UNIVERSITÀ DEGLI STUDI DI FIRENZE
Facoltà di Scienze Matematiche, Fisiche e Naturali

Dottorato in Scienze Chimiche, XXI Ciclo

Simulations of rare events in chemistry

Tesi di Dottorato di
Simone Marsili

Tutore:
Prof. **Piero Procacci**

Coordinatore:
Prof. **Gianni Cardini**

Settore disciplinare CHIM02
Firenze, Dicembre 2008

Contents

Introduction **iv**

1 Equilibrium averages from non-equilibrium measurements **1**

1.1 Crooks equation for steered molecular dynamics using a Nosé-Hoover thermostat 4

1.1.1 Work fluctuation theorem for Hamiltonian equations of motion 7

1.1.2 Steered Molecular Dynamics simulations of the folding and unfolding reactions of decaalanine 13

1.2 Generalization of the work fluctuation theorem in molecular dynamics simulations 25

1.A Proof of Eq.1.14 33

1.B Work fluctuation theorem using Nosé-Hoover chains 34

2 Improving history-dependent methods **35**

2.1 Metadynamics under control 37

2.1.1 Metadynamics and the Gillespie algorithm 38

2.1.2 Metadynamics simulation of an isomerization model 40

2.2 Self-Healing Umbrella Sampling 44

2.2.1 An history-dependent umbrella sampling algorithm 45

2.2.2 SHUS simulations of alanine dipeptide isomerization reaction 47

2.A Proof of convergence for the SHUS algorithm 51

2.B Reference free energy surface for the alanine dipeptide isomerization reaction 54

2.C The optimal shape of the hills 55

3 From instantaneous to reversible replica exchanges 57

3.1 The Replica Exchange Method and the work fluctuation theorem . . 59

Introduction

The molecular mechanism of chemical reactions is captured by their transition states and reaction coordinates. Reaction coordinates measure the progress of a reaction from the reactant to the product states, and transition states are high-energy intermediates on reactive paths. This simple description is made possible by the selection of a small set of variables (the reaction coordinates) in such a way that they obey approximately an autonomous set of deterministic equations, such as Ohm's law or the Langevin description of Brownian motion. The enormous number of eliminated microscopic variables are assumed to vary so much more rapidly than the few macroscopic ones that they are able to reach almost instantaneously their equilibrium distribution, the equilibrium that belongs to the instantaneous value of the reaction coordinates as it these were fixed[1]. This property is known with the name of Markov property. In principle, the dynamics of any closed isolated physical system can be described as a Markov process by introducing all microscopic variables in its description. As a matter of fact, the microscopic motion in phase space is deterministic and therefore Markovian. However, understanding a chemical or a physical problem means exactly to find its simplest description starting from its elusive microscopic nature.

Rare events are processes that occur infrequently due to dynamical bottlenecks that separate stable states. Once this threshold is crossed, however, a trajectory will move quickly to the reaction products. Clearly, a rare event cannot be neglected when attempting a coarse-grained description of a complex process and the correct reaction coordinate and transition state must be found. For example, in liquid carbon disulfide, a cyclohexane solute molecule will undergo isomerization roughly once every 0.1s[2] while, in the same liquid, a molecule will diffuse at a distance of one molecular diameter in roughly 10^{-11}s. Isomerization of cyclohexane is therefore a rare event, as it can be understood looking at the high free energy barriers separating

the chair and boat conformers, along a single collective variable[3] that is a function of the torsional angles of the molecule.

Many interesting physical, chemical and biological processes occur on time scales that exceed those accessible by molecular dynamics simulation by orders of magnitude. From this point of view, we can effectively define a rare event as a dynamical process that occur so infrequently that it is impractical to obtain quantitative information about it through straightforward trajectory calculations. In general, a rare event arises when a free energy barrier lies between two metastable states, and therefore the study of the free energy profile along a reaction coordinate is the first step in deriving a coarse-grained description of a reaction. The computation of free energy differences by means of atomistic simulations is the central issue of this thesis. Three different approaches are explored, to the purpose of accelerating the sampling of a rare event and improving the computation of free energy differences.

The first chapter is dedicated to free energy calculations through far from equilibrium measurements. A way to calculate the free energy difference between two states is simply to steer the system from one state to the other by means of a reversible transformation. The reversible work spent in the process is a state function and equals the free energy difference between these two states. However, a reversible transformation is an ideal process and in practice one has always to deal with non-equilibrium processes occurring in a finite time. Moreover, if a transition between these states is a rare event in the time scale of the steering process, the typical realization of the experiment will give a work value very different from the reversible work, and its arithmetical average on an ensemble of realizations will give only an upper bound to the free energy difference, as stated from the second law of thermodynamics, $\langle W \rangle \geq \Delta A$. Jarzynski has shown that the second law can be rewritten as an equality[4], $\langle \exp(-W/k_B T) \rangle = \exp(-\Delta A/k_B T)$, if the arithmetical average is changed in an exponential average. An even more primitive relation, the work fluctuation theorem[5], was proven by Crooks, that relates the probability of observing a certain work value performing a transformation or its time reversed. In this thesis, the validity of both relations for non Hamiltonian dynamical systems is discussed[6], along with the range of validity of the common "Gaussian approximation" and its relation to the Markovian character of the dynamics along a reaction coordinate. As shown in the last part of the chapter, these relations still hold for changes in the temperature of a thermostat or in the pressure of a barostat coupled to the system[7], for a system evolved in the NPT ensemble according to the Martyna-Tobias-Klein equations of motion[8].

In the second chapter, the so-called history dependent methods are discussed. These methods represent a dynamical evolution of the standard Umbrella Sampling technique[9]. In the latter, sampling is improved by lowering the free energy of the transition state between two stable states, adding an external unphysical potential term to the Hamiltonian of the system and thus "flattening" the free energy profile. However, in order to obtain an uniform sampling one must know *a priori* the position and the height of the free energy barrier, that is, the quantity we are trying to determine. History dependent methods, as the Wang-Landau algorithm[10] or metadynamics[11], solve this problem trying to determine the optimal biasing potential on the fly, and give dynamical rules to evolve it during a simulation. The external potential becomes a non Markovian term added to the original Hamiltonian, being a functional of the whole trajectory of the system in the space of the reaction coordinates. Given that all the relevant reaction coordinates are considered, and in the limit of slow evolution, the external potential converges to the free energy inverted in sign. In real cases, the potential does not converge but oscillates around the correct free energy[12]. An algorithm is presented, Self-Healing Umbrella Sampling[13], that solves this convergence problem of the previous approaches. Moreover, it is shown how to control the evolution of a generic potential so as to focus the computational effort on the physically relevant states in the space of the reaction coordinates.

The Replica Exchange Method[14, 15, 16, 17] (REM) provides a simple solution to the problem of sampling a rare event. In REM, several independent trajectories, called replicas, are simultaneously generated in different thermodynamic conditions. Usually, these conditions are chosen so as to span homogeneously the thermodynamic space from the ensemble of interest to a different ensemble with enhanced transition rates. During the simulation, neighboring replicas are allowed to exchange their ensemble, subject to specific acceptance criteria. In this fashion, a trajectory is no longer bound to an unique given equilibrium ensemble but can randomly walk in a thermodynamic space of different equilibrium conditions, visiting ensembles where an ergodic sampling is possible, and then going back to the quasiergodic ensemble of interest. In the third chapter, a novel and more general formalism for REM is presented, that permits replica exchanges in an arbitrary length of time and shows the deep connections between the algorithm and the work fluctuation theorem.

1

Equilibrium averages from non-equilibrium measurements

The behaviour of large, near-equilibrium systems is explained by classical thermodynamics. Let us consider a gas enclosed in a vessel of volume V in contact with a thermal bath at temperature T. The state of the gas can be modified by changing isothermally the volume of the container of ΔV. If the transformation is slow enough then the process goes through a sequence of equilibrium states and the process is called reversible. However, in general the gas will be driven out of equilibrium. Part of the total work W spent in the process will be lost as heat, resulting in entropy production. This is the content of the second law of thermodynamics as stated by Clausius

$$\Delta S \geq Q/T = -\Delta S_B \tag{1.1}$$

where Q is the heat exchanged with the bath and $\Delta S_B = -Q/T$ is the entropy change of the bath during the process. The second law can be rewritten as a relation between the work and free energy difference ΔA between the final and the initial state of the gas. Since $\Delta A = \Delta E - T\Delta S = W + Q - T\Delta S$, using Eq.1.1 one finds that

$$W \geq \Delta A \tag{1.2}$$

that is, the work spent during a transformation is an upper bound to the free energy difference ΔA.

The general procedure of statistical mechanics consists in studying the behaviour of a collection or *ensemble of systems* identical to the system of actual interest,

distributed over a range of different precise states. As we go to large numbers of degrees of freedom, the average behaviour in the appropriate ensemble is found to explain the macroscopic behaviour of the individual system, and the classical thermodynamics point of view is recovered[18].

Therefore, we can introduce a non-equilibrium experiment as follows: imagine an ensemble of systems whose evolution depends on an externally controlled parameter λ. The value of this parameter is switched from A to B in a time t, starting from equilibrium conditions. Work W is an ensemble property, and is given by

$$W = \int_0^t \mathrm{d}\tau \left\langle \frac{\partial H}{\partial \lambda} \right\rangle \dot{\lambda} = \langle w \rangle \tag{1.3}$$

where the symbol $\langle\rangle$ indicate an average over the ensemble of systems and w is the work exerted upon one of the systems, to which we will refer as microscopic work, to be distinguished from the macroscopic work W. In a reversible transformation, for each of the systems we will spent the same amount of work $w = W = W_{\mathrm{rev}}$, the reversible work necessary to bring the system from state A to state B. For finite-time processes, however, the microscopic work is a fluctuating quantity that can be characterized by its probability distribution $P(w)$. Since the second law $W \geq \Delta A$ is an average property of the systems, for some realizations the microscopic work can even be smaller than the free energy difference, $w < \Delta A$. The work fluctuation theorem[5] precisely quantifies the probability of these "transient" violations of the second law, and, together with the Jarzynski equality[4] provide exact relations to extract a free energy difference from an ensemble of non-equilibrium experiments. After the experiment is concluded, we can imagine to wait a time sufficient for the ensemble of systems to relax to equilibrium in state B and then to drive them back to the initial state A, changing the parameter λ with a time-reversed protocol. The work fluctuation theorem states that the probability $P(w)$ of measuring a work value w in the $A \to B$ transformation is related to the probability $P(-w)$ of observing a value $-w$ in the $B \to A$ transformation as

$$P(w)e^{-w/k_{\mathrm{B}}T} = P(-w)e^{-\Delta A/k_{\mathrm{B}}T} \tag{1.4}$$

The Jarzynski equality states that the exponential average over the work values measured in the initial non-equilibrium experiment satisfies

$$\int \mathrm{d}w P(w) e^{-w/k_{\mathrm{B}}T} = e^{-\Delta A/k_{\mathrm{B}}T} \tag{1.5}$$

and is obtained from the work fluctuation theorem integrating on all the possible work values.

The range of validity of these equalities is still under debate[19, 20, 21], and the central question is how to to model the *environment* and its coupling to the system during the non-equilibrium process. In Sec.1.1, a proof of the work fluctuation theorem (or Crooks equation) is provided, in the context of constant volume, constant temperature steered molecular dynamics[22] simulations of systems thermostated by means of the Nosé-Hoover method (and its variant using a chain of thermostats). As a numerical test the folding and unfolding processes of decaalanine in vacuo at finite temperature is used. The distribution of the irreversible work for the folding process is shown to be markedly non-Gaussian thereby implying, according to Crooks equation, that also the work distribution of the unfolding process must be inherently non-Gaussian. The clearly asymmetric behavior of the forward and backward irreversible work distributions is a signature of a non-Markovian regime for the folding/unfolding of decaalanine. In Sec.1.2, the proof is extended to changes in the temperature or the pressure of the environment.

1.1 Crooks equation for steered molecular dynamics using a Nosé-Hoover thermostat

Among the methods devised for calculating free energy surfaces, the Jarzynski equality[4, 23] (JE) and the correlated Crooks equation[5] (CE), are perhaps some of the most intriguing because of their far reaching theoretical implications. In fact they establish a strict correlation between two seemingly unrelated physical quantities, *i.e.* the work done on a system during irreversible (or better, dissipative) transformations and the free energy difference between the final and the initial state of the transformations. According to Crooks[5], the JE appears to follow from a more general equation (that will be referred as CE), that is Eq. 10 of Ref. [5] (see also Eq. 1.7 of the present paper). The CE is in fact a point by point relation involving statistical distributions of the work, while the JE regards average values. If, on the one side, the JE appears to be less general than the CE, on the other side it was derived using more general assumptions with respect to CE. The JE is indeed essentially based on the canonical distribution (*i.e.* the basic statistical postulate) and on the Liouville theorem[4, 24]. The CE, in its original formulation, is instead based on the microscopic reversibility and on the Markov chain assumption used, *e.g.*, in Monte Carlo simulations[25]. Crooks himself made a step forward generalizing the equation to dynamical Markovian systems[26] (*e.g.*, those obeying the over-damped Langevin equation). More recently, Evans[27], starting from the transient fluctuation theorem[28], demonstrated the CE for general (not necessarily Markovian) dynamical systems in the isokinetic thermodynamical ensemble.

From the experimental point of view, both the JE[29] and, more recently CE[30] have been verified using atomic force microscopy. However, as pointed out by several authors[31, 32, 30], these experiments have been all conducted in conditions in which the system is close to equilibrium with Gaussian or nearly Gaussian fluctuations around the mean dissipated work[31].

Recently Park and Schulten[22] have performed extensive computer experiments using steered molecular dynamics (SMD) simulations on decaalanine aimed at numerically verifying the JE and CE. In agreement with early studies[33, 34], Park and Schulten showed the statistical difficulties of estimating the free energy along the unfolding coordinate by using the JE. Nonetheless, in the forced unfolding of the α-helix form of decaalanine, they obtained seemingly Gaussian work distributions. The Gaussian shape of the work distribution was put forward as an evidence of the Markovian nature of the unfolding process. As remarked in several

studies[33, 22, 31, 35], when the work distribution in the one direction, $P_f(W)$, is Gaussian, then the CE sets strict constraints for the work distribution of the backward transformation[33, 35], $P_b(-W)$. In particular if $P_f(W)$ is Gaussian, then i) $P_b(-W)$ must also be Gaussian with identical width; ii) the intersection point of $P_b(-W)$ and $P_f(W)$ falls at $W = \Delta F$, ΔF being the free energy difference for the forward transformation; iii) the average work in the forward transformation $\overline{W}$, the variance σ of the work distributions, and the free energy difference ΔF obey the equation[33, 35]

$$\overline{W} = \Delta F - \frac{\sigma^2}{2k_B T}. \tag{1.6}$$

Applying Eq. 1.6, Park and Schulten[22] found quite contradictory results. On the one hand, their SMD simulations provided almost perfect Gaussian work distributions for two very different steering velocities. On the other hand the free energy curve calculated using the CE at the greatest steering velocity differs from the exact curve by about 20 % (in the final state of the transformation). These results put some doubt either on the validity of the CE in the context of SMD simulations or on the Gaussian (and hence Markovian) nature of the unfolding transformation. Park and Schulten did not calculate the work distribution in the backward direction (refolding process) and hence they did not fully test the CE.

In the present work the CE will be derived for a general system for which the irreversible transformation is performed by SMD simulations with stiff spring approximation and the temperature is kept fixed with a Nosé-Hoover (NH) thermostat[36, 37] and with a chain of NH thermostats[38, 39]. In a recent article[40] Jarzynski proved that the CE is valid in the context of a procedure where the initial microstates for the forward and backward transformations are taken from canonical distributions, and the transformation is performed removing the heat reservoir. The Jarzynski's proof follows straightforwardly from this demonstration by simply setting the mass of the thermostat to infinity during the transformation, that is removing the heat exchange between system and thermal bath.

Recently, Cuendet published a statistical mechanical route to the JE based on the equations of motions for the non-Hamiltonian NH dynamics[41]. The present derivation of the CE uses the same strategy of Cuendet based on the equations of motions. In this sense this study can be considered as an extension of Cuendet's work. The main difference between this derivation of the CE and the Cuendet's demonstration of the JE consists in the initial step of the proof. In the present case the starting point is the fluctuation theorem[28] that holds for a single transformation (and its time reversal), while in Ref. [41] the starting point is the ensemble average

of the exponential of the work done during a transformation. From this second point of view, since JE can be trivially derived from the CE but not viceversa, the Cuendet's derivation can be considered less general.

As exemplary system, the widely studied process of helix-coil folding of decaalanine in vacuo at finite temperature[22, 42] was considered. The two work distributions, $P_f(W)$ and $P_b(-W)$, indeed obey the CE, irrespective of the steering velocity. In addition, contrary to what is generally assumed[22, 35], such work distributions are inherently non-Gaussian. Since a Gaussian work distribution is generated when the process is Markovian[22], the observed non-Gaussian shape for the refolding transformation, far from disproving the CE, could provide additional information on the dynamical regime of decaalanine, indicating a finite damping behavior along the folding/unfolding reaction coordinate.

1.1.1 Work fluctuation theorem for Hamiltonian equations of motion

The CE has been originally derived[5] for microscopically reversible Markovian systems in the context of Monte Carlo simulations[25]. If we define a generic reaction coordinate as a function of the Cartesian coordinates of the particles of a system (*e.g.*, a distance between two atoms or a torsional angle), we can characterize every point along the reaction coordinate path by a parameter λ, such that $\lambda = 0$ and $\lambda = 1$ correspond to two ensembles of microstates (from now on indicated as macrostates $\mathcal{A}$ and $\mathcal{B}$, respectively) for which the reaction coordinate is constrained to different values. A dynamical process where λ is externally driven from zero to one, according to an arbitrary time scheduling, will be referred as *forward transformation*, while the time reversal path will be indicated as *backward transformation*. Given these definitions, the CE sets a relation between the following four quantities:

1. $P(A \rightarrow B)$, *i.e.* the joint probability of taking a microstate A from the macrostate $\mathcal{A}$ (through a canonical sampling) and of performing the forward transformation to the microstate B belonging to the macrostate $\mathcal{B}$;

2. $P(A \leftarrow B)$, *i.e.* the joint probability of taking the microstate B from the macrostate $\mathcal{B}$ (through a canonical sampling) and performing the backward transformation to the microstate A;

3. W_{AB}, *i.e.* the work done on the system during the forward transformation (from A to B);

4. $\Delta F = F(\mathcal{B}) - F(\mathcal{A})$, *i.e.* the free energy difference between the macrostates $\mathcal{A}$ and $\mathcal{B}$.

The CE reads as follows:

$$\frac{P(A \rightarrow B)}{P(A \leftarrow B)} = \exp[\beta(W_{AB} - \Delta F)] \tag{1.7}$$

where $\beta = (k_B T)^{-1}$, k_B being the Boltzmann constant and T the temperature. In the previous equation the difference $W_{AB} - \Delta F$ corresponds to the work dissipated in the forward transformation. Using the relation $W_{AB} = -W_{BA}$ (where W_{BA} is the work done on the system in the backward transformation), and grouping together all the trajectories yielding the same work (in the forward and backward transformation), the following relation can be recovered[43]

$$P_{\mathcal{A}\rightarrow\mathcal{B}}(W) = P_{\mathcal{A}\leftarrow\mathcal{B}}(-W)\ \exp[\beta(W - \Delta F)], \tag{1.8}$$

where $P_{\mathcal{A}\to\mathcal{B}}(W)$ and $P_{\mathcal{A}\leftarrow\mathcal{B}}(-W)$ are the work distribution functions obtained from the forward and backward transformations, respectively. Here W is intended to be the work done on the system in the forward transformation.

