Exploring chemo-mechanical transduction in the myosin molecular motor through computer simulations

Florian Blanc

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE 222

Équipe Ingénierie des fonctions moléculaires, Institut de Chimie, UMR 7177

THÈSE

présentée par :

Florian BLANC

soutenue le : **25 Septembre 2018**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Chimie/Chimie théorique et informatique

Exploration de la transduction chimio-mécanique chez le moteur moléculaire myosine par simulations numériques

THÈSE dirigée par :

Dr CECCHINI Marco	Maître de Conférence, université de Strasbourg
Dr HOUDUSSE Anne	Directrice de Recherche, institut Curie

RAPPORTEURS :

Dr HÉNIN Jérôme	Chargé de Recherche, institut de Biologie Physico-Chimique
Pr HUMMER Gerhard	Professor, Max Planck Institute of Biophysics

---

AUTRES MEMBRES DU JURY :

Dr PROST Jacques	Directeur de Recherche émérite, institut Curie

Exploring chemo-mechanical transduction in the myosin molecular motor through computer simulations

Abstract

Life relies on free energy conversions performed by molecular machines. Among them, the myosin molecular motor couples the hydrolysis of ATP to force production on actin through a swing of a « lever-arm ». Completing the cycle requires a regeneration step, the recovery stroke, in which the motor returns to its armed configuration and hydrolyzes ATP, which makes it crucial for chemo-mechanical transduction. In this thesis, we investigate the mechanism of the recovery stroke using molecular simulations. Capitalizing on a new crystal structure of myosin VI, we propose an original mechanism for the transition in which the re-priming of the lever arm is only loosely coupled to ATPase activation. Rather, our calculations suggest it is driven by thermal fluctuations in a ratchet-like manner, as opposed to previous models predicting strong coupling. Our results hint at how molecular motors may exploit spontaneous conformational fluctuations to produce work in an isothermal environment.

La vie repose sur des conversions d'énergie libre assurées par des machines moléculaires. Parmi elles, le moteur moléculaire myosine couple l'hydrolyse de l'ATP à la production de force sur l'actine par basculement d'un « bras de levier ». Compléter le cycle requiert une étape de régénération, ou *recovery stroke*, où le moteur retourne dans sa configuration armée et hydrolyse l'ATP, ce qui est crucial pour la transduction chimio-mécanique. Cette thèse étudie le mécanisme du *recovery stroke* par des simulations moléculaires. Partant d'une nouvelle structure cristallographique de la myosine VI, nous proposons un mécanisme original pour la transition dans lequel la remise en place du bras de levier n'est que faiblement couplée à l'activation de l'ATPase. En fait, nos calculs suggèrent qu'elle est déclenchée par les fluctuations thermiques de manière *ratchet-like*, et en contradiction avec des modèles précédents prédisant un couplage fort. Nos résultats suggèrent comment les moteurs moléculaires pourraient exploiter les fluctuations conformationnelles spontanées pour produire du travail dans un environnement isotherme.

Contents

Outline of the thesis

We outline the structure of the thesis and summarize the content of each chapter.

The *Résumé détaillé* (page 19) is a detailed summary of the entire thesis written in French. Afterwards, the content of the thesis is in English and organized as follows.

In Chapter 1, we give an informal overview of molecular machines, at the cross-road between statistical mechanics and nano-biophysics. We discuss Brownian motion, thermal agitation and the apparent paradox that molecular motors such as myosins can produce directed movement in an isothermal, highly fluctuating environment. Chapter 2 provides general information on the myosin superfamily, including its biological roles and the current consensus as to the motor mechanism. Chapter 3 briefly reviews Molecular Dynamics (MD) simulations, introducing the potential energy function and important notions of statistical mechanics. Then, Chapter 4 dives deeper into the formalism and methodology of free energy calculations and reviews popular methods, several of which are used throughout this thesis. Chapters 1 to 4 represent a general introduction and need not be read in details to understand the thesis.

Chapters 5 to 12, by constrast, present original results, some of which are yet unpublished. Chapter 5 comes back to myosin and begins by a review of the existing literature on the recovery stroke; in particular, a critical discussion of previously published models of the recovery stroke is initiated. Then, at the end of the chapter, the new Pre-Transition State (PTS) crystal structure is introduced, and compared to the end-points of the recovery stroke. The PTS hypothesis, *i.e.* the putative relevance of the PTS structure as an intermediate along the recovery stroke, is proposed.

Chapter 6 presents the characterization of the dynamical behaviour of the motor domain's conformational states, including PTS, by unbiased MD simulations. Notably, in this chapter are presented several geometrical observables designed to describe the recovery stroke. Chapter 7 reports on the use of Accelerated MD techniques to explore the conformational space of the recovery stroke. In chapter 8, free energy calculations with the ABF strategy are used to probe the energetics of ATPase activation as a function of the conformational state of the motor domain; in support of the PTS hypothesis, it is argued that the free energy cost for early ATPase activation makes such a scenario unrealistic. Chapters 9 and 10 report on the use of biased simulations (Steered Molecular Dynamics) and various free energy calculation protocols to investigate the mechanisms of the PR $\rightarrow$ PTS and PTS $\rightarrow$ PPS transitions. Combining the results presented in Chapters 5 to 10, we arrive at a novel, so-called ratchet-like model for the recovery stroke of myosin VI, summarized at the beginning of Chapter 11. Then, in this chapter, we critically discuss the emerging scenario in light of available experimental results on the recovery stroke. We show that our model is not contradicted by existing mutational data, but that it is also the case for the competing model by Fischer and co-workers (Stefan Fischer, Windshügel, et al. 2005). Settling the debate would instead require a detailed energetic comparison of the two models; we delineate a strategy to perform this analysis using the string method in collective variables, and report on its first results. Finally, chapter 12 reports on results unrelated to the recovery stroke, obtained respectively on the tail domain of *Dictyostelium discoideum* Myosin VII and the lever-arm domain of Myosin X. This chapter illustrates how simple simulation techniques (conventional MD) can complement various experimental approaches for the study of biomolecules.

Appendix A outlines some complementary theoretical notions which were deemed superfluous for the main text, but were included for the sake of completeness. Appendix B reports on the self-contained computational study of a prototypical, artificial molecular machine, by the means of free energy calculations and path-optimization methods. It illustrates how the approaches used for the investigation of myosin can be applied, beyond biophysics, to molecular machines in general - and, in turn, how simple molecular machines may be used as model systems. Finally, Appendix C is the manuscript and supplementary information of (Blanc et al. 2018), the main publication reporting on the results presented in this thesis.

List of publications

An intermediate along the recovery stroke of Myosin VI revealed by X-ray crystallography and Molecular Dynamics
Florian Blanc, Tatiana Isabet, Hannah Benisty, H. Lee Sweeney, Marco Cecchini, Anne Houdusse
PNAS, 2018

Myosin MyTH4-FERM structures highlight important principles of convergent evolution
Vicente José Planelles-Herrero, **Florian Blanc**, Serena Sirigu, Helena Sirkia, Jeffrey Clause, Yannick Sourigues, Daniel O. Johnsrud, Béatrice Amigues, Marco Cecchini, Susan P. Gilbert, Anne Houdusse, and Margaret A. Titus
PNAS, 2016

The myosin X motor is optimized for movement on actin bundles
Virginie Ropars, Zhaohui Yang, Tatiana Isabet, **Florian Blanc**, Kaifeng Zhou, Tianming Lin, Xiaoyan Liu, Pascale Hissier, Frédéric Samazan, Béatrice Amigues, Eric D. Yang, Hyokeun Park, Olena Pylypenko, Marco Cecchini, Charles Sindelar, H. Lee Sweeney and Anne Houdusse
Nature Communications, 2016

List of communications

A novel intermediate along the recovery stroke of Myosin VI revealed by X-ray crystallography and Molecular Dynamics
Florian Blanc, Tatiana Isabet, Marco Cecchini, Anne Houdusse
Congrès du Groupe de Graphisme et Modélisation Moléculaires (GGMM) 2017, Reims, France (Poster presentation)

A novel intermediate along the recovery stroke of Myosin VI revealed by X-ray crystallography and Molecular Dynamics
Florian Blanc, Tatiana Isabet, Marco Cecchini, Anne Houdusse
European Molecular Biology Organization (EMBO) Meeting, 2016, Mannheim, Germany (Poster presentation)

Remerciements

J'ouvre ces remerciements en exprimant ma plus sincère gratitude aux membres de mon jury de thèse, à savoir le Professeur Jacques Prost, examinateur, le Docteur Jérôme Hénin, rapporteur, et le Professeur Gerhard Hummer, rapporteur et président. Ce fut pour moi un honneur de voir des scientifiques de ce calibre évaluer mon travail avec tant de soin (et ce, malgré mon retard...). Je remercie aussi chaleureusement Fabienne Penner et Muriel Muzet, grâce auxquelles soutenance et pot de thèse ont pu prendre place dans les magnifiques salles de l'Institut de Sciences et d'Ingénierie Supramoléculaires.

Cette thèse, je crois, est l'aboutissement d'une trajectoire entamée il y a très longtemps - aussi longtemps que je me souvienne, en fait. Cela veut dire, bien sûr, qu'au fil des années j'ai rencontré de nombreux enseignants et mentors qui ont nourri ma vocation pour la recherche. Qu'il me soit ici permis de les remercier. Tous et toutes m'ont apporté, mais il en est certains qui ont eu une importance particulière.

Je pense à mon institutrice du cours préparatoire, Micheline (impossible de me rappeler comment orthographier son nom de famille...), qui m'offrit des clichés de la surface lunaire pris par son mari astronome devant lesquelles je passai de longues heures à rêver; à Messieurs Bouvier, Bain, respectivement professeurs d'histoire-géographie et de latin au collège puis au lycée, et Madame Bugnet, également professeure de latin, tous trois des modèles de cette belle érudition rencontrée, hélas, bien plus souvent chez les littéraires; à mes professeurs de sciences au lycée, tout particulièrement Madame Bruni, professeure de SVT (qui fut la première à me conseiller d'envisager l'ENS), Monsieur Coz, professeur de physique, et Monsieur Vivès, professeur de mathématiques, grâce auxquels j'ai commencé à entrevoir que la science est bien plus qu'une collection de faits.

Ensuite, les classes préparatoires, Lycée Thiers à Marseille, ce rythme infernal et cette grisante impression de fuite en avant alors que ce que je sais encore aujourd'hui être le socle ferme de mes connaissances scientifiques est bâti, parfois dans la douleur, par des enseignants aussi compétents qu'exigeants. Je demeure ainsi profondément reconnaissant envers Madame Le Bars, Monsieur Chauvet, Madame Grossetête, professeurs de SVT; Madame Pagano, professeure de mathématiques en spé; et tout particulièrement Messieurs Rédoglia et Jaubert (professeurs de physique-chimie en sup et spé, respectivement), qui poussèrent l'exigence encore plus loin lors de la préparation aux Olympiades de chimie 2009 pour l'un, et de la préparation spécifique aux ENS pour l'autre. Enfin, j'adresse mes plus sincères remerciements à Monsieur Reissman, professeur de mathématiques en sup. En plus du pédagogue génial mariant rigueur et humour, je me souviendrai toujours de cette après-midi de janvier 2009 où, la neige (!) ayant paralysé le tram marseillais, nous fûmes contraints de marcher du centre-ville jusqu'aux Caillols. Jamais, je crois, n'ai-je appris autant sur autant de sujets différents en l'espace d'une seule conversation que lors de celle que nous partageâmes ce jour-là. Un de ces évènements improbables qui marquent une vie.

Après la prépa, l'École Normale Supérieure, rue d'Ulm, Paris, objectif tant rêvé et finalement atteint. On sort de prépa la tête très pleine, de connaissances et de certitudes. L'ENS se charge d'élargir les premières et d'ébranler les secondes. Aux cours magistraux denses et structurés succèdent les séminaires de recherche, et pour la première fois l'accent est mis sur les problèmes non résolus et les questions ouvertes. Les conditions d'étude toutes particulières, la proximité des équipes de recherche, la confiance accordée aux élèves dans la construction de leur programme d'études, tout cela contribue à nous préparer au métier de chercheur, et par bien des aspects, cette thèse, c'est rue d'Ulm qu'elle a commencé. Que soient donc remerciés celles et ceux qui font de l'École cet endroit incroyablement éclairé: ses enseignants, ses chercheurs, son personnel, et ses étudiants. Plus prosaïquement, c'est également l'ENS qui a permis le financement de la majorité de mes études, dont la thèse, grâce au salaire de fonctionnaire-stagiaire afférent au statut de normalien, puis à une allocation doctorale spécifique. Cet argent, c'est celui des contribuables français, et je sais la dette que leur dois. Puisqu'il est question de financement, la Fondation pour la Recherche Médicale a également beaucoup contribué à ce travail *via* une dotation sur projet, qui m'a notamment permis d'effectuer une quatrième année de thèse durant laquelle tout a finalement abouti. Mes sincères remerciements à la FRM et ses donateurs.

Les cours, les séminaires, les articles, tout cela est important, mais *in fine* le travail de recherche s'apprend au laboratoire. En tant que stagiaire, j'ai eu la chance d'évoluer sous la direction de scientifiques expérimentés et compétents, avec lesquels j'ai beaucoup appris sur la pratique de la science. Tout particulièrement, je remercie chaleureusement Catherine Etchebest, qui m'a doté de bases solides tant dans la théorie que la pratique de la dynamique moléculaire, pendant mon stage de M1; ainsi que Nohad Gresh, qui fit de même pour la chimie quantique et les arcanes des champs de force polarisables pendant mon stage de M2. Mais, même si j'ai travaillé avec eux sur des projets passionnants en compagnie de personnes qui l'étaient tout autant, demeurait toutefois un sujet bien particulier vers lequel je finissais toujours par retourner. Un sujet qui semblait exister, comme par magie, à l'intersection de tous mes champs d'intérêt scientifiques, ou presque. Un sujet vaste, ancien, intimidant, sur lequel beaucoup a déjà été dit, écrit, débattu. Ce sujet, ce problème, c'est celui des principes de fonctionnement des moteurs moléculaires, au premier rang desquels la myosine et sa majestueuse complexité.

Ma première rencontre significative avec les moteurs moléculaires eut lieu pendant une belle journée de printemps, en 2010, à Marseille, pendant l'épreuve écrite de physique du concours d'entrée à l'ENS. Un bon tiers du sujet portait sur un modèle mathématique très simple, et élégant, du déplacement de la kinésine sur un microtubule, au moyen de marches aléatoires. Quelques heures plus tard, je quittai la salle en ayant beaucoup appris (ce qui est assez paradoxal!) et avec deux nouvelles convictions personnelles. La première, que l'approche consistant à dériver des propriétés macroscopiques à partir de considérations microscopiques, cette fameuse physique statistique, à laquelle je ne m'étais pas véritablement frotté jusqu'alors, était incroyablement séduisante. Je voulais continuer à explorer ce domaine. La seconde, que les moteurs moléculaires étaient des objets si fascinants que ni la biologie ni la physique, prises séparément, ne pouvaient suffire à en élucider les principes. Mon choix d'étudier la biologie avait été motivé tant par ma passion pour cette discipline que par le refus de cesser d'étudier les autres domaines des sciences: inconcevable d'arrêter l'étude de la chimie, de la physique ou des mathématiques sous prétexte que je voulais être biologiste. Il fallait simplement trouver un sujet qui permette de concilier tous ces aspects. De tels sujets ne manquent pas, mais je l'ignorais largement à l'époque - et avec cette histoire de moteurs moléculaires, on était en plein dedans.

Un peu plus d'un an plus tard, à la recherche d'un stage d'été, je tombai sur une offre intéressante: une étude cristallographique du domaine moteur de la myosine X. Tiens tiens. De la biologie structurale, qui m'attirait depuis le lycée et la première fois que j'avais contemplé les fascinantes intrications d'une structure 3D de protéine; et qui plus est, sur un de ces fameux moteurs moléculaires... Et c'est ainsi que je poussai la porte du laboratoire d'Anne Houdusse à l'Institut Curie.

J'ai une immense dette envers Anne.

C'est dans le groupe d'Anne que j'ai mené mon premier véritable travail de recherche. Que j'ai connu pour la première fois l'excitation des résultats positifs. Comme encadrante de stage d'abord, co-directrice de thèse ensuite, je me sais privilégié d'avoir pu travailler sous la férule d'une chercheuse de classe mondiale. C'est Anne qui m'a appris presque tout ce que je sais sur les myosines et qui a achevé de me convaincre que le sujet était passionnant. Elle qui m'a montré la quantité d'informations que l'on peut obtenir en analysant une structure de protéine, lorsque l'on sait où chercher. Elle qui m'a fait prendre conscience des nombreuses zones d'ombre qui demeuraient autour d'un sujet pourtant en apparence aussi rebattu que celui des myosines. C'est Anne qui m'a mis le pied à l'étrier, et a ensuite continué de me guider par ses conseils, généreux, ses critiques, toujours fondées, et ses idées, nombreuses. Merci.

C'est également Anne qui, voyant que je délaissais la cristallographie pour la modélisation, me mit en contact avec Marco Cecchini à l'ISIS, Strasbourg, dans le groupe duquel je passai d'abord un stage d'été en 2013, et qui fut par la suite mon directeur de thèse. Un groupe appliquant modélisation et mécanique statistique à un large éventail de sujets alliant physico-chimie et biologie, qui plus est avec un axe sur la myosine: vous imaginez mon enthousiasme ! C'est avec Marco que j'ai réellement appris à transformer des idées vagues en arguments structurés et étayés par des données. Je lui sais gré de n'avoir jamais transigé avec son exigence de rigueur et d'ambition. Cela fut parfois frustrant ("I'm not convinced" - Marco Cecchini). Je me rends surtout compte aujourd'hui à quel point ce fut crucial pour espérer produire de la science de qualité. Sous la direction de Marco, je me suis emparé de mon sujet de thèse, et c'est parce qu'il aura su quand me guider par ses conseils et ses idées, et quand me laisser libre de chercher, d'expérimenter. Avec le recul, je m'aperçois qu'il m'a donné juste assez de structure pour me permettre de m'épanouir comme chercheur, sans m'étouffer, tout en me mettant au défi de toujours aller un peu plus loin. Je lui en suis à jamais reconnaissant.

Un dernier point, d'importance. Tant avec Anne qu'avec Marco, j'ai toujours eu le sentiment que mes idées et propositions, même les plus naïves, étaient écoutées comme l'auraient été celles d'un pair. Je crois que c'est chose rare, et je voulais donc le souligner et vous exprimer ma gratitude. Pour cela et le reste, je n'aurais pu souhaiter meilleurs encadrants, et je me souviendrai de ma thèse avec vous comme d'une période d'épanouissement intellectuel complet. Merci, merci, merci.

Les collègues, maintenant. Je remercie en premier lieu les membres de l'équipe Motilité Structurale, tout spécialement Tatiana Isabet et Virginie Ropars pour leur encadrement et leur aide, très appréciés, et Vicente Planelles-Herrero et Julien Robert-Paganin pour les échanges scientifiques que nous avons eus. Ensuite, mes remerciements aux membres, présents et passés, de l'équipe Ingénierie des Fonctions Moléculaires. L'ancienne génération d'abord: Nicolas Calimet, Jérémy Esque et Nicolas Merstorf; merci pour l'ambiance, les moments de rigolade, et le soutien.

Viennent les compagnons d'armes: Simone Conti, Nicolas Martin, Joel Montalvo-Acosta, et Adrien Cerdan, qui avons commencé nos thèses environ au même moment (à plus/moins un an près). En-

semble, nous avons appris, avancé, reculé, partagé l'excitation de la recherche et la frustration qui en est indissociable. Messieurs, si j'ai malheureusement échoué à vous convaincre que la myosine est le seul système digne d'intérêt, les autres n'étant que des *toy-systems*, je resterai immensément fier d'avoir finalement mérité ce doctorat en votre compagnie. De Simone, je retiendrai le souci de rigueur et les impressionnantes qualités d'analyse; de Joel, la créativité bouillonnante, et la bonne humeur permanente; de Nicolas, le charisme, le talent pour l'organisation, et le professionnalisme (ou, plus clairement, comment Nico était de loin le seul adulte responsable du groupe); et d'Adrien, l'inspiration et la persévérance (il en aura fallu pour stabiliser cette fameuse structure de l'état actif, et ce n'était que le début). Sachez que j'ai cherché à m'inspirer de toutes vos qualités dans mon propre travail de chercheur, et que cette thèse vous doit beaucoup. Sachez aussi que je suis heureux et fier de pouvoir me dire votre ami.

Merci beaucoup à Diego Gomes, dont les amples compétences techniques m'ont plus d'une fois été d'un grand secours, comme au reste de l'équipe. Les stagiaires: Cédric Bouysset, Paulina Paçak, Rifat Gimatev, et Donatienne de Francquen. Merci pour votre enthousiasme au labo et en dehors ! Un merci spécial à Dona pour son aide lors de l'impression finale du manuscrit... Enfin, la nouvelle génération, celles et ceux qui ont rejoint le groupe juste avant ou juste après la soutenance de ma thèse. Gilberto Pereira, Marion Sisquellas, et plus récemment Alison Popp et Katia Galenti. Savoir que c'est vous qui représentez la relève tempère la peine que j'ai à finalement quitter l'équipe.

Après notre départ de l'ISIS, nous avons rejoint l'Institut de Chimie, et je remercie ses membres pour l'accueil chaleureux qu'ils nous ont fait. Tout particulièrement Jean Weiss, Jennifer Wytko, Romain Rupert, Mary-Ambre Carvalho, Valérie Heitz et Henri-Pierre Jacquot de Rouville. Merci pour la bonne humeur, les mots-fléchés, et les conversations scientifiques ou pas.

Au tour de la famille, à présent.

Ma belle-famille d'abord, au sens large, pour leur soutien constant qui prit une forme très concrète lorsque la rédaction du manuscrit entra en collision avec les vacances d'été. Gisèle, Philippe et Loïc; Christian et Carole; ma profonde gratitude, doublée de mes excuses, pour avoir toléré que je passe mes journées à écrire quand le temps en votre compagnie est déjà si rare. Je crois au moins pouvoir dire que ça a valu le coup. Alice et Gilbert, merci pour vos mots d'encouragement; René et Gabrielle, et Christian et Carole, encore, merci d'avoir poussé le soutien jusqu'à venir assister à ma soutenance (et aider à son bon déroulement!). J'en suis honoré.

À mes deux grand-mères, Jeanine et Lucette, merci pour votre gentillesse et votre amour permanents. Je n'ai pas été le plus assidu pour donner fréquemment des nouvelles, mais je sais que vous suiviez l'affaire de près, et j'en suis très heureux. Papy, les dernières années ont été dures, et tu es finalement parti, la veille de la soutenance de cette thèse. Mais je me souviens. Et j'ignore si, comme tu me l'avais dit un jour, je suis la fierté de la famille; mais je sais que jusqu'à la fin tu étais fier de moi. Merci.

À mon frère, Mathieu, je dis un merci particulier. On est pas trop du genre démonstratif, tous les deux. Mat, on a pris des chemins différents, mais sache que je suis admiratif et fier de tes réussites; et je crois -je sais- que la réciproque est vraie. Et autre chose. Partir à l'autre bout de la France pour faire une thèse, c'est bien beau. Dans notre cas, cela a impliqué que c'est toi qui es resté et qui as du gérer, notamment quand la santé de Papy s'est dégradée. Je ne l'oublierai pas. Cela vaut aussi pour ma belle-soeur préférée, Faustine.

C'est certainement à mes parents, Christian et Patricia, que je dois le plus. Eux qui ont d'abord su comment répondre aux questions insistantes d'un enfant curieux (parfois trop, ils sauront de quoi je parle), puis lui donné les outils pour aller chercher les réponses tout seul. Mes parents, qui bien que n'étant pas scientifiques eux-mêmes, n'ont jamais parus dérangés par l'idée que je veuille consacrer ma vie à la recherche - malgré quelques tentatives pour la forme (non, Papa, ouvrir un camion de pizzas n'est toujours pas dans mes projets). Qui, au contraire, m'ont toujours épaulé, soutenu (notamment financièrement et logistiquement, ce qui fut plusieurs fois crucial, jusqu'à la toute fin), et plus que tout, m'ont fait confiance. C'est cette confiance toute simple qui a été mon moteur, et je vous en remercie, du fond du cœur. Vingt-huit ans après je vous dis ceci: vous n'auriez pas pu mieux faire.

Il était bien sûr impossible de ne pas réserver les dernières lignes de ces remerciements à celle qui est ma compagne, ma partenaire et l'amour de ma vie. Odyssée, tu as complété ton propre doctorat deux ans avant que je ne fasse de même, ce qui m'a permis de voir de première main ce qu'est une thèse brillante, et brillamment soutenue. Sache que la scientifique et l'enseignante que tu es sont autant d'inspirations pour moi, et que je suis coupable de ne pas te l'avoir assez dit. Pendant toutes ces années, nous avons changé ensemble, mûri ensemble, mais tu as gardé cette joie de vivre authentique et simple dont je me moque gentiment et qui, à la vérité, illumine ma vie. Tu es ma constante, mon "si et seulement si", dans les moments de réussite et les moments de doute. Je sais que cette thèse n'aurait pas abouti sans ta patience, ta force et ton amour. Ou, pour reprendre sans scrupules tes propres mots, *tout aurait été beaucoup moins bien sans toi*.

Florian Edmond Charles Blanc, Strasbourg, le 29 avril 2019

Résumé détaillé

Introduction

Machines moléculaires et fluctuations

Le fonctionnement cellulaire met en jeu une myriade de macromolécules biologiques, en particulier des protéines, qui assurent les diverses tâches nécessaires à la vie au niveau moléculaire: catalyse enzymatique, signalisation, transport transmembranaire... Les progrès de la biologie structurale ont révélé les architectures remarquablement complexes que peuvent prendre les protéines, et comment la forme tridimensionnelle d'une protéine est essentielle pour sa capacité à remplir sa fonction. Dans de nombreux cas, cette fonction met aussi en jeu des changements conformationnels de grande amplitude et divers partenaires moléculaires, à tel point que l'on parle de *machine moléculaire*. S'il est tentant de faire l'analogie entre ces machines moléculaires et les machines de fabrication humaine, il existe une différence fondamentale entre les deux, qui tient essentiellement à la disparité d'échelles. À l'échelle macroscopique, familière, la gravité et l'inertie dominent et les objets ont un comportement déterministe. En revanche, à l'échelle nanoscopique (1×10^{-9} m), les fluctuations thermiques, stochastiques, dominent: les machines moléculaires opèrent efficacement en dépit des collisions incessantes avec les autres molécules environnantes, dans des conditions qui ont été comparées à une *tempête moléculaire* (Hoffmann 2012). Une idée générale, maintenant bien admise, est que les machines biomoléculaires fonctionnent en fait en exploitant ces fluctuations - ce qui n'est finalement guère surprenant étant donné que les machines moléculaires ont évolué dans ces conditions. Ainsi, il semble raisonnable de s'attendre à ce que, contrairement aux machines macroscopiques, les machines moléculaires présentent une certaine "mollesse" (*softness*), c'est-à-dire que le couplage entre leurs différents domaines structuraux doit être faible pour que les changements de conformations à l'origine de la fonction puissent être déclenchés par les fluctuations thermiques.

Cependant, il reste à fournir la description détaillée de la manière dont ces principes sont effectivement implémentés dans une machine biomoléculaire donnée. Outre son intérêt fondamental, une telle description ouvrirait la voie à la conception rationnelle de machines moléculaires synthétiques fonctionnant selon les mêmes principes, ce qui est un enjeu crucial en nanotechnologies. Dans cette thèse, nous présentons des résultats de simulations moléculaires qui contribuent à expliciter le mécanisme par lequel la myosine, une importante machine moléculaire biologique, exploite les fluctuations thermiques pour assurer sa fonction de production de force et de mouvement directionnel sur le cytosquelette d'actine.

Les myosines

Les myosines sont une superfamille de protéines motrices, associées au cytosquelette d'actine, et sont présentes dans toutes les cellules eucaryotes. Les myosines jouent un rôle fondamental dans plusieurs fonctions biologiques critiques telles que la motilité cellulaire, le trafic intracellulaire, l'endocytose, et sont à la base de la contraction musculaire. En contexte pathologique, les myosines sont causalement

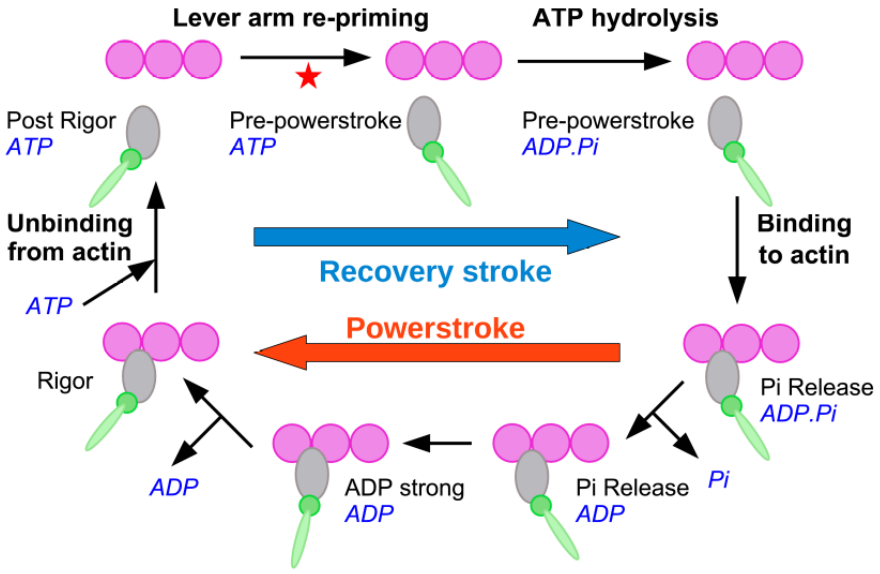


Figure 0.1.: Cycle moteur de la myosine, adapté de (Blanc et al. 2018). L'étoile rouge représente la position (hypothétique) de l'état PTS discuté dans cette thèse.

associées à des maladies graves comme la cardiomyopathie hypertrophique, la défaillance cardiaque, plusieurs formes de surdité héréditaire, et certains cancers (notamment à cause de leur effet positif sur la motilité cellulaire, à l'origine de métastases).

Les myosines sont qualifiées de *moteurs moléculaires*, parce qu'elles convertissent l'énergie libre chimique fournie par l'hydrolyse de l'ATP en travail mécanique et déplacement directionnel le long des filaments d'actine. Malgré des différences entre isoformes liées aux particularités des rôles biologiques spécifiques qu'elles peuvent remplir, le mécanisme de production de force est remarquablement conservé. Le cycle moteur, ou cycle actomyosine, met en jeu des changements conformationnels de grande ampleur dans le domaine moteur de la myosine (Figure 0.1). Pendant la phase de production de force, la liaison du domaine moteur à l'actine est couplée avec la libération des produits d'hydrolyse (Phosphate inorganique, puis ADP) et le basculement vers l'avant du "bras de levier", un domaine structural de forme allongée situé à la suite du domaine moteur. Cette étape du cycle est appelée *powerstroke*, habituellement traduit par "coup de force". Pour que le moteur puisse fonctionner de manière cyclique, une étape de remise en place du bras de levier doit avoir lieu; de plus, cette étape doit prendre place lorsque la myosine n'est pas liée à l'actine, sans quoi le mouvement produit pendant le *powerstroke* sera annulé. Cette étape de régénération est qualifiée de *recovery stroke* et constitue le sujet d'étude principal de cette thèse[1].

Le recovery stroke: anciens modèles et nouvelle structure

Pendant le *recovery stroke*, la remise en place du bras de levier est couplée à l'activation de l'activité ATPase du domaine moteur, et donc à l'hydrolyse de l'ATP (Figure 0.2). Ainsi, bien que cette étape

1. À la différence du *powerstroke* traduit par "coup de force", il n'existe pas de traduction généralement admise pour le *recovery stroke*. "Transition de remise en place" ou "ré-armement" pourraient par exemple être proposés, mais il nous a semblé qu'une traduction systématique se ferait au détriment de la fluidité de lecture. Par conséquent, nous conservons le terme anglophone *recovery stroke* dans la partie francophone de cette thèse.

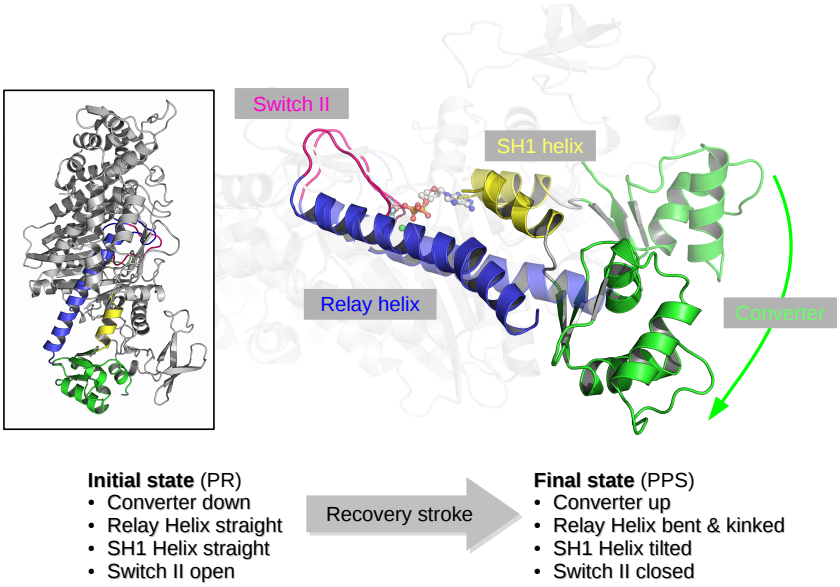


Figure 0.2.: Résumé des éléments structuraux et de leurs réarrangements pendant le *recovery stroke*.

ne corresponde pas à celle de production de force, elle est cruciale pour comprendre comment la myosine peut agir comme un convertisseur d'énergie chimique en énergie mécanique - un processus qualifié de *transduction chimio-mécanique*. Si les bases thermodynamiques de la transduction chimio-mécanique sont relativement bien comprises (Hill 2005), c'est nettement moins le cas en ce qui concerne la manière dont des machines biomoléculaires comme la myosine peuvent réaliser, dans les faits, la conversion énergétique. Une description détaillée des transitions fonctionnelles, associant mécanisme décrit à résolution atomique et profil énergétique, est nécessaire pour cela. Dans ce contexte, élucider le mécanisme du *recovery stroke* de la myosine représenterait un progrès significatif dans la compréhension des principes de fonctionnement des moteurs moléculaires, notamment parce que les principes mis au jour pourraient être réinvestis pour la conception rationnelle de moteurs moléculaires synthétiques.

Depuis le début des années 1990, des études structurales (cristallographie aux rayons X et cryo-microscopie électronique) ont révélé les diverses conformations adoptées par la myosine au cours de son cycle, mettant notamment en lumière le fait que leurs caractéristiques générales sont essentiellement invariantes d'une isoforme à l'autre (Sweeney and Houdusse 2010b). En particulier, l'état Post Rigor State (PR) et l'état Pre-Powerstroke State (PPS), respectivement identifiés comme l'état initial et l'état final du *recovery stroke*, ont été résolus pour plusieurs myosines. Dès les années 2000, plusieurs groupes de recherche ont utilisé ces structures comme points de départ d'études numériques visant à modéliser le mécanisme du *recovery stroke* (principalement, pour la myosine II de *Dictyostelium discoideum*). Ces modèles, bien qu'obtenus à partir de méthodes variées et différant dans les détails de leurs conclusions, ont presque tous en commun de proposer que l'activation de l'activité ATPase (*via* la fermeture d'une boucle appelée *switch II* sur l'ATP dans le site actif) représente l'événement initiateur du *recovery stroke*. Ainsi, dans le modèle de Stefan Fischer, Windshügel, et al. (2005) (le plus cité), la fermeture du *switch II* déclenche une séquence de changements conformationnels locaux qui aboutissent, de proche en proche, à la rotation du bras de levier. Une caractéristique importante de

ce modèle est qu'il propose un mécanisme *fortement couplé* dans lequel les réarrangements individuels des différents sous-domaines sont étroitement coordonnées, laissant peu de place aux événements stochastiques.

À l'inverse, une nouvelle structure cristallographique de la myosine VI a récemment été résolue dans le groupe d'Anne Houdusse (Institut Curie, Paris). Cette structure, baptisée *Pre-Transition state* ou PTS, présente des caractéristiques frappantes qui suggèrent qu'elle est représentative d'un intermédiaire structural du *recovery stroke*, jusque-là inconnu; de plus, ces caractéristiques sont en contradiction avec les modèles précédemment publiés de la transition. En effet, dans la structure PTS, le *switch II* est ouvert, mais le bras de levier est presque complètement ré-armé: cela suggère d'une part l'absence d'un couplage fort entre ces deux réarrangements, et d'autre part que c'est le mouvement du bras de levier, plutôt que la fermeture du site actif, qui est l'événement initiateur du *recovery stroke*. Nous sommes ainsi amenés à formuler "l'hypothèse PTS" (*PTS hypothesis*), selon laquelle la structure PTS représente bel et bien un intermédiaire du *recovery stroke*. Dans cette thèse, les méthodes de la biophysique numérique (simulations de dynamique moléculaire et calculs d'énergie libre) sont utilisées pour explorer les implications de l'hypothèse PTS vis-à-vis du mécanisme de la transduction chimio-mécanique; *in fine*, nous proposons une stratégie, reposant sur des méthodes récentes de calcul de chemin optimal, pour tester directement l'hypothèse.

Résultats et discussion

Dynamique moléculaire et observables pour le recovery stroke

La dynamique moléculaire (MD) est une technique de simulation numérique qui consiste à intégrer les équations du mouvement (classiques) pour un système moléculaire décrit par un potentiel approximatif, ou "champ de forces", également classique. En général, les équations du mouvement sont modifiées pour s'assurer que les trajectoires observées sont représentatives de la dynamique qu'aurait le système à température constante. Ainsi, la dynamique moléculaire permet d'obtenir une image des fluctuations conformationnelles des protéines sous l'effet de l'agitation thermique. Lorsqu'elle est appliquée à des systèmes de grande taille (à l'échelle moléculaire), comme une protéine de plusieurs centaines de résidus entourée de molécules d'eau explicites, la dynamique moléculaire est une technique coûteuse en calculs qui requiert l'usage prolongé de supercalculateurs. C'est notamment le cas pour la présente étude de la myosine.

Une trajectoire de dynamique moléculaire correspond à un grand nombre (typiquement plusieurs milliers) de configurations atomiques du système étudié. Contrairement à la biologie structurale expérimentale, il n'est pas possible de mener une étude détaillée de chaque configuration; à la place, il est d'usage de choisir un petit nombre d'*observables*, ou *variables collectives*, qui décrivent les réarrangements pertinents pour le système en question. La projection de la trajectoire sur ces observables offre une vue résumée du comportement dynamique du système, et permet de caractériser les différents états conformationnels en fonction des valeurs typiques prises par les observables. Pour le *recovery stroke* de la myosine, nous avons introduit une série d'observables visant à capturer les réarrangements des sous-domaines individuels durant la transition. Les observables les plus importantes sont:

1. Les trois coordonnées cartésiennes X', Y', Z' du barycentre du sous-domaine *convertisseur* par rapport aux axes principaux du domaine moteur de la myosine. Le convertisseur représente un

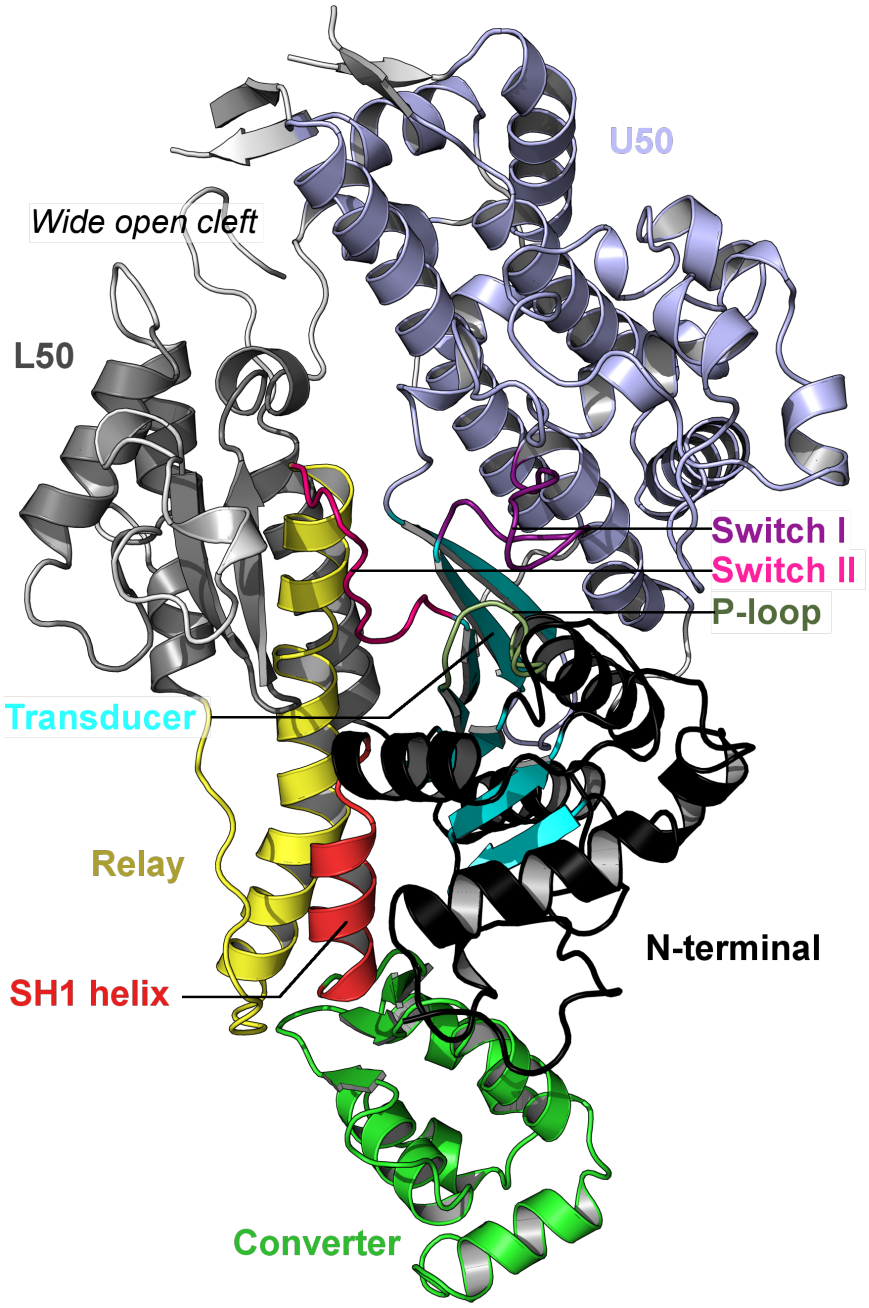


Figure 0.3.: Structure cristallographique PTS de la myosine VI, adaptée de (Blanc et al. 2018).

connecteur entre le domaine moteur et le bras de levier; en fait, le basculement du bras de levier est directement dirigé par une rotation de grande amplitude du convertisseur. Caractériser la position relative du convertisseur par rapport au domaine moteur est donc crucial pour décrire le *recovery stroke*, et c'est la fonction de ces trois observables.

2. Deux angles θ_{RH} et θ_{SH1} introduits respectivement pour rendre compte du développement d'une courbure (*bending*) et d'un coude (*kinking*) dans l'hélice Relais, et du basculement (*tilting*) de l'hélice SH1. Les hélices Relais et SH1 représentent deux connecteurs flexibles entre le domaine moteur et le convertisseur (et donc, le bras de levier). Les réarrangements de ces hélices, décrits par les angles susmentionnés, sont couplés avec la rotation du convertisseur pendant le *recovery stroke*.

3. Le $\Delta RMSD_{kink}$ de la région de l'hélice Relais qui développe un coude (*kink*). La formation du coude passe par une ré-organisation des liaisons hydrogènes intra-hélicales au niveau du squelette peptidique des résidus 485 à 493. La différence de RMSD, pour une configuration donnée, par rapport à l'état PTS (coude présent) et l'état PR (coude absent) permet de caractériser la conformation locale de l'hélice. Pour une hélice non coudée, similaire à la structure PR, cette observable vaut typiquement 1.4 Å; par symétrie, une hélice complètement coudée correspond à des valeurs de l'ordre de -1.4 Å.

4. Deux distances sont introduites pour rendre compte de l'état du site actif, en particulier vis-à-vis de la position de la boucle *switch II* cruciale pour l'activation de l'activité ATPase. La première distance, notée d_1, est définie comme la distance entre les atomes R205CZ et E461CD; elle permet de décrire la formation (ou non) du "pont salin critique" (*critical salt-bridge*) entre les résidus R205 et E461. La seconde distance, notée d_γ, est définie comme la distance entre les atomes G459N et ATP:O1G. Cette observable décrit la formation d'une liaison hydrogène entre la boucle *switch II* (*via* l'atome d'azote du squelette peptidique de G459) et le groupement phosphate γ de l'ATP. Ces deux interactions, pont-salin critique et liaison hydrogène switch II-ATP, sont requises pour la fermeture du *switch II* et la catalyse de l'hydrolyse de l'ATP. On supposera donc que la myosine peut être considérée catalytiquement active lorsque ces deux interactions sont formées.

Caractérisation de la dynamique du domaine moteur de la myosine VI par dynamique moléculaire

À l'aide de simulations de dynamique moléculaire (MD) non-biasées sur des échelles de temps de quelques centaines de nano-secondes, nous avons étudié la stabilité et les caractéristiques dynamiques de la structure PTS; de plus, nous avons pu les comparer avec les états extrémaux du *recovery stroke*, à savoir les structures PR et PPS.

Dans l'état PTS, les simulations révèlent que le sous-domaine convertisseur occupe une position clairement intermédiaire entre les états PR et PPS, ce qui est en accord avec l'hypothèse PTS (Figure 0.4). De plus, on observe que le convertisseur présente des fluctuations de position bien plus importantes dans l'état PTS que dans les états PR et PPS, parce qu'il est essentiellement découplé du domaine moteur (Blanc et al. 2018). Notamment, une transition du convertisseur vers une position très proche de celle qu'il occupe en PPS, a été observée de manière réversible à l'échelle de temps 300 ns. Cette observation renforce l'hypothèse que le PTS est bien un intermédiaire entre PR et PPS, puisqu'une transition vers PPS, partielle mais spontanée, est capturée. De plus, elle suggère que le

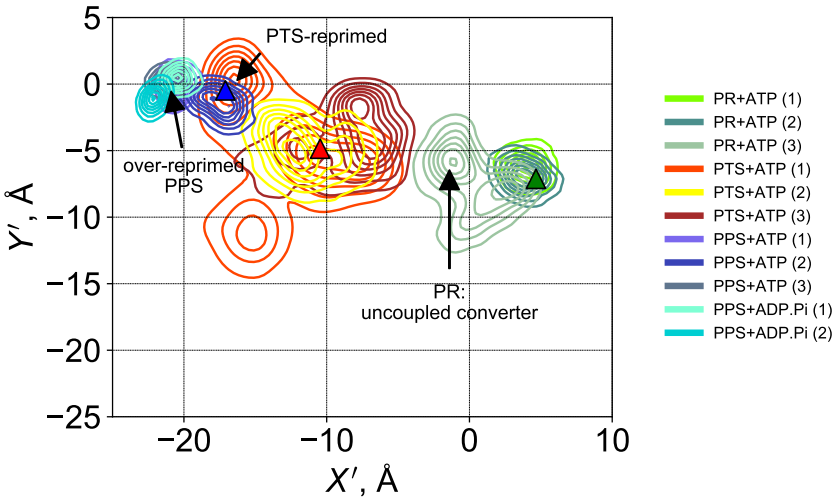


Figure 0.4.: Distributions statistiques de la position du convertisseur, décrite par les coordonnées locales X' et Y', observées dans les simulations de dynamique moléculaire du domaine moteur de la myosine VI.

PTS représente bien un bassin conformationnel avec une stabilité intrinsèque, comme illustré par la réversibilité de la transition. Finalement, aucun mouvement significatif du *switch II* n'est relevé dans la simulation du PTS, en dépit des importants mouvements du convertisseur. Si elle n'a pas valeur de preuve, cette observation est certainement cohérente avec l'idée que *switch II* et convertisseur (et par extension, bras de levier) ne sont que faiblement couplés pendant le *recovery stroke*.

Dans l'état PR, trois simulations indépendantes sur la même échelle de temps montrent que le convertisseur est moins dynamique qu'en PTS, parce qu'il est stabilisé par des contacts avec le domaine moteur absents en PTS. Cependant, dans une des simulations, un découplage spontané du convertisseur est observé - et est irréversible sur l'échelle de temps 300 ns. Lors de ce découplage, le convertisseur se déplace vers une nouvelle position, intermédiaire entre celle observée dans les structures cristallographiques PR et PTS. Cette observation est également cohérente avec l'hypothèse PTS, mais n'est certainement pas suffisante pour la valider. De plus, on remarque que ce mouvement de grande amplitude du convertisseur n'a aucun effet détectable sur le *switch II* - une nouvelle observation en faveur d'un couplage faible entre convertisseur et site actif.

Dans l'état PPS, dans la majorité des simulations le convertisseur relaxe vers une nouvelle position encore plus "basculée" que celle observée dans la structure cristallographique; les causes de cette relaxation ne sont pas établies pour l'instant. On note, cependant, que même dans cette nouvelle position le convertisseur présente des fluctuations positionnelles de faible amplitude, similaires à celles observées dans le bassin PR, et qui contrastent avec le "dynamisme" observé en PTS. L'observation la plus frappante émergeant des simulations de l'état PPS (lié à l'ATP) est celle d'une ré-ouverture du *switch II*; dans les trois simulations (indépendantes) de l'état PPS+ATP, la liaison hydrogène switch II-ATP se rompt rapidement, suivie dans deux simulations sur trois par une rupture du pont-salin critique.

En principe, une longue simulation de dynamique moléculaire non-biaisée initiée à partir de la structure PR liée à l'ATP devrait permettre de capturer une transition spontanée vers le PPS, c'est-à-dire un évènement de *recovery stroke*. En étudiant le mécanisme de cette transition, il serait possible de savoir si la structure PTS correspond ou non à un intermédiaire, et donc de valider ou invalider

l'hypothèse PTS. Cependant, les études expérimentales montrent que l'échelle de temps du *recovery stroke* est de l'ordre de la milliseconde, un ordre de grandeur typique pour les transitions conformationnelles complexes dans les protéines. Une telle durée est hors de portée d'une simulation non-biaisée compte-tenu de nos moyens de calculs actuels, d'autant plus si on prend en compte le fait qu'une étude robuste nécessiterait plusieurs répliques de cette simulation. En pratique, l'approche directe n'est donc pas réalisable du fait de la trop longue durée de la transition étudiée. Globalement, les simulations "non-biaisées" présentées ici ont permis de capturer des transitions partielles qui sont cohérentes avec l'hypothèse PTS, mais insuffisantes pour la démontrer. De même, l'absence d'observation de transitions directes PR ↔ PPS sans passer par PTS (ce qui invaliderait l'hypothèse PTS) pourrait s'expliquer par l'échelle de temps, trop courte, plutôt que par la très faible probabilité de telles transitions. Il est donc clair que les simulations non-biaisées ne permettent pas d'étudier efficacement le mécanisme du *recovery stroke*. Cependant, elles permettent de déterminer les caractéristiques dynamiques (distribution des observables pertinentes) des états extrémaux (PR et PPS) ainsi que l'état intermédiaire putatif (PTS), ce qui se révélera crucial pour analyser les résultats de simulations plus élaborées employées par la suite pour surmonter (ou esquiver) le problème de l'échelle de temps.

Exploration du paysage conformationnel du recovery stroke par "échantillonnage amélioré"

Les méthodes dites d'échantillonnage amélioré (*enhanced sampling*) permettent de circonvenir le problème d'échelle de temps *via* diverses stratégies. Notamment, la dynamique moléculaire accélérée (*Accelerated Molecular Dynamics*, aMD) introduit un terme supplémentaire dans le potentiel, ou *boost*, de manière à déstabiliser les configurations basses en énergie. Par conséquent, les barrières énergétiques séparant les minima locaux définissant les conformations se trouvent abaissées; et donc, les transitions sont plus rapides. À l'aide de cette méthode, des transitions conformationnelles de durée normalement milliseconde sont explorées dans des simulations sub-microsecondes. Cela fait de l'aMD une méthode de choix pour espérer capturer une transition spontanée tout en conservant un coût en calculs acceptable. Néanmoins, l'amélioration de l'échantillonnage permis par l'aMD nécessite généralement l'introduction d'un *boost* très élevé dans le potentiel, qui, en pratique, compromet la possibilité de re-scorer les configurations visitées selon leur poids de Boltzmann. Par conséquent, l'exploration facilitée permise par l'aMD est qualitative. Une variante plus récente de la méthode, la dynamique moléculaire accélérée gaussienne (*Gaussian aMD*, GaMD), tente de résoudre ce problème en construisant le *boost* de manière à préserver à la fois la capacité à re-scorer et l'amélioration de l'échantillonnage.

Dans un premier temps, nous avons mené des simulations de dynamique moléculaire accélérée à partir des états PR, PTS et PPS du domaine moteur de la myosine VI liés à l'ATP. Pour chaque état, les paramètres du *boost* aMD ont été estimés à partir des simulations conventionnelles évoquées plus haut; puis, deux simulations indépendantes ont été menées pour des durées de l'ordre de 100 ns. Ces simulations révèlent une grande richesse de comportements; notamment, après projection des trajectoires sur les observables caractéristiques du *recovery stroke* définies précédemment, on observe que l'espace conformationnel visité par les états PR et PTS, et PTS et PPS, se recouvrent. En fait, dans une des simulations de l'état PPS, une transition "inverse" vers le PTS est observée de manière spontanée (Figure 0.5). Dans cette simulation, les observables prennent des valeurs compatibles avec celles observées dans la structure cristallographique et les simulations non-biaisées du PTS. En outre, les simulations initiées depuis l'état PR explorent des configurations où le convertisseur a accompli un mouvement en direction du bassin PTS, et où l'hélice relais présente un coude (*kink*), mais où le switch

II n'est pas totalement fermé. La conclusion générale de ces simulations de dynamique moléculaire accélérée semble donc être que des transitions spontanées de PR vers PTS et PPS vers PPS, partielles ou totales, ont été capturées; à l'inverse, aucune transition directe entre PR et PPS n'est observée sur ces échelles de temps. Donc, ces simulations appuient l'hypothèse que le PTS est un intermédiaire du *recovery stroke*.

Cependant, le *boost* introduit par le protocole aMD a pour conséquence une certaine instabilité des structures secondaires (notamment, des dé-pliements d'hélices α). Ainsi, il est difficile de tirer des conclusions sur le mécanisme de la transition en conditions réelles à partir des trajectoires de dynamique accélérée.

La variante gaussienne, déjà mentionnée, repose sur l'introduction d'un *boost* nettement plus faible tel que l'analyse quantitative (calcul des poids de Boltzmann "débiaisés") soit possible. On s'attend donc à ce que des trajectoires de dynamique moléculaire accélérée gaussiennes permettent une étude mécanistique de la transition. Dans cette perspective, nous avons mené des simulations de GaMD à partir de l'état PR afin d'explorer de possibles transitions spontanées vers l'état PTS. Dans un premier temps, nous avons mené une "équilibration" avec le protocole spécifique de GaMD. Cette étape d'équilibration, détaillée dans le texte principal, est nécessaire pour obtenir un *boost* au potentiel qui vérifie les propriétés exigées. Dans notre cas, le temps total d'équilibration est de 100 ns, valeur élevée qui devrait assurer une estimation fiable des paramètres nécessaires à l'obtention du *boost*. Partant de la structure équilibrée, 5 simulations de GaMD indépendantes, chacune de 100 ns, ont été réalisées. Les résultats montrent que l'échantillonnage n'est essentiellement pas modifié par rapport aux simulations non-biaisées, et, en particulier, aucune transition hors du bassin PR n'est observée. Il semble que la préservation de la possibilité de *rescoring* se fasse au prix d'un *boost* si faible (quelques kcal mol^{-1}) que l'échantillonnage n'est pas presque pas perturbé; GaMD ne semble pas être une bonne stratégie pour l'exploration conformationnelle non-dirigée.

En dynamique moléculaire accélérée, l'amélioration de l'échantillonnage est obtenue en modifiant le paysage d'énergie potentielle. Il s'agit donc d'une stratégie "non-dirigée" au sens où le biais n'est pas appliqué sur une (ou plusieurs) variables collectives spécifiquement choisies pour décrire la transition étudiée. En aMD, les barrières de potentiel sont abaissées pour accélérer la dynamique, mais cette dernière n'est pas perturbée par ailleurs. Cela implique notamment qu'il faut attendre que la transition d'intérêt (ici, le *recovery stroke*) prenne place au gré des fluctuations, ce qui n'est pas différent de l'approche naïve basée sur la dynamique moléculaire conventionnelle à ceci près que le temps d'attente devrait être réduit. Par contraste, des approches de simulation plus directes peuvent être employées dans lesquelles un biais est appliqué pour explorer une transition donnée (par le truchement d'observables décrivant la transition). Les approches de ce type seront privilégiées dans la suite.

Paysage énergétique de la fermeture du switch II

Les calculs d'énergie libre géométriques sont une catégorie de méthodes numériques, inspirées par la mécanique statistique, qui visent à estimer le potentiel de force moyenne, ou profil d'énergie libre, le long d'une (ou deux, rarement plus) variable collective. Le profil d'énergie libre $F(\xi)$ le long d'une variable collective $\hat{\xi}(x)$ (fonction des coordonnées atomiques x) est défini, à une constante près, comme:

$$F(\xi) = -k_B T \ln \int e^{-\beta U(x)} \delta(\hat{\xi}(x) - \xi) \mathrm{d}x \tag{0.1}$$

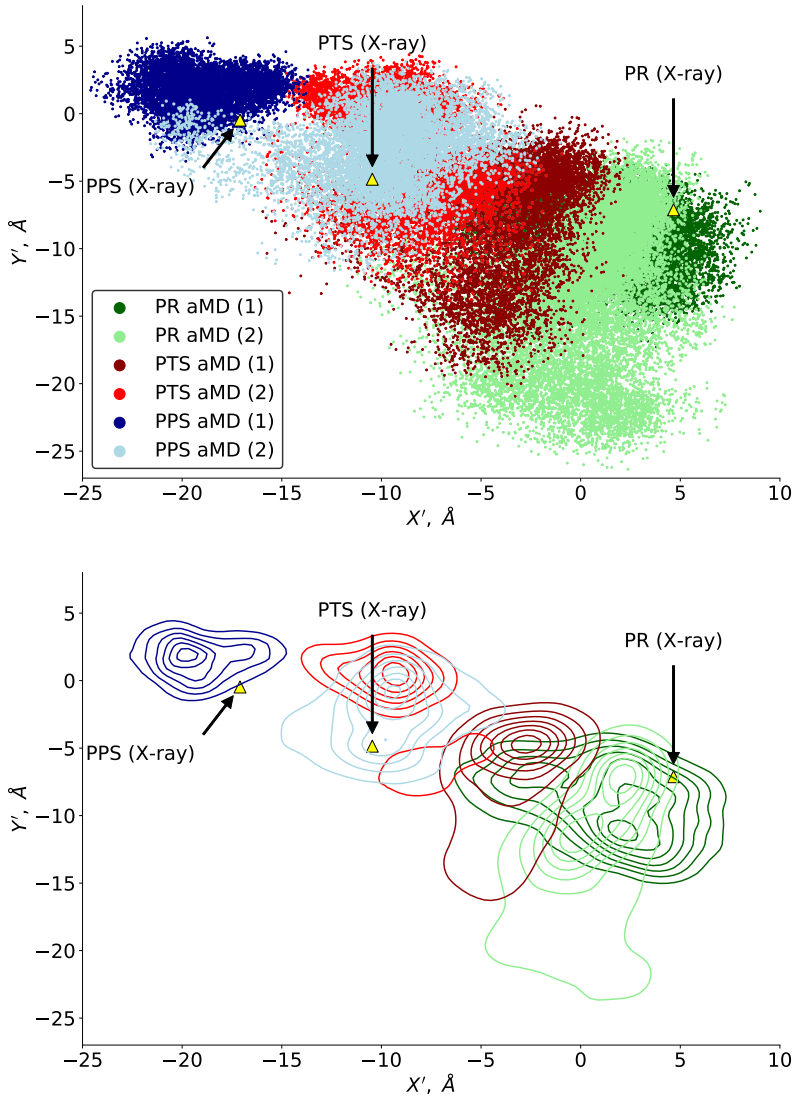


Figure 0.5.: Distribution de la position du convertisseur dans les simulations de dynamique moléculaire accélérée. Le recouvrement entre les différents états est très clair; de plus, on voit que la simulation PPS aMD (2), bien qu'initiée depuis la structure PPS, se stabilise dans le bassin PTS.

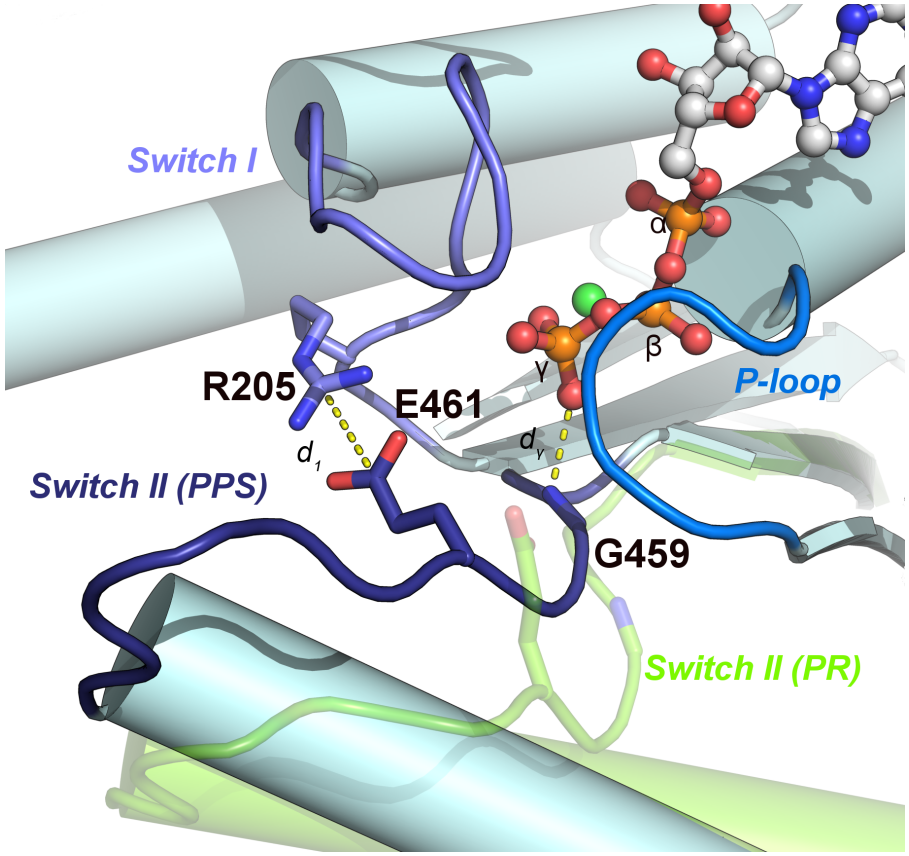


Figure 0.6.: Site actif de la myosine et définition des observables d_1 et d_γ caractérisant la fermeture du *switch II*.

où les quantités impliquées sont définies dans le chapitre 4 du texte principal. Le profil d'énergie libre représente le potentiel effectif dans lequel évolue la variable collective (traitée comme une variable dynamique) et contient toute l'information sur sa distribution d'équilibre; notamment, les minima représentent des états (méta-) stables. De plus, la hauteur des barrières d'énergie libre séparant les minima est reliée au taux de la transition entre ces minima; ainsi, si la transition entre deux états R et P présente une barrière $\Delta F^\ddagger$, le taux cinétique de la transition prendra la forme:

$$k_{R\to P} = A.e^{-\beta\Delta F^\ddagger} \tag{0.2}$$

où A est un facteur pré-exponentiel. Ainsi, les calculs d'énergie libre donnent accès aux états stables et à la cinétique associés à la transition décrite par la variable collective étudiée.

Grâce à des calculs d'énergie libre, nous avons caractérisé le paysage d'énergie libre gouvernant l'activation de l'activité ATPase (fermeture du *switch II*) dans les différents états conformationnels adoptés par le domaine moteur de la myosine. Ces calculs, réalisés avec la stratégie ABF (*Adaptive Biasing Force*) le long des observables d_1 et d_γ introduites plus haut (Figure 0.6), nous ont permis d'estimer le coût énergétique pour réaliser la fermeture du *switch II*, en fonction de la conformation générale du domaine moteur - et, donc, crucialement, de la position du convertisseur/bras de levier.

Le résultat le plus important émergeant de cette étude est que ce coût est de l'ordre de 10 kcal mol^{-1} dans les états PR et PTS, valeur élevée. Cette valeur correspond à la différence estimée d'énergie libre entre l'état le plus stable (identifié comme *switch II* totalement ouvert en PR, partiellement ouvert en PTS) et l'état *switch II* fermé (malgré tout identifié comme un minimum local par les calculs ABF),

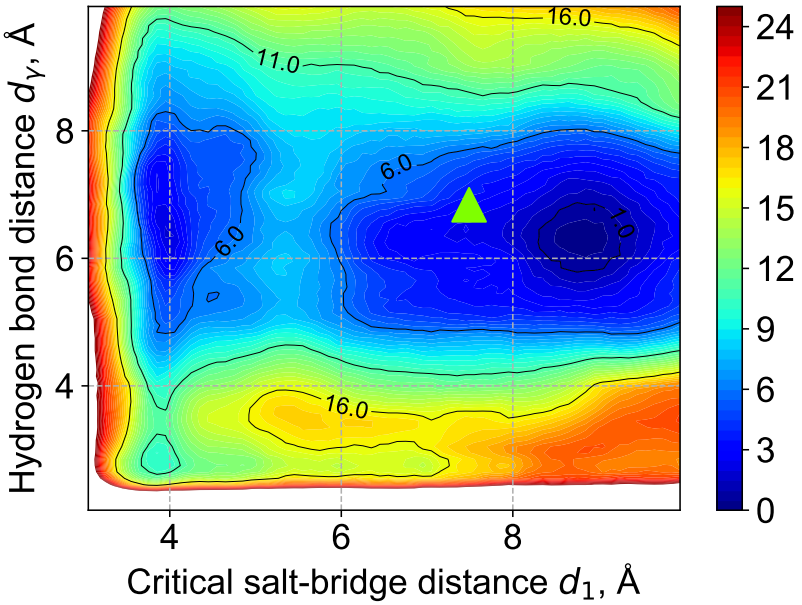


Figure 0.7.: Paysage énergétique associé à la fermeture du *switch II* dans l'état PR obtenu par des calculs ABF. Le bassin correspondant au *switch II* complètement fermé (et donc à une myosine catalytiquement active), identifié dans le coin inférieur gauche du graphe, est environ 10 kcal mol^{-1} plus haut énergie libre que l'état fondamental. Le triangle vert représente les valeurs des observables dans la structure cristallographique.

voir Figure 0.7 pour le PR.

On remarque donc que la fermeture du *switch II* est presque aussi défavorable en PTS qu'en PR, et ce malgré la rotation déjà significative du convertisseur en PTS. Cela confirme quantitativement l'hypothèse émise sur la base de la structure PTS et des simulations non-biaisées, à savoir que site actif et convertisseur sont faiblement plutôt que fortement couplés. Par ailleurs, ces calculs permettent dans le même temps d'explorer directement la possibilité, contraire à l'hypothèse PTS, que le *recovery stroke* est initié par la fermeture du *switch II*: ce scénario correspond à la fermeture du *switch II* dans l'état PR. En plus de la grande différence d'énergie libre entre les états ouvert et fermé en faveur du premier, nos calculs suggèrent que la barrière d'énergie libre pour la fermeture du *switch II* en PR est de l'ordre de 12 kcal mol^{-1}, valeur élevée. Nos résultats suggérent donc que la fermeture du *switch II* dans l'état PR est rare et transitoire, et donc qu'elle ne représente pas l'évènement initiateur du *recovery stroke*, comme nous le proposons dans (Blanc et al. 2018).

Cependant, pour valider cette conclusion, il reste à vérifier que le chemin de transition alternatif, à savoir celui qui implique une transition précoce vers le PTS suivie d'une fermeture tardive du *switch II*, est bel et bien moins coûteux énergétiquement que la fermeture précoce du *switch II*, prévue par exemple dans le modèle de Fischer. Pour cela, une compréhension plus fine du mécanisme de la transition PR→ PTS est nécessaire.

Mécanisme de la transition PR→ PTS et barrière dĕnergie libre

Le mécanisme de la transition PR→ PTS a été étudié dans un premier temps à l'aide de simulations de dynamique moléculaire ciblée (*Targeted Molecular Dynamics*, TMD) et dirigée (*Steered Molecular*

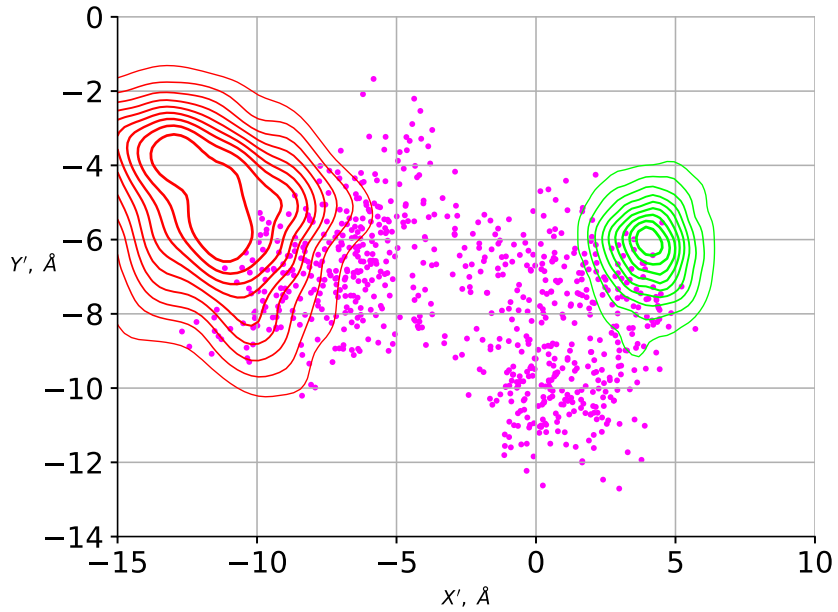


Figure 0.8.: Biaiser le réarrangement du sous-domaine Relais-SH1 est suffisant pour déplacer le convertisseur (points roses) de la position PR (lignes de densité vertes) à la position PTS (lignes de densité rouges).

Dynamics, SMD). Dans ces simulations, un biais harmonique dépendant du temps est appliqué sur une observable décrivant l'un des réarrangements élémentaires prenant place pendant la transition PR → PTS. L'intérêt de cette approche est que d'une part, le biais permet de s'affranchir du temps d'attente, puisque la transition est directement "pilotée"; d'autre part, en appliquant le biais seulement sur un sous-domaine, on peut observer la réponse structurale des autres sous-domaines au réarrangement du sous-domaine biaisé. Dans un premier temps, nous avons mené des simulations de TMD partant du PR et dans lesquelles le sous-domaine Relais-SH1 (hélice Relais, boucle Relais, hélice SH1) est biaisé vers sa conformation PTS (en appliquant un bias harmonique sur RMSD du squelette peptidique de ce sous-domaine). Nous observons qu'à l'échelle de temps 15 ns, le changement de conformation du sous-domaine Relais-SH1 est suffisant pour déclencher la rotation du convertisseur de la position PR vers la position PTS (Figure 0.8). Cette observation est robuste à des changements dans le protocole et la durée de simulation. Par contraste, si la rotation du convertisseur de PR vers PTS est biaisée, ce n'est pas suffisant en général pour déclencher le changement de conformation du sous-domaine Relais-SH1 (en particulier la formation du coude dans l'hélice relais). Finalement, dans aucune de ces simulations n'est observée une fermeture du *switch II*, appuyant encore une fois l'existence d'un couplage faible entre le site actif et la région "génératrice de force" composée du sous-domaine Relais-SH1 et du convertisseur.

Les informations collectées à l'aide de ces simulations biaisées ont ensuite été ré-investies pour explorer le paysage énergétique de la transition PR→ PTS. À cette fin, le paysage d'énergie libre conjoint pour la position du convertisseur, décrite par la variable X', et la conformation locale de l'hélice Relais dans la région du *kink*, décrite par la variable $\Delta RMSD_{kink}$, a été déterminée en utilisant une approche récente dite *extended ABF*, Figure 0.9.

D'une part, ces calculs, initiés de la structure PR, identifient un bassin d'énergie libre correspondant à l'état PTS (vérifié par inspection visuelle et comparaison avec les données de dynamique moléculaire

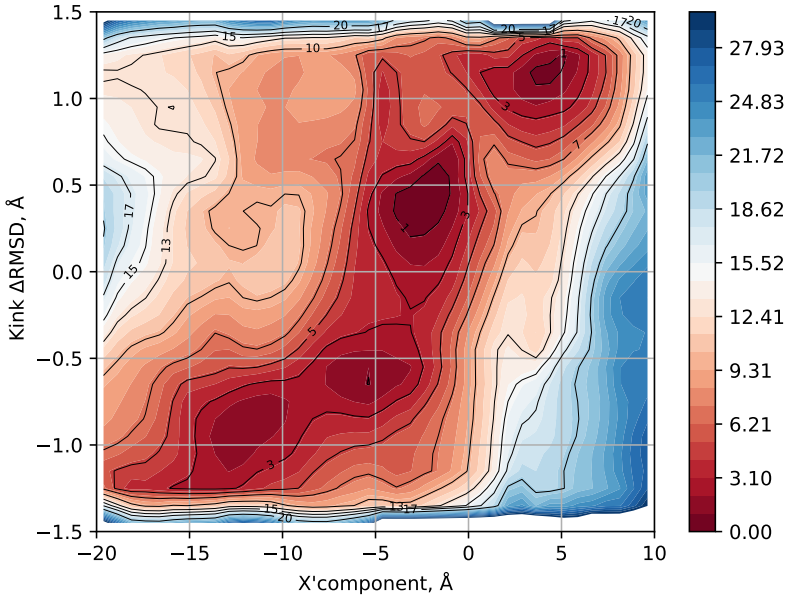


Figure 0.9.: Paysage d'énergie libre associé à la transition PR → PTS calculé par eABF.

non-biaisées); ils apportent donc un soutien indépendant à la pertinence de la structure PTS comme état conformationnel accessible depuis le PR et intermédiaire putatif du *recovery stroke*. En outre, les calculs révèlent que la plus haute barrière d'énergie libre pour la transition PR→ PTS n'excède pas 7 kcal mol^{-1}, contre 12 kcal mol^{-1} pour le mécanisme concurrent qui correspondrait à une fermeture précoce du *switch II* dans l'état PR. Dans la mesure où c'est la hauteur de la plus haute barrière qui contrôle la cinétique de la transition, nous en déduisons que le chemin PR→ PTS est exploré bien plus rapidement que le chemin qui commence avec la fermeture du *switch II*. Un calcul rapide, négligeant le facteur pré-exponentiel, montre ainsi que le taux attendu de la transition PR→ PTS est environ 4000 fois plus élevé que celui de la fermeture du *switch II* en PR. Ce point important appuie ainsi la pertinence du PTS comme intermédiaire - cinétiquement favorisé- du *recovery stroke*, et appuie du même coup le modèle à fermeture tardive du *switch II*.

Il faut néanmoins apporter plusieurs réserves. D'une part, pour des raisons techniques détaillées dans le corps principal du manuscrit, des méthodes de calculs d'énergie libre différentes, bien que proches, sont utilisées pour caractériser le paysage énergétique de la fermeture du *switch II* et celui de la transition PR→ PTS. De plus, les variables collectives utilisées comme supports des calculs ne sont pas les mêmes dans les deux cas. Il est donc possible que ces deux barrières ne soient pas directement comparables. Par ailleurs, une description complète de la cinétique devrait aussi inclure une estimation du facteur pré-exponentiel (qui est fonction d'un coefficient de diffusion dépendant de la position), que nous n'avons pas réalisée. Finalement, même dans le cas où notre estimation du rapport des taux cinétiques rend compte du comportement réel du système, elle ne concerne que les phases précoces de la transition. Or, dans le cadre de l'hypothèse PTS comme dans celui des modèles concurrents commençant par la fermeture du *switch II*, des réarrangements supplémentaires sont requis pour atteindre l'état PPS et terminer le *recovery stroke*.

En fait, même en l'absence de couplage fort, le modèle de Fischer et collaborateurs semble impliquer l'existence d'un intermédiaire structural du *recovery stroke* dans lequel le *switch II* fermé, mais le convertisseur n'aurait pas encore basculé (ou très partiellement) (Blanc et al. 2018; Koppole,

J. C. Smith, and Stefan Fischer 2007). Nous avons baptisé cet intermédiaire hypothétique FPI, pour *Fischer's putative intermediate* (intermédiaire hypothétique de Fischer). Dans le modèle de Fischer, une fois l'état FPI atteint, la formation du *kink* dans l'hélice Relais et le basculement de l'hélice SH1 (entre autres) conduisent à l'achèvement de la rotation du convertisseur. Par contraste, dans le modèle de la transition émergeant de l'hypothèse PTS, ces réarrangements ont lieu pendant la transition PR→ PTS et d'autres réarrangements, qui restent à caractériser, constituent la transition PTS→ PPS. Les résultats présentés ci-dessus suggèrent que la transition PR→ PTS est plus rapide que la transition PR→ FPI. Il est néanmoins possible que l'inverse soit vrai pour la seconde transition, *i.e.* que FPI→ PPS soit plus rapide que PTS→ PPS. Si c'est le cas, le chemin de transition correspondant au modèle de Fischer peut alors être cinétiquement privilégié par rapport au modèle incluant le PTS. Pour clarifier la situation et discriminer entre les deux modèles, il faut dans un premier lieu disposer d'une description de la transition PTS→ PPS. Dans un second temps, il faudra mettre en place un protocole pour comparer directement, et par la même méthode, les barrières d'énergie libre correspondant aux deux modèles concurrents.

Mécanisme de la transition PTS→ PPS

La comparaison des structures PTS et PPS permet d'identifier les réarrangements impliqués dans la transition PTS→ PPS. Les plus apparents sont la fermeture du *switch II* et la complétion de la rotation du convertisseur. De plus, la "crevasse" (*cleft*) entre les domaines U50 et L50 se ferme partiellement par une rotation du domaine L50, et l'hélice Relais subit un mouvement de corps rigide de type "chaise-à-bascule" (*seesaw*). Par des simulations de dynamique moléculaire ciblée (TMD), nous avons observé que forcer la fermeture du *switch II* depuis l'état PTS n'est pas suffisant pour induire les autres réarrangements et atteindre l'état PPS. Une autre série de simulations biaisées nous a cependant permis d'identifier un réarrangement jusque-là non-décrit (à notre connaissance) qui assiste la fermeture du *switch II*. Ce réarrangement consiste en un "échange" des interactions entre les brins-β 4, 5 et 6 du feuillet β central de la myosine, aussi appelé transducteur (*transducer*). Chacun de ces brins est directement lié à l'une des boucles du site actif, et le mouvement du brin 5 associé au *switch II* semble stabiliser un changement de conformation de ce dernier, favorisant l'établissement de la liaison hydrogène *switch II*-ATP. Cependant, cette transition locale du transducteur ne semble pas non plus couplée, sur l'échelle de temps de la simulation, avec d'autres réarrangements constitutifs de la transition globale vers le PPS. Afin d'explorer plus finement chacun de ces réarrangements, des variables collectives spécifiques ont été développées et chaque réarrangement a été "guidé" en SMD. Comme précédemment, et contrairement au cas de la transition PR→ PTS, cela ne suffit pas à déclencher les autres réarrangements.

Vers la validation de lhypothèse PTS

Modèle ratchet-like

De notre étude numérique de la structure PTS émerge un nouveau modèle mécanistique pour le *recovery stroke*, qui admet le PTS comme intermédiaire structural. Dans ce modèle, la transition est initiée par un mouvement du convertisseur, déclenché par les fluctuations thermiques et couplé au réarrangement du sous-domaine Relais-SH1. Le domaine moteur subit donc une première transition vers l'état PTS dans laquelle le convertisseur explore dynamiquement une variété de positions et d'états métastables partiellement ré-armés alors que le *switch II* est toujours ouvert. Par un mécanisme non-totalement élucidé, le domaine moteur subit alors une seconde transition, du PTS vers le

PPS, pendant laquelle la complétion de la rotation du convertisseur est couplée au mouvement de "*see-saw*" de l'hélice Relais, à la fermeture partielle de la crevasse L50/U50 et finalement à la fermeture du switch II qui rend possible la catalyse de l'hydrolyse de l'ATP. De plus, pendant cette transition, la fermeture du *switch II* est assistée par le réarrangement des brins β associés aux boucles du site actif. Donc, de l'hypothèse PTS découle un scénario original pour le *recovery stroke*, que nous qualifions de *ratchet-like*, parce qu'il implique que le "ré-armement" du moteur (rotation du convertisseur/bras de levier) est déclenché par les fluctuations thermiques et est essentiellement terminé lorsque le *switch II* se ferme et permet l'hydrolyse de l'ATP, qui stabiliserait l'état PPS. Ce modèle suggère donc un mécanisme possible par lequel la myosine capture des fluctuations conformationnelles productives pour progresser le long de son cycle moteur, ce qui contraste avec la vision classique où le *recovery stroke* est initié dans le site actif en réponse à l'attraction électrostatique exercée par l'ATP sur le *switch II* (modèle "fortement couplé" de Fischer). Ce modèle *ratchet-like* apparaît donc comme une alternative crédible aux propositions précédentes quant au mécanisme du *recovery stroke*. Plus généralement, il pourrait illustrer à l'échelle atomique un principe de fonctionnement des moteurs moléculaires.

Modèle ratchet-like vs modèle de Fischer

Si la comparaison des barrières d'énergie libre associées aux évènements initiateurs de la transition dans le cadre des deux modèles concurrents tend à favoriser notre modèle *ratchet-like*, il n'est pas établi à ce stade qu'il représente bien le modèle le plus probable. D'autres approches sont requises pour trancher - à commencer par la confrontation du scénario *ratchet-like* avec les données expérimentales disponibles sur le *recovery stroke*. Plusieurs études ont rapporté et caractérisé des mutants du *recovery stroke*, *i.e.* des mutations du domaine moteur affectant les transitions élémentaires impliquées dans le *recovery stroke*. De manière générale, les phénotypes de ces mutants sont en accord avec les prédictions du modèle de Fischer du *recovery stroke*, et sont donc avancés pour appuyer ce scénario mécanistique. Cependant, et comme nous le discutons dans notre publication (Blanc et al. 2018), ces résultats expérimentaux ne sont pas non plus en désaccord avec l'hypothèse PTS et le modèle *ratchet-like*. La raison en est que notre modèle implique les mêmes transitions élémentaires que le modèle de Fischer, mais dans un ordre différent et avec un couplage également différent. Les études de mutants peuvent montrer qu'une mutation donnée perturbe le *recovery stroke*, mais n'ont pas la résolution suffisante pour trancher entre les deux modèles concurrents parce que ces derniers correspondent à des prédictions essentiellement identiques. Il est important de rappeler, à ce stade, que le modèle de Fischer est tout-à-fait plausible, en particulier s'il est interprété de manière stochastique (ce qui est esquissé par Fischer et collaborateurs dans (Koppole, J. C. Smith, and Stefan Fischer 2007)) plutôt qu'en supposant un couplage fort. En fait, il semble raisonnable de penser que le modèle de Fischer dans son acception stochastique et le modèle *ratchet-like* construit à l'issue de nos propres travaux représentent deux chemins possibles pour le *recovery stroke*, dont il s'agit maintenant de décider lequel est le plus probable. Du point de vue de la mécanique statistique, cela correspond à l'idée que chacun de ces chemins peut être exploré par la myosine au gré des fluctuations thermiques, mais que l'un d'eux correspond au chemin dominant parce qu'il présente le taux le plus rapide, ou de manière (presque) équivalente les barrières d'énergie libre les plus basses. Le problème est donc maintenant d'évaluer ces barrières pour conclure. Par opposition aux calculs d'énergie libre déjà présentés, qui ont permis une comparaison préliminaire des barrières pour l'évènement initiateur de la transition, une comparaison suffisamment fiable pour trancher entre les deux modèles devrait prendre en compte toute la transition et utiliser une même méthode avec les mêmes variables collectives pour le calcul des barrières. Étant donné la complexité du *recovery stroke*, on ne peut pas utiliser un calcul d'énergie libre

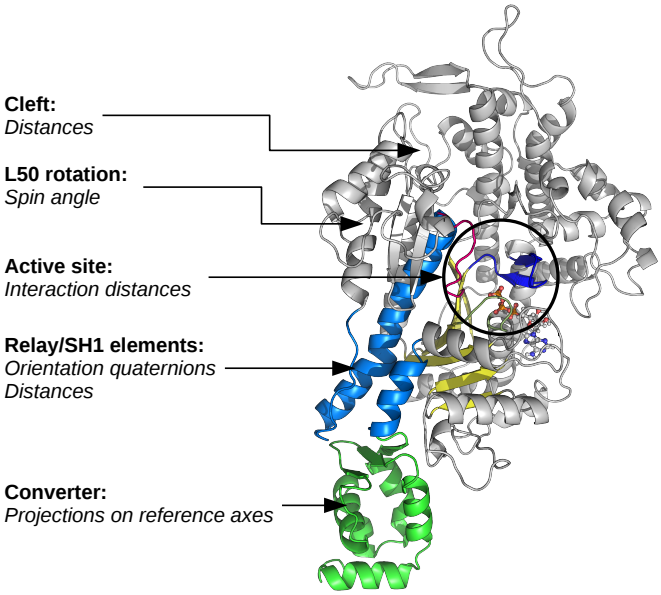


Figure 0.10.: Vue résumée des variables collectives utilisées pour décrire les réarrangements élémentaires prenant place pendant le *recovery stroke* pour les calculs de *string method*.

le long d'une ou deux variables collectives pour caractériser la transition. Au contraire, il faudrait, idéalement, utiliser au moins une variable collective par réarrangement élémentaire de sous-domaine, *i.e.* le mouvement du convertisseur, la fermeture du site actif, le changement de conformation du sous-domaine Relais-SH1, etc. Il est illusoire d'espérer obtenir le paysage d'énergie libre complet avec une si haute dimensionnalité, mais cela n'est en fait pas nécessaire. Seuls les chemins d'énergie libre minimale (qui sont des courbes uni-dimensionnelles définies dans l'espace de toutes les variables collectives choisies), correspondant au modèle de Fischer et au modèle *ratchet-like*, sont nécessaires; et des approches récentes permettent de les calculer.

Stratégie numérique pour la comparaison des modèles

La méthode de la corde (*string method*) s'est imposée ces dernières années comme la technique de référence pour déterminer le chemin de transition optimal associé à un changement conformationnel - ce qui est un préalable au calcul du profil d'énergie libre le long de la transition, et donc des barrières cinétiques. Cette méthode utilise une approche itérative pour relaxer une première approximation du chemin transitionnel, défini dans un espace de variables collectives de dimension arbitraire, vers un chemin d'énergie libre (localement) minimale. On peut montrer que ce chemin d'énergie libre minimale représente alors un excellent modèle du mécanisme de transition le plus probable, et constitue le point de départ optimal pour des calculs d'énergie libre et de taux cinétique. Cependant, la *string method* implique la simulation parallèle de nombreuses copies du système d'intérêt, et nécessite donc l'utilisation de ressources en calculs très importantes.

Concernant le *recovery stroke*, nous avons identifié au fil de notre étude une série de 25 observables qui caractérisent les différentes transitions élémentaires, résumées à la Figure 0.10 et détaillées dans le texte principal. Nous pouvons donc mener des calculs "*string method*" pour étudier le *recovery stroke*.

Cela suggère la stratégie suivante pour comparer les deux modèles; pour chacun des deux modèles:

1. Génération d'une première approximation du chemin de transition par SMD
2. Optimisation vers le chemin d'énergie libre minimale correspondant par la *string method*.
3. Calcul du profil d'énergie libre le long du chemin (voir détails dans le texte principal)

Une fois cela fait, les barrières d'énergie libre peuvent être directement comparées et on peut conclure quant au taux cinétique relatif des deux chemins de transition, et donc quant à celui qui décrit le mécanisme dominant. Il s'agit néanmoins d'une série de calculs très coûteux, pour lesquels nous avons récemment reçu une allocation PRACE de 17 millions d'heures scalaires. L'exploration du modèle *ratchet-like* par la *string method* est en cours. Les premiers résultats suggèrent que ce modèle admet bien un chemin d'énergie libre minimale, dans lequel le PTS est un intermédiaire et la fermeture du *switch II* intervient à la fin de la transition. Cela nous rassure quant à la plausibilité du modèle, mais n'apporte pas plus d'information quant à sa comparaison avec le modèle de Fischer. L'étude du modèle de Fischer impliquera notamment de construire un modèle de l'état FPI à partir de ses caractéristiques attendues (switch II fermé, hélice Relais non-coudée mais ayant subi la "chaise-à-bascule", convertisseur partiellement ré-armé). Nous prévoyons de construire ce modèle à l'aide de longues simulations de SMD.
En conclusion, nous avons étudié par des simulations moléculaires les implications de l'hypothèse PTS pour la transduction chimio-mécanique chez le moteur myosine. De cette hypothèse découle un mécanisme original pour le *recovery stroke*, que nous avons appelé le modèle "*ratchet-like*" pour insister sur l'importance qu'il donne aux fluctuations conformationnelles. Ce modèle est plausible, cohérent avec les observations expérimentales, mais ces dernières ne permettent pas de le discriminer par rapport aux modèles concurrents. Cependant, nous avons mis en place une stratégie prometteuse, mais coûteuse, pour trancher entre les modèles concurrents du *recovery stroke*. En outre, on peut remarquer que notre approche fournira la séquence détaillée des transitions élémentaires impliquées dans le *recovery stroke*, ainsi que le profil d'énergie libre associé; le résultat final sera donc une description complète (structurale et énergétique) de la transduction chimio-mécanique de la myosine.

Au-delà du domaine moteur: études numériques du bras de levier et du domaine queue

Hormis l'étude du *recovery stroke*, qui représente le sujet principal de cette thèse, nous avons également étudié deux problèmes liés aux myosines à l'aide de simulations moléculaires. Nos résultats ont contribué à des publications (Planelles-Herrero, Blanc, et al. 2016; Ropars et al. 2016).

Flexibilité du bras de levier de la myosine X

La myosine X est une myosine processive, impliquée dans plusieurs fonctions importantes comme la division cellulaire, et qui a la particularité de marcher préférentiellement sur des filaments d'actine réticulés (par une autre protéine, la fascine) que sur les filaments seuls comme le font les autres isoformes. La détermination des bases structurales de cette processivité différentielle est un sujet de recherche actif. L'équipe d'Anne Houdusse, au sein d'une collaboration impliquant cristallographes et spécialistes des expériences de motilité en molécule-unique, ont décrit comment la structure de la myosine X est adaptée à la marche sur l'actine réticulée, notamment grâce à la résolution de la structure cristallographique du dimère du bras de levier (Ropars et al. 2016). Cependant, les modèles

structuraux du dimère complet (*i.e.* incluant le dimère bras de levier mais aussi les deux domaines moteurs) construits à partir de cette nouvelle structure ne permettaient pas de rendre compte de certaines tailles de pas observées en molécule-unique. Nous avons alors mené une simulation de dynamique moléculaire de 100 ns, du dimère du bras de levier, en solvant implicite, pour en évaluer la flexibilité et caractériser les conformations alternatives potentiellement adoptée par la structure en solution. Les résultats montrent que le dimère relaxe vers une configuration plus allongée; si cette conformation est utilisée pour construire un modèle du dimère complet (incluant les deux têtes), il permet de rendre compte de l'une des tailles de pas observées expérimentalement. Donc, nos résultats de simulations aident à réconcilier la structure cristallographique avec les observations expérimentales (Ropars et al. 2016).

Dynamique conformationnelle du domaine MyTH-FERM dans la queue de la myosine

Le domaine queue est variable au sein de la superfamille des myosines, et permet la liaison de chaque isoforme avec ses partenaires cellulaires spécifiques. Les domaines MyTH-FERM se trouvent dans les queues de certaines isoformes, comme la myosine VII ou la myosine X, et sont importants pour plusieurs processus cellulaires critiques notamment grâce à leur capacité à lier les microtubules. Partant de nouvelles structures de domaines MyTH-FERM récemment résolues dans l'équipe d'Anne Houdusse, nous avons utilisé des simulations de dynamique moléculaire en solvant explicite pour caractériser leur dynamique conformationnelle. Les résultats montrent que le domaine MyTH-FERM tel qu'observé dans de la myosine, avec une organisation caractéristique dite "feuille-de-trèfle", présente une conformation stable à l'échelle de temps 30 ns. Par contraste, dans la taline, une protéine de jonction entre les intégrines et le cytosquelette qui présente un domaine FERM organisé linéairement plutôt qu'en feuille de trèfle, le domaine FERM fluctue beaucoup plus sur la même échelle temporelle. Ces résultats suggèrent que les différences de séquence entre boucles connectant les sous-domaines FERM peuvent en changer les caractéristiques dynamiques, ce qui pourrait contrôler la spécificité de liaison à différents partenaires (Planelles-Herrero, Blanc, et al. 2016).

1. Molecular machines and fluctuations

1.1. Brownian motion and thermal fluctuations at the nanoscale

Brownian motion was discovered by Scottish botanist Robert Brown in 1827, as the incessant and disordered movement of small (colloidal) particles in water under the microscope (Duplantier 2005). After considerable theoretical and experimental progress, it was recognized that this motion originates from the frequent collisions of the particles with the surrounding solvent molecules. Brownian motion is a manifestation of thermal molecular agitation. At finite temperature (*i.e.* non-zero[1]), molecules are animated by a fast, disordered motion (in the sense that the averages of the separate velocity components are zero); Brownian motion results from the frequent collisions of solvent molecules with the heavier - and observable through a microscope - colloidal particles. Although these collisions should average to a zero neat displacement of the particle (by isotropy of the solution, and if the particle is still with respect to the solvent), the force felt by the particle at time t is generally non-zero due to statistical fluctuations. This idea was notably formalized in the Langevin theory of Brownian motion, where the force exerted by the successive collisions is separated into an average force (which is zero if the particle is still) and a random component accounting for the statistical fluctuations. The Langevin equation for Brownian motion reads (in one-dimension):

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -\gamma\frac{\mathrm{d}x}{\mathrm{d}t} + L(t) \tag{1.1}$$

where m is the mass of the particle, x is its position, t is the time, γ is a friction coefficient and the random force $L(t)$ is a Gaussian white noise, satisfying:

$$\langle L(t)\rangle = 0 \quad \langle L(t)L(t+\tau)\rangle = 2\gamma k_B T\delta(\tau) \tag{1.2}$$

The first condition expresses that the fluctuations should average out; in fact, the average contribution is the friction term in equation 1.1. The second condition is a so-called *fluctuation-dissipation relation*; it relates the fluctuations (variance) of the random force to the dissipation property γ. Also, we find that the variance of the force is proportional to temperature - as is expected for a random force emerging from thermal fluctuations. Overall, the Langevin equation summarizes the most important features of physical behaviour at the nanoscale: friction and fluctuations - the ”molasses” and ”hurricane” in the words of Prof. Dean Astumian (Astumian 2007).

The development of a theory of Brownian motion (with seminal work by Einstein and Smoluchowski besides Langevin’s contribution) was one of the triumphs of the nascent statistical mechanics, and eventually led directly to the experimental demonstration of the existence of atoms by Jean Perrin (Perrin 2014, 1909). Arguably, this cemented statistical mechanics as one of the most relevant theories to investigate the behaviour of matter at the molecular scale. This remains true one century later, after decades of progress which notably witnessed the development of molecular simulation techniques (McCammon, Gelin, and Karplus 1977) and single-molecule experimental approaches allowing to di-

1. We will leave low-temperature quantum behaviour out of the discussion.

rectly visualize the thermal fluctuations of a molecular system (*e.g.* Rief 1997). In parallel, the rise of molecular biology over the 20th century shed light onto the molecular organization of the cell (Alberts 2008), and proteins with complex functions like transmembrane pumps or molecular motors were discovered, leading to the concept of *molecular machine*. Unsurprisingly, statistical mechanics provides the interpretative framework to make sense of experimental results and justify computational approaches to the study of molecular machines. In this thesis, we report on an investigation of the functional mechanism of the myosin molecular motor, an important molecular machine, by the means of molecular simulations.

1.2. Thermal fluctuations and biomolecular function

1.2.1. Functional dynamics

Obviously, biological systems when observed at the molecular scale have no reason not to undergo thermal fluctuations. The folding process itself (the phenomenon by which some proteins acquire their native three-dimensional structure) can be described as a diffusive walk on a multi-dimensional free energy landscape ending in a minimum corresponding to the folded state ("folding funnel" picture) (Karplus 2011; Socci, Onuchic, and Wolynes 1996). This does not exclude the possibility of 1) small oscillations around the minimum and 2) existence of several local minima, *i.e.* alternative conformations. These ideas seem to predict the existence of a conformational dynamics on at least two different time-scales: a fast, local, "in-basin" dynamics, and less frequent global conformational transitions corresponding to thermally-activated stochastic jumps between basins (see also 4.4). In fact, recent experimental results have shed light on the existence of Intrinsically Disordered Proteins (IDPs), for which no privileged tertiary structure is detected; rather, IDPs exist as ensembles of conformations (Uversky 2002). Unlike what a strict application of the structure-function relationship paradigm may imply, IDPs are not devoid of biological function; on the contrary, it appears that their flexibility and dynamical nature lie at the heart of their function by allowing, for example, their interaction with a wide variety of partners. More generally, a growing body of evidence shows that dynamics is functionally relevant even for folded proteins, that is, not only do proteins fluctuate (which is not so surprising), but these fluctuations are required for function (Henzler-Wildman and Kern 2007). Examples include the involvement of vibrational motions in enzyme catalysis (Hammes-Schiffer and Benkovic 2006; Hay and Scrutton 2012) and the importance of pre-existing equilibrium between an ensemble of conformations for biomolecular recognition (Boehr, Nussinov, and Wright 2009).

As reported in several places, notably by Yon-Kahn (2006) and Karplus (2006), it seems that the idea that proteins can exhibit large-amplitude, functionally relevant conformational fluctuations initially encountered some opposition in the structural biology community[2]. This is attributed to the

2. On this topic, the following anecdote is related by Lisa Pollack (Schlick 2012) about the career of Prof. Klaus Schulten, a pioneer in the field of biomolecular simulations:

> In fact, selling the usefulness of the computational microscope and the molecular dynamics approach in its very early stages was also a battle Schulten sometimes had to wage. In 1985, while still a professor in Munich, Schulten went to a supercomputing center in Illinois to run some calculations, and returned to Germany with a movie illustrating a protein in motion, based on molecular dynamics simulations. When Schulten showed the movie, one of his colleagues became quite enraged. "He got so upset when he saw it, he almost wanted to physically attack me," Schulten recounts. "He told everybody this is the greatest rubbish he'd ever seen in his life. He was a crystallographer who thought basically of proteins as some kind of Gothic cathedral that were cast in stone."

crucial importance of X-ray crystallography - a technique inherited from solid-state physics - in unraveling the structures of proteins. It makes sense that the very evocative, but static pictures of proteins obtained by crystallography may have biased the community towards a static representation. Similarly, in solution biochemistry experiments, the very large number of copies of the protein under study averages out the fluctuations, *e.g.* in the rate of an enzymatic reaction. The development of new experimental (Nuclear Magnetic Resonance, Small-angle X-ray scattering (SAXS), single-molecules) and computational/theoretical techniques (first and foremost, Molecular Dynamics simulations) nonetheless permitted to reconcile the static and dynamical points of view.

We may also conjecture that some biologists had a hard time accepting the fluctuating nature of proteins because the apparent lack of stability which it implies was seen as incompatible with the protein fulfilling its function[3]. This apparent paradox comes from a naive extrapolation of the functioning principles of macroscopic objects: if one compares a protein to a macroscopic device, such as a car, it is clear that having a dynamic, fluctuating structure can only be a disadvantage. However, this only points to the irrelevance of such a comparison: thermal agitation is an unavoidable phenomenon at the nanoscopic scale, which is the scale at which biological evolution started (Dawkins 1976; Hoffmann 2012). In fact, recent studies show that dynamical features can be optimized under selective pressure (Campbell et al. 2018). In some sense, what is surprising is not that fluctuating proteins perform their function at the nanoscale, but that the evolutionary process could build upon them to yield (relatively) stable living beings at the macroscopic scale. Thus, the consideration of structural fluctuations should be at the heart of an atomic description of protein function.

1.3. Molecular motors - from biology to chemistry

Several families of proteins, characterized in a wide range of species, have been experimentally demonstrated to exhibit directed movement along a filamentous track. Prominent examples include polymerases and helicases (which move along nucleic acid molecules), and cytoskeletal motors such as kinesins and dyneins (microtubules-related) and, of course, actin-related myosins (Howard 2001; Schliwa and Woehlke 2003; Ronald D Vale 2003; Ronald D. Vale and Milligan 2000). In other cases, consistent rotary motion is observed, such as in the Fo-F1 ATPase complex (Noji et al. 1997). The observation of such a directed movement is in apparent contradiction with the isotropic character of thermal agitation. Indeed, one may imagine using the directed motion to work against a force, as has been done in single-molecule experiments (Karagiannis, Ishii, and Yanagida 2014). Thus, the second-law of thermodynamics dictates that this motion be coupled with an exergonic process which provides the thermodynamic driving force for the movement.

Macroscopic engines, as idealized by the reversible Carnot engine, operate by exploiting the spontaneous (exergonic) flow of heat from a hot source to a cold source to generate work. By analogy, may molecular motors also exploit a temperature difference? The answer is negative, as there is no known biological process that could maintain a stable temperature gradient along a molecular filament. Cells are isothermal environments. Rather, it has been shown that these motors are chemically-fuelled: the directional motion is coupled with an exergonic chemical reaction, the hydrolysis of Adenosine Tri-Phosphate (ATP) in the case of cytoskeletal motors. This is called *chemo-mechanical coupling*, or *chemo-mechanical transduction*.

Other types of couplings are found in biological systems (Hill 2005). For example, in chemo-osmotic coupling, the free energy from an exergonic chemical reaction is used to drive solute molecules

3. Imagine trying to tighten a screw with a wobbly, constantly fluctuating screwdriver.

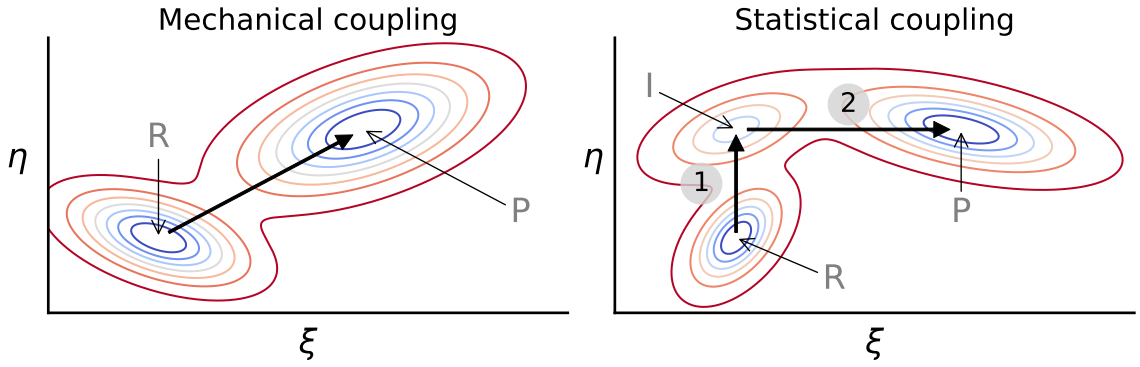


Figure 1.1.: Schematic illustration of mechanical (left) and statistical (right) couplings for the conformational transitions of molecular machines. Black arrows materialize (qualitatively) the privileged directions of transitions. R, reactant state; P, product state; I, intermediate state.

across a membrane against their concentration gradient. The only thermodynamic requirement is that the overall free energy change be negative. This thermodynamic criterion should not make us overlook that a physical agent, *i.e.* a molecular machine, should exist for the coupling to actually take place. Thus, the question becomes as to how the thermodynamic principles of free energy transduction are effectively realized in given molecular machines; and, after our above discussion, we may expect conformational fluctuations and transitions to play an important role (Hoffmann 2012). In this thesis, by focusing on a particular step of the myosin motor cycle, we aim at furthering our understanding of the structural mechanism of chemo-mechanical transduction in ATP-powered motors. As we will see, this will entail the detailed mechanistic description of the associated conformational transitions.

1.3.1. Strong coupling and statistical coupling in conformational transitions

We now introduce an important concept in the description of conformational transitions, *i.e.* tertiary rearrangements of the atomic structure between local minima on the free energy surface. We consider a transition of an un-specified molecular machine between two states, conventionally called Reactant (initial state) and Product (final state). To achieve a proper energetic and structural description of the transition, a customary first step is to introduce *collective variables* which describe its various aspects: for example, if two subdomains undergo an internal rearrangement during the overall conformational transition, one should choose two collective variables (say ξ and η), each describing a sub-domain. In this context, the question arises of the coupling between these two (or more) conformational degrees of freedom. Figure 1.1 represents schematically two extreme cases for the features of the free energy landscape along ξ and η. On the left, the minimal free energy path connecting the reactant and product states (which describes the dominant transition pathway) is diagonal, in such a way that any change along ξ is accompanied by a change along η; in this case, the conformational transition is *strongly coupled* or *mechanically coupled*. Mechanical coupling is how most macroscopic machines operate. By contrast, on the right the change along η is nearly complete before changes along ξ become energetically favorable, and there exists an intermediate state I where η is rearranged but not ξ. In this case, the coupling is referred to as *statistical*, because the probability of capturing a transition along ξ is conditioned by the value of η.

When studying a functional conformational transition, the nature of the coupling is an important

question which should be addressed for a proper description. Of course, the scenarios illustrated on Figure 1.1 are limiting cases, and intermediate cases are expected in real systems. Also, in the most frequent situation where more than two collective variables are needed to describe the transition, the nature of the coupling may differ between individual pairs of observables.

Let us note that the notion of strong coupling may also have another meaning in the context of molecular machines (Hill 2005; Hoffmann 2012; Karplus and Gao 2004). From a thermodynamic point of view, molecular machines operate by coupling an endergonic process (non-spontaneous) with an exergonic process (spontaneous), such that the overall free energy change is negative. For example, we may consider a hypothetical transmembrane pump whose functional cycle involves the translocation of a single molecule L against its concentration gradient (endergonic) for one ATP molecule hydrolyzed (exergonic in cellular conditions). This molecular machine will be said *strongly coupled* or *tightly coupled* if the translocation of one L molecule is always accompanied by the hydrolysis of one ATP molecule. Because of thermal fluctuations, it is nevertheless expected that some events of wasteful hydrolysis will happen (*i.e.* one ATP hydrolyzed but no L translocated). These events, referred to as "slippage", make it so that the average number of translocated L molecules per hydrolyzed ATP is lower than one. If this average number is significantly lower than one, the machine will be termed *weakly coupled*. Although this definition of strong vs weak coupling is of interest in the study of molecular motors, we will focus on the other definition, as we are interested in elucidating the mechanism of conformational transitions and as such investigating the coupling between relevant degrees of freedom. Thus, unless stated otherwise, any references to strong/statistical coupling is to be understood in terms of the definition introduced at the beginning of this paragraph. As a conclusion to this discussion, let us remark that these two distinct concepts are not unrelated; to understand this, let us consider what happens during a slippage (wasteful hydrolysis) at the conformational level. In a simple structural model, the pump exhibits a nucleotide binding site which can exist in an inactive (open) and an active (closed) configuration, and described by a collective variable ξ; and a L-binding site, which can be either inward facing (say, before translocation) or outward facing (after translocation), described by a collective variable η. If ξ and η are strongly coupled in the sense of Figure 1.1, the inward $\rightarrow$ outward transition is strongly coupled to the open-close transition of the active site, in such a way that the hydrolysis of ATP always results in the translocation of one L molecule; thus the machine is strongly coupled in the second sense. If ξ and η are statistically coupled, there will exist an intermediate I (Figure 1.1, right) where the active site is closed but the inward $\rightarrow$ outward transition has not happened. If there exists a pathway for the release of the hydrolysis products which is itself not strongly coupled with the inward $\rightarrow$ outward transition, there will be a wasteful hydrolysis; thus, statistical coupling of the conformational degrees of freedom would result in weak coupling in the overall cycle.

1.3.2. Artificial molecular machines

The design and synthesis of artificial molecular machines, exhibiting comparable characteristics to biological ones, has been a long-standing goal in chemistry and in fact forms one of the bases of nanotechnology (Kay and Leigh 2015). These last decades, considerable progresses have been made in supra-molecular chemistry, notably with the invention of mechanically interlocked chemical motifs (catenanes and rotaxanes) which provide basic ingredients to elicit relative motions of structural elements in synthetic chemical architectures (Dietrich-Buchecker et al. 2003; Sauvage 1998). The novelty and promising nature of these works has been recognized by the recent award (2016) of the Nobel Prize in Chemistry to Jean-Pierre Sauvage, Sir J. Fraser Stoddart, and Ben Feringa (Feringa

2016; Sauvage 2016; Stoddart 2016). Yet, no human-made molecular machine comes near the complexity and efficiency of their biological counterparts.

1.3.2.1. Molecular machines, molecular motors, molecular switches

To discuss artificial molecular machines, and compare them to biological ones, we need to clarify the difference between a *molecular machine*, a *molecular motor* and a *molecular switch*. A molecular machine may be loosely defined as a molecular assembly capable to perform complex operations like realizing some form of free energy transduction or generating movement. A molecular *motor* is defined as a molecular machine capable of *sustained* production of *net* mechanical force and/or directed displacement (linear or rotary) over repeated transitions. By contrast, a molecular *switch* is defined as a molecular machine capable of force/motion generation, but not of net displacement over repeated transitions. Let us clarify. Myosin is a molecular *motor*, because repeated cycles of ATP hydrolysis drive processive (net) displacement along actin, as demonstrated for several isoforms. This is possible because, in the motor cycle of myosin (Figure 2.5), the actin-bound force generating step (powerstroke) is compensated by an *off-actin* re-priming step (recovery stroke); if the recovery stroke happened while myosin were bound to actin, the forward movement generated during the powerstroke would be cancelled. So, net displacement over repeated cycles (and by extension, production of work) requires that the movement-generating step and the re-priming step occur by distinct pathways (linear motors) or that periodic motion be possible (rotary motors). On the opposite, if we consider a muscle-like artificial architecture such as the one discussed in Appendix B (Figure B.1), the relative sliding of the two sub-units upon pH change can produce force; but, the only way to ”re-prime” the assembly for a new force-producing transition is to undo the sliding by the same pathway. Thus, such an architecture cannot produce sustained work by repeated transitions; it is a *switch*, rather than a motor. While switches are relatively simple to realize (one essentially needs a bistable architecture with a way of modulating the relative stabilities of the two conformations, usually by the means of a chemical modification), motors are more challenging. We note that switches by themselves open exciting technological possibilities, *e.g.* for binary information storage.

1.3.2.2. Externally-controlled and autonomous molecular machines

In general, artificial molecular machines are driven by an external, human-controlled power input. For instance, in the rotary motors by Feringa and co-workers, continuous light irradiation is used to trigger the isomerization of a central double bond, followed by unidirectional relaxation of the unstable conformation so-generated through the use of asymmetric blocking groups (Feringa 2001). If irradiation is stopped, rotation stops as well. Contraction or extension in artificial muscle-like systems is triggered by an external change in chemical conditions (*e.g.* pH). Repeated sequences of contraction/extension ”cycles” are possible (although they will not generate net work), but require that the pH of the solution be externally modulated accordingly. Thus, such molecular machines are not *autonomous*. This is in stark contrast with the ability of biological molecular motors to operate continuously in homogeneous steady-state conditions, *i.e.* while the concentration of the ”fuel” molecule (ATP) is kept constant. It is only very recently that an artificial, autonomous chemically-driven molecular motor has been synthesized by the group of David Leigh (M. R. Wilson et al. 2016). The theoretical analysis of the design principles of this motor sheds light on general principles for molecular motor operation (Astumian 2016). Similarly, it is expected that a better understanding of the functioning principles of biological molecular machines, whose ability to perform efficient free energy transduction by exploiting fluctuations has been optimized through billions of years of evolution, will translate into novel design

strategies for artificial machines. And, while theoretical considerations are certainly of tremendous interest for this purpose (Astumian 2016, 2012; Hill 2005), they should be complemented by detailed case-studies illustrating at atomic-resolution the structural basis for free energy transduction. By their unique ability to give access to the thermal dynamics and energetics of complex molecular systems, molecular dynamics simulations and free energy calculations are ideally suited to such a goal (Singharoy and Chipot 2016). In this thesis, we explore the structural mechanism of chemo-mechanical transduction in the myosin motor using these computational strategies.

2. The myosin superfamily

Myosins are a wide superfamily of actin-based, ATP-powered motor proteins involved in a range of crucial functions including muscle contraction, endocytosis, cell migration and motility or intracellular cargo traffic. This thesis aims to address some aspects of the myosin motor mechanism by the means of molecular computational biophysics approaches.

2.1. Brief historical perspective

Initial studies of myosin are indissociable from early research in muscle biology and the elucidation of the muscular contraction mechanism. As reported by Andrew G. Szent-Györgyi, myosin was first isolated by Kühne in 1864 as a protein extract from muscle - hence the name myosin, which derives from the Greek root for muscle (Szent-Györgyi 2004). Adenosine Tri-Phosphate (ATP), first identified in 1929, was proposed in 1934 to be the energy source for muscular contraction. As such, the discovery in 1939 by Engelhardt and Lyubimova of the ATPase activity of myosin was critical. Albert Szent-Györgyi and co-workers later showed (1942) that the so-called myosin was in fact a mixture of two proteins, one of which retained the name myosin and the other was called actin. They showed that the addition of ATP to an actin/myosin (actomyosin) extract significantly decreased its viscosity; also, they demonstrated that actomyosin threads shortened in presence of ATP. These important observations were crucial in establishing myosin and actin as key players in muscle contraction. Subsequent work on skeletal muscle revealed the cellular organization of myocytes: actin and myosin are found within organized subcellular structures called sarcomeres and belong to different types of ”bands” observed in optical microscopy. In 1954, H.E. Huxley and E. Jean Hanson, and A.F. Huxley and R. Niedergerke, independently showed how muscular contraction proceeds by relative sliding of actin and myosin filaments in a sarcomere, leading to the now widely accepted *sliding filament theory* (A. F. Huxley and Niedergerke 1954; H. E. Huxley and Hanson 1954). These investigators speculated on the existence of an interaction between actin and myosin, responsible for force generation by the muscle, but the molecular details were unclear at this time. Slightly later A.F. Huxley proposed a model of muscle contraction which postulated the existence of a ”myosin side-piece” protruding from the myosin filament towards actin, and behaving as an elastic element able to store thermal energy and interact with actin (A. F. Huxley 1957). At about the same time, H.E. Huxley discovered the actomyosin cross-bridges, and identified them as protrusions from the myosin filament interacting with actin (H. E. Huxley 1957). Some years later, it was recognized by Reedy and co-workers that the cross-bridges could rotate, as they adopt a 45° angle relative to actin when bound to actin in the absence of ATP (*rigor*), but a 90° angle in relaxed muscle (Reedy, Holmes, and Tregear 1965). This eventually led to the *swinging cross-bridge model*, which explains sarcomere filament relative sliding by the rotation of the myosin head while it is bound to actin.

Finally, the resolution in the early 1990s of the first cryo-EM and crystallographic structures of the motor domain revealed that the swinging element was made only of the extended ”lever-arm” domain rather than the entire myosin head (Rayment, Holden, et al. 1993; Rayment, Rypniewski, et al. 1993; Schröder et al. 1993). The *swinging lever-arm theory*, which represents the current general consensus

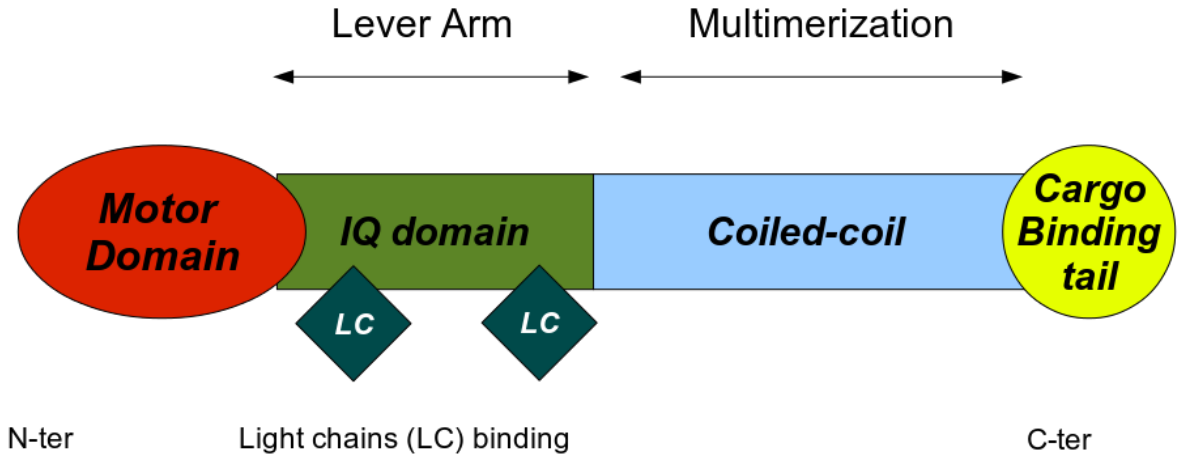


Figure 2.1.: Schematic primary structure of a myosin.

as to the functioning principle of the myosin motor, states that ATP binding, hydrolysis and the release of hydrolysis products are coupled to local structural rearrangements in the active site and its vicinity, which are amplified into larger swings of the lever-arm (Holmes 1997). These findings turned out to be general for the myosin superfamily as structural and functional studies of non-muscular myosins progressed.

2.2. Generalities on the myosin superfamily

2.2.1. Unconventional Myosins

Myosins are first and foremost known as a fundamental player of muscle contraction. However, after decades of study of muscular myosins, it was realized that the myosin superfamily actually includes many non-muscular members with various fascinating properties. These isoforms are referred to as unconventional myosins, and turn out to be involved in a very wide range of processes in Eukaryotic cells, including cell motility, intracellular cargo transport, endocytosis, phagocytosis, cell division, and so on (Batters and Veigel 2016; M. A. Hartman et al. 2011; Ross, Ali, and Warshaw 2008). In fact, phylogenetic analyses concluded that no less that 35 classes of myosins exist (Foth, Goedecke, and Soldati 2006; Odronitz and Kollmar 2007; Richards and Cavalier-Smith 2005).

The myosin superfamily is characterized by a well conserved motor domain (Sweeney and Houdusse 2010b), carrying the ATPase activity, and generally exhibits a consensus primary structure with, besides the motor domain, a lever-arm (at least partially made of IQ motifs which bind calmodulin light-chains), a dimerization/multimerization region (typically a coiled-coil), and a tail-domain involved in specific binding to cellular partners (Figure 2.1).

Among unconventional myosins, *processive* myosins exhibit the striking capacity to generate sustained directional motion on actin, see Figure 2.2. Consistently, these myosins are key players in intra-cellular cargo transport. All processive myosins except myosin VI are plus-directed, *i.e.* they progress towards the plus-end of the actin filament. The peculiarities which allow minus-directed dis-

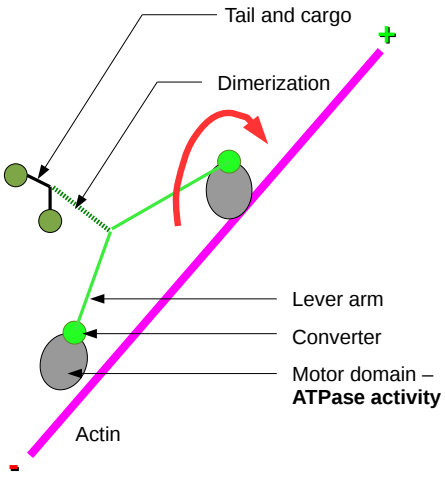


Figure 2.2.: Schematic organization of a processive myosin.

placement for myosin VI will be reviewed later on. Other myosins, such as some myosin I isoforms, have been shown to represent force-sensors rather than transporters, and are involved in the biochemical responses of the cell to changes in applied mechanical forces (Greenberg and Ostap 2013).

2.2.2. Myosins in pathological contexts and myosin-targeting drugs

Unsurprisingly given their biological importance, mutated (defective or over-expressed) myosins are found to be involved in several serious pathologies such as hypertrophic cardiomyopathy (Geisterfer-Lowrance et al. 1990), hereditary deafness (Boëda et al. 2002), and several forms of cancer, among others (Preller and Holmes 2013).

Consequently, a wide diversity of small-molecule allosteric effectors of myosins have been identified (reviewed in Preller and Holmes 2013). Among these, the recently discovered *Omemcamtiv Mecarbil* stands out as a (cardiac) myosin activator, *i.e.* it increases the power output of the cardiac muscle, which has important implications for the treatment of heart failure (Malik et al. 2011; Morgan et al. 2010). Also, at high concentration, it has been shown to rescue processivity on a mutated form of myosin VI (Pylypenko et al. 2015). The exact mechanism by which this activation is achieved is not yet completely understood (Hashem, Tiberti, and Fornili 2017; Planelles-Herrero, J. J. Hartman, et al. 2017; Rohde, Thomas, and Muretta 2017; Winkelmann et al. 2015).

2.3. General structure of the motor domain

The structure of the motor domain is well conserved between the different classes of myosins (Sweeney and Houdusse 2010b). It is composed of four large subdomains connected by loops and flexible joints, see Figure 2.3.

