Rare-Event Sampling of Ligand Transport in Proteins

A dissertation presented
by

Jakub Rydzewski

to
Institute of Physics
Faculty of Physics, Astronomy and Informatics
Nicolaus Copernicus University

in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

in the subject of Biophysics

under the supervision of Prof. dr hab. W. Nowak

Nicolaus Copernicus University
Toruń
March 2018

Acknowledgments

My gratitude to:

- Prof. W. Nowak (Faculty of Physics, Institute of Physics, Astronomy and Informatics, Nicolaus Copernicus University, Toruń, Poland);
- Prof. H. Grubmüller (Department of Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany);
- Prof. M. Parrinello (Department of Chemistry and Applied Biosciences, Swiss Federal Institute of Technology in Zurich c/o Institute of Computational Science, Università della Svizzera italiana, Lugano, Switzerland);
- Prof. H. Nakatsuji (Quantum Chemistry Research Institute, Kyoto, Japan).

This research was funded by:

- National Science Centre: grant Preludium (2015/19/N/ST3/02171);
- National Science Centre: grant Etiuda (2016/20/T/ST3/00488);
- National Science Centre: grant Opus (2016/23/B/ST4/01770);
- Ministry of Science and Higher Education: grant (2012/05/N/ST3/03178);
- Ministry of Science and Higher Education: scholarship for outstanding achievements (2016/2017 and 2017/2018);
- Kuyavian-Pomeranian Voivodeship Marshal (2015/2016).

For providing computer facilities, I am indebted to Interdisciplinary Centre for Modern Technologies at Nicolaus Copernicus University in Toruń and Swiss National Supercomputing Centre at Swiss Federal Institute of Technology in Zurich.

Toruń, 6 XII 2017 JR

Contents

1 Abstract

Molecular dynamics (MD) allows an insight into dynamics of biological macromolecules on the atomic scale. There are instances, however, in which the sampled physical process occurs rarely. The problem of rare-event sampling is associated with large energy barriers ($\gg kT$) separating important potential energy minima.[1] When an energy barrier is difficult to overcome, barrier crossing becomes a rare event, and the probability that a thermal fluctuation will drive a system from one minimum to another over such a barrier, becomes exponentially small with the ratio of the barrier height to kT. Nonetheless, rare transitions through disjoint configurational regions are possible if the system displays ergodicity.[1,2] This means that in an infinite time limit, the time spent by the system in some region of configuration space is proportional to the volume of this region, and thus, all accessible intermediate states are populated. In the context of kinetics, an event becomes rare if the time scale of this event is much higher than the time scale of molecular motions.

Transport processes in heterogeneous media (e.g., ion diffusion, ligand migration) are rare events.[3] Although accounting for protein dynamics at all the stages of transport processes presents a challenge for both experimental and computational approaches,[4] general questions about protein coupling are related to the nature of transport,[5] e.g., passive or active, which is crucial for the rational design of compound with improved selectivity and specificity.[6] The experimental methods used currently to quantify ligand diffusion, e.g., time-resolved crystallography,[7] spectroscopy[8,9] and xenon binding,[10,11] focus primarily on gaseous species, providing indirect evidence for the migration of larger ligands and limited knowledge about their binding stability, lifetime and specificity.[12,13]

The dynamics of protein binding pockets and transient channels/tunnels is important for ligand transport,[12,14] because interior topological features and their structural flexibility allow proteins to facilitate binding processes by adapting to their individual binding partners. This includes transport along possibly multiple binding pathways to the binding pocket, recognition at the binding site and advancing ligand through transient exit routes to solvent.[3] For instance, among biophysical processes governing the general relationship between ligand binding and protein dynamics are allosteric signaling[15,16] and ligand-induced stability.[17] The residence time is dependent on structural features of both the ligand and the protein, therefore

such aspects as substrate specificity, and tunnel geometry cannot be neglected in quantifying the efficacy of ligand recognition.[4,18]

Such studies are important also from the point of view of drug discovery. The response to a drug is not only associated with the topology of egress pathways, but also related to their binding kinetics and thermodynamics.[19,20] After the ligand reaches the docking pocket, it resides in its destination for a period of time, which is of pivotal importance in all regulatory processes in biology.[21,22] The need for understanding how ligands migrate through protein tunnels has spurred the development of several methods, including steered MD,[23] locally enhanced sampling,[24,25] and metadynamics.[26–31]

This thesis is devoted to the development of a new method for the optimal reconstruction of ligand transport pathways and free-energy profiles along these pathways. The methodology presented here is general and may be adopted to other transport processes. The following theses are considered in this dissertation:

- Binding mechanisms underlying ligand transport in proteins are related to the degree of coupling between protein dynamics and ligand migration. Protein dynamics is paramount for the prediction of the network of tunnels and channels;
- Accounting for internal topological features of proteins (e.g., channels, tunnels and cavities) and their thermodynamics is necessary to effectively model the optimal pathways of ligand transport;
- Statistical learning (e.g., machine learning and dimensionality reduction) offers methods to reduce 'big data' stemming from simulations to mine meaningful generalized coordinates.

Abstract (Polish)

Dynamika molekularna pozwala na obserwowanie ewolucji biologicznych makromolekuł na poziomie atomowym. Niestety, w niektórych przypadkach symulowane zdarzenie jest niewystarczająco próbkowane. Problem próbkowania rzadkiego zdarzenia jest związany z wysokimi barierami energetycznymi ($\gg kT$) separującymi ważne dla tego zdarzenia minima energii potencjalnej.[1] Pokonanie takiej bariery staje się rzadkim zdarzeniem, gdy bariera jest nieprzekraczalna w badanej skali czasowej. Prawdopodobieństwo pokonania bariery energii poprzez termiczną fluktuację staje się eksponencjalnie niskie z wykładnikiem równym stosunkowi energii bariery do kT. Takie przejście przez rozłączne regiony przestrzeni konfiguracyjnej jest teoretycznie możliwe, jeśli badany układ spełnia hipotezę ergodyczną.[1,2] Oznacza to, że gdy czas symulacji dąży do nieskończoności, czas spędzony przez układ w danym regionie przestrzeni konfiguracyjnej jest proporcjonalny do objętości tego regionu, więc wszystkie dostępne stany przejściowe są równoprawdopodobne. W kontekście kinetyki, zdarzenie staje się rzadkie, jeśli jego skala czasowa jest o wiele większa niż skala czasowa podstawowych ruchów molekularnych.

Zjawiska transportowe w heterogenicznych ośrodkach (np. dyfuzja jonów, migracja ligandów) są procesami rzadkimi.[3] Uwzględnienie dynamiki białek na każdym poziomie procesu transportowego jest wyzwaniem nie tylko dla eksperymetów, ale również dla symulacji.[4] Poziom sprzężenia dynamiki układów biologicznych z procesami transportowymi jest zależny od fundametalnych zasad kierujących transportem biologicznym.[5,6] Techniki eksperymentalne używane obecnie do badania dyfuzji ligandów, np. krystalografia w domenie czasowej,[7] spektroskopia[8,9] oraz wiązanie ksenonu,[10,11] dostarczają pośrednie dowody na migrację większych ligandów oraz ograniczoną wiedzę o stabilności wiązania, czasie życia oraz specyficzności.[12,13]

Dynamika krótkotrwało otwartych białkowych tuneli oraz kanałów jest ważna dla transportu ligandów.[12,14] Ich topologiczne własności oraz strukturalna mobilność pozwalają białkom na wspomaganie procesów wiązania poprzez dostosowanie się do indywidualnych partnerów, włączając w to transport wzdłuż wielu ścieżek wiązania do kieszeni wiążącej, rozpoznanie w kieszeni wiążacej oraz kierowanie ligandów przez kanały wyjściowe do rozpuszczalnika.[3] Przykładowo, wymienione trudności są kluczowe dla biofizyczneych procesów, takich jak sygnalizacja allosteryczna[15,16] oraz stabilizacja wywołana asocjacją kompleksu ligand–białko.[17] Czas życia liganda w kieszeni wiążącej zależy zarówno od liganda, jak i białka, co sugeruje, że specyficzność substratu i struktura tunelu nie mogą być pominięte w opisie rozpoznawania

liganda.[4,18]

