



Sorbonne Université

École doctorale 397 - Physique et Chimie des Matériaux
Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie (IMPMC)

Development of predictive approaches for biomolecular association kinetics

Thèse de Doctorat de Physique

Par Karen Palacio-Rodriguez

Dirigée par

Fabio Pietrucci & Alessandro Barducci

Présentée et soutenue publiquement le 23/09/2022

Devant un jury composé de :

| | | |
|-----------------------------|----------------------|--|
| <i>Présidente du jury :</i> | Rodolphe Vuilleumier | École Normale Supérieure, Paris |
| <i>Rapporteurs :</i> | Edina Rosta | University College London, London |
| | Jérôme Hénin | Institut de Biologie Physico-Chimique, Paris |
| <i>Directeur :</i> | Fabio Pietrucci | Sorbonne Université, Paris |
| <i>Co-directeur :</i> | Alessandro Barducci | Centre de Biologie Structurale, Montpellier |
| <i>Examinateurs :</i> | Isabelle Callebaut | Sorbonne Université, Paris |
| | A. Marco Saitta | Sorbonne Université, Paris |

2022

 Except where otherwise noted, this work is licensed under
<http://creativecommons.org/licenses/by-nc-nd/3.0/>

To my family.

Acknowledgments

To the Doctoral School ED 397 Physique et Chimie des Matériaux and Sorbonne Université for the funding and support to this doctoral project. To the Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie (IMPMC) for the support of the Ph.D. project and the doctoral missions. To the members of the jury for kindly accepting the invitation to review this thesis

I am grateful to my advisor, Fabio Pietrucci, for his constant support, his careful guidance, for inspiring me with his example to do quality and rigorous science, but above all for always finding the right words to explain the most complex concepts in the clearest way; my journey through the Ph.D. would not have been so pleasant without the guidance of such a good teacher. To my co-advisor, Alessandro Barducci, for sharing with me his knowledge, his patience, his guidance, and his support beyond the Ph.D. project requirements that motivated me to continue on the academic path. A special thank is due to Pilar Cossío, my former master's supervisor, for helping me make connections with the best in the field, her support in getting this Ph.D. thesis, and for her continued support throughout this project both personally and intellectually.

I thank the members of the IMPMC and the theory section of the PHYSIX team for the constant feedback, especially to Hadrien Vroylandt for his advice on the theory challenges encountered during the development of this thesis. To all the collaborators of the project, for enriching my work and my career with their knowledge.

I also thank to my friends in academia at Sorbonne Université, University of Antioquia and those that are spread throughout the world for the intellectual and emotional support: a Ph.D. project can be challenging but with the right company it is very enjoyable.

I would like to thank to my family, for being my continual motivation. Finally, I must express my very profound gratitude to Julián Mejía, for everything.

Abstract

Atomistic computer simulations of rare events have three paramount goals: predicting detailed mechanisms, free energy landscapes, and kinetic rates of transformation processes like phase transitions, chemical reactions, biomolecular folding or association. In real-life applications, all of these tasks are cumbersome and require intensive human and computer effort, especially the calculation of rates. The difficulty resides in the gap between the long time scale associated to such processes, also known as rare events, and the short time scale that is accessible by molecular dynamics simulations. Enhanced sampling techniques can accelerate the exploration of high-free energy regions, adding external forces to the system to pull it out of free energy basins or focusing sampling on the transition region and efficiently exploring transition paths. These techniques allow to reconstruct mechanisms and free energy landscapes for a wide range of activated processes in physics, chemistry and biology. Methods aimed at accurate kinetic rates are at present less mature and still require large computational effort and/or rely on ideal collective variables.

We developed two efficient methodologies for the prediction of transition rates from molecular dynamics simulations in combination with enhanced sampling techniques. Both strategies only require sets of short simulations, which allows exploiting the parallel capabilities of current supercomputers. On one side, we use metadynamics, a widely used enhanced sampling technique that adds a time-dependent bias potential to the system, disrupting its dynamics. We overcome this limitation by developing a method based on Kramers' theory for calculating the barrier-crossing rate when a time-dependent bias is added to the system. We tested this method in a benchmark system and apply it to complex all-atoms simulations, showing that we are able to extract the rate and measure at the same time the quality of the collective variables for processes where Kramers' theory holds. On the other side, transition path sampling trajectories are the golden standard to access mechanistic information: we demonstrate that they also encode accurate thermodynamic and kinetic information, that can be extracted by training a data-driven overdamped Langevin model of the dynamics projected on a collective variable. We also tested this method over benchmark systems to establish a validation criteria for the accurate time resolutions that yields markovian behavior and apply it to complex all-atoms simulations to recover free energies, position-dependent diffusion coefficients, and rates. Overall, these new theoretical tools that can be freely downloaded from GitHub make efficient use of computing resources providing simple procedures to accurately predict kinetic rates and could be suitable for applications far beyond the field of biomolecular association.

Contents

| | |
|---|-----------|
| DEDICATION | 3 |
| ACKNOWLEDGMENTS | 4 |
| ABSTRACT | 5 |
| I Background | 15 |
| 1 RÉSUMÉ | 16 |
| 2 INTRODUCTION | 24 |
| 2.1 On the importance of kinetic rates estimation | 24 |
| 2.2 Kinetic rates from atomistic computer simulations | 25 |
| 2.2.1 Molecular dynamics | 25 |
| 2.2.2 Enhanced sampling methods | 26 |
| 2.3 Open challenges | 27 |
| 2.4 Our proposal | 28 |
| II Theory and Methods | 30 |
| 3 KINETIC RATES ESTIMATION | 31 |
| 3.1 Macroscopic perspective | 31 |
| 3.2 Direct calculation of the rate constant | 33 |
| 3.3 Estimation of the mean first-passage time | 33 |
| 3.3.1 The Kramers' problem | 33 |
| 3.3.2 Mean lifetime method | 34 |
| 3.4 Transition state theory | 36 |
| 3.5 Bennett-Chandler method | 37 |
| 4 ATOMISTIC COMPUTER SIMULATIONS | 40 |
| 4.1 Background | 40 |
| 4.2 Classical molecular dynamics simulations | 41 |

| | | |
|------------|---|-----------|
| 4.2.1 | Time evolution and statistical principles | 41 |
| 4.2.2 | Integration of equation of motion | 43 |
| 4.2.3 | Force fields | 44 |
| 4.2.4 | Temperature coupling | 47 |
| 4.2.5 | Pressure coupling | 48 |
| 4.2.6 | Other practical considerations | 50 |
| 4.3 | Following transformations in reduced dimensionality | 52 |
| 4.3.1 | Collective variables, order parameters and reaction coordinate | 52 |
| 4.3.2 | Quality of collective variables | 52 |
| 4.3.3 | The committor | 54 |
| 4.3.4 | Heuristic collective variables | 54 |
| 4.3.5 | Path collective variables | 55 |
| 4.3.6 | Dimensionality reduction and machine learning based classification . | 56 |
| 5 | ENHANCED SAMPLING METHODS | 57 |
| 5.1 | Background | 57 |
| 5.2 | Biased enhanced sampling methods | 58 |
| 5.2.1 | Umbrella sampling | 59 |
| 5.2.2 | Metadynamics | 60 |
| 5.3 | Unbiased enhanced sampling methods | 61 |
| 5.3.1 | Transition path sampling | 61 |
| III | Results and Discussion | 65 |
| 6 | TRANSITION RATES AND EFFICIENCY OF COLLECTIVE VARIABLES FROM TIME-DEPENDENT BIASED SIMULATIONS | 66 |
| 6.1 | Introduction | 66 |
| 6.1.1 | Kinetic rate estimation from time-dependent biased simulations . . . | 67 |
| 6.1.2 | Transition rates from MetaD and current limitations | 67 |
| 6.1.3 | Our proposal | 67 |
| 6.2 | Theoretical methods | 68 |
| 6.2.1 | Kramers time-dependent rate | 68 |
| 6.2.2 | Rate-acceleration factor: efficiency of CVs | 69 |
| 6.2.3 | Survival probability from metadynamics simulations | 72 |
| 6.2.4 | Maximization of the likelihood function | 73 |
| 6.2.5 | Practical considerations | 74 |
| 6.2.6 | Simulation setup for the 2D double-well potential | 74 |
| 6.2.7 | Rate from infrequent metadynamics | 75 |
| 6.3 | Results and discussion | 76 |
| 6.3.1 | Benchmark system: 2D double-well potential | 76 |
| 6.3.2 | Escape rates from Kramers Time-dependent Rate vs Infrequent Metadynamics | 77 |

| | | |
|----------|--|------------|
| 6.3.3 | Cumulative distribution functions | 77 |
| 6.3.4 | Collective variable efficiency | 77 |
| 6.4 | Conclusions and outlook | 78 |
| 7 | KINETIC RATES FROM METADYNAMICS: APPLICATIONS | 80 |
| 7.1 | Introduction | 80 |
| 7.2 | Interaction of fullerene C ₂₄₀ dimer in water solution | 80 |
| 7.2.1 | Description of the system | 80 |
| 7.2.2 | Simulation methods | 81 |
| 7.2.3 | Results and discussion | 82 |
| 7.3 | Protein-ligand unbinding: CDK2 protein-CS3 ligand | 83 |
| 7.3.1 | Description of the system | 83 |
| 7.3.2 | Simulation methods | 84 |
| 7.3.3 | Results and discussion | 86 |
| 7.4 | Conclusions and outlook | 89 |
| 8 | FREE ENERGY LANDSCAPES, DIFFUSION COEFFICIENTS AND KINETIC RATES FROM DATA-DRIVEN LANGEVIN MODELS | 91 |
| 8.1 | Introduction | 91 |
| 8.1.1 | Langevin models from MD trajectories | 92 |
| 8.1.2 | Challenges in modelling real-systems dynamics with the Langevin equation | 93 |
| 8.1.3 | Our proposal | 93 |
| 8.2 | Theoretical methods | 93 |
| 8.2.1 | Mathematical models of projected dynamics | 93 |
| 8.2.2 | Optimization of Langevin models (parameter estimation) | 95 |
| 8.2.3 | Integration of the Langevin equation of motion | 97 |
| 8.2.4 | Optimal time resolution of the reduced model | 98 |
| 8.2.5 | Kinetic rates | 99 |
| 8.3 | Results and discussion | 100 |
| 8.3.1 | Langevin models from overdamped trajectories | 100 |
| 8.3.2 | Langevin models from underdamped trajectories | 101 |
| 8.4 | Conclusions and outlook | 104 |
| 9 | KINETIC RATES FROM DATA-DRIVEN LANGEVIN MODELS: APPLICATIONS | 106 |
| 9.1 | Introduction | 106 |
| 9.2 | Interaction of C ₆₀ and C ₂₄₀ fullerene dimer in water solution | 106 |
| 9.2.1 | Description of the system | 106 |
| 9.2.2 | Simulation methods | 107 |
| 9.2.3 | Results and discussion | 108 |
| 9.3 | Crystallization of Germanium-Telluride | 112 |
| 9.3.1 | Description of the system | 112 |
| 9.3.2 | Simulation methods | 113 |

| | | |
|-----------|--|------------|
| 9.3.3 | Results and discussion | 114 |
| 9.4 | Protein-protein interactions in barnase-barstar complex | 116 |
| 9.4.1 | Description of the system | 116 |
| 9.4.2 | Simulation methods | 117 |
| 9.4.3 | Results and discussion | 119 |
| 9.5 | Conclusions and outlook | 122 |
| 10 | COLLECTIVE VARIABLE OPTIMIZATION USING DATA-DRIVEN LANGEVIN MODELS | 124 |
| 10.1 | Introduction | 124 |
| 10.2 | Theory and methods | 125 |
| 10.2.1 | Variational principle for rate minimization | 125 |
| 10.2.2 | Algorithm for collective variable optimization | 125 |
| 10.2.3 | Simulation methods | 126 |
| 10.3 | Results and discussion | 128 |
| 10.3.1 | Benchmark system: double-well potential | 128 |
| 10.3.2 | Application: fullerene dimers | 130 |
| 10.4 | Conclusions and outlook | 132 |
| IV | Conclusions and Perspectives | 134 |
| 11 | CONCLUSIONS AND PERSPECTIVES | 135 |
| V | Appendix | 138 |
| | APPENDIX A SUPPLEMENTARY INFORMATION: KINETICS FROM METADYNAMICS | 139 |
| | APPENDIX B SUPPLEMENTARY INFORMATION: KINETICS FROM LANGEVIN MODELS | 142 |
| | APPENDIX C DISSEMINATION OF RESEARCH RESULTS | 147 |
| VI | References | 150 |

List of figures

| | | |
|-----|---|-----|
| 3.1 | Schematic illustration of the Kramers and Bennett-Chandler approaches | 35 |
| 4.1 | Schematic representation of the terms involved in the force field | 45 |
| 4.2 | Schematic illustration of the CV quality | 53 |
| 5.1 | Schematic illustration of the time scales issue in MD simulations | 58 |
| 5.2 | Schematic representation of the umbrella sampling and metadynamics methods | 59 |
| 5.3 | Schematic representation of the transition path sampling method | 62 |
| 5.4 | Comparison between aimless shooting algorithms | 64 |
| 6.1 | Kramers time-dependent rate for biased simulations and assessment of the efficiency of the CVs | 68 |
| 6.2 | Cusp-like harmonic double-well potential | 70 |
| 6.3 | Illustration of the efficiency of the CVs on 2D double-well potential | 72 |
| 6.4 | MetaD Monte Carlo simulations on a 2D double-well potential. | 76 |
| 6.5 | CDFs of simulation and rescaled jump times and fits to theoretical distributions for the 2D double-well simulations | 78 |
| 7.1 | Fullerene C ₂₄₀ dimer representation and PMF | 81 |
| 7.2 | Fullerene dimer dissociation biased simulations | 83 |
| 7.3 | CDK2 biological function | 84 |
| 7.4 | CDK2-ligand in binding pose | 85 |
| 7.5 | CDK2-ligand CVs as a function of time | 87 |
| 7.6 | Multi-state dynamics in the CDK2 simulations | 88 |
| 7.7 | CDK2-ligand unbinding biased simulations | 89 |
| 8.1 | Schematic illustration of the approach proposed to optimize Langevin models of projected dynamics | 97 |
| 8.2 | Optimal overdamped Langevin models trained on overdamped Langevin trajectories | 101 |
| 8.3 | Input trajectories generated with a non-overdamped Langevin integrator for double-well benchmark system | 102 |
| 8.4 | Optimal overdamped Langevin models for double-well benchmark system . . | 102 |

| | | |
|------|---|-----|
| 9.1 | Characteristic atomic configurations for the C ₆₀ dimer and initial input MD trajectories | 109 |
| 9.2 | Example of the convergence of $-\log \mathcal{L}(\theta)$ | 110 |
| 9.3 | Optimal Langevin models for C ₆₀ and C ₂₄₀ fullerene dimers | 111 |
| 9.4 | MFPTs computed from the optimal Langevin models of fullerene dimers | 111 |
| 9.5 | Langevin models from association trajectories: fullerene dimer | 112 |
| 9.6 | Amorphous to crystalline phase transition of the GeTe | 113 |
| 9.7 | Trajectories for the optimization of the Langevin models: GeTe crystallization | 114 |
| 9.8 | Optimal Langevin models for the GeTe crystallization | 115 |
| 9.9 | Crystallographic structure of the barnase-barstar complex | 117 |
| 9.10 | Representative frames from the initial and final states of the Barnase-Barstar simulations | 119 |
| 9.11 | Barnase-barstar association trajectories projected over different CVs | 120 |
| 9.12 | Optimal Langevin models for the barnase-barstar association | 121 |
| 10.1 | 2D double-well benchmark system for CV optimization | 128 |
| 10.2 | Kinetic rate from CV optimization: 2D double well potential | 129 |
| 10.3 | Application of CV optimization to fullerene dimers | 130 |
| 10.4 | Kinetic rate from CV optimization: fullerene dimers | 131 |
| A1 | Average accumulated MetaD bias for simulations with different bias deposition time | 140 |
| A2 | γV_{MB} as a function of the normalized time t/t_d for the 2D double-well simulations | 140 |
| A3 | iMetaD results for the CDK2 | 141 |
| B1 | Effect of the inclusion of D(q) in the OLE | 143 |
| B2 | Autocorrelation function and distribution of the effective noise G for optimal Langevin models of different systems | 144 |
| B3 | Optimal overdamped Langevin models trained on underdamped Langevin trajectories | 145 |
| B4 | Diffusion profiles from umbrella sampling: fullerene dimers | 145 |
| B5 | Propagator accuracy test | 146 |
| B6 | Effect of the number of input trajectories | 146 |

List of Tables

| | | |
|-----|---|-----|
| 8.1 | Comparison between MFPTs estimated from Eq. 8.19 and from reactive flux | 104 |
| 9.1 | Brute force vs Langevin model based estimated of the MFPT: GeTe crystallization. | 115 |
| 9.2 | Langevin model based estimated of the dissociation rate: Barnase-barstar interaction. | 121 |
| B1 | Brute force vs reactive flux estimate of the MFPT. | 142 |

List of Abbreviations

| | |
|----------------|--|
| AS | Aimless Shooting |
| BB | Barnase-Barstar |
| CDF | Cumulative Distribution Function |
| CDKs | Cyclin-Dependent Kinases |
| CDK2 | Cyclin-Dependent Kinase 2 |
| CV | Collective Variable |
| ECDF | Empirical Cumulative Distribution Function |
| FES | Free Energy Surface |
| FF | Force Field |
| FFS | Forward Flux Sampling |
| GLE | Generalized Langevin Equation |
| KS-test | Kolmogorov-Smirnov test |
| KTR | Kramers' Time-dependent Rate |
| iMetaD | Infrequent Metadynamics |
| LJ | Lennard-Jones |
| MSM | Markov State Models |
| MC | Monte Carlo |
| MD | Molecular Dynamics |
| MFPT | Mean-First Passage Time |
| MetaD | Metadynamics |
| ML | Maching Learning |
| NNP | Neural Network Potential |

| | |
|-----------------|--|
| OLE | Overdamped Langevin Equation |
| OP | Order Parameter |
| PBC | Periodic Boundary Conditions |
| PCA | Principle Component Analysis |
| PDB | Protein Data Bank |
| PES | Potential Energy Surface |
| PIV | Permutation Invariant Vector |
| PME | Particle-Mesh Ewald |
| PMF | Potential of Mean Force |
| PPI | Protein-Protein Interactions |
| RC | Reaction Coordinate |
| RMSD | Root Mean Square Deviation |
| SLE | Standard Langevin Equation |
| TCDF | Theoretical Cumulative Distribution Function |
| TIS | Transition Interface Sampling |
| TPE | Transition Path Ensemble |
| TPS | Transition Path Sampling |
| TS | Transition State |
| TSE | Transition State Ensemble |
| TST | Transition State Theory |
| TST-BC | Bennett-Chandler approach |
| US | Umbrella Sampling |
| WHAM | Weighted Histogram Analysis Method |
| WT-MetaD | Well-Tempered Metadynamics |

Part I

Background

1

Résumé

Nous avons décidé d'écrire cette thèse en anglais, car tous les membres du jury ne sont pas francophones, mais aussi par soucis d'accessibilité au plus grand nombre. Ce choix est rendu possible par le règlement de l'école doctorale 397, sous réserve qu'un "résumé substantiel 5-10 pages" soit écrit en français. C'est ce résumé qui est présenté dans les pages suivantes, en suivant la structure du texte principal en anglais.

We chose to write this thesis in English as some of the jury's members are not francophone, and to make it accessible to a greater number. This choice is possible in accordance to the doctoral school 397, if a substantial french summary of 5-10 pages is provided. This summary is presented in the following, keeping the same structure as the main text in English.

La prédiction des vitesses cinétiques est essentielle pour comprendre les phénomènes physico-chimiques. Des processus tels que les transitions de phase, les réactions chimiques ou les interactions biomoléculaires et les changements de conformation sont caractérisés par un réseau d'états métastables souvent séparés par des barrières d'énergie libre élevées. Chaque état métastable correspond à un minimum local dans une surface d'énergie libre (FES) qui piège la dynamique du système pendant un temps long devant le temps intrinsèque des vibrations de liaison jusqu'à ce que le système puisse s'échapper via des fluctuations aléatoires vers un autre état métastable. Une barrière d'énergie libre élevée séparant deux états génère un goulot d'étranglement dynamique entre eux, rendant une transition entre bassins moins probable, par conséquent, ces transitions sont présentées comme des événements rares.

Les événements rares présentent une cinétique lente. Dans les exemples de processus physico-chimiques mentionné ci-dessus, une transition typique entre deux états intéressants, par exemple, la transition de la phase liquide à la phase solide, la conversion d'un réactif en produit, l'association protéine-ligand ou changement de conformation d'une protéine peut prendre des échelles de temps allant de quelques microsecondes à quelques secondes et au-

delà. L'estimation des taux de transition (c'est-à-dire l'inverse du temps de transition) nous permet de prédire les échelles de temps de ces événements, comprendre les mécanismes moléculaires et à un niveau plus pratique, prendre des décisions éclairées en fonction des propriétés cinétiques du système.

Puisque cette dynamique lente est typique des événements rares, l'importance de la cinétique ne se limite pas à la découverte de médicaments. Entre autres exemples, la prédiction des constantes de vitesse pour les transitions de phase des matériaux inorganiques joue un rôle clé dans la conception de nouveaux dispositifs avec des applications technologiques. Dans le domaine de la biologie moléculaire, l'étude des interactions protéines-protéines (PPI), le repliement des protéines et la cinétique d'agrégation des protéines aident à élucider les causes des maladies et créer des alternatives pour les traiter et les prévenir. Par conséquent, les méthodes permettant la prédiction précise des taux cinétiques ont une utilité large allant de la conception de matériaux à la thérapeutique. Dans ce qui suit, nous présenterons les méthodologies actuelles utilisées pour estimer les vitesses de réaction *in silico*.

Les simulations informatiques atomistiques d'événements rares ont trois objectifs primordiaux : prédire le mécanisme détaillé, calculer le paysage d'énergie libre et les taux de réaction. Toutes ces tâches sont, dans de nombreux cas, lourdes et nécessitent un effort humain et informatique intensif, le calcul des taux de réactions étant le plus difficile. Le défi du calcul des taux est donné par la complexité de simuler l'évolution d'un système dans le temps. Les propriétés dépendant du temps ne sont donc pas accessibles avec la plupart des méthodes de simulations informatiques atomistiques disponibles. Néanmoins, la dynamique moléculaire (MD) permet de suivre le comportement en fonction du temps d'un système en propageant les positions et les vitesses de tous les atomes du système à travers le solution des équations du mouvement. Les simulations MD ont été largement utilisées car elles permettent d'étudier à la fois le comportement thermodynamique et cinétique d'un système moléculaire.

Les simulations de dynamique moléculaire prédisent comment chaque atome d'un système moléculaire se déplacera dans le temps par intégration numérique des équations du mouvement. Positions et vitesses des atomes sont prédits selon un modèle d'interactions interatomiques, le champ de force. Les forces dans le système sont échantillonnées à chaque pas de temps en évaluant le champ de force sur les positions actuelles des atomes. Pour assurer la stabilité numérique dans l'intégration des équations de mouvement, un très petit pas de temps doit être utilisé (égal ou inférieur au temps caractéristique de la vibration la plus rapide). Puisque les simulations MD évaluent la position des atomes à chaque pas de temps, cette technique est un outil puissant pour élucider le mécanisme moléculaire à des résolutions atomiques et à des échelles de temps courtes difficilement accessibles par des techniques expérimentales. Le résultat des simulations MD est une collection de positions et de vitesses d'atomes dans le temps. Afin d'étudier en détail les mécanismes moléculaires, on peut se focaliser sur un ou quelques degrés de liberté du système. Par conséquent, on peut utiliser une variable collective (CV), c'est-à-dire un descripteur qui peut identifier les mouvements collectifs du système au cours du processus de transformation.

Puisque les CV permettent de suivre la progression du processus de transformation, avec les simulations MD, on pourrait estimer directement les taux cinétiques en mesurant le temps moyen que le système passe dans un certain état dans la limite d'un temps d'échantillonnage infini. D'autres approches théoriques visent à estimer la cinétique à partir de simulations MD en utilisant différents cadres mathématiques. D'un côté, la dynamique complète du système peut être projetée sur des CV de faible dimension conduisant à des équations différentielles stochastiques de Langevin. D'autre part, l'espace de configuration peut être partitionné en états discrets et les transitions entre les états peuvent être comptés pour construire des modèles d'états de Markov (MSM).

Bien que les simulations MD semblent être une voie naturelle pour étudier le comportement cinétique d'un système moléculaire, comme expliqué ci-dessus, les événements rares sont présents dans de nombreux processus moléculaires intéressants. Par conséquent, l'estimation des propriétés thermodynamiques et cinétiques à l'aide de MD implique l'évaluation des équations de mouvement des millions ou des milliards de fois. Cela conduit à un écart entre les échelles de temps des événements rares décrites précédemment et celles accessibles avec MD, même en utilisant de puissants supercalculateurs. Les méthodes d'échantillonnage avancées sont devenues des alternatives utiles pour accélérer l'échantillonnage en MD. Celles-ci ont été développées pour accélérer l'échantillonnage de l'espace conformationnel, qui se caractérise par des paysages accidentés avec de hautes barrières d'énergie libre.

Les méthodes d'échantillonnage avancées en combinaison avec les simulations MD sont devenues des alternatives utiles pour étudier les événements qui se produisent à de longues échelles de temps. Ces méthodes ont été développées pour accélérer l'échantillonnage de l'espace conformationnel des processus impliquant des événements rares, généralement caractérisés par des paysages accidentés et des barrières à haute énergie. D'une manière générale, les méthodes d'échantillonnage avancées peuvent être regroupées en trois catégories :

- Les méthodes qui ajoutent une force artificielle externe (biais) au système, de sorte que l'exploration des régions à haute énergie libre est favorisée, poussant le système à explorer l'espace conformationnel plus rapidement.
- Les méthodes qui concentrent l'échantillonnage sur la région de transition (la barrière d'énergie libre), explorant un ensemble de chemins de transition au lieu d'explorer des conformations spécifiques.
- Les méthodes qui explorent l'hamiltonien à différentes températures où, compte tenu de la distribution de Boltzmann, l'occupation des barrières d'énergie par rapport aux minima augmente avec la température, donc l'échantillonnage est plus efficace.

Dans ce travail, nous nous concentrerons sur les deux premières stratégies répertoriées. Les méthodes qui entrent dans la première catégorie nécessitent l'utilisation d'une ou des CV pour ajouter un potentiel basé fonction des CV à l'énergie potentielle du système. Ces CV doivent décrire les degrés de liberté considérés comme les plus pertinents pour le mécanisme de transition. Dans ces méthodes, le potentiel de biais peut être constant ou dépendant

du temps. Le biais ajouté à l'énergie potentielle étant connu, on peut en principe reconstruire les profils d'énergie libre projetés sur les CV biaisés directement ou au moyen de méthodes de pondération. Parmi ces méthodes, la métodynamique (MetaD) est une technique d'échantillonnage avancée où la recherche conformationnelle est accélérée en ajoutant au champ de force un potentiel basé dépendant de l'histoire d'une ou de quelques CV. La métodynamique a l'avantage que, pour une simulation convergée et des CV appropriés, il est possible de récupérer directement le profil d'énergie libre du système à partir du biais ajouté par la métodynamique. Cependant, la métodynamique, ainsi que d'autres méthodes d'échantillonnage avancées similaires ont comme inconvénients que à cause de l'accélération de l'échantillonnage les informations sur la dynamique du système simulé sont corrompues. Plusieurs méthodes ont été développées pour estimer les coefficients de vitesse à partir des simulations de métodynamique, certaines impliquent le calcul de coefficients de diffusion et la construction de modèles d'états de Markov ; d'autres méthodes redéfinissent la notion de facteur d'accélération pour l'appliquer aux simulations de métodynamique. Bien que ces méthodes aient été appliquées avec succès dans certains cas, certaines d'entre elles sont incapables de reproduire la cinétique des transitions plus rapide que le temps de décalage utilisé pour le calcul du coefficient de diffusion, tandis que d'autres nécessitent des CV idéales pour éviter la corruption de la dynamique dans l'état de transition (TS).

La deuxième catégorie des méthodes d'échantillonnage avancées décrites ci-dessus sont celles qui se concentrent sur l'exploration d'un grand nombre de chemins de transition reliant deux états prédéfinis. La définition des états nécessite une connaissance préalable du processus de transformation. Parmi ces approches, la plus utilisée est le Transition Path Sampling (TPS). Dans cette méthode, une collection de trajectoires réactives est générée à partir d'un chemin réactif unique. Le chemin initial peut être obtenu, par exemple, en utilisant une autre méthode d'échantillonnage avancée comme la métodynamique. Un avantage du TPS est qu'il exploite le fait que le temps de chemin de transition est plus rapide que le temps caractéristique dans les états, par conséquent, l'échantillonnage est plus efficace. En utilisant le TPS, plusieurs méthodes permettent l'estimation des taux cinétiques à partir de la dynamique d'équilibre, par exemple, l'échantillonnage d'interface de transition (TIS) et la dynamique de non-équilibre, par exemple, l'échantillonnage de flux direct (FFS). Bien qu'en théorie le TPS permette un accès direct aux informations cinétiques du système, en pratique ces techniques nécessitent un échantillonnage poussé pour que l'estimation de vitesse atteigne la convergence, ce qui la rend inaccessible en termes de coût de calcul.

Compte tenu de cela, il existe des questions ouvertes dans le domaine : une compréhension complète d'un processus physico-chimique comprend un compte rendu détaillé de la cinétique. Étant donné que la thermodynamique peut être obtenue à partir de la cinétique, mais pas l'inverse, des stratégies simples et efficaces capables d'extraire la cinétique des simulations informatiques atomistiques sont toujours nécessaires. Outre la précision intrinsèque du champ de force interatomique (un domaine très actif, qui dépasse cependant le cadre de la présente thèse), l'échantillonnage extensif requis pour la prédiction des taux cinétiques à partir de simulations MD reste l'un des principaux défis actuels, en dépit de l'avancée des capacités informatiques actuelles. Les méthodes capables d'extraire des informations

thermodynamiques et cinétiques fiables à partir de trajectoires d'échantillonnage avancées font l'objet de recherches intenses. De plus, afin d'exploiter de manière efficace les supercalculateurs actuels, les méthodes qui peuvent utiliser des ensembles de trajectoires courtes obtenues en parallèle sont plus optimales que celles qui nécessitent de longues trajectoires d'équilibre.

Dans le même temps, les CV sont essentiels pour comprendre les mécanismes moléculaires des événements rares. Ils sont utilisés dans l'échantillonnage avancé, soit comme directions de biais, soit comme indicateurs de l'état actuel du système. De nos jours, la prédiction des taux cinétiques repose souvent sur l'hypothèse que les CV utilisés sont idéales. Dans les versions de base du TPS comme l'aimless shooting où le but est d'échantillonner les chemins de transition, la dynamique n'étant pas biaisée, différentes CV peuvent être définis a posteriori pour l'analyse. D'autre part, les méthodes de TPS avancées telles que TIS et FFS permettent de calculer la cinétique, au prix du choix a priori d'une CV idéale. Dans ces dernières techniques, ainsi comme dans la métadynamique et les approches d'échantillonnage biaisées associées, un mauvais choix de CV entraîne une sous-estimation ou une surestimation grave du taux cinétique et peut également conduire à échantillonner des chemins de transition défavorables.

De plus, améliorer les CV dans le but d'avoir une meilleure estimation du taux de transition nécessiterait alors d'exécuter un nouvel ensemble de simulations de TPS et de métadynamique. Par conséquent, les méthodes capables d'estimer la cinétique via une analyse post-traitement et qui permettent soit de comparer efficacement des ensembles de CV soit de corriger l'effet de mauvaises directions de biais seront utiles.

En résumé, dans ce manuscrit, nous tentons de répondre aux questions suivantes : Pouvons-nous développer des méthodologies efficaces pour le calcul des taux cinétiques d'événements rares à l'aide de trajectoires des méthodes d'échantillonnage avancées ? Ces méthodes permettent-elles d'évaluer la qualité des CV a posteriori évitant d'avoir à lancer de nouvelles simulations lorsqu'une mauvaise CV est utilisé au premier abord? Pouvons-nous établir des critères pour tester la qualité de nos résultats lorsque la cinétique exacte n'est pas disponible pour la comparaison?

Dans cette thèse nous contribuons à avancer l'état de l'art dans le développement de méthodes pour l'estimation des vitesses de réaction à partir de méthodes liées à la dynamique moléculaire. Les plus grands défis dans ce domaine sont entre autres: (i) la limitation des échelles de temps accessibles avec les simulations MD qui rendent difficile l'étude des événements rares, (ii) le besoin des méthodes d'échantillonnage avancé d'avoir une CV idéal pour le calcul des constantes de réactions, et (iii) la nécessité des critères pour diagnostiquer la fiabilité des prédictions. Afin de remédier à ces problèmes, nous proposons ici deux nouvelles méthodes alternatives pour prédire les taux de transition. Ces méthodes théoriques ont été conçues à partir de la compréhension fondamentale de la dynamique modélisée via la dynamique moléculaire et des simulations d'échantillonnage avancée. L'analyse post-traitement permet alors de récupérer les statistiques de franchissement de la barrière en utilisant des trajectoires d'équilibre ou hors d'équilibre. Elles permet également d'exploiter les capacités de calcul parallèle des ordinateurs actuels favorisant l'utilisation d'ensembles

relativement petits de simulations courtes (par rapport aux échelles de temps des événements rares) au lieu de longues trajectoires ergodiques.

Dans le chapitre 6, nous avons développé une nouvelle méthodologie basée sur la théorie de Kramers pour estimer les taux de transition à partir d'un ensemble de simulations courtes soumis à un biais dépendant du temps. Nous nous sommes inspirés des méthodes utilisées par la communauté des forces spectroscopie qui étudie les expériences de molécules uniques, et nous améliorons considérablement une approche similaire la métadynamique infréquente.

Nos résultats à la fois sur des systèmes de référence (chapitre 6) et sur des simulations MD (chapitre 7) indiquent que la nouvelle approche fournit simultanément à la fois les taux de transition (même quand des CV non idéales sont utilisés) à partir de simulations plusieurs ordres de grandeur plus courtes que le temps moyen de transition des processus, ainsi qu'une estimation quantitative de la qualité d'une CV. Ce dernier résultat ouvre une perspective intéressante d'optimisation automatique des CV en utilisant, par exemple, les approches d'apprentissage automatique avec le paramètre de qualité de CV comme fonction d'optimisation.

Nous avons utilisé des exemples de simulations de métadynamique, cependant, la méthode KTR que nous avons développée est utilisable pour toute simulation biaisée par des CV dépendante du temps, comme entre autres la force de polarisation adaptative ou MD à biais adaptatif. Les méthodes limitant le niveau de remplissage du biais pourraient également être utilisées dans le cadre de notre méthode. Ce qui aurait l'avantage de simplifier le calcul du potentiel biaisé et de la probabilité de survie. De plus, notre méthode propose une nouvelle mesure de l'efficacité du biais ajouté pour accélérer la transition (liée à l'efficacité des CV). Cela surmonte plusieurs limitations sévères rencontrées avec les approches précédentes où le biais devait être déposé très rarement sur des CV idéales. Plusieurs aspects de la théorie KTR peuvent être améliorés dans le futur. Des méthodes automatisées pour déterminer quand le franchissement de barrière se produit pourraient être utiles. Aussi, il serait utile d'extrapoler la théorie KTR aux cas de non-équilibre, en effet comme nous l'avons montré la théorie de Kramers et l'hypothèse quasi-adiabatique que nous prenons ne sont plus valide dans ce cas.

Enfin, la principale limite de notre approche actuelle, au delà l'imprécision du champ de force, semble être l'approximation implicite consistant à considérer le processus de transition comme l'échappement d'un seul puits de "réactifs" vers un seul puits de "produits": une généralisation à plusieurs bassins et dimensions, comme cela a été fait pour la spectroscopie de force pour prendre explicitement en compte un réseau de chemins de transition éventuellement multiples à travers plusieurs états métastables serait très pratique pour tenir compte de la dynamique multi-états dans des simulations complexes.

Dans le chapitre 8, nous développons et appliquons un deuxième cadre méthodologique visant à l'estimation précise des paysages d'énergie libre, des taux cinétiques ainsi que l'optimisation automatique des coordonnées de réaction. Nous adoptons un schéma basé sur les équations de Langevin modélisant la dynamique projetée sur une CV : ces équations apportent une interprétation physique en termes de force moyenne et de bruit qui fait défaut dans des approches plus abstraites axées sur les données comme les modèles d'état de Markov

ou les prédicteurs basés sur l'apprentissage automatique.

Comme premier résultat important, notre travail démontre la possibilité d'obtenir simultanément les énergies libres et les taux cinétiques d'une manière conceptuellement simple, en optimisant un modèles de Langevin par maximum de vraisemblance à partir d'une centaine de trajectoires, cela quelle que soit la hauteur de la barrière. Remarquons que deux tests sont également présentés qui permettent d'identifier la résolution temporelle optimale pour la construction du modèle stochastique : c'est un point crucial pour permettre l'application robuste à des systèmes dont le comportement n'est pas connu à l'avance.

Les modèles de Langevin suramortis reproduisent bien les propriétés thermodynamiques et dynamiques quantitatives du système original, y compris les taux cinétiques précis, malgré un écart de plusieurs ordres de grandeur en temps par rapport aux trajectoires MD courtes utilisées pour l'entraînement. Incidemment, nous notons que notre approche ne se limite pas à l'échantillonnage des chemins de transition : Les modèles de Langevin peuvent être optimisés à partir de n'importe quel ensemble de trajectoires non biaisées explorant la région de transition.

Dans des travaux futurs, cette approche d'optimisation des modèles de Langevin pourra être généralisée à différents types d'équations de Langevin (généralisées ou standard) permettant d'étendre son utilisation à des applications telles que les réactions chimiques en solution. Jusqu'à présent, l'approche se limite à des modèles de Langevin unidimensionnels : compte tenu de la complexité des processus physico-chimiques et de la difficulté de trouver une CV qui les décrit, une extension de l'approche à plus de dimensions, qui permet l'utilisation simultanée de plusieurs CV, pourrait également être utile.

Ayant ouvert la possibilité d'obtenir des modèles Langevin proches de la dynamique projetée le long de CV différentes, une direction naturelle de recherche, que nous avons poursuivie au chapitre 10, consiste à effectuer une variation systématique de la définition de la CV pour obtenir une coordonnée de réaction optimale. Comme principe variationnel guidant l'algorithme d'optimisation, nous avons systématiquement réduit le taux de transition : l'estimation directe de ce dernier par quadrature numérique est une ingrédient important pour l'efficacité numérique.

Considérant un CV d'essai formé par une combinaison linéaire d'un ensemble de CV putatifs, comme une preuve de concept, nous démontrons que la coordonnée de réaction idéale est rapidement et de manière fiable récupérée dans un simple système de benchmark 2D. Remarquablement, nos résultats montrent qu'un modèle de Langevin à maximum de vraisemblance le long d'un CV non idéal tend à sous-estimer la vraie barrière d'énergie libre dans la même direction : ce phénomène ne limite cependant pas l'efficacité de l'algorithme d'optimisation du CV, puisqu'il préserve le principe variationnel. Une première application à la dissociation d'un dimère de fullerène dans l'eau confirme la capacité de la nouvelle méthode à obtenir en même temps une coordonnée de réaction optimale et une estimation précise du taux de transition. Cependant, des travaux sont en cours pour étendre l'étude à un ensemble plus large de CV putatifs (dont la construction est un problème intéressant en soi) et à différents systèmes. Une autre amélioration importante pourrait être l'inclusion de la non-linéarité dans la combinaison de CV grâce à l'intelligence artificielle et à l'amélioration du

calcul efficacité des codes numériques utilisés pour l'optimisation des modèles de Langevin.

Nous avons fourni ici un résumé des avancées théoriques les plus importantes et des résultats de cette thèse. Les deux méthodes sont des voies alternatives avec un objectif commun : la prédiction des cinétique de réaction. La possibilité d'aborder le problème de la cinétique sous différents angles montre sa complexité. L'utilisation de l'une des méthodes proposées ou de l'autre dépendent des données disponibles et de l'application voulue. Cette thèse ouvre une pléthore de possibilités pour l'exploitation de l'information disponible dans d'échantillonnage de chemin de transitions, ou la possibilité de récupérer des informations utiles à partir de simulations de métodynamique avec biais sur des CV non idéales, à l'utilisation de ces approches dans l'optimisation de variables collectives. Alors que la communauté de la simulation informatique atomistique progresse et plus loin dans l'étude de la cinétique, nous attendons avec impatience toutes les améliorations et applications que peuvent avoir nos nouvelles approches.

2

Introduction

2.1 ON THE IMPORTANCE OF KINETIC RATES ESTIMATION

The prediction of kinetic rates is key for understanding physical-chemical phenomena. Processes like phase transitions, chemical reactions, or biomolecular interactions and conformational changes are characterized by a network of metastable states often separated by high free energy barriers.¹ Each metastable state corresponds to a local minimum in the Free Energy Surface (FES) which traps the system's dynamics for a time much longer than the intrinsic time of fast bond vibrations until the system can escape via random fluctuations to another metastable state. A high free energy barrier separating two states generates a dynamical bottleneck between them, making a transition between basins less probable, therefore, these transitions are presented as rare events.²

Rare events display a slow kinetics. In the examples of physicochemical processes mentioned above, a typical transition between two interesting states, e.g., liquid to solid phase transition, reactant to product conversion, protein-ligand association or unfolded to folded conformational change can feature timescales from microseconds to seconds and beyond.^{1,2} The estimation of the transition rates (i.e. the inverse of transition time) allows us to predict the timescales of these events, understand molecular mechanisms and in a more practical level, make informed decisions based on the kinetic properties of the system.

One remarkable example on the importance of rare events kinetics is the field of drug discovery. Due to the high cost of experimental trials, computer-assisted methods have gained relevance in recent decades in the prediction of drug candidates.^{3–6} During decades, efforts in drug discovery were directed to the estimation of affinities of small molecules to protein targets. However, promising compounds in the early stages of affinity-based drug screening get often discarded in clinical trials due to a lack of *in vivo* efficacy.⁷ Recently,

it has been shown that kinetic rates can correlate better than affinity with drug efficacy, possibly because *in vivo* complex non-equilibrium interactions make drug residence time play a more important role than *in vitro* equilibrium binding affinity.^{7–9} This finding has shifted the paradigm in drug discovery towards the accurate prediction of kinetic rates.^{7,10}

In the latter example, the drug residence time is quantified by means of its reciprocal, the dissociation rate constant k_{off} . Clinically relevant off-rates of drug candidates range from $10^0 - 10^{-5} \text{ s}^{-1}$, i.e., residence times from seconds to hours.⁷ In this context, being able to predict off-rates opens the possibility to find better drug candidates. Therefore, the prediction of these kinetic rates must involve the accurate characterization of long-time dynamics.

Since this slow dynamics is typical of rare events, examples of the importance of kinetics are not restricted to drug discovery. Among other examples, the prediction of rate constants for phase transitions of inorganic materials plays a key role in the design of new devices with technological applications. Also, in the field of molecular biology, the study of Protein-Protein Interactions (PPI), protein folding, and protein aggregation kinetics helps to elucidate the causes of diseases and create alternatives to treat and prevent them. Therefore, methods allowing the accurate prediction of kinetic rates could be useful in broad applications from material design to therapeutics. In the following, we will introduce the current methodologies used to estimate kinetic rates *in silico*.

2.2 KINETIC RATES FROM ATOMISTIC COMPUTER SIMULATIONS

Atomistic computer simulations of rare events have three paramount goals: predicting detailed mechanisms, free energy landscapes, and kinetic rates. All of these tasks are, in many cases, cumbersome and require intensive human and computer effort, the calculation of rates being the most difficult.^{1,2} The challenge of calculating rates is given by the complexity of simulating the evolution of a system over time. Time-dependent properties are therefore not accessible with most of the atomistic computer simulation methods available. Nevertheless, all-atoms Molecular Dynamics (MD) allows following the time-dependent behavior of a system by propagating the positions and velocities of all atoms in the system through the solution of equations of motion. MD simulations have been widely used because they allow studying both the thermodynamic and kinetic behavior of a molecular system.^{11,12}

2.2.1 MOLECULAR DYNAMICS

Molecular dynamics simulations predict how each atom in a molecular system will move over time by numerical integration of the equations of motion. Positions and velocities of the atoms are predicted according to a model of interatomic interactions: the force field. The forces in the system are sampled at each time step by evaluating the force field on the current atoms' positions.¹³ To ensure numerical stability in the integration of the equations of motion, a very small time step has to be used (shorter than the fastest bond vibrations). Since MD simulations evaluate the position of the atoms at each time step, this technique is

a powerful tool to elucidate molecular mechanisms at atomic resolutions and short timescales that are barely accessible through experimental techniques.¹¹

The result of MD simulations is a collection of positions and velocities of atoms over time. In order to study in detail molecular mechanisms one can focus on one or a few degrees of freedom of the system. Therefore, one can use a Collective Variable (CV), i.e., a descriptor that can identify collective movements of the system over the transformation process.¹⁴ Since CVs allow us to follow the progress of the transformation process, with MD simulations one could directly estimate kinetic rates in the infinite sampling limit, by measuring the average time that the system spends in a certain state. Other theoretical approaches aim to estimate kinetics from MD simulations making use of different mathematical frameworks.² On one side, the full dynamics of the system can be projected onto low-dimensional CVs leading to Langevin stochastic differential equations.¹⁵⁻¹⁷ On the other hand, the configuration space can be partitioned into discrete states and the transition between states can be counted to construct Markov State Models (MSM).^{18,19}

Although MD simulations seem like a natural path to study the kinetic behavior of a molecular system^{11,12,20}, as explained above, many interesting molecular processes present rare events, therefore, the estimation of meaningful thermodynamic and kinetics properties using MD involve the evaluation of the equations of motion more than billions of times.¹¹ This leads to a gap between the timescales of rare events described before and those accessible with MD, even using powerful supercomputers.²¹⁻²³ Therefore, enhanced sampling methods have become useful alternatives to speed up the sampling in MD.

2.2.2 ENHANCED SAMPLING METHODS

In combination with MD simulations, enhanced sampling methods have become useful alternatives for studying events that occur at long timescales.¹ These methods have been developed to accelerate the sampling of the conformational space of processes involving rare events, typically characterized by rugged landscapes and high energy barriers.²⁴ Broadly speaking, enhanced sampling methods can be grouped into three categories:¹

- Methods that add an external artificial force (bias) to the system, so that the exploration of the regions with high free energy is favored, pushing the system to explore the conformational space faster.
- Methods that focus the sampling on the transition region (the free energy barrier region), exploring an ensemble of transition paths instead of exploring specific conformations.
- Methods that explore the Hamiltonian at different temperatures where, given the Boltzmann distribution, the population of energy barriers relative to minima increases with temperature, hence sampling is more effective.

Note that in this work we focus on the first two strategies listed in the list, so we will elaborate further on them.

The methods that fall into the first category require the use of CVs to add a bias potential to the potential energy of the system as a function of them. These CVs should describe the degrees of freedom considered most relevant to the transition mechanism.²⁵ In these methods, the bias potential can be constant or time-dependent. Since the bias added to the potential energy is known, in principle one can reconstruct the free energy profiles projected on the biased CVs directly or by means of re-weighting methods. Among these methods, Metadynamics (MetaD)²⁶, is an enhanced sampling technique where the conformational search is accelerated by adding a history-dependent bias potential to the force-field as a function of one or few CVs.

MetaD has the advantage that for a converged simulation and appropriate CVs²⁷, it is possible to directly recover the free energy profile of the system from the MetaD bias.²⁸ However, a disadvantage of MetaD, as well as other related enhanced sampling methods, is that information about the dynamics of the simulated system is corrupted, due to the sampling acceleration.²⁷ Several methods have been developed for estimating the rate coefficients from MetaD simulations²², some involve the calculation of diffusion coefficients^{29,30} and the construction of Markov State Models;³¹ other methods redefine the concept of the acceleration factor to apply it to MetaD simulations.³² Although these methods have been successfully applied in some cases^{33–35}, some of them are unable to reproduce the kinetics of transitions faster than the lag time assigned for the calculation of the diffusion coefficient²⁹, while others, require ideal CVs to avoid the corruption of the dynamics over the Transition State (TS).³²

In the second category of enhanced sampling methods described above are those that focus on the exploration of a large number of transition pathways connecting two predefined states.¹ The definition of the states requires some prior knowledge of the transformation process. Among these approaches, the most widely used is Transition Path Sampling (TPS).³⁶ In this method, a collection of reactive trajectories is generated starting from a single reactive pathway. The initial pathway can be obtained, for instance, using a biasing enhanced sampling method.

An advantage of TPS is that it exploits the fact that the transition path time is faster than the waiting time, therefore, the sampling is more efficient. Within the TPS framework³⁷ several methods allow the estimation of kinetic rates for instance, Transition Interface Sampling (TIS)³⁸ and Forward Flux Sampling (FFS).^{39,40} Although in theory TPS allows direct access to the kinetic information of the system, in practice these techniques are complex to apply and require extensive sampling for the rate estimate to reach convergence, making it very expensive in terms of computational and human time.

2.3 OPEN CHALLENGES

A complete understanding of a physicochemical process include a detailed account of kinetics. Since thermodynamics can be obtained from kinetics, but not vice versa, simple and efficient strategies that are able to extract kinetics from atomistic computer simulations are still needed. Apart from the intrinsic accuracy of the interatomic force field (a very active field,

which however is beyond the scope of the present thesis), the extensive sampling required for the prediction of kinetic rates from MD simulations remains one of the major open challenges, notwithstanding the advance in current computer capabilities. Methods able to extract reliable thermodynamic and kinetic information from enhanced sampling trajectories are at the focus of intense research and are highly desirable. Additionally, in order to exploit in an efficient way current supercomputers, methods that can use ensembles of short trajectories obtained in parallel are more desirable than those that require long equilibrium trajectories.

At the same time, CVs are key to understand molecular mechanisms of rare events. They are used in enhanced sampling either as biasing directions or as indicators of the current state of the system. Nowadays the prediction of kinetic rates often relies on the use of optimal CVs. In basic versions of TPS like aimless shooting where the aim is to sample transition paths, since the dynamics is unbiased, different CVs can be defined *a posteriori* for the purpose of analysis.

