Molecular basis of protein protein interactions during synaptogenesis

Emma Goulard Coderc de Lacam

UNIVERSITÉ DE LORRAINE

Ć2MP Doctoral School

Molecular basis of protein protein interactions during synaptogenesis

THÈSE

présenté et défendu publiquement le 20 Novembre 2023

pour obtenir le titre de

Docteur de l'Université de Lorraine

(mention Chimie théorique)

par

Emma Goulard Coderc de Lacam

Composition du jury

Président :	Serge Antonczak	Université Côte d'Azur, Nice, France
Rapporteurs :	Thomas Simonson	Ecole Polytechnique,Palaiseau, France
	Sophie Sacquin-Mora	Université Paris Cité, Paris, France
Examinateurs :	Serge Antonczak	Université Côte d'Azur, Nice, France
	Emmanuelle Bignon	Université de Lorraine, Nancy, France
Encadrant :	Christophe Chipot	

Laboratoire de Physique et Chimie Théoriques - UMR CNRS 7019

Mis en page avec la classe thesul.

Acknowledgements

Completing this doctoral journey has been a monumental endeavor, and I am deeply grateful to all those who have accompagnied me along the way. This achievement would not have been possible without the collective support of all you.

Foremost, I express my heartfelt gratitude to my supervisor, Christophe Chipot, for his unwavering guidance and invaluable mentorship. I am truly fortunate to have learn from you. François Dehez, your kindness and support were also pivotal in my PhD journey.

I am indebted to my colleagues and fellow researchers for their constant support and discussions that made me grow through this journey.

Margaret Blazhynska, I could not have wished for a better colleague and now friend to work side by side. Thank you for your constant faith in me and great scientific discussions.

Flo, thanks for sharing an office with such a great vibe whenever you are not wandering in the US. I wish you all the best for your future defense.

Haochuan Chen, you help me so much when needed, and I thank you for your kindness.

Séverine Bonneberger, you are the best secretary, always having answers to our problems super fast. I will miss our little talk at the coffee machine.

To all my fellow PhDs in the lab, past and present (Valentin, Vishal, Attila, Leila, Daniel, Andy, Rohith, Julia, Michele ...) I thank you for all the good times we shared together and the many lunches at our favorite restaurant, the CROUS!

You all contributed to making this lab a formidable place to conduct my PhD in.

I would also like to express my gratitude towards the University of Lorraine and the DREAM fellowship for sponsoring my trip to the SimBac lab in Atlanta. This trip was a formidable opportunity to discover and apprehend the research culture in the US. I really appreciate how J.C. Gumbart and his team (Katie, Gwantsa, Andrew, Yang Li, Vincent, Anna) integrated me and made me feel part of it.

Mom and Dad, I could not have done it without our support either. Your belief in my abilities and your understanding during the challenging moments have been a cornerstone of my perseverance.

Gabriel, you have been my rock in difficult times. You always knew what to say to cheer me up.

Contents

List of Figures

List of Tables

Table of Abbreviations

ACF Auto-correlation function

ABF Adaptive Biasing Force

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

SPR Surface plasmon resonance

TI Thermodynamic integration

VMD Visual molecular dynamics, University of Illinois

WTM-eABF Well-tempered metadynamics extended adaptive biasing force

DIP Dpr intercating proteins

Dpr Defective proboscis extension response

CSPs Cell-surface proteins

ML Machine Learning

LDA Linear Discriminant Analysis

RF Random Forest

Introduction (English version)

1 Synaptogenesis

To understand how neurodevelopmental and neurodegenerative diseases are working to develop an adapted treatment, it is imperative to go back to the mechanism behind the brain organization.[1,2] Synaptogenesis is the creation of synapses, which are the connections between neurons. Synapses can also be found at the neuromuscular junction. Their role is to pass a signal in the form of an action potential from one neuron to another. An action potential is a quick change in the membrane's voltage, determined by the ratio between intra- and extracellular ions. The two types of synapses are electrical and chemical, as presented in Figure 1 below.

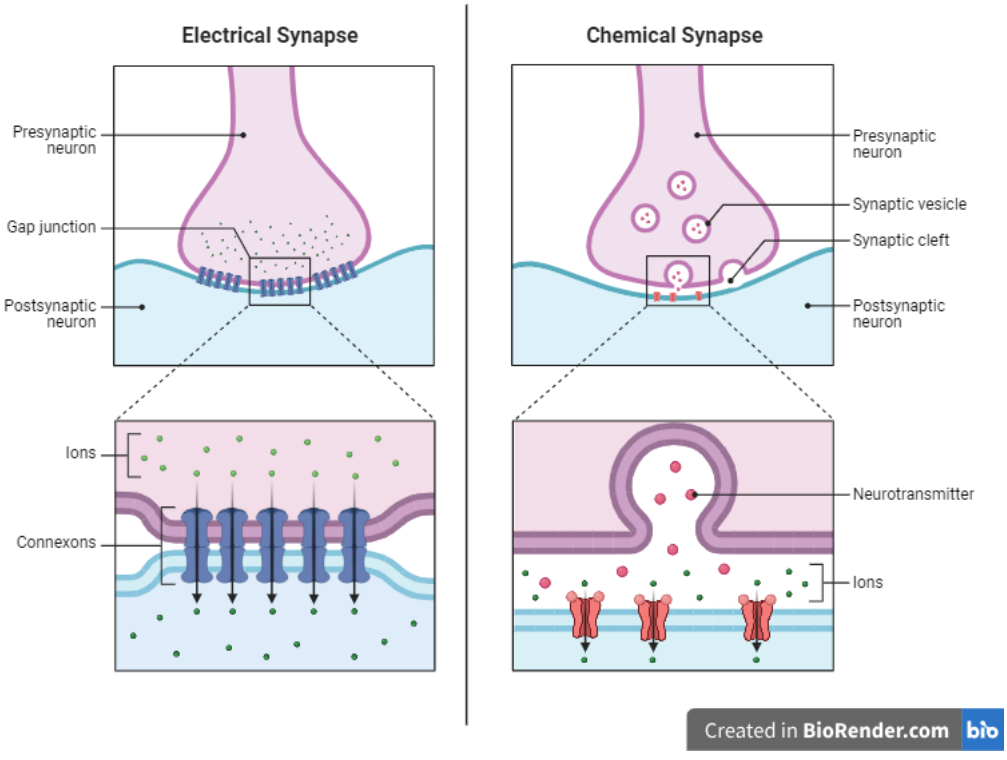

Figure 1: Schematic of a chemical and electrical synapse. Created with BioRender.com

The most well-known chemical synapse involves a particular cellular and membrane organization

triggered by upstream cellular communication. It is composed of a pre-synaptic part, specialized in Ca^{2+}-mediated release of neurotransmitters, and the post-synaptic part, which is a very dynamic membrane with many receptors dedicated to the capture of neurotransmitters (see Figure 1). These two parts are separated by an inter-cellular space called the synaptic cleft, which measures 15-20 nm.[3] The capture of the neurotransmitters triggers the opening of ion channels, which generate the voltage change and allow the action potential to pass. In contrast, the space separating the pre- and post-synaptic neurons is much smaller in the electrical synapse. The gap junction, formed by connexons, allows for a straightforward flow of ions, leading to the direct depolarization of the membrane and the propagation of the action potential.

In 1963, Roger Sperry postulated the chemoaffinity hypothesis: "Neurons make connections relying on the specific interactions of cell-surface proteins (CSPs) based on their affinity."[4] Therefore, synaptogenesis should be considered solely as an affinity problem. Ozkan et al. identified these CSPs in the model organism of a fruit fly, *Drosophila Melanogaster*.[5] These CSPs correspond to two families called the Defective proboscis extension response (Dpr) and the Dpr Interacting Proteins (DIP), composed of 21 and 11 members, respectively. The Dpr and DIP proteins belong to the immunoglobin super-family composed of three and two globular domains, respectively. They interact through their first globular domain only (see Figure 2). The similarity inside every family is very high, which can be explained by the strong likelihood of evolutionary duplication events in CSPs.[6,7] Additionally, each member of the family bears the canonical binding interface.[6]

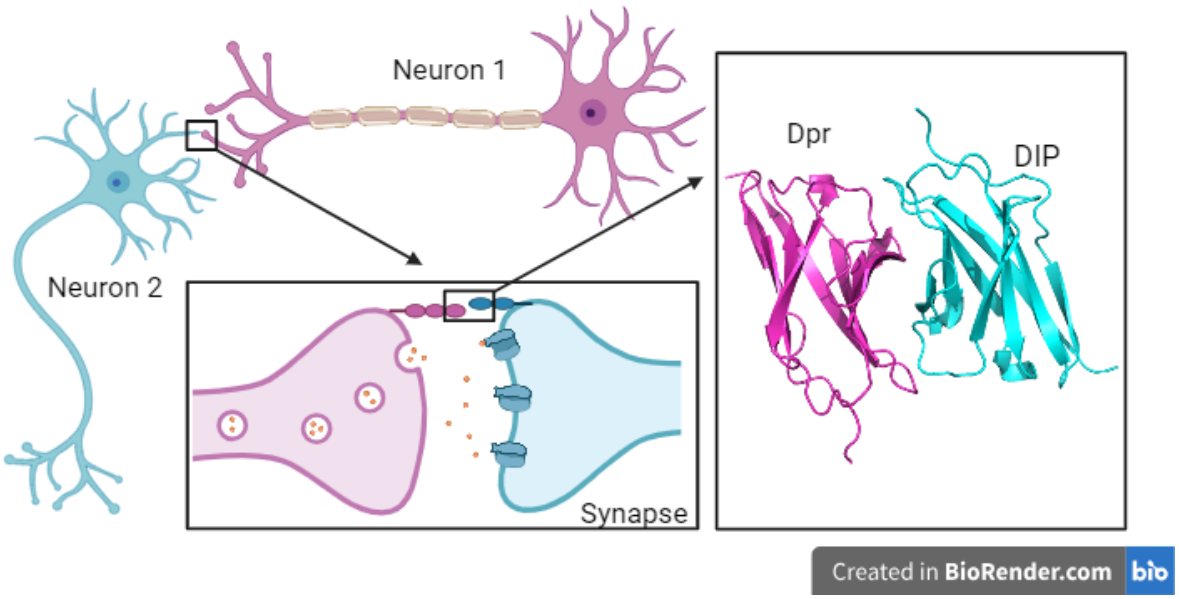

Figure 2: Schematic of neurons and a chemical synapse with the first domain of DIP/Dpr proteins. Created with BioRender.com

Out of the 231 crossing possibilities, only 57 complexes were detected below the threshold of -5.1 kcal/mol, meaning their binding is strong enough to lead to a synapse in downstream events.[8,9] Additional phylogenetic studies and similarity led to the following interactome depicted by Figure 3[9]

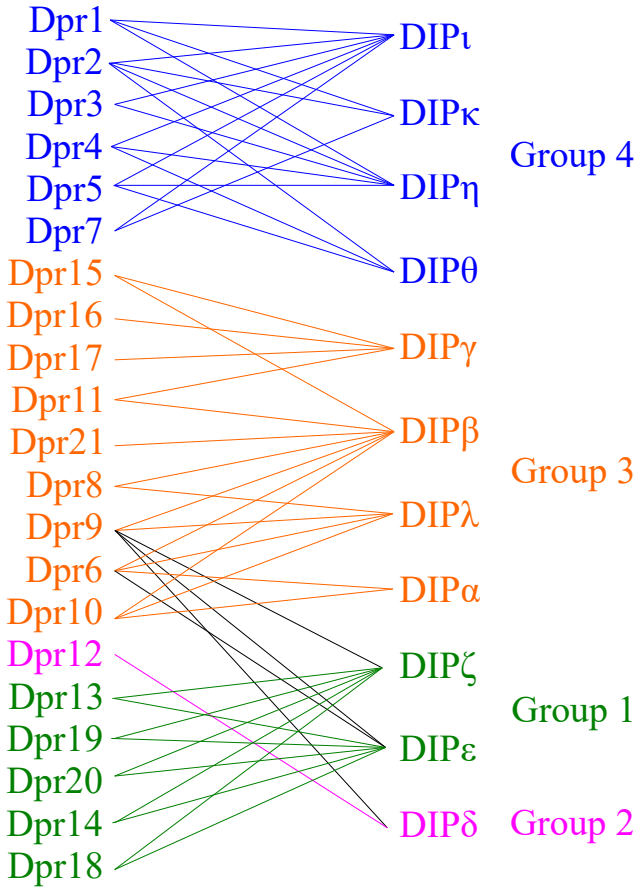

Figure 3: DIP-DPR interactome using data from reference[8]

However, the absolute binding free energy remains the only definitive criterion to classify a complex into cognate (strong binders) and non-cognate (weak binders).

The objective of my Ph.D. is, therefore, to develop a robust methodology able to predict the affinity of the two proteins and determine if they interact together while reasoning about the molecular basis of the interaction in the context of Dpr-DIP interactome.

2 Assessing binding affinity between proteins

The absolute binding free energy of a protein complex reflects the stability and the diverse interactions of a protein with another quantitatively, allowing for a direct comparison between complexes. Computing the affinity requires the structure of the bound complex, which several methods, such as homology modeling or docking, can provide. Homology modeling allows solving the structure of complex or individual partners using another 3D structure referred to as the template, given that they share a sequence with a similarity score superior to 30% to guarantee an identical fold. Docking, however, is used to solve the bound form of the complex using partners with isolated structures. It uses complementarity as the first criterion to place the partners before using pre-establish potentials to score the pose and provide an estimate of the binding affinity. Docking usually involves treating the partners as rigid entities, neglecting the conformational changes linked to binding. Furthermore, the accuracy greatly depends on the scoring function.[10]

In our case with the Dpr-DIP interactome, some complexes have been resolved using X-ray diffrac-

tion[8] and can serve as templates given the high amino acid similarity within the two protein families. Nandigrami et al.[11] successfully generated the 231 Dpr-DIP hetero dimers full atoms models using homology modeling and further refined the structures by a 200 ns molecular dynamics equilibrium simulation. The access to the structure enables the use of techniques leveraging force field and simulation-based methods.

In principle, the binding of proteins could be observed in long enough brute-force MD simulations. The range of protein dissociation required to extract the binding free energy is within minutes, which is currently out of the scope of MD simulations.[12] Endpoint methods such as molecular mechanics - Poisson Boltzmann surface area (MM-PBSA) can be used to alleviate this limitation.[13,14,15,16,17] The binding affinity is computed here as the sum of its gas-phase energy (MM), the solvation free-energy (PBSA), and a contribution due to the configurational entropy of the solute extracted from molecular dynamics (MD) simulations. Despite its popularity,[18] the energies obtained can show significant errors on account of the numerous approximations (implicit solvation, constant dielectric medium, numerical subtraction of large mean quantities to obtain small ones).[12,19,20] Alternatively, steered molecular dynamics (SMD), where one partner is pulled away from the binding site in non-equilibrium simulations, can be employed. The method relies upon the idea that the non-equilibrium work to dissociate the partners reflects the strength of their interaction. Therefore, application of the Jarzynski identity[21] would allow recovery of the binding affinity. In practice, multiple realizations of binding and unbinding in a near-equilibrium regime are required to obtain the desired quantities..[22] However, the pulling pathway, as well as the magnitude of the force, can drastically affect the convergence and prevent an accurate reproduction of the binding affinity.[23,24]The reversible association of a protein-protein complex is characterized by large configurational changes upon binding, encumbering the convergence of simulations. The idea emerged to restrain the movement of the partners while estimating the cost of those restraints prior to separating the proteins to improve the convergence. Woo and Roux proposed the introduction of restraints on a set of collective variables (CVs) representing the slow degrees of freedom of the binding partner to limit its movement and ensure converged configurational ensembles. This approach was further implemented in the so-called geometrical and alchemical routes,[25] both detailed in chapter 1 and used in this work. This restraint-based approach inspired other methods such as attach-pull-release[26] and funnel metadynamics[27,28]

Alternative approaches using machine learning (ML) to predict or classify binding partners at a lower computational cost have emerged in the last decade and are mainly directed at protein-ligand complexes due to the interest of the pharmaceutical industry in the drug design field and the numerous descriptors available for ligand properties.[29,30,31,32,33,34] However, in rare cases, ML has been used to predict protein-protein binding free energies. Moal et al. have successfully developed molecular descriptors and used four ML algorithms (Random Forest,[35] M5' Regression tree,[36] Multivariate Adaptive Regression Splines,[37] and Radial Basis Function Interpolation[38]) to predict the binding affinity for a subset of 57 complexes.[39] They obtained a correlation with leave-one-out validation of 0.70, 0.69, 0.68, and 0.75, respectively. However, the subset of descriptors affect significantly the quality of the models. Vangone et al. designed a model based on the number of interacting contacts and the non-interacting surface. They obtained a high accuracy (R = -0.73, $\rho < 0.0001$; RMSE = 1.89 kcal/mol^{-1}), using a dataset of 81 complexes..[40] These approaches are very limited and only trained on a minimal amount of complexes, which can be explained by difficulties in obtaining both accurate structural and thermodynamic data. Many approaches are dedicated to interaction predictions rather than predicting the binding free energy or classifying it into binding or non-binding complexes.[41,42]Pavlova et al. used ML to treat MD data to determine residue mutations to strengthen the binding affinity of the SARS-CoV-2 virus to the human receptor ACE2 before appearing in dangerous variants, e.g., N501Y..[43]

3 Outline

To predict the binding free energy of protein-protein complexes, I first participated in elaborating a detailed protocol of our methods using the alchemical and geometrical routes, whose theoretical background is detailed in chapter 1.[25] My contribution is presented in Chapter 3 and was published in the following article:

"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. *J. Chem. Theory Comput.*, Doi.org/10.1038/s41596-021-00676-1."

This protocol, however, only referred to protein-ligand complexes, and we needed to expand it to protein-protein complexes. At times, the COVID syndemic was still ongoing, and we decided to contribute to the research effort on the virus by computing and assessing the molecular basis of the affinity from the diverse variants of concerns (VOCs). This work is reported in chapter 4 and can be found under the following reference :

"**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 "

In searching for accurate COVID data, we discovered that the reported binding estimates span a wide range of kcal/mol, primarily due to poor structural data and needed to fast-track the research due to public pressure. In that spirit, many methodological shortcuts appeared in the literature and were employed. We demonstrated in the publication below and Chapter 5 that these shortcuts do not produce accurate and reproducible estimates contrary to the rigorous approach of the geometrical route.

"Blazhynska, M., **Goulard Coderc de Lacam, E.**, Chen, H., Roux, B. Chipot, C. Hazardous shortcuts in standard binding–free energy calculations, *J. Phys. Chem.*, 2022, 13, 27, 6250–6258, Doi: 10.1021/acs.jpclett.2c01490"

The geometric route remains criticized on account of the numerous long-biased simulations needed to obtain the binding affinity estimate. We searched for an alternative to speed up those calculations. Multiple-time stepping was recently made available and tested with the HMR trick on the ABlSH3-p41 and the MDM2-p53:NVP-CGM097 complexes. The paper resulting from this work is presented in chapter 6 and is available under:

"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"

Having gained experience in binding free-energy computation for protein-protein, I used the geometrical route on the target system of my PhD, the Dpr-DIP interactome. However, treating the complete dataset in a reasonable time and bereft of any human intervention was impossible with the geometrical route. Therefore, I relied on ML approaches (theoretical basis in chapter 2) to classify the complexes into cognate and non-cognate partners while reasoning about the molecular basis of their interactions. The resulting work is presented in chapter 7 and is currently under review at J. Chem. Inf. Model.

Introduction (version française)

1 La synaptogenèse

Comprendre le fonctionnement des maladies neurodégénératives et des troubles de neurodéveloppement, dans l'espoir de trouver des traitements adaptés, implique d'élucider les mécanismes responsables de l'organisation du tissu nerveux.[1,2] La synaptogenèse correspond à la formation des synapses qui sont les connexions entre les neurones. Certaines synapses se trouvent également à la jonction neuro-musculaire. Elles permettent de passer un signal sous la forme d'un potentiel d'action d'un neurone à un autre. On distingue deux types de synapses, électriques et chimiques représentés dans la Figure 1 ci-dessous.

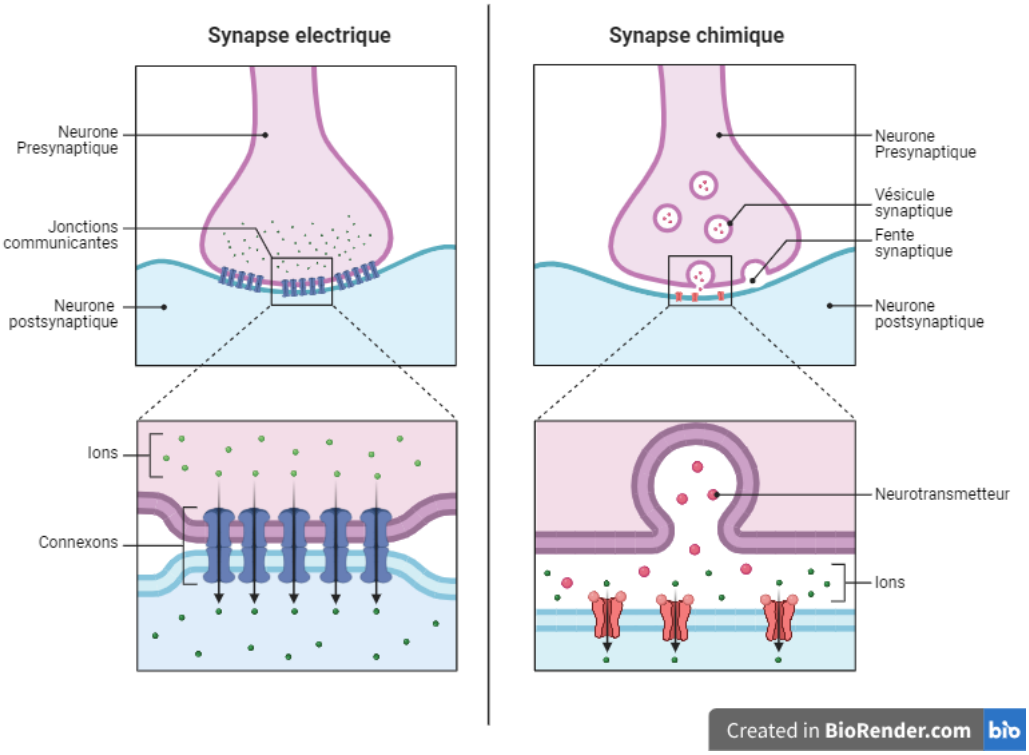

Figure 1: Schéma d'une synapse chimique et électrique. Crée avec BioRender.com

La synapse chimique est la plus connue et se compose d'une organisation cellulaire et membranaire très spécifique déclenchée par des communications cellulaires en ammont. Elle se compose d'une partie

pré-synaptique appartenant à un premier neurone, spécialisée dans la libération des neurotransmetteurs en réponse à un afflux d'ions calcium, Ca^{2+} (voir Figure 1). La partie post-synaptique, appartenant au second neurone, est une membrane très dynamique remplie de récepteurs dédiés à la capture des neurotransmetteurs. Les neurones sont séparés par un espace intercellulaire appelé la fente synaptique mesurant entre 15 et 20 nm..[3]La capture des neurotransmetteurs déclenche l'ouverture de canaux ioniques qui permettent le changement de voltage nécessaire pour générer le potentiel d'action. Pour la synapse électrique, l'espace entre les neurones est très réduit. Les jonctions communicantes au niveau des membranes permettent un flux direct d'ions d'une cellule à l'autre ce qui permet de propager la dépolarisation membranaire et le potentiel d'action.

L'hypothèse de la chemoaffinité énoncée par Roger Sperry en 1963, stipule que : "les neurones réalisent des connexions reposant sur les interactions spécifiques de protéines cellulaires de surface basées sur leur affinité."[4] La synaptogenèse est donc un problème d'affinité. Chez l'organisme modèle la mouche, *Drosophila Melanogaster*, ces protéines de surface ont été identifiées.[5] Il s'agit de deux familles de protéines appelées Defective proboscis extension response (Dpr) et Dpr Interacting Proteins (DIP) composées de 21 et 11 membres, respectivement.

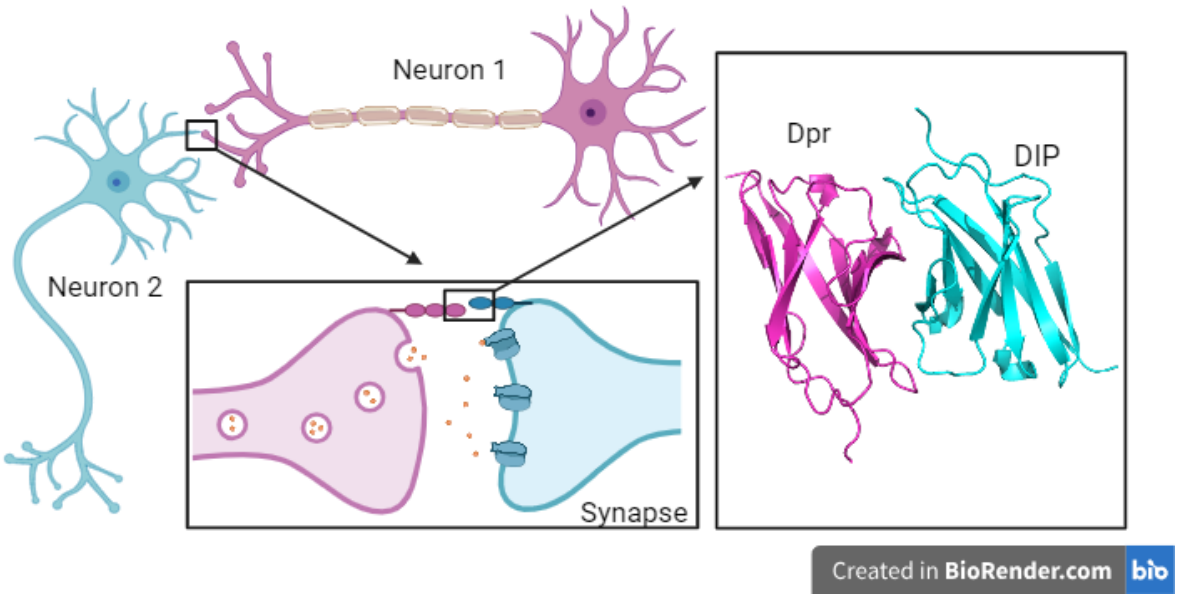

Figure 2: Schema de neurones et d'une synapse chimique avec le premier domaine d'un complex DIP/Dpr. Crée avec BioRender.com

Parmi les 231 croisements possibles, seul un petit nombre de complexes, 57, ont été détectés en dessous du seuil de $-$ 5.1 kcal/mole, ce qui signifie que leur liaison est suffisamment forte pour déclencher l'apparition future d'une synapse.[8,9] Des études supplémentaires phylogénétiques et la similarité ont conduit à l'interactome Dpr-DIP représenté dans la Figure 3.[9]

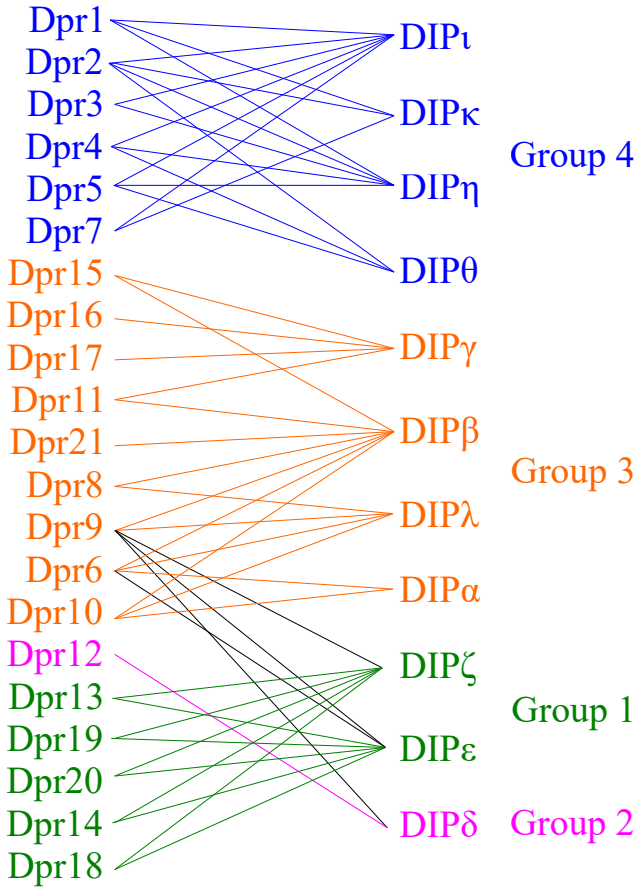

Figure 3: Dpr-DIP interactome avec les données de[8]

L'objectif de ma thèse est donc de trouver une méthode fiable capable de prédire l'affinité de deux protéines pour déterminer si elles interagissent entre elles tout en résonnant sur les bases moléculaires responsables de cette interaction dans le cas de l'interatome Dpr-DIP.

2 Calcul de l'affinité proteine-protéine

L'énergie libre de liaison d'une protéine représente quantitativement la stabilité et les diverses interactions avec son partenaire, ce qui permet une comparaison directe entre complexes. Une structure du complexe formé par deux protéines est requise pour calculer leur affinité de liaison. Différentes méthodes telles que la modélisation par homologie et le docking permettent de générer ces structures en absence de données structurales expérimentales L'homologie permet de générer une structure à partir d'une structure de référence appelée template si ces structures partagent une séquence similaire à plus de 30%, ce qui garantit un repliement identique. Cette méthode peut s'appliquer aux complexes et aux protéines isolées. Au contraire, le docking permet d'obtenir la structure d'un complexe à partir de celle des protéines isolées. La complémentarité est le critère principal pour placer les protéines. Un potentiel prédéfini est ensuite appliqué pour estimer le score de la pose et donc l'affinité de liaison. Le docking traite souvent les entités moléculaires comme des objets rigides, négligeant les changements conformationnels liés à leur liaison. De plus, la précision de l'estimation dépend grandement de la fonction de score employée.[10]

Dans le cas de l'interactome Dpr-DIP, quelques complexes ont une structure résolue expérimentale-

ment par diffraction aux rayons X et peuvent servir de templates grâce à la forte similarité de séquences au sein des deux familles. Nandigrami et al.[11] ont généré les modèles de tous les complexes (231) à l'aide de la modélisation par homologie. Ces modèles ont ensuite été raffinés par une simulation à l'équilibre de 200 ns. L'accès à l'ensemble de ces structures permet l'utilisation de techniques se reposant sur la simulation numérique et les champs de forces (voir Chapitre 1)

La liaison de protéines peut, en principe, être observée dans des simulations suffisamment longues. Cependant, le temps caractéristique pour observer la dissociation de deux protéines se situe dans les minutes, actuellement hors de la portée de la simulation par dynamique moléculaire.[12] La technique MMPBSA permet de contourner cette limitation..[13,14,15,16,17] L'affinité de liaison y est calculée par la somme de son énergie en phase gazeuse (MM), l'énergie de solvatation (PBSA) en une contribution pour l'entropie configurationnelle du soluté extraite de la simulation (MD). Malgré sa popularité,,[18] les énergies de liaison obtenues peuvent présenter des erreurs dues aux approximations de la méthode (solvatation implicite, constantes diélectriques, extraction de petites quantités à partir de quantités plus larges)..[12,19,20] Une alternative est la *steered molecular dynamics* (SMD), où un des partenaires est extrait du site de liaison dans une simulation hors équilibre. L'idée qui soutient cette méthode est que le travail nécessaire à l'extraction reflète la force de l'interaction entre les partenaires. En appliquant l'équation de Jarzynski,[21] on peut donc théoriquement retrouver l'énergie libre de liaison. Cependant, en pratique, il faut de nombreuses réalisations dans un régime de quasi-équilibre pour obtenir les quantités désirées De plus le chemin choisi pour l'extraction, qui ne correspond pas nécessairement au chemin réel, ainsi que l'ampleur de la force exercée peuvent affecter grandement la reproduction de l'énergie libre de liaison.[23,24]

L'association réversible d'un complexe protéine-protéine se caractérise par de larges changements configurationnels lors de la liaison, ralentissant la convergence de la simulation. Pour résoudre ce souci, l'idée de restreindre les mouvements des partenaires tout en évaluant le coût des restreints a émergé. Woo et Roux[44] ont proposé d'introduire une restreinte harmonique sur un ensemble de variables collectives représentant les dégrées de liberté lents du ligand pour limiter son mouvement et garantir des ensembles configurationnels convergés. Cette approche a été implémentée dans la "route géométrique" et la "route alchimique" par Gumbart et al.[25] Ces deux routes, ainsi que plus d'informations sur le calcul d'énergie libre par dynamique moléculaire sont présentées dans le chapitre 1.

Avec l'avancement et la généralisation de l'apprentissage automatique, des méthodes ont émergées pour prédire et classifier des complexes avec un coût informatique plus faible. Cependant, elles sont majoritairement dédiées aux complexes protéine-ligand, dues à l'intérêt porté l'industrie pharmaceutique, le drug design et l'existence de nombreux descripteurs pour les ligands.[29,30,31,32,33,34] Quelques rares applications concernent la prédiction de l'affinité protéine-protéine. Moal et al.[39] ont développé des descripteurs moléculaires et utilisé quatre algorithmes (Random Forest,[35] M5' Regression tree,[36] Multivariate Adaptive Regression Splines,[37] and Radial Basis Function Interpolation[38]) pour prédire l'affinité de 57 complexes. Ils ont obtenu des corrélations avec une validation par leave-one-out de 0.70, 0.69, 0.68, et 0.75, respectivement. Cependant, le choix des descripteurs affecte grandement les corrélations obtenues. Vangone et al. ont créé un modèle basé sur le nombre de contacts en interaction et une contribution de la surface non en interaction. Ils ont obtenu une corrélation forte ($R = -0.73$, $\rho < 0.0001$; RMSE = 1.89 kcal/mol^{-1}), sur un ensemble de 81 complexes..[40] Ces approches sont limitées et seulement entraînées sur un petit ensemble de structures, ce qui s'explique par la difficulté d'obtenir des données qualitatives à la fois structurelles et thermodynamiques.

La majorité des applications se tournent vers la prédiction des interactions entre protéines.[41,42] Pavlova et al. ont utilisé l'apprentissage automatique pour prédire les mutations les plus favorables pour augmenter l'affinité de SARS-CoV-2 à son récepteur humain et pu prédire certaines qui ont été retrouvées dans des variants plus tardifs, e. g. N501Y.[43]

3 Plan

Pour prédire l'énergie libre de liaison de complexes, j'ai participé premièrement à l'élaboration d'un protocole détaillé pour réaliser ces calculs à l'aide des routes alchimique et géométrique. Les fondations théoriques de ces méthodes sont détaillées dans le chapitre **??**. Ce travail a débouché par la publication suivante et est présenté dans le chapitre 3 : "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. *J. Chem. Theory Comput.*, Doi.org/10.1038/s41596-021-00676-1"

Cependant, ce protocole n'est destiné qu'aux complexes protéines-ligands. Lors de la syndémie de COVID-19, nous avons décidé de contribuer à l'effort de recherche sur ce virus en calculant l'affinité des variants préoccupants en étendant la route géométrique et notre protocole aux complexes protéines-protéines. Les résultats de cette investigation sont publiés dans l'article suivant et se trouvent dans le chapitre 4:

"**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 "

Au cours de nos recherches dans le cadre de la syndémie de COVID, nous avons découvert des estimations de l'affinité de liaison entre SARS-CoV-2 et le RBD dans un large intervalle de kcal/mol, principalement expliqué par le peu de données structurales et la nécessité d'obtenir des résultats fiables et rapides sous la pression de l'opinion publique. Dans cet esprit, des raccourcis méthodologiques ont vu le jour au sein de la littérature. Nous avons démontré dans l'article ci-dessous, et qui constitue le chapitre 5, que ces raccourcis ne permettent pas d'obtenir des estimations précises et reproductibles contrairement à notre approche rigoureuse avec la route géométrique.

"Blazhynska, M., **Goulard Coderc de Lacam, E.**, Chen, H., Roux, B. Chipot, C. Hazardous shortcuts in standard binding–free energy calculations, *J. Phys. Chem.*, 2022, 13, 27, 6250–6258, Doi: 10.1021/acs.jpclett.2c01490"

La route géométrique est critiquée au vu des nombreuses simulations biaisées nécessaires à l'obtention de l'affinité de liaison. Nous avons donc cherché une alternative à ces raccourcis pour accélérer ces calculs. Nous avons utilisé une combinaison entre le MTS et HMR sur les complexes ABlSH3-p41 et MDM2-p53-NVP-CGM097. Le Chapitre 6 présente ses résultats, également disponibles dans la publication suivante :

"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"

Avec toute cette expérience acquise dans le domaine du calcul de l'énergie de liaison et leur base moléculaire, j'ai appliqué cette méthode de la route géométrique aux complexes Dpr-DIP ainsi qu'une approche alternative utilisant l'apprentissage automatique, (fondation théorique dans le chapitre 2) pour distinguer les complexes cognate des non-cognates. Cette investigation est présentée dans le chapitre 7 et est actuellement en révision chez J. Chem. Inf. Model.

Chapter 1

Methods for binding free energies calculations using molecular dynamics simulations

This chapter recaps the theoretical underpinnings of standard binding free-energy calculations using molecular dynamics simulations. We will first discuss the definition of molecular dynamics (MD) and the force field. Then, we will move to enhanced sampling with biased MD before diving into the standard binding free-energy calculation methods employed throughout this thesis.

1.1 Classical molecular dynamics simulations

1.1.1 Definition

Molecular dynamics monitors the time evolution of a molecular system, generating trajectories (positions and velocities) of the ensemble of particles by numerically integrating Newton's equation of motion. For each particle i, we can write the force ($\vec{F_i}$) as the mass (m_i) multiplied by the acceleration, which is the second derivative of the position (r_i) on time.

$$m_i \frac{\mathrm{d}^2 \vec{r_i}}{\mathrm{d}t^2} = \vec{F_i} \tag{1.1}$$

By deriving the force $\vec{F}$, we gain access to the potential energy (U) of the system for each component:

$$\vec{F_i} = -\Delta U_i \tag{1.2}$$

The potential energy function is derived from the force field, which will be described in the next section. To solve these equations for each particle, we resolve to a numerical way using a finite method. This method relies on the knowledge of the position, velocities, and acceleration of every particle at a given time t to estimate the values of those same variables at time $t + \delta t$, where δt stands for a tiny time interval referred to as the timestep. The timestep used in the simulation has to match the fastest motion of the system, e.g., the vibration of the bond between the hydrogen and carbon atoms, to guarantee energy conservation. The use of algorithms such as RATTLE[45] and SHAKE[46] to constrain this bond allows for the use of a time step of 2 femtoseconds instead of 1. Another option to get a larger timestep is to use Hydrogen Mass Repartitioning (HMR),[47] where the mass of the hydrogen atoms is artificially increased by transferring some out of the bonded hetero atoms, thus conserving the total mass of the system.

1.1.2 Force field

The force field is an approximation used to compute the potential energy between particles. It can be decomposed into several terms: stretching, bending, torsion (bonded terms), Van der Walls and electrostatic (Non-bonded terms), presented below.

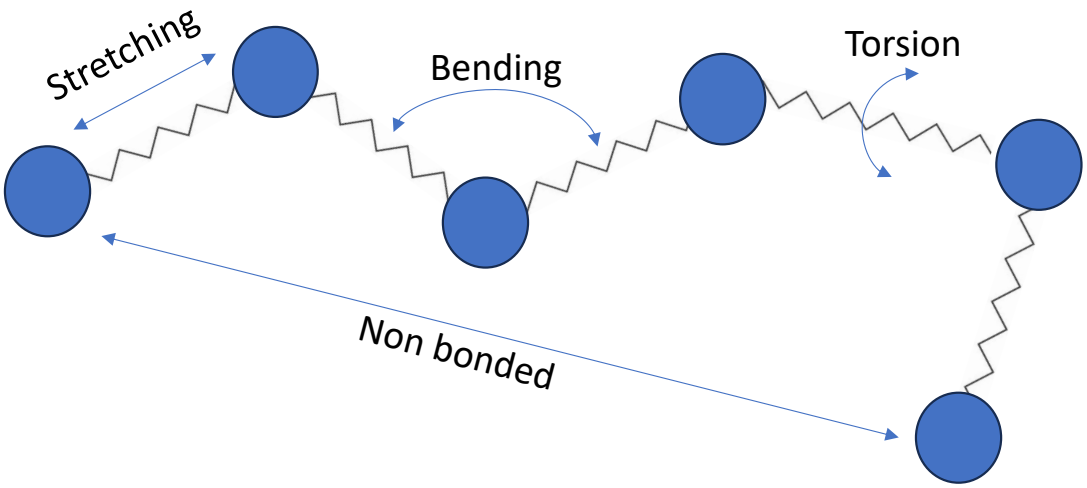

Figure 1.1: Schematic representation of a system composed of atoms (blue spheres) with the bond represented by a spring with the different types of interactions in a force field

Streching term

Let us consider a diatomic system. When attempting to bring the two atoms closer together, the protective repulsive shell generated by the electronic cloud will increase the energy to prevent steric clashes. Conversely, pushing the atoms further apart will also increase energy proportionally to the distance between atoms until the bond breaks when the dissociation energy (D_e) is reached. This potential energy can be described using a Morse potential :

$$U(r) = D_e \{1 - exp[-a(r - r_0)]\}^2 \tag{1.3}$$

with

$$a = \sqrt{\frac{k}{2D_e}} \tag{1.4}$$

, k corresponding to the stiffness of the spring, and r_0 the natural bond length at equilibrium.

However, evaluating the exponential part of eq 1.3 can be computationally expensive, especially when considering many pairs. Since we work with systems at equilibrium, the relevant region is close to the Morse potential minima, which a harmonic potential can approach:

$$U_{streching} = k(r - r_0)^2 \tag{1.5}$$

Bending term

The bending term represents the bending angle between three bonded atoms. It has a similar behavior to the stretching term and can be described by a harmonic approximation.

$$U_{bending} = k(\theta - \theta_0)^2 \tag{1.6}$$

Torsion term

The torsion term is linked to the potential rotation around the middle bond in a system composed of four bonded atoms. It is a periodic function due to its angular nature and returns to its original position after 360 °. This term is expressed as:

$$U_{torsion} = \sum V_n \cos(n\omega) \tag{1.7}$$

with V_n the height of the energy barrier of order n and ω the angle value.

Non bonded interations

Non-bonded interactions occur between atoms that are either not directly bonded to each other or separated by more than three bonds. They are composed of Van der Walls and electrostatics terms.

Van der Waals term The Van der Waals term accounts for the interaction between non-charged atoms and is usually represented by a Lennard-Jones potential:

$$U_{vdW} = \sum_i \sum_{j \neq i} 4\pi\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{1.8}$$

, where r_{ij} is the distance between atoms i and j. σ and ε are both parameters specific to the chosen force field representing the minimum distance and the specific dielectric constant for the atoms pair, respectively.

Electrostatic term Electrostatic interactions involve charged atoms and are included in the force field and follow Coulomb's law :

$$U_{elec} = \sum_i \sum_{j \neq i} \frac{q_i q_j}{R_{ij}} \tag{1.9}$$

where q_i and q_j are partial charge associated with atoms i and j ,respectively, and R_ij is the distance between the two atoms.

For all the simulations reported in this thesis, we chose the CHARMM36[48] force field as it is an all-atom and robust force field parameterized for proteins.

Obtaining binding-free energies requires long simulations to observe binding and unbinding events, which are, in the timescale of minutes, currently out of reach for classical MD.[12] To favor the appearance of binding and unbinding events, we need to enhance the exploration of the relevant regions of the free energy landscape by biasing the MD. These methods are called Enhanced sampling. The next section will present the Well-tempered metadynamics extended Adaptive Biasing Force (WTM-eABF) method, which we selected for the studies reported in this manuscript.

1.2 Biased MD and Enhanced sampling with the well-tempered metadynamics extended ABF

The WTM-eABF[49] combines a well-tempered metadynamics[50] and extended ABF.[51] We will briefly go through the principle of metadynamics and ABF before delving into WTM-eABF.

1.2.1 Metadynamics.

In metadynamics, the goal is to boost the sampling of the Potential Energy Surface (PES) through a collective variable (CV). A CV is a variable that can be monitored and controlled during MD. The choice of an appropriate CV is crucial as it must describe the system at a lower complexity and be able to establish a difference between the initial state A and the target state B. The biasing potential, here, comprises a series of Gaussian functions deposited over time in the visited states. Each Gaussian function adds a repulsive contribution to the potential energy, effectively flattening the free energy landscape and reducing energy barriers, thus encouraging the system to explore a broader range of conformations and facilitating transitions between different states (see Figure 1.2).[52]

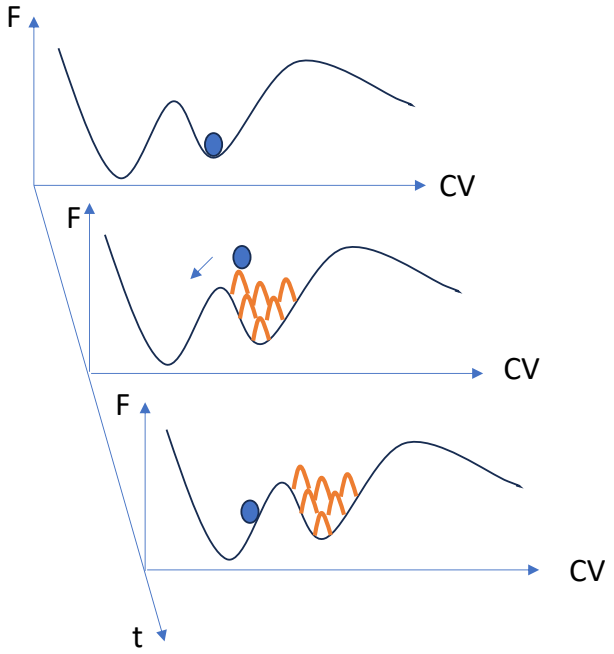

Figure 1.2: Schematic representation of metadynamics

Metadynamics can be dampened by many factors, particularly the height of the Gaussian. A high Gaussian can favor a rapid visit of a region of interest but can lead to significant errors in computing the PES. Controversially, a lower Gaussian will ensure less error in the estimation of the PES at the cost of an increased time to visit regions of the PES.[53] In the well-tempered variant, the controlled temperature fluctuation facilitates barrier crossing. It prevents going into regions with an energy too high to be of interest, limiting the risk of overfilling and ensuring maximum optimization of the computational time.[50]

1.2.2 Adaptive Biasing Force

Adaptive Biasing Force (ABF), as stated by its name, relies on applying an adapted biasing force throughout the simulation and aims at improving the efficiency of sampling of the PES when the system is constrained by high free-energy barriers preventing ergodic sampling.

One advantage of the method is that it does not require prior knowledge of the system since the bias is applied once enough samples have been gathered to estimate a mean force. Furthermore, the process is done adaptively with a running average of the instantaneous force acting along the coordinate to obtain an on-the-fly estimate. Concomitantly, an external potential is applied, canceling the average estimated force and allowing convergence toward a flat potential of mean force, which can be thoroughly and quickly explored.

Mathematically, standard ABF can be defined by the following equations:

$$\frac{\mathrm{d}A}{\mathrm{d}\xi} = -\left\langle F_\xi \right\rangle_\xi = \left\langle \frac{\partial U(\mathbf{x})}{\partial \xi} \right\rangle_\xi - \left\langle \frac{1}{\beta} \frac{\partial \ln \left[\mathbf{J}\right]}{\partial \xi} \right\rangle \tag{1.10}$$

$$\mathbf{F}_{\text{bias}} = -\left\langle F_\xi \right\rangle_\xi \nabla_{\mathbf{x}} \xi \tag{1.11}$$

where $\frac{\mathrm{d}A}{\mathrm{d}\xi}$ is the free energy derivative, $|\mathbf{J}|$ corresponds to the determinant of Jacobian when going from generalized to Cartesian coordinates, $\mathbf{F}_{\text{bias}}$ is the biasing force acting on the system.

1.2.3 Well tempered metadynamics extended ABF

Extended ABF implicates using an extended degree of freedom to transition coordinates, rendering the CV independent and removing the need to explicitly determine $|\mathbf{J}|$.Therefore, the free energy can be estimated simply with the corrected z-averaged restraint (CZAR) estimator[54] or umbrella integration (UI). The well-tempered metadynamics variant solves the metadynamics issue with convergence and focuses solely on relevant free energy surface regions, as discussed above. ABF flattens the free energy landscape, while metadynamics allows sampling of less frequently visited states, making both approaches complementary.[55] The biasing potential in WTM-eABF is:

$$-\frac{1}{\beta\sigma^2}\left(\langle \xi(r) - \lambda \rangle_\lambda\right) + \frac{\partial}{\partial \lambda} h_0 \exp\left(-\frac{V_m\left(\lambda, t\right)}{k_B \Delta T}\right) V_m\left(\lambda, t\right) \tag{1.12}$$

where $V_m(\lambda, t)$ is a growing biasing potential with respect to stimulation time t, h_0 is the height of the Gaussian functions for the metadynamics part, and β is the inverse of the Boltzmann constant multiplied by the system temperature, T. σ is the standard deviation between $\xi(r)$ and λ. Thus, the equations of motion are modified to the following:

$$\begin{cases} U(\mathbf{x}, \lambda) & = U_{\text{FF}}(\mathbf{x}) + \dfrac{1}{2} k \big(\xi(\mathbf{x}) - \lambda\big)^2 + U_{\text{bias}}(\lambda) \\ \mathbf{m}_{\mathbf{x}} \ddot{\mathbf{x}} & = -\nabla_{\mathbf{x}} U(\mathbf{x}, \lambda) - \gamma_{\mathbf{x}} \mathbf{m}_{\mathbf{x}} \dot{\mathbf{x}} + \sqrt{2\gamma_{\mathbf{x}} \frac{1}{\beta}}\, \mathbf{m}_{\mathbf{x}}^{1/2}\, \dot{\mathbf{W}}(t) \\ m_\lambda \ddot{\lambda} & = -\nabla_\lambda U(\mathbf{x}, \lambda) - \gamma_\lambda m_\lambda \dot{\lambda} + \sqrt{2\gamma_\lambda \frac{1}{\beta}}\, m_\lambda^{1/2}\, \dot{\mathbf{W}}(t) \end{cases} \tag{1.13}$$

where $\xi(\mathbf{x})$ corresponds to the real CV. The extended degree of freedom, λ, represents the fictitious particle. The spring connecting the real CV to the fictitious particle is characterized by a force constant k and has to be strong enough to ensure proper coupling. 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.

Two additional parameters can affect the coupling and have been explored in detail in Chapter 6. They are the oscillation period (τ) and the coupling width (σ). They affect the mass of the fictitious particle and the spring constant as per:

$$\begin{cases} m_\lambda &= -\frac{1}{\beta}\left(\frac{\tau}{2\pi\sigma}\right)^2 \\ k &= \frac{1}{\beta\sigma^2} \end{cases} \tag{1.14}$$

In this work, we choose to use the CZAR estimator to estimate the free energy gradients from the WTM-eABF using:

$$A'(\xi) = -\frac{1}{\beta}\frac{\mathrm{d}(\ln\tilde{\rho}(\xi))}{\mathrm{d}\xi} + k\left(\langle\lambda - \xi\rangle_\xi\right) \tag{1.15}$$

where $\tilde{\rho}(\xi)$ is the biased distribution of ξ, λ is a value of the fictitious particle, k is the corresponding spring force constant, and $\langle\rangle_{\boldsymbol{\xi}}$ is the ensemble average. Once the gradients are estimated, the PMF is integrated numerically with standard numerical integration techniques, such as the trapezoidal rule or Simpson's rule, for a unique CV.

1.3 Binding free-energy calculations theoretical background

The reversible association of a protein with a ligand can be explained by the following equation:

$$\mathrm{Protein} + \mathrm{Ligand} \rightleftharpoons \mathrm{Protein} : \mathrm{Ligand} \tag{1.16}$$

where Protein:Ligand is the bound form of the complex. The equilibrium association constant K_{eq} can be defined as the ratio between the concentration of the free and the bound entities.

$$K_{_{\mathrm{eq}}} = \frac{[\mathrm{protein} : \mathrm{ligand}]}{[\mathrm{protein}][\mathrm{ligand}]} \tag{1.17}$$

Let us assume that the probability of having no ligand bound to the protein is p_0 and that p_1 is the probability that at least one ligand is bound to the protein. The equilibrium association constant can now be seen as:

$$K_{_{\mathrm{eq}}} = \frac{p_1}{p_0}\frac{1}{[\mathrm{ligand}]} \tag{1.18}$$

If ligand binding modes are consistent and non-cooperative, the logarithmic ratio p_1/p_0 can be expressed as the reversible work required to remove one free ligand from its solvated environment ("bulk") and to bring it to the protein at the binding site ("site"). Then, the equilibrium constant can be written as configurational integrals of a ligand in the binding site normalized by all the ligands in the bulk:[56,57]

$$K_{eq} = \frac{1}{[\mathrm{ligand}]}\frac{N\int_{\mathrm{site}}\mathrm{d}\mathbf{1}\int\mathrm{d}x\; e^{-\beta U}}{\int_{\mathrm{bulk}}\mathrm{d}\mathbf{1}\int\mathrm{d}x\; e^{-\beta U}} \tag{1.19}$$

where $\mathbf{1}$ corresponds to the ligand at the binding site, N is the number of ligands, and U is the potential energy.

A ligand can be either in the bound state corresponding to the distance x_1, or in the unbound state with a distance x_1^* far away from the binding site, allowing for the introduction of a Dirac function in the equilibrium constant expression. Since the concentration of the ligand is equal to the ratio of the total number of ligands, N, to the bulk volume V, the expression simplifies to:[58,44]

$$K_{eq} = \frac{\int_{\text{site}} \mathrm{d}\mathbf{1} \int dx\, e^{-\beta U}}{\int_{\text{bulk}} \mathrm{d}\mathbf{1}\, \delta(x_1 - x_1^*) \int dx\, e^{-\beta U}}, \tag{1.20}$$

where the numerator and the denominator represent the bound and the unbound states, respectively.

However, the determination of configurational integrals is not achievable in a reasonable time using brute-force MD. To alleviate this limitation, Gumbart et al., based on the theoretical framework of Woo and Roux,[44] proposed to decompose the reversible association into several steps, the contribution of which can be evaluated separately into the geometrical and alchemical route.[25]

Once we obtained the equilibrium constant, the binding free energy is recovered using the following equation:

$$\Delta G^\circ = -\frac{1}{\beta} \ln(K_{\text{eq}} C^\circ) \tag{1.21}$$

where C° corresponds to the standard concentration of 1/1661 Å^3.[56]

1.3.1 Geometrical route

From the geometric standpoint, the binding of two molecular objects, a protein and another protein or a ligand, can be monitored by following the distance separating the two centers of mass as a CV, and the equilibrium constant could be seen as a one-dimensional PMF, $w(r)$:

$$K_{eq} = 4\pi \int_0^R \mathrm{drr}^2 \mathrm{e}^{-\beta \mathrm{w(r)}} \tag{1.22}$$

where R stands for the limit distance of association.

Equation eq 1.22 implies that to compute the equilibrium constant, all the conformational space around the ligand has to be sampled. Given the timescale of the simulations, this assumption is unlikely to be true, even with enhanced sampling methods introduced above. To alleviate this limitation the geometrical route proposes a gradual introduction of restraints on the two molecular partners to freeze their conformation and their relative position and orientation. These restraints are enforced via harmonic potential applied on CVs. These CVs are detailed in the following table:

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

Step	CVs	Partner movement	Representation*	Restraints
1	RMSD	Conformational	G_c^{site}, G_c^{bulk}	
2	Θ		G_Θ^{site}	RMSD
3	Φ	Orientational	G_Φ^{site}	RMSD, Θ
4	Ψ		G_Ψ^{site}	RMSD, Θ, Φ
5	θ		G_θ^{site}	RMSD, Θ, Φ, Ψ
6	ϕ	Positional	G_ϕ^{site}	RMSD, Θ, Φ, Ψ, θ
7	r		$w(r)$	RMSD, Θ, Φ, Ψ, θ, ϕ

*The superscripts "site" and "bulk" refer to the bound and the unbound states, respectively.

These restraints have a cost since they result in a loss of the conformational (ΔG_c), orientational (ΔG_o), and positional (ΔG_p) entropy. Therefore, their contribution must be considered in the binding free energy and estimated independently both in bulk (unbound state), and at the site (bound state). The equilibrium constant with the geometrical route can now be computed using:

$$
\begin{aligned}
K_{\text{eq}}^{\text{geom}} = & \frac{\int_{site} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x} e^{-\beta U}}{\int_{site} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x} e^{-\beta(U+u_c)}} \\
\times & \frac{\int_{site} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x} e^{-\beta(U+u_c)}}{\int_{site} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x} e^{-\beta(U+u_c+u_o)}} \\
\times & \frac{\int_{site} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x} e^{-\beta(U+u_c+u_o)}}{\int_{site} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x} e^{-\beta(U+u_c+u_o+u_a)}} \\
\times & \frac{\int_{site} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x} e^{-\beta(U+u_c+u_o+u_a)}}{\int_{site} \mathrm{d}\mathbf{1} \left(\mathbf{x_1} - \mathbf{x_1^*}\right) \int \mathrm{d}\mathbf{x} e^{-\beta(U+u_c+u_o)}} \\
\times & \frac{\int_{bulk} \mathrm{d}\mathbf{1} \left(\mathbf{x_1} - \mathbf{x_1^*}\right) \int \mathrm{d}\mathbf{x} e^{-\beta(U+u_c+u_o)}}{\int_{bulk} \mathrm{d}\mathbf{1} \left(\mathbf{x_1} - \mathbf{x_1^*}\right) \int \mathrm{d}\mathbf{x} e^{-\beta(U+u_c)}} \\
\times & \frac{\int_{bulk} \mathrm{d}\mathbf{1} \left(\mathbf{x_1} - \mathbf{x_1^*}\right) \int \mathrm{d}\mathbf{x} e^{-\beta(U+u_c)}}{\int_{bulk} \mathrm{d}\mathbf{1} \left(\mathbf{x_1} - \mathbf{x_1^*}\right) \int \mathrm{d}\mathbf{x} e^{-\beta U}}
\end{aligned}
\tag{1.23}
$$

where $u_0 = u_\Theta + u_\Phi + u_\Psi$ for the orientational contribution and $u_a = u_\theta + u_\phi$ for the position contribution of polar angles.

The three initial contributions correspond to conformational, orientational, and positional restraints, respectively, and are computed with six independent PMFs. The fourth contribution is the reversible separation of the ligand from the protein toward the bulk aqueous environment. This specific contribution is computed in the presence of all the afforded mentioned restraints, and can be computed using $S^* \times I^*$ where I^* is the separation term, and S^* is a surface term representing the fraction of a sphere of radius r^*, centered at the binding site of the reference protein, accessible to its partner.

$$
\begin{cases}
I^* = \displaystyle\int_{\text{site}} \mathrm{d}r \; \mathrm{e}^{-\beta\left(w(r) - w(r^*)\right)} \\
S^* = r^{*2} \displaystyle\int_0^\pi \mathrm{d}\theta \; \sin\theta \int_0^{2\pi} \mathrm{d}\phi \; \mathrm{e}^{-\beta u_a}
\end{cases}
\tag{1.24}
$$

with r^*, a point far from the binding site, where the proteins no longer interact with each other.I* is the separation term, and S* is a surface term, which represents the fraction of a sphere of radius r^*, centered at the binding site of the reference protein, accessible to its partner.

The contribution highlighted in cyan corresponds to the possible reorientation of the rigid-body lig-

and in its reference conformation when bound the protein, and is evaluated analytically using:

$$e^{-\beta\Delta G_o^{\text{bulk}}} = \frac{1}{8\pi^2}\int_0^{\pi} d\Theta \int_0^{2\pi} d\Phi \int_0^{2\pi} d\Psi e^{-\beta u_o} \tag{1.25}$$

The last contribution corresponds to restrain the ligand in its reference conformation in the bulk. Simplifying the eq 1.22 and using eq 1.21, the binding free energy can be seen the different PMFs contributions :

$$\Delta G_{\text{b}}^{\circ} = -\Delta G_c^{\text{site}} - \Delta G_o^{\text{site}} - \Delta G_a^{\text{site}} - \frac{1}{\beta}\ln(S^* I^* C^{\circ}) + \Delta G_o^{\text{bulk}} + \Delta G_c^{\text{bulk}} \tag{1.26}$$

1.3.2 Alchemical route

The second technique to compute the binding affinity of two molecular objects is based on a series of alchemical transformations.

The main idea is to reversibly decouple one of the partners from the rest of the system. This technique is unsuitable for protein-protein complexes due to the extensive perturbation involved, which can be challenging to converge and estimate.[57] Therefore, I will only refer to protein: ligand complex for the molecular partners in this part. To assess the convergence and micro-reversibility of the decoupling, the transformation is performed bi-bidirectionally. When performing the backward transformation, making the ligand reappear to the rest of the system by slowly reintroducing interactions, it is doubtful to replace it in the correct native bound state, thus breaking the thermodynamic micro-reversibility of the transformation. This is referred to as the wandering ligand problem in the literature.[59] To prevent this behavior, the alchemical route introduced restraints to freeze the ligand's relative conformation, orientation, and position to maintain it in the native bound conformation throughout the decoupling. Since the free energy is a state function, its value is independent of the path taken, so by computing the cost of the restraints and the free energy of annihilation of the restrained ligand (ligand*, (see the thermodynamic cycle of Figure 1.3) we can obtain the estimate of the standard absolute binding affinity. The restraints employed are the same as those presented for the geometrical route in the previous section.

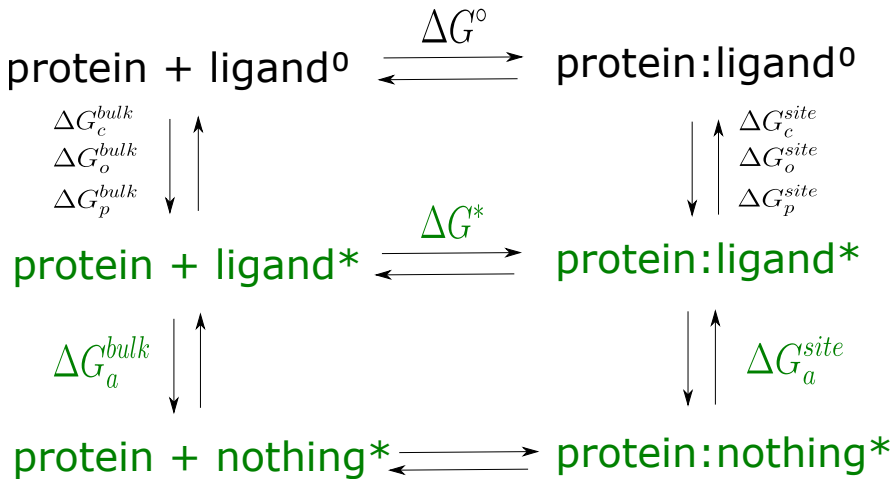

Figure 1.3: Complete thermodynamic cycle of the alchemical route, with the alchemical part colored in green. * stands for a restrained ligand , 0 stands for the standard state

The equilibrium constant can now be written as :

$$
\begin{aligned}
K_{\mathrm{eq}}^{\mathrm{alch}} = & \frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int d\mathbf{x}\, e^{-\beta U_1}}{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, e^{-\beta[U_1+u_c]}} \\
& \times \frac{\int_{\mathrm{site}} d\mathbf{1} \int \mathrm{d}\mathbf{x}\, e^{-\beta[U_1+u_c]}}{\int_{\mathrm{site}} d\mathbf{1} \int \mathrm{d}\mathbf{x}\, e^{-\beta[U_1+u_c+u_o]}} \\
& \times \frac{\int_{\mathrm{site}} d\mathbf{1} \int \mathrm{d}\mathbf{x}\, e^{-\beta[U_1+u_c+u_o]}}{\int_{\mathrm{site}} d\mathbf{1} \int \mathrm{d}\mathbf{x}\, e^{-\beta[U_1+u_c+u_o+u_a]}} \\
& \times \frac{\int_{\mathrm{site}} d\mathbf{1} \int \mathrm{d}\mathbf{x}\, e^{-\beta[U_1+u_c+u_o+u_a]}}{\int_{\mathrm{site}} d\mathbf{1} \int \mathrm{d}\mathbf{x}\, e^{-\beta[U_1+u_c+u_o+u_a+u_r]}} \\
& \times \frac{\int_{\mathrm{site}} d\mathbf{1} \int \mathrm{d}\mathbf{x}\, e^{-\beta[U_1+u_c+u_o+u_a+u_r]}}{\int_{\mathrm{site}} d\mathbf{1} \int \mathrm{d}\mathbf{x}\, e^{-\beta[U_0+u_c+u_o+u_a+u_r]}} \\
& \times \frac{\int_{\mathrm{bulk}} d\mathbf{1} \int \mathrm{d}\mathbf{x}\, e^{-\beta[U_0+u_c+u_o+u_a+u_r]}}{\int_{\mathrm{bulk}} d\mathbf{1}\delta\left(\mathbf{x_1}-\mathbf{x_1^*}\right) \int \mathrm{d}\mathbf{x}\, e^{-\beta[U_0+u_c+u_o]}} \\
& \times \frac{\int_{\mathrm{bulk}} d\mathbf{1}\delta\left(\mathbf{x_1}-\mathbf{x_1^*}\right) \int \mathrm{d}\mathbf{x}\, e^{-\beta[U_0+u_c+u_o]}}{\int_{\mathrm{bulk}} d\mathbf{1}\delta\left(\mathbf{x_1}-\mathbf{x_1^*}\right) \int \mathrm{d}\mathbf{x}\, e^{-\beta[U_0+u_c]}} \\
& \times \frac{\int_{\mathrm{bulk}} d\mathbf{1}\delta\left(\mathbf{x_1}-\mathbf{x_1^*}\right) \int \mathrm{d}\mathbf{x}\, e^{-\beta[U_0+u_c]}}{\int_{\mathrm{site}} d\mathbf{1}\delta\left(\mathbf{x_1}-\mathbf{x_1^*}\right) \int \mathrm{d}\mathbf{x}\, e^{-\beta[U_1+u_c]}} \\
& \times \frac{\int_{\mathrm{bulk}} d\mathbf{1}\delta\left(\mathbf{x_1}-\mathbf{x_1^*}\right) \int \mathrm{d}\mathbf{x}\, e^{-\beta[U_1+u_c]}}{\int_{\mathrm{bulk}} d\mathbf{1}\delta\left(\mathbf{x_1}-\mathbf{x_1^*}\right) \int \mathrm{d}\mathbf{x}\, e^{-\beta[U_1]}}
\end{aligned}
\tag{1.27}
$$

The potential energy U_0 characterizes the fully decoupled state of the ligand, while U_1 represents the state in which it bound and interacts with the environment. The above equilibrium constant in (eq 1.27) precisely follows the various steps described in the thermodynamic cycle depicted in Figure 1.3. The first two contributions arise from the conformational, u_c, and orientational, u_o, harmonic restraints imposed on the ligand. The third contribution corresponds to the polar angles term, u_a, representing the positional restraints. The fourth contribution corresponds to the translational term, u_r, of the positional restraints. The fifth and eighth contributions, colored in orange, correspond to the alchemical transformations in which the ligand is reversibly decoupled from its environment, respectively, in the bound and unbound states as stated in the thermodynamic cycle of Figure 1.3. The sixth and seventh contributions, highlighted in cyan, are analytical and account for the reorientation and translation of a rigid body, e.g., the

ligand, in a homogeneous bulk liquid. The final contribution represents the conformational changes of the ligand in the free, unbound state..[60,57,61]

The decoupling of the interaction and restraint is performed by coupling the nonbonded interactions (vdW and electrostatics) of the ligand to a parameter, λ, defined such as $\lambda = 0$ correspond to interactions fully turned on, and $\lambda = 1$ fully turn off. Estimating the free energy cost for the restraints is performed using thermodynamic integration (TI) formalism (see next paragraph) and the annihilation of the ligand using free energy perturbations (FEP).

Evaluation of the convergence of the alchemical transformations requires examination of the hysteresis between the backward and forward transformation of a given λ, checking both the free-energy changes and the overlap of the probability distribution functions for FEP (see initial figures in Chapter 3)

Thermodynamical Integration

Thermodynamic integration (TI) is one of the most used methods for free energy calculation. It relies on computing, before integrating, the free energy derivatives with respect to an order parameter along a transformation path connecting two states. The order parameter can be just a parameter in the Hamiltonian or function of the coordinates..[62]

The derivative of the free energy can, therefore, be obtained by the following equation :

$$\Delta A = \int_0^1 \frac{\partial A_\lambda}{\partial \lambda} d\lambda = \int_0^1 \left\langle \frac{\partial U(x,\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \tag{1.28}$$

In practice, the continuous integral (eq 1.28) is computed with techniques such as the trapezoidal rule, the accuracy of which depends on the smoothness of the integrand and the number of evaluation points.[63] Furthermore, TI offers flexibility in defining the diverse λ , allowing for fine control over the transformation between states.[62]

Free Energy Perturbation

FEP is one of the oldest formalists to compute free-energy differences, initially developed by Zwanzig.[64] First, a reference state (state 0) is defined by the potential energy $U_0(x)$ where x represents the coordinates. The target state (state 1), $U_1(x)$, is the sum of the reference state and the perturbation ($\Delta U(x)$):

$$U_1(x) = U_0(x) + \Delta U(x) \tag{1.29}$$

In our case, we can define the reference state as the bound state of molecular objects at play and the target state as the unbound state. The difference in free energy can be defined as the difference between the two states following:

$$\Delta A = -\frac{1}{\beta} \ln \frac{Z_1}{Z_0} \tag{1.30}$$

where β is the inverse of the Boltzmann constant k_B multiplied by the temperature, Z_0 and Z_1 correspond to the partition function of state 0 and 1, respectively. The partition function can be written as:

$$Z = \frac{1}{h^{3N} N!} \int \int \beta U] \mathrm{d}\mathbf{x} \tag{1.31}$$

Exploiting the partition function and density probability equations, ΔA can be simplified into:

$$\Delta A = -\frac{1}{\beta} \ln \langle \exp[-\beta \Delta U(x)] \rangle_0 \tag{1.32}$$

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.[62] In practice, it is rare for single-step transformations between highly distinctive states to meet the requirement of small changes and micro-reversibility.[62]

Therefore, the reaction pathway connecting states 0 and 1 must be divided into multiple intermediate states, even though these states may not correspond to actual physical configurations. Bennett first introduced this idea of introducing intermediate states, referred to as "overlap sampling".[65] By incorporating these additional states in the transformation 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..[66,67] To achieve this, the energy is formulated as a function of a coupling parameter,λ, which governs the transformation between states. The optimum choice of the number of intermediate states is essential. An insufficient number can result in inadequate exploration of specific energy regions, thus introducing bias and inaccuracies in the free-energy difference estimation. On the contrary, an excessive number increases the computational cost without a gain in accuracy. The selection of the appropriate number of intermediate λ states must take into account the available computational resources, the perturbation at play, and the complexity of the system.[68,62]

Chapter 2

Methods for machine learning

Sommaire

2.1 Introduction

Machine Learning (ML) usually starts with a collection of observables called the training dataset and a model. The model is a mathematical function, or a set of functions, of the training data points with extra parameters. One can use an algorithm (machine part) to adjust the model's parameters (learning part) to optimize the outcome. ML can be used for several tasks: regression, classification, clustering, and pattern recognition. ML algorithms can be divided into three classes: supervised, unsupervised, and reinforcement learning. In supervised learning, the desired outcome called the target, is already known, and the goal of the model is to learn how to predict this expected result when presented with new unseen data. In unsupervised learning, only the initial data points are known, and the objective is to find patterns or relationships in the dataset in order to be able to generalize without explicit guidance. The last class, reinforcement learning, involves learning how to interact with the environment to maximize a reward. For example, learning a car itinerary with a reward every time the path reaches the destination to get the fastest route.

In predicting binding free energies with ML, the focus is primarily on supervised learning, since some affinity values are already known. Due to the difficulties in predicting this thermodynamic quantity, we set up our approach towards classification methods belonging to supervised learning. More specifically,

we chose two different algorithms, Random Forest[35] and Linear Discriminant Analysis,[69] which we will detail in the following sections of this chapter.

The most complicated part of ML in structural biology is the design of input features,[41] which is the data used to train and classify protein complexes. This topic is explored in the first section of this chapter.

2.2 Input features design

Traditionally, the input features for proteins belong to two categories: sequence features extracted purely from the amino acid sequence or structural features from the 3D structure of molecular objects.

2.2.1 Sequence-based features

Sequence-based features are only extracted from the primary structure of proteins, that is, their amino acid sequence. They can be categorized into four distinct categories: (i) residue primary encoding, (ii) evolutionary-based information, (iii) residue physicochemical properties, and (iv) predicted structural features.[41]

For primary encoding, one can think of traditional bio-informatics techniques such as one hot encoding, a vector of length twenty containing 19 zeros except for one position with the number 1 representing the residue at hand in the sequence. However, this representation is basic and unsuitable when dealing with protein interactions.[41] The second category, evolutionary-based representation, is, therefore, more employed for protein-protein interactions. It relies on computational sequence alignments to generate features such as sequence profiles, position-specific scoring matrices, or a conservation score.[70] Residue physicochemical properties, such as their hydrophobicity, charge or volume, or conformational properties, have also been routinely included in diverse prediction algorithms related to protein-protein interactions.[70] The last category of predicted structural features, encompasses properties such as relative solvent surface accessibility, secondary structure, and protein flexibility.[41] It often reinforces purely sequence-based features.

2.2.2 Structure-based features

Using structure-based features requires prior knowledge of the 3D atomic structure of the protein or protein complex, making them dependent on the availability of such structural data. They can be geometrical descriptors, such as indexes describing the surface shape or curvature, Zernicke descriptor of the voxelized protein surface representation,[71] geometric invariant fingerprints.[34,72] Conventionally, measures of flexibility using crystallographic B-factors, secondary structure motifs, and residue solvent accessibility are also used in structure-based features.[41]

2.2.3 Design of input features used in this work

Sequence-based features were used for the work presented in this thesis involving ML (covered in Chapter 7) since we have access to all the sequences from both target protein families (Dpr-DIP).[11] We determined evolutionary conserved residues using a multiple sequence alignment for each family with CLUSTAL -W[73] and a scoring metric for the conservation called the Mutual Information (MI). The MI is defined by the difference between the individual entropy at the position, H_i and H_j, and the joint entropy H_{ij}.

$$\mathrm{MI} = -H_{ij} + H_i + H_j = \sum_{x,y} p_{ij}(x,y) \log_2 \frac{p_{ij}(x,y)}{p_i(x)p_j(y)} \tag{2.1}$$

where $p_i(x)$ and $p_j(y)$ stands for the occurrence probability of a given amino acid at position i or j, $p_{ij}(x, y)$ is the joint probability of the amino acids in the multiple sequence alignment. This metric allows estimation of how much information is gained by accounting for the covariance of residues of DIP and Dpr rather than treating them separately. The MI between non-cognate pairs is treated as noise since they are not supposed to have co-evolved and are, therefore, generated by simple combinatorics. Consequently, the MI generated by non-cognate pairs is subtracted from the one obtained by the cognate pairs to select interacting pairs with non-zero MI values. To account solely for interacting residues at the interface, the MI is then weighted by an inverse distance, either from a backbone carbon α extracted from the crystal structure of the Dpr10DIPα[74] since every complex shares the same fold, or from the COM of side chains extracted from MD equilibrium trajectories taken from reference.[11] To consider the energetics involved in residue interactions, a pairwise amino potential is added to form an interacting score by summing it for every amino-acid (AA) pair of a complex:

$$\text{score} = \sum_{AApairs} \text{MI} \times \text{distance} \times \text{potential} \tag{2.2}$$

These input features combine sequence (MI), structure (distance filter), and physics-based metrics (pairwise potential), each bringing a different type of information to reinforce the feature's ability to describe the complex at play best. This input feature generation process is easily transferable to any protein-protein interactome and shows an excellent potency for future use in the protein-protein interaction prediction field.

2.3 Linear Discriminant Analysis

2.3.1 Principle and Mathematical formalism

Linear Discriminant Analysis (LDA)[69] is a classification method used to obtain the best linear combinations of input features to discriminate between classes and draw the linear decision boundaries to separate the different classes (see Figure 2.1).

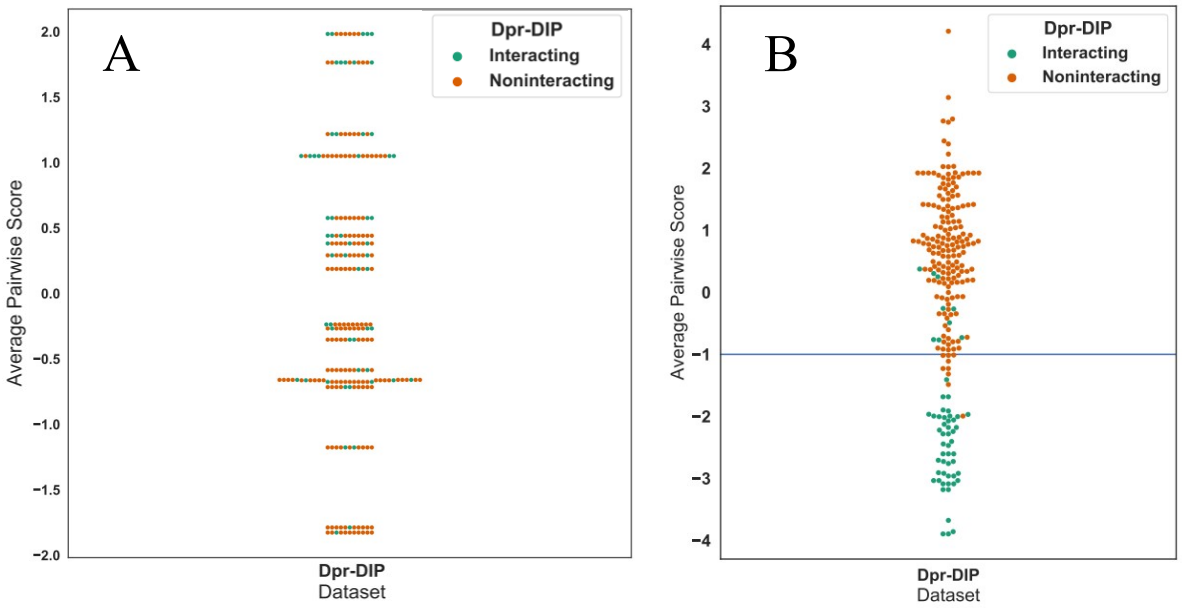

Figure 2.1: Scoring of complexes DIP-Dpr A) prior and B) after LDA using Drosophila Melanogaster dataset

Let us define the classifier/predictor $G(x)$. It takes discrete values represented by the set $\mathcal{G}$, and therefore, the input space can be separated into regions labeled according to their class. The specificity of LDA resides in the linearity of those boundaries.[75] LDA comes down to model the class conditional distribution of data $P(X|y=k)$ for the k classes. Using Bayes rule, predictions for training samples x can be obtained with the following:

$$P(y=k|x) = \frac{P(x|y=k)P(y=k)}{P(x)} \tag{2.3}$$

with the associated Gaussian distribution density:

$$P(y=k|x) = \frac{1}{(2\pi)^{d/2}\left|\sum_k\right|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^t \sum_k^{-1}(x-\mu_k)\right) \tag{2.4}$$

In the LDA, the covariance matrix of every class, $\sum$, is assumed to be identical. Therefore, the log-posterior is reduced to the following:

$$\log P(x=k|y) = -\frac{1}{2}(x-\mu_x)^t \sum^{-1}(x-\mu_x) + \log P(y=k) + cst \tag{2.5}$$

The term $(x-\mu_x)^t \sum^{-1}(x-\mu_x)$ referred to the Mahalanobis distance[76] between the sample x and the mean μ_x. It is a proxy to estimate how close the sample x is to mu_x while considering the variance. Ledoit and Wolf proposed a different notation of the log-posterior using shrinkage of the covariance matrix:[77]

$$\log P(y=k|x) = \omega_k^t x + \omega_{k_0} + cst \tag{2.6}$$

with $\omega_k^t = \sum^{-1}\mu_k$, the coefficients and $\omega_{k_0} = -\frac{1}{2}\mu_k^t \sum^{-1}\mu_k + \log P(y=k)$, the intercept of the linear boundary.

2.3.2 Scikit-learn implementation and tuning

We used the scikit-learn[78] implementation of the LDA in Python, which proposed three different estimation methods:

- 'svd solver' corresponds to the default method associated with this LDA implementation. The advantage is that the solver does not require compute the covariance matrix, making it an appropriate choice for a large amount of data.

- 'lsqr solver' requires computation of the covariance matrix $\sum$, without having to compute explicitly the inverse matrix $\sum^{-1}$, using $\sum \omega = \mu_k$, thus being more efficient than the 'eigen solver' in the case of classification.

- 'eigen solver'. It involves complete computation of the covariance matrix, rendering it unsuitable in the case of many input features.

In chapter 7, while using LDA I chose the 'svd solver' as it provided the best results and the least amount of computational time.

2.4 Random Forest

2.4.1 Principle

Random Forest (RF) is an ensemble method and consists of a forest of random decision trees. It was initially put forth by Breiman in 2001,[35] although the term "random forest" was first introduced by Ho in 1995.[79] RF is a variation of bagging with de-correlated trees averaged at the end.[80] Due to its easy tuning and interpretability, it is widely used.

The idea is to remove the noise from the data and reduce the variance, which is performed well by decision trees since they can capture complex patterns. However, trees remain noisy themselves and benefit significantly from the final averaging..[80] Individual trees generated by bagging are identically distributed, and the average obtained from a subset of tree B is identical to one of any of them. Therefore, the only improvement is through variance reduction.[80] Specifically, in RF, the variance reduction is improved by reducing the correlation between trees without over-increasing the variance through the random selection of the input variables in the tree growing state process (see (b) in the algorithm 1 below).[80]

2.4.2 Algorithm

Algorithm 1 Random Forest for Classification, taken from reference[80]

1. For $b = 1$ to B :
   (a) Draw a bootstrap sample **Z** of size N from the training data
   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
      (i) Select m variables at random from the p variables
      (ii) Pick the best variable/split point among the m
      (iii) Split the node into two daughter nodes.
2. Output the ensemble of trees.

To make a prediction at a new point x : Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{rf}^{B}(x) = majority\ vote \left\{\hat{C}_b(x)\right\}_1^B$.

2.4.3 Parameters and Scikit-learn formalism

We use the scikit-learn implementation of the RF Classifier in this work.[78] In this particular implementation of the RF, the data at node m is represented by Q_m with n_m samples. For split at a node $\theta = (j, t_m)$, j being a feature and t_m a threshold, the goal is to partition the data into a right ($Q_m^{right}(\theta)$)and left ($Q_m^{left}(\theta)$) subsets, with

$$Q_m^{left}(\theta) = \{(x, y) \,|x_j \leq t_m\} \tag{2.7}$$

$$Q_m^{right}(\theta) = Q_m / Q_m^{left}(\theta) \tag{2.8}$$

where x is the training data vector and y is the class label.

The quality assessment of a potential split is estimated with an impurity function or loss function H. The scikit-learn implementation offers two diverse methods to estimate the impurity :

- the Gini:

$$H(Q_m) = \sum_k p_{mk}(1 - p_{mk}) \tag{2.9}$$

- the Entropy (can also be referred to as the log loss)

$$H(Q_m) = -\sum_k p_{mk} log(p_{mk}) \tag{2.10}$$

with p_{mk} the proportion of k classes observable at node m.

This process is repeated for each split until reaching either the fixed length of the tree or there is only one sample remaining.

I used the entropy criteria as it provided better results in my case. Furthermore, we already use an entropic metric to compute the MI in the input features, thus justifying continuing to use entropy criteria to measure the quality of the split at a node in the RF.

Chapter 3

Detailed Protocol for Standard Binding Free-energy

Sommaire

This chapter consists into the detailed protocol I used for the diverse binding free-energy calculations throughout this thesis, which was published in 2022 in the Nature Protocol Journal under the following reference:

"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. *J. Chem. Theory Comput.*, Doi.org/10.1038/s41596-021-00676-1."

Firstly, I explain the inclusion of this work in my thesis as well as my detailed contribution. Then, I provide a summary of this publication before moving on to the entire original text.

3.1 Motivations and personal contribution

During the first year of my Ph.D., I embarked on learning how to conduct binding free-energy calculations. In that spirit, I participated in the development of a protocol for protein-ligand, leveraging the BFEE2 tool previously developed in the group.[81] This work resulted in the following publication[82] and the entire original text is presented in the following sections. I focused on performing calculations for three specific complexes involving the T4 Lysozyme L99A and three different ligands, namely ethyl benzene (PDBID: 1NHB),[83] paraxylene (PDB: 187L),[83] and n-butyl-benzene (PDB: 186L).[83] Their chemical structure is shown below in Figure 3.1. Lysozyme T4 is an enzyme involved in the degradation of peptidoglycans to ensure the release of mature viral particles produced in a host cell.

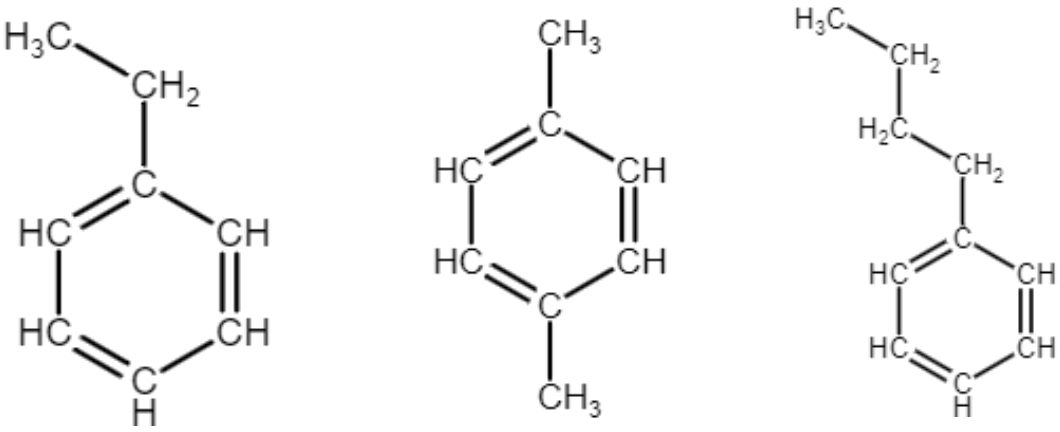

Figure 3.1: Chemical structure of A) ethylbenzene B) para-xylene C) n-butylbenzene

All the ligands were buried in a hydrophobic pocket, making the alchemical route the only option for estimating the standard binding free energy. The specific parameters for the ligand, not included in the selected force field we used (for protein, CHARMM36[84] and TIP3 water model[85]), were generated individually through the CGenFF webserver.[48] To estimate how reliable the parameters provided are, the server gives "penalties". None of the ligands generated penalties (score of 0), showing that we could use the parameters safely in our calculations.

Before using BFEE2 pretreatment options to generate our simulation inputs, all complexes were equilibrated in NVT ensemble with NAMD 3.0 program[86] by following this protocol: 500 steps of minimization, 5 ns of equilibration with restraints on the protein backbone and ligand with a force of 1 kcal/mol. This force was then gradually decreased with a scaling of the force to 0.75, 0.5, and finally 0.25 for 500 ps at each step. The restraints were then completely removed, leading up to a total of 100 ns of equilibration.