As stated in the Introduction, Eq. 1.7 sets strong limitations to the behavior of the forward and backward work distributions (Eq. 1.8). In particular, since $\int P_{\mathcal{A}\to\mathcal{B}}(W)\ \mathrm{d}W = 1$, $P_{\mathcal{A}\leftarrow\mathcal{B}}(-W)$ will vanish for $W \to \infty$ at a faster rate than $\exp(-\beta W)$, so that the integrand function decays to zero. Correspondingly, since $\int P_{\mathcal{A}\leftarrow\mathcal{B}}(-W)\ \mathrm{d}W = 1$, $P_{\mathcal{A}\to\mathcal{B}}(W)$ will decay to zero faster than $\exp(\beta W)$ for $W \to -\infty$. A Gaussian distribution indeed satisfies this condition. In this respect, Park and Schulten showed[22] that, under the assumption that the system is Markovian, SMD simulations with stiff springs result in Gaussian work distributions. However not all Gaussian distributions are permissible. In fact Eq. 1.8 establishes a relation between the moments of the normal distribution (in particular $\overline{W}$ and $\sigma^2 = \overline{W^2} - \overline{W}^2$) and ΔF. Suppose that, for a given velocity of the forward transformations $\mathcal{A} \to \mathcal{B}$, the work distribution $P_{\mathcal{A}\to\mathcal{B}}(W)$ is a (normalized) Gaussian function[22]. Then, according to Eq. 1.8, we have that

$$P_{\mathcal{A}\leftarrow\mathcal{B}}(-W) = \frac{1}{\sigma\sqrt{2\pi}}\ \exp\left[\frac{-(W-\overline{W}_{\mathcal{AB}})^2}{2\ \sigma^2}\right]\exp\left[\beta(\Delta F - W)\right] \tag{1.9}$$

where $\overline{W}_{\mathcal{AB}}$ is the average work done on the system in the forward transformations $\mathcal{A} \to \mathcal{B}$. The above equation may be rearranged as follows

$$P_{\mathcal{A}\leftarrow\mathcal{B}}(-W) = \frac{1}{\sigma\sqrt{2\pi}}\ \exp\left[\beta\left(\Delta F - \overline{W}_{\mathcal{AB}} + \frac{\beta\sigma^2}{2}\right)\right]\exp\left[\frac{-(-W+\overline{W}_{\mathcal{AB}}-\beta\sigma^2)^2}{2\ \sigma^2}\right]. \tag{1.10}$$

From the previous equation we conclude that $P_{\mathcal{A}\to\mathcal{B}}(W)$ and $P_{\mathcal{A}\leftarrow\mathcal{B}}(-W)$ are Gaussian functions with identical width. The center of $P_{\mathcal{A}\leftarrow\mathcal{B}}(-W)$ falls at $\overline{W}_{\mathcal{BA}} = -\overline{W}_{\mathcal{AB}} + \beta\sigma^2$. Moreover the intersection point of the two work distributions occurs at $W = \Delta F$. Considering that $P_{\mathcal{A}\leftarrow\mathcal{B}}(-W)$ must be normalized to one, the following equations hold

$$\Delta F = \overline{W}_{\mathcal{AB}} - \frac{\beta\sigma^2}{2} \tag{1.11}$$

$$\Delta F = -\overline{W}_{\mathcal{BA}} + \frac{\beta\sigma^2}{2}. \tag{1.12}$$

Summing term by term Eqs. 1.11 and 1.12, we get[35]

$$\Delta F = \frac{1}{2}\left(\overline{W}_{\mathcal{AB}} - \overline{W}_{\mathcal{BA}}\right). \tag{1.13}$$

Eq. 1.11 (or Eq. 1.12) can in principle be used to recover the entire free energy of the system along the λ coordinate. Interestingly, if one of the forward or backward

work distributions is not Gaussian then the other one cannot be Gaussian either. In such cases Eq. 1.13 could be used as an approximation. Alternatively, one could use directly Eq. 1.8 and histogram methods to calculate ΔF.

In deriving the CE for the case of constant volume, constant temperature SMD simulations using a NH thermostat[36, 37], we start from considering the ratio between the probability of observing a given phase space trajectory from a microstate A to a microstate B, $p[A(\mathbf{x}(0)) \rightarrow B(\mathbf{x}(\tau))]$, and the probability of observing the time-reversal trajectory, $p[A(\mathbf{x}(0)) \leftarrow B(\mathbf{x}(\tau))]$:

$$\frac{p[A(\mathbf{x}(0)) \rightarrow B(\mathbf{x}(\tau))]}{p[A(\mathbf{x}(0)) \leftarrow B(\mathbf{x}(\tau))]} = \frac{p[A(\mathbf{x}(0))]}{p[B(\mathbf{x}(\tau))]} \exp\left(-\int_0^\tau \nabla_{\mathbf{x}} \cdot \dot{\mathbf{x}} \, \mathrm{d}t\right) \tag{1.14}$$

where τ is the duration of the irreversible process, $\mathbf{x}$ is a vector in the multi-dimensional phase space, $p[A(\mathbf{x}(0))]$ and $p[B(\mathbf{x}(\tau))]$ are the probabilities (not necessarily at equilibrium) of the phase space points $\mathbf{x}(0)$ and $\mathbf{x}(\tau)$, respectively. The function $\nabla_{\mathbf{x}} \cdot \dot{\mathbf{x}}$ is the divergence of the phase space velocity, the so-called compressibility of the system[44]. Eq. 1.14 was derived by Evans[45, 28] and is extraordinarily general. In fact, it holds for both Hamiltonian and non Hamiltonian systems with time-reversal invariant equation of motions. A proof of Eq. 1.14 in the case that $p[A(\mathbf{x}(0))]$ and $p[B(\mathbf{x}(\tau))]$ are equilibrium probabilities is given in appendix 1.A.

We now assume that in the time τ the system is driven from the microstate A, characterized by the reaction coordinate $\zeta_{\mathcal{A}}$, to the microstate B, characterized by the reaction coordinate $\zeta_{\mathcal{B}}$, using a time dependent harmonic potential

$$V(\zeta(\mathbf{q}), t) = \frac{k}{2}\left[\zeta(\mathbf{q}) - \zeta_{\mathcal{A}} + (\zeta_{\mathcal{A}} - \zeta_{\mathcal{B}})\frac{t}{\tau}\right]^2 \tag{1.15}$$

The functional form of this potential implies that the reaction coordinate evolves with constant velocity. However, since an explicit expression of $V(\zeta(\mathbf{q}), t)$ is not required in the following proof, the use of a more complex time scheduling function would not change the final result. We must consider that, when $V(\zeta(\mathbf{q}), t)$ is added to the Hamiltonian of the system, the thermal energy provided by the thermostat can flow, not only from and to the physical system, but also from and to the additional potential term. The total energy of this extended system (physical system plus guiding potential) at time t is

$$H(t) = H_0 + V(\zeta, t) \tag{1.16}$$

where H_0 is the total energy of the physical system (kinetic energy plus internal potential energy). In the previous equation (and in the following) the dependence

on $\mathbf{q}$ of the reaction coordinate is omitted for simplicity of notation. The total energy change in the $A \rightarrow B$ transformation can thus be calculated as follows

$$Q_{AB} + W_{AB} = \int_0^\tau \dot{H}(t)\ \mathrm{d}t = H(\tau) - H(0) \tag{1.17}$$

where Q_{AB} and W_{AB} are the heat entering the system and the work done on the system during the transformation, respectively. The only allowed heat flow from and to the system occurs through the thermostat. Moreover, since we are dealing with a constant volume system, the work done on the system can only be performed through the guiding potential $V(\zeta, t)$. Considering Eq. 1.16, the total time derivative of $H(t)$ is

$$\dot{H}(t) = \frac{\partial V(\zeta, t)}{\partial t} + \nabla_{\mathbf{x}} V(\zeta, t) \cdot \dot{\mathbf{x}} + \nabla_{\mathbf{x}} H_0 \cdot \dot{\mathbf{x}}. \tag{1.18}$$

Substituting Eq. 1.18 into Eq. 1.17 and taking into account that the work performed on the system in the $A \rightarrow B$ transformation is

$$W_{AB} = \int_0^\tau \frac{\partial V(\zeta, t)}{\partial t}\ \mathrm{d}t, \tag{1.19}$$

one obtains

$$Q_{AB} = \int_0^\tau \nabla_{\mathbf{x}} H_0 \cdot \dot{\mathbf{x}}\ \mathrm{d}t + \int_0^\tau \nabla_{\mathbf{x}} V(\zeta, t) \cdot \dot{\mathbf{x}}\ \mathrm{d}t. \tag{1.20}$$

In this equation the integral involving H_0 corresponds to the heat provided by the thermostat to the physical system, while the other integral is the heat related to the guiding potential term. Eq. 1.20 can be written as follows

$$Q_{AB} = \int_0^\tau \sum_{i=1}^{3N} \left[\left(\frac{\partial H_0}{\partial q_i} + \frac{\partial V(\zeta, t)}{\partial q_i} \right) \dot{q}_i + \frac{\partial H_0}{\partial p_i} \dot{p}_i \right]\ \mathrm{d}t \tag{1.21}$$

where we have considered that $V(\zeta, t)$ does not depend explicitly on the momenta. Eq. 1.21 can be rearranged using the equations of motion that in the case of a system with a NH thermostat and with the guiding potential $V(\zeta, t)$ are[36, 37]

$$\begin{aligned}
\dot{q}_i &= \frac{\partial H_0}{\partial p_i} \\
\dot{p}_i &= -\frac{\partial H_0}{\partial q_i} - \frac{\partial V(\zeta, t)}{\partial q_i} - \dot{\eta} p_i \\
\dot{\eta} &= \frac{p_\eta}{M_\eta} \\
\dot{p}_\eta &= \sum_{i=1}^{3N} \frac{p_i^2}{m_i} - \frac{3N}{\beta}
\end{aligned} \tag{1.22}$$

where η and p_η are the thermostat variable and its conjugate momentum, respectively, and M_η is the related inertia factor. Using the equations of motion into Eq. 1.21, we obtain

$$Q_{AB} = -\frac{3N}{\beta}\int_0^\tau \dot{\eta}\ \mathrm{d}t + \frac{p_\eta^2(0) - p_\eta^2(\tau)}{2M_\eta}. \tag{1.23}$$

For a system coupled to a NH thermostat the compressibility is[44]

$$\nabla_{\mathbf{x}} \cdot \dot{\mathbf{x}} = -3N\dot{\eta}. \tag{1.24}$$

Substituting Eq. 1.24 into Eq. 1.23 we get

$$\int_0^\tau \nabla_{\mathbf{x}} \cdot \dot{\mathbf{x}}\ \mathrm{d}t = \beta Q_{AB} + \beta\ \frac{p_\eta^2(\tau) - p_\eta^2(0)}{2M_\eta}. \tag{1.25}$$

The next ingredient needed in Eq. 1.14 is the ratio between the equilibrium probabilities $p[A(\mathbf{x}(0))]$ and $p[B(\mathbf{x}(\tau))]$. For a system coupled to a NH thermostat, the $6N$-dimensional phase space ($3N$ particle coordinates and $3N$ conjugate momenta) is augmented by the two degrees of freedom of the thermostat, *i.e.* $\mathbf{x} = (\mathbf{q}, \mathbf{p}, \eta, \mathbf{p}_\eta)$. The ratio of the equilibrium probabilities of the microstates A and B is given by [44, 36]

$$\frac{p[A(\mathbf{x}(0))]}{p[B(\mathbf{x}(\tau))]} = \exp\left[\beta\ \frac{p_\eta^2(\tau) - p_\eta^2(0)}{2M_\eta}\right]\ \exp[\beta(H(\tau) - H(0) - \Delta F)] \tag{1.26}$$

where $H(0)$ is the energy of the physical system plus the guiding potential energy in the microstate A (Eq. 1.16). $H(\tau)$ is the same quantity for the microstate B. It is worthwhile to note that in Eq. 1.26 $\Delta F = F(\zeta = \zeta_{\mathcal{B}}) - F(\zeta = \zeta_{\mathcal{A}}) \equiv F(\mathcal{B}) - F(\mathcal{A})$ refers to equilibrium states whose Hamiltonian includes also the harmonic potential at fixed reaction coordinates $\zeta(0) \equiv \zeta_{\mathcal{A}}$ and $\zeta(\tau) \equiv \zeta_{\mathcal{B}}$. Instead, our target would be that of getting free energy differences along the reaction coordinate for a system whose Hamiltonian includes *only* the kinetic energy of the particles and the real interparticle potential energy. With this respect, Park and Schulten[22] have shown that the free energy $F(\zeta)$ of a guided system becomes identical to the true free energy of the system in the stiff spring approximation, that is for an infinite force constant k (see Eq. 1.15).

Exploiting Eqs. 1.17, 1.25 and 1.26 into Eq. 1.14, on eobtains

$$\frac{p[A(\mathbf{x}(0)) \to B(\mathbf{x}(\tau))]}{p[A(\mathbf{x}(0)) \leftarrow B(\mathbf{x}(\tau))]} = \exp[\beta(W_{AB} - \Delta F)]. \tag{1.27}$$

Eq. 1.27 is identical to Eq. 1.7 (originally derived for Markovian systems) and it has been derived for all dynamical systems coupled to a NH thermostat. As can be

seen in appendix 1.B, the demonstration reported above can be straightforwardly extended to the context of SMD simulations where the temperature is kept fixed with a NH chain algorithm[38, 39].

1.1.2 Steered Molecular Dynamics simulations of the folding and unfolding reactions of decaalanine

CE guarantees that, if the forward work distribution is Gaussian, then the backward work distribution must also be Gaussian. From a computational standpoint this fact is extremely important since it would give a practical way to compute the free energy along a reaction coordinate with the simple Eq. 1.11 (or Eq. 1.12). Gaussian work distributions were actually found in the context of SMD simulations[22], for the limited but significant case[22, 46, 47] of the unfolding of decaalanine. It is remarkable that in Ref. [22] almost Gaussian work distributions were observed for two very different steering velocities, *i.e.* $v = 10$ Å ns^{-1} and $v = 100$ Å ns^{-1}. Nonetheless, application of Eq. 1.11 provided a very good estimate of the free energy curve only for $v = 10$ Å ns^{-1}, whereas for $v = 100$ Å ns^{-1} a significant divergence from the exact result[22] was found (mainly for large end-to-end distances). This observation raises some doubts about either the validity of the CE in the context of SMD simulations, or about the Gaussian nature of the underlying distributions, that indeed for the unfolding of decaalanine "look" Gaussian[22]. In order to shed further light on this issue, the numerical experiment by Park and Schulten[22] has been repeated. In particular we have carried out SMD simulations of decaalanine at finite temperature, but focusing on both forward (unfolding) and backward (folding) trajectories.

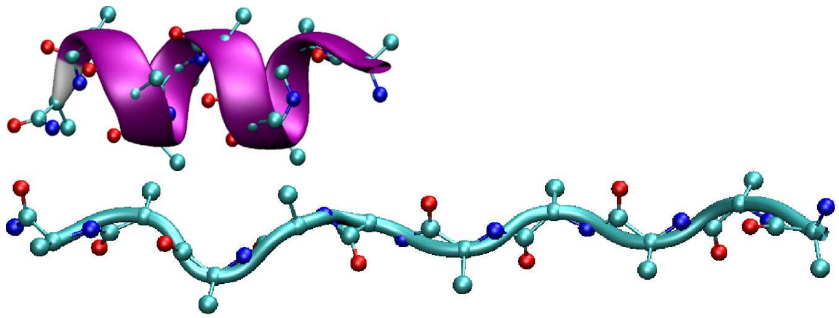

Figure 1.1: Representation of folded and unfolded conformations of the decaalanine molecule.

The N atom of the N-terminus residue has been constrained to a fixed position, while the N atom of the C-terminus residue has been constrained to move along a given fixed direction. The reaction coordinate ζ is hence taken to be the distance between the N atoms of the two terminal amide groups. Therefore the guiding potential for SMD has the form of Eq. 1.15, where $\zeta_{\mathcal{A}}$ and $\zeta_{\mathcal{B}}$ are the initial and final values of the reaction coordinate and τ is the total (simulation) time of the transformation. In the present study the stretching of decaalanine, that is, the evolution from a α-helix ($\zeta_{\mathcal{A}} = 15.5$ Å) to an elongated configuration ($\zeta_{\mathcal{B}} = 31.5$ Å), has been arbitrarily considered as the forward process. It should be noted that in general the end-to-end distance does not uniquely determine the configurational state of polypeptides. However, the equilibrium distribution at $\zeta_{\mathcal{A}} = 15.5$ Å corresponds to an ensemble of microstates tightly peaked around the α-helix structure, as for this end-to-end distance alternative structures are virtually impossible.[22, 46] The same holds true for the final totally stretched state at $\zeta_{\mathcal{B}} = 31.5$ Å. So these two equilibrium ensembles are well determined and can be effectively sampled using relatively few microstates.

The force constant used for guiding the processes (Eq. 1.15) is 800 kcal mol^{-1} Å^{-2}, which is about 100 times larger than that used in Ref. [22]. This allows to minimize the possible negative impact of the stiff spring approximation[22] on the free energy calculation. The force field for decaalanine is taken from Ref. [48]. The starting configurations of decaalanine for the forward and backward trajectories were randomly picked from standard molecular dynamics simulations of the molecule using a harmonic potential (force constant $k = 800$ kcal mol^{-1} Å^{-2}) on the end-to-end distance. The equilibrium value of the end-to-end distance was fixed to 15.5 and 31.5 Å for generating the initial configurations of the forward and backward trajectories, respectively. Constant temperature in both molecular dynamics and SMD simulations was enforced using a NH thermostat[36, 37] at the temperature of 300 K. The considered steering velocities, expressed as the simulation time τ, are 10, 20, 30, 50, 100, and 200 ps. For each steering regime 10^4 forward trajectories and 10^4 backward trajectories were generated. Such a sampling allows the quantities considered in the present study to reach a good convergence. All calculations were done with the program ORAC[49], properly modified for performing SMD simulations.

Fig. 1.2 shows the normalized work distributions $P_f(W)$ and $P_b(-W)$ for the forward and backward transformation, respectively. In agreement with Ref. [22], $P_f(W)$ indeed looks Gaussian for all steering velocities. On the contrary $P_b(-W)$ appears to deviate from the Gaussian trend for all steering regimes, the largest

deviation occurring for the slower transformations ($\tau = 100$ ps and $\tau = 200$ ps). In order to quantify this observation, in Table 1.1 we report the first four moments of $P_f(W)$ and $P_b(-W)$. For a Gaussian function the expected value of s_3 and s_4 is zero, while from the table we see significant deviations from zero for both s_3 and s_4 at all steering regimes. In particular, while for $P_f(W)$ there is a general increase of the Gaussian character with the slowing down of the transformation, the $P_b(-W)$ distributions unexpectedly (see Eq. 1.10) show the opposite behavior. Moreover, the width of $P_f(W)$ differs significantly from that of $P_b(-W)$ at all steering velocities.

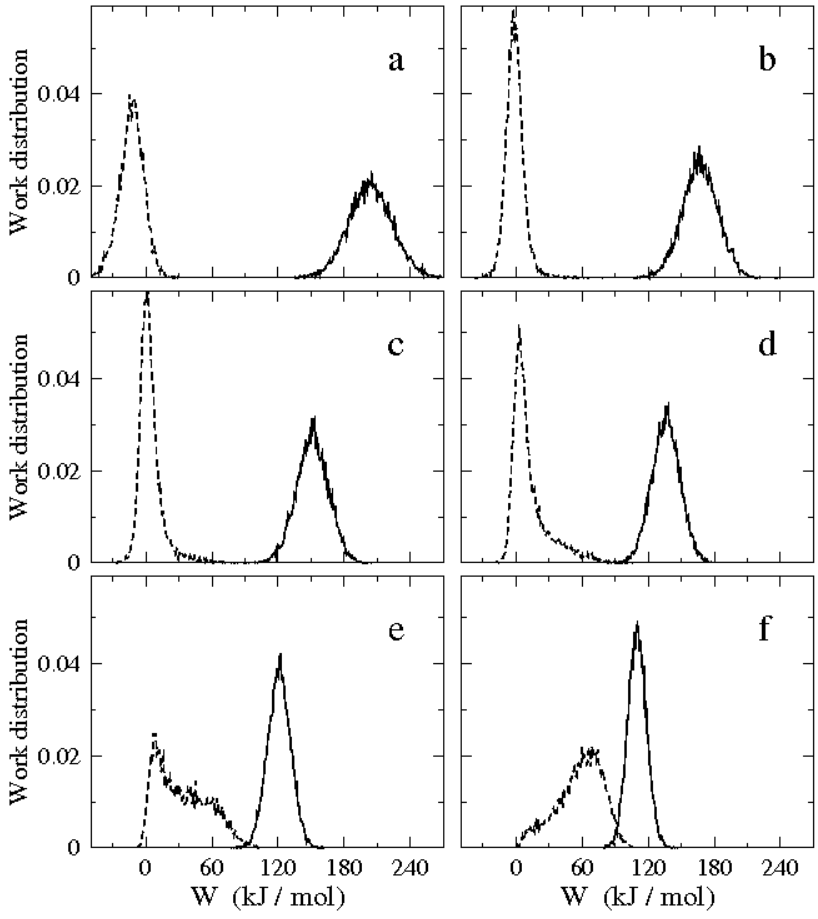

Figure 1.2: $P_f(W)$ and $P_b(-W)$ work distribution functions (solid and dashed lines, respectively) for various steering velocities ($\tau = 10$, 20, 30, 50, 100, and 200 ps from panel *a* to panel *f*).

Regarding the trends of the average values of the irreversible work ($\overline{W}_f$ and $-\overline{W}_b$ in Table 1.1), we see that, as the process is slowed down, they tend to approach each

other (see also Fig. 1.2) and will eventually become superimposed when the quasi-reversible regime is attained.

	$P_f(W)$				$P_b(-W)$			
τ (ps)	$\overline{W}_f$	σ	s_3	s_4	$-\overline{W}_b$	σ	s_3	s_4
10	204.1	20.1	7.9	12.3	-13.1	11.3	7.6	9.6
20	167.5	16.2	5.9	8.3	-1.1	8.7	8.8	13.4
30	151.8	14.6	6.3	2.7	4.1	10.8	13.9	18.4
50	136.6	12.8	4.6	6.0	12.9	16.8	20.1	22.4
100	121.2	10.5	3.3	5.1	33.6	23.6	18.5	22.3
200	110.8	8.6	3.2	5.2	58.5	20.6	16.3	15.1

Table 1.1: First four moments (in kJ mol^{-1}) of the work distributions for the forward [$P_f(W)$] and backward [$P_b(-W)$] transformations at various steering velocities.

