# Enhanced Conformational Sampling of Proteins Using TEE-REX

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen

Vorgelegt von
**Marcus Kubitzki**
aus Rottweil

Göttingen 2007

D7
Referent: Prof. Dr. Helmut Grubmüller
Koreferent: Prof. Dr. Tim Salditt
Tag der mündlichen Prüfung: 11.12.2007

# Contents

# 1 Introduction

*The beginner... should not be discouraged if... he finds that he does not have the prerequisites for reading the prerequisites.*
— Paul Halmos

In all living organisms, a large diversity of processes critically depend on the activity of proteins, biological macromolecules which are mainly composed of polymeric chains of amino acids [1]. Although in many of these processes the mere *structure* of a protein dominates its function (e.g. collagen in tissues, $\alpha$-keratin in hair, histones as spools around which DNA winds), protein *dynamics* is fundamental to many others. Virtually all biological processes involving motion find their origin in protein dynamics, i.e. the proteins' ability to adopt different conformations. Muscle contraction, for instance, is based on the combined action of actin and myosin. Other examples are the molecular motors kinesin and F1-ATPase. The inherent conformational flexibility of proteins is not only restricted to mobility as the primary function, but is essential for the function of many transport proteins, proteins involved in signal transduction, cellular recognition (e.g. in the immune system), and numerous enzymes [1]. In many enzymes, conformational changes serve to enclose the substrate, thereby preventing its release from the protein and ideally positioning it for the protein to perform its function, as e.g. in lysozyme. Allosteric proteins, such as hemoglobin in red blood cells, employ another special class of conformational transitions. Here, substrate binding to one subunit of these multimeric proteins triggers a conformational change that alters the substrate affinity of the other subunits, thereby sharpening the switching response of these proteins. Generally, in allosteric proteins binding of an effector molecule alters, via certain conformational changes, the binding affinity of one or more binding sites which are distinct from the effector binding site.

The range of conformational change encountered in nature varies from very subtle, local changes, as in the case of myoglobin, to global conformational changes, involving motions of significant amplitude for large parts of a protein such as adenylate kinase or importin-$\beta$. Furthermore, dynamics not only plays an important role in the functional

native state of many proteins, but also the mechanism by which proteins reach that native conformation, the folding process, is of course a dynamic one.

Experimental techniques have made substantial progress in revealing protein structure (especially X-ray crystallography [2, 3], nuclear magnetic resonance (NMR) spectroscopy [4, 5], cryo-electron microscopy [6]) and conformational flexibility (e.g. NMR relaxation [7], fluorescence spectroscopy [8], electron paramagnetic resonance [9], neutron scattering [10, 11]). In some instances, different functional states of proteins were structurally characterized by trapping them in certain substates [12]. Furthermore, time-resolved X-ray crystallography [13, 14] allows to follow the conformational protein motion with picoseconds time resolution. Wide-spread use of this technique is impeded, though, by the massive experimental effort involved.

Despite this enormous variety, experimental techniques having spatio-temporal resolution in the nanosecond as well as the nanometer regime are not available, and thus information on the conformational space accessible to proteins *in vivo* often remains obscure. In particular, details on the pathways between different known conformations are usually unknown. Computer simulation techniques provide the only possibility to obtain dynamic information on proteins at atomic resolution in the nanosecond to microsecond time range. Out of all possible ways of simulating protein motions, molecular dynamics (MD) techniques are among the most popular. MD tries to describe the time evolution of molecular systems as realistically as possible. In a typical simulation, an experimentally determined configuration is put into an environment that best mimics its natural environment. Once started, the time evolution of the system is described by integrating Newton's equations of motion for all atoms, treated as point masses interacting via simple force terms. The method operates in the full $3N$ dimensional configuration space of the protein and the surrounding solvent molecules (where $N$ is the number of atoms). The large number of pairwise interactions to be evaluated and the short femtosecond time steps enforced by the fastest motions (O-H bond vibrations) entail very long computation times, limiting MD at present to systems of $10^5$-$10^6$ atoms and to timescales of several 100 ns. Apart from a few exceptions, however, relevant biological processes such as the gating of ion channels, allosteric interactions, ligand binding, enzymatic activity or protein folding occur on the microsecond to seconds timescale, and still remain out of reach for conventional MD.

An efficient exploration of the vast configuration space spanned by all molecular conformations, therefore, proves to be a challenging endeavor. The numerous interactions present in the system give rise to a complex $3N$-dimensional rugged free energy land-

scape [15, 16] whose global shape is supposed to be funnel-like with the native state populating the global minimum [17]. A more detailed look reveals a multitude of almost iso-energetic minima separated by energy barriers of various heights. Each of these minima corresponds to one particular conformational substate, with neighboring minima corresponding to similar conformations. Within this picture, structural transitions are barrier crossings, with the transition rate depending on the height of the barrier. The largest barriers are rarely traversed, and thus are hardly observed in MD simulations under standard conditions (*sampling problem*).

Efforts to bridge the gap towards all-atom simulations on biologically relevant length and time scales are manifold and numerous algorithms have been developed to enhance conformational sampling (see [18, 19] for recent reviews). Conceptually, three categories of methods can be distinguished: (1) those that try to mimic biological systems as realistically as possible and focus on sophisticated (mathematical) methods to enhance computational efficiency, affecting the dynamics and thermodynamics as little as possible, (2) those that gain computation time by simplifying the molecular models involved, and (3) those algorithms that make use of special properties of the simulated system to describe the latter in more appropriate, internal coordinates. The above division is not exclusive and some methods cannot be assigned to either category whereas others are hybrid methods based on principles from several categories. A number of examples from all categories will be treated in the following paragraphs before introducing two methods from the first and third category, namely replica exchange and essential dynamics, that play a key role throughout this thesis.

Algorithms to speed up the core MD algorithm, especially the calculation of long-range Coulomb forces, belong to the first category. Besides parallelization, recent developments include efficient methods such as multiple time step algorithms [20, 21, 22, 23, 24], fast multipole methods [25, 26, 27, 28], and Ewald summation techniques [29, 30]. Also, the use of constraints [31, 32, 33] helps to increase efficiency by allowing a longer time step. Other approaches to reach equilibrium conformational properties at an enhanced sampling rate deal with the problem of high frequency vibration of hydrogen atoms [34, 35]. Available methods to study functional transitions of a protein usually require prior knowledge of the transition (i.e. an appropriate reaction coordinate) and thus mainly differ in the definition of the transition coordinate and the way the system is forced to proceed along this coordinate. The method of umbrella sampling [36, 37] requires a pre-defined reaction coordinate, whereas no such limitation is given e.g. in targeted MD [38], essential dynamics [39] and force probe MD [40], mimicking atomic

force microscopy (AFM) single-molecule experiments. In conformational flooding [41], the potential energy landscape of the system is adaptively modified, thereby enabling the system to escape from local minima.

Simplified or coarse grained models of biomolecular systems belong to the second category of methods designed to enhance conformational sampling (see [42] for a comprehensive review). Representations of several atoms up to complete amino acids by single beads allow a drastic reduction of computational means, thereby enabling the simulation of large macromolecular aggregates on micro- to millisecond timescales. This gain in efficiency, however, comes with an inherent lack of accuracy compared to all-atom descriptions of proteins, restricting current models to semi-quantitative statements. Essential in this respect are the parametrizations of used force fields that are both accurate and transferable—that is, force fields capable of describing the general dynamics of systems having different compositions and configurations. As the graining becomes coarser this process becomes increasingly difficult, since more specific interactions must effectively be included in fewer parameters and functional forms. This has led to a variety of models representing different compromises between accuracy and transferability. Apart from using simplified models for proteins and lipids, also surrounding solvent molecules are subject to either coarse graining [43] or complete omission [44, 45, 46, 47]. In the latter case, mostly used in simulations where the solute is represented at atomic resolution, solvent effects are implicitly modeled by additional terms in the force field.

As opposed to the usual Cartesian representation, the efficiency of computer simulations of proteins can be increased by describing the system in their internal degrees of freedom. The use of torsion angles is a natural choice in this respect, since dihedral angles are the main degrees of freedom, of which the $\phi$ and $\psi$ backbone dihedrals play the largest role. The advantage of applying torsion angles in the study of protein dynamics [48, 49] again comes from the larger time steps that can be taken during the simulation (factors of up to 6.5 have been reported [49]). However, a number of problems are encountered when protein dynamics is described in torsion angle space. Solving the equations of motion in these internal coordinates requires the inverse of the moments of inertia tensor at each step. Since matrix inversion scales with the third power of the number of matrix elements in terms of computation time, application of such methods is limited to small systems. However, a method to get around this problem has been proposed [27], reducing the computational cost to order $N$ instead of $N^3$. A second problem connected with torsion angle dynamics is the absence of bond-angle fluctuations, which severely restricts protein dynamics. This results in an overestima-

tion of conformational barriers, thus making the method most useful for simulations at elevated temperatures, used e.g. in the field of NMR structure determination [50] and refinement [51]. Another way to define internal coordinates in proteins is based on the notion that most positional fluctuations occur along only a few collective degrees of freedom. This was first realized from normal mode analysis (NMA) of the small protein bovine pancreatic trypsin inhibitor [52, 53, 54]. Since then, a number of studies [55, 56, 57, 58, 59, 60, 61, 62, 63] have shown that protein dynamics is dominated by a very limited number of collective coordinates, even beyond the harmonic approximation made in NMA. This has led to the development of several simulation methods employing such collective coordinates [39, 64, 65, 66, 67] to drive the dynamics. Although sampling efficiency is increased by these methods, such algorithms often do not reproduce a canonical ensemble.

### Replica Exchange

As a method from the first category introduced above, the replica exchange (REX) method [68] produces correct Boltzmann ensembles for the simulated system, however at the cost of losing dynamical information. The method belongs to the class of so-called generalized ensemble algorithms, which have gained increasing attention in recent years (see [69, 70] for a review). In the REX formalism, enhanced conformational sampling is achieved by simulating in parallel multiple copies (called *replicas*) of the system having a different Hamiltonian, which get exchanged according to a Monte Carlo criterion with a certain probability. Temperature as the discriminating property is used in most applications, but variants using different variables such as hydrophobicity or atomic overlap [71] have been developed. For temperature REX simulations, the large exploratory power of the high temperature replicas is hereby aiding—via exchanges—the low temperature replicas to overcome local energy minima, while the latter achieve a canonical sampling of these newly reached regions of the free energy landscape.

In the context of all-atom simulations, the application of REX is severely hampered by the large number of degrees of freedom associated with these systems. The probability of exchange $P$, essential for the method, approximately scales exponentially with the number of degrees of freedom of the system, $N_{df}$, and the temperature difference $\Delta T = T_m - T_{m-1}$ between successive replicas, $P \sim \exp\{-N_{df}\Delta T\}$. Consequently, applying REX requires a large ($> 30$) number of replicas to bridge a sufficient temperature gap of several hundred Kelvin. Thus, considerable computational effort is needed even for small systems containing only a few thousand atoms.

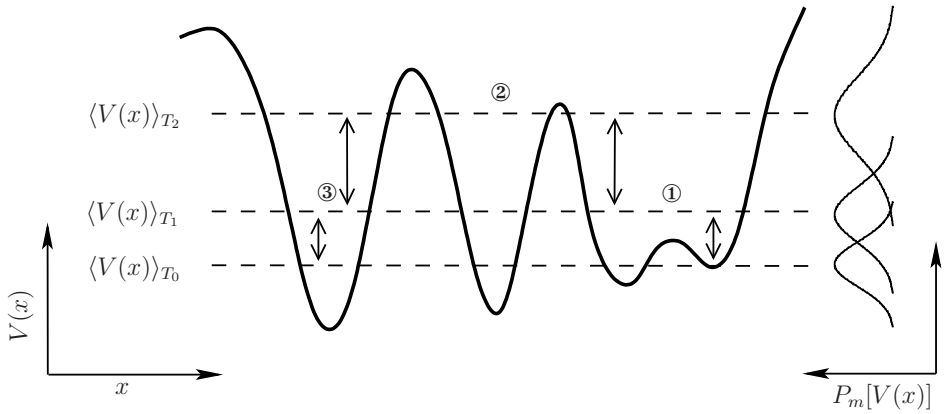


*Figure 1.1:* Temperature REX scheme. Provided sufficient overlap of potential energy distributions $P_m[V(x)]$ (right), replica $m = 0$ moves around configuration space $x$ (①→③) using the higher mobility of auxiliary replicas $m > 0$ within the potential energy landscape $V(x)$ (left).

### Essential Dynamics

To overcome this inherent limitation of REX, a reduction of the simulated number of degrees of freedom is a promising possibility to explore. As already stated above, the use of internal coordinates for the description of protein dynamics can offer computational advantages. As a representative of the third category, essential dynamics (ED) sampling [39, 61] has been successfully applied in recent years. The method excites, following different selectable protocols, collective modes obtained from a principal component analysis (PCA) [59, 60, 61, 72]. Thereby, a considerable enhancement in sampling along these modes of motion is observed, however at the cost of losing dynamical and thermodynamical information. The advantages offered by PCA modes nevertheless render ED an interesting sampling technique.

In PCA, those collective degrees of freedom are selected which contribute most to the atomic motion seen in an ensemble of structures by diagonalizing the covariance matrix of atomic fluctuations. In contrast to NMA, no assumption about the harmonicity of the underlying potential energy landscape is thereby made. Moreover, the first 5-10 % of all PCA modes usually suffice to describe more than 90 % of all fluctuations, and these principal modes were shown to represent biologically relevant motions in several cases [61, 73, 74, 75].

Given the specific advantages of essential dynamics, PCA and replica exchange, the idea comes up of merging the REX and ED approaches into a new and efficient algorithm.

### Aim of this Thesis

In this thesis a new simulation method is developed which combines the advantages of collective coordinate algorithms with the thermodynamical accuracy of the replica exchange formalism. In particular, the standard temperature REX algorithm is brought together with the specific excitation of functionally relevant modes used in the ED protocol to enhance conformational sampling. The modes are thereby constructed from a PCA of an ensemble of structures.

The main tasks which need to be addressed in this work are (1) the incorporation of both methodologies into a coherent algorithm, (2) its implementation and (3) validation, as well as (4) its application. After a brief introduction into the principles of molecular dynamics simulations in chapter 2, the foundation for the new algorithm is laid subsequently. We discuss in detail the ideas behind replica exchange (chapter 3) and principal component analysis (chapter 4), which lies at the heart of the ED algorithm. Chapter 5 is devoted to the new TEE-REX algorithm, synthesizing both approaches. In order to validate the algorithm, its accuracy and performance with respect to REX and MD is evaluated in chapter 6. In particular, the statistical properties of the generated ensemble as well as the sampling performance are assessed quantitatively. To demonstrate the sampling power of TEE-REX, the algorithm is applied to adenylate kinase (chapter 7). For this experimentally well studied enzyme, exhibiting very large conformational motions crucial for its catalytic function, the transition pathway between the two crystallographically resolved structures has not been resolved on atomic level until now. With TEE-REX, a possible transition pathway elucidating the underlying atomic mechanism was observed for the first time.

# 2 Molecular Dynamics Simulations

*Go ahead, make my day.*
— Harry Callahan/Dirty Harry

## 2.1 Principles

This thesis is concerned with the development and application of a new simulation method for which classical molecular dynamics (MD) provides the foundation. Thus, the principles and approximations on which MD simulations rest are briefly outlined in this chapter. For a comprehensive description of MD I refer to [18, 76, 77, 78], and the reference manual [79] of the GROMACS simulation package [80] used in this work.

The exact description of any physical system requires the solution of the time-dependent Schrödinger equation for the $N$-particle wave function $\psi(\boldsymbol{r}, \boldsymbol{R})$ of the system, having $\boldsymbol{r}$ nuclear and $\boldsymbol{R}$ electronic degrees of freedom,

$$i\hbar\frac{\partial}{\partial t}\psi(\boldsymbol{r}, \boldsymbol{R}) = H\psi(\boldsymbol{r}, \boldsymbol{R})\,. \tag{2.1}$$

Here, $H$ denotes the Hamiltonian and $\hbar = h/2\pi$ is the reduced Planck constant. Due to the large number of about $10^3$ to $10^7$ interacting particles for currently simulated biomolecular systems, any attempt at solving such systems via Eq. (2.1) is prohibitive. Approximations are therefore needed to reduce computational demands on current available hardware.

### Born-Oppenheimer

Due to the much lower mass and consequently much higher velocity of the electrons compared to the nuclei, electrons can often be assumed to instantaneously follow the motion of the nuclei. Thus, in the Born-Oppenheimer approximation the total wave

function is separated into the nuclear $\psi_n$ and the electronic wave function $\psi_{el}$,

$$\psi(\boldsymbol{r}, \boldsymbol{R}) = \psi_n(\boldsymbol{r})\psi_{el;\boldsymbol{r}}(\boldsymbol{R})\,.$$

The electronic wave function $\psi_{el;\boldsymbol{r}}(\boldsymbol{R})$ now only parametrically depends on the position $\boldsymbol{r} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N)$, but not on the dynamics of the nuclei. As a result of this approximation, Eq. (2.1) separates into a time-dependent Schrödinger equation for the motion of the nuclei and a time-independent Schrödinger equation for the electron dynamics.

### Newtonian Dynamics

Via the Born-Oppenheimer approximation only the nuclear motion has to be considered, with the electronic degrees of freedom influencing the dynamics of the nuclei in the form of a potential energy surface $V(\boldsymbol{r})$. The second essential approximation is to describe the motion of the nuclei in this potential energy surface classically by Newton's equations of motion

$$m_i \frac{d^2\boldsymbol{r}_i}{dt^2} = -\nabla_i V(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N)\,,$$

where $m_i$ and $\boldsymbol{r}_i$ are the mass and the position of the $i$-th nucleus.

### Force Fields

With the nuclear motion described classically, the Schrödinger equation for the electronic degrees of freedom has to be solved to obtain the potential energy $V(\boldsymbol{r})$. Due to the large number of electrons a further simplification is necessary. Therefore, a semi-empirical force field is introduced which approximates $V(\boldsymbol{r})$ by a large number of functionally simple energy terms for bonded and non-bonded interactions

$$\begin{aligned} V(\boldsymbol{r}) &= V_{\text{bonds}} + V_{\text{angles}} + V_{\text{dihedrals}} + V_{\text{improper}} + V_{\text{Coul}} + V_{\text{LJ}} \\ &= \sum_{\text{bonds}} \frac{1}{2} k_i^l (l_i - l_{i,0})^2 + \sum_{\text{angles}} \frac{1}{2} k_i^\theta (\theta_i - \theta_{i,0})^2 \\ &\quad + \sum_{\text{dihedrals}} \frac{V_n}{2}(1 + \cos(n\varphi - \delta)) + \sum_{\text{improper}} \frac{1}{2} k_\xi (\xi_{ijkl} - \xi_0)^2 \\ &\quad + \sum_{\text{pairs } i,j} \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_{ij}} + \sum_{\text{pairs } i,j} 4\epsilon_{ij} \left[ \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right]. \end{aligned}$$

Parameters for bonded interactions comprise equilibrium bond lengths $l_{i,0}$ and angles $\theta_{i,0}$, the respective force constants $k_i^l$ and $k_i^\theta$, the torsional barrier height $V_n$ with mul-

tiplicity $n$ and the phase $\delta$ of the dihedrals[1]. Improper dihedral angles $\xi_{ijkl}$ between planes $(ijk)$ and $(jkl)$ are needed to keep planar groups planar (e.g. aromatic rings) and to preserve chirality in tetrahedral groups. Non-bonded interactions are parametrized in terms of partial charges $q_i$ for Coulombic interactions while $\epsilon_{ij}$ and $\sigma_{ij}$ define the depth and width of the Lennard-Jones potential, summarizing the short-range Pauli-repulsion and induced dipole-dipole van der Waals interaction between uncharged atoms. All these parameters are determined using *ab initio* quantum chemical calculations or comparisons of structural or thermodynamical data with suitable averages of small molecule MD ensembles. Between different force fields the number of energy terms, their functional form and their individual parameters can differ considerably. From the numerous force fields developed, e.g., CHARMM [81], AMBER [82], GROMOS [83] and OPLS [84], the latter was used throughout this work.

## 2.2 Validity of MD

Although MD simulations have become an established tool in the study of biomolecules, the validity of this approach—like any other scientific model—has to be kept in mind when using MD simulations. The description of a biomolecular system as point masses moving classically in an effective potential breaks down as soon as quantum effects such as electronic reorganizations or very low temperatures (few K) are considered. In such cases a combined quantum mechanical and classical mechanical (QM/MM) approach, originally proposed by Warshel and Levitt [85], may be taken that allows for an accurate description of electronic excitations, charge-fluctuations and -transfer and the forming and breaking of chemical bonds.

In classical MD simulations, depending on the chosen force field and the type of compound studied, each atom is assigned a partial charge that reflects the polarity and approximately models effective polarization. Throughout the simulation these charges are kept constant thereby excluding explicit polarization effects. Nowadays, several polarizable water models and force fields exist, see [86] for a recent review.

The approximation of the potential energy surface $V(\boldsymbol{r})$ by some empirical force field naturally raises the question of how accurate physical quantities are modeled. Each force field has its own strengths in reproducing certain observations due to the data that were specifically used to parameterize it. Consequently, the choice of a particular force field will depend on the property and level of accuracy one is interested in.

[1] In a four atom system $ABCD$, the angle between the two planes $ABC$ and $BCD$ defines the dihedral.

## 2.3 Implementation Details and Simulation Setup Protocol

The above approximations lay the foundation for a practical realization of MD simulations of proteins, as it is done in the GROMACS simulation suite which was used here and whose algorithms and methods will be sketched in the following.

Newton's equations of motion are solved iteratively in discrete steps by means of the leap-frog algorithm [87], which has the advantage that the computationally intensive force calculations need to be done only once per integration step. The length of one time step has to be chosen such that it is small in comparison to the fastest motions of the system. Bond vibrations involving hydrogen occur within several femtoseconds, restricting the time step to 0.5 fs. A number of algorithms to constrain covalent bond lengths have been developed that allow larger time steps. All MD simulations presented in this thesis use the LINCS [31] and SETTLE [32] algorithms, allowing a time step of 2 fs.

Besides interactions with membranes and other macromolecules, water is the natural environment for proteins. For a simulation of a model system that matches the *in vivo* system as close as possible, water molecules and sodium chloride in physiological concentration are added to the system in order to solvate the protein. Having a simulation box filled with solvent and solute, artifacts due to the boundaries of the system may arise, such as evaporation, high pressure due to surface tension and preferred orientations of solvent molecules on the surface. To avoid such artifacts, periodic boundary conditions are applied. In this way, the simulation system does not have any surface. This, however, may lead to new artifacts if the molecules artificially interact with their periodic images due to e.g. long-range electrostatic interactions. These periodicity artifacts are minimized by increasing the size of the simulation box. Different choices of unit cells, e.g., cubic, dodecahedral or truncated octahedral allow an improved fit to the shape of the protein, and, therefore, permit a substantial reduction of the number of solvent molecules while simultaneously keeping the crucial protein-protein distance high. Long-range Coulomb interactions in periodic systems are treated by the Particle-Mesh-Ewald (PME) method [88, 89], which, in contrast to simple cut-off methods [90, 91], allows their correct and computationally efficient evaluation.

A solution of Newton's equations of motion conserves the total energy of the system, resulting in a microcanonical *NVE* ensemble. However, real biological subsystems of the size studied in simulations constantly exchange energy with their surrounding.

Furthermore, a constant pressure of usually 1 bar is present. To account for these features, algorithms are introduced which couple the system to a temperature and pressure bath. From the many proposed *thermostats* [92, 93, 94, 95, 96], the popular Berendsen thermostat is used which simply rescales the velocities in each step using

$$v' = \lambda v\,, \qquad \lambda = \left[1 + \frac{\Delta t}{\tau_T}\left(\frac{T_0}{T} - 1\right)\right]^{1/2}\,. \tag{2.2}$$

Here, $T_0$ denotes the reference temperature of the heat bath, $\tau_T$ the coupling constant, and $\Delta t$ the integration time step. Pressure coupling in this work is done by the Berendsen barostat [94], which rescales the coordinates at each step. Thus, isobaric-isothermal $NPT$ ensembles are created.

# 3 Replica Exchange Molecular Dynamics

*A child of five would understand this. Send someone to fetch a child of five.*

— Groucho Marx

## 3.1 Conformational Sampling

The aim of computer simulations of molecular systems is to calculate macroscopic behavior from microscopic interactions. Thus, in order to describe the thermodynamics and kinetics of proteins, a thorough sampling of the conformational space of the system is required. Following equilibrium statistical mechanics, any observable that can be connected to macroscopic experiments is defined as an ensemble average $\langle A \rangle_{\text{ensemble}}$ over all possible realizations of the system. For a protein simulation described by MD, the ensemble average cannot be computed directly from a single trajectory. However, the ergodic hypothesis, which is generally assumed to apply for protein dynamics, allows the indirect computation of ensemble averages as time averages from such a single trajectory produced by MD simulations,

$$\langle A \rangle_{\text{time}} = \lim_{T \to \infty} \frac{1}{T} \int_{t=0}^{T} A(\boldsymbol{r}(t), \boldsymbol{p}(t)) dt \,.$$

Given current computer hardware, a fully converged sampling of all possible conformational states with the respective Boltzmann weight is attainable for simple systems comprising several amino acids (see, e.g. [97]). For proteins, consisting of hundreds to thousands of amino acids, conventional MD simulations often do not converge and reliable estimates of experimental quantities can not be calculated.

#### Energy Landscape

The inefficiency in sampling is a result of the ruggedness of the systems' energy landscape, a concept put forward by Frauenfelder [15, 16]. The term energy landscape is ambiguously[1] used within the literature, defined either as the potential or the free energy of the system as a function of all structural degrees of freedom $N_{df}$. The global shape of the free energy landscape is supposed to be funnel-like, with the native state populating the global energy minimum [17]. Looking in more detail, the complex high-dimensional free energy landscape is characterized by a multitude of almost iso-energetic minima, separated from each other by energy barriers of various heights. Each of these minima corresponds to one particular conformational substate, with neighboring minima corresponding to similar conformations. Within this picture, structural transitions are barrier crossings, with the transition rate depending on the height of the barrier.