I will go into more detail about the results for every complex I studied in the following parts since they are not detailed in the manuscript, which focuses more on the practical aspects of running these simulations.

3.1.1 Ethylbenzene

Presentation and computational details

This complex can be described as a buried ligand/globular protein system. The ligand is located inside a hydrophobic cavity of 153 A3 and contains no unpaired hydrogen bond donors or acceptors except for M102 sulfurs, as presented in Figure 3.2. The structure was obtained from PDB entry 1NHB resolved at 1.80 Å.[83] The experimental binding affinity was obtained wih calorimetric analysis[87] and reached −5.76 kcal/mol.

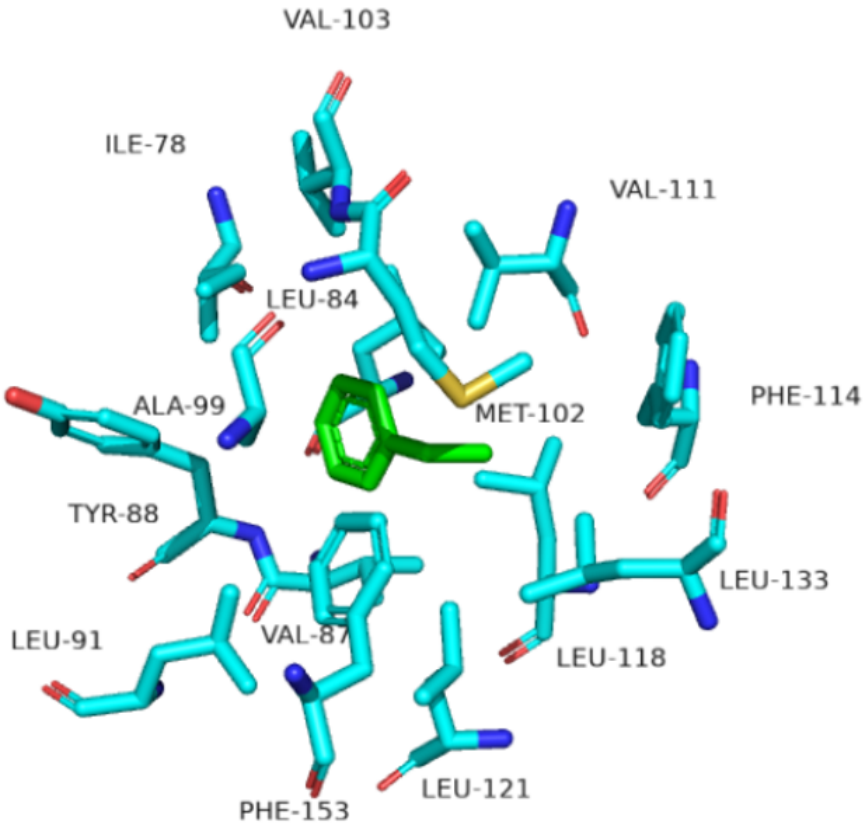

Figure 3.2: Binding pocket of ethylbenzene in lysozyme TA L99A

The crystallographic waters were retained to build the complex for the following simulations. The complex was then solvated in a cubic box and neutralized with Na^+, Cl^- counter ions leading to a system composed of 36678 atoms (2604 atoms for the protein, 18 for the ligand, 11 349 water molecules, and 9 Cl^- ions). The dimension of the periodic cell was 68 x 72 x 80 $Å^3$.

The number of windows was set in advanced settings to 50 for the coupling/decoupling parts and 15 for TI calculations. The lambda schedule used for the TI calculation in the bound state was changed to: "0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99, 0.999, 0.9999, 0.99999, 1". The time per window was also modified with 0.4 ns of equilibration and 1.6 ns of data collection for a total of 2 ns per window. For the FEP calculations, the sampling time was up to 3 ns per window instead of the default of 1ns and cut into ten blocks of 0.1 with $\Delta\lambda = 0.02$. The starting point of each block was a configuration at a higher lambda to prevent sampling from being stuck in a kinetic trap. The margin parameter was modified to 8, the steps per cycle to 400, and the pairlistpercycle to 40 to consider the use of GPUs. The rest of the parameters were left to the default provided by BFEE2.

Figure 3.3 presents the restraints cost obtained with TI for ethylbenzene. The hysteresis between the forward and backward transformation is low for all the components, which demonstrates convergence of the simulations.

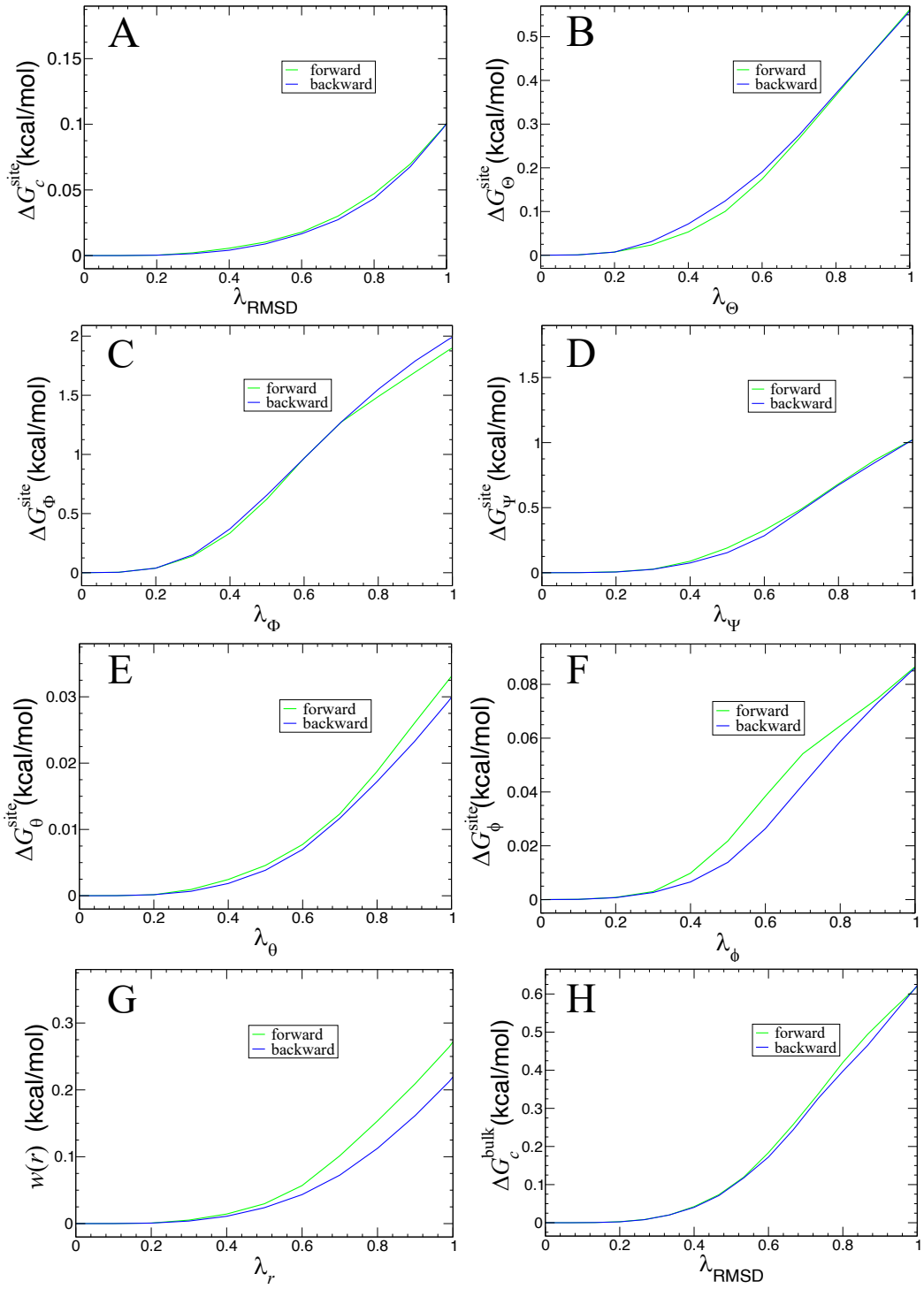

Figure 3.3: Restraints contributions for ethylbenzene in the bound state, with (A) the RMSD, (B) Θ, (C) Φ, (D) Ψ , (E) θ, (F) ϕ and (G) the distance between COMs. (H) corresponds to the RMSD contribution in the unbound state

Figure 3.4, Figure 3.5, Figure 3.6, and Figure 3.7 are displaying the free energy change and probability distributions obtained for coupling/decoupling of the ligand in the bound and unbound state respectively.

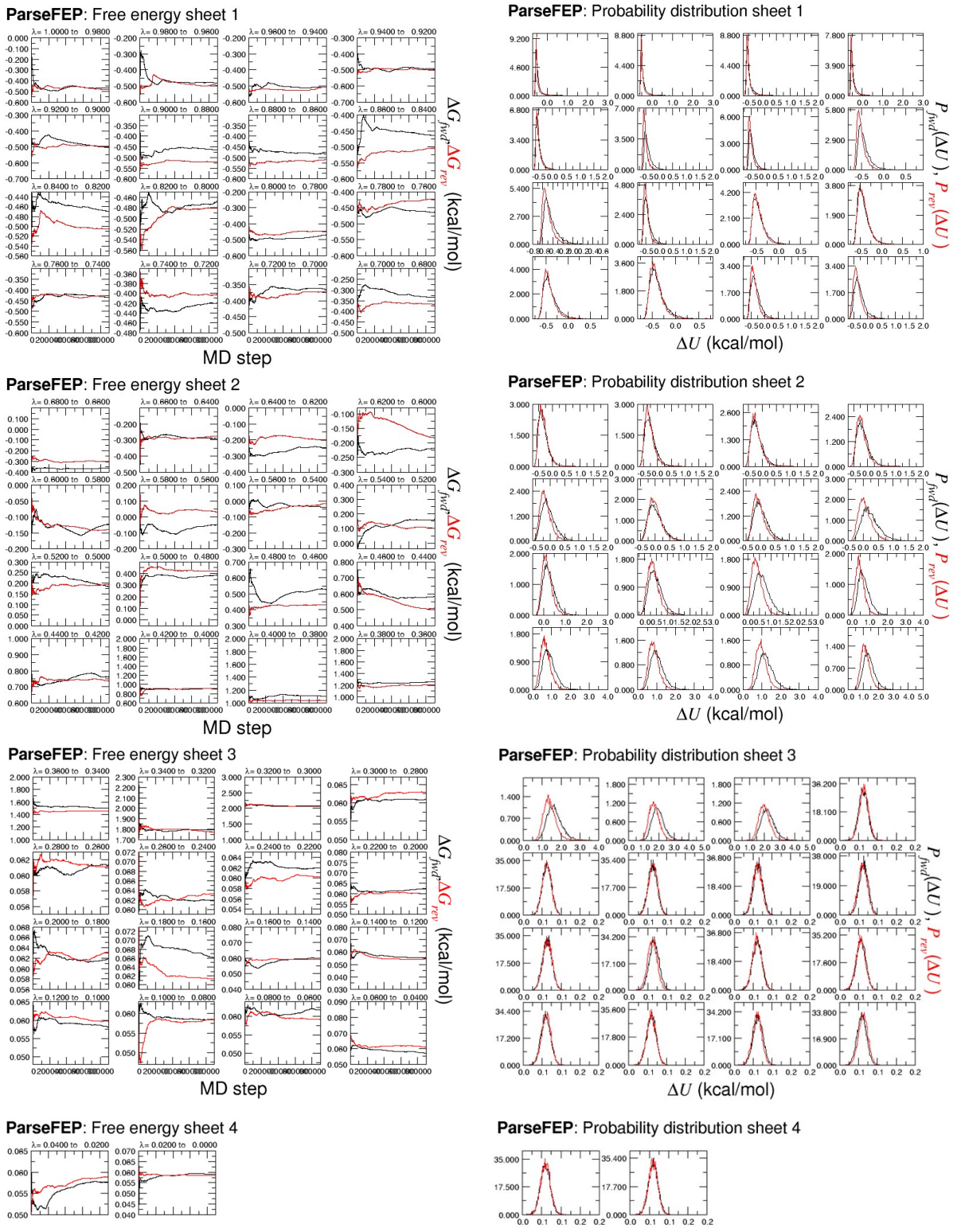

Figure 3.4: Free energies and probability distributions with hysteresis for coupling/ decoupling steps in FEP calculations for the bound state generated by ParseFEP[88]

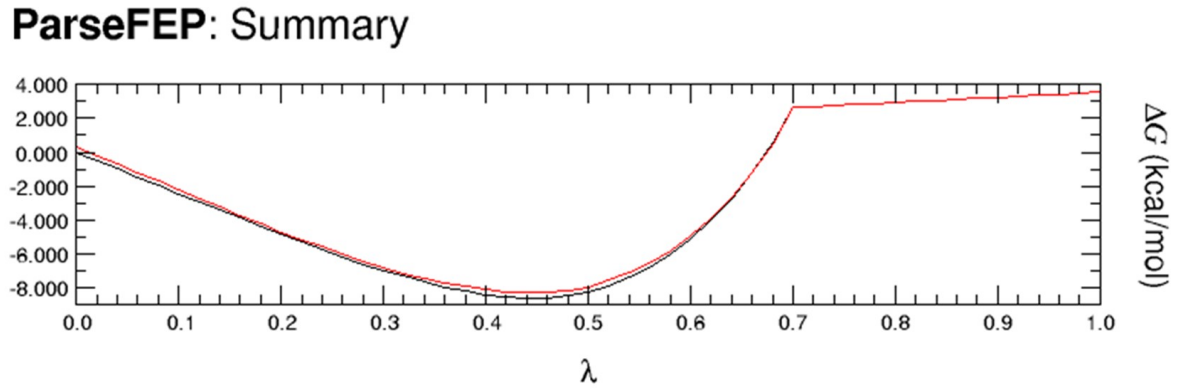

Figure 3.5: Summary of the full transformation for the reversible annihilation of the restrained ligand performed with FEP and obtained with the ParseFEP module of NAMD in the bound state

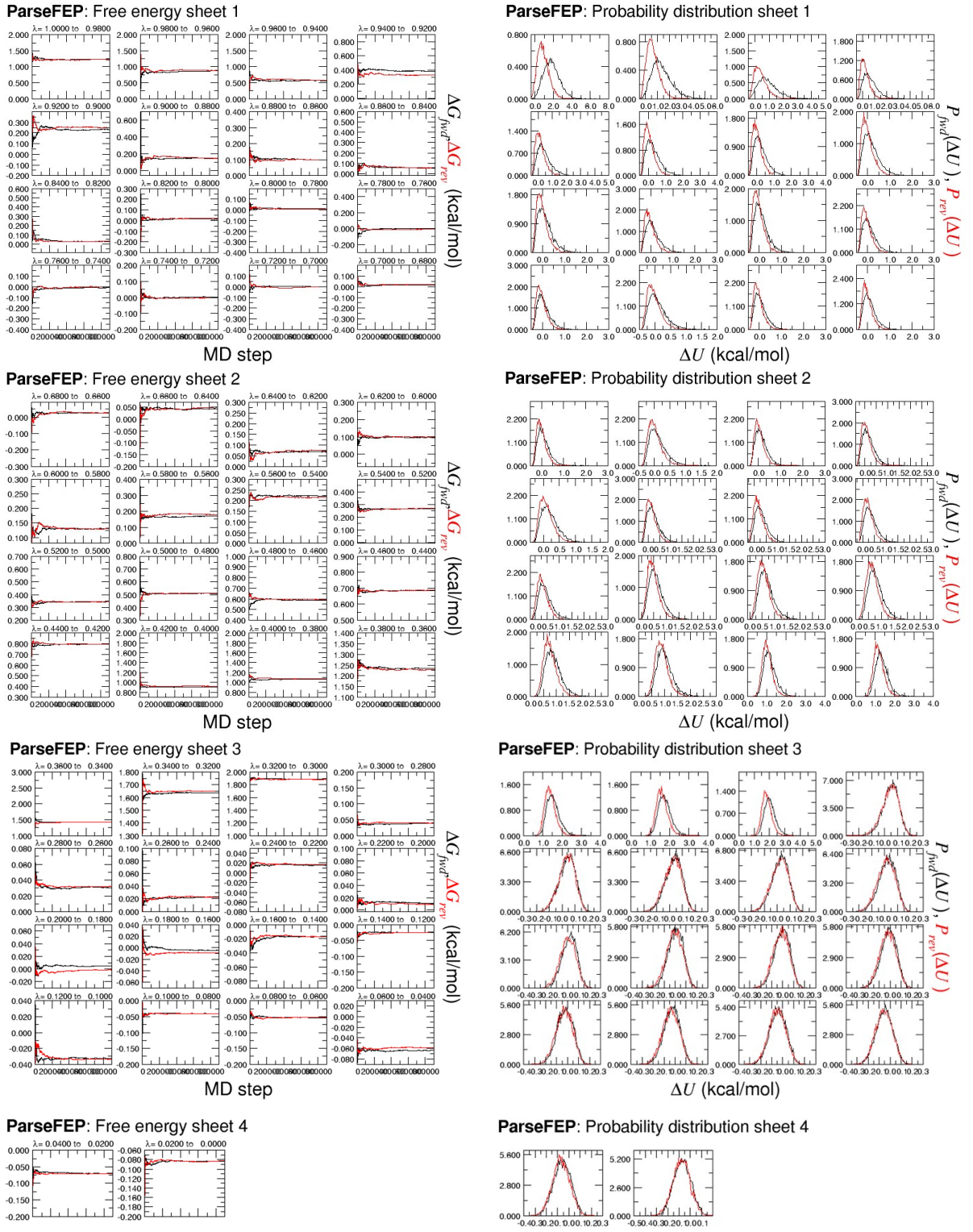

Figure 3.6: Free energies and probability distribution functions with hysteresis for coupling/ decoupling steps in FEP calculations for the unbound state generated by ParseFEP[88]

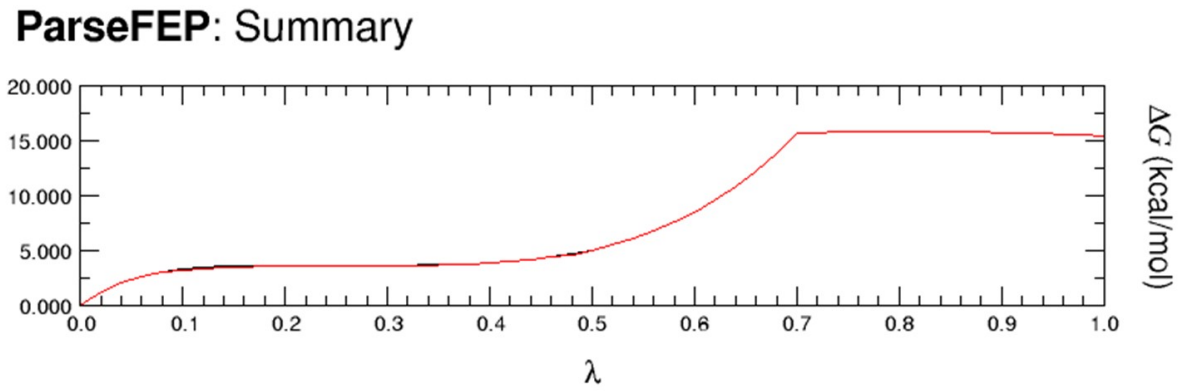

Figure 3.7: Summary of the full transformation for the reversible annihilation of the restrained ligand performed with FEP and obtained with the ParseFEP module of NAMD[88] in the unbound state

In Figure 3.5, Figure 3.4, Figure 3.6 and Figure 3.7, we observe that the probabililty distribution and free energies are overlapping for the forward (black) and backward (red) transformations, demonstrating convergence of the diverse simulations.

The following table reports all the contributions for the computation of the final estimate along with the speed and the chosen stratification strategy.

Contribution	Alch. (kcal/mol)	Alch. (ns)	Timing (ns/day)	Stratification
$\Delta G^{site}_{decouple}$	2.7 ± 0.3	300	20.57 (3.5h/window of 3ns)	50 windows
ΔG^{site}_{c}	-0.1 ± 0.0	60	22.96	15 windows
ΔG^{site}_{Θ}	-0.6 ± 0.0			
ΔG^{site}_{Φ}	-2.0 ± 0.1			
ΔG^{site}_{Ψ}	-1.1 ± 0.0			
ΔG^{site}_{θ}	-0.03 ± 0.0			
ΔG^{site}_{ϕ}	-0.01 ± 0.0			
ΔG^{site}_{r}	-0.3 ± 0.1			
$\Delta G^{bulk}_{decouple}$	-15.5 ± 0.1	200	80(0.6h/window of 2ns)	50 windows
ΔG^{bulk}_{c}	0.6 ± 0.0	30	107.35	15 windows
ΔG^{bulk}_{o+a+r}	11.2			
ΔG°_{b}	-5.2 ± 0.2	$\Delta G^{\text{exp}}_{b}$: -5.76 kcal/mol[87]		

Table 3.1: Detailed of the diverse binding free-energy contributions to the final ΔG°_{b} for the ethylbenzene

Our estimates fall close the experimental binding affinity, proving the protocol's success. Furthermore, binding free energy for this complex had already been performed by Deng and Roux, and they obtained a binding affinity of -5.04 kcal/mol.[89] Their estimate is slightly further away from the experimental value than our estimate, thus enforcing the potency of the protocol.

3.1.2 Paraxylene

This complex comprises the same Lyzozyme T4 L99A protein bound to paraxylene and uses the same binding pocket. The binding pose is represented in Figure 3.8.

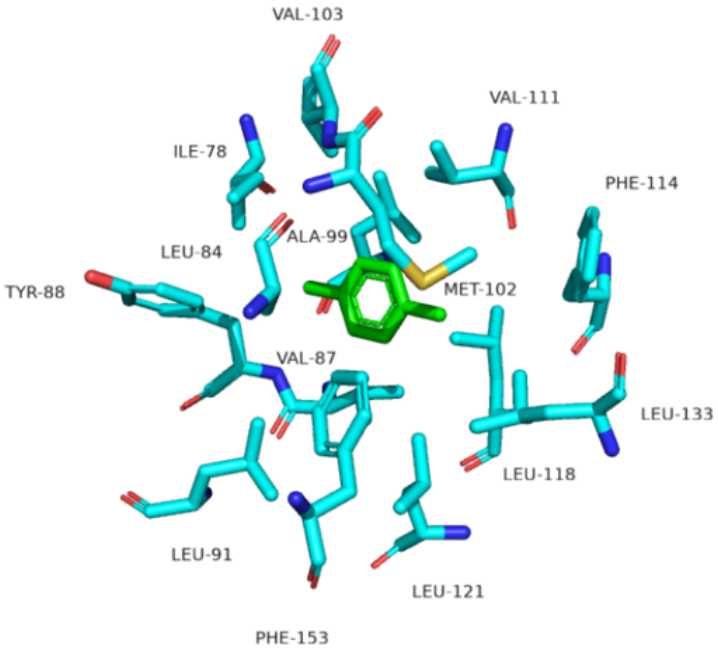

Figure 3.8: Binding pocket of paraxylene in lysozyme T4 L99A

The X-ray structure was taken from the PDB using the code 187l.[87] The crystallographic waters were retained to build the complex for the following simulations. The complex was then solvated and neutralized with NA^+, Cl^- counter ions, leading to a system composed of 36153 atoms (2604 atoms for the protein, 18 for the ligand, 11 174 water molecules, and 9 Cl^- ions). The dimension of the periodic cell was 68 x 72 x 80 Å3. The starting coordinates for BFEE2[81] were taken at the end of equilibration when the ligand had the same conformation as in the crystallographic structure and saved into a PDB file using VMD[90] The parameters for TI and GPU use correspond to those used for ethylbenzene. For the FEP calculations, the sampling time was up to 3 ns per window instead of the default of 1ns and cut into ten blocks of 0.1 with $\Delta\lambda = 0.02$. The rest of the parameters were left to the default provided by BFEE2.

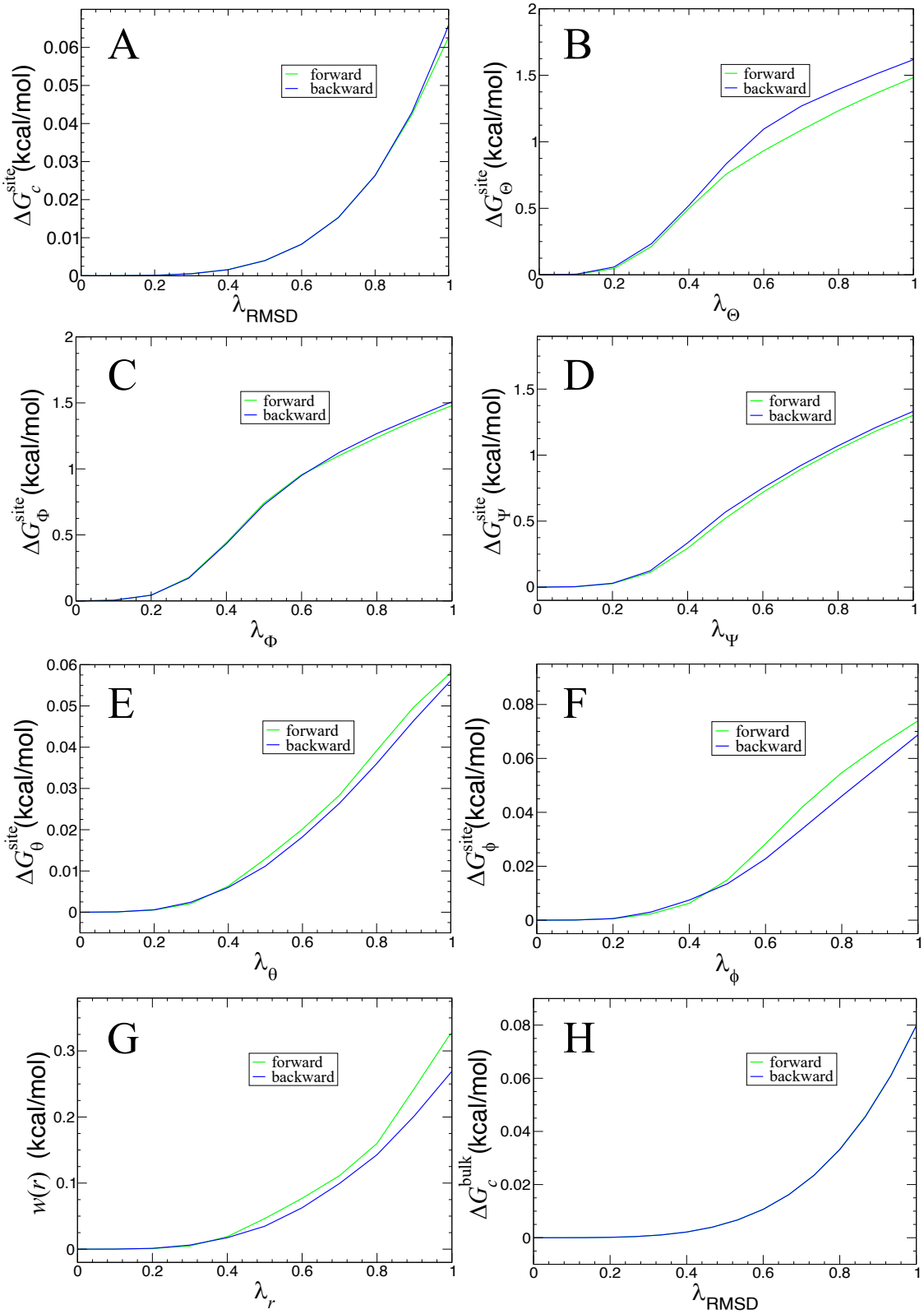

Figure 3.9: Restraints contributions for paraxylene in the bound state, with (A) the RMSD, (B) Θ, (C) Φ, (D) Ψ , (E) θ, (F) ϕ and (G) the distance between COMs. (H) corresponds to the RMSD contribution in the unbound state.

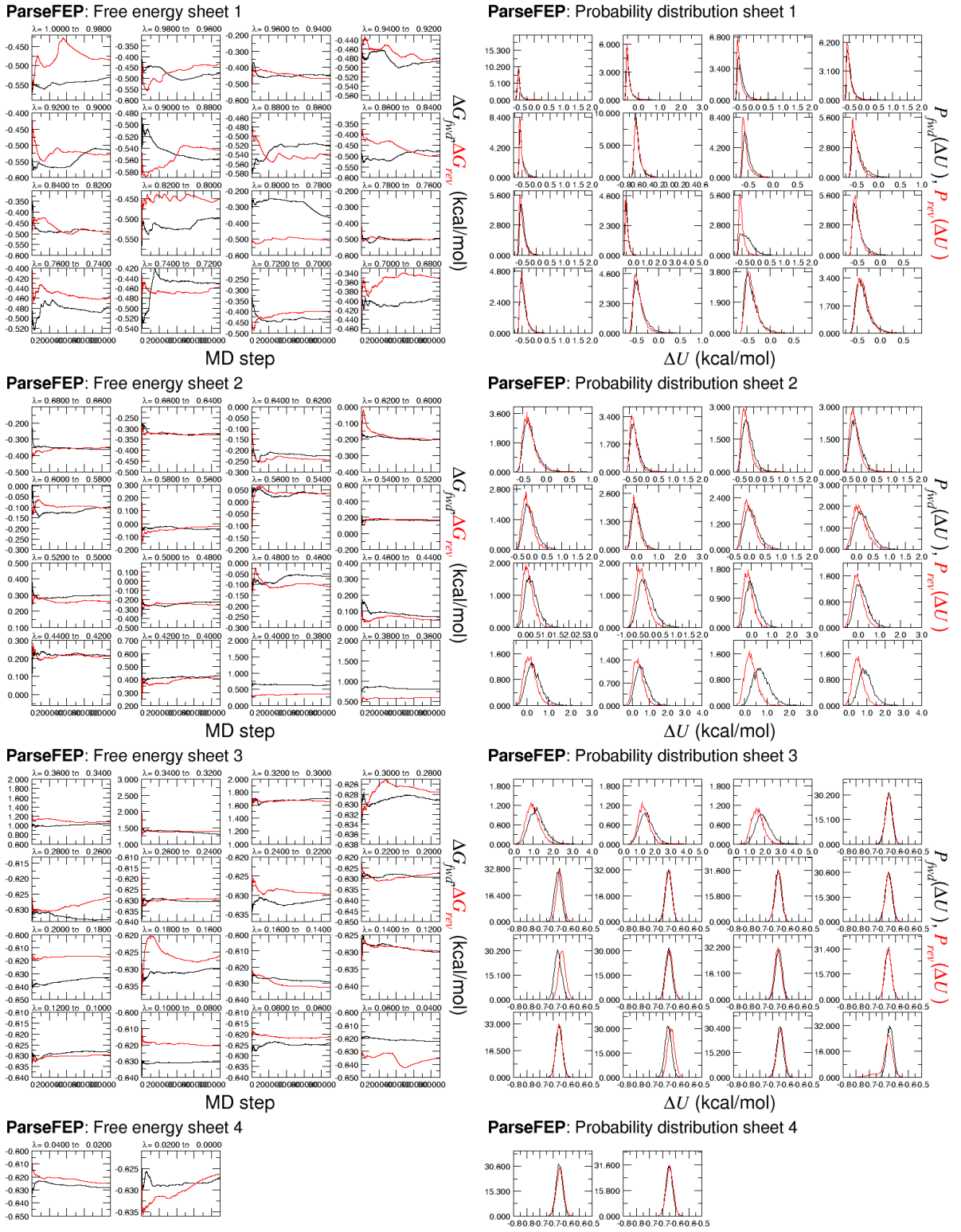

Figure 3.10: Free energies and probability distributions with hysteresis for coupling/ decoupling steps in FEP calculations for the bound state of paraxylene generated by ParseFEP[88]

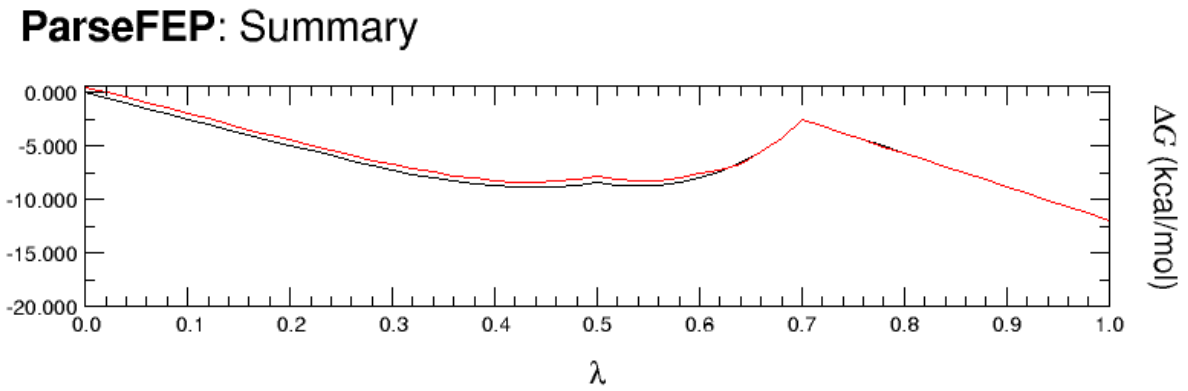

Figure 3.11: Summary of the full transformation for the reversible annihilation of the restrained paraxylene performed with FEP and obtained with the ParseFEP module of NAMD in the bound state

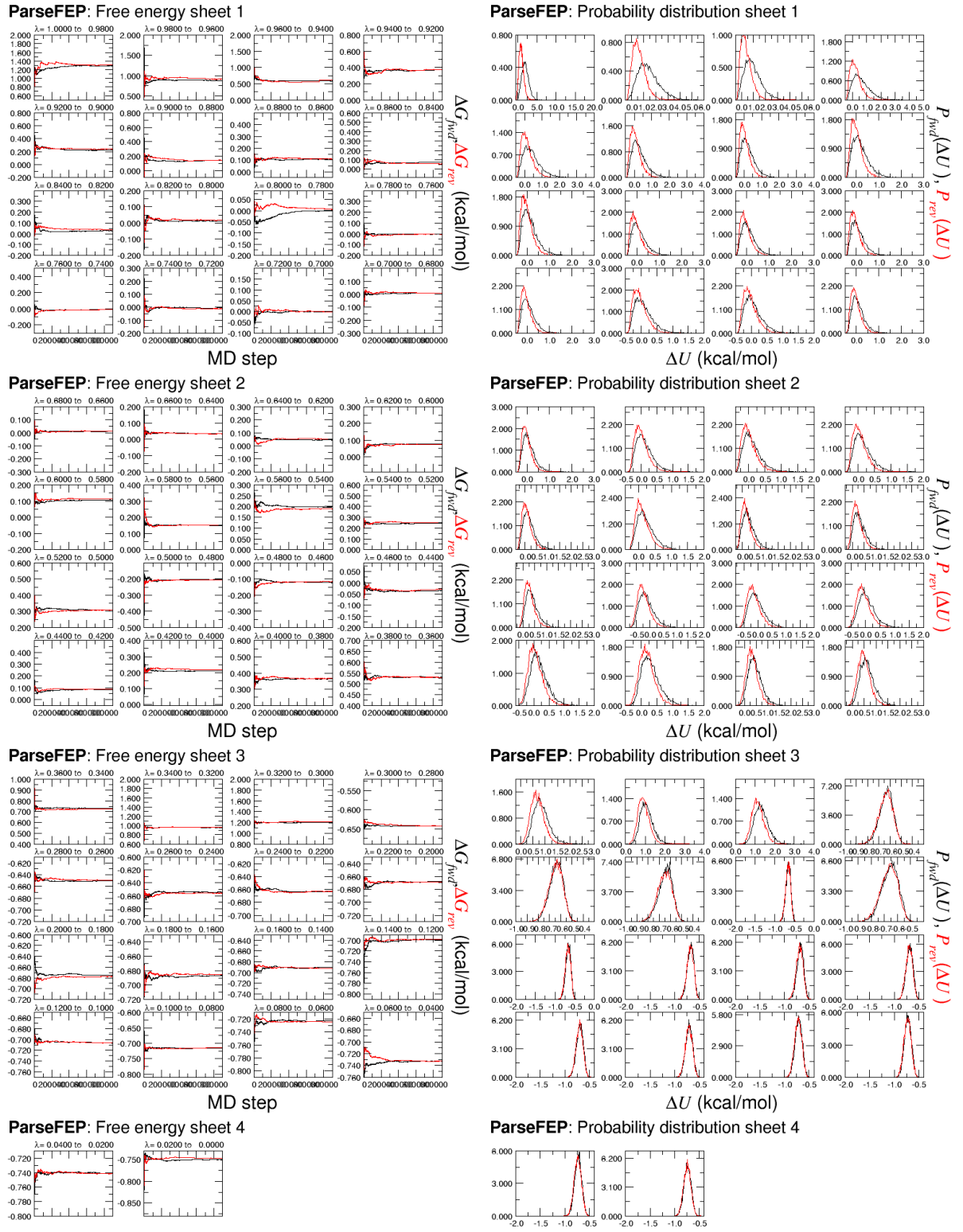

Figure 3.12: Free energies and probability distribution functions with hysteresis for coupling/ decoupling steps in FEP calculations for the unbound state of paraxylene generated by ParseFEP[88]

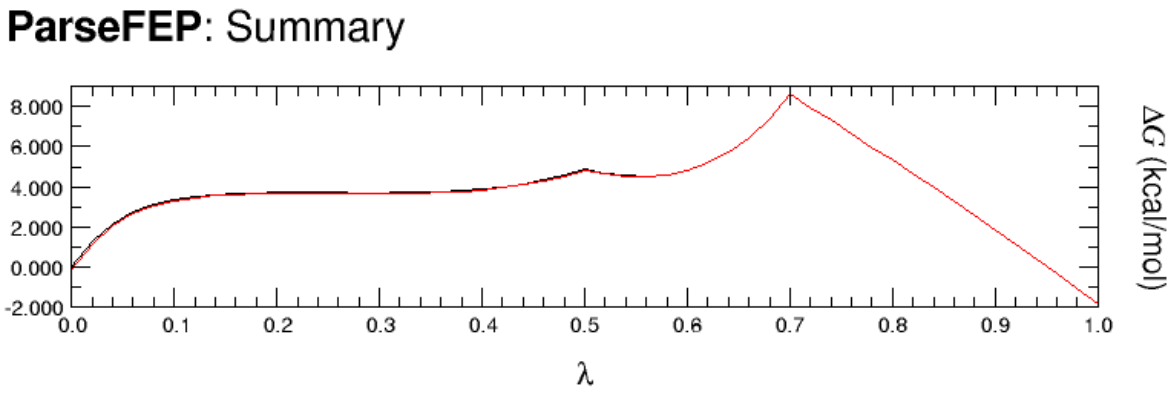

Figure 3.13: Summary of the full transformation for the reversible annihilation of the restrained paraxylene performed with FEP and obtained with the ParseFEP module of NAMD in the unbound state

The following table reports all the contributions to the computation of the final estimate, the speed, and the chosen stratification strategy.

Contribution	Alch. (kcal/mol)	Alch. (ns)	Timing (ns/day)	Stratification
$\Delta G^{site}_{decouple}$	-12.2 ± 0.3	200	24 (2h/window of 2ns)	50 windows
ΔG^{site}_{c}	-0.1 ± 0.0	60	28.83	15 windows
ΔG^{site}_{Θ}	-1.5 ± 0.0			
ΔG^{site}_{Φ}	-1.5 ± 0.0			
ΔG^{site}_{Ψ}	-1.4 ± 0.0			
ΔG^{site}_{θ}	-0.1 ± 0.0			
ΔG^{site}_{ϕ}	-0.1 ± 0.0			
ΔG^{site}_{r}	-0.4 ± 0.0			
$\Delta G^{bulk}_{decouple}$	1.8 ± 0.1	100	60(0.8h/window of 2ns)	50 windows
ΔG^{bulk}_{c}	-0.1 ± 0.0	30	107.35	15 windows
ΔG^{bulk}_{o+a+r}	11.2			
ΔG°_{b}	-4.3 ± 0.4	$\Delta G^{\text{exp}}_{b}$: -4.67 kcal/mol[87]		

Table 3.2: Detailed of the diverse binding free-energy contributions to the final ΔG°_{b} for the paraxylene.

The estimate obtained with our protocol falls really close to the experimental value. Furthermore, Deng et al.,[89] have obtained an estimate of $-$ 9.06 kcal/mol using a similar approach, which is further away from experimental value of -4.67 kcal/mol.[87] This results further demonstrate the potency of the protocol.

3.1.3 N-butylbenzne

This complex comprises the same Lyzozyme T4 L99A protein bound to n-butylbenzene and uses the same binding pocket as previously. The binding pose is represented in Figure 3.14.

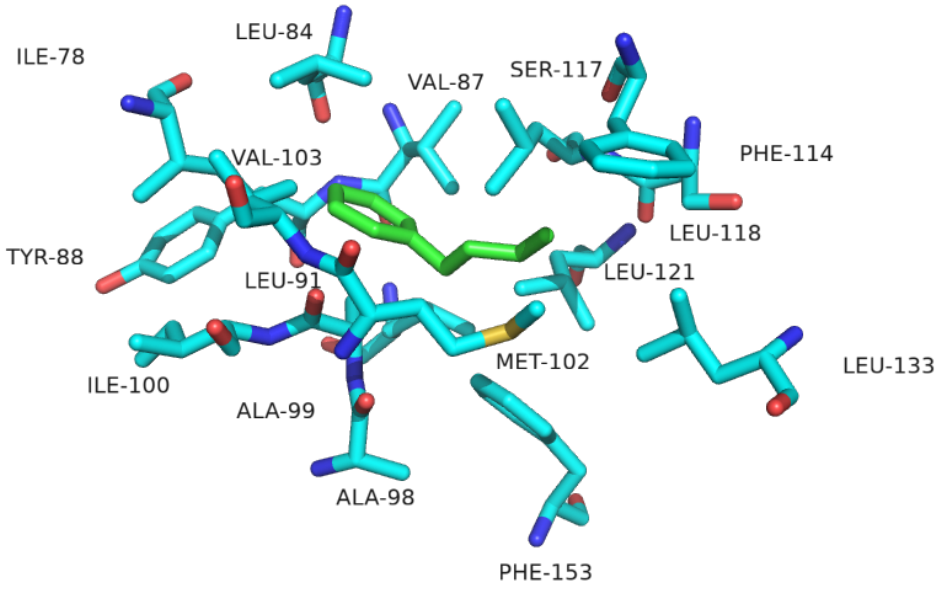

Figure 3.14: Binding pocket of n-butyl-benzene in lysozyme T4 L99A

The X-ray structure was taken from the PDB using the code 186l.[87] The crystallographic waters were retained to build the complex for the following simulations. The complex was then solvated and neutralize with Na^+, Cl^- counter ions, leading to a system composed of 36927 atoms (2604 atoms for the protein, 24 for the ligand, 11 430 water molecules, and 9 Cl- ions).The starting coordinates for BFEE2[81] were taken during equilibration when the ligand had the same conformation as in the crystallographic structure and saved into a PDB file using VMD[90]

In the case of n-butylbenzene, the estimated value was initially far off the experimental target value despite reasonable restraints convergence and very low hysteresis. Our strategy was to change the starting point and use the apo structure for the protein extracted from the end of the previously obtained FEP transformation. This trick allowed us to get closer to the desired value, only one kcal/mol away, and the detailed contributions are presented in Table 3.3 The need to use the apo structure can be explained by the slight deformation of the binding pocket located at helix F caused by the long carbon chain on the aromatic cycle, as demonstrated by Morton et al.[83]

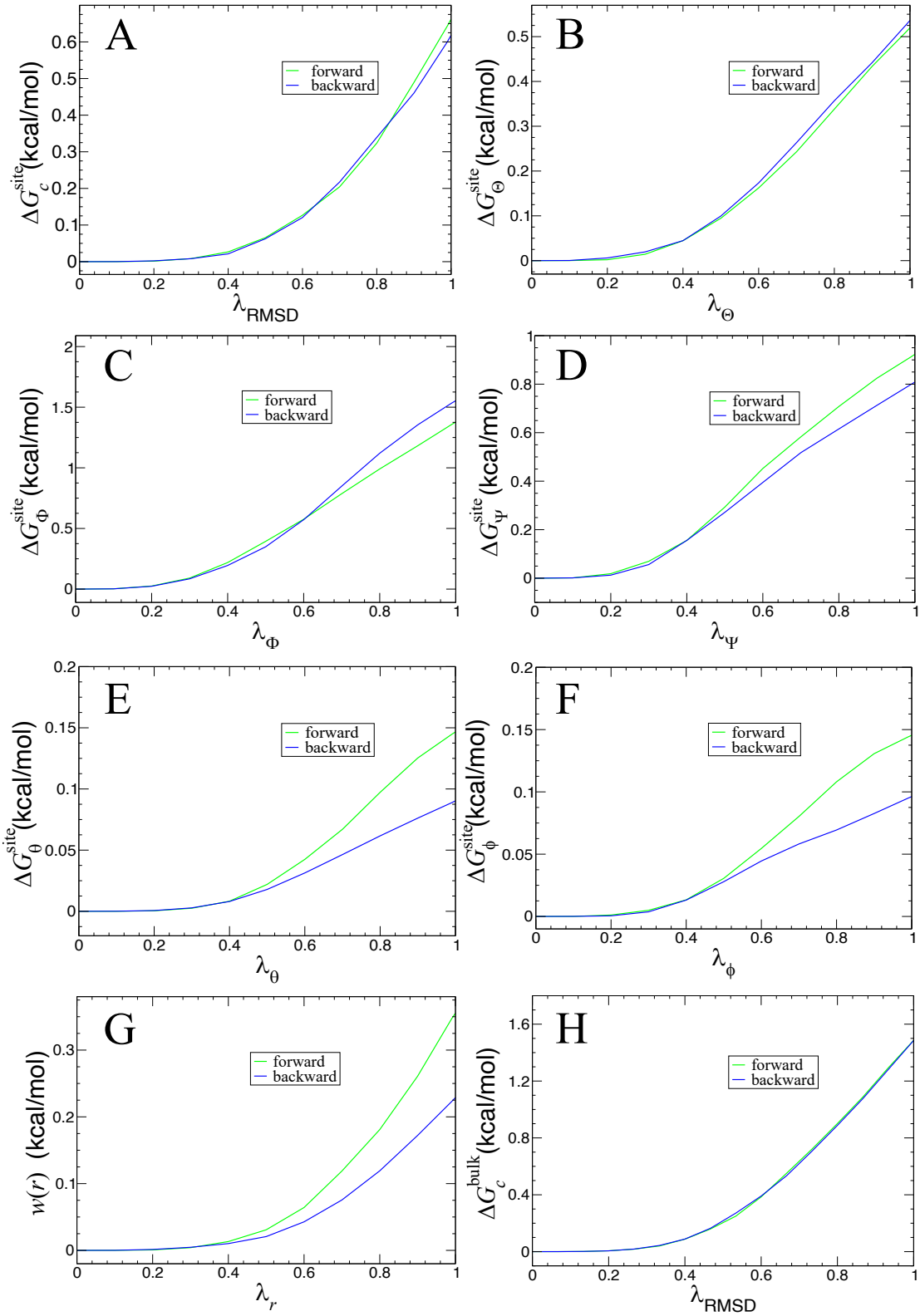

Figure 3.15: Restraints contributions for n-butylbenzene in the bound state, with (A) the RMSD, (B) Θ, (C) Φ, (D) Ψ , (E) θ, (F) ϕ and (G) the distance between COMs. (H) corresponds to the RMSD contribution in the unbound state.

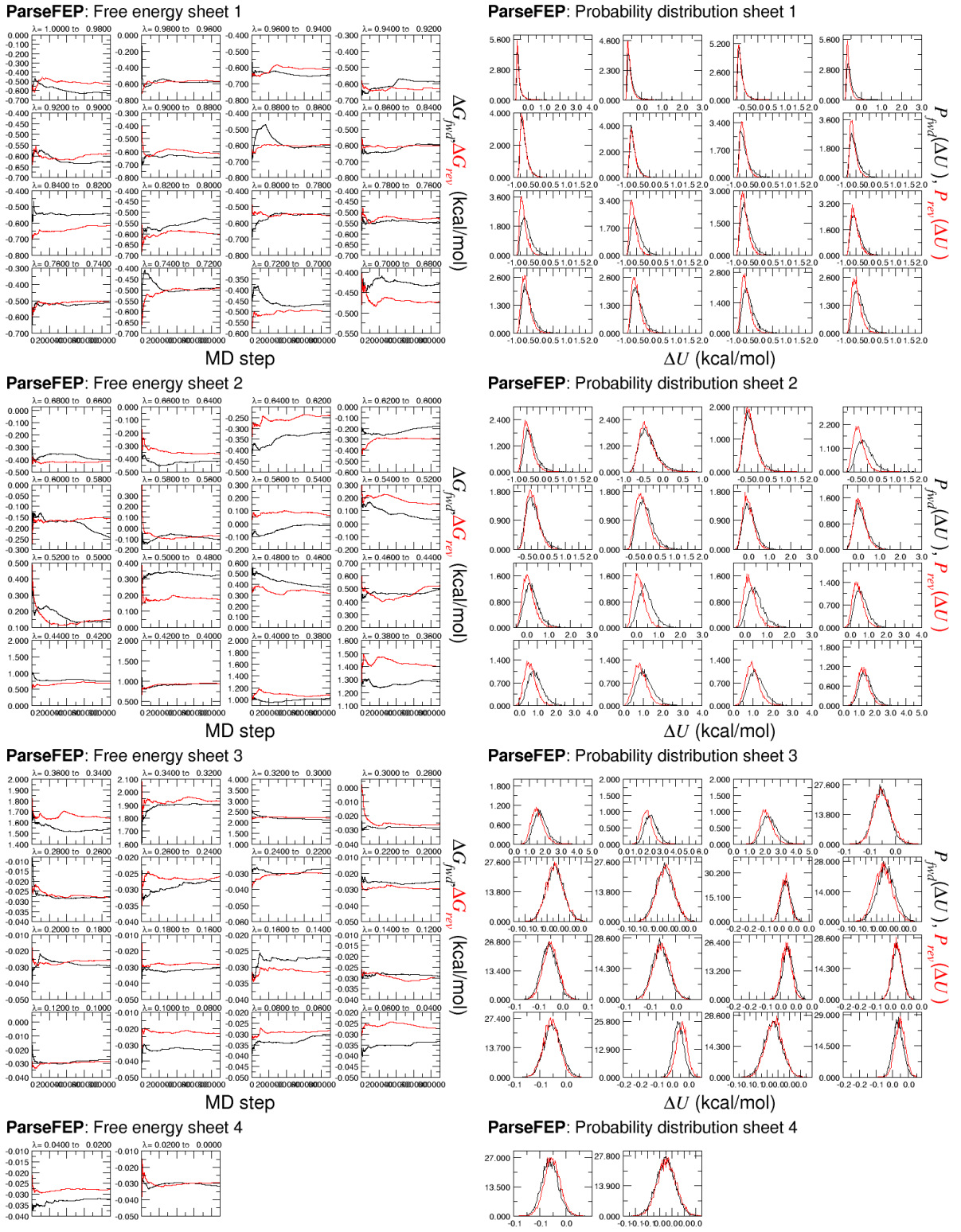

Figure 3.16: Free energies and probability distribution functions with hysteresis for coupling/ decoupling steps in FEP calculations for the bound state of n-butylbenzene generated by ParseFEP[88]

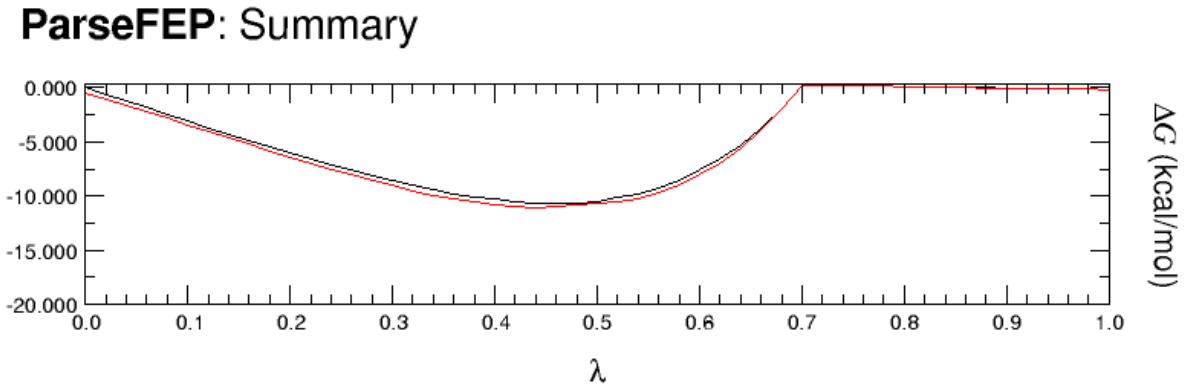

Figure 3.17: Summary of the full transformation for the reversible annihilation of the restrained n-butylbenzene performed with FEP and obtained with the ParseFEP module of NAMD in the bound state

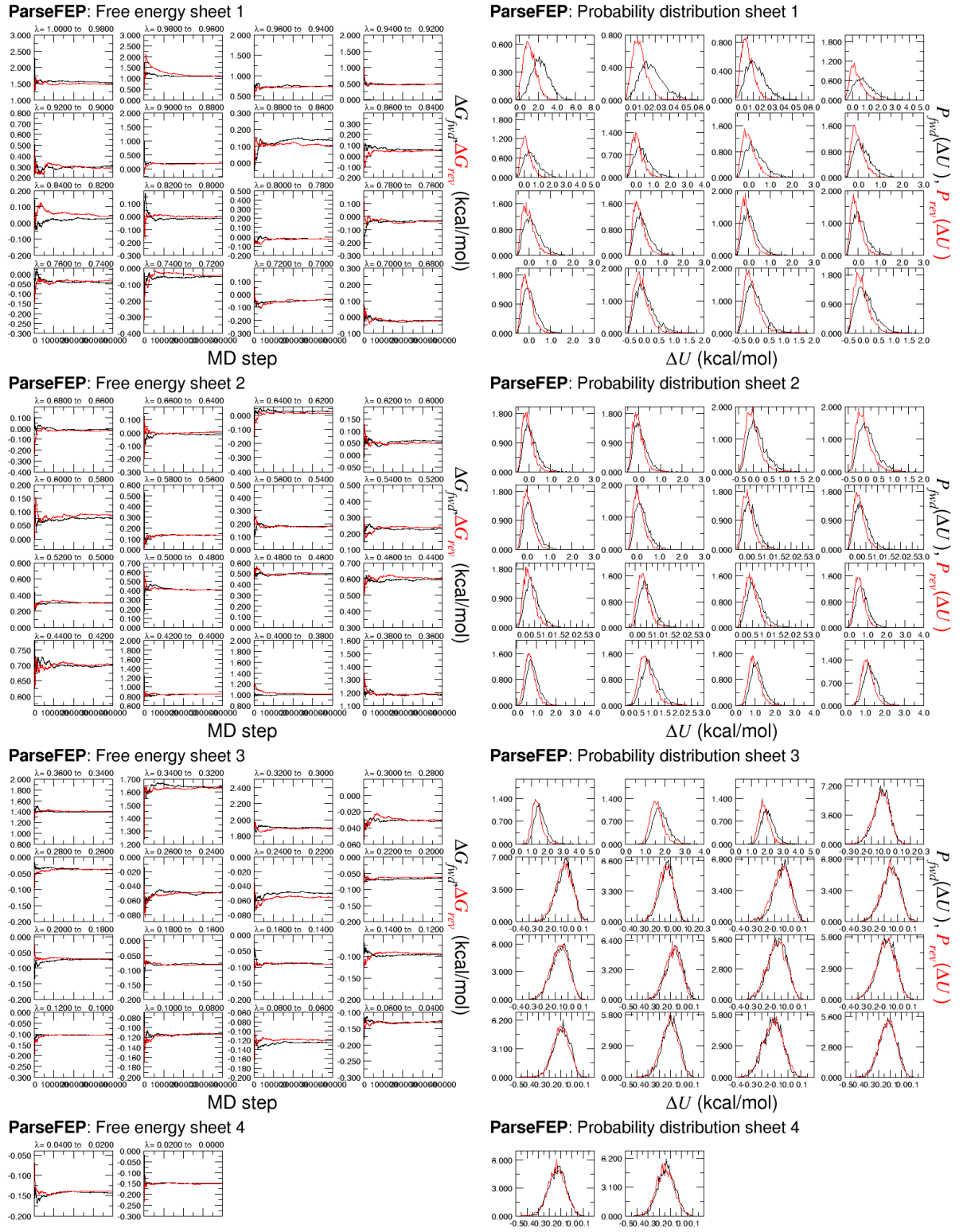

Figure 3.18: Free energies and probability distribution functions with hysteresis for coupling/ decoupling steps in FEP calculations for the unbound state of n-butylbenzene generated by ParseFEP[88]

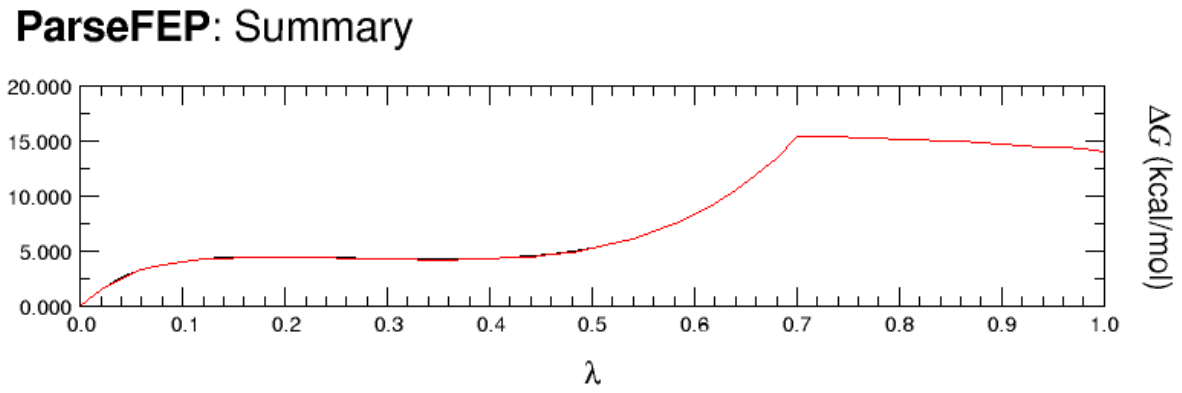

Figure 3.19: Summary of the full transformation for the reversible annihilation of the restrained n-butylbenzene performed with FEP and obtained with the ParseFEP module of NAMD in the unbound state

Contribution	Alch. (kcal/mol)	Alch. (ns)	Timing (ns/day)	Stratification
$\Delta G^{site}_{decouple}$	0.1 ± 0.4	200	24 (2h/window of 2ns)	50 windows
ΔG^{site}_{c}	-0.1 ± 0.0	60	28.83	15 windows
ΔG^{site}_{Θ}	-0.6 ± 0.0			
ΔG^{site}_{Φ}	-0.5 ± 0.0			
ΔG^{site}_{Ψ}	-1.5 ± 0.2			
ΔG^{site}_{θ}	-0.1 ± 0.1			
ΔG^{site}_{ϕ}	-0.2 ± 0.1			
ΔG^{site}_{r}	-0.4 ± 0.1			
$\Delta G^{bulk}_{decouple}$	-14.1 ± 0.1	100	60(0.4h/window of 1ns)	50 windows
ΔG^{bulk}_{c}	-1.6 ± 0.0	30	103.4	15 windows
ΔG^{bulk}_{o+a+r}	11.1			
ΔG°_{b}	-5.7 ± 0.4	$\Delta G^{\text{exp}}_{b}$: -6.70 kcal/mol[87]		

Table 3.3: Detailed of the diverse binding free-energy contributions to the final ΔG°_{b} for the n-butylbenzene

3.2 Summary

The protocol described in this article leverages the BFEE2 software[81] and can be summarized into four steps to obtain an accurate binding free energy estimate.

1. **Acquiring structure.** The initial step involves acquiring a three-dimensional structure of the system. It can either be from a database, such as the PDB,[91] a model, or obtained by docking. The structure has to be generated using a coordinates file and a topology file compatible with BFEE2 (NAMD[86] or GROMACS[92] files format)

2. **Input files generation.** This second step consists of generating all input files by the BFEE2 tool to run the diverse steps and simulation to obtain the binding free energy. In addition to selecting the geometrical or alchemical route, the user can fine-tune advanced features such as the pulling

direction in the reversible separation or the presence of a membrane in the system. In the alchemical route setup, the user can specify the lambda schedule in the TI part of the alchemical route, controlling the gradual transformation of the ligand or specifying a number of windows. These advanced features provide greater control and customization of the free-energy calculation process by the end user to ensure optimal sampling and accurate estimation of protein-ligand binding free energies.

3. **Running Simulations.** This step involves running the diverse MD simulations using the MD engine (NAMD[86] or GROMACS[92]), compatible with the use of Colvars.[93] During this step, the end-user is responsible for checking that the simulations have correctly sampled the proper CV space and are converged. To that end, BFEE2[81] provides additional modules in the Graphical User Interface(GUI). Furthermore, for the alchemical route, the ParseFEP,[88] module, part of the VMD analysis toolkit,,[90] can be used for the automated generation of probability distribution function (PDF) profiles and free-energy profiles. Visual inspection of PDF overlap and the hysteresis associated with the free-energy changes can be used to ascertain convergence.

4. **Post-treatment and final estimate.** BFEE2 is called to perform the post-treatments by analyzing the output files generated in the different MD simulations without human intervention to obtain the final binding affinity estimate.

This protocol encompassed several advantages: (i) Minimum human intervention through the use of BFEE2,[81] (ii) robust theoretical framework using the geometrical route or the alchemical route, (iii) wide range of applications (globular or membrane proteins, surface or buried/semi-buried ligands), and (iv) reproducibility of results is guarantee thought the standardization of CVs and free-energies calculations.

However, it remains limited by the force field accuracy and quality, as the calculations are based on MD simulations. Furthermore, the utilization of ligands with high conformational variability between the bound and the unbound state can be limited due to the chosen CV, RMSD, which might not be appropriate to characterize the isomerization of the ligand. Lastly, the high computational cost required to obtain converged ensemble in all the simulations remains a heavy limitation.

3.3 Original text

This paper enclosed a detailed protocol on how to run binding free energy calculation using the methodology developed in the group and used throughout my thesis. This methodology is applied to the two different ways discussed above, namely the geometrical and the alchemical route.

nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-021-00676-1

Check for updates

Accurate determination of protein:ligand standard binding free energies from molecular dynamics simulations

Haohao Fu[1], Haochuan Chen[1], Marharyta Blazhynska[2], Emma Goulard Coderc de Lacam[2], Florence Szczepaniak[2,3], Anna Pavlova[4], Xueguang Shao[1], James C. Gumbart[4], François Dehez[2], Benoît Roux[3,5,6], Wensheng Cai[1]✉ and Christophe Chipot[2,7,8]✉

Designing a reliable computational methodology to calculate protein:ligand standard binding free energies is extremely challenging. The large change in configurational enthalpy and entropy that accompanies the association of ligand and protein is notoriously difficult to capture in naive brute-force simulations. Addressing this issue, the present protocol rests upon a rigorous statistical mechanical framework for the determination of protein:ligand binding affinities together with the comprehensive Binding Free-Energy Estimator 2 (BFEE2) application software. With the knowledge of the bound state, available from experiments or docking, application of the BFEE2 protocol with a reliable force field supplies in a matter of days standard binding free energies within chemical accuracy, for a broad range of protein:ligand complexes. Limiting undesirable human intervention, BFEE2 assists the end user in preparing all the necessary input files and performing the post-treatment of the simulations towards the final estimate of the binding affinity.

Introduction

Complete understanding and prediction of the recognition and association of a protein and a ligand is of paramount importance in chemistry and biology, most notably in the field of protein engineering and pharmaceutical sciences. The binding free energy directly mirrors the ability of the ligand to interact with the protein, and is, therefore, regarded as the key quantity in studies of molecular recognition and association phenomena. However, the experimental determination of the binding affinity of any prospective compound is often costly in terms of synthesis time and money. To alleviate this important hurdle in drug design and lead optimization, much attention and effort have been devoted to development of computational methodologies for the accurate estimation of binding free energies in silico[1].

The main challenge posed by the estimation of a binding free energy by means of computer simulations is to capture the considerable change in configurational enthalpy and entropy corresponding to the conformational, orientational and positional movements of the ligand with respect to the protein in the course of their reversible association[2,3]. One of the strategies that has proven reliable consists in turning to the combination of molecular dynamics (MD) and advanced-sampling techniques to guarantee the adequate exploration of each relevant degree of freedom[1,4,5]. Still, the reversible association events of protein:ligand complexes are very difficult to capture in advanced-sampling simulations. A widely adopted strategy to overcome the sampling issue consists in introducing a restraining potential at an intermediate step to control the motion of the ligand relative to the protein binding site during the geometric separation or alchemical decoupling of the partner. This strategy, introduced by Hermans and Shankar[6], has been progressively enriched over the years by a number of additional developments and variants[7–12]. The restraining potential is introduced at one

[1]Research Center for Analytical Sciences, Frontiers Science Center for New Organic Matter, College of Chemistry, Nankai University, Tianjin Key Laboratory of Biosensing and Molecular Recognition, State Key Laboratory of Medicinal Chemical Biology, Tianjin, China. [2]Laboratoire International Associé CNRS and University of Illinois at Urbana-Champaign, UMR 7019, Université de Lorraine, Vandœuvre-lès-Nancy, France. [3]Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL, USA. [4]School of Physics, Georgia Institute of Technology, Atlanta, GA, USA. [5]Department of Chemistry, University of Chicago, Chicago, IL, USA. [6]Center for Nanoscale Materials, Argonne National Laboratory, Argonne, IL, USA. [7]Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [8]Theoretical and Computational Biophysics Group, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA. ✉e-mail: wscai@nankai.edu.cn; chipot@illinois.edu

Table 1 | CVs considered in binding free-energy calculations

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

end-point to 'confine' the uncoupled ligand within the binding site, and is then 'released' at the other end-point, where this step can be carried out analytically[6,7,10]. Gilson et al. called free-energy calculations in which there is no translational restraint the 'double annihilation method', and calculations in which there is a translational restraint the 'double decoupling method'[13].

Based on the idea of introducing restraints to confine the ligand with respect to the protein during separation or decoupling of the partner and then estimating the contribution to the binding affinity through post-treatments, since 1996, we have developed a number of numerical schemes[14–17], simulation strategies[7,18–21] and software[22,23] to facilitate in silico determination of the standard binding free energy. These successive developments are encapsulated in an automated, streamlined, general and accurate methodology[23] put forth to calculate the binding affinity of a flexible ligand with respect to a protein, as detailed hereafter.

Development of the protocol

All the algorithms and numerical strategies described below have been automated and implemented in the latest version of the Binding Free Energy Estimator 2 (BFEE2) open-source and user-friendly software[23], which can be used in conjunction with the popular visualization platform Visual Molecular Dynamics (VMD)[24].

Except perhaps for the simplest systems, unbiased brute-force MD is largely unable to sample the large conformational, translational and orientational changes that accompany the reversible association of a drug-like ligand with a protein. To circumvent this problem, we proposed to deliberately control the sampling through the use of configurational restraints and formulate the calculation in terms of geometric or alchemical transformations[18]. In 2013, it was shown that both routes lead to equivalent results[19]. Depending on the particular situation, features of either route may be advantageously exploited and combined for the accurate determination of the standard binding free energy between a drug-like molecule and a protein. Irrespective of the chosen route, restraints may be added and accounted for rigorously, using a set of template-based and generalized collective variables (CVs) describing the slow degrees of freedom of the reversible association (Table 1)[10,20].

Geometrical route

In the geometrical route, restraints are introduced one by one to progressively focus the conformational and orientational movements of the ligand with respect to the protein before their complete separation through a rectilinear pathway. The free energy associated with such a transformation of the object at hand can be expressed in terms of the potential of mean force (PMF). The contributions of the various restraints to the binding free energy are estimated via PMF calculations using WTM-eABF[15,16], a variant of the adaptive biasing force (ABF) algorithm (Fig. 1a)[25]. The theoretical underpinnings of the geometrical route are detailed in refs. [18,19].

Alchemical route

In the alchemical route, simulations are performed according to the thermodynamic cycle shown in Fig. 1b. The ligand is decoupled reversibly from its environment, i.e., the protein or the bulk, using the so-called alchemical free-energy perturbation (FEP) method, with its position, orientation and conformation restrained to those of the native state[26,27]. The energetic cost to enforce these restraints is then estimated through thermodynamic integration, zeroing out the associated force constant[28,29].

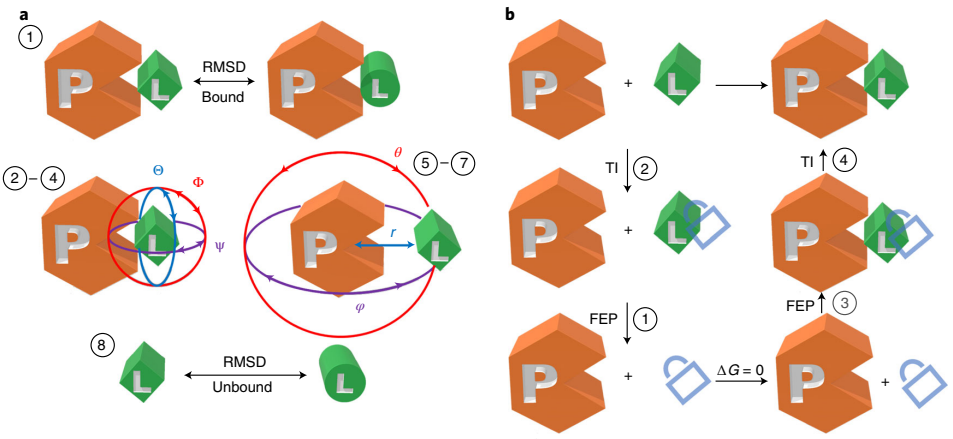

Fig. 1 | Illustration of the geometrical and alchemical routes. a, Geometrical route. **b**, Alchemical route. The numbers indicate the order of free-energy calculations set up using BFEE2. The lock represents the restraints applied to the conformational, orientational and positional degrees of freedom of the ligand with respect to the protein.

To improve the reliability of the free-energy estimates, the simulations are performed bidirectionally for each step depicted in Fig. 1b. The theoretical underpinnings of the alchemical route can be found in ref. [19].

Applications of the methodology

Since the first publication in 1996, our methodology has been applied to many molecular assemblies. For instance, we have accurately evaluated the binding affinity of three flexible decapeptides to the SH3 domain of the Abl kinase[19,20]. The standard binding free energy of the netropsin–DNA complex, wherein the conformational change of the host molecule is noteworthy, has been determined within chemical accuracy[30]. We have also employed our methodology blindly in protein engineering and shown that corannulene is more effective than perylene in inhibiting protein activity, which was later confirmed by experiments[31]. All in all, given the appropriate choice of a route, either geometrical or alchemical, our methodology has proven able to estimate the binding affinity of a large variety of protein:ligand complexes—including, but not limited to rigid and flexible ligands, ligands interred and lying at the surface of the protein, and combination thereof—with suitable accuracy.

Apart from the aforementioned applications, our methodology has been adopted by different research groups[32–37]. It is noteworthy that an independent benchmarking study has specifically underscored the remarkable reliability of both the geometrical and alchemical routes for the estimation of binding free energies[33]. Moreover, our methodology has been successfully used in force-field parameterization and validation, which requires very accurate free-energy estimates.[38] A selection of success stories of our methodology is given in Table 2.

Comparison with other methodologies

There is a variety of methodologies designed to estimate the absolute binding free energy of a ligand and a protein. Embodying different schools of thought, they have their own merits and drawbacks, to the extent that there is in general no optimal choice far superior to the others. Still, selecting the most appropriate methodology to address a specific problem remains of paramount importance.

Free-energy calculations based on an end-point approximation, like molecular mechanics Poisson–Boltzmann surface area (MM-PBSA)[39], only require equilibrium trajectories of the solvated protein:ligand complex and, thus, constitute an attractive option for high-throughput screening. The reliability of MM-PBSA free-energy estimates remains, however, highly uncertain due to the use of an implicit solvent and the inadequacy of the approximation to capture the effects of large conformational changes.

Funnel metadynamics[40], as a combination of geometric restraints with the metadynamics technique[41], offers a good balance between accuracy and efficiency and is particularly well suited for the study of rigid ligands lying at the surface of a protein. For flexible or buried ligands, or possibly the

Table 2 | Success stories of restraint-based binding free-energy calculations

Complex	PDB ID[a]	Error (kcal/mol)[b]	Remarks	Reference
Abl kinase-SH3:p41	1BBZ	0.5[c], 0.4[d]	Large, flexible ligand	[19,20] and this work
Abl kinase-SH3:p5	—	0.1	Large, flexible ligand	[20]
Abl kinase-SH3:p24	—	0.3	Large, flexible ligand	[20]
Human $p56^{lck}$-SH2:pYEEI	1LKK	~0.5[e]		[18]
FKBP12:ligand3	—	0.6		[21]
FKBP12:ligand5	—	2.1		[21]
FKBP12:ligand6	—	1.1		[21]
FKBP12:ligand8	1FKG	0.6		[21]
FKBP12:ligand9	1FKH	0.5		[21]
DIAP1-BIR1:Grim peptide	1SE0	0.8	Large, flexible ligand	[38] and this work
Trypsin:benzamidine	3ATL	~1.1[e]	Semi-buried ligand	This work
HBV Cp:NVR-010-001-E2	5E0I	1.6[f]	Semi-buried ligand	This work
T4 lysozyme L99A:benzene	4W52	0.8	Buried ligand	[11] and this work
T4 lysozyme L99A:ethyl benzene	1NHB	0.6	Buried ligand	[11] and this work
T4 lysozyme L99A:paraxylene	187L	0.4	Buried ligand	[11] and this work
T4 lysozyme L99A:n-butyl benzene	186L	1.0	Buried ligand	[11] and this work
Human CREBBP- Bromodomain: dihydroquinoxalinone	4NYX	0.3	Cation–π interaction dominates	[38]
Human CREBBP- Bromodomain: isoxazolyl-benzimidazole	4NR7	0.4	Cation–π interaction dominates	[38]
Factor Xa:cation inhibitor	2JKH	0.9	Cation–π interaction dominates	[38]
Factor Xa:cation inhibitor	2Y5H	1.1	Cation–π interaction dominates	[38]
Factor Xa:cation inhibitor	2Y5G	2.1	Cation–π interaction dominates	[38]
Factor Xa:quaternary ammonium	2BOK	0.3	Cation–π interaction dominates	[38]
DNA:netropsin	—	0.5	Large conformational change of the host	[30]
MDM2-p53:NVP-CGM097	4ZYF	0.5	Semi-buried ligand	This work
MUP-I:2-methoxy-3-isopropylpyrazine	1QY2	0.0	Buried ligand	This work
MUP-I:6-hydroxy-6-methyl-3-heptanone	1I05	0.5	Buried ligand	This work
β1-adrenergic receptor:4-methyl-2-(piperazin-1-yl) quinoline	3ZPR	1.0	Membrane protein:ligand complex	This work
V1-ATPase:ATP (tightly bound)	—	1.6[g]		[75]
V1-ATPase:ATP (bound)	—	1.1[g]		[75]
V1-ATPase:ATP (empty)	—	2.0[g]		[75]
V1-ATPase:ADP+Pi (tightly bound)	—	2.1[g]		[75]
V1-ATPase:ADP+Pi (empty)	—	2.2[g]		[75]

[a]A dash means that crystal structure is either not available or not used in the free-energy calculation. [b]Unsigned errors are provided. If the binding free energy of a complex is recalculated as an example in this study, the error of this calculation is provided. [c]Alchemical route. [d]Geometrical route. [e]Different experimental values are available. The average of them is used to calculate the errors. [f]Experimental value is not available. Simulation results in ref. [78] are used as the reference. [g]The experimental estimates were obtained with F_1-ATPase. Detailed experimental and estimated binding free energies are provided in Supplementary Information

combination thereof, one may, however, face convergence issues and difficulty to select an appropriate set of CVs to describe the relative movement of the ligand with respect to the protein. A useful companion guide to funnel metadynamics can be found in ref. [42].

Several alternative methodologies are inherently similar to the strategy presented in this protocol. For example, the double decoupling method[13] and the alchemical route can be regarded as variants of each other. Conversely, attach–pull–release[37] is akin to the geometrical route, which separates the protein and the ligand by means of a PMF calculation. Confine and release[43] uses PMF calculations to account for conformational changes of the protein while performing alchemical FEP to decouple the ligand from the protein. One particular advantage that distinguishes our methodology amid its analogs is the streamlining and automation of the complete free-energy calculation process, from the input generation to the post-treatment, by means of a user-friendly software package, BFEE2[23].

Standard binding free energies can be also evaluated through direct simulation of the reversible association using advanced-sampling techniques such as Gaussian accelerated MD[44] and replica

exchange with solute scaling[45]. Such approaches do not require a priori knowledge of the native state of a protein:ligand complex and CVs describing the slow degrees of freedom of the reversible association. No simulation strategy relying on spontaneous events, however, can guarantee that the correct binding model can be sampled reversibly within the timescale amenable to MD simulations.

Advantages and limitations of this protocol

Advantages

- Theoretically rigorous. Some approaches introduce various approximation to trade accuracy for efficiency, just like MM-PBSA[39]. We guarantee that our methodology, which builds upon a theoretical framework introduced in 2005[18], is formally rigorous, to the extent that free-energy calculations following this protocol are expected to converge within force-field accuracy
- Available for a wide range of protein:ligand and host–guest complexes because their nature, namely globular versus membrane protein, buried versus semi-buried ligand and ligand lying at the surface of the protein, rigid versus flexible ligand, is taken into consideration in our methodology and software
- Able to circumvent the difficulty of sampling configurational space. In our methodology, either one-dimensional PMF calculations or alchemical transformations are performed, with a reduced configurational space to sample by virtue of the introduction of geometric restraints. This methodology avoids the dimensionality issue in multidimensional free-energy calculations and obviates the need to capture the large change in protein:ligand configuration by means of advanced-sampling simulations
- Minimal human intervention. Our implementation of the methodology in a user-friendly software, BFEE2, streamlines the overall standard binding free-energy calculation. From the definition of the CVs, the preparation of the different input files, to the post-treatment of the simulations, each step is automated in BFEE2, hence minimizing human intervention. Still, monitoring the trajectories and tuning up the simulation parameters as a function of the protein:ligand complex at hand, e.g., simulation length, is advisable
- Easy assessment of convergence. Convergence of the PMF calculations, or of the alchemical transformations, can be directly assessed through graphical user interface (GUI)-based tools
- Robustness and reproducibility of results. Since the definition of the CVs (Table 1) and free-energy calculations (Fig. 1) are both standardized[19,20], the protocol will yield the same binding affinity for a given protein:ligand complex in replicated simulations

Limitations

- Force-field dependence. The overall accuracy—as opposed to the statistical precision[25,46]—of binding free-energy calculations rests upon the quality of the force field utilized, as our methodology is based on MD simulations. The end user is, therefore, strongly suggested to have some level of expertise in the selection and validation of force fields, especially for drug-like molecules. This force-field dependence exists in all MD-based methodologies aimed at estimating binding affinities. To make this protocol self-contained, we explain hereafter how to set up molecular assemblies using the CHARMM[47] and Amber[48] force fields. It should be made very clear that our method itself is rigorously formulated, independently of the force field, as opposed to approximate schemes like MM/PBSA, and hence, the users can improve the reliability of the standard binding free-energy calculation by turning to a more accurate model (e.g., polarizable force fields such as Drude[49] and AMOEBA[50]) if necessary
- Extremely deeply buried ligands. Our methodology can be employed to determine the binding affinity of ligands buried in a protein. For deeply buried ligands, however, capturing the solvent reorganization, that is, water entering and exiting the binding site as the substrate is decoupled reversibly from it, requires extensive simulation times, which might induce convergence issues in the free-energy calculations. In some cases, this issue can be circumvented by treating some water molecules as part of the protein or the ligand. Under the premise that the exchange of water inside and outside the binding site is essential, MD experts may want to turn to advanced-sampling techniques, such as REST2[45,51] or Monte-Carlo based methods[52,53], combined with FEP to address this issue. These approaches are, however, not yet proposed in BFEE2, and the users must modify manually the inputs generated by the software
- Ligands with notable conformational variability between the bound and the unbound state. In many instances, the bound- and unbound-state conformations of the ligand may be different. In the workflow presented below, the root-mean-square deviation (RMSD) with respect to the bound-state conformation is used as a CV to characterize the conformational change of the ligand. Although this choice is anticipated to be reasonable for most drug-like molecules, the conformation of some ligands,

such as those with ribose ring puckering, may not be sensitive to the change of the RMSD with respect to its bound-state conformation. Under such circumstances, the experienced end user may want to change the default definition of the RMSD to another CV that can characterize the isomerization of the ligand, adapting the relevant configurational files generated by BFEE2. Still, the examples presented below are indicative that the use of the RMSD with respect to the bound-state conformation is sufficient to describe the conformational change of ligands as flexible as a heptapeptide (DIAP1-BIR1: Grim peptide) and proline-rich decapeptides (Abl kinase-SH3:p41/p5/p24)

- Computational cost. Our methodology is by and large more expensive than those based on end-point approximations, which is the price to guarantee reliability of the free-energy estimates. Typically, a timescale of a few microseconds, possibly a few hundreds of nanoseconds, is sufficient to determine the binding affinity of a protein:ligand complex with the desired accuracy. In practice, with the development of graphics processing unit (GPU)-based architectures and GPU-accelerated MD engines, microsecond simulations are now routinely performed in the field of drug design. Furthermore, it is worth noting that some of the subprocesses (Fig. 1) in our methodology can be performed in parallel, thereby taking advantage of multi-GPU machines to reduce the total wall-clock time devoted to free-energy calculations. It should be clarified that a priori estimation of the computational time needed to overcome possible bottlenecks, or kinetic traps, such as conformational rearrangements in the binding pocket, is not always possible. Hence, additional computational effort may be sometimes required to guarantee suitable convergence of the simulations

Other points

- Requirement of the native binding motif. It is difficult to ascertain how resilient the methodology is to an assumed initial binding pose. Since the free-energy method rests on configurational sampling, in principle, this procedure by itself has the ability to explore the configurations in the neighborhood of the assumed pose. Simulations may either lead to an improved binding pose or lead to a completely different one. If the assumed initial pose is very inaccurate, the outcome is more uncertain. However, any alternative simulation strategy that does not rely on prior knowledge of the native binding pose implicitly assumes that it has the ability to discover the latter from scratch, which represents a considerable computational effort of its own. In practice, we believe that a sufficiently accurate approximation of the native state can be obtained at a reduced computational cost by means of molecular docking when no experimentally determined structure is available

Overview

In this protocol, we first introduce both the geometrical and alchemical routes for the calculation of the standard binding free energy of a flexible ligand, p41, to the SH3 domain of tyrosine kinase Abl[19] using NAnoscale Molecular Dynamics (NAMD) as our MD engine (see Box 1 for the use of Gromacs as the MD engine). To demonstrate the performance of our methodology when applied to different classes of protein:ligand complexes, we provide the following additional examples: (i) DIAP1-BIR1: Grim peptide, illustrating the use of Gromacs as the MD engine (Box 2), (ii) trypsin:benzamidine, the application of the geometrical route to a semi-buried ligand (Box 3), (iii) β1-adrenergic receptor: 4-methyl-2-(piperazin-1-yl) quinoline, the study of a membrane protein (Box 4), and (iv) Factor Xa: quaternary ammonium, analyzing the driving force underlying the protein:ligand association (Box 5).

Estimating protein:ligand standard binding free energies using our methodology includes four stages, irrespective of the chosen route, namely,

1. Modeling. This stage consists in preparing the structure and topology files readable by MD engines, which can be carried out by modeling tools such as CHARMM-GUI[54] and AmberTools[55].
2. Input files generation. This stage consists in preparing all the configurational files for the multistep free-energy calculation, automated in BFEE2.
3. Simulation. This stage consists in carrying out the different MD simulations. This stage requires human intervention to monitor the convergence of the latter, and, if need be, tune up selected parameters. Molecular visualization software, like VMD[24], is required to visualize the trajectory. Part of the convergence analysis is available in the latest version of BFEE2 and in ParseFEP[56].
4. Post-treatment. This stage consists in calculating the binding affinity on the basis of the output files generated in the different MD simulations. Bookkeeping and evaluation of the different configurational integrals are handled by BFEE2.

For first-time users of BFEE2, the additional step of software installation is evidently required, as detailed in the Materials section. After installation, we suggest that the end user start learning from

Box 1 | Analysis of a PMF calculation

In this box, we show how to check the convergence of a PMF calculation and extend the simulation if convergence is not achieved. We assume that the reader has already completed at least one PMF calculation introduced in the main text.

Procedure ● Timing 10 min

1 Open the BFEE2 plug-in. Switch to the 'Post-treatment→Quick-plot' tab. Load the `*_abf_i.abf1.czar.pmf` files (if a stratification strategy is used, all the `*.pmf` files should be loaded) into the 'Plot (stratified) PMFs' section, and click the 'Plot' button to display the PMF.

2 Convergence of a PMF calculation manifests itself in an approximate plateau of the time evolution of the PMF RMSD with respect to its initial value (usually a zero vector). To plot this curve, load `*_abf_i.abf1.hist.czar.pmf` into the 'Calculate PMF RMSD convergence' section, and click the 'Plot' button (see Fig. 4 as an example). If a stratification strategy is used, the convergence of each window should be analyzed independently.

3 If a stratification strategy is used, we strongly advise checking the continuity of the gradient across adjacent windows. To achieve this, open `*_abf_i.abf1.czar.grad` and `*_abf_(i+1).abf1.czar.grad` using a text editor and look at the last value of the gradient in the former and the first value in the latter. One can also plot the gradient curve following the procedure of step 1, loading `*_abf_i.abf1.czar.grad` instead.

4 If convergence is not satisfied, use `*_abf_i.extend.conf` to extend the simulation. By looking at the time evolution of the PMF RMSD with respect to its initial value, one can estimate the time required to achieve convergence and set an appropriate simulation time in `*_abf_i.extend.conf` before running the simulation.

It should be noted that, in some cases, simply adding simulation time may not guarantee convergence. Instead, the result deteriorates with the increase of simulation time. Under these premises, we suggest analyzing the trajectory to see whether partial denaturation of the protein has occurred as a result of introducing biasing forces. If so, the end user may want to improve the initial structure of the simulation, use a stratification strategy, or add apposite restraints to maintain the proper conformation of the protein. Increasing the value of `fullSamples` in *.in files has also proven to help reduce deleterious nonequilibrium effects through accruing more samples and providing a more accurate estimate of the initial biasing force prior to applying it[79].

Box 2 | Use BFEE2 with Gromacs

Apart from NAMD, BFEE2 supports binding free-energy calculations through the geometrical route using Gromacs[58] and is, therefore, compliant with different schools of thought for the determination of protein–ligand binding affinities. Some steps in the preparation of the required input files have to be carried out manually, however, since third-party software, such as VMD and MDAnalysis cannot directly generate files in the desired Gromacs format.

Procedure ● Timing depends on the available computational resources

1 Compile Gromacs with the Colvars support; see Supplementary Information for more details.

2 Open the BFEE2 plug-in. Switch to the 'Pre-treatment→Gromacs' tab. Load the topology (TOP) and structure (PDB) files of the solvated complex and ligand, respectively.