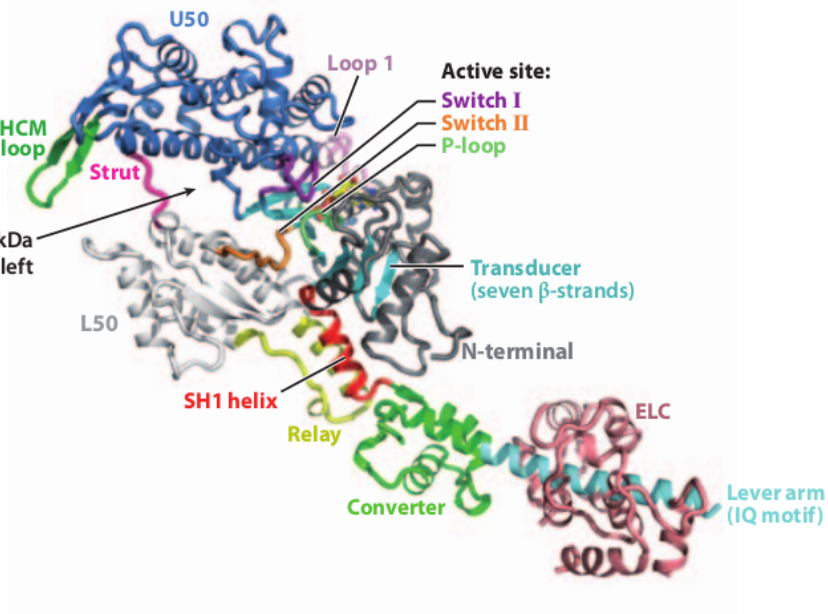


Figure 2.3.: Structural elements of the myosin motor domain. Image taken without modification from (Sweeney and Houdusse 2010b).

The active site is located at the interface between the Upper 50 kDa (U50) and the N-terminal subdomain. It is formed by three nucleotide-binding loops: the P-loop (that binds the Mg^{2+} ion and the phosphate tail), switch I, and switch II, which binds the γ-phosphate in its closed conformation, allowing catalysis. This pattern is called a Walker-motif and is shared notably by G-proteins and kinesin (R. D. Vale 1996; Walker et al. 1982). It is extremely well-conserved among myosin motors. Switch II closure upon the γ-phosphate of ATP turns on the catalytic activity. The structure of the active site of *Dictyostelium discoideum* myosin II (Dd myo2) in the Pre-Powerstroke State (PPS) state is detailed on figure 2.4.

This structure (1VOM) was the first published Pre-Powerstroke State (PPS) structure and was solved with ADP+Vanadate (C. A. Smith and Ivan Rayment 1996). The ADP.Vanadate "molecule" is generally considered to mimic ADP.Pi. The active site of 1VOM is thus representative of a post-hydrolysis state. Upper 50 kDa (U50) and Lower 50 kDa (L50) are linked by several loops and separated by a large cleft, whose closure is involved in the binding to actin through the interaction with four loops: HCM loop (Hypertrophic Cardiomyopathy loop), loop 2, loop 3 and loop 4.

The fourth subdomain, the converter, is linked rather loosely to the rest of the protein *via* the Relay (a structural element formed by the Relay Helix, RH, and the Relay Loop, RL) and the SH1 helix. The rotation of the converter is key for translating and amplifying small structural changes in the motor domain into a large swing of the lever-arm. This latter, in the direct continuity of the converter, is an extended subdomain typically formed by repeated IQ motifs capable of binding calmodulin (or calmodulin-like) light chains. The swinging lever arm theory predicts that the step size of a (processive) myosin is positively correlated to the length of its lever, *i.e.* the number of IQ motifs (Purcell et al. 2002). Myosins VI and X were at some point believed to challenge the theory, as they exhibit small numbers of IQ motifs but large step sizes; in fact, it was realized that other structural elements may contribute to extending the reach of the lever arm (Ropars et al. 2016; James A. Spudich and Sivaramakrishnan 2010).

Since 1993, dozens of myosin structures have been published. They span many isoforms, ligands

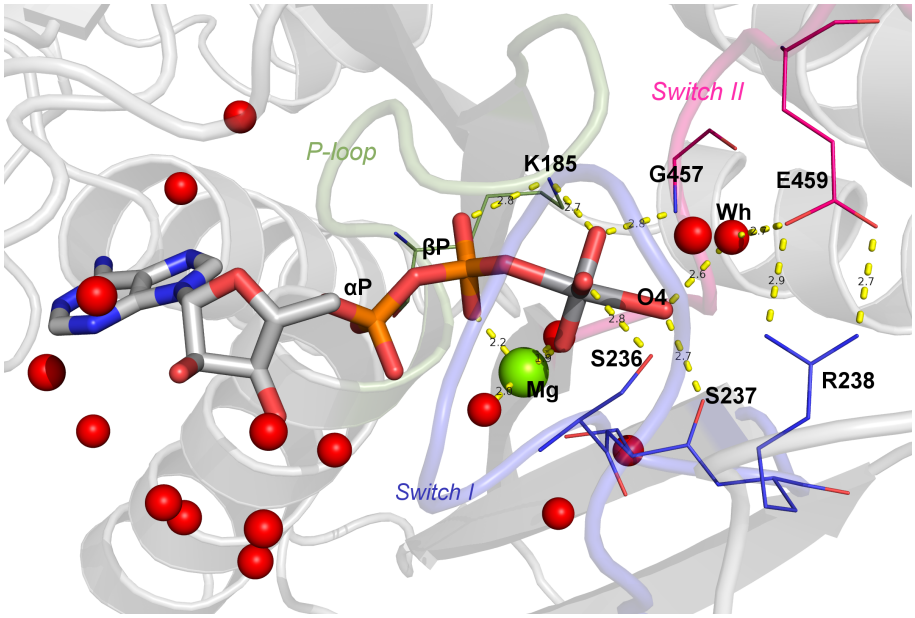


Figure 2.4.: Close-up on the active site of myosin II in the PPS state (1VOM).

and crystallization conditions but can be mapped to a limited number of conformations. Most of these conformations are considered to reflect the actual functional states assumed by the motor during the motor cycle. They differ by the relative arrangement and the conformation of the above mentioned structural elements.

2.4. The motor cycle of myosin

By combining solution and structural data, an integrated picture of the myosin motor cycle - also called Lymn-Taylor cycle - was progressively obtained. The widely accepted view of the cycle describes how myosin explores a series of structural states with variable affinity for both actin and ATP/ADP.Pi, in such a way that a forward swing of the lever-arm is performed when myosin is strongly bound to actin - the *powerstroke*. The cycle is closed by an off-actin re-priming step, the *recovery stroke*, in which the reverse rotation of the lever-arm ”re-cocks” the motor in preparation for the next powerstroke. ATP hydrolysis occurs at the end of the recovery stroke.

Historically, states with low-affinity for actin were crystallized first; this allowed the early identification of the Post Rigor State (PR) and Pre-Powerstroke State (PPS) structures and the first insight into the coupling between the catalytic site and the rotation of the converter during the recovery stroke (Chapter 5), see Geeves and Holmes 1999, and references therein. Also, the description of the myosin active site initiated the debate as to the phosphate exiting mechanism after hydrolysis, which is still not settled today (Cecchini, Alexeev, and Karplus 2010; Llinas et al. 2015; Preller and Holmes 2013; Rayment, C. Smith, and R G Yount 1996; Ralph G. Yount, Lawson, and Ivan Rayment 1995).

A crystal structure of the actomyosin complex is still unavailable. However, in 2003 was solved the *Rigor-like* structure of myosin V (in absence of actin) which was shown to be representative of the Rigor state (Coureux et al. 2003; Sweeney and Houdusse 2004). This structure revealed that the U50 - L50 cleft closes upon binding to actin. Upon binding of ATP to the Rigor state, the cleft re-opens, nullifying the affinity for actin and releasing the motor. This Rigor to Post-Rigor conformational transition has been well described by computational approaches, which have shown how the binding of

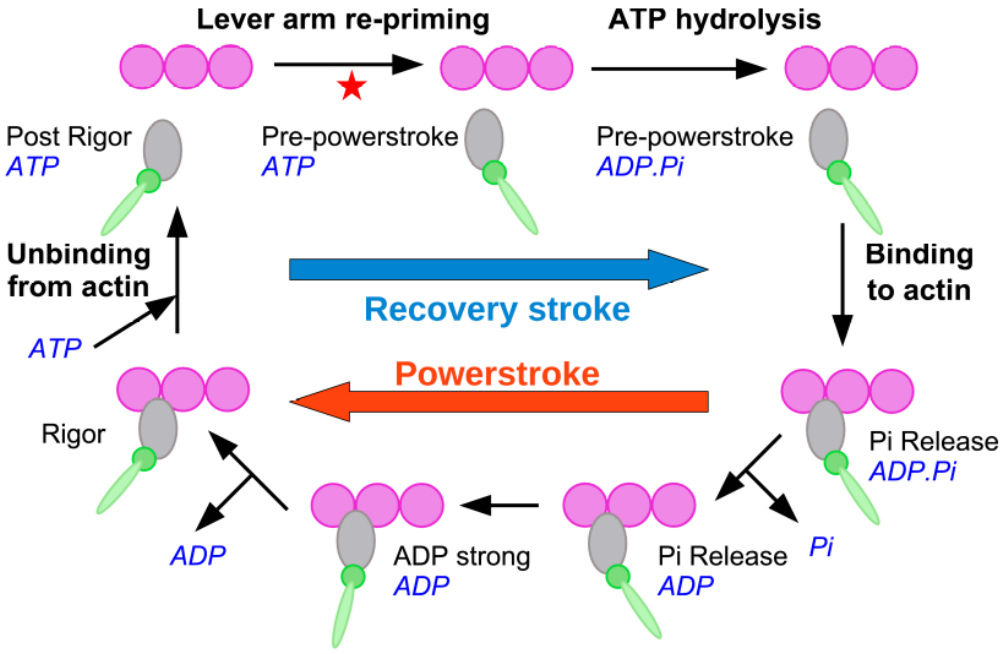


Figure 2.5.: Motor cycle of myosin, taken from (Blanc et al. 2018). The red star indicates the putative position of the PTS structure, investigated in this thesis.

ATP triggers a movement of the P-loop and switch I coupled to the opening of the cleft (Cecchini, Houdusse, and Karplus 2008; Ovchinnikov, Trout, and Karplus 2010). Also, the central β-sheet or transducer untwists, which is believed to introduce a mechanical strain in the motor domain, subsequently relieved during the powerstroke.

2.4.1. The powerstroke and the mechanism of force production

As per the current consensus, myosin generates force upon strong interaction with actin (Sweeney and Houdusse 2010b). The sequential release of the hydrolysis products - Pi, then ADP - is coupled to the forward swing of the lever-arm sub-domain, which is thought to be responsible for generating directed motion. The mechanistic interplay between the formation of the interaction interface with actin, the release of the products, and the swing of the lever is still the topic of considerable debate (Houdusse and Sweeney 2016; Sweeney and Houdusse 2010b). In addition, the very idea of a powerstroke, *i.e.* that force is produced through the elastic relaxation of the ”cocked” lever-arm, has been challenged in favour of a purely Brownian mechanism (Astumian 2015; Karagiannis, Ishii, and Yanagida 2014).

Note that although the motor cycle is shared by all myosin isoform, the transition rates and in particular the rate-limiting steps are isoform-specific and allow for specific adaptations for a given function (De La Cruz et al. 1999; Howard 2001). For example, muscular myosin II spends a short-fraction of its cycle strongly bound to actin (low duty-ratio); since muscular force is generated by the addition of elementary steps performed by an array of myosin heads, it ensures that each head can contribute to the overall shortening of the sarcomere without interfering with the other heads. By contrast, two-headed processive myosins like myosin V exhibit a high-duty ratio (close to 0.5), ensuring that one head remains attached to actin at all time. Overall, Bloemink and Geeves proposed a functional clas-

sification of myosin isoforms based on the general biophysical properties of their motor cycle (duty ratio, load dependence, etc); they identified 4 categories, *i.e.* fast movers (*e.g.* muscular myosins for fast muscle fibers), slow movers (*e.g.* muscular myosins for slow fibers), sensors (*e.g.* myosin Ib) and processive transporters (*e.g.* myosins V and VI) (Bloemink and Geeves 2011). Interestingly this functional classification does not match the phylogenetic repartition in myosin families.

2.5. Myosin VI, a minus-directed processive motor

2.5.1. Structural bases of minus-directed motion

The reverse directionality of the myosin VI (myo6) isoform was discovered in 1999 (Wells et al. 1999); in this study, it was found that the rotation of the lever-arm in myosin VI occurs in the opposite direction as compared to other known isoforms. This striking observation initiated the search for the structural bases of reverse directionality, which were finally elucidated through a combination of structural and single-molecule studies, reviewed in (Sweeney and Houdusse 2010a). Briefly, sequence analyses revealed two unique myosin VI insertions, insert 1 near the nucleotide-binding-site and insert 2 at the junction between the converter and lever-arm. It was eventually showed that insert 2 is responsible for minus-directed motion (Bryant, Altman, and J. A. Spudich 2007; H. Park et al. 2007), while insert 1 is involved in modulating the access of ATP and ADP to the nucleotide-binding site. Crystal structures of myosin VI in the Rigor and PPS states revealed that insert 2 introduces a sharp turn at the beginning of the lever-arm, which re-orients it with respect to the standard direction observed in other myosin isoforms (Ménétrey, Bahloul, et al. 2005; Ménétrey, Llinas, Mukherjea, et al. 2007). This reverses the direction of the lever-arm swing during the powerstroke, driving the motor's displacement towards the minus-end. In addition, a unique conformational transition of the converter was discovered upon resolution of the PPS structure: while the converter takes on the canonical "R-fold" observed in virtually all myosin isoforms in the Rigor and Post Rigor State (PR) states (Ménétrey, Llinas, Cicolari, et al. 2008), it is found in a novel "P-fold" in the PPS state, along with the recently solved Pi-Release state (Llinas et al. 2015). It was shown that this transition assists in increasing the amplitude of the powerstroke, allowing for a larger step size, by contributing an extra component to the lever-arm swing. Moreover, computational studies of this transition have suggested that it is responsible for the experimentally observed variability of step-size distribution in myosin VI (Ovchinnikov, Cecchini, Vanden-Eijnden, et al. 2011). Finally, we note that the timing of the P $\rightarrow$ R transition of the converter during the powerstroke is still unclear (Ménétrey, Isabet, et al. 2012; Mugnai and Thirumalai 2017).

2.5.2. Biological roles of myosin VI

Myosin VI can play the role of a transporter, but also an actin-based anchor (Sweeney and Houdusse 2010a, 2007). Myosin VI, initially discovered in *D. melanogaster*, was found to play a key role during spermatogenesis and cell division in this organism. In Mammalians, myosin VI is required for normal endocytosis and protein secretion. Also, it is involved in regulating epithelial cell migration and the maintenance of stereocilia in sensory hair cells. The reader is referred to (Sweeney and Houdusse 2010a, 2007, and references therein) for a detailed overview. From a pathological perspective, myosin VI has been implicated in ovarian and prostate cancers (Dunn et al. 2006; Yoshida et al. 2004). Moreover, mutated, defective myosins cause hereditary deafness in Human and Mouse (Ahmed et al. 2003; Sweeney and Houdusse 2007).

3. Molecular Dynamics simulations

We introduce the Molecular Dynamics (MD) simulation methodology, in which the equations of motion of each atom of a molecular system are numerically integrated so as to investigate its dynamic and thermodynamic properties. First, strategies for the numerical integration of the equations of motion are briefly discussed. Second, the functional form of the interaction potential (force-field) is reviewed, and we illustrate how *ad hoc* modifications to the potential can be used to enhance the sampling. Third, essential notions of statistical mechanics are introduced to show how constant-temperature simulations can be achieved. We aim for a brief overview rather than comprehensiveness; the interested reader will find deeper presentations of MD and related methods in (Frenkel and Smit 2002; Leach 2001; Tuckerman 2010, among others).

3.1. Numerical integrators for Molecular Dynamics

At the heart of MD is the numerical integration of the equations of motion; it is required that the simulated trajectory be time-reversible and energy-conserving, so as to mimic as much as possible the properties of Hamiltonian dynamics. It turns out that the most straightforward scheme for numerical integration of differential equations, (explicit) Euler method, does not meet these requirements. Instead, more robust integrators must be developed. We will detail the popular Verlet's integrator along with the Brooks-Brünger-Karplus (BBK) integrator, suited for stochastic dynamics. Theoretical accounts on the integration of Hamiltonian dynamics can be found in (E. Hairer, Lubich, and Wanner 2006; Tuckerman 2010).

3.1.1. Verlet integrator

Numerical integration of ordinary differential equations begins with a discretization of time, introducing a time-step Δt. Then, the point of the integration algorithm is to allow for the iterative approximation of the solution to the equation, starting from the known initial conditions. Typically, the derivation of an integrator starts with a Taylor expansion. To derive Verlet's integrator, let us thus expand the atomic position to second order:

$$x(t + \Delta t) = x(t) + \Delta t \dot{x}(t) + \frac{1}{2}\Delta t^2 \ddot{x}(t) + \mathcal{O}(\Delta t^3) \tag{3.1}$$

and, backward in time:

$$x(t - \Delta t) = x(t) - \Delta t \dot{x}(t) + \frac{1}{2}\Delta t^2 \ddot{x}(t) + \mathcal{O}(-\Delta t^3) \tag{3.2}$$

Summing up equations 3.1 and 3.2 yields:

$$x(t + \Delta t) = 2x(t) - x(t - \Delta t) + \Delta t^2 \ddot{x}(t) + \mathcal{O}(\Delta t^4) \tag{3.3}$$

which, after truncation at the 2nd order, defines the Verlet integrator. Using Newton's second law (A.4), one can replace $\ddot{x}(t)$ by $f(t)/m$ where the force f derives from a user-defined interaction potential called force-field (see next section). The computation of the forces, rather than numerical integration, is the most computationally expensive part of MD.

Several remarks are in order:

- **Error** The 3rd-order terms cancel out since $(-\Delta t)^3 = -\Delta t^3$; only even powers are left in the expansion. This means that the error on the integration is of 4th order; by contrast, the error is 3rd-order with the explicit Euler method, because there is no such cancellation.
- **Reversibility** The integrator is invariant under the time reversal transformation $\Delta t \to -\Delta t$, which shows that it is time-reversible.
- **Velocity** The velocity v does not appear explicitly; if needed, it can be calculated using a centered finite-difference scheme, $v(t) = \frac{1}{2\Delta t}\left(x(t+\Delta t) - x(t-\Delta t)\right)$.
- **Energy conservation**: it can be shown that Verlet's integrator conserves energy very well for small enough Δt, see (Frenkel and Smit 2002).

This integrator was popularized in the MD community by Verlet (1967), hence its designation. However, it has a more ancient history. Norwegian physicist C. Störmer used it in 1907 to study *aurora borealis*, and French astronomer J.B. Delambre used it (around 1792) to construct astronomical tables. Also, it seems that I. Newton himself uses the method in the *Principia*. These historical points are related in more details by Ernst Hairer, Lubich, and Wanner (2003).

We note that several variants of the Verlet integrator exist, the most popular ones being velocity-Verlet and leap-frog; see *e.g.* (Frenkel and Smit 2002) for review. Also, note that it can be shown that the maximal allowed time-step is conditioned by the fastest oscillatory motions of the system under study. For molecular systems, this generally restricts the time-step to 1 fs if hydrogen atoms are allowed to oscillate about covalent bonds. By using a constraining algorithm, one can freeze these covalent bonds involving hydrogens and use a 2 fs time-step, accelerating the simulation (Ryckaert, Ciccotti, and Berendsen 1977).

3.1.2. Brünger-Brooks-Karplus integrator for Langevin dynamics

As we will see later (section 3.3.4.1), it is of interest to simulate the following dynamics:

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -\nabla_x U - \gamma\frac{\mathrm{d}x}{\mathrm{d}t} + L(t) \tag{3.4}$$

where $L(t)$ is a random force. Given the presence of a velocity-dependent force, Verlet's integrator must be modified. A popular integrator for equation 3.4 is the BBK integrator, introduced in (Brünger, C. L. Brooks, and Karplus 1984).

To derive BBK, we start from the Verlet integrator and insert the expression for the acceleration:

$$\ddot{x}(t) = \frac{1}{m}\left[-\frac{\partial U}{\partial x} - \gamma\dot{x}(t) + L(t)\right] \tag{3.5}$$

$$x(t+\Delta t) = 2x(t) - x(t-\Delta t) + \frac{\Delta t^2}{m}\left(-\frac{\partial U}{\partial x} - \gamma\dot{x}(t) + L(t)\right) \tag{3.6}$$

Then, we insert the finite-difference estimate of $\dot{x}$:

$$x(t+\Delta t) = 2x(t) - x(t-\Delta t) + \frac{\Delta t^2}{m}\left(-\frac{\partial U}{\partial x} - \gamma\left(\frac{x(t+\Delta t) - x(t-\Delta t)}{2\Delta t}\right) + L(t)\right) \quad (3.7)$$

Rearranging yields:

$$x(t+\Delta t)\left[1+\frac{\gamma\Delta t}{2m}\right] = 2x(t) - x(t-\Delta t)\left[1-\frac{\gamma\Delta t}{2m}\right] - \frac{\partial U}{\partial x}\frac{\Delta t^2}{m} + L(t)\frac{\Delta t^2}{m} \quad (3.8)$$

and we finally arrive at the BBK integrator:

$$x(t+\Delta t) = \frac{1}{1+\frac{\gamma\Delta t}{2m}} \cdot \left[2x(t) - x(t-\Delta t)\left[1-\frac{\gamma\Delta t}{2m}\right] - \frac{\partial U}{\partial x}\frac{\Delta t^2}{m} + L(t)\frac{\Delta t^2}{m}\right] \quad (3.9)$$

Note that the integration of the stochastic equation 3.4 by the means of integrator 3.9 requires the use of a random number generator to draw values for the random force $L(t)$.

3.2. Classical energy models for molecular simulations

A detailed description of a molecular system, including its electronic properties, should be treated by quantum mechanics (Szabo and Ostlund 1996). However, quantum methods require a very large amount of computational resources, which generally makes them unsuitable for simulations of large molecules and/or long timescales. Simplified, classical, functional forms for the potential energy are usually adopted instead - this approach is called molecular mechanics. A *force field* refers to a given functional form, and the associated parameters (Leach 2001). Perhaps the most conceptually important difference from quantum approaches is that there is no explicit treatment of the electronic degrees of freedom - atoms are represented as point-like particles without finer structure. The assumption is that the potential energy of the system can be written as a function of the nuclear coordinates only, which is a consequence of the Born-Oppenheimer approximation. This notably implies that, unless specific refinements are introduced, force-fields cannot account for covalent bond breaking/formation.

Unlike quantum methods which are (at least in principle) parameter-free, force fields need to be parametrized. This involves adjusting the parameters so as to reproduce reference experimental results and/or quantum calculations. A requirement for a good force-field is that the parameters derived from fitting to a reference data set are transferable to different (generally larger and more complex) systems, for which the reference quantum calculations are impossible to perform in the first place. Usually, the simplicity of the force field expression comes at the expense of generality; one cannot expect a force field to be accurate in reproducing all properties of interest. Instead, a given force field will generally be parametrized to be applied on a certain category of molecules (*e.g.* proteins, sugars), and/or to yield accurate predictions for a given quantity of interest (*e.g.* the condensed phase properties of a pure liquid). A wide variety of force-fields have been developed to model as wide a range of molecular systems. Notably, several force-fields dedicated to the simulation of biomolecules, particularly proteins, have been developed and improved over the years. Popular examples include AMBER (Cornell et al. 1995), OPLS (Jorgensen, Maxwell, and Tirado-Rives 1996), GROMOS (Oostenbrink et al. 2004), and CHARMM (A. D. MacKerell et al. 1998). Although most force-fields share a common "backbone" (*e.g.* the use of harmonic potentials to model covalent bonds), there exists par-

ticular refinements which constitute each force-field's identity. In the following, we will focus on the CHARMM force-field, since it is the one used in this thesis.

The full CHARMM potential U takes the form:

$$U(x) = U_b(x) + \sum_{\text{atom pairs}} (U_{LJ} + U_{elec}) + U_{CMAP} \tag{3.10}$$

The meaning of each term is explained below. To perform an MD simulation, one needs the forces deriving from the potential U, given by $f = -\nabla U$.

3.2.1. Bonded terms

Most often, a force-field represents the interactions associated to covalent bonds (the so-called "bonded terms", also "internal terms") by harmonic potentials applied on inter-atomic distances and angles. The force constant is determined so that the harmonic potential fits the local minimum of the reference interaction energy (as computed, for example, by quantum mechanics) which corresponds to the covalently bound state. Reference (or "equilibrium") values of each term are also obtained in this manner. In the CHARMM potential function, bonded terms read (B. R. Brooks et al. 2009; Bernard R. Brooks et al. 1983):

$$\begin{aligned} U_b(x) = & \sum_{\text{bonds}} K_b \left(b - b_0\right)^2 \\ & + \sum_{\text{angles}} K_\theta \left(\theta - \theta_0\right)^2 \\ & + \sum_{\text{Urey-Bradley}} K_{UB} \left(S - S_0\right)^2 \\ & + \sum_{\text{dihedrals}} K_\varphi \left(1 + \cos(n\varphi + \delta)\right) \\ & + \sum_{\text{impropers}} K_\omega \left(\omega - \omega_0\right)^2 \end{aligned} \tag{3.11}$$

b is the bond length between two covalently joined atoms; θ is the angle formed by three consecutive atoms; φ is the dihedral angle formed by four consecutive atoms. ω is the "improper" dihedral angle between non-consecutive atoms; its essential purpose is to maintain planarity, for example in the case of benzene rings. Finally, the Urey-Bradley term applies to the distance S between two atoms separated by a third atom, and thus involved in an angle θ. Its purpose is to restrict the movement of the two bonds around the central atom. All the reference values for these potential functions are *parameters* of the force-field, and their values depends on the specifics of the atoms involved in the bond, angle, etc.

3.2.2. Non-bonded terms

Non-bonded terms refer to the components of the force field which model non-covalent interactions, *i.e.* Van der Waals interactions and electrostatic interactions.

3.2.2.1. Lennard-Jones potential for Van der Waals interactions

The Van der Waals interactions are traditionally modelled using the Lennard-Jones potential, or 6-12 potential. In the case of two identical atoms separated by a distance r, the Lennard-Jones potential $U_{LJ}(r)$ writes:

$$U_{LJ}(r) = \frac{A}{r^{12}} - \frac{B}{r^6} \tag{3.12}$$

where A and B are constants that will be discussed shortly. The $1/r^6$ corresponds to the distance-dependence of the three types of Van der Waals interaction energies (Keesom, Debye and London)[1]. The repulsive term in $1/r^{12}$ has no theoretical justification, but arguably reproduces well-enough the steep increase in potential energy when two atoms are brought close enough that their electronic clouds repel each other (because of the Pauli exclusion principle). Also, it is often stated that the choice of $1/r^{12}$ was a computing trick to accelerate calculations (because $1/r^{12} = (1/r^6)^2$) used in the early days of molecular simulations.

Another, equivalent formulation of the Lennard-Jones potential is in terms of the "energy wells depth" ε and the distance r^{min} which minimizes the interaction energy:

$$U_{LJ}(r) = \varepsilon \left[\left(\frac{r^{min}}{r} \right)^{12} - 2 \left(\frac{r^{min}}{r} \right)^{6} \right] \tag{3.13}$$

ε and r^{min} are characteristic of a given atom type. When atoms of different types i and j are interacting through the Lennard-Jones potential, a combination rule should be chosen. In CHARMM, $\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}$ (geometric mean) and $r_{ij}^{min} = (r_i^{min} + r_j^{min})/2$ (arithmetic mean).

Due to its long-range dependence in $1/r^6$, the Lennard-Jones potential becomes quickly negligible as distance increases. As such, it is customary to truncate the Lennard-Jones interactions, *i.e.* to compute the interactions only within a cut-off. Another reason for this truncation is that one should avoid interactions between periodic images when a molecular simulation is performed using periodic boundary conditions; thus, (formally) infinite range interactions should either be truncated or treated by special methods. In practice, the truncation is not abrupt; rather, the potential function is smoothly switched so as to ensure that the (modified) interaction energy is exactly zero at the cut-off distance. This is required for energy conservation, and to avoid artefactual forces at the cut-off distances. Also, note that the introduction of the cut-off *per se* does little for performance improvement, because deciding whether two atoms are within cut-off distance or not still requires the full calculation of the pairwise distance list, which is the most expansive part of the calculation. Instead, neighbour lists are introduced and updated for only a fraction of the time-steps (*e.g.* , one step over 20), see (Leach 2001).

3.2.2.2. Electrostatics

In most force-fields including CHARMM, electrostatics is described only in terms of Coulombic interactions between fixed atomic charges. The interaction potential takes the usual form:

$$U_{elec}(r) = \frac{q_i q_j}{4\pi\epsilon_0 \epsilon r} \tag{3.14}$$

1. To be more rigorous, one should mention that Keesom forces (permanent dipole - permanent dipole interactions) and Debye forces (permanent dipole - induced dipole interactions) exhibit the typical dependence in $1/r^6$ only after thermal averaging over the rotational degrees of freedom.

where q_i and q_j are the respective charges of atoms i and j, r the separation distance between them, $\epsilon_0 = 8.85 \times 10^{-12}\,\mathrm{F\,m^{-1}}$ is the vacuum permittivity, and ϵ is the relative dielectric constant of the medium (set to 1 in explicit solvent simulations). Computational procedures exist to assign the partial charges, *e.g.* in such a way that the molecular electrostatic potential computed by quantum techniques is reproduced.

Unlike the Van der Waals case, the $1/r$ behaviour of the Coulombic interaction makes it impossible to use a truncation without introducing a sizeable error. This represents a challenge for MD simulations as it suggests that the calculation of the full pairwise interatomic distance list is required at each integration step, introducing an $O(N^2)$ algorithmic scaling. To alleviate this problem, alternative approaches to efficiently compute the electrostatic contribution have been developed. Ewald sums, introduced in the 20$^{\text{th}}$ century for the study of crystals, provide a way to compute the electrostatic interactions in periodic systems by going to Fourier space (Leach 2001). This makes them suited for usage in molecular simulations, since periodic boundary conditions are very frequently used. The modern implementation of Ewald summation, Particle Mesh Ewald (PME), computes the Fourier sum on a lattice using the Fast Fourier Transform algorithm (Darden, York, and Pedersen 1993). The $O(N \ln N)$ scaling of this algorithm represents a significant improvement over the $O(N^2)$ scaling of a naive, complete pairwise calculation. Note that Ewald summation/PME require the splitting of the electrostatic interaction into a short-ranged and a long-ranged components; specifically, PME handles the long-ranged part. The short-ranged component is computed using direct pairwise summation, and is defined using a distance cut-off similarly to the Lennard-Jones case.

3.2.3. The CMAP correction

The CMAP correction, introduced in (Mackerell, Feig, and C. L. Brooks 2004), is a CHARMM-specific grid-based correction aimed at improving the accuracy in the treatment of backbone energetics in proteins. It adds so-called "cross-terms" in the potential function to match the (φ, ψ) (the protein backbone dihedral angles) energy surfaces to reference quantum calculations.

3.2.4. Modifying the potential to enhance the sampling: Accelerated Molecular Dynamics

Accelerated Molecular Dynamics (aMD), and its more recent Gaussian variant, are enhanced sampling methods developed by the team of McCammon (Hamelberg, Oliveira, and McCammon 2007; Miao, Feher, and McCammon 2015; Pierce et al. 2012). Unlike some other enhanced sampling methods which will be outlined later (chapter 4), they do not rely on collective variables and the form of the perturbing potential is somewhat arbitrary, in the sense that it is not obtained from first-principle considerations grounded in statistical mechanics. As such, we introduce them here as an example of *ad-hoc* modification to the force-field function.

The general idea is to apply a so-called "boost-potential" to stable conformations (*i.e.* atomic configurations of potential energy below a given threshold), while leaving high-energy configurations untouched. This reduces the range of accessible energy values and, mechanically, lowers the energy barriers, leading to faster conformational exploration. Thus, aMD enhances the sampling by flattening the potential energy landscape. In principle, unbiased Boltzmann-weights can be recovered since the value of the boost is known. In practice, reweighting is difficult due to the typically large absolute energy values encountered in large, solvated systems (Miao, Sinko, et al. 2014). To remedy this, Gaussian Accelerated Molecular Dynamics (GaMD) was developed. In GaMD, the form of the boost

is designed such that the cumulant expansion of the Boltzmann-factor is valid to a good approximation. This is possible using a harmonic boost, leading to a Gaussian Boltzmann-factor, hence the name Gaussian aMD.

3.2.4.1. Formalism of Accelerated MD

In aMD, a boost potential ΔV is added to the force-field V.

$$V^*(x) = V(x) + \Delta V(x) \tag{3.15}$$

Given an energy threshold E, the boost takes the form:

$$\Delta V(x) = \frac{(E - V(x))^2}{\alpha + E - V(x)} \mathbb{I}(V(x) < E) \tag{3.16}$$

$\mathbb{I}(V(x) < E)$ is the indicator function for the condition $V(x) < E$, *i.e.* this function takes value 1 if $V(x) < E$ (in which case the boost is applied) and 0 otherwise. α and E are the two adjustable parameters. α is the acceleration factor which determines the extent to which the potential energy surface is smoothed (high values of α lead to a decreased boost, and conventional MD is recovered for $\alpha = +\infty$.)

The aMD boost can be applied on the total potential energy and/or the dihedral potential energy. The most popular approach is to apply the two boosts, with different parameters (dual-boost approach). A procedure for parameter estimation is provided by the authors. For the dihedral boost, one should use:

$$E = \bar{E}_{dihedral} + 4N_{res} \text{ and } \alpha = \frac{4}{5} N_{res} \tag{3.17}$$

with N_{res} the number of amino-acid residues in the protein under study. For the total potential energy boost, one should use:

$$E = \bar{E}_{potential} + 0.16N_{atoms} \text{ and } \alpha = 0.16N_{atoms} \tag{3.18}$$

with N_{atoms} the total number of protein atoms in the system. The average potential energies $\bar{E}_{dihedral}$ and $\bar{E}_{potential}$ should be estimated from unbiased MD simulations. The procedure seems to be based on empirical considerations rather than theoretical principles, and it is unclear how one should go about adjusting the parameters if the above prescriptions yield unstable simulations (typically, helical unfolding). Also, as explained in Appendix (A.4), the large values and non-Gaussian distribution of the boost prevents re-weighting of the simulation data, and as such makes aMD ill-suited to quantitative studies. GaMD has been developed as an attempt to overcome these limitations.

3.2.4.2. Gaussian AMD

The GaMD boost takes a harmonic form:

$$\Delta V(x) = \frac{1}{2} k (E - V(x))^2 \mathbb{I}(V(x) \leq E) \tag{3.19}$$

It is clear from equation 3.19 that a boost will be applied on configurations of potential energy lower than E, which represents the threshold. The harmonic "force" constant k controls the intensity of the destabilization of the energetically favorable configurations. E and k are free parameters; in the GaMD procedure, they are determined in such a way that re-weightability up to a given energy

cut-off σ_0 (usually a small integer multiple of k_BT) is preserved. The systematic determination of E and k rests on several requirements (Miao, Feher, and McCammon 2015; Pang et al. 2017):

- The boost should preserve order: if $V(x_1) < V(x_2)$, then the boost should be such that $V^*(x_1) < V^*(x_2)$
- The boost should contract energy differences, *i.e.* ensure that $V^*(x_2) - V^*(x_1) < V(x_2) - V(x_1)$

The authors show that this implies the following condition:

$$V_{max} \leq E \leq V_{min} + \frac{1}{k} \tag{3.20}$$

where V_{max} and V_{min} are respectively the maximal and minimal possible potential energy values for the system under study. Equation 3.20 notably shows that the lower bound of the energy threshold is V_{max}. Also, it implies that k must verify:

$$k \leq \frac{1}{V_{max} - V_{min}} \tag{3.21}$$

or, with $k_0 \equiv k(V_{max} - V_{min})$, $0 < k_0 \leq 1$. Finally, to preserve re-weightability, the standard deviation of the boost $\sigma_{\Delta V}$ must be smaller than the cut-off σ_0. When all these conditions are combined, and E is set to its lower-bound V_{max} (which ensures maximum acceleration), k_0 must be set to:

$$k_0 = \min\left(1.0, \frac{\sigma_0}{\sigma_V} \cdot \frac{V_{max} - V_{min}}{V_{avg} - V_{min}}\right) \tag{3.22}$$

where σ_V and V_{avg} are respectively the standard deviation and average of V. Thus, in GaMD, the boost parameters k_0 (or k) and E are set as functions of V_{min}, V_{max}, V_{avg} and σ_V, *i.e.* statistics about the potential energy which can be estimated along an unbiased trajectory. In practice, the parameters are estimated through a "GaMD equilibration" procedure. First, a short unbiased MD simulation is performed in which no statistics are collected, so as to relax the system. Second, another unbiased MD simulation is performed, but statistics about the potential energy are collected. At the end of stage 2, a first GaMD bias is constructed from the collected data. During stage 3, this bias is applied to the system and another MD run is launched, during which new potential energy statistics are collected for the boosted dynamics. Finally in stage 4, the GaMD boost is applied and updated on-the-fly with the data collected during the simulation. At the end of stage 4, the boost is "equilibrated" and can be used without further modification for production dynamics.

3.3. Statistical mechanics of thermostatted systems: theory and algorithms

3.3.1. The canonical ensemble

Systems of interest in the laboratory are frequently exchanging energy -and possibly matter- with their surroundings. In particular, a thermostatte system is constantly exchanging heat with its thermostat. These situations are not accounted for in the microcanonical ensemble (see Appendix, A.3.1). The introduction of a so-called *canonical* ensemble is required to handle constant-temperature systems. It

turns out that the canonical ensemble can be straightforwardly derived from the microcanonical one; this derivation is given in Appendix (A.3.2). Its result is the celebrated canonical distribution:

$$\boxed{P_l = \frac{e^{-\beta E_l}}{\sum\limits_{l'} e^{-\beta E_{l'}}}} \tag{3.23}$$

where P_l is the probability for the system to be in micro-state l, and E_l is the energy of l. One also introduces the canonical partition function $Q(\beta)$ as:

$$Q(\beta) \equiv \sum_{l'} e^{-\beta E_{l'}} \tag{3.24}$$

The fundamental formula 3.23 shows that, in a thermalized system, the probability to observe a micro-state is a decreasing exponential function of its energy; low energy states will be visited more frequently. This is a central result, which explains the connection between energy and probability which lies at the heart of our understanding of molecular systems: it explains why the exploration of high-energy states are *rare events*. It is also apparent that the probability of observing a high energy state increases when temperature increases, as more energy is available from the thermostat to explore otherwise low-probability states.

3.3.2. Canonical partition function and free energy

The canonical partition function $Q(\beta)$ introduced above can be interpreted as an effective number of micro-states accessible to the system at a given temperature. Q plays the role of a generating function for thermodynamic observables, thus providing a connection between statistical mechanics and thermodynamics. A *canonical free energy* F can be introduced, which reads:

$$\boxed{F(\beta) = -k_B T \ln Q(\beta)} \tag{3.25}$$

F plays the same role and has the same information content as Q, but arguably makes the connection with thermodynamics even more intuitive. Notably, using Shannon's entropy formula applied to the canonical distribution, one can show that $F = \langle E \rangle - TS$, which is the usual thermodynamic free energy (Diu 1989).

In Appendix A.3.2.1, we explain how the canonical distribution can be translated to the case of a classical system described by a Hamiltonian $\mathcal{H}$. The classical canonical distribution is:

$$\rho(p,q) = \frac{1}{Q_{cl}} e^{-\beta \mathcal{H}(p,q)} \tag{3.26}$$

where the classical canonical partition function Q_{cl} is given by:

$$Q_{cl}(\beta) = \frac{1}{h^{3N}} \int \mathrm{d}p \mathrm{d}q e^{-\beta \mathcal{H}(p,q)} \tag{3.27}$$

Finally, we conclude this paragraph by considering the most frequent case of a separable Hamiltonian of the form $\mathcal{H}(p,q) = \sum\limits_{i=1}^{3N} \frac{p_i^2}{2m_i} + V(q)$. In this case, the canonical partition function is factorized

into separate momentum and position contributions:

$$Q_{cl} = \frac{1}{h^{3N}} \int \mathrm{d}p e^{-\beta \sum_{i=1}^{3N} \frac{p_i^2}{2m_i}} \int \mathrm{d}q e^{-\beta V(q)} \tag{3.28}$$

The position contribution, also called configurational integral, depends on the specifics of the interaction potential. It is usually denoted as Z:

$$Z(\beta) \equiv \int \mathrm{d}q e^{-\beta V(q)} \tag{3.29}$$

The momentum contribution, however, can be analytically expressed as a product of Gaussian integrals. We get:

$$\frac{1}{h^{3N}} \prod_{i=1}^{3N} \int \mathrm{d}p_i e^{-\beta \frac{p_i^2}{2m_i}} = \frac{1}{h^{3N}} \prod_{i=1}^{3N} \sqrt{\frac{2\pi m_i}{\beta}} \tag{3.30}$$

Introducing the thermal De Broglie wavelength $\Lambda_i \equiv \sqrt{\frac{\beta h^2}{2\pi m_i}}$, the momentum integral reads:

$$\frac{1}{h^{3N}} \prod_{i=1}^{3N} \int \mathrm{d}p_i e^{-\beta \frac{p_i^2}{2m_i}} = \prod_{i=1}^{N} \Lambda_i^{-3} \tag{3.31}$$

and the canonical partition is rewritten in the separated form:

$$Q_{cl} = \frac{Z}{\prod_{i=1}^{N} \Lambda_i^3} \tag{3.32}$$

This separation of contributions is important in the context of molecular simulations, because it shows that there is no strict need to sample from the momentum distribution (since its contribution is analytically known). This is notably what makes Monte-Carlo approaches viable.

3.3.3. The equipartition theorem

We now present an important result deriving from the canonical distribution, namely the **equipartition theorem**, which provides an intuitive microscopic interpretation of temperature. Considering a separable Hamiltonian, we are interested in the average kinetic energy $\langle E_c \rangle$. One has:

$$\langle E_c \rangle = \sum_{i=1}^{3N} \frac{\langle p_i^2 \rangle}{2m_i} \tag{3.33}$$

At equilibrium, each p_i is Gaussian-distributed, with zero-mean (by isotropy of space). Therefore $\langle p_i^2 \rangle$ is the variance of the associated Gaussian probability distribution, that is, $m/\beta = m_i k_B T$. A more usual way to write it is in term of the average-squared velocity, $\langle v_i^2 \rangle = \frac{k_B T}{m_i}$. Thus, the average kinetic energy satisfies:

$$\langle E_c \rangle = \frac{3N}{2} k_B T \tag{3.34}$$

Similar expressions exist for other quadratic degrees of freedom (*e.g.* the positions if the interaction

potential is harmonic). Equation 3.34 shows that each translational degree of freedom, on average, is thermalized to the extent of $\frac{1}{2}k_BT$ when the system is in equilibrium with a thermostat at temperature T. This explains why k_BT is taken as the typical order of magnitude for the energy of thermal agitation.

3.3.4. Achieving canonical sampling in simulations

The Verlet integration algorithm presented in section 3.1 allows for the simulation of Hamiltonian dynamics, which is energy preserving. This means that Hamiltonian simulations sample from the microcanonical distribution (NVE ensemble). Sampling from the NVT ensemble, that is, from the canonical distribution, requires appropriate modifications of the integration algorithm, notably so as to keep the temperature constant. These modifications should model the influence of the bath. Several schemes to that effect have been proposed over the years. Note, however, that merely maintaining the temperature constant is *not* sufficient to sample from the canonical distribution. Rather, one should justify that the stationary phase-space (or configuration-space, since atomic velocities are arguably less interesting) probability density under the modified dynamics is actually of the form prescribed by equation 3.23.

3.3.4.1. Langevin dynamics

Langevin dynamics is a stochastic thermostatting method which relies on the Langevin description of Brownian motion. It is the primary approach used in NAMD (Phillips et al. 2005) and as such, is used for virtually all thermostatted simulations reported in this thesis. In Langevin dynamics, the Newtonian equations of motion (which describe the constant-energy situation) are modified by the addition of a velocity-dependent friction and a random force. The new dynamics for a cartesian coordinate x_i reads:

$$m_i\frac{\mathrm{d}^2x_i}{\mathrm{d}t^2} = -\nabla_{x_i}U - \gamma_i\frac{\mathrm{d}x}{\mathrm{d}t} + L(t) \tag{3.35}$$

where m_i is the mass of the atom whose x_i is a coordinate (hereafter called "atom i" to simplify), γ_i is a friction coefficient applied to atom i, and $L(t)$ is a Langevin random force. $L(t)$ must satisfy certain statistical properties for equation 3.35 to effectively act as a thermostat to the temperature T. We take $L(t)$ as a Gaussian white noise, *i.e.* :

$$\begin{aligned}\langle L(t)\rangle &= 0\\ \langle L(t+\tau)L(t)\rangle &= C\delta(\tau)\end{aligned} \tag{3.36}$$

where $\langle ...\rangle$ refers to the time average and C is a positive constant. To the stochastic differential equation on the trajectory (equation 3.35) can be associated a (deterministic) partial differential equation on the probability $\rho(v,x,t)$ to observe the velocity v and the position x at time t under the dynamics 3.35 (Zwanzig 2001) (the index has been dropped for simplicity). This partial differential equation is the forward Kolmogorov equation, or Fokker-Planck equation. In this case it reads (Zwanzig 2001):

$$\frac{\partial\rho}{\partial t} = -\frac{\partial}{\partial x}[v\rho] + \frac{1}{m}\frac{\partial}{\partial v}\left[(\nabla U + \gamma v)\rho\right] + \frac{C}{2m^2}\frac{\partial^2\rho}{\partial v^2} \tag{3.37}$$

At equilibrium, $\frac{\partial\rho}{\partial t} = 0$, so we are looking for stationary solutions of 3.37. It can be shown that the stationary solution ρ_{eq} is canonical if $C = 2\gamma k_BT$ (fluctuation-dissipation theorem), which is

the requirement on the friction and random force for equation 3.35 to sample from the canonical distribution.

Note that the philosophy here is a bit different from the original Langevin approach. Unlike the Langevin theory of Brownian motion, one cannot assume that the thermostatted particle is much larger in size than the solvent molecules to justify such a coarse-grained description - in MD simulations, each individual atom will be thermostatted through equation 3.35. Rather, equation 3.35 should be understood in a more abstract sense, as a device to couple the system to the bath. Consistently, γ_i should be set to the smallest value which ensures accurate coupling to the thermostat. In this case, the Langevin terms can be seen as small perturbations of the Hamiltonian dynamics, and thus it is reasonable to assume that trajectories retain their physical significance (which is less clear with other thermostatting methods). Other advantages of Langevin dynamics are:

- The friction term will dampen large variations of velocity, which improves numerical stability of the equation of motion (Phillips et al. 2005)
- The use of a random force adds some noise to the simulation, which arguably should assist in having an effective sampling of the configurational space.

The BBK algorithm introduced in the previous section is used for the numerical integration of Langevin dynamics.

3.3.4.2. Velocity-rescaling

Another popular approach to thermostatting is velocity-rescaling, in which temperature control is achieved through a modification of the atomic velocities. Indeed, the temperature T ands average squared velocities are connected (at equilibrium) through the equipartition theorem:

$$\frac{1}{2}m\langle v^2\rangle = \frac{3}{2}k_B T \tag{3.38}$$

For a given atomic velocity $v(t)$, a "kinetic temperature" $\Theta(t)$ can be defined as $\Theta(t) = \frac{mv(t)^2}{3k_B}$. A thermostat should notably ensure that the average kinetic temperature $\bar{\Theta} = T$ where T is the target temperature and that the fluctuations of Θ are consistent with what is expected from the canonical ensemble.

It is clear that if the kinetic temperature is $\Theta \neq T$, one can reach T by rescaling each atomic velocity v_i by $\alpha_i \equiv \sqrt{T/\Theta}$. In a naive implementation, this would be done after each propagation step of the dynamics. This of course would lead to abrupt velocity changes and an unstable simulation. Instead, the popular coupling algorithm of Berendsen (Berendsen et al. 1984) uses an exponential relaxation with a time constant τ to ensure more gentle coupling.

Despite its popularity, Berendsen's thermostat does not sample from the canonical distribution, notably because it does not reproduce the correct energy fluctuations. This was reported to lead to spectacular violations of equipartition, or "flying ice-cube", in which all the kinetic energy got concentrated into one translational (or rotational) degree of freedom (Harvey, Tan, and Cheatham 1998). A true canonical velocity-rescaling algorithm was proposed in 2008 by Bussi and co-workers (Bussi, Donadio, and Parrinello 2007). This thermostat introduces a stochastic term to make sure to sample from the theoretical, equilibrium distribution of the kinetic energy.

3.3.4.3. Velocity-reassignment

Another very popular approach to thermostatting is the one of Andersen, also known as velocity-reassignment (Andersen 1980). In this scheme, the effect of the bath is modelled by stochastic "collisions" , *i.e.* random events in which one (or several) particles of the system are randomly selected and have their velocities re-assigned by drawing them from a Maxwell-Boltzmann distribution at the target temperature (Andersen 1980; Frenkel and Smit 2002). It can be shown that this procedure indeed generates a canonical distribution. The strength of the coupling to the thermostat is set by a parameter ν, which represents the probability per unit of time for a collision to happen.

3.3.4.4. Nosé-Hoover extended dynamics

Note that other families of numerical thermostats exist. In particular, in *extended degrees of freedom* techniques, additional (fictitious) degrees of freedom are added to the system so as to model the effect of the bath. They are endowed with specific equations of motions designed in such a way that the atomic degrees of freedom sample from the canonical distribution. The most-popular approach of this class is the Nosé-Hoover thermostat (Hoover 1985; Nosé 1984). These methods are reviewed in (Tuckerman 2010) to which the interested reader is referred.

3.3.5. Pressure control

The canonical ensemble corresponds to the constant temperature, constant volume situation. To be even closer to laboratory conditions, one should instead work at fixed pressure, *i.e.* in the so-called isothermal-isobaric ensemble (NPT ensemble). Similarly to thermostats, computational *barostats* have been developed to ensure constant pressure in MD simulations. We will not discuss these methods; the reader is instead referred to (Frenkel and Smit 2002; Tuckerman 2010, for review).

4. Free energy calculations: an overview

4.1. Why is it interesting, why is it challenging?

Molecular simulations emerged as computational tools for the estimation of thermodynamic averages in classical interacting systems of high dimension, for which analytical calculations quickly show their limits. The original publication of the Metropolis algorithm (Metropolis et al. 1953) and the first reported Molecular Dynamics simulations (*e.g.* Alder and Wainwright 1957, 1959) deal with such problems. Both the Metropolis and constant-temperature MD methods can be seen as (primitive) enhanced sampling techniques, because they are designed to generate configurations drawn in the Boltzmann distribution, rather than uniformly. However, it became apparent that this was often insufficient to ensure convergence of thermodynamic estimates from simulations. This is due to to the metastable character of the dynamics in thermalized systems: the trajectory tends to explore the vicinity of local potential energy minima for a long time (which leads to apparent convergence), before jumping in a stochastic manner to a different basin. Transitions are *rare events*. In situations where the proper convergence of thermodynamic quantities requires the exploration of all the basins, waiting for the stochastic transitions to happen is impractical.

For example, one may be interested in the free energy difference between two conformations A and B of a protein. Here A and B refer to local energetic minima or basins, within which the protein fluctuates. Given operational definitions of A and B (for example using an order parameter), a naive scheme would be to run a constant-temperature MD simulation of the protein and measure the times τ_A and τ_B that the protein spends in each basin. Time-averaged occupancy probabilities are then readily computed as $P_A = \tau_A/(\tau_A + \tau_B)$ and similarly for τ_B.

Finally, assuming ergodicity of the simulation, one can equate time-averaged probabilities with ensemble probabilities and compute the free energy difference as:

$$\Delta F_{A\to B} = -k_B T \ln \frac{P_B}{P_A} = -k_B T \ln \frac{\tau_B}{\tau_A} \tag{4.1}$$

However, consider the situation where the simulation is so short that the system, initially in A, never leaves this basin: it is impossible to obtain an estimate of ΔF. Similarly, if only one crossing event is captured, the estimated occupancy probabilities may be very different from their equilibrium values. Typically, a reasonable number of recrossing events should be observed (the definition of reasonable being system-dependent) for this approach to be considered; and this is generally possible only with very simple systems, such as alanine dipeptide. This problem is referred to by some authors as *quasi non-ergodicity* of the simulated dynamics (Comer, Gumbart, et al. 2015). Indeed, as the simulation is of finite time, the sampling is not full (*i.e.* not all regions of configurational space are visited according to their Boltzmann weight). This leads to an apparent breaking of ergodicity, in spite of the dynamics used for sampling being ergodic in the mathematical sense (*i.e.* ergodicity would hold for an infinitely long trajectory - see also Appendix, A.3.3).

Consequently, a wide variety of numerical methods have been proposed to overcome the limitations of equilibrium sampling. These methods generally combine an enhanced sampling strategy (whose

purpose is to allow for the exploration of low-probability/high-energy configurations) with a numerical procedure to estimate the sought-after free energy difference. In many cases, they involve the parallel simulation of several replicas of the system, possibly communicating (*e.g.* through exchange of temperature).

In the following we review some of the most popular methods, with a special emphasis on geometrical free energy calculations, *i.e.* methods to compute the free energy along a given geometrical observable (or collective variable, CV). Overall, these methods rely on applying a specific bias on the CV of interest to enhance the sampling. If the bias is well-designed, it will attenuate or suppress the metastable character of the dynamics along this particular CV - however, this is not true for the so-called *orthogonal degrees of freedom*. Orthogonal degrees of freedom refer to independent collective variables which are also metastable, and remain so since no explicit bias is applied onto them. Metastability along orthogonal degrees of freedom can impair convergence of a free energy calculation and it may be very difficult to identify the faulty degree(s) of freedom. They represent one of the biggest challenge for geometrical free energy calculations.

Another challenge lies in the interpretation of the biased trajectories. It is largely accepted that trajectories extracted from conventional MD simulations can be used to infer mechanistic details, because it is considered that the physics of the simulation matches reasonably well that of the real molecular system. However, if an external bias is added to enhance the sampling, the physics is changed and it becomes difficult to infer mechanistic details from the trajectory. In particular, this problem arises when one wishes to obtain kinetic information from the biased simulation. We will review several approaches to extract kinetic information from biased simulations. Finally, the rare events -stochastic transitions- are frequently of interest in and of themselves: most relevant for this thesis, conformational transitions of proteins are at the heart of their functional mechanism (see Chapter 1). However, the interest for rare events is by no means limited to molecular biophysics; other important examples of rare events include nucleation processes, allelic fixation in population genetics, chemical reactions observed at the molecular level, financial crises, climatic transitions... Even if it is not necessarily possible to establish a unifying formalism for all these examples, a common feature is the importance of noise (temperature in molecular systems, genetic drift in population genetics...) in triggering the rare events in a stochastic manner. Coming back to molecular systems, several methods for the determination of the optimal (in a sense to be defined) transition pathway(s) have been developed, and some will also be presented at the end of this chapter.

4.2. Alchemical free energy calculations

Free energy calculations are termed alchemical when they are aimed at evaluating the free energy difference between systems described by different (although often close) Hamiltonians. For example, one may seek the difference of solvation free energy between CH_3F and CH_3Cl. The transformation of F to Cl is akin to a transmutation in the alchemical sense, hence the name for this category of calculations. Of course, this transformation is not permitted by a classical force field and the two molecules are formally described by two different Hamiltonians $\mathcal{H}_0$ and $\mathcal{H}_1$. The sought after free energy difference is then:

$$\Delta F_{0\to 1} = -k_B T \ln \frac{Q_1}{Q_0} \tag{4.2}$$

Possibly the most widespread use of alchemical calculations in nowadays applications is the computation of binding affinities, in which the interaction terms between the ligand and the receptor are

alchemically turned off. In the present work, we are not concerned with such problems, and so we will keep the description of alchemical calculations to a minimum. Excellent reviews on these questions are available for a more detailed description (*e.g.* Montalvo-Acosta and Cecchini 2016). We nonetheless made the choice to expose them, first because of their importance in the field, second because they provide the opportunity to introduce some concepts (such as thermodynamic integration), which are also important for geometrical calculations.

4.2.1. Transformation of the Hamiltonian

In the study of alchemical transformations, it is customary to introduce an adimensional parameter λ which defines a family of Hamiltonians $\mathcal{H}_\lambda$ such that:

$$\mathcal{H}_\lambda = (1-\lambda)\mathcal{H}_0 + \lambda\mathcal{H}_1 \tag{4.3}$$

$\lambda \in [0,1]$ represents a "progress variable" parametrizing the alchemical transformation from $\mathcal{H}_0$ to $\mathcal{H}_1$. λ does not represent a measurable quantity of the system, rather it is an auxiliary variable whose introduction will prove useful later on. As such, its choice is not unique and one may replace the λ prefactors in equation 4.3 by any increasing function $g(\lambda) \in [0,1]$ if such a choice were more convenient for the problem at hand. For simplicity, we will stick to a linearly increasing λ on $[0,1]$ except if stated otherwise.

4.2.2. Free energy perturbation

Let us consider the (classical) partition function Q_1 of $\mathcal{H}_1$. It writes:

$$Q_1 = \int \mathrm{d}p\mathrm{d}x e^{-\beta\mathcal{H}_1} \tag{4.4}$$

where the pre-factor involving h has been omitted. Introducing $1 = e^{-\beta\mathcal{H}_0}e^{+\beta\mathcal{H}_0}$ in the integral of equation 4.4 yields:

$$Q_1 = \int \mathrm{d}p\mathrm{d}x e^{-\beta\mathcal{H}_0}e^{+\beta(\mathcal{H}_0-\mathcal{H}_1)} \tag{4.5}$$

which is:

$$Q_1 = \langle e^{+\beta(\mathcal{H}_0-\mathcal{H}_1)}\rangle_0 Q_0 \tag{4.6}$$

where $\langle ...\rangle_0$ is the average with respect to (the canonical distribution generated by) $\mathcal{H}_0$. Rearranging 4.6 and plugging in 4.2 yields the so-called Free Energy Perturbation (FEP) formula:

$$\boxed{e^{-\beta\Delta F_{0\to1}} = \langle e^{-\beta\Delta\mathcal{H}_{0\to1}}\rangle_0} \tag{4.7}$$

where $\Delta\mathcal{H}_{0\to1} = \mathcal{H}_1 - \mathcal{H}_0$. The name "free energy perturbation" originates in the early use of this formula as a starting point for a perturbative expansion:

$$\langle e^{-\beta\Delta\mathcal{H}}\rangle_0 = \sum_{k=0}^{+\infty}\frac{(-\beta)^k}{k!}\langle\Delta\mathcal{H}^k\rangle_0 \tag{4.8}$$

If the perturbation $\Delta\mathcal{H}$ to $\mathcal{H}_0$ is weak, expansion 4.8 may be truncated at a finite order, which may allow an approximate analytical calculation of the free energy difference. To the best of our knowledge, the earliest use of this approach (and the first reported derivation of equation 4.7) is due to Zwanzig (1954). Zwanzig used the perturbation technique to study the equation of state of a Lennard-Jones gas, using a hard-sphere gas as the reference, unperturbed system.

It ought to be noted that modern applications generally use formula 4.7 directly; as such, it could be more appropriate to refer to this free energy calculation framework as the **exponential formula** formalism.

4.2.3. Thermodynamic integration

Thermodynamic integration (TI) refers to all free energy calculations in which the free energy is computed as the integral of its derivative. The method was initially introduced by Kirkwood (1935) as a route to the calculation of the chemical potential in interacting gases. Omitting momenta for simplicity (which is not always possible, *e.g.* if the alchemical transformation changes atomic masses), we consider the family of λ-parametrized potential energy functions $U_\lambda = U + \lambda V$.

We introduce the λ-dependent (configurational) partition function:

$$Z(\lambda) = \int \mathrm{d}x e^{-\beta(U(x)+\lambda V(x))} \tag{4.9}$$

In the alchemical setting, one simply writes:

$$\Delta F = \int_0^1 \frac{\mathrm{d}F}{\mathrm{d}\lambda}\mathrm{d}\lambda \tag{4.10}$$

Actually using equation 4.10 requires knowing the free energy derivative. We now proceed to establish its expression.

$$-\beta\frac{\mathrm{d}F}{\mathrm{d}\lambda} = \frac{\mathrm{d}}{\mathrm{d}\lambda}\ln Z(\lambda) = \frac{1}{Z(\lambda)}\frac{\mathrm{d}Z}{\mathrm{d}\lambda} \tag{4.11}$$

$$\frac{\mathrm{d}Z}{\mathrm{d}\lambda} = \frac{\mathrm{d}}{\mathrm{d}\lambda}\int \mathrm{d}x e^{-\beta(U(x)+\lambda V(x))} \tag{4.12}$$

$$\frac{\mathrm{d}Z}{\mathrm{d}\lambda} = -\beta\int \mathrm{d}x V(x) e^{-\beta(U(x)+\lambda V(x))} \tag{4.13}$$

Thus:

$$\frac{1}{Z(\lambda)}\frac{\mathrm{d}Z}{\mathrm{d}\lambda} = -\beta\frac{1}{Z(\lambda)}\int \mathrm{d}x V(x) e^{-\beta(U(x)+\lambda V(x))} = -\beta\langle V(x)\rangle_\lambda \tag{4.14}$$

And finally:

$$\boxed{\frac{\mathrm{d}F}{\mathrm{d}\lambda} = \langle V(x)\rangle_\lambda} \tag{4.15}$$

or equivalently:

$$\boxed{\frac{\mathrm{d}F}{\mathrm{d}\lambda} = \langle\frac{\partial U_\lambda}{\partial\lambda}\rangle_\lambda} \tag{4.16}$$

Thus, in the alchemical setting, the derivative of the free energy with respect to the control parameter takes a very simple -and elegant- expression. We will see that the corresponding expression is more complicated when the free energy is differentiated with respect to a function of the internal degrees of freedom of the system.

4.2.3.1. An application of thermodynamic integration: the confinement method for absolute chemical potential

The confinement method is a recent approach for the computation of "absolute" chemical potential, as opposed to above methods that aim to estimate free energy differences. The idea of the confinement stems from the fact that the analytical expression of the partition function of the harmonic oscillator is known: the free energy can be computed exactly for harmonic systems. The confinement method uses high force constant harmonic restraints to transform the energy landscape into a harmonic one in the vicinity of the basin under study (Cecchini, S. V. Krivov, et al. 2009; Esque and Cecchini 2015; Ovchinnikov, Cecchini, and Karplus 2013; Tyka, Clarke, and Sessions 2006). The reversible work done during this "confinement" procedure is evaluated by thermodynamic integration. The absolute chemical potential of the "harmonicized" state is then computed analytically by normal mode analysis performed in the highly restrained potential. Although it is not explicitly formulated as such, we argue that the transformation of the Hamiltonian to the strongly harmonic state is akin to an alchemical transformation, which is why we categorize confinement as an alchemical method.

4.3. Geometrical free energy calculations

Geometrical free energy calculations refer to the family of methods aimed at computing the free energy profile, or Potential of Mean Force (PMF) along a given collective variable (possibly high dimensional). Unlike the alchemical case, we are not interested in computing the free energy difference between two systems with two different Hamiltonians, but rather in mapping the (relative) free energy of different phase space regions for a given Hamiltonian by projecting them upon a lower dimension Collective Variable (CV). These methods are of critical importance for the computational study of conformational transitions, and as such are widely used throughout the present thesis. For this reason, we take a particular care in exposing the most popular methods and their underlying theory, even though we do not aim for a comprehensive list. The derivations of several important relations are given in Appendix, A.5.

4.3.1. Collective variables and potential of mean force

For an atomic configuration described by its $3N$-dimensional cartesian coordinates vector x, we define a collective variable (CV, also colvar or observable) as a function $\hat{\xi}$ such that $\hat{\xi} : \mathbb{R}^{3N} \to \mathbb{R}^n, x \mapsto \hat{\xi}(x)$, where n is an integer, small (generally 1 or 2) in practical cases. Collective variables typically represent geometrical measurements on the system under study, such as an inter-atomic distance, the Root-Mean-Square Deviation (RMSD) with respect to a particular conformation, a gyration radius...

The Potential of Mean Force (PMF) $\Delta F(\xi)$ is an important quantity associated with a collective variable. It corresponds to the free energy of the system where all degrees of freedom except the Collective Variable (CV) are allowed to equilibrate, and the CV is fixed at a given value ξ. In mathematical terms, we introduce the restricted (configurational) partition function $Z(\xi)$ such that:

$$Z(\xi) \equiv \int e^{-\beta U(x)} \delta(\hat{\xi}(x) - \xi) \mathrm{d}x \tag{4.17}$$

where $U(x)$, as previously, is the potential energy function of the system. It is clear that, with Z the full configurational partition function, one has:

$$P(\xi) = \langle \delta(\hat{\xi}(x) - \xi) \rangle = \frac{Z(\xi)}{Z} \tag{4.18}$$

where $P(\xi)$ is the equilibrium (canonical) probability density of $\hat{\xi}(x)$. Then, with ξ_0 the value of maximal probability, it is customary to define the PMF as:

$$\Delta F(\xi) \equiv -k_B T \ln \frac{P(\xi)}{P(\xi_0)} = -k_B T \ln \frac{Z(\xi)}{Z(\xi_0)} \tag{4.19}$$

or, without the reference level, as:

$$F(\xi) \equiv -k_B T \ln Z(\xi) \tag{4.20}$$

All these alternative definitions differ only by an additive constant, which drops when taking free energy differences - which are the only relevant quantities.

Physical interpretations of the potential of mean force

From the definition of equation 4.18, it is clear that the PMF provides the same information as the equilibrium probability density. As such, it virtually contains all the information about the equilibrium properties of the collective variable: if one knows the PMF, one can compute the average value, the standard deviation, the occupancy ratio between two values of $\hat{\xi}(x)$, etc. In this picture, the PMF does not provide knowledge about the dynamical properties of the system, notably its kinetics. To make the connection with dynamical properties, it is customary to recognize the PMF as the effective potential for the evolution of the CV treated as a dynamical variable. Most often, this dynamics is assumed to take the form of an overdamped Langevin equation:

$$\gamma_\xi \dot{\xi}(t) = -\frac{\mathrm{d}F(\xi)}{\mathrm{d}\xi} + L(t) \tag{4.21}$$

where γ_ξ is an effective friction coefficient, $F(\xi) = -k_B T \ln Z(\xi) = \Delta F(\xi) + cst$ and $L(t)$ is a random Langevin force of zero average and which obeys a fluctuation-dissipation relation, $\langle L(t) L(t+\tau) \rangle = 2\gamma_\xi k_B T \delta(\tau)$. This assumption is justified much like in the same way as the original Langevin approach, except that the "solvent" (*i.e.* the degrees of freedom treated effectively by the friction and random forces) corresponds to the orthogonal degrees of freedom to $\hat{\xi}$. More general (non-Markovian) types of stochastic dynamics may also be used, *i.e.* the generalized Langevin equation (Zwanzig 2001).

As such, the determination of the PMF along a collective variable is usually the first step towards understanding its effective dynamics, which can yield precious insight into the behaviour of the full system if the collective variable is well-chosen. We now review numerical schemes to access the PMF.

4.3.2. Steered Molecular Dynamics and non-equilibrium methods

Steered Molecular Dynamics (SMD) is a popular method in which a moving harmonic bias is applied on a collective variable during the simulation, typically to drive a conformational change in a protein. Many Steered Molecular Dynamics (SMD) simulations on a variety of collective variables are reported throughout this thesis.

Early SMD simulations were introduced as a computational counterpart to single-molecule experiments, such as force-spectroscopy. For instance, SMD was used to probe unbinding between molecular partners (Grubmüller, Heymann, and Tavan 1996; Izrailev et al. 1997) or the force-induced unfolding of titin (Paci and Karplus 1999).

Moreover, Targeted MD (TMD), introduced around the same time, can be seen as a special case of SMD in which the biased collective variable is the RMSD with respect to a target structure (Schlitter, Engels, and Krüger 1994; Schlitter, Engels, Krüger, et al. 1993). TMD was instrumental for early investigations of large-scale conformational transitions in proteins, as it allowed to probe these otherwise rare events while using finite-temperature sampling. However it became apparent that the transition pathways thus generated were generally unreliable, because the largest changes tended to happen first. A modified, restricted-TMD approach was published to remedy this problem (Vaart and Karplus 2005).

In its modern implementation, SMD uses a harmonic potential (unlike a constraint like in initial TMD), whose center is moved with a constant velocity v between user-defined values $\xi_{initial}$ and ξ_{final} of the collective variable.

$$V_{SMD}(x,t) = \frac{1}{2}k\left(\hat{\xi}(x) - \xi_{initial} - vt\right)^2 \tag{4.22}$$

v is fixed such that $\xi_{initial} + vt_{total} = \xi_{final}$, where t_{total} is the total duration of pulling. During an SMD simulation, the accumulated work performed by the moving restraint on the system can be collected using the "time-integral of velocity $\times$ force" formula:

$$W(\xi) = -k\int_0^t v \cdot (\hat{\xi}(x(t')) - \xi(t'))\mathrm{d}t' \tag{4.23}$$

where $\xi(t)$ refers to the position of the moving center at time t, *i.e.* $\xi(t) = \xi_{initial} + vt$.

$W(\xi)$ provides an estimate of the PMF along ξ (between $\xi_{initial}$ and ξ_{final}). However, since the pulling occurs at finite, non-zero velocity, this estimate is generally poor because the system is not in equilibrium with the restraint all along the simulation. The second law of thermodynamics, applied to constant-temperature transformations, establishes that $\Delta F(\xi) \leq \langle W(\xi)\rangle$ (taking the starting value of the collective variable in the SMD protocol as the reference level, and where the average is taken over an ensemble of independent SMD simulations). As such, the average non-equilibrium work profile collected during a series of SMD simulations can be used as an upper-bound estimate of the free energy difference, especially if the pulling is done reasonably slowly.

Theoretical advances in non-equilibrium statistical mechanics, in particular Jarzynski's theorem, show that the exact free energy difference can be computed by exponential averaging of the work profiles obtained from several independent SMD simulations (Chipot and Pohorille 2007; Jarzynski 1997a,b):

$$e^{-\beta\Delta F} = \langle e^{-\beta W}\rangle \tag{4.24}$$

In equation 4.24, the average is taken over a collection of independent, finite-time pulling simula-

tions, initiated from a system in thermal equilibrium, and with the same prescribed path for the pulling bias. Equation 4.24 is of considerable fundamental significance, as it shows that an equilibrium quantity (the free energy difference) can be obtained by averaging over non-equilibrium trajectories. As pointed out by Hummer and Szabo (Hummer and Szabo 2005), Jarzynski's theorem covers the intermediate situation between infinitely slow pulling (in which case thermodynamic integration can be used to recover the free energy difference) and infinitely fast pulling (in which case the pulling is seen as an instantaneous change in the Hamiltonian and the free energy difference is estimated by free energy perturbation). We note that some modifications of equation 4.24 are required to use it for PMF estimation; the reader is referred to (Chipot and Pohorille 2007, Chapter 5) for details. However, experience shows that this approach is very slow to converge. This is due to the exponentially rare occurrence of negative work trajectories, which still contribute significantly to the exponential average.

4.3.3. Umbrella sampling

Umbrella sampling is one of the oldest methods for free energy calculations, and the form it takes has evolved over time. The initial formulation is due to Torrie and Valleau (1977). We focus here on its modern and most widely used sense, that is, stratified harmonically restrained sampling. The idea is to divide the collective variable space into so-called "windows" in which the collective variable of interest is harmonically restrained to a given value, biasing the sampling in its vicinity. This notably allows the sampling of high free energy regions. The use of windows ensures (or should so) full coverage of the collective variable space of interest. Recovering the full PMF requires an unbiasing procedure in which, additionally, the data obtained from all the windows are combined to provide a single estimate of the free energy profile. Several such procedures have been proposed. We now review a very popular one, called Weighted Histogram Analysis Method (WHAM), which arrives at an unbiased estimate of the free energy profile by combining the biased histograms from the parallel windows (Kumar et al. 1992).

4.3.3.1. Unbiasing probability distributions

For a system described by an unbiased potential $U(x)$ and a collective variable $\hat{\xi}$ for which we want to estimate a PMF, we devise a set of M independent simulations (usually termed "windows") in which a biasing potential is added. The biasing potential depends on x only through $\hat{\xi}(x)$ and harmonically restrains $\hat{\xi}(x)$ to a given value $\bar{\xi}_i$ for window i.

$$V_i\left(\hat{\xi}(x)\right) = \frac{1}{2}k_i\left(\hat{\xi}(x) - \bar{\xi}_i\right)^2 \tag{4.25}$$

$\widetilde{P}_i(\xi)$ is the *biased* probability density of ξ, *i.e.* the distribution of ξ in the canonical ensemble generated by the biased potential $U(x) + V_i\left(\hat{\xi}(x)\right)$. Our goal is to "correct" $\widetilde{P}_i(\xi)$ to recover the unbiased, canonical distribution $P(\xi)$.

Let $\widetilde{Z}_i$ be the configurational partition function for the biased potential of window i:

$$\widetilde{Z}_i \equiv \int \mathrm{d}x e^{-\beta\left[U(x)+V_i\left(\hat{\xi}(x)\right)\right]} \tag{4.26}$$

and $\widetilde{Z}_i(\xi)$ the restricted configurational biased partition function for a given value of the collective variable:

$$\widetilde{Z}_i(\xi) \equiv \int \mathrm{d}x e^{-\beta\left[U(x)+V_i\left(\hat{\xi}(x)\right)\right]} \delta\left(\hat{\xi}(x)-\xi\right) \tag{4.27}$$

The biased probability $\widetilde{P}_i(\xi)$ can be expressed as the ratio of these two partition functions $\widetilde{P}_i(\xi) = \widetilde{Z}_i(\xi)/\widetilde{Z}_i$. Also, let Z be the configurational partition function with respect to the unbiased potential $U(x)$:

$$Z \equiv \int \mathrm{d}x e^{-\beta U(x)} \tag{4.28}$$

We know that $P(\xi)$ reads:

$$P(\xi) = \frac{1}{Z} \int \mathrm{d}x e^{-\beta U(x)} \delta\left(\hat{\xi}(x)-\xi\right) \tag{4.29}$$

By insertion, we see that:

$$P(\xi) = \frac{\widetilde{Z}_i}{Z} \frac{1}{\widetilde{Z}_i} \int \mathrm{d}x e^{-\beta\left[U(x)+V_i\left(\hat{\xi}(x)\right)\right]} e^{+\beta V_i\left(\hat{\xi}(x)\right)} \delta\left(\hat{\xi}(x)-\xi\right) \tag{4.30}$$

which can be rewritten as:

$$P(\xi) = e^{-\beta \Delta F_i} \langle e^{+\beta V_i\left(\hat{\xi}(x)\right)} \delta\left(\hat{\xi}(x)-\xi\right) \rangle_i \tag{4.31}$$

where the free energy difference ΔF_i has been defined as $\Delta F_i \equiv -k_B T \ln\left(\widetilde{Z}_i/Z\right)$ and $\langle ... \rangle_i$ is the canonical average with respect to the biased potential in window i. Using a conditional average, one can write:

$$\langle e^{+\beta V_i\left(\hat{\xi}(x)\right)} \delta\left(\hat{\xi}(x)-\xi\right) \rangle_i = \langle e^{+\beta V_i\left(\hat{\xi}(x)\right)} | \hat{\xi}(x) = \xi \rangle_i \langle \delta\left(\hat{\xi}(x)-\xi\right) \rangle_i \tag{4.32}$$

where $\langle ... | \hat{\xi}(x) = \xi \rangle_i$ is the conditional, biased canonical average in window i, knowing that $\hat{\xi}(x) = \xi$.

By definition $\langle \delta\left(\hat{\xi}(x)-\xi\right) \rangle_i = \widetilde{P}_i(\xi)$. Furthermore, since V_i depends on x only through $\hat{\xi}(x)$, the conditional average $\langle e^{+\beta V_i\left(\hat{\xi}(x)\right)} | \hat{\xi}(x) = \xi \rangle_i$ can directly be written as $e^{+\beta V_i(\xi)}$. Gathering everything, one can express the unbiased probability as a function of the biased one:

$$P(\xi) = e^{-\beta \Delta F_i} e^{+\beta V_i(\xi)} \widetilde{P}_i(\xi) \tag{4.33}$$

Equation 4.33 connects the unbiased and biased probability distributions and forms the basic tool to recover an unbiased PMF from a statically biased simulation of known bias. We note that this analysis could not be used if the bias were a function of the history of the system, as is the case in adaptive methods (see 4.3.4.3 and 4.3.5).

4.3.3.2. PMF reconstruction by the Weighted Histogram Analysis Method (WHAM)

Although exact, equation 4.33 generally cannot be used directly since values of ξ that are very different from the center of the biasing potential will be poorly sampled in an actual simulation, leading to inaccurate estimation of their free energy. This is the reason why several windows are run, such that together they cover the full range of interesting values for ξ. In this case, equation 4.33 can be applied for each window j, yielding the unbiased probability distributions $P_j(\xi)$. Let us remark that in the case of perfect sampling, $\forall j,\ \forall \xi, P_j(\xi) = P(\xi)$.

Obviously this will not be the case in practical scenarios. Instead, one may aim to reconstruct the full unbiased probability $P(\xi)$ by combining the unbiased distributions obtained from the windows. This may be achieved using ξ-dependent coefficients $c_j(\xi)$ to build up a linear combination:

$$P(\xi) = \sum_{j=1}^{M} c_j(\xi) P_j(\xi) \tag{4.34}$$

$$P(\xi) = \sum_{j} c_j(\xi) e^{-\beta \Delta F_j} e^{+\beta V_j(\xi)} \widetilde{P}_j(\xi) \tag{4.35}$$

Equation 4.35 makes it clear that the calculation of the unbiased probability distribution from the biased simulations requires the knowledge of the coefficients c_j along with the **free energy offset** values ΔF_j. The latter may be seen as the reversible work required to place the system in the biasing potential $U + V_j$. The coefficients c_j shall also be supplemented with a normalization condition, $\sum c_j(\xi) = 1$.

WHAM is an iterative algorithm for the self-consistent determination of the c_j and ΔF_j unknown quantities. For that purpose, two coupled equations are derived using an error minimization argument (Tuckerman 2010):

$$P(\xi) = \frac{\sum_{j=1}^{M} N_j P_j(\xi)}{\sum_{j=1}^{M} N_j e^{\beta \Delta F_j} e^{-\beta V_j(\xi)}} \tag{4.36}$$

$$e^{-\beta \Delta F_i} = \int \mathrm{d}\xi P(\xi) e^{-\beta V_i(\xi)} \tag{4.37}$$

where N_j is the number of frames in the trajectory of window j. Starting from a guess of $P(\xi)$ (*i.e.* the c_j) and the ΔF_j, these equations are iterated in turn until a self-consistent solution to both is found. At this stage, the unbiased probability density $P(\xi)$ is obtained.

Another post-processing method for umbrella sampling, the Multistate Bennett Acceptance Ratio, has been more recently proposed; the reader is referred to the corresponding publications (Shirts and Chodera 2008). Finally, in section 4.3.4.5, we outline yet another approach, called Umbrella Integration (UI), which combines Umbrella Sampling with Thermodynamic Integration.

4.3.4. Gradient-based approaches

A priori, there is no fundamental obstacle to using thermodynamic integration to compute a PMF, writing $F(\xi_2) - F(\xi_1) = \int_{\xi_1}^{\xi_2} F'(\xi) \mathrm{d}\xi$. However, the expression for the free energy derivative with respect to ξ is more complicated than in the alchemical case. In the following, we give its expression for the case of a one-dimensional CV and review some of the numerical techniques to exploit it as a route to the potential of mean force. These techniques are referred to as *gradient-based approaches*. Once the gradient profile is obtained, numerical procedures for integration can be used to obtain the PMF, such as Simpson's method. These methods will not be reviewed here; for instance, the reader may instead refer to (Press 2007, Chapter 4).

4.3.4.1. Generalized thermodynamic force along a collective variable

By a direct analogy with classical mechanics, where the (conservative) force f_x applied on a cartesian coordinate x derives from the potential energy U as $f_x = -\partial_x U$, the quantity $f_\xi = -F'(\xi)$ represents a generalized thermodynamic force acting on ξ treated (at least formally) as a dynamical degree of freedom. Generalized, because $\xi = \hat{\xi}(x)$ can be seen as a generalized coordinate in the sense of analytical mechanics (see Appendix, A.2); thermodynamic, because this force derives from a free energy rather than a potential energy. As such, it represents a thermal average of the effect the other degrees of freedom of the system have on the dynamics of ξ. In fact, it is the *mean force* deriving from the potential of mean force, *i.e.* the free energy profile. This idea is already present in the effective evolution equation for ξ (equation 4.21).

It can be shown (see Appendix, A.5.1) that $F'(\xi)$ reads:

$$\boxed{F'(\xi) = \left\langle \frac{\partial \tilde{U}}{\partial q_1} - k_B T \frac{\partial}{\partial q_1} \ln J(q) \right\rangle_{\hat{\xi}(x)=q_1=\xi}} \tag{4.38}$$

where a complete change from cartesian to generalized coordinates $x \to q$, such that $q_1 = \hat{\xi}(x)$, has been introduced. $J(q)$ is the Jacobian of this transformation, formally written as $|\partial x / \partial q|$. $\tilde{U}$ is the potential energy of the system expressed as a function of the generalized coordinates, and $\langle ... \rangle_{\hat{\xi}(x)=q_1=\xi}$ refers to the conditional canonical average when the first generalized coordinate q_1 is restricted to a given value ξ. The term between brackets, which is a function of the cartesian coordinates through the generalized coordinates, is usually called the *instantaneous force* (up to a sign inversion).

Upon comparison with the corresponding expression for the alchemical setting (equation 4.16), an extra term involving the Jacobian appears. It corresponds to a geometric entropy contribution, describing how the volume element in configurational space changes under the variable change.

As it stands, equation 4.38 could be used to estimate $F'(\xi)$ from simulation data given analytical expressions for $\tilde{U}$ and $J(q)$. However, it turns out that practical use is difficult, notably because the Jacobian term involves the cumbersome manipulation of second derivatives with respect to the generalized coordinates.

In addition, the procedure would involve a complete coordinate change for mathematical reasons, but the remaining $3N-1$ generalized coordinates (*i.e.* $q_2, ..., q_{3N}$) are irrelevant - their explicit specification may seem like an unnecessary hassle. Following this line of reasoning, den Otter (Otter 2000) and independently Ciccotti and co-workers (Ciccotti, Kapral, and Vanden-Eijnden 2005) showed than it is possible to avoid the complete coordinate change by proving the following formula:

$$\boxed{F'(\xi) = \left\langle \nabla U \cdot \frac{\boldsymbol{w}}{\boldsymbol{w} \cdot \nabla \hat{\xi}} - k_B T \nabla \cdot \frac{\boldsymbol{w}}{\boldsymbol{w} \cdot \nabla \hat{\xi}} \right\rangle_\xi} \tag{4.39}$$

where ∇ is the gradient operator with respect to the cartesian coordinates and $\boldsymbol{w}$ is an *arbitrary* vector-field satisfying (Ciccotti, Kapral, and Vanden-Eijnden 2005):

1. $\boldsymbol{w} \cdot \nabla \hat{\xi} \neq 0$.
2. For any holonomic constraint $\sigma(x) = 0$, $\boldsymbol{w} \cdot \nabla \sigma = 0$

In the case where $\hat{\xi}$ is a vectorial CV, *i.e.* $\hat{\xi} = (\hat{\xi}_j)_{j=1,..,n}$, n vector-fields $\boldsymbol{w}_j$ must be introduced and these requirements become:

1. $\boldsymbol{w}_j \cdot \nabla\hat{\xi}_i = \delta_{ij}$.
2. For any holonomic constraint $\sigma(x) = 0$, $\boldsymbol{w}_j \cdot \nabla\sigma = 0$

A pedagogical proof of equation 4.39 is provided in (Chipot and Pohorille 2007, Chapter 4).

One can see that equation 4.39 is actually a generalization of equation 4.38, which is recovered for the choice $w_i = (\boldsymbol{w})_i \equiv \partial_\xi x_i$ (the so-called "inverse gradient"). The interest of equation 4.39 is that one is at freedom to choose another expression for $\boldsymbol{w}$ purely as a matter of convenience, and without having to define an explicit coordinate change. However, mathematical limitations to the use of formula 4.39 remain because of the orthogonality conditions listed above.

4.3.4.2. Constrained dynamics - blue-moon sampling

One of the oldest PMF calculation scheme, and the first to use thermodynamic integration for that purpose, blue moon sampling was introduced by Carter et al. (1989). In this technique, the CV of interest is discretized and parallel simulations are performed in which the value of CV is *constrained* (rather than restrained). That is, a specific algorithm is used to ensure that any time, the value of the collective variable is kept equal to some constant value ξ_0. This contrasts with the "restrained" case, in which a harmonic potential is used to keep $\hat{\xi}(x)$ close to ξ_0, but free to fluctuate around it. The constraint (see also Appendix, A.2.2.3) acts as a force of the form $\lambda\nabla\hat{\xi}$ applied on the collective variable, where λ is a Lagrange multiplier. The Lagrange multiplier is computed on-the-fly during the constrained simulation. Then, one can show (Chipot and Pohorille 2007, Chapter 4) that the generalized thermodynamic force is related to λ by:

$$F'(\xi) = \frac{\left\langle Z_\xi^{-1/2}\left(\lambda + \frac{1}{2\beta Z_\xi}\sum_i \frac{1}{m_i}\frac{\partial\hat{\xi}}{\partial x_i}\frac{\partial \ln Z_\xi}{\partial\xi}\right)\right\rangle_{\xi,\dot{\xi}=0}}{\left\langle Z_\xi^{-1/2}\right\rangle_{\xi,\dot{\xi}=0}} \tag{4.40}$$

where:

$$Z_\xi \equiv \sum_i \frac{1}{m_i}\left(\frac{\partial\hat{\xi}}{\partial x_i}\right)^2 \tag{4.41}$$

and $\langle ... \rangle_{\xi,\dot{\xi}=0}$ refers to the canonical average when the value of $\hat{\xi}(x)$ is fixed to ξ and its time-derivative is fixed to zero. Briefly, this rather complicated expression arises because applying a constraint on $\hat{\xi}(x)$ automatically freezes its time-derivative to zero; this needs to be corrected to recover the correct, unbiased canonical average. The reader should refer for example to (Chipot and Pohorille 2007; Tuckerman 2010) for a detailed treatment.

4.3.4.3. Adaptive Biasing Force

The adaptive biasing force (ABF) method is an elegant unconstrained thermodynamic integration procedure in which an estimate of the free energy gradient is built on-the-fly and used to bias the dynamics of the system, enhancing the sampling (Chipot and Hénin 2005; Comer, Gumbart, et al. 2015; Darve and Pohorille 2001; Darve, Rodríguez-Gómez, and Pohorille 2008; Darve, M. A. Wilson, and Pohorille 2002; Hénin and Chipot 2004; Hénin, Fiorin, et al. 2010). Assuming that the collective variable under study behaves according to equation 4.21, *i.e.* overdamped Langevin dynamics, it is clear that the metastable nature of the dynamics is due to the existence of local minima of $F(\xi)$, which

are felt by the CV through the term $F'(\xi)$. If the underlying free energy landscape were completely flat ($F'(\xi) = 0\ \forall\xi$), then the dynamics would be a free diffusion without metastability. Thus, erasing the gradient term in equation 4.21 would lead to enhanced sampling. In probabilistic terms, if $F(\xi)$ is used as a biasing potential (*i.e.* if the potential energy $U(x)$ of the system is changed into $U(x) - F(\hat{\xi}(x))$), the probability distribution of ξ becomes uniform, which implies that the dynamics (along ξ) is no longer metastable. $F(\xi)$ thus appears as an excellent choice of biasing potential[1].

Of course, if $F(\xi)$ were known, there would be no point in doing a free energy calculation. Instead, an estimate $A_t(\xi)$ at time t of $F(\xi)$ should be used, provided that it satisfies $\lim_{t\to+\infty} A_t(\xi) = F(\xi)$. In adaptive methods, of which ABF is a prominent example, $A_t(\xi)$ is built on-the-fly during the simulation and simultaneously used to further bias the dynamics. Specifically, ABF uses equation 4.38 (or variants thereof) to progressively build an estimate $\Gamma_t(\xi)$ of the free energy gradient by averaging over the instantaneous force values visited by the dynamics. In turn, this gradient estimate is used to bias the dynamics. In CV-space, the ABF dynamics thus reads:

$$\begin{cases} \gamma_\xi \frac{\mathrm{d}\xi}{\mathrm{d}t} = -\frac{\mathrm{d}F}{\mathrm{d}\xi} + \Gamma_t(\xi) + L(t) \\ \Gamma_t(\xi) = \frac{1}{t}\int_0^t \mathrm{d}t' \phi(\xi(t'))\delta(\xi(t') - \xi) \end{cases} \tag{4.42}$$

where ϕ is the negative instantaneous force, *i.e.* the argument of the average in equation 4.38, and $\xi(t) = \hat{\xi}(x(t))$. The second line in equation 4.42 is the updating rule for $\Gamma_t(\xi)$, which is built as a time average over the trajectory.

Using the chain rule to compute the biasing force acting on the cartesian coordinates, one obtains an equation of motion of the form:

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -\gamma\frac{\mathrm{d}x}{\mathrm{d}t} - \nabla_x U + \Gamma_t(\hat{\xi}(x(t)) \cdot \nabla_x\hat{\xi}(x(t)) + L_x(t) \tag{4.43}$$

where a Langevin dynamics has been assumed for the cartesian coordinates, which is customary in molecular simulations.

The use of equation 4.39 to estimate the thermodynamic force requires orthogonality with holonomic constraints, and mutual orthogonality of CV components in multi-dimensional situations. Furthermore, a suitable expression for the vector-field $\boldsymbol{w}$ of equation 4.39 is not necessarily available for all types of CVs (see (Fiorin, Klein, and Hénin 2013)). These conditions represent specific restrictions to the usage of the Adaptive Biasing Force (ABF) methodology.

Consistency and convergence of ABF Formula 4.38 and its generalization (equation 4.39) contain a canonical average, *i.e.* an average taken with respect the Boltzmann distribution of the unbiased potential. Yet, in ABF the bias is estimated by a time-average over the biased trajectory. As such, it may seem that this time-average will not correctly estimate the un-perturbed canonical average. In fact, one notices that the bias acts only on ξ, and thus does not affect the statistical distribution of the other degrees of freedom (*i.e.* the remaining $3N - 1$ generalized coordinates). Since the canonical average at hand is conditioned by the value of ξ, it is independent on any bias applied to it. Mathematical proofs of the convergence of the ABF bias to the true free energy gradient exist and are outlined in (Lelièvre, Stoltz, and Rousset 2010).

1. It is nonetheless pointed out in (Lelièvre, Stoltz, and Rousset 2010) that there is currently no proof that it is an *optimal* bias, for example the one that would maximize the convergence speed. The possibility exists that other choices for the biasing potential may lead to faster convergence.

Numerical considerations In this paragraph, we explain how the ABF dynamics is implemented for usage in simulations. In practice, the collective variable of interest is discretized between user-defined boundaries ξ_{min} and ξ_{max}, with a bin width $\delta\xi$. For each bin, an estimate of the mean force is built by collecting the values of the instantaneous force each time the bin is visited by the simulation, and averaging them. Experience shows that it is ill-advised to apply the bias immediately, as its value is expected to be very sensitive to noise in the initial stages of the simulation. Applying the bias in this case would result in strong non-equilibrium effects that could impede convergence and adversely affect the stability of the simulation. Instead, the standard protocol is to wait until a given bin has been visited for a pre-defined number of times before fully applying the bias. This number is called the *fullSamples* parameter; while the "number of counts" (*i.e.* the number of times a bin has been visited) is inferior to *fullSamples*, a scaled-down bias is applied. For instance, in the implementation used in this work, no bias is applied as long as the number of counts is below $fullSamples/2$, then the applied bias is linearly ramped up so as to be fully applied when the number of counts reaches *fullSamples*. In standard ABF, *fullSamples* and $\delta\xi$ are the only free parameters (and a $\delta\xi$ should anyway be chosen for any free energy calculation scheme). For a given $\delta\xi$, taking a large value of *fullSamples* should ensure that the bias is already reasonably converged when it starts to be applied, which should favor smooth convergence of the calculation. However, it will require longer simulations.

Multi-dimensional case In the multidimensional case ($n \geq 2$), several subtleties arise regarding ABF. First, there is an orthogonality condition which must be satisfied by the CV-components, see above (4.3.4.1). In practice, this notably translates into the impossibility to use CVs defined with overlapping sets of atoms. Second, unlike the one-dimensional case, there is no guarantee that the average force estimate $\Gamma_t(\xi)$ (where in this case $\xi = (\xi_1, ..., \xi_n)$) is a gradient, *i.e.* that there exists a function G such that $\Gamma_t(\xi) = \nabla_\xi G$ and extra post-processing steps must be taken to ensure that the final bias estimate is indeed a gradient.

Generalized ABF As for all geometrical free energy calculation schemes, it is very challenging to go beyond small values of n (say 2 or 3), because the volume of configurational space to be explored scales exponentially with n, and so does the memory required to store the grid. Generalized Adaptive Biasing Force (gABF) (Chipot and Lelièvre 2011; Lelièvre, Stoltz, and Rousset 2010) replaces the n-dimensional biasing function by a sum of n 1-dimensional biasing functions, *i.e.* the total gABF bias reads:

$$\Gamma_t^{gABF}(\xi_1, ..., \xi_n) = \sum_{i=1}^{n} \Gamma_t^i(\xi_i) \tag{4.44}$$

where each Γ_t^i is built as a gradient estimate along ξ_i the same way it would be for a 1-dimensional ABF calculation.

The rationale behind this expression is the following. Taking $n = 2$ to simplify, let us assume that ξ_1 and ξ_2 are statistically independent. Then, the joint probability density $P_{joint}(\xi_1, \xi_2)$ equals the product of the individual densities: $P_{joint}(\xi_1, \xi_2) = P_1(\xi_1)P_2(\xi_2)$. Equivalently, the corresponding two-dimensional PMF $F_{joint}(\xi_1, \xi_2)$ will satisfy:

$$
\begin{aligned}
F_{joint}(\xi_1, \xi_2) &= -k_B T \ln P_{joint}(\xi_1, \xi_2) \\
&= -k_B T \ln P_1(\xi_1) P_2(\xi_2) \\
&= -k_B T \ln P_1(\xi_1) - k_B T \ln P_2(\xi_2) \\
&= F_1(\xi_1) + F_2(\xi_2)
\end{aligned}
\tag{4.45}
$$

This shows why, in the case were the two CV-components are independent, the total bias can be written as a sum of individual biases. Thus, gABF is expected to perform well when weakly coupled collective variables are used. We note that it is somewhat unclear for us whether the method will converge even in the case of coupled collective variables. At any rate, for small n, it seems that gABF can be used to quickly obtain a first approximation of the n-dimensional gradient, to be used as the starting bias in a conventional ABF simulation. This reportedly led to faster convergence than running a full conventional 2D ABF calculation, in the case of alanine dipeptide (Chipot and Lelièvre 2011).

Conclusion on ABF ABF appears as a natural (*i.e.* , firmly grounded in statistical mechanics) method for free energy calculations. Consequently, the method is nearly parameter-free. The form of the bias is both rigorously justified and easy to grasp intuitively. In addition, a mathematical proof of convergence is available. These points make ABF a very appealing strategy. However, in the formulation we have outlined above, it still suffers from a number of drawbacks - mostly, the incompatibility with holonomic constraints, the requirement of orthogonality for multidimensional calculations, and the unavailability of suitable expressions for the instantaneous force for several classes of CVs. A more recent formulation, termed extended ABF (eABF), alleviates most of these problems while retaining most of the strengths of the original method. eABF is presented below (see 4.3.6.2).

4.3.4.4. Gradient estimation by harmonic restraints

We now outline a completely different route to the estimation of the free energy gradient, which relies on harmonic restraints. Let us consider a point ξ in colvar space at which we want to estimate the free energy gradient, and add a harmonic restraining potential centred in ξ, defining a new potential $V_k(x) = U(x) + \frac{1}{2}k(\hat{\xi}(x) - \xi)^2$. The free energy F_k associated with this potential is parametrized by ξ and will be denoted as $F_k(\xi)$:

$$
e^{-\beta F_k(\xi)} \equiv Z_k(\xi) = \int \mathrm{d}x e^{-\beta V_k(x)} \tag{4.46}
$$

$F_k(\xi)$ is sometimes called the "mollified" free energy profile along ξ.

It turns out that the following convergence property holds (Maragliano, A. Fischer, et al. 2006; Maragliano and Vanden-Eijnden 2006):

$$
\frac{\mathrm{d}F_k}{\mathrm{d}\xi} \underset{k \to +\infty}{\to} \frac{\mathrm{d}F}{\mathrm{d}\xi} \tag{4.47}
$$

where as usual $F(\xi) = -k_B T \ln \int \mathrm{d}x e^{-\beta U(x)} \delta(\hat{\xi}(x) - \xi)$. Intuitively, this property is expected to hold because the term $e^{-\frac{1}{2}\beta k(\hat{\xi}(x)-\xi)^2}$ in V_k will get closer and closer to a δ function as k increases. This qualitative reasoning can be made more rigorous by going to the Fourier space to obtain a proof of convergence, which is given in Appendix, A.5.2. This procedure also yields the error, which behaves as $\mathcal{O}(1/k)$.

Remark There is a somewhat counter-intuitive character to property 4.47, because one may expect that if $k \to +\infty$, one creates a free energy minimum in ξ and as such the free energy gradient should tend to 0, rather than $\mathrm{d}F/\mathrm{d}\xi$. This problem is easily solved when one remarks that F_k is *not* the PMF along ξ for the full Hamiltonian (force-field + harmonic restraint, *i.e.* $V_k(x)$). This PMF, say $G_k(\xi)$, would rather be defined as $G_k(\xi) = -k_B T \ln \int \mathrm{d}x e^{-\beta V_k(x)} \delta(\hat{\xi}(x) - \xi)$, which clearly is different from equation 4.46. $G_k(\xi)$ is indeed expected to admit a minimum in ξ if k is large enough.

Use for free energy calculations Property 4.47 opens the way to the estimation of the PMF (through its gradient) from harmonically restrained simulations with a high force constant. Indeed, one notices that:

$$\frac{\mathrm{d}F_k}{\mathrm{d}\xi} = -\frac{k_B T}{Z_k} \cdot \frac{\mathrm{d}Z_k}{\mathrm{d}\xi} \tag{4.48}$$

$$= -\frac{k_B T}{Z_k} \int \mathrm{d}x e^{-\beta U(x)} \frac{\partial}{\partial \xi} e^{-\frac{\beta k}{2}(\hat{\xi}(x)-\xi)^2} \tag{4.49}$$

$$= -\frac{k_B T}{Z_k} \int \mathrm{d}x e^{-\beta U(x)} \beta k \left(\hat{\xi}(x) - \xi\right) e^{-\frac{\beta k}{2}(\hat{\xi}(x)-\xi)^2} \tag{4.50}$$

which leads to:

$$\frac{\mathrm{d}F_k}{\mathrm{d}\xi} = -k\langle \hat{\xi}(x) - \xi \rangle_k \tag{4.51}$$

where $\langle ... \rangle_k$ is the average with respect to the Boltzmann distribution generated by the potential V_k. Numerically, this average can be readily estimated as a time average over a simulation of the potential V_k, assuming ergodicity.

The above considerations suggest a straightforward PMF calculation protocol, nearly identical in terms of set-up to umbrella sampling:

- Discretize the CV of interest into centers ξ_i
- For each window (*i.e.* each ξ_i value), run an MD simulation with the potential: $V_{k,i} = U(x) + \frac{1}{2}k(\hat{\xi}(x) - \xi_i)^2$
- For each window, estimate $\mathrm{d}F_k(\xi_i)/\mathrm{d}\xi$ from the trajectory average using equation 4.51
- Reconstruct the PMF by numerical integration of the gradient profile

To the best of our knowledge, the earliest use of this procedure is reported in (Van Eerden et al. 1989). Unlike conventional umbrella sampling, this procedure does not require overlap between adjacent windows. However, given the $k \to +\infty$ limit in property 4.47, it is expected to hold only for rather high force constants, and only approximately. In fact, resonance problems with the numerical integration of the equations of motion prevent the use of an arbitrarily high force constant (see for instance Conti and Cecchini 2018). Correcting procedures should be used to account for the finiteness of k. Umbrella integration (UI), which we now introduce, is one such procedure.

4.3.4.5. Umbrella Integration

Umbrella Integration (UI) builds upon the WHAM-idea of piecing together the information of parallel windows, but applies it to the estimation of the gradient profile, rather the free energy profile itself (Kästner and Thiel 2005). Taking the logarithm of equation 4.33 in a window simulated under the harmonic restraining potential of equation 4.25, one can write:

$$-k_B T \ln P_i(\xi) = \Delta F_i - V_i(\xi) - k_B T \ln \widetilde{P}_i(\xi) \tag{4.52}$$

where $P_i(\xi)$ is the unbiased distribution in window i. Equation 4.52 is reformulated as:

$$F_i(\xi) = -k_B T \ln \widetilde{P}_i(\xi) - V_i(\xi) + C_i \tag{4.53}$$

where $F_i(\xi) = -k_B T \ln P_i(\xi)$ is the unbiased PMF from window i, and we have renamed the offset ΔF_i to C_i to avoid confusion and highlight its status as an unknown constant. With WHAM, biased distributions from all the windows are combined to obtain C_i iteratively. Umbrella Integration takes a different route. We differentiate equation 4.53:

$$\frac{\mathrm{d}F_i(\xi)}{\mathrm{d}\xi} = -k_B T \frac{\mathrm{d} \ln \widetilde{P}_i(\xi)}{\mathrm{d}\xi} - \frac{\mathrm{d}V_i(\xi)}{\mathrm{d}\xi} \tag{4.54}$$

which makes C_i disappear. We now make the hypothesis that $\widetilde{P}_i(\xi)$ is Gaussian, which is justified in the case of a high enough harmonic restraint. This is the main assumption underlying the UI method, and it allows to propose an analytical expression for $\widetilde{P}_i(\xi)$:

$$\widetilde{P}_i(\xi) = \frac{1}{\widetilde{\sigma}_i \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\xi - \widetilde{\xi}_i}{\widetilde{\sigma}_i}\right)^2} \tag{4.55}$$

where $\widetilde{\xi}_i$ is the (biased) average value of ξ in window i, and $\widetilde{\sigma}_i$ its standard deviation. Plugging equation 4.55 in equation 4.54 (and differentiating the bias) yields:

$$\frac{\mathrm{d}F_i(\xi)}{\mathrm{d}\xi} = -k_B T \frac{\xi - \widetilde{\xi}_i}{\widetilde{\sigma}_i} - k_i(\xi - \bar{\xi}_i) \tag{4.56}$$

Equation 4.56 contains only known quantities, as $\tilde{\xi}_i$ and $\widetilde{\sigma}_i$ can be estimated from the trajectory of window i. However, it suffers from the same problem as 4.33, namely poor sampling of regions far from $\bar{\xi}_i$. Similar in spirit to WHAM, UI uses a weighted average over the windows to build a "global" gradient estimate $\frac{\mathrm{d}F(\xi)}{\mathrm{d}\xi}$ as:

$$\left.\frac{\mathrm{d}F(\xi)}{\mathrm{d}\xi}\right|_{\xi_{bin}} = \sum_i p_i(\xi_{bin}) \left.\frac{\mathrm{d}F_i(\xi)}{\mathrm{d}\xi}\right|_{\xi_{bin}} \tag{4.57}$$

for each bin center ξ_{bin}. The weights p_i are given by:

$$p_i(\xi) = \frac{N_i \widetilde{P}_i(\xi)}{\sum_i N_i \widetilde{P}_i(\xi)} \tag{4.58}$$

where N_i is the number of frames sampled in window i. Umbrella integration can straightforwardly be extended to dimensions higher than one (Kästner 2009). With respect to WHAM, it does not

require overlap between the adjacent windows - but it seems to us that taking very largely spaced, non-overlapping windows somewhat defeats the purpose of equation 4.57.

4.3.5. Metadynamics

Metadynamics is an adaptive biasing potential method in which a locally repulsive potential is gradually deposited in regions previously visited by the dynamics (Barducci, Bonomi, and Parrinello 2011; Laio and Parrinello 2002; Valsson, Tiwary, and Parrinello 2016). The ideas behind the technique can be traced back to previous methods such as local elevation (Huber, Torda, and Gunsteren 1994), conformational flooding (Grubmüller 1995) or hyperdynamics (Voter 1997). Like its predecessors, metadynamics "fills" the free energy basins and eventually pushes the system out of them, enhancing the sampling. At any time, an estimate of the potential of mean force can be obtained by taking the opposite of the biasing potential function.

In the modern formulation of the method (Iannuzzi, Laio, and Parrinello 2003), the repulsive potential is implemented as a sum of Gaussian functions in collective variable space. Gaussians are deposited every time τ at the current point.

At time t, the biasing potential takes the form:

$$V_{meta}(\xi, t) = \sum_k W_0 \exp - \frac{\left(\xi - \xi(k\tau)\right)^2}{2\sigma_\xi} \tag{4.59}$$

Since it is expected that the bias will "fill" the free energy landscape upon convergence, the following relation must hold in the long-time limit:

$$F(\xi) + V_{meta}(\xi, t) \underset{t \to +\infty}{=} C \tag{4.60}$$

where C is a constant.

The term "metadynamics" refers to the dynamic update of the bias as the simulation progresses, which represents another dynamics, concurrent to the Molecular Dynamics evolution of the system itself (Laio and Parrinello 2002).

In addition to τ, free parameters include the height W_0 and width σ_ξ of the Gaussian. A popular rule of thumb for the determination of the parameters is to choose W_0 and τ such that:

$$\frac{W_0}{\tau} \ll \frac{k_B T}{t_\xi} \tag{4.61}$$

where t_ξ refers to the correlation time of ξ (in unbiased dynamics). This criterion should ensure that the system can quickly relax to a new equilibrium after deposition of a new Gaussian. A more detailed discussion of the functional dependence of the residual error of the free energy estimate on the parameters can notably be found in (Bussi, Laio, and Parrinello 2006).

Well-tempered metadynamics In standard metadynamics, the height of the deposited Gaussians is constant. As a result, once the free energy landscape has been effectively flattened, any bias update creates a locally repulsive "bump" in the PMF. This leads to a phenomenon of oscillation of the PMF estimate, and makes it unclear when the run should be stopped (Barducci, Bonomi, and Parrinello 2011). In fact, a theoretical analysis of metadynamics showed that the biasing potential in the long-time limit indeed provides an estimate of the free energy profile, but with a non-vanishing quadratic error (Bussi, Laio, and Parrinello 2006).

This problematic behaviour is corrected in the so-called well-tempered formulation of metadynamics, in which a smoothly decreasing term is added to improve convergence (Barducci, Bussi, and Parrinello 2008). The Gaussian height W_0 in equation 4.59 is now rescaled as function of time, according to:

$$W(t) = W_0 \tau e^{-\frac{V_{wt}(\xi,t)}{k_B \Delta T}} \tag{4.62}$$

where $V_{wt}(\xi, t)$ is the history-dependent metadynamics bias collected during the well-tempered run (*i.e.* , it corresponds to equation 4.59, but with $W(t)$ replacing W_0) and ΔT is a parameter with the dimension of a temperature. In well-tempered metadynamics, as a given point of the free energy surface is visited, smaller and smaller Gaussians will be added at this point. It can be shown that under these conditions, the long-time limit of the bias satisfies:

$$V_{wt}(\xi, t) \underset{t \to +\infty}{=} -\frac{\Delta T}{T + \Delta T} F(\xi) + C \tag{4.63}$$

but this time, as the added bias is exponentially decreased, convergence is achieved in a single-run and easier to assess. Also, the biased probability distribution $P_b(\xi)$ upon convergence reads (Barducci, Bonomi, and Parrinello 2011):

$$P_b(\xi) \propto e^{-\frac{F(\xi)}{k_B (T + \Delta T)}} \tag{4.64}$$

which shows that standard MD (canonical sampling) is recovered for $\Delta T = 0$ and standard metadynamics (flat landscape in the long-time limit, $P_b(\xi) = cst$) for $\Delta T = +\infty$. Thus, ΔT appears as a tuning parameter to balance speed of the configurational exploration (higher ΔT) and preservation of smooth convergence (lower ΔT). Recent mathematical work has provided the explicit (time-dependent) form for C in equation 4.63, allowing for a derivation of a time-independent estimator for the free energy in well-tempered metadynamics simulations (Bonomi, Barducci, and Parrinello 2009; Tiwary and Parrinello 2015; Valsson, Tiwary, and Parrinello 2016). Notably, this allowed the derivation of a re-weighting procedure to compute the probability distribution of a non-biased collective variable, which is a non-trivial task for history-dependent biases.

ABF and metadynamics rest on different approaches to the estimation of the free energy, but also share a number of features. Both are examples of Adaptive Biasing methods, because the bias is constructed in a manner dependent on the system's history (Lelièvre, Stoltz, and Rousset 2010). Unlike ABF, it seems to us that the form of the bias in metadynamics is more arbitrary and requires more tunable parameters. However, metadynamics also has its advantages; in addition to the unbiasing scheme of orthogonal observables mentioned above, a procedure to estimate kinetic rates directly from a well-tempered metadynamics simulation has been developed (Tiwary and Parrinello 2013). To the best of our knowledge, equivalent techniques (in particular for the re-weighting of orthogonal observables) do not exist for ABF.

4.3.6. Extended degrees of freedom and extended dynamics

The idea of "extending" the system by adding artificial degrees of freedom is a common trick at the heart of many computational techniques, notably the Nosé-Hoover thermostat and Car-Parrinello first-principle Molecular Dynamics. Several similar methods have been proposed for free energy calculations (see Tuckerman 2010, for review). Here, we will focus on the subset of these methods in which the additional degree of freedom is harmonically coupled to the system. We will show how this

provides a powerful framework for the formulation of enhanced sampling and free energy calculations techniques. Mostly, we follow the presentation by Maragliano and Vanden-Eijnden (2006), but using the notations from the ABF community (Lesage et al. 2017).

We apply a harmonic restraining potential of center λ and force constant k on the collective variable $\hat{\xi}(x)$:

$$V_k(x,\lambda) = U(x) + \frac{1}{2}k\left(\hat{\xi}(x) - \lambda\right)^2 \tag{4.65}$$

The mollified free energy $F_k(\lambda)$, depending on λ and k, can then be obtained by partial integration over the cartesian coordinates:

$$F_k(\lambda) = -k_B T \ln \int \mathrm{d}x e^{-\beta V_k(x,\lambda)} \tag{4.66}$$

Now, let us treat λ as a dynamical variable, called the extended degree of freedom. Its "fictitious" dynamics will typically be taken of Langevin (possibly overdamped) type, that is, the extended degree of freedom is coupled to a thermostat[2]. Its equation of motion is thus (in the overdamped case):

$$\bar{\gamma}\dot{\lambda} = k\left(\hat{\xi}(x) - \lambda\right) + L_\lambda(t,\beta) \tag{4.67}$$

where $\bar{\gamma}$ is a fictitious friction coefficient, L_λ is the random Langevin force and β is the (inverse) temperature of the thermostat. Finally, we define the PMF $F(z)$ along the collective variable $\hat{\xi}(x)$ using the classical definition (equations 4.17 and 4.18) as:

$$F(z) = -k_B T \ln \int \mathrm{d}x e^{-\beta U(x)} \delta\left(\hat{\xi}(x) - z\right) \tag{4.68}$$

The PMF $F(z)$ introduced in equation 4.68 is the same as the one discussed throughout this chapter; the change of notation in the variable ($\xi \mapsto z$) is of no mathematical consequence, but will be useful to clarify its distinction from the λ variable.

The cornerstone of using harmonically extended dynamics for free energy calculation is that, under certain conditions discussed below, the evolution equation of λ takes the form:

$$\bar{\gamma}\dot{\lambda} = -\frac{\mathrm{d}F}{\mathrm{d}z} + L_\lambda(t,\beta) \tag{4.69}$$

In other words, it means that λ will undergo an effective dynamics over the 1D free energy landscape $F(z)$. There are two conditions required for equation 4.69 to be an effective equation for 4.67, that we now outline.

The first condition is that one must have:

$$k\left(\hat{\xi}(x) - \lambda\right) \simeq -\partial_\lambda F_k(\lambda) = k\langle\hat{\xi}(x) - \lambda\rangle_{k,\lambda} \tag{4.70}$$

where $\langle ... \rangle_{k,\lambda}$ is the canonical average with respect to the conditional canonical distribution generated by $V_k(x,\lambda)$ for a fixed λ. For any observable $B(x)$ it is defined as:

2. Other types of thermostatted dynamics can be used as well, but overdamped Langevin dynamics makes calculations easier.

$$\langle B \rangle_{k,\lambda} = \frac{1}{Z_k(\lambda)} \int \mathrm{d}x B(x) e^{-\beta V_k(x,\lambda)} \tag{4.71}$$

where $Z_k(\lambda) \equiv \exp(-\beta F_k(\lambda))$.

Equation 4.70 expresses that the dynamics of λ must be sufficiently slower than the one of the cartesian coordinates x in such a way that these latter equilibrate with the value of λ at all time. For instance, if the cartesian coordinates undergo Langevin dynamics with a friction coefficient γ (which is an usual situation in molecular simulations), this condition is satisfied in the limit $\bar{\gamma} \gg \gamma$. When this is the case, equation 4.67 can be averaged with respect to 4.71, leading to:

$$\bar{\gamma}\dot{\lambda} = k \left\langle \left(\hat{\xi}(x) - \lambda \right) \right\rangle_{k,\lambda} + L_\lambda(t,\beta) = -\frac{\partial F_k}{\partial \lambda} + L_\lambda(t,\beta) \tag{4.72}$$

Thus, if λ is a slow variable, it evolves according to the effective potential $F_k(\lambda)$. This leads us to the second condition, namely $\partial_\lambda F_k(\lambda) \simeq \partial_z F(z)$. Or, we have already established in section 4.3.4.4 that $\partial_\lambda F_k(\lambda) \to \partial_z F(z)$ when $k \to +\infty$ (strong coupling). When λ is a slow variable ($\bar{\gamma} \gg \gamma$) and is strongly coupled to the collective variable ($k \to +\infty$), its dynamics samples the PMF $F(z)$.

These observations shows that the adjunction of a harmonically coupled extended degree of freedom opens the way to the estimation of $\partial_z F(z)$, and thus of $F(z)$. However, there is no enhanced sampling in the dynamics of equation 4.69; as such, the exploration of the free energy profile will not be improved with respect to a regular Molecular Dynamics simulation. It turns out, nevertheless, that the extended approach can be readily combined with enhanced sampling approaches. We will review two such approaches, Temperature-Accelerated Molecular Dynamics (TAMD) and Extended Adaptive Biasing Force (eABF). Other combinations are also possible, such as with metadynamics (Ensing et al. 2006; Iannuzzi, Laio, and Parrinello 2003), which actually predates both TAMD and eABF.

4.3.6.1. Temperature-accelerated Molecular Dynamics

TAMD was introduced by Maragliano and Vanden-Eijnden (2006), and turns equation 4.69 into an enhanced sampling scheme simply by coupling λ to a higher-temperature thermostat than the one of the non-extended system. The equation of motion for λ reads:

$$\bar{\gamma}\dot{\lambda} = -\frac{\mathrm{d}F}{\mathrm{d}z} + \bar{L_\lambda}(t,\bar{\beta}) \tag{4.73}$$

where the extended friction coefficient $\bar{\gamma}$ and extended Langevin force $\bar{L_\lambda}(t,\bar{\beta})$ define a Langevin thermostat for the inverse temperature $\bar{\beta} < \beta$. In intuitive terms, TAMD is a practical way to "thermalize" a given collective variable at a higher temperature to specifically enhance the sampling along it. In principle one may increase the fictitious temperature so that $\bar{\beta}\Delta F^\ddagger = \mathcal{O}(1)$, where $\Delta F^\ddagger$ is the highest free energy barrier along $F(z)$, making the evolution of λ barrier-less and allowing for un-hindered sampling.

In addition, more than one CV can be included in the protocol (see (Maragliano and Vanden-Eijnden 2006)), in which case the enhanced sampling properties are preserved, even though the reconstruction of the multi-dimensional PMF is generally not feasible in practice for more than 3 or 4 CVs. Multi-CV TAMD has notably been used to capture large-scale conformational transitions in protein systems, providing insight into their functional mechanisms (Abrams and Vanden-Eijnden 2010).

4.3.6.2. Extended ABF

Extended ABF (eABF) combines the ideas of ABF and harmonically extended dynamics into a powerful free energy calculation framework which, it turns out, alleviates many of the limitations of the original method while retaining its elegance and simplicity. To the best of our knowledge, the first published formulation of extended ABF is found in Lelièvre, Stoltz, and Rousset (2010); it has since been the object of several methodological publications (notably Fu et al. 2016; Lesage et al. 2017).

The idea of eABF is to apply the ABF dynamics on the extended degree of freedom λ rather than directly on the collective variable of interest. Given the extended system of cartesian coordinates x and extended degree of freedom λ, a collective variable $\hat{\xi}_{ext}(x, \lambda) = \lambda$ is formally defined; a standard ABF dynamics is then applied on λ. This has the major advantage that the generalized force applied on λ is already known analytically: it is simply the harmonic force of the restraint. Consequently, there is no need for the (sometimes convoluted) expression of the mean force estimate; also, usage restrictions related to the mean force estimation (orthogonality with constraints and between biased CVs) are alleviated. In addition, it has been shown that the mean force estimate from eABF is typically much smoother than in ABF, which arguably accelerates convergence because the applied bias is less sensitive to noise (Lesage et al. 2017). This comes at the price of the introduction of a new parameter, namely the coupling force constant k.

Using the notations of section 4.3.4.3, the eABF dynamics in CV-space thus reads:

$$\begin{aligned} \bar{\gamma}\dot{\lambda} &= k\left(\hat{\xi}(x) - \lambda\right) + \Gamma_t(\lambda) + L_\lambda(t, \beta) \\ \Gamma_t(\lambda) &= \frac{1}{t}\int_0^t \mathrm{d}t' k\left(\lambda(t') - \hat{\xi}(x(t'))\right)\delta(\lambda(t') - \lambda) \end{aligned} \tag{4.74}$$

The second line of equation 4.74 is simply the time-average estimate of:

$$\Gamma(\lambda) = \langle k\left(\lambda' - z\right)\rangle_\lambda \tag{4.75}$$

where λ' is a dummy variable and $\langle ... \rangle_\lambda$ is the canonical average generated by the extended potential at a fixed value λ of the extended degree of freedom. Upon convergence of the eABF bias, the system thus evolves in the extended potential:

$$\tilde{U}(x, \lambda) = U(x) + \frac{1}{2}k\left(\hat{\xi}(x) - \lambda\right)^2 - A(\lambda) \tag{4.76}$$

where $A(\lambda)$ is the converged eABF *potential* bias, *i.e.* satisfying $A'(\lambda) = \Gamma(\lambda)$.

4.3.6.3. Corrected z-averaged restraint (CZAR) estimator

In practice, the strong coupling limit $k \to +\infty$ cannot be realized. As such, the reconstructed PMF is $F_k(\lambda)$ rather than $F(z)$. A deconvolution procedure is required to recover $F(z)$. We now outline the Corrected z-averaged restraint (CZAR) estimator, initially implemented in the *colvars* module (Fiorin, Klein, and Hénin 2013) and derived in (Lesage et al. 2017), see also Appendix, A.5.3. This estimator of $F'(z)$ is given by:

$$F'(z) = -\frac{1}{\beta}\frac{\mathrm{d}\ln\tilde{P}(z)}{\mathrm{d}z} + k\left(\langle\lambda\rangle_z - z\right) \tag{4.77}$$

$\tilde{P}(z)$ is the distribution of $z = \hat{\xi}(x)$ during the eABF trajectory. $\langle\lambda\rangle_z$ is the average value of

Free energy estimator	Sampler
Naive histogram	Conventional MD
Pieced histograms (WHAM, MBAR)	Static harmonic restraints with stratification (US)
Thermodynamic (umbrella) integration	Static harmonic restraints with stratification (US)
Thermodynamic integration	Adaptive Biasing Force
Thermodynamic integration	Blue-Moon Sampling
Non-equilibrium identities	Moving harmonic restraints (SMD)
Metadynamics estimator	History-dependent metadynamics bias

Table 4.1.: Common geometrical free energy calculations strategies: estimator + sampler.

λ, conditioned by the value of z. Both can be computed from a trajectory. Then, the Corrected z-Averaged Restraint (CZAR) estimator 4.77 can be used to reconstruct $F'(z)$, and numerical integration yields $F(z)$. Equation 4.77 can be readily extended to the multi-dimensional case (Lesage et al. 2017). Also, another estimator exists for eABF post-processing, which is based on Umbrella Integration (Fu et al. 2016; Lesage et al. 2017).

4.3.7. General perspective on geometrical free energy calculations

After this overview of geometrical free energy calculations, we are in a position to identify the common features to this rather wide variety of approaches. A successful geometrical free energy calculation scheme combines a "sampler" (*i.e.* a procedure designed to enhance the sampling by allowing for the exploration of low probability regions in configurational space) and a free energy estimator (*i.e.* a procedure to reconstruct the free energy profile from the data acquired during the dynamics). In table 4.1, we have classified most of the aforementioned free energy methods according to this framework. It is apparent from the table that a given estimator can be used to post-process data coming from various sampling strategies (*e.g.* ABF vs blue-moon sampling), and reciprocally that several estimators can be used to analyze the results of a given sampler (*e.g.* WHAM vs UI for umbrella sampling data). In addition, one may use a TI-estimator to post-process a conventional MD simulation (which would be equivalent to performing an ABF simulation without applying the ABF bias). Recent publications also report on the use of the metadynamics bias in combination with the TI-based estimator (Fiorin, Klein, and Hénin 2013; Mones, Bernstein, and Csányi 2016). In principle, one may thus combine the most appropriate sampler and estimator for the problem at hand.

Error and convergence analysis Assessing the error in a free energy calculation is a difficult task, as improper equilibration of the orthogonal degrees of freedom may lead to apparent convergence. Typical methods to evaluate convergence involve following the relative variation of the quantity of interest over time (*e.g.* the PMF or its gradient) and assessing whether a plateau is reached. After convergence is attained (or assumed), evaluation of the residual error may involve bootstrapping procedures (in which a randomly selected subset of the collected data are used to re-compute the PMF, assessing its robustness with respect to what should be akin to equilibrium fluctuations). Also, specific error estimators may be developed for specific free energy estimators.