On the other hand, advanced TPS methods like TIS and FFS allow computing kinetic rates, at the price of a careful *a priori* choice of a CV. In these latter techniques, as well as in MetaD and related biased sampling approaches, a poor CV choice yields severe under- or overestimation of the kinetic rate and it can also lead to sample unfavorable pathways. As a consequence, improving the CV to improve the kinetic rate would require to run new sets of simulations in both TPS and MetaD.⁴¹ Therefore, methods that are able to estimate the kinetics in post-processing analysis and that allow either to efficiently screening pools of CVs or to correct the effect of poor biasing directions will be useful.

In summary, in this manuscript we attempt to answer the following questions: Can we develop efficient methodologies for the calculation of kinetic rates of rare events using short trajectories from enhanced sampling methods? Can these methods assess the quality of the CV *a posteriori* avoiding the need to run new simulations when a poor CV is used in the first place? Can we establish criteria to test the quality of our results when the exact kinetic rate is not available for comparison?

2.4 OUR PROPOSAL

To answer the questions above we propose two strategies: i) Estimation of transition rates from MetaD trajectories using Kramers' theory, assessing and correcting simultaneously the efficacy of the CVs. ii) Optimization of Langevin models to access transition rates, free energy and diffusion profiles from unbiased MD and TPS trajectories. Both strategies are used in post-processing mode and only require small sets of short trajectories (short compared to the waiting time of the respective rare events) that can be obtained in parallel, reducing the wall-clock time needed for computation.

This manuscript is organized as follows: First, we show an overview of rate theory and current computational methods used to estimate kinetic rates in the methods section. Then, we introduce in detail the theory, validation and applications of the methods proposed in this Ph.D. project. Finally, we present the conclusions and perspectives for applications and extensions of our proposed methods. A summary of the dissemination of research results

(publications, presentations at conferences, etc.) completed during the time of the PhD can be found in Appendix C.

Part II

Theory and Methods

3

Kinetic rates estimation

3.1 MACROSCOPIC PERSPECTIVE

Transitions between metastable states in rare events, that is, transitions between regions of configuration space with low free energy (high probability), separated by barriers, i.e., interface regions of high free energy (low probability) are stochastic processes:¹ the times needed for a transition are different for multiple samples of the system prepared in exactly the same macroscopic conditions.⁴² The average time needed to see a transition is related to the transition rate k via

$$\langle t \rangle = \frac{1}{k} . \quad (3.1)$$

At the macroscopic level, the kinetics of rare events is well captured by the kinetic equations. Let us take as an example of a rare event a molecule in solution with two different conformations, in which the transformation of one conformation or reactant (A) to other conformation or product (B) requires overcoming a sizeable barrier or activation energy. The kinetic equations describe the the time evolution of the populations (concentrations) in the the reactant and product states. The concentrations for each conformation c_A and c_B evolve in time following the expressions

$$\begin{aligned} \frac{dc_A(t)}{dt} &= -k_{A \rightarrow B} c_A(t) + k_{A \leftarrow B} c_B(t) , \\ \frac{dc_B(t)}{dt} &= -k_{A \leftarrow B} c_B(t) + k_{A \rightarrow B} c_A(t) , \end{aligned} \quad (3.2)$$

where $k_{A \rightarrow B}$ and $k_{A \leftarrow B}$ are the forward and backward rate constants or transition rates for the reaction $A \rightleftharpoons B$, respectively. The equations above assume that the number of transitions per unit time from A to B is proportional to the concentration (population) in A and vice versa, with a proportionality constant given by the rate constant. We remark that these equations are general for any two-state system, i.e., a system that can exist in two different states.

Equilibrium is reached when the concentrations of the two states do not change with time. Under these conditions, the ratio of the concentrations is equal to the ratio of the backward and forward rate constants:

$$\frac{\langle c_A \rangle}{\langle c_B \rangle} = \frac{k_{A \leftarrow B}}{k_{A \rightarrow B}}, \quad (3.3)$$

where the angular brackets $\langle \dots \rangle$ refer to the equilibrium concentrations of each species.

In computer simulations, one way to determine transition rates is by preparing the system in one state and monitoring its transformation or decay to the other state over time. For example, starting from state A , the concentrations relax to equilibrium following the equations

$$\begin{aligned} c_A(t) &= \langle c_A \rangle + [c_A(0) - \langle c_A \rangle] \exp(-[k_{A \rightarrow B} + k_{A \leftarrow B}]t), \\ c_B(t) &= \langle c_B \rangle + [1 - \exp(-[k_{A \rightarrow B} + k_{A \leftarrow B}]t)]. \end{aligned} \quad (3.4)$$

Therefore, for a reversible reaction, i.e. where the system can constantly fluctuate between states A and B , the inverse of the reaction time τ can be written as

$$\tau^{-1} = k_{A \rightarrow B} + k_{A \leftarrow B} = k_{A \rightarrow B} \frac{\langle c_A \rangle + \langle c_B \rangle}{\langle c_B \rangle}. \quad (3.5)$$

On the other hand, in the case of irreversible transitions, the evolution of the concentrations is determined only by the forward rate $k_{A \rightarrow B}$, as the backward rate $k_{A \leftarrow B}$ effectively vanishes

$$\begin{aligned} c_A(t) &= c_A(0) \exp(-k_{A \rightarrow B} t), \\ c_B(t) &= c_A(0)[1 - \exp(-k_{A \rightarrow B} t)]. \end{aligned} \quad (3.6)$$

Therefore, the reaction time correspond to the average waiting time (Eq. 3.1) for the decay of the initial state

$$\tau = k_{A \rightarrow B}^{-1}. \quad (3.7)$$

In principle, we can use computational experiments to calculate transition rate constants. For this purpose, in the next sections, we will introduce the basics to calculate the forward rate constant $k_{A \rightarrow B}$, keeping in mind that the same procedure can be used to compute the

backward rate $k_{A \leftarrow B}$.

3.2 DIRECT CALCULATION OF THE RATE CONSTANT

If one can follow the evolution of a system over time by means of atomistic computer simulations, for example, by using MD simulations, a way to estimate the rate constant is by analyzing a long simulation. If the free energy barriers between the two states can be overcome in much less time than the total simulation time, then the system will be able to go back and forth between the two metastable states A and B within the simulation time. Therefore, it is possible to count the number of transitions (jumps) from one state to another and the time the system spends in each state. With this information and following the definition of the rate constant, it can be estimated as

$$k_{A \rightarrow B} = \frac{n_{A \rightarrow B}}{t_A} , \quad (3.8)$$

where $n_{A \rightarrow B}$ is the number of jumps out of state A and t_A is the total time spent in the state A . As mentioned above, Eq. 3.8 can only be applied in practice if it is possible to observe enough transitions between the states, that is, if the free energy barriers are not very large. Therefore, transition rates on rare events cannot generally be calculated in this way by means of conventional MD simulations. The details of this limitation in MD simulations and methods aimed to speed up the sampling, i.e., enhanced sampling methods are discussed in Chapters 4 and 5, respectively.

3.3 ESTIMATION OF THE MEAN FIRST-PASSAGE TIME

The Mean-First Passage Time (MFPT) is the time it takes for the system starting from state A to first arrive at state B . In the following sections we will define two methods to calculate the MFPT from atomistic computer simulations.

3.3.1 THE KRAMERS' PROBLEM

Imagine the escape of a Brownian particle from a free energy well. In 1940 Kramers^{43,44} introduced the idea of viewing chemical reactions as diffusion processes by proposing a model to estimate the escape rate of the particle from the bottom of the well. The position of the particle is described by a variable q . The particle moves diffusively in a free energy surface $F(q)$ with a position dependent coefficient $D(q)$. The motion of the particle is described by the overdamped Langevin equation:^{17,45}

$$\dot{q} = -\beta D(q) \frac{\partial F(q)}{\partial q} + \frac{\partial D(q)}{\partial q} + \sqrt{2D(q)} \eta(t) , \quad (3.9)$$

where $\eta(t)$ is a Gaussian white noise. The Langevin equations are discussed in detail in Chapter 8, section 8.2.1. Following the model, the probability density ρ of the particle position q moving on the free energy $F(q)$ evolves according to the Smoluchowski equation:⁴⁶

$$\frac{\partial}{\partial t} \rho = \frac{\partial}{\partial q} \left(\beta D(q) \frac{\partial F(q)}{\partial q} \rho - \frac{\partial D(q)}{\partial q} \rho \right) + \frac{\partial^2}{\partial q^2} (D(q) \rho) \quad (3.10)$$

By analogy, we can model the transition of the system from a metastable state A to a metastable state B as the escape of this Brownian particle.

Now, consider the MFPT $\tau(q^*)$ to reach the top of the barrier, q^* . The barrier top q^* is the location of the TS, where the probability to advance towards the products state B is equal to the probability to return to the reactants state A , therefore, the reaction rate constant is given by

$$k_{A \rightarrow B} = \frac{1}{2} \tau(q^*)^{-1} \quad (3.11)$$

Using the Smoluchowski equation, one can arrive to a general expression for the MFPT $\tau(q)$ to reach any point q starting from any point q_0 ¹⁶

$$\tau(q, q_0) = \int_{q_0}^q dy \frac{1}{D(y)} \exp(\beta F(q)) \int_a^y dz \exp(-\beta F(z)) , \quad (3.12)$$

where a is a reflective wall located at the left of the free energy well (see Fig. 3.1a). Under this framework, if the relaxation time in the well is fast compared to the escape time, starting from any initial point q_0 at the bottom of the well, the MFPT to reach a point q near the barrier is the same. Assuming that the shape of the barrier is close to parabolic near its top (the potential energy can be expanded around q^* and truncated after the quadratic term), the MFPT for values q near to the barrier-top are given by⁴²

$$\tau(q) = \frac{\tau_j}{2} [1 + \text{erf}([q - q^*]c)] , \quad (3.13)$$

where $\text{erf}(z) = (1/\sqrt{\pi}) \int_0^z \exp(-y^2) dy$ is the error function and $c = \sqrt{\beta|U''(q^*)|/2}$ is the respective local curvature. In practice, Eq. 3.13 it is useful to estimate the MFTP when the free energy barrier is not very high. For that, several simulations can be launched from different configurations at the bottom of well A and follow them until the barrier is crossed with τ_j defined as this jump time, i.e., the simulation time need to jump to state B . Then one can fit the Eq. 3.13 to the simulation data with τ_j , q^* , and c as variational parameters. Finally, the rate constant is given by

$$k_{A \rightarrow B} = \tau_j^{-1} . \quad (3.14)$$

3.3.2 MEAN LIFETIME METHOD

The direct observation method⁴⁷ or mean lifetime method⁴⁸ allows estimating the MFPT with the advantage that it does not rely on the exact definition of the transition state⁴⁹. In this method it is assumed that the times to reach a particular state are Poisson distributed,

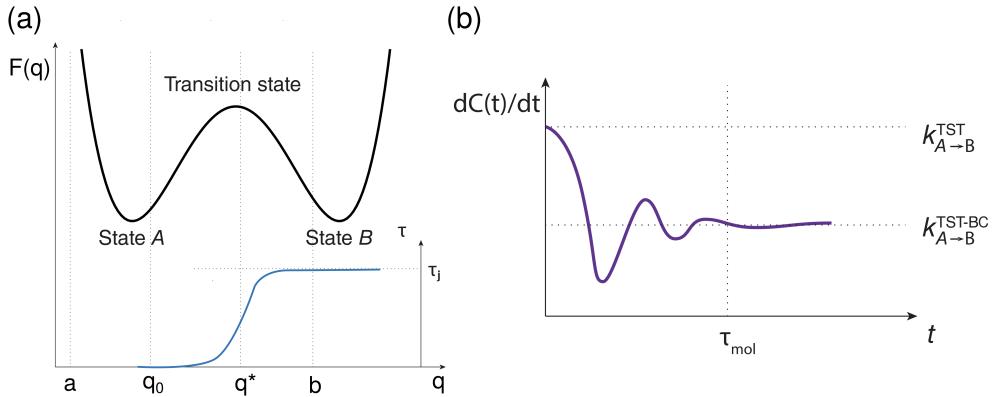


Figure 3.1: Schematic illustration of the Kramers and Bennett-Chandler approaches. (a) Kramers approach: the system diffuses in a free energy potential $F(q)$ along a coordinate q within a reflective boundary a and a moving absorbing boundary b . The transition rate is inversely proportional to twice the mean time required to escape through $b = q^*$ when starting at q_0 , see Eq. 3.11. In the high barrier limit, the MFPT can be estimated with Eq. 3.13. (b) Bennett-Chandler approach: the rate constant is given by the time derivative of the correlation function. On short time scales $t < \tau_{mol}$, correlated recrossings of the dividing surface may occur, effectively reducing the flux out of A. $k_{A \rightarrow B}^{TST}$ overestimates the rate by neglecting those recrossings. The Bennett-Chandler approach corrects the rate for such recrossings, $k_{A \rightarrow B}^{TST-BC}$. Adapted from Ref. 42.

therefore, the probability of observing a transition to state B in a given time interval are fitted to the distribution

$$H(t|q \in B) = gt \exp(-ht) , \quad (3.15)$$

where g and h are fitting parameters. Here, the rate constant is given by h

$$k_{A \rightarrow B} = h . \quad (3.16)$$

To estimate the rate constant with this method, different simulations must be launched from state A with random velocities drawn from a Boltzmann distribution and initial configurations at the bottom of the well. These simulations can be stopped once they reach state B and the simulation times are used to make the fit in Eq. 3.16. A more detailed discussion of this method can be found in the Chapter 6.

In comparison to the method presented in section 3.2 where a single long MD simulation is used, the methods presented in this section use multiple shorter MD simulations. However, these MFPT estimation methods are still limited to small free energy barriers, where at least one transition can be observed in a simulation time achievable through MD simulations. When the location of the TS is known, one possibility to deal with this issue is to use other methods as shown below.

3.4 TRANSITION STATE THEORY

An alternative view to the problem of estimating the rate constant is given by the Transition State Theory (TST)^{50–52}, which was developed almost in parallel with Kramers' theory. The central concept in the TST is the existence of the transition state (TS), which is characterized by a saddle point in the Potential Energy Surface (PES). This activated state has a single imaginary frequency or negative curvature in only one degree of freedom: the reaction coordinate. The TS does not correspond to a single configuration of the system, in fact all the configurations that fall within the surface perpendicular to the direction of the reaction coordinate, here called dividing surface, belong to the Transition State Ensemble (TSE).

In TST, the reaction rate is given by

$$k_{A \rightarrow B}^{TST} = \kappa \langle \dot{q}^* \rangle \frac{Z^*}{Z_A} , \quad (3.17)$$

where $\langle \dot{q}^* \rangle$ is the average velocity with which the activated complex slides across the dividing surface, Z^* is the probability density in the TS,

$$Z^* = \exp(-\beta F(q^*)) , \quad (3.18)$$

and Z_A is the partition function of the initial state,

$$Z_A = \int_{-\infty}^{q^*} \exp(-\beta F(q)) dq . \quad (3.19)$$

Finally, κ is a prefactor which can account for the possibility of recrossings of the dividing surface. In the TST formulation κ is set to 1. Note that in Eq. 3.17 only trajectories that successfully cross the dividing surface and transition to state B are considered.

In practice, the estimation of the TST rate involves two sets of simulations that are performed separately: (i) the free energy profile as a function of coordinate q is estimated using MD or enhanced sampling methods (see Chapters 4 and 5), thus allowing the estimation of the ratio of the partition functions. (ii) the average velocity of the activated complex is calculated from a set of multiple simulations launched from configurations at the top of the barrier belonging to the TSE. The expression used for the numerical estimate of the rate constant becomes

$$k_{A \rightarrow B}^{TST} = \frac{1}{2} \langle |\dot{q}| \rangle_{q=q^*} \frac{\exp(-\beta F(q^*))}{\int_{-\infty}^{q^*} \exp(-\beta F(q)) dq} . \quad (3.20)$$

The main drawback of TST is non-recrossings assumption, i.e., assuming that all trajectories coming from the reactant state A and passing through the dividing surface relax into the product state B . In reality, the system can fluctuate over the barrier and cross back to the reactant state, increasing the average transition time. Therefore, TST overestimates the rate, being $k_{A \rightarrow B}^{TST}$ the maximum possible rate (see Fig. 3.1b). An accurate estimation of the rate constant requires considering recrossings, which has been the main motivation for the

development of the strategy explained in the next section.

3.5 BENNETT-CHANDLER METHOD

A computational method that allows the accurate estimation of the transition rate is the Bennett-Chandler approach (TST-BC)^{53,54}, also called reactive flux method. Given a reaction coordinate q , the characteristic functions that indicate the state where the system is can be defined as

$$h_A(q) = \theta(q^* - q) = \begin{cases} 1, & \text{if } q < q^* , \\ 0, & \text{else ,} \end{cases} \quad (3.21)$$

$$h_B(q) = \theta(q - q^*) = 1 - h_A(q) ,$$

where $\theta(q)$ is the Heaviside theta function. The dividing surface is located at $q = q^*$, therefore, the functions h_A and h_B indicate whether a configuration q is in state A or B . The conditional probability of finding the system in state B at time t given that the system was at state A at time 0 is given by the time correlation function $C(t)$

$$C(t) = \frac{\langle h_A[q(0)]h_B[q(t)] \rangle}{\langle h_A \rangle} , \quad (3.22)$$

where $\langle h_A \rangle = \langle c_A \rangle / (\langle c_A \rangle + \langle c_B \rangle)$ connects to fraction of equilibrium concentrations in A (Eq. 3.3). Hence, the time evolution of the product concentration for a system where $c_B(0) = 0$ can be written in terms of the correlation function as

$$c_B(t) = (\langle c_A \rangle + \langle c_B \rangle)C(t) . \quad (3.23)$$

Now, let us call the typical time scale of molecular fluctuations τ_{mol} . For time scales larger than τ_{mol} the populations evolve following Eqs. 3.4. Thus, the combination of Eq. 3.4 and Eq. 3.23 leads to

$$C(t) = \langle h_B \rangle [1 - \exp(-[k_{A \rightarrow B} + k_{A \leftarrow B}]t)] . \quad (3.24)$$

The latter expression can be reduced when taking any time t in the time interval $\tau_{mol} < t \ll \tau$, with $\tau = (k_{A \rightarrow B} + k_{A \leftarrow B})^{-1}$

$$C(t) \approx k_{A \rightarrow B}t . \quad (3.25)$$

Therefore, for $t > t_{mol}$, after all the correlated recrossings have occurred, the rate constant is given by the derivative of the correlation function:

$$k_{A \rightarrow B}^{TST-BC} = \left. \frac{dC(t)}{dt} \right|_{t>t_{mol}} . \quad (3.26)$$

The crucial difference from the TST is encoded in Eq. 3.26, since in this framework recrossings are explicitly taken into account, while in the TST they are ignored (see Fig. 3.1b). It is possible to rewrite the TST rate $k_{A \rightarrow B}^{TST}$ in this framework using procedure below.

The correlation function is time translational invariant, as a consequence, one can use the definition in Eq. 3.22 to express $\dot{C}(t)$ as

$$\frac{dC(t)}{dt} = \frac{\langle \dot{q}(0) \delta[q(0) - q^*] h_B[q(t)] \rangle}{\langle h_A \rangle}. \quad (3.27)$$

Since in TST the recrossings are neglected, one can take the limit $t \rightarrow 0$ in the last equation by assuming that all the passages through the dividing surface become uncorrelated in time. Thus, the time dependency is eliminated as the difference in $h_B[q(t)] = \theta(q(t) - q^*)$ is replaced by $\theta[\dot{q}(0)]$,

$$k_{A \rightarrow B}^{TST} = \frac{\langle \dot{q}(0) \delta[q(0) - q^*] \theta[\dot{q}(0)] \rangle}{\langle \theta(q^* - q) \rangle}. \quad (3.28)$$

The analogy with the Eq. 3.20 can be clearly seen multiplying the equation above by the factor $\delta(q^* - q)$.

$$k_{A \rightarrow B}^{TST} = \frac{\langle \dot{q}(0) \delta[q(0) - q^*] \theta[\dot{q}(0)] \rangle}{\langle \delta(q^* - q) \rangle} \times \frac{\langle \delta(q^* - q) \rangle}{\langle \theta(q^* - q) \rangle}. \quad (3.29)$$

With this transformation we can again calculate the rate from equilibrium averages, since the first factor on the right hand side corresponds to the expectation value of the positive velocity and the second term is the conditional probability to find the system at q^* . Finally, we can measure the effect of ignoring recrossings and the deviation of the rate from TST with respect to the rate with the Bennett-Chandler approach (i.e., the exact rate up to statistical errors) with the transmission coefficient κ ,

$$\kappa = \frac{k_{A \rightarrow B}^{TST-BC}}{k_{A \rightarrow B}^{TST}}, \quad (3.30)$$

where κ is a number between 0 and 1. Roughly, κ measures the fraction of crossings of the dividing surface that lead to successful transitions from reactants to products. A larger number of recrossing leads to a smaller value of κ . Let us remember that in the TST κ is always set to 1 (see Eq. 3.17). In the following chapters of the methods part we will explain in depth where these challenges arise from and in the chapters of the results part we will introduce the strategies that we propose to tackle these difficulties. Therefore, TST always overestimates the reaction rate constant.

The procedure used for the practical application of the Bennett-Chandler/reactive flux method is described in the Chapter 8, section 8.2.5. Although the Bennett-Chandler approach discussed in this section allows in theory the precise calculation of the rate constant, in practice it is computationally very expensive because large amounts of trajectories are required for the estimation of the correlation functions (see chapter 8, section 8.2.5). Addi-

tionally, it relies on the use of a good reaction coordinate capable of describing the progress of the transition. Finding a good reaction coordinate is an open challenge even today.⁴² Therefore, this method cannot be easily used in applications to complex systems.

In this chapter we have discussed the different statistical methods available for the computational estimation of rate constants. In addition, we mention some of the challenges (without going into details) that have stopped the extensive application of these methods, such as: (*i*) the computational cost associated with carrying out long simulations and (*ii*) the difficulty in finding a coordinate of reaction that describes the transition process. In the following chapters of the methods part we will explain in depth where these challenges arise from. Then, in the chapters of the results part we will introduce the strategies that we propose to tackle these difficulties.

4

Atomistic computer simulations

4.1 BACKGROUND

The detailed study of physicochemical transformation processes requires a thorough understanding of the motion of all atoms in a system over time, in particular, when the objective is to measure the kinetic properties of a system and study the mechanism of the process. Currently, even the most advanced experimental techniques (with few exceptions in development such as ultrafast X-ray^{55,56}) lack the time resolution needed to study transformation mechanisms at the atomistic level. Therefore, Molecular Dynamics (MD) simulations emerged as an alternative to elucidate the behavior of a molecular system on spatial and temporal scales that experiments cannot access.

MD simulations are numerical experiments that aim to predict how every atom in a molecular system moves over time. These predictions are based on a general model, so-called Force Field (FF), describing the interatomic interactions.¹³ Thus, we can think on MD simulations as a high-resolution numerical microscope. However, just as the accuracy of the experiments is given by the quality of the instruments and subject to systematic and statistical errors, the accuracy of the MD predictions is given by the quality of the FF and limited by numerical errors which we will discuss later.

Nowadays, MD simulations are increasingly used to understand molecular mechanisms in combination with a wide variety of experimental techniques such as X-ray crystallography, cryo-electron microscopy (cryo-EM), nuclear magnetic resonance (NMR), electron paramagnetic resonance (EPR), and Förster resonance energy transfer (FRET).¹¹ Therefore, we will focus on classical molecular dynamics simulations to develop methods that allow us to obtain insights also on thermodynamics (free energies) and kinetics (diffusion coefficients and transition rates) in addition to the study of the transformation mechanisms.

4.2 CLASSICAL MOLECULAR DYNAMICS SIMULATIONS

By construction, molecular dynamics simulations model the time-dependent behavior of a system, thus, given a pressure and a temperature, they are ideal to study the mechanisms, thermodynamics, and kinetics of transformation processes.¹ In a nutshell, we can describe the idea behind molecular dynamics as follows: Starting from an initial configuration of the system, given by the positions of all the atoms, we estimate the force experienced by each atom given the interaction with the other atoms. The spatial position of the atoms can be predicted by solving numerically the equations of motion as a function of time, using small time steps to propagate the motion. Once the positions are updated, the forces are recomputed and the procedure is repeated. The result of this repetitive operation is a three-dimensional movie or trajectory containing the positions of all the atoms at each time step.¹¹ The following is a detailed description of the process to follow the time evolution of the system.

4.2.1 TIME EVOLUTION AND STATISTICAL PRINCIPLES

PHASE SPACE

In classical MD simulations a molecular system is made up of point-like particles representing each atom. The dynamics of the system is given by the motion of these particles in space. Therefore, to reconstruct a particle's trajectory we need its position $\mathbf{r}(t)$ and momentum $\mathbf{p}(t)$ at each time. We can define the phase space of the system at each time t , as the set of coordinates $x_i(t)$, $y_i(t)$, $z_i(t)$, $p_{x_i}(t)$, $p_{y_i}(t)$ and $p_{z_i}(t)$, for each particle i . For a system composed of N particles we then get a $6N$ dimensional space. The system's dynamics is described by the collection of positions in this phase space, i.e., by the trajectory in abstract space. The microscopical state of the system is then defined by the position at each point of the phase space. On the other hand, the system's macroscopic equilibrium state is given by an average over the microscopical states accessible by the system, as we will describe below, but first, we will focus on monitoring the dynamics of the system by tracking the time evolution of $\mathbf{r}(t)$ and $\mathbf{p}(t)$.

TIME EVOLUTION

The time evolution of the particles of the components of a system can be followed by solving the equations of motion. The analytical solution of these equations if only possible for very simple examples, therefore, for real systems it is necessary to solve them using approximate numerical methods. One of these methods is molecular dynamics.

Consider an isolated system (constant energy E) with a fix number of N particles in a box of constant volume V , i.e., a system in the micro-canonical or NVE ensemble. The Lagrangian of the system is given by the sum of the kinetic energy given by the movement of each atom and the potential energy due to the interaction between particles

$$\mathcal{L} = \sum_{i=1}^N \frac{1}{2} m_i \mathbf{v}_i^2 - U(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) , \quad (4.1)$$

where m , \mathbf{r} , and \mathbf{v} are the mass, the position, and the velocity of each particle in the system, respectively. The first term on the left hand side of Eq. 4.1 represents the kinetic energy of our system and the second term represent the potential energy due to interaction between the particles. The Legendre transformation allows to go from the Lagrangian to the Hamiltonian ($\mathcal{L}(\mathbf{r}, \mathbf{v}, t) \rightarrow H(\mathbf{r}, \mathbf{p}, t)$), given the momentum $\mathbf{p}_k = \partial_{\mathbf{v}_k} \mathcal{L}$

$$H = \sum_{i=1}^N \mathbf{p}_i \left(\frac{\partial \mathcal{L}}{\partial \mathbf{v}_i} \right) - \mathcal{L} . \quad (4.2)$$

By applying the least action principle to the Lagrangian, we can deduce the Euler-Lagrange equation

$$\frac{d}{dt} \left(\frac{\partial \mathcal{L}}{\partial \mathbf{v}_i} \right) = \frac{\partial \mathcal{L}}{\partial \mathbf{r}_i} . \quad (4.3)$$

Finally, from the last expression we can write the equation of motion

$$\mathbf{F}_i = m_i \frac{d\mathbf{v}_i}{dt} = - \frac{\partial U}{\partial \mathbf{r}_i} , \quad (4.4)$$

where \mathbf{F}_i is the force that act on the particle i . The potential energy U is a complex function that is currently approximated by using force fields (see next section). The analytical solution of Eq. 4.4 is not possible. To find the positions of the atoms in time, Eq. 4.4 must be solved numerically, discretizing time in small intervals, i.e., using a time step Δt . Then, at each iteration after one Δt , we calculate \mathbf{F} by means of the derivative of U and from this we use the equation of motion to update the positions and velocities of the atoms. To accurately preserve the dynamics of the system, Δt must be small enough so that the force can be considered constant between two time steps. This is achieved by using a Δt equal to or smaller than the fastest motion of the system (usually linked to fast bond vibrations). The need for such a small time step becomes a limitation for molecular dynamics simulations, since, as we will see below, in order to directly estimate equilibrium properties using molecular dynamics it is necessary to sample the system for large enough time, the extension depending on the phenomenon under study (ergodic principle, see below). These long time scales are unachievable in practice since they require the solution of the equation of motion such a large number of times that it is impossible even using current supercomputers.²³ Now we will we describe the ergodic hypothesis and the most common algorithms used to integrate the equation of motion, i.e., to update numerically the velocities and positions in a system.

ERGODIC PRINCIPLE

Understanding the usefulness of MD simulations involves understanding the statistical ideas behind them. The most important is the ergodic hypothesis which grants the access to ensemble averages in phase space by means of time averages. The ergodic hypothesis ensures that a macroscopical state of a system can be measured by either of these two routes: (*i*) by taking one sample and measuring its states over a long period of time or, (*ii*) by measuring the states of a large number of samples of the system from the same initial conditions. Now, if we want to estimate a physical quantity A , following route (*i*) we can compute the time average as

$$\bar{A} = \frac{1}{N} \sum_k A_k N_k , \quad (4.5)$$

where k is the observed microscopical state and N_k is the number of times we have measured the microscopical state A_k .

On the other hand, following route (*ii*)

$$\langle A \rangle = \sum A_k p_k , \quad (4.6)$$

where p_k is the probability to be in a microscopic state A_k according to statistical mechanics. In the canonical ensemble (*NVT*), p_k is proportional to the Boltzmann factor $p_k \propto \exp(-U/k_B T)$, with k_B and T as the Boltzmann constant and temperature, respectively.

At large N we have

$$\lim_{N \rightarrow +\infty} \frac{N_k}{N} = p_k . \quad (4.7)$$

As a consequence, the ergodic principle is constructed, i.e., the average over a long period of time is equal to ensemble average: $\langle A \rangle = \bar{A}$

This result is central to molecular dynamics since it tells us that we can reconstruct macroscopic properties of any system by simulating long enough trajectory. Now, let's go to the more technical aspects of MD describing the algorithms behind, defining the potential energy function and the description of temperature and pressure in MD.

4.2.2 INTEGRATION OF EQUATION OF MOTION

In MD the integration of the equation of motion is achieved numerically. Here we will briefly describe the most commonly used algorithms for this task: the Verlet and leap-frog algorithm. These algorithms display the most important properties required for an integrator to grant ergodicity:⁵⁷ conservation of the phase space volume, the energy and total momentum, and time reversibility.

Verlet and leap-frog are equivalent algorithms based on a second order Taylor expansion of the position (note that for simplicity we omit the sub-index i , but each expression is acting

over each particle)

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \frac{d\mathbf{r}(t)}{dt} \Delta t + \frac{1}{2} \frac{d^2\mathbf{r}(t)}{dt^2} \Delta t^2 + O(\Delta t^3). \quad (4.8)$$

Let's take the same expansion for $(r)(t - \Delta t)$ and sum the two expressions

$$\mathbf{r}(t + \Delta t) = 2\mathbf{r}(t) - \mathbf{r}(t - \Delta t) + \frac{d^2\mathbf{r}(t)}{dt^2} \Delta t^2 + O(\Delta t^3). \quad (4.9)$$

This leads to two ways of defining the velocities. The first gives rise to the Verlet algorithm:

$$\mathbf{v}(t) = \frac{\mathbf{r}(t + \Delta t) - \mathbf{r}(t - \Delta t)}{2\Delta t} + O(\Delta t^2). \quad (4.10)$$

The second defines the leap-frog algorithm:

$$\mathbf{v}\left(t - \frac{1}{2}\Delta t\right) = \frac{\mathbf{r}(t) - \mathbf{r}(t - \Delta t)}{2\Delta t} + O(\Delta t^2) \quad (4.11)$$

For instance, combining the leap-frog algorithm and the equation of motion, Eq. 4.4, we obtain

$$\mathbf{v}\left(t + \frac{1}{2}\Delta t\right) = \mathbf{v}\left(t - \frac{1}{2}\Delta t\right) + \frac{1}{m}\mathbf{F}(t)\Delta t + O(\Delta t^2), \quad (4.12)$$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}\left(t + \frac{1}{2}\Delta t\right)\Delta t + O(\Delta t^2). \quad (4.13)$$

Therefore, after initialization, in MD the position of the particle i is computed at time $t + \Delta t$ from its velocity at time $t + 1/2\Delta t$.

4.2.3 FORCE FIELDS

In MD simulations the true potential of a system is modeled through a Force Field (FF). Force fields are mathematical functions describing the dependence of the energy of a system on the coordinates of its particles and seek to reproduce the properties of a system. These models try to maintain the trade-off between simplicity to allow fast computation of the equations and a good level of detail to reproduce the intricate properties of the system. In these models, molecules are defined as sets of atoms that are held together by elastic forces corresponding to the different interactions between the atoms. The FFs then correspond to the analytical form of the interatomic potential $U(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$ with some parameters integrated into this form. The FF parameters are usually obtained either through computations (*ab initio* and semiempirical potentials) or by means of experiments such as X-ray, NMR, infrared, etc. Currently there are many FFs reported in the literature and available in MD packages that vary in complexity and in the way of obtaining the parameters, but in general they are based on this type of equation⁵⁸

$$\begin{aligned}
U = & \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2 + \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2 + \sum_{torsions} \frac{V_n}{2} [1 + \cos(n\phi - \delta)] \\
& + \sum_{impropers} V_{imp} + \sum_{LJ} 4\epsilon_{ij} \left(\left(\frac{\sigma_{ij}}{rij} \right)^{12} - \left(\frac{\sigma_{ij}}{rij} \right)^6 \right) + \sum_{elec} \frac{q_i q_j}{r_{ij}} .
\end{aligned} \tag{4.14}$$

In the Eq. 4.14 the first four terms refer to intramolecular interactions: bond stretching, angle bending, dihedral torsions and improper torsions, while the last two terms refer to repulsive and Van der Waals interactions (the Lennard-Jones (LJ) potential) and the Coulombic interactions. A schematic representation of these terms is shown in Fig. 4.1. Below we will briefly describe these terms.

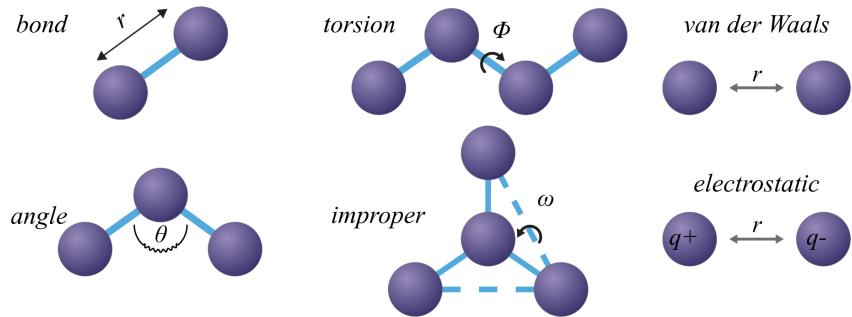


Figure 4.1: Schematic representation of the terms involved in the force field, Eq. 4.14.

INTRAMOLECULAR TERMS

The bond stretching is typically represented with a simple harmonic function (see first term in Eq. 4.14). This function controls the length of the covalent bond in the molecule. The spring constant k_b is usually parameterized with *ab initio* calculations or infrared and Raman experiments while r_0 is typically taken from X-ray experiments. For instance, considering common FFs for biomolecular simulations, one of the major limitations of MD lies in the description of the bonds by means of an harmonic function as it implies that bonds cannot be broken and therefore physicochemical processes involving bond breaking cannot be studied with classical MD simulations. Methods devised to overcome this limitation are beyond the scope of this manuscript.

In similarity to the bond stretching, angle bending (second term in Eq. 4.14) is also typically represented with a harmonic potential where the spring constant k_a is commonly parameterized from *ab initio* calculations or less frequently with experiments that allow to obtain the vibrational spectra. In addition to the bonds and angles between atoms, for molecules with more than four atoms in a row it is also necessary to include dihedrals and torsion terms. These terms are responsible for modeling the degree of rigidity of a molecule and are therefore essential for describing conformational changes.

The torsional energy is typically represented by a cosine function (third term in Eq. 4.14) where ϕ is the torsional angle, δ is the phase, n defines the number of minima or maxima between 0 and 2π , and V_n determines the height of the potential barrier. Usually, two or more torsional terms with different n can be combined to improve the description of complex systems allowing having minima with different depths.

The last term of the intramolecular interactions (fourth term in Eq. 4.14) is related to improper angles, accounting for the planarity of functional groups such as aromatic rings, sp^2 hybridized carbons, etc. A typical expression to describe the improper torsion is

$$U_{imp} = \sum_{impropers} \frac{k_{imp}}{2} (\omega - \omega_0)^2 , \quad (4.15)$$

where ω is the improper angle describing the deviation from planarity.

INTERMOLECULAR TERMS

The 12-6 LJ potential is often used to represent Van der Waals interactions (second last term in Eq. 4.14). Van der Waals forces can act between any pair of atoms belonging to different molecules (or to the same molecule if they are sufficiently separated). These forces result from the balance between repulsive interactions given by the overlap of the electron clouds of both atoms and attractive forces due to the interaction between induced dipoles.

The last term in Eq. 4.14 corresponds to the electrostatic interactions. The simulation of the electronic density in classical MD is typically approached by assigning a partial atomic charge to each atom, then the electrostatic contribution to the total energy is computed using Coulomb's law. The partial charges are typically derived from *ab initio* calculations. Note that Van der Waals and electrostatic interactions are long-ranged, therefore, they have to be computed for any pair of atoms of the system. The direct implication of this computation is a performance issue. In section 4.2.6 we show how to deal with these terms to maintain a good balance between accuracy and performance. We will now briefly mention the FFs most commonly used in practice.

POPULAR FORCE FIELDS

The first FFs appeared in the 1960s around the development of the molecular mechanics method. Since then force fields have been improved and generalized for application to large groups of materials. For example, some of the most popular families of FFs for the simulation of biomolecules have been CHARMM⁵⁹, AMBER⁶⁰, GROMOS⁶¹, while other FFs such as OPLS⁶² and COMPASS⁶³ are used for condensed matter simulations. These FFs have been improved over the years, leading to different versions of them.

There are also some FFs designed to describe particular systems. For instance, often when we simulate biomolecules in water, we typically use two different FFs, one for the biomolecule and another one for the water. A large number of water models has been proposed over the years. In this manuscript we use some of the rigid non-polarizable water potentials^{64,65}

such as TIP3P, TIP4P, TIP5P, SPC and SPC/E. In addition, another type of FFs has emerged: machine learning potentials.^{66,67} Machine Learning (ML)-based force fields aim to reduce the gap between the accuracy of *ab initio* methods and the efficiency of classical FFs. However, they lack the generality and transferability of classical FFs, and their applicability is restricted to the particular type of systems where they are trained.

Going back to the theory behind MD, so far we have been talking about the *NVE* ensemble, however, if we want to compare our MD measurements with experiments we need to go beyond and introduce control over the temperature and the pressure of the system. The next two sections focus on that.

4.2.4 TEMPERATURE COUPLING

The temperature of a system of N particles is given by its kinetic energy,

$$E_k = \frac{1}{2} \sum_{i=1}^N m_i \mathbf{v}_i^2 , \quad (4.16)$$

where m are the masses and v the velocities of each particle i . One can also write the kinetic energy in terms of the degrees of freedom of the system

$$E_k = \frac{1}{2} K_B T N_{df} , \quad (4.17)$$

where k_B is the Boltzmann constant and N_{df} is the number of degrees of freedom: $N_{df} = 3N - N_c - N_{com}$, determined by the number of constraints imposed over the system N_c , and the number of translational and rotational degree of freedom accessible by the center of mass N_{df} .

The temperature in MD simulation is fixed by coupling the system with a thermostat at constant temperature T . This setup leads us to the *NVT* ensemble. We will now describe two of the most commonly used thermostats.

BERENDSEN THERMOSTAT

To keep a fixed temperature, the Berendsen thermostat works in trying to suppress the fluctuation of the kinetic energy.⁶⁸ Note that this leads to a regime outside the NVT assembly where the sampling is not correct. The distributions of the kinetic energy and other fluctuation properties are incorrect. However, the error scales as $1/N$ so for large systems this effect is negligible.

The kinetic energy is scaled every time step with a time dependent factor λ as

$$\Delta E_k = (\lambda - 1) E_k , \quad (4.18)$$

where the scaling factor λ is given by

$$\lambda = \left[1 + \frac{\Delta t}{\tau_T} \left(\frac{T_0}{T(t - 1/2\Delta t)} - 1 \right) \right]^{1/2}, \quad (4.19)$$

where the parameter τ_T is directly related with the time constant of the temperature coupling τ_c

$$\tau_c = 2C_v \frac{\tau_T}{N_{df} k_B}, \quad (4.20)$$

where C_v is the total heat capacity of the system. As a consequence, when the velocity is rescaled, the energy difference is distributed between the kinetic and the potential energy.⁶⁸

The deviation of the system's temperature T from the thermostat temperature T_0 is corrected slowly by

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau_c}, \quad (4.21)$$

which means that the deviation between the two temperatures decays exponentially with the time constant τ_c . Therefore, the strength of the coupling can be defined through τ_c . In practice, one talks about short coupling time for $\tau_c \approx 0.05$ ps and long coupling time for $\tau_c \geq 2.0$ ps.

STOCHASTIC VELOCITY-RESCALING THERMOSTAT

The velocity-rescale thermostat uses the same principle that the Berendsen thermostat, adding a random term to ensure a correct energy distribution⁶⁹

$$dE_k = (E_k^0 - E_k) \frac{dt}{\tau_T} + 2 \left(\frac{E_k E_k^0}{N_{df}} \right)^{1/2} \frac{dW}{\sqrt{\tau_T}}. \quad (4.22)$$

In the last equation, E_k^0 corresponds to the thermostat kinetic energy and dW corresponds to a Wiener process (Brownian motion), i.e., a continuous random process with Gaussian distribution and no memory. In contrast to the Berendsen thermostat, the velocity-rescale thermostat preserves the correct NVT ensemble.⁷⁰

4.2.5 PRESSURE COUPLING

Another useful ensemble to compare with experiments is the *NPT* ensemble, where the number of particles N , the temperature T and the pressure P are fixed. Following analogous ideas to those used for the temperature, to set the pressure we couple our system to a barostat.

To set up a barostat, first we define an expression for the pressure. In a simulated system the pressure P can be calculate in terms of the kinetic energy and the internal pair potential using the virial theorem

$$P = \frac{2}{3V}(E_k - \Xi) , \quad (4.23)$$

where Ξ is the virial of the system

$$\Xi = -\frac{1}{2} \sum_{i < j} \mathbf{F}_{ij}(\mathbf{r}_i - \mathbf{r}_j) , \quad (4.24)$$

with \mathbf{F}_{ij} being the force on particle i due to particle j . As we did for thermostats, we now describe two of the most used barostats.

BERENDSEN BAROSTAT

As the pressure is defined in terms of the virial, the latter can be modified to fix the pressure. Therefore, in the Berendsen barostat the interparticle distance is scaled every step.⁶⁸ The scaling is anisotropic leading to a scaling matrix μ

$$\mu_{ij} = \left[\delta_{ij} - \frac{\Delta t}{3\tau_P} \beta_{ij} (P_{ij}^0 - P_{ij}(t)) \right]^{1/3} . \quad (4.25)$$

Here β is the isothermal compressibility of the system (for water $\beta = 4.6 \times 10^{-5}$ bar⁻¹, at 1 atm and 300 K) and \mathbf{P}^0 the fixed pressure of the barostat. The Berendsen barostat adds an extra term to the equations of motion which effects the pressure change, therefore, the energy has to be corrected by the work that the barostat applies over the system. Consequently, this barostat leads to large oscillations of the pressure and volume.

The equation of motion followed by the system is then

$$\frac{d\mathbf{P}}{dt} = \frac{\mathbf{P}^0 - \mathbf{P}}{\tau_P} , \quad (4.26)$$

where \mathbf{P}^0 is the reference pressure. The relaxation time τ_P can be set for a short or long time of relaxation, as for the thermostat explained above. Note that by construction, this barostat does not yield a true NPT ensemble because we are rescaling positions and not adding term to the hamiltonian⁷⁰, therefore more accurate barostats are required, as the one explained below.

PARRINELLO-RAHMAN BAROSTAT

In contrast to the Berendsen barostat, the Parrinello-Rahman barostat uses a new hamiltonian yielding a correct NPT ensemble⁷¹

$$H = \sum_{i=1}^N \frac{\mathbf{P}_i^2}{2m_i} + U(\mathbf{r}_i) + \sum_i P_{ii}V + \frac{1}{2} \sum_{ij} W_{ij} \left(\frac{db_{ij}}{dt} \right)^2 . \quad (4.27)$$

Giving the equation of motion

$$\frac{d^2\mathbf{r}_i}{dt^2} = \frac{\mathbf{F}_i}{m_i} - \mathbf{M} \frac{d\mathbf{r}_i}{dt}, \quad (4.28)$$

where,

$$\mathbf{W} = \mathbf{b}^{-1} \left(\mathbf{b} \frac{d\mathbf{b}'}{dt} + \frac{d\mathbf{b}}{dt} \mathbf{b}' \right) \mathbf{b}'^{-1}. \quad (4.29)$$

Here the matrix parameter \mathbf{W} determines the strength of the coupling, V the volume of the box and \mathbf{b} is the box vector represented as a matrix.

$$\frac{d^2\mathbf{b}}{dt^2} = \mathbf{V}\mathbf{W}^{-1}\mathbf{b}'^{-1}(\mathbf{P} - \mathbf{P}^0). \quad (4.30)$$

This barostat displays oscillations during the relaxation towards the reference pressure. As a consequence, if the current pressure is too far away from the reference temperature, the box may oscillate a lot, leading the crash of the simulation. Therefore, in practice one can start by balancing with the Berendsen barostat to prevent the simulation from collapsing and then in production one can use the Parrinello-Rahman barostat.

4.2.6 OTHER PRACTICAL CONSIDERATIONS

In this section we describe other key ingredients for accurate and efficient MD simulations.

PERIODIC BOUNDARY CONDITIONS

The molecular systems we are often interested in are composed of large numbers of atoms. For example, if we wanted to model water, one gram of water has about 3×10^{22} molecules, i.e., $3 \times 3 \times 10^{22}$ atoms. Modeling this large number of particles is, even today, very difficult. However, a realistic modeling of these molecular systems requires taking into consideration their large size. Since molecular systems are extremely large and expensive to simulate using MD, one can apply Periodic Boundary Conditions (PBC). The PBC enable us to mimic an infinite system by modelling a relatively small part of it. For this purpose, in MD the molecular system of finite size is modeled in a simulation box (unit cell) that is replicated in adjacent cells in all directions. During the simulation, all the particles are free to move in the original (central) cell. At the same time, their periodic images of the adjacent cells move in an identical way. For instance, for a cubic box of size $\mathbf{L} = (L_x, L_y, L_z)$, an atom located on $\mathbf{r} = (x, y, z)$ has a periodic image in $\mathbf{r} = (x + n_x L_x + y + n_y L_y + z + n_z L_z)$, with $\mathbf{n} = (n_x, n_y, n_z) \in \mathbb{N}^3$. Therefore, any particle that crosses one boundary of the cell, will reappear on the opposite side of this cell.

Although this scheme eliminates the edge effects that would be present when simulating the finite-size system, other unwanted artifacts are introduced. For example, in systems that are non-periodic in nature such as particles immersed in liquids (e.g., proteins in water), the periodicity will cause errors due to the un-physical nature of the replication. These artifacts

are reduced with increasing box size and are generally less severe than those caused by edge effects.

LONG AND SHORT-RANGE TERMS

As explained in section 4.2.3, the interaction between the atoms of the system is modeled through a force field that corresponds to the potential energy as a function of the positions of the atoms $U(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$. This potential is composed of terms involving pair and triplet particle interactions. Furthermore, we can classify these interactions into short and long-range interactions. Short-range interactions are those that decay faster than $1/r^3$ and long-range interactions are all other interactions with slower decay.

To improve performance one may require that these interactions are not calculated for all pairs of atoms but for the closest pairs, as follows: for short-range terms, we define a cutoff radius after which the interaction is zero. For particles farther away than this cutoff radius these interactions are not calculated, which reduces the computational time by reducing the number of calculations required. The minimum image convention is also used: only the closest particles, whether image or real, will be considered. In practice it should be noted that the cutoff radius used to truncate short-range interactions cannot exceed half of the shortest box vector, or else there would be more than one image within the cutoff distance of the potential.

In contrast, long-range terms should not be cut. For the long-range terms we use the Ewald summation scheme⁷², in which the long-range interaction sum is decomposed to achieve faster convergence

$$E_{long-range} = \sum_{\mathbf{n}} \sum_{i,j}^N \phi(\mathbf{r}_{ij,n}) = E_{dir} + E_{rec} + E_0 , \quad (4.31)$$

where ϕ represents the long-range term, E_{dir} is a sum in real space which contains screened short range interactions, and E_{rec} is a sum in reciprocal space which contains the long range interactions. The Fourier transform is used to communicate between the real and the reciprocal space. In this scheme, the sum term in reciprocal space scales with the square of the number of particles N^2 , therefore, for a large molecular system this term represents a limiting factor in terms of performance. The Particle-Mesh Ewald (PME) scheme⁷³ allows to improve the performance of the computation of the reciprocal term. To do so, the charge of the particles is assigned to a grid. Then, the Fourier transformation of the grid is computed and the reciprocal energy term is obtained by a single sum over the grid. The forces on each atom are obtained by performing the inverse transformation of the potential at the grid points. We highlight that this scheme is more suitable for large systems as it scales with $N \log(N)$.

So far we have discussed the technical details of MD simulations and how to obtain trajectories of the time evolution of a system. However, these trajectories are collections of atomic positions that, by themselves, do not provide much information about the system.

To interpret these trajectories and obtain the thermodynamic and kinetic information we are interested in, we often have to reduce their dimensionality, as we explain below.

4.3 FOLLOWING TRANSFORMATIONS IN REDUCED DIMENSIONALITY

Classical molecular dynamics simulations sample the phase space of the system and provide trajectories that describe the positions $x \in \mathbb{R}^{3N}$ of all the atoms in the system. In principle, at the long-time limit this collection of coordinates encodes all the properties of the system. However, the analysis of a high-dimensional phase space is intrinsically difficult. Therefore to allow interpretability of the data, we often aim to reduce the dimensionality of the problem. Working on an appropriate low-dimensional phase space allow us to gain insights about the mechanisms of the transformation processes, as well as to compute thermodynamic and kinetic properties.