For MD simulations at room temperature, only those barriers are easily overcome that are smaller than or comparable to the thermal energy $k_B T$ and the observed structural changes are small, e.g. side chain rearrangements. Most of its time the system will spend in locally stable states (*kinetic trapping*). Of higher interest are—due to their connection to biological function—the exploration of different conformational states and the mechanism of global conformational transitions, which require the system to overcome large energy barriers. Since MD simulations are mostly restricted to the nanosecond timescale, functionally relevant conformational changes are rarely observed.

A plethora of enhanced sampling methods have been developed to tackle this multi-minima problem, see e.g. [18, 19, 76] and references therein. Among them, generalized ensemble algorithms have been widely used in recent years (for a review, see e.g. [69, 70]).

## 3.2 Replica Exchange

### 3.2.1 Generalized Ensemble Algorithms

Generalized ensemble algorithms sample an artificial ensemble that is either constructed from compositions or extensions of the original ensemble. The multicanonical algorithm [98] and its variant simulated tempering Monte-Carlo (MC) [99, 100] are examples of this second category.

[1] In this thesis, the terms 'potential energy landscape' and 'free energy landscape' are used to avoid possible misunderstandings.

In the multicanonical algorithm, the bell-shaped canonical distribution of the potential energy $p(E)$ is modified by a so-called multicanonical weight factor $w(E)$ making the resulting distribution uniform ($p(E)w(E) = \text{const}$). In a single multicanonical simulation this flat distribution can then be sampled extensively by MD or MC because potential energy barriers are no longer present. For simulated tempering, the temperature is no longer fixed but becomes a dynamical variable, and both the configuration and the temperature are updated during a single MC simulation with a weight factor. The latter is chosen such that the probability distribution of temperature is constant ($p(T) = \text{const}$). Hence, a random walk in temperature space is realized, which in turn induces a random walk in potential energy space and allows the system to escape from local energy minima. In both algorithms, estimates for canonical ensemble averages of physical quantities are obtained by reweighting techniques [37, 101].

The main problem with these algorithms, however, is the non-trivial determination of the different multicanonical weight factors by an iterative process involving short trial simulations. For complex systems this procedure can be very tedious and attempts have been made to accelerate convergence of the iterative process [102, 103, 104, 105, 106].

The replica exchange (REX) algorithm, developed as an extension of simulated tempering, removes the problem of finding correct weight factors. It belongs to the first category of algorithms where a composition of the original ensemble is sampled. The standard temperature formulation of replica exchange MD, as detailed in [68], constitutes the main building block of the *Temperature Enhanced Essential dynamics Replica EXchange* (TEE-REX) algorithm developed in chapter 5. The standard temperature REX algorithm is reviewed in the following section to introduce the concept and clarify notation.

### 3.2.2 Temperature Replica Exchange

Consider a simulation system of $N$ atoms of mass $m_k$ ($k = 1, \ldots, N$) with their coordinate and velocity vectors denoted by $x := (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \in \mathbb{R}^{3N}$, $\boldsymbol{x}_i \in \mathbb{R}^3$ and $v := (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N) \in \mathbb{R}^{3N}$, $\boldsymbol{v}_i \in \mathbb{R}^3$, respectively. The Hamiltonian $H(x, v) = E(x) + K(v)$ is given by the sum of the potential energy $E(x)$ and the kinetic energy $K(v) = \sum_{k=1}^{N} m_k \boldsymbol{v}_k^2/2$. In the canonical ensemble at temperature $T$, each state $s := (x, v)$ with the Hamiltonian $H(s)$ has a probability given by its Boltzmann factor $W(s) = \exp\{-\beta H(s)\}$, with the inverse temperature $\beta^{-1} = k_B T$ and the Boltzmann constant $k_B$. Via the equipartition theorem, the average kinetic energy is linked to the number

of degrees of freedom $N_{df}$ of the system,

$$\langle K(v)\rangle_T = \frac{N_{df}}{2} k_B T \,. \tag{3.1}$$

Usually $N_{df} \ll 3N$ since constraint algorithms [31, 32] considerably restrict the number of degrees of freedom. As soon as flexible bonds are simulated, $N_{df} = 3N$ and the standard textbook expression for a free $N$-particle system is recovered.

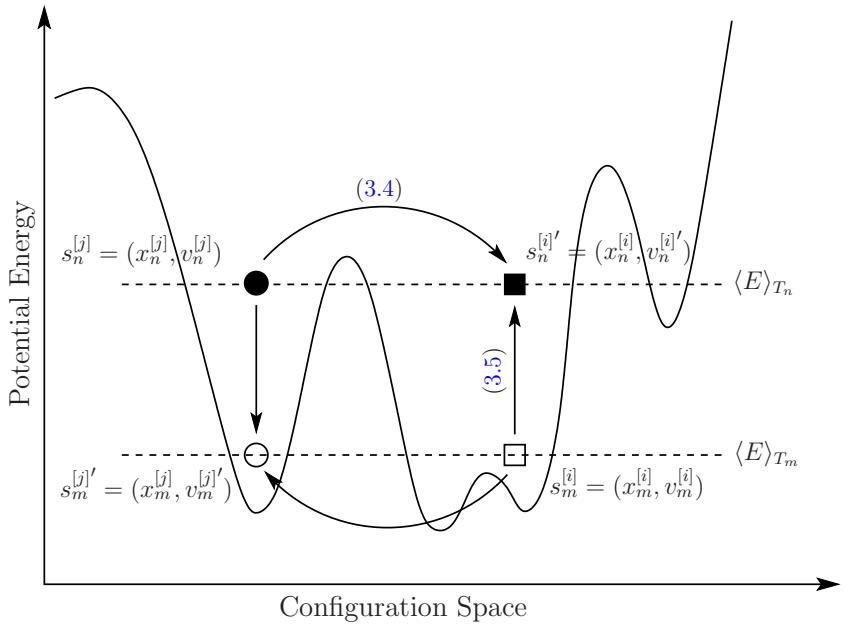


*Figure 3.1:* Schematic illustration of exchange $S \to S'$, cp. equation (3.3). Configuration and temperature of each replica icon (e.g. □) are represented by shape and fill-color, respectively. The average potential energy of each replica is shown as dashed line. Configuration (3.4) or temperature (3.5) exchange corresponds to a horizontal or vertical movement, respectively.

In REX, a generalized ensemble is constructed, which consists of $M+1$ *non-interacting* copies (or *replicas*) of the original system in the canonical ensemble with temperatures $\{T_0, T_1, \ldots, T_M\}$ and $T_m \leq T_{m+1}$ $(m = 0, \ldots, M)$. Equal temperatures for two or more replicas are possible, but seldomly used. A state of this generalized ensemble is characterized by

$$S = \{\ldots, s_m^{[i]}, \ldots\} \,,$$

where the configuration $s_m^{[i]} := (x_m^{[i]}, v_m^{[i]})$ represents the coordinates $x_m^{[i]}$ and velocities $v_m^{[i]}$ of all atoms of the $i$th replica at temperature $T_m$. Because the replicas are non-interacting, the statistical weight of a state $S$ of this generalized ensemble is given by

the product of Boltzmann factors for each replica,

$$W(S) = \exp\left\{ -\sum_{m=0}^{M} \beta_m H(s_m^{[i]}) \right\}. \tag{3.2}$$

We now consider an exchange between a pair of replicas $i$ and $j$,

$$S = \{\ldots, s_m^{[i]}, \ldots, s_n^{[j]}, \ldots\} \quad \rightarrow \quad S' = \{\ldots, s_m^{[j]\prime}, \ldots, s_n^{[i]\prime}, \ldots\}. \tag{3.3}$$

In more detail, (3.3) reads

$$\begin{aligned} s_m^{[i]} &= (x_m^{[i]}, v_m^{[i]}) \quad &\rightarrow \quad s_m^{[j]\prime} &= (x_m^{[j]}, v_m^{[j]\prime}) \\ s_n^{[j]} &= (x_n^{[j]}, v_n^{[j]}) \quad &\rightarrow \quad s_n^{[i]\prime} &= (x_n^{[i]}, v_n^{[i]\prime}), \end{aligned} \tag{3.4}$$

which corresponds to an exchange of configurations of the two replicas. It is interesting to note that this configuration exchange is equivalent to exchanging a pair of temperatures[2] $T_m$ and $T_n$ (Fig. 3.1),

$$\begin{aligned} s_m^{[i]} &= (x_m^{[i]}, v_m^{[i]}) \quad &\rightarrow \quad s_n^{[i]\prime} &= (x_n^{[i]}, v_n^{[i]\prime}) \\ s_n^{[j]} &= (x_n^{[j]}, v_n^{[j]}) \quad &\rightarrow \quad s_m^{[j]\prime} &= (x_m^{[j]}, v_m^{[j]\prime}). \end{aligned} \tag{3.5}$$

Unlike the original implementation of REX using Monte-Carlo updating steps [107, 108, 109], REX MD requires a rescaling of velocities, indicated by the primes in Eq. (3.4) and Eq. (3.5). Velocity rescaling is done in such a way, that the equipartition theorem (3.1) holds for each replica at all times. To this end, we look at the situation immediately after a temperature exchange $S \rightarrow S'$. Starting with replica $s_m^{[i]} = (x_m^{[i]}, v_m^{[i]})$, the equipartition theorem reads $2\langle K(v_m^{[i]})\rangle_{T_m} = N_{df} k_B T_m$. Upon exchange, replica $s_n^{[i]\prime}$ receives the rescaled velocities of replica $s_m^{[i]}$, thus $2\langle K(v_n^{[i]\prime})\rangle_{T_n} = N_{df} k_B T_n$. Combining both expressions yields

$$\langle K(v_n^{[i]\prime})\rangle_{T_n} = \frac{T_n}{T_m} \langle K(v_m^{[i]})\rangle_{T_m}.$$

Since $K(v) \sim v^2$, we arrive at the primed velocities

$$v_n^{[i]\prime} = \sqrt{\frac{T_n}{T_m}} v_m^{[i]}, \qquad v_m^{[j]\prime} = \sqrt{\frac{T_m}{T_n}} v_n^{[j]}, \tag{3.6}$$

[2] In a parallel computing environment, this exchange protocol requires much less network communication between replicas.

where all atoms of the replicas are rescaled uniformly.

For the exchange process to converge towards the equilibrium distribution (3.2), it is sufficient to impose the detailed-balance condition on the transition probability (*exchange/acceptance* probability) $P(S \to S')$,

$$W(S)P(S \to S') = W(S')P(S' \to S)\,.$$

It is

$$\begin{aligned}
\frac{P(S \to S')}{P(S' \to S)} &= \frac{\exp\{-\beta_m H(s_m^{[j]\prime})\}\exp\{-\beta_n H(s_n^{[i]\prime})\}}{\exp\{-\beta_m H(s_m^{[i]})\}\exp\{-\beta_n H(s_n^{[j]}\}} \\
&= \exp\{-\beta_m[H(s_m^{[j]\prime}) - H(s_m^{[i]})] - \beta_n[H(s_n^{[i]\prime}) - H(s_n^{[j]})]\} \\
&\overset{(3.6)}{=} \exp\{\beta_m[E(x_m^{[i]}) - E(x_m^{[j]})] - \beta_n[E(x_n^{[i]}) - E(x_n^{[j]})]\} \\
&\overset{(*)}{=} \exp\{(\beta_m - \beta_n)[E(x_m^{[i]}) - E(x_n^{[j]})]\}
\end{aligned}$$

In the last step $(*)$ we used the fact that the potential energy of the system immediately after exchange solely depends on the respective conformation of the system an not on the temperature; thus, $x_m^{[j]} = x_n^{[j]}$ and $x_n^{[i]} = x_m^{[i]}$. Detailed-balance can be satisfied by the usual Metropolis Monte-Carlo criterion:

$$P(S \to S') = \min\left\{1, \exp\{(\beta_m - \beta_n)[E(x_m^{[i]}) - E(x_n^{[j]})]\}\right\}. \tag{3.7}$$

For simulations performed in the $NPT$-ensemble, Eq. (3.7) is modified by a pressure correction term [110]. Putting together everything, a simulation using the REX algorithm is realized by alternately performing the following two steps:

(1) *simultaneous* and *independent* simulation of each replica for a certain number of MD[3] steps

(2) exchange of two replicas according to the Metropolis criterion (3.7).

In practice, only neighboring replicas are exchanged since the acceptance probability (3.7) exponentially decreases with the temperature and potential energy difference $\Delta\beta\Delta E$. Within the generalized ensemble $S$ a random walk in temperature space is performed, translating into a random walk in potential energy space for a single replica.

[3] Monte-Carlo updating can also be used.

This facilitates an efficient and statistically correct conformational sampling of the rugged free energy landscape of the system.

### 3.2.3 Algorithm Performance

The appropriate choice of temperatures is crucial for an optimal performance of the REX algorithm. Depending on the problem under study, the properties of the system at the lowest temperature $T_0$ within the replica setup are usually of particular interest. Therefore, replica temperatures have to be chosen such that (a), the lowest temperature sufficiently samples energy states of interest; (b), the highest temperature is large enough to overcome energy barriers of the system; and (c), the acceptance probability $P(S \rightarrow S')$ is sufficiently high, requiring an adequate overlap of potential energy distributions for neighboring replicas, see Eq. (3.7). Protocols for this task can be found in [71, 111]. Once a temperature distribution is established, the frequency with which exchanges between replicas are attempted (*exchange attempt frequency*) needs to be fixed. There is some discussion within the literature [112, 113, 114, 115, 116, 117] analyzing the interplay between acceptance probability and sampling efficiency. Likewise, finding an appropriate criterion for judging REX efficiency is still a matter of controversy [115, 116, 118, 119, 120].

Besides its simplicity and ease of implementation, the REX algorithm has some advantages which is reflected in the widespread use of this method over the last few years. The main advantage of REX over other generalized ensemble methods lies in the fact that the weight factor $W(S)$ is known *a priori* and does not have to be determined by a tedious and time-consuming procedure. Furthermore, each replica $s_m$ samples from a Boltzmann ensemble having a temperature $T_m$. Using the weighted histogram analysis method (WHAM) [37], thermodynamic quantities as a function of temperature can be calculated from the simulated generalized ensemble $S$. This property of REX is often used in the study of phase transitions [121] such as folding/unfolding simulations [111, 122, 123, 124, 125, 126, 127, 128] or aggregation phenomena [129]. In particular, the temperature dependence of calculated free energies $\Delta G = \Delta H - T\Delta S$ allows inferences about enthalpic and entropic contributions, assuming that $\Delta H$ and $\Delta S$ do not depend on temperature [130].

## 3.3 All-Atom Explicit Solvent REX Simulations

Although REX has advantages over previously used algorithms, the method quickly becomes inapplicable for simulations at full atomic resolution using explicit solvent. Because the number of replicas needed to span a given temperature range scales with the square root of the number of degrees of freedom of the system, $N_{df}^{1/2}$, many replicas need to be simulated to span a temperature range that includes significantly higher temperatures than the reference temperature $T_0$ [71, 131]. Even for small peptide simulations [132] several thousand degrees of freedom are present in the system, requiring already $> 20$ replicas to obtain exchange probabilities of $\sim 25\%$. For a given temperature range $T_M - T_0$, a slight decrease in the necessary number of replicas can be achieved by accepting lower exchange probabilities. However, temperature differences of more than a few K are usually not possible within this setup.

The reason for this limitation is the large number of explicitly simulated solvent molecules. A simple estimate [71, 131] shows that the potential energy difference $\Delta E \sim N_{df}\Delta T$ is dominated by the contribution from the solvent degrees of freedom $N_{df}^{\mathrm{sol}}$, constituting the largest fraction of the total number of degrees of freedom, $N_{df}$, of the system. Thus, the acceptance probability $P(S \to S') \sim \exp\{-\Delta E\}$ is dramatically decreased, which in turn enormously increases computational demands.

A promising point of attack for the improvement of this class of algorithms is therefore the reduction of the number of degrees of freedom used in the calculation of the exchange probability (3.7). Implicit solvent models such as the semianalytical generalized Born model [44, 45] or more rigorous models based on Poisson-Boltzmann equations [46, 47] are often used in this respect. Here, the free energy of solvation of the solute is estimated based on coordinates of the solute. Although this neglect of explicit solvent molecules significantly reduces computational costs, simulations using such models do not have as good a balance between protein-protein and protein-solvent interactions as explicit solvent models [123, 124, 133, 134, 135].

Besides implicit solvent models, a lot of research has been carried out over the last few years to reduce the number of degrees of freedom in simulations [70, 71, 131, 132, 136, 137, 138, 139, 140, 141]. Overall, each of these techniques has its own strengths and weaknesses, and the best choice typically depends on the nature of the physical system being studied and on the available computing resources. The TEE-REX algorithm, discussed in chapter 5, takes an altogether different approach to the reduction of $N_{df}$ by employing generalized coordinates.

# 4 Principal Component Analysis

*If you can't convince 'em, confuse 'em.*
— Harry Truman

Apart from the desire to enhance conformational sampling and/or reduce computational costs by reducing the number of degrees of freedom of the simulated replicas, standard REX and all its variants use an *undirected* excitation scheme such as temperature [68, 70, 107, 108, 110, 131, 132, 136, 137, 139, 141, 142, 143], hydrophobicity [71], potential energy [132, 136, 140], atomic overlap [71] or degree of coarse-graining [138, 144]. In this context, undirected means that the parameter, discriminating different replicas, acts *uniformly* and *uncorrelated* on all degrees of freedom affected by this parameter. In standard REX simulations this parameter is temperature: for replica $s_m$, the kinetic energy provided by temperature $T_m > T_0$ is, via Eq. (2.2), evenly distributed over all solvent and solute atoms of the system and no preference is given to any specific degrees of freedom. Thus, the majority of motions excited by this scheme are of limited interest since they involve only small uncorrelated fluctuations, e.g. rearrangements of side chain atoms. Often, however, the focus of interest lies on correlated large-scale motions of the system such as the opening and closing of protein domains relative to each other.

A further enhancement of sampling can therefore be achieved by combining a REX-based simulation protocol with a *directed* excitation scheme. Naturally, important functional motions of the considered protein lend themselves to this task. With the TEE-REX algorithm, a formal realization of this idea is presented in chapter 5. As a first step, the notion of collective motions of proteins is introduced in this chapter.

## 4.1 Internal Coordinate Description of Protein Dynamics

There exist two major techniques to extract and classify relevant information about large conformational changes from an ensemble of protein structures, generated either

experimentally or theoretically: normal mode analysis (NMA) and principal component analysis (PCA). Here, we focus on the latter.

Both NMA and PCA are based on the notion that by far the largest fraction of positional fluctuations in proteins occurs along only a small subset of collective degrees of freedom. This was first realized from normal mode analysis of a small protein [52, 53, 54]. In NMA, the potential energy surface is assumed to be harmonic and collective variables are obtained by diagonalization of the Hessian[1] matrix in a local energy minimum. Quasi-harmonic analysis [55, 56, 57, 58], PCA [59, 60, 61, 72] and singular-value decomposition [62, 63] of MD trajectories of proteins that do not assume harmonicity of the dynamics, have shown that indeed protein dynamics is dominated by a limited number of collective coordinates, even though the major modes are frequently found to be largely anharmonic. These methods identify those collective degrees of freedom that best approximate the total amount of fluctuation. The subset of largest-amplitude variables form a set of generalized internal coordinates that can be used to effectively describe the dynamics of a protein. Often, a small subset of 5-10 % of the total number of degrees of freedom yields a remarkably accurate approximation. As opposed to torsion angles as internal coordinates, these collective internal coordinates are not known beforehand but must be defined either using experimental structures or an ensemble of simulated structures. Once an approximation of the collective degrees of freedom has been obtained, this information can be used for the analysis of simulations as well as in simulation protocols designed to enhance conformational sampling [39, 41, 64, 65, 145].

## 4.2 Theoretical Background

A detailed mathematical treatment of PCA can be found in [61, 72, 145]. Here, we give a comprehensive description based on [146]. In essence, a principal component analysis is a multi-dimensional linear least squares fit procedure in configuration space. The structure ensemble of a molecule, having $N$ particles, can be represented in $3N$-dimensional configuration space as a distribution of points with each configuration represented by a single point. For this cloud, always one axis can be defined along which the maximal fluctuation takes place. As illustrated for a two-dimensional example (Fig. 4.1), if such a line fits the data well, all data points can be approximated by only the projection onto that axis, allowing a reasonable approximation of the position even when neglecting the

[1] second derivative $\sum_{i,j} \frac{\partial^2 V}{\partial x_i \partial x_j}$ of the potential energy

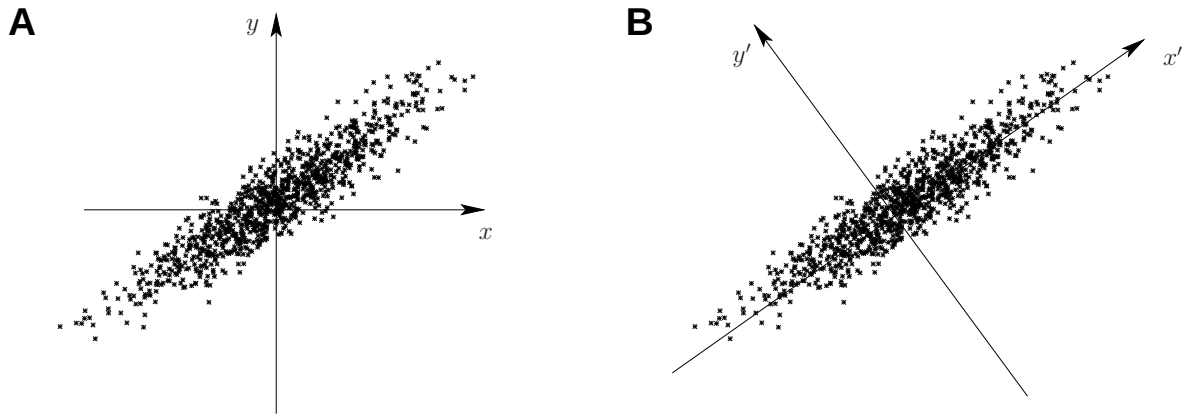


*Figure 4.1:* Illustration of PCA in two dimensions. Two coordinates $(x, y)$ are required to identify a point in the ensemble in panel $A$, whereas one coordinate $x'$ approximately identifies a point in panel $B$.

position in all directions orthogonal to it. If this axis is chosen as coordinate axis, the position of a point can be represented by a single coordinate. The procedure in the general $3N$-dimensional case works similarly. Given the first axis that best describes the data, successive directions orthogonal to the previous set are chosen such as to fit the data second-best, third-best, and so on (the *principal components*). Together, these directions span a $3N$-dimensional space. Applications of such a multidimensional fit procedure on protein configurations from MD simulations of several proteins have proven that typically the first ten to twenty principal components are responsible for 90 % of the fluctuations of a protein (Fig. 4.2) [59, 60, 61]. These principal components correspond to collective coordinates, containing contributions from every atom of the protein. In a number of cases these principal modes were shown to be involved in the functional dynamics of the studied proteins [61, 73, 74, 75]. Hence, the subspace responsible for the majority of all fluctuations has been referred to as the *essential subspace* [61].

Only internal fluctuations are usually of interest in the study of protein dynamics. Thus, the first step of a PCA is to remove overall rotation and translation of each configuration of the ensemble: after a translation of the center of mass of every configuration to the origin, a rotational least squares fit of the atoms onto a reference structure is performed. Next, the variance-covariance matrix of positional fluctuations is constructed and diagonalized. Let $x(t)$ describe the fitted trajectory (ensemble) of internal motions of the protein, where $x \in \mathbb{R}^{3N}$ is a column vector describing the coordinates of $N$ protein atoms and the time index $t$ identifies each member of the ensemble. Although

ensemble configurations $x(t)$ are denoted as a function of time, they may be provided in any order. In general, a PCA is not restricted to the complete protein but may be performed on any subset of atoms (e.g. the protein backbone). The variance-covariance matrix now reads

$$C = \langle (x(t) - \langle x \rangle)(x(t) - \langle x \rangle)^T \rangle \,,$$

with the angle brackets $\langle \cdot \rangle$ representing an ensemble average. Particles moving in a correlated fashion correspond to positive matrix elements (positive correlation) or negative elements (negative correlation), and those that move independently to small matrix elements. The symmetric matrix $C \in \mathbb{R}^{3N} \times \mathbb{R}^{3N}$ can always be diagonalized by an orthogonal coordinate transformation $T$ which transforms $C$ into a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{3N})$, containing the eigenvalues $\lambda_i$ of $C$,

$$\Lambda = T^T C T \quad \text{or} \quad C = T \Lambda T^T \,.$$

The $i$th column of $T = (\mu_1, \mu_2, \ldots, \mu_{3N})$ contains the normalized eigenvector (principal component) $\mu_i \in \mathbb{R}^{3N}$ of $C$ belonging to $\lambda_i$. When a sufficient number of independent configurations (at least[2] $3N + 1$) are available to evaluate $C$, there will be $3N - 6$ eigenvectors with non-zero eigenvalues. Six eigenvalues should be exactly zero, of which the corresponding eigenvectors describe the overall rotation and translation that was eliminated by the fitting procedure beforehand. If only $K < 3N + 1$ independent configurations are available then at most $K - 1$ non-zero eigenvalues with corresponding eigenvectors will result.

The eigenvalues correspond to the mean square positional fluctuation along the respective eigenvector, and therefore contain the contribution of each principal component to the total fluctuation. When the eigenvectors are sorted to decreasing eigenvalue, the first eigenvectors describe those collective motions that best approximate the sum of fluctuations and the last eigenvectors correspond to the most constrained degrees of freedom. The characteristics of these collective motions can be studied by projecting the ensemble onto single eigenvectors, yielding the principal coordinates $p_i(t) \in \mathbb{R}$,

$$p_i(t) = \mu_i \cdot (x(t) - \langle x \rangle) \,.$$

[2] An intuitive argument for this number is the following: Imagine a single configuration as one point, then two independent configurations (points in a plane) can be described by one collective coordinate, going through these two points. Given three independent points, two collective coordinates are needed. Hence, for a description of a protein having $3N$ degrees of freedom, $3N+1$ independent configurations are necessary.