Addressing orthogonal degrees of freedom The schemes described above enhance the sampling along the collective variable being biased, but strictly speaking their validity rests on the assumption that all

other degrees of freedom (*i.e.* orthogonal degrees of freedom) are properly equilibrated (*i.e.* Boltzmann-distributed). This is rarely the case in practice, and is sometimes a major hindrance to the proper convergence of a free energy calculation. To alleviate this, strategies for enhanced sampling of the orthogonal degrees of freedom have been proposed. In some cases, it may possible to identify the problematic degree of freedom and bias it explicitly, adding 1 to the dimensionality of the free energy calculation. However, when it is not clear which degree of freedom is faulty, or if it is clear that many of them are at play, alternate approaches must be considered. A very popular family of methods is based on replica-exchange (Sugita and Okamoto 1999). In replica-exchange, a collection of replicas of the system are simulated in parallel. The replicas differ by the value of one (or several) control parameter, which can for instance be the temperature (parallel tempering), the reference center of an umbrella-sampling bias (bias-exchange umbrella sampling) or another parameter of the Hamiltonian, etc (Fukunishi, Watanabe, and Takada 2002; Moradi and Tajkhorshid 2013; S. Park, Kim, and Im 2012; Sugita, Kitao, and Okamoto 2000). Thanks to this variation in the control parameter, each replica is expected to sample a different region of the configurational space and to capture independent transitions along the orthogonal degrees of freedom. The second ingredient in this type of calculation is the exchange between replicas. Every so often, an exchange of values for the control parameter is attempted between two replicas. Importantly, this exchange is accepted according to a (generalized) Metropolis criterion, which ensures that the data will be in a position to be re-weighted according to the Boltzmann distribution, despite the change of a parameter during the dynamics. Using this approach should ensure that favourable transitions along orthogonal degrees of freedom are propagated along the replicas. Note that it is possible to perform multidimensional replica-exchange using several control parameters (Sugita, Kitao, and Okamoto 2000).

Another approach, somewhat conceptually similar, is used in the context of ABF calculations and is called *shared ABF*(Comer, Gumbart, et al. 2015; Comer, Phillips, et al. 2014; Lelièvre, Stoltz, and Rousset 2010). In shared ABF, several independent replicas of the ABF dynamics are initiated; every so often, the estimated gradient bias accumulated in each replica is shared between the replicas. Several replicas can thus explore different regions of the configurational space at the same time, and capture different orthogonal transitions; the resulting ”consensus” ABF bias results from the data obtained in all the replicas. In more advanced implementations, selection procedures can even be used to eliminate replicas which are likely to sample irrelevant regions of the configurational space. Along the same lines, ”multiple-walkers” metadynamics is also used to accelerate the estimation of the free energy surface (Piana and Laio 2007).

4.4. Transition pathways and kinetics

In the previous section, we have detailed methods for the computation of the free energy profile along collective variables, but we have left out the discussion of the choice of the CVs. It seems clear that the PMF along $\hat{\xi}$ will be of interest only if $\hat{\xi}$ accounts, in one way or another, for an interesting feature of the system under study. Most often, *transitions* between metastable states are of central interest. For example, conformational transitions of proteins are crucial for their function and the detailed description of the rearrangements with atomic resolution can be a significant step towards the rational development of active molecules or the rationalization of mutant phenotypes. In addition, as we will outline, the study of transitions is closely intertwined with the prediction of kinetic rates; this is key, because kinetic rates are usually the dynamical quantities available experimentally, which allows for the comparison of the model to experience. By contrast, obtaining the atomically-detailed picture of a conformational transition still requires, generally, a significant amount of computational work.

We are interested in the study of a transition between two metastable basins, conventionally called R (reactant, starting state) and P (product, final state). Typically, a collective variable $\hat{\xi}$ (possibly multi-dimensional) is introduced to describe the transition, and must be chosen with care. The least stringent requirement for $\hat{\xi}$ is that it take significantly different values for the two basins. If this is the case, $\hat{\xi}$ is called an *order parameter*. An order parameter gives a distinction criterion between reactant and product states, but does not necessarily provide any information on the actual mechanism of the transition. We will see that this mechanism, under certain hypotheses, can be described by a so-called *optimal path*, described by a one-dimensional curve in the space of cartesian coordinates or collective variables, and connecting the reactant and product states. Various relevant definitions for this optimal path, and computational methods to determine it in molecular systems, are reviewed in this section. If such an optimal path is built, then we can introduce a progress parameter α, such that $\alpha(R) = 0$, $\alpha(P) = 1$, and α changes from 0 to 1 as one progresses along the path. α represents a *reaction coordinate* (or *transition coordinate*), *i.e.* a collective variable which parametrizes the actual mechanism of the transition. Thus, optimal path construction methods are actually techniques to obtain reaction coordinates, or good approximations thereof. Note that the use of advanced techniques for the construction of reaction coordinates is not always strictly required, as chemical/physical intuition may be enough to suggest good candidates especially for simple systems. However, when tackling conformational transitions of proteins, it is almost always necessary. Also, there are alternative approaches to the construction of reaction coordinates, which rely on different underlying principles, such as the minimum-cut method of Krivov and co-workers (Banushkina and Sergei V. Krivov 2016; S. V. Krivov and Karplus 2008; Sergei V. Krivov and Karplus 2006).

Finally, an alternative approach to the investigation of reaction mechanisms exists, which relies on *sampling of reactive trajectories*. This so-called *transition path sampling* family of methods will not be discussed here; the interested reader may instead refer to (Bolhuis et al. 2002).

4.4.1. Minimum energy paths

Let us a consider a system evolving under over-damped Langevin dynamics in a potential U:

$$\gamma \dot{x} = -\nabla U(x) + L(t) \tag{4.78}$$

where x is the $3N$-dimensional vector of cartesian coordinates, and other quantities should be interpreted accordingly; and where $L(t)$ is a Langevin random force satisfying a fluctuation-dissipation relation. For low temperatures, the noise is small and the trajectories will nearly behave according to the steepest-descent dynamics, up to small stochastic perturbations:

$$\gamma \dot{x} = -\nabla U(x) \tag{4.79}$$

Let us consider a transition between the reactant R and product P states (which both correspond to minima of U), under the stochastic dynamics 4.78. A Minimum Energy Path (MEP) connecting R to P is defined as a curve $\varphi(x)$ such that (E, Ren, and Vanden-Eijnden 2002):

$$\left[\nabla U\right]^{\perp}\left(\varphi(x)\right) = 0 \tag{4.80}$$

Along the Minimum Energy Path (MEP), the orthogonal component of the potential energy gradient (as thus, the orthogonal force) is 0; the only force is co-linear to the MEP (and is of course 0 at stationary points of the potential energy surface). See Figure 4.1 for a simple example of MEP. At low noise, typical reactive trajectories (*i.e.* trajectories in which a transition from R to P is realized) should

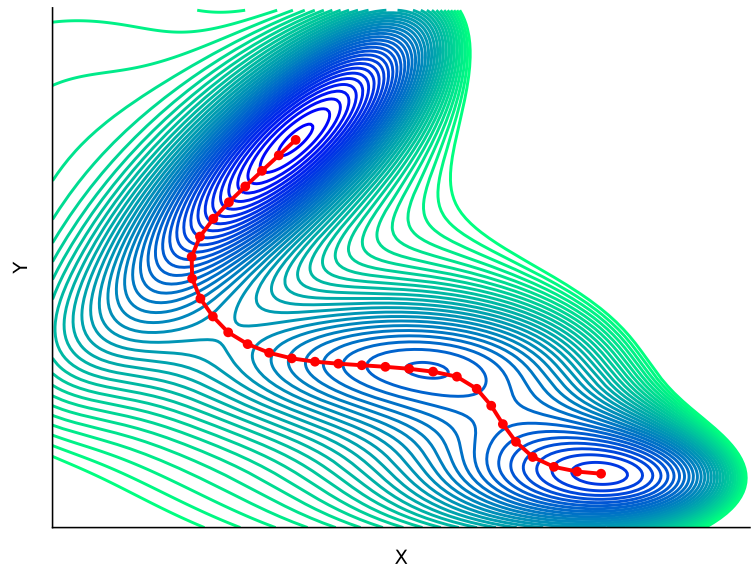


Figure 4.1.: Minimum Energy Path computed with the Zero-Temperature String Method (ZTS) on the Müller-Brown potential (Müller and Brown 1979).

be close to $\varphi(x)$, as this path entails the minimal increase in potential energy (and thus, the highest probability). Thus, the determination of an MEP is a strategy to investigate transition mechanisms, notably in molecular systems. Several schemes to that effect have been proposed, some of which we will briefly outline.

4.4.1.1. Conjugate Peak Refinement

Conjugate Peak Refinement (CPR) was introduced by Fischer and Karplus in 1992 (Stefan Fischer and Karplus 1992). The method focuses on the identification of saddle-points along the transition pathway, *i.e.* points in which the energy is maximal in one-direction and minimal in the others. These represent the transition states of the reaction. Briefly, Conjugate Peak Refinement (CPR) exploits an Hessian-independent conjugate minimization procedure to locate the saddle-points of the transition pathway, starting from a guess (usually a linear interpolation). The results is an ordered sequence of atomic configurations following the MEP connecting R and P. Fischer and co-workers have applied CPR to the recovery stroke of myosin (Stefan Fischer, Windshügel, et al. 2005), see Chapter 5.

4.4.1.2. Functional Optimization

The approach of functional optimization, introduced by Ölender and Elber (Ölender and Elber 1997), relies on the numerical minimization of the following functional:

$$\mathcal{F}(x(\varphi)) \equiv \int_{x_R}^{x_P} \sqrt{\nabla U^\top \nabla U}\,\mathrm{d}q \tag{4.81}$$

where U is the potential energy function and $\mathrm{d}q$ a distance element along the path. As proved in the original publication, the minimizing solution to equation 4.81 is a minimum energy path. Elber and

West use this approach to build a transition path of the myosin recovery stroke (Elber and A. West 2010), see Chapter 5.

4.4.1.3. Zero-temperature string method

The Zero-Temperature String Method (ZTS) was introduced by E, Ren and Vanden-Eijnden (E, Ren, and Vanden-Eijnden 2002). The idea is to relax a guess path by steepest descent along the orthogonal component of ∇U; thus, the converged path will be such that $\nabla U^{\perp}$ is everywhere 0, *i.e.* a MEP according to equation 4.80. Building upon the ZTS, several variants of the string methods were published, which exploit the same philosophy but aim at building different kinds of optimal paths. These variants will be reviewed below.

In the ZTS, the path is discretized into M images, or *string*, *i.e.* configurations of the system along the guess path connecting R and P. To relax this path into a MEP, the most recent version of the method (E, Ren, and Vanden-Eijnden 2007) uses the following two-step iterative procedure:

1. Each image is moved on the potential energy surface by steepest descent following the *full* gradient (not just its orthogonal component). At iteration τ, image $\varphi_i(\tau)$ is evolved according to:

$$\tilde{\varphi}_i(\tau+1) = \varphi_i(\tau) - \nabla U(\varphi_i(\tau))\Delta\tau \tag{4.82}$$

   where $\Delta\tau$ is a step size chosen by the user.

2. The $\tilde{\varphi}_i(\tau+1)$ are redistributed along the string so as to enforce a given parametrization; most often, equal spacing of the images is chosen. This step, called *reparametrization*, completes the iteration and yields the new, updated images $\varphi_i(\tau+1)$ along the string.

Reparametrization During reparametrization, equal spacing of the images along the string is enforced[3]. The purpose of this step is to compensate for the co-linear component of image displacement upon steepest descent; indeed, if images were allowed to move along the string, each image would eventually converge to the potential energy minimum whose attraction basin it belongs to. This is of course not desired when looking for transition pathways. The reparametrization step proceeds as follows:

1. The total "length" L of the string is computed by the following formula:

$$L = \sum_{i=1}^{M-1} |\tilde{\varphi}_{i+1} - \tilde{\varphi}_i| \tag{4.83}$$

   that is, the string length is taken as the sum of the individual lengths of the straight line segments joining successive images. The length of the string up to image i, l_i, is similarly defined:

$$l_i = \sum_{j=1}^{i-1} |\tilde{\varphi}_{j+1} - \tilde{\varphi}_j| \tag{4.84}$$

3. In fact, one may consider arbitrary parametrizations of the string, as discussed in (E, Ren, and Vanden-Eijnden 2007). We limit ourselves to the equal-spacing case, which is the most frequently encountered in applications.

2. The purpose of reparametrization is to move the images in such a way that the reparametrized images are equally-spaced along the new string. To that end, an interpolation procedure is used so as to compute the new positions of the images: an analytical expression $\varphi(\alpha)$ is constructed. Using this expression, generally a piecewise linear interpolation, one can redistribute the images along the string in an equidistant fashion without changing the shape of the string (Maragliano, A. Fischer, et al. 2006).

4.4.2. String methods in collective variables

For large molecular systems, the potential energy landscape is expected to be very rugged and as such, many alternate pathways can contribute to an overall transition between the reactants and the products. Given their local nature, it is highly likely that any of the methods presented above will only yield a *locally* minimal energy path, which may or may not be representative of the dominant pathway. In addition, when using the complete set of cartesian coordinates, the resulting MEP is expected to exhibit an irrelevant degree of details, because it will include very local atomic motions which are not interesting for a general description of the transition. Just like in the case of geometrical free energy calculations, it is convenient to reduce the complexity of the problem by projecting it onto a lower-dimensional collective variable space; the defining collective variables should be chosen in such a way that the relevant, global features of the transition (*e.g.* large-scale domain relative motions) are captured, but the irrelevant local atomic motions are filtered out. Upon performing this projection, it is expected that alternative MEPs which differ only through these irrelevant motions will coalesce, thereby offering a clearer, coarse-grained picture of the transition. When considering systems at finite-temperature[4], projecting upon the collective-variable space corresponds to studying the reduced dynamics on the free energy landscape; thus, optimal transition paths are now Minimum Free Energy Path (MFEP).

By analogy with the ZTS, one may evolve a string following the orthogonal component of the free energy gradient. This is the essence of the String Method in Collective Variables (CVSM), first published by (Maragliano, A. Fischer, et al. 2006).

4.4.2.1. Mean-force formulation

As before, given a (molecular) system described by its $3N$-dimensional cartesian coordinates vector x, we introduce a n-dimensional vectorial CV $\hat{\xi}(x) = (\hat{\xi}_1(x), ..., \hat{\xi}_n(x))$. This time, n needs not be restricted to small values. The free energy profile along $\hat{\xi}$ $F(\xi) = F(\xi_1, ..., \xi_n)$ is defined as usual from formula 4.20:

$$F(\xi) = -k_B T \ln \int e^{-\beta U(x)} \prod_{l=1}^{n} \delta\left(\hat{\xi}_l(x) - \xi_l\right) \mathrm{d}x \quad \equiv -k_B T \ln Z(\xi) \tag{4.85}$$

In this context, the reactant R and product P states are defined as local minima of F. An Minimum Free Energy Path (MFEP) is a one-dimensional path φ in collective variable space (*i.e.* expressed as a function of the ξ_l) connecting R and P, such that the orthogonal component of the free energy gradient vector is everywhere 0. Mathematically, this reads:

4. At finite-temperature, the irrelevant degrees of freedom are eliminated by Boltzmann-averaging them out. One may also imagine performing an elimination while remaining at zero-temperature, by "freezing" each irrelevant degree of freedom in its minimum energy value, which one could call "adiabatic elimination". Since we are dealing with thermalized systems, we will not be interested in these cases.

$$[\mathcal{M}(\xi)\nabla_\xi F]^\perp = 0 \tag{4.86}$$

As before, it is convenient to introduce the progress parameter α which goes from 0 in R to 1 in P along the path.

Because of the possibly curvilinear nature of the collective variable components, a position-dependent *metric tensor* $\mathcal{M}(\xi)$ must be introduced to define orthogonality. As discussed in Appendix (A.2.2.1), a metric tensor arises when a variable change to generalized coordinates is performed. In the present case, since $n < 3N$, there is no bijective mapping between the cartesian coordinates and the collective variable components. An averaging of the eliminated degrees of freedom is thus expected in the expression for the tensor. Indeed, it is shown in (Maragliano, A. Fischer, et al. 2006) that the generic element of the metric tensor reads:

$$\mathcal{M}_{ij}(\xi) = Z(\xi)^{-1} \int \sum_{k=1}^{3N} \frac{\partial \hat{\xi}_i(x)}{\partial x_k} \frac{\partial \hat{\xi}_j(x)}{\partial x_k} e^{-\beta V(x)} \prod_{l=1}^{n} \delta\left(\hat{\xi}_l(x) - \xi_l\right) \mathrm{d}x \quad = \sum_{k=1}^{3N} \left\langle \frac{\partial \hat{\xi}_i(x)}{\partial x_k} \frac{\partial \hat{\xi}_j(x)}{\partial x_k} \right\rangle_\xi \tag{4.87}$$

where $\langle ... \rangle_\xi$ is the canonical average conditioned by $\hat{\xi}(x) = \xi$.

If the free energy gradient and metric tensor are known, a steepest-descent in collective variable space can be formulated by a direct analogy with the ZTS, giving the so-called "mean force" variant of the String Method in Collective Variables (CVSM). A guess path in collective variable space is discretized into M images (which are now n-dimensional vectors) and is relaxed iteratively along the free energy gradient by steepest descent, while reparametrization is used to enforce equal spacing. With the notations of section 4.4.1.3:

$$\tilde{\varphi}_i(\tau + 1) = \varphi_i(\tau) - \mathcal{M}(\xi_i(\tau))\nabla_\xi F(\xi_i(\tau))\Delta\tau \tag{4.88}$$

followed by the re-parametrization step, which is essentially identical to the cartesian case. Note however that in practical uses, investigators typically normalize each collective variable component by its total variation along the string before re-parametrization, to ensure that each component is given equal weight (see for instance Lev et al. 2017; Takemoto et al. 2018). Finally, it is customary to smooth the string before re-parametrization to reduce the noise coming from the estimation of the metric tensor and the free energy gradient (see below); generally, local averaging is used. A small, positive smoothing parameter s is introduced and the updated images are modified according to $\tilde{\varphi}_i \leftarrow (1 - s)\tilde{\varphi}_i + \frac{s}{2}(\tilde{\varphi}_{i-1} + \tilde{\varphi}_{i+1})$.

Estimation of the free energy gradient and metric tensor The usage of equation 4.88 for string optimization requires the knowledge of $\mathcal{M}(\xi)$ and $\nabla_\xi F(\xi)$. Unlike the ZTS case, these are thermal quantities which must be estimated from MD simulations. For each image along the string, an MD simulation is performed with harmonic restraints (of force constant k_l for CV component ξ_l) centered on the image. Then, the procedure described in section 4.3.4.4 is used to obtain an estimate of $\nabla_\xi F$, *i.e.* using the following time-average:

$$\frac{\partial F}{\partial \xi_l} \simeq \frac{k_l}{t} \int_0^t \mathrm{d}t' \left(\xi_l - \hat{\xi}(x(t'))\right) \tag{4.89}$$

with an error in $\mathcal{O}(1/k_l)$ and $\mathcal{O}(1/\sqrt{t})$ (Maragliano, A. Fischer, et al. 2006). By a similar argument, the metric tensor is estimated as:

$$\mathcal{M}_{ij} \simeq \frac{1}{t}\sum_{k=1}^{3N}\int_0^t \mathrm{d}t' \frac{\partial\hat{\xi}_i(x(t'))}{\partial x_k}\frac{\partial\hat{\xi}_j(x(t'))}{\partial x_k} \tag{4.90}$$

Algorithm 1: The mean-force string method

Input: Number of images M, Initial path in CV-space $(\varphi_i)_{i=1,..,M}$, Force constants (k_l), Length of harmonic run t_1, Smoothing parameter s, Steepest-descent step size $\Delta\tau$

while String not converged **do**

  for $i = 1, ..., M$ **do**

    Run harmonically restrained MD (force constants (k_l)) at the image center φ_i for t_1;

    Evaluate the free-energy gradient by time-averaging

    $\partial_{\xi_l} F(i) \leftarrow \frac{k_l}{t_1}\int_0^{t_1} \mathrm{d}t' \left(\xi_l - \hat{\xi}(x(t'))\right)$;

    Evaluate the metric tensor by time-averaging

    $\mathcal{M}_{lm}(i) \leftarrow \frac{1}{t_1}\sum_{k=1}^{3N}\int_0^{t_1} \mathrm{d}t' \frac{\partial\hat{\xi}_l(x(t'))}{\partial x_k}\frac{\partial\hat{\xi}_m(x(t'))}{\partial x_k}$;

    Update the image : $\varphi_i \leftarrow \varphi_i - \mathcal{M}(i)\nabla_\xi F(i)\Delta\tau$;

  Smooth the string with parameter s;

  Re-parametrize the string;

Algorithm 1 summarizes the procedure of the mean-force string method. More recently, a variant of this method, called "on-the-fly" CVSM, was proposed in which the string is evolved concurrently with the images undergoing MD by the means of a harmonic extended dynamics (see 4.3.6), *i.e.* the string is treated as an extended degree of freedom (Maragliano and Vanden-Eijnden 2007).

4.4.2.2. Transition path theory and committor function

Beyond our qualitative justification of the MFEP, its significance for the study of reaction paths can be justified rigorously in the context of *transition path theory*, *i.e.* the statistical mechanical theory of paths in complex systems (E, Ren, and Vanden-Eijnden 2005b; E and Vanden-Eijnden 2010).

The central object of transition path theory is the *committor function* $q(x, v)$ (where v are the atomic velocities), which is defined as the probability that a trajectory initiated from the point (x, v) will reach the product state P *before* it reaches the reactant state R. In some sense, this quantity represents the "perfect" reaction coordinate for the R →P transition, notably because, by construction, $q = 0$ in R, $q = 1$ in P, and $q = 1/2$ at the transition state. In the mathematical theory of stochastic processes, it is shown that q satisfies a so-called backward Kolmogorov equation, a complicated partial differential equation which is, practically, impossible to solve analytically. Instead, approximations and numerical procedures must be used, of which the string method in collective variables is a prominent example. In this case, it is assumed that (Ovchinnikov, Karplus, and Vanden-Eijnden 2011):

1. The committor is independent of the velocities: $q(x, v) = q(x)$

2. There exists a (vectorial) collective variable $\hat{\xi}(x)$ such that $q(x) \simeq Q(\hat{\xi}(x))$ to a good approximation. This is actually what is meant by the requirement that the collective variables should be "good" descriptors of the transition.

3. Most of the reactive flux in CV-space is concentrated into a narrow channel, the transition tube.

Under these assumptions, it is first shown that the MFEP in terms of $\hat{\xi}(x)$ represents the most likely path for stochastic transitions between R and P; second, that the isocommittor surfaces can be approximated by the orthogonal hyperplanes to the MFEP in CV-space (Maragliano, A. Fischer, et al. 2006). The orthogonal hyperplane at progress parameter α along the string is defined by all the values of x such that (Ovchinnikov, Karplus, and Vanden-Eijnden 2011):

$$\sum_{l,m=1}^{n} \frac{\mathrm{d}\varphi_l(\alpha)}{\mathrm{d}\alpha} \mathcal{M}^{-1}_{lm}(\varphi(\alpha))(\hat{\xi}_m(x) - \varphi_j(\alpha)) = 0 \tag{4.91}$$

which thus gives the (approximate) expression for the isocommittor surfaces. Note that formula 4.91 expresses a simple orthogonality condition, but is more complicated than the familiar Euclidean case because the metric tensor must be accounted for.

Finally, going beyond the picture of a single-dimensional transition path, we note again that the MFEP represents the maximum probability path among a family of similar, roughly parallel paths in CV-space which together form the "transition tube". Implicitly in the justification of the MFEP, it is assumed that the transition tube is *narrow*. The Finite-Temperature String Method (FTS) was introduced to account for situations in which this assumption is not valid, see (E, Ren, and Vanden-Eijnden 2005a; Vanden-Eijnden and Venturoli 2009b).

4.4.2.3. Harmonic relaxation and swarms-of-trajectories variants

Among others, two slightly different variants of the CVSM have been proposed; the so-called "string optimization with swarms-of-trajectories" approach (Pan, Sezer, and Benoît Roux 2008), and a method by Zhu and Hummer (Zhu and Hummer 2010), which we deem "harmonic relaxation" method. Interestingly in these two methods, the string is evolved following its observed drift in CV-space, rather than by an explicit steepest descent along the orthogonal free energy gradient.

In the harmonic relaxation method, this is achieved by updating each image to its final position after a harmonically restrained run to the image center, before reparametrization (algorithm 2). In the swarms-of-trajectories method, each image is updated according to the ensemble-averaged drift in CV-space measured on a "swarm" of very short (typically a few ps) unbiased MD simulations, initiated from a previous harmonically restrained run at the image center (algorithm 3).

4.4.2.4. Brief comparison of the mean-force vs drift-based methods

There is a conceptual difference between the optimal pathways generated by mean-force CVSM, and these generated by drift-based CVSM (Johnson and Hummer 2012; Maragliano, Benoît Roux, and Vanden-Eijnden 2014). In the mean-force formulation, explicit steepest-descent along an estimate of the free energy gradient is used. This estimate is independent of the *dynamic* properties of the CV evolution, *i.e.* it does not depend on the diffusion coefficient of the collective variable. The only optimized quantity is the free energy, and this variant yields a true minimum free energy path. By contrast, in drift-based methods, the drift vector used to update the positions of the images depends both on the local free energy gradient and the diffusion coefficient. If this latter is position-dependent,

Algorithm 2: The "harmonic relaxation" string method

Input: Number of images M, Initial path in CV-space $(\varphi_i)_{i=1,..,M}$, Force constants k_l, Length of harmonic run t_1, Smoothing parameter s

while String not converged **do**

  for $i = 1, ..., M$ **do**

    Run harmonically restrained MD (force constants k_l) at the image center φ_i for t_1;

    Store final value of the CV vector $\xi_i(t_1)$;

    Update the image : $\varphi_i \leftarrow \xi_i(t_1)$;

  Smooth the string with parameter s;

  Re-parametrize the string;

Algorithm 3: The string method with swarms-of-trajectories

Input: Number of images M, Initial path in CV-space $(\varphi_i)_{i=1,..,M}$, Force constants k_l, Length of harmonic run t_1, Length of free run t_2, Number of trajectories within a swarm S, Smoothing parameter s

while String not converged **do**

  for $i = 1, ..., M$ **do**

    Run harmonically restrained MD (force constants k_l) at the image center φ_i for t_1;

    for $p = 1, ..., S$ **do**

      Run free MD for t_2 initiated from the last frame of the harmonically restrained run;

      Store final value of the CV vector $\xi_p(t_2)$

    Compute average new position over the swarm $\langle \xi_p \rangle(i) = \frac{1}{S} \sum_{p=1}^{S} \xi_p(t_2)$;

    Update the image with average new position: $\varphi_i \leftarrow \langle \xi_p \rangle(i)$;

  Smooth the string with parameter s;

  Re-parametrize the string;

the resulting optimal path after convergence will not be a *bona fide* MFEP. Rather, it may be called a Most Probable Transition Path (MPTP) (Pan, Sezer, and Benoît Roux 2008). Johnson and Hummer showed that the Most Probable Transition Path (MPTP) is less sensitive to changes in defining CVs than the MFEP (Johnson and Hummer 2012).

The detailed mathematical comparison between the two types of paths is discussed in the aforementioned references and the reader is referred to these publications for a rigorous treatment. However, we can intuitively illustrate this concept with the following metaphor. Imagine that we are trying to cross a mountain range. In principle, we should follow the minimal altitude path, that is, the path along which the orthogonal component of the gravitational potential energy gradient is everywhere zero. More clearly, this would entail following the bottom of valleys and crossing elevated regions at passes (saddle-points). But, perhaps the actual minimal energy path is actually not so practicable; for example, the bottom of a valley is frequently occupied by a river. Although one could walk in the river, it would be probably faster to walk along it on the bank. In this case, because the mobility along the minimal (free) energy path is a lot less than for a close, but energetically suboptimal path, the former ends up not being the optimal pathway to cross the mountain.

Due to this difference between MFEP and MPTP, in the following we will refer to the result of a string method calculation as an "optimal path" when the type of paths is not relevant for the discussion. Note that it is rather unclear whether this distinction is of critical importance in practical cases; in this thesis, the "swarms-of-trajectories" variant of the CVSM is used, more out of convenience of implementation than because of a motivated choice regarding the nature of the sought-after pathway.

4.4.3. Free energy along the path

Once an optimal path has been identified, one may be interested in evaluating the free energy profile along it, notably because it may provide valuable insight into the kinetics of the transition (see 4.4.4) and allow for the discovery of on-pathway intermediates. In the following, we will define more properly the free energy along the path, and we will see that it actually encompasses two different quantities which we will discuss. Also, note that unless specified otherwise, the notions and techniques presented in this subsection apply to arbitrary paths in CV space (*i.e.* non-necessarily optimal ones).

A path is defined as a parametrized curve $\varphi(\alpha)$ for $\alpha \in [0, 1]$ the progress parameter. Each point along φ has n coordinates (i.e. the path lives in the n-dimensional CV space) and we assume the existence of n smooth functions ξ_l such that $\varphi_l(\alpha) = \xi_l(\alpha)$. The full PMF $F(\xi)$ is obtained by application of 4.85. It is a function from $\mathbb{R}^n$ to $\mathbb{R}$, *i.e.* any point of the CV-space is mapped onto a single, real free energy value - it is not restricted to the path. This function is of tremendous interest, but is impossible to compute in practice. However, one can introduce a free energy *along the path* $F(\alpha)$ (chosen such that $F(\alpha = 0) = 0$ to get rid of the reference constant), such that:

$$F(\alpha) \equiv F(\xi(\alpha)) = F(\xi_1(\alpha), ..., \xi_n(\alpha)) \tag{4.92}$$

and, by the chain rule, its derivative reads:

$$\frac{\mathrm{d}F}{\mathrm{d}\alpha} = \sum_{l=1}^{n} \frac{\partial F}{\partial \xi_l} \cdot \frac{\mathrm{d}\xi_l}{\mathrm{d}\alpha} \tag{4.93}$$

where $\frac{\mathrm{d}\xi_l}{\mathrm{d}\alpha} = \varphi_l'(\alpha)$.

The reconstruction of $\frac{\mathrm{d}F}{\mathrm{d}\alpha}$ from simulation data requires the estimation of $\frac{\partial F}{\partial \xi_l}$ and $\frac{\mathrm{d}\xi_l}{\mathrm{d}\alpha}$. Section 4.3.4.4 shows how to estimate the $\frac{\partial F}{\partial \xi_l}$ using harmonically restrained simulations; and, for a given path defined by a set of discrete images, analytical $\varphi_l(\alpha)$ functions can be obtained using fitting procedures (Maragliano, A. Fischer, et al. 2006). Finally, the free energy profile is recovered by numerically integrating $\frac{\mathrm{d}F}{\mathrm{d}\alpha}$. This procedure is equivalent to a pseudo-n-dimensional "on-the-path" umbrella sampling (pseudo, because n collective variables are used to sample what is actually a one-dimensional curve). As such, the gradients can also be estimated by Umbrella Integration (UI), and bias-exchange can be used to enhance the sampling (Ovchinnikov, Karplus, and Vanden-Eijnden 2011).

By construction, $F(\alpha)$ represents the free energy of the single point at position α along the string. As such, it omits the contribution coming from the finiteness of the transition tube surrounding the optimal path, in which reactive trajectories also have a significant probability to take place. To account for this contribution -which can be seen as entropic as it corrects for the local configurational space volume around the optimal path-, a new free energy must be defined.

Such a free energy is proposed by Vanden-Eijnden and co-workers, based on the considerations that 1) the committor is the "perfect" reaction coordinate and 2) that orthogonal hyperplanes to the MFEP approximate isocommittor surfaces (see 4.4.2.2) (Ovchinnikov, Karplus, and Vanden-Eijnden 2011; Vanden-Eijnden and Venturoli 2009b). For notational convenience, we rewrite equation 4.91 as (Ovchinnikov, Karplus, and Vanden-Eijnden 2011):

$$g(x, \alpha) = 0 \tag{4.94}$$

For $g \in \mathbb{R}$, the canonical average over the atomic coordinates $\langle \delta(g(x, \alpha) - g) \rangle$ corresponds to the probability $P_\alpha(g)$ that $g(x, \alpha)$ takes on a given value g. For $g = 0$, x belongs to an isocommittor surface; that is, by the second assumption presented in 4.4.2.2 (*i.e.* $q(x) \simeq Q(\hat{\xi}(x))$), the $g(x, \alpha) = 0$ hyperplane approximates the isocommittor surface such that $Q(\hat{\xi}(x)) = Q(\varphi(\alpha))$.

Thus, the probability $P_\alpha(g = 0) = \langle \delta(g(x, \alpha)) \rangle$ gives the equilibrium probability as a function of the (approximation of) the committor function $Q(\varphi(\alpha))$ (Ovchinnikov, Karplus, and Vanden-Eijnden 2011). The associated free energy $G(\alpha)$ is defined as:

$$G(\alpha) \equiv -k_B T \ln \langle \delta(g(x, \alpha)) \rangle \tag{4.95}$$

$G(\alpha)$ is defined with respect to the complete isocommittor hypersurfaces, rather than simply points along the MFEP. Thus, it accounts for the transition tube - although in an approximate manner. Importantly, since the committor function is independent from the particular choice of CVs used to construct the string, so should $G(\alpha)$ provided that the chosen CVs are good in the sense outlined above.

4.4.3.1. Voronoi cell sampling

Vanden-Eijnden and Venturoli (2009b) provide a procedure to compute $G(\alpha)$. This approach relies on the *Voronoi tessellation* of the CV-space along the discretized reference path $(\varphi_i)_{i=1,\dots,M}$, *i.e.* the determination of polygonal cells, such that cell i encloses the region consisting of all points closer to φ_i than any other $\varphi_{j \neq i}$.

The Voronoi tessellation along the MEP for the Müller-Brown potential is shown on Figure 4.2. On this figure, it is clearly visible that the Voronoi tessellation constitutes a *partition* of CV-space; as such, sampling within a cell is by no-means restricted to the reference path, which, as we will see, opens the way to an estimation of the contribution of the transition tube to the overall free energy.

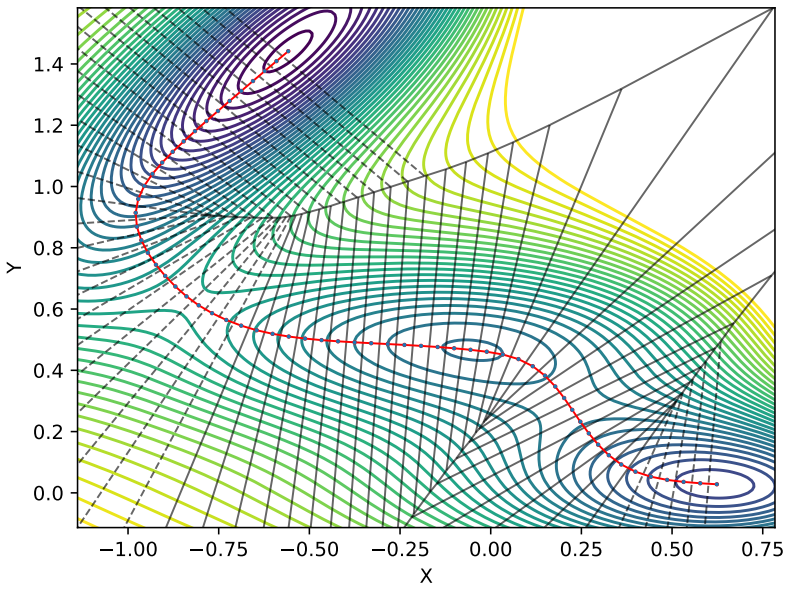


Figure 4.2.: Voronoi tessellation (computed with SciPy) along the MEP of the Müller-Brown potential.

The Voronoi cells form a discrete collection of states in CV-space, which can be assigned equilibrium occupancy probabilities (under ergodicity, these are the fractions of time the system spends in a given cell, for an infinitely long trajectory). We call π_i this occupancy probability for cell B_i centered on image φ_i of the reference path.

$$\pi_i = \frac{1}{Z} \int_{B_i} \mathrm{d}x e^{-\beta U(x)} \tag{4.96}$$

Also, it can be shown from their definition that the boundary between two adjacent cells is an orthogonal hyperplane to the reference path, *i.e.* an approximate isocommittor surface by equation 4.91. This lets Ovchinnikov, Karplus, and Vanden-Eijnden (2011) conclude that the following formula holds for π_i:

$$\pi_i \simeq M^{-1} e^{-\beta G_i} \tag{4.97}$$

where M is the number of images along the path and $G_i = G\left(\alpha = \frac{i}{M}\right)$. This shows that $G(\alpha)$ can be reconstructed if the π_i are known, for a fine enough discretization along the path. The method introduced in (Vanden-Eijnden and Venturoli 2009b) serves precisely this purpose. A restricted sampling procedure is introduced, such that the dynamics is confined within a cell, and a series of individual simulations are initiated along the cells. In the original implementation, confinement is achieved through momentum reversal upon collision with the cell boundary; however, this is rather cumbersome as it requires an *ad hoc* modification of the MD integrator. An equivalent, but more convenient procedure based on half-harmonic restraints was later introduced (Maragliano, Vanden-Eijnden, and Benoît Roux 2009).

For each cell, the number of collision events with the cell boundaries during the simulation (of length t_i) are recorded. A transition rate $\nu_{i,j}$ from cell B_i to cell B_j can then be estimated:

$$\nu_{i,j} = \frac{N_{i,j}}{t_i} \tag{4.98}$$

where $N_{i,j}$ is the number of collisions of the simulation confined in cell B_i with the boundary separating it from cell B_j. At equilibrium, conservation of probability implies:

$$\begin{cases} \sum_i \pi_i = 1 \\ \sum_{i,j} \pi_i \nu_{i,j} = \sum_{i,j} \pi_j \nu_{j,i} \end{cases} \tag{4.99}$$

which can be solved for the π_i, giving access to $G(\alpha)$. Ovchinnikov, Karplus, and Vanden-Eijnden (2011) performed this computation for the conformational transition of the myosin VI converter, and compared it with the free energy along the path $F(\alpha)$; it was found, unsurprisingly, that $G(\alpha) \neq F(\alpha)$, indicating that the width of the transition tube cannot be neglected or taken as constant along the transition path[5]. In fact, understandably, they found that $G(\alpha) < F(\alpha)$, as is expected because $G(\alpha)$ includes the missing entropic correction.

4.4.3.2. Path collective variables

As already explained, the path produced by a string method optimization is a single-dimensional object, whereas a complete description of the transition pathway should account for the transition tube. In addition, the string method is local and will relax towards the closest optimal path to the initial guess; there is no direct procedure to search for relevant pathways that are qualitatively different from the guess. Path collective variables, introduced in (Branduardi, Gervasio, and Parrinello 2007), aim at solving both problems: sampling around the reference path to probe the transition tube, and possibly identifying alternate paths. To that end, the authors introduce a pair of collective variables, termed path CVs, such that one (s) represents the position along the path, and the other (z) represents the orthogonal distance with respect to it. $s \in [0, 1]$ and can be seen as a "dynamical" version of the progress parameter α. Both s and z are defined for any point in CV space and are amenable to enhanced sampling strategies (*e.g.* metadynamics, umbrella sampling, ABF), which allows for an efficient exploration of the free energy surface $F(s, z)$. On this surface, the region of small z values corresponds to the immediate vicinity of the reference path, *i.e.* the transition tube.

Considering an arbitrary point $\Xi = (\xi_1, ..., \xi_n)$ in CV-space, and the α-parametrized reference path $\Gamma(\alpha) = (g_1(\alpha), ..., g_n(\alpha))$, the path CVs are defined as follows:

$$s(\Xi) \equiv \lim_{\lambda \to +\infty} \frac{\int_0^1 \mathrm{d}\alpha \alpha e^{-\lambda ||\Xi - \Gamma(\alpha)||^2}}{\int_0^1 \mathrm{d}\alpha e^{-\lambda ||\Xi - \Gamma(\alpha)||^2}} \tag{4.100}$$

$$z(\Xi) \equiv \lim_{\lambda \to +\infty} -\frac{1}{\lambda} \ln \int_0^1 \mathrm{d}\alpha e^{-\lambda ||\Xi - \Gamma(\alpha)||^2} \tag{4.101}$$

where $||...||$ is the norm in CV-space. The free energy surface $F(s, z)$ is then defined as usual, for example by:

$$F(s, z) = -k_B T \ln \langle \delta(s(\Xi) - s) \delta(z(\Xi) - z) \rangle \tag{4.102}$$

From then on, there are two ways of eliminating the variable z so as to compute a "free energy along the path":

5. With respect to this publication, we note that have inverted the notations for G and F.

- $F(s, z = 0)$ represents the free energy on the one-dimensional path, in the same sense as the "on-the-path" PMF defined above. It would be very interesting to investigate the relationship between these two quantities, which are ultimately related to the same notion.
- Writing $e^{-\beta F(s)} = \int \mathrm{d}z e^{-\beta F(s,z)}$, $F(s)$ (up to a constant) is the free energy profile along the path accounting for the entropic contribution of the transition tube, since points of arbitrary distance from the path are included in the integral. It gives the same information as the $G(\alpha)$ profile evaluated by Voronoi tesselation, and again, it is certainly worth investigating the connection between these two quantities.

Moreover, when using free energy calculations techniques to map $F(s, z)$, one may identify a low free energy region for z significantly larger than 0, and separated from the reference transition tube by a free energy barrier. This is the sign that an alternative transition pathway may have been located, and illustrates how path CVs can be employed to perform enhanced sampling in path space. Note however that z suffers from the same limitation as any "distance-like" CV (such as inter-atomic distances or RMSD): it becomes more and more degenerate as it increases. Because of this, enhanced sampling along path CV likely represents a good first step to identify competing pathways, but these latter will then have to be studied in more details with a more robust method like the string method.

Finally, as shown by the authors, the path CV can also be used as the basis for a variational path optimization scheme, by recognizing that the "path tension" $\mathcal{T}[\gamma(\alpha)] \equiv \int_0^1 \mathrm{d}s F(s, z = 0)$ is minimal if the path γ used to define the path CVs is a minimal free energy path. However, it is unclear whether this particular approach to path optimization compares favorably to the string method, and we will not detail it further here.

4.4.4. From free energy to kinetics

An optimal path provides a picture of the most probable mechanism for a given transition; then, the computation of a free energy profile along it will reveal the positions of possible intermediates (local minima), and free energy barriers between them. It turns out that the rate of a transition depends on the free energy barriers by a Boltzmann-type formula. As such, the determination of an optimal transition path by the string method, followed by a calculation of the free energy profile along it, provides a route to the computational determination of the rate, which can then be compared with experimental data.

4.4.4.1. Succinct introduction to reaction rate theory

The connection between the kinetics of a process and the height of energy barriers dates back to the empirical Arrhenius law for chemical reactions, which reads:

$$k(T) = A \exp\left(-\frac{E_a}{k_B T}\right) \tag{4.103}$$

In equation 4.103, $k(T)$ is the temperature-dependent rate constant (in s^{-1} for a first order reaction) and E_a is the so-called *activation energy*. The usual interpretation is that E_a represents the height of the potential energy barrier which must be crossed (through thermal activation) for the reaction to take place. These empirical considerations can be clarified by statistical mechanics.

Eyring-Polanyi formula and the quasi-thermodynamic model of reaction rate From a molecular point of view, the potential energy barrier (*i.e.* a potential energy maximum, or a saddle-point in more than one dimension) points to the existence of an energetically unstable atomic configuration called the transition state. If this transition state is considered, at least in thought, as an actual intermediate species along the reaction mechanism then one may study the equilibrium between it and the reactants. This approach, generally associated with the names of Eyring and Polanyi, leads to the derivation - from first principles- of a rate formula very similar in form to the Arrhenius law 4.103 (Eyring 1935). A pedagogical derivation using this quasi-equilibrium approach can be found in Hill (1986).

$$k_{Eyring}(T) = \frac{k_B T}{h} \exp\left(-\frac{\Delta G^{\ddagger}}{k_B T}\right) \tag{4.104}$$

where h is Planck's constant, and $\Delta G^{\ddagger}$ is the height of the free energy barrier for the transition initiated in the Reactant state.

Kramers theory The derivation of the Eyring formula 4.104 relies on several approximations that severely limit its applicability, even in the case of simple chemical reactions. *A fortiori*, it seems unclear how one may translate these ideas to more complex problems (of higher dimension) such as conformational changes of large molecules (and notably, protein folding). An alternative approach, due to Kramers (1940), relies on the theory of stochastic processes, and has found a wide range of applications for the study of complex molecular systems (Hänggi, Borkovec, and Talkner 1990). The system is modelled as a particle undergoing a diffusive dynamics on a free-energy surface; considering the statistical properties of this dynamics (using the Fokker-Planck equation), one can arrive at a general expression for the rate of escape from a free energy basin. In the high-viscosity limit and assuming local harmonicity for the reactant and transition states, Kramers found:

$$k_{Kramers}(T) = \frac{2\pi m \omega_1 \omega_2}{\gamma_\xi} e^{-\beta \Delta G^{\ddagger}} \tag{4.105}$$

with ω_1 the harmonic pulsation of oscillation in the reactant well, ω_2 an effective pulsation related to the curvature of the free energy surface in the vicinity of the transition state, m is the mass of the particle and γ_ξ the effective friction coefficient. See for instance (Zwanzig 2001) for a derivation. Despite being obtained by completely different routes, the above formulas have in common that the rate is given, up to a pre-exponential factor, by the Boltzmann exponential of a (free) energy barrier. This shows why the knowledge of the PMF grants some insight into the kinetics, at least qualitatively. A more quantitative estimation of the kinetic rate requires the estimation of the pre-factor; in Kramers' theory, relevant for conformational transitions, the pre-factor is controlled by the friction-coefficient, or equivalently the diffusion coefficient (By the Stokes-Einstein relation, one has $D_\xi = 1/\beta\gamma_\xi$ where D_ξ is the diffusion-coefficient). How can it be estimated?

4.4.4.2. Estimation of the position-dependent diffusion coefficient

When studying the kinetics of transitions in the collective variable space, it is customary to assume that the dynamics of the collective variables are of overdamped-Langevin (or diffusive) type, to a good approximation. If this is the case a position-dependent diffusion coefficient is well-defined, and influences the kinetics of stochastic transitions between metastable states. However, Kramers' formula 4.105 involves a constant diffusion (or friction) coefficient and must be generalized to allow compu-

tation of the rate. We introduce the Fokker-Planck equation associated with the diffusive dynamics of CV $\hat{\xi}(x) = \xi$ (or Smoluchowski equation) (Ovchinnikov, Nam, and Karplus 2016; Zwanzig 2001):

$$\frac{\partial P(\xi,t)}{\partial t} = \frac{\partial}{\partial \xi}\left[D(\xi)\left[-\beta f(\xi)P(\xi,t) + \frac{\partial P}{\partial \xi}\right]\right] \tag{4.106}$$

where $f(\xi) = -F'(\xi)$ is the free energy gradient, and $P(\xi,t)$ is the time-dependent probability distribution of ξ. Then, it can be shown that the rate k for a transition from R to P is given by:

$$k^{-1} = \int_R^P \mathrm{d}y \int_R^y \mathrm{d}z D^{-1}(y) e^{\beta[F(y)-F(z)]} \tag{4.107}$$

which reduces to the original Kramers's formula 4.105 for constant $D_\xi = 1/\beta\gamma_\xi$ and using an harmonic approximation for the reactant well and the transition state.

We note that while the rate is controlled exponentially by the free energy profile, it depends only linearly on the pre-factor; as such, the error made in estimating the diffusion coefficient is expected to have less dramatic repercussions on the rate than the error on the PMF. Several schemes have been proposed to estimate the diffusion coefficient, some of which we now briefly outline.

Autocorrelation function methods In non-equilibrium statistical mechanics, it is shown that a direct relation holds between the diffusion coefficient and the velocity-autocorrelation function $C_{\dot{\xi}}(\tau) = \langle\dot{\xi}(t)\dot{\xi}(t+\tau)\rangle$ ($\langle ...\rangle$ is the time average) (Zwanzig 2001).

This relation can be exploited to compute the diffusion coefficient by estimating $C_{\dot{\xi}}(\tau)$ from harmonically-restrained simulation trajectories (Hummer 2005; Woolf and Benoit Roux 1994). A harmonic restraint centered on the value ξ_i of the CV is applied, and the velocity-autocorrelation function in the window $C_{\dot{\xi}}(\tau,\xi_i) = \langle\dot{\xi}(t)\dot{\xi}(t+\tau)\rangle_i$ is computed.

Introducing the Laplace transform of the velocity-autocorrelation function:

$$\tilde{C}_{\dot{\xi}}(s,\xi_i) = \int_0^\infty e^{-s\tau} C_{\dot{\xi}}(\tau,\xi_i)\mathrm{d}\tau \tag{4.108}$$

it can be shown that the following quantity can be computed (Woolf and Benoit Roux 1994):

$$D(s,\xi_i) = -\frac{\tilde{C}_{\dot{\xi}}(s,\xi_i)\langle\delta\xi^2\rangle_i\langle\dot{\xi}\rangle_i}{\tilde{C}_{\dot{\xi}}(s,\xi_i)[s\langle\delta\xi^2\rangle_i + (\langle\dot{\xi}\rangle_i/s)] - \langle\delta\xi^2\rangle_i\langle\dot{\xi}\rangle_i} \tag{4.109}$$

(with $\delta\xi = \xi - \langle\xi\rangle_i$) such that the diffusion coefficient D satisfies $D(\xi_i) = \lim_{s\to 0} D(s,\xi_i)$.

Hummer (2005) provides a simplified expression using the position-autocorrelation function in the window $C_\xi(\tau,\xi_i) = \langle\delta\xi(t)\delta\xi(t+\tau)\rangle_i$. Using the CV correlation time τ_i given by:

$$\tau_i = \frac{\int_0^{+\infty}\langle\delta\xi(t)\delta\xi(t+\tau)\rangle_i \mathrm{d}t}{\langle\delta\xi^2\rangle_i} \tag{4.110}$$

Hummer finds:

$$D(\xi_i) = \frac{\langle\delta\xi^2\rangle_i}{\tau_i} \tag{4.111}$$

which can be used to reconstruct $D(\xi)$ using a stratification strategy. It may seem at first that this approach does not entail any additional calculations with respect to a traditional umbrella sampling

run. However, formula 4.111 is exact only if the dynamics of the restrained collective variable is harmonic and overdamped (Hummer 2005). This is *not* a requirement for the PMF calculation in umbrella sampling. In practice, this implies that the force constant must be high enough that the actual free energy gradient is negligible as compared to the harmonic force. This requires the use of large force constants (typically larger than for umbrella sampling), which will often involve extra calculations. In addition, the integration time step may have to be reduced if the force constant is too large.

Flat-bottom potential approach The flat-bottom potential approach by Ovchinnikov and co-workers is a simple method to compute at the same time the PMF and the diffusion coefficient (Ovchinnikov, Nam, and Karplus 2016). It relies on the usage of half-harmonic potentials, of the form:

$$V(\xi, a, b) = \begin{cases} \frac{1}{2}k(\xi - a)^2 \text{ if } \xi < a \\ 0 \text{ if } a < \xi < b \\ \frac{1}{2}k(\xi - b)^2 \text{ if } \xi > b \end{cases} \tag{4.112}$$

That is, the dynamics is confined between harmonic "walls" at positions a and b, but is otherwise unbiased. The authors show that, if $\Delta \equiv |b-a|$ is small enough that both the free energy profile $F(\xi)$ can be considered linear (*i.e.* $F(\xi) \simeq g\xi$, with g a constant) and the diffusion-coefficient $D(\xi)$ can be considered constant, then both can be estimated easily. When a simulation is performed under the biasing potential $V(\xi, a, b)$, the trajectory-average of the CV *excluding all frames which fall outside of the interval* $[a, b]$ is given by (assuming ergodicity):

$$\bar{\xi}_{MD} = \frac{\int_a^b \xi \exp(-\beta g \xi) \mathrm{d}\xi}{\int_a^b \exp(-\beta g \xi) \mathrm{d}\xi} \tag{4.113}$$

which can be solved for g using numerical procedures. Then, from $g \simeq F'(\xi)$, the PMF can be reconstructed by Thermodynamic Integration (TI). Regarding the diffusion coefficient, it is shown using equation 4.106 that the mean *round-trip time* T_{rt} satisfies:

$$T_{rt} = \frac{1}{Dg^2}\left[e^{g\Delta} + e^{-g\Delta} - 2\right] \tag{4.114}$$

Measuring the average round-trip time over the simulation (*i.e.* the average time taken to touch boundary a, then b, then a - or conversely) thus gives access to the value of D. Using a stratification strategy to cover the range of ξ values, both $F(\xi)$ and $D(\xi)$ are eventually recovered.

Bayesian analysis Bayesian approaches have emerged as a powerful framework for statistical inference in a variety of scientific fields, including the estimation of free energies from molecular simulation data (Habeck 2012). The general problem is to infer a set of parameters $\mathcal{P}$ (corresponding to a prescribed model) given experimental observations (in the broad sense) $\mathcal{O}$. For that purpose, Bayes' theorem on conditional probabilities is used:

$$\mathbb{P}(\mathcal{P}|\mathcal{O}) = \frac{\mathbb{P}(\mathcal{O}|\mathcal{P})\mathbb{P}(\mathcal{P})}{\mathbb{P}(\mathcal{O})} \tag{4.115}$$

and the set of parameters which maximizes the *a posteriori* probability $\mathbb{P}(\mathcal{P}|\mathcal{O})$ is determined, typically using Markov-Chain Monte-Carlo sampling techniques. This approach notably requires an

expression for the so-called *likelihood function* $\mathcal{L}(\mathcal{P};\mathcal{O}) = \mathbb{P}(\mathcal{O}|\mathcal{P})$ along with an assumption on the *a priori* probability (or prior) of the parameters $\mathbb{P}(\mathcal{P})$. For technical reasons, the remaining term $\mathbb{P}(\mathcal{O})$ usually needs not be computed for Bayesian estimation.

In the context of diffusion-coefficient estimation, the parameters to be estimated are of course the diffusion-coefficient $D(\xi)$ and sometimes also the PMF $F(\xi)$, and the observations are the time-series of CV values over the simulation. A prior on D and F is introduced, which may be flat or reflect some prior knowledge/intuition about these quantities. If diffusive dynamics is assumed, the likelihood function, which gives the probability of observing a given sequence $(\xi(t))$ for *specified* D and F, can be computed explicitly. In this context, Bayesian inference will yield the diffusion-coefficient and free energy profile which best match the observed trajectory under the assumed dynamics.

Hummer (2005) develops such an approach based on short unbiased MD runs used to construct a (fine) discrete-state approximation of the diffusive dynamics in CV-space, described by a transition rate matrix whose entries are estimated by Bayesian inference.

A similar-in-spirit method, proposed by Comer, Chipot, and González-Nilo (2013), was originally formulated for ABF simulations. Intuitively, upon convergence of the ABF bias, the system diffuses upon the flattened free energy landscape and its dynamics is controlled only by the position-dependent diffusion coefficient. Thus, after ABF convergence, a space-resolved (in CV-space) analysis of the local diffusivity (for example using mean-square displacements as a function of time) can lead to an estimate of D_ξ, as done for instance by Wereszczynski and McCammon (2012). However, this approach can be improved, because even before convergence of the ABF bias, the dynamics is affected by the diffusion coefficient. The idea behind the method by Comer and co-workers is to use a Bayesian estimation of the diffusion coefficient, knowing the trajectory, and the time-evolution of the ABF bias. It seems that the approach could also be used with other time-dependent methods such as metadynamics.

An attractive feature of Bayesian approaches is that it does not rely on the assumption of a diffusive dynamics; instead, any type of dynamics for which the likelihood function can be computed may be considered. Chipot and Comer (2016) illustrate this by studying sub-diffusive dynamics.

4.4.5. Direct estimation of the rate by milestoning

The assumption of diffusive dynamics which underlines rate calculations based on Kramers' theory is rather strong (it amounts to an assumption of Markovian behaviour) and may not be valid in practice. Also, the estimation of a kinetic rate from the joint knowledge of the free energy barrier and the position-dependent pre-factor is arguably not the most direct route. There are methods which allow for the determination of the rate directly, which grants them the attractive feature of being independent on the assumptions of transition-state theory and its generalizations. Of course, they still require some other assumptions to be valid. Markov State Models (MSM) are a popular approach along these lines, in which rates (and free energies) are estimated by measuring, in simulations, the number of transitions between discrete configurational sub-states of the system, assuming that these transitions are Markovian (Pande, Beauchamp, and Bowman 2010). These methods will not be reviewed further. Instead, we briefly outline a different method, *milestoning*, introduced by Elber and co-workers for rate computations (Faradjian and Elber 2004).

Milestoning is a reaction-coordinate-based method to estimate the timescale of the transition between two metastable states, introduced by Faradjian and Elber (2004). One needs a reaction-coordinate model ξ whose typical timescale is longer than any other timescale in the transition. A series of orthogonal hypersurfaces (complementary sets) are chosen along ξ. These are the *milestones*. Then, at

milestone s, several unbiased MD simulations are launched (from the equilibrium ensemble limited to the milestone). A simulation is stopped when it reaches $s \pm 1$, which should be the case in a reasonable time. One is then able to estimate the distributions $K_s^+(\tau)$ and $K_s^-(\tau)$ of the first-passage times to milestones $s+1$ and $s-1$, respectively. We have:

$$K_s^+(\tau)\mathrm{d}\tau = \mathbb{P}(T_{s\to s+1} \in [\tau, \tau + \mathrm{d}\tau]) \tag{4.116}$$

and a similar equation for $s \to s-1$ transitions. Mathematical procedures allow to compute the probability $P_s(t)$ for the system to be in milestone s at time t, from the knowledge of the K_s distributions. This is notably sufficient to compute the mean first passage time (MFPT) associated with the transition, whose inverse is the rate of the transition. The reader is referred to the corresponding publications for mathematical details (Faradjian and Elber 2004; A. M. A. West, Elber, and Shalloway 2007). We note that milestoning was used in a computational study of the myosin recovery stroke and that the predicted timescale agreed well with experimental data (Elber and A. West 2010).

Moreover, an alternate formulation called *Markovian milestoning* was proposed by Vanden-Eijnden and Venturoli (2009a). This procedure requires the same set-up as the Voronoi-restricted sampling introduced in (Vanden-Eijnden and Venturoli 2009b) and used to compute the corrected free energy along a path, $G(\alpha)$. In this formulation, the boundaries between Voronoi cells are used as milestones and statistics about the collisions with boundaries from restricted sampling are used to estimate the rate; see the corresponding publication for details (Vanden-Eijnden and Venturoli 2009a).

Milestoning with isocommittor surfaces Intuitively, one expects the quality of the rate prediction from milestoning calculations to be dependent on the quality of the chosen reaction-coordinate model ξ. In fact, one may wonder if there exists an *optimal* choice of ξ to obtain the best possible rate estimate. Vanden-Eijnden, Venturoli, et al. (2008) demonstrate that milestoning is optimal if the milestones are chosen to be the isocommittor surfaces, which is an important finding because these can be estimated using the CVSM. Thus, a good strategy for the study of a conformational transition seems to be 1) CVSM optimization 2) computation of the free energy profiles $F(\alpha)$ (on-the-path profile accessed *e.g.* by Umbrella Sampling (US)) and $G(\alpha)$ (corrected profile obtained by Voronoi-restricted sampling) and 3) rate estimation by milestoning along the optimal path (using the same Voronoi-restricted simulations). This is precisely the strategy used in (Ovchinnikov, Cecchini, Vanden-Eijnden, et al. 2011; Ovchinnikov, Karplus, and Vanden-Eijnden 2011) for the study of the fold transition in the myosin VI converter.

5. The recovery stroke of myosin and the PTS crystal structure

Summary The recovery stroke corresponds to the off-actin step of the motor cycle, in which the lever-arm is reprimed in the armed, pre-powerstroke configuration. It is also the step during which ATP is hydrolysed, which highlights why a proper description of the recovery stroke is crucial to understand chemo-mechanical transduction by the myosin motor. In this chapter, we review the existing literature on the recovery stroke of myosin, with a special emphasis on mechanistic models emerging from computational studies by various research groups. We arrive at the conclusion that existing data is not sufficient to conclude as to the most probable mechanism for the recovery stroke conformational transition. Then, we describe the novel PTS structure, and explain why it is consistent with a previously un-recognized intermediate along the recovery stroke, and that it suggests a novel mechanism for the transition. The PTS structure and its description are part of our publication (Blanc et al. 2018).

5.1. Structural changes in the motor domain upon the recovery stroke

The comparison of the PR and PPS crystallographic structures reveals the extent of the structural changes which characterize the recovery stroke. Such structures have been solved first for *Dictyostelium discoideum* Myosin II (Fisher et al. 1995; C. A. Smith and Ivan Rayment 1996), but other isoforms have followed over the years. X-ray structures of the Post-Rigor state and the Pre-Powerstroke state have been obtained for *Dictyostelium discoideum* Myosin II using ATP or ADP+Pi analogues (Fisher et al. 1995; C. A. Smith and Ivan Rayment 1996). The comparison of these structures shows that the converter rotates by about 60° during the recovery stroke. This conformational change is accompanied by a set of structural modifications within the main body (the motor domain excluding the converter). The Relay helix, which is one of the two connectors between the main body and the converter, bends during the transition, and forms a kink about halfway along its length by a rearrangement of backbone hydrogen bonds. The other connector, the SH1 Helix, secludes from the Relay helix and tilts in-place. These features seem to be common to all myosins, as they are observed on other crystallized myosins such as myosin VI (Ménétrey, Llinas, Cicolari, et al. 2008; Ménétrey, Llinas, Mukherjea, et al. 2007). Finally in the active site, switch II closes upon ATP, mostly with the formation of two important interactions: the so-called ”critical salt-bridge” between R238 and E459, and a hydrogen bond between the backbone nitrogen of G457 on Switch II and the γ-phosphate of ATP. Both these interactions are required to turn on the ATPase activity (Kiani and S. Fischer 2014; Li and Cui 2004; Onishi, Kojima, et al. 1998; Onishi, Ohki, et al. 2002), even though recent computational work has suggested that they may not be sufficient (Lu et al. 2017).

The main structural elements involved in the recovery stroke are illustrated on figure 5.1. Figure 5.2 shows the most important rearrangements taking place during the recovery stroke.

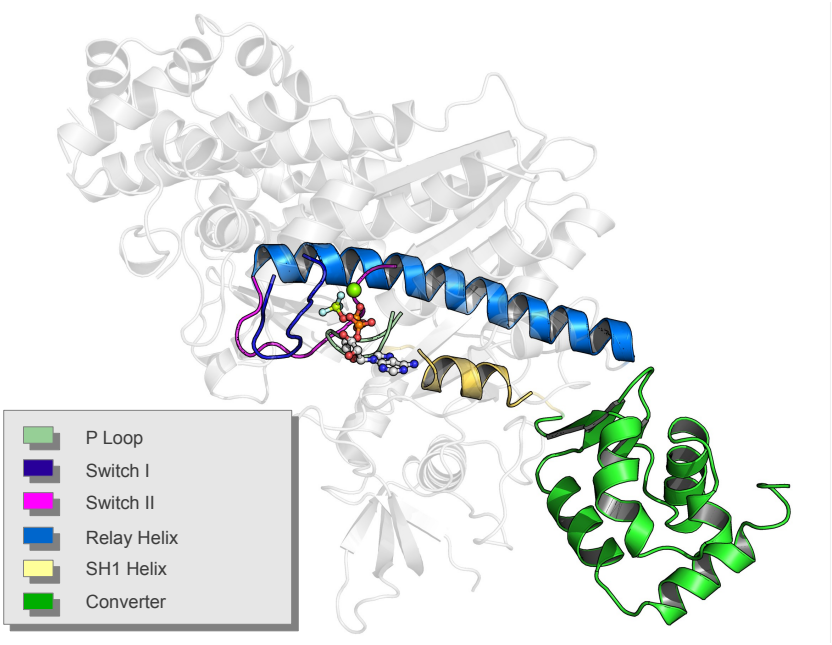


Figure 5.1.: Main structural elements involved in the recovery stroke, illustrated on the Post-Rigor structure of myosin VI.

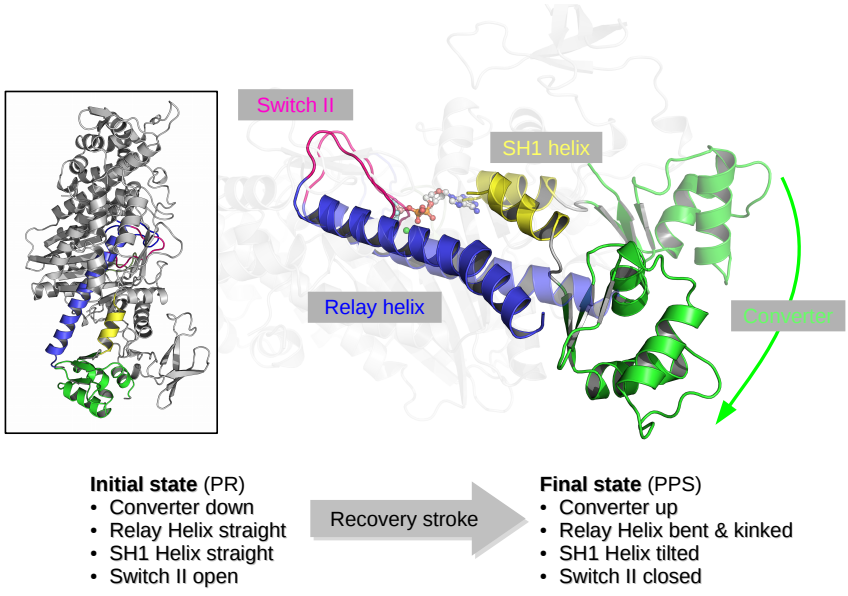


Figure 5.2.: Summary of the rearrangements taking place during the recovery stroke (myosin VI PR and PPS structures). The inset shows the full PR structure for reference.

5.2. Insights from solution and single-molecule experiments

Solution experiments performed on the Heavy Mero-Myosin (HMM) fragment showed that the hydrolysis of ATP by the motor domain was much faster than the release of ADP and Pi in solution in absence of actin (Adelstein and Eisenberg 1980; Lymn and Edwin W. Taylor 1970), and also that the hydrolysis happened off-actin (Edwin William Taylor, Lymn, and Moll 1970). Actin was proposed to increase the overall ATPase rate by accelerating the release of hydrolysis products upon interaction with the motor domain, rather than by a direct effect on the hydrolysis step. As early as 1980 it was suspected that the off-actin hydrolysis step was coupled with the angular swing of the cross-bridge, which would later be identified with that of the lever-arm. The resolution of myosin crystal structures showed that the converter/lever-arm could adopt a variety of angular positions. Eventually, the swing of the lever-arm during the recovery stroke was detected in solution by several independent investigators, using spectroscopic approaches such as resonant energy transfer (see for instance Shih et al. 2000, and references therein).

Also, it was observed that the addition of ATP triggered an increase in the intrinsic fluorescence of the motor domain, which was attributed to a conformational change affecting the W501 residue in *Dictyostelium discoideum* myosin II. Crystallography showed that this tryptophan residue was located at the Relay/converter interface, and this fluorescence increase was thus interpreted as the signature of the bending/kinking of the relay helix, *i.e.* the conformational transition of the force-generating region. This allowed Málnási-Csizmadia et al. (2001) to perform a kinetic study of the recovery stroke in solution conditions. The results showed that the conformational transition detected by fluorescence measurement (which the authors termed "open-close transition") was coupled to but distinct from ATP hydrolysis, and took place on the millisecond timescale. Moreover, they showed that the equilibrium ratio of the open-close transition (*i.e.* of the recovery stroke) was dependent on the nucleotide analog bound to the motor domain, with ADP.Pi analog ADP.AlF_4 favoring the post-recovery state, and ATP analogs AMPPNP and ADP.BeF_x giving an equilibrium constant close to unity.

Later on, time-resolved spin-labelling experiments by Agafonov and co-workers were designed to detect the rearrangement of the force-generating region and independently confirmed that both the pre- and post-recovery conformations are compatible either with ATP analogs or ADP.Pi analogs, even though ATP analogs (respectively ADP.Pi analogs) favor the pre-recovery conformation (respectively post-recovery) (Agafonov et al. 2009). These findings are also consistent with the observation that the motor domain can be crystallized in the pre-powerstroke configuration with either ATP or ADP.Pi analogs. Subsequent time-resolved Förster resonant energy transfer (FRET) studies by the same group suggested that the recovery stroke as a conformational transition preceded the hydrolysis of ATP (Y. E. Nesmelov et al. 2011). The model emerging from these studies is that the conformational transition of the motor domain is only loosely coupled to the hydrolysis of ATP. In addition, in absence of actin, no reversal of the recovery stroke was observed while myosin was bound to ADP.Pi (which would correspond to an unproductive ATP hydrolysis event). This suggests that the barrier for the reverse transition is high, as is expected since the PPS.ADP.Pi state must be of long-enough lifetime to find actin and continue the cycle.

Shiroguchi and co-workers were able to directly observe the recovery stroke of myosin Va at the single-molecule level (Shiroguchi et al. 2011). They followed the movement of a silicon bead attached to the lever-arm upon short, UV-controlled bursts of ATP in the system. This study emphasizes the importance of the recovery stroke in correctly orienting the detached head of the motor, which is crucial to find the next binding site of actin and partially explains the stability of the step-length. The recovery stroke is proposed to provide a forward bias that complements the one of the power stroke, and as such

is instrumental in allowing movement of the motor, especially under load. This contrasts with the more usual way of seeing the recovery stroke as a simple ”re-priming” of the lever-arm. Kinetic analysis gives an average lifetime of 40 s for the PPS state in bulk conditions (actin-free), before reversal of the recovery stroke. This points out the stability of the PPS state, which must be destabilized by actin binding to release the phosphate and proceed along the power stroke, in agreement with the previous proposal of (Y. E. Nesmelov et al. 2011).

Finally, in a recent time-resolved FRET study of myosin V, Trivedi and co-workers measured the kinetic rate of reverse lever-arm swing in the recovery stroke to be close to 300 s^{-1} (Trivedi et al. 2015), which puts the structural transition on the millisecond timescale, as already anticipated by previous studies. Separate tryptophan fluorescence measurements yielded a very close rate, which the authors intriguingly interpreted as supporting a tight-coupling between lever-arm swing and ATPase activation in the recovery stroke. However, it seems that the authors arrive at this conclusion by assuming that the change in tryptophan fluorescence is a proxy for ATPase activation, whereas it is more widely accepted that it measures the conformational rearrangement of the Relay helix. Thus, another interpretation of this result is that the movement of the converter/lever-arm and the rearrangement of the Relay helix are tightly coupled during the recovery stroke, which appears reasonable since they are in direct contact in the crystal structure.

Overall, time-resolved spectroscopy and single-molecule experiments confirm that the conformational changes anticipated by crystallography actually take place in solution during the recovery stroke. Furthermore, they put forward a dynamic view of the recovery stroke in which the coupling between the *structural state* (*i.e.* the conformation of the motor domain) and the *biochemical state* (*i.e.* the nature of the nucleotide bound in the active site) is loose, and possibly complete re-priming of the lever-arm may be explored before ATP hydrolysis takes place. However, despite their tremendous interest, these approaches are limited in temporal and spatial resolution. Most importantly, no reported solution study to date could investigate *directly* the rearrangement of the active site during the recovery stroke (switch II closure), and as such there is a shortage of experimental data on the kinetics of switch II closure and its coupling with the lever-arm swing, or its timing relative to the hydrolysis step. Unfortunately, it seems likely that introducing a FRET probe to monitor switch II closure would represent an intrusive modification affecting the kinetic rates and order of events during the transition.

Computational techniques, on the other hand, provide the opportunity to study the recovery stroke with an atomic level of detail and with total control on the perturbation introduced in the system - at the cost of the approximations and limitations intrinsic to using a numerical model rather than the real system.

5.3. Computational Models

As shown by independent experiments, the timescale for the recovery stroke transition is typically 1 ms. Considering the large size of the myosin motor domain (800 residues), this is far beyond the reach of unbiased all-atoms molecular dynamics, whose accessible timescales range from ns to µs. Rather, a description of the transition requires specialized computational approaches which make longer timescales accessible by introducing rather strong approximations. Several groups have applied a variety of enhanced sampling and transition path exploration methods to the recovery stroke of myosin (mostly Dd myo2). In the following we review these previously proposed models. This section builds upon the Supplementary Text 1 in (Blanc et al. 2018), which can be found in Appendix (C).

5.3.1. Fischer and co-workers (2005-2007)

To our knowledge, the first attempt to model the recovery stroke transition in myosin II has been made by Stefan Fischer, Windshügel, et al. (2005). Using the Conjugate Peak Refinement (CPR) method, these authors computed a series of structural intermediates along the minimum energy path connecting the PR state to the PPS state of Dd myo2. Based on these calculations, they propose that the recovery stroke starts with the spontaneous formation of the critical salt bridge between switch I (R238) and switch II (E459) (in myosin VI, the corresponding residues are R205 and E461), which brings the backbone nitrogen of G457 close enough to the γ-phosphorous of ATP to hydrogen-bond with it. Thus in this model, the overall transition starts with switch II closure. As the backbone oxygen of G457 is also involved in a hydrogen bond with the side chain of N475 on the Relay helix, switch II closure was proposed to introduce strain on the Relay helix. This mechanical pull on the Relay helix drives a seesaw motion of the helix about an aromatic fulcrum formed by the intertwined side chains of F481, F482 and F652. This seesaw motion consists in a rigid-body ”rocking” of the Relay helix relative to the N-terminal subdomain of myosin. Since the non-covalent bonds between the C-terminal part of the Relay helix and the converter are maintained during the transition, the rocking of the Relay helix induces the rotation of the converter. Later on, Fischer and co-workers built on this initial proposal with additional calculations. Most importantly, a second CPR study proposed a more detailed scenario involving two main stages (Koppole, J. C. Smith, and Stefan Fischer 2007). The first stage corresponds to the original findings described in (Stefan Fischer, Windshügel, et al. 2005), *i.e.* the seesaw motion of the Relay helix driven by the pull of switch II closure and driving roughly half of the converter rotation. In the second stage, the formation of an additional interaction between switch II and the P-loop (hydrogen bond between F458-S181) induces the rotation of the wedge loop, a rather conserved β-hairpin motif part of the L50 subdomain in the vicinity of switch II. Upon rotating, the wedge loop is observed to push against the SH2/SH1 junction and trigger a tilting motion of the SH1 helix (distinct from the seesaw motion of the Relay helix, which happens in the first stage). This tilting movement of the SH1 helix accounts for the remainder of the converter rotation and promotes the rearrangement of side-chains at the Relay/SH1 interface, which allows for the formation of the kink in the Relay helix. This region of the Relay/SH1 interface is proposed to be critical in stabilizing either the straight or kinked state of the Relay helix; as such, it is called the aromatic switch, or hydrophobic lock (residues F487, F506 and I687). The rearrangement of the aromatic switch involves the threading of the bulky F487 side-chain between the Relay helix and the Relay loop. This movement, which is required to form the kink and thus complete the recovery stroke, is claimed to be sterically hindered as long as the Relay helix has not undergone its seesaw motion. Thus, in Fischer’s model, the aromatic switch determines the sequentiality of the transition, as it forces the kink formation to take place after the seesaw motion of the Relay helix.

The main features of Fischer’s model are summarized on Figures 5.3 and 5.4. Though plausible, the conclusions of Fischer and co-workers are based on a zero-temperature single path which provides a strongly coupled picture of the transition, with rearrangements progressing deterministically in a smooth, progressive and concerted manner. This neglects the role of entropy in the recovery stroke at room temperature and may mask the stochastic character of the process.

The CPR studies were subsequently complemented by short MD simulations of the end-states. In the first study, short unbiased MD simulations were used to investigate the dynamics of the active site (Koppole, J. C. Smith, and Stefan Fischer 2006); it was found that switch I and switch II undergo a dramatic decrease in flexibility upon ATP binding, but are more flexible with ADP.Pi than ATP. This would support a role as sensors of the nucleotide state in the active site for these two important loops. The authors also proposed a mechanism to explain how the PPS configuration of the motor domain

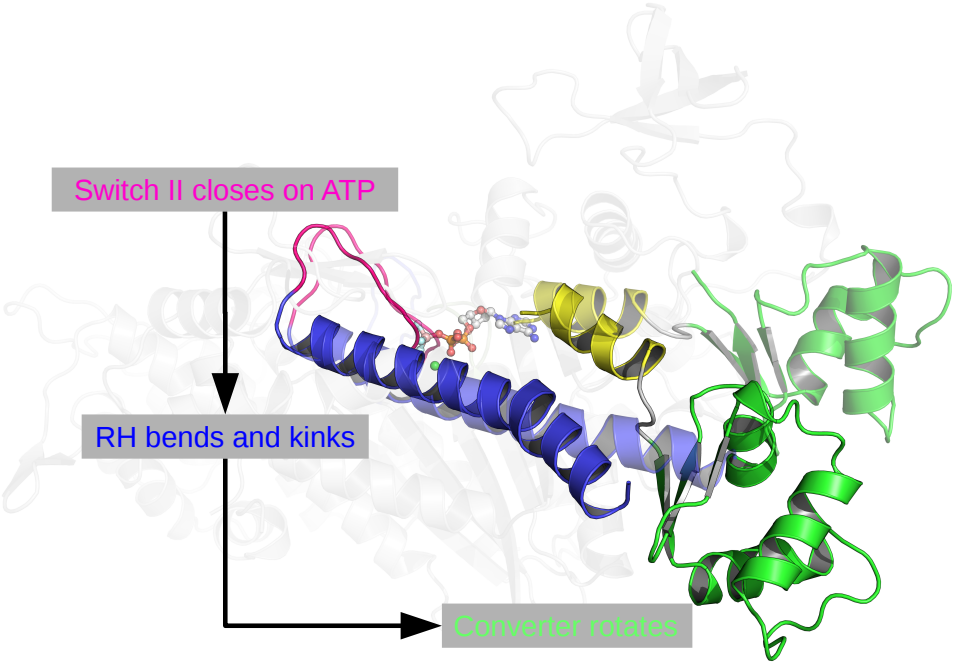


Figure 5.3.: Summary of the strongly-coupled model of Fischer and co-workers.

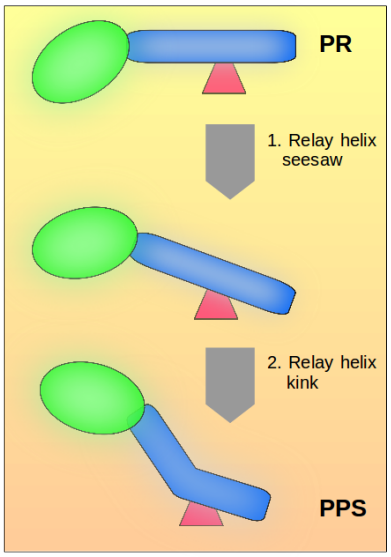


Figure 5.4.: Fischer's model corresponds to a two-stage mechanism in which the seesaw motion of the Relay helix precedes the formation of the kink.

may be stabilized in presence of ADP.Pi. With ADP.Pi, the hydrogen-bond between the carbonyl of G457 and the side chain of N475 was observed to break, uncoupling switch II from the Relay helix. More precisely, after ATP hydrolysis, the γ-phosphate is free to move away from the β-phosphate towards switch II. Switch II pushes on N475, which flips and forms a new hydrogen-bond with the side chain of Y573 (on the wedge loop), rather than with G457. This contributes to locking the Relay helix and the converter in the PPS configuration, in addition to explaining the enhanced flexibility of the switch II with ADP.Pi, because this latter is now uncoupled from the Relay helix. In the second study, principal component analysis of MD simulations shows that some of the functional predicted by CPR motions are detectable within the equilibrium fluctuations of the end states structures (Mesentean et al. 2007). Interestingly, the seesaw motion of the Relay helix was detected in the PR state simulation (*i.e.* the initial state), whereas SH1 helix movements (apparently akin to a "piston-like" motion slightly different from the tilting reported in the second CPR study) coupled to wedge loop displacements were observed in PPS. These findings seem to provide independent support to the order of events emerging from the CPR analysis. However, these conclusions rely on very short (about 5 ns) simulations and may suffer from convergence issues. In addition, there is no report of a tendency to switch II closure in the PR state, which would be expected in the framework of Fischer's model because it represents the initiating event of the transition.

5.3.2. Woo and co-workers (2007-2008)

To our knowledge, Woo's work is the first attempt to evaluate the free energy profile along the recovery stroke (Woo 2007). Woo used two-dimensional umbrella sampling, using the RMSD from PR and PPS structures as reaction coordinates, to study the recovery stroke of scallop myosin. Intriguingly, it is found that the transition from PR to PPS occurs downward a monotonic free-energy gradient, with the PPS state corresponding to a free-energy minimum. However, the very high reported free-energy difference between the PR and PPS ($-30\,\mathrm{kcal\,mol^{-1}}$), along with the fact that the PR state is not identified as a free-energy minimum, cast serious doubt onto the proper convergence of the umbrella sampling calculations. Notably, this latter observation can be explained by an improper relaxation of the PR-like structures, which were generated by applying a harmonic restraint on the PPS structure. An alternative explanation is a that it is a specific feature of scallop myosin, which has been reported to crystallize in PPS rather than PR with ATP analogs (Houdusse, Szent-Györgyi, and Cohen 2000), suggesting the PPS is the ground state even in presence of ATP for this isoform. By contrast, Dd Myo2 and Myo6 crystallize in PR with ATP analogs.

Although the estimate of the free-energy difference is probably unreliable, the umbrella sampling windows taken together constitute a "staged" Targeted Molecular Dynamics (TMD) simulation from the PPS to the PR. As such, it can still provide qualitative information on a possible transition mechanism between these two states. Notably, the simulations exhibit a gradual rotation of the converter and a gradual straightening of the relay helix as the system gets restrained towards the PR. Strikingly, the critical salt-bridge (R242-E465 in scallop myosin) is essentially always formed, except for the window restrained to the PR conformation. As the author points out, this is consistent with the early and spontaneous closure of this salt-bridge in Fischer's work, supporting its interpretation as the initiating event of the recovery stroke. Another possibility, however, is that the windows were too short in time (about 3 ns per window) to sample the re-opening of the salt-bridge upon going to the PR state from the PPS state. Lack of proper equilibration may lead to over-estimating the stability of the salt-bridge. Nevertheless, the overall sequence of events observed along the umbrella sampling windows is consistent with Fischer's proposal. However, it ought to be noted that no sampling is reported

for the off-diagonal regions of the free energy landscape; *i.e.* , the sampling is concentrated around a concerted pathway, which is expected to resemble Fischer's model. The possibility of an off-diagonal transition pathway, which would deviate strongly from Fischer's proposal, is thus not eliminated by Woo's study.

This point was partially addressed in a follow-up study by Harris and Woo (2008), in which the $\Delta RMSD$ between PR and PPS is used as a single-dimensional reaction-coordinate model (rather than the RMSDs from PR and PPS as a 2D reaction coordinate model, as in the first study), but in which the free energy landscapes along other collective variables, *e.g.* the Relay helix bending angle, are computed. Notably, the reported free energy landscape along ($RMSD_{PR}, RMSD_{PPS}$) covers a larger extent of the configurational space than the one of the previous paper. However, the observation that the minimal free energy transition path is diagonal is not changed. Interestingly, the authors report on infrequent "excursions" of the converter to a detached state, which is proposed to highlight the flexibility of the converter/motor domain connectors.

Overall, the work by Woo and Harris constitutes an interesting take on a quantitative study of the recovery stroke, but its significance is limited by the likely insufficient convergence of the PMF. Although the findings are mostly consistent with Fischer's model, it does not seem that this study was really in a position to provide a stringent test for this earlier proposal.

5.3.3. Cui and co-workers (2007)

Cui and co-workers proposed a model of the recovery stroke by combining an impressive diversity of techniques, most importantly umbrella sampling and targeted Molecular Dynamics (Yu et al. 2007a,b). Umbrella sampling calculations performed on the end-states of the recovery stroke pointed to a model in which ATP hydrolysis and the swing of the converter are statistically coupled to the open/close transition of switch II. In the presence of ATP in the active site, these calculations allowed them to probe the effective free energy landscape for switch II closure depending on the position of the converter; they found that a closed switch II is thermodynamically favoured in PPS (*i.e.* with a re-primed converter), while the open and closed states are accessible in PR (*i.e.* with an un-primed converter). Interestingly, TMD simulations of the transition showed an early rotation of the converter and a late switch II closure, thus suggesting a different order of events from the proposals by Fischer's and Woo's models. In comparison to CPR, TMD natively accounts for thermal fluctuations, but a well-known artifact is that larger rearrangements tend to happen first in the simulated transition. This artifact may explain the fact that converter rotation (a rather global change) precedes switch II closure (a more local change). Also, the observation from free energy calculations that switch II may be able to reversibly close and open while the motor domain is in the pre-recovery (PR) configuration does not exclude the possibility that switch II closure is actually an early event.

5.3.4. Elber and co-workers (2010)

Another attempt to model the recovery stroke of myosin was published by Elber and A. West (2010). In this study, the authors first use a functional optimization approach to obtain a minimum energy path for the recovery stroke, then use the milestoning procedure to compute the rate of the transition. The minimum energy path is mostly consistent with the order of events proposed by Fischer and co-workers - early rearrangements in the active site and later movements in the Relay/SH1/converter region. Probably the most striking result of this study is the remarkable agreement between the rate estimated by milestoning (0.5 ms) and the experimentally measured time-scale for the transition. By

contrast, the proposed transition mechanism derives from a zero-temperature path determination technique and as such is likely to suffer from the same limitations as Fischer's CPR results, even though the agreement with the experimental rate supports the biological relevance of the model.

In addition to the overall rate, the milestoning calculations allow the estimation of time delays between successive structural transitions; West and Elber conclude that the structural "response" of the force-generating area to the closure of switch II happens with a delay of about 400 ns. This observation goes against the strongly coupled mechanism of Fischer and co-workers, and is instead reminiscent of a statistically coupled mechanism (see discussion in 1.3.1).

5.3.5. Baumketner and co-workers (2011-2012)

The most recent model of the recovery stroke is due to Baumketner and co-workers and was outlined in a series of papers (Baumketner 2012a,b; Baumketner and Y. Nesmelov 2011).

The initial study by Baumketner and Y. Nesmelov (2011) used implicit solvent MD simulations to investigate the dynamics of the PR and PPS states. High temperature (350 K) simulations on the 10 ns timescale showed spontaneous closure of switch II without rotation of the converter in PR, supporting the conclusions that switch II closure is the initiating event of the recovery stroke, and that the force-generating region is statistically coupled to the active site. However, no such observation was made at 300 K. An interesting finding is that there is no strict coupling between the formation of the critical salt-bridge and the switch II-ATP hydrogen bond: configurations in which one of these interactions was formed and not the other were sampled. This is in contradiction with the concerted nature of switch II closure in Fischer's model.

Two follow-up papers focused on detailing the mechanism and energetics of the rearrangements in the force-generating region, primarily the formation of the kink in the Relay helix and the rotation of the converter (Baumketner 2012a,b). Using temperature replica-exchange applied on a minimal model comprising the Relay (helix and loop), SH1 helix and converter, Baumketner mapped the free energy landscape for the PR $\rightarrow$ PPS transition restricted to this fragment. In the first study (Baumketner 2012a), the position of the SH1 helix was restrained either in PR (non-tilted) or PPS (tilted). Consistently, the corresponding free energy landscape for the Relay and converter exhibited a minimum in the PR configuration with a PR-like SH1 helix, and in the PPS configuration with a PPS-like SH1 helix. Based on these results, Baumketner argues that 1) the recovery stroke proceeds by a population shift mechanism and 2) the rearrangement of the force-generating region is primarily controlled by the SH1 helix. In the second study (Baumketner 2012b), the same method was applied without restraining the SH1 helix. The minimal model of the force-generating region was found to admit two free energy minima, corresponding to PR and PPS configurations. This demonstrated that the minimal model exhibits recovery-stroke-like transitions, and the analysis of the free energy barriers revealed that in the most likely scenario, the tilting of SH1 precedes the rotation of the converter. Note that the mechanism by which SH1 tilting stabilizes the PPS-configuration of the Relay helix and loop is essentially identical to Fischer's proposal of the aromatic switch rearrangement.

Baumketner's results give a central role to the SH1 helix in controlling the recovery stroke, and it is argued that the SH1 helix represents the main route of allosteric communication between the active site and the converter (as opposed to the Relay helix in Fischer's picture). However, no mechanism for the coupling between the active site and the SH1 helix (*i.e.* how switch II closure drives SH1 helix tilting) is proposed, which prevents Baumketner's model from offering a self-contained picture of the recovery stroke.

5.4. PTS crystal structure and PTS hypothesis for the recovery stroke mechanism

5.4.1. General discussion of the recovery stroke models

Two important areas of disagreement regarding the recovery stroke mechanism emerge when comparing the studies outlined above. First, there is the question of the initiating event. Second, the question as to whether the coupling between the converter and the active site is mechanical or statistical. A model of the recovery stroke should provide an answer to both these questions along with a description of the individual rearrangements and their most likely chronological order.

In several of the previously proposed mechanisms, the recovery stroke is initiated by the closure of switch II onto ATP (Fischer, Woo, Elber, Baumketner). Only the model of Cui and co-workers suggests that the rotation of the converter could be an early event, and that closure of switch II could occur at the end of the transition - but the picture is not entirely clear.

Regarding the coupling, Fischer's model points to a strong, mechanical coupling, but the main method used to obtain the results is intrinsically biased towards such coupling. However, the results by Woo obtained with finite-temperature sampling are mostly in agreement. By contrast, Cui, Elber and Baumketner propose a statistically-coupled model, albeit with different details.

Fischer and co-workers recognize that a "loosely' coupled (*i.e.* statistically coupled) version of their model is also possible; and they acknowledge that the recovery stroke could be initiated by a movement of the converter rather than the closure of switch II (Stefan Fischer, Windshügel, et al. 2005; Koppole, J. C. Smith, and Stefan Fischer 2007). However, they assert that the sequence of events (seesaw motion of the Relay helix, followed by seesaw motion of the SH1 helix and formation of the kink in the Relay helix) should be respected even in the case of statistical coupling, because the aromatic switch controls the sequentiality of the transition (Figure 5.4). If we expand upon their reasoning, this implies that, in the case of statistical coupling, Fischer's model predicts the existence of a structural intermediate in which the converter is partially rotated, switch II is (mostly) closed, and the Relay helix has undergone the seesaw motion, but is not yet kinked. We term this hypothetical conformation **Fischer's putative intermediate** (FPI). To our knowledge, such an intermediate has never been characterized experimentally.

By contrast, Baumketner's model suggests that the movement of the SH1 helix could happen early in the transition, unlike Fischer's conclusions. However, since Baumketner used a minimal model of the motor domain, he is not in a position to discuss the full sequence of events. In fact, since the main body of the motor domain is not completely included, there is no way to account for the seesaw motion of the Relay helix in Baumketner's model. This is the weakness of Baumketner's otherwise ambitious and informative analysis.

To summarize, the models previously proposed for the recovery stroke of myosin disagree on several fundamental points; yet, most are consistent with available experimental data. The solution experiments summarized earlier overall point to the fact that the hydrolysis of ATP happens after the recovery stroke; this may suggest that switch II closure is a late event (otherwise, hydrolysis could happen while the remainder of the motor domain has not yet undergone the transition), but it certainly does not prove it. Indeed, ATP hydrolysis could also be slower than the structural rearrangements associated with the rotation of the converter. For reasons explained above, this is a difficult question to tackle experimentally; in addition, computational studies of the hydrolysis step by the means of quantum methods still seem rather far from achieving quantitative prediction of the detailed reaction mechanism and associated kinetics (Grigorenko et al. 2007; Kiani and S. Fischer 2014, 2013; Li and

Cui 2004). Similarly, although it has been reported that the motor domain seems able to explore the recovery stroke transition dynamically regardless of the bound nucleotide, there is not enough resolution to delineate the interplay between the conformation of the force-generating region and the state of switch II.

Finally, and as will be discussed later on, the available mutational and pharmaceutical data on the recovery stroke mechanism are not sufficient to discriminate between competing models; see (Blanc et al. 2018, Supplementary Text 1) in Appendix C and section 11.1.3.

Thus, both the initiating event of the recovery stroke, and the nature of the coupling between the individual transitions, are still unresolved. We note that the type of coupling and the nature of the initiating event are independent: one may have a statistically coupled switch II-initiated pathway, a strongly-coupled converter-initiated pathway, or inversely[1]. More data, in particular structural, are required to further the debate. In this thesis, we report on the characterization of a novel myosin VI crystal structure (termed Pre-Transition State or PTS), whose features are consistent with a previously un-recognized intermediate along the recovery stroke (Blanc et al. 2018). Assuming that this is indeed the case, the PTS structural and dynamical characteristics, which we investigate, suggest a novel *ratchet-like* model for the recovery stroke. In this model, the transition is not initiated by the closure of switch II, but rather by a thermally-activated movement of the converter; in addition, the coupling between the active site and the converter is statistical, such that the re-priming of the force generating region proceeds up to completion before being stabilized by the closure of switch II and the hydrolysis of ATP. Finally, this model entails a different order of events relative to previous proposals.

5.4.2. Crystallization conditions and resolution of the crystal structure

The protein expression, crystallization, X-ray diffraction and structural refinement which led to the resolution of the PTS crystal structure have been performed by Dr Tatiana Isabet, with assistance by Hannah Benisty and under the supervision of Dr Anne Houdusse, in the Houdusse team at Institut Curie, Paris (Structural Motility Team, UMR 144 - Institut Curie - Paris Sciences et Lettres University). The so-called MD-insert 2 recombinant porcine myosin VI construct was expressed using the baculovirus expression system. The MD-insert 2 (Motor Domain - insert 2) construct contains the motor domain of myosin VI, the converter, and is truncated after residue I789, *i.e.* at the end of the first helix of insert 2. A purification Flag tag was appended to the N-terminus.

Crystals of the MD-insert 2 construct were obtained with 2 mM MgADP.BeF_x with the hanging drop vapor diffusion method. Spontaneous nucleation of small crystals was observed at 277 K for equal amounts of reservoir solution (7% PEG 8000, 50 mM Tris, 1 mM TCEP, 15% glycerol) and protein stock solution (10 mg mL^{-1} of protein in 10 mM Hepes, pH 7.5, 50 mM NaCl, 1 mM TCEP, 1 mM NaN_3, 1 mM EDTA). Usable crystals for X-ray diffraction were then grown by seeding and cryocooled; X-ray data collection was performed at the European Synchrotron Radiation Facility (ESRF, Grenoble, France). Diffraction patterns were processed with XDS (Kabsch 2010). The initial structural model was obtained by molecular replacement from the myosin VI PPS structure (2V26) with Phaser (McCoy et al. 2007). Then, refinement was performed at 2.2 Å resolution with Coot (Emsley and Cowtan 2004) and BUSTER (G. Bricogne et al. 2011). The final structure revealed a previously unseen conformation of the motor domain; for reasons that will become apparent later, this new structure was termed the Pre-Transition State or PTS. It is deposited in the Protein Data Bank under the code **5O2L**.

1. One should note, however, that the definition of an initiating event for a strongly coupled pathway is more ambiguous, since everything moves together at the same time.

5.4.3. Overall description of the PTS crystal structure

The PTS crystal structure is showed on Figure 5.5. It exhibits the following major features:

- Open switch II; the distance between G459 (homologous to G457 of *Dictyostelium discoideum* myosin II) and the fluoride atom of BeF_x (7 Å) is too large for hydrogen-bonding, and the critical salt-bridge (R205-E461 in myosin VI) is not formed.
- The Relay-SH1 elements are in a post-recovery configuration; most notably, the SH1 helix is tilted and the Relay helix exhibits a kink (Figure 5.7)
- The converter is partially re-primed, in an intermediate position between the PR and PPS positions (Figure 5.8); moreover, it is in the canonical R-fold.

In addition, the cleft between the L50 and U50 subdomains is wide-open. Further comparison with the end-points of the recovery stroke (PR and PPS) shows how the PTS structure could be consistent with a previously unrecognized intermediate along the recovery stroke. Indeed, it exhibits both PR-like (active site) and PPS-like (Relay-SH1 elements) characteristics, along with a clearly intermediate converter position. In the active site, the conformation of switch II is close to PR (Figure 5.6). Also, super-imposition onto the structurally invariant N-terminal subdomain (residues 50-172 and 662-670) reveals that no rigid-body motion (*i.e.* , seesaw) of the Relay helix has occurred in PTS, just like in PR. By contrast, in PPS, the seesaw motion is clearly seen as the inward shift of the Relay helix towards the inside of the active site (Figures 5.6 and 5.9). The internal conformation of the Relay and SH1 helices is very close between PTS and PPS (0.75 Å CA-RMSD), despite the lack of seesaw motion of the Relay helix (Figure 5.7). Notably, the SH1 helix is mostly tilted in PTS.

5.4.4. The PTS suggests a novel mechanism for the recovery stroke

The most striking feature of the PTS structure is the co-existence of a significantly (but not completely) re-primed converter along with an open, ATPase-inactive nucleotide binding site. This contradicts both claims by Fischer and co-workers, namely that there exists a strong coupling between the closure of the nucleotide-binding site and the movement of the converter, and that the closure of the nucleotide-binding site initiates the recovery stroke (Figure 5.10).

Indeed, strong coupling would predict that the rearrangements of the converter and the active site are at all time tightly coordinated; thus, if it were true, the PTS structure should exhibit a partially closed switch II on account of the observed partially moved converter. Rather, the PTS structure points to a statistical coupling, because it suggests that large movements of the converter can take place without mechanically closing switch II in response.

Regarding the initiating event, a switch II-initiated, statistically coupled pathway would predict an intermediate with a (mostly) closed switch II and a down converter in the PR position. Conversely, the PTS structure points to a converter-initiated, statistically coupled pathway. We can now formulate the **PTS hypothesis**:

The PTS structure is representative of an on-pathway intermediate along the recovery stroke of myosin.

The final state of the recovery stroke (PPS) was previously termed the ”Transition State”; as such, we named the new structure Pre-Transition State to highlight its putative status as an intermediate preceding the PPS.

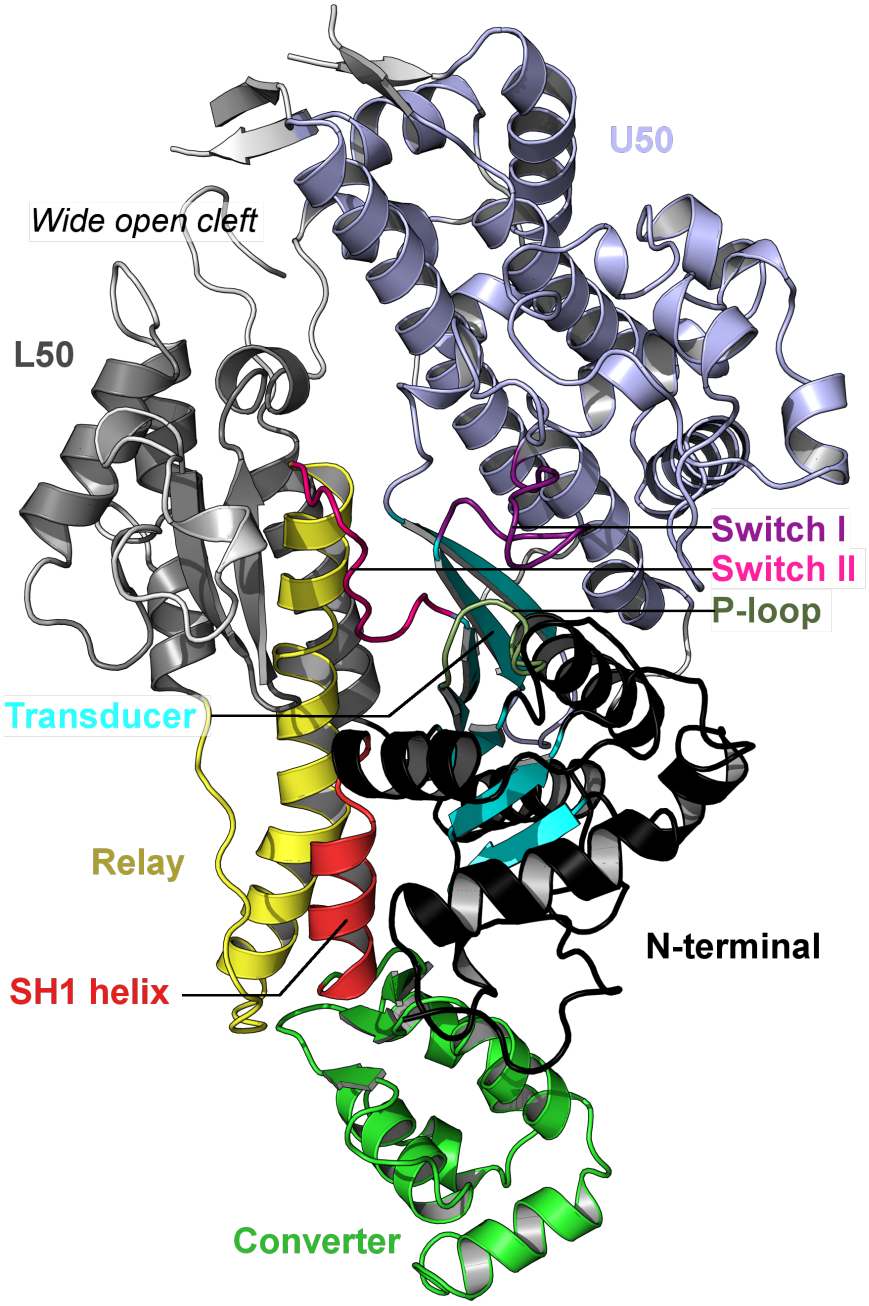


Figure 5.5.: The PTS crystal structure. For clarity, the SH3 domain in N-terminus is not represented. This figure is adapted from (Blanc et al. 2018).

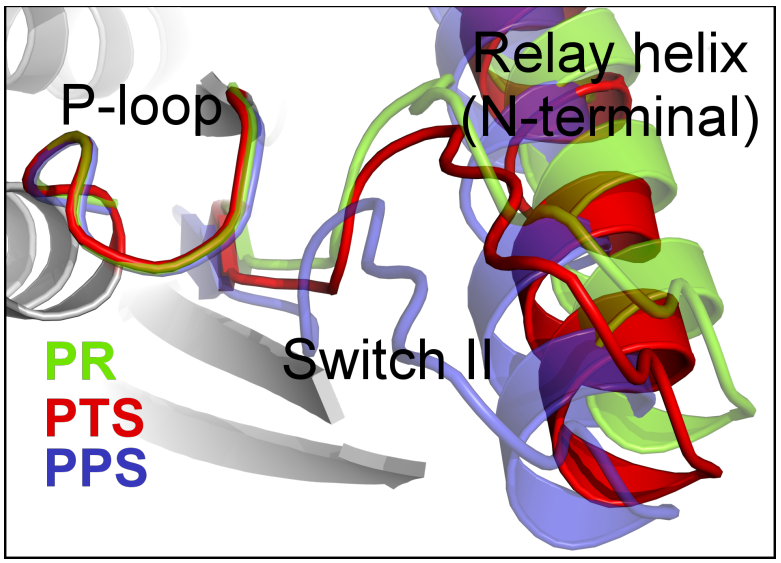


Figure 5.6.: Comparison of switch II conformation and position between PR, PTS and PPS. This figure is adapted from (Blanc et al. 2018).

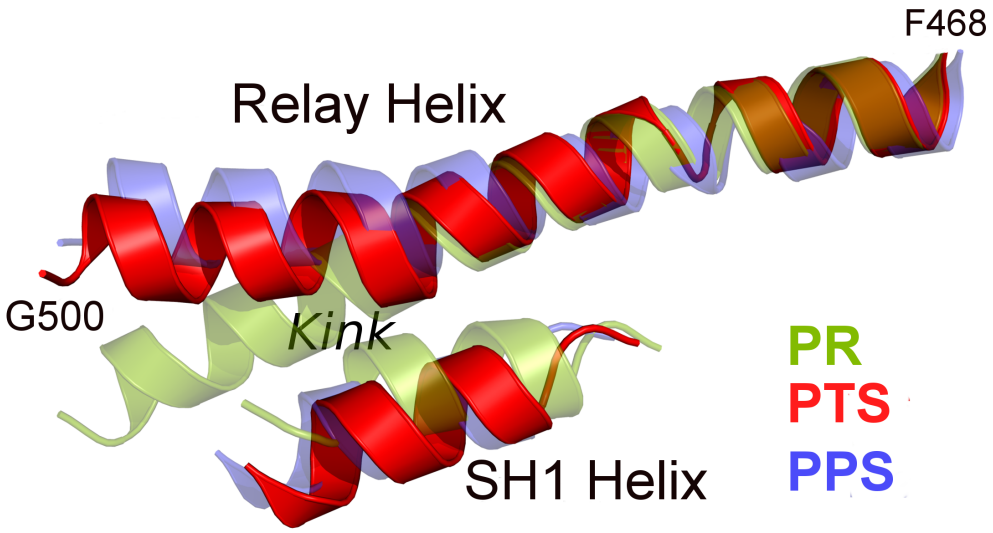


Figure 5.7.: Comparison of the internal conformation of the Relay-SH1 elements between PR, PTS and PPS. In PTS, the Relay helix is kinked and the SH1 is tilted, taking a conformation very close to PPS. This figure is adapted from (Blanc et al. 2018).

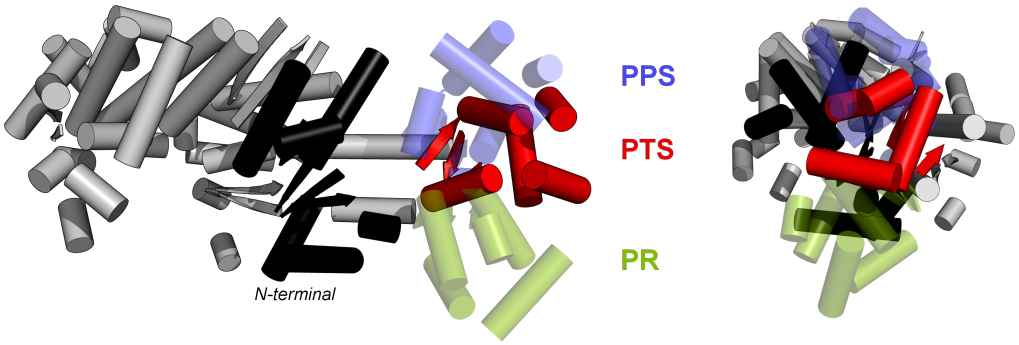


Figure 5.8.: Comparison of the converter position relative to the main body of the motor domain between PR, PTS and PPS. This figure is adapted from (Blanc et al. 2018).

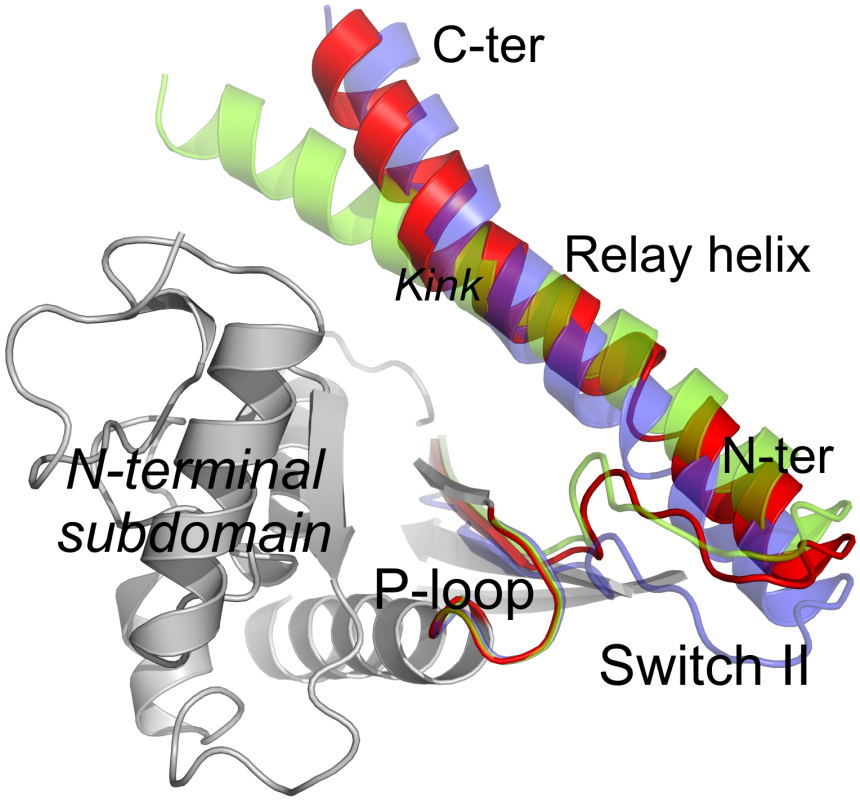
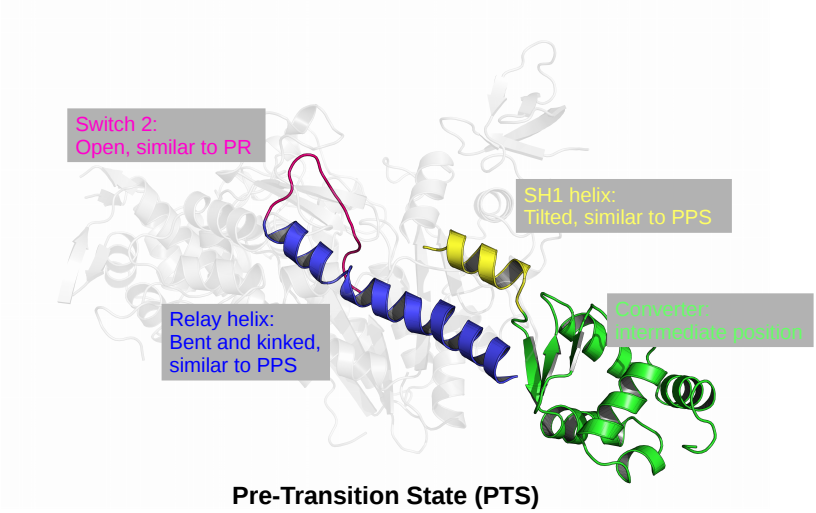


Figure 5.9.: Seesaw motion of the Relay helix from PR/PTS to PPS (*i.e.* inward motion of the N-terminal region of the helix upon going to PPS).



1

Figure 5.10.: The PTS structure contradicts the strong coupling model.

Most of the present thesis is devoted to the analysis of the PTS structure by molecular simulations and to the study of the novel recovery stroke mechanistic scenario emerging from this analysis. Using computational approaches, we explore the implications of the PTS hypothesis for our understanding of chemo-mechanical transduction in myosin, outline the mechanistic scenario emerging from it, and eventually propose a strategy to directly test the hypothesis.

6. Conformational dynamics of the motor domain characterized by unbiased simulations

Summary In this chapter, we first present the protocol we followed to prepare structural models of the myosin motor domain for MD simulation, along with the simulation parameters. Then, we detail several non-trivial geometrical observables used to describe some aspects of the recovery stroke transition. Finally, we report on the behaviour of the motor domain in the PR, PTS and PPS states as characterized by unbiased MD. These simulations are part of our publication (Blanc et al. 2018).

6.1. Simulation protocol

6.1.1. Structure preparation

Solvated structural models of the myosin VI motor domain in the PR, PTS and PPS conformations were prepared for MD simulation using the following protocol. Missing residues[1] (in particular flexible loops) were modelled using template-based homology modelling (when a template was available) or *de novo* reconstruction using MODELLER (Fiser and Šali 2003; Webb and Sali 2002).

We selected the best model among 10 based on the minimum DOPE (Discrete Optimized Potential Energy) score. Special care was taken for loop reconstruction, since we observed that MODELLER can produce knotted loop conformations with a low DOPE score. Nucleotide analogues in the binding pocket were replaced by ATP or ADP.Pi depending on the case. In the case of ADP.Pi, the doubly protonated form $H_2PO_4^-$ was used in accordance with recent quantum investigations of the hydrolysis mechanism (Kiani and S. Fischer 2014).

Each model was submitted to the MolProbity web-server (Davis et al. 2007) to optimize the rotameric states of side chains so as to avoid clashes, followed by visual inspection. After this, the model was processed by CHARMM (B. R. Brooks et al. 2009, versions c38b1 or c40b1) to add hydrogen atoms and relax the geometry of the nucleotide using a short energy minimization in vacuum.

These models were then submitted to Poisson-Boltzmann/Monte Carlo calculations so as to determine the most probable protonation states of histidines at 300 K and neutral pH. A multisite titration approach was used (Bashford and Karplus 1991), in which the solvent and protein interior are treated as continua of respective relative dielectric constants 80.0 and 4.0. Ions in the solution were accounted for by a Boltzmann-distributed continuous charge density corresponding to 150 mM at 300 K. The electrostatic potential was determined by numerical resolution of the Poisson-Boltzmann equation, using the Adaptive Poisson-Boltzmann Solver (APBS, Baker et al. 2001) through the tAPBS front-end. Finally, Monte Carlo sampling was used to evaluate the protonation probabilities with the Karlsberg2 program (Kieseritzky and Knapp 2007; Rabenstein and Knapp 2001).

Since protonation states are not treated dynamically by classical force fields, having different protonation states between the myosin conformations would introduce imppassable barriers. This is to

1. Missing residues for PR: 1-3; 353-367; 394-409; 623-638. Missing residues for PTS: 1-4; 356-360; 397-405; 624-631. Missing residues for PPS: 1-4; 174-180; 396-404; 622-637.

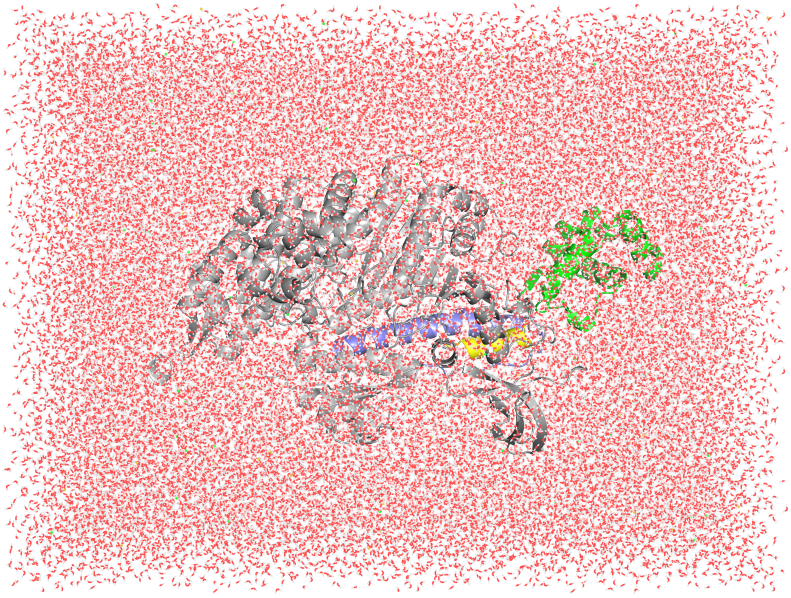


Figure 6.1.: Explicitly solvated structural model of the motor domain of myosin VI in the simulation box.

be avoided since we eventually want to study conformational transitions; as such, when discrepancies arose as to the protonation states between myosin conformers, the PTS ones were (arbitrarily) retained.

The final protonated model were again processed with CHARMM to add the crystallographic water molecules. In the case of the PPS+ATP structure, which was modelled from the PPS+ADP.VO_4 crystal structure, an additional water molecule was added corresponding to the position of the outermost vanadate oxygen atom with respect to ADP. Then, each model was solvated by TIP3P water molecules in a 144 Å $\times$ 108 Å $\times$ 96 Å orthorhombic box using the *solvate* plugin of VMD (Humphrey, Dalke, and Schulten 1996, versions 1.9.2 and 1.9.3). Sodium and Chloride ions were added so as to ensure electroneutrality of the system and a total salt concentration of 150 mM with the *ionize* plugin of VMD.

6.1.2. Energy model

Energetics were modelled using the CHARMM36 classical force-field (Huang and Alexander D. MacKerell 2013; A. D. MacKerell et al. 1998) and the CMAP correction (Mackerell, Feig, and C. L. Brooks 2004). Dispersion interactions were treated using a Lennard-Jones potential cut off at 12 Å with a switching function starting at 10 Å. Short range electrostatics were also cut off at 12 Å. Long range electrostatics were treated using the PME method, with a 6^{th} order spline interpolation and a 1 Å-spaced grid.

After solvation and addition of ions, each system was energy-minimized for 5000 steps in NAMD with absolute positional harmonic restraints of force constant 10 kcal/mol/Å^2 on heavy protein atoms and 5 kcal/mol/Å^2 on the oxygen atoms of crystallographic water molecules (Phillips et al. 2005).

6.1.3. Dynamics parameters and preparation

Minimized systems were submitted to 1 ns of heating simulation, with active restraints, up to 300 K. During heating, temperature control was achieved using the Andersen thermostat and the box volume was kept constant. Heating was followed by a 2 ns equilibration run for which temperature control

was switched to Langevin dynamics (with a 1 ps^{-1} friction coefficient) and pressure control was turned on with a Berendsen barostat (time constant 400 ps). During equilibration, harmonic restraints were smoothly turned down following a cubic scaling function to be nullified at the end of the run, see equation 6.1.

$$k(n) = k_0 \left(1 - \frac{n}{n_{total}}\right)^3 \tag{6.1}$$

where n refers to the current MD step, n_{total} is the total number of MD steps, k_0 is the harmonic force constant at the beginning of equilibration and $k(n)$ is the scaled force constant at step n. This cubic scaling function was initially implemented in a custom NAMD build by Dr. Nicolas Calimet.

Production runs were then immediately launched with no absolute positional restraints, a 0.1 ps^{-1} Langevin friction coefficient, and otherwise the same parameters. Due to the use of an anisotropic box, the protein was kept aligned with the box using a harmonic restraint on its orientation quaternion as allowed by the colvars module in NAMD (Fiorin, Klein, and Hénin 2013). The orientation quaternion was computed with respect to the CA atoms of the (reconstructed) crystal structure and a 1×10^4 kcal mol^{-1} force constant was used; this was done using the *colvars* module (Fiorin, Klein, and Hénin 2013). Short NVE simulations were used to make sure that this force constant was low enough not to compromise energy conservation in the limit of the integrator accuracy.

Moreover, in the most recent simulations, this orientational restraint was complemented (purely for commodity) by a positional restraint to keep the protein center of geometry at the center of the box. Note that since both the orientational and positional biases do not affect the internal conformational dynamics of the protein, we will refer to MD simulations in which no other bias than these two are active as *unbiased* simulations.

Equations of motion were integrated using the BBK integrator and a multiple time-step scheme, allowing to use different time-steps for the various terms of the total force. The basal time step was 2 fs (allowed by the use of SHAKE to constrain covalent bonds involving hydrogen atoms, (Ryckaert, Ciccotti, and Berendsen 1977)); bonded and short-range interactions were computed every time step, and full electrostatics every 2 time steps. We note that this differs from the popular 1/2/4 scheme; test runs showed that our system was unstable when submitted to this dynamics.

Unless otherwise stated the above simulation parameters were used for all simulations reported in this thesis. Finally, structural and trajectory data were analyzed using Pymol (pymol.org), VMD (Humphrey, Dalke, and Schulten 1996), Wordom (M. Seeber et al. 2007; Michele Seeber et al. 2011) and in-house Python scripts relying heavily on NumPy (Walt, Colbert, and Varoquaux 2011), SciPy (Jones, Oliphant, and Peterson 2001) along with scikit-learn (Pedregosa et al. n.d.) and pandas (McKinney 2010). Interactive data analysis was performed with IPython (Perez and Granger 2007), and plotting with the matplotlib graphical library (Hunter 2007).

6.2. Collective variables to analyze the conformation and dynamics of myosin

The motor domain of myosin includes nearly 800 residues organized in a complex tertiary structure which rearranges dramatically during the recovery stroke. Although one may carry on a detailed structural comparison between a limited amount of conformations (*e.g.* the crystal structures of PR and PPS), it is not practical to do so when analyzing the thousands of frames generated by a Molecular Dynamics simulation. Instead, geometric observables (collective variables) must be designed which

capture and summarize the conformation and dynamics of important subdomains. We now introduce some of them; others will be discussed later on.

6.2.1. Position of the converter by projection upon reference axes

Characterizing the position of the converter with respect to the main body of the motor domain is necessary to describe the recovery stroke. Although it is easy to visualize the difference in converter position between PR and PPS, providing a quantitative description proved surprisingly difficult due to the complexity of the converter movement, especially when the PTS state is included in the picture. Previous investigators used angle-based variables, and notably Baumketner introduced a pair of dihedral angles which were acceptable order parameters to distinguish between PR and PPS in Dd Myo2 (Baumketner 2012a,b). However, the same dihedrals were not successful in completely discriminating between PTS and PPS regarding the position of the converter (data not shown). Thus, we developed an original set of observables to characterize the position of the converter.

We define observables X', Y' and Z' as the projections of the center of geometry of the CA atoms of the converter onto the principal axes of the main body of the motor domain, see Figure 6.2. The idea was to express the position of the converter in a local frame, co-moving with the entire protein. The use of cartesian coordinates appeared the most practical although cylindrical coordinates can also be used, to isolate the rotational component of the converter's motion.

The computation of the principal axes turned out to be somewhat problematic. Our initial idea was to recompute the basis-set on-the-fly for each conformation using principal component analysis, however eigenvalue exchange was found to be very frequent, leading to extremely noisy time series with jumps associated to basis vector permutations that were very cumbersome to de-convolute. Thus, we resorted to a "reference axes" approach in which the basis set was computed only once on a reference myosin structure, after which each myosin conformation was aligned with this reference basis using least-square fitting. The reference structure was taken as the averaged coordinates of the main body atoms between the myosin VI PR, PTS and PPS crystal structures. In hindsight, this formulation - originally implemented using a Python program- turned out to be very practical since its subsequent implementation directly in the *colvars* module of NAMD was possible, using *distanceZ* components (Fiorin, Klein, and Hénin 2013). This later opened the way to biased simulations using these observables as collective variables to drive the movement of the converter.

6.2.2. Conformation of the Relay-SH1 elements

The extensive conformational rearrangement of the Relay-SH1 elements during the recovery stroke involves the formation of a kink in the Relay helix, an angular displacement of the C-terminal tip of this helix, and an in-place tilting movement of the SH1 helix. Several collective variables were introduced to describe this conformational change, either in a global or local manner.

6.2.2.1. $\Delta RMSD$ of the Relay and SH1 helices

The $\Delta RMSD_{R/SH1}$ between the PR and PTS conformations of the Relay-SH1 elements was used to describe the rearrangement in a global manner. Given its flexibility, the Relay loop was excluded from the definition, and the observable is defined by reference to the CA atoms of residues 468-499 (Relay helix) and 693-704 (SH1 helix). $\Delta RMSD_{R/SH1} = RMSD(\cdot, PTS) - RMSD(\cdot, PR)$ takes values close to 2.06 Å for PR-like configurations of the Relay and SH1 helices, and -2.06 Å for PTS/PPS-like configurations.