The large and unexpected difference between the work distribution functions in the forward and backward direction (see discussion above) could be related to incomplete statistical sampling. In order to show the statistical quality of our numerical tests, in Fig. 1.3 we report the $P_f(W)$ and $P_b(-W)$ work distributions calculated for the steering velocity corresponding to $\tau = 200$ ps using 10^4 and 2 10^3 trajectories. In spite of the large difference in terms of number of considered trajectories, the two sets of distributions are very similar except for the expected noise effects. The similarity of the work distributions calculated with different sampling is also confirmed numerically by the nearly coincidence of the four moments of the distributions (data not shown). This fact suggests that the non Gaussian character of the backward work distributions has to be ascribed to the physics of the transformations, which in turn must be related to the non Markovian character of the transformations themselves.

Although the work distributions reported in Fig. 1.2 are in general not Gaussian, we could tentatively use the equations for Gaussian distributions (Eqs. 1.11 and 1.12) as done in Ref. [22], for reconstructing the potential of mean force, $F_f(\zeta)$, in the full interval spanned by the reaction coordinate. The free energy profile $F_f(\zeta)$ for the forward (unfolding) process is reported in Fig. 1.4 for the steering velocities corresponding to $\tau = 20$ ps and $\tau = 200$ ps. The exact free energy curve reported in Fig. 1.4 is calculated using the thermodynamic integration method. In order to show the amount of dissipated work along the reaction coordinate, the curve relative to the mean irreversible work is also shown. Comparing the mean irreversible work

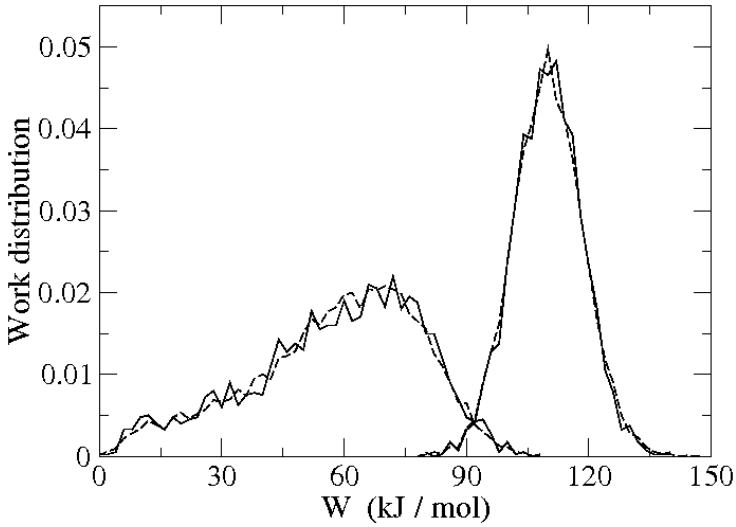

Figure 1.3: $P_f(W)$ and $P_b(-W)$ work distribution functions (curves on the right and left part of the graph, respectively) calculated for the slowest steering velocity ($\tau = 200$ ps) using 10^4 and 2 10^3 trajectories (dashed and solid lines, respectively).

at the two steering velocities (Figs. 1.4a and 1.4b), we csan appreciate the large dependence of the dissipated work on the steering regime. For $\tau = 20$ ps, the mean irreversible work deviates from the exact free energy curve for all values of ζ. For $\tau = 200$ ps we see instead that in the first stages of the transformation, *i.e.* for $15.5 < \zeta < 20$ Å, the mean irreversible work almost coincides with the exact free energy, implying a negligible dissipated work. The implications of this fact on $F_f(\zeta)$ are evident. At the lowest steering velocity the agreement between $F_f(\zeta)$ and the exact free energy is good, being less satisfactory for $\zeta > 25$ Å. In general the faster the process, the larger the deviation of the Gaussian approximant $F_f(\zeta)$ from the exact free energy.

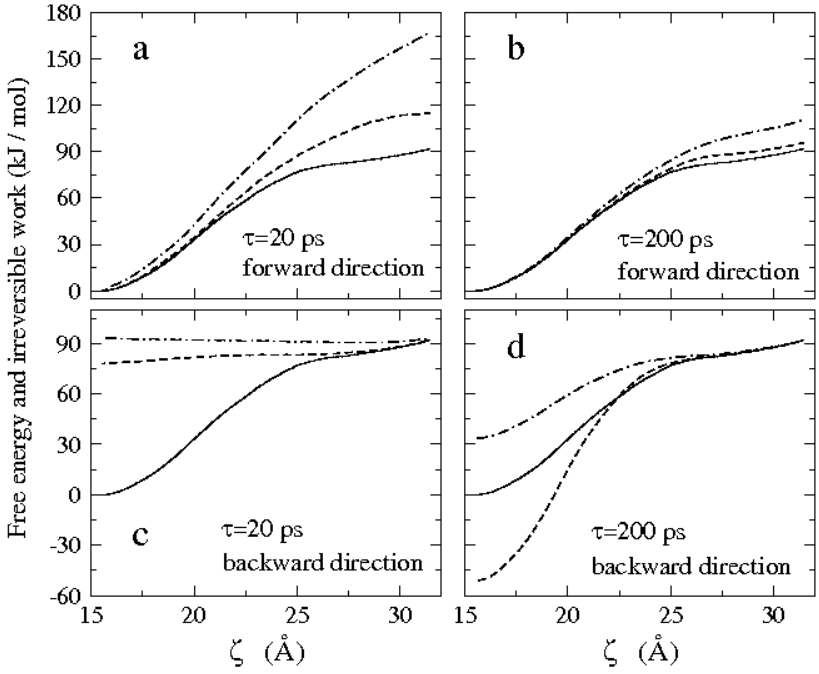

Figure 1.4: Free energy and mean irreversible work as a function of the reaction coordinate ζ for the forward and backward transformations and two steering velocities. The exact free energy and the mean irreversible work are reported with solid and dot-dashed lines, respectively. The free energy $F_f(\zeta)$ for the forward direction (dashed lines in panels *a* and *b*) is calculated using Eq. 1.11. The free energy $F_b(\zeta)$ for the backward direction (dashed lines in panels *c* and *d*) is calculated using Eq. 1.12. Panel *a*: forward direction and $\tau = 20$ ps; panel *b*: forward direction and $\tau = 200$ ps; panel *c*: backward direction and $\tau = 20$ ps; panel *d*: backward direction and $\tau = 200$ ps.

When the free energy is calculated using the data for the backward (folding) transformation (Figs. 1.4c and 1.4d), the agreement between the Gaussian approximant $F_b(\zeta)$ and the exact free energy profile becomes very unsatisfactory for both steering velocities. In particular, for the slowest quasi-reversible pulling ($\tau = 200$ ps, Fig. 1.4d), the free energy difference $F_b(31.5) - F_b(15.5)$ surprisingly differs by as much as 30 % from the exact value.

A summary of the performance of Eqs. 1.11 and 1.12 in the free energy estimate as a function of the steering regime is given in Fig. 1.5, where the free energy difference between the unfolded and folded state ($\Delta F = \Delta F_f = F_f(31.5) - F_f(15.5)$ for the forward transformation and $\Delta F = \Delta F_b = F_b(31.5) - F_b(15.5)$ for the backward transformation) is shown. ΔF_f is clearly convergent to the exact value with decreasing the steering velocity, whereas no clear trend can be extrapolated for ΔF_b.

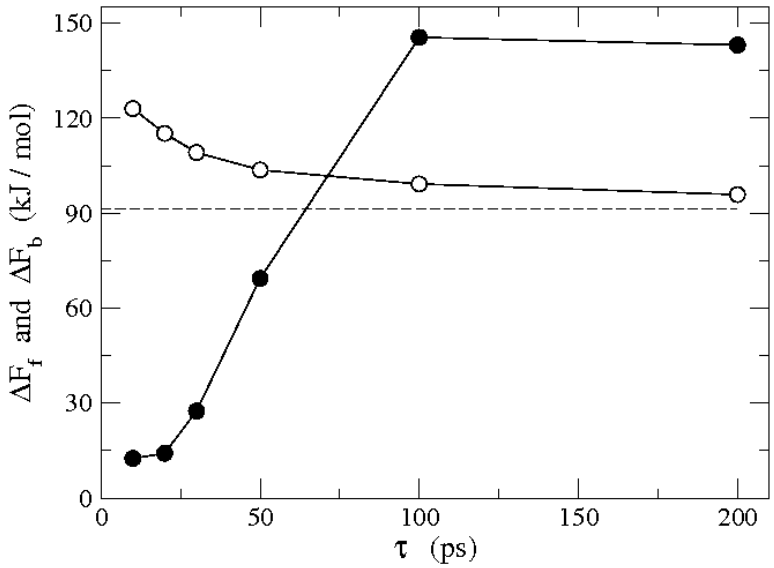

Figure 1.5: Free energy difference of the unfolded and folded states of decaalanine as a function of the steering velocity (in terms of the simulation time τ). Open circles: ΔF_f calculated from the forward trajectories using Eq. 1.11. Full circles: ΔF_b calculated from the backward trajectories using Eq. 1.12. The horizontal dashed line indicates the exact ΔF calculated through thermodynamic integration. The lines are drawn as a guide for eyes.

These results reveal a striking asymmetry of the forward and backward transformations. In fact, assuming the validity of Eq. 1.10 in the case of Gaussian (or nearly Gaussian) work distributions, it remains completely unclear why $P_f(W)$ and $P_b(-W)$ are so different (relatively narrow and apparently Gaussian $P_f(W)$, broad and strongly asymmetric $P_b(-W)$) even for slow steering velocities. As a matter of fact, as stated in the first paragraph of this section, the validity of the CE implies that if one work distribution is Gaussian, then the work distribution relative to the inverse transformation must be Gaussian too. It seems therefore reasonable to assume that the same statement should hold true for *nearly* Gaussian work distributions. On the contrary, our results clearly indicate that a nearly Gaussian (Markovian) process in one direction can be markedly non Gaussian in the reverse direction.

We may then try to calculate the free energy difference ΔF directly from Eq. 1.8, using exclusively $P_f(W)$ and $P_b(-W)$ making no assumption or approximation about their shape. According to Eq. 1.8, we note that ΔF corresponds exactly to the work, say W_x, at which the intersection of $P_f(W)$ and $P_b(-W)$ occurs. For fast transformations this point falls on the tails of the work distributions (see Fig. 1.2), that are invariably the left tail of $P_f(W)$ and the right tail of $P_b(-W)$. From a computational standpoint, the determination of ΔF becomes more and more difficult with increasing the mean dissipated work. If the steering velocity is too large, the two work distributions are far apart and W_x cannot be reliably determined (see Figs. 1.2a, 1.2b, 1.2c, and 1.2d). For low steering velocity, W_x can instead be determined even quite precisely (Figs. 1.2e and 1.2f). This scenario can be better appreciated in Fig. 1.6, where we report a zoomed view of Fig. 1.2. From the work distributions obtained at the two lowest steering velocities (Figs. 1.6e and 1.6f), we recover the almost exact ΔF. It is indeed remarkable that the estimate of ΔF using directly Eq. 1.8 with no assumption on the work distributions is much better than those reported in Fig. 1.5 where the Gaussian shape assumption was used.

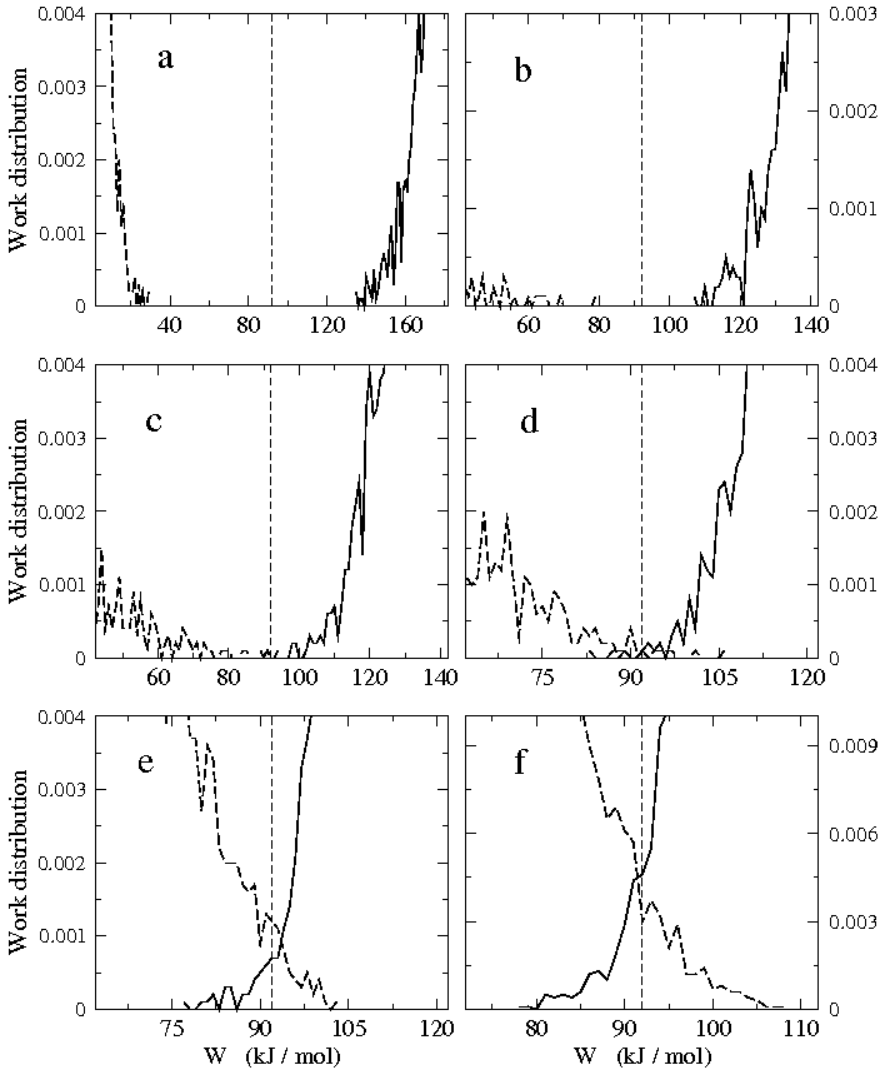

Figure 1.6: Zoomed view (from Fig. 1.2) of the $P_f(W)$ and $P_b(-W)$ work distribution functions (solid and dashed lines, respectively) for various steering velocities (τ = 10, 20, 30, 50, 100, and 200 ps from panel *a* to panel *f*). The vertical dashed lines show the value of ΔF obtained from the thermodynamic integration method.

Although Fig. 1.6 furnishes a clear and quite conclusive numerical demonstration of the validity of the CE for NH molecular dynamics, the test of the CE we provide above is essentially based on a specific and very limited aspect of the equation, namely that $W_x = \Delta F$. CE actually implies much more than this. In fact Eq. 1.8 (and our specific application to SMD simulations), if physically true, must hold for any W. We can thus in principle recover $P_b(-W)$ (or $P_f(W)$) from the knowledge of the only quantities $P_f(W)$ (or $P_b(-W)$) and ΔF. We report this test in Fig. 1.7 for the steering time $\tau = 200$ ps. The agreement between the (forward or backward) work distribution as observed in the simulations and the one derived from its counterpart (backward or forward) via CE is very good, demonstrating numerically the validity beyond any reasonable doubt of the CE in the context of NH SMD simulations. The noise observed in the retrieved work distributions is due to the unavoidable poor statistics in the tails of the original work distributions. As previously noted, the distribution $P_b(-W)$ at $\tau = 200$ ns has an unexpected non Gaussian character compared to the seemingly Gaussian shape of the corresponding $P_f(W)$ distribution. Nonetheless, as shown in Fig. 1.7, it is possible to reconstruct an important part of the non Gaussian backward distribution using the nearly Gaussian distribution of the forward process. This result points to the following conclusion: $P_f(W)$ is not Gaussian in its left tail, *i.e.* there where the body of the backward work distribution $P_b(-W)$ is carved[40]. Correspondingly, $P_b(-W)$ is approximately Gaussian only in its right tail, *i.e.* for processes ending up, in average, with a successful reforming of the α-helix. This remarkable intertwined behavior of the forward and backward work distributions is compactly and elegantly accounted for by the CE.

In the present study a theoretical proof and numerical tests of the Crooks equation (CE) have been provided, in the context of constant volume, constant temperature steered molecular dynamics simulations where the Nosé-Hoover thermostat is used. The generalization of the CE to Nosé-Hoover dynamical systems (not necessarily Markovian), along with the previous generalization provided by Evans for the isokinetic ensemble[27], strengthens the idea[27] that the Jarzynski equality and the CE have general validity, both being a manifestation of the fluctuation theorem[45].

In order to numerically verify the CE, tests on an isolated decaalanine peptide at finite temperature have been performed. Although the CE adapted to Gaussian work distributions (Eqs. 1.11 and 1.12) does not yield satisfactory results for the unfolding and (especially) folding of decaalanine, the use of the CE without any assumption on the shape of the work distributions (Eq. 1.8) allows to recover very

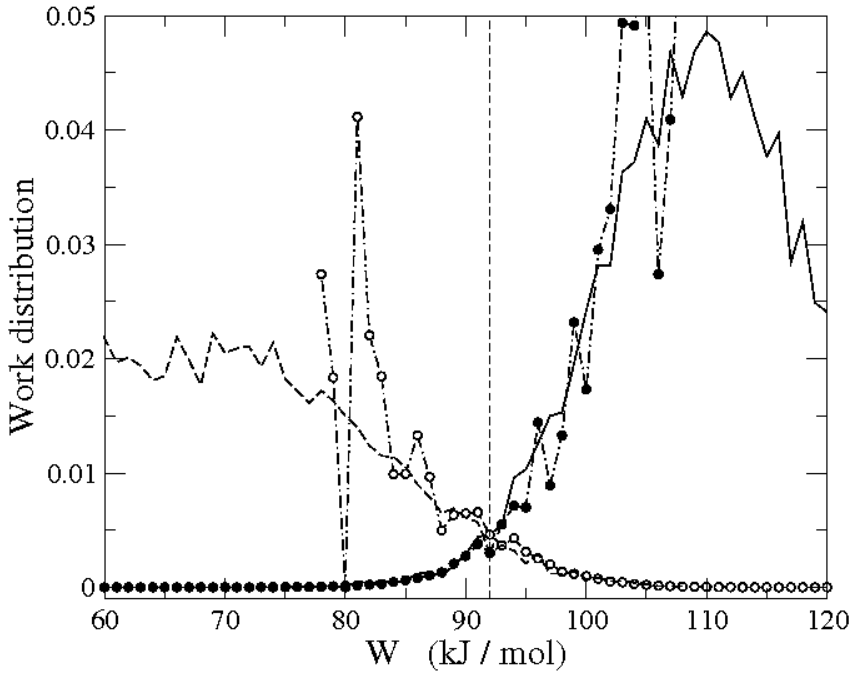

Figure 1.7: $P_f(W)$ and $P_b(-W)$ work distribution functions (solid and dashed line, respectively) for the steering velocity corresponding to $\tau = 200$ ps. The open circled line is the backward work distribution obtained from the forward work distribution via CE. The full circled line is the forward work distribution obtained from the backward work distribution via CE.

precisely the exact folding/unfolding free energy. These results i) show that the dynamics of decaalanine is far from being Markovian and ii) provide a convincing numerical test of the validity of the CE for non-Markovian systems. The left tail of the forward work distribution turns out to be a crucial feature, since it is in the left tail that, according to the CE, the shape of the backward work distribution is carved. For the same reasons, particular importance is also to be ascribed to the right tail of the backward work distribution. The behavior of the tails of the work distributions conveys a great deal of thermodynamical information, and valuable clues about the dynamical regime at the equilibrium typical of the underlying reaction coordinate.

From a practical standpoint, the CE expressed in terms of work distribution functions (Eq. 1.8), cannot be applied, as such, for reconstructing the whole free energy profile along a given reaction path. As a matter of fact the CE allows to recover only free energy differences between two well defined macrostates. A

full reconstruction of the free energy profile would require to split the interval of the reaction coordinate into several segments, where the CE machinery is applied independently. The determination of the whole free energy profile using one set of trajectories alone would be possible if the forward and backward transformation can be described by a Markovian process. However, this is not true for decaalanine and probably it is not true in general for biomolecules.

1.2 Generalization of the work fluctuation theorem in molecular dynamics simulations

In the field of molecular dynamics simulations, several routes to the free energy calculation along selected collective variables have been opened in the last decade. At variance with classical methods such as thermodynamic integration or free energy perturbation, which are based on equilibrium dynamics, most of these new approaches rely on the production of non-equilibrium trajectories. Typical examples are the recently developed adaptive bias potential methods, such as metadynamics[11] and self-healing umbrella sampling[13]. Since these techniques must in principle bring to a final equilibrium sampling in the subspace of the collective variables, they are[11] or must eventually mutate[13] into quasi-equilibrium methodologies. A substantially different scenario has been shown at the end of 90th by Jarzynski[4] and Crooks[5], who introduced "truly" non-equilibrium strategies for determining free energy differences. In particular they proposed a way to relate free energy differences between two thermodynamic states, differing in at least one (mechanical) collective variable, to the external work done on the system in an ensemble of non-equilibrium trajectories switching between the two states.