Note that the variance $\langle p_i^2 \rangle$ equals the eigenvalues $\lambda_i$, $\langle p_i^2 \rangle = \lambda_i$. Often, two- or three-dimensional projections along the major principal components are used to allow a representation of the sampled distribution in configuration space or to compare multiple ensembles along the principal modes of collective fluctuation. A translation of these projections back into Cartesian space can be used to visualize the atomic displacements associated with a particular eigenvector,

$$x_i'(t) = p_i(t)\mu_i + \langle x \rangle \, .$$

### Annotations

The fact that a small subset of the total number of degrees of freedom (essential subspace) dominates the molecular dynamics of proteins (Fig. 4.2) originates from the presence of a large number of internal constraints and restrictions defined by the atomic interactions present in a biomolecule. These interactions range from strong covalent bonds to weak non-bonded interactions, whereas the restrictions are given by the dense packing of atoms in native-state structures.

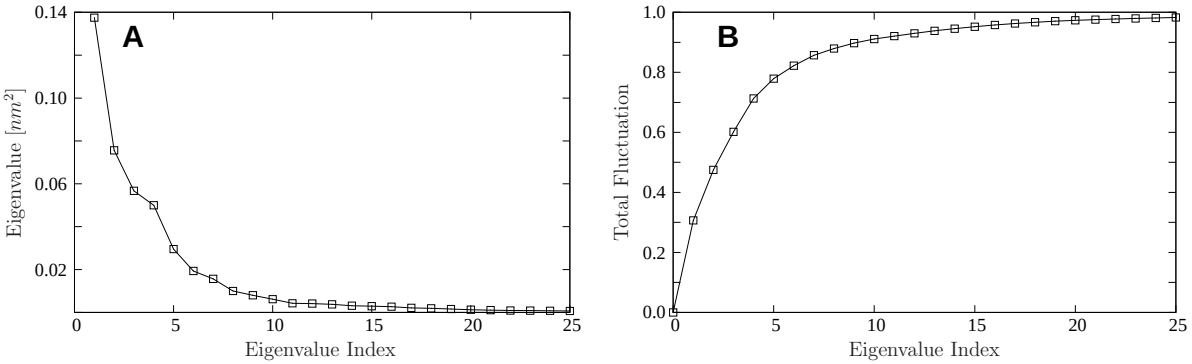


*Figure 4.2:* Typical PCA eigenvalue spectrum (MD ensemble of guanylin backbone structures). The first five eigenvectors (panel *A*) cover 80% of all observed fluctuations (panel *B*).

Overall, protein dynamics at physiological temperatures has been described as diffusion among multiple minima [147, 148, 149]. The dynamics on short timescales is dominated by fluctuations within a local minimum, corresponding to eigenvectors having low eigenvalues. On longer timescales large fluctuations are dominated by a largely anharmonic diffusion between multiple wells. These slow dynamical transitions are usually represented by the largest-amplitude modes of a PCA.

In contrast to normal mode analysis, PCA of a MD simulation trajectory does not rest on the assumption of a harmonic potential. In fact, PCA can be used to study

the degree of anharmonicity in the molecular dynamics of the simulated system. For proteins, it was shown that at physiological temperatures, especially the major modes of collective fluctuation are dominated by anharmonic fluctuations [61, 150].

By definition, PCA is a *linear* analysis, i.e. only linear correlations between atomic displacements enter the covariance matrix $C$. This means that non-linear correlations between atom movements may be overlooked as they get spread out across multiple collective coordinates. In practice, this is usually not a big problem, except for systems that undergo large-scale rotations. In such cases, several eigenvectors are needed for a description of these motions.

# 5 Temperature Enhanced Essential Dynamics Replica Exchange

*All great ideas look like bad ideas to people who are losers.*
*It's always good to test a new idea with known losers to make sure they don't like it.*
— Dogbert

As we have seen in chapter 3, generalized ensemble algorithms such as REX provide a means to tackle the notorious sampling problem encountered in all-atom simulations of biomolecular systems using explicit solvent. Focusing on standard temperature REX, the method quickly becomes computationally prohibitive for all but the smallest systems. The main bottleneck is the large number of simulated degrees of freedom. From PCA (chapter 4) we know that the configurational dynamics of proteins is dominated by a small number of collective degrees of freedom. When using sampling techniques based on a selective excitation of such collective coordinates [39, 41, 64, 65], a significant increase of sampling efficiency can be achieved [39, 151, 152]. However, systems simulated with such methods are always in a non-equilibrium state, rendering it difficult to extract thermodynamic properties of the system from such simulations.

With the newly developed *Temperature Enhanced Essential dynamics Replica EXchange* (TEE-REX) algorithm the favorable properties of REX are now combined with those resulting from a specific excitation of functionally relevant modes, while at the same time avoiding the drawbacks of both approaches. In the following, we sketch the algorithm and discuss in depth the crucial parts of the simulation protocol, namely temperature coupling and calculation of exchange probability.

## 5.1 Algorithm

The basis for TEE-REX is given by the replica framework, i.e. $M+1$ replicas ($m = 0, \ldots, M$) of the system under study are simulated simultaneously and independently

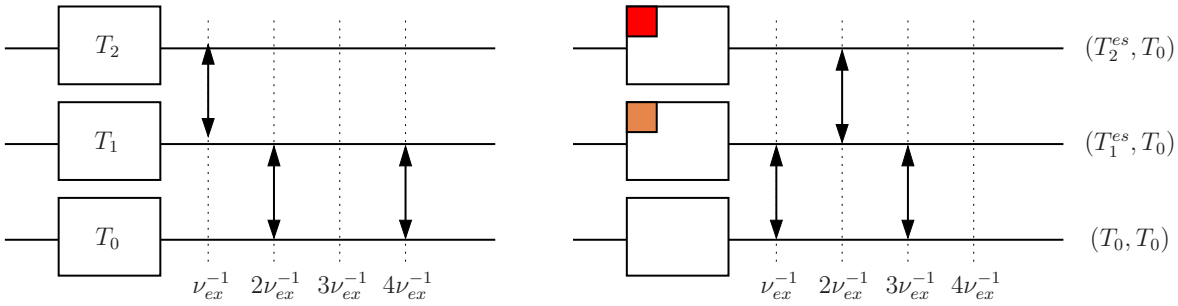


*Figure 5.1:* Comparison of standard temperature REX (left panel) and the TEE-REX algorithm (right panel) for a three-replica simulation. Temperatures are sorted in increasing order, $T_{i+1} > T_i$. Exchanges ($\leftrightarrow$) are attempted ($\cdots$) with frequency $\nu_{ex}$. Unlike REX, only an essential subspace $\{es\}$ (colored boxes) containing a few collective modes is excited within each TEE-REX replica. Reference replica $(T_0, T_0)$, containing an approximate Boltzmann ensemble, is used for analysis.

by MD with periodic exchange attempts every $\nu_{ex}^{-1}$ time steps. In contrast to standard replica exchange, TEE-REX replicas $m = 1, \ldots, M$ are divided into an *essential subspace* and its complement (Fig. 5.1). The essential subspace

$$\{es\} := \{\mu_i \in \mathbb{R}^{3N_I} \,|\, i = 1, \ldots, N_{es}\}$$

is defined by a subset of $N_{es}$ eigenvectors $\{\mu_k \in \mathbb{R}^{3N_I} \,|\, i = 1, \ldots, 3N_I\}$, describing collective modes of a subsystem of interest having $N_I$ atoms. A loop region or the protein backbone could be such a subsystem. The collective degrees of freedom $\{\mu_k\}$ can be obtained in a variety of ways. Here, we use a PCA of an ensemble of structures (e.g. NMR or X-ray data or a previous simulation), but also eigenvectors from a normal mode analysis of a single structure can be used. Between exchanges, the essential subspace of replicas $m = 1, \ldots, M$ is coupled to a temperature bath $T_m^{es} > T_0$ with the rest of the simulation system staying at the reference temperature $T_0$. For replica $m = 0$, no partition into $\{es\}$ and its complement is applied and all degrees of freedom are coupled to the same temperature, $T_0^{es} = T_0$. The ensemble generated by this reference replica is later on used for analysis.

Before discussing algorithmic details, let us summarize the idea behind the TEE-REX algorithm. In essence, TEE-REX combines the REX approach with essential dynamics. A thermal stimulation of *only* those degrees of freedom that contribute significantly to the total fluctuations of the system is carried out, and multiple simulations at different temperature levels are coupled to each other via a REX framework. This way, several benefits are combined and drawbacks avoided. In contrast to standard temperature

REX, the specific excitation of collective coordinates promotes sampling along these often functionally relevant modes of motion, i.e. the advantages of essential dynamics (ED) [61] are used. To counterbalance the disadvantages associated with such a specific excitation, i.e. the construction of biased ensembles, the scheme is embedded within the REX protocol. Thereby ensembles are obtained having approximate Boltzmann statistics and the enhanced sampling properties of REX are utilized. The exchange probability $P(S \to S') \sim \exp\{-N^*_{df}\Delta T\}$ between two replicas crucially depends on the excited number of degrees of freedom $N^*_{df}$ (chapter 3.3). With the stimulated essential subspace $\{es\}$ containing only a *minute* fraction of the total number of degrees of freedom of the system, $\dim\{es\} = N^*_{df} \ll N_{df}$, the bottleneck of low exchange probabilities in all-atom REX simulations is bypassed. For given exchange probabilities, large temperature differences $\Delta T$ can thus be used, such that only a few replicas are required.

Next, the specific protocol used to excite $\{es\}$ modes is discussed which requires substantial changes in the temperature coupling protocol of standard MD simulations.

## 5.2 Temperature Coupling

The temperature[1] coupling of the essential subspace $\{es\}$ is carried out in the following way: Let $N_I$ be the number of atoms of the subsystem of interest by which the eigenvectors $\{\mu_k \in \mathbb{R}^{3N_I} \,|\, k = 1, \ldots, 3N_I\}$ are defined. We denote these atoms *index atoms* to distinguish them from the remaining atoms of the system. The total number of atoms in the system is thus given by $N = N_I + N_R$ ($R$ for *remaining*). By introducing the eigenvectors $\{\mu_k\}$, a second orthonormal basis set besides the usual Cartesian reference frame is established for the description of index atoms. Throughout the following paragraph all vectors (PCA modes, velocities) are written solely within the Cartesian reference frame.

At each timestep, the velocity vector $v_m(t) \in \mathbb{R}^{3N}$ of each non-reference replica $m = 1, \ldots, M$ is split into two parts, describing index atoms and their complement

$$v_m(t) = \begin{pmatrix} v^I_m(t) \\ v^R_m(t) \end{pmatrix},$$

with $v^I_m(t) \in \mathbb{R}^{3N_I}$ and $v^R_m(t) \in \mathbb{R}^{3N_R}$. Next, the velocity vector $v^I_m(t)$ of the index group

[1] Due to the unique assignment of temperatures with replicas in all TEE-REX simulations reported here, the replica index $[i]$, introduced in chapter 3, is dropped henceforth.

is decomposed within the new eigenvector basis set $\{\mu_k\}$ into an essential subspace part $v_m^{es}(t)$ and its complement $\tilde{v}_m^{es}(t)$ (Fig. 5.2)

$$\begin{aligned} v_m^I(t) &= \sum_{i=1}^{3N_I} \left(v_m^I(t)\cdot\mu_i\right)\mu_i \\ &= \sum_{i=1}^{N_{es}} \left(v_m^I(t)\cdot\mu_i\right)\mu_i + \sum_{i=N_{es}+1}^{N_I} \left(v_m^I(t)\cdot\mu_i\right)\mu_i \\ &=: v_m^{es}(t) + \tilde{v}_m^{es}(t)\,. \end{aligned}$$

The velocity projection onto $\{es\}$, $v_m^{es}(t)$, is then coupled to the respective essential subspace temperature $T_m^{es}$ using a Berendsen thermostat,

$$v_m^{es\prime}(t) = \lambda_m v_m^{es}(t)\,, \qquad \lambda_m = \left[1 + \frac{\Delta t}{\tau_m^{es}}\left\{\frac{T_m^{es}}{T_m^{es}(t-\frac{\Delta t}{2})}\right\}\right]^{1/2}. \tag{5.1}$$

All velocity components not coupled to the essential subspace, i.e. $\tilde{v}_m^{es}(t) = v_m^I(t) - v_m^{es}(t)$ and $v_m^R(t)$, are coupled to the reference temperature $T_0$ using any standard coupling algorithm [92, 93, 94]. For the Berendsen thermostat used here, the coupling of the non-essential velocity components within the Cartesian reference frame is given by $\tilde{v}_m^{es\prime}(t) = \lambda_0 \tilde{v}_m^{es}(t)$ and $v_m^{R\prime}(t) = \lambda_0 v_m^R(t)$. Thus, after temperature coupling, the velocity vector $v_m'(t) \in \mathbb{R}^{3N}$ of the full system reads

$$v_m(t) \to v_m'(t) = \begin{pmatrix} v_m^{I\,\prime}(t) \\ v_m^{R\prime}(t) \end{pmatrix} = \begin{pmatrix} \lambda_m v_m^{es}(t) + \lambda_0 \tilde{v}_m^{es}(t) \\ \lambda_0 v_m^R(t) \end{pmatrix}.$$

The reference replica $m = 0$ undergoes a standard MD simulation, since $v_0'(t) = \lambda_0 v_0(t)$. A two dimensional illustration of the temperature coupling of index atoms is given in Fig. 5.2.

## 5.3 Exchange Probability

The coupling of different degrees of freedom to heat baths of different temperature $(T_m^{es}, T_0)$ creates an inherent non-equilibrium situation. Except for the reference replica $m = 0$, the statistical weight of each state in replica $m > 0$ is therefore no longer known. To account for this new situation, the acceptance probability of Eq. (3.7) used for standard REX is modified. The additional kinetic energy (5.1) put locally into the

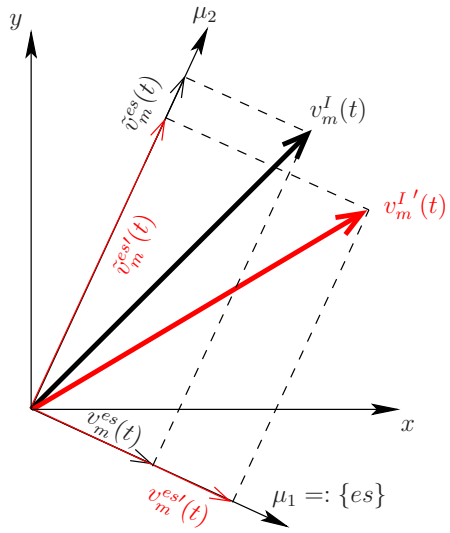


*Figure 5.2:* Essential subspace temperature coupling of the index group atoms, $v_m^I(t) \to v_m^{I\,\prime}(t)$, visualized in two dimensions. After projection onto $\{es\} := \mu_1$, the projected velocity vector of index atoms $v_m^{es}(t)$ is coupled to $T_m^{es}$, resulting in $v_m^{es\prime}(t)$ (red arrow). Velocity components $\tilde{v}_m^{es}(t)$ orthogonal to $\{es\}$ are coupled to the reference temperature $T_0$. The resulting back-projected velocities $v_m^{I\,\prime}(t)$ are biased towards all collective modes of motion set up in $\{es\}$.

few essential degrees of freedom ($N_{es} \ll N_{df}$) is now *thought* to be evenly distributed over the whole system, thus defining an *effective* temperature. Starting from the kinetic energy of replica $m > 0$, $K(v_m) = K^I(v_m^{es}) + K^R(v_m^R, \tilde{v}_m^{es}(t))$, and using the equipartition theorem

$$2K = N_{df} k_B T_{eff}\,, \quad 2K^I = N_{es} k_B T_m^{es}\,, \quad 2K^R = (N_{df} - N_{es}) k_B T_0$$

for the different contributions, we arrive at the effective temperature for each non-reference replica,

$$T_m^{eff} = \left(1 - \frac{N_{es}}{N_{df}}\right) T_0 + \frac{N_{es}}{N_{df}} T_m^{es} = T_0 + \frac{N_{es}}{N_{df}} (T_m^{es} - T_0)\,. \tag{5.2}$$

$N_{df}$ hereby denotes the degrees of freedom of the complete system. Given Eq. (5.2), the modified acceptance criterion used in TEE-REX thus reads

$$P(S \to S') = \min\left\{1, \exp\left[(\beta_m^{eff} - \beta_n^{eff})\ (E(x_m) - E(x_n))\right]\right\}. \tag{5.3}$$

It is $\beta_0^{eff} \equiv \beta_0$ for the reference replica $m = 0$. By replacing $\beta_m \to \beta_m^{eff}$ in Eq. (3.7) of the standard replica exchange criterion, one implicitly assumes that the ensemble created by each non-reference replica can be described by an equilibrium Boltzmann distribution at the effective temperature introduced in Eq. (5.2). Since each non-reference replica by construction samples some unknown non-equilibrium distribution, this approximation introduces—upon exchange with the reference replica—some bias in the statistics of the reference ensemble $m = 0$. However, the number of degrees of freedom of the

complete system is much larger than the few excited degrees of freedom comprising the essential subspace $\{es\}$ ($N_{df} \gg N_{es}$). Hence, the approximation made in Eq. (5.2) can be considered a small deviation from an equilibrium distribution and, therefore, can be expected to be valid for all but the smallest systems simulated with TEE-REX. Moreover, to quantify the bias introduced by this scheme and to assess the extent to which the TEE-REX algorithm approximates a Boltzmann distribution, extensive tests were carried out with a test system for which a converged Boltzmann distribution was available for comparison (chapter 6).

## 5.4 Essential Subspace Composition

The composition of the essential subspace (i.e. what modes have been chosen) is in principle irrelevant with respect to the definition of $T_m^{eff}$. However, the excitations obtained using a specific $\{es\}$ naturally depend on the choice of modes. Each PCA mode represents a single (collective) degree of freedom, contributing via equipartition—like any other degree of freedom—to the kinetic energy. This is independent of whether the respective mode describes a global transition or a more localized motion (e.g. involving a loop). Here, it is important to note that PCA modes describe linearly independent collective modes, thereby neglecting non-linear couplings. If one specific eigenvector is excited, several other modes are indirectly excited, either outside the $\{es\}$ (like side chains) or inside the essential subspace. This feature has influence on sampling performance, not only along $\{es\}$ modes but also along PCA modes indirectly linked to the former (chapter 6.3 and Table 6.1).

The fact that the choice of modes for $\{es\}$ is not restricted in any way makes the TEE-REX algorithm quite versatile with regard to biomolecular applications. In chapter 7, the algorithm is successfully applied to the problem of simulating large conformational transitions in proteins, a challenging task for all-atom MD simulations. Along the same line are questions concerning allostery, where conformational changes in tertiary and quarternary structure are important. These may be addressed in future applications. Besides the composition of the essential subspace, also the specific nature of the chosen modes can be varied. In this work, principal component analysis is used for the calculation of modes, but also NMA or the recently developed full correlation analysis [145] can be utilized for TEE-REX, offering modes possessing different properties.

# 6 Benchmarking TEE-REX

*I think animal testing is a terrible idea;*
*they get all nervous and give the wrong answers.*
— Unknown

With the TEE-REX algorithm in place, its accuracy and performance with respect to REX and MD has to be evaluated. In particular, two things need attention: First, in order to validate the ensemble approximation made in Eq. (5.2), extensive tests of the TEE-REX protocol were made using a dialanine peptide (Fig. 6.1) to rigorously validate TEE-REX generated ensembles. As a converged MD ensemble is available for this system, it allows us to quantitatively assess any systematic deviations from a canonical ensemble possibly introduced by the TEE-REX protocol. Second, the sampling efficiency of TEE-REX was assessed using a small peptide. We start by reporting on simulation details.

## 6.1 Simulation Details

All simulations were carried out using the MD software package GROMACS 3.3.1 [80], supplemented by the TEE-REX module. The OPLS-all atom force field [84] was used for proteins and TIP4P was used as a water model [153]. All simulations were performed in the *NPT* ensemble. In all MD simulations the temperature was kept constant at $T = 300\,\mathrm{K}$ by coupling to an isotropic Berendsen thermostat [94] with a coupling time of $\tau_t$=0.1 ps. The pressure was coupled to a Berendsen barostat [94] with $\tau_p$=1 ps and an isotropic compressibility of $4.5 \cdot 10^{-5}$ bar$^{-1}$ in the $x$, $y$ and $z$ directions. All bonds were constrained by using the LINCS algorithm [31]. An integration time step of $\Delta t = 2\,\mathrm{fs}$ was used. Lennard-Jones and Coulombic interactions were calculated explicitly at a distance smaller than 10 Å; above 10 Å, long-range electrostatic interactions were calculated by particle mesh Ewald (PME) summation [88], with a reciprocal grid spacing of 0.12 nm and fourth-order B-spline interpolation.

### MD Simulations

The dialanine reference simulation system was set up as follows. PyMOL [154] was used to build an $N$-acetylated dialanine to prevent electrostatic attraction between the $N$- and the $C$-terminus. The peptide was solvated in a rhombic dodecahedral box with box vectors of 2.35 Å length. The system comprised ∼1200 atoms. $Na^+$ ions were added accordingly to neutralize the system. Energy minimization of the solvated system using the steepest descent algorithm was followed by a 100 ps MD simulation at the target temperature using harmonic position restraints on the heavy atoms of the peptide with a force constant of $k = 1000\,\mathrm{kJmol^{-1}nm^{-2}}$ to equilibrate the solvent. After one ns of equilibration, a 4.1 $\mu$s trajectory was produced by unbiased MD simulation. Structures were saved every 1 ps for further analysis.

### TEE-REX Simulations

Four 210 ns TEE-REX simulations of dialanine starting from different equilibrated MD structures were performed. Each TEE-REX simulation consisted of two replicas, with an essential subspace temperature of 500 K for the second replica whereas the first, reference replica was run at 300 K. A PCA was performed on the first 1.87 $\mu$s of the reference MD trajectory, taking all backbone atoms into account. The first two eigenvectors, describing 92% of all backbone fluctuations, were defined to describe the essential subspace. The $\{es\}$ was coupled to a Berendsen thermostat with a coupling time of $\tau_m^{es} = \Delta t = 2\,\mathrm{fs}$. Exchanges between replicas were attempted every $\nu_{ex}^{-1} = 140\,\mathrm{ps}$ and were accepted on average with 97.7% probability. Structures were saved every 1 ps. After each successful exchange 40 ps of trajectory were discarded from analysis.

### Free Energy

Free energy landscapes of dialanine were calculated in the subspace spanned by the first two eigenvectors (essential subspace $\{es\}$). Assuming equilibrated ensembles, the relative Gibbs free energy

$$\Delta G(x_i, y_j) = -k_B T \ln \left[ \frac{P(x_i, y_j)}{P_{\min}} \right] \tag{6.1}$$

was calculated for discrete grid points $(x_i, y_j)$ using a $k$-nearest neighbor scheme [72] for the spatial probability function $P(x_i, y_j)$. It is $P_{\min} := \min \{P(x_i, y_j)\}$.

## 6.2 Statistical Ensemble

To probe the ensemble generated by TEE-REX, a 4.1 $\mu$s explicit-solvent MD simulation of the dialanine peptide (Fig. 6.1) was compared to four 210 ns TEE-REX simulations of the same system (see chapter 6.1 for computational details).

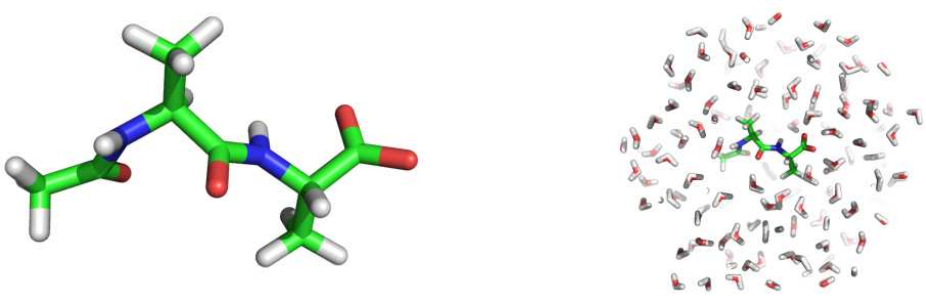

*Figure 6.1:* *N*-acetylated dialanine molecule (left) and corresponding simulation system (right), containing ∼1200 atoms.

Dialanine was chosen since it constitutes one of the smallest systems with a non-trivial configuration space. Because of its small size extensive trajectories can be generated within a reasonable amount of time. The main motions of dialanine occur around its ($\phi$,$\psi$)-pair of dihedrals, hence, the available configuration space of the system is very limited. This increases chances to achieve *complete* sampling with conventional MD simulations, i.e. a full coverage of the available phase space. Furthermore, deviations from the equilibrium distribution of the reference replica due to the input of configurations excited by the essential subspace $\{es\}$ can be expected to be largest for small systems. This is because the influence exerted by the stimulated $\{es\}$ increases with the number of modes, $N_{es} = \dim\{es\}$. The fraction $N_{es}/N_{df}$ can thus be taken as a measure for the influence caused by the excitation of modes. For dialanine, the fraction $N_{es}/N_{df} \sim 10^{-3}$ is at least one order of magnitude larger than for systems usually simulated.

The thermodynamic behavior of a system is completely known once a thermodynamic potential such as the Gibbs free energy is available. Comparing free energies thus enables us to decide to which degree ensembles created by both methods coincide. However, calculating relative free energies according to Eq. (6.1) requires a converged ensemble. Therefore, as a first step, we checked whether the MD reference trajectory yielded a converged ensemble, i.e. if a complete sampling of the configuration space of the system was obtained.

### 6.2.1 Convergence of the MD Reference

As a first test, structural convergence was examined using the eigenvector inner product matrix $\mu_i^A \cdot \mu_j^B$. Backbone eigenvectors $\{\mu_i^A\}$, obtained from a PCA of the full 4.1 $\mu$s MD trajectory, were compared to eigenvector sets $\{\mu_i^B\}$, calculated from trajectory fragments of length 180 ns to 1.87 $\mu$s (Fig. 6.2). Then, subspaces spanned by the first four eigenvectors of each set were constructed. Therein, 97 % of all backbone fluctuations are covered. Next, overlaps of these different subspaces with the subspace of the full trajectory were calculated. The subspace overlap between the $m$ orthonormal (ON) vectors $\{w_1, \ldots, w_m\}$ and the reference subspace spanned by the $n$ ON vectors $\{v_1, \ldots, v_n\}$ is given by $n^{-1} \sum_{i=0}^{n} \sum_{j=0}^{m} (v_i \cdot w_j)^2$. The overlap will increase with increasing $m$ and equals one when the set $\{v_i\}$ is a subspace of set $\{w_j\}$. Results indicate that structural convergence is reached for trajectory fragments of lengths ≥400 ns (measured subspace overlap of > 99 %).