3 Set the temperature of the simulation, and define the protein and ligand moieties (following the MDAnalysis syntax).

4 Perform the following simulations sequentially. Each job must be submitted manually only after the previous one is completed.

```
000_eq/000_eq.mdp
001_RMSD_bound/001_PMF.mdp
002_euler_theta/002_PMF.mdp
003_euler_phi/003_PMF.mdp
004_euler_psi/004_PMF.mdp
005_polar_theta/005_PMF.mdp
006_polar_phi/006_PMF.mdp
007_r/007_Minimize.mdp
007_r/007_Equilibration.mdp
007_r/007_PMF.mdp
008_RMSD_unbound/008_Equilibration.mdp
008_RMSD_unbound/008_PMF.mdp
```

Note that the definition of `centers` for the restraints declared in the `*.in` files must be revised after the PMF calculation over an angle, Euler or spherical coordinate, is completed. See Steps 7–12 of Procedure 1 for the detail of setting the `centers` keyword and understanding each step. Compilation of the configuration file is required before running a simulation using Gromacs. We have automated the compilation and update of `centers` by making shell (`*.sh`) scripts available to the end user. If a stratification strategy is used, however, the end user is invited to update the input files manually.

5 Perform the post-treatment corresponding to Steps 15–17 of Procedure 1. Note that the force constants ought to be set to the corresponding values in the `*_colvars.dat` files, as the unit used internally by Gromacs is different from that used by NAMD.

We have provided examples of input files (PDB ID: 1SE0 and 1BBZ) for Gromacs in Supplementary Data to be tested by the end user.

the Abl-SH3:p41 illustration to grasp the gist of the methodology, focusing specifically on the 'Simulation' stage, which requires the greatest human intervention and expertise. Each example mentioned in this protocol, other than Abl-SH3:p41, highlights a specificity that ought to be paid attention to in practice. We also provide a lookup table to troubleshoot possible issues encountered while applying this protocol.

Box 3 | Handling a semi-buried ligand through BFEE2

As explained in the main text, the default direction along which the ligand is separated reversibly from the protein is the vector connecting their respective centers of mass. If the ligand is semi-buried, the end user may want to choose another direction to separate the two partners. We want to emphasize here that this treatment is only useful for the geometrical route—the alchemical route can be adopted without any specific setting, irrespective of the ligand being buried or not.

Procedure ● Timing depends on the available computational resources

1 Open the structure file of the protein:ligand complex (PDB) with VMD. Observe the structure of the complex, and find a possible path along which the barrier of separating the ligand from the protein is minimal. Then, define the path as the line connecting the center of mass of a manually chosen moiety and that of the ligand. This step is critical, and does directly affect the convergence of the binding free-energy calculation.
2 Follow Steps 1–4 of Procedure 1.
3 Define the moiety chosen above as the reference by setting 'User-defined separation direction→Reference' (following the MDAnalysis syntax) in the 'Advanced settings' menu.
4 Follow Steps 6–12 and 15–17 of Procedure 1.

The end user is required to monitor the trajectory of the simulation characterizing the separation of the ligand from the protein to ensure that the selected path is appropriate. A suboptimal path may lead to fraying—possibly partial denaturation of the protein in the course of separation. To reduce the trial-and-error overhead, the user can run `007_r/007.1_eq.conf` and `007_r/007.2_abf_1.conf` and examine the selected path upstream from the formal standard binding free-energy calculation.

Box 4 | Handling a membrane-protein:ligand complex through BFEE2

BFEE2 can be used to calculate the standard binding affinity of a ligand towards a membrane protein. If the geometrical route is adopted, the end user should ascertain that the ligand is appropriately separated from both the protein and its membrane environment by selecting a suitable direction of the translation of the ligand. Should this requirement be overly difficult to satisfy, the alchemical route is preferred. Such is the case of ligand deeply buried in the binding pocket of the membrane protein, e.g., in a G-protein-coupled receptor.

Procedure ● Timing depends on the available computational resources

1 During modeling, ensure that the membrane is perpendicular to the *z* axis to comply with the `flexibleCell` option of NAMD.
2 Select either the geometrical (Procedure 1) or the alchemical (Procedure 2) route. Follow Steps 2–4 of Procedure 1 or Step 1 of Procedure 2.
3 Check 'Model→Membrane Protein' in the 'Advanced settings' menu. If necessary, define an appropriate separation direction by setting 'User-defined separation direction→Reference', as detailed in Box 3.
4 Perform a very long equilibration to ensure proper contact of the lipids with the membrane protein, and proper hydration of the protein's soluble parts, which can be achieved by manually increasing the simulation time in `000_eq/000_eq.conf` and `007_r/007.1_eq.conf` (geometrical route) or `000_eq/000.1_eq.conf` and `000_eq/000.2_eq_ligandOnly.conf` (alchemical route). Follow Step 7 of Procedure 1 or Step 4 of Procedure 2.
5 Complete the free-energy calculation by following the remaining steps detailed in Procedure 1 or 2.

We want to emphasize that thorough exploration of the hydration state of the binding site is challenging, as the exchange of water molecules inside and outside the membrane protein is slow. One can equilibrate the membrane protein and its environment without the ligand, allowing water to diffuse inside the protein as a preamble to the PMF calculations—or alchemical transformations, should it prove necessary. During the free-energy calculation, sampling in each window should be sufficient to allow the reversible hydration of the ligand and its binding site to be captured. More elaborate techniques, relying, for instance, on grand-canonical MD, fall beyond the scope of the present contribution[52,53,80].

Expertise needed to complete the protocol

The end user is expected to know how to run MD simulations using either NAMD[57] or Gromacs[58]. Moreover, some experience of CV-based free-energy calculations using Colvars[29] is desirable, though not mandatory. Complete knowledge of the structural detail to define the slow degrees of freedom underlying the reversible association of the protein:ligand complex is in principle not necessary, but the end user is expected to know the experimental conditions of the three-dimensional structure determination, that is, the ionic strength and pH, which are crucial for setting the correct protonation state of the protein and of the ligand and carrying out the simulations with the appropriate salinity.

Experimental design

The workflow of our methodology is shown in Fig. 2 and explained below.

Modeling

The end user is required to generate structure and topology files readable for MD engines, starting ordinarily from a PDB ID, should there be an experimentally determined three-dimensional structure available, or from a PDB-formatted file obtained using molecular docking. Although we provide an introduction on the use of CHARMM-GUI[54], a web-based server, for setting up simulations involving

Box 5 | Analyzing the enthalpic driving force underlying the protein:ligand association

The pair-interaction calculation feature available in NAMD can be utilized to estimate the interaction energy and forces between any two moieties of the computational assay over an MD trajectory. The enthalpic driving force underlying protein:ligand association can, therefore, be analyzed with exquisite detail in this post-hoc treatment of the trajectory describing the separation of the ligand from the protein. This feature is only available when the geometrical route is followed, as no geometric pathway is determined in the alchemical route.

Procedure ● Timing depends on the available computational resources

1 Generate all the input files for a binding free-energy calculation following the geometrical route, as documented in Steps 1–6 of Procedure 1.
2 Reduce the value of `dcdFreq` in `007_r/abf_*.conf`, and set `colvarsTrajFrequency` in `colvars_*.in` to that of `dcdFreq`. This step makes NAMD write the trajectory file more frequently than the default, as the trajectory will be used subsequently for analysis purposes. The choice of an appropriate value of `dcdFreq` is subservient to the available disk space. A `dcdFreq` of 500 may be acceptable in many cases.
3 Complete the free-energy calculation following Steps 7–12 introduced in Procedure 1.
4 Prepare a PDB file, indicating the interaction of which two moieties will be calculated by setting the beta value of the two moieties as 1 and 2, respectively. If the self-interaction energies of a moiety are required, set the beta value of the moiety as 1.
5 An example of the configuration file of pair-interaction calculation is provided in Supplementary Information. Run the pair-interaction calculation just like a standard NAMD job. It should be noted that, to this date, GPU acceleration is not supported for this task. A CPU version of NAMD is required.
6 The van der Waals and electrostatic interaction energies and forces between the two selected moieties are written for each frame of the trajectory in the NAMD output (`log`) file. The corresponding value of the CV, i.e., protein:ligand center-of-mass distance, *r*, for each frame can be found in the `output/*.colvars.traj`. Hence, the user can calculate the average interaction energy and force between the two selected moieties in each bin. The end user can then either get the contribution to the PMF of the interaction of the two moieties by integration of the force along the CV or analyze the pair-interaction energy profile directly. Some smoothing is sometimes necessary to remove the noise in the obtained profiles due to lack of data in the saved trajectories.
7 Repeat Steps 4–6 for any interesting moiety pairs.

An example of the analysis of the enthalpic driving force underlying protein:ligand association is provided in Fig. 11 (PDB ID: 2BOK). In this case, the association of the protein and the ligand, promoted by a cation–π interaction, is hindered by the desolvation of a quaternary ammonium, as the latter is highly hydrophilic. In stark contrast, the protein:water interaction is favorable for association, as the aromatic moieties of the protein are shielded from the solvent. Interestingly enough, when the ligand is very close to the protein, the protein:water interaction becomes unfavorable, as the charged quaternary ammonium perturbs the network of hydrogen bonds formed between the protein and the water. The balance of the aforementioned contributions determines the optimal distance separating the protein from the ligand in the bound state.

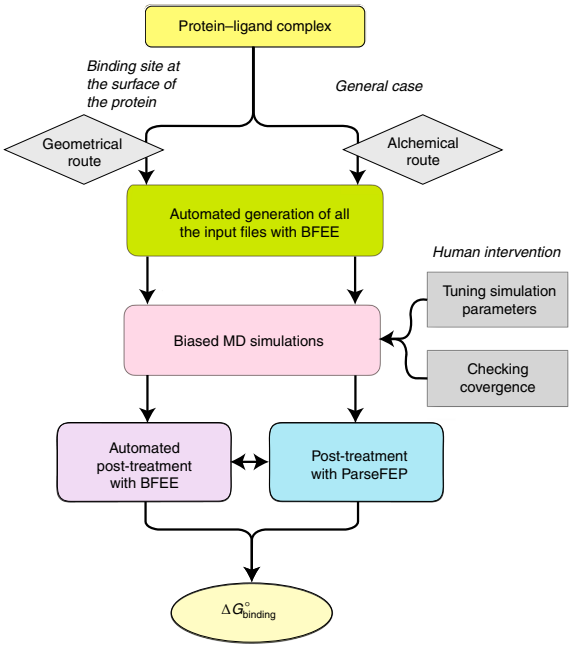

Fig. 2 | Workflow of our methodology. The light-yellow box represents the 'Modeling' stage (Procedures 1 and 2, Step 1); the green box, the 'Input files generation' stage (Procedure 1, Steps 2–6 and Procedure 2, Steps 1–3); the pink box, the 'Simulation' stage (Procedure 1, Steps 7–12 and Procedure 2, Steps 4 and 5); and the purple and cyan boxes, the 'Post-treatments' stage (Procedure 1, Steps 13–15 and Procedure 2, Steps 6–8).

Box 6 | Analyzing alchemical free-energy calculations using ParseFEP

ParseFEP is a powerful tool for the analysis of FEP calculations. It can be used to improve the reliability of the free-energy estimate and detect possible convergence issues in FEP calculations. The use of ParseFEP as a post-hoc tool is, therefore, highly recommended in the standard binding free-energy calculation that follows an alchemical route. Note that ParseFEP only supports Linux and MacOS operating systems, wherein XMGrace can be installed[56].

Procedure ● Timing 20 min

1 Complete the binding free-energy calculation following the procedure introduced in the main text.
2 Open ParseFEP through 'VMD→Extensions→Analysis→Analyze FEP simulation'.
3 Load `fep_forward.fepout` and `fep_backward.fepout` into ParseFEP.
Set parameters, as detailed in https://www.ks.uiuc.edu/Research/vmd/plugins/parsefep/. We suggest using the Bennett acceptance ratio estimator[63].
Click 'Run FEP parsing', and wait for the computation to complete in the background. The free-energy estimate and statistical error will be displayed in the VMD terminal.
It is noteworthy that ParseFEP will output a series of figures (Fig. 10). For the probability distributions of the perturbation, ΔU, from a bidirectional calculation, the end user is recommended to ascertain that the forward and the backward simulations sample a similar configurational space, mirroring the microstate reversibility of the transformation at hand. The end user can also monitor the evolution of the free-energy estimates in each window to verify the convergence of the bidirectional transformation[61].

protein:ligand complexes (Procedure 1 and 2, Step 1), discussing the detail of the molecular-modeling stage prefacing the determination of the binding affinity falls beyond the scope of this protocol.

Input-file generation

In this stage, all the configurational files required for a complete binding free-energy calculation are generated using BFEE2. At this stage, the end user must determine whether the geometrical or alchemical route will be followed. Whereas the former is only germane to complexes in which the ligand lies at the surface of the protein, the latter is suited to any protein:ligand complex. We provide in this protocol examples for both routes (Procedure 1, Steps 2-6 and Procedure 2, Steps 1-3).

Although the generation of input files is almost completely automated in BFEE2, given that the appropriate route is selected as a function of the nature of the protein:ligand complex at hand, the end user may find it necessary to tune some parameters, most notably the simulation length and whether or not a stratification strategy[25] ought to be used, which requires some prior experience with MD simulations. To give the non-expert an idea of how to tune simulation parameters, we detail the simulation times and stratification strategies for all the examples reported in this protocol and rationalize these settings (Procedure 1, Step 5 and Procedure 2, Step 2).

Simulations

All the simulations are carried out in this stage. It is important that the MD engine, either NAMD or Gromacs, be patched with the latest version of the Colvars module (Materials). It is noteworthy that the end user has the burden of assessing convergence of the different free-energy calculations with the help of BFEE2 and any other graphical-interface-based tool like ParseFEP[56] (Procedure 1, Steps 7–12 and Procedure 2, Steps 4 and 5).

Post-treatments

All the post-treatments—bookkeeping of the free-energy calculations and evaluation of the configurational integrals—can be performed automatically using BFEE2, without any human intervention (Procedure 1, Steps 13–15 and Procedure 2, Steps 6–8). In addition to the standard binding free-energy calculation, we provide guidelines to analyze the interaction of the partners at play—or moieties thereof—to identify the driving forces that underlie molecular association (Box 6).

Materials

Example data

- The structure files of the protein:ligand complexes examined in this protocol are accessible in the Protein Data Bank (www.rcsb.org) using a PDB ID or obtained from Supplementary Data
- Output files from the geometrical and alchemical routes (Supplementary Data)

Hardware and software

- In principle, any computer can be used to run the simulations. However, a Linux-based computer with at least one discrete graphics card is recommended for the 'Simulation' stage, considering the computational cost and software compatibility. For the other stages, computers running Windows, Linux or Mac OS are appropriate
- VMD 1.9.3 or later (www.ks.uiuc.edu/Research/vmd)
- NAMD 3.0 alpha or later (www.ks.uiuc.edu/Research/namd) or Gromacs 2020.4 or later (www.gromacs.org) patched with Colvars (colvars.github.io)
- Python 3.7 or later (https://www.python.org/). The use of conda https://docs.conda.io/en/latest/) is recommended
- BFEE2 (github.com/fhh2626/BFEE2)

The installation guidelines of these pieces of software are provided in Supplementary Information.

Procedure 1: the geometric route

▲ **CRITICAL** We assume here that NAMD is the MD workhorse. The reader is referred to Box 2 to learn the use of Gromacs.

Modeling ● Timing 10 min

1 Generate the topology and coordinate files readable by the MD engines. Detail of the procedure using CHARMM-GUI is provided in Supplementary Information (see also the tutorials for the CHARMM, AmberTools and Gromacs environments).

Input-file generation ● Timing 10 min

2 Open BFEE by typing

```
BFEE 2Gui.py
```

in the terminal (Linux and Mac OS environment) or PowerShell (Windows environment). Optionally, one can link BFEE2 with VMD through File→Settings. If BFEE2 is not linked with VMD, some input files, e.g., the structure file of the extended water box, cannot be generated automatically. Under these premises, scripts will be generated automatically and can be run manually within VMD to create all the necessary files for the binding free-energy calculation.
? TROUBLESHOOTING

3 Set the path to the topology (PSF or PARM) and the structure (PDB or RST) files. If the CHARMM force field is used, the path to the CHARMM force field files (PRM or STR) is also required. For the Abl-SH3:p41 example, the relative paths to the topology and the structure files are `workflow/complex.psf` and `workflow/complex.pdb`, respectively. The relative paths to the force-field files are `workflow/par_all36m_prot.prm` and `workflow/toppar_water_ions.prm`.
! CAUTION The non-bonded fix (NBFIX) terms of the CHARMM general force field for organic molecules (`par_all36_cgenff.prm`) rely on the definition of atom types in other force-field files. Either include `par_all36m_prot.prm`, `par_all36_na.prm` and `par_all36_carb.prm` whenever `par_all36_cgenff.prm` is required, or remove the unnecessary NBFIX terms in `par_all36_cgenff.prm` manually to satisfy the dependency.

4 Set the temperature of the simulation, and define the protein and ligand moieties (following the MDAnalysis syntax, as documented in docs.mdanalysis.org/stable/documentation_pages/selections.html). In the Abl-SH3:p41 example, enter `300`, `segid SH3D` and `segid PPRO`, respectively.
! CAUTION If the CHARMM force field is utilized, it is convenient to make the selection with the keyword `segid`. Conversely, if the Amber force field is utilized, `resid` ought to be preferred because Amber PARM files do not have segment information.

5 Optionally, Set the parameters of the 'Advanced settings' menu. A short description of these parameters is presented hereafter.
- *User-defined Separation direction→Reference*. By default, the separation of the ligand from the protein proceeds along the direction of the vector connecting the centers of mass. By defining an additional object as the reference (following the MDAnalysis syntax), the separation is along the direction of the line connecting the centers of mass of the ligand and the reference. This option is useful for handing semi-buried ligands using the geometrical route (see also Box 3)

- *User-defined large box*. By default, when the CHARMM force field is utilized, BFEE2 automatically expands the TIP3P water box of the molecular assembly to dimensions germane for the separation of the ligand from the protein. If the simulation is to be performed in a non-aqueous environment, if a water model different from the standard TIP3P model is utilized, or if the Amber force field is chosen in lieu of the CHARMM force field, the end user is mandated to provide the molecular assembly with the relevant solvent box of suitable dimensions
- *Stratification*. This option specifies whether the reaction path is decomposed into strata or windows. A value >1 indicates the use of a stratification strategy. For example, for Step 2, exploring the Euler angle Θ ranging from −10 to 10 degrees, *Stratification* set to 2 indicates two separate simulations whereby Θ∈[−10, 0] and Θ∈[0, 10]
- *Compatibility→Pinning down the protein*. This option ensures that the protein remains at the center of the simulation box through adding restraints. By default, the Euler and spherical-coordinate (polar and azimuthal) angles are defined relying on the 'Orientation (quaternion)' CV in the Colvars module. Under these premises, enforcement of roto-translational restraints is required to avoid a net angular acceleration of the protein–ligand complex
- *Compatibility→Use quaternion-based CVs*. The default definitions of Euler and spherical-coordinate angles rely on the 'Orientation (quaternion)' CV in the Colvars module (supported by NAMD 2.13 and later), requiring pinning down the protein, as discussed above. If this option is unchecked, a new, hard-coded definition of Euler and spherical-coordinate angles is utilized (supported by the git version of NAMD patched with the git version of Colvars), which circumvents the requirement of pinning down the protein, as a torque is added internally to the protein to prevent it from tumbling and drifting. We, nevertheless, recommend that the quaternion-based definition of the angles be utilized, as it has been thoroughly tested, and it is fully compatible with the recent official releases of NAMD
- *Model→Membrane protein*. If this option is checked, BFEE2 will assume that the molecular assembly contains a membrane protein. Semi-isotropic pressure coupling will be adopted, and, for the unbound-state simulations, the ligand will be re-solvated and re-neutralized (see also Box 4). At present, this option is only available when the CHARMM force field is used
- *Parallel runs*. We recommend estimating the error associated with a binding free-energy calculation through the geometrical route by running in parallel independent simulations with distinct random-number-generator seeds and computing the standard deviation over the different results obtained. This option determines how many independent simulations are carried out in parallel

▲ **CRITICAL STEP** *Stratification* subsumes a set of parameters crucial for the convergence rate of the free-energy calculations at hand. Within the geometric route, the PMF calculations handling the separation of the ligand from the protein may require three to five windows, or strata, whereas those describing the conformational change of the ligand, either in the bound or in the unbound state, may require one to five windows, depending on the flexibility of the substrate. In the Abl-SH3:p41 example, (3, 1, 1, 1, 1, 1, 5, 3) is a good choice for the number of windows for each step of the geometrical route.

6 Click on the 'Generate Inputs' button, and choose the directory where all the input files will be located. Figure 3 shows the recommended settings for the geometrical route in the case of the Abl-SH3:p41 example.

Simulation ● Timing depends on the available computational resources

7 Equilibrate the molecular assembly by executing

```
cd 000_eq
namd2 +p8 +idlepoll +devices 0 000_eq.conf > 000_eq.log &
```

in terminal (Linux and Mac OS) or PowerShell (Windows). `+p8` means that eight CPU cores will be used, and `+idlepoll +devices 0` indicates that the simulation will be run on GPU 0. The end user can modify these parameters depending on the available computational resources.

? TROUBLESHOOTING

8 Perform the free-energy calculations dealing with the conformational change of the ligand in the bound state by executing

```
cd 001_RMSDBound
namd2 +p8 +idlepoll +devices 0 001_abf_1.conf > 001_abf_1.log &
```

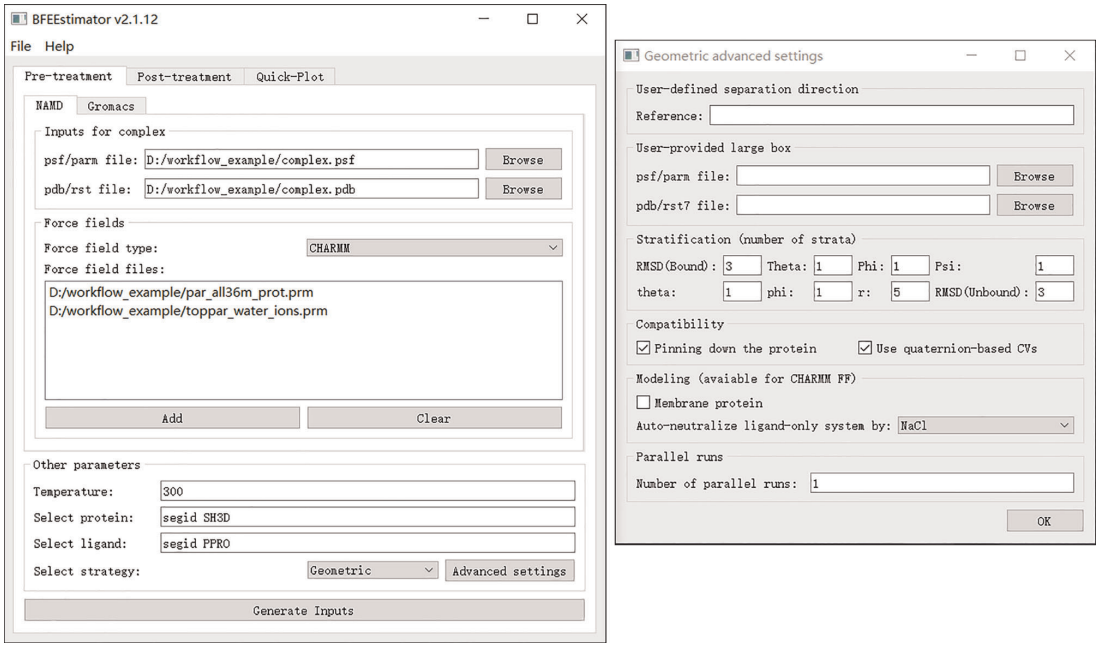

Fig. 3 | Settings for the generation of inputs for the Abl-SH3:p41 case example following the geometrical route. Left, main window of BFEE2; right, advanced settings. Figures 3 and 4 show the BFEE2 interface under Windows 10.

If the transformation is stratified, the simulations of the different windows can be performed in parallel. It must be noted, however, that the restart files of window *i*, which are generated shortly after starting the simulation, are required as the starting point of window $i + 1$. We suggest monitoring the value of the CV in `output/001_abf_i.colvars.traj` to ensure that it is appropriate for window $i + 1$. The user needs to ascertain that the simulation is suitably converged (Fig. 4).

▲**CRITICAL STEP** If the stimulation is not converged, extend it by using, e.g., `001_abf_1.extend.conf`, with an appropriate simulation time. We suggest running at least (100, 10, 10, 10, 10, 10, 200, 100) nanoseconds for each step of the Abl-SH3:p41 example. These simulation times are merely indicative, and ought to be adjusted as a function of the nature of the protein:ligand complex.

? TROUBLESHOOTING

9 Perform the free-energy calculations characterizing the change of Euler angle Θ by executing

```
cd 002_EulerTheta
namd2 +p8 +idlepoll +devices 0 002_abf_1.conf > 002_abf_1.log &
```

Similar to the previous step, ascertain that the simulation is converged by setting an adequate simulation time. In the geometrical route, the change in the three Euler angles in the PMF calculations does not induce any marked variation of the polar and azimuthal angles, but the opposite is not true. Euler angles must, therefore, be handled prior to the polar and azimuthal angles. There is, however, no particular order for the treatment of the three Euler angles.

? TROUBLESHOOTING

10 Let us suppose that the value of Θ corresponding to $\Delta G = 0$ is $-2°$ in `002_EulerTheta/output/abf_1.abf1.czar.pmf`—or, alternatively, in the file where the different contributions of a stratified free-energy calculation are merged. In the latter case, select the 'Merge PMFs' option in the 'Quick-plot' tab to see the complete PMF.

Then open `003_EulerPhi/colvars_1.in`, and change,

```
harmonic {
colvars eulerTheta
forceConstant 0.1
```

```
centers 0.0
}
```

to

```
harmonic {
colvars eulerTheta
forceConstant 0.1
centers -2.0
}
```

to guarantee an optimal value for Θ when handing the other CVs.
Next, perform the free-energy calculation characterizing the change of Euler angle Φ by executing

```
cd 003_EulerPhi
namd2 +p8 +idlepoll +devices 0 003_abf_1.conf > 003_abf_1.log &
```

? TROUBLESHOOTING

11 Run the following simulations sequentially, which characterizes, respectively, the changes in Euler angle Ψ and in spherical-coordinate angles θ and φ, the equilibration of the protein:ligand in an extended water box and the separation of the ligand from the protein, namely

```
004_EulerPsi/004_abf_1.conf
005_PolerTheta/005_abf_1.conf
006_PolerPhi/006_abf_1.conf
007_r/007.1_eq.conf
007_r/007.2_abf_1.conf
```

! CAUTION Similar to Step 10, prior to the simulations, revise the definition of `centers` for the restraints declared in the `*.in` files.
? TROUBLESHOOTING

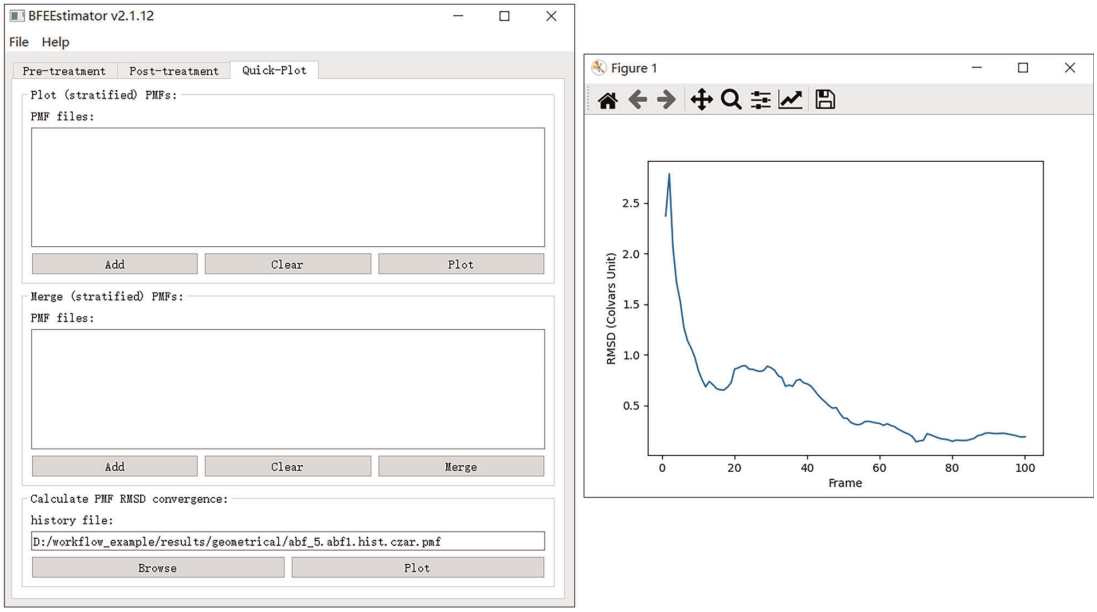

Fig. 4 | Monitoring the convergence of a PMF calculation. An approximate plateau of the time evolution of the PMF RMSD with respect to its initial value usually is suggestive of a satisfactory convergence. Note that some minor fluctuations of the time-evolution curve are common and can usually be ignored.

12 Equilibrate the ligand-only computational assay and perform the PMF calculation describing the conformational change of the substrate in bulk water by running the following simulations sequentially,

```
008_RMSDUnbound/008.1_eq.conf
008_RMSDUnbound/008.2_abf_1.conf
```

Then go to Step 13.

▲ **CRITICAL STEP** The simulations corresponding to `001_RMSDBound` and `008_RMSDUnbound` are independent and can, therefore, be performed in parallel, or concurrently, while those corresponding to `002_EulerTheta` through `007_r` ought to be carried out in a sequential order, due to the need of updating the value of the harmonic-restraint centers.

? TROUBLESHOOTING

Post-treatment ● Timing 10 min

13 Open the BFEE2 window, and switch to the 'Post-treatment→Geometric' tab. Then load the PMF files of each step.

14 Set the force constants of CVs. The value of these force constants corresponds to, for example, the `forceConstant` option in the following block.

```
harmonic {
colvars eulerTheta
forceConstant 0.1
centers 0.0
}
```

If the `forceConstant` options in `*.in` files are not changed manually during the multistep free-energy calculation, the default force constants provided in the 'Post-treatment' tab of BFEE2 GUI can be directly adopted.

15 Set the temperature of simulations and r^* of the integration. The choice of r^* should not affect the result of the free-energy calculation, as long as it is sufficiently large (i.e., the PMF curve of `007_r` is flat for $r > r^*$). Then click on the 'Calculate binding free energy' button. Figure 5 shows how to do post-treatment of the geometrical route for the Abl-SH3:p41 example.

Procedure 2: the alchemical route

Modeling and input-file generation ● Timing 20 min

1 Perform Steps 1–4 from Procedure 1.

2 Optionally, set the parameters of the 'Advanced settings' menu. A short description of these parameters is presented hereafter.

- *Stratification*. Similar to the corresponding option for the geometrical route
- *Double-wide sampling simulation*. If this option is checked, a double-wide sampling simulation in lieu of explicit forward and backward transformations will be performed. The end user must know how to parse the output files of double-wide simulations
- *Compatibility→Pinning down the protein*. Similar to the corresponding option for the geometrical route
- *Compatibility→Use quaternion-based CVs*. Similar to the corresponding option for the geometrical route
- *Minimization before sampling*. This option allows NAMD to perform an energy minimization prior to an FEP calculation at a given value of the coupling parameter
- *Model→Membrane protein*. Similar to the corresponding option in the geometrical route

▲ **CRITICAL STEP** *Stratification* subsumes a set of parameters crucial for the convergence rate of the free-energy calculations at hand. Within the alchemical route, as many as 20–400 intermediate states may be required for each transformation, depending on the flexibility, when adding reversibly the relevant geometric restraints, and the size of the ligand, when decoupling it reversibly from its environment. In the Abl-SH3:p41 example, (100, 200,

100, 200) × 2 are good choices for the number of windows for each step of the alchemical route.

3 Click on the 'Generate Inputs' button, and choose the directory where all the input files will be located. Figure 6 shows the recommended settings for the alchemical route in the case of the Abl-SH3:p41 example.

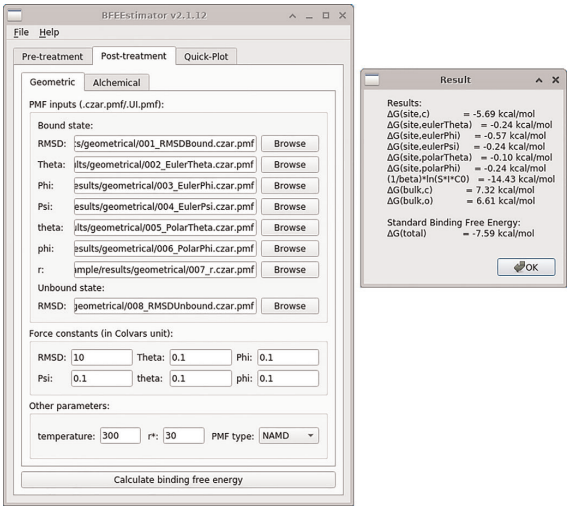

Fig. 5 | Settings for the post-treatment of the Abl-SH3:p41 example following the geometrical route. Left, main window of BFEE2; right, the results. The contributions from all the substeps are supplied, and the standard binding free energy is the sum of them. This figure shows the BFEE2 graphical interface under Ubuntu 20.04.

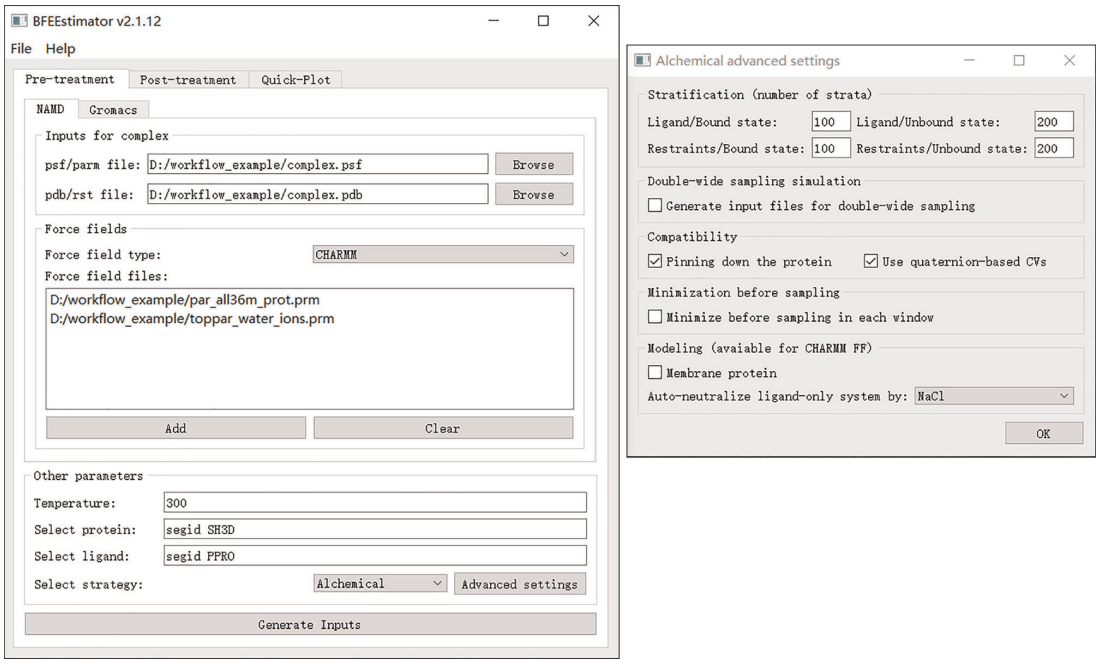

Fig. 6 | Settings for the generation of inputs for the Abl-SH3:p41 case example following the alchemical route. Left, main window of BFEE2; right, advanced settings.

Simulation ● Timing depends on the available computational resources

4 Equilibrate the molecular assembly by executing

```
cd 000_eq
namd2 +p8 +idlepoll +devices 0 000.1_eq.conf > 000.1_eq.log &
namd2 +p8 +idlepoll +devices 1 000.2_eq_ligandOnly.conf >
000.2_eq_ligandOnly.log &
```

in terminal (Linux and Mac OS) or PowerShell (Windows). The `+p8` means using eight CPU cores and `+idlepoll +devices 0` represents the use of GPU 0. The end user can adjust these parameters to adapt their own computers.

? TROUBLESHOOTING

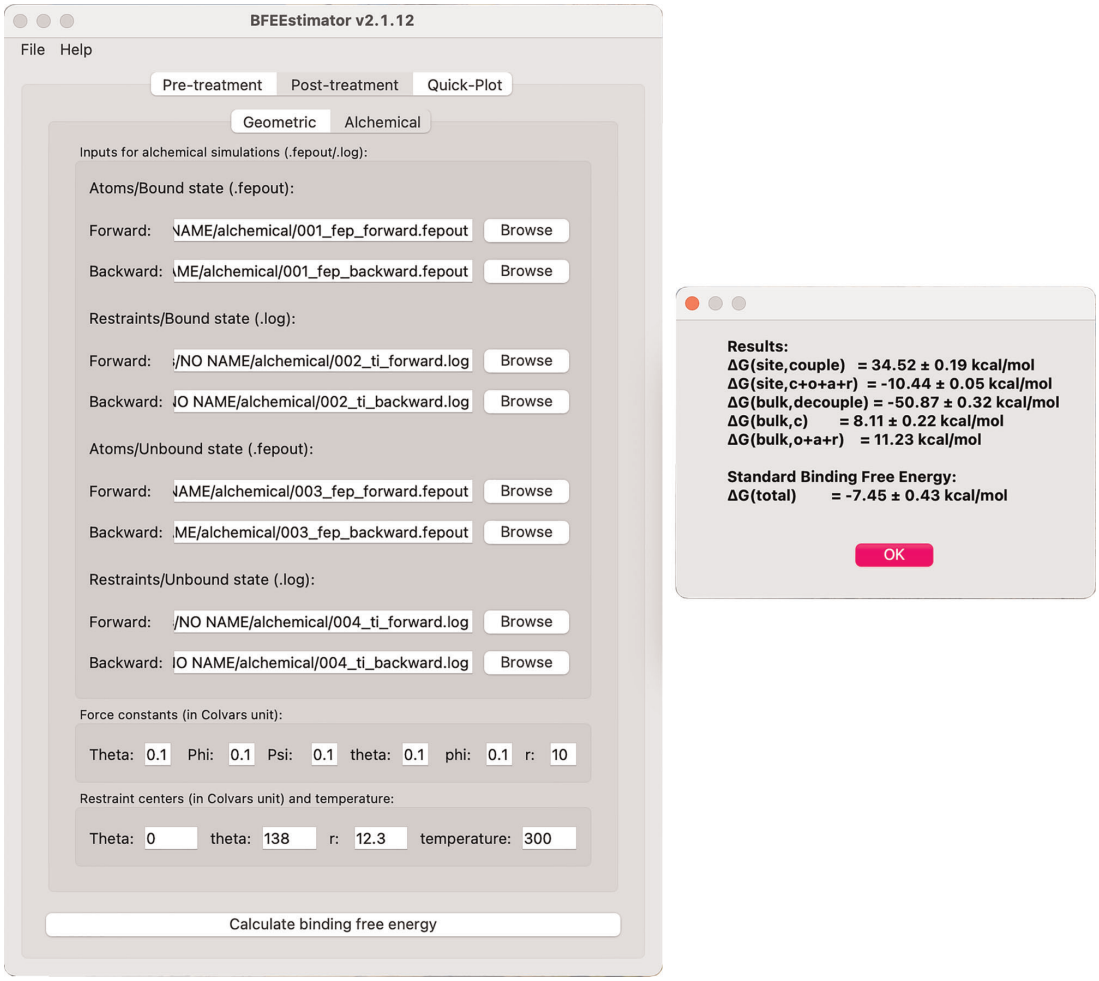

Fig. 7 | Settings for post-treatment of the Abl-SH3:p41 example for the alchemical route. Left, main window of BFEE2; right, the results. The contributions of all the substeps are provided, and the standard binding free energy is the sum of them. The reported errors show the hysteresis between the forward and backward simulations, which corresponds to an approximate measure of the reliability of the free-energy calculation. This figure shows the BFEE2 graphical interface under Mac OS 11.4.

5 Run the following simulations:

```
001_MoleculeBound/001.1_fep_backward.conf
001_MoleculeBound/001.2_fep_forward.conf
002_RestraintBound/002.1_ti_backward.conf
002_RestraintBound/002.2_ti_forward.conf
003_MoleculeUnbound/003.1_fep_backward.conf
003_MoleculeUnbound/003.2_fep_forward.conf
004_RestraintUnbound/004.1_ti_backward.conf
004_RestraintUnbound/004.2_ti_forward.conf
```

▲ **CRITICAL STEP** These simulations can be carried out in parallel, for as long as the backward alchemical simulations are performed prior to the corresponding forward ones. The end user should ascertain that the simulation is converged, which is mirrored in a small error between the forward and the backward simulations, as detailed below in the 'Post-treatment' section and Box 5. Should the simulation not be converged, one needs to increase either the simulation time or the number of windows, or possibly both, and restart the corresponding simulations. We suggest to run at least (100, 400, 100, 400) × 2 nanoseconds for each step of the Abl-SH3:p41 example, prior to jumping to Step 6.
? TROUBLESHOOTING

Post-treatment ● Timing 10 min

6 Open the BFEE2 window, and switch to the 'Post-treatment→Alchemical' tab. Then load the `*.fepout` or `*.log` files of each step.

7 Similar to the Geometrical route (Step 14 of Procedure 1), set force constants of CVs.

8 Set the centers of restraints and the temperature of the simulations. The centers of restraints correspond to, for example, the `centers` option in the following block.

```
harmonic {
colvars eulerTheta
forceConstant 0.1
centers 0.0
}
```

Then click the 'Calculate binding free energy' button. Figure 7 shows how to perform post-treatment of the Abl-SH3:p41 example for the alchemical route. The errors reflecting the hysteresis between the forward and backward simulations are provided and indicate the convergence of the multistep free-energy calculation. The end user should extend the simulation time or increase the number of windows, or both, if a large hysteresis is measured between the forward and backward transformations.
▲ **CRITICAL STEP** The ParseFEP plugin of VMD can be used to improve the reliability of alchemical free-energy calculations. See Box 6 for more details.

Troubleshooting

Troubleshooting advice is provided in Table 3. Most of the issues are relevant to the use of NAMD. The end user is therefore strongly advised to become familiar with the latter through NAMD User's Guide and tutorials. We also suggest asking for help through the NAMD mailing list (https://www.ks.uiuc.edu/Research/namd/mailing_list/).

Table 3 | Troubleshooting table

Procedure 1 step	Procedure 2 step	Problem	Possible reason	Solution
2	1	Executing BFEE2 fails with error message 'TypeError: "Shiboken.ObjectType" object is not iterable'	Some pip versions of PySide2 are not stable in Windows	Use conda to install BFEE2

Table continued

Table 3 (continued)

Procedure 1 step	Procedure 2 step	Problem	Possible reason	Solution
7	4	Simulation crashes with error message 'FATAL ERROR: DIDN'T FIND vdW PARAMETER FOR ATOM TYPE *****'	Some atom types used in the topology file (PSF) are not found in the provided force field files (PRM)	Determine the necessary force field files, and include all of them in the 'input generation' step
			The dependency of NBFIX terms of par_all36_cgenff.prm is not satisfied	Either include par_all36m_prot.prm, par_all36_na.prm and par_all36_carb.prm whenever par_all36_cgenff.prm is required, or remove the unnecessary NBFIX terms in par_all36_cgenff.prm manually to satisfy the dependency
		Simulation crashes with error message 'ERROR: Atoms moving too fast; simulation has become unstable'	The initial structure provided by the user is problematic, e.g., with bad initial contacts of atoms	Check the initial structure before loading it in BFEE2. Remodel the molecular assembly if necessary
		Simulation crashes with error message 'ERROR: Periodic cell has become too small for original patch grid'	The initial structure provided by the user is problematic, e.g. with the wrong density of bulk water	Check the initial structure before loading it to BFEE2. Remodel the molecular assembly if necessary
11	—	Simulation crashes with error message 'ERROR: Atoms moving too fast; simulation has become unstable'	A bug of VMD 1.9.3 may set the coordinates of some atoms to (0, 0, 0) when parsing the PDB file of a large molecular assembly	Use VMD 1.9.4, or find atoms with coordinates of (0, 0, 0) in complex_largeBox.pdb, copy the corresponding lines from the original PDB file and overwrite the lines of atoms with coordinates of (0, 0, 0) in complex_largeBox.pdb
9–12	—	Simulation crashes with error message 'Info: EXTENDED SYSTEM FILE output/abf_1.restart.xscFATAL ERROR: Unable to open extended system file'	The restart files of window 1 are missing, but they are required as the starting point of window 2	Run the simulation of window 1 (***_abf_1.conf) until output/abf_1.restart.coor, output/abf_1.restart.vel and output/abf_1.restart.xsc are generated
11–12	4–5	Simulation crashes with error message 'FATAL ERROR: UNABLE TO OPEN.psf FILE *****.psf'	Some input files are not generated automatically because BFEE2 is not linked with VMD	Run *.tcl through VMD manually to generate the necessary files
–	5	Simulation crashes with error message 'Info: EXTENDED SYSTEM FILE output/fep_backward.xscFATAL ERROR: Unable to open extended system file'	The restart files of the backward simulation are missing, but they are required as the starting point of the forward alchemical transformation	Finish running fep_backward.conf before starting the forward alchemical transformation
–		Simulation crashes with error message 'ERROR: Alchemical free-energy perturbation is not supported in CUDA version'	GPU acceleration is not supported by NAMD version <3.0	Use NAMD 3.0 or a later version as the MD engine

Timing

Procedure 1: the geometrical route

Step 1, modeling: 10 min

Steps 2–6, generating the required input files: 10 min

Steps 7–12, performing the simulations of the geometrical route and monitoring the convergence: depends on the available computational resources. As a reference, on a GTX 2070, the simulation speeds for the Abl-SH3:p41 case example are 50 ns/d for the PMF calculation characterizing the separation of the ligand from the protein and 90 ns/d for all other PMF calculations. Hence, the total time required by these steps for the Abl-SH3:p41 case example may be <1 d

Steps 13–15, post-treatments of the geometrical route: 10 min

Procedure 2: the alchemical route

Steps 1–3, modeling and generating the required input files: 20 min

Steps 4 and 5, performing the simulations of the alchemical route and monitoring the convergence: depends on the available computational resources. As a reference, on a GTX 2070, the simulation speeds

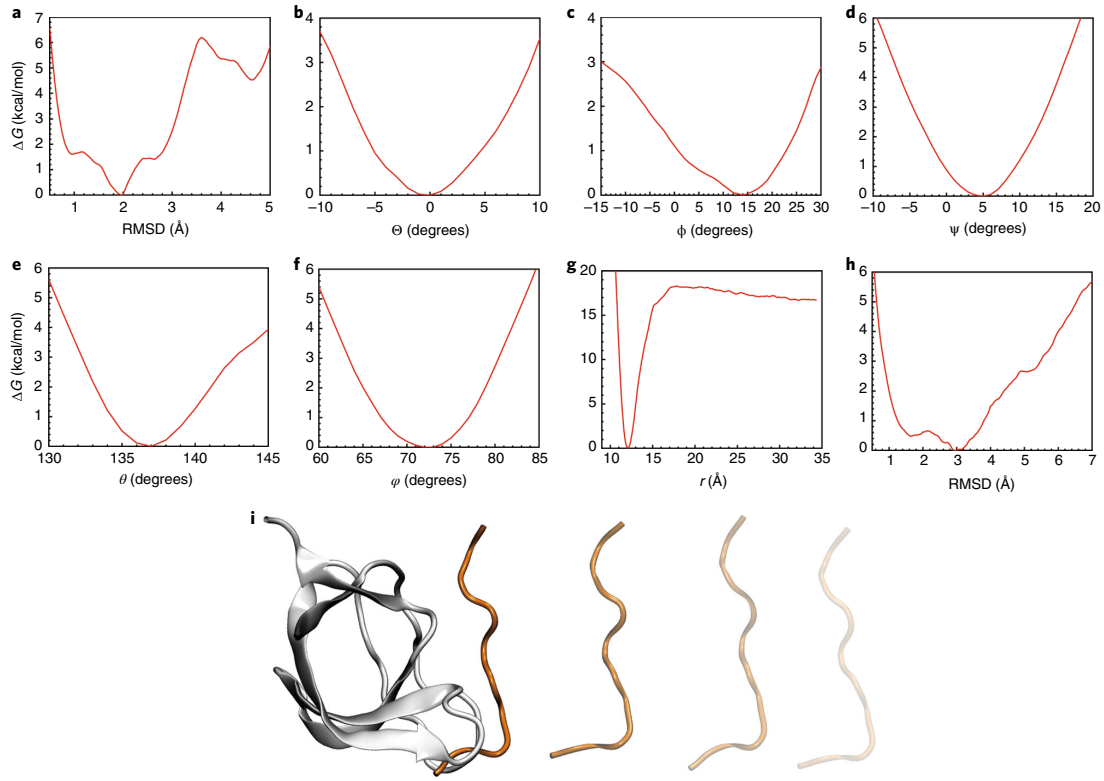

Fig. 8 | Results of PMF calculations in the geometrical route for the Abl-SH3:p41 example. a–h, The PMF calculations using as the CV the RMSD of the ligand with respect to its native, bound-state conformation (**a**), the three Euler angles, Θ (**b**), Φ (**c**) and Ψ (**d**), the polar, θ (**e**), and azimuth, φ (**f**), angles and the distance between the center of mass of the ligand and that the protein (**g**), and the RMSD of the ligand in a bulk environment, i.e., in the unbound state, with respect to its native, bound-state conformation (**h**). **i**, The protein:ligand separation follows an unphysical, rectilinear pathway, owing to the restraints acting on all the other CVs, which diminishes the difficulty of capturing the change in configurational entropy as the two partners of the complex associate.

for the Abl-SH3:p41 case example are 40, 60, 80 and 100 ns/d for the reversible decoupling of the ligand from the protein, adding restraints on the ligand in its bound state, decoupling the ligand from the bulk water, and adding restraints on the ligand in its unbound state, respectively. Hence, the total time required by these steps for the Abl-SH3:p41 case example may <2 d
Steps 6–8, post-treatments of the alchemical route: 10 min

Boxes
Box 1, analysis of a PMF calculation: 20 min
Boxes 2–4, additional examples of binding free-energy calculations: depends on the available computational resources
Box 5, analyzing the enthalpic driving force underlying protein:ligand association: depends on the available computational resources
Box 6, analyzing alchemical free-energy calculations using ParseFEP: 20 min

Anticipated results

The Abl-SH3:p41 case example—geometrical route. Eight PMF calculations were performed to estimate the standard binding free energy of the complex (Fig. 8).

It should be noted that, for flexible ligands, the thermalized bound-state conformation provided to BFEE2 as the reference may slightly differ between independent runs owing to the chaotic nature of MD. This difference might affect the outcome of the PMF calculations, especially those using as a CV the

RMSD of the ligand with respect to the reference native conformation. However, the final standard binding free energy is, in principle, not affected by such minute conformational differences between the references provided to BFEE2, assuming that deviation in the protein:ligand structures remains moderate.

The following prerequisites of final PMFs can be used to validate the correctness and convergence of a standard binding free-energy calculation through the geometrical route:

- From the simulations using the Euler and spherical-coordinate angles as CVs, the PMFs are generally pseudo-quadratic. If this is not the case, we suggest extending the range of the CV accordingly
- The simulation describing the reversible separation of the ligand from the protein is the key step of the geometrical route. Starting from the global minimum, the free energy usually increases sharply until a pseudo-plateau is reached. The slight decay in the free energy at large separations stems from the contribution of the geometric entropy, or the Jacobian[59], which is evaluated analytically throughout the simulation, and subtracted from the PMF[29] by BFEE2. In the Abl-SH3:p41 case example, a pseudo-plateau corresponding to a free energy ranging from 15 to 20 kcal/mol is reasonable
- The final result of the PMF calculations using as a CV the RMSD of the ligand with respect to its native, bound-state conformation may acutely rely on the structure of the reference (bound-state conformation). Generally, the PMF does not necessarily consist of a single well, and is often skewed

Using the PMFs shown in Fig. 8 (data provided in Supplementary Data), the estimated standard binding free energy of the Abl-SH3:p41 complex is −7.6 kcal/mol, in excellent agreement with the experimental value, i.e., −7.99 kcal/mol[60], as depicted in Fig. 6. If independent simulations are run in parallel to calculate the standard error, the latter should be within 0.5 kcal/mol, considering that combination of our parallel runs in refs. [16,20,22] yields a standard error of 0.2 kcal/mol. As a post-hoc treatment of the free-energy calculation, the driving force underlying protein:ligand association can be studied with exquisite detail using the pair-interaction calculation feature available in NAMD, as described in Box 5.

The Abl-SH3:p41 case example—alchemical route. Following Procedure 2, using bidirectional alchemical transformations, the standard binding free energy of the Abl-SH3:p41 complex was estimated to be −7.5 kcal/mol (data provided in Supplementary Data), with an approximate error of 0.4 kcal/mol based on the hysteresis between the forward and backward transformations.

It is noteworthy that performing bidirectional transformations along the alchemical route is a convenient way to verify thermodynamic micro-reversibility, mirrored in the absence of a hysteresis between the forward and backward simulations, while allowing the corresponding statistical data to be combined to yield a maximum-likelihood estimator of the free energy[61]. A thorough analysis is observing the free-energy change with respect to λ of the bidirectional simulations, which can be extracted from the FEPOUT or the LOG files by

```
grep "#Free energy change for lambda" 001_fep_forward.fepout > 001_lambda_forward.dat
```

or

```
grep "dA/dLambda" 002_ti_forward.log > 002_lambda_forward.log
```

The overlap of the free energy profiles as a function of the coupling parameter, λ, of the forward and backward transformations, is a necessary, albeit not sufficient, condition for convergence of

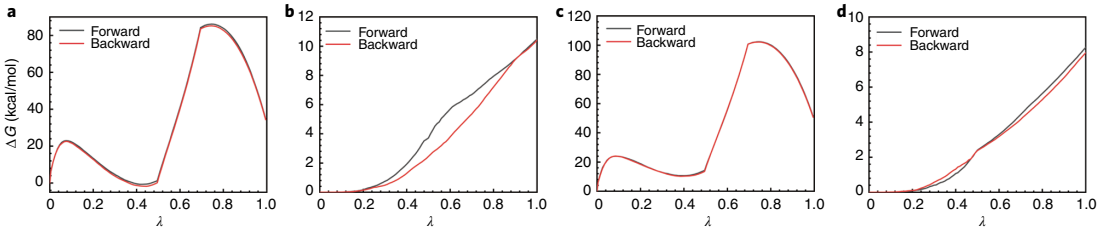

Fig. 9 | Free-energy changes with respect to λ in the alchemical route. a–d, Free-energy change accounting for reversibly decoupling the ligand from the protein (**a**), adding restraints on the bound-state ligand (**b**), decoupling the ligand from the bulk water (**c**) and adding restraints on the unbound-state ligand (**d**), respectively.

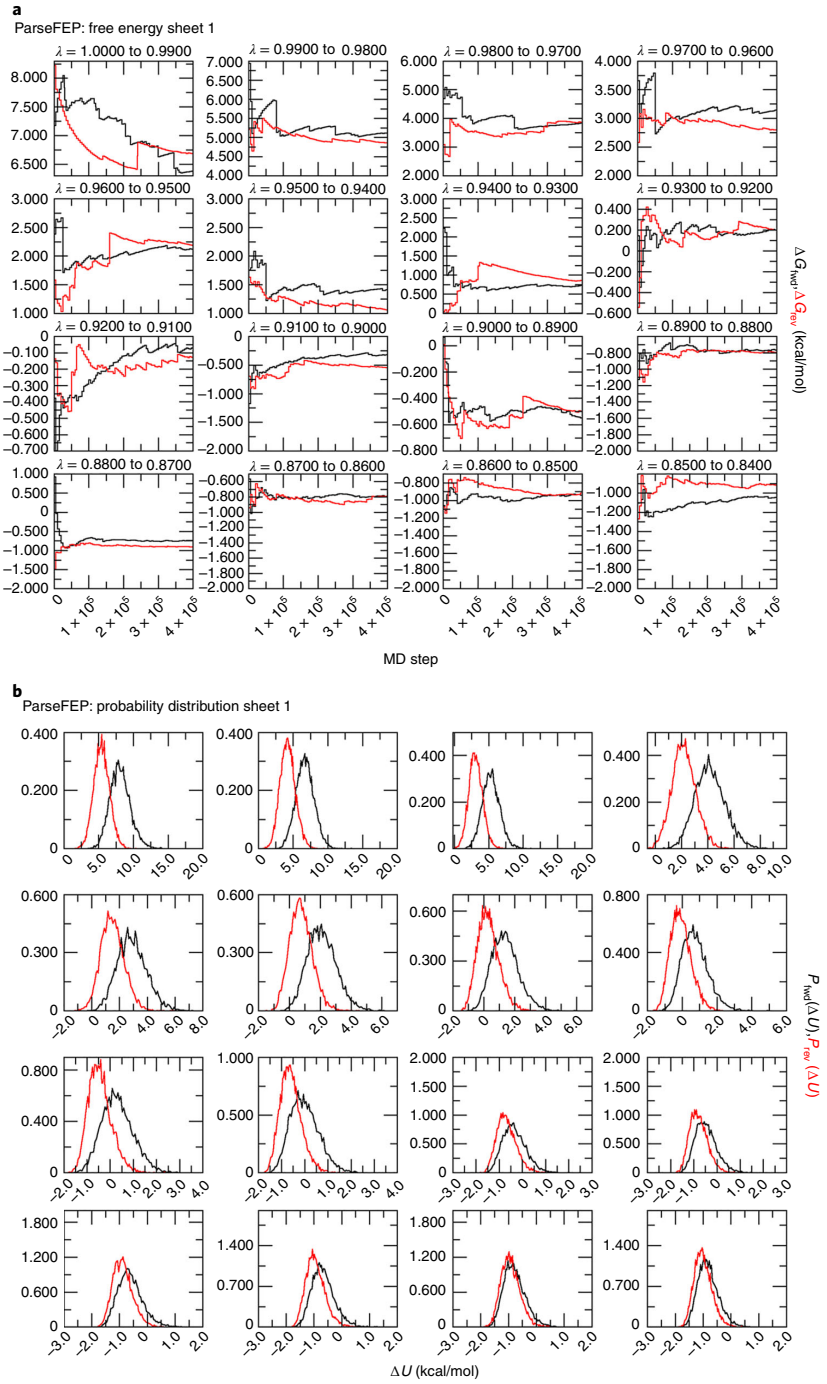

Fig. 10 | Example of the outputs of ParseFEP. a,**b**, Evolution of the free energy in each window of a stratified, bidirectional calculation (**a**), and associated probability distributions of the potential-energy difference, ΔU (**b**). See Box 6 for more details.

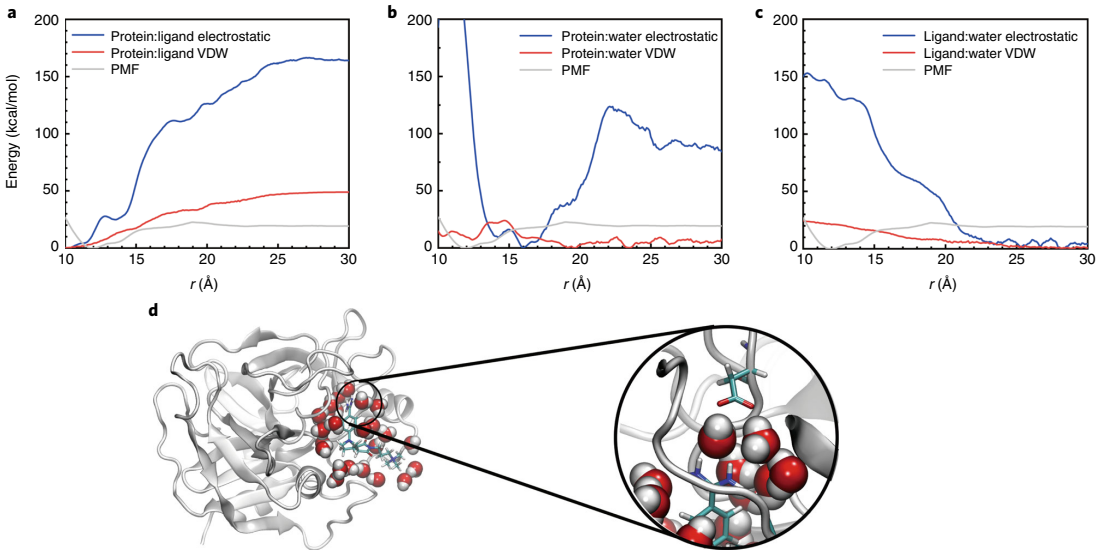

Fig. 11 | Analysis of the enthalpic driving force of the association of Factor Xa: quaternary ammonium. a–c, The contributions of protein:ligand (**a**) protein:water (**b**) and ligand:water (**c**) are analyzed. The electrostatic and van der Waals (VDW) energies are shown, respectively. **d**, The structure corresponding to the most favorable protein–water interaction (r = 16 Å). Water-mediated salt bridges are found in this structure.

bidirectional simulations[61] and, therefore, provides a rough estimate of the reliability of the calculation. In the Abl-SH3:p41 case example, the curves characterizing the forward and backward simulations are very close, suggesting a suitable convergence of our simulations (Fig. 9). A more thorough analysis of the convergence, and of the statistical and systematic errors associated to the simulations, can be performed using ParseFEP[56,61,62]. (Box 6 and Fig. 10). If the Bennett acceptance ratio estimator[63] implemented in ParseFEP[56] is used to improve the precision of free energy calculations, the calculated standard binding free energy of p41 to Abl-SH3 is −7.4 ± 0.3 kcal/mol.

Other examples. Apart from Abl-SH3:p41, we have calculated the standard binding free energies of the following examples to illustrate the use of BFEE2, that is,

- DIAP1-BIR1: Grim peptide. This example was chosen as an illustration of the use of Gromacs as the MD workhorse. Following the procedure depicted in Box 2, the standard binding free-energy estimate was found to be equal to −8.7 ± 0.7 kcal/mol, in excellent agreement with the experimental value of −9.5 kcal/mol[64]
- T4 lysozyme L99A:benzene. Since the ligand molecule, benzene, is deeply buried in the protein, the alchemical route was adopted to estimate the binding free energy. In stark contrast with the Abl-SH3:p41 complex, long simulation times are required to decouple reversibly the ligand from its binding site, as capturing the exchange of water molecules between the cavity and the bulk is admittedly slow. The binding free energy estimated in this study, namely −6.0 ± 1.0 kcal/mol, agrees well with the experimental measurement of −5.2 kcal/mol[65]
- Trypsin:benzamidine. As benzamidine is semi-buried in the protein, a well-defined pathway is required to separate the former from the latter, should the geometrical route be chosen. Following the guideline provided in Box 3, the computed binding free energy, −7.8 ± 0.6 kcal/mol, only differs slightly from the experimental value (−7.2 to −6.3 kcal/mol)[66–68]
- Factor Xa:quaternary ammonium. The choice of an appropriate force field is crucial for the estimation of standard binding free energies. The standard CHARMM36m force field[47] yields a binding free-energy estimate of −3.7 ± 0.5 kcal/mol, differing markedly from the experimental value, −9.0 kcal/mol[69]. By grossly ignoring induction phenomena, pairwise additive force fields, like the CHARMM36m force field, are notorious for misrepresenting cation–π interactions,[70] which drive association in the factor Xa:quaternary ammonium complex. Switching to a force field germane to cation–π interactions[38,71], the computed binding free energy now amounts to −8.7 ± 0.4 kcal/mol, in excellent agreement with the experimental value. This example is also used to illustrate how to analyze the driving force underlying protein:ligand association, as detailed in Box 5 and Fig. 11

Table 4 | Pitfalls and caveats of examples shown in this protocol

Examples	Pitfalls and caveats	See also
Abl-SH3:p41	Captures the conformational change of the flexible ligand. Usually requires a stratification strategy in free-energy calculations	Main procedure and refs. [19,20]
DIAP1-BIR1:Grim peptide	—	Box 2
T4 lysozyme L99A:benzene	Buried ligand. Requires long simulation time to capture the water exchange inside and outside the binding cavity	Refs. [11]
Trypsin:benzamidine	Semi-buried ligand. Requires manual definition of the separation direction if the geometrical route is adopted	Box 3
Factor Xa:quaternary ammonium	Require careful choice of the force field to correctly model cation–π interaction	Ref. [38]
MUP-I:2-methoxy-3-isopropylpyrazine and MUP-I:6-hydroxy-6-methyl-3-heptanone	Bound water in the binding site may drift away during the free energy calculation. Requires definition of additional restraints for these water molecules	Supplementary Information
β1-adrenergic receptor:4-methyl-2-(piperazin-1-yl) quinoline	Captures the reversible hydration of the binding site for a membrane protein. Requires equilibration of the membrane protein and its environment without the ligand, allowing water to diffuse inside the protein as a preamble to the free-energy calculations	Box 4
V1-ATPase:nucleotide (ATP or ADP + P_i)	Practically challenging example. Requires the combination of multiple operations mentioned above	Ref. [75]

- MUP-I:2-methoxy-3-isopropylpyrazine and MUP-I:6-hydroxy-6-methyl-3-heptanone. Protein:ligand affinity ranking is an important task in pharmaceutical sciences and is usually performed through relative binding free-energy calculations. In this protocol, we show that protein:ligand affinity ranking can be easily achieved through standard binding free-energy calculations. Through the alchemical route, the standard binding free energies of MUP-I:2-methoxy-3-isopropylpyrazine and MUP-I:6-hydroxy-6-methyl-3-heptanone are estimated at −7.8 ± 1.0 and −5.5 ± 0.7 kcal/mol, respectively, in good agreement with the experiment (−7.8 and −6.0 kcal/mol)[72,73]
- β1-adrenergic receptor:4-methyl-2-(piperazin-1-yl) quinoline. BFEE2 can be used to predict the binding affinity of a ligand to a membrane protein, which has traditionally been seen as a daunting challenge. One of the difficulties is to capture the reversible hydration of the binding site during the PMF calculations or alchemical transformations, as the membrane protein is fully immersed in its lipid environment. Following the procedure of the alchemical route introduced in Box 4, the calculated standard binding free energy of β1-adrenergic receptor:4-methyl-2-(piperazin-1-yl) quinoline of −8.1 ± 1.0 kcal/mol agrees well with the experimental value, namely −9.07 kcal/mol[74]
- V_1-ATPase:nucleotide (ATP or ADP + P_i). One of the most challenging applications of our methodology is the determination of the binding affinity of ATP and ADP + P_i bound towards V_1-ATPase in its distinct conformational states, following the alchemical route[75]. The molecular assemblies at hand, of dimensions on the order of 170 × 170 × 190 $Å^3$, are particularly complex compared with most of the biological objects reported herein. Our results indicate that ATP association in the tight site (−11.6 ± 0.8 kcal/mol) is energetically more favorable than that of ADP + P_i (−8.3 ± 0.9 kcal/mol), and that binding affinities of both of the nucleotides in the empty site are nearly identical (ATP: −4.1 ± 1.1 kcal/mol, ADP: −4.3 ± 0.8 kcal/mol) and less favorable compared with the tight or bound sites. This trend is in good agreement with the experimental measurements carried out on F_1-ATPase[76]

The pitfalls and caveats of the aforementioned examples are summarized in Table 4. Additional success stories of the methodology are provided in Table 2. To make this protocol completely transparent to the end user, we provide in Supplementary Information a detailed description of the complexes, the parameters of the binding free-energy calculation and the results of the different substeps for three practical examples, namely MDM2-p53:NVP-CGM097, MUP-I:2-methoxy-3-isopropylpyrazine and MUP-I:6-hydroxy-6-methyl-3-heptanone, in addition to the Abl-SH3:p41 case example detailed in the text.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The input and output files of BFEE2 of examples are provided in Supplementary Data. The data shown in Figs. 8–11 were obtained from new simulations, as a way to verify and guarantee the reproducibility of our protocol. Input files for these simulations are available from the corresponding authors upon request.

Code availability

The Python package of BFEE2 can be installed through pip (https://pypi.org/project/BFEE2/) and conda (https://anaconda.org/conda-forge/bfee2). The source code of BFEE2 is available on GitHub (https://github.com/fhh2626/BFEE2)[77].

References

1. Limongelli, V. Ligand binding free energy and kinetics calculation in 2020. *WIREs Comput. Mol. Sci.* **10**, e1455 (2020).
2. Chodera, J. D. & Mobley, D. L. Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design. *Annu. Rev. Biophys.* **42**, 121–142 (2013).
3. Li, A. & Gilson, M. K. Protein-ligand binding enthalpies from near-millisecond simulations: analysis of a preorganization paradox. *J. Chem. Phys.* **149**, 72311 (2018).
4. de Ruiter, A. & Oostenbrink, C. Advances in the calculation of binding free energies. *Curr. Opin. Struct. Biol.* **61**, 207–212 (2020).
5. Chipot, C. Frontiers in free-energy calculations of biological systems. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **4**, 71–89 (2014).
6. Hermans, J. & Shankar, S. The free energy of xenon binding to myoglobin from molecular dynamics simulation. *Isr. J. Chem.* **27**, 225–227 (1986).
7. 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.* **71**, 670–681 (1996).
8. 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.* **119**, 2707–2714 (1997).
9. Mann, G. & Hermans, J. Modeling protein–small molecule interactions: structure and thermodynamics of noble gases binding in a cavity in mutant phage T4 lysozyme L99A. *J. Mol. Biol.* **302**, 979–989 (2000).
10. Boresch, S., Tettinger, F., Leitgeb, M. & Karplus, M. Absolute binding free energies: a quantitative approach for their calculation. *J. Phys. Chem. B* **107**, 9535–9551 (2003).
11. Deng, Y. & Roux, B. Calculation of standard binding free energies: aromatic molecules in the T4 lysozyme L99A mutant. *J. Chem. Theory Comput.* **2**, 1255–1273 (2006).
12. Mobley, D. L., Chodera, J. D. & Dill, K. A. On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. *J. Chem. Phys.* **125**, 84902 (2006).
13. 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.* **72**, 1047–1069 (1997).
14. 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.* **12**, 3506–3513 (2016).
15. Fu, H. et al. Zooming across the free-energy landscape: shaving barriers, and flooding valleys. *J. Phys. Chem. Lett.* **9**, 4738–4745 (2018).
16. Fu, H., Shao, X., Cai, W. & Chipot, C. Taming rugged free energy landscapes using an average force. *Acc. Chem. Res.* **52**, 3254–3264 (2019).
17. Fu, H. et al. Finding an optimal pathway on a multidimensional free-energy landscape. *J. Chem. Inf. Model.* **60**, 5366–5374 (2020).
18. Woo, H.-J. & Roux, B. Calculation of absolute protein–ligand binding free energy from computer simulations. *Proc. Natl Acad. Sci. USA* **102**, 6825–6830 (2005).
19. Gumbart, J. C., Roux, B. & Chipot, C. Standard binding free energies from computer simulations: what is the best strategy? *J. Chem. Theory Comput.* **9**, 794–802 (2013).
20. 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.* **13**, 5173–5178 (2017).
21. Wang, J., Deng, Y. & Roux, B. Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. *Biophys. J.* **91**, 2798–2814 (2006).
22. Fu, H. et al. BFEE: a user-friendly graphical interface facilitating absolute binding free-energy calculations. *J. Chem. Inf. Model.* **58**, 556–560 (2018).
23. Fu, H., Chen, H., Cai, W., Shao, X. & Chipot, C. BFEE2: automated, streamlined, and accurate absolute binding free-energy calculations. *J. Chem. Inf. Model.* **61**, 2116–2123 (2021).
24. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
25. Comer, J. et al. The adaptive biasing force method: everything you always wanted to know but were afraid to ask. *J. Phys. Chem. B* **119**, 1129–1151 (2015).

26. Zwanzig, R. W. High-temperature equation of state by a perturbation method. I. Nonpolar gases. *J. Chem. Phys.* **22**, 1420–1426 (1954).
27. Chen, H. et al. Boosting free-energy perturbation calculations with GPU-accelerated namd. *J. Chem. Inf. Model.* **60**, 5301–5307 (2020).
28. Kirkwood, J. G. Statistical mechanics of fluid mixtures. *J. Chem. Phys.* **3**, 300–313 (1935).
29. Fiorin, G., Klein, M. L. & Hénin, J. Using collective variables to drive molecular dynamics simulations. *Mol. Phys.* **111**, 3345–3362 (2013).
30. Zhang, H. et al. Accurate estimation of the standard binding free energy of netropsin with DNA. *Molecules* **23**, 228 (2018).
31. Du, S. et al. Curvature of buckybowl corannulene enhances its binding to proteins. *J. Phys. Chem. C* **123**, 922–930 (2019).
32. Sun, H., Li, Y., Tian, S., Wang, J. & Hou, T. P-loop conformation governed crizotinib resistance in G2032R-mutated ROS1 tyrosine kinase: clues from free energy landscape. *PLOS Comput. Biol.* **10**, e1003729 (2014).
33. Deng, N. et al. Comparing alchemical and physical pathway methods for computing the absolute binding free energy of charged ligands. *Phys. Chem. Chem. Phys.* **20**, 17081–17092 (2018).
34. Kuusk, A. et al. Adoption of a turn conformation drives the binding affinity of p53 C-terminal domain peptides to 14-3-3σ. *ACS Chem. Biol.* **15**, 262–271 (2020).
35. Qian, Y. et al. Absolute free energy of binding calculations for macrophage migration inhibitory factor in complex with a druglike inhibitor. *J. Phys. Chem. B* **123**, 8675–8685 (2019).
36. Comer, J. et al. Beta-1,3 oligoglucans specifically bind to immune receptor CD28 and may enhance T cell activation. *Int. J. Mol. Sci.* **22**, 3124 (2021).
37. Velez-Vega, C. & Gilson, M. K. Overcoming dissipation in the calculation of standard binding free energies by ligand extraction. *J. Comput. Chem.* **34**, 2360–2371 (2013).
38. Liu, H., Fu, H., Chipot, C., Shao, X. & Cai, W. Accuracy of alternate nonpolarizable force fields for the determination of protein–ligand binding affinities dominated by cation−π interactions. *J. Chem. Theory Comput.* **17**, 3908–3915 (2021).
39. 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.* **120**, 9401–9409 (1998).
40. Limongelli, V., Bonomi, M. & Parrinello, M. Funnel metadynamics as accurate binding free-energy method. *Proc. Natl Acad. Sci. USA* **110**, 6358–6363 (2013).
41. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl Acad. Sci. USA* **99**, 12562–12566 (2002).
42. Raniolo, S. & Limongelli, V. Ligand binding free-energy calculations with funnel metadynamics. *Nat. Protoc.* **15**, 2837–2866 (2020).
43. Mobley, D. L., Chodera, J. D. & Dill, K. A. Confine-and-release method: obtaining correct binding free energies in the presence of protein conformational change. *J. Chem. Theory Comput.* **3**, 1231–1235 (2007).
44. Miao, Y., Bhattarai, A. & Wang, J. Ligand Gaussian accelerated molecular dynamics (LiGaMD): characterization of ligand binding thermodynamics and kinetics. *J. Chem. Theory Comput.* **16**, 5526–5547 (2020).
45. Wang, L., Friesner, R. A. & Berne, B. J. Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). *J. Phys. Chem. B* **115**, 9431–9438 (2011).
46. Kofke, D. A. & Cummings, P. T. Precision and accuracy of staged free-energy perturbation methods for computing the chemical potential by molecular simulation. *Fluid Phase Equilib* **150–151**, 41–49 (1998).
47. Huang, J. et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).
48. Tian, C. et al. ff19SB: amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *J. Chem. Theory Comput.* **16**, 528–552 (2020).
49. Lemkul, J. A., Huang, J., Roux, B. & MacKerell, A. D. An empirical polarizable force field based on the classical drude oscillator model: development history and recent applications. *Chem. Rev.* **116**, 4983–5013 (2016).
50. Ponder, J. W. et al. Current status of the AMOEBA polarizable force field. *J. Phys. Chem. B* **114**, 2549–2564 (2010).
51. Jo, S. & Jiang, W. A generic implementation of replica exchange with solute tempering (REST2) algorithm in NAMD for complex biophysical simulations. *Comput. Phys. Commun.* **197**, 304–311 (2015).
52. Deng, Y. & Roux, B. Computation of binding free energy with molecular dynamics and grand canonical monte carlo simulations. *J. Chem. Phys.* **128**, 115103 (2008).
53. Ben-Shalom, I. Y., Lin, C., Kurtzman, T., Walker, R. C. & Gilson, M. K. Simulating water exchange to buried binding sites. *J. Chem. Theory Comput.* **15**, 2684–2691 (2019).
54. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865 (2008).
55. Case, D. A. et al. Amber 2021 (University of California, San Francisco, 2021).
56. Liu, P., Dehez, F., Cai, W. & Chipot, C. A toolkit for the analysis of free-energy perturbation calculations. *J. Chem. Theory Comput.* **8**, 2606–2616 (2012).
57. Phillips, J. C. et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **153**, 44130 (2020).
58. Abraham, M. J. et al. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).

59. Hénin, J. & Chipot, C. Overcoming free energy barriers using unconstrained molecular dynamics simulations. *J. Chem. Phys.* **121**, 2904–2914 (2004).
60. Pisabarro, M. T. & Serrano, L. Rational design of specific high-affinity peptide ligands for the Abl-SH3 domain. *Biochemistry* **35**, 10634–10640 (1996).
61. Pohorille, A., Jarzynski, C. & Chipot, C. Good practices in free-energy calculations. *J. Phys. Chem. B* **114**, 10235–10253 (2010).
62. Hahn, A. M. & Then, H. Characteristic of Bennett's acceptance ratio method. *Phys. Rev. E* **80**, 031111 (2009).
63. Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **22**, 245–268 (1976).
64. Brown, S. P. & Muchmore, S. W. Large-scale application of high-throughput molecular mechanics with Poisson–Boltzmann surface area for routine physics-based scoring of protein–ligand complexes. *J. Med. Chem.* **52**, 3159–3165 (2009).
65. Morton, A. & Matthews, B. W. Specificity of ligand binding in a buried nonpolar cavity of T4 lysozyme: linkage of dynamics and structural plasticity. *Biochemistry* **34**, 8576–8588 (1995).
66. Mares-Guia, M., Nelson, D. L. & Rogana, E. Electronic effects in the interaction of para-substituted benzamidines with trypsin: the involvement of the π-electronic density at the central atom of the substituent in binding. *J. Am. Chem. Soc.* **99**, 2331–2336 (1977).
67. Katz, B. A. et al. Structural basis for selectivity of a small molecule, S1-binding, submicromolar inhibitor of urokinase-type plasminogen activator. *Chem. Biol.* **7**, 299–312 (2000).
68. Schwarzl, S. M., Tschopp, T. B., Smith, J. C. & Fischer, S. Can the calculation of ligand binding free energies be improved with continuum solvent electrostatics and an ideal-gas entropy correction? *J. Comput. Chem.* **23**, 1143–1149 (2002).
69. Schärer, K. et al. Quantification of cation–π interactions in protein–ligand complexes: crystal-structure analysis of Factor Xa bound to a quaternary ammonium ion ligand. *Angew. Chemie Int. Ed.* **44**, 4400–4404 (2005).
70. Khan, H. M., MacKerell, A. D. & Reuter, N. Cation–π interactions between methylated ammonium groups and tryptophan in the CHARMM36 additive force field. *J. Chem. Theory Comput.* **15**, 7–12 (2019).
71. Liu, H., Fu, H., Shao, X., Cai, W. & Chipot, C. Accurate description of cation–π interactions in proteins with a nonpolarizable force field at no additional cost. *J. Chem. Theory Comput.* **16**, 6397–6407 (2020).
72. Bingham, R. J. et al. Thermodynamics of binding of 2-methoxy-3-isopropylpyrazine and 2-methoxy-3-isobutylpyrazine to the major urinary protein. *J. Am. Chem. Soc.* **126**, 1675–1681 (2004).
73. 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). *Protein Sci* **10**, 997–1004 (2001).
74. Christopher, J. A. et al. Biophysical fragment screening of the β1-adrenergic receptor: identification of high affinity arylpiperazine leads using structure-based drug design. *J. Med. Chem.* **56**, 3446–3455 (2013).
75. Singharoy, A., Chipot, C., Moradi, M. & Schulten, K. Chemomechanical coupling in hexameric protein–protein interfaces harnesses energy within V-type atpases. *J. Am. Chem. Soc.* **139**, 293–310 (2017).
76. Adachi, K., Oiwa, K., Yoshida, M., Nishizaka, T. & Kinosita, K. Controlled rotation of the F1-ATPase reveals differential and continuous binding changes for ATP synthesis. *Nat. Commun.* **3**, 1022 (2012).
77. Fu, H. et al. Accurate determination of protein:ligand standard binding free energies from molecular dynamics simulations. *BFEE2: Binding free energy estimator 2*. https://doi.org/10.5281/zenodo.5501842 (2021).
78. Liu, H., Okazaki, S. & Shinoda, W. Heteroaryldihydropyrimidines alter capsid assembly by adjusting the binding affinity and pattern of the hepatitis B virus core protein. *J. Chem. Inf. Model.* **59**, 5104–5110 (2019).
79. Miao, M. et al. Avoiding non-equilibrium effects in adaptive biasing force calculations. *Mol. Simul.* **47**, 390–394 (2021).
80. Samways, M. L., Bruce Macdonald, H. E. & Essex, J. W. Grand: a Python module for grand canonical water sampling in OpenMM. *J. Chem. Inf. Model.* **60**, 4436–4441 (2020).

Acknowledgements

This study was supported by the National Natural Science Foundation of China (22073050, 22174075 and 22103041), the China Post-doctoral Science Foundation (bs6619012), Frontiers Science Center for New Organic Matter, Nankai University (63181206), the US National Institutes of Health (R01-AI148740), the National Science Foundation (NSF) through grant no. MCB-1517221, the France and Chicago Collaborating in The Sciences (FACCTS) program, and the Agence Nationale de la Recherche (ProteaseInAction). J.C.G. acknowledges computational resources provided through the Extreme Science and Engineering Discovery Environment (XSEDE; TG-MCB130173). The paper is dedicated to the 100th anniversary of Chemistry at Nankai University.

Author contributions

H.F., X.S., W.S. and C.C. conceived the project. H.F. designed the BFEE2 software and implemented the workflow of binding free-energy calculations. H.C. implemented the Gromacs support of BFEE2. H.F., M.B., F.S., E.G.C.D, A.P., F.D. and J.C.G. tested the software. H.F., B.R., W.S. and C.C. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41596-021-00676-1.

Correspondence and requests for materials should be addressed to Wensheng Cai or Christophe Chipot.

Peer review information *Nature Protocols* thanks Nanjie Deng and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 13 November 2020; Accepted: 7 December 2021;
Published online: 11 March 2022

Related links

Key references using this protocol

Woo, H. et al. *Proc. Natl Acad. Sci. USA* **102**, 6825–6830 (2005): https://doi.org/10.1073/pnas.0409005102
Gumbart, J. C. et al. J. *Chem. Theory Comput.* **9**, 794–802 (2013): https://doi.org/10.1021/ct3008099
Fu, H. et al. *J. Chem. Theory Comput.* **13**, 5173–5178 (2017): https://doi.org/10.1021/acs.jctc.7b00791
Fu, H. et al. *Acc. Chem. Res.* **52**, 3254–3264 (2019): https://doi.org/10.1021/acs.accounts.9b00473
Fu, H. et al. *J. Chem. Inf. Model.* **61**, 2116–2123 (2021): https://doi.org/10.1021/acs.jcim.1c00269

Chapter 4

Calculation of Binding Affinities for SARS-CoV-2 Variants

Sommaire

This chapter will present the work I performed on related SARS-CoV-2 complexes. Firstly I will start with how this work is related to my thesis and my personal contribution. Then I will provide a summary of the related publication, prefacing the entire original text, which can also be found under the following reference:

"**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
"

4.1 Motivations and personal contribution

During my PhD, the COVID-19 syndemic was still a major source of concern. The emergence of new worrisome variants such as Delta and Omicron exerted tremendous pressure on the scientific community to provide insight into the virus function to help healthcare professionals in the fight against the virus. With the growing availability of models and structures, we sought to compute with our theoretical approach the binding affinity of all the variants of concerns (VOCs) with their human receptor ACE2 to elucidate their strategies of high infectiosity. Furthermore, the binding affinities reported in the literature were inconsistent with one another, using different techniques and spanning a wide range of kcal/mol. I computed the affinity of Alpha, Beta, Delta, and Omicron with ACE2. I analyzed their interactions at the interface, which gave me experience with absolute binding free energy in the case of protein-protein complexes.

4.2 Summary

In this work, we focused on the SARS-CoV-2 original strain (WT) and the variants called Variants of Concern (VOCs): Alpha, Beta, Delta, and Omicron.[94,95] A variant has two mechanisms to increase its contagiosity: increasing the affinity for the human receptor or escaping the immune system. We computed the binding free energy of the VOCs Receptor Binding Domain (RBD) with the Angiotensin Conversion Enzyme 2 (ACE2) using the geometrical route approach, adapted to protein-protein complexes, to investigate whether an increase in the affinity for ACE2 could explain the rise of those variants. Our calculations also include two neutralizing antibodies, S2E12[96] and H11-D4 to explore potential treatments.[97]

The results of our calculations are detailed below in Table 4.1.

Complexes	$\Delta G^{\circ}_{\mathrm{b}}$ (kcal/mol)	$\Delta G^{\circ}_{\mathrm{exp}}$ (kcal/mol)
$\mathrm{WT_{crystal}}$:ACE2	-11.5 ± 0.3	-11.4[98]
$\mathrm{WT_{model}}$:ACE2	-6.7 ± 2.3	-11.4[98]
$\mathrm{Alpha_{model}}$:ACE2	-12.3 ± 1.2	-11.6[99]
$\mathrm{Beta_{model}}$:ACE2	-10.0 $\pm$ 1.2	-11.1[99]
$\mathrm{Beta_{Cryo\text{-}EM}}$:ACE2	-11.0 ± 1.6	-11.1[99]
$\mathrm{Delta_{model}}$:ACE2	-9.6 ± 0.5	-9.9[97]
Omicron BA.2:ACE2	-11.4 ± 1.3	-11.5[100]
WT:ACE2 (full-length glycans)	-10.8 ± 0.3	-11.4[98]
S2E12:Delta	-12.5 ± 0.3	-12.0[101]
H11-D4:WT	-9.4 ± 0.5	-9.9[102]

Table 4.1: Computed binding free energies against experimental values of all studied complexes

While the calculations with our models were ongoing, new experimental structures for the Alpha, Beta, and Delta variants were reported.[97,103,104] As a measure of precaution, we checked our models against these new structures to ensure that the altered interfaces were consistent with all available experimental data. The Delta and the Alpha variant models were in agreement with experimental data.[97,103] However, an alignment revealed a local rearrangement at the interface centered on residue H34 of ACE2 in the experimental structure of the Beta variant (7SY6),[104] and absent in our model. This conformational change is specific to the Beta variant and is not present in other VOCs, as documented by the authors.[104]

While our initial model of the Beta variant, on the one hand, and the early structure of the WT taken from the Covid-19 Charmm GUI archive, on the other hand, depart from the more recent experimental

structures, we have reported the corresponding free-energy calculations to showcase the influence of the starting structure on the binding affinity using both models (WT_{model}, $Beta_{model}$) and experimental structures ($WT_{crystal}$, $Beta_{Cryo\text{-}EM}$). Overall, we obtained results in good agreement with experimental thermodynamic data for nearly all complexes, except for the early WT and Beta variant models. This showcases the crucial importance of an accurate starting structure to perform this type of calculation and one of the method's limitations. Comparing the affinity obtained for all VOCs, using WT:ACE2 as a reference, we established that both the Alpha and the Beta variants rely on an increased affinity for ACE2. In contrast, the Delta variant has a lower affinity for ACE2 and has spread widely owing to its immune-escape properties. The Omicron variant has a binding affinity similar to that of the WT, albeit harboring a different interaction pattern at the interface, thereby explaining its high degree of immune escape and the rapid prevalence of this variant. The antibody S2E12, despite Delta string immune escape properties, was able to bind with a good affinity of -12.5 kcal/mol, showing promising results for future therapies.

