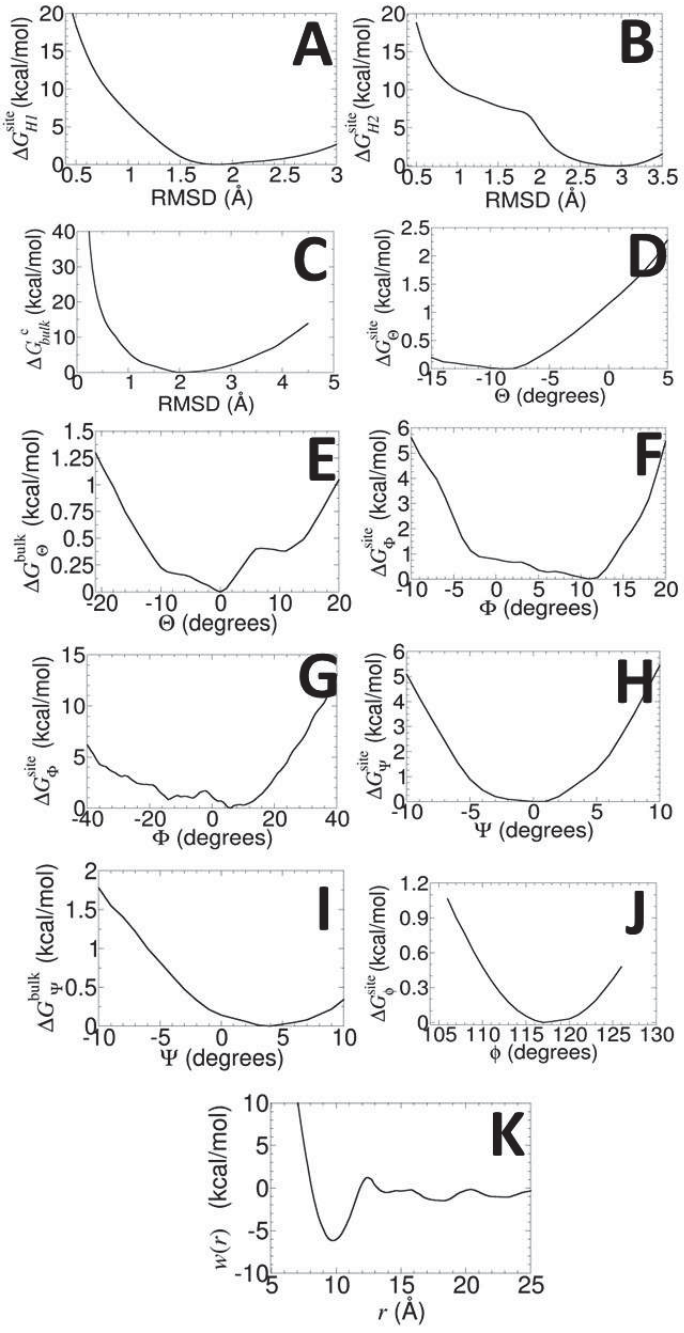

Figure 7.7: Individual PMFs for all components. The RMSDs in the bound state of H_1 (A) and H_2 (B) helices, RMSD of the unbound state (C), Θ in the bound state (D), Θ in the unbound state (E), Φ in the bound state (F), Φ in the unbound state (G), Ψ in the bound state (H), Ψ in the unbound state (I), azimuthal angle ϕ (J), and physical separation PMF $w(r)$ (K).

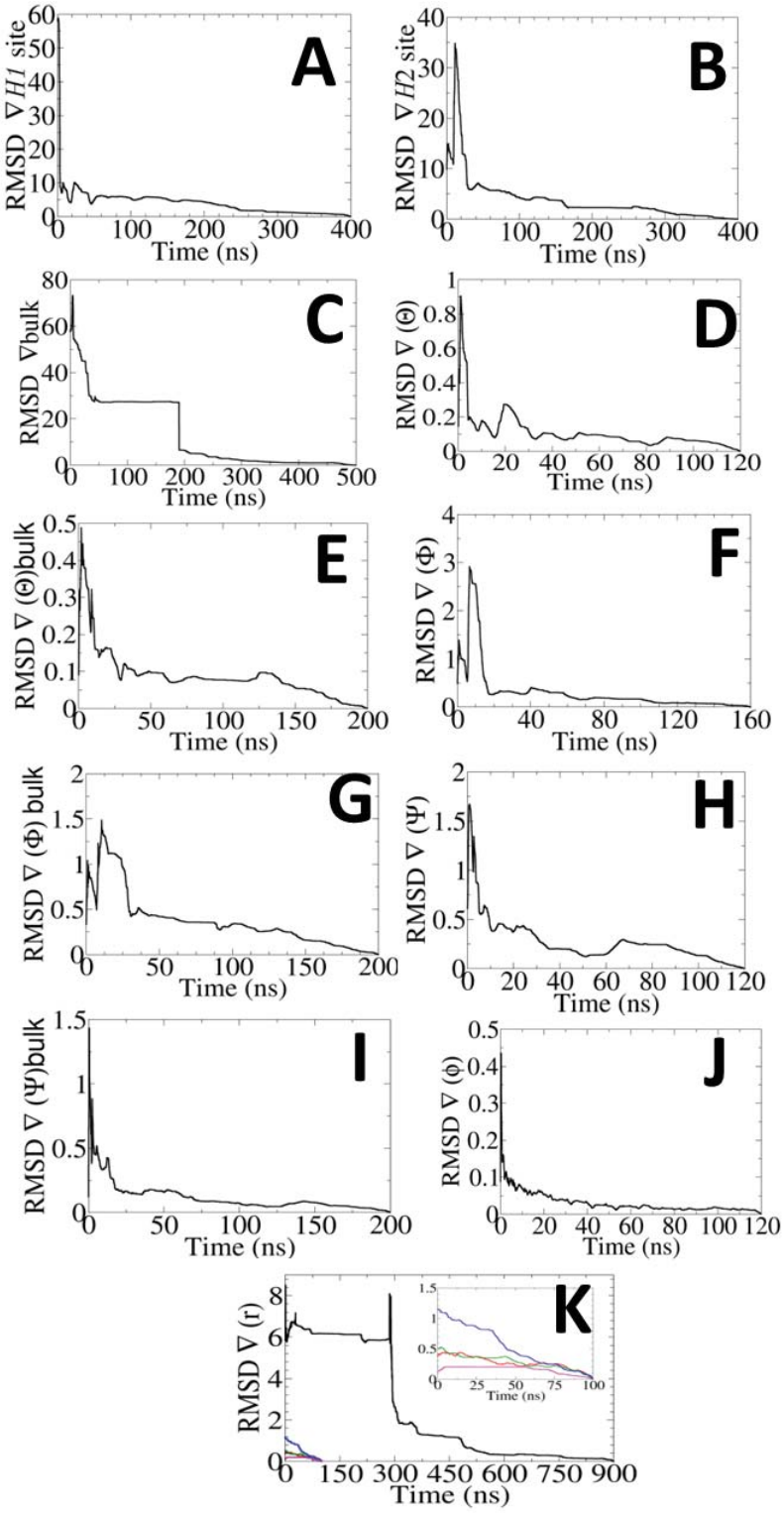

Figure 7.8: Convergence curves for individual PMFs for all components. The RMSDs in the bound state of H_1 (A) and H_2 (B) helices, RMSD of the unbound state (C), Θ in the bound state (D), Θ in the unbound state (E), Φ in the bound state (F), Φ in the unbound state (G), Ψ in the bound state (H), Ψ in the unbound state (I), azimuthal angle ϕ (J), and physical separation PMF $w(r)$ (K).

to Cartesian coordinates (x, y, z) are:

$$x = r\cos(\phi),\ y = r\sin(\phi),\ z = z.$$

When performing a coordinate transformation from Cartesian to cylindrical coordinates, we need to consider the Jacobian determinant introduced by the transformation. In this case, the determinant is equal to r, which corresponds to the radial coordinate in cylindrical coordinates, thus, the Jacobian is equal $-k_B T \ln(r^*)$.

Additional analysis of the metastable state

The analysis of the separation simulation showed that GpA at reversible association exhibits a metastable state. To further investigate this phenomenon, we conducted a comprehensive comparative analysis of the prevalent hydrogen bonds and their respective occupancies in both the metastable and the native structures, as depicted in Figure 7.9. The highest occupancy for both states relates to the nitrogen atom of SER69 of H_1 helix (donor) and oxygen of the GLU72 side-chain of H_2 helix (acceptor). However, the O—H $\cdots$ O of the SER69–GLU72 is observed only in the native structure of the GpA, which highlights the conformational differences of both states.

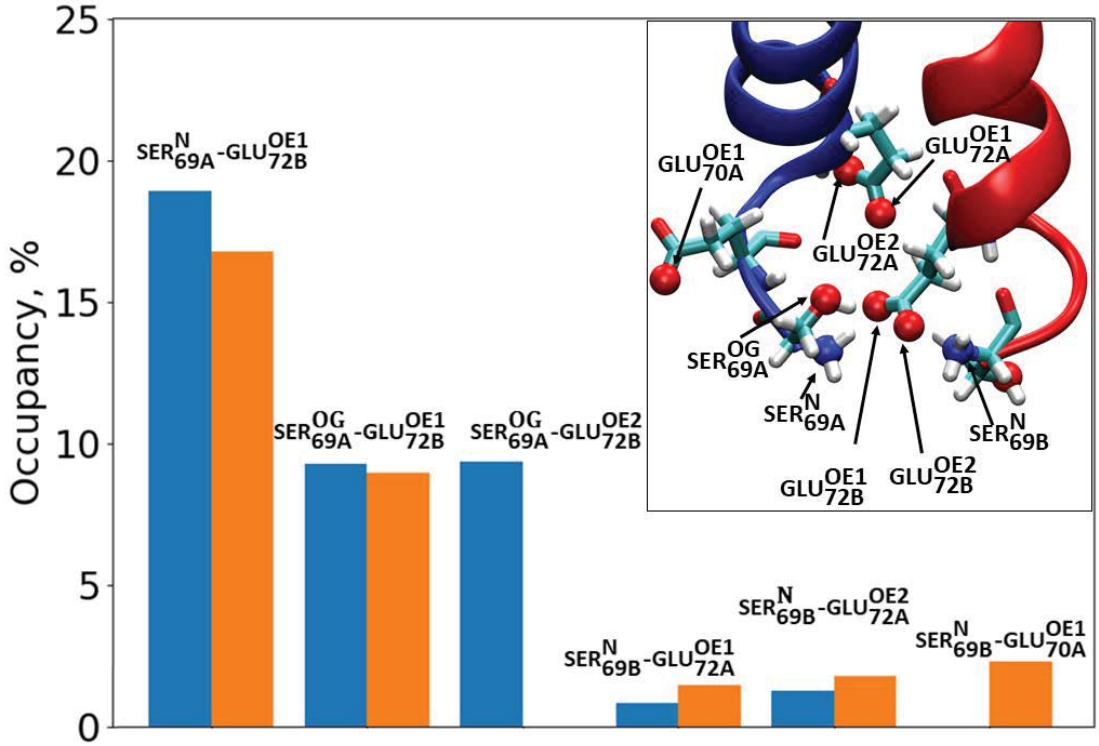

Figure 7.9: Occupancy of hydrogen bonds of native (blue) and non–native (orange) in 1 μs–equilibrium simulation. The superscripts correspond to the donor and acceptor atoms participating in hydrogen bond formation. The subscripts correspond to the number of the residue with index "A" for H_1 helix and "B" for H_2.

Chapter 8

Improving Speed and Affordability without Compromising Accuracy

8.1 Ways to Mitigate Computational Burdens and Enhancing Efficiency

Being a seamlessly reliable and robust approach for standard binding free energies calculation of different protein-ligand and protein-protein complexes,[74,75] the stepwise geometrical route has been often criticized for its high computational cost and time-consuming nature, hindering its widespread adoption in practical applications.[155,156,158,160] In the following study, we aimed to alleviate the burden of the substantial computational resources typically required by this methodology. Specifically, our focus was to investigate the application of strategies that might accelerate PMF calculations while preserving the accuracy and reproducibility of the geometrical route.

One of the reasons for the deceleration of the PMF simulations is the integration of the force-field terms. To reduce the computational cost of the MD simulations, two perspective approaches can be adopted. Firstly, the implementation of

Shake/Rattle/Settle algorithms,[343–345] effectively suppresses high-frequency fluctuations within hydrogen-containing bonds, thereby enabling swifter simulations. Secondly, the employment of a hydrogen-mass repartitioning (HMR) technique permits the redistribution of the mass of heavy atoms into the bonded hydrogen atoms, conserving the overall molecular mass.[288,346–351] In the HMR, the hydrogen atoms in methyl and methylene groups are typically assigned a mass of three g/mol to slow down their vibration during MD simulations.[347] This adjustment enables the utilization of longer integration time steps of up to four fs, without compromising energy conservation or thermodynamic properties.[288] Moreover, employing multiple time-stepping methods, such as Verlet-I/r-RESPA,[352,353] can further enable simulations with larger time steps (to eight fs) by updating an effective integration of short-range and long-range forces at different time scales.

The incorporation of CVs introduces additional reasons for the computational slowdown of PMF calculations. As discussed in Chapter 1, the evaluation of CVs typically involves the calculation of geometric properties, such as distances or angles, which need to be performed at every time step, adding an extra computational burden. Moreover, biases or restraints, applied to the CVs to drive the system towards desired states, can be in the form of different potentials (e.i., harmonic or more complex), which require additional energy evaluations and force calculations, contributing further to the computational overhead. To mitigate the impact of CVs on simulation speed, a multiple-time-stepping (MTS) strategy can be used.[354] The MTS allows computing CVs and corresponding biasing forces at a lower frequency, rather than every single MD step.[352,354,355] This is particu-

larly useful when evaluating CVs that involve the roto-translational alignments of large numbers of atoms, which can be computationally expensive.[67,356] In this technique, the CVs and biasing forces are calculated every N_{MTS} MD steps, where N_{MTS} is chosen by the user.

In this study, we implemented HMR in conjunction with the MTS technique to accelerate the PMF calculations. The value of N_{MTS}, indicating the frequency of CV and biasing force calculations, was set to every 2 or 4 steps. Our focus was specifically directed toward the most time-consuming aspect of the geometrical route, namely the physical separation PMF calculations. To ensure accuracy, sampling uniformity, and convergence rates, we fine-tuned the extended-Lagrangian parameters of the WTM-eABF[117] enhanced sampling algorithm (see Chapter 3). To evaluate the effectiveness of our accelerated calculations, we performed a total of fifty triplicate physical separation PMF simulations for the Abl kinase-SH3 domain:p41 complex[12,75,162] using various combinations of HMR and MTS techniques and compared the accuracy and convergence of these accelerated calculations with those of the classical approach of the geometrical route. In order to validate the results obtained with the best-performing configurations, we conducted five additional simulations. Furthermore, to showcase the combinations applicability to other protein-ligand complexes, we triplicated a 200-nanosecond separation simulation employing nine protocols for the MDM2-p53:NVP-CGM097 complex (see Chapter 4).[74,75,131] Our simulations, which cumulatively amounted to 14.4 μs of simulation time, allowed us to identify an optimal parameter set that achieved a threefold increase in convergence rate without significant loss of accuracy. The re-

sults of this study were published as Blazhynska et al., *J. Chem. Theory Comput.*, 2023.

8.2 Application of Acceleration Schemes to Abl kinase-SH3 domain:p41

As the technical part of our investigation, we streamlined the workflow by omitting the input-file preparation step in the BFEE2 tool[75,123] and focused on optimizing the prepped physical separation PMF calculations along the COM distance between the protein and the ligand. The energetic contributions from other degrees of freedom in the bound and unbound states were obtained from a previous study.[75] The input files were modified to incorporate the association of HMR and MTS along with three extended-Lagrangian parameters introduced previously in Chapter 2 (σ, γ_λ, and τ), as summarized in Table 8.1.

For the calculation of the physical separation PMFs, we employed the Colvars module,[67] dividing the COM distance between the two binding partners into 0.1-Å bins. The resulting separation PMFs were combined with the previously obtained conformational, orientational, and positional PMFs[75] and used as input for the post-treatment analysis using the BFEE2 software[75,357] to compute the final standard binding affinity, $\Delta G^\circ_\mathrm{b}$.

Table 8.1: Summary of Simulation Setups and Parameters

Scheme	HMR	N_{MTS}	σ (Å)	γ_λ (ps^{-1})	τ (fs)
1	No	2	0.1	1.0	200
2	No	4	0.1	1.0	200
3	Yes	1	0.1	1.0	200
4	Yes	2	0.1	1.0	200
5	Yes	4	0.1	1.0	200
Protocol	HMR	N_{MTS}	σ (Å)	γ_λ (ps^{-1})	τ (fs)
6	applied on scheme 4		0.1	3.0	200
7			0.1	5.0	200
8			0.1	7.0	200
9			0.1	10.0	200
10			0.01	1.0	200
11			0.05	1.0	200
12			0.5	1.0	200
13			0.1	1.0	100
14			0.1	1.0	300
15	applied on scheme 1,2,3,5		0.1	1.0	300
16			0.05	1.0	200
17			0.1	10.0	200
18			0.1	7.0	200
Scheme	HMR	N_{MTS}	σ (Å)	γ_λ (ps^{-1})	τ (fs)

19	applied on scheme 1–5	0.05	10.0	300
20		0.1	10.0	300
21		0.05	7.0	300
22		0.1	7.0	300

Probing the acceleration schemes. The outcomes of applying the acceleration schemes using the default extended-Lagrangian parameters (Table 8.1 (schemes 1-5)) are presented in Table 8.2. To ensure the reliability of the results, we performed three independent 50-ns physical-separation PMF calculations for the Abl kinase-SH3 domain:p41 complex for each scheme. The average PMFs are depicted in Figure 8.1A, and the uncertainty values are reported in Table 8.2 as standard deviations of the binding free energy calculated from the three independent replicas. As observed in Table 8.2, the faster schemes provide significant computational acceleration, achieving up to three times faster calculations compared to the reference PMF calculation. However, such increased speed comes at the cost of a noticeable loss in the accuracy of the final binding free-energy estimate. Scheme 4, for example, exhibits a large variance in the obtained PMFs and deviates from the reference $\Delta G^{\circ}_{\mathrm{b}}$ by 1.5 kcal/mol. Notably, the average $\Delta G^{\circ}_{\mathrm{b}}$ for schemes 1 and 3 are comparable and only about $k_{\mathrm{B}}T$ away from the experimental value.[162]

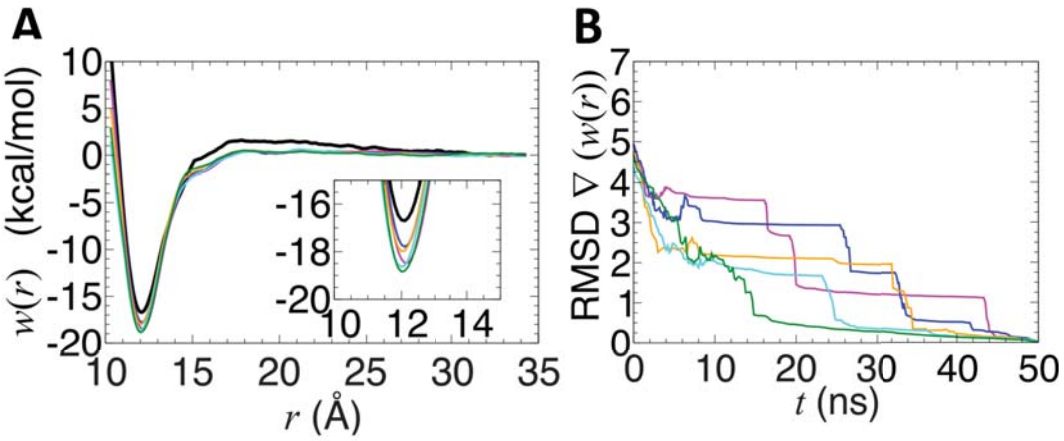

Figure 8.1: (A) Averaged physical-separation PMFs for three replicas obtained after individual 50–ns simulations. All the PMFs were determined within the separation distance range of [10.3; 34.3] Å. (B) Averaged convergences for the physical-separation PMFs. The curves correspond to the different calculation schemes: Reference (black), scheme 1 (blue), scheme 2 (magenta), scheme 3 (orange), scheme 4 (cyan), and scheme 5 (green). Reproduced with permission from *J. Chem. Theory Comput.*, **2023**, 19(11), 3091-3101, Copyright 2022 American Chemical Society.

The convergence progress was assessed by computing the RMSD between the free-energy gradients, as shown in Figure 8.1B. It can be observed that scheme 3 exhibits slightly faster convergence compared to scheme 1. Although schemes 4 and 5 show significant deviations from the experimental results, they converge more smoothly and rapidly than any other scheme, which can be attributed to a cumulative effect of suboptimal extended-Lagrangian parameters choice for this particular protein-ligand complex. To investigate this idea further, we analyzed the impact of individual extended-Lagrangian parameters on the physical separation PMF calculation for scheme 4, as it exhibits lower accuracy compared to other schemes, making it more sensitive to variations in the free-energy calculation parameters.

Table 8.2: Results of binding free–energy estimations within the averaged from three replicas 50–ns separation PMF simulation applying different computational schemes.

Scheme	Separation Contribution (kcal/mol)[a]	$\Delta G^{\circ}_{\mathrm{b}}$ (kcal/mol)	Speed (ns/day)
reference[b]	− 14.4[75]	-7.6 ± 0.4[75]	52
①			
1	−15.0		
2	−16.3	−8.5 ± 0.8	69
3	−14.7		
②			
1	−16.5		
2	−15.5	−9.1± 0.5	85
3	−15.9		
③			
1	− 15.7		
2	−15.2	−8.5 ± 0.3	105
3	−15.4		
④			
1	− 18.0		
2	−14.2	−9.1 ± 1.9	137
3	−15.7		
⑤			
1	− 16.0		
2	−16.9	−9.5 ± 0.5	163
3	−15.9		

[a]The separation contribution is mathematically derived[12] from the physical separation PMF calculated along the COM distance coordinate. [b]corresponds to the standard binding free-energy evaluation via the geometrical route without HMR or MTS

Damping factor for the Langevin dynamics of the extended variable.

The damping factor, γ_{λ}, was investigated as the extended-Lagrangian parameter, which is capable to enhance sampling efficiency.[122,358] Optimal selection of the damping factor plays a crucial role in accurately reproducing the self-diffusion

properties of the system and maintaining temperature conservation, thereby preventing any potential overheating of the extended variable.[358] To explore the impact of different damping factors on computational scheme 4, we conducted simulations while keeping all other extended-Lagrangian parameters at their default values. The results, reported in Table 8.3 (protocols 6-9), present the values of $\Delta G^{\circ}_{\mathrm{b}}$ obtained as the average over three independent replicas, accompanied by their corresponding standard deviations. Among these protocols, protocol 8 exhibited the best agreement with experimental data[162] and demonstrated a low standard deviation, indicating the high accuracy of this theoretical prediction.