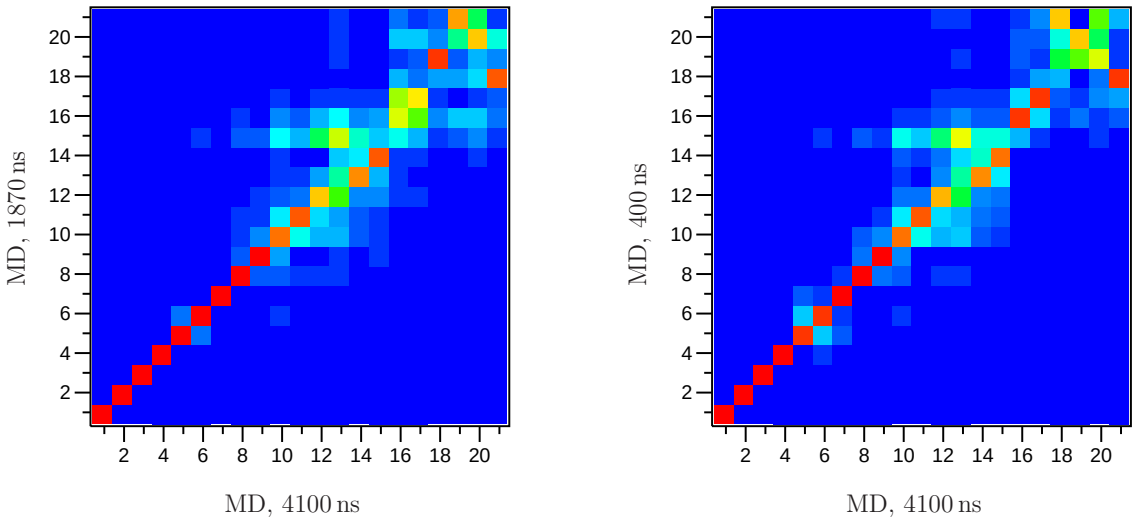


*Figure 6.2:* Eigenvector (EV) inner product matrices $\mu_i^A \cdot \mu_j^B$ for EV sets $\{\mu_i^{A/B}\}$ derived from different MD trajectory fragments. Color coding of matrix elements ranges from blue (orthogonal EV, $\mu_i^A \perp \mu_j^B$) to red (parallel EV, $\mu_i^A \parallel \mu_j^B$). Left: fully converged EV sets from a 1.87 $\mu$s MD piece ($y$ axis) vs. the 4.1 $\mu$s MD trajectory ($x$ axis). Right: EV set from a 400 ns MD piece ($y$ axis) vs. the 4.1 $\mu$s MD reference. Subspace overlaps are calculated for the subspace comprising the first four EVs (lower left red diagonal elements).

As a second test for convergence, transitions between the two main dialanine conformations were counted. Fig. 6.3*B* shows representative structures found along the system path overlayed onto a two-dimensional free energy surface derived from a 420 ns

MD trajectory piece. The main motion of the system is a rotation around its only dihedral pair around the $C_\alpha - C$ bond between the $C_\alpha$ atom of $\mathrm{Ala}_1$ and the carbon atom of the second peptide unit. Starting from an "open" conformation (with respect to the distance of the $N$- and $C$-termini) in the left basin (principal component $p_1 \leq -0.1$), a transition to a "closed" conformation in the right basin (principal component $p_1 \geq 0.2$) takes place. During the 4.1 $\mu$s of MD simulation time, more than 900 transitions between the "open" and the "closed" conformation were observed, giving further evidence for a converged ensemble covering the complete configuration space.

As a third test for convergence we evaluated relative free energy landscapes for dialanine ensembles generated by MD and TEE-REX (see below).

### 6.2.2 Ensemble Comparison – Free Energy Landscape

Ensembles generated by MD and TEE-REX were compared in terms of relative Gibbs free energy landscapes $\Delta G(x, y)$ calculated from trajectory projections onto the two-dimensional essential subspace $\{es\}$ excited in all dialanine TEE-REX simulations (Fig. 6.3). A 1870 ns piece of the full 4.1 $\mu$s reference MD trajectory was used to define the $\{es\}$ eigenvectors (see chapter 6.1). We used the information that ensembles from trajectory parts of length $\geq$400 ns are converged (chapter 6.2.1) to define nine independent non-overlapping 420 ns MD trajectory fragments out of the full 4.1 $\mu$s MD reference. To guarantee equal computational effort, the length of a single two-replica TEE-REX simulation was thus set to 210 ns. In this way we compare ensembles that were generated with the same computational effort. Four 210 ns two-replica TEE-REX simulations with replica temperatures $(T_m^{es}, T_0)$ of (300 K, 300 K) for the first and (500 K, 300 K) for the second replica were started from different MD snapshots taken from the full MD trajectory to check for any dependence of the sampling on the starting structure.

The upper panels of Fig. 6.3 show typical Gibbs relative free energy surfaces (units of kJ/mol) for TEE-REX (panel *A*) and MD (panel *B*) ensembles with respect to the first two backbone eigenvectors comprising the essential subspace $\{es\}$. The observed ring structure seen in all ensembles is due to the fact that a non-linear dihedral rotation is described by two orthogonal linear PCA coordinates. Two distinct conformations are distinguishable, an "open" conformation located in the left minimum of the $\Delta G$ surface and a "closed" conformation located in the right minimum. Transitions between the two conformations occur along the free energy "valley" (upper pathway), illustrated by representative structures shown in Fig. 6.3*B*. A free energy barrier of $\sim$15 kJ/mol

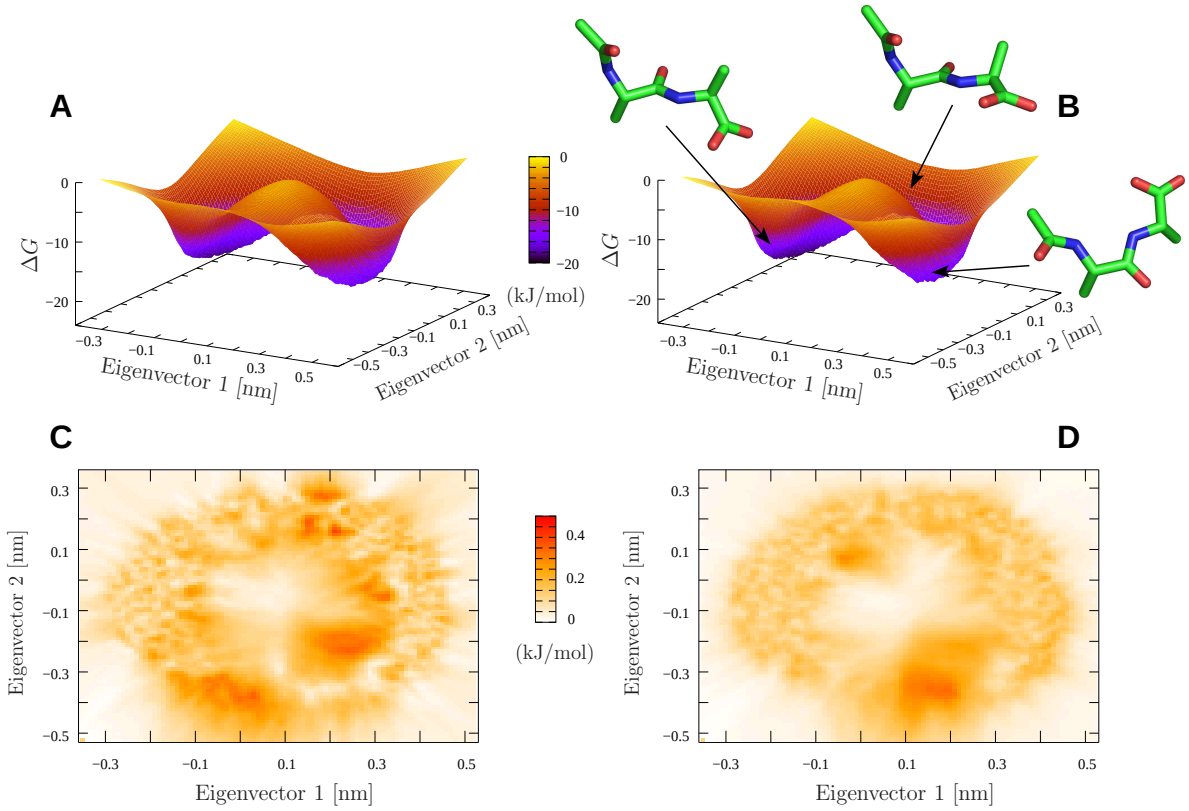


*Figure 6.3:* Comparison of dialanine ensembles generated by TEE-REX and MD. Gibbs relative free energy surfaces (in units of kJ/mol) projected onto the first two MD-derived backbone eigenvectors ($\{es\}$) are shown for a TEE-REX ensemble (panel *A*) and an ensemble from a 420 ns MD trajectory piece (panel *B*), overlaid by representative structures found along the system pathway. All calculations were carried out on an equal number of 40000 samples. Panels *C* and *D*: standard deviations (top view, units of kJ/mol) $\sigma_{\text{TEE-REX}}$ (panel *C*) and $\sigma_{\text{MD}}$ (panel *D*), calculated for all four TEE-REX and all nine MD free energy surfaces, respectively.

(saddle) impedes the conformational transition along the lower pathway. No apparent visual difference between the free energy surfaces determined by the two methods is seen, indicating that TEE-REX creates ensembles very similar to that created by MD.

Fig. 6.3*C-D* display standard deviations $\sigma_{\text{TEE-REX}}$ and $\sigma_{\text{MD}}$ (in units of kJ/mol), calculated from all four TEE-REX and all nine MD $\Delta G$ surfaces, respectively. The statistical error is less than 0.4 kJ/mol≈0.15 $k_B$T for both methods and thus very low with respect to the absolute $\Delta G$ values. This further supports the assumption of converged ensembles in both cases. In the case of MD (panel *D*), the largest statistical errors are found in the saddle region, which hinders conformational transitions along the lower pathway. These comparatively large errors are due to the poor sampling in this part of the configuration space, since barrier heights of 15 kJ/mol are rarely overcome by MD during

420 ns of simulation time. While the central region is not sampled by MD (Fig. 6.3*D*), panel *C* shows that TEE-REX explores this region, indicating the ability of the latter to sample high-energy regions more frequently than MD. When comparing Fig. 6.3*C*-*D*, it is important to note that $\sigma_{\text{TEE-REX}}$ was constructed using four samples, whereas nine MD samples were used for $\sigma_{\text{MD}}$.

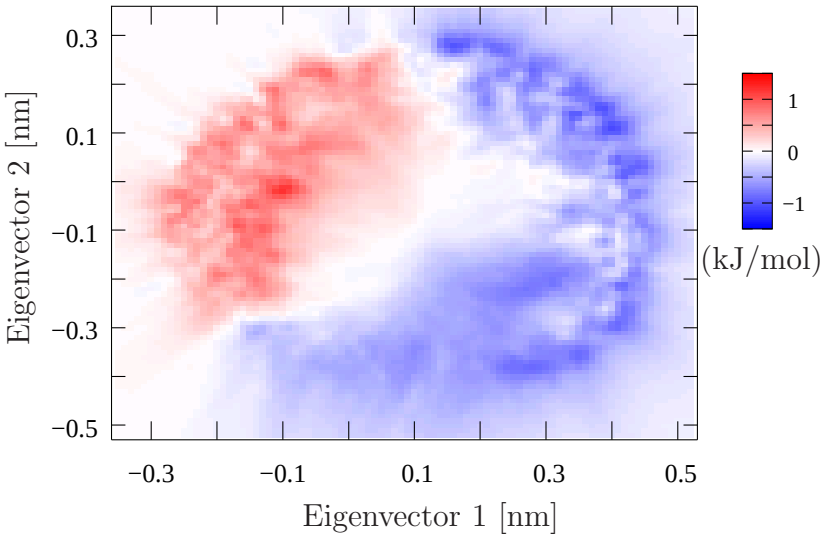


*Figure 6.4:* Top view of the difference in free energy $\langle \Delta G_{\text{TEE-REX}} - \Delta G_{\text{MD}} \rangle$, averaged over all combinations $\Delta G^i_{\text{TEE-REX}} - \Delta G^j_{\text{MD}}$. See text for details.

To investigate the shape of the free energy surfaces generated by both methods in detail, Fig. 6.4 displays the difference $\langle \Delta G_{\text{TEE-REX}} - \Delta G_{\text{MD}} \rangle$ averaged over all combinations $\Delta G^i_{\text{TEE-REX}} - \Delta G^j_{\text{MD}}$ ($i = 1, \ldots, 4$; $j = 1, \ldots, 9$). Areas colored in blue are sampled more frequently by TEE-REX than by MD, since $\Delta G_{\text{TEE-REX}} < \Delta G_{\text{MD}}$ in these areas. The maximum absolute deviations of $1.5\,\text{kJ/mol} \simeq 0.6\,k_B\text{T}$ from the ideal case $\Delta G_{\text{TEE-REX}} - \Delta G_{\text{MD}} = 0$ are commensurate with the maximum statistical errors of $0.15\,k_B\text{T}$ (Fig. 6.3) found for each method. As can be seen from the distribution of blue regions, high energy configurations are more frequently sampled by TEE-REX, whereas MD sampling focuses on the stretched low energy basin containing the open conformation. Thus, the excitation of essential subspace modes allows the TEE-REX reference replica to explore high energy configurations usually not available to a normal MD sampling at the same temperature.

## 6.3 Sampling Efficiency

To analyze the sampling efficency of the TEE-REX algorithm, the 13 amino-acid peptide hormone guanylin [155] was simulated by both MD and TEE-REX. To provide meaningful statements about sampling efficiency, two independent 60 ns trajectory fragments from the 130 ns TEE-REX reference replica were compared to four independent 180 ns=3×60 ns MD trajectory fragments taken from one 800 ns MD trajectory. Besides employing projections onto eigenvectors drawn from the essential subspace $\{es\}$, both methods were compared using $(\phi, \psi)$ dihedral space.

It is generally accepted that standard REX improves sampling efficiency over classical MD. However, the computational effort associated with explicit solvent simulations is often very high with respect to the gain in sampling. Initial tests with standard temperature REX simulations of guanylin showed only a slight increase in sampling performance over classical MD. In particular, the computational effort of a 4-replica REX simulation of 60 ns length was 25 % larger than the 3-replica TEE-REX simulations of comparable length. Therefore we omit REX and directly compare results from MD with TEE-REX.

### 6.3.1 Simulation Details

#### MD Simulations

The MD reference simulation system of guanylin was set up as follows. From a standard REX simulation a snapshot of the 300 K reference replica served as the MD starting structure. The simulation system is based on the protonated NMR structure (Protein Data Bank (PDB) entry 1GNA), solvated in a rhombic dodecahedral box and neutralized adding $Na^+$ ions accordingly. The system comprised ∼6000 atoms. Energy minimization of the solvated system with the steepest descent algorithm was followed by a 100 ps MD simulation at the target temperature using harmonic position restraints on the heavy atoms of the protein with a force constant of $k = 1000\,\mathrm{kJ\,mol^{-1}nm^{-2}}$ to equilibrate the solvent. After one ns of equilibration, a 800 ns trajectory was produced by free MD simulation. Structures were saved every 2 ps for further analysis.

#### TEE-REX Simulations

One 130 ns TEE-REX simulation of guanylin starting from an equilibrated MD structure was performed. Three replicas were simulated, having essential subspace tem-

peratures of 450 K and 800 K. A PCA of a 50 ns MD trajectory fragment taking all backbone atoms into account was performed. The first six eigenvectors, describing 87 % of all backbone fluctuations, defined the essential subspace $\{es\}$. Exchanges were attempted every $\nu_{ex}^{-1} = 160$ ps and were accepted with 97.8 % probability. Structures were saved every 1 ps. A 40 ps trajectory snippet was discarded after each successful exchange to yield equilibrated ensembles.

### 6.3.2 Essential Subspace

Every MD and TEE-REX reference ensemble was projected onto the first two backbone eigenvectors of the six-dimensional essential subspace $\{es\}$ used in the TEE-REX simulation. Together, both eigenvectors describe 64% of all backbone fluctuations of the system. In Fig. 6.5, all of these projections are displayed, together with their respective starting structures (red diamonds). Fig. 6.5*A* shows the configuration space sampled by a 180 ns fragment of MD trajectory ranging from 20-200 ns. The intensely sampled region in the upper half of the $\mu_1\mu_2$-plane indicates a pronounced local minimum in the free energy surface of the system. For the remaining 600 ns of simulation time, the MD simulation gets trapped in this region of configuration space (Panel *B*-*D*). Projections of both 60 ns fragments of the 130 ns TEE-REX reference replica trajectory, ranging from 5-65 ns and 70-130 ns, are shown in Fig. 6.5*E*-*F*. Although the starting structure lies within the local minimum amply sampled by MD (Panel *E*), the space captured by TEE-REX not only covers that explored by MD, but extends beyond that. This result is independent from the starting structure, as a projection of the second 60 ns TEE-REX reference trajectory fragment confirms (Panel *F*).

To quantify TEE-REX sampling performance, the time evolution of sampled configuration space volumes $V_i(\tau)$ was measured using projections of all MD and TEE-REX trajectory fragments along the first two eigenvectors of the six-dimensional essential subspace $\{es\}$ excited in the TEE-REX simulation. In order to monitor time evolution, the $\mu_1\mu_2$-plane (Fig. 6.5) was discretized by a grid with a spacing of 0.01 nm. At each time step, the number of occupied grid cells was recorded. Conversion of time into computational effort $\tau$ (measured in units of 180 ns MD simulation time) yielded the $V_i(\tau)$ curves shown in Fig. 6.6. Panel *A* compares TEE-REX sampling performance $V_{\text{TEE-REX}}(\tau)$ (solid lines) against MD sampling curves $V_{\text{MD}}(\tau)$ (dotted lines) for all 180 ns MD trajectory fragments of the 800 ns reference MD simulation.

Apart from the the first 200 ns of simulation time, the sampling performance of MD is quite limited compared to TEE-REX. Here, the dependence of the MD sampling on

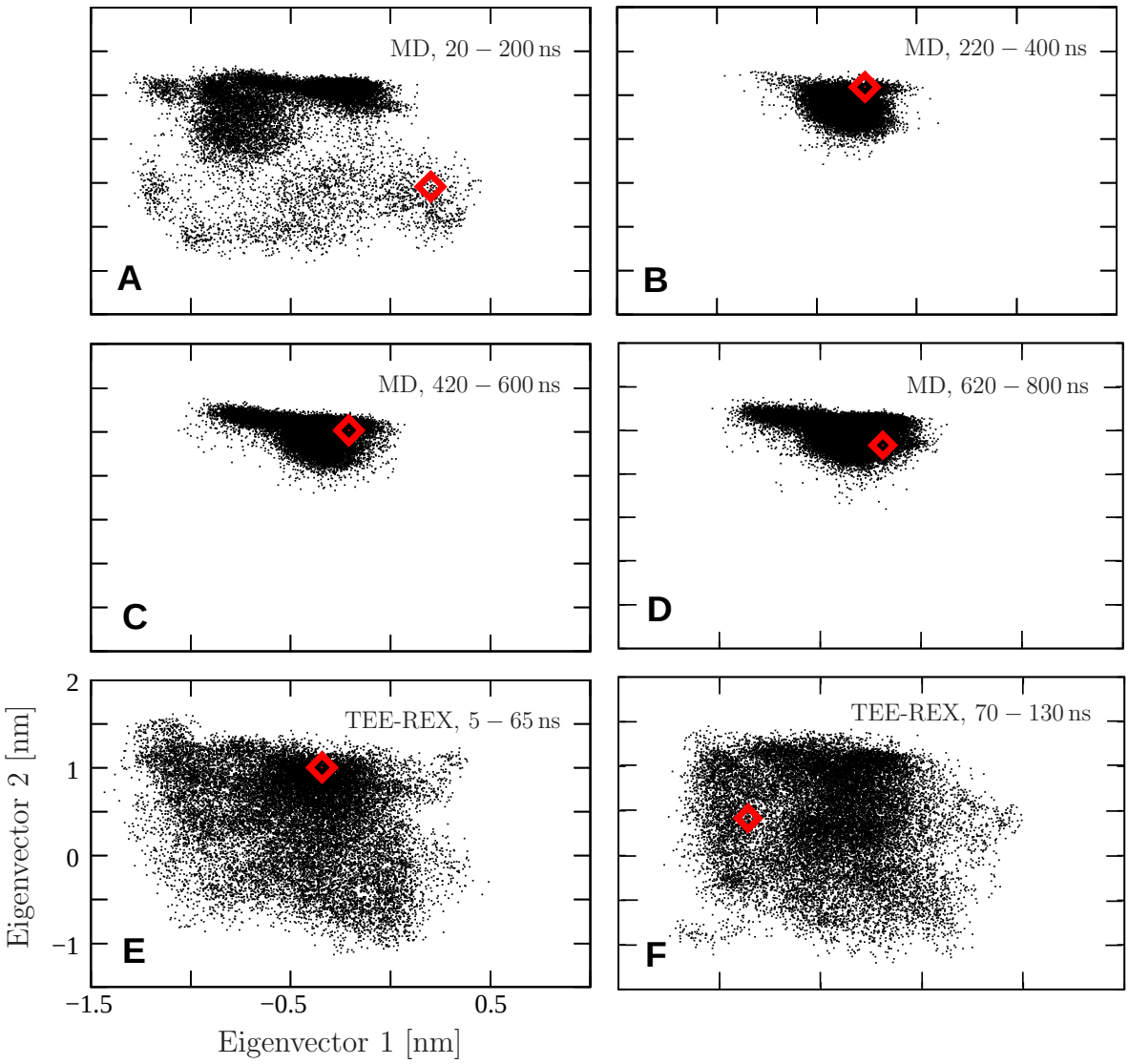


*Figure 6.5:* Trajectory projections of guanylin MD and TEE-REX simulations on the first two eigenvectors used in the TEE-REX simulation (axes' labels are only shown for panel *E*). Red diamonds represent the starting structure of each simulation window. Panels *A*-*D*: projection of MD ensembles for four different time spans. Panel *E*-*F*: TEE-REX ensemble for both 60 ns pieces.

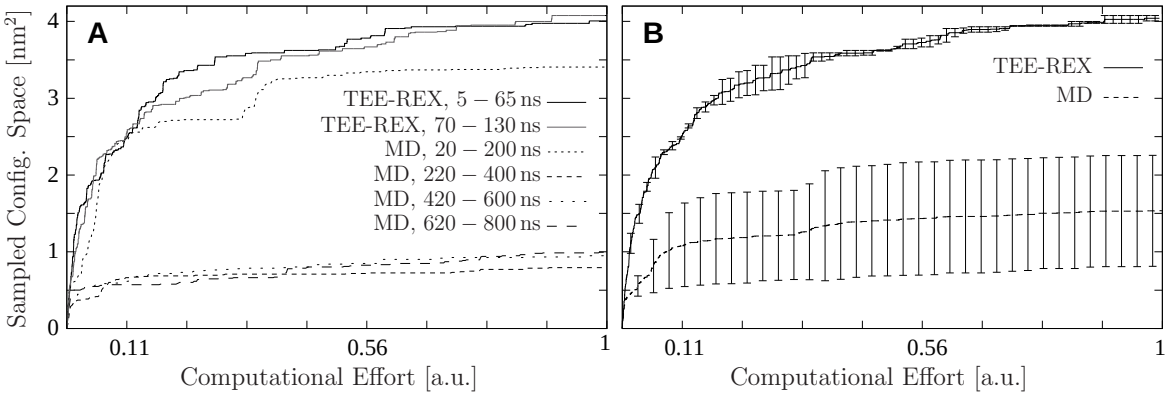


*Figure 6.6:* Quantitative comparison of TEE-REX sampling performance with respect to MD for a guanylin test system. Sampled configuration space volumes $V_i(\tau)$ are measured versus computational effort $\tau$ (in units of 180 ns MD simulation time) for trajectory projections onto the first two eigenvectors of the six-dimensional $\{es\}$ excited in the TEE-REX simulation of guanylin. Panel *A*: TEE-REX performance (dark and light solid lines) versus MD performance (dashed lines). Panel *B*: Average TEE-REX (solid) and MD (dashed) sampling performance $\langle V_i(\tau)\rangle \pm \sigma_i$ with error bars denoting standard deviations $\sigma_i$.

the starting structure becomes clearly visible. For TEE-REX, sampling performance is independent of the starting structure, displaying the ability of the method to efficiently explore large regions of configuration space within short simulation times. Fig. 6.6*B* summarizes the results of Fig. 6.6*A*, showing average TEE-REX (solid line) and MD (dashed line) performance $\langle V_i(t)\rangle \pm \sigma_i$, with errorbars representing standard deviations $\sigma_i$. In the 180 ns MD simulation windows of guanylin, on average only 10% ($\tau = 0.1$) of the total computational effort is necessary to sample 80% of the configuration space available to MD. Thus, exploring the remaining 20% of configuration space is computationally very expensive. For TEE-REX, we see a 3.6-fold[1] increase in sampled configuration space using the same computational effort $\tau = 0.1$. Although the sampling rate of TEE-REX decreases with increasing $\tau$, it outperforms the MD sampling rate by a factor of three.

### 6.3.3 Dihedral Space

In order to evaluate the sampling performance of TEE-REX in subspaces not related to the essential subspace $\{es\}$, ensembles of both methods were compared within full $(\phi, \psi)$

[1] A different measure for sampling performance is given by the fact that TEE-REX requires around 5 % of the computational effort to sample $V_{\mathrm{MD}}(\tau = 1)$, resulting in a more than 20-fold gain in sampling over MD.

dihedral space. Panels *A*-*C* of Fig. 6.7 show Ramachandran plots [156] of several 180 ns fragments of MD trajectory. In all three fragments the left half plane $\phi \in [-180°, 0°[$ is well-sampled by MD, whereas moderate sampling is achieved in the remaining half plane $\phi \in [0°, 180°[$. For the corresponding TEE-REX ensemble (Fig. 6.7*D*) a substantial increase in sampling is seen, where a notably broader range of $\psi$ values is sampled by TEE-REX.