Analyzing in greater detail the binding interface interaction network of VOCs during separation trajectories, we were able to establish that the WT, the Alpha, and Delta variants were all forming salt-bridges, e.g., D30:K417 and K31:E484, with higher occupancies in the case of the Alpha variant. It can partially explain why the Alpha variant has the highest binding affinity towards the human receptor among all VOCs. Interestingly, the Delta variant and WT occupancies are similar insofar as these two salt bridges are concerned. This resemblance could rationalize why the affinity did not increase when the Delta variant emerged. Our theoretical results confirm the observation of Bhattarai et al.[105] on these salt bridges' importance in the WT and the Alpha variants. Overall, this investigation has put forth the potency of the geometrical route to provide valuable insights into the underlying recognition and association processes of a protein-protein complex.

4.3 Original article

4.3.1 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), originated from Wuhan, China, in late December 2019,[106] is the virus that causes coronavirus disease 2019, or COVID-19, which has quickly escalated into a worldwide syndemic[107] and resulted in more than six million casualties.[108] This virion particle is composed of a nucleocapsid, membrane, spike (S), and envelope structural proteins. The cell entry is mediated through the S protein, which can be decomposed into two subunits, S1 and S2.[109,110,111] The first subunit contains the receptor-binding domain (RBD) responsible for the binding to the human receptor, the angiotensin convertor enzyme 2 (ACE2),[112,98] whereas the second subunit contains the cell fusion machinery and serves as an anchor to the membrane. A cleavage between the two subunits is necessary for infection, and is performed in host cells by enzymes, e.g., furin, prior to release.[113,114]

The virus keeps on mutating with new variants appearing on the World Health Organization (WHO) watch list, referred to as variants of concern (VOCs), thereby hampering the discovery of efficient therapies.[94,115,116] According to the Centers for Disease Control and Prevention (CDC), VOCs are characterized by either an increase in transmissibility, a more severe disease (i.e., more casualties, higher hospitalization rates), a reduced neutralization by antibodies, and reduced effectiveness of treatments, vaccines, or detection failures.[95]

Several mutations that confer a higher reproduction rate have been reported for the different variants. D614G, in particular, is known to promote S1/S2 cleavage, allowing a more exposed RBD, which favors binding to ACE2.[117,118] However, this mutation has been found in many variants, including all the VOCs, and cannot be the sole cause of the fitness difference between them. In general, mutations in the virus can

improve two properties, pivotal for higher infectivity, namely binding to ACE2 via its RBD,[109,119,112,98] or escaping the immune response.[120,121]

State-of-the-art free-energy calculations could play a crucial role in the quantitative prediction of binding affinities, and, thereupon, their comparison for the different VOCs binding through their RBDs to either ACE2, or a series of antibodies. One possible route to investigate how structural differences between VOCs affect recognition and association to the host cell consists in assessing relative binding affinities to ACE2 through in-silico point mutations[122,123,124] by means of alchemical transformations.[125,126] For instance, one of the rigorous studies of the mutation of N501 was performed by Fratev[123] using free-energy perturbation (FEP) calculations.[127,128,64] In this computational investigation, the N501Y substitution, observed in the Alpha variant, led to an enhanced affinity by 1.8 kcal/mol, compared to the wild-type (WT), a finding in line with that of Pavlova et al. using a similar strategy.[43] Additional predictive in-silico point mutations were carried out in VOCs, although they have yet to be documented thoroughly at the experimental level.[129,130,109]

The limitation of relative binding free-energy calculations lies in their provision of a binding-affinity difference based on a reference state. In other words, for a relative binding free-energy calculation, the reference state must be known beforehand. To circumvent this limitation, resorting to absolute binding free-energy calculations might be desirable. For instance, as a first step towards the quantitative assessment of the thermodynamics that underlies molecular association, Kim et al. turned to steered molecular dynamics (SMD)[131] to estimate the difference in binding strength for all the VOCs available at the time of their study.[132] They showed that the force needed to separate the Alpha variant RBD from ACE2 is the strongest amongst all investigated VOCs. In contrast, the difference in the force profiles between the WT and the Beta or Delta variant is marginal. These observations were experimentally validated by Koehler et al.[133] employing atomic force microscopy,[134] barring the Delta variant, which was not considered in their study. It is worth noting that SMD can, in principle, offer access to the binding free energy at the price of multiple realizations in a near-equilibrium regime, and application of the Jarzynski identity,[21,22] which was not performed by Kim et al.[132]

Conversely, potential-of-mean-force (PMF) calculations represent another option relying on first principles to obtain accurate binding affinities. García-Iriepa et al.[135] used a 1–μs simulation with the metadynamics-extended adaptive biasing force (meta-eABF) algorithm[136] to determine a binding free-energy of about -2.6 kcal/mol between the WT RBD and ACE2, suggestive of a rather innocuous viral load. Ngo et al.[137] estimated the energetic cost to fully dissociate the RBD from ACE2 to be 15 kcal/mol, while Chakraborty obtained a binding free-energy estimate of -34 kcal/mol,[138] turning to umbrella sampling[139] combined with the weighted histogram analysis method.[140,141] These discrepant results, at variance with the experimental findings,[98] may be ascribed to insufficient sampling, most notably to account for the slow reorientation of the binding partners, as well as premature and possibly incomplete structural data.

Another popular approach for the determination of absolute binding free energies is provided by molecular mechanics/Poisson-Boltzmann surface area (MM/PBSA), or molecular mechanics/ generalized Born surface area (MM/GBSA)[13,14,15,16,17] because of its inexpensive appealing nature. However, this methodology has proven at times to grossly overestimate standard binding affinities.[142,143,144] For instance, in the case of the WT:ACE2 complex examined by Khan et al., the binding affinity amounted to -64.0 kcal/mol,[145] at variance with the experimental data obtained by Lan et al. of -11.4 kcal/mol.[98] In all likelihood, both MM/PBSA and MM/GBSA are also vulnerable to the sampling issues and limitations of the structural data mentioned previously.

The present contribution aims to predict from first principles, with accurate structural data, the absolute binding free energies between the RBD of a number of SARS-CoV-2 VOCs and (i) the ACE2 protein and (ii) neutralizing antibodies, following the rigorous theoretical framework of the geometrical route,[44,57] to elucidate the role of point mutations in viral contagiousness. The RBDs investigated here

are those of the WT, the Alpha variant (December 2020), the Beta variant (December 2020), the Delta variant (May 2021), and Omicron BA.2 (November 2021).[94] The latter is known to spread faster than Delta and was categorized as a VOC within only a few days of its appearance.[94] Omicron BA.2 has 16 mutations in its RBD alone. Such a high number of mutations is unusual and has been predicted to change the antibody epitopes, and, therefore, to confer significantly higher immune escape properties than the Delta variant.[146]

Two antibody candidates, namely S2E12,[101] a neutralizing antibody bound to the Delta variant, dominant at the time this investigation was initiated, and H11–D4, which is bound to the WT,[102] were chosen to examine the immune–escape properties of the VOCs, as well as potential therapies. Additionally, the importance of using fully glycosylated models was investigated to assess whether the complete retinue of glycans is required to reproduce the experimental binding affinity, considering that these polysaccharides have proven crucial in prior studies for recognition and association to the host cell, as well as for escaping the immune response.[147,148]

4.3.2 Methods

The following subsections briefly recap the methodology employed in this work and its theoretical underpinnings, and describe the protocols of the simulations reported herein.

Binding free-energy calculations

The formation of a protein–protein complex involves significant conformational changes, encumbering the ergodic sampling of configurations within the simulation time amenable to brute-force MD. Importance sampling algorithms[62] can be used to accelerate the sampling of rare events, such as spontaneous binding. In these algorithms, external forces are applied onto collective variables (CVs), which consist of essential degrees of freedom involved in the reversible association, and can be controlled and monitored in a simulation. For instance, to speed up the reversible binding of two proteins, biasing forces can be applied onto the CV of their separation—i.e., the Euclidean distance between their centers of mass (COMs).

However, considering only the distance between the two COMs does not preclude random tumbling of the two binding partners across the reaction pathway, slowing down the convergence of the separation PMF calculation.[62] To circumvent this shortcoming, appropriate restraints acting on a set of additional CVs, representing the slow degrees of freedom of the complex, are applied in the course of the separation, as prescribed in the geometrical route introduced by Gumbart et al.,[57] the foundation of which can be found in reference.[44] These additional CVs are the backbone distance root-mean-square deviations (RMSDs) of the two proteins with respect to the reference, native conformation—i.e., in the complex, the three Euler angles describing their relative orientation, and two additional angles (polar and azimuth) for their relative position. Applying geometrical restraints in the form of soft, harmonic potentials onto these CVs is tantamount to a loss in configurational entropy, corresponding to conformational (ΔG_c^{site}), orientational (ΔG_o^{site}) and positional (ΔG_a^{site}) free-energy contributions, which must be accounted for in the computation of the standard binding free energy, both in the "bulk" (unbound state) and at the "site" (bound state). Therefore, the geometrical route consists of a series of independent PMF calculations determined sequentially with the progressive introduction of restraints as a preamble to performing the separation PMF calculation. The binding free energy can then be expressed as a sum of these different free-energy contributions:

$$\Delta G_{\text{b}}^{\circ} = \Delta G_c^{\text{bulk}} - \Delta G_c^{\text{site}} - \Delta G_o^{\text{site}} - \Delta G_a^{\text{site}}(\theta,\phi) + \Delta G_o^{\text{bulk}} - \frac{1}{\beta}\ln(S^* I^* C^{\circ}) \tag{4.1}$$

where $\beta = (k_{\mathrm{B}}T)^{-1}$, with k_{B}, the Boltzmann constant, and T, the temperature. C° denotes the standard concentration of 1 M, that is C° = 1/1661 (Å^3).[89] I^* is the separation term, and S^* is a surface term, which represents the fraction of a sphere of radius r^*, centered at the binding site of the reference protein, accessible to its partner, that is,

$$\begin{cases} I^* &= \displaystyle\int_{\mathrm{site}} \mathrm{d}r\ \mathrm{e}^{-\beta\left(w(r)-w(r^*)\right)} \\ S^* &= r^{*2}\displaystyle\int_0^{\pi} \mathrm{d}\theta\ \sin\theta \int_0^{2\pi} \mathrm{d}\phi\ \mathrm{e}^{-\beta u_a} \end{cases} \tag{4.2}$$

Here, r^* is a point far from the binding site, where the proteins no longer interact with each other, u_a is the sum of the harmonic restraint potentials of polar angles θ and ϕ.

In some cases, additional RMSD restraints acting on protein–protein interfacial side chains may be required due to their exposure to the solvent, and the possibility of their isomerization in the course of the separation, resulting in a loss of interaction, and a progressive deterioration of the binding free-energy estimate.[57]

Computational assays

The structure of the WT:ACE2 complex was taken from the Protein Data Bank (PDB) source 6M0J.[98] Our structures of the Alpha, Beta and Delta variants were constructed by introducing single-point mutations in the WT RBD, namely (i) N501Y for the Alpha variant, (ii) N501Y, E484K, K417N for the Beta variant, and (iii) T487K and L452R for the Delta variant[149,109] using VMD[90] and called, respectively, Alpha$_{\mathrm{model}}$, Beta$_{\mathrm{model}}$ and Delta$_{\mathrm{model}}$.

The Omicron BA.2 variant differs significantly from its fellow VOCs and contains 16 mutations in its RBD alone (G333D, S371F, S373P, S375F, T376A, D405N, R408S, K417N, N440K, S477N, T478K, E484A, Q493R, Q498R, N501Y, and Y505H). A configurational change with a better packing of the N-terminal chain of ACE2 in comparison to the WT has been observed, and could thermodynamically favor intermolecular interactions.[150] Such discrepancy with the WT strain dictated the choice of using the experimental X-ray structure of the BA.2 sub-variant deposited in the PDB (7ZF7).[100]

As glycans were shown to play an important role in recognition and association,[147,148] the consequence of their presence on the binding affinity was also investigated in this work. A model of the WT:ACE2 complex with full-length polysaccharide chains was generated using a previously constructed model (see ACE2 Scheme #1 in reference[151]) relying on the initial structure 6M17.[152] We note that the two complexes in PDBs 6M17 and 6M0J are nearly identical (RMSD of 1.2 Å) and likely represent the same energetic minimum. The antibody assays were prepared with the Charmm-GUI webserver.[153] Point mutations in the RBD (viz. T478K, L452R) were introduced to generate the S2E12:Delta complex by using the WT structure deposited in the PDB (7R6X) as a template.[101] The H11-D4:WT complex was obtained from the PDB (6YZ5).

While the calculations with our models were ongoing, new experimental structures for the Alpha, Beta, and Delta variants were reported.[97,103,104] As a measure of precaution, we checked our models against these new structures to ensure that the altered interfaces were consistent with all available experimental data. For the Delta variant, the interacting pattern shown by McCallum et al. was identical to ours and the Delta$_{\mathrm{model}}$ retained.[97] The experimental structure of the Alpha variant was published in the PDB (7EDJ) by Yang et al.[103] A short SMD simulation was performed to correct our interface, and mirror the interaction pattern of their structure in our Alpha$_{\mathrm{model}}$ structure. The Beta variant structure was deposited in the PDB 7SY6 in late December 2021 by Mannar et al.[104] An alignment revealed a local rearrangement at the interface centered about residue H34 of ACE2 in the experimental structure,

and absent in our model. This conformational change is specific to the Beta variant, and is not present in other VOCs, as documented by the authors.[104] While our initial model of the Beta variant, on the one hand, and the early structure of the WT taken from the Covid-19 Charmm-GUI archive, on the other hand, depart from the more recent experimental structures, we have decided to report the corresponding free-energy calculations to showcase the influence of the starting structure on the binding affinity using both models ($\mathrm{WT}_{\mathrm{model}}$, $\mathrm{Beta}_{\mathrm{model}}$) and experimental structures ($\mathrm{WT}_{\mathrm{crystal}}$, $\mathrm{Beta}_{\mathrm{Cryo\text{-}EM}}$).

Molecular dynamics simulations

The macromolecular CHARMM36[84] force field and the TIP3P model[85] were used to describe the proteins and the water, respectively. The zinc ion present in the ACE2 protein was kept due to its catalytic importance in the ACE2 function.[154] The parameters of the zinc ion were taken from the zinc AMBER force field (ZAFF)[155] for a cation chelated by two histidine and two glutamate residues, corresponding best to the coordination pattern in our ACE2 structure.

All simulations were performed using the NAMD 3.0 MD engine.[86] All computational assays corresponded to a physiological concentration of NaCl of 0.15 M. They were minimized for 500 steps, prior to a 100–ns pre-equilibration in the isothermal-isobaric ensemble, keeping the temperature (300 K) and the pressure (1 atm) constant by means of a Langevin thermostat[156] and the Langevin piston algorithm,[157] respectively. The particle-mesh Ewald (PME) algorithm[158] was utilized to handle long-range electrostatic interactions. Van der Waals and short-range electrostatic interactions were truncated with a smoothed 12-Å spherical cutoff. The equations of motion were integrated with a 2-fs time step. The coordinates and force-field parameters from the pre-equilibration steps were used as inputs in the binding free-energy estimator 2 (BFEE2),[81] a tool for streamlining and automating the setup of binding free-energy calculations, originally designed to tackle protein–ligand complexes. To expand the BFEE2 applicability to protein–protein complexes, RMSD calculations of the backbone of each protein were included, both in the bulk aqueous medium and at the binding site. The importance-sampling algorithm employed for the computation of the PMFs was the well-tempered extended adaptive biasing force algorithm (WTM-eABF)[49] as implemented in the collective variable module (Colvars) of NAMD.[93] The PMFs were run sequentially for all complexes, starting from the distance RMSD of the proteins with respect to the native conformation up to the physical separation of two proteins. Only the bound state RMSDs of the $\mathrm{Alpha}_{\mathrm{model}}$ and both Beta variants required modifications of the default parameters to address convergence issues (see Supporting Information (SI) for additional details). Once all the simulations were completed, BFEE2[81] was invoked again for the post-treatment of the PMFs to extract the individual contributions of the binding affinity, and to infer the final binding free-energy estimate (see SI for details of the computation).

4.3.3 Results and Discussion

Binding free-energy calculations

The binding free-energy estimates for the different complexes are gathered in Table 4.2, and nearly all match the experimental measurements within chemical accuracy, except for the early WT:ACE2 model. The reason for the significant discrepancy between the theoretical and experimental $\Delta G^{\circ}_{\mathrm{b}}$ stems from local differences between the model and the experimental structure. Since models can miss crucial interactions, their use magnifies the vulnerability of free-energy calculations to inadequate initial structures, and the likelihood of erroneous binding-affinity estimates. However, in some cases, such as the Beta variant model in complex with ACE2, the structural differences discussed previously did not affect at first sight the theoretical $\Delta G^{\circ}_{\mathrm{b}}$. Replacing key interactions with a number of poorly predicted side-chain interactions could result in a fortuitous cancellation of errors, with no guarantee of recovering the correct

network of nonbonded interactions, and, hence, the correct standard binding free energy. It is noteworthy that accurate structure modeling is not only a prerequisite to the geometrical route, but, in general, to all MD-based binding free-energy strategies.

Complexes	$\Delta G_{\mathrm{b}}^{\circ}$ (kcal/mol)	$\Delta G_{\mathrm{exp}}^{\circ}$ (kcal/mol)
$\mathrm{WT_{crystal}}$:ACE2	-11.5 ± 0.3	-11.4[98]
$\mathrm{WT_{model}}$:ACE2	-6.7 ± 2.3	-11.4[98]
$\mathrm{Alpha_{model}}$:ACE2	-12.3 ± 1.2	-11.6[99]
$\mathrm{Beta_{model}}$:ACE2	-10.0 ± 1.2	-11.1[99]
$\mathrm{Beta_{Cryo\text{-}EM}}$:ACE2	-11.0 ± 1.6	-11.1[99]
$\mathrm{Delta_{model}}$:ACE2	-9.6 ± 0.5	-9.9[97]
Omicron BA.2:ACE2	-11.4 ± 1.3	-11.5[100]
WT:ACE2 (full-length glycans)	-10.8 ± 0.3	-11.4[98]
S2E12:Delta	-12.5 ± 0.3	-12.0[101]
H11-D4:WT	-9.4 ± 0.5	-9.9[102]

Table 4.2: Computed binding free energy against experimental values of all studied complexes

Comparing the $\Delta G_{\mathrm{b}}^{\circ}$ of all VOCs in complex with ACE2, it is apparent that the Delta variant possesses the lowest binding affinity for the receptor (viz. -9.6 kcal/mol), which is almost 2 kcal/mol weaker than that of the WT (viz. -11.5 kcal/mol) (see Table 4.2). Moreover, Mlcochova et al.[159] showed experimentally that the Delta variant does not exhibit a higher affinity towards human ACE2 than both the Alpha variant (viz. -11.6 kcal/mol)[99] and the WT. Their findings corroborate our calculations, from whence we can conclude that the Delta variant increases its fitness over other VOCs by relying more on immune escape than on increased affinity. This result explains the rapid prevalence of the Delta variant over previous VOCs, notwithstanding the increased vaccination rate amid the population.[160,118]

The Omicron BA.2 variant has been reported both by experimental and theoretical studies to have either an enhanced affinity, or an affinity similar to that of the Delta variant, causing some debate on the actual affinity of Omicron for ACE2.[161,162,163,164,165,113] Our results confirm an affinity close to that of the WT, notwithstanding their very different binding interfaces. As can be seen from Figure 4.1, the separation PMFs for Delta and Omicron BA.2 are strikingly similar, which could explain the uncertainty on the reported binding affinity of the two VOCs.[161,162,163,164,165,113]

As depicted in Figure 4.1, among all PMFs underlying the separation of the VOCs from ACE2, Delta possesses the largest well depth, in excess of -26 kcal/mol, which is almost 2.5 times the absolute value of the final estimate (viz. -9.6 kcal/mol). This result underscores the importance of accounting for all the degrees of freedom other than the physical separation of the binding partners in standard binding free-energy calculations. More specifically, the PMFs along the RMSDs (see Table S6) contribute significantly to the binding affinity. It is also worth noting that most of the computational effort (viz. 60% of the total simulation time for the WT) was invested in recovering the conformational entropy contributions in the bulk and at the binding site. The Alpha and model Beta (referred to as $\mathrm{Beta_{model}}$) variants correspond to similar PMFs (see Figure 4.1). Their well depths differ by less than 2 kcal/mol. A possible reason for this discrepancy lies in using the same initial template and the shared N501Y mutation, which significantly increases the binding affinity, as reported in the literature.[123,119,43]

Furthermore, when comparing the different free-energy contributions for the variants (see Table S11), we notice that orientational and positional contributions in the bound states are both similar and small when compared to the other contributions. The conformational and the physical separation vary the most between VOCs. The difference in the conformational contribution arises from the mutations that confers differences in the inner flexibility/stability of the proteins. The $\mathrm{WT_{model}}$ conformational contribution is

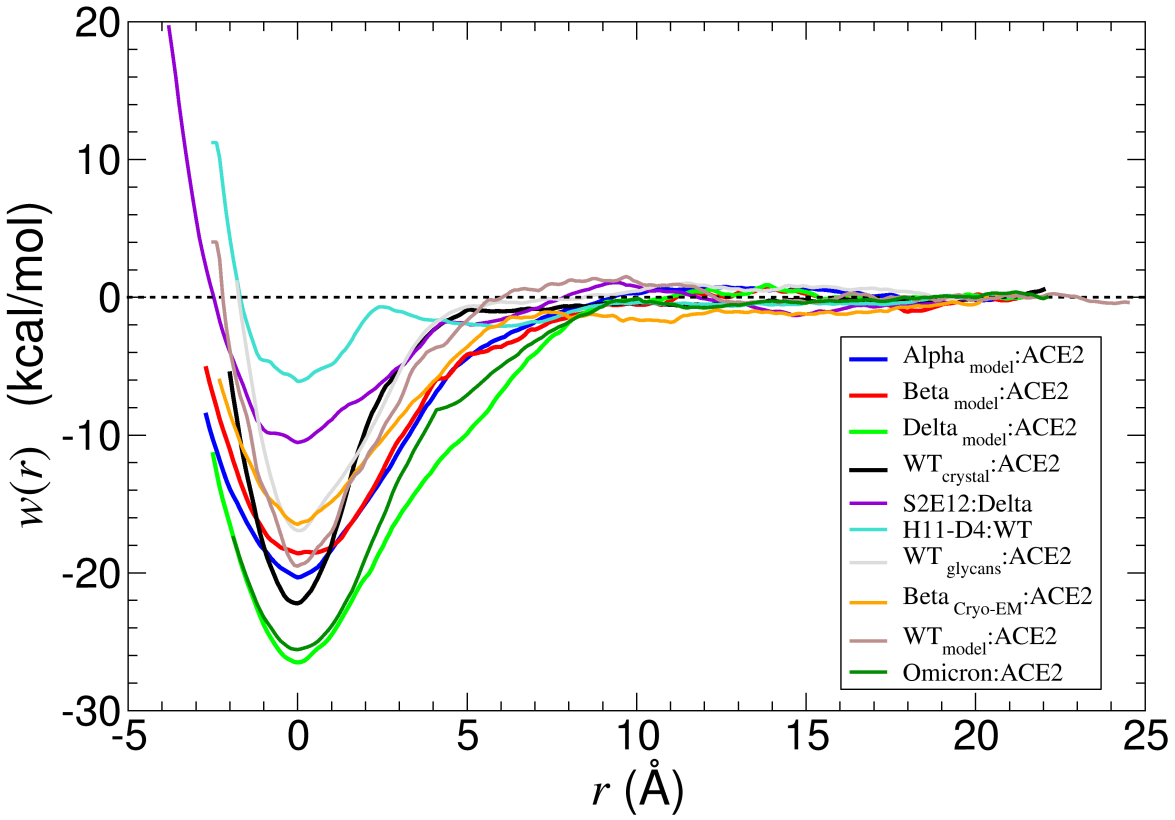

Figure 4.1: PMFs obtained during the reversible separation of ACE2 and the RBD of the WT (with minimal glycans in black, fully glycosylated in grey and the early model in taupe) and the different variants ($\text{Alpha}_{\text{model}}$: blue, $\text{Beta}_{\text{model}}$: red, $\text{Beta}_{\text{Cryo-EM}}$: 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 $r = 0$.

significantly higher than that for the rest of the variants due to the need of stronger restraints to assuage convergence issues, as stated in the SI.

The WT:ACE2 complex structure used here is minimally glycosylated, with only one glycan found at four glycosylation sites on ACE2 (i.e., N53, N90, N322, and N546) and one site on the RBD (N343). Additional glycans, while present, were not resolved, likely due to their flexibility. To determine the effect of full-length glycans on binding, if any, we repeated the calculation of the binding free energy between the RBD and ACE2 using a fully glycosylated model of the complex constructed previously (see ACE2 Scheme #1 in reference[151]). This model has 8–10 sugars at each site—the ones mentioned already, as well as two others on ACE2 (namely, at N103 and N432), as depicted in Figure Figure 4.2A. Following the same steps as in the other calculations, we found a slightly smaller (about -1 kcal/mol) binding free energy at -10.8 kcal/mol, compared to the minimally glycosylated model (-11.5 kcal/mol). This is in subtle contrast to recent experiments, in which the presence of glycans was 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).[166] Regardless, our calculated binding free energy for the fully glycosylated complex falls within the range observed experimentally, which spans $4k_{\text{B}}T$ (-9.0 kcal/mol,[167] -10.3 kcal/mol,[166] and -11.4 kcal/mol[98]).

To determine the efficiency of antibodies against SARS-CoV-2 variants, we selected two complexes,

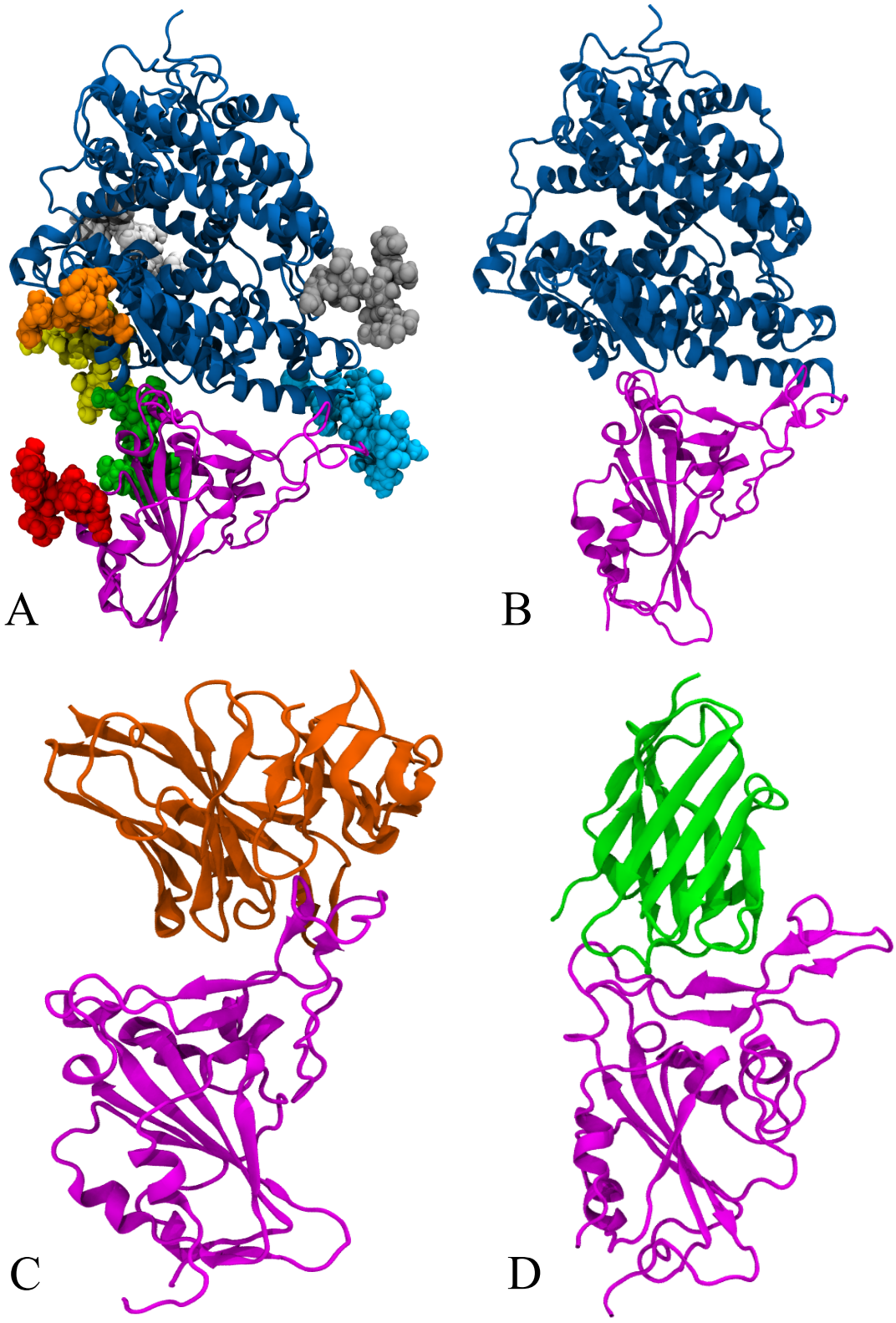

Figure 4.2: Representation of the binding mode of (A) WT:ACE2 with glycans, (B) WT:ACE2, (C) WT:H11-D4, (D) Delta:S2E12.

namely the neutralizing nanobody H11-D4 bound to the WT,[102] and the human antibody S2E12 bound to the Delta variant.[101] The experimental binding free energy taken as the reference in Table 4.2 for the S2E12:Delta complex was inferred from a neutralization curve reported by Mlcochova et al.,[159] 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. The binding poses of the antibody complexes are depicted in Figure 4.2. H11-D4 binds the WT by forming several hydrogen bonds (S494:V102, E484:S57, E484:R52, F490:S104, Q493:S104), and a salt bridge (R52:E484). Stacking of Y449 onto N101 also participates in the binding.[102] The S2E12 binding site is centered about P486, with a cavity lined with aromatic residues[101] as shown in Figure 4.3.

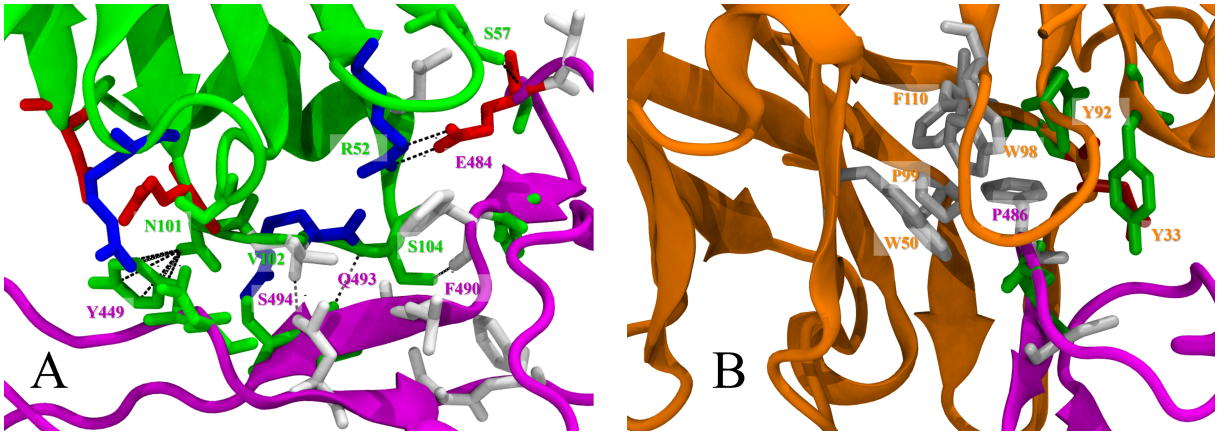

Figure 4.3: Interaction interfaces of (A) antibody H11-D4 and the WT RBD and (B) antibody S2E12 and the Delta RBD, with salt bridges and hydrogen bonds highlighted with dotted lines.

The calculated binding free energies for the antibody complexes matched the experimental data within chemical accuracy (see Table 4.2). The binding affinity for the H11-D4:WT complex amounts to -9.4 kcal/mol, which is less than that for WT:ACE2 (i.e., -11.5 kcal/mol). These results emphasize the insufficient neutralizing activity of H11-D4 against SARS-CoV-2 in its WT strain, at least absent for significantly higher concentrations than ACE2. Huo et al.[102] reached the same conclusion based on the nonconservation of the H11-D4 epitope between SARS-CoV and SARS-CoV-2, from whence they recommended the use of H11-D4 in cocktails of antibodies binding different regions of SARS-CoV-2 to boost the efficiency of the treatment. The binding affinity of S2E12 to the Delta variant reported by Mlcochova et al.,[159] was confirmed by our binding free-energy calculations. The large $\Delta G^{\circ}_{\mathrm{b}}$ for S2E12:Delta complex, amounting to -12.5 kcal/mol, is appreciably greater than that for Delta:ACE2 (-9.6 kcal/mol). Put together, although the Delta variant does not lean on a strong binding to ACE2, but rather on immunity escape,[159,118] the S2E12 antibody seems to be effective against SARS-CoV-2 infection induced by the Delta variant. Starr et al.[101] showed that unlike other antibodies examined in their study, S2E12 was able to bind to a gamut of SARS-CoV-2 variants. They stated that its efficiency could be linked to the scarcity of S2E12 in polyclonal sera, as well as to the lack of evolutionary pressure by this antibody to SARS-CoV-2. In summary, S2E12 is a therapeutic candidate that could potentially withstand the appearance of new variants, while retaining a reasonable efficacy against the virus.[96] 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),[168] confirming our previous statement.

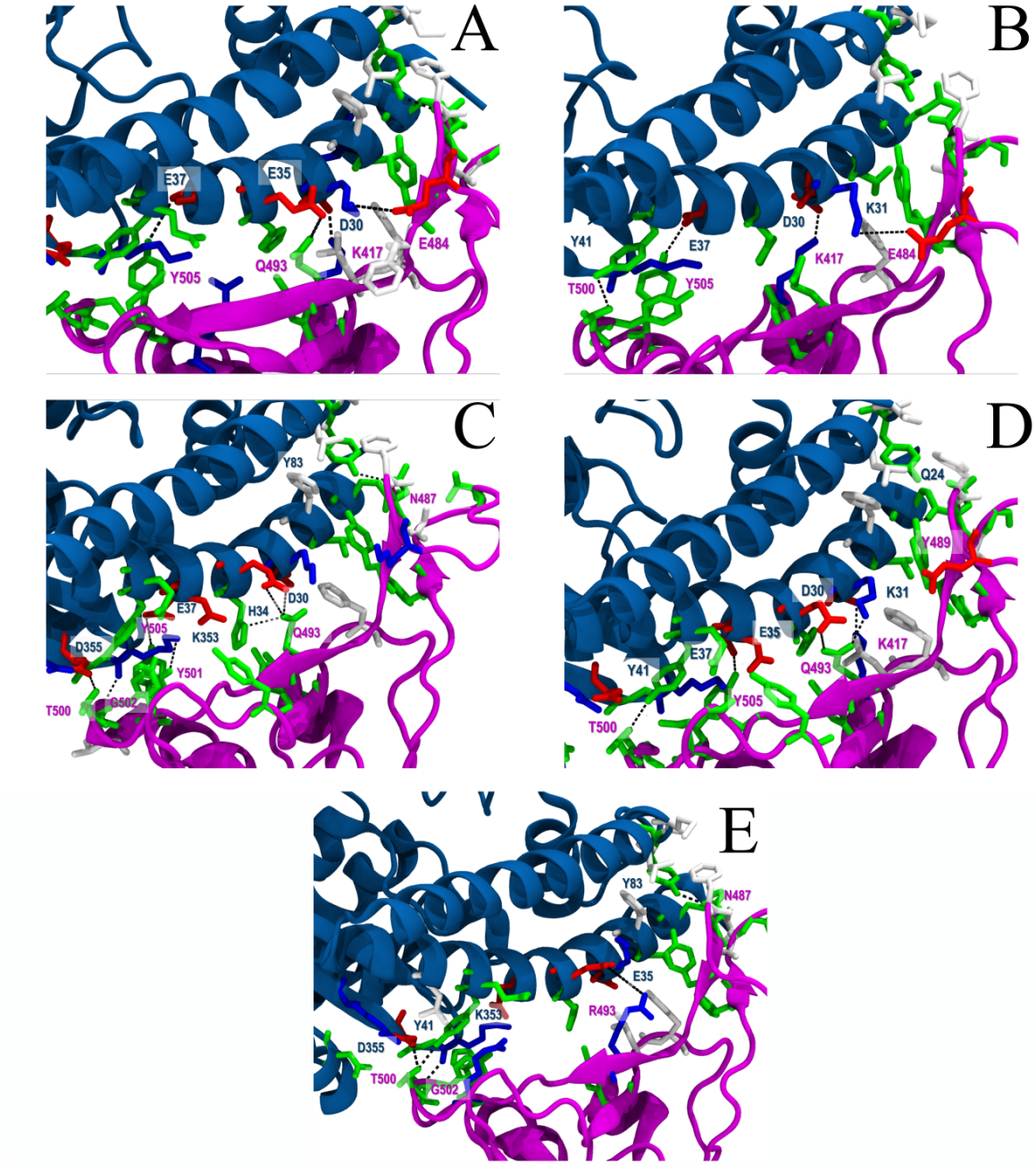

Figure 4.4: Interaction interface with ACE2 of (A) WT, (B) Alpha, (C) Beta, (D) Delta, and (E) Omicron BA.2 variants with salt bridges and hydrogen bonds highlighted with dotted lines.

Protein–protein interaction networks

To understand the consequences of mutations in terms of binding affinity, we analyzed the separation trajectories in greater detail, focusing on the networks of interactions consisting of salt bridges and hydrogen bonds. In our simulations, we observed the formation of D30:K417 and K31:E484 salt bridges (see Figure 4.4 and Figure 4.5) in Alpha:ACE2, Delta:ACE2 and WT:ACE2 complexes when the two proteins are in intimate contact (i.e., COM distance < 50 Å). As shown in Figure 4.5, the Alpha variant

exhibits higher occupancy for both salt bridges. It can partially explain why the Alpha variant has the highest binding affinity towards the human receptor among all VOCs. Interestingly enough, the histograms for the Delta variant and the WT are almost similar insofar as these two salt bridges are concerned. This resemblance could rationalize why the affinity did not increase when the Delta variant emerged. Our theoretical results confirm the observation of Bhattarai et al.[105] on the importance of the presence of these salt bridges in the case of the WT and the Alpha variant. However, these non-covalent interactions are no longer present in the Beta variant owing to the K417N and E484K mutations. According to the experimental data—and to our simulations, the binding free energies of the Alpha and Beta variants are similar.[99] The loss of the D30:K417 and K31:E484 salt bridges in the Beta variant, either in the Cryo-EM structure (referred to as $\text{Beta}_{\text{Cryo-EM}}$) or in the model ($\text{Beta}_{\text{model}}$), might be compensated by the formation of novel interactions, e.g., a salt bridge detected by Socher et al.[169] between K484 and E75. This particular salt bridge was only detected in the case of the $\text{Beta}_{\text{Cryo-EM}}$:ACE2 complex and in a small number of configurations in our separation trajectory. Luan et al. have also observed the formation of this salt bridge after 190 ns of simulation, and reasoned that exposure to water weakened and even broke this interaction,[121] which rationalizes its scarcity in our own simulations. This new interaction could also explain why the loss of K31:E484 and D30:K417 did not result in any significant decrease in the binding affinity for the $\text{Beta}_{\text{Cryo-EM}}$ variant. This discrepancy in the detection of salt bridges between the two Beta variant structures underscores again the need for an appropriate starting structure.

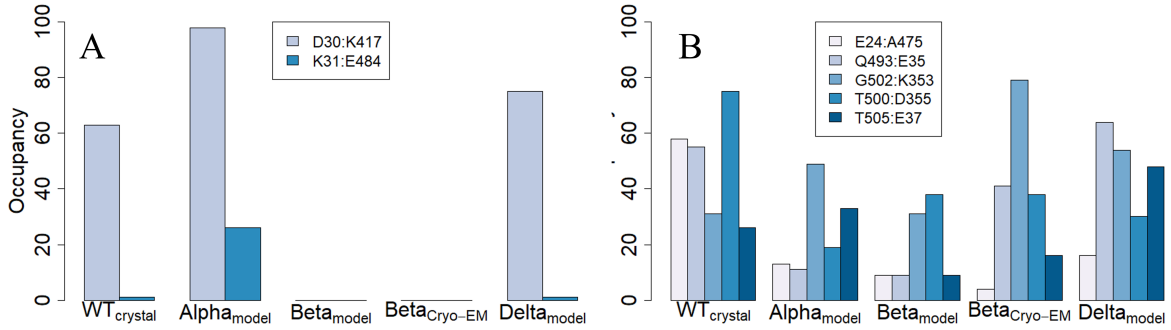

Figure 4.5: (A) Occupancy of the salt bridges in the separation trajectories for the WT and the studied variants computed at a close center-of-mass distance (<50 Å). (B) Occupancy of the hydrogen bonds in the separation trajectories for the WT and the studied variants computed at a close center-of-mass distance (<50 Å).

The most important hydrogen-bond interactions that occurred at the interface between ACE2 and the different VOCs are depicted in Figure 4.5, where substantial discrepancies in occupancy are visible when the partners are in intimate contact (distance between COMs < 50 Å). For instance, the population of the E24:A475 hydrogen bond is strongly reduced in the different variants (viz. 58% for the WT against 13%, 9%, 4% and 16% for the Alpha, $\text{Beta}_{\text{model}}$, $\text{Beta}_{\text{Cryo-EM}}$, and Delta variants, respectively). The Beta variant model exhibits a hydrogen-bond pattern similar to that of its Alpha counterpart for the first three bonds (i.e., E24:A475, Q493:E35, and G502:K353), which is another argument in favor of the similarity of the PMFs, and the use of the same initial template. The discrepant hydrogen-bond occupancy between the two Beta structures is evidenced in Figure 4.5, which highlights a lower occupancy for all bonds in the case of the model structure, barring the initial one, that is E24:ALA475. This result implies that misplaced side chains in the model may result in a destabilized interface, and explain the ca. 1 kcal/mol difference between their $\Delta G_{\text{b}}^{\circ}$'s (see Table 4.2).

The high number of mutations harbored by Omicron in the RBD is responsible for a totally different binding interface, thus preventing a direct comparison of interaction networks between this variant and the other VOCs. A compensation between favorable, e.g., N501Y, and destabilizing mutations, e.g., K417N, is responsible for maintaining appropriate binding to ACE2, while generating significant immune escape, as revealed by structural analysis.[170,100] Omicron BA.2 does not possess any of the salt bridges previously mentioned owing to the Q493R, K417N and E484A mutations. However, a new salt bridge formed between E35 and R493 was detected in our simulations. Some hydrogen bonds were conserved when compared to the WT, like G502:K353 and T500:D355.

4.3.4 Conclusions

The COVID-19 crisis has exerted tremendous pressure on the scientific community to obtain fast and accurate results to help understand and battle the SARS-CoV-2 virus. The syndemic started over two years ago, and a wealth of data has been published during this time interval. Determination of new structures of the VOCs bound to either ACE2 or antibodies has made possible the theoretical investigation reported herein, offering a critical assessment of the predictive power of a state-of-the-art methodology for protein–protein standard binding free-energy calculations. We have applied a rigorous computational protocol leaning on the so-called geometrical route to determine the binding free energies of VOC:ACE2 complexes, where VOCs stand for the RBD of the WT, Alpha, Beta, Delta, and Omicron BA.2 variants, and antibodies:VOC (H11-D4:WT, S2E12:Delta) complexes, and underscored the importance of the contribution arising from all degrees of freedom other than the physical separation of the binding partners, most notably the conformational ones. Except for the model WT:ACE2 complex, the experimental binding free energies of all complexes were reproduced in approximately 1–μs–long simulations up to chemical accuracy, emphasizing both the robustness and the potency of the method. The Beta variant model failed to yield the experimental affinity, albeit falling within the $k_{\mathrm{B}}T$ margin, which stems from fortuitous cancellations of errors rooted in an incorrect starting structure. The discrepancies between the model and the Cryo-EM structures underscore the paramount importance of the starting point. Models—based, for instance, on simple amino-acid replacements, as well as docked structures—are often produced in the absence of structural data, but, owing to their questionable accuracy, are of limited use, regardless of how predictive the methodology at hand is, as was demonstrated cogently in this work. Comparing the affinity obtained for all VOCs, using WT:ACE2 as a reference, we established that both the Alpha and the Beta variants rely on an increased affinity for ACE2. In contrast, the Delta variant has a lower affinity for ACE2, and has spread widely owing to its immune-escape properties. The Omicron variant has a binding affinity similar to that of the WT, albeit harboring a different interaction pattern at the interface, thereby explaining its high degree of immune escape and the rapid prevalence of this variant. The Alpha and the Beta variant share a common mutation, namely N501Y, which explains, at least in part, the resemblance of their PMFs, and, thus, their similar binding affinities. However, the Beta variant has additional specific mutations, viz. E484K and K417N, found to compensate each other,[171] and, as a result, these mutations do not significantly affect its binding free energy when compared to the Alpha variant.[171] Inasmuch as the complexes with an antibody are concerned, the present investigation indicates that S2E12 has a strong affinity for the Delta variant and, thus, represents a potential candidate for COVID-19 therapies, irrespective of the variant at play, being in principle able to withstand the emergence of new mutations.[96] Notwithstanding the vulnerability to the initial structure of MD-based free-energy calculations, in general, and of the geometrical route, in particular, the latter methodology, when applied rigorously, devoid of shortcuts, constitutes a reliable approach to predict the binding affinity of existing and ever-emerging variants of SARS-CoV-2 towards the host cell, whilst offering valuable atomistic insights into the underlying recognition and association processes.

4.4 Supporting Informations

The supporting information comprises the description of the molecular assemblies, computational details (individual contributions and convergence) of all free-energy calculations.

4.4.1 WT SARS-CoV-2 RBD : ACE2 (6m0j)

Molecular assembly details

The starting coordinates used to build the molecular assembly were taken from the crystallographic entry 6m0j in the PDB resolved at 2.5 Å.[98] All amino acids were protonated in accordance with a pH of 7. To resolve the histidine protonation ambiguity linked to its pK_a (around 6) and determine the appropriate position of the proton between the delta and epsilon nitrogen atoms of the imidazole ring, additional visual inspection of the structure and probability of hydrogen bond was considered. The structure was then solvated in a cubic box, and an ionic strength of 0.15 NaCl was added to mimic physiological conditions resulting in a total of 270176 atoms. The dimensions of the periodic cell are 121 x 133 x 175 $Å^3$.

4.4.2 Result details

The detailed results of the diverse contributions is presented in Table 4.3 and individual components PMFs in Figure 4.6 with the associated convergence in Figure 4.7.

Table 4.3: 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)
$\Delta G^{\text{site}}_{c(\text{RBD})}$	-7.0 ± 0.1	120
$\Delta G^{\text{site}}_{c(\text{ACE})}$	-6.8 ± 0.1	120
$\Delta G^{\text{site}}_{\Theta}$	-0.3 ± 0.0	80
$\Delta G^{\text{site}}_{\Phi}$	-0.1 ± 0.0	40
$\Delta G^{\text{site}}_{\Psi}$	-0.1 ± 0.0	40
$\Delta G^{\text{site}}_{\theta}$	-0.3 ± 0.0	90
$\Delta G^{\text{site}}_{\phi}$	-0.3 ± 0.0	50
$(1/\text{beta}) * \ln(\text{S} * \text{I} * \text{C}_0)$	-21.7 ± 0.0	120
$\Delta G^{\text{bulk}}_{c(\text{RBD})}$	5.7 ± 0.1	200
$\Delta G^{\text{bulk}}_{c(\text{ACE2})}$	12.8 ± 0.0	200
$\Delta G^{\text{bulk}}_{o}$	6.6	
ΔG_{b}	-11.5 ± 0.3 (calculation) -11.4 (experiment)[98]	1060

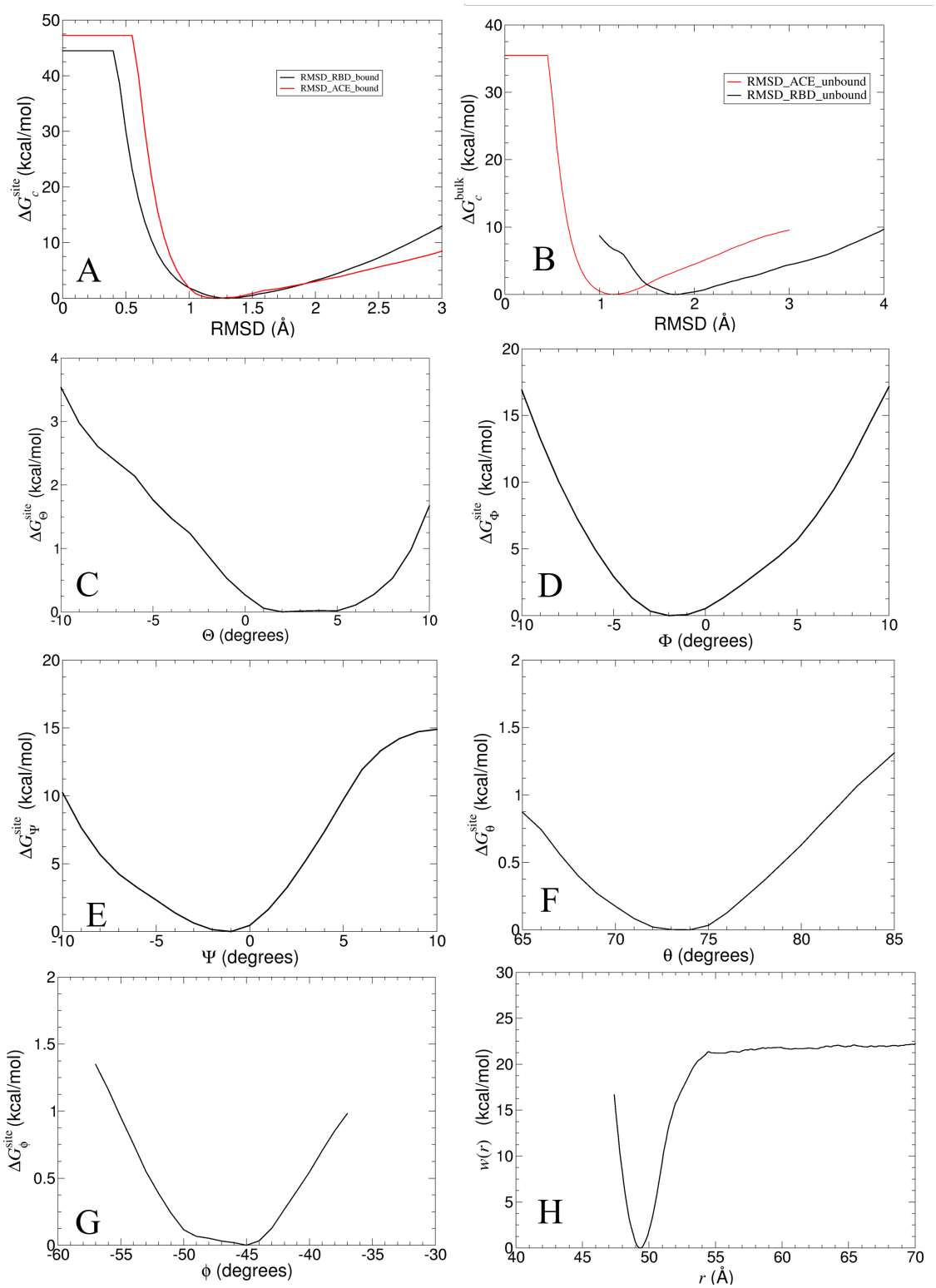

Figure 4.6: Individual PMFs for all components. The PMF calculations using RMSDs of the WT RBD and ACE2 proteins in the bound (A), and unbound state (B), Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

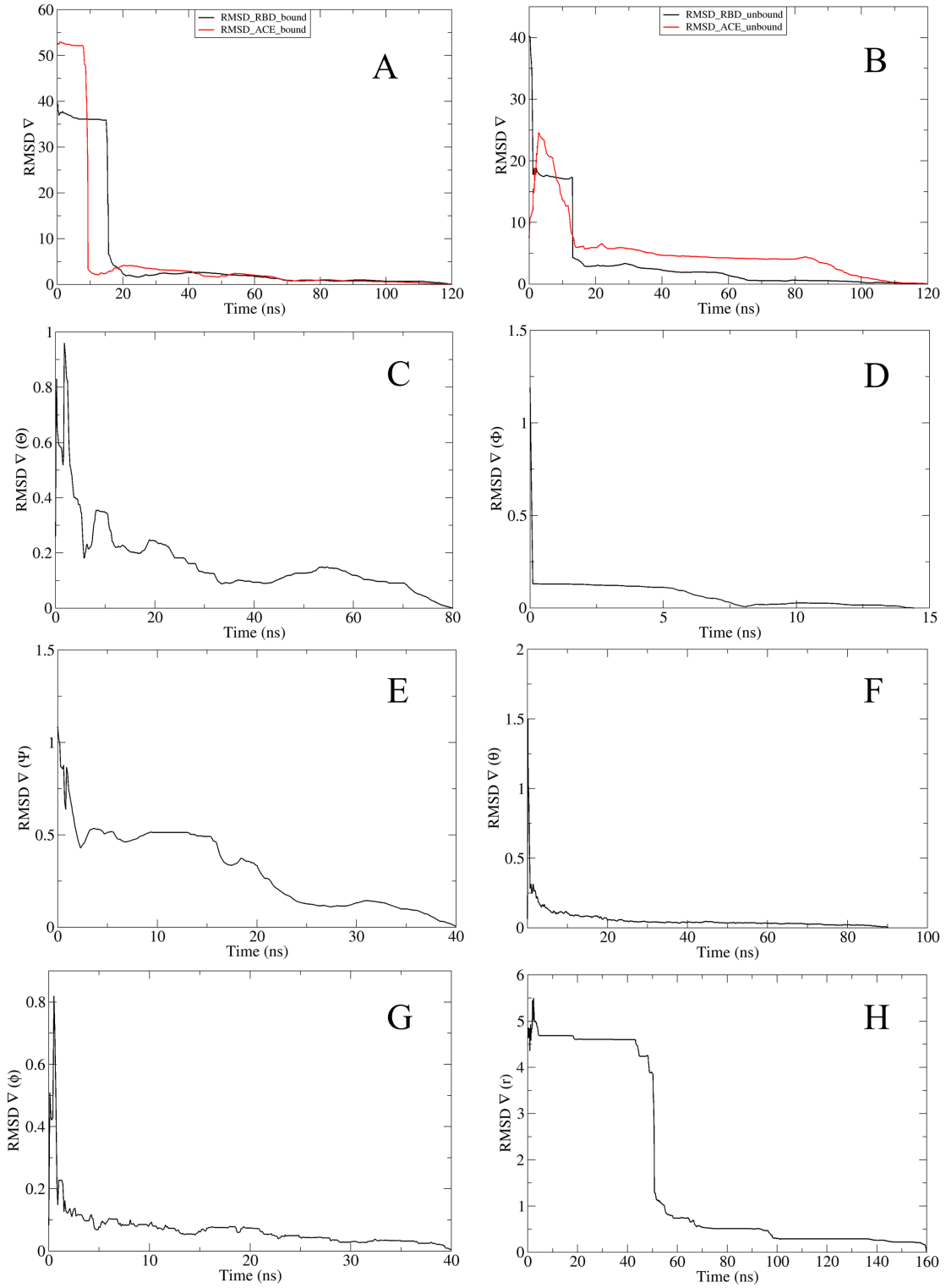

Figure 4.7: Convergence curve for individual PMFs for all components using RMSDs of the WT RBD and ACE2 proteins in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

4.4.3 WT SARS-CoV-2 RBD : ACE2 (model)

Molecular assembly details

The molecular assembly of the model WT with ACE2 was built using 6M017 as a template and taken from the CHARMM GUI archive. The molecular assembly was then solvated in cubic box with an ionic strength of 0.15 NaCl to mimic physiologic conditions. The dimension of the periodic cell are 130 x 130 x 128 $Å^3$ and the total number of atoms is 205925. The use of a stronger force constant (100 kcal/mol) was needed to prevent convergence and sampling issues in the case of the RMSDs.

Result details

The detailed results of the diverse contributions is presented in Table 4.4 and individual components PMFs in Figure 4.8 with the associated convergence in Figure 4.9.

Table 4.4: 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)
$\Delta G^{\text{site}}_{c(\text{RBD})}$	-37.0 ± 0.1	800
$\Delta G^{\text{site}}_{c(\text{ACE})}$	-64.5 ± 1.1	200
$\Delta G^{\text{site}}_{\Theta}$	-0.3 ± 0.0	200
$\Delta G^{\text{site}}_{\Phi}$	-0.5 ± 0.1	40
$\Delta G^{\text{site}}_{\Psi}$	-0.2 ± 0.3	40
$\Delta G^{\text{site}}_{\theta}$	-0.3 ± 0.2	200
$\Delta G^{\text{site}}_{\phi}$	-0.9 ± 0.0	184
$(1/\text{beta}) * \ln(\text{S} * \text{I} * \text{C}_0)$	$-14{,}3 \pm 0.2$	339
$\Delta G^{\text{bulk}}_{c(\text{RBD})}$	36.1 ± 0.1	200
$\Delta G^{\text{bulk}}_{c(\text{ACE2})}$	66.9 ± 0.2	200
$\Delta G^{\text{bulk}}_{o}$	8.28	
ΔG_{b}	-6.7 ± 2.3 (calculation) -11.4 (experiment)[98]	2403

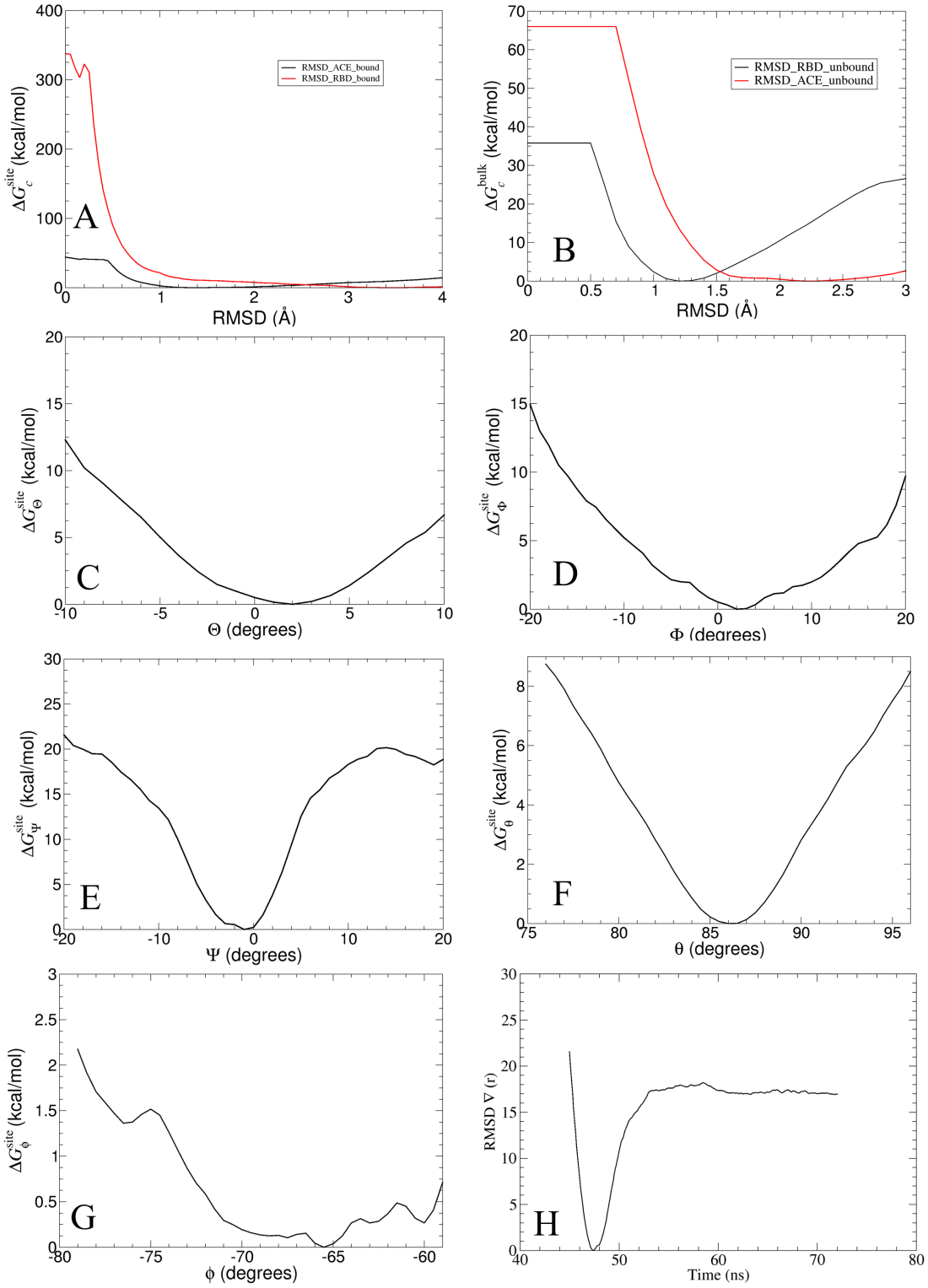

Figure 4.8: Individual PMFs for all components. The PMF calculations using RMSDs of the WT RBD model and ACE2 proteins in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

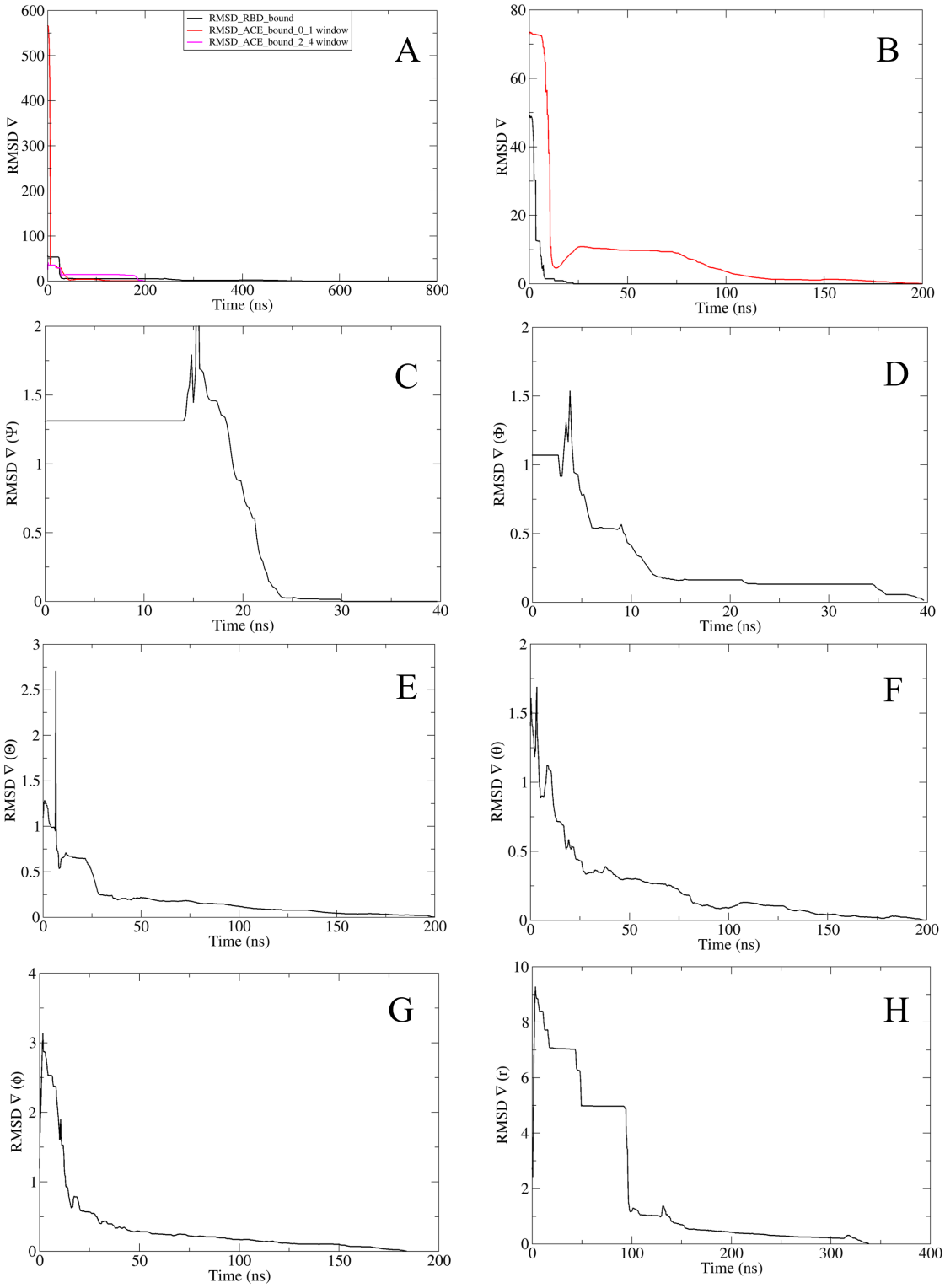

Figure 4.9: Convergence curve for individual PMFs for all components using RMSDs of the WT RBD model and ACE2 proteins in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

4.4.4 Alpha SARS-CoV-2 RBD : ACE2

Molecular assembly details

The single points mutation N501Y was introduced in the WT RBD in complex with ACE2 to generate the Alpha RBD:ACE2 variant. The complex was then solvated in a cubic box of water with an ionic strenght of 0.15 NaCl to mimic physiologic conditions resulting in a periodic cell of 130 x 130 x 128 $Å^3$ and 205932 atoms.

An experimental structure of the alpha variant was published by Yang et al. in the PDB (7EDJ).[103] An alignment with the model reveal missing interaction in the model. A short SMD was performed to correct our interface and mirror the interaction patterns displayed in the crystal structure.

Convergence issue

The bound RMSDs, both the ACE2 and the RBD, were poorly converged with the default parameters provided by BFEE2. Adjustment of the extended fluctuation to 0.01 instead of 0.05 was required to obtain the results presented below.

Result details

The detailed results of the diverse contributions is presented in Table 4.5 and individual components PMFs in Figure 4.10 with the associated convergence in Figure 4.11.

Table 4.5: Results for each contribution to the binding free energy of the Alpha SARS-CoV-2 variant spike RBD:ACE2 in the geometrical route.

Contribution	PMF (kcal/mol)	PMF (ns)
$\Delta G^{\text{site}}_{c(\text{RBD})}$	-4.5 ± 0.4	200
$\Delta G^{\text{site}}_{c(\text{ACE})}$	-8.5 ± 0.4	200
$\Delta G^{\text{site}}_{\Theta}$	-0.3 ± 0.0	40
$\Delta G^{\text{site}}_{\Phi}$	-0.1 ± 0.1	40
$\Delta G^{\text{site}}_{\Psi}$	-0.2 ± 0.1	40
$\Delta G^{\text{site}}_{\theta}$	-0.7 ± 0.0	40
$\Delta G^{\text{site}}_{\phi}$	-0.7 ± 0.0	50
$(1/\text{beta}) * \ln(\text{S} * \text{I} * \text{C}_0)$	-18.4 ± 0.0	175
$\Delta G^{\text{bulk}}_{c(\text{RBD})}$	6.0 ± 0.1	150
$\Delta G^{\text{bulk}}_{c(\text{ACE2})}$	8.5 ± 0.1	100
$\Delta G^{\text{bulk}}_{o}$	6.6	
ΔG_{b}	-12.3 ± 1.2 (calculation) -11.6 (experiment)[99]	1035

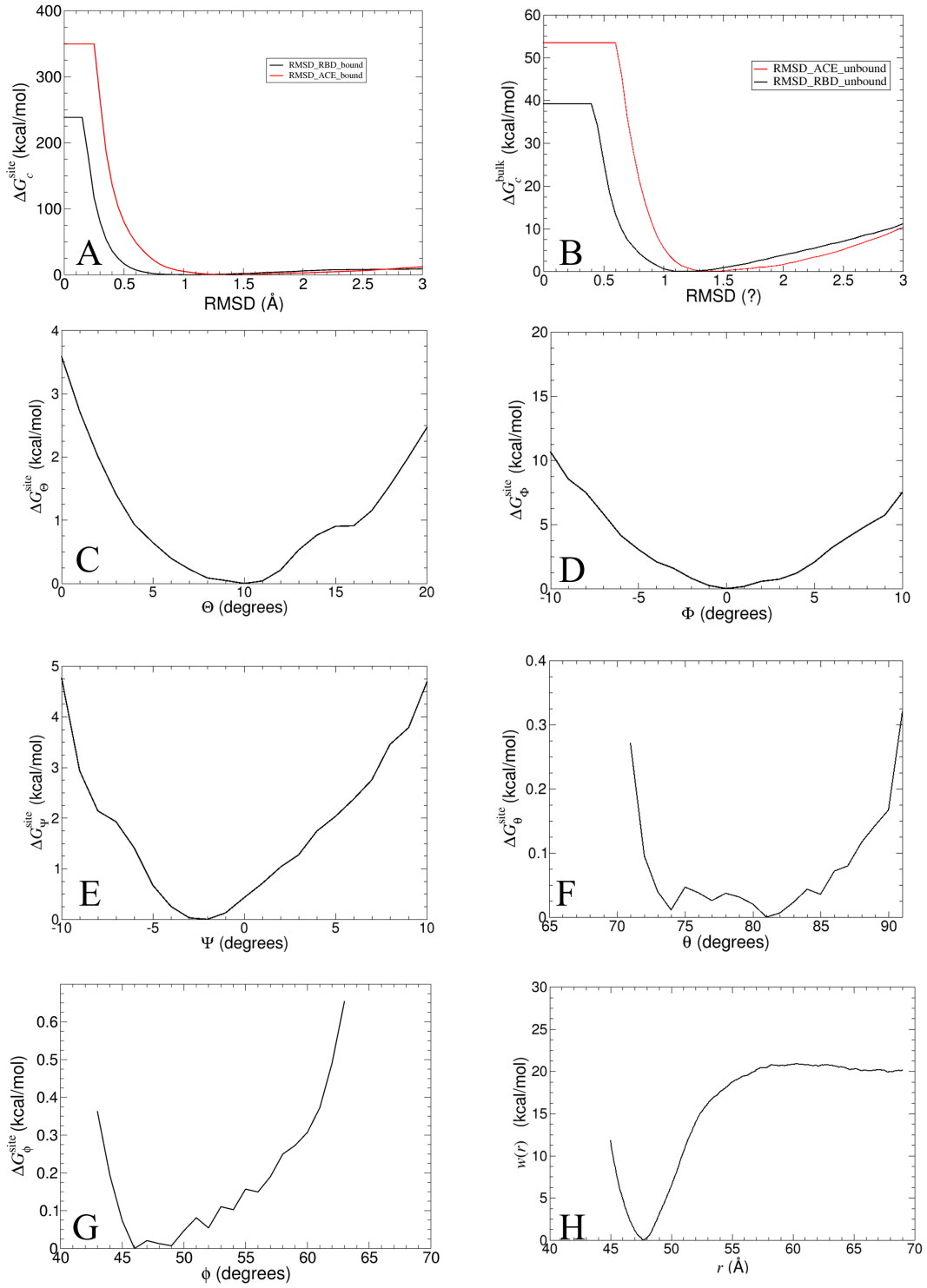

Figure 4.10: Individual PMFs for all components. The PMF calculations using RMSDs of the Alpha RBD and ACE2 proteins in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

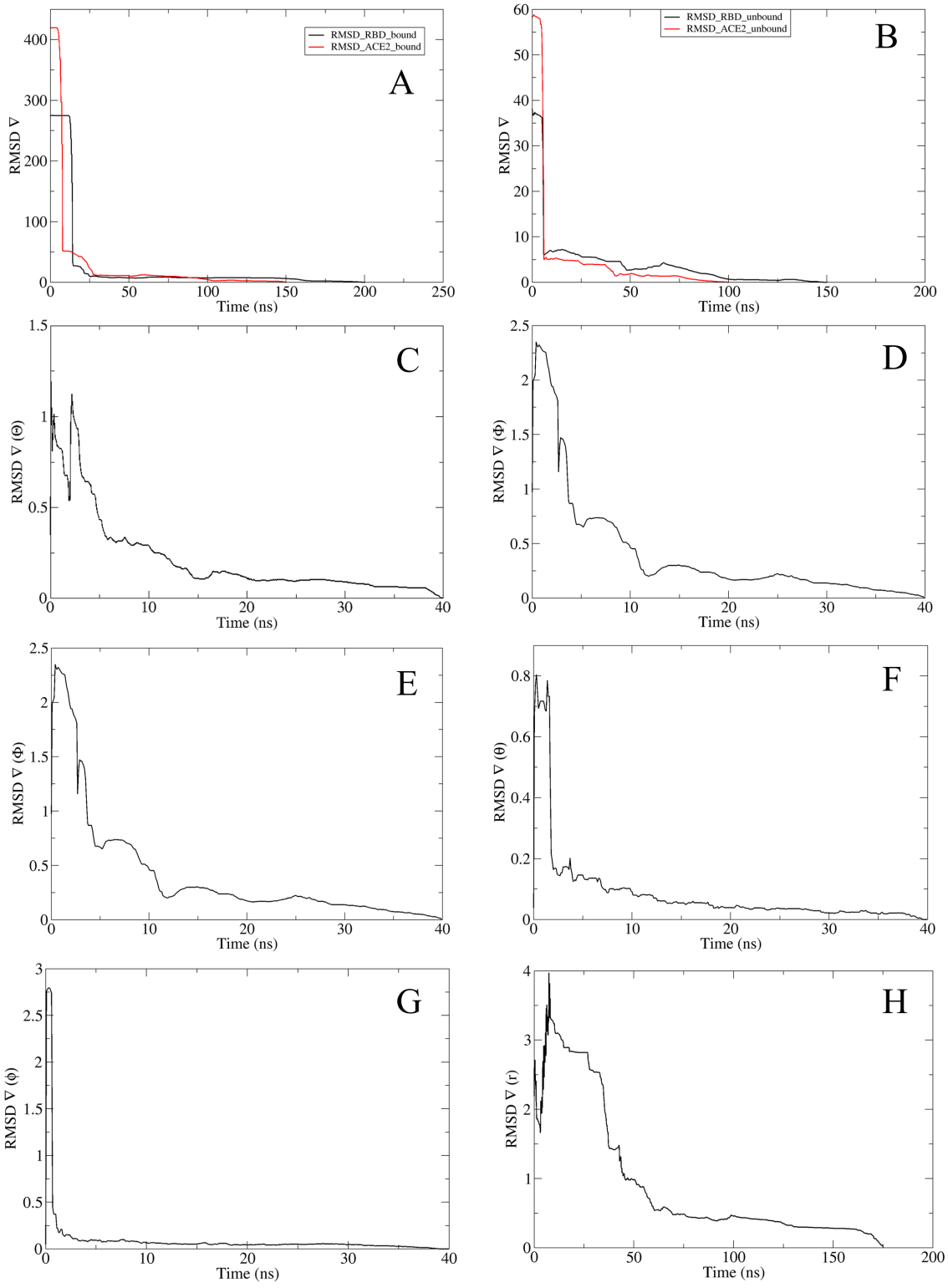

Figure 4.11: Convergence curve for individual PMFs for all components using RMSDs of the Alpha RBD and ACE2 proteins in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

4.4.5 Beta SARS-CoV-2 RBD : ACE2 (model)

Molecular assembly details

The molecular assembly was built by introducing the point mutations N501Y, E484K and K417N into the WT model. It was solvated and an ionic concentration of 0.15 M was added to mimic physiological concentration resulting in a periodic cell of dimension 129 x 128 x 127 $Å^3$ and 205931 atoms.

Convergence issue

The bound RMSDs, both the ACE2 and the RBD, were poorly converged with the default parameters provided by BFEE2. Adjustment of the extended fluctuation to 0.01 instead of 0.05 was required to obtain the results presented below.

Result details

The detailed results of the diverse contributions is presented in Table 4.6 and individual components PMFs in Figure 4.12 with the associated convergence in Figure 4.13.

Table 4.6: Results for each contribution to the binding free energy of the model Beta SARS-CoV-2 variant spike RBD:ACE2 in the geometrical route.

Contribution	PMF (kcal/mol)	PMF (ns)
$\Delta G^{\text{site}}_{c(\text{RBD})}$	-4.9 ± 0.5	150
$\Delta G^{\text{site}}_{c(\text{ACE})}$	-6.0 ± 0.5	150
$\Delta G^{\text{site}}_{\Theta}$	-0.2 ± 0.0	40
$\Delta G^{\text{site}}_{\Phi}$	-0.1 ± 0.1	60
$\Delta G^{\text{site}}_{\Psi}$	-0.2 ± 0.1	40
$\Delta G^{\text{site}}_{\theta}$	-0.6 ± 0.0	60
$\Delta G^{\text{site}}_{\phi}$	-0.7 ± 0.0	40
$(1/\text{beta}) * \ln(\text{S} * \text{I} * \text{C}_0)$	-22.9 ± 0.0	150
$\Delta G^{\text{bulk}}_{c(\text{RBD})}$	7.4 ± 0.1	150
$\Delta G^{\text{bulk}}_{c(\text{ACE2})}$	11.6 ± 0.1	200
$\Delta G^{\text{bulk}}_{o}$	6.6	
ΔG_{b}	-10.0 ± 1.2 (calculation) -11.1 (experiment)[99]	990

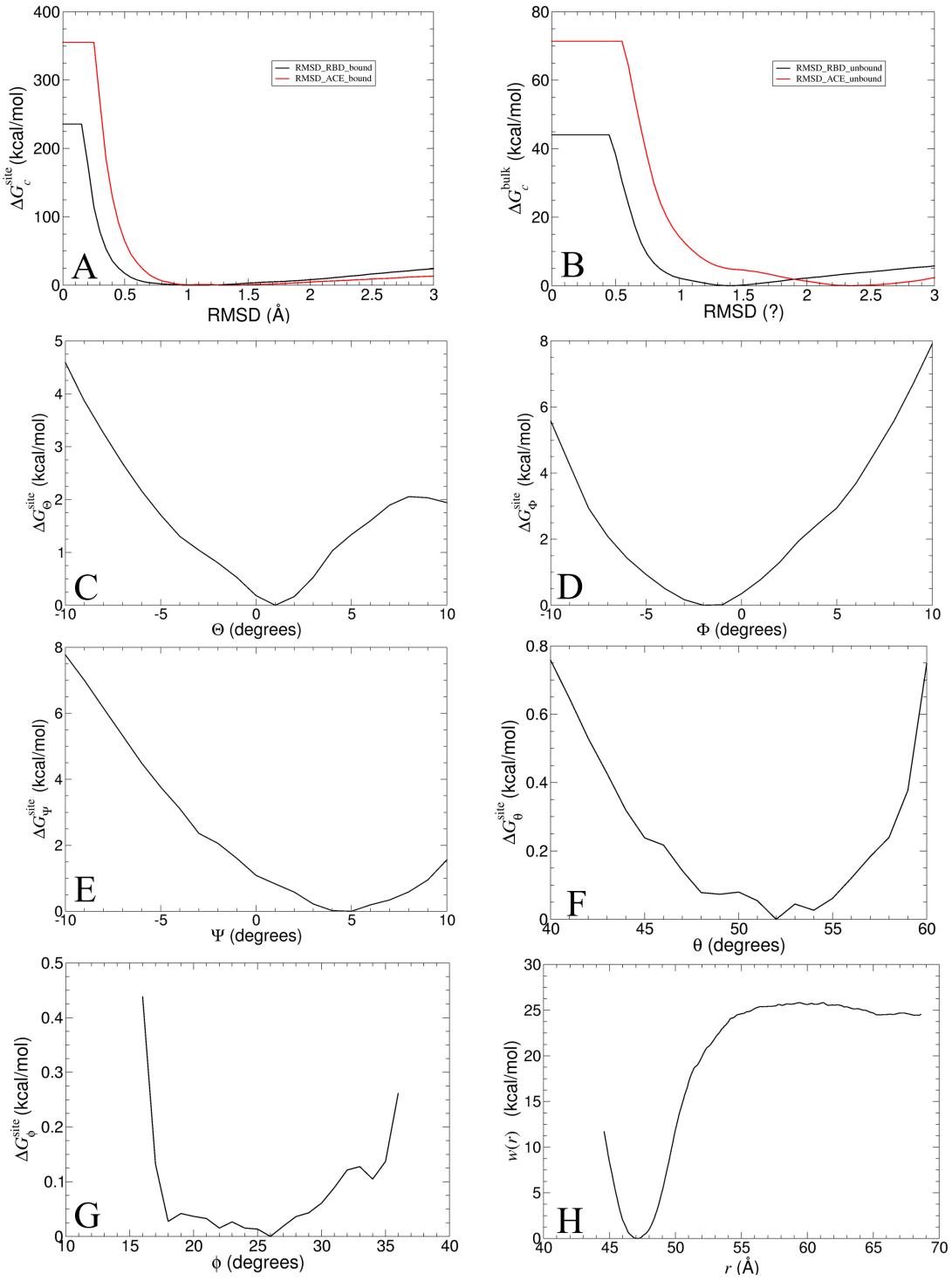

Figure 4.12: Individual PMFs for all components. The PMF calculations using RMSDs of the Beta model RBD and the ACE2 proteins in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

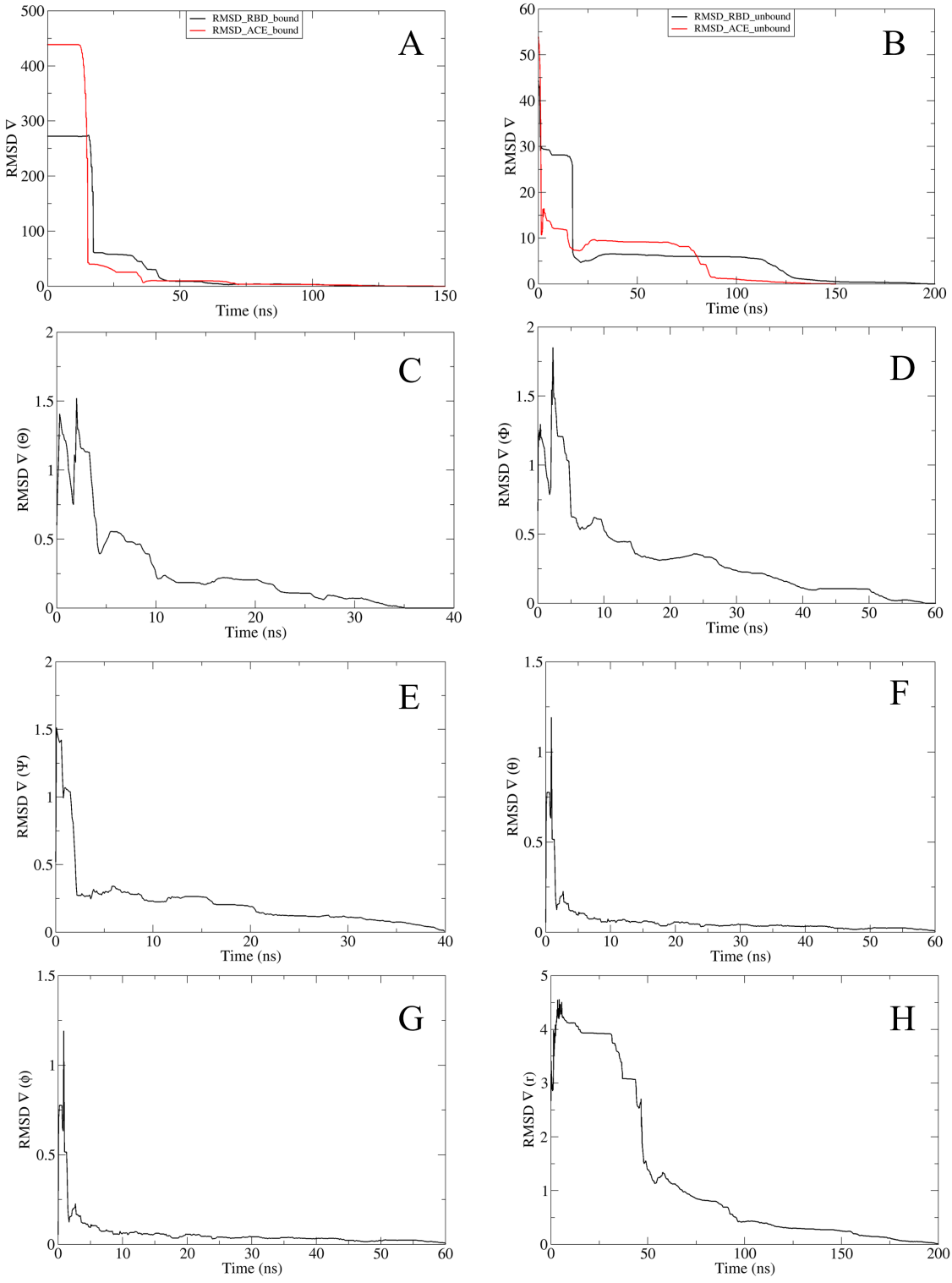

Figure 4.13: Convergence curve for individual PMFs for all components using RMSDs of the Beta model RBD and ACE2 proteins in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

4.4.6 Beta SARS-CoV-2 RBD : ACE2 (7ys6)

Molecular assembly details

The starting coordinates used to build the molecular assembly were taken from the PDB entry 7sy6 resolved at 2.8 Å.[104] The structure was then solvated in a cubic box and a ionic strength of 0.15 NaCl was added to mimic physiological conditions resulting in a total of 209982 atoms. The dimensions of the periodic cell are 126 x 115 x 148 $Å^3$.

Convergence issue

The bound RMSDs, both the ACE2 and the RBD, were poorly converged with the default parameters provided by BFEE2. Adjustment of the extended fluctuation to 0.01 instead of 0.05 and of the FullSamples to 20000 in place of 10000 was required to obtain the presented results.

Result details

The detailed results of the diverse contributions is presented in Table 4.7 and individual components PMFs in Figure 4.14 with the associated convergence in Figure 4.15.

Table 4.7: Results for each contribution to the binding free energy of the Beta SARS-CoV-2 variant spike RBD:ACE2 (7ys6) in the geometrical route.