This low-dimensional projections of the phase space are especially important in the study of rare events, i.e., processes that involve transitions between metastable states separated by sizeable free energy barriers, not only for the analysis, but also to improve the sampling of transitions between metastable states.⁷⁴ Details about using CVs to improve sampling are discussed in the Chapter 5. In the following, we will define these projections by introducing the difference between Collective Variables (CVs), Order Parameter (OP), and Reaction Coordinate (RC). In addition, we will give some examples of the CVs typically used for the study of processes such as conformational changes, protein folding, nucleation, molecular association and dissociation, etc., including the definition of the ideal CV: the committor.

4.3.1 COLLECTIVE VARIABLES, ORDER PARAMETERS AND REACTION COORDINATE

As explained above, the number of degrees of freedom in an MD simulation is extremely large. Therefore, defining a few CVs is useful to understand the transition mechanisms and to control and accelerate the phase space sampling.¹⁴ Following the definitions adapted by Peters⁷⁵ and Rogal⁷⁴, we define any function of the full phase space coordinates as a collective variable. However, a useful CV is a low-dimensional projection able to identify interesting collective phenomena over long timescales.¹⁴ On the other hand, an order parameter OP is a CV or combination of CVs able to distinguish between different metastable states of the system. Finally, a reaction coordinate RC is a function of atomic coordinates able to accurately capture the progress of a transition between two metastable states, i.e., the RC is a CV that simultaneously allows to separate two metastable states and captures the slow degrees of freedom of the transition between them.

4.3.2 QUALITY OF COLLECTIVE VARIABLES

The choice of the CV to study the transitions between states has a great impact on the interpretation of the mechanisms and on the accuracy of the estimated free energies and transition rates. For instance, for fixed P and T we can estimate the free energy profile as a function of a CV q , as the marginal of the equilibrium probabilities,

$$F(q) = -k_B T \log(\rho(q)) = -k_B T \log \left(\int dx \rho(x) \delta(q - q(x)) \right). \quad (4.32)$$

A free energy profile computed with Eq. 4.32 is simpler to analyze than the complex high-dimensional PES, however, depending on the quality of the CV the insights we gain from the interpretation of this profile are more or less useful. On the other hand, the kinetic estimates also depend crucially on the use of a good CV, as reaction-rate theories often rely on the identification of a suitable RC able to give good estimates of the free energy barrier and the progress of the transition.⁷⁴

To identify how good a CV is, let's take as an example the double-well potentials in Fig. 4.2. If we want to study the transition between A and B, depending on the coordinate we choose to project the free energy we will be able or not to discriminate between the two states through the free energy profile. In Fig. 4.2a, q_2 represents a bad reaction coordinate as it is not able to distinguish between the two states, while q_1 clearly separates both states and preserves the true height of the free energy barrier. On the other hand, in Fig. 4.2b, we show how even if the CV is able to properly distinguish the two states, two different landscapes can yield to the same free energy profile. However, only the CV q_1 on the left in Fig. 4.2b represents a good RC as the configurations with $q_1 = q^*$ are representative of the transition state, while on the right, configurations with $q_1 = q^*$ and small values of q_2 are committed to state A and configurations with $q_1 = q^*$ and large values of q_2 are committed to state B. As configurations at q^* on the right in 4.2b do not necessarily belong to the transition state ensemble, q_1 in this case is a poor CV that will not catch the proper dynamics of the system.

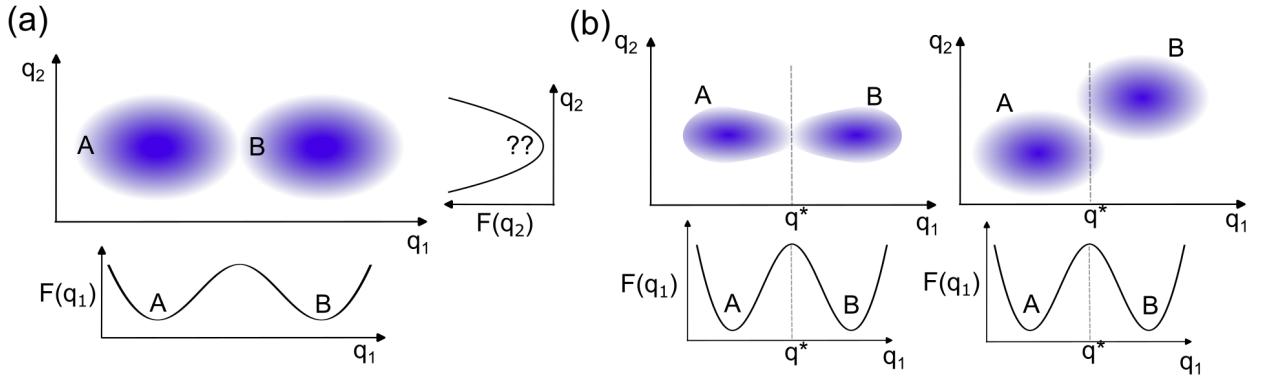


Figure 4.2: Schematic illustration of the CV quality. A schematic 2D double-well potentials illustrating different projections of the free energy landscape over a CV q . (a) 2D double-well free energy landscape and its projections over q_1 and q_2 . (b) Schematic representation of two free energy landscapes yielding the same free energy profile.

With this in mind, we ask ourselves: what is a good collective variable? A suitable RC to describe the progress of a transition between two states follow three main criteria proposed in ref. 75

- The RC is a function of only the instantaneous point in configuration space (velocities

are not included in the definition).

- The value of the RC should change monotonically between two states and the corresponding isosurfaces yield a set of non-intersecting dividing surfaces in configuration space.
- A free energy profile can be projected along the RC and the reduced dynamics along this RC are still consistent with the dynamics in the full phase space.

These criteria are general independently of the type of dynamics⁷⁴ (ranging from over-damped to inertial dynamics). In practice, defining a proper CV can be tricky. Traditionally, CVs based on physical intuition have been mostly used because of their ease of interpretation.¹⁴ However, physics-motivated CVs may not capture all relevant degrees of freedom for the transition between states. On the other hand, RC derived mathematically by construction encodes the dynamics of the process and is very accurate in describing the transition but they are difficult to interpret.⁷⁴ A combination of both approaches that maintains a trade-off between interpretability and accuracy may be useful. Different types of CVs are described below with some representative examples.

4.3.3 THE COMMITTOR

For tracking the transition between two states A and B, the collective variable defined as the ideal RC is the committor.⁷⁶ Also called commitment probability or p_{fold} (in the protein folding context), the committor $p_B(\mathbf{r})$ is defined as the probability that trajectories started from a given configuration \mathbf{r} will reach state B before reaching state A . Therefore, the committor provides a measure of the progress of the transition between states taking values from 0 (configurations are fully committed to A) to 1 (configurations are fully committed to B). For instance, the configurations with $p_B(\mathbf{r}) = 0.5$ belong to the TSE, as they have an equal probability to reach state A or B .

In principle, the committor fulfills all the criteria of an optimal RC presented in the previous section.⁷⁴ However, in practice this ideal RC poses some challenges: (i) from a computational point of view, the committor is computationally expensive. Its estimation requires running a large number of MD simulations starting from each configuration \mathbf{r} . (ii) By definition, the committor is almost identical to 0 and 1 except in a region spanning few $k_B T$ from the barrier top. This poses a major problem in its estimation except in the case of small barriers. (iii) It is difficult to interpret since it is not directly linked to any observable. These drawbacks limit their routine use and therefore, other types of CVs are defined.

4.3.4 HEURISTIC COLLECTIVE VARIABLES

Since we are usually interested not only in obtaining accurate results of thermodynamic and kinetic properties, but also in being able to interpret them mechanistically, heuristic or CVs based on previous knowledge of the system/process (here also called physically motivated CVs) serve as a first approximation to the description of the transition. Intuition-based CVs

have the advantage of being usually simple defined and easy to interpret. They can be defined by prior knowledge of the system or by similarity with other processes. For example, in the study of interactions between protein complexes one may be interested in their interpretation through interatomic distances, angles between groups of atoms, formation of specific contacts between amino acids, etc.¹⁴ In the case of protein folding some physically motivated CVs may be the dihedral angles, the number of native contacts or the Root Mean Square Deviation (RMSD). On the other hand, while nucleation can be described by the cluster sizes, the description of transformations in solids, is better described by changes in symmetry or local coordination.⁷⁴ In short, this type of CV depends entirely on the process under study.

It is important to mention another relevant aspect when choosing a CV: In addition to being used for trajectory analysis, collective variables are often used to accelerate sampling through enhanced sampling methods. This will be the subject of the next chapter. For now, we can advance that the derivation of intuition-based CVs for enhanced sampling methods requires that the CVs be differentiable with respect to the atom positions. Therefore, both the differentiability of the CVs as well as the computational cost of obtaining these derivatives play an essential role in its choice.⁷⁴

4.3.5 PATH COLLECTIVE VARIABLES

Another interesting group of CVs are the path collective variables.⁷⁷ The basic idea behind path CVs is to start from a sequence of n reference configurations, representing a path connecting A and B by means of equally spaced configurations across the transition. Then the progress along the path and the distance to the path can be tracked from the CVs $S(\mathbf{r})$ and $Z(\mathbf{r})$, respectively. The path CV $S(\mathbf{r})$ represents the progress along the reference path: $S(\mathbf{r}) = 0$ indicates that the current configuration is in state A , while $S(\mathbf{r}) = 1$ means that the current configuration is in state B . The S path CV is defined as

$$S(\mathbf{r}) = \frac{1}{n-1} \frac{\sum_{i=1}^n (i-1) \exp(-\lambda D(\mathbf{r}^i, \mathbf{r}))}{\sum_{i=1}^n \exp(-\lambda D(\mathbf{r}^i, \mathbf{r}))}, \quad (4.33)$$

where $D(\mathbf{r}^i, \mathbf{r})$ is a distance metric between the reference configuration \mathbf{r}^i and the instantaneous configuration \mathbf{r} . On the other hand, $Z(\mathbf{r})$ measures the cumulative distance from the reference path. The reference path is not necessarily the minimum free energy path, so the trajectory can move outside this path if another path is more favorable. Therefore, large values of $Z(\mathbf{r})$ mean that the system is moving away from the reference path. the Z path CV is defined as

$$Z(\mathbf{r}) = -\frac{1}{\lambda} \log \left(\sum_{i=1}^n \exp(-\lambda D(\mathbf{r}^i, \mathbf{r})) \right). \quad (4.34)$$

In both cases, λ is a parameter that controls the shape of the collective variable space, enlarging or reducing the width of the space along S when studying the transition process. In practice, a general rule of thumb is to choose it such that $\lambda D(\mathbf{r}^i, \mathbf{r}^{i+1}) \approx 2.3$. Different

distance D metrics can be used depending on the problem under study such as the mean square displacement or the Permutation Invariant Vector (PIV).

4.3.6 DIMENSIONALITY REDUCTION AND MACHINE LEARNING BASED CLASSIFICATION

Since the use of CVs implies a projection of the high-dimensional phase space in a low-dimensionsal one, the problem of finding a good CV is essentially a problem of dimensionality reduction. Therefore, dimensionality reduction schemes can be applied to identify CVs. The simplest dimensionality reduction method is Principle Component Analysis (PCA), in which the target CV is the linear combination of coordinates that best captures the variance of the dataset. Other dimensionality reduction techniques used to find CVs are: multidimensional scaling⁷⁸, kernel PCA⁷⁹, Isomap⁸⁰, Sketch-maps⁸¹, diffusion maps⁸², among others, which are based on different distance spaces or introduce non-linearity.⁷⁴ These schemes usually require MD data spanning over all the transition. However, the availability of a large dataset is not always guaranteed. To overcome this disadvantage, some ML methods have been developed.

Some ML-based approaches take as training data only the sampling performed in metastable states. Normally, doing MD simulations at these free energy minima is easier than sampling the transition. From this data, the ML algorithms can learn to classify the two states and give clues as to how to go from one state to another. These classifiers can be used as CVs. Although these approaches are flexible in the type of structures that can be used for learning⁸³, e.g., support vector machines, logistic regression, and neural networks; the trajectory data still has to be initially projected into a set of descriptors which can be more general and easier to define than a CV, but still require some prior knowledge of the system. The CVs resulting from ML-based methods are usually not optimal because they are trained only with data from the metastable states, however they can be a good starting point for enhanced sampling strategies.

Although the CVs discussed in this section are well defined, they lack physical interpretability. This shows that the problem of finding optimal collective variables is still open. Our strategy to optimize CVs that preserve the dynamics of the system and allow physical interpretation of the transition mechanism will be discussed in the Chapter 10.

5

Enhanced sampling methods

5.1 BACKGROUND

The characterization of transitions in rare events poses a challenge for MD simulations. As we explained in the previous chapter, the sampling of the phase space in MD is achieved through the integration of the equations of motion at each time step. This operation is computationally expensive, limiting the time scales accessible with MD.^{12,21} In Fig. 5.1 we show approximately how time scales change with the height of the free energy barrier. Let's take as an approximate example of the computational effort needed a transition between metastable states separated by a free energy barrier of $30 k_B T$ (a rare event). Doing a back of envelope calculation using the TST, at room temperature would take an average transition time of the order of seconds. Using a typical time step of 2 fs it translates to 10^{14} integration steps. Furthermore, typical events of our interest occur on even longer timescales (see Fig. 5.1). In practice, classical all-atom MD simulations are limited to the microsecond time scale.¹

Enhanced sampling methods have been developed to overcome this limitation. In general, depending on the enhanced sampling method used, it is possible to access three different features, the last one being the most complex: (i) Quick exploration of the configurations corresponding to metastable states; (ii) reconstruction of free energy landscapes by recovering the original statistical ensembles; (iii) characterization of rates and transition times.

The exploration of the phase space and in particular the reconstruction of the free energy profiles can be achieved with methods such as: Adaptive Biasing Force^{86,87}, Umbrella Sampling⁸⁸, Metadynamics²⁶, among others. A recent review with details on these methods can be found in Ref. 89. Methods aimed at the last feature, i.e., accurate estimation of kinetic rates are at present not routinely used due to their complexity.² Among them are MSM^{18,19}

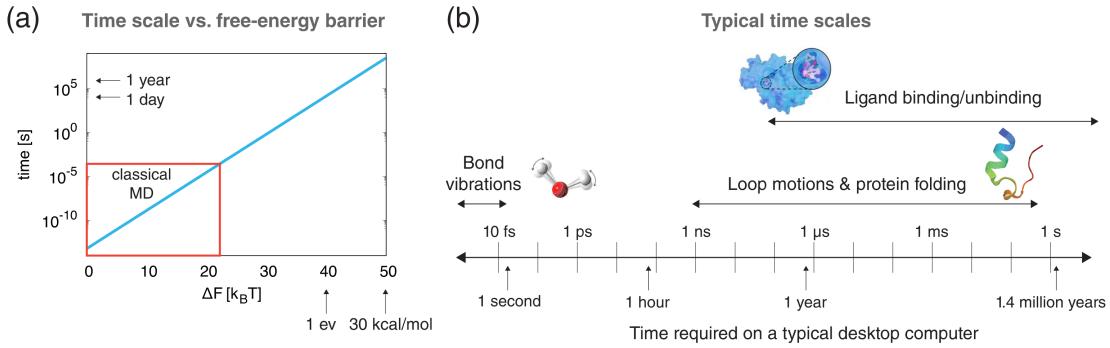


Figure 5.1: Schematic illustration of the time scales issue in MD simulation. (a) Approximate average time scale needed to simulate as a function of the height of the free energy barrier (adapted from Ref. 84). The red square indicates the time scales reachable with classical MD simulations. (b) Examples of time scales of typical interesting events compared with the time required to simulate with MD them in a desktop computer. Adapted from Refs. 23,85.

and TPS-based methods such as: TIS³⁸ and FFS.^{39,40}

In the following sections we focus on the methods used in this thesis. We divided them into two classes: the enhanced sampling methods that add bias on a CV, section 5.2; and unbiased enhanced sampling methods focused on the sampling of transition paths, section 5.3.

5.2 BIASED ENHANCED SAMPLING METHODS

In section 4.3.2 we defined the free energy landscape $F(q)$, Eq. 4.32, as a function of $q(x)$ with $x \in \mathbb{R}^{3N}$. For a suitable RC, the free energy landscape features the regions of CV space that are least likely to be explored (regions with high free energy as the barrier). Therefore, starting from the notion of $F(q)$ one can develop methods to enhance the sampling in poorly populated regions at the equilibrium: an external potential energy or bias $V_B(q)$ can be added to the free energy landscape to increase the probability of sampling high free energy regions. As a result of the addition of the bias potential, a force $-dV_B(q)/dq$ acts on the system counteracting the average force $-dF(q)/dq$. The aim is that the trajectory over the new artificial potential $F(q) + V_B(q)$ behaves like a random walk in a flat FES.

In theory, this procedure would result in a uniform sampling of the entire FES in a shorter computation time (since there is no need to overcome large free energy barriers). However, in practice the implementation of biased enhanced sampling methods is more complicated because the shape of the free energy landscape $F(q)$ is not known in advance. Therefore, different strategies have been developed to build bias potentials that effectively counteract $F(q)$. These enhanced sampling methods are divided into two categories: methods with static bias potential and methods where the bias potential is built on-the-fly. In this thesis we use a method from each of these categories, as explained below.

5.2.1 UMBRELLA SAMPLING

Umbrella Sampling (US)⁸⁸ is an enhanced sampling method where the bias potential is constant during the MD simulation. In US a quadratic potential (umbrella) is often used to restrict a trajectory in a predefined region of configuration space. The bias potential is defined as

$$V_i(q) = \frac{k_u}{2}(q - q_i)^2 . \quad (5.1)$$

For strong k_u , the trajectory will be forced to sample only a small region (window) centered around q_i . Therefore, several MD simulations are launched spanning the q space to ensure proper sampling. A representation of the US potential and a trajectory restricted to a window i is shown in Fig. 5.2a. The sampled distribution in window i is then

$$\rho_i(q) = \frac{\rho(q) \exp(-V_i(q)/k_B T)}{\int dq' \rho(q') \exp(-V_i(q')/k_B T)} . \quad (5.2)$$

The combination of the distributions ρ_i for all windows allows to recover the underlying unbiased distribution ρ , yielding $F(q)$. A popular method used to combine the US distribution and reconstruct free energy profiles is the Weighted Histogram Analysis Method (WHAM)⁹⁰. In WHAM a self-consistent estimator of the free energy is build based on the sampled distributions ρ_i . We will not go into the details behind WHAM since in this thesis, we use US for the estimation of diffusion coefficients (see section 9.2.2), instead of using it to estimate free energy profiles. For the latter purpose, we use metadynamics as explained in the next section.

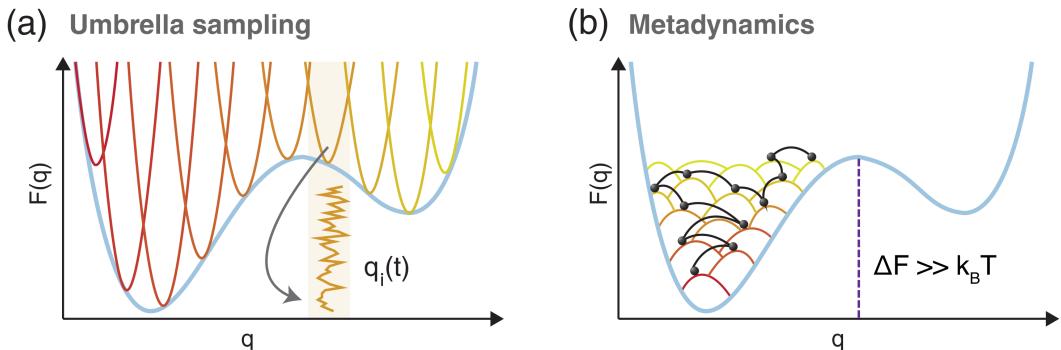


Figure 5.2: Schematic representation of the umbrella sampling and metadynamics methods. (a) Representation of the umbrella sampling bias potential over a 1D free energy landscape. Each color represents a quadratic potential (umbrella) that restrain the system into a region of the CV phase space (window). An illustration of a typical trajectory in one window k is also shown. (b) Representation of the metadynamics bias potential over the same 2D free energy potential as in (a). The bias potential is built on-the-fly as a sum of Gaussian functions. Each color represents a different Gaussian added to $F(q)$. An schematic illustration of the position of the system after the addition of a new Gaussian is also shown. Figures are adapted from Ref. 84.

5.2.2 METADYNAMICS

Metadynamics²⁶ (MetaD), is an enhanced sampling method where the sampling is accelerated by adding a history-dependent bias potential to the force-field as a function of one or few CVs^{26,91}. The bias potential corresponds to a sum of Gaussians added to the positions of the CVs at regular intervals of time ($t_d, 2t_d, \dots, nt_d$), i.e., for a CV q and a bias deposition time t_d , a Gaussian function of height ω and width σ_q is added to the system every t_d time units (Fig. 5.2b). Each Gaussian is centered at the current position of the system $q(t = kt_d)$, as described by the following equation

$$V(q, t) = \omega \sum_{k=1}^n \exp\left(-\frac{(q - q(kt_d))^2}{2\sigma_q^2}\right). \quad (5.3)$$

The bias potential in Eq. 5.3 progressively builds a potential that fills or counteracts the underlying unbiased potential. As a consequence, the system can escape from the local minima and overcome free energy barriers in a computationally efficient manner. In the long time limit for a converged simulation and appropriate CVs²⁷, it is possible to recover the free energy of the system within fluctuations (that can be reduced by time averaging the V_B profiles in a time interval at convergence)⁹² by changing the sign of the MetaD bias.⁹³

$$F(q) = -V_B(q, t) + C, \quad (5.4)$$

where C is a constant resulting for the accumulation of bias over time. A key test to assess convergence in MetaD simulations consists in detecting whether the q histograms are flat in different time intervals after the filling. In general, convergence in MetaD is not assured in a finite and affordable simulation time, as it depends on the choice of the CVs and on the detailed features of the process. Moreover, the time required for convergence increases exponentially with the number of relevant degrees of freedom of the system.²⁸ Therefore, variants to the standard MetaD (fixed Gaussian height and width) have been developed. The most widely used is explained in the next section.

WELL-TEMPERED METADYNAMICS

In Well-Tempered Metadynamics (WT-MetaD)⁹⁴ the Gaussian height ω is reduced as the simulation progresses

$$\omega = \omega_0 \exp\left(\frac{-V_B(q, t)}{k_B \Delta T}\right), \quad (5.5)$$

where ω_0 is the initial Gaussian height provided by the user. The parameter ΔT controls the scaling of the Gaussian height. For $\Delta T \rightarrow 0$ the sampling is equivalent to unbiased MD, while $\Delta T \rightarrow \infty$ correspond to standard MetaD, i.e., fixed Gaussian height. With WT-MetaD, at convergence, the bias potential stops fluctuating and the free energy multiplied by the bias factor $\Delta T/(T + \Delta T)$ can be recovered after checking that the q histogram is flat in proper time intervals.

PRACTICAL CONSIDERATIONS

Regardless of whether we use MetaD or WT-MetaD, in practice, no one can predict is a given simulation will converge, and in fact, often MetaD simulations do not converge for several reasons: (i) Depending on the CVs used, the estimation of their derivatives can be more or less complex, this being a limiting factor for the time scales accessible with MetaD. (ii) Although one might be tempted to use a small deposition time (adding Gaussians very frequently) to speed up the scan, t_d must be large enough to allow relaxation of the system between two bias depositions. (iii) The height and width of the Gaussians cannot be increased indefinitely either to speed up the convergence, since the resolution of the reconstructed free energy landscape depends on these parameters. (iv) the extra step of adding the bias potential and estimating the forces associated with it has the consequence that the computational performance per time step is always less than that of pure MD simulations. (v) Above all, any sampling and estimation in MetaD depends on the CVs used. When a poor CV is used, that is, a variable in directions orthogonal to the transition we want to model, the added bias does not help speed up the transition (see Chapter 6 for a detailed discussion) and artificial hysteresis can be observed, as a consequence of lack of convergence.

In addition to the convergence problem in MetaD, although static information like equilibrium populations can be easily recover in post-processing, time-related information like transition rates is nor directly available. In MetaD, the dynamic information of the system is corrupted due to the added bias, since the system explores an artificial potential. Given the difficulty of achieving convergence in MetaD simulations and on recovering the dynamics information of the system, methods that are capable of extracting kinetic information from short non-converged MetaD simulations are useful. To advance the state of the art in this direction, in the Chapter 6 we present a novel method for this purpose.

5.3 UNBIASED ENHANCED SAMPLING METHODS

5.3.1 TRANSITION PATH SAMPLING

In the previous section we introduced examples of enhanced sampling methods that increase the probability of sampling high free energy regions by adding a bias that modifies the shape of the sampled potential. Now we present a different class of methods, that focus on sampling and collecting a large number of transition paths, i.e., trajectories connecting a initial state A with a final state B. Specifically, in this section we will talk about Transition Path Sampling (TPS)³⁶ which is the most widely used method for sampling transition pathways.¹ The basic principle of TPS is to concentrate the sampling in the transition region. Usually, in a conventional MD simulation, most of the time the system remains stuck in the minima. From time to time (with a frequency that decreases with increasing barrier size), the system may jump to another state. In rare events, the average waiting time in the minima is much larger than the transition time. Therefore, the exploration of transition path ensembles only requires obtaining short transition paths which is cheaper than extensively sampling the

entire phase space.

The general TPS algorithm works as follows: starting from an initial reactive trajectory (a trajectory that goes from A to B), a configuration within this trajectory is chosen. From this configuration, two new trajectories (shootings) are initiated with equal velocities but in opposite directions, i.e., one trajectory is propagated backward and the other forward in time. The initial velocities for these new trajectories are obtained by making small perturbations of the velocity in the previous trajectory or are randomly generated from a Boltzmann distribution. If by joining these two new trajectories, the resulting trajectory connects states A and B, then it is accepted, otherwise it is rejected, and the procedure is iterated, see Fig. 5.3. A transition path ensemble can be generated through a simple Monte Carlo (MC) scheme in path space in which new shooting points are randomly generated. The resulting Transition Path Ensemble (TPE) is meant to correspond to the equilibrium one. Therefore it contains the information about the mechanism and dynamics of the activated process. In this wealth of information lies the usefulness of TPS.

In practice, there are several considerations to take into account to implement TPS. It is evident that the algorithm needs a preliminary definition of states A and B: although in principle, TPS is not a CV-based method, for the definition of the metastable states some notion of a low-dimensional space is needed. Additionally, as explained above, an input reactive trajectory is required for the initialization of the algorithm: usually other FES exploration techniques such as those explained in section 5.2 are required to obtain it. Moreover, in TPS it is not trivial how to obtain computational efficiency and assess the convergence. In the next section we describe in detail the specific TPS algorithm used in this work to generate collections of transition paths, other algorithms are reviewed in Ref. 37.

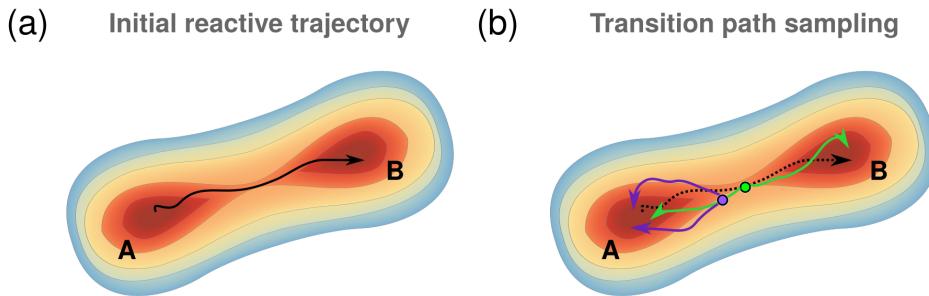


Figure 5.3: Schematic representation of the transition path sampling method. (a) Initial reactive trajectory generated with other sampling methods. This trajectory connects states A and B previously defined by the user. (b) Illustration of the general TPS algorithm. The initial trajectory is discretized (black arrow) and a shooting point is chosen (green dot). Two trajectories are propagated backward and forward from the shooting point. If the trajectories connect A and B are accepted (green arrows). A new shooting point is chosen from the last accepted trajectory (purple dot) and the process is repeated. If trajectories fall into the same state they are rejected (purple arrows).

AIMLESS SHOOTING ALGORITHM

Aimless Shooting (AS)^{95,96} is a procedure for the automatic generation of transition path trajectories. In AS, a trial shooting point is drawn from the vicinity of the last accepted point, leading to one of the biggest advantages of AS: it is able to find new shooting points without needing to use an order parameter to describe the transition region. Also to one of its biggest disadvantages: in some systems the generated transition path sequence tends to pass through a narrow region, therefore, usually a large number of iterations are needed to reach decorrelation of the sampled transition paths.

The AS algorithm starts from an initial reactive trajectory of length t_{total} and follows an iteration of the next steps:

1. The initial (or the last accepted trajectory) is discretized in a series of candidate shooting points with a separation in time $\Delta t \ll t_{total}$.
2. From the last shooting point, frame i , select randomly as a new shooting point with equal probability the frame $i + \Delta t$ or $i - \Delta t$.
3. From the new shooting point, propagate the system backward and forward in time, with momenta drawn from a Boltzmann distribution. The trajectories run until they reach state A or B or until they reach a predefined maximum duration.
4. Accept the new trajectory (combination of the backward and forward shootings) if it connects states A and B.

The acceptance rate in this algorithm is correlated with Δt . For $\Delta t \rightarrow 0$, the acceptance rate r goes to 100%, as the same region of the TSE as in the initial trajectory is being sampled. In contrast, when $\Delta t \rightarrow t_{total}$, $r \rightarrow 0$ as the trajectories are mostly committed to the final state where initial trajectory end up. Therefore Δt should be chosen aiming the acceptance rate that yields the largest sampling of decorrelated paths for a fixed amount of computational time.

Improvements to the AS algorithm to achieve faster decorrelation have been proposed. For example, aimless shooting within a range⁹⁷ that aims to reach decorrelation more quickly, while maintaining a high acceptance rate. The biggest difference of the latter algorithm with respect to traditional AS is that the new shooting point is chosen from a range of points close to the last accepted shooting point. The range is defined from the values of a CV that describes the transition, this being the greatest drawback of the algorithm since it assumes prior knowledge of the RC. Then the new trajectory is accepted with probability $p_{acc} = \min(1, n/n')$. Where n and n' are the number of shooting points that are within the shooting range, for the new and the old trajectory, respectively. The efficiency of this algorithm depends on the size of the chosen range. A scheme with the comparison of the algorithms to choose new sampling points is shown in Fig. 5.4.

We remark that the TPS techniques presented here are valuable in providing unbiased transition paths. However, to obtain free energies and rates using TPS other more sophisticated TPS-based algorithms as TIS or FFS are required. These algorithms are not trivial to

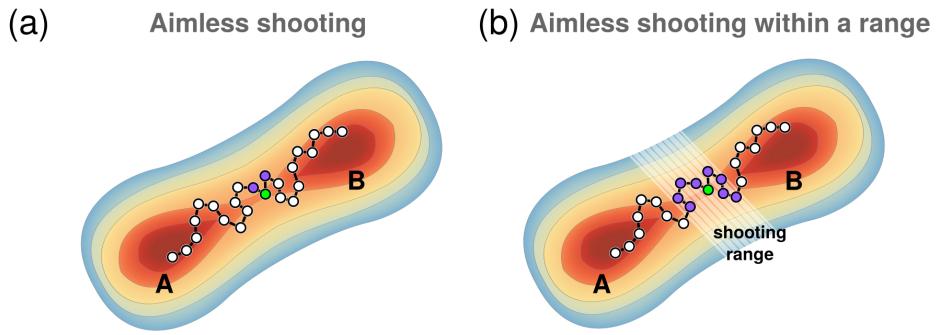


Figure 5.4: Comparison between aimless shooting algorithms. Green dots represent the last accepted shooting point, while purple dots represent the new possible shooting points (a) Schematic representation of the AS algorithm. Only frames $i + \Delta t$ or $i - \Delta t$ can be taken as new shooting points (b) Schematic representation of the AS within a range algorithm. All frames inside the predefined range (white pattern) can be chosen as new shooting points.

use, require large amounts of computational time and often resort to CVs. Therefore, methods that are capable of extracting thermodynamic and kinetic information from transition paths generated with, for instance AS, are desirable.

In this chapter we explored the enhanced sampling methods used in this thesis to effectively sample the FES. Despite their usefulness, directly accessing the kinetic information of activated processes through these enhanced sampling methods is still a challenge. In the following chapters we will show the state of the art methods to estimate transition rates from these enhanced sampling techniques and their drawback, as well as the two new methods we propose to overcome those flaws and extract the dynamic information from MetaD and TPS.

Part III

Results and Discussion

6

Transition Rates and Efficiency of Collective Variables from Time-dependent Biased Simulations

6.1 INTRODUCTION

Simulations with an adaptive time-dependent bias, such as metadynamics, enable an efficient exploration of the conformational space of a system. However, the dynamic information of the system is altered by the bias. Infrequent Metadynamics ([iMetaD](#)) allows to recover the transition rate of crossing a barrier by rescaling the simulation transition times with an acceleration factor accounting for the bias speed up and comparing the average time of escape from a fit to a Poisson distribution. [iMetaD](#) is based on [TST](#), therefore it requires the use of ideal collective variables and no bias deposition near the transition state. Unfortunately, for simulations of complex molecules, these conditions are not always fulfilled. To overcome these limitations, and inspired by single-molecule force spectroscopy, in this chapter we introduce a method that we recently developed based on Kramers' theory for calculating the barrier-crossing rate when a time-dependent bias is added to the system. We assess the *efficiency of the collective variables* parameter by measuring how efficiently the bias accelerates the transitions compared to ideal behavior. We present approximate analytical expressions of the survival probability that accurately reproduce the barrier-crossing time statistics, and enable the extraction of the unbiased transition rate even for challenging cases, where previous methods fail.

6.1.1 KINETIC RATE ESTIMATION FROM TIME-DEPENDENT BIASED SIMULATIONS

Kinetic rate constants are of fundamental importance by quantifying the speed of interconversion between metastable states in the description of physical phenomena. However, their estimation is not a trivial task. Computer-assisted methods, in particular MD simulations, have gained relevance in recent decades for the prediction of thermodynamic and kinetic properties of complex processes such as protein folding or ligand binding.^{10–12,20,98–101} However, MD has the limitation that many interesting events occur at long timescales (rare events) not accessible by standard simulations, even using powerful supercomputers.²³

Enhanced sampling methods such as metadynamics²⁶ in combination with MD simulations have become useful alternatives for studying rare events.¹ MetaD accelerates the conformational sampling by adding a history-dependent bias potential to the force-field. The biasing potential is a function of CVs chosen to describe the degrees of freedom considered most relevant to the transition mechanism.^{26,91} For converged simulations and appropriate CVs, MetaD allows the direct estimation of the free-energy profile of the system using the MetaD bias.⁹¹ A disadvantage of MetaD, as well as other enhanced sampling methods, is that information about the dynamics of the simulated system is corrupted due to the sampling acceleration.²⁷ Therefore, the estimation of the rate coefficients from biased simulations is a difficult target.

Several methods have been developed to estimate rate coefficients from enhanced-sampling simulations.^{2,22,102,103} Some involve the calculation of diffusion coefficients and the construction of Markov State Models.^{29–32,104–109}

6.1.2 TRANSITION RATES FROM METAD AND CURRENT LIMITATIONS

In the MetaD framework, infrequent metadynamics³² (iMetaD), has been widely used in recent years.^{34,110–115} iMetaD aims to recover the escape time by rescaling the simulation time with an acceleration factor.^{32,116,117} This method requires a slow biasing frequency and small set of CVs that determine the relevant states and pathways of the system.²² When these conditions are satisfied, the distribution of escape times follows a Poisson behavior.¹¹⁸ Despite its usefulness, a major limitation of iMetaD is to rely on ideal CVs, where the bias potential is zero at all the dividing surfaces.^{119,120} For example, some iMetaD tests failed for MD simulations of the biotin-streptavidin complex for biases along non-ideal CVs at different pulling forces.¹¹⁵ Modifications of iMetaD have thus been proposed.^{121,122} However, these rescaling-based methods do not directly compute the time-dependent rate or barrier-crossing survival probably due to the bias.

6.1.3 OUR PROPOSAL

Here we introduce our recently published method¹²³, in which we take inspiration from dynamic force-spectroscopy experiments, where a force is ramped up with time similar to the simulations with a dynamic bias. For these experiments, accurate kinetic predictions of the force-dependent rates and transition probabilities have been derived.^{124–127} By considering

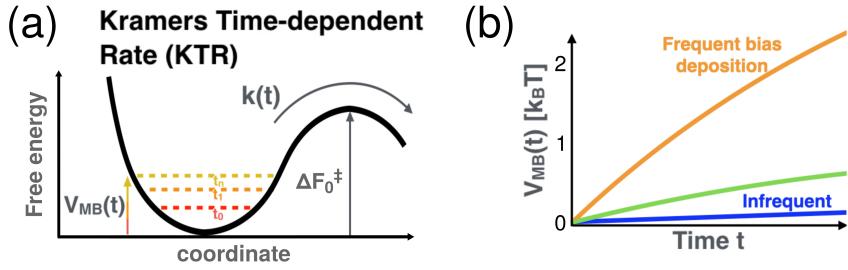


Figure 6.1: Kramers time-dependent rate for biased simulations and assessment of the efficiency of the CVs. (a) Schematic representation of the escape from a bottom-well with a time-dependent bias. $\gamma V_{MB}(t)$ measures the effective contribution of the added bias-height toward lowering the effective barrier. (b) Examples of $V_{MB}(t)$ for frequent and infrequent bias-deposition times. The results are, for the double-well potential example, $V_{MB}(t)$ along y with bias-deposition times $t_d = 1$ (orange), 5 (green), and 20 (blue) ($\times 10^3$) steps within the first 3×10^5 steps.

the biasing potential analogous to an external force, we introduce a physical model of barrier-crossing events in time-dependent biased simulations for computing directly the transition statistics. The major advantages are that one can extract the unbiased rate and, at the same time, assess the efficiency of the CVs in terms of their contribution to the bias acceleration. This chapter is organized as follows: i) we describe in detail our barrier-crossing model and our adaptation of Kramers' theory to metadynamics simulations, ii) we explain the technical details and simulations methods, and iii) we test the method on a 2D potential that allows to study the effect of the quality of the CV, finding transition rates with equal or better accuracy than those extracted from iMetaD and gaining insights about the quality of the CV at the same time.

6.2 THEORETICAL METHODS

6.2.1 KRAMERS TIME-DEPENDENT RATE

For diffusive dynamics, and high barriers, Kramers' theory^{43,44} is used to calculate the rate (i.e., inverse of the mean residence time) to cross a barrier along a coordinate

$$k_0 = k_{\text{pre}} e^{-\beta \Delta F_0^\ddagger}, \quad (6.1)$$

where ΔF_0^\ddagger is the barrier height (Fig. 6.1a), $\beta = 1/k_B T$ is the inverse temperature, and k_{pre} is the pre-exponential factor that depends on the diffusion coefficient, shape of the bottom-well and barrier-top.

In MetaD short-range repulsive functions are deposited at regular time intervals within a low-dimensional CV space. Over time, the bias fills the well. This reduces the effective barrier experienced by the system which is thus crossed more rapidly. To describe this reduction of the barrier height, we use the time-dependent maximum bias (MB) averaged

over multiple runs

$$V_{\text{MB}}(t) = \frac{1}{R} \sum_r \max_{t' \in [0, t]} V_B^r(t') , \quad (6.2)$$

where $V_B^r(t)$ is the instantaneous bias at time t for simulation run r , and R is the total number of runs. $V_{\text{MB}}(t)$ is the average maximum height of the biasing potential (i.e., the level of bias added to the bottom-well) up to time t . In the case of an ideal CV, $\Delta F_0^\ddagger - V_{\text{MB}}(t)$ would be the effective time-dependent barrier experienced by the biased system.

$V_{\text{MB}}(t)$ depends on the shape of the potential surface, the bias-deposition time (t_d), and bias-deposition height, among others. In Fig. 6.1b, we present some examples of $V_{\text{MB}}(t)$ for several t_d . The blue line shows the case for which iMetaD is valid. We will compute directly the statistics for barrier-crossing times from biased simulations, covering a wider range of $V_{\text{MB}}(t)$, by explicitly taking into account their time dependence.

Assuming a quasi-adiabatic bias deposition, we apply Kramers' theory over the potential presented in Fig. 6.1a, to calculate the time-dependent rate of escape due to the bias acceleration

$$k(t) = k_{\text{pre}} e^{-\beta \Delta F_0^\ddagger + \beta \gamma V_{\text{MB}}(t)} = k_0 e^{\beta \gamma V_{\text{MB}}(t)} , \quad (6.3)$$

where k_0 is the intrinsic rate (Eq. 6.1), and we are assuming that k_{pre} does not change due to the bias. We introduce $\gamma \in [0, 1]$ as an additional parameter that measures how much of the bias contributes to the acceleration. For ideal CVs, i.e., where the added bias acts along the direction of the true transition and helps to lower the effective barrier, we expect $\gamma = 1$. By contrast, we expect $\gamma \approx 0$ for poorly chosen CVs, i.e., where the bias acts in directions orthogonal to the transition.

6.2.2 RATE-ACCELERATION FACTOR: EFFICIENCY OF CVs

We illustrate that γ is related to a physical parameter, which controls the efficiency of the CV, for a cusp-like harmonic double-well potential (see Eq. 6.4).

Consider the diffusive escape from a 2D double-well potential over a cusp-like barrier. The potential is given by

$$V(x, y) = \begin{cases} \frac{y^2}{2} + \frac{(x-ay)^2}{2}, & \text{for } y \leq 1 , \\ -\infty, & \text{for } y > 1 , \end{cases} \quad (6.4)$$

where parameter a controls the separation between the wells along x (see Fig. 6.2). The bottom-minimum is at $x = y = 0$, and the barrier for escape at $x = a$ and $y = 1$. The barrier height is $\Delta F_0^\ddagger = \frac{1}{2}$. Fig. 6.2 shows examples of the potential for $a = 0.5$ and $a = 1 \times 10^{-5}$. We consider diffusion on this potential with a uniform and isotropic diffusion coefficient at an inverse temperature $\beta \gg 1$, i.e., in the high-barrier limit. Then, for $a \approx 0$, y is a good CV because it reports on the approach to the barrier, whereas x is a poor CV because it reports on motions orthogonal to the escape over the barrier.

Now, consider that metadynamics is performed with x as the chosen CV. We assume that bias deposition is very slow such that we have quasi-equilibrium conditions. Then,

metadynamics flattens the potential of mean force (PMF) along x up to a preset level ΔF . The PMF along x is defined as

$$e^{-\beta F(x)} = \int_{-\infty}^1 dy e^{-\beta V(x,y)} , \quad (6.5)$$

up to an additive constant, chosen such that the minimum is at a PMF value of zero. For small a and x , the PMF along x can be approximated as

$$F(x) \approx \frac{x^2}{2(1+a^2)} . \quad (6.6)$$

For a given metadynamics bias level ΔF , the bias acts on the range $|x| < \sqrt{2(1+a^2)\Delta F}$. The combined potential including the metadynamics bias is then

$$U(x, y | \Delta F) = \begin{cases} V(x, y) - F(x), & \text{for } |x| < \sqrt{2(1+a^2)\Delta F} , \\ V(x, y) - \Delta F, & \text{otherwise .} \end{cases} \quad (6.7)$$

The potential $U(x, y)$ has a minimum in the bound well for $|a| < 1/\sqrt{2\Delta F - 1}$. The minimum is located at $x = \sqrt{2(1+a^2)\Delta F}$ and $y = a\sqrt{2\Delta F/(1+a^2)}$. The lowest energy barrier to escape on the combined potential is located at $x = \sqrt{2(1+a^2)\Delta F}$ and $y = 1$. For $|a| < 1/\sqrt{2\Delta F - 1}$, the height of the barrier on the potential $U(x, y)$ including the metadynamics bias is

$$\Delta F^\ddagger = \frac{1 + a[a - \sqrt{8(1+a^2)\Delta F} + 2a\Delta F]}{2} . \quad (6.8)$$

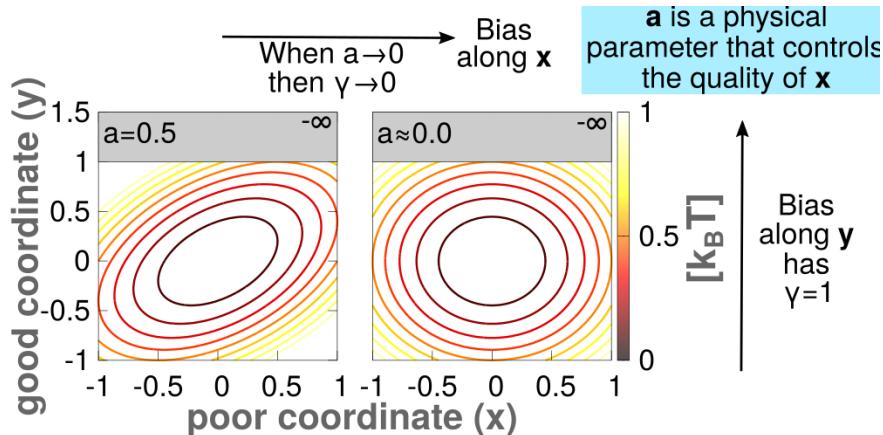


Figure 6.2: Cusp-like harmonic double-well potential. The potential is given in Eq. 6.4 where a is a physical parameter that controls the quality of coordinate x . Examples for $a = 0.5$ (left) and 1×10^{-5} (right) are shown. If the bias is along y , we show that $\gamma_y = 1$ for all a . However, if the bias is along x and $a \rightarrow 0$, we derive $\gamma_x \rightarrow 0$.

Now if metadynamics has reached a level of $\Delta F = \Delta F_0^\ddagger = 1/2$, i.e., the level of the barrier in the potential $V(x, y)$, the potential well is nominally filled. However, as the above calculation shows, for $|a| < 1/\sqrt{2\Delta F - 1}$ a barrier ΔF_0^\ddagger remains. For small a and $\Delta F = \Delta F_0^\ddagger = 1/2$, the height of this remaining barrier is

$$\Delta F^\ddagger \approx \frac{1}{2} - a . \quad (6.9)$$

If one now uses the Kramers' approximation for the rate of escape from the potential well without bias (Eq. 6.1), and with a bias potential filled up along x to a height of $\Delta F = \Delta F_0^\ddagger$, then the rate accelerates to

$$k_{\text{MetaD}} = k_{\text{pre}} e^{-\beta \Delta F^\ddagger} \approx k_{\text{pre}} e^{-\beta \Delta F_0^\ddagger(1-2a)} . \quad (6.10)$$

This means that we have not gotten the full boost from metadynamics. In this work, we correct for this reduced boost by the factor γ . For the problem here, this factor is obtained from

$$k_{\text{MetaD}} \approx k_{\text{pre}} e^{-\beta \Delta F_0^\ddagger(1-2a)} = k_{\text{pre}} e^{-\beta \Delta F_0^\ddagger(1-\gamma)} . \quad (6.11)$$

Therefore, for the escape from this double well and $a \approx 0$, the efficiency of the CV is thus

$$\gamma \approx 2a . \quad (6.12)$$

In other words, for $a \rightarrow 0$, when x is orthogonal to the escape flux, we have no acceleration from metadynamics, and from $0 < a \ll 1$, we have only a much smaller acceleration than what would be achieved by biasing along a well-chosen CV.

If instead y is chosen as the CV, metadynamics flattens the PMF along y . The PMF along y is given, up to a constant, by

$$F(y) = \frac{y^2}{2} . \quad (6.13)$$

The combined potential is

$$U(x, y | \Delta F) = \begin{cases} V(x, y) - F(y), & \text{for } |y| < \sqrt{2\Delta F} \\ V(x, y) - \Delta F, & \text{otherwise} . \end{cases} \quad (6.14)$$

For $\Delta F = \Delta F_0^\ddagger = 1/2$, the barrier to the exit at $x = a$ and $y = 1$ vanishes. Therefore, the choice of y as a CV results in a boost factor $\gamma = 1$ for all values of a .

In conclusion, if $a \approx 0$, x is a poor coordinate because the wells are not separated when projecting along this direction. For this case, if we bias along x , we demonstrate that $\gamma \approx 2a$ (and therefore $\gamma \rightarrow 0$). In contrast, when biasing along the good CV (y), $\gamma = 1$ is derived for any a . These results show that γ is not simply an *ad hoc* bias factor, but it measures a physical property that can be related to the efficiency of the CV.

This can be further exemplified for a 2D double-well potential, $\exp(-F(x, y)) = \exp(-(y^2 + x^2)/2) +$

$\exp(-((y-2)^2+x^2)/2)$, and biasing a collective variable $z = x \cos(\varphi) + y \sin(\varphi)$ that depends on an angle φ between x and y (Fig. 6.3a). Note that $\varphi = 0$ ($\pi/2$) is a poor (ideal) CV that biases along x (y , respectively). In Fig. 6.3a, we present examples of the effective potential surface (true potential plus V_{MB}) at different filling levels (solid surface in Fig. 6.3a) $V_{\text{MB}} = 0.05, 0.15, 0.3 \text{ k}_B T$ for $\varphi = 0, \pi/10, \pi/2$. Comparing these biasing directions at the same filling level, we find that the effective barriers are different, and that there is φ -dependent fraction of the bias (directly related to γ) which is effectively filling the well (Fig. 6.3b).