Jarzynski Identity (JI). The JI relates an exponential average of the work W to drive the system from the state A to the state B at constant temperature to the free energy difference $\Delta F = F(B) - F(A)$ between the two states,

$$\langle \mathrm{e}^{-\beta W} \rangle = \mathrm{e}^{-\beta \Delta F} \tag{1.28}$$

where $\beta = (k_B T)^{-1}$, k_B being the Boltzmann constant and T the temperature. The average quantity $\langle \exp(-\beta W) \rangle$ is calculated over different non-equilibrium phase space trajectories whose initial points are canonically distributed. Note that, since we are dealing with non-equilibrium (irreversible) trajectories, the final phase space points are not canonically distributed. After the first demonstration[4], the JI has been proved for a variety of cases from Hamiltonian and non-Hamiltonian dynamics[23, 50, 41, 51], to Langevin[52] and Markov-chain[43] dynamics. The first experimental test of the JI was published by Liphardt *et al.*[29], who applied Eq. 1.28 to measurements of the irreversible work done to mechanically stretch a single molecule of RNA.

Crooks Equation (CE). The CE relates the probability of a non-equilibrium phase space trajectory, $\Gamma_0 \to \Gamma_\tau$, to the probability of its time reversal, $\Gamma_0^* \leftarrow \Gamma_\tau^*$. The phase space points Γ_0 and Γ_τ may refer to different Hamiltonians (where, e.g.,

the distance between two molecules is constrained to different values). The CE establishes that

$$\frac{p(\Gamma_0 \to \Gamma_\tau)}{p(\Gamma_0^* \leftarrow \Gamma_\tau^*)} = \exp[\beta(W_{\Gamma_0 \to \Gamma_\tau} - \Delta F)], \tag{1.29}$$

where $W_{\Gamma_0 \to \Gamma_\tau}$ is the work done on the system during the trajectory $\Gamma_0 \to \Gamma_\tau$. In Eq. (1.29), $p(\Gamma_0 \to \Gamma_\tau)$ is the joint probability of taking the microstate Γ_0 from a canonical distribution with initial Hamiltonian and of performing the forward transformation to the microstate Γ_τ. $p(\Gamma_0^* \leftarrow \Gamma_\tau^*)$ is the analogous joint probability for the time reversal path. ΔF is the free energy difference between the final and initial thermodynamic states. In Monte Carlo or molecular dynamics simulations, a more manageable but less general form of Eq. (1.29) is used[43]. This form is easily obtained from Eq. (1.29) by summing the probabilities of all possible trajectories during which the same amount of work W is done on the system. It reads as follows:

$$\frac{P_F(W)}{P_R(-W)} = \exp[\beta(W - \Delta F)], \tag{1.30}$$

where $P_F(W)$ is the probability distribution of the work done on the system during all possible forward trajectories, while $P_R(-W)$ is the analogous distribution for the reverse paths. Note that the JI (Eq. 1.28) may be trivially recovered by rearranging Eq. 1.30 and by integrating over W. The CE was originally derived[5] for microscopically reversible Markovian systems in the context of Monte Carlo simulations. Several other proofs followed[26, 27, 6, 50, 53]. In particular, in Ref. [27] Evans pointed out the connection between the CE and the fluctuation theorem[45, 54, 55]. From the experimental point of view the CE has been verified using atomic force microscopy for the process of unfolding and refolding of a small RNA hairpin and an RNA three-helix junction[30].

The Jarzynski and Crooks theorems share the fact that the external work W is of mechanical nature and the thermodynamic conditions of the initial and final states are the same. These are indeed basic assumptions in the various proofs of the theorems and in the available computational[6] and experimental[29, 30] tests. In the present paper we propose a generalization of the JI and CE to realizations that drive the system out of equilibrium, not only using a mechanical force acting on the physical system, but also irreversibly changing the thermodynamic conditions of the physical system. This provides the opportunity of determining the relevant equilibrium quantities (the free energy difference or the ratio between the partition functions) of the initial and final states, that may differ in the basic thermodynamic quantities (P, T and V). We will also derive Eqs. 1.28 and 1.30 as special cases of such generalized equations.

In the following, we specifically address the proof for the generalized CE and we derive the corresponding generalized JI as stated above, that is by integration on the overall work variable. We start by considering a dynamic system with a given initial energy that evolves according to the Martyna-Tobias-Klein (MTK) equations of motion[8] (NPT-based dynamics). It has been shown[44] that for stationary systems such equations yield the proper NPT partition function both with and without momentum conservation. Here, we limit ourselves to the latter case, since it can be easily proved that the result does not change when the momentum conservation is applied. Suppose to drive such system out of equilibrium by an arbitrary combination of the following mechanisms: 1) introduction of some time-dependent external potential $U(t)$ that produces mechanical work on the system; 2) temperature variation through the thermostat; 3) external pressure variation through the barostat. The time schedules for the mechanical work and the pressure and temperature variations are arbitrary and mutually independent. The effect of such a transformation is to change the energy of the global system from

$$\begin{aligned} \mathcal{H}(0) &= H + U(0) + \psi_{\rm bar} + VP(0) + \\ &+ \psi_{\rm th} + \left[(3N+1)\eta_1 + \sum_{k=2}^{M} \eta_k\right] \beta^{-1}(0) \end{aligned} \tag{1.31}$$

to

$$\begin{aligned} \mathcal{H}(\tau) &= H + U(\tau) + \psi_{\rm bar} + VP(\tau) + \\ &+ \psi_{\rm th} + \left[(3N+1)\eta_1 + \sum_{k=2}^{M} \eta_k\right] \beta^{-1}(\tau), \end{aligned} \tag{1.32}$$

where H is the (potential plus kinetic) energy of the physical system, $\psi_{\rm bar} = p_\epsilon^2/(2M_b)$ is the kinetic energy associated to the barostat with mass M_b and $\psi_{\rm th} = \sum_{k=1}^{M} p_{\eta_k}^2/(2Q_k)$ is the kinetic energy associated to the thermostat (according to the MTK algorithm we use a Nosé-Hoover chain[39] with M coupled thermostats). It is important to note that, in Eqs. 1.31 and 1.32, the external potential $U(t)$, the external pressure $P(t)$, and the temperature $[k_B\beta(t)]^{-1}$, depend explicitly on time. For convenience we separate the total energy of the global system at time t, $\mathcal{H}(t)$, into two terms: the energy of the *physical system + barostat* (from now on called extended system) and the energy of the thermostat:

$$\mathcal{H}(t) = \mathcal{H}_{\rm es}(t) + \mathcal{H}_{\rm th}(t) \tag{1.33}$$

where

$$\mathcal{H}_{\rm es}(t) = H + U(t) + \psi_{\rm bar} + VP(t) \tag{1.34}$$

and

$$\mathcal{H}_{\rm th}(t) = \psi_{\rm th} + \left[(3N+1)\eta_1 + \sum_{k=2}^{M}\eta_k\right]\beta^{-1}(t). \tag{1.35}$$

For simplicity of notation, in Eqs. 1.34 and 1.35 we have expressed the dependence on t only for those quantities that depend explicitly on time. The work done on the global system during the transformation is

$$W = \int_0^\tau \frac{\partial \mathcal{H}(t)}{\partial t}\,{\rm d}t. \tag{1.36}$$

As stated above (see also Eqs. 1.34 and 1.35), three terms of the total energy $\mathcal{H}(t)$ depend explicitly on time. Correspondingly, W is given by the sum of three terms, namely

$$\begin{aligned} W &= W_{\rm m} + W_{\rm bar} + W_{\rm th} \\ &= \int_0^\tau \frac{\partial U(t)}{\partial t}\,{\rm d}t + \int_0^\tau V\frac{\partial P(t)}{\partial t}\,{\rm d}t + \\ &+ \int_0^\tau \left[(3N+1)\eta_1 + \sum_{k=2}^{M}\eta_k\right]\frac{\partial \beta^{-1}(t)}{\partial t}\,{\rm d}t, \end{aligned} \tag{1.37}$$

where $W_{\rm m}$, $W_{\rm bar}$ and $W_{\rm th}$ are the mechanical work on the physical system, the work done to produce a pressure change and the work done to produce a temperature change, respectively. The quantities $W_{\rm m}$, $W_{\rm bar}$ and $W_{\rm th}$ can be directly calculated from molecular dynamics simulations, since the time schedules of $U(t)$, $P(t)$ and $\beta(t)$ are given.

For a thermostated system with an incorporated barostat, the thermal energy provided by the thermostat during the transformation can flow, not only from and to the physical system, but also from and to the barostat. The total energy change of the extended system can thus be expressed as

$$\mathcal{H}_{\rm es}(\tau) - \mathcal{H}_{\rm es}(0) = Q + W_{\rm m} + W_{\rm bar}, \tag{1.38}$$

where Q is the heat flowing in the extended system from the thermostat and $W_{\rm m} + W_{\rm bar}$ is the total work done on the extended system. Analogously, from Eq. 1.35 we can derive the energy change of the thermostat during the transformation:

$$\begin{aligned} \mathcal{H}_{\rm th}(\tau) - \mathcal{H}_{\rm th}(0) &= W_{\rm th} + \int_0^\tau \dot{\psi}_{\rm th}\,{\rm d}t + \\ &+ \int_0^\tau \left[(3N+1)\dot{\eta}_1 + \sum_{k=2}^{M}\dot{\eta}_k\right]\beta^{-1}(t)\,{\rm d}t. \end{aligned} \tag{1.39}$$

Since in Eq. 1.38 we have arbitrarily assumed that the heat entering into the extended system is positive, the sum of the last two terms of Eq. 1.39 corresponds to $-Q$. Therefore:

$$\mathcal{H}_{\rm th}(\tau) - \mathcal{H}_{\rm th}(0) = W_{\rm th} - Q. \tag{1.40}$$

The proof proceeds by considering the so-called transient fluctuation theorem by Evans[45], that correlates the joint probabilities of Eq. 1.29 to the compressibility $\nabla_\Gamma \cdot \dot{\Gamma}$ of the system and to the probabilities $p(\Gamma_0)$ and $p(\Gamma_\tau)$ of the initial and final phase space points:

$$\frac{p(\Gamma_0 \to \Gamma_\tau)}{p(\Gamma_0^* \leftarrow \Gamma_\tau^*)} = \frac{p(\Gamma_0)}{p(\Gamma_\tau)}\ \mathrm{e}^{-\int_0^\tau \nabla_\Gamma \cdot \dot{\Gamma}\ \mathrm{d}t}. \tag{1.41}$$

In our case the probabilities $p(\Gamma_0)$ and $p(\Gamma_\tau)$ are canonically distributed. Therefore considering the expression of the canonical probability of a phase space point $\Gamma \equiv (\mathbf{p}, \mathbf{r}, p_\epsilon, V, p_\eta)$ provided by the MTK algorithm for a momentum conserving system[44], we can write:

$$\begin{aligned}\frac{p(\Gamma_0)}{p(\Gamma_\tau)} &= \frac{\mathrm{e}^{-\beta(0)[\psi_{\rm bar}(0)+\psi_{\rm th}(0)+V(0)P(0)]}}{\mathrm{e}^{-\beta(\tau)[\psi_{\rm bar}(\tau)+\psi_{\rm th}(\tau)+V(\tau)P(\tau)]}} \times \\ &\times \frac{\mathrm{e}^{-\beta(0)[H(0)+U(0)]}}{\mathrm{e}^{-\beta(\tau)[H(\tau)+U(\tau)]}}\ \frac{\omega_{P,T}^{(\tau)}}{\omega_{P,T}^{(0)}}.\end{aligned} \tag{1.42}$$

where $\omega_{P,T}^{(\tau)}$ and $\omega_{P,T}^{(0)}$ are the partition functions of the final and initial thermodynamic states in the Γ phase space, respectively. In order to obtain the partition functions in the phase space of the coordinates and momenta of the physical system, the integrals over p_ϵ and p_η in $\omega_{P,T}^{(\tau)}$ and $\omega_{P,T}^{(0)}$ must be calculated. Hence, using Eq. 1.34, we rewrite Eq. 1.42 as follows:

$$\frac{p(\Gamma_0)}{p(\Gamma_\tau)} = \frac{\mathrm{e}^{-\beta(0)[\mathcal{H}_{\rm es}(0)+\psi_{\rm th}(0)]}}{\mathrm{e}^{-\beta(\tau)[\mathcal{H}_{\rm es}(\tau)+\psi_{\rm th}(\tau)]}} \left[\frac{\beta(0)}{\beta(\tau)}\right]^m \frac{\Omega_{P,T}^{(\tau)}}{\Omega_{P,T}^{(0)}}. \tag{1.43}$$

where $\Omega_{P,T}^{(\tau)}$ and $\Omega_{P,T}^{(0)}$ are the partition functions in the phase space of the physical system and $m = (M+1)/2$. To obtain the final expression for the ratio $p(\Gamma_0 \to \Gamma_\tau)/p(\Gamma_0^* \leftarrow \Gamma_\tau^*)$, we need to determine the exponential function in Eq. 1.41. The MTK equations of motion for a momentum conserving system give rise to the following compressibility[44]:

$$\nabla_\Gamma \cdot \dot{\Gamma} = -(3N+1)\ \dot{\eta}_1 - \sum_{k=2}^{M} \dot{\eta}_k. \tag{1.44}$$

Using Eq. 1.44, the exponential function in Eq. 1.41 can be written as

$$\mathrm{e}^{-\int_0^\tau \nabla_\Gamma \cdot \dot{\Gamma}\ \mathrm{d}t} = \frac{\mathrm{e}^{-(3N+1)\ \eta_1(0) - \sum_{k=2}^{M} \eta_k(0)}}{\mathrm{e}^{-(3N+1)\ \eta_1(\tau) - \sum_{k=2}^{M} \eta_k(\tau)}}. \tag{1.45}$$

Upon substitution of Eqs. 1.43 and 1.45 into Eq. 1.41 and using Eqs. 1.33 and 1.35, we obtain

$$\frac{p(\Gamma_0 \to \Gamma_\tau)}{p(\Gamma_0^* \leftarrow \Gamma_\tau^*)} = \frac{\mathrm{e}^{-\beta(0)\mathcal{H}(0)}}{\mathrm{e}^{-\beta(\tau)\mathcal{H}(\tau)}} \left[\frac{\beta(0)}{\beta(\tau)}\right]^m \frac{\Omega_{P,T}^{(\tau)}}{\Omega_{P,T}^{(0)}}. \tag{1.46}$$

By using Eqs. 1.33, 1.38 and 1.40, $\mathcal{H}(\tau)$ can be expressed as a function of the quantities $\mathcal{H}(0)$, W_m, W_bar, and W_th:

$$\mathcal{H}(\tau) = \mathcal{H}(0) + W_\mathrm{m} + W_\mathrm{bar} + W_\mathrm{th}. \tag{1.47}$$

Upon substitution of Eq. 1.47 into Eq. 1.46 we finally get

$$\frac{p(\Gamma_0 \to \Gamma_\tau)}{p(\Gamma_0^* \leftarrow \Gamma_\tau^*)} = \frac{\Omega_{P,T}^{(\tau)}}{\Omega_{P,T}^{(0)}}\ \mathrm{e}^{\beta(\tau)W + [\beta(\tau)-\beta(0)]\mathcal{H}(0) + m \ln \frac{\beta(0)}{\beta(\tau)}}, \tag{1.48}$$

where $W = W_\mathrm{m} + W_\mathrm{bar} + W_\mathrm{th}$ and must be calculated following Eq. 1.37. Eq. 1.48 relates the probability of a general non-equilibrium transformation (i.e., involving mechanical work, and pressure and temperature changes) and its time reversal, to the total work done on the global system in the forward process and to the partition functions of the initial and final states. Eq. 1.46 (or equivalently Eq. 1.48) is the generalized form of Eq. 1.29 and is the central result of the present paper.

The extension of the CE to systems where the volume (instead of the external pressure) and the temperature change during the transformation due to external work is straightforward. In such case the MTK equations of motion reduce to the Nosé-Hoover chain equations[39] (NVT dynamics) and the energy of the global system is

$$\begin{aligned} \mathcal{H}(t) &= H[V(t)] + U(t) + \psi_\mathrm{th} + \\ &+ \left[(3N+1)\eta_1 + \sum_{k=2}^{M} \eta_k\right] \beta^{-1}(t). \end{aligned} \tag{1.49}$$

In Eq. 1.49, the dependence of the energy of the physical system on the volume is explicitly given because the volume, and hence the energy, may be arbitrarily changed during the transformation. Moreover, since the external pressure is constant, the

relation $W_{\rm bar} = 0$ holds. However, as previously stated, the physical system may undergo additional work, say $W_{\rm vol}$, during the transformation

$$W_{\rm vol} = \int_0^\tau \frac{\partial H(V)}{\partial V}\ \dot{V}\ {\rm d}t. \tag{1.50}$$

By following the guideline that brought to Eq. 1.48, we may recover the generalized CE for NVT dynamic systems:

$$\frac{p(\Gamma_0 \to \Gamma_\tau)}{p(\Gamma_0^* \leftarrow \Gamma_\tau^*)} = \frac{\Omega_{V,T}^{(\tau)}}{\Omega_{V,T}^{(0)}}\ {\rm e}^{\beta(\tau)W + [\beta(\tau)-\beta(0)]\mathcal{H}(0) + m\ln\frac{\beta(0)}{\beta(\tau)}}, \tag{1.51}$$

where $\mathcal{H}(0)$ is given by Eq. 1.49 (with $t = 0$), $m = M/2$ (only the integrals over p_η are present in $\omega_{V,T}^{(\tau)}$ and $\omega_{V,T}^{(0)}$) and $W = W_{\rm m} + W_{\rm vol} + W_{\rm th}$.

In order to recover the generalized versions of Eqs. 1.28 and 1.30, we define the following adimensional functional of a generic trajectory that brings the system from the state A at time 0 to the state B at time τ:

$$\mathcal{W} \equiv \mathcal{W}_{AB} = \beta_B W + (\beta_B - \beta_A)\mathcal{H}_A + m\ln\frac{\beta_A}{\beta_B}, \tag{1.52}$$

where $(k_B\beta_A)^{-1}$ and $(k_B\beta_B)^{-1}$ are the temperatures of the initial and final states, respectively, and $\mathcal{H}_A$ is the energy of the global system (Eq. 1.31) in the initial state. Using the above definition, collecting all trajectories yielding the same $\mathcal{W}$, and exploiting the fact that $\mathcal{W}_{BA} = -\mathcal{W}_{AB} \equiv -\mathcal{W}$, Eq. 1.48 transforms as follows

$$\frac{P_F(\mathcal{W})}{P_R(-\mathcal{W})} = {\rm e}^{\mathcal{W}}\ \frac{\Omega_B}{\Omega_A}, \tag{1.53}$$

where $P_F(\mathcal{W})$ and $P_R(-\mathcal{W})$ are the normalized distribution functions of $\mathcal{W}$ and $-\mathcal{W}$ for the forward and backward transformations, respectively. Multiplying both sides of Eq. 1.53 by ${\rm e}^{-\mathcal{W}}P_R(-\mathcal{W})$ and integrating the resulting equation over $\mathcal{W}$, the generalized JI is obtained

$$\langle {\rm e}^{-\mathcal{W}} \rangle = \frac{\Omega_B}{\Omega_A}. \tag{1.54}$$

With analogous considerations, it can be shown that the functional relations of Eqs. 1.53 and 1.54 are also valid for NVT-based dynamics (Eq. 1.51). The difference between NVT and NPT dynamic systems stems from the meaning of the quantities W, $\mathcal{H}_A$, Ω_A and Ω_B as discussed above.

It is now straightforward to derive Eqs. 1.28 and 1.30 as special cases of the non-equilibrium work theorems expressed by Eqs. 1.54 and 1.53. To this aim we consider the particular case in which the temperature of the thermodynamic states

A and B is the same. Such condition implies that $\beta_B = \beta_A = \beta$ in Eq. 1.52 and therefore $\mathcal{W} = \beta W$. The same condition allows us to relate the ratio Ω_B/Ω_A to the free energy difference between the states A and B. In particular: $\Omega_B/\Omega_A = e^{-\beta\Delta G}$ for NPT dynamic systems and $\Omega_B/\Omega_A = e^{-\beta\Delta F}$ for NVT dynamic systems. The CE and JI are easily recovered using the relations obtained for $\mathcal{W}$ and Ω_B/Ω_A into Eqs. 1.53 and 1.54, respectively.

In conclusion the generalized non-equilibrium relations we present could be fruitfully exploited, not only for the direct determination of free energy differences, but also used in thermodynamic cycles. Our results may open interesting perspectives either into computational or into experimental field, providing a framework where both intensive and extensive thermodynamic variables can be freely manipulated during the non-equilibrium measurements.