Studia nad transportem ligandów w białkach są także ważne w kontekście odkrywania nowych leków, ponieważ odpowiedź białka na wiązanie liganda jest powiązana nie tylko z topologią ścieżek wiązania, ale także z termodynamiką oraz kinetyką takich procesów.[19,20] Po tym jak ligand osiąga miejsce wiązania, zostaje tam na określony okres, który jest kluczowy dla wszystkich regulacyjnych procesów w biologii.[21,22] Potrzeba zrozumienia jak ligandy migrują przez tunele białkowe wymusiła na badaczach stworzenie nowych metod obliczeniowych, takich jak sterowana dynamika molekularna,[23] lokalnie wzmocnione probkówanie[24,25] oraz metadynamika.[26–31]

Niniejsza praca doktorska jest poświęcona rozwojowi nowych metod symulacyjnych do optymalnej rekonstrukcji ścieżek transportu ligandów oraz profilów energii swobodnej wzdłuż tych scieżek. Metodologia zaprezentowana tutaj jest ogólna i może zostać użyta do symulowania innych procesów transportowych w materii biologicznej. Poniższe tezy są rozważone w niniejszej dyzertacji:

- Mechanizmy wiązania leżące u podstaw transportu liganda są związane z poziomem sprzężenia pomiędzy dynamiką białka a migracją liganda. Dynamika jest kluczowa dla rozpoznania sieci tuneli i kanałów w białkach;
- Uzwlędenienie wewnętrznej topologii białek (np. kieszeni, kanałów i tuneli) oraz ich termodynamiki jest konieczne do efektywnego modelowania optymalnych ścieżek transportu ligandów;
- Uczenie statystyczne (np. uczenie maszynowe oraz redukcja wymiarowości) oferuje metody do redukcji *big data* z symulacji do obliczenia ważnych zmiennych uogólnionych.

2 Introduction

This section introduces concepts underpinning MD and a new methodology to study ligand recognition and transport. This short review should serve as an introduction to methods used in the publications comprising this thesis. The main emphasis of this section is on ergodicity of a dynamical system, and biased sampling of rare events, particularly ligand transport and unbinding.

2.1 Dynamics

Consider a system of N point particles in $\mathbb{R}^d$ whose dynamics is given by the following equations:[32]

$$\begin{cases} \dot{r} = \nabla_p H(r,p) \\ \dot{p} = -\nabla_r H(r,p), \end{cases} \tag{2.1}$$

where $r \equiv (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N) \in \mathbb{R}^{dN}$ and $p \equiv (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N) \in \mathbb{R}^{dN}$ represent the positions and the momenta of the system, respectively. The smooth function $H : \Gamma \to \mathbb{R}$ is the Hamiltonian, and $\Gamma = \mathbb{R}^{dN} \times \mathbb{R}^{dN}$ denotes the phase space. The Hamiltonian has the form $H(r,p) = T(p) + U(r)$, where $T(p)$ is the kinetic energy, and $U(r)$ is the potential energy. $U(r)$ is smooth and typically assumed to be a pairwise radial function, which means that:

$$U(r) = \sum_{i=1}^{N} \sum_{j>i}^{N} U_{ij}(r_{ij}), \tag{2.2}$$

where $r_{ij} = \|r_i - r_j\|$ and $U_{ij} : \mathbb{R}_{>0} \to \mathbb{R}$. The potential energy used in MD simulations is typically the following:

$$\begin{aligned} U(r) = & \underbrace{\sum_i \frac{k_{\alpha_i}}{2}(\alpha_i - \alpha_{i_0})^2}_{\text{angle vibrations}} \\ & + \underbrace{\sum_i \frac{k_{d_i}}{2}(1 + \cos(n\phi_i - \phi_{i_0}))^2}_{\text{dihedral vibrations and rotations}} \\ & + \underbrace{\sum_i \frac{k_{r_i}}{2}(r_i - r_{i_0})^2}_{\text{bond vibrations}} \\ & + \underbrace{\sum_i \sum_{j>i} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}}_{\text{pairwise electrostatic interactions}} \\ & + \sum_i \sum_{j>i} 4\epsilon_{ij} \left[\underbrace{\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}}_{\text{Pauli repulsion}} - \underbrace{\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}}_{\text{van der Waals attraction}} \right]. \end{aligned} \tag{2.3}$$

The force acting on kth point particle is calculated from the partial potential energy $U_k(r)$ by $\boldsymbol{F}_k(r) = -\nabla_r U_k(r)$, where $U_k(r) = \sum_{j=1}^{N} U_{kj}(r_{kj})$ and the forces acting on each point particle are represented by $F = (\boldsymbol{F}_1, \ldots, \boldsymbol{F}_N) \in \mathbb{R}^{dN}$. This formula is used to calculate unbiased forces acting on the system during MD simulations.

2.2 Ergodicity

The equations of motion given by Eq. 2.1 conserve the total energy, i.e., $\mathrm{d}H/\mathrm{d}t = 0 \implies H(r) = const.$ Therefore, a trajectory (a chronological sequence of configurations for a system) will generate microscopic configurations belonging to a microcanonical ensemble with energy E and volume V. The probability density function of the microcanonical ensemble is given by:

$$P_{NVE}(r, p) = \frac{C_N}{Q(N, V, E)} \delta(H(r, p) - E), \tag{2.4}$$

where C_N is a dimensionless N-dependent factor that accounts for the nature of the particles. When the particles are identical $C_N = 1/N!h^{3N}$, where h is the Planck constant. If the system consists of many components with N_A particles of type A, N_B particles of B and so forth, and N total particles, C_N is equal to $1/h^{3N}N_A!N_B!\ldots$. The microcanonical partition function is denoted by $Q(N, V, E)$.

Another important probability measure is the canonical (or Gibbs) measure, given by the probability density function:

$$P_{NVT}(r,p) = \frac{C_N}{Q(N,V,T)} \exp(-\beta H(r,p)), \tag{2.5}$$

where $\beta > 0$ is the inverse temperature and $Q(N,V,T)$ is the canonical partition function.

The canonical measure accounts for the assembly of all microstates with fixed N and volume. The energy can fluctuate, however, and the system is kept at equilibrium by being in contact with a heat bath at temperature T (or the inverse temperature β). The microcanonical and canonical measures become equivalent asymptotically as N goes to infinity.[1]

Suppose the system with energy E is given an infinite amount of time to sample all configurations in the phase space characterized with the constant energy hypersurface. A system with such property displays ergodicity, which means that the sampled configurations can be used to generate a microcanonical ensemble needed to calculate its associated averages.

A basic procedure, given an ergodic trajectory generated by a Hamiltionian $H(r,p)$, is to replace microcanonical phase space averages by time averages over the trajectory according to:

$$\langle x \rangle \equiv \frac{\int \mathrm{d}r\mathrm{d}p\, x(r,p)\delta(H(r,p)-E)}{\int \mathrm{d}r\mathrm{d}p\, \delta(H(r,p)-E)} = \lim_{\tau\to\infty} \frac{1}{\tau} \int_0^\tau \mathrm{d}t\, x(r_t,p_t) \equiv \bar{x}, \tag{2.6}$$

which clearly states that the ergodic hypothesis will not hold if the sampling of the system is not sufficient either due to the limited computational time or the inaccessibility of configurations. For instance, if the potential energy $U(r)$ has high barriers, the dynamics of the system will be limited to regions where $U(r) < E$ which will result in separatrices in the phase space. Dynamical systems that obey the equivalence shown in Eq. 2.6 are ergodic.