*Table 6.1:* Average TEE-REX sampling efficiency for guanylin, calculated in different two-dimensional subspaces. The efficiency measured in parts of the excited essential subspace, $\{\mu_1, \mu_2\} \subset \{es\}$, is shown for comparison.

| Subspace | Efficiency Gain |
|---|---|
| $(\phi, \psi)$ | 2.43 |
| $\{\mu_7, \mu_8\}$ | 2.80 |
| $\{\mu_{14}, \mu_{15}\}$ | 2.62 |
| $\{\mu_1, \mu_2\} \subset \{es\}$ | 3.65 |

For a more detailed analysis the volume $V(\tau = 1)$ explored in dihedral space was calculated for each of the eleven pairs of dihedrals in all four MD and two TEE-REX ensembles. The average gain in sampling efficiency $\langle V_{\text{TEE-REX}}/V_{\text{MD}} \rangle$ for $(\phi, \psi)$ space is shown in Table 6.1 together with results from additional analyses, made on two PCA subspaces linearly independent from the $\{\mu_1, \mu_2\} \subset \{es\} = \{\mu_1, \ldots, \mu_6\}$ space, namely $\{\mu_7, \mu_8\}$ and $\{\mu_{14}, \mu_{15}\}$. For all subspaces independent from $\{es\}$, sampling performances are comparable, yielding an approximately 2.5-fold gain in TEE-REX sampling efficiency over classical MD. Although these values are lower than the observed 3.6-fold performance gain measured in the $\{\mu_1, \mu_2\}$ subspace, it clearly demonstrates the capability of TEE-REX as an efficient sampling method.

## 6.4 Defining $\{es\}$ Using Sparse Structure Information

The sampling enhancement in TEE-REX is largely due to excitations of the essential subspace $\{es\}$. Hence, the question arises how sampling performance is influenced by the definition of $\{es\}$, i.e. the amount of available structural information from which collective coordinates are constructed. Often, such information is limited and hard to get, either through experimental (few or no X-ray/NMR structures) and/or computational[2] restrictions. We therefore tested the algorithm's sampling power under the condition that only few configurations of a system are available.

[2]Starting from a single structure, the CONCOORD method [157] allows to generate structural ensembles at low computational costs, thereby alleviating lack of structural information. However, the statistical weight of each structure is thereby unknown.

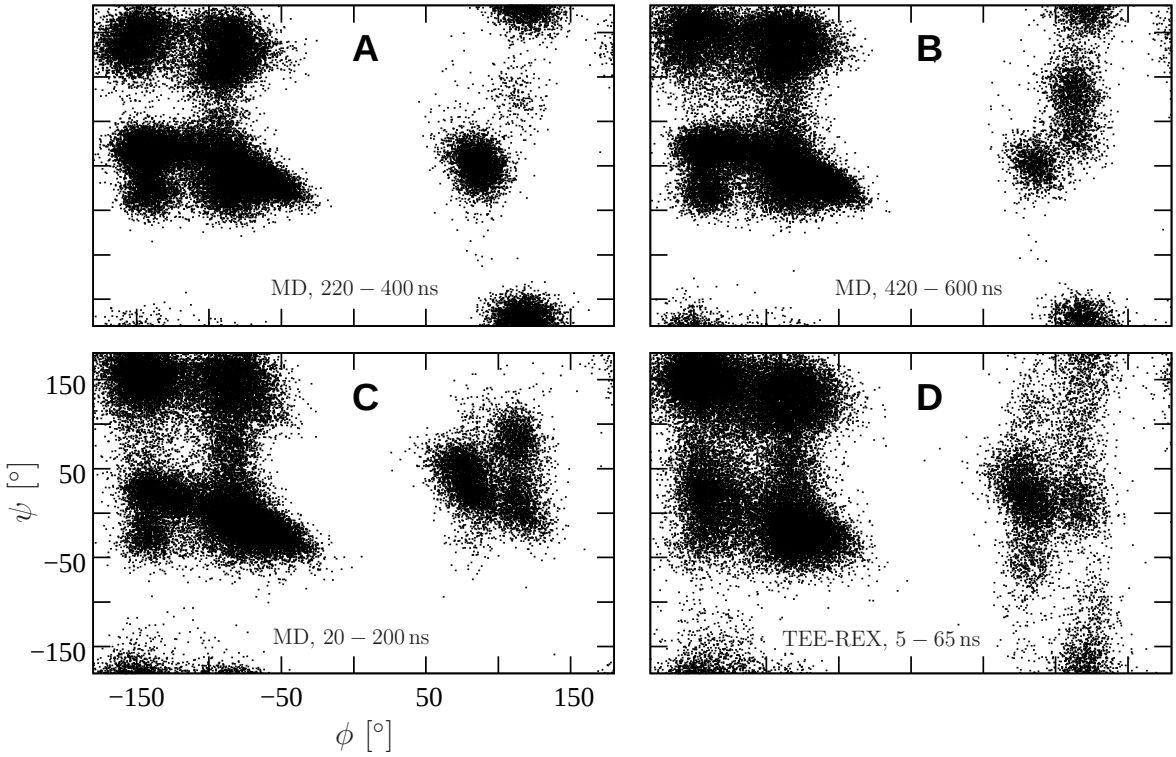


*Figure 6.7:* Ramachandran plots of different guanylin MD and TEE-REX ensembles (axes' labels are only shown for panel *C*). MD ensembles for different 180 ns time windows (panels *A*-*C*) are compared with a TEE-REX ensemble, ranging 5-65 ns (panel *D*). Enhanced sampling of TEE-REX with respect to MD is observed.

To mimic sparse structural information, a 130 ns TEE-REX simulation of guanylin was performed using an essential subspace $\{es\}'$, constructed using eigenvectors obtained from a PCA on the backbone atoms of an MD trajectory spanning only 1 ns. This corresponds to a 50-fold reduction of available structures with respect to the original definition of $\{es\}$. Compared to the six eigenvectors used originally, the first ten eigenvectors were necessary in the construction of $\{es\}'$ to account for the same 87 % of all observed backbone fluctuations. Calculating the subspace overlap between both essential subspaces shows that $\{es\}'$ can reproduce about 70 % of the configurational space covered by $\{es\}$. Looking at the modes itself, we found that collective modes $\mu'_j \in \{es\}'$ differ markedly from the well-defined modes $\mu_i \in \{es\}$ (Fig. 6.8). When this subspace $\{es\}'$ is taken as a basis for TEE-REX, despite the substantial reduction of structural information, projections of 60 ns trajectory pieces from both TEE-REX simulations onto the first two eigenvectors of $\{es\}$ reveal only minor differences in sampled regions of configuration space (Fig. 6.9) as compared to the original. Comparing sam-

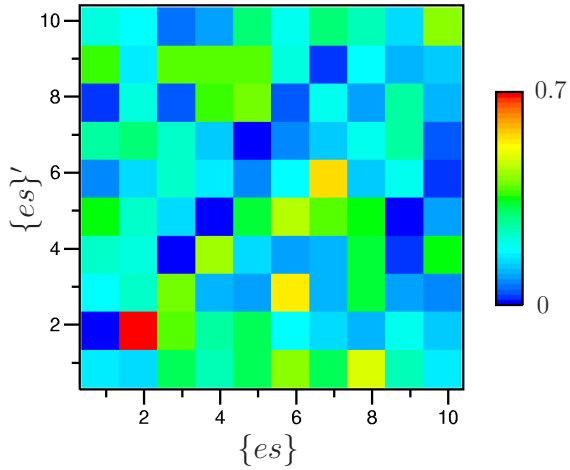


*Figure 6.8:* Inner product matrix $\mu_i \cdot \mu'_j$, showing the difference of collective modes of motion present in $\{es\}'$ and $\{es\}$. Compared to $\{es\}$, different modes are contained within $\{es\}'$, resulting from lack of structural information available for the construction of $\{es\}'$.

pled configuration space volumes measured over computational effort yields an average difference of 7 % in sampling efficiency within the essential subspaces. These results indicate that TEE-REX sampling efficiency is hardly sensitive to the choice of the essential subspace. To further validate these findings the overlap of both ensembles in

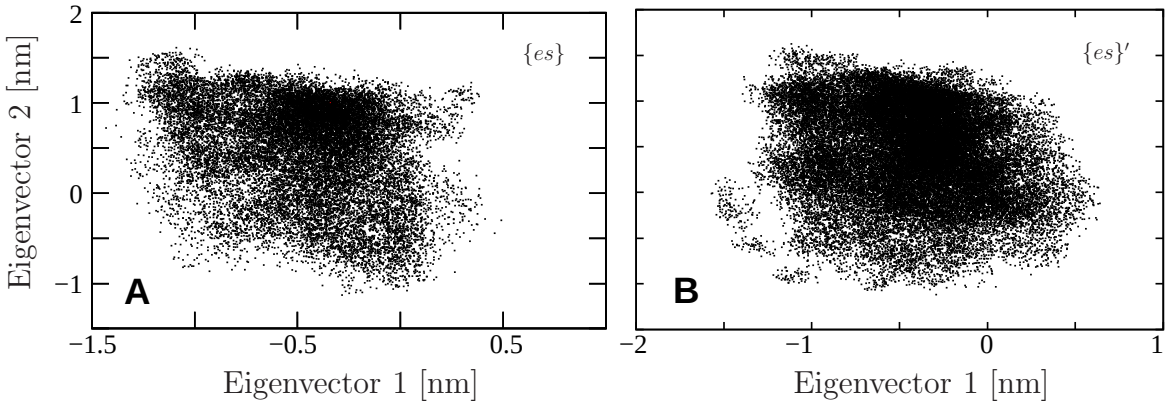


*Figure 6.9:* Influence of TEE-REX sampling performance on the amount of structural information present in the essential subspace. Only minor differences in sampling are seen between the structurally well-defined $\{es\}$ (panel *A*) and $\{es\}'$ (panel *B*), derived from only 2 % of the structures used for $\{es\}$.

full $(\phi, \psi)$ dihedral space was estimated to have a PCA-independent measure. To this end, the $(\phi, \psi)$ plane was discretized by a grid of size 1° and the grid cells shared by both ensembles were counted, yielding an overlap of more than 84 %.

## 6.5 Algorithm Sensitivity

During development, extensive tests were made with the algorithm to elucidate its sensitivity with respect to the three main parameters: essential subspace temperature $T_m^{es}$, size of the essential subspace $N_{es}$, and exchange attempt frequency $\nu_{ex}$.

Excitations of the chosen $\{es\}$ are controlled by $T_m^{es}$ and the corresponding coupling constant $\tau_m^{es}$ (Eq. (5.1), page 32), defining the coupling strength. Both parameters are not independent of each other since for a weak coupling $\tau_m^{es} \gg \Delta t$, dissipation of the excitation energy to colder degrees of freedom (i.e. degrees of freedom coupled to $T_0$) leads to a lower $\{es\}$ temperature and hence reduced efficiency in sampling. Thus, a higher subspace temperature needs to be chosen to achieve the same sampling efficiency as with a tight coupling and a lower $\{es\}$ temperature. Values for both of these parameters depend on the particular system and were chosen such as to find an optimal compromise between sampling efficiency and accuracy. Except for applications dealing with folding/unfolding, combining a high excitation temperature $T_m^{es} \gtrsim 700\,\mathrm{K}$ with a strong coupling constant $\tau_m^{es} \approx \Delta t$ should be avoided. In such cases the coupling to the heat bath is very strong and only a small amount of the excitation energy put into the $\{es\}$ is actually dissipated through other modes. As a consequence, a rapid loss in tertiary and secondary structure is observed within a few ns of simulation time. In general, increasing $T_m^{es}$ to arbitrarily high values (in combination with a reasonable choice of $\tau_m^{es}$) may allow sampling of configurations having a low Boltzmann factor at the reference temperature $T_0$, leading either to slow convergence of the reference ensemble, or to a bias of the latter (in case convergence is not reached). As already mentioned above, for very high excitation temperatures, care has to be taken regarding (partial) unfolding events. In such cases the available configurational space drastically increases, rendering it improbable to return to folded states.

| Method | attempt $\nu_{ex}$ | acceptance prob. | actual $\nu'_{ex}$ |
|---|---|---|---|
| REX | 1 ps | 20 % | 5 ps |
| TEE-REX | 160 ps | 97 % | ∼160 ps |

*Table 6.2:* Typical REX and TEE-REX exchange frequencies.

The time between two exchange events should be chosen long enough to (1) allow equilibration of the reference replica after each exchange, and (2) to enable the system to adequately sample this new part of configuration space. Values of $\sim 100\,\mathrm{ps}$ were chosen because autocorrelation functions of different structural and energetical properties, such as velocity, RMSD and the short-range contribution to the Coulomb energy

(cp. chapter 6.1), showed correlation times on the order of several picoseconds. The exchange frequencies thus employed are one to two orders of magnitude lower than those routinely used in REX simulations. In TEE-REX simulations the exchange *attempt* frequency $\nu_{ex}$ is almost identical to the *actual* exchange frequency $\nu'_{ex}$ due to high acceptance probabilities of $> 95\,\%$. Values for the latter in standard REX simulations usually lie between 10-50 %, resulting in actual exchange frequencies one order of magnitude higher than in TEE-REX (see Table 6.2 for an example).

In this study the size of the essential subspace, $N_{es}$, was always chosen such that around 90 % of the total mean square fluctuations of the respective atoms was included. In general, the composition (i.e. number and type of modes) are chosen according to the specific problem under study.

## 6.6 Conclusions

The applicability of standard REX to all-atom simulations of biomolecular systems using explicit solvent becomes computationally prohibitive for systems comprising more than a few thousand atoms. Due to the large number of degrees of freedom involved, numerous replicas are needed to span a given temperature range. To overcome this inherent limitation we developed a new algorithm combining the replica exchange framework with the idea of essential dynamics. In each TEE-REX replica only a selection of essential collective modes of a subsystem of interest is excited, with the rest of the system staying at a reference temperature. The collective modes are taken from a PCA of a subsystem of interest. This selective excitation of functional relevant motions within the replica framework overcomes the computational limitations inherent to replica exchange and at the same time efficiently samples the configurational space of the system.

For a dialanine test system TEE-REX ensembles agree favorably with converged reference MD ensembles, making TEE-REX an efficient method for the study of thermodynamic properties of biomolecular systems.

The algorithm can easily be applied to larger systems. Because only a small fraction $N_{es} \ll N_{df}$ of the degrees of freedom of the system are excited in each TEE-REX replica, the exchange probability $P(S \rightarrow S')$ is no longer dominated by the solvent contribution to the potential energy. This drastically cuts down computational demands with respect to conventional REX, enabling TEE-REX to address problems currently not readily accessible to MD or other ensemble-preserving methods.

The superior sampling performance of TEE-REX with respect to MD was demonstrated using guanylin as a test system. Here, the degree to which modes are defined, i.e. the amount of *a priori* structural information, was shown to have only a minor influence on sampling performance.

The choice of the essential subspace degrees of freedom prior to any TEE-REX simulation renders the method very suitable to address questions related to structural and dynamical properties of biomolecular systems.

# 7 Simulating Large Conformational Transitions – Application to Adenylate Kinase

*Life is pleasant. Death is peaceful. It's the transition that's troublesome.*
— Isaac Asimov

## 7.1 Introduction

Experimentally determined structures of protein conformations have become increasingly available over the last years. Such structures, which are commonly determined by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy offer atomic resolution but only provide static pictures of different conformational states of proteins. Often, one is interested in the transitions between such conformations as they frequently form the basis for protein function. Examples are motor proteins such as myosin and kinesin, the bacterial flagella, ATP synthase [1, 12, 158], the chaperonin GroEL [159] or nuclear transport proteins [160]. However, questions related to the underlying transition pathway often remain open [161]. Despite recent advances in time-resolved X-ray crystallography [13, 14], it remains a challenge to elucidate the pathways and mechanisms of protein conformational dynamics.

Among the family of nucleoside triphosphate (NTP) kinases, *Escherichia coli* adenylate kinase (ADK) is a structurally well studied protein exhibiting large conformational motions crucial for its catalytic function [162]. ADK is a monomeric ubiquitous enzyme that plays a key role in energy maintenance within the cell, controlling cellular ATP levels by catalyzing the reaction

$$\mathrm{Mg^{2+}{:}\,ATP} + \mathrm{AMP} \rightleftharpoons \mathrm{Mg^{2+}{:}\,ADP} + \mathrm{ADP}\,.$$

Structurally, the enzyme consists of three domains (Fig. 7.1): the large central "CORE" domain (blue), an AMP binding domain referred to as "AMPbd" (red), and a lid-shaped ATP-binding domain termed "LID" (green), which covers the phosphate groups at the active center [162]. In an unligated structure of ADK the LID and AMPbd adopt an open conformation, whereas they assume a closed conformation in a structure crystallized with the transition state inhibitor $Ap_5A$ [163]. Here, the ligands are contained in a highly specific environment required for catalysis. Recent $^{15}N$ nuclear magnetic resonance spin relaxation studies [164] have shown the existence of catalytic domain motions in the flexible AMPbd and LID domains on the nanosecond time scale, while the relaxation in the CORE domain is on the picosecond time scale [165, 166].

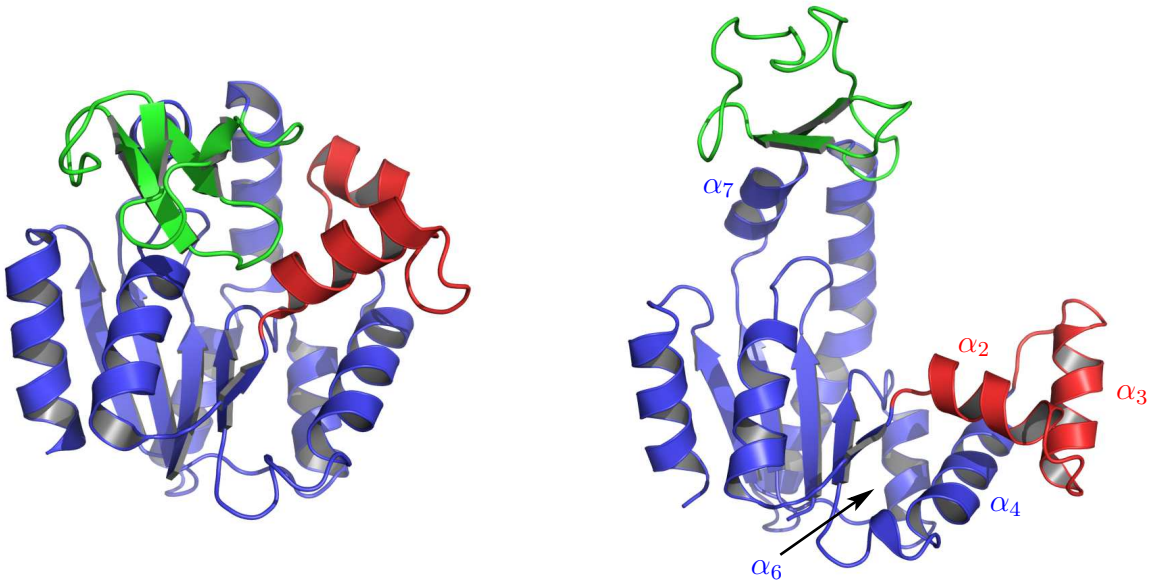


*Figure 7.1:* Closed (left) and open (right) crystal structures of *E. coli* adenylate kinase (ADK) having domains CORE (blue), AMPbd (red) and LID (green). The transition state inhibitor $Ap_5A$ is removed in the closed crystal structure (left). Assignment of secondary structure elements according to [163].

For ADK, several computational studies have addressed its large conformational flexibility [167, 168, 169]. Maragakis and Karplus [170] used a coarse-grained plastic network model to generate a minimum energy path between the open and closed crystal structures. Their study predicts that the motion of the LID region precedes the motion of the AMPbd when going from the open to the closed form. In a very recent study, essential dynamics simulations have been performed to determine whether the closed conformation is accessible to the open unligated enzyme [171]. The study suggests

that the open conformation is at a slightly lower free energy than the closed conformation, as a consequence of hinge motions responsible for domain closure. Although considerable progress has been made towards the understanding of these large conformational changes in ADK, a detailed picture of the open to closed transition at full atomic resolution has not yet been achieved.

Here, we employed TEE-REX to facilitate the spontaneous transition between the open and closed structures of ADK, achieving a fully atomistic description of the transition pathway and its underlying mechanics.

## 7.2 Simulation Details

All simulations were carried out using the MD software package GROMACS 3.3.1 [80], supplemented by the TEE-REX module [97]. Eight TEE-REX simulations were carried out, four starting from the equilibrated open and four form the closed crystal structures. Additionally, two reference MD simulations were started from the closed and open conformation. The OPLS-all atom force field [84] was used for the protein and TIP4P was used as a water model [153]. All simulations were performed in the *NPT* ensemble. The pressure was coupled to a Berendsen barostat [94] with $\tau_p = 1.0\,\text{ps}$ and an isotropic compressibility of $4.5 \cdot 10^{-5}\,\text{bar}^{-1}$. All bonds were constrained using the LINCS algorithm [31]. An integration time step of 2 fs was used. Lennard-Jones interactions were calculated explicitly with a 10 Å cutoff. Coulombic interactions were calculated explicitly at a distance smaller than 10 Å; above 10 Å, long-range electrostatic interactions were calculated with the PME method [88], using a direct space cutoff of 10 Å and a reciprocal grid spacing of 0.12 nm and fourth order B-spline interpolation.

### MD Simulations

Two reference simulations one of 92 ns and one of 109 ns length were started from the closed (MDc) and open (MDo) conformation, respectively. In all MD simulations the temperature was kept constant at $T = 300\,\text{K}$ by coupling to an isotropic Berendsen thermostat [94] with a coupling time of $\tau_t = 0.01\,\text{ps}$ for the MDc and $\tau_t = 0.1\,\text{ps}$ for the MDo simulation.

The MDc simulation system was set up as follows. From the protonated crystal structure (PDB entry 1AKE) [163] the two-substrate-mimicking inhibitor $P^1$,$P^5$-bis(adenosine-5')pentaphosphate ($Ap_5A$) was removed. The protein was then solvated in a rhombic dodecahedral box with box vectors of 74.775 Å length. The system com-

prised 37 965 atoms. Four $Na^+$ ions were added to neutralize the system. The energy of the solvated system was minimized using the steepest descent algorithm. Subsequently, a 100 ps MD simulation at the target temperature was carried out using harmonic position restraints on the heavy atoms of the protein with a force constant of $k = 1000\,\mathrm{kJmol^{-1}nm^{-2}}$ to equilibrate water and ions. Next, a trajectory of 109 ns length was produced by a free (unbiased) MD simulation. Structures were recorded every 1 ps for subsequent analysis.

For the MDo simulation system the protonated crystal structure (PDB entry 4AKE) [162] was solvated in a rectangular box having a size of 63.309 Å × 83.52 Å × 77.031 Å. The system comprised 53 195 atoms. To mimic physiological conditions, 38 $Na^+$ and 35 $Cl^-$ ions were added. Energy minimization of the solvated system using the steepest descent algorithm was followed by a 500 ps MD simulation at the target temperature using harmonic position restraints on the heavy atoms of the protein with a force constant of $k = 1000\,\mathrm{kJmol^{-1}nm^{-2}}$ to equilibrate water and ions. Subsequently, a trajectory of 92 ns length was produced by a free MD simulation. Structures were recorded every 3 ps for subsequent analysis.

### TEE-REX Simulations

Eight 20 ns TEE-REX simulations were performed, starting from the equilibrated open (TRo 1-4) and closed (TRc 1-4) starting structures used in the MDo and MDc simulations, respectively. Each TEE-REX simulation consisted of three replicas, having temperatures $(T_m^{es},T_0)$ of (300 K, 300 K) and (320 K, 300 K) for the reference $m = 0$ and first excited replica $m = 1$, respectively. Excitation temperatures of (550 K, 300 K) or (650 K, 300 K) were used for the second excited replica $m = 2$ (see Table 7.1).

Three different eigenvector sets were used in the construction of the essential subspaces $\{es\}$ (Fig. 7.2). A PCA was performed on the first 5 ns of the MDo simulation, taking all backbone atoms into account. The first five eigenvectors, describing 92 % of the backbone fluctuations, defined the essential subspace $\{\mu_1,\ldots,\mu_5\}$. Repeating this procedure for the first 5 ns of the MDc simulation yields the eigenvector set $\{\nu_1,\ldots,\nu_5\}$, describing 92 % of the respective backbone fluctuations. A PCA on the combined MDo and MDc ensemble results in $\{es\} = \{\kappa_1,\ldots,\kappa_5\}$. The latter was used for simulations TRo 1-2 and TRc 1-2. Thereby, also the eigenvector directly connecting the open and the closed structure is excited (Fig. 7.2*C*). For the second set of eigenvectors, $\{es\} = \{\mu_1,\ldots,\mu_5,\nu_1,\ldots,\nu_5\}$, used in simulations TRo 3-4 and TRc 3-4, only eigenvectors describing the local fluctuations of the respective conformation are excited.

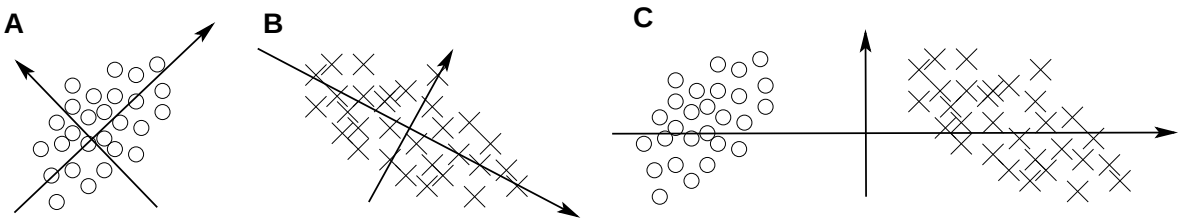


*Figure 7.2:* Schematic two-dimensional visualization of the procedure to construct the essential subspaces $\{\mu_i\}$, $\{\nu_i\}$ and $\{\kappa_i\}$ used in TEE-REX simulations of ADK, see also Table 7.1. Principal component analyses performed on MD ensembles (scatter plots $\circ$ and $\times$) are depicted as coordinate axes ($\dashv\vdash$). Essential subspaces $\{\mu_i\}$ (Panel *A*) and $\{\nu_i\}$ (Panel *B*) describe only local fluctuations around the respective conformation, whereas the eigenvector directly connecting the open and closed form is excited in $\{es\} = \{\kappa_i\}$ (Panel *C*).