To examine the sampling uniformity of our calculations, we examined the number of force samples per bin for the selected schemes at the end of the separation PMF simulation (Figure 8.2). Although the sampling was relatively uniform along the studied cases, it remained low, considering the length of the simulation. The smallest number of samples was observed for protocol 6, ($N \approx 20{,}000$), and the largest ones, for protocols 7 and 8 ($N \approx 32{,}000$). The observed difference could impact the convergence rate for these simulations (Figure 8.2C). Protocols 7 and 8 corresponded to the fastest convergence, whereas protocol 6 was unlikely to be converged. Overall, protocol 8 was found to be the best choice for the damping factor in accelerated calculations with HMR, using a 4-fs time step and MTS, updating the CVs every two steps.

Spring stiffness and oscillation period. Another way to enhance the accuracy of the standard binding free-energy calculations within proposed acceleration schemes is to tune a so-called "extended fluctuation" parameter, also known

as the coupling width, σ,[67,109] which regulates the standard deviation of the fictitious particle linked to the real CV. By changing this parameter, the stiffness of the spring that connects the real and fictitious degree of freedom and the mass of the latter can be altered (see eq. 2.4, Chapter 2). Lesage et al., by studying the reversible folding of deca-alanine in a vacuum using standard eABF, showed that overlarge σ values could lead to accelerated convergence yet at the price of undersampling of the reaction pathway regions.[109] Herein, we examined three different extended fluctuations values, namely 0.01, 0.05, and 0.5 Å, in combination with HMR and MTS with a time interval of 2 (protocols 10, 11, and 12), and compared our results with those obtained with the reference value, i.e., $\sigma = 0.1$ Å of scheme 4. We found that protocol 11 ($\sigma = 0.05$ Å) was the most consistent with the experimental target value. Further inspection revealed that protocol 12 ($\sigma = 0.5$ Å) had the fastest convergence rate and the deepest valley among the schemes tested, but it suffered from insufficient sampling at the 10–20–Å separation range (Figure 8.2D-F). Our research findings contradict those of Lesage et al.,[109] who found that decreasing the extended fluctuation parameter could improve the convergence rate in binding free-energy calculations. The differences in results could be attributed to various factors, including the differences in the environments, the use of the MTS algorithm, the number of restrained CVs used in calculating the physical separation PMF, and the use of different enhanced-sampling algorithms—namely, eABF vs. WTM-eABF.

The oscillation period, τ, is another parameter that influences the coupling between the extended degree of freedom and the CV. We conducted additional

examinations with two values of τ, namely 100 and 300 fs, while employing HMR and MTS with a time interval of 2 (protocols 13 and 14). These results were then compared to the reference value of $\tau = 200$ fs from scheme 4 (see Table 8.3 and Figures 8.2 G, H, I). Among the examined protocols, protocol 14 (with $\tau = 300$ fs) exhibited an improved accuracy compared to scheme 4. However, for protocol 13, a detailed analysis of the number of samples per bin revealed insufficient sampling at small separations ($10 \leq r \leq 17$ Å). Additionally, we observed that the convergence was notably slower in protocols 13 and 14 compared to reference scheme 4. Although the accuracy of $\Delta G^{\circ}_{\mathrm{b}}$ was enhanced in protocol 14, our findings suggest that $\tau = 200$ fs can be considered a reliable and safe choice for the current study-case, considering both accuracy and computational efficiency.

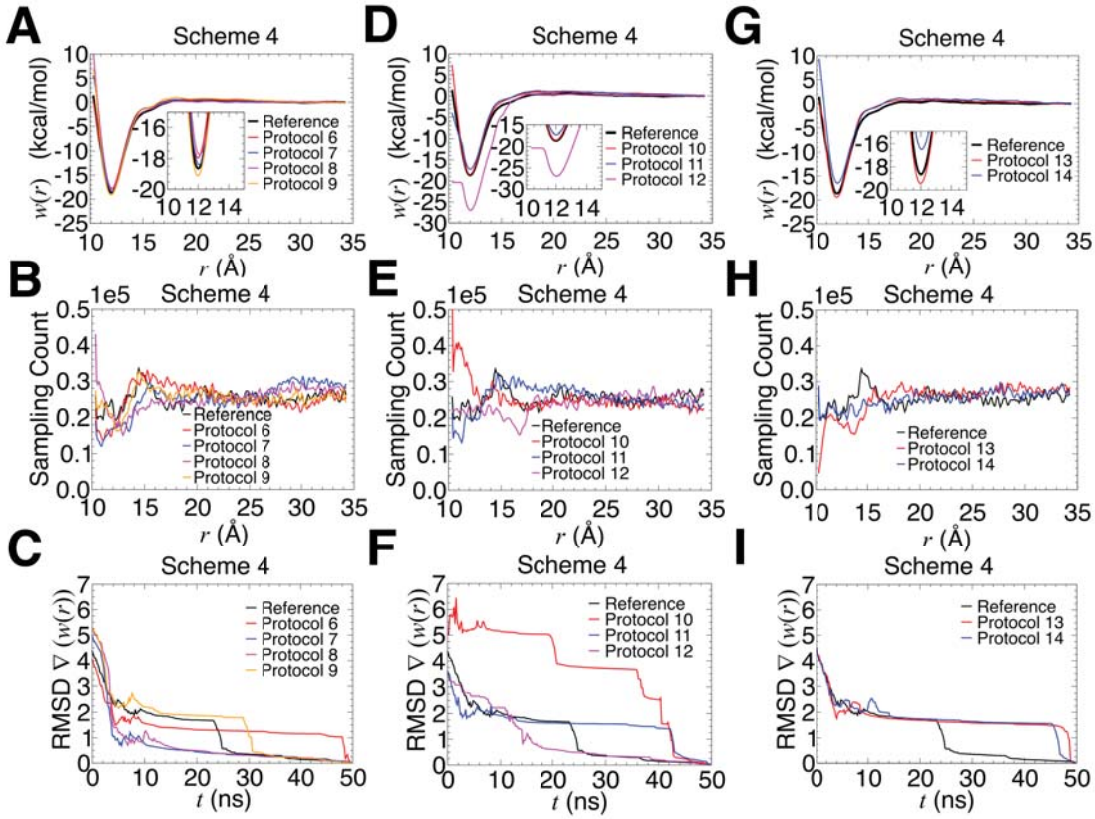

Figure 8.2: Panels A–C correspond to scheme 4 (denoted as Reference), and protocols 6, 7, 8, and 9 with $\gamma_\lambda = 1$ (black), 3 (red), 5 (blue), 7 (magenta), and 10 (orange) ps^{-1}. Panels D–F correspond to scheme 4, and protocols 10, 11, 12 with $\sigma = 0.1$ (black), 0.01 (red), 0.05 (blue) and 0.5 (magenta) Å. G–I correspond to scheme 4, and protocols 13, 14 with $\tau = 200$ (black), 100 (red), and 300 (blue) fs, respectively. (A, D, G) Averaged separation PMFs obtained after a triplicated 50–ns simulation. (B, E, H) An average number of samples per bin achieved in the simulations. (C, F, I) Average convergence rates achieved in the simulations. Reproduced with permission from *J. Chem. Theory Comput.*, **2023**, 19(11), 3091-3101, Copyright 2022 American Chemical Society.

Mixing different extended-Lagrangian parameters. After investigating the impact of individual extended-Lagrangian parameters on scheme 4, we selected the parameters that had the most significant influence and applied them independently and simultaneously to schemes 1, 2, 3, and 5. The details of these selected protocols (15–22) applied to the corresponding schemes (1–5) are summarized in Table 8.4, along with the default parameter values from Table 8.1. We examined the

Table 8.3: Results of binding free–energy estimations within averaged from three replicas 50–ns separation PMF simulation applying different damping factors, extended fluctuations, and oscillation periods to scheme 4.

Protocol	k (kcal/(mol· Å^2))	m_λ (kcal/(mol· ps^2 · Å^2))	ΔG°_b (kcal/mol)	Speed (ns/day)
Scheme 4[a]			-9.1 ± 1.9	137
⑥	59.6	6.0×10^4	-8.6 ± 1.0	142
⑦			-8.8 ± 0.5	141
⑧			-8.4 ± 0.1	141
⑨			-7.8 ± 0.9	140
⑩	5961.6	6.0×10^6	-9.5 ± 1.3	140
⑪	238.5	2.4×10^5	-8.1 ± 0.5	141
⑫	2.4	2.4×10^3	-17.6 ± 1.8	141
⑬	59.6	1.5×10^4	-10.2 ± 0.2	143
⑭		1.4×10^5	-8.4 ± 0.8	141

[a] the results of scheme 4 are duplicated from Table 8.2 to facilitate the comparison with the results of the application of protocols.

effects of the parameters $\tau = 300$ fs, $\sigma = 0.05$ Å, $\gamma_\lambda = 10$ and 7 ps^{-1} individually, while keeping all other parameters at their default values (protocols 15, 16, 17, and 18, respectively), as well as in various combinations, such as $\sigma = 0.05$ Å, $\tau = 300$ fs, $\gamma_\lambda = 10$ ps^{-1} (protocol 19), $\sigma = 0.1$ Å, $\tau = 300$ fs, $\gamma_\lambda = 10$ ps^{-1} (protocol 20), $\sigma = 0.05$ Å, $\tau = 300$ fs, $\gamma_\lambda = 7$ ps^{-1} (protocol 21), and $\sigma = 0.1$ Å, $\tau = 300$ fs, $\gamma_\lambda = 7$ ps^{-1} (protocol 22). the updated extended-Lagrangian parameter sets applied to schemes 1–5 are denoted with a slash (/), for example, combination 15/1 represents the application of protocol 15 to scheme 1 (-9.2 ± 0.4 kcal/mol

in Table 8.4).

According to Table 8.4, protocol 16 yields the highest accuracy for most of the schemes. Furthermore, combinations 18/3 and 18/4, as well as 19/2, 21/1, and 22/2, provide better agreement in the estimation of $\Delta G_{\mathrm{b}}^{\circ}$. However, only a few combinations yield results close to the experimental binding affinity, yet associated to high standard deviations, particularly 17/4, 21/2, and 22/3. These findings indicate that simultaneous changes in multiple parameters may lead to a decrease in the accuracy of the calculated binding free energy, whereas individual changes may significantly improve agreement with experimental data.

In order to assess the accuracy and precision of our results, we conducted further simulations by increasing the number of replicas for protocols 16, 18, and 21 up to five, since these protocols showed the most promising agreement with the available standard binding free-energy estimates (see Table 8.5). By performing quintuplicate simulations, we aimed to reduce the uncertainties in the calculated free energy differences and improve the overall accuracy of our predictions. The results of the quintuplicate simulations revealed smaller uncertainties for certain combinations, such as 18/2, 18/3, 18/5, 21/2, 21/4, and 21/5, compared to the triplicated simulations. However, for other combinations, the standard deviations remained within the same range or slightly exceeded it, indicating similar or slightly larger uncertainty. It is important to note that despite the increase in the number of replicas, the overall conclusion regarding the comparison of these protocols remains unchanged.

Table 8.4: Results of binding free–energy estimations within averaged from three replicas 50–ns separation PMF simulation applying selected protocols (15–22) to schemes 1–5.

Scheme	ΔG_b° (kcal/mol)			
	⑮	⑯	⑰	⑱
①	-9.2 ± 0.4	-8.9 ± 0.4	-7.5 ± 1.6	-8.3 ± 0.7
②	-9.4 ± 1.8	-8.1 ± 0.4	-6.2 ± 1.6	-10.3 ± 1.9
③	-9.1 ± 0.5	-7.4 ± 0.6	-8.4 ± 0.9	-7.9 ± 0.5
④	-8.4 ± 0.8	-8.1 ± 0.5	-7.8 ± 0.9	-8.4 ± 0.1
⑤	-9.1 ± 1.4	-9.5 ± 1.1	-9.2 ± 0.4	-8.6 ± 1.1
Scheme	⑲	⑳	㉑	㉒
①	-9.9 ± 0.4	-9.5 ± 1.1	-7.7 ± 0.7	-6.9 ± 0.6
②	-8.5 ± 0.7	-8.7 ± 1.0	-7.9 ± 1.0	-8.0 ± 0.6
③	-9.4 ± 0.9	-10.0 ± 1.5	-7.2 ± 0.3	-7.9 ± 0.9
④	-9.5 ± 0.5	-10.8 ± 0.2	-8.7 ± 1.2	-7.4 ± 1.2
⑤	-9.6 ± 0.6	-8.6 ± 1.5	-7.7 ± 0.9	-8.1 ± 1.0

An examination of the sampling behavior in the context of the extended-Lagrangian parameters (see Figure 8.3) reveals that schemes 1 and 3 have the highest number of samples per bin across most protocols (15-22), while scheme 5 has the lowest number of samples per bin due to its association of HMR with MTS, resulting in updating the CVs at a low frequency. However, at small separation distances, some combinations (15/2, 16/4, 17/5, 18/3, 19/5, 20/5, 20/3, 21/1, 21/4, and 22/5) exhibit non-uniformity in sampling. Interestingly, schemes 2 and 4 display nearly identical sampling efficiency for all protocols except for protocol

Table 8.5: Results of binding free–energy estimations within averaged from five replicas 50–ns separation PMF simulation applying selected protocols (16, 18, 21) to schemes 1–5.

Scheme	ΔG_b° (kcal/mol)		
	⑯	⑱	㉑
①	-9.2 ± 0.5	-7.8 ± 0.9	-7.9 ± 1.0
②	-7.9 ± 0.5	-9.6 ± 1.5	-8.1 ± 0.8
③	-7.4 ± 0.6	-7.7 ± 0.4	-7.9 ± 0.9
④	-8.1 ± 0.5	-8.0 ± 0.5	-8.6 ± 0.8
⑤	-8.6 ± 1.3	-8.3 ± 1.0	-7.7 ± 0.8

18, where scheme 2 shows the most uniform sampling with the largest number of samples per bin compared to other schemes (Figure 8.3C). Furthermore, the convergence analysis (as shown in Figure 8.4) indicates that scheme 2 has slower convergence than the other schemes, leading to an underestimation of binding affinity (e.g., protocol 17/2, -6.2 ± 1.6 kcal/mol, Table 8.5).

The convergence rates were found to be the highest for protocols 16 and 21, while convergence was not achieved for combinations 19/3, 15/4, and 20/1. Our investigation of parameter combinations showed that the damping coefficient, γ_λ, and the extended-fluctuation term, σ, can significantly affect the convergence of simulations (see Figure 8.4 E–H). To explore how changing the protocol can impact the efficiency of acceleration schemes, we analyzed the probability distributions of the differences between the real, ξ, and fictitious, λ, variables for physical separation simulations using combinations 21/3 and 20/4, which produced significantly different binding affinities of -7.2 ± 0.3 and -10.8 ± 0.2 kcal/mol, respectively

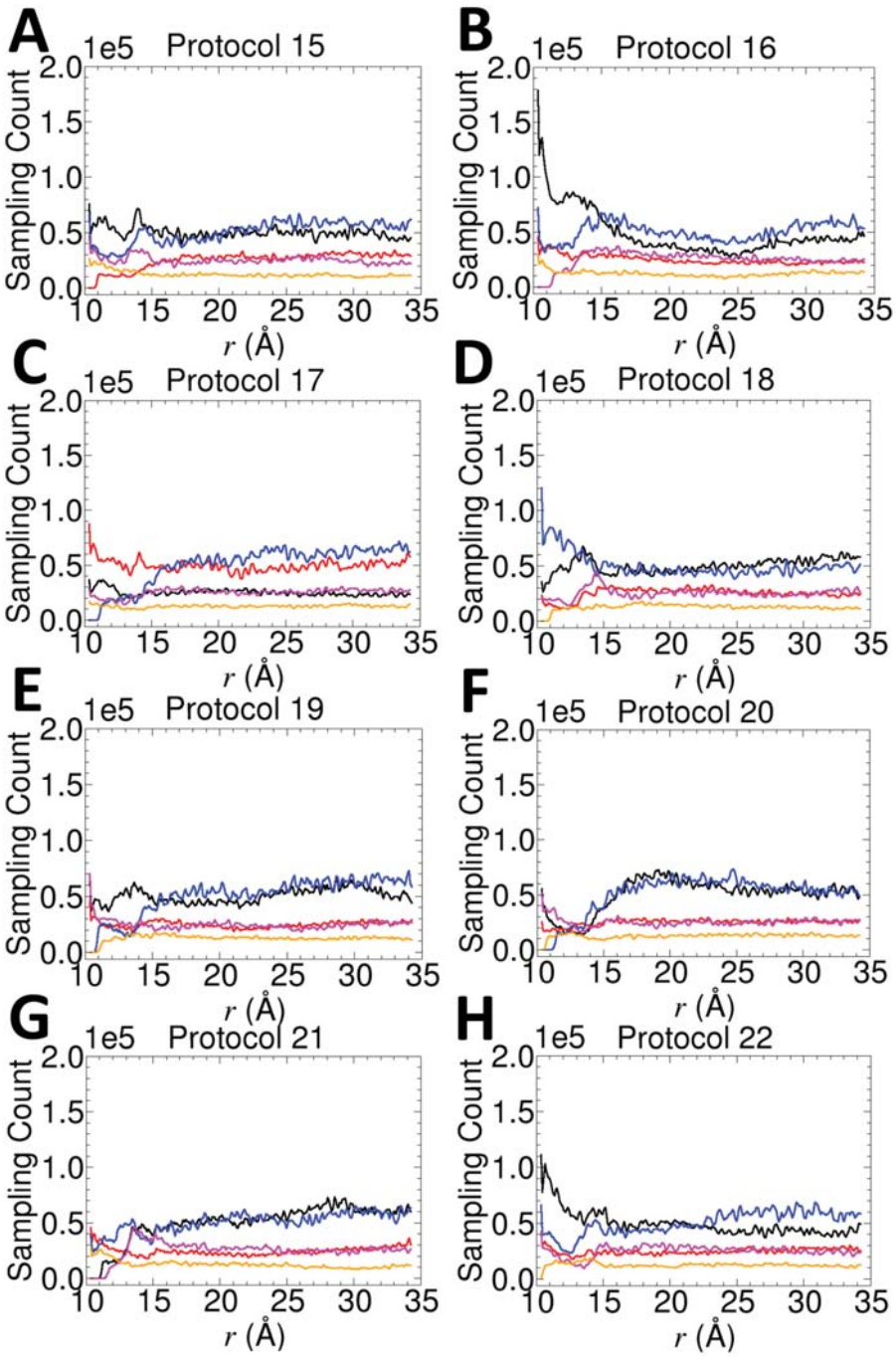

Figure 8.3: Number of samples per bin at the end of the 50–ns separation PMF simulation obtained for combination of protocols (15-22) with schemes 1–5: (A) protocol 15 (τ = 300 fs), (B) protocol 16 (σ = 0.05 Å), (C) protocol 17 (γ_λ = 7 ps^{-1}), (D) protocol 18 (γ_λ = 10 ps^{-1}), (E) protocol 19 (τ = 300 fs, σ = 0.05 Å and γ_λ = 10 ps^{-1}), (F) protocol 20 (τ = 300 fs, σ = 0.1 Å and γ_λ = 10 ps^{-1}), (G) protocol 21 (τ = 300 fs, σ = 0.05 Å and γ_λ = 7 ps^{-1}), (H) protocol 22 (τ = 300 fs, σ = 0.1 Å and γ_λ = 7 ps^{-1}). The different curves correspond to: Scheme 1 (black), scheme 2 (red), scheme 3 (blue), scheme 4 (magenta), and scheme 5 (orange). Reproduced with permission from *J. Chem. Theory Comput.*, **2023**, 19(11), 3091-3101, Copyright 2022 American Chemical Society.

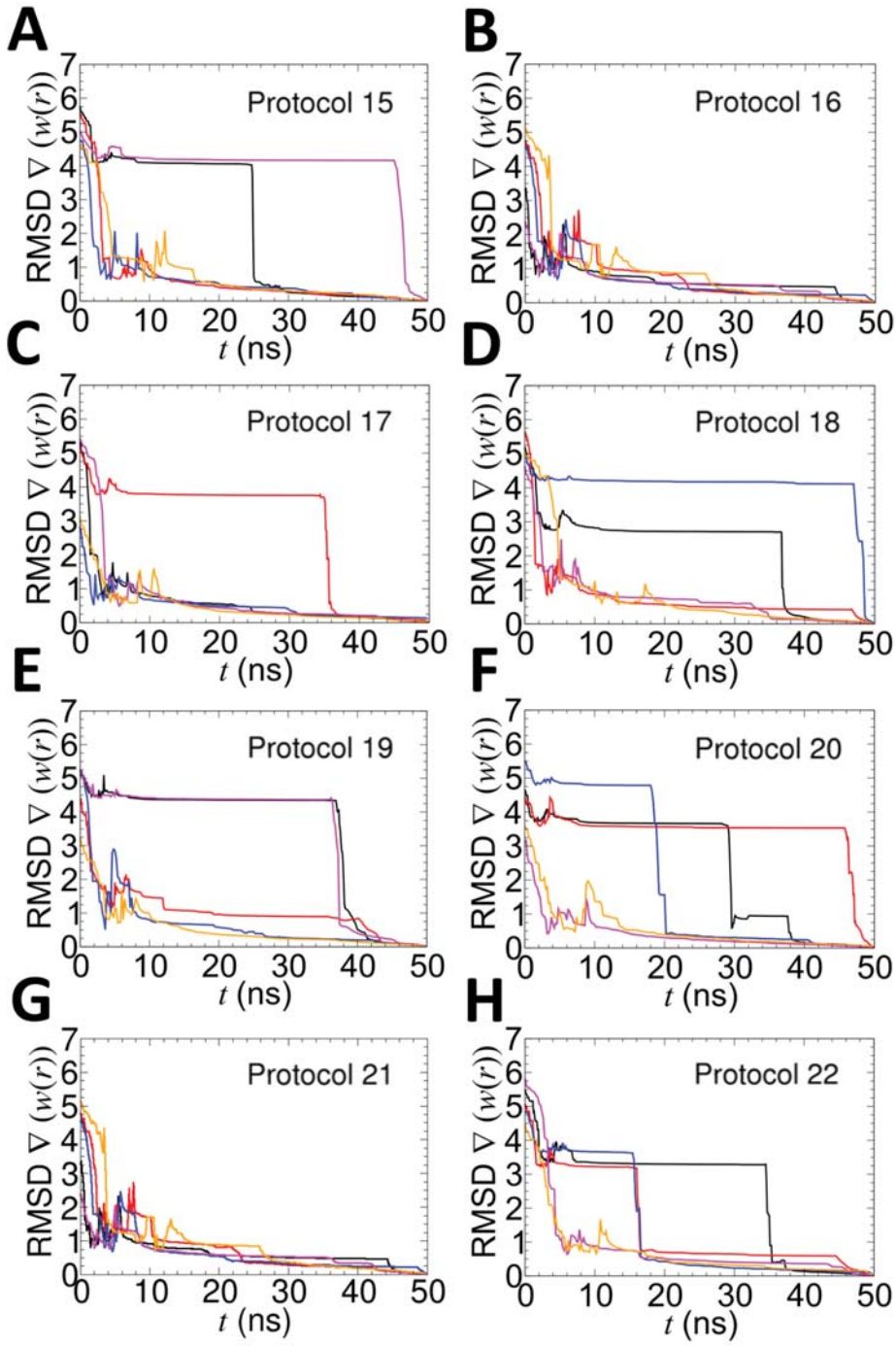

Figure 8.4: Convergence properties of the PMF calculations. (A) protocol 15 (τ = 300 fs), (B) protocol 16 (σ = 0.05 Å), (C) protocol 17 (γ_λ = 7 ps^{-1}), (D) protocol 18 (γ_λ = 10 ps^{-1}), (E) protocol 19 (τ = 300 fs, σ = 0.05 Å and γ_λ = 10 ps^{-1}), (F) protocol 20 (τ = 300 fs, σ = 0.1 Å and γ_λ = 10 ps^{-1}), (G) protocol 21 (τ = 300 fs, σ = 0.05 Å and γ_λ = 7 ps^{-1}), (H) protocol 22 (τ = 300 fs, σ = 0.1 Å and γ_λ = 7 ps^{-1}). The curves correspond to: Scheme 1 (black), scheme 2 (red), scheme 3 (blue), scheme 4 (magenta), and scheme 5 (orange). Reproduced with permission from *J. Chem. Theory Comput.*, **2023**, 19(11), 3091-3101, Copyright 2022 American Chemical Society.