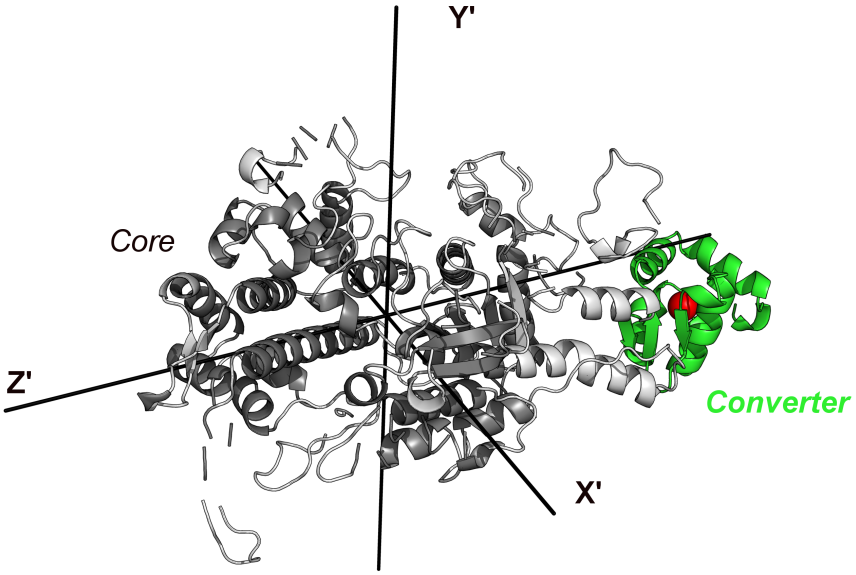


Figure 6.2.: Illustration of the reference principal axes used to compute the X',Y' and Z' components of the converter (in green) position with respect to the main body of the motor domain. The "core" residues (in dark gray) are the ones used for the calculation of the axes. The red sphere materializes the position of the converter center of geometry.

6.2.2.2. Backbone $\Delta RMSD$ of the kink region

To isolate the formation of the kink in the Relay helix from other rearrangements in the Relay-SH1 elements, we introduced another, local $\Delta RMSD_{kink}$ defined with the C, CA, O, N atoms (backbone atoms) of residues 485 to 493. This corresponds to the region which rearranges its backbone hydrogen-bonding pattern upon the formation of the kink, going from 1.4 Å for a straight Relay helix to -1.4 Å for a kinked Relay helix.

6.2.2.3. Angular description of the rearrangement

The rearrangement of the Relay-SH1 elements involves bending, tilting and more generally complex re-orientation of helical segments with respect to one another. Although the global $\Delta RMSD$ is a very useful "summarizing" variable (*i.e.* an order parameter), a finer description of the transition requires separate observables to account for the individual motions of the SH1 helix and the C-terminal fragment of the Relay helix. We used orientation quaternions to describe these motions (Fiorin, Klein, and Hénin 2013), see Figure 6.3.

Tilting angle of the SH1 helix To describe the tilting of the SH1 helix, we used an orientation quaternion with respect to the PR crystal structure, computed from the CA atomic coordinates of the SH1 helix (693-704) and expressed in the reference frame formed by a large subset of the main body residues CA (50-650). This set of reference residues was chosen by trial and error. The associated orientation angle θ_{SH1} has values 0° in PR (by construction), 21.6° in PTS, and 31.4° in PPS crystal structures.

Kink angle of the Relay helix Similarly, the kink angle of the Relay helix is described as the orientation angle made between the C-terminal fragment of the Relay helix (residues 490 to 499) and its position

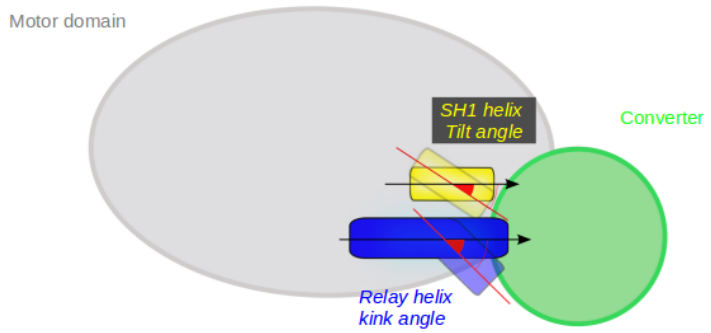


Figure 6.3.: Schematic definition of the Relay-SH1 angular descriptors θ_{RH} and θ_{SH1}.

in the PR crystal structure. This angle measures the re-orientation of the tip of the Relay helix as it bends and kinks during the recovery stroke. Low values are associated with a non-kinked, weakly bent helix in a pre-recovery configuration. Higher values correspond to a kinked and bent helix, as seen in PTS and even more in PPS. This angle θ_{RH} takes values 0° in PR, 53.1° in PTS and 61.9° in PPS crystal structures.

6.2.3. Active site interactions

The opening state of switch II in the active site was monitored by considering a pair of distances describing the formation of the two important interactions for switch II closure, namely the critical salt-bridge and the switch II- ATP hydrogen bond. The critical salt-bridge is described by the distance d_1 between atoms R205CZ and E461CD; the switch II-ATP hydrogen bond is described by the distance d_γ between the atom G459N on switch II, and the atom ATPO1G on the γ-phosphate of ATP (or of Pi in the case of ADP.Pi). Note that because of the possibility of a rotation of the γ-phosphate, a more proper definition of this observable would have been in terms of the minimal distance between all three hydrogen atoms of the γ-phosphate and G459N. However, such a rotation was never observed in MD simulations with ATP and the more convenient above definition of d_γ was retained in these cases. By contrast, for the PPS+ADP.Pi simulations, d_γ is defined as the distance between G459N and the closest oxygen atom of Pi, since phosphate rotation is observed. These two observables are illustrated on Figure 8.1, Chapter 8.

6.3. Dynamics of the motor domain in unbiased simulations

We now report on the results of unbiased Molecular Dynamics simulations of the motor domain of myosin VI. The simulations discussed in this section are summarized in Table 6.1. For a given conformational state and ligand, each production simulation was initiated from independent equilibration

Simulation	Length (ns)	Starting conformation
PR+ATP (1)	101	2VAS+ATP
PR+ATP (2)	100	2VAS+ATP
PR+ATP (3)	300	2VAS+ATP
PTS+ATP (1)	306	5O2L+ATP
PTS+ATP (2)	100	5O2L+ATP
PTS+ATP (3)	100	5O2L+ATP
PPS+ATP (1)	100	2V26+ATP
PPS+ATP (2)	100	2V26+ATP
PPS+ATP (3)	100	2V26+ATP
PPS+ADP.Pi (1)	101	2V26+ADP.Pi
PPS+ADP.Pi (2)	100	2V26+ADP.Pi

Table 6.1.: List of unbiased MD simulations of the myosin VI motor domain.

runs, but these equilibration runs were launched from the same heating simulation. Since we are using Langevin dynamics (which includes a random force) and different random generator seeds, it is found that the various simulations initiated from the same heated structure exhibit different trajectories and can be treated as independent replicates.

6.3.1. Dynamical features of the Post-Rigor state

In two out of three simulations (simulations PR+ATP (1) and (2), see Table 6.1), the PR state exhibits remarkable conformational stability, see Figure 6.4. The converter and Relay-SH1 elements observables fluctuate around pre-recovery values, in a rather narrow conformational basin as evidenced by the small amplitude of the fluctuations on the 100 ns timescale. Interestingly in the third simulation replica (PR+ATP (3)), a spontaneous uncoupling of the converter from the N-terminal domain is captured at $t = 50$ ns (see Figure 6.4, the "PR: uncoupled converter" local density maximum, and Figure 6.5). This uncoupling happens without formation of a kink in the Relay helix (despite a certain amount of bending of the helix) and is not observed to revert after extending the simulation to 300 ns, see Figure 6.5. Interestingly, the time-series of the θ_{SH1} angle shows that a tilting of the SH1 helix precedes the movement of the converter, suggesting that the former may actually initiate the latter, see Figure 6.5.

In the active site, the two important interactions for the closure of switch II, namely the critical salt-bridge and the switch II-ATP hydrogen bond, remain broken similar to what is seen in the crystal structure (Figure 6.5). In particular, this remains the case after the uncoupling of the converter captured in simulation 3. This observation provides some circumstantial support to 1) the lack of strong coupling between the active site and the converter and 2) the idea that the recovery stroke transition is not initiated by the closure of switch II, but rather by spontaneous rearrangements in the force-generating regions. Moreover, we note that this observation is obtained in total independence from the PTS structure.

6.3.2. Dynamical features of the Pre-Powerstroke state

Three runs of PPS+ATP and two runs of PPS+ADP.Pi were performed. In all but one simulation, the converter is observed to undergo a relaxation towards a position which is slightly more re-primed than

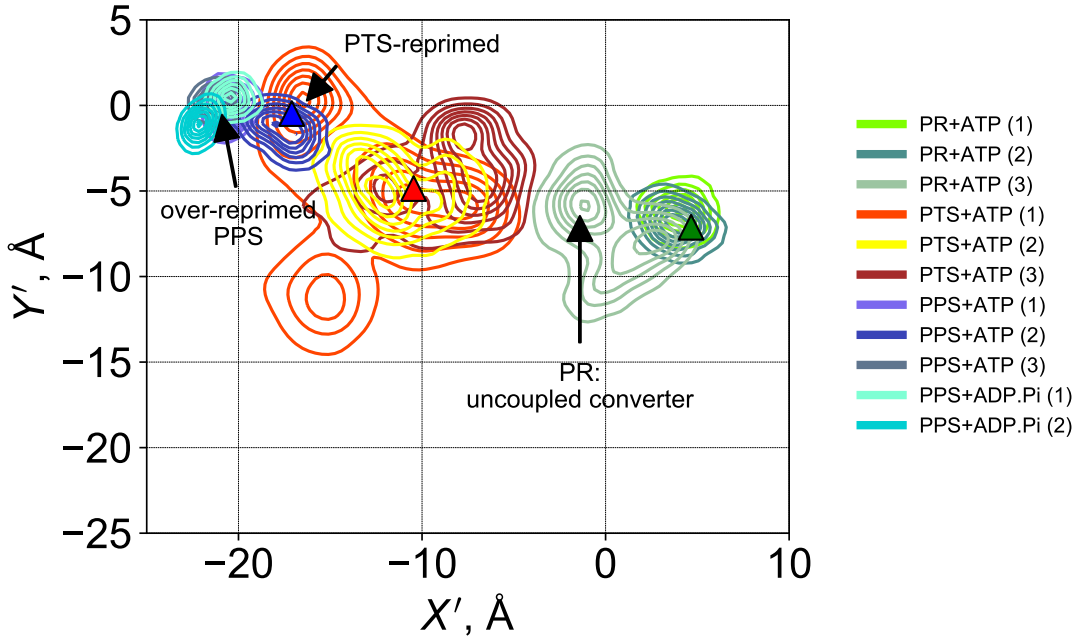


Figure 6.4.: Statistical distributions of the converter position characterized by X',Y' in all unbiased simulations. The triangles represent crystallographic values (green, PR; red, PTS; blue, PPS). Some of the apparent metastable states emerging from the simulations are highlighted.

in the crystal structure, and the positional distribution does not overlap with the crystallographic position, see Figure 6.4 (PPS "over-reprimed" basin). However in simulation PPS+ATP (2), the converter explores positions compatible with the crystallographic one (Figures 6.4 and 6.7). The cause of the converter relaxation is unclear; we note however that the extent of converter positional fluctuations is similar between PPS simulations, and is also comparable to PR simulations.

With ATP, a striking and consistent re-opening of switch II (breaking of the switch II-ATP hydrogen bond; also, breaking of the critical salt-bridge in two out of three simulations) is observed whereas switch II is closed in the starting conformation (Figure 6.7). A more detailed study of the re-opening of switch II will be given later, in the context of the PTS → PPS transition. Also, in one simulation (PPS+ATP (1)), a re-opening of the L50/U50 cleft is observed, see Figure 6.6.

By contrast, when ADP.Pi is present, the cleft remains closed. Regarding switch II, the results are inconsistent, with one simulation in which the critical salt-bridge spends most of the time broken (but reforms towards the end of the simulation, PPS+ADP.Pi (1)) and one where it is maintained all along (PPS+ADP.Pi (2)), Figure 6.7. The hydrogen bond between switch II and the γ-phosphate remains formed, but the γ-phosphate is now free to relax towards switch II, relieving the strain which may be associated with a closed switch II onto ATP and would account for switch II re-opening when ATP is present. Overall, the MD results point to a surprising instability of the PPS-like configuration of the active site when ATP is present, which is consistent with previous experimental results suggesting that the recovery stroke is reversible with ATP. The re-opening of the cleft observed in one ATP-bound simulation, but not with ADP.Pi, is also consistent with this idea, and suggests that ADP.Pi indeed contribute to stabilizing the PPS state. However, the lack of statistics makes it premature to draw

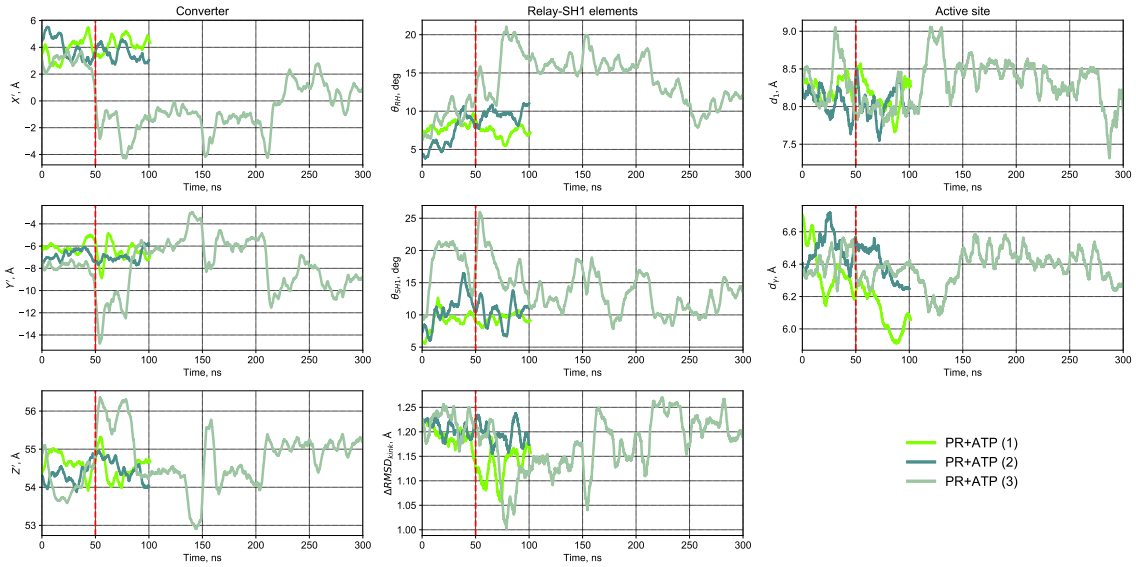


Figure 6.5.: Evolution of the recovery stroke observables in unbiased simulations of the PR state. The dotted red line materializes the beginning of the large converter movement captured in simulation PR+ATP (3). For clarity the 5 ns running average is shown.

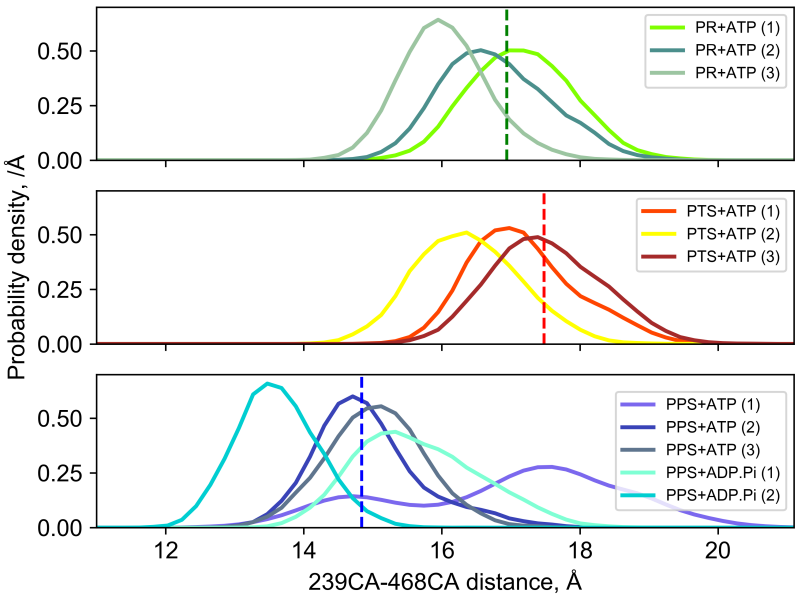


Figure 6.6.: Statistical distributions of a typical cleft distance during unbiased MD. This distance measures the distance between the L50 and U50 subdomain. The results show that the cleft is wide-open in PR and PTS, and closed (partially) for all but one PPS simulations. By contrast, in PPS+ATP (1) a rotation movement of the L50 leads to the re-opening of the cleft.

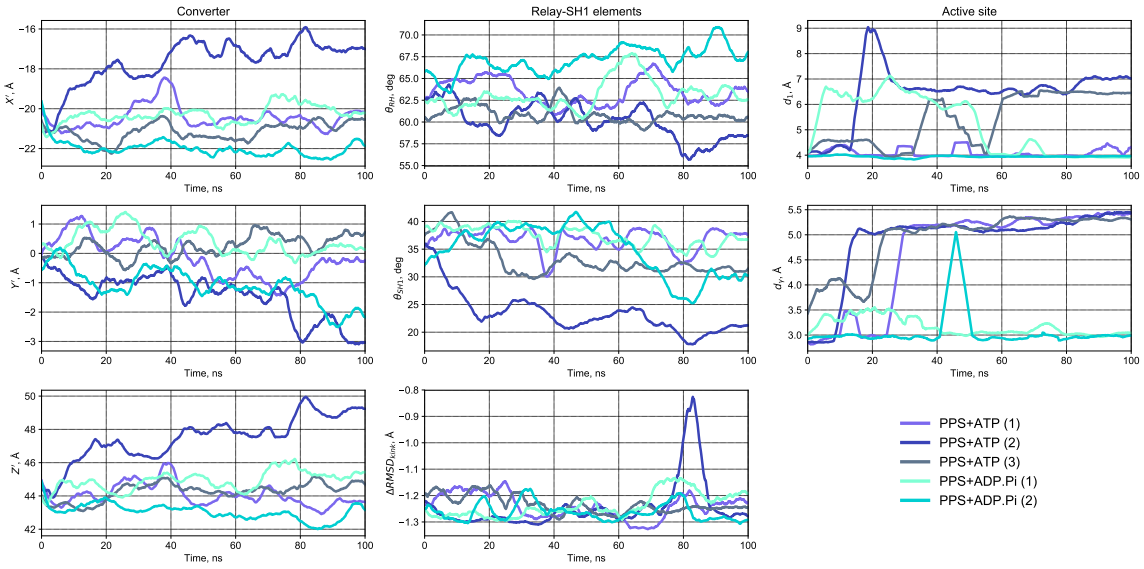


Figure 6.7.: Evolution of the recovery stroke observables in unbiased simulations of the PPS state. For clarity the 5 ns running average is shown.

definitive conclusions about this point based on the present simulations.

6.3.3. Dynamical features of the Pre-Transition state

As compared to PR and PPS, PTS simulations consistently exhibit extensive positional fluctuations of the converter, as shown on Figure 6.4. This is explained by the fact that the converter is essentially *uncoupled* from the N-terminal subdomain in PTS, *i.e.* it is connected to the motor domain only through the Relay and SH1 helices. Statistical distributions (Figure 6.4) show that all three PTS simulations, despite being in overlap, explore various regions of the (X', Y') space, and identify several potential metastable states. Interestingly, in the PTS+ATP (1) simulation, a spontaneous movement of the converter towards a metastable basin in overlap with PPS-like converter positions is sampled, see Figures 6.4 and 6.8, "PTS-reprimed" basin. This partial re-priming of the converter, which starts around t=50 ns, is eventually reversed as the conveter returns to its initial basin. Thus, this simulation demonstrates that PTS can sample spontaneous, extensive, reversible transitions towards a configuration in which it is closer to PPS and stabilized on the N-terminal domain through new contacts, see Figure 6.9. Whether this "PTS-reprimed" state observed in simulation is functionally relevant is unclear and would require more investigation. More generally, it is interesting to remark that the positional distribution of the converter in PTS does occupy an intermediate position between these of PR and PPS, as would be expected for an intermediate (Figure 6.4). Notably, if this observation is certainly not sufficient to rule-out PTS being off-pathway, it seems to us that it justifies the proposal that the PTS structure belongs to a configurational basin, or *state*, distinct from PR and PPS. This motivates our past (and future) use of the term "PTS state".

Interestingly, the large movements of the converter sampled in PTS are *not* accompanied by switch II closure events (even partial), supporting the absence of strong coupling between converter and active site as already suggested by the crystal structure, see Figure 6.8. By contrast, the converter dynamics seems coupled to that of the Relay-SH1 elements, notably since an evolution of θ_{RH} and θ_{SH1} towards PPS-compatible values is observed when the converter swings to the PTS-reprimed state (Figure 6.8).

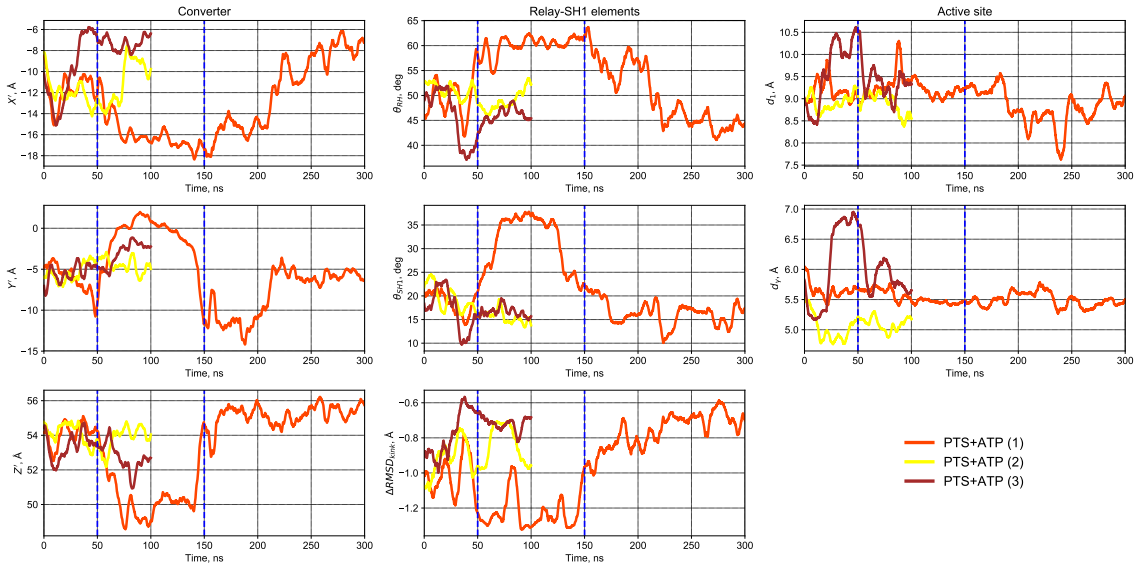


Figure 6.8.: Evolution of the recovery stroke observables in unbiased simulations of the PTS state. The dotted blue lines materialize the beginning and the end of the converter movement to the ”PTS-reprimed” basin sampled in simulation PTS+ATP (1). For clarity the 5 ns running average is shown.

6.3.4. Conclusion

The unbiased simulations reveal interesting points which are consistent with the PTS hypothesis, such as the observation of partial transitions along the PR → PTS and PTS → PPS directions, consisting mostly of converter movements but without effect on the dynamics of switch II. The positions occupied by the converter in PTS (described by the X', Y' plane) are indeed in-between PR and PPS ones, as would be expected for an intermediate. But, none of that is enough to establish (or refute) the PTS hypothesis; in the subsequent chapters we will develop more advanced strategies, using a range of biased and enhanced simulations, to aim for a more quantitative approach to the testing of the PTS hypothesis. In this context, the present unbiased simulations, which provide a characterization of the PR, PTS and PPS states beyond the crystal structure, will be invaluable as reference points to analyze configurations produced by other simulation approaches.

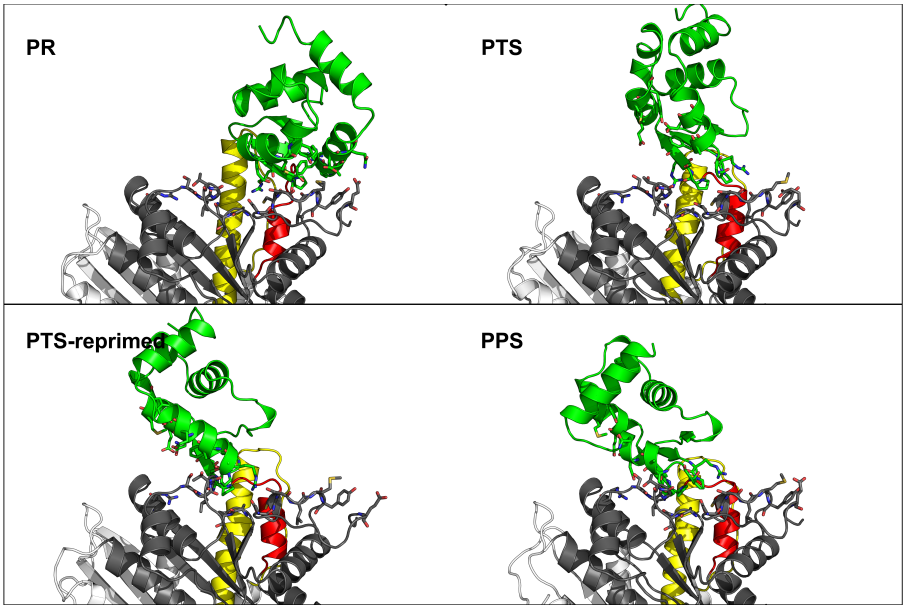


Figure 6.9.: Comparison of the converter/N-terminal interface between PR, PTS, PPS and the PTS-reprimed states.

7. Exploration of the conformational landscape of the recovery stroke by enhanced sampling

Summary We used aMD and GaMD simulations to explore the conformational landscape of the recovery stroke. In support of the PTS hypothesis, aMD reveals overlap in converter positions between PR/PTS and PTS/PPS, but nearly no PR/PPS overlap. Also, a reverse transition from PPS to PTS is captured. However, the large boost used in aMD simulations precludes their quantitative analysis (re-weighting). Thus, we turn to GaMD simulations, but these fail in exhibiting significant enhanced sampling. These results are unpublished.

Unbiased simulations reported in the previous Chapter have revealed a great deal about the dynamical characteristics of the motor domain, and the results regarding the PTS are *consistent* with the hypothesis that it is an on-pathway intermediate. However, these simulations only provide circumstantial support for this hypothesis, most notably because no complete, spontaneous transition is captured either from PR to PTS, or from PTS to PPS.

Of course, if PTS were indeed an intermediate, we would expect the transition to take place on the 1 ms timescale, out of reach of unbiased MD. Thus, to capture a spontaneous transition in a reasonable simulation time, an enhanced sampling technique should be used. This chapter reports on the use of Accelerated and Gaussian Accelerated MD simulations as an attempt to demonstrate the full transition. Since its introduction, aMD has been used several times for this purpose; notably, the technique was employed to probe the functional transitions of small GTPases (Grant, Gorfe, and McCammon 2009), and later of the kinesin molecular motor (Scarabelli and Grant 2013).

7.1. Accelerated molecular dynamics simulations

7.1.1. Set up

Accelerated MD (aMD) simulations of the motor domain were run using the dual-boost approach with NAMD (Wang et al. 2011). For each state of the motor domain, potential energy statistics were collected from the last 25 ns of unbiased MD trajectories (reported in the previous chapter), according to the procedure described in section 3.2.4.1. In the case of the PR state, we used the PR+ATP (3) unbiased trajectory, *i.e.* the simulation which captured a spontaneous uncoupling of the converter - which corresponds to a higher potential energy state (data not shown). We made this choice so that the bias applied in the PR aMD simulations would be as high as possible, while still being estimated from an unbiased trajectory, in the hope to efficiently enhance the sampling. Two independent Accelerated MD simulations were run per state of the motor domain for about 100 ns and projected on the observables characterizing the recovery stroke (Table 7.1).

Simulation	Length (ns)	Starting conformation
PR aMD (1)	200	2VAS+ATP
PR aMD (2)	100	2VAS+ATP
PTS aMD (1)	100	5O2L+ATP
PTS aMD (2)	100	5O2L+ATP
PPS aMD (1)	100	2V26+ATP
PPS aMD (2)	80	2V26+ATP

Table 7.1.: List of accelerated MD simulations of the myosin VI motor domain.

7.1.2. Results and Discussion

7.1.2.1. Enhanced converter fluctuations

Figures 7.1 top (scatter plot), 7.1 bottom (density lines) and 7.2 (time-series) show the distribution and evolution of the converter descriptive observables X', Y' and Z' during the aMD simulations. As compared to unbiased simulations, it is seen that the aMD boost expectedly results in an enhanced positional dynamics of the converter, which explores regions of the (X', Y') map left un-sampled by unbiased runs. Notably, the following observations are made:

- In the PR (1) simulation, a movement of the converter in the direction of PTS is captured.
- In the PTS (1) simulation, a movement of the converter in the direction of PR is captured.
- In the PTS (2) simulation, a movement of the converter in the direction of PPS is captured.
- Most strikingly, in the PPS (2) simulation, the converter quickly relaxes to a PTS-like position, which overlaps clearly with the PTS crystal structure and the region sampled by unbiased PTS simulations.

By contrast, the PR (2) simulation explores converter fluctuations which are essentially orthogonal to the PR → PTS transition, and the PPS (1) simulation remains in the PPS basin for most of the simulation. However, the overall picture emerging from aMD simulations is that partial or complete transitions between the PTS and the end-states are captured, but, interestingly, no direct PR → PPS transition nor the other way around. Thus, the aMD simulations support the relevance of the PTS structure as an intermediate along the recovery stroke, at least regarding the position of the converter.

7.1.2.2. Behaviour of switch II

The behaviour of switch II described by the d_1 and d_γ observables is reported on Figures 7.3 and 7.4. All aMD simulations, regardless of the starting conformation, spend most of the time with a formed critical salt-bridge (d_1 = 4 Å). In PR, a rapid formation of the critical salt-bridge is observed. In PTS, reversible opening and closing of the salt-bridge is observed on the simulation time-scale. Finally in PPS (which exhibits a closed salt-bridge in the starting configuration), the salt-bridge is maintained for most of the simulation (PPS (1)) or all of it (PPS (2)).

By contrast, the switch II-ATP hydrogen bond (monitored by d_γ) remains in its starting state (*i.e.* open in PR and PTS and closed in PPS) for all but one simulation. But, the PPS (2) simulations spends longer with a broken hydrogen bond, although a reversible formation event is captured. Strikingly,

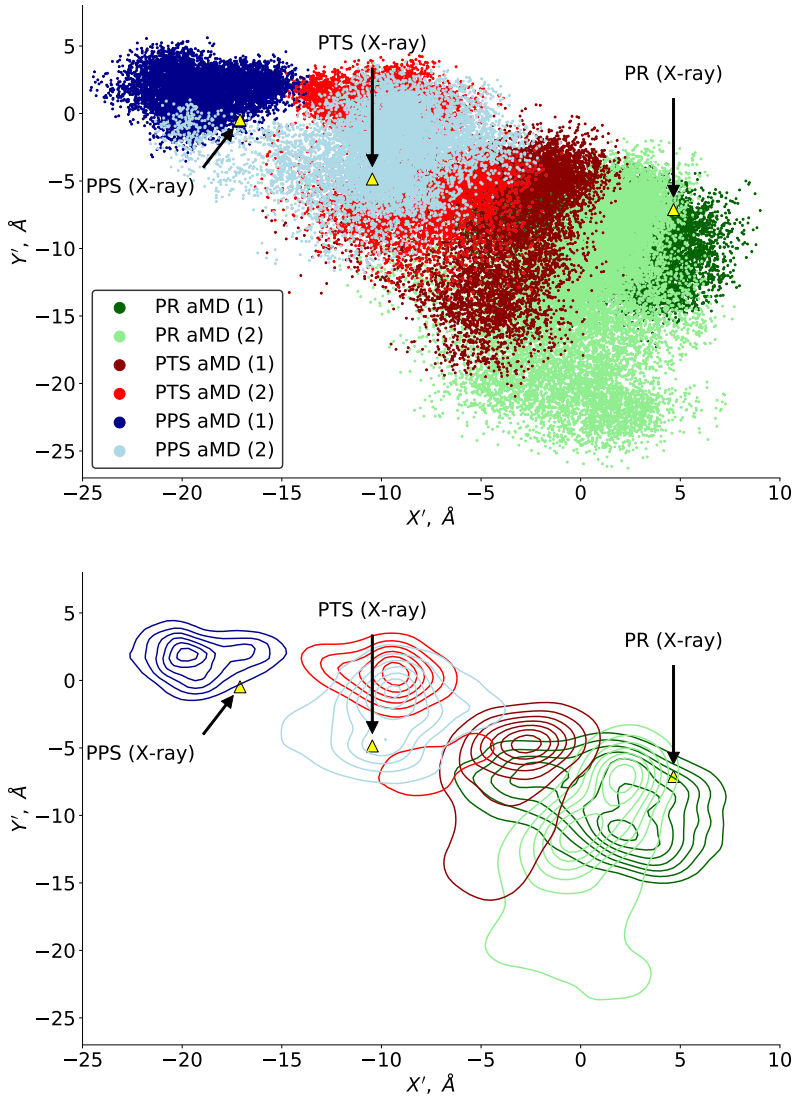


Figure 7.1.: Distribution of converter positions in aMD simulations

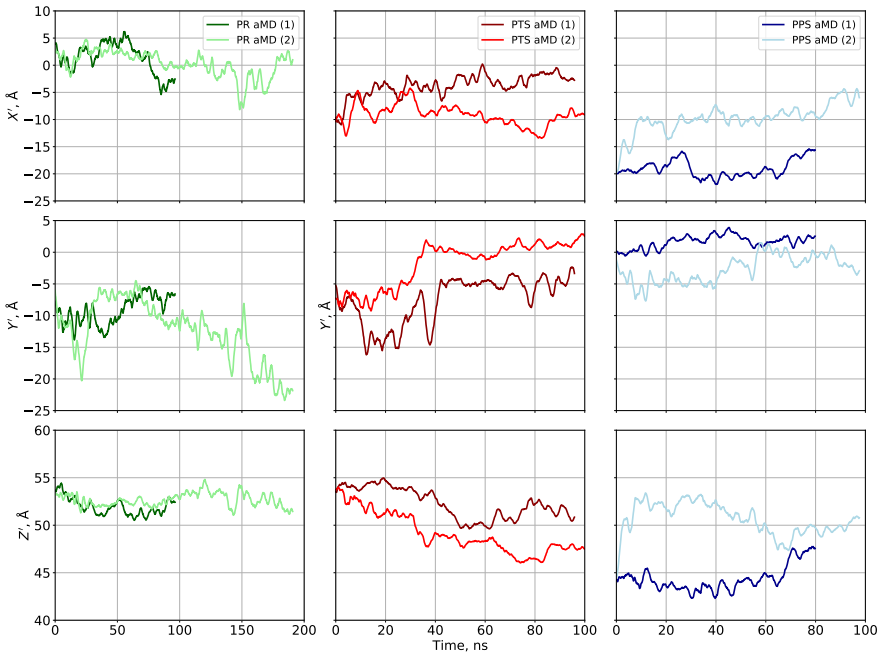


Figure 7.2.: Time-series of converter position in aMD simulations. The 2 ns running-average is shown for clarity.

this is the PPS simulation in which the converter rapidly takes on PTS-like positions. Thus, in the PPS (2) simulation, both the partial re-opening of switch II and the movement of the converter strongly suggest that this trajectory captured an event of PPS → PTS transition, which would support the PTS hypothesis.

7.1.2.3. Relay-SH1 elements

Similarly to the converter, the aMD boost enhances the conformational fluctuations of the Relay and SH1 helices as described by the θ_{RH} and θ_{SH1} orientation angles, see Figures 7.5 and 7.6.

Both PTS simulations and the PPS (1) simulations explore regions compatible with their respective crystal structures; the PPS (2) simulation, in agreement with the behaviour for other observables reported above, explores PTS-like values of the angles. The PR simulations exhibit a significantly enhanced sampling of the (θ_{RH}, θ_{SH1}) map as compared to unbiased MD, suggesting that events of Relay helix bending/kinking (simulation PR (1)) or SH1 helix tilting (simulation PR (2)) are captured. Visual inspection of the trajectories reveal that this is in fact not always the case. Rather, several simulations (PPS (1) and PR (1),(2)) explore disordered-SH1 configurations in which the SH1 helix unfolds, as evidenced by the time-series of its internal CA RMSD reported on Figure 7.7.

Regarding the Relay helix, the time-series of the local, kinking region $\Delta RMSD$ (see chapter 6) presented on Figure 7.8 suggest that some partially kinked configurations may be explored in the PR (1) simulation. Visual inspection shows that a kink is indeed observed in the Relay helix, but it is not located at the same place as the kink observed in the PTS/PPS crystal structures (”pseudo-kinked” configuration). An example conformation of the motor domain with a pseudo-kinked Relay helix and a disordered SH1 helix, extracted from the PR (1) simulation, is shown on Figure 7.9.

The observation of a disordered SH1 helix is reminiscent of an earlier proposal that SH1-unwound states (”internally uncoupled”) may be explored during the recovery stroke, which was formulated on

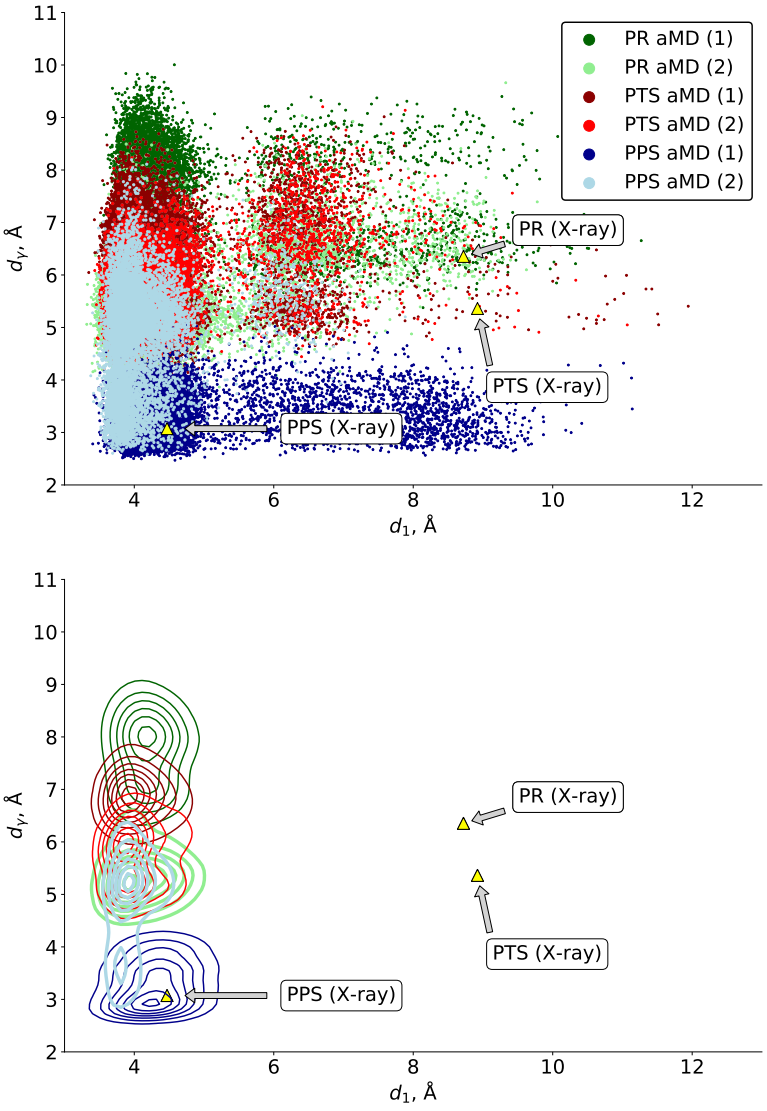


Figure 7.3.: Distribution of the active-site distances d_1 and d_γ in aMD simulations.

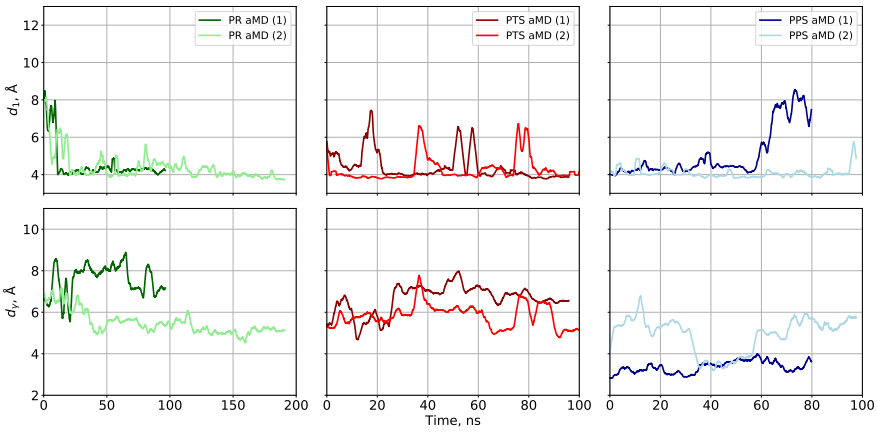


Figure 7.4.: Time-series of the active-site distances d_1 and d_γ in aMD simulations (1 ns running averages).

the basis of cross-linking and crystallography results (Himmel et al. 2002; Houdusse, Kalabokis, et al. 1999; Nitao and Reisler 1998). In addition, the spontaneous formation of a kink in PR, while switch II is not completely closed, would support the idea that the recovery stroke is initiated by the rearrangement of the force-generating region rather than the closure of the active site. However, instability of secondary structures, in particular helical ones, is a well-known artifact of aMD. Without the ability to reweight the trajectories, it is essentially impossible to determine if the explored disordered configurations could represent biologically meaningful intermediates of the recovery stroke, or are just spurious, high-energy configurations. Thus, even though several observations from aMD apparently support the PTS hypothesis, these are not enough to draw more than tentative conclusions.

In an attempt to preserve the enhanced sampling properties while allowing for a more robust interpretation of the simulations, we turned to Gaussian Accelerated MD, for which reweighting is possible.

7.2. Gaussian Accelerated MD

7.2.1. Set up

For the GaMD simulations, we decided to focus on the PR state in a first time, as the PTS hypothesis predicts that it should undergo a transition to PTS. As such, an accelerated simulation of PR could reach the PTS basin without prior knowledge of its existence, which would represent a strong supporting argument for the PTS hypothesis. All GaMD simulations were performed using the dual-boost approach and the lower bound threshold value with the NAMD implementation (Pang et al. 2017).

GaMD simulations must be preceded by a so-called "GaMD equilibration" in which statistics about the potential energy of tGaMD equilibration, see section 3.2.4.2. In our set-up, stage 1 was 1 ns long, stage 2 was 19 ns long, stage 3 was 5 ns long and stage 4 was 75 ns long for a total of 100 ns of equilibration, which is similar (slightly longer) to the values reported for a system of comparable size (Palermo et al. 2017)). The cut-off for energy fluctuations of the boosting potential σ_0 was set to 6 kcal mol^{-1} for both the total potential energy boost and the dihedral boost.

From the GaMD equilibrated structure and boost, 5 100 ns independent production simulations were run. The results of these simulations follow.

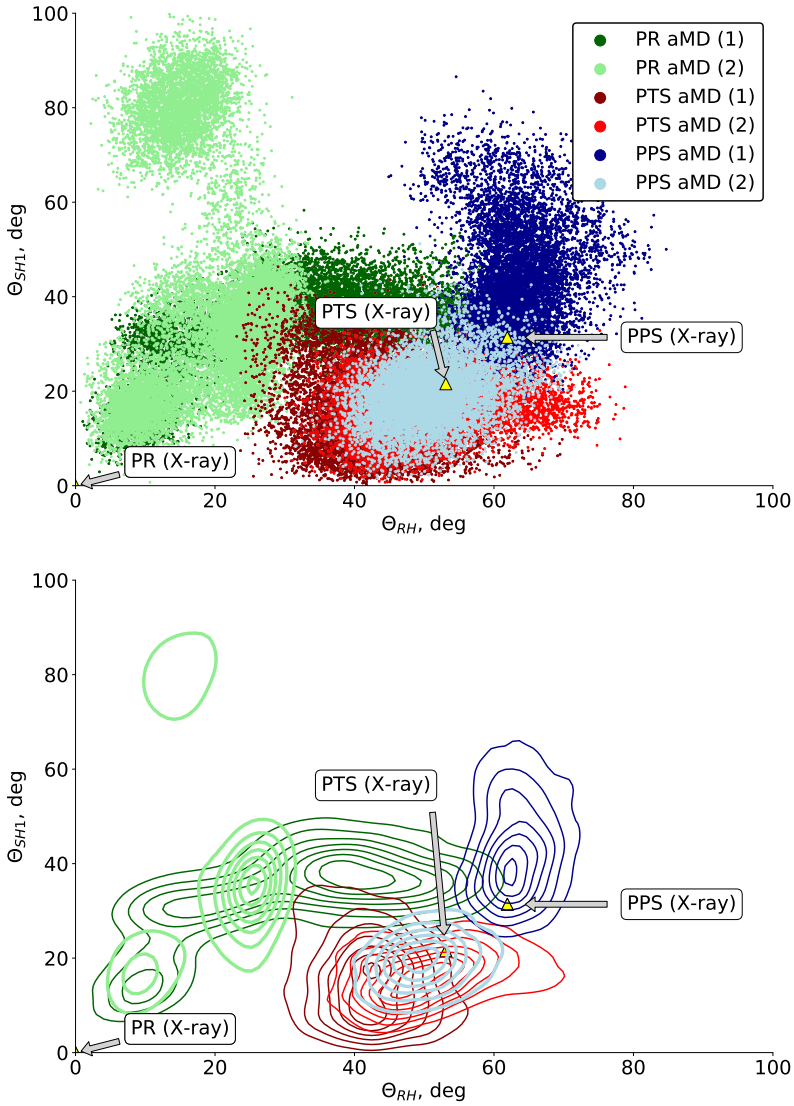


Figure 7.5.: Distribution of the Relay-SH1 elements angular descriptors θ_{RH} and θ_{SH1} in aMD simulations.

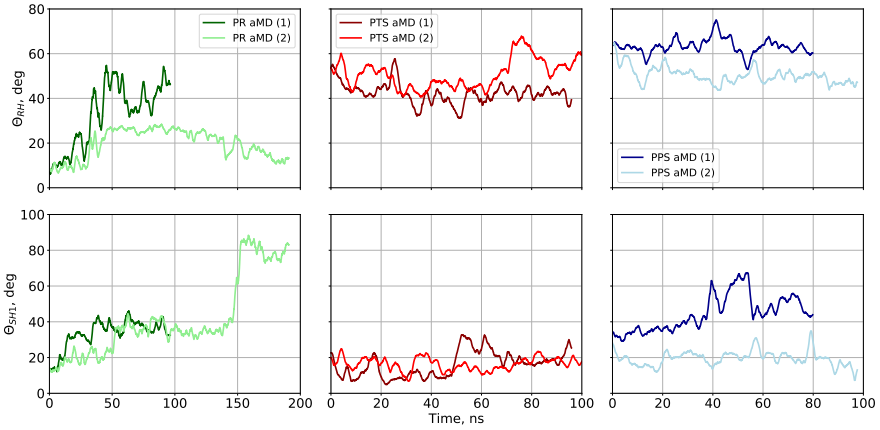


Figure 7.6.: Time-series of the Relay-SH1 elements angular descriptors θ_{RH} and θ_{SH1} in aMD simulations (1 ns running averages).

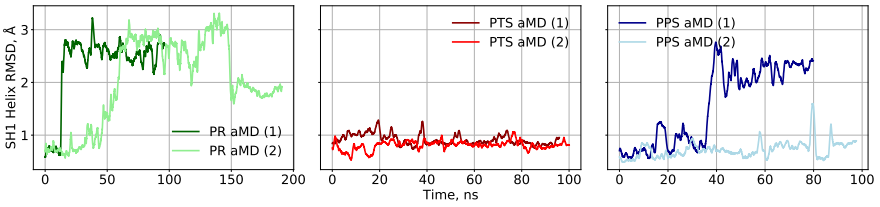


Figure 7.7.: Evolution of the internal CA RMSD of the SH1 helix during aMD simulations (2 ns running average). Abrupt increases in RMSD correspond to helix unfolding.

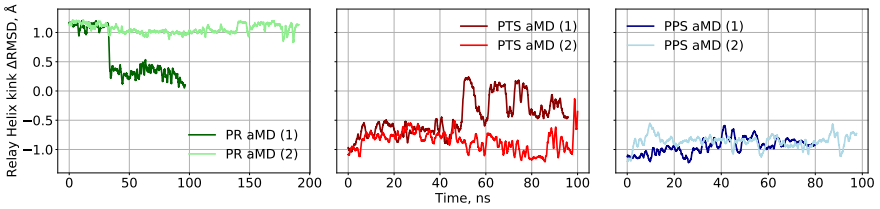


Figure 7.8.: Evolution of the Relay helix kink local $\Delta RMSD$ during aMD simulations (2 ns running average).

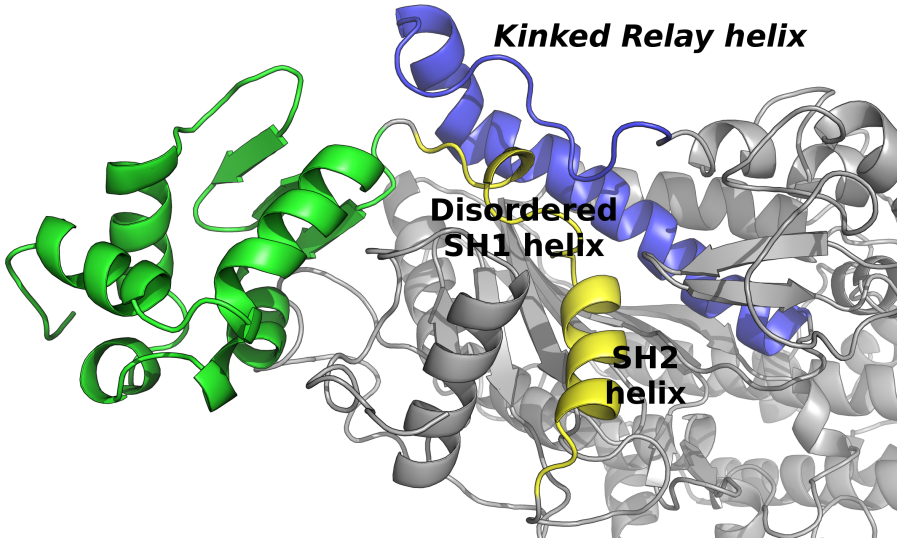


Figure 7.9.: Example of pseudo-kinked Relay helix, disordered SH1 helix sampled in PR aMD simulation.

7.2.2. Results and discussion

The results of the GaMD simulations are disappointing in the sense that the observed behaviour, regarding the average and fluctuations of relevant observables, barely seems to deviate from what is observed in unbiased simulations. Despite the care taken in constructing the GaMD boost (with a very long equilibration to ensure a robust estimation of the parameters), it does not seem that the sampling is enhanced. This is illustrated on Figure 7.10 for the converter position; similar patterns are observed for other observables. Consistently, re-weighting the statistical distribution of the observable X' obtained by pooling the data from the 5 simulations (using the cumulant expansion to second order), barely affects the distribution (Figure 7.11).

In fact, it seems that preserving the ability to reweight in a large system entails using a bias several orders of magnitude smaller than in regular aMD. Consequently, the extent of the conformational exploration seems seriously reduced. Notably, the recent study by McCammon's team of the CRISPR-Cas9 functional mechanism demonstrated the ability of GaMD to yield a free energy landscape, but relied on prior Targeted Molecular Dynamics (TMD) calculations to produce a guess path of the conformational transition under study (Palermo et al. 2017). In our case, using TMD would defeat the purpose of the accelerated simulations, as they are intended to capture spontaneous transitions. Alternatively, much longer simulations could have been used, but this approach was not retained since its likelihood of success was not deemed high enough to justify its potential cost in computational resources.

Obviously, aMD/GaMD is not the only existing enhanced sampling strategy, even if we limit ourselves to those independent on the definition of reaction coordinates. For instance, coarse-grained force-field approaches were suggested by the second Reviewer of (Blanc et al. 2018) as a way to accelerate the sampling. Nevertheless, we chose not to attempt coarse-grained simulations, as we considered it highly likely that the simplification to the potential energy surface introduced by coarse-graining may change the nature and/or energetics of the transition pathways. Other strategies, such as replica-exchange, may also have been attempted and could still be should we decide to further this

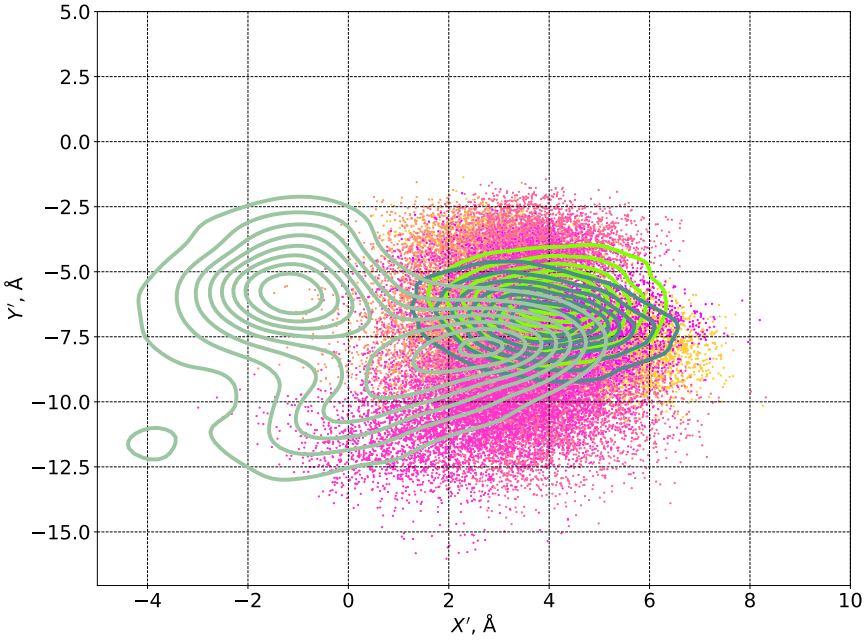


Figure 7.10.: Positional fluctuations of the converter in GaMD simulations (scatter dots) as compared to the density lines from unbiased simulations of the PR state (green lines). There is no apparent enhancement of sampling in GaMD.

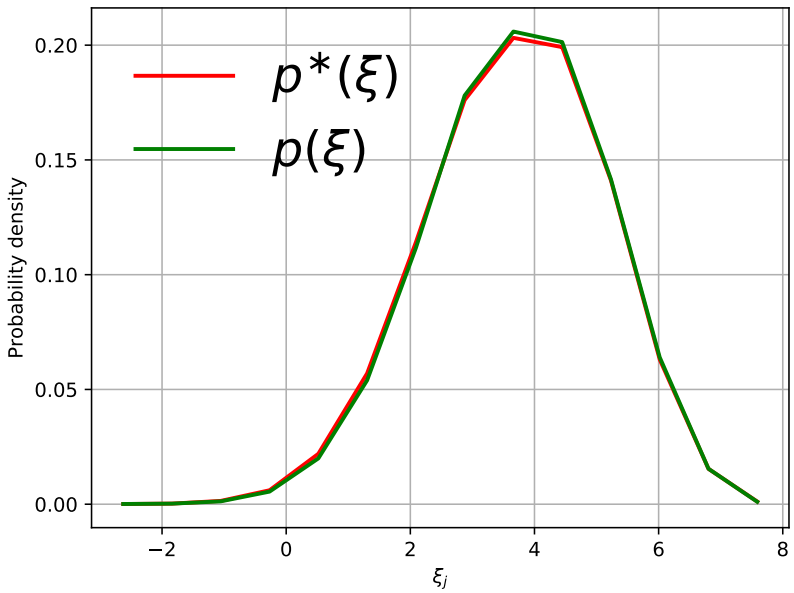


Figure 7.11.: Biased (p^*) and unbiased (p) statistical distributions for observable $\xi_j = X'$, in Å.

direction of research.

To summarize, the aMD simulations do capture transitions predicted by the PTS hypothesis (including a rather clear PPS → PTS transition), but do not offer the quantitative precision which would be required to confidently validate the hypothesis. By contrast, GaMD simulations may be suited to achieve such precision, but nothing meaningful -either in favour of or against the PTS hypothesis- is observed.

Overall, this suggests that the investigation of the PTS hypothesis by the means of ”agnostic” approaches (*i.e.* approaches in which one hopes to retrieve the PTS state without injecting knowledge about it) is not an efficient strategy. Thus, in the coming chapters, this strategy will be mostly abandoned in favour of approaches in which transitions involving PTS are explored using biases, and compared to the predictions of alternative models.

8. Energetics of ATPase activation in myosins

Summary In this chapter, we explore the effective free energy landscape of ATPase activation through switch II closure for the known conformations of the motor domain in the recovery stroke (PR, PTS and PPS). For this purpose, we employ ABF free energy calculations along two important interactions that must form for switch II to close. For myosin VI, the most crucial finding is that for both PR and PTS the fully closed switch II state is about 10 kcal mol higher in free energy than the open state, further supporting 1) the irrelevance of switch II-initiated models of the recovery stroke and 2) the existence of a statistical coupling between the converter motion and the closure of Switch II. Preliminary calculations on *Dictyostelium discoideum* myosin II suggest that the cost of closure in PR is is also very high in this isoform, suggesting it may not follow a switch II-initiated mechanism either. Finally, ABF calculations on the PTS of myosin VI reveal a stable state in which switch II is partially closed through an uncoupling from the Relay helix, pointing to a possible picture for the mechanism of late switch II closure. The results on myosin VI are part of our publication (Blanc et al. 2018).

8.1. ATPase activation through switch II closure

Switch II closure entails the formation of two important interactions: the critical salt-bridge (R205-E461 in Myo6) and the switch II-ATP hydrogen bond (G459 backbone nitrogen - oxygen on ATP γ phosphate). Inter-atomic distances representative of these two interactions can be used as reaction coordinates to probe the energetics of switch II closure using free energy calculations. We use the distance d_1 between R205CZ and E461CD to describe the critical salt-bridge, and the distance d_γ between G459N and ATP O1G to describe the hydrogen bond, see Figure 8.1. We operate under the assumption that the myosin motor domain is catalytically active when both interactions are formed, which allows us to study ATPase activation without resorting to quantum-mechanical simulations of the ATP hydrolysis reaction. This assumption is generally consistent with the literature (see chapter 5), despite a very recent QM/MM study which suggested that other, yet unidentified residues might also be involved in promoting the hydrolysis (Lu et al. 2017).

8.2. Free energy calculation strategy

For this study, we used the Adaptive Biasing Force (ABF) framework (see 4.3.4.3) as implemented in NAMD/colvars (Fiorin, Klein, and Hénin 2013; Phillips et al. 2005). Among the variety of free energy calculations techniques, the choice of ABF was motivated by the following considerations. First, it is a ”non-directed” strategy in the sense that a time-dependent biasing force is constructed on-the-fly and allows a gentle exploration of the configurational space. This seems preferable to the Umbrella Sampling family of methods (4.3.3), for which starting conformations for the unlikely states must be produced beforehand. Second, as compared with metadynamics (which also employs a time-dependent bias to enhance sampling), the ABF formalism is less parameter-dependent; in addition, the functional form of the bias is firmly grounded in statistical mechanics as argued previously.

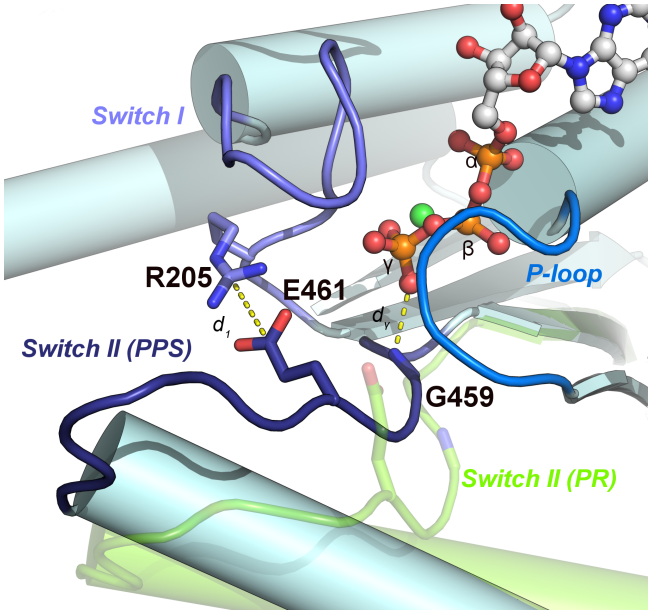


Figure 8.1.: Definition of the collective variables used to probe ATPase activation in myosin VI.

The two collective variables were discretized on a rectangular grid with a 0.1 Å spacing. As the G459N atom is connected to an hydrogen atom, it feels the force applied by the SHAKE constraining algorithm if this latter is active, which is incompatible with the computation of the generalized force estimate by ABF. As such, SHAKE was turned off for protein atoms (but kept on for water molecules) and a 1 fs timestep was used. The *fullSamples* parameter was set to 200, which is advised in the case of 2D ABF calculations.

We employed the following 2-step strategy. In a first time, a regular ABF run (hereafter termed "exploratory run") is initiated from the equilibrated structure of the conformational state under study. As the ABF bias starts to act, the system escapes the initial basin in configurational space and start to explore the grid. When most points of the grid have been sampled at least *fullSamples* times, this simulation is stopped. Then, the grid is divided into 56 1 Å x 1 Å non-overlapping windows to be used in a stratified, parallel ABF calculation. Each window is initialized starting from the ABF bias estimated during the exploratory run. The initial coordinates are also extracted from the exploratory run. At any time, the full estimate of the free energy gradient can be reconstructed by piecing together the data from the windows, and numerically integrated to obtain the two-dimensional PMF estimate (with the $abf_integrate$ tool).

Significance of the calculations Rigorously, our approach does not represent a proper calculation of the free energy profile along $(d_1,\ d_\gamma)$, because this would imply that all the degrees of freedom orthogonal to these two collective variables are equilibrated, and so, reversibly sampled to their full extent during the simulations. This would notably require reversible exchange between the various conformational states of the myosin motor domain. In this limit of perfect sampling, the resulting free energy landscape would be independent of the atomic structures used to initiate the calculations. This is not the case here; rather, we compute state-dependent effective PMFs when the global conformation of the protein is fluctuating *locally* within a given basin. In essence, this amounts to assuming that the dynamics of d_1 and d_γ is faster than the typical timescale of the global conformational transitions, such that d_1 and d_γ can reach local equilibrium within a given global conformational basin. The effective

PMF represents (up to a Boltzmann inversion) the probability distribution of $(d_1,\ d_\gamma)$ in these local equilibrium conditions. Thus, it is a tool to assess how the overall conformation of the motor domain affects the dynamics and accessible states of switch II, notably by revealing population shifts as the motor progresses along the recovery stroke[1].

The assumption that $(d_1,\ d_\gamma)$ equilibrate faster than the global conformation may certainly be false. However, if this were the case, this would rule out switch II-initiated scenarios as these latter precisely rely on the closure of switch II being faster than the overall conformational transition. In fact, switch II-initiated scenarios would be most consistent with the existence of a thermodynamically accessible "closed switch II" basin in the effective PMF observed in the PR state. We will see that this pattern is not observed, for myosin VI at least.

8.3. Effective free energy landscape of ATPase activation in myosin VI

We now discuss the results of the ABF calculations on ATPase activation. The convergence assessment, error analysis and some control simulations are presented later on.

8.3.1. End-states of the recovery stroke

PR state The free energy landscape for the PR state (Figure 8.2) exhibits a global minimum corresponding to a fully open switch II, to which the equilibrated structure belongs. The fully-closed switch II state is detected as a metastable basin, however the free energy difference from the ground state is about 11 kcal mol^{-1}. Another metastable basin is identified corresponding to a formed critical salt-bridge, but a broken hydrogen bond between ATP and switch II. This state is only about 2 kcal mol^{-1} higher in free energy than the ground state, from which it is separated by an approximately 7 kcal mol^{-1} barrier (in the direction of salt-bridge formation). This result suggests that the formation of the salt-bridge may initiate the closure of switch II, as previously proposed, (Stefan Fischer, Windshügel, et al. 2005). If this were the case, the minimal free energy pathway seems would involve 1) the formation of the critical salt-bridge and 2) the formation of the switch II-ATP hydrogen bond. From the "formed salt-bridge" basin, the barrier for this latter transition is of order 10 kcal mol^{-1} as estimated by reading the free energy values on Figure 8.2. Alternatively, effective barriers can be computed by using Boltzmann-integration to eliminate one of the collective variables (Figure 8.5). For the PR state, the effective barrier along d_γ is a good estimator for the overall barrier, because in this state the only basin in which the switch II-hydrogen bond is formed is the fully closed switch II state. This effective barrier is of the order of 12 kcal mol^{-1} from the fully open state, which we take as the estimate of the free energy barrier to switch II closure in PR. The large measured free energy level for the fully closed state shows that this latter may not be explored with significant probability while the motor domain is in the PR state. Rather, extensive conformational changes are likely to be required for a closed switch II to be stabilized.

PPS state In PPS, the fully closed switch II state is the global minimum, confirming that this conformational state is catalytically active, see Figure 8.3. Interestingly, a basin corresponding to a formed salt-bridge but broken hydrogen bond is detected with a 1 kcal mol^{-1} free energy difference from the

1. We note that this approach was used by Cui and co-workers to study the same problem on Dd Myo2 (Yu et al. 2007a, see also 5.3.3). In comparison, we use ABF rather than umbrella sampling, and the collective variables are not the same: Cui and co-workers used the $RMSD$ with respect to the closed and open configurations of switch II.

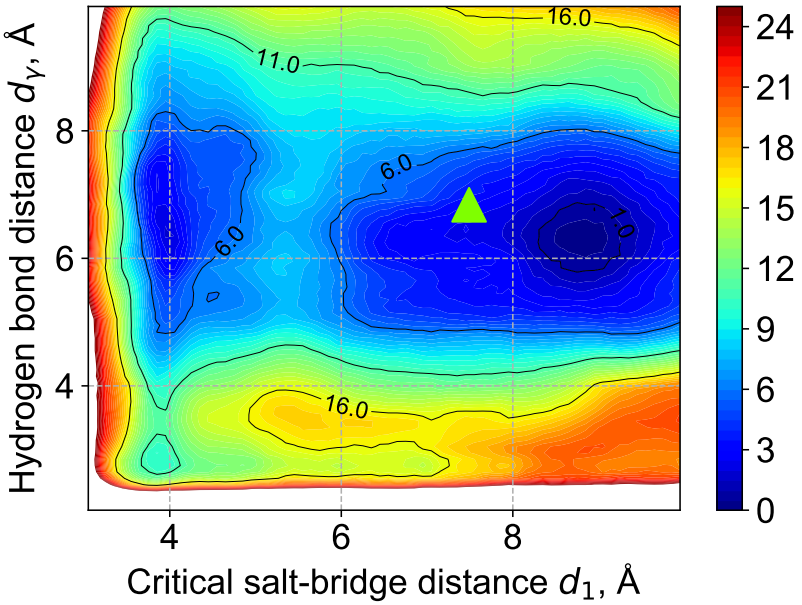


Figure 8.2.: Effective free energy landscape of switch II closure in the PR state of myosin VI. The triangle represents the position of the equilibrated structure. Free energy in kcal mol^{-1}.

ground state and separated from it by a 2 kcal mol^{-1} barrier. These low values suggest that the fully closed switch II may not be completely stable in presence of ATP, consistently with unbiased simulations (Chapter 6). Finally, a wide basin corresponding to a fully open switch II is found at about 6 kcal mol^{-1}.

Overall, the comparison of PR and PPS reveals how the global conformational change of the motor domain along the recovery stroke shifts the ground state of switch II from open to close. The very large reported free energy level of the closed switch II state in PR (10 kcal/mol), along with the associated barrier (12 kcal/mol), makes it unlikely that switch II closure initiates the recovery stroke.

8.3.2. PTS state

The effective free energy landscape for the PTS state exhibits two striking features, see Figure 8.4. First and foremost, the fully closed switch II state is still significantly higher in free energy than any other detected metastable state, as its free energy relative to the ground state is about 9 kcal mol^{-1}. Also, a new basin appears which corresponds to a previously un-described configuration in which the switch II-ATP hydrogen bond is formed, but the critical salt-bridge is not. Surprisingly, this state is identified as the global minimum. When the PMF projected onto d_γ, this basin, which represents the dominant contribution to the ”formed hydrogen-bond” ensemble of configurations, is found to be about 2 kcal mol^{-1} lower in free energy than the ”open hydrogen-bond” ensemble (Figure 8.5).

Visual inspection of the ABF trajectory shows that this basin corresponds to an ensemble of configurations in which switch II uncouples from the Relay helix and undergoes a movement towards ATP, described on Figure 8.6.

In this previously undescribed configuration of the active site, switch II undergoes a large motion along with a local conformational change, allowing for the formation of the switch II-ATP hydrogen-bond while seemingly disfavoring that of the critical salt-bridge. Switch II uncouples from the Relay

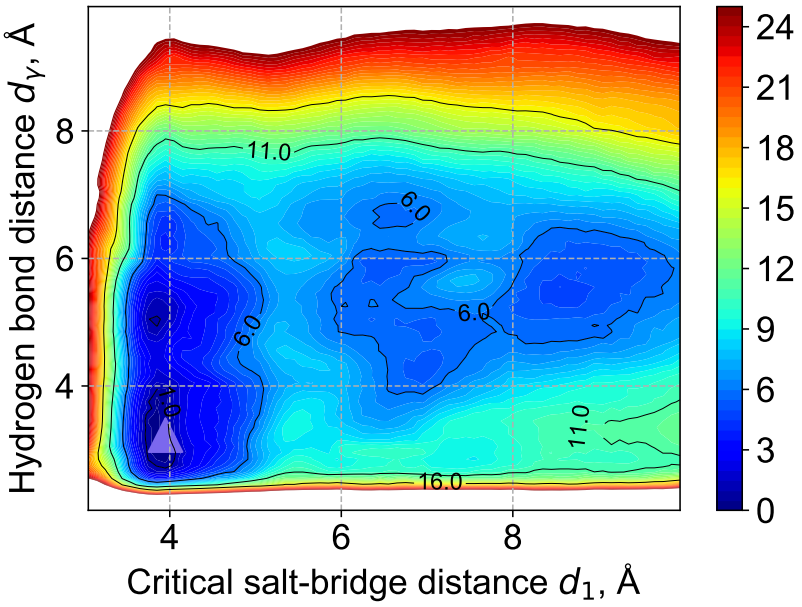


Figure 8.3.: Effective free energy landscape of switch II closure in the PPS state of myosin VI. The triangle represents the position of the equilibrated structure. Free energy in kcal mol^{-1}.

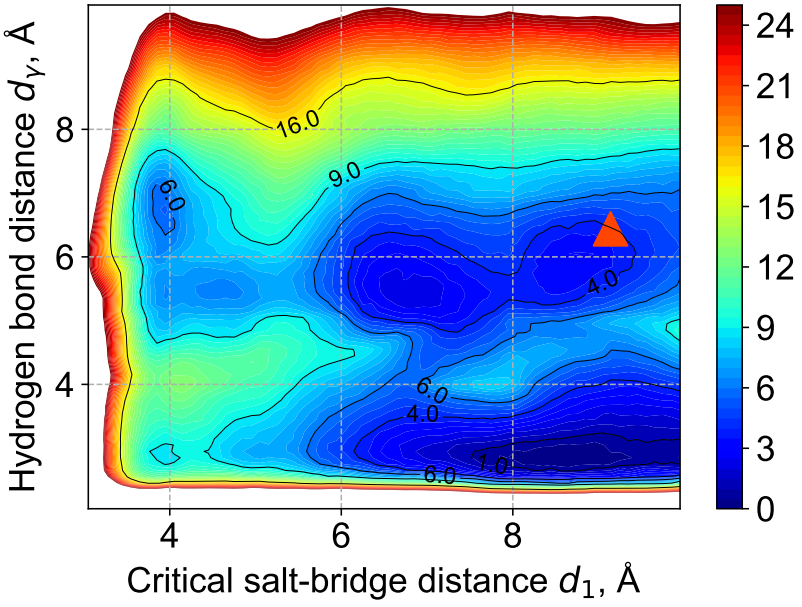


Figure 8.4.: Effective free energy landscape of switch II closure in the PTS state of myosin VI. The triangle represents the position of the equilibrated structure. Free energy in kcal mol^{-1}.

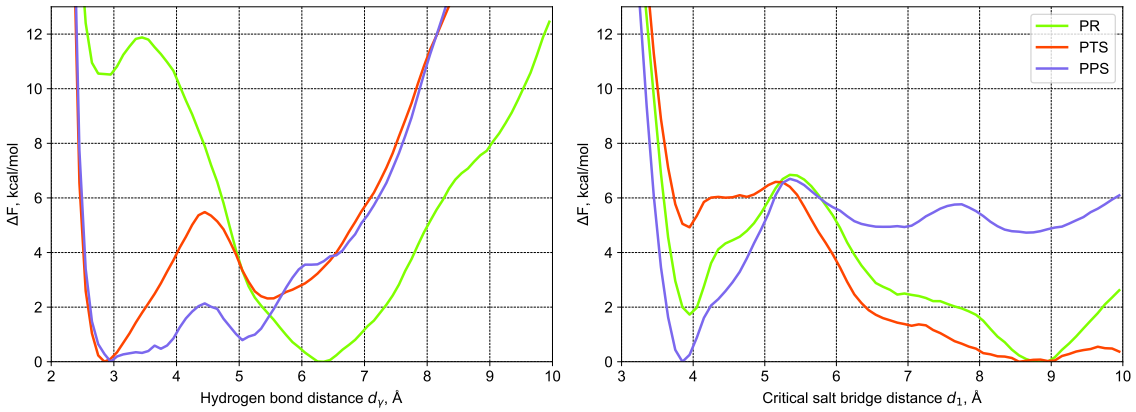


Figure 8.5.: Single-dimensional PMF obtained by Boltzmann-integration along the second degree of freedom.

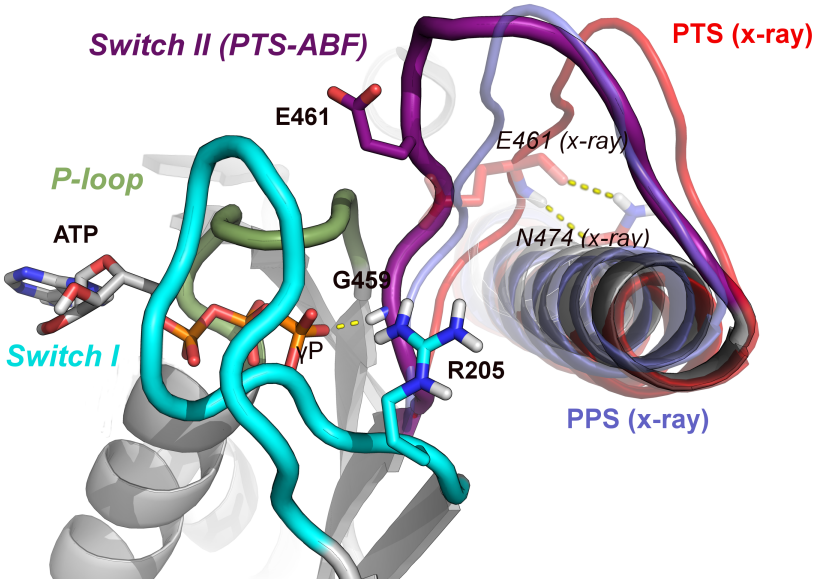


Figure 8.6.: Representative conformation of the ”uncoupled switch II” state captured in the ABF simulation of PTS, as compared to the crystallographic PTS and PPS configurations.

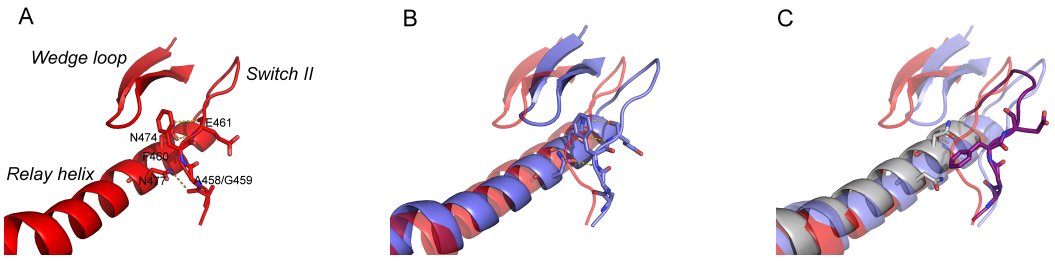


Figure 8.7.: Switch II is also uncoupled from the wedge loop in the ”uncoupled switch II” states captured in ABF simulations. A. Interaction of switch II with the Relay helix and the wedge loop in PTS. B. In PPS, the seesaw motion of the Relay helix and the rotation of L50 allow switch II to close while maintaining its interactions with these structural elements. C. In the ”switch II uncoupled” configuration, the hydrogen bond between switch II and the Relay helix is broken, as well as the hydrophobic contact between F460 and the wedge loop.

helix by breaking the N474(side chain):E461(backbone) hydrogen bond formed at the beginning of the simulation. In addition, this uncoupling entails the ”extraction” of the switch II residue F460 from an ”hydrophobic cradle” formed by surrounding side-chains of the wedge loop (residues 581-585, part of the L50), see Figure 8.7; in other words, switch II uncouples from the L50 as well as from the Relay helix. By contrast, in PPS, where the most stable configuration also exhibits a formed switch II-ATP hydrogen-bond, the coupling of switch II to both the Relay helix and the L50 is maintained (and may be required to also form the critical salt-bridge). We may speculate that this is made possible by the inward seesaw motion of the Relay helix and the inward rotation of the L50, which would yield a more confined organization of the active site. These two motions take place during the PTS → PPS transition. The identification of this ”uncoupled switch II” as the global minimum reveals that in PTS, the barrier for full switch II closure from the ground state entails a barrier of about 10 kcal mol^{-1}. Alternatively, if the path in which the formation of the critical salt-bridge occurs first, followed by the formation of the hydrogen bond, this latter partial transition would exhibit a rather close 9 kcal mol^{-1} barrier. 9-10 kcal mol^{-1} thus seems to be the range for the barrier to switch II closure in PTS, which is still a high value.

The transition towards a decoupled switch II was captured in a non-reversible manner during the exploratory phase of the ABF calculation, and corresponds to an orthogonal transition to the collective variables undergoing ABF dynamics. As a result, the exact value of the free energy of the decoupled switch II state relative to the fully open and fully closed states may be affected by an error coming from the fact that different regions of the orthogonal space are being sampled. This may explain why this state is identified as the ground state in PTS, which is inconsistent with the crystal structure. Alternatively, the fully open switch II state may have been selected by crystallization (considering the small 2 kcal mol^{-1} free energy difference).

To better assess the free energy difference between this basin and the other possible configurations of switch II, we could have performed a three-dimensional free energy calculation, supplementing d_1 and d_γ with a distance d_{RH} accounting for the hydrogen bond(s) coupling switch II and the Relay helix. Another strategy (perhaps less direct and as such less likely to succeed in our opinion) would have been to resort to shared ABF with bias exchange so as to enhance the sampling in orthogonal space. For lack of computer resources, we did not perform these expensive calculations.

We note however that 1) the error analysis of the ABF calculations (see below) did not highlight

any particular issue in PTS as compared to PR and PPS; 2) although the region of the configurational space with a formed hydrogen bond but an open salt-bridge is sampled in all three ABF simulations, only in PTS is the uncoupling of switch II captured. This would suggest that the coupling between the Relay helix and switch II is weakened in PTS in a such a way that the attraction of ATP takes over.

Whether this is an artifact of the free energy protocol, a non-functional but extant off-pathway state, or actually indicates that switch II closure from PTS proceeds by first uncoupling it from the Relay helix is presently unclear and would require a more detailed study. Nevertheless, the most important result of this calculation is that the fully closed switch II state is thermodynamically unfavored in PTS - demonstrating that despite the significant movement of the converter, the nucleotide-binding site remains in the catalytically-inactive state. This shows how, under the hypothesis that PTS actually represents a functional intermediate, the recovery stroke is not initiated by the closure of switch II, and the position of the converter and the state of the active site are statistically rather than strongly coupled.

8.3.3. Convergence and error analysis

The convergence of the calculations and the residual statistical error on the resulting PMF were analyzed as follows (Blanc et al. 2018, Supplementary Information). First, we made sure that each point of the $(d_1,\ d_\gamma)$ grid was visited significantly more times than *fullSamples*. Then, we computed the time-evolution of the RMSD of the gradient estimate within each window, with respect to the final gradient estimate, and checked that a plateau to near-zero values was achieved in most windows. To evaluate the residual statistical error, a bootstrapping-like approach inspired by (Wereszczynski and McCammon 2012) was used. Finally, we performed a separate set of free energy calculations using d_1 and an auxiliary collective variable d_2 corresponding to the distance between atoms R199CZ and E461CD, *i.e.* a secondary salt-bridge involving E461. Upon elimination of the second variable (either d_2 or d_γ) by Boltzmann-integration, the resulting one-dimensional PMFs along d_1 are very similar, which supports the proper convergence of our calculations. The complete error analysis is detailed in (Blanc et al. 2018, Supplementary Information), which can be found in Appendix C.

8.4. ATPase activation in Dictyostelium discoideum myosin II

The same approach to the elucidation of the energetics of ATPase activation was applied to the PR state of Dd myo2. Starting from the corresponding crystal structure (1MMR), an explicitly solvated equilibrated structure was prepared following the same protocol as for myosin VI (see Chapter 6). The collective variables d_1 and d_γ are defined in the same way as for myo6, using the corresponding Dd myo2 residues (*i.e.* R238/E459 for the critical salt-bridge and G457 for the hydrogen bond with ATP). A two-step ABF calculation was performed using the same parameters, grid and window definitions as for myosin VI. The exploratory run was 76.2 ns long; then, each window was simulated for 7.5 ns. The resulting two-dimensional PMF is presented on Figure 8.8.

8.4.1. Results and discussion

The potential of mean force exhibits two intriguing features. First, in the global minimum, the critical salt-bridge is formed (but not the switch II-ATP hydrogen bond). This is in contradiction with the crystal structure, but is consistent with the reported tendency of the critical salt-bridge to form easily in Dd myo2, (see for instance Baumketner and Y. Nesmelov 2011; Stefan Fischer, Windshügel, et al.

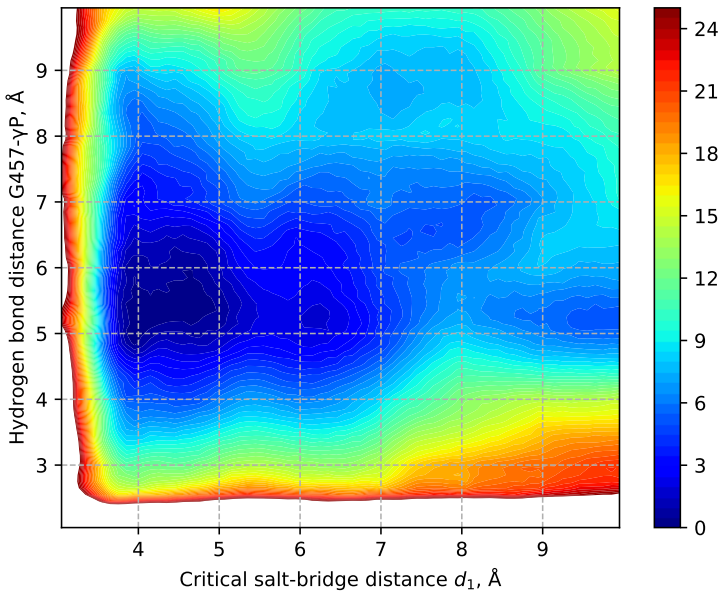


Figure 8.8.: Effective PMF for switch II closure in the PR state of Dd myo2. Free energy in kcal mol^{-1}.

2005). We also observed spontaneous closure of the critical salt-bridge on timescales less than 100 ns in unbiased simulations (data not shown). Second, no minimum is identified where the fully-closed switch II configuration should be. This may be the sign of an imperfect convergence of the ABF calculation. However, assuming that the calculation is meaningful, this would suggest that direct switch II closure from the PR conformation of Dd myo2 is extremely unlikely, because it does not even correspond to a transition to a metastable state. As such, the result of this calculation goes against a switch II-initiated mechanism for Dd myo2, like in myosin VI (myo6). This is an important observation, because Dd myo2 is the ”prototypical” myosin used for most of the previously published models of the recovery stroke (Chapter 5). Thus, our results directly challenge the previous proposals of a switch II-initiated recovery stroke.

8.4.2. Convergence analysis

As for myosin VI, we checked that 1) the configurational space is sampled by orders of magnitude more than $fullSamples$=200 (Figure 8.9), and 2) that the per-window-RMSD of the gradient estimate stabilizes towards near zero values (Figure 8.10). This suggests that the calculation may be properly converged in the local basin corresponding to the PR state.

8.5. Conclusion: Unfavorable energetics for early ATPase activation in the myosin superfamily?

The free energy exploration of the activation of ATPase in myo6 shows that the fully-closed switch II configuration is about 10 kcal mol^{-1} above the ground state in both PR and PTS, strongly suggesting that an early switch II-closure mechanism for the recovery stroke is very unlikely. Identical calculations on the PR state of Dd myo2 yield qualitatively similar results, challenging more directly previous

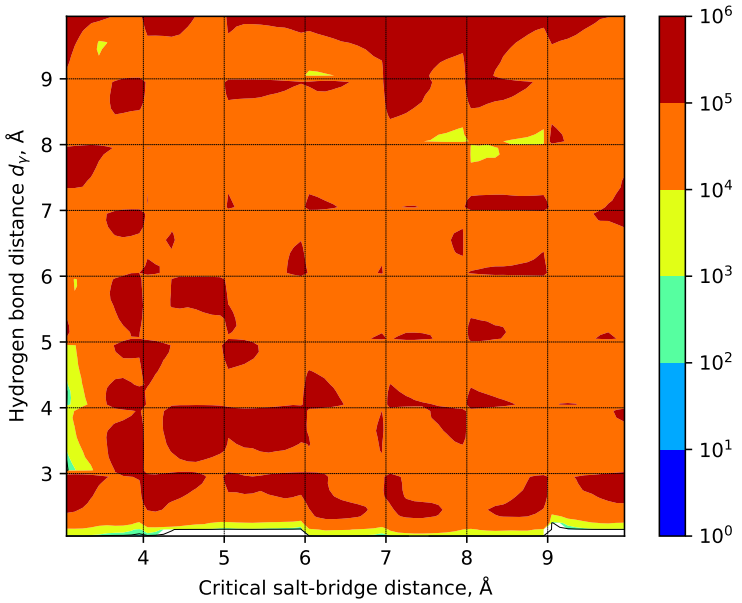


Figure 8.9.: Each grid point is visited orders of magnitude more than *fullSamples* during stratified ABF calculations on the PR state of Dd myo2.

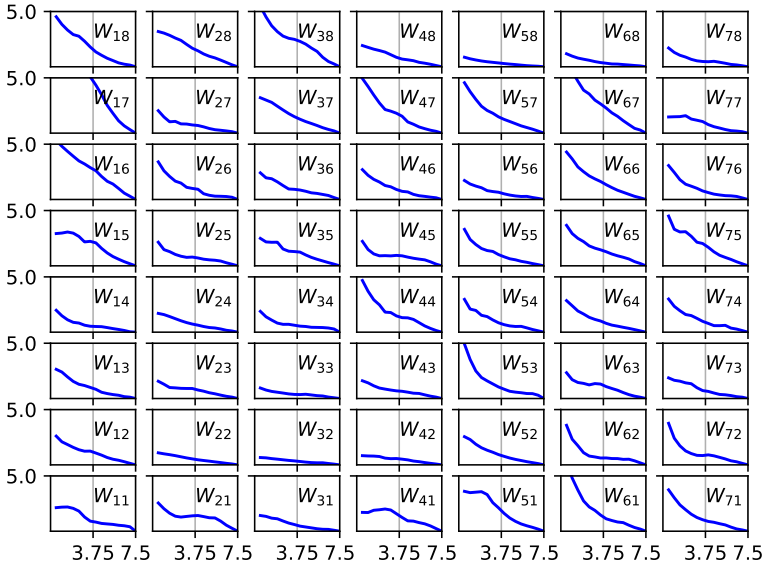


Figure 8.10.: Convergence of the gradient estimate RMSD with respect to its final value during stratified ABF calculations on the PR state of Dd myo2. 78.6% of windows are converged to within 0.2 kcal mol^{-1} Å^{-1}, and 100% within 0.5 kcal mol^{-1} Å^{-1}.

mechanistic proposals, and suggesting (in a preliminary manner) that the recovery stroke may not be switch II-initiated in the entire myosin superfamily. Of course, substantiating this claim would require performing the same calculations on a wider sample of myosin isoforms. At the time of writing, calculations on myosin Va, scallop muscular myosin and Smooth Muscle Myosin II (SMM2) are either in preparation or in the exploratory ABF stage.

There is a more fundamental limitation to our approach. Using ABF, we have probed the free energy barriers and differences corresponding to the switch II-initiated pathway. We conclude that this pathway is unlikely to be explored on the basis of the high reported free energy costs, but this is actually not enough to settle the question. Instead, one should make sure that the alternative pathway predicted by the PTS hypothesis (*i.e.* an early transition from PR to PTS) indeed exhibits a smaller free energy barrier than early switch II closure. To that end, a better understanding of the PR $\rightarrow$ PTS transition must be achieved, so as to design suitable collective variables to perform the free energy calculation. This is the focus of the next chapter.

9. Mechanism of the PR → PTS transition

Summary In this chapter, we first use Targeted Molecular Dynamics (TMD) and Steered Molecular Dynamics (SMD) simulations to study the mechanism of the PR → PTS transition, which represents the first major transition of the recovery stroke assuming the PTS hypothesis. The results show that this transition involves the concerted rearrangement of the Relay-SH1 elements and the rotation of the converter to an intermediate position, without coupling to switch II. Then, using eABF, we explore the two-dimensional free energy landscape along the movement of the converter and the formation of the kink in the Relay helix; this calculation, initiated from the PR state, is successful in identifying PTS as a free energy minimum and allows the estimation of the free energy barrier for the transition. A 7 kcal mol^{-1} barrier is found, which is smaller than the one reported for early switch II closure in PR, in support the PTS hypothesis. These results are unpublished.

9.1. Mechanistic study of the transition by Targeted/Steered MD

9.1.1. Specialized TMD of the Relay-SH1 elements

To study the PR → PTS transition, we first used "specialized" TMD simulations (Ovchinnikov, Trout, and Karplus 2010). In a specialized TMD simulation, only a sub-domain is submitted to the biasing potential, rather than the entire protein. As such, this method allows for the analysis of the structural response of other sub-domains to the driven transition of the biased sub-domain; it is thus a powerful tool to investigate the coupling between elementary rearrangements of sub-domains, which together make up the full transition.

A 15 ns specialized TMD simulation of the Relay-SH1 elements from PR to PTS was performed, using the native TMD module of NAMD and a force constant of 200 kcal/mol/Å^2. The results reveal that targeting these elements to their PTS conformation is sufficient to drive the swing of the converter from its PR to its PTS position (Figure 9.1), and in particular to break the N-terminal/converter interactions seen in PR (Figure 9.3, bottom). Strikingly, no effect on switch II is detected during the simulations, which remains open (see Figure 9.2). A replica of this simulation performed with identical parameters yields very similar results (data not shown).

Defining a converter polar rotation angle θ ($\theta = \arctan(Y'/X')$), one can monitor the movement of the converter as a function of time during the TMD simulation (Figure 9.3, top panel). It is seen that the movement of the converter occurs in two stages: first, a rather smooth, progressive rotation up to $t \simeq 8$ ns; then, an abrupt movement occurs and the converter reaches a state in which its average position does not seem to change, but with extensive positional fluctuations which overlap with the distribution observed in PTS unbiased simulation. The analysis of the number of contacts[1] between the converter and the N-terminal sub-domain as a function of time provides additional information, as it reveals that the converter first rapidly (from $t = 0$ to $t \simeq 2$ ns) reaches a state in which only about 20 contacts are maintained (versus 80 in the PR state); then, upon the abrupt movement at $t \simeq 8$ ns, most

1. We define a contact as any pair of heavy atoms within 4 Å of each other

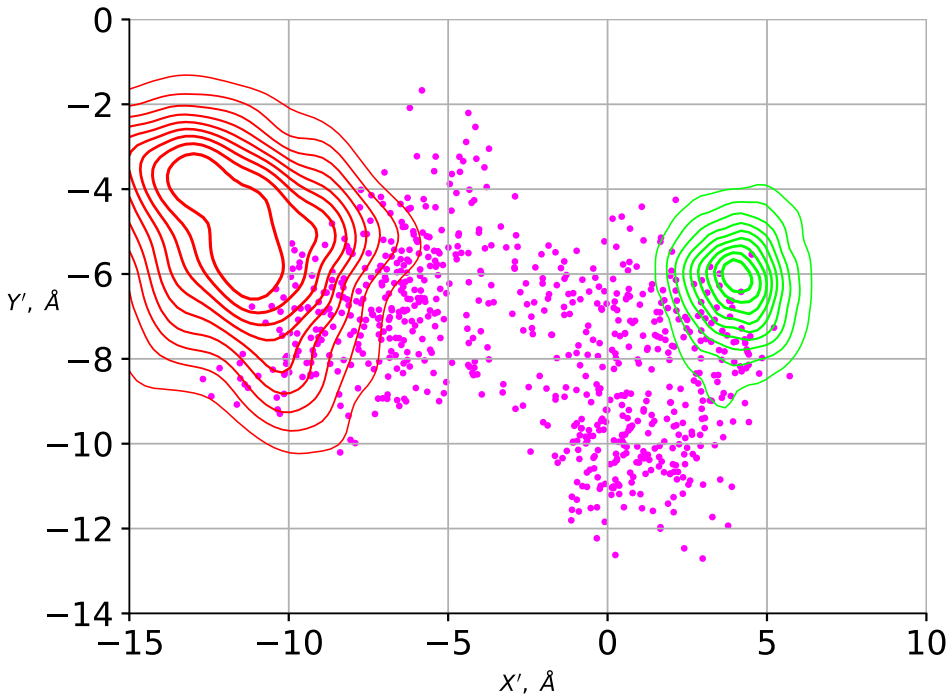


Figure 9.1.: Driving the rearrangement of the Relay-SH1 elements drives the movement of the converter from PR to PTS. The converter position during TMD, described by the X' and Y' observables, is in pink. Green and red density lines represent the statistical distributions of converter positions in the first 40 ns of PR+ATP (1) and PTS+ATP (1) unbiased simulations, respectively.

contacts are broken, similar to what is observed in the PTS unbiased simulation, which explains the increased positional fluctuations.

Visual inspection of the trajectory allows to relate the behaviour of the converter to the rearrangement of the Relay-SH1 elements driven by the TMD bias. It is seen that the progressive rotation of the converter in the first stage is associated with a progressive bending of the Relay helix; then, interestingly, the abrupt converter positional shift at $t \simeq 8$ ns corresponds to the formation of the kink in the Relay helix, illustrated on Figure 9.4. By contrast, no detectable effect on switch II is observed (data not shown). Simultaneously, hydrophobic contacts involved in maintaining the position of the converter in PR break, see Figure 9.5.

The results of this specialized TMD simulation are summarized on Figure 9.6.

9.1.2. Behaviour of the hydrophobic lock

The rearrangement of the hydrophobic lock (called aromatic switch in (Stefan Fischer, Windshügel, et al. 2005)), and corresponding in myosin VI to residues L489 on the Relay helix, Y508 on the Relay loop, L700 on the SH1 helix) has been put forward as an important elementary transition to stabilize the post-recovery configurations of the Relay-SH1 elements by several investigators (Baumketner 2012a; Stefan Fischer, Windshügel, et al. 2005), see also Chapter 5. It corresponds to a double rotameric transition of the side chains of L489 and L700, which exchange their positions while remaining in

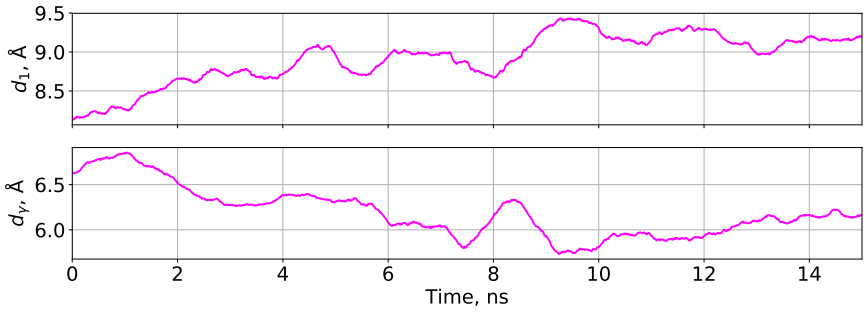


Figure 9.2.: Evolution of the active site distances d_1 and d_γ during specialized TMD simulation (1 ns running average). Switch II remains open.

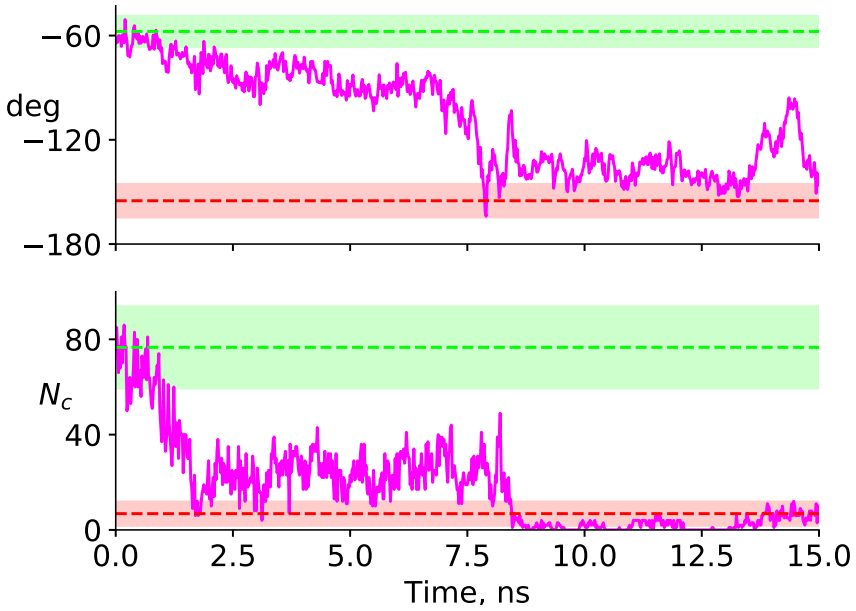
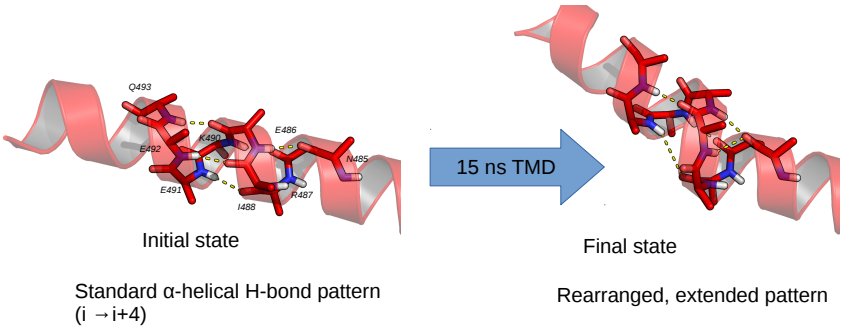
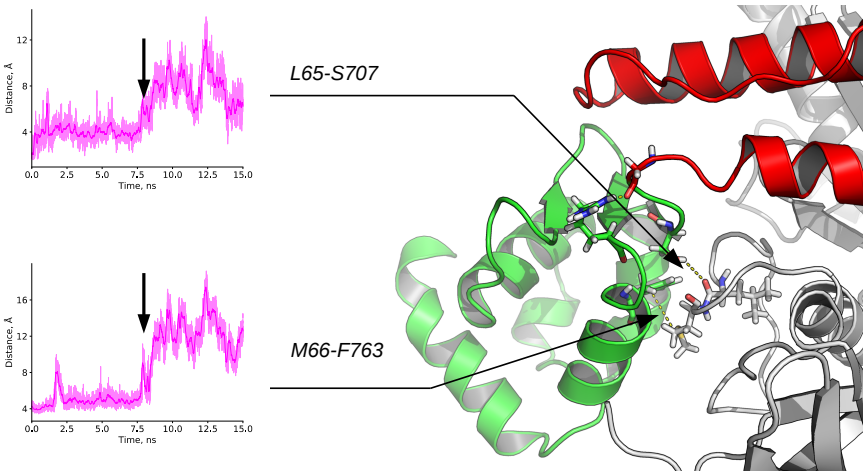


Figure 9.3.: Evolution of the converter rotation angle θ (top) and the number of converter/N-terminal contacts (bottom) during specialized TMD simulation. The green dotted lines materialize the average values in the first 40 ns of the PR+ATP (1) unbiased simulation, and the green transparent layers gives the corresponding fluctuations ($\pm$ standard-deviation). The red line and transparent layers represent the same quantities measured on the first 40 ns of the PTS+ATP (1) unbiased simulation.



1

Figure 9.4.: Formation of the kink in the Relay helix during the specialized TMD simulation.



1

Figure 9.5.: Evolution of hydrophobic N-ter/converter contacts during specialized TMD simulation.

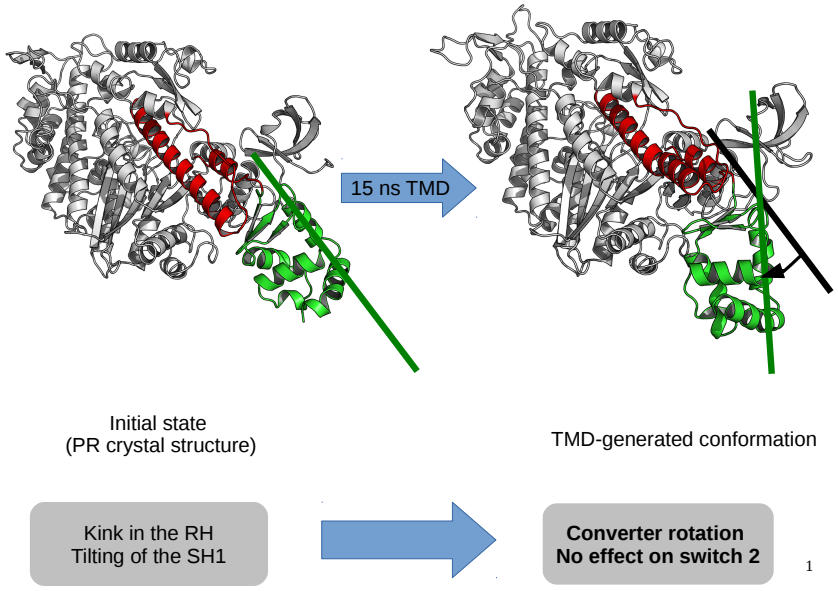


Figure 9.6.: Summary of the specialized TMD simulation: biasing the rearrangement of the Relay-SH1 elements (red) drives the rotation of the converter (green) to PTS.

hydrophobic contact.

The analysis of the TMD trajectory shows that the hydrophobic lock does rearrange, despite no explicit bias being applied on the side chains (Figure 9.7). Visual inspection shows that when the kink forms, the side chain of L489 is pushed towards the SH1 helix as the C-terminal part of the Relay helix undergoes a « corkscrew » like motion. As L489 is in hydrophobic contact with L700, this pushing movement drives the tilting of the SH1 helix and, incidentally, the rotation of the converter.

Interestingly, at the beginning of the simulation, the movement of L489 is sterically hindered by the bulky side chain of Y508 on the Relay loop. This sterical hindrance is relaxed in the early stages of the simulation by a near-rigid body motion of the relay loop, motion that is driven by the TMD forces as the relay loop is part of the restrained set of atoms in this simulation. This observation is of significant interest, as it directly challenges a proposal by Fischer and co-workers (Stefan Fischer, Windshügel, et al. 2005; Koppole, J. C. Smith, and Stefan Fischer 2007). In Fischer's model, the hydrophobic lock is presented as the determinant of the sequentiality of the recovery stroke (Chapter 5). Indeed, it is argued that the seesaw motion of the Relay helix is *required* to relieve the sterical hindrance to the hydrophobic lock rearrangement, which in turn is necessary for the kink to form. Thus, Fischer and co-workers conclude that the kink must form *after* the seesaw motion of the Relay helix. The PTS structure, which exhibits a kinked, but "un-seesawed" Relay helix, challenges this conclusion; in addition the present TMD trajectory reveals that a rearrangement of the hydrophobic lock is possible without seesaw through a seclusion of the Relay loop.

9.1.3. Robustness with respect to changes in the protocol

In addition to 2 TMD simulations which gave consistent results, we also ran 2 15 ns-long SMD simulations using the $\Delta RMSD$ of the Relay-SH1 elements (defined in chapter 6, 6.2.2.1) as a collective variable, with a 200 $\mathrm{kcal/mol/Å^2}$ force constant. One forward simulation (PR to PTS) and one back-

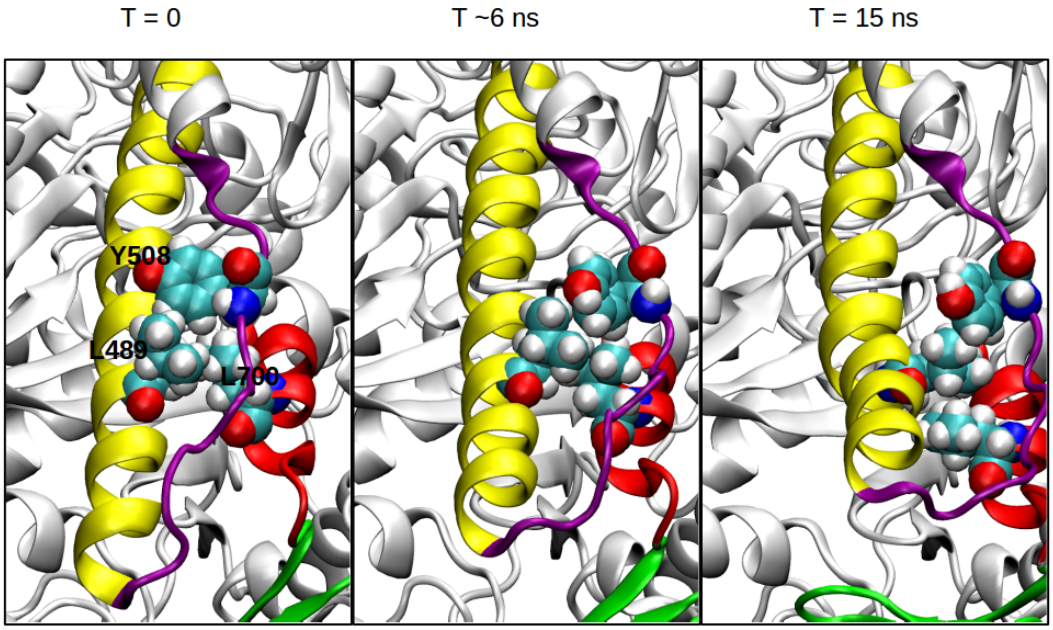


Figure 9.7.: Sequence of events in the rearrangement of the hydrophobic lock observed in the specialized TMD simulation.

ward simulation (PTS to PR) were performed. Although the backward simulation unambiguously produced a PR-like converter, the converter position explored at the end of the forward simulation is not that observed in the PTS crystal structure; rather, it corresponds to an alternate, metastable position sampled in the PTS+ATP (1) unbiased simulation (from 150 to 210 ns), see Figure 9.8. Thus, it is still a PTS-like position. Overall, and despite this surprising observation, this new set of simulations thus confirms that the rearrangement of the Relay-SH1 elements is sufficient to drive the swing of the converter from PR to PTS.

Finally, we performed two independent replicates of a 100 ns multi-CV SMD simulation, driving the collective variables listed in Table 9.1 from PR to PTS. These variables are designed to account for independent aspects of the global rearrangement of the Relay-SH1 elements[2].

The results of these simulations are shown on Figure 9.9. In agreement with the previous results, in simulation 2 the converter indeed reaches PTS. In simulation 1, however, this is not the case and the Relay-SH1 elements rearrange, but the converter still explores PR-like positions by the end of the simulation.

Visual inspection reveals that this is because the interactions between the Relay and the converter (which are always maintained in unbiased MD, regardless of the state), break upon the formation of

2. We note that we also performed individual SMD along subsets of these variables (*i.e.* only the quaternion angles, or only the kink distances), but these simulations were not successful in producing PTS-like positions of the converter (data not shown).

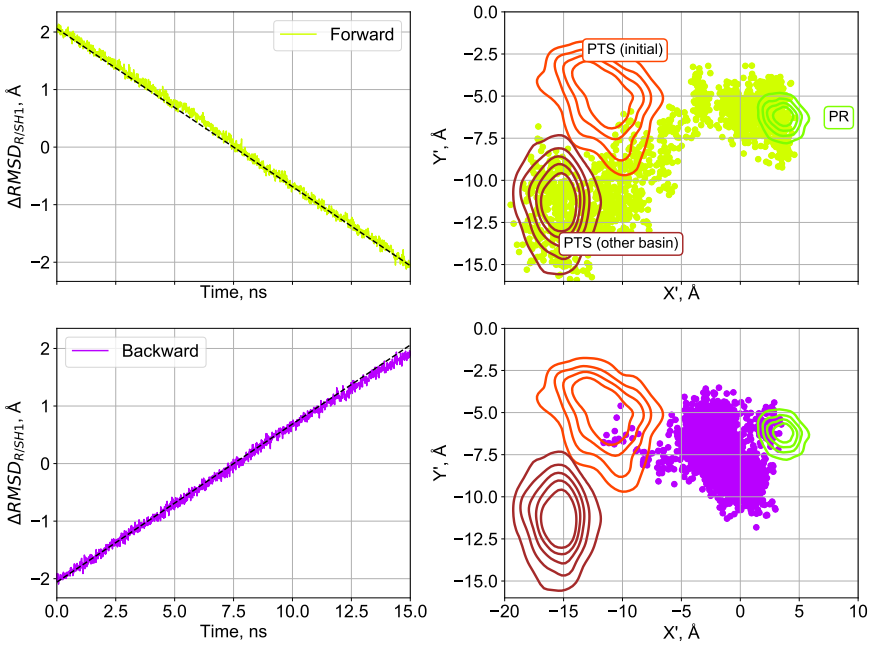


Figure 9.8.: SMD simulations along the $\Delta RMSD_{R/SH1}$ collective variable. Top, forward simulation. Bottom, backward simulation. The "PR" and "PTS (initial)" density lines on the (X', Y') map correspond, as previously, to statistical distributions during the first 40 ns of unbiased simulations PR+ATP (1) and PTS+ATP (1), respectively. The "PTS (other basin)" density lines represent the statistical distribution from t=150 ns to t=210 ns of unbiased simulation PTS+ATP (1), and correspond to a metastable converter position explored in this simulation.

Collective variable	**Associated rearrangement**	**Force constant**
Orientation quaternion L_{RH}	Bending/re-orientation of the Relay helix	1000 kcal mol^{-1}
Orientation quaternion L_{SH1}	Tilting/re-orientation of the SH1 helix	1000 kcal mol^{-1}
k_1 (486O:490N)	Kink in the Relay helix	10 kcal/mol/Å^2
k_2 (485O:489N)	Kink in the Relay helix	10 kcal/mol/Å^2
k_3 (485O:490N)	Kink in the Relay helix	10 kcal/mol/Å^2
k_4 (486O:491N)	Kink in the Relay helix	10 kcal/mol/Å^2
$d_{R/SH1}$ (469-482CA:693-703CA)	Seclusion of Relay and SH1 helices	10 kcal/mol/Å^2

Table 9.1.: Summary of the biased CVs in the multi-CV SMD simulations of the Relay-SH1 rearrangement. L_{RH} and L_{SH1} are defined similarly to their *orientation angle* counterparts (6.2.2.3) except that the full quaternion is used instead of only the angle. The k_i distances describe the backbone hydrogen bonds which rearrange during the formation of the kink in the Relay Helix, as illustrated on Figure 9.4. Finally, $d_{RH/SH1}$ measures the distances between the N-terminal region of the Relay helix and the SH1 helix. It is introduced to account for the seclusion motion reported by Baumketner in his study of the transition (Baumketner 2012b).

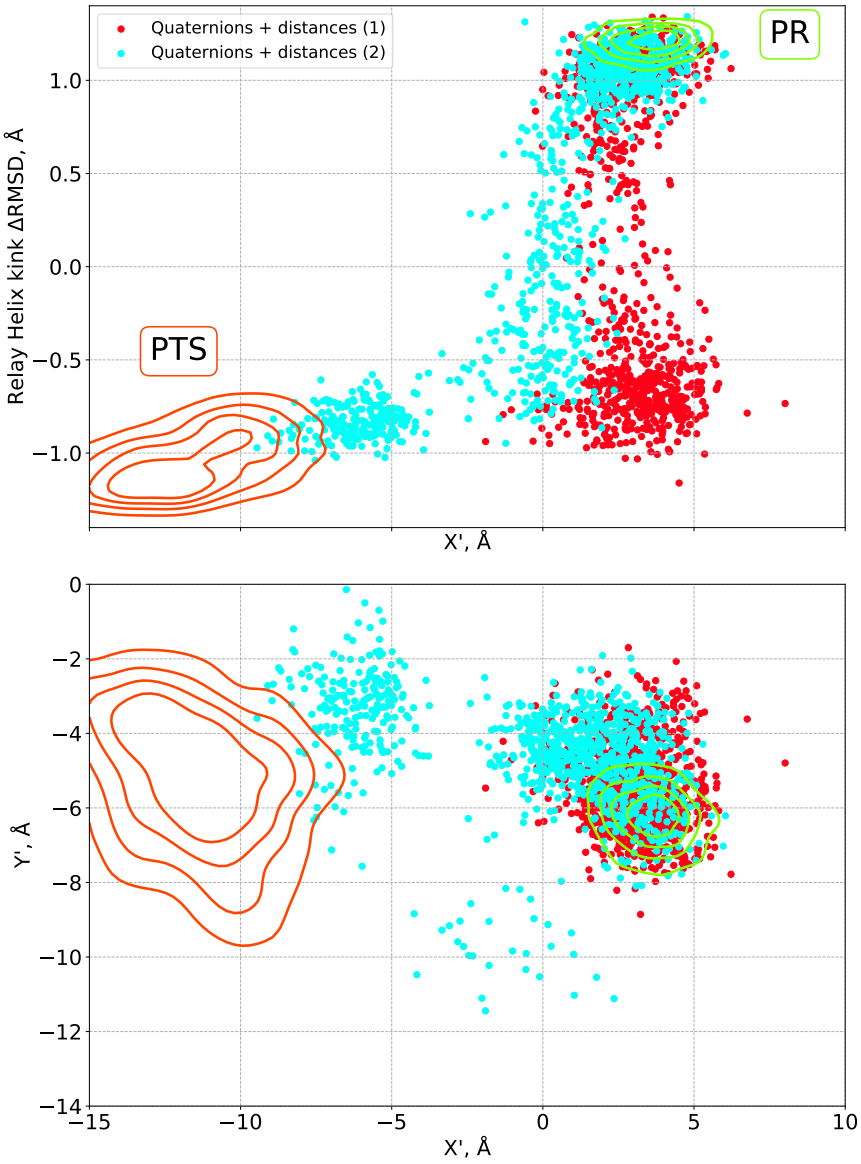


Figure 9.9.: Projection of the 100 ns multi-CV simulations onto selected observables of the recovery stroke. Top panel, the projection upon the $(X',\ \Delta RMSD_{kink})$ map allows the analysis of the coupling between the formation of the kink in the Relay helix and the position of the converter. Bottom panel, the projection upon the $(X',\ Y')$ allows the visualization of the converter movement in response to the SMD bias on the Relay-SH1 elements.

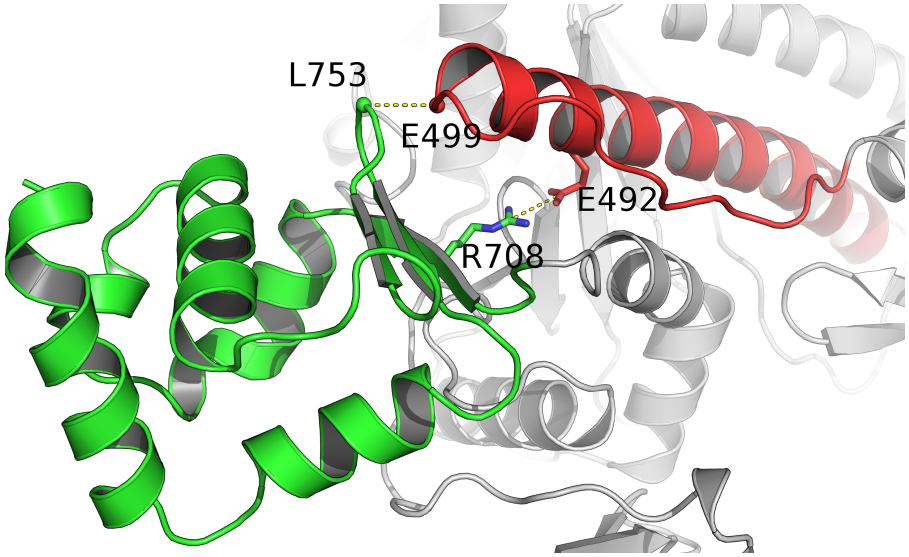


Figure 9.10.: Overview of the converter/Relay interface in PR and definition of the distances reported on Figure 9.11.

the kink in the Relay helix, whereas the contacts with the N-terminal sub-domain are maintained; see Figure 9.10 for an illustration of the (PR-like) converter/Relay interface, and Figure 9.11 for the evolution of distances describing the converter/Relay contacts and the converter/N-ter contacts during these SMD simulations. Interestingly, the time-series also reveal a transient breaking of the converter/Relay interactions upon the formation of the kink in simulation 2, suggesting that the converter may transiently uncouple from the Relay even in the case of a successful swing to PTS positions.

9.1.3.1. Pulling on the converter

Using SMD on the X', Y', Z' collective variables defined to characterize the position of the converter, we moved the converter from its PR to its PTS position, see Figure 9.12.

Most simulations revealed that it is possible to move the converter (breaking the contacts with the N-terminal sub-domain) without altering the state of the Relay helix, see Figure 9.13. This would suggest that the most energetically costly rearrangement is the formation of the kink, rather than the movement of the converter. Notably, it is consistent with the observation of a spontaneous, transient decoupling of the converter from the N-terminal sub-domain in one unbiased PR simulation (Chapter 6) and may indicate that the highest barrier to the PR to PTS transition is the formation of the kink in the Relay helix.

9.1.4. Comparison with the results of Baumketner

In his study of the recovery stroke of Dd myo2, Baumketner already pointed out the importance of the Relay-SH1 elements in controlling the rotation of the converter (Baumketner 2012a,b, see also 5.3.5). One may thus legitimately wonder as to the novelty of the present results, which is what we will now attempt to clarify.

As compared to Baumketner's work, our own results are obtained on an explicitly solvated, full-size myosin motor domain. In addition to providing a more realistic description of the system, our setup has the distinct advantage of making it possible to assess the coupling of the Relay-SH1 conformational

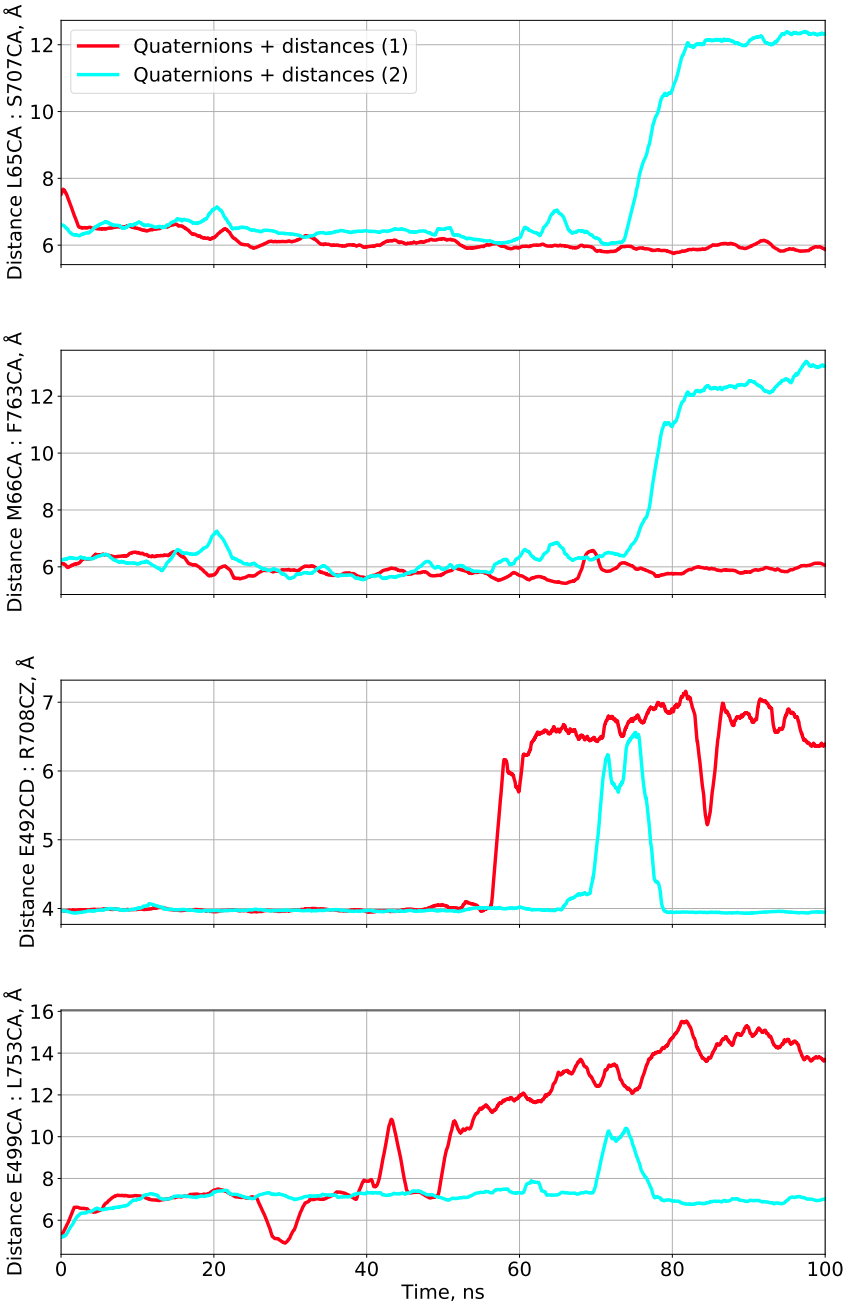


Figure 9.11.: Evolution of a set of converter/main body distances during multi-CV SMD simulations. Top panels, N-ter/converter distances. Bottom panels, Relay/converter distances.