The used essential subspaces were coupled to a Berendsen thermostat with a coupling time of $\tau_t = 0.05\,\text{ps}$, with all other degrees of freedom coupled with $\tau_t = 0.1\,\text{ps}$. Exchanges between replicas were attempted every 160 ps and were accepted with $> 95\,\%$ probability. Structures were recorded every 2 ps for further analysis. A summary of all parameter combinations used is given in Table 7.1.

| Simulation | Starting Structure | Time | Essential Subspace | Temperature |
|---|---|---|---|---|
| MDo | open | 109 ns | | $T = 300\,\text{K}$ |
| MDc | closed | 92 ns | | $T = 300\,\text{K}$ |
| TRo 1 | open | 20 ns | $\{\kappa_1, \ldots, \kappa_5\}$ | $\max\{T^{es}\} = 550\,\text{K}$ |
| TRo 2 | open | 20 ns | $\{\kappa_1, \ldots, \kappa_5\}$ | $\max\{T^{es}\} = 650\,\text{K}$ |
| TRo 3 | open | 20 ns | $\{\mu_1, \ldots, \mu_5, \nu_1, \ldots, \nu_5\}$ | $\max\{T^{es}\} = 550\,\text{K}$ |
| TRo 4 | open | 20 ns | $\{\mu_1, \ldots, \mu_5, \nu_1, \ldots, \nu_5\}$ | $\max\{T^{es}\} = 650\,\text{K}$ |
| TRc 1 | closed, no ligand | 20 ns | $\{\kappa_1, \ldots, \kappa_5\}$ | $\max\{T^{es}\} = 550\,\text{K}$ |
| TRc 2 | closed, no ligand | 20 ns | $\{\kappa_1, \ldots, \kappa_5\}$ | $\max\{T^{es}\} = 650\,\text{K}$ |
| TRc 3 | closed, no ligand | 20 ns | $\{\mu_1, \ldots, \mu_5, \nu_1, \ldots, \nu_5\}$ | $\max\{T^{es}\} = 550\,\text{K}$ |
| TRc 4 | closed, no ligand | 20 ns | $\{\mu_1, \ldots, \mu_5, \nu_1, \ldots, \nu_5\}$ | $\max\{T^{es}\} = 650\,\text{K}$ |

*Table 7.1:* TEE-REX and MD simulation details. Shown simulation parameters are: simulation length (ns), used essential subspace $\{es\}$, maximum $\{es\}$ excitation temperature (K).

## 7.3 Results

### 7.3.1 Conformational Transition

To analyze whether any spontaneous transition between the open and closed conformation was observed, simulations were analyzed in terms of their root mean square deviation (RMSD) of all backbone atoms with respect to the experimental structure that was not the starting structure of the simulation (i.e. the closed structure for simulations starting from the open X-ray structure and vice versa). A threshold of 0.3 nm, derived from averaging equilibrium fluctuations relative to the respective starting structures (Fig. 7.3*D*), was used to identify transitions. Panels *A*-*C* of Fig. 7.3 show the RMSD as a function of simulation time for all TEE-REX and MD simulations, with the transition threshold of 0.3 nm indicated by a dashed line.

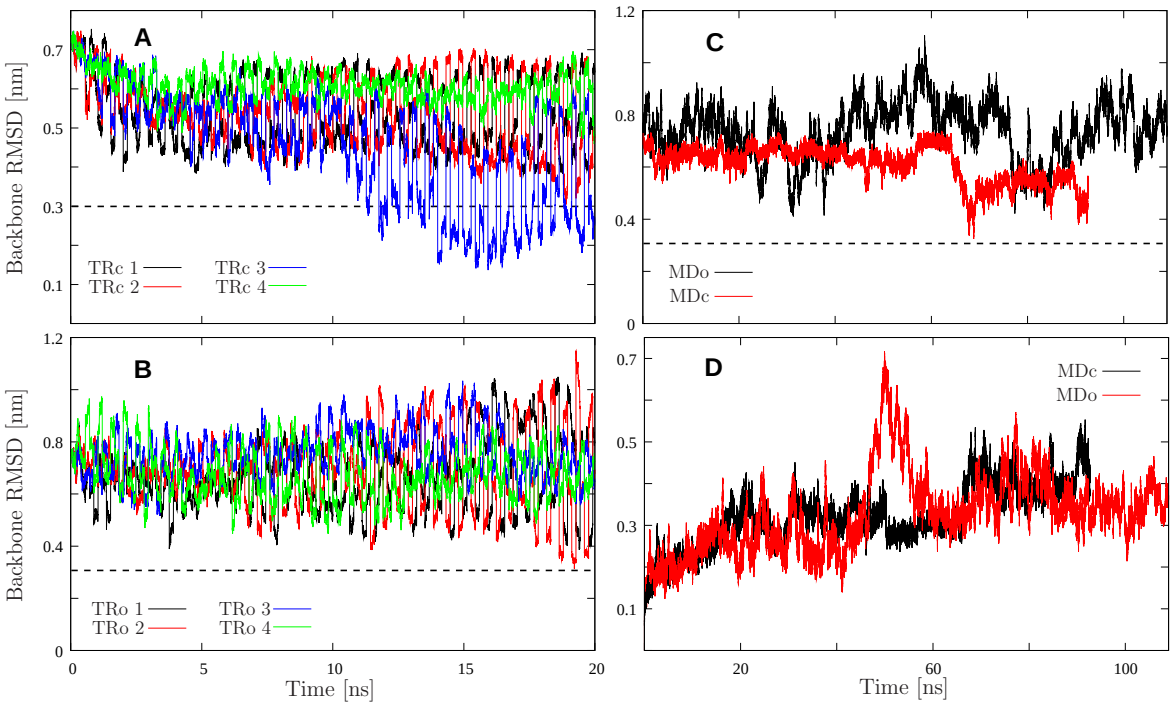


*Figure 7.3:* Backbone RMSD versus time for all simulations reported. Panel *A*: RMSD of TEE-REX simulations started from the closed conformation with regard to the open structure: TRc 1 (black), TRc 2 (red), TRc 3 (blue), TRc 4 (green). Panel *B*: RMSD of TEE-REX simulations started from the open form relative to the closed conformation: TRo 1 (black), TRo 2 (red), TRo 3 (blue), TRo 4 (green). Panel *C*: RMSD of MD simulations MDo (black) and MDc (red), started from the open and closed structure, with respect to the closed and open conformation, respectively. Panel *D*: Backbone RMSD of MD simulations MDc (black) and MDo (red) relative to their starting structure.

A complete transition event is clearly observed for simulation TRc 3 (blue line in Panel *A*), as evident from the fact that this simulation repeatedly approaches the open conformation as close as 0.14 nm in terms of RMSD. Interestingly, the modes excited in the TRc 3 simulation did not include the difference X-ray mode connecting the open and closed experimental structures (Fig. 7.2). The TRc 2 simulation also seems to briefly sample the open conformation. However, strictly this is not the case (chapter 7.3.3). In all four TRc simulations, a constant drop in RMSD by 0.15 nm during the first three nanoseconds of simulation time is observed, putatively indicating a relaxation from internal strain energy. For TEE-REX simulations TRo 1-4 (Fig. 7.3*B*), starting from the open conformation, only a slight excursion towards the closed conformation is observed in the TRo 2 simulation during the 20 ns of simulation time. In comparison, none of the MD simulations shows a complete transition between the two conformations. The MDc simulation approaches the open conformation up to 0.33 nm RMSD briefly during the 92 ns of simulation time, in accordance with measured rates of ∼52 ns for ADK domain motions [164].

Throughout all MD and TEE-REX simulations a preference for the open conformation is observed, consistent with a suggested lower free energy for the open conformation [171]. However, the inability of several TRc and TRo simulations to reach the open and closed conformation, respectively, indicates a free energy barrier additional to the suggested monotonic profile of Snow *et al.* [171].

### 7.3.2 Pathway Characterization

The TRc 3 ensemble was used to analyze the conformational transition in more detail. In contrast to conventional MD, dynamical information on the transition pathway is lost in TEE-REX due to frequent exchanges between all replicas. However, a pathway can easily be constructed from the TRc 3 reference ensemble using a RMSD distance measure described below.

#### Construction of the Transition Pathway

The construction of a pathway between the open and closed conformation of ADK for the TRc 3 simulation is based on the following idea: in going from point $A$ to some distant point $B$ one increases the distance from $A$ while at the same time approaching the target point $B$ (thereby neglecting detour routes). The backbone RMSD curves with respect to the open (RMSDo) and closed (RMSDc) structures are used as the

distance measure. In doing so, a continuous pathway is defined between the open and the closed state, commensurate with the simulated ensemble. In a first step, RMSD differences RMSDo − RMSDc were calculated for each frame and sorted in decreasing order. A large distance in the RMSD space thereby corresponds to a structure close to the beginning of the path, whereas a low RMSD distance points to a structure in the vicinity of the target conformation. Representative structures for a given number of evenly spaced waypoints in the constructed RMSD distance space were chosen to visualize the pathway (Fig. 7.4).

To probe for possible detour routes neglected in the algorithm described above, the TRc 3 ensemble was projected into the plane spanned by the two RMSD coordinates given in Fig. 7.3*A* and *B*, i.e., the RMSD with respect to the open and closed structure, respectively. A narrow distribution along the diagonal connecting both end states of the pathway is found, implying only limited variability for ADK along the transition path.

### Transition Pathway

Figure 7.4 shows a complete pathway (yellow line) overlaying the TRc 3 ensemble (black) as well as the MDo (red) and MDc (cyan) ensembles. Two PCA analyses were carried out on the TRc 3 ensemble to define the $x$ and $y$ coordinate. The $x$-axis is given by the first eigenvector describing the AMPbd motion with respect to the CORE. Similarly, the first eigenvector depicting the LID motion with respect to the CORE defines the $y$-axis. The crystal structure of the open conformation (PDB code 4AKE) is indicated by a green square, together with four cartoon representations visualizing different structures along the pathway (insets *A*-*D*, magenta triangles). The closed crystal structure (PDB code 1AKE, ligand removed) is shown in inset *D*, corresponding to the lower right triangle at one end of the pathway. Secondary structure assignments were taken from Müller and Schulz [163].

A secondary structure analysis of the structures along the pathway was performed using DSSP [172] to check for structural stability during the transition. For the CORE (residues 1-29, 60-121, 160-214) and AMPbd (residues 30-59), no significant change in secondary structure is seen despite the large conformational change of the latter. As for the LID (residues 122-159), both $\beta$-sheets are stable, with only small conversions among residues constituting bends and turns of the domain. Overall, ADK strongly maintains its integrity, showing only minute changes in secondary structure. Thus, the system essentially behaves like a rigid body with flexible domains.

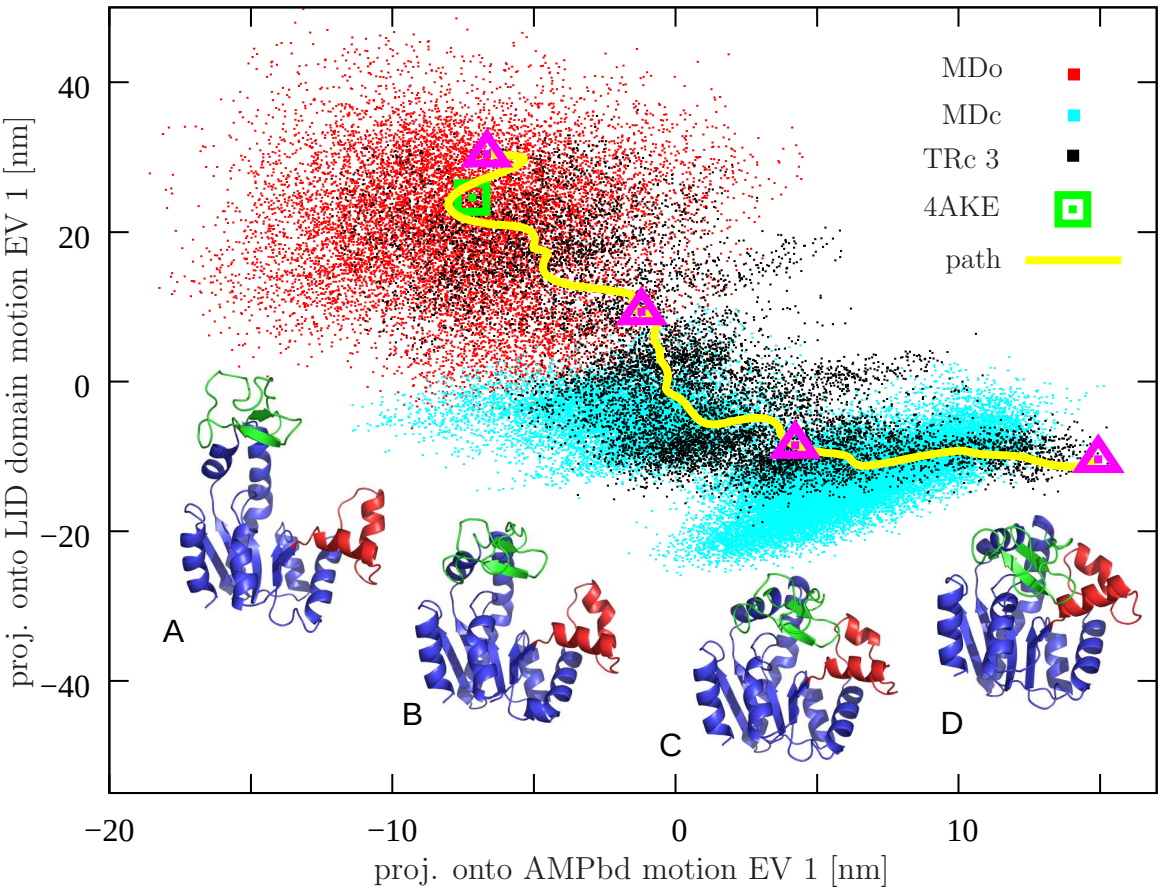


*Figure 7.4:* Two-dimensional projection of the complete transition path (yellow line), overlaying the TRc 3 (black dots), MDo (red dots) and MDc (cyan dots) ensembles, as well as the crystal structure of the open (green square) and closed conformation (inset *D*, rightmost magenta triangle). Colored cartoon representations *A*-*D* visualize different structures along the pathway (magenta triangles). Color coding follows Fig. 7.1.

Unligated ADK can sample a wide range of conformations between the open and closed structure, offering ligands several favorable structures for binding. This result is indicative of the *conformational selection* view of ligand binding proposed for ADK [173, 174, 175, 176]. The sampled regions of configurational space suggest a preferential transition pathway of the unligated enzyme. The complete transition from the unligated closed (structure *D*) to the open conformation (structure *A*) can be characterized by two phases. During the first phase (insets *C* and *D*), the LID remains essentially closed while the AMPbd, comprising helices $\alpha_2$ and $\alpha_3$ (Fig. 7.1), assumes a half-open conformation (structure *C*). In doing so, $\alpha_2$ bends towards helix $\alpha_4$ of the CORE by 15° with respect to $\alpha_3$ (Fig. 7.5*A*). This opening of the AMP binding cleft could facilitate

an efficient release of the formed product. For the second phase, a partially correlated opening of the LID domain together with the AMPbd is observed. Halfway towards the open conformation (structure $A$), an intermediate half-open structure (transition state $B$) at the contact interface of both MDc and MDo ensembles is visited. During ∼100 ns of simulation time, both MD simulations were unable to pass beyond this interface in either direction, suggesting a substantial free energy barrier along the pathway.

A domain motion analysis using DynDom [177] of the second phase of the transition shows that the LID opening cannot be described by a pure hinge-bending motion. Only in the last part of phase two (going from $B$ to $A$) the LID motion follows a pure hinge-bending motion of approximately 30°, with the hinge axis given by residues L115, I116 and R167, L168. At the beginning of phase two, a combination of a hinge-bending motion and an outward translation away from the AMPbd characterizes the pathway towards transition state $B$.

For phase one, an opening motion of the AMPbd is found, with bending residues S30, T31 and K69-R71. An interesting observation in this respect is the formation of a highly stable salt bridge D118-K136, connecting the LID and CORE domains (Fig. 7.5*B*). Estimating the total non-bonded interaction between LID and CORE, it was found that this salt bridge contributes substantially to the total interaction energy between the two domains. From a comparison of thirteen[1] PDB structures from yeast, maize, human and bacterial adenylate kinase, ten structures feature such a salt bridge motif at the LID-CORE interface.

### 7.3.3 Alternative Pathways

The observed transition pathway of the TRc 3 ensemble corresponds to a particular sequence of events: starting from the closed conformation, a subsequent half-opening of the AMPbd is followed by a partially correlated opening of the LID/AMPbd complex. It is an interesting question whether this pathway is the only possibility or whether alternative pathways are also possible.

An investigation of all TEE-REX simulations, conducted within the subspace shown in Fig. 7.4, reveals that simulations TRc 1 and TRc 2 sample the first part of an alternative transition pathway: starting from the closed state, the LID opens independently from the AMPbd, with the latter remaining between the closed and a half-open conformation. Inside the depicted PCA subspace of Fig. 7.4, this motion corresponds to

[1] PDB structures examined (* contain salt bridge motif): 4AKE*, 1AK2*, 1AKY, 1P3J*, 1S3G, 1ZAK*, 1ZD8, 1ZIO*, 1ZIP*, 2AK3*, 2AR7*, 2C9Y*, 2ECK*

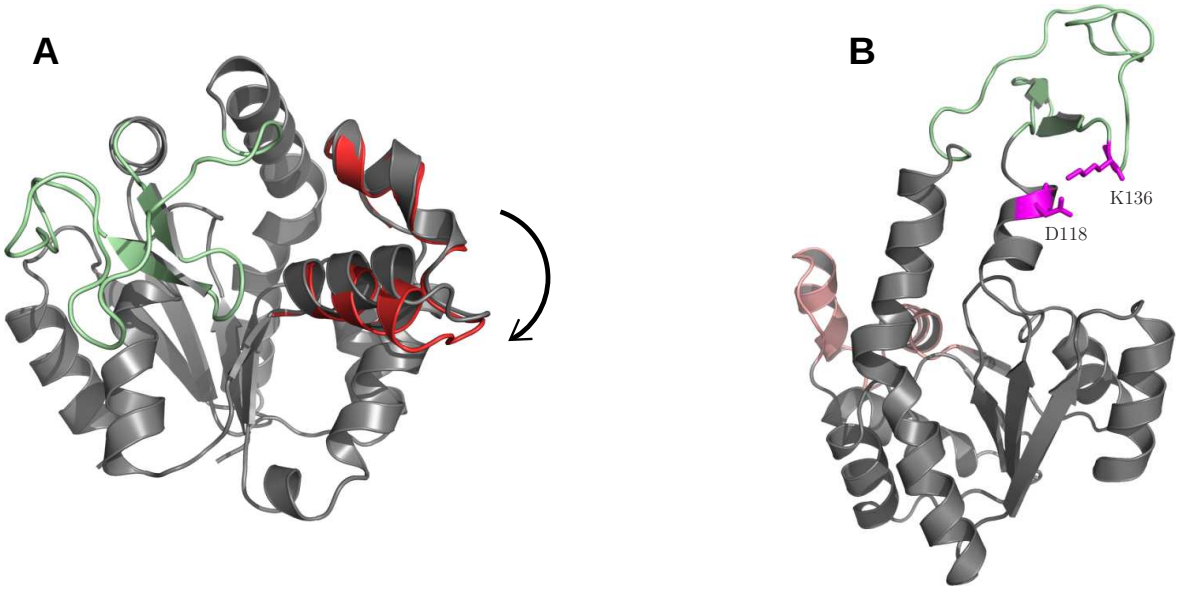


*Figure 7.5:* Prominent features observed along the ADK transition path. Panel *A*: Opening of the AMPbd cleft (grey→red) during phase one. Panel *B*: Rear view of the highly stable salt bridge motif D118-K136 (magenta), connecting the LID (green) and CORE (grey) domains.

a basically vertical path parallel to the $y$-axis, with the starting point being the closed conformation. For TEE-REX simulation TRc 2, the RMSD analysis (Fig. 7.3) suggests a full transition event. However, despite the RMSD of only 0.3 nm, the final conformation deviates considerably from the open crystal structure within the structurally resolved PCA subspace (Fig. 7.6), mainly along the $x$-axis describing the AMPbd motion. To complete a transition along this pathway, the AMPbd would be required to fully open up. Yet, both the TRc 1 and TRc 2 simulations fail to complete this last step, possibly indicating a substantial free energy barrier in this part of configuration space.

A comparison of the final structures from the TRc 2 and TRc 3 simulations shows that conformational deviations are located in helix $\alpha_3$ of the AMPbd, the LID, helix $\alpha_5$ of the CORE domain as well as helices $\alpha_6$ and $\alpha_7$ connecting the LID and CORE domain. To investigate structural features in more detail, the pathway obtained from the TRc 2 ensemble was compared to the transition pathway of the TRc 3 ensemble (Fig. 7.6). After initial relaxation (Fig. 7.3*A*), the TRc 2 pathway diverges from the TRc 3 case. In the former, the LID domain fully opens with the AMPbd remaining closed. Here, the LID domain motion occurs independently from the AMPbd, in contrast to the observed transition in TRc 3, where opening of the AMPbd precedes LID movement. While the LID assumes a half-open conformation in TRc 3, the AMPbd gains flexibility

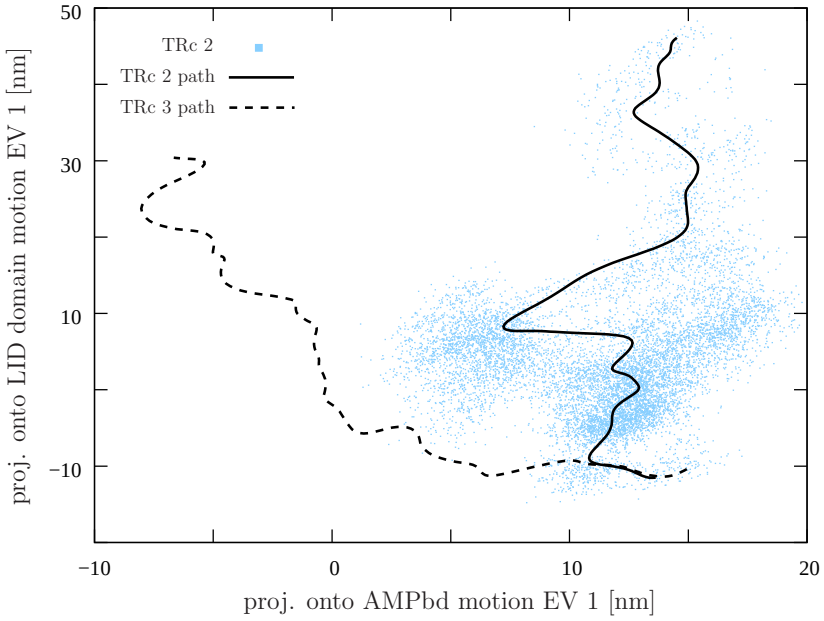


*Figure 7.6:* Two-dimensional projection of the complete TRc 3 transition path (dashed line) and the TRc 2 pathway (solid line), overlaying the TRc 2 ensemble (grey dots).

via bending of helix $\alpha_2$ towards $\alpha_4$ of the CORE, resembling the motion encountered in the transition pathway. However, a full opening of the AMPbd is prevented by helix $\alpha_3$, keeping its relative position to helix $\alpha_7$ throughout the TRc 2 pathway. LID flexibility is comparable for both pathways, although the end configurations assume a slightly different aperture angle with regard to the CORE domain. A secondary structure assignment of the TRc 2 pathway using DSSP shows larger fluctuations compared to the TRc 3 transition, the differences concentrating in the CORE and AMPbd. For the latter, residues of helices $\alpha_2$ and $\alpha_3$ adjacent to the helix-connecting loop exhibit the largest structural variability.

## 7.4 Discussion

With the newly developed TEE-REX algorithm, the spontaneous domain conformational transition of *E. coli* adenylate kinase was simulated for the first time at full atomic resolution. In contrast to coarse-grained models [167, 169, 170] originally designed for such tasks, TEE-REX combines the advantage of atomic detail with a highly efficient and approximately ensemble-preserving algorithm.

From a series of eight TEE-REX simulations, complemented by two MD references,

a possible transition pathway was found. A truly spontaneous transition was induced, since in the TRc 3 simulation, showing a full transition event, the temperature-enhanced essential subspace did not contain the transition mode connecting the open and closed conformations (see Fig. 7.2 and Table 7.1). The pathway could be characterized by two phases. Starting from the closed conformation, a half-opening of the AMPbd is followed by a partially correlated opening motion of the LID/AMPbd complex towards the open state. This sequence of events exceeds findings of a study by Maragakis and Karplus [170]. From a minimum free energy path, retrieved from a coarse-grained model of ADK, Maragakis and Karplus found that the LID motion occurs independently from AMPbd motion.

Along the transition pathway we identified two prominent features (Fig. 7.5). First, during phase one, opening of the AMPbd domain occurs via bending of the $\alpha_2$ helix towards $\alpha_4$ of the CORE domain by approximately 15° with respect to helix $\alpha_3$. This opening of the AMP binding cleft might be involved in facilitating an efficient release of the formed product after catalysis. However, since all simulations were carried out in the absence of any ligand, no conclusions can be drawn with respect to ligand behavior. Second, a stable salt bridge, D118-K136, connecting the LID and CORE domains forms that strongly contributes to the total enthalpic interaction between both domains, suggesting a stabilizing function for the open conformation. The occurrence of such a salt bridge motif in several adenylate kinase structures of different species supports the hypothesis. Breaking this salt bridge via mutation, e.g. D118A, should thus be expected to decrease the stability of the open state.

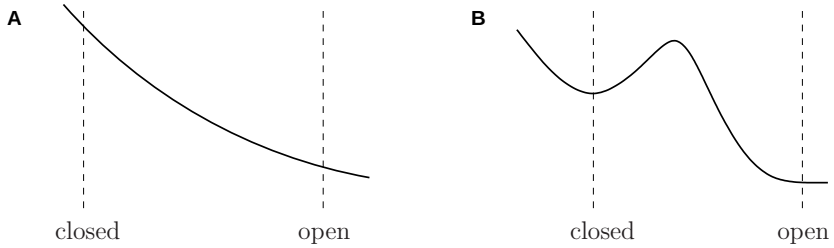


*Figure 7.7:* Schematic representation of suggested free energy profiles for unligated ADK. Monotonic profile by Snow *et al.* [171] (Panel *A*); suggested profile by the author (Panel *B*).