(see Table 8.4 and Figure 8.5A, B). In combination 21/3, the Gaussian probability distribution indicates the synchronization of ξ and λ along the reaction pathway (see Figure 8.5A). The time-evolution of the separation coordinate, as well as that of the harmonically restrained degrees of freedom Θ and Φ (see Figure 8.5C), confirm the expected behavior of the CV throughout the physical separation. Since convergence was not fully attained within 50 ns for combination 21/3, we assume that additional sampling could improve the estimation of $\Delta G^{\circ}_{\mathrm{b}}$ (see Figure 8.4G, blue).

Contrary, combination 20/4 results in a positively skewed distribution that indicates weak coupling between the real and fictitious variables, which is further supported by the drift in the time-evolution of the separation coordinate, r, and its impact on the harmonically restrained angular CVs, Θ and Φ, between 30–40 ns as shown in Figure 8.4D. Notably, the convergence for this combination was achieved at an early stage of 40 ns, as demonstrated in Figure 8.4F (magenta).

In summary, the results of this study underscore the substantial impact that the choice of protocols and acceleration schemes can have on the accuracy, convergence rate, and number of samples per bin required for physical-separation PMF calculations. For instance, for scheme 1, all the modifications of the extended Langevin parameters did not enhance any of the discussed criteria, emphasizing the fragility of the application of only MTS with a frequency of 2 for the calculation of the CV's energetic contributions. In contrast, the combination of MTS with a frequency of 4 and a small value of σ (0.05 Å) in scheme 2 yields accurate results i.e., combination 16/2, while keeping all the remainder parameters as the

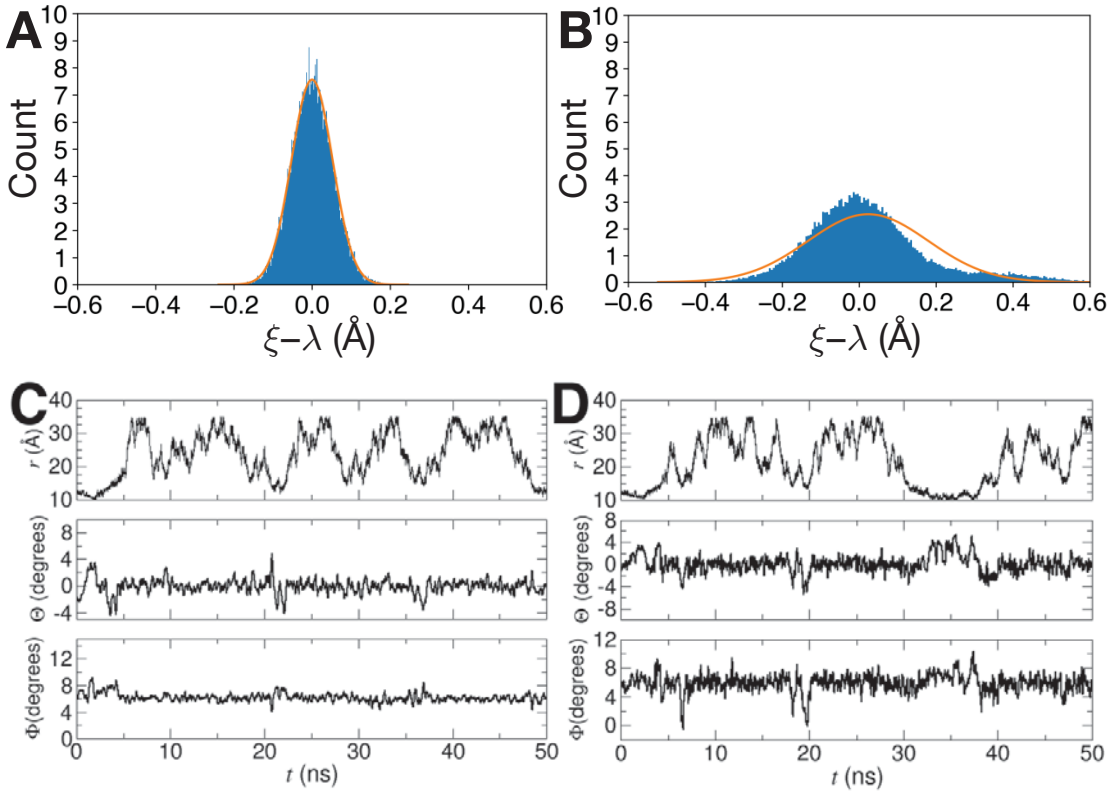

Figure 8.5: (A, B) Probability distribution of the difference between the real, ξ, and the fictitious, λ, particles. The orange curves correspond to a true Gaussian distribution fitted to the data. The plots are built with the help of the matplotlib Python library.[359] (C, D) Running averages of the CVs, namely, r, the separation, Θ, and Φ, the Euler angles. (A, C) correspond to combination 21/3 (τ = 300 fs, γ_λ = 7 ps^{-1} and σ = 0.05 Å) and (B, D) to combination 20/4 (τ = 300 fs, γ_λ = 10 ps^{-1} and σ = 0.1 Å) (see Table 8.4). Reproduced with permission from *J. Chem. Theory Comput.*, **2023**, 19(11), 3091-3101, Copyright 2022 American Chemical Society.

default values (γ_λ = 1.0 ps^{-1}, τ = 200 fs). However, modifying the oscillation period to 300 fs (combination 15/2) or damping factor to 7, or 10 ps^{-1} (combinations 17/2 or 18/2) alone leads to erroneous estimates due to increased inertial mass or overdamping. The use of updated extended-Lagrangian parameters such as combination 19/2 (σ= 0.05 Å and τ = 300 fs, and γ_λ = 10 ps^{-1}) and combination 22/2 (σ = 0.1 Å, γ_λ = 7.0 ps^{-1} and τ = 300 fs) can improve accuracy, but careful use is needed as they may result in fewer force samples and slower convergence.

For scheme 3 (only HMR application), the default values for the extended parameters are the most suitable, as they align with the goal of the HMR method to maintain simulation stability and complex structural properties.[288,347] In contrast, for schemes 4 and 5, utilizing different protocols may lead to a compromised reproduction of the physical-separation PMF. While in scheme 4, reducing the extended fluctuation (σ= 0.05 Å) or increasing the damping factor for extended-Lagrangian dynamics (up to γ_λ = 7.0 ps^{-1}) when keeping τ at its default value can mitigate this issue. It's important to note that fine-tuning the relevant parameters alone may not be sufficient to obtain an accurate binding free-energy estimate. Additional efforts and considerations, such as the choice of acceleration schemes and protocols, are likely necessary to achieve reliable results.

8.3 Application of Acceleration Schemes to MDM2-p53:NVP-CGM097

To further validate our findings, we applied the same methodology to the MDM2-p53:NVP-CGM097 protein-ligand complex that has been previously studied in our group.[75,131] We performed physical separation PMF calculations for this complex using various computational schemes, the detailed analysis of which is presented in the next subsection. The HMR method with default values for the extended-Lagrangian parameters (scheme 3) was found to provide the closest agreement with the experimental data and was 1.5 times faster than using MTS alone (schemes 1 and 2). It should be noted that the efficiency of the MTS algo-

rithm may depend on the complexity of the protein-ligand complex being studied. We also investigated the impact of selected extended-Lagrangian parameters on the accuracy/speed ratio for scheme 4 (protocols 7, 8, 11, and 14) and found that a damping factor of 7 ps^{-1} (protocol 8) was the best choice for this acceleration scheme applied to the MDM2-p53:NVP-CGM097 complex. Our results emphasize the importance of exercising caution when adjusting the extended-fluctuation parameter (protocol 11) as it could overly restrain the CV and affect association during the simulation, leading to a loss of accuracy in the standard binding free-energy calculations. Conversely, increasing the oscillation period value to 300 fs (protocol 14) increased the efficiency of the simulation at the expense of accuracy for this particular complex. It should be emphasized that the results obtained from the MDM2-p53:NVP-CGM097 complex may not be applicable to other protein-ligand complexes. Therefore, it is crucial to explore various acceleration schemes and protocols systematically to achieve dependable and precise outcomes in standard binding free-energy calculations.

Detailed Analysis for MDM2-p53:NVP-CGM097 complex. The results of binding free-energy estimations for the MDM2-p53 protein-ligand complex using different computational schemes are summarized in Table 8.6. The reference in the table corresponds to the result obtained from the geometrical route, without any acceleration.[75] The results obtained from the different schemes (circled 1–5, 7, 8, 11, 14) presented in this study. The simulations were carried out for a duration of 200 ns, and the outcomes were averaged over three replicas. We chose the simulation length of 200 ns based on the computational time required for convergence

of the separation PMF calculations in the reference standard geometrical route.[75]

In contrast to the Abl-SH3:p41 complex discussed earlier, the results for the MDM2-p53:NVP-CGM097 complex suggest that schemes 1 and 2, which accelerate MTS by a factor of 2 and 4, respectively, provide only minimal acceleration of the separation PMF simulations. These observations suggest that the efficiency of the MTS algorithm for accelerating standard binding free-energy calculations may depend on various factors, including the complexity of the protein-ligand complex, the number of atoms in the system, and the nature of interactions between the moieties. Larger and more complex biological objects like the MDM2-p53:NVP-CGM097 complex require extensive sampling to reach convergence and capture the protein-ligand interactions accurately (200 ns), which may limit the potential benefits of MTS acceleration compared to smaller systems like Abl-SH3:p41 complex (50 ns).

The most accurate agreement with the reference value for the MDM2-p53:NVP-CGM097 protein-ligand complex was achieved by scheme 3, which only uses HMR (Figure 8.6). This scheme provides a binding free-energy estimation of -11.3 ± 0.5 kcal/mol, falling within $k_{\mathrm{B}}T$ of the experimental value (-11.8 kcal/mol).[75] In contrast to the other schemes, scheme 3 has reached convergence at the end of the 200-ns simulations (Figure 8.6B), demonstrating its effectiveness in accurately capturing the protein-ligand interactions. These results suggest that default values for the extended-Lagrangian parameters work best with HMR, which is consistent with the objective of HMR to maintain the stability and structural properties of the complex during simulations.[288,347]

Furthermore, we investigated the impact of various extended-Lagrangian parameters on the accuracy/speed ratio of scheme 4 by applying protocols 7, 8, 11, and 14 (as shown in Figures 8.6, 8.8, and 8.9). Our results suggest that protocol 8, which employs a γ_λ value of 7 ps^{-1}, provides the best balance between accuracy and efficiency. To specify, protocol 8 exhibited the most uniform sampling across the reaction path (Figure 8.8) and showed the smoothest and fastest convergence (Figure 8.9), explaining its superior performance in predicting the experimental and reference standard binding free energies.

In contrast, the results of the impact of the selected extended-Lagrangian parameters on the accuracy and efficiency of scheme 4 showed that adjusting the extended fluctuation parameter from the default value of 0.1 to 0.05 Å(protocol 11) could potentially affect the accuracy of the standard binding free-energy calculations by tightening the applied restraint on the COM distance CV. The observed relatively high standard deviation in the calculated ΔG°_b, despite the close agreement with the reference values, suggests that caution should be taken when adjusting the σ parameter. On the other hand, applying protocol 14 with $\tau = 300$ fs increased efficiency but resulted in the underestimation of the calculated free-energy due to the more inertial fictitious particle affecting the sampling of the free-energy landscape, as shown in Figure 8.8C and Figure 8.7 and summarized in Table 8.6.

Table 8.6: Results of binding free–energy estimations within the averaged from three replicas 200–ns separation PMF simulation applying different computational schemes.

Scheme	Separation Contribution (kcal/mol)	$\Delta G^{\circ}_{\mathrm{b}}$ (kcal/mol)	Speed (ns/day)
reference[a]	−17.9[75]	−11.3 ± 0.9[75]	36.0
①			
1	−13.3		
2	−11.6	−7.0 ± 2.0	38.5
3	−15.9		
②			
1	−15.9		
2	−17.4	−8.8 ± 2.3	42.2
3	−12.8		
③			
1	−18.0		
2	−18.3	−11.3 ± 0.5	63.5
3	−17.3		
④			

Table 8.6 Continued from previous page

Scheme	Separation Contribution (kcal/mol)	$\Delta G_{\mathrm{b}}^{\circ}$ (kcal/mol)	Speed (ns/day)
1	−18.8		
2	−18.3	−11.2 ± 1.3	76.7
3	−16.3		
⑤			
1	− 18.3		
2	−17.4	−10.6 ± 1.2	88.8
3	−15.8		
⑦			
1	−17.3		
2	−18.7	−11.0 ± 1.0	77.6
3	−16.8		
⑧			
1	−18.6		
2	−18.3	−11.6 ± 0.4	75.3
3	− 17.8		
⑪			
1	−16.7		
		−10.5 ± 0.6	74.9

Table 8.6 Continued from previous page

Scheme	Separation Contribution (kcal/mol)	$\Delta G^{\circ}_{\mathrm{b}}$ (kcal/mol)	Speed (ns/day)
2	−17.7		
3	−16.8		
⑭			
1	−15.5		
2	−15.2	-8.6 ± 0.4	78.3
3	−14.8		

[a]corresponds to the standard binding free-energy evaluation via the geometrical route without HMR or MTS

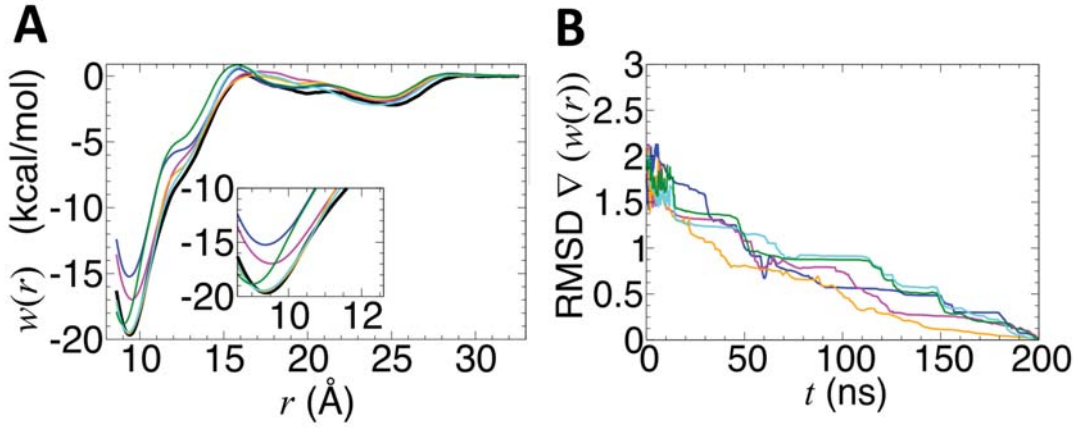

Figure 8.6: (A) Averaged physical separation PMFs for three replicas obtained after individual 200–ns simulations. All the PMFs were determined within the separation distance range of [8.6; 32.6] Å(B) Averaged convergences for the physical separation PMFs. The curves correspond to the different calculation schemes: reference (black),[75] scheme 1 (blue), scheme 2 (magenta), scheme 3 (orange), scheme 4 (cyan), and scheme 5 (green). Reproduced with permission from *J. Chem. Theory Comput.*, **2023**, 19(11), 3091-3101, Copyright 2022 American Chemical Society.

8.4 Unveiling the Potential of HMR and MTS in the Geometrical Route

Overall, our study demonstrated that HMR alone can significantly accelerate binding affinity calculations, achieving nearly a twofold speedup compared to the physical separation PMF calculation following a classical setup. Importantly, this acceleration does not compromise the accuracy or trajectory stability of the simulations, effectively reproducing the standard binding affinity.

On the other hand, the combined schemes of HMR and MTS, as well as MTS alone, require careful tuning of the extended-Lagrangian parameters to ensure the accurate reproduction of experimental results. By increasing the damping factor or

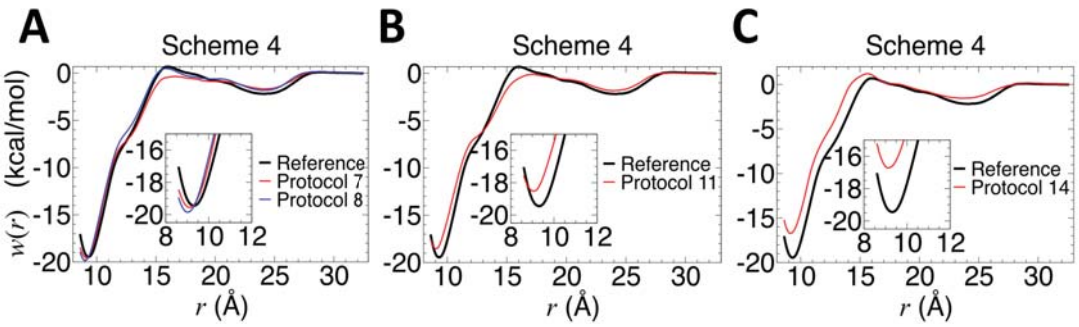

Figure 8.7: Averaged physical separation PMFs for three replicas obtained after individual 200–ns simulations. All the PMFs were determined within the separation distance range of [8.6; 32.6] Å(A) correspond to scheme 4 (black–denoted as Reference), and protocol 7 and 8 with $\gamma_\lambda = 5$ (red) or 7 (blue) ps^{-1}, respectively. (B) corresponds to scheme 4 (black–denoted as Reference) and protocol 11 with $\sigma = 0.05$ Å (red) and (C) corresponds to scheme 4 (black–denoted as Reference) and protocol 14 with $\tau = 300$ fs (red). Reproduced with permission from *J. Chem. Theory Comput.*, **2023**, 19(11), 3091-3101, Copyright 2022 American Chemical Society.

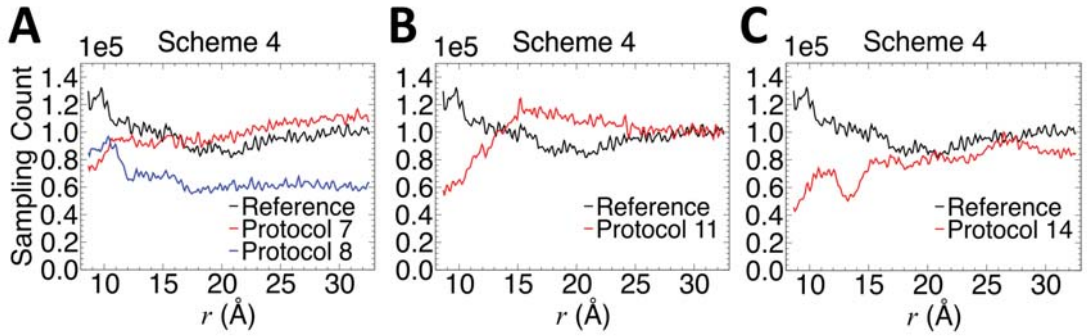

Figure 8.8: Average number of samples per bin achieved for three replicas obtained after individual 200–ns simulations. (A) correspond to scheme 4 (black–denoted as Reference) and protocol 7 and 8 with $\gamma_\lambda = 5$ (red) or 7 (blue) ps^{-1}, respectively. (B) corresponds to scheme 4 (black–denoted as Reference) and protocol 11 with $\sigma = 0.05$ Å (red) and (C) corresponds to scheme 4 (black–denoted as Reference) and protocol 14 with $\tau = 300$ fs (red). Reproduced with permission from *J. Chem. Theory Comput.*, **2023**, 19(11), 3091-3101, Copyright 2022 American Chemical Society.

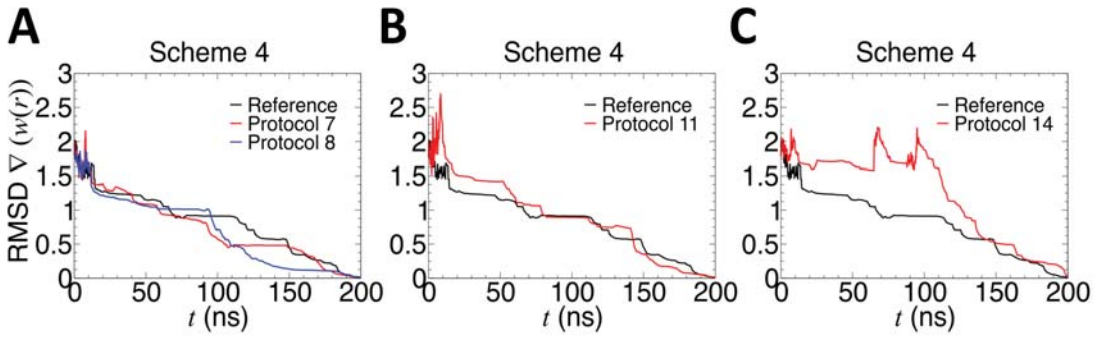

Figure 8.9: Average convergence rates achieved for three replicas obtained after individual 200–ns simulations. (A) correspond to scheme 4 (black–denoted as Reference) and protocol 7 and 8 with $\gamma_\lambda = 5$ (red) or 7 (blue) ps^{-1}, respectively. (B) corresponds to scheme 4 (black–denoted as Reference) and protocol 11 with $\sigma = 0.05$ Å (red) and (C) corresponds to scheme 4 (black–denoted as Reference) and protocol 14 with $\tau = 300$ fs (red). Reproduced with permission from *J. Chem. Theory Comput.*, **2023**, 19(11), 3091-3101, Copyright 2022 American Chemical Society.

decreasing the extended fluctuation, we observed a significant improvement in the performance of free-energy calculations, resulting in a nearly threefold speedup compared to the reference physical separation PMF simulation. However, the efficiency improvement with MTS alone was more modest.