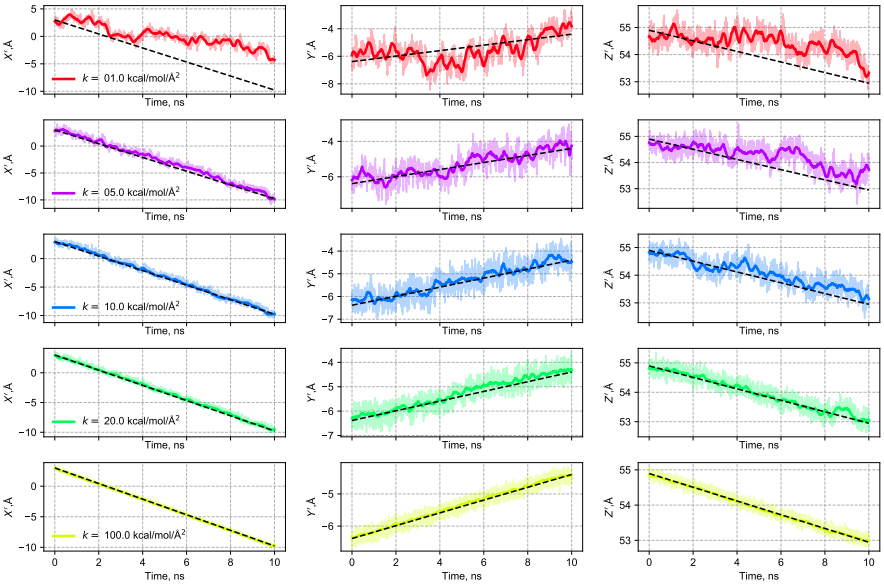


Figure 9.12.: Evolution of the converter observables X', Y', Z' under SMD bias for a range of force constants.

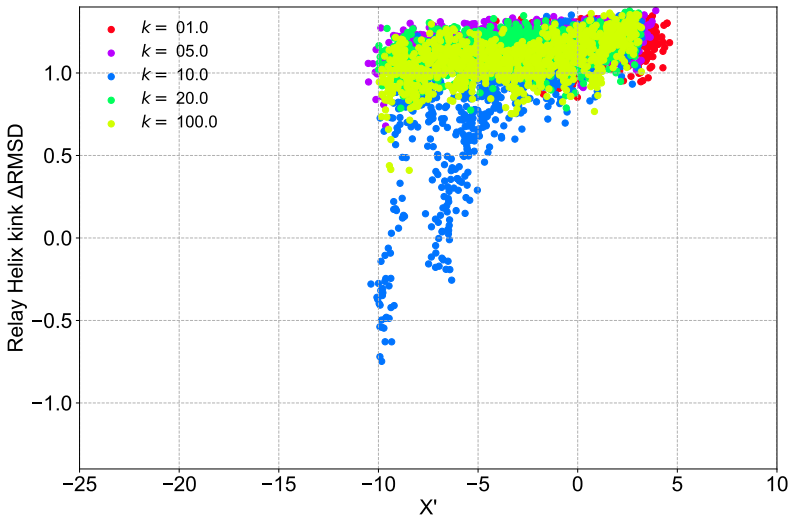


Figure 9.13.: Projection of the converter pulling simulations onto the X', $\Delta RMSD_{kink}$ plane. The results show that in all but one simulation, the Relay helix stays straight despite the converter moving to PTS-compatible positions. Force constants are in kcal/mol/Å^2.

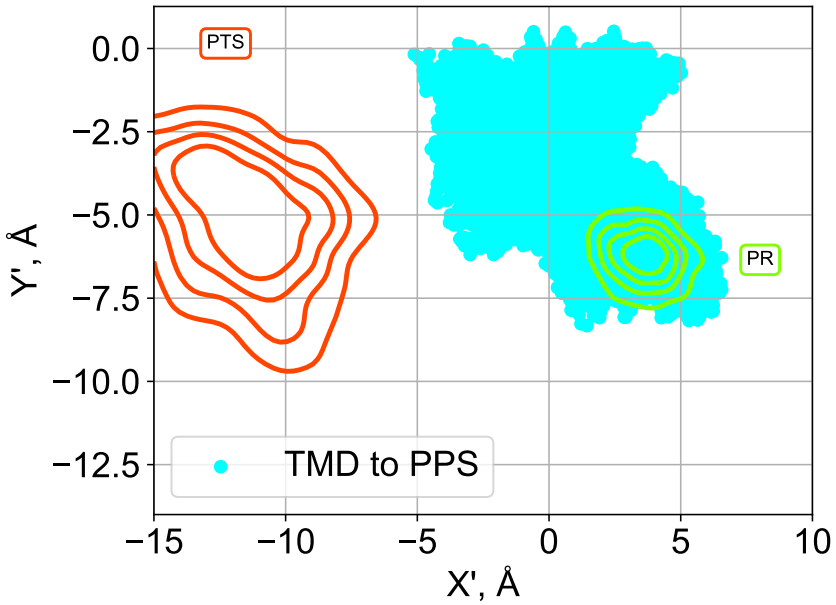


Figure 9.14.: A specialized TMD of the Relay-SH1 elements targeted from PR to PPS fails in yielding either PTS or PPS-like converter positions. A 15 ns specialized TMD was performed with the same protocol as the PR to PTS one reported above. Although the converter does move, it does not reach PTS, nor PPS.

transition with sub-domains other than the converter. As such, the absence of observation of any significant change in the remainder of the motor domain following the Relay-SH1 rearrangement - most importantly, in switch II- is to be understood as an actual, novel result.

We note again that a frustrating aspect of Baumketner's work was the lack of a proposed mechanism for the coupling between the rearrangements in the Relay-SH1 elements and switch II closure. In light of the PTS structure in general, and of the presently discussed results in particular, the reason appears clear: there is no such coupling at this stage of the recovery stroke. This is at least in qualitative agreement with the ABF results presented in chapter 8 regarding the PTS state, for which the fully closed switch II is not the global minimum despite the rearrangement of the Relay-SH1 elements.

Another important point is the subtle difference in conformation between the PTS and PPS Relay-SH1 elements. Despite an overall similarity, the kink angle of the Relay helix and the tilt angle of the SH1 helix change from PTS to PPS (see Chapter 6). Surprisingly, when the specialized TMD protocol described at the beginning of this section is applied from PR to PPS, it fails in yielding either a PPS or PTS-compatible position of the converter, see Figure 9.14. Both the knowledge of the PTS structure, and a complete description of the myosin motor domain, are required to meaningfully characterize the early stages of the recovery stroke.

9.2. Energetics of the transition probed by Umbrella Sampling

To go beyond the kinematic description emerging from the analysis of biased trajectories, one may investigate the energetics of the transition, *i.e.* the free energy profile along it. Such an analysis is expected to yield at least qualitative insights into the relative free energy between the PR and PTS states,

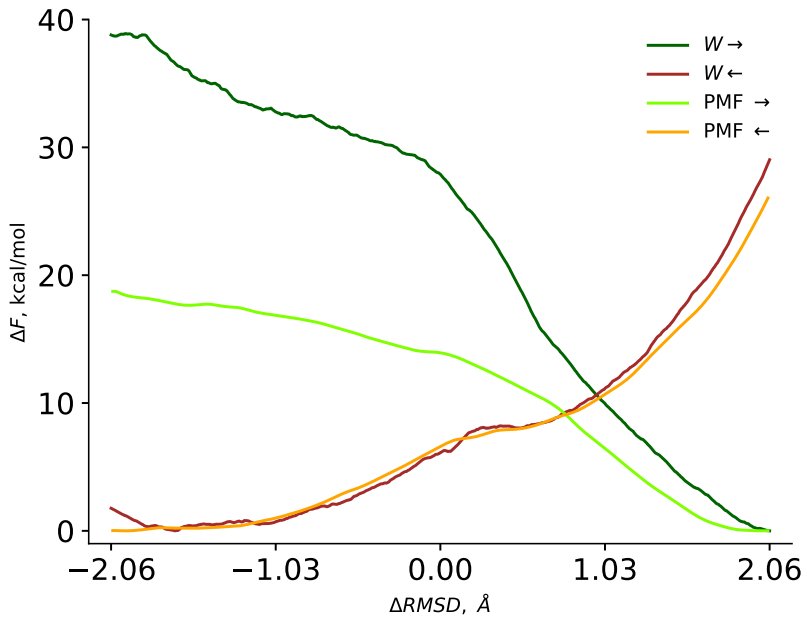


Figure 9.15.: Non-equilibrium work profiles and PMFs along the $\Delta RMSD$ of the Relay-SH1 elements in the PR → PTS transition, for forward (→) and backward (←) simulations.

and to allow the identification of the most costly-rearrangements, which are the ones representing the rate-determining free energy barriers.

9.2.1. Protocol and PMF calculation

Two independent umbrella sampling calculations using the $\Delta RMSD$ on the Relay-SH1 elements were performed: one using starting conformations extracted from the forward SMD simulation along $\Delta RMSD$, and one using conformations extracted from the corresponding backward simulation. In each case, the range of the collective variable was discretized into 67 windows. In each window, a (time-independent) harmonic potential with force constant $k = 200.0$ kcal/mol/Å^2 centered on the window center was applied.

The free energy profile along $\Delta RMSD$ was computed either using WHAM or our own implementation of UI, with gave virtually indistinguishable results. Two-dimensional WHAM was used to evaluate free energy profiles along unbiased collective variables.

9.2.2. Lack of convergence of the individual profiles

As shown on Figure 9.15, it is clear that the umbrella sampling simulations are not converged since the forward and backward profiles are completely different,despite the satisfactory overlap between adjacent windows (Figure 9.17). For the forward simulation, Umbrella sampling is indeed successful in relaxing the non-equilibrium work profile, but only up to a certain point; in particular, no free energy minimum is clearly identified where the PTS state should be located. For the backward simulation, there is no striking difference between the profiles before and after relaxation by US, and no minimum is located at the PR state.

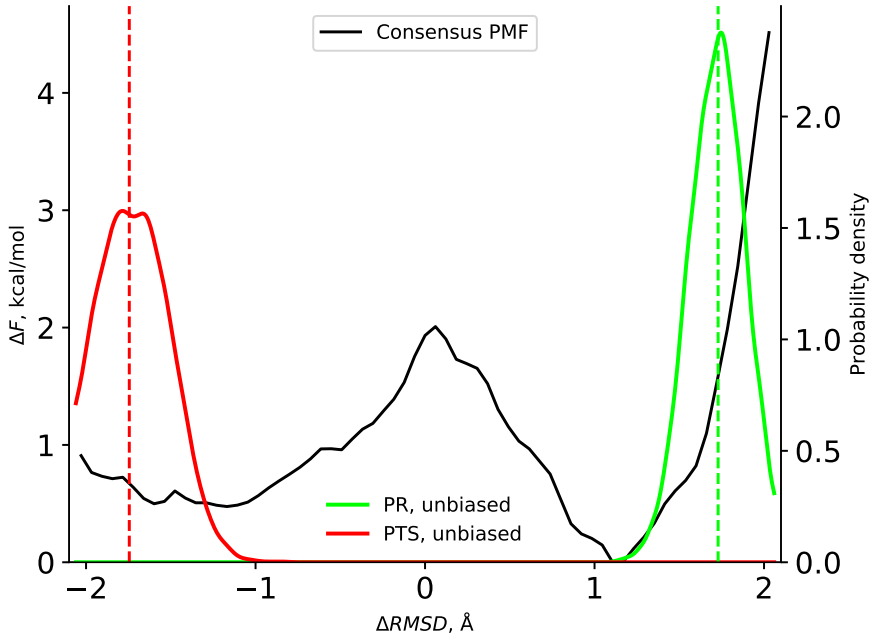


Figure 9.16.: "Consensus" potential of mean force along $\Delta RMSD$ of the Relay-SH1 elements for the PR → PTS transition, along with distributions of $\Delta RMSD$ from unbiased simulations of PR and PTS.

9.2.3. Combining the two profiles?

In the forward and backward umbrella sampling simulations, the same observable is biased on the same system; one may also decide that these two simulations are actually only one and combine their outcome so as to compute a "consensus" free energy profile. This profile is shown on Figure 9.16; interestingly, it is a double well potential with a small free energy barrier from PR (2 kcal mol^{-1}). In addition, the PTS basin appears wider than the PR one, which suggests entropic stabilization of PTS.

However, comparison with distributions of $\Delta RMSD$ from unbiased MD simulations shows that the consensus PMF is inconsistent with these latter, since the identified minima do not match the highest probability values measured in unbiased simulations. As seen on Figure 9.16, this is particularly true for the PR state.

In practice (and as evidenced by the significant difference between the individual free energy profiles, Figure 9.15) it is highly likely that both sets of simulations are sampling independent regions of the configurational space along the orthogonal degrees of freedom. This is very clear when one looks, for example, at the number of contacts between the converter and the N-terminal subdomain: if the forward and backward simulations were properly converged they should exhibit the same contact pattern as a function of $\Delta RMSD$, but this is not the case (data not shown). Arguably, the combination of the two sets of US simulations would be relevant if transitions along these orthogonal degrees of freedom were sampled. A possible strategy to achieve this would be to use bias-exchange umbrella sampling on the full set of 2x67 windows. We did not try this approach, in part because of practical considerations, *e.g.* the fact that the number of water molecules and ions in the PR and PTS simulation boxes are not identical, which affects the value of the total potential energy. It is unclear whether applying the Metropolis criterion for replica exchange in this situation is relevant.

At any rate, the above results showed that the global $\Delta RMSD$ is not a good transition coordinate

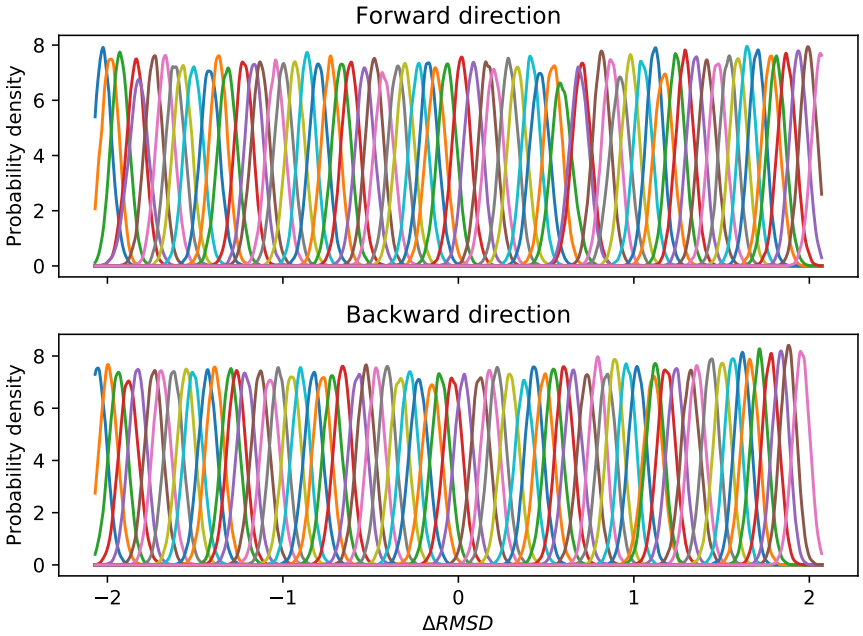


Figure 9.17.: Statistical distributions of the $\Delta RMSD$ collective variable in Umbrella Sampling demonstrate overlap between adjacent windows.

model to describe the PR →PTS transition. Indeed, the $\Delta RMSD$ of an entire sub-domain is probably too degenerate a reaction coordinate to ensure that the sampling is restricted in the vicinity of the minimal free energy path.

Consequently, we further tackled the problem with alternative strategies: two-dimensional extended ABF on a different set of transition coordinate models (next subsection), and more globally, with the string method in collective variables (see Chapter 11).

9.3. Extended ABF analysis of the coupling between converter swing and formation of the kink

Beyond its mechanistic interest, the computational exploration of the PR →PTS transition is expected to clarify whether this transition is more likely to initiate the recovery stroke than switch II closure. If this were the case, it would provide strong support for the PTS hypothesis since it would suggest that the PTS basin is attained faster than the hypothetical *Fischer putative intermediate* (FPI) state, in support of the PTS hypothesis.

Structural comparison complemented by the previously outlined TMD/SMD study have shown that the PR →PTS transition mostly consists in the conformational transition of the Relay-SH1 elements and the movement of the converter. As such, an appropriate pair of collective variables, each describing one of these rearrangements, could be used to compute the free energy landscape of the transition. We now report on this calculation, performed with the extended ABF approach.

9.3.1. Choice of collective variables

This calculation is challenging, notably because both rearrangements are complex and require more than one collective variable each for a proper description. Regarding the converter, it is apparent that most of the positional variation is accounted for by the X' variable, which changes by about 15 Å from PR to PTS, while Y' and Z' are nearly left unchanged on average. Thus, X' is a good candidate for a transition coordinate model for the converter swing in the PR →PTS transition. The case of the Relay-SH1 elements is more delicate: its transition consists in two global angular re-orientation of helices and a dramatic rearrangement of backbone hydrogen bonds (formation of the kink in the Relay helix); furthermore, our past attempt with umbrella sampling has revealed that using a global $\Delta RMSD$ is improper. Thus, short of using several collective variables for the Relay-SH1, one has to make a choice as to which elementary rearrangement is most important[3]. Based on the above discussion of the TMD/SMD trajectories and physical intuition, we reasoned that the rate-limiting rearrangement for the Relay-SH1 transition was likely to correspond to the formation of the kink in the Relay helix. Indeed, it involves the breaking of several backbone hydrogen bonds and one could argue that its transition state probably exhibits a locally unfolded backbone, where the pre-kink hydrogen bonding pattern is disrupted, but the post-kink pattern is not formed yet. Such a transition state is expected to be high in free energy. Based on these considerations, we chose to focus on a local description of the kink rather than a full description of the Relay-SH1 transition. Even then, at least 4 distances (the 4 k_i reported in Table 9.1) are needed to properly account for the hydrogen-bonding exchange during the kink formation (2 bonds form and 2 bonds are broken); this is still out-of-reach. Instead, we used the local $\Delta RMSD_{kink} = RMSD_{kink,PTS} - RMSD_{kink,PR}$, taken with respect to the backbone atoms in the kinking regions (see 6.2.2.2). Although this collective variable may still suffer from the issues associated with $\Delta RMSD$-type variables, it is defined in a very local manner on a rather small set of atoms. We expect that it will provide a reasonable description of the kink, which is actually confirmed by preparatory SMD simulations (data not shown).

9.3.2. Free energy calculation protocol

Initially, we wanted to use conventional ABF to map the free energy landscape over $(X', \Delta RMSD_{kink})$. This choice was essentially motivated by 1) the "good properties" of ABF already discussed in Chapter 8 and 2) the will to have as close a set-up as possible to the calculations on ATPase-activation, so as to justify the comparison of the free energy barriers between the two sets of calculations. Unfortunately, $\Delta RMSD_{kink}$ as it is defined involves overlapping sets of atoms (since the same set of atoms is used to compute the RMSD with respect to the PR and PTS references), which violates the orthogonality requirement of two-dimensional conventional ABF. Moreover, the use of a 1 fs timestep for our previous ABF calculations on switch II arguably entailed a significant cost in computational resources. Both these restrictions are however lifted with extended ABF[4], at the expense of having to choose harmonic coupling constants. SMD simulations (reported above) showed that a 10 kcal/mol/Å^2 force

3. Another possibility may have been to use the Generalized Adaptive Biasing Force (gABF) approach, which allows for the use of many collective variables in a PMF calculation, see 4.3.4.3. We were not aware of this method when these calculations were prepared; in addition, it is unclear how gABF would have performed in this situation as the various CV used to describe the Relay-SH1 (Relay helix and SH1 helix orientation angles and some local observable describing the kink) are not expected to be weakly coupled. Finally, our subsequent string method analysis of the transition, which is in progress at the time of writing, explicitly includes all these degrees of freedom.

4. eABF could very well have been used to perform the PMF calculations on switch II. The reason it was not is that at the time, we did not understand its underlying formalism. The eABF calculations discussed here were undertaken after we had improved our theoretical understanding of the method.

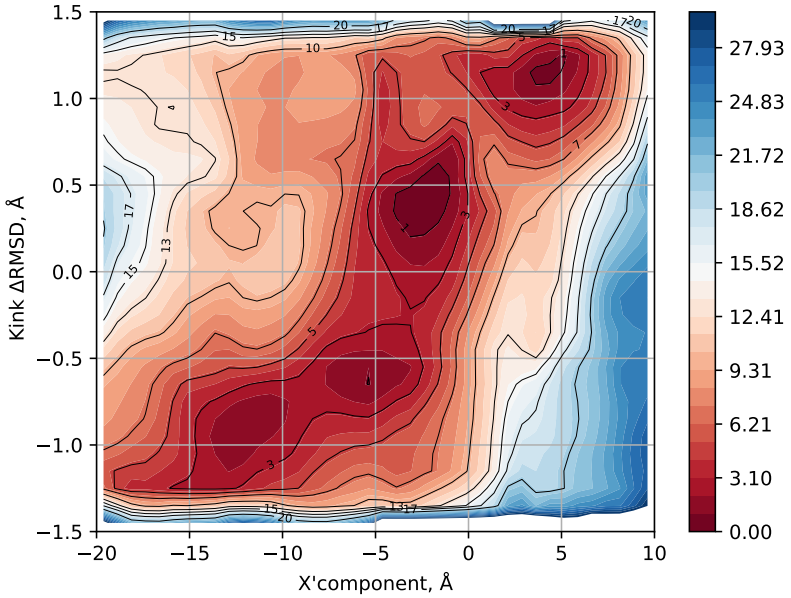


Figure 9.18.: Two-dimensional PMF along X' and $\Delta RMSD_{kink}$ obtained by eABF calculation. Free energies are given in kcal mol^{-1}.

constant was enough to drive the movement of the converter along X', and this value was retained for the coupling constant of this CV. Regarding $\Delta RMSD_{kink}$, short SMDs using a range of force constants were launched and we chose the minimal force constant capable of driving the formation of the kink, *i.e.* 125 kcal/mol/Å^2 (data not shown). The friction constant for the extended dynamics was set to 10 ps^{-1} for both extended degrees of freedom.

Like before, we applied a two-step strategy, *i.e.* an exploratory run followed by stratification. Since the conformational changes under study are large-scale domain movements/rearrangements, we performed a significantly longer exploratory run (700 ns) than in the ATPase activation case (70 ns). This exploratory run was started from the equilibrated PR structure, and the *fullSamples* parameter was set to 2000.

The configurational space was then divided into 4x3 non-overlapping windows of identical size; each window was initialized with the gradient estimate obtained from the exploratory run and atomic coordinates from the configuration closest to the window center sampled during the exploratory run. Also, an ”ABF equilibration” run in which the eABF bias accumulated during the exploratory run was applied but not updated was performed for 1 ns for each window, after which standard eABF was run for 220 ns per window. Harmonic walls acting on the extended degrees of freedom were used to confine the dynamics in each window. The full gradient estimate acting on the extended degrees of freedom is obtained by piecing together the gradient estimates from each window. Then, the CZAR estimator is used to recover the gradient estimate acting on the collective variables of interest.

9.3.3. Results and Discussion

The free energy landscape resulting of the two-step eABF calculation is shown on Figure 9.18. It reveals several local free energy minima, including the PR basin (upper right corner).

9.3.3.1. "Agnostic" identification of the PTS state

Strikingly, a wide free energy basin encompassing at least two local minima is detected in the lower left region of the free energy landscape, *i.e.* where one expects to find the PTS state.

X' is used as an order parameter to distinguish between PR and PTS, because its typical value goes from 5 Å in PR to -12 Å in PTS. However, PTS and PPS exhibit similar values of X'. Thus, the local free energy minimum located at $X' \simeq -12$ Å could encompass not only PTS-like conformations, but may also include contributions from PPS-like conformations. Y' and Z', rather than X', account for the PTS → PPS transition. Thus, to assess whether PPS-like configurations are sampled in the free energy basin, one should monitor the distribution of Y' and Z'. Figure 9.21 (bottom left and right) shows that, for windows which correspond to the putative PTS-free energy basin, PPS-like values of Y' and Z' are sampled only rarely, which shows that this basin is likely not to be "contaminated" by PPS-like conformations.

Furthermore, the projection of the unbiased MD trajectories onto the PMF shows that the putative PTS basin is indeed consistent with converter positions sampled by unbiased PTS simulations (Figure 9.22). By contrast, this is clearly not the case for PPS. Finally, switch II remains open throughout the eABF simulation, for all windows including the ones belonging to the putative PTS-basin (data not shown). These observations support the conclusion that the free energy basin identified by eABF calculations is indeed representative of the PTS state. This is an important result, because the simulation was initiated from the PR structure (rather than PTS) and is nearly "PTS-agnostic", *i.e.* almost no beforehand knowledge of PTS is injected in the simulation design. The only knowledge of PTS comes from the definition of $\Delta RMSD_{kink}$, where the PTS structure of the kinked region is used as a reference. However, the configurations adopted by the atoms involved in PTS and PPS are virtually indistinguishable, as the RMSD between PTS and PPS is 0.32 Å. It seems reasonable to expect that the same calculation devised using PPS instead of PTS as a reference for $\Delta RMSD_{kink}$ would have yielded very similar results. As such, the "*de novo*" identification through free energy calculations of PTS as a metastable conformational state accessible from PR is a strong argument in support of the PTS hypothesis. Also, it suggests that the PTS "state" sampled in unbiased MD may in fact correspond to at least two distinct metastable intermediates, separated by a low 3 kcal mol^{-1} barrier.

9.3.3.2. Dominant pathway and barrier of the PR→PTS transition

The free energy landscape also reveals that the transition from PR to PTS proceeds through a nearly diagonal transition tube, in which the movement of the converter and the formation of the kink are rather strongly coupled. Interestingly, yet another intermediate is predicted along the PR to PTS transition (X'=$-$2.5 Å, $\Delta RMSD_{kink}$=0.5 Å) which seems to be close in free energy to the PR state, but separated from it by a rather large 7 kcal mol^{-1} barrier. Consistently, this intermediate, whose structural characterization would be a natural direction of investigation, is never explored in unbiased simulations. Finally, we remark that the aforementioned barrier happens to be the highest one along the dominant pathway for the PR → PTS transition; and, we immediately notice that it is significantly lower than the reported 12 kcal mol^{-1} free energy barrier for the closure of switch II from PR (see Chapter 8). Thus, the present eABF calculations suggest that when the motor domain is in the PR state, it is faster to jump to PTS through a concerted rearrangement of the converter and Relay helix, than to close switch II. In other words, these results support the ratchet-like mechanism of the recovery stroke and challenge switch II-initiated scenarios. Also, we note that the proposal that the movement of the converter precedes the rearrangement of the Relay helix, made earlier on the basis of SMD simulations, is not confirmed; instead, eABF calculations suggest that both rearrangements are tightly

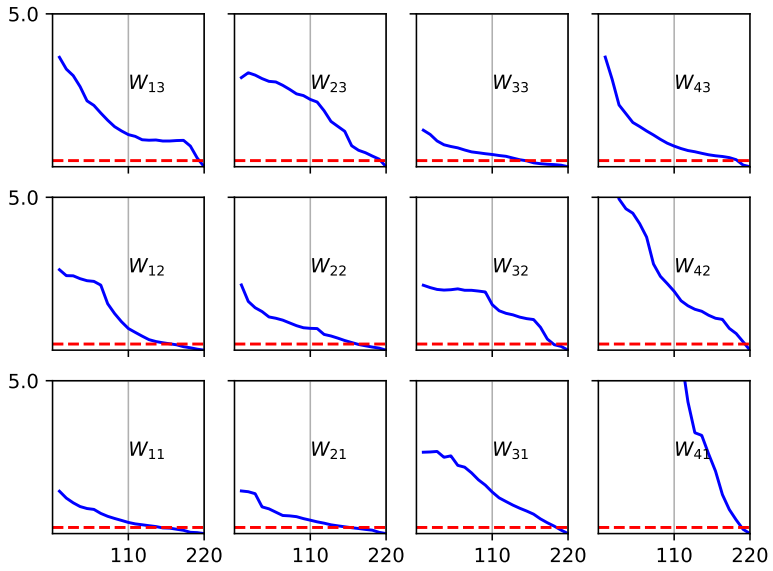


Figure 9.19.: RMSD of the gradient estimate for each window during the stratified run. X-axis, time of stratified simulation in ns; Y-axis, gradient RMSD, in kcal mol^{-1} Å^{-1}. The red dotted lines show the 0.5 kcal mol^{-1} Å^{-1} cut-off which may be used to evaluate convergence.

coupled.

9.3.4. Convergence and error analysis

9.3.4.1. Gradient convergence

To evaluate the convergence of the average force estimate, which conditions that of the resulting PMF, we computed the Root-Mean-Square Deviation (RMSD) of the gradient over each full window, with respect to its value at the end of the simulation. The results are reported on Figure 9.19.

Clearly, several windows did not achieve proper convergence as the gradient RMSD does not plateau at the end of the simulation. However, it seems that most windows corresponding to the ”upper right to lower left” diagonal, *i.e.* these which correspond to the dominant transition pathway identified by the calculations, exhibit reasonable convergence as the RMSD stabilizes to values lower than 0.5 kcal mol^{-1} Å^{-1} in the latest stages of the simulation. This suggests that extending the calculations may be required to obtain a fully converged free energy landscape, but that the gradient estimate in the relevant region of the PMF is reliable. We note that this analysis is rather preliminary, notably because averaging over large windows results in loss of resolution; inhomogeneous convergence within a window cannot be assessed by this approach.

9.3.4.2. Configurational space coverage and orthogonal degrees of freedom

The inspection of the number of counts reveals that the full grid is everywhere sampled by at least one order of magnitude more than *fullSamples* (2000). However, it also reveals a rather imbalanced

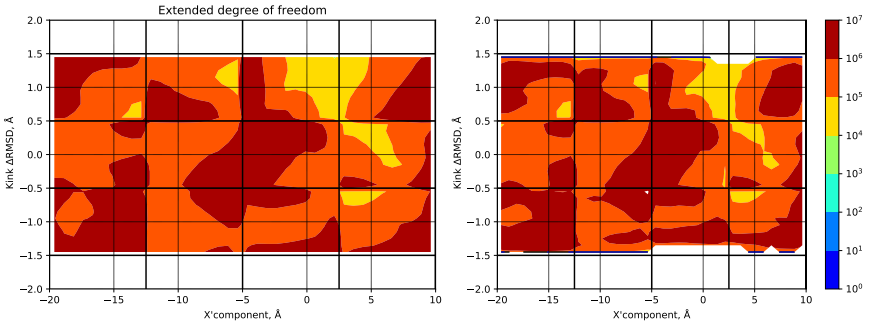


Figure 9.20.: Coverage of the configurational space during eABF simulation. The number of counts, *i.e.* the number of times a given grid point has been visited, is represented in logarithmic scale. Left, extended degree of freedom; Right, actual values of the collective variables. The thick black lines materialize the window boundaries.

sampling which remains even after stratification; in particular, the PR basin seems less sampled than other regions of the grid, despite the use of the stratification strategy.

In fact, the projection of the stratified trajectories onto the $(X', \Delta RMSD)$ map (Figure 9.21, upper left) reveals that some windows do not achieve full coverage, suggesting that the diffusive regime (*i.e.* convergence) has not yet been attained in these windows. On the other hand, we note that orthogonal degrees of freedom which are nonetheless important to describe the PR → PTS transition are "well-behaved" in the sense they take on values consistent with the expectations (Figure 9.21) ; for example, in the lower left window of the $(X', \Delta RMSD_{kink})$ map, which is one of these which represent the PTS state, the distributions of Y', Z', θ_{RH} and θ_{SH1} are indeed consistent with PTS values, suggesting that a proper transition to PTS has been captured.

9.3.4.3. Comparison with unbiased MD

The projection of the statistical distributions of $(X', \Delta RMSD_{kink})$ from unbiased simulations reveals very reasonable agreement with the free energy basins predicted by the eABF calculation. Notably, unbiased PTS simulations match remarkably well with the extended free energy basin located by eABF. Moreover, the projection of a PPS simulation shows clearly that this free energy basin is incompatible with PPS; we conclude that the likelihood for this basin to actually correspond to PTS is very high.

We note that one PTS simulation exhibits a density maximum which does not correspond to a free energy minimum according to the eABF results (simulation PTS+ATP (1), maximum at X'=−16 Å and $\Delta RMSD_{kink} \simeq -1.4$ Å). But, this state corresponds in fact to the PTS-reprimed state, in which the converter is docked onto the N-terminal sub-domain (see Chapter 6). No such (orthogonal) transition to the PTS-reprimed state is explored in the eABF trajectory, which explains why it is not located as a free energy minimum. Also, we recall that in the PR+ATP (3) simulation, the converter spontaneously uncouples from the N-terminal sub-domain and reaches a new position, stable on the simulation timescale (>100 ns) but with a straight Relay helix. Arguably, this state should have been located as a local minimum by the eABF calculation, which is actually not the case. This may indicate lack of convergence of the PMF in the vicinity of the PR basin, which was also suggested by the relative lack of sampling in this region evident on Figure 9.20.

Overall, the agreement of the free energy landscape and the unbiased simulation data, and the "correct" behaviour of the important orthogonal degrees of freedom, are in favour of the proper convergence of our calculations even though some inconsistencies remain. Extra analyses (starting with a

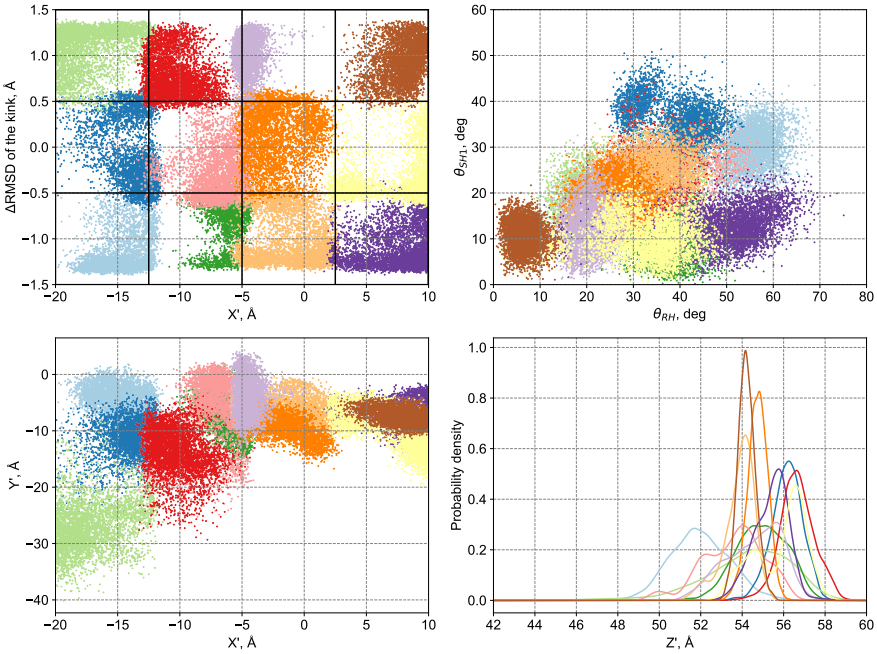


Figure 9.21.: Behaviour of selected orthogonal observables during the stratified eABF runs.

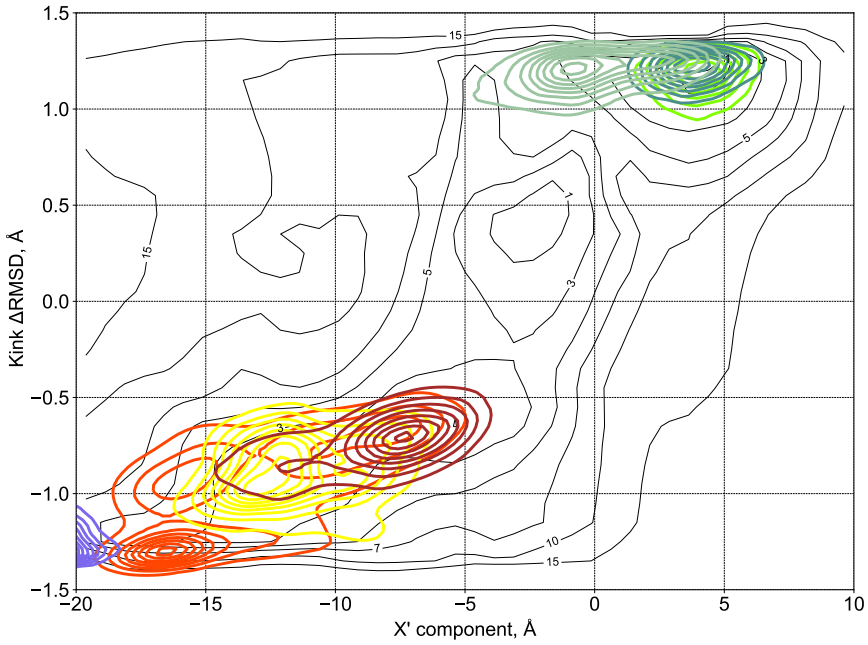


Figure 9.22.: Projection of unbiased MD trajectories onto the free energy landscape. The color code matches that of Chapter 6; green density lines correspond to PR simulations, orange/yellow lines to PTS, and blue lines to PPS.

deeper assessment of the convergence of the gradient estimate as a function of simulation time), and possibly independent replica-simulations are nonetheless needed to make sure that this is the case.

9.4. Conclusion

The eABF calculations discussed above are the outcome of our study of the PR → PTS transition mechanism. Despite their proper convergence being somewhat uncertain, their remarkable agreement with unbiased simulations of the motor domain supports their significance. In summary, these calculations highlight that the PR → PTS proceeds by a concerted converter swing/kinking of the Relay helix rearrangement as already foreshadowed by TMD/SMD simulations; crucially, they provide strong support for the PTS hypothesis both through a *de novo* identification of the PTS state, and the finding that the PR → PTS transition is likely to be faster than switch II closure. Whether this mechanism is specific or not to myosin VI is unclear; equivalent eABF calculations on Dd myo2 are ongoing at the time of writing to investigate this important question.

Our interpretation, nevertheless, suffers from several limitations even assuming that the eABF calculations are perfectly converged. First, the *de novo* identification of PTS is encouraging as it provides independent validation for the crystal structure; but, since our calculations do not (by design) account for the PTS → PPS transition, we are not in a position to claim that PTS is indeed an *intermediate* along the recovery stroke, and not an off-pathway state. Recall, however, that a spontaneous PPS → PTS transition was observed in aMD simulations (Chapter 7), suggesting that transitions between PTS and PPS are possible. More importantly, our kinetic argument in favour of the ratchet-like scenario rests on a comparison of free energy barriers obtained using different sets of collective variables; whether such a comparison is in fact legal is unclear. Also, even if the PR → PTS transition were indeed faster than early switch II closure (*i.e.* the transition from PR to the Fischer Putative Intermediate (FPI) state), the possibility is not eliminated that the PTS → PPS transition be much slower than the FPI → PPS transition, in such a way that Fischer's switch-II-initiated mechanism would represent the fastest pathway for the complete recovery stroke.

Settling the question thus requires a comparison on an equal footing (same collective variables) and taking into account the full recovery stroke transition. Optimal pathway calculations using the String Method in Collective Variables (CVSM) offer an attractive way to fulfill these requirements, but a better understanding of the PTS → PPS transition is needed before progressing further. This is the topic of the next chapter.

10. Mechanism of the PTS → PPS transition

Summary After our extensive study of the PR → PTS transition, we now turn to the second part of the recovery stroke in our emerging model, namely the PTS → PPS transition. Unlike the first step, we fail to identify a single sub-domain whose rearrangement would drive the full transition. Arguably, this is because this transition is more complex and involves virtually all the sub-domains of the motor domain. We report on the design of structural observables to characterize the individual rearrangements and drive them in SMD simulations. Combining these SMDs with previous unbiased simulations, we identify a novel rearrangement involving a switching in the interaction pattern of β-strands 4,5 and 6 of the central myosin β-sheet (or transducer) controlling switch II closure, revealing a previously undescribed role for the transducer in the recovery stroke. These results are unpublished.

10.1. Overview of the transition

Comparison of crystal structures shows that the PTS to PPS transition includes the following sub-transitions:

- Closure of switch II
- Inward rotation of the L50 subdomain and closure of the inner L50/U50 cleft
- Completion of converter rotation
- Seesaw motion of the Relay helix
- Internal conformational transition of the converter

This global transition involves essentially all the relevant sub-domains for the recovery stroke, and is expected to be more complex than the PR → PTS transition (which essentially involves only the converter and the Relay-SH1 elements, as shown in the previous chapter). Encouraged by the success of specialized TMD/SMD for this previous transition, we attempted similar approaches to unravel the coupling between the various elementary rearrangements. However, this was unsuccessful in identifying a particular sub-domain which would control the full transition. For instance, driving the closure of the active site from PTS using a TMD bias does result in some L50 rotation, but in no detectable movement of the converter (see summary on Figure 10.1).

These negative results suggest that no strong coupling may exist between the elementary rearrangements in the PTS → PPS transition, and that they should rather be investigated individually.

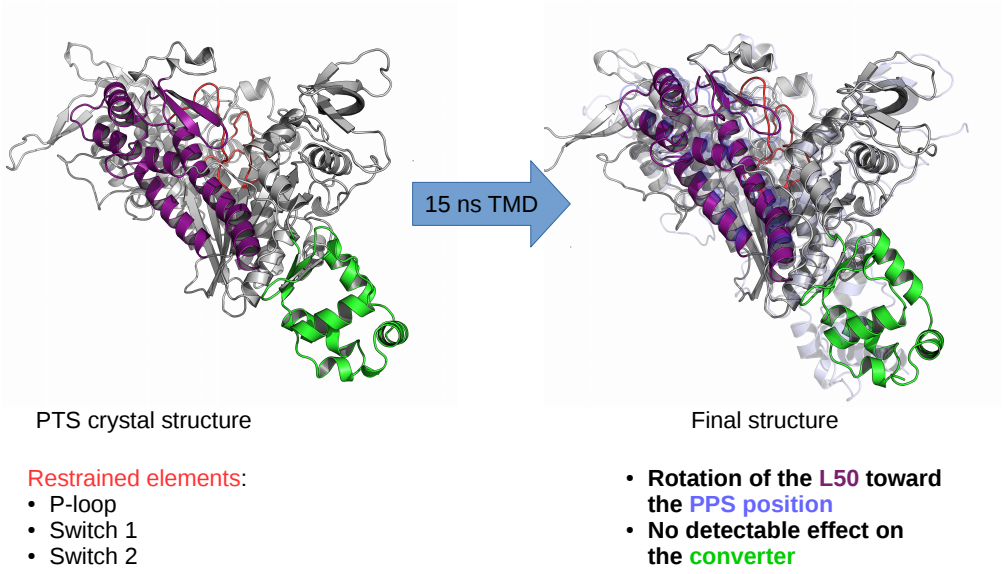


Figure 10.1.: Closure of the active site from PTS by TMD does not result in PPS-like converter positions. From PTS, the flexible loops of the active site were restrained towards the PPS conformation over a 15 ns TMD simulation with force constant 200 kcal/mol/Å^2. Comparison with the PPS crystal structure shows a rotation of the L50 towards PPS, but no clear converter movement.

10.2. A rearrangement of the β-sheet interaction pattern controls the closure of Switch II

10.2.1. Insight from unbiased MD trajectory

As reported in Chapter 6, unbiased MD simulations of the PPS+ATP state show a re-opening of switch II. By the principle of microscopic reversibility, we may argue that the sequence of events for this re-opening, when reversed, provides a possible mechanism for the closure of switch II during the PTS to PPS transition. Thus, in the following, we describe this sequence of events and analyze its implications for the PTS to PPS transition.

For this purpose, we consider the CA-RMSD of the active site with respect to the PPS crystal structure, where the active site includes the three consensus elements of the Walker motif along with neighbouring secondary structure elements, notably the associated β-strands of the central β-sheet:

- P-loop, preceding strand (strand 4, $\beta 4$) and following helix: residues 145 to 165
- Switch I, preceding helix and following strand (strand 6, $\beta 6$): residues 190 to 215
- Switch II and preceding strand (strand 5, $\beta 5$): residues 450 to 468

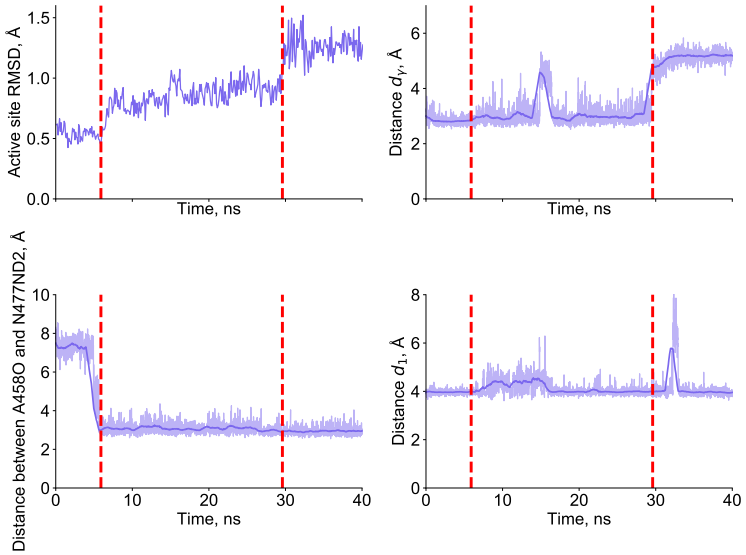


Figure 10.2.: Evolution of the RMSD of the active site and of several active site distances during unbiased MD simulation PPS+ATP (1). The dotted lines materialize events 1 and 2. For distances, the 1 ns running average is shown (thick line) over the raw data.

We analyze the time-series of this RMSD from unbiased simulation PPS+ATP (1) (Figure 10.2, top panel) and describe the corresponding structural rearrangements.

At the beginning of the simulation, the conformation of the active site is very close to the one observed in the crystal structure of the PPS state. On the RMSD time series, one distinguishes three main stages separated by transition events at $t \simeq 6$ ns and $t \simeq 30$ ns, each producing an increase in RMSD of about 0.3 Å. Also, between these two events, the RMSD progressively increases from about 0.8 Å to 1.0 Å. This suggests that there are two major rearrangements of the active site elements (termed event 1 and event 2) which, as illustrated on Figure 10.2 (middle panel), ultimately lead to the breaking of the switch II-ATP hydrogen bond.

Clearly, the second event (at $t \simeq 30$ ns) corresponds to a 2 Å seclusion of G459N from ATP which completely breaks the hydrogen bond. However, a closer inspection reveals that the first event at $t \simeq 6$ ns seems to produce 1) a slight increase, on average, of the distance d_γ (2.8 Å vs 3.0 Å) and 2) an increase in its fluctuations (Figure 10.2, top right, between events 1 and 2). Even more, the hydrogen bond is transiently broken for a few ns at $t \simeq 15$ ns. These observations suggest that the first event actually destabilizes the hydrogen bond.

Visual inspection shows that event 1 begins with the seclusion of the C-terminal of switch II from switch I. Its first consequence is the breaking of the Y462/E152 side chain/side chain interaction, *i.e.* a P-loop/switch II interaction. This seems to weaken the hydrogen bond between Y462O and the side chain of R199 (a switch II/switch I interaction). Eventually, this hydrogen bond breaks.

Interestingly, at the beginning of the simulation, N477 on the Relay helix is in such a rotameric state that it does not interact with switch II. A rotameric transition of N477 to form an interaction with the backbone of G459-F460 is concomitant to event 1, see Figure 10.2, bottom left. We note that upon event 1, the side chains of R205 and E461 are slightly displaced, but the critical salt-bridge is maintained (Figure 10.2, bottom right). Overall, event 1 seems to correspond to a change in the interaction pattern between switch II and its surroundings, with the breaking of P-loop/switch II and switch I/switch II interactions, and the formation of a switch II/Relay helix interaction.

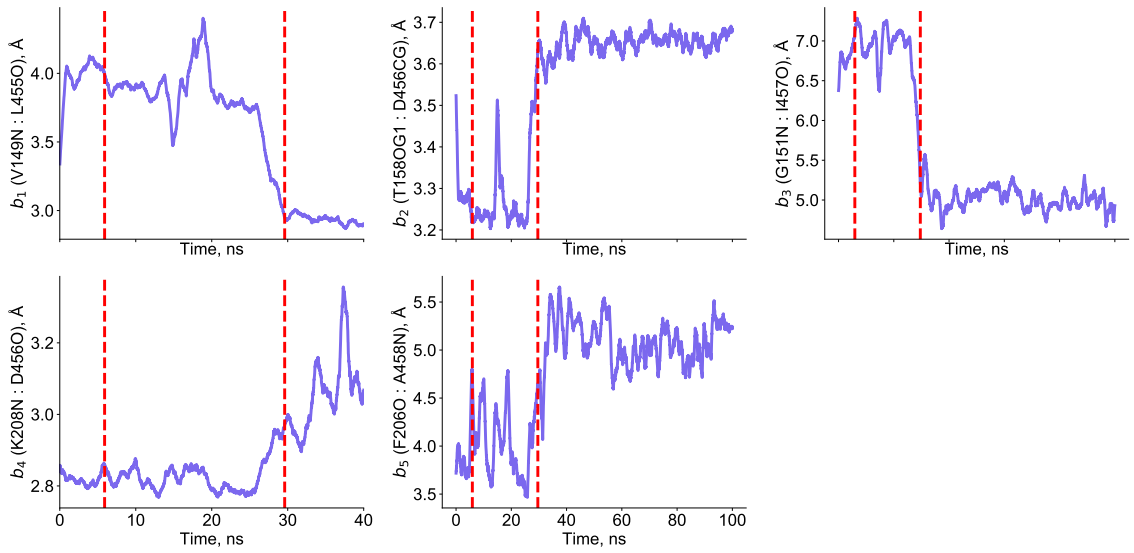


Figure 10.3.: Evolution of the b_i distances (see main text) during unbiased MD simulation PPS+ATP (1). The dotted lines materialize events 1 and 2. The 1 ns running average is shown.

At $t \simeq 30$ ns (event 2), the β-sheet backbone hydrogen bond between V149N and L455O, which was broken at the beginning of the simulation, reforms, effectively re-coupling β-strands 4 and 5 of the transducer (distance b_1 on Figure 10.3). Simultaneously, the switch II-ATP hydrogen bond breaks. Upon the reformation of the $\beta4/\beta5$ interaction, there is a clear seclusion between $\beta5$ and $\beta6$, as evidenced by distances b_4 and b_5 on Figure 10.3). This suggests that $\beta5$, the strand which precedes switch II, cannot establish an optimal hydrogen bonding pattern with both $\beta4$ and $\beta6$ at the same time; local rearrangements of the transducer by "switching" the interactions of adjacent strands seem to control the position of switch II. "Passing the movie in reverse", this suggests that the formation of the switch II-ATP hydrogen bond (while the critical salt-bridge is already formed) would be driven by the un-coupling of $\beta5$ from $\beta4$ and its subsequent coupling to $\beta6$. Also, it suggests that further rearrangements are needed to form the hydrogen bond between switch II and the Relay helix while switch II is closed on ATP.

10.2.2. Biasing the β-strand switching

10.2.2.1. SMD protocol

The analysis of spontaneous switch II opening from PPS points to a previously un-described role of the β-strands of the transducer in controlling the state of switch II. It seems that the switch II-ATP hydrogen bond cannot be formed while the $\beta4/\beta5$ interaction (V149:L455) is formed, and reciprocally. This prediction can be tested by SMD: driving the disruption of the $\beta4/\beta5$ interaction from PTS should result in the formation of the switch II-ATP hydrogen bond. However, when we performed this simulation, nothing clear happened to switch II (data not shown). We decided to widen the set of biased distances by considering distances in the active site region (as defined above) which change significantly during switch II re-opening, and which involve residues belonging to $\beta5$ and either $\beta4$ or $\beta6$. 5 such distances were identified, as reported on Figure 10.3; these distances are called b_i.

The b_i distances were simultaneously biased with the following SMD protocol. Starting from the PTS equilibrated structure, a 15 ns constant-velocity SMD was applied, targeting each distance to

Distance	**Initial value (Å)**	**Target value (Å)**	**Force constant (**$\mathrm{kcal/mol/Å^2}$**)**
b_1 (V149N:L455O)	3.01	3.91	49.3827
b_2 (T158OG1:D456CG)	3.51	3.26	640
b_3 (G151N:I457O)	4.52	6.95	6.77404
b_4 (K208N:D456O)	2.97	2.82	1777.78
b_5 (F206O:A458N)	6.59	3.99	5.91716

Table 10.1.: Description of the biased distances in SMD simulation of the β-sheet rearrangements.

its average value measured during the first 30 ns of the unbiased PPS simulation, *i.e.* before the re-opening of switch II, see Table 10.1. Thus, it is expected that this simulation will capture the formation of the switch II-ATP hydrogen bond. Finally, to ensure that each distance feels a comparable biasing force, the individual harmonic force constants were rescaled according to the difference between the initial and target values of the corresponding distances, see Table 10.1.

Before describing the results of this simulation, several remarks are in order. First, for two distances (b_2, b_4) the range of variation is actually very small. This is because the interactions described by these distances are actually formed or nearly formed in the PTS equilibrated structure, whereas they are totally broken after switch II re-opening in the PPS unbiased simulation. Clearly, this challenges their relevance as determinants of switch II closure, but may also indicate imperfect equilibration of the PTS structure. We chose to keep these distances in the biasing set so as not to miss potentially important rearrangements. As a consequence of the very small variation, the rescaled force constants are very large for these distances, which may cause integration errors. No numerical instability of the simulation was observed, perhaps due to the stabilizing effect of the Langevin friction. Furthermore, we did not check whether such high force constants did preserve energy conservation in NVE dynamics, and we acknowledge that this is a likely possibility. However, the purpose of this simulation was to observe the structural response to the driven rearrangement in a qualitative (the switch II-ATP hydrogen bond forms, or not) rather than quantitative manner (we do not seek to estimate the associated free energy cost). Thus, we may argue that such a local perturbation, if it may change the details of the energetics, is unlikely to dramatically change the specifics of the structural response. Finally, several distances do not correspond to actual interactions (for example, b_3 or G151N:I457O, a P-loop/switch II distance, is never small enough that a hydrogen bond is actually formed). It is possible that other, actual interactions rearrange during the transition and that their values are correlated to the distances we have chosen to bias. Arguably, such distances would represent better candidates for biased simulation.

10.2.2.2. Results

The 15 ns SMD simulation did not result in the formation of the switch II-ATP hydrogen bond. However, this simulation was extended by a 100 ns "static bias" simulation, in which each biased distance was harmonically restrained to its target value with the same force constant as in SMD, using a time-independent potential. The behaviour of distances b_i during SMD and static bias simulations is shown on Figure 10.4.

Strikingly, at about t =60 ns, the switch II-ATP forms and remains stable for the remaining of the simulation (Figure 10.5). Interestingly, in the early stage of the simulation the distance between switch II and ATP decreases progressively, and stabilizes around 4.5 Å, which is also the value observed after switch II re-opening in the PPS unbiased simulation (Figure 10.2). Overall, these results support our

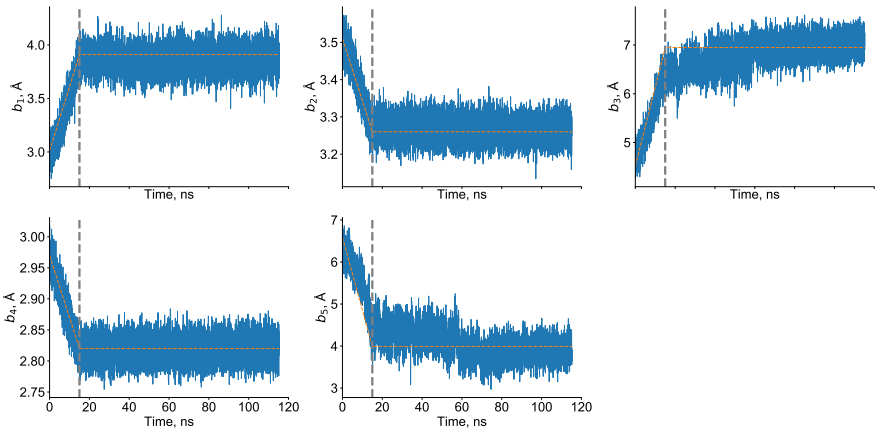


Figure 10.4.: Evolution of the biased distances b_i during SMD, then static bias simulation. The dotted line marks the transition to a static harmonic bias; the orange lines materialize the center of the bias. On average, each b_i follows the biasing potential.

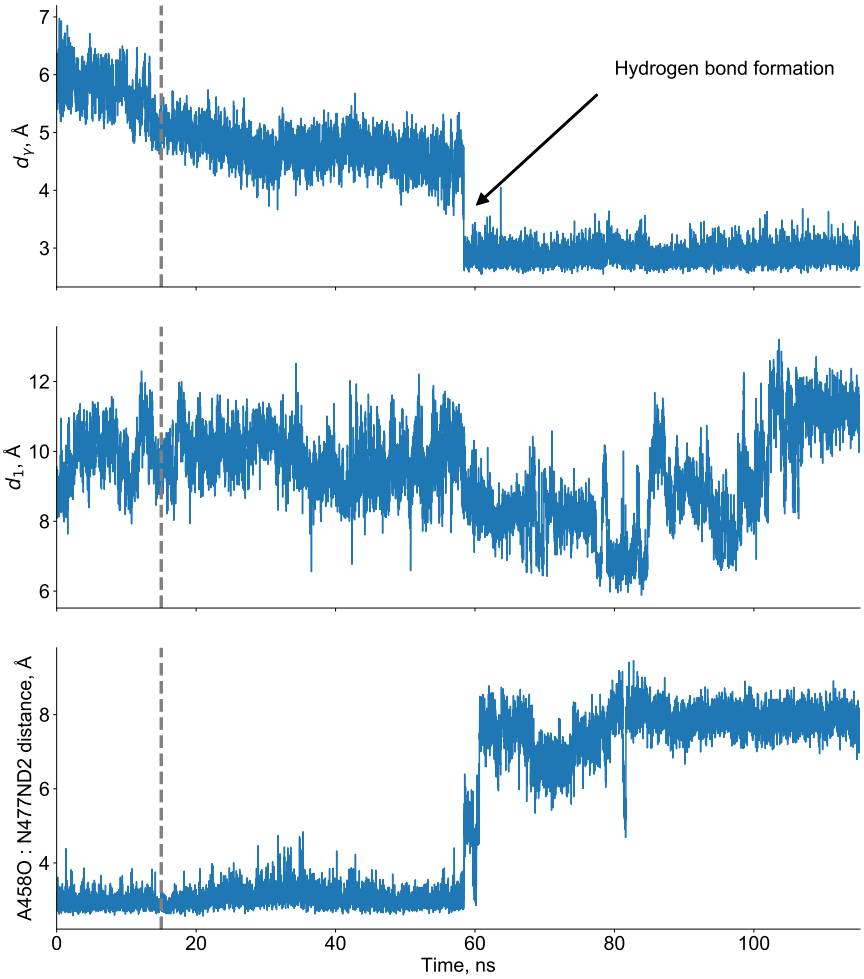


Figure 10.5.: Evolution of distances in the active site during SMD, then static bias simulation. The dotted line marks the transition to a static harmonic bias.

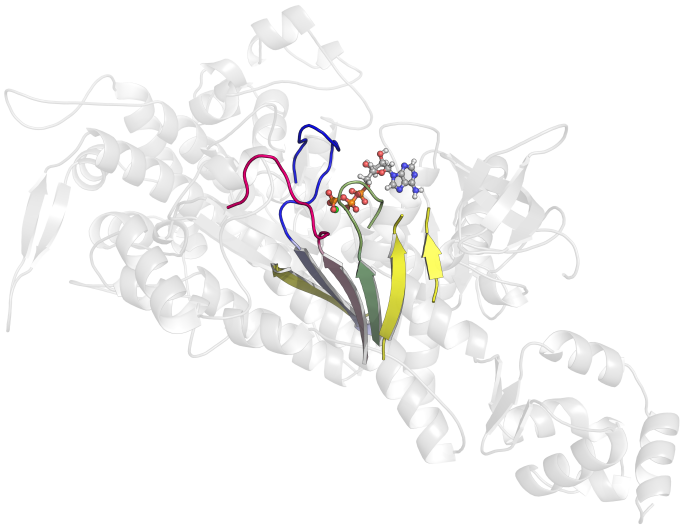


Figure 10.6.: Central myosin β-sheet and active site in the motor domain.

initial interpretation that the "switching" of the interaction of $\beta 5$ from $\beta 4$ to $\beta 6$ drives the formation of the switch II-hydrogen bond. See Figures 10.6 and 10.7 for a visual insight into the rearrangement.

Importantly, the formation of the hydrogen bond is not accompanied by that of the critical salt-bridge (Figure 10.5, middle panel) but is associated with the breaking of the polar contact between the side of N477 on the Relay helix and the backbone oxygen of A458 on switch II (Figure 10.5, bottom panel). In other words, the formation of the switch II-ATP hydrogen bond from PTS in this simulation entails the uncoupling of switch II from the Relay helix, similarly to what was observed in the ABF calculation on PTS reported in Chapter 8.

10.2.3. Conclusion

The model emerging from our analysis is that the inward motion of switch II which is necessary for its closure is facilitated by "switching" the position of the switch II-associated β-strand ($\beta 5$), from interacting with $\beta 4$ (P-loop-associated), to interacting with $\beta 6$ (switch I-associated), as sketched on Figure 10.8.

Rearranging the β-sheet interaction pattern probably represents a (the?) free energy barrier to switch II closure during the final stages of the PTS → PPS transition, because it is likely to involve a fully-uncoupled $\beta 5$ as a transition state. If this is indeed the case, it highlights a previously un-reported (to our knowledge) role for the transducer in the recovery stroke. Interestingly in the biased simulation discussed above, no clear rearrangement of the other structural elements involved in the recovery stroke (converter, Relay helix, etc) is detected. Thus, the mechanism by which the β-sheet rearrangement may be coupled to other elementary rearrangements remains to be understood. Assuming that the initiating event is the completion of the converter rotation (which is purely speculative, but may be justified by analogy with the PR → PTS transition), one may imagine the following scenarios, which are not mutually exclusive:

- Transducer-controlled pathway: the movement of the converter is transmitted to the transducer (for example through the Relay and SH1 helices) whose twisting state locally changes, leading

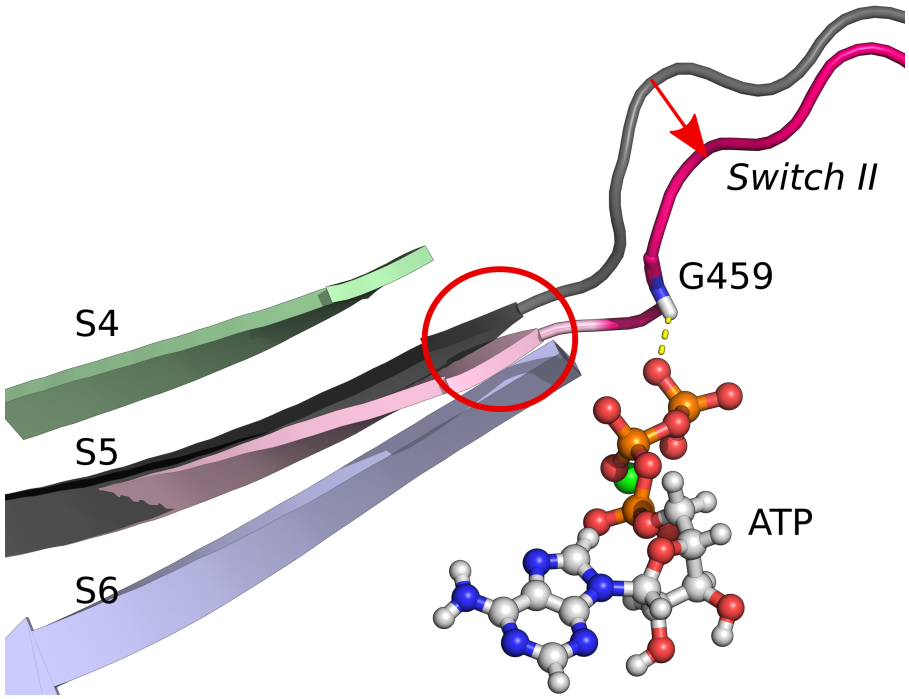


Figure 10.7.: Final conformation of the active site in the biased simulation from PTS. The arrow materializes the movement of the C-terminal tip of strand $\beta 5$ (S5), which translates into a larger motion of switch II allowing the formation of the hydrogen bond with ATP. In black is represented the position of strand $\beta 5$ and switch II at the beginning of the simulation (PTS state).

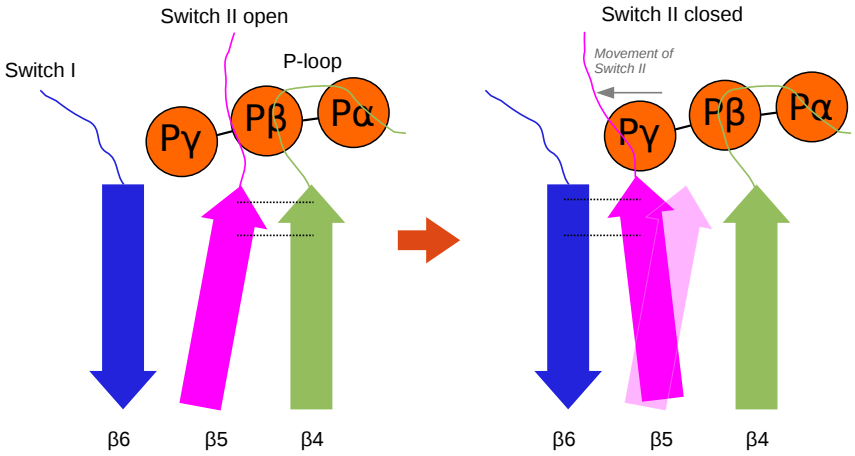


Figure 10.8.: The ”switching” of the interaction of $\beta 5$ from $\beta 4$ to $\beta 6$ controls the formation of the switch II-ATP hydrogen bond.

to β-strand switching.

- L50-controlled pathway: the movement of the converter is transmitted to the L50 subdomain (through the Relay Helix and Loop), whose inward rotation exerts a sufficient force on switch II to drive its full closure by breaking the β-sheet hydrogen bond.

Also, we note that one should explain the formation of the critical salt-bridge, which is not observed in the biased simulation and as such is unlikely to be controlled by the β-strand switching mechanism.

10.3. Seesaw motion of the Relay helix and L50 rotation

In this section, we report on the development of observables to describe two other important rearrangements during the recovery stroke, namely the seesaw motion of the Relay helix and the rotation of the L50 subdomain.

10.3.1. Seesaw motion of the Relay helix

The seesaw motion of the Relay helix is an important rearrangement of the recovery stroke, first proposed by Fischer and co-workers (Stefan Fischer, Windshügel, et al. 2005), see also Chapter 5. It corresponds to a rigid-body motion of the Relay helix, which notably brings its N-terminal region (residues 468 to 480) towards the inside of the nucleotide binding site. Thus, we defined observables s_1 and s_2 as distances between backbone atoms of the N-ter Relay helix and either the P-loop or switch I. More precisely, s_1 is the distance between the CA atoms of residues 151-153 (P-loop) and 475-479 (distal N-ter Relay helix), and s_2 is the distance between the CA atoms of residues 194-196 (switch I) and 468-472 (proximal N-ter Relay helix), see Figure 10.9.

The evolution of s_1 and s_2 during the unbiased MD simulations of Chapter 6 are shown on Figure 10.10. These results confirm that the Relay helix has not undergone the seesaw motion in PTS; in addition, they show that the seesaw tends to partially reverse in PPS, even though it is complete at the beginning of the simulations. This participates of the general instability of the PPS active site with ATP, as already reported (see Chapter 6 and the previous discussion of the β-strand switching mechanism).

When considering the PPS+ATP (1) unbiased simulation (already discussed at the beginning of this chapter), one sees that the reversal of the seesaw motion is clearly correlated to the breaking of the switch II-ATP hydrogen bond, but not particularly with the breaking of the critical salt-bridge (Figure 10.11).

To further investigate this correlation, 10 ns-long SMD simulations along s_1 and s_2 were performed from PTS to PPS with a range of force constants (Figure 10.12). Intriguingly, despite the seesaw motion of the Relay helix being successfully driven, no events of switch II closure are observed (data not shown).

Even though we have developed a set of observables to describe (and drive) the seesaw motion of the Relay helix, no coupling of this movement with switch II closure is detected, which contrasts with the findings of Fischer and co-workers, see Chapter 5 and (Stefan Fischer, Windshügel, et al. 2005; Koppole, J. C. Smith, and Stefan Fischer 2007). We reserve further investigations of this problem for later.

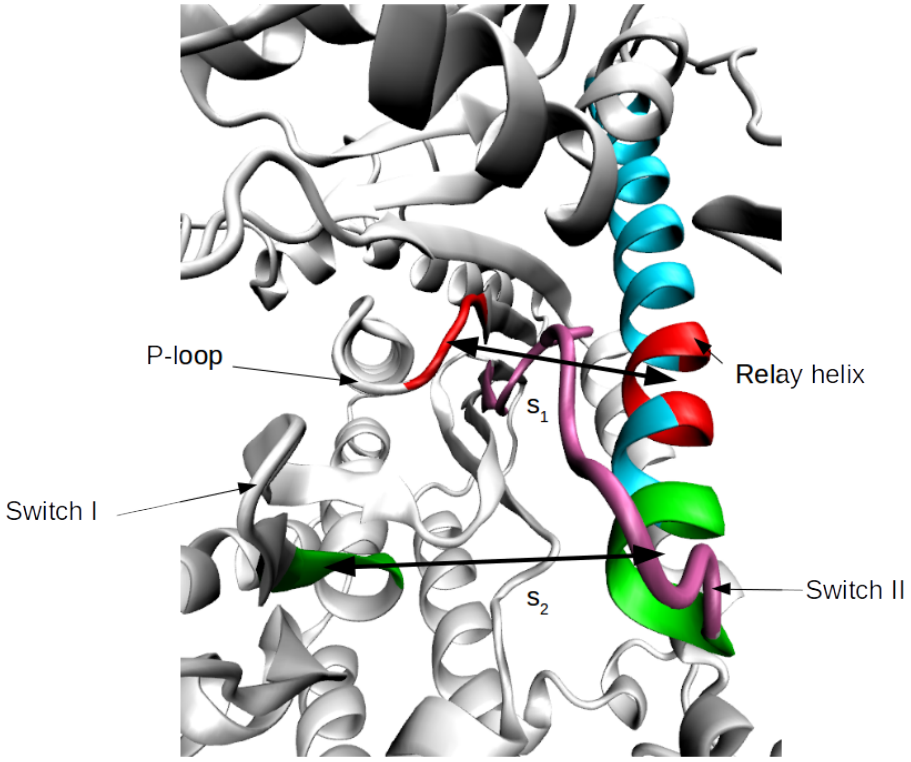


Figure 10.9.: Definition of the s_1 and s_2 distances to describe the seesaw motion of the Relay helix.

10.3.2. L50 rotation

Several attempts to drive the rotation of the L50 subdomain[1] using distances, orientation quaternions or rigid-body RMSD were performed, but were unsuccessful (data not shown), in the sense that SMD simulations did result either in unfolding of the motor domain or in no clear effect. Finally, we defined the L_1 collective variable as the *spinAngle* of rotation of the L50 subdomain (Fiorin, Klein, and Hénin 2013), using the following procedure.

The PR (pre-L50 rotation) and PPS (post-L50 rotation) crystal structures of myo6 were superimposed on the N-terminal subdomain, which notably isolates the movement of the L50. The orientation quaternion describing the orientation of the L50 in PR with respect to its position in PPS was computed with *colvars*; then, the corresponding rotation axis was computed. This axis, expressed in the reference frame of the N-terminal subdomain, is used to define the spin angle of the L50 subdomain, using the PPS crystallographic configuration as reference. This is illustrated on Figure 10.13.

L_1 was used to drive the rotation of the L50 in a series of 10 ns SMD simulations from PR ($L_1 \simeq 6°$) to PPS ($L_1 = 0$ by construction), trying a range of force constants. We used PR rather than PTS as the starting structure arbitrarily, as the purpose of these simulations was essentially to design an observable describing L50 rotation. The time-series of L_1 during SMD are reported on Figure 10.14.

Although visual inspection indeed shows an inward motion of the L50, no other striking rearrangement (*e.g.* switch II closure or a movement of the converter) is detected in the simulations (data not shown).

1. The L50 is defined as the CA atoms of residues 513 to 520, 524 to 535, 539 to 552 and 576 to 590.

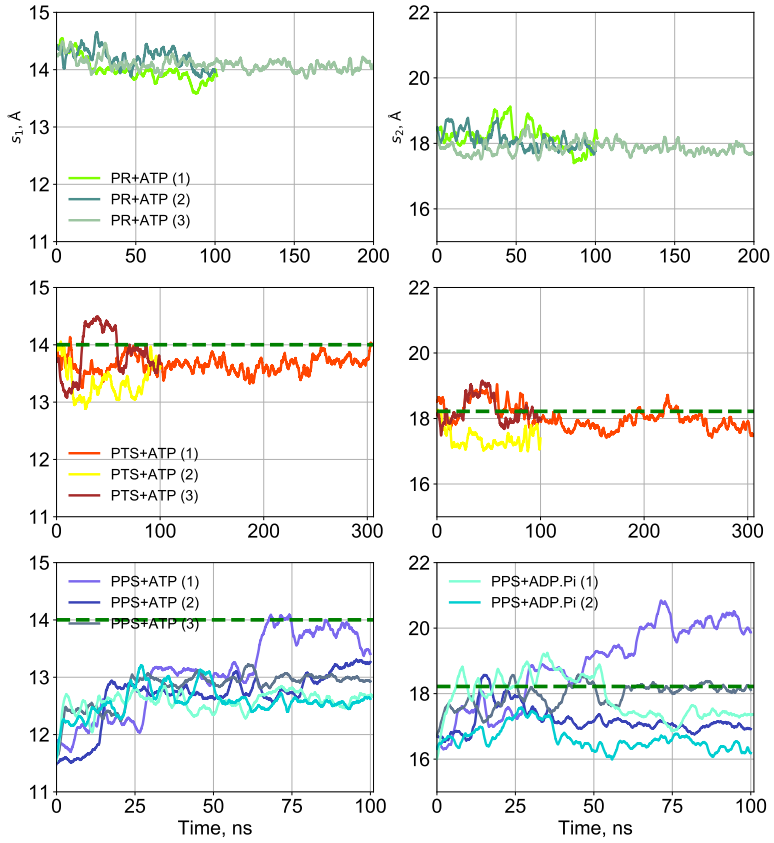


Figure 10.10.: Time-series of the distances s_1 and s_2 describing the seesaw motion of the Relay helix in unbiased simulations (2 ns running average). The dotted green line materializes the PR average value, which represents an "un-seesawed" Relay helix.

10.4. Conclusion

Arguably, the only mechanistic insight reported in this chapter is the putative mechanism of completion of switch II by switching of the β-strands interaction pattern. By contrast, the SMD simulations along CVs developed to drive the rotation of the L50 subdomain and the seesaw motion of the Relay helix do not reveal much of the overall transition mechanism, because no clear structural response to the perturbation is observed in other regions of the motor. In the next chapter, we outline a strategy to test the PTS hypothesis by the means of String Method in Collective Variables (CVSM) calculations. Incidentally, these calculations, which will notably make use of the CVs reported above, are in particular expected to yield a model of the PTS → PPS transition. This explains why the discussion of L50 rotation and Relay helix seesaw has been kept to a minimum in the present chapter, and why the movement of the converter (for which descriptive observables are already available) was not discussed.

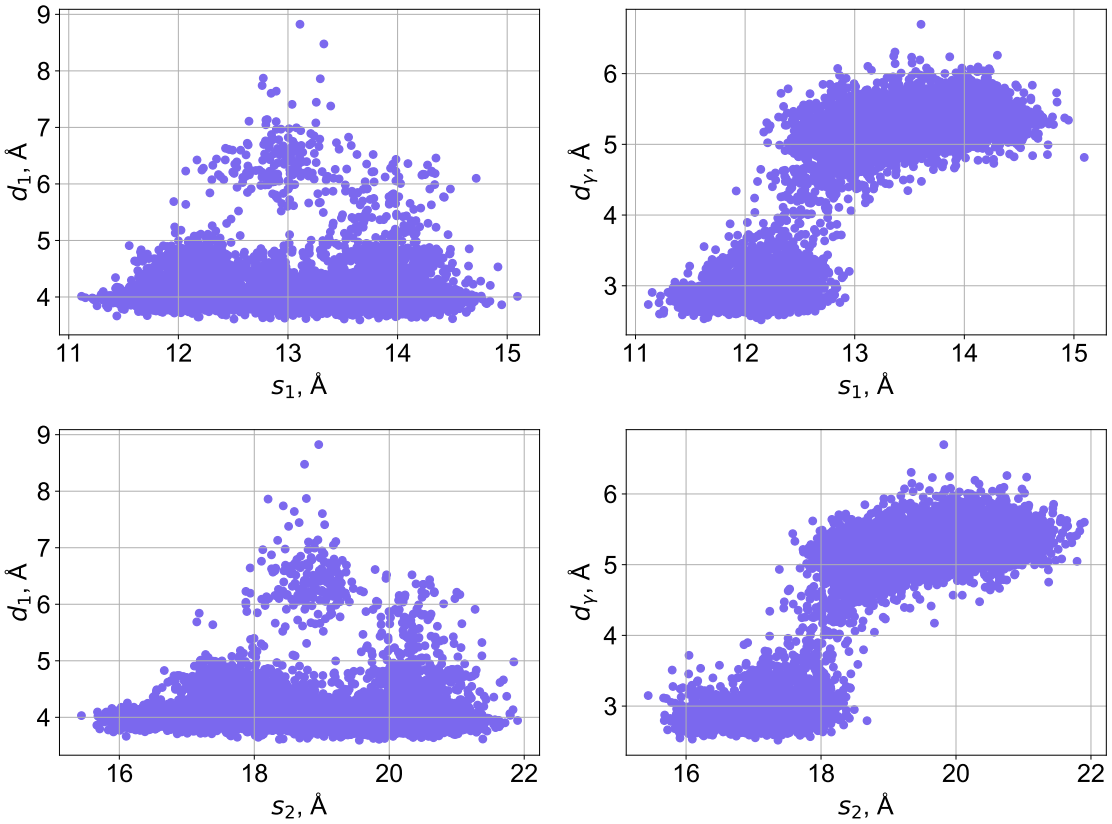


Figure 10.11.: Pairwise scatter plots of the two seesaw descriptive observables (s_1 and s_2) and the two descriptive switch II observables (d_1 and d_γ) for the PPS+ATP (1) unbiased MD simulation.

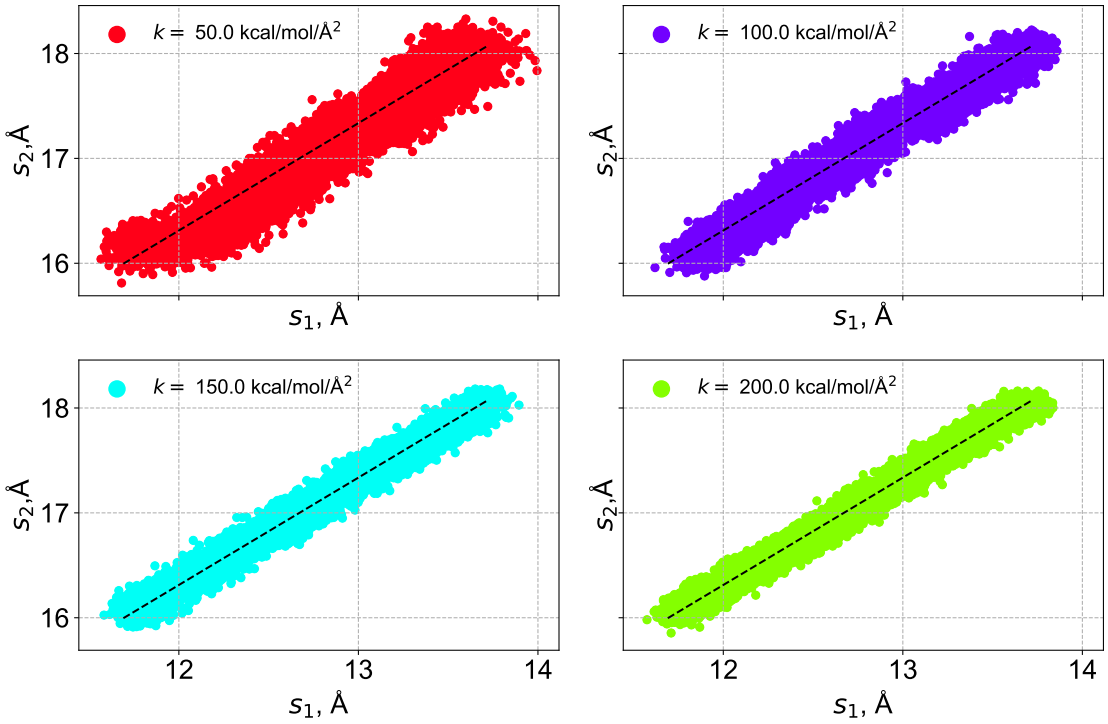


Figure 10.12.: Evolution of the biased seesaw distances s_1 and s_2 during SMD simulations from PTS to PPS. The black dotted lines represent the linear SMD bias.

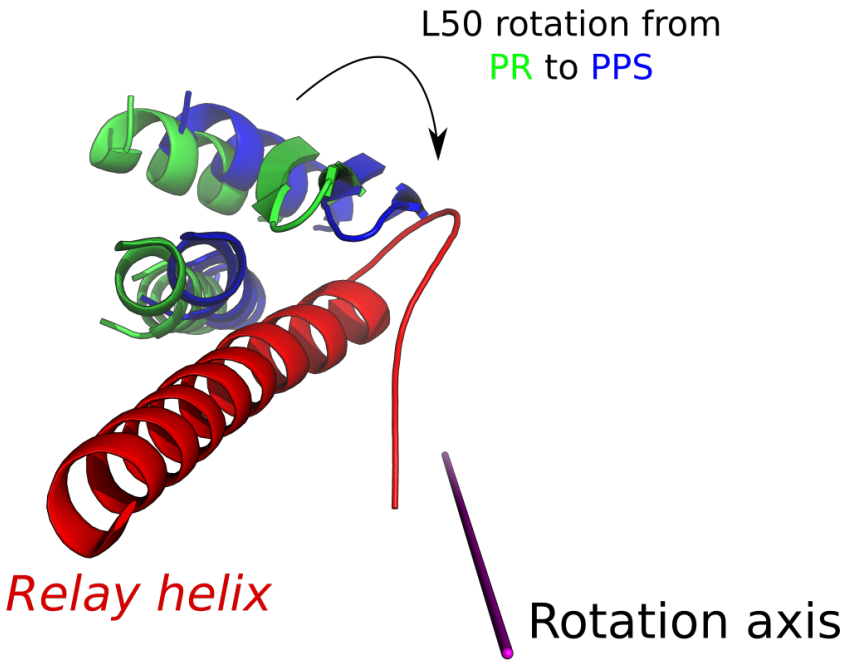


Figure 10.13.: Visualization of the L50 subdomain rotation from PR to PPS and of the rotation axis.

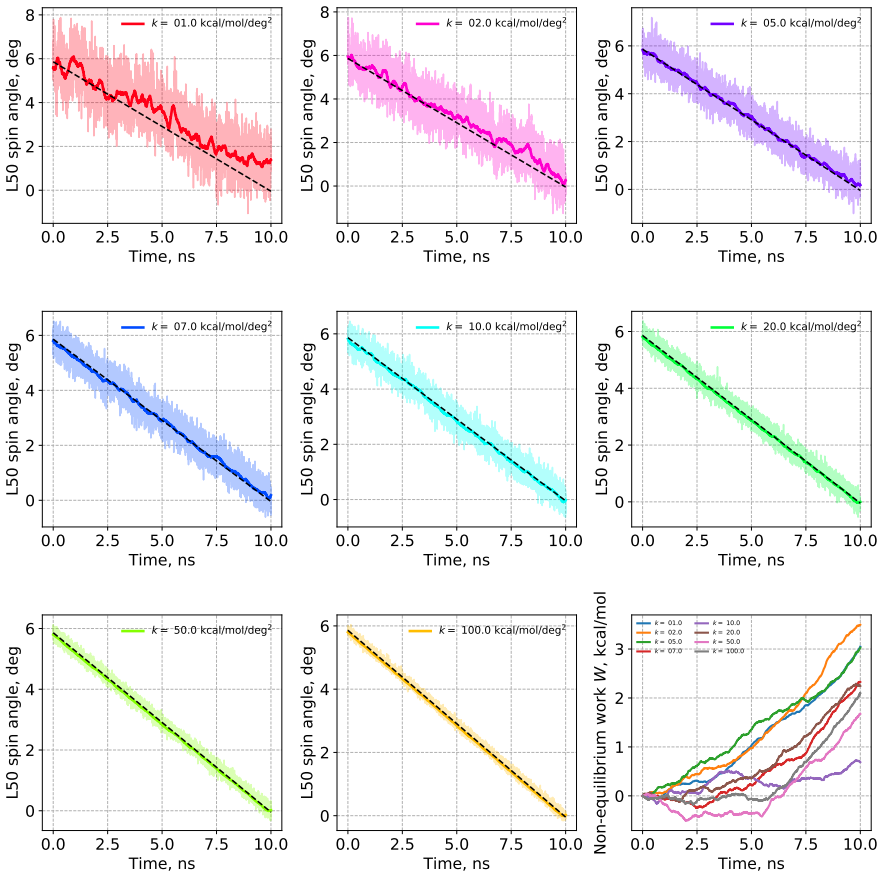


Figure 10.14.: Evolution of the biased observable L_1 during SMD simulations from PR to PPS, for a range of force-constants. In the lower right corner, the non-equilibrium work profiles are shown.

11. The ratchet-like model: overview, supporting arguments and missing pieces

Summary In this chapter, we first assemble the pieces of information obtained in chapters 5 to 10 into a global overview of the ratchet-like model emerging from the PTS structure and associated simulations. Despite some remaining unclear points, it is our conviction that this novel mechanistic scenario provides a realistic and consistent picture of the conformational transition. However, the value of a model is decided by its experimental verification. Thus, we turn to the existing experimental literature and critically review how the PTS hypothesis is consistent, or not, with the existing data. We argue that, provided that the overall mechanistic features are transferable between myosin II and myosin VI, there exists no clear refutation of the PTS model. Crucially, we explain why this point stands also for experimental results put forward in support of Fischer's alternative model. As such, we narrow down the discussion to the question of deciding which model, between the ratchet-like and Fischer's, is most likely to represent the pathway for the recovery stroke. Our overall conclusion is that available experimental techniques cannot resolve the details of the transition with atomic resolution in such a way that deciding between the two models becomes possible. Rather, one must turn to simulation approaches. Thus, we outline a strategy to directly compare the two mechanistic proposals using the string method in collective variables combined with free energy calculations. At the time of writing, these calculations are ongoing thanks to a 17,000,000 CPU-hours PRACE allocation. The first results of this analysis are presented. Finally, we conclude with a more general discussion of our findings in the context of molecular motors functioning principles. Apart from the discussion of experimental data, the results of this chapter are unpublished.

11.1. Overall summary of the ratchet-like model

11.1.1. Supporting arguments and main findings from the present work

The ratchet-like model emerges from the hypothesis that the PTS structure represents an on-pathway intermediate along the recovery stroke. The intriguing features of this structure suggest a mechanism in which the recovery stroke is initiated by a movement of the converter, and ATPase activation through switch II closure is statistically (rather than mechanically) coupled to the re-priming of the converter (Chapter 5). In Chapter 6, unbiased MD simulations of the myo6 motor domain yield the following observations in support of the ratchet-like model:

- A spontaneous uncoupling of the converter is captured in PR (simulation PR+ATP (3)). This supports the proposal that the recovery stroke is initiated by a movement of the converter.
- A spontaneous, reversible partial re-priming of the converter in PTS towards a position closer to PPS is captured (simulation PTS+ATP (1)). This supports the proposal that the PTS is an intermediate between PR and PPS; however, there is no further evidence that this "PTS-reprimed" configuration is on-pathway.

- At least in PR and PTS, all observed converter motions (which are extensive in PTS) have no detectable effect on switch II, which remains open. This supports the absence of strong, mechanical coupling between switch II and converter (at least in the early stages of the recovery stroke).

In addition, we observe from unbiased PTS simulations that the PTS converter, being largely uncoupled from the motor domain, undergoes extensive positional fluctuations which may facilitate the completion of the swing during the PTS → PPS transition. Accelerated Molecular Dynamics (aMD) simulations reported in Chapter 7 most notably capture a backward PPS → PTS transition, but fail to exhibit any direct PR ↔ PPS transition; this also supports the PTS being an intermediate of the recovery stroke. The analysis of the energetics of ATPase activation (switch II closure) by ABF calculations in Chapter 8 reveals that both in PR and PTS the fully-closed switch II is a high-in-free-energy metastable state separated by a high-barrier from partially closed configurations; in particular in PR, the barrier is about 12 kcal mol^{-1}. By contrast, in PPS, the fully closed switch II is the ground state. These findings strongly support a statistical coupling between switch II closure and converter rotation. Also, the calculations on PTS highlight a possible pathway for switch II closure involving its transient uncoupling from the Relay helix. Biased simulations and free energy calculations reveal the mechanism of the PR → PTS transition in Chapter 9. This transition is seen to proceed in a rather strongly coupled manner and involves the conformational isomerization of the Relay-SH1 elements, and the swing of the converter to the PTS basin. Whether the initiating event is the movement of the converter or the rearrangement of the Relay-SH1 elements is presently unclear, but this transition certainly occurs without detectable coupling to switch II, further supporting the statistical coupling of converter and switch II. Also, the estimated free energy barrier for the PR → PTS transition, 7 kcal mol^{-1}, makes this transition seemingly more likely than the early switch II closure predicted by alternative scenarios. Finally, the importance of local transducer rearrangements for switch II closure in the PTS → PPS transition is described in Chapter 10.

11.1.2. Sequence of events in the ratchet-like model

Assembling the previous findings, we can propose the following sequence of events for the recovery stroke in the ratchet-like scenario:

1. From the PR state, thermally activated swing of the converter and/or formation of the kink in the Relay helix and/or tilting of the SH1 helix. A seclusion of the Relay loop allows the rearrangement of the hydrophobic lock without seesaw motion of the Relay helix at this stage (9.1.2).

2. The motor domain reaches the PTS state. The converter, only coupled to the Relay and SH1 helices, fluctuates extensively while switch II is still open - which may provide entropic stabilization to the state. Possibly, switch II uncouples from the Relay helix N-terminal to form the switch II-ATP hydrogen bond, but the critical salt-bridge does not form yet.

3. Among these converter fluctuations, one is captured which allows the docking of the converter in the PPS position. Alternatively, the converter first explores a partially re-primed, bound state such as the ”PTS-reprimed” conformation captured in unbiased MD (Chapter 6), before jumping to the PTS interface by an unknown mechanism.

4. By an unknown mechanism, the completion of the converter swing drives the seesaw motion of the Relay helix, the inward rotation of the L50 (including the wedge loop), the local rearrangement of the transducer, and finally the full closure of switch II. Whether this latter starts by the formation of the critical salt-bridge or the switch II-ATP hydrogen bond is also unclear.

It is apparent that this picture is not yet complete as many details, particularly about the end of the rearrangement, are speculative. But, the defining feature of this ratchet-like model is that rearrangements of the force-generating region (converter+Relay/SH1) are driven by thermal fluctuations and precede ATPase activation. This model illustrates how a molecular motor may operate through *loose* (statistical) coupling between re-priming of the mechanical elements and chemical step. As such, it provides an example of how a motor works by exploiting spontaneous, thermally-driven fluctuations, rather than by strongly coupled rearrangements like in the case of macroscopic devices. This might be a general functioning principle for biomolecular motors, along with a design principle for artificial ones (Astumian 2007).

11.1.3. Critical assessment of the ratchet-like model against experimental data

As already mentioned in Chapter 5, many experimental studies of the recovery stroke have been performed. In the Supplementary Text 1 of our publication (Blanc et al. 2018), in Appendix (C) of the present thesis, we show how the phenotypes of characterized mutants of the recovery stroke which were proposed in support of Fischer's model are in fact not inconsistent with our mechanistic proposal (Batra, Geeves, and Manstein 1999; Kintses, Yang, and Málnási-Csizmadia 2008; Murphy, Rock, and James A. Spudich 2001; Patterson et al. 1997; Sasaki, Shimada, and Sutoh 1998; Sirigu et al. 2016; Tsiavaliaris et al. 2002). The reader is referred to this text for a detailed discussion, but the basic argument is actually rather simple. The set of elementary rearrangements which have to occur during the recovery stroke is fixed by the observed differences between its end-points, *i.e.* the PR and PPS states. As such, any mechanistic model of the full transition is bound to involve these rearrangements and may differ from alternative proposals only by 1) their order of occurrence (sequence of events) and 2) the nature of the coupling between them (mechanical vs statistical). Targeted mutational or pharmaceutical perturbations disrupt a given rearrangement; as a result, the overall transition is impaired, regardless of the specifics of the mechanism. In fact, some mutations may still in principle allow for the discrimination between competing mechanisms, for example by trapping particular intermediates predicted in only some of the mechanistic proposals. In our discussion in (Blanc et al. 2018), we show that this is not likely to be the case regarding the available data on the recovery stroke of myosin. We note that regardless of the validity of previous mechanistic proposals (Chapter 5), credit should be given to their authors for identifying and describing the elementary rearrangements involved the recovery stroke. Finally, we stress that virtually all the experimental results discussed above have been obtained for myosin isoforms other than myosin VI; although we have assumed that they can be generalized to myosin VI, the possibility is left open that the mechanism of the recovery stroke may differ between isoforms. Despite some preliminary results suggesting that Dd myo2 (the primary isoform used in experiments) may follow a similar mechanism, this point will have to be investigated in details in the future.

11.1.4. Unified picture of competing models for the recovery stroke

Considering the sequence of events for the ratchet-like model (11.1.2), it seems that the seesaw motion of the Relay helix has to be coupled to the completion of the converter swing and the closure of switch

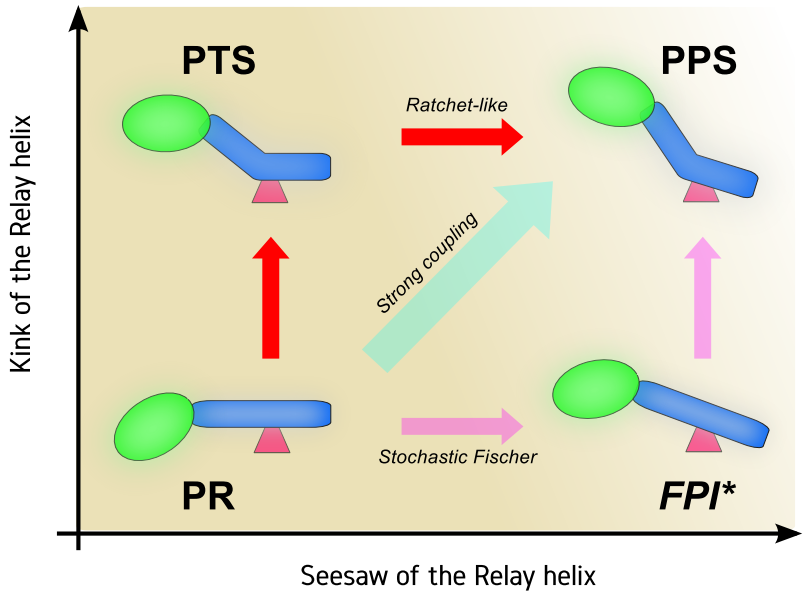


Figure 11.1.: Unified picture of the recovery stroke mechanistic proposals. The two rearrangements involving the Relay helix are coupled to different elementary transitions in the rest of the motor. Limiting cases for the recovery stroke are the PTS-ratchet-like model (statistical coupling, the kink precedes the seesaw, PTS is the intermediate), the strongly coupled model by Fischer (concerted transition) and the statistically-coupled interpretation of Fischer's model (statistical coupling, the seesaw precedes the kink, prediction of "Fischer's putative intermediate" FPI.)

II during the PTS to PPS transition, although the mechanistic details are unclear at the time. Moreover, the formation of the kink in the Relay helix is coupled to the tilting of the SH1 helix and a partial swing of the converter. Interestingly, Fischer's model yields essentially similar predictions (but with a different timing and a different conclusion as to the nature of the coupling).

In fact, we recognize that a unified picture of competing mechanistic proposals for the recovery stroke is obtained by considering the two following "reaction coordinates": formation of the kink in the Relay helix, and seesaw motion of the Relay helix; see Figure 11.1. In this framework, the existence of PTS is actually expected if the kink in the Relay helix forms before the seesaw motion happens.

In this picture, our own ratchet-like model, the original strongly coupled interpretation of Fischer's model, and the alternative statistically coupled interpretation of Fischer's model appear as limiting cases for the recovery stroke pathway. This restricts the range of possible mechanisms, and allows us to focus on a comparison of the alternative scenarios to determine the most probable one. Our estimates of the free energy barriers for the initiating events in both models (Figure 11.2) suggests that the ratchet-like pathway is more probable, since, assuming a constant and identical pre-exponential factor, we find:

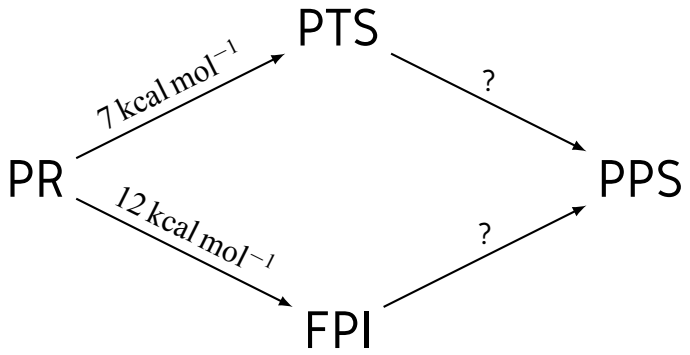


Figure 11.2.: Transition diagram of the recovery stroke

$$\frac{k_{ratchet-like}}{k_{statistical-Fischer}} \simeq \frac{e^{-\beta \Delta F^{\ddagger}_{ratchet-like}}}{e^{-\beta \Delta F^{\ddagger}_{statistical-Fischer}}} \simeq 4000 \tag{11.1}$$

for the relative initiation rate. But, this result relies on strong approximations (diffusion-coefficient neglected and barriers estimated with different sets of CVs) and includes only the initial stages of the full transition. Thus, it is not sufficient to demonstrate that the ratchet-like scenario is most likely. More robust approaches must be used for a truly quantitative insight, as we show now.

11.2. String method strategy for the comparison of mechanistic proposals

The string method in collective variables (see 4.4.2) has emerged in recent years as the state-of-the-art method for the study of conformational transitions in biomolecular machines including motors (Das et al. 2017; Lev et al. 2017; Ma and Schulten 2015; Ovchinnikov, Karplus, and Vanden-Eijnden 2011; Singharoy, Chipot, et al. 2017; Zhu and Hummer 2010). And, Chipot and co-workers have pioneered the investigation of the functional mechanisms of *artificial* molecular machines (see citations in Appendix B along with our own study of a prototypical rotaxane-based molecular switch by the string method).

In fact, Singharoy and Chipot (2016) have proposed a general strategy to investigate the energetics of functional cycles in molecular motors. Interestingly, their strategy combines the usage of the string method and bias-exchange umbrella sampling along the functional transitions, so as to reveal the free energy profiles, with alchemical (Free Energy Perturbation (FEP)) calculations to evaluate the free energy changes upon the hydrolysis of ATP, in such a way that a closed picture of the cycle is obtained. This approach is quite computationally expensive when applied on large systems, but if successful, it will in principle provide a full thermodynamic and kinetic characterization of the motor cycle, notably opening the way to experimental testing of quantitative predictions.

The application of this method to myosin is a tempting perspective, but a premature one, first and foremost because no high-resolution structure of the actomyosin complex is available. This precludes any string method study of the on-actin branch of the cycle. Rather, the string method may be used to derive the most probable pathway for the recovery stroke, which is the question we are addressing here. A naïve application of the string method towards this aim would involve the generation of a guess path between PR and PPS, followed by its relaxation to an optimal path by string iterations.

But this approach is most likely bound to fail in revealing the most probable mechanism, because the resulting path will *a priori* represent a *locally* optimal path. Arguably, a better approach is to optimize, then compare two (or more) alternative pathways for which prior knowledge suggests they represent reasonable candidates for the *globally* optimal pathway, *i.e.* , in the case of the recovery stroke, the ratchet-like scenario vs the statistical Fischer scenario. The strategy for such a comparison would proceed as follows. For each scenario, a guess path is generated and optimized by the string method using the same set of collective variables (4.4.2). Then, free energy calculations along the path yield the free energy profile, ideally including the entropic correction (4.4.3). Finally, rate calculations by milestoning or any other suitable method give a quantitative prediction of the respective rates for each scenario (4.4.4). Once these are obtained, the most likely transition path corresponds to the one with the fastest predicted rate, which may possibly be compared to experimental measurements. Also, we note that the situation in which the two pathways have close rates and as such both contribute to the overall flux would be detected.

The string calculations are required because accurate rate estimations must be performed along an optimal path; also, they will yield the atomically-detailed mechanism of the transition in each scenario as a by-product. Possibly, the evaluation of the free energy profile along the path could be bypassed (because it is not needed if the rate is estimated by milestoning), but the free energy profiles will reveal the potential intermediates along the paths, and the free energy barriers, which are interesting *per se*. Notably, if it is found at this stage that one pathway entails significantly lower barriers than the other, one may already conclude that it is the most likely one, on account of the exponential dependence of the rate on the barrier height (4.4.4).

We apply this strategy to the elucidation of the optimal pathway for the recovery stroke of myosin VI by comparing the ratchet-like scenario and the statistical-Fischer scenario; in a second time, the strongly-coupled scenario may also be considered. For that purpose, guess-path generation, string optimizations, free energy calculations and rate calculations along both pathways are required, which represents an extensive computational cost. We have recently been awarded a 17,000,000 CPU-hours PRACE allocation to perform these challenging calculations. At the time of writing, the analysis of the ratchet-like scenario is ongoing, and we now outline the first results.

11.2.1. Choice of supporting collective variables

11.2.1.1. Rationale

The choice of the supporting set of collective variables, *i.e.* the set of collective variables which will be used for string optimization, is crucial. Previous string method studies in the literature typically resort to preliminary validation procedures, using SMD simulations to ascertain the capacity of a given collective variable to drive the particular rearrangement it is supposed to describe. For instance, in the work by Victor Ovchinnikov and co-workers (Ovchinnikov, Karplus, and Vanden-Eijnden 2011), SMD is used to identify a minimal set of distances which are sufficient to drive the full conformational change under study (the R to P transition of the myosin VI converter). An extreme example is the recent study by Takemoto and co-workers (Takemoto et al. 2018), in which a single collective variable is used to drive the global conformational transition of a triose-phosphate/phosphate antiporter; this CV is then combined with the position of the transported ligand (phosphate) along the pore for string method optimizations, eventually leading to a converged string in a 2-dimensional space which describes the coupling between the phosphate translocation and the global conformational changes of the protein. Thus, in these studies, the general philosophy regarding the choice of CVs seems to be one of parsimony.

In the case of the recovery stroke of myosin, we chose a different approach. Instead of being parsimonious, we decided to go for at least one collective variable for each sub-transition that was deemed important based on our previous analyses of the recovery stroke. For example, we explicitly included different CVs for both the Relay-SH1 elements and the converter, despite SMD evidence showing that driving the rearrangement of the Relay-SH1 is sufficient to drive the rotation of the converter (see Chapter 9). These CVs were validated using SMD simulations, and the resulting supporting set includes 25 independent CVs.

Generally speaking, we made this choice 1) to ensure as fine a resolution as possible in terms of the sequence of structural events along the computed optimal path, and 2) to make sure that each degree of freedom identified as (potentially) relevant is explicitly accounted for (and not only through passive coupling with a separate biased degree of freedom).

It is thus likely that our supporting set of CV is redundant, *i.e.* it includes more CVs than would have been strictly required to capture the recovery stroke. Fundamentally, we do not believe this to be a problem, since there is no theoretical limit (other than the dimension of the full system) to the number of CVs which can be included in the definition of the string. In practice, it may nonetheless pose two problems. First, there is an extra computational cost associated with a rather high number of CVs. Second, one would intuitively expect that in a high-dimensional space (high number of CVs), the free energy surface over which the string is evolved towards convergence is more rugged (*i.e.* exhibits many local minima) than in low-dimension. This is one of the reasons the string method in collective variables is preferred to the zero-temperature string method (which optimizes the string in cartesian coordinates) for large systems. A more rugged free energy landscape may slow down the relaxation of the string towards the (locally) optimal solution.

The influence of the supporting set dimension on the convergence behaviour of the string is a difficult question to tackle systematically, because one would need to compare the convergence of strings computed on the same system for very different numbers of supporting CVs. Yet, if a system is large enough for this to be possible, it is probably too large for a systematic string optimization to be undertaken (*i.e.* it is not a good toy-model). Moreover, one may remember that it is natural to expect the convergence of the string to be dependent on the particular nature of CVs and not only their number. Overall, it seems that a compromise should be found between a supporting set of CVs which encompasses all relevant degrees of freedom, but of small enough dimension that the free energy surface smoothness favors convergence. Parsimony prioritizes smoothness and is certainly a good approach, but there also have been reported cases of successful string method studies using very high numbers of CVs (see for instance Miller, Vanden-Eijnden, and Chandler 2007).

11.2.1.2. List of collective variables

Throughout this thesis, we have introduced a collection of collective variables to describe the various elementary sub-transitions taking place during the recovery stroke. Most of these CVs were validated using SMD/TMD simulations, see Chapters 9 and 10. It was thus natural to use these observables in the supporting set for the string method. In addition, we added a series of distances, suggested by Dr Anne Houdusse, to account for the closing of the 50 kDa cleft - another sub-transition taking place during the PTS → PPS transition. Overall, the full supporting set includes 25 collective variables which are summarized on Figure 11.3 and Table 11.1.

We note that no explicit collective variable was introduced to account for the internal conformational transition of the converter. We made this choice because it is a myosin VI-specific sub-transition, which is not suited for a general description of the recovery stroke. A spontaneous conformational

Collective variable	Associated rearrangement	Force constant
Distance d_1 (R205CZ:E461CD)	Formation of the critical salt-bridge	20.0 kcal/mol/Å^2
Distance d_γ (G459N:ATPO1G)	Formation of the switch II-ATP hydrogen bond	20.0 kcal/mol/Å^2
Distance d_2 (R199CZ:E461CD)	Formation of the secondary salt-bridge[a]	20.0 kcal/mol/Å^2
Distance b_1 (V149N:L455O)	β-strands switching	20.0 kcal/mol/Å^2
Distance b_2 (T158OG1:D456CG)	β-strands switching	20.0 kcal/mol/Å^2
Distance b_3 (G151N:I457O)	β-strands switching	20.0 kcal/mol/Å^2
Distance b_4 (K208N:D456O)	β-strands switching	20.0 kcal/mol/Å^2
Distance b_5 (F206O:A458N)	β-strands switching	20.0 kcal/mol/Å^2
Distance s_1 (151-153CA:475-479CA)	Seesaw motion of the Relay helix	60.0 kcal/mol/Å^2
Distance s_2 (194-196CA:468-472CA)	Seesaw motion of the Relay helix	60.0 kcal/mol/Å^2
Distance c_1 (239CA:468CA)	Cleft closure	10.0 kcal/mol/Å^2
Distance c_2 (199CA:464CA)	Cleft closure	10.0 kcal/mol/Å^2
Distance c_3 (357CA:537CA	Cleft closure	10.0 kcal/mol/Å^2
Distance c_4 (419CA:599CA)	Cleft closure	10.0 kcal/mol/Å^2
Spin angle L_1	L50 rotation	5 kcal/mol/°^{2}
Distance k_1 (486O:490N)	Kink in the Relay helix	50 kcal/mol/Å^2
Distance k_2 (485O:489N)	Kink in the Relay helix	50 kcal/mol/Å^2
Distance k_3 (485O:490N)	Kink in the Relay helix	50 kcal/mol/Å^2
Distance k_4 (486O:491N)	Kink in the Relay helix	50 kcal/mol/Å^2
Distance $d_{R/SH1}$ (469-482CA:693-703CA)	Seclusion of Relay and SH1 helices	20 kcal/mol/Å^2
Orientation angle θ_{RH}	Bending/re-orientation of the Relay helix	5 kcal/mol/°^{2}
Orientation angle θ_{SH1}	Tilting/re-orientation of the SH1 helix	5 kcal/mol/°^{2}
Orthogonal projection X'	Converter swing	10 kcal/mol/Å^2
Orthogonal projection Y'	Converter swing	10 kcal/mol/Å^2
Orthogonal projection Z'	Converter swing	10 kcal/mol/Å^2

Table 11.1.: List of collective variables used as the supporting set for string method optimizations of the recovery stroke pathways.

a. The secondary salt-bridge, R199-E461, corresponds to a possible salt-bridge between switch I and switch II, distinct from the critical one, which was observed to form rapidly in the PTS equilibration simulation (not shown). This salt-bridge was used as an auxiliary reaction coordinate for the control ABF calculations reported in Chapter 8.

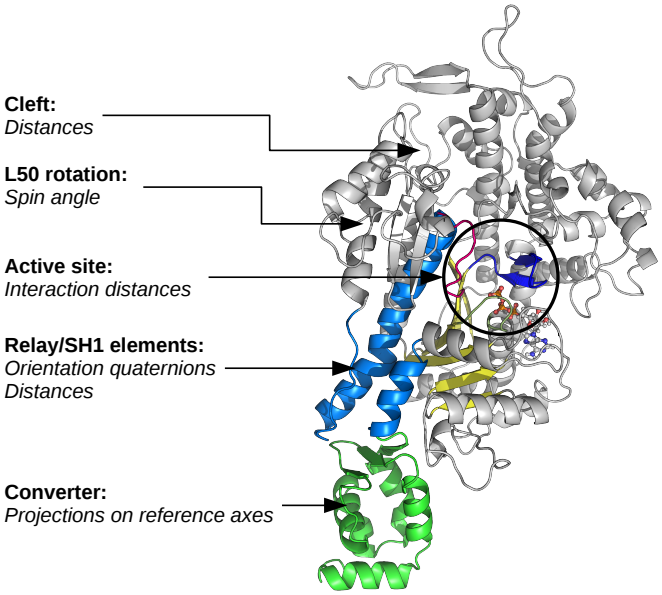


Figure 11.3.: Summary of the collective variables used as the supporting set for string method calculations on the recovery stroke of myosin.

transition of the converter may still be captured along the string if this transition is indeed coupled to the general rearrangements occurring during the recovery stroke.

11.2.2. First results: string method study of the ratchet-like model in Myosin VI

We report on the first results of the string method study, namely the ongoing optimization of two strings describing the ratchet-like model in myosin VI.

11.2.2.1. Construction and choice of a guess path

To construct a guess path to initialize the string calculations, we used Steered Molecular Dynamics along the 25 supporting CVs defined above. The following protocol was used: first, a 50 ns SMD simulation acting on all the CVs with their respective force constants was run from the PR (equilibrated structure) to the PTS (equilibrated structure); then, the last frame of this simulation was used to initiate a 50 ns SMD targeting the PPS state (equilibrated structure). This ”staged” SMD protocol ensures that the PTS state is indeed visited, so that the guess path is an acceptable approximation of a transition consistent with the ratchet-like model. Then, 60 equally spaced frames were extracted from the full (*i.e.* concatenated) SMD trajectory. The trace of these frames in CV-space represents a path, which deviates from the linear path of the SMD bias because it also contains the fluctuations sampled during the simulation. This path was normalized (as described below) and re-parametrized, after which it was used as reference to run on-the-path Umbrella Sampling. This serves two purposes. First, the 60 conformations extracted from SMD cannot be directly used as a starting point for string optimization, because they retain the “memory” of the non-equilibrium pulling of the SMD protocol (imperfect relaxation). A static relaxation in the harmonic restraining potential, or ”pre-equilibration”, is required. Second, this allows to obtain a first estimate of the PMF along the path. For each window, 4.5 ns of

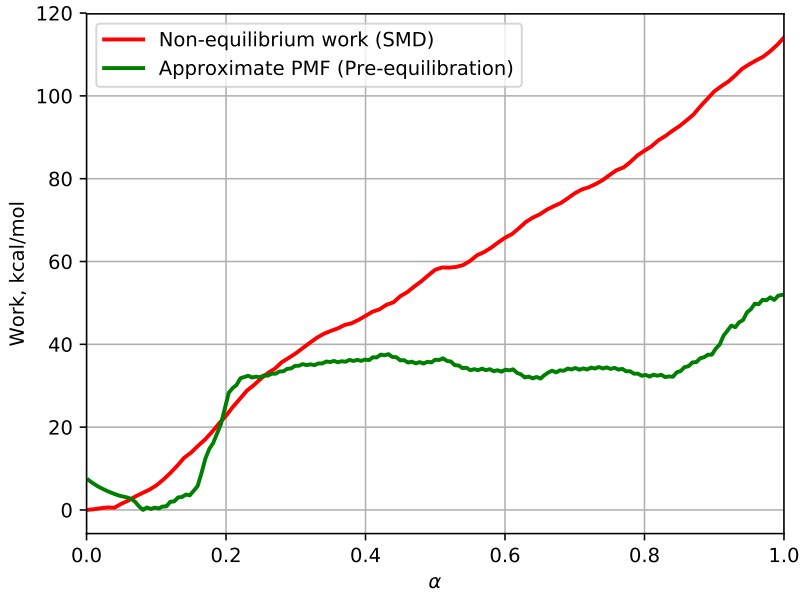


Figure 11.4.: Comparison of the irreversible work accumulated during the SMD and the (approximate) PMF obtained by static relaxation of 60 images along the SMD path 1.

harmonically restrained sampling was performed. Then, the PMF along the path was computed by Umbrella Integration with the chain rule and spline-fitting (see 4.4.3).

Two independent staged-SMD simulations were performed and compared. We made sure that in each case all the CVs had not ended up too far away from their respective target values. Then, we chose the path which exhibited the smallest free energy change after relaxation by umbrella sampling (pre-equilibration), because this path is arguably the closest to a minimum free energy path (Figures 11.4 and 11.5). Thus, this path (path 1) was retained as the guess for the string calculations.

11.2.2.2. String method calculations: run parameters

The string calculations were run with our NAMD implementation of the string method with swarms-of-trajectories, which simulates each image in parallel, but each run of a given swarm sequentially. The string calculations were run with 60 images and 20 short MD runs within a given swarm. At each iteration, the harmonically restrained stage lasts 100 ps and is followed by 20×10 ps of free swarm simulations. Then, image positions are updated using the average drift from the swarm of trajectories, the string is re-parametrized after normalization of CVs, and smoothed by local averaging using a 0.1 smoothing parameter. The force constants used for the harmonically restrained stage are reported in Table 11.1.

The Molecular Dynamics parameters were the same as before except that a 1 fs time-step was used and a 1.0 ps friction coefficient was used.

Since the string is define in a space of heterogeneous CVs (different mathematical natures and units), a normalization procedure must be used before re-parametrization to prevent the string parametrization from being overly affected by the specifics of the numerical values of each CV. To that end, before reparametrization, each CV was normalized by its total variation (*i.e.* the difference between its maximal and minimal values) to make sure that all CVs exhibit comparable range of variations.

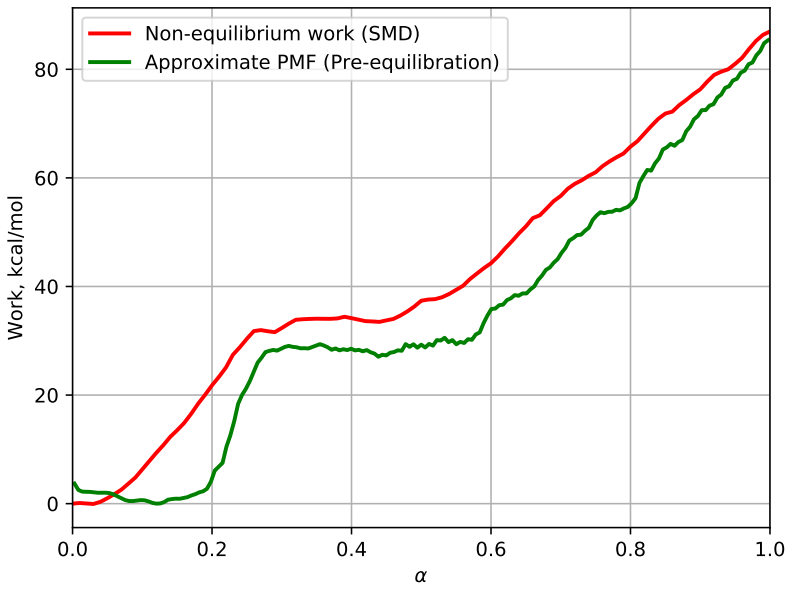


Figure 11.5.: Comparison of the irreversible work accumulated during the SMD and the (approximate) PMF obtained by static relaxation of 60 images along the SMD path 2.

String method iterations were initiated from the configurations obtained at the end of pre-equilibration. Two independent strings were relaxed: a ”free end string”, in which the end-points (first and last images) are allowed to relax following the drift vector as the others, and a ”fixed end string” where the end-points are kept at the PR and PPS equilibrated structures, respectively. At the time of writing, 76 iterations have been performed for the ”fixed end” string, and 68 for the ”free end” string.

11.2.2.3. String method calculations: convergence

The convergence of the strings was evaluated using several definitions of the RMSD. After normalization of CVs by their total variation, one can compute 1) the RMSD with respect to the final string, 2) the ”progressive” RMSD (*i.e.* the RMSD of the current string with respect to the previous string) and 3) the RMSD with respect to the initial string (guess path). All these quantities are expected to plateau upon convergence of the string. Figures 11.6 and 11.7 show the evolution of these three quantities for each string, over string iterations. In each case, a tendency to convergence is observed, suggesting that the strings are indeed relaxing towards optimal paths. However, large fluctuations remain which suggests that convergence is not yet achieved. Consequently, these string calculations are still being extended at the time of writing.

11.2.2.4. Behaviour of selected observables

We now present the evolution of selected observables along the relaxed paths, comparing their behaviour between the two paths and with respect to the initial guess. For analysis, the averaged string over the last 10 iterations are considered.

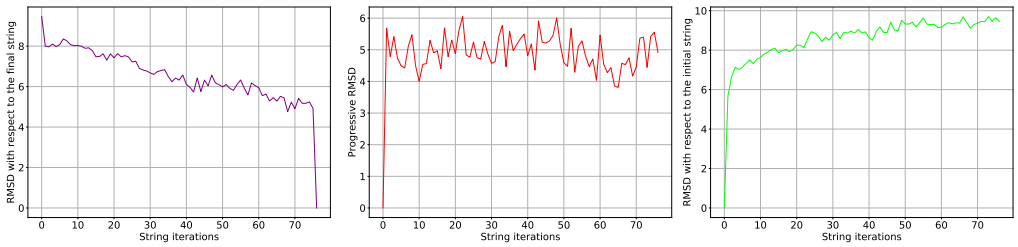


Figure 11.6.: Convergence behaviour of the fixed end-points string after 76 iterations.

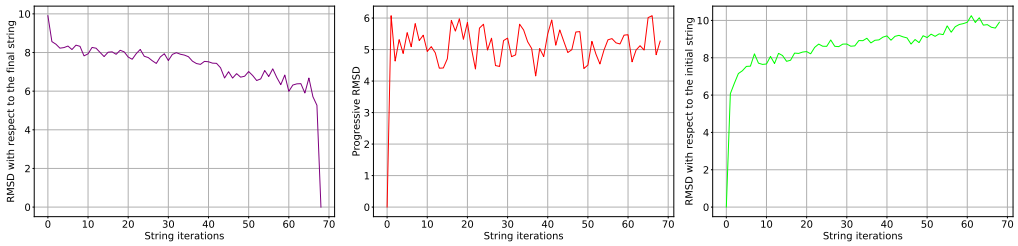


Figure 11.7.: Convergence behaviour of the free end-points string after 68 iterations.

Converter position The movement of the converter along the recovery stroke in the $X'Y'$ plane is shown on Figure 11.8. Strikingly, both relaxed paths do not wander too far away from the initial guess from the SMD trajectory. The metastable converter positions probed by the PTS unbiased simulations (see also Chapter 6) are not visited by the string simulations (e.g. $X' = -7.5$ Å, $Y' = -2$ Å). When looking at individual components (Figure 11.9), it is apparent that the trend set by the SMD simulation is not changed, except perhaps for Z' which seems to reach lower values faster after string relaxation than during the SMD. One may also point out the remarkable similarity of the two independently evolved strings in the second part of the transition ($\alpha > 0.6$), which suggests that the pathway for completing the converter swing from PTS to PPS is robust. Also, note that for X' and Z' the string reaches the same values at $\alpha = 1$ regardless as to whether the ends are held fixed or not. This is not the case for Y', as fixing the end enforces a large jump from a plateau in $Y' = -3.5$ Å to the fixed value of $Y' \simeq 0$ Å. The plateau at $Y' = -3.5$ Å, also probed by the free-end string optimization, may represent the actual stable basin of the converter in the PPS state when the converter has not undergone the R to P fold transition (which is not observed in either string).