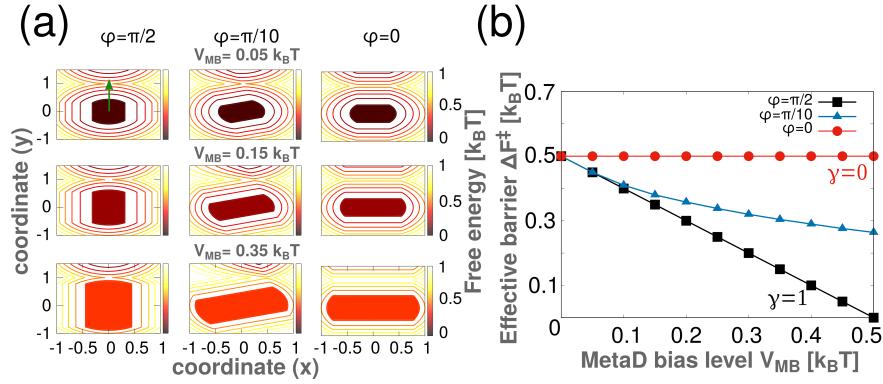


Figure 6.3: Illustration of the efficiency of the CVs on 2D double-well potential. (a) 2D double-well potential, $\exp(-F(x,y)) = \exp(-(y^2+x^2)/2) + \exp(-((y-2)^2+x^2)/2)$, with MetaD bias added over a collective variable $z = x \cos(\varphi) + y \sin(\varphi)$ that depends on an angle φ between x and y for $\varphi = 0, \pi/10, \pi/2$ at different MetaD filling levels $V_{\text{MB}} = 0.05, 0.15, 0.30 \text{ k}_B T$ (solid surface). The green arrow in the first panel indicates the transition under study. (b) Effective barrier $\Delta F^\ddagger = \Delta F_0^\ddagger - \gamma V_{\text{MB}}$ as a function of the filling level V_{MB} for $\varphi = 0, \pi/10, \pi/2$.

6.2.3 SURVIVAL PROBABILITY FROM METADYNAMICS SIMULATIONS

Using Kramers' Time-dependent Rate (KTR), from Eq. 6.3, we calculate the survival probability by adapting the methods used for the analysis of single-molecule force spectroscopy experiments¹²⁴,

$$\begin{aligned} S(t) &= \exp \left(- \int_0^t k(t') dt' \right) \\ &= \exp \left(-k_0 \int_0^t e^{\beta \gamma V_{\text{MB}}(t')} dt' \right). \end{aligned} \quad (6.15)$$

The survival probability depends on $V_{\text{MB}}(t)$, which we can fit numerically, or analytically (see Appendix A, section A), and use it to calculate $S(t)$ (Eq. 6.15). Note that the Cumulative Distribution Function (CDF) of the escape times is $1 - S(t)$.

6.2.4 MAXIMIZATION OF THE LIKELIHOOD FUNCTION

Let us assume there are $M + N$ independent biased simulations that start from the well bottom where M cross the barrier and N remain in the basin. The optimal transition rate k_0 and γ can be extracted by maximizing the likelihood function¹²⁸ (i.e. given our model, maximizing the probability to observe M events at times t_i and N non-events up to time t_j),

$$\mathcal{L} = \prod_{i \in \text{events}}^M -\frac{dS(t)}{dt} \Big|_{t=t_i} \prod_{j \in \text{non-events}}^N S(t_j) , \quad (6.16)$$

where i and j account for events and non-events, respectively, and t_j is the total simulation time for run j that did not transition. In practice, we first numerically fit $V_{\text{MB}}(t)$ using a spline representation with a time grid that spans the original time range.

From Eq. 6.16, we obtain an analytical expression of the rate in terms of γ and $V_{\text{MB}}(t)$ that maximizes it, as follows:

For a general survival model, the likelihood is given by the product of the probability of transition at time t_i , $-\frac{dS(t)}{dt} \Big|_{t=t_i}$, and the probability of not having transitioned at final time t_j , $S(t_j)$ ¹²⁸.

Combining the expressions of the likelihood from Eq. 6.16 and the survival function from Eq. 6.15, we obtain an expression for the log-likelihood

$$\begin{aligned} \ln \mathcal{L}(k_0, \gamma) &= -k_0 \sum_i^{M+N} \int_0^{t_i} e^{\beta \gamma V_{\text{MB}}(t')} dt' \\ &\quad + M \ln k_0 + \sum_{i \in \text{events}}^M \beta \gamma V_{\text{MB}}(t_i) , \end{aligned} \quad (6.17)$$

where M is the number of events and N the number of non-events. Our aim is to extract parameters k_0^* and γ^* that maximize this function. The derivative with respect to k_0 of the log-likelihood is

$$\frac{\partial}{\partial k_0} \ln \mathcal{L}(k_0, \gamma) = - \sum_i^{M+N} \int_0^{t_i} e^{\beta \gamma V_{\text{MB}}(t')} dt' + \frac{M}{k_0} . \quad (6.18)$$

To solve for the optimal intrinsic rate, we set the previous derivative to zero, and find it as a function of γ

$$k_0^*(\gamma) = \frac{M}{\sum_i^{M+N} \int_0^{t_i} e^{\beta \gamma V_{\text{MB}}(t')} dt'} . \quad (6.19)$$

Using this function, we search for the optimal parameters γ^* and $k_0^* = k_0^*(\gamma^*)$ through

numerical maximization of $\ln \mathcal{L}(k_0^*(\gamma), \gamma)$. The code in our github repository (see below) uses the Brent method for bounded optimization.

Using this expression and the numerical fit, we numerically search for the optimal γ . Because one only has to run a 1D numerical optimization over γ , we note that this is more stable and robust than a 2D numerical fit directly of $S(t)$ using both k_0 and γ . To validate the statistics of the simulation barrier-crossing times, bootstrap analysis and Kolmogorov-Smirnov test (KS-test) can be performed.

6.2.5 PRACTICAL CONSIDERATIONS

CUMULATIVE DISTRIBUTION FUNCTION AND KOLMOGOROV-SMIRNOV TEST:

The KS-test^{129,130} is a non-parametric test that enables comparing the similarity between one-dimensional distribution functions. This test compares the Empirical Cumulative Distribution Function (ECDF) to the Theoretical Cumulative Distribution Function (TCDF). The ECDF is given by a step function that jumps by $1/(N + M)$ at each t_i of the M events, which takes care of the non-event data N points. On the other hand, the TCDF is related to the survival function via $1 - S(t)$, where the survival function $S(t)$ is computed from Eq. 6.15, using a spline representation of $V_{\text{MB}}(t)$ and fitted parameters k_0^*, γ^* (as described above). One can perform the KS-test using the TCDF, comparing it to the ECDF (calculated using the simulation-jump times), and extracting a p -value that indicates if the KS-test passed or not. A p -value threshold of 0.05 is typically used.

BOOTSTRAP ANALYSIS:

To estimate the errors of the extracted parameters, we performed a bootstrapping analysis. The number of bootstrapping samples is 100 for the 2D double-well potential. We use the samples that pass the KS-test, and calculate the distribution for each parameter. From these we can estimate the standard deviation or the percentile error, for example, by taking the values at the 30, 50 and 70 percentiles of the distribution.

CODE AND DATA AVAILABILITY:

The codes to obtain the parameters from likelihood maximization, compute cumulative distribution function and perform the KS-tests are available on GitHub: <https://github.com/kpalaciorodr/KTR>.

6.2.6 SIMULATION SETUP FOR THE 2D DOUBLE-WELL POTENTIAL

To analyse the dependence of the results on the efficiency of the CVs, we used a 2D double-well potential $U(x, y) = -k_B T \ln[e^{(-20 \cdot (x-0.2)^2 - 100 \cdot (y-0.2)^2)} + e^{(-20 \cdot (x-0.8)^2 - 100 \cdot (y-0.8)^2)}]$ (similar to that in ref. 131). For this potential, the projection of the free energy surface along the x coordinate leads to an underestimation of the barrier between the wells, while the projection along the y coordinate represents faithfully the underlying 2D barrier of around $8 k_B T$. The

MC step was performed randomly from a uniform distribution for each **CV** in the range [-0.005, 0.005] around the current point inside a grid between -0.4 and 1.4 for x and y . These trial moves were accepted according to the Metropolis criterion. A bias height of 0.04 and a bias width of 0.04 were used. To asses the effect of the **CV**, the simulations were performed adding bias only along a single coordinate (i.e., either the x or y coordinate, in independent simulations). The bias-deposition time was varied and 100 simulations were launched for each **CV**. To obtain the reference rate coefficient k_0^{unbiased} for the double-well potential, we performed 1000 unbiased runs with 3×10^7 MC steps. We calculated $k_0^{\text{unbias}} = M / (\sum_i t_i + \sum_j t_j)$ where M is the total number of events, t_i is the escape time observed in run i , and t_j is the total simulation time for run j that did not have a transition (for the unbiased case around 30% of non-events).

6.2.7 RATE FROM INFREQUENT METADYNAMICS

iMetaD³², is a simple and commonly applied approach that starts by repeating a series of identical MetaD simulations where Gaussians are deposited relatively slowly, in order to reduce the probability of biasing the transition state region. Under these conditions, ideas from Grubmüller¹¹⁶ and Voter¹¹⁷ based on TST are exploited to recover the unbiased dynamics of the system by rescaling the simulation time with an acceleration factor correcting for the bias. In practice, we rescale the times from MetaD by¹¹¹

$$t_{\text{res}} = \sum_{i=1}^{N_t} dt e^{\beta V_i^B}, \quad (6.20)$$

where dt is the stride to print the MetaD bias, which should be no greater than the bias deposition time t_d and no less than the MD time step, i is the step number, N_t is the total number of steps until an event occurred or the simulation stopped, V_i^B is the instantaneous MetaD biasing potential at step i and $\beta = 1/(k_B T)$ with k_B the Boltzmann's constant and T the absolute temperature.

In iMetaD, the reliability of the rescaled escape (residence) times is quantitatively assessed by comparing their **ECDF** to the **TCDF** of a homogeneous Poisson process using a two-sample KS test¹¹⁸. A p -value threshold of 0.05 is typically used¹¹⁸. The **ECDF** is fitted to the **TCDF** of a homogeneous Poisson process, $P(t) = 1 - \exp(-t/\tau)$. Following ref. 118, a **TCDF** is built from a large number of sample times (e.g., 10^6) randomly generated according to $P(t)$. Then, a two sample KS test is applied to assess the similarity between **ECDF** and **TCDF**, with the null hypothesis that the two sampled distributions share the same underlying distribution. If the p -value from the **KS-test** is higher than 0.05, ignoring the fact that τ has been fitted, then the null hypothesis is accepted and a Poissonian process is considered appropriate to describe the statistics of the process. The rate coefficient is then calculated as the inverse of the average time from the fit of the **ECDF**, $k_{\text{iMetaD}} = 1/\tau$.

We adapted a Matlab code provided in ref. 118 to python3.6 for the calculation of τ using iMetaD and the **KS-test** (for both methods, iMetaD and KTR). The histograms of the CDFs

are built using the number of bins equal the total number of simulations (this was tested empirically to obtain the best fits of the CDF for the 2D system). We use logarithmic scale to create the histograms for iMetaD and a linear scale for KTR .

6.3 RESULTS AND DISCUSSION

6.3.1 BENCHMARK SYSTEM: 2D DOUBLE-WELL POTENTIAL

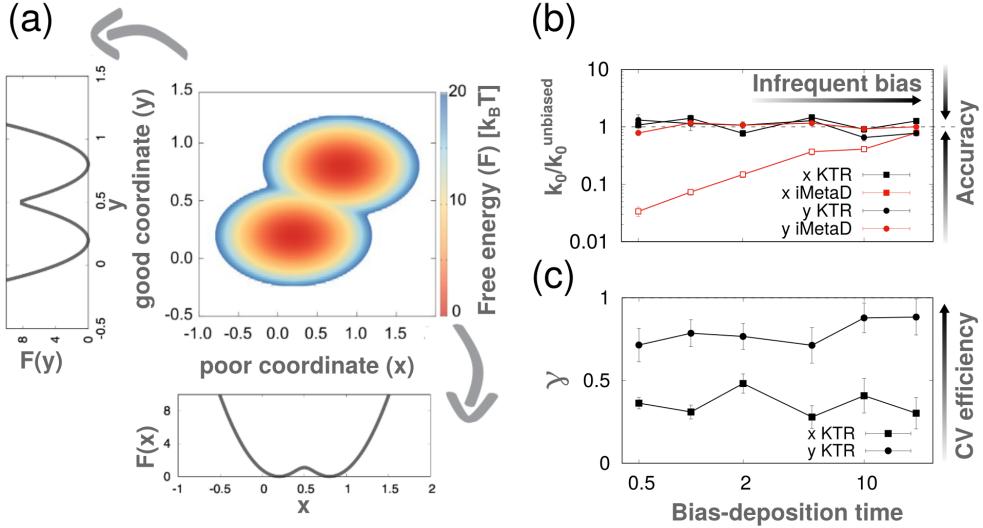


Figure 6.4: MetaD Monte Carlo simulations on a 2D double-well potential. (a) Two dimensional free-energy surface $F(x, y)$, where the bias is deposited along a poor (x) or good (y) CV with $\exp(-\beta F(x, y)) = \exp((-20 \cdot (x - 0.2)^2 - 100 \cdot (y - 0.2)^2)) + \exp((-20 \cdot (x - 0.8)^2 - 100 \cdot (y - 0.8)^2))$ and $\beta = 1/k_B T = 1$. White regions have values $> 20 k_B T$. Free energy profiles $F(x)$ and $F(y)$ are also shown. (b) Maximum likelihood extracted intrinsic rate (k_0) normalised by the true rate (calculated from unbiased simulations) as a function of the bias-deposition time ($t_d \times 10^3$ steps) for the KTR (black) and iMetaD (red) methods for biases along x (squares) and y (circles). Empty squares indicate cases where the KS-test failed for more than 50% of the bootstrap trials. (c) Maximum likelihood extracted efficiency of the CV (γ) as a function of the bias-deposition time for biases along x (squares) and y (circles) using the KTR method. Error bars show the standard deviation using the distribution from bootstrap analysis (see section 6.2.5).

To study the effect of the CVs, we tested the theory on a 2D double-well potential (Fig. 6.4a). The free energy profile projected along the “ x ” coordinate shows that it is a poor coordinate because it underestimates the real barrier (see profile at the bottom in Fig. 6.4a.). However, when projecting the free energy along the “ y ” coordinate the true barrier is well determined (see profile at the left in Fig. 6.4a.). For the transition rate estimation, an important parameter in the MetaD setup is the bias-deposition time. The bias-deposition time corresponds to the inverse of the frequency to update the bias potential, i.e., to add a new Gaussian function. In iMetaD, the bias-deposition time is large (compared to a standard MetaD setup), to avoid adding bias over the TS region. In our KTR formalism we also expect to recover better results at large deposition times as it will allow the system to equilibrate

better and will avoid overbiasing (we will discuss this topic in depth in the next chapter). We ran MetaD Monte Carlo simulations over y or x CV (see section 6.2.6). We estimated the rate (k_0) and CV efficiency (γ) for setups with different bias-deposition times. We started 100 simulations from the lower well and counted a transition when the system reached the top well ($y \geq 0.8$). Fig. 6.1b shows examples of the average $V_{\text{MB}}(t)$ over the runs biased along y .

6.3.2 ESCAPE RATES FROM KRAMERS TIME-DEPENDENT RATE VS INFREQUENT METADYNAMICS

In Fig. 6.4b, we present the extracted rates from numerical integration of Eq. 6.15 with maximum likelihood (Eq. 6.16) and compare them to those extracted with iMetaD by rescaling the times using the bias acceleration (Eq. 6.20). We find that the KTR method estimates accurate unbiased rates, even for the challenging cases of fast deposition times and poor reaction coordinates. We note that for some of these cases (empty squares in Fig. 6.4b) the iMetaD KS-test fails, indicating that the iMetaD estimate should not be used.

6.3.3 CUMULATIVE DISTRIBUTION FUNCTIONS

In Fig. 6.5-left, we show the CDFs for the simulation-jump times and their fits using $1 - S(t)$ (from Eq. 6.15) for both coordinates and the different bias-deposition times. These results show that x is a poor coordinate because it does not accelerate as much the barrier-crossing events. However, in all cases there is good agreement between the theoretical and empirical CDFs. In Fig. 6.5-right, we show the CDFs for the rescaled times using iMetaD (see section 6.2.7 for details) together with their theoretical fits. We find that along the good CV (y) this method works well. However, along the poor CV (x) with frequent bias-deposition times the KS-test fails and therefore iMetaD fails¹¹⁸. By adding bias along the poor CV, we are accumulating bias along the wrong direction, so the system is not aided to transition and spends more time than necessary in the reactant state, resulting in an overestimation of residence times and a non-Poissonian distribution. Additionally, at low deposition times we are adding bias over the TS (see Appendix A, Fig. A1), so the underlying assumptions of the iMetaD do not hold.

6.3.4 COLLECTIVE VARIABLE EFFICIENCY

Importantly, the KTR theory enables extracting information about the efficiency of the biased CVs. Fig. 6.4c shows the extracted γ as a function of the bias-deposition time for both coordinates. As expected, biasing along the poor (x) CV leads to a lower γ , while using a good (y) biasing direction yields γ closer to 1. We note that for y , γ is not exactly equal to 1 because y despite being a better CV than x is still not perfect. Nonetheless, it describes accurately the transition between the two basins and the barrier region.

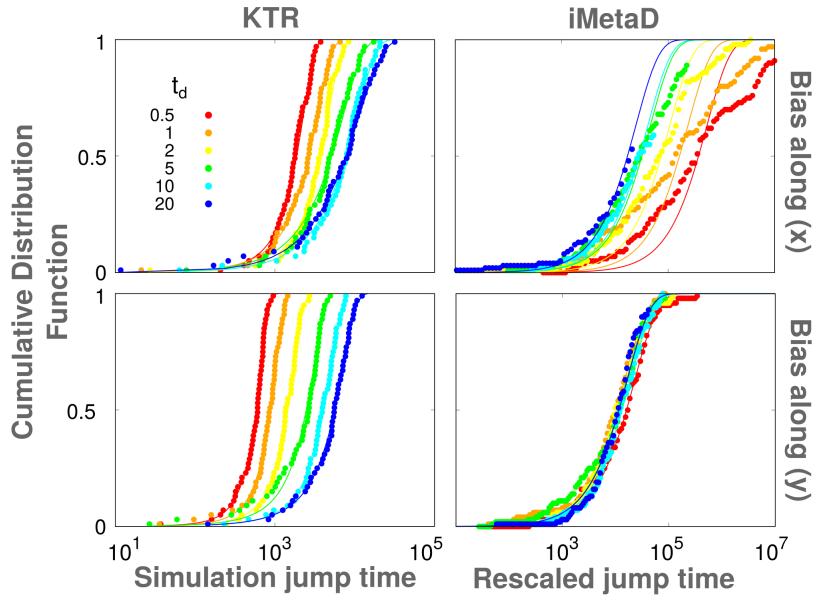


Figure 6.5: CDFs of simulation and rescaled jump times and fits to theoretical distributions for the 2D double-well simulations. Left: Cumulative distribution function (CDFs) of the simulation barrier-crossing (jump) times from KTR . Runs with bias along x (top) or y (bottom) for different bias-deposition times (t_d). Fits to the respective theoretical cumulative distribution function are shown as solid lines. Right: CDFs of the rescaled jump times from iMetaD. Runs with bias along x (top) or y (bottom) for different bias-deposition times (t_d). Fits to the respective Poisson distribution are shown as solid lines.¹¹⁸ Red to blue points go from frequent to infrequent t_d .

6.4 CONCLUSIONS AND OUTLOOK

Inspired by the methods from the force spectroscopy community^{124–127} that calculate barrier-crossing rates induced by forces acting on single molecules, in this chapter, we used Kramers’ theory to calculate the time-dependent transition rates and survival probabilities from time-dependent biased simulations. These results indicate that the KTR method is able to extract accurate unbiased rates and assess the efficiency of a CV along non-ideal biasing directions.

We have used examples from MetaD simulations, however, the KTR method is general for any time-dependent biased simulation along some CVs, such as, adaptive biasing force¹³² or adaptively biased MD¹³³, among others. Methods restricting the bias filling level^{134,135} could also be used within the KTR framework, which would have the advantage of simplifying bookkeeping of the bias potential and survival probability. We highlight our method enables estimating the unbiased intrinsic rate even if the biasing direction is poor. Moreover, it provides a novel measure for the effectiveness of the added bias to accelerate the transition (related to the efficiency of the CVs). This overcomes severe limitations encountered with previous approaches where the bias had to be deposited very infrequently over ideal CVs.

In the next chapter we will investigate in detail the limitations of the method and the regime of bias-deposition times where the method is valid through its application to all-atoms MD-MetaD simulations of more complex systems.

ACKNOWLEDGMENTS

The method presented in this chapter was developed in co-authorship with Hadrien Vroylandt, Lukas S. Stelzl, Fabio Pietrucci, Gerhard Hummer, and Pilar Cossio. Additionally, we gratefully acknowledge Attila Szabo, Erik Thiede and Marylou Gabrié for useful discussions, as well as David Silva for code optimization advice.

7

Kinetic rates from metadynamics: Applications

7.1 INTRODUCTION

After benchmarking our KTR method in a double-well potential, we now apply the KTR theory to study three challenging systems using all-atom MD simulations. Initially, we explore the limits of our method and provide convergence criteria to assess its validity through its application to the dissociation of the fullerene C₂₄₀ dimer in water solution. Then, we apply the method to a more complex process: the unbinding of the ligand CS3 from Cyclin-Dependent Kinase 2 (CDK2) protein. Overall, these applications show the robustness of the method and the points that could be improved in the future.

7.2 INTERACTION OF FULLERENE C₂₄₀ DIMER IN WATER SOLUTION

7.2.1 DESCRIPTION OF THE SYSTEM

Fullerenes are allotropes of carbon that are characterized by having a closed spherical mesh of carbon atoms. They have attracted the attention of the scientific community since their discovery in the 1980s.¹³⁶ These carbon structures have unique properties that make them interesting for diverse applications. Their use has been proposed in biomedical areas due to their ability to trap metals and small molecules, as well as their ability to permeate biological membranes and their antioxidant properties.¹³⁷ Also, in the area of new materials, fullerenes offer great opportunities due to the possibility of surface functionalization and encapsulation of various molecules for applications in catalysis, photovoltaics, photodetection.¹³⁸

We studied the dissociation of a fullerene C₂₄₀ dimer in water solution, see Fig. 7.1a. In water, the associated state of the dimer is more favorable since fullerene molecules are hydrophobic. Despite the simplicity of the process: the symmetry of the fullerene molecules

makes different binding conformations of the dimer indistinguishable, it still retains the complexity of the interaction with water molecules and a large free energy barrier separating the associated and dissociated states (see Fig. 7.1b). In the following, we describe the simulation methods used to study fullerene dimer dissociation, as well as the procedure used to estimate the dissociation rate using the KTR method.

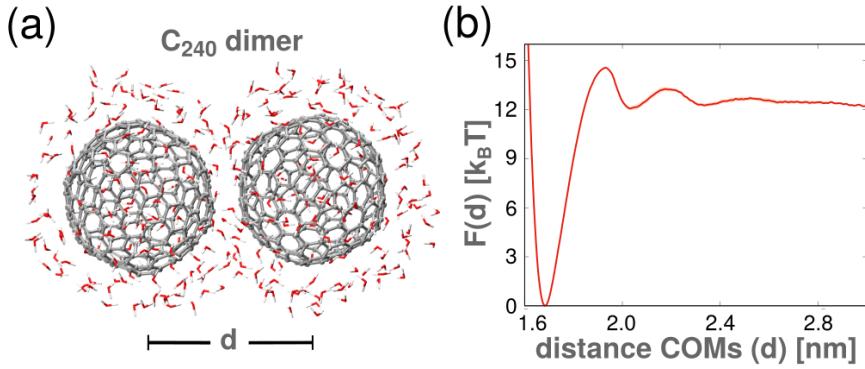


Figure 7.1: Fullerene dimer system's representation and free energy profile. (a) Representative snapshot of the associated C_{240} fullerene dimer complex. We use as CV the distance between the center of mass (d) of each fullerene molecule. (b) Free-energy profile as a function of the distance between the center of mass of the C_{240} molecules.

7.2.2 SIMULATION METHODS

To simulate the interaction of a C_{240} fullerene dimer in a water solution, the dimer is solvated with 5375 water molecules in a simulation box of $5.22 \times 5.22 \times 5.22$ nm with periodic boundary conditions. MD simulations are carried out using GROMACS v2019.4^{139,140} patched with PLUMED 2.5.3.¹⁴¹ We adopted the SPC water model and the OPLS-AA force-field for carbon. Geometry minimization exploited the steepest descent algorithm, stopped when the maximum force was ≤ 50 kJ/mol · nm. We used the leapfrog algorithm to propagate the equations of motion and the nonbonding interactions were calculated using a PME scheme with a 1.2 nm cutoff for the part in real space. We performed a 100 ps equilibration in an NVT ensemble with a stochastic velocity rescaling scheme followed by a 100 ps equilibration in an NPT ensemble using the Parrinello-Rahman barostat with a time step of 1 fs. We generated MD production trajectories without restraints, with a time step of 1 fs in the NPT ensemble at 298 K and 1 atm.

We estimated the free energy profile in Fig. 7.1b using WT-MetaD⁹⁴, adding bias over the distance between the center of mass (d) of each fullerene with an initial Gaussian height of 1.0 kJ/mol and width of 0.01 nm. The bias factor was set to 8, and Gaussians were deposited every 1 ps. 5 replicas of 500 ns each were simulated. The average and standard deviation of the free energy profiles was estimated using these replicas.

To compute the dissociation rate using the KTR method, we performed WT-MetaD simulations adding bias over d with an initial Gaussian height of 1.0 kJ/mol and width of 0.01

nm. The bias factor was set to 8. We used bias deposition times t_d between 0.5 to 30 ps. For each t_d we started 50 trajectories from the associated state ($d \approx 1.68$ nm) with different initial velocities and stopped them when they reached the barrier ($d = 1.93$ nm).

To estimate the errors of the extracted parameters, we performed a bootstrapping analysis. The number of bootstrapping samples is 50 for the fullerene. We use the samples that pass the KS-test, and calculate the distribution for each parameter. From these we can estimate the standard deviation or the percentile error, for example, by taking the values at the 30, 50 and 70 percentiles of the distribution.

7.2.3 RESULTS AND DISCUSSION

We assessed the dissociation rate of a C_{240} fullerene dimer in water solution (Fig. 7.1a). The free-energy profile as a function of the distance between the centers of mass (d) estimated using WT-MetaD (Fig. 7.1b) shows a large activation barrier of around $14.5 k_B T$. The free energy landscape is characterized by a deep single-well free energy minimum corresponding to the associated state and an almost flat dissociated state with some bumps corresponding to solvation layers. The shape of the free energy landscape makes this fullerene dissociation process an ideal target to test our method to compute escape rates. Since we know the reference free energy barrier, we can explore extreme biasing conditions, e.g. fast deposition times and study their effect in the accuracy of the estimated rate.

To obtain k_0 and γ , we performed 50 WT-MetaD simulations starting from the bound state for a wide range of t_d . Fig. 7.2a shows $V_{MB}(t/t_d)$ as a function of the time normalized by the bias-deposition time t/t_d (i.e., the number of Gaussians added in the MetaD simulation). For $t_d \leq 5$ ps the cumulative bias is much larger than the barrier and therefore we observe overbiasing. Interestingly, for this reduced time scale, all curves collapse, showing that the shape of the filling potential should be the same regardless of these parameter.

In Fig. 7.2c and Fig. 7.2d, we present the extracted k_0 and γ as a function of t_d , respectively. We find that for the simulations where the accumulated bias is much larger than the barrier, i.e. the shaded region, the rate is overestimated and γ is underestimated. The KTR theory fails for this overbiasing regime because the exponent in Eq. 6.3 should be strictly negative, and to compensate for the exceeding bias γ is artificially lowered. The plateau in Figs. 7.2c and 7.2d shows that, in the converged regime, k_0 and γ should not dependent on the bias-deposition time (nor on other biasing conditions). Next, we plot $\gamma V_{MB}(t/t_d)$ as a function of t/t_d (Fig. 7.2b). In this case, only curves where γ is not artificially lowered collapse, as happens (within error) for the 2D double-well system along both coordinates (see Appendix A, Fig. A2). Note that, in Fig. 7.2d, if the converged $\gamma \sim 0.83$ were used for the $t_d = 0.5$ ps case, then the effective bias would be larger than the true barrier at long times. These results led us to establish two convergence criteria: *i*) $V_{MB}(t/t_d)$ and $\gamma V_{MB}(t/t_d)$ curves should collapse, and *ii*) the extracted k_0 and γ should be within error for different setups.

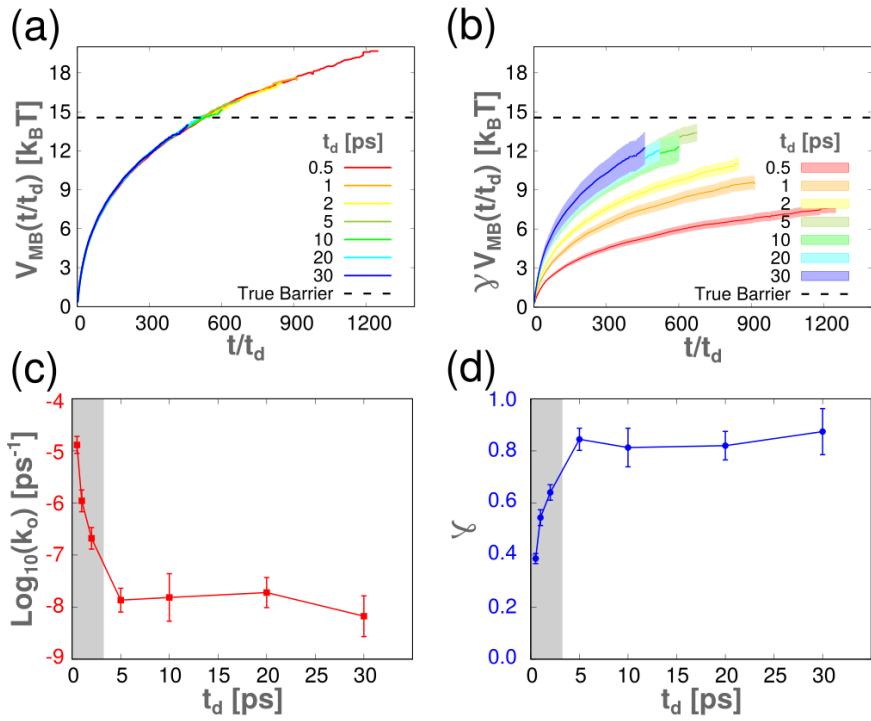


Figure 7.2: Fullerene dimer dissociation biased simulations. (a) V_{MB} as a function of the time normalized by the bias deposition time (t/t_d) for the different simulation setups, resulting in a collapse of the curves. (b) V_{MB} scaled by γ as a function of the time normalized by the bias deposition time (t/t_d) for the different simulation setups. (c) Maximum likelihood extracted intrinsic rate (k_0). Error bars show the 30 and 70 percentiles from bootstrap analysis that pass the KS-test. (d) Efficiency of the CV (γ) as a function of the bias-deposition time. Error bars show the 30 and 70 percentiles from bootstrap analysis that pass the KS-test.

7.3 PROTEIN-LIGAND UNBINDING: CDK2 PROTEIN-CS3 LIGAND

7.3.1 DESCRIPTION OF THE SYSTEM

Cyclin-Dependent Kinases (CDKs) are a family of protein kinases first discovered for their role in regulating the eukaryotic cell cycle, Fig. 7.3a. They are present in all known eukaryotes.¹⁴² To date, 13 CDK family members (CDK1–CDK13) have been identified.¹⁴³ In this section, the biological protein target CDK2 is used as test system. CDK2 is one of the kinases with more structural and pharmacological information available.¹⁴⁴ This biological target is involved in the cycle, where, when bound to cyclin proteins (cyclin E and cyclin A, respectively) facilitates the transition between G1/S phases and the progression of the S phase of the cell cycle^{145,146}, see Fig. 7.3a, i.e., it is necessary to move from DNA synthesis to replication. CDK2 is also involved in regulating transcription, mRNA processing, and the differentiation of nerve cells. Due to its important role in the cell cycle, CDK2 has been considered as a target for anticancer drugs.^{147–149}

CDKs are kinases, namely, enzymes that phosphorylate (attach phosphate groups to)

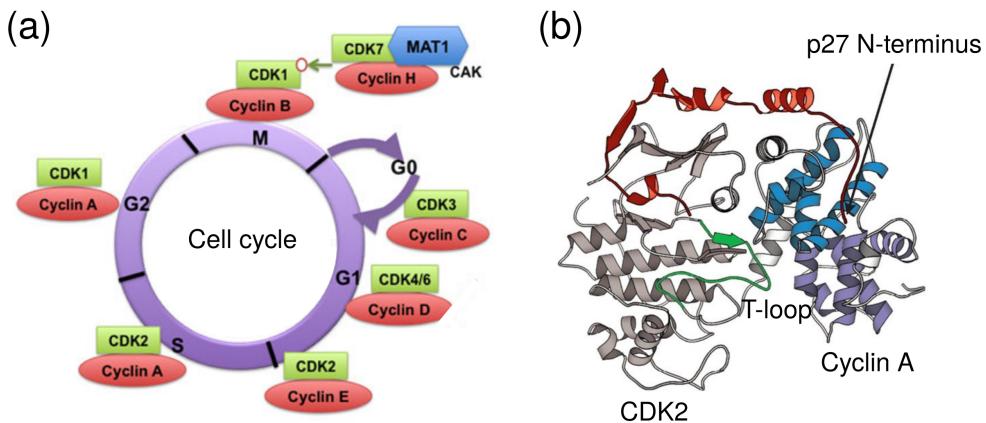


Figure 7.3: (a) Representation of cyclin-dependent kinases and their cyclin regulatory subunits. CDK-cyclin complexes have direct functions in regulating the cell cycle. The biological function of CDK2 in the step from stage G₁ to S in the cell cycle. Adapted from ref. 150 (b) Representation of the structure for the complex CDK2/Cyclin A. Adapted from ref. 142.

specific target proteins. The attached phosphate group acts like a switch, making the target protein more or less active. The regulation of its activity is given by the location of the T-loop, see Fig. 7.3b, because this loop opens and closes the entrance to the active site. Furthermore it is well known that monomeric CDK2 lacks regulatory activity, for this reason it needs to be activated by its positive regulators, cyclins A and E, or be phosphorylated on the catalytic segment of the protein.^{143,151} There are various Cyclin-dependent kinase inhibitors that interact by hydrophobic interactions and by making hydrogen bonds with the kinase, especially with the backbone of Glu 81 and Leu 83 in the structure of the apoenzyme.¹⁵² Other residues where hydrogen bonds can be formed are Asp 86 and His 84.¹⁴⁸ Lys 33 and Asp 145 residues display a bimodal distribution, between two predominant side-chain conformations so they can interact with different ligands.¹⁵³ Recently, the CDK2-ligand unbinding of ligands with Protein Data Bank (PDB) IDs 18K and 62K has been studied with machine learning transition state analysis by simultaneously obtaining free energy profiles and rates.¹⁵⁴ Here, we use an alternative approach to estimate the unbinding rate of the CDK2-CS3 complex, PDB structure 4EK5 (see Fig. 7.4a). This system allows us to evaluate the performance of the KTR method in studying an unbinding event. The CS3 is a 3-aminopyrazole inhibitor with PDB ID: 03K (see Fig. 7.4b). In the binding pose, this ligand interacts with the residues Glu 81, Leu 83 and Gln 85, see Fig. 7.4c. In the following, we describe the simulation methods used to study the unbinding of the CS3 ligand to the CDK2 protein, as well as the results of the application of the KTR method on this process.

7.3.2 SIMULATION METHODS

WT-MetaD simulations were carried out using the GROMACS v2019.4 program patched with PLUMED 2.5.3. The complex was solvated with a cubic water box, centered at the

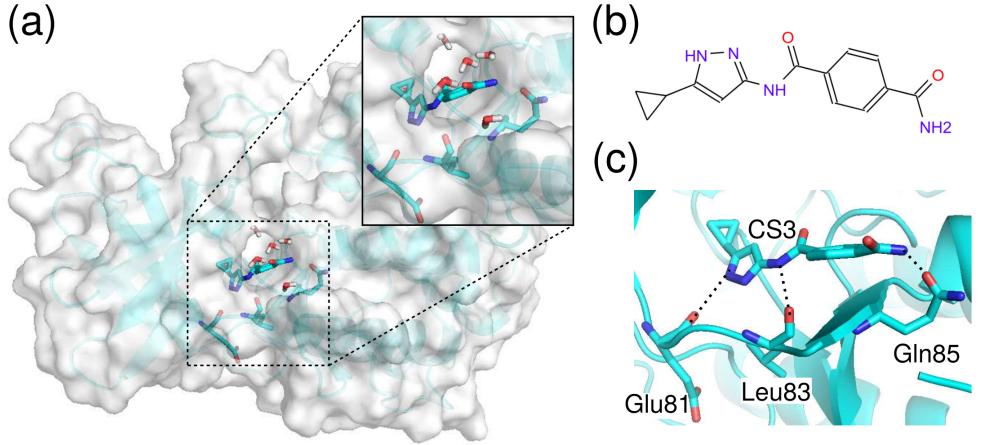


Figure 7.4: (a) Global view of the protein and of the binding pocket. We highlight the residues and water molecules that interact with the ligand inside the binding pocket. (b) Structure of the ligand 03K. (c) Main interaction between the ligand (03K) and the target (CDK2). The residues involved on this interaction are labeled.

geometric center of the complex with at least 2.0 nm between any two periodic images. The AMBER99SB-ILDN¹⁵⁵ force field was used to model the system with the TIP3P water model.¹⁵⁶ The ligand was parameterized using antechamber¹⁵⁷ with GAFF¹⁵⁸. The parameters found were converted into GROMACS format using ACPYPE.¹⁵⁹ A minimization was done with the steepest descent algorithm and stopped when the maximum force was ≤ 1000 kJ/mol·nm. Periodic boundary conditions were considered. We used the leapfrog algorithm to propagate the equations of motion and the nonbonding interactions were calculated using a PME scheme with a 1.0 nm cutoff for the part in real space. We performed a 100 ps equilibration in an NVT ensemble using the velocity rescaling thermostat⁶⁹ followed by a 100 ps equilibration in an NPT ensemble using Parrinello-Rahman barostat⁷¹ with a time step of 2 fs. The MD production was performed without restraints, with a time step of 2 fs in an NPT ensemble at 300.15 K and 1 atm.

We chose two CVs for the WT-MetaD simulations. The first CV was the solvation state of the ligand (w), calculated as the coordination number between two groups

$$w = \sum_{i \in A} \sum_{j \in B} w_{ij}, \quad (7.1)$$

with

$$w_{ij} = \frac{1 - \left(\frac{r_{ij} - d_0}{r_0} \right)^n}{1 - \left(\frac{r_{ij} - d_0}{r_0} \right)^m}, \quad (7.2)$$

where $d_0 = 0$, $r_0 = 0.3$, $n = 6$ and $m = 10$. In the sum of Eq. 7.1, group A is the center of mass (COM) of the ligand and group B are the oxygen atoms of all water molecules at a distance shorter than 5 Å from the pocket. The second CV was the distance between binding

pocket and ligand (d). We define d as the distance between the COM of the heavy atoms in the ligand and the COM of the α -carbons in the binding pocket, i.e., at a distance of 5 Å from the ligand in the binding pose.

WT-MetaD is performed with an initial Gaussian height of 1.5 kJ/mol. The width (σ) of the Gaussians was $\sigma_w = 0.13$ and $\sigma_d = 0.02$ nm, for the w and d CVs, respectively. We used a bias factor of 15. We performed independent simulations where Gaussians were deposited every 1, 10, 100 and 500 ps with total simulation time of 10, 40, 300 and up to 1500 ns, respectively. 50 simulations per Gaussian deposition time were performed.

To monitor the escape of the ligand from the binding pocket, we followed (without biasing it) the evolution of an additional CV, the crystallographic contacts (c). With c we can easily distinguish between the bound and unbound state of the ligand. Analogous to w , the number of crystallographic contacts is defined as $c = \sum_{i \in A} \sum_{j \in B} c_{ij}$. We use Eq. 7.2 replacing w_{ij} by c_{ij} with the same parameters and with atoms groups responsible for the main interactions between the ligand and the binding pocket, i.e., group A are the nitrogen atoms of the ligand that form hydrogen bonds – in the binding pose – with the atoms in the group B: O-Glu81, O-Leu83 and O-Gln85. These interactions are shown in Fig. 7.4c.

7.3.3 RESULTS AND DISCUSSION

The unbinding of a ligand from a protein-ligand complex is an important field in drug discovery. In particular, the calculation of the unbinding rates and residence times have taken especial relevance in recent years^{160,161}, due to their better correlation with the drug efficacy of a molecule in *in vivo* assays compared to affinity energies.¹⁰³ We studied the unbinding of the ligand CS3 to the CDK2 protein using WT-MetaD. A global view of the protein-ligand complex and its interactions is shown in Fig. 7.4. We ran four sets of WT-MetaD simulations starting from the bound conformation with 50 replicas each of 10, 40, 300 and up to 1500 ns and bias-deposition time of 1, 10, 100 and 500 ps, respectively. We biased two CVs that caused the ligand to unbind during the simulations in most cases. CV1 (d) is the distance between the center of mass of the ligand and the pocket and CV2 (w) tracks the solvation state of the ligand (see Fig. 7.5 and section 7.3.2).

When monitoring the solvation state of the ligand, Fig. 7.5b, we found that the corresponding CV has a lower value when the ligand is inside the binding pocket than when the ligand is outside the pocket, as expected. However, the fluctuations of the CV complicate the identification of the states. A similar situation is found when monitoring the distance between the ligand and the pocket (Fig. 7.5a). When the distance begins to increase, there is no simple way to define when the ligand breaks its main interaction with the pocket. Therefore, as an alternative to clearly identify the unbound state, we monitor an additional CV: the number of crystallographic contacts, c , (i.e., the coordination number between the atoms that made contact in the crystal structure) without adding bias to it (Fig. 7.5c). CV c approaches zero when the interactions between the ligand and the pocket are broken and the ligand reaches the unbound state.

We note that the unbinding of the ligand from CDK2 involves several metastable states

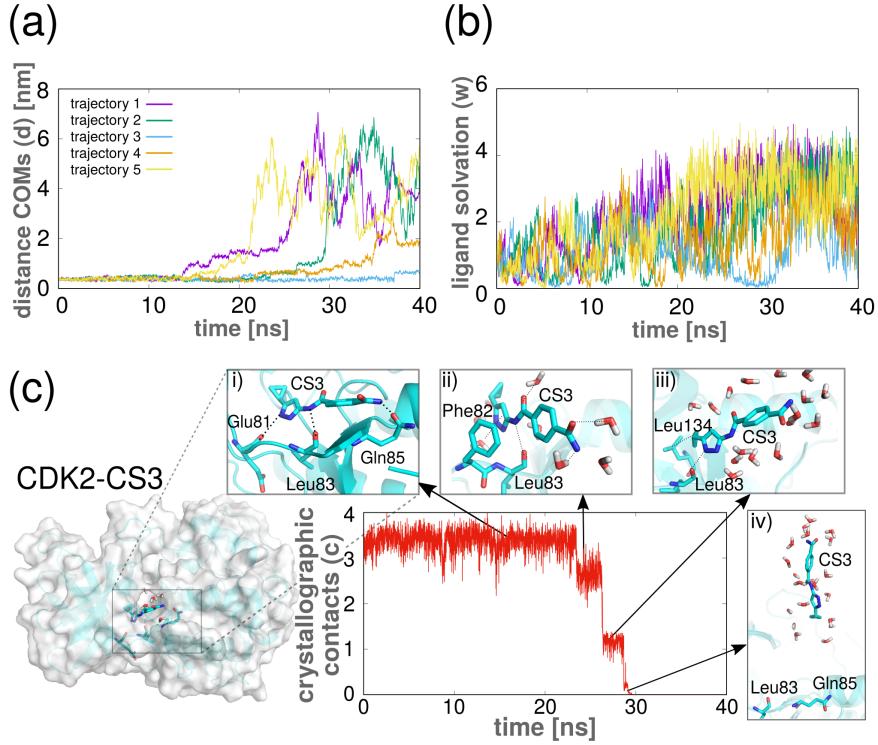


Figure 7.5: CDK2-ligand CVs as a function of time. (a) and (b) Five example trajectories of the biasing CVs for $t_d = 10$ ps. Each color represents a different trajectory: (a) the distance between the center of mass of the ligand and the center of mass of the binding pocket d and (b) the solvation state of the ligand w are shown as a function of time. (c) Global view of the protein, binding pocket and representative snapshots of metastable states. We highlight the residues that interact with the ligand inside the binding pocket. Example of the CV measuring the crystallographic contacts between the ligand and the binding pocket (c) as a function of time. There are several metastable states explored before the ligand reaches the unbound state where it is fully solvated with no interactions inside the binding site. For our analysis, we defined the unbound state when the crystallographic contacts c is < 0.01 .

before the unbound state is reached. A representation of the different states found according to the value of the collective variable is shown in Fig. 7.5c. In the crystallographic pose, the ligand interacts primarily through hydrogen bonds with the pocket residues Glu 81, Leu 83 and Gln 85. Then, when the ligand begins to leave the active site, it adopts another conformation interacting with the binding pocket residues through a stacking interaction with Phe 82 and H-bond with Leu 83. Before completely leaving the binding site, a hydrophobic contact with Leu 134 and an H-bond with Leu 83 are formed. At the same time, the solvation of the ligand increases. Finally, when the crystallographic contacts go to zero, the ligand is in a fully solvated, unbound state, or diffusing elsewhere on the protein surface far from the active site. Therefore, we defined a transition on the final dissociation event when CV c is less than 0.01. We also found some simulations where the system spends some time in some on and off-pathway metastable state before jumping back to the bound state, followed by the dissociation transition (Fig. 7.6a). This situation violates the assumption of an

ever-increasing maximum bias until escape (see Fig. 7.6b), therefore, we discarded these simulations from further analysis.

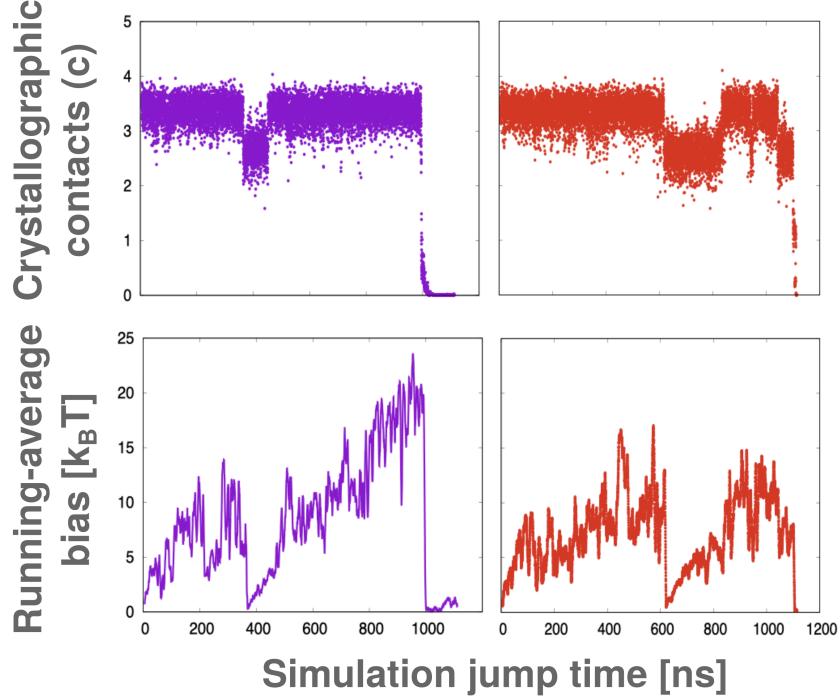


Figure 7.6: Multi-state dynamics in the CDK2 simulations. Example of two trajectories where the system spends some time in an on- or off-pathway metastable state before jumping back to the bound state from the $t_d = 500$ ps setup. (top) Number of crystallographic contacts $CV(c)$ as a function of the time. The systems jumps out and then back to the bound state. (bottom) Running-average of the bias as a function of time. The window average was 50 time-strides. When the system jumps out of the bound state there is a drastic fall in the bias. In practice, we defined a transition back to a metastable state when the running-average bias went below $5 k_B T$, the V_{MB} did not change for X time, and after that the system jumped back to the bound state. $X=0.2, 2, 30$, and 90 ns for the $t_d=1, 10, 100$ and 500 ps setups, respectively. Because the well is partially filled when a transition back to a metastable state occurs, and the KTR theory does not take this into account, trajectories that presented these kind of transitions were discarded from the CDK2 analysis ($\sim 50\%$ for all setups).

We monitored V_{MB} as a function of t/t_d (Fig. 7.7a), and noted that, for the 1 and 10 ps setups, the curves do not collapse as for the systems studied above. These results show that for these setups, the system does not have enough time to locally-equilibrate on the biased free-energy landscape; instead, it is surfing on the biasing potential, conveying that they are in the overbiasing regime and that the rates will be overestimated. Therefore, we discard these simulations from further analysis.

In Fig. 7.7b, we show the empirical CDFs using the simulation barrier-crossing times with fits of the KTR formalism for the 100 and 500 ps setups that are in the converged regime. The estimated γ and k_0 and their errors are also shown. Note that the errors are extracted from bootstrap analysis of the passed KS-trials, using the same procedure as for

the fullerene dimer in section 7.2.2. Importantly, both setups have extracted parameters within error: $\log(k_0[s^{-1}]) \approx \log(8.4) \pm \log(5.1)$ and $\log(9.5) \pm \log(5.6)$; and $\gamma = 0.65 \pm 0.09$ and 0.64 ± 0.11 for the $t_d = 100$ and 500 ps setups, respectively.

In comparison to the experimental rate $\log(k_{\text{exp}}[s^{-1}]) \approx \log(0.26)^{162}$ (see details in Appendix A, section A), the extracted median rate is around one order of magnitude larger than the experiment. This is probably because we are not taking into account the multi-well dynamics and force-field inaccuracy.² Nonetheless, this is a promising result since, we are using simulations of length many orders magnitude shorter than the estimate of seconds. Moreover, the KTR method proves valuable when predicting the underlying statistics of barrier-crossing times in comparison to attempted CDF-fits using the iMetaD rescaled times, where more than 74% of the bootstrap samples failed the KS-test, therefore iMetaD results are not valid in this case¹¹⁸ (see Appendix A, Fig. A3).