It is crucial to note that despite having an accurate force field and reliable structural data, the specificity of each protein-ligand or protein-protein complex should not be overlooked. Therefore, an optimization of the extended-Lagrangian parameters discussed in this study should be conducted as a preliminary step for accelerated standard binding free-energy calculations using the geometrical route for obtaining accurate and reliable results.

Conclusion and Perspectives (English version)

In the inaugural year of my doctoral studies, I embarked on a journey to comprehend the nuances of binding free-energy alterations in protein-ligand complexes. My primary objective was to gain an understanding of standard binding free-energy calculations, and in pursuance of this goal, I participated in the elaboration of a protocol for standard binding free-energy calculations leveraging the BFEE2 software. I executed this innovative tool on three distinct complexes, and the results were promising, to say the least.

The statistical mechanical framework of this protocol is based on the restraint MD-based approach, initially proposed by Woo and Roux,[29] which has been subsequently developed into geometrical and alchemical routes.[12] The salient feature of this approach is the reduction of the numerous configurational changes occurring during binding to a few structural parameters, CVs, corresponding to the slow degrees of freedom associated with the relative movements of the binding partners during the MD simulation. The geometrical restraints acting on these CVs restrict the configurational space, leading to the accelerated convergence of the free-energy calculation.

With the experimental knowledge of the protein-ligand complex's bound state, the BFEE2 application, coupled with a reliable force field, delivers standard bind-

ing free energies with chemical accuracy. Moreover, this protocol significantly reduces the need for human intervention, as it efficiently helps the end-user to prepare all the necessary input files and to perform the post-treatment to obtain the final estimate of the binding affinity.

Since I worked with restraint–based binding free–energy approaches for protein-ligand complexes, specifically with alchemical and geometrical routes,[12,75] I decided to apply this strategy to more difficult cases, such as protein-protein interactions. In the case of protein-protein complexes, the application of the alchemical route is no longer possible because of the large perturbations of the protein-protein complexes involved in the decoupling of the protein, thus, creating difficulties in reaching convergence of the binding free-energy calculations.[46] Instead, the geometrical route, where the two molecular objects are progressively separated from one another in the presence of orientational, positional, and conformational restraints on both proteins introduced gradually, serves to control the change in configurational entropy that accompanies the dissociation process, thus, allowing the computations to converge within affordable simulation times. As my first case study, I examined a pig insulin dimer, in which the dimerization is driven by hydrophobic interactions of both monomers' interfaces and the experimental binding affinity was already established in the literature, so I was able to compare the calculated estimate with the well-documented one.[143,144]

To explore the accuracy and convergence of binding free-energy calculations for protein-ligand and protein-protein complexes, I conducted methodological research on standard binding free-energy calculations and compared the rigorous

theoretical framework of the geometrical route with its computational shortcut, where all other degrees of freedom were unrestrained during the physical separation of the partners. The results of this research revealed that the geometrical route, despite its relatively high computational cost, provides full control over every degree of freedom in the reversible association process through a universal set of CVs. By analyzing the corresponding PMF, the contribution of these degrees of freedom to the binding free energy can be determined with a high level of accuracy. In contrast, the shortcut approach lacks systematic theoretical justification and often leads to deviations from experimental values.

Although setting up the various steps of the protocol, especially defining the relevant CVs, may seem challenging and time-consuming, tools such as the CHARMM-GUI[207] and the earlier discussed BFEE2 software can alleviate the burden.[75,357] In contrast to the robust geometrical route, its shortcut approach lacks systematic theoretical justification, leading to results that deviate from experimental values. Our study also highlights that MD simulations of the typical length of one μs are insufficient for achieving adequate sampling of available configurational space in protein-ligand and protein-protein complexes, resulting in inaccurate binding affinity estimates.

It should be noted that while the geometrical route ensures optimal convergence properties of binding free-energy calculations and fully reproducible results, the limited accuracy of the force field and the reliability of the native binding motif are potential sources of discrepancies with experiment, making it the weak point of all binding free-energy calculations in general. One important takeaway from

this study is that, despite its allure of simplicity, the shortcut approach can be problematic, while the computational cost of the rigorous theoretical framework of the geometrical route is significantly more manageable. Given the lack of theoretical justification for the simplified approaches, there is no compelling reason not to use the geometrical route to obtain precise estimates of binding affinity.

Building upon the focus of my Ph.D. research on evaluating the efficacy of different approaches to protein-ligand and protein-protein binding free-energy calculations, it is worth mentioning that my work has found its application in the context of the recent global COVID-19 health crisis. Given the demonstrated robustness and accuracy of the geometrical route, we were confident that this approach could offer accurate and reliable binding affinities responsible for the COVID-19 syndemic, which placed immense pressure on the scientific community to quickly obtain accurate results to help understand and combat the disease.[171,360]

In our group, we aimed to answer topical questions, such as whether the mutations impact the infectivity or immunity evasion of COVID-19 variants. More specifically, we studied the interactions of RBDs of the following SARS-CoV-2 VOCs: wild type (WT), Alpha, Beta, Delta, and Omicron BA.2 in complex with ACE2.[196,215,361,362] To learn about antibody-based treatment mechanisms for COVID-19, I also investigated some VOCs' RBDs in complex with several antibodies, such as WT bound to a neutralizing nanobody H11-D4[201] and Delta variant[202] in complex with a human neutralizing antibody S2E12.[210] Our simulations successfully reproduced the experimental binding free energies of most complexes, except for the modeled WT and Beta variants, which were hindered

by an incorrect starting structure. The discrepancies between the model and the experimental structures observed in this study underscore the paramount importance of the starting point to correctly reproduce the binding pose and ensure the accuracy of the final results. Inaccurate starting structures can lead to fortuitous cancellations of errors, resulting in misleading or incorrect predictions of binding affinity. Our results showed that both Alpha and Beta variants have increased affinities for ACE2, while Delta and Omicron BA.2 variants have a lower affinity and possess immune-escape properties. Furthermore, the S2E12 antibody was found to have a strong affinity for the Delta variant, making it a potential candidate for COVID-19 therapies. Our study highlights the reliability of the geometrical route for predicting binding affinity and providing atomistic insights into the recognition and association processes of SARS-CoV-2 variants with the host cell.

Furthermore, my research aimed to expand the application of the geometrical route not only to biological complexes in water but also in a non-isotropic environment. For this purpose, I investigated the transmembrane (TM) protein binding mechanism of a membrane protein-protein complex of glycophorin A (GpA), which forms a non-covalent homodimer.[234] Various protocols were considered for this study, encompassing the careful selection of force fields, rigorous standardization to reflect experimental conditions, appropriate sampling algorithms, and the incorporation of anisotropic environmental factors.

I introduced the methodological underpinning of the tailored geometrical route and the unrestrained physical separation PMF calculation strategy, con-

sidering the homodimer symmetry and its reversible association in the cylindrical coordinate frame of the membrane environment. While lipid bilayers impose natural orientational restraints on membrane proteins, it remains crucial to consider these CVs during the investigation of the binding process due to their substantial reduction in computational time required for achieving convergence.

Moreover, the geometrical route provided a comprehensive understanding of the GpA binding mechanism. It was discovered that this mechanism aligns perfectly with the two-stage model,[225,230,231] which serves as a conceptual framework for analyzing the folding of integral membrane proteins. The association between GpA helices is driven by solvent interactions and subsequently stabilized through the formation of helix-helix interactions.

Notably, the versatility of the geometrical route was demonstrated by its applicability in calculating binding free energies within any inhomogeneous and anisotropic environment. The methodology employed in this part of my Ph.D. studies offers novel insights into the mechanism of α-helix dimerization and holds promise for theoretically predicting the binding affinity of more intricate membrane protein assemblies. It was also shown that the geometrical route can be generalized to calculate binding free energies in any inhomogeneous and anisotropic environment. The methodology employed here provides new insight into the mechanism of TM α–helix dimerization, and holds promise for the theoretical prediction of binding affinity of more complex membrane protein assemblies.

In addition to the methodological investigations of the geometrical route, we

have also explored methods for accelerating the performance of such calculations. As mentioned earlier, one drawback of the geometrical route lies in its relatively high computational cost, primarily attributed to the numerous biased molecular simulations that must be conducted before obtaining the final binding free energy estimate. In this work, we aimed at reducing the computational investment of these binding-affinity calculations through improving their efficiency, without sacrificing accuracy, convergence, and sampling uniformity.

To achieve this goal, we performed physical-separation PMF calculations on the Abl kinase-SH3:p41 and the MDM2-p53:NVP-CGM097 complexes following the geometrical route, and employed MTS option[352,354,355] for CV and biasing-force computations, with and without HMR.[288,347] The MTS strategy is based on the fact that biasing forces have a smoother dependence on atomic positions and vary slower than physical force-field derived forces, allowing for a larger time step during integration. MTS enables CVs and time-dependent biases to be evaluated less frequently, resulting in a net increase in computational efficiency.[67,354] Additionally, the HMR slows down the highest-frequency motions of the molecular objects at play to increase the integration time step of the MD simulation.[347]

Our results confirm that application of HMR alone accelerates calculations by nearly a factor of two compared to the reference physical-separation PMF calculation, without compromising accuracy or trajectory stability,[288] and reproduces the standard binding affinity appropriately. The combination of MTS and HMR are, therefore, promising strategies for accelerating binding-affinity calculations without sacrificing accuracy and convergence, which is of great importance for the

accurate prediction of binding affinities in drug design and discovery. However, even armed with an accurate force field and reliable structural data, it is important to note that the specificity of each protein-ligand or protein-protein complex cannot be overlooked, and it is recommended that the extended-Lagrangian parameters be optimized prior to the application of the accelerated computational schemes for standard binding free-energy calculations following the geometrical route.

Looking forward, the standard binding free-energy calculations for protein-ligand and protein-protein complexes remain an active area of research with significant potential for future developments. One potential avenue for exploration is the incorporation of machine learning and artificial intelligence techniques to enhance the accuracy and speed of the calculations. As these technologies become increasingly integrated into the pharmaceutical industry, it is essential to ensure that they are used responsibly and with proper oversight.

Additionally, the recent SARS-CoV-2 syndemic has highlighted the importance of standard binding free energy calculations for drug discovery and the development of effective therapeutics. The emergence of new Covid-19 variants as well as other causative agents of diseases underscores the need for continued research on the mechanisms of viral infection and the development of new treatments. By leveraging the predictive power of computational approaches such as geometrical route, we can accelerate the discovery of new drugs and therapies to combat different diseases.

Conclusion and Perspectives (Version française)

Au cours de la première année de mes études doctorales, j'ai cherché à comprendre les nuances des altérations de l'énergie libre de liaison dans les complexes protéine-ligand. Mon premier objectif était d'entreprendre des calculs de l'énergie libre standard de liaison et, pour ce faire, j'ai participé à l'élaboration d'un protocole pour réaliser ces calculs difficiles à l'aide de l'outil BFEE2. Je me suis plus particulièrement intéressée à trois complexes et les resultats étaient pour le moins prometteurs.

Au cœur de ce protocole se trouve l'approche théorique proposée par Woo et Roux,[29] basée sur l'application de restreintes au cours de la dynamique moléculaire, MD, qui a évolué dans le formalisme de la "*geometrical*" et "*alchemical*" routes.[12] La caractéristique principale de cette approche est qu'elle réduit les nombreux changements de conformation qui se produisent pendant la liaison à quelques paramètres structurels, les variables collectives, CVs. Ils correspondent aux degrés de liberté lents associés aux mouvements relatifs des partenaires de liaison au cours de la simulation MD. Les restraintes géométriques agissant sur ces CVs restreignent l'espace configurationnel, ce qui permet d'accélérer la convergence du calcul de l'énergie libre et de capturer avec précision les degrés de liberté lents de l'association réversible.

La connaissance de la structuré liée d'un complexe, obtenue par des données expérimentales, et l'utilisation de BFEE2, couplée à un champs de force fiable, permet l'obtention d'énergies libres standard de liaison avec la précision chimique (+/- 1 kcal/mol). En outre, ce protocole réduit considérablement le besoin d'intervention humaine, car il aide efficacement l'utilisateur à préparer tous les fichiers initiaux nécessaires et à effectuer le post-traitement pour obtenir l'estimation finale de l'affinité de liaison.

Pour approfondir mes recherches sur les calculs d'énergie libre de liaison basés sur la restrainte pour les complexes protéine-ligand, j'ai étendu mes investigations à des cas plus délicats tels que les interactions protéine-protéine. En raison des larges perturbations des complexes protéine-protéine, l'utilisation de la voie alchimique n'est plus possible. Au lieu de cela, j'ai employé la *geometrical route*, qui consiste en séparation des deux moitiés moléculaires en présence de restraintes d'orientation, de position et de conformation sur les deux protéines introduites progressivement pour contrôler les changements d'entropie configurationnelle au cours du processus de dissociation. La *geometrical route* permet aux calculs de converger dans des temps de simulation raisonnables.

Comme étude de cas, j'ai examiné un dimère d'insuline de porc, dans lequel la dimérisation est controlée par les interactions hydrophobes à l'interface des deux monomères, et l'affinité de liaison expérimentale a déjà été établie dans la littérature.[143,144]

Dans le cadre de mes recherches, j'ai cherché à approfondir ma compréhension

de la précision et de la convergence des calculs de l'énergie libre de liaison pour les complexes protéine-ligand et protéine-protéine. Pour ce faire, j'ai mené des recherches méthodologiques sur les calculs de l'énergie libre standard de liaison et j'ai comparé le cadre théorique rigoureux de la *geometrical route* avec son raccourci computationnel, dans lequel tous les autres degrés de liberté ne sont pas restreints pendant la séparation physique des partenaires.

Nos résultats montrent que la *geometrical route* offre un contrôle total de chaque degré de liberté dans le processus d'association réversible par le biais d'un ensemble universel de CV, et que la contribution de ces degrés de liberté à l'énergie libre de liaison peut être déterminée avec un niveau de précision approprié à partir des forces moyennes potentielles, PMFs, correspondants. Bien que la *geometrical route* entraîne un coût de calcul relativement élevé, son cadre théorique rigoureux offre des résultats fiables et reproductibles, ce qui en fait une approche préférable pour des estimations précises de l'affinité de liaison. Le protocole de la *geometrical route* peut être appliqué à n'importe quel complexe où la liaison se produit à la surface d'une protéine. Bien que la mise en place des différentes étapes du protocole, en particulier la définition de CV pertinentes, puisse sembler difficile et longue, des outils tels que les logiciels CHARMM-GUI[207] et BFEE2,[75,357] peuvent automatiser la définition des CV et générer tous les fichiers nécessaires pour la *geometrical route*, ce qui rend le protocole plus facile à utiliser.

Contrairement à la *geometrical route* robuste, son approche raccourcie manque de justification théorique systématique, ce qui conduit à des résultats qui s'écartent des valeurs expérimentales. Notre étude souligne également que les simulations

non-restreintes MD d'une longueur typique d'un μs sont insuffisantes pour obtenir un échantillonnage adéquat de l'espace configurationnel disponible dans les complexes protéine-ligand et protéine-protéine, ce qui entraîne des estimations inexactes de l'affinité de liaison.

Il convient de noter que si la *geometrical route* garantit des propriétés de convergence optimales pour les calculs d'énergie libre de liaison et des résultats entièrement reproductibles, les approximations du champ de force et la dépendance au motif de liaison de départ sont des sources potentielles de divergences avec l'expérience, ce qui en fait le point faible de tous les calculs d'énergie libre de liaison.

L'un des enseignements importants de cette étude est que, malgré son attrait par sa simplicité, l'approche par raccourci peut être problématique, alors que le coût de calcul de la *geometrical route* suivie rigoureusement est nettement moindre. Étant donné l'absence de justification théorique des approches simplifiées, il n'y a pas de raison impérieuse de ne pas utiliser la *geometrical route* pour obtenir des estimations précises de l'affinité de liaison.

Bien que mon doctorat ait porté sur l'évaluation de l'efficacité de différentes approches des calculs de l'énergie libre de la liaison protéine-ligand et protéine-protéine, il convient de noter que nos résultats ont des implications plus larges pour la récente crise de santé publique COVID-19. La syndromie a exercé une pression énorme sur la communauté scientifique pour qu'elle obtienne rapidement des résultats précis afin d'aider à comprendre et à combattre la maladie.[171,360]

Dans notre groupe, nous avons cherché à répondre à des questions d'actualité, telles que l'impact des mutations sur l'infectivité ou l'évasion immunitaire des variantes COVID–19. Plus précisément, nous avons étudié les interactions des RBD des variants préoccupants, VOC, SARS-CoV-2 suivants : type sauvage (WT), Alpha, Beta, Delta, et Omicron BA.2 en complexe avec ACE2.[196,215,361,362] Pour en savoir plus sur les mécanismes de traitement basés sur les anticorps pour COVID-19, j'ai également étudié les RBD de certains VOC en complexe avec plusieurs anticorps, tels que WT lié à un nanocorps neutralisant H11-D4[201] et la variante Delta[202] en complexe avec un anticorps neutralisant humain S2E12.[210] En nous appuyant sur nos travaux antérieurs qui ont démontré la robustesse et la précision de la *geometrical route* pour les complexes protéine-ligand et protéine-protéine, nous étions convaincus que la *geometrical route* pouvait fournir des résultats précis et fiables pour les interactions plus compliquées dans les complexes COVID-19.

Nos simulations ont reproduit avec succès les énergies libres de liaison expérimentales de la plupart des complexes, à l'exception des variantes WT et Beta modélisés, entravés par une structure de départ incorrecte. Les divergences entre le modèle et les structures expérimentales observées dans cette étude soulignent l'importance primordiale du point de départ pour reproduire correctement la pose de liaison et garantir l'exactitude des résultats finaux. Des structures de départ inexactes peuvent conduire à des annulations fortuites d'erreurs, ce qui entraîne des prédictions trompeuses ou incorrectes de l'affinité de liaison.

Nos résultats ont montré que les variantes Alpha et Beta ont des affinités accrues pour l'ACE2, tandis que les variantes Delta et Omicron BA.2 ont une

affinité plus faible et possèdent des propriétés accrues d'évasion immunitaire. En outre, l'anticorps S2E12 présente une forte affinité pour la variante Delta, ce qui en fait un candidat potentiel pour les thérapies COVID-19. Notre étude met en évidence la fiabilité de la *geometrical route* pour prédire l'affinité de liaison et fournir des informations atomistiques sur les processus de reconnaissance et d'association des variantes du SARS-CoV-2 avec la cellule hôte.

De plus, ma recherche visait à étendre l'application de la méthode géométrique non seulement aux complexes biologiques en solution aqueuse, mais également dans un environnement non isotrope.

Dans ce but, j'ai étudié le mécanisme de liaison des protéines transmembranaires (TM) d'un complexe protéine-protéine de la glycophorine A (GpA), formant un homodimère non covalent.[234] Divers protocoles ont été pris en compte pour cette étude, incluant une sélection minutieuse des champs de force, une standardisation rigoureuse reflétant les conditions expérimentales, l'utilisation d'algorithmes d'échantillonnage appropriés, ainsi que l'intégration de facteurs environnementaux anisotropes.

J'ai introduit les fondements méthodologiques de la méthode géométrique adaptée et de la stratégie de calcul PMF (potentiel de force moyenne) de séparation physique non contrainte, en prenant en compte la symétrie de l'homodimère et son association réversible dans le cadre des coordonnées cylindriques de l'environnement membranaire. Bien que les bicouches lipidiques imposent des contraintes orientantes naturelles aux protéines membranaires, il est essentiel de considérer ces

variables collectives (CVs) lors de l'étude du processus de liaison en raison de leur réduction substantielle du temps de calcul nécessaire à la convergence.

De plus, la méthode géométrique a permis de comprendre de manière approfondie le mécanisme de liaison de GpA. Il a été découvert que ce mécanisme s'aligne parfaitement sur le modèle à deux étapes,[225,230,231] qui sert de cadre conceptuel pour analyser le repliement des protéines membranaires intégrales. L'association entre les hélices de GpA est induite par des interactions avec le solvant et stabilisée ultérieurement par la formation d'interactions hélice-hélice. Il est important de souligner la polyvalence de la méthode géométrique démontrée par son applicabilité dans le calcul des énergies libres de liaison dans tout environnement inhomogène et anisotrope. La méthodologie utilisée dans le cadre de cette partie de mes études doctorales offre de nouvelles perspectives sur le mécanisme de dimérisation de l'α-hélice et ouvre la voie à la prédiction théorique de l'affinité de liaison de montages protéiques membranaires plus complexes. De plus, il a été démontré que la méthode géométrique peut être généralisée pour calculer les énergies libres de liaison dans tout environnement inhomogène et anisotrope. La méthodologie employée ici apporte de nouvelles perspectives sur le mécanisme de dimérisation de l'α-hélice TM et offre des possibilités prometteuses pour la prédiction théorique de l'affinité de liaison de montages protéiques membranaires plus complexes.

En complément des investigations méthodologiques de la *geometrical route*, nous avons également exploré des méthodes visant à accélérer les calculs de ce type. Comme mentionné précédemment, l'un des inconvénients de la *geometrical*

route réside dans son coût computationnel relativement élevé, principalement dû aux nombreuses simulations moléculaires biaisées qui doivent être effectuées avant d'obtenir l'estimation finale de l'énergie libre de liaison. Dans ce travail, notre objectif était de réduire l'investissement computationnel de ces calculs d'affinité de liaison en améliorant leur efficacité, sans compromettre l'exactitude, la convergence et l'uniformité de l'échantillonnage.