1.A Proof of Eq.1.14

The proof of Eq. 1.14 proceeds as follows. Since the dynamics is deterministic, the probability ratio of the $A \to B$ and $A \leftarrow B$ transformations is simply given by the ratio between the number of initial points of the $A \to B$ process and the number of initial points of the time-reversal $A \leftarrow B$ process:

$$\frac{p[A(\mathbf{x}(0)) \to B(\mathbf{x}(\tau))]}{p[A(\mathbf{x}(0)) \leftarrow B(\mathbf{x}(\tau))]} = \frac{p[A(\mathbf{x}(0))]\ \delta\mathbf{x}(0)}{p[B(\mathcal{M}\mathbf{x}(\tau))]\ \mathcal{M}\delta\mathbf{x}(\tau)} \tag{1.55}$$

where $\mathcal{M}$ is the time-reversal operator such that $\mathcal{M}(\mathbf{q}, \mathbf{p}) = (\mathbf{q}, -\mathbf{p})$. In the previous equation $\delta\mathbf{x}(0)$ and $\delta\mathbf{x}(\tau)$ are the volume elements of the phase space at the points $\mathbf{x}(0)$ and $\mathbf{x}(\tau)$, respectively. $p[A(\mathbf{x}(0))]$ and $p[B(\mathcal{M}\mathbf{x}(\tau))]$ are the equilibrium probabilities of the states $\mathbf{x}(0)$ and $\mathcal{M}\mathbf{x}(\tau)$. Since the equilibrium probability of a phase space state is independent on the sign of the momenta, the time-reversal operator does not affect the probability at the denominator, *i.e.* $p[B(\mathcal{M}\mathbf{x}(\tau))] = p[B(\mathbf{x}(\tau))]$. For the time-reversal trajectory, the volume element $\mathcal{M}\delta\mathbf{x}(\tau)$ is related to the volume element $\mathcal{M}\delta\mathbf{x}(0)$ through the Jacobian $J = \exp\left(-\int_0^\tau \nabla_{\mathbf{x}} \cdot \dot{\mathbf{x}}\ \mathrm{d}t\right)$ of the transformation $\mathcal{M}\delta\mathbf{x}(0) \leftarrow \mathcal{M}\delta\mathbf{x}(\tau)$,

$$\mathcal{M}\delta\mathbf{x}(0) = \exp\left(-\int_0^\tau \nabla_{\mathbf{x}} \cdot \dot{\mathbf{x}}\ \mathrm{d}t\right)\ \mathcal{M}\delta\mathbf{x}(\tau). \tag{1.56}$$

Substituting Eq. 1.56 into Eq. 1.55 and exploiting the invariance of the phase space volume elements upon application of the time-reversal operator, *i.e.* $\mathcal{M}\delta\mathbf{x} = \delta\mathbf{x}$, we recover Eq. 1.14.

1.B Work fluctuation theorem using Nosé-Hoover chains

The demonstration of the CE for the case of a chain of M NH thermostats whose inertial factors are M_{η_1}, M_{η_2}, ..., M_{η_M} follows the guideline we have described in Sec. 1.1.1. The substantial difference occurs in the equations of motion that in the case of the NH chain algorithm[38, 39] coupled to a guiding potential are:

$$\begin{aligned}
\dot{q}_i &= \frac{\partial H_0}{\partial p_i} \\
\dot{p}_i &= -\frac{\partial H_0}{\partial q_i} - \frac{\partial V(\zeta,t)}{\partial q_i} - \dot{\eta}_1 p_i \\
\dot{\eta}_k &= \frac{p_{\eta_k}}{M_{\eta_k}}, \qquad k = 1, ..., M \\
\dot{p}_{\eta_1} &= \sum_{i=1}^{3N} \frac{p_i^2}{m_i} - 3N\beta^{-1} - \frac{p_{\eta_2}}{M_{\eta_2}} p_{\eta_1} \\
\dot{p}_{\eta_k} &= \frac{p_{\eta_{k-1}}^2}{M_{\eta_{k-1}}} - \beta^{-1} - \frac{p_{\eta_{k+1}}}{M_{\eta_{k+1}}} p_{\eta_k}, \qquad k = 2, ..., M-1 \\
\dot{p}_{\eta_M} &= \frac{p_{\eta_{M-1}}^2}{M_{\eta_{M-1}}} - \beta^{-1}.
\end{aligned} \tag{1.57}$$

Combining the equations of motion reported above with Eq. 1.21, we recover the analog of Eq. 1.23:

$$Q_{AB} = -3N\beta^{-1} \int_0^\tau \dot{\eta}_1 \, \mathrm{d}t - \beta^{-1} \sum_{k=2}^{M} \int_0^\tau \dot{\eta}_k \, \mathrm{d}t + \sum_{k=1}^{M} \frac{p_{\eta_k}^2(0) - p_{\eta_k}^2(\tau)}{2M_{\eta_k}}. \tag{1.58}$$

It is easy to prove that the compressibility of the system (analog of Eq. 1.24) is

$$\nabla_{\mathbf{x}} \cdot \dot{\mathbf{x}} = -3N\dot{\eta}_1 - \sum_{k=2}^{M} \dot{\eta}_k. \tag{1.59}$$

Combining Eqs. 1.58 and 1.59, we get the analog of Eq. 1.25

$$\int_0^\tau \nabla_{\mathbf{x}} \cdot \dot{\mathbf{x}} \, \mathrm{d}t = \beta Q_{AB} + \beta \sum_{k=1}^{M} \frac{p_{\eta_k}^2(\tau) - p_{\eta_k}^2(0)}{2M_{\eta_k}}. \tag{1.60}$$

Finally, for a system coupled to a NH chain of thermostats, the ratio between the equilibrium probabilities $p[A(\mathbf{x}(0))]$ and $p[B(\mathbf{x}(\tau))]$ is (analog of Eq. 1.26):

$$\frac{p[A(\mathbf{x}(0))]}{p[B(\mathbf{x}(\tau))]} = \exp\left[\beta \sum_{k=1}^{M} \frac{p_{\eta_k}^2(\tau) - p_{\eta_k}^2(0)}{2M_{\eta_k}}\right] \exp[\beta(H(\tau) - H(0) - \Delta F)]. \tag{1.61}$$

Upon substitution of Eqs. 1.61, 1.60 and 1.17 into Eq. 1.14, we recover the CE (Eq. 1.27).

2

Improving history-dependent methods

In a simulation, a possible strategy to enhance the transition rates between two metastable states and to focus attention on a rare event is simply to change the original free energy landscape by adding an artificial biasing potential. This solution is known as Non Boltzmann or Umbrella Sampling[56]. If the transition state can be identified and located at some value of the reaction coordinate, the biasing potential $V(s)$ can be used to balance the activation barrier so as to flatten the free energy landscape along the reaction coordinate s. However, the height and the position of the transition state, as well as the free energy difference between the two metastable states, are more the results of a simulation than known *a priori* parameters, and therefore a great effort was devoted to develop adaptive biasing methods [57, 58, 10, 11, 59], that improve the biasing potential "on the fly" as the simulation proceeds, using the previous part of the trajectory to build up an history-dependent potential.

An history-dependent potential $V(s,t)$ is an artificial potential that changes in time according to the trajectory followed by the system, disfavoring already visited configurations. Let us consider an ensemble of replicas of the system, spread along the reaction coordinate s with a probability distribution $p(s,t)$. The evolution of the biasing potential is then given by

$$\frac{\partial V(s,t)}{\partial t} = \omega p(s,t) \tag{2.1}$$

where ω has the dimensions of an energy rate. If this rate is sufficiently slow, we can suppose that $p(s,t)$ is the equilibrium distribution for the biasing potential $V(s,t)$

at time t. Then, when $V(s,t)$ flattens the free energy, the systems in the ensemble are distributed uniformly along the reaction coordinate and the biasing potential is stationary within an addictive constant.

In this chapter, two aspects of this approach are addressed. In Sec.2.1, an alternative form of Eq.2.1 is proposed which contains the additional term p^-

$$\frac{\partial V(s,t)}{\partial t} = \omega \left[p(s,t) - p^-(s,t) \right] \tag{2.2}$$

that determines the probability to lower the potential in state s at time t. Such term is exploited to prevent the biasing potential to grow indefinitely. In Sec.2.2 an algorithm is introduced whose biasing potential depends logarithmically on an estimate of the equilibrium distribution $\rho(s)$ along s,

$$V(s,t) = \log \rho(s,t) \tag{2.3}$$

ρ is calculated from the previous history of the trajectory as

$$\rho(s,t) = \frac{\tilde{N}(s,t)}{A(t)} \tag{2.4}$$

where $\tilde{N}$ is a reweighted histogram, and $A(t)$ a time dependent normalization constant. At variance with previous approaches, the algorithm is shown to converge *quantitatively* to the free energy surface of the system.

2.1 Metadynamics under control

Metadynamics[11] is an example of algorithm that relies on an history-dependent potential. A metadynamics simulation consists in two steps. In the first one, a set of reaction coordinates is chosen whose dynamics describes the process under study. Such a procedure requires an high degree of chemical and physical intuition for its application to complex molecular system, since these variables cannot be obviously determined from the molecular structure, and, as outlined in the Introduction, it should be considered more the result of a simulation than its starting point. The second step is the metadynamics simulation, during which an history-dependent potential is constructed by summing up potential terms, commonly referred as "hills", at regular time intervals along the trajectory in the space of the reaction coordinates. The shape of these hills can be different, and the optimal choice from the computational point of view will be described in section 2.C. This non-Markovian potential term pushes the system to visit new states at a faster rate with respect to a standard dynamics, while at the same time reconstructing the free energy landscape as the biasing potential inverted in sign.

In this section a simple argument is proposed to evolve the metadynamics potential, while keeping fixed the number of potential terms or "hills". Roughly speaking, looking at a metadynamics run as a "flooding" of a free energy surface with the biasing potential as a fluid, this would corresponds to stop the flooding and look at the fluid waving. Such a procedure permits to restrict the sampling to the accessible states given such a fixed "volume" of the biasing potential. From the computational point of view, since it stops the number of hills to grow indefinitely during a metadynamics run, this method permits to fix *a priori* the computational cost of integrating the forces from the biasing potential.

2.1.1 Metadynamics and the Gillespie algorithm

In a metadynamics simulation, the algorithm keeps on adding terms to the history-dependent potential, until the simulations stops. However, it is often difficult to decide when to terminate a metadynamics run. As a matter of fact, in a single run, even if all the relevant slow modes of the system are accelerated by the biasing potential, the free energy does not converge to a definite value but fluctuates around the correct result, leading to an average error which is proportional to the square root of the bias potential deposition rate [12, 60]. Therefore, averages on an ensemble of independent realizations are needed. Furthermore, in practical applications, continuing a run carries the risk that the system is irreversibly pushed in regions of configurational space which are not physically relevant. These issues have already been recognized and different solutions have been proposed to alleviate these problems [12, 13, 61, 62].

In a standard metadynamics simulation, the evolution of the biasing potential is given on average by

$$\frac{\partial V(s,t)}{\partial t} = \omega p(s,t) \tag{2.5}$$

where ω is the potential deposition rate, and $p^+(s)$ denotes the probability of deposing an hill in state s, and therefore coincides with the probability of finding the system in the state s at time t. In the slow deposition limit, one can assume that the system is in equilibrium with the external potential , and therefore rewrite the previous equation as

$$\frac{\partial V(s,t)}{\partial t} = \omega \frac{e^{-\beta(F(s)+V(s,t))}\mathrm{ds}}{Z_\Omega(t)} \tag{2.6}$$

where

$$Z_\Omega(t) = \int_\Omega e^{-\beta(F(s)+V(s,t))}\mathrm{ds} \tag{2.7}$$

is a time-dependent normalization. The subscript Ω denotes the subset of the state space accessible to the system, given the biasing potential at time t is $V(t)$, that is, those states separated by free energy barriers not greater than a few β^{-1} from the current state of the system.

Even if the set Ω is clearly defined by the topology of the free energy landscape, and the latter is locally flattened by the added potential for all states included in Ω, the potential $V(s)$ keeps on growing with a rate independent of s:

$$\frac{\partial V(s,t)}{\partial t} = \omega \frac{\mathrm{ds}}{\Omega} \tag{2.8}$$

since for $V(s,t) = -F(s)$ on Ω one has $Z_\Omega(t) = \Omega$. Since the potential grows indefinitely, the system will be indefinitely pushed to visit new, less probable states,

eventually sampling configurations outside of the collection of metastable states and transition states we are interested in. Moreover, the potential will be updated slower and slower with the increase of the number of accessible states, Ω, and its computational burden will increase.

This "overfilling" problem can be circumvented by adding a new term to Eq.2.5

$$\frac{\partial V(s,t)}{\partial t} = \omega[p(s,t) - p^{-}(s,t)] \tag{2.9}$$

where $p^{-}(s,t)$ denotes the probability of deposing a negative hill (or "subtract" an hill) at point s. The simplest choice is an uniform probability on the set Ω, $p^{-}(s,t) = \mathrm{ds}/\Omega$. Then, when $V(s) = -F(s)$ on Ω, now we have

$$\frac{\partial V(s,t)}{\partial t} = \omega\left(\frac{\mathrm{ds}}{\Omega} - \frac{\mathrm{ds}}{\Omega}\right) = 0 \tag{2.10}$$

and the potential is stationary.

Such a procedure can be illustrated as moving a preexistent hill to a new position, establishing a dynamics of the hills and conserving their total number. How to choose an hill such that its position is randomly distributed on Ω? Random hills clearly will not be distributed uniformly. However, the number of hills deposed in a state s, $n(s,t)$ is linearly related to the biasing potential in that point by the relation $V(s,t) \simeq n(s,t)h$ where h is the average height of an hill. Therefore, a correct procedure is to choose an hill with a probability that is inversely proportional to the biasing potential in s. This can be achieved by placing, for each hill, a segment on a line, of length proportional to $1/V(s)$. Choosing a single random position along the length of this stack of segments will select the segment corresponding to a specific hill, as in the Gillespie algorithm[63] it selects a specific pathway for the system to follow among a set of possible pathways with different rate constants.

2.1.2 Metadynamics simulation of an isomerization model

In order to illustrate the efficiency of the algorithm, we have considered a one-dimensional isomerization model where the molecular potential is a double-minimum potential given by

$$V_{\mathrm{mol}}(q) = a(q^2 - 1)^2 + \frac{b}{4}(q - 1)^2 \tag{2.11}$$

and shown in the top panel of Fig.2.1, and the interactions of the molecule with the surroundings are modeled with Brownian dynamics. The parameters $a = 12$ kJ mol^{-1} and $b = 2$ kJ mol^{-1} correspond approximately to the height of the barrier separating the minima and to the energy difference between them, respectevely.

The system has been simulated at a temperature of 300 K. In a first reference simulation, the dynamics along the reaction coordinate q has been accelerated through standard metadynamics. A hill of height $h = 0.1$ kJ mol^{-1} and width $w = 0.05$ has been deposed along q every 100 time steps. The trajectory of the system is shown in the middle panel of Fig.2.1. After a transient period during which the history dependent potential fills the two minima, the system diffuses along the reaction coordinate and starts exploring a greater and greater region ($|q| > 1.5$) outside of the two basins.

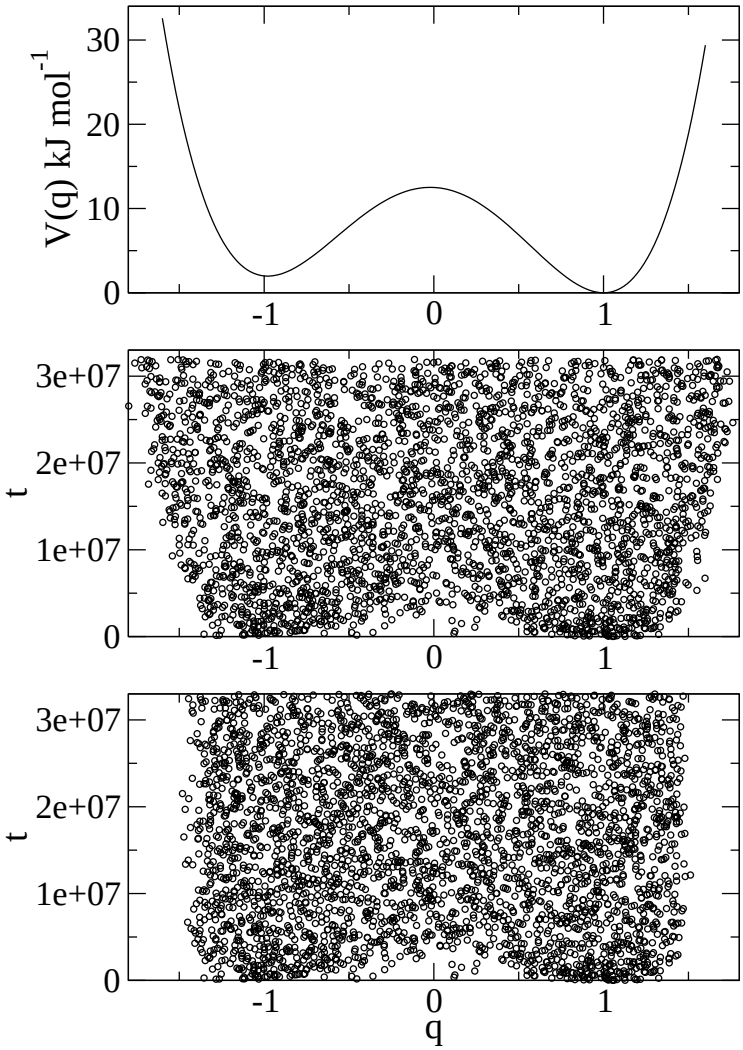

Figure 2.1: A trajectory from a standard metadynamics simulation on the model potential (middle panel) is compared to a metadynamics simulation with a constant number of hills (bottom panel); the model potential is shown in the top panel

In the second simulation, after this first transient period of $4 \cdot 10^5$ time steps, the number of hills has been fixed to 4000 by means of the algorithm presented in the previous section. As we can see in the lower panel of Fig.2.1, the system diffuses in a well-defined region that corresponds to the region of the two minima and the transitions state. In Fig.2.2 the energy profile computed by inverting the biasing potential at a random time during the simulation is compared with the exact potential. The reconstructed potential presents typical "bumps" of a metadynamics due to a finite deposition rate. If we average in time the biasing potential on a longer simulation, this unpleasent feature disappears (Fig.2.2).

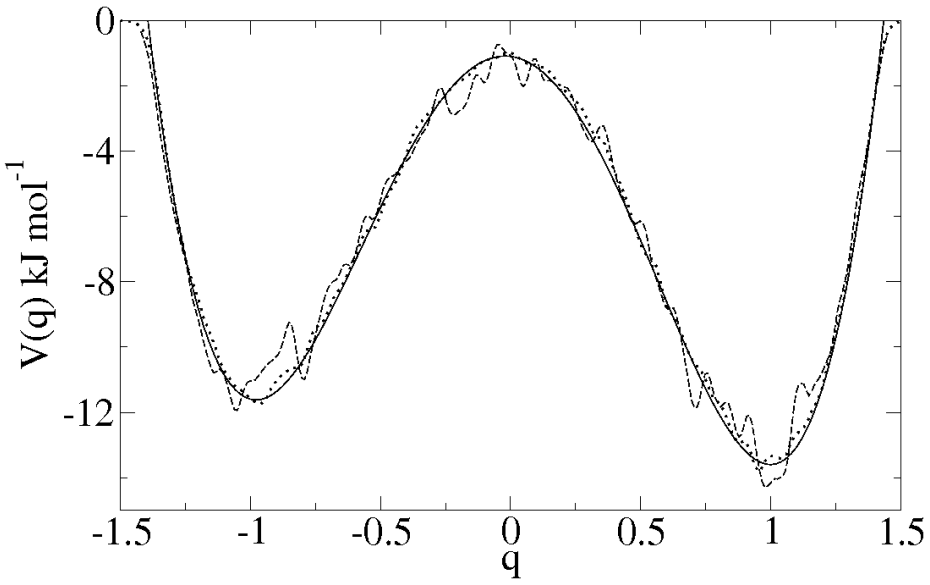

Figure 2.2: The model potential (straight line)is shown as a reference along with the potential reconstructed inverting in sign the history-dependent potential at a random time during the simulation (dashed line), and with the potential reconstructed by averaging over all the simulation time(dotted line)

It's remarkable that results from metadynamics simulations are typically computed over sets of different independent trajectories[12], and not as time averages along a single longer trajectory, since in standard metadynamics a run longer than the transient filling period will push the system in irrelevant regions of configurational space. However, in this section we have seen how to control the accessible configurational space during a metadynamics run, by using a Gillespie-like algorithm to appropriately move the deposed hills and fixing their number.