In general, the ergodicity of a dynamical system is limited mainly locally, since the simulated system has often many degrees of freedom (proteins, glasses and polymers).[33] The average energy for a single degree of freedom at equilibrium is kT[a] if the system is able to equipartition the energy. Therefore, if a high energy barrier is intrinsically inseparable from a given mode of dynamics, a very long time is needed to promote barrier crossing through thermal fluctuations (Fig. 2.1).

The ergodicity problem is severe in systems where the process falls into the category of rare events, and thus, to study such problems unbiased MD methods are not enough. Biased MD, however, requires generalized coordinates describing the progress of the process under study.

[a] The result of the virial theorem, i.e., $\left\langle x_i \frac{\partial H}{\partial x_j} \right\rangle = kT\delta_{ij}$, where x_i and x_j are specific components of the phase space vector

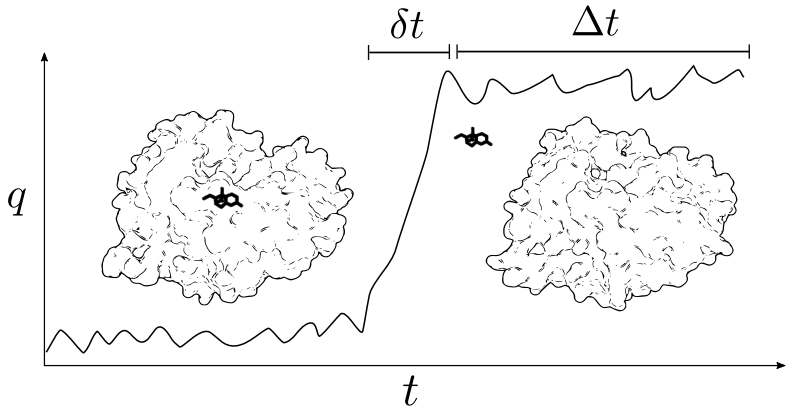

Figure 2.1: Schematic representation of the ligand–protein dynamics from the bound to the dissociated state. The transition between the two states becomes rare if the time scale of this event is much higher than the time scale of molecular motions, i.e., $\delta t/\Delta t \approx 0$.

2.3 Generalized Coordinates

Generalized coordinates monitor the progress of some chemical, mechanical, or thermodynamics process. Whenever referring to these generalized coordinates, they will be named as reaction coordinates (RCs), although other designations are also used in the literature; for instance, the term collective variable (CV) is common. There is no consensus in the nomenclature with some researchers referring to RCs as a one-dimensional CVs,[34,35] despite the contrast with naming a one-dimensional CV also a CV.[36–38] In this thesis, CVs are referred to as functions of RCs, which are in the current context a set of Cartesian coordinates. Examples of trivial CVs that can be used in modeling frequently observed events are root-mean-square distance, coordination number, dipole moment, etc.

If CVs are computed, MD can be used to recover free energies along the pathways (Sec. 2.4 Biased Sampling of Rare Events). Thus, for rare events such as ligand transport, calculating CVs describing the process optimally is very difficult. Particularly, the sampled trajectory should be optimal from the energetic point of view, since optimized RCs often correspond to a reaction as it is found experimentally. It is worth to mention that RCs should be used with care because a biased potential applied to a non-optimal pathway can drive the system in a misleading way, generating erroneous predictions of free energy barriers and kinetic rates.

Typical methods used to discover the reaction pathways (RPs) of ligand binding/unbinding include, but are not limited to, steered MD,[23] random acceleration MD,[39,40] or locally enhanced sampling.[24] RPs should not just measure the progress of generalized coordinates describing a molecular transition, but also be useful in quantifying the process in a methodical manner that would provide a basic understanding of the dynamical process.

The following subsection introduces a new biased MD method for finding and biasing RPs that can be used to describe processes like ligand association, dissociation, diffusion and migration.

2.3.1 Memetic Sampling

Memetic sampling (MS) proposed in this thesis exploits biased sampling of the configurational space of the ligand within the protein matrix to reconstruct Cartesian RPs, thereby overcoming the problem of transient curved tunnels/channels and overestimation of energy barriers and mean first passage times.[3,41–43]

In simple biasing procedures, the system is biased by an external harmonic potential, typically introduced as $\frac{k}{2}(vt - \boldsymbol{\delta}(r) \cdot \boldsymbol{n})^2$, where the deviation $\boldsymbol{\delta}$ is calculated as the difference between the current and the initial position of the biased subsystem, n is the biasing direction, t denotes time, k is a force constant, and v is the biasing velocity. In such methods, the biasing direction is either linear (steered MD[23]) or random (random acceleration MD[39,40]) which is very unlikely to be optimal in terms of interaction energy in proteins with curved channels and tunnels.[3] Therefore, in MS, the biasing direction $\boldsymbol{n}(\Lambda)$ is a function of the effective interaction energy Λ subject to minimization, $\boldsymbol{n}(\Lambda) : \min_r \Lambda(r)$.

In MS, biased trajectories are sampled by minimizing the effective interaction energy $\Lambda(r)$ between the ligand and the protein:

$$\Lambda(r) = V(r^{(\alpha)}, r^{(\beta)}) + \sum_{i=1}^{N_\alpha} \sum_{j>i}^{N_\beta} h_{ij} \exp\left(-\|\boldsymbol{r}_i^{(\alpha)} - \boldsymbol{r}_j^{(\beta)}\|^2 / 2\sigma^2\right) \tag{2.7}$$

comprises the van der Waals, electrostatic and hydrogen-bond interactions denoted jointly as $V(r^{(\alpha)}, r^{(\beta)})$ and the term describing the conformational flexibility of the biased ligand in the protein matrix. The notation $r^{(\alpha)}$ and $r^{(\beta)}$ indicates the positions of subsystems consisting of the ligand and the protein, respectively. In Eq. 2.7, the sum is iterated over all pairs of atoms between the subsystems. The Gaussian width is denoted as σ and its height $h_{ij} = s_i v_j + s_j v_i$ is given by the linear combination of solvation coefficients s and atomic volumes v.

The biasing procedure in MS[b] consists of multiple stages, each maintained during m time steps of an MD simulation. At the beginning of each stage, the biasing direction is recalculated by minimizing the effective interaction energy Λ in the neighborhood of the current position of the ligand; the end of each stage represents an intermediate state. For instance, the kth state is determined during MD as follows (Fig. 2.2):

1. The kth ligand conformation in the neighborhood of the $(k-1)$th ligand conformation is found by minimizing Λ. The neighborhood is assured by sampling ligand conformations from the $(k-1)$th conformation with a given cut-off;

2. The ligand is biased in m steps of the MD simulation in the direction of the kth ligand conformation;

[b] The algorithms published in Article I[41] are collectively termed MS, and although they differ in the optimization procedure to minimize the effective interaction energy (immune algorithms, simulated annealing, etc.), the biasing procedure is general.