Pour atteindre cet objectif, nous avons effectué des calculs PMF de séparation physique sur les complexes Abl kinase-SH3:p41 et MDM2-p53:NVP-CGM097 en suivant la *geometrical route*, et nous avons utilisé l'option MTS pour les calculs de CV et des forces appliqués pour biaiser le système , avec et sans HMR.[288,347] La stratégie MTS est basée sur le fait que les biais ont une dépendance plus faibles aux positions atomiques et varient plus lentement que les forces physiques dérivées du champ de force, ce qui permet un pas de temps plus important lors de l'intégration. La stratégie MTS permet d'évaluer moins fréquemment les CVs et les biais dépendant du temps, ce qui se traduit par une augmentation nette de l'efficacité des calculs. En outre, le HMR ralentit les mouvements de plus haute fréquence des objets moléculaires en jeu afin d'augmenter le pas de temps d'intégration de la simulation MD.[347] Nos résultats confirment que l'application de la HMR seule accélère les calculs de près d'un facteur deux par rapport au calcul PMF de référence lors de la séparation physique, sans compromettre la précision ou la stabilité de la trajectoire,[288] et reproduit l'affinité standard de liaison de manière appropriée. La combinaison de MTS et de HMR est donc une stratégie prometteuse pour accélérer les calculs d'affinité de liaison sans sacrifier la précision

et la convergence, ce qui revêt une importance cruciale pour les prédictions dans le cadre de la conception et de la découverte de médicaments. Cependant, même avec un champ de force précis et des données structurelles fiables, il est important de noter que la spécificité de chaque complexe protéine-ligand ou protéine-protéine ne peut pas être négligée, et il est recommandé que les paramètres Lagrangien étendu soient optimisés avant l'application des schémas de calcul accélérés pour les calculs standard de l'énergie libre de liaison suivant la *geometrical route*.

Les calculs de l'énergie libre de liaison absolue pour les complexes protéine-ligand et protéine-protéine restent un domaine de recherche actif avec un potentiel important pour les développements futurs. Une voie potentielle d'exploration est l'incorporation de techniques d'apprentissage automatique et d'intelligence artificielle pour améliorer la précision et la rapidité des calculs. Ces technologies étant de plus en plus intégrées dans l'industrie pharmaceutique, il est essentiel de veiller à ce qu'elles soient utilisées de manière responsable et avec un contrôle adéquat.

En outre, la récente épidémie de SARS-CoV-2 a mis en évidence l'importance des calculs standard de l'énergie libre de liaison pour la découverte de médicaments et le développement de thérapies efficaces. L'émergence de nouvelles variantes de Covid-19 ainsi que d'autres agents responsables de maladies souligne la nécessité de poursuivre la recherche sur les mécanismes de l'infection virale et de mettre au point de nouveaux traitements. En exploitant le pouvoir prédictif d'approches informatiques telles que la *geometrical route*, nous pouvons accélérer la découverte de nouveaux médicaments et de nouvelles thérapies pour lutter contre différentes maladies.

Abstract (English version)

During my Ph.D. thesis, I investigated standard binding free-energy calculations in protein-ligand complexes using the restraint MD-based approaches of alchemical and geometrical routes. By studying three different protein-ligand complexes, I contributed to creating a protocol leveraging the BFEE2 software for automating these calculations. Expanding on this work, I applied the strategy to intricate protein-protein interactions with more complex recognition and association phenomena. As a case study, I first examined a pig insulin dimer driven by hydrophobic interactions at the monomers' interface, comparing the calculated estimate with experimental binding affinity. To deepen my understanding of binding free-energy calculations for protein-ligand and protein-protein complexes, I further conducted methodological research, comparing the rigorous geometric route with its shortcut, where other degrees of freedom remained unrestrained during the physical separation of the partner, thus, emphasizing their need. After demonstrating the accuracy of the geometric route, its applicability was extended to predicting and evaluating binding affinities of SARS-CoV-2 variants in complex with a human receptor and antibodies, allowing identification of the main mechanism responsible for their high infectiousness. Specific considerations linked to the protein's structural specificity and its non-homogeneous environment were also investigated with the reversible separation of the GpA dimer in the lipid bilayer, thus broadening the scope of the geometrical route. Additionally, I explored acceleration strategies using the multiple time-stepping (MTS) option available in

the Colvars module with and without the hydrogen-mass repartitioning (HMR) trick. By tuning parameters, I achieved nearly threefold computation acceleration without compromising the accuracy of the binding free energy calculations. Overall, my in-depth methodological survey opens up opportunities for faster and more accessible standard binding affinity evaluation for more intricate biocomplexes.

Résumé (Version française)

Durant ma thèse, j'ai étudié les calculs d'énergie libre de liaison absolue dans des complexes protéine-ligand, en utilisant des approches basées sur la dynamique moléculaire avec des restraintes, notamment *alchemical* et *geometrical routes*. En étudiant trois complexes protéine-ligand différents, j'ai contribué à la création d'un protocol utilisant le logiciel BFEE2 crée pour l'automatisationlogiciel BFEE2 pour l'automatisation de ces calculs. Pour approfondir ce travail , j'ai appliqué cette stratégie à des interactions protéine-protéine avec des phénomènes de reconnaissance et d'association plus recherchés. Comme cas pratique, j'ai examiné un dimère d'insuline porcine, où la dimérisation était induite par des interactions hydrophobes à l'interface des monomères. J'ai par la suite comparé l'estimation d'énergie libre calculée à celle de l'experience. Pour élargir ma compréhension des calculs d'énergie libre de liaison aux complexes protéine-ligand et protéine-protéine, j'ai réalisé une recherche méthodologique. J'ai comparé la robustesse de la *geometrical route* à son raccourci, où les autres degrés de liberté restaient non-restraints pendant la séparation physique des partenaires renforçant leur nécessité. Après avoir démontré l'exactitude de la *geometrical route*, son application a été étendue à la prédiction et à l'évaluation des affinités de liaison des variants du SARS-CoV-2 en association avec un récepteur humain et des anticorps, nous permettant de comprendre le mécanisme derrière la contagiosité élevée de ces variants. Les considerations spécifiques liées à la structure de la protéine et à son environnement non-homogène, ont également été explorés avec la sep-

aration reversible du dimer de GpA dans une bicouche lipidique, enlargissant le champ d'application de la *geometrical route*. De plus, j'ai exploré des stratégies d'accélération des calculs en utilisant l'option *" multiple time-stepping"* (MTS) disponible dans le module Colvars, avec et sans la technique de *"hydrogen-mass repartitioning"* (HMR). En ajustant les paramètres, j'ai obtenu une accélération des calculs presque triplée, sans compromettre l'exactitude des calculs d'énergie libre de liaison. Dans l'ensemble, mon étude méthodologique poussée ouvre des nouvelles possibilités pour des calculs d'énergie libre standard de liaison plus rapides et accessibles destinés aux biocomplexes plus délicats.

Bibliography

[1] Pebay-Peyroula, E., Biophysical Analysis of Membrane Proteins: Investigating Structure and Function, Wiley-Blackwell: Germany, 2007.

[2] Chipot, C.; Pohorille, A., Free energy calculations. Theory and applications in chemistry and biology, Springer Verlag: Berlin, 2007.

[3] Marsh, J. A.; Teichmann, S. A., Structure, Dynamics, Assembly, and Evolution of Protein Complexes, *Annu. Rev. Biochem.* **2015**, *84*, 551–575.

[4] Nooren, Ine M. A.; Thornton, J. M., Diversity of Protein-Protein Interactions, *EMBO J.* **2003**, *22*, 3486–3492.

[5] Wereszczynski, J.; McCammon, J. A., Statistical Mechanics and Molecular Dynamics in Evaluating Thermodynamic Properties of Biomolecular Recognition, *Q. Rev. Biophys.* **2012**, *45*, 1–25.

[6] Lee, T.-S. et al., Alchemical Binding Free Energy Calculations in AMBER20: Advances and Best Practices for Drug Discovery, *J. Chem. Inf. Model.* **2020**, *60*, 5595–5623.

[7] Mobley, D. L.; Gilson, M. K., Predicting Binding Free Energies: Frontiers and Benchmarks, *Annu. Rev. Biophys.* **2017**, *46*, 531–558.

[8] Chipot, C., Free Energy Methods for the Description of Molecular Processes, *Ann. Rev. Biophys.* **2023**, *52*, 113–138.

[9] Woods, C. J.; Malaisree, M.; Hannongbua, S.; Mulholland, A. J., A Water-Swap Reaction Coordinate for the Calculation of Absolute Protein-Ligand Binding Free Energies., *J. Chem. Phys.* **2011**, *134*, 054114 – 054114–13.

[10] Cole, D. J.; Tirado-Rives, J.; Jorgensen, W. L., Molecular Dynamics and Monte Carlo Simulations for Protein-Ligand Binding and Inhibitor Design., *Biochim. Biophys. Acta* **2015**, *1850*, 966–971.

[11] De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A., Role of Molecular Dynamics and Related Methods in Drug Discovery, *J. Med. Chem.* **2016**, *59*, 4035–4061.

[12] Gumbart, J. C.; Roux, B.; Chipot, C., Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy?, *J. Chem. Theory Comput.* **2013**, *9*, 794–802.

[13] Chipot, C., Frontiers in Free-Energy Calculations of Biological Systems, *WIREs Comput. Mol. Sci.* **2014**, *4*, 71–89.

[14] Landau, L. D, Statistical physics, The Clarendon Press: Oxford, 1938.

[15] Kirkwood, J. G., Statistical Mechanics of Fluid Mixtures, *J. Chem. Phys.* **1935**, *3*, 300–313.

[16] Kirkwood, J. G., Theory of Liquids, Gordon and Breach, Science Publishers, 1968.

[17] Zwanzig, R. W., High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases, *J. Chem. Phys.* **1954**, *22*, 1420–1426.

[18] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of Simple Potential Functions for Simulating Liquid Water, *Chem. Phys.* **1983**, *79*, 926–935.

[19] Beveridge, W. L.; Jorgensen, D. L., Computer simulation of chemical and biomolecular systems, New York Academy of Sciences, 1986.

[20] Jorgensen, W. L., Free energy calculations: a breakthrough for modeling organic chemistry in solution, *Acc. Chem. Res.* **1989**, *22*, 184–189.

[21] Kollman, P. A., Free energy calculations: Applications to chemical and biochemical phenomena, *Chem. Rev.* **1993**, *93*, 2395–2417.

[22] Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W. et al., Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models, *Acc. Chem. Res.* **2000**, *33*, 889–897.

[23] Chipot, C.; Kollman, P. A.; Pearlman, D. A., Alternative Approaches to Potential of Mean Force Calculations: Free Energy Perturbation versus Thermodynamic Integration. Case Study of Some Representative Nonpolar Interactions, *J. Comput. Chem.* **1996**, *17*, 1112–1131.

[24] Darve, E.; Pohorille, A., Calculating free energies using average force, *J. Chem. Phys.* **2001**, *115*, 9169–9183.

[25] Chipot, C.; Pearlman, D. A., Free Energy Calculations. The Long and Winding Gilded Road, *Mol. Simul.* **2002**, *28*, 1–12.

[26] Hénin, J.; Chipot, C., Overcoming free energy barriers using unconstrained molecular dynamics simulations, *J. Chem. Phys.* **2004**, *121*, 2904–2914.

[27] Simonson, T.; Archontis, G.; Karplus, M., Free Energy Simulations Come of Age: Protein-Ligand Recognition, *Acc. Chem. Res.* **2002**, *35*, 430–437.

[28] Berne, B. J.; Straub, J. E., Novel methods of sampling phase space in the simulation of biological systems, *Curr. Opin. Struct. Biol.* **1997**, *7*, 181–189.

[29] Woo, H.-J.; Roux, B., Calculation of Absolute Protein-Ligand Binding Free Energy from Computer Simulations., *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6825–6830.

[30] Freire, E.; Mayorga, O. L.; Straume, M., Isothermal titration calorimetry, *Anal. Chem.* **1990**, *62*, 950A–959A.

[31] Grolier, J.-P. E.; del Río, J. M., Isothermal titration calorimetry: A thermodynamic interpretation of measurements, *J. Chem. Thermodyn.* **2012**, *55*, 193–202.

[32] Singh, P., SPR Biosensors: Historical Perspectives and Current Challenges, *Sens. Actuators B: Chem.* **2016**, *229*, 110–130.

[33] Rognan, D., The impact of in silico screening in the discovery of novel and safer drug candidates, *Pharmacol. Ther.* **2017**, *175*, 47–66.

[34] Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv. Drug Deliv. Rev.* **2001**, *46*, 3–26.

[35] Pohorille, A.; Jarzynski, C.; Chipot, C., Good Practices in Free-Energy Calculations., *J. Phys. Chem. B* **2010**, *114*, 10235–53.

[36] Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W. et al., Scalable molecular dynamics on CPU and GPU architectures with NAMD., *J. Chem. Phys.* **2020**, *153*, 044130.

[37] Alder, J.; Wainwright, T.E., Phase Transition for a Hard Sphere System, *J. Chem. Phys.* **1957**, *27*, 1208.

[38] McCammon, J.; Gelin, B.; Karplus, M., Dynamics of Folded Proteins, *Nature* **1977**, *267*, 585–590.

[39] Chen, H.; Maia, J. D. C.; Radak, B. K.; Hardy, D. J.; Cai, W.; Chipot, C.; Tajkhorshid, E., Boosting Free-Energy Perturbation Calculations with GPU-Accelerated NAMD., *J. Chem. Inf. Model.* **2020**, *60*, 5301–5307.

[40] Pan, A. C.; Jacobson, D.; Yatsenko, K.; Sritharan, D.; Weinreich, T. M.; Shaw, D. E., Atomic-Level Characterization of Protein-Protein Association, *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 4244–4249.

[41] Buch, I.; Giorgino, T.; De Fabritiis, G., Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations, *Proc. Natl. Acad. Sci.* **2011**, *108*, 10184–10189.

[42] Noé, F.; Fischer, S., Transition Networks for Modeling the Kinetics of Conformational Change in Macromolecules, *Curr. Opin. Struct. Biol.* **2008**, *18*, 154–162.

[43] Pande, V. S.; Beauchamp, K.; Bowman, G. R., Everything You Wanted to Know about Markov State Models but Were Afraid to Ask, *Methods* **2010**, *52*, 99–105.

[44] Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F., Markov Models of Molecular Kinetics: Generation and Validation, *J. Chem. Phys.* **2011**, *134*, 174105.

[45] Husic, B. E.; Pande, V. S., Markov State Models: From an Art to a Science, *J. Am. Chem. Soc.* **2018**, *140*, 2386–2396.

[46] Gumbart, J. C.; Roux, B.; Chipot, C., Efficient Determination of Protein-Protein Standard Binding Free Energies from First Principles, *J. Chem. Theory Comput.* **2013**, *9*, 3789–3798.

[47] Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A., Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices, *J. Am. Chem. Soc.* **1998**, *120*, 9401–9409.

[48] Massova, I.; Kollman, P. A., Combined Molecular Mechanical and Continuum Solvent Approach (MM-PBSA/GBSA) to Predict Ligand Binding, *Perspect. drug discov. des.* **2000**, *18*, 113–135.

[49] Homeyer, N.; Gohlke, H., Free Energy Calculations by the Molecular Mechanics Poisson-Boltzmann Surface Area Method., *Mol. Inform.* **2012**, *31*, 114–22.

[50] Gohlke H., Case D.A., Converging Free Energy Estimates: MM-PB(GB)SA Studies on the Protein-Protein Complex Ras-Raf., *J. Comput. Chem.* **2004**, *25(2)*, 238–50.

[51] Genheden, S.; Ryde, U., The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities, *Expert Opin. Drug Discov.* **2015**, *10*, 449–461.

[52] Swanson, J. M. J.; Henchman, R. H.; McCammon, J. A., Revisiting Free Energy Calculations: a Theoretical Connection to MM/PBSA and Direct Calculation of the Association Free Energy., *Biophys. J.* **2004**, *86*, 67–74.

[53] Pearlman, D. A., Evaluating the Molecular Mechanics Poisson-Boltzmann Surface Area Free Energy Method Using a Congeneric Series of Ligands to p38 MAP Kinase, *J. Med. Chem.* **2005**, *48*, 7796–7807.

[54] Izrailev, S.; Stepaniants, S.; Isralewitz, B.; Kosztin, D.; Lu, H.; Molnar, F.; Wriggers, W.; Schulten, K., Computational Molecular Dynamics: Challenges, Methods, Ideas: Steered Molecular Dynamics, Springer: Berlin, Heidelberg, 1999.

[55] Isralewitz, B.; Gao, M.; Schulten, K., Steered Molecular Dynamics and Mechanical Functions of Proteins, *Curr. Opin. Struct. Biol.* **2001**, *11*, 224–230.

[56] Gu, J.; Li, H.; Wang, X., A Self-Adaptive Steered Molecular Dynamics Method Based on Minimization of Stretching Force Reveals the Binding Affinity of Protein–Ligand Complexes, *Molecules* **2015**, *20*, 19236–19251.

[57] Potterton, A.; Husseini, F. S.; Southey, M. W. Y.; Bodkin, M. J.; Heifetz, A.; Coveney, P. V.; Townsend-Nicholson, A., Ensemble-Based Steered Molecular Dynamics Predicts Relative Residence Time of A2A Receptor Binders, *J. Chem. Theory Comput.* **2019**, *15*, 3316–3330.

[58] Gong, Q.; Zhang, H.; Zhang, H.; Chen, C., Calculating the Absolute Binding Free Energy of the Insulin Dimer in an Explicit Solvent, *RSC Adv.* **2020**, *10*, 790–800.

[59] Jarzynski, C., Nonequilibrium Equality for Free Energy Differences., *Phys. Rev. Lett.* **1997**, *78*, 2690–2693.

[60] Park, S.; Khalili-Araghi, F.; Tajkhorshid, E.; Schulten, K., Free Energy Calculation from Steered Molecular Dynamics Simulations Using Jarzynski's Equality, *J. Chem. Phys.* **2003**, *119*, 3559–3566.

[61] Yang, L.-J.; Zou, J.; Xie, H.-Z.; Li, L.-L.; Wei, Y.-Q.; Yang, S.-Y., Steered Molecular Dynamics Simulations Reveal the Likelier Dissociation Pathway of Imatinib from its Targeting Kinases c-Kit and Abl., *PloS one* **2009**, *4*, e8470–e8470–8.

[62] Baştuğ, T.; Chen, P.-C.; Patra, S. M.; Kuyucak, S., Potential of Mean Force Calculations of Ligand Binding to Ion Channels from Jarzynski's Equality and Umbrella Sampling, *J. Chem. Phys.* **2008**, *128*, 155104–155104–9.

[63] Siebenmorgen, T.; Zacharias, M., Evaluation of Predicted Protein-Protein Complexes by Binding Free Energy Simulations, *J. Chem. Theory Comput.* **2019**, *15*, 2071–2086.

[64] Hermans, J.; Shankar, S., The free energy of xenon binding to myoglobin from molecular dynamics simulation, *Isr. J. Chem.* **1986**, *27*, 225–227.

[65] Roux, B.; Nina, M.; Pomès, R.; Smith, J. C., Thermodynamic stability of water molecules in the bacteriorhodopsin proton channel: A molecular dynamics free energy perturbation study, *Biophys. J.* **1996**, *71*, 670–681.

[66] Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M., Absolute Binding Free Energies: a Quantitative Approach for Their Calculation, *J. Phys. Chem. B* **2003**, *107*, 9535–9551.

[67] Fiorin, G.; Klein, M. L.; Hénin, J., Using collective variables to drive molecular dynamics simulations, *Mol. Phys.* **2013**, *111*, 3345–3362.

[68] Limongelli, V.; Bonomi, M.; Parrinello, M., Funnel Metadynamics as Accurate Binding Free-Energy Method, *Proc. Natl. Acad. Sci.* **2013**, *110*, 6358–6363.

[69] Heinzelmann, G.; Henriksen, N. M.; Gilson, M. K., Attach-Pull-Release Calculations of Ligand Binding and Conformational Changes on the First BRD4 Bromodomain, *J. Chem. Theory Comput.* **2017**, *13*, 3260–3275.

[70] Deng, Y.; Roux, B., Calculation of Standard Binding Free Energies: Aromatic Molecules in the T4 Lysozyme L99A Mutant, *J. Chem. Theory Comput.* **2006**, *2*, 1255–1273.

[71] Wang, J.; Deng, Y.; Roux, B., Absolute Binding Free Energy Calculations Using Molecular Dynamics Simulations with Restraining Potentials., *Biophys. J.* **2006**, *91*, 2798–814.

[72] Mobley, D. L.; Chodera, J. D.; Dill, Ken A., The Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change, *J. Chem. Theory Comput.* **2007**, *3*, 1231–1235.

[73] Shoup, D.; Szabo, A., Role of Diffusion in Ligand Binding to Macromolecules and Cell-Bound Receptors, *Biophys. J.* **1982**, *40*, 33–39.

[74] Blazhynska, M.; Goulard Coderc de Lacam, E.; Chen, H.; Roux, B.; Chipot, C., Hazardous Shortcuts in Standard Binding Free Energy Calculations, *J. Phys. Chem. Lett.* **2022**, *13*, 6250–6258.

[75] Fu, H.; Chen, H.; Blazhynska, M.; Goulard Coderc de Lacam, E.; Szczepaniak, F.; Pavlova, A.; Shao, X.; Gumbart, J. C.; Dehez, F.; Roux, B.; Cai, W.; Chipot, C., Accurate Determination of Protein:Ligand Standard Binding Free Energies from Molecular Dynamics Simulations, *Nat. Protoc.* **2022**, *17*, 1114–1141.

[76] Liu, P.; Dehez, F.; Cai, W.; Chipot, C., A Toolkit for the Analysis of Free-Energy Perturbation Calculations., *J. Chem. Theory Comput.* **2012**, *8*, 2606–2616.

[77] Hermans, J.; Wang, L., Inclusion of Loss of Translational and Rotational Freedom in Theoretical Estimates of Free Energies of Binding. Application to a Complex of Benzene and Mutant T4 Lysozyme, *J. Am. Chem. Soc.* **1997**, *119*, 2707–2714.

[78] Deng, Y. Q.; Roux, B., Computation of Binding Free Energy with Molecular Dynamics and Grand Canonical Monte Carlo Simulations, *J. Chem. Phys.* **2008**, *128*, 115103.

[79] Straatsma, T. P.; Berendsen, H. J. C., Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations, *J. Chem. Phys.* **1988**, *89*, 5876–5886.

[80] Born, M., Volumen und Hydratationswärme der Ionen, *Z. Phys.* **1920**, *1*, 45.

[81] Bennett, C. H., Efficient estimation of free energy differences from Monte Carlo data, *J. Comput. Phys.* **1976**, *22*, 245–268.

[82] Hahn, A. M.; Then, H., Characteristic of Bennett's acceptance ratio method, *Phys. Rev. E* **2009**, *80*, 031111.

[83] Lu, N.; Kofke, D. A.; Woolf, T. B., Staging Is More Important than Perturbation Method for Computation of Enthalpy and Entropy Changes in Complex Systems, *J. Phys. Chem. B* **2003**, *107*, 5598–5611.