From our simulation data, a qualitative picture of the underlying free energy landscape of unligated ADK can be estimated (Fig. 7.7). All TEE-REX and MD simulations, starting from the closed crystal structure, show a preference for the open conformation, whereas no such preference for the closed state is seen for simulations starting from the

open state. This finding is consistent with a declining free energy profile deduced from simulations that induced the transition from the closed to the open conformation [171] (Fig. 7.7*A*). However, the inability of both MD and several TRc and TRo simulations to reach the open and closed structure, respectively, indicates a free energy barrier in addition to the suggested monotonic profile (Fig. 7.7*B*). Thus, the following picture emerges supporting the *conformational selection* view of ligand binding: in equilibrium, unligated ADK can sample both—open and closed—conformations, as observed in the spontaneous transition of the TRc 3 simulation from the closed to the open structure. Here, the closed state has a higher free energy with regard to the open state. Upon ligand binding the closed structure is stabilized by protein-ligand interactions for catalysis. From the behavior of both MD and TEE-REX simulations, the coarse location in configuration space of the additional free energy barrier can be estimated, corresponding to a half-open conformation of ADK.

Transition pathways other than the one characterized seem possible, as an analysis of all TEE-REX simulations suggests. Although a complete transition was not observed, an independent opening of the LID domain with respect to the AMPbd was found. In both pathways the characteristic half-opening of the AMPbd and the stable salt bridge motif are present, underlining their significance for the atomistic mechanics of the transition. Full opening of the AMPbd would complete the alternative route. However, this was not observed, possibly indicating an additional free energy barrier in this part of configuration space. Together with the observed larger fluctuations in secondary structure elements, indicating high internal strain energies, the enthalpic penalty along this route possibly renders it unfavorable as a transition pathway of ADK. However, the fact that no full transition events were observed along this pathway could also be due to limited sampling in our simulations. Therefore, it can presently not be ruled out that transitions also occur along this pathway.

# 8 Summary and Conclusions

*It's hard to be nostalgic when you can't remember anything.*
— Unknown

A major goal of protein science is to explore the coupling of protein motion and function. Whereas the underlying molecular basis in terms of protein dynamics is often not directly accessible by experiments, molecular dynamics (MD) simulations have shown to be a valuable microscopic complement. In this respect, ongoing efforts are needed in the development of algorithms aimed at an enhanced and efficient sampling of the conformational space of proteins.

In this thesis, the new temperature enhanced essential dynamics replica exchange (TEE-REX) method is developed, combining the ideas of essential dynamics (ED) with the temperature replica exchange (REX) formalism.

Enhanced sampling in REX MD is achieved by simulating in parallel a number of copies (replicas) of the system having different temperatures. Large free energy barriers in conformational space are overcome in low-temperature replicas via exchange with high-temperature copies, thereby utilizing the larger mobility of the latter. As a crucial factor for sampling performance, the exchange probability scales inversely exponential with the temperature difference among adjacent replicas and the excited number of degrees of freedom per copy. Consequently, computational demands (i.e. number of replicas) drastically limit REX performance when already applied to medium sized (few thousand particles) systems simulated in full atomic detail.

To improve REX, a reduction in the number of degrees of freedom excited per replica is thus a promising means to overcome this computational bottleneck. In the ED framework collective coordinates, describing functional modes of motion, are excited to yield an enhanced sampling of conformational space. These collective modes of motion are obtained from a principal component analysis (PCA) of the covariance matrix of atomic fluctuations. In the ED protocol the system is constantly driven along selected collective modes, irrespective of the topology of the underlying free energy landscape. Although the sampling is enhanced by this driving process, sampled structures are distributed

evenly over conformational space which results in a wrong statistical ensemble. Despite this lack of thermodynamical accuracy of ED, its usefulness originates from the fact that around 5-10 % of the first collective PCA modes describe all observed fluctuations by more than 90 %.

With the newly developed TEE-REX method, the REX and ED approaches are brought together in a consistent framework. In contrast to standard REX only a few predefined—collective—degrees of freedom, called essential subspace $\{es\}$, are excited in each but the reference replica, with the remaining degrees of freedom kept at the reference temperature throughout all copies. A substantial reduction in computational effort (brought about by the larger temperature steps in TEE-REX) is thus combined with the specific excitation of important modes of motion. Additionally, the REX framework ensures an approximate correct statistical weighting of each structure within the generated TEE-REX reference ensemble.

To assess algorithmic performance, statistical properties of TEE-REX and MD reference ensembles were investigated, as well as the sampling efficiency with respect to MD. Statistical properties were directly probed by calculating a thermodynamic potential, namely the relative Gibbs free energy (Eq. (6.1), page 36) for a dialanine peptide using extended multi-microsecond MD and sub-microsecond TEE-REX simulations. Since conclusive statements can only be given on the basis of converged ensembles, different convergence measures were applied to the MD reference ensemble, validating convergence of the latter. A comparison of free energy landscapes showed favorable agreement within the respective statistical errors of $\sigma_{\text{TEE-REX}} \approx \sigma_{\text{MD}} \leq 0.15\, k_B\text{T}$ at T= 300 K. The deviations from a correct statistical ensemble introduced by exchange of non-Boltzmann structures into the TEE-REX reference replica are thereby largest for small systems such as dialanine, due to the large fraction of excited $\{es\}$ degrees of freedom. Hence, the observed statistical differences of a TEE-REX generated ensemble constitute an upper bound.

TEE-REX and MD simulations of a guanylin test system were performed to compare the sampling efficiency of both methods. A direct comparison to REX simulations was discarded because the computational effort involved with REX considerably exceeded that by TEE-REX while only yielding a slight increase (few %) in sampling efficiency over MD. As a measure for efficiency, projections of the sampled $3N$-dimensional configuration space of the system onto different 2-dimensional subspaces were calculated as a function of computational effort. Within the essential subspace $\{es\}$, TEE-REX outperforms MD on average by more than a factor of three. Sampling efficiency in sub-

spaces independent from $\{es\}$ is slightly lower, but still an 2.5-fold gain in TEE-REX sampling efficiency over classical MD is seen.

In contrast to other simulation protocols based on REX, the TEE-REX algorithm can easily be applied to larger systems. Because only a small fraction $N_{es} \ll N_{df}$ of the degrees of freedom of the system are excited in each TEE-REX replica, the exchange probability $P(S \rightarrow S')$ is no longer dominated by the solvent contribution to the potential energy. This drastically cuts down computational demands (three replicas suffice regardless of system size) with respect to conventional REX, enabling TEE-REX to address problems currently not readily accessible to MD or other ensemble-preserving methods.

Information about different protein configurations is a necessary prerequisite for the construction of the essential subspace $\{es\}$. Experimental or theoretical limitations can severely restrict the available information. Using the guanylin test system, the effect on the sampling efficiency of TEE-REX was investigated using only 2 % of the structural information available (and originally applied) for the construction of the essential subspace. Although both subspaces differed markedly, only minor differences in the sampled configuration space between the structurally well-defined $\{es\}$ and the poorly defined $\{es\}'$ were found. These results indicate that TEE-REX sampling efficiency is hardly sensitive on the *a priori* available structural information.

In a first application, the sampling power of TEE-REX was used to study adenylate kinase (ADK), an experimentally well studied monomeric enzyme playing a key role in energy maintenance within the cell. ADK exhibits very large conformational motions crucial for its catalytic function of phosphorylation and de-phosphorylation of ADP. Despite considerable computational effort, a transition pathway between the two crystallographically known end states of the catalytic cycle at atomic resolution was still missing. Using different essential subspaces derived from MD simulations of the two crystal structures, a possible transition pathway was characterized for the first time using TEE-REX. In addition to the identification of experimentally verifiable structural features, a qualitative picture of the underlying free energy landscape was proposed.

## Outlook

Numerous applications of the TEE-REX algorithm to questions concerning structural, dynamical and thermodynamical properties of biomolecules can profit from the advantages offered by the method. The all-atom description combined with an highly efficient and statistically accurate sampling provides a strong footing, e.g. for the calculation of

relative free energy differences between different protein conformers or to study conformational transitions at atomic resolution. Unlike other REX-based schemes, systems of arbitrary size can thereby be treated with little additional computational effort. This opens up the possibility to investigate e.g. protein docking or allostery of large protein complexes. Here, the enhanced sampling allows for the investigation of small conformational changes, which require a large signal-to-noise ratio in terms of configurational fluctuations.

The specific excitation introduced by the essential subspace degrees of freedom add a flexible element to the algorithm. As demonstrated for adenylate kinase, the exploration of unknown transition pathways is one application. Although the pre-defined $\{es\}$ modes are fixed throughout a TEE-REX simulation, new regions of conformational space are explored by the unbiased reference replica. Thus, a PCA on this reference ensemble results in new $\{es\}'$ modes which can be fed back into the algorithm, yielding a further exploration of conformational space.

The temperature excitation of the $\{es\}$ is an important ingredient for the enhancement in sampling. However, adjusting the strength of the excitation depends on the studied system and the choice of the $\{es\}$ modes. Leaving aside folding/unfolding studies, an appropriate combination of essential subspace temperatures $T_m^{es}$ and coupling constant $\tau_m^{es}$ needs to be chosen. Depending on the experience of the user, various trial simulations are necessary to determine suitable simulation parameters. To avoid this additional computational effort, an adaptive $T_m^{es}$ temperature control scheme is proposed, based upon on-the-fly calculations of the diffusion constant of the system within the essential subspace. Preliminary work by the author indicates that such a scheme indeed allows self-regulation of the essential subspace temperature. So far, this approach is not entirely free of parameters since a cut-off for anomalous diffusive behavior, indicating unfolding, must be specified. However, the trial phase is replaced by setting up various TEE-REX simulations with different cut-off parameters, of which most will actively contribute to the sampling of conformational space.

# A How to Set Up a TEE-REX Simulation

*Never start a calculation before you know the answer.*

— John Archibald Wheeler

Here we describe the *general* protocol for setting up a TEE-REX simulation using the GROMACS simulation package. Implementation details specific to a certain version of the software can be found in the corresponding `README` file. As an example, a single protein (guanylin) solvated in water and ions is considered.

## A.1 Construction of $\{es\}$

Before a TEE-REX simulation can be started, the essential subspace modes $\{es\}$ have to be constructed from structural information. Here we use PCA for this task so an ensemble of structures is necessary, either experimental (X-ray, NMR) or theoretical (MD/REX simulations, CONCOORD, homology modeling, ...). In a first step, the group of index atoms has to be chosen (`make_ndx`) for which PCA modes are calculated. In general, *any* subset of protein atoms can be used. For monomeric proteins, backbone atoms are routinely used since most of the conformational flexibility of a protein is determined by the backbone. We adopt this choice in our example. In multimeric proteins (e.g. hemoglobin) or systems containing several proteins and/or ligands, the subset of index atoms can belong to one (e.g. ligand, binding pocket) or to all constituents (e.g. all backbone atoms of a tetramer). Once the PCA is performed on the chosen index group (`g_covar`), the desired essential subspace is built using the `make_edi` tool. The $\{es\}$ modes are stored in the `sam.edi` file.

## A.2 Recalculation of Degrees of Freedom

After the construction of the $\{es\}$, the simulation system is partitioned into at least three groups used for temperature coupling: index atoms, remaining atoms of the protein, and solvent. If no constraint algorithm for covalent bonds of the protein is used, the actual TEE-REX simulation can be started. In case bonds are constrained (e.g. using the LINCS algorithm), a recalculation of the number of degrees of freedom (DOF) for the two protein temperature coupling groups is necessary due to the partitioning in index and remaining atoms.

When different temperature coupling groups are used, the `grompp` preprocessor estimates the number of DOF of the respective group $i$ to ensure a correct calculation of the group temperature via the equipartition theorem $2K_i = N_i k_B T_i$. However, this estimate only works if atoms of different groups do not share covalent bonds, i.e. are not connected to each other. Ordinary MD setups use a protein-solvent topology; in our example, the protein is partitioned into `Backbone` and `Protein_&_!Backbone`, i.e. both groups share a lot of bonds and the DOF estimate for these groups leads to wrong temperatures. A two-step process is thus needed to recalculate the number of DOF such that the preset temperature values are reached during the TEE-REX simulation.

In the first step, a short MD simulation of the system is made at the reference temperature $T_0$ with the two standard temperature coupling groups `Protein` and `Other` (containing solvent & ions). The obtained trajectory (`traj.trr`) serves as a reference. Next, a rerun (`mdrun -rerun traj.trr`) over this reference trajectory is performed at $T_0$ using the TEE-REX coupling groups `Backbone`, `Protein_&_!Backbone` and `Other`. Out of the obtained deviations from the reference temperature $T_0$, the correct number of DOF for each protein temperature coupling group is calculated and stored in the `mdp` file. As a check, a short TEE-REX simulation is made with the new split-protein topology and the correct number of group DOF.

We demonstrate the procedure by going through the example. The reference temperature is set to $T_0 = 300\,\mathrm{K}$ for all three coupling groups `Backbone`, `Protein_&_!Backbone` and `Other`. GROMACS is abbreviated by GMX.

#### Step 1 - Create Short MD Reference at $T_0$

A short ($\sim 100\,\mathrm{ps}$) MD simulation at $T_0 = 300\,\mathrm{K}$ is performed which serves as the reference. Degrees of freedom calculated by `grompp` are denoted by $\mathrm{DOF}_{\mathrm{GMX}}$:

| | $T_0$ | $T_{\mathrm{GMX}}$ | $\mathrm{DOF}_{\mathrm{GMX}}$ | $\mathrm{DOF}_{\mathrm{NEW}}$ |
|---|---|---|---|---|
| `Protein` | 300 K | 300.2 K | 324 | – |
| `Other` | 300 K | 300.0 K | 8670 | – |
| `SYSTEM` | 300 K | 300.0 K | 8994 | – |

### Step 2 - MD Rerun Using New Topology

From the reference trajectory `traj.trr`, a rerun is performed using the split topology `Backbone`, `Protein_&_!Backbone`, `Other` and the wrongly calculated $\mathrm{DOF}_{\mathrm{GMX}}$:

| | $T_0$ | $T_{\mathrm{GMX}}$ | $\mathrm{DOF}_{\mathrm{GMX}}$ | $\mathrm{DOF}_{\mathrm{NEW}}$ |
|---|---|---|---|---|
| `Backbone` | 300 K | 377.5 K | 51.5 | – |
| `Protein_&_!Backbone` | 300 K | 286.5 K | 272.5 | – |
| `Other` | 300 K | 300.1 K | 8670 | – |
| `SYSTEM` | 300 K | 300.0 K | 8994 | – |

### Step 3 - Recalculating $\mathrm{DOF}_{\mathrm{GMX}}$

$\mathrm{DOF}_{\mathrm{GMX}}$ are recalculated such that $T_{\mathrm{group}} = T_0$. For this, we use the equipartition theorem. According to GMX, the kinetic energy is distributed over all DOF as $2K = N_{df}^{\mathrm{GMX}} k_B T_{\mathrm{GMX}}$, with the wrong number of $\mathrm{DOF}_{\mathrm{GMX}}$, $N_{df}^{\mathrm{GMX}}$, for the respective group (`Backbone` and `Protein_&_!Backbone`). On the other hand, the correct distribution of kinetic energy reads $2K = N_{df}^{\mathrm{NEW}} k_B T_0$. Equating both expressions yields

$$N_{df}^{\mathrm{GMX}} T_{\mathrm{GMX}} = N_{df}^{\mathrm{NEW}} T_0 \quad \Leftrightarrow \quad N_{df}^{\mathrm{NEW}} = N_{df}^{\mathrm{GMX}} \frac{T_{\mathrm{GMX}}}{T_0} \,. \tag{A.1}$$

For our example, we get (solvent DOF are not affected):

| | $\mathrm{DOF}_{\mathrm{GMX}}$ | $\mathrm{DOF}_{\mathrm{NEW}}$ |
|---|---|---|
| `Backbone` | 51.5 | 64.87 |
| `Protein_&_!Backbone` | 272.5 | 259.13 |
| `Protein` | 324 | 324 |

Because of Eq. (A.1), non-integer values for the number of DOF for a group are possible. However, the total number of protein DOF is a constant which we can use as a verification.

#### Step 4 - Check $\mathrm{DOF}_{\mathrm{NEW}}$ Assignment

To check the new DOF assignment, we perform a short TEE-REX simulation (without exchanges!) using the new topology and $\mathrm{DOF}_{\mathrm{NEW}}$:

| | $T_0$ | $T_{\mathrm{GMX}}$ | $\mathrm{DOF}_{\mathrm{NEW}}$ |
|---|---|---|---|
| `Backbone` | 300 K | 301.2 K | 64.87 |
| `Protein_&_!Backbone` | 300 K | 299.9 K | 259.13 |
| `Other` | 300 K | 300.1 K | 8670 |
| `SYSTEM` | 300 K | 300.1 K | 8994 |

## A.3 Start TEE-REX Run

We have now set up the desired essential subspace $\{es\}$ in `sam.edi` and calculated the correct number of degrees of freedom for a split-protein topology if necessary. For each replica, a separate `tpr` file must be made, with the corrected $\mathrm{DOF}_{\mathrm{NEW}}$ provided in the respective `mdp` file. A TEE-REX simulation is then started by invoking both the REX and the essential dynamics option of GMX, `mdrun -ei -replex`. TEE-REX specific information is stored in the log file of the corresponding replica and can be post-processed using e.g. shell scripts.

# Bibliography


[1] J. M. Berg, J. L. Tymoczko, and L. Stryer. *Biochemistry.* W. H. Freeman an Co., New York, fifth edition, 2002.

[2] W. Hoppe, W. Lohmann, H. Markl, and H. Ziegler. *Biophysik.* Springer, 1982.

[3] J. W. Jung and W. Lee. Structure-based functional discovery of proteins: Structural proteomics. *J. Biochem. Mol. Biol.*, 37:28–34, 2004.

[4] A. T. Brunger and M. Nilges. Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR spectroscopy. *Q. Rev. Biophys.*, 26:49–125, 1993.

[5] M. Nilges. Structure calculation from NMR data. *Curr. Opin. Struct. Biol.*, 6:617–623, 1996.

[6] R. J. McIntosh. Electron microscopy of cells: a new beginning for a new century. *J. Cell. Biol.*, 153:F25–F32, 2001.

[7] J. G. Kempf and J. P. Loria. Protein dynamics from solution NMR theory and applications. *Cell Biochem. Biophys.*, 37:187–211, 2003.

[8] S. Weiss. Fluorescence spectroscopy of single biomolecules. *Science*, 283:1676–1683, 1999.

[9] H. J. Steinhoff. Methods for study of protein dynamics and protein-protein interaction in protein-ubiquitination by electron paramagnetic resonance spectroscopy. *Frontiers in Bioscience*, 7:C97–C110, 2002.

[10] J. C. Smith. Protein dynamics – comparison of simulations with inelastic neutron-scattering experiments. *Q. Rev. Biophys.*, 24:227–291, 1991.

[11] F. Gabel, D. Bicout, U. Lehnert, M. Tehei, M. Weik, and G. Zaccai. Protein dynamics studied by neutron scattering. *Q. Rev. Biophys.*, 35:327–367, 2002.

[12] M. Gerstein, A. M. Lesk, and C. Chothia. Structural mechanisms for domain movements in proteins. *Biochemistry*, 33:6739–6749, 1994.

[13] K. Moffat. The frontiers of time-resolved macromolecular crystallography: movies and chirped X-ray pulses. *Faraday Discuss.*, 122:65–77, 2003.

[14] F. Schotte, M. Lim, T. A. Jackson, A. V. Smirnov, J. Soman, J. S. Olson, G. N. Phillips Jr. and M. Wulff, and P. A. Anfinrud. Watching a protein as it functions with 150 ps time-resolved X-ray crystallography. *Science*, 300:1944–1947, 2003.

[15] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. The energy landscapes and motions of proteins. *Science*, 254:1598–1603, 1991.

[16] H. Frauenfelder and D. T. Leeson. The energy landscape in non-biological and biological molecules. *Nature Struct. Biol.*, 5:757–759, 1998.

[17] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[18] S. A. Adcock and J. A. McCammon. Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chem. Rev.*, 106:1589–1615, 2006.

[19] K. Tai. Conformational sampling for the impatient. *Biophys. Chem.*, 107:213–220, 2004.

[20] R. H. Zhou, E. Harder, H. F. Xu, and B. J. Berne. Efficient multiple time step method for use with Ewald and particle mesh Ewald for large biomolecular systems. *J. Chem. Phys.*, 115:2348–2358, 2001.

[21] M. E. Tuckerman, B. J. Berne, and G. J. Martyna. Molecular-dynamics algorithm for multiple time scales – systems with long-range forces. *J. Chem. Phys.*, 94:6811–6815, 1991.

[22] M. E. Tuckerman, B. J. Berne, and G. J. Martyna. Reversible multiple time scale molecular-dynamics. *J. Chem. Phys.*, 97:1990–2001, 1992.

[23] P. Minary, M. E. Tuckerman, and G. J. Martyna. Long time molecular dynamics for enhanced conformational sampling in biomolecular systems. *Phys. Rev. Lett.*, 93:150201, 2004.

[24] J. L. Scully and J. Hermans. Multiple time steps – limits on the speedup of molecular-dynamics simulations of aqueous systems. *Mol. Simul.*, 11:67–77, 1993.

[25] L. Greengard and V. Rokhlin. On the evaluation of electrostatic interactions in molecular modeling. *Chem. Scr.*, 29A:139–144, 1989.

[26] J. A. Board, J. W. Csey, J. F. Leathrum, A. Windemuth, and K. Schulten. Accelerated molecular-dynamics simulation with the parallel fast multipole algorithm. *Chem. Phys. Lett.*, 198:89–94, 1992.

[27] A. M. Mathiowetz, A. Jain, N. Karasawa, and W. A. Goddard. Protein simulations using techniques suitable for very large systems – the cell multipole method for nonbond interactions and the Newton-Euler inverse mass operator method for internal coordinate dynamics. *Proteins*, 20:227–247, 1994.

[28] M. Eichinger, H. Grubmüller, H. Heller, and P. Tavan. FAMUSAMM: an algorithm for rapid evaluation of electrostatic interactions in molecular dynamics simulations. *J. Comput. Chem.*, 18:1729–1749, 1997.

[29] A. Y. Toukmaji and J. A. Board. Ewald summation techniques in perspective: A survey. *Comput. Phys. Commun.*, 95:73–92, 1996.

[30] R. W. Hockney and J. W. Eastwood. *Computer Simulation Using Particles.* McGraw-Hill, New York, 1981.

[31] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *J. Comp. Chem.*, 18:1463–1472, 1997.

[32] S. Miyamoto and P. A. Kollman. SETTLE: An analytical version of the SHAKE and RATTLE algorithms for rigid water models. *J. Comp. Chem.*, 13:952–962, 1992.

[33] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23:327–341, 1977.

[34] C. M. Bennet. Mass tensor molecular dynamics. *J. Comput. Phys.*, 19:267–279, 1975.

[35] K. A. Feenstra, B. Hess, and H. J. C. Berendsen. Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J. Comput. Chem.*, 20:786–798, 1999.

[36] G. M. Torrie and J. P. Valleau. Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid.

[37] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecules. I. the method. *J. Comp. Chem*, 13:1011–1021, 1992.

[38] J. Schlitter, M. Engels, and P. Krüger. Targeted molecular dynamics: A new approach for searching pathways of conformational transitions. *J. Mol. Graph.*, 12:84–89, 1994.

[39] A. Amadei, A. B. M. Linssen, B. L. de Groot, D. M. F. van Aalten, and H. J. C. Berendsen. An efficient method for sampling the essential subspace of proteins. *J. Biom. Str. Dyn.*, 13:615–626, 1996.

[40] H. Grubmüller. Force probe molecular dynamics simulations. *Methods Mol. Biol.*, 305:493–515, 2005.

[41] H. Grubmüller. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Phys. Rev. E.*, 52:2893–2906, 1995.

[42] V. Tozzini. Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.*, 15:144–150, 2005.

[43] S. J. Marrink, A. H. de Vries, and A. E. Mark. Coarse grained model for semi-quantitative lipid simulations. *J. Phys. Chem. B*, 108:750–760, 2004.

[44] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, 112:6127–6129, 1990.

[45] A. Gosh, C. S. Rapp, and R. A. Friesner. Generalized Born model based on a surface integral formulation. *J. Phys. Chem. B*, 102:10983–10990, 1998.

[46] A. Jean-Charles, A. Nicholls, K. Sharp, B. Honig, A. Tempczyk, T. F. Hendrickson, and W. C. Still. Electrostatic contributions to solvation energies: comparison of free energy perturbation and continuum calculations. *J. Am. Chem. Soc.*, 113:1454–1455, 1991.

[47] R. Luo, L. David, and M. L. Gilson. Accelerated Poisson-Boltzmann calculations for static and dynamic systems. *J. Comput. Chem.*, 23:1244–1253, 2002.

[48] T. Noguti and N. Gō. Efficient Monte Carlo method for simulation of fluctuating conformations of native proteins. *Biopolymers*, 24:527–546, 1985.

[49] A. K. Mazur, V. E. Dorofeev, and R. A. Abagyan. Derivation and testing of explicit equations of motion for polymers described by internal coordinates. *J. Comput. Phys.*, 92:261–272, 1991.

[50] R. Abagyan and M. Totrov. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.*, 235:983–1002, 1994.

[51] E. G. Stein, L. M. Rice, and A. T. Brünger. Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation. *J. Magn. Reson.*, 124:154–164, 1997.

[52] B. Brooks and M. Karplus. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci.*, 80:6571–6575, 1983.

[53] N. Gō, T. Noguti, and T. Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci.*, 80:3696–3700, 1983.

[54] M. Levitt, C. Sander, and P. S. Stern. Normal-mode dynamics of a protein: Bovine pancreatic trypsin inhibitor. *Int. J. Quant. Chem: Quant. Biol. Symp.*, 10:181–199, 1983.

[55] M. Karplus and J. N. Kushick. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, 14:325–332, 1981.

[56] R. M. Levy, A. R. Srinivasan, W. K. Olsen, and J. A. McCammon. Quasi-harmonic method for studying very low frequency modes in proteins. *Biopolymers*, 23:1099–1112, 1984.