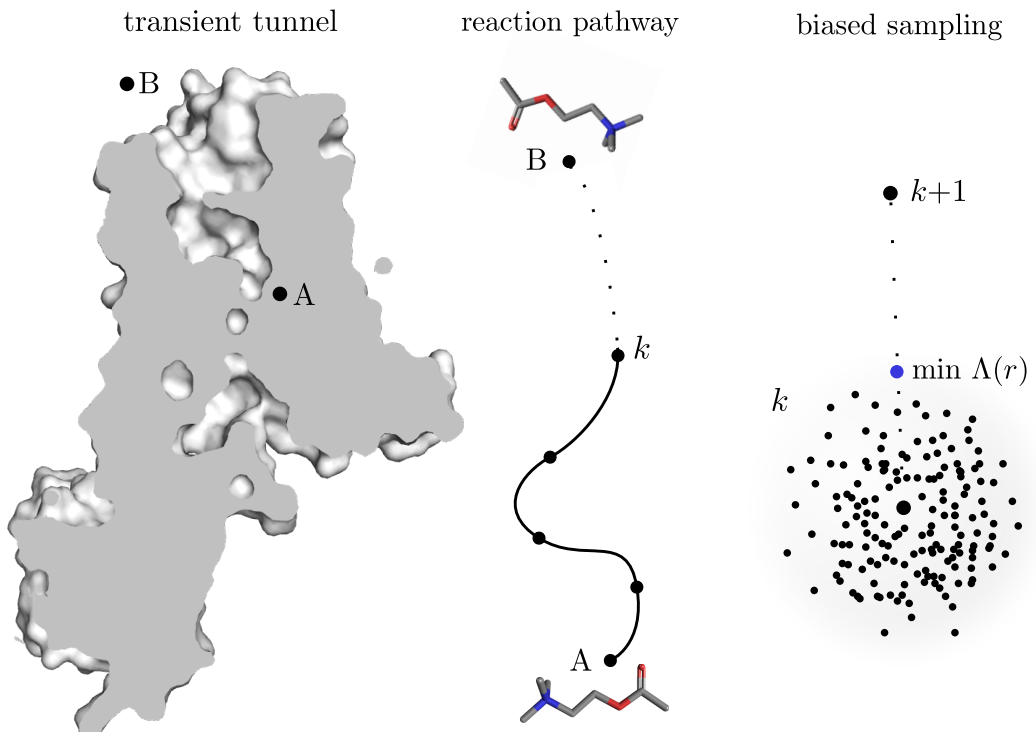

Figure 2.2: Sampling of ligand unbinding pathways using MS. As an example, the unbinding of acetycholine from M1 muscarinic receptor (PDB ID: 5cxv) is shown. The unbinding is initiated from the bound state (A) of the M1–acetylcholine complex, and ends once the ligand reaches solvent (B). The X-ray structure of M1 muscarinic receptor indicates that there is a transient tunnel with a narrow gorge along the exit route from A to B. The RP characterizing atomistically the unbinding along the exit tunnel is identified and sampled using MS. Namely, to determine the $(k+1)$th intermediate, the conformations of acetylcholine are sampled in the neighborhood of the kth intermediate. The optimal direction of biasing is calculated by minimizing the effective interaction energy $\Lambda(r)$ between M1 muscarinic receptor and acetylcholine.

3. After m steps of the MD simulation, the conformation of the system comprising the kth ligand conformation and the protein adjusted to the moving ligand, is saved as the kth intermediate state.

The sampling is initiated from the bound state of the system, and ends when the system dissociates (Fig. 2.2). For the optimization procedure, see Article I.[41] Although Eq. 2.7 represents the functional $\Lambda(r)$ used to sample optimal RPs in ligand–protein complexes, a functional for other processes (e.g., protein unfolding) can be used also, because the minimization procedure is general.

The optimal trajectory sampled by MS can be used as a RP, $q \equiv \{\boldsymbol{q}_i\}, i = 1, \ldots, M$, where M is the number of the optimal protein–ligand conformations $\boldsymbol{q}_i \in \mathbb{R}^{3N}$, i.e., the

bound state, the M-2 intermediate states, and the unbound state. To monitor the progress along and the distance from the RP, the path CVs[44,45] are introduced as follows:

$$s(r) = \lim_{\lambda\to\infty} \frac{\int_0^1 \mathrm{d}t\, t \exp(-\lambda\|r - q(t)\|^2)}{\int_0^1 \mathrm{d}t\, \exp(-\lambda\|r - q(t)\|^2)} \tag{2.8}$$

and

$$z(r) = -\lim_{\lambda\to\infty} \lambda^{-1} \ln \int_0^1 \mathrm{d}t\, \exp(-\lambda\|r - q(t)\|^2), \tag{2.9}$$

where a configuration of the system r is compared to the RP q using a metric $\|\cdot\|$. Introducing these variables is very helpful in reconstructing free energies along RPs.

The RPs of ligand transport are reconstructed in MS by a single sweep, as in string methods.[46–49] At the current stage of the development, the sampling is sufficient to find optimal RPs characterizing ligand recognition, but not to estimate thermodynamic and kinetic quantities. For this reason, MS was combined with other approaches described in detail in the next section.

2.4 Biased Sampling of Rare Events

Suppose a process of interest can be monitored by a subset of n CVs that are obtained from a transformation of the Cartesian RCs by $q_i = f_i(r)$, $i = 1, \ldots, n$. Then the probability density that these n CVs have values $q_i = \gamma_i$ in the canonical ensemble is:

$$P(\gamma) = \frac{C_N}{Q(N,V,T)} \int \mathrm{d}p\mathrm{d}r\, \exp\left(-\beta H(r,p)\right) \prod_{i=1}^{n} \delta(f_i(r) - \gamma_i), \tag{2.10}$$

where $\gamma = (\gamma_1, \ldots, \gamma_n)$, $\beta = 1/kT$ is the inverse temperature of the system, and the δ-functions are introduced to fix the CVs q_i at γ_i. The free-energy hypersurface associated with $P(\gamma)$ is given by:

$$F(\gamma) = -\beta^{-1} \ln P(\gamma). \tag{2.11}$$

Although the reconstruction of free energy is a very difficult task, there are approaches that can be used for this purpose, e.g., blue-moon ensemble methods,[48] umbrella sampling,[50] adiabatic dynamics,[51] and metadynamics.[31] Because a part of this thesis is devoted to estimating free energy along the RPs of ligand transport, and was created in the group of Prof. M. Parrinello, MS was joined with metadynamics.

2.4.1 Metadynamics

Metadynamics[52] is a method developed by Laio and Parrinello to enhance rare-event sampling and reconstruct a free-energy landscape of complex systems during simulations. In metady-

namics energy minima are filled in using a time-dependent biasing potential in form of the Gaussian functions. When a minimum is filled in, the system is driven into the next minima which are subsequently filled. This procedure is repeated until the entire energy landscape is flat, i.e., when the dynamics of the system is diffusive (Fig. 2.3). From the accumulated biasing potential, the free energy along the RPs can be estimated.

Consider Eq. 2.10 in the ensemble notation, i.e., where $\langle\cdot\rangle$ denotes the canonical average:

$$P(\gamma) = \left\langle \prod_{i=1}^{n} \delta(f_i(r) - \gamma_i) \right\rangle, \tag{2.12}$$

which can be replaced with a time average over a trajectory as shown in Eq. 2.6:

$$P(\gamma) = \lim_{\tau\to\infty} \frac{1}{\tau} \int_0^\tau \mathrm{d}t \prod_{i=1}^{n} \delta(f_i\left(r(t)\right) - \gamma_i), \tag{2.13}$$

under the assumption that the trajectory is ergodic. In metadynamics, the Dirac δ-functions are expressed as the limit of the Gaussian function as the width goes to 0, i.e., $\delta(x-a) = \lim_{\sigma\to 0}(2\pi\sigma^2)^{-1/2}\exp(-(x-a)^2/2\sigma^2)$. Using this expression, the time average from Eq. 2.13 becomes:

$$P(\gamma) = \lim_{\tau\to\infty}\lim_{\sigma\to 0} \frac{1}{\tau}\frac{1}{\sqrt{2\pi\sigma^2}} \int_0^\tau \mathrm{d}t \prod_{i=1}^{n} \exp\left[-\frac{\left(f_i\left(r(t)\right) - \gamma_i\right)^2}{2\sigma^2}\right], \tag{2.14}$$

which means that Eq. 2.14 can be used as an approximation to $P(\gamma)$ for sufficiently small Gaussian width to hold ergodicity. This relation is discretized usually as:

$$P(\gamma) \approx \frac{1}{\tau}\frac{1}{\sqrt{2\pi\sigma^2}} \sum_{k=0}^{K-1} \exp\left[-\sum_{i=1}^{n} \frac{\left(f_i\left(r(k\Delta t)\right) - \gamma_i\right)^2}{2\sigma^2}\right], \tag{2.15}$$

where Δt is the time step and the iteration over k goes from the beginning of a trajectory at time 0 to its end at time $(K-1)\Delta t$.