2.2 Self-Healing Umbrella Sampling

The ability of Molecular Dynamics (MD) simulations to study differences of entropy-related thermodynamic potentials of complex systems is strongly limited by the time needed to perform an ergodic sampling of the configurational space. In the original Umbrella Sampling (US) approach[56], an enhanced sampling of slow degrees of freedom is achieved by performing simulations in an artificial ensemble, obtained by adding an external potential V to the real Hamiltonian H. This potential has to be chosen so as to flatten the free energy surface (FES) along a selected multi-dimensional reaction coordinate $\mathbf{s}(\mathbf{r})$ ($\mathbf{r}$ is the vector of the $3N$ coordinates of the system), preventing the system from being trapped in a local minimum. The probability density for the unbiased system is recovered from the biased one through the relation[64]

$$\rho(\mathbf{s}) = \langle \delta(\mathbf{s} - \mathbf{s}(\mathbf{r})) \rangle = \frac{\langle \delta(\mathbf{s} - \mathbf{s}(\mathbf{r}))\ \mathrm{e}^{\beta V(\mathbf{s}(\mathbf{r}))} \rangle'}{\langle \mathrm{e}^{\beta V(\mathbf{s}(\mathbf{r}))} \rangle'} \tag{2.12}$$

where $\delta(...)$ is the Dirac function, $\beta = (k_B T)^{-1}$ with k_B being the Boltzmann constant. In Eq. 2.12 the primed angular brackets stand for a canonical average in the thermodynamic ensemble governed by the Hamiltonian

$$H' = H + V(\mathbf{s}(\mathbf{r})) \tag{2.13}$$

Ideally, in order to obtain an uniform sampling, one must choose a bias potential equal to the free energy inverted in sign, *i.e.* the very quantity we are trying to determine. A common solution to solve this circular problem is to perform a series of subsequent, quasi-equilibrium simulations as prescribed by the adaptive US method[65, 66, 67]. The bias potential is updated at the beginning of each simulation by matching the statistics resulting from all the previous runs. Recently, different approaches to reconstruct the FES self consistently have been proposed. These methods are based on a history-dependent bias potential (metadynamics[11]) or force (adaptive biasing force method[68, 69]) that is continuously varied during a single non equilibrium trajectory.

Inspired by the self-healing capabilities of the metadynamics of a non stationary probability distribution, the "inexact" non equilibrium nature of the adaptive US methodology is fully exploited, leading a parameter-free self consistent algorithm where improved estimates of the probability are determined "on the fly" with no need for *a posteriori* analysis for combining the statistics resulting from different bias potentials.

2.2.1 An history-dependent umbrella sampling algorithm

Consider a system in the canonical ensemble. Given a generic n-dimensional reaction coordinate $\mathbf{s}$ depending on the atomic coordinates (*e.g.*, a distance in a dissociation reaction or a dihedral angle in an isomerization process), the free energy $A(\mathbf{s})$ is defined in terms of the probability density of $\mathbf{s}$, as

$$A(\mathbf{s}) = -\beta^{-1} \ln \rho(\mathbf{s}). \tag{2.14}$$

If the ergodic hypothesis applies, $\rho(\mathbf{s})$, and hence $A(\mathbf{s})$, can be calculated by means of a time average over an equilibrium trajectory. In order to overcome the slow convergence of such average, we can generate a perturbed trajectory of the original system under the action of an external potential, providing that a relation is given to recover the correct statistics for the unperturbed system. In the case of an external potential $V(\mathbf{s})$ not explicitly dependent on time, as in the standard US method[56], this relation is Eq. 2.12. The natural choice for a history-dependent biased dynamics is to use a logarithmic relation between the time-dependent bias potential $V(\mathbf{s}, t)$ and some estimate of the real probability density $\rho(\mathbf{s})$ at time t, $\rho(\mathbf{s}, t)$

$$V(\mathbf{s}, t) = \beta^{-1} \ln \rho(\mathbf{s}, t) \tag{2.15}$$

where $\rho(\mathbf{s}, t)$ is a normalized function at each t, such that $0 < \rho(\mathbf{s}, t) < R$ for any $\mathbf{s}$ and any arbitrary value of R. This definition of bias potential automatically leads to a fast sampling of the reaction coordinate, exhorting the system to visit configurational states for which $\rho(\mathbf{s}, t)$ is small. In this case the dynamics of the system is governed by the time-dependent Hamiltonian

$$H' = H + \beta^{-1} \ln \rho(\mathbf{s}, t). \tag{2.16}$$

However, we have not yet exactly defined the function $\rho(\mathbf{s}, t)$. If there can be found a definition such that $\rho(\mathbf{s}, t)$, expressed as a time average, converges to the correct ensemble average for the probability density $\rho(\mathbf{s})$ in the long time limit, then $\rho(\mathbf{s}, t)$ can be taken as a correct estimate of $\rho(\mathbf{s})$. We can start by noticing that, in the hypothesis that the biased system is ergodic, the ensemble averages in Eq. 2.12 can be expressed as time integrals, such that

$$\rho(\mathbf{s}) = \lim_{t \to \infty} \frac{\int_0^t \delta[\mathbf{s} - \mathbf{s}(\tau)] \ \mathrm{e}^{\beta V(\mathbf{s}(\tau))} d\tau}{\int_0^t \mathrm{e}^{\beta V(\mathbf{s}(\tau))} d\tau}. \tag{2.17}$$

where the dynamics is driven by the Hamiltonian of Eq. 2.13. For the explicitly time dependent bias potential of Eq. 2.15, we define $\rho(\mathbf{s}, t)$ in a similar fashion, *i.e.*,

$$\rho(\mathbf{s}, t) = \frac{\int_0^t \delta[\mathbf{s} - \mathbf{s}(\tau)] \ \mathrm{e}^{\beta V(\mathbf{s}(\tau), \tau)} d\tau}{\int_0^t \mathrm{e}^{\beta V(\mathbf{s}(\tau), \tau)} d\tau} \tag{2.18}$$

where the dynamics is generated by the time-dependent Hamiltonian of Eq. 2.16. Substituting Eq. 2.15 into Eq. 2.18 we obtain a recursive relation for the probability density

$$\rho(\mathbf{s}, t) = \frac{\int_0^t \delta[\mathbf{s} - \mathbf{s}(\tau)]\ \rho(\mathbf{s}, \tau) d\tau}{\int_0^t \rho(\mathbf{s}(\tau), \tau)\ d\tau}. \tag{2.19}$$

Eq. 2.19 can be easily implemented in standard MD simulation programs with minor modifications. Starting from *any* initial arbitrary non zero density, it can be shown (see appendix 2.A) that the resulting non equilibrium dynamics automatically evolves to a stationary state where the bias potential nullifies the underlying free energy and the probability density converges to the exact solution. As in metadynamics, any kind of discrepancy between the biasing potential and the FES inverted in sign will be corrected by the subsequent dynamics. In the algorithm summarized by Eqs. 2.19 and 2.15, the evolution of the time dependent Hamiltonian stems exclusively from the dynamics of the system and *viceversa*. Therefore, the method does not involve system dependent parameters or corrections, reducing user intervention to a minimum.

2.2.2 SHUS simulations of alanine dipeptide isomerization reaction

In order to highlight the power and reliability of the algorithm, the FES using Eq. 2.19 of the solvated alanine dipeptide as a function of the dihedral angles Φ and Ψ was investigated.

The simulation of one dipeptide molecule and 288 water molecules was performed in the constant volume (cubic box of 21 Å side-length with standard periodic boundary conditions), constant temperature (300 K) thermodynamic ensemble using the program ORAC[70]. The temperature control was achieved using a Nosé-Hoover thermostat[36]. The dipeptide is modeled using the Amber03 force field[71]. For water we used the TIP3P potential[72]. Electrostatics has been accounted for by the smooth particle mesh Ewald method[73] using a fourth order B-spline interpolation polinomial for the charges, an Ewald α parameter of 0.43 $Å^{-1}$, and a grid spacing of $\sim$1 Å for the fast Fourier transform calculation of the charge weighted structure factor. A cutoff distance of 10 Å has been used for the nonbonded interactions. The reaction coordinate bin width is 5° for both the dihedral angles Φ and Ψ. The bias potential has been updated every 0.5 ps.

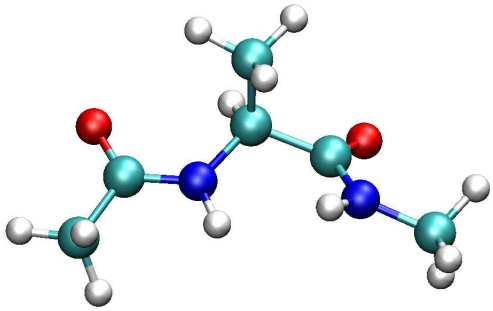

Figure 2.3: Representation of the alanine dipeptide molecule.

The FES of the solvated alanine dipeptide depending on Φ and Ψ has already been investigated in several computational studies (see Refs. [11, 71] and references therein). The FES of isomerization of the alanine dipeptide obtained from our methodology is reported in Figure 2.4 for two sampling simulation times, *i.e.* 1 and 10 ns. The reference FES obtained from standard US technique[56] is reported in

Figure S1 of Appendix B. After only 1 ns, the system has scanned the *entire* domain of the bi-dimensional reaction coordinate with a good accuracy. In particular the transition path between the C_{7eq} and α_R free energy minima[11] can be clearly seen. The three main free energy minima are located at $\Phi = -70°$, $\Psi = -20°$ (α_R), $\Phi = -70°$, $\Psi = 155°$ (C_{7eq}) and $\Phi = -155°, \Psi = 155°$ (C_5). Setting the free energy of the deeper minimum (C_{7eq}) as the zero point, the relative depth of the α_R minimum is ~ 0.5 kJ mol^{-1} and the transition state between C_{7eq} and α_R is located at $\Phi \simeq -80°$, $\Psi = 70°$ with activation energy of about 10 kJ mol^{-1}. These results agree with previous calculations obtained with the same force field[71], showing the balance between the extended and the helical FES regions of alanine dipeptide.

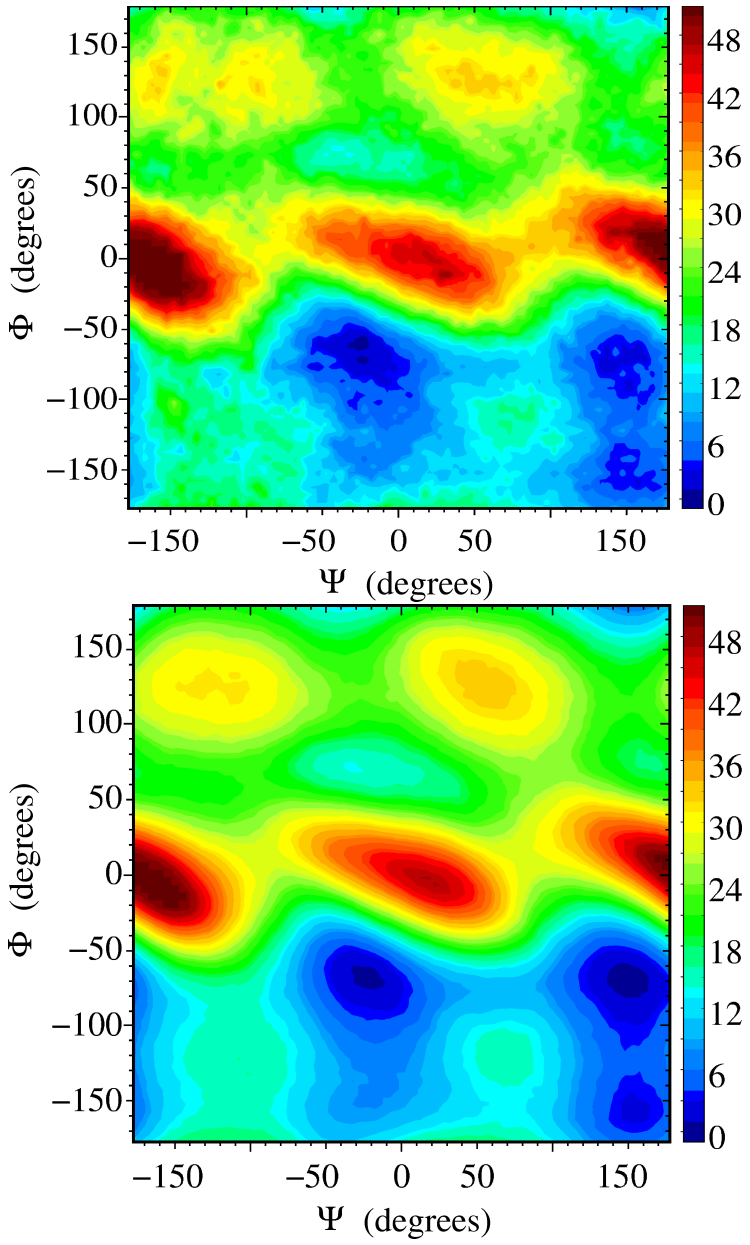

Figure 2.4: FES of the alanine dipeptide system as a function of Ψ and Φ torsional angles, estimated after a simulation time of 1 ns (top panel) and 10 ns (bottom panel). The free energy scale is in kJ mol^{-1}. The zero free energy is set in the absolute minimum of each surface. The reference FES (obtained by standard US) is shown in Appendix B

The convergence of the algorithm can be appreciated in Figure 2.5, where the evolution of the root mean square deviation of the calculated FES with respect to the reference one is reported. After 1 ns of simulation (see FES in the top panel of Figure 2.4) the average error is less than 2 kJ mol^{-1}. After 10 ns of simulation (see FES in the bottom panel of Figure 2.4) the average error is as small as 0.5 kJ mol^{-1}.

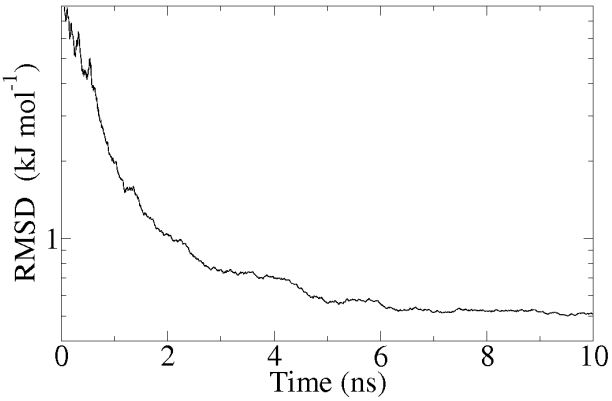

Figure 2.5: Root mean square deviation of the estimated FES of the alanine dipeptide from the reference FES (see appendix 2.B.

In conclusion, this method utilizes, in the spirit of the adaptive US techniques, a history-dependent bias potential that is progressively updated in order to flatten the FES, eventually leading to an uniform sampling along the chosen reaction coordinate. The novelty of such an approach with respect to standard US methods is the introduction of a history-dependent bias potential that is *continuously* varied during a single simulation on the basis of "on the fly" evaluations of the probability density function. The non-equilibrium probability density of the biased sampling is indeed used to obtain, virtually at every step of the simulation, a new estimate of the FES thus allowing a self-healing updating of the bias potential as the simulation proceeds. Our non-equilibrium approach avoids altogether the problem of connecting statistics collected in independent equilibrium simulations, resulting in a parameter-free, general, and highly efficient self consistent algorithm.

2.A Proof of convergence for the SHUS algorithm

In order to evaluate equation

$$\rho(\mathbf{s},t) = \frac{\int_0^t \delta[\mathbf{s}-\mathbf{s}(\tau)]\ \rho(\mathbf{s},\tau)d\tau}{\int_0^t \rho(\mathbf{s}(\tau),\tau)d\tau}. \tag{2.20}$$

in a numerical simulation, we have to discretize the n components of the reaction coordinate $\mathbf{s} \equiv (s_1, s_2, ..., s_i, ..., s_n)$ into a number of intervals or bins through the relations $s_i = b_i \Delta s_i$, where Δs_i and b_i are the step size and the bin index related to the component s_i, respectively. The range spanned by s_i is limited, *i.e.* $b_i^{(min)} \leq b_i \leq b_i^{(max)}$. In this picture the running (simulation) time t is also a discrete quantity that can be represented as $t = j\Delta t$, where Δt is the simulation time step and j is the simulation step index ($j = 0, 1, 2, ..., N_{step}$). In the following treatment, the dependence of $\rho(\mathbf{s},t)$ on t is expressed by the step index j, while the dependence on $\mathbf{s}$ is expressed by the vector $\mathbf{b} \equiv (b_1, b_2, ..., b_n)$ whose components are the bin indexes of the s_i's. We then define the adimensional quantity $\rho[\mathbf{b}, j]$ as a discrete approximation of $\rho(\mathbf{s},t)\Delta\mathbf{s}$ on the connected and bounded domain $\mathcal{S}$

$$\rho[\mathbf{b}, j] = \frac{\sum_{l=0}^{j} \delta_{\mathbf{b},\mathbf{b}(l)}\ \rho[\mathbf{b}, l-1]}{\sum_{\mathbf{m}} \sum_{l=0}^{j} \delta_{\mathbf{m},\mathbf{m}(l)}\ \rho[\mathbf{m}, l-1]} \tag{2.21}$$

where $\delta_{\mathbf{b},\mathbf{b}(j)}$ is a Kronecker delta and $\mathbf{b}(j)$ identifies the vector $\mathbf{b}$ pointing to the n-dimensional bin sampled at the jth time step. With this definition, the quantity $\rho[\mathbf{b}, j]$ is supported on the bounded domain $\mathcal{S}$ and normalized at each time j. As $\rho[\mathbf{b}, j]$ is defined by a recursive relation (Eq. 2.21), we need an initial (arbitrary) distribution to start up the algorithm. To this end, we assume $\delta_{\mathbf{b},\mathbf{b}(0)} = c$ and $\rho[\mathbf{b}, -1] = 1$ for all $\mathbf{b}$ vectors spanning the domain $\mathcal{S}$ with c being an arbitrary constant. This choice amounts to start with a uniform distribution on $\mathcal{S}$. Given the above initial conditions and given that $\sum_{\mathbf{b}} \rho[\mathbf{b}, j] = 1$, it can be trivially shown that the relation

$$0 < \rho[\mathbf{b}, j] < 1 \tag{2.22}$$

holds for all times j and all $\mathbf{b}$ vectors spanning the $\mathcal{S}$ domain. We now show that the quantity $\rho[\mathbf{b}, j]$ defined above converges to a time independent solution for all $\mathbf{b}$, namely in the domain of definition of $\mathbf{s}$. To this aim it is sufficient to show that

$$\lim_{j\to\infty} \rho[\mathbf{b}, j+J] = \lim_{j\to\infty} \rho[\mathbf{b}, j] \tag{2.23}$$

for any arbitrarly large J. From Eq. 2.21 we get the following equation

$$\rho[\mathbf{b}, j+J] = \frac{F[\mathbf{b},j]\left(1+\frac{\sum_{l=j+1}^{j+J}\delta_{\mathbf{b},\mathbf{b}(l)}\ \rho[\mathbf{b},l-1]}{F[\mathbf{b},j]}\right)}{G[j]\left(1+\frac{\sum_{\mathbf{m}}\sum_{l=j+1}^{j+J}\delta_{\mathbf{m},\mathbf{m}(l)}\rho[\mathbf{m},l-1]}{G[j]}\right)} \tag{2.24}$$

where we have defined

$$F[\mathbf{b}, j] = \sum_{l=1}^{j} \delta_{\mathbf{b},\mathbf{b}(l)}\ \rho[\mathbf{b}, l-1] \tag{2.25}$$

and

$$G[j] = \sum_{\mathbf{b}} F[\mathbf{b}, j]. \tag{2.26}$$

Exploiting the condition of Eq. 2.22 and the definitions of Eqs. 2.25 and 2.26, it can be shown that

$$\begin{aligned} \lim_{j\to\infty} \frac{\sum_{l=j+1}^{j+J}\delta_{\mathbf{b},\mathbf{b}(l)}\ \rho[\mathbf{b},l-1]}{F[\mathbf{b},j]} &= 0 \\ \lim_{j\to\infty} \frac{\sum_{\mathbf{m}}\sum_{l=j+1}^{j+J}\delta_{\mathbf{m},\mathbf{m}(l)}\ \rho[\mathbf{m},l-1]}{G[j]} &= 0. \end{aligned} \tag{2.27}$$

Taking the limit $j \to \infty$ of Eq. 2.24 and using Eqs. 2.25, 2.26 and 2.27, we recover Eq. 2.23.