Contribution	PMF (kcal/mol)	PMF (ns)
$\Delta G^{\text{site}}_{c(\text{RBD})}$	-9.5 ± 0.5	4x 100
$\Delta G^{\text{site}}_{c(\text{ACE})}$	-11.6 ± 1.1	400
$\Delta G^{\text{site}}_{\Theta}$	-0.3 ± 0.0	40
$\Delta G^{\text{site}}_{\Phi}$	-0.2 ± 0.0	40
$\Delta G^{\text{site}}_{\Psi}$	-0.1 ± 0.0	40
$\Delta G^{\text{site}}_{\theta}$	-0.7 ± 0.0	60
$\Delta G^{\text{site}}_{\phi}$	-0.8 ± 0.0	60
$(1/\text{beta}) * \ln(\text{S} * \text{I} * \text{C}_0)$	-12.9 ± 0.0	100
$\Delta G^{\text{bulk}}_{c(\text{RBD})}$	8.0 ± 0.0	100
$\Delta G^{\text{bulk}}_{c(\text{ACE2})}$	10.5 ± 0.0	100
$\Delta G^{\text{bulk}}_{o}$	6.6	
ΔG_{b}	-11.0 ± 1.6 (calculation) -11.1 (experiment)[99]	1040

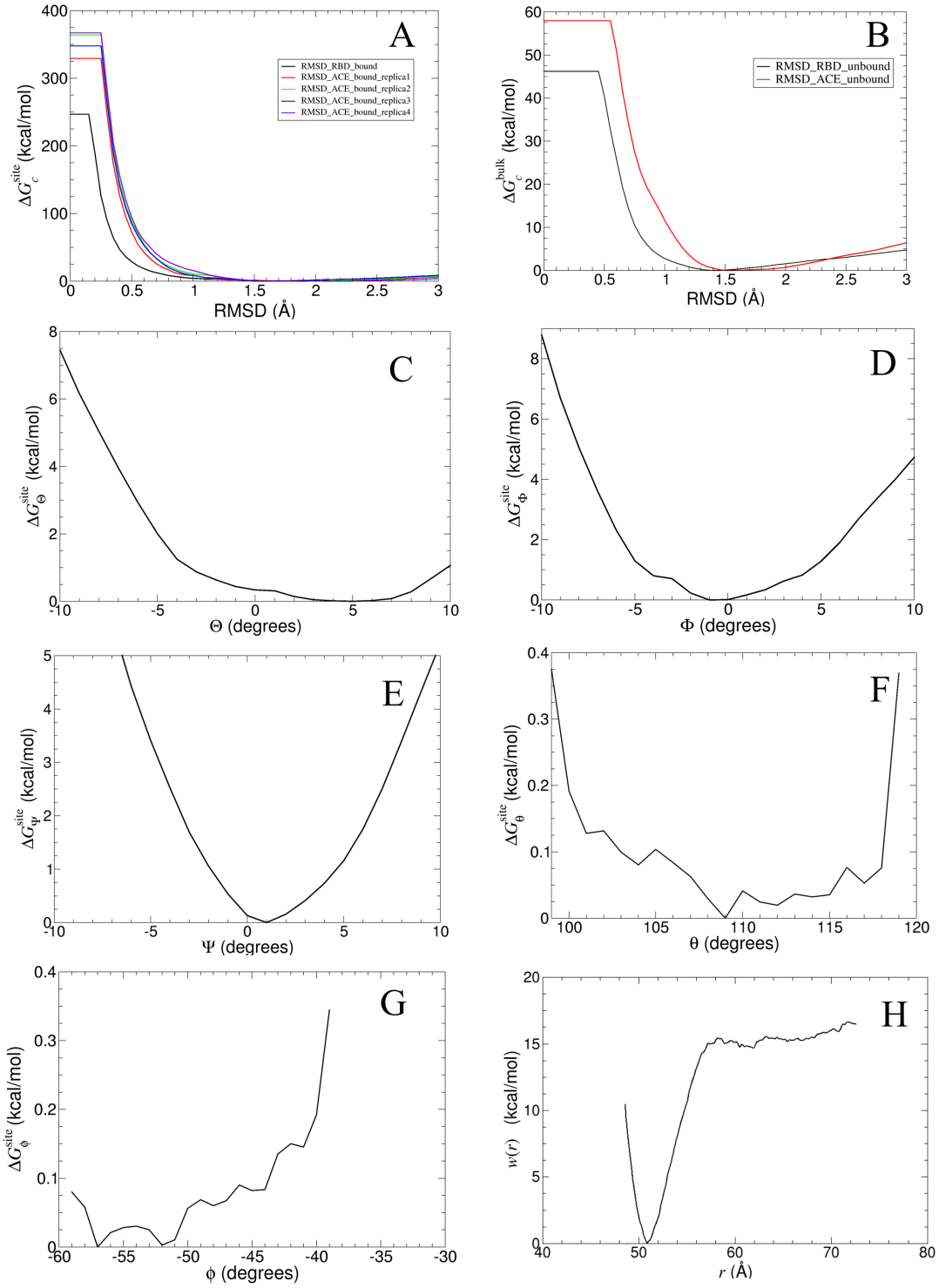

Figure 4.14: Individual PMFs for all components. The PMF calculations using RMSDs of the $\text{Beta}_{\text{Cryo-EM}}$ RBD and the ACE2 proteins in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

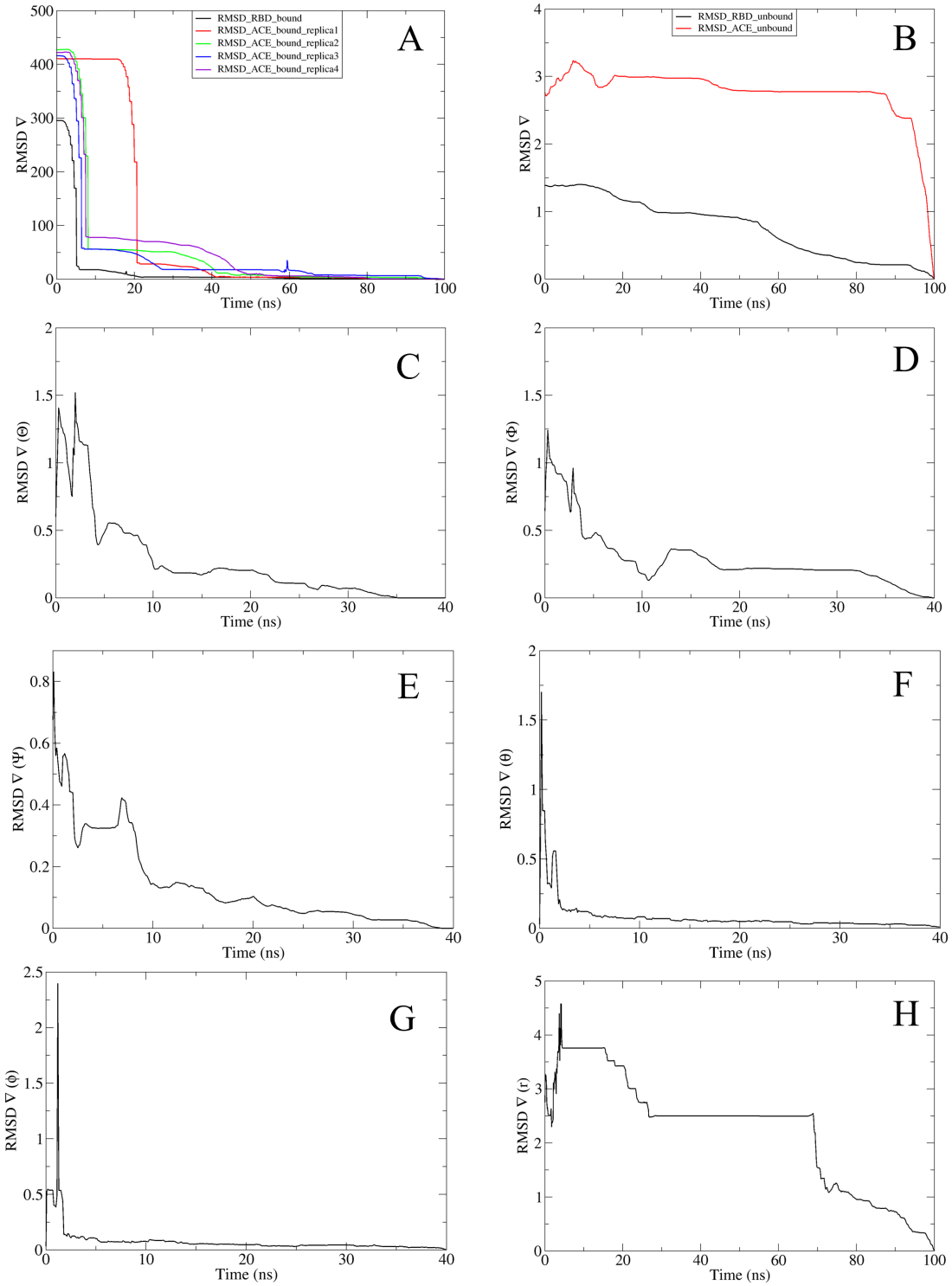

Figure 4.15: Convergence curve for individual PMFs for all components using RMSDs of the $\text{Beta}_{\text{Cryo-EM}}$ RBD and ACE2 proteins in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

4.4.7 Delta SARS-CoV-2 RBD : ACE2

Molecular assembly details

The molecular assembly was built by introducing the point mutations L452K and T487K into the WT model. It was solvated and an ionic concentration of 0.15 M was added to mimic physiological concentration resulting in a periodic cell of dimension 131 x 131 x 130 $Å^3$ and 205938 atoms.

Result details

The detailed results of the diverse contributions is presented in Table 4.8 and individual components PMFs in Figure 4.16 with the associated convergence in Figure 4.17.

Table 4.8: Results for each contribution to the binding free energy of the Delta SARS-CoV-2 variant spike RBD:ACE2 in the geometrical route.

Contribution	PMF (kcal/mol)	PMF (ns)
$\Delta G^{\text{site}}_{c(\text{RBD})}$	-7.9 ± 0.1	250
$\Delta G^{\text{site}}_{c(\text{ACE})}$	-12.6 ± 0.1	200
$\Delta G^{\text{site}}_{\Theta}$	-0.1 ± 0.0	60
$\Delta G^{\text{site}}_{\Phi}$	-0.3 ± 0.0	40
$\Delta G^{\text{site}}_{\Psi}$	-0.3 ± 0.0	40
$\Delta G^{\text{site}}_{\theta}$	-0.6 ± 0.0	60
$\Delta G^{\text{site}}_{\phi}$	-0.7 ± 0.0	40
$(1/\text{beta}) * \ln(\text{S} * \text{I} * \text{C}_0)$	-24.7 ± 0.0	100
$\Delta G^{\text{bulk}}_{c(\text{RBD})}$	16.5 ± 0.2	100
$\Delta G^{\text{bulk}}_{c(\text{ACE2})}$	14.5 ± 0.1	150
$\Delta G^{\text{bulk}}_{o}$	6.6	
ΔG_{b}	-9.6 ± 0.5 (calculation) -9.9 (experiment)[97]	1040

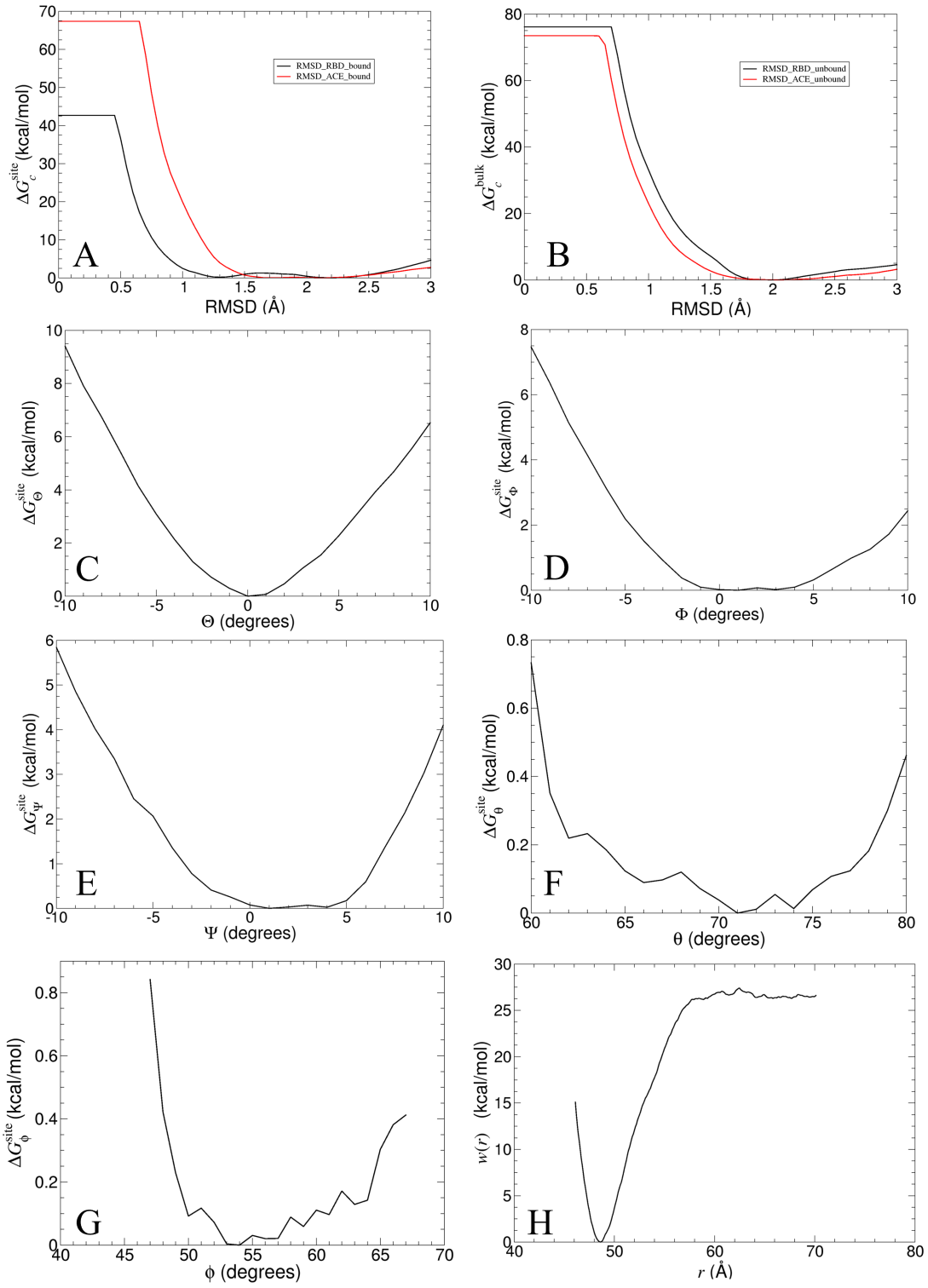

Figure 4.16: Individual PMFs for all components. The PMF calculations using RMSDs of the Delta RBD and the ACE2 proteins in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

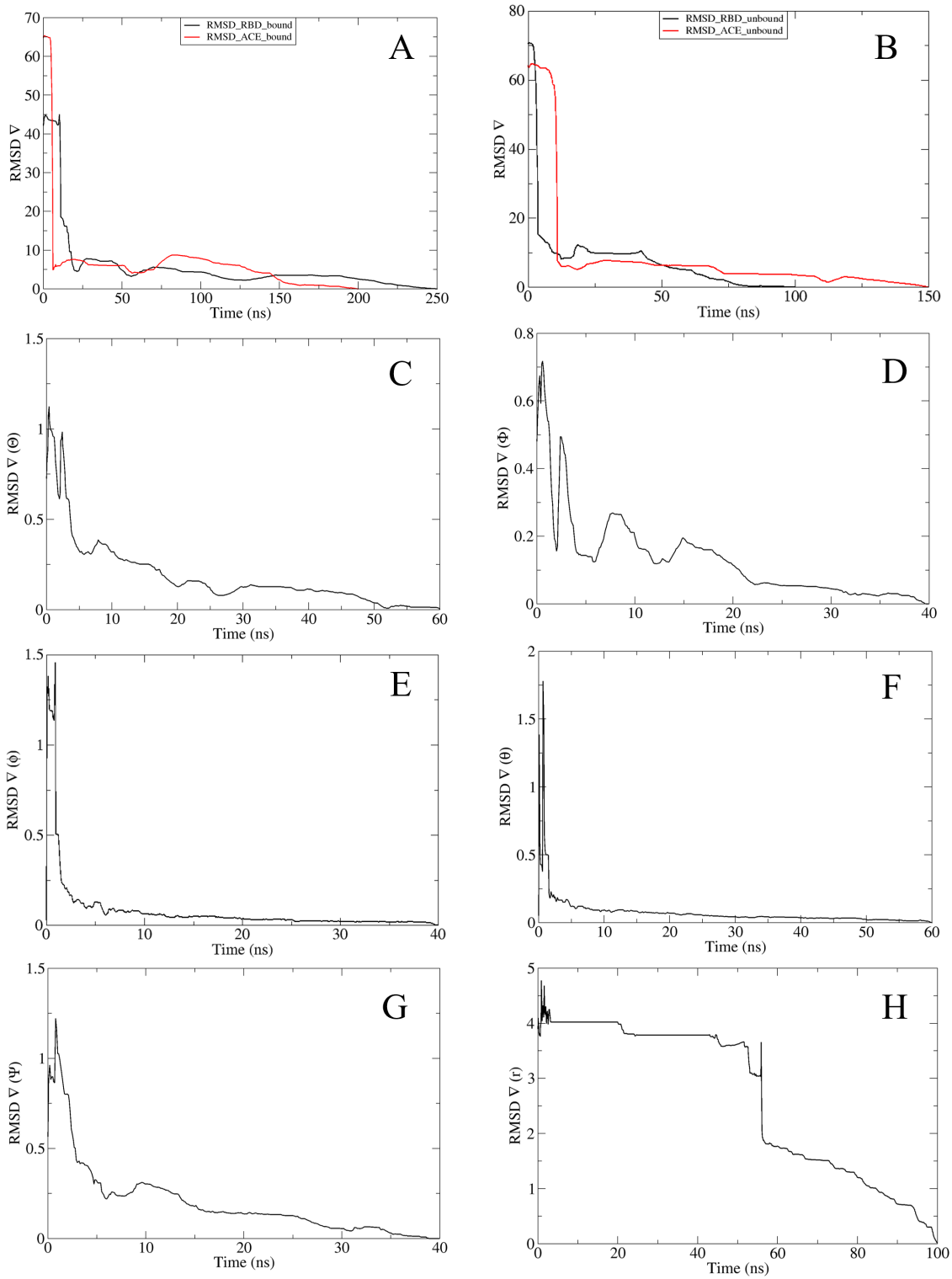

Figure 4.17: Convergence curve for individual PMFs for all components using RMSDs of the Delta RBD and ACE2 proteins in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

4.4.8 Omicron BA2 SARS-CoV-2 RBD : ACE2

Molecular assembly details

The starting coordinates used to build the molecular assembly were taken from the PDB entry (7ZF7[100]). 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 167397 atoms.

Result details

The detailed results of the diverse contributions is presented in Table 4.9 and individual components PMFs in Figure 4.18 with the associated convergence in Figure 4.19.

Table 4.9: 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
$(1/\text{beta}) * \ln(\text{S} * \text{I} * \text{C}_0)$	-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	
ΔG_{b}	-11.4 ± 1.3 (calculation) $-$ 11.5[100]	1268

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

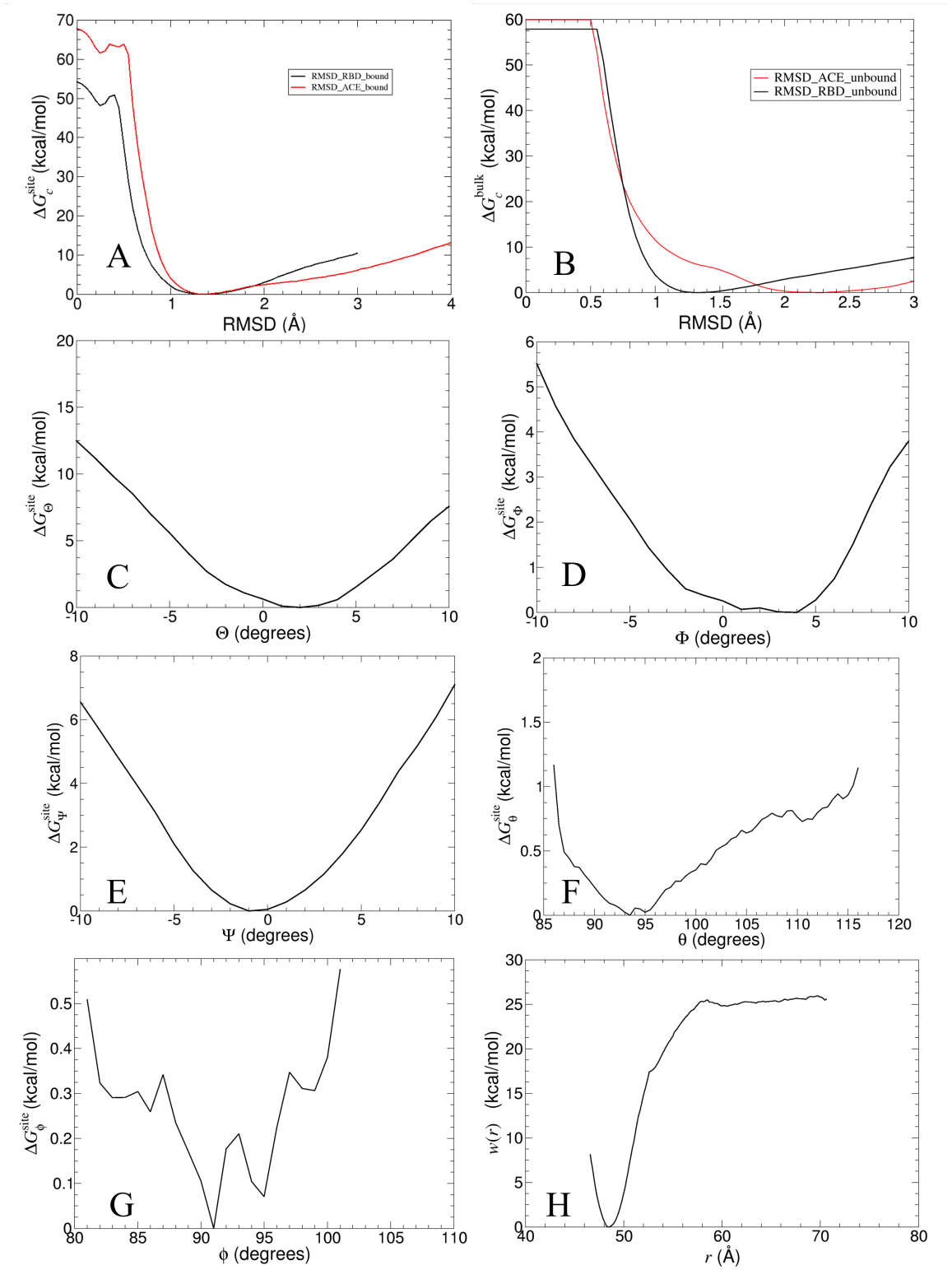

Figure 4.18: Individual PMFs for all components. The PMF calculations using RMSDs of the Omicron RBD and the ACE2 proteins in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

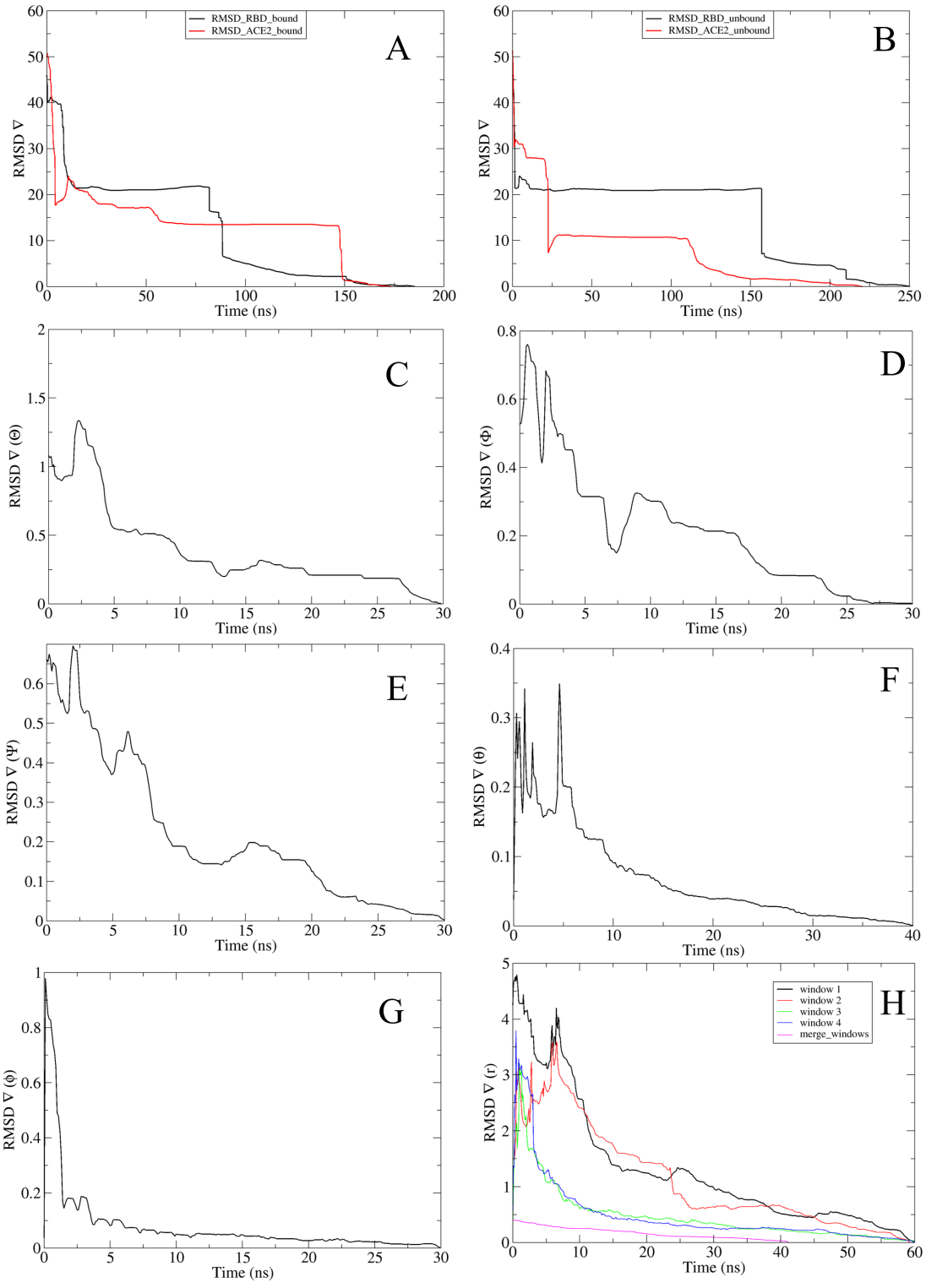

Figure 4.19: Convergence curve for individual PMFs for all components using RMSDs of the Omicron RBD and ACE2 proteins in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

4.4.9 SARS-CoV-2 RBD : ACE2 with glycans

Molecular assembly details

The molecular assembly was built using the crystal WT structure (6m17) as a template. It was solvated and an ionic concentration of 0.15 M was added to mimic physiological concentration resulting in a periodic cell of dimension 141 x 141 x 140 $Å^3$ and 265882 atoms.

Result details

The detailed results of the diverse contributions is presented in Table 4.10 and individual components PMFs in Figure 4.20 with the associated convergence in Figure 4.21.

Table 4.10: Results for each contribution to the binding free energy of the SARS-CoV-2 spike RBD:ACE2 with glycans present in the geometrical route.

Contribution	PMF (kcal/mol)	PMF (ns)
$\Delta G^{\text{site}}_{c(\text{RBD})}$	-10.7 ± 0.2	130
$\Delta G^{\text{site}}_{c(\text{ACE2})}$	-8.8 ± 0.1	200
$\Delta G^{\text{site}}_{\Theta}$	-0.1 ± 0.0	80
$\Delta G^{\text{site}}_{\Phi}$	-0.1 ± 0.1	40
$\Delta G^{\text{site}}_{\Psi}$	-0.1 ± 0.1	40
$\Delta G^{\text{site}}_{\theta}$	-0.4 ± 0.0	16
$\Delta G^{\text{site}}_{\phi}$	-0.4 ± 0.0	16
$(1/\text{beta}) * \ln(\text{S} * \text{I} * \text{C}_0)$	-15.6 ± 0.1	215
$\Delta G^{\text{bulk}}_{c(\text{RBD})}$	11.3 ± 0.1	120
$\Delta G^{\text{bulk}}_{c(\text{ACE2})}$	7.6 ± 0.1	200
$\Delta G^{\text{bulk}}_{o}$	6.6	
ΔG_{b}	-10.8 ± 0.3 (calculation) -11.4 (experiment)[98]	1057

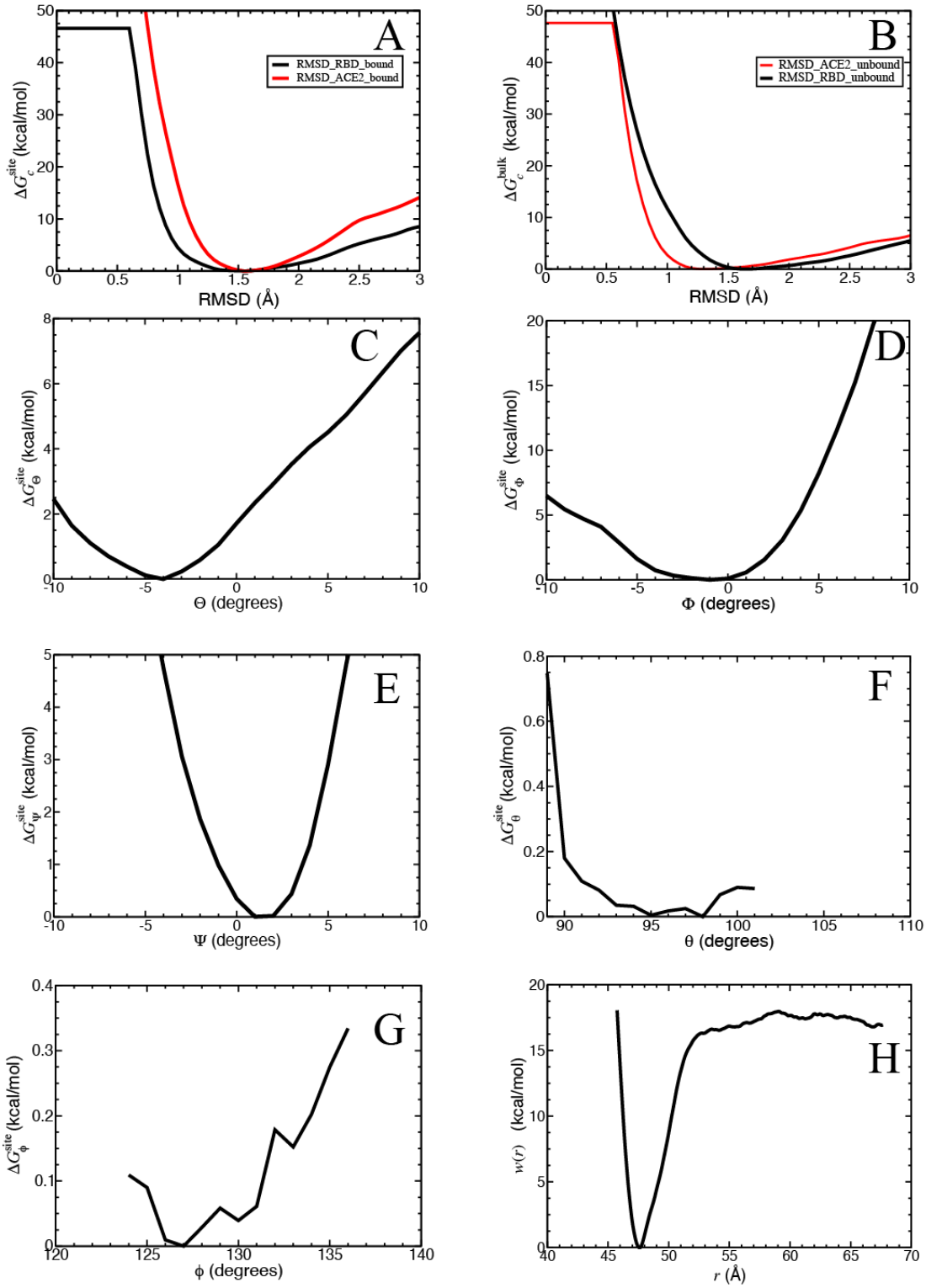

Figure 4.20: Individual PMFs for all components. The PMF calculations using RMSDs of the RBD and the ACE2 proteins with glycans present in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

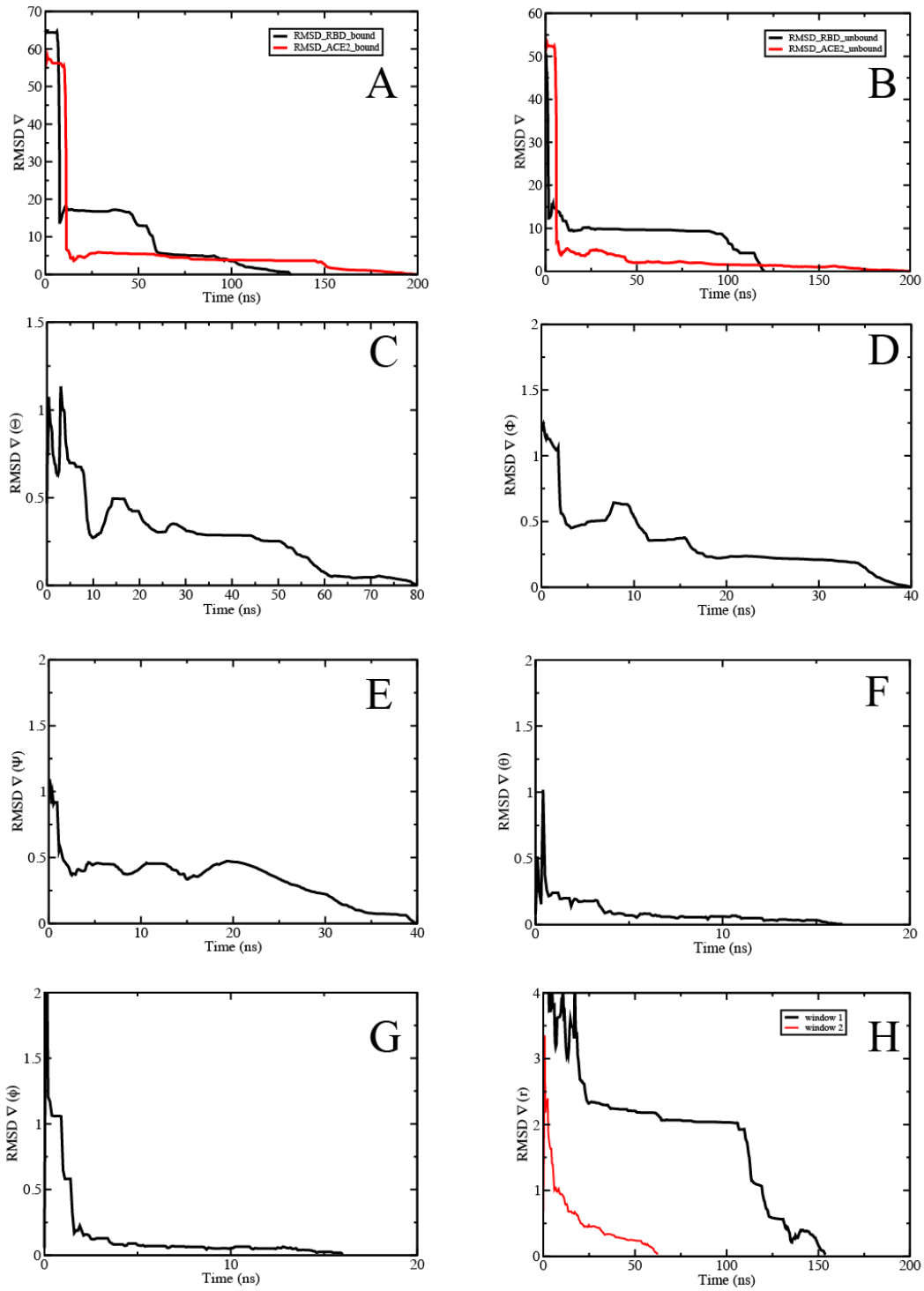

Figure 4.21: Convergence curve for individual PMFs for all components using RMSDs of the RBD and ACE2 proteins with glycans present in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

4.4.10 WT SARS-CoV-2 RBD : H11-D4

Molecular assembly details

The starting coordinates were taken from the crystallographic structure 6YZ5 resolved at 1.80 Å. Glycans in the crystal structure were retained. The assembly was then solvated in cubic 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 73977 atoms.

Result details

The detailed results of the diverse contributions is presented in Table 4.11 and individual components PMFs in Figure 4.22 with the associated convergence in Figure 4.23.

Table 4.11: 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
$\Delta G^{\text{site}}_{c(\text{H11}-\text{D4})}$	-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
$(1/\text{beta}) * \ln(\text{S} * \text{I} * \text{C}_0)$	-4.3 ± 0.0	140
$\Delta G^{\text{bulk}}_{c(\text{RBD})}$	9.3 ± 0.1	160
$\Delta G^{\text{bulk}}_{c(\text{H11}-\text{D4})}$	7.3 ± 0.2	160
$\Delta G^{\text{bulk}}_{o}$	6.6	
ΔG_{b}	-9.4 ± 0.5 (calculation) -9.9 (experiment)[102]	1290

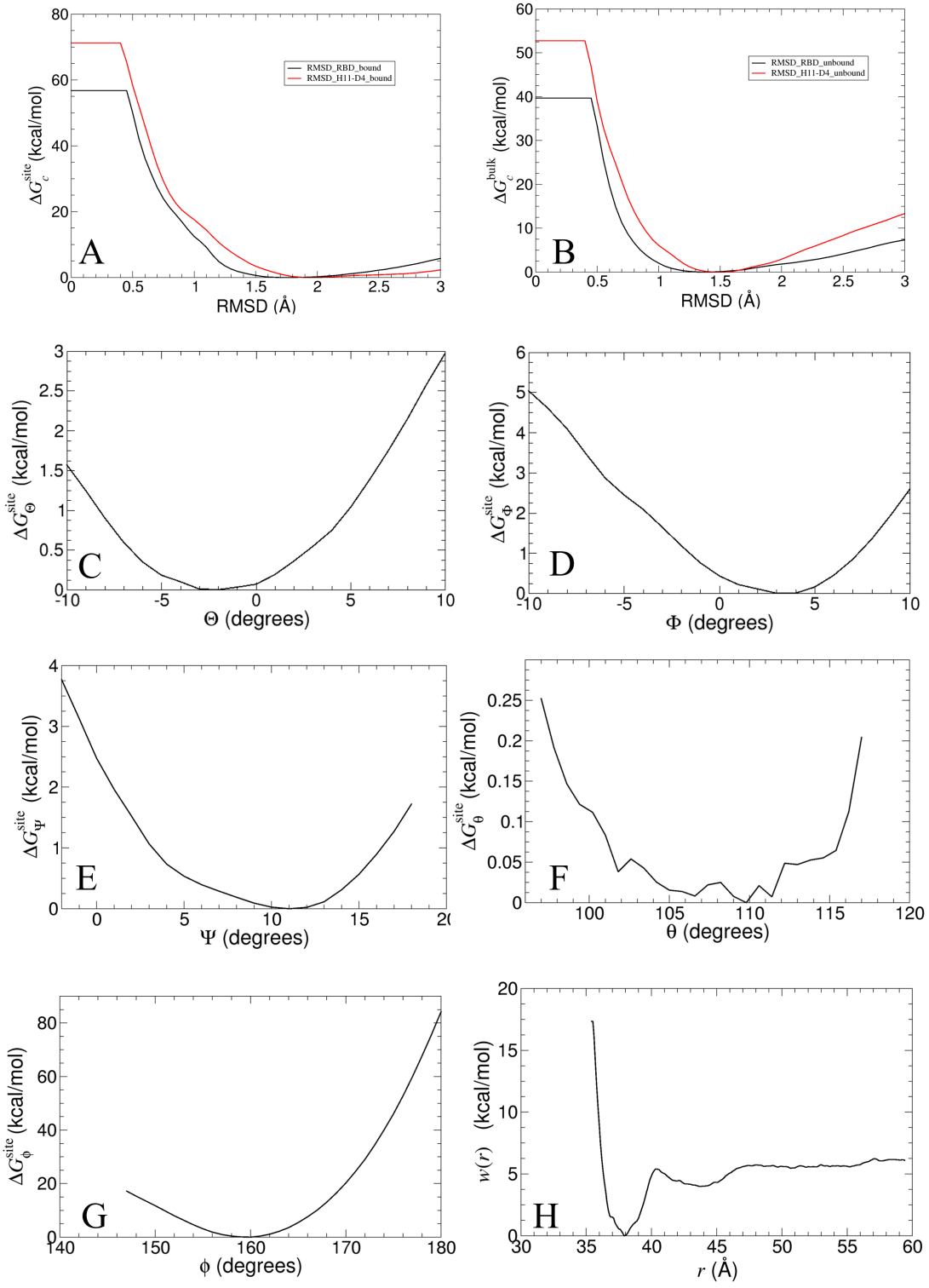

Figure 4.22: Individual PMFs for all components. The PMF calculations using RMSDs of the RBD and the H11-D4 chains in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

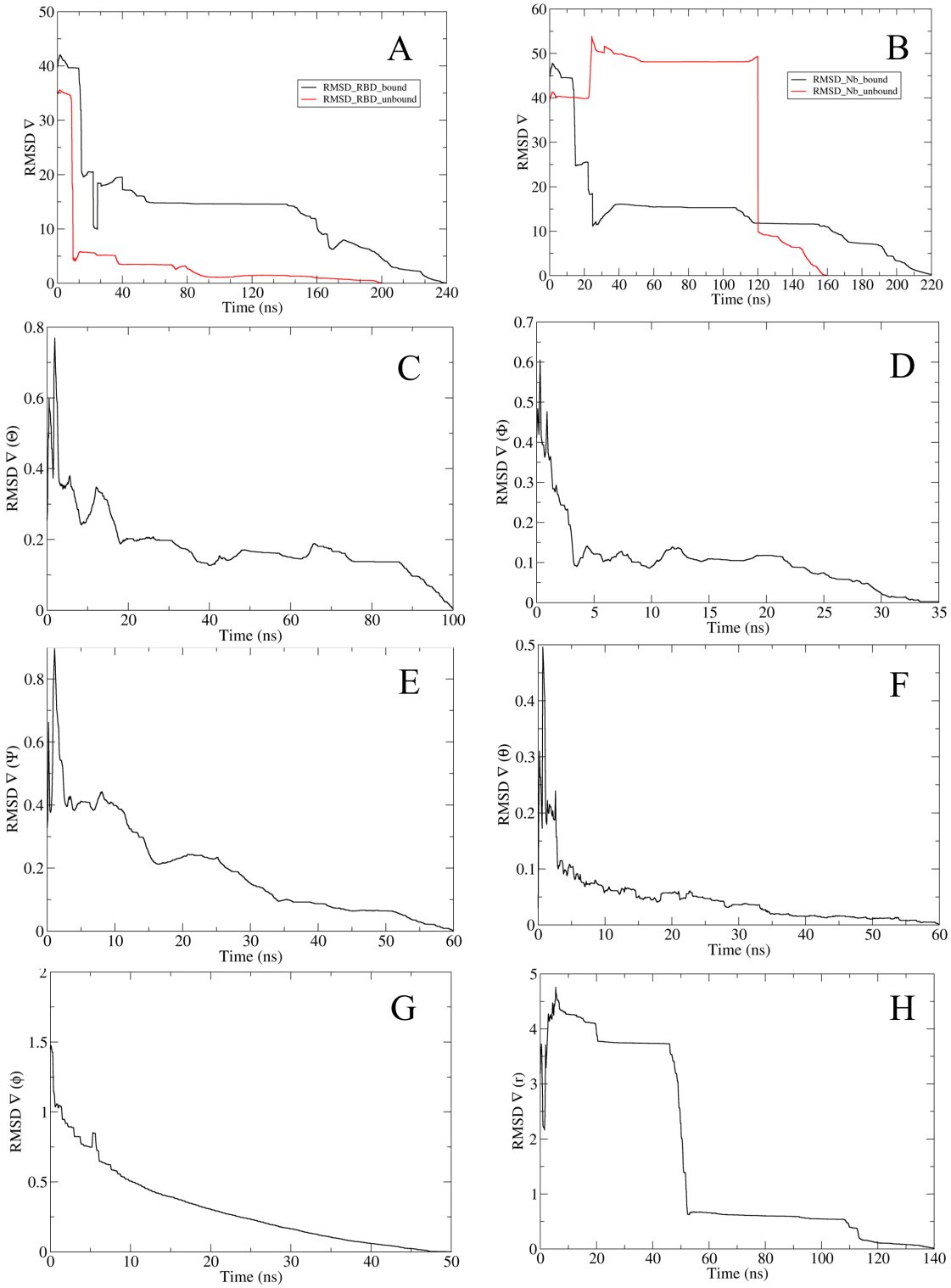

Figure 4.23: Convergence curve for individual PMFs for all components using RMSDs of the RBD and H11-D4 proteins in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

Contributions (kcal/mol)	Variants						
	WT_{model}	$WT_{crystal}$	Alpha	$Beta_{model}$	$Beta_{Cryo-EM}$	Delta	Omicron
ΔG_c^{site}	− 101.5	− 13.8	− 13	− 10.9	− 21.1	− 20.5	− 15.3
ΔG_o^{site}	− 1	− 0.5	− 0.6	− 0.5	− 0.6	− 0.7	− 0.5
ΔG_a^{site}	− 1.2	−0.6	− 1.4	− 1.3	− 1.5	− 1.3	− 1.2
ΔG_c^{bulk}	103	18.5	14.5	19	18.5	31	23
ΔG_o^{bulk}	8.3	6.6	6.6	6.6	6.6	6.6	6.6
$(1/beta) * ln(S * I * C_0)$	− 14.3	− 21.7	− 18.4	− 22.9	− 12.9	− 24.7	− 24
ΔG_b	−6.7	− 11.4	− 12.3	− 10.0	− 11.0	− 9.6	− 11.4

Table 4.13: Comparison of the decomposition of the binding free-energy between VOCs

4.4.11 Delta SARS-CoV-2 RBD : S2E12

The starting coordinates were taken from the pdb struture of the WT RBD in complex with S2E12 (7R6X) resolved at 2.95 Å. Point mutations L452R and T478K were introduced in the WT RBD to generate the δ : S2E12 model by using CharmmGUI.[153] The assembly was then solvated in cubic 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 103029 atoms.

Result details

The detailed results of the diverse contributions is presented in Table 4.12 and individual components PMFs in Figure 4.24 with the associated convergence in Figure 4.25.

Table 4.12: 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)
$\Delta G_{c(RBD)}^{site}$	-8.2 ± 0.1	100
$\Delta G_{c(S2E12)}^{site}$	-21.7 ± 0.1	100
ΔG_{Θ}^{site}	-0.4 ± 0.0	100
ΔG_{Φ}^{site}	-0.2 ± 0.0	60
ΔG_{Ψ}^{site}	-0.4 ± 0.0	120
ΔG_{θ}^{site}	-0.6 ± 0.0	130
! ΔG_{ϕ}^{site}	-0.3 ± 0.0	60
$(1/beta) * ln(S * I * C_0)$	-8.0 ± 0.0	130
$\Delta G_{c(RBD)}^{bulk}$	8.1 ± 0.1	100
$\Delta G_{c(S2E12)}^{bulk}$	12.6 ± 0.2	100
ΔG_o^{bulk}	6.6	
ΔG_b	-12.5 ± 0.3 (calculation) -12.0 (experiment)[101]	900

Comparison of the decomposition of the biding free-energy between variants

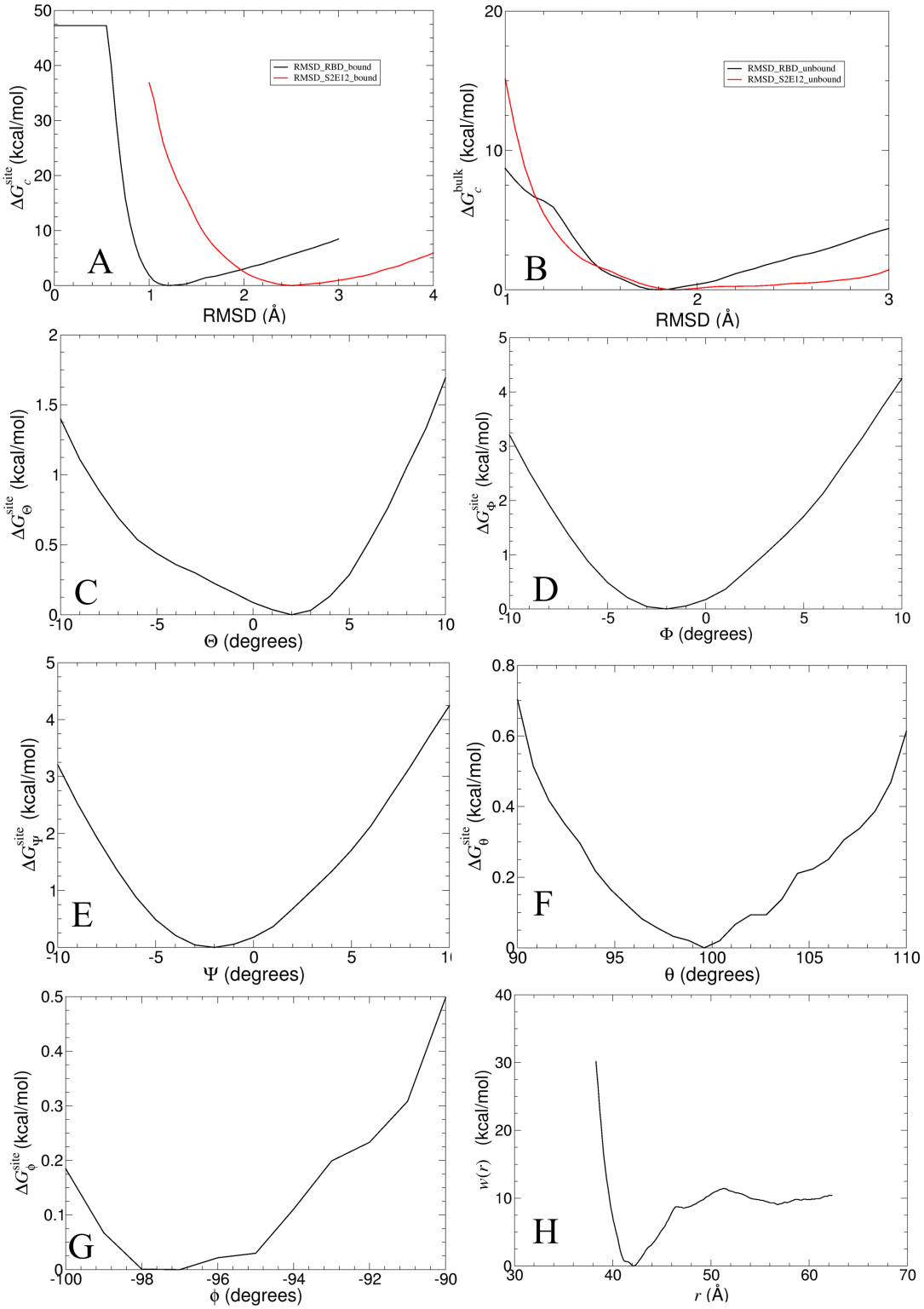

Figure 4.24: Individual PMFs for all components. The PMF calculations using RMSDs of the Delta RBD and the S2E12 chains in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

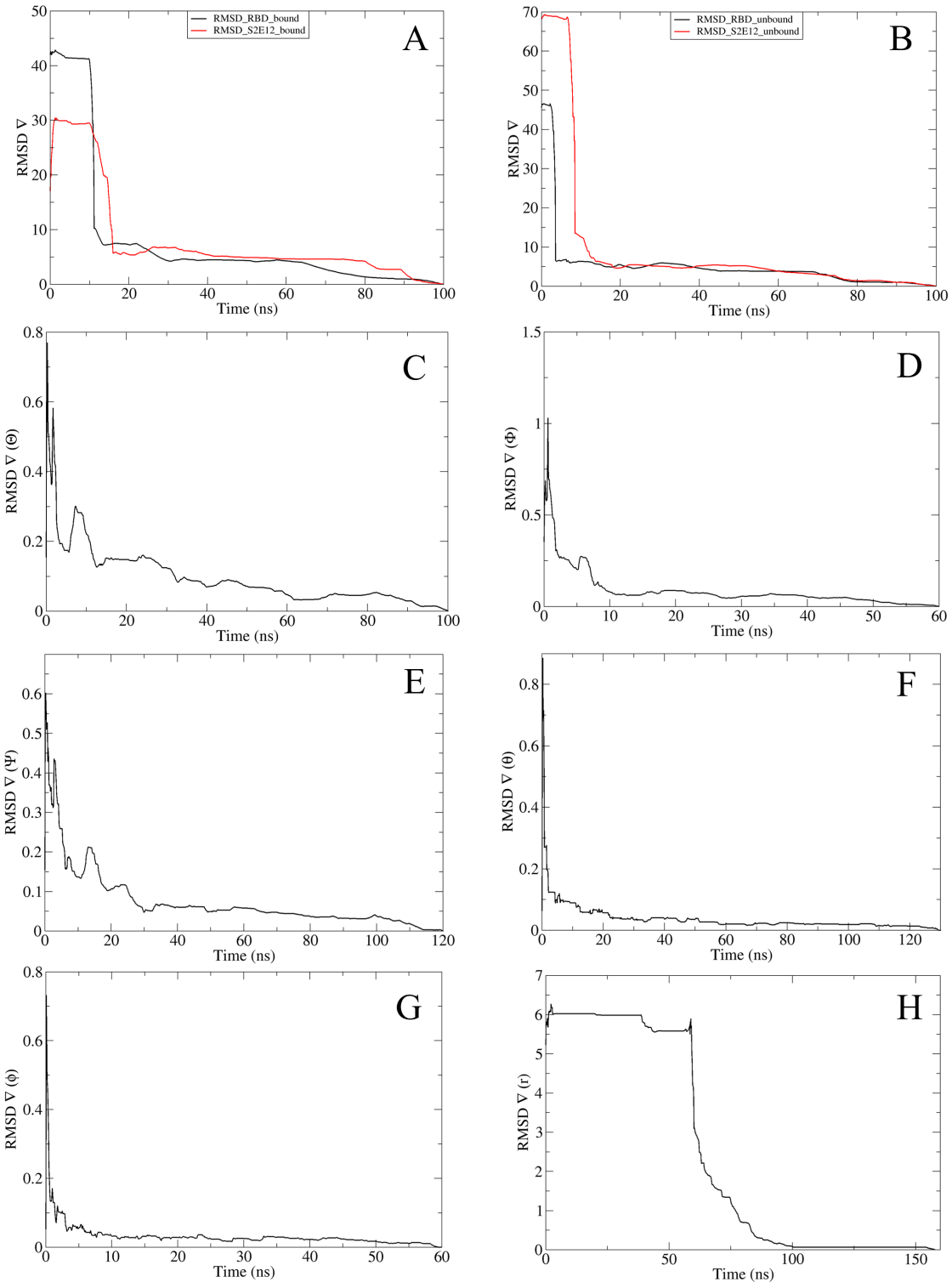

Figure 4.25: Convergence curve for individual PMFs for all components using RMSDs of the Delta RBD and S2E12 chains in the bound (A), and unbound state (B) , Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

Chapter 5

Hazardous Shortcuts in Standard Binding Free Energy Calculations

Sommaire

This chapter presents the study of the danger of performing shortcuts of the geometrical route. I will first introduce the motivations behind this work and detail my personal contribution before providing a summary of the publication. The original text is presented in the last section and can be found under:

"Blazhynska, M., **Goulard Coderc de Lacam, E.**, Chen, H., Roux, B. Chipot, C. Hazardous shortcuts in standard binding–free energy calculations, *J. Phys. Chem.*, 2022, 13, 27, 6250–6258, Doi: 10.1021/acs.jpclett.2c01490"

5.1 Motivations and personal contribution

This work was motivated by the presence in the literature of free-energy calculations using shortcuts of the geometrical route.[172,135] We wanted to enforce the need to account for the restraints both in terms of accuracy and reproducibility of the estimates. My contribution to this article was the protein-complex cheA-cheY, for which I ran both the rigorous geometrical route and its shortcut. I will go into detail about the complex and my specific results in the next part since they are not part of the article's main body.

5.1.1 Presentation and computation details

The chemotaxis histidine kinase CheA P2 domain binds to its phosphorylation target CheY forming five hydrogen bonds, seven hydrophobic contacts, and one π stacking.[173]

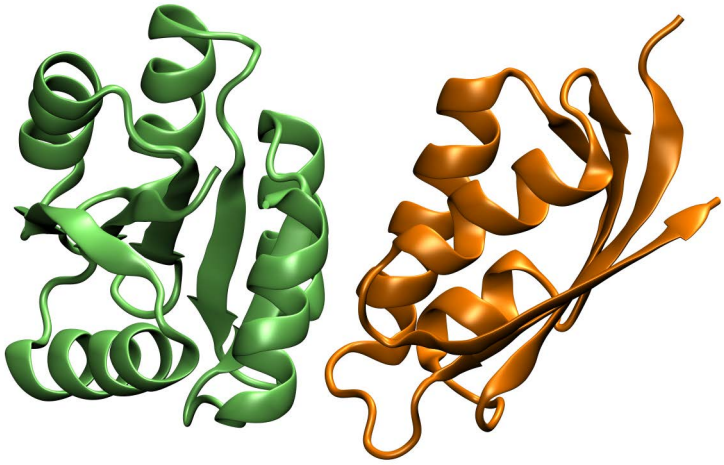

Figure 5.1: The CheA kinase-P2:CheY structure (PDB: 1U0S)[173]

The starting coordinates were taken from the X-ray diffraction structure of the CheA kinase-P2:CheY complex resolved at 1.9 Å (PDB entry 1U0S)..[173] They were initially prepared with CHARMM-GUI webserver,[153,174] and were described with the all-atom CHARMM36m force field[175] and the TIP3P water model.[176] Sodium and chloride ions were added to ensure electric neutrality and a 150 mM salt concentration, which corresponds to physiological conditions.[177] The obtained system consists of 94323 atoms in total. The dimensions of the periodic cell were $93 \times 116 \times 95$ Å^3.

Before the free-energy calculations, the complex was equilibrated for 100 ns in the NVT ensemble using NAMD 3.0.[86] The temperature and the pressure were kept at 300 K and 1 atm using overdamped Langevin dynamics and the Langevin piston, respectively.[157,156] A time step of 2 fs was used to integrate the equations of motion. Short-range interactions were smoothly turned to zero between 10 and 12 Å. The pair-list distance was 14 Å. The PME algorithm[158] was used for the long-range electrostatic interactions.

The topology files and Cartesian coordinates from the equilibrium simulations were then piped into the Binding Free-Energy Estimator 2 (BFEE2) software,[81,82] which automatically generates the input files for binding free-energy calculations with the geometrical route associated to the well-tempered meta-eABF (WTM-eABF) algorithm.[49,136]

5.1.2 Results

The following table recaps the diverse contributions to the binding free-energy obtained with the rigorous geometrical route.

Table 5.1: Results for each contribution to the binding free energy of the cheA:cheY in the geometrical route.

Contribution	PMF (kcal/mol)	PMF (ns)
$G^{\text{site}}_{c(cheA)}$	-5.0 $\pm$ 0.1	250
$G^{\text{site}}_{c(cheY)}$	-2.4 $\pm$ 0.1	100
G^{site}_{Θ}	-0.1 $\pm$ 0.0	40
G^{site}_{Φ}	-0.3 $\pm$ 0.0	60
G^{site}_{Ψ}	-0.2 $\pm$ 0.0	40
G^{site}_{θ}	-0.7 $\pm$ 0.0	50
G^{site}_{ϕ}	-0.6 $\pm$ 0.0	50
$(1/\text{beta}) * \ln(\text{S} * \text{I} * \text{C}_0)$	-14.3 $\pm$ 0.0	150
$G^{\text{bulk}}_{c(cheA)}$	3.5 $\pm$ 0.0	200
$G^{\text{bulk}}_{c(cheY)}$	4.9 $\pm$ 0.1	100
G^{bulk}_{o}	6.6	
ΔG°_{b}	-8.9 $\pm$ 0.3 (calculation) -9.1 (experiment)[173]	Total: 1040 ns

The following graphs represent the diverse PMFs' contributions to the final estimates in Figure 5.2 along with the associated convergence in Figure 5.3.

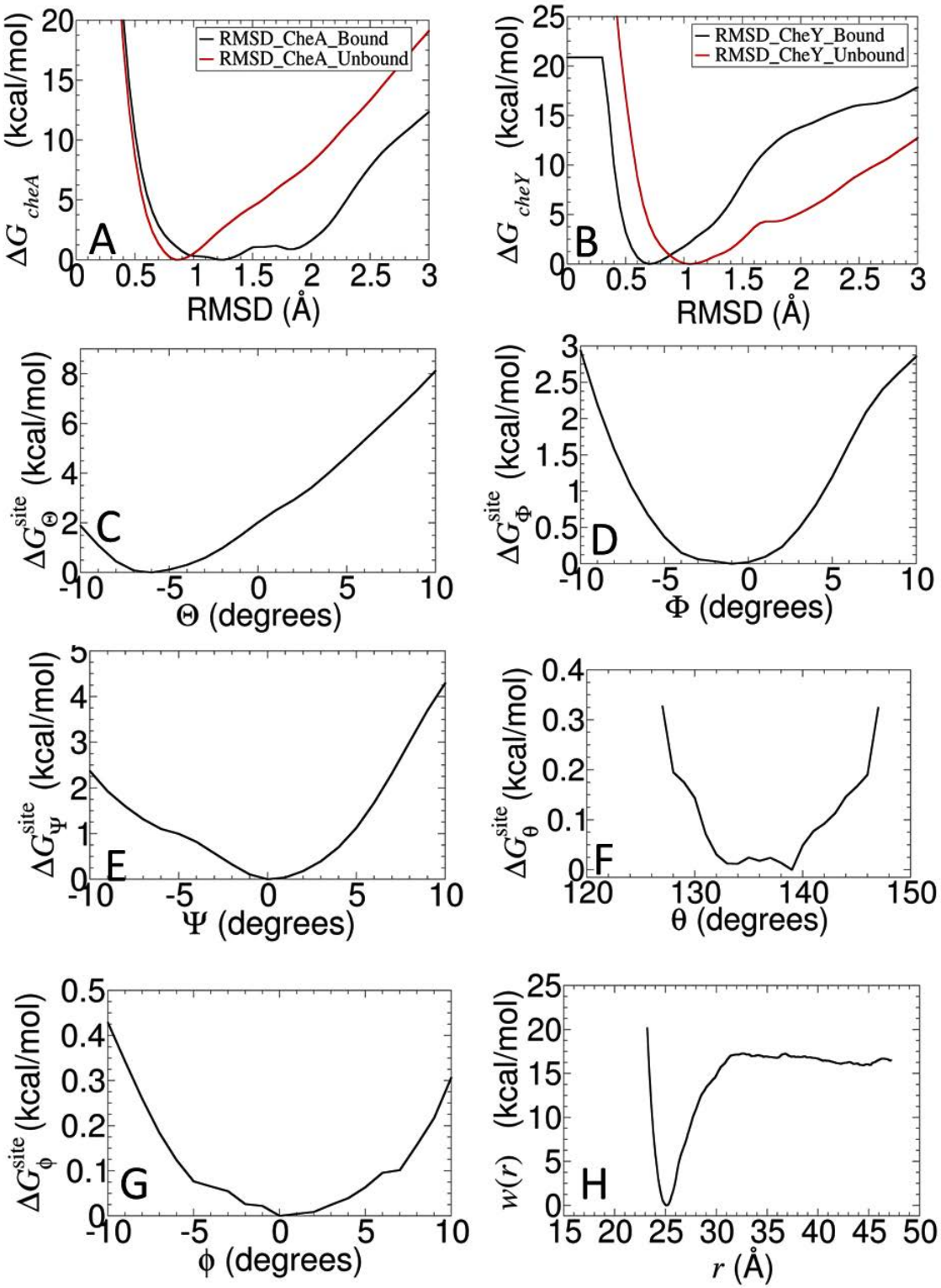

Figure 5.2: Individual PMFs for all components. The PMF calculations using RMSDs of the cheA protein in the bound and unbound state (A), the RMSDs of the cheY protein in the bound and unbound states (B), Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

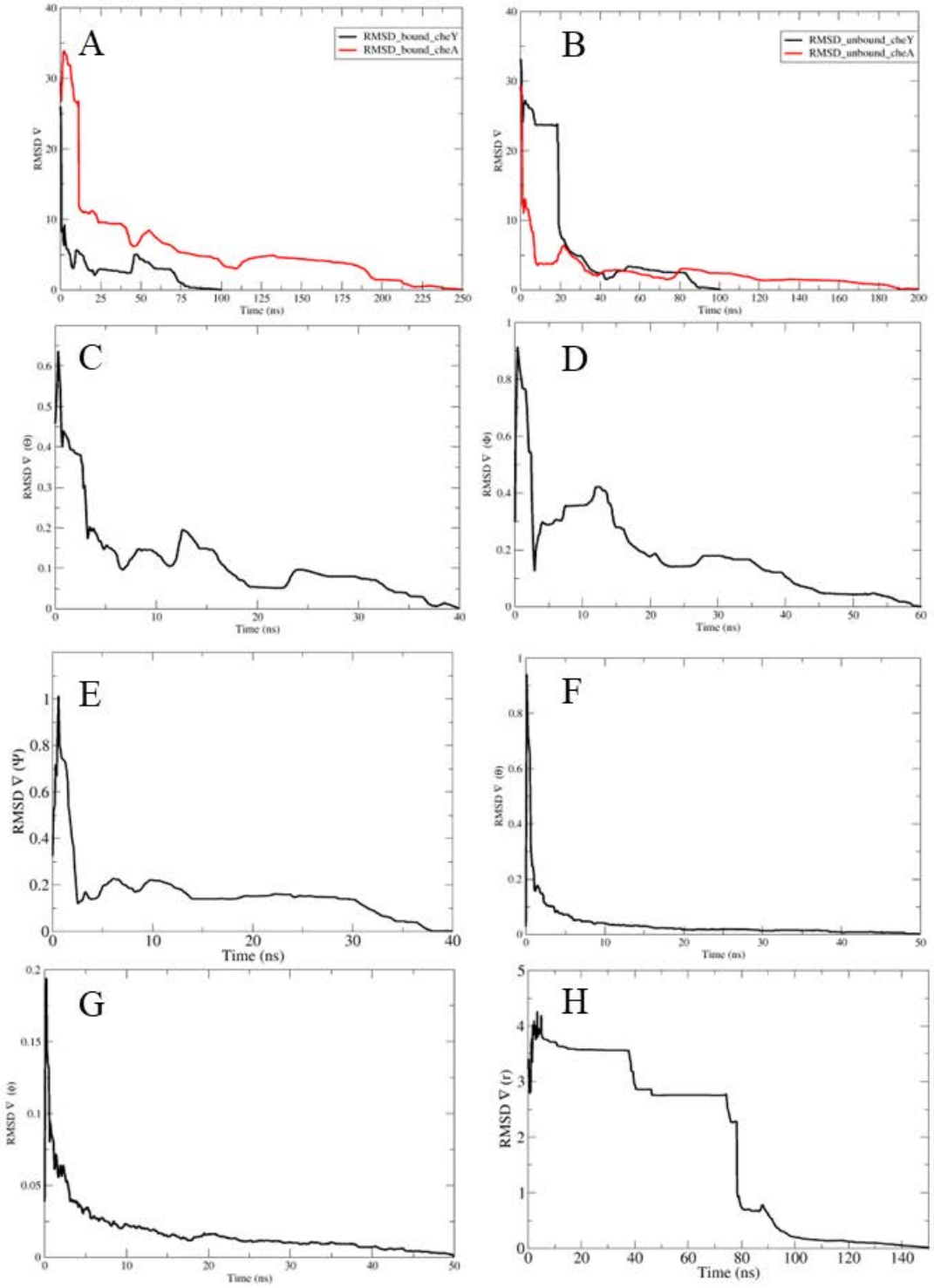

Figure 5.3: Convergence curves for individual PMFs for all components using RMSDs of the CheA kinase-P2:CheY in bound (A) and unbound (B) states, Θ (C), Φ (D), Ψ (E), θ (F), ϕ (G), and the centers-of-mass distance between two molecular entities (H), as the collective variable, respectively.

In Figure 5.4, we can observe that the separation curves of the shortcut, devoid of any restraints except from pinning down one of the proteins at the origin of the simulation box, do not showcase a clear minimum compared to the rigorous geometrical route. Their binding affinity amounted to -3.5 and -3.9 kcal/mol, respectively. They are demonstrating fluctuations throughout the 1μs simulations, showing their unreliability.

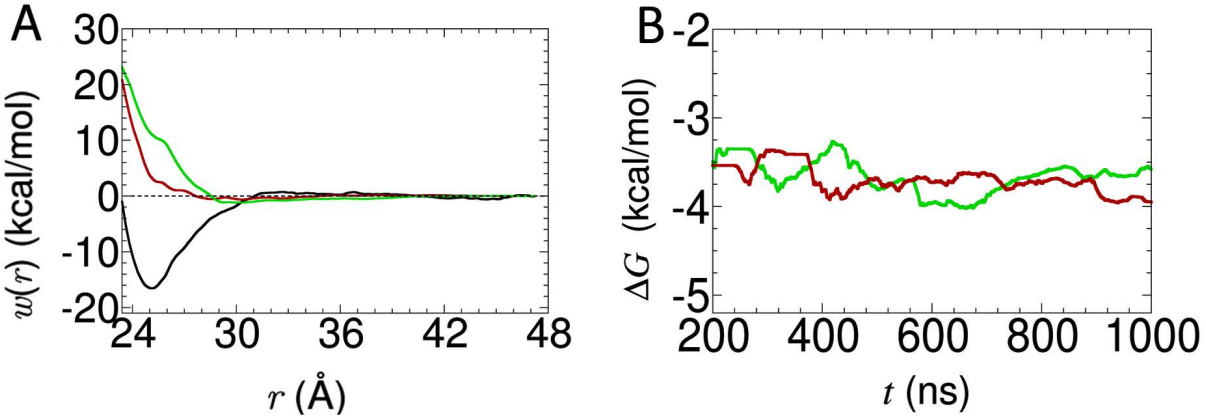

Figure 5.4: (A) Normalized separation PMFs for cheA:cheY complex calculated in the shortcut of the geometrical route for two replicas (the first replica is colored in red and the second one is in green) in comparison to the separation PMF obtained in the geometrical route (black). All the PMFs were obtained within the separation distance range of [23.2; 47.2] Å, (B) 1 μs evolution of binding free energy values of cheA:cheY complex obtained in unrestrained computations for two replicas (in red and in green, respectively).

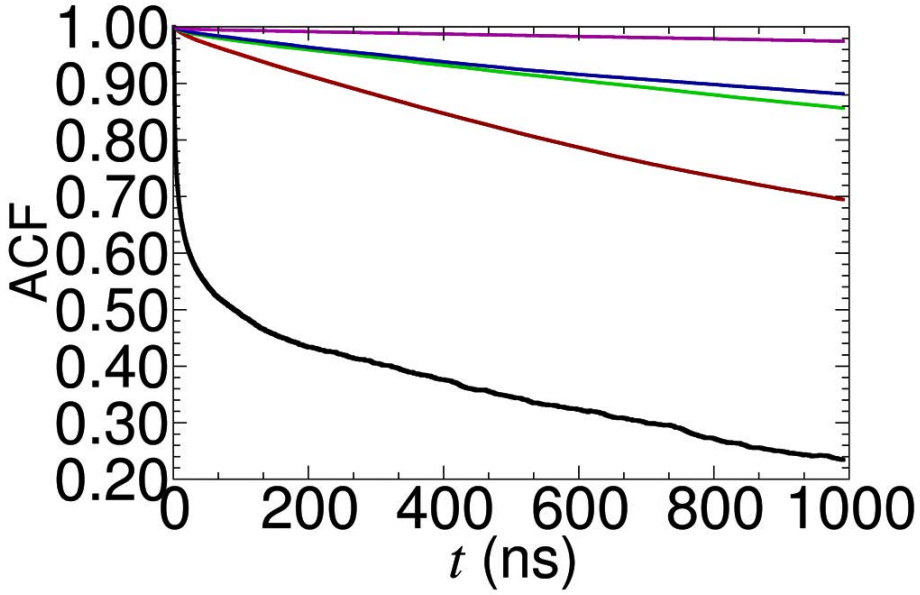

Figure 5.5: ACF of unrestrained orientational and polar angles collective variables Θ, Φ, Ψ, θ, ϕ are colored in black, red, green, blue and magenta, respectively for the red replica of Figure 5.4 A.

Examining the ACF curves computed based on the red replica of the shortcut, we observed that none of the collective variables reached 0, which is excepted when all the space available has been explored. This observation is related to the unreproducible and unreliable binding free energy estimates obtained with the shortcut of the geometrical route.

5.2 Summary

To satisfy the need for restraints in the course of the partners' separation, we carefully compared the binding free energy obtained with the full geometrical route to the ones obtained when considering only the separation in 1 μs long simulations for five systems:

- Abl kinase-SH3:p41 (4 replicas)
- MDM2-p53:NVP-CGM097 (4 replicas)
- pig insulin dimer (2 replicas)
- SARS-CoV-2 spike RBD:ACE2 (2 replicas)
- CheA kinase-P2:CheY (2 replicas)

Since the convergence of the PMF calculations following the entire geometrical route is commonly achieved in finite-length simulations by virtue of the geometrical restraints acting on the CVs, our protocol ordinarily supplies very similar estimates in replicated simulations of protein-ligand and protein-protein complexes.[178,82] Under these premises, only a single value and its total simulation time of the restrained $\Delta G^{\circ}_{\mathrm{b}}$ is reported in Table 5.2 below.

Table 5.2: 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	$\Delta G^{\circ}_{\mathrm{b}}$		
	(shortcut)	(geometrical route)	(experiment)
Abl kinase-SH3:p41[179]			
1	−5.4		
2	−4.6	−7.9± 0.2 (0.19 μs)[57,178]	−8.0[179]
3	−4.4		
4	−3.7		
MDM2-p53:NVP-CGM097[180]			
1	−5.4		
2	−5.6	−11.3 ± 0.9 (0.48 μs)[82]	−11.8[180]
3	−6.0		
4	−6.1		
Pig Insulin Dimer[181]			
1	−4.3	−7.0 ± 0.2 (0.88 μs)	−7.2[182]
2	−8.5		
SARS-CoV-2 spike RBD:ACE2[98]			
1	−4.2	−11.5 ± 0.3 (1.07 μs)	−11.4[98]
2	−7.9		
CheA kinase-P2:CheY[173]			
1	−3.5	−8.9 ± 0.3 (1.04 μs)	−9.1[173]
2	−3.9		

As can be inferred from Table 5.2, estimates obtained with the geometrical route match the experimental value within chemical accuracy, whereas shortcuts, devoid of restraints, estimates are poorly converged, inconsistent, and at variance with experimental data.

We estimated the simulation-time estimation necessary to achieve quasi-ergodic sampling along orientational and positional CVs for the simplest protein-protein study complex, pig insulin dimer. We fitted autocorrelation functions (ACFs), $C(t)$, of angular CVs, with a simple exponential function,[183] $C(t) = \mathrm{e}^{-t/\tau}$, 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 sampling along Θ, Φ, Ψ, θ, and ϕ for the simple protein-protein complex formed by the pig insulin dimer are 0.22, 0.28, 0.28, 35.09, and 13.74 μs, respectively. The maximum value of 35 μs obtained for the polar, θ, is likely to constitute 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 (shortcut). This result demonstrates that the computational cost incurred in following the rigorous theoretical framework of the geometrical route will substantially be more affordable than its shortcuts in the long run.

5.3 Original article

5.3.1 Introduction

Understanding how proteins form stable complexes with ligands—or with other proteins, is of fundamental importance in a variety of research areas, from pharmaceutical sciences to protein engineering. As a measurement of the interaction strength, the standard binding free energy quantifies the reversible association of molecular moieties, and has proven to be of great practical interest in drug design, self-assembly, or catalysis.[184,185,186,62,25,187] Over the years, a number of different molecular-dynamics (MD) based methods have been developed to determine standard binding free energies.[188,59,14,189,60,125,184] The most widely used approach is alchemical free energy perturbation (FEP), wherein the ligand is progressively decoupled from its environment by means of unphysical intermediate states.[64,62] This approach, however, is essentially limited to the treatment of relatively small ligands (e.g., drug-like) and cannot be used to treat the binding of two macromolecules.[57] To tackle this class of problems, one of the most popular avenues is the molecular mechanics/Poisson-Boltzmann and surface area (MM-PBSA) method,[14,190,191,17] where the binding free energy of a moiety is estimated as the sum of its gas-phase energy (MM), the solvation free energy (PBSA), and a contribution due to the configurational entropy of the solute extracted from MD simulations.[14,192] Unfortunately, despite its appealing inexpensive nature and success stories, the MM-PBSA approach is built upon a number of uncontrolled approximations of unknown validity, such as implicit solvation and ignorance of large structural changes upon binding,[193,19] thereby calling into question its significance and general applicability to protein-ligand and protein-protein complexes.[25,20] An alternative approach is furnished by steered MD (SMD),[131,194] whereby the binding partners are separated along an arbitrary direction via non-equilibrium pulling simulations.[195,196,197] The simple 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. It is in principle possible to recover rigorously the binding free energy through the application of the Jarzynski identity.[21] In practice, however, accurate determination of the target quantity would require multiple realizations in a near-equilibrium regime.[22] Convergence of the method is affected by the magnitude of the applied pulling force and the fact that it may not always correspond to the most physically favorable separation pathway[24] needed to guarantee a faithful reproduction of the binding affinity. Ultimately, a non-equilibrium SMD strategy necessitates substantial computational resources to obtain well-converged estimates of the binding free energy.[23] Simple unbiased brute-force MD, perhaps, provide the most straightforward approach to determine the binding free energy. Here, the starting point is the separated partners of the complex, expected to reversibly unbind in the course of a finite-length simulation. Following this strategy, Pan et al.[198] were able to characterize several protein complexes by turning to replica-exchange simulation techniques. Leaning on a brute-force approach, Buch et al.[199] exploited multiple short unbiased trajectories, to statistically reconstruct the binding process of the enzyme-inhibitor complex trypsin-benzamidine by means of a Markov-state model.[200,201,202,203] This computationally intensive approach has, however, proven to be only amenable to relatively weak (millimolar) binders, thus limiting its appeal.[187] Addressing the conceptual and computational challenges posed by strategies for the accurate estimation of binding free energies that rest on MD simulations requires the problem at hand to be considered from a different perspective. At the core of the issues at hand is the accurate reproduction of the change in configurational entropy arising from large conformational, translational, and orientational movements that accompany the reversible association of the two partners, underscoring the need for converged configurational ensemble averages.[25,204] However, due to the size, complexity, and large solvation free energy of the partners, MD-based approaches have proven vulnerable to convergence issues.[57] To ensure converged configurational ensemble averages, Woo and Roux proposed to introduce geometrical restraints acting on a set of collective variables (CVs),[44] assumed to represent the slow degrees of freedom associated with the relative

movements of the two partners, thereby reducing the configurational space to be sampled, and resulting in accelerated convergence of the free-energy calculation. Directly inspired by this restraint-based strategy are the funnel metadynamics[27] and attach-pull-release (APR) methods,[26] as well as various restraints-based approaches introduced in alchemical free-energy methodologies.[89,58,205] From a fundamental perspective, narrowing the accessible configurational space with restraining potentials during a physical transformation associated with a free energy difference is a standard variance-reduction strategy, with the understanding that the biasing effect of such restraints must be evaluated precisely at the two end-states to yield a final unbiased estimate of the binding affinity.[44,25] In this sense, the introduction of restraining potentials is sometimes referred to as "chaperoned" free-energy calculations.[206]

The original strategy put forth by Woo and Roux was generalized in the so-called "geometrical route", where the selected CVs are the Euclidean distance between the centers of mass of the two objects at play, as well as the Euler angles, the polar and azimuthal angles describing, respectively, the relative orientation and position of these molecular objects.[25,57,178] Under these premises, the binding affinity is evaluated by means of a sequential series of geometrical transformations and potential-of-mean-force (PMF) calculations, culminating with the physical separation of the two partners.[44,57,25] The geometrical restraints applied on the CVs prevent the two molecular objects from randomly tumbling while controlling their separation. While calculating a radial separation PMF without additional geometrical restraints can yield valid results as long as the sampling is sufficient,[207,208,209,210] in practice the presence of such restraints helps accelerate the convergence of the separation PMF calculation by reducing the conformational space that needs to be sampled. It is incorrect, however, to calculate a separation PMF in the presence of geometrical restraints without accounting for their biasing effects. Still, a number of attempts have been made to determine standard binding free energies following elements of the geometrical route with various computational shortcuts,[211,213,214,135,172,215,216,217,218,212] not all of which were formally justified. It is the aim of the present letter to demonstrate the necessity to follow the complete geometrical route involving (i) the PMF calculations along all CVs measuring the conformational changes as well as the relative orientation and position of the two partners, and (ii) all the relevant geometrical restraints in the determination of the PMF underlying their physical separation, to guarantee both accurate and reproducible standard binding free-energy estimates for a variety of protein-ligand and protein-protein complexes.

5.3.2 Theoretical background and methodology

The theoretical underpinnings of the methodology upon which the protein-ligand and protein-protein binding free-energy calculations of this contribution lean are recapped below. Further details of this methodology can be found elsewhere.[44,25,57,178] The standard equilibrium binding free energy of two molecular entities can be expressed as:

$$\Delta G_{\mathrm{b}}^{\circ} = -k_{\mathrm{B}}T \ln(K_{\mathrm{eq}}\, C^{\circ}), \tag{5.1}$$

where k_{B} is the Boltzmann constant, T is the temperature, K_{eq} is the equilibrium binding constant, and C° denotes the standard concentration of 1 M (C° = 1/1661 Å^3).[89] Assuming that all the degrees of freedom other than the physical separation are suitably averaged, the equilibrium binding constant can be written as:[56]

$$K_{\mathrm{eq}} = 4\pi \int_{\mathrm{site}} \mathrm{d}r\; r^2\; \mathrm{e}^{-\beta[w(r)-w(\mathrm{r}^*)]}, \tag{5.2}$$

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. In practice, the separation distance r^* must be chosen such that $w(r)$ reaches a plateau for $r \geq r^*$.

As mentioned earlier, the application of restraining potentials on a set of selected CVs reduces the configurational space of the complex to be sampled, which results in an accelerated convergence of the binding free-energy calculations.[44] The optimal choice of these CVs is presented schematically in Figure 5.6 and described in Table 5.3. In the geometrical route, the separation PMF, $w(r)$, is prefaced by a series of PMF calculations along the selected CVs. This sequential series of geometrical transformations could be visualized as nesting dolls, whereby the restraints are introduced one by one. Specifically, the molecular flexibility described by means of the distance root-mean-square deviation (RMSD) of one partner with respect to a reference conformation (e.g., coordinates of the crystallographic structure of the complex) can be seen as the outermost doll, while the inner dolls represent the orientational Euler-angle (Θ, Φ, and Ψ) and the positional polar-angle (θ and ϕ) movements. In the case of protein-protein complexes, the flexibility of the proteins at play may entail significant deviations from the bound-state structure. As a result, the contribution arising from the distance RMSD over backbone atoms for each protein has to be taken into account in the free-energy calculations. Moreover, in particular instances, to prevent spurious isomerization of amino-acid side chains at the interface of the complex arising from solvent exposure, additional restraints along the distance RMSDs of these side chains ought to be introduced.[57] Each energetic contribution due to the restraints acting on the CVs is estimated independently in a one-dimensional free-energy calculation.

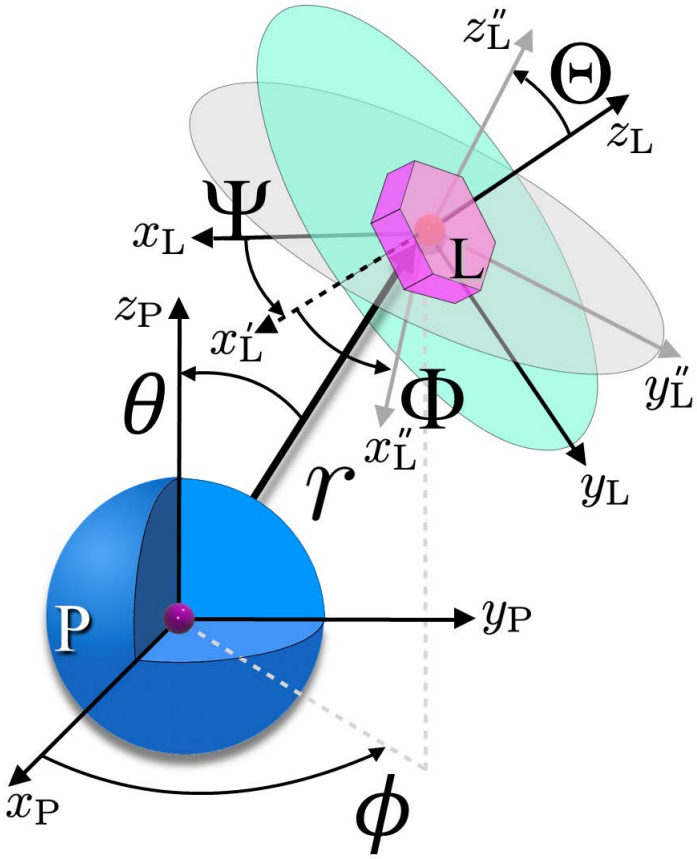

Figure 5.6: Schematic representation of the reference coordinates used to define the orientational and positional restraints, where P and L correspond to protein and ligand, respectively. The P-L center-of-mass distance is represented as r. θ and ϕ relate to the position of L with respect to P. The Euler angles (roll angle Θ, pitch angle Φ, and yaw angle Ψ) determine the relative orientation from the bound state.[178]

Table 5.3: Collective variables used in the standard binding free-energy calculations

Step	CVs	Partner movement	Representation[a]	Restraints
1	RMSD	Conformational	G_c^{site}, G_c^{bulk}	
2	Θ		G_Θ^{site}	RMSD
3	Φ	Orientational	G_Φ^{site}	RMSD, Θ
4	Ψ		G_Ψ^{site}	RMSD, Θ, Φ
5	θ		G_θ^{site}	RMSD, Θ, Φ, Ψ
6	ϕ	Positional	G_ϕ^{site}	RMSD, Θ, Φ, Ψ, θ
7	r		$w(r)$	RMSD, Θ, Φ, Ψ, θ, ϕ

[a]The superscripts "site" and "bulk" refer to the bound and the unbound states, respectively.

When the different free-energy contributions are determined, the reversible separation is simulated keeping all other conformational, positional, and orientational CVs at the value corresponding to the minimum of their PMFs by means of suitable harmonic potentials.