[84] Lu, N.; Kofke, D. A.; Woolf, T. B., Improving the efficiency and reliability of free energy perturbation calculations using overlap sampling methods, *J. Comput. Chem.* **2004**, *25*, 28–40.

[85] Humphrey, W.; Dalke, A.; Schulten, K., VMD: Visual molecular dynamics, *J. Mol. Graph.* **1996**, *14*, 33–38.

[86] Zacharias, M.; Straatsma, T. P.; McCammon, J. A., Separation-shifted scaling, a new scaling method for Lennard-Jones interactions in thermodynamic integration, *J. Chem. Phys.* **1994**, *100*, 9025–9031.

[87] Beutler, T. C.; Mark, Alan E.; Vanschaik, Robert C.; Gerber, Paul R.; Vangunsteren, Wilfred F., Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations, *Chem. Phys. Lett.* **1994**, *222*, 529–539.

[88] Pitera, J. W.; van Gunsteren, W. F., A Comparison of Non-Bonded Scaling Approaches for Free Energy Calculations, *Mol. Simulat.* **2002**, *28*, 45–65.

[89] Hénin, J.; Lelièvre, T.; Shirts, M. R.; Valsson, O.; Delemotte, L., Enhanced Sampling Methods for Molecular Dynamics Simulations [Article v1.0], *J. Comp. Mol. Sci.* **2021**, *4*, 1–60.

[90] Chen, H.; Chipot, C., Enhancing Sampling with Free-Energy Calculations, *Curr. Opin. Struct. Biol.* **2022**, *77*, 102497.

[91] Torrie, G. M.; Valleau, J. P., Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling, *J. Comput. Phys.* **1977**, *23*, 187–199.

[92] Chandler, D., Introduction to Modern Statistical Mechanics, Oxford University Press: New York, 1987.

[93] Bartels, C.; Karplus, M., Multidimensional adaptive umbrella sampling: applications to main-chain and side-chain peptide conformations, *J. Comput. Chem.* **1997**, *18*, 1450–1462.

[94] Marsili, S.; Barducci, A.; Chelli, R.; Procacci, P.; Schettino, V., Self-healing umbrella sampling: a non-equilibrium approach for quantitative free energy calculations, *J. Phys. Chem. B* **2006**, *110*, 14011–14013.

[95] Wojtas-Niziurski, W.; Meng, Y.; Roux, B.; Bernèche, S., Self-learning adaptive umbrella sampling method for the determination of free energy landscapes in multiple dimensions, *J. Chem. Theory Comput.* **2013**, *9*, 1885–1895.

[96] Boczko, E. M.; Brooks 3rd, Charles L., First-Principles Calculation of the Folding Free Energy of a Three-Helix Bundle Protein, *Science* **1995**, *269*, 393–396.

[97] Bernèche, S.; Roux, B., Energetics of ion conduction through the K+ channel, *Nature* **2001**, *414*, 73–77.

[98] Iannuzzi, M.; Laio, A.; Parrinello, M., Efficient exploration of reactive potential energy surfaces using Car-Parrinello molecular dynamics., *Phys. Rev. Lett.* **2003**, *90*, 238302.

[99] Car, R.; Parrinello, M., Unified Approach for Molecular Dynamics and Density-Functional Theory, *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.

[100] Laio, A.; Gervasio, F. L., Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science, *Rep. Prog. Phys.* **2008**, *71*, 126601.

[101] Barducci, A.; Bussi, G.; Parrinello, M., Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method, *Phys. Rev. Lett.* **2008**, *100*, 020603.

[102] Dama, J. F.; Parrinello, M.; Voth, G. A., Well-Tempered Metadynamics Converges Asymptotically, *Phys. Rev. Lett.* **2014**, *112*, 240602.

[103] Barducci, A.; Bonomi, M.; Parrinello, M., Linking well-tempered metadynamics simulations with experiments., *Biophys. J.* **2010**, *98*, L44–L46.

[104] Darve, E.; Rodríguez-Gómez, D.; Pohorille, A., Adaptive biasing force method for scalar and vector free energy calculations., *J. Chem. Phys.* **2008**, *128*, 144120.

[105] Comer, J.; Gumbart, J. C.; Hénin, J.; Lelièvre, T.; Pohorille, A.; Chipot, C., The Adaptive Biasing Force Method: Everything You Always Wanted to Know but Were Afraid to Ask, *J. Phys. Chem. B* **2015**, *119*, 1129–1151.

[106] Fu, H.; Shao, X.; Cai, W.; Chipot, C., Taming Rugged Free Energy Landscapes Using an Average Force, *Acc. Chem. Res.* **2019**, *52*, 3254–3264.

[107] Lelièvre, T.; Rousset, M.; Stoltz, G., Computation of free energy profiles with parallel adaptive dynamics, *J. Chem. Phys.* **2007**, *126*, 134111.

[108] Fu, H.; Shao, X.; Chipot, C.; Cai, W., Extended Adaptive Biasing Force Algorithm. An On–the–Fly Implementation for Accurate Free–Energy Calculations, *J. Chem. Theory Comput.* **2016**, *12*, 3506–3513.

[109] Lesage, A.; Lelièvre, T.; Stoltz, G.; Hénin, J., Smoothed Biasing Forces Yield Unbiased Free Energies with the Extended-System Adaptive Biasing Force Method, *J. Phys. Chem. B* **2017**, *121*, 3676–3685.

[110] Lelièvre, T.; Rousset, M.; Stoltz, G., Free Energy Computations, Imperial college press, 2010.

[111] Zheng, L.; Yang, W., Practically Efficient and Robust Free Energy Calculations: Double-Integration Orthogonal Space Tempering, *J. Chem. Theory Comput.* **2012**, *8*, 810–823.

[112] Hénin, J.; Fiorin, G.; Chipot, C.; Klein, M. L., Exploring Multidimensional Free Energy Landscapes Using Time-Dependent Biases on Collective Variables, *J. Chem. Theory Comput.* **2010**, *6*, 35–47.

[113] Cao, L.; Stoltz, G.; Lelièvre, T.; Marinica, M.-C.; Athènes, M., Free energy calculations from adaptive molecular dynamics simulations with adiabatic reweighting, *J. Chem. Phys.* **2014**, *140*, 104108.

[114] Mones, L.; Bernstein, N.; Csányi, G., Exploration, Sampling, And Reconstruction of Free Energy Surfaces with Gaussian Process Regression, *J. Chem. Theory Comput.* **2016**, *12*, 5100–5110.

[115] Fu, H.; Zhang, H.; Chen, H.; Shao, X.; Chipot, C.; Cai, W., Zooming across the Free-Energy Landscape: Shaving Barriers, and Flooding Valleys, *J. Phys. Chem. Lett.* **2018**, *9*, 4738–4745.

[116] Chen, H.; Fu, H.; Chipot, C.; Shao, X.; Cai, W., Overcoming Free-Energy Barriers with a Seamless Combination of a Biasing Force and a Collective Variable-Independent Boost Potential, *J. Chem. Theory Comput.* **2021**, *17*, 3886–3894.

[117] Fu, H.; Chen, H.; Wang, X.; Chai, H.; Shao, X.; Cai, W.; Chipot, C., Finding an Optimal Pathway on a Multidimensional Free-Energy Landscape, *J. Chem. Inf. Model.* **2020**, *60*, 5366–5374.

[118] Blazhynska, M.; Goulard Coderc de Lacam, E.; Chen, H.; Chipot, C., Improving Speed and Affordability without Compromising Accuracy: Standard Binding Free-Energy Calculations Using an Enhanced Sampling Algorithm, Multiple-Time Stepping, and Hydrogen Mass Repartitioning, *J. Chem. Theory Comput.* **2023**, *19*, 3091–3101.

[119] Roberts, G. O.; Stramer, O., Langevin Diffusions and Metropolis-Hastings Algorithms, *Methodol. Comput. Appl. Probab.* **2002**, *4*, 337–357.

[120] Robert, C. P.; Elvira, V.; Tawn, N.; Wu, C., Accelerating MCMC algorithms, *WIREs Comp. Stats.* **2018**, *10*, e1435–14.

[121] Betancourt, M., The Convergence of Markov Chain Monte Carlo Methods: From the Metropolis Method to Hamiltonian Monte Carlo, *Ann. Phys. (Berlin)* **2019**, *531*, 1700214.

[122] Skeel, R. D.; Hartmann, C., Choice of Damping Coefficient in Langevin Dynamics, *Eur. Phys. J. B* **2021**, *94*, 178.

[123] Fu, H.; Chen, H. Binding Free Energy Estimator 2. https://github.com/fhh2626/BFEE2, 2023.

[124] Zheng, L.; Chen, M.; Yang, W., Random walk in orthogonal space to achieve efficient free-energy simulation of complex systems, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 20227–20232.

[125] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E., The Protein Data Bank (2000), *Nucleic Acids Res.* **2000**, *28*, 235–242.

[126] Berman, H. M. et al., RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning (2023), *Nucleic Acids Res.* **2023**, *51*, D488–D508.

[127] Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E., GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers, *SoftwareX* **2015**, *1-2*, 19–25.

[128] Jo, S.; Kim, T.; Iyer, V. G.; Im, W., CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM, *J. Comput. Chem.* **2008**, *29*, 1859–1865.

[129] Case, D. A.; Belfon, K.; Ben-Shalom, I. Y.; Brozell, S. R.; Cerutti, D. S.; Cheatham III, T.E.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Giambasu, G. et al. Amber 2020, 2020.

[130] Flyvbjerg, H.; Petersen, H. G., Error estimates on averages of correlated data, *J. Chem. Phys.* **1989**, *91*, 461–466.

[131] Holzer, P.; Masuya, K.; Furet, P.; Kallen, J.; Valat-Stachyra, T.; Ferretti, S.; Berghausen, J.; Bouisset-Leonard, M.; Buschmann, N.; Pissot-Soldermann, C. et al., Discovery of a Dihydroisoquinolinone Derivative (NVP-CGM097): A Highly Potent and Selective MDM2 Inhibitor Undergoing Phase 1 Clinical Trials in p53wt Tumors., *J. Med. Chem.* **2015**, *58*, 6348–58.

[132] Bauer, S.; Demetri, G. D.; Halilovic, E.; Dummer, R.; Meille, C.; Tan, D. S. W.; Guerreiro, N.; Jullion, A.; Ferretti, S.; Jeay, S. et al., Pharmacokinetic-pharmacodynamic guided optimisation of dose and schedule of CGM097, an HDM2 inhibitor, in preclinical and clinical studies, *Br. J. Cancer* **2021**, *125*, 687–698.

[133] Zhang, M.; Chen, X.; Dong, X.; Wang, J.; Feng, W.; Teng, Q.; Cui, Q.; Li, J.; Li, X.; Chen, Z., NVP-CGM097, an HDM2 Inhibitor, Antagonizes ATP-Binding Cassette Subfamily B Member 1-Mediated Drug Resistance, *Front. Oncol.* **2020**, *10*, 1219.

[134] Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, J. A. D., CHARMM36m: an Improved Force Field for Folded and Intrinsically Disordered Proteins., *Nat. methods* **2017**, *14*, 71–73.

[135] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, J. A. D., CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields., *J. Comput. Chem.* **2010**, *31*, 671–90.

[136] Vanommeslaeghe, K.; MacKerell, J. A. D., Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing, *J. Chem. Inf. Model.* **2012**, *52*, 3144–3154.

[137] Uhlenbeck, G. E.; Ornstein, L. S., On the Theory of the Brownian Motion, *Phys. Rev.* **1930**, *36*, 823–841.

[138] Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R., Constant Pressure Molecular Dynamics Simulation: The Langevin Piston Method, *J. Chem. Phys.* **1995**, *103*, 4613–4621.

[139] Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An N .log(N Method for Ewald Sums in Large Systems, *Chem. Phys.* **1993**, *98*, 10089–10092.

[140] Bingham, R. J.; Findlay, J. B.C.; Hsieh, S.-Y.; Kalverda, A. P.; Kjellberg, A.; Perazzolo, C.; Phillips, S. E.V.; Seshadri, K.; Trinh, C. H.; Turnbull, W. B.; Bodenhausen, G.; Homans, S. W., Thermodynamics of binding of 2-methoxy-3-isopropylpyrazine and 2-methoxy-3-isobutylpyrazine to the major urinary protein, *J. Am. Chem. Soc.* **2004**, *126*, 1675–1681.

[141] Timm, D. E.; Baker, L. J.; Mueller, H.; Zidek, L.; Novotny, M. V., Structural basis of pheromone binding to mouse major urinary protein (MUP-I), *Prot. Sci.* **2001**, *10*, 997–1004.

[142] Fu, H.; Cai, W.; Hénin, J.; Roux, B.; Chipot, C., New Coarse Variables for the Accurate Determination of Standard Binding Free Energies, *J. Chem. Theory Comput.* **2017**, *13*, 5173–5178.

[143] Baker, E. N.; Blundell, T. L.; Cutfield, J. F.; Dodson, E. J.; Dodson, G. G.; Hodgkin, D.; Crowfoot, M.; Hubbard, R. E.; Isaacs, N. W.; Reynolds, C. D. et al., The Structure of 2Zn Pig Insulin Crystals at 1.5 Å Resolution, *Phil. Trans. R. Soc. Lond.* **1988**, *319*, 369–456.

[144] Zoete, V.; Meuwly, M.; Karplus, M., Study of the Insulin Dimerization: Binding Free Energy Calculations and per-Residue Free Energy Decomposition., *Proteins* **2005**, *61*, 79–93.

[145] Ganim, Z.; Jones, K. C.; Tokmakoff, A., Insulin dimer dissociation and unfolding revealed by amide I two-dimensional infrared spectroscopy., *Phys. Chem. Chem. Phys.* **2010**, *12*, 3579–88.

[146] Rege, N. K.; Wickramasinghe, N. P.; Tustan, A. N.; Phillips, N. F. B.; Yee, V. C.; Ismail-Beigi, F.; Weiss, M. A., Structure-Based Stabilization of Insulin as a Therapeutic Protein Assembly via Enhanced Aromatic-Aromatic Interactions., *J. Biol. Chem.* **2018**, *293*, 10895–10910.

[147] Banerjee, P.; Mondal, S.; Bagchi, B., Insulin Dimer Dissociation in Aqueous Solution: A Computational Study of Free Energy Landscape and Evolving Microscopic Structure along the Reaction Pathway., *J. Chem. Phys.* **2018**, *149*, 114902.

[148] Tse, C.; Wickstrom, L.; Kvaratskhelia, M.; Gallicchio, E.; Levy, R.; Deng, N., Exploring the Free-Energy Landscape and Thermodynamics of Protein-Protein Association, *Biophys. J.* **2020**, *119*, 1226–1238.

[149] Perthold, J. W.; Oostenbrink, C., Simulation of Reversible Protein-Protein Binding and Calculation of Binding Free Energies Using Perturbed Distance Restraints, *J. Chem. Theory Comput.* **2017**, *13*, 5697–5708.

[150] Joshi, D. C.; Lin, J.-H., Delineating Protein-Protein Curvilinear Dissociation Pathways and Energetics with Naïve Multiple-Walker Umbrella Sampling Simulations., *J. Comput. Chem.* **2019**, *40*, 1652–1663.

[151] Wang, J.; Ishchenko, A.; Zhang, W.; Razavi, A.; Langley, D., A Highly Accurate Metadynamics-Based Dissociation Free Energy Method to Calculate Protein-Protein and Protein-Ligand Binding Potencies, *Sci. Rep.* **2022**, *12*, 2024.

[152] Doudou, S.; Burton, N. A.; Henchman, R. H., Standard Free Energy of Binding from a One-Dimensional Potential of Mean Force, *J. Chem. Theory Comput.* **2009**, *5*, 909–918.

[153] Patel, J. S.; Ytreberg, F. M., Fast Calculation of Protein-Protein Binding Free Energies Using Umbrella Sampling with a Coarse-Grained Model, *J. Chem. Theory Comput.* **2018**, *14*, 991–997.

[154] Niu, Y.; Shi, D.; Li, L.; Guo, J.; Liu, H.; Yao, X., Revealing Inhibition Difference between PFI-2 Enantiomers against SETD7 by Molecular Dynamics Simulations, Binding Free Energy Calculations and Unbinding Pathway Analysis, *Sci. Rep.* **2017**, *7*, 46547–46547–11.

[155] García-Iriepa, C.; Hognon, C.; Francés-Monerris, A.; Iriepa, I.; Miclot, T; Barone, G.; Monari, A.; Marazzi, M., Thermodynamics of the Interaction between the Spike Protein of Severe Acute Respiratory Syndrome Coronavirus-2 and the Receptor of Human Angiotensin-Converting Enzyme 2. Effects of Possible Ligands, *J. Phys. Chem. Lett.* **2020**, *11*, 9272–9281.

[156] Francés-Monerris, A.; Hognon, C.; Miclot, T.; García-Iriepa, C.; Iriepa, I.; Terenzi, A.; Grandemange, S.; Barone, G.; Marazzi, M.; Monari, A., Molecular Basis of SARS-CoV-2 Infection and Rational Design of Potential Antiviral Agents: Modeling and Simulation Approaches, *J. Proteome Res.* **2020**, *19*, 4291–4315.

[157] Lapelosa, M., Conformational Dynamics and Free Energy of BHRF1 Binding to Bim BH3, *Biophys. Chem.* **2018**, *232*, 22–28.

[158] Zuo, Z.; Wang, B.; Weng, J.; Wang, W., Stepwise Substrate Translocation Mechanism Revealed by Free Energy Calculations of Doxorubicin in the Multidrug Transporter AcrB, *Sci. Rep.* **2015**, *5*, 13905–13911.

[159] Jin, X.; Bai, Q.; Xue, W.; Liu, H.; Yao, X., Computational Study on the Inhibition Mechanism of a Cyclic Peptide MaD5 to PfMATE: Insight from Molecular Dynamics Simulation, Free Energy Calculation and Dynamical Network Analysis, *Chemom. Intell. Lab. Syst.* **2015**, *149*, 81–88.

[160] Hernández-Alvarez, L.; B., Oliveira J. A.; Hernández-González, J. E.; Chahine, J.; Pascutti, P. G.; de Araujo, A. S.; de Souza, F. P., Computational Study on the Allosteric Mechanism of Leishmania Major IF4E-1 by 4E-Interacting Protein-1: Unravelling the Determinants of m7GTP Cap Recognition, *Comput. Struct. Biotechnol. J.* **2021**, *19*, 2027–2044.

[161] Hu, F.; Liu, X.-T.; Zhang, J.-L.; Zheng, Q.-C.; Eglitis, R. I.; Zhang, H.-X., MD Simulation Investigation on the Binding Process of Smoke-Derived Germination Stimulants to Its Receptor, *J. Chem. Inf. Model.* **2019**, *59*, 1554–1562.

[162] Pisabarro, M. T.; Serrano, L.; Wilmanns, M., Crystal Structure of the Abl-SH3 Domain Complexed with a Designed High-Affinity Peptide Ligand: Implications for SH3-Ligand Interactions, *J. Mol. Biol.* **1998**, *281*, 513–521.

[163] Strazza, S.; Hunter, R.; Walker, E.; Darnall, D. W., The thermodynamics of bovine and porcine insulin and proinsulin association determined by concentration difference spectroscopy, *Arch. Biochem. Biophys.* **1985**, *238*, 30–42.

[164] Lan, J.; Ge, J.; Yu, J.; Shan, S.; Zhou, H.; Fan, S.; Zhang, Q.; Shi, X.; Wang, Q. et al., Structure of the SARS-CoV-2 Spike Receptor-Binding Domain Bound to the ACE2 Receptor, *Nature* **2020**, *581*, 215–220.

[165] Massart, D. L.; Vandeginste, B. G. M.; Deming, S. N.; Michotte, Y.; Kaufman, L., Data Handling in Science and Technology: Chapter 14 - Correlation Methods, Elsevier: Amsterdam, 2003.

[166] Palencia, A.; Camara-Artigas, A.; Pisabarro, M. T.; Martinez, J. C.; Luque, I., Role of Interfacial Water Molecules in Proline-Rich Ligand Recognition by the Src Homology 3 Domain of Abl., *J. Biol. Chem.* **2010**, *285*, 2823–33.

[167] Zafra-Ruano, A.; Luque, I., Interfacial Water Molecules in SH3 Interactions: Getting the Full Picture on Polyproline Recognition by Protein-Protein Interaction Domains, *FEBS Lett.* **2012**, *586*, 2619–2630.

[168] Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S. et al., Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein, *ACS Cent. Sci.* **2020**, *6*, 1722–1734.

[169] Harvey, W. T.; Carabelli, A. M.; Jackson, B.; Gupta, R. K.; Thomson, E. C.; Harrison, E. M.; Ludden, C.; Reeve, R.; Rambaut, A.; Peacock, S. J.; Robertson, D. L., SARS-CoV-2 variants, Spike Mutations and Immune Escape, *Nat. Rev. Microbiol.* **2021**, *19*, 409–424.

[170] McCarthy, K. R.; Rennick, L. J.; Nambulli, S.; Robinson-McCarthy, L. R.; Bain, W. G.; Haidar, G.; Duprex, W. P., Recurrent Deletions in the SARS-CoV-2 Spike Glycoprotein Drive Antibody Escape, *Science* **2021**, *371*, 1139.

[171] Horton, R., Offline: COVID-19 is not a pandemic, *Lancet* **2020**, *396*, 874.

[172] World Health Organization. WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int.

[173] Tsang, J. L. Y.; Binnie, A.; Fowler, R. A., Twenty articles that critical care clinicians should read about COVID-19, *Intensive Care Med* **2021**, *47*, 337–341.

[174] Nalbandian, A.; Sehgal, K.; Gupta, A. et al., Post-acute COVID-19 syndrome, *Nat. Med.* **2021**, *27*, 601–615.

[175] Hadfield, J.; Megill, C.; Bell, S. M.; Huddleston, J.; Potter, B.; Callender, C.; Sagulenko, P.; Bedford, T.; Neher, R. A., Nextstrain: real-time tracking of pathogen evolution, *Bioinformatics* **2018**, *34*, 4121–4123.

[176] Jackson, C. B.; Farzan, M.; Chen, B.; Choe, H., Mechanisms of SARS-CoV-2 entry into cells, *Nat. Rev. Mol. Cell. Biol.* **2022**, *23*, 3–20.

[177] Yao, H.; Song, Y.; Chen, Y.; Wu, N.; Xu, J.; Sun, C.; Zhang, J.; Weng, T.; Zhang, Z.; Wu, Z. et al., Molecular Architecture of the SARS-CoV-2 Virus, *Cell* **2020**, *183*, 730–738.e13.

[178] Walls, A. C.; Park, Y-J.; Tortorici, M. A.; Wall, A.; McGuire, A. T.; Veesler, D., Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein, *Cell* **2020**, *181*, 281–292.e6.