The next step is to show that the time independent solution of our problem, *i.e.* $\lim_{j\to\infty}\rho[\mathbf{b}, j]$, gives exactly the probability density of the reaction coordinate $\mathbf{s}$. This can be proved by showing that, in the limit of large simulation times (large j), the biased distribution function $f[\mathbf{b}, j, J]$ calculated in the time interval from j to $j+J$ for arbitrary large J is uniform on the $\mathcal{S}$ domain. $f[\mathbf{b}, j, J]$ is defined as follows

$$f[\mathbf{b}, j, J] = \frac{\sum_{l=j}^{j+J}\delta_{\mathbf{b},\mathbf{b}(l)}}{\sum_{\mathbf{m}}\sum_{l=j}^{j+J}\delta_{\mathbf{m},\mathbf{m}(l)}}. \tag{2.28}$$

From the definition of $\rho[\mathbf{b}, j]$ (Eq. 2.21) and the definition of $G[j]$ (Eq. 2.26), one obtains the following relation for $\delta_{\mathbf{b},\mathbf{b}(j)}$:

$$\delta_{\mathbf{b},\mathbf{b}(j)} = \frac{\rho[\mathbf{b}, j]}{\rho[\mathbf{b}, j-1]}\ G[j] - G[j-1]. \tag{2.29}$$

Substituting the previous equation into Eq. 2.28, we get

$$f[\mathbf{b}, j, J] = \frac{\sum_{l=j}^{j+J}\left(\frac{\rho[\mathbf{b},l]}{\rho[\mathbf{b},l-1]}G[l] - G[l-1]\right)}{\sum_{\mathbf{m}}\sum_{l=j}^{j+J}\left(\frac{\rho[\mathbf{m},l]}{\rho[\mathbf{m},l-1]}G[l] - G[l-1]\right)} \tag{2.30}$$

Using the relation demontrated above (Eq. 2.23) and provided that $\lim_{j\to\infty} \rho[\mathbf{b}, j] \neq 0$, for all $\mathbf{b}$ in $\mathcal{S}$, we obtain

$$\lim_{j\to\infty} f[\mathbf{b}, j, J] = \frac{\sum_{l=j}^{j+J} (G[l] - G[l-1])}{\sum_{\mathbf{m}} \sum_{l=j}^{j+J} (G[l] - G[l-1])} = \frac{1}{N} \tag{2.31}$$

where N is the number of bins. Therefore, *provided that* $\lim_{j\to\infty} \rho[\mathbf{b}, j] \neq 0$ *for all* $\mathbf{b}$ *in* $\mathcal{S}$, the limiting sampling of the reaction coordinate is uniform. This means that, in the long time limit, the external potential $V(\mathbf{s}, t)$ generated by the function $\rho(\mathbf{s}, t)$ nullifies the underlying free energy $A(\mathbf{s})$, *i.e.* $\lim_{t\to\infty} V(\mathbf{s}, t) = -A(\mathbf{s})$.

2.B Reference free energy surface for the alanine dipeptide isomerization reaction

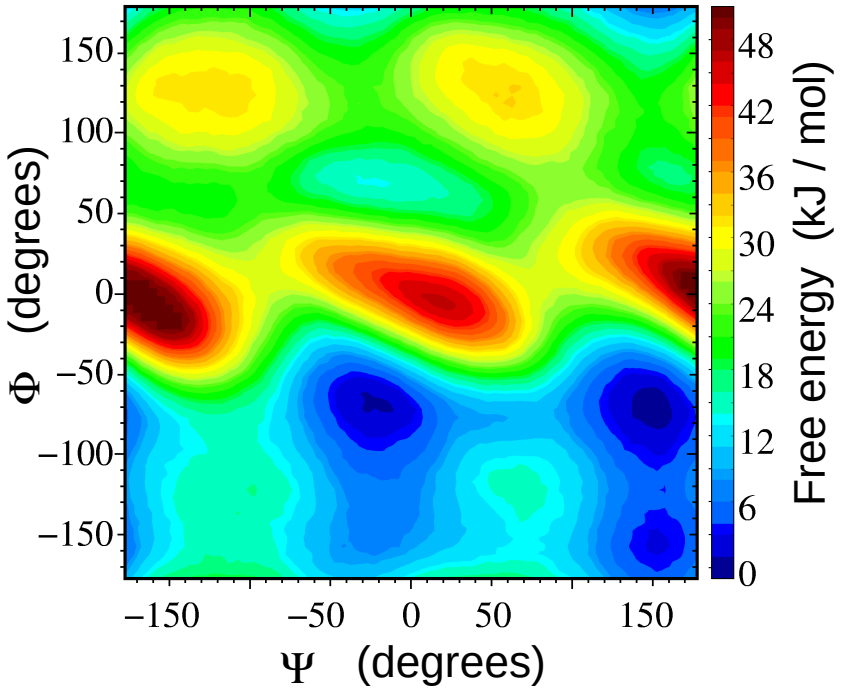

Figure 2.6: Reference free energy surface for the alanine dipeptide, obtained from standard US technique

2.C The optimal shape of the hills

In its standard implementation, the history-depnendent potential of metadynamics is given by a sum of Gaussian functions. Some variants have been introduced, with the intent of improving the accuracy or the efficiency of the method[74, 61]. In this section we present Lucy's function[75, 76] (LF) as a very efficient alternative to the use of Gaussians.

The LF is the *simplest* function satisfying the following four conditions:
(i) it is normalizable
(ii) it has a finite range w
(iii) it has a maximum at the origin
(iv) it has $n-1$ continuous derivative everywhere
It is defined for a generic order n as

$$f(x) = \begin{cases} h\left(1+2\frac{|x|}{w}\right)\left(1-\frac{|x|}{w}\right)^n & \text{if } -w \le x \le w \\ 0 & \text{if } x < w, x > w \end{cases} \tag{2.32}$$

whith the origin at $x = 0$. The symbols h and w denote the height and the width, respectevely. Since this function will play the role of a potential, we need one continuous derivative only. The explicit form of the LF and its derivative for $n = 2$ is given by:

$$f(-w \le x \le w) = h\left(1+2\frac{|x|}{w}\right)\left(1-\frac{|x|}{w}\right)^2 \tag{2.33}$$

$$\frac{\partial f(x)}{\partial x} = \frac{6h}{w^3}x(|x|-w). \tag{2.34}$$

Such a simple derivative is particularly fit for the computation of the history dependent forces during a metadynamics run. Moreover, since LF has a finite range by definition, it does not need to be smoothly truncated[61], and the contribution to the forces from hills farther than the width w can be correctly neglected. A LF with $h = w = 1$ and a Gaussian function with the same height and standard deviation $\sigma = 1/3$ are compared in Fig.2.7.

For N-dimensional spaces, one can use the product of monodimensional LFs as a N-dimensional hill of height h:

$$f_N(\mathbf{x}) = (1/h)^{N-1}\prod_{i=1}^{N} f_i(x_i) \tag{2.35}$$

$$\frac{\partial f_N(\mathbf{x})}{\partial x_i} = (1/h)^{N-1}\frac{\partial f(x_i)}{\partial x_i}\prod_{j\neq i}^{N} f_j(x_j) \tag{2.36}$$

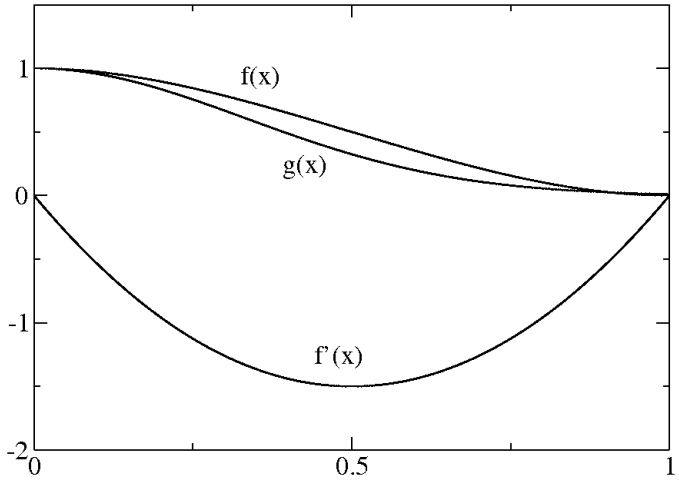

Figure 2.7: Lucy's weight function f with $h = w = 1$, along with its first derivative f' and a Gaussian function g with the same height and $\sigma = 1/3$.

In more than one dimension, the contribution to the forces from a given hill vanishes if at least one component of the vector going from the origin of the hill to the actual position of the system is greater in absolute value than the width of the LF along that CV.

3

From instantaneous to reversible replica exchanges

One of the most important problems of scientific computing is large-scale parallelization of algorithms. Traditional parallelization schemes require extremely fast communication between multiple processors and frequently do not scale to large numbers of processors. As an alternative, it is possible to *statistically* couple many simulations run in parallel to obtain a result that is equivalent to a longer simulation. For example, the Replica Exchange Method[14, 15, 16, 17] (REM) consists in performing a series of indipendent simulations of the same system, each in different equilibrium conditions. In its most common implementation, simulations differ in temperature: at variance with Non-Boltzmann Sampling and history-dependent methods, where an energy barrier between two metastable states is flattened, the transition rate is enhanced using an higher temperature. Then, to transfer the barrier-crossing efficiency from runs at high temperature to the target temperature, temperature exchanges are attempted periodically, moving through a series of exchanges high-temperature configurations to the lowest temperature simulation.

The temperature exchanges, or replica exchanges, are the crux of the algorithm. If two simulations at temperatures T_j and T_i have energies E_j and E_i, they exchange their temperatures with the Metropolis probability

$$p(\text{acc}) = \min\left\{1, e^{\Delta\beta\Delta E}\right\} \tag{3.1}$$

where $\Delta\beta = \beta_j - \beta_i = (k_B T_j)^{-1} - (k_B T_i)^{-1}$ is their inverse temperature difference and $\Delta E = E_j - E_i$ is their energy difference. This simple recipe guarantees a

correct equilibrium sampling for each simulation at each temperature. If we rewrite the acceptance probability as

$$p(\text{acc}) = \min\left\{1, e^{\Delta\beta E_j - \Delta\beta E_i}\right\} \tag{3.2}$$

and interpret $\Delta\beta$ as a displacement in the space of temperature, and the energy of a replica as a conjugated force, then the argument of the exponential function have the physical meaning of a sum two "works": the work spent in moving a replica with energy E_i from a temperature β_i to a temperature β_j and a replica with energy E_j from a temperature β_j to a temperature β_i

$$p(\text{acc}) = \min\left\{1, e^{-(W_{i\to j} + W_{j\to i})}\right\}. \tag{3.3}$$

This argument shows that the REM acceptance probability favours low-dissipative exchanges, and suggests a strong connection between this algorithm and the work fluctuation theorem[5]. In this chapter, starting from this connection the REM algorithm is extended from instantaneous exchanges to general non-equilibrium transformations, in which the temperatures (or more generally the thermodynamical conditions) of two replicas are exchanged with an arbitrary protocol.

3.1 The Replica Exchange Method and the work fluctuation theorem

In a computer simulation, a quasi-ergodic behavior occurs when the simulated system gets trapped into one or several basins of the energy surface, giving rise to biased statistics. In general, such a problem arises from the difference in time scale between the affordable observation time and the long characteristic time of transition among different important low-energy regions. "Broken ergodicity" is characteristic of complex systems with rugged potential energy landscapes, like spin glasses, atomic clusters and biomolecules. The Replica Exchange Method (REM) [14, 15, 16, 17] provides an elegant and simple solution to quasi-ergodic sampling. In REM, several independent trajectories, called replicas, are simultaneously generated in different thermodynamic conditions. Usually, these conditions are chosen so as to span homogeneously the thermodynamic space from the ensemble of interest to a different ensemble with enhanced transition rates, where the sampling is ergodic. During the simulation, neighboring replicas are allowed to exchange their ensemble, subject to specific acceptance criteria. In this fashion, a trajectory is no longer bound to an unique given equilibrium ensemble but can randomly walk in a thermodynamic space of different equilibrium conditions, visiting ensembles where an ergodic sampling is possible, and then going back to the quasi-ergodic ensemble of interest. Therefore, REM is an algorithm which employs an extended ensemble formalism in order to overcome slow relaxation. The gain in sampling efficiency with respect to a series of uncoupled parallel trajectories comes from the exchange of information between trajectories, and the replica exchange process is the tool by which "information" (e.g. a particular configuration) is carried, for example, from an high to a low temperature.

With the aim of developing a novel and more general formalism for REM, in this letter we highlight the non equilibrium character of a replica exchange process. In recent years, some key relations concerning the statistical mechanics of non equilibrium processes have been derived [45, 77, 54, 55, 4, 5, 78]. Anticipating our results, we show that a work fluctuation theorem is valid in the context of a replica exchange. More precisely, interpreting an exchange as an externally driven process, the work spent in performing an exchange will be shown to obey exactly to a fluctuation symmetry relation,

$$\frac{P(W_{xc} = w)}{P(W_{xc} = -w)} = \exp(\beta w). \tag{3.4}$$

The above equation quantifies the probability ratio of dissipating a work w and $-w$ while exchanging the thermodynamic states of two replicas both starting from equi-

librium conditions. While the above equation can be straightforwardly derived for a standard REM simulation, where replica exchanges are instantaneous, we demonstrate its validity for exchanges performed *in an arbitrary duration time* τ. We also show that relation 3.4 implies that, in a REM simulation, an attempted replica exchanges must be accepted with probability

$$\min\{1, \exp(-\beta W_{xc})\} \tag{3.5}$$

in order to satisfy detailed balance and preserve the equilibrium distribution of the extended system of replicas. The above rule is shown to coincide with the ordinary REM acceptance probability for instantaneous exchanges.

This result expands the physical meaning of the "replica exchange" from the instantaneous switches entailing maximum dissipation of the standard algorithm up to ideal non dissipative reversible exchange having unitary acceptance probability.

Consider two replicas evolving in the canonical ensemble. The dynamics of the replicas are supposed to be stochastic and Markovian. For simplicity and with no loss of generality, we will consider the Hamiltonian REM algorithm[79], in which the two replicas have the same temperature β but differ in the energy function. We define an energy function $E(x, \lambda)$ where x is a microstate of the system and λ is an externally driven parameter. Each of the two replicas refers to a different value of the λ parameter, $\lambda = a$ for one replica and $\lambda = b$ for the other replica. Here, λ is a state parameter, since to each λ value corresponds a unique equilibrium distribution in the space of the microstates, $P(x, \lambda) = Z(\lambda)^{-1}\exp(-\beta \mathrm{E(x, \lambda)})$, where the normalization constant $Z(\lambda)$ is the system partition function for a given λ. The free energy ΔA between states a and b is then given by $\Delta A = -\beta^{-1}\ln[Z(b)/Z(a)]$. An extended microstate of the collection of replicas is indicated by the vector $\mathbf{x} = \{x_1, x_2\}$ where x_1 is the microstate for the replica in state a and x_2 is the microstate in state b. In this notation, the probability of a configuration $\mathbf{x}$ of the extended system of replicas is given by the joint probability $P(\mathbf{x}) = P(x_1, a)P(x_2, b)$. In a Markov chain in the ensemble of microstates of the extended system, there are two transition process that leaves the distribution $P(\mathbf{x})$ invariant: i) a conventional transition scheme applied independently on each replica and obeying the detailed balance condition for corresponding equilibrium distribution, and ii) the replica exchange process. In this latter scheme, the values of the λ parameter for the two replicas are instantaneously switched. In Hamiltonian REM, this corresponds to instantaneously switch the two energy functions of the replicas. An exchange is then accepted with probability

$$\min\{1, \exp\left(-\beta(\Delta E_{a\to b} + \Delta E_{b\to a})\right)\} \tag{3.6}$$

where $\Delta E_{a\to b} = E(x_1, b) - E(x_1, a)$ and $\Delta E_{b\to a} = E(x_2, a) - E(x_2, b)$ are the energy changes for the two replicas following from the exchange process.

From a physical point of view, an exchange should be seen as an externally driven process, in which we externally switch the energy functions of two replicas. Since the latter are non interacting, the two processes are independent and the total work done in exchanging replicas is simply given by the sum of the works

$$W_{xc} = W_{a\to b} + W_{b\to a} \tag{3.7}$$

where W_{xc} is the work spent on the ensemble of two replicas and $W_{a\to b}$ and $W_{b\to a}$ are the work spent to change the state of one replica from $\lambda = a$ to $\lambda = b$ and the state of the other replica from $\lambda = b$ to state $\lambda = a$, respectively. Moreover, since an exchange process leaves the extended ensemble distribution invariant, the total work W_{xc} coincides with the work dissipated during the exchange process. In the standard REM, the exchange process is instantaneous, *i.e.*, adiabatic, and the work spent for each replicas equals the energy change

$$W_{a\to b} = \Delta E_{a\to b} \qquad W_{b\to a} = \Delta E_{b\to a}. \tag{3.8}$$

Using Eqs. 3.7 and 3.8 into Eq. 3.6, the acceptance probability for an Hamiltonian REM move can therefore be rewritten as a function of the dissipated work W_{xc}, i.e.

$$\min\left\{1, \exp\left(-\beta W_{xc}\right)\right\}. \tag{3.9}$$

We have seen how the REM algorithm, in its standard scheme, exploits instantaneous driven transformations to generate configurations in a space of different thermodynamic conditions. It is now natural to ask whether one can devise a REM algorithm in which the states of two replicas are exchanged in an arbitrary duration time τ, rather than instantaneously. Such an exchange can be performed by switching with a given protocol $\lambda_{a\to b,t}$ the state of one replica from $\lambda_{a\to b}(0) = a$ to $\lambda_{a\to b}(\tau) = b$, and changing similarly the state of the other replica with a protocol $\lambda_{b\to a,t}$ from b at time 0 to a at time τ. Given the energy function $E(x, \lambda)$, the works performed on each replica are given by the integrals

$$W_{a\to b} = \int_0^\tau \dot{\lambda}_{a\to b}\partial_\lambda E dt \quad W_{a\to b} = \int_0^\tau \dot{\lambda}_{b\to a}\partial_\lambda E dt. \tag{3.10}$$

Again, the total work spent for an exchange process is given by $W_{xc} = W_{a\to b} + W_{b\to a}$. However, at variance with the instantaneous exchange case, Eq. 3.7, the work is now a *functional* of the paths sampled from the initial states x_1 and x_2 in the space of

the microstates x during the process. If we denote the path for the $a \to b$ process with $x_{\alpha,t}$ and for the $b \to a$ process with $x_{\beta,t}$, this can be expressed by writing

$$W_{xc}[x_{\alpha,t}, x_{\beta,t}] = W_{a\to b}[x_{\alpha,t}] + W_{b\to a}[x_{\beta,t}]. \tag{3.11}$$

The Crooks work theorem[5] is of special interest in this context. Given two states of a system whose evolution is Markovian, this theorem relates the probability of observing a path in state space in a driven process to the probability of observing the reversed path when the system is driven according to a reverse time schedule, both processes being started from equilibrium conditions:

$$\frac{P_{\lambda_t}[x_t]}{P_{\bar{\lambda}_t}[\bar{x}_t]} = \exp\left(\beta(W[x_t] - \Delta A)\right). \tag{3.12}$$

Here, ΔA is the canonical free energy difference between the initial and final states at $\lambda(0) = a$ and at $\lambda(\tau) = b$, λ_t and $\bar{\lambda}_t$ are the forward and reversed time schedules of the externally driven state parameter and x_t and $\bar{x}_t$ are the resulting forward and reversed paths of the system.

As seen, a replica exchange is composed of two independent, opposite transformations of the states of two replicas. We introduce a more compact notation for handling pairs of paths of a two-replica system, by looking to the ensemble of replicas as an extended system, and using a vectorial notation. Supposing to start from an extended microstate $\mathbf{x} = \{x_1, x_2\}$, we denote with the symbol $\mathbf{x}_t = \{x_{\alpha,t}, x_{\beta,t}\}$ an extended path. $x_{\alpha,t}$ is the path associated with the $a \to b$ transformation, starting from $x_\alpha(0) = x_1$ and ending in $x_\alpha(\tau) = x_1'$, obtained with a switching protocol $\lambda_{a\to b,t}$, and $x_{\beta,t}$ is the path associated with the $b \to a$ transformation, starting from $x_\beta(0) = x_2$ and ending in $x_\beta(\tau) = x_2'$ and obtained with the protocol $\lambda_{b\to a,t}$. The final extended microstate is therefore given by $\mathbf{x}' = \{x_2', x_1'\}$, where replicas are reordered accordingly to their λ value. We assume that replicas are exchanged with time reversed protocols, such that $\lambda_{b\to a,t} = \bar{\lambda}_{a\to b,t}$, and drop the subscripts such that $\lambda_t \equiv \lambda_{a\to b,t}$ and $\bar{\lambda}_t \equiv \lambda_{b\to a,t}$. The reversed exchange can be constructed by inverting both the processes: one obtains the extended path $\bar{\mathbf{x}}_t = \{\bar{x}_{\beta,t}, \bar{x}_{\alpha,t}\}$, starting from $\mathbf{x}'$ and ending in $\mathbf{x}$, where the replica path $\bar{x}_{\alpha,t} \equiv \bar{x}_\alpha(t) = x_\alpha(\tau - t)$ is sampled while changing the state of the system from b to a with a protocol $\bar{\lambda}_t$ and the path $\bar{x}_{\beta,t} \equiv \bar{x}_\beta(t) = x_\beta(\tau - t)$ is sampled during a $a \to b$ transformation with a protocol λ_t.