Metadynamics uses the discretized $P(\gamma)$ shown in Eq. 2.15 to enhance on-the-fly sampling of the configurational space represented by the RPs during simulations. Consider a bias potential of the following form

$$U_G(r;t) = w \sum_{t=\tau_G, 2\tau_G, \ldots,} \exp\left[-\sum_{i=1}^{n} \frac{\left(f_i\left(r_G(t)\right) - f_i(r)\right)^2}{2\sigma^2}\right], \tag{2.16}$$

where $r_G(t)$ is the time evolution of the Cartesian coordinates of the system under the action of the potential $U + U_G$ (a biased trajectory) up to time t, and τ_G is the time interval of Gaussian

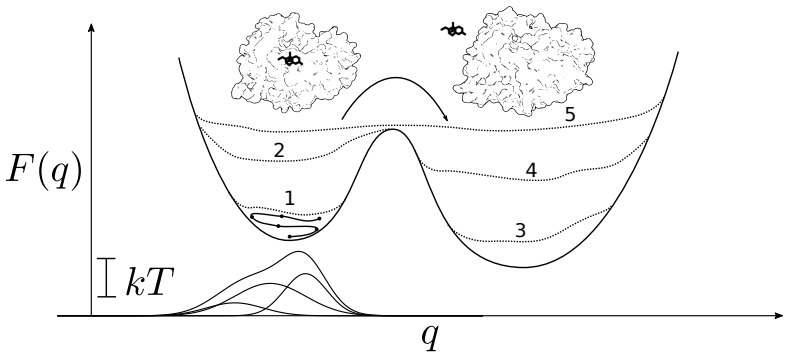

Figure 2.3: Schematic representation of the metadynamics bias potential accumulated after a short trajectory. The bias is built on-the-fly from accumulated repulsive Gaussians.

deposition. Therefore, the potential U is biased during the ongoing simulation.

Eq. 2.16 makes sense intuitively as follow. The Gaussians of height w and width σ are deposited along the CVs $f_i(r)$, $i = 1, \ldots, n$ at intervals τ_G, i.e., added to the unbiased potential U defined by a force field. For instance, if the system lies initially within a deep energy minimum ($\gg kT$), then by adding U_G the energy minimum is filled in, and the system elevated until it can escape to the next energy minimum. The procedure is stopped and converged when the energy landscape is flat, i.e., when the dynamics is diffusive.

Following Laio et al.,[53] Eq. 2.16 over a long trajectory (theoretically $t \to \infty$) converges to the free energy

$$F(q) \approx -w \sum_{t=\tau_G, 2\tau_G, \ldots,} \exp\left[-\sum_{i=1}^{n} \frac{\left(f_i\left(r_G(t)\right) - q_i\right)^2}{2\sigma^2}\right]. \tag{2.17}$$

In this dissertation, the free-energy profile is reconstructed to understand thermodynamics of ligand transport in proteins.

2.5 Implementation: maze

The MS method is implemented in the C++11 programming language. The program uses boost libraries and the Mersenne Twister random number generator.[72] MD simulations are performed by interfacing the program with NAMD 2.9 via the external program forces feature.[73] Every given number of time steps NAMD 2.9 calls `maze` to calculate the biasing potential acting on a ligand. `maze` implements also random acceleration MD for sampling random pathways, and simulated annealing and swarm optimization for sampling of ligand transport pathways. Apart from Article I where these methods were used as reference pathways, the simulations using MS (Articles I–V) were done via memetic algorithms. `maze` is available from the author upon request.

3 Conclusions

Protein conformational flexibility modulates kinetics and thermodynamics of ligand recognition and transport.[5] The degree of coupling between protein dynamics and ligand migration is responsible for mechanisms underlying these phenomena. Although there are many methods for the prediction of protein tunnels based on a single structure,[55] in contrast to some opinions[a], protein dynamics is necessary to observe conformational flexibility of transient routes.[41,54] Unless these methods are used on an ensemble of protein structures, no evidence showing that tunnels reconstructed geometrically represent an ergodic ensemble of ligand–protein structures can be provided. The MS biasing procedure introduced in this dissertation is able to find transient protein tunnels and probe the RPs of ligand transport during MD simulations, thereby alleviating the problem of insufficient sampling of ligand–protein conformations. The code for the MS simulations is available upon request.

Moreover, many proteins exhibit internal topological features with deeply buried binding sites, and multiple curved exit routes, in contrast to proteins, such as HIV proteases, in which the binding site is relatively exposed on the protein surface, as underlined in Articles III and IV. For instance, the cytochrome P450 family is known for multiple transient migration tunnels.[39,56,57] In Article I, MS succeeded in finding complex exit routes for cytochrome P450cam, and identifying additional pathways for camphor that have been not yet discussed.[41,43] The free-energy profiles along new camphor unbinding pathways enabled us to suggest that cytochrome P450cam may exhibit pathway hopping,[58] a phenomenon based on jumping between possible ligand binding pathways characterized by similar free energies of binding intermediates. This was previously suggested for protein folding.[59]

Analyzing the 'big data' from simulations is virtually impossible to perform without statistical learning (e.g., machine learning and dimensionality reduction). Such methods are used usually to reduce dimensionality of the studied physical problem, e.g., ligand transport, and to mine or construct important generalized coordinates.[60] Currently, statistical learning is becoming more popular in simulations.[36,37,61] In the Articles II and V,[42,43] machine-learning techniques allowed us to successfully reduce the dimensionality of camphor migration in cytochrome P450cam, and to propose new collective variables in form of coarse trajectories

[a] A private communication from Prof. M. Cieplak.

that may be used to analyze the process of ligand migration and to reconstruct free energy.

The recent developments in predicting transient ligand tunnels in proteins,[3,6,41,42,62] their thermodynamics[43,63] and kinetics[5,12,58,64–66] encouraged Magistrato to term these studies "the next-generation frontier of computational biology."[67–71] This may come true because free-energy related quantities depend on short-lived intermediate states, making them demanding to observe via X-ray crystallography or NMR. This makes RPs and thermodynamic quantities indispensable information for ligand transport, and any rare event in general.

One of the greatest limitations of the current methodology for studying ligand transport processes is its inability to sample via nonadiabatic MD simulations, as explained in detail in Article VI. The evolution of a ligand–protein complex along multiple diabatic energy curves is nearly impossible, because the size of such systems makes it inefficient to model using quantum dynamics. There are, however, classical approaches suitable for modeling photoexcitation, photodissociation, etc. In Article VII, using the Landau-Zener model, we successfully sampled CO diffusion pathways in a mutated neuroglobin, recently suggested to act like a CO scavenger. The results indicated that the degree of coupling between the diffusing ligand and the dynamics of neuroglobin binding pocket cannot be ignored in quantifying ligand photodissociation.

The theses postulated in this dissertation (Abstract) are strongly supported by Articles I–VII.

Bibliography

[1] M. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford University Press, 2010.

[2] D. Chandler. *Introduction to Modern Statistical Mechanics* Oxford University Press, 1987.