The equilibrium constant following the geometrical route can be expressed as:[44]

$$K_{\text{eq}} = S^* I^* e^{-\beta(G_c^{\text{bulk}} - G_c^{\text{site}} + G_o^{\text{bulk}} - G_o^{\text{site}} - G_a^{\text{site}})} \tag{5.3}$$

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

$$S^* = (r^*)^2 \int_0^\pi \mathrm{d}\theta \, \sin\theta \int_0^{2\pi} \mathrm{d}\phi \, \mathrm{e}^{-\beta u_a} \tag{5.4}$$

Here, u_a is the sum of harmonic restraint potentials of θ and ϕ. I^* is a one-dimensional integral over r, defined in terms of the separation PMF:

$$I^* = \int_{\text{site}} \mathrm{d}r \, \mathrm{e}^{-\beta[w(r) - w(r^*)]} \tag{5.5}$$

the range of integration of which is strictly defined over the bound state only.[25,57] The free energy terms G_o^{site} and G_a^{site} are the sums of the orientational (i.e., G_Θ^{site}, G_Φ^{site}, and G_Ψ^{site}), and positional (i.e., G_θ^{site}, and G_ϕ^{site}) angle contributions in the bound state. G_o^{bulk} corresponds to the orientational movement of the unbound partner in an isotropic medium, and can be determined analytically as an angular integral. In practice, K_{eq} in eq 5.3 reaches a plateau at the separation distance r^*.

In the absence of conformational and orientational restraints, the configurational space available to the complex covers all possible orientational and conformational degrees of freedom at every stages during the physical separation of the molecular objects. In practice, sampling along all degrees of freedom other than the distance separating their centers of mass is markedly slowed down by the numerous free-energy barriers to overcome. This might require extensively long simulations, depending upon the height of the free-energy barriers to overcome in order to achieve proper convergence at each separation. To characterize the convergence of the shortcut geometrical route, we have performed only separation PMF calculations, monitoring all the other CVs described in Table 5.3, yet applying no harmonic potential onto them.

For comparison purposes, the binding free energies of the five complexes listed in Table 5.4 were calculated following both the geometrical route and its shortcut, whereby 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 computational assays were initially prepared with CHARMM-GUI webserver,[153,174] and were described with the all-atom CHARMM36m force field[175] and the TIP3P water model.[176] Sodium and chloride ions were added to ensure electric neutrality and a 150 mM salt concentration, which corresponds to physiological conditions.[177] In the case of the protein-ligand complexes, the substrates were parameterized, using the CHARMM general force field (CGenFF)[48] through the CGenFF web server. The topology files and Cartesian coordinates from the equilibrium simulations were then piped into the Binding Free-Energy Estimator 2 (BFEE2) software,[81,82] which automatically generates the input files for binding free-energy calculations with the geometrical route associated to the well-tempered meta-eABF (WTM-eABF) algorithm.[136,49] For a fair comparison, all the separation PMFs of the shortcut route were obtained from 1–μs long simulations split in a set of sequential 100–ns subruns. To make our point cogent, the length of these simulations was chosen purposely to exceed that of the complete geometrical route, amounting to about 600 ns for protein-ligand complexes, and to about 900 ns for protein-protein complexes. The standard binding free-energy estimates reported in this work were compared with the available theoretical and experimental measurements of references.[179,180,182,98,173] Detail of all MD simulation setups is relegated to the Supporting Information (SI).

5.3.3 Results and Discussion

The results of our simulations of the two protein-ligand and three protein-protein complexes are presented in Table 5.4. To reinforce the conclusions of this work, the binding affinities were determined from several replicas (i.e., four for the protein-ligand complexes, and two for the protein-protein complexes), yielding the uncertainties on the calculated association constant and the standard binding free energy.

Since convergence of the PMF calculations following the geometrical route is commonly achieved in finite-length simulations by virtue of the geometrical restraints acting on the CVs (Table 5.3), our protocol ordinarily supplies very similar estimates in replicated simulations of protein-ligand and protein-protein complexes.[178,82] Under these premises, only a single value and its total simulation time of the restrained $\Delta G^\circ_\mathrm{b}$ is reported in Table 5.4. As can be observed in Table 5.4, the different values of the unrestrained $\Delta G^\circ_\mathrm{b}$ do not fall within $k_\mathrm{B}T$ from the experimental measurements. In the following discussion, we primarily focus on two protein-protein complexes, namely the SARS-CoV-2 spike RBD:ACE2 and the pig insulin dimers. Detail of the complexes examined herein, i.e., the structures and computational assays, can be found in the SI.

As shown in Table 5.4, following the geometrical route allowed the binding free energy of the SARS-CoV-2 spike RBD:ACE2 complex to be determined with remarkable accuracy (i.e., the restrained $\Delta G^\circ_\mathrm{b}$ is equal to -11.5 ± 0.3 kcal/mol). By way of surface plasmon resonance experiment, Lan et al.[98] reported a binding affinity of -11.4 kcal/mol. In the present work, use was made of their crystal structure available in the Protein Data Bank (PDB) with the code 6M0J.[98] In stark contrast, simplification of the geometric route to a 1–μs separation PMF calculation, akin to the shortcut followed by Francés-Monerris et al.,[135,172] led to the unrestrained $\Delta G^\circ_\mathrm{b}$ of -4.2 and -7.9 kcal/mol, obtained from two independent replicas. Similarly discrepant binding free energies were found for the pig insulin dimer. In this case, the values of the unrestrained $\Delta G^\circ_\mathrm{b}$ for the two replicas are -4.3 and -8.5 kcal/mol. The marked difference between these estimates suggests that the shortcut of the geometrical route is unable to supply trustworthy and reproducible results. Moreover, the deviation observed for nearly all complexes between the unrestrained $\Delta G^\circ_\mathrm{b}$ estimates and their experimental counterpart is substantial, indicative that the shortcut of the geometrical route constitutes in general an unreliable strategy to predict standard binding free energies. While the second replica of the separation PMF calculation for the pig insulin dimer, yielding a binding affinity equal to -8.5 kcal/mol, is reasonably close to the experimental value, we believe this

Table 5.4: 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	$\Delta G_{\mathrm{b}}^{\circ}$		
	(shortcut)	(geometrical route)	(experiment)
Abl kinase-SH3:p41[179]			
1	−5.4		
2	−4.6	−7.9± 0.2 (0.19 μs)[25,178]	−8.0[179]
3	−4.4		
4	−3.7		
MDM2-p53:NVP-CGM097[180]			
1	−5.4		
2	−5.6	−11.3 ± 0.9 (0.48 μs)[82]	−11.8[180]
3	−6.0		
4	−6.1		
Pig Insulin Dimer[181]			
1	−4.3	−7.0 ± 0.2 (0.88 μs)	−7.2[182]
2	−8.5		
SARS-CoV-2 spike RBD:ACE2[98]			
1	−4.2	−11.5 ± 0.3 (1.07 μs)	−11.4[98]
2	−7.9		
CheA kinase-P2:CheY[173]			
1	−3.5	−8.9 ± 0.3 (1.04 μs)	−9.1[173]
2	−3.9		

result represents the fortuitous exception that confirms the rule.

The PMFs determined for the pig insulin dimer in the geometrical route and its shortcut, and the time-evolution of the unrestrained $\Delta G_{\mathrm{b}}^{\circ}$ in the course of the 1–μs simulation are shown in Figure 5.7. Both replicas of the shortcut route exhibit fluctuations in the first half of the free-energy calculation. It is noteworthy that for one of the replicas, the unrestrained $\Delta G_{\mathrm{b}}^{\circ}$ reaches -16 kcal/mol, before plateauing at a value roughly $2k_{\mathrm{B}}T$ away from the experimental value. The observed abrupt variations in the PMFs cast further doubt on the reliability of the shortcut route.

In sharp contrast with the geometrical route, in its shortcut version, which imposes sampling the entire configurational space available to the dimer, the contribution of the configurational entropy to the binding affinity at large separations, namely the Jacobian term, cannot be adequately evaluated. 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 (see Figure 5.7A).

Sampling of all possible configurations, which is imposed by the shortcut of the geometrical route, would require substantial simulation times to ensure adequate convergence of the binding free-energy

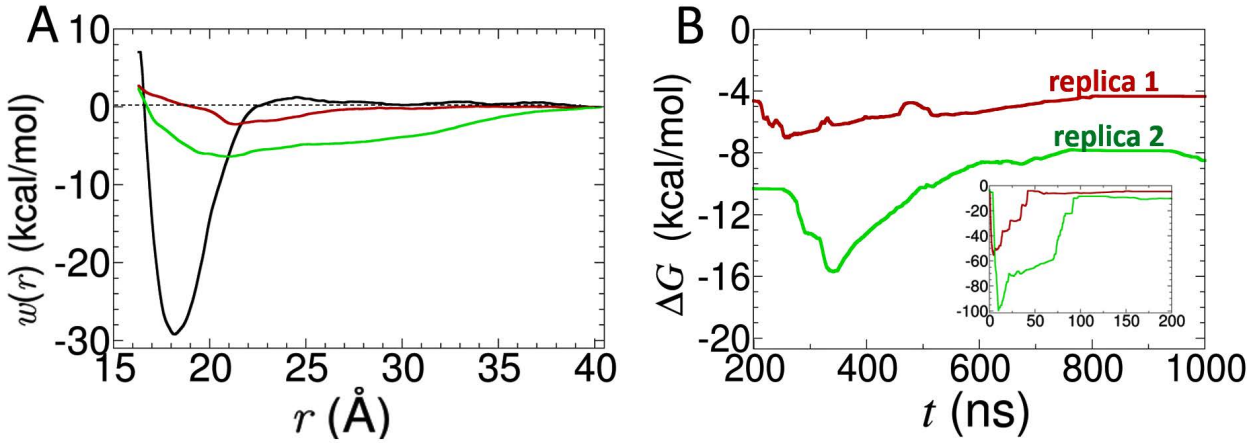

Figure 5.7: (A) Normalized separation PMFs for insulin dimer calculated in the shortcut of the geometrical route for two replicas (in red and green, respectively) in comparison to the separation PMF of the geometrical route (black). All the PMFs were obtained within the separation distance range of [16.3; 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.

calculation. In practice, such simulation times are incompatible with the contingencies of a reasonably fast and predictive method based on first principles. To demonstrate the impossibility to sample the full range of positional and orientational angles accessible to the molecular objects as they dissociate reversibly over the course of the 1 μs–timescale simulation, we determined the autocorrelation functions (ACFs) of the Θ, Φ, Ψ, θ and ϕ angles in the case of the pig insulin dimer (Figure 5.8A). These ACFs were compared to those obtained with the angular CVs restrained during the converged 100–ns separation PMF calculation following the geometrical route (Figure 5.8B).

As is apparent from Figure 5.8A, none of the ACFs decay to zero during the unrestrained separation PMF calculation, indicative that the simulation time is grossly insufficient to witness complete decorrelation of the variables at play. This observation is particularly true for the ACFs of θ and ϕ, the decay of which are much slower than that of the three Euler angles. In contrast, the angular ACFs of the separation PMF (Figure 5.8B) following the geometrical route decay rapidly in the first nanoseconds of the simulation and continuously fluctuate around zero afterwards, demonstrating the total decorrelation of the CVs in the course of the PMF calculation.

A rough evaluation of the computational time necessary to achieve complete decorrelation in the shortcut of the geometrical route was estimated by fitting the ACFs with a simple exponential function,[183] $C(t) = \mathrm{e}^{-t/\tau}$, 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 sampling along Θ, Φ, Ψ, θ, and ϕ for the simple protein-protein complex formed by the pig insulin dimer are 0.22, 0.28, 0.28, 35.09, and 13.74 μs, respectively. It is expected 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.

To examine the evolution of the angular CVs of a fully unrestrained separation PMF calculation, where the reference protein is free to tumble and reorient in the box, we performed two independent, 1–μs long simulations of the pig insulin dimer (Figure 5.9A).

The final binding free-energy estimates of both replicas are equal -3.9 and -4.6 kcal/mol, high-

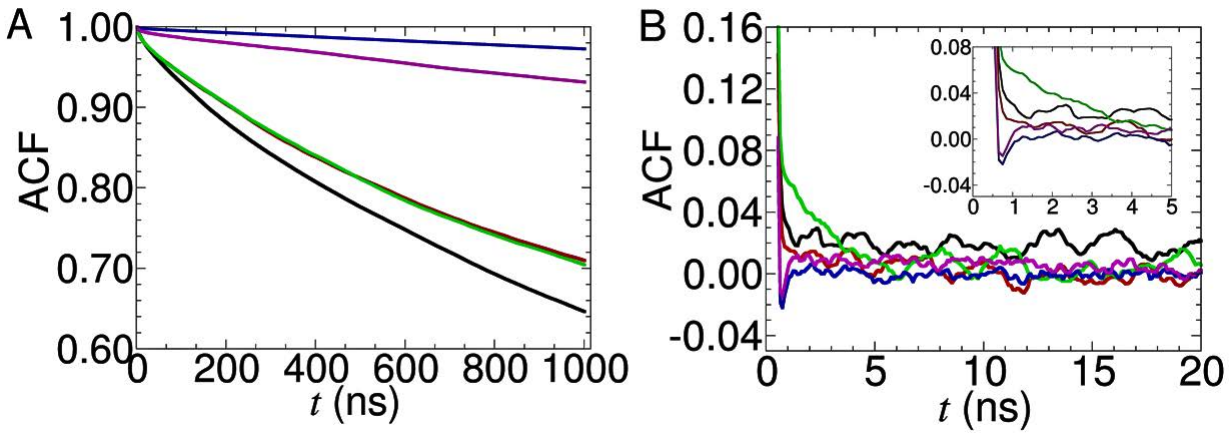

Figure 5.8: (A) Autocorrelation function (ACF) of unrestrained orientational and polar angles collective variables of the pig insulin dimer (unrestrained $\Delta G^{\circ}_{\mathrm{b}}$ equal to -8.5 kcal/mol). (B) Variation of the ACF of restrained orientational and polar angles collective variables 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.

lighting the diversity of the obtained results unless the geometrical route is applied. As well as in the case of the above-discussed shortcut of the geometrical route, we reconstructed the ACFs of the Θ, Φ, Ψ, θ, ϕ of the totally unrestrained separation PMF calculations (Figure 5.9B). Herein, the both replicas' relaxation times necessary to achieve quasi-ergodic sampling along Θ, Φ, Ψ, θ, and ϕ for the pig insulin dimer are 2.63 and 1.52, 4.60 and 2.78, 3.93 and 1.67, 6.89 and 6.69, 4.91 and 4.53 μs, respectively. Under these premises, the relevant time to reach convergence of this strategy is diverse and nearly six times faster than the unrestrained simulations with a pinned protein and seven times longer than the application of the geometrical route for the binding free-energy estimation. While a reasonably converged free-energy profile, mirroring adequate orientational and positional averaging, can be determined faster, the presented shortcuts remain poorly competitive with the full geometrical route, which, for the same protein-protein complex, supplies the correct answer within 0.88 μs (see Table 5.4).

5.3.4 Concluding remarks

This letter examined critically two different avenues commonly followed for the estimation of binding affinities, namely (i) the so-called geometrical route,[44,57,25] whereby geometrical restraints act on the relative orientation and position of the binding partners, and (ii) shortcuts of this strategy,[211,213,214,135,172,215,216,217,218,212] devoid of geometrical restraints, and assesses their respective merits and drawbacks for two protein-ligand and three protein-protein complexes.

Our results show that the binding free energies obtained using the geometrical route reliably and quantitatively match the values inferred from experiment, whereas those predicted from the simplified shortcut routes systematically depart from them. This quantitative disagreement raises legitimate questions about the relevance of approaches oversimplified without theoretical justification to predict with appropriate accuracy standard binding free energies. While the geometrical route could potentially be reduced to a simple separation PMF calculation in the case of rapidly relaxing, small binding partners, as would be the case, for instance, in the benzene dimer,[63] our study demonstrates that in protein-ligand and protein-protein complexes, whereby the position and the orientation decorrelate over the μs–timescale, adequate sampling of the available configurational space is not achieved by simulations of typical length,

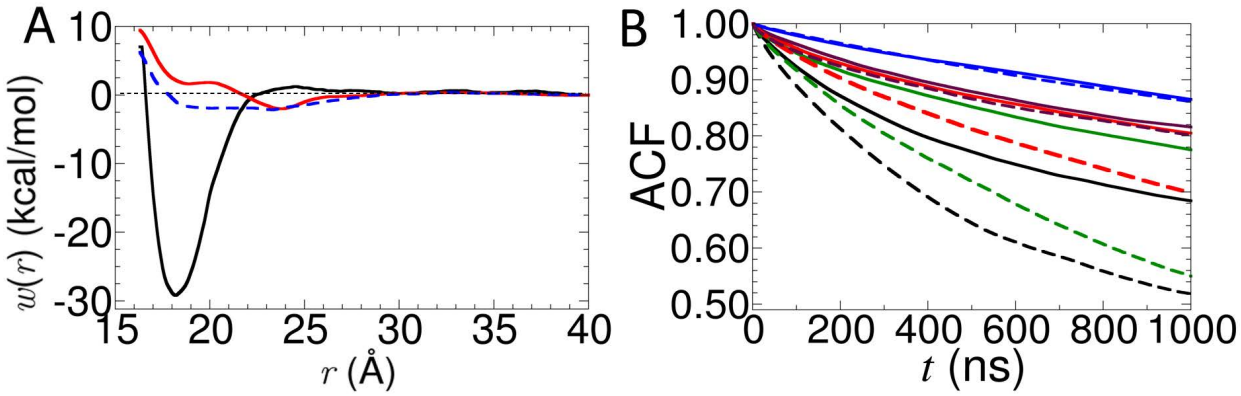

Figure 5.9: (A) Normalized 1–μs separation PMFs for insulin dimer calculated in the totally unrestrained simulation for two replicas (replica 1 in solid red and replica 2 in dashed blue) in comparison to the separation PMF of the geometrical route (black). (B) Autocorrelation functions (ACF) of unrestrained orientational and polar angles collective variables 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.

thus, leading to erroneous estimates of the binding affinity.

In sharp contrast, in the rigorous geometrical route, the PMFs underlying the internal conformational and the relative orientational and positional changes within the binding partners are determined first, as a preamble to the separation PMF calculation, performed with suitable restraining potentials, which remarkably accelerate the convergence and reduces the required computational time. The geometrical route offers full control of every degree of freedom in the reversible association process through a universal set of CVs, and their contribution to the binding free energy can be determined with an appropriate level of accuracy from the corresponding PMFs. At this juncture, it is fair to recognize that however robust and guaranteeing optimum convergence properties of the free-energy calculations, as well as fully reproducible results, the Achilles heel of the geometrical route remains that of all binding free-energy calculations in a general sense, namely the limited accuracy of the force field, and the reliability of the native binding motif, which are both potential sources of discrepancies with experiment.

The protocol of the full geometrical route can be generalized to any complex, whereby binding, either of a protein or a ligand, occurs at the surface of the protein. The burden of setting up the different steps of the protocols, most notably the definition of the relevant CVs, may be viewed as overly daunting and time consuming, thus, enticing the end-user to follow seemingly more appealing and simpler routes. However, a number of tools have been devised to help prepare the free-energy calculations by automating the definition of the CVs and generating all the necessary files for the geometrical route. Such is the case of the CHARMM-GUI[219] and the BFEE2[81,82] software, designed to streamline the entire protocol.

Aside from the seduction of simplicity and its potential dangers, one of the lessons taught by the simulations reported herein is that the computational cost incurred in the rigorous theoretical framework of the geometrical route turns out to be substantially more affordable than that of its shortcuts. In the absence of theoretical substantiation of the latter oversimplified approach, there is no cogent justification to not resort to the geometrical route to obtain accurate binding-affinity estimates.

Chapter 6

Improving Speed and Affordability of Binding Free-Energy Calculations

Sommaire

This chapter presents my work to speed up the binding free-energy calculations using the geometrical route. Firstly, I will start with how this work is related to my thesis and my contribution in a preamble. Then, I will move forward with a summary of the article before providing the original entire publication, which can also be found under the reference:

"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"

6.1 Preamble

The geometrical route is often criticized on account of numerous biased simulations leading to a high computational cost. Our goal was to speed up the calculations without sacrificing the accuracy and reliability of the estimates obtained with this strategy. The first trick consists of using HMR. The second resides in computing the CVs and the forces exerted along these CVs, employing an MTS scheme similar in spirit,[220] due to the substantial computational cost incurred in the evaluation of the CVs, most specifically those involving roto-translational alignments of large numbers of atoms.[221,93] Under these premises, the CVs and the biasing forces are determined every N_{MTS} MD steps, where N_{MTS} is fixed by the end-user, which appreciably augments the computational performance of enhanced-sampling algorithms..[222,223,220] In this work, we combine HMR and MTS to improve the speed of the reversible association calculation in the geometrical route while tuning the extended-Lagrangian parameters and ensuring that the simulations' accuracy, sampling uniformity, and adequate convergence rates are preserved. This investigation was conducted in tandem with my colleague Marharyta Blazhynska, and no specific individual contributions can be determined.

6.2 Summary

In this article, we set up calculations of the physical separation of the Abl-SH3:p41 complex since it is the most sensitive part of the binding free-energy calculation. We took into account the cost of restraints, as they are essential,[224] by using already performed calculations with the geometrical route,[82] in the post-treatment of simulations. We combined (i) a longer time step for the integration of the equations of motion with hydrogen-mass repartitioning (HMR) and (ii) multiple time-stepping (MTS) for collective-variable and biasing-force evaluation in five different schemes:

- no HMR no MTS denoted as our reference
- no HMR with an MTS of 2 (scheme 1)
- no HMR with an MTS of 4 (scheme 2)
- HMR only (scheme 3)
- HMR with an MTS of 2 (scheme 4)
- HMR with an MTS of 4 (scheme 5)

The results of probing these combinations are displayed in the following Table 6.1

Table 6.1: 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[82]	-7.6 ± 0.4[82]	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[57] from the PMF calculated along the physical separation coordinate.

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

The faster schemes can speed up the simulation by a factor of three compared to the reference PMF calculation. However, this gain in computational efficiency comes at a price—the accuracy of the final binding free–energy estimate has perceptibly deteriorated. For instance, scheme 4 is 1.5 kcal/mol away from the reference $\Delta G^{\circ}_{\mathrm{b}}$, corresponding to a significant variance in the obtained PMFs. Further analysis revealed that schemes 4 and 5 suffered from a cumulative effect of suboptimal extended-Lagrangian parameters. To test this idea, we investigated the influence of the individual extended-Lagrangian parameters (damping factor, extended fluctuation, and oscillation period) on the physical-separation PMF calculation for scheme 4. Since this scheme represents the lowest accuracy among the other schemes, the impact of tuning up the parameters of the free-energy calculations is expected to be more glaring compared to the alternate schemes. Keeping all other parameters equal to their reference values, we found that an increase of the damping factor, or a decrease of the extended fluctuation, is prone to enhance significantly the performance of the free-energy calculation, speeding it up nearly three times with the

application of the combined schemes, compared to the reference physical-separation simulation, while improvement of the efficiency when using MTS alone was more modest. Although we were able to find the specific parameters for the Abl kinase-SH3:p41 complex, we probed the selected schemes on an additional system, MDM2-p53:NVP-CGM097,[180] to validate our results further. As a take-home message, we draw attention to each complex's specificity and recommend optimizing the extended-Lagrangian parameters before using the proposed accelerated schemes.

6.3 Original article

6.3.1 Introduction

Quantifying the standard binding free energies of protein-ligand and protein-protein complexes by molecular dynamics (MD) simulations has been attracting great interest in rational drug design.[25,187,227,228,229,230,231,232,233,225,226] Although alchemical transformations,[125] employing, for instance, free-energy perturbation (FEP),[234,64] have successfully predicted the binding affinities of small molecules towards proteins, they suffer from poor convergence in the event of large perturbations, precluding their use for protein–protein complexes.

Another popular approach for standard binding free-energy calculations of protein-ligand and protein-protein complexes is molecular mechanics/Poisson-Boltzmann and surface area (MM-PBSA), or generalized Born surface area (MM-GBSA).[14,190,191,17] This method evaluates the binding energy of a complex by adding up its gas-phase energy, solvation free energy, and a contribution of the conformational entropy of the solute determined from MD simulations.[14,192] Despite its wide use and relatively low cost, the MM-PBSA approach is built upon several uncontrolled approximations of unknown validity.[19,20,25] For instance, the solvation free energy is often calculated using the Poisson-Boltzmann equation, which assumes that the solvent is a continuous dielectric medium.[14,190,191,17,192] However, this assumption may not hold for complexes featuring regions of high dielectric constant.[193,235] Additionally, the conformational entropy of the solute is estimated based on a single MD simulation, which may not accurately capture the entropic contribution of the binding process.[19,20,25,193,235]

To alleviate the aforementioned limitations, one may turn to a so-called "geometrical route", or method based on a physical pathway,[44,27,26] wherein the potential of mean force (PMF) along an arbitrary dissociation path is determined. Gumbart et al.[25] put forth a full geometrical route that evaluates the binding free energy using stepwise PMF calculations along the collective variables[236,237] (CVs) proposed by Woo and Roux,[44] which describe the different stages of the reversible separation of a protein-ligand complex.

Determination of CVs can be refined[178] and automated through, for instance, the binding free-energy estimator (BFEE) plug-in, and its updated standalone version BFEE2,[238,81,82] available online.[239] making use of the efficient sampling algorithm well-tempered meta-extended adaptive biasing force[49] (WTM-eABF) to speed up the convergence of the PMF calculations. Such a stepwise geometrical route has proven accurate for the estimation of standard binding free energies of protein-ligand complexes,[82] e.g., Src homology 3 (SH3) domain of tyrosine kinase Abl and a family of decapeptides.[240,179] Moreover, this method has also been extended to the determination of the binding affinities of protein-protein complexes,[57] e.g., the receptor binding domain of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants and its dedicated target, the human angiotensin conversion enzyme 2.[241] The total computational cost incurred in this geometrical route remains, however, substantial, which has led shortcuts to be sought.[218,216,135,172] These shortcuts are aimed at simplifying the stepwise PMF calculations into a single one for the physical dissociation of the complex, ignoring the contributions from the relative orientation, position and conformational changes of the binding partners, an approximation that has proven unreliable.[224]

The evaluation of the force-field terms and the CVs are particularly time-consuming in PMF calculations. To reduce the computational cost of the simulation, these force-field terms could be integrated with a longer time step, given that fast degrees of freedom are suitably handled, on the one hand, by the Shake/Rattle/Settle algorithms,[46,45,242] and, on the other hand, by the hydrogen-mass repartitioning (HMR) trick,[47,243] which is broadly used in MD-based investigations.[244,245,246,247] Concomitantly, the CVs and the corresponding biasing forces could be computed at a lower frequency in lieu of every single MD step, turning to a multiple-time-stepping (MTS) strategy.[220] Put together, to assuage the computational burden of binding-affinity calculations, we propose to meld an MTS scheme for CV and biasing-force evaluation with HMR in order to accelerate PMF calculations, while tuning the WTM-eABF parameters to preserve reliability of our simulations. To illustrate the efficiency of the proposed methodology, we carried out fifty independent WTM-eABF simulations in triplicates for the physical separation of the SH3 domain of the Abl tyrosine kinase in complex with the p41 ligand,[179] examining different associations of MTS, HMR and enhanced-sampling parameters to discover the optimal combination ensuring fast and accurate standard binding free-energy estimations.

6.3.2 Methods

The daunting computational challenge posed by standard binding free-energy calculations is rooted in the significant configurational entropy change occurring upon reversible association of the binding partners, difficult to sample in brute-force MD simulations. In this work, we applied the MD-based approach proposed by Woo and Roux,[44] where the chosen CVs represent the slow degrees of freedom associated with the relative movements of the molecular objects at play during their physical separation. These CVs can be measured and controlled throughout the MD simulation. The introduction of geometrical restraints acting in the form of soft harmonic potentials on the relevant CVs allows the available configurational space to be reduced, and the slow degrees of freedom of the reversible association to be adequately captured, thus resulting in accelerated convergence of the free-energy calculation.[25,57,178]

A formalization of this strategy as well as the mathematical underpinnings for the standard binding free-energy estimation, referred to as the "geometrical route" was introduced by Gumbart et al.,[25] and considers that the physical separation of the two molecular objects is performed in the presence of orientational (Θ, Φ, and Ψ Euler angles), positional (polar, θ, and azimuthal, ϕ, angles), and conformational (distance root-mean-square deviation, or RMSD, of the substrate) restraints to control the change in configurational entropy that accompanies the dissociation process. Additional detail about the application of this methodology to protein–ligand complexes can be found in references,[44,25,57,178,82] and.[224]

To sample efficiently the relevant regions of the CV space that contribute significantly to the binding free energy of the complex, turning to a suitable enhanced-sampling algorithm[248,249] is strongly recommended to perform the PMF calculation. In the present work, we employed the WTM-eABF scheme.[49,82] WTM-eABF combines the strength of the extended-Lagrangian[250] variant of ABF[251,51,54] (eABF) and the well-tempered variant of metadynamics[252,50,253] to effectively sample the free-energy landscape without risking to be trapped in local minima.

In this algorithm, the molecular system is evolving under Langevin dynamics at a temperature T, and the total energy, $U(\mathbf{x},\lambda)$, depends not only on the force-field term, $U_{\mathrm{FF}}(\mathbf{x})$, but also, first, on the spring connecting the real CV to a fictitious particle, and, second, on additional biasing potentials, $U_{\mathrm{bias}}(\lambda)$. The corresponding equations of motion are,

$$\begin{cases} U(\mathbf{x},\lambda) & = U_{\mathrm{FF}}(\mathbf{x}) + \dfrac{1}{2}k\big(\xi(\mathbf{x})-\lambda\big)^2 + U_{\mathrm{bias}}(\lambda) \\ \mathbf{m}_{\mathbf{x}}\ddot{\mathbf{x}} & = -\nabla_{\mathbf{x}}U(\mathbf{x},\lambda) - \gamma_{\mathbf{x}}\mathbf{m}_{\mathbf{x}}\dot{\mathbf{x}} + \sqrt{2\gamma_{\mathbf{x}}k_{\mathrm{B}}T}\,\mathbf{m}_{\mathbf{x}}^{1/2}\,\dot{\mathbf{W}}(t) \\ m_\lambda\ddot{\lambda} & = -\nabla_\lambda U(\mathbf{x},\lambda) - \gamma_\lambda m_\lambda\dot{\lambda} + \sqrt{2\gamma_\lambda k_{\mathrm{B}}T}\,m_\lambda^{1/2}\,\dot{\mathbf{W}}(t) \end{cases} \tag{6.1}$$

where $\xi(\mathbf{x})$ is the real CV, a function of Cartesian coordinates $\mathbf{x}$, and λ is the extended degree of freedom, i.e., the fictitious particle. k is the force constant of the spring. $\mathbf{W}(t)$ is a time-dependent Wiener process. $\mathbf{m_x}$ and m_λ are the masses of the atoms and the extended variable, respectively. $\gamma_\mathbf{x}$ and γ_λ are the friction coefficients of the atoms and the extended variable, respectively. k_B is the Boltzmann constant.

The coupling between the fictitious and the real particles has to be strong enough to transfer the applied forces to the actual CV. Two essential parameters that control the spring and, therefore, play an important role in the sampling are the oscillation period, τ, and the coupling width, σ. Both of them affect the inertial mass of the fictitious particle, as per,

$$\begin{cases} m_\lambda &= k_\mathrm{B}T\left(\dfrac{\tau}{2\pi\sigma}\right)^2 \\ k &= \dfrac{k_\mathrm{B}T}{\sigma^2} \end{cases} \tag{6.2}$$

Another extended Langevin dynamics parameter is the damping factor, γ_λ, which, depending upon its value, could slow down, or accelerate diffusive sampling[254,255,256,257] in the PMF calculation. How the PMF can be reconstructed from the WTM-eABF simulation is described in the Supporting Information (SI).

Since in the geometrical route, the energetic contributions of each applied restraint onto the CVs have to be evaluated via PMF calculations, a correct binding free–energy estimation requires a significant number of independent simulations—i.e., eight for protein-ligand binding,[82] and up to fourteen for protein-protein binding[57]—and computational time to reach convergence. For this reason, acceleration of these simulations without loss of accuracy is much desirable and of paramount importance.

One way to speed up the MD simulations consists in applying a longer time step for the force-field terms with the HMR trick. HMR artificially increases the mass of hydrogen atoms by distributing that of the heavier atoms they are chemically bonded to, conserving the overall molecular mass.[258,47,243] The recommended mass for methyl and methylene hydrogen atoms is 3 g/mol.[47] In this case, the fastest motion of the molecular objects, i.e., the vibration of chemical bonds involving hydrogen atoms, is slowed down, thereby allowing an integration time step of up to 4 fs to be used without affecting energy conservation in the MD simulation, and reproduction of thermodynamic quantities.[243] If multiple time-stepping, for instance, in the form of the Verlet-I/r-RESPA algorithm,[222] is applied, effective time steps as large as 8 fs can be used for updating the long-range contribution of the force field.

Another avenue to accelerate biased molecular simulations consists in computing the CVs and the forces exerted along these CVs, employing an MTS scheme similar in spirit.[220] The rationale for turning to this functionality is the substantial computational cost incurred in the evaluation of the CVs, most specifically those involving roto-translational alignments of large numbers of atoms.[221,93] Under these premises, the CVs and the biasing forces are determined every N_MTS MD steps, where N_MTS is fixed by the end-user—2 or 4 in the present work, which appreciably augments the computational performance of enhanced-sampling algorithms.[222,223,220] In this work, we combine HMR and MTS to improve the speed of the reversible association calculations through the geometrical route, while tuning the extended-Lagrangian parameters and ensure that accuracy, sampling uniformity, and adequate convergence rates of the simulations are preserved.

6.3.3 Computational assays

As the model biological complex for our investigation, we used the Abl-SH3 domain in complex with the proline-rich peptide p41 (APSYSPPPPP), which plays an essential role in intracellular signaling.[240,179] The structure of the Abl kinase-SH3:p41 dimer has been solved with a resolution of 1.65 Å by Pisabarro et al., using X-ray crystallography, and was deposited in the Protein Data Bank with entry code

1BBZ.[179] The stability of the complex is maintained by means of an intermolecular hydrogen bond between the carbonyl groups of Y4 and P8 of p41, on the one hand, and W36 and Y52 of the protein, on the other hand.[179,259,260] The standard binding free energy was measured experimentally as -8.0 ± 0.1 kcal/mol.[179]

The computational assay consisted of 73,054 atoms in total. The dimensions of the periodic cell were $94 \times 88 \times 96$ Å^3. The complex was described by the all-atom CHARMM36m force field,[175] with the TIP3P water model.[176] Sodium and chloride ions were added to provide electric neutrality and a 150 mM salt concentration, which corresponds to physiological conditions.[177] The full binding free–energy estimation for the Abl kinase-SH3:p41 complex via the geometrical route was described previously in references,[25,178,82] and.[224] Here, using BFEE2,[81,82,239] we bypassed the step of input-file preparation, and focused primarily on a modification of already prepped physical-separation PMF calculations along the center-of-mass distance between the protein and the ligand, where all the energetic contributions from the other degrees of freedom in the bound and the unbound states are known from a previous study.[82] More specifically, we modified our input files to reflect association of HMR and MTS, as well as the three extended-Lagrangian parameters, σ, γ_λ, and τ, as reported in Table 6.2. An example of a modified Colvars[93] input file for scheme 21, and the necessary files to perform the accelerated PMF calculations, are supplied in the SI.

For consistency, all the simulations were performed on servers equipped with 32 CPU cores and two GPU cards (GeForce RTX 2080 Ti), using the NAMD 3.0 MD engine.[86] The temperature and the pressure were kept at 300 K and 1 atm, using Langevin dynamics and the Langevin piston, respectively.[157,156] Short-range interactions were smoothly truncated between 10 and 12 Å. The pair-list distance was 14 Å. For the long-range electrostatic interactions, the particle-mesh Ewald (PME) algorithm[158] was used. The equations of motion were integrated with an effective time step of 4 and 2 fs, using or not HMR, respectively.

The physical separation PMFs were computed with the Colvars module.[93] The distance between the center of mass of the two binding partners ranged from 10.3 to 34.3 Å, and was discretized in 0.1–Å bins. The resulting separation PMFs, in combination with the remainder conformational, orientational, and positional PMFs obtained previously[82] were fed to the BFEE2 software[81,82] to compute the final standard binding affinity, $\Delta G^\circ_\mathrm{b}$.

Table 6.2: 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
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

6.3.4 Results and Discussions

Probing the acceleration schemes. The results of applying our acceleration schemes with the default extended-Lagrangian parameters (see Table 6.2, schemes 1-5) are gathered in Table 6.3. For each scheme, to make our results unassailable, we triplicated the 50–ns physical-separation PMF calculations, and an averaged PMF is shown in Figure 6.1A. The uncertainty values reported in Table 2 represent the standard deviation associated with the binding free energy, which is calculated from the three independent replicas. As can be noticed from Table 6.3, usage of the faster schemes can speed up the simulation by a factor of three, compared to the reference PMF calculation. However, this gain in computational efficiency comes at a price—the accuracy of the final binding free–energy estimate is perceptibly deteriorated. For instance, scheme 4 is 1.5 kcal/mol away from the reference $\Delta G^{\circ}_{\mathrm{b}}$, and corresponds to a large variance in the obtained PMFs. Interestingly enough, the average $\Delta G^{\circ}_{\mathrm{b}}$'s for schemes 1 and 3 are similar, and only about $k_{\mathrm{B}}T$ away from the experimental value.[179]

Progress of the convergence was monitored by calculating the RMSD between the free–energy gradients, as depicted in Figure 6.1B. Examination of this figure reveals that scheme 3 converges slightly faster than scheme 1. Although the binding affinity for schemes 4 and 5 departs markedly from experiment, these schemes correspond to faster and smoother convergence than any other scheme, which stems from a cumulative effect of suboptimal extended-Lagrangian parameters for this particular protein-ligand complex. To test this idea, we decided to investigate the influence of the individual extended-Lagrangian

Table 6.3: 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[82]	-7.6 ± 0.4[82]	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[25] from the PMF calculated along the physical separation coordinate.

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

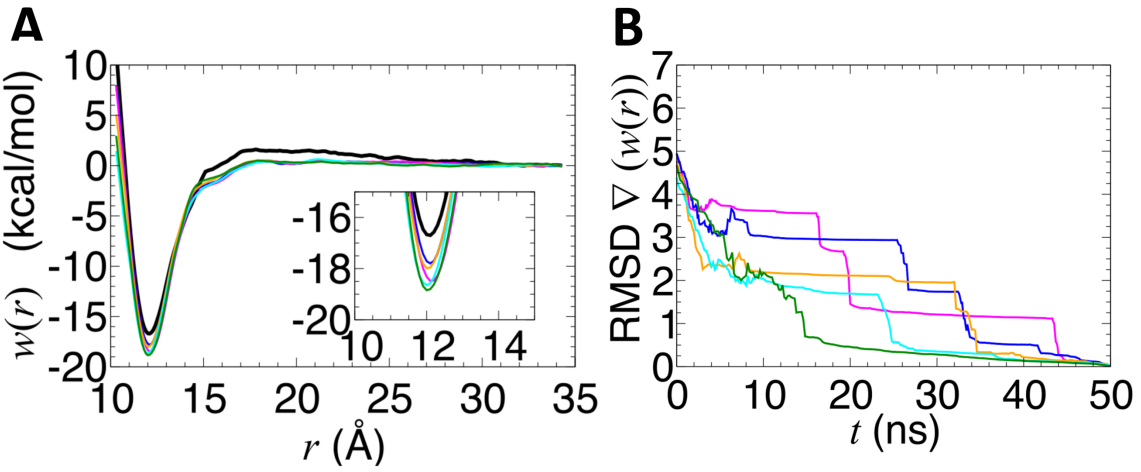

Figure 6.1: (A) Averaged physical-separation PMFs for three replicas obtained after individual 50–ns simulations. All the PMFs were determined within the separation distance range of [10.3; 34.3] Å. (B) Averaged convergences for the physical-separation PMFs. The curves correspond to the different calculation schemes: Reference (black), scheme 1 (blue), scheme 2 (magenta), scheme 3 (orange), scheme 4 (cyan), and scheme 5 (green).

parameters on the physical-separation PMF calculation for scheme 4. Since this scheme represents the lowest accuracy amid the other schemes, the impact of tuning up the parameters of the free-energy calculations is expected to be more glaring, compared to the alternate schemes.

Damping factor for the Langevin dynamics of the extended variable. As the first extended-Lagrangian parameter to be explored, we selected the damping factor, γ_λ, the appropriate choice of which is likely to enhance sampling efficiency.[261,257] Correctly optimized, this parameter also controls the reproduction of the self-diffusion properties of the system, and enforces temperature conservation, preventing overheating of the extended variable.[261] The results of applying different values of γ_λ to computational scheme 4, keeping all other extended-Lagrangian parameters untouched, are reported in Table 6.4 (protocols 6–9). Each value of ΔG°_b was obtained as the average over three replicas, and supplied with the measured standard deviation. Among the protocols whereby γ_λ was varied, two, namely 8 and 9, best match the experimental binding affinity.[179] However, only protocol 8 possesses a low standard deviation, suggestive of a noteworthy accuracy of this theoretical prediction. The averaged PMFs for protocols 6–9, and the corresponding averaged number of samples per bin, are displayed in Figure 6.2A.

To examine the sampling uniformity of our calculations, we analyzed the number of force samples per bin for the selected schemes at the end of the 50–ns separation PMF calculations (Figure 6.2B). Being relatively uniform along ξ for all the studied cases, the sampling, however, remains relatively low, considering the length of the simulation. The smallest number of samples was obtained 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 6.2C). Protocols 7 and 8 correspond to the fastest convergence, whereas protocol 6 is unlikely to be converged. Put together, protocol 8 appears to reflect the best choice for γ_λ 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 avenue to improve the binding free–energy calculations within the proposed acceleration schemes consists in tuning the so-called "extended fluctuation",

also known as the coupling width, σ, which controls the standard deviation of the fictitious degree of freedom coupled to the real CV. This parameter alters the stiffness of the spring connecting the real and the fictitious particle and the mass of the latter (see eq 6.2). Studying in vacuum the reversible folding of deca–alanine along its end–to–end distance using standard eABF, Lesage et al. showed that when overlarge, σ could lead to accelerated convergence, yet at the price of undersampled regions of the reaction pathway.[54] We examined three extended-fluctuation 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. Only protocol 11 turned out to be within $k_{\mathrm{B}}T$ from the experimental target value,[179] and corresponded to the greatest consistency over the triplicates. Visual inspection of the physical-separation PMFs reveals that protocol 12 (σ = 0.5 Å) yields the deepest valley (Figure 6.2D) and the fastest convergence rate among the other schemes (Figure 6.2F), which could stem from insufficient sampling over the 10–20–Å separation range (Figure 6.2E). Our results further show that the convergence rate slows down with decreasing σ. This statement is in contrast with that of Lesage et al.,[54] which could be ascribed to the application of the MTS algorithm, the number of restrained CVs involved in the physical separation PMF calculations, as well as the use of distinct enhanced-sampling algorithms—namely, eABF versus WTM-eABF.

The other parameter that controls the coupling between the extended degree of freedom and the CV is the oscillation period, τ. We tested two more values for this parameter, specifically 100 and 300 fs, in combination with HMR and MTS, with a time interval of 2 (protocols 13 and 14), and compared our results to those obtained with the reference value of $\tau = 200$ fs of scheme 4 (see Table 6.4, Figure 6.2G, H, I). Protocol 14, with τ = 300 fs, improved the accuracy compared to scheme 4. Detailed analysis of the number of samples per bin reveals for protocol 13 that sampling is insufficient at small separations ($10 \leq r \leq 17$ Å). In addition, convergence is notably slowed down in protocols 13 and 14, compared to the reference (see Figure 6.2I). Despite the gain in accuracy of $\Delta G^{\circ}_{\mathrm{b}}$, our results suggest that τ = 200 fs could be considered as a safe choice for the present investigation.

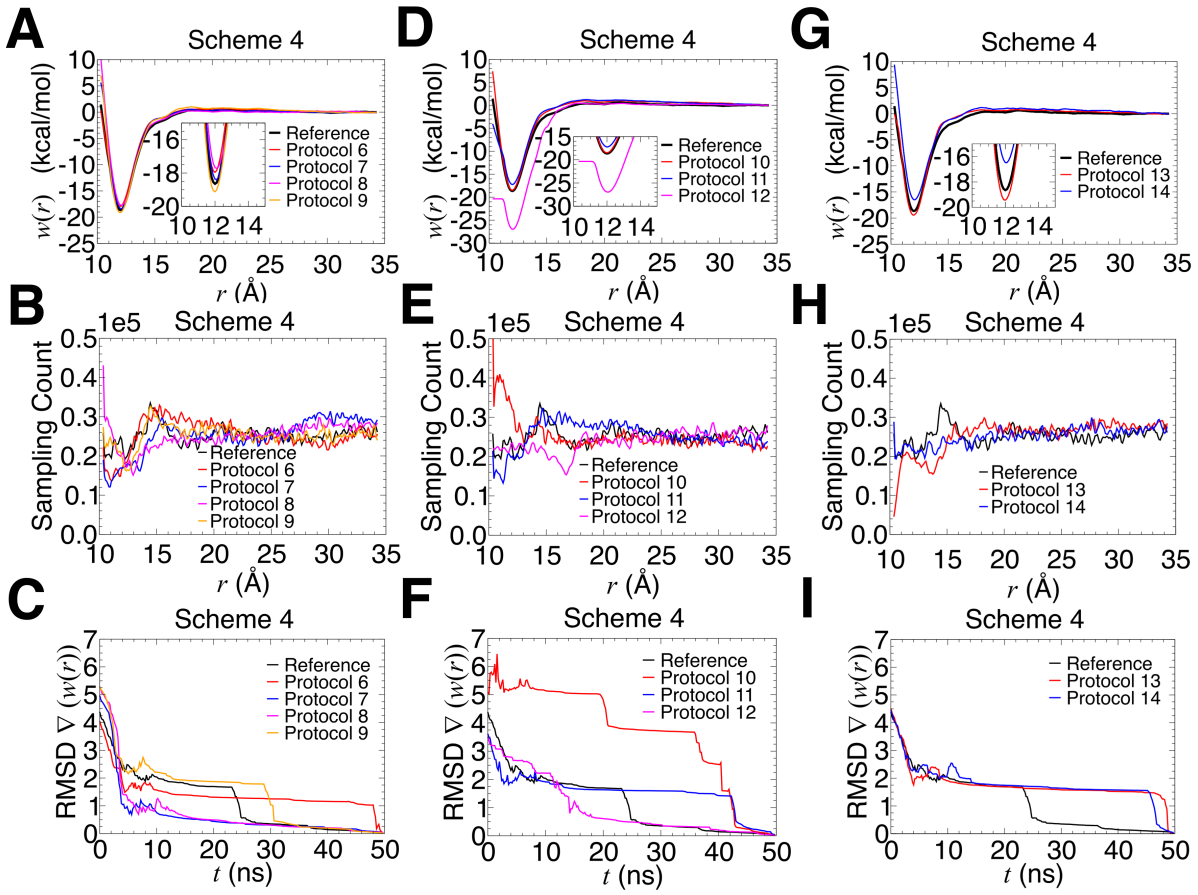

Figure 6.2: Panels A–C correspond to scheme 4 (denoted as Reference), and protocols 6, 7, 8, and 9 with γ_λ = 1 (black), 3 (red), 5 (blue), 7 (magenta), and 10 (orange) ps^{-1}. Panels D–F correspond to scheme 4, and protocols 10, 11, 12 with σ = 0.1 (black), 0.01 (red), 0.05 (blue) and 0.5 (magenta) Å. G–I correspond to scheme 4, and protocols 13, 14 with τ = 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) Average number of samples per bin achieved in the simulations. (C, F, I) Average convergence rates achieved in the simulations.

Table 6.4: 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))	$\Delta G^\circ_{\mathrm{b}}$ (kcal/mol)	Speed (ns/day)
Scheme 4[a]			-9.1 ± 1.9	137
⑥			-8.6 ± 1.0	142
⑦	59.6	6.0×10^4	-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 6.3 to facilitate the comparison with the results of the application of protocols.

Mixing different extended-Lagrangian parameters. After probing the influence of each of the extended-Lagrangian parameters of scheme 4, we selected those that had the biggest impact on our criteria for PMF calculations, namely accuracy, sampling uniformity, and convergence rate, and applied them both independently and simultaneously to the schemes corresponding to different HMR and MTS parameters (see Table 6.5, schemes 1, 2, 3, and 5). Particularly, we investigated the impact of the selected protocols 15–22 onto schemes 1–5 (Table 6.2). Herein, the chosen parameters, τ = 300 fs, σ = 0.05 Å, γ_λ = 10 and 7 ps^{-1}, were changed individually, keeping all other parameters equal to their default value (protocols 15, 16, 17, and 18, respectively). Alternatively, we changed the parameters in a series of combinations, viz., σ = 0.05 Å, τ = 300 fs, γ_λ = 10 ps^{-1} (protocol 19), σ = 0.1 Å, τ = 300 fs, γ_λ = 10 ps^{-1} (protocol 20), σ = 0.05 Å, τ = 300 fs, γ_λ = 7 ps^{-1} (protocol 21), and σ = 0.05 Å, τ = 300 fs, γ_λ = 7 ps^{-1} (protocol 22). Henceforth, application of an updated extended-Lagrangian parameter set to the acceleration schemes 1–5 will be shown with a slash (/), e.g., protocol 15 applied to scheme 1 will be referred to as combination 15/1 (-9.2 ± 0.4 kcal/mol from Table 6.5).

It can be noticed from Table 6.5, that protocol 16 provides the best accuracy for most schemes. Furthermore, we observe better agreement in the estimation of $\Delta G^\circ_\mathrm{b}$ with combinations 18/3 and 18/4, as well as 19/2, 21/1, and 22/2. Among all possible combinations, only in a few cases are we close to the experimental binding affinity, yet associated to high standard deviations (especially 17/4, 21/2, and 22/3). These results demonstrate that concomitant change of different parameters may lead to deterioration in the reproduction of the standard binding free energy—although when performed independently, this change might improve the agreement.

To test the accuracy and precision of our results, we increased the number of replicas for protocols 16, 18, and 21 up to five, as these protocols showed the most promising agreement with the available standard binding free-energy estimates. The results of this study are gathered in Table 1 of the SI. Specifically, for some combinations (namely 18/2, 18/3, 18/5, 21/2, 21/4, and 21/5), the quintuplicate separation simulation led to smaller uncertainties than those presented in Table 6.6, thereby suggesting improved accuracy. In other combinations, the standard deviations fell within the same $k_\mathrm{B}T$ range as the triplicated simulations, or slightly exceeded it, suggestive of a similar, or slightly larger degree of uncertainty. Specifically, combinations 18 and 21 consistently outperformed combination 16 in predicting the experimental binding free energy value of -7.99 kcal/mol.[179] However, despite the increase in the number of replicas, our overall conclusion regarding the comparison of these protocols remains unchanged.

Analysis of the sampling (see Figure 6.3) indicates that for most protocols (15–22), schemes 1 and 3 correspond to the largest number of samples per bin across the reaction pathway. In contrast, scheme 5 corresponds to the smallest number of samples per bin, which stems from the association of HMR and MTS, updating the CVs at a low frequency. At small separation, r, combinations 15/2, 16/4, 17/5, 18/3, 19/5, 20/5, 20/3, 21/1, 21/4, and 22/5 result in sampling non-uniformity. Interestingly enough, the sampling efficiency for schemes 2 and 4 are nearly identical for all protocols examined here, except for 18 (see Figure 6.3C). In this particular case, scheme 2 yields the most uniform sampling, with the largest number of samples per bin, compared to any other scheme. Moreover, from the analysis of the convergence plots (see Figure 6.4), it can be noticed that convergence is markedly slowed down compared to the other schemes, mirrored in the poor estimate of the binding affinity (-6.2 ± 1.6 kcal/mol).

The fastest convergence rates are obtained with protocols 16 and 21. Conversely, convergence was not achieved for combinations 19/3, 15/4, and 20/1. The present exploration of parameter combinations (see Figure 6.4 E–H) demonstrates how the choice of the damping coefficient, γ_λ, and the extended-fluctuation term, σ, can affect the convergence of the simulations. To understand how a change in the protocol can modulate the efficiency of the acceleration schemes, we performed an analysis of the probability distributions of the differences between the real, ξ, and fictitious, λ, variables for the physical-separation simulation, following combinations 21/3 and 20/4, for which the reproduced binding affinities vary significantly, i.e., -7.2 ± 0.3 and -10.8 ± 0.2 kcal/mol, respectively (see Table 6.5 and Figure 6.5A,

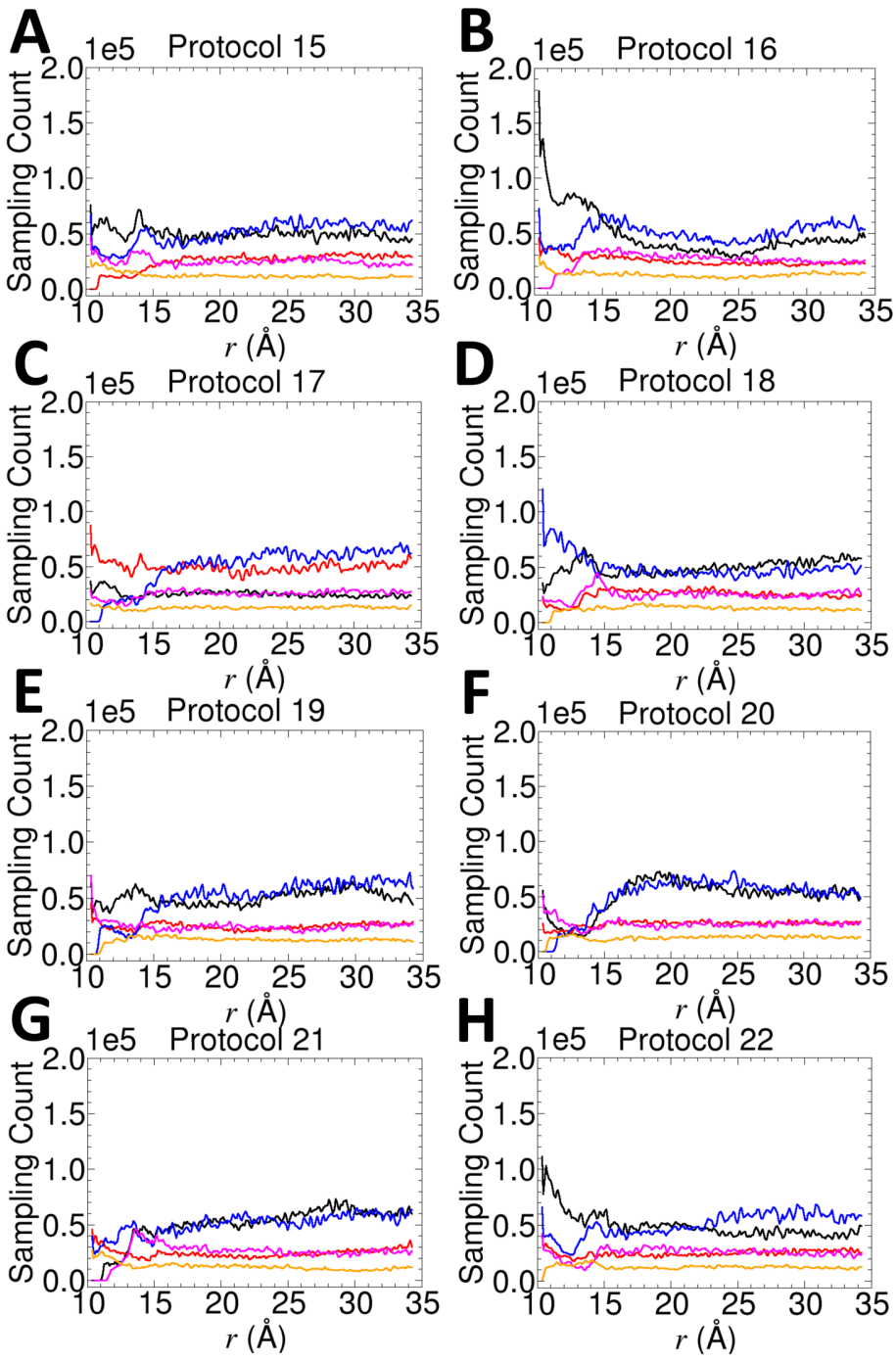

Figure 6.3: Number of samples per bin at the end of the 50–ns separation PMF simulation obtained for combination of protocols (15-22) with schemes 1–5: (A) protocol 15 (τ = 300 fs), (B) protocol 16 (σ = 0.05 Å), (C) protocol 17 (γ_λ = 7 ps^{-1}), (D) protocol 18 (γ_λ = 10 ps^{-1}), (E) protocol 19 (τ = 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).

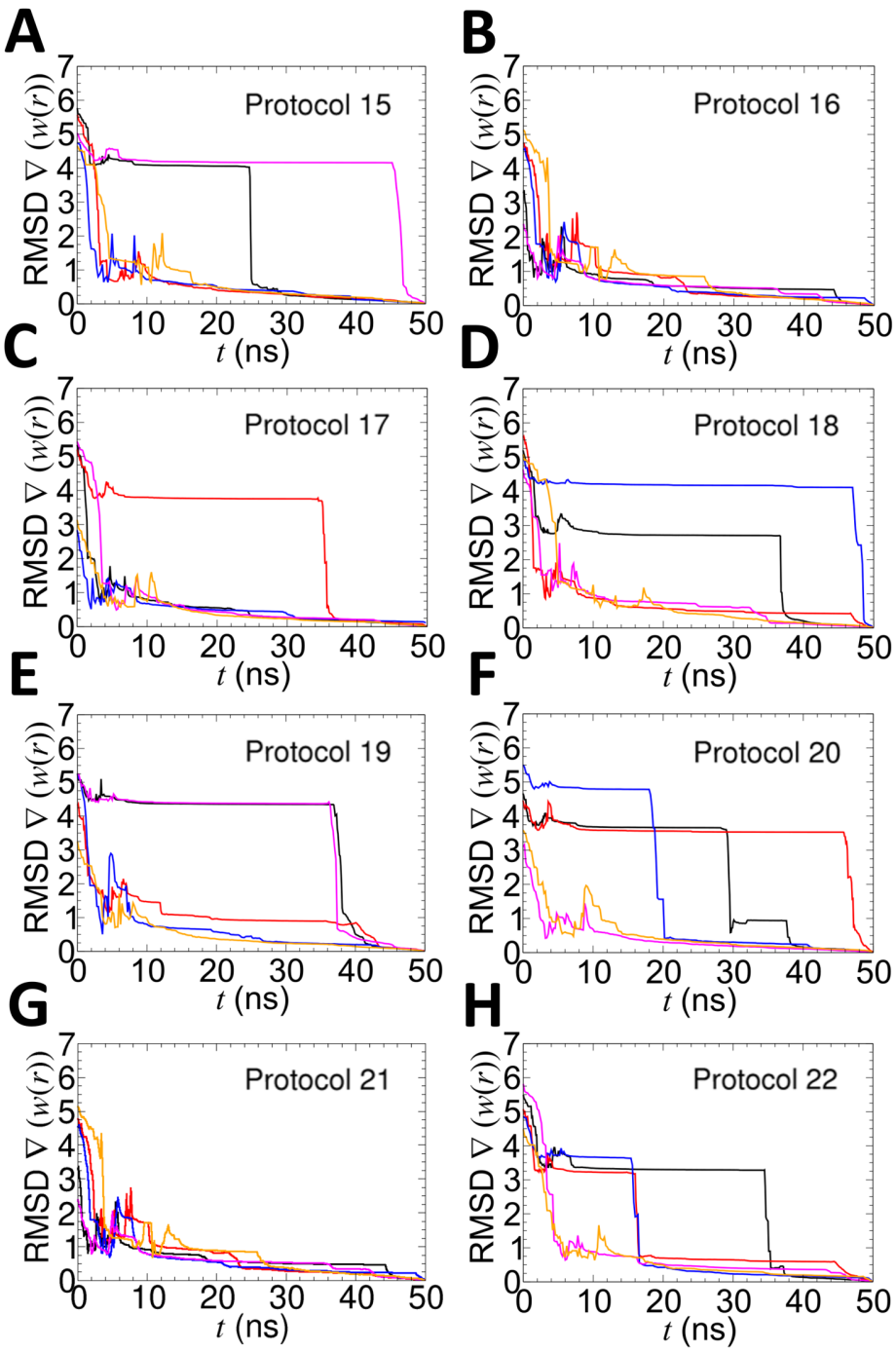

Figure 6.4: Convergence properties of the PMF calculations. (A) protocol 15 (τ = 300 fs), (B) protocol 16 (σ = 0.05 Å), (C) protocol 17 (γ_λ = 7 ps^{-1}), (D) protocol 18 (γ_λ = 10 ps^{-1}), (E) protocol 19 (τ = 300 fs, σ = 0.05 Å and γ_λ = 10 ps^{-1}), (F) protocol 20 (τ = 300 fs, σ = 0.1 Å and γ_λ = 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).

Table 6.5: 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

B). In the case of combination 21/3, the Gaussian probability distribution reflects the synchronization of ξ and λ along the reaction pathway (see Figure 6.5A). The time-evolution of the separation coordinate, as well as that of the harmonically restrained degrees of freedom Θ and Φ (see Figure 6.5C) confirm the expected behavior of the CV throughout the physical separation. Since in combination 21/3, convergence is not fully attained within 50 ns, there is still a possibility to improve the estimation of ΔG_b° with additional sampling (see Figure 6.4G, blue).

In contrast, in combination 20/4, the positively skewed distribution reflects the weak coupling between the variables, which is also mirrored in the drift in the time-evolution of variable r, and impacts the harmonically restrained Θ and Φ angular CVs at time interval 30–40 ns (Figure 6.4D). Interestingly enough, the convergence for this combination was reached as early as 40 ns (see Figure 6.4F, magenta).

Put together, the results obtained here show that the choice of protocols in association with the acceleration schemes may significantly modulate the number of samples per bin, convergence rate, and accuracy of the 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, application of MTS with a frequency of 4 (scheme 2) with σ= 0.05 Å, i.e.,

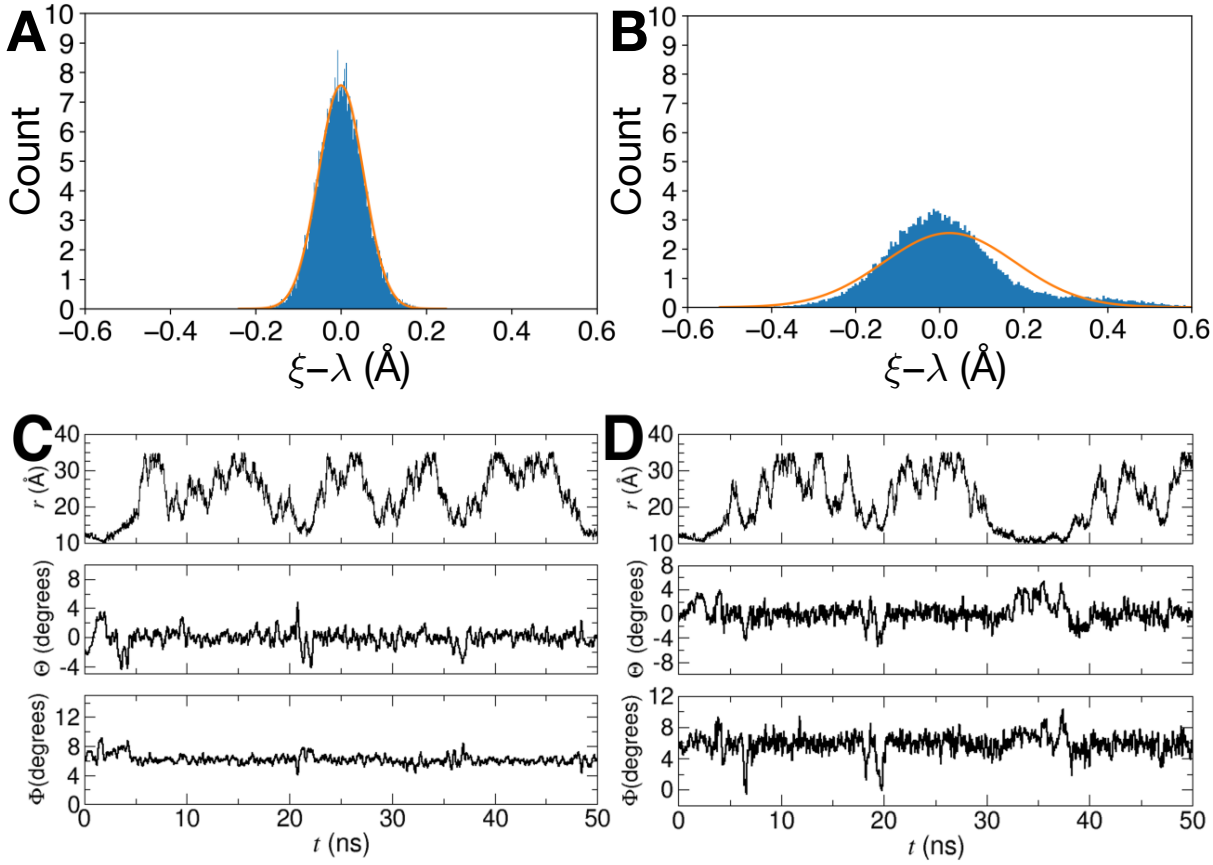

Figure 6.5: (A, B) Probability distribution of the difference between the real, ξ, and the fictitious, λ, particles. The orange curves correspond to a true Gaussian distribution fitted to the data. The plots are built with the help of the matplotlib python library.[262] (C, D) Running averages of the 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 6.5).

combination 16/2, while keeping all the remainder parameters as the default values (γ_λ = 1.0 ps^{-1}, τ = 200 fs), enables the experimental estimate to be reproduced accurately. However, the increase of the oscillation period to 300 fs (combination 15/2), or of the damping factor to 7, or 10 ps^{-1} (combinations 17/2 or 18/2) alone results in erroneous estimates. This shortcoming can be ascribed to the increased inertial mass of the fictitious particle, or to overdamping, leading to inaccurate integration of the equations of motion for the extended coordinate. Despite the favorable application of the series 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), the attention of the end–user ought to be drawn to the careful use of these combinations, which can result in lesser force samples, and possible slowed convergence.

For scheme 3 (only HMR application), the best combination of the extended parameters is the use of the default values, which is in accordance with the purpose of the HMR method to not affect the stability of the simulations, nor the structural properties of the complex.[47,243] For the rest of the acceleration schemes (4 and 5), applying different protocols, there is a high possibility of a deteriorated reproduction of the physical-separation PMF. If in scheme 4, this issue could be overcome through lowering of the extended fluctuation (σ= 0.05 Å), or increasing the damping factor for the extended-Lagrangian dynamics (up to γ_λ = 7.0 ps^{-1}), keeping τ to its default value, the problem cannot be addressed by simply tuning the relevant parameters, and necessitates additional effort to yield a reasonable binding free–energy estimate.

To further support the findings of our study, we expanded our investigation using the same strategy to the MDM2-p53:NVP-CGM097 protein-ligand complex,[180] which has been previously studied in our group.[82] The results of our standard binding free-energy estimations for this complex, employing a variety of computational schemes, are summarized in the SI. We found that the application of HMR (scheme 3) with default values for the extended-Lagrangian parameters provided the closest agreement with the experimental data, and was 1.5 times faster than using MTS alone (schemes 1 and 2). However, the efficiency of the MTS algorithm may depend on the complexity of the protein-ligand complex examined. We also tested 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 option for this acceleration scheme applied to the complex at hand. Our results show that caution should be exercised when adjusting the extended-fluctuation parameter (protocol 11), as it could potentially overly restrain the CV and affect association in the course of 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) for this particular complex increased the efficiency at the expense of accuracy, leading to a more inertial of the fictitious particle, which, in turn, impacted sampling of the free-energy landscape, and resulted in the underestimation of the free energy. However, it is worth noting that this observation is specific to the MDM2-p53:NVP-CGM097 complex, and may not necessarily transpose to other protein-ligand complexes. Overall, our study underscores the importance of exploring systematically different acceleration schemes and protocols to ensure reliable and accurate results in standard binding free-energy calculations.

6.3.5 Conclusion

We have shown in previous theoretical investigations,[25,238,81,82] that protein–ligand standard binding free energies determined with the geometrical route can reliably and quantitatively match experiment, provided suitable convergence of the different PMF calculations. Still, this approach, however theoretically sound and robust, has often been the object of criticism on account of its computational cost, rooted in the number of required biased molecular simulations to be performed. In this work, we aimed at reducing the computational investment of these binding-affinity calculations through improvement of their

efficiency, yet without sacrificing accuracy, convergence, and sampling uniformity. Towards this end, we have carried out physical-separation PMF calculations on the Abl kinase-SH3:p41 and the MDM2-p53:NVP-CGM097 complexes following the geometrical route, and leveraging MTS[220,223,222] for CV and biasing-force computations, with and without HMR.[47,243] This strategy is warranted by the fact that biasing forces have a smoother dependence on atomic positions, and vary slower than the physical—force-field-derived forces, to the extent that they can be integrated with a larger time step. The objective of MTS is to distinguish the forces at play, thereby allowing CVs and time-dependent biases to be evaluated less frequently,[93,220] which results in a net increase of the computational efficiency. Additional increase can be achieved through HMR by slowing down the highest-frequency motions of the molecular objects at play to maximize the integration time step of the MD simulation.[47] Our results confirm that application of HMR alone helps accelerate calculations by nearly a factor a two compared to the reference physical-separation PMF calculation, without affecting accuracy nor stability of the trajectories,[243] and reproduce the standard binding affinity appropriately. In stark contrast, combinations of HMR and MTS, as well as MTS alone require careful tuning of the extended-Lagrangian parameters to guarantee suitable reproduction of the experiment. For instance, keeping all other parameters equal to their reference values, we found that an increase of the damping factor, or a decrease of the extended fluctuation is prone to enhance significantly the performance of the free-energy calculation, speeding it up nearly three times with the application of the combined schemes, compared to the reference physical-separation simulation, while improvement of the efficiency when using MTS alone was more modest. Still, even armed with an accurate force field and reliable structural data, the specificity of each protein–ligand, or protein–protein complex cannot be overstated, and the end-user is, therefore, advised to optimize the extended-Lagrangian parameters discussed herein as a preamble to the accelerated standard binding free–energy calculations following the geometrical route.

6.4 Supplementary Information for the article

6.4.1 Additional data for the Abl-SH3:p41 complex

This section reports the results obtained with the additional replicas—i.e., five in total—for selected combinations (16,18,21) for the AblSH3:p41 complex.

6.4.2 Data for the extra complex: MDM2-p53:NVP-CGM097

Description of the molecular assembly

The molecular assembly consists of a protein called MDM2-p53 which targets p53 for degradation, and a ligand called NVP-CGM097, which is a drug currently undergoing clinical trials.[180,263,264] The ligand is a dihydroisoquinolinone derivative, semi-buried inside the protein, and occupies the middle of the binding site, reaching the three critical binding pockets of the p53 residues Leu26, Trp23, and Phe19. The binding free energy of the ligand was measured to be -11.8 kcal/mol, using isothermal calorimetry (ITC) experiment, and X-ray crystallography was employed to obtain an experimental structure of the binding pose with a 1.80 Å–resolution.[180]

Computational assays The initial coordinates for the simulation were obtained from PDB entry 4ZYF.[180] The protein, ligand, and water were modeled using the all-atom CHARMM36m force field,[175] CHARMM general force field[48] and the TIP3P water model,[176] respectively. The parameters for the ligand were generated using the CGenFF program v2.4.04,[48]

Table 6.6: 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

Results using HMR and MTS scheme Table 6.7 summarizes the results of our binding free-energy estimations for the MDM2-p53 protein-ligand complex, using different computational schemes. The reference presented in Table 6.7 corresponds to the result of the straightforward application of the geometrical route—i.e., devoid of acceleration—detailed in our previous work,[82] and is compared to the schemes (circled 1–5, 7, 8, 11, 14) described in this study. The simulations were run for 200 ns, and the results were averaged over three replicas. The choice of 200 ns as the simulation length is connected with the computational time needed in the reference standard geometrical route to converge the separation PMF calculations.[82]

In contrast to the Abl-SH3:p41 complex discussed in the main text, schemes 1 and 2 (where MTS acceleration is by a factor of 2 and 4, respectively) provide only marginal acceleration of the separation PMF simulations. Based on these observations, we assume that while the MTS algorithm is a useful one for accelerating standard binding free-energy calculations, its efficiency, however, may depend on the complexity of the protein-ligand complex, including factors such as the size of the system (number of atoms) and the nature of the interactions at play between its moieties. Larger and more complex biological objects, such as the MDM2-p53:NVP-CGM097 complex, require more extensive sampling to reach convergence, and to accurately capture the protein-ligand interactions (200 ns), compared to the smaller Abl-SH3:p41 complex (50 ns), which may somewhat limit the benefit of MTS acceleration.

As shown in Figure 6.6, scheme 3, which only involves the use of HMR, provides the most accurate agreement (-11.3 ± 0.5 kcal/mol) with the reference value (-11.3 ± 0.9 kcal/mol),[82] and falls within $k_B T$ with the experiment (-11.8 kcal/mol).[82] Interestingly, scheme 3 is the only scheme that has reached convergence at the end of the 200-ns simulations (see Figure 6.6B), demonstrating its effectiveness in accurately capturing the protein-ligand interactions. These results suggest that application of HMR works best when the default values are used for the extended-Lagrangian parameters, and is consistent with the objective of HMR to maintain the stability and structural properties of the complex in the course simulations.[243,47]

Additionally, we tested the impact of selected extended-Lagrangian parameters on the accuracy/speed ratio of scheme 4 by applying protocols 7, 8, 11, and 14 (Figure 6.7, Figure 6.8, and Figure 6.9). Our findings indicate that among protocol 7 ($\gamma_\lambda = 5$ ps^{-1}) and 8 ($\gamma_\lambda = 7$ ps^{-1}), protocol 8 offers the best accu-

racy and efficiency for the acceleration scheme. Examination of the convergence of the different tested protocols revealed that protocol 8 exhibited the most homogeneous sampling across the reaction path (Figure 6.8), and reflected the smoothest and fastest convergence (Figure 6.9), rationalizing the observed excellent agreement between the experimental and reference standard binding free energies.

In contrast, reducing the extended fluctuation parameter from the default value of 0.1 to 0.05 Å (protocol 11) could potentially tighten the applied restraint on the center-of-mass distance CV, which, in turn, could affect the behavior of the ligand and the receptor during the simulation, and, thus, impact the accuracy of our standard binding free-energy calculations.[54] Despite the close agreement between the calculated $\Delta G^{\circ}_{\mathrm{b}}$ and the reference values, we recommend caution to the end-users when adjusting this parameter, due to the observed relatively high standard deviation. We also found that protocol 14, with $\tau = 300$ fs, increased efficiency at the expense of accuracy. Its application led to a more inertial fictitious particle, which, in turn, affected the sampling of the free-energy landscape (see Figure 6.8C), and resulted in the underestimation of the calculated free-energy (see Figure 6.7 and Table 6.7).

Table 6.7: 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)	ΔG°_{b} (kcal/mol)	Speed (ns/day)
reference[a]	−17.9[82]	−11.3 ± 0.9[82]	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			
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	−11.6 ± 0.4	75.3
2	−18.3		
3	− 17.8		
⑪			
1	−16.7	−10.5 ± 0.6	74.9
2	−17.7		
3	−16.8		

Continued on next page

Table 6.7 Continued from previous page

Scheme	Separation Contribution (kcal/mol)	ΔG°_{b} (kcal/mol)	Speed (ns/day)
(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

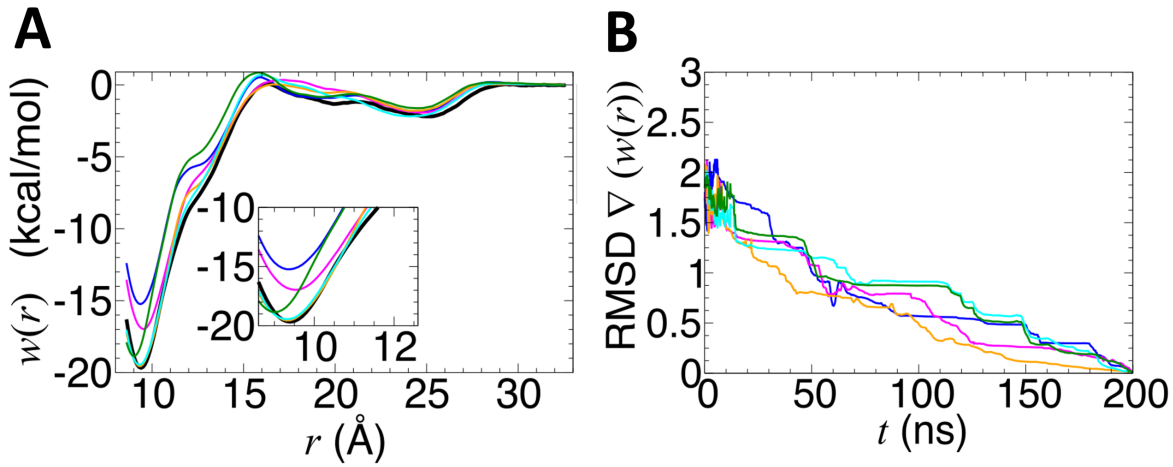

Figure 6.6: (A) Averaged physical separation PMFs for three replicas obtained after individual 200–ns simulations. All the PMFs were determined within the separation distance range of [8.6; 32.6] Å(B) Averaged convergences for the physical separation PMFs. The curves correspond to the different calculation schemes: reference (black),[82] scheme 1 (blue), scheme 2 (magenta), scheme 3 (orange), scheme 4 (cyan), and scheme 5 (green).

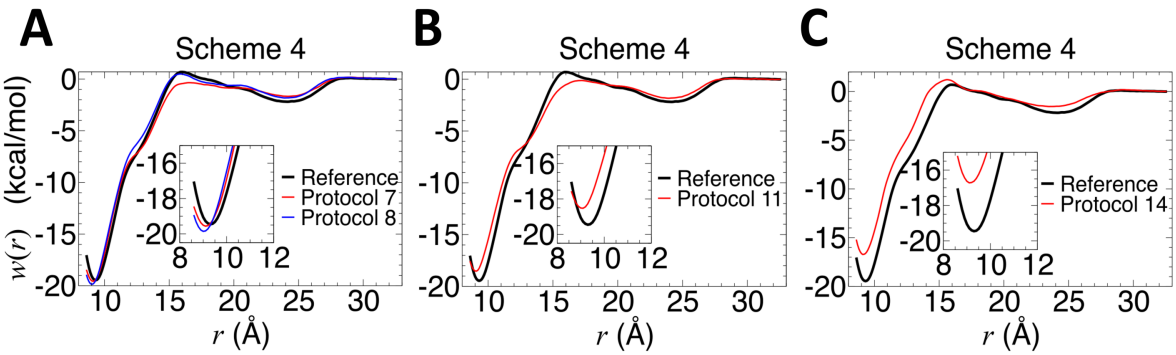

Figure 6.7: Averaged physical separation PMFs for three replicas obtained after individual 200–ns simulations. All the PMFs were determined within the separation distance range of [8.6; 32.6] Å(A) correspond to scheme 4 (black–denoted as Reference), and protocol 7 and 8 with $\gamma_\lambda = 5$ (red) 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).

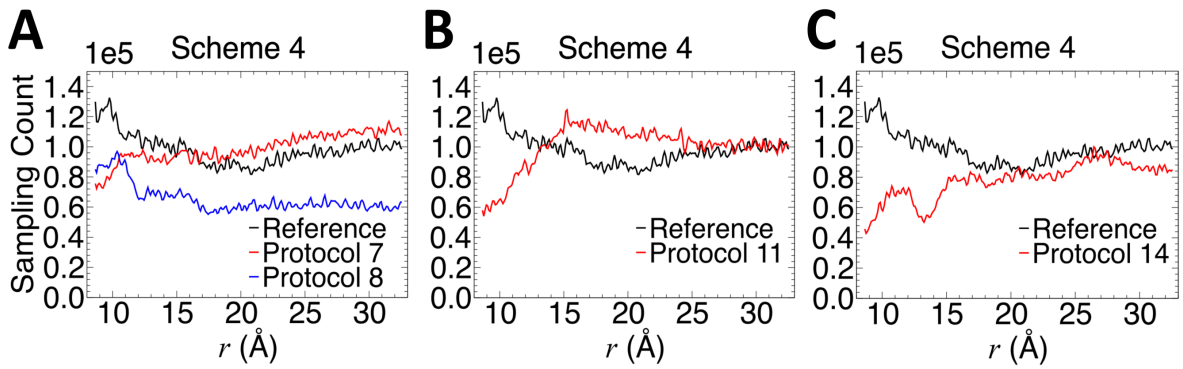

Figure 6.8: Average number of samples per bin achieved for three replicas obtained after individual 200–ns simulations. (A) correspond to scheme 4 (black–denoted as Reference) and protocol 7 and 8 with $\gamma_\lambda = 5$ (red) or 7 (blue) ps^{-1}, respectively. (B) corresponds to scheme 4 (black–denoted 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).

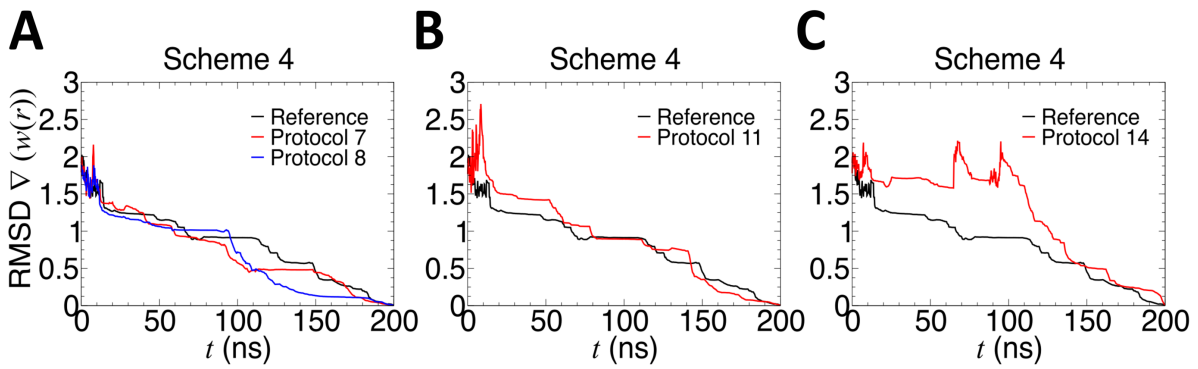

Figure 6.9: Average convergence rates achieved for three replicas obtained after individual 200–ns simulations. (A) correspond to scheme 4 (black–denoted as Reference) and protocol 7 and 8 with $\gamma_\lambda = 5$ (red) or 7 (blue) ps^{-1}, respectively. (B) corresponds to scheme 4 (black–denoted as 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).

Chapter 7

Protein-protein classification using the Dpr-DIP interactome

Sommaire

This chapter recaps the study of the molecular basis of the interaction between Dpr and DIP proteins in the context of synaptogenesis using the geometrical route and ML approaches. This works has been submitted at the JCIM journal under the following :

"**Goulard Coderc de Lacam, E.**, Roux, B. and Chipot, C. Classifying protein-protein binding affinity with free-energy calculations and machine-learning approaches, *J. Chem. Inf. Model*, 2023 "

and is currently under review. I will provide a summary of the publication before providing the full-lenght text.

7.1 Summary

The Dpr-DIP interactome has been identified to play an important role in synaptogenesis in *Drosophila Melanogaster*. However, experimental data suggest that merely a limited subset of complexes, essentially 57 out of a total of 231 heterodimers, exhibit strong binding affinity. In this work, we sought to identify the residue-level molecular basis underlying the difference in binding affinity using a state-of-the-art methodology consisting of standard binding free-energy calculations with the geometrical route and machine learning (ML) techniques. We determined the binding affinity for two complexes using statistical mechanics simulations, achieving an excellent reproduction of the experimental data. Moreover, we predicted the binding free energy for two additional low-affinity complexes, devoid of experimental precise estimation while simultaneously identifying key residues for the binding (see Table 7.1 below).

Complex	ΔG_b° (kcal/mol)	ΔG_b°, exp (kcal/mol)	Simulation time (μs)
Dpr6-DIPα	-6.7 ± 0.1	-7.7[8]	2.1
Dpr6-DIPβ	-5.4 ± 0.4	-6.5[8]	2.3
Dpr6-DIPγ	-4.5 ± 0.4	< -4.1[8]	2.2
Dpr6-DIPθ	-2.3 ± 0.6	< -4.5[8]	1.5

Table 7.1: Binding free energy estimates using the geometrical route[57]

We discovered a sensitivity of interfacial side-chains to isomerization in the presence of water, thus causing deterioration of the binding estimate and requiring additional customs restraints for every complex based on visual inspection of interfaces. This specific tuning required to perform these calculations cannot be easily implemented for all Dpr-DIP complexes within a reasonable time and bereft of human intervention.Table 7.1 shows that in an ideal world, MD could be able to discriminate the complexes, but it is computationally prohibiting to do all of the complexes now.

Leveraging the documented affinities and a pre-establish classification inferred from SPR assays,[8] we turned next to supervised ML techniques to gain additional insight into the distinction between the two types of complexes. We designed specifiic input features, leveraging a multiple sequence alignment (sequence-based approach) with a structure-based distance and a pre-established pairwise interaction potential (physicochemical compatibility).

LDA and RF achieved remarkable accuracy, as high as 0.99, in discerning between strong (cognate) and weak (non-cognate) binders using diverse energetic potentials (see Table 7.2)

Potential	LDA		RF	
	training	validation	training	validation
Elementary[11]	0.8667	0.7707	0.9814	0.9252
Dill[265]	0.9241	0.8283	0.9923	0.9302
Dima[266]	0.9109	0.7849	0.9925	0.9339
Dosztányi[267]	0.8901	0.7471	0.9924	0.9311
Betancourt[268]	0.9164	0.8127	0.9922	0.9251

Table 7.2: Accuracy results using different pairs amino acids potentials for LDA and RF for *Drosophila Melanogaster*

The presented ML approach encompasses easily transferable input features, enabling its broad application to any interactome, while facilitating the identification of pivotal residues critical for binding interactions. The predictive power of the generated model was probed on similar protein families from thirteen diverse *Drosophila* species as well as for homodimers and human homologous, IgLONs.

Table 7.3: Accuracy results for trained LDA and RF using the different species and complexes

Species	complexes	LDA accuracy	RF accuracy
Drososphila Busckii	Dpr-DIP	0.92	1.0
Drososphila Sechellia	Dpr-DIP	0.95	0.98
Drososphila Persimillis	Dpr-DIP	0.99	1.0
Drososphila Simulans	Dpr-DIP	0.98	0.99
Drososphila Rhopaloa	Dpr-DIP	0.94	0.99
Drososphila Guanche	Dpr-DIP	0.96	0.99
Drososphila Pseudoobscura Pseudoobscura	Dpr-DIP	0.92	0.995
Drososphila Ananassae	Dpr-DIP	0.92	0.99
Drososphila Kikkawai	Dpr-DIP	0.93	0.98
Drososphila Virilis	Dpr-DIP	0.94	1.0
Drososphila Grimshawi	Dpr-DIP	0.95	0.96
Drososphila Willistoni	Dpr-DIP	0.96	0.99
Drososphila Mojavensis	Dpr-DIP	0.93	0.98
Human	IgLONS	1.0	0.0
Drosophila Melanogaster	DIP-DIP, Dpr-Dpr	0.90	0.54

Our ML model exhibited commendable performance on these additional datasets, showcasing its reliability and robustness across the species barrier, even if the RF does not performed well on homodimers and IgLONs. The binding affinities of the homodimers are also weaker than those characteristic of the heterodimers formed between the proper partners, which can explain the difficulty of RF to accurately identify the strong binders amongst homodimers. However, when aiming at identifying the key residues for binding in a particular dataset, with no interest in transferability, RF outperforms LDA by learning better the specific patterns and leveraging the RF ensemble properties to obtain the relative residue importance in the classification.