[179] Shang, J.; Ye, G.; Shi, K.; Wan, Y.; Luo, C.; Aihara, H.; Geng, Q.; Auerbach, A.; Li, F., Structural basis of receptor recognition by SARS-CoV-2, *Nature* **2020**, *581*, 221–224.

[180] Lan, J.; Ge, J.; Yu, J.; Shan, S.; Zhou, H.; Fan, S.; Zhang, Q.; Shi, X.; Wang, Q.; Zhang, L.; Wang, X., Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor, *Nature* **2020**, *581*, 215–220.

[181] Hoffmann, M.; Kleine-Weber, H.; Pöhlmann, S., A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells, *Mol. Cell* **2020**, *78*, 779–784.e5.

[182] Benton, D. J.; Wrobel, A. G.; Xu, P.; Roustan, C. et al., Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion, *Nature* **2020**, *588*, 327–330.

[183] World Health Organization. Tracking SARS-CoV-2 variants. https://www.who.int/health-topics/typhoid/tracking-SARS-CoV-2-variants, accessed January 28, 2022.

[184] Liu, C.; Zhou, Q.; Li, Y.; Garner, L. V.; Watkins, S. P.; Carter, L. J.; Smoot, J.; Gregg, A. C. et al., Research and Development on Therapeutic Agents and Vaccines for COVID-19 and Related Human Coronavirus Diseases, *ACS Cent. Sci.* **2020**, *6*, 315–331.

[185] Volkan, E., COVID-19: Structural Considerations for Virus Pathogenesis, Therapeutic Strategies and Vaccine Design in the Novel SARS-CoV-2 Variants Era, *Mol. Biotechnol.* **June 2021**, *63*, 885–897.

[186] Alaofi, A. L.; Shahid, M., Mutations of SARS-CoV-2 RBD May Alter Its Molecular Structure to Improve Its Infection Efficiency, *Biomolecules* **2021**, *11*, 1273.

[187] Luan, B.; Huynh, T., Insights into SARS-CoV-2's Mutations for Evading Human Antibodies: Sacrifice and Survival, *J. Med. Chem.* **2022**, *65*, 2820–2826.

[188] Gobeil, S. M.-C.; Janowska, K.; McDowell, S.; Mansouri, K.; Parks, R.; Stalls, V.; Kopp, M. F.; Manne, K.; Li, D.; Wiehe, K. et al., Effect of natural mutations of SARS-CoV-2 on spike structure, conformation, and antigenicity, *Science* **2021**, *373*, eabi6226.

[189] Zhang, J.; Xiao, T.; Cai, Y.; Lavine, C. L.; Peng, H.; Zhu, H.; Anand, K.; Tong, P.; Gautam, A.; Mayer, M. L. et al., Membrane fusion and immune evasion by the spike protein of SARS-CoV-2 Delta variant, *Science* **2021**, *374*, 1353–1360.

[190] Kim, S.; Liu, Y.; Lei, Z.; Dicker, J.; Cao, Y.; Zhang, X. F.; Im, W., Differential Interactions between Human ACE2 and Spike RBD of SARS-CoV-2 Variants of Concern, *J. Chem. Theory Comput.* **2021**, *17*, 7972–7979.

[191] Khan, A.; Wei, D.-Q.; Kousar, K.; Abubaker, J.; Ahmad, S.; Ali, J.; et al., Preliminary Structural Data Revealed That the SARS-CoV-2 B.1.617 Variant's RBD Binds to ACE2 Receptor Stronger Than the Wild Type to Enhance the Infectivity, *ChemBioChem* **2021**, *22*, 2641–2649.

[192] Chakraborty, S., E484K and N501Y SARS-CoV 2 spike mutants Increase ACE2 recognition but reduce affinity for neutralizing antibody, *Int. Immunopharmacol.* **2022**, *102*, 108424.

[193] Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M., The weighted histogram analysis method for free energy calculations on biomolecules. I. The method, *J. Comput. Chem.* **1992**, *13*, 1011–1021.

[194] Souaille, M.; Roux, B., Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations, *Comput. Phys. Commun.* **2001**, *135*, 40–57.

[195] Nutalai, R.; Zhou, D.; Tuekprakhon, A.; Ginn, H. M.; Supasa, P.; Liu, C. et al., Potent cross-reactive antibodies following Omicron breakthrough in vaccinees, *Cell* **2022**, *185*, 2116–2131.e18.

[196] Mannar, D.; Saville, J. W.; Zhu, X.; Srivastava, S. S.; Berezuk, A. M. et al., Structural analysis of receptor binding domain mutations in SARS-CoV-2 variants of concern that modulate ACE2 and antibody binding, *Cell Rep.* **2021**, *37*, 110156.

[197] McCallum, M.; Walls, A. C.; Sprouse, K. R.; Bowen, J. E.; Rosen, L. E.; Dang, H. V. et al., Molecular basis of immune evasion by the delta and kappa SARS-CoV-2 variants, *Science (N.Y.)* **2021**, *374(6575)*, 1621–1626.

[198] Yang, T.-J.; Yu, P.-Y.; Chang, Y.-C.; Liang, K.-H.; Tso, H.-C.; Ho, M.-R.; Chen, W.-Y.; Lin, H.-T.; Wu, H.-C.; Hsu, S.-T. D., Effect of SARS-CoV-2 B.1.1.7 mutations on spike protein structure and function, *Nat. Struct. Mol. Biol.* **2021**, *28*, 731–739.

[199] Han, P.; Su, C.; Zhang, Y.; Bai, C.; Zheng, A.; Qiao, C.; Wang, Q.; Niu, S.; Chen, Q.; Zhang, Y. et al., Molecular insights into receptor binding of recent emerging SARS-CoV-2 variants, *Nat. Commun.* **2021**, *6103*.

[200] Starr, T.N.; Czudnochowski, N.; Liu, Z. *et al.*, SARS-CoV-2 RBD antibodies that maximize breadth and resistance to escape., *Nature* **2021**, *597*, 97—-102.

[201] Huo, J.; Le Bas, A.; Ruza, R. R.; Duyvesteyn, H. M. E.; Mikolajek, H. et al., Neutralizing nanobodies bind SARS-CoV-2 spike RBD and block interaction with ACE2, *Nat. Struct. Mol. Biol.* **2020**, *27*, 846–854.

[202] Mlcochova, P.; Kemp, S. A.; Dhar, M. S.; Papa, G.; Meng, B.; Ferreira, I. A. T. M.; Datir, R.; Collier, D. A.; Albecka, A.; Singh, S. et al., SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion, *Nature* **2021**, *599*, 114–119.

[203] Acharya, A.; Lynch, D. L.; Pavlova, A.; Pang, Y. T.; Gumbart, J. C., ACE2 glycans preferentially interact with SARS-CoV-2 over SARS-CoV, *Chem. Commun.* **2021**, *57*, 5949–5952.

[204] Huang, Y.; Harris, B. S.; Minami, S. A.; Jung, S.; Shah, P. S.; Nandi, S.; McDonald, K. A.; Faller, R., SARS-CoV-2 spike binding to ACE2 is stronger and longer ranged due to glycan interaction, *Biophys. J.* **2022**, *121*, 79–90.

[205] Wrapp, D.; Wang, N.; Corbett, K. S.; Goldsmith, J. A.; Hsieh, C. L.; Abiona, O.; Graham, B. S.; McLellan, J. S., Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation, *Science* **2020**, *367*, 1260–1263.

[206] Edara, V.-V.; Pinsky, B. A.; Suthar, M. S.; Lai, L.; Davis-Gardner, M. E. et al., Infection and Vaccine-Induced Neutralizing-Antibody Responses to the SARS-CoV-2 B.1.617 Variants, *N. Engl. J. Med.* **2021**, *385*, 664–666.

[207] Zhang, H.; Kim, S.; Giese, T. J.; Lee, T-S.; Lee, J.; York, D. M.; Im, W., CHARMM-GUI Free Energy Calculator for Practical Ligand Binding Free Energy Simulations with AMBER, *J. Chem. Inf. Model.* **2021**, *61*, 4145–4151.

[208] Fratev, F., N501Y and K417N Mutations in the Spike Protein of SARS-CoV-2 Alter the Interactions with Both hACE2 and Human-Derived Antibody: A Free Energy of Perturbation Retrospective Study, *J. Chem. Inf. Model.* **2021**, *61*, 6079–6084.

[209] Pavlova, A.; Zhang, Z.; Acharya, A.; Lynch, D. L.; Pang, Y-T.; Mou, Z.; Parks, J. M.; Chipot, C.; Gumbart, J. C., Machine Learning Reveals the Critical Interactions for SARS-CoV-2 Spike Protein Binding to ACE2, *J. Phys. Chem. Lett.* **2021**, *12*, 5494–5502.

[210] Starr, T. N; Greaney, A. J.; Addetia, A.; Hannon, W. W; Choudhary, M. C.; Dingens, A. S.; Li, J. Z.; Bloom, J. D., Prospective mapping of viral mutations that escape antibodies used to treat COVID-19, *Science (N.Y.)* **2021**, *371*, 850–854.

[211] Tortorici, M. A.; Beltramello, M.; Lempp, F. A.; Pinto, D.; Dang, H. V.; Rosen, L. E.; McCallum, M. et al., Ultrapotent human antibodies protect against SARS-CoV-2 challenge via multiple mechanisms, *Science* **2020**, *370*, 950–957.

[212] Huang, M.; Wu, L.; Zheng, A.; Xie, Y. et al., Atlas of currently available human neutralizing antibodies against SARS-CoV-2 and escape by Omicron sub-variants BA.1/BA.1.1/BA.2/BA.3, *Immunity* **2022**, *55(8)*, 1501–1514.e3.

[213] Bhattarai, N.; Baral, P.; Gerstman, B. S.; Chapagain, P. P., Structural and Dynamical Differences in the Spike Protein RBD in the SARS-CoV-2 Variants B.1.1.7 and B.1.351, *J. Phys. Chem. B* **2021**, *125*, 7101–7107.

[214] Socher, E.; Conrad, L.; Paulsen, F.; Sticht, H.; Zunke, F.; Arnold, P., Computational decomposition reveals reshaping of the SARS-CoV-2–ACE2 interface among viral variants expressing the N501Y mutation, *J. Cell. Biochem.* **2021**, *122*, 1863–1872.

[215] Chen, J.; Wang, R.; Gilby, N. B.; Wei, G. W., Omicron Variant (B.1.1.529): Infectivity, Vaccine Breakthrough, and Antibody Resistance, *J. Chem. Inf. Model.* **2022**, *62*, 412–422.

[216] Wu, L.; Zhou, L.; Mo, M.; Liu, T.; Wu, C.; Gong, C.; Lu, K.; Gong, L.; Zhu, W.; Xu, Z., SARS-CoV-2 Omicron RBD shows weaker binding affinity than the currently dominant Delta variant to human ACE2, *Sig. Transduct. Target. Ther.* **2022**, *7*, 8.

[217] Kim, S.; Liu, Y.; Ziarnik, M.; Cao, Y.; Zhang, X. F.; Im, W., Binding of Human ACE2 and RBD of Omicron Enhanced by Unique Interaction Patterns Among SARS-CoV-2 Variants of Concern, *bioRxiv* **2022**, 2022.01.24.477633.

[218] Nguyen, H. L.; Thai, N. Q.; Nguyen, P. H.; Li, M. S., SARS-CoV-2 Omicron Variant Binds to Human Cells More Strongly than the Wild Type: Evidence from Molecular Dynamics Simulation, *J. Phys. Chem. B* **2022**, *126*, 4669–4678.

[219] Hong, Q.; Han, W.; Li, J.; Xu, S.; Wang, Y.; Xu, C.; Li, Z. et al., Molecular basis of receptor binding and antibody neutralization of Omicron, *Nature* **2022**, *604*, 546–552.

[220] Cao, Y.; Wang, J.; Jian, F. et al., Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies, *Nature* **2022**, *602*, 657–663. Article.

[221] Zhou, J.; Sukhova, K.; McKay, P. F.; et al., A. Kurshanand, Omicron breakthrough infections in vaccinated or previously infected hamsters, *bioRxiv* **2022**. Preprint at bioRxiv.

[222] Willett, B. J.; Grove, J.; MacLean, O. A. et al., SARS-CoV-2 Omicron is an immune escape variant with an altered cell entry pathway, *Nat Microbiol* **2022**, *7*, 1161–1179.

[223] Carabelli, A. M.; Peacock, T. P.; Thorne, L. G. et al., SARS-CoV-2 variant biology: immune escape, transmission and fitness, *Nat Rev Microbiol* **2023**, *21*, 162–177.

[224] Adams, P. D.; Engelman, D. M.; Brünger, A. T., Improved prediction for the structure of the dimeric transmembrane domain of glycophorin A obtained through global searching, *Proteins* **1996**, *26*, 257–261.

[225] Popot, J.-L.; Engelman, D. M., Helical membrane protein folding, stability, and evolution., *Annu. Rev. Biochem.* **2000**, *69*, 881–922.

[226] Booth, P. J.; Templer, R. H.; Meijberg, W.; Allen, S. J.; Curran, A. R.; Lorch, M., In vitro studies of membrane protein folding., *Crit. Rev. Biochem. Mol.* **2001**, *36*, 501–603.

[227] Arkin, I. T., Structural aspects of oligomerization taking place between the transmembrane α-helices of bitopic membrane proteins, *Biochim. Biophys. Acta - Biomembr.* **2002**, *1565*, 347–363.

[228] Chin, C.-N.; von Heijne, G.; de Gier, J.-W. L., Membrane proteins: shaping up, *Trends Biochem. Sci.* **2002**, *27*, 231–234.

[229] Engelman, D. M.; Chen, Y.; Chin, C.-N.; Curran, A. R.; Dixon, A. M.; Dupuy, A. D.; Lee, A. S.; Lehnert, U.; Matthews, E. E.; Reshetnyak, Y. K.; Senes, A.; Popot, J.-L., Membrane protein folding: beyond the two-stage model., *FEBS Lett.* **2003**, *555*, 122–5.

[230] Engelman, D.M.; Steitz, T.A., The spontaneous insertion of proteins into and across membranes: The helical hairpin hypothesis, *Cell* **1981**, *23*, 411–422.

[231] Popot, J.-L.; Engelman, D. M., Membrane protein folding and oligomerization: the two-stage model., *Biochemistry* **1990**, *29*, 4031–7.

[232] Bormann, B. J.; Engelman, D. M., Intramembrane Helix-Helix Association in Oligomerization and Transmembrane Signaling, *Annu. Rev. Biophys. Biomol. Struct.* **1992**, *21*, 223–242.

[233] Engelman, D. M.; Adair, B.; Brunger, A.; Hunt, J.; Kahn, T.; Lemmon, M.; MacKenzie, K.; Treutlein, H., in *Biochemistry of Cell Membranes: A Compendium of Selected Topics*, Birkhäuser Basel: Basel, 1995, pp. 297–310.

[234] MacKenzie, K. R.; Prestegard, J. H.; Engelman, D. M., A transmembrane helix dimer: structure and implications., *J. Sci.* **1997**, *276*, 131–133.

[235] Bormann, B. J.; Knowles, W. J.; Marchesi, V. T., Synthetic peptides mimic the assembly of transmembrane glycoproteins., *J. Biol. Chem.* **1989**, *264*, 4033–4037.

[236] Lemmon, M. A.; Flanagan, J. M.; Treutlein, H. R.; Zhang, J.; Engelman, D. M., Sequence specificity in the dimerization of transmembrane alpha-helices., *Biochemistry* **1992**, *31*, 12719–12725.

[237] Smith, S. O.; Song, D.; Shekar, S.; Groesbeek, M.; Ziliox, M.; Aimoto, S., Structure of the transmembrane dimer interface of glycophorin A in membrane bilayers., *Biochemistry* **2001**, *40*, 6553–6558.

[238] Lemmon, M. A.; Flanagan, J. M.; Hunt, J. F.; Adair, B. D.; Bormann, B. J.; Dempsey, C. E.; Engelman, D. M., Glycophorin A dimerization is driven by specific interactions between transmembrane alpha-helices., *J. Biol. Chem.* **1992**, *267*, 7683–7689.

[239] Lemmon, M. A.; Treutlein, H. R.; Adams, P. D.; Brünger, A. T.; Engelman, D. M., A dimerization motif for transmembrane alpha-helices., *Nat. Struct. Mol. Biol.* **1994**, *1*, 157–63.

[240] Lemmon, M. A.; Engelman, D. M., Specificity and promiscuity in membrane helix interactions., *Q. Rev. Biophys.* **1994**, *27*, 157–218.

[241] Langosch, D.; Brosig, B.; Kolmar, H.; Fritz, H.-J., Dimerisation of the Glycophorin A Transmembrane Segment in Membranes Probed with the ToxR Transcription Activator, *J. Mol. Biol.* **1996**, *263*, 525–530.

[242] Mingarro, I.; Whitley, P.; Lemmon, M. A.; von Heijne, G., Ala-insertion scanning mutagenesis of the glycophorin A transmembrane helix: a rapid way to map helix-helix interactions in integral membrane proteins., *Protein Sci.* **1996**, *5*, 1339–1341.

[243] Brosig, B.; Langosch, D., The dimerization motif of the glycophorin A transmembrane segment in membranes: importance of glycine residues., *Protein Sci.* **1998**, *7*, 1052–6.

[244] Russ, W. P.; Engelman, D. M., TOXCAT: a measure of transmembrane helix association in a biological membrane., *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 863–8.

[245] Fleming, K. G.; Engelman, D. M., Specificity in transmembrane helix-helix interactions can define a hierarchy of stability for sequence variants., *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14340–14344.

[246] Fisher, L. E.; Engelman, D. M.; Sturgis, J. N., Effect of detergents on the association of the glycophorin a transmembrane helix., *Biophys. J.* **2003**, *85*, 3097–3105.

[247] Doura, A. K.; Fleming, K. G., Complex Interactions at the Helix–Helix Interface Stabilize the Glycophorin A Transmembrane Dimer, *J. Mol. Biol.* **2004**, *343*, 1487–1497.

[248] Doura, A. K.; Kobus, F. J.; Dubrovsky, L.; Hibbard, E.; Fleming, K. G., Sequence context modulates the stability of a GxxxG-mediated transmembrane helix-helix dimer., *J. Mol. Biol.* **2004**, *341*, 991–998.

[249] Smith, S. O.; Bormann, B. J., Determination of helix-helix interactions in membranes by rotational resonance NMR, *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 488–491.

[250] Trenker, R.; Call, M. E.; Call, M. J., Crystal Structure of the Glycophorin A Transmembrane Dimer in Lipidic Cubic Phase, *J. Am. Chem. Soc.* **2015**, *137*, 15676–15679.

[251] Janosi, L.; Prakash, A.; Doxastakis, M., Lipid-modulated sequence-specific association of glycophorin A in membranes., *Biophys. J.* **2010**, *99*, 284–292.

[252] Souza, P. C. T.; Alessandri, R.; Barnoud, J.; Thallmair, S.; Faustino, I.; Grünewald, F.; Patmanidis, I.; Abdizadeh, H.; Bruininks, B. M. H.; Wassenaar, T. A. et al., Martini 3: a general purpose force field for coarse-grained molecular dynamics, *Nat. Methods* **2021**, *18*, 382–388.

[253] Fleming, K. G.; Ackerman, A. L.; Engelman, D. M., The effect of point mutations on the free energy of transmembrane alpha-helix dimerization., *J. Mol. Biol.* **1997**, *272*, 266–75.

[254] Fleming, K. G., Standardizing the free energy change of transmembrane helix-helix interactions., *J. Mol. Biol.* **2002**, *323*, 563–71.

[255] Fleming, K. G.; Ren, C.-C.; Doura, A. K.; Eisley, M. E.; Kobus, F. J.; Stanley, A. M., Thermodynamics of glycophorin A transmembrane helix dimerization in C14 betaine micelles., *Biophys. Chem.* **2004**, *108*, 43–49.

[256] Schneider, D.; Engelman, D. M., GALLEX, a Measurement of Heterologous Association of Transmembrane Helices in a Biological Membrane, *J. Biol. Chem.* **2003**, *278*, 3105–3111.

[257] Nash, A.; Notman, R.; Dixon, A. M., De novo design of transmembrane helix–helix interactions and measurement of stability in a biological membrane, *Biochim. Biophys. Acta - Biomembr.* **2015**, *1848*, 1248–1257.

[258] Hong, H.; Blois, T. M.; Cao, Z.; Bowie, J. U., Method to measure strong protein–protein interactions in lipid bilayers using a steric trap, *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 19802–19807.

[259] Chen, L.; Novicky, L.; Merzlyakov, M.; Hristov, T.; Hristova, K., Measuring the Energetics of Membrane Protein Dimerization in Mammalian Membranes, *J. Am. Chem. Soc.* **2010**, *132*, 3628–3635.

[260] Sarabipour, S.; Hristova, K., Glycophorin A transmembrane domain dimerization in plasma membrane vesicles derived from CHO, HEK 293T, and A431 cells., *Biochim. Biophys. Acta* **2013**, *1828*, 1829–1833.

[261] Hénin, J.; Pohorille, A.; Chipot, C., Insights into the Recognition and Association of Transmembrane α-Helices. The Free Energy of α-Helix Dimerization in Glycophorin A, *J. Am. Chem. Soc.* **2005**, *127*, 8478–8484.

[262] Domański, J.; Sansom, M. S. P.; Stansfeld, P. J.; Best, R. B., Balancing Force Field Protein-Lipid Interactions To Capture Transmembrane Helix-Helix Association., *J. Chem. Theory Comput.* **2018**, *14*, 1706–1715.

[263] Domański, J.; Hedger, G.; Best, R. B.; Stansfeld, P. J.; Sansom, M. S. P., Convergence and Sampling in Determining Free Energy Landscapes for Membrane Protein Association, *J. Phys. Chem. B* **2017**, *121*, 3364–3375.

[264] Majumder, A.; Kwon, S.; Straub, J. E., On Computing Equilibrium Binding Constants for Protein-Protein Association in Membranes, *J. Chem. Theory Comput.* **2022**, *18*, 3961–3971.

[265] Sengupta, D.; Marrink, S. J., Lipid-mediated interactions tune the association of glycophorin A helix and its disruptive mutants in membranes, *Phys. Chem. Chem. Phys.* **2010**, *12*, 12987–12996.

[266] Chadda, R.; Cliff, L.; Brimberry, M.; Robertson, J. L., A model-free method for measuring dimerization free energies of CLC-ec1 in lipid bilayers, *J. Gen. Physiol.* **2018**, *150*, 355–365.

[267] Lomize, A. L.; Pogozheva, I. D.; Mosberg, H. I., Quantification of helix-helix binding affinities in micelles and lipid bilayers., *Protein Sci.* **2004**, *13*, 2600–2612.