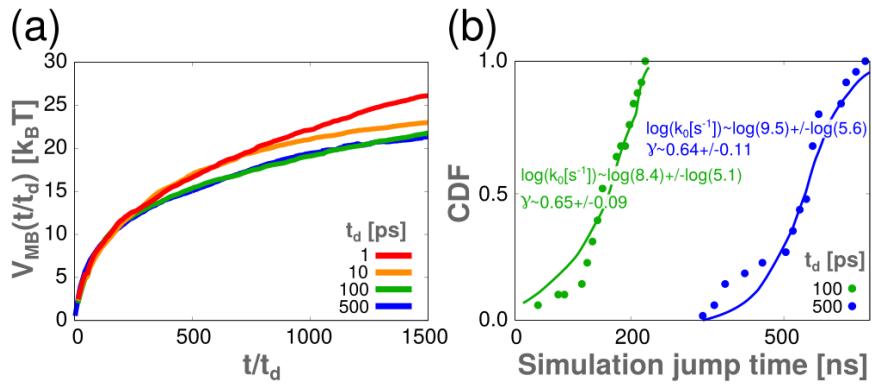


Figure 7.7: CDK2-ligand unbinding biased simulations. (a) V_{MB} as a function of the time normalized by the bias deposition time (t/t_d) for the different simulation setups. Note that the $t_d = 1$ and 10 ps curves do not collapse. (b) CDF of the simulation barrier-crossing (jump) times for the WTMetaD simulations for the converged setups: 100 (green) and 500 ps (blue). The solid lines show the fits of the KTR method using the numerical $S(t)$ and LM formalism to estimate the optimal parameters. The extracted rate k_0 and efficiency of CV γ are within error (shown with the same color scheme), and $> 90\%$ of KS-tests pass for these setups. Errors were estimated from the range of the 30 and 70 percentiles of bootstrap samples.

7.4 CONCLUSIONS AND OUTLOOK

In this chapter we demonstrated that the KTR method works for cases where the system has enough time to equilibrate between bias depositions. The applications presented here also allow us to show the limits of our method and provide convergence criteria to assess its validity. There are several points of the KTR theory that can be improved in future work. Automatized methods to determine when the barrier-crossing occurs might be helpful. γ could be used in iterative approaches to refine the biasing direction for the cases where the barrier-crossing events are too slow along an initial poor CV. We have shown that Kramers' theory and the quasi-adiabatic assumption breakdown in the overbiasing regime; therefore, it would be useful to extrapolate the theory to non-equilibrium conditions.¹⁶³ The main

limitation of the current approach, besides force-field inaccuracy, appears to be the implicit approximation of considering the transition process as the escape from a single “reactants” well to a single “products” well: future work will focus on extending the methodology to take explicitly into account a network of possibly multiple transition pathways across several metastable states, yielding improved accuracy for complex systems.

8

Free energy landscapes, diffusion coefficients and kinetic rates from data-driven Langevin models

8.1 INTRODUCTION

In this chapter we address the problem of constructing accurate mathematical models of the dynamics of complex systems projected on a collective variable. To this aim we introduce a conceptually simple yet effective algorithm for estimating the parameters of Langevin and Fokker-Planck equations from a set of short, possibly out-of-equilibrium molecular dynamics trajectories, obtained for instance from transition path sampling or as relaxation from high free-energy configurations. The approach maximizes the model likelihood based on any explicit expression of the short-time propagator, hence it can be applied to different evolution equations. Here we demonstrate the numerical efficiency and robustness of the algorithm in a set of double-well potentials used as benchmark systems. Our methodology allows reconstructing the accurate thermodynamics and kinetics of activated processes, namely free energy landscapes, diffusion coefficients, and kinetic rates. Compared to existing enhanced sampling methods, we directly exploit short unbiased trajectories, at a competitive computational cost.

Projecting the dynamics of the system on one (or a few) CV q leads to a theoretical framework describing equilibrium thermodynamics through free energy landscapes, and kinetics (including out-of-equilibrium relaxation) through Langevin and Fokker-Planck equations.^{16,164} Depending on the nature of the physical system, and on the observational time resolution τ , three main types of Langevin equations appear in the literature. In the non-

Markovian case, a generalized Langevin equation is customarily employed, including a deterministic force (the gradient of the free energy $-dF/dq$) and time-correlated friction and noise, connected by the fluctuation-dissipation theorem.^{16,165} Such time correlation, or memory effects can be significant for the short-time dynamics of small solutes immersed in a liquid bath.^{166–168}

However, in many applications in physics, chemistry or biology an observational time resolution τ coarser than the memory timescale is pertinent, so that a Markovian Langevin equation accurately describes the projected dynamics. Depending on the intensity of the friction, or, in other words, on the extent of its characteristic time compared to τ , which depends on the physical process and on the choice of CVs, a second-order underdamped Langevin equation featuring inertia (and reminiscent of Hamilton's equations) or a first-order overdamped equation are appropriate.^{17,45,169}

8.1.1 LANGEVIN MODELS FROM MD TRAJECTORIES

A number of algorithms aim at parametrizing Langevin models starting from trajectories of many-particle systems.^{167,168,170–197} More in details, non-Markovian friction (the so-called memory kernel) can be reconstructed in different ways, for instance based on a set of time-correlation functions and Volterra integral equations (see, e.g., Refs. 167,168,171), or via likelihood maximization.¹⁹⁷ In the case of Markovian (overdamped) Langevin equations, the latter approach^{176,181,183}, together with direct estimation of Kramers-Moyal coefficients.^{183,189} have been employed

Often, only friction is addressed, assuming that the free-energy profile is known in advance, or that it can be directly estimated from the histogram of the CV based on long equilibrium trajectories reversibly sampling all relevant states (see, e.g., Refs. 168,178,181). Clearly, this severely limits the scope of the techniques, given that ergodic molecular dynamics (MD) trajectories can be affordably generated only in the case of low kinetic barriers (a few $k_B T$).

In the case of rare events, some approaches are able to exploit enhanced sampling simulations, like umbrella sampling, to alleviate the timescale problem^{175,183,190}, even though the cost of such simulations to estimate free-energy and diffusion profiles remains considerable in presence of high barriers. Moreover, such acceleration techniques usually rely on an initial guess of CVs to be biased: a bad choice can lead to poor convergence and to explore unfavorable transition mechanisms compared to brute force MD, while a change of CVs requires in general to generate new biased trajectories, multiplying the costs. At variance with previous attempts, in this work we address this issue by basing our Langevin optimization approach on unbiased TPS-like³⁶ trajectories, spontaneously relaxing from barrier tops, and faithfully reproducing brute force transitions thanks to the reliability of TPS techniques. In our case, the choice of the CV is performed *a posteriori*, and multiple CVs can be tested at negligible computational cost since no new MD trajectories are required.

8.1.2 CHALLENGES IN MODELLING REAL-SYSTEMS DYNAMICS WITH THE LANGEVIN EQUATION

Whether the projected dynamics of a given system can be accurately modeled by a specific flavor of Langevin equation depends upon two main observational choices: the CV q and the time resolution τ of the trajectory. In particular, it is expected that memory effects decrease when CVs become similar to the committor function.^{110,169,175} For a fixed CV, too small τ can render visible non-Markovian effects. Albeit acknowledged in the literature, these principles have seldom been applied to critically assess how accurately a given Langevin model reproduces the original dynamics for different CVs and different time resolutions.^{181,189,194} In many cases a single system is studied and Langevin predictions at fixed time resolution are compared with independent estimates of free energies and rates. We also remark a further important point: τ cannot be freely increased to enhance Markovian behavior, since too large values rarefy the sampled trajectory to the point of lacking information on barrier regions.

As a result of all the previous concerns and difficulties, to date CV-based Langevin equations have been used sporadically and most often heuristically, and are not systematically applied to access the thermodynamics and kinetics of a wide range of complex physico-chemical processes. It is our aim to improve the state of the art in this respect, with particular care to the assessment of time resolution effects on the accuracy of the reduced-dimensionality model, as detailed below.

8.1.3 OUR PROPOSAL

Here we discuss our recently published method¹⁹⁸, a conceptually simple and computationally efficient method to construct data-driven Langevin models of rare events in complex systems. Since equilibrium MD data is not necessary for training, the method can be applied irrespective of the height of the barriers separating metastable states. We also provide an explicit procedure to assess the optimal resolution τ leading to accurate Langevin models, and to predict their reliability. The chapter is organized as follows: first, we describe the algorithm used to optimize the models and the computational details of the MD simulations, then, we test the algorithm on a double-well potential with known free energy, diffusion profiles and kinetic rates. Finally, we discuss the scope of the method, its limitations and future application perspectives.

8.2 THEORETICAL METHODS

8.2.1 MATHEMATICAL MODELS OF PROJECTED DYNAMICS

Langevin equations can be obtained from Hamilton's equations of motion by projecting the high-dimensional deterministic dynamics on a subset of the phase-space variables^{16,197}. When a single generalized coordinate q and its conjugated momentum $p = mq$ are considered,

the Generalized Langevin Equation (GLE) is customarily obtained:

$$m\ddot{q} \equiv \dot{p} = -\frac{\partial F(q)}{\partial q} - \int_0^\infty dt' \Gamma(t') p(t-t') + R(t) , \quad (8.1)$$

where $F(q) = -k_B T \log \rho_{\text{eq}}(q)$ is the free-energy landscape (potential of mean force) corresponding to the stationary distribution $\rho_{\text{eq}}(q)$, and the fluctuation-dissipation theorem $\langle R(0)R(t) \rangle = k_B T m \Gamma(t)$ connects the friction $\Gamma(t)$ and the random force $R(t)$ to guarantee thermal equilibrium in the long-time limit^{199,200}. There are two restrictive hypotheses behind Eq. 8.1, often tacitly assumed: 1) q is a linear function of Cartesian coordinates, hence the mass m is q -independent¹⁷⁸, and 2) the projected-out bath coordinates are linearly coupled to the q coordinate, hence $\Gamma(t)$ is q -independent^{46,201,202}. In many applications, Eq. 8.1 is further simplified by assuming a fast decay of the memory function $\Gamma(t)$ (where "fast" refers to the observational timescale of interest), so that

$$-\int_0^\infty dt' \Gamma(t') p(t-t') \approx -p(t) \int_0^\infty dt' \Gamma(t') = -\gamma p(t) , \quad (8.2)$$

leading to the memory-less Standard Langevin Equation (SLE)

$$\dot{p} = -\frac{\partial F(q)}{\partial q} - \gamma p + \sqrt{2k_B T m \gamma} \eta(t) , \quad (8.3)$$

where the fluctuation-dissipation theorem simplifies to $\langle R(0)R(t) \rangle = 2k_B T m \gamma \delta(t)$, with $\eta(t)$ a Gaussian white noise: $\langle \eta(t) \rangle = 0$, $\langle \eta(0)\eta(t) \rangle = \delta(t)$.

In the limit of strong friction, velocity fluctuations away from the equilibrium distribution are quickly damped (where "quick", once again, refers to the observational timescale of interest), suggesting to assume $\dot{p} \approx 0$ yielding the widely-employed Overdamped Langevin Equation (OLE)^{17,45}:

$$\dot{q} = -\frac{1}{m\gamma} \frac{\partial F(q)}{\partial q} + \sqrt{\frac{2k_B T}{m\gamma}} \eta(t) \equiv -\beta D \frac{\partial F(q)}{\partial q} + \sqrt{2D} \eta(t) , \quad (8.4)$$

where we introduced the diffusion coefficient $D = k_B T / m\gamma = 1/\beta m\gamma$, which allows to avoid the explicit appearance of the mass in the equation. Here $\eta(t)$ is a Gaussian white noise. As remarked above, a generic coupling between q and the bath coordinates leads to position-dependent $\gamma(q)$ and $D(q)$, therefore a realistic description of the dynamics generally requires a non-constant $D(q)$ ¹⁷⁶, and the use of the overdamped equation can be formally justified (for instance, it provides exact mean first passage times (MFPT)) when q is the optimal CV, i.e., any monotonic one-to-one function of the committor¹⁶⁹. It can be shown that, as a consequence of position dependence in D , an additional force $\partial D(q)/\partial q$ must be included

to recover the correct $\rho_{\text{eq}}(q)$ ²⁰³:

$$\dot{q} = -\beta D(q) \frac{\partial F(q)}{\partial q} + \frac{\partial D(q)}{\partial q} + \sqrt{2D(q)} \eta(t) , \quad (8.5)$$

$$\rho_{\text{eq}}(q) = \frac{C}{D(q)} \exp \left[\int^q dq' \frac{d}{dq'} (-\beta F + \log D) \right] \quad (8.6)$$

$$= C e^{-\beta F(q)} . \quad (8.7)$$

We illustrated the need of adding this extra “D” force in the Appendix B, Fig. B1.

The corresponding Fokker-Planck equation is the Smoluchowski equation⁴⁶, describing the evolution of the probability density $\rho(q, t)$ from an initial distribution $\rho(q, 0)$ (typically $= \delta(q - q_0)$) under given boundary conditions:

$$\begin{aligned} \frac{\partial}{\partial t} \rho &= \frac{\partial}{\partial q} \left(\beta D \frac{\partial F}{\partial q} \rho - \frac{\partial D}{\partial q} \rho \right) + \frac{\partial^2}{\partial q^2} (D\rho) \\ &= \frac{\partial}{\partial q} \left[D e^{-\beta F} \frac{\partial}{\partial q} (e^{\beta F} \rho) \right] . \end{aligned} \quad (8.8)$$

The latter equation can be seen as a continuity equation $\partial_t \rho(q, t) = -\partial_q J(q, t)$, where $\rho_{\text{eq}}(q)$ is obtained by directly integrating the equation $J = 0$ (Eq. 8.6).

Even though a coordinate transformation from q to $s(q)$ with $\dot{s} = \dot{q}/\sqrt{2D(q)}$ would allow to work with a position-independent diffusion coefficient (see Ref. 164), we do not pursue this route since we aim to build Langevin models for given physically-motivated coordinates q .

8.2.2 OPTIMIZATION OF LANGEVIN MODELS (PARAMETER ESTIMATION)

To construct an optimal model of the q -projected dynamics, we adopt an efficient algorithm that maximizes the likelihood $\mathcal{L}(\theta)$ to observe the trajectory $\{q_i \equiv q(t_i)\}_{i=1,\dots,N}$, $t_{i+1} - t_i = \tau$ given the Langevin model Eq. 8.5. The parameters θ encode the shape of $F(q)$ and $D(q)$. To obtain an explicit form for the likelihood, we consider the formal solution of the Fokker-Planck equation $\partial_t \rho = L\rho$ (Eq. 8.8), calling L the Fokker-Planck operator (the adjoint of the generator of the stochastic process)^{13,164}:

$$\begin{aligned} p(q', t + \tau | q, t) &= e^{L\tau} \delta(q' - q) \\ &= \left[1 + L\tau + \frac{1}{2} L^2 \tau^2 + \dots \right] \delta(q' - q) . \end{aligned} \quad (8.9)$$

For small τ , the short-time transition probability density p , also called propagator, is an ingredient of path-integral formulations of long-time transition probabilities, and, neglecting

terms of the order of τ^2 , can be explicitly obtained (using Fourier transforms^{164,183)} in Gaussian form:

$$p(q', t + \tau | q, t) \approx \frac{1}{\sqrt{2\pi\mu}} e^{-(q' - q - \phi)^2 / 2\mu}, \quad (8.10)$$

$$\phi = a\tau = (-\beta DF' + D')\tau, \mu = 2D\tau, \quad (8.11)$$

where the prime indicates d/dq , and a is the drift term in Eq. 8.5 (note that this propagator is exact for any $\tau > 0$ if F' and D are constant). A better approximation, still retaining the Gaussian form Eq. 8.10, can be obtained by keeping terms up to order τ^2 in Eq. 8.9, applying for instance Drozdov's approach based on the cumulant-generating function:²⁰⁴

$$\phi = a\tau + \frac{1}{2}(aa' + Da'')\tau^2, \mu = 2D\tau + (aD' + 2a'D + DD'')\tau^2. \quad (8.12)$$

Armed with this explicit expression for the short-time propagator, it is possible to write the likelihood of the observed trajectory $\{q_i\}_{i=1,\dots,N}$:

$$-\log \mathcal{L}(\theta) = \sum_{i=1}^{N-1} \frac{1}{2} \log[2\pi\mu_i(\tau)] + \sum_{i=1}^{N-1} \frac{[q_{i+1} - q_i - \phi_i(\tau)]^2}{2\mu_i(\tau)}, \quad (8.13)$$

where $\phi_i(\tau) \equiv \phi(q_i, \tau)$, $\mu_i(\tau) \equiv \mu(q_i, \tau)$. The optimization of the Langevin model is achieved by minimizing $-\log \mathcal{L}(\theta)$ as a function of the parameters, directly yielding the free energy and diffusion profiles. Eq. 8.13 is minimized using a simple iterative stochastic algorithm: starting from an initial guess for the free energy and diffusion profiles (represented on a discrete grid of 1000 points), there are two types of trial moves (with probabilities 20% and 80%, respectively): i) adding to the profiles a Gaussian at a random position, with a random width between $(q_{max} - q_{min})/20$ and $(q_{max} - q_{min})$ and height between $-k_B T/2$ and $k_B T/2$ (with 20% probability to modify either profile, and 80% to modify both, at each optimization step), or ii) scaling the profiles by a Gaussian-distributed random factor with mean 1 and standard deviation 0.3. The moves are accepted or rejected on the basis of a Metropolis criterion on $(-\log \mathcal{L}_{new} + \log \mathcal{L}_{old})/\tilde{T}$. The effective temperature \tilde{T} is automatically scaled during the iterations to keep the acceptance close to a target value of 5%.

The algorithm we propose for the optimization of the likelihood function (Fig. 8.1) is straightforward:

1. Start with trajectories $q(t)$ from MD and with an initial guess for the free energy and diffusion profiles.
2. Compute $-\log \mathcal{L}(\theta)$ for the initial data.
3. Make a random change of the free energy and diffusion profiles and compute the new $-\log \mathcal{L}(\theta)$.
4. Use a Metropolis criterion to accept or reject the new parameters.
5. Repeat N_{iter} times the procedure.

Note that alternatives to stochastic Monte Carlo optimization could be used, however we

appreciate the simplicity and generality of this approach, that does not require to estimate derivatives $\partial\mathcal{L}/\partial\theta$.

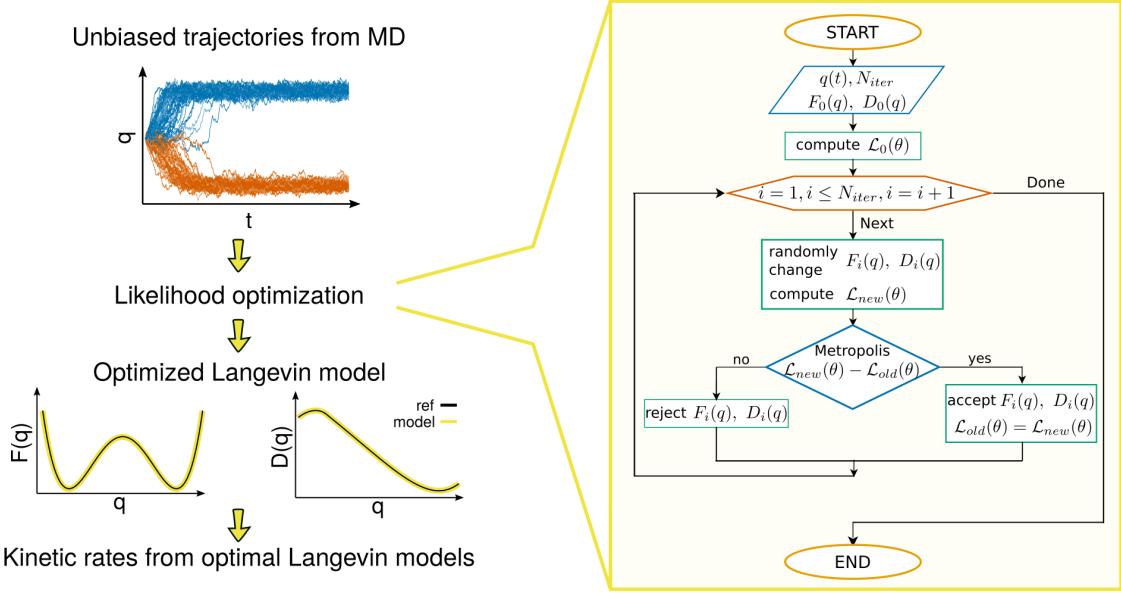


Figure 8.1: Schematic illustration of the approach proposed to optimize Langevin models of projected dynamics. The input data is formed by unbiased MD trajectories projected on a selected collective variable, $q(t)$. A likelihood function $\mathcal{L}(\theta)$ is maximized (i.e., $-\log \mathcal{L}(\theta)$ is minimized) by randomly varying the parameters encoding the free energy $F(q)$ and diffusion profile $D(q)$ of an overdamped Langevin equation, until optimally modeling the original MD data. At each iteration, the new parameters are accepted or rejected according to a Metropolis criterion. The optimal Langevin model corresponds to the last accepted $F(q)$ and $D(q)$ profiles, and is used for further estimation of kinetic rates (see Methods for details).

The initial guess of the parameters can facilitate the optimization of the likelihood function: without requiring any additional data besides the projected MD trajectories, an approximate value of D in a free-energy minimum can be estimated from the variance of q divided by its autocorrelation time.¹⁷⁶ As sketched in Fig. 8.1 the input trajectories $q(t)$ typically contain a final part fluctuating in the bottom of a well, that can be employed in the latter formula. The profile $D(q)$ is simply set equal to the constant estimated value. As for the initial $F(q)$, we found that the specific form is largely irrelevant. For the sake of simplicity we start from a flat profile $F(q) = 0$.

The code and input files are freely available on GitHub (<https://github.com/physix-repo/optLE>).

8.2.3 INTEGRATION OF THE LANGEVIN EQUATION OF MOTION

To generate input data for the benchmark systems, as well as to test the quality of the approximate propagator used in the likelihood function (see previous section), we numerically integrate Langevin trajectories using the Milstein algorithm, an order 1 strong Taylor scheme

superior to Euler-Maruyama in the case of multiplicative noise (i.e., position-dependent diffusion):²⁰⁵

$$\begin{aligned}\Delta q = & \left[-\beta D(q)F'(q) + \frac{1}{2}D'(q) \right] \Delta t \\ & + \sqrt{2D(q)\Delta t} G + \frac{1}{2}D'(q)\Delta t G^2 ,\end{aligned}\tag{8.14}$$

where G is a Gaussian-distributed random number with zero mean and unit variance. The integration timestep Δt is chosen after testing the observed long-time histogram of q against $e^{-\beta F(q)}$, where $F(q)$ is the exact landscape used in the Langevin equation.

8.2.4 OPTIMAL TIME RESOLUTION OF THE REDUCED MODEL

The accuracy of the optimized Langevin model with respect to the original MD trajectory $q(t)$ crucially depends on the time resolution τ adopted, $q(t) \equiv q(k\tau)$, $k = 1, \dots, N$. In fact, it is expected that only for a sufficiently large τ the overdamped equation can model accurately the MD data. To guide the choice of τ we introduce diagnostics able to predict what choices lead to accurate models.

In general, for given profiles $F(q)$ and $D(q)$, a numerical Langevin integrator can be “inverted” to estimate the value of an effective noise $G(t)$ corresponding to each observed displacement $\Delta q = q(t + \tau) - q(t)$ in the input MD trajectory. The simplest formula corresponds to the Euler-Maruyama integrator:

$$G = \frac{\Delta q + (\beta DF' - D')\tau}{\sqrt{2D\tau}} .\tag{8.15}$$

Another option is to invert the Milstein integrator, Eq. 8.14:

$$G = \frac{1}{D'\tau} \left(-\sqrt{2D\tau} + \sqrt{(2D\tau) - (2D'\tau) \left(\frac{1}{2}D'\tau - \beta DF'\tau - \Delta q \right)} \right) .\tag{8.16}$$

If the resulting effective noise would have been generated by the model, the $G(t)$ trajectory would be random, with zero mean, unit variance and no time correlation: $\langle G(t)G(t') \rangle = \delta_{t,t'}$ (the latter symbol indicating the Kronecker delta, since time is discretized). Therefore, to estimate the expected accuracy of the optimal Langevin model it is possible to inspect the mean, variance, and correlation time τ_{noise} of the effective noise¹⁹⁴. For the latter quantity, we adopt here a simple operative definition, as the first time where the autocorrelation function $C(t) = \langle G(0)G(t) \rangle / \langle G^2 \rangle$ drops below 1/100. Note that in case of oscillatory behavior of $C(t)$ it is important to inspect the latter to assess if the estimate is acceptable. We find that the expressions for the noise from Euler-Maruyama (Eq. 8.15) or Milstein (Eq. 8.16) integrators yield similar results (see Appendix B, Fig. B2), therefore, for simplicity, in what follows we use Eq. 8.15 to determine τ_{noise} .

As a second metric to diagnose problems in model accuracy we generate (using the integrator Eq. 8.14) $M = 100$ Langevin trajectories of duration τ and timestep $\Delta t \ll \tau$ starting from each point of the input trajectory $\{q_i\}_{i=1,\dots,N}$. The resulting final displacements $\tilde{q}_{i+1}^k - q_i$ (with $k = 1, \dots, M$) are shifted and scaled according to the theoretical propagator of the model, Eq. 8.10, by subtracting ϕ_i and dividing by $\sqrt{\mu_i}$. In this way, their distribution would be Gaussian with zero mean and unit variance if the propagator was exact, and an average likelihood can be evaluated to estimate deviations from such ideal behavior:

$$-\overline{-\log \mathcal{L}_{\text{prop}}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{1}{M} \sum_{k=1}^M \frac{1}{2} \left[\log 2\pi + \left(\frac{\tilde{q}_{i+1}^k - q_i - \phi_i}{\sqrt{\mu_i}} \right)^2 \right], \quad (8.17)$$

which yields $\frac{1}{2}[\log 2\pi + 1] \approx 1.419$ in the ideal case. For increasing τ , the propagator becomes unreliable, reducing the likelihood and hence allowing to predict, once again, the quality of the model (see Fig. 8.4).

8.2.5 KINETIC RATES

In general cases, the kinetic rate for a transition A→B can be computed by means of the reactive flux (Bennett-Chandler) technique, that includes recrossing effects beyond transition state theory⁴²

$$k(t) = \frac{\langle \dot{q}(0) \delta[q(0) - q^*] h_B[q(t)] \rangle}{\langle h_A \rangle},$$

where $h_A(q)$ and $h_B(q)$ are the indicator functions of metastable states A and B. Under the hypothesis of rare transition events, there is a plateau in the values of k for a broad interval of t values larger than the typical microscopic timescale (e.g., of molecular vibrations) and smaller than the MFPT. q^* is any point close to the separatrix: only efficiency is affected if they differ significantly.

The rate can be computed numerically by shooting N_s MD simulations from q^* with random atomic velocities drawn for the Maxwell-Boltzmann distribution: for a given t , $k(t)$ is the product of the sum of initial velocities for trajectories ending up in B, divided by the total number of shots N_s , and of a term containing information about the free energy barrier

$$k(t) = \frac{\sum_{j=1}^{N_s} \dot{q}_j(0) h_B[q_j(t)]}{N_s} \frac{e^{-\beta F(q^*)}}{\int_{\Omega_A} dq e^{-\beta F(q)}}. \quad (8.18)$$

We remark that for rare transitions with overdamped one-dimensional dynamics it is also possible to compute the MFPT (the inverse of the rate) as an integral over the free energy and diffusion landscapes:

$$k^{-1}(q_0) = \int_{q_0}^b dy \frac{e^{\beta F(y)}}{D(y)} \int_a^y dz e^{-\beta F(z)}, \quad (8.19)$$

where, in the typical case of the escape from a well (even though the formula is general),

q_0 is the initial position in the well, a is the position of a reflective boundary located on the opposite side as the transition barrier, and b is the position of an absorbing boundary located beyond the barrier.⁴² In the following, we adopt as location of the absorbing boundary b the first minimum in $\frac{d}{dq}k^{-1}$ beyond the free energy barrier.

8.3 RESULTS AND DISCUSSION

In the new approach, the Langevin model is parametrized from unbiased MD trajectories spanning the relevant CV region (transition region). Such trajectories can be easily obtained as spontaneous out-of-equilibrium relaxation from initial high free-energy configurations, e.g., the top of a barrier. Several effective techniques are available to discover transition states and reactive paths at a moderate computational cost^{26,36,97,206–208}: although not trivial, this task is generally much less involved than reconstructing accurate free energy profiles. In the following we use as input CV-projected MD trajectories $q(t)$ obtained from TPS or by shootings from the barrier top and relaxing into the minima. Note that such trajectories are short, since in rare events the transition path time is typically very fast compared to the waiting time in a minimum, i.e., the MFPT.

Initially, we will test our method on a simple test system: a 1D double-well potential. Using this benchmark system we have full control over the dynamics of the system. In the following, we vary the shape of the free energy and diffusion profiles to demonstrate the applicability of the method independent of the shape of these profiles.

8.3.1 LANGEVIN MODELS FROM OVERDAMPED TRAJECTORIES

As a first necessary test, we verified that our method yields Langevin models accurately reproducing the correct $F(q)$ and $D(q)$ when using as input 50 ps long overdamped trajectories relaxing from the barrier top of double-well landscapes. We vary the shape of the free energy profile starting from symmetric or asymmetric profiles, and also vary the height of the free energy barrier as well as the shape of the diffusion profile. The Langevin dynamics trajectories used as input are obtained using the overdamped integrator, Eq. 8.14, having as initial configuration $q = 0$. We optimize langevin models for different time resolutions. For each system, we perform 10 independent optimizations to obtain the average and standard error.

We show in Fig. 8.2 that the reconstructed free energy profile and position-dependent diffusion coefficient deviate less than $1 k_B T$ and 10%, respectively, from the reference exact profiles, regardless of τ , within the time resolution where the quality of the approximate propagator is acceptable. This test leaves us with two important conclusions: *i*) The optimization of overdamped Langevin models through Likelihood maximization, as expected, leads to accurate free energy and diffusion profiles using overdamped trajectories as initial data, i.e., the method is self-consistent. *ii*) When the input trajectories are Markovian, the smaller the τ the more accurate the optimal Langevin model. When the time resolution is large the propagator loses accuracy and the initial data set used for optimization is reduced,

leading to models with more fluctuations compared to the reference profiles. However, even for $\tau = 1$ ps, the optimal Langevin models reproduce accurately the shape of the free energy and diffusion profiles. With this result in mind, we will study in detail the importance of time resolution in the next section.

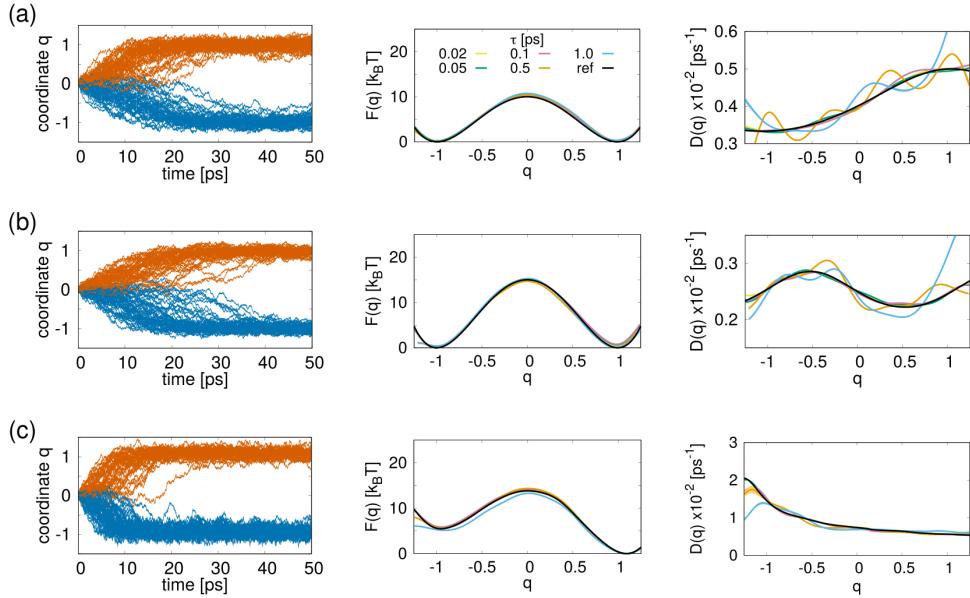


Figure 8.2: Optimal overdamped Langevin models trained on overdamped Langevin trajectories. (a) Symmetric double-well potential with $10 k_B T$ barrier and diffusion coefficient $0.003 \leq D(q) \leq 0.005 \text{ ps}^{-1}$. (b) Symmetric double-well potential with $15 k_B T$ barrier and diffusion coefficient $0.0022 \leq D(q) \leq 0.0028 \text{ ps}^{-1}$. (c) Asymmetric double-well potential with $13 k_B T$ barrier and diffusion coefficient $0.005 \leq D(q) \leq 0.025 \text{ ps}^{-1}$. Left: 100 input trajectories of 50 ps relaxing from the barrier top. Center: optimal free energy profiles. Right: optimal diffusion profiles. The different time resolutions τ employed to optimize the models are shown in colors. The exact profiles are shown with black lines. The thickness of the lines correspond to standard error over 10 independent stochastic optimization runs.

8.3.2 LANGEVIN MODELS FROM UNDERDAMPED TRAJECTORIES

As a second, more difficult benchmark we verify whether our method yields Langevin models accurately reproducing the correct $F(q)$ and $D(q)$ starting from underdamped Langevin trajectories. The input data is composed by 100 short underdamped Langevin trajectories $q(t)$, each one 20 ps long, relaxing from the barrier top of a double-well landscape (see Fig. 8.3). The exact free energy and diffusion profiles are shown in black in Fig. 8.4.

Since the initial trajectories are generated with an underdamped integrator, building the overdamped model entails projecting out the momentum p conjugated to the CV position q . As a result, a q -based Langevin model is expected to display memory (time correlation) in the friction and noise for τ comparable to the integration timestep $\Delta t = 10^{-4}$ ps. Adopting coarse-enough τ resolutions, on the other hand, should lead to memory-less behavior, well reproduced by an overdamped model.

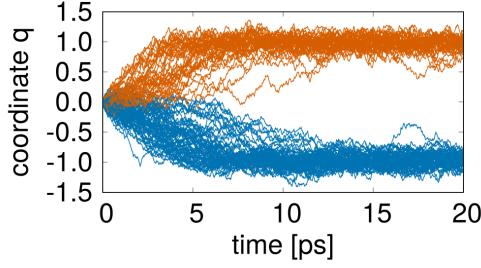


Figure 8.3: 100 input trajectories generated with a non-overdamped Langevin integrator, using the standard Langevin equation: $\dot{p} = -\frac{\partial F(q)}{\partial q} - \gamma p + \sqrt{2k_B T m \gamma} \eta(t)$. The benchmark system is a symmetric double-well free-energy landscape with $10 k_B T$ barrier ($m = 1 k_B T \cdot \text{ps}^{-2}$). 53% of the trajectories ended up in the right basin and 47% in the left basin.

To test this hypothesis, in Fig. 8.4 we generate Langevin models using $0.01 \text{ ps} \leq \tau \leq 1 \text{ ps}$: for the smallest τ value, the model strongly overestimates the barrier and underestimates the diffusion coefficient. As τ increases the reconstructed model becomes more accurate (see Fig. 8.4) until at large values of τ the error increases again. We found that there is a relatively narrow range of τ values for which the Langevin overdamped model is accurate: to identify

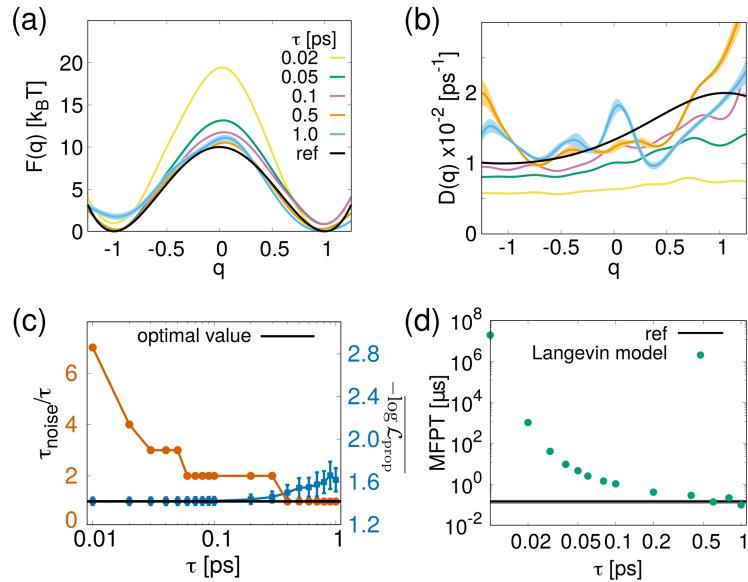


Figure 8.4: Optimal overdamped Langevin models (Eq. 8.5) for different resolutions τ starting from underdamped trajectories on a double-well free-energy landscape. (a) $F(q)$ and (b) $D(q)$ profiles (exact ones are shown with black lines); vertical bars correspond to standard error over 10 independent stochastic optimizations. (c) Autocorrelation time of the effective noise τ_{noise} (Eq. 8.16) normalized by τ (i.e., number of correlated steps), and average likelihood quantifying the accuracy of the propagator (Eq. 8.17). The black line represents the ideal values. Vertical bars are the standard deviation over all trajectory steps Δq . (d) MFPTs for the optimal overdamped Langevin models shown in (a) and (b), computed with Eq. 8.19: the reference value in black is computed with reactive flux on the exact profiles in black in panels (a) and (b). Vertical bars are the standard error over 10 independent Langevin models (hardly visible at this scale).

the optimal range in real-world applications we propose two complementary tests that allow identifying the upper and lower boundaries of τ bracketing high-quality Langevin models.

The lower acceptable τ is the smallest time displaying no memory effects, measured through the autocorrelation of the effective Langevin noise, i.e., the noise extracted a posteriori by analyzing the original trajectory with the optimal Langevin model, Eq. 8.16 (see the Appendix B, Fig. B2 for details). The upper limit on τ corresponds to the onset of a significant deviation between the approximate and the exact propagators (the latter estimated from accurate numerical integration, see the Methods section 8.2.4 and Eq. 8.17). Both diagnostics are reported in Fig. 8.4c, and together they predict that the Langevin model should attain the best accuracy for τ values close to 0.5 ps: large enough to avoid memory, but small enough for the short-time propagator to remain reliable.

Overall, when selecting τ based on the proposed diagnostics, models converge smoothly and rapidly ($< 10^6$ Monte Carlo steps) to a good approximation of the exact results: as few as 100 reference trajectories are sufficient to reconstruct the free energy profile to within 1 $k_B T$ and $D(q)$ to within 15% error (we remark that fluctuations in $D(q)$ appear converged with respect to the different parameters used in the reconstruction of the models and irrespective of the input trajectories). The same conclusions can be derived regardless of the shape of the diffusion and free energy profiles (see the Appendix B, Fig. B3).

Besides estimating free energy and diffusion profiles, a key objective of the new method is the accurate prediction of kinetic rates at a moderate computational cost. Fig. 8.4d shows the MFPTs obtained with Eq. 8.19 using the optimal Langevin models. Too small values of τ lead to strongly overestimated MFPT, as expected from the overestimated barriers, while the rate becomes accurate for τ values within the predicted optimal window (Fig. 8.4c). On the other hand, for τ larger than optimal, MFPTs in the same order of magnitude as the reference value are still obtained. We highlight that the input trajectories for the optimization of the Langevin models are 20 ps each (an aggregate time of 2 ns for the 100 trajectories), while the MFPT is $0.14 \pm 0.08 \mu\text{s}$, meaning that we can recover MFPTs orders of magnitude larger than the input trajectories.

We note that, alternatively, we can compute the MFPT using the reactive flux technique^{54,209}, free from transition state approximations. This traditionally requires two distinct sets of expensive simulations, one to compute the free energy barrier and another to compute a correlation function accounting for recrossings (see the Methods section 8.2.4 and Appendix B, Table B1). However, this technique becomes inexpensive if combined with our approach, since the free energy barrier is directly obtained from likelihood optimization on short trajectories, while recrossing statistics can be accurately estimated from cheap synthetic Langevin trajectories. We remark that the rates computed from Eq. 8.19 are similar to those of reactive flux (see Table 8.1). Because Eq. 8.19 gives us the MFPT directly from the free energy and diffusion profiles of the optimal Langevin models (without the need to launch new synthetic trajectories) while being as accurate as reactive flux, we choose to estimate the MFPT with this equation in the rest of this and the next chapters.

| τ | MFPT (μs) from Eq. 8.19 | MFPT (μs) from Eq. 8.18 |
|--------|--------------------------------------|--------------------------------------|
| 0.02 | 1052 ± 51 | 1398 ± 11 |
| 0.05 | 4.6 ± 0.3 | 3.1 ± 0.3 |
| 0.10 | 1.1 ± 0.1 | 0.83 ± 0.02 |
| 0.50 | 0.27 ± 0.03 | 0.20 ± 0.05 |
| 1.00 | 0.10 ± 0.01 | 0.13 ± 0.01 |

Table 8.1: Comparison between MFPTs estimated from Eq. 8.19 and from reactive flux for the double-well potential in Fig. 8.4. The brute force estimate is $0.14 \pm 0.08 \mu\text{s}$. The reactive flux estimate is obtained with Eq. 8.18 and 10000 trajectories, each one 20 ps-long.

8.4 CONCLUSIONS AND OUTLOOK

The calculation of free energy landscapes and kinetic rates are key tasks of computer simulations of complex systems. Even though these two tasks are usually tackled using different *ad hoc* techniques², the main contribution of the present work consists in demonstrating that they can be *simultaneously* achieved in a conceptually simple way, by estimating Langevin models starting from a relatively inexpensive set of about one hundred TPS-like trajectories, regardless of the barrier height. Likelihood maximization – an efficient parameter estimation technique – is combined with a double test bracketing the time resolution τ , granting control over the accuracy of the model. The choice of a suitable τ is especially important for accurately estimating the free energy, because the rate has an exponential dependence on the barrier while only a linear dependence on D .

The resulting Markovian Langevin equations reproduce well the quantitative thermodynamic and dynamic properties of the original benchmark system starting from both, over-damped and underdamped Langevin dynamics trajectories, yielding accurate kinetic rates, despite a gap of many orders of magnitude with respect to the short trajectories used for training. We also remark that equilibrium properties are systematically recovered from out-of-equilibrium data: standard transition path sampling trajectories are the golden standard for the study of transformation mechanisms, however, since such data set is a small subset of all possible (reactive and non-reactive) pathways, lacking Boltzmann distribution of the configurations, it cannot be used for the *direct* estimate of equilibrium histograms (i.e., free energies) and rate matrices (e.g., in Markov state models) by simple averaging.^{37,210} Here we show that the contrary is true, provided bare transition paths are employed to train a suitable stochastic model (see also Refs. 175,176,183,211–213 for related or alternative ideas). Note that Langevin equations of motion, compared to other machine-learning tools, retain a direct physical interpretation, with the separation between a systematic average force and friction/noise effects describing the projected-out degrees of freedom (commonly referred to as “the bath”).

In the next chapter we will extend the application of this method to the analysis of all-atoms MD simulations of complex systems. We will demonstrate that processes such as association and dissociation of large molecules (compared to the size of the solvent),

liquid-solid phase transition of inorganic materials, conformational changes of small proteins and protein-protein association, among others, can be accurately described by means of Markovian Langevin models.

Then, we will introduce a second issue that concerns the effect of projecting the dynamics on sub-optimal CVs, rather than on the ideal reaction coordinate, commonly identified with the committor function^{75,169,214}. In principle, the CV definition can be systematically optimized with an iterative scheme, where a single initial set of MD trajectories can be employed to build optimal Langevin models for progressively improving CVs. A related idea was recently proposed in the case of discrete Markov state models.^{41,215} Therefore, our new approach represents a starting point both for the systematic modelling of MD data through physically-motivated stochastic equations as well as for machine-learning approaches to reaction coordinate optimization.

ACKNOWLEDGMENTS

The method introduced in this chapter was developed in co-authorship with Fabio Pietrucci. We gratefully acknowledge very insightful discussions with Alessandro Barducci, Christoph Dellago, Gerhard Hummer, Gerhard Stock, A. Marco Saitta, Rodolphe Vuilleumier, Hadrien Vroylandt, Riccardo Ferrando, and Andrea Pérez-Villa.

9

Kinetic rates from data-driven Langevin models: Applications

9.1 INTRODUCTION

The double-well potentials in the previous chapter allowed us to benchmark the scope and limitations of the method while having full control over the system. To assess the method in realistic complex systems we study the interaction of the fullerene C₆₀ and C₂₄₀ dimer in water solution. As explained in Chapter 7, the fullerene dimer has a simple enough free energy landscape for benchmarking, while dissociation of the dimer remains a rare event. In this chapter we also show how the dynamics of various processes can be modeled by means of the overdamped Langevin equation: from phase transitions in metal mixtures to protein-protein interactions we extract free energies profiles and rates from limited MD data. The diversity of these applications demonstrates the usefulness of our method and at the same time allows us to determine its limitations and aspects to improve in the future.

9.2 INTERACTION OF C₆₀ AND C₂₄₀ FULLERENE DIMER IN WATER SOLUTION

9.2.1 DESCRIPTION OF THE SYSTEM

In Chapter 7, section 7.2, we introduced the fullerene C₂₄₀ dimer system in water solution. In this section, we use both the fullerene C₆₀ and the fullerene C₂₄₀ dimer to benchmark our method. Changes in the size of the fullerene molecule are reflected in a change in the height of the free energy barrier to move from the associated state to the dissociated state. In this particular case, the C₆₀ dimer in water presents a dissociation free energy barrier of around 7.0 $k_B T$ compared to a barrier of around 14.5 $k_B T$ for the C₂₄₀ dimer. These

systems then allow us to study the effect of the height of the free energy barrier (the higher the barrier, the lower the transition rate), and illustrate the usefulness of our method in the study of rare events. In the following, we show the characterization of the free energy and diffusion profiles for these systems. Additionally, we diagnose the best τ resolution to study the dissociation dynamics of the fullerene dimers and estimate the transition rates.

9.2.2 SIMULATION METHODS

FULLERENE C₆₀ DIMER

We performed MD simulations to study the interaction between two fullerene C₆₀ molecules solvated by 2398 water molecules, in a simulation box of 3.607 × 3.607 × 3.607 nm with periodic boundary conditions. The initial fullerene coordinates and topology are taken from Ref. 137. MD simulations are carried out using GROMACS v2019.4^{139,140} patched with PLUMED 2.5.3¹⁴¹. We adopted the SPC water model⁶⁸ and the OPLS-AA force-field⁶² for carbon. Geometry minimization exploited the steepest descent algorithm, stopped when the maximum force was ≤ 50 kJ/mol·nm. We used the leapfrog algorithm to propagate the equations of motion and the nonbonding interactions were calculated using a PME scheme with a 1.2 nm cutoff for the part in real space. We performed a 100 ps equilibration in an NVT ensemble with a stochastic velocity rescaling scheme²¹⁶ followed by a 100 ps equilibration in an NPT ensemble using the Parrinello-Rahman barostat⁷¹ with a time step of 1 fs. We generated MD production trajectories without restraints, with a time step of 1 fs in the NPT ensemble at 298 K and 1 atm. The reference free-energy profile of the association/dissociation of the fullerene C₆₀ dimer in water as a function of the distance between the centers of mass (d) was computed from 5 unbiased simulations of 500 ns each, from the population histogram: $F(d) = -k_B T \log \rho_{\text{eq}}(d)$. We estimated the standard deviation of each histogram bin using the 5 simulations. The resulting profile is consistent with the free-energy profile in Ref. 137.

The diffusion profile used for comparison was computed from umbrella sampling simulations where we restrict d with a harmonic potential around a reference value d_i over different windows i ,

$$U(d; d_i) = \frac{c}{2}(d - d_i)^2 \quad (9.1)$$

Then, the diffusion coefficient in the window i is estimated as the ratio between the variance of the variable ($\sigma_{d_i}^2$) and the autocorrelation time of the variable itself (τ^{corr}):¹⁷⁶

$$D(d_i) = \frac{\sigma_i^2}{\tau_i^{\text{corr}}} \quad (9.2)$$

We setup a window every 0.1 nm in a range of d between 1.0 and 1.8 nm. We performed 1 ns MD simulation for equilibration and 10 ns MD simulation production in each window with a spring constant $c = 1000 \frac{\text{kJ}}{\text{mol}}$. For each window, we split the 10 ns MD simulations in 10 independent blocks of 1 ns each. We estimated the variance of d and the autocorrelation time in each block. Finally, we report the average $D(d_i)$ value and its standard error over

the blocks. We note that $D(d)$ does not show large fluctuations with respect to the choice of c (see Appendix B Fig. B4).

To generate the input trajectories for the construction of Langevin models, we employed aimless-shooting (AS)^{95,96} simulations. We obtained a first reactive trajectory by shooting from randomly-picked configurations of the unbiased trajectories with d between 1.2 nm and 1.3 nm. We then performed AS using a separation $\delta t = 0.1$ ps between successive shooting points, and a total length of 20 ps for each trajectory. We obtained a total of 1110 accepted trajectories relaxing from the transition state region with an acceptance rate of 14%. The script we developed to perform AS with GROMACS is publicly available at <https://github.com/physix-repo/aimless-shooting>.

To estimate the dissociation rate we use the reactive flux formalism over 1000 aimless shooting trajectories. We define the dissociated state as $d \geq 1.34$ nm and the associated state as $d \leq 1.17$ nm (see Fig. 9.1a). The value of the MFPT obtained with reactive flux was validated with the brute force estimate from 100 unbiased MD trajectories launched from the bound state: 6.1 ± 1.2 ns versus 6.5 ± 0.6 ns, respectively. Finally, we estimated the transmission coefficient as⁴²:

$$\kappa = \frac{k_{A \rightarrow B}^{RF}}{k_{A \rightarrow B}^{TST}}, \quad (9.3)$$

where $k_{A \rightarrow B}^{RF}$ is the dissociation rate constant from reactive flux (Eq. 8.18), and $k_{A \rightarrow B}^{TST}$ is the dissociation rate constant from transition state theory. The estimated transmission coefficient for the dissociation of the fullerene C₆₀ dimer is 0.37, meaning that the number of recrossings before relaxation is small.