7.2 Original article

7.2.1 Introduction

The brain comprises a vast array of neurons linked by a complex synaptic connection network. Understanding how neurites distinguish synaptic partners in densely packed neuronal tissue remains a central question in neurobiology.[269,270] Roger Sperry's chemoaffinity hypothesis proposed in the 1960s stated that neurons make specific connections based on the affinity of cell-surface labels.[4] This concept was further refined into the idea that cell adhesion-like cell-surface proteins expressed on neuron surfaces bind to each other and trigger downstream events leading to synapse formation between appropriate partners.[271] The realm of cell recognition encompasses cell adhesion molecules known as cell-surface proteins (CSPs), which frequently arise from gene duplication events occurring throughout evolutionary processes. Such duplications give rise to multiple members within the CSP family, with each protein exhibiting a characteristic canonical binding interface indicative of the particular family.[8] In this case, the binding strength is solely responsible for the different functions of each protein.[6,7,8] In the fruit fly, the Defective proboscis extension response (Dpr)-Dpr Interacting Proteins(DIP) interactome has been identified by Ozkan et al.[5] to play that role. Out of the 231 heterodimers formed by the 11 DIP and 21 Dprs, respectively, only 57 displayed high affinity by surface plasmon resonance (SPR)[272] assay,[8] assuming that the strong binding correspond to cognate complexes leaving the remaining 174 complexes

as non-cognate (weak binders). Due to the high sequence similarity of these complexes, special efforts are required to understand the mechanism underlying the specific association of the binding partners.

The availability of comprehensive structural information regarding our protein–protein complexes has opened avenues for the use of empirical structure-based scoring methods that leverage force fields and thermodynamics properties.[273,17,274] In the MM/PBSA approximation, the binding affinity is computed as the sum of its gas-phase energy (MM), the solvation free energy using Poisson-Boltzmann continuum electrostatics and the solvent-exposed surface area (PBSA), and a contribution due to the configurational entropy of the solute extracted from molecular dynamics (MD) simulations.[13,14,15,16,17] The MM/PBSA was tested for the Dpr-DIP interaction, resulting in a maximum distinguishability of 30 % between cognate and non-cognate complexes.[11] However, the numerous underlying approximations (implicit solvation and constant dielectric medium),[19,20,57] as well as the poor resulting discrimination obtained, call into question its general applicability for protein–protein complexes. Since implicit solvation approaches cannot provide reliable estimates for the Dpr-DIP interactome, the idea is to use an explicit solvent model with enhanced sampling to favor rare events such as binding or unbinding. Steered MD involves one protein being pulled away from the other.[131] This method allows the recovery of the binding free energy through the Jarzynski equation at the hefty cost of multiple realizations in a near-equilibrium regime,[21,22] under the questionable assumption that binding depends solely on the physical separation.[68] The utilization of this strategy necessitates a significant allocation of computational resources, rendering it incompatible with the processing of large datasets. The geometrical route,[57] conceptualized by Woo and Roux[44] and generalized for protein–protein complexes by Gumbart et al.,[57] offers a framework with potential-of-mean force (PMF) calculations, in which the slow degrees of freedom of the reversible association are gradually restrained to enhance convergence and reduce the computational cost while preserving estimates reaching chemical accuracy.

Alternative approaches using machine learning (ML) to predict, or classify binding partners at a lower computational cost have emerged in the last decade, and are mostly directed at protein–ligand complexes due to the interest of the pharmaceutical industry in the drug design field.[29,30,31,32,33,34] Only a few applications have been dedicated to predicting the binding of protein–protein complexes.[275,276,42,11] The methodologies employed to address this challenge can be broadly categorized into two distinct categories, leveraging either sequence or structural characteristics.[41] In the case of sequence-based approaches, features rely on evolutionarily conserved residues, hypothesized to be essential for the function, and, therefore, for binding.[277] The various types of input features include (i) residue encoding, (ii) evolutionary information, (iii) residue physicochemical properties, and (iv) predicted structural features.[41] Residue encoding, using traditional techniques such as the one-hot representation (vector of length twenty containing zeros representing the twenty amino acids. The position in that vector representing the encoded residue will have a value of one to differentiate from the surrounding zeros), is insufficient to predict the protein interactions efficiently.[41] Consequently, evolutionary descriptors, like mutual information,[277] are preferred. Inasmuch as structural features are concerned, they usually consist of geometrical descriptors extracted from the structure, such as the surface accessible to the solvent, the curvature of a set of atoms, or a set of invariant geometrical fingerprints.[41,72] An inherent limitation of structure-based approaches is the quality and availability of reliable structures linked to the difficulties to solve structures either experimentally or theoretically.

In this work, we aimed at predicting amid a series of homologous protein–protein complexes which ones are cognate, and at identifying the molecular basis of their formation at the residue level. Towards this end, we first used the geometrical route[57] as a precise method based on first principles, and applied to a few complexes. Then, we turned to ML schemes using sequence-based features to treat the whole Dpr-DIP dataset and identify the key residues responsible for differentiating cognate and non-cognate complexes.

7.2.2 Methods

The following subsections briefly recap the methodology employed in this work, its theoretical underpinnings, and describe the protocols of the calculations reported herein.

Binding Free-Energy Calculations

The formation of a protein–protein complex involves significant conformational changes, hardening the ergodic sampling of configurational space within the simulation time amenable to MD. In the realm of computational techniques for protein binding, importance sampling algorithms[62] emerge as powerful tools. These algorithms operate by introducing external forces onto collective variables (CVs) to accelerate the sampling of rare events, such as protein binding. These CVs are essential degrees of freedom involved in the reversible association, and can be controlled and monitored during the course of the simulation.[93] One of the straightforward CVs to control the reversible binding of two proteins is the Euclidean distance between their centers of mass (COMs). However, it does not prevent random tumbling of the two molecular objects at play, nor conformational changes upon association, slowing down the convergence of the separation potential of mean force (PMF) calculation.[62,224] To alleviate this shortcoming, a set of restraints using additional CVs, representing the minimum information to define one partner with respect to the other one, are applied in the course of the separation as prescribed in the geometrical route introduced by Gumbart et al.[57] These CVs are the backbone distance root-mean-square deviations (RMSDs) of the two proteins with respect to the reference, native conformation—i.e., in the bound state, the three Euler angles describing their relative orientation, and two additional angles (polar and azimuth) for their relative position. Applying geometrical restraints in the form of harmonic potentials onto these CVs results in an effective loss in configurational entropy, corresponding to conformational (ΔG_c^{site}), orientational (ΔG_o^{site}), and positional (ΔG_a^{site}) free-energy contributions, which must be accounted for in the computation of the standard binding free energy, both in the "bulk" (unbound state) and "site" (bound state). Therefore, the geometrical route involves a series of independent PMF calculations determined sequentially with the progressive introduction of restraints, prefacing the separation PMF calculation restrained along the specific axis "a". The binding free energy can then be expressed as a sum of these different free-energy contributions, namely,

$$\Delta G_{\text{b}}^{\circ} = \Delta G_c^{\text{bulk}} - \Delta G_c^{\text{site}} - \Delta G_o^{\text{site}} - \Delta G_a^{\text{site}}(\theta,\phi) + \Delta G_o^{\text{bulk}} - \frac{1}{\beta}\ln(S^* I^* C^{\circ}) \tag{7.1}$$

where $\beta = (k_{\text{B}}T)^{-1}$, with k_{B}, the Boltzmann constant, and T, the temperature. C° represents the standard concentration of 1 M, which corresponds to 1/1661 (Å^3).[89] I^* is the separation term, and S^* is a surface term, which represents the fraction of a sphere of radius r^*, centered at the binding site of the reference protein, accessible to its partner, that is,

$$\begin{cases} I^* &= \displaystyle\int_{\text{site}} \mathrm{d}r\ \mathrm{e}^{-\beta\left(w(r)-w(r^*)\right)} \\ S^* &= r^{*2}\displaystyle\int_0^{\pi}\mathrm{d}\theta\ \sin\theta\int_0^{2\pi}\mathrm{d}\phi\ \mathrm{e}^{-\beta u_a} \end{cases} \tag{7.2}$$

Here, r^* is a point located far away from the binding site, where the proteins no longer interact with each other, and u_a is the sum of the harmonic restraint potentials imposed on polar angles θ and ϕ to restrain the separation along the axis a.

In some cases, additional RMSD restraints acting on protein–protein interfacial side chains are required due to their exposure to the solvent and the possibility of their local conformational change in the

course of the separation,[57] For the four selected complexes, specific side chains from the residues displayed in Table 7.4 were restrained based on visual inspection of the starting structures. The numbering of residues defined from the position in multiple sequence alignment (MSA) to facilitate direct comparison between complexes. ARG150 is, however, present in a region that lacks conservation between Dprs, hence, it is not assigned a number within the MSA.

Complexes	Dpr restrained residues	DIP restrained residues
Dpr6-DIPα	ARG150 (not in MSA),30,36	8, 9, 42, 44
Dpr6-DIPβ	10, 11, 22	9, 42, 44
Dpr6-DIPγ	20, 21, 32, 34	42
Dpr6-DIPθ	7, 10, 11, 20, 21, 30, 32	9, 40, 42

Table 7.4: Side chains restrained residue numbers during the geometrical route using their position in the MSA.

Computational assays

All simulations were performed using the NAMD3 program.[86] The all-atoms macromolecular CHARMM36[84] force field and the TIP3P model[85] were used to describe the proteins and the water, respectively. All computational assays corresponded to a physiological concentration of NaCl of 0.15 M. The temperature (300 K) and the pressure (1 atm) were kept constant employing a Langevin thermostat[156] and the Langevin piston algorithm,[157] respectively. Long-range electrostatic interactions were handled by the particle-mesh Ewald (PME) algorithm.[158] Van der Waals and short-range electrostatic interactions were truncated with a smoothed 12-Å spherical cutoff. The equations of motion were integrated every 2 fs.

Theoretical models of each interacting moiety, namely their first immunoglobulin domain, were generated using homology modeling and simulated for 200 ns at equilibrium.[11] The starting structures for the binding free energy calculations were the resulting equilibrated complexes.[11] They were piped into the binding free-energy estimator 2 (BFEE2),[81] a tool designed to help set up binding free-energy calculations for protein–ligand complexes,[82] and expanded to protein–protein complexes by including RMSD calculations of the backbone of each protein, both in the bulk and at the binding site. The well-tempered extended adaptive biasing force algorithm (WTM-eABF)[49] as implemented in the collective variable module (Colvars) of NAMD[93] was chosen as the importance-sampling algorithm. The PMFs were run sequentially for all complexes, starting from the distance RMSD of the proteins with respect to the native conformation up to the physical separation of the two proteins. Once all the PMF calculations were completed, BFEE2[81] was invoked again at the post-processing stage, to extract the individual contributions to the binding affinity from each PMF, and to infer the final binding free-energy estimate.

Machine Learning strategies

Input features generation

One of the most demanding and pivotal aspects of ML lies in the generation of the input features. These features serve to describe our protein–protein complex interacting regions (interfaces) for every complex in a suitable format for ML algorithms[41] We chose to rely on evolutionary features, better at describing the binding interface at the residue level.[41] The mutual information (MI) was obtained based on a multi-sequence alignment of the first immunoglobulin domain for each protein family[11] performed with CLUSTAL-W.[73] The MI is defined as the difference between individual entropy at the residue position in the MSA, H_i and H_j, and the joint entropy H_{ij}.

$$\mathrm{MI} = -H_{ij} + H_i + H_j = \sum_{x,y} p_{ij}(x,y) \log_2 \frac{p_{ij}(x,y)}{p_i(x) p_j(y)} \tag{7.3}$$

where $p_i(x)$ and $p_j(y)$ stand for the occurrence probability of a given amino acid (x or y) at position i or j, $p_{ij}(x,y)$ is the joint probability of the amino acids at the positions i and j in the multiple sequence alignment. This metric is utilized to estimate how much information is gained by accounting for the covariance of residues of DIP and Dpr rather than by treating them separately. The MI between non-cognate pairs is treated as noise, since they are not supposed to have co-evolved, and are, therefore, generated by simple combinatorial. Consequently, the MI generated by non-cognate pairs is subtracted from the one obtained for cognate pairs. Only residue pairs with a non-zero MI value were kept for the input features. To account solely for interacting residues at the interface, the MI is then weighted by an inverse distance, either from $\alpha-$carbon atoms extracted from the crystal structure of the Dpr10-DIPα[74] since every complex shares the same fold, or from the inverse distance between the COM of the side chains extracted from MD equilibrium trajectories taken from reference.[11] To consider the energetics associated with residue interactions, a pairwise amino-acid (AA) potential is added to the MI and distance for every residue pair of the complex. The sum of these features for all residues forms a score for the complex.

$$\mathrm{score} = \sum_{\mathrm{AApairs}} (\mathrm{MI} \times \mathrm{distance} \times \mathrm{potential}) \tag{7.4}$$

ML algorithms and parameters

To distinguish between cognate and non-cognate complexes, two different ML algorithms were selected for their simplicity and their interpretability properties, namely linear discriminant analysis (LDA)[69] and random forest (RF).[35] The scikit-learn library[78] and the specific functions *LinearDiscriminantAnalysis* and *RandomForestClassifier* in Python 3.8.5[278]were employed. The selection criteria selected for RF was the entropy with no limit on the entropy value or tree length to limit tree growth. The training test consisted of 80% of the full dataset picked randomly, leaving the remaining 20% as the validation set, and was repeated a thousand times to ascertain the robustness of our results.

Extraction of sequences from the database

To test out the trained models on similar proteins from different *Drosophila* species, we searched in the Uniprot database,[279] for each DIP and Dpr, the most similar proteins based on their amino-acids sequences, using BLAST (basic local alignment searching tool) version 2.9.0+.[280] Each recovered sequence was assigned to a specific DIP and Dpr based on similarity scores, the maximum being 100%, and the minimum 47.6%. In case of a conflict of assignment between two recovered sequences, the choice was made towards the one with the highest identity score. To define this score, BLAST was used in combination with the BLOSSUM62 substitution matrix,[281] recommended for queries longer than 85 amino acids with a threshold of 10. Not all the DIPs and Dprs were recovered in the different selected *Drosophila* species, and the names of the missing proteins are reported in Table 7.5.

Table 7.5: Summary of the Blast search reporting the name of proteins for which the sequence was not recovered in Uniprot[279] for all 14 *Drosophila* species considered in this study

Species (*Drosophila...*)	**Missing DIPs**	**Missing Dprs**	**Cognate complexes**	**Non − cognate complexes**
... Melanogaster	-	-	57	174
... Busckii	κ	19, 21	51	139
... Sechellia	β, λ	7	42	138
... Persimillis	β, ι, λ, κ	7,8,12,21	34	85
... Simulans	β, δ, λ	5,7,12,21	37	99
... Rhopaloa	β, η	8,11,21	42	120
... Guanche	β, λ	-	46	143
... Pseudoobscura Pseudoobscura	-	-	57	174
... Ananassae	-	-	57	174
... Kikkawai	λ	-	53	157
... Virilis	β	3	47	153
... Grimshawi	β, λ	-	46	143
... Willistoni	α	8,12,13,21	49	121
... Mojavensis	-	3	54	166

7.2.3 Results and Discussion

Binding Free-Energy Calculations

The binding free-energy estimates, $\Delta G_{\mathrm{b}}^{\circ}$, of four Dpr-DIP complexes, were determined using the geometrical route,[57] and are reported in Table 7.6. Dpr6-DIPβ and Dpr6-DIPα are both cognate complexes, for which the exact binding free energies are known experimentally.[8] Conversely, Dpr-6DIPθ and Dpr6-DIPγ are both non-cognate complexes, for which only a rough estimation of their binding affinity is available.

Complex	$\Delta G_{\mathrm{b}}^{\circ}$ (kcal/mol)	$\Delta G_{\mathrm{b}}^{\circ}$, exp (kcal/mol)	Simulation time (μs)
Dpr6-DIPα	-6.7 ± 0.1	-7.7[8]	2.1
Dpr6-DIPβ	-5.4 ± 0.4	-6.5[8]	2.3
Dpr6-DIPγ	-4.5 ± 0.4	< -4.1[8]	2.2
Dpr6-DIPθ	-2.3 ± 0.6	< -4.5[8]	1.5

Table 7.6: Binding free energy estimates using the geometrical route,

During the physical separation simulation, a progressive decay of the PMF in the shape of successive plateaus was observed, mirroring a loss of interaction at the interface, most likely due to side-chain isomerization in response to solvent exposure. As stated in the Methods section, introducing additional restraints on the RMSD of those interacting side chains was sufficient to correct this artificial behavior, and recover the experimental value.[8] The same procedure was applied to the non-cognate complexes, and, as anticipated, binding free energies lower than those of the cognate ones were obtained.

In the SPR assays,[8] complexes with a dissociation constant ($\mathrm{K_d}$) above 200 μM, which corresponds to binding affinities lower than -5.1 kcal/mol, were considered as non-cognate. Nevertheless, Cosmanescu et al. provided an estimation of the binding affinity, namely lower than -4.1 and -4.5 for Dpr6-DIPγ and Dpr6-DIPθ, respectively. Our $\Delta G_{\mathrm{b}}^{\circ}$ for Dpr6-DIPγ is slightly above the experimental threshold, while remaining below the limit separating cognate from non-cognate complexes (-5.1 kcal/mol). Dpr6-DIPθ computed binding affinity, on the other hand, is under the expected value of -4.5 kcal/mol. The lesser accuracy of the dissociation constant measurements due to less sensitive sensor chips[8] precludes a more precise validation of our predictions for the non-cognate complexes.

Comparing the different physical separation PMFs (see Figure 7.1), the depth of the wells reveals a major difference between the cognate and non-cognate complexes. The non-cognate complexes are close to -10 kcal/mol, and the cognates are in the -24 to -30 kcal/mol region, demonstrating a stark contrast in their separation behavior, which could help roughly categorize the complexes without the necessity to

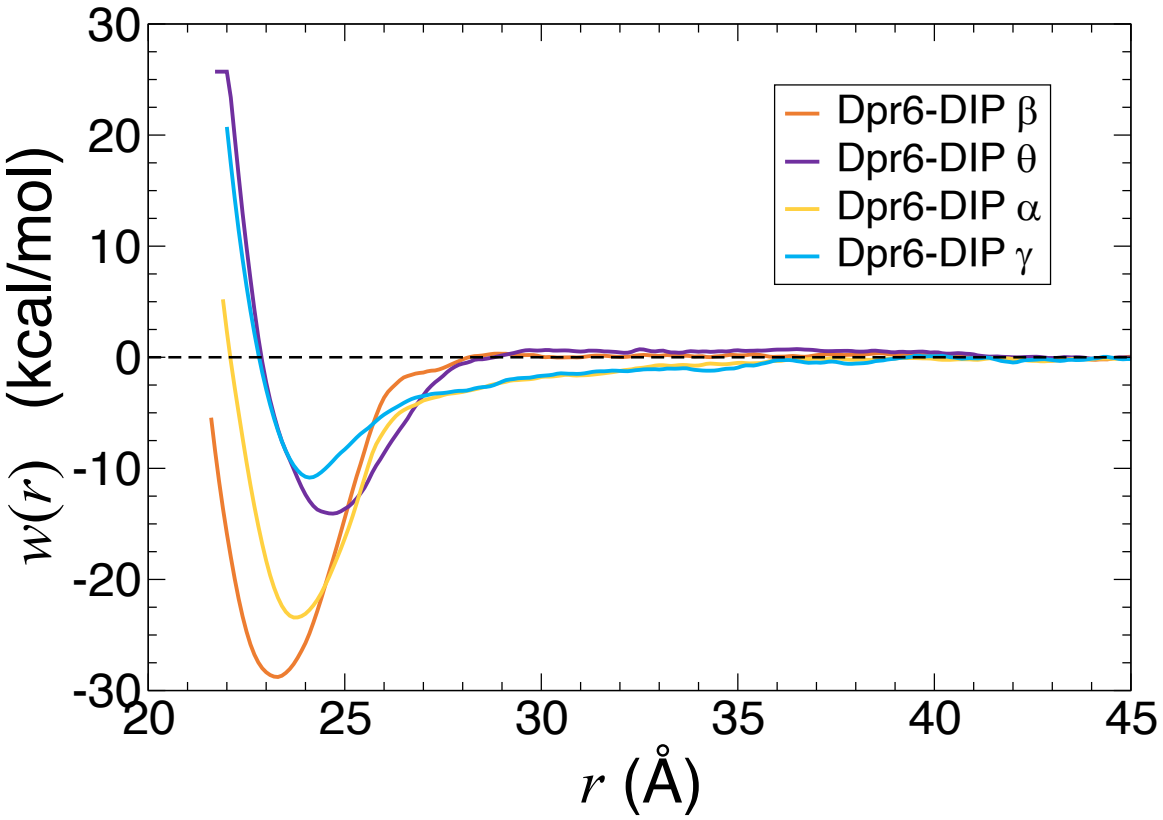

Figure 7.1: Separation PMFs of all Dpr6 complexes examined with the geometrical route. r stands for the Euclidean distance between the protein COMs.

follow the entire geometrical route.

Looking into the protein–protein interactions of simulated complexes, the hydrogen-bond occupancies (see Figure 7.2) for residue pairs detected by VMD[90] were examined. A clear difference arises between cognate (α, β) and non-cognate (γ, θ) complexes for the 10-40 and 30-9 pairs (using the MSA Dpr-DIP residue positions). Furthermore, RMSD side-chain restraints are present for β and θ in pair 10-40 for α and θ in pair 30-9, and yet a clear difference with a lower occupancy in the case of the non-cognate complexes is observed in the separation simulations. This result confirms that cognate complexes form more stable interactions than non-cognate ones. The close binding affinity estimates for β (cognate) and γ (non-cognate) might stem from similar high occupancies for several pairs (20-0, 34-19). Overall, the remarkable accord observed between the experimental and the predicted binding affinities obtained using the geometrical route offers a compelling evidence for the reliability and robustness of this methodology in effectively classifying complexes and providing a more precise binding affinity than the experimental assay. Nevertheless, the specific tuning required to perform these calculations cannot be easily implemented for all Dpr-DIP complexes within a reasonable time and bereft of human intervention.Table 7.6 shows that in an ideal world, MD could be able to discriminate the complexes, but it is computationally prohibiting to do all of the complexes now. Leveraging the documented affinities and a pre-establish classification inferred from SPR assays,[8] we turned next to supervised ML techniques to gain additional insight into the distinction between the two types of complexes.

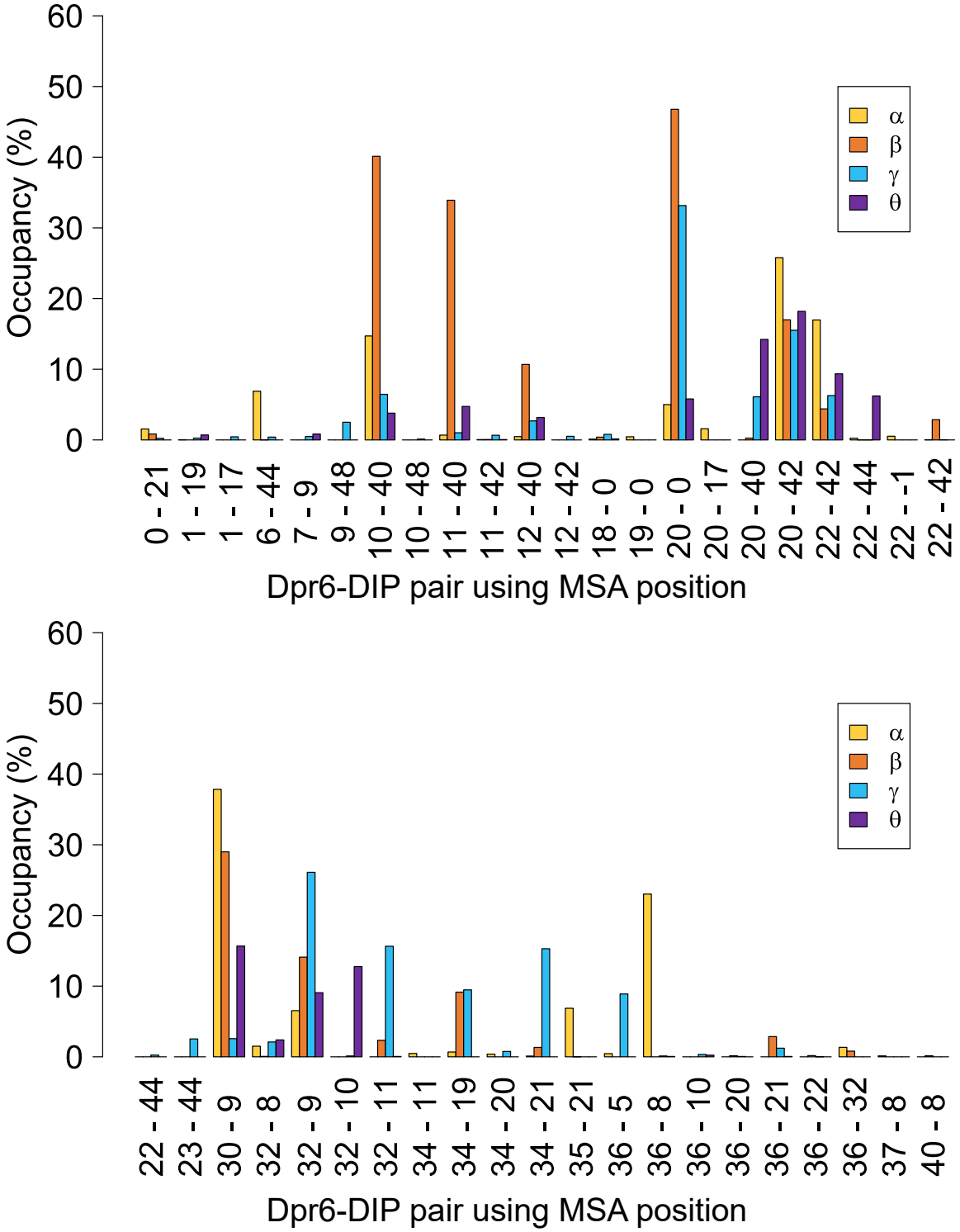

Figure 7.2: Hydrogen-bonds occupancies obtained during the separation trajectories of the geometrical route when the proteins are in close contact (in the 4 Å interval of the distance corresponding to their PMF minima).

Machine Learning Strategies

Nandigrami et al.[11] established that an LDA-weighted AIMS algorithm was able to discriminate between cognate and non-cognate partners in a set of 231 complexes, achieving an accuracy of approximately 0.88. In their work, they applied a pairwise amino-acids elementary potential, assigning a value of $+2$ to strong interactions (e.g., R and D) while highly repulsive interactions (e.g., K and K), were given a -2 value. However, this initial approach has certain limitations. Specifically, it disregards the contribution of several amino acids (e.g., alanine, threonine, serine, and glycine) by attributing them a value of 0, thereby treating them as non-interacting residues. Due to the inherent simplifications in this elementary pairwise potential, we sought a more accurate scoring metric to describe the interaction of residue pairs. In particular, we considered four different potentials obtained with different approaches, all presented in the Supporting Information (SI). Dill and Thomas proposed a pairwise potential for amino acids, which was constructed iteratively until it could effectively distinguish between native and decoy conformations.[265] Following the same aim, Dima et al.[266] proposed a pairwise amino-acids potential using a coarse-grained representation of the amino acids, therefore reducing their complexity while still capturing the essential interactions. Similarly, Dosztányi et al.[267] developed their pairwise interaction potential by employing a quadratic form of the amino acids composition. This potential was derived through statistical analysis of a comprehensive database of globular proteins, allowing residue-residue interactions that contribute to the stability and folding of proteins to be identified. Furthermore, Betancourt et al. conducted all-atom simulations in a water environment to extract the interaction potential of all amino acids using radial density functions, which were then utilized to build a coarse-grained model.[268]

Exchanging the elementary potential with those based either on calculations or experiments[268,266,267,265] merely improved the accuracy by 0.04 at best, as reported in Table 7.7. Interestingly, the exclusion of the four amino acids discussed previously did not adversely affect the final classification, hence indicating that their contribution to the binding distinguishability was minimal and suggesting that modification of the pairwise potential was not the crucial factor in enhancing the performance of the ML algorithm. However, when using a neutral value of 1 as the interacting potential for all amino acids as a control, the accuracy dropped to 0.5 when using LDA, emphasizing the importance of incorporating an interaction potential in input features for the classification algorithm. Besides, this observation underscores that the elementary potential, despite its simplicity, contains sufficient information for an accurate classification. Since the affinity between proteins is non-linear, we replaced LDA with a non-linear classification algorithm, namely RF (see Methods section for details). The RF model exhibited remarkable performance by achieving an accuracy of up to 0.98 when utilizing the elementary potential, surpassing the LDA's accuracy by 0.1. Furthermore, testing each potential using the RF model revealed comparable accuracies across different potential conditions, reaffirming our earlier assertion regarding the limited impact of distinct potentials on classification accuracy.

Potential	LDA		RF	
	training	validation	training	validation
Elementary[11]	0.8667	0.7707	0.9814	0.9252
Dill[265]	0.9241	0.8283	0.9923	0.9302
Dima[266]	0.9109	0.7849	0.9925	0.9339
Dosztányi[267]	0.8901	0.7471	0.9924	0.9311
Betancourt[268]	0.9164	0.8127	0.9922	0.9251

Table 7.7: Accuracy results using different pairs amino acids potentials for LDA and RF

A thorough examination of the feature importance mapped back to specific residues in the RF model provided insights into the key residues contributing to distinguishing between cognate and non-cognate protein–protein complexes (Figure 7.3 B and C). However, upon further investigations of the interactions mediated by these residues using the available MD equilibrium trajectories for randomly selected complexes, no direct distinction could be established between cognate and non-cognate complexes.

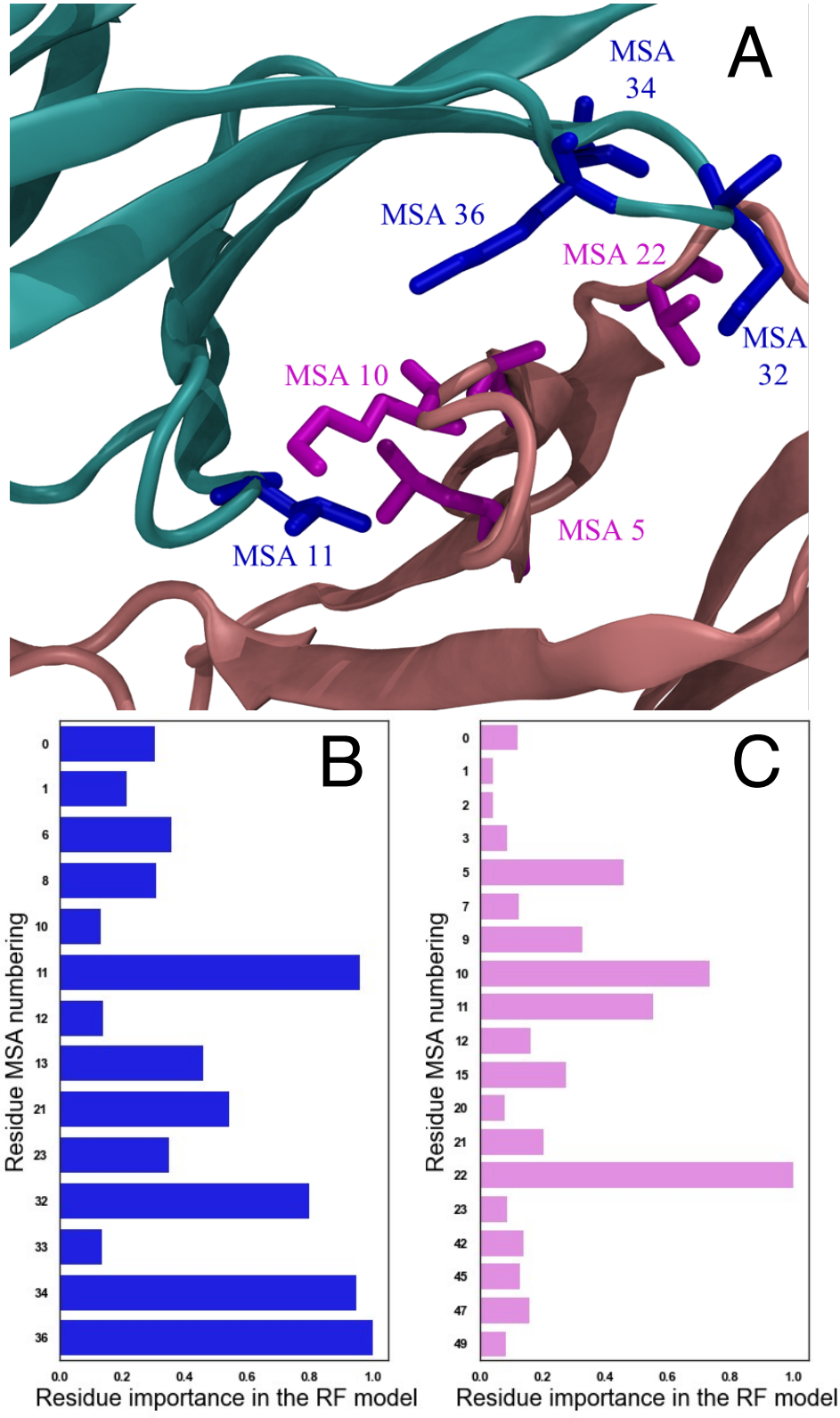

Figure 7.3: A) Interface pattern of Dpr1-DIPα with MSA important residue B) Dpr per-residue importance according to the RF model and MSA position C) DIP per-residue importance according to the RF model with backbone distance and MSA position.

In addition to the potential, another parameter that can be fine-tuned to improve the description of the complexes at the residue level is the distance embedded in the input features. To reflect the actual interacting distance more accurately in the score than with a simple $\alpha-$carbon atom distance between pairs identical for all complexes, we took advantage of the equilibrium MD simulations. Specifically, we extracted the distances separating the side-chains COMS from the MD trajectories and calculated an average distance over the last 100 ns of the simulation. The updated ML results are gathered in Table 7.8 below.

Potential	LDA		RF	
	training	validation	training	validation
Elementary[11]	0.8230	0.6956	0.9998	0.8364
Dill[265]	0.9333	0.8082	0.9999	0.8650
Dima[266]	0.9267	0.7857	0.9999	0.8849
Dosztányi[267]	0.8886	0.7493	0.9999	0.8627
Betancourt[268]	0.8805	0.7200	0.9999	0.8390

Table 7.8: Accuracy results for LDA or RF with side chains COM mean distance

Incorporating the COM mean distance instead of that of $\alpha-$carbon atoms, lead to similar results for the training set when using both LDA and RF. However, a slight decrease in performance was observed for RF, with a validation set loss of approximately 0.1, indicating a possible overfitting problem, meaning that the model has overlearned some patterns from the training that are not transferable to new data and can cause significant generalization problems. Surprisingly, the COM mean distance from MD simulations did not provide any significant information to distinguish between cognate and non-cognate complexes. Moreover, the identification of important residues was found to be identical to that obtained using the $\alpha-$carbon atom distance, neither showing gain of additional structural information nor contributing to a better understanding of the distinctive characteristics between cognate and non-cognate complexes.

Potential	LDA		RF	
	training	validation	training	validation
Elementary[11]	0.9888	0.7596	1.0	0.8364
Dill[265]	0.9751	0.7811	1.0	0.873
Dima[266]	0.9682	0.7873	0.9999	0.8753
Dosztányi[267]	0.9677	0.7921	0.9999	0.8747
Betancourt[268]	0.9626	0.7450	0.9999	0.8671

Table 7.9: Accuracy results for LDA or RF with side chains COM mean distance and all pairs at the interface

In an effort to obtain a more comprehensive description of the binding interface and identify critical residues, we extended our ML approach to include all residue pairs rather than focusing on MI preselected residue pairs. This broader strategy resulted in an improved accuracy for the LDA model, with values similar to those for the RF model with $\alpha-$carbon distance. For the RF model, the results are similar to the previous RF with MD distance showing no improvement. LDA, with a more detailed and tailored description of the complexes, was able to perform on par with RF, incorporating less information, emphasizing the importance of feature engineering when developing ML models for the classification of protein–protein complexes.

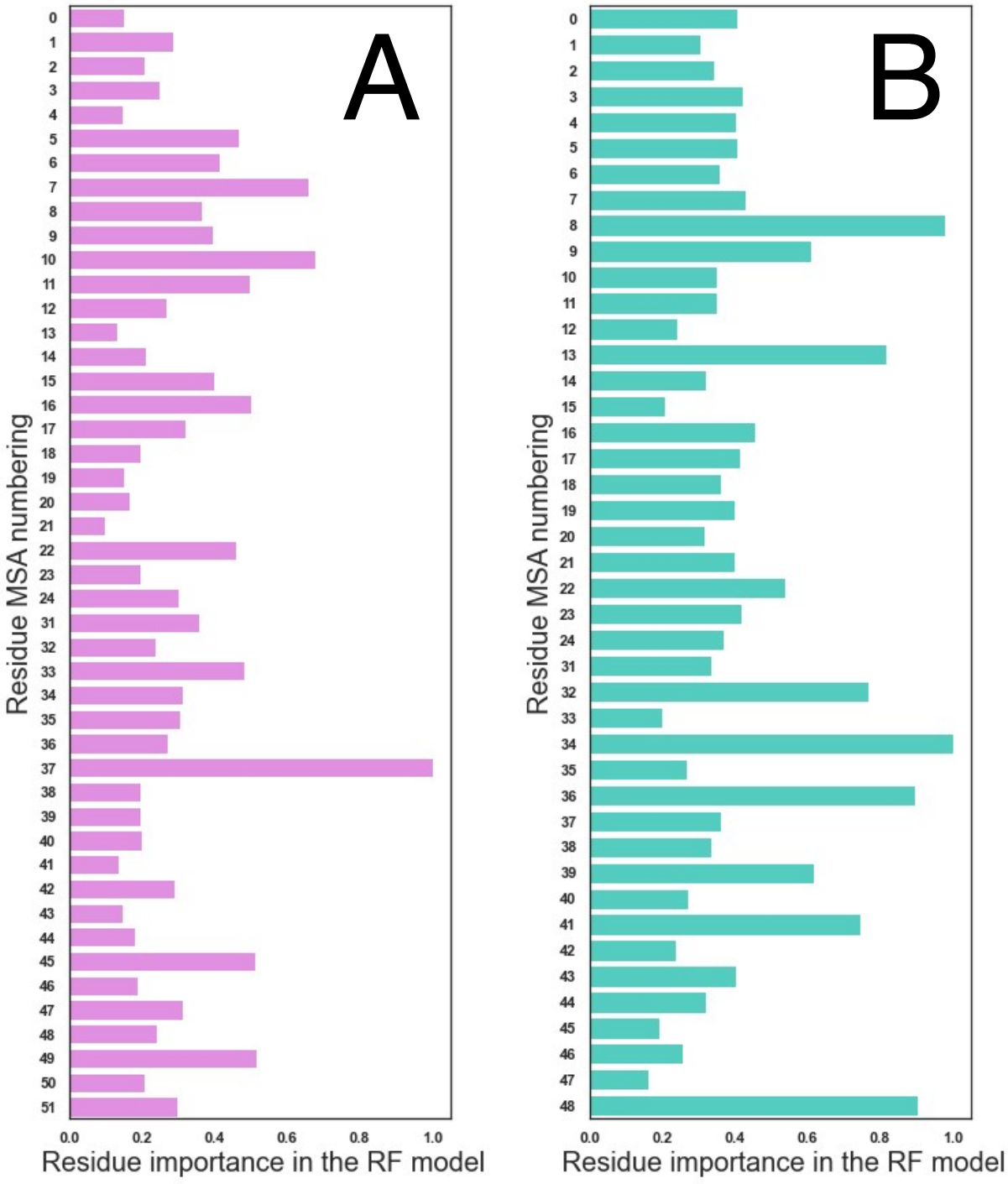

Figure 7.4: Per-residues importance according to MSA position for the RF model using molecular dynamics distance, all pairs at the interface, and the Dima potential[266] for A) the Dprs and B) the DIPs.

Analyzing the two per-residue profiles (see Figure 7.3B and Figure 7.4A), we still identify identical critical residues as when using MI interacting pairs, such as 5, 10, and 22 for the Dprs, with high importance scores above 0.5 in both cases. Residue 10 was identified as well through the hydrogen-bond analysis with high occupancies for the cognate complexes in the geometrical route, demonstrating its

specific importance. In the case of the DIPs, residues 32 and 34 exhibited both a relative importance above 0.5 for both the MI and all pairs. Interestingly, residue 32 was observed to form hydrogen bonds with weak occupancy in the cognate complexes examined in the geometrical route (see Figure 7.2)

While extending the analysis to include all residue pairs validated some critical residues identified through MI pairs, our investigation revealed that direct classification relying on these critical residues interactions in the modeled complex structures is impossible. This finding is in line with the study of Sergeeva et al., which demonstrates that the compatibility between complexes is predominantly governed by negative constraints rather than by the formation of new and stronger interactions. These constraints were found to be specific to each binding subgroup defined by Sergeeva et al.,[282] emphasizing the need for careful consideration of subgroup-specific constraints and, highlighting the challenge of generalizing the findings to other complexes.

Expanding to Drosophila Family Homodimers and to Other Species

To assess the robustness and further validate our ML approach differently, we applied the trained ML algorithms (LDA or RF) to other Dpr-DIP complexes from other close species. Thirteen species from the Drosophila family with the most DIP and Dpr matches in the Uniprot database[279] were considered, namely *Drosophila buskii*, *Drosophila sechellia*, *Drosophila persimillis*, *Drosophila simulans*, *Drosophila rhopaloa*, *Drosophila guanche*, *Drosophila pseudoobscura pseudoobscura*, *Drosophila ananassae*, *Drosophila kikkawai*, *Drosophila virilis*, *Drosophila grimshawi*, *Drosophila Willistoni*, and *Drosophila mojavensis* (see the detailed procedure in the Methods section).The level of annotation in the Uniprot database for the recovered protein sequences of these thirteen species is thin (score of one or two out of five in the majority), and several were only referred to as their genetic loci (genetic position). This result can be explained by the focus on *Drosophila Melanogaster* as a biologically relevant model, given the short lifetime, the high numbers of larvas, and the low cost associated with their breeding.[283,284]

[ht!]

Table 7.10: Accuracy results for trained LDA and RF using the different species of *Drosophila*.

Species (*Drososphila ...*)	LDA accuracy	RF accuracy
... Busckii	0.92	1.0
... Sechellia	0.95	0.98
... Persimillis	0.99	1.0
... Simulans	0.98	0.99
... Rhopaloa	0.94	0.99
... Guanche	0.96	0.99
... Pseudoobscura Pseudoobscura	0.92	0.995
... Ananassae	0.92	0.99
... Kikkawai	0.93	0.98
... Virilis	0.94	1.0
... Grimshawi	0.95	0.96
... Willistoni	0.96	0.99
... Mojavensis	0.93	0.98

Due to the high percentage of identity for the found protein matches (the lowest being 44.4% for Dpr12 in *Drosophila Grimshawi*, and 99% for the highest, see the SI), we assume that cognate complexes in *Drosophila Melanogaster* were conserved in the different *Drosophila* species, and that the fold was identical to reuse the Dpr10-DIPα $\alpha-$carbon atoms distance. We chose the Betancourt potential,[268] as it provided the best results with these conditions for both RF and LDA (see Table 7.7).

We obtained an accuracy in the range of 0.92−0.96 and 0.98−1.0, for LDA and RF, respectively. It is noteworthy that the accuracies observed in our study align closely with the training phase conducted on *Drosophila Melanogaster*, which demonstrates the remarkable consistency and robustness of our strategy. We, therefore, anticipate that our ML models will exhibit a similar level of proficiency when presented with novel Dpr-DIP complexes. Thus, the models will be able to effectively distinguish and accurately classify these complexes as either cognate or non-cognate.

The Dpr-DIP interactome comprises also a few homodimers formed by DIPα, DIPη, DIPθ, DIPζ, Dpr8, Dpr12 and Dpr21. We generated the input features for all possible homodimers using a distance taken from one specific DIP homodimer (DIPα, PDB: 6EFY). Using the already trained LDA model, we successfully predicted the correct classes of such homodimers, except for DIPζ, which corresponds to the weakest binding affinity amongst the three available measurements, namely for DIPα, DIPη and DIPζ. This result underscores the predictive power of the present methodology generalized from hetero– to homodimers, yet training only on a single type of complexes. It is noteworthy, however, that RF fails to predict with the expected high accuracy (0.54) the strongest candidates amongst the DIPs homodimers, which may stem from a potential overfitting, as has been detected in previous analysis (see section Machine Learning Strategies). The binding affinities of the homodimers are also weaker than those characteristic of the heterodimers formed between the proper partners, which can explain the difficulty of RF to accurately identify the strong binders. The homophilic interactions are hypothesized to play a role in rescuing a default in heterophilic interactions.[285]

Members of the Dpr and DIP families are homologous to those of the IgLONs family in humans, which is composed of the opioid-binding protein/cell adhesion molecule-like (OPCML/OB-CAM/IgLON1), the neurotrimin (NTM/IgLON2), the limbic system associated membrane protein (LSAMP/IgLON3), the neuronal growth regulator 1 (NEGR1/IgLON4), and the IGLON5. They can form hetero– and homodimers with a high affinity, i.e., −9.9 kcal/mol for the weakest complex, the IGLON4 homodimer.[286] We tested our model on the 25 interaction combinations, considering all complexes as cognate, given their affinity for one another.[286] The distance feature was extracted from the IgLON5/NEGR1 heterodimer structure with PDB accession number 6DLD.[286] The model was modified to use the MI directly in lieu of the MI difference between cognate and non-cognate, since we do not have any non-cognate IgLON complex. The LDA model predicted all the complexes as cognate, whereas RF failed to classify them accurately. To further verify the LDA results, we assigned IgLON complexes with affinity below −11 kcal/mol as non-cognate complexes to use the MI difference in the input feature. This LDA model still classifies all complexes as cognate, which is expected given the greater affinity of IgLONs compared to that of the strongest Dpr-DIP complex (−8.2 kcal/mol for Dpr9-DIPλ).[8] The success of the LDA model for the IgLONS further validates our findings that the algorithm is applicable to different species, or complexes. In stark contrast, RF fails to predict on a different dataset, thus highlighting the limitations of the algorithm. However, when aiming at identifying the key residues for binding in a particular dataset, with no interest in transferability, RF outperforms LDA by learning better the specific patterns and leveraging the RF ensemble properties to obtain the relative residue importance in the classification.

7.2.4 Conclusion

In this work, we sought to elucidate the molecular basis differentiating strong (cognate) and weak (non-cognate) binders of the Dpr-DIP interactome with state-of-the-art approaches resting on a restraint-based method, the so-called geometrical route,[57] and ML with RF[35] and LDA.[69] The excellent agreement between experimental and computational binding free energy estimates obtained with the geometrical route[57] for two cognate complexes underscores the predictive power of the methodology. The estimation of low binding affinities for non-cognate complexes further validated the method by providing low estimates that fit within the expected range. We emphasize the need for specific side chain restraints

in this particular family of proteins due to the exposure of the interface to water, although it may only sometimes be necessary.[57] However, the geometrical route has limitations of its own when dealing with a large number of congeneric complexes due to its computational cost and unique tuning requirements, incompatible with high-throughput predictions. As an alternative, we turned to ML methods, which are well-suited for handling large datasets. The difficulty resides in choosing an appropriate description of the binding interface in a proper format for the algorithm.[41] We used input features composed of a multiple sequence alignment (sequence-based approach) with a structure-based distance and a pre-established pairwise interaction potential (physicochemical compatibility), proving that combining different types of information is vital for accurate predictions. These three components are non-specific, so that the input features can be seamlessly transferred to any similar biological interacting problem, like the binding specificity of cell-adhesion proteins, such as cadherins,[6] as we demonstrated the transferability of the trained LDA model to homologous proteins, namely IgLONs and Dpr-DIP homodimers. Furthermore, using per-residue importance extracted from the RF model, we were able to identify key residues necessary for Dpr-DIP binding. However, they are insufficient to use directly as a criterion to separate randomly selected cognate and non-cognate complexes. It should be noted that this observation is specific to the Dpr-DIP interactome,[282] and should not hinder the identification of essential residues for a different set of protein complexes. Additionally, we acknowledge that our classification of Dpr-DIP complexes was based solely on in-vitro assays,[8] overlooking crucial biological factors such as cell expression and localization, the importance of which has been investigated previously.[287,288] These factors should be taken into account for a more comprehensive understanding of the Dpr-DIP interactome at a biological level. Overall, our study provides insights into the molecular basis of binding specificity in the Dpr-DIP interactome and highlights the potential of combining physics-based and ML approaches to analyze protein–protein interactions and quickly separate the wheat from the chaff—i.e., the strong binders from the weaker ones, acting like a selection filter for high throughput. Our method is primarily aimed at protein families with evolutionarily linked members due to the MSA requirement for the MI in the input features. Furthermore, supervised ML requires prior knowledge of the "target" value to learn how to predict in the training process. Datasets with robust structural and thermodynamic information are scarce, and mainly directed at protein–ligand complexes, e.g., PDBbind,[289] which contains 19,433 protein–ligand complexes versus 2,852 protein–protein complexes in the current version (2020), thereby limiting the use of supervised ML algorithms for protein–protein standard binding free-energy predictions.

7.3 Supporting Informations

7.3.1 Detailled PMFs contributions for each Dpr-DIP complexes

Table 7.11: Detailed results of the different contributions to the binding free energy for Dpr6-DIPβ

Contribution	PMF (kcal/mol)	PMF (ns)
$\Delta G^{\text{site}}_{c(\text{Dpr6backbone})}$	$-$ 6.0 $\pm$ 0.0	250
$\Delta G^{\text{site}}_{c(\text{DIP}\beta\text{backbone})}$	$-$ 7.3 $\pm$ 0.1	200
$\Delta G^{\text{site}}_{c(\text{Dpr6sidechains})}$	$-$ 2.7 $\pm$ 0.0	150
$\Delta G^{\text{site}}_{c(\text{DIP}\beta\text{sidechains})}$	$-$ 3.1 $\pm$ 0.1	200
$\Delta G^{\text{site}}_{\Theta}$	$-$ 0.1 $\pm$ 0.1	40
$\Delta G^{\text{site}}_{\Phi}$	$-$ 0.2 $\pm$ 0.0	40
$\Delta G^{\text{site}}_{\Psi}$	$-$ 0.2 $\pm$ 0.0	40
$\Delta G^{\text{site}}_{\theta}$	$-$ 0.5 $\pm$ 0.0	120
$\Delta G^{\text{site}}_{\phi}$	$-$ 0.5 $\pm$ 0.0	70
$(1/\beta) * \ln(\text{S} * \text{I} * \text{C}_0)$	$-$ 26.7 $\pm$ 0.0	150
$\Delta G^{\text{bulk}}_{c(\text{Dpr6backbone})}$	11.1 $\pm$ 0.0	300
$\Delta G^{\text{bulk}}_{c(\text{DIP}\beta\text{backbone})}$	12.6 $\pm$ 0.1	550
$\Delta G^{\text{bulk}}_{c(\text{Dpr6sidechains})}$	3.8 $\pm$ 0.0	100
$\Delta G^{\text{bulk}}_{c(\text{DIP}\beta\text{sidechains})}$	7.7 $\pm$ 0.1	100
$\Delta G^{\text{bulk}}_{o}$	6.6	
$\Delta G^{\circ}_{\text{b}}$	-5.4 ± 0.5 (calculation) -6.5 (experiment)[8]	

Table 7.12: Detailed results of the different contributions to the binding free energy for Dpr6-DIPθ

Contribution	PMF (kcal/mol)	PMF (ns)
$\Delta G^{\text{site}}_{c(\text{Dpr6backbone})}$	$-\ 6.2 \pm 0.1$	200
$\Delta G^{\text{site}}_{c(\text{DIP}\theta\text{backbone})}$	$-\ 5.2 \pm 0.1$	100
$\Delta G^{\text{site}}_{c(\text{Dpr6sidechains})}$	$-\ 7.3 \pm 0.1$	100
$\Delta G^{\text{site}}_{c(\text{DIP}\theta\text{sidechains})}$	$-\ 2.6 \pm 0.0$	100
$\Delta G^{\text{site}}_{\Theta}$	$-\ 0.2 \pm 0.0$	40
$\Delta G^{\text{site}}_{\Phi}$	$-\ 0.2 \pm 0.0$	60
$\Delta G^{\text{site}}_{\Psi}$	$-\ 0.2 \pm 0.0$	40
$\Delta G^{\text{site}}_{\theta}$	$-\ 0.6 \pm 0.0$	40
$\Delta G^{\text{site}}_{\phi}$	$-\ 0.6 \pm 0.0$	40
$(1/\beta) * \ln(\text{S} * \text{I} * \text{C}_0)$	$-\ 10.3 \pm 0.0$	100
$\Delta G^{\text{bulk}}_{c(\text{Dpr6backbone})}$	5.9 ± 0.1	200
$\Delta G^{\text{bulk}}_{c(\text{DIP}\theta\text{backbone})}$	5.2 ± 0.1	200
$\Delta G^{\text{bulk}}_{c(\text{Dpr6sidechains})}$	7.8 ± 0.1	150
$\Delta G^{\text{bulk}}_{c(\text{DIP}\theta\text{sidechains})}$	5.4 ± 0.1	100
$\Delta G^{\text{bulk}}_{o}$	6.6	
$\Delta G^{\circ}_{\text{b}}$	-2.3 ± 0.6 (calculation) < -4.5[8]	

Table 7.13: Detailed results of the different contributions to the binding free energy for Dpr6-DIPα

Contribution	PMF (kcal/mol)	PMF (ns)
$\Delta G^{\text{site}}_{c(\text{Dpr6backbone})}$	$-$ 6.4 $\pm$ 0.0	250
$\Delta G^{\text{site}}_{c(\text{DIP}\alpha\text{backbone})}$	$-$ 3.4 $\pm$ 0.0	150
$\Delta G^{\text{site}}_{c(\text{Dpr6sidechains})}$	$-$ 6.3 $\pm$ 0.0	250
$\Delta G^{\text{site}}_{c(\text{DIP}\alpha\text{sidechains})}$	$-$ 2.7 $\pm$ 0.0	150
$\Delta G^{\text{site}}_{\Theta}$	$-$ 0.1 $\pm$ 0.1	60
$\Delta G^{\text{site}}_{\Phi}$	$-$ 0.2 $\pm$ 0.0	60
$\Delta G^{\text{site}}_{\Psi}$	$-$ 0.2 $\pm$ 0.0	40
$\Delta G^{\text{site}}_{\theta}$	$-$ 0.7 $\pm$ 0.0	70
$\Delta G^{\text{site}}_{\phi}$	$-$ 0.5 $\pm$ 0.0	40
$(1/\beta) * \ln(\text{S} * \text{I} * \text{C}_0)$	$-$ 21.1 $\pm$ 0.0	100
$\Delta G^{\text{bulk}}_{c(\text{Dpr6backbone})}$	6 $\pm$ 0.0	200
$\Delta G^{\text{bulk}}_{c(\text{DIP}\alpha\text{backbone})}$	7.1 $\pm$ 0.0	200
$\Delta G^{\text{bulk}}_{c(\text{Dpr6sidechains})}$	7.2 $\pm$ 0.0	250
$\Delta G^{\text{bulk}}_{c(\text{DIP}\alpha\text{sidechains})}$	7.8 $\pm$ 0.0	150
$\Delta G^{\text{bulk}}_{o}$	6.6	
ΔG_{b}	-6.7 ± 0.1 (calculation) -7.7 (experiment)[8]	2120

Table 7.14: Detailed results of the different contributions to the binding free energy for Dpr6-DIPγ

Contribution	PMF (kcal/mol)	PMF (ns)
$\Delta G^{\text{site}}_{c(\text{Dpr6backbone})}$	$-3.3 \pm 0{,}1$	200
$\Delta G^{\text{site}}_{c(\text{DIP}\gamma\text{backbone})}$	-9.6 ± 0.0	250
$\Delta G^{\text{site}}_{c(\text{Dpr6sidechains})}$	-4.5 ± 0.0	250
$\Delta G^{\text{site}}_{c(\text{DIP}\gamma\text{sidechains})}$	-0 ± 0.1	100
$\Delta G^{\text{site}}_{\Theta}$	-0.2 ± 0.0	40
$\Delta G^{\text{site}}_{\Phi}$	-0.4 ± 0.0	60
$\Delta G^{\text{site}}_{\Psi}$	-0.3 ± 0.0	60
$\Delta G^{\text{site}}_{\theta}$	-0.3 ± 0.0	40
$\Delta G^{\text{site}}_{\phi}$	-0.5 ± 0.0	40
$(1/\beta) * \ln(\text{S} * \text{I} * \text{C}_0)$	-8.4 ± 0.0	200
$\Delta G^{\text{bulk}}_{c(\text{Dpr6backbone})}$	$5{,}4 \pm 0.0$	350
$\Delta G^{\text{bulk}}_{c(\text{DIP}\gamma\text{backbone})}$	$5{,}8 \pm 0.1$	350
$\Delta G^{\text{bulk}}_{c(\text{Dpr6sidechains})}$	5.3 ± 0.0	200
$\Delta G^{\text{bulk}}_{c(\text{DIP}\gamma\text{sidechains})}$	0 ± 0.1	100
$\Delta G^{\text{bulk}}_{o}$	6.6	
$\Delta G^{\circ}_{\text{b}}$	-4.5 ± 0.4 (calculation) < -4.1 kcal[8]	2240

7.3.2 Pairwise amino acids potentials

Residue	A	G	L	M	F	W	K	Q	E	S	P	V	I	C	Y	H	R	N	D	T
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	1	1	1	1	-2	-1	-2	0	1	1	1	-0.5	-0.5	-0.5	-2	-1	-2	0
M	0	0	1	1	1	1	-2	-1	-2	0	1	1	1	-0.5	-0.5	-0.5	-2	-1	-2	0
F	0	0	1	1	1	1	-2	-1	-2	0	1	1	1	-0.5	-0.5	-0.5	-2	-1	-2	0
W	0	0	1	1	1	1	-2	-1	-2	0	1	1	1	-0.5	-0.5	-0.5	-2	-1	-2	0
K	0	0	-2	-2	-2	-2	-2	1	2	0	-1	-2	-2	0.5	1	1	-2	1	2	0
Q	0	0	-1	-1	-1	-1	1	1	1	0	-1	-1	-1	0.5	1	1	1	1	1	0
E	0	0	-2	-2	-2	-2	2	1	-2	0	-1	-2	-2	0.5	1	1	2	1	-2	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	1	1	1	1	-1	-1	-1	0	0	1	1	-0.5	-0.5	-0.5	-2	-1	-2	0
V	0	0	1	1	1	1	-2	-1	-2	0	1	1	1	-0.5	-0.5	-0.5	-2	-1	-2	0
I	0	0	1	1	1	1	-2	-1	-2	0	1	1	1	-0.5	-0.5	-0.5	-2	-1	-2	0
C	0	0	-0.5	-0.5	-0.5	-0.5	0.5	0.5	0.5	0	-0.5	-0.5	-0.5	2	0.5	0.5	0.5	0.5	0.5	0
Y	0	0	-0.5	-0.5	-0.5	-0.5	1	1	1	0	-0.5	-0.5	-0.5	0.5	1	1	1	1	1	0
H	0	0	-0.5	-0.5	-0.5	-0.5	1	1	1	0	-0.5	-0.5	-0.5	0.5	1	1	1	1	1	0
R	0	0	-2	-2	-2	-2	-2	1	2	0	-2	-2	-2	0.5	1	1	-2	1	2	0
N	0	0	-1	-1	-1	-1	1	1	1	0	-1	-1	-1	0.5	1	1	1	1	1	0
D	0	0	-2	-2	-2	-2	2	1	-2	0	-2	-2	-2	0.5	1	1	2	1	-2	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 7.15: Elementary pairwise amino acids potential[11]

Residue	A	G	L	M	F	W	K	Q	E	S	P	V	I	C	Y	H	R	N	D	T
A	-0.08	-0.09	-0.57	0.05	-0.05	-0.08	0.0	-0.11	0.03	-0.01	0.41	-0.60	-0.64	-0.30	-0.37	-0.15	0.07	-0.14	0.10	-0.22
G	-0.09	0.04	-0.08	0.00	0.21	-0.14	0.10	0.12	0.40	-0.04	0.40	-0.20	0.40	-0.42	-0.32	0.0	-0.15	-0.18	-0.06	0.13
L	-0.57	-0.08	-1.14	-0.31	-1.02	-0.97	0.10	-0.13	-0.05	-0.07	0.09	-1.03	-1.04	-0.69	-0.60	-0.36	-0.10	-0.10	0.50	-0.39
M	0.05	0	-0.31	0.36	-1.03	-0.07	0.55	-0.54	0.02	-0.47	-0.25	-0.94	-0.41	-1.23	-1.10	-0.35	-0.43	0.31	1.07	0.06
F	-0.05	0.21	-1.02	-1.03	-0.61	-0.89	-0.17	0.10	0.19	0.14	-0.43	-0.78	-0.66	-0.98	-0.82	-0.75	-0.22	-0.02	0.20	-0.19
W	-0.08	-0.14	-0.97	-0.07	-0.89	0.02	-0.30	0.40	0.32	-0.20	-0.44	-0.6	-0.89	-0.3	-0.99	-0.41	-0.78	-0.68	0.24	0.07
K	0.0	0.10	0.10	0.55	-0.17	-0.30	1.45	0.04	-0.09	0.18	0.51	-0.08	-0.26	0.0	-0.20	0.14	0.30	0.18	-0.09	0.09
Q	-0.11	0.12	-0.13	-0.54	0.10	0.40	0.04	-0.08	0.62	0.40	-0.21	-0.09	-0.39	-0.49	-0.39	0.06	0.62	-0.05	0.46	-0.17
E	0.03	0.40	-0.05	0.02	0.19	0.32	-0.09	0.62	0.21	0.30	0.33	-0.02	-0.20	0.04	0.22	-0.53	-0.26	-0.25	0.68	0.15
S	-0.01	-0.04	-0.07	-0.47	0.14	-0.20	0.18	0.40	0.30	-0.13	0.44	-0.31	-0.13	-0.20	0.25	0.05	0.61	0.37	-0.09	0.05
P	0.41	0.40	0.09	-0.25	-0.43	-0.44	0.51	-0.21	0.33	0.44	0.28	-0.08	0.25	0.07	-0.45	-0.22	-0.02	0.11	0.84	0.36
V	-0.60	-0.20	-1.03	-0.94	-0.78	-0.60	-0.08	-0.09	-0.02	-0.31	-0.08	-1.15	-0.98	-0.94	-0.70	-0.35	-0.48	-0.24	0.25	0.06
I	-0.64	0.40	-1.04	-0.41	-0.66	-0.89	-0.26	-0.39	-0.20	-0.13	0.25	-0.98	-0.71	-0.48	-0.87	-0.52	-0.08	0.39	0.04	-0.29
C	-0.30	-0.42	-0.69	-1.23	-0.98	-0.30	0.0	-0.49	0.04	-0.20	0.07	-0.94	-0.48	-1.79	-0.96	-0.82	-0.40	-0.32	0.55	-0.38
Y	-0.37	-0.32	-0.60	-1.10	-0.82	-0.99	-0.20	-0.39	0.22	0.25	-0.45	-0.70	-0.87	-0.96	0.35	-0.67	0.21	-0.74	0.11	-0.23
H	-0.15	0.00	-0.36	-0.35	-0.75	-0.41	0.14	0.05	-0.53	0.05	-0.22	-0.35	-0.52	-0.82	-0.67	0.14	-0.01	0.06	-0.06	-0.27
R	0.07	-0.15	-0.10	-0.43	-0.22	-0.78	0.30	0.62	-0.26	0.61	-0.02	-0.48	-0.08	-0.40	0.21	-0.01	0.23	0.04	-0.15	-0.17
N	-0.14	-0.18	-0.10	0.31	-0.02	-0.68	0.18	-0.05	-0.25	0.37	0.11	-0.24	0.39	-0.32	-0.74	0.06	0.04	-0.86	-0.12	-0.27
D	0.10	-0.06	0.50	1.07	0.20	0.24	-0.09	0.46	0.68	-0.09	0.84	0.25	0.04	0.55	0.11	-0.06	-0.15	-0.12	0.60	-0.03
T	-0.22	0.13	-0.39	0.06	-0.19	0.07	0.09	-0.17	0.15	0.05	0.36	0.06	-0.29	-0.38	-0.23	-0.27	-0.17	-0.27	-0.03	0.26

Table 7.16: Dill pairwise amino acids potential[265]

Residue	A	G	L	M	F	W	K	Q	E	S	P	V	I	C	Y	H	R	N	D	T
A	-0.7070	0.5198	-0.7009	0.4328	1.0697	0.1185	0.8608	-0.7140	0.1305	0.1172	0.1165	0.1134	-0.3824	-0.1557	-1.3848	-0.0224	0.9747	-0.4957	-0.1949	0.1776
G	0.5198	-0.2049	0.2492	-0.1603	-0.7531	0.3134	-0.8587	-0.1131	-0.2682	-1.2261	0.4681	0.3017	1.3344	0.8033	0.6074	-0.0566	0.1600	-1.0720	-0.3322	0.6343
L	-0.7009	0.2492	-1.6636	0.3821	-0.6108	1.4267	-0.4350	0.0831	0.3980	0.4718	0.5276	-0.9996	-1.3772	-0.6380	0.8716	-0.0386	-0.8436	0..8640	0.7433	1.0923
M	0.4328	-0.1603	0.3821	0.5294	-1.1503	0.3734	-0.6290	0.0490	0.2455	-0.1111	-0.0859	0.6549	0.2443	-0.5745	-0.6279	-0.3334	-0.0258	0.5457	-0.0525	0.2743
F	1.0697	-0.7531	-0.6108	-1.1503	-1.3629	0.4297	0.2475	2.0038	0.7324	0.9225	-0.1826	-0.9834	0.0699	0.6468	-0.6891	-0.4500	-0.9308	0.9747	0.6461	-1.2400
W	0.1185	0.3134	1.4267	0.3734	0.4297	0.1134	0.1976	-0.9070	0.3664	0.0699	-0.3461	-0.7498	-0.3289	-0.9614	-0.7988	0.0669	-0.1341	-0.1321	0.1901	0.0238
K	0.8608	-0.8587	-0.4350	-0.6290	0.2475	0.1976	0.3968	-0.0936	-1.3099	0.6963	0.2947	-0.2309	0.2210	0.7892	0.4131	-0.1756	1.2654	-0.5092	-1.2961	0.5704
Q	-0.7140	-0.1131	0.0831	0.0490	2.0038	-0.9070	-0.0936	-0.5176	-1.3927	0.1719	-0.2668	0.0301	-0.1433	0.2150	0.0515	0.6932	0.2842	-0.4228	0.1232	0.8629
E	0.1305	-0.2682	0.3980	0.2455	0.7324	0.3664	-1.3099	-1.3927	0.8208	-0.6826	0.6673	0.1672	0.2700	0.7040	0.6113	-0.4849	-1.9339	-0.5738	1.7472	0.2074
S	0.1172	-1.2261	0.4718	-0.1111	0.9225	0.0699	0.6963	0.1719	-0.6826	0.1131	1.1270	0.5035	-0.3377	0.1736	0.2182	-0.2969	-0.2162	-0.8147	-0.4756	-0.3658
P	0.1165	0.4681	0.5276	-0.0859	-0.1826	-0.3461	0.2947	-0.2668	0.6673	1.1270	-0.1151	-0.0721	-0.9709	-1.1189	0.3825	0.5118	0.6611	0.0212	-0.0663	-0.9949
V	0.1134	0.3017	-0.9996	0.6549	-0.9834	-0.7498	-0.2309	0.0301	0.1672	0.5035	-0.0721	-0.8266	-0.0977	0.1563	0.2563	0.7549	-0.1950	-0.1169	0.3597	0.6555
I	-0.3824	1.3344	-1.3772	0.2443	0.0699	-0.3289	0.2210	-0.1433	0.2700	-0.3377	-0.9709	-0.0977	-0.9924	0.0656	-0.7972	-0.0868	1.0170	1.2717	0.6350	-0.2305
C	-0.1557	0.8033	-0.6380	-0.5745	0.6468	-0.9614	0.7892	0.2150	0.7040	0.1736	-1.1189	0.1563	0.0656	-1.6688	-0.2008	-0.0542	0.2098	-0.1620	0.6354	0.4686
Y	-1.3848	0.6074	0.8716	-0.6279	-0.6891	-0.7988	0.4131	0.0515	0.6113	0.2182	0.3825	0.2563	-0.7972	-0.2008	1.1437	0.2091	0.2496	-0.6676	-0.0322	-0.2384
H	-0.0224	-0.0566	-0.0386	-0.3334	-0.4500	0.0669	-0.1756	0.6932	-0.4849	-0.2969	0.5118	0.7549	-0.0868	-0.0542	0.2091	0.4672	-0.3751	0.3610	0.0667	-0.9127
R	0.9747	0.1600	-0.8436	-0.0258	-0.9308	-0.1341	1.2654	0.2842	-1.9339	-0.2162	0.6611	-0.1950	1.0170	0.2098	0.2496	-0.3751	0.9061	0.4824	-1.7453	0.4926
N	-0.4957	-1.0720	0..8640	0.5457	0.9747	-0.1321	-0.5092	-0.4228	-0.5738	-0.8147	0.0212	-0.1169	1.2717	-0.1620	-0.6676	0.3610	0.4824	-0.2057	0.7542	-0.0858
D	-0.1949	-0.3322	0.7433	-0.0525	0.6461	0.1901	-1.2961	0.1232	1.7472	-0.4756	-0.0663	0.3597	0.6350	0.6354	-0.0322	0.0667	-1.7453	0.7542	-0.0953	-1.0131
T	0.1776	0.6343	1.0923	0.2743	-1.2400	0.0238	0.5704	0.8629	0.2074	-0.3658	-0.9949	0.6555	-0.2305	0.4686	-0.2384	-0.9127	0.4926	-0.0858	-1.0131	-0.3140

Table 7.17: Dima pairwise amino acids potential[266]

Residue	A	G	L	M	F	W	K	Q	E	S	P	V	I	C	Y	H	R	N	D	T
A	-0.2	-0.26	-0.36	-0.22	-0.46	0.05	0.1	0.01	0.26	-0.09	0.14	-0.42	-0.57	-0.44	-0.5	0.5	0.2	0.07	0.16	-0.05
G	-0.26	-0.12	0.24	0.02	0.17	0.51	0.19	0.46	0.65	0.28	0.6	0.27	0.24	-0.34	-0.35	0.18	0.5	-0.04	0.35	0.28
L	-0.36	0.24	-1.1	-0.5	-1.08	-0.95	0.13	-0.01	0.33	0.17	0.42	-0.97	-1.25	-0.53	-0.63	-0.24	-0.07	0.21	0.28	0.07
M	-0.22	0.02	-0.5	-0.74	-0.78	-0.56	0.48	0.26	0.29	0.48	0.01	-0.73	-0.58	-0.43	-1.02	-0.07	0.15	0.32	0.35	0.16
F	-0.46	0.17	-1.08	-0.78	-0.94	-0.6	0.01	0.26	0.55	-0.22	0.2	-1.15	-0.88	-0.88	-0.88	-0.4	-0.19	0.22	0.38	0.02
W	0.05	0.51	-0.95	-0.56	-0.6	-0.05	0.1	0.18	0.03	-0.22	-0.37	-0.56	-0.68	0.24	-1.41	-0.85	-0.73	-0.12	0.04	0.11
K	0.1	0.19	0.13	0.48	0.01	0.1	0.42	0.15	-0.47	0.1	0.5	0.1	0.15	-0.09	0.04	0.79	0.53	0.26	-0.39	-0.19
Q	0.01	0.46	-0.01	0.26	0.26	0.18	0.15	-0.12	0.04	0.29	1.02	-0.11	-0.08	-0.43	0.11	0.69	0.24	0.37	0.49	0.04
E	0.26	0.65	0.33	0.29	0.55	0.03	-0.47	0.04	0.6	0.18	0.69	0.39	0.37	0.19	0.17	0.18	-0.52	0.01	0.55	0.37
S	-0.09	0.28	0.17	0.48	-0.22	-0.22	0.1	0.29	0.18	-0.06	0.54	0.12	0.32	0.13	-0.14	0.21	0.27	0.15	0.19	0.08
P	0.14	0.6	0.42	0.01	0.2	-0.37	0.5	1.02	0.69	0.54	0.27	-0.02	0.36	-0.14	-0.12	0.25	0.47	0.27	1.03	0.88
V	-0.42	0.27	-0.97	-0.73	-1.15	-0.56	0.1	-0.11	0.39	0.12	-0.02	-0.89	-1.06	-0.62	-0.71	0.16	0.01	0.4	0.69	-0.01
I	-0.57	0.24	-1.25	-0.58	-0.88	-0.68	0.15	-0.08	0.37	0.32	0.36	-1.06	-1.16	-0.36	-0.85	-0.0	0.14	-0.09	0.44	-0.27
C	-0.44	-0.34	-0.53	-0.43	-0.88	0.24	-0.09	-0.43	0.19	0.13	-0.14	-0.62	-0.36	-2.99	-0.79	-1.11	-0.24	-0.52	0.21	-0.22
Y	-0.5	-0.35	-0.63	-1.02	-0.88	-1.41	0.04	0.11	0.17	-0.14	-0.12	-0.71	-0.85	-0.79	-0.76	-0.26	0.01	0.32	0.43	-0.32
H	0.5	0.18	-0.24	-0.07	-0.4	-0.85	0.79	0.69	0.18	0.21	0.25	0.16	-0.0	-1.11	-0.26	0.42	0.24	0.2	-0.23	0.11
R	0.2	0.5	-0.07	0.15	-0.19	-0.73	0.53	0.24	-0.52	0.27	0.47	0.01	0.14	-0.24	0.01	0.24	0.17	0.13	-0.37	0.45
N	0.07	-0.04	0.21	0.32	0.22	-0.12	0.26	0.37	0.01	0.15	0.27	0.4	-0.09	-0.52	0.32	0.2	0.13	0.14	-0.02	0.1
D	0.16	0.35	0.28	0.35	0.38	0.04	-0.39	0.49	0.55	0.19	1.03	0.69	0.44	0.21	0.43	-0.23	-0.37	-0.02	0.17	-0.12
T	-0.05	0.28	0.07	0.16	0.02	0.11	-0.19	0.04	0.37	0.08	0.88	-0.01	-0.27	-0.22	-0.32	0.11	0.45	0.1	-0.12	-0.03

Table 7.18: Dosztányi pairwise amino acids potential[267]

Residue	A	G	L	M	F	W	K	Q	E	S	P	V	I	C	Y	H	R	N	D	T
A	-0.53	-0.39	-0.4	-0.62	-0.65	-0.83	-0.65	-0.68	0.38	-0.38	-0.43	-0.6	-0.65	-0.44	-0.66	-0.51	-0.39	-0.71	0.39	-0.39
G	-0.39	-0.37	-0.58	-0.74	-0.57	-0.68	-0.56	-0.68	0.44	-0.37	-0.52	-0.62	-0.6	-0.41	-0.8	-0.5	-0.48	-0.58	0.36	-0.35
L	-0.4	-0.58	0.14	-0.55	-0.71	-1.08	-0.53	-1.0	-0.98	-0.42	-0.47	-0.55	-0.57	-0.37	-1.04	-0.77	0.5	-0.99	-0.9	-0.42
M	-0.62	-0.74	-0.55	-0.9	-1.13	-1.48	-0.85	-1.09	0.45	-0.63	-0.69	-0.87	-0.91	-0.59	-1.41	-0.76	-0.51	-1.09	0.42	-0.6
F	-0.65	-0.57	-0.71	-1.13	-1.32	-1.77	-1.09	-1.5	0.23	-0.51	-0.79	-1.01	-1.02	-0.7	-1.61	-1.22	-1.05	-1.34	0.29	-0.67
W	-0.83	-0.68	-1.08	-1.48	-1.77	-2.28	-1.35	-1.83	-0.21	-0.65	-1.19	-1.19	-1.3	-1.02	-2.0	-1.47	-1.7	-1.6	-0.01	-0.93
K	-0.65	-0.56	-0.53	-0.85	-1.09	-1.35	-0.89	-1.1	0.51	-0.54	-0.67	-0.77	-0.8	-0.66	-1.4	-0.74	-0.38	-1.02	0.51	-0.59
Q	-0.68	-0.68	-1.0	-1.09	-1.5	-1.83	-1.1	-1.24	-0.08	-0.63	-0.85	-0.97	-1.09	-0.86	-1.48	-1.08	-0.96	-1.13	0.09	-0.71
E	0.38	0.44	-0.98	0.45	0.23	-0.21	0.51	-0.08	1.78	0.05	0.46	0.52	0.58	0.23	-0.27	0.04	-1.09	-0.14	1.89	0.14
S	-0.38	-0.37	-0.42	-0.63	-0.51	-0.65	-0.54	-0.63	0.05	-0.22	-0.39	-0.43	-0.45	-0.44	-0.66	-0.46	-0.36	-0.47	0.03	-0.23
P	-0.43	-0.52	-0.47	-0.69	-0.79	-1.19	-0.67	-0.85	0.46	-0.39	-0.53	-0.64	-0.65	-0.47	-1.05	-0.66	-0.37	-0.76	0.43	-0.39
V	-0.6	-0.62	-0.55	-0.87	-1.01	-1.19	-0.77	-0.97	0.52	-0.43	-0.64	-0.75	-0.73	-0.64	-1.14	-0.65	-0.5	-0.97	0.48	-0.49
I	-0.65	-0.6	-0.57	-0.91	-1.02	-1.3	-0.8	-1.09	0.58	-0.45	-0.65	-0.73	-0.86	-0.52	-1.24	-0.72	-0.29	-1.02	0.5	-0.49
C	-0.44	-0.41	-0.37	-0.59	-0.7	-1.02	-0.66	-0.86	0.23	-0.44	-0.47	-0.64	-0.52	-0.52	-1.02	-0.63	-0.46	-0.78	0.39	-0.45
Y	-0.66	-0.8	-1.04	-1.41	-1.61	-2.0	-1.4	-1.48	-0.27	-0.66	-1.05	-1.14	-1.24	-1.02	-1.9	-1.29	-1.13	-1.39	-0.1	-0.89
H	-0.51	-0.5	-0.77	-0.76	-1.22	-1.47	-0.74	-1.08	0.04	-0.46	-0.66	-0.65	-0.72	-0.63	-1.29	-0.75	-0.58	-0.88	0.04	-0.45
R	-0.39	-0.48	0.5	-0.51	-1.05	-1.7	-0.38	-0.96	-1.09	-0.36	-0.37	-0.5	-0.29	-0.46	-1.13	-0.58	0.96	-1.01	-1.14	-0.4
N	-0.71	-0.58	-0.99	-1.09	-1.34	-1.6	-1.02	-1.13	-0.14	-0.47	-0.76	-0.97	-1.02	-0.78	-1.39	-0.88	-1.01	-1.08	0.05	-0.65
D	0.39	0.36	-0.9	0.42	0.29	-0.01	0.51	0.09	1.89	0.03	0.43	0.48	0.5	0.39	-0.1	0.04	-1.14	0.05	1.88	0.1
T	-0.39	-0.35	-0.42	-0.6	-0.67	-0.93	-0.59	-0.71	0.14	-0.23	-0.39	-0.49	-0.49	-0.45	-0.89	-0.45	-0.4	-0.65	0.1	-0.24

Table 7.19: Betancourt pairwise amino acids potential[268]

Conclusion and Perspectives (English version)

This thesis has contributed to assessing whether a protein-protein complex interacts strongly using traditional and chemically accurate binding free-energies methods and ML approaches.

We successfully developed a protocol[82] to compute absolute binding free energies of protein-ligand complexes using the method of the geometrical and alchemical route leveraging the software BFEE2.[81] My contribution consisted of three complexes: N-butyl-benzene, paraxylene, and ethylbenzene, all bound to the Lyzozyme T4L99A. These complexes involved buried ligands, so I used the alchemical route to develop the protocol. We recover binding affinity up to chemical accuracy for each complex. This protocol encompasses the use of BFEE2[81] and is constructed to limit human intervention and facilitate both the setup and the post-treatment for those tedious calculations. I hope this protocol will democratize free-energy calculations and guide non-specialists to enlarge the scope of binding free-energy calculations, which are crucial in discovering drugs and new treatments.

This protocol was, however, limited to protein-ligand complexes. The alchemical route cannot be extended to a protein-protein complex on account of the significant perturbation to dissipate all the interactions of one of the two proteins, too big to converge in characteristic simulation time. During my PhD, we were still in the middle of the COVID-19 syndemic, so we studied the variants of concern and assessed their different binding affinities to confirm their strategy for high transmission: immune escape or higher affinity for ACE2. We extended the protocol to protein-protein by adding additional restraints to the second partner, RMSD, both in the bound and unbound form during the geometrical route. We obtained binding free energies at chemical accuracy compared to experiment data except when using modeled structures proposed in the absence of an experimental structure for the WT and the Beta variant. The binding free energy of models differed for more than a kcal/mol for both the experimental data and the estimate obtained with accurate structures. This work demonstrated the need for a proper structure to match experimental data, the limitations of models, the vulnerability of the method to the initial structural data, and the extension of our protocol to protein-protein complexes. Furthermore, these results demonstrate the potency of an in-sillico approach like the geometrical route in fighting emerging diseases.

In gathering data on COVID-19 simulations and binding affinity, we discovered that our method's shortcuts were employed to gain computational time and simplify calculations.[135,172] Simulations of the reversible separation of protein-protein or protein-ligand were computed bereft of restraints to obtain the binding free energy. We carefully assess the ability of such simulations to produce accurate and reproducible estimates on five different complexes, both protein-protein and protein-ligand. We demonstrated the shortcomings of these methods and their lack of reproducibility and agreement with experimental data contrary to the rigorous geometrical route. This methodological investigation emphasizes the need for restraints and their contribution to speeding up the convergence of the reversible separation of the two molecular partners. In the case of the pig insulin dimer, without restraints, we estimate the convergence

time to fully explore the configurational space based on the CV used in the geometrical to be 35 μs, so almost 35 times the computational time required for the geometrical route. These results allowed us to validate our approach and confirmed the absolute necessity of restraints in our binding free-energy calculations. No matter how tedious and lengthy the preparation of the diverse biased simulations takes, the geometrical route is faster, in the long run, than shortcuts to obtain converged and accurate simulations. Furthermore, tools such as BFEE2,[81] as proposed in our protocol, can help prepare those tedious calculations.

We, therefore, searched for an alternative solution to speed up the calculations that will maintain the excellent accuracy and reproducibility of the estimate obtained with the geometrical route. We took advantage of both MTS and HMR tricks[258] to reduce the effective simulation time. HMR, by artificially increasing the mass of hydrogen atoms without changing the system's total mass, allows for a timestep of 4 fs instead of 2. On the other hand, MTS allows the computation and application of biasing forces on the system to be decoupled from atomic forces and applied every n step (n is user-chosen). This timestep change is possible because biasing forces are less dependent on the current position than atomic forces. Using the model system AblSH3;p41 complex, we demonstrated that HMR alone can speed up the reversible separation in the presence of restraints by a factor of 2 without deterioration of the estimate. However, we observed a degradation of the accuracy when using MTS or MTS with HMR. We investigated the origin of this phenomenon and determined that the extended Lagrangian was perturbed and that the coupling of the extended particle to the actual CV was no longer Gaussian distributed, which is required when using the WTM-eABF algorithm.[49] To fix the issue, we investigated several parameters: the damping factor, the oscillation period, and the coupling width. We found a combination of those parameters to rescue the binding estimate, which allowed us to speed up the calculation by a factor of three. However, this parametrization is system-dependent and cannot be overlooked when performing binding free-energy calculations with WTM-eABF. By speeding up these calculations, performing binding free energy estimation should be less time-consuming and computationally expensive. This methodological survey will help broaden the scope of potential applications of the geometrical route by reducing the computational cost and wall-clock time.

For the Dpr-DIP interactome, we were able to use the protocol developed and extended for protein-protein for binding free-energy calculations to four complexes, two cognates (Dpr6DIPα, Dpr6DIPβ) and two non-cognates (Dpr6-DIPθ, Dpr6-DIPγ). During the computation of the free energy associated with the reversible separation of the proteins, we discovered a sensibility of interfacial interacting amino acids to isomerization in the presence of water, which causes a progressive deterioration of the estimate. To fix the issue, we designed custom restraints to block the motion of those interacting side chains in their initial position using additional RMSD restraints. This study revealed the potential need for supplemental restraints on top of those added to account for the protein-protein nature of the complex. This behavior was already observed in the case of the barnstar-barnase complex,[57] but no restraints were needed in the case of COVID-19 variants.[241] Before the computation of the separation PMF, there is no indication of whether such custom restraints are needed. Although we successfully recovered binding affinity in the expected range for all complexes, generalizing this method to all 231 complexes in a reasonable time and without human intervention was impossible. Given ML's ability to predict properties on large datasets, we turn to these approaches. Using LDA and RF, selected for their interpretability, we predicted with an accuracy of 0.99 the cognate or non-cognate nature of complexes. We used specific input features, which are a unique combination of genetic (MI), physicochemical (pairwise amino acids potential), and structural (distance coming from MD or reference structure) information. These input features are very general and can be used in similar problems. We demonstrated that the LDA could transfer to diverse homologous proteins (homodimer vs. heterodimers) or different species (IgLONs in humans, 13 different *Drosophila* species). We were not able to detect specific interactions to assess the compatibility between partners for the Dpr-DIP directly at the structural level, but this is due mainly to the high similarity

of protein and the presence of negative constraints amongst sub-groups according to Sergeeva et al..[282] However, we demonstrated the potency of the ML approach by obtaining an accurate classification of homologous proteins for diverse species and, more specifically, on the IgLON proteins in humans.

Conclusion et perspectives (version française)

Nous avons réussi à développer notre protocole[82] pour calculer les énergies libres de liaison absolue des complexes protéines-ligand en utilisant la méthode de la route géométrique et alchimique et en exploitant le logiciel BFEE2.[81] Ma contribution se constitue de trois complexes différents : n-butylbenzène, paraxylène et éthylbenzène, tous liés au lysozyme T4L99A. Tous ces complexes impliquaient des ligands enfouis, j'ai donc uniquement utilisé la route alchimique dans le cadre du développement de ce protocole. Nous avons réussi à récupérer l'affinité de liaison avec une précision chimique pour chaque complexe. Ce protocole implique l'utilisation de BFEE2[81] et est conçu pour limiter l'intervention humaine et faciliter à la fois la préparation et le post-traitement de ces calculs fastidieux. Mon espoir est que ce protocole démocratise les calculs d'énergie libre et serve de guide aux non-spécialistes pour élargir le champ d'application des calculs d'énergie libre de liaison, qui sont cruciaux dans la découverte de nouveaux médicaments.