Relay-SH1 elements Much like what is observed for the converter, the relaxed paths projected onto the $\theta_{RH}, \theta_{SH1}$ plane do not drastically deviate from the SMD guess (even though some local departures are observed), see Figures 11.10 and 11.11. θ_{RH} exhibits a rather smooth progression, with a steeper increase during the first part of the full transition corresponding to the formation of the kink in the Relay helix ($\alpha \simeq 0.3$). By contrast, θ_{SH1} fluctuates more, in particular after the formation of the kink ($\alpha \simeq 0.3$). It is not clear at this point whether this fluctuating behaviour is a feature of the optimal pathway or a memory of the initial guess, which exhibited similar fluctuations.

We note that the free-end string relaxes towards the center of a basin sampled in unbiased MD of the PPS state (Figure 11.10). By contrast, the fixed-end string ends on the value reported for the equilibrated PPS structure, which appears not to belong to any metastable state identified by unbiased MD. This would suggest that the CV values observed on the center of the most populated cluster should be used as targets, rather than those of the equilibrated structure. However, note that for the

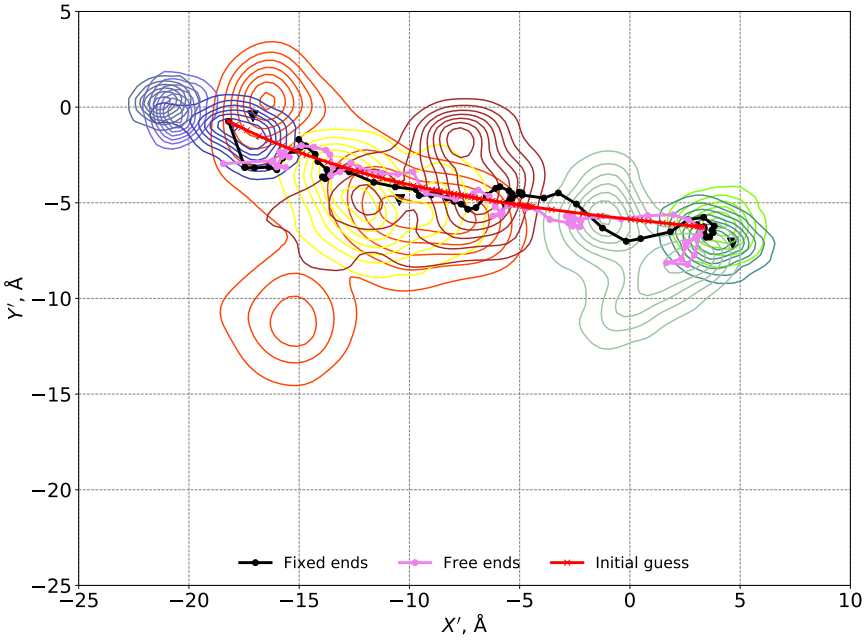


Figure 11.8.: Path of the converter in the X', Y' plane as observed from string calculations along the ratchet-like model. Density lines from unbiased MD simulations are shown for comparison (green, PR+ATP; yellow/orange, PTS+ATP; blue, PPS+ATP).

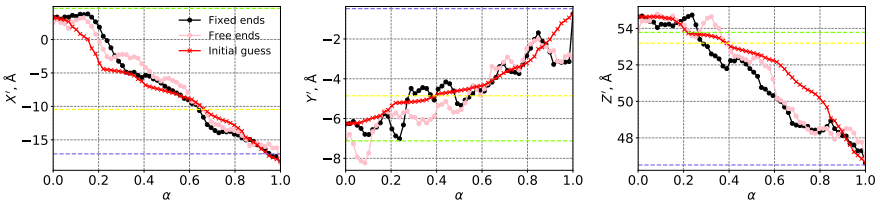


Figure 11.9.: Evolution along the string (progress parameter α) of the converter position components. The dotted lines materialize the corresponding values on the PR (green), PTS (yellow) and PPS (blue) equilibrated structures.

converter, the opposite behavior is observed (the free-end string is out of the PPS basin, Figure 11.8).

Multiple PPS basins The density lines for the PPS simulations (blue lines on figure 11.10) show that at least two distinct metastable states are explored in PPS regarding the Relay-SH1. These states are characterized by a common value of the θ_{RH} angle, but different θ_{SH1} values corresponding to different tilting states of the SH1 helix. The most tilted states are explored in PPS simulations where an "over-repriming" of the converter is also captured, i.e. where the converter is seen to relax towards positions even more re-primed than is observed in the PPS crystal structure (see also Figure 11.8, the PPS basin at $X' = -21$ Å, and Chapter 6). Similar orientations of the SH1 helix are also probed in unbiased PTS simulation during the transient, partial movement of the converter towards the PTS position (Chapter 6).

Regarding the Relay helix kink hydrogen-bonding distances, three out of four distances show significant departure from the guess path to adopt a clear sigmoid shape consistent with an "all-or-none"

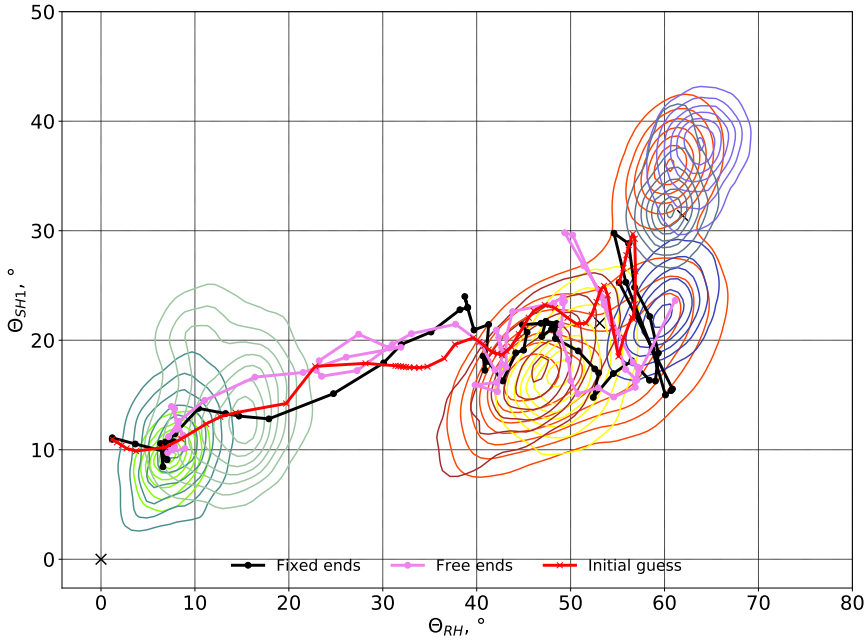


Figure 11.10.: Projection of the strings onto the $\theta_{RH}, \theta_{SH1}$ plane. Density lines from unbiased MD simulations are shown for comparison (green, PR+ATP; yellow/orange, PTS+ATP; blue, PPS+ATP).

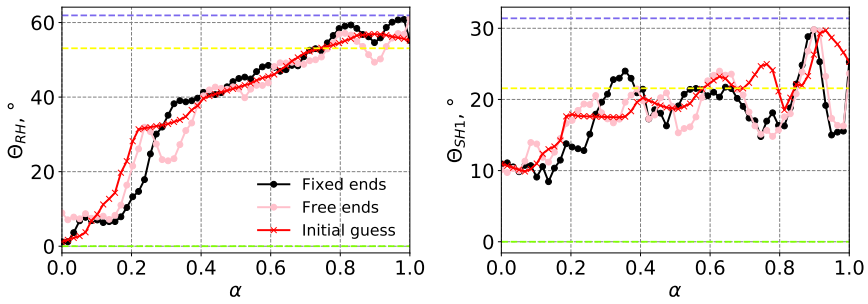


Figure 11.11.: Evolution along the strings (progress parameter α) of the Relay-SH1 orientation angles. The dotted lines materialize the corresponding values on the PR (green), PTS (yellow) and PPS (blue) equilibrated structures.

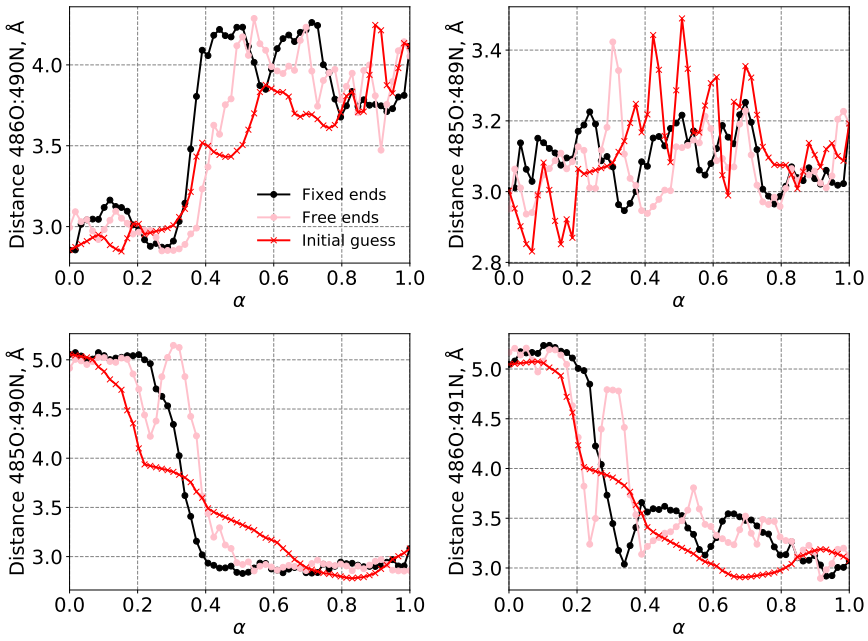


Figure 11.12.: Evolution along the strings (progress parameter α) of the Relay helix kink backbone hydrogen-bonding distances.

behaviour for the exchange in hydrogen bonds (see Figure 11.12). This is a novel result -although one we suspected- and reassures us that the string optimization is actually doing something to relax the path away from the guess where appropriate. The remaining distance (485O:489N) exhibits a more noisy behavior, reminiscent of what was already seen in the guess path. It is likely due to the very small range of variation of this distance (about 0.2 Å). The associated hydrogen bond may be weakened rather than completely broken during the formation of the kink.

Active site interactions and switch II closure As before, switch II closure is studied through observables d_1 (critical salt-bridge distance) and d_γ (switch II-ATP hydrogen bond distance). These interactions are probably the observables exhibiting the most markedly diverging behaviour with respect to the guess path. Whereas the guess path is extremely smooth and takes a turn to visit PTS-compatible values, both relaxed paths are bypassing PTS to visit intermediate basins which are sampled in PPS simulations (Figure 11.13). Moreover, in both relaxed paths, switch II appears to remain open, with PR-like values, for most of the transition. This suggests that the late-switch II closure feature of the ratchet-like mechanism is preserved when the pathway is optimized, but also that the conformation of the active site explored in PTS is possibly off-pathway (as opposed to the conformation of the Relay-SH1 elements and the position of the converter). The final closure of switch II occurs for both relaxed paths between the 59th and 60th images along the string (so, the last two), see Figure 11.14. For the fixed-end string, this was first interpreted as a sign that the closed-switch II PPS configuration is unstable with ATP (which is consistent with our observations in unbiased MD, chapter 6). However, surprisingly, a very similar behaviour is observed in the free-end string, where the end state is free to relax to a more favorable basin. It is thus possible that the abrupt, ”one-step” behaviour of the formation of the switch II-ATP hydrogen bond observed in both strings is an actual feature of the transition. If this were confirmed, it would represent a strong supporting argument for the late closure

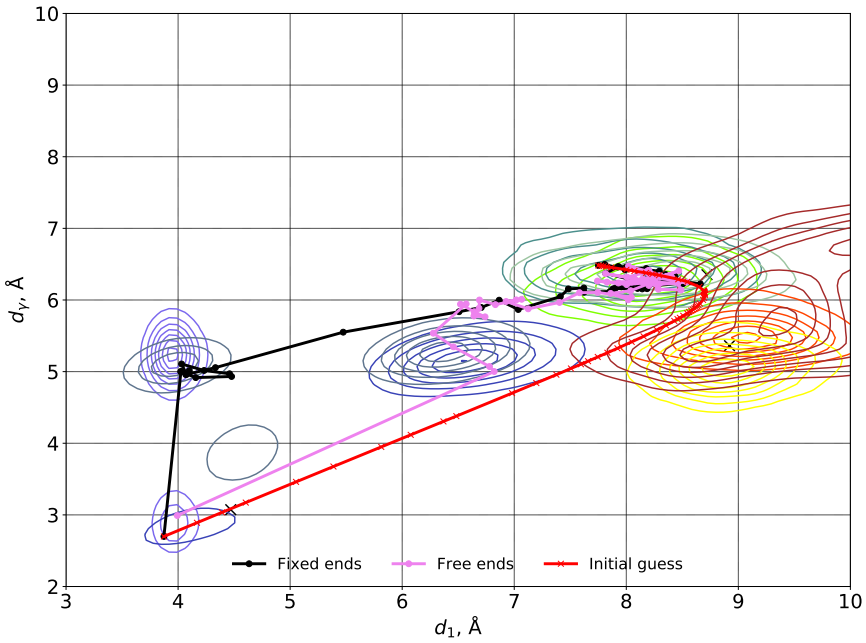


Figure 11.13.: Projection of the strings onto the d_1, d_γ plane to reveal the pathways of switch II closure in the ratchet-like model of the recovery stroke (green, PR+ATP; yellow/orange, PTS+ATP; blue, PPS+ATP).

of switch II.

Also, interestingly, both strings explore alternative metastable basins for switch II configurations, which correspond respectively to a concerted formation of the salt-bridge and the hydrogen bond (violet curve = free-end string) and a sequential mechanism where the formation of the salt-bridge precedes that of the hydrogen bond (black curve = fixed-ends), see Figure 11.13. Thus, it is possible that multiple pathways exist for the closure of switch II. In this respect, we note that the pathway where switch II decouples from the Relay Helix to form the hydrogen bond with ATP before the salt-bridge, which was sampled by ABF simulations (see Chapter 8), is not sampled by the string calculations.

The determinants of stability of the closed-switch II configuration in PPS+ATP are still unclear (see discussion in Chapter 10), and an exploration of the latest stages of switch II closure in presence of ADP.Pi in the active site may be required. To that end, one may use a ”double string” strategy,

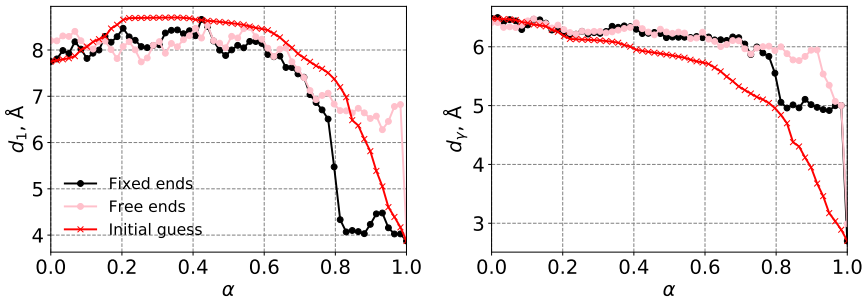


Figure 11.14.: Evolution along the strings (progress parameter α) of the active site distances d_1 and d_γ.

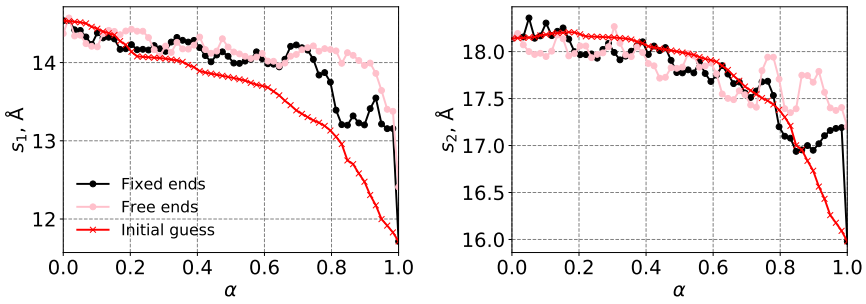


Figure 11.15.: Evolution along the string (progress parameter α) of the seesaw distances s_1 and s_2.

first optimizing the PR → "PPS with open switch II" transition in presence of ATP, followed by the optimization of the "PPS with open switch II" → "PPS with closed switch II" transition in presence of ADP.Pi. The "PPS with open switch II" can be identified as the center of the most populated cluster of a long PPS unbiased MD simulation where switch II is observed to re-open, as is the case in the simulations we have run to date. These calculations are in preparation.

Seesaw motion of the Relay helix The evolution of the distances s_1 and s_2 describing the seesaw motion of the Relay helix is shown on Figure 11.15. As compared to the rather progressive evolution observed in the guess path, the seesaw motion in relaxed paths is seen to occur with a delay and with different modalities between the two paths. In the fixed-end paths, an abrupt seesawing event is detected around $\alpha = 0.8$ (so, near the end of the recovery stroke), which occurs simultaneously to the formation of the critical salt bridge and the shortening of the d_γ distance in the active site (compare Figure 11.15 with Figure 11.14). This suggests that the completion of the seesaw and the formation of the critical salt-bridge are coupled, which was already somewhat hinted in the ABF calculations on PPS (chapter 8) but was not observed in SMD simulations of the seesaw (chapter 10). In the free end string, there seems to be a relaxation towards an even later, and incomplete, seesaw motion; this, in consistency with our observations on the active site, again points towards the relative instability of the seesaw in PPS where ATP is present, which certainly deserves further investigation.

11.2.3. Ruggedness of the projected paths

A striking common feature of the projected paths is their significant ruggedness. Paths appear to be irregular, exhibiting kinks and re-crossings. This is not unexpected as the object under study is a 2-dimensional projection of a curve lying in a 25 dimensional space; details about the topography are lost upon dimensionality reduction. As such, one needs not be concerned by the facts that 1) equal spacing is not satisfied (it is satisfied in the 25-dimensional space) and 2) re-crossings are observed (they are a projection artifact).

However, the irregular aspect of the paths is problematic because it will be challenging to obtain a sensible approximation of the path with a smooth function (B-spline). This operation is required for the evaluation of the PMF along the path by Umbrella Sampling, and the noise in the path is likely to propagate to the free energy gradient estimate and thus, to the final PMF. It may be a good idea to consider increasing the level of smoothing (by local averaging) of the strings for the subsequent iterations.

11.2.4. Conclusions and future calculations

Even if their proper convergence to optimal pathways is uncertain, the strings reported above seem sufficient to draw general conclusions about the ratchet-like model of the recovery stroke. Essentially, it is observed that the features of the (imperfectly) relaxed paths are still consistent with the predictions of the ratchet-like model, *i.e.* early formation of the kink in the Relay helix and converter movement, existence of a PTS state as an intermediate, and late seesaw motion of the Relay helix/closure of switch II.

Since these string calculations were specifically designed to probe the ratchet-like scenario, this is not really surprising; however, it is encouraging because it suggests that there indeed exists a locally optimal transition pathway corresponding to the ratchet-like model. If this were not the case, string iterations would have been expected to quickly relax away from the ratchet-like pathway. Thus, the present string calculations, upon convergence, will arguably be well suited to provide the long sought-after atomically detailed description of the recovery stroke in the ratchet-like model, and will provide adequate starting points for the evaluation of the associated free energy barriers (umbrella sampling/Voronoi sampling) and kinetic rate (milestoning). In particular, as the PTS state seems to be entropically stabilized by a wide distribution of converter positions, the inclusion of the entropic correction to the free energy profile is likely to be essential to obtain a reliable PMF.

Of course this will not be sufficient to conclude that the ratchet-like pathway is the most probable; instead, our strategy involves performing a set of similar calculations along the model of Fischer and co-workers (in its statistical interpretation, at least in a first time). A required step towards such calculations will be the generation of guess paths by a sequential SMD protocol along PR → FPI → PPS. But, at the time of writing, no structural characterization of the hypothetical FPI is available. Thus, a model of the FPI structure should be produced, according to its hypothetical characteristics (straight but see-sawed Relay helix, closed switch II, etc). As the significance of the subsequent string calculations is expected to rely heavily on the quality of the FPI model, particular care will be taken in its production. As a first approach, we plan to build this model using long SMD simulations (>100 ns) followed by a long relaxation (same time-scale at least) with static harmonic restraints, and finally an assessment of its structural stability using unbiased MD.

Once an acceptable model of FPI is obtained, string method optimizations and free energy calculations will reveal the atomically detailed sequence of events and kinetic rate entailed by the statistical scenario of Fischer and co-workers (and, more generally, by switch II-initiated scenarios). The comparison of the predicted rate with that of the ratchet-like scenario is expected to be decisive in elucidating the most probable mechanism of the recovery stroke, and eventually validate or not the PTS hypothesis for myosin VI. Subsequently, the same calculations will be considered for Dd myo2 rather than myosin VI, so as to assess the generality of the predicted mechanism.

11.3. A discussion of the term ”ratchet”

The main results of this thesis, namely the proposal of a novel mechanistic model for the recovery stroke of myosin, has now been outlined. We have argued that this new model is consistent with what would be expected for a molecular motor, as it gives a dominant role to conformational fluctuations (see also Chapter 1). Notably, we have used the term ”ratchet-like” to characterize the model. The notion of ratchet, and its relevance to describe the functioning principles of molecular motors, is an ancient and still unresolved topic. Over the years, the term ratchet seems to have taken on several related, but distinct acceptances, which we believe may be source of confusion. This section is in-

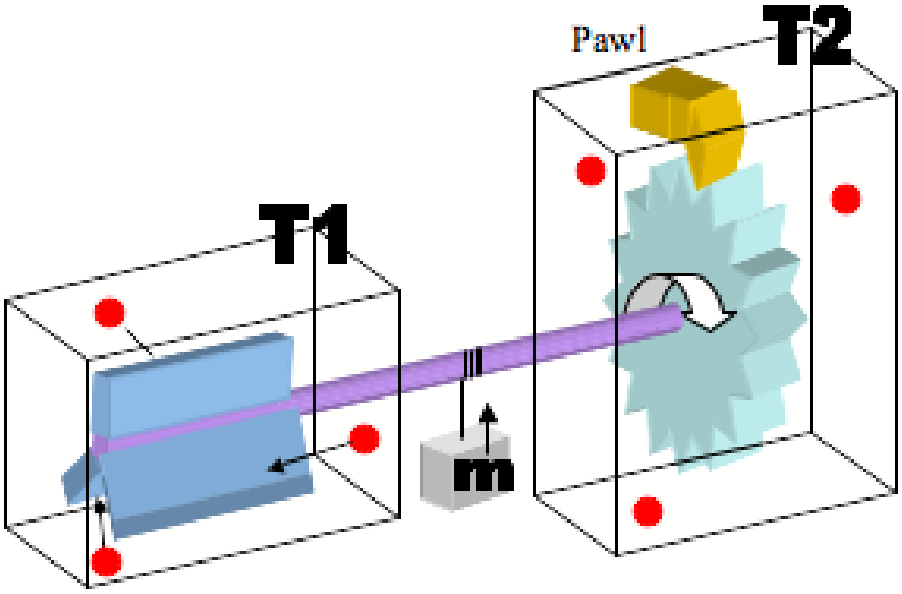


Figure 11.16.: Illustration of the ”ratchet-and-pawl” device. Credits: Wikipedia (image free of rights)

tended as a short critical discussion of our findings in the broader context of molecular motors and their functioning principles.

11.3.1. Mechanical analogies and the classical picture of molecular motor operation

In the classical picture, first proposed by A.F. Huxley in 1957, motor proteins such as myosin can produce force by releasing conformational free energy that has been stored at another step of the cycle. The powerstroke corresponds to the conformational change in which the stored conformational free energy is elastically dissipated in a productive manner, i.e. through the forward swing of the lever arm; the recovery stroke is thus the free energy storage step. In the powerstroke framework, it is hypothesized that some sub-domains can undergo a conformational change to a strained state, akin to the loading of a spring - these hypothetical subdomains are referred to as *hidden springs* (Houdusse and Sweeney 2001).

11.3.2. The concept of ratchet and its relevance for molecular motors

The concept of Brownian ratchet originates in a famous thought experiment by Marian von Smoluchowski (1912), in which he describes the following device (see Feynman, Leighton, and Sands 2011). A paddle-wheel is immersed in a gas at temperature $T > 0$. Through an axle, the paddle-wheel is connected to a dented wheel in a different compartment, such that the rotation of the paddle-wheel is freely transmitted to the dented wheel, see Figure 11.16. If the device is of microscopic size, the collisions of the gas molecules on the paddle will cause rotation events and move the dented wheel in the other compartment; in this situation, the average rotation is zero. However, if one adds a pawl to the dented wheel such that one direction of motion (reverse) is prevented (ratchet-and-pawl mechanism), only fluctuations which happen to push the paddle in the forward direction will produce actual rotation. With this system, we expect a one-way rotation (although in a stochastic manner) of the wheels. If one attaches a small mass to the central axle, we anticipate that the rotation of the axle will lead to the progressive elevation of the mass, demonstrating how such a device would produce work.

This idea is of course wrong, as it predicts that work can be extracted from isothermal random fluctuations in contradiction with the Second Law. As such, the ratchet-design is the prototype of a perpetual motion machine of the second kind. The practical explanation of why it does not work is that if the system is small enough that its rotation can be driven by Brownian fluctuations, this will also apply to the pawl. The pawl will also fluctuate, sometimes opening and permitting a reverse rotation - such that the average displacement is zero.

For this reason, the Smoluchowski ratchet cannot represent a good model for the functioning principles of molecular machines. However, it highlights a seductive possible mechanism: the rectification of isotropic fluctuations. Essentially, the idea is that molecular machines continuously undergo fluctuations, some of them corresponding to productive motions: rectification refers to any mechanism by which these productive fluctuations are captured, and/or unproductive (backward) fluctuations are excluded. How exactly this takes place is to be determined for the system of interest, but thermodynamics requires that it be coupled to an exergonic process.

Theoretical studies have put forward the concept of ”flashing ratchet” in which the moving particle undergoes Brownian dynamics while switching between two potential surfaces: a flat one, onto which the particle diffuses freely and isotropically (*e.g.* when a motor is detached from its filamentous track), and a periodic, asymmetric *ratcheting potential* (reflecting the periodic and polar nature of the track) (Jülicher, Ajdari, and Prost 1997). When the stochastic switching between the two potential surfaces is averaged out, a single, effective potential surface is obtained, but, as justified by Jülicher, Ajdari, and Prost (1997), this effective potential is also periodic and thus, does not exhibit a net gradient; so, no directional motion is achieved. Instead, directional motion requires the breaking of the detailed balance (*i.e.* microscopic reversibility) associated with the transition between the two potential surfaces, see (Jülicher, Ajdari, and Prost 1997) for details. It is proposed that the hydrolysis of ATP in non-equilibrium conditions (*i.e.* when ATP, ADP and Pi concentrations are maintained away from their equilibrium values by active cellular processes) provides this breaking of detailed balance. So, in this picture, the effectively irreversible hydrolysis step drives the transition from the flat to the ratcheting potential surfaces. A way to apply this to myosin is by postulating that ATP hydrolysis at the end of the recovery stroke ”locks” the motor in the PPS state which will subsequently interact with actin (*i.e.* , jump from the freely diffusing state to the asymmetric potential).

Our usage of the term ratchet in ”ratchet-like” scenario is related, but slightly different. We do not consider the rectification of isotropic positional fluctuations, but of *conformational* fluctuations. Namely, in our model, instead of being driven by the perturbation of ATP binding, the transition occurs first with large thermally-activated rearrangements of the force-generating region of the motor domain (*i.e.* a productive conformational fluctuation), which precede the formation of specific interactions within the active site. These latter, by promoting the hydrolysis of ATP, stabilize the system in the PPS state. Note that the ratchet-like model is not inconsistent with the powerstroke framework; rather, it indicates how myosin stores conformational free energy in preparation for the powerstroke by capturing spontaneous fluctuations which go towards the high-in-free-energy PPS state. In the terminology of J. Howard, this would correspond to a global, Kramers-like mechanism (see Howard 2001, pages 268-269) as opposed to a local, Eyring-like mechanism.

11.3.3. Kinetic asymmetry in chemically-fuelled motors

Several investigators, among which Prof. Dean Astumian, have recently challenged the previous ideas of conformational free energy storage and irreversible ”locking” steps on theoretical grounds (Astumian 2007, 2015, 2010). Regarding free energy storage, we described how, in the powerstroke

framework, the PPS is a ”mechanically strained” high-in-free energy state which relaxes elastically upon interaction with actin. In other words, this implies that the PPS motor domain is out of mechanical equilibrium at the very beginning of its interaction with actin, and that the powerstroke corresponds to the return to equilibrium. By a rather convincing statistical-mechanical argument, Astumian shows that this picture does not hold because mechanical equilibrium is virtually always satisfied for autonomous molecular machines (Astumian 2007). More precisely, while Astumian does not challenge the fact that some of the conformational transitions during the cycle may happen by relaxation along a conformational free energy gradient, he finds that they play no role in determining motor properties such as directionality and stopping force, see (Astumian 2015) for details.

The only ”non-equilibrium” aspect of an autonomous biomolecular motor operation is the fact that ATP, ADP and Pi concentrations are never allowed to reach their equilibrium values, but are actively kept constant. And, obviously, a single motor cannot sense these bulk concentrations (Astumian 2012). The conformational fluctuations of the motor, including productive ones, will be exactly the same regardless of the fuel concentrations. Instead, these concentrations only affect the relative probabilities with which ATP or ADP, Pi will bind to the motor. Based on these considerations, Astumian proposed a different mechanism by which forward motion and force generation is achieved in autonomous biomolecular motors.

Briefly, this mechanism, which also provides the basis for a recently reported autonomous chemically-fuelled motor, is that of a so-called *information ratchet* (Astumian 2016, 2012; M. R. Wilson et al. 2016). In this picture, all fluctuations are explored by the system in mechanical equilibrium, but kinetic barriers are modulated in such a way that the reversal of a forward (productive) fluctuations is kinetically disfavoured relative to progressing along the cycle. Thus, directionality is controlled by the ratio of forward and backward free energy barriers, rather than state-to-state free energy differences. Modulation of kinetic barriers is achieved by allosteric coupling between the conformational state of the motor and the rate of binding/release of the reactants and products of the catalyzed reaction (Astumian 2010).

The structural mechanism by which kinetic asymmetry may be achieved in biological motors such as myosin is unclear. In a 2013 study, Mukherjee and Warshel reported on a coarse-grained molecular mechanics-based energetic description of the motor cycle of myosin V (Mukherjee and Warshel 2013). Strikingly, and in disagreement with the prediction of the powerstroke framework, they found that the recovery stroke actually happens with a negative change in conformational free energy, which is proposed to compensate for the loss of the stabilizing interaction with actin upon the Rigor to PR transition. Then, by correcting the calculated free energy levels for a single, isolated myosin head with bulk chemical potentials of ATP, ADP, Pi and actin, they arrived at a global thermodynamic description of the stepping cycle for a two-headed motor. Interestingly, they concluded, in agreement with Astumian’s prediction, that the forward pathway does correspond to the one with the lowest free energy barriers. Unfortunately, the lack of details about the used energy function casts doubt as to the quantitative precision of these results.

Although this may simply point to our own imperfect understanding of Astumian’s picture, it seems to us that a complete irrelevance of mechanical/structural features in determining motor properties is unlikely, notably because in myosin VI, directionality reversal has been experimentally proved to be achieved by a re-orientation of the lever-arm (see 2.5). It is unclear how this finding fits in the ”information ratchet” picture.

We note that all-atom molecular dynamics simulations, in particular the string method and free energy calculations which were used in this thesis, provide an unmatched opportunity to compute 1) the ligand-dependent free energy barriers central in the information ratchet view, and 2) the conforma-

tional free energy levels of the conformational states of the motor, which may be used to assess the free energy storage predicted by the powerstroke framework. Thus, we expect molecular simulations to be instrumental in bridging the gap between theoretical and structural descriptions of molecular motors in the coming years.

12. Myosin is more than the motor domain: computational investigations of the lever and tail domains of myosins

Summary This chapter presents computational results obtained on problems unrelated to the recovery stroke. First, the flexibility of the lever-arm of Myosin X is analyzed by implicit-solvent simulations. Conformations extracted from simulations help resolve an apparent discrepancy between crystallography and single-molecule experiments. Second, the conformational dynamics of MyTH-FERM domains in the myosin tail is studied by MD and compared to that of the talin homolog. The observed difference in flexibility helps to understand the difference in affinity and binding modes to cellular partners. Both these projects led to (non-first author) publications (Planelles-Herrero, Blanc, et al. 2016; Ropars et al. 2016).

12.1. Flexibility of the lever-arm domain of Myosin X

12.1.1. Myosin X in the cell

Myosin X is a processive, plus-directed, Vertebrate-specific, 240 kDa Myosin (Berg et al. 2001; Yonezawa et al. 2000). It is expressed in most tissues with typically low levels. In Myosin X, the lever arm is made of 3 IQ domains (binding calmodulin (CAM)-like light chains) extended by a Single Alpha Helix (SAH) region and a coiled-coil dimerization domain. Its tails notably exhibits a MyTH4-FERM domain.

Myosin X is involved in a variety of cellular processes including filopodia formation and cell division (Bohil, Robertson, and Cheney 2006). During meiosis, Myosin X plays a key role in coupling the actin and microtubule cytoskeletons, and is required for properly setting up the meiotic spindle (Weber et al. 2004). Myosin X is also involved in phagocytosis (Chavrier 2002) and binds to the plasma membrane via integrins (Zhang et al. 2004).

Filopodia are thin cellular projections which are notably involved in cell migration and environmental exploration (Mattila and Lappalainen 2008). Filopodia are internally structured by parallel, fascin-reticulated actin bundles. Myosin X is mostly localized at the tip of filopodia and has been shown to exhibit a bi-directional intra-filopodia motility. Centripetal movement is a passive consequence of the inward actin flow, and only movement towards the tip of the filopodia corresponds to processive displacement of Myosin X on actin. Myosin X is involved in cargo transport towards the filopodial tip and seems to be required for filopodia formation and elongation, which makes it a crucial player in neuron extension and tumor invasion.

12.1.2. Myosin X exhibits selectivity for actin bundles

Early attempts to characterize the motor properties (processivity, duty-ratio) of Myosin X by the means of motility assays led to an apparent contradiction: despite its well documented role as a processive cargo transporter in the cell, Myosin X exhibited poor processivity *in vitro*, especially as compared to Myosin V. The group of R. Rock shed light on this discrepancy by showing that Myosin X recovers a high *in vitro* processivity on reticulated actin filaments - which are precisely the type of actin filaments found in filopodia (Nagy et al. 2008). Importantly, Myosin X exhibits processivity on fascin-reticulated filaments (*in vivo*-like) but also on artificially methylcellulose-reticulated filaments. The bundled nature of actin appears to be crucial in allowing Myosin X to walk processively, but the structural bases of this differential processivity remained to be elucidated.

Through an approach combining structural biology and single-molecule motility assays, the Houdusse team, as part of a collaboration, provided new insight into the functional adaptations of myosin X to preferential processive displacement on bundled actin (Ropars et al. 2016). Intriguingly, single-molecule experiments reveal the existence of four possible step sizes when myosin X steps on bundled actin. A crystal structure of the myosin X lever-arm dimer solved by Dr Virginie Ropars and reported in this study offers the possibility to rationalize the step sizes, because it contains two complete SAH domains interacting through their coiled-coil dimerization region (along with the last CAM-binding motif, termed IQ3). As such, it represents the region connecting the two myosin heads in a processive myosin X - the ”legs” of the motor. However, being a static structure, it only accounts for one of the observed step sizes. In this context, we performed an MD simulation of the dimer to evaluate its stability and flexibility, and assess whether alternative conformations compatible with other observed step sizes may be explored. The results indeed show that the dimer takes on a more extended conformation in simulation, which is compatible with one of the measured step sizes.

12.1.3. Molecular Dynamics of the lever arm dimer

12.1.3.1. Simulation setup

We used MD to assess the flexibility of the lever arm. Given the very extended shape of the structure, it appeared more reasonable to use an implicit treatment of the solvent. In particular, the use of the FACTS implicit solvent model (Haberthür and Caflisch 2008) was previously validated by Wolny et al. (2014) on the SAH domain of Myosin X. We used FACTS to run a 100 ns MD simulation of the Myosin X lever arm dimer at 300K, in complex with two calmodulin chains (Figure 12.1).

Simulations were run with CHARMM (version c38b1) (B. R. Brooks et al. 2009). FACTS parameters were taken from (Wolny et al. 2014), *i.e.* Tfps $= 3$, dielectric constant $= 1.0$, $\kappa = 4.0$ and $\gamma = 0.015$ (see the FACTS publication (Haberthür and Caflisch 2008) for definition of these parameters). Before the production simulation, the structure was energy minimized (1000 steps of steepest-descent and 2000 steps of Adopted-Basis Newton-Raphson method) under 10 kcal/mol/Å^2 positional harmonic restraints on the heavy atoms, using the crystal structure as a reference. The minimized structure was heated to 300 K using successive 100 ps-long Langevin dynamics runs at 50 K, 100 K, 150 K, 200 K and 300 K with a friction coefficient of 10 ps^{-1}, with active restraints. Then, the heated structure was equilibrated for 1 ns, with a 10 ps^{-1} friction coefficient and positional restraints of force constant 5 kcal/mol/Å^2. A 100 ns production simulation, without restraints, was performed at 300 K with a 1 ps^{-1} friction coefficient.

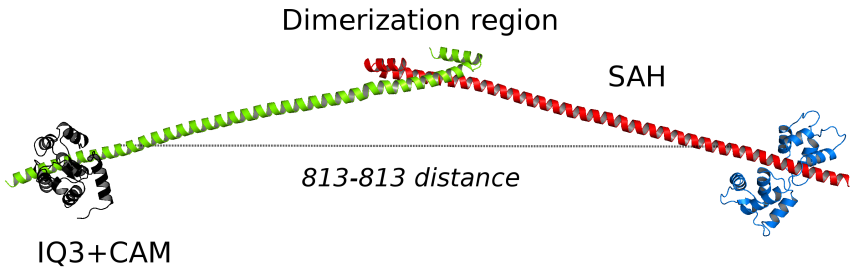


Figure 12.1.: Model of the IQ3-SAH-CC-SAH-IQ3 dimer, with CAM chains added to the IQ3 regions. The distance between E813 residues CA atoms of each chain (813-813), used to evaluate the extension of the structure, is shown.

12.1.3.2. Results

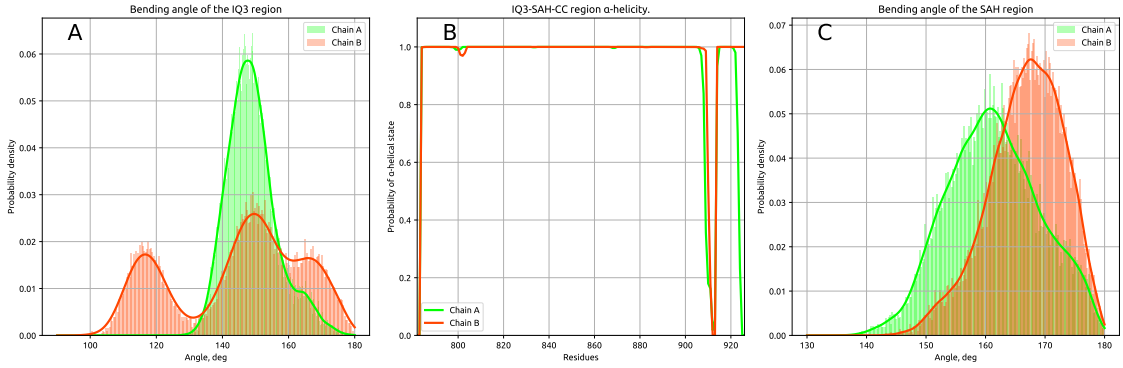


Figure 12.2.: Conformational dynamics of the lever-arm dimer during MD simulation. **A.** Bending angle of the IQ3 region. **B.** α-Helical content of each chain during the simulation. **C.** Bending angle of the SAH region. Adapted from (Ropars et al. 2016).

The simulation indeed reveals considerable flexibility in the overall shape and bending state of the lever-arm dimer, but no melting of the SAH domain; see Figure 12.2. Rearrangements in the coiled-coil dimerization region were observed that led to a rotation of each individual chain, changing the conformation of the whole assembly (Figure 12.3). Even though the explored conformations differ by the relative orientation of the two chains, they have in common to represent more extended structures than the crystallographic one, as measured by the distance between residues 813 of chain A and B, see Figure 12.4. Crucially, this new spreading distance is consistent with the step size measured by single molecule experiments and reveals a novel configuration of the motor dimer that allows it to take extended steps, see (Ropars et al. 2016, Figure 5).

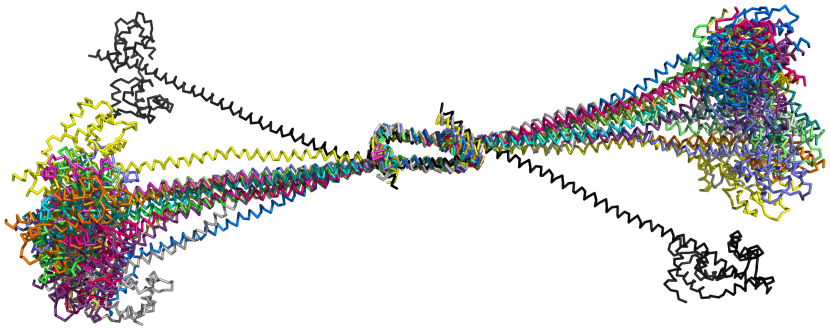


Figure 12.3.: Conformational ensemble explored during MD simulation (colored structures), as compared to the crystal structure (black). An in-place rotation of each chain at the level of the coiled-coil domain re-orients them and allows for the exploration of more extended structures.

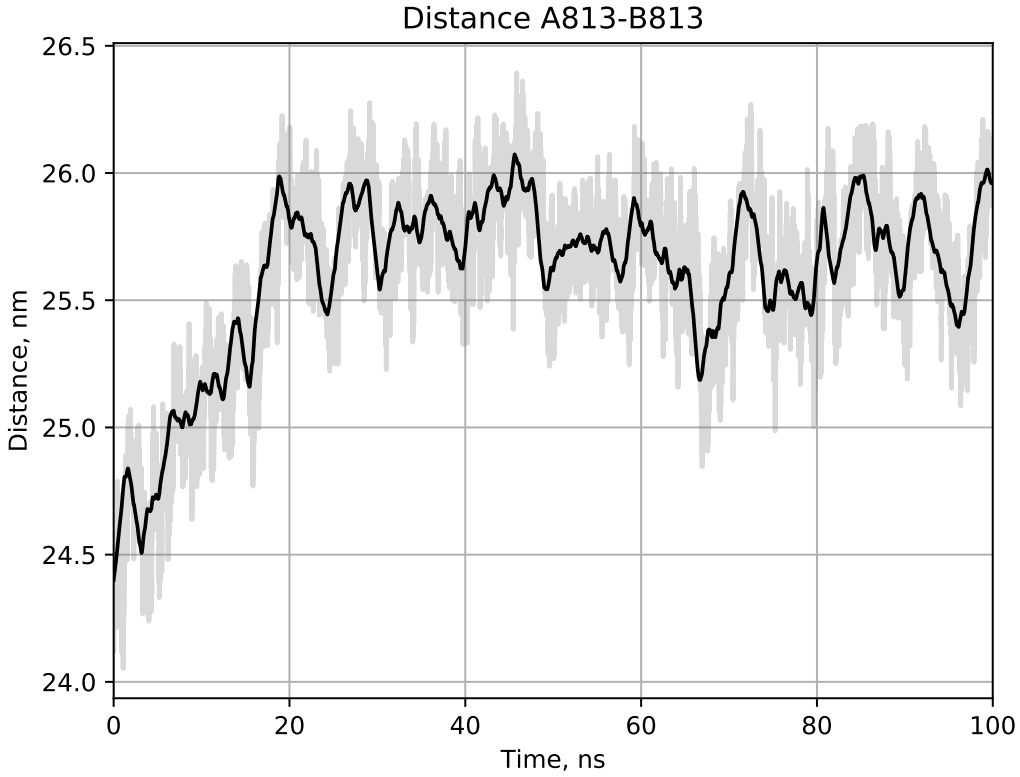


Figure 12.4.: Evolution of the distance between residues 813 of chains A and B during MD simulation. Adapted (Ropars et al. 2016).

12.2. Conformational dynamics of MyTH-FERM domains, an important category of myosin tail domains

12.2.1. Context

Myosins exhibiting MyTH-FERM domains in their tail, such as myosin VII or myosin X, are involved in a wide range of crucial cellular processes; notably, their ability to bind microtubules makes them a major link between the actin and microtubule cytoskeletons (Planelles-Herrero, Blanc, et al. 2016). MyTH-FERM myosins are widely conserved and found to fulfill similar functions in evolutionary distant organisms, *e.g.* when comparing *Dictyostelium discoideum* to Mammalians. Novel crystal structures of the two MyTH-FERM (MF) domains of *Dictyostelium discoideum* myosin VII have been solved by Dr Vincente Planelles-Herrero (Houdusse group), see Figure 12.5. Interestingly, the FERM lobes exhibit a so-called ”clover-leaf” spatial organization which is also observed on previously reported Mammalian MF structures (Hirano et al. 2011). Interestingly, the specifics of the relative orientation of the lobes differ between MF isoforms, which is due to sequence divergence in the connecting loops. The general idea defended in (Planelles-Herrero, Blanc, et al. 2016) is that the clover-leaf organization provides a ”multifunctional platform” for the binding to cellular partners, in which small re-orientations of the FERM lobes can modulate the binding affinity, or create binding site for new partners. Thus, over evolutionary time, new functions for the MF domains emerge by ”molecular tinkering” as the molecular evolution of the connecting loops gives rise to new binding opportunities by slightly rearranging the lobes, while the overall clover-leaf organization is maintained.

By contrast, talin, a 3-FERM lobes protein involved in connecting integrins to the cytoskeleton, exhibits a linear (rather than clover-leaf) arrangement of the FERM lobes, see Figure 12.6. This different arrangement is attributed to drastically different connecting loops from the ones observed in myosin MF domains. In this context, we use MD simulations in explicit solvation to compare the flexibility of the clover-leaf and linear organizations of the FERM lobes.

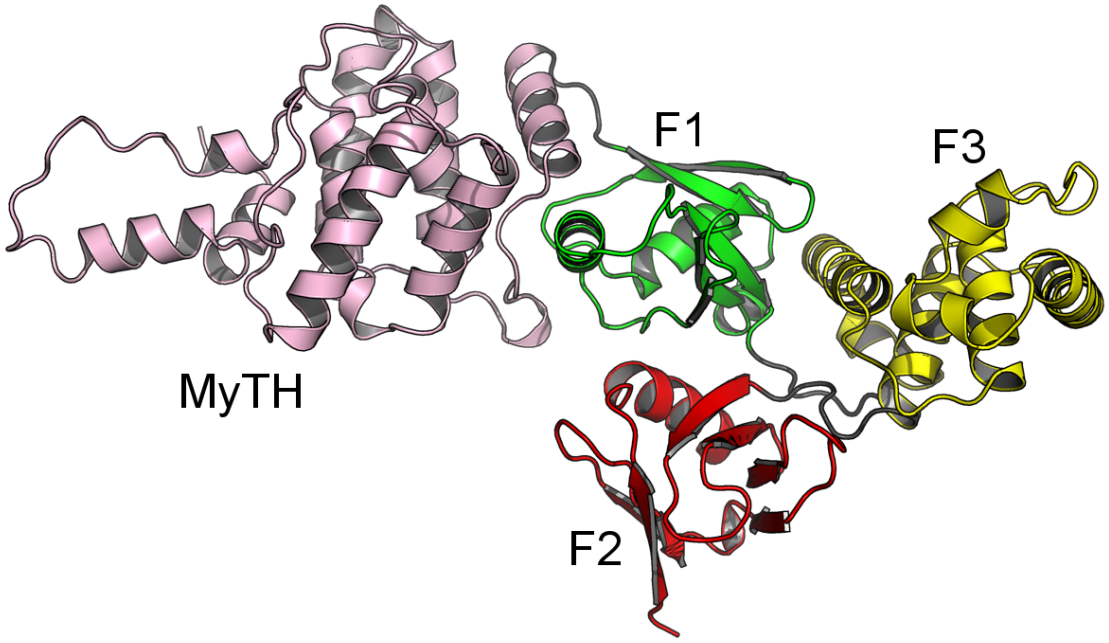


Figure 12.5.: Structure of the MyTH-FERM domain 1 of *Dictyostelium discoideum* Myosin VII. The FERM lobes are arranged in a cloverleaf organization.

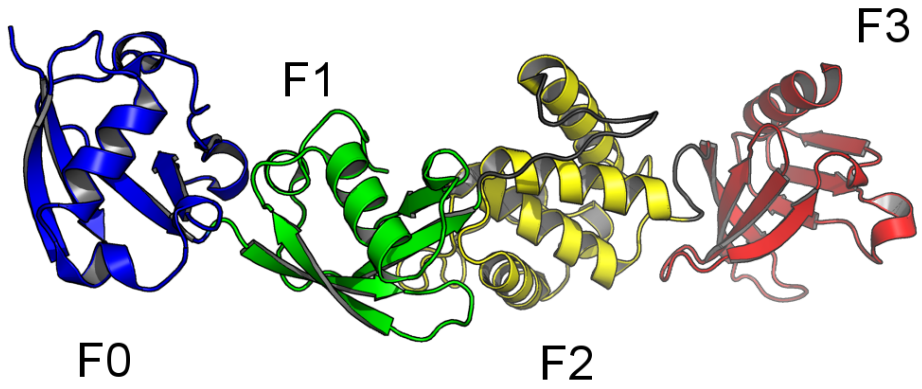


Figure 12.6.: Structure of the FERM domain of talin. The lobes are arranged in a linear organization.

12.2.2. Simulation results

Explicitly solvated models of the two MF domains of *Dictyostelium discoideum* myosin VII and of talin were prepared, minimized, heated and equilibrated according to the protocol described for the

myosin VI motor domain in chapter 6. Production MD simulations were performed on the 30 ns timescale. The RMSD of the backbone atoms of the FERM lobes from their respective crystal structures is shown on Figure 12.7. Briefly, the data shows that the RMSD of the FERM lobes in the MF domains is significantly smaller than that of talin; thus, it suggests that the cloverleaf organization of the FERM lobes in the myosin tail is structurally stable over time, while the linear arrangement of talin is not.

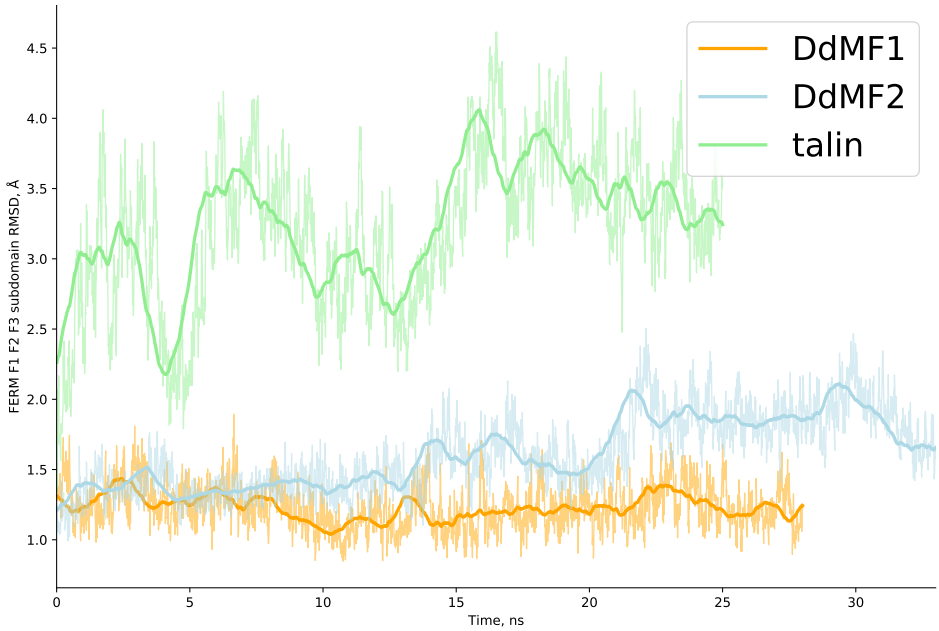


Figure 12.7.: Evolution of the RMSD of the FERM lobes during explicit solvent MD simulation.

These results illustrate how linker sequences between FERM lobes may impart different flexibility patterns to the entire FERM domain, and notably highlights that myosin MyTH-FERM domains in the clover-leaf organization are rather rigid. It is proposed that this stability is important in favoring the molecular recognition of the binding partners to myosin tails; the reader is referred to (Planelles-Herrero, Blanc, et al. 2016) for a more precise discussion of the complete results.

Bibliography

Abrams, C. F. and Eric Vanden-Eijnden (2010). “Large-scale conformational sampling of proteins using temperature-accelerated molecular dynamics”. In: Proceedings of the National Academy of Sciences 107.11. DOI: 10.1073/pnas.0914540107.

Adelstein, R S and E Eisenberg (1980). “Regulation and Kinetics of the Actin-Myosin-ATP Interaction”. In: Annual Review of Biochemistry 49.1. DOI: 10.1146/annurev.bi.49.070180.004421.

Agafonov, Roman V., Igor V. Negrashov, Yaroslav V. Tkachev, Sarah E. Blakely, Margaret A. Titus, David D. Thomas, and Yuri E. Nesmelov (2009). “Structural dynamics of the myosin relay helix by time-resolved EPR and FRET”. In: Proceedings of the National Academy of Sciences 106.51. DOI: 10.1073/pnas.0909757106.

Ahmed, Zubair M., Robert J. Morell, Saima Riazuddin, Andrea Gropman, Shahzad Shaukat, Mussaber M. Ahmad, Saidi A. Mohiddin, Lameh Fananapazir, Rafael C. Caruso, Tayyab Husnain, et al. (2003). “Mutations of MYO6 are associated with recessive deafness, DFNB37”. In: The American Journal of Human Genetics 72.5.

Alberts, Bruce, ed. (2008). Molecular biology of the cell. 5th ed. New York: Garland Science. 1 p. ISBN: 978-0-8153-4105-5.

Alder, B. J. and T. E. Wainwright (1957). “Phase Transition for a Hard Sphere System”. In: The Journal of Chemical Physics 27.5. DOI: 10.1063/1.1743957.

Alder, B. J. and T. E. Wainwright (1959). “Studies in Molecular Dynamics. I. General Method”. In: The Journal of Chemical Physics 31.2. DOI: 10.1063/1.1730376.

Andersen, Hans C. (1980). “Molecular dynamics simulations at constant pressure and/or temperature”. In: The Journal of Chemical Physics 72.4. DOI: 10.1063/1.439486.

Ashley, Steven (2015). “Core Concept: Ergodic theory plays a key role in multiple fields:” in: Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.1500429112.

Astumian, R. Dean (2016). “Artificial molecular motors: Running on information”. In: Nature nanotechnology 11.7.

Astumian, R. Dean (2007). “Design principles for Brownian molecular machines: how to swim in molasses and walk in a hurricane”. en. In: Physical Chemistry Chemical Physics 9.37. DOI: 10.1039/b708995c.

Astumian, R. Dean (2015). “Irrelevance of the Power Stroke for the Directionality, Stopping Force, and Optimal Efficiency of Chemically Driven Molecular Machines”. en. In: Biophysical Journal 108.2. DOI: 10.1016/j.bpj.2014.11.3459.

Astumian, R. Dean (2012). “Microscopic reversibility as the organizing principle of molecular machines”. In: Nature Nanotechnology 7.11. DOI: 10.1038/nnano.2012.188.

Astumian, R. Dean (2010). “Thermodynamics and Kinetics of Molecular Motors”. en. In: Biophysical Journal 98.11. DOI: 10.1016/j.bpj.2010.02.040.

Baker, Nathan A., David Sept, Simpson Joseph, Michael J. Holst, and J. Andrew McCammon (2001). “Electrostatics of nanosystems: Application to microtubules and the ribosome”. In: Proceedings of the National Academy of Sciences 98.18. DOI: 10.1073/pnas.181342398.

Banushkina, Polina V. and Sergei V. Krivov (2016). "Optimal reaction coordinates". In: Wiley Interdisciplinary Reviews: Computational Molecular Science. DOI: 10.1002/wcms.1276.

Barducci, Alessandro, Massimiliano Bonomi, and Michele Parrinello (2011). "Metadynamics". In: Wiley Interdisciplinary Reviews: Computational Molecular Science 1.5. DOI: 10.1002/wcms.31.

Barducci, Alessandro, Giovanni Bussi, and Michele Parrinello (2008). "Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method". In: Physical Review Letters 100.2. DOI: 10.1103/PhysRevLett.100.020603.

Bashford, Donald and Martin Karplus (1991). "Multiple-site titration curves of proteins: an analysis of exact and approximate methods for their calculation". In: The Journal of Physical Chemistry 95.23. DOI: 10.1021/j100176a093.

Batra, Renu, Michael A. Geeves, and Dietmar J. Manstein (1999). "Kinetic Analysis of *Dictyostelium discoideum* Myosin Motor Domains with Glycine-to-Alanine Mutations in the Reactive Thiol Region †". In: Biochemistry 38.19. DOI: 10.1021/bi982251e.

Batters, Christopher and Claudia Veigel (2016). "Mechanics and activation of unconventional myosins". In: Traffic. DOI: 10.1111/tra.12400.

Baumketner, Andrij (2012a). "Interactions between relay helix and Src homology 1 (SH1) domain helix drive the converter domain rotation during the recovery stroke of myosin II". In: Proteins: Structure, Function, and Bioinformatics 80.6. DOI: 10.1002/prot.24051.

Baumketner, Andrij (2012b). "The mechanism of the converter domain rotation in the recovery stroke of myosin motor protein". In: Proteins: Structure, Function, and Bioinformatics 80.12. DOI: 10.1002/prot.24155.

Baumketner, Andrij and Yuri Nesmelov (2011). "Early stages of the recovery stroke in myosin II studied by molecular dynamics simulations". In: Protein Science 20.12. DOI: 10.1002/pro.737.

Berendsen, Herman J.C, J.P.M. Postma, Wilfred F. van Gunsteren, A. DiNola, and J.R. Haak (1984). "Molecular dynamics with coupling to an external bath". In: The Journal of Chemical Physics 81.8. DOI: 10.1063/1.448118.

Berg, JS, BH Derfler, CM Pennisi, DP Corey, and RE Cheney (2001). "Myosin-X, a novel myosin with pleckstrin homology domains, associates with regions of dynamic actin". In: Journal of Cell Science 113.

Blanc, Florian, Tatiana Isabet, Hannah Benisty, H. Lee Sweeney, Marco Cecchini, and Anne Houdusse (2018). "An intermediate along the recovery stroke of myosin VI revealed by X-ray crystallography and molecular dynamics". In: Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.1711512115.

Bloemink, Marieke J. and Michael A. Geeves (2011). "Shaking the myosin family tree: Biochemical kinetics defines four types of myosin motor". In: Seminars in Cell & Developmental Biology 22.9. DOI: 10.1016/j.semcdb.2011.09.015.

Boëda, Batiste, Aziz El□Amraoui, Amel Bahloul, Richard Goodyear, Laurent Daviet, Stéphane Blanchard, Isabelle Perfettini, Karl R. Fath, Spencer Shorte, Jan Reiners, Anne Houdusse, Pierre Legrain, Uwe Wolfrum, Guy Richardson, and Christine Petit (2002). "Myosin VIIa, harmonin and cadherin 23, three Usher I gene products that cooperate to shape the sensory hair cell bundle". In: The EMBO Journal 21.24. DOI: 10.1093/emboj/cdf689.

Boehr, David D, Ruth Nussinov, and Peter E Wright (2009). "The role of dynamic conformational ensembles in biomolecular recognition". In: Nature Chemical Biology 5.11. DOI: 10.1038/nchembio.232.

Bohil, A. B., B. W. Robertson, and R. E. Cheney (2006). "Myosin-X is a molecular motor that functions in filopodia formation". In: Proceedings of the National Academy of Sciences 103.33. DOI: 10.1073/pnas.0602443103.

Bolhuis, Peter G., David Chandler, Christoph Dellago, and Phillip L. Geissler (2002). "TRANSITION PATH SAMPLING: Throwing Ropes Over Rough Mountain Passes, in the Dark". In: Annual Review of Physical Chemistry 53.1. DOI: 10.1146/annurev.physchem.53.082301.113146.

Bonomi, Massimiliano, Alessandro Barducci, and Michele Parrinello (2009). "Reconstructing the equilibrium Boltzmann distribution from well-tempered metadynamics". In: Journal of Computational Chemistry 30.11. DOI: 10.1002/jcc.21305.

Branduardi, Davide, Francesco Luigi Gervasio, and Michele Parrinello (2007). "From A to B in free energy space". In: The Journal of Chemical Physics 126.5. DOI: 10.1063/1.2432340.

Brooks, B. R., C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, Victor Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and Martin Karplus (2009). "CHARMM: The biomolecular simulation program". In: Journal of Computational Chemistry 30.10. DOI: 10.1002/jcc.21287.

Brooks, Bernard R., Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus (1983). "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations". In: Journal of Computational Chemistry 4.2. DOI: 10.1002/jcc.540040211.

Brünger, Axel, Charles L. Brooks, and Martin Karplus (1984). "Stochastic boundary conditions for molecular dynamics simulations of ST2 water". In: Chemical Physics Letters 105.5. DOI: 10.1016/0009-2614(84)80098-6.

Bryant, Z., D. Altman, and J. A. Spudich (2007). "The power stroke of myosin VI and the basis of reverse directionality". In: Proceedings of the National Academy of Sciences 104.3. DOI: 10.1073/pnas.0610144104.

Bussi, Giovanni, Davide Donadio, and Michele Parrinello (2007). "Canonical sampling through velocity rescaling". In: The Journal of Chemical Physics 126.1. DOI: 10.1063/1.2408420.

Bussi, Giovanni, Alessandro Laio, and Michele Parrinello (2006). "Equilibrium Free Energies from Nonequilibrium Metadynamics". In: Physical Review Letters 96.9. DOI: 10.1103/PhysRevLett.96.090601.

Campbell, Eleanor C, Galen J Correy, Peter D Mabbitt, Ashley M Buckle, Nobuhiko Tokuriki, and Colin J Jackson (2018). "Laboratory evolution of protein conformational dynamics". In: Current Opinion in Structural Biology 50. DOI: 10.1016/j.sbi.2017.09.005.

Carter, E.A., Giovanni Ciccotti, James T. Hynes, and Raymond Kapral (1989). "Constrained reaction coordinate dynamics for the simulation of rare events". In: Chemical Physics Letters 156.5. DOI: 10.1016/S0009-2614(89)87314-2.

Castiglione, Patrizia, Massimo Falcioni, Annick Lesne, Angelo Vulpiani, and Cambridge University Press (2008). Chaos and Coarse Graining in Statistical Mechanics. Cambridge: Cambridge University Press.

Cecchini, Marco, Yuri Alexeev, and Martin Karplus (2010). "Pi Release from Myosin: A Simulation Analysis of Possible Pathways". In: Structure 18.4. DOI: 10.1016/j.str.2010.01.014.

Cecchini, Marco, Anne Houdusse, and Martin Karplus (2008). "Allosteric Communication in Myosin V: From Small Conformational Changes to Large Directed Movements". In: PLoS Computational Biology 4.8. Ed. by Matthew P. Jacobson. DOI: 10.1371/journal.pcbi.1000129.

Cecchini, Marco, S. V. Krivov, M. Spichty, and Martin Karplus (2009). "Calculation of Free-Energy Differences by Confinement Simulations. Application to Peptide Conformers". In: The Journal of Physical Chemistry B 113.29. DOI: 10.1021/jp9020646.

Chavrier, Philippe (2002). "May the force be with you: Myosin-X in phagocytosis." In: Nature Cell Biology 4.7. DOI: 10.1038/ncb0702-e169.

Chipot, Christophe and Jeffrey Comer (2016). "Subdiffusion in Membrane Permeation of Small Molecules". In: Scientific Reports 6.1. DOI: 10.1038/srep35913.

Chipot, Christophe and Jérôme Hénin (2005). "Exploring the free-energy landscape of a short peptide using an average force". In: The Journal of Chemical Physics 123.24. DOI: 10.1063/1.2138694.

Chipot, Christophe and Tony Lelièvre (2011). "Enhanced Sampling of Multidimensional Free-Energy Landscapes Using Adaptive Biasing Forces". In: SIAM Journal on Applied Mathematics 71.5. DOI: 10.1137/10080600X.

Chipot, Christophe and Andrew Pohorille, eds. (2007). Free energy calculations: theory and applications in chemistry and biology. Springer series in chemical physics 86. Berlin ; New York: Springer. 517 pp. ISBN: 978-3-540-38447-2.

Ciccotti, Giovanni, Raymond Kapral, and Eric Vanden-Eijnden (2005). "Blue Moon Sampling, Vectorial Reaction Coordinates, and Unbiased Constrained Dynamics". In: ChemPhysChem 6.9. DOI: 10.1002/cphc.200400669.

Ciccotti, Giovanni and Eric Vanden-Eijnden (2015). "The trees and the forest: Aims and objectives of molecular dynamics simulations". In: The European Physical Journal Special Topics 224.12. DOI: 10.1140/epjst/e2015-02537-1.

Cohen-Tannoudji, Claude, Bernard Diu, and Franck Laloë (2008). Mécanique quantique. OCLC: 717703469. Paris: Hermann.

Comer, Jeffrey, Christophe Chipot, and Fernando D. González-Nilo (2013). "Calculating Position-Dependent Diffusivity in Biased Molecular Dynamics Simulations". In: Journal of Chemical Theory and Computation 9.2. DOI: 10.1021/ct300867e.

Comer, Jeffrey, James C. Gumbart, Jérôme Hénin, Tony Lelièvre, Andrew Pohorille, and Christophe Chipot (2015). "The Adaptive Biasing Force Method: Everything You Always Wanted To Know but Were Afraid To Ask". In: The Journal of Physical Chemistry B 119.3. DOI: 10.1021/jp506633n.

Comer, Jeffrey, James C. Phillips, Klaus Schulten, and Christophe Chipot (2014). "Multiple-Replica Strategies for Free-Energy Calculations in NAMD: Multiple-Walker Adaptive Biasing Force and Walker Selection Rules". In: Journal of Chemical Theory and Computation 10.12. DOI: 10.1021/ct500874p.

Conti, Simone and Marco Cecchini (2018). "Modeling the adsorption equilibrium of small-molecule gases on graphene: effect of the volume to surface ratio". In: Physical Chemistry Chemical Physics 20.15. DOI: 10.1039/C7CP08047F.

Cornell, Wendy D., Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman (1995). "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules". In: Journal of the American Chemical Society 117.19. DOI: 10.1021/ja00124a002.

Coureux, Pierre-Damien, Amber L. Wells, Julie Menetrey, Christopher M. Yengo, Carl A. Morris, H. Lee Sweeney, and Anne Houdusse (2003). "A structural state of the myosin V motor without bound nucleotide". In: Nature 425.6956. DOI: 10.1038/nature01927.

Darden, Tom, Darrin York, and Lee Pedersen (1993). "Particle mesh Ewald: An N□log(N) method for Ewald sums in large systems". In: The Journal of Chemical Physics 98.12. DOI: 10.1063/1.464397.

Darve, Eric and Andrew Pohorille (2001). “Calculating free energies using average force”. In: The Journal of Chemical Physics 115.20. DOI: 10.1063/1.1410978.

Darve, Eric, David Rodríguez-Gómez, and Andrew Pohorille (2008). “Adaptive biasing force method for scalar and vector free energy calculations”. In: The Journal of Chemical Physics 128.14. DOI: 10.1063/1.2829861.

Darve, Eric, Michael A. Wilson, and Andrew Pohorille (2002). “Calculating Free Energies Using a Scaled-Force Molecular Dynamics Algorithm”. In: Molecular Simulation 28.1. DOI: 10.1080/08927020211975.

Das, Avisek, Huan Rui, Robert Nakamoto, and Benoît Roux (2017). “Conformational Transitions and Alternating-Access Mechanism in the Sarcoplasmic Reticulum Calcium Pump”. In: Journal of Molecular Biology 429.5. DOI: 10.1016/j.jmb.2017.01.007.

Davis, Ian W., Andrew Leaver-Fay, Vincent B. Chen, Jeremy N. Block, Gary J. Kapral, Xueyi Wang, Laura W. Murray, W. Bryan Arendall, Jack Snoeyink, Jane S. Richardson, and David C. Richardson (2007). “MolProbity: all-atom contacts and structure validation for proteins and nucleic acids”. In: Nucleic Acids Research 35 (suppl 2). DOI: 10.1093/nar/gkm216.

Dawkins, Richard (1976). The selfish gene. Oxford: Oxford University Press. 224 pp.

De La Cruz, E. M., A. L. Wells, S. S. Rosenfeld, E. M. Ostap, and H. Lee Sweeney (1999). “The kinetic mechanism of myosin V”. In: Proceedings of the National Academy of Sciences 96.24. DOI: 10.1073/pnas.96.24.13726.

Dietrich-Buchecker, C. O., M. C. Jimenez-Molero, V. Sartor, and J.-P. Sauvage (2003). “Rotaxanes and catenanes as prototypes of molecular machines and motors”. In: Pure and Applied Chemistry 75.10. DOI: 10.1351/pac200375101383.

Diu, Bernard (1989). Eléments de physique statistique. Paris: Hermann.

Dunn, Thomas A., Shenglin Chen, Dennis A. Faith, Jessica L. Hicks, Elizabeth A. Platz, Yidong Chen, Charles M. Ewing, Jurga Sauvageot, William B. Isaacs, Angelo M. De Marzo, and Jun Luo (2006). “A Novel Role of Myosin VI in Human Prostate Cancer”. In: The American Journal of Pathology 169.5. DOI: 10.2353/ajpath.2006.060316.

Duplantier, Bertrand (2005). “Le Mouvement Brownien, Divers et Ondoyant”. In: Séminaire Poincaré.

E, Weinan, Weiqing Ren, and Eric Vanden-Eijnden (2005a). “Finite Temperature String Method for the Study of Rare Events”. In: The Journal of Physical Chemistry B 109.14. DOI: 10.1021/jp0455430.

E, Weinan, Weiqing Ren, and Eric Vanden-Eijnden (2007). “Simplified and improved string method for computing the minimum energy paths in barrier-crossing events”. In: The Journal of Chemical Physics 126.16. DOI: 10.1063/1.2720838.

E, Weinan, Weiqing Ren, and Eric Vanden-Eijnden (2002). “String method for the study of rare events”. In: Physical Review B 66.5. DOI: 10.1103/PhysRevB.66.052301.

E, Weinan, Weiqing Ren, and Eric Vanden-Eijnden (2005b). “Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes”. In: Chemical Physics Letters 413.1. DOI: 10.1016/j.cplett.2005.07.084.

E, Weinan and Eric Vanden-Eijnden (2010). “Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events”. In: Annual Review of Physical Chemistry 61.1. DOI: 10.1146/annurev.physchem.040808.090412.

Elber, Ron and Anthony West (2010). “Atomically detailed simulation of the recovery stroke in myosin by Milestoning”. In: Proceedings of the National Academy of Sciences 107.11. DOI: 10.1073/pnas.0909636107.

Emsley, P. and K. Cowtan (2004). "Coot: model-building tools for molecular graphics". In: Acta Crystallographica Section D: Biological Crystallography 60.12. DOI: 10.1107/S0907444904019158.

Ensing, Bernd, Marco De Vivo, Zhiwei Liu, Preston Moore, and Michael L. Klein (2006). "Metadynamics as a Tool for Exploring Free Energy Landscapes of Chemical Reactions". In: Accounts of Chemical Research 39.2. DOI: 10.1021/ar040198i.

Esque, Jeremy and Marco Cecchini (2015). "Accurate Calculation of Conformational Free Energy Differences in Explicit Water: The Confinement–Solvation Free Energy Approach". In: The Journal of Physical Chemistry B 119.16. DOI: 10.1021/acs.jpcb.5b01632.

Eyring, Henry (1935). "The Activated Complex in Chemical Reactions". In: The Journal of Chemical Physics 3.2. DOI: 10.1063/1.1749604.

Faradjian, Anton K. and Ron Elber (2004). "Computing time scales from reaction coordinates by milestoning". In: The Journal of Chemical Physics 120.23. DOI: 10.1063/1.1738640.

Feringa, Ben L. (2001). "In Control of Motion: From Molecular Switches to Molecular Motors †". In: Accounts of Chemical Research 34.6. DOI: 10.1021/ar0001721.

Feringa, Ben L. (2016). "The Art of Building Small: from Molecular Switches to Motors". In:

Feynman, Richard P., Robert B. Leighton, and Matthew L. Sands (2011). The Feynman lectures on physics. New millennium ed. OCLC: ocn671704374. New York: Basic Books. 3 pp.

Fiorin, Giacomo, Michael L. Klein, and Jérôme Hénin (2013). "Using collective variables to drive molecular dynamics simulations". In: Molecular Physics 111.22. DOI: 10.1080/00268976.2013.813594.

Fischer, Stefan and Martin Karplus (1992). "Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom". In: Chemical Physics Letters 194.3. DOI: 10.1016/0009-2614(92)85543-J.

Fischer, Stefan, Björn Windshügel, Daniel Horak, Kenneth C. Holmes, and Jeremy C. Smith (2005). "Structural mechanism of the recovery stroke in the Myosin molecular motor". In: Proceedings of the National Academy of Sciences of the United States of America 102.19. DOI: 10.1073/pnas.0408784102.

Fiser, András and Andrej Šali (2003). "Modeller: Generation and Refinement of Homology-Based Protein Structure Models". In: Methods in Enzymology. Vol. 374. Elsevier, pp. 461–491. ISBN: 978-0-12-182777-9. DOI: 10.1016/S0076-6879(03)74020-8.

Fisher, Andrew J., Clyde A. Smith, James Thoden, Robert Smith, Kazuo Sutoh, Hazel M. Holden, and Ivan Rayment (1995). "X-ray Structures of the Myosin Motor Domain of Dictyostelium discoideum Complexed with MgADP.BeFx and MgADP.AlF4-". In: Biochemistry 34.28. DOI: 10.1021/bi00028a004.

Foth, Bernardo J., Marc C. Goedecke, and Dominique Soldati (2006). "New insights into myosin evolution and classification". In: Proceedings of the National Academy of Sciences of the United States of America 103.10.

Frenkel, Daan and Berend Smit (2002). Understanding molecular simulation: from algorithms to applications. 2nd ed. 1. San Diego: Academic Press. 638 pp.

Fu, Haohao, Xueguang Shao, Christophe Chipot, and Wensheng Cai (2016). "Extended Adaptive Biasing Force Algorithm. An On-the-Fly Implementation for Accurate Free-Energy Calculations". In: Journal of Chemical Theory and Computation 12.8. DOI: 10.1021/acs.jctc.6b00447.

Fukunishi, Hiroaki, Osamu Watanabe, and Shoji Takada (2002). "On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction". In: The Journal of Chemical Physics 116.20. DOI: 10.1063/1.1472510.

G. Bricogne, E. Blanc, M. Brandl, C. Flensburg, P. Keller, W. Paciorek, P. Roversi, A. Sharff, O.S. Smart, C. Vonrhein, and T.O. Womack (2011). BUSTER version 2.11. 2. Cambridge, UK.

Gallavotti, Giovanni (1998). "Chaotic dynamics, fluctuations, nonequilibrium ensembles". In: Chaos: An Interdisciplinary Journal of Nonlinear Science 8.2. DOI: 10.1063/1.166320.

Geeves, Michael A. and Kenneth C. Holmes (1999). "Structural Mechanism of Muscle Contraction". In: Annual Review of Biochemistry 68.1. DOI: 10.1146/annurev.biochem.68.1.687.

Geisterfer-Lowrance, Anja A.T., Susan Kass, Gary Tanigawa, Hans-Peter Vosberg, William McKenna, Christine E. Seidman, and J.G. Seidman (1990). "A molecular basis for familial hypertrophic cardiomyopathy: A β cardiac myosin heavy chain gene missense mutation". In: Cell 62.5. DOI: 10.1016/0092-8674(90)90274-I.

Grant, Barry J., Alemayehu A. Gorfe, and J. Andrew McCammon (2009). "Ras Conformational Switching: Simulating Nucleotide-Dependent Conformational Transitions with Accelerated Molecular Dynamics". In: PLoS Computational Biology 5.3. Ed. by James M. Briggs. DOI: 10.1371/journal.pcbi.1000325.

Greenberg, Michael J. and E. Michael Ostap (2013). "Regulation and control of myosin-I by the motor and light chain-binding domains". In: Trends in Cell Biology 23.2. DOI: 10.1016/j.tcb.2012.10.008.

Grigorenko, B. L., A. V. Rogov, Igor A. Topol, S. K. Burt, H. M. Martinez, and A. V. Nemukhin (2007). "Mechanism of the myosin catalyzed hydrolysis of ATP as rationalized by molecular modeling". In: Proceedings of the National Academy of Sciences 104.17. DOI: 10.1073/pnas.0701727104.

Grubmüller, Helmut (1995). "Predicting slow structural transitions in macromolecular systems: Conformational flooding". In: Physical Review E 52.3. DOI: 10.1103/PhysRevE.52.2893.

Grubmüller, Helmut, B. Heymann, and P. Tavan (1996). "Ligand Binding: Molecular Mechanics Calculation of the Streptavidin-Biotin Rupture Force". In: Science 271.5251. DOI: 10.1126/science.271.5251.997.

Habeck, Michael (2012). "Bayesian Estimation of Free Energies From Equilibrium Simulations". In: Physical Review Letters 109.10. DOI: 10.1103/PhysRevLett.109.100601.

Haberthür, Urs and Amedeo Caflisch (2008). "FACTS: Fast analytical continuum treatment of solvation". In: Journal of Computational Chemistry 29.5. DOI: 10.1002/jcc.20832.

Hairer, E., Christian Lubich, and Gerhard Wanner (2006). Geometric numerical integration: structure-preserving algorithms for ordinary differential equations. 2nd ed. 31. Berlin ; New York: Springer. 644 pp.

Hairer, Ernst, Christian Lubich, and Gerhard Wanner (2003). "Geometric numerical integration illustrated by the Störmer-Verlet method". In: Acta Numerica 12. DOI: 10.1017/S0962492902000144.

Hamelberg, Donald, César Augusto F. de Oliveira, and J. Andrew McCammon (2007). "Sampling of slow diffusive conformational transitions with accelerated molecular dynamics". In: The Journal of Chemical Physics 127.15. DOI: 10.1063/1.2789432.

Hammes-Schiffer, Sharon and Stephen J. Benkovic (2006). "Relating protein motion to catalysis". In: Annu. Rev. Biochem. 75.

Hänggi, Peter, Michal Borkovec, and Peter Talkner (1990). "Reaction-rate theory: fifty years after Kramers". In: Reviews of Modern Physics 62.2. DOI: 10.1103/RevModPhys.62.251.

Hanwell, Marcus D., Donald E. Curtis, David C. Lonie, Tim Vandermeersch, Eva Zurek, and Geoffrey R. Hutchison (2012). "Avogadro: an advanced semantic chemical editor, visualization, and analysis platform". In: Journal of Cheminformatics 4.1. DOI: 10.1186/1758-2946-4-17.

Harris, Michael J. and Hyung-June Woo (2008). "Energetics of subdomain movements and fluorescence probe solvation environment change in ATP-bound myosin". In: European Biophysics Journal 38.1. DOI: 10.1007/s00249-008-0347-3.

Hartman, M. Amanda, Dina Finan, Sivaraj Sivaramakrishnan, and James A. Spudich (2011). "Principles of Unconventional Myosin Function and Targeting". In: Annual Review of Cell and Developmental Biology 27.1. DOI: 10.1146/annurev-cellbio-100809-151502.

Harvey, Stephen C., Robert K.-Z. Tan, and Thomas E. Cheatham (1998). "The flying ice cube: Velocity rescaling in molecular dynamics leads to violation of energy equipartition". In: Journal of Computational Chemistry 19.7. DOI: 10.1002/(SICI)1096-987X(199805)19:7<726::AID-JCC4>3.0.CO;2-S.

Hashem, S., M. Tiberti, and A. Fornili (2017). "Allosteric modulation of cardiac myosin dynamics by omecamtiv mecarbil." In: PLoS computational biology 13.11.

Hay, Sam and Nigel S. Scrutton (2012). "Good vibrations in enzyme-catalysed reactions". In: Nature Chemistry 4.3. DOI: 10.1038/nchem.1223.

Hénin, Jérôme and Christophe Chipot (2004). "Overcoming free energy barriers using unconstrained molecular dynamics simulations". In: The Journal of Chemical Physics 121.7. DOI: 10.1063/1.1773132.

Hénin, Jérôme, Giacomo Fiorin, Christophe Chipot, and Michael L. Klein (2010). "Exploring Multidimensional Free Energy Landscapes Using Time-Dependent Biases on Collective Variables". In: Journal of Chemical Theory and Computation 6.1. DOI: 10.1021/ct9004432.

Henzler-Wildman, Katherine and Dorothee Kern (2007). "Dynamic personalities of proteins". In: Nature 450.7172. DOI: 10.1038/nature06522.

Hill, Terrell L. (1986). An introduction to statistical thermodynamics. New York: Dover Publications. 508 pp.

Hill, Terrell L. (2005). Free energy transduction and biochemical cycle kinetics. Mineola, N.Y: Dover Publications. 119 pp.

Himmel, D. M., S. Gourinath, L. Reshetnikova, Y. Shen, A. G. Szent-Györgyi, and C. Cohen (2002). "Crystallographic findings on the internally uncoupled and near-rigor states of myosin: Further insights into the mechanics of the motor". In: Proceedings of the National Academy of Sciences 99.20. DOI: 10.1073/pnas.202476799.

Hirano, Yoshinori, Taiki Hatano, Aya Takahashi, Michinori Toriyama, Naoyuki Inagaki, and Toshio Hakoshima (2011). "Structural basis of cargo recognition by the myosin-X MyTH4-FERM domain: Myosin-X binding to DCC, integrin and microtubule". In: The EMBO Journal 30.13. DOI: 10.1038/emboj.2011.177.

Hoffmann, Peter M. (2012). Life's ratchet: how molecular machines extract order from chaos. New York: Basic Books. 278 pp.

Holmes, Kenneth C. (1997). "The swinging lever-arm hypothesis of muscle contraction". In: Current Biology 7.2.

Hoover, William G. (1985). "Canonical dynamics: Equilibrium phase-space distributions". In: Physical Review A 31.3. DOI: 10.1103/PhysRevA.31.1695.

Houdusse, Anne, Vassilios N. Kalabokis, Daniel Himmel, Andrew G. Szent-Györgyi, and Carolyn Cohen (1999). "Atomic structure of scallop myosin subfragment S1 complexed with MgADP: a novel conformation of the myosin head". In: Cell 97.4.

Houdusse, Anne and H. Lee Sweeney (2016). "How Myosin Generates Force on Actin Filaments". In: Trends in Biochemical Sciences 41.12. DOI: 10.1016/j.tibs.2016.09.006.

Houdusse, Anne and H. Lee Sweeney (2001). “Myosin motors: missing structures and hidden springs”. In: Current Opinion in Structural Biology 11.2. DOI: 10.1016/S0959-440X(00)00188-3.

Houdusse, Anne, Andrew G. Szent-Györgyi, and Carolyn Cohen (2000). “Three conformational states of scallop myosin S1”. In: Proceedings of the National Academy of Sciences 97.21.

Howard, Jonathon (2001). Mechanics of motor proteins and the cytoskeleton. eng. Nachdr. Sunderland, Mass: Sinauer.

Huang, Jing and Alexander D. MacKerell (2013). “CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data”. In: Journal of Computational Chemistry 34.25. DOI: 10.1002/jcc.23354.

Huber, Thomas, Andrew E. Torda, and Wilfred F. van Gunsteren (1994). “Local elevation: A method for improving the searching properties of molecular dynamics simulation”. In: Journal of Computer-Aided Molecular Design 8.6. DOI: 10.1007/BF00124016.

Hummer, Gerhard (2005). “Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations”. In: New Journal of Physics 7. DOI: 10.1088/1367-2630/7/1/034.

Hummer, Gerhard and Attila Szabo (2005). “Free Energy Surfaces from Single-Molecule Force Spectroscopy”. In: Accounts of Chemical Research 38.7. DOI: 10.1021/ar040148d.

Humphrey, William, Andrew Dalke, and Klaus Schulten (1996). “VMD: Visual molecular dynamics”. In: Journal of Molecular Graphics 14.1. DOI: 10.1016/0263-7855(96)00018-5.

Hunter, John D. (2007). “Matplotlib: A 2D Graphics Environment”. In: Computing in Science & Engineering 9.3. DOI: 10.1109/MCSE.2007.55.

Huxley, A. F. (1957). “Muscle structure and theories of contraction”. In: Progress in Biophysics and Biophysical Chemistry 7.

Huxley, A. F. and R. Niedergerke (1954). “Structural Changes in Muscle During Contraction: Interference Microscopy of Living Muscle Fibres”. In: Nature 173.4412. DOI: 10.1038/173971a0.

Huxley, H. E. (1957). “THE DOUBLE ARRAY OF FILAMENTS IN CROSS-STRIATED MUSCLE”. In: The Journal of Cell Biology 3.5. DOI: 10.1083/jcb.3.5.631.

Huxley, H. E. and Jean Hanson (1954). “Changes in the Cross-Striations of Muscle during Contraction and Stretch and their Structural Interpretation”. In: Nature 173.4412.

Iannuzzi, Marcella, Alessandro Laio, and Michele Parrinello (2003). “Efficient Exploration of Reactive Potential Energy Surfaces Using Car-Parrinello Molecular Dynamics”. In: Physical Review Letters 90.23. DOI: 10.1103/PhysRevLett.90.238302.

Izrailev, S., S. Stepaniants, M. Balsera, Y. Oono, and K. Schulten (1997). “Molecular dynamics study of unbinding of the avidin-biotin complex”. In: Biophysical Journal 72.4. DOI: 10.1016/S0006-3495(97)78804-0.

Jarzynski, C. (1997a). “Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach”. In: Physical Review E 56.5. DOI: 10.1103/PhysRevE.56.5018.

Jarzynski, C. (1997b). “Nonequilibrium Equality for Free Energy Differences”. In: Physical Review Letters 78.14. DOI: 10.1103/PhysRevLett.78.2690.

Johnson, Margaret E. and Gerhard Hummer (2012). “Characterization of a Dynamic String Method for the Construction of Transition Pathways in Molecular Reactions”. In: The Journal of Physical Chemistry B 116.29. DOI: 10.1021/jp212611k.

Jones, Eric, Travis Oliphant, and Pearu Peterson (2001). “{SciPy}: Open source scientific tools for {Python}”. In:

Jorgensen, William L., David S. Maxwell, and Julian Tirado-Rives (1996). "Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids". In: Journal of the American Chemical Society 118.45. DOI: 10.1021/ja9621760.

Jülicher, Frank, Armand Ajdari, and Jacques Prost (1997). "Modeling molecular motors". en. In: Reviews of Modern Physics 69.4. DOI: 10.1103/RevModPhys.69.1269.

Kabsch, W. (2010). "XDS". In: Acta Crystallographica Section D: Biological Crystallography 66.2. DOI: 10.1107/S0907444909047337.

Karagiannis, Peter, Yoshiharu Ishii, and Toshio Yanagida (2014). "Molecular Machines Like Myosin Use Randomness to Behave Predictably". In: Chemical Reviews 114.6. DOI: 10.1021/cr400344n.

Karplus, Martin (2011). "Behind the folding funnel diagram". In: Nature chemical biology 7.7.

Karplus, Martin (2006). "SPINACH ON THE CEILING: A Theoretical Chemist's Return to Biology". In: Annual Review of Biophysics and Biomolecular Structure 35.1. DOI: 10.1146/annurev.biophys.33.110502.133350.

Karplus, Martin and Yi Qin Gao (2004). "Biomolecular motors: the F1-ATPase paradigm". In: Current Opinion in Structural Biology 14.2. DOI: 10.1016/j.sbi.2004.03.012.

Kästner, Johannes (2009). "Umbrella integration in two or more reaction coordinates". In: The Journal of Chemical Physics 131.3. DOI: 10.1063/1.3175798.

Kästner, Johannes and Walter Thiel (2005). "Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method: "Umbrella integration"". In: The Journal of Chemical Physics 123.14. DOI: 10.1063/1.2052648.

Kay, Euan R. and David A. Leigh (2015). "Rise of the Molecular Machines". In: Angewandte Chemie International Edition 54.35. DOI: 10.1002/anie.201503375.

Khinchin, A. IA. (1949). Mathematical foundations of statistical mechanics; New York: Dover Publications.

Kiani, F. A. and S. Fischer (2014). "Catalytic strategy used by the myosin motor to hydrolyze ATP". In: Proceedings of the National Academy of Sciences 111.29. DOI: 10.1073/pnas.1401862111.

Kiani, F. A. and S. Fischer (2013). "Stabilization of the ADP/Metaphosphate Intermediate during ATP Hydrolysis in Pre-power Stroke Myosin: QUANTITATIVE ANATOMY OF AN ENZYME". In: Journal of Biological Chemistry 288.49. DOI: 10.1074/jbc.M113.500298.

Kieseritzky, Gernot and Ernst-Walter Knapp (2007). "Optimizing pKA computation in proteins with pH adapted conformations". In: Proteins: Structure, Function, and Bioinformatics 71.3. DOI: 10.1002/prot.21820.

Kintses, Bálint, Zhenhui Yang, and András Málnási-Csizmadia (2008). "Experimental Investigation of the Seesaw Mechanism of the Relay Region That Moves the Myosin Lever Arm". In: Journal of Biological Chemistry 283.49. DOI: 10.1074/jbc.M805848200.

Kirkwood, John G. (1935). "Statistical Mechanics of Fluid Mixtures". In: The Journal of Chemical Physics 3.5. DOI: 10.1063/1.1749657.

Koppole, Sampath, Jeremy C. Smith, and Stefan Fischer (2006). "Simulations of the Myosin II Motor Reveal a Nucleotide-state Sensing Element that Controls the Recovery Stroke". In: Journal of Molecular Biology 361.3. DOI: 10.1016/j.jmb.2006.06.022.

Koppole, Sampath, Jeremy C. Smith, and Stefan Fischer (2007). "The Structural Coupling between ATPase Activation and Recovery Stroke in the Myosin II Motor". In: Structure 15.7. DOI: 10.1016/j.str.2007.06.008.

Kramers, H.A. (1940). "Brownian motion in a field of force and the diffusion model of chemical reactions". In: Physica 7.4. DOI: 10.1016/S0031-8914(40)90098-2.

Krivov, S. V. and Martin Karplus (2008). “Diffusive reaction dynamics on invariant free energy profiles”. In: Proceedings of the National Academy of Sciences 105.37. DOI: 10.1073/pnas.0800228105.

Krivov, Sergei V. and Martin Karplus (2006). “One-Dimensional Free-Energy Profiles of Complex Systems: Progress Variables that Preserve the Barriers”. In: The Journal of Physical Chemistry B 110.25. DOI: 10.1021/jp060039b.

Kumar, Shankar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman (1992). “The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method”. In: Journal of computational chemistry 13.8.

Laio, Alessandro and Michele Parrinello (2002). “Escaping free-energy minima”. In: Proceedings of the National Academy of Sciences 99.20. DOI: 10.1073/pnas.202427399.

Leach, Andrew R. (2001). Molecular modelling: principles and applications. 2nd ed. Harlow, England ; New York: Prentice Hall. 744 pp.

Lelièvre, Tony, Gabriel Stoltz, and Mathias Rousset (2010). Free energy computations: a mathematical perspective. London ; Hackensack, N.J: Imperial College Press. 458 pp.

Lesage, Adrien, Tony Lelièvre, Gabriel Stoltz, and Jérôme Hénin (2017). “Smoothed Biasing Forces Yield Unbiased Free Energies with the Extended-System Adaptive Biasing Force Method”. In: The Journal of Physical Chemistry B 121.15. DOI: 10.1021/acs.jpcb.6b10055.

Lev, Bogdan, Samuel Murail, Frédéric Poitevin, Brett A. Cromer, Marc Baaden, Marc Delarue, and Toby W. Allen (2017). “String method solution of the gating pathways for a pentameric ligand-gated ion channel”. In: Proceedings of the National Academy of Sciences 114.21. DOI: 10.1073/pnas.1617567114.

Li, Guohui and Qiang Cui (2004). “Mechanochemical Coupling in Myosin: A Theoretical Analysis with Molecular Dynamics and Combined QM/MM Reaction Path Calculations”. In: The Journal of Physical Chemistry B 108.10. DOI: 10.1021/jp0371783.

Liu, Peng, Christophe Chipot, Wensheng Cai, and Xueguang Shao (2014). “Unveiling the Underlying Mechanism for Compression and Decompression Strokes of a Molecular Engine”. In: The Journal of Physical Chemistry C 118.23. DOI: 10.1021/jp503241p.

Liu, Peng, Xueguang Shao, and Wensheng Cai (2015). “Deciphering the Mechanism Involved in the Switch On/Off of Molecular Pistons”. In: Chinese Journal of Chemistry 33.10. DOI: 10.1002/cjoc.201500402.

Liu, Peng, Xueguang Shao, Christophe Chipot, and Wensheng Cai (2016). “The true nature of rotary movements in rotaxanes”. In: Chem. Sci. 7.1. DOI: 10.1039/C5SC03022F.

Llinas, Paola, Tatiana Isabet, Lin Song, Virginie Ropars, Bin Zong, Hannah Benisty, Serena Sirigu, Carl Morris, Carlos Kikuti, Dan Safer, H. Lee Sweeney, and Anne Houdusse (2015). “How Actin Initiates the Motor Activity of Myosin”. In: Developmental Cell 33.4. DOI: 10.1016/j.devcel.2015.03.025.

Lu, Xiya, Victor Ovchinnikov, Darren Demapan, Daniel Roston, and Qiang Cui (2017). “Regulation and Plasticity of Catalysis in Enzymes: Insights from Analysis of Mechanochemical Coupling in Myosin”. In: Biochemistry 56.10. DOI: 10.1021/acs.biochem.7b00016.

Lymn, Richard W. and Edwin W. Taylor (1970). “Transient state phosphate production in the hydrolysis of nucleoside triphosphates by myosin”. In: Biochemistry 9.15. DOI: 10.1021/bi00817a007.

Ma, Wen and Klaus Schulten (2015). “Mechanism of Substrate Translocation by a Ring-Shaped ATPase Motor at Millisecond Resolution”. In: Journal of the American Chemical Society 137.8. DOI: 10.1021/ja512605w.

MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and Martin Karplus (1998). "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins". In: The Journal of Physical Chemistry B 102.18. DOI: 10.1021/jp973084f.

Mackerell, Alexander D., Michael Feig, and Charles L. Brooks (2004). "Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations". In: Journal of Computational Chemistry 25.11. DOI: 10.1002/jcc.20065.

Malik, F. I., J. J. Hartman, K. A. Elias, B. P. Morgan, H. Rodriguez, K. Brejc, R. L. Anderson, S. H. Sueoka, K. H. Lee, J. T. Finer, R. Sakowicz, R. Baliga, D. R. Cox, M. Garard, G. Godinez, R. Kawas, E. Kraynack, D. Lenzi, P. P. Lu, A. Muci, C. Niu, X. Qian, D. W. Pierce, M. Pokrovskii, I. Suehiro, S. Sylvester, T. Tochimoto, C. Valdez, W. Wang, T. Katori, D. A. Kass, Y.-T. Shen, S. F. Vatner, and D. J. Morgans (2011). "Cardiac Myosin Activation: A Potential Therapeutic Approach for Systolic Heart Failure". In: Science 331.6023. DOI: 10.1126/science.1200113.

Málnási-Csizmadia, András, David S. Pearson, Mihály Kovács, Robert J. Woolley, Michael A. Geeves, and Clive R. Bagshaw (2001). "Kinetic Resolution of a Conformational Transition and the ATP Hydrolysis Step Using Relaxation Methods with a *Dictyostelium* Myosin II Mutant Containing a Single Tryptophan Residue †". In: Biochemistry 40.42. DOI: 10.1021/bi010963q.

Maragliano, Luca, Alexander Fischer, Eric Vanden-Eijnden, and Giovanni Ciccotti (2006). "String method in collective variables: Minimum free energy paths and isocommittor surfaces". In: The Journal of Chemical Physics 125.2. DOI: 10.1063/1.2212942.

Maragliano, Luca, Benoît Roux, and Eric Vanden-Eijnden (2014). "Comparison between Mean Forces and Swarms-of-Trajectories String Methods". In: Journal of Chemical Theory and Computation 10.2. DOI: 10.1021/ct400606c.

Maragliano, Luca and Eric Vanden-Eijnden (2006). "A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations". In: Chemical Physics Letters 426.1. DOI: 10.1016/j.cplett.2006.05.062.

Maragliano, Luca and Eric Vanden-Eijnden (2007). "On-the-fly string method for minimum free energy paths calculation". In: Chemical Physics Letters 446.1. DOI: 10.1016/j.cplett.2007.08.017.

Maragliano, Luca, Eric Vanden-Eijnden, and Benoît Roux (2009). "Free Energy and Kinetics of Conformational Transitions from Voronoi Tessellated Milestoning with Restraining Potentials". In: Journal of Chemical Theory and Computation 5.10. DOI: 10.1021/ct900279z.

Mattila, Pieta K. and Pekka Lappalainen (2008). "Filopodia: molecular architecture and cellular functions". In: Nature Reviews Molecular Cell Biology 9.6. DOI: 10.1038/nrm2406.

McCammon, J. Andrew, Bruce R. Gelin, and Martin Karplus (1977). "Dynamics of folded proteins". In: Nature 267.5612. DOI: 10.1038/267585a0.

McCoy, A. J., R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, and R. J. Read (2007). "Phaser crystallographic software". In: Journal of Applied Crystallography 40.4. DOI: 10.1107/S0021889807021206.

McKinney, Wes (2010). "Data Structures for Statistical Computing in Python". In:

Ménétrey, Julie, Amel Bahloul, Amber L. Wells, Christopher M. Yengo, Carl A. Morris, H. Lee Sweeney, and Anne Houdusse (2005). "The structure of the myosin VI motor reveals the mechanism of directionality reversal". In: Nature 435.7043. DOI: 10.1038/nature03592.

Ménétrey, Julie, Tatiana Isabet, Virginie Ropars, Monalisa Mukherjea, Olena Pylypenko, Xiaoyan Liu, Javier Perez, Patrice Vachette, H. Lee Sweeney, and Anne M. Houdusse (2012). "Processive Steps in the Reverse Direction Require Uncoupling of the Lead Head Lever Arm of Myosin VI". In: Molecular Cell 48.1. DOI: 10.1016/j.molcel.2012.07.034.

Ménétrey, Julie, Paola Llinas, Jérome Cicolari, Gaëlle Squires, Xiaoyan Liu, Anna Li, H. Lee Sweeney, and Anne Houdusse (2008). "The post-rigor structure of myosin VI and implications for the recovery stroke". In: The EMBO Journal 27.1. DOI: 10.1038/sj.emboj.7601937.

Ménétrey, Julie, Paola Llinas, Monalisa Mukherjea, H. Lee Sweeney, and Anne Houdusse (2007). "The Structural Basis for the Large Powerstroke of Myosin VI". In: Cell 131.2. DOI: 10.1016/j.cell.2007.08.027.

Mesentean, Sidonia, Sampath Koppole, Jeremy C. Smith, and Stefan Fischer (2007). "The Principal Motions Involved in the Coupling Mechanism of the Recovery Stroke of the Myosin Motor". In: Journal of Molecular Biology 367.2. DOI: 10.1016/j.jmb.2006.12.058.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller (1953). "Equation of State Calculations by Fast Computing Machines". In: The Journal of Chemical Physics 21.6. DOI: 10.1063/1.1699114.

Miao, Yinglong, Victoria A. Feher, and J. Andrew McCammon (2015). "Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation". In: Journal of Chemical Theory and Computation 11.8. DOI: 10.1021/acs.jctc.5b00436.

Miao, Yinglong, William Sinko, Levi Pierce, Denis Bucher, Ross C. Walker, and J. Andrew McCammon (2014). "Improved Reweighting of Accelerated Molecular Dynamics Simulations for Free Energy Calculation". In: Journal of Chemical Theory and Computation 10.7. DOI: 10.1021/ct500090q.

Miller, T. F., Eric Vanden-Eijnden, and D. Chandler (2007). "Solvent coarse-graining and the string method applied to the hydrophobic collapse of a hydrated chain". In: Proceedings of the National Academy of Sciences 104.37. DOI: 10.1073/pnas.0705830104.

Mones, Letif, Noam Bernstein, and Gábor Csányi (2016). "Exploration, Sampling, And Reconstruction of Free Energy Surfaces with Gaussian Process Regression". In: Journal of Chemical Theory and Computation 12.10. DOI: 10.1021/acs.jctc.6b00553.

Montalvo-Acosta, Joel José and Marco Cecchini (2016). "Computational Approaches to the Chemical Equilibrium Constant in Protein-ligand Binding". In: Molecular Informatics 35.11. DOI: 10.1002/minf.201600052.

Moore, Calvin C. (2015). "Ergodic theorem, ergodic theory, and statistical mechanics". In: Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.1421798112.

Moradi, M. and E. Tajkhorshid (2013). "Mechanistic picture for conformational transition of a membrane transporter at atomic resolution". In: Proceedings of the National Academy of Sciences 110.47. DOI: 10.1073/pnas.1313202110.

Morgan, Bradley P., Alexander Muci, Pu-Ping Lu, Xiangping Qian, Todd Tochimoto, Whitney W. Smith, Marc Garard, Erica Kraynack, Scott Collibee, Ion Suehiro, Adam Tomasi, S. Corey Valdez, Wenyue Wang, Hong Jiang, James Hartman, Hector M. Rodriguez, Raja Kawas, Sheila Sylvester, Kathleen A. Elias, Guillermo Godinez, Kenneth Lee, Robert Anderson, Sandra Sueoka, Donghong Xu, Zhengping Wang, Nebojsa Djordjevic, Fady I. Malik, and David J. Morgans (2010). "Discovery of Omecamtiv Mecarbil the First, Selective, Small Molecule Activator of Cardiac Myosin". In: ACS Medicinal Chemistry Letters 1.9. DOI: 10.1021/ml100138q.

Morrow, Timothy I. and Edward J. Maginn (2002). "Molecular Dynamics Study of the Ionic Liquid 1- *n* -Butyl-3-methylimidazolium Hexafluorophosphate". In: The Journal of Physical Chemistry B 106.49. DOI: 10.1021/jp0267003.

Mugnai, Mauro L. and D. Thirumalai (2017). "Kinematics of the lever arm swing in myosin VI". In: Proceedings of the National Academy of Sciences 114.22. DOI: 10.1073/pnas.1615708114.

Mukherjee, S. and A. Warshel (2013). "Electrostatic origin of the unidirectionality of walking myosin V motors". In: Proceedings of the National Academy of Sciences 110.43. DOI: 10.1073/pnas.1317641110.

Müller, Klaus and Leo D. Brown (1979). "Location of saddle points and minimum energy paths by a constrained simplex optimization procedure". In: Theoretica Chimica Acta 53.1. DOI: 10.1007/BF00547608.

Murphy, Coleen T., Ronald S. Rock, and James A. Spudich (2001). "A myosin II mutation uncouples ATPase activity from motility and shortens step size". In: Nature Cell Biology 3.3.

Nagy, S., B. L. Ricca, M. F. Norstrom, D. S. Courson, C. M. Brawley, P. A. Smithback, and R. S. Rock (2008). "A myosin motor that selects bundled actin for motility". In: Proceedings of the National Academy of Sciences 105.28. DOI: 10.1073/pnas.0802592105.

Nesmelov, Y. E., R. V. Agafonov, I. V. Negrashov, S. E. Blakely, M. A. Titus, and D. D. Thomas (2011). "Structural kinetics of myosin by transient time-resolved FRET". In: Proceedings of the National Academy of Sciences 108.5. DOI: 10.1073/pnas.1012320108.

Nitao, Lisa K. and Emil Reisler (1998). "Probing the Conformational States of the SH1−SH2 Helix in Myosin: A Cross-Linking Approach †". In: Biochemistry 37.47. DOI: 10.1021/bi9817212.

Noji, Hiroyuki, Ryohei Yasuda, Masasuke Yoshida, and Kazuhiko Kinosita (1997). "Direct observation of the rotation of F1-ATPase". In: Nature 386.6622. DOI: 10.1038/386299a0.

Nosé, Shūichi (1984). "A molecular dynamics method for simulations in the canonical ensemble". In: Molecular Physics 52.2. DOI: 10.1080/00268978400101201.

Odronitz, Florian and Martin Kollmar (2007). "Drawing the tree of eukaryotic life based on the analysis of 2,269 manually annotated myosins from 328 species". In: Genome biology 8.9.

Ölender, Roberto and Ron Elber (1997). "Yet another look at the steepest descent path". In: Journal of Molecular Structure (Theochem) 398.399.

Onishi, Hirofumi, Shin-ichiro Kojima, Kazuo Katoh, Keigi Fujiwara, Hugo M. Martinez, and Manuel F. Morales (1998). "Functional transitions in myosin: Formation of a critical salt-bridge and transmission of effect to the sensitive tryptophan". In: Proceedings of the National Academy of Sciences 95.12.

Onishi, Hirofumi, Takashi Ohki, Naoki Mochizuki, and Manuel F. Morales (2002). "Early stages of energy transduction by myosin: Roles of Arg in Switch I, of Glu in Switch II, and of the salt-bridge between them". In: Proceedings of the National Academy of Sciences 99.24. DOI: 10.1073/pnas.242604099.

Oostenbrink, Chris, Alessandra Villa, Alan E. Mark, and Wilfred F. Van Gunsteren (2004). "A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6". In: Journal of Computational Chemistry 25.13. DOI: 10.1002/jcc.20090.

Otter, W. K. den (2000). "Thermodynamic integration of the free energy along a reaction coordinate in Cartesian coordinates". In: The Journal of Chemical Physics 112.17. DOI: 10.1063/1.481329.

Ovchinnikov, Victor, Marco Cecchini, and Martin Karplus (2013). "A Simplified Confinement Method for Calculating Absolute Free Energies and Free Energy and Entropy Differences". In: The Journal of Physical Chemistry B 117.3. DOI: 10.1021/jp3080578.

Ovchinnikov, Victor, Marco Cecchini, Eric Vanden-Eijnden, and Martin Karplus (2011). “A Conformational Transition in the Myosin VI Converter Contributes to the Variable Step Size”. In: Biophysical Journal 101.10. DOI: 10.1016/j.bpj.2011.09.044.

Ovchinnikov, Victor, Martin Karplus, and Eric Vanden-Eijnden (2011). “Free energy of conformational transition paths in biomolecules: The string method and its application to myosin VI”. In: The Journal of Chemical Physics 134.8. DOI: 10.1063/1.3544209.

Ovchinnikov, Victor, Kwangho Nam, and Martin Karplus (2016). “A Simple and Accurate Method To Calculate Free Energy Profiles and Reaction Rates from Restrained Molecular Simulations of Diffusive Processes”. In: The Journal of Physical Chemistry B. DOI: 10.1021/acs.jpcb.6b02139.

Ovchinnikov, Victor, Bernhardt L. Trout, and Martin Karplus (2010). “Mechanical Coupling in Myosin V: A Simulation Study”. In: Journal of Molecular Biology 395.4. DOI: 10.1016/j.jmb.2009.10.029.

Paci, Emanuele and Martin Karplus (1999). “Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations”. In: Journal of Molecular Biology 288.3. DOI: 10.1006/jmbi.1999.2670.

Palermo, Giulia, Yinglong Miao, Ross C. Walker, Martin Jinek, and J. Andrew McCammon (2017). “CRISPR-Cas9 conformational activation as elucidated from enhanced molecular simulations”. In: Proceedings of the National Academy of Sciences 114.28. DOI: 10.1073/pnas.1707645114.

Pan, Albert C., Deniz Sezer, and Benoît Roux (2008). “Finding Transition Pathways Using the String Method with Swarms of Trajectories”. In: The Journal of Physical Chemistry B 112.11. DOI: 10.1021/jp0777059.

Pande, Vijay S., Kyle Beauchamp, and Gregory R. Bowman (2010). “Everything you wanted to know about Markov State Models but were afraid to ask”. In: Methods 52.1. DOI: 10.1016/j.ymeth.2010.06.002.

Pang, Yui Tik, Yinglong Miao, Yi Wang, and J. Andrew McCammon (2017). “Gaussian Accelerated Molecular Dynamics in NAMD”. In: Journal of Chemical Theory and Computation 13.1. DOI: 10.1021/acs.jctc.6b00931.

Park, H., A. Li, L.-Q. Chen, Anne Houdusse, P. R. Selvin, and H. Lee Sweeney (2007). “The unique insert at the end of the myosin VI motor is the sole determinant of directionality”. In: Proceedings of the National Academy of Sciences 104.3. DOI: 10.1073/pnas.0610066104.

Park, Soohyung, Taehoon Kim, and Wonpil Im (2012). “Transmembrane Helix Assembly by Window Exchange Umbrella Sampling”. In: Physical Review Letters 108.10. DOI: 10.1103/PhysRevLett.108.108102.

Patterson, Bruce, Kathleen M. Ruppel, Yuan Wu, and James A. Spudich (1997). “Cold-sensitive Mutants G680V and G691C ofDictyostelium Myosin II Confer Dramatically Different Biochemical Defects”. In: Journal of Biological Chemistry 272.44.

Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau (n.d.). “Scikit-learn: Machine Learning in Python”. In: MACHINE LEARNING IN PYTHON.

Perez, Fernando and Brian E. Granger (2007). “IPython: A System for Interactive Scientific Computing”. In: Computing in Science & Engineering 9.3. DOI: 10.1109/MCSE.2007.53.

Perrin, Jean (2014). Les atomes. OCLC: 899155978. Paris: Flammarion.

Perrin, Jean (1909). “Mouvement brownien et réalité moléculaire”. In: Annales de Chimie et de Physique 18.

Phillips, James C., Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten (2005). "Scalable molecular dynamics with NAMD". In: Journal of Computational Chemistry 26.16. DOI: 10.1002/jcc.20289.

Piana, Stefano and Alessandro Laio (2007). "A Bias-Exchange Approach to Protein Folding". In: The Journal of Physical Chemistry B 111.17. DOI: 10.1021/jp0678731.

Pierce, Levi C.T., Romelia Salomon-Ferrer, Cesar Augusto F. de Oliveira, J. Andrew McCammon, and Ross C. Walker (2012). "Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics". In: Journal of Chemical Theory and Computation 8.9. DOI: 10.1021/ct300284c.

Planelles-Herrero, Vicente José, Florian Blanc, Serena Sirigu, Helena Sirkia, Jeffrey Clause, Yannick Sourigues, Daniel O. Johnsrud, Beatrice Amigues, Marco Cecchini, Susan P. Gilbert, Anne Houdusse, and Margaret A. Titus (2016). "Myosin MyTH4-FERM structures highlight important principles of convergent evolution". In: Proceedings of the National Academy of Sciences 113.21. DOI: 10.1073/pnas.1600736113.

Planelles-Herrero, Vicente José, James J. Hartman, Julien Robert-Paganin, Fady I. Malik, and Anne Houdusse (2017). "Mechanistic and structural basis for activation of cardiac myosin force production by omecamtiv mecarbil". In: Nature Communications 8.1. DOI: 10.1038/s41467-017-00176-5.

Preller, Matthias and Kenneth C. Holmes (2013). "The myosin start-of-power stroke state and how actin binding drives the power stroke: The Myosin Start-of-Power Stroke State". In: Cytoskeleton 70.10. DOI: 10.1002/cm.21125.

Press, William H., ed. (2007). Numerical recipes: the art of scientific computing. 3rd ed. OCLC: ocn123285342. Cambridge, UK ; New York: Cambridge University Press. 1235 pp. ISBN: 978-0-521-88068-8 978-0-521-88407-5 978-0-521-70685-8.

Purcell, T. J., C. Morris, J. A. Spudich, and H. Lee Sweeney (2002). "Role of the lever arm in the processive stepping of myosin V". In: Proceedings of the National Academy of Sciences 99.22. DOI: 10.1073/pnas.182539599.

Pylypenko, Olena, Lin Song, Ai Shima, Zhaohui Yang, Anne M. Houdusse, and H. Lee Sweeney (2015). "Myosin VI deafness mutation prevents the initiation of processive runs on actin". In: Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.1420989112.

Rabenstein, Björn and Ernst-Walter Knapp (2001). "Calculated pH-Dependent Population and Protonation of Carbon-Monoxy-Myoglobin Conformers". In: Biophysical Journal 80.3. DOI: 10.1016/S0006-3495(01)76091-2.

Raiteri, Paolo, Giovanni Bussi, Clotilde S. Cucinotta, Alberto Credi, J. Fraser Stoddart, and Michele Parrinello (2008). "Unravelling the Shuttling Mechanism in a Photoswitchable Multicomponent Bistable Rotaxane". In: Angewandte Chemie International Edition 47.19. DOI: 10.1002/anie.200705207.

Rayment, I, H. Holden, M Whittaker, C. Yohn, M Lorenz, Kenneth C. Holmes, and R. Milligan (1993). "Structure of the actin-myosin complex and its implications for muscle contraction". In: Science 261.5117. DOI: 10.1126/science.8316858.

Rayment, I, W. Rypniewski, K Schmidt-Base, R Smith, D. Tomchick, M. Benning, D. Winkelmann, G Wesenberg, and H. Holden (1993). "Three-dimensional structure of myosin subfragment-1: a molecular motor". In: Science 261.5117. DOI: 10.1126/science.8316857.

Rayment, I, C Smith, and R G Yount (1996). "The Active Site of Myosin". In: Annual Review of Physiology 58.1. DOI: 10.1146/annurev.ph.58.030196.003323.

Reedy, M. K., Kenneth C. Holmes, and R. T. Tregear (1965). “Induced Changes in Orientation of the Cross-Bridges of Glycerinated Insect Flight Muscle”. In: Nature 207.5003. DOI: 10.1038/2071276a0.

Richards, Thomas A. and Thomas Cavalier-Smith (2005). “Myosin domain evolution and the primary divergence of eukaryotes”. In: Nature 436.7054. DOI: 10.1038/nature03949.

Rief, M. (1997). “Reversible Unfolding of Individual Titin Immunoglobulin Domains by AFM”. In: Science 276.5315. DOI: 10.1126/science.276.5315.1109.

Rohde, John A., David D. Thomas, and Joseph M. Muretta (2017). “Heart failure drug changes the mechanoenzymology of the cardiac myosin powerstroke”. In: Proceedings of the National Academy of Sciences 114.10. DOI: 10.1073/pnas.1611698114.

Ropars, Virginie, Zhaohui Yang, Tatiana Isabet, Florian Blanc, Kaifeng Zhou, Tianming Lin, Xiaoyan Liu, Pascale Hissier, Frédéric Samazan, Béatrice Amigues, Eric D. Yang, Hyokeun Park, Olena Pylypenko, Marco Cecchini, Charles V. Sindelar, H. Lee Sweeney, and Anne Houdusse (2016). “The myosin X motor is optimized for movement on actin bundles”. In: Nature Communications 7. DOI: 10.1038/ncomms12456.

Ross, Jennifer L, M Yusuf Ali, and David M Warshaw (2008). “Cargo transport: molecular motors navigate a complex cytoskeleton”. In: Current Opinion in Cell Biology 20.1. DOI: 10.1016/j.ceb.2007.11.006.

Ryckaert, Jean-Paul, Giovanni Ciccotti, and Herman J.C Berendsen (1977). “Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes”. In: Journal of Computational Physics 23.3. DOI: 10.1016/0021-9991(77)90098-5.

Sasaki, N., T. Shimada, and K. Sutoh (1998). “Mutational Analysis of the Switch II Loop ofDictyostelium Myosin II”. In: Journal of Biological Chemistry 273.32. DOI: 10.1074/jbc.273.32.20334.

Sauvage, Jean-Pierre (2016). “From Chemical Topology to Molecular Machines”. In:

Sauvage, Jean-Pierre (1998). “Transition Metal-Containing Rotaxanes and Catenanes in Motion: Toward Molecular Machines and Motors”. In: Accounts of Chemical Research 31.10. DOI: 10.1021/ar960263r.

Scarabelli, Guido and Barry J. Grant (2013). “Mapping the Structural and Dynamical Features of Kinesin Motor Domains”. In: PLoS Computational Biology 9.11. Ed. by Emad Tajkhorshid. DOI: 10.1371/journal.pcbi.1003329.

Schlick, Tamar, ed. (2012). Innovations in biomolecular modeling and simulations. Vol. 1: ... RSC biomolecular sciences 23. OCLC: 930833891. Cambridge: Royal Soc. of Chemistry. 355 pp. ISBN: 978-1-84973-461-5.

Schlitter, J., M. Engels, and P. Krüger (1994). “Targeted molecular dynamics: A new approach for searching pathways of conformational transitions”. In: Journal of Molecular Graphics 12.2. DOI: 10.1016/0263-7855(94)80072-3.

Schlitter, J., M. Engels, P. Krüger, E. Jacoby, and A. Wollmer (1993). “Targeted Molecular Dynamics Simulation of Conformational Change-Application to the T ↔ R Transition in Insulin”. In: Molecular Simulation 10.2. DOI: 10.1080/08927029308022170.

Schliwa, Manfred and Günther Woehlke (2003). “Molecular motors”. In: Nature 422.6933. DOI: 10.1038/nature01601.

Schröder, Rasmus R., Dietmar J. Manstein, Werner Jahn, Hazel Holden, Ivan Rayment, Kenneth C. Holmes, and James A. Spudich (1993). “Three-dimensional atomic model of F-actin decorated with Dictyostelium myosin S1”. In: Nature 364.6433. DOI: 10.1038/364171a0.

Seeber, M., M. Cecchini, F. Rao, G. Settanni, and A. Caflisch (2007). "Wordom: a program for efficient analysis of molecular dynamics simulations". In: Bioinformatics 23.19. DOI: 10.1093/bioinformatics/btm378.

Seeber, Michele, Angelo Felline, Francesco Raimondi, Stefanie Muff, Ran Friedman, Francesco Rao, Amedeo Caflisch, and Francesca Fanelli (2011). "Wordom: A user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces". In: Journal of Computational Chemistry 32.6. DOI: 10.1002/jcc.21688.

Shih, William M., Zygmunt Gryczynski, Joseph R. Lakowicz, and James A. Spudich (2000). "A FRET-based sensor reveals large ATP hydrolysis–induced conformational changes and three distinct states of the molecular motor myosin". In: Cell 102.5.

Shiroguchi, Katsuyuki, Harvey F. Chin, Diane E. Hannemann, Eiro Muneyuki, Enrique M. De La Cruz, and Kazuhiko Kinosita (2011). "Direct Observation of the Myosin Va Recovery Stroke That Contributes to Unidirectional Stepping along Actin". In: PLoS Biology 9.4. Ed. by James Spudich. DOI: 10.1371/journal.pbio.1001031.

Shirts, Michael R. and John D. Chodera (2008). "Statistically optimal analysis of samples from multiple equilibrium states". In: The Journal of Chemical Physics 129.12. DOI: 10.1063/1.2978177.

Singharoy, Abhishek and Christophe Chipot (2016). "Methodology for the Simulation of Molecular Motors at Different Scales". In: The Journal of Physical Chemistry B. DOI: 10.1021/acs.jpcb.6b09350.

Singharoy, Abhishek, Christophe Chipot, Mahmoud Moradi, and Klaus Schulten (2017). "Chemomechanical Coupling in Hexameric Protein–Protein Interfaces Harnesses Energy within V-Type ATPases". In: Journal of the American Chemical Society 139.1. DOI: 10.1021/jacs.6b10744.

Sirigu, Serena, James J. Hartman, Vicente José Planelles-Herrero, Virginie Ropars, Sheila Clancy, Xi Wang, Grace Chuang, Xiangping Qian, Pu-Ping Lu, Edward Barrett, Karin Rudolph, Christopher Royer, Bradley P. Morgan, Enrico A. Stura, Fady I. Malik, and Anne M. Houdusse (2016). "Highly selective inhibition of myosin motors provides the basis of potential therapeutic application". In: Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.1609342113.

Smith, Clyde A. and Ivan Rayment (1996). "X-ray Structure of the Magnesium(II).ADP.Vanadate Complex of the Dictyostelium discoideum Myosin Motor domain to 1.9 Å Resolution". In: Biochemistry 35.

Socci, N. D., J. N. Onuchic, and P. G. Wolynes (1996). "Diffusive dynamics of the reaction coordinate for protein folding funnels". In: The Journal of Chemical Physics 104.15. DOI: 10.1063/1.471317.

Spudich, James A. and Sivaraj Sivaramakrishnan (2010). "Myosin VI: an innovative motor that challenged the swinging lever arm hypothesis". In: Nature Reviews Molecular Cell Biology 11.2. DOI: 10.1038/nrm2833.

Sternberg, Shlomo (2010). Dynamical systems. Mineola, N.Y: Dover Publications. 265 pp.

Stoddart, J Fraser (2016). "Mechanically Interlocked Molecules (MIMs)—Molecular Shuttles, Switches, and Machines". In:

Sugita, Yuji, Akio Kitao, and Yuko Okamoto (2000). "Multidimensional replica-exchange method for free-energy calculations". In: The Journal of Chemical Physics 113.15. DOI: 10.1063/1.1308516.

Sugita, Yuji and Yuko Okamoto (1999). "Replica-exchange molecular dynamics method for protein folding". In: Chemical Physics Letters 314.1. DOI: 10.1016/S0009-2614(99)01123-9.

Sweeney, H. Lee and Anne Houdusse (2010a). "Myosin VI Rewrites the Rules for Myosin Motors". In: Cell 141.4. DOI: 10.1016/j.cell.2010.04.028.

Sweeney, H. Lee and Anne Houdusse (2010b). “Structural and Functional Insights into the Myosin Motor Mechanism”. In: Annual Review of Biophysics 39.1. DOI: 10.1146/annurev.biophys.050708.133751.

Sweeney, H. Lee and Anne Houdusse (2004). “The motor mechanism of myosin V: insights for muscle contraction”. In: Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 359.1452. DOI: 10.1098/rstb.2004.1576.

Sweeney, H. Lee and Anne Houdusse (2007). “What can myosin VI do in cells?” In: Current Opinion in Cell Biology 19.1. DOI: 10.1016/j.ceb.2006.12.005.

Szabo, Attila and Neil S. Ostlund (1996). Modern quantum chemistry: introduction to advanced electronic structure theory. Mineola, N.Y: Dover Publications.

Szent-Györgyi, Andrew G. (2004). “The Early History of the Biochemistry of Muscle Contraction”. In: The Journal of General Physiology 123.6. DOI: 10.1085/jgp.200409091.

Takemoto, Mizuki, Yongchan Lee, Ryuichiro Ishitani, and Osamu Nureki (2018). “Free Energy Landscape for the Entire Transport Cycle of Triose-Phosphate/Phosphate Translocator”. In: Structure 0.0. DOI: 10.1016/j.str.2018.05.012.