FULLERENE C₂₄₀ DIMER

The MD simulations were performed using the same protocol as in Chapter 7, section 7.2. The description of the estimation of the reference free energy profile is also presented in section 7. We obtained the diffusion profiles for comparison from umbrella sampling simulations following the same procedure and with the same parameters used for C₆₀ fullerene dimer. We placed a window every 0.1 nm in a range of d from 1.6 to 2.6 nm. We performed AS simulations following the same procedure used for C₆₀ fullerenes. We obtained 1515 accepted trajectories of 100 ps each, with an acceptance rate of 15%. We calculate the dissociation rate using 1000 trajectories and reactive flux: $9.4 \pm 1.1 \mu\text{s}$. We define the dissociated state as $d \geq 2.01$ nm and the associated state as $d \leq 1.9$ nm. The estimated transmission coefficient is 0.6.

9.2.3 RESULTS AND DISCUSSION

We modeled the dynamics of two fullerene molecules in explicit water, considering two different sizes C₆₀ and C₂₄₀ (Fig. 9.1). The free-energy landscape as a function of the distance d between fullerenes' centers of mass features a minimum for the bound complex (with dif-

ferent depth for the two sizes) and a relatively flat region for the dissociated state, see black lines in Fig. 9.3. The resulting rare dissociation events and diffusion-controlled association events^{137,217–219} are analogous to other processes such as protein-protein interaction. We obtained reference free-energy profiles using 2.5 μ s-long brute-force MD trajectories for the C₆₀ dimer, while we employed WT-MetaD⁹⁴ for the C₂₄₀ dimer due to its very low dissociation rate. The diffusion profiles are obtained from sets of 10 ns umbrella sampling simulations every 0.1 nm in distance between the center of mass (see section 9.2.2 for details).

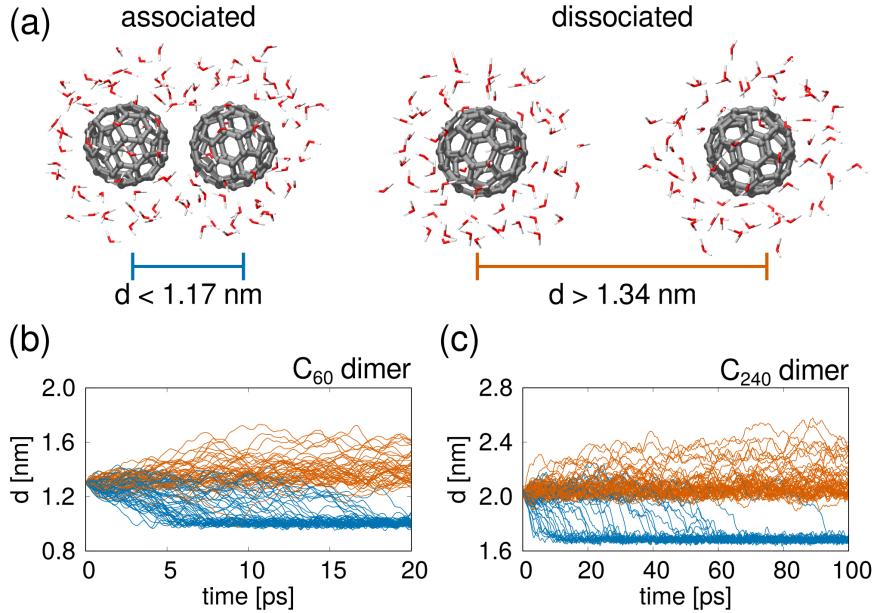


Figure 9.1: (a) Characteristic atomic configurations for the C₆₀ dimer in bulk water solution; only waters distant less than 5 Å from carbon atoms are shown. The associated and dissociated states are defined in terms of the distance d between the centers of mass of the fullerenes. (b) 100 MD trajectories (from aimless shooting) starting from the transition state ensemble and featuring the association (blue) or dissociation (orange) of the C₆₀ dimer, and (c) of the C₂₄₀ dimer, respectively.

We maximized the likelihood of Langevin models following the procedure presented in Fig. 8.1, starting from short aimless shooting⁹⁵ TPS trajectories (20/100 ps each, cumulative duration 2/10 ns for C₆₀/C₂₄₀, see Fig. 9.1). Since the Langevin model optimization is performed through a Monte Carlo procedure, we run 10 independent Langevin model optimizations per system. An example of the convergence of $-\log \mathcal{L}(\theta)$ as a function of the number of iterations is shown in Fig. 9.2.

We adopted a range of time resolutions τ between 0.1 and 2 ps: for fast time scales, memory is expected to arise from the projection of the solvated fullerenes many-body dynamics on a single CV. Sizable memory effects are evident in the effective noise (Eq. 8.16) for $\tau < 0.5$ ps (see also Appendix B, Fig. B2), severely affecting the accuracy of $F(d)$. For larger τ values, similarly to the case of the analytic double-well potential in Chapter 8, the Langevin models become increasingly accurate: memory decays after about 0.5 ps for the smaller fullerenes and after about 1.0 ps for the bigger ones. As for the diffusion profile,

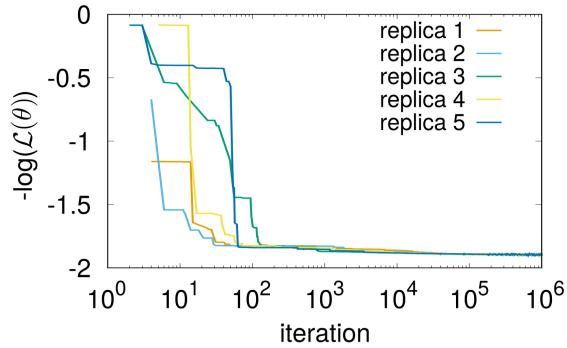


Figure 9.2: Example of the convergence of $-\log \mathcal{L}(\theta)$ as a function of the number of iterations for the C₆₀ fullerene dimer in solution. Colors represent different independent optimizations. A similar behavior with a smooth convergence is observed in all the Langevin models presented in this work.

all models show $D(d)$ in the same order of magnitude, in good agreement with the diffusion coefficients estimated from umbrella sampling¹⁷⁶ (see section 9.2.2). We remark that generating 100 short TPS trajectories for Langevin optimization can be computationally faster than generating a sufficient amount of umbrella sampling trajectories to estimate the diffusion profile.

On the other hand, the quality of the approximate short-time propagator Eq. 8.12 decreases as τ increases. Accurate results are clearly identified in the range of τ that minimizes both non-Markovian behavior and error in the propagator (see Fig. 9.3). Moreover, the short time transition probability $p(q', t + \tau|q, t)$ predicted by the optimal Langevin model (Eq. 8.10) is in good agreement with a distribution collected from short MD shootings of length τ (see example in Appendix B, Fig. B5). Interestingly, optimizing the models based on 1000 reference trajectories instead of 100 yields minimal improvements (see Appendix B, Fig. B6).

Finally, we assessed the accuracy of kinetic predictions given by the models. The dissociation MFPT is on the nanosecond and microsecond scales, respectively, for C₆₀ and C₂₄₀ dimers, i.e., 2 and 5 orders of magnitude slower than the out-of-equilibrium MD trajectories used for training. For each value of τ , we computed MFPTs using Eq. 8.19 on the optimal models, yielding Fig. 9.4. We emphasize that the latter approach allows to compute the MFPT without generating any additional expensive MD trajectories.

For too small values of τ , the barrier is overestimated (see Fig. 9.3), leading to overestimate the MFPTs by several orders of magnitude. For the C₆₀ dimer the MFPT becomes accurate (compared to the brute force result) for $\tau > 0.5$ ps, while for the C₂₄₀ dimer this level of accuracy requires $\tau > 1$ ps: once again, these are the optimal time resolutions predicted by the diagnostics in Fig. 9.3. Taken together, all these results demonstrate that the approximations inherent in the Langevin models are under control, leading to accurate predictions about the thermodynamics and kinetics of complex systems.

We emphasize that our approach is not restricted to transition path sampling: in principle, any set of unbiased trajectories spanning the regions of interest could be used to train the

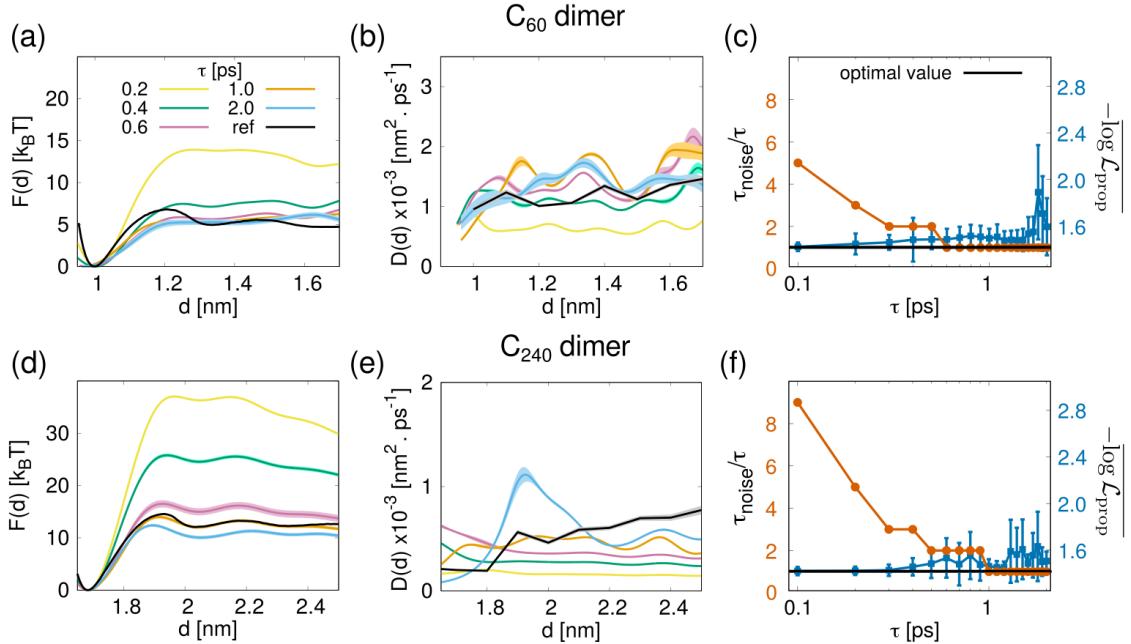


Figure 9.3: Optimal Langevin models for the interaction of C_{60} and C_{240} fullerene dimers in explicit water, for different values of the time resolution τ . The input data set consists in 100 MD trajectories (aimless shooting) relaxing from the transition state ensemble in Fig. 9.1. (a), (d) Free-energy profiles $F(q)$, compared to the reference ones in black: vertical bars correspond to standard deviations over 10 independent stochastic optimizations. (b), (e) Diffusion profiles $D(q)$, compared to the umbrella sampling ones in black: vertical bars correspond to standard deviations over 10 independent stochastic optimizations. (c), (f) Autocorrelation time of the effective noise (Eq. 8.16) normalized by τ (i.e., number of correlated steps), and average likelihood quantifying the accuracy of the propagator (Eq. 8.17). The black line represents the ideal value for both quantities; vertical bars refers to the standard deviation over all trajectory steps Δq .

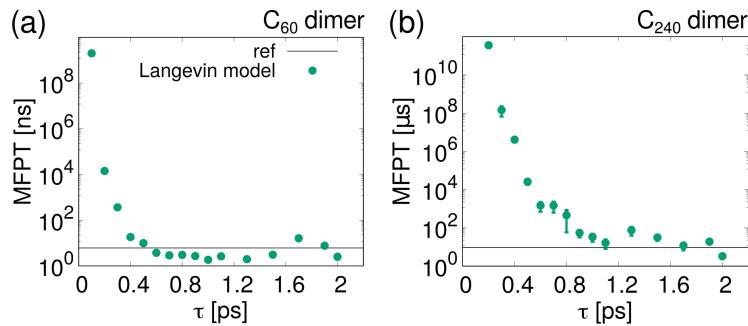


Figure 9.4: MFPTs computed from the optimal Langevin models generated with different time resolutions τ , for the dissociation of fullerene (a) C_{60} , and (b) C_{240} dimer, in water solution. MFPTs are estimated from Eq. 8.19 (reference value in black). Vertical bars are the standard error over 10 independent Langevin models.

Langevin model. For example, in Fig. 9.5 we optimized Langevin models using as input trajectories that start from the fullerene C_{60} dimer dissociated state and reach the associated

state. The C_{60} dimer association process presents a much lower free energy barrier than the dissociation process, therefore obtaining trajectories that present an association event is less computationally expensive than obtaining dissociation trajectories. We found accurate free energy and diffusion profiles using only association trajectories at an appropriate τ . This result is important because it means that our method can be applied in processes where one direction of the transition is more favorable than another and therefore equilibrium properties could be obtained from unidirectional trajectories. We will illustrate this advantage of the method in the following sections.

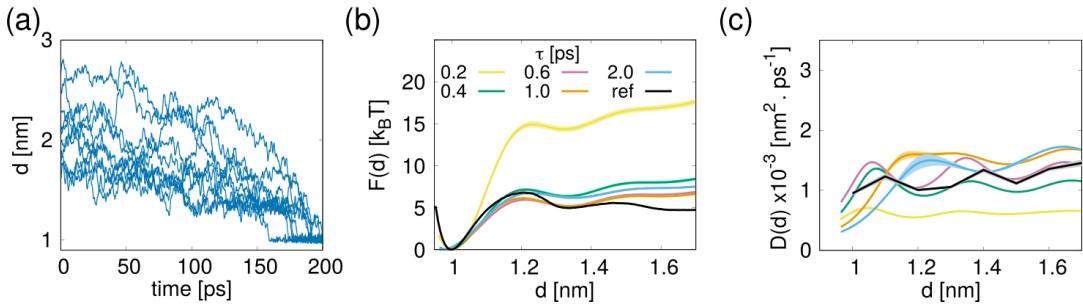


Figure 9.5: Langevin models from association trajectories. (a) Examples of 100 association trajectories used to optimized the (b) free energy and (c) diffusion profiles from 200 ps MD trajectories of the C_{60} fullerene dimer. The different time resolutions τ employed to optimize the models are shown in colors. The reference free energy profile (computed from brute force MD) and diffusion profile (computed from umbrella sampling) are shown with black lines. The thickness of the lines correspond to standard error over 10 independent optimization runs.

Since the dynamics of various activated processes can be modeled with the overdamped Langevin equation, in the following sections we will present some preliminary results in which we optimize Langevin models to recover free energy profiles and rates in two important processes: the crystallization of Germanium-Telluride and the interaction of the Barnase-Barstar protein complex. We emphasize that although these applications of our method are still work in progress, the results obtained so far are promising.

9.3 CRYSTALLIZATION OF GERMANIUM-TELLURIDE

9.3.1 DESCRIPTION OF THE SYSTEM

Phase change materials can display different optical and electrical properties depending on the phase they are in. This characteristic makes fast crystallization materials of great interest to the electronic and optical industry, especially for their use as active layers in rewritable optical disks and in phase change memories.^{220,221} The crystallization rate is a critical parameter of these type of materials, as phase change memories made with materials of crystallization rate < 10 ns become competitive in comparison with others kinds of memories available in the market. One of these materials that has already been used in phase

change memories is the pseudo-binary compound GeTe-Sb₂Te₃ (GST), as it displays rapid phase-switching, low power consumption, high thermal stability, and long cyclability.²²⁰

Experimental evidence indicates that the presence of GeTe core network in GST promotes the fast crystallization.^{221,222} From a computational point of view, the fast phase-switching of GeTe allows its study on time scales that can be achieved through *ab initio* MD with density functional theory, but still limiting the size of the system to few hundreds of atoms. In order to reproduce the experimentally observed bulk properties, the simulated system must have a larger size (thousands of atoms). Machine learning potentials emerge as an alternative to overcome this computational limitation. In particular, Lee, *et al.*²²⁰ developed a Neural Network Potential (NNP) for the GeTe system that is able to reproduce the kinetics of the crystallization in good agreement with experiments. Therefore, in the following we study the amorphous to crystalline phase transition of the GeTe alloy from MD simulation using this NNP, see Fig. 9.6.

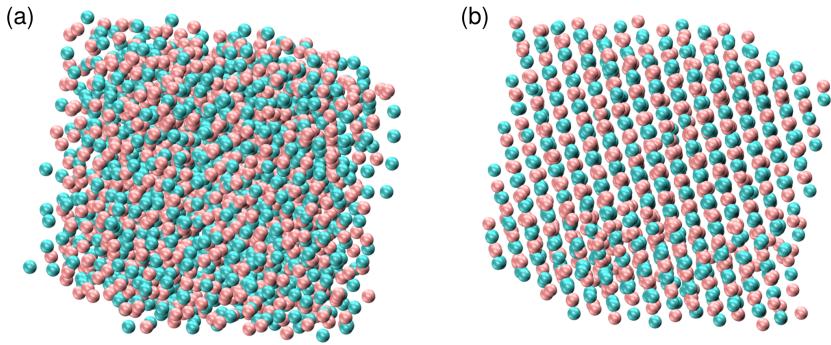


Figure 9.6: Amorphous to crystalline phase transition of the GeTe. Two representative frames corresponding to (a) amorphous and (b) crystalline phase of GeTe compound, in a MD simulation box. Germanium and Telluride atoms are represented in pink and cyan, respectively.

9.3.2 SIMULATION METHODS

In this section, we used as input data a set 32 brute force MD simulations provided by Julien Lam (Center for Materials Elaboration and Structural Studies, France). The MD simulations were performed in LAMMPS²²³ using the NNP in Ref. 220. A cubic simulation box was used with 1000 Germanium and 1000 Tellurium atoms. The simulations were performed at temperature $T = 600$ K, which is in the middle of the relevant temperature range (500 – 700 K) for the programming protocols of the electronic memories.²²¹ The simulations were started from the amorphous phase (see Fig. 9.6a). The total simulation time was 4 ns for each trajectory.

Using this set of MD simulations we projected the trajectories on two different CVs: the total potential energy (U) and the Parrinello entropy²²⁴ (S), see Fig. 9.7. Then, we optimize 10 independent overdamped Langevin models for each CV following the same procedure as in the previous section. We performed 1×10^6 iterations in each optimization. The initial

free energy profile to start the optimizations is set to $F(q) = 0$ $k_B T$; $q = S, U$ and the initial diffusion profile is set to $D(q) = 2 \times 10^{-5}$ ps $^{-1}$; $q = S, U$. The time resolution for the optimization of the Langevin models was set to $\tau = 2$ ps.

The reference crystallization rate was computed from the set of brute force trajectories using the distribution of jump times and the mean lifetime method, explained in Chapter 3, section 3.3.2.

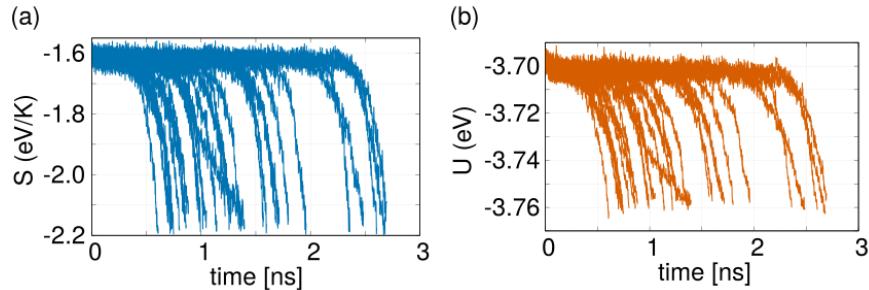


Figure 9.7: Trajectories for the optimization of the Langevin models. (a) Total potential energy (U) and (b) Parrinello entropy (S). We computed the CVs from the beginning of the trajectories that start at the amorphous phase and stopped when they reach the crystalline phase. A total of 32 unbiased MD trajectories are presented.

9.3.3 RESULTS AND DISCUSSION

In many crystallization processes of inorganic materials, the transition path time scales are long enough to allow its study with relative long time resolutions. Therefore, an overdamped regime is likely the appropriate description, and the method proposed in Chapter 8 for the optimization of Langevin models to access free energy, diffusion profiles, and kinetic rates, could be an ideal candidate to study crystallization processes. In particular, the fast crystallization of materials as GeTe offer an attractive playground for testing this method in real-life applications.

In what follows, we study the crystallization process of GeTe, by optimizing Langevin models from brute force MD simulations. The GeTe amorphous to crystalline phase transition has been studied both experimentally and computationally.^{220,221,225} Since our method can also be used with trajectories going in only one direction, MD simulations that start from the amorphous phase and reach the crystalline phase are, in principle, sufficient to apply the method (see last part of section 9.2.3). To optimize the Langevin models we follow this phase transition by means of two CVs: the total potential energy (U) and the Parrinello entropy²²⁴ (S). Both CVs are in principle good order parameters, since crystallization from a disorder state is expected to decrease U and S . However, it is of course unclear how closely such CVs approximate the ideal reaction coordinate or the committor. Although each MD trajectory have a total duration of 4 ns, we only take for analysis the part of the trajectories until it reaches the crystalline state ($S < -2.2$ eV/K) (see Fig. 9.7). Using these trajectories

as input data, we optimize 10 independent Langevin models for each CV. The resulting free energy and diffusion profiles are presented in Fig. 9.8.

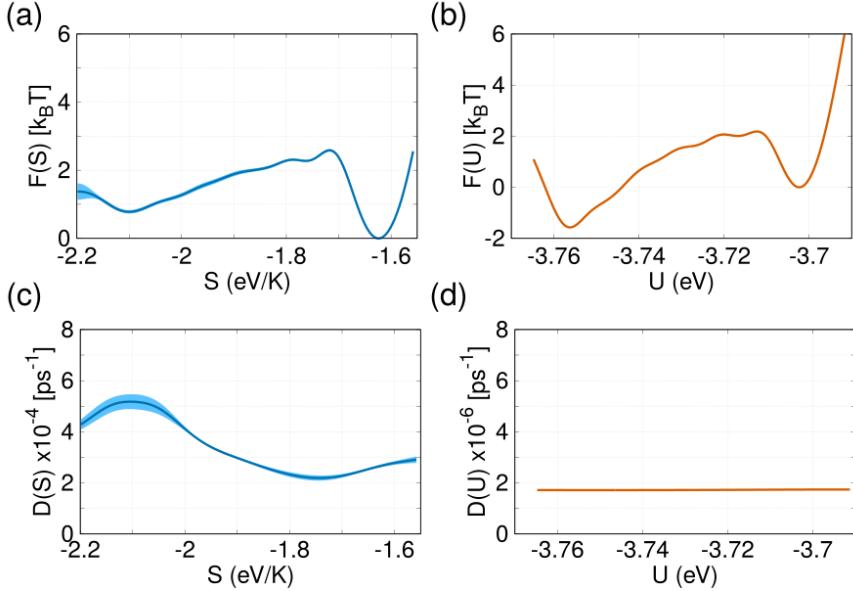


Figure 9.8: Optimal Langevin models for the Parrinello entropy S (left) and the total potential energy U (right). (a) and (b) optimal free energy profiles for S and U , respectively. (c) and (d) optimal diffusion profiles for S and U , respectively. The thickness of the lines correspond to standard error over 10 independent optimization runs.

In Fig. 9.8a and Fig. 9.8b, the free energy minimum on the right corresponds to the amorphous phase. Both free energy profiles display a typical landscape with two minima separated by a free energy barrier. As explained in Chapter 4, the height of the free energy barrier in the free energy profile depends on the CV used for the projection. In this case, the free energy barrier for each CV is: $\Delta F(S) = 2.58 \pm 0.03 k_B T$ and $\Delta F(U) = 2.066 \pm 0.008 k_B T$. We also note that the diffusion coefficient is higher for the variable S than for U .

Given these optimal Langevin models, we can estimate the rate for the transition from amorphous state to crystalline state using Eq. 8.19. For this purpose, we invert the free energy and diffusion profiles (in order to continue integrating from left to right in Eq. 8.19), using as absorbing boundary a the minimum value of the respective CV, q_0 is the position of the minimum in the amorphous state, and b is the position of the minimum after the free energy barrier. The crystallization rates obtained with this method are shown in Table 9.1.

We note that both the rate from brute force MD simulations and the rate obtained from the optimal Langevin models are in good agreement. We note that the rates are also in good agreement with those reported by Bruns, *et al.*²²⁵ for a 20-nm thick GeTe layer in a phase change. However, current devices use phase change memories where the GeTe has a different density and a crystallization time of about 30 ns (see Ref. 220), therefore the direct value of comparison for our Langevin models is the brute force rate instead the experimental value. Based on these results we conclude that both U and S represent good approximations of the

| Method | MFPT (ns) |
|-----------------------------|---------------|
| Experimental ²²⁵ | ~ 1 |
| Brute force | 1.3 ± 0.1 |
| Langevin models S | 0.9 ± 0.1 |
| Langevin models U | 1.1 ± 0.1 |

Table 9.1: Brute force vs Langevin model based estimated of the MFPT for the GeTe crystallization. The experimental value correspond to a crystallization rate for a phase change memory with 20-nm thick GeTe reported in Ref. 225. Brute force estimate is obtained from the distribution of jump times in brute force trajectories. The MFPT from the Langevin models is estimated using Eq. 8.19.

ideal reaction coordinate, at least at the conditions investigated.

The results of the GeTe fast crystallization study presented in this section are promising, however, they represent a preliminary test, the main interest of our approach being the estimation of slower nucleation rates, in a regime where brute force MD simulations are expensive. We will work to validate in-depth and extend the results in the near future as follows: (i) expanding the range of temperatures studied, doing more MD simulations for temperatures between 500 K to 700 K (to mimic experimental conditions); (ii) for those temperatures where it is not possible to observe phase transitions in a reasonable computational time we plan to use AS and optimize Langevin models from these trajectories as done for the fullerenes; (iii) possibly we will also study the system at other densities to assess the effect of density on the crystallization rate. Furthermore, in the same spirit, we plan to extend the applications of our Langevin models to estimate the kinetics of the phase transitions in the following systems: liquid-solid phase transition in Binary Lennard-Jones liquids²²⁶ and water nucleation using the mW water model.²²⁷

9.4 PROTEIN-PROTEIN INTERACTIONS IN BARNASE-BARSTAR COMPLEX

9.4.1 DESCRIPTION OF THE SYSTEM

Despite the paramount importance of PPI in controlling most of biological functions, including pathological processes from cancer to Alzheimer, today it is routinely impossible to predict with accuracy whether two proteins bind together, the corresponding free energy gain, and the association/dissociation rate constants. Some experimental techniques allow to characterize the most stable structures, association kinetics and binding affinities. However, the experimental study of the interaction mechanism is still limited only to certain types of proteins that exhibit, for example, paramagnetic properties, allowing the use of specific experimental techniques, such as double mutation cycles and paramagnetic relaxation enhancement.²²⁸ Therefore, all-atoms MD simulations combined with data-driven Langevin models could emerge as a useful tool to reveal complete mechanisms and predict free energies and rates.

In this section we study a prototypical example of protein-protein association: the bacterial ribonuclease barnase with its inhibitor barstar (see Fig. 9.9). The bacterium *Bacillus amy-*

loliquefaciens synthesize the barstart protein to inhibit the ribonuclease activity of its binding partner barnase. The Barnase-Barstar (BB) complex bounds tightly allowing the retention of barnase until its secretion.^{229,230} This complex presents a diffusion-limited association well-characterized experimentally with a binding free energy of $\Delta G_b = -19.0 \pm 0.2 \text{ kcal} \cdot \text{mol}^{-1}$, a $k_{on} = 6.0 \times 10^8 \text{ s}^{-1} \cdot M^{-1}$ and a $k_{off} = 8.0 \times 10^{-6} \text{ s}^{-1}$.^{231,232}

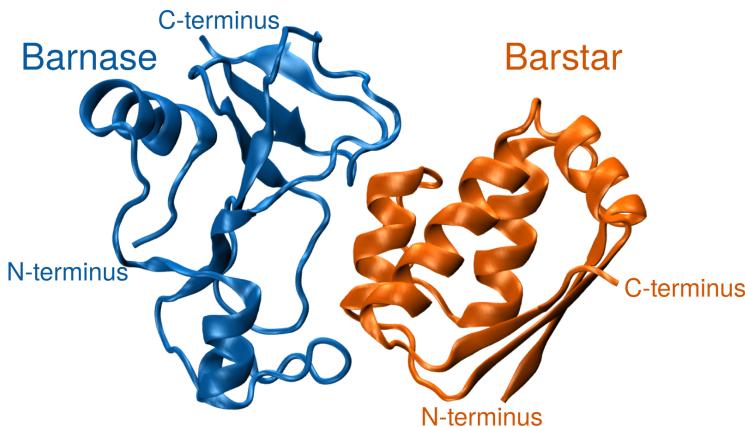


Figure 9.9: Crystallographic structure of the barnase-barstar complex. PDB code: 1BRS.²³³ Barnase (blue) and barstar (orange) are shown in cartoon representation. Terminal C and N tails are also shown.

From the computational point of view, binding free energies for this complex has been computed from atomistic simulations starting from the bound pose and using a Potential of Mean Force (PMF)-based enhanced sampling method.²³⁴ Also, the binding free energy landscape has been reconstructed using metadynamics.²³⁵ Regarding kinetics, the estimation of the association rate requires expensive sets of MD simulations. Plattner, *et al.*⁹⁹ built a kinetic model of BB association using MSM with an aggregate simulation time of 1.7 milliseconds. On the other hand, Pan, *et al.*²²⁸ estimated the association k_{on} and k_{off} rate for the BB complex combining unbiased MD simulations and the tempered binding enhanced sampling method, with aggregate simulation times on the order of hundreds of microseconds.

In what follows, we will use Langevin models to study the association of the BB complex. This process entails a high complexity given the number of intermediate states and the different interactions that these proteins can present. In particular, we will focus on the issue of finding the CV that best describes the process, since we will show that the accuracy of Langevin models crucially depends on the CV quality.

9.4.2 SIMULATION METHODS

MOLECULAR DYNAMICS SIMULATIONS

In this section, we used as input data a set unbiased MD simulations provided by D. E. Shaw Research. The details of the MD simulations are reported in the methods section in ref. 228. The data set consisted of 61 MD trajectories (including 28 successful association

events) with an average length of $5 \mu s$ each, and an aggregated simulation time of $212.7 \mu s$ (see Table 1, condition BB_2 in ref. 228). Each trajectory contains the positions of all the atoms of the complex in frames printed every 180 ps. No water molecules were provided inside the trajectories.

COLLECTIVE VARIABLE DEFINITIONS

Starting from the raw MD data, we projected the dynamics of the system in a set of simple CVs to be able to distinguish the associated and dissociated state and to optimized Langevin models. All the CVs values from the trajectories were computed using PLUMED 2.5.3¹⁴¹. Below we describe each CV in detail.

Native contacts: we defined the total native contacts as the contacts between the α -carbon atoms (C_α) at the protein-protein interface. The protein-protein interface is defined as any pair of C_α atoms, one from each protein monomer, within 10 Å of each other in the experimentally determined complex (PDB code: 1BRS). Then, the native contacts were calculated as the coordination number between two groups

$$c = \sum_{i \in A} \sum_{j \in B} c_{ij} , \quad (9.4)$$

with

$$c_{ij} = \frac{1 - \left(\frac{r_{ij} - d_0}{r_0} \right)^n}{1 - \left(\frac{r_{ij} - d_0}{r_0} \right)^m} , \quad (9.5)$$

where $d_0 = 0$, $r_0 = 0.3$ nm, $n = 6$ and $m = 8$. In the sum of Eq. 9.4, group A are the C_α atoms at the interface of Barnase, while the group B are the C_α atoms at the interface of Barstar.

Pair contacts: the pair contacts are defined in the same way as the total native contacts (using Eq. 9.4) but instead of computing the sum for all the atoms in the group, we computed the sum by pairs of atoms. The parameters d_0 , r_0 , n and m stay the same as for the total native contacts.

Interface RMSD: We define the interface RMSD as the distance RMSD of the C_α atoms at the protein-protein interface of the complex with respect to the same atoms in the experimental structure. The distance RMSD between two structures \mathbf{X}^A and \mathbf{X}^B is calculated as:

$$d(\mathbf{X}^A, \mathbf{X}^B) = \sqrt{\frac{1}{N(N-1)} \sum_{i \neq j} [d(\mathbf{x}_i^a, \mathbf{x}_j^a) - d(\mathbf{x}_i^b, \mathbf{x}_j^b)]^2} , \quad (9.6)$$

where N is the number of C_α atoms at the protein-protein interface of the experimental complex and $d(\mathbf{x}_i, \mathbf{x}_j)$ represents the distance between atoms i and j .

Path CV: Using the interface RMSD, we defined a RMSD-based path CV. For a detailed

explanation of path CVs see Chapter 4, section 4.3.2. For this purpose, we take 14 reference frames from one of the trajectories, spanning from the dissociated complex to the complex in the associated state. The frames were equidistant in the RMSD. The distance RMSD between the reference frames was 4 Å.

To recall the equation in section 4.3.2, the S path CV is defined as

$$S(\mathbf{X}) = \frac{1}{n-1} \frac{\sum_{i=1}^n (i-1) \exp(-\lambda D(\mathbf{X}^i, \mathbf{X}))}{\sum_{i=1}^n \exp(-\lambda D(\mathbf{X}^i, \mathbf{X}))} \quad (9.7)$$

Using as distance metric $D(\mathbf{X}^i, \mathbf{X})$ the interface RMSD. Here, λ was set to 5.6 Å⁻¹, such that $\lambda D(\mathbf{X}^i, \mathbf{X}) = 2.3$.

LANGEVIN MODELS OPTIMIZATION

Using the set of MD simulations with successful association events we optimize 10 independent overdamped Langevin models for each CV following the same procedure as in the previous section. We performed 1×10^5 iterations in each optimization. The time resolution for the optimization of the Langevin models was set to $\tau = 180$ ps.

9.4.3 RESULTS AND DISCUSSION

The experimental dissociation rate constant for the Barnase-Barstar complex is $k_{off} = 8.0 \times 10^{-6}$ s⁻¹.²³¹ Therefore, starting from the associated complex, the dissociation would take an average time of 125000 s, or approximately 1.5 days. This time scale is well beyond what is achievable today with MD simulations. Therefore, simulating the dissociation of the BB complex starting from the associated state is not possible with standard MD, as it involves overcoming high free energy barriers. However, the association process of the complex is much more favorable, as it mainly is a down-hill process, so if you have good computational resources (like D. E. Shaw Research has with Anton supercomputer) it is possible to collect an ensemble of association trajectories starting from the dissociated state in time scales of hundreds of μ s. Taking into account that our method allows to optimize Langevin models from unidirectional trajectories, e.g., association trajectories, we characterize the association of the BB complex using the unbiased MD trajectories obtained by Pan, *et al.*²²⁸ These trajectories start from the dissociated state and end up in the associated complex, see Fig. 9.10.

The process of association of a protein complex can involve many intermediate states, since different interactions between amino acids can occur, leading to encounter complexes in non-native poses, dissociation and subsequent association (to the native state or to other metastable states). The trajectories analyzed in this section are a clear example of this. Most of these trajectories present complex pathways with one or more events where the two proteins encounter in different metastable states before dissociating and re-associating to the native complex.²²⁸ Therefore, finding a one-dimensional CV to describe the dissociation process and optimizing the Langevin models is not a trivial task. We follow a set of CVs

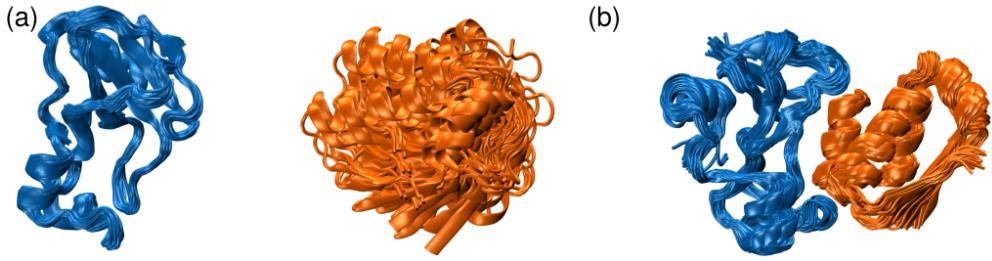


Figure 9.10: Initial and final frames for each trajectory of the Barnase-Barstar simulations. The barnase (blue) and barstar (orange) proteins are shown in cartoon representation. 28 association events are registered starting from (a) dissociated state, and ending up in (b) associated complex.

composed by: native contacts, pair contacts, interface RMSD and a path CV. To easily compare the free energy profiles, we normalized these CVs between 0 and 1 (see Fig. 9.11).

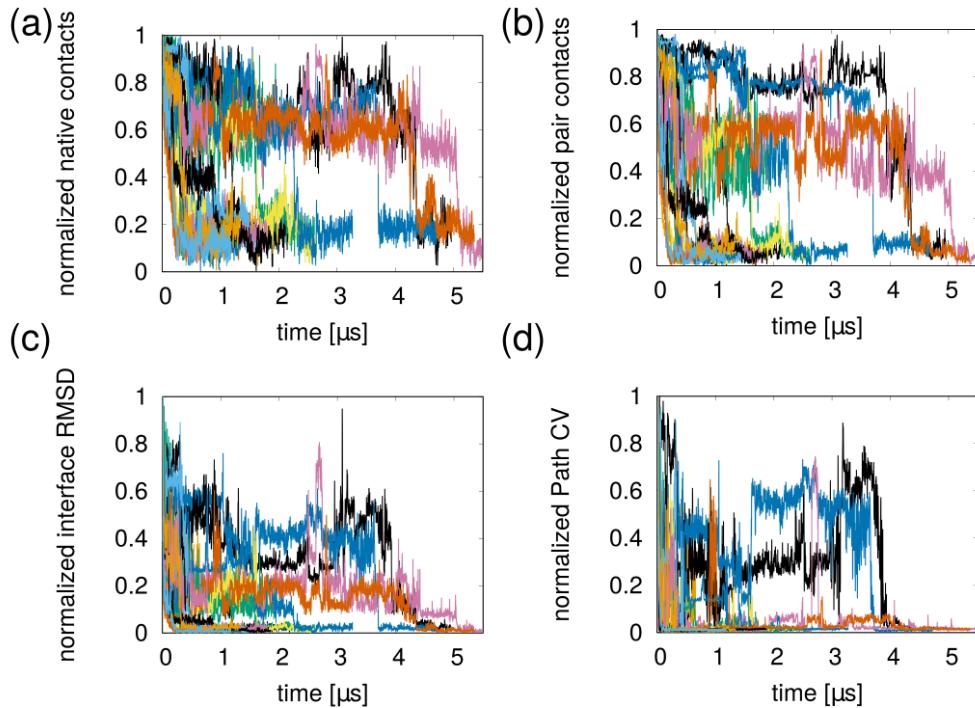


Figure 9.11: Barnase-barstar association trajectories projected over different CVs: (a) native contacts, (b) pair contacts, (c) interface RMSD, and (d) path CV. The CVs are normalized between 0 and 1 to facilitate the comparison of the free energy profiles below. Different colors represent different MD trajectories.

In Fig. 9.11 we reduced the complexity of the trajectories by concentrating only on the part of the MD simulation that leads to a successful association with the native complex. Once arriving at the native complex we also follow projected trajectories up to 1 μ s of the MD simulation within this state. Even with these considerations, in Fig. 9.11 we can see that all trajectories display metastable states between the dissociated state ($\text{CV} \rightarrow 1$) and

the associated state ($\text{CV} \rightarrow 0$).

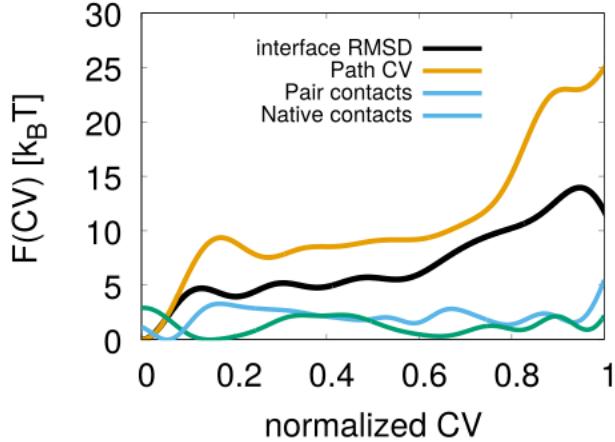


Figure 9.12: Free energy profiles from optimal Langevin models using as CVs (a) native contacts, (b) pair contacts, (c) interface RMSD, and (d) path CV . The thickness of the lines correspond to standard error over 10 independent optimization runs.

We use the trajectories in Fig. 9.11 as input data to optimize Langevin models and obtain free energy profiles (see Fig. 9.12). We remark that the input set contains only 28 association trajectories with a coarse time resolution between frames $\tau = 180$ ps. Therefore, we are pushing to the limit our optimization machinery to make the most out of the limited data. As mentioned above, typical protein complex dissociation processes are characterized by rough up-hill free energy landscapes with possible metastable states. Hence, it is difficult to pinpoint exactly where the free energy barrier between the associated and dissociated state is located. The experimental binding free energy is about $-19.0 \text{ kcal} \cdot \text{mol}^{-1} \approx 32 \text{ } k_B T$. In our Langevin model however, simple variables like native and pair contacts displays free energy barriers of less than $3 \text{ } k_B T$. A slightly more specific variable, the interface RMSD displays a barrier of about $13 \text{ } k_B T$. Finally, the path CV which was designed from reference frames to follow the dissociation process presents a free energy barrier of about $25 \text{ } k_B T$.

In previous chapters we have shown how depending on the CV in which the FES is projected, the free energy profile can be different and as sup-optimal CVs tend to underestimate the height of the free energy barrier. The results in Fig. 9.12 indicate that the same is true for the BB Langevin models. Regarding the shape of the free energy profiles, all of them seem to indicate the presence of metastable states between the associated and dissociated state (also visible from the trajectories in Fig. 9.11) that we have not yet characterized.

Finally, in Table 9.2 we show the dissociation rates from the different Langevin models obtained using Eq. 8.19. Here, we set the absorbing boundary $a = 0$, $q_0 = 0$ and $b = 1$. We expect the FF to give a rate relatively close to the experimental one, event though the trajectories provided by D.E. Shaw do not have an estimated rate in the article; such estimate is available for a slightly modified FF.²²⁸ Again, a trend with a slower rate is observed when improving the definition of the CV from total contacts to path CV . Although we are still far from the experimental value, these preliminary results indicate that the Langevin models

| Method | k_{off} (s ⁻¹) |
|----------------|---|
| Experimental | 8×10^{-6} |
| Path CV | $5.4 \times 10^{-2} \pm 1.2 \times 10^{-2}$ |
| Interface RMSD | $2.8 \times 10^2 \pm 1.1 \times 10^2$ |
| Pair contacts | $8.1 \times 10^5 \pm 2.9 \times 10^5$ |
| Total contacts | $2.9 \times 10^6 \pm 7.9 \times 10^6$ |

Table 9.2: Langevin model based estimated of the dissociation rate: Barnase-barstar interaction. The experimental value is reported in Ref. 231. The MFPT from the Langevin models suing the different CVs is estimated using Eq. 8.19.

are consistent and show a solid trend: increasing accuracy of the CV in tracing the detailed unbinding process corresponds to smaller rates, approaching closer and closer the expected value.

We are not surprised by the underestimation of the free energy barrier and the overestimation of the rate, considering the limited amount of input data and its nature (only association trajectories). On the contrary, we find very encouraging the fact that just by making an effort to improve the definition of CVs we are improving our thermodynamic and kinetic predictions by orders of magnitude. Therefore, our method to optimize Langevin models is emerging as a promising tool for post-analysis CV optimization, as we will discuss in Chapter 10.

In the near future, we are working on several ways to improve the results presented in this section: (i) we are going to characterize the different intermediate metastable states via cluster analysis and define the different association pathways followed by the trajectories. An isolation of the different pathways will allow us to improve the reaction coordinate used to study the process. (ii) we will try to include the role of water, as it likely gives an important contribution to the binding/unbinding kinetics.²³⁶ by generating more MD simulations of spontaneous association with a shorter time resolution to increase the input data set. (iii) As we are aware of the complexity of the PPI, in parallel we will work on the validation of the method in a smaller system but still challenging process, the tryptophan-cage folding. The last system is attractive since we have TPS trajectories provided by Peter Bolhuis (University of Amsterdam, Netherlands) and we can also easily generate MD trajectories of the entire folding process.

9.5 CONCLUSIONS AND OUTLOOK

Our results on different systems indicate that, whenever sufficient trajectory data is available together with a good definition of the reaction coordinate, overdamped Langevin models reproduce well the quantitative thermodynamic and dynamic properties of the original many-body system, including accurate kinetic rates, despite a gap of many orders of magnitude with respect to the short MD trajectories used for training. Incidentally, we note that our approach is not restricted to transition path sampling: Langevin models can be optimized

from any set of unbiased trajectories exploring the transition region (see example in Fig. 9.5 and Fig. 9.8).

In future applications of the new approach, two main issues have to be taken into consideration. First, different kinds of Langevin equation (generalized, standard or overdamped) can be necessary to faithfully reproduce the projected dynamics of a complex system, depending on the physical process, the choice of the **CV**, and the observational timescale. For chemical reactions in water the memory timescale could be comparable to the transition path time, requiring a non-Markovian equation, whereas for crystal nucleation or protein folding Markovian models are customarily invoked.

Finally, it should be noted that the last application in thesis shows that the overdamped Langevin models for the study of the last processes require the optimization of a good collective variable. However, obtaining a good **CV** in general is difficult. To address this issue, in the next chapter we will focus on challenge.

ACKNOWLEDGMENTS

The applications presented in this chapter were part of several collaborations: (*i*) The crystallization of Germanium-Telluride is being studied in collaboration with Julien Lam, Christoph Dellago, and Fabio Pietrucci. Input files for the **MD** simulations of Germanium-Telluride phase transition were provided by Julien Lam. (*ii*) The dimerization of the barnase-barstar complex is being studied in collaboration with Alessandro Barducci, Peter Bolhuis and Fabio Pietrucci. Barnase-barstar **MD** simulations were preformed by D. E. Shaw Research.

10

Collective variable optimization using data-driven Langevin models

10.1 INTRODUCTION

While collective variables (CVs) are ubiquitously used to model physico-chemical transformations, finding optimal reaction coordinates (RCs) that yield accurate thermodynamic and kinetic properties is a fundamental challenge in the field of atomistic simulations. As shown in the last two chapters, the dynamics of a system projected on a CV can be model through Langevin equations. We demonstrated that accurate free energies and rates can be extracted from transition path sampling trajectories projected on a good RC by optimizing overdamped Langevin models via likelihood maximization.¹⁹⁸ However, in the last chapter we showed that the quality of the reaction coordinate defines how accurately the Langevin model can reproduce the original dynamics. Here, we are proposing a computationally affordable method to optimize RCs in post-processing analysis. The approach is based on a variational principle stating that the dynamics projected on a CV yields a kinetic rate greater or equal to that from the full dynamics.²³⁷ Therefore, minimizing the kinetic rate of the effective dynamics from Langevin models yields the optimal RCs. Having as input a set of short molecular dynamics trajectories from aimless-shooting transition path sampling⁹⁵, we project the dynamics on a trial CV obtained as linear combination of a pool of physically based CVs. We optimize the RC via kinetic rate minimization using a Monte Carlo-Metropolis procedure. We test the validity of the algorithm on a benchmark two-dimensional double well potential. Then, we apply the method on the same system that we used for the validation of the methods presented in previous chapters: the interaction of a C₆₀ and C₂₄₀ fullerene dimer in water. We find optimal reaction coordinates simultaneously

with accurate kinetic rates for both systems.

This chapter is organized as follows: (*i*) we explain the algorithm developed to optimize the CVs and the theory behind it. (*ii*) We test the validity of the model by applying it on a 2D double-well potential with known free energy and diffusion coefficient. (*iii*) We apply the algorithm on a system of fullerene dimers in water. (*iv*) Finally, the conclusions and perspectives of this work in progress are presented.

10.2 THEORY AND METHODS

10.2.1 VARIATIONAL PRINCIPLE FOR RATE MINIMIZATION

Consider two states A and B whose full dynamics is reversible. The transition rate from state A to B is denoted by $k_{A \rightarrow B}$. The reaction rate of the effective dynamics between states \tilde{A} and \tilde{B} , defined in a lower dimensional space corresponding to a coordinate q , is $\tilde{k}_{\tilde{A} \rightarrow \tilde{B}}$. Zhang, *et al.*²³⁷ proved that the transition rate of the full dynamics is always less than or equal than the one computed using effective dynamics. In other words, the optimal reaction coordinate yields a minimal rate,

$$k_{A \rightarrow B} \leq \tilde{k}_{\tilde{A} \rightarrow \tilde{B}} \quad (10.1)$$

Here, we make use of this variational principle in Eq. 10.1 to optimize a RC based on the kinetic rate. This optimal RC is generally defined as a monotonic one-to-one function of the committor. We remark that using the committor we expect the dynamics to be successfully described by overdamped equation.¹⁶⁹ Moreover, when using the committor as RC, the rate constant is preserved by the effective dynamics:²³⁷

$$k_{A \rightarrow B} = \tilde{k}_{\tilde{A} \rightarrow \tilde{B}} \quad (10.2)$$

However, as explained in chapter 4 estimating the committor is very computationally expensive. Some approaches have been proposed for CV optimization aiming at committor optimization⁹⁵, even some recent approaches that use artificial intelligence.²³⁸ However, its cost remains prohibitive. Therefore, in our algorithm we propose an alternative procedure to committor optimization, based on the effective dynamics in a Langevin model, as we will explain below.

10.2.2 ALGORITHM FOR COLLECTIVE VARIABLE OPTIMIZATION

The main idea of the algorithm is to minimize the effective rate computed from optimal Langevin models as follows: starting from a set N of potentially relevant collective variables $\mathbf{q}(t) = \{q_1(t), q_2(t) \dots q_N(t)\}$, one generates a random linear combination of the set with different weights $q_{\text{initial}} = w_1 q_1(t) + w_2 q_2(t) + \dots + w_N q_N(t)$ with $\vec{w} = \{w_1, w_2 \dots w_N\}$. The weights of the linear combination are normalized such that $\|\vec{w}\| = 1$. The optimization is done using a Metropolis-Monte Carlo based algorithm. Thus, a trial move is proposed by adding small increments $\delta \vec{w} = \{\delta w_1, \delta w_2 \dots \delta w_N\}$ such that $q_{\text{trial}} = (w_1 + \delta w_1) q_1(t) + \dots + (w_N + \delta w_N) q_N(t)$. The

increments are randomly drawn from a uniform distribution between $[-0.05, 0.05]$. Then, one computes the free energy and the diffusion profiles from the likelihood maximization as well as the kinetic rate (Eq. 8.19) using the same procedure as in the last two chapters.¹⁹⁸ If the rate is reduced ($k_{\text{new}} < k_{\text{old}}$), the trial move is accepted and the increments are added to the initial weights.