[3] J. Rydzewski and W. Nowak. Ligand Diffusion in Proteins via Enhanced Sampling in Molecular Dynamics. *Phys. Life Rev.* 22:58, 2017.

[4] R. Elber. Ligand Diffusion in Globins: Simulations versus Experiment. *Curr. Op. Struct. Biol.* 20:162, 2010.

[5] M. Amaral, D. Kokh, J. Bomke, A. Wegener, H. Buchstaller, H. Eggenweiler, P. Matias, C. Sirrenberg, R. Wade, and M. Frech. Protein Conformational Flexibility Modulates Kinetics and Thermodynamics of Drug Binding. *Nat. Comm.* 8:2276, 2017.

[6] N. Bruce, G. Ganotra, D. Kokh, S. Sadiq, and R. C. Wade. New Approaches for Computing Ligand–Receptor Binding Kinetics. *Curr. Op. Struct. Biol.* 49:1, 2018.

[7] Z. Ren, S.-I. Adachi, W. Schildkamp, D. Bourgeois, M. Wulff, and K. Moffat. Photolysis of the Carbon Monoxide Complex of Myoglobin: Nanosecond Time-Resolved Crystallography. *Science* 274:1726, 1996.

[8] A. Rupenyan, J. Commandeur, and M. L. Groot. CO Photodissociation Dynamics in Cytochrome P450bm3 Studied by Subpicosecond Visible and Mid-Infrared Spectroscopy. *Biochemistry* 48:6104, 2009.

[9] S. Kim and M. Lim. Protein Conformation-Induced Modulation of Ligand Binding Kinetics: A Femtosecond Mid-IR Study of Nitric Oxide Binding Trajectories in Myoglobin. *J. Am. Chem. Soc.* 127:8908, 2005.

[10] C. Tetreau, L. Mouawad, S. Murail, P. Duchambon, Y. Blouquit, and D. Lavalette. Disentangling Ligand Migration and Heme Pocket Relaxation in Cytochrome P450cam. *Biophys. J.* 88(2):1250–1263, 2005.

[11] R. C. Wade, P. J. Winn, I. Schlichting. A Survey of Active Site Access Channels in Cytochromes P450. *J. Inorg. Biochem.* 98:1175, 2004.

[12] A. Stank, D. B. Kokh, J. C. Fuller, and R. C. Wade. Protein Binding Pocket Dynamics. *Acc. Chem. Res.* 49:809, 2016.

[13] J. Ikebe, K. Umezawa, and J. Higo. Enhanced Sampling Simulations to Construct Free-Energy Landscape of Protein–Partner Substrate Interaction. *Biophys. Rev.* 8:45, 2016.

[14] R. Baron and J. McCammon. Molecular Recognition and Ligand Association. *Annu. Rev. Phys. Chem.* 64:151, 2013.

[15] J.-P. Changeux and S. Edelstein. Allosteric Mechanisms of Signal Transduction. *Science* 308:1424, 2005.

[16] S. Lu, S. Li, and J. Zhang. Harnessing Allostery: A Novel Approach to Drug Discovery. *Med. Res. Rev.* 34:1242, 2014.

[17] A. Moorhouse, A. Santos, M. Gunaratnam, M. Moore, S. Neidle, and J. Moses. Stabilization of G-Quadruplex DNA by Highly Selective Ligands via Click Chemistry. *J. Am. Chem. Soc.* 128:15972, 2006.

[18] D. L. Mobley and K. A. Dill. Binding of Small-Molecule Ligands to Proteins: "What You See" is Not Always "What You Get". *Structure* 17:489, 2009.

[19] R. Zhang and F. Monsma. Binding Kinetics and Mechanism of Action: Toward the Discovery and Development of Better and Best in Class Drugs. *Expert Opin. Drug Discov.* 5:1023, 2010.

[20] D. C. Swinney. Biochemical Mechanisms of Drug Action: What Does it Take for Success? *Nat. Rev. Drug Discov.* 3:801, 2004.

[21] F. Bai, Y. Xu, J. Chen, Q. Liu, J. Gu, X. Wang, J. Ma, H. Li, J. N. Onuchic, and H. Jiang. Free Energy Landscape for the Binding Process of Huperzine A to Acetylcholinesterase. *Proc. Natl. Acad. Sci. U.S.A.* 110:4273, 2013.

[22] R. A. Copeland, D. L. Pompliano, and T. D. Meek. Drug–Target Residence Time and its Implications for Lead Optimization. *Nat. Rev. Drug Discov.* 5:730, 2006.

[23] H. Grubmüller, B. Heymann, and P. Tavan. Ligand Binding: Molecular Mechanics Calculation of the Streptavidin-Biotin Rupture Force. *Science* 271:997, 1996.

[24] R. Elber and M. Karplus. Enhanced Sampling in Molecular Dynamics: Use of the Time-Dependent Hartree Approximation for a Simulation of Carbon Monoxide Diffusion Through Myoglobin. *J. Am. Chem. Soc.* 112:9161, 1990.

[25] W. Nowak, R. Czerminski, and R. Elber. Reaction Path Study of Ligand Diffusion in Proteins: Application of the Self Penalty Walk (SPW) Method to Calculate Reaction Coordinates for the Motion of CO Through Leghemoglobin. *J. Am. Chem. Soc.* 113:5627, 1991.

[26] G. Bussi, A. Laio, and M. Parrinello. Equilibrium Free Energies from Nonequilibrium Metadynamics. *Phys. Rev. Lett.* 96:090601, 2006.

[27] L. Sutto, S. Marsili, and F. L. Gervasio. New Advances in Metadynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 2:771, 2012.

[28] N. Bešker and F. L. Gervasio. Using Metadynamics and Path Collective Variables to Study Ligand Binding and Induced Conformational Transitions. *Methods Mol. Biol.* 819:501, 2012.

[29] A. Barducci, M. Bonomi, and M. Parrinello. Metadynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 1:826, 2011.

[30] A. Laio and F. L. Gervasio. Metadynamics: A Method to Simulate Rare Events and Reconstruct the Free Energy in Biophysics, Chemistry and Material Science. *Rep. Prog. Phys.* 71:126601, 2008.

[31] O. Valsson, P. Tiwary, and M. Parrinello. Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annu. Rev. Phys. Chem.* 67:159, 2016.

[32] W. Rubinowicz and W. Królikowski. *Theoretical Mechanics. (Polish).* Państwowe Wydawnictwo Naukowe, 2000.

[33] R. Elber and M. Karplus. Multiple Conformational States of Proteins: A Molecular Dynamics Analysis of Myoglobin. *Science* 235:318, 1987.

[34] R. Elber, J. M. Bello-Rivas, P. Ma, A. E. Cardenas, and A. Fathizadeh. Calculating Iso-Committor Surfaces as Optimal Reaction Coordinates with Milestoning. *Entropy* 19:219, 2017.

[35] R. Elber. A New Paradigm for Atomically Detailed Simulations of Kinetics in Biophysical Systems. *Q. Rev. Biophys.* 50:8, 2017.

[36] A. Ardevol, G. A. Tribello, M. Ceriotti, and M. Parrinello. Probing the Unfolded Configurations of a β-Hairpin using Sketch-Map. *J. Chem. Theory Comput.* 11:1086, 2015.

[37] M. Ceriotti, G. Tribello, and M. Parrinello. Simplifying the Representation of Complex Free-Energy Landscapes using Sketch-Map. *Proc. Natl. Acad. Sci. U.S.A.* 108:13023, 2011.

[38] G. A. Tribello, M. Ceriotti, and M. Parrinello. Using Sketch-Map Coordinates to Analyze and Bias Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U.S.A.* 109:5196, 2012.