Taylor, Edwin William, Richard W. Lymn, and George Moll (1970). “Myosin-product complex and its effect on the steady-state rate of nucleoside triphosphate hydrolysis”. In: Biochemistry 9.15. DOI: 10.1021/bi00817a008.

Tiwary, Pratyush and Michele Parrinello (2015). “A Time-Independent Free Energy Estimator for Metadynamics”. In: The Journal of Physical Chemistry B 119.3. DOI: 10.1021/jp504920s.

Tiwary, Pratyush and Michele Parrinello (2013). “From Metadynamics to Dynamics”. In: Physical Review Letters 111.23. DOI: 10.1103/PhysRevLett.111.230602.

Torrie, G.M. and J.P. Valleau (1977). “Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling”. In: Journal of Computational Physics 23.2. DOI: 10.1016/0021-9991(77)90121-8.

Trivedi, Darshan V., Joseph M. Muretta, Anja M. Swenson, Jonathon P. Davis, David D. Thomas, and Christopher M. Yengo (2015). “Direct measurements of the coordination of lever arm swing and the catalytic cycle in myosin V”. In: Proceedings of the National Academy of Sciences 112.47. DOI: 10.1073/pnas.1517566112.

Tsiavaliaris, Georgios, Setsuko Fujita-Becker, Renu Batra, Dmitrii I. Levitsky, F. Jon Kull, Michael A. Geeves, and Dietmar J. Manstein (2002). “Mutations in the relay loop region result in dominant-negative inhibition of myosin II function in Dictyostelium”. In: EMBO reports 3.11.

Tuckerman, Mark E. (2010). Statistical mechanics: theory and molecular simulation. Oxford ; New York: Oxford University Press. 696 pp.

Tuckerman, Mark E., B. J. Berne, and G. J. Martyna (1992). “Reversible multiple time scale molecular dynamics”. In: The Journal of Chemical Physics 97.3. DOI: 10.1063/1.463137.

Tyka, Michael D., Anthony R. Clarke, and Richard B. Sessions (2006). “An Efficient, Path-Independent Method for Free-Energy Calculations”. In: The Journal of Physical Chemistry B 110.34. DOI: 10.1021/jp060734j.

Uversky, V. N. (2002). “Natively unfolded proteins: A point where biology waits for physics”. In: Protein Science 11.4. DOI: 10.1110/ps.4210102.

Vaart, Arjan van der and Martin Karplus (2005). “Simulation of conformational transitions by the restricted perturbation–targeted molecular dynamics method”. In: The Journal of Chemical Physics 122.11. DOI: 10.1063/1.1861885.

Vale, R. D. (1996). “Switches, latches, and amplifiers: common themes of G proteins and molecular motors”. In: The Journal of Cell Biology 135.2. DOI: 10.1083/jcb.135.2.291.

Vale, Ronald D (2003). “The Molecular Motor Toolbox for Intracellular Transport”. In: Cell 112.4. DOI: 10.1016/S0092-8674(03)00111-9.

Vale, Ronald D. and Ronald A. Milligan (2000). “The way things move: looking under the hood of molecular motor proteins”. In: Science 288.5463.

Valsson, Omar, Pratyush Tiwary, and Michele Parrinello (2016). “Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint”. In: Annual Review of Physical Chemistry 67.1. DOI: 10.1146/annurev-physchem-040215-112229.

Van Eerden, J., W.J. Briels, S. Harkema, and D. Feil (1989). “Potential of mean force by thermodynamic integration: Molecular-dynamics simulation of decomplexation”. In: Chemical Physics Letters 164.4. DOI: 10.1016/0009-2614(89)85222-4.

Vanden-Eijnden, Eric and Maddalena Venturoli (2009a). “Markovian milestoning with Voronoi tessellations”. In: The Journal of Chemical Physics 130.19. DOI: 10.1063/1.3129843.

Vanden-Eijnden, Eric and Maddalena Venturoli (2009b). “Revisiting the finite temperature string method for the calculation of reaction tubes and free energies”. In: The Journal of Chemical Physics 130.19. DOI: 10.1063/1.3130083.

Vanden-Eijnden, Eric, Maddalena Venturoli, Giovanni Ciccotti, and Ron Elber (2008). “On the assumptions underlying milestoning”. In: The Journal of Chemical Physics 129.17. DOI: 10.1063/1.2996509.

Vanommeslaeghe, K., E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell (2010). “CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields”. In: Journal of Computational Chemistry 31.4. DOI: 10.1002/jcc.21367.

Verlet, Loup (1967). “Computer ”Experiments” on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules”. In: Physical Review 159.1. DOI: 10.1103/PhysRev.159.98.

Voter, Arthur F. (1997). “Hyperdynamics: Accelerated Molecular Dynamics of Infrequent Events”. In: Physical Review Letters 78.20. DOI: 10.1103/PhysRevLett.78.3908.

Walker, J. E., M. Saraste, M. J. Runswick, and N. J. Gay (1982). “Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold”. In: The EMBO journal 1.8.

Walt, Stéfan van der, S Chris Colbert, and Gaël Varoquaux (2011). “The NumPy Array: A Structure for Efficient Numerical Computation”. In: Computing in Science & Engineering 13.2. DOI: 10.1109/MCSE.2011.37.

Wang, Y, C B Harrison, K Schulten, and J A McCammon (2011). “Implementation of accelerated molecular dynamics in NAMD”. In: Computational Science & Discovery 4.1. DOI: 10.1088/1749-4699/4/1/015002.

Webb, Benjamin and Andrej Sali (2002). “Comparative Protein Structure Modeling Using MODELLER”. In: Current Protocols in Bioinformatics. John Wiley & Sons, Inc. ISBN: 978-0-471-25095-1.

Weber, Kari L., Anna M. Sokac, Jonathan S. Berg, Richard E. Cheney, and William M. Bement (2004). “A microtubule-binding myosin required for nuclear anchoring and spindle assembly”. In: Nature 431.7006.

Wells, Amber L., Abel W. Lin, Li-Qiong Chen, Daniel Safer, Shane M. Cain, Tama Hasson, Bridget O. Carragher, Ronald A. Milligan, and H. Lee Sweeney (1999). “Myosin VI is an actin-based motor that moves backwards”. In: Nature 401.6752. DOI: 10.1038/46835.

Wereszczynski, Jeff and J. Andrew McCammon (2012). “Nucleotide-dependent mechanism of Get3 as elucidated from free energy calculations”. In: Proceedings of the National Academy of Sciences 109.20.

West, Anthony M. A., Ron Elber, and David Shalloway (2007). “Extending molecular dynamics time scales with milestoning: Example of complex kinetics in a solvated peptide”. In: The Journal of Chemical Physics 126.14. DOI: 10.1063/1.2716389.

Wilson, Miriam R., Jordi Solà, Armando Carlone, Stephen M. Goldup, Nathalie Lebrasseur, and David A. Leigh (2016). “An autonomous chemically fuelled small-molecule motor”. In: Nature 534.7606. DOI: 10.1038/nature18013.

Winkelmann, Donald A., Eva Forgacs, Matthew T. Miller, and Ann M. Stock (2015). “Structural basis for drug-induced allosteric changes to human β-cardiac myosin motor activity”. In: Nature Communications 6. DOI: 10.1038/ncomms8974.

Wolny, M., M. Batchelor, P. J. Knight, E. Paci, L. Dougan, and M. Peckham (2014). “Stable single alpha helices are constant-force springs in proteins”. In: Journal of Biological Chemistry. DOI: 10.1074/jbc.M114.585679.

Woo, Hyung-June (2007). “Exploration of the conformational space of myosin recovery stroke via molecular dynamics”. In: Biophysical Chemistry 125.1. DOI: 10.1016/j.bpc.2006.07.001.

Woolf, Thomas B. and Benoit Roux (1994). “Conformational flexibility of o-phosphorylcholine and o-phosphorylethanolamine: a molecular dynamics study of solvation effects”. In: Journal of the American Chemical Society 116.13.

Wu, Jishan, Ken Cham-Fai Leung, Diego Benítez, Ja-Young Han, Stuart J. Cantrill, Lei Fang, and J. Fraser Stoddart (2008). “An Acid-Base-Controllable [c2]Daisy Chain”. In: Angewandte Chemie International Edition 47.39. DOI: 10.1002/anie.200803036.

Yonezawa, Satoshi, Atsushi Kimura, Seizo Koshiba, Shigeo Masaki, Takao Ono, Atsuko Hanai, Shinichi Sonta, Takashi Kageyama, Takayuki Takahashi, and Akihiko Moriyama (2000). “Mouse Myosin X: Molecular Architecture and Tissue Expression as Revealed by Northern Blot and in Situ Hybridization Analyses”. In: Biochemical and Biophysical Research Communications 271.2. DOI: 10.1006/bbrc.2000.2669.

Yon-Kahn, Jeannine (2006). Histoire de la science des protéines. Les Ulis, France: EDP Sciences.

Yoshida, Hiroyuki, Wenjun Cheng, Jamie Hung, Denise Montell, Erika Geisbrecht, Daniel Rosen, Jinsong Liu, and Honami Naora (2004). “Lessons from border cell migration in the Drosophila ovary: A role for myosin VI in dissemination of human ovarian cancer”. In: Proceedings of the National Academy of Sciences of the United States of America 101.21. DOI: 10.1073/pnas.0400400101.

Yount, Ralph G., J. David Lawson, and Ivan Rayment (1995). “Is myosin a ”back door” enzyme?” In: Biophysical Journal 68.

Yu, Haibo, Liang Ma, Yang Yang, and Qiang Cui (2007a). “Mechanochemical Coupling in the Myosin Motor Domain. I. Insights from Equilibrium Active-Site Simulations”. In: PLoS Computational Biology 3.2. DOI: 10.1371/journal.pcbi.0030021.

Yu, Haibo, Liang Ma, Yang Yang, and Qiang Cui (2007b). “Mechanochemical Coupling in the Myosin Motor Domain. II. Analysis of Critical Residues”. In: PLoS Computational Biology 3.2. DOI: 10.1371/journal.pcbi.0030023.

Zhang, Hongquan, Jonathan S. Berg, Zhilun Li, Yunling Wang, Pernilla Lång, Aurea D. Sousa, Aparna Bhaskar, Richard E. Cheney, and Staffan Strömblad (2004). “Myosin-X provides a motor-based link between integrins and the cytoskeleton”. In: Nature Cell Biology 6.6. DOI: 10.1038/ncb1136.

Zhao, Yan-Ling, Rui-Qin Zhang, Christian Minot, Klaus Hermann, and Michel A. Van Hove (2015). "Revealing highly unbalanced energy barriers in the extension and contraction of the muscle-like motion of a [c2]daisy chain". In: Phys. Chem. Chem. Phys. 17.28. DOI: 10.1039/C5CP00315F.

Zhu, F. and Gerhard Hummer (2010). "Pore opening and closing of a pentameric ligand-gated ion channel". In: Proceedings of the National Academy of Sciences 107.46. DOI: 10.1073/pnas.1009313107.

Zwanzig, Robert W. (1954). "High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases". In: The Journal of Chemical Physics 22.8. DOI: 10.1063/1.1740409.

Zwanzig, Robert W. (2001). Nonequilibrium statistical mechanics. Oxford ; New York: Oxford University Press. 222 pp.

Acronyms

ABF Adaptive Biasing Force.

aMD Accelerated Molecular Dynamics.

ATP Adenosine Tri-Phosphate.

BBK Brooks-Brünger-Karplus.

CAM calmodulin.

CPR Conjugate Peak Refinement.

CV Collective Variable.

CVSM String Method in Collective Variables.

CZAR Corrected z-Averaged Restraint.

Dd myo2 *Dictyostelium discoideum* myosin II.

eABF Extended Adaptive Biasing Force.

FEP Free Energy Perturbation.

FPI Fischer Putative Intermediate.

FTS Finite-Temperature String Method.

gABF Generalized Adaptive Biasing Force.

GaMD Gaussian Accelerated Molecular Dynamics.

HMM Heavy Mero-Myosin.

IDP Intrinsically Disordered Protein.

L50 Lower 50 kDa.

MD Molecular Dynamics.

MEP Minimum Energy Path.

MFEP Minimum Free Energy Path.

MPTP Most Probable Transition Path.

MSM Markov State Models.

myo6 myosin VI.

PME Particle Mesh Ewald.

PMF Potential of Mean Force.

PPS Pre-Powerstroke State.

PR Post Rigor State.

PTS Pre-Transition State.

RMSD Root-Mean-Square Deviation.

SAH Single Alpha Helix.

SAXS Small-angle X-ray scattering.

SMD Steered Molecular Dynamics.

SMM2 Smooth Muscle Myosin II.

TAMD Temperature-Accelerated Molecular Dynamics.

TI Thermodynamic Integration.

TMD Targeted Molecular Dynamics.

U50 Upper 50 kDa.

UI Umbrella Integration.

US Umbrella Sampling.

WHAM Weighted Histogram Analysis Method.

ZTS Zero-Temperature String Method.

Appendices

A. Complementary theoretical notions

A.1. Justification of a classical description

In this section, we justify why quantum mechanics is not needed in the description of molecular machines.

At the beginning of the 20$^{\text{th}}$ century, statistical mechanics was still grounded in classical mechanics, *i.e.* Newtonian mechanics and its generalizations. However, it became apparent that classical mechanics was inadequate to properly describe microscopic objects such as electrons or atoms. This motivated the development of a novel theory, quantum mechanics, which was built upon classical mechanics but introduced a number of new concepts which were generally considered counter-intuitive (Cohen-Tannoudji, Diu, and Laloë 2008).

Essentially, quantum mechanics abandons the point-like particle idealization in favour of a wave-like representation. A complex wavefunction φ is introduced, whose square-modulus represents the positional probability density of the particle.

This allows the theory to account for phenomena such as delocalization and interference, which are observed with photons but also with "particles of matter" such as electrons, atoms or even molecules. A remarkable result is Heisenberg's theorem, which states that one cannot simultaneously achieve arbitrary precision on the position and momentum of a particle. Denoting by Δx the standard deviation of the position (with respect to the quantum probability density), and Δp the standard deviation of the momentum, Heisenberg's theorem states that:

$$\Delta x \Delta p \geq \frac{\hbar}{2} \tag{A.1}$$

where $\hbar = h/2\pi$ is the reduced Planck constant.

Importantly, quantum mechanics could be used to describe the behaviour of (non-relativistic macroscopic) objects, but it is usually not worth the trouble. Classical mechanics works fine on the macroscopic scale; in fact, surprisingly it is still mostly applicable even for some nanoscopic phenomena. This claim seems to contradict the very reason why quantum mechanics was introduced, so let us substantiate it qualitatively using order-of-magnitude approaches.

We rewrite Heisenberg's relation as an order-of-magnitude approximation $\Delta x \Delta p \sim \hbar$. Considering a nanoscopic object at temperature T, (classical) statistical mechanics reveals that the dispersion of its velocity v is of order $\sqrt{\frac{k_B T}{m}}$, with m the mass of the object, and k_B Boltzmann's constant. Combining the two relations and with $\Delta p = m\Delta v$, one finds:

$$\Delta x \sim \frac{\hbar}{\sqrt{m k_B T}} \tag{A.2}$$

Δx is the typical extent of spatial quantum delocalization expected for the system under study, *i.e.* it tells us whether the system exhibits "quantum behaviour" or not: this will be the case if Δx is of the same order of magnitude (or larger) as the size of the object. Δx is called the De Broglie thermal

wavelength[1] and emerges naturally when studying equilibrium statistical mechanics (see Appendix, A.3).

A protein the size of myosin has roughly 1.2×10^4 atoms; taking half of these as hydrogen atoms and the rest as an equal mix of carbon, nitrogen and oxygen yields a total mass of about (6000 + 12x2000 + 14x2000 + 16x2000) atomic mass units, *i.e.* 9×10^4 u or about 1.3×10^{-22} kg. At T =300 K, k_BT =4×10^{-21} J. With $\hbar \simeq 1 \times 10^{-34}$ J s, we finally arrive at $\Delta x \sim 1 \times 10^{-13}$ m. By contrast, the typical size of a protein like myosin is $l \sim 1$ nm. Clearly, we have $\Delta x \ll l$: quantum de-localization effects are not expected to play a major role at the nanometric scale, and classical mechanics is a good approximation for the description of molecular machines. In addition, this somewhat justifies the use of classical potentials and motion equations to investigate (bio)-molecular systems by computer simulations at atomic resolution. Of course, certain phenomena, first and foremost chemical reactions involving electronic rearrangements, are purely quantum in nature; these would have to be treated by the appropriate level of theory.

A.2. Classical mechanics

Although initial progress in classical mechanics can be traced back to Antiquity, the most crucial advances are arguably due to Galileo and Isaac Newton. Classical mechanics is concerned with the movement of bodies and has proved adequate to account for the behavior of macroscopic objects. It can notably be used with remarkable accuracy to understand celestial dynamics. During the 18th and 19th centuries, classical mechanics was reformulated in a more flexible framework, analytical mechanics, which allows for more generality in tackling difficult problems (e.g. with non-cartesian coordinates).

A.2.1. Newtonian mechanics

A.2.1.1. Forces

Newtonian mechanics is built around the concept of force, *i.e.* a vectorial quantity which is used to describe physical interactions between objects. The specific nature of the force depends on the problem at-hands; Newtonian mechanics is concerned with providing a general framework to study the movement of an object (dynamics) under the influence of the forces applied on it. To this end, the theory rests on three fundamental laws, Newton's laws of motion, which are sufficient to prescribe the equation of dynamics. These laws are:

1. Inertia principle: in an inertial frame of reference, if no force is applied on an object, its movement is rectilinear (straight line) and uniform (zero acceleration).

2. Fundamental principle of dynamics: in an inertial frame of reference, if a force $\boldsymbol{F}$ is applied on an object of constant mass m, then its acceleration $\boldsymbol{a}$ satisfies:

$$m\boldsymbol{a} = \boldsymbol{F} \tag{A.4}$$

1. Its actual definition differs from equation A.2 by an irrelevant multiplicative constant:

$$\Delta x = \frac{h}{\sqrt{2\pi m k_B T}} \tag{A.3}$$

3. Action/Reaction principle: if an object a exerts a force $\boldsymbol{F}_{a/b}$ on an object b, then b exerts a force $\boldsymbol{F}_{b/a} = -\boldsymbol{F}_{a/b}$ on a.

The description of motion requires the introduction of a reference frame; in a rather tautological manner, we will define an *inertial* reference frame as one in which the inertia principle applies. The existence of such frames is postulated, and we will always assume that we are working in a properly defined, inertial "laboratory reference frame".

Given a reference frame, we introduce the position-vector $\boldsymbol{r}$ of the object in motion. Its successive time-derivatives are the velocity $\boldsymbol{v} = \dot{\boldsymbol{r}}$ and acceleration $\boldsymbol{a} = \ddot{\boldsymbol{r}}$. The problem of dynamics is to obtain the trajectory $\boldsymbol{r}(t)$ by resolving the second-order differential equation A.4.

A.2.1.2. Energy

Rather than focusing on the notion of force, one can instead focus on the notion of *energy*. Instead of trying to give a verbal definition of energy, which is surprisingly not easy, we will stick to mathematical definitions. For notational convenience, we now drop the bold-face when writing vectors.

Kinetic energy We first introduce the *kinetic energy* K defined as:

$$K(\dot{r}) = \frac{1}{2} m \dot{r}^2 \tag{A.5}$$

For a system of N particles, each with position-vector r_i and mass m_i, the kinetic energy is additive:

$$K(\dot{r_1}, ..., \dot{r_N}) = \frac{1}{2} \sum_{i=1}^{N} m_i \dot{r}_i^2 \tag{A.6}$$

Work For a force F acting on an object of mass m along the infinitesimal displacement $\mathrm{d}l$ (which is a vectorial quantity), we define the (infinitesimal) *work* δW of F as:

$$\delta W = F \cdot \mathrm{d}l \tag{A.7}$$

The total work between points A and B is obtained by integration:

$$W = \int_A^B F \cdot \mathrm{d}l \tag{A.8}$$

Just like the kinetic energy, the work is an energy, having dimensional force $\times$ displacement.

Potential Energy We define a *conservative force* as a force F such that:

$$F \equiv -\nabla U \tag{A.9}$$

where ∇ refers to the gradient operator in cartesian coordinates and U is a function of position (*i.e.* the three cartesian coordinates), called the *potential energy function* or simply *potential*. If there are N particles, U is a function of their $3N$ cartesian coordinates. For a one-dimensional problem, let us derive the work W_c done by a conservative force:

$$W_c = \int_A^B F \mathrm{d}r = -\int_A^B \frac{\mathrm{d}U}{\mathrm{d}x}\mathrm{d}x = -\int_A^B \mathrm{d}U = U_A - U_B \tag{A.10}$$

The work done by a conservative force is equal to the potential energy difference between the end-points of the displacement; it is independent of the path followed during the movement. Notably, this implies that the work done by a conservative force over a closed trajectory is always zero.

Mechanical energy For a point of mass m with position r and velocity $\dot{r}$, the mechanical energy E_m is defined as the sum of the potential and kinetic energies:

$$E_m = \frac{1}{2}m\dot{r}^2 + U(r) \tag{A.11}$$

This definition is naturally extended to the case of N particles.

Some theorems about energy in Newtonian mechanics We now derive some useful properties of the mechanical energy and its kinetic and potential components. Let us consider a particle of mass m undergoing a conservative force $F = -\nabla U$, and thus following a trajectory $r(t)$ solution of the equation of dynamics A.4. Introducing its mechanical energy $E_m(r(t), \dot{r}(t))$, we have:

$$\begin{aligned} \frac{\mathrm{d}E_m}{\mathrm{d}t} &= \frac{\partial K}{\partial \dot{r}}\frac{\mathrm{d}\dot{r}}{\mathrm{d}t} + \frac{\partial U}{\partial r}\frac{\mathrm{d}r}{\mathrm{d}t} \\ &= (m\ddot{r} - F)\,\dot{r} \\ &= 0 \end{aligned} \tag{A.12}$$

where the fundamental principle of dynamics (equation A.4) has been used to go from the second to third lines. Thus, if the system is acted upon only by conservative forces, its mechanical energy remains constant - it is conserved, hence the name *conservative*. What if there are also non-conservative forces at play? In one-dimension for simplicity, and along an infinitesimal displacement $\mathrm{d}r$, one writes:

$$\begin{aligned} &(m\ddot{r} - F_c - F_{nc})\,\mathrm{d}r = 0 \\ &m\ddot{r}\mathrm{d}r + \mathrm{d}U = F_{nc}\mathrm{d}r \\ &\frac{1}{2}m\mathrm{d}(\dot{r}^2) + \mathrm{d}U = \delta W_{nc} \\ &\mathrm{d}\left(K + U\right) = \delta W_{nc} \\ &\mathrm{d}E_m = \delta W_{nc} \end{aligned} \tag{A.13}$$

where F_c is the conservative force, F_{nc} the non-conservative force, and δW_{nc} the infinitesimal work of the non-conservative force. Thus, the variation of mechanical energy equals the work of non-conservative forces, a result sometimes called the *mechanical energy theorem*. A force which tends to decrease the mechanical energy is termed *dissipative*; this is the case, for example, for a friction force of the form $F_{nc} = -\gamma\dot{r}$, with $\gamma > 0$. Finally, taking the differential of equation A.10, we have $\delta W_c = -\mathrm{d}U$, where δW_c is the work of the conservative force. The final line of equation A.13 becomes:

$$\mathrm{d}K = \delta W_{nc} + \delta W_c \tag{A.14}$$

which shows that the variation of kinetic energy equals the work of all forces acting on the system: this is the *kinetic energy theorem*. In Newtonian mechanics, exploiting the energy theorems and the conservation law A.12 is sometimes a more straightforward route to problem resolution than the direct resolution of the equation of dynamics A.4. Later formulations of classical mechanics, which we now briefly outline, similarly treat energy as the fundamental quantity.

A.2.2. Lagrangian formulation and the extremal-action principle

The Newtonian formulation of mechanics is generally suited to the study of problems in cartesian coordinates, but may become hardly tractable if different coordinate systems are used. Lagrangian mechanics is a complete reformulation of classical mechanics with conservative forces, introduced by J.L. Lagrange in the 18$^{\text{th}}$ century. As we illustrate below, this formalism provides a natural framework for changes of coordinate systems.

The fundamental object of Lagrangian mechanics is the Lagrangian function, or Lagrangian $\mathcal{L}$. With the notations of the chapter, one has:

$$\mathcal{L}(\{r_i\},\{\dot{r}_i\}) = K(\{\dot{r}_i\}) - U(\{r_i\}) \tag{A.15}$$

where a system of N particles is considered. Clearly, the Lagrangian is homogeneous to an energy.

A.2.2.1. Generalized coordinates

Let us now introduce a new set of $3N$ *generalized coordinates* $\{q_\alpha\}_{\alpha=1,...,3N}$ such that $q_\alpha = f_\alpha(r_1,...,r_N)$.

Furthermore, we assume the existence of the inverse transformation, that is, there is a family of (vector) functions g_i satisfying $r_i = g_i(q_1,...,q_{3N})$. Note that unlike the $(r_i)_{i=1,...,N}$, the $(q_\alpha)_{\alpha=1,...,3N}$ are scalar numbers.

How does the Lagrangian transform under the change of coordinates?

Mass metric tensor We begin by deriving the new form of the kinetic energy $K(r_1,...,r_N)$. Using the chain rule, we obtain

$$\dot{r}_i = \sum_{\alpha=1}^{3N} \frac{\partial r_i}{\partial q_\alpha} \cdot \dot{q}_\alpha \tag{A.16}$$

Inserting equation A.16 into the kinetic energy definition A.5 yields:

$$K(r_1,...,r_N) = \frac{1}{2}\sum_{\alpha=1}^{3N}\sum_{\beta=1}^{3N}\sum_{i=1}^{N} m_i \frac{\partial r_i}{\partial q_\alpha}\frac{\partial r_i}{\partial q_\beta}\cdot \dot{q}_\alpha \dot{q}_\beta \equiv \tilde{K}(\dot{q}_1,...,\dot{q}_{3N},q_1,...,q_{3N}) \tag{A.17}$$

We call G the matrix of generic element:

$$G_{\alpha\beta} \equiv \sum_{i=1}^{N} m_i \frac{\partial r_i}{\partial q_\alpha}\frac{\partial r_i}{\partial q_\beta} \tag{A.18}$$

G is called the *mass metric tensor*, and is symmetric. The new expression for the kinetic energy is thus:

$$\tilde{K}(\dot{q}_1, ..., \dot{q}_{3N}, q_1, ..., q_{3N}) = \frac{1}{2}\sum_{\alpha=1}^{3N}\sum_{\beta=1}^{3N} G_{\alpha\beta}\dot{q}_\alpha\dot{q}_\beta \tag{A.19}$$

Remark We note that the definition of kinetic energy in cartesian coordinates A.5 appears as a special case of A.17 with G the diagonal matrix with each mass repeated three times.

We can now write down the Lagrangian in generalized coordinates:

$$\mathcal{L} = \frac{1}{2}\sum_{\alpha=1}^{3N}\sum_{\beta=1}^{3N} G_{\alpha\beta}\dot{q}_\alpha\dot{q}_\beta - U\left(r_1(q_1, ..., q_{3N}), ..., r_N(q_1, ..., q_{3N})\right) \tag{A.20}$$

Or, in condensed notation with $\tilde{U}(q_1, ..., q_{3N}) \equiv U(r_1(q_1, ..., q_{3N}), ..., r_N(q_1, ..., q_{3N}))$:

$$\mathcal{L} = \tilde{K} - \tilde{U} \tag{A.21}$$

A.2.2.2. Action extremization and Euler-Lagrange equations

The above considerations have so far not given anything interesting regarding the study of physical systems. We have introduced the Lagrangian as the difference of the kinetic and potential energies, and discussed its behavior under an invertible change to a set of generalized coordinates. In the following we show how this formalism, in combination with an important physical principle, allows the derivation of equation of motions for the generalized coordinates.

Let us assume that the system, described by a set of generalized coordinates $q = \{q_\alpha\}$, undergoes a physical transformation from an initial state (t_1, q_1) to a final state (t_2, q_2) (t is the time). We define the *action integral* $\mathcal{S}$ as:

$$\mathcal{S} \equiv \int_{t_1}^{t_2} \mathcal{L}(q(t), \dot{q}(t))\mathrm{d}t \tag{A.22}$$

$\mathcal{S}$ is a functional of the trajectory $q(t)$ followed by the system from the initial to the final state. Accordingly, we write $\mathcal{S} = \mathcal{S}[q(t)]$.

Our goal is to determine this trajectory using the Lagrangian formalism. This is possible by invoking the **action extremization principle**, which states that the realized trajectory renders $\mathcal{S}$ extremal.

This suggests that the trajectory may be calculated by taking the "derivative" of the action with respect to the trajectory, which is a function. Such a quantity is defined in variational calculus. Consider a second trajectory $q'(t)$ with the same end-points as $q(t)$, and define $\delta q \equiv q' - q$. We define $\delta\mathcal{S}$ as:

$$\delta\mathcal{S} \equiv \mathcal{S}[q'] - \mathcal{S}[q] \tag{A.23}$$

$$\delta\mathcal{S} = \int_{t_1}^{t_2} \left[\mathcal{L}(q'(t), \dot{q}'(t)) - \mathcal{L}(q(t), \dot{q}(t))\right] \mathrm{d}t \tag{A.24}$$

For $\delta q \to 0$, we can expand $\mathcal{L}(q'(t), \dot{q}'(t)$ to first order, which yields (formally):

$$\delta\mathcal{S} = \int_{t_1}^{t_2} \left[\frac{\partial\mathcal{L}}{\partial q}\cdot\delta q + \frac{\partial\mathcal{L}}{\partial\dot{q}}\cdot\delta\dot{q}\right] \mathrm{d}t \tag{A.25}$$

The second term of the integral can be integrated by parts:

$$\delta\mathcal{S} = \int_{t_1}^{t_2} \frac{\partial \mathcal{L}}{\partial q} \cdot \delta q \mathrm{d}t + \left[\frac{\partial \mathcal{L}}{\partial \dot{q}} \cdot \delta q\right]_{t_1}^{t_2} - \int_{t_1}^{t_2} \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial \dot{q}} \cdot \delta q \mathrm{d}t \tag{A.26}$$

The term between brackets vanishes because $\delta q(t_1) = \delta q(t_2) = 0$ by construction. Thus, the functional differential of the action $\delta\mathcal{S}$ reads:

$$\delta\mathcal{S} = \int_{t_1}^{t_2} \left[\frac{\partial \mathcal{L}}{\partial q} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial \dot{q}}\right] \cdot \delta q \mathrm{d}t \tag{A.27}$$

In accordance with the action extremization principle, the actual trajectory satisfies $\delta\mathcal{S}[q] = 0$. This yields the **Euler-Lagrange equation**:

$$\boxed{\frac{\partial \mathcal{L}}{\partial q} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial \dot{q}} = 0} \tag{A.28}$$

Note that equation A.28 is equivalent to Newton's second law (equation A.4), as can be seen by simple substitution. The Lagrangian formalism is a reformulation of Newton's mechanics, and does not introduce new physical results; however, it makes calculations simpler in many cases.

A.2.2.3. Constraints

The Lagrangian formulation is well-suited to constrained dynamics, *e.g.* if the dynamics is forced to remain on a given surface, or if the distance between moving points is kept constant. This latter case is of interest for Molecular Dynamics simulations, because constraining the length of covalent bonds involving hydrogen atoms allows for the elimination of the fastest motions in the system and as such for the use of a larger integration timestep (Ryckaert, Ciccotti, and Berendsen 1977). Constraints can also be used for free energy calculations (see 4.3.4.2). We will limit ourselves to constraints of the form:

$$\sigma_k(r_1, ..., r_N, t) = 0 \tag{A.29}$$

where $k = 1, ..., N_c$, N_c being the number of constraints applied on the system.

Such a constraint does not involve the momenta and is called *holonomic*. For example, a distance constraint would take the form $||r_1 - r_2||^2 - d^2 = 0$. Dealing with an holonomic constraint typically involves the introduction of a Lagrange multiplier λ, and the constraining force takes the form $f_c = \lambda\nabla\sigma$ along with the condition $\nabla\sigma \cdot \dot{r} = 0$. The procedure to solve for λ, and thus for the constrained dynamics, is presented in (Tuckerman 2010) to which the reader is referred.

A.2.3. Hamiltonian formulation

The Hamiltonian formalism is a popular alternative to the Lagrangian one, in which the $3N$ second-order differential equations are replaced by $6N$ first order equations. It was proposed in the 19$^{\text{th}}$ century by William R. Hamilton.

A.2.3.1. Hamiltonian function

We first introduce the generalized momenta $p_\alpha \equiv \dfrac{\partial \mathcal{L}}{\partial \dot{q_\alpha}}$. Inserting it into the Euler-Lagrange equation A.28 (going to a single dimension for simplicity), one gets:

$$\dot{p} = \frac{\partial \mathcal{L}}{\partial q} \tag{A.30}$$

We now define the Hamiltonian function, or simply Hamiltonian, $\mathcal{H}$, as:

$$\mathcal{H}(q,p) = p\dot{q} - \mathcal{L}(q,\dot{q}) \tag{A.31}$$

Taking the differential of A.31, one gets:

$$\mathrm{d}\mathcal{H} = \mathrm{d}(p\dot{q}) - \mathrm{d}\mathcal{L} \tag{A.32}$$

$$\mathrm{d}\mathcal{H} = p\mathrm{d}\dot{q} + \dot{q}\mathrm{d}p - \frac{\partial \mathcal{L}}{\partial q}\mathrm{d}q - \frac{\partial \mathcal{L}}{\partial \dot{q}}\mathrm{d}\dot{q} \tag{A.33}$$

Since $p \equiv \dfrac{\partial \mathcal{L}}{\partial \dot{q}}$, terms in $\mathrm{d}\dot{q}$ drop from equation A.33, effectively eliminating explicit dependence of $\mathcal{H}$ on generalized velocities $\mathrm{d}\dot{q}$ and replacing it by dependence on the generalized momenta p. This is an example of Legendre transform. It yields:

$$\mathrm{d}\mathcal{H} = \dot{q}\mathrm{d}p - \frac{\partial \mathcal{L}}{\partial q}\mathrm{d}q \tag{A.34}$$

After some algebra (either by direct substitution in A.31 or by integrating A.34 once the expression is simplified), one gets:

$$\mathcal{H}(q,p) = \frac{p^2}{2m} + U(q) \tag{A.35}$$

A.2.3.2. Hamiltonian equations of motion

Previously, we introduced the Lagrangian formalism as a way to derive generalized equations of motion. The motivation for the Hamiltonian formalism is the same and we now proceed to establish the form taken by the motion equations.

Taking the differential of $\mathcal{H}$ yields:

$$\mathrm{d}\mathcal{H} = \frac{\partial \mathcal{H}}{\partial q}\mathrm{d}q + \frac{\partial \mathcal{H}}{\partial p}\mathrm{d}p \tag{A.36}$$

By term-by-term identification with equation A.34, and remembering that $\dot{p} = \partial_q \mathcal{L}$, one directly gets:

$$\boxed{\begin{cases} \dot{q} = \frac{\partial \mathcal{H}}{\partial p} \\ \dot{p} = -\frac{\partial \mathcal{H}}{\partial q} \end{cases}} \tag{A.37}$$

As planned, the second-order Euler-Lagrange equation has been turned into two coupled first order equations, Hamilton's equations.

A.2.3.3. N-particle case

We now return to the case where N particles are described by $3N$ generalized coordinates q_α and show that the previous relations still hold. We follow the derivation given in (Tuckerman 2010). The definition of the Hamiltonian A.31 now reads:

$$\mathcal{H} = \sum_{\alpha=1}^{3N} p_\alpha \dot{q_\alpha} - \mathcal{L} \tag{A.38}$$

This time, we have:

$$\begin{aligned} p_\alpha &= \frac{\partial \mathcal{L}}{\partial \dot{q}_\alpha} \\ &= \sum_{\beta=1}^{3N} G_{\alpha\beta}\dot{q}_\beta \end{aligned} \tag{A.39}$$

where the symmetry of G has been used to obtain the last line. This relation can be inverted to obtain:

$$\dot{q}_\alpha = \sum_{\beta=1}^{3N} G^{-1}_{\alpha\beta} p_\beta \tag{A.40}$$

with

$$G^{-1}_{\alpha\beta} = \sum_{i=1}^{N} \frac{1}{m_i} \left(\frac{\partial q_\alpha}{\partial r_i}\right)\left(\frac{\partial q_\beta}{\partial r_i}\right) \tag{A.41}$$

Inserting equation A.40 into equation A.38 yields the expression for the Hamiltonian:

$$\mathcal{H} = \frac{1}{2}\sum_{\alpha=1}^{3N}\sum_{\beta=1}^{3N} p_\alpha G^{-1}_{\alpha\beta} p_\beta + U\left(r_1(q_1, ..., q_{3N}), ..., r_N(q_1, ..., q_{3N})\right) \tag{A.42}$$

Hamilton's equations are left unchanged:

$$\boxed{\begin{cases} \dot{q}_\alpha = \frac{\partial \mathcal{H}}{\partial p_\alpha} \\ \dot{p}_\alpha = -\frac{\partial \mathcal{H}}{\partial q_\alpha} \end{cases}} \tag{A.43}$$

In the Hamiltonian framework, the state of the system is described by a point in the $6N$-dimensional space of generalized positions and momenta. This space is called the *phase space* and Hamilton's equations prescribe the movement of the point in phase-space.

A.2.4. Classical propagator, Liouville theorem and Liouville equation

A.2.4.1. Evolution of a phase-space observable under Hamiltonian dynamics

Let us consider a function A of the generalized positions and momenta $A = A(q, p, t)$. Functions of this form are called *phase-space observables*. We are interested in the evolution of A as the system undergoes Hamiltonian dynamics. We can write:

$$\mathrm{d}A = \partial_q A \mathrm{d}q + \partial_p A \mathrm{d}p + \partial_t A \mathrm{d}t \tag{A.44}$$

which leads to:

$$\dot{A} = \partial_q A \dot{q} + \partial_p A \dot{p} + \partial_t A \tag{A.45}$$

Inserting the Hamilton equations A.37 in A.45 yields the evolution equation for A:

$$\frac{\mathrm{d}A}{\mathrm{d}t} = \frac{\partial A}{\partial q}\frac{\partial \mathcal{H}}{\partial p} - \frac{\partial A}{\partial p}\frac{\partial \mathcal{H}}{\partial q} + \frac{\partial A}{\partial t} \tag{A.46}$$

Introducing the Poisson bracket $\{\cdot, \mathcal{H}\} = \partial_p \mathcal{H} \partial_q - \partial_q \mathcal{H} \partial_p$, we can rewrite A.46 as:

$$\frac{\mathrm{d}A}{\mathrm{d}t} = \{A, \mathcal{H}\} + \frac{\partial A}{\partial t} \tag{A.47}$$

Some authors also define the Liouville operator L such that $iL \equiv \{\cdot, \mathcal{H}\}$. With this notation equation A.47 rewrites:

$$\frac{\mathrm{d}A}{\mathrm{d}t} = iLA + \frac{\partial A}{\partial t} \tag{A.48}$$

If A has no explicit time-dependence ($\partial_t A = 0$), equation A.48 has the formal solution:

$$A(t) = e^{iLt} A(0) \tag{A.49}$$

or, for two times t, t':

$$A(t') = e^{iL(t'-t)} A(t) \tag{A.50}$$

Applying the operator $U(t'-t) \equiv e^{iL(t'-t)}$ on a phase-space observable translates it in time by $t'-t$. For this reason. this operator is called the **classical propagator**, or classical evolution operator. It is a useful object to systematically derive numerical integrators for classical dynamics, which are built by splitting the propagator into separate components, then taking finite-order expansions of each (Tuckerman 2010; Tuckerman, Berne, and Martyna 1992).

Finally, we note that if A is a conserved quantity, it satisfies the equation:

$$\{A, \mathcal{H}\} = 0 \tag{A.51}$$

A.2.4.2. Liouville theorem - Volume-preservation in phase space

We now demonstrate a very important property of Hamiltonian dynamics, conservation of volume in phase space.

Let us consider the "point in phase-space" $\Gamma(t) = (q(t), p(t))$. By Hamiltonian evolution:

$$\Gamma(t + \mathrm{d}t) = (q(t + \mathrm{d}t), p(t + \mathrm{d}t)) = (q + \dot{q}\mathrm{d}t, p + \dot{p}\mathrm{d}t) \tag{A.52}$$

Substituting Hamilton's equations A.37 into A.52:

$$\Gamma(t + \mathrm{d}t) = (q + \partial_p \mathcal{H}\mathrm{d}t, p - \partial_q \mathcal{H}\mathrm{d}t) \tag{A.53}$$

It is possible to look at the translation in time from t to $t + \mathrm{d}t$ under Hamiltonian dynamics as a variable change in phase-space from $q = q(t),\ p = p(t)$ to $q' = q(t + \mathrm{d}t),\ p' = p(t + \mathrm{d}t)$. The

Jacobian matrix of this transformation reads:

$$J = \begin{bmatrix} \frac{\partial q'}{\partial q} & \frac{\partial q'}{\partial p} \\ \frac{\partial p'}{\partial q} & \frac{\partial p'}{\partial p} \end{bmatrix} \tag{A.54}$$

From equation A.53, J rewrites:

$$J = \begin{bmatrix} 1 - \partial^2_{pq}\mathcal{H}\mathrm{d}t & \partial^2_{p^2}\mathcal{H}\mathrm{d}t \\ -\partial^2_{q^2}\mathcal{H}\mathrm{d}t & 1 + \partial^2_{pq}\mathcal{H}\mathrm{d}t \end{bmatrix} \tag{A.55}$$

From expression A.55, it comes that the Jacobian determinant $|J|$ of the transformation can be written:

$$|J| = 1 + O(\mathrm{d}t^2) \tag{A.56}$$

Thus, the coordinate transformation in phase-space under Hamiltonian dynamics has a unit Jacobian determinant; this shows that it preserves the volume in phase-space. This result is called **Liouville's theorem**.

A.2.4.3. Liouville equation and phase-space probability density

As a first step towards statistical mechanics, we introduce a probability density in phase-space $\rho(p, q, t)$, which measures the probability of finding the system at the phase-space point (p, q) at time t (a more proper definition of ρ will be given in section A.3.3). ρ is a phase-space observable, so if p and q are taken as the solutions to Hamilton's equations, its time-evolution can be written:

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \{\rho, \mathcal{H}\} + \frac{\partial \rho}{\partial t} \tag{A.57}$$

It can be shown by Liouville's theorem that the total time derivative $\dfrac{\mathrm{d}\rho}{\mathrm{d}t}$ is 0. This yields Liouville's equation:

$$\frac{\partial \rho}{\partial t} = -\{\rho, \mathcal{H}\} \tag{A.58}$$

Liouville's equation gives the time-evolution of the phase-space probability density under Hamiltonian dynamics. It is a fundamental equation for statistical mechanics, as 1) stationary solution(s) describe the equilibrium probability distribution of the system and 2) the time-dependent behaviour describes relaxation towards equilibrium. However, this equation cannot be solved directly for systems with large numbers of degrees of freedom.

A.3. Complements on classical statistical mechanics

In this section, we first outline the microcanonical formalism. Although of limited interest in simulation applications, it represents the starting point for all of equilibrium statistical mechanics, and allows for the rigorous definition of thermodynamic quantities. In a second time, we briefly outline the conceptual foundations of classical statistical mechanics by discussing the use of probabilistic approaches for deterministic dynamics, the ergodic hypothesis, and the importance of chaos.

A.3.1. Phase equiprobability and the microcanonical ensemble

Let us consider an isolated system of N particles, volume V and energy E. Since the system is isolated these values are constant and define the macroscopic state of the system. The associated statistical ensemble (see below, A.3.3) is the so-called *microcanonical* ensemble, or NVE ensemble. Although its practical relevance is rather limited, the microcanonical ensemble forms the starting point for the derivation of other ensembles and the definition of thermodynamic properties (such as temperature) in statistical mechanical terms.

In the classical setting, the system is described by a Hamiltonian $\mathcal{H}(q,p)$ defined over the $6N$-dimensional phase space. Since the total energy is fixed to E, the trajectories are confined to the $6N-1$ dimensional hypersurface $\Sigma(E)$ such that $\mathcal{H}(q,p) = E$ (and such that the system also satisfies the additional constraints, like constant volume V). Actually, especially for large systems which are the main concern of statistical mechanics, the total energy is not known exactly but up to some uncertainty $\Delta E \ll E$, such that a more appropriate definition of $\Sigma(E)$ is $\Sigma(E) = \{(p,q), E \leq \mathcal{H}(q,p) \leq E + \Delta E\}$.

The properties of the system are entirely contained in the phase-space probability distribution $\rho(q,p)$, whose evolution is given by Liouville's equation A.58. At equilibrium, one has:

$$\{\rho, \mathcal{H}\} = 0 \tag{A.59}$$

A priori, there are an infinity of possible solutions of equation A.59. We choose to focus on distribution functions of the form $\rho(q,p) = \rho(\mathcal{H}(q,p))$, *i.e.* functions which depend on the point in phase-space only through the value of the Hamiltonian. Clearly, by relation A.51, such functions are stationary solutions of Liouville's equation A.58. Since only the constant-energy hypersurface is physically accessible, $\rho(q,p)$ should be non-zero only if $(q,p) \in \Sigma(E)$. And, since by hypothesis ρ depends only on $\mathcal{H} = E$, it is constant on $\Sigma(E)$. Thus, this hypothesis is equivalent to assume that at equilibrium all the micro-states compatible with the condition $\mathcal{H}(q,p) = E$ have equal probability. The microcanonical distribution thus has the form:

$$\begin{cases} \rho(q,p) = cst \text{ if } (q,p) \in \Sigma(E) \\ \rho(q,p) = 0 \text{ otherwise.} \end{cases} \tag{A.60}$$

This is the so-called **phase equiprobability postulate** and it is one of the only two physical postulates that has to be invoked in order to derive equilibrium statistical mechanics (the other being the ergodic hypothesis, although whether this latter is an absolute requirement is controversial).

Introducing the number $\Omega(E,V,N)$ of micro-states belonging to $\Sigma(E)$, the microcanonical distribution A.60 has constant value $1/\Omega(E,V,N)$ on $\Sigma(E)$. It is clear that $\Omega(E,V,N)$ is expected to be proportional to the "volume" $\Delta\Gamma$ occupied in phase-space by $\Sigma(E)$:

$$\Delta\Gamma = \int_{(p,q)\in\Sigma(E)} \mathrm{d}p\mathrm{d}q \tag{A.61}$$

Since the generalized positions and momenta $q = q_1, ..., q_{3N}$ and $p = p_1, ..., p_{3N}$ are continuous variables, we need to introduce an arbitrary "length scale" in phase space $\delta q = \prod_{i=1}^{3N} \delta q_i$ and $\delta p = \prod_{i=1}^{3N} \delta p_i$ to compute a (finite) number of (discrete) micro-states. The product $u \equiv (\delta q \delta p)^{3N}$ represents the volume of a micro-state. In this picture, the phase space is tessellated by hypercubic "phase space elementary cells" of volume u. With this definition, Ω becomes the ratio of the total volume of $\Sigma(E)$ and the volume of an elementary cell, that is:

$$\Omega(E,V,N) = \frac{1}{u}\int_{(p,q)\in\Sigma(E)} \mathrm{d}p\mathrm{d}q \tag{A.62}$$

The choice of an elementary phase-space volume can be guided by quantum mechanics. Let us consider two points in phase space (with one particle and one spatial dimension to simplify) (q,p) and $(q'=q+\delta q, p'=p+\delta p)$. What is the condition on δq and δp for these two points to belong to the same micro-states? Using Heisenberg's relation, we see that (q,p) and (q',p') belong to different micro-states only if δq and δp are chosen such that $\delta q \delta p \geq \frac{\hbar}{2}$. This suggests that the natural volume of a phase space cell is of the order of the Planck constant h. Thus (coming back to N particles and 3 spatial dimensions) it is customary to take $u = h^{3N}$. $\Omega(E,V,N)$ is sometimes called the **microcanonical partition function**.

A.3.1.1. Microcanonical entropy and microcanonical quantities

Boltzmann introduced a microcanonical entropy S as follows:

$$S = k_B \ln \Omega(E,V,N) \tag{A.63}$$

Let us note that with this definition, it seems that the value of S depends on the choice of the elementary phase space volume u and the energy uncertainty ΔE. In fact, the dependence on u is irrelevant since almost only entropy variations will be of interest; in addition, it can be justified that ΔE appears only through an $\mathcal{O}(\ln \Delta E)$ contribution, which is negligible (see for instance Diu 1989). The choice of $u = h^{3N}$, already justified above, notably allows to recover the thermodynamic entropy for the ideal gas, starting from formula A.63 (Diu 1989).

Temperature The (microcanonical) temperature T is *defined* as:

$$\frac{1}{T} \equiv \frac{\partial S}{\partial E}\Big)_{V,N} \tag{A.64}$$

Thus, temperature measures how the number of accessible states of a system changes when its energy changes.

Pressure The (microcanonical) pressure P is defined as:

$$\frac{P}{T} \equiv \frac{\partial S}{\partial V}\Big)_{E,N} \tag{A.65}$$

Chemical potential The (microcanonical) chemical potential μ is defined as:

$$\frac{\mu}{T} \equiv -\frac{\partial S}{N}\Big)_{E,V} \tag{A.66}$$

It is thus clear that entropy acts as a generating function for thermodynamic quantities.

A.3.2. Derivation of the canonical ensemble from the microcanonical ensemble

The canonical ensemble applies for the case of constant temperature, rather than internal energy. The probability distribution associated with the canonical ensemble, called canonical distribution or Boltzmann's distribution, is thus central to the study of thermalized systems. We give an elementary derivation of this distribution, which follows directly from the phase equiprobability postulate. The proof is adapted from (Diu 1989).

First let us consider an isolated system $\mathcal{S}_{tot}$ of energy E_{tot}, volume V_{tot} and number N_{tot} of particles. We can apply the microcanonical approach to this system, computing its microcanonical partition function $\Omega(E_{tot}, N_{tot}, V_{tot})$ as explained previously. Now let us divide $\mathcal{S}_{tot}$ into two subsystems: a small subsystem $\mathcal{S}$ on which we will focus, and its surroundings $\mathcal{S}_{bath}$ within $\mathcal{S}_{tot}$, usually called the bath. Regardless of the micro-states assumed by $\mathcal{S}_{bath}$ and $\mathcal{S}$, $\mathcal{S}_{tot}$ being isolated ensures that:

$$E_{\mathcal{S}} + E_{bath} = E_{tot} \tag{A.67}$$

Importantly, although E_{tot} is constant, $E_{\mathcal{S}}$ and E_{bath} are free to fluctuate. We are now asking: what is the probability P_l for $\mathcal{S}$ to be found in a given microstate l of energy ε_l?

By the phase equiprobability principle, all configurations of the full system compatible with total energy E_{tot} have the same probability. Thus, it is enough to enumerate the number of micro-states $\Omega_{tot}(E_{tot}|\text{state of } \mathcal{S} = l)$, *i.e.* the number of micro-states of the full system such that the sub-system $\mathcal{S}$ is in micro-state l. This number divided by the total number of accessible micro-states gives the sought after probability:

$$P_l = \frac{\Omega_{tot}(E_{tot}|\text{state of } \mathcal{S} = l)}{\Omega(E_{tot})} \tag{A.68}$$

Since the state of $\mathcal{S}$ is specified, one has to account only for the undetermined micro-state of the bath, which has energy $E_{tot} - \varepsilon_l$:

$$\Omega_{tot}(E_{tot}|\text{state of } \mathcal{S} = l) = \Omega_{bath}(E_{tot} - \varepsilon_l) \tag{A.69}$$

Thus:

$$P_l \propto \Omega_{bath}(E_{tot} - \varepsilon_l) \tag{A.70}$$

Using Boltzmann's formula (equation A.63) we switch to entropy:

$$P_l \propto e^{-\frac{1}{k_B} S_{bath}(E_{tot} - \varepsilon_l)} \tag{A.71}$$

Since we assumed that the size of $\mathcal{S}$ is small as compared to the bath, $\varepsilon_l \ll E_{tot}$. We take the first-order Taylor expansion of the bath entropy:

$$S_{bath}(E_{tot} - \varepsilon_l) = S_{bath}(E_{tot}) - \frac{\partial S_{bath}}{\partial E}\varepsilon_l + o(\varepsilon_l) \tag{A.72}$$

and, by definition of the microcanonical temperature:

$$S_{bath}(E_{tot} - \varepsilon_l) = S_{bath}(E_{tot}) - \frac{1}{T}\varepsilon_l + o(\varepsilon_l) \tag{A.73}$$

where $T = T_{bath}$ is the microcanonical temperature of the bath. Plugging equation A.73 into equation A.71 yields:

$$P_l = Ce^{-\frac{\varepsilon_l}{k_B T}} = Ce^{-\beta\varepsilon_l} \tag{A.74}$$

where C is a multiplicative constant and we have introduced the inverse temperature $\beta = 1/k_B T$. We now evaluate C by requiring that P_l, as a probability, be normalized to 1 upon summation over all states: $\sum_l P_l = 1 \Rightarrow C^{-1} = \sum_l e^{-\beta\varepsilon_l}$. C^{-1}, usually denoted as Q, is called the *canonical partition function* and depends on the temperature (along with the volume and number of particles). We arrive at the canonical distribution:

$$\boxed{P_l = \frac{e^{-\beta\varepsilon_l}}{\sum_{l'} e^{-\beta\varepsilon_{l'}}}} \tag{A.75}$$

A.3.2.1. Classical case - phase space integrals

So far, we have reasoned using discrete sums over micro-states. This may be fully justified in quantum mechanics, where micro-states can be taken as (discrete) eigenstates of the Hamiltonian operator, but should be adapted to account for classical situations. This requires some care. It is natural to establish a correspondence between the pair (l, E_l) (*i.e.* a discrete micro-state and its energy) to $((p,q), \mathcal{H}(p,q))$ (*i.e.* a point in phase-space and its energy, given by the value of the Hamiltonian function at this point). Since the momenta p and positions q are continuous variables, the sum over-states should be turned into a phase-space integral:

$$\sum_l e^{-\beta E_l} \to \int \mathrm{d}p\mathrm{d}q e^{-\beta\mathcal{H}(p,q)} \tag{A.76}$$

Following the same line of reasoning as for the microcanonical ensemble (Appendix, A.3.1), we introduce the volume of an elementary cell in phase-space h^{3N} to arrive at the classical canonical partition function:

$$Q_{cl}(\beta) = \frac{1}{h^{3N}} \int \mathrm{d}p\mathrm{d}q e^{-\beta\mathcal{H}(p,q)} \tag{A.77}$$

We note that in the case where the particles of the system are indistinguishable (*e.g.* in the case of the ideal gas), a pre-factor $1/N!$ should be added in front of equation A.77; indeed, indistinguishability implies that the current micro-state is left unchanged by the permutation of two particles. The contribution of this pre-factor usually drops upon considering free energy differences, and it is ignored in this thesis.

A.3.3. Ensemblist view, ergodic hypothesis and justification of the statistical approach: a short discussion

Considering any macroscopic system, *e.g.* a gaz enclosed in a recipient, the number of microscopic degrees of freedom is of order of the Avogadro number; even if there were a way for the observer to know all the positions and momenta at a given time, it would be impossible to integrate the equations of motion either analytically or numerically. Both the imperfect knowledge of the microscopic configuration and the impossibility to solve the equations of motion motivate the introduction of a sta-

tistical approach, called statistical mechanics. Indeed, intuitively, the macroscopic, observable state of a system is obtained as an average over its accessible micro-states.

This entails two equally important aspects. First, the very notion of average implies the existence of a probability distribution with respect to which the average is computed - the function ρ already discussed above. Thus, statistical mechanics should be concerned with providing an appropriate conceptual framework in which this distribution can be properly defined. This framework is the so-called *ensemblist* view, generally attributed to Gibbs. Second, independently of the mathematics, there must exist a physical phenomenon through which a system can actually switch between available micro-states, effectively realizing the calculation of the average that would represent the result of a macroscopic measurement. Intuitively, this phenomenon is thermal agitation; because of temperature, a system comprising a large number of molecules will sample its accessible microstates. Depending on the external constraints applied on the system, the occupancy probability of a given micro-state changes. Statistical mechanics is concerned with providing ways to compute these probabilities. By its predictive power and consistency with the older results of thermodynamics, there is no doubt left that statistical mechanics *works*, that is, can make testable predictions subsequently validated by experiment. However, there is a fundamental discrepancy in the usage of probabilistic tools to describe deterministic processes; the question as to *why* statistical mechanics works is not yet settled (Castiglione et al. 2008). In the following we give a brief overview of the existing lines of thought regarding this matter, with no particular regard for mathematical rigour.

In the ensemblist view, one imagines that M copies of the system of interest are prepared in such a way that macroscopic constraints (*e.g.* fixed volume V) are satisfied, but that nothing is known on the actual microscopic states of each system replica. This collection of replicas forms a *statistical ensemble*, and the limit $M \to +\infty$ is considered. To obtain the value of any observable A, a simple arithmetic average is used:

$$\langle A \rangle = \lim_{M \to +\infty} \frac{1}{M} \sum_{i=1}^{M} A_i \tag{A.78}$$

where A_i is the value of A in replica i. Notably (in the classical case), the probability density in phase-space $\rho(p, q, t)$ (with t the time) is similarly defined; for an infinitesimal phase space volume $\mathrm{d}p\mathrm{d}q$, $\rho(p, q, t)\mathrm{d}p\mathrm{d}q$ is the fraction of the M systems in the micro-state (p, q) up to $(\mathrm{d}p, \mathrm{d}q)$ at time t. At equilibrium, $\rho(p, q, t)$ converges to a stationary distribution $\rho_{eq}(p, q)$. In this case, $\langle A \rangle$ is rewritten:

$$\langle A \rangle = \int \rho_{eq}(p, q) A(p, q) \mathrm{d}p\mathrm{d}q \tag{A.79}$$

The ensemblist view provides a convenient way of introducing a statistical approach, by the means of a mental construction which allows the use of probabilities through the law of large numbers (equation A.78). However, it is *a priori* unclear how this approach is applicable to real systems. Indeed, experimental measurements correspond to time-averages over a single system, rather than ensemble averages over a collection thereof. This discrepancy is solved by the *ergodic hypothesis*, which states that time-averages and ensemble-averages are assumed to be equal:

$$\int \rho_{eq}(p, q) A(p, q) \mathrm{d}p\mathrm{d}q = \lim_{t \to +\infty} \frac{1}{t} \int \mathrm{d}t' A(p(t'), q(t')) \tag{A.80}$$

where $(p(t), q(t))$ are solutions of the Hamiltonian equations of motion A.37.

The study of the significance and applicability of the ergodic hypothesis sparked an entire research

domain mathematics, ergodic theory, which can be seen as a sub-field of dynamical systems and measure theory (Ashley 2015; Moore 2015). The ergodic theorem, proved in the 1930s (Moore 2015), establishes the conditions for equation A.80 to hold. The main condition is called *metric indecomposability of phase-space* and, intuitively, expresses that any two regions in phase-space must be accessible from one another under the dynamics. Proving this property for a given Hamiltonian dynamics is in general not doable; rather, it is generally assumed. The reason generally advanced is that Hamiltonian dynamics is chaotic (for interacting systems with a high number of elements). In this context, chaotic means sensitivity to initial conditions (or Lyapunov instability): the difference between two trajectories which start from two very close points in phase space increases exponentially. Thus, under the ergodic hypothesis, statistical mechanics works because of particular properties of the *dynamics* of the system. Interestingly, this suggests that the statistical approach may not be limited to systems with large numbers of degrees of freedom; in fact, there are known cases of low dimensional deterministic dynamical systems for which a non-singular stationary probability density can be found, like the logistic map (Sternberg 2010). Also, recent theoretical work has shown that the assumption of a particular case of chaotic dynamics, the so-called *chaotic hypothesis*, is sufficient to derive many important results of non-equilibrium statistical mechanics, suggesting that the assumption of chaos is an appropriate foundation for statistical mechanics (Gallavotti 1998).

By contrast, another justification to statistical mechanics, notably defended by Khinchin, puts the emphasis on the large number of degrees of freedom (Khinchin 1949). The argument is essentially that any "relevant" function of the $6N$ degrees of freedom will take nearly constant values everywhere on the constant-energy hypersurface, because the relative fluctuations go to 0 when $N \to +\infty$. If this is the case, phase-space averages and time-averages coincide without assumption on the dynamics. Khinchin proved this result for the case of system of non-interacting particles (it was later extended to the weakly interacting case). However, the assumption of non-interacting particles seems quite strong; if it is made, it amounts to claiming that the actual system itself behaves as a statistical ensemble (because the non-interacting elements can be seen as the non-interacting replicas of the system which make up the ensemble).

To summarize, it is still unclear as to whether statistical mechanics is valid due to the large number of degrees of freedom or due to the particular properties of the microscopic dynamics. Note that we have barely touched upon the depth of the debate; the interested reader may for example refer to (Castiglione et al. 2008, and references therein).

Implications for Molecular Dynamics simulations

In an isolated, classical system of N particles the dynamics is prescribed by the Hamiltonian $\mathcal{H}(q, p)$ which generates the motion equations, as explained above (A.2). In practical situations, the interaction potential which defines the potential energy in the Hamiltonian makes it impossible to obtain an analytical expression for the trajectory, because it is too complex. Consequently, one has to resort to a numerical integration procedure. When applied to a molecular system, this approach is called Molecular Dynamics. Thus, with respect to the physical behaviour of the real system, there are two sources of error: the use of an approximate interaction potential (typically a classical force-field) and the use of a numerical approximation of the dynamics. For our discussion, we will forget that the force-field is not an accurate depiction of reality, and call the "real dynamics" the dynamics generated by the Hamiltonian equations, *i.e.* the *exact* dynamics generated by the force-field.

This real dynamics is confined to the hypersurface of energy $E = \mathcal{H}(q_0, p_0)$ and is time-reversible. Therefore one should seek integration algorithms which reproduce these properties as accurately as

possible. Furthermore, it is desirable for the trajectory obtained by numerical integration to be as close as possible to the real one; however this requirement is arguably less important. The reason is that Molecular Dynamics initially emerged as a *sampling* method to study the thermodynamic properties of systems intractable by analytical methods (Alder and Wainwright 1957, 1959). Historically, the point of MD was to compute averages, rather than produce trajectories. This contrasts for example with celestial mechanics simulations. Just like the case of justifying statistical mechanics as a whole, it may seem counter-intuitive to use a simulation methodology relying on a purely deterministic formalism to evaluate statistical quantities. It is justified, as above, by the chaotic nature of dynamics (Frenkel and Smit 2002). This underlines why it is somewhat hopeless to obtain a good approximation of the real trajectory, as the simulated one will eventually drift away from it as small numerical errors accumulate. In fact, the very reason that makes MD unsuited to accurate prediction of trajectories makes it, paradoxically, a good sampling method despite its deterministic roots.

Practically, in Molecular Dynamics simulations, the constant-energy dynamics is of less importance than the constant-temperature one, because it more closely corresponds to actual laboratory experiments. MD should sample from the canonical distribution, rather than the microcanonical one. From a numerical point of view, this requires a modification of Hamilton's equations (see 3.3.1) such that constant-temperature MD trajectories exhibit the wanted statistical properties. Depending on the method chosen to achieve canonical sampling, one may end up with a completely unrealistic dynamics (Ciccotti and Vanden-Eijnden 2015; Tuckerman 2010). This must be kept in mind when analyzing MD trajectories. Notably, it provides a strong motivation for the use of Langevin dynamics, which can be seen as a small random perturbation of Hamiltonian dynamics and as such, can be expected to yield reasonably realistic trajectories.

A.4. Re-weighting of Accelerated MD simulations

A.4.1. Re-weighting and cumulant expansion

As the boost potential is known, it is in principle possible to reweight the probabilities measured from the accelerated simulation to obtain unbiased canonical probabilities. Considering a collective variable $\xi = \hat{\xi}(x)$ (see also 4.3), the biased probability distribution $p^*(\xi)$ can be directly estimated from the accelerated simulation. Our purpose is to recover the unbiased distribution $p(\xi)$.

A.4.1.1. Boltzmann re-weighting of a collective variable

Anticipating on 4.3, one has:

$$p^*(\xi) = \frac{\int \mathrm{d}x e^{-\beta V^*(x)} \delta(\hat{\xi}(x) - \xi)}{\int \mathrm{d}x e^{-\beta V^*(x)}} \tag{A.81}$$

where x represents the vector of atomic coordinates.

The unbiased probability $p(\xi)$ writes:

$$p(\xi) = \frac{\int \mathrm{d}x e^{-\beta V(x)} \delta(\hat{\xi}(x) - \xi)}{\int \mathrm{d}x e^{-\beta V(x)}} \tag{A.82}$$

or:

$$p(\xi) = \frac{\int \mathrm{d}x e^{-\beta V^*(x)} e^{+\beta \Delta V(x)} \delta(\hat{\xi}(x) - \xi)}{\int \mathrm{d}x e^{-\beta V^*(x)} e^{+\beta \Delta V(x)}} \tag{A.83}$$

which is rewritten as:

$$p(\xi) = \frac{\langle e^{+\beta \Delta V(x)} \delta(\hat{\xi}(x) - \xi) \rangle^*}{\langle e^{+\beta \Delta V(x)} \rangle^*} \tag{A.84}$$

where $\langle ... \rangle^*$ is the canonical average with respect to the aMD potential; $\langle ... \rangle^*$ can be estimated (assuming ergodicity) as a time average along the aMD simulation. Equation A.84 is exact, but of little practical interest. It must be converted to a form usable with data from numerical simulations. To that end, we discretize ξ into M bins of width $\delta\xi$ and introduce the family of indicator functions $\mathbb{I}_j(x)_{j=1,...,M}$. $\mathbb{I}_j(x) = 1$ if $\hat{\xi}(x) \in$ bin j, 0 otherwise. Let us re-express the numerator and denominator of equation A.84 with these new notations.

$$\langle e^{+\beta \Delta V(x)} \delta(\hat{\xi}(x) - \xi) \rangle^* = \frac{1}{N} \sum_x \mathbb{I}_j(x) e^{+\beta \Delta V(x)} \tag{A.85}$$

where the sum is taken on all the frames sampled by the simulation, and N is the total number of frames. We introduce N_j (the number of frames belonging to bin j) and multiply equation A.85 by N_j/N_j:

$$\frac{1}{N} \sum_x \mathbb{I}_j(x) e^{+\beta \Delta V(x)} = \frac{N_j}{N} \cdot \frac{1}{N_j} \sum_x \mathbb{I}_j(x) e^{+\beta \Delta V(x)} \tag{A.86}$$

We recognize that $N_j/N = p^*(\xi_j)$ (assuming ergodicity). Also, we introduced the biased average restricted to bin j, $\langle ... \rangle_j^*$, *i.e.* this average is taken only over the configurations belonging to bin j. Thus equation A.85 rewrites:

$$\langle e^{+\beta \Delta V(x)} \delta(\hat{\xi}(x) - \xi) \rangle^* = p^*(\xi_j) \langle e^{+\beta \Delta V(x)} \rangle_j^* \tag{A.87}$$

We now turn to the denominator of equation A.84.

$$\langle e^{+\beta \Delta V(x)} \rangle^* = \frac{1}{N} \sum_x e^{+\beta \Delta V(x)} \tag{A.88}$$

$$= \frac{1}{N} \sum_{j=1}^{M} \sum_x \mathbb{I}_j(x) e^{+\beta \Delta V(x)} \tag{A.89}$$

$$= \sum_{j=1}^{M} \frac{N_j}{N} \frac{1}{N_j} \sum_x \mathbb{I}_j(x) e^{+\beta \Delta V(x)} \tag{A.90}$$

yielding:

$$\langle e^{+\beta \Delta V(x)} \rangle^* = \sum_{j=1}^{M} p^*(\xi_j) \langle e^{+\beta \Delta V(x)} \rangle_j^* \tag{A.91}$$

Combining equations A.87 and A.91 we get the reweighting formula:

$$p(\xi_j) = \frac{p^*(\xi_j)\langle e^{+\beta\Delta V(x)}\rangle_j^*}{\sum_{j=1}^{M} p^*(\xi_j)\langle e^{+\beta\Delta V(x)}\rangle_j^*} \tag{A.92}$$

All terms of equation A.92 can be estimated from the aMD simulation. However, it does not work well in practice, as we now discuss.

A.4.1.2. Energetic noise and cumulant expansion

The Boltzmann factors $e^{+\beta\Delta V(x)}$ which appear in equation A.92 are problematic when the boost potential is large, as taking the exponential of a large number frequently leads to numerical overflow. A number of strategies may be attempted to alleviate this problem, such as using a Taylor expansion of the exponential. According to McCammon and co-workers, the most accurate approach is to use a second order cumulant expansion (Miao, Sinko, et al. 2014). The cumulants of $\beta\Delta V$ (formally treated as a random variable) are the coefficients C_k such that:

$$\langle e^{+\beta\Delta V}\rangle = \sum_{k=1}^{+\infty} \frac{\beta^k}{k!} C_k \tag{A.93}$$

The first cumulants are given by:

$$C_1 = \langle \Delta V\rangle \tag{A.94}$$

$$C_2 = \langle \Delta V^2\rangle - \langle \Delta V\rangle^2 = \sigma_{\Delta V}^2 \tag{A.95}$$

$$C_3 = \langle \Delta V^3\rangle - 3\langle \Delta V^2\rangle\langle \Delta V\rangle + 2\langle \Delta V\rangle^3 \tag{A.96}$$

Second order cumulant expansion is exact if the distribution of ΔV is perfectly Gaussian (this is because a Gaussian has zero cumulants starting at order 3). The anharmonicity γ allows for estimating how different the actual ΔV distribution is from a Gaussian.

$$\gamma = \frac{1}{2}\ln(2\pi e\sigma_{\Delta V}^2) + \int_0^{+\infty} p(\Delta V)\ln p(\Delta V)\mathrm{d}\Delta V \tag{A.97}$$

γ is simply the difference between the entropy of a perfect Gaussian with same standard deviation, and the actual statistical entropy of the distribution of ΔV. It is 0 if ΔV actually follows a Gaussian distribution. For big systems, this condition is typically not met with aMD, and accurate reweighting is essentially impossible, making aMD a qualitative exploration tool at best. GaMD aims at solving this problem by applying a boost designed to follow a Gaussian distribution to a good approximation, that is, a harmonic boost.

A.5. Derivations of some relations for free energy calculations

A.5.1. Free energy gradient with respect to a collective variable

We follow the derivation as it is outlined in Tuckerman (2010). For a one-dimensional CV $\hat{\xi}$, we want to establish an analytical expression for $F'(\xi)$.

From $F(\xi) = -k_B T \ln P(\xi)$ it comes that:

$$-\beta F'(\xi) = \frac{1}{P(\xi)} \cdot \frac{\mathrm{d}P(\xi)}{\mathrm{d}\xi} = \frac{1}{Z(\xi)} \cdot \frac{\mathrm{d}Z(\xi)}{\mathrm{d}\xi} \tag{A.98}$$

Then:

$$\frac{\mathrm{d}Z(\xi)}{\mathrm{d}\xi} = \int \mathrm{d}x e^{-\beta U(x)} \frac{\partial}{\partial \xi} \delta(\hat{\xi}(x) - \xi) \tag{A.99}$$

We now introduce a full variable change from the $3N$ cartesian coordinates x to a set of $3N$ generalized coordinates q such that $q_1 = \hat{\xi}(x)$. The remaining $3N - 1$ variables are left unspecified. We introduce the Jacobian $J(q)$ of the transformation. Under this variable change, the integral in equation A.99 transforms as follows:

$$\int \mathrm{d}x e^{-\beta U(x)} \frac{\partial}{\partial \xi} \delta(\hat{\xi}(x) - \xi) = \int \mathrm{d}q J(q) e^{-\beta \tilde{U}(q)} \frac{\partial}{\partial \xi} \delta(q_1 - \xi) \tag{A.100}$$

where $\tilde{U}(q(x)) = U(x)$.

For an arbitrary function f, one has $\partial_y f(x-y) = -\partial_x f(x-y)$. Applied to the δ in equation A.100, this yields:

$$\int \mathrm{d}q J(q) e^{-\beta \tilde{U}(q)} \frac{\partial}{\partial \xi} \delta(q_1 - \xi) = -\int \mathrm{d}q J(q) e^{-\beta \tilde{U}(q)} \frac{\partial}{\partial q_1} \delta(q_1 - \xi) \tag{A.101}$$

Next, an integration by part is performed to obtain:

$$-\int \mathrm{d}q J(q) e^{-\beta \tilde{U}(q)} \frac{\partial}{\partial q_1} \delta(q_1 - \xi) = +\int \mathrm{d}q \frac{\partial}{\partial q_1} \left[J(q) e^{-\beta \tilde{U}(q)} \right] \delta(q_1 - \xi) \tag{A.102}$$

We expand the derivative:

$$\frac{\partial}{\partial q_1} \left[J(q) e^{-\beta \tilde{U}(q)} \right] = J(q) e^{-\beta \tilde{U}(q)} \left(\frac{\partial}{\partial q_1} \ln J(q) - \beta \frac{\partial \tilde{U}}{\partial q_1} \right) \tag{A.103}$$

Combining the previous results, we get:

$$\frac{\mathrm{d}Z(\xi)}{\mathrm{d}\xi} = \int \mathrm{d}q \left(\frac{\partial}{\partial q_1} \ln J(q) - \beta \frac{\partial \tilde{U}}{\partial q_1} \right) e^{-\beta \left(\tilde{U}(q) - k_B T \ln J(q) \right)} \delta(q_1 - \xi) \tag{A.104}$$

where the Jacobian has been exponentiated. We finally obtain:

$$\frac{1}{P(\xi)} \cdot \frac{\mathrm{d}P(\xi)}{\mathrm{d}\xi} = \frac{1}{Z(\xi)} \cdot \frac{\mathrm{d}Z(\xi)}{\mathrm{d}\xi} = \frac{\int \mathrm{d}q \left(\frac{\partial}{\partial q_1} \ln J(q) - \beta \frac{\partial \tilde{U}}{\partial q_1} \right) e^{-\beta \left(\tilde{U}(q) - k_B T \ln J(q) \right)} \delta(q_1 - \xi)}{\int \mathrm{d}q e^{-\beta \left(\tilde{U}(q) - k_B T \ln J(q) \right)} \delta(q_1 - \xi)} \tag{A.105}$$

where the $x \mapsto q$ variable change has also been applied in the denominator. Next, we note that for a given observable A, the canonical average must be left invariant by the variable change, which implies that:

$$\langle A \rangle = \frac{\int \mathrm{d}x A(x) e^{-\beta U(x)}}{\int \mathrm{d}x e^{-\beta U(x)}} = \frac{\int \mathrm{d}q J(q) \tilde{A}(q) e^{-\beta \tilde{U}(q)}}{\int \mathrm{d}q J(q) e^{-\beta \tilde{U}(q)}} \tag{A.106}$$

In other words, when A is expressed as a function of q (*i.e.* as $\tilde{A}(q)$), the configurational space average along the q-coordinates with respect to the modified potential $\tilde{U}(q) - k_B T \ln J(q)$ is the canonical average. Taking $\tilde{A} = \left(\frac{\partial}{\partial q_1} \ln J(q) - \beta \frac{\partial \tilde{U}}{\partial q_1} \right) \delta(q_1 - \xi)$, one obtains for the numerator of equation A.105:

$$\int \mathrm{d}q \left(\frac{\partial}{\partial q_1} \ln J(q) - \beta \frac{\partial \tilde{U}}{\partial q_1} \right) e^{-\beta \left(\tilde{U}(q) - k_B T \ln J(q) \right)} \delta(q_1 - \xi) = Z \left\langle \frac{\partial}{\partial q_1} \ln J(q) - \beta \frac{\partial \tilde{U}}{\partial q_1} \delta(q_1 - \xi) \right\rangle \tag{A.107}$$

where $Z = \int \mathrm{d}x e^{-\beta U(x)} = \int \mathrm{d}q J(q) e^{-\beta \tilde{U}(q)}$ is the canonical (configurational) partition function. And, taking $\tilde{A} = \delta(q_1 - \xi)$, the denominator of equation A.105 becomes:

$$\int \mathrm{d}q e^{-\beta \left(\tilde{U}(q) - k_B T \ln J(q) \right)} \delta(q_1 - \xi) = Z \left\langle \delta\left(q_1 - \xi\right) \right\rangle \tag{A.108}$$

Combining A.107 and A.108, we recognize a conditional canonical average, and obtain:

$$\begin{aligned} -\beta F'(\xi) &= \frac{\left\langle \frac{\partial}{\partial q_1} \ln J(q) - \beta \frac{\partial \tilde{U}}{\partial q_1} \delta(q_1 - \xi) \right\rangle}{\left\langle \delta\left(q_1 - \xi\right) \right\rangle} \\ &= \frac{\left\langle \frac{\partial}{\partial q_1} \ln J(q) - \beta \frac{\partial \tilde{U}}{\partial q_1} \middle| q_1 = \xi \right\rangle}{\left\langle \delta\left(q_1 - \xi\right) \right\rangle} \left\langle \delta\left(q_1 - \xi\right) \right\rangle \\ &= \left\langle \frac{\partial}{\partial q_1} \ln J(q) - \beta \frac{\partial \tilde{U}}{\partial q_1} \middle| q_1 = \xi \right\rangle \end{aligned} \tag{A.109}$$

Changing to the notation $\langle ... | q_1 = \xi \rangle = \langle ... \rangle_{\hat{\xi}(x) = q_1 = \xi}$ we finally arrive at:

$$\boxed{-\beta F'(\xi) = \left\langle \frac{\partial}{\partial q_1} \ln J(q) - \beta \frac{\partial \tilde{U}}{\partial q_1} \right\rangle_{\hat{\xi}(x) = q_1 = \xi}} \tag{A.110}$$

Equation 4.38 immediately follows.

A.5.2. Free energy gradient from harmonic restraint

We derive property 4.47. To the best of our knowledge, this proof is first due to Maragliano, A. Fischer, et al. (2006) and Maragliano and Vanden-Eijnden (2006). We follow their approach for the derivation, which has the advantage of also providing the behaviour of the error. For notational simplicity, we set $h_k(\xi) \equiv e^{-\frac{1}{2}\beta k (\hat{\xi}(x) - \xi)^2}$. We introduce the Fourier transform $\tilde{h}_k(\omega)$ of $h_k(\xi)$:

$$\tilde{h}_k(\omega) = \int h_k(\xi) e^{-i\xi\omega} \mathrm{d}\xi \tag{A.111}$$

with i the imaginary unit. Reciprocally, h_k can be written in terms of $\tilde{h}_k(\omega)$ using the inverse Fourier transform:

$$h_k(\xi) = \frac{1}{2\pi} \int \tilde{h}_k(\omega) e^{+i\xi\omega} \mathrm{d}\omega \tag{A.112}$$

We start from the expression of $Z_k(\xi)$ and replace h_k by expression A.112:

$$Z_k(\xi) = \int e^{-\beta U(x)} h_k(\xi) \mathrm{d}x = \int e^{-\beta U(x)} \frac{1}{2\pi} \int \tilde{h}_k(\omega) e^{+i\xi\omega} \mathrm{d}\omega \, \mathrm{d}x \tag{A.113}$$

We notice that $\tilde{h}_k(\omega)$ (equation A.112) is a Gaussian integral[2], and as such can be analytically solved. We have:

$$\tilde{h}_k(\omega) = \int e^{-\frac{1}{2}\beta k(\hat{\xi}(x)-\xi)^2 - i\xi\omega} \mathrm{d}\xi \tag{A.115}$$

and, after expanding $-\frac{1}{2}\beta k(\hat{\xi}(x) - \xi)^2 - i\xi\omega$, we obtain by application of formula A.114:

$$\tilde{h}_k(\omega) = \sqrt{\frac{2\pi}{\beta k}} e^{\frac{(\beta k \hat{\xi}(x) - i\omega)^2}{2\beta k}} e^{-\frac{\beta k}{2}\hat{\xi}(x)^2} \tag{A.116}$$

which yields after simplification:

$$\tilde{h}_k(\omega) = \sqrt{\frac{2\pi}{\beta k}} e^{-i\omega\hat{\xi}(x)} e^{-\frac{\omega^2}{2\beta k}} \tag{A.117}$$

The analytical expression for $\tilde{h}_k(\omega)$ can now be inserted into equation A.113:

$$Z_k(\xi) = \int e^{-\beta U(x)} \frac{1}{2\pi} \int \sqrt{\frac{2\pi}{\beta k}} e^{-i\omega\hat{\xi}(x)} e^{-\frac{\omega^2}{2\beta k}} e^{+i\xi\omega} \mathrm{d}\omega \, \mathrm{d}x \tag{A.118}$$

After simplification:

$$Z_k(\xi) = \sqrt{\frac{1}{2\pi\beta k}} \int e^{-\beta U(x)} \int e^{+i\omega(\xi - \hat{\xi}(x))} e^{-\frac{\omega^2}{2\beta k}} \mathrm{d}\omega \, \mathrm{d}x \tag{A.119}$$

We can now take the Taylor expansion of $e^{-\frac{\omega^2}{2\beta k}}$ in equation A.119:

$$e^{-\frac{\omega^2}{2\beta k}} = \sum_{n>0} \frac{(-1)^n}{n!} \left(\frac{\omega^2}{2\beta k}\right)^n = 1 + \mathcal{O}(\frac{1}{k}) \tag{A.120}$$

which leads after insertion in equation A.119:

$$Z_k(\xi) = \sqrt{\frac{1}{2\pi\beta k}} \int e^{-\beta U(x)} \int e^{+i\omega(\xi - \hat{\xi}(x))} \left(1 + \mathcal{O}(\frac{1}{k})\right) \mathrm{d}\omega \, \mathrm{d}x \tag{A.121}$$

After some manipulations, we recognize the Fourier transform of a Dirac function in the leading term of equation A.121:

2. Given a, b, c constant numbers, one has:

$$\int_{-\infty}^{+\infty} e^{-ax^2 + bx + c} \mathrm{d}x = \sqrt{\frac{\pi}{a}} e^{\frac{b^2}{4a} + c} \tag{A.114}$$

$$\int e^{-i\omega(\hat{\xi}(x)-\xi)}\mathrm{d}\omega = \delta\left(\hat{\xi}(x)-\xi\right) \tag{A.122}$$

After insertion in equation A.121, this leads to:

$$Z_k(\xi) = \sqrt{\frac{1}{2\pi\beta k}}\int e^{-\beta U(x)}\left(\delta(\hat{\xi}(x)-\xi)+\mathcal{O}(\frac{1}{k})\right)\mathrm{d}x \tag{A.123}$$

And so:

$$\boxed{Z_k(\xi) = \sqrt{\frac{1}{2\pi\beta k}}\left(Z(\xi)+\mathcal{O}(\frac{1}{k})\right)} \tag{A.124}$$

We are now in a position to complete the proof:

$$\frac{\mathrm{d}F_k}{\mathrm{d}\xi} = -k_BT\frac{\mathrm{d}}{\mathrm{d}\xi}\ln Z_k(\xi) = -k_BT\frac{\mathrm{d}}{\mathrm{d}\xi}\ln Z + \mathcal{O}(\frac{1}{k}) \tag{A.125}$$

in which the prefactor $\sqrt{\frac{1}{2\pi\beta k}}$ drops upon differentiation as it does not depend on ξ; the remaining equation reads:

$$\boxed{\frac{\mathrm{d}F_k}{\mathrm{d}\xi} = \frac{\mathrm{d}F}{\mathrm{d}\xi}+\mathcal{O}(\frac{1}{k})} \tag{A.126}$$

which implies property 4.47. In addition, we find that the error due to the finiteness of k scales as $1/k$. In the Appendix of Maragliano, A. Fischer, et al. (2006), more precise formulas for the error are derived; the reader is referred to this publication for more details.

A.5.3. Derivation of the CZAR estimator

We derive the CZAR estimator for eABF calculations following (Lesage et al. 2017).

From the mollified free energy $F_k(\lambda)$ (equation 4.66), we introduce the canonical distribution $P_k(\lambda)$ in the extended potential:

$$P_k(\lambda) = \frac{1}{Q_k}\int \mathrm{d}x e^{-\beta U(x)}e^{-\frac{1}{2}\beta k\left(\hat{\xi}(x)-\lambda\right)^2} \tag{A.127}$$

with:

$$Q_k \equiv \int \mathrm{d}x\mathrm{d}\lambda e^{-\beta U(x)}e^{-\frac{1}{2}\beta k\left(\hat{\xi}(x)-\lambda\right)^2} \tag{A.128}$$

As λ is a *bona fide* degree of freedom, this marginal probability distribution is obtained by partial integration without the need of a δ function. We may nonetheless introduce $\int \mathrm{d}z\delta\left(\hat{\xi}(x)-z\right)=1$ in equation A.127, to obtain:

$$P_k(\lambda) = \frac{1}{Q_k}\int \mathrm{d}x\mathrm{d}z e^{-\beta U(x)}e^{-\frac{1}{2}\beta k\left(\hat{\xi}(x)-\lambda\right)^2}\delta\left(\hat{\xi}(x)-z\right) \tag{A.129}$$

Up to a normalization constant, the term $\int \mathrm{d}x e^{-\beta U(x)} \delta\left(\hat{\xi}(x) - z\right)$ in equation A.129 is the *unbiased* canonical distribution $P(z)$ of z (see also equation 4.68). As such, equation A.127 is rewritten as:

$$P_k(\lambda) \propto \int P(z) e^{-\frac{1}{2}\beta k(z-\lambda)^2} \mathrm{d}z \tag{A.130}$$

So, $P_k(\lambda)$ (the probability distribution/free energy profile obtained from the eABF run at finite k) is given by the convolution of $P(z)$ (the distribution/free energy profile of interest) and a Gaussian kernel of variance $1/\beta k$ (Lesage et al. 2017).

We introduce the un-normalized joined distribution $P_k(\lambda, z)$ by removing the integral in A.130:

$$P_k(\lambda, z) \propto P(z) e^{-\frac{1}{2}\beta k(z-\lambda)^2} \tag{A.131}$$

Note that at this stage, we have considered only extended dynamics, rather than eABF dynamics. We now turn to this case. Assuming that the eABF bias has reached full convergence, a biased canonical distribution is obtained from the potential 4.76:

$$\tilde{P}_k(x, \lambda) = \frac{1}{\tilde{Q}_k} \exp -\beta \left(U(q) + \frac{1}{2} k(\hat{\xi}(x) - \lambda)^2 - A(\lambda) \right) \tag{A.132}$$

where the tilde refers to quantities biased by the eABF bias $-A(\lambda)$, and $\tilde{Q}_k$ is a normalization factor. By similar manipulations as before in the unbiased case (derivation of equation A.131), we arrive at the biased joint distribution $\tilde{P}_k(\lambda, z)$ such that:

$$\tilde{P}_k(\lambda, z) \propto P(z) e^{-\frac{1}{2}\beta k(z-\lambda)^2} e^{+\beta A(\lambda)} \tag{A.133}$$

Taking the logarithm of A.133 and rearranging yields:

$$\ln P(z) = \ln \tilde{P}_k(\lambda, z) - \beta A(\lambda) + \frac{1}{2} k(z-\lambda)^2 + cste \tag{A.134}$$

and, recalling that $F(z) = -k_B T \ln P(z) + cste$ (which is the PMF we are looking for), we arrive at:

$$F(z) = -k_B T \ln \tilde{P}_k(\lambda, z) + A(\lambda) - \frac{k}{2}(z-\lambda)^2 + cste \tag{A.135}$$

Now, we rewrite $\tilde{P}_k(\lambda, z)$ using a conditional probability:

$$\tilde{P}_k(\lambda, z) = \tilde{P}_k(\lambda | z) \cdot \tilde{P}(z) \tag{A.136}$$

where:

$$\tilde{P}(z) \propto P(z) \int \mathrm{d}\lambda e^{-\frac{1}{2}\beta k(z-\lambda)^2} e^{+\beta A(\lambda)} \tag{A.137}$$

Differentiating A.135 with respect to z yields:

$$F'(z) = -k_B T \frac{\mathrm{d} \ln \tilde{P}_k(\lambda|z)}{\mathrm{d}z} - k_B T \frac{\mathrm{d} \ln \tilde{P}(z)}{\mathrm{d}z} + k(\lambda - z) \tag{A.138}$$

Multiplying equation A.138 by $\tilde{P}_k(\lambda|z)$ yields:

$$F'(z)\tilde{P}_k(\lambda|z) = -k_BT\frac{\mathrm{d}\tilde{P}_k(\lambda|z)}{\mathrm{d}z} - k_BT\frac{\mathrm{d}\ln\tilde{P}(z)}{\mathrm{d}z}\tilde{P}_k(\lambda|z) + k(\lambda - z)\tilde{P}_k(\lambda|z) \tag{A.139}$$

An integration over λ can now be performed, recalling that $\int \mathrm{d}\lambda\tilde{P}_k(\lambda|z) = 1$ by normalization:

$$F'(z) = -k_BT\int\frac{\mathrm{d}\tilde{P}_k(\lambda|z)}{\mathrm{d}z}\mathrm{d}\lambda - k_BT\frac{\mathrm{d}\ln\tilde{P}(z)}{\mathrm{d}z} + \int k(\lambda - z)\tilde{P}_k(\lambda|z)\mathrm{d}\lambda \tag{A.140}$$

Under technical conditions, which we assume are met, one may write:

$$\int\frac{\mathrm{d}\tilde{P}_k(\lambda|z)}{\mathrm{d}z}\mathrm{d}\lambda = \frac{\mathrm{d}}{\mathrm{d}z}\int\tilde{P}_k(\lambda|z)\mathrm{d}\lambda = 0. \tag{A.141}$$

such that we finally arrive at the CZAR estimator:

$$F'(z) = -\frac{1}{\beta}\frac{\mathrm{d}\ln\tilde{P}(z)}{\mathrm{d}z} + k\left(\langle\lambda\rangle_z - z\right) \tag{A.142}$$

B. String method study of a prototypical molecular switch

As another application of the string method, we studied the force-producing conformational transition of a rotaxane-based molecular switch initially synthesized by the Stoddart group (Wu et al. 2008). The switch is made of two intertwined rotaxane subunits which can slide past each other in a piston-like fashion; however, the sliding is blocked by on-axle blocker groups in the form of phenyl rings which sterically clash with the inner ring atoms. As such, there is no free sliding; rather the assembly is a *molecular switch* which can exist in two conformations, an extended and a contracted one (see Figures B.1 and B.2). This architecture, called a [c2]-daisy chain, is thus reminiscent of a molecular muscle. As showed in (Wu et al. 2008), at neutral pH the contracted conformation is favoured. However, the extension motion can be actuated by acidifying the solution, in such a way that the outermost station gets protonated. In these conditions, the extended form becomes thermodynamically favoured thanks to the strong Coulombic attraction of the electronegative ether oxygen atoms for the positive charge.

In the following, we report on a preliminary string method study of the sliding transition in neutral pH conditions, as an opportunity to perform a computational exploration of a simpler molecular machine than myosin. These results are unpublished.

B.1. Model construction

We obtained parameters for the system using the CGENFF automated parameter assignment procedure (Vanommeslaeghe et al. 2010). The molecular model was built manually using Avogadro (Hanwell et al. 2012) and CHARMM. After a brief energy minimization *in vacuo*, the model was placed in an orthorhombic box of acetonitrile molecules; 6 PF_6^- ions were added to mimic experimental conditions. A pre-equilibrated acetonitrile box was kindly provided by Joel Montalvo-Acosta. The parameters for the PF_6^- ions were adapted from (Morrow and Maginn 2002). As compared to this publication, we used identical partial charges for all the fluorine atoms; also, we observed that it was necessary to change the reference values of the F-P-F angles from 90° to 180° for the ions to have a stable geometry in MD simulations. We acknowledge that no significant effort was undertaken to ensure that the used parameters quantitatively reproduce the behaviour of the real system. Although this certainly challenges any validation of the results by comparison with experiments, it is not crucial for our purpose here, because we expect that the general features of the system (*e.g.* bistability) will be preserved.

Using a protocol similar to that reported for myosin (*i.e.* heating then equilibration under restraints, see Chapter 6), equilibrated structures of both the contracted and extended forms were prepared. Unbiased MD simulations initiated from either of these conformations showed an absence of spontaneous transitions in about 100 ns, illustrating their kinetic stability on this timescale (data not shown). We thus resorted to the string method to investigate the transition from contracted to extended.

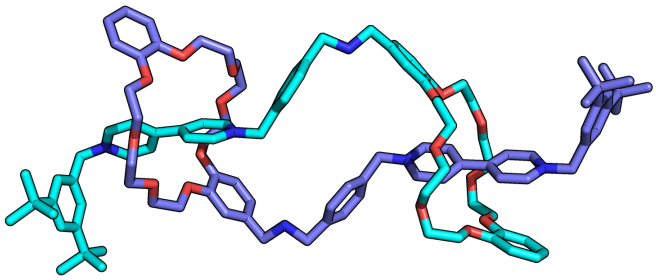


Figure B.1.: Switchable rotaxane-based [c2]-daisy chain synthesized and characterized by Stoddart's group.

B.2. Collective variables for the description of the conformational change

We introduced 4 observables to characterize the conformation of the system. The *extension* L was defined as the distance between the central carbon atoms of the two terminal groups. The inner distance l was defined as the distance between the centers of geometry of the two ether rings. Finally, to characterize local sliding of each ring along its axle, we drew inspiration from (Liu, Chipot, et al. 2014) and introduced observables P_1 and P_2, defined as the projection of the center of geometry of the ring on the vector formed by the two "axial" carbon atoms of the bulky stopper groups. In other words, treating each axle as a proper geometric axis, P_i represents the position of the corresponding ring along this axis. Furthermore, P_1 and P_2 exhibit the following interesting property: by construction, they take an approximately 0 value when the ring is located on the stopper group - which arguably represents the highest free energy barrier to ring sliding. So, these observables take the value 0 on the barriers.

B.3. String optimization

A guess path was generated by a 1 ns SMD simulation along P_1 and P_2 starting from the extended configuration. Then, 32 frames equally spaced in time were extracted from the trajectory and the corresponding values of the supporting CVs L, l, P_1 and P_2 were gathered to obtain a non-reparametrized guess string. This string was then reparametrized without normalizing the CVs by their total variation, as they all are distances expressed in angstrom with comparable total variation. After a 1 ns pre-equilibration (*i.e.* harmonically restrained run), the string method with swarms of trajectories was used to relax the guess path towards a locally minimum free energy path. For each iteration, and for

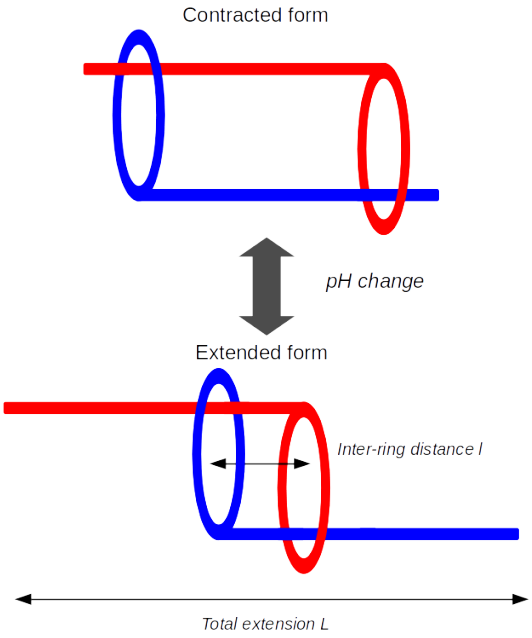


Figure B.2.: Schematic depiction of the relative ring sliding and definition of the collective variables l and L.

each of the 32 images along the string, a 1 ps harmonically restrained run was followed by a swarm of 10 500 fs runs initiated from the final configuration from the harmonically restrained run. Then, the string was evolved following the average drift per CV over the swarm, reparametrized, and linearly smoothed with a 0.1 smoothing parameter. The ends of the string were left free to relax. 571 iterations of this string method protocol were performed. For all simulations (SMD, pre-equilibration and string iterations), a 1 fs timestep was used (with no rigid bonds), along with a 300 K temperature fixed by a Langevin thermostat of damping 0.5 ps^{-1} and a 1 bar pressure fixed by a Berendsen barostat; we took care to use the compressibility of acetonitrile (8.17×10^{-5} bar^{-1}). When applicable, a 4.0 $kcal/mol/Å^2$ force-constant was used for all collective variables.

After roughly 500 iterations the string attained apparent convergence (Figure B.3) despite some residual fluctuations; to remove them, we computed the average string over the last 60 iterations. This average string, simply termed final string and presented in Figure B.4, will be used for the mechanistic

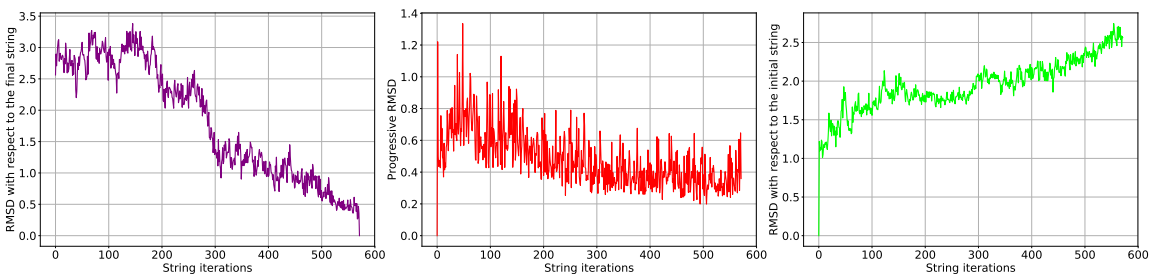


Figure B.3.: Convergence behaviour of the string. Left panel, RMSD of the string with respect to the last iteration. Middle panel, Progressive RMSD, *i.e.* RMSD with respect to the previous iteration; Right panel, RMSD with respect to the initial string.

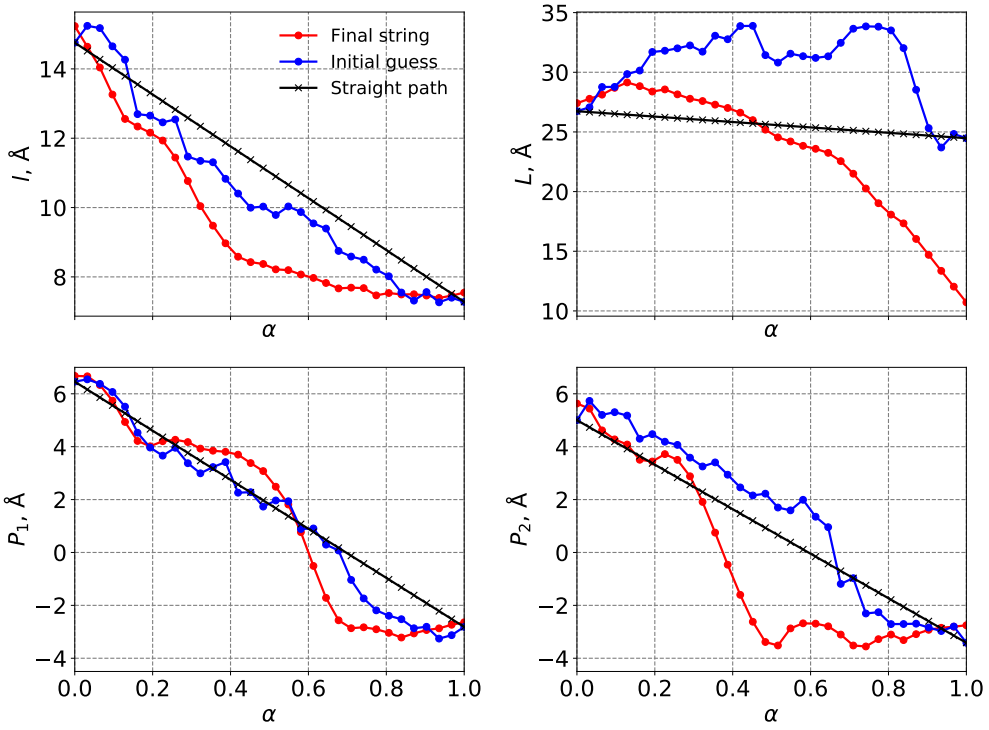


Figure B.4.: Progress along the string of the 4 supporting collective variables.

analysis discussed below.

B.4. ABF calculations

As a complement to the string optimization, we performed two-dimensional ABF calculations to map the free energy landscape along P_1 and P_2. Conventional ABF was used with a 1 fs timestep; P_1 and P_2 are defined with non-overlapping sets of atoms and can be used in an ABF calculation. A non-stratified simulation initiated from the extended configuration was run for 191 ns. The *fullSamples* parameter was set to 200 and a 0.5 Å-spaced grid was used, with each transition coordinate ranging from -8 Å to 8 Å. The resulting free energy landscape (along with its comparison to the final string) is presented on Figure B.5 (right panel). For reasons that will be discussed shortly, it is clear that this PMF is not converged, although it exhibits interesting and probably robust qualitative features. Consequently, no error analysis was performed.

B.5. Results and discussion: Asymmetric mechanism for ring sliding

The system exhibits a rotational symmetry as the two subunits are identical[1]. Consequently, one may imagine two limiting cases for the extended → contracted transition. The concerted, or symmetric pathway would involve the simultaneous inward sliding of both rings; by contrast, in the asymmetric mechanism, one ring would slide and cross the barrier before the other. As the two subunits are iden-

1. The model of the system was actually constructed in CHARMM by applying a rotation to one of the subunits.

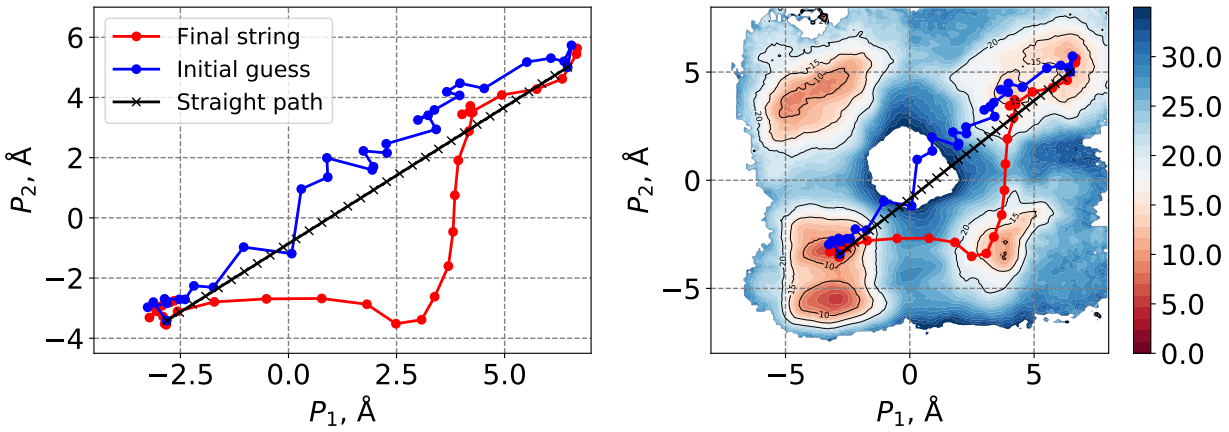


Figure B.5.: Asymmetric sliding mechanism in $(P_1,\ P_2)$ space. Left, projection of the final string, guess path and straight path onto P_1 and P_2. It is seen that the final string strongly deviates from the straight line and explores an asynchronous transition mechanism. Right, projection of the paths onto the free energy landscape computed with ABF along P_1 and P_2. The relaxed path agrees surprisingly well with the metastable basins detected from the ABF analysis.

tical, the concerted mechanism would entail simultaneous crossing of the blocking barriers, which seems less likely than successive crossing events. As we will see, free energy calculations and string method optimizations confirm this intuition. Figure B.5 summarizes the most important results emerging from the string method/ABF study of the system. The left panel shows that the final string indeed predicts an asymmetric sliding in the extended → contracted transition, as P_1 undergoes a nearly full change while the value of P_2 remains constant. One may point out that this asynchronous ring sliding mechanism is reminiscent of the *statistically coupled* picture presented in Chapter 1 (Figure 1.1), and discussed for myosin throughout this thesis. By contrast, the SMD straight path along with the initial guess was closer to a *mechanically coupled* mechanism. Thus, the study of the rotaxane suggests that the string method is indeed able to relax away from the mechanically coupled path, and towards the statistically coupled path, if the latter is indeed of lower free energy than the former. This is an encouraging result, because this is precisely the strategy we are currently using to discriminate between strongly and statistically coupled mechanisms in the recovery stroke of myosin (see Chapter 11).

On the right panel of Figure B.5, the final string is projected onto the (yet un-converged) ABF free energy landscape. As expected considering the symmetry of the system (and of P_1 and P_2), the free energy landscape is globally symmetric with respect to the main diagonal ($P_1 = P_2$). Strikingly, 4 free energy basins are detected: a basin around $P_1 = P_2 = 5$ Å corresponding to the extended state; a basin around $P_1 = P_2 = -4$ Å corresponding to the contracted state; and two off-diagonal, symmetrical basins ($P_1 \simeq 5$ Å, $P_2 \simeq -4$ Å and conversely), each corresponding to intermediate states in which only one ring has slid. Remarkably, the final string explores the sub-diagonal intermediate basin ($P_1 \simeq 5$ Å, $P_2 \simeq -4$ Å); the consistency between two independent sets of calculations supports the validity of our results, at least from a qualitative point of view. Nevertheless, the lack of perfect symmetry of the free energy landscape points to its imperfect convergence, and suggests that longer calculations, possibly using a stratification strategy, should be used.

Also, it seems that both the extended and contracted configurations each actually encompass at least two sub-basins separated by smaller barriers. The molecular interpretation of these substates is

unclear and deserves further investigation; we note that the transition between the two sub-basins of the extended state (upper-right corner of Figure B.5 right) is also explored by the final string, suggesting it is not an artifact due to the imperfect convergence of the ABF calculation. Interestingly, another study of the same system using quantum methods reported on a folding/unfolding transition of the terminal stoppers during the sliding, which may explain the existence of the observed substates (Zhao et al. 2015). However, to our knowledge, these investigators did not report on the asymmetric sliding mechanism.

More generally, free energy calculations have been used in the past to study the functional mechanisms of artificial molecular machines (see for instance Liu, Chipot, et al. 2014; Liu, Shao, and Cai 2015; Raiteri et al. 2008). Notably, a recent study highlighted the complexity of the sliding movement of a rotaxane along a single axle, revealing an unexpected rotation motion associated with the translation (Liu, Shao, Chipot, et al. 2016). In our case, the mechanical interlocking of the two sub-units prevents such a rotation motion from taking place. Nevertheless, our findings are in line with the emerging consensus that the transition mechanisms in artificial molecular machines can be surprisingly more complex than anticipated by chemical intuition. Proper parameter validation, longer ABF calculations with a proper error analysis, and possibly independent replicates of the string calculations, are now required to further this promising preliminary analysis.

C. Manuscript of ”An intermediate of the recovery stroke of myosin VI revealed by X-ray crystallography and molecular dynamics”

C.1. Main Text

Classification: BIOLOGICAL SCIENCES: Biophysics and Computational Biology

An intermediate along the recovery stroke of myosin VI revealed by X-ray crystallography and molecular dynamics

Florian Blanc[1,2,4,5], Tatiana Isabet[1,2], Hannah Benisty[1,2], H. Lee Sweeney[3], Marco Cecchini[4,5,&] and Anne Houdusse[1,2,&]

Affiliations:

[1]Structural Motility, Institut Curie, PSL Research University, CNRS, UMR 144, F-75005 Paris, France

[2]Sorbonne Universités, UPMC University Paris 06, CNRS, UMR 144, Paris, France

[3]Department of Pharmacology & Therapeutics and the Myology Institute, University of Florida College of Medicine, PO Box 100267, Gainesville, FL 32610-0267, USA

[4]ISIS, UMR 7006 CNRS, Université de Strasbourg, F-67083 Strasbourg Cedex, France

[5]Institut de Chimie de Strasbourg, UMR 7177 CNRS, Université de Strasbourg, F-67083 Strasbourg Cedex, France

[&]**Correspondence to :**

Anne Houdusse – Institut Curie CNRS, UMR144, 26 rue d'Ulm, 75248 Paris cedex 05, France. Tel : 33-(0)1-56-24-63-95 – Fax: 33-(0)1-56-24-63-82 – E-mail: anne.houdusse@curie.fr

Marco Cecchini – Institut de Chimie de Strasbourg, UMR 7177 CNRS, Université de Strasbourg, 4 Rue Blaise Pascal, F-67083 Strasbourg Cedex, France. Tel : 33-(0)3-68-58-51-25 – Email: mcecchini@unistra.fr

KEYWORDS: Myosin, Recovery Stroke, Molecular Motor, Large conformational change, Molecular dynamics, Free energy calculations

Manuscript information:

Text pages: 19

Number of Figures: 4

Number of Tables : 0 –

Supplemental Information: Texts: 4, Figures: 15, Tables: 4, Movies: 0

ABSTRACT

Myosins form a class of actin-based, ATPase motor proteins that mediate important cellular functions such as cargo transport and cell motility. Their functional cycle involves two large-scale swings of the lever arm: the force-generating powerstroke, which takes place on actin, and the recovery stroke during which the lever arm is reprimed into an armed configuration. Previous analyses of the pre-recovery (post-rigor) and post-recovery (pre-powerstroke) states predicted that closure of switch II in the ATP binding site precedes the movement of the converter and the lever arm. Here, we report on a crystal structure of myosin VI, called Pre-Transition State (PTS), which was solved at 2.2 Å resolution. Structural analysis and all-atom Molecular Dynamics are consistent with PTS being an intermediate along the recovery stroke, where the Relay/SH1 elements adopt a post-recovery conformation and switch II remains open. In this state, the converter appears to be largely uncoupled from the motor domain and explores an ensemble of partially reprimed configurations through extensive, reversible fluctuations. Moreover, we found that the free energy cost of hydrogen-bonding switch II to ATP is lowered by more than 10 kcal/mol compared to the pre-recovery state. These results support the conclusion that closing of switch II does not initiate the recovery stroke transition in myosin VI. Rather, they suggest a mechanism in which lever arm repriming would be mostly driven by thermal fluctuations and eventually stabilized by the switch II interaction with the nucleotide in a ratchet-like fashion.

SIGNIFICANCE STATEMENT

Myosins are motor proteins involved in the transport of cellular cargoes and muscle contraction. Upon interaction with actin, the motor domain undergoes a conformational transition, called powerstroke, in which the lever arm is swung to generate force and directional motion. The recovery stroke re-primes the motor by coupling the reverse swing of the lever arm to ATP hydrolysis. Using X-ray crystallography and molecular simulations, we characterize a putative intermediate along the recovery stroke of myosin VI, which challenges existing models of myosin chemomechanical transduction. Intriguingly, the new structure suggests that the repriming of the lever arm would be uncoupled from ATPase activity until the very end of the recovery stroke and mostly driven by thermal fluctuations.

\body

INTRODUCTION

Myosins are a wide superfamily of molecular motor proteins involved in a number of vital processes as diverse as intracellular cargo transport, endocytosis, muscle contraction and cell motility (1). Defective myosins were found to be implicated in severe pathologies in humans such as hypertrophic cardiomyopathy (2) and deafness (3), while others including myosin VI were shown to have a role in cancer cell proliferation and metastasis (4). Recent studies highlighted the therapeutic potential of small-molecule inhibitors (5) and activators (6–8) targeting myosin, demonstrating that a detailed knowledge of the force-production mechanism in this motor family would facilitate the rational design of drug candidates.

Myosin motors work through a complex cycle of conformational transitions that couple ATP hydrolysis with force production on actin (see Figure 1). Previous analyses characterized the conformational states of the motor domain during the cycle and the kinetics of the transitions between them (reviewed in (9, 10), see also (11–13)). These studies along with measurements of the stroke size are consistent with the swinging lever arm hypothesis, in which the structural changes in the ATP-binding or the actin-binding sites are amplified into a large swing of the extended lever arm region through the rotation of the converter subdomain (14). In this framework, two major events occur: a force-generating step taking place on actin, which corresponds to the large-amplitude swing of the lever arm termed powerstroke, and an off-actin reverse transition called recovery stroke in which the motor and the lever arm return to their primed configuration. This latter is crucial for chemo-mechanical transduction as it couples the repriming of the lever arm with ATP hydrolysis. Also, this step occurs entirely off-actin and therefore represents an interesting target for pharmacological regulation (5).

Early crystallographic studies on *Dictyostelium discoideum* myosin II (Dd Myo2) and other myosins with various ATP-analogs have trapped the motor domain in the pre-recovery (also called Post-Rigor state, PR) and the post-recovery (also called Pre-Powerstroke state, PPS) conformations. Their comparison revealed that key structural changes accompany the reverse swing of the lever arm : 1) closure of the inner cleft via the formation of critical interactions near the active site (e.g. switch II

closure on the γ-phosphate of ATP) and 2) a major conformational change of the flexible connectors between the motor domain and the converter, (i.e. the Relay helix, the Relay loop and SH1 helix). Importantly, the latter rearrangement involves the formation of a kink in the Relay helix. Computational studies started from these high-resolution structures were instrumental for the development of mechanistic models of the transition between the initial PR state and the final PPS state. Based on various computational strategies (15–24), several models were proposed (SI Appendix, Supplementary Text 1). A common feature of these models (with the notable exception of Cui and co-workers (17, 18)) is that switch II closure is presented as the initiating event of the recovery stroke, which triggers the large-amplitude rotation of the converter. Although the specifics of the coupling between switch II closure and converter repriming are still under debate, the most accepted view (first proposed by Fischer and co-workers (15)) is that closing of switch II exerts strain on the Relay helix that bends and kinks in response, driving the converter rotation. Importantly, none of the existing models predicts the occurrence of intermediates where the converter is uncoupled from the motor domain.

Here, we report on the structural and dynamic characterization of a putative intermediate along the recovery stroke of myosin VI by X-ray crystallography and Molecular Dynamics, which we call pre-transition state (PTS); see Figure 1. The structure, solved at 2.2 Å resolution, reveals a configuration of the motor domain in which the Relay/SH1 elements adopt a nearly post-recovery (PPS-like) configuration while switch II is open as in PR. Using molecular simulations, we explore the implications of the PTS structure for the recovery stroke mechanism. Our results indicate that if PTS were on-path to the post-recovery state, switch II closure would occur at the end of the recovery stroke with the lever arm being essentially reprimed by thermal fluctuations. The isolation of the PTS structure thus suggests the existence of statistical, rather than mechanical, coupling between ATP hydrolysis and the backward swing of the lever arm, in contrast with existing models of the recovery stroke.

RESULTS

Overall description of the PTS Myosin VI crystal structure

From extensive crystallization screens, a previously uncharacterized conformation of the motor

domain of myosin VI has been determined at 2.2 Å resolution (SI Appendix, Supplementary Table S1). Crystals of this structural state were produced with the ATP analogue ADP.BeF_x and could not be obtained with the ADP.Pi analogues ADP.VO_4 or ADP.AlF_4. The crystal structure (Figure 2) reveals a conformation of the motor domain that differs significantly from the Post-Rigor (PR, PDB code: 2VAS) and the Pre-Powerstroke (PPS, PDB code: 2V26) states previously reported for myosin VI. Most importantly, although the converter is partially reprimed, the structural features around the nucleotide, in particular switch II, are not in position to promote hydrolysis of ATP (Figure 2B).

In the active site, switch II is slightly shifted towards the “closed” position found in the PPS state but still exhibits the structural features of an “open” state; the distance between the Beryllium atom of the -BeF_x group and the amide nitrogen of G459 (7.0 Å) is too large for hydrogen-bonding, and the critical salt-bridge R205 – E461, which is required to promote ATP hydrolysis (25), is not formed. The U50/L50 actin-binding cleft is wide open and exhibits minimal deviation from the PR conformation (see also SI Appendix, Supplementary Figure S1). However, the position of the converter indicates that the motor is in a partially primed configuration (Figure 2D). Finally, the Relay/SH1 elements strongly resemble the conformation adopted in the PPS structure, most prominently because of the kinked Relay helix (Figure 2C). Since this new structure is compatible with an ATP bound state, it is likely to represent a state that myosin adopts in complex with ATP when the motor is detached from actin. Furthermore, its structural features are consistent with an intermediate state on the way to the hydrolysis-competent PPS state. Since PPS was referred to as the Transition State of myosin hydrolysis, we name this structure of myosin VI the Pre-Transition State or PTS.

However, important differences from PPS still exist. Although the internal RMSD of the Relay-SH1 elements between PTS and PPS is quite small (0.75 Å excluding the Relay loop), structural alignment onto the N-terminal subdomain reveals that these elements undergo a rigid-body motion to complete the recovery stroke. This global movement, which brings the N-terminal region of the Relay helix towards the inside of the nucleotide-binding site, is consistent with the “seesaw” motion originally proposed by Fischer and co-workers (15), see Figure 2B. Also, the converter subdomain adopts an intermediate position between PR and PPS and displays the canonical R-fold, which was observed in the PR and Rigor structures of myosin VI and virtually every crystal structure for other myosin isoforms; i.e. the converter adopts an unconventional P-fold only in the reprimed PPS and Pi

Release structures of myosin VI (11, 26). Interestingly, most of the contacts between the converter and the motor domain in either PR or PPS of myosin VI, are not formed in the PTS structure, suggesting that this latter might represent a "decoupled converter" state (SI Appendix, Supplementary Tables S2 and S3).

In summary, the PTS structure exhibits a mostly open switch II (PR-like) in a motor with nearly rearranged Relay/SH1 elements (PPS-like) and a converter in an intermediate position. These features are consistent with a conformational state of the motor representative of a previously un-described intermediate along the recovery stroke of myosin VI.

Unbiased Molecular Dynamics simulations reveal a dynamic converter in PTS

To explore the significance of the PTS structure, we performed sub-μs Molecular Dynamics simulations (> 1.4 μs of cumulated simulation time) with an explicit treatment of the solvent starting from the PR (2x100 ns, 1x200 ns), PTS (1x306 ns and 2x100 ns) and PPS (3x100 ns with ATP, 2x100 ns with ADP.Pi) structures of myosin VI; see SI Appendix, Supplementary Table S4). The resulting trajectories were analyzed by monitoring structural observables that describe the conformation of the various elements involved in the recovery stroke. The results of the analysis follow.

The projection of the center of geometry of the converter on the plane defined by the two transverse principal axes on the motor domain (defined in SI Appendix, Supplementary Text 2 and Supplementary Figure S2) shows that the PTS converter is highly dynamic and explores a significantly larger volume than in PR or PPS, where it is confined in proximity to the crystallographic position by specific interactions with the N-terminal domain; see Figure 3; SI Appendix, Supplementary Figures S3, S4, S5 and Supplementary Table S2. In the 306 ns PTS simulation, the time series of the longitudinal component of the converter fluctuations shows that from 50 to 70 ns the converter undergoes a spontaneous swing towards a new position that is closer to PPS (Figure 3). After the swing, the converter appears to be as confined as in PR and PPS, although the new position is not equivalent to a PPS state. This is due to the formation of new contacts between the converter and the N-terminal domain, some of them being absent both in the PR and PPS states (SI Appendix, Supplementary Table S2 and Supplementary Figure S4). The new position of the converter is stable for ~ 75 ns, after which the converter unbinds from the N-terminal domain and eventually returns in vicinity of its initial

position after 50 more nanoseconds (SI Appendix, Supplementary Figure S3). In addition, our analysis shows that both the Relay and SH1 helices, which adopt intermediate orientations in PTS (0-40 ns), move towards PPS upon the partial swing of the converter (SI Appendix, Supplementary Figure S6) and return to their initial conformation when the converter moves back to its initial position (SI Appendix, Supplementary Figure S7). Strikingly, no change in the conformational fluctuations of switch II was detected during the simulation of the PTS structure. In fact, the position of switch II remains close to that in PR for the entire trajectory and corresponds to an "open" state; neither the switch II – γ-phosphate interaction nor the critical salt-bridge (R205-E461) are formed (SI Appendix, Supplementary Figure S8). Finally, a transient uncoupling of the converter from the motor domain was captured in one simulation repeat of the PR state (SI Appendix, Supplementary Figure S3). During this event, while the SH1 helix largely re-orients in coordination to the converter movement, the Relay helix does not (SI Appendix, Supplementary Figure S7). This observation suggests that the formation of the kink in the Relay helix, which is characteristic of PTS and the post-recovery stroke states, may be rate limiting in the PR to PTS isomerization.

Overall, both the crystal structures of myosin VI and the corresponding MD consistently support the conclusion that the PTS structure is representative of an intermediate of the recovery stroke in which the converter subdomain is free to explore a wide range of positions through thermal fluctuations and its conformational dynamics is coupled to the Relay/SH1 elements but is (still) uncoupled from the rest of the motor, including switch II.

Free energy calculations highlight a late switch II closure

Switch II closure is a hallmark of the myosin recovery stroke and occurs through the formation of a hydrogen bond between the amide nitrogen of G459 and the γ-phosphate of ATP, and the catalytically essential salt-bridge between E461 (switch II) and R205 (switch I) (25). As described above, these critical interactions are not formed in the PTS crystal structure. To explore the energetics of switch II closure along the recovery stroke of myosin VI, we performed ABF free energy calculations using the critical salt-bridge separation (d_1) and the hydrogen-bonding distance with the γ-phosphate of ATP (d_γ) as reaction coordinates in PR+ATP, PTS+ATP and PPS+ATP states; see SI Appendix. The goal of these calculations was to probe the energetics of closing switch II in the "mean-field" of the rest of the

protein, which depends on the global conformation of the motor domain that is assumed to be stable on the ABF simulation timescale. To ensure convergence of the free energy calculations, a two-step ABF strategy was adopted, which includes a final stratification over 56 non-overlapping windows; see SI Appendix. The completeness of sampling (SI Appendix, Supplementary Figure S10), the smooth convergence of the free energy gradient per window (SI Appendix, Supplementary Figure S11), and the small statistical errors on the resulting PMF (SI Appendix, Supplementary Figure S12) all suggest converged free energy results. The results in Figure 4A, 4C show that the position of the converter and/or the conformation of the Relay/SH1 elements effectively shift the equilibrium from an open switch II in PR+ATP to a closed switch II in PPS+ATP. Also, they indicate that a "partially closed" switch II state with a formed G459-ATP hydrogen bond but an open salt-bridge may be stabilized in PTS, which remains catalytically inactive. Visual inspection of the ABF trajectory in PTS shows that the partially closed state with a formed G459-ATP hydrogen bond is reached by uncoupling switch II from the Relay helix, which involves the breaking of a pair of hydrogen bonds between N474 on the Relay helix and the backbone of E461 on switch II (Figure 4C) as well as the extraction of the side chain of F460 from a hydrophobic cavity belonging to the L50 subdomain (see SI Appendix, Supplementary Figure S15). However, since the formation of the critical salt-bridge is still disfavored in PTS, the free energy results in Figure 4B indicate that supplementary rearrangements are required to complete the recovery stroke. As in PTS the "seesaw" motion of the Relay helix is incomplete (see *Overall Description of the PTS Myosin VI Crystal Structure*), we infer that this global movement is crucial to produce an ATPase competent state. Thus, the present ABF calculations suggest that two distinct pathways exist to reach the final PPS state: one in which the switch II – ATP hydrogen bond is formed in PTS via the uncoupling of switch II from the L50 subdomain; and another one in which the seesaw motion of the Relay helix with a fully coupled switch II results in the formation of the critical salt-bridge interaction with switch I. Although we cannot conclude on which pathway is kinetically preferred for the PTS to PPS transition, we note that both of them are consistent with a late closure of switch II during the recovery stroke transition, which is the most important result emerging from the simulations. Finally, the ABF results (Figure 4B) suggest that the partially closed switch II with a formed G459-ATP hydrogen bond would be most favored in the PTS state, which is actually not observed in the crystal structure. Since the free energy difference between the broken and formed hydrogen bond

configurations probed by ABF along d_γ is small (~ 2 kcal/mol, see SI Appendix, Supplementary Figure S9B), both states are likely populated in PTS with the fully open state possibly selected on crystallization.