[57] R. M. Levy, M. Karplus, J. Kushick, and J. Perahia. Evaluation of the configurational entropy for proteins: application to molecular dynamics of an $\alpha$-helix. *Macromolecules*, 17:1370–1374, 1984.

[58] M. M. Teeter and D. A. Case. Harmonic and quasi harmonic descriptions of crambin. *J. Phys. Chem.*, 94:8091–8097, 1990.

[59] A. Kitao, F. Hirata, and N. Gō. The effects of solvent on the conformation and the collective motions of proteins - normal mode analysis and molecular-dynamics simulations of melittin in water and vacuum. *Chem. Phys.*, 158:447–472, 1991.

[60] A. E. García. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.*, 68:2696–2699, 1992.

[61] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen. Essential dynamics of proteins. *Proteins*, 17:412–425, 1993.

[62] T. D. Romo, J. B. Clarage, D. C. Sorensen, and G. N. Philipps Jr. Automatic identification of discrete substates in proteins: singular value decomposition analysis of time-averaged crystallographic refinements. *Proteins*, 22:311–321, 1995.

[63] I. Bahar, B. Erman, T. Haliloglu, and R. L. Jernigan. Efficient characterization of collective motions and interresidue correlations in proteins by low-resolution simulations. *Biochemistry*, 36:13512–13523, 1997.

[64] Z. Zhang, Y. Shi, and H. Liu. Molecular dynamics simulations of peptides and proteins with amplified collective motions. *Biophys. J.*, 84:3583–3593, 2003.

[65] J. He, Z. Zhang, Y. Shi, and H. Liu. Efficiently explore the energy landscape of proteins in molecular dynamics simulations by amplifying collective motions. *J. Chem. Phys.*, 119:4005–4017, 2003.

[66] D. M. F. van Aalten, D. A. Conn, B. L. de Groot, H. J. C. Berendsen, J. B. C. Findlay, and A. Amadei. Protein dynamics derived from clusters of crystal structures. *Biophys. J.*, 73:2891–2896, 1997.

[67] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, 80:505–515, 2001.

[68] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314:141–151, 1999.

[69] Y. Iba. Extended ensemble Monte Carlo. *Int. J. Mod. Phys. C*, 12:623–656, 2001.

[70] A. Mitsutake, Y. Sugita, and Y. Okamoto. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers*, 60:96–123, 2001.

[71] H. Fukunishi, O. Watanabe, and S. Takada. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.*, 116:9058–9067, 2002.

[72] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., New York, 2d edition, 2001.

[73] D. M. F. van Aalten, A. Amadei, G. Vriend, A. B. M. Linssen, G. Venema, H. J. C. Berendsen, and V. G. H. Eijsink. The essential dynamics of thermolysin – confirmation of hinge-bending motion and comparison of simulations in vacuum and water. *Prot. Eng.*, 8:1129–1136, 1995.

[74] D. M. F. van Aalten, J. B. C. Findlay, A. Amadei, and H. J. C. Berendsen. Essential dynamics of the cellular retinol binding protein – evidence for ligand induced conformational changes. *Prot. Eng.*, 8:1129–1136, 1995.

[75] B. L. de Groot, S. Hayward, D. M. F. van Aalten, A. Amadei, and H. J. C. Berendsen. Domain motions in bacteriophage T4 lysozyme: a comparison between molecular dynamics and crystallographic data. *Proteins*, 31:116–127, 1998.

[76] W. F. van Gunsteren and H. J. C. Berendsen. Computer-simulation of molecular-dynamics - methodology, applications, and perspectives in chemistry. *Angew. Chem. Int. Edit. Engl.*, 29:992–1023, 1990.

[77] T. Hansson, C. Oostenbrink, and W. F. van Gunsteren. Molecular dynamics simulations. *Curr. Opin. Struct. Biol.*, 12:190–196, 2002.

[78] A. R. Leach. *Molecular Modelling: Principles and Applications*. Pearson Education Limited, Essex, England, second edition, 2001.

[79] D. van der Spoel, E. Lindahl, B. Hess, A. R. van Buuren, E. Apol, P. J. Meulenhoff, D. P. Tieleman, A. L. T. M. Sijbers, K. A. Feenstra, R. van

Drunen, and H. J. C. Berendsen. Gromacs user manual version 3.2, 2004. `http://www.gromacs.org`.

[80] E. Lindahl, B. Hess, and D. van der Spoel. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.*, 7:306–317, 2001. Internet: `http://www.gromacs.org`.

[81] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, 4:187–217, 1983.

[82] S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case. An all atom force field for simulations of proteins and nucleic acids. *J. Comp. Chem.*, 7:230–252, 1986.

[83] W. F. van Gunsteren and H. J. C. Berendsen. *Groningen Molecular Simulation (GROMOS) Library Manual*. Biomos, Groningen, 1987.

[84] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118:11225–11236, 1996.

[85] A. Warshel and M. Levitt. Theoretical studies of enzymatic reactions: dielectric electrostatic and steric stabilization of carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, 103:227–249, 1976.

[86] A. Warshel, M. Kato, and A. V. Pisliakov. Polarizable force fields: history, test cases and prospects. *J. Chem. Theory Comput.*, 2007.

[87] R. W. Hockney, S. P. Goel, and J. W. Eastwood. A 10000 particle molecular dynamics model with long-range forces. *Chem. Phys. Lett.*, 21:589–591, 1973.

[88] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98:10089–10092, 1993.

[89] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. A smooth particle mesh Ewald potential. *J. Chem. Phys.*, 103:8577–8592, 1995.

[90] C. L. Brooks III, B. M. Pettitt, and M. Karplus. Structural and energetic effects of truncating long ranged interactions in ionic and polar fluids. *J. Chem. Phys.*, 83:5897–5908, 1985.

[91] P. J. Steinbach and B. R. Brooks. New spherical-cutoff methods for long-range forces in macromolecular simulation. *J. Comp. Chem.*, 15:667–683, 1994.

[92] H. C. Anderson. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.*, 72:2384–2393, 1980.

[93] S. Nose. A unified formulation of the constant temperature molecular dynamics method. *J. Chem. Phys.*, 81:511–519, 1984.

[94] H. J. C. Berendsen, J. P. M. Postma, A. Di Nola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

[95] S. M. Kast, K. Nicklas, H.-J. Bär, and J. Brickmann. Constant temperature molecular dynamics simulations by means of a stochastic collision model. 1. non-interacting particles. *J. Chem. Phys.*, 100:566–576, 1994.

[96] S. M. Kast and J. Brickmann. Constant temperature molecular dynamics simulations by means of a stochastic collision model. 2. the harmonic oscillator. *J. Chem. Phys.*, 104:3732–3741, 1996.

[97] M. B. Kubitzki and B. L. de Groot. Molecular dynamics simulations using temperature-enhanced essential dynamics replica exchange. *Biophys. J.*, 92:4262–4270, 2007.

[98] B. A. Berg and T. Neuhaus. Multicanonical algorithms for first-order phase transitions. *Phys. Lett.*, 267:249–253, 1991.

[99] A. P. Lyubartsev, A. A. Martinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov. New approach to Monte Carlo calculations of the free energy: Method of expanded ensembles. *J. Chem. Phys.*, 96:1776–1783, 1992.

[100] E. Marinari and G. Parisi. Simulated tempering: A new Monte Carlo scheme. *Europhys. Lett.*, 19:451–458, 1992.

[101] J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok, and K. A. Dill. Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations. *J. Chem. Theory Comput.*, 3:26–41, 2007.

[102] B. A. Berg and T. Celik. New approach to spin-glass simulations. *Phys. Rev. Lett.*, 69:2292–2295, 1992.

[103] S. Kumar, P. W. Payne, and M. Vásquez. Method for free-energy calculations using iterative techniques. *J. Comput. Chem.*, 17:1269–1275, 1996.

[104] G. R. Smith and A. D. Bruce. Multicanonical Monte Carlo study of solid-solid phase coexistence in a model colloid. *Phys. Rev. E*, 53:6530–6543, 1996.

[105] U. H. E. Hansmann. Effective way for determination of multicanonical weights. *Phys. Rev. E*, 56:6200–6203, 1997.

[106] C. Bartels and M. Karplus. Probability distributions for complex systems: Adaptive umbrella sampling of the potential energy. *J. Phys. Chem. B*, 102:865–880, 1998.

[107] K. Hukushima and K. Nemoto. Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. (Jap.)*, 65:1604–1608, 1996.

[108] R. H. Swendsen and J. S. Wang. Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.*, 57:2607–2609, 1986.

[109] M. C. Tesi, E. J. J. van Rensburg, E. Orlandini, and S. G. Whittington. Monte Carlo study of the interacting self-avoiding walk model in three dimensions. *J. Stat. Phys.*, 82(1-2):155–181, 1996.

[110] T. Okabe, M. Kawata, Y. Okamoto, and M. Mikami. Replica-exchange Monte Carlo method for the isobaric-isothermal ensemble. *Chem. Phys. Lett.*, 335:435–439, 2001.

[111] K. Y. Sanbonmatsu and A. E. García. Structure of met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. *Proteins*, 46:225–234, 2002.

[112] D. A. Kofke. On the acceptance probability of replica-exchange Monte Carlo trials. *J. Chem. Phys.*, 117:6911–6914, 2002.

[113] D. A. Kofke. Erratum: On the acceptance probability of replica-exchange Monte Carlo trials. *J. Chem. Phys.*, 120:10852, 2004.

[114] C. Predescu, M. Predescu, and C. V. Ciobanu. On the efficiency of exchange in parallel tempering Monte Carlo simulations. *J. Phys. Chem. B*, 109:4189–4196, 2005.

[115] D. M. Zuckerman and E. Lyman. A second look at canonical sampling of biomolecules using replica exchange simulation. *J. Chem. Theory Comput.*, 2:1200–1202, 2006.

[116] D. M. Zuckerman and E. Lyman. Erratum: A second look at canonical sampling of biomolecules using replica exchange simulation. *J. Chem. Theory Comput.*, 2:1693, 2006.

[117] X. Periole and A. E. Mark. Convergence and sampling efficiency in replica exchange simulations of peptide folding in explicit solvent. *J. Chem. Phys.*, 126:014903, 2007.

[118] E. Lyman and D. M. Zuckerman. Ensemble-based convergence analysis of biomolecular trajectories. *Biophys. J.*, 91:164–172, 2006.

[119] S. E. Murdock, K. Tai, M. H. Ng, S. Johnston, B. Wu, H. Fangohr, C. A. Laughton, J. W. Essex, and M. S. P. Sansom. Quality assurance for biomolecular simulations. *J. Chem. Theory Comput.*, 2:1477–1481, 2006.

[120] E. Lyman and D. M. Zuckerman. The structural de-correlation time: A robust statistical measure of convergence of biomolecular simulations. *arXiv:q-bio*, 0607037v2, 2007.

[121] A. M. Ferrenberg and R. H. Swendsen. New Monte Carlo technique for studying phase transitions. *Phys. Rev. Lett.*, 61:2635–2638, 1988.

[122] R. Zhou, B. J. Berne, and R. Germain. The free energy landscape for $\beta$-hairpin folding in explicit water. *Proc. Natl. Acad. Sci.*, 98:14931–14936, 2001.

[123] R. Zhou and B. J. Berne. Can a continuum solvent model reproduce the free energy landscape of a $\beta$-hairpin folding in water? *Proc. Natl. Acad. Sci.*, 99:12777–12782, 2002.

[124] R. Zhou. Free energy landscape of protein folding in water: Explicit vs. implicit solvent. *Proteins*, 53:148–161, 2003.

[125] F. Rao and A. Caflisch. Replica exchange molecular dynamics simulations of reversible folding. *J. Chem. Phys.*, 119:4035–4042, 2003.

[126] A. E. García and J. N. Onuchic. Folding a protein in a computer: An atomic description of the folding/unfolding of protein A. *Proc. Natl. Acad. Sci.*, 100:13898–13903, 2003.

[127] H. Kokubo and Y. Okamoto. Prediction of membrane protein structures by replica-exchange Monte Carlo simulations: Case of two helices. *J. Chem. Phys.*, 120:10837–10847, 2004.

[128] M. M. Seibert, A. Patriksson, B. Hess, and D. van der Spoel. Reproducible polypeptide folding and structure prediction using molecular dynamics simulations. *J. Mol. Biol.*, 354:173–183, 2005.

[129] M. Cecchini, F. Rao, M. Seeber, and A. Caflisch. Replica exchange molecular dynamics simulations of amyloid peptide aggregation. *J. Chem. Phys.*, 121:10748–10756, 2004.

[130] P. H. Nguyen, Y. Mu, and G. Stock. Structure and energy landscape of a photoswitchable peptide: A replica exchange molecular dynamics study. *Proteins*, 60:485–494, 2005.

[131] X. Cheng, G. Cui, V. Hornak, and C. Simmerling. Modified replica exchange simulation for local structure refinement. *J. Phys. Chem. B*, 109:8220–8230, 2005.

[132] P. Liu, B. Kim, R. A. Friesner, and B. J. Berne. Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proc. Natl. Acad. Sci.*, 102:13749–13754, 2005.

[133] C. Simmerling, B. Stockbine, and A. J. Roitberg. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.*, 124:11258–11259, 2002.

[134] J. W. Pitera and W. Swope. Understanding folding and design: Replica-exchange simulations of "Trp-cage" miniproteins. *Proc. Natl. Acad. Sci.*, 100:7587–7592, 2003.

[135] H. Nymeyer and A. E. García. Simulation of the folding equilibrium of $\alpha$-helical peptides: A comparison of the generalized Born approximation with explicit solvent. *Proc. Natl. Acad. Sci.*, 100:13934–13939, 2003.

[136] Y. Sugita, A. Kitao, and Y. Okamoto. Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.*, 113:6042–6051, 2000.

[137] Y. M. Rhee and V. S. Pande. Multiplexed-replica exchange molecular dynamics method for protein folding simulations. *Biophys. J.*, 84:775–786, 2003.

[138] M. Christen and W. F. van Gunsteren. Multigraining: An algorithm for simultaneous fine-grained and corase-grained simulation of molecular systems. *J. Chem. Phys.*, 124:154106, 2006.

[139] A. Okur, L. Wickstrom, M. Layten, R. Geney, K. Song, V. Hornak, and C. Simmerling. Improved efficiency of replica exchange simulations through use of a hybrid explicit/implicit solvation model. *J. Chem. Theory Comput.*, 2:420–433, 2006.

[140] R. Affentranger, I. Tavernelli, and E. di Iorio. A novel Hamiltonian replica exchange MD protocol to enhance protein conformational space sampling. *J. Chem. Theory Comput.*, 2:217–228, 2006.

[141] S. Trebst, M. Troyer, and U. Hansmann. Optimized parallel tempering simulations of proteins. *J. Chem. Phys.*, 124:174903, 2006.

[142] U. H. E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, 281:140–150, 1997.

[143] M. K. Fenwick and F. A. Escobedo. Expanded ensemble and replica exchange methods for simulation of protein-like systems. *J. Chem. Phys.*, 119:11998–12010, 2003.

[144] E. Lyman, F. M. Ytreberg, and D. M. Zuckerman. Resolution exchange simulation. *Phys. Rev. Lett.*, 96:028105, 2006.

[145] O. Lange. *Collective Langevin Dynamics of Conformational Motions in Proteins*. PhD thesis, Georg-August-Universität Göttingen, 2005.

[146] B. L. de Groot. *Native State Protein Dynamics: A Theoretical Approach*. PhD thesis, Rijksuniversiteit Groningen, 1999.

[147] A. Kitao, S. Hayward, and N. Gō. Energy landscape of a native protein: Jumping-among-minima model. *Proteins*, 33:496–517, 1998.

[148] A. Amadei, B. L. de Groot, M.-A. Ceruso, M. Paci, A. Di Nola, and H. J. C. Berendsen. A kinetic model for the internal motions of proteins: Diffusion between multiple harmonic wells. *Proteins*, 35:283–292, 1999.

[149] A. Kitao and N. Gō. Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.*, 9:143–281, 1999.

[150] S. Hayward, A. Kitao, and N. Gō. Harmonicity and anharmonicity in protein dynamics: A normal mode analysis and principal component analysis. *Proteins*, 23:177–186, 1995.

[151] B. L. de Groot, A. Amadei, R. M. Scheek, N. A. J. van Nuland, and H. J. C. Berendsen. An extended sampling of the configurational space of HPr from *E. coli*. *Proteins*, 26:314–322, 1996.

[152] B. L. de Groot, A. Amadei, D. M. F. van Aalten, and H. J. C. Berendsen. Towards an exhaustive sampling of the configurational spaces of the two forms of the peptide hormone guanylin. *J. Biomol. Str. Dyn.*, 13:741–751, 1996.

[153] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926–935, 1983.

[154] W. L. DeLano. The PyMOL molecular graphics system, 2002. `http://www.pymol.org`.

[155] M. G. Currie, K. F. Fok, J. Kato, R. J. Moore, F. K. Hamra, K. L. Duffin, and C. E. Smith. Guanylin: an endogenous activator of intestinal guanylate cyclase. *Proc. Natl. Acad. Sci.*, 89:947–951, 1992.

[156] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, 7:95–99, 1963.

[157] B. L. de Groot, D. M. F. van Aalten, R. M. Scheek, A. Amadei, G. Vriend, and H. J. C. Berendsen. Prediction of protein conformational freedom from distance constraints. *Proteins*, 29:240–251, 1997.

[158] M. Karplus and Y. Q. Gao. Biomolecular motors: the F1-ATPase paradigm. *Curr. Opin. Struct. Biol.*, 14:250–259, 2004.

[159] Z. Xu, A. L. Horwich, and P. B. Sigler. The crystal structure of the asymmetric Gro-EL-GroES-$(ADP)_7$ chaperonin complex. *Nature*, 388:741–750, 1997.

[160] U. Zachariae and H. Grubmüller. A highly strained nuclear conformation of the exportin Cse1p revealed by molecular dynamics simulations. *Structure*, 14:1469–1478, 2006.

[161] S. C. Harrison. Whither structural biology? *Nature Struct. Mol. Biol.*, 11:12–15, 2004.

[162] C. W. Müller, G. Schlauderer, J. Reinstein, and G. E. Schulz. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure*, 4:147–156, 1996.

[163] C. W. Müller and G. E. Schulz. Structure of the complex between adenylate kinase from Eschericia coli and the inhibitor $Ap_5A$ refined at 1.9Å resolution: a model for a catalytic transition state. *J. Mol. Biol.*, 224:159–177, 1992.

[164] Y. E. Shapiro and E. Meirovitch. Activation energy of catalysis-related domain motion in E. coli adenylate kinase. *J. Phys. Chem. B*, 110:11519–11524, 2006.

[165] V. Tugarinov, Y. E. Shapiro, Z. Liang, J. H. Freed, and E. Meirovitch. A novel view of domain flexibility in E. coli adenylate kinase based on structural mode-coupling $^{15}$N NMR spin relaxation. *J. Mol. Biol.*, 315:155–170, 2002.

[166] Y. E. Shapiro, E. Kahana, V. Tugarinov, Z. Liang, J. H. Freed, and E. Meirovitch. Domain flexibility in ligand-free and inhibitor bound Eschericia coli adenylate kinase based on a mode-coupling analysis of $^{15}$N spin relaxation. *Biochemistry*, 41:6271–6281, 2002.

[167] N. A. Temiz, E. Meirovitch, and I. Bahar. Eschericia coli adenylate kinase dynamics: comparison of elastic network model modes with mode-coupling $^{15}$N-NMR relaxation data. *Proteins*, 57:468–480, 2004.

[168] H. Lou and R. I. Cukier. Molecular dynamics of apo-adenylate kinase: a distance replica exchange method for the free energy of conformational fluctuations. *J. Phys. Chem. B*, 110:24121–24137, 2006.

[169] P. C. Whitford, O. Miyashita, Y. Levy, and J. N. Onuchic. Conformational transitions of adenylate kinase: switching by cracking. *J. Mol. Biol.*, 366:1661–1671, 2007.

[170] P. Maragakis and M. Karplus. Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. *J. Mol. Biol.*, 352:807–822, 2005.

[171] C. Snow, G. Qi, and S. Hayward. Essential dynamics sampling study of adenylate kinase: comparison to citrate synthase and implication for the hinge and shear mechanisms of domain motion. *Proteins*, 67:325–337, 2007.

[172] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[173] H. J. Zhang, X. R. Sheng, W. D. Niu, X. M. Pan, and J. M. Zhou. Evidence for at least two native forms of rabbit muscle adenylate kinase in equilibrium in aqueous solution. *J. Biol. Chem.*, 273:7448–7456, 1998.

[174] Y. Han, X. Li, and X. M. Pan. Native states of adenylate kinase are two active sub-ensembles. *FEBS Lett.*, 528:161–165, 2002.

[175] E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay, and D. Kern. Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, 438:117–121, 2005.

[176] D. Tobi and I. Bahar. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc. Natl. Acad. Sci.*, 102:18908–18913, 2005.

[177] S. Hayward and H. J. C. Berendsen. Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme. *Proteins*, 30:144–154, 1998.

# Acknowledgements

Am Ende dieser arbeitsreichen Zeit ist es mir ein Vergnügen, all jenen meinen Dank auszusprechen, die auf vielfältige Art und Weise zum Gelingen dieser Arbeit beigetragen haben. Besonders herzlich bedanken möchte ich mich bei Dr. Bert de Groot – für eine Betreuung wie ich sie mir nicht besser hätte wünschen können. Mit seiner freundlichen, offenen und ruhigen Art hatte er nicht nur immer ein offenes, geduldiges Ohr für sämtliche Fragen, Vorschläge und Ideen meinerseits, sondern gab mir auch sämtliche Freiheiten diese zu verfolgen. Zu jeder Zeit(!) war es möglich Probleme und Ideen mit ihm zu diskutieren, wobei er stets seinen immensen Wissens- und Erfahrungsschatz mit mir teilte. Die Art und Weise dieser Gespräche hat nachhaltigen Eindruck bei mir hinterlassen.

Prof. Dr. Tim Salditt danke ich, daß er – trotz seines übervollen Terminkalenders – bereit war, Gutachter für diese Arbeit zu sein. Ebenso danke ich Prof. Dr. Helmut Grubmüller, für seine Bereitschaft als Erstgutachter dieser Arbeit zu fungieren und darüber hinaus für das interdisziplinäre Arbeitsumfeld und exzellente Arbeitsklima in seiner Abteilung, das anderswo schwer zu finden sein dürfte; als 'ursprünglicher' Quantenphysiker habe ich sehr vom biologischen und chemischen Wissen meiner Kollegen profitiert. Als Sekretärin hat Eveline Heinemann wesentlichen Anteil am positiven Arbeitsumfeld. Das erste was mir beim Besuch Ihres Büros auffiel war die Tatsache, daß Ihr Sekretariat neben der üblichen Büroausstattung eine – wenn auch kleine – Werkstatt inklusive Drehbank und Lötkolben enthält. Auf Ihre (nicht nur handwerkliche) Kompetenz und Organisationsfähigkeit, die sich in einer unübersehbar großen Ansammlung gelber Post-Its manifestiert, war stets Verlaß. Ein weiteres Standbein der Abteilung ist die ausgezeichnete Arbeit der Systemadministratoren Dr. Ansgar Esztermann, Martin Fechner, Ingo Hoffman und Oliver Slawik. Ohne sie wären die gelegentlichen Wartungs- und Erweiterungsarbeiten am Cluster sowie an unseren Arbeitsplatzrechnern schlicht undenkbar. Obwohl im Laufe der Jahre das eine oder andere Teil meines Arbeitsplatzrechners seinen Geist aufgab, dauerte es nie länger als ein paar Stunden bis man wieder weiterarbeiten konnte. Vielen Dank dafür.

Einer Vielzahl von (ehemaligen) Kollegen verdanke ich allerlei Ratschläge, Hilfe, Kritik und Diskussionen, die wesentlichen Anteil an dieser Arbeit haben. Ohne Dr. Jürgen Haas und Dr. Carsten Kutzner wäre so manche programmiertechnische Untiefe von GROMACS und diversen Shell-Skripten nicht umschifft worden. Dr. Ira Tremmel, Dr. Carsten Kutzner und Dr. Matthias Müller danke ich für die große Sorgfalt, mit der sie meine Manuskripte gelesen haben. Mein besonderer Dank geht an Dr. Ulrich Zachariae, für unzählige Diskussionen über Wissenschaft und die Menschen die sie betreiben, das Leben im Allgemeinen sowie viel detektivischen Spürsinn.

Ohne das Wissen und die Tatsache, daß man sich immer auf die Unterstützung seiner Famile und Freunde verlassen kann, würde diese Arbeit nicht existieren. Vielen Dank. Ganz zum Schluß möchte ich Irena danken. Weil das Leben ohne Dich nicht dasselbe wäre.

# Lebenslauf von Marcus Kubitzki

## Persönliche Daten

| | |
|---|---|
| Name: | Marcus Kubitzki |
| Adresse: | Nikolaikirchhof 12, 37073 Göttingen |
| Geburtsdatum/-ort: | 14. Januar 1977, Rottweil |
| Nationalität: | Deutsch |

## Ausbildung

| | |
|---|---|
| 06/1996 | Abitur am Leibniz-Gymnasium, Rottweil<br>Prüfungsfächer: Mathematik, Physik, Deutsch und Religion<br>Note: 1.6 |
| 09/1996 – 09/1997 | Zivildienst |
| 10/1997 – 06/2003 | Universität Konstanz<br>Immatrikuliert für Diplomstudiengang Physik |
| 10/1999 | Vordiplom in Physik<br>Note: gut |
| 08/2000 – 06/2001 | Wesleyan University, Middletown CT (USA) |
| 05/2002 – 06/2003 | Diplom in Physik<br>Diplomarbeit: *State and Parameter Estimation in Quantum Theory*<br>Betreuer: Prof. Dr. J. Audretsch<br>Note: sehr gut |
| 10/2003 – 06/2004 | Universität Konstanz<br>Wissenschaftlicher Mitarbeiter, Lehrstuhl Prof. Dr. E. Bohl, Fakultät für Mathematik und Statistik |
| 07/2004 – 12/2007 | Max-Planck Institut für Biophysikalische Chemie, Göttingen<br>Doktorarbeit: *Enhanced Conformational Sampling of Proteins Using TEE-REX*<br>Betreuer: Dr. Bert de Groot<br>Note: summa cum laude |