[39] S. K. Lüdemann, V. Lounnas, and R. C. Wade. How Do Substrates Enter and Products Exit the Buried Active Site of Cytochrome P450cam? 1. Random Expulsion Molecular Dynamics Investigation of Ligand Access Channels and Mechanisms. *J. Mol. Biol.* 303:797, 2000.

[40] S. K. Lüdemann, V. Lounnas, and R. C. Wade. How Do Substrates Enter and Products Exit the Buried Active Site of Cytochrome P450cam? 2. Steered Molecular Dynamics and Adiabatic Mapping of Substrate Pathways. *J. Mol. Biol.* 303:813, 2000.

[41] J. Rydzewski and W. Nowak. Memetic Algorithms for Ligand Expulsion from Protein Cavities. *J. Chem. Phys.* 143:124101, 2015.

[42] J. Rydzewski and W. Nowak. Machine Learning Based Dimensionality Reduction Facilitates Ligand Diffusion Paths Assessment: A Case of Cytochrome P450cam. *J. Chem. Theory Comput.* 12:2110, 2016.

[43] J. Rydzewski and W. Nowak. Thermodynamics of Camphor Migration in Cytochrome P450cam by Atomistic Simulations. *Scientific Reports* 7:7736, 2017.

[44] D. Branduardi, F. Gervasio, and M. Parrinello. From A to B in Free Energy Space. *J. Chem. Phys.* 126:054103, 2007.

[45] J. Fidelak, J. Juraszek, D. Branduardi, M. Bianciotto, and F. L. Gervasio. Free-Energy-Based Methods for Binding Profile Determination in a congeneric Series of CDK2 Inhibitors. *J. Phys. Chem. B* 114:9516, 2010.

[46] E. Weinan, W. Ren, and E. Vanden-Eijnden. Finite Temperature String Method for the Study of Rare Events. *J. Phys. Chem. B* 109:6688, 2005.

[47] L. Maragliano and E. Vanden-Eijnden. On-the-Fly String Method for Minimum Free Energy Paths Calculation. *Chem. Phys. Lett.* 446:182, 2007.

[48] E. Weinan, W. Ren, and E. Vanden-Eijnden. String Method for the Study of Rare Events. *Phys. Rev. B* 66:052301, 2002.

[49] L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti. String Method in Collective Variables: Minimum Free Energy Paths and Isocommittor Surfaces. *J. Chem. Phys.* 125:024106, 2006.

[50] G. M. Torrie and J. P. Valleau. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* 23:187, 1977.

[51] L. Rosso, P. Mináry, Z. Zhu, and M. E. Tuckerman. On the Use of the Adiabatic Molecular Dynamics Technique in the Calculation of Free Energy Profiles. *J. Chem. Phys.* 116:4389, 2002.

[52] A. Laio and M. Parrinello. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U.S.A.* 99:12562, 2002.

[53] A. Laio, A. Rodriguez-Fortea, F. L. Gervasio, M. Ceccarelli, and M. Parrinello. Assessing the Accuracy of Metadynamics. *J. Phys. Chem. B* 109:6714, 2005.

[54] E. Chovancova, A. Pavelka, P. Benes, O. Strnad, J. Brezovsky, B. Kozlikova, A. Gora, V. Sustr, M. Klvana, P. Medek. Caver 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLoS Comput. Biol.* 8:e1002708, 2012.

[55] M. Petřek, P. Košinová, J. Koča, and M. Otyepka. Mole: A Voronoi Diagram-Based Explorer of Molecular Channels, Pores, and Tunnels. *Structure* 15:1357, 2007.

[56] V. Cojocaru, P. J. Winn, and R. C. Wade. The Ins and Outs of Cytochrome P450s. *Biochim. Biophys. Acta* 1770:390, 2007.

[57] P. J. Winn, S. K. Lüdemann, R. Gauges, V. Lounnas, and R. C. Wade. Comparison of the Dynamics of Substrate Access Channels in Three Cytochrome P450s Reveals Different Opening Mechanisms and a Novel Functional Role for a Buried Arginine. *Proc. Natl. Acad. Sci. U.S.A.* 99:5361, 2002.

[58] S. D. Lotz and A. Dickson. Unbiased Molecular Dynamics of 11 min Timescale Drug Unbinding Reveals Transition State Stabilizing Interactions. *J. Am. Chem. Soc.* 140:618, 2018.

[59] J. S. Butler and S. N. Loh. Kinetic Partitioning During Folding of the P53 DNA Binding Domain. *J. Mol. Biol.* 350:906, 2005.

[60] M. A. Rohrdanz, W. Zheng, and C. Clementi. Discovering Mountain Passes via Torchlight: Methods for the Definition of Reaction Coordinates and Pathways in Complex Macromolecular Reactions. *Annu. Rev. Phys. Chem.* 64:295, 2013.

[61] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande. Moleculenet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* 9:513, 2018.

[62] Y. Wang, E. Papaleo, and K. Lindorff-Larsen. Mapping Transiently Formed and Sparsely Populated Conformations on a Complex Energy Landscape. *eLife* 5:e17505, 2016.

[63] A. Magistrato, J. Sgrignani, R. Krause, and A. Cavalli. Single or Multiple Access Channels to the Cyp450s Active Site? An Answer from Free Energy Simulations of the Human Aromatase Enzyme. *J. Phys. Chem. Lett.* 8:2036, 2017.

[64] B. Voß, R. Seifert, U. B. Kaupp, and H. Grubmüller. A Quantitative Model for CAMP Binding to the Binding Domain of MLOK1. *Biophys. J.* 111:1668, 2016.

[65] A. Dickson and S. Lotz. Multiple Ligand Unbinding Pathways and Ligand-Induced Destabilization Revealed by WExplore. *Biophys. J.* 112:620, 2017.

[66] A. Dickson, P. Tiwary, and H. Vashisth. Kinetics of Ligand Binding Through Advanced Computational Approaches: A Review. *Curr. Top. Med. Chem.* 17:2626, 2017.

[67] A. Magistrato. Direct in Silico Visualization of Ligands Channelling Through Proteins: The Next-Generation Frontier of Computational Biology: Comment on "Ligand Diffusion via Enhanced Sampling Molecular Dynamics" by J. Rydzewski and W Nowak. *Phys. Life Rev.* 22:82, 2017.

[68] J. Rydzewski and W. Nowak. Rare-event Sampling in Ligand Diffusion: Reply to Comments on "Ligand Diffusion in Proteins via Enhanced Sampling in Molecular Dynamics". *Phys. Life Rev.* 22:85, 2017.

[69] A. Kolinski. Toward More Efficient Simulations of Slow Processes in Large Biomolecular Systems: Comment on "Ligand Diffusion in Proteins via Enhanced Sampling in Molecular Dynamics" by J. Rydzewski and W. Nowak. *Phys. Life Rev.* 22:75, 2017.

[70] K. Kuczera. Finding Optimal Paths Through Biomolecular Mazes: Comment on: "Lgand Diffusion in Proteins via Enhanced Sampling in Molecular Dynamics" by J. Rydzewski and W. Nowak. *Phys. Life Rev.* 22:77, 2017.

[71] M. S. Li. Ligand Migration and Steered Molecular Dynamics in Drug Discovery: Comment on "Ligand Diffusion in Proteins via Enhanced Sampling in Molecular Dynamics" by J. Rydzewski and W. Nowak. *Phys. Life Rev.* 22:79, 2017.

[72] M. Matsumoto, and T. Nishimura. Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator. *ACM T. Model. Comput. S.* 8:3, 1998.

[73] J. C. Phillips, C. James, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, Klaus. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* 26:1781, 2005.

4 Publications

Published Articles Comprising the Thesis

I. J. Rydzewski, W. Nowak. Memetic Algorithms for Ligand Expulsion from Protein Cavities. *J. Chem. Phys.* 143:124101, 2015.