Next, new increments are generated and new trial moves are tested until the process is repeated for a total number of iterations set by the user. Trial move configurations are rejected if they yield a larger rate, but some could be accepted based on a metropolis criterion,

$$P = \left(\frac{k_{\text{old}}}{k_{\text{new}}}\right)^{\alpha} \times \frac{\tau_{\text{noise old}}}{\tau_{\text{noise new}}}, \quad (10.3)$$

τ_{noise} is the characteristic time of the auto-correlation function of the effective noise (see Chapter 8, section 8.2.4). If P is less than a random number r uniform between 0 and 1, the move is rejected and the configuration is discarded and a new one is tested. In a nutshell, the algorithm does the following:

1. Read the input transition path trajectories
2. Construct an initial RC as a sum of the CVs with random weights
3. Randomly change the weight of the CV and find the free energy and diffusion profiles of the configuration from the optimal Langevin models using the likelihood maximization
4. Calculate the MFPT using the integral in equation 8.19.
5. Accept the configuration based on a Metropolis criterion
6. Repeat until N iterations are reached

The integration limits in Eq. 8.19 are automatically estimated: a is the minimum value of CV q , q_0 is the average value of CV q for all trajectories ending in the left basin, and b is the average value of CV q for all trajectories ending in the basin on the right. The biggest advantage of this algorithm is that the optimization of the CVs is done in post-analysis: the pool of CVs is flexible, since new CVs can be extracted from the trajectories without the need to run new expensive MD simulations.

10.2.3 SIMULATION METHODS

2D DOUBLE-WELL POTENTIAL

The benchmark 2D double-well free energy surface is defined as the sum two Gaussian functions:

$$F(x, y) = Ae^{-\mu\left[\frac{(x+a)*c-x_0)^2}{\sigma_x^2} + \frac{(y+a)*c-y_0)^2}{\sigma_y^2}\right]} + Be^{-\mu\left[\frac{(x+a)*c-x_1)^2}{\sigma_x^2} + \frac{(y+a)*c-y_1)^2}{\sigma_y^2}\right]}, \quad (10.4)$$

where A and B are the respective heights of the Gaussian functions, $A = B = -20 \text{ } k_BT$. $x_0 = 1.2$ and $y_0 = 2.0$, $x_1 = 2.8$ and $y_1 = 2.0$ are the coordinates of the Gaussian centers. The width of the Gaussian is $\sigma_x = \sigma_y = 0.5$, and $\mu = 0.5$. $a = 0.28$ and $c = 2.6$ are scaling

factors, so that $x \in [0, 1]$ and $y \in [0, 1]$, see Fig. 10.1a. The diffusion coefficient in this case was set to a constant value $D(x, y) = [0.015, 0.015]$

The MFPT was estimated from brute force simulations using the direct estimation method (see Chapter 3, section 3.2). For this purpose, 10 long overdamped Langevin dynamics simulations of $1 \mu s$, with a time step $\Delta t = 1 \times 10^{-4} ps$ each were performed. For the aggregate simulation time ($10 \mu s$), 16039 jumps from state A to B where observed, then we found a MFPT = $5600 ps$. The transition rate $k_{A \rightarrow B}$ was estimated as the inverse of the MFPT, $k_{A \rightarrow B} = 1.8 \times 10^{-4} ps^{-1}$.

The input trajectories for the optimization of the CVs were obtained by shooting 100 overdamped Langevin trajectories from the barrier top (see Chapter 8, section 8.2.3). From this trajectories, 51 relaxed to state A (left well) and 49 relaxed to state B (right well).

FULLERENE C₆₀ AND C₂₄₀ DIMERS

In the previous chapters we have used the interaction between the fullerene dimer in water solution to validate our methods. Here, we exploit the data already generated for these systems to test the CV optimization protocol. The description of these systems and the methodology used to generate the input trajectories can be found in section 9.2.2.

We used as input data the same set of a 100 aimless shooting trajectories generated in the previous chapter. These trajectories were projected on three different CVs: (i) the distance between the centers of mass of the fullerene molecules, defined in previous chapters, (ii) the carbon-carbon contacts (C-C contacts), and (iii) the water-carbon contacts (O-C contacts).

C-C contacts: the contacts between the carbon atoms of each fullerene were calculated as the coordination number between two groups

$$c = \sum_{i \in A} \sum_{j \in B} c_{ij}, \quad (10.5)$$

with

$$c_{ij} = \frac{1 - \left(\frac{r_{ij} - d_0}{r_0} \right)^n}{1 - \left(\frac{r_{ij} - d_0}{r_0} \right)^m}, \quad (10.6)$$

where $d_0 = 0$, $r_0 = 0.35$ nm, $n = 6$ and $m = 10$. In the sum of Eq. 10.5, the carbon atoms in one of the fullerene molecules of the dimer is group A , while the the carbon atoms in the other fullerene is group B .

O-C contacts: Analogous to c , the number of water-carbon contacts is defined as $w = \sum_{i \in A} \sum_{j \in B} w_{ij}$. We use Eq. 10.6 replacing c_{ij} by w_{ij} with $d_0 = 0$, $r_0 = 0.6$ nm, $n = 6$ and $m = 10$. Group A is the carbon atoms of the fullerene molecules while group B are the oxygen atoms of the water molecules in the simulation box.

In all cases, for each putative CV, the optimal Langevin model was optimized for 1×10^6 iteration steps, with a time resolution $\tau = 1$ ps, the latter was tested as sufficient to yield time-decorrelated noise when using the dimer center-of-mass distance as CV.

10.3 RESULTS AND DISCUSSION

10.3.1 BENCHMARK SYSTEM: DOUBLE-WELL POTENTIAL

To test the validity of the algorithm, we applied it to a theoretical benchmark model: a 2D double-well potential Eq. 10.4. We remark that in this case the input simulations correspond to overdamped Langevin dynamics. However the projected dynamics is not expected to be very strongly different from the original one, and for the optimal (x) coordinate we expected the overdamped model to be a good one. The optimal collective variable is x while y could be considered a bad variable as the projection on y , yields the wrong free energy landscape, that is, a single well instead of a double well (see Fig. 10.1). Moreover, a projection in y leads to trajectories that are not able to identify the state where the system ends up after relaxation (see Fig. 10.1b). Linear combinations of x and y yield to sub-optimal CVs, as shown in the central panels of Fig. 10.1.

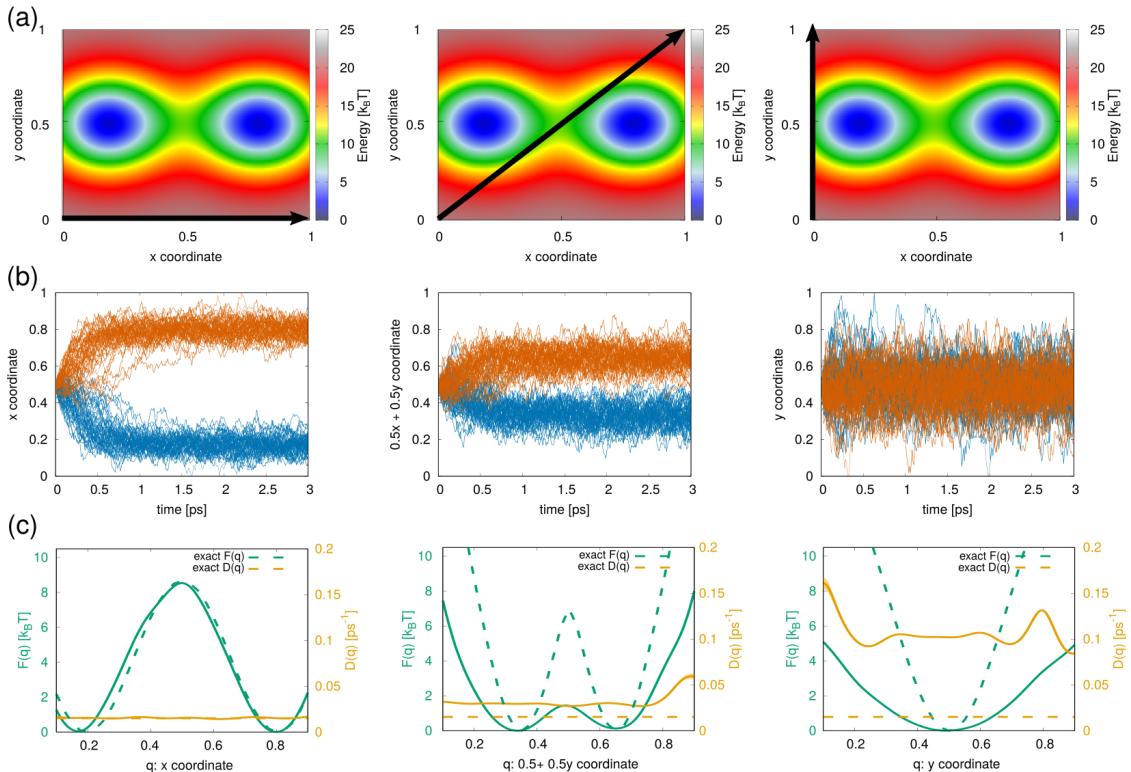


Figure 10.1: 2D double-well benchmark system for CV optimization. (a) 2D double-well free energy landscape used for the simulations. (b) Trajectories projected in the CVs indicated by the respective black arrows in panel (a). Trajectories ending up in state A (left basin) are colored in blue, while trajectories ending up in state B (right basin) are colored in orange. (c) Dashed lines: exact free energy and diffusion profiles for different CVs q . Solid lines: average free energy and diffusion profiles from 10 independent optimal Langevin models optimized using the trajectories in panel (b). The width of the solid lines correspond to the standard error.

When the trajectories are projected into the ideal CV x , the free energy and diffusion

profiles are accurately reconstructed by the optimal Langevin models, see left hand side in Fig. 10.1. In contrast, when non-ideal CVs are used (center and right hand sides in Fig. 10.1), not only is the free energy barrier underestimated with respect to the exact one of the ideal variable, but the free energy profiles of the optimal Langevin models tend to underestimate this barrier even more. At the same time, the diffusion profile is overestimated. We note that our Langevin model optimization method seeks to reproduce the dynamics of the input trajectories, “no matter the cost”, even if it must lose precision in $F(q)$ and $D(q)$ to do so. However, as the free energy barrier is still underestimated in this effective dynamics, the variational principle still holds. Our in-depth understanding of this “compensation” phenomenon between the free energy and the projected diffusion is still in progress.

After showing the effect of projection on non-ideal CVs, we can use the CV optimization algorithm. Starting from random linear combinations of the variables x and y , we optimize Langevin models to calculate the transition rate and use the variational principle for the MC optimization of the CV. We performed 10 CV optimizations of 300 MC iteration steps each, see Fig. 10.2. We can see how at the beginning of the optimization, some of the replicas start with a very high value of k , however, after about 150 iteration steps, all the replicas converge to the reference k value (up to statistical error). This variation in orders of magnitude shows the importance of using an optimal variable to accurately estimate the rate. At the last iteration the average optimal CV is $q = (0.98 \pm 0.12)x + (0.07 \pm 0.15)y$. This is a successful result, since the optimal CV corresponds well with the ideal CV $1x + 0y$.

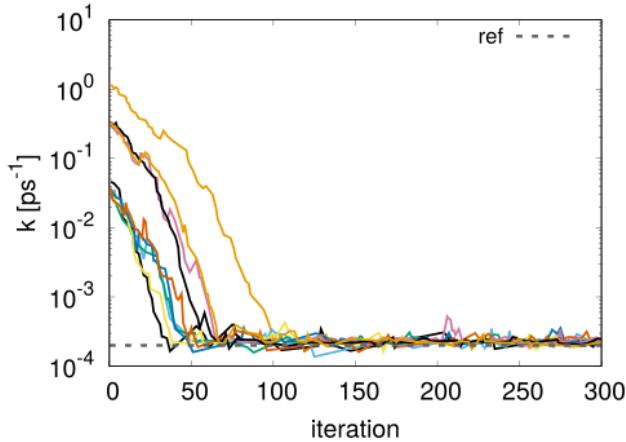


Figure 10.2: Kinetic rate from CV optimization: 2D double well potential. Transition rate k as a function of the number of optimization iterations. The dashed line corresponds to the reference rate estimated from brute force simulations. Colored lines correspond to 10 different CV optimizations starting from different random linear combinations of x and y .

Now we test the method on a set of all-atoms MD simulations, where, as explained in the last chapter, memory effects are expected upon projection on 1D, and the original dimensionality is $\sim 10^4$. Hence a more complex system becomes a more interesting challenge for CV optimization.

10.3.2 APPLICATION: FULLERENE DIMERS

In previous chapters, we have already discussed the advantages that the interaction of a fullerene dimer in water gives us to validate our methods. Here, we take as input data for the CV optimization the same trajectories used in the Chapter 9. We project these trajectories into different CVs: the distance between the centers of mass of each fullerene, the carbon-carbon contacts and the water-carbon contacts, as shown in Fig. 10.3.

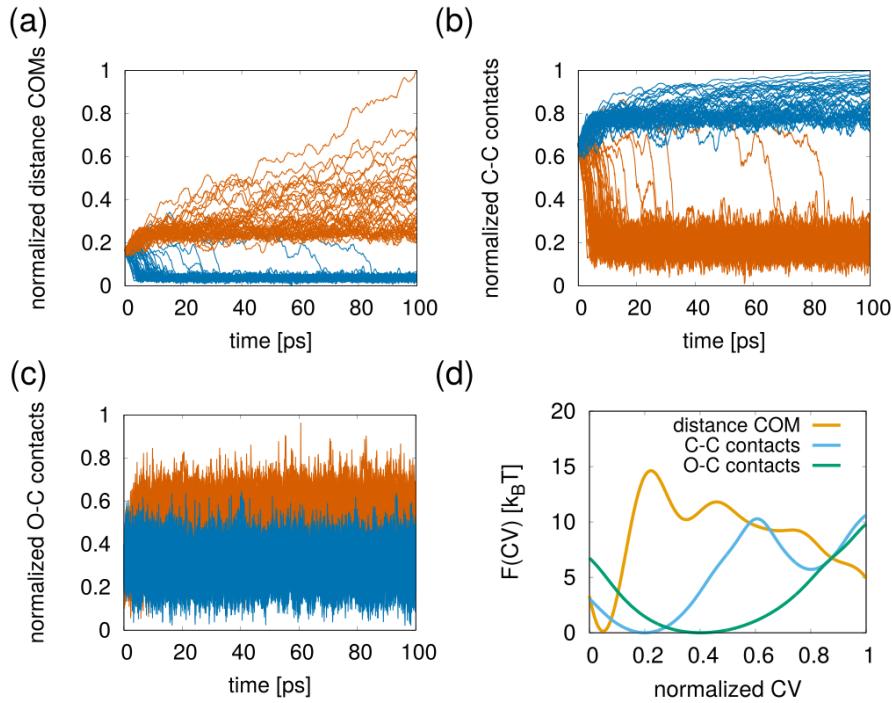


Figure 10.3: Application of CV optimization to fullerene dimers in water, using as example the data of the C_{240} fullerene dimer. (a), (b), and (c) Input trajectories for the CV optimization projected in the distance between the centers of mass, the carbon-carbon contacts and the water-carbon contacts, respectively. The values of the CVs are normalized between 0 and 1. Trajectories ending up in the associated state are colored in blue, and those ending up in the dissociated state are in orange. (d) Average free energy profiles from optimal Langevin models using the trajectories in panels (a), (b), and (c). 10 independent Langevin optimizations were performed. The standard error correspond to the width of the lines.

These variables will be the pool of CVs used for the linear combinations in the CV optimization algorithm. In Chapter 9 we show that the distance between the centers of mass captures well the dynamics of the dimer association and dissociation process. Here, we add another variable that depends on the distances of the carbon atoms, but this time in the form of a switching function: the number of C-C contacts between the two fullerenes. Additionally, since we are studying the interaction of fullerenes in water, one might think that a variable involving the atoms of the water molecules may be useful to describe the solvation shells between the two fullerene molecules that are involved in the association and dimer

dissociation. For this purpose, we define the contacts between the oxygen molecules of the water and the carbon atoms of the fullerenes (O-C contacts).

In Fig. 10.3 we take as an example the trajectories of the C_{240} fullerene dimer. The variables related to the distance between the carbon atoms (distance between centers of mass and C-C contacts) are able to distinguish between the associated and dissociated states. However, when we use these variables separately to optimize Langevin models, the C-C contacts display a lower free energy barrier, see Fig. 10.3d. On the other hand, the variable contacts between water and carbon atoms is not a good order parameter since it is barely capable of distinguishing between the two states, see Fig. 10.3c. This is reflected in the free energy profile of the optimal Langevin model, which presents a single minimum (Fig. 10.3d).

This pool of three CVs was entered as input for CV optimization. We note that, since Eq. 8.19 is written for a left to right transition, the values of the C-C contacts variable were inverted to obtain a rate consistent with the others. However, Eq. 8.19 is general and a simple modification to the integration limits allows the same equation to be used for inverse transitions. In Fig. 10.4, we present the dissociation rate constant of the C_{60} and C_{240} fullerene dimers during the CV optimization process.

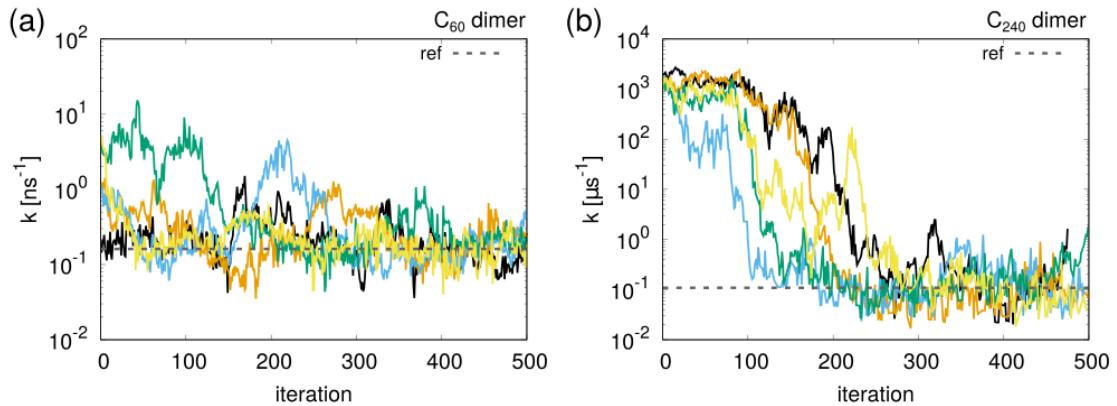


Figure 10.4: Kinetic rate from CV optimization: C_{60} and C_{240} fullerene dimers. Transition rate k as a function of the number of optimization iterations for (a) C_{60} and (b) C_{240} fullerene dimer. The dashed line corresponds to the reference rate estimated from brute force simulations (see Chapter 9). Colored lines correspond to 5 different CV optimizations starting from (a) different random linear combinations of the distance between the centers of mass, the carbon-carbon contacts and the water-carbon contacts, (b) a linear combination with the same weight for all variables.

When the CV formulation is changed there are three possible sources of rate reduction: the first is the improvement of the CV towards an optimal RC, that we seek. The second, however, is the spurious overestimation of the barrier height due to a worse CV having longer characteristic timescale for Markovianity (see fullerene results in chapter 9). The third is the overestimation or underestimation of the barrier due to a non-converged optimization of the Langevin model. We notice this effect when, to improve the computational performance of the CV optimization, we optimize the Langevin models only for 1×10^5 iterations. Clearly, we want to avoid the last two effects, to have a genuine optimization of the quality of the CV.

We therefore monitor the auto-correlation time of the “observed noise” at each optimization step, refusing CV changes that increase it and run longer optimizations of the Langevin models (1×10^6 iterations) to ensure convergence of the free energy and diffusion profiles.

Once we had these factor under control, we find how starting the CV optimization from different random linear combinations of the CVs or from a linear combination with the same weight for all variables, at convergence the rate fluctuates around the reference rate, showing that the algorithm is able to successfully optimize the transition rate and reaction coordinate. Interestingly, we note that the variable that includes water, the O-C contacts, is the CV that contributes the least to the optimal RC, while the distance between the centers of mass is the variable that has the greatest weight. However, a more detailed analysis of these results, as well as a more specific tuning of the algorithm is still needed in order to reduce the large fluctuations that we currently see in the converged regime.

10.4 CONCLUSIONS AND OUTLOOK

The reaction coordinate optimization problem is an open challenge in our field. Here, we introduce a computationally simple method for the optimization of reaction coordinates based on a kinetic rate minimization from optimal Langevin models. We show that the algorithm can successfully optimize the reaction coordinate based on an accurate estimate of the transition rate in the 2D double-well potential benchmark system. In addition, we found promising results for the C₆₀ and C₂₄₀ fullerene dimers in water. However, some improvements can still be implemented. For example, considering that the current implementation of the algorithm does not restrict the number of CVs to be used, a greater number of CVs can be introduced in the pool to increase the probability of including the relevant degrees of freedom for the transition process. Although we are initially testing physically motivated CVs, machine learning-based CVs could also be included in the optimization, as done recently using also pools of CVs to optimize reaction coordinates in ligand-protein unbinding.¹⁵⁴

Another important improvement could be the inclusion of non-linearity in the combination of CVs through artificial intelligence. Regarding computational cost, although the algorithm works in post-processing and does not require running new expensive MD simulations, the current bottleneck is in the optimization of the Langevin models in each iteration of the CV optimization. To improve the performance of this process and taking advantage of the formulation of the likelihood function, a parallel implementation of the latter is devised. We also want to investigate more in detail the relationship between the optimized CV from our algorithm and the committor.

Finally, considering that the Langevin model framework can be applied to a wide variety of physicochemical processes, as illustrated in chapter 9, the CV optimization algorithm presented here could also be applied to different systems. Taking into account all these possibilities, the improvement and extension of this reaction coordinate optimization algorithm gave rise to a new Ph.D. project that will be carried out by Line Mouaffac under the supervision of Fabio Pietrucci, at the Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie (IMPMC), Sorbonne Université.

ACKNOWLEDGMENTS

The method introduced in this chapter was developed in co-authorship with Line Mouaffac and Fabio Pietrucci. LM implemented the code for CV optimization and performed the simulations of the 2D double-well potential. We gratefully acknowledge very insightful discussions with Christoph Dellago, Gerhard Hummer, A. Marco Saitta, Hadrien Vroylandt, and Alessandro Barducci.

Part IV

Conclusions and Perspectives

11

Conclusions and perspectives

In this thesis we worked on moving forward the state of the art in the development of methods for kinetic rate prediction from molecular dynamics related methods. Among the biggest challenges in this field are: (i) the limitation of the time scales accessible with MD simulations that make it difficult to study rare events, (ii) the dependence of enhanced sampling methods for the calculation of rate constants on ideal collective variables, and (iii) the need for criteria to diagnose the reliability of the predictions. In order to address these issues, here we propose two alternative new methods for predicting transition rates. These theoretical methods were designed from the fundamental understanding of the dynamics modeled through MD and enhanced sampling simulations to recover the statistics of barrier crossing events from equilibrium or out-of-equilibrium trajectories in post-processing analysis and thinking of exploiting the parallel calculation capabilities of current computers, favoring the use of relatively small ensembles of short simulations (compared to the time scales of rare events) instead of long ergodic trajectories.

In chapter 6 we developed a new methodology based on Kramers' theory of activated barrier crossing to estimate escape rates starting from an ensemble of short simulations subject to a time-dependent bias. We took inspiration from the methods used by the force spectroscopy community that address single molecule experiments, and we significantly improve over a related approach, infrequent metadynamics.

Our results both on benchmark systems (chapter 6) and on all-atom MD simulations (chapter 7) indicate that the new approach simultaneously provides rates (even when sub-optimal CVs are biased) from simulations orders of magnitude shorter than the average waiting time of the processes, as well as a quantitative estimation of the quality of a CV. This latter result opens an interesting perspective of automatic CV-optimization using, for instance, machine learning approaches with the *quality-of-CV* parameter as loss function.

We have used examples from MetaD simulations, however, the KTR method is general for any time-dependent biased simulation along some CVs, such as, adaptive biasing force or adaptively biased MD, among others. Methods restricting the bias filling level could also be used within the KTR framework, which would have the advantage of simplifying bookkeeping of the bias potential and survival probability. Additionally, our method provides a novel measure for the effectiveness of the added bias to accelerate the transition (related to the efficiency of the CVs). This overcomes severe limitations encountered with previous approaches where the bias had to be deposited very infrequently over ideal CVs.

There are several points of the KTR theory that can be improved in future work. Automatized methods to determine when the barrier-crossing occurs might be helpful. Also, it would be useful to extrapolate the theory KTR to non-equilibrium conditions, as we have shown that Kramers' theory and the quasi-adiabatic assumption breakdown in the over-biasing regime. Finally, the main limitation of the current approach, besides force-field inaccuracy, appears to be the implicit approximation of considering the transition process as the escape from a single "reactants" well to a single "products" well: a generalization to multiple-basins and dimensions, as was done for force-spectroscopy to take explicitly into account a network of possibly multiple transition pathways across several metastable states would be very convenient to account for the multi-state dynamics in complex simulations.

In chapter 8 we develop and apply a second methodological framework aimed to the estimation of accurate free-energy landscapes, kinetic rates and also to the automatic optimization of reaction coordinates. We adopt a scheme based on Langevin equations modeling the dynamics projected on a CV: such equations bear a physical interpretation in terms of average force and noise that is lacking in more abstract data driven approaches like Markov state models or machine learning-based predictors.

As a first important result, our work demonstrates the possibility to obtain simultaneously free energies and kinetic rates in a conceptually simple way, by optimizing overdamped Langevin models via maximum likelihood starting from about one hundred path sampling trajectories, regardless of the barrier height. We remark that two tests are also presented that allow to identify the optimal time resolution for the construction of the stochastic model: this is a crucial point to allow the robust application to systems whose behavior is not known in advance.

Overdamped Langevin models reproduce well the quantitative thermodynamic and dynamic properties of the original many-body system, including accurate kinetic rates, despite a gap of many orders of magnitude with respect to the short MD trajectories used for training. Incidentally, we note that our approach is not restricted to transition path sampling: Langevin models can be optimized from any set of unbiased trajectories exploring the transition region.

In future work, this approach to optimize Langevin models may be generalized to different kinds of Langevin equations (generalized or standard) allowing its use to be extended to applications such as chemical reactions in solution. Until now, the approach is restricted to one dimensional Langevin models: taking into account the complexity of the physicochemical processes and how difficult it is to find a RC that describes them, an extension of the approach

to more dimensions, which allows the simultaneous use of several CVs, could also be useful.

Having opened the possibility to obtain faithful Langevin models of the projected dynamics along different CVs, a natural direction of research, that we pursued in chapter 10, consists in performing a systematic variation of the CV definition to obtain an optimal reaction coordinate. As a variational principle guiding the optimization algorithm, we systematically reduce the transition rate: direct estimation of the latter by numerical quadrature is an important ingredient for numerical efficiency.

Considering a trial CV formed by a linear combination of a pool of putative CVs, as a proof of concept we demonstrate that the ideal reaction coordinate is quickly and reliably recovered in a simple 2D benchmark system. Remarkably, our results show that a maximum-likelihood Langevin model along a non-ideal CV tends to underestimate the true free-energy barrier along the same direction: this phenomenon does not limit the efficiency of the CV-optimization algorithm, however, since it preserves the variational principle.

A first application to the dissociation of a fullerene dimer in water confirms the ability of the new method to yield at the same time an optimal reaction coordinate and an accurate estimation of the transition rate. However, work is in progress to extend the study to a large pool of putative CVs (whose construction is an interesting problem per se) and to different systems. Another important improvement could be the inclusion of non-linearity in the combination of CVs through artificial intelligence and the improvement of the computational efficiency of the numerical codes used for Langevin models optimization.

Here we have provided a summary of the most important theoretical advances and results from this thesis. Both methods are alternative pathways with a common goal: kinetic rate prediction. The possibility of approaching the problem of kinetics from different perspectives demonstrates its complexity. The use of one of the methods proposed or the other will depend on the available data and the application at hand. This thesis opens up a plethora of possibilities from the exploitation of the wealth of information available in transition path sampling trajectories, or the possibility of recovering useful information from MetaD simulations with bias over non-ideal CVs, to the use of these approaches in the optimization of collective variables. As the atomistic computer simulation community moves further and further into the study of kinetics, we are looking forward to any improvements and applications that our new approaches may have.

Part V

Appendix

A

Supplementary information: kinetics from metadynamics

SUPPLEMENTARY TEXT

ANALYTICAL $S(t)$ FOR A LOGARITHMIC AND LINEAR TIME-DEPENDENT BIAS

If the bias presents a logarithmic time dependence, $V_{\text{MB}}(t) = a \log(1 + b t)$ (*e.g.*, orange line in Fig. 1b), then the survival probability can be analytically calculated using Eq. 6.15

$$S(t) = \exp \left(\frac{k_0}{b(\beta\gamma a + 1)} \left(1 - (1 + b t)^{\beta\gamma a + 1} \right) \right). \quad (\text{A.1})$$

where γ and k_0 are the efficiency of the CV and intrinsic transition rate, respectively. If there is a linear time-dependence of the bias $V_{\text{MB}}(t) = a t$, and using 6.15, the analytical expression for the survival probability is

$$S(t) = \exp \left(\frac{k}{\beta\gamma a} \left(1 - e^{\beta\gamma a t} \right) \right). \quad (\text{A.2})$$

These analytic expressions of $S(t)$ can be then used within the likelihood expression (Eq. 6.16) to extract the relevant parameters.

EXPERIMENTAL UNBINDING RATE CDK2-CS3 COMPLEX

In the unbinding simulations of the CS3 ligand and the CDK2 protein we use as a reference value for comparison the experimental unbinding rate of this complex reported by Dunbar

Jr, *et al.*, in ref. 162. The authors measured separately the unbinding rate of the CS3 ligand from the protein CDK2 and from the CDK2-CyclinA complex. In the Supporting Information of ref. 162 only the values of CDK2-CyclinA k_{off} are reported, with the CS3 ligand having an off-rate of 0.061 s^{-1} . However, they specify in the Main Text that all the data is available in: <http://www.csardock.org/> (see dataset 2: 2012 CSAR Exercise, subfolder CDK2_data/CDK2_Master Kd sheet .xlsx). In this dataset, the CS3 ligand's k_{off} from CDK2 (without CyclinA) is reported as 0.259 s^{-1} . Because the MD simulations have been performed only with CDK2 (no CyclinA), we compared our results with the latter value.

SUPPLEMENTARY FIGURES

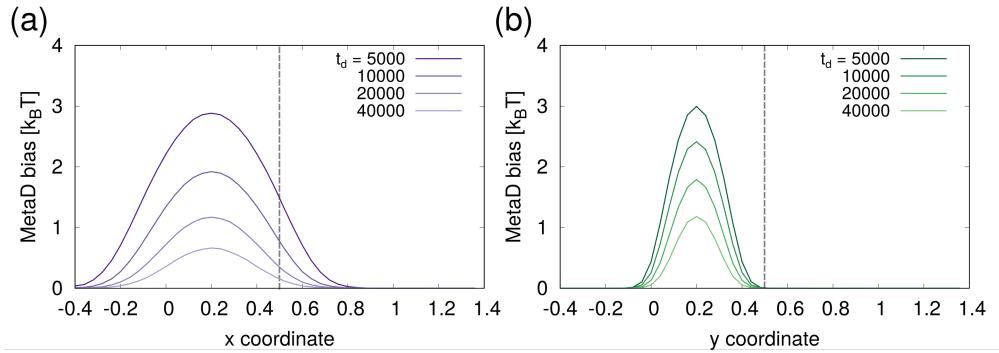


Figure A1: Average accumulated MetaD bias for simulations with different bias deposition time. a) MetaD bias is added over the x coordinate; b) MetaD bias is added over the y coordinate. The position of the TS corresponds to the gray dashed line. We observe that when bias is deposited along the x coordinate, a large quantity of bias is deposited over the TS region. On the other hand, by depositing bias along the y coordinate, no bias is deposited in the TS region.

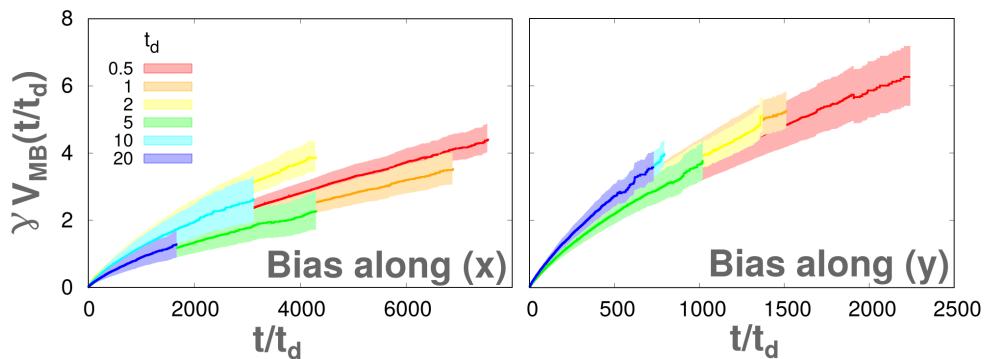


Figure A2: γV_{MB} as a function of the normalized time t/t_d for the 2D double-well simulations. γV_{MB} as a function of the time normalized by the bias deposition time (t/t_d) for the different simulation setups basing the x -coordinate (left) and y -coordinate (right). Red to blue lines go from frequent to infrequent t_d .

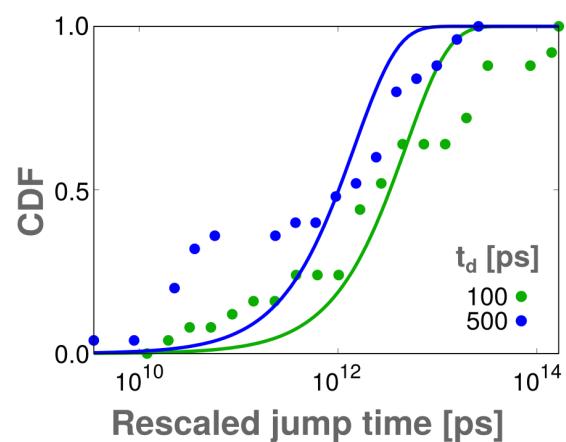


Figure A3: iMetaD results for the CDK2. Cumulative distribution function (CDF) of the iMetaD rescaled times from the same CDK2 simulations as shown in the Main Text with $t_d=100$ and 500 ps. Attempted fits using a Poisson distribution are shown as solid lines. 90% and 74%, respectively, of the bootstrap samples failed the KS-test.

B

Supplementary information: kinetics from Langevin models

SUPPLEMENTARY TABLES

| Method | N. traj | MFPT (ns) |
|-------------|---------|------------------|
| Brute force | 100 | 18.2 ± 0.1 |
| | 1000 | 90 ± 69 |
| | 2000 | 25 ± 4 |
| | 5000 | 27 ± 2 |
| | 8000 | 20.8 ± 0.3 |
| | 10000 | 19.1 ± 0.1 |
| | 12000 | 19.07 ± 0.09 |
| | 15000 | 18.24 ± 0.05 |
| | 18000 | 18.17 ± 0.04 |
| | 20000 | 18.20 ± 0.07 |

Table B1: Brute force vs reactive flux estimate of the MFPT. We used as a benchmark system a double-well potential with a $7 k_B T$ barrier and $50 \leq \gamma(q) \leq 150$ ps. Brute force estimate is obtained as the average escape time of 100 trajectories (at least 50 ns-long) starting from the minimum. Reactive flux estimate is obtained with Eq. 8.18, and different number of trajectories. Each trajectory is 40 ps-long. We note that at least 15000 trajectories are needed to converge the reactive flux estimate. We highlight that in practical applications, the need of large amount of trajectories to converge the correlation function of Eq. 8.18 restricts the use of reactive flux.

SUPPLEMENTARY FIGURES

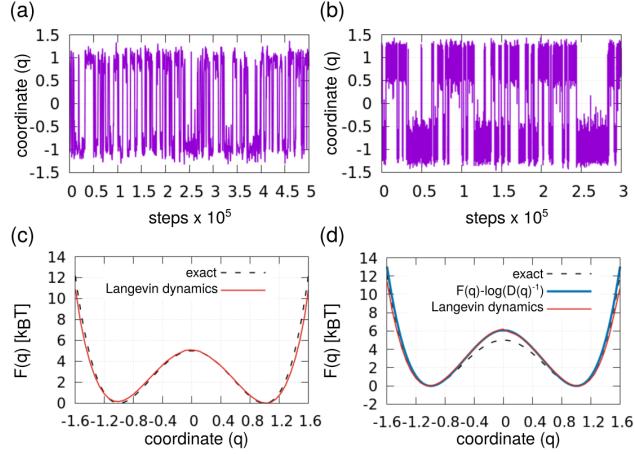


Figure B1: Effect of the inclusion of $D(q)$ in the OLE. We performed long Langevin dynamics simulations over a 1D double-well model with a $5 k_BT$ free energy barrier ($m = 1 k_BT \cdot \text{ps}^{-2}$), using Eq. 8.4, *i.e.*, ignoring the term $\frac{\partial D(q)}{\partial q}$ in Eq. 8.5. (left) Double-well with constant $D(q)$. (right) Double-well with q -dependent $D(q) = (25(x+1)^2(x-1)^2(x-2)(x+2) + 150)^{-1}$. (a), (b) Langevin dynamics trajectories from double-well with and without constant $D(q)$, respectively. (c), (d) Free energy profiles recover from the population histogram for constant and non-constant $D(q)$, respectively. The dashed lines correspond to the exact double-well used to perform the Langevin dynamics simulations. The solid red lines correspond to the population histogram. The blue line in panel (c) shows the exact result expected by neglecting the q -dependent $D(q)$ in Eq. 8.5. Since we miss the extra “ D ” force, the equilibrium distribution should be $\rho = \frac{C}{D(q)} e^{-\beta F} = D(q)^{-1} e^{-\beta F}$, so the recovered $F(q)$ is in this case $-kT \log \rho = F(q) - \log(D(q)^{-1})$. Panel (d) illustrates the importance of including $\frac{\partial D(q)}{\partial q}$ in Eq. 8.5, without this term, in this case, the free energy barrier is overestimated about $1 k_BT$.

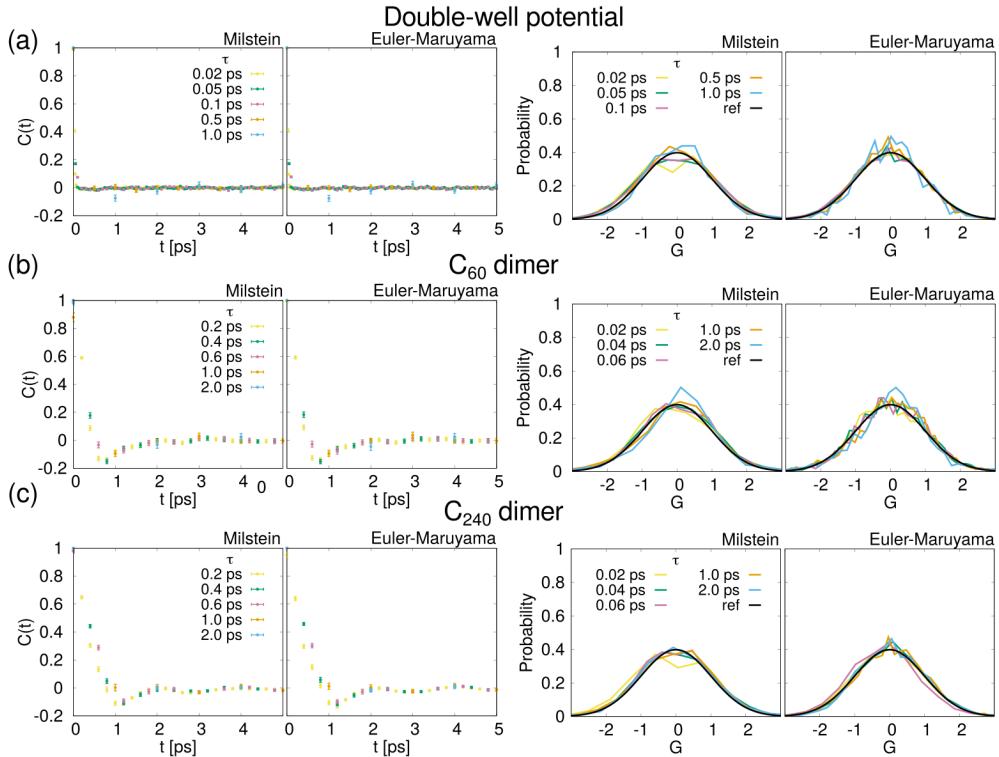


Figure B2: Autocorrelation function $C(t)$ (left panels) and distribution of the effective noise G (right panels) for optimal Langevin models of different systems. The expression used to estimate the noise is found inverting the Euler-Maruyama, or the Milstein integrator, see the Methods section C, in the Main Text. (a) Symmetric double-well potential (non-overdamped input trajectories) with $10 k_B T$ barrier and diffusion coefficient $0.01 \leq D(q) \leq 0.02 \text{ ps}^{-1}$ (see Fig. 8.3 and Fig. 8.4). (b) C₆₀ fullerene dimer in solution. (c) C₂₄₀ fullerene dimer in solution. The different time resolutions τ used to optimize the overdamped Langevin models are shown in colors. The expected zero-mean, unit-variance Gaussian distribution of the noise is shown in black lines on the right panels. Error bars correspond to standard error over all the trajectories data-points.

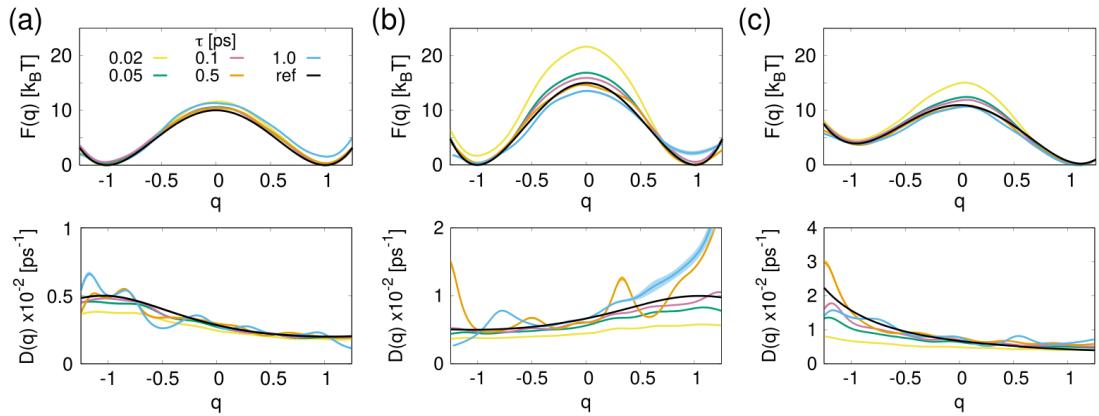


Figure B3: Optimal overdamped Langevin models, each one trained on 100 non-overdamped Langevin trajectories of 50 ps relaxing from the barrier top. (a) Symmetric double-well potential with $10 k_B T$ barrier and diffusion coefficient $0.002 \leq D(q) \leq 0.005 \text{ ps}^{-1}$. (b) Symmetric double-well potential with $15 k_B T$ barrier and diffusion coefficient $0.005 \leq D(q) \leq 0.01 \text{ ps}^{-1}$. (c) Asymmetric double-well potential with $10 k_B T$ barrier and diffusion coefficient $0.005 \leq D(q) \leq 0.03 \text{ ps}^{-1}$. The different time resolutions τ employed to optimize the models are shown in colors. The exact profiles are shown with black lines. The thickness of the lines correspond to standard error over 10 independent stochastic optimization runs.

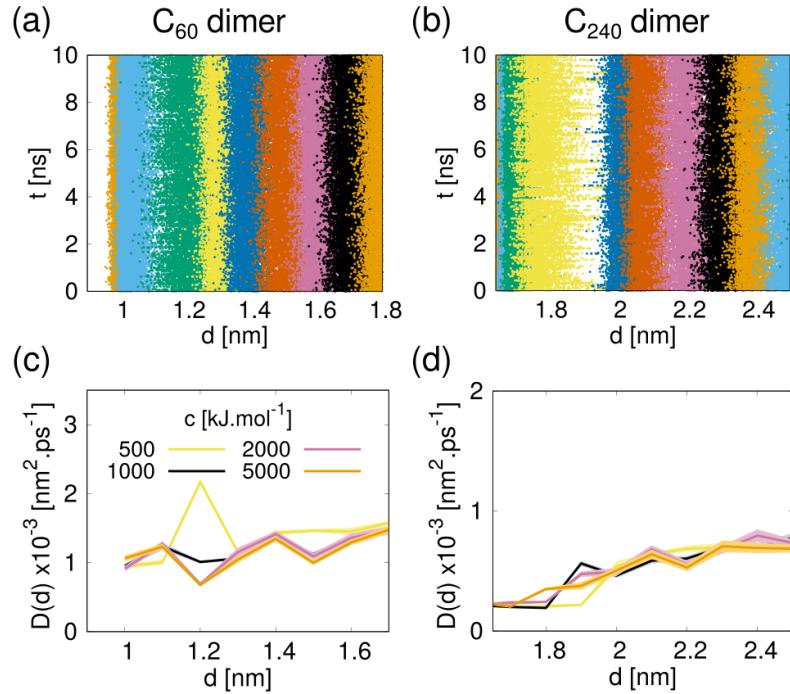


Figure B4: Diffusion profiles from umbrella sampling. Trajectory on each window for (a) C_{60} fullerene dimer and (b) C_{240} fullerene dimer using a spring constant $c = 1000 \text{ kJ} \cdot \text{mol}^{-1}$. Note that for the calculation of the diffusion coefficient we do not required converged umbrella sampling calculations. Diffusion profiles $D(d)$ for different values of the spring constant c for (c) C_{60} fullerene dimer and (d) C_{240} fullerene dimer.

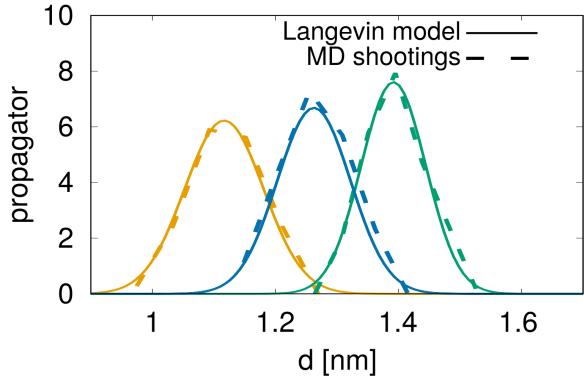


Figure B5: Propagator test. Comparison between $p(q', t + \tau | q, t)$ predicted by the optimal Langevin model and a distribution of MD shootings trajectories for the C_{60} fullerene dimer. The time resolution was set to $\tau = 1$ ps. The optimized free energy and diffusion profiles (Fig. 9.3, orange profiles) were used to estimate the short time transition probability (Eq. 8.10 predicted by the Langevin model (solid lines). The distribution from MD correspond to the histograms of the displacement after 1 ps from sets of 500 shooting trajectories (dashed lines). We use 5 shooting points extracted from a long unbiased trajectory for each location: before the barrier (orange), at the barrier top (blue), and after the barrier (green).

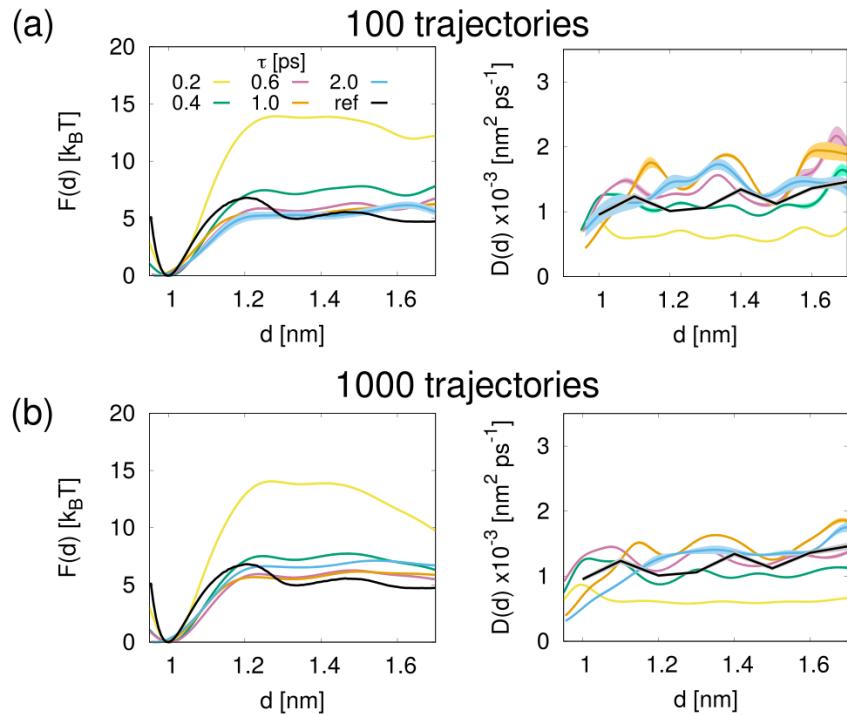


Figure B6: Effect of the number of input trajectories. Optimal free energy and diffusion profiles reconstructed from (a) 100 and (b) 1000 initial MD trajectories of 20 ps for the C_{60} fullerene dimer. The different time resolutions τ employed to optimize the models are shown in colors. The reference free energy profile (computed from brute force MD) and diffusion profile (computed from umbrella sampling) are shown with black lines. The thickness of the lines correspond to standard error over 10 independent optimization runs.