DISCUSSION

Biomolecular motors like myosin harness and transduce the chemical energy of ATP by cycling through a series of complex conformational transitions. The structural characterization of all the relevant steps with atomic resolution is critical for the elucidation of the mechanism that steers function. Nonetheless, it is not sufficient. High-resolution dynamical and most importantly energetic information is needed to assess the significance of the structural states, infer on the sequence of events, and explain why alternative and potentially meaningful pathways are actually not explored. By focusing on the recovery stroke of myosin VI, we demonstrate that the synergistic use of X-ray crystallography and all-atom Molecular Dynamics provides a powerful approach to explore protein function with atomic resolution.

The recovery stroke is a critical step of the myosin cycle in which the repriming of the lever arm is coupled to ATP hydrolysis. Providing a detailed understanding of this large isomerization of the motor domain is of fundamental importance, in particular to elucidate how chemical energy may be stored in preparation for the powerstroke. However, its characterization by solution experiments is challenging. First, this transition occurs on the millisecond timescale (27), which makes it difficult for time-resolved analyses. Second, it corresponds to the largest isomerization of the motor domain, which cannot be easily correlated with a unique biophysical signal such as ATP binding, which precedes it, or ATP hydrolysis, which occurs after it. Last, it is a reversible process.

In this work, we report on the structural and dynamical characterization of a putative intermediate along the recovery stroke of myosin VI, which we term Pre-Transition State or PTS. Comparison of the PTS structure with the Post-Rigor (PR) and the Pre-Powerstroke (PPS) states reveals a previously unreported configuration of the motor domain in which the Relay/SH1 elements adopt a nearly post-recovery (PPS-like) configuration, the converter is in an intermediate position, and switch II is open. Corresponding MD simulations support the conclusion that switch II and the converter are not

mechanically coupled in PTS, with the motor domain remaining catalytically inactive even if the converter has departed from the initial pre-recovery position. Most importantly, the discovery of the PTS structure suggests a new mechanism for the recovery stroke in myosin. In the emerging scenario, the repriming of the motor head to the armed pre-powerstroke configuration would be mediated by: 1) the spontaneous isomerization (kinking/tilting) of the Relay/SH1 elements coupled with a converter swing to an intermediate position; 2) closing of switch II over the nucleotide via the seesaw motion of the Relay helix; and 3) completion of the converter swing. Intriguingly, this interpretation is consistent with a mechanism in which lever arm repriming would be initiated by thermal fluctuations and proceed through a restricted random search, with the converter probing an ensemble of configurations compatible with a kinked Relay helix until it finds its way to the post-recovery binding interface.

Free energy calculations on the closure of switch II in PR, PTS and PPS provide additional information. The results indicate that spontaneous closure of switch II is essentially impossible in PR, because it is thermodynamically disfavored, such that a transition towards an intermediate state similar to PTS would be required at the beginning of the recovery stroke. Also, they indicate that the formation of the catalytically essential salt-bridge is still unfavorable in PTS. Therefore, our analysis supports the conclusion that switch II closure is a late event of the recovery stroke, which requires an additional rearrangement of the motor domain that is not sampled yet in PTS. Finally, the results indicate that the formation of a hydrogen bond between switch II and the γ-phosphate of ATP is energetically favorable in PTS and can be formed upon breaking of interactions between switch II and the Relay helix. Hence, these free energy results are consistent with the existence of two distinct pathways to close switch II, which involve or not an uncoupling of switch II from the L50 subdomain. Assuming that the PTS structure is on-path to the post-recovery state, these results provide an understanding of the recovery stroke mechanism in myosin VI. Whether or not the emerging scenario is specific to myosin VI is presently unclear. We note, however, that the mechanism above is consistent with our recent finding that smooth muscle myosin II can be effectively trapped in a pre-hydrolysis state by binding of an allosteric inhibitor (5), whose negative modulatory activity may precisely block the conformational transition of the Relay/SH1 elements at the beginning of the recovery stroke; see SI Appendix, Supplementary Text 1.

The mechanistic interpretation of PTS emerging from X-ray crystallography and MD simulations is in clear disagreement with existing models of the recovery stroke (15, 21, 22) which were obtained for Dd Myo2, see SI Appendix, Supplementary Text 1. In the most accepted view, the recovery stroke starts with the spontaneous closure of switch II via the formation of the critical salt-bridge with switch I, which promotes a 60 degrees rotation of the converter by pulling on the Relay helix (15). This model assumes strong, mechanical coupling between the configuration of the active site (in particular the position of switch II) and the converter swing, with the Relay helix acting as a mechanical connector. In sharp contrast, our analysis of myosin VI supports the existence of statistical coupling between the re-orientation of the converter and ATP hydrolysis, suggesting a mechanism in which the repriming of the converter is mostly driven by thermal fluctuations and ultimately stabilized by closing of switch II over the nucleotide in a “ratchet-like” fashion. Since these two scenarios involve the same elementary sub-transitions, albeit with different timing, discriminating between the two would require time-resolved experiments able to deconvolute the sequence of structural events with atomic resolution, which are currently unavailable. Note, for instance, that the mutagenesis experiments in support of Fischer’s interpretation (28) cannot really distinguish between the “strongly-coupled” and the “ratchet-like” models because both of them involve the same seesaw motion of the Relay helix; see SI Appendix, Supplementary Text 1. To the best of our knowledge, only advanced simulation techniques for path optimization in free energy space, such as the string method in collective variables (29, 30), would allow for sufficient time and space resolution to determine which pathway is kinetically preferred. These challenging calculations are left for the future.

Finally, a striking peculiarity of myosin VI is the existence of two stable conformations for the converter (26, 31). As the PR structure of myosin VI exhibits the canonical R-fold converter, an internal conformational transition of the converter must take place during the recovery stroke of myosin VI. The presence of an R-fold converter in the new PTS structure is consistent with the picture that the converter isomerization takes place at the end of the recovery stroke, as previously suggested (32, 33). Also, it suggests that the P-fold is unstable when the converter does not occupy a fully reprimed PPS position. Whether the isomerization to the P-fold is required to complete switch II closure and/or full converter repriming is presently unclear and requires further investigation.

MATERIALS AND METHODS

Expression Constructs, Production, and Purification

Recombinant DNA of porcine myosin VI was generated to express a truncated myosin VI construct containing the motor domain using the baculovirus expression system. A C-terminal truncation was made at I789, creating the MD construct. This truncation is at the end of the first (proximal) helix of insert 2. In addition, the construct had a Flag tag (encoding DYKDDDDK) appended via a glycine to the N terminus to facilitate purification. Expressed myosin molecules were purified as previously described (26, 34).

Crystallization and Data Collection

Crystals of myosin VI in the PTS state were obtained with the MD construct incubated with 2mM MgADP-BeF_X using the hanging-drop vapor-diffusion method. Spontaneous nucleation occurred at 277 K with equal amounts of reservoir solution (containing 7% polyethylene glycol [PEG] 8000, 50 mM TRIS, pH 7.5, 1 mM TCEP, 15% glycerol) and stock solution of the protein (10 mg/ml in 10 mM HEPES, pH 7.5, 50 mM NaCl, 1 mM TCEP, 1 mM NaN3 with 1mM EDTA). The best crystals were obtained using seeding. Crystals of proteins were cryo-cooled prior to data collection at the European Synchrotron Radiation Facility (ESRF). The data sets were processed with XDS (35). Statistics on the data collection and the final models are given in SI Appendix, Supplementary Table S1. The myosin VI MD PTS was solved by molecular replacement with the myosin VI MD pre-powerstroke (PPS) model (PDB code 2V26) using the program Phaser (36). Refinement was performed at 2.20 Å resolution using Coot (37) and BUSTER (38). The atomic coordinates and structure factors have been deposited in the Protein Data Bank, www.pdb.org, with accession number **5O2L**.

Explicit solvent unbiased MD simulations

PR, PTS and PPS structural models were solvated in orthorhombic boxes of TIP3P water (supplemented with 150 mM NaCl) and minimized under harmonic restraints. Minimized, restrained systems were heated up to 300 K for 1 ns at constant volume. Then, 2 ns equilibration dynamics were

run at constant pressure during which the harmonic restraints were smoothly turned down. Production dynamics were launched from the resulting coordinates and velocities. Simulations were run with NAMD 2.10 (39) using the CHARMM36 force field (40). Short-range electrostatics and Van der Waals interactions were cut-off at 12 Å. Long-range electrostatics was treated by the Particle Mesh Ewald method. The length of bonds involving hydrogen atoms was constrained with RATTLE, and a 2 fs integration time step was used. See SI Appendix for details.

Potential of Mean Force calculations with the ABF method

Bi-dimensional potentials of mean force were computed along the distances d_1 between R205CZ/E461CD (critical salt-bridge) and d_γ between G459N/ATPO1G (Switch II/ATP hydrogen bond) using the adaptive biasing force (ABF) algorithm (41) as implemented in NAMD 2.10 (42). See SI Appendix for details.

ACKNOWLEDGEMENTS

We thank the beamline scientists of the beamline ID23-1 (ESRF synchrotron) for excellent support during data collection. This work was granted access to the HPC resources of CCRT/CINES under the allocation 2016-[076644] made by GENCI (Grand Equipement National de Calcul Intensif). HLS was supported by National Institutes of Health (NIH) grant DC009100. The A.H. and M.C. teams were jointly supported by the Fondation pour la Recherche Médicale (DBI20141231319). M.C. was supported by the Agence Nationale de la Recherche (ANR) through the LabEx project Chemistry of Complex Systems (CSC-MCE-13), and the International Center for Frontier Research in Chemistry (icFRC). A.H. was supported by a grant from AFM 17235. The A.H. team is part of Labex CelTisPhyBio:11-LBX-0038, which is part of the IDEX PSL (ANR-10-IDEX-0001-02 PSL). F.B. received support from the French Ministry of Higher Education and Research.

ACCESSION CODES

Coordinates and structure factors have been deposited in the Protein Data Bank under accession code **5O2L**.

AUTHOR CONTRIBUTIONS

A.H., F.B. and M.C. designed research; T.I., H.B. and A.H. were involved in crystallization and T.I. solved the X-ray structure; F.B. performed the Molecular Dynamics simulations; all authors analyzed the data; F.B., M.C. and A.H. wrote the manuscript with the help of the other authors.

References

1. Schliwa, M. (2003) *Molecular motors* ed Schliwa M (Wiley-VCH, Weinheim).

2. Geisterfer-Lowrance AAT, Kass S, Tanigawa G, Vosberg H-P, McKenna W, Seidman CE, Seidman JG (1990) A molecular basis for familial hypertrophic cardiomyopathy: A β cardiac myosin heavy chain gene missense mutation. *Cell* 62(5):999–1006.

3. Melchionda S, Ahituv N, Bisceglia L, Sobe T, Glaser F, Rabionet R, Arbones ML, Notarangelo A, Di Iorio E, Carella M, others (2001) MYO6, the human homologue of the gene responsible for deafness in Snell's waltzer mice, is mutated in autosomal dominant nonsyndromic hearing loss. *Am J Hum Genet* 69(3):635–640.

4. Makowska KA, Hughes RE, White KJ, Wells CM, Peckham M (2015) Specific Myosins Control Actin Organization, Cell Morphology, and Migration in Prostate Cancer Cells. *Cell Rep* 13(10):2118–2125.

5. Sirigu S, Hartman JJ, Planelles-Herrero VJ, Ropars V, Clancy S, Wang X, Chuang G, Qian X, Lu P-P, Barrett E, Rudolph K, Royer C, Morgan BP, Stura EA, Malik FI, Houdusse AM (2016) Highly selective inhibition of myosin motors provides the basis of potential therapeutic application. *Proc Natl Acad Sci*:201609342.

6. Malik FI, Hartman JJ, Elias KA, Morgan BP, Rodriguez H, Brejc K, Anderson RL, Sueoka SH, Lee KH, Finer JT, Sakowicz R, Baliga R, Cox DR, Garard M, Godinez G, Kawas R, Kraynack E, Lenzi D, Lu PP,

Muci A, Niu C, Qian X, Pierce DW, Pokrovskii M, Suehiro I, Sylvester S, Tochimoto T, Valdez C, Wang W, Katori T, Kass DA, Shen Y-T, Vatner SF, Morgans DJ (2011) Cardiac Myosin Activation: A Potential Therapeutic Approach for Systolic Heart Failure. *Science* 331(6023):1439–1443.

7. Pylypenko O, Song L, Shima A, Yang Z, Houdusse AM, Sweeney HL (2015) Myosin VI deafness mutation prevents the initiation of processive runs on actin. *Proc Natl Acad Sci*. doi:10.1073/pnas.1420989112.

8. Planelles-Herrero VJ, Hartman JJ, Robert-Paganin J, Malik FI, Houdusse A (2017) Mechanistic and structural basis for activation of cardiac myosin force production by omecamtiv mecarbil. *Nat Commun* 8(1). doi:10.1038/s41467-017-00176-5.

9. Geeves MA, Holmes KC (1999) Structural Mechanism of Muscle Contraction. *Annu Rev Biochem* 68(1):687–728.

10. Sweeney HL, Houdusse A (2010) Structural and Functional Insights into the Myosin Motor Mechanism. *Annu Rev Biophys* 39(1):539–557.

11. Llinas P, Isabet T, Song L, Ropars V, Zong B, Benisty H, Sirigu S, Morris C, Kikuti C, Safer D, Sweeney HL, Houdusse A (2015) How Actin Initiates the Motor Activity of Myosin. *Dev Cell* 33(4):401–412.

12. Ecken J von der, Heissler SM, Pathan-Chhatbar S, Manstein DJ, Raunser S (2016) Cryo-EM structure of a human cytoplasmic actomyosin complex at near-atomic resolution. *Nature* 534(7609):724–728.

13. Wulf SF, Ropars V, Fujita-Becker S, Oster M, Hofhaus G, Trabuco LG, Pylypenko O, Sweeney HL, Houdusse AM, Schröder RR (2016) Force-producing ADP state of myosin bound to actin. *Proc Natl Acad Sci* 113(13):E1844–E1852.

14. Warshaw DM (2004) Lever arms and necks: a common mechanistic theme across the myosin superfamily. *J Muscle Res Cell Motil* 25(6):467–474.

15. Fischer S, Windshügel B, Horak D, Holmes KC, Smith JC (2005) Structural mechanism of the recovery stroke in the Myosin molecular motor. *Proc Natl Acad Sci U S A* 102(19):6873–6878.

16. Woo H-J (2007) Exploration of the conformational space of myosin recovery stroke via molecular dynamics. *Biophys Chem* 125(1):127–137.

17. Yu H, Ma L, Yang Y, Cui Q (2007) Mechanochemical Coupling in the Myosin Motor Domain. I. Insights from Equilibrium Active-Site Simulations. *PLoS Comput Biol* 3(2):e21.

18. Yu H, Ma L, Yang Y, Cui Q (2007) Mechanochemical Coupling in the Myosin Motor Domain. II. Analysis of Critical Residues. *PLoS Comput Biol* 3(2):e23.

19. Mesentean S, Koppole S, Smith JC, Fischer S (2007) The Principal Motions Involved in the Coupling Mechanism of the Recovery Stroke of the Myosin Motor. *J Mol Biol* 367(2):591–602.

20. Koppole S, Smith JC, Fischer S (2007) The Structural Coupling between ATPase Activation and Recovery Stroke in the Myosin II Motor. *Structure* 15(7):825–837.

21. Elber R, West A (2010) Atomically detailed simulation of the recovery stroke in myosin by Milestoning. *Proc Natl Acad Sci* 107(11):5001–5005.

22. Baumketner A, Nesmelov Y (2011) Early stages of the recovery stroke in myosin II studied by molecular dynamics simulations. *Protein Sci* 20(12):2013–2022.

23. Baumketner A (2012) Interactions between relay helix and Src homology 1 (SH1) domain helix drive the converter domain rotation during the recovery stroke of myosin II. *Proteins Struct Funct Bioinforma* 80(6):1569–1581.

24. Baumketner A (2012) The mechanism of the converter domain rotation in the recovery stroke of myosin motor protein. *Proteins Struct Funct Bioinforma* 80(12):2701–2710.

25. Onishi H, Kojima S, Katoh K, Fujiwara K, Martinez HM, Morales MF (1998) Functional transitions in myosin: Formation of a critical salt-bridge and transmission of effect to the sensitive tryptophan. *Proc Natl Acad Sci* 95(12):6653–6658.

26. Ménétrey J, Llinas P, Mukherjea M, Sweeney HL, Houdusse A (2007) The Structural Basis for the Large Powerstroke of Myosin VI. *Cell* 131(2):300–308.

27. Trivedi DV, Muretta JM, Swenson AM, Davis JP, Thomas DD, Yengo CM (2015) Direct measurements of the coordination of lever arm swing and the catalytic cycle in myosin V. *Proc Natl Acad Sci* 112(47):14593–14598.

28. Kintses B, Yang Z, Málnási-Csizmadia A (2008) Experimental Investigation of the Seesaw Mechanism of the Relay Region That Moves the Myosin Lever Arm. *J Biol Chem* 283(49):34121–34128.

29. Maragliano L, Fischer A, Vanden-Eijnden E, Ciccotti G (2006) String method in collective variables: Minimum free energy paths and isocommittor surfaces. *J Chem Phys* 125(2):024106.

30. Pan AC, Sezer D, Roux B (2008) Finding Transition Pathways Using the String Method with Swarms of Trajectories. *J Phys Chem B* 112(11):3432–3440.

31. Ménétrey J, Isabet T, Ropars V, Mukherjea M, Pylypenko O, Liu X, Perez J, Vachette P, Sweeney HL, Houdusse AM (2012) Processive Steps in the Reverse Direction Require Uncoupling of the Lead Head Lever Arm of Myosin VI. *Mol Cell* 48(1):75–86.

32. Ménétrey J, Llinas P, Cicolari J, Squires G, Liu X, Li A, Sweeney HL, Houdusse A (2008) The post-rigor structure of myosin VI and implications for the recovery stroke. *EMBO J* 27(1):244–252.

33. Ovchinnikov V, Cecchini M, Vanden-Eijnden E, Karplus M (2011) A Conformational Transition in the Myosin VI Converter Contributes to the Variable Step Size. *Biophys J* 101(10):2436–2444.

34. Sweeney HL, Rosenfeld SS, Brown F, Faust L, Smith J, Xing J, Stein LA, Sellers JR (1998) Kinetic Tuning of Myosin via a Flexible Loop Adjacent to the Nucleotide Binding Pocket. *J Biol Chem* 273(11):6262–6270.

35. Kabsch W (2010) XDS. *Acta Crystallogr D Biol Crystallogr* 66(2):125–132.

36. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ (2007) Phaser crystallographic software. *J Appl Crystallogr* 40(4):658–674.

37. Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D*

Biol Crystallogr 60(12):2126–2132.

38. G. Bricogne, E. Blanc, M. Brandl, C. Flensburg, P. Keller, W. Paciorek, P. Roversi, A. Sharff, O.S. Smart, C. Vonrhein, T.O. Womack (2011) *BUSTER version 2.11. 2* (Global Phasing Ltd, Cambridge, UK).

39. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kalé L, Schulten K (2005) Scalable molecular dynamics with NAMD. *J Comput Chem* 26(16):1781–1802.

40. Huang J, MacKerell AD (2013) CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J Comput Chem* 34(25):2135–2145.

41. Comer J, Gumbart JC, Hénin J, Lelièvre T, Pohorille A, Chipot C (2015) The Adaptive Biasing Force Method: Everything You Always Wanted To Know but Were Afraid To Ask. *J Phys Chem B* 119(3):1129–1151.

42. Fiorin G, Klein ML, Hénin J (2013) Using collective variables to drive molecular dynamics simulations. *Mol Phys* 111(22–23):3345–3362.

Figure legends

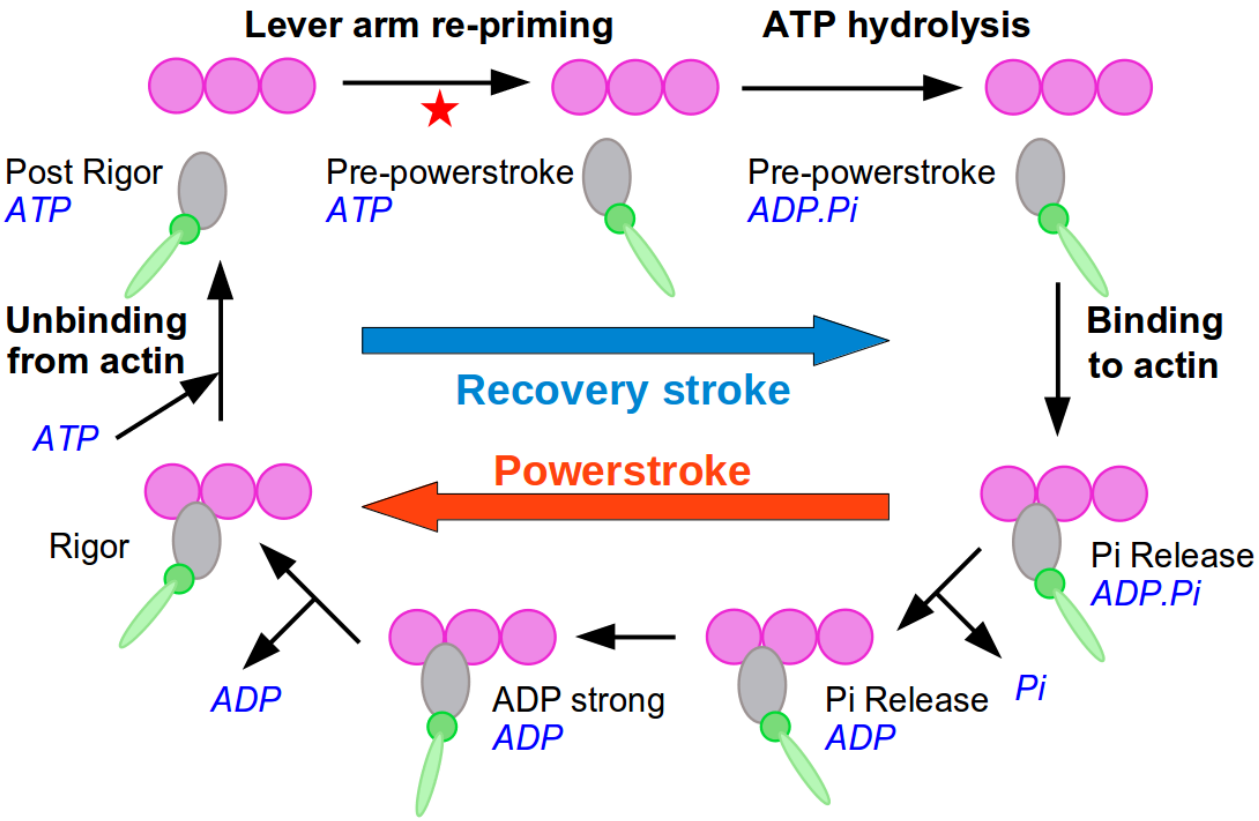
Figure 1: Overview of the actomyosin cycle. When ATP is bound, the motor undergoes a fast and reversible transition known as the recovery stroke that re-primes the lever arm in preparation for force production. The red star materializes the putative position in the cycle of the new Pre-Transition State (PTS) reported in this study.

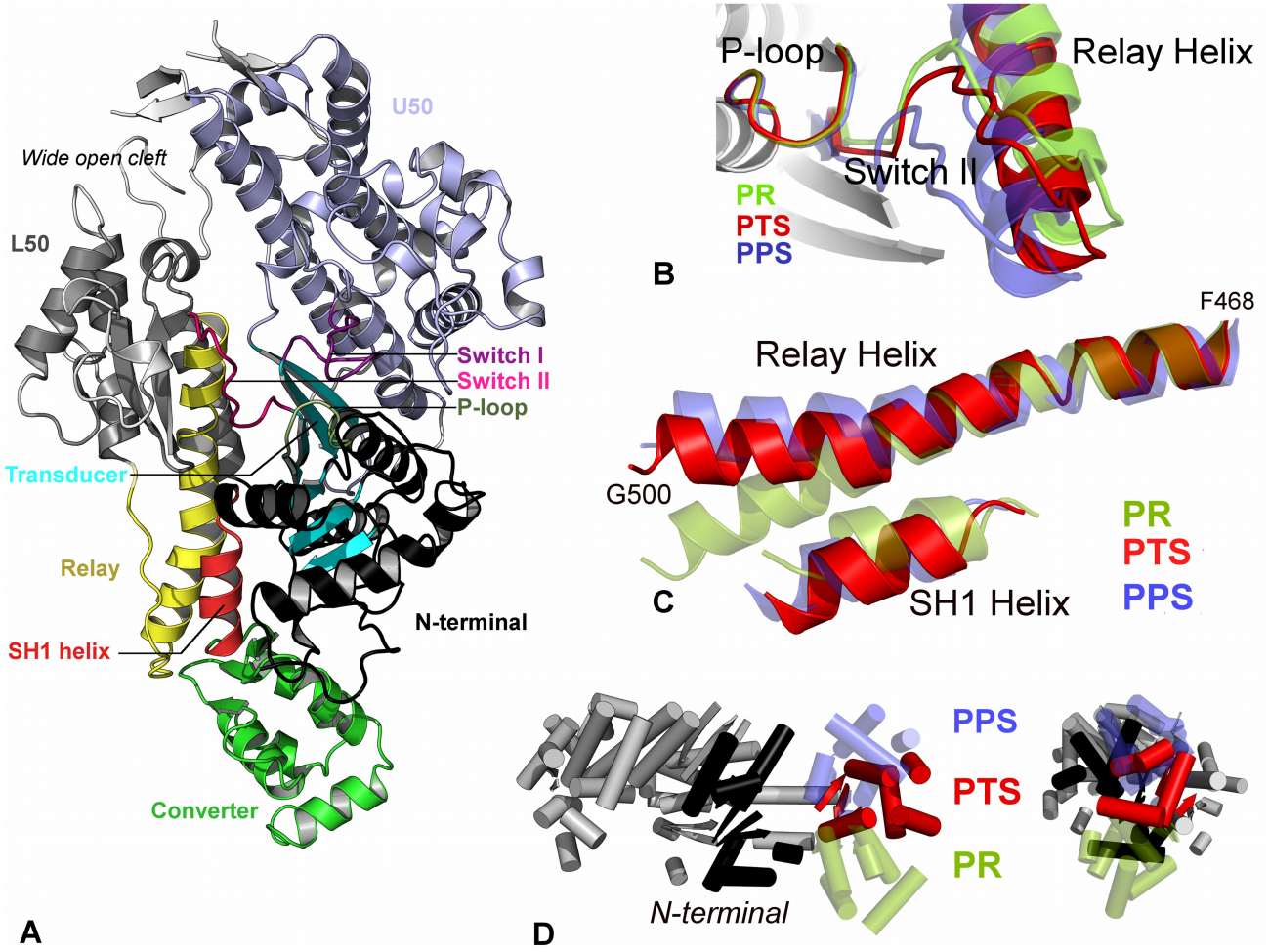
Figure 2: The PTS crystal structure reveals original structural features consistent with an on-pathway intermediate of the recovery stroke. A. General view of the PTS crystal structure of myosin VI. For clarity, the SH3 motif is not represented. **B**. Switch II adopts an “open” position closer to PR than PPS. A global movement of the Relay helix, i.e. the seesaw motion proposed by Fischer and co-workers, is also required to reach PPS. **C**. Comparison of the Relay and SH1 helices. Unlike in PR, the Relay helix exhibits a kink and the SH1 helix is tilted downward. However, the orientation of the post-kink

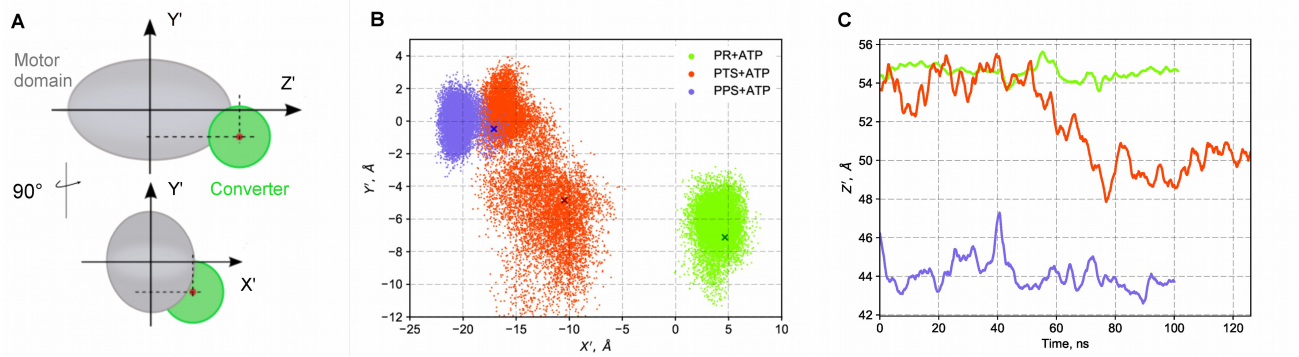
fragment in the Relay helix and the degree of tilting of the SH1 helix differ from PPS. **D**. The converter adopts an intermediate position between PR and PPS.

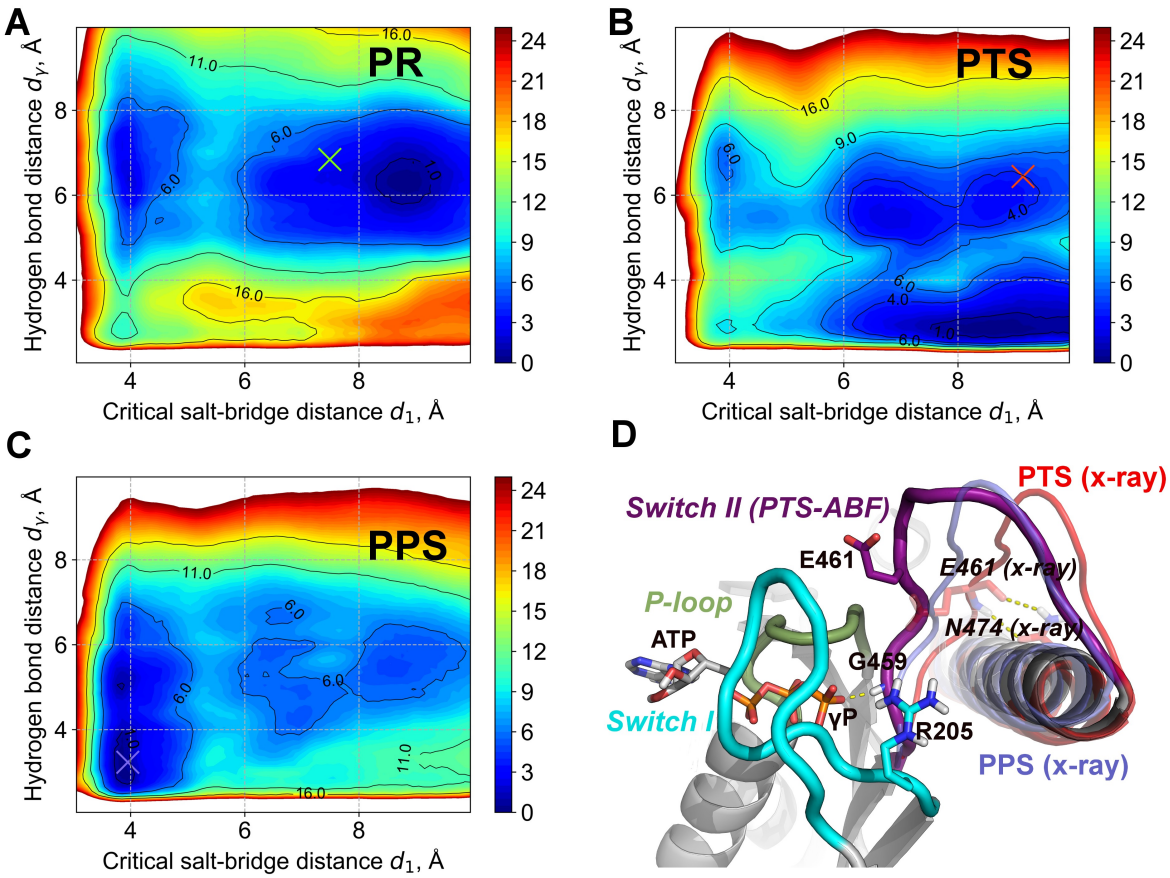
Figure 3: Positional dynamics of the converter in MD. A. Geometric observables to monitor the position of the converter in simulation. By projecting the center of geometry of the converter Cα atoms on the principal axes of the motor domain, the components *X'*, *Y'* and *Z'* provide a convenient representation of the converter position relative to the motor domain; see SI Appendix for details. **B**. Positional dynamics of the converter on the transverse plane *X'Y'*. Data points for PTS correspond to the first 125 ns; see SI Appendix, Supplementary Figure S3 for the complete data. Crosses indicate the crystallographic values. The data show the existence of two positional states for the converter in PTS: one widely distributed and centered on (−12 Å, −6 Å); one more confined and in slight overlap with PPS centered on (−15 Å, 0 Å). **C**. Time-series of the *Z'* component. The decrease in *Z'* starting at t=50 ns in the PTS simulation corresponds to a partial repriming towards the PPS position. For clarity, the running average over 2 ns is plotted.

Figure 4: State-dependent free energy landscape of switch II closure in the myosin VI motor domain. A. PR state. **B.** PTS state. **C.** PPS state. Crosses indicate values from MD-equilibrated structures, which are very similar to the crystal structures. All free energies are given in kcal/mol. **D.** Representative configuration of the "partially closed" switch II state sampled by the ABF simulation of PTS. As compared to the PTS crystal structure (in red), switch II uncouples from the Relay helix and undergoes a large motion to form the hydrogen bond with ATP. Interestingly, this new configuration is distinct from PPS (in blue), notably because the critical salt-bridge is disfavored.


Lever arm re-priming
ATP hydrolysis
Post Rigor
ATP
Pre-powerstroke
ATP
Pre-powerstroke
ADP.Pi
Unbinding
from actin
ATP
Recovery stroke
Binding
to actin
Powerstroke
Rigor
Pi Release
ADP.Pi
ADP
ADP strong
ADP
Pi Release
ADP
Pi


U50
Wide open cleft
L50
Switch I
Switch II
P-loop
Transducer
Relay
N-terminal
SH1 helix
Converter
A
P-loop
Relay Helix
Switch II
PR
PTS
PPS
B
F468
Relay Helix
G500
PR
PTS
PPS
C
SH1 Helix
PPS
PTS
PR
N-terminal
D


A
Y'
Motor domain
Z'
90°
Y'
Converter
X'
B
PR+ATP
PTS+ATP
PPS+ATP
Y', Å
X', Å
C
Z', Å
Time, ns


A
PR
B
PTS
C
PPS
D
Hydrogen bond distance d_γ, Å
Critical salt-bridge distance d_1, Å
PTS (x-ray)
Switch II (PTS-ABF)
E461
E461 (x-ray)
P-loop
ATP
N474 (x-ray)
G459
Switch I
γP
R205
PPS (x-ray)

C.2. Supplementary Information

Supplementary Information

for "**An intermediate along the recovery stroke of myosin VI revealed by X-ray crystallography and Molecular Dynamics**" by Florian Blanc, Tatiana Isabet, Hannah Benisty, H. Lee Sweeney, Marco Cecchini and Anne Houdusse

Supplementary Text 1: Competing models of the recovery stroke

Overview of previously proposed mechanisms for the recovery stroke

In the following, we outline some of the models of the recovery stroke of myosin, which we compare with the mechanistic scenario emerging from the discovery of the PTS structure. The first, and very popular, model relies on a minimum potential-energy path (obtained by conjugate peak refinement) between the end-points of the recovery stroke of Dd Myo2 (1). Based on these calculations, Fischer and co-workers proposed that the recovery stroke starts with the spontaneous formation of the critical salt-bridge between switch I and switch II (Dd Myo2 R238-E459; equivalent to Myo6 R205-E461). In this view, the formation of the critical salt-bridge brings the backbone nitrogen of G457 (G459 in Myo6) into hydrogen-bonding distance with the γ-phosphate of ATP and pulls on the Relay helix, whose bending and kinking cause the swing of the converter subdomain. Therefore, the converter swing is proposed to result from strong interactions initiated from switch II closure in the active site. Later on, Fischer and coworkers expanded on this initial model using additional conjugate peak refinement calculations and proposed that the rotation of the converter actually happens in two steps (2). The first step corresponds to the seesaw motion of the Relay helix, resulting directly from the pulling of switch II, which drives roughly half of the rotation of the converter. In the second step, the formation of an additional interaction between switch II and P-loop triggers the rotation of the so-called wedge loop (part of the L50 subdomain), which by pushing onto the SH2/SH1 helical junction, causes a seesaw motion of the SH1 helix (distinct from that of the Relay helix) which is responsible for completion of the converter swing along with the formation of the kink in the Relay helix. Although this model provides a plausible mechanistic picture and is consistent with independent mutagenesis experiments (3), there exists no experimental evidence of how the transition is initiated and how the allosteric coordination leads the motor domain to the PPS state. Furthermore, being a zero-temperature analysis, this model does not capture entropic effects by definition, and may overlook

the stochastic nature of the transition. Refinements of the model based on (short) finite-temperature Molecular Dynamics simulations did not change the overall picture (4, 5).

A second approach is the one of Woo and Harris using umbrella sampling, which represents a pioneering attempt to perform a free energy calculation along the recovery stroke (6, 7). However, the very high free energy difference predicted between the PR and PPS states (30 kcal/mol) and the short duration of the simulations cast doubt on the accuracy of these results.

Cui and co-workers tackled the problem using a wide range of computational methods (including umbrella sampling and targeted Molecular Dynamics) (8, 9). This ambitious endeavor suggested a model in which ATP hydrolysis and the swing of the converter are statistically rather than mechanically coupled to the open/close transition of switch II. Interestingly, targeted Molecular Dynamics (TMD) showed a different order of events with respect to the one proposed by Fischer, with a late closing of switch II and late kinking of the relay helix both preceded by a large rotation of the converter. Although this approach natively accounts for thermal fluctuations, TMD is known to bias the largest changes to occur first, which could lead to an unrealistic picture of the transition.

A remarkable attempt to model the recovery stroke is the work of West and Elber (10). Using milestoning, these authors proposed a sequential picture of the recovery stroke consistent with the model of Fischer and co-workers, in which the early rearrangements in the active site are followed by Relay/SH1/converter motions delayed by several hundreds of nanoseconds. Although the estimated rate (~0.5 ms) is consistent with experimental measurements, these calculations rely on a pre-defined transition path that is based on a potential-energy optimization and the proposed mechanism may retain memory of the initial zero-temperature path.

Finally, Baumketner and Nesmelov (11), and later Baumketner (12, 13), proposed the most recent model of the recovery stroke by combining unbiased Molecular Dynamics of the full motor domain with replica-exchange simulations of smaller fragments only. Interestingly, implicit solvent simulations at high temperature (350 K) of the PR state of Dd Myo2 showed spontaneous closure of switch II with no rotation of the converter, in support to the conclusion that switch II closure is the initiating event of the recovery stroke. Further studies have led to a refined model in which both the kinking of the relay helix and the converter swing are driven by the displacement of the SH1 helix, which is reminiscent of the second step in the most recent model by Fischer and co-workers. Yet, no mechanistic interpretation has been proposed for the coupling between switch II closure and the displacement of the SH1 helix.

The ratchet-like mechanism emerging form the discovery of the PTS state in myosin VI and corresponding simulations (see *Main Text*) challenges the broad consensus above and is in clear disagreement with the popular view that postulates switch II closure as the initiating event of the recovery stroke (Fischer, Woo, Elber, Baumketner).

Although Fischer and co-workers did recognize that a "loosely coupled" or stochastic version of their model would be possible and did acknowledge that the initiating event of the recovery stroke could be the converter rotation driven by thermal fluctuations rather than switch II closure, it was asserted that the sequence of events underlying the recovery stroke (i.e. seesaw motion of the Relay helix, seesaw motion of the SH1 helix and formation of the kink in the Relay helix) would be broadly respected. If so, the loosely coupled version of Fischer's model would predict the existence of a structural intermediate on the recovery stroke with a motor domain in which the Relay helix is *seesawed*, the converter is partially rotated, but the SH1 helix is not tilted and the kink in the Relay helix is not formed. Additionally, this hypothetical intermediate should have an almost closed switch II. To the best of our knowledge, there is no structural (crystallographic) evidence supporting the existence of such an intermediate. In sharp contrast, the PTS structure of myosin VI exhibits exact opposite features: in PTS the Relay helix has a kink, the SH1 helix is tilted, but the Relay helix has not done the seesaw motion, and switch II is open (see Figure 2 of *Main Text*). Hence, the mechanistic scenario emerging from the analysis of the PTS structure strongly challenges the order of events in Fischer's interpretation. Finally, although the model of Cui and co-workers is closer to a ratchet-like mechanism, the PTS structure suggests that kinking of the Relay helix would be an early event in the recovery stroke, which is inconsistent with the sequence of events proposed by these authors.

Experimental support for Fischer's model does not disprove the ratchet-like mechanism

In this section we review a series of experiments that have been used to support the "strong coupling" model of Fischer et al. (3) and argue that this evidence does not disprove the "ratchet-like" mechanism emerging from the discovery of the PTS state.

Defective seesaw mutants In 2008, Kintses and co-workers characterized a pair of Dd Myo2 mutants designed to alter the so-called "aromatic fulcrum" described by Fischer, i.e. the cluster of residues over which the Relay helix pivots during the seesaw motion (double mutant F481A, F482A , and mutant F652A) (3). These mutants exhibited a reduced ATPase activity with respect to the wild type

(apparent hydrolysis rate decreased by a factor of 2-3) and fluorescence experiments suggested that repriming of the Relay/converter region (as measured by the fluorescence of W501) was incomplete. These results were interpreted as providing evidence of the importance of the seesaw motion in the recovery stroke mechanism, supporting Fischer's interpretation. However, since the seesaw motion is as critical in the ratchet-like model, we would also expect these mutants to display a hindered recovery stroke. Moreover, we note that the structural state with a partially rearranged force-generating area detected by fluorescence measurements on the fulcrum mutants could be actually consistent with the PTS state.

Mutants to uncouple switch II and the Relay helix As pointed out by Fischer and co-workers, the hydrogen bond between the peptide group of S456-G457 (A458-G459 in Myo6) and N475 (N477 in Myo6) mediates the coupling between switch II and the Relay helix. In Fischer's model, this coupling is critical to translate the pull on switch II into driving a seesaw rigid-body motion of the Relay helix. Thus, as proposed by Fischer et al. the characterization of a mutant myosin defective for this hydrogen bond (e.g. N475A in Dd Myo2/N477A in Myo6) would be informative. We now discuss what is predicted for such mutants in the context of the ratchet-like model. Our ABF calculations indicate that a "partially uncoupled" switch II state, in which the hydrogen bond between switch II and ATP is formed (but not the critical salt-bridge) and the coupling with the Relay helix is disrupted (see Figure 4 of Main Text), may be explored in PTS. This observation suggests that a possible pathway for switch II closure would involve its transient uncoupling from the Relay helix with the motor domain remaining in the PTS conformation. Arguably, such an uncoupling would be favored rather than impaired by the N475A mutation (or N477A in Myo6) and our model would predict this mutation to be benign, against Fischer's conclusion. To the best of our knowledge, such a mutant has not yet been characterized. We note, however, that the same ABF calculation in PPS also indicates that the inward displacement of the Relay helix caused by its seesaw motion seems required to stabilize the closed state of switch II (see Figure 4 of Main Text). If so, the N475A mutant can still impair the recovery stroke by destabilizing the post-recovery state, even if the inward movement of switch II precedes that of the Relay helix.

The Dd Myo2 S456L mutation (which would be A458L in Myo6) was reported to lead to a decreased step size (14). In this mutant, the introduction of a bulky leucine is expected to hinder or even prevent a complete seesaw of the Relay helix regardless as to whether switch II uncouples or not from the Relay helix. If so, this mutation is expected to reduce the amplitude of the converter swing

during the recovery stroke, thus potentially limiting the forward swing of the lever arm during the powerstroke, which would explain the reduced step size observed experimentally (14).

Coupling between the re-priming of the converter and the rotation of the L50 subdomain In phase 2 of Fischer's model, an inward rotation of the L50 subdomain (driven by the formation of a hydrogen bond between switch II and the P-loop) triggers a seesaw motion (or tilting) of the SH1 helix, which was proposed to drive the second half of the converter rotation. In the PTS structure, the L50 subdomain has not undergone this rotation, which is required to reach the PPS state. Yet, the SH1 helix is tilted, which suggests that coupling between the movement of the L50 and the seesaw of the SH1 helix is weak, if not absent. Mutations in the SH1-SH2 helical junction (G680A and G680V) in Dd Myo2 were shown to produce lower motility and lower ATPase rates, and were put forward by Fischer *et al* to support their model (15, 16). Since the tilting of the SH1 helix hinges on the SH1-SH2 junction, we predict that mutations in this area will affect the transition to PTS, and thus potentially affect the recovery stroke kinetics even in the absence of coupling between the SH1 helix and the L50 subdomain.

In Dd Myo2, residue F458 was proposed by Fischer et al to act as a physical connector between switch II and the wedge loop, thereby providing mechanical coupling between these two elements. Our ABF calculations on Myo6 are consistent with this interpretation and show that "unbinding" of the homologous residue (F460) from a hydrophobic cavity formed by residues belonging to the wedge loop promotes uncoupling of switch II from the Relay helix. The mutation F458A in Dd Myo2 was shown to prevent actin-activated Pi release, while retaining (and even increasing) the basal ATPase activity (17). Since this mutation is likely to decrease the strength of interaction between switch II and L50, it will favor the exploration of a conformational state in which switch II is closed (ATPase active) but the actin-binding cleft is open, consistent with the experimental observations.

A striking rearrangement of the hydrophobic cluster in the Relay-SH1 region (involving F487, F506 and I687 in Dd Myo2/L489, Y508 and L700 in Myo6), which is observed by comparing the pre- and post-recovery states of myosin, was proposed to act as an "aromatic switch" to stabilize either the straight or the kinked state of the Relay helix (1, 13).

It was argued by Fischer and co-workers that such a rearrangement, which involves the threading of the bulky side chain of F487 (L489 in Myo6) between the Relay helix and the Relay loop, is sterically hindered as long as the seesaw motion of the Relay helix is incomplete, which lets them

conclude that the seesaw motion occurs before kinking of the of the Relay helix. Characterized mutations in this region, i.e. F487A and F506G, were shown to uncouple ATPase activity from lever arm motion, which underlined the importance of this hydrophobic cluster on the mechanism of the recovery stroke (18). However, these mutants provide no information on the sequentiality of the recovery stroke transition, as the aromatic switch will be implicated in the stabilization of a kinked Relay helix independently of whether this rearrangement is an early or late event. Strikingly, the PTS structure of myosin VI exhibits a rearranged hydrophobic cluster, which demonstrates that a kinked, pre-seesaw Relay helix is possible.

In conclusion, since both the strong coupling (Fischer's) and ratchet-like (ours) models of the recovery stroke share the same elementary steps, although with a different timing, the current experimental support in favor of the former does not disprove the latter. By contrast, observations made on the PTS structure of myosin VI are in clear contradiction with the model of Fischer, thus challenging the strong coupling view of the recovery stroke. Whether the mechanistic scenario emerging from the PTS structure is specific to myosin VI or generalizable to the superfamily is still unclear and requires further investigation. Finally, we note that the early work of Fischer *et al*. and the subsequent computational studies by other investigators were instrumental and helped characterizing the details of the individual sub-steps the motor takes during the recovery stroke.

Insight from the Smooth Muscle Myosin II structure solved in complex with an allosteric inhibitor of the recovery stroke

The recent resolution of a crystal structure of Smooth Muscle Myosin II (SMM2) in complex with an allosteric inhibitor by us provides additional information (19). The inhibitor binds in a pocket between the Relay and SH1 helices and stabilizes an intermediate of the recovery stroke where switch II is open, the Relay helix is straight and not seesawed, and the converter has moved by less than 10° relative to the pre-recovery conformation. Thus, this structure suggests that the inhibitor hinders the first major transition in the recovery stroke.

Because the "strong coupling" (Fischer's) and the ratchet-like (our) models disagree on the nature of the initiating event (in the former it would be the seesaw motion of the Relay helix; in the latter, the kinking in the Relay helix and/or tilting of the SH1 helix), this structure may be used to distinguish between the two mechanistic scenarios. In fact, if the inhibitor blocked the formation of

the kink in the Relay helix, the ratchet-like model would explain why it traps a near-pre-recovery state. Conversely, if the inhibitor hindered the seesaw motion, Fischer's model would be most consistent with this recent crystallographic evidence. The precise mechanism of action of the SMM2 inhibitor is presently unclear. We note, however, that the effect of inhibitor binding on the basal ATPase activity is a 100-fold reduction, in contrast to the fulcrum mutants of Dd Myo2 that exhibited a 2-3-fold reduction (3) Although this comparison must be taken with care as it involves two different isoforms analyzed with different experimental techniques, this observation suggests that the inhibitor would not affect the seesaw motion, thus favoring a ratchet-like mechanism for the recovery stroke.

More generally, the discussion above highlights how a time-resolved, atomistic description of the recovery stroke transition would provide a fundamental understanding of chemo-mechanical transduction in molecular motors and the mechanism of action of positive/negative allosteric modulators. For this purpose, free energy calculations started from the structure of SMM2 in complex with the inhibitor could be used to compare the impact of drug binding onto the free energy barriers associated with the elementary rearrangements (i.e. seesaw vs kinking of the Relay helix) so as to distinguish between competing models. These calculations are left for the future.

Supplementary Text 2. Position of the converter characterized by projections on the reference axes.

The position of the converter relative to the motor domain was characterized by projecting its center of geometry on the three principal axes of the latter. These axes were computed by principal component analysis of the averaged Cα coordinates of the structural core of the motor domain, which includes residues 70-83; 88-90; 93-96; 109-115; 127-141; 147-149; 158-171 of the N-terminal subdomain; residues 177-193 (loop 1); residues 207-215; 218-227; 449-455 of the transducer; residues 232-236; 246-263; 268-276; 285-290; 298-303; 313-327; 331-349; 369-379; 383-396; 406-410; 413-442; 603-611; 615-623; 625-628 of the U50 subdomain; and residues 468-480; 512-519; 525-534; 540-551; 560-562; 566-570; 576-581; 584-589; 593-597; 643-660; 662-669; 682-690 of the L50 subdomain. These residues were chosen for their structural invariance between the PR, PTS and PPS crystal structures. The longitudinal axis of the motor domain (i.e. the direction of maximal extension) is referred to as Z'. The transverse axes are called X' and Y'. See Supplementary Figure S2 for

illustration. X',Y' and Z' were expressed as *distanceZ* collective variables in the *colvars* module of NAMD for monitoring during the MD simulations.

Supplementary Text 3. Conformational dynamics of the Relay/SH1 elements in MD simulations.

The orientation of the Relay helix was monitored by measuring the angle formed by the C-terminal region of the helix (residues 490 to 499) with its configuration in the PR crystal structure, after structural alignment on the main body of the motor domain (residues 50 to 650), see Supplementary Figure S6. This observable measures the kink of the Relay helix, with small angles (<10 degrees) corresponding to a straight helix and higher values to a kinked helix. Similarly, the orientation of the SH1 helix was monitored by measuring the angle formed relative to its configuration in PR. These two angles were measured using *orientationAngle* collective variables from the *colvars* module in NAMD.

The results shown on Supplementary Figure S6 demonstrate that both the Relay and SH1 helices adopt an intermediate orientation in the PTS state (0-40 ns) and that their re-orientation is coupled to the partial swing of the converter to the PPS state. The scatter plots, probability distributions and full time-series for all independent unbiased MD repeats are shown on Supplementary Figure S7.

The large amplitude movements of the SH1 helix (but not the Relay helix) observed in simulations PR+ATP (3) and PPS+ATP (2), correspond to the rather large movements of the converter sampled in these simulation repeats (Supplementary Figure S3).

Supplementary Text 4. Detailed Procedures.

Preparation of structural models for simulations

The structural models were prepared from the corresponding crystal structures for PTS, PR (PDB code: 2VAS) and PPS (PDB code: 2V26) of myosin VI. For each model, the motor domain was truncated after residue I789 to match the MD construct.

Missing fragments reconstruction

Coordinates for missing fragments (PR: residues 1-3, 353-367, 394-409, 623-638; PTS: residues 1-4, 356-360, 397-405, 624-631; PPS: residues 1-4, 174-180, 396-404, 622-637) were obtained by homology modeling from the pool of available crystal structures, mostly but not exclusively from myosin VI. The program *MODELLER* was used (20). In each case, ten loop models were generated and

the best was selected based on the DOPE (Discrete Optimized Potential Energy) score. Each structure was then submitted to the MolProbity server to ensure reasonable rotameric states for the residue side-chains (21). The nucleotides (ATP or ADP.Pi) were modelled by substituting the corresponding atoms from the nucleotide analogues present in the crystal structures. In the case of the PPS+ATP model (obtained from a PPS structure solved with the ADP.Pi analogue ADP.vanadate), a water molecule was placed on the position occupied by the vanadate oxygen most remote from the β-phosphate.

pKa calculations and protonation states assignment

The most likely protonation states of the histidines at neutral pH were determined by pK_A calculations using a multisite titration approach (22). In this approach, the solvent and protein interior are treated as continua (of respective dielectric constants of 80.0 and 4.0). Ions were modeled by a Boltzmann-distributed charge density corresponding to 150 mM NaCl at 300 K. The electrostatic potential was computed by solving the Poisson-Boltzmann equation numerically with the Adaptive Poisson-Boltzmann Solver (23), using the tAPBS front-end (http://agknapp.chemie.fu-berlin.de/karlsberg/). Finally, a Monte Carlo sampling was used to evaluate the protonation probabilities with the Karlsberg2 program (24, 25). Since protonation states are not treated dynamically by the force-field, different protonation states in PR, PTS and PPS would introduce insurmountable barriers and potentially hinder spontaneous conformational transitions between the structures. Thus, when different protonation states were predicted for the same histidine in PR, PTS or PPS, we retained the PTS prediction for all three structures. All non-histidine titratable residues were assumed to be in their standard protonation state.

Unbiased Molecular Dynamics simulations

Each structure was placed in a 144 Å x 108 Å x 96 Å orthorhombic box and solvated with TIP3P water molecules along with Na and Cl ions to ensure both electroneutrality and a salt concentration of 150 mM. The programs CHARMM (version c38b1) and VMD (version 1.9.2) were used for the preparation of the simulation boxes (26, 27). The CHARMM36 force field was used to model the energetics. After solvation, each system was subjected to 5000 steps of energy minimization with NAMD (version 2.10) with 10 kcal/mol/$Å^2$ harmonic restraints on the heavy atoms of the protein and 5 kcal/mol/$Å^2$ ones on

the oxygen atoms of the crystallographic water molecules. Each minimized system was heated to 300 K over 1 ns using Molecular Dynamics with active harmonic restraints. Then, an NPT equilibration was performed for 2 ns with a Langevin thermostat set to 300 K (friction coefficient of 1 ps^{-1}) and a Berendsen barostat set to 1 bar (with a 400 fs time-constant) (28). During equilibration, the harmonic restraints were smoothly removed following a cubic scaling to yield a free structure at the end. NPT production simulations were then launched from the equilibrated systems with the same parameters except for the Langevin friction, which was set to 0.1 ps^{-1}. Covalent bonds involving hydrogen atoms were constrained using RATTLE, which allows for a 2 fs time step (29). Short-range electrostatic and Van der Waals interactions were cut-off at 12 Å. The Particle Mesh Ewald (PME) method was used for long-range electrostatics, along with a 6-th order spline interpolation and a 1 Å-spaced grid. The r-RESPA multiple-stepping scheme was used with an evaluation of both bonded and Van der Waals forces at every step, and an evaluation of electrostatic forces every 2 steps (30). During the production runs, the protein was kept parallel to the box using a harmonic restraint (of force constant 1000 kcal/mol) on the *orientation* quaternion as implemented in the *colvars* module of NAMD. Simulation analysis and visualization were performed using the *colvars* module in NAMD, Wordom (31, 32), VMD, and Pymol (www.pymol.org) along with in-house Python scripts (33–35).

Adaptive Biasing Force (ABF) calculations

Briefly, ABF records the average generalized force experienced by the collective variable (i.e. minus the free energy gradient), then it applies an exact opposite force so as to locally flatten the free energy landscape (36). The estimate of the average force is refined as the simulation progresses and the system explores previously inaccessible regions. At convergence, the ABF bias cancels out the generalized force felt by the collective variable, whose dynamics becomes diffusive, which allows for efficient sampling. In this work, we used ABF to probe the energetics of switch II closure using two distances (d_1 and d_γ, see Main Text and Supplementary Figure S9A) as collective variables. At any time, the 2D PMF estimate can be recovered by thermodynamic integration from the free energy gradient profile collected by the simulation:

$$F(d_1, d_\gamma) = \int\int \nabla F_{ABF} \mathrm{d}d_1 \mathrm{d}d_\gamma$$

The integration of the gradient profile into a 2D PMF was performed using the *abf_integrate* script provided with the *colvars* module.

The 1D PMF along d_1 was then computed from the 2D PMF $F(d_1, d_\gamma)$ by averaging out d_γ as

$$F(d_1) = -k_B T \ln \int e^{-F(d_1, d_\gamma)/k_B T} \mathrm{d}d_\gamma$$

where the integration is performed over the entire domain sampled by the ABF simulation. Finally, the potential of mean force along d_1 was shifted so that the minimum free energy state is assigned a zero ΔF. The same procedure was used to obtain the PMF along d_γ. Boltzmann averaging was performed by Simpson integration using Scipy (35). The configurational space defined by the two reaction coordinates (d_1, d_γ) was discretized on a square grid of spacing 0.1 Å between 3 and 10 Å for d_1 and 2 and 10 Å for d_γ, resulting in a 70x80 grid. At any given grid point, the biasing force was applied after the grid point was visited more than 200 times over the course of the simulation (*fullSamples* parameter).

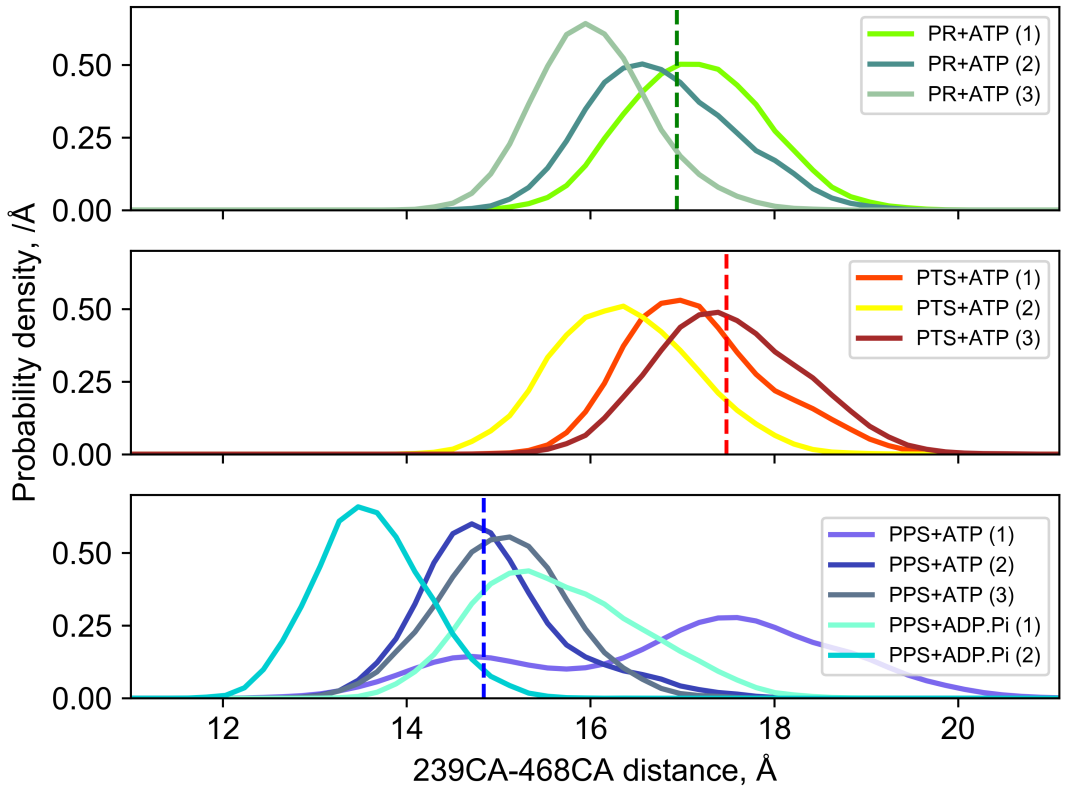
To ensure convergence of the calculations, a two-step strategy was adopted. First, a non-stratified ABF bidimensional free energy calculation was carried out until full coverage of the configurational space was achieved. Then, the configurational space was divided into 56 (1 Å x 1 Å) non-overlapping windows delimited by half-harmonic potentials (force constant 300 kcal/mol/Å^2) and a stratified ABF calculation was performed. The starting configuration and bias for each window were extracted from the previous non-stratified run. Smooth convergence of the 2D free-energy gradient in the stratified calculation was achieved by increasing sampling by approximately one order of magnitude with respect to the initial non-stratified sampling. Simulation lengths are reported in Supplementary Table S4 and amounted to more than 400 ns cumulated simulation time per myosin structure. ABF simulations were run with the same parameters as in production runs, except that RATTLE was disabled for the protein because it is incompatible with the calculation of the generalized force. Consistently, a 1 fs time step was used for the ABF calculations. In addition, no harmonic restraint was applied on the orientation of the protein. Supplementary Figure S10 shows that stratification ensures nearly uniform sampling of the relevant configurational space. In addition, analysis of the root mean square deviation (RMSD) of the per-window free energy gradient shows variations of < 0.3 kcal/mol/Å on the last 500 ps in more than 90% of the windows (Supplementary Figure S11), and < 0.9 kcal/mol/Å in all windows. Finally, statistical errors on the resulting 2D and 1D free energy profiles, which were estimated using a Gaussian perturbation approach inspired of (37), are < 1.5 kcal/mol almost everywhere; see Supplementary Figure S12. For this purpose, the standard deviation of the point-by-point 2D free-energy gradient was estimated by generating a thousand gradient profiles by adding

random numbers drawn from a Gaussian distribution with zero mean and standard deviation equal to the standard error of the mean of the gradient estimate to the stratified simulation data. A 1 ps decorrelation time was used to compute the standard error of the mean. Then, point-by-point statistical errors on the 2D free energy profile were estimated as the standard deviation of the thousand 2D PMFs resulting from the integration of the perturbed 2D gradient profiles. The statistical errors on the 1D free energy profiles were estimated similarly upon integration of the perturbed 2D gradient profiles, followed by Boltzmann averaging. The uniform coverage of the configurational space (Supplementary Figure S10), the smooth and nearly uniform convergence of the free energy gradient per window (Supplementary Figure S11), and the small errors of the resulting potentials of mean force (Supplementary Figure S12) suggest converged ABF calculations. Finally, to demonstrate the robustness of the free energy results, an independent ABF analysis of the PR+ATP, PTS+ATP and PPS+ATP myosin states was carried out using the distance between the side chains of R199 and E461 as a secondary reaction coordinate *(d_2)*; these residues are close in space and were found to form a salt-bridging interaction during the MD equilibration of the PTS structure that was stable for > 100 ns. The ABF calculations were carried out using the two-step strategy described above. Strikingly, the results in Supplementary Figure S13 show that the potentials of mean force projected on the primary salt-bridge distance *(d_1)* are barely affected by the use of a non-correlated secondary reaction coordinate. Although the height of the barriers as well as the fine structure of the free energy basins (e.g. the open salt-bridge state in PPS) are not exactly the same, the gross features appear remarkably conserved. These simulation results thus indicate that independent ABF calculations using different pairs of reaction coordinates converged to the same free energy result, which strongly supports the relevance of the numerical results in Figure 4 of the *Main Text*. In addition, the statistical distribution of the collective variables sampled in unbiased simulations is in agreement with the predicted basins from the ABF calculations (Supplementary Figure S14).

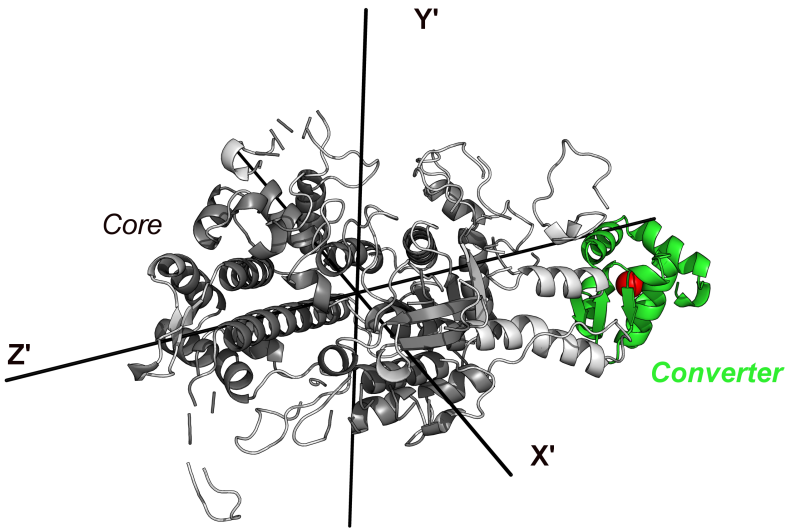
Last, we note that the three myosin structures explored by ABF calculations (i.e. PR+ATP, PTS+ATP and PPS+ATP) actually represent the same chemical state of the motor domain. Therefore, in the limit of infinite sampling, these calculations should yield the same, global PMF, which is not the case here (Figure 4). We stress that the goal of our analysis was evaluating the “local” PMF felt by the collective variables d_1 and d_γ in a given conformational basin of the motor domain (PR, PTS or PPS). In this sense, the potentials of mean force reported in Figure 4 of the *Main Text* represent effective free energies

computed in the mean field of the motor domain, whose conformation is assumed to be globally stable on the simulation timescale. To obtain the “global” PMF, the large-amplitude conformational transitions of the motor domain should be averaged out, which requires significantly longer ABF calculations.

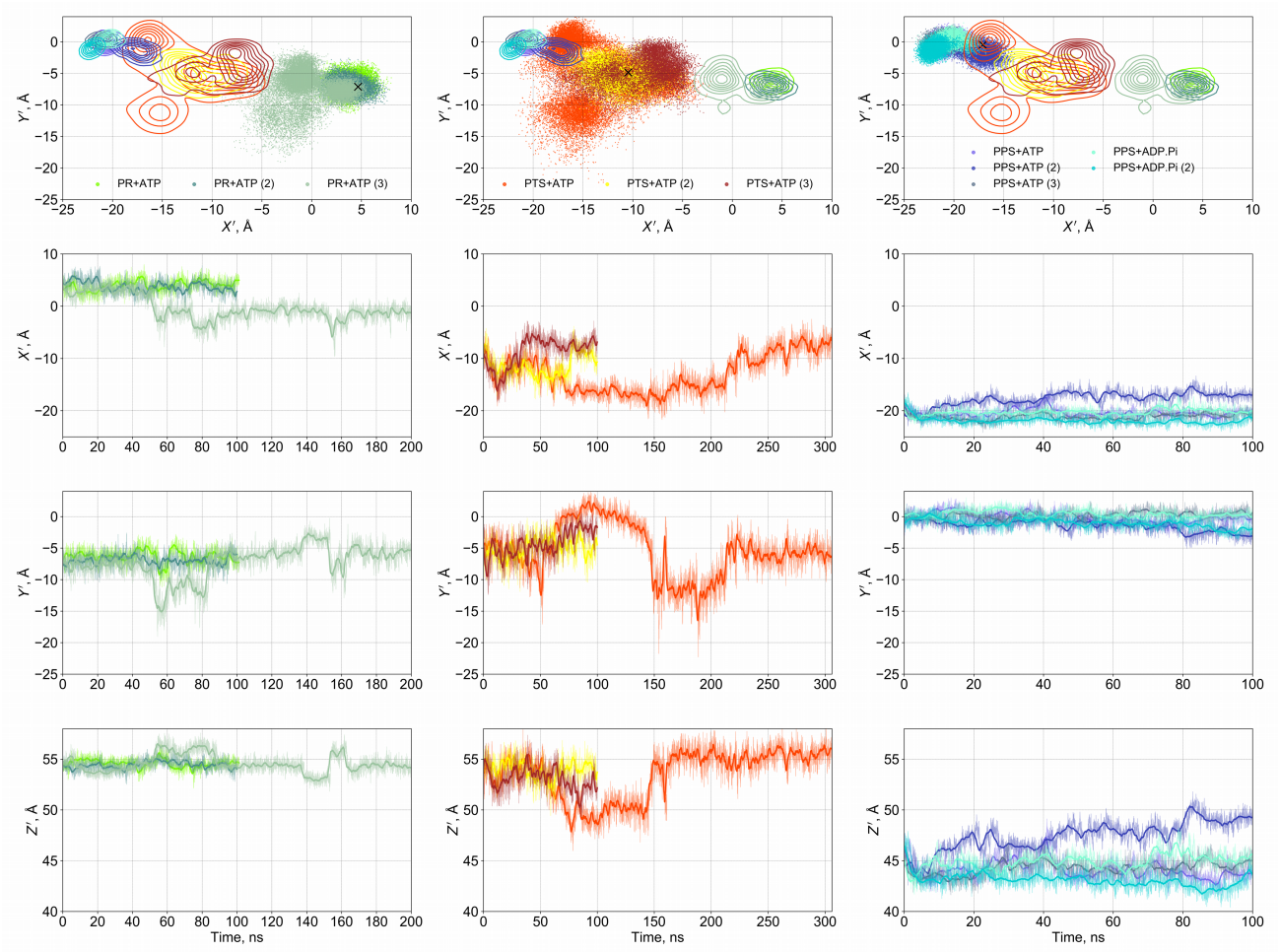
Supplementary Figures



Supplementary Figure S1: Opening state of the cleft during independent unbiased MD simulations. The distance between the Cα atoms of residues 239 (U50) and 468 (L50) was used to characterize the opening state of the cleft. Statistical distributions from unbiased MD repeats are given for the PR+ATP (top panel, green curves), PTS+ATP (middle panel, red/yellow curves), PPS+ATP (bottom panel, blue curves) and PPS+ADP.Pi (bottom panel, cyan curves). The same color code is used throughout the paper. The dotted lines indicate the crystallographic values. In both PR and PTS, the cleft is wide open and the configurations sampled in MD are consistent with the corresponding crystal structures. In PPS, the cleft is partially closed and remains so, except for one PPS+ATP simulation in which a re-opening is observed, possibly due to the presence of ATP in the active site.

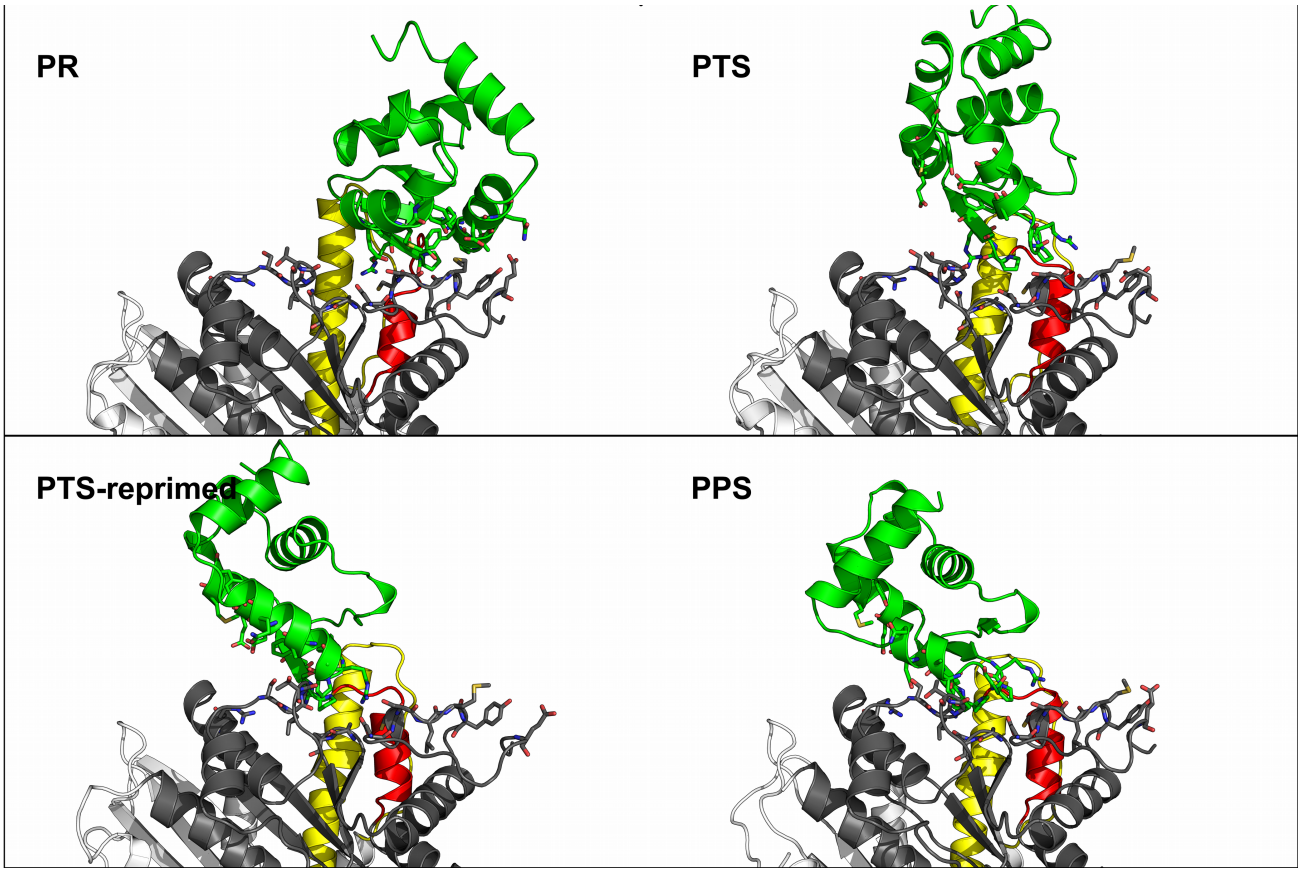


Supplementary Figure S2: Residues and axes used to define the converter projection. Black lines, principal axes. Dark gray, residues of the core used to compute the principal axes. Green, converter residues. Red sphere, center of geometry of the converter which is projected onto the principal axes.

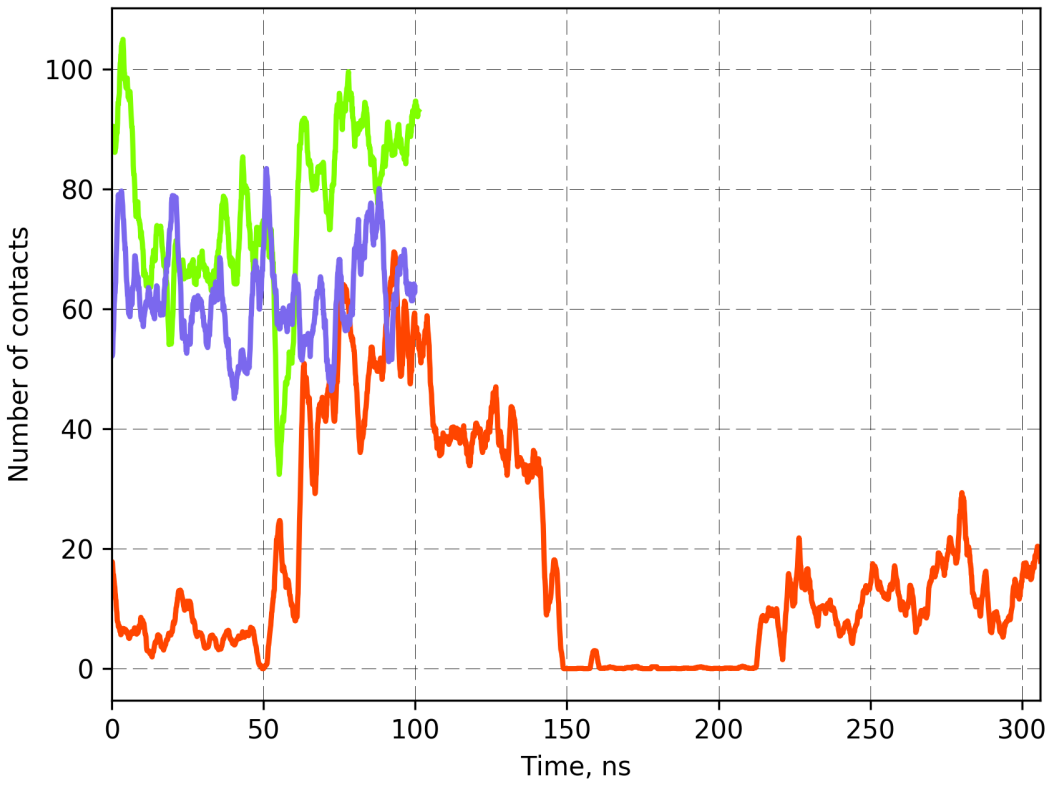


Supplementary Figure S3: Evolution of the converter projections on the motor domain reference axes during independent unbiased MD. First row: scatter plots of the X' component vs the Y' component for PR (left), PTS (middle) and PPS (right) unbiased simulations. For comparison purposes, the density lines of the statistical distributions (obtained by Kernel Density Estimation) of the other two states are shown as thick lines. Black crosses materialize the crystallographic values. Second row: evolution of the X' component in PR (left), PTS (middle) and PPS (right) unbiased simulations. Third row: evolution of the Y' component in PR (left), PTS (middle) and PPS (right) unbiased simulations. Fourth line: evolution of the Z' component in PR (left), PTS (middle) and PPS (right) unbiased simulations. For clarity the 2 ns running average (thick line) is superimposed to the raw data. The results reveal different dynamic behaviors of the converter in PTS (large fluctuations due to the converter being uncoupled from the N-terminal) and in PR/PPS (restricted fluctuations in most simulations). Interestingly, the PR+ATP simulation repeat 3 and PPS+ATP simulation repeat 2 sample

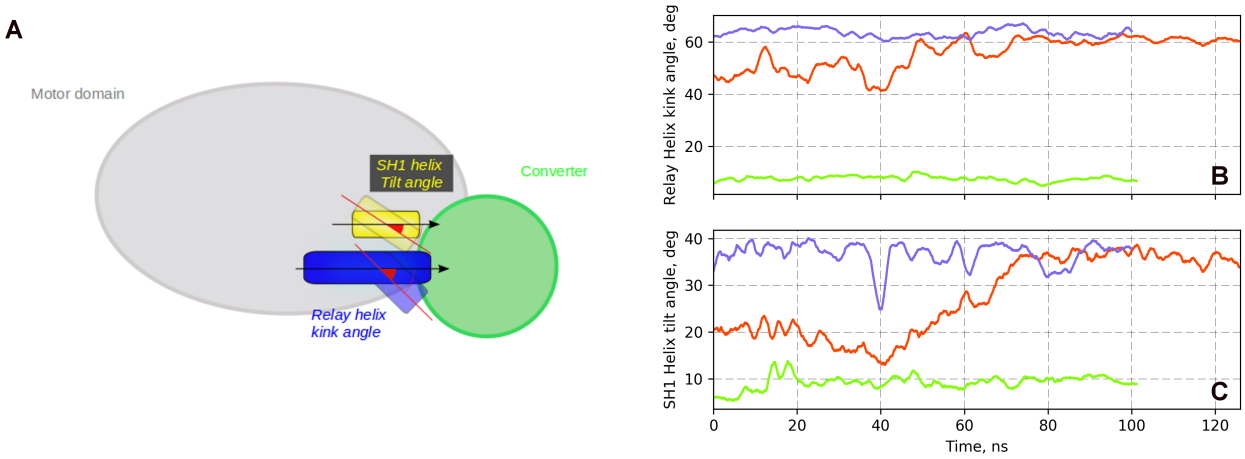
larger movements of the converter that deviate from the trends observed in other repeats of the same structural states. We note that these movements occur towards PTS-compatible positions.



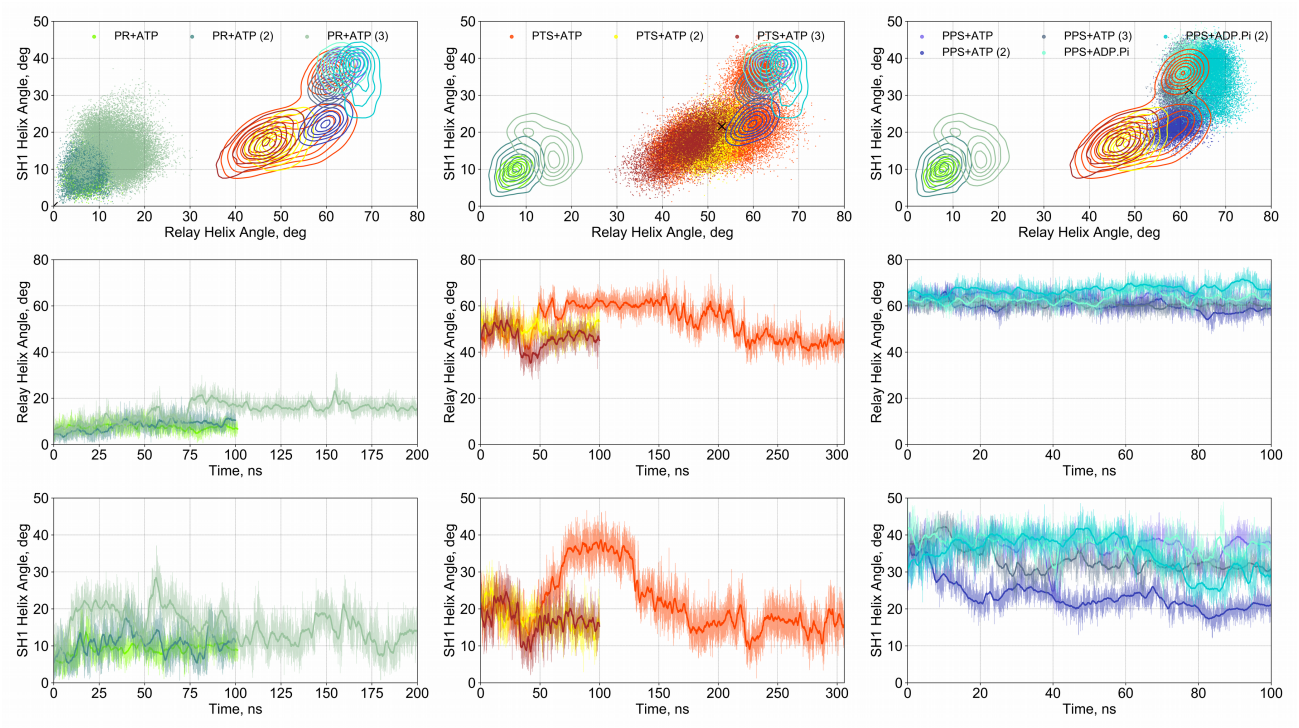
Supplementary Figure S4: Visualization of the N-terminal/converter interface for various structures of the motor domain of myosin VI. The converter exhibits only a few contacts with the N-terminal subdomain in PTS, unlike in PR and PPS in which more contacts are formed. The PTS-reprimed state, which is reversibly sampled in the simulation, exhibits a comparable (although lower) number of contacts as in PR and PPS, but a different interface. See also Supplementary Table S2. For clarity, the SH3 motif is not represented.



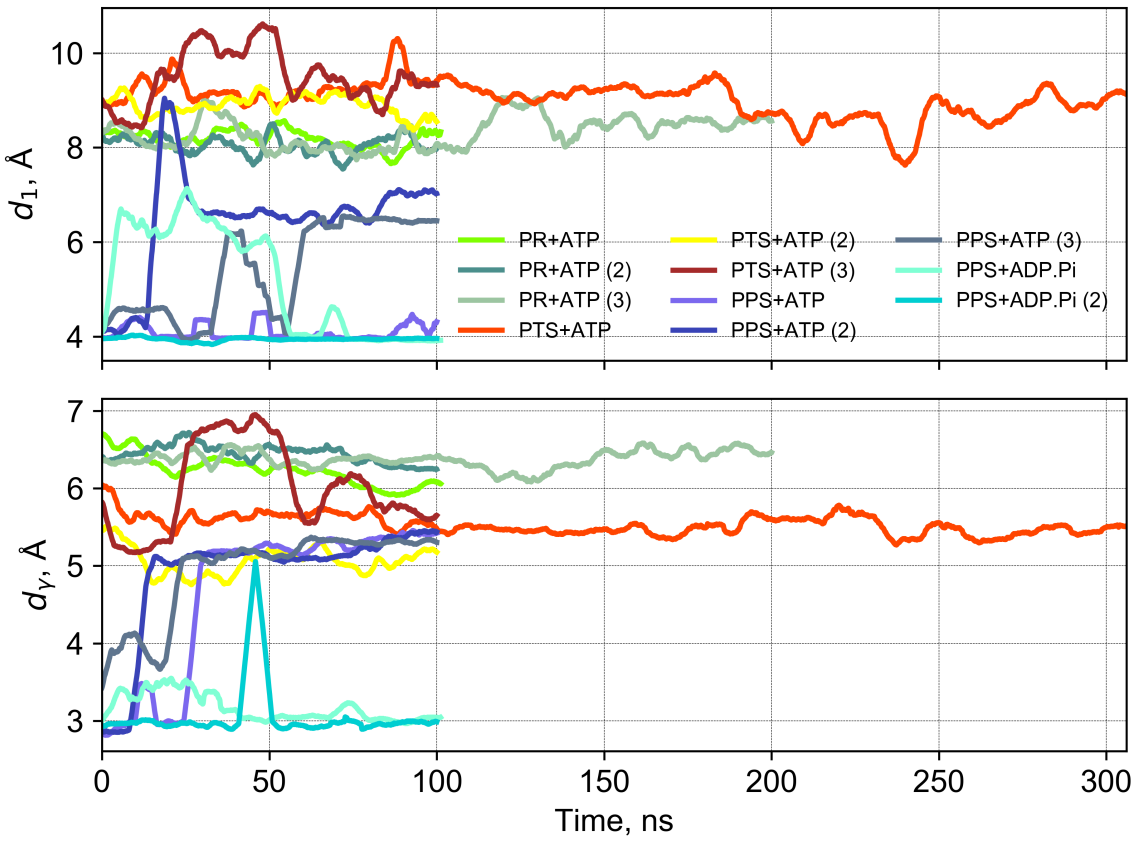
Supplementary Figure S5: Evolution of the number of contacts between the N-terminal and converter subdomains during independent unbiased MD. Green, PR+ATP (1); Blue, PPS+ATP (1); Red, PTS+ATP (1). A contact is considered formed between two atoms if they are within 4.5 Å from each other. For clarity the 2 ns running average is shown. The results demonstrate that the converter and N-terminal subdomains maintain a high number of contacts throughout the PR and PPS simulations, whereas the PTS is initially in a "decoupled converter" state. Upon partial re-priming of the converter at t=50 ns in the PTS simulation, new contacts are formed which stabilize its position (PTS-reprimed state). Finally, the converter in PTS spontaneously goes back to its initial decoupled state. See also Supplementary Table S2.



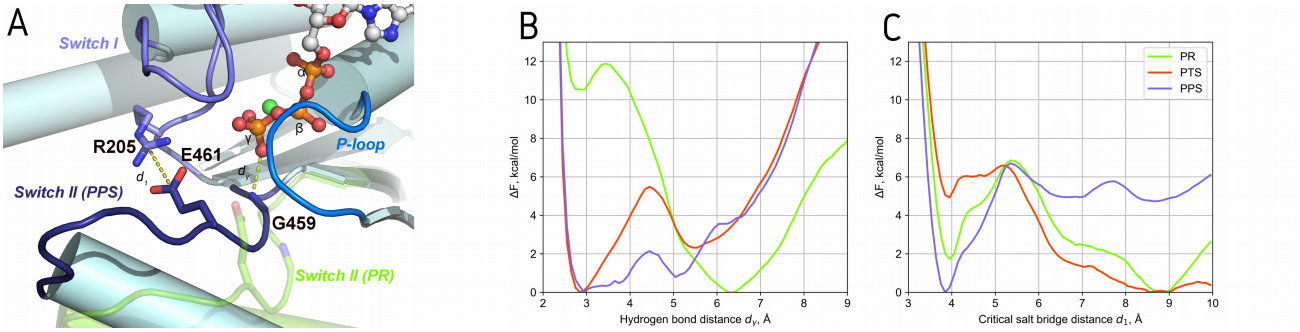
Supplementary Figure S6: The rearrangement of the Relay/SH1 elements is coupled to the partial converter swing observed in the PTS+ATP simulation. A. Definition of the Relay helix kink angle and the SH1 helix tilt angle; see Supplementary Text 3. **B.** Evolution of the Relay helix kink angle. **C.** Evolution of the SH1 tilt helix angle. Green, PR+ATP (1); Blue, PPS+ATP (1); Red, PTS+ATP (1). For clarity the 2 ns running average is shown and only the first 125 ns of the PTS+ATP simulation are represented.



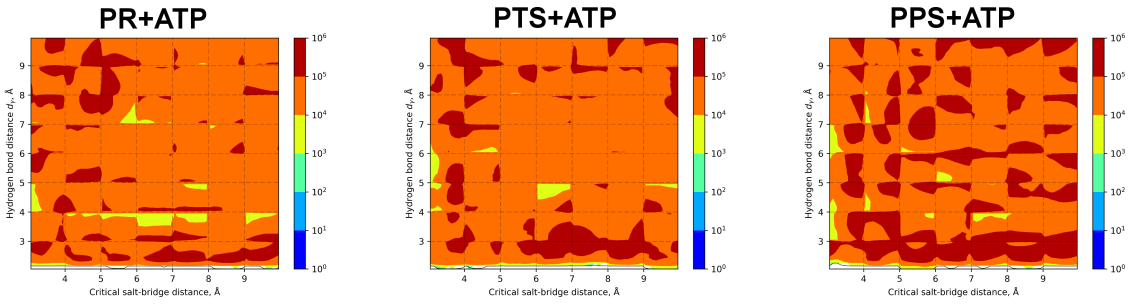
Supplementary Figure S7: Conformational dynamics of the Relay/SH1 elements during independent unbiased MD. First row: scatter plots of the Relay helix kink angle vs the SH1 helix tilt angle for PR (left), PTS (middle) and PPS (right) unbiased simulations. For comparison purposes, the density lines of the statistical distributions (obtained by Kernel Density Estimation) of the other two states are shown as thick lines. Black crosses materialize the crystallographic values. In PR, the black cross is on (0,0) because the PR crystal structure was used as reference to compute the angles. Second row: evolution of the Relay helix kink angle in PR (left), PTS (middle) and PPS (right) unbiased simulations. Third row: evolution of the SH1 helix tilt angle in PR (left), PTS (middle) and PPS (right) unbiased simulations. For clarity the 2 ns running average (thick line) is superimposed to the raw data. See Supplementary Text 3.



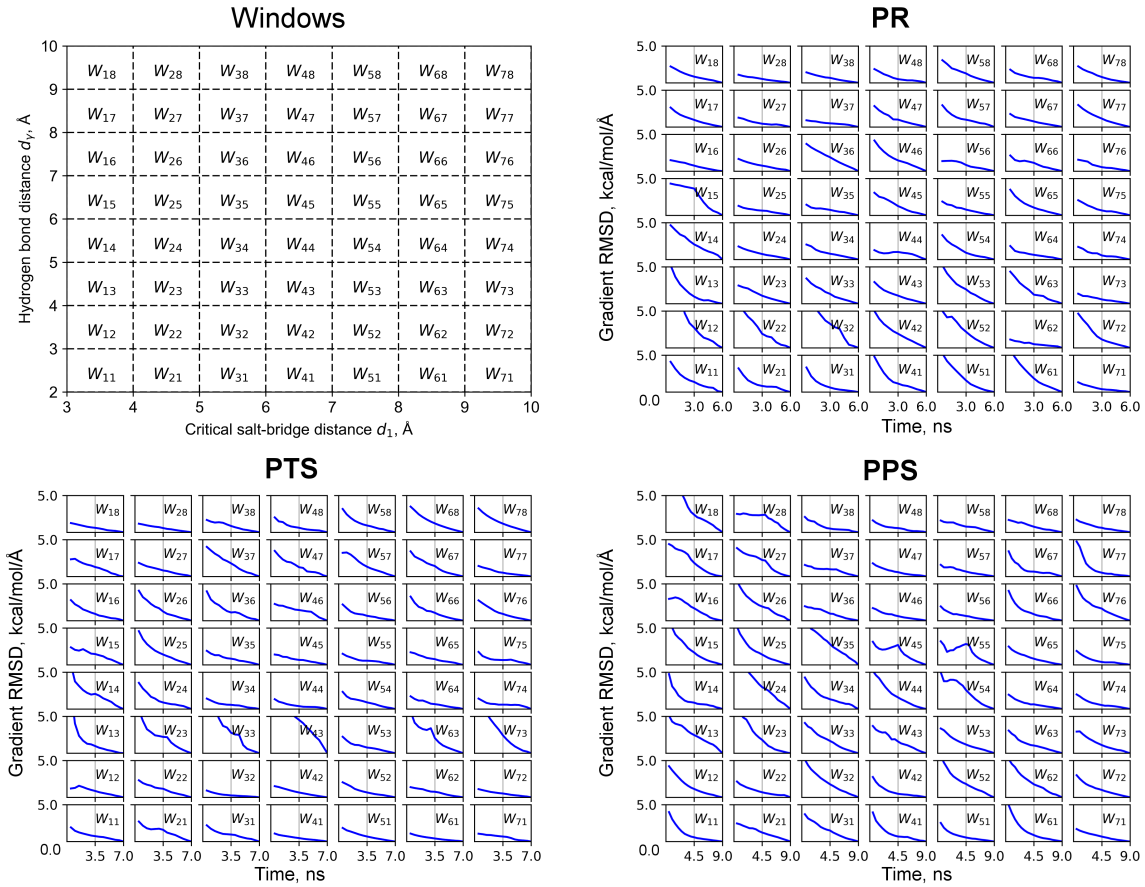
Supplementary Figure S8: Evolution of critical distances for switch II closure during unbiased MD. A. Distance d_1 between R205CZ and E461CD (critical salt-bridge). **B.** Distance d_γ between G459N (on switch II) and the closest oxygen atom of ATP γ-phosphate (or Pi in the case of PPS+ADP.Pi). For clarity the 5 ns running average is shown. The color code is the same as in Supplementary Figure S1. Strikingly, the partial re-priming of the converter in the PTS+ATP simulation at t=50 ns has no detectable effect on these important interactions in the active site, suggesting that the converter and the active site are uncoupled at this stage of the recovery stroke.



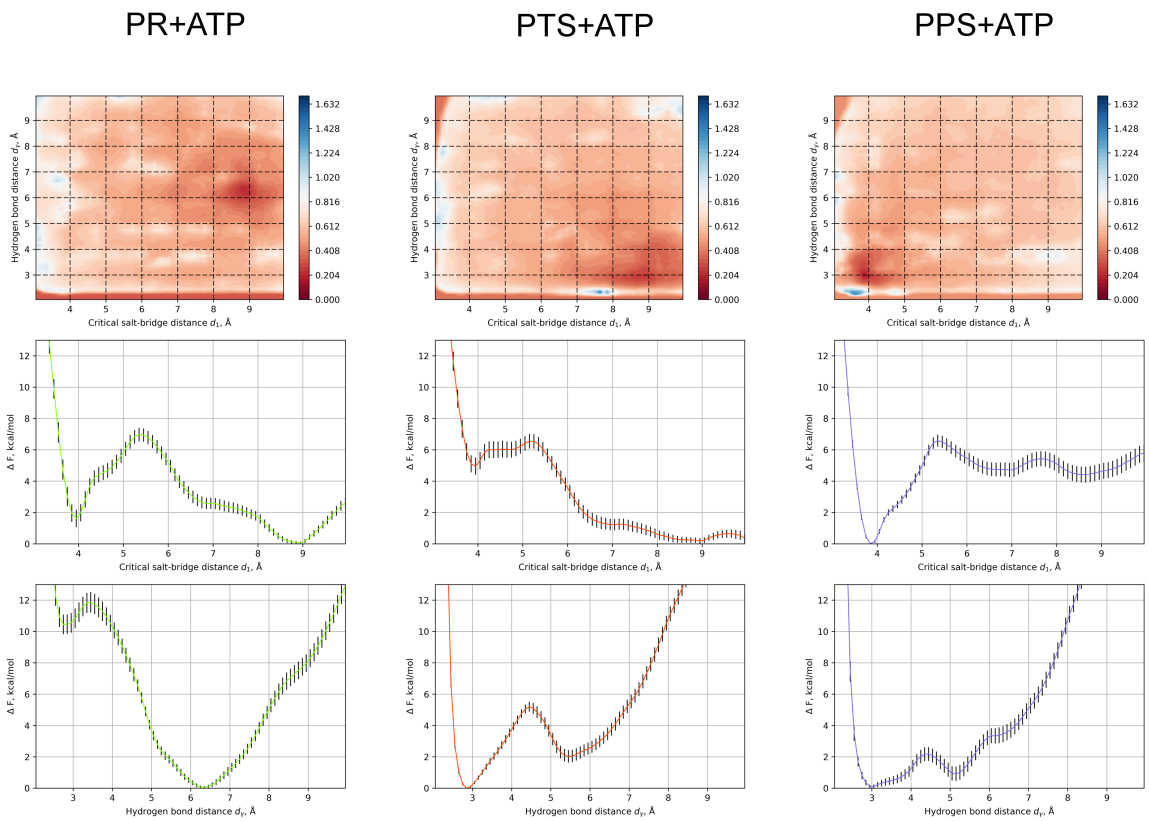
Supplementary Figure S9: State-dependent 1D potentials of mean force of switch II closure in the myosin VI motor domain. A. Close-up on the myosin VI active site and definition of the reaction coordinates probed by the ABF calculations. The open (PR) and closed (PPS) positions of switch II are shown. **B.** Potential of mean force for the formation of the G459-ATP hydrogen bond. **C.** Potential of mean force for the formation of the critical salt-bridge R205-E461. All free energies are given in kcal/mol. One-dimensional PMFs are computed by Boltzmann integration from the two-dimensional free energy profiles, see Supplementary Text 4.



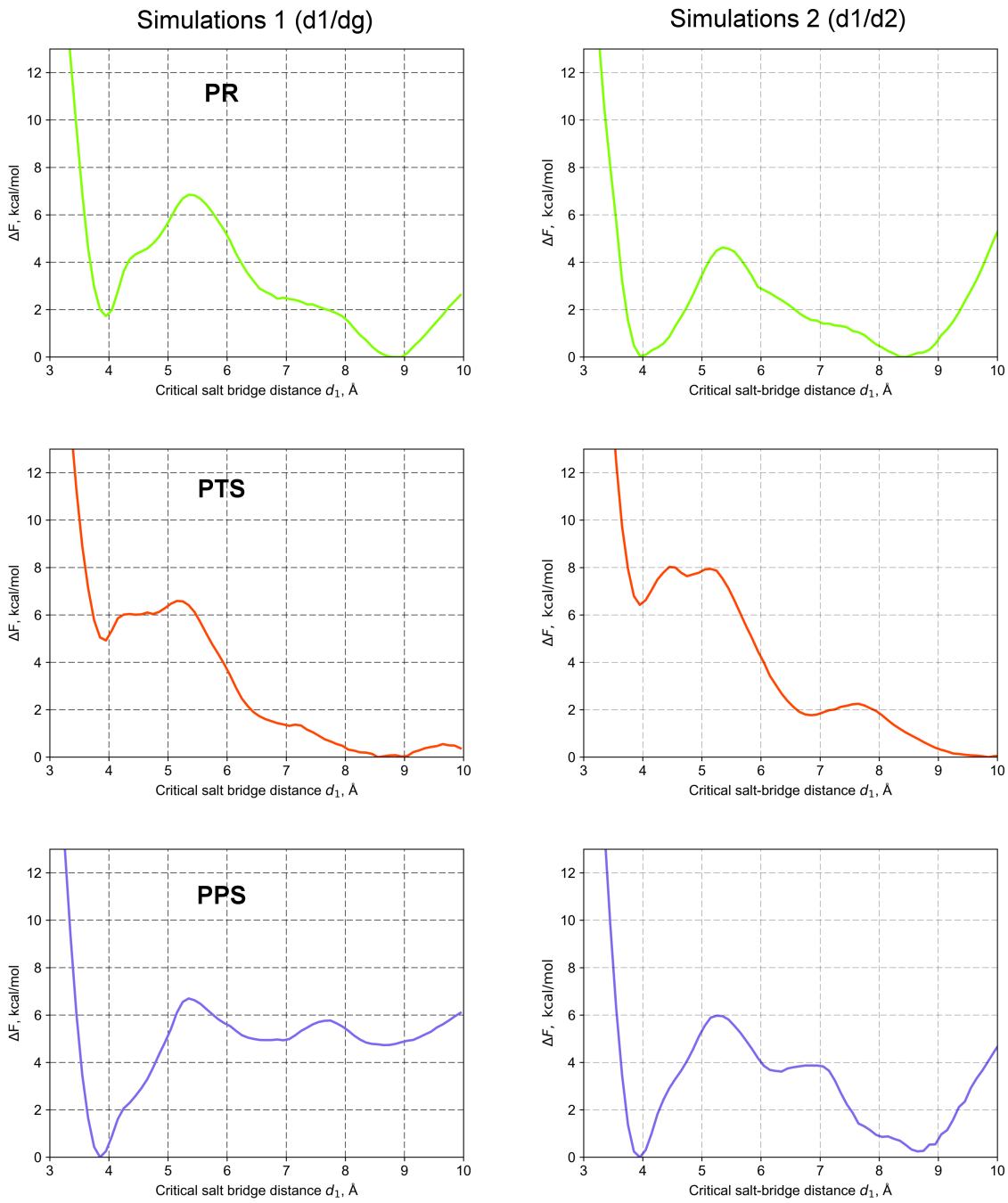
Supplementary Figure S10: Sampling of the configurational space in stratified ABF calculations. The degree of sampling (i.e. the number of times a given grid point is visited in the simulation) over the full 2D grid is represented on a logarithmic scale. The results demonstrate that the configurational space of interest is significantly sampled in all 3 simulations, supporting the relevance of the corresponding free energy surfaces in Figure 4, Main Text.



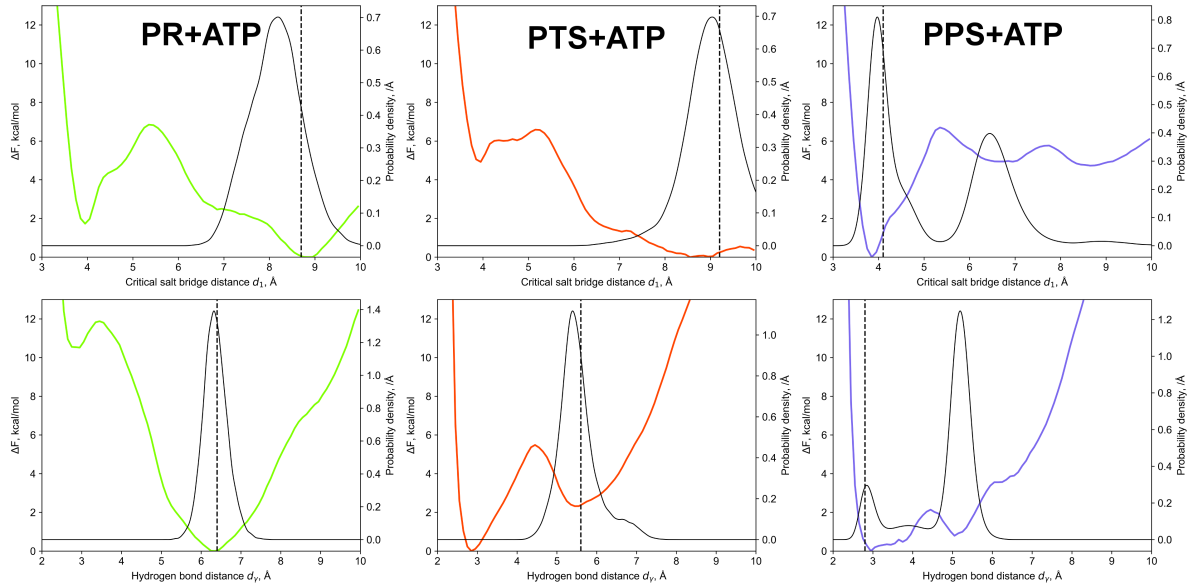
Supplementary Figure S11: Convergence of the per-window free energy gradient RMSD per window in stratified ABF calculations. For each window, the root-mean-square deviation (RMSD) of the gradient estimate with respect to its final value is shown as a function of time. For windows bordering the lower bounds of the collective variable, a 0.5 Å margin from the border was excluded from the RMSD calculation; in these regions the free energy gradient probed by ABF corresponds to the very steep inter-atomic repulsion forces, yielding artificially large gradient variations that are physically irrelevant.



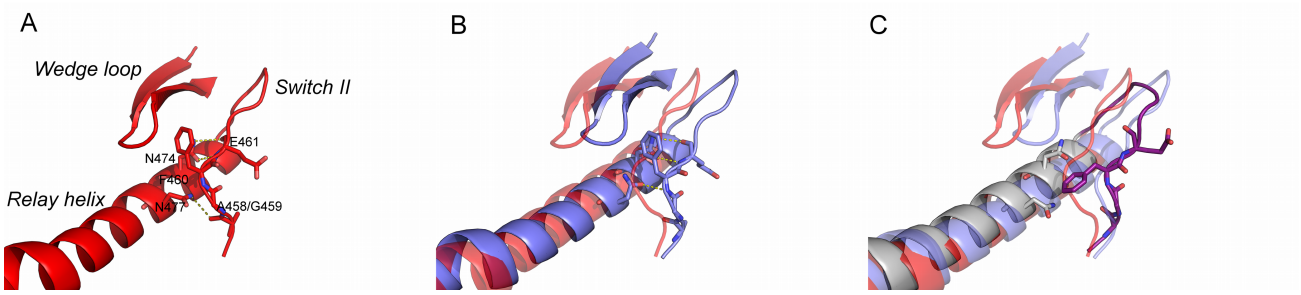
Supplementary Figure S12: Error analysis of the ABF results by Gaussian perturbation. Upper panel: the standard deviation of the 2D free energy profiles is shown. Lower panel: the standard deviation of the 1D free energy profiles along d_1 and d_γ are shown. All free energy values are given in kcal/mol. See Supplementary Text 4.



Supplementary Figure S13: Comparison of one-dimensional PMFs obtained with two different auxiliary reaction coordinates. See Supplementary Text 4.



Supplementary Figure S14: Comparison of unbiased MD with ABF results. The 1D potentials of mean force obtained by ABF (in colors) are compared with the statistical distributions of the corresponding observables obtained from unbiased Molecular Dynamics (black curves) and X-ray structures (black dotted lines). The results show that unbiased simulations primarily explore regions identified as free energy basins by ABF, supporting the consistency between these two complementary simulation approaches. Regarding d_γ, in PTS and PPS, the system is found to spend more time in the least stable state (broken hydrogen bond). In PTS, this may be explained by considering that the minimum free energy path to form the hydrogen bond goes through a metastable intermediate corresponding to a semi-open critical salt-bridge (see Figure 4B, the local free energy minimum at d_1~6.8 Å and d_γ~5.5 Å) which is not sampled by the PTS unbiased MD. In PPS, the instability of the hydrogen bond between switch II and ATP in unbiased simulations is striking and may be due to the corresponding crystal structure being solved with an ADP.Pi rather than an ATP analogue.



Supplementary Figure S15: Breaking of the switch II – Relay helix/L50 interactions upon switch II uncoupling in the ABF simulation of the PTS structure. A. Interactions between switch II and the Relay helix (N477-A458/G459 and N474-E461) and between switch II and the L50 subdomain (F460-wedge loop) in the PTS crystal structure. **B.** In PPS (blue), these interactions are maintained despite the closure of switch II thanks to the seesaw motion of the Relay helix and an inward rotation of the wedge loop; these movements are apparent upon comparison with the PTS (transparent red). **C.** In the "uncoupled switch II" state sampled by ABF simulations (Relay helix in grey, switch II in purple), these interactions are broken. Notably the side chain of F460 flips and is extracted from the pocket formed by the wedge loop. The conformation is the same as one Figure 4D of Main Text; PTS (transparent red) and PPS (transparent blue) crystal structures are represented for easier comparison.

Supplementary Tables

Supplementary Table S1: Data collection and refinement statistics for the PTS crystal structure

Crystal	Myosin VI PTS
Data collection	
Beamline	ESRF ID23-1
Space Group	$P2_12_12_1$
Unit cell *a*,*b*,*c* [Å] *α*, *β*, *γ* [°]	72.34, 83.83, 177.66
Molecules per asymmetric unit	1
Resolution [Å]	30.0 – 2.20
R_{sym} [%]	6.3 (113.4)
$I/\sigma I$	15.57 (1.29)
Completeness [%]	99.6 (99.1)
Redundancy	5.1 (5.1)
Refinement	
Resolution [Å]	22.18-2.20 (2.26-2.20)
No. of reflections	55469
R_{work}/R_{free} [%]	18.23/22.68 (26.05-31.94)
Average B-factor [Å^2]	67.33
r.m.s.d. bond lengths [Å], angles [°]	0.010/1.06
PDB ID	**5O2L**

Supplementary Table S2: Contacts between the converter and N-terminal subdomains in various structures of the motor domain of myosin VI

A contact is considered formed when two atoms are within 4.5 Å of each other.

State	**Post-rigor** (2VAS)	**Pre-transition State** (5O2L)	**PTS-reprimed**	**Pre-powerstroke** (2V26)
	N-ter aa	N-ter aa	N-ter aa	N-ter aa
Converter (F705-I773)				
P706	-	C63 (1)	A91-N92 (2-8)	A91-N92 (2-4)
S707	L65-M66 (1-2)	-	-	-
R708	C63-S64-A91 (12-1-1)	-	-	-
E713	S64 (6)	-	-	-
M717	S64 (1)	-	-	-
Y718	M66 (1)	-	-	-
R759	-	-	D61-C63-S64-A91 (9-4-2-2)	-
P760	-	C63 (8)	L120-G121 (2-6)	D61-C63 (1-3)
G761	-	C63 (1)	G121 (3)	-
F763	S64-L65-M66 (4-4-9)	-	-	N92-S119-L120 (14-3-8)
A764	M66-Y67 (6-4)	-	S119-G121 (1-2)	S11-L120-T122 (6-1-1)
D767	E53-M66 (1-8)	-	-	S119-R136 (8-2)
Q768	E53 (4)	-	-	-
Overall	65 contacts	10 contacts	41 contacts	53 contacts

Supplementary Table S3: Contacts between the converter last helix and the SH1 helix in various structures of the motor domain of myosin VI

State	**Post-rigor** (2VAS)	**Pre-transition State** (5O2L)	**PTS-reprimed**	**Pre-powerstroke** (2V26)
	SH1 aa	SH1 aa	SH1 aa	SH1 aa
Converter last helix (F759-I773)				
R759	G704 (7)	Q702-G703-G704 (4-8-4)	G704 (1)	M701-Q702-G703-G704 (1-3-12-10)
P760	G704 (3)	M701-Q702-G704 (2-1-3)	-	M701-G704 (7-2)
Overall	10 contacts	22 contacts	1 contact	35 contacts

Supplementary Table S4: List of Molecular Dynamics (MD) simulations

See Supplementary Text 4 for details on the simulations.

Simulation	**Algorithm**	**Length**	**Starting conformation**
PR+ATP (1)	Conventional MD	101 ns	2VAS + ATP
PR+ATP (2)	Conventional MD	100 ns	2VAS + ATP
PR+ATP (3)	Conventional MD	200 ns	2VAS + ATP
PTS+ATP (1)	Conventional MD	306 ns	5O2L + ATP
PTS+ATP (2)	Conventional MD	100 ns	5O2L + ATP
PTS+ATP (3)	Conventional MD	100 ns	5O2L + ATP
PPS+ATP (1)	Conventional MD	100 ns	2V26 + ATP
PPS+ATP (2)	Conventional MD	100 ns	2V26 + ATP
PPS+ATP (3)	Conventional MD	100 ns	2V26 + ATP
PPS+ADP.Pi (1)	Conventional MD	101 ns	2V26 + ADP.Pi
PPS+ADP.Pi (2)	Conventional MD	100 ns	2V26 + ADP.Pi
PR+ATP	ABF	76.3 ns + 6 ns/window	2VAS + ATP
PTS+ATP	ABF	68 ns + 7 ns/window	5O2L + ATP
PPS+ATP	ABF	68 ns + 9 ns/window	2V26 + ATP

Supplementary References

1. Fischer S, Windshügel B, Horak D, Holmes KC, Smith JC (2005) Structural mechanism of the recovery stroke in the Myosin molecular motor. *Proc Natl Acad Sci U S A* 102(19):6873–6878.

2. Koppole S, Smith JC, Fischer S (2007) The Structural Coupling between ATPase Activation and Recovery Stroke in the Myosin II Motor. *Structure* 15(7):825–837.

3. Kintses B, Yang Z, Málnási-Csizmadia A (2008) Experimental Investigation of the Seesaw Mechanism of the Relay Region That Moves the Myosin Lever Arm. *J Biol Chem* 283(49):34121–34128.

4. Mesentean S, Koppole S, Smith JC, Fischer S (2007) The Principal Motions Involved in the Coupling Mechanism of the Recovery Stroke of the Myosin Motor. *J Mol Biol* 367(2):591–602.

5. Koppole S, Smith JC, Fischer S (2006) Simulations of the Myosin II Motor Reveal a Nucleotide-state Sensing Element that Controls the Recovery Stroke. *J Mol Biol* 361(3):604–616.

6. Woo H-J (2007) Exploration of the conformational space of myosin recovery stroke via molecular dynamics. *Biophys Chem* 125(1):127–137.

7. Harris MJ, Woo H-J (2008) Energetics of subdomain movements and fluorescence probe solvation environment change in ATP-bound myosin. *Eur Biophys J* 38(1):1–12.

8. Yu H, Ma L, Yang Y, Cui Q (2007) Mechanochemical Coupling in the Myosin Motor Domain. I. Insights from Equilibrium Active-Site Simulations. *PLoS Comput Biol* 3(2):e21.

9. Yu H, Ma L, Yang Y, Cui Q (2007) Mechanochemical Coupling in the Myosin Motor Domain. II. Analysis of Critical Residues. *PLoS Comput Biol* 3(2):e23.

10. Elber R, West A (2010) Atomically detailed simulation of the recovery stroke in myosin by Milestoning. *Proc Natl Acad Sci* 107(11):5001–5005.

11. Baumketner A, Nesmelov Y (2011) Early stages of the recovery stroke in myosin II studied by molecular dynamics simulations. *Protein Sci* 20(12):2013–2022.

12. Baumketner A (2012) Interactions between relay helix and Src homology 1 (SH1) domain helix drive the converter domain rotation during the recovery stroke of myosin II. *Proteins Struct Funct Bioinforma* 80(6):1569–1581.

13. Baumketner A (2012) The mechanism of the converter domain rotation in the recovery stroke of myosin motor protein. *Proteins Struct Funct Bioinforma* 80(12):2701–2710.

14. Murphy CT, Rock RS, Spudich JA (2001) A myosin II mutation uncouples ATPase activity from motility and shortens step size. *Nat Cell Biol* 3(3):311.

15. Patterson B, Ruppel KM, Wu Y, Spudich JA (1997) Cold-sensitive Mutants G680V and G691C ofDictyostelium Myosin II Confer Dramatically Different Biochemical Defects. *J Biol Chem* 272(44):27612–27617.

16. Batra R, Geeves MA, Manstein DJ (1999) Kinetic Analysis of *Dictyostelium discoideum* Myosin Motor Domains with Glycine-to-Alanine Mutations in the Reactive Thiol Region †. *Biochemistry (Mosc)* 38(19):6126–6134.

17. Sasaki N, Shimada T, Sutoh K (1998) Mutational Analysis of the Switch II Loop ofDictyostelium Myosin II. *J Biol Chem* 273(32):20334–20340.

18. Tsiavaliaris G, Fujita-Becker S, Batra R, Levitsky DI, Kull FJ, Geeves MA, Manstein DJ (2002) Mutations in the relay loop region result in dominant-negative inhibition of myosin II function in Dictyostelium. *EMBO Rep* 3(11):1099–1105.

19. Sirigu S, Hartman JJ, Planelles-Herrero VJ, Ropars V, Clancy S, Wang X, Chuang G, Qian X, Lu P-P, Barrett E, Rudolph K, Royer C, Morgan BP, Stura EA, Malik FI, Houdusse AM (2016) Highly selective inhibition of myosin motors provides the basis of potential therapeutic application. *Proc Natl Acad Sci*:201609342.

20. Webb B, Sali A (2002) Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics* (John Wiley & Sons, Inc.). Available at: http://dx.doi.org/10.1002/0471250953.bi0506s47.

21. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB, Snoeyink J, Richardson JS, Richardson DC (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* 35(suppl 2):W375–W383.

22. Bashford D, Karplus M (1991) Multiple-site titration curves of proteins: an analysis of exact and approximate methods for their calculation. *J Phys Chem* 95(23):9556–9561.

23. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc Natl Acad Sci* 98(18):10037–10041.

24. Rabenstein B, Knapp E-W (2001) Calculated pH-Dependent Population and Protonation of Carbon-Monoxy-Myoglobin Conformers. *Biophys J* 80(3):1141–1150.

25. Kieseritzky G, Knapp E-W (2007) Optimizing pKA computation in proteins with pH adapted conformations. *Proteins Struct Funct Bioinforma* 71(3):1335–1348.

26. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M (2009) CHARMM: The biomolecular simulation program. *J Comput Chem* 30(10):1545–1614.

27. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. *J Mol Graph* 14(1):33–38.

28. Berendsen HJ., Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. *J Chem Phys* 81(8):3684–3690.

29. Ryckaert J-P, Ciccotti G, Berendsen HJ. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* 23(3):327–341.

30. Tuckerman M, Berne BJ, Martyna GJ (1992) Reversible multiple time scale molecular dynamics. *J Chem Phys* 97(3):1990.

31. Seeber M, Cecchini M, Rao F, Settanni G, Caflisch A (2007) Wordom: a program for efficient analysis of molecular dynamics simulations. *Bioinformatics* 23(19):2625–2627.

32. Seeber M, Felline A, Raimondi F, Muff S, Friedman R, Rao F, Caflisch A, Fanelli F (2011) Wordom: A user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces. *J Comput Chem* 32(6):1183–1194.

33. Hunter JD (2007) Matplotlib: A 2D Graphics Environment. *Comput Sci Eng* 9(3):90–95.

34. Perez F, Granger BE (2007) IPython: A System for Interactive Scientific Computing. *Comput Sci Eng* 9(3):21–29.

35. Jones E, Oliphant T, Peterson P (2001) {SciPy}: Open source scientific tools for {Python}. Available at: http://www.scipy.org [Accessed June 1, 2017].

36. Comer J, Gumbart JC, Hénin J, Lelièvre T, Pohorille A, Chipot C (2015) The Adaptive Biasing Force Method: Everything You Always Wanted To Know but Were Afraid To Ask. *J Phys Chem B* 119(3):1129–1151.

37. Wereszczynski J, McCammon JA (2012) Nucleotide-dependent mechanism of Get3 as elucidated from free energy calculations. *Proc Natl Acad Sci* 109(20):7759–7764.

Florian BLANC

Exploration de la transduction chimio-mécanique chez le moteur moléculaire myosine par simulations numériques

Résumé

La vie repose sur des conversions d'énergie libre assurées par des machines moléculaires. Parmi elles, le moteur moléculaire myosine couple l'hydrolyse de l'ATP à la production de force sur l'actine par basculement d'un « bras de levier ». Compléter le cycle requiert une étape de régénération, ou *recovery stroke,* où le moteur retourne dans sa configuration armée et hydrolyse l'ATP, ce qui est crucial pour la transduction chimio-mécanique. Cette thèse étudie le mécanisme du *recovery stroke* par des simulations moléculaires. Partant d'une nouvelle structure cristallographique de la myosine VI, nous proposons un mécanisme original pour la transition dans lequel la remise en place du bras de levier n'est que faiblement couplée à l'activation de l'ATPase. En fait, nos calculs suggèrent qu'elle est déclenchée par les fluctuations thermiques de manière *ratchet-like,* et en contradiction avec des modèles précédents prédisant un couplage fort. Nos résultats suggèrent comment les moteurs moléculaires pourraient exploiter les fluctuations conformationnelles spontanées pour produire du travail dans un environnement isotherme.

Résumé en anglais

Life relies on free energy conversions performed by molecular machines. Among them, the myosin molecular motor couples the hydrolysis of ATP to force production on actin through a swing of a « lever-arm ». Completing the cycle requires a regeneration step, the recovery stroke, in which the motor returns to its armed configuration and hydrolyzes ATP, which makes it crucial for chemo-mechanical transduction. In this thesis, we investigate the mechanism of the recovery stroke using molecular simulations. Capitalizing on a new crystal structure of myosin VI, we propose an original mechanism for the transition in which the re-priming of the lever arm is only loosely coupled to ATPase activation. Rather, our calculations suggest it is driven by thermal fluctuations in a ratchet-like manner, as opposed to previous models predicting strong coupling. Our results hint at how molecular motors may exploit spontaneous conformational fluctuations to produce work in an isothermal environment.