II. J. Rydzewski, W. Nowak. Machine Learning Based Dimensionality Reduction Facilitates Ligand Diffusion Paths Assessment: A Case of Cytochrome P450cam. *J. Chem. Theory Comp.* 12:2110, 2016.

III. J. Rydzewski, W. Nowak. Ligand Transport in Proteins via Enhanced Sampling in Molecular Dynamics. *Phys. Life Rev.* 22:58, 2017.

IV. J. Rydzewski, W. Nowak. Rare-Event Sampling in Ligand Diffusion. *Phys. Life Rev.* 22:85, 2017.

V. J. Rydzewski, W. Nowak. Thermodynamics of Camphor Migration in Cytochrome P450cam by Atomistic Simulations. *Sci. Rep.* 7:7736, 2017.

VI. J. Rydzewski, W. Nowak. Molecular Dynamics Simulations of Large Systems in Electronic Excited States. *Handbook of Computational Chemistry.* Springer Berlin, 2016.

VII. J. Rydzewski, W. Nowak. Photoinduced Transport in an H64Q Neuroglobin Antidote for Carbon Monoxide Poisoning. *J. Chem. Phys.* 148, 2018.

Manuscripts Related to the Thesis under Review

VIII. J. Rydzewski, R. Jakubowski, H. Grubmüller, W. Nowak. Kinetics of Huperzine A Dissociation from Acetylcholinesterase via Multiple Unbinding Pathways.

IX. J. Rydzewski. Conformational Collapse of the Distal Heme Binding Pocket in Neuroglobin during Ligand Diffusion.

Other Published Articles

X. J. Rydzewski, R. Jakubowski, G. Nicosia, W. Nowak. Conformational Sampling of a Biomolecular Rugged Energy Landscape. *IEEE/ACM Trans. Comput. Biol. Bioinform.* PP:1, 2017.

XI. J. Rydzewski, W. Nowak, G. Nicosia. Inferring Pathological States in Cortical Neuron Microcircuits. *J. Theor. Biol.* 386:34, 2015.

XII. J. Rydzewski, R. Jakubowski, W. Nowak. Communication: Entropic Measure to Prevent Energy Over-Minimization in Molecular Dynamics Simulations. *J. Chem. Phys.* 143:171103, 2015.

XIII. J. Rydzewski, W. Strzałka, W. Nowak. Nanomechanics of PCNA: A Protein-Made DNA Sliding Clamp. *Chem. Phys. Lett.* 634:263, 2015.

4.1 Summary of the Publications

A transient nature of protein channels/tunnels often leads to difficulties in modeling ligand transport pathways using MD. In Article I, the MS method to sample ligand unbinding pathways, and to explore protein cavities is proposed. MS is tested on three proteins with an increasing complexity of channels/tunnels: M2 muscarinic G-protein-coupled receptor, nitrile hydratase, and cytochrome P450cam. In each case, MS outperforms standard MD methods that have been used to sample ligand unbinding pathways so far. In the most difficult case, MS predicts a pathway for camphor that has not been discussed so far.

Selecting CVs to describe quantitatively physical processes is indispensable in analyzing high-dimensional data, for instance, from the MS simulations of ligand transport and recognition. Following the research presented in Article I, a nonlinear dimensionality-reduction method is applied to reduce the high-dimensional configuration space of ligand–protein dissociation and calculate new low-dimensional coarse pathways (Article II). This mapping retains main conformational changes that occur during dissociation. The topological similarity of the coarse pathways is studied using the Fréchet metric, which results in facilitating machine-learning classification of the camphor dissociation pathways in cytochrome P450cam.

Article III presents a modern review of ligand transport processes in heterogeneous media. The main emphasis is on the dynamics of protein channels and tunnels, which pose difficulties in both experiments and simulations. Article III reviews the current literature and introduces the recent methodology to reconstruct the RPs of ligand diffusion and the free-energy profiles along the RPs using biased sampling of the conformational space. Presented methods are illustrated on several ligand–protein systems, for instance, cytochromes and G-protein-coupled receptors. Article IV and the comments on Article III (Refs. 67–71) show main problems in the modeling of ligand transport, and possible applications in biophysics to come.

In Article V, the MD methodology introduced in Articles I–IV is used to study the

camphor binding to cytochrome P450cam. The RPs of camphor unbinding from the active site of cytochrome P450cam to solvent are observed via three egress routes during microsecond-long simulations. Analysis of the simulations is facilitated by a machine-learning technique. Metadynamics is used to estimate free energies along the RPs and indicate diverse camphor binding configurations to cytochrome P450cam. The results suggest that the unbinding of camphor along the pathway near the substrate recognition site is preferred thermodynamically. The corresponding conformational change is characterized by the retraction of the F and G helices and the disorder of the B' helix. These results are corroborated by experimental studies and provide detailed insight into the ligand binding and conformational behavior of the cytochrome family.

The difficulty of modeling ligand transport increases tremendously if a ligand–protein complex is able to evolve along multiple diabatic energy curves. Although studying such large biological systems is nearly impossible using quantum dynamics, nonadiabatic MD for ligand transport would be of high interest to study at the atomic level, e.g., in optogenetics. Article VI introduces many concepts (e.g., nonadiabatic MD, conical intersection) important for photoexcited ligand transport in proteins, which are used to study CO diffusion within neuroglobin (Article VII).

CO is a leading cause of poisoning deaths worldwide, without available antidotal therapy. Binding of CO by neuroglobin (Ngb) with a mutated distal histidine was recently proposed as a potential therapy for CO poisoning. In Article VII, an atomistic mechanism of CO diffusion in H64Q Ngb is revealed by nonadiabatic MD simulations. Our results demonstrate that the distribution of CO within the proteins differs substantially because of the rearrangement of amino acids surrounding the distal heme pocket. The mutation leads to the shortening of the time scale of CO geminate recombination, making H64Q Ngb 2.7 times more frequent binder than WT Ngb.

4.2 Declaration of Contribution

Date: March 12, 2018

Prof. dr hab. Wiesław Nowak
Department of Biophysics and Medical Physics
Institute of Physics
Nicolaus Copernicus University
E-mail: wiesiek@fizyka.umk.pl

Statement

In the following articles:

I. J. Rydzewski, W. Nowak. Memetic Algorithms for Ligand Expulsion from Protein Cavities. *J. Chem. Phys.* 143:124101, 2015.

II. J. Rydzewski, W. Nowak. Machine Learning Based Dimensionality Reduction Facilitates Ligand Diffusion Paths Assessment: A Case of Cytochrome P450cam. *J. Chem. Theory Comp.* 12:2110, 2016.

III. J. Rydzewski, W. Nowak. Ligand Transport in Proteins via Enhanced Sampling in Molecular Dynamics. *Phys. Life Rev.* 22:58, 2017.

IV. J. Rydzewski, W. Nowak. Rare-Event Sampling in Ligand Diffusion. *Phys. Life Rev.* 22:85, 2017.

V. J. Rydzewski, W. Nowak. Thermodynamics of Camphor Migration in Cytochrome P450cam by Atomistic Simulations. *Sci. Rep.* 7:7736, 2017.

VI. J. Rydzewski, W. Nowak. Molecular Dynamics Simulations of Large Systems in Electronic Excited States. *Handbook of Computational Chemistry.* Springer Berlin, 2016.

VII. J. Rydzewski, W. Nowak. Photoinduced Transport in an H64Q Neuroglobin Antidote for Carbon Monoxide Poisoning. *J. Chem. Phys.* 148, 2018,

presented as the Ph.D. thesis by Jakub Rydzewski, my contribution consisted of a discussion of the scientific problems undertaken in the dissertation, including methods, results, manuscripts, and peer review. My contribution to every article does not exceed 10%.

W. Nowak