C

Dissemination of research results

PUBLISHED PAPERS

- Palacio-Rodriguez, K., Vroylandt, H., Stelzl, L.S., Pietrucci, F., Hummer, G., Cossio, P. Transition Rates and Efficiency of Collective Variables from Time-Dependent Biased Simulations. *The Journal of Physical Chemistry Letters*. In press.
DOI: <https://doi.org/10.1021/acs.jpclett.2c01807>
- Palacio-Rodriguez, K. & Pietrucci, F. Free Energy Landscapes, Diffusion Coefficients, and Kinetic Rates from Transition Paths. *Journal of Chemical Theory and Computation*. In press. DOI: <https://doi.org/10.1021/acs.jctc.2c00324>

PUBLICATIONS IN PREPARATION

- Mouaffac, L., Palacio-Rodriguez, K., & Pietrucci, F. Optimal reaction coordinates from effective dynamics of transition paths. *In prep.*
- Palacio-Rodriguez, K., Brotzakis, F., Bolhuis, P., Barducci, A., & Pietrucci, F. Assessing free energy landscapes and kinetic rates for protein-protein interactions via Langevin models. *In prep.*
- Lam, J., Palacio-Rodriguez, K., Dellago, C., & Pietrucci, F. Prediction of crystallization rates from short molecular dynamics trajectories. *In prep.*
- Palacio-Rodriguez, K., Paloni, M., Barducci, A., & Pietrucci, F. Unraveling the interaction between folded proteins and membraneless organelles with coarse-grained dynamics. *In prep.*

PUBLICATIONS IN COLLABORATION

- Ochoa, R., Palacio-Rodriguez, K., Clemente, C. M., & Adler, N. S. (2021). dockECR: Open consensus docking and ranking protocol for virtual screening of small molecules. *Journal of Molecular Graphics and Modelling*, 109, 108023.
- Giraldo-Barreto, J., Ortiz, S., Thiede, E. H., Palacio-Rodriguez, K., Carpenter, B., Barnett, A. H., & Cossio, P. (2021). A Bayesian approach to extracting free-energy profiles from cryo-electron microscopy experiments. *Scientific Reports*, 11(1), 1-15.
- Lans, I., Palacio-Rodríguez, K., Cavasotto, C. N., & Cossio, P. (2020). Flexi-pharma: a molecule-ranking strategy for virtual screening using pharmacophores from ligand-free conformational ensembles. *Journal of computer-aided molecular design*, 34(10), 1063-1077.
- Lans, I., Anoz-Carbonell, E., Palacio-Rodríguez, K., Aínsa, J. A., Medina, M., & Cossio, P. (2020). In silico discovery and biological validation of ligands of FAD synthase, a promising new antimicrobial target. *PLoS computational biology*, 16(8), e1007898.

PARTICIPATION TO CONFERENCES

Invited talk

- Data-driven Langevin equations from transition path sampling trajectories. *Memory effects in dynamical processes*. 23-25 June, (2021). CECAM-ESI, Online.

Contributed talk

- Transition rates and efficiency of collective variables from time-dependent biased simulations. *20 years of Metadynamics*. 5-8 September, (2022). Lausanne, Switzerland.
- Development of predictive approaches for biomolecular association kinetics. *Biophysics at the Dawn of Exascale Computers*. 16-20 May, (2022). Hamburg, Germany.

Poster

- Development of predictive approaches for molecular association kinetics. *Gordon Research Conference: Multiscale Modeling of Complex Systems: Methods and Applications*. 17-22 July, (2022). Barcelona, Spain.
- Strategies to predict kinetic rates and quality of collective variables. *Chasing CVs using Machine Learning: from methods development to biophysical applications*. 28-30 June, (2022). Paris, France. **Poster Award Winner**
- Data-driven Langevin equations from transition paths. *Atomistic simulations of interfacial processes in energy materials*. 13-15 June, (2022). Paris, France.

- Free energy landscapes and kinetic rates from transition paths. *Paris International School on Advanced Computational Materials*. 29 Aug-5 Sept, (2021). Paris, France. **Poster Award Winner**
- Kinetic rates from time-dependent bias potentials. *Summer School on Biomolecular Simulations*. 7-11 June, (2021). BioExcel, Online.
- Free energy landscapes, diffusion coefficients and kinetic rates from transition paths. *CECAM Mixed-Gen series, session 1 on activated events*. 28 January, (2021). CECAM, Online.

GitHub repositories

- Free Energy Landscapes, Diffusion Coefficients, and Kinetic Rates from Transition Paths: <https://github.com/physix-repo/optLE>
- Transition Rates and Efficiency of Collective Variables from Time-dependent Biased Simulations: <https://github.com/kpalaciorodr/KTR>

Part VI

References

References

- [1] Fabio Pietrucci. Strategies for the exploration of free energy landscapes: Unity in diversity and challenges ahead. *Reviews in Physics*, 2:32–45, 2017.
- [2] Carlo Camilloni and Fabio Pietrucci. Advanced simulation techniques for the thermodynamic and kinetic characterization of biological systems. *Advances in Physics: X*, 3(1):1477531, 2018.
- [3] Joerg Holenz and Patrick Stoy. Advances in lead generation. *Bioorganic & medicinal chemistry letters*, 29(4):517–524, 2019.
- [4] Didier Rognan. The impact of in silico screening in the discovery of novel and safer drug candidates. *Pharmacology & therapeutics*, 175:47–66, 2017.
- [5] Tanaji T Talele, Santosh A Khedkar, and Alan C Rigby. Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. *Current topics in medicinal chemistry*, 10(1):127–141, 2010.
- [6] William L Jorgensen. The many roles of computation in drug discovery. *Science*, 303(5665):1813–1818, 2004.
- [7] Doris A Schuetz, Wilhelmus Egbertus Arnout de Witte, Yin Cheong Wong, Bernhard Knasmueller, Lars Richter, Daria B Kokh, S Kashif Sadiq, Reggie Bosma, Indira Nederpelt, Laura H Heitman, et al. Kinetics for drug discovery: an industry-driven effort to target drug residence time. *Drug Discovery Today*, 22(6):896–911, 2017.
- [8] David C Swinney. The role of binding kinetics in therapeutically useful drug action. *Current opinion in drug discovery & development*, 12(1):31–39, 2009.
- [9] Albert C Pan, David W Borhani, Ron O Dror, and David E Shaw. Molecular determinants of drug–receptor binding kinetics. *Drug discovery today*, 18(13-14):667–673, 2013.
- [10] Zhiye Tang, Si-Han Chen, and Chia-En A Chang. Transient states and barriers from molecular simulations and the milestones theory: Kinetics in ligand–protein recognition and compound design. *Journal of chemical theory and computation*, 16(3):1882–1895, 2020.

- [11] Scott A Hollingsworth and Ron O Dror. Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143, 2018.
- [12] Aravindh Ganesan, Michelle L Coote, and Khaled Barakat. Molecular dynamics-driven drug discovery: leaping forward with confidence. *Drug discovery today*, 22(2):249–269, 2017.
- [13] Ben Leimkuhler and Charles Matthews. *Molecular dynamics*, volume 36. Springer, 2015.
- [14] Frank Noé and Cecilia Clementi. Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. *Current opinion in structural biology*, 43:141–147, 2017.
- [15] Daniel T Gillespie. The multivariate langevin and fokker–planck equations. *American Journal of Physics*, 64(10):1246–1257, 1996.
- [16] Robert Zwanzig. *Nonequilibrium statistical mechanics*. Oxford University Press, 2001.
- [17] William Coffey and Yu P Kalmykov. *The Langevin equation: with applications to stochastic problems in physics, chemistry and electrical engineering*. World Scientific, 2012.
- [18] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D Chodera, Christof Schütte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *The Journal of chemical physics*, 134(17):174105, 2011.
- [19] Brooke E Husic and Vijay S Pande. Markov state models: From an art to a science. *Journal of the American Chemical Society*, 140(7):2386–2396, 2018.
- [20] Mattia Bernetti, Matteo Masetti, Walter Rocchia, and Andrea Cavalli. Kinetics of drug binding and residence time. *Annual review of physical chemistry*, 70:143–171, 2019.
- [21] Marco De Vivo, Matteo Masetti, Giovanni Bottegoni, and Andrea Cavalli. Role of molecular dynamics and related methods in drug discovery. *Journal of medicinal chemistry*, 59(9):4035–4061, 2016.
- [22] Alex Dickson, Pratyush Tiwary, and Harish Vashisth. Kinetics of ligand binding through advanced computational approaches: a review. *Current topics in medicinal chemistry*, 17(23):2626–2641, 2017.
- [23] Joao Marcelo Lamim Ribeiro, Sun-Ting Tsai, Debabrata Pramanik, Yihang Wang, and Pratyush Tiwary. Kinetics of ligand–protein dissociation from all-atom simulations: Are we there yet? *Biochemistry*, 58(3):156–165, 2018.

- [24] Rafael C Bernardi, Marcelo CR Melo, and Klaus Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1850(5):872–877, 2015.
- [25] Samia Aci-Sèche, Sonia Ziada, Abdennour Braka, Rohit Arora, and Pascal Bonnet. Advanced molecular dynamics simulation methods for kinase drug discovery. *Future medicinal chemistry*, 8(5):545–566, 2016.
- [26] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.
- [27] Andrea Cavalli, Andrea Spitaleri, Giorgio Saladino, and Francesco L Gervasio. Investigating drug–target association and dissociation mechanisms using metadynamics-based algorithms. *Accounts of chemical research*, 48(2):277–285, 2015.
- [28] Alessandro Laio and Francesco L Gervasio. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics*, 71(12):126601, 2008.
- [29] Fabrizio Marinelli, Fabio Pietrucci, Alessandro Laio, and Stefano Piana. A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations. *PLoS computational biology*, 5(8):e1000452, 2009.
- [30] Gerhard Hummer. Position-dependent diffusion coefficients and free energies from bayesian analysis of equilibrium and replica molecular dynamics simulations. *New Journal of Physics*, 7(1):34, 2005.
- [31] Luca Donati and Bettina G Keller. Girsanov reweighting for metadynamics simulations. *The Journal of chemical physics*, 149(7):072335, 2018.
- [32] Pratyush Tiwary and Michele Parrinello. From metadynamics to dynamics. *Physical review letters*, 111(23):230602, 2013.
- [33] Fabio Pietrucci, Fabrizio Marinelli, Paolo Carloni, and Alessandro Laio. Substrate binding mechanism of hiv-1 protease from explicit-solvent atomistic simulations. *Journal of the American Chemical Society*, 131(33):11811–11818, 2009.
- [34] Pratyush Tiwary, Jagannath Mondal, and Bruce J Berne. How and when does an anticancer drug leave its binding site? *Science advances*, 3(5):e1700014, 2017.
- [35] Huiyong Sun, Youyong Li, Mingyun Shen, Dan Li, Yu Kang, and Tingjun Hou. Characterizing drug–target residence time with metadynamics: How to achieve dissociation rate efficiently without losing accuracy against time-consuming approaches. *Journal of Chemical Information and Modeling*, 57(8):1895–1906, 2017.

- [36] Peter G Bolhuis, David Chandler, Christoph Dellago, and Phillip L Geissler. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.*, 53(1):291–318, 2002.
- [37] Peter G Bolhuis and David WH Swenson. Transition path sampling as markov chain monte carlo of trajectories: Recent algorithms, software, applications, and future outlook. *Advanced Theory and Simulations*, 4(4):2000237, 2021.
- [38] Titus S Van Erp, Daniele Moroni, and Peter G Bolhuis. A novel path sampling method for the calculation of rate constants. *The Journal of chemical physics*, 118(17):7762–7774, 2003.
- [39] Rosalind J Allen, Patrick B Warren, and Pieter Rein Ten Wolde. Sampling rare switching events in biochemical networks. *Phys. Rev. Lett.*, 94(1):018104, 2005.
- [40] Rosalind J Allen, Daan Frenkel, and Pieter Rein Ten Wolde. Forward flux sampling-type schemes for simulating rare events: Efficiency analysis. *The Journal of chemical physics*, 124(19):194111, 2006.
- [41] Pratyush Tiwary and B. J. Berne. Spectral gap optimization of order parameters for sampling complex molecular systems. *Proc. Natl. Acad. Sci. U.S.A.*, 113(11):2839–2844, 2016.
- [42] Swetlana Jungblut and Christoph Dellago. Pathways to self-organization: Crystallization via nucleation and growth. *The European Physical Journal E*, 39(8):1–38, 2016.
- [43] Hendrik Anthony Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940.
- [44] Peter Hänggi, Peter Talkner, and Michal Borkovec. Reaction-rate theory: fifty years after kramers. *Reviews of modern physics*, 62(2):251, 1990.
- [45] Nicolaas Godfried Van Kampen. *Stochastic processes in physics and chemistry*, volume 1. Elsevier, 1992.
- [46] Benny Carmeli and Abraham Nitzan. Theory of activated rate processes: Position dependent friction. *Chem. Phys. Lett.*, 102(6):517–522, 1983.
- [47] Guram Chkonia, Judith Wölk, Reinhard Strey, Jan Wedekind, and David Reguera. Evaluating nucleation rates in direct simulations. *The Journal of chemical physics*, 130(6):064505, 2009.
- [48] Vladimir G Baidakov and Azat O Tipeev. On two approaches to determination of the nucleation rate of a new phase in computer experiments. *Thermochimica acta*, 522(1-2):14–19, 2011.

- [49] Peter Talkner. Mean first passage time and the lifetime of a metastable state. *Zeitschrift für Physik B Condensed Matter*, 68(2):201–207, 1987.
- [50] Henry Eyring. The activated complex in chemical reactions. *The Journal of Chemical Physics*, 3(2):107–115, 1935.
- [51] Henry Eyring. The theory of absolute reaction rates. *Transactions of the Faraday Society*, 34:41–48, 1938.
- [52] Eugene Wigner. The transition state method. *Transactions of the Faraday Society*, 34:29–41, 1938.
- [53] Charles H Bennett. Molecular dynamics and transition state theory: the simulation of infrequent events. In *Molecular dynamics and transition state theory: the simulation of infrequent events*. ACS Publications, 1977.
- [54] David Chandler. Statistical mechanics of isomerization dynamics in liquids and the transition state approximation. *J. Chem. Phys.*, 68(6):2959–2970, 1978.
- [55] Peter M Kraus, Michael Zürch, Scott K Cushing, Daniel M Neumark, and Stephen R Leone. The ultrafast x-ray spectroscopic revolution in chemical dynamics. *Nature Reviews Chemistry*, 2(6):82–94, 2018.
- [56] Margherita Maiuri, Marco Garavelli, and Giulio Cerullo. Ultrafast spectroscopy: state of the art and open challenges. *Journal of the American Chemical Society*, 142(1):3–15, 2019.
- [57] Herman JC Berendsen. Practical algorithms for dynamic simulations. *Molecular-dynamics simulation of statistical-mechanical systems*, 1986.
- [58] Miguel A. González. Force fields and molecular dynamics simulations. *École thématique de la Société Française de la Neutronique*, 12:169–200, 2011.
- [59] Alex D MacKerell Jr, Donald Bashford, MLDR Bellott, Roland Leslie Dunbrack Jr, Jeffrey D Evanseck, Martin J Field, Stefan Fischer, Jiali Gao, H Guo, Sookhee Ha, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry B*, 102(18):3586–3616, 1998.
- [60] Wendy D Cornell, Piotr Cieplak, Christopher I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, David C Spellmeyer, Thomas Fox, James W Caldwell, and Peter A Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

- [61] Chris Oostenbrink, Alessandra Villa, Alan E Mark, and Wilfred F Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: the gromos force-field parameter sets 53a5 and 53a6. *Journal of computational chemistry*, 25(13):1656–1676, 2004.
- [62] William L Jorgensen, David S Maxwell, and Julian Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996.
- [63] Huai Sun. Compass: an ab initio force-field optimized for condensed-phase applications overview with details on alkane and benzene compounds. *The Journal of Physical Chemistry B*, 102(38):7338–7364, 1998.
- [64] Pekka Mark and Lennart Nilsson. Structure and dynamics of the tip3p, spc, and spc/e water models at 298 k. *The Journal of Physical Chemistry A*, 105(43):9954–9960, 2001.
- [65] Jan Zielkiewicz. Structural properties of water: Comparison of the spc, spce, tip4p, and tip5p models of water. *The Journal of chemical physics*, 123(10):104501, 2005.
- [66] Jörg Behler. Perspective: Machine learning potentials for atomistic simulations. *The Journal of chemical physics*, 145(17):170901, 2016.
- [67] Oliver T Unke, Stefan Chmiela, Huziel E Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T Schütt, Alexandre Tkatchenko, and Klaus-Robert Müller. Machine learning force fields. *Chemical Reviews*, 121(16):10142–10186, 2021.
- [68] Herman JC Berendsen, JPM van Postma, Wilfred F van Gunsteren, ARHJ DiNola, and Jan R Haak. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, 81(8):3684–3690, 1984.
- [69] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.
- [70] Michael R Shirts. Simple quantitative tests to validate sampling from thermodynamic ensembles. *Journal of chemical theory and computation*, 9(2):909–926, 2013.
- [71] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182–7190, 1981.
- [72] Paul P Ewald. Die berechnung optischer und elektrostatischer gitterpotentiale. *Annalen der physik*, 369(3):253–287, 1921.
- [73] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An $n \log(n)$ method for ewald sums in large systems. *The Journal of chemical physics*, 98(12):10089–10092, 1993.

- [74] Jutta Rogal. Reaction coordinates in complex systems-a perspective. *The European Physical Journal B*, 94(11):1–9, 2021.
- [75] Baron Peters. Reaction coordinates and mechanistic hypothesis tests. *Annu. Rev. Phys. Chem.*, 67(1):669–690, 2016.
- [76] Lars Onsager. Initial recombination of ions. *Physical Review*, 54(8):554, 1938.
- [77] Davide Bauduardi, Francesco Luigi Gervasio, and Michele Parrinello. From a to b in free energy space. *The Journal of chemical physics*, 126(5):054103, 2007.
- [78] Michael AA Cox and Trevor F Cox. Multidimensional scaling. In *Handbook of data visualization*, pages 315–347. Springer, 2008.
- [79] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer, 1997.
- [80] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.
- [81] Gareth A Tribello, Michele Ceriotti, and Michele Parrinello. Using sketch-map coordinates to analyze and bias molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, 109(14):5196–5201, 2012.
- [82] Ronald R Coifman, Ioannis G Kevrekidis, Stéphane Lafon, Mauro Maggioni, and Boaz Nadler. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Modeling & Simulation*, 7(2):842–864, 2008.
- [83] Mohammad M Sultan and Vijay S Pande. Automated design of collective variables using supervised machine learning. *The Journal of chemical physics*, 149(9):094106, 2018.
- [84] Alexandre Jedrecy. *Study of phase transformation of matter through topological coordinates*. PhD thesis, Sorbonne université, 2020.
- [85] Matthew C Zwier and Lillian T Chong. Reaching biological timescales with all-atom molecular dynamics simulations. *Current opinion in pharmacology*, 10(6):745–752, 2010.
- [86] Eric Darve and Andrew Pohorille. Calculating free energies using average force. *The Journal of chemical physics*, 115(20):9169–9183, 2001.
- [87] Jeffrey Comer, James C Gumbart, Jérôme Hénin, Tony Lelièvre, Andrew Pohorille, and Christophe Chipot. The adaptive biasing force method: Everything you always wanted to know but were afraid to ask. *The Journal of Physical Chemistry B*, 119(3):1129–1151, 2015.

- [88] Glenn M Torrie and John P Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.
- [89] Jérôme Hénin, Tony Lelièvre, Michael R Shirts, Omar Valsson, and Lucie Delemonette. Enhanced sampling methods for molecular dynamics simulations. *arXiv preprint arXiv:2202.04164*, 2022.
- [90] Shankar Kumar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of computational chemistry*, 13(8):1011–1021, 1992.
- [91] Giovanni Bussi and Alessandro Laio. Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics*, 2:200–212, 2020.
- [92] Yanier Crespo, Fabrizio Marinelli, Fabio Pietrucci, and Alessandro Laio. Metadynamics convergence law in a multidimensional system. *Physical Review E*, 81(5):055701, 2010.
- [93] Giovanni Bussi, Alessandro Laio, and Michele Parrinello. Equilibrium free energies from nonequilibrium metadynamics. *Physical review letters*, 96(9):090601, 2006.
- [94] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical review letters*, 100(2):020603, 2008.
- [95] Baron Peters and Bernhardt L Trout. Obtaining reaction coordinates by likelihood maximization. *The Journal of chemical physics*, 125(5):054108, 2006.
- [96] Baron Peters, Gregg T Beckham, and Bernhardt L Trout. Extensions to the likelihood maximization approach for finding reaction coordinates. *The Journal of chemical physics*, 127(3):034109, 2007.
- [97] Hendrik Jung, Kei ichi Okazaki, and Gerhard Hummer. Transition path sampling of rare events by shooting from the top. *J. Chem. Phys.*, 147(15):152716, 2017.
- [98] Steffen Wolf, Benjamin Lickert, Simon Bray, and Gerhard Stock. Multisecond ligand dissociation dynamics from atomistic simulations. *Nature communications*, 11(1):1–8, 2020.
- [99] Nuria Plattner, Stefan Doerr, Gianni De Fabritiis, and Frank Noé. Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and markov modelling. *Nature chemistry*, 9(10):1005–1011, 2017.
- [100] John D Chodera and Frank Noé. Markov state models of biomolecular conformational dynamics. *Current opinion in structural biology*, 25:135–144, 2014.

- [101] Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2011.
- [102] Daria B Kokh, Marta Amaral, Joerg Bomke, Ulrich Grädler, Djordje Musil, Hans-Peter Buchstaller, Matthias K Dreyer, Matthias Frech, Maryse Lowinski, Francois Vallee, et al. Estimation of drug-target residence times by τ -random acceleration molecular dynamics simulations. *Journal of chemical theory and computation*, 14(7):3859–3869, 2018.
- [103] Ariane Nunes-Alves, Daria B Kokh, and Rebecca C Wade. Recent progress in molecular simulation methods for drug binding kinetics. *Current Opinion in Structural Biology*, 64:126–133, 2020.
- [104] Lukas S Stelzl and Gerhard Hummer. Kinetics from replica exchange molecular dynamics simulations. *Journal of chemical theory and computation*, 13(8):3927–3935, 2017.
- [105] Lukas S Stelzl, Adam Kells, Edina Rosta, and Gerhard Hummer. Dynamic histogram analysis to determine free energies and rates from biased simulations. *Journal of chemical theory and computation*, 13(12):6328–6342, 2017.
- [106] Magd Badaoui, Adam Kells, Carla Molteni, Callum J Dickson, Viktor Hornak, and Edina Rosta. Calculating kinetic rates and membrane permeability from biased simulations. *The Journal of Physical Chemistry B*, 122(49):11571–11578, 2018.
- [107] Timo M Schäfer and Giovanni Settanni. Data reweighting in metadynamics simulations. *Journal of chemical theory and computation*, 16(4):2042–2052, 2020.
- [108] Stefanie Kieninger, Luca Donati, and Bettina G Keller. Dynamical reweighting methods for markov models. *Current opinion in structural biology*, 61:124–131, 2020.
- [109] Stephanie M Linker, R Gregor Weiß, and Sereina Riniker. Connecting dynamic reweighting algorithms: Derivation of the dynamic reweighting family tree. *The Journal of Chemical Physics*, 153(23):234106, 2020.
- [110] Pratyush Tiwary, Vittorio Limongelli, Matteo Salvalaglio, and Michele Parrinello. Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proceedings of the National Academy of Sciences*, 112(5):E386–E391, 2015.
- [111] Rodrigo Casasnovas, Vittorio Limongelli, Pratyush Tiwary, Paolo Carloni, and Michele Parrinello. Unbinding kinetics of a p38 map kinase type ii inhibitor from metadynamics simulations. *Journal of the American Chemical Society*, 139(13):4780–4788, 2017.
- [112] Debabrata Pramanik, Zachary Smith, Adam Kells, and Pratyush Tiwary. Can one trust kinetic and thermodynamic observables from biased metadynamics simulations?: Detailed quantitative benchmarks on millimolar drug fragment dissociation. *The Journal of Physical Chemistry B*, 123(17):3672–3678, 2019.

- [113] João Marcelo Lamim Ribeiro, Davide Provasi, and Marta Filizola. A combination of machine learning and infrequent metadynamics to efficiently predict kinetic rates, transition states, and molecular determinants of drug dissociation from g protein-coupled receptors. *The Journal of Chemical Physics*, 153(12):124105, 2020.
- [114] Mrinal Shekhar, Zachary Smith, Markus Seeliger, and Pratyush Tiwary. Protein flexibility and dissociation pathway differentiation can explain onset of resistance mutations in kinases. *bioRxiv*, 2021.
- [115] Willmor J Peña Ccoa and Glen M Hocky. Assessing models of force-dependent unbinding rates via infrequent metadynamics. *The Journal of Chemical Physics*, 156(12):125102, 2022.
- [116] Helmut Grubmüller. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Physical Review E*, 52(3):2893, 1995.
- [117] Arthur F Voter. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Physical Review Letters*, 78(20):3908, 1997.
- [118] Matteo Salvalaglio, Pratyush Tiwary, and Michele Parrinello. Assessing the reliability of the dynamics reconstructed from metadynamics. *Journal of chemical theory and computation*, 10(4):1420–1425, 2014.
- [119] Bradley M Dickson. Erroneous rates and false statistical confirmations from infrequent metadynamics and other equivalent violations of the hyperdynamics paradigm. *Journal of chemical theory and computation*, 15(1):78–83, 2018.
- [120] Salman A Khan, Bradley M Dickson, and Baron Peters. How fluxional reactants limit the accuracy/efficiency of infrequent metadynamics. *The Journal of Chemical Physics*, 153(5):054125, 2020.
- [121] Donatella Callegari, Alessio Lodola, Daniele Pala, Silvia Rivara, Marco Mor, Andrea Rizzi, and Anna Maria Capelli. Metadynamics simulations distinguish short-and long-residence-time inhibitors of cyclin-dependent kinase 8. *Journal of chemical information and modeling*, 57(2):159–169, 2017.
- [122] Yong Wang, Omar Valsson, Pratyush Tiwary, Michele Parrinello, and Kresten Lindorff-Larsen. Frequency adaptive metadynamics for the calculation of rare-event kinetics. *The Journal of chemical physics*, 149(7):072309, 2018.
- [123] Karen Palacio-Rodriguez, Hadrien Vroylandt, Lukas S Stelzl, Fabio Pietrucci, Gerhard Hummer, and Pilar Cossio. Transition rates and efficiency of collective variables from time-dependent biased simulations. *The journal of physical chemistry letters*, In press, 2022.

- [124] Gerhard Hummer and Attila Szabo. Kinetics from nonequilibrium single-molecule pulling experiments. *Biophysical journal*, 85(1):5–15, 2003.
- [125] Olga K Dudko, Gerhard Hummer, and Attila Szabo. Intrinsic rates and activation free energies from single-molecule pulling experiments. *Physical review letters*, 96(10):108101, 2006.
- [126] Olga K Dudko, Gerhard Hummer, and Attila Szabo. Theory, analysis, and interpretation of single-molecule force spectroscopy experiments. *Proceedings of the National Academy of Sciences*, 105(41):15755–15760, 2008.
- [127] Pilar Cossio, Gerhard Hummer, and Attila Szabo. Kinetic ductility and force-spike resistance of proteins from single-molecule force spectroscopy. *Biophysical journal*, 111(4):832–840, 2016.
- [128] Jerald F. Lawless. Observation Schemes, Censoring, and Likelihood. In *Statistical Models and Methods for Lifetime Data*, chapter 2, pages 49–78. John Wiley & Sons, Ltd, 2002.
- [129] Nickolay Smirnov. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281, 1948.
- [130] Andrey Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91, 1933.
- [131] Edina Rosta and Gerhard Hummer. Free energies from dynamic weighted histogram analysis using unbiased markov state model. *Journal of chemical theory and computation*, 11(1):276–285, 2015.
- [132] Jerome Henin, Giacomo Fiorin, Christophe Chipot, and Michael L Klein. Exploring multidimensional free energy landscapes using time-dependent biases on collective variables. *Journal of chemical theory and computation*, 6(1):35–47, 2010.
- [133] Volodymyr Babin, Christopher Roland, and Celeste Sagui. Adaptively biased molecular dynamics for free energy calculations. *The Journal of chemical physics*, 128(13):134101, 2008.
- [134] James F Dama, Glen M Hocky, Rui Sun, and Gregory A Voth. Exploring valleys without climbing every peak: More efficient and forgiving metabasin metadynamics via robust on-the-fly bias domain restriction. *J. Chem. Theory Comput.*, 11(12):5638–5650, 2015.
- [135] Bradley M Dickson. Overfill protection and hyperdynamics in adaptively biased simulations. *J. Chem. Theory Comput.*, 13(12):5925–5932, 2017.
- [136] Harold W Kroto, James R Heath, Sean C O’Brien, Robert F Curl, and Richard E Smalley. C60: Buckminsterfullerene. *nature*, 318(6042):162–163, 1985.

- [137] Luca Monticelli. On atomistic and coarse-grained models for C₆₀ fullerene. *J. Chem. Theory Comput.*, 8(4):1370–1378, 2012.
- [138] Yves Noël, Marco De La Pierre, Claudio M Zicovich-Wilson, Roberto Orlando, and Roberto Dovesi. Structural, electronic and energetic properties of giant icosahedral fullerenes up to c6000: insights from an ab initio hybrid dft study. *Physical Chemistry Chemical Physics*, 16(26):13390–13401, 2014.
- [139] Herman JC Berendsen, David van der Spoel, and Rudi van Drunen. Gromacs: a message-passing parallel molecular dynamics implementation. *Computer physics communications*, 91(1-3):43–56, 1995.
- [140] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.
- [141] Gareth A Tribello, Massimiliano Bonomi, Dawide Branduardi, Carlo Camilloni, and Giovanni Bussi. Plumed 2: New feathers for an old bird. *Computer Physics Communications*, 185(2):604–613, 2014.
- [142] David Owen Morgan. *The cell cycle: principles of control*. New science press, 2007.
- [143] Marco De Vivo, Giovanni Bottegoni, Anna Berteotti, Maurizio Recanatini, Francesco Luigi Gervasio, and Andrea Cavalli. Cyclin-dependent kinases: bridging their structure and function through computations. *Future medicinal chemistry*, 3(12):1551–1559, 2011.
- [144] George Kontopidis, Campbell McInnes, Sravan R Pandalaneni, Iain McNae, Darren Gibson, Mokdad Mezna, Mark Thomas, Gavin Wood, Shudong Wang, Malcolm D Walkinshaw, et al. Differential binding of inhibitors to active and inactive cdk2 provides insights for drug design. *Chemistry & biology*, 13(2):201–211, 2006.
- [145] Yan Li, Jingxiao Zhang, Weimin Gao, Lilei Zhang, Yanqiu Pan, Shuwei Zhang, and Yonghua Wang. Insights on structural characteristics and ligand binding mechanisms of cdk2. *International journal of molecular sciences*, 16(5):9314–9340, 2015.
- [146] Philip D Jeffrey, Alicia A Russo, Kornelia Polyak, Emma Gibbs, Jerard Hurwitz, Joan Massagué, and Nikola P Pavletich. Mechanism of cdk activation revealed by the structure of a cyclin-cdk2 complex. *Nature*, 376(6538):313–320, 1995.
- [147] C Benson, S Kaye, P Workman, M Garrett, M Walton, and J De Bono. Clinical anticancer drug development: targeting the cyclin-dependent kinases. *British journal of cancer*, 92(1):7–12, 2005.

- [148] Abdulilah Ece and Fatma Sevin. The discovery of potential cyclin a/cdk2 inhibitors: a combination of 3d qsar pharmacophore modeling, virtual screening, and molecular docking studies. *Medicinal chemistry research*, 22(12):5832–5843, 2013.
- [149] Geoffrey I Shapiro. Cyclin-dependent kinase pathways as targets for cancer treatment. *Journal of clinical oncology*, 24(11):1770–1783, 2006.
- [150] Eiman Aleem and Robert J Arceci. Targeting cell cycle regulators in hematologic malignancies. *Frontiers in cell and developmental biology*, 3:16, 2015.
- [151] Bartłomiej Bartkowiak, Pengda Liu, Hemali P Phatnani, Nicholas J Fuda, Jeffrey J Cooper, David H Price, Karen Adelman, John T Lis, and Arno L Greenleaf. Cdk12 is a transcription elongation-associated ctd kinase, the metazoan ortholog of yeast ctk1. *Genes & development*, 24(20):2303–2316, 2010.
- [152] S Vadivelan, Barij Nayan Sinha, Sheeba Jem Irudayam, and Sarma ARP Jagarlapudi. Virtual screening studies to design potent cdk2-cyclin a inhibitors. *Journal of chemical information and modeling*, 47(4):1526–1535, 2007.
- [153] José S Duca, Vincent S Madison, and Johannes H Voigt. Cross-docking of inhibitors into cdk2 structures. 1. *Journal of chemical information and modeling*, 48(3):659–668, 2008.
- [154] Magd Badaoui, Pedro J Buigues, Dénes Berta, Gaurav M Mandana, Hankang Gu, Tamas Foldes, Callum J Dickson, Viktor Hornak, Mitsunori Kato, Carla Molteni, et al. Combined free-energy calculation and machine learning methods for understanding ligand unbinding kinetics. *Journal of chemical theory and computation*, 18(4):2543–2555, 2022.
- [155] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L Klepeis, Ron O Dror, and David E Shaw. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8):1950–1958, 2010.
- [156] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79(2):926–935, 1983.
- [157] Junmei Wang, Wei Wang, Peter A Kollman, and David A Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of molecular graphics and modelling*, 25(2):247–260, 2006.
- [158] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of computational chemistry*, 25(9):1157–1174, 2004.

- [159] Alan W Sousa Da Silva and Wim F Vranken. Acpype-antechamber python parser interface. *BMC research notes*, 5(1):1–8, 2012.
- [160] Rutger HA Folmer. Drug target residence time: a misleading concept. *Drug Discovery Today*, 23(1):12–16, 2018.
- [161] Neil J Bruce, Gaurav K Ganatra, Daria B Kokh, S Kashif Sadiq, and Rebecca C Wade. New approaches for computing ligand–receptor binding kinetics. *Current opinion in structural biology*, 49:1–10, 2018.
- [162] James B Dunbar Jr, Richard D Smith, Kelly L Damm-Ganamet, Aqeel Ahmed, Emilio Xavier Esposito, James Delproposto, Krishnapriya Chinnaswamy, You-Na Kang, Ginger Kubish, Jason E Gestwicki, et al. Csar data set release 2012: ligands, affinities, complexes, and docking decoys. *Journal of chemical information and modeling*, 53(8):1842–1852, 2013.
- [163] Jakob T Bullerjahn, Sebastian Sturm, and Klaus Kroy. Theory of rapid force spectroscopy. *Nature communications*, 5(1):1–10, 2014.
- [164] Hannes Risken. *The Fokker-Planck Equation*. Springer, 1996.
- [165] J Luczka. Non-markovian stochastic processes: Colored noise. *Chaos*, 15(2):026107, 2005.
- [166] Richard F Grote and James T Hynes. The stable states picture of chemical reactions. ii. rate constants for condensed and gas phase reaction models. *J. Chem. Phys.*, 73(6):2715–2732, 1980.
- [167] Hee Sun Lee, Surl-Hee Ahn, and Eric F Darve. Building a coarse-grained model based on the mori-zwanzig formalism. *MRS Online Proceedings Library Archive*, 1753, 2015.
- [168] Jan O Daldrop, Julian Kappler, Florian N Brünig, and Roland R Netz. Butane dihedral angle dynamics in water is dominated by internal friction. *Proc. Natl. Acad. Sci. U.S.A.*, 115(20):5169–5174, 2018.
- [169] Jianfeng Lu and Eric Vanden-Eijnden. Exact dynamical coarse-graining without time-scale separation. *The Journal of Chemical Physics*, 141(4):044109, 2014.
- [170] John E Straub, Michal Borkovec, and Bruce J Berne. Calculation of dynamic friction on intramolecular degrees of freedom. *J. Phys. Chem.*, 91(19):4995–4998, 1987.
- [171] M Tuckerman and BJ Berne. Vibrational relaxation in simple fluids: Comparison of theory and simulation. *The Journal of chemical physics*, 98(9):7301–7318, 1993.
- [172] J Timmer. Parameter estimation in nonlinear stochastic differential equations. *Chaos, Solitons & Fractals*, 11(15):2571–2578, 2000.

- [173] Janez Gradišek, Silke Siegert, Rudolf Friedrich, and Igor Grabec. Analysis of time series from stochastic processes. *Phys. Rev. E*, 62(3):3146, 2000.
- [174] Alexandre J Chorin, Ole H Hald, and Raz Kupferman. Optimal prediction with memory. *Phys. D*, 166(3-4):239–257, 2002.
- [175] Gerhard Hummer and Ioannis G Kevrekidis. Coarse molecular dynamics of a peptide fragment: Free energy, kinetics, and long-time dynamics computations. *J. Chem. Phys.*, 118(23):10762–10773, 2003.
- [176] Gerhard Hummer. Position-dependent diffusion coefficients and free energies from bayesian analysis of equilibrium and replica molecular dynamics simulations. *New J. Phys.*, 7(1):34, 2005.
- [177] Robert B Best and Gerhard Hummer. Diffusive model of protein folding dynamics with kramers turnover in rate. *Phys. Rev. Lett.*, 96(22):228104, 2006.
- [178] Oliver F Lange and Helmut Grubmüller. Collective langevin dynamics of conformational motions in proteins. *J. Chem. Phys.*, 124(21):214903, 2006.
- [179] Illia Horenko, Carsten Hartmann, Christof Schütte, and Frank Noe. Data-based parameter estimation of generalized multidimensional langevin processes. *Phys. Rev. E*, 76(1):016706, 2007.
- [180] Eric Darve, Jose Solomon, and Amirali Kia. Computing generalized langevin equations and generalized fokker–planck equations. *Proc. Natl. Acad. Sci. U.S.A.*, 106(27):10884–10889, 2009.
- [181] Cristian Micheletti, Giovanni Bussi, and Alessandro Laio. Optimal langevin modeling of out-of-equilibrium molecular dynamics simulations. *J. Chem. Phys.*, 129(7):074105, 2008.
- [182] Alexander Berezhkovskii and Attila Szabo. Time scale separation leads to position-dependent diffusion along a slow coordinate. *The Journal of chemical physics*, 135(7):074108, 2011.
- [183] Qi Zhang, Jasna Brujić, and Eric Vanden-Eijnden. Reconstructing free energy profiles from nonequilibrium relaxation trajectories. *J. Stat. Phys.*, 144(2):344–366, 2011.
- [184] Daan Crommelin and Eric Vanden-Eijnden. Diffusion estimation from multiscale data by operator eigenpairs. *Multiscale Modeling & Simulation*, 9(4):1588–1623, 2011.
- [185] Daan Crommelin. Estimation of space-dependent diffusions and potential landscapes from non-equilibrium data. *Journal of Statistical Physics*, 149(2):220–233, 2012.

- [186] Norbert Schaudinnus, Björn Bastian, Rainer Hegger, and Gerhard Stock. Multidimensional langevin modeling of nonoverdamped dynamics. *Phys. Rev. Lett.*, 115(5):050602, 2015.
- [187] Norbert Schaudinnus, Benjamin Lickert, Mithun Biswas, and Gerhard Stock. Global langevin model of multidimensional biomolecular dynamics. *J. Chem. Phys.*, 145(18):184114, 2016.
- [188] Dominika Lesnicki, Rodolphe Vuilleumier, Antoine Carof, and Benjamin Rotenberg. Molecular hydrodynamics from memory kernels. *Phys. Rev. Lett.*, 116(14):147804, 2016.
- [189] Roberto Meloni, Carlo Camilloni, and Guido Tiana. Properties of low-dimensional collective variables in the molecular dynamics of biopolymers. *Phys. Rev. E*, 94(5):052406, 2016.
- [190] Mithun Biswas, Benjamin Lickert, and Gerhard Stock. Metadynamics enhanced markov modeling of protein dynamics. *The Journal of Physical Chemistry B*, 122(21):5508–5514, 2018.
- [191] Frederico Campos Freitas, Angelica Nakagawa Lima, Vinicius de Godoi Contessoto, Paul C Whitford, and Ronaldo Junio de Oliveira. Drift-diffusion (drdiff) framework determines kinetics and thermodynamics of two-state folding trajectory and tunes diffusion models. *The Journal of chemical physics*, 151(11):114106, 2019.
- [192] Feliks Nüske, Lorenzo Boninsegna, and Cecilia Clementi. Coarse-graining molecular systems by spectral matching. *The Journal of chemical physics*, 151(4):044116, 2019.
- [193] Marco Baldovin, Fabio Cecconi, and Angelo Vulpiani. Effective equations for reaction coordinates in polymer transport. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(1):013208, 2020.
- [194] Benjamin Lickert and Gerhard Stock. Modeling non-markovian data using markov state and langevin models. *The Journal of Chemical Physics*, 153(24):244112, 2020.
- [195] Shu Wang, Zhan Ma, and Wenxiao Pan. Data-driven coarse-grained modeling of polymers in solution with structural and dynamic properties conserved. *Soft Matter*, 16(36):8330–8344, 2020.
- [196] Cihan Ayaz, Lucas Tepper, Florian N Brünig, Julian Kappler, Jan O Daldrop, and Roland R Netz. Non-markovian modeling of protein folding. *Proceedings of the National Academy of Sciences*, 118(31):e2023856118, 2021.
- [197] Hadrien Vroylandt, Ludovic Goudenège, Pierre Monmarché, Fabio Pietrucci, and Benjamin Rotenberg. Likelihood-based non-markovian models from molecular dynamics. *Proceedings of the National Academy of Sciences*, 119(13):e2117586119, 2022.

- [198] Karen Palacio-Rodriguez and Fabio Pietrucci. Free energy landscapes, diffusion coefficients, and kinetic rates from transition paths. *Journal of chemical theory and computation*, In press, 2022.
- [199] Rep Kubo. The fluctuation-dissipation theorem. *Rep. Prog. Phys.*, 29(1):255, 1966.
- [200] Robert Zwanzig. Nonlinear generalized langevin equations. *J. Stat. Phys.*, 9(3):215–220, 1973.
- [201] George R Haynes, Gregory A Voth, and Eli Pollak. A theory for the activated barrier crossing rate constant in systems influenced by space and time dependent friction. *The Journal of chemical physics*, 101(9):7811–7822, 1994.
- [202] Hadrien Vroylandt and Pierre Monmarché. Position-dependent memory kernel in generalized langevin equations: theory and numerical estimation. *The Journal of Chemical Physics*, 2022.
- [203] Andy WC Lau and Tom C Lubensky. State-dependent diffusion: Thermodynamic consistency and its path integral formulation. *Phys. Rev. E*, 76(1):011123, 2007.
- [204] Alexander N Drozdov. High-accuracy discrete path integral solutions for stochastic processes with noninvertible diffusion matrices. *Phys. Rev. E*, 55(3):2496, 1997.
- [205] Peter Eris Kloeden, Eckhard Platen, and Henri Schurz. *Numerical solution of SDE through computer experiments*. Springer Science & Business Media, 2012.
- [206] S Izrailev, S Stepaniants, B Isralewitz, D Kosztin, H Lu, F Molnar, W Wriggers, and K Schulten. Computational molecular dynamics: Challenges, methods. In *Ideas*, volume 4, page 39, 1998.
- [207] Robert B Best and Gerhard Hummer. Reaction coordinates and rates from transition paths. *Proceedings of the National Academy of Sciences*, 102(19):6732–6737, 2005.
- [208] Amit Samanta, Mark E Tuckerman, Tang-Qing Yu, and E Weinan. Microscopic mechanisms of equilibrium melting of a solid. *Science*, 346(6210):729–732, 2014.
- [209] Peter Hänggi, Peter Talkner, and Michal Borkovec. Reaction-rate theory: fifty years after kramers. *Rev. Mod. Phys.*, 62:251–341, 1990.
- [210] Peter G Bolhuis and Christoph Dellago. Trajectory-based rare event simulations. *Rev. Comput. Chem.*, 27:111, 2010.
- [211] Saravanapriyan Sriraman, Ioannis G Kevrekidis, and Gerhard Hummer. Coarse master equation from bayesian analysis of replica molecular dynamics simulations. *The Journal of Physical Chemistry B*, 109(14):6479–6484, 2005.

- [212] Max Innerbichler, Georg Menzl, and Christoph Dellago. State-dependent diffusion coefficients and free energies for nucleation processes from bayesian trajectory analysis. *Mol. Phys.*, 116:2987–2997, 2018.
- [213] Z Faidon Brotzakis and Peter G Bolhuis. Approximating free energy and committor landscapes in standard transition path sampling using virtual interface exchange. *The Journal of chemical physics*, 151(17):174111, 2019.
- [214] Polina V Banushkina and Sergei V Krivov. Optimal reaction coordinates. *WIREs: Comput. Mol. Sci.*, 6(6):748–763, 2016.
- [215] Wei Chen, Aik Rui Tan, and Andrew L Ferguson. Collective variable discovery and enhanced sampling using autoencoders: Innovations in network architecture and error function design. *The Journal of chemical physics*, 149(7):072312, 2018.
- [216] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.
- [217] Soumik Banerjee. Molecular dynamics study of self-agglomeration of charged fullerenes in solvents. *J. Chem. Phys.*, 138(4), 2013.
- [218] Liwei Li, Dmitry Bedrov, and Grant D. Smith. A molecular-dynamics simulation study of solvent-induced repulsion between C 60 fullerenes in water. *J. Chem. Phys.*, 123(20):1–8, 2005.
- [219] Chi Cheng Chiu, Russell DeVane, Michael L. Klein, Wataru Shinoda, Preston B. Moore, and Steven O. Nielsen. Coarse-grained potential models for phenyl-based molecules: II. Application to fullerenes. *J. Phys. Chem. B*, 114(19):6394–6400, 2010.
- [220] Dongheon Lee, Kyeongpung Lee, Dongsun Yoo, Wonseok Jeong, and Seungwu Han. Crystallization of amorphous ge te simulated by neural network potential addressing medium-range order. *Computational Materials Science*, 181:109725, 2020.
- [221] Gabriele C Sosso, Giacomo Miceli, Sebastiano Caravati, Federico Giberti, Jörg Behler, and Marco Bernasconi. Fast crystallization of the phase change compound ge te by large-scale molecular dynamics simulations. *The journal of physical chemistry letters*, 4(24):4241–4246, 2013.
- [222] Koji Ohara, László Temleitner, Kunihisa Sugimoto, Shinji Kohara, Toshiyuki Matsunaga, László Pusztai, Masayoshi Itou, Hiroyuki Ohsumi, Rie Kojima, Noboru Yamada, et al. The roles of the ge-te core network and the sb-te pseudo network during rapid nucleation-dominated crystallization of amorphous ge₂sb₂te₅. *Advanced Functional Materials*, 22(11):2251–2257, 2012.

- [223] Aidan P Thompson, H Metin Aktulga, Richard Berger, Dan S Bolintineanu, W Michael Brown, Paul S Crozier, Pieter J in't Veld, Axel Kohlmeyer, Stan G Moore, Trung Dac Nguyen, et al. Lammps-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications*, 271:108171, 2022.
- [224] Pablo M Piaggi and Michele Parrinello. Predicting polymorphism in molecular crystals using orientational entropy. *Proceedings of the National Academy of Sciences*, 115(41):10251–10256, 2018.
- [225] G Bruns, P Merkelbach, C Schlockermann, M Salinga, M Wuttig, TD Happ, JB Philipp, and M Kund. Nanosecond switching in gete phase change memory cells. *Applied physics letters*, 95(4):043108, 2009.
- [226] Sarah EM Lundrigan and Ivan Saika-Voivod. Test of classical nucleation theory and mean first-passage time formalism on crystallization in the lennard-jones liquid. *The Journal of Chemical Physics*, 131(10):104503, 2009.
- [227] Valeria Molinero and Emily B Moore. Water modeled as an intermediate element between carbon and silicon. *The Journal of Physical Chemistry B*, 113(13):4008–4016, 2009.
- [228] Albert C Pan, Daniel Jacobson, Konstantin Yatsenko, Duluxan Sritharan, Thomas M Weinreich, and David E Shaw. Atomic-level characterization of protein–protein association. *Proceedings of the National Academy of Sciences*, 116(10):4244–4249, 2019.
- [229] Robert W Hartley. Barnase and barstar: expression of its cloned inhibitor permits expression of a cloned ribonuclease. *Journal of molecular biology*, 202(4):913–915, 1988.
- [230] Robert W Hartley. Barnase and barstar: two small proteins to fold and fit together. *Trends in biochemical sciences*, 14(11):450–454, 1989.
- [231] Gideon Schreiber and Alan R Fersht. Interaction of barnase with its polypeptide inhibitor barstar studied by protein engineering. *Biochemistry*, 32(19):5145–5150, 1993.
- [232] Gideon Schreiber and Alan R Fersht. Rapid, electrostatically assisted association of proteins. *Nature structural biology*, 3(5):427–431, 1996.
- [233] Ashley M Buckle, Gideon Schreiber, and Alan R Fersht. Protein-protein recognition: Crystal structural analysis of a barnase-barstar complex at 2.0-. ang. resolution. *Biochemistry*, 33(30):8878–8889, 1994.
- [234] James C Gumbart, Benoît Roux, and Christophe Chipot. Efficient determination of protein–protein standard binding free energies from first principles. *Journal of chemical theory and computation*, 9(8):3789–3798, 2013.

- [235] Alessandro Barducci, Massimiliano Bonomi, Meher K Prakash, and Michele Parrinello. Free-energy landscape of protein oligomerization from atomistic simulations. *Proceedings of the National Academy of Sciences*, 110(49):E4708–E4713, 2013.
- [236] Fabio Pietrucci and Alessandro Laio. A collective variable for the efficient exploration of protein beta-sheet structures: application to sh3 and gb1. *Journal of Chemical Theory and Computation*, 5(9):2197–2201, 2009.
- [237] Wei Zhang, Carsten Hartmann, and Christof Schütte. Effective dynamics along given reaction coordinates, and reaction rate theory. *Faraday discussions*, 195:365–394, 2016.
- [238] Hendrik Jung, Roberto Covino, and Gerhard Hummer. Artificial intelligence assists discovery of reaction coordinates and mechanisms from molecular dynamics simulations. *arXiv preprint arXiv:1901.04595*, 2019.