The Crooks theorem, Eq. 3.12 can now be exploited to quantify the ratio between the probability of observing an extended path $\mathbf{x}_t$, $P[\mathbf{x}_t]$, and the probability of observing the time reversal extended path $\bar{\mathbf{x}}_t$, $P[\bar{\mathbf{x}}_t]$, while exchanging the states of

the replicas. Such probabilities are defined as

$$P[\mathbf{x}_t] = P_{\lambda_t}[x_{\alpha,t}]P_{\bar{\lambda}_t}[x_{\beta,t}] \tag{3.13}$$

$$P[\bar{\mathbf{x}}_t] = P_{\lambda_t}[\bar{x}_{\beta,t}]P_{\bar{\lambda}_t}[\bar{x}_{\alpha,t}]. \tag{3.14}$$

It is worthwhile to note that the forward and reversed exchanges share the same overall exchange protocol, that is, the same pair of protocols λ_t and $\bar{\lambda}_t$.

Using relation 3.12 we can write the pair of equations

$$\frac{P_{\lambda_t}[x_{\alpha,t}]}{P_{\bar{\lambda}_t}[\bar{x}_{\alpha,t}]} = \exp\left(\beta(W[x_{\alpha,t}] - \Delta A)\right) \tag{3.15}$$

$$\frac{P_{\bar{\lambda}_t}[x_{\beta,t}]}{P_{\lambda_t}[\bar{x}_{\beta,t}]} = \exp\left(\beta(W[x_{\beta,t}] + \Delta A)\right) \tag{3.16}$$

since $\Delta A = \Delta A_{ab} = -\Delta A_{ba}$. Therefore, multiplying Eq. 3.15 and Eq. 3.16, and using Eqs. 3.13 and Eq. 3.14, we can write an equation that correlates the probability of observing an extended path while performing a replica exchange to the probability of observing its reversed extended path

$$\frac{P[\mathbf{x}_t]}{P[\bar{\mathbf{x}}_t]} = \exp\left(\beta W_{xc}[\mathbf{x}_t]\right) \tag{3.17}$$

where again $W_{xc}[\mathbf{x}_t] = W_{a\to b}[x_{\alpha,t}] + W_{b\to a}[x_{\beta,t}]$. Summing over all extended paths $\mathbf{x}_t$ that yield the the work $W_{xc} = w$, we recover equation 3.4, that quantifies the probability of a second law violation in a replica exchange process, that is, the probability of observing a work value $w < 0$, as in this case, $W_{rev} = \Delta A_a \to b + \Delta A_b \to a = 0$.

We now finally show that a replica exchange performed with an arbitrary switching time satisfies the detailed balance condition if the acceptance probability is given by equation 3.5. The algorithm, given an initial extended state $\mathbf{x}$, will propose a new extended state $\mathbf{x}'$ switching the state of replicas in a time τ. The protocol λ_t for a $a \to b$ transformation is established at the beginning of the simulation, and, again, we assume that the protocol for the $b \to a$ transformation is reversed in time, $\bar{\lambda}_t$. Detailed balance in the extended system of replicas is expressed by

$$P(\mathbf{x})T(\mathbf{x}'|\mathbf{x}) = P(\mathbf{x}')T(\mathbf{x}|\mathbf{x}') \tag{3.18}$$

where $T(\mathbf{x}|\mathbf{x}')$ and $T(\mathbf{x}'|\mathbf{x})$ are the transition rates between the states $\mathbf{x}$, $\mathbf{x}'$ and $\mathbf{x}'$, $\mathbf{x}$, respectively. Here, the extended microstates $\mathbf{x} = \{x_1, x_2\}$ and $\mathbf{x}' = \{x_2', x_1'\}$ are connected by two sets of extended paths. We define the set $\mathcal{X}$ whose members are all the possible extended paths $\mathbf{x}_t$ going from $\mathbf{x}$ to $\mathbf{x}'$, and the set $\mathcal{X}'$ of all the extended

paths $\mathbf{x}'_t$ going from $\mathbf{x}'$ to $\mathbf{x}$. The transition rate $T(\mathbf{x}'|\mathbf{x})$ can be rewritten as a sum of separated transition rates over all possible paths

$$T(\mathbf{x}'|\mathbf{x}) = \sum_{\mathbf{x}_t \in \mathcal{X}} T[\mathbf{x}_t|\mathbf{x}] \tag{3.19}$$

where $T[\mathbf{x}_t|\mathbf{x}]$ is the probability that, starting from $\mathbf{x}$, the replicas will exchange their states *following the particular path* $\mathbf{x}_t$. The detailed balance condition, Eq. 3.18, can therefore be rewritten as

$$P(\mathbf{x}) \sum_{\mathbf{x}_t \in \mathcal{X}} T[\mathbf{x}_t|\mathbf{x}] = P(\mathbf{x}') \sum_{\mathbf{x}'_t \in \mathcal{X}'} T[\mathbf{x}'_t|\mathbf{x}']. \tag{3.20}$$

Given an extended path $\mathbf{x}_t$, the conjugated time reversed path $\bar{\mathbf{x}}_t$ can be found by time reversing independently each replica path: if $\mathbf{x}_t = \{x_{1,t}, x_{2,t}\}$, then $\bar{\mathbf{x}}_t = \{\bar{x_{2t}}, \bar{x_{1t}}\}$. Is is easy to see that the members of the set $\mathcal{X}'$ can be found by reversing the members of the set $\mathcal{X}$, and *viceversa*, that is, there is a one to one correspondence between the terms of the sums of equation 3.20. Therefore, the detailed balance condition expressed by equation 3.18 is satisfied by exploiting this one to one correspondence, and imposing the following detailed balance over each pair of conjugated extended paths

$$P(\mathbf{x})T[\mathbf{x}_t|\mathbf{x}] = P(\mathbf{x}')T[\bar{\mathbf{x}}_t|\mathbf{x}']. \tag{3.21}$$

As is commonly done in Monte Carlo schemes[80, 81], it is useful to express the transition rate as

$$T[\mathbf{x}_t|\mathbf{x}] = P[\mathbf{x}_t|\mathbf{x}]\mathrm{acc}[\mathbf{x}_t|\mathbf{x}] \tag{3.22}$$

where the proposal function $P[\mathbf{x}_t|\mathbf{x}]$ corresponds to the conditional probability, given $\mathbf{x}$, that the algorithm will generate the extended path $\mathbf{x}_t$, and $\mathrm{acc}[\mathbf{x}_t|\mathbf{x}]$ is the probability of accepting such an exchange. Using the definition Eq. 3.22, the detailed balance condition, Eq. 3.21, can be expressed as

$$\frac{\mathrm{acc}[\mathbf{x}_t|\mathbf{x}]}{\mathrm{acc}[\bar{\mathbf{x}}_t|\mathbf{x}']} = \frac{P[\bar{\mathbf{x}}_t|\mathbf{x}']}{P[\mathbf{x}_t|\mathbf{x}]} \frac{P(\mathbf{x}')}{P(\mathbf{x})}. \tag{3.23}$$

Using the fact that $P[\mathbf{x}_t] = P(\mathbf{x})P[\mathbf{x}_t|\mathbf{x}]$ and $P[\bar{\mathbf{x}}_t] = P(\mathbf{x}')P[\bar{\mathbf{x}}_t|\mathbf{x}']$ and exploiting Eq. 3.17, we finally find that

$$\frac{\mathrm{acc}[\mathbf{x}_t|\mathbf{x}]}{\mathrm{acc}[\bar{\mathbf{x}}_t|\mathbf{x}']} = \exp\left(-\beta W_{xc}[\mathbf{x}_t]\right). \tag{3.24}$$

such that a replica exchange process performed in an *arbitrary time duration* τ is accepted with probability given by Eq. 3.5. Clearly, for $\tau = \infty$, i.e. for infinitely

slow (reversible) exchange processes $W_{a\to b}[x_{\alpha,t}] = \Delta A = -W_{b\to a}[x_{\beta,t}]$ such that $W_{xc}[\mathbf{x}_t] = 0$ and the exchanges are *always* accepted.

Although we restricted the analysis to replicas in the canonical ensemble that refer to different energy functions, these results can be straightforwardly generalized. The Crooks relation can easily be rewritten for a *generic* driven process[78], in which, given a parameter dependent equilibrium distribution $P(x,\lambda)$, starting from equilibrium conditions, the thermodynamic state of a system is changed irreversibly from $\lambda = a$ to $\lambda = b$, and *viceversa*. We introduce a generic "weight" $\omega(x,\lambda)$, such that $P(x,\lambda) = Z^{-1}(\lambda)\exp(\omega(x,\lambda))$. Then, the general form of the work fluctuation theorem of equation 3.12 is given by

$$\frac{P_{\lambda_t}[x_t]}{P_{\bar{\lambda}_t}[\bar{x}_t]} = \frac{Z(b)}{Z(a)}\exp\left(\Omega[x_t]\right)) \tag{3.25}$$

where we have defined the "action" $\Omega[x_t]$ as

$$\Omega[x_t] = -\int_0^\tau \dot{\lambda}\partial_\lambda\omega(x,\lambda)\mathrm{d}t \tag{3.26}$$

evaluated over the path x_t with schedule λ_t. As a quick consistency check, we observe that in the case where the initial and final canonical states differ in the energy functions, $\omega(x,\lambda) = -\beta E(x,\lambda)$, the path functional $\Omega[x_t]$ is given by $\Omega[x_t] = \beta\int_0^\tau \dot{\lambda}\partial_\lambda E(x,\lambda)dt = \beta W[x_t]$, thereby recovering Eq. 3.12.

Following the same line of the previous demonstration and using Eq. 3.25 in place of Eq. 3.12, the following general acceptance probability for an exchange of states between replicas performed in an arbitrary duration time τ is derived

$$\min\{1, \exp(-\Omega[\mathbf{x}_t])\} \tag{3.27}$$

where, as before, $\Omega[\mathbf{x}_t] = \Omega_{a\to b}[x_{\alpha,t}] + \Omega_{b\to a}[x_{\beta,t}]$

For a process in which the bath temperature changes in time, and the initial and final states are described by the distributions $P(x,\beta_a) = Z(\beta_a)^{-1}\exp(-\beta_\mathrm{a}\mathrm{E(x)})$ and $P(x,\beta_b) = Z(\beta_b)^{-1}\exp(-\beta_b E(x))$, we define a weight $\omega(x,\beta) = -\beta E(x)$ where $\lambda \equiv \beta$. Then, $\Omega[x_t] = \int_0^\tau \dot{\beta}E(x)dt$ [82]. In such case, one finds that $\lim_{\tau\to 0}\Omega[\mathbf{x}_t] = -\Delta\beta\Delta E$, where $\Delta\beta = \beta_b - \beta_a$, and $\Delta E = E(x_2) - E(x_1)$, thereby recovering the standard parallel tempering acceptance ratio, $\min\{1, \exp(\Delta\beta\Delta E)\}$.

The work fluctuation theorem has been extensively used as a method for calculating free energy differences from non equilibrium trajectories. In this contribution, we exploit this theorem in order to obtain an equilibrium sampling of different states, connecting them through non equilibrium transformations. The standard

REM scheme turns out to be the limiting case of a more general algorithm, in which the states of two replicas are exchanged in a given time τ. While for $\tau \to \infty$ such an exchange process is an equilibrium process, and the acceptance probability is unitary, for a generic duration time τ the acceptance probability is a function of the work W_{xc} dissipated during the process. Clearly, with respect to an algorithm employing slow, quasi-reversible replica exchanges, a lower acceptance ratio is expected for faster, out-of-equilibrium exchanges. For instantaneous, strongly dissipative processes ($\tau = 0$), the original algorithm is recovered. In the standard approach the poor scaling of REM sampling efficiency with system size is solved, for example, reducing the temperature difference between replicas and therefore increasing the number of parallel trajectories to be simulated. Here, we show that another solution is possible by introducing the "switching time" τ as an effective parameter of a REM simulation, in order to minimize the dissipative character of replica exchange processes and obtain an efficient walk in the space of the different thermodynamic conditions of the replicas.

Bibliography

[1] N. G. Van Kampen. *Stochastic Processes in Physics and Chemistry.* North-Holland, 1992.

[2] J. Jonas D. L. Hasha, T. Eguchi. *J. Am. Chem. Soc.*, 104:2290, 1982.

[3] D. Chandler. World Scientific Singapore, 1998.

[4] C. Jarzynski. *Phys. Rev. Lett.*, 78:2690, 1997.

[5] G. E. Crooks. *J. Stat. Phys.*, 90:1481, 1998.

[6] P. Procacci, S. Marsili, A. Barducci, G. F. Signorini, and R. Chelli. *J. Chem. Phys.*, 125:164101, 2006.

[7] R. Chelli, S. Marsili, A. Barducci, and P. Procacci. *Phys. Rev. E*, 75:050101, 2007.

[8] G. J. Martyna, D. J. Tobias, and M. L. Klein. *J. Chem. Phys.*, 101:4177, 1994.

[9] G. M. Torrie and J. P. Valleau. *J. Comp. Chem.*, 23:187, 1977.

[10] F. Wang and D. P. Landau. *Phys. Rev. Lett.*, 86:2050, 2001.

[11] A. Laio and M. Parrinello. *Proc. Natl. Acad. Sci. USA*, 99:12562, 2002.

[12] A. Laio, A. Rodriguez-Fortea, F. L. Gervasio, M. Ceccarelli, and M. Parrinello. *J. Phys. Chem. B*, 109:6714, 2005.

[13] S. Marsili, A. Barducci, R. Chelli, P. Procacci, and V. Schettino. *J. Phys. Chem. B*, 110:14011, 2006.

[14] R. H. Swendsen and J. S. Wang. *Phys. Rev. Lett.*, 57:2607, 1986.

[15] C. J. Geyer. Markov chain monte carlo maximum likelihood. In E. M. Keramidas, editor, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*. Interface Foundation, Fairfax Station, 1991.

[16] E. Marinari and G. Parisi. *Europhys. Lett.*, 19:451, 1992.

[17] K. Hukushima and K. Nemoto. *J. Phys. Soc. Jpn.*, 65:1604, 1996.

[18] R. C. Tolman. *The principles of statistical mechanics*. Oxford University Press, 1938.

[19] T. Speck and U. Seifert. *J. Stat. Mech*, page L09002, 2007.

[20] P. Pradhan, Y. Kafri, and D. Levine. *Phys. Rev. E*, 77:041129, 2008.

[21] R. J. Harris and G. M. Schutz. *J. Stat. Mech.*, page P07020, 2007.

[22] S. Park and K. Schulten. *J. Chem. Phys.*, 120:5946, 2004.

[23] C. Jarzynski. *Phys. Rev. E*, 56:5018, 1997.

[24] C. Jarzynski. *J. Stat. Mech.*, page P09005, 2004.

[25] M. P. Allen and D. J. Tildesley. *Computer simulation of liquids*. Clarendon Press, Oxford, 1987.

[26] G. E. Crooks. *Phys. Rev. E*, 61:2361, 2000.

[27] D. J. Evans. *Mol. Phys.*, 101:1551, 2003.

[28] D. J. Evans and D. J. Searles. *Advances in Physics*, 51:1529, 2002.

[29] J. Liphardt, S. Dumont, S. B. Smith, I. Tinoco, and C. Bustamante. *Science*, 296:1832, 2002.

[30] D. Collin, F. Ritort, C. Jarzynski, S. B. Smith, I. Tinoco, and C. Bustamante. *Nature*, 437:231, 2005.

[31] R. D. Astumian. *Am. J. Phys.*, 74:683, 2006.

[32] E. G. D. Cohen and D. Mauzerall. *J. Stat. Mech.*, page P07006, 2004.

[33] G. Hummer. *J. Chem. Phys.*, 114:7330, 2001.

[34] J. Gore, F. Ritort, and C. Bustamante. *Proc. Natl. Acad. Sci. USA*, 100:12564, 2003.

[35] I. Kosztin, B. Barz, and L. Janosi. *J. Chem. Phys.*, 124:064106, 2006.

[36] W. G. Hoover. *Phys. Rev. A*, 31:1695, 1985.

[37] W. G. Hoover. *Phys. Rev. A*, 34:2499, 1986.

[38] G. J. Martyna. *Phys. Rev. E*, 50:3234, 1995.

[39] M. E. Tuckerman, B. J. Berne, G. J. Martyna, and M. L. Klein. *J. Chem. Phys.*, 99:2796, 1993.

[40] C. Jarzynski. *Phys. Rev. E*, 73:046105, 2006.

[41] M. A. Cuendet. *Phys. Rev. Lett.*, 96:120602, 2006.

[42] S. Park, F. Khalili-Araghi, E. Tajkhorshid, and K. Shulten. *J. Chem. Phys.*, 119:3559, 2003.

[43] G. E. Crooks. *Phys. Rev. E*, 60:2721, 1999.

[44] M. E. Tuckerman, Yi Liu, G. Ciccotti, and G. J. Martyna. *J. Chem. Phys.*, 115:1678, 2001.

[45] D. J. Evans, E. G. D. Cohen, and G. P. Morriss. *Phys. Rev. Lett.*, 71:2401, 1993.

[46] J. Hénin and C. Chipot. *J. Chem. Phys.*, 121:2904, 2004.

[47] C. Chipot and J. Hénin. *J. Chem. Phys.*, 123:244906, 2005.

[48] A.D Mackerell, D. Bashford, M Bellot, R.L. Dunbrack, J.D. Evanseck, M.J. Field, J. Gao, H. guo, S. Ha, D. Joseph-Mcarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Nog, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, and M. Karplus. *J. Phys. Chem. B*, 102:3586, 1998.

[49] P. Procacci, T. A. Darden, E. Paci, and M. Marchi. *J. Comp. Chem.*, 18:1848, 1997.

[50] E. Schöll-Paschinger and C. Dellago. *J. Chem. Phys.*, 125:054105, 2006.

[51] M. A. Cuendet. *J. Chem. Phys.*, 125:144109, 2006.

[52] J. M. Schurr and B. S. Fujimoto. *J. Phys. Chem. B*, 107:14007, 2003.

[53] S. R. Williams, D. J. Searles, and Denis J. Evans. *arXiv:cond-mat/0611541*, 2006.

[54] G. Gallavotti and E. G. D. Cohen. *Phys. Rev. Lett.*, 74:2694, 1995.

[55] G. Gallavotti and E. G. D. Cohen. *J. Stat. Phys.*, 80:931, 1995.

[56] G. M. Torrie and J. P. Valleau. *Chem. Phys. Lett.*, 28:578, 1974.

[57] T. Huber, A. E. Torda, and W. F. van Gunsteren. *J. Comput.-Aided Mol. Des.*, 8:695, 1994.

[58] H. Grubm*ü*ller. *Phys. Rev. E*, 52:2893, 1995.

[59] G. Hummer and I. Kevrekidis. *J. Chem. Phys.*, 118:10762, 2003.

[60] G. Bussi, A. Laio, and M. Parrinello. *Phys. Rev. Lett.*, 96:09061, 2006.

[61] V. Babin, C. Roland, T. A. Darden, and C. Sagui. *J. Chem. Phys.*, 125:204909, 2006.

[62] A. Barducci, G. Bussi, and M. Parrinello. *Phys. Rev. Lett.*, 100:02603, 2008.

[63] D. T. Gillespie. *J. Phys. Chem.*, 81:2340, 1977.

[64] D. Chandler. Oxford University Press, 1987.

[65] M. Mezei. *J. Comput. Phys.*, 68:237, 1987.

[66] G. H. Paine and H. A. Scheraga. *Biopolymers*, 24:1391, 1985.

[67] R. W. W. Hooft, B. P. van Eijck, and J. Kroon. *J. Chem. Phys.*, 97(9):6690, 1992.

[68] E. Darve and A. Pohorille. *J. Chem. Phys.*, 115:9169, 2001.

[69] D. Rodriguez-Gomez, E. Darve, and A. Pohorille. *J. Chem. Phys.*, 120:3563, 2004.

[70] P. Procacci, T. A. Darden, E. Paci, and M. Marchi. *J. Comp. Chem.*, 18:1848, 1997.

[71] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. M. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. M. Wang, and P. Kollman. *J. Comput. Chem.*, 24:1999, 2003.

[72] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. *J. Chem. Phys.*, 79:926, 1983.

[73] U. Essmann, M. L. Perera, M. L. Berkovitz, T. Darden, H. Lee, and G. L. Pedersen. *J. Chem. Phys.*, 103:8577, 1995.

[74] M. Iannuzzi, A. Laio, and M. Parrinello. *Phys. Rev. Lett.*, 90:238302, 2003.

[75] L. B. Lucy. *Astronom. J.*, 82:1013, 1977.

[76] W. G. Hoover and C. G. Hoover. *Phys. Rev. E*, 73:016702, 2006.

[77] D. J. Evans and D. J. Searles. *Phys. Rev. E*, 50:1645, 1994.

[78] S. Sasa T. Hatano. *Phys. Rev. Lett*, 86:3463, 2001.

[79] H. Fukunishi, O. Watanabe, and S. Takada. *J. Chem. Phys.*, 116:9058, 2002.

[80] W. K. Hastings. *Biometrika*, 57:97, 1970.

[81] M. E. J. Newman and G. T. Barkema. *Monte Carlo Methods in Statistical Physics*. Oxford University Press, Oxford, 1999.

[82] C. Chatelain. *J. Stat. Mech.*, page P04011, 2007.