[268] Chadda, R.; Bernhardt, N.; Kelley, E. G.; Teixeira, S. C.M.; Griffith, K.; Gil-Ley, A.; Öztürk, T. N.; Hughes, L. E. et al., Membrane transporter dimerization driven by differential lipid solvation energetics of dissociated and associated states, *Elife* **2021**, *10*, e63288.

[269] Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A., The statistical-thermodynamic basis for computation of binding affinities: a critical review, *Biophys. J.* **1997**, *72*, 1047–1069.

[270] Fisher, L. E.; Engelman, D. M.; Sturgis, J. N., Detergents modulate dimerization, but not helicity, of the glycophorin A transmembrane domain., *J. Mol. Biol.* **1999**, *293*, 639–651.

[271] Le Maire, M.; Champeil, P.; Moller, J. V., Interaction of membrane proteins and lipids with solubilizing detergents., *Biochim. Biophys. Acta* **2000**, *1508*, 86–111.

[272] Zhang, J.; Lazaridis, T., Calculating the Free Energy of Association of Transmembrane Helices, *Biophys. J.* **2006**, *91*, 1710–1723.

[273] Booth, P. J., Sane in the membrane: designing systems to modulate membrane proteins, *Curr. Opin. Struct. Biol.* **2005**, *15*, 435–440.

[274] Findlay, H. E.; Booth, P. J., The biological significance of lipid–protein interactions, *J. Condens. Matter Phys.* **2006**, *18*, S1281.

[275] Harris, N. J.; Pellowe, G. A.; Blackholly, L. R.; Gulaidi-Breen, S.; Findlay, H. E.; Booth, P. J. ., Methods to study folding of alpha-helical membrane proteins in lipids, *Open Biol.* **2022**, *12*, 220054.

[276] Allen, S. J.; Curran, A. R.; Templer, R. H.; Meijberg, W.; Booth, P. J., Controlling the Folding Efficiency of an Integral Membrane Protein, *J. Mol. Biol.* **2004**, *342*, 1293–1304.

[277] Meijberg, W.; Booth, P. J., The Activation Energy for Insertion of Transmembrane α-Helices is Dependent on Membrane Composition, *J. Mol. Biol.* **2002**, *319*, 839–853.

[278] Bogdanov, M.; Heacock, P. N.; Dowhan, W., A polytopic membrane protein displays a reversible topology dependent on membrane lipid composition, *EMBO J.* **2002**, *21*, 2107–2116.

[279] Dowhan, W.; Vitrac, H.; Bogdanov, M., Lipid-Assisted Membrane Protein Folding and Topogenesis, *Protein J.* **2019**, *38*, 274–288.

[280] Hong, H.; Bowie, J. U., Dramatic Destabilization of Transmembrane Helix Interactions by Features of Natural Membrane Environments, *J. Am. Chem. Soc.* **2011**, *133*, 11389–11398.

[281] Booth, P. J.; Curnow, P., Folding scene investigation: membrane proteins, *Curr. Opin. Struct. Biol.* **2009**, *19*, 8–13.

[282] Lipinski, K.; McKay, M. J.; Afrose, F.; Martfeld, A. N.; Koeppe, R. E. 2nd; Greathouse, D. V., Influence of Lipid Saturation, Hydrophobic Length and Cholesterol on Double-Arginine-Containing Helical Peptides in Bilayer Membranes., *ChemBioChem.* **2019**, *20*, 2784–2792.

[283] Ding, W.; Palaiokostas, M.; Shahane, G.; Wang, W.; Orsi, M., Effects of High Pressure on Phospholipid Bilayers, *J. Phys. Chem. B* **2017**, *121*, 9597–9606.

[284] Javanainen, M.; Hammaren, H.; Monticelli, L.; Jeon, J.-H.; Miettinen, M. S.; Martinez-Seara, H.; Metzler, R.; Vattulainen, I., Anomalous and normal diffusion of proteins and lipids in crowded lipid membranes, *Faraday Discuss.* **2013**, *161*, 397–417.

[285] Bandara, A.; Panahi, A.; Pantelopulos, G. A.; Nagai, T.; Straub, J. E., Exploring the impact of proteins on the line tension of a phase-separating ternary lipid mixture., *J. Chem. Phys.* **2019**, *150*, 204702.

[286] Huang, J.; D., MacKerell J. A., CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data., *J. Comput. Chem.* **2013**, *34*, 2135–45.

[287] Klauda, J. B.; Venable, R. M.; Freites, J. A.; O'Connor, J. W.; Tobias, D. J.; Mondragon-Ramirez, C.; Vorobyov, I.; MacKerell, J. A. D.; Pastor, R. W., Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types, *J. Phys. Chem. B* **2010**, *114*, 7830–7843.

[288] Balusek, C.; Hwang, H.; Lau, C. H.; Lundquist, K.; Hazel, A.; Pavlova, A.; Lynch, D. L.; Reggio, P. H.; Wang, Y.; Gumbart, J. C., Accelerating Membrane Simulations with Hydrogen Mass Repartitioning., *J. Chem. Theory Comput.* **2019**, *15*, 4673–4686.

[289] Booth, P. J., Folding α-helical membrane proteins: kinetic studies on bacteriorhodopsin, *Fold Des.* **1997**, *2*, R85–R92.

[290] Neumann, J.; Klein, N.; Otzen, D. E.; Schneider, D., Folding energetics and oligomerization of polytopic α-helical transmembrane proteins, *Arch. Biochem. Biophys.* **2014**, *564*, 281–296.

[291] MacKenzie, K. R., Folding and Stability of α-Helical Integral Membrane Proteins, *Chem. Rev.* **2006**, *106*, 1931–1977.

[292] Marx, D. C.; Fleming, K. G., Local Bilayer Hydrophobicity Modulates Membrane Protein Stability, *J. Am. Chem. Soc.* **2021**, *143*, 764–772.

[293] Alenghat, F. J.; Golan, D. E., Membrane protein dynamics and functional implications in mammalian cells, *Curr. Top. Membr.* **2013**, *72*, 89–120.

[294] Treutlein, H. R.; Lemmon, M. A.; Engelman, D. M.; Brünger, A. T., The glycophorin A transmembrane domain dimer: Sequence-specific propensity for a right-handed supercoil of helices, *Biochemistry* **1992**, *31*, 12726–12732.

[295] Goulard Coderc de Lacam, E.; Blazhynska, M.; Chen, H.; Gumbart, J. C.; Chipot, C., When the dust has settled: Calculation of binding affinities from first principles for SARS-CoV-2 variants with quantitative accuracy, *J. Chem. Theory Comput.* **2022**, *18*, 5890–5900.

[296] Hénin, J.; Leliévre, T.; Shirts, M. R.; Valsson, O.; Delemotte, L., Enhanced Sampling Methods for Molecular Dynamics Simulations, *Living J. Comput. Mol. Sci.* **2022**, *4*, 1583.

[297] Carter, E. A.; Ciccotti, G.; Hynes, J. T.; Kapral, R., Constrained reaction coordinate dynamics for the simulation of rare events, *Chem. Phys. Lett.* **1989**, *156*, 472–477.

[298] Khavrutskii, I. V.; Dzubiella, J.; McCammon, J. A., Computing accurate potentials of mean force in electrolyte solutions with the generalized gradient-augmented harmonic Fourier beads method, *J. Chem. Phys.* **2008**, *128*, 044106.

[299] Chasis, J. A.; Reid, M. E.; Jensen, R. H.; Mohandas, N., Signal transduction by glycophorin A: role of extracellular and cytoplasmic domains in a modulatable process., *J. Cell Biol.* **1988**, *107*, 1351–1357.

[300] Pluhackova, K.; Kirsch, S. A.; Han, J.; Sun, Z.; Unruh, T.; Böckmann, R. A., A Critical Comparison of Biomembrane Force Fields: Structure and Dynamics of Model DMPC, POPC, and POPE Bilayers, *J. Phys. Chem. B* **2016**, *120*, 3888–3903.

[301] Hansen, S. K.; Vestergaard, M.; Thøgersen, L.; Schiøtt, B.; Nielsen, N. C.; Vosegaard, T., Lipid Dynamics Studied by Calculation of 31P Solid-State NMR Spectra Using Ensembles from Molecular Dynamics Simulations, *J. Phys. Chem. B* **2014**, *118*, 5119–5129.

[302] van Meer, G.; Voelker, D. R.; Feigenson, G. W., Membrane lipids: where they are and how they behave., *Nat. Rev. Mol.* **2008**, *9*, 112–24.

[303] Poger, D.; Caron, B.; Mark, A. E., Validating lipid force fields against experimental data: Progress, challenges and perspectives, *Biochim. Biophys. Acta - Biomembr.* **2016**, *1858*, 1556–1565.

[304] Feller, S. E.; MacKerell, J. A. D., An Improved Empirical Potential Energy Function for Molecular Simulations of Phospholipids, *J. Phys. Chem. B* **2000**, *104*, 7510–7515.

[305] MacKerell, J. A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S. et al., All-atom empirical potential for molecular modeling and dynamics studies of proteins., *J. Phys. Chem. B* **1998**, *102*, 3586–616.

[306] Klauda, J. B.; Monje, V.; Kim, T.; Im, W., Improving the CHARMM Force Field for Polyunsaturated Fatty Acid Chains, *J. Phys. Chem. B* **2012**, *116*, 9424–9431.

[307] Pastor, R. W.; MacKerell, J. A. D., Development of the CHARMM Force Field for Lipids, *J. Phys. Chem. Lett.* **2011**, *2*, 1526–1532.

[308] Wang, F.; Stuart, S. J.; Latour, R. A., Calculation of adsorption free energy for solute-surface interactions using biased replica-exchange molecular dynamics, *Biointerphases* **2008**, *3*, 9–18.

[309] Sugita, Y.; Okamoto, Y., Replica-exchange molecular dynamics for protein folding, *Chem. Phys. Lett.* **1999**, *314*, 141–151.

[310] Sugita, Y.; Kamiya, M.; Oshima, H.; Re, S.-I., Replica-Exchange Methods for Biomolecular Simulations, *Methods mol. biol.* **2019**, *2022*, 155–177.

[311] Bernardi, R. C.; Melo, M. C. R.; Schulten, K., Enhanced sampling techniques in molecular dynamics simulations of biological systems, *Biochim. Biophys. Acta Gen. Subj.* **2015**, *1850*, 872–877.

[312] Rosta, E.; Hummer, G., Error and efficiency of replica exchange molecular dynamics simulations, *J. Chem. Phys.* **2009**, *131*, 165102.

[313] Beck, D. A. C.; White, G. W. N.; Daggett, V., Exploring the energy landscape of protein folding using replica-exchange and conventional molecular dynamics simulations, *J. Struct. Biol.* **2007**, *157*, 514—523.

[314] Lelièvre, T.; Rousset, M.; Stoltz, G., Computation of free energy differences through nonequilibrium stochastic dynamics: The reaction coordinate case, *J. Comput. Phys.* **2007**, *222*, 624–643.

[315] Minoukadeh, K.; Chipot, C.; Lelièvre, T., Potential of Mean Force Calculations: A Multiple-Walker Adaptive Biasing Force Approach, *J. Chem. Theory Comput.* **2010**, *6*, 1008–1017.

[316] Comer, J.; Phillips, J. C.; Schulten, K.; Chipot, C., Multiple-Replica Strategies for Free-Energy Calculations in NAMD: Multiple-Walker Adaptive Biasing Force and Walker Selection Rules, *J. Chem. Theory Comput.* **2014**, *10*, 5276–5285.

[317] Lupas, A. N.; Bassler, J.; Dunin-Horkawicz, S., The Structure and Topology of α-Helical Coiled Coils, *Subcell. Biochem.* **2017**, *82*, 95–129.

[318] Rhys, G. G.; Wood, C. W.; Beesley, J. L.; Zaccai, N. R.; Burton, A. J.; Brady, R. L.; Thomson, A. R.; Woolfson, D. N., Navigating the Structural Landscape of De Novo α-Helical Bundles, *J. Am. Chem. Soc.* **2019**, *141*, 8787–8797.

[319] Mueller, B. K.; Subramaniam, S.; Senes, A., A frequent, GxxxG-mediated, transmembrane association motif is optimized for the formation of interhelical Cα–H hydrogen bonds, *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, E888–E895.

[320] Urano, R.; Kokubo, H.; Okamoto, Y., Predictions of Tertiary Structures of α-Helical Membrane Proteins by Replica-Exchange Method with Consideration of Helix Deformations, *J. Phys. Soc. Japan* **2015**, *84*, 084802.

[321] Domański, J.; Sansom, M. S. P.; Stansfeld, P. J.; Best, R. B., Atomistic mechanism of transmembrane helix association, *PLoS Comput. Biol.* **2020**, *16*, e1007919.

[322] Kučerka, N.; Nieh, M.-P.; Katsaras, J., Fluid phase lipid areas and bilayer thicknesses of commonly used phosphatidylcholines as a function of temperature, *Biochim. Biophys. Acta - Biomembr.* **2011**, *1808*, 2761–2771.

[323] MacKenzie, K. R.; Engelman, D. M., Structure-based prediction of the stability of transmembrane helix-helix interactions: the sequence dependence of glycophorin A dimerization, *Proc. Natl. Acad. Sci. U.S.A* **1998**, *95*, 3583–3590.

[324] Zhuang, X.; Makover, J. R.; Im, W.; Klauda, J. B., A systematic molecular dynamics simulation study of temperature dependent bilayer structural properties, *Biochim. Biophys. Acta - Biomembr.* **2014**, *1838*, 2520–2529.

[325] Kucerka, N.; Tristram-Nagle, S.; Nagle, J. F., Structure of fully hydrated fluid phase lipid bilayers with monounsaturated chains., *J. Membr. Biol.* **2005**, *208*, 193–202.

[326] Sun, D.; Peyear, T. A.; Bennett, W. F. D.; Andersen, O. S.; Lightstone, F. C.; Ingólfsson, H. I., Molecular Mechanism for Gramicidin Dimerization and Dissociation in Bilayers of Different Thickness, *Biophys. J.* **2019**, *117*, 1831–1844.

[327] Andersen, O. S.; Koeppe, R. E., Bilayer Thickness and Membrane Protein Function: An Energetic Perspective, *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 107–130.

[328] Marčelja, S., Lipid-mediated protein interaction in membranes, *Biochim. Biophys. Acta - Biomembr.* **1976**, *455*, 1–7.

[329] Marsh, D., Protein modulation of lipids, and vice-versa, in membranes, *Biochim. Biophys. Acta - Biomembr.* **2008**, *1778*, 1545–1575.

[330] Mondal, S.; Khelashvili, G.; Weinstein, H., Not Just an Oil Slick: How the Energetics of Protein-Membrane Interactions Impacts the Function and Organization of Transmembrane Proteins, *Biophys. J.* **2014**, *106*, 2305–2316.

[331] Mouritsen, O. G.; Bloom, M., Models of Lipid-Protein Interactions in Membranes, *Annu. Rev. Biophys. Biomol. Struct.* **1993**, *22*, 145–171.

[332] Goforth, R. L.; Chi, A. K.; Greathouse, D. V.; Providence, L. L.; Koeppe, R. E. II; Andersen, O. S., Hydrophobic Coupling of Lipid Bilayer Energetics to Channel Function, *J. Gen. Physiol.* **2003**, *121*, 477–493.

[333] Sparr, W. L.; Nazarov, P. V.; Rijkers, D. T. S.; Hemminga, M.A.; Tieleman, D. P.; Killian, J. A., Self-association of Transmembrane α-Helices in Model Membranes: Importance of helix orientation and role of hydrophobic mismatch, *J. Biol. Chem.* **2005**, *280*, 39324–39331.

[334] Sperotto, M. M.; Mouritsen, O. G., Mean-field and Monte Carlo simulation studies of the lateral distribution of proteins in membranes, *Eur. Biophys. J.* **1991**, *19*, 157–168.

[335] Sperotto, M. M.; Mouritsen, O. G., Lipid enrichment and selectivity of integral membrane proteins in two-component lipid bilayers, *Eur. Biophys. J.* **1993**, *22*, 323–328.

[336] Soubias, O.; Teague, W. E. J.; Hines, K. G.; Gawrisch, K., Rhodopsin/lipid hydrophobic matching-rhodopsin oligomerization and function., *Biophys. J.* **2015**, *108*, 1125–32.

[337] Grau-Campistany, A.; Strandberg, E.; Wadhwani, P.; Reichert, J.; Bürck, J.; Rabanal, F.; Ulrich, A. S., Hydrophobic mismatch demonstrated for membranolytic peptides and their use as molecular rulers to measure bilayer thickness in native cells, *Sci. Rep.* **2015**, *5*, 9388.

[338] Hsin, J.; Chipot, C.; Schulten, K., A glycophorin A-like framework for the dimerization of photosynthetic core complexes., *J. Am. Chem. Soc.* **2009**, *131*, 17096–8.

[339] Mingarro, I.; Elofsson, A.; von Heijne, G., Helix-helix packing in a membrane-like environment11Edited by F. E. Cohen, *J. Mol. Biol.* **1997**, *272*, 633–641.

[340] Park, S. H.; Opella, S. J., Tilt Angle of a Trans-membrane Helix is Determined by Hydrophobic Mismatch, *J. Mol. Biol.* **2005**, *350*, 310–318.

[341] Gofman, Y.; Haliloglu, T.; Ben-Tal, N., The Transmembrane Helix Tilt May Be Determined by the Balance between Precession Entropy and Lipid Perturbation, *J. Chem. Theory Comput.* **2012**, *8*, 2896–2904.

[342] Moore, R. D.; Morrill, G. A., A Possible Mechanism for Concentrating Sodium and Potassium in the Cell Nucleus, *Biophys. J.* **1976**, *16*, 527–533.

[343] Ryckaert, J.; Ciccotti, G.; Berendsen, H. J. C., Numerical Integration of the Cartesian Equations of Motion for a System with Constraints: Molecular Dynamics of n–Alkanes, *J. Comput. Phys.* **1977**, *23*, 327–341.

[344] Andersen, H. C., Rattle: A "Velocity" Version of the Shake Algorithm for Molecular Dynamics Calculations, *J. Comput. Phys.* **1983**, *52*, 24–34.

[345] Miyamoto, S.; Kollman, P. A., settle: An Analytical Version of the Shake and Rattle Algorithms for Rigid Water Models, *J. Comput. Chem.* **1992**, *13*, 952–962.

[346] Feenstra, K. A.; Hess, B.; Berendsen, H. J. C., Improving Efficiency of Large Time-scale Molecular Dynamics Simulations of Hydrogen-rich Systems, *J. Comput. Chem.* **1999**, *20*, 786–798.

[347] Hopkins, C. W.; Le Grand, S.; Walker, R. C.; Roitberg, A. E., Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning, *J. Chem. Theory Comput.* **2015**, *11*, 1864–1874.

[348] Jeckelmann, J.-M.; Lemmin, T.; Schlapschy, M.; Skerra, A.; Fotiadis, D., Structure of the human heterodimeric transporter 4F2hc-LAT2 in complex with Anticalin, an alternative binding protein for applications in single-particle cryo-EM, *Sci. Rep.* **2022**, *12*, 18269.

[349] Pang, Y. T.; Acharya, A.; Lynch, D. L.; Pavlova, A.; Gumbart, J. C., SARS-CoV-2 spike opening dynamics and energetics reveal the individual roles of glycans and their collective impact, *Commun. Biol.* **2022**, *5*, 1170.

[350] Takeda, H.; Busto, J. V.; Lindau, C.; Tsutsumi, A.; Tomii, K.; Imai, K.; Yamamori, Y.; Hirokawa, T.; Motono, C.; Ganesan, I. et al., A multipoint guidance mechanism for β-barrel folding on the SAM complex, *Nat. Struct. Mol. Biol.* **2023**, *30*, 176–187.

[351] Kalbermatter, D.; Jeckelmann, J.-M.; Wyss, M.; Shrestha, N.; Pliatsika, D.; Riedl, R.; Lemmin, T.; Plattet, P.; Fotiadis, D., Structure and supramolecular organization of the canine distemper virus attachment glycoprotein., *Proc. Natl. Acad. Sci. U.S.A.* **2023**, *120*, e2208866120.

[352] Tuckerman, M.; Berne, B. J.; Martyna, G. J., Reversible Multiple Time Scale Molecular Dynamics, *J. Chem. Phys.* **1992**, *97*, 1990–2001.

[353] Jung, J.; Kasahara, K.; Kobayashi, C.; Oshima, H.; Mori, T.; Sugita, Y., Optimized Hydrogen Mass Repartitioning Scheme Combined with Accurate Temperature/Pressure Evaluations for Thermodynamic and Kinetic Properties of Biological Systems, *J. Chem. Theory Comput.* **2021**, *17*, 5312–5321.

[354] Ferrarotti, M. J.; Bottaro, S.; Pérez-Villa, A.; Bussi, G., Accurate Multiple Time Step in Biased Molecular Simulations, *J. Chem. Theory Comput.* **2015**, *11*, 139–146.

[355] Sexton, J. C.; Weingarten, D. H., Hamiltonian Evolution for the Hybrid Monte Carlo Algorithm, *Nucl. Phys. B* **1992**, *380*, 665–677.

[356] Coutsias, E. A.; Seok, C.; Dill, K. A., Using Quaternions to Calculate RMSD, *J. Comput. Chem.* **2004**, *25*, 1849–1857.

[357] Fu, H.; Chen, H.; Cai, W.; Shao, X.; Chipot, C., BFEE2: Automated, Streamlined, and Accurate Absolute Binding Free-Energy Calculations, *J. Chem. Inf. Model.* **2021**, *61*, 2116–2123.

[358] Chipot, C.; Hénin, J., Exploring the Free-energy Landscape of a Short Peptide Using an Average Force, *J. Chem. Phys.* **2005**, *123*, 244906.

[359] Hunter, J. D., Matplotlib: A 2D Graphics Environment, *Comput. Sci. Eng.* **2007**, *9*, 90–95.

[360] Hulvej-Rod, M.; Hulvej-Rod, N., Towards a syndemic public health response to COVID-19., *Scand. J. Public Health* **2021**, *49(1)*, 14–16.

[361] Nguyen, H. L.; Thai, N. Q.; Nguyen, P. H.; Li, M. S., SARS-CoV-2 Omicron Variant Binds to Human Cells More Strongly than the Wild Type: Evidence from Molecular Dynamics Simulation, *J. Phys. Chem. B* **2022**, *126*, 4669–4678.

[362] Korber, B.; Fischer, W. M.; Gnanakaran, S.; Yoon, H.; Theiler, J.; Abfalterer, W.; Hengartner, N.; Giorgi, E. E.; Bhattacharya, T.; Foley, B. et al., Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus, *Cell* **2020**, *182(4)*, 812–827.e19.