Cependant, ce protocole était limité aux complexes protéines-ligand. La route alchimique ne peut pas être étendue à un complexe protéine-protéine en raison de l'importance de la perturbation nécessaire pour dissiper toutes les interactions de l'une des deux protéines, trop importante pour converger dans les temps caractéristiques de la simulation. Pendant ma thèse, nous étions en plein milieu de la syndémie COVID-19, nous avons donc étudié les variants préoccupants et évalué leurs affinités de liaison différente pour confirmer leur stratégie de transmission élevée : évasion immunitaire ou haute affinité pour ACE2. Nous avons étendu le protocole aux protéines-protéines en ajoutant des contraintes supplémentaires au second partenaire sous la forme de restreintes additionnelles au niveau de son RMSD, à la fois dans sa forme liée et non liée, dans le cadre de la route géométrique. Nous avons réussi à obtenir une affinité de liaison avec une précision chimique par rapport aux données expérimentales, sauf lorsque nous avons utilisé des structures modélisées, proposées en l'absence d'une structure expérimentale dans le cas de la souche WT et la souche Beta. L'énergie libre de liaison des modèles différait de plus d'une kcal/mol par rapport aux données expérimentales et à l'estimation obtenue avec des structures résolues expérimentalement. Ce travail a montré la nécessité d'une structure appropriée pour obtenir des affinités correspondant aux données expérimentales, les limites des modèles et la vulnérabilité de la méthode aux données structurales initiales ainsi que la généralisation du protocole aux complexes protéines-protéines. De plus, ces résultats démontrent l'apport de méthodes computationnelles comme la route géométrique dans la lutte contre des maladies émergentes.

Dans le processus de collecte de données sur les simulations COVID-19 et les calculs d'affinités de liaison, nous avons découvert, dans l'espoir de simplification et de gain de temps de calcul, que des raccourcis de notre méthode étaient utilisés. Les simulations de séparation réversible de protéines-protéines ou de protéines-ligands ont été réalisées sans aucune contrainte pour obtenir l'énergie libre de liaison. Nous avons soigneusement évalué la capacité de telles simulations à produire des estimations précises et reproductibles sur 5 complexes différents, à la fois protéines-protéines et protéines-ligands. Nous avons

pu démontrer les lacunes de ces méthodes, c.-à-d. l'absence de reproductibilité et d'accord avec les données expérimentales contrairement à la route géométrique rigoureuse. Ce travail souligne également la nécessité des contraintes et leur contribution à accélérer la convergence de la séparation réversible des deux partenaires moléculaires. Dans le cas du dimère d'insuline porcine, sans contrainte, nous avons estimé que le temps de convergence nécessaire pour explorer pleinement l'espace de configuration basé sur les CV utilisées dans la route géométrique était de 35 μs, soit presque 35 fois le temps de calcul requis pour la route géométrique. Cela nous a permis de valider notre approche et de confirmer la nécessité absolue des contraintes dans nos calculs d'énergie libre de liaison. Le temps passé à préparer l'ensemble des fichiers initiaux avec la route géométrique est un investissement qui, à long terme, permet d'obtenir des simulations convergées plus rapidement qu'avec les approches qui sont dépourvues de restreintes (approche avec raccourci).

Les raccourcis n'étant pas une option pour réduire le coût computationnel, nous avons cherché une solution qui ne compromettrait pas l'excellente précision et reproductibilité de l'estimation obtenue avec la route géométrique. Nous avons profité à la fois du MTS (Multiple Time Stepping) et de l'astuce HMR (Hydrogen Mass Repartition) pour réduire le temps de simulation effectif. HMR, en augmentant artificiellement la masse des atomes d'hydrogène tout en conservant la masse totale du système, permet d'utiliser un pas de temps de 4 fs au lieu de 2. MTS, en revanche, permet de calculer et d'appliquer les forces sur le système de manière déconnectée des forces atomiques tous les n pas de temps (n est choisi par l'utilisateur). Les forces de biais sont moins dépendantes de la position actuelle que les forces atomiques, rendant possible cette astuce. Nous avons démontré en utilisant le système modèle AblSH3 ; p41 que HMR seul était capable d'accélérer la séparation réversible en présence de contraintes d'un facteur de 2 sans détérioration de l'estimation. Cependant, nous avons observé une dégradation de la précision lors de l'utilisation de MTS ou de MTS en combinaison avec HMR. Nous avons cherché la cause de ce phénomène et avons déterminé que le Lagrangien étendu était perturbé et que la particule fictive n'était plus rattachée à la variable collective par une distribution gaussienne, ce qui est requis lorsque l'on utilise l'algorithme WTM-eABF. Pour résoudre le problème, nous avons étudié un certain nombre de paramètres : le facteur d'amortissement, la période d'oscillation et la fluctuation associée au couplage. Nous avons trouvé une combinaison de ces paramètres capable de restaurer une estimation de l'énergie libre de liaison en accord avec les données expérimentales tout en accélérant le calcul par un facteur trois. Cependant, cette paramétrisation dépend beaucoup du système et ne peut pas être négligée lors de la réalisation de calculs d'énergie libre de liaison avec WTM-eABF.

Pour l'interaction Dpr-DIP, nous avons pu utiliser le protocole développé et étendu pour les protéines-protéines pour les calculs d'énergie libre de liaison de quatre complexes, deux cognates (Dpr6-DIPα, Dpr6-DIPβ) et deux non-cognates (Dpr6-DIPθ, Dpr6-DIPγ). Dans le processus de calcul de l'énergie libre associée à la séparation réversible des protéines, nous avons découvert une sensibilité des acides aminés interfaciaux à l'isomérisation en présence d'eau, ce qui provoque une détérioration progressive de l'estimation. Pour résoudre le problème, nous avons conçu des restreintes personnalisées pour bloquer le mouvement de ces chaînes latérales dans leur position initiale en utilisant des RMSD supplémentaires tout en évaluant leurs coûts. Cette étude a révélé la nécessité potentielle de restreintes supplémentaires en plus de celles ajoutées pour prendre en compte la nature protéine-protéine du complexe. Cela a déjà été observé dans le cas du complexe barnstar-barnase,[57] mais aucune restreinte supplémentaire n'était nécessaire dans le cas des variants de la COVID-19. Avant simulation, il n'y avait aucun moyen de prédire le besoin de ces restreints, qui imposent un réglage spécifique pour chaque complexe. Bien que nous ayons réussi à récupérer l'affinité de liaison dans la plage attendue pour tous les complexes, généraliser la méthode à l'ensemble des 231 complexes n'était pas une option dans un temps raisonnable. Étant donné la capacité de l'apprentissage automatique à prédire des propriétés sur de grands ensembles de données, nous nous sommes tournés vers ces approches. En utilisant la LDA (Analyse Discriminante Linéaire) et le RF (Forêts Aléatoires), nous avons prédit avec une précision de 0,99 la nature cognate

ou non-cognate des complexes. Nous avons utilisé des descripteurs d'entrée spécifiques, qui sont une combinaison unique d'informations génétiques (MI), physicochimiques (potentiel d'interaction entre les acides aminés par paire) et structurales (distance provenant de la dynamique moléculaire ou de la structure de référence). Ces caractéristiques d'entrée sont très générales et peuvent être utilisées dans de nombreux problèmes similaires. Nous n'avons pas pu détecter d'interactions spécifiques pour évaluer la compatibilité entre les partenaires directement au niveau structurel, mais cela est principalement dû à la grande similarité entre les protéines et à la présence de contraintes négatives entre les sous-groupes selon Sergeeva et al..[282] Cependant, nous avons pu démontrer la force de cette méthode en obtenant une classification correcte à l'aide du modèle entraîné pour des protéines homologues chez d'autres espèces et notamment chez l'homme avec les IgLONS.

Appendix A

Example of a colvars file used in chapter 5

An example of a Colvars configurational file employed for scheme 21, in chapter 5 where $N_{\mathrm{MTS}} = 2$, $\sigma = 0.05$ Å, $\gamma_\lambda = 7$ ps^{-1}, and $\tau = 300$ fs, is depicted below.

```
colvarsTrajFrequency      500
colvarsRestartFrequency   50000
indexFile                 ../../complex.ndx
colvar {
    name RMSD
    TimeStepFactor  2   # MTS
    rmsd {
        atoms {
            indexGroup  ligand
        }
        refpositionsfile  ../../complex_largeBox.xyz
    }
}
colvar {
    name eulerTheta
    TimeStepFactor  2    # MTS
    customFunction asin(2 * (q1*q3-q4*q2)) * 180 / 3.1415926
    Orientation {
        name  q
        atoms {
            indexGroup  ligand
            centerReference    on
            rotateReference    on
        enableFitGradients no
            fittingGroup {
                indexGroup  protein
            }
            refpositionsfile  ../../complex_largeBox.xyz
         }
         refpositionsfile  ../../complex_largeBox.xyz
    }
}
colvar {
    name eulerPhi
    TimeStepFactor  2   # MTS
    customFunction atan2(2*(q1*q2+q3*q4), 1-2*(q2*q2+q3*q3)) * 180 / 3.1415926
    Orientation {
        name  q
        atoms {
```

```
            indexGroup  ligand
            centerReference    on
            rotateReference    on
        enableFitGradients no
            fittingGroup {
                indexGroup  protein
            }
            refpositionsfile  ../../complex_largeBox.xyz
         }
         refpositionsfile  ../../complex_largeBox.xyz
    }
}
colvar {
    name eulerPsi
    TimeStepFactor  2     # MTS
    customFunction atan2(2*(q1*q4+q2*q3), 1-2*(q3*q3+q4*q4)) * 180 / 3.1415926
    Orientation {
        name  q
        atoms {
            indexGroup  ligand
            centerReference    on
            rotateReference    on
        enableFitGradients no
            fittingGroup {
                indexGroup  protein
            }
            refpositionsfile  ../../complex_largeBox.xyz
         }
         refpositionsfile  ../../complex_largeBox.xyz
    }
}
colvar {
    name polarTheta
    TimeStepFactor  2       # MTS
    customFunction acos(-i2) * 180 / 3.1415926
    distanceDir {
        name  i
        group1 {
            indexGroup  reference
            centerReference    on
            rotateReference    on
            enableFitGradients no
            fittingGroup {
                indexGroup  protein
            }
            refpositionsfile  ../../complex_largeBox.xyz
        }
        group2 {
            indexGroup  ligand
            centerReference    on
            rotateReference    on
            enableFitGradients no
            fittingGroup {
                indexGroup  protein
            }
            refpositionsfile  ../../complex_largeBox.xyz
        }
    }
}
```

```
colvar {
    name polarPhi
    TimeStepFactor  2       # MTS
    customFunction atan2(i3, i1) * 180 / 3.1415926
    period  360
    wrapAround 0.0
    distanceDir {
        name  i
        group1 {
            indexGroup  reference
            centerReference    on
            rotateReference    on
            enableFitGradients no
            fittingGroup {
                indexGroup  protein
            }
            refpositionsfile  ../../complex_largeBox.xyz
        }
        group2 {
            indexGroup  ligand
            centerReference    on
            rotateReference    on
            enableFitGradients no
            fittingGroup {
                indexGroup  protein
            }
            refpositionsfile  ../../complex_largeBox.xyz
        }
    }
}
colvar {
    name    r
    TimeStepFactor  2  # MTS
    width 0.1
    lowerboundary 10.3
    upperboundary 34.3
    subtractAppliedForce on
    expandboundaries  on
    extendedLagrangian on
    extendedFluctuation 0.05
    ExtendedLangevinDamping 7
    ExtendedTimeConstant 300
    distance {
        forceNoPBC       yes
        group1 {
            indexGroup  reference
    }
        group2 {
            indexGroup  ligand
        }
    }
}
abf {
    colvars        r
    TimeStepFactor  2
    FullSamples    10000
    historyfreq    50000
    writeCZARwindowFile
}
```

```
metadynamics {
    colvars           r
    TimeStepFactor  2       # MTS
    hillWidth         3.0
    hillWeight        0.05
    wellTempered      on
    biasTemperature   4000
}
harmonicWalls {
    colvars           r
    TimeStepFactor  2           # MTS
    lowerWalls        10.3
    upperWalls        34.3
    lowerWallConstant 0.2
    upperWallConstant 0.2
}
# Conformational restraints
harmonic {
    colvars         RMSD
    TimeStepFactor  2            # MTS
    forceConstant   10.0
    centers         0.0
}
# Orientational angles restraints
harmonic {
    colvars         eulerTheta
    TimeStepFactor  2             # MTS
    forceConstant   0.1
    centers         0.0
}
harmonic {
    colvars         eulerPhi
    TimeStepFactor  2             # MTS
    forceConstant   0.1
    centers         6.0
}
harmonic {
    colvars         eulerPsi
    TimeStepFactor  2             # MTS
    forceConstant   0.1
    centers         6.0
}
# Positional angles restraints
harmonic {
    colvars         polarTheta
    TimeStepFactor  2              # MTS
    forceConstant   0.1
    centers         127.0
}
harmonic {
    colvars         polarPhi
    TimeStepFactor  2               # MTS
    forceConstant   0.1
    centers         -75.0
}
colvar {
  name translation
  TimeStepFactor  2                # MTS
  distance {
```

```
    group1 {
      indexGroup  protein
    }
    group2 {
      dummyAtom (-0.49201327562332153, 0.07457345724105835, 0.7762426137924194)
    }
  }
}
harmonic {
  colvars       translation
  TimeStepFactor  2                   # MTS
  centers       0.0
  forceConstant 100.0
}

colvar {
  name orientation
  TimeStepFactor  2                    # MTS
  orientation {
    atoms {
      indexGroup  protein
    }
    refPositionsFile   ../../complex_largeBox.xyz
  }
}
harmonic {
  colvars       orientation
  TimeStepFactor  2                     # MTS
  centers       (1.0, 0.0, 0.0, 0.0)
  forceConstant 2000.0
}
```

Bibliography

[1] S. Batool et al. “Synapse formation: from cellular and molecular mechanisms to neurodevelopmental and neurodegenerative disorders”. In: *J. Neurophysiol.* 121.4 (Apr. 2019), pp. 1381–1397. DOI: 10.1152/jn.00833.2018.

[2] T.C. Südhof. “Towards an Understanding of Synapse Formation”. In: *Neuron* 100.2 (Oct. 2018), pp. 276–293. DOI: 10.1016/j.neuron.2018.09.040.

[3] T.C. Südhof. “The cell biology of synapse formation”. In: *J. Cell. Biol.* 220.7 (June 2021), e202103052. DOI: 10.1083/jcb.202103052.

[4] R. W. Sperry. “Chemoaffinity in the orderly growth of nerve fiber patterns and connections”. In: *Proc. Natl. Acad. Sci.* 50.4 (Oct. 1963), pp. 703–710. DOI: 10.1073/pnas.50.4.703.

[5] E. Özkan et al. “An extracellular interactome of immunoglobulin and LRR proteins reveals receptor-ligand networks”. In: *Cell* 154.1 (July 2013), pp. 228–239. DOI: 10.1016/j.cell.2013.06.006.

[6] J. Brasch et al. “Homophilic and heterophilic interactions of type II cadherins identify specificity groups underlying cell-adhesive behavior”. In: *Cell Reports* 23.6 (May 2018), pp. 1840–1852. DOI: 10.1016/j.celrep.2018.04.012.

[7] K.M. Goodman et al. “Molecular basis of sidekick-mediated cell-cell adhesion and specificity”. In: *eLife* 5 (Sept. 2016), e19058. DOI: 10.7554/eLife.19058.

[8] F. Cosmanescu et al. “Neuron-subtype-specific expression, interaction affinities, and specificity determinants of DIP/Dpr cell recognition proteins”. In: *Neuron* 100.6 (Dec. 2018), 1385–1400.e6. DOI: 10.1016/j.neuron.2018.10.046.

[9] S. Cheng et al. “Molecular basis of synaptic specificity by immunoglobulin superfamily receptors in Drosophila”. In: *eLife* 8 (Jan. 2019), e41028. DOI: 10.7554/eLife.41028.

[10] X. Meng et al. “Molecular Docking: A powerful approach for structure-based drug discovery”. In: *Curr. Comput. Aided Drug Des.* 7.2 (June 2011), pp. 146–157.

[11] P. Nandigrami et al. “Computational assessment of protein–protein binding specificity within a family of synaptic surface receptors”. In: *J. Phys. Chem. B* (July 2022). DOI: 10.1021/acs.jpcb.2c02173.

[12] T. Siebenmorgen and M. Zacharias. “Computational prediction of protein–protein binding affinities”. In: *Wiley Interdiscip. Rev. Comput. Mol. Sci* 10.3 (2020), e1448. DOI: 10.1002/wcms.1448.

[13] Alexey V. Onufriev and David A. Case. “Generalized Born implicit solvent models for biomolecules”. In: *Annu. Rev. Biophys.* 48.1 (May 2019), pp. 275–296. DOI: 10.1146/annurev-biophys-052118-115325.

[14] J. Srinivasan et al. “Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices”. In: *J. Am. Chem. Soc.* 120.37 (Sept. 1998), pp. 9401–9409. DOI: 10.1021/ja981844+.

[15] W. Im, D. Beglov, and B. Roux. “Continuum solvation model: computation of electrostatic forces from numerical solutions to the Poisson-Boltzmann equation”. In: *Comput. Phys. Commun.* 111.1 (June 1998), pp. 59–75. DOI: 10.1016/S0010-4655(98)00016-2.

[16] M.S. Lee and M.A. Olson. “Calculation of absolute protein-ligand binding affinity using path and endpoint approaches”. In: *Biophys. J.* 90.3 (Feb. 2006), pp. 864–877. DOI: 10.1529/biophysj.105.071589.

[17] N. Homeyer and H. Gohlke. “Free Energy Calculations by the Molecular Mechanics Poisson-Boltzmann Surface Area Method”. In: *Molecular Informatics* 31.2 (2012), pp. 114–122. DOI: 10.1002/minf.201100135.

[18] C. Wang et al. “Recent Developments and Applications of the MMPBSA Method”. In: *Front. Mol. Biosci.* 4 (2018). DOI: 10.3389/fmolb.2017.00087.

[19] J. Swanson, R. Henchman, and J. McCammon. “Revisiting Free Energy Calculations: A Theoretical Connection to MM/PBSA and Direct Calculation of the Association Free Energy”. In: *Biophys. J.* 86.1 (Jan. 2004), pp. 67–74. DOI: 10.1016/S0006-3495(04)74084-9.

[20] D.A. Pearlman. “Evaluating the Molecular Mechanics Poisson-Boltzmann Surface Area Free Energy Method Using a Congeneric Series of Ligands to p38 MAP Kinase”. In: *J. Med. Chem.* 48.24 (Dec. 2005), pp. 7796–7807. DOI: 10.1021/jm050306m.

[21] C. Jarzynski. “Nonequilibrium Equality for Free Energy Differences”. In: *Phys. Rev. Lett.* 78.14 (Apr. 1997), pp. 2690–2693. DOI: 10.1103/PhysRevLett.78.2690.

[22] S. Park et al. “Free energy calculation from steered molecular dynamics simulations using Jarzynski’s equality”. In: *J. Chem. Phys.* 119.6 (Aug. 2003), pp. 3559–3566. DOI: 10.1063/1.1590311. (Visited on 05/11/2022).

[23] T. Baştuğ et al. “Potential of mean force calculations of ligand binding to ion channels from Jarzynski’s equality and umbrella sampling”. In: *J. Chem. Phys.* 128.15 (Apr. 2008), p. 155104. DOI: 10.1063/1.2904461.

[24] L. Yang et al. “Steered Molecular Dynamics Simulations Reveal the Likelier Dissociation Pathway of Imatinib from Its Targeting Kinases c-Kit and Abl”. In: *PLOS ONE* 4.12 (Dec. 2009), e8470. DOI: 10.1371/journal.pone.0008470.

[25] J.C. Gumbart, B. Roux, and C. Chipot. “Standard Binding Free Energies from Computer Simulations: What is the Best Strategy?” In: *J. Chem. Theory Comput.* 9.23794960 (Jan. 2013), pp. 794–802. DOI: 10.1021/ct3008099.

[26] G. Heinzelmann, N.M. Henriksen, and M.K. Gilson. “Attach-Pull-Release Calculations of Ligand Binding and Conformational Changes on the First BRD4 Bromodomain”. In: *J. Chem. Theory Comput.* 13.7 (July 2017), pp. 3260–3275. DOI: 10.1021/acs.jctc.7b00275. (Visited on 02/28/2022).

[27] V. Limongelli, M. Bonomi, and M. Parrinello. “Funnel Metadynamics as Accurate Binding Free-Energy Method”. In: *Proc. Natl. Acad. Sci.* 110.16 (Apr. 2013), pp. 6358–6363. DOI: 10.1073/pnas.1303186110.

[28] S. Raniolo and V. Limongelli. “Ligand binding free-energy calculations with funnel metadynamics.” In: *Nat. Protoc.* 15 (9 Sept. 2020), pp. 2837–2866. DOI: 10.1038/s41596-020-0342-4.

[29] J. Jiménez et al. "KDEEP: protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks". In: *J. Chem. Inf. Model.* 58.2 (Feb. 2018), pp. 287–296. DOI: 10.1021/acs.jcim.7b00650.

[30] J. Jiménez et al. "DeepSite: protein-binding site predictor using 3D-convolutional neural networks". In: *Bioinformatics* 33.19 (Oct. 2017), pp. 3036–3042. DOI: 10.1093/bioinformatics/btx350.

[31] H. Hassan-Harrirou, C. Zhang, and T. Lemmin. "RosENet: improving binding affinity prediction by leveraging molecular mechanics energies with an ensemble of 3D convolutional neural networks". In: *J. Chem. Inf. Model.* 60.6 (June 2020), pp. 2791–2802. DOI: 10.1021/acs.jcim.0c00075.

[32] G. Bitencourt-Ferreira and W.F. de Azevedo. "Development of a machine-learning model to predict Gibbs free energy of binding for protein-ligand complexes". In: *Biophysical Chemistry* 240 (Sept. 2018), pp. 63–69. DOI: 10.1016/j.bpc.2018.05.010.

[33] J. Gebhardt et al. "Combining molecular dynamics and machine learning to predict self-solvation free energies and limiting activity coefficients". In: *J. Chem. Inf. Model.* 60.11 (Nov. 2020), pp. 5319–5330. DOI: 10.1021/acs.jcim.0c00479.

[34] S. Riniker. "Molecular Dynamics Fingerprints (MDFP): machine learning from MD data to predict free-energy differences". In: *J. Chem. Inf. Model.* 57.4 (Apr. 2017), pp. 726–741. DOI: 10.1021/acs.jcim.6b00778.

[35] L. Breiman. "Random Forests". In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[36] Y. Wang and I.H. Witten. *Induction of model trees for predicting continuous classes*. Working Paper. 1996. URL: https://hdl.handle.net/10289/1183.

[37] J.H. Friedman. "Multivariate Adaptive Regression Splines". In: *The Annals of Statistics* 19.1 (1991), pp. 1–67. DOI: 10.1214/aos/1176347963.

[38] R.L. Hardy. "Multiquadric equations of topography and other irregular surfaces". In: *J. Geophys. Res.* 76.8 (1971), pp. 1905–1915. DOI: 10.1029/JB076i008p01905.

[39] I.H. Moal, R. Agius, and P.A. Bates. "Protein–protein binding affinity prediction on a diverse set of structures". In: *Bioinformatics* 27.21 (Nov. 2011), pp. 3002–3009. DOI: 10.1093/bioinformatics/btr513.

[40] A. Vangone and A. Bonvin. "Contacts-based prediction of binding affinity in protein–protein complexes". In: *eLife* 4 (July 2015), e07454. DOI: 10.7554/eLife.07454.

[41] R. Casadio, P.L. Martelli, and C. Savojardo. "Machine learning solutions for predicting protein–protein interactions". In: *Wiley Interdiscip. Rev. Comput. Mol. Sci* 12.6 (2022), e1618. DOI: 10.1002/wcms.1618.

[42] S. Das and S. Chakrabarti. "Classification and prediction of protein–protein interaction interface using machine learning algorithm". In: *Sci. Rep.* 11.1 (Jan. 2021), p. 1761. DOI: 10.1038/s41598-020-80900-2.

[43] A. Pavlova et al. "Machine learning reveals the critical interactions for SARS-CoV-2 Spike protein binding to ACE2". In: *J. Phys. Chem. Lett.* 12.23 (June 2021), pp. 5494–5502. DOI: 10.1021/acs.jpclett.1c01494.

[44] H. Woo and B. Roux. "Calculation of absolute protein–ligand binding free energy from computer simulations". In: *Proc. Natl. Acad. Sci.* 102.19 (May 2005), pp. 6825–6830. DOI: 10.1073/pnas.0409005102.

[45] H. C. Andersen. "RATTLE: a "velocity" version of the Shake algorithm for molecular dynamics calculations". In: *J. Comput. Phys.* 52 (Oct. 1983), pp. 24–34. DOI: 10.1016/0021-9991(83)90014-1.

[46] J. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. "Numerical Integration of the Cartesian Equations of Motion for a System with Constraints: Molecular Dynamics of n–Alkanes". In: *J. Comput. Phys.* 23 (1977), pp. 327–341. DOI: https://doi.org/10.1016/0021-9991(77)90098-5.

[47] C.W. Hopkins et al. "Long-time-step molecular dynamics through hydrogen mass repartitioning". In: *J. Chem. Theory Comput.* 11.4 (Apr. 2015), pp. 1864–1874. DOI: 10.1021/ct5010406.

[48] K. Vanommeslaeghe et al. "CHARMM General Force Field (CGenFF): A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields". In: *J Comput Chem* 31.4 (Mar. 2010), pp. 671–690. DOI: 10.1002/jcc.21367.

[49] H. Fu et al. "Taming Rugged Free Energy Landscapes Using an Average Force". In: *Acc. Chem. Res.* 52.11 (Nov. 2019), pp. 3254–3264. DOI: 10.1021/acs.accounts.9b00473.

[50] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. "Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method". In: *Phys. Rev. Lett.* 100.2 (Jan. 2008), p. 020603. DOI: 10.1103/PhysRevLett.100.020603.

[51] H. Fu et al. "Extended adaptive biasing force algorithm. An on-the-fly implementation for accurate free-energy calculations". In: *J. Chem. Theory Comput.* 12.8 (Aug. 2016), pp. 3506–3513. DOI: 10.1021/acs.jctc.6b00447.

[52] A. Laio et al. "Assessing the Accuracy of Metadynamics". In: *J. Phys. Chem. B* 109.14 (Apr. 2005), pp. 6714–6721. DOI: 10.1021/jp045424k.

[53] D. Min et al. "On the convergence improvement in the metadynamics simulations: A Wang-Landau recursion approach". In: *J. Chem. Phys.* 126.19 (May 2007), p. 194104. DOI: 10.1063/1.2731769.

[54] A. Lesage et al. "Smoothed biasing forces yield unbiased free energies with the extended-system adaptive biasing force method". In: *J. Phys. Chem. B* 121.15 (Apr. 2017), pp. 3676–3685. DOI: 10.1021/acs.jpcb.6b10055.

[55] J. Comer et al. "The Adaptive Biasing Force Method: Everything You Always Wanted To Know but Were Afraid To Ask". In: *J. Phys. Chem. B* 119.3 (Jan. 2015), pp. 1129–1151. DOI: 10.1021/jp506633n.

[56] D. Shoup and A. Szabo. "Role of diffusion in ligand binding to macromolecules and cell-bound receptors". In: *Biophys. J.* 40.1 (Oct. 1982), pp. 33–39. DOI: 10.1016/S0006-3495(82)84455-X.

[57] J.C. Gumbart, B. Roux, and C. Chipot. "Efficient determination of protein-protein standard binding free energies from first principles". In: *J. Chem. Theory Comput.* 9.8 (Aug. 2013), 10.1021/ct400273t. DOI: 10.1021/ct400273t.

[58] J. Wang, Y. Deng, and B. Roux. "Absolute Binding Free Energy Calculations Using Molecular Dynamics Simulations with Restraining Potentials". In: *Biophys. J.* 91.8 (Oct. 2006), pp. 2798–2814. DOI: 10.1529/biophysj.106.084301.

[59] J. Hermans and L. Wang. "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". In: *J. Am. Chem. Soc.* 119.11 (Mar. 1997), pp. 2707–2714. DOI: 10.1021/ja963568+.

[60] Y. Deng and B. Roux. "Computation of binding free energy with molecular dynamics and grand canonical Monte Carlo simulations". In: *J. Chem. Phys.* 128.11 (Mar. 2008), p. 115103. DOI: 10.1063/1.2842080.

[61] J. C. Gumbart, B. Roux, and C. Chipot. "Protein:ligand standard binding free energies: A tutorial for alchemical and geometrical transformations". In: *tutorial, www.ks.uiuc.edu* (2018).

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

[63] C. Chipot, P.A. Kollman, and D.A. Pearlman. "Alternative Approaches to Potential of Mean Force Calculations: Free Energy Perturbation versus Thermodynamic Integration. Case Study of Some Representative Nonpolar Interactions". In: *J. Comput. Chem.* 17.9 (July 1996), pp. 1112–1131. DOI: 10.1002/(SICI)1096-987X(19960715)17:9<1112::AID-JCC4>3.0.CO;2-V.

[64] R.W. Zwanzig. "High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases". In: *J. Chem. Phys.* 22.8 (Aug. 1954), pp. 1420–1426. DOI: 10.1063/1.1740409.

[65] C.H. Bennett. "Efficient estimation of free energy differences from Monte Carlo data". In: *J. Comput. Phys.* 22.2 (Oct. 1976), pp. 245–268. DOI: 10.1016/0021-9991(76)90078-4.

[66] Nandou Lu, David A. Kofke, and Thomas B. Woolf. "Staging Is More Important than Perturbation Method for Computation of Enthalpy and Entropy Changes in Complex Systems". In: *J. Phys. Chem. B* 107.23 (2003), pp. 5598–5611. DOI: 10.1021/jp027627j.

[67] N. Lu, D.A. Kofke, and T.B. Woolf. "Improving the efficiency and reliability of free energy perturbation calculations using overlap sampling methods". In: *J. Comput. Chem.* 25.1 (2004), pp. 28–40. DOI: 10.1002/jcc.10369.

[68] Christophe Chipot. "Free Energy Methods for the Description of Molecular Processes". In: *Ann. Rev. Biophys.* 52 (2023), pp. 3.1–3.26.

[69] M. Barker and W. Rayens. "Partial least squares for discrimination". In: *J. Chemom.* 17.3 (Mar. 2003), pp. 166–173. DOI: 10.1002/cem.785.

[70] J. Zhang and L. Kurgan. "SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences". In: *Bioinformatics* 35.14 (July 2019), pp. i343–i353. DOI: 10.1093/bioinformatics/btz324.

[71] D. Kihara et al. "Molecular surface representation using 3D Zernike descriptors for protein shape comparison and docking". In: *Curr Protein Pept Sci* 12.6 (Sept. 2011), pp. 520–530. DOI: 10.2174/138920311796957612.

[72] S. Yin et al. "Fast screening of protein surfaces using geometric invariant fingerprints". In: *Proc. Natl. Acad. Sci.* 106.39 (Sept. 2009), pp. 16622–16626. DOI: 10.1073/pnas.0906146106.

[73] J.D. Thompson, D.G. Higgins, and T.J. Gibson. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". In: *Nucleic Acids Res.* 22.22 (Nov. 1994), pp. 4673–4680. DOI: 10.1093/nar/22.22.4673.

[74] S. Cheng et al. "Family of neural wiring receptors in bilaterians defined by phylogenetic, biochemical, and structural evidence". In: *Proc. Natl. Acad. Sci.* 116.20 (May 2019), pp. 9837–9842. DOI: 10.1073/pnas.1818631116.

[75] T. Hastie, R. Tibshirani, and J. Friedman. "Random Forests". In: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer New York, 2009, pp. 587–604. ISBN: 978-0-387-84858-7. DOI: 10.1007/978-0-387-84858-7_15.

[76] G.J. McLachlan. "Mahalanobis distance". In: *Reson* 4.6 (June 1999), pp. 20–26. DOI: 10.1007/BF02834632.

[77] O. Ledoit and M. Wolf. "Honey, I Shrunk the Sample Covariance Matrix". In: *UPF Economics and Business Working Paper* 691 (June 2003). DOI: 10.2139/ssrn.433840.

[78] F. Pedregosa et al. "Scikit-learn: machine learning in Python". In: *J. Mach. Learn. Res.* 12 (Oct. 2011), pp. 2825–2830.

[79] T.K. Ho. "Random decision forests". In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 1. 1995, 278–282 vol.1. DOI: 10.1109/ICDAR.1995.598994.

[80] T. Hastie, R. Tibshirani, and J. Friedman. "Random Forests". In: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer New York, 2009, pp. 587–604. ISBN: 978-0-387-84858-7. DOI: 10.1007/978-0-387-84858-7_15.

[81] H. Fu et al. "BFEE2: Automated, Streamlined, and Accurate Absolute Binding Free-Energy Calculations". In: *J. Chem. Inf. Model.* 61.5 (May 2021), pp. 2116–2123. DOI: 10.1021/acs.jcim.1c00269.

[82] H. Fu et al. "Accurate determination of protein:ligand standard binding free energies from molecular dynamics simulations". In: *Nat Protoc* (Mar. 2022), pp. 1–28. DOI: 10.1038/s41596-021-00676-1.

[83] A. Morton and B.W. Matthews. "Specificity of ligand binding in a buried nonpolar cavity of T4 lysozyme: Linkage of dynamics and structural plasticity". In: *Biochemistry* 34.27 (July 1995), pp. 8576–8588. DOI: 10.1021/bi00027a007.

[84] J. Huang and A.D. MacKerell. "CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data". In: *J. Comput. Chem.* 34.25 (2013), pp. 2135–2145. DOI: 10.1002/jcc.23354.

[85] P. Mark and L. Nilsson. "Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K". In: *J. Phys. Chem. A* 105.43 (Nov. 2001), pp. 9954–9960. DOI: 10.1021/jp003020w.

[86] J.C. Phillips et al. "Scalable molecular dynamics on CPU and GPU architectures with NAMD". In: *J. Chem. Phys.* 153.4 (July 2020), p. 044130. DOI: 10.1063/5.0014475.

[87] A. Morton, W.A. Baase, and B.W. Matthews. "Energetic origins of specificity of ligand binding in an interior nonpolar cavity of T4 lysozyme". In: *Biochemistry* 34.27 (July 1995), pp. 8564–8575. DOI: 10.1021/bi00027a006.

[88] P. Liu et al. "A Toolkit for the Analysis of Free-Energy Perturbation Calculations". In: *J. Chem. Theory Comput.* 8.8 (Aug. 2012), pp. 2606–2616. DOI: 10.1021/ct300242f.

[89] Y. Deng and B. Roux. "Calculation of Standard Binding Free Energies: Aromatic Molecules in the T4 Lysozyme L99A Mutant". In: *J. Chem. Theory Comput.* 2.5 (Sept. 2006), pp. 1255–1273. DOI: 10.1021/ct060037v.

[90] W. Humphrey, A. Dalke, and K. Schulten. "VMD – Visual Molecular Dynamics". In: *Journal of Molecular Graphics* 14 (1996), pp. 33–38.

[91] H.M. Berman et al. "The Protein Data Bank". In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 235–242. DOI: 10.1093/nar/28.1.235.

[92] S. Pronk et al. “GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit”. In: *Bioinformatics* 29.7 (Apr. 2013), pp. 845–854. DOI: 10.1093/bioinformatics/btt055.

[93] G. Fiorin, M.L. Klein, and J. Hénin. “Using collective variables to drive molecular dynamics simulations”. In: *Mol. Phys.* 111.22-23 (Dec. 2013), pp. 3345–3362. DOI: 10.1080/00268976.2013.813594.

[94] WHO. *Tracking SARS-CoV-2 variants*. accessed January 28, 2022. URL: https://www.who.int/health-topics/typhoid/tracking-SARS-CoV-2-variants.

[95] CDC. *Coronavirus Disease 2019 (COVID-19)*. en-us. accessed February 21, 2022. Feb. 2020. URL: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html.

[96] M.A. Tortorici et al. “Ultrapotent human antibodies protect against SARS-CoV-2 challenge via multiple mechanisms”. In: *Science* 370.6519 (Nov. 2020), pp. 950–957. DOI: 10.1126/science.abe3354.

[97] M. McCallum et al. “Molecular basis of immune evasion by the Delta and Kappa SARS-CoV-2 variants”. In: *Science* (Nov. 2021). DOI: 10.1126/science.abl8506.

[98] J. Lan et al. “Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor”. In: *Nature* 581.7807 (May 2020), pp. 215–220. DOI: 10.1038/s41586-020-2180-5.

[99] P. Han et al. “Molecular insights into receptor binding of recent emerging SARS-CoV-2 variants”. In: *Nat Commun* 12.1 (Oct. 2021), p. 6103. DOI: 10.1038/s41467-021-26401-w.

[100] Rungtiwa Nutalai et al. “Potent cross-reactive antibodies following Omicron breakthrough in vaccinees”. In: *Cell* 185.12 (June 2022), 2116–2131.e18. DOI: 10.1016/j.cell.2022.05.014.

[101] T.N. Starr et al. “SARS-CoV-2 RBD antibodies that maximize breadth and resistance to escape”. In: *Nature* 597.7874 (Sept. 2021), pp. 97–102. DOI: 10.1038/s41586-021-03807-6.

[102] J. Huo et al. “Neutralizing nanobodies bind SARS-CoV-2 spike RBD and block interaction with ACE2”. In: *Nat Struct Mol Biol* 27.9 (Sept. 2020), pp. 846–854. DOI: 10.1038/s41594-020-0469-6.

[103] T. Yang et al. “Effect of SARS-CoV-2 B.1.1.7 mutations on spike protein structure and function”. In: *Nat. Struct. Mol. Biol.* (Aug. 2021). DOI: 10.1038/s41594-021-00652-z.

[104] D. Mannar et al. “Structural analysis of receptor binding domain mutations in SARS-CoV-2 variants of concern that modulate ACE2 and antibody binding”. In: *Cell Reports* 37.12 (Dec. 2021). DOI: 10.1016/j.celrep.2021.110156.

[105] N. Bhattarai et al. “Structural and Dynamical Differences in the Spike Protein RBD in the SARS-CoV-2 Variants B.1.1.7 and B.1.351”. In: *J. Phys. Chem. B* (June 2021). DOI: 10.1021/acs.jpcb.1c01626.

[106] C. Wang et al. “A novel coronavirus outbreak of global health concern”. In: *Lancet* 395.10223 (2020), pp. 470–473. DOI: 10.1016/S0140-6736(20)30185-9. (Visited on 01/06/2022).

[107] R. Horton. “Offline: COVID-19 is not a pandemic”. In: *The Lancet* 396.10255 (Sept. 2020), p. 874. DOI: 10.1016/S0140-6736(20)32000-6.

[108] WHO. *WHO Coronavirus (COVID-19) Dashboard*. accessed January 28, 2022. URL: https://covid19.who.int.

[109] C.B. Jackson et al. "Mechanisms of SARS-CoV-2 entry into cells". In: *Nat Rev Mol Cell Biol* 23.1 (Jan. 2022), pp. 3–20. DOI: 10.1038/s41580-021-00418-x.

[110] H. Yao et al. "Molecular Architecture of the SARS-CoV-2 Virus". In: *Cell* 183.3 (Oct. 2020), 730–738.e13. DOI: 10.1016/j.cell.2020.09.018.

[111] A.C. Walls et al. "Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein". In: *Cell* 181.2 (Apr. 2020), 281–292.e6. DOI: 10.1016/j.cell.2020.02.058.

[112] J. Shang et al. "Structural basis of receptor recognition by SARS-CoV-2". In: *Nature* 581.7807 (May 2020), pp. 221–224. DOI: 10.1038/s41586-020-2179-y.

[113] M. Hoffmann, H. Kleine-Weber, and S. Pöhlmann. "A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells". In: *Molecular Cell* 78.4 (May 2020), 779–784.e5. DOI: 10.1016/j.molcel.2020.04.022.

[114] D.J. Benton et al. "Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion". In: *Nature* 588.7837 (Dec. 2020), pp. 327–330. DOI: 10.1038/s41586-020-2772-0. (Visited on 05/25/2022).

[115] C. Liu et al. "Research and Development on Therapeutic Agents and Vaccines for COVID-19 and Related Human Coronavirus Diseases". In: *ACS Cent. Sci.* 6.3 (Mar. 2020), pp. 315–331. DOI: 10.1021/acscentsci.0c00272. (Visited on 05/25/2022).

[116] E. Volkan. "COVID-19: Structural Considerations for Virus Pathogenesis, Therapeutic Strategies and Vaccine Design in the Novel SARS-CoV-2 Variants Era". In: *Mol Biotechnol* (June 2021), pp. 885–897. DOI: 10.1007/s12033-021-00353-4.

[117] S.M. Gobeil et al. "Effect of natural mutations of SARS-CoV-2 on spike structure, conformation, and antigenicity". In: *Science* 373.6555 (Aug. 2021), eabi6226. DOI: 10.1126/science.abi6226.

[118] J. Zhang et al. "Membrane fusion and immune evasion by the spike protein of SARS-CoV-2 Delta variant". In: *Science* 374.6573 (2021), pp. 1353–1360. DOI: 10.1126/science.abl9463.

[119] A.L. Alaofi and M. Shahid. "Mutations of SARS-CoV-2 RBD May Alter Its Molecular Structure to Improve Its Infection Efficiency". In: *Biomolecules* 11.9 (Aug. 2021), p. 1273. DOI: 10.3390/biom11091273.

[120] W.T. Harvey et al. "SARS-CoV-2 variants, spike mutations and immune escape". In: *Nat. Rev. Microbiol.* 19.7 (July 2021), pp. 409–424. DOI: 10.1038/s41579-021-00573-0.

[121] B. Luan and T. Huynh. "Insights into SARS-CoV-2's mutations for evading human antibodies: sacrifice and survival". In: *J. Med. Chem.* 65.4 (Feb. 2022), pp. 2820–2826. DOI: 10.1021/acs.jmedchem.1c00311.

[122] C. Chen et al. "Computational prediction of the effect of amino acid changes on the binding affinity between SARS-CoV-2 spike RBD and human ACE2". In: *Proc. Natl. Acad. Sci.* 118.42 (Oct. 2021). DOI: 10.1073/pnas.2106480118.

[123] F. Fratev. "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". In: *J. Chem. Inf. Model.* 61.12 (Nov. 2021), pp. 6079–6084. DOI: 10.1021/acs.jcim.1c01242.

[124] B.O. Villoutreix et al. "In silico investigation of the new UK (B.1.1.7) and South African (501Y.V2) SARS-CoV-2 variants with a focus at the ACE2–Spike RBD interface". In: *Int. J. Mol. Sci.* 22.4 (Jan. 2021), p. 1695. DOI: 10.3390/ijms22041695.

[125] J.D. Chodera et al. "Alchemical free energy methods for drug discovery: progress and challenges". In: *Curr. Opin. Struct. Biol.* 21.2 (2011), pp. 150–160. DOI: https://doi.org/10.1016/j.sbi.2011.01.011.

[126] A. Pohorille, C. Jarzynski, and C. Chipot. "Good practices in free-energy calculations". In: *J. Phys. Chem. B* 114.32 (Aug. 2010), pp. 10235–10253. DOI: 10.1021/jp102971x.

[127] W.L. Jorgensen and L.L. Thomas. "Perspective on free-energy perturbation calculations for chemical equilibria". In: *J. Chem. Theory. Comput.* 4.6 (June 2008), pp. 869–876. DOI: 10.1021/ct800011m.

[128] W.L. Jorgensen and C. Ravimohan. "Monte Carlo simulation of differences in free energies of hydration". In: *J. Comput. Phys.* 83.6 (Sept. 1985), pp. 3050–3054. DOI: 10.1063/1.449208.

[129] M.R. Smaoui and H. Yahyaoui. "Unraveling the stability landscape of mutations in the SARS-CoV-2 receptor-binding domain". In: *Sci. Rep.* 11.1 (Apr. 2021), p. 9166. DOI: 10.1038/s41598-021-88696-5.

[130] J. Zou et al. "Computational prediction of mutational effects on SARS-CoV-2 binding by relative free energy calculations". In: *J. Chem. Inf. Model.* 60.12 (Dec. 2020), pp. 5794–5802. DOI: 10.1021/acs.jcim.0c00679.

[131] S. Izrailev et al. *Steered Molecular Dynamics*. Ed. by P. Deuflhard et al. Lecture Notes in Computational Science and Engineering. Berlin, Heidelberg: Springer, 1999, pp. 39–65. DOI: 10.1007/978-3-642-58360-5_2.

[132] S. Kim et al. "Differential interactions between human ACE2 and Spike RBD of SARS-CoV-2 variants of concern". In: *J. Chem. Theory Comput.* 17.12 (Dec. 2021), pp. 7972–7979. DOI: 10.1021/acs.jctc.1c00965.

[133] M. Koehler et al. "Molecular insights into receptor binding energetics and neutralization of SARS-CoV-2 variants". In: *Nat. Commun.* 12.1 (Nov. 2021), p. 6977. DOI: 10.1038/s41467-021-27325-1.

[134] D. Fotiadis et al. "Imaging and manipulation of biological structures with the AFM". In: *Micron* 33.4 (Jan. 2002), pp. 385–397. DOI: 10.1016/S0968-4328(01)00026-9.

[135] C. García-Iriepa et al. "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". In: *J. Phys. Chem. Lett.* 11.21 (Nov. 2020), pp. 9272–9281. DOI: 10.1021/acs.jpclett.0c02203.

[136] H. Fu et al. "Zooming across the Free-Energy Landscape: Shaving Barriers, and Flooding Valleys". In: *J. Phys. Chem. Lett.* 9.16 (Aug. 2018), pp. 4738–4745. DOI: 10.1021/acs.jpclett.8b01994.

[137] V.A. Ngo and R.K. Jha. "Identifying key determinants and dynamics of SARS-CoV-2/ACE2 tight interaction". In: *PLOS ONE* 16.9 (Sept. 2021), e0257905. DOI: 10.1371/journal.pone.0257905.

[138] S. Chakraborty. "E484K and N501Y SARS-CoV 2 spike mutants increase ACE2 recognition but reduce affinity for neutralizing antibody". In: *Int. Immunopharmacol.* 102 (Jan. 2022), p. 108424. DOI: 10.1016/j.intimp.2021.108424.

[139] G.M. Torrie and J.P. Valleau. "Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling". In: *J. Comput. Phys.* 23.2 (Feb. 1977), pp. 187–199. DOI: 10.1016/0021-9991(77)90121-8.

[140] S. Kumar et al. "The weighted histogram analysis method for free energy calculations on biomolecules. I. The method". In: *J. Comput. Chem.* 13.8 (Feb. 1992), pp. 1011–1021. DOI: 10.1002/jcc.540130812.

[141] M. Souaille and B. Roux. "Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations". In: *Comput. Phys. Commun.* 135.1 (Mar. 2001), pp. 40–57. DOI: 10.1016/S0010-4655(00)00215-0.

[142] E.S. Istifli et al. "Understanding the molecular interaction of SARS-CoV-2 spike mutants with ACE2 (angiotensin converting enzyme 2)". In: *J. Biomol. Struct. Dyn.* 40.23 (Sept. 2021), pp. 12760–12771. DOI: 10.1080/07391102.2021.1975569.

[143] W. Zhou et al. "N439K variant in Spike protein alter the infection efficiency and antigenicity of SARS-CoV-2 based on molecular dynamics simulation". In: *Front. Cell Dev. Biol.* 9 (2021), p. 2071. DOI: 10.3389/fcell.2021.697035.

[144] J. Verma and N. Subbarao. "Insilico study on the effect of SARS-CoV-2 RBD hotspot mutants' interaction with ACE2 to understand the binding affinity and stability". In: *Virology* 561 (Sept. 2021), pp. 107–116. DOI: 10.1016/j.virol.2021.06.009.

[145] A. Khan 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". In: *ChemBioChem* 22.16 (2021), pp. 2641–2649. DOI: 10.1002/cbic.202100191.

[146] D. Xiong et al. "Immune escape mechanisms of SARS-CoV-2 Delta and Omicron variants against two monoclonal antibodies that received emergency use authorization". In: *J. Phys. Chem. Lett.* 13.26 (July 2022), pp. 6064–6073. DOI: 10.1021/acs.jpclett.2c00912.

[147] L. Casalino et al. "Beyond shielding: the roles of glycans in the SARS-CoV-2 Spike protein". In: *ACS Cent. Sci.* 6.10 (Oct. 2020), pp. 1722–1734. DOI: 10.1021/acscentsci.0c01056.

[148] T. Sztain et al. "A glycan gate controls opening of the SARS-CoV-2 Spike protein". In: *Nat. Chem.* 13.10 (Oct. 2021), pp. 963–968. DOI: 10.1038/s41557-021-00758-3.

[149] C. Chakraborty, M. Bhattacharya, and A.R. Sharma. "Present variants of concern and variants of interest of severe acute respiratory syndrome coronavirus 2: Their significant mutations in S-glycoprotein, infectivity, re-infectivity, immune escape and vaccines activity". In: *Rev. Med. Virol.* 32.2 (June 2021), e2270. DOI: 10.1002/rmv.2270.

[150] X. Zhao et al. "Origin of the tight binding mode to ACE2 triggered by multi-point mutations in the Omicron variant: a dynamic insight". In: *Phys. Chem. Chem. Phys* 24.15 (Mar. 2022), pp. 8724–8737. DOI: 10.1039/D2CP00449F.

[151] A. Acharya et al. "ACE2 glycans preferentially interact with SARS-CoV-2 over SARS-CoV". In: *Chem. Commun.* 57.48 (June 2021), pp. 5949–5952. DOI: 10.1039/D1CC02305E.

[152] R. Yan et al. "Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2". In: *Science* 367.6485 (Mar. 2020), pp. 1444–1448. DOI: 10.1126/science.abb2762.

[153] S. Jo et al. "CHARMM-GUI: A web-based graphical user interface for CHARMM". In: *Journal of Computational Chemistry* 29.11 (2008), pp. 1859–1865. DOI: 10.1002/jcc.20945.

[154] Y. Polak and R.C. Speth. "Metabolism of angiotensin peptides by angiotensin converting enzyme 2 (ACE2) and analysis of the effect of excess zinc on ACE2 enzymatic activity". In: *Peptides* 137 (Mar. 2021), p. 170477. DOI: 10.1016/j.peptides.2020.170477.

[155] M.B. Peters et al. "Structural survey of zinc-containing proteins and development of the Zinc AMBER Force Field (ZAFF)". In: *J. Chem. Theory Comput.* 6.9 (Sept. 2010), pp. 2935–2947. DOI: 10.1021/ct1002626.

[156] G.E. Uhlenbeck and L.S. Ornstein. "On the theory of the brownian motion". In: *Phys. Rev.* 36.5 (Sept. 1930), pp. 823–841. DOI: 10.1103/PhysRev.36.823.

[157] S.E. Feller et al. "Constant pressure molecular dynamics simulation: the Langevin piston method". In: *J. Chem. Phys.* 103.11 (Sept. 1995), pp. 4613–4621. DOI: 10.1063/1.470648.

[158] T. Darden, D. York, and L. Pedersen. "Particle Mesh Ewald: An *N* -log(*N*) method for Ewald sums in large systems". In: *Chem. Phys.* 98.12 (June 1993), pp. 10089–10092. DOI: 10.1063/1.464397.

[159] P. Mlcochova et al. "SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion". In: *Nature* (Sept. 2021), pp. 1–8. DOI: 10.1038/s41586-021-03944-y.

[160] V. Edara et al. "Infection and vaccine-induced neutralizing-antibody responses to the SARS-CoV-2 B.1.617 variants". In: *N. Engl. J. Med.* 385.7 (Aug. 2021), pp. 664–666. DOI: 10.1056/NEJMc2107799.

[161] L. Wu et al. "SARS-CoV-2 Omicron RBD shows weaker binding affinity than the currently dominant Delta variant to human ACE2". In: *Sig. Transduct. Target Ther.* 7.1 (Jan. 2022), pp. 1–3. DOI: 10.1038/s41392-021-00863-2.

[162] X. Zhang et al. "SARS-CoV-2 Omicron strain exhibits potent capabilities for immune evasion and viral entrance". In: *Sig Transduct Target Ther* 6.1 (Dec. 2021), pp. 1–3. DOI: 10.1038/s41392-021-00852-5.

[163] S. Kim et al. "Binding of human ACE2 and RBD of Omicron enhanced by unique interaction patterns among SARS-CoV-2 variants of concern". In: *J. Comput. Chem.* 44.4 (Feb. 2023), pp. 594–601. DOI: 10.1002/jcc.27025.

[164] H.L. Nguyen et al. "SARS-CoV-2 Omicron variant binds to human cells more strongly than the wild type: evidence from molecular dynamics Simulation". In: *J. Phys. Chem. B* 126.25 (June 2022), pp. 4669–4678. DOI: 10.1021/acs.jpcb.2c01048.

[165] Q. Hong et al. "Molecular basis of receptor binding and antibody neutralization of Omicron". In: *Nature* 604.7906 (Apr. 2022), pp. 546–552. DOI: 10.1038/s41586-022-04581-9.

[166] Y. Huang et al. "SARS-CoV-2 spike binding to ACE2 is stronger and longer ranged due to glycan interaction". In: *Biophys. J.* 121 (2022), pp. 79–90. DOI: 10.1016/j.bpj.2021.12.002.

[167] D. Wrapp et al. "Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation". In: *Science* 367 (2020), pp. 1260–1263. DOI: 10.1126/science.abb2507.

[168] M. Huang 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". In: *Immunity* (June 2022). DOI: 10.1016/j.immuni.2022.06.005.

[169] E. Socher et al. "Computational decomposition reveals reshaping of the SARS-CoV-2–ACE2 interface among viral variants expressing the N501Y mutation". In: *J. Cell. Biochem.* 122.12 (Dec. 2021), pp. 1863–1872. DOI: 10.1002/jcb.30142.

[170] D. Mannar et al. "SARS-CoV-2 Omicron variant: Antibody evasion and cryo-EM structure of spike protein–ACE2 complex". In: *Science* 375.6582 (Feb. 2022), pp. 760–764. DOI: 10.1126/science.abn7760.

[171] M.I. Barton et al. "Effects of common mutations in the SARS-CoV-2 Spike RBD and its ligand, the human ACE2 receptor on binding affinity and kinetics". In: *eLife* 10 (Aug. 2021), e70658. DOI: 10.7554/eLife.70658.

[172] A. Francés-Monerris et al. “Molecular Basis of SARS-CoV-2 Infection and Rational Design of Potential Antiviral Agents: Modeling and Simulation Approaches”. In: *J. Proteome Res.* 19.11 (Nov. 2020), pp. 4291–4315. DOI: 10.1021/acs.jproteome.0c00779.

[173] S. Park et al. “In different organisms, the mode of interaction between two signaling proteins is not necessarily conserved”. In: *Proc. Natl. Acad. Sci.* 101.32 (2004), pp. 11646–11651. DOI: 10.1073/pnas.0401038101.

[174] J. Lee et al. “CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field”. In: *J. Chem. Theory Comput.* 12.1 (Jan. 2016), pp. 405–413. DOI: 10.1021/acs.jctc.5b00935.

[175] J. Huang et al. “CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins.” eng. In: *Nat. Methods.* 14 (2017), pp. 71–73. DOI: 10.1038/nmeth.4067.

[176] W.L. Jorgensen et al. “Comparison of simple potential functions for simulating liquid water”. In: *Chem. Phys.* 79.2 (July 1983), pp. 926–935. DOI: 10.1063/1.445869.

[177] R.D. Moore and G.A. Morrill. “A Possible Mechanism for Concentrating Sodium and Potassium in the Cell Nucleus”. In: *Biophys. J.* 16.1276381 (May 1976), pp. 527–533. DOI: 10.1016/S0006-3495(76)85707-4.

[178] H. Fu et al. “New Coarse Variables for the Accurate Determination of Standard Binding Free Energies”. In: *J. Chem. Theory Comput.* 13.11 (Nov. 2017), pp. 5173–5178. DOI: 10.1021/acs.jctc.7b00791.

[179] M.T. Pisabarro, L. Serrano, and M. Wilmanns. “Crystal Structure of the Abl-SH3 Domain Complexed with a Designed High-Affinity Peptide Ligand: Implications for SH3-Ligand Interactions”. In: *J. Mol. Biol.* 281.3 (1998), pp. 513–521. DOI: 10.1006/jmbi.1998.1932.

[180] P. Holzer et al. “Discovery of a Dihydroisoquinolinone Derivative (NVP-CGM097): A Highly Potent and Selective MDM2 Inhibitor Undergoing Phase 1 Clinical Trials in p53wt Tumors.” eng. In: *J. Med. Chem.* 58 (Aug. 2015), pp. 6348–6358. DOI: 10.1021/acs.jmedchem.5b00810.

[181] E.N. Baker et al. “The Structure of 2Zn Pig Insulin Crystals at 1.5 Å Resolution”. In: *Phil. Trans. R. Soc. Lond.* 319 (1988), pp. 369–456. DOI: 10.1098/rstb.1988.0058.

[182] V. Zoete, M. Meuwly, and M. Karplus. “Study of the Insulin Dimerization: Binding Free Energy Calculations and per-Residue Free Energy Decomposition.” eng. In: *Proteins* 61 (Oct. 2005), pp. 79–93. DOI: 10.1002/prot.20528.

[183] D. L. Massart et al. Data Handling in Science and Technology: Chapter 14 - Correlation Methods. Amsterdam: Elsevier, 2003. DOI: 10.1016/S0922-3487(08)70227-2.

[184] C.J. Woods et al. “A water-swap reaction coordinate for the calculation of absolute protein–ligand binding free energies”. In: *J. Chem. Phys.* 134.5 (Feb. 2011), p. 054114. DOI: 10.1063/1.3519057.

[185] D.J. Cole, J. Tirado-Rives, and W.L. Jorgensen. “Molecular dynamics and Monte Carlo simulations for protein–ligand binding and inhibitor design”. In: *Biochimica et Biophysica Acta (BBA) - General Subjects*. Recent developments of molecular dynamics 1850.5 (May 2015), pp. 966–971. DOI: 10.1016/j.bbagen.2014.08.018.

[186] M. De Vivo et al. “Role of Molecular Dynamics and Related Methods in Drug Discovery”. In: *J. Med. Chem.* 59.9 (May 2016), pp. 4035–4061. DOI: 10.1021/acs.jmedchem.5b01684.

[187] C. Chipot. "Frontiers in Free-Energy Calculations of Biological Systems". In: *WIREs Comput. Mol. Sci.* 4.1 (Jan. 2014), pp. 71–89. DOI: 10.1002/wcms.1157.

[188] S. Miyamoto and P.A. Kollman. "What determines the strength of noncovalent association of ligands to proteins in aqueous solution?" In: *Proc. Natl. Acad. Sci.* 90.18 (Sept. 1993), pp. 8402–8406. DOI: 10.1073/pnas.90.18.8402.

[189] S.B. Dixit and C. Chipot. "Can Absolute Free Energies of Association Be Estimated from Molecular Mechanical Simulations? The Biotin-Streptavidin System Revisited". In: *J. Phys. Chem. A* 105.42 (Oct. 2001), pp. 9795–9799. DOI: 10.1021/jp011878v.

[190] I. Massova and P.A. Kollman. "Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding". In: *Perspectives in Drug Discovery and Design* 18.1 (June 2000), pp. 113–135. DOI: 10.1023/A:1008763014207.

[191] P.A. Kollman et al. "Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models". In: *Acc. Chem. Res.* 33.12 (Dec. 2000), pp. 889–897. DOI: 10.1021/ar000033j.

[192] H. Gohlke and D.A. Case. "Converging free energy estimates: MM-PB(GB)SA studies on the protein–protein complex Ras–Raf". In: *Journal of Computational Chemistry* 25.2 (2004), pp. 238–250. DOI: 10.1002/jcc.10379.

[193] S. Genheden and U. Ryde. "The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities". In: *Expert Opin Drug Discov* 10.5 (May 2015), pp. 449–461. DOI: 10.1517/17460441.2015.1032936.

[194] B. Isralewitz et al. "Steered molecular dynamics investigations of protein function". In: *Journal of Molecular Graphics and Modelling* 19.1 (Feb. 2001), pp. 13–25. DOI: 10.1016/S1093-3263(00)00133-9.

[195] J. Gu, H. Li, and X. Wang. "A Self-Adaptive Steered Molecular Dynamics Method Based on Minimization of Stretching Force Reveals the Binding Affinity of Protein–Ligand Complexes". In: *Molecules* 20.10 (2015), pp. 19236–19251. DOI: 10.3390/molecules201019236.

[196] A. Potterton et al. "Ensemble-Based Steered Molecular Dynamics Predicts Relative Residence Time of A2A Receptor Binders". In: *J. Chem. Theory Comput.* 15.5 (May 2019), pp. 3316–3330. DOI: 10.1021/acs.jctc.8b01270.

[197] Q. Gong et al. "Calculating the Absolute Binding Free Energy of the Insulin Dimer in an Explicit Solvent". In: *RSC Adv.* 10 (2 2020), pp. 790–800. DOI: 10.1039/C9RA08284K.

[198] A.C. Pan et al. "Atomic-Level Characterization of Protein-Protein Association". In: *Proc. Natl. Acad. Sci.* 116.10 (Feb. 2019), pp. 4244–4249. DOI: 10.1073/pnas.1815431116.

[199] I. Buch, T. Giorgino, and G. De Fabritiis. "Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations". In: *Proc. Natl. Acad. Sci.* 108.25 (2011), pp. 10184–10189. DOI: 10.1073/pnas.1103547108.

[200] F. Noé and S. Fischer. "Transition networks for modeling the kinetics of conformational change in macromolecules". In: *Current Opinion in Structural Biology*. Theory and simulation / Macromolecular assemblages 18.2 (Apr. 2008), pp. 154–162. DOI: 10.1016/j.sbi.2008.01.008.

[201] V.S. Pande, K. Beauchamp, and G.R. Bowman. "Everything you wanted to know about Markov State Models but were afraid to ask". In: *Methods*. Protein Folding 52.1 (Sept. 2010), pp. 99–105. DOI: 10.1016/j.ymeth.2010.06.002.

[202] J. Prinz et al. “Markov models of molecular kinetics: Generation and validation”. In: *J. Chem. Phys.* 134.17 (May 2011), p. 174105. DOI: 10.1063/1.3565032.

[203] B.E. Husic and V.S. Pande. “Markov State Models: From an Art to a Science”. In: *J. Am. Chem. Soc.* 140.7 (Feb. 2018), pp. 2386–2396. DOI: 10.1021/jacs.7b12191.

[204] T. Siebenmorgen and M. Zacharias. “Evaluation of Predicted Protein–Protein Complexes by Binding Free Energy Simulations”. In: *J. Chem. Theory Comput.* 15.3 (Mar. 2019), pp. 2071–2086. DOI: 10.1021/acs.jctc.8b01022.

[205] D.L. Mobley, J.D. Chodera, and K.A. Dill. “Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change”. In: *J. Chem. Theory Comput.* 3.4 (July 2007), pp. 1231–1235. DOI: 10.1021/ct700032n.

[206] W. Yang, R. Bitetti-Putzer, and M. Karplus. “Chaperoned Alchemical Free Energy Simulations: a General Method for QM, MM, and QM/MM Potentials”. In: *J. Chem. Phys.* 120.20 (May 2004), pp. 9450–9453. DOI: 10.1063/1.1738106.

[207] C. Tse et al. “Exploring the Free-Energy Landscape and Thermodynamics of Protein-Protein Association”. In: *Biophys. J.* 119.6 (Sept. 2020), pp. 1226–1238. DOI: 10.1016/j.bpj.2020.08.005.

[208] J. Perthold and C. Oostenbrink. “Simulation of Reversible Protein-Protein Binding and Calculation of Binding Free Energies Using Perturbed Distance Restraints”. In: *J. Chem. Theory Comput.* 13.11 (Nov. 2017), pp. 5697–5708. DOI: 10.1021/acs.jctc.7b00706.

[209] D.C. Joshi and J. Lin. “Delineating Protein-Protein Curvilinear Dissociation Pathways and Energetics with Naïve Multiple-Walker Umbrella Sampling Simulations.” In: *J. Comput. Chem.* 40 (17 June 2019), pp. 1652–1663. DOI: 10.1002/jcc.25821.

[210] Jing Wang et al. “A highly accurate metadynamics-based Dissociation Free Energy method to calculate protein–protein and protein–ligand binding potencies”. In: *Sci Rep* 12.1 (Feb. 2022), p. 2024. DOI: 10.1038/s41598-022-05875-8.

[211] S. Doudou, N.A. Burton, and R.H. Henchman. “Standard Free Energy of Binding from a One-Dimensional Potential of Mean Force”. In: *J. Chem. Theory Comput.* 5.4 (Apr. 2009), pp. 909–918. DOI: 10.1021/ct8002354.

[212] F. Hu et al. “MD Simulation Investigation on the Binding Process of Smoke-Derived Germination Stimulants to Its Receptor”. In: *J. Chem. Inf. Model.* 59.4 (Apr. 2019), pp. 1554–1562. DOI: 10.1021/acs.jcim.8b00844.

[213] J. Patel and F.M. Ytreberg. “Fast Calculation of Protein–Protein Binding Free Energies Using Umbrella Sampling with a Coarse-Grained Model”. In: *J. Chem. Theory Comput.* 14.2 (Feb. 2018), pp. 991–997. DOI: 10.1021/acs.jctc.7b00660.

[214] Y. Niu et al. “Revealing inhibition difference between PFI-2 enantiomers against SETD7 by molecular dynamics simulations, binding free energy calculations and unbinding pathway analysis”. In: *Sci Rep* 7.1 (Apr. 2017), p. 46547. DOI: 10.1038/srep46547.

[215] M. Lapelosa. “Conformational dynamics and free energy of BHRF1 binding to Bim BH3”. In: *Biophysical Chemistry* 232 (Jan. 2018), pp. 22–28. DOI: 10.1016/j.bpc.2017.11.001.

[216] Z. Zuo et al. “Stepwise substrate translocation mechanism revealed by free energy calculations of doxorubicin in the multidrug transporter AcrB”. In: *Sci Rep* 5.1 (Sept. 2015), p. 13905. DOI: 10.1038/srep13905.

[217] X. Jin et al. "Computational study on the inhibition mechanism of a cyclic peptide MaD5 to Pf-MATE: Insight from molecular dynamics simulation, free energy calculation and dynamical network analysis". In: *Chemometrics and Intelligent Laboratory Systems* 149 (Dec. 2015), pp. 81–88. DOI: 10.1016/j.chemolab.2015.10.013.

[218] L. Hernández-Alvarez et al. "Computational study on the allosteric mechanism of Leishmania major IF4E-1 by 4E-interacting protein-1: Unravelling the determinants of m7GTP cap recognition". In: *Computational and Structural Biotechnology Journal* 19 (Jan. 2021), pp. 2027–2044. DOI: 10.1016/j.csbj.2021.03.036.

[219] Yuzhao Zhang et al. "In silico binding profile characterization of SARS-CoV-2 spike protein and its mutants bound to human ACE2 receptor". In: *Briefings in Bioinformatics* 22.6 (Nov. 2021), bbab188. DOI: 10.1093/bib/bbab188.

[220] M.J. Ferrarotti et al. "Accurate multiple time step in biased molecular simulations". In: *J. Chem. Theory Comput.* 11.1 (Jan. 2015), pp. 139–146. DOI: 10.1021/ct5007086.

[221] E.A. Coutsias, C. Seok, and K.A. Dill. "Using quaternions to calculate RMSD". In: *J. Comput. Chem.* 25 (2004), pp. 1849–1857. DOI: 10.1002/jcc.20110.

[222] M. Tuckerman, B.J. Berne, and G.J. Martyna. "Reversible multiple time scale molecular dynamics". In: *J. Chem. Phys.* 97.3 (Mar. 1992), pp. 1990–2001. DOI: 10.1063/1.463137.

[223] J.C. Sexton and D.H. Weingarten. "Hamiltonian evolution for the hybrid Monte Carlo algorithm". In: *Nucl. Phys. B* 380.3 (Aug. 1992), pp. 665–677. DOI: 10.1016/0550-3213(92)90263-B.

[224] M. Blazhynska et al. "Hazardous shortcuts in standard binding free energy calculations". In: *J. Phys. Chem. Lett.* (June 2022), pp. 6250–6258. DOI: 10.1021/acs.jpclett.2c01490.

[225] U. Raucci et al. "Enhanced sampling aided design of molecular photoswitches". In: *J. Am. Chem. Soc.* 144.42 (Oct. 2022), pp. 19265–19271. DOI: 10.1021/jacs.2c04419.

[226] D. Ray et al. "Rare event kinetics from adaptive bias enhanced sampling". In: *J. Chem. Theory Comput.* 18.11 (Oct. 2022), pp. 6500–6509. DOI: 10.1021/acs.jctc.2c00806.

[227] D.L. Mobley and M.K. Gilson. "Predicting binding free energies: frontiers and benchmarks". In: *Annu. Rev. Biophys.* 46 (May 2017), pp. 531–558. DOI: 10.1146/annurev-biophys-070816-033654.

[228] N. Deng et al. "Comparing alchemical and physical pathway methods for computing the absolute binding free energy of charged ligands". In: *Phys. Chem. Chem. Phys.* 20.29896599 (June 2018), pp. 17081–17092. DOI: 10.1039/c8cp01524d.

[229] A. de Ruiter and C. Oostenbrink. "Advances in the calculation of binding free energies". In: *Curr. Opin. Struct. Biol.* 61 (Apr. 2020), pp. 207–212. DOI: 10.1016/j.sbi.2020.01.016.

[230] Maria M. Reif and Martin Zacharias. "Improving the potential of mean force and nonequilibrium pulling simulations by simultaneous alchemical modifications". In: *J. Chem. Theory Comput.* 18.6 (2022), pp. 3873–3893. DOI: 10.1021/acs.jctc.1c01194.

[231] M. Invernizzi and M. Parrinello. "Exploration vs convergence speed in adaptive-bias enhanced sampling". In: *J. Chem. Theory Comput.* 18.6 (May 2022), pp. 3988–3996. DOI: 10.1021/acs.jctc.2c00152.

[232] M. Bertazzo et al. "Machine learning and enhanced sampling simulations for computing the potential of mean force and standard binding free energy". In: *J. Chem. Theory Comput.* 17.8 (July 2021), pp. 5287–5300. DOI: 10.1021/acs.jctc.1c00177.

[233] H. Fu et al. "Meta-analysis reveals that absolute binding free-energy calculations approach chemical accuracy". In: *J. Med. Chem.* 65.19 (Sept. 2022), pp. 12970–12978. DOI: 10.1021/acs.jmedchem.2c00796.

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

[235] D. Hao et al. "How well does the extended linear interaction energy method perform in accurate binding free energy calculations?" In: *J. Chem. Inf. Model.* 60.12 (Nov. 2020), pp. 6624–6633. DOI: 10.1021/acs.jcim.0c00934.

[236] J. Rogal. "Reaction coordinates in complex systems–a perspective". In: *Eur. Phys. J. B* 94.11 (Nov. 2021), p. 223. DOI: 10.1140/epjb/s10051-021-00233-5.

[237] B. Peters. "Reaction coordinates and mechanistic hypothesis tests". In: *Annu. Rev. Phys. Chem.* 67.1 (May 2016), pp. 669–690. DOI: 10.1146/annurev-physchem-040215-112215.

[238] H. Fu et al. "BFEE: a user-friendly graphical interface facilitating absolute binding free-energy calculations". In: *J. Chem. Inf. Model.* 58 (Feb. 2018), pp. 556–560. DOI: 10.1021/acs.jcim.7b00695.

[239] H. Fu and H. Chen. *Binding Free Energy Estimator 2*. https://github.com/fhh2626/BFEE2. Version 2.3.2. Feb. 2023. DOI: 10.5281/zenodo.7623921.

[240] J. Y. Wang. "Abl tyrosine kinase in signal transduction and cell-cycle regulation." eng. In: *Curr. Opin. Genet. Dev.* 3 (1 Feb. 1993), pp. 35–43. DOI: 10.1016/s0959-437x(05)80338-7.

[241] E. Goulard Coderc de Lacam et al. "When the dust has settled: calculation of binding affinities from first principles for SARS-CoV-2 variants with quantitative accuracy". In: *J. Chem. Theory Comput.* 18.10 (Oct. 2022), pp. 5890–5900. DOI: 10.1021/acs.jctc.2c00604.

[242] S. Miyamoto and P. A. Kollman. "SETTLE: an analytical version of the SHAKE and RATTLE algorithms for rigid water models". In: *J. Comput. Chem.* 13 (Oct. 1992), pp. 952–962.

[243] C. Balusek et al. "Accelerating membrane simulations with hydrogen mass repartitioning". In: *J Chem Theory Comput* 15.8 (Aug. 2019), pp. 4673–4686. DOI: 10.1021/acs.jctc.9b00160.

[244] J. Jeckelmann et al. "Structure of the human heterodimeric transporter 4F2hc-LAT2 in complex with Anticalin, an alternative binding protein for applications in single-particle cryo-EM". In: *Sci. Rep.* 12.1 (Oct. 2022), p. 18269. DOI: 10.1038/s41598-022-23270-1.

[245] Y. Pang et al. "SARS-CoV-2 spike opening dynamics and energetics reveal the individual roles of glycans and their collective impact". In: *Commun. Biol.* 5.1 (Nov. 2022), p. 1170. DOI: 10.1038/s42003-022-04138-6.

[246] H. Takeda et al. "A multipoint guidance mechanism for β-barrel folding on the SAM complex". In: *Nat. Struct. Mol. Biol.* 30.2 (Jan. 2023), pp. 176–187. DOI: 10.1038/s41594-022-00897-2.

[247] D. Kalbermatter et al. "Structure and supramolecular organization of the canine distemper virus attachment glycoprotein." eng. In: *Proc. Natl. Acad. Sci.* 120 (6 Feb. 2023), e2208866120. DOI: 10.1073/pnas.2208866120.

[248] H. Chen and C. Chipot. "Enhancing sampling with free-energy calculations". In: *Curr. Opin. Struct. Biol.* 77 (Nov. 2022), p. 102497. DOI: 10.1016/j.sbi.2022.102497.

[249] C. Chipot. "Free energy methods for the description of molecular processes". In: *Ann. Rev. Biophys.* 52 (May 2023), pp. 113–138. DOI: 10.1146/annurev-biophys-062722-093258.

[250] M. Iannuzzi, A. Laio, and M. Parrinello. "Efficient exploration of reactive potential energy surfaces using Car-Parrinello molecular dynamics". In: *Phys. Rev. Lett.* 90.23 (June 2003), p. 238302. DOI: 10.1103/PhysRevLett.90.238302.

[251] T. Lelièvre, M. Rousset, and G. Stoltz. "Computation of free energy profiles with adaptive parallel dynamics". In: *J. Chem. Phys.* 126.13 (Apr. 2007), p. 134111. DOI: 10.1063/1.2711185.

[252] A. Laio and M. Parrinello. "Escaping free-energy minima". In: *Proc. Natl. Acad. Sci.* 99.20 (Oct. 2002), pp. 12562–12566. DOI: 10.1073/pnas.202427399.

[253] J.F. Dama, M. Parrinello, and G.A. Voth. "Well-tempered metadynamics converges asymptotically". In: *Phys. Rev. Lett.* 112 (June 2014), p. 240602. DOI: 10.1103/PhysRevLett.112.240602.

[254] G.O. Roberts and O. Stramer. "Langevin diffusions and Metropolis-Hastings algorithms". In: *Methodol. Comput. Appl. Probab.* 4.4 (Dec. 2002), pp. 337–357. DOI: 10.1023/A:1023562417138.

[255] C.P. Robert et al. "Accelerating MCMC algorithms". In: *WIREs Comp. Stats.* 10.5 (2018), e1435–14. DOI: https://doi.org/10.1002/wics.1435.

[256] M. Betancourt. "The convergence of Markov chain Monte Carlo methods: from the Metropolis method to Hamiltonian Monte Carlo". In: *Ann. Phys. (Berlin)* 531.3 (2019), p. 1700214. DOI: https://doi.org/10.1002/andp.201700214.

[257] R.D. Skeel and C. Hartmann. "Choice of damping coefficient in Langevin dynamics". In: *Eur. Phys. J. B* 94.9 (Sept. 2021), p. 178. DOI: 10.1140/epjb/s10051-021-00182-z.

[258] K.A. Feenstra, B. Hess, and H.J.C. Berendsen. "Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems". In: *J. Comput. Chem.* 20.8 (1999), pp. 786–798. DOI: 10.1002/(SICI)1096-987X(199906)20:8<786::AID-JCC5>3.0.CO;2-B.

[259] A. Palencia et al. "Role of interfacial water molecules in proline-rich ligand recognition by the Src homology 3 domain of Abl." In: *J. Biol. Chem.* 285 (Jan. 2010), pp. 2823–33. DOI: 10.1074/jbc.M109.048033.

[260] A. Zafra-Ruano and I. Luque. "Interfacial water molecules in SH3 interactions: getting the full picture on polyproline recognition by protein-protein interaction domains". In: *FEBS Lett.* 586 (May 2012), pp. 2619–2630. DOI: 10.1016/j.febslet.2012.04.057.

[261] C. Chipot and J. Hénin. "Exploring the free-energy landscape of a short peptide using an average force". In: *J. Chem. Phys.* 123.24 (Dec. 2005), p. 244906. DOI: 10.1063/1.2138694.

[262] J. D. Hunter. "Matplotlib: A 2D Graphics Environment". In: *Comput. Sci. Eng.* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[263] S. Bauer et al. "Pharmacokinetic-pharmacodynamic guided optimisation of dose and schedule of CGM097, an HDM2 inhibitor, in preclinical and clinical studies". In: *Br. J. Cancer* 125 (2021), pp. 687–698. DOI: 10.1038/s41416-021-01444-4.

[264] H. Zhang et al. "CHARMM-GUI Free Energy Calculator for Practical Ligand Binding Free Energy Simulations with AMBER". In: *J. Chem. Inf. Model.* 61.9 (Sept. 2021), pp. 4145–4151. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.1c00747.

[265] P.D. Thomas and K.A. Dill. "An iterative method for extracting energy-like quantities from protein structures." In: *Proc. Natl. Acad. Sci.* 93.21 (Oct. 1996), pp. 11628–11633. DOI: 10.1073/pnas.93.21.11628.

[266] R.I. Dima et al. “Extraction of interaction potentials between amino acids from native protein structures”. In: *J. Chem. Phys.* 112.20 (May 2000), pp. 9151–9166. DOI: 10.1063/1.481525.

[267] Z. Dosztányi et al. “The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins”. In: *J. Mol. Biol.* 347.4 (Apr. 2005), pp. 827–839. DOI: 10.1016/j.jmb.2005.01.071.

[268] M.R. Betancourt and S.J. Omovie. “Pairwise energies for polypeptide coarse-grained models derived from atomic force fields”. In: *J. Chem. Phys.* 130.19 (May 2009), p. 195103. DOI: 10.1063/1.3137045.

[269] M. Winding et al. “The connectome of an insect brain”. In: *Science* 379.6636 (Mar. 2023), eadd9330. DOI: 10.1126/science.add9330.

[270] B.L. Chen, D.H. Hall, and D.B. Chklovskii. “Wiring optimization can relate neuronal structure and function”. In: *Proc. Natl. Acad. Sci.* 103.12 (Mar. 2006), pp. 4723–4728. DOI: 10.1073/pnas.0506806103.

[271] S. Xu et al. “Affinity requirements for control of synaptic targeting and neuronal cell survival by heterophilic IgSF cell adhesion molecules”. In: *Cell Reports* 39.1 (Apr. 2022), p. 110618. DOI: 10.1016/j.celrep.2022.110618.

[272] P. Singh. “SPR biosensors: historical perspectives and current challenges”. In: *Sens. Actuators B: Chem.* 229 (June 2016), pp. 110–130. DOI: 10.1016/j.snb.2016.01.118.

[273] R.F. Alford et al. “The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design”. In: *J. Chem. Theory Comput.* 13.6 (June 2017), pp. 3031–3048. DOI: 10.1021/acs.jctc.7b00125.

[274] R. Chen, L. Li, and Z. Weng. “ZDOCK: an initial-stage protein-docking algorithm”. In: *Proteins* 52.1 (July 2003), pp. 80–87. DOI: 10.1002/prot.10389.

[275] S. Hashemifar et al. “Predicting protein–protein interactions through sequence-based deep learning”. In: *Bioinformatics* 34.17 (Sept. 2018), pp. i802–i810. DOI: 10.1093/bioinformatics/bty573.

[276] S. Conti, V. Ovchinnikov, and M. Karplus. “ppdx: Automated modeling of protein–protein interaction descriptors for use with machine learning”. In: *J. Comput. Chem.* 43.25 (2022), pp. 1747–1757. DOI: 10.1002/jcc.26974.

[277] S. Ovchinnikov, H. Kamisetty, and D. Baker. “Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information”. In: *eLife* 3 (May 2014), e02030. DOI: 10.7554/eLife.02030.

[278] G. van Rossum. *Python tutorial*. Tech. rep. CS-R9526. Amsterdam: Centrum voor Wiskunde en Informatica (CWI), May 1995.

[279] The UniProt Consortium. “UniProt: the universal protein knowledgebase in 2021”. In: *Nucleic Acids Res.* 49.D1 (Jan. 2021), pp. D480–D489. DOI: 10.1093/nar/gkaa1100.

[280] S.F. Altschul et al. “Basic local alignment search tool”. In: *J. Mol. Biol.* 215.3 (Oct. 1990), pp. 403–410. DOI: 10.1016/S0022-2836(05)80360-2.

[281] S. Henikoff and J.G. Henikoff. “Amino acid substitution matrices from protein blocks.” In: *Proc. Natl. Acad. Sci.* 89.22 (Nov. 1992), pp. 10915–10919. DOI: 10.1073/pnas.89.22.10915.

[282] A.P. Sergeeva et al. "DIP/Dpr interactions and the evolutionary design of specificity in protein families". In: *Nat. Commun.* 11.1 (Dec. 2020), p. 2125. DOI: 10.1038/s41467-020-15981-8.

[283] N.S. Tolwinski. "Introduction: Drosophila—A Model System for Developmental Biology". In: *J. Dev. Biol.* 5.3 (Sept. 2017), p. 9. DOI: 10.3390/jdb5030009.

[284] B.H. Jennings. "Drosophila – a versatile model in biology & medicine". In: *Mater. Today* 14.5 (May 2011), pp. 190–195. DOI: 10.1016/S1369-7021(11)70113-4.

[285] S. Xu et al. "Interactions between the Ig-Superfamily Proteins DIP-α and Dpr6/10 Regulate Assembly of Neural Circuits". In: *Neuron* 100.6 (2018), 1369–1384.e6. DOI: https://doi.org/10.1016/j.neuron.2018.11.001.

[286] F.M. Ranaivoson et al. "A Proteomic Screen of Neuronal Cell-Surface Molecules Reveals IgLONs as Structurally Conserved Interaction Modules at the Synapse". In: *Structure* 27.6 (June 2019), 893–906.e9. DOI: 10.1016/j.str.2019.03.004.

[287] M. Morey. "Dpr-DIP matching expression in Drosophila synaptic pairs". In: *Fly (Austin)* 11.1 (July 2016), pp. 19–26. DOI: 10.1080/19336934.2016.1214784.

[288] Y. Wang et al. "Systematic expression profiling of Dpr and DIP genes reveals cell surface codes in Drosophila larval motor and sensory neurons". In: *Development* 149.10 (May 2022), dev200355. DOI: 10.1242/dev.200355.

[289] R. Wang et al. "The PDBbind Database Methodologies and Updates". In: *J. Med. Chem.* 48.12 (June 2005), pp. 4111–4119. DOI: 10.1021/jm048957q.

Résumé

La capacité à déterminer rapidement et avec certitude si la liaison protéine-protéine est suffisamment forte pour avoir une fonction biologique est d'une importance primordiale dans le contexte de la découverte de médicaments in silico et des études fonctionnelles biologiques comme la synaptogénèse. Dans ma thèse, j'ai participé à l'établissement d'un protocole clair et complet utilisant la voie alchimique et géométrique pour les calculs standard de l'énergie libre de liaison absolue, qui reste le seul critère définitif pour estimer quantitativement l'affinité de liaison. J'ai étendu ce protocole aux complexes protéines-protéines dans le contexte de la syndémie COVID-19, contribuant à élucider la stratégie choisie par les variants préoccupants pour augmenter leur contagiosité tout en analysant la base moléculaire responsable de leurs affinités diverses pour ACE2. J'ai également renforcé les bases méthodologiques des calculs standard de l'énergie libre de liaison en démontrant la nécessité de restreintes pour obtenir des estimations précises en comparant la voie géométrique aux stratégies dépourvues de restreintes. De plus, j'ai participé à une exploration systématique méthodologique de la combinaison entre répartition de la masse de l'hydrogène et un pas de temps multiple pour le calcul des forces dans le but d'accélérer les calculs des énergies libres de liaison standard en utilisant l'approche de la voie géométrique. Cette étude a révélé la nécessité d'optimiser les paramètres impliqués dans le Lagrangien étendu de l'algorithme d'échantillonnage amélioré pour garantir la précision tout en améliorant la vitesse par un facteur de trois. Enfin, j'ai appliqué la voie géométrique à des complexes de synaptogenèse, avec la famille des Defective proboscis response (Dpr) et Dpr intercating proteins (DIP). J'ai pu retrouver des affinités correspondant aux données expérimentales avec la précision chimique. Cependant, ces calculs ont révélé la nécessité d'un réglage individuel et ne peuvent pas être réalisés dans l'ensemble de l'interactome Dpr-DIP en un temps raisonnable et sans intervention humaine. J'ai donc eu recours à des approches d'apprentissage automatiques, telles que l'analyse discriminante linéaire et la forêt aléatoire, en raison de leur capacité à traiter de grandes quantités de données et de leur interprétabilité. J'ai classifié avec précision les liaisons faibles et fortes dans cet ensemble de données, ce qui peut être utile dans l'analyse à haut débit. J'ai proposé des descripteurs d'entrée spécifiques, en exploitant les propriétés séquentielles, structurelles et physico-chimiques des protéines. Ces descripteurs d'entrée sont très généraux et pourraient être utilisés pour aborder des problèmes d'interaction entre protéines.

Abstract

The ability to rapidly ascertain with accuracy whether the binding of protein-protein is strong enough to have a biologically relevant function is of paramount importance in the context of in silico drug discovery and biological functional studies like those involve in synaptogenesis. In the thesis, I participated in establishing a clear and comprehensive protocol leveraging the alchemical and geometrical route for standard absolute binding free-energy calculations, which is the only definitive criterion to estimate the binding affinity quantitatively. I extended this protocol to protein-protein complexes in the context of the COVID-19 syndemic, helping to elucidate the strategy the variants of concerns had chosen to increase contagiosity while reasoning on the molecular basis responsible for their diverse affinities for ACE2. I also reinforce the methodological basis of the standard binding free-energy calculations by demonstrating the need for restraint to obtain accurate estimates in comparing restraint-based methods (the geometrical route) and unrestraint strategies. Furthermore, I participate in a methodological systematic exploration of the combination of hydrogen mass repartitioning and multiple time stepping to speed up

the calculations of standard binding free energies using the geometrical route approach, revealing the need for carefully tuned hyperparameters involved in the extended Lagrangian of the enhanced sampling algorithm to guarantee accuracy while improving speed by a factor of three. Lastly, I employed the geometrical route on complexes from synaptogenesis, the Defective proboscis extended response (Dpr), and the Dpr interacting proteins (DIP). I could recover experimental binding free-energy matching experimental data up to chemical accuracy. However, these calculations revealed a need for individual tunning and cannot be performed to the full extent of the Dpr-DIP interactome in a reasonable time. I resorted to machine learning approaches, Linear Discriminant Analysis, and Random Forest for their ability to handle large amounts of data and interpretability. I accurately classified weak and strong binders in this dataset, which can be helpful in high-throughput analysis. I proposed specific input features, leveraging sequence, structural, and physicochemical properties. These input features are very general and could be used to tackle protein-interacting problems.