Development and Application of Accurate Molecular Mechanics Sampling Methods: From Atomic Clusters to Protein Tetramers

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von
Florent Henri René Hédin
von Frankreich

Basel, 2019

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von:

Prof. Dr. Markus Meuwly
Prof. Dr. Anatole von Lilienfeld

Basel, den 20. September 2016

Prof. Dr. Jörg Schibler
Dekan

"*Une seule partie de la physique occupe la vie de plusieurs hommes, et les laisse souvent mourir dans l'incertitude.* "

"*A single part of physics occupies the lives of many men, and often leaves them dying in uncertainty.* "

François-Marie Arouet, a.k.a. Voltaire : "A Madame la Marquise du Châtelet, Avant-Propos," Eléments de Philosophie de Newton (1738).

UNIVERSITÄT BASEL

Abstract

Philosophisch-Naturwissenschaftliche Fakultät

Departement Chemie

Doctor of Philosophy

Development and Application of Accurate Molecular Mechanics Sampling Methods: From Atomic Clusters to Protein Tetramers

by Florent Henri René Hédin

In this PhD Thesis molecular systems off increasing size and complexity are investigated, using both *standard sampling* and *advanced sampling* methods. The implementation and validation of two of those *rare events sampling methods* is described, namely the SA-MC and PINS algorithm. The development and use of a toolkit for fitting force fields parameters (for the Lennard-Jones and Multipoles parameters), the Fitting Wizard, is presented. The stability of the Hæmoglobin tetramer is also investigated in solution using standard Molecular Dynamics. The two first Chapters introduce the necessary theoretical background, and are followed by the results sections containing the articles written during this PhD.

Acknowledgements

I would like first of all to thank Prof. Markus Meuwly who offered me this opportunity of PhD: I learned a lot working for you, and you pushed me to not only *use* the methods, but to *understand* them properly, which is infinitely more important in my opinion.

Then I would like to acknowledge all the past and present members of the Meuwly and von Lilienfeld groups, for scientific discussions and exchange, but not only *scientific*: I want to mention particularly the never-ending discussions with Pierre-André, Maksym, Juan-Carlos, Juvenal, and the others …around the coffee machine and the sofa in our laboratory !

And last of all I want to express all my gratitude to my family and friends: my parents Philippe and Christine, my sister Clémence, and all the others (my two grand-mothers, my aunts, …and more) who always supported me in my study and professional choices. And also to those old and best friends that supported me for at least one decade (and more for some of them, how brave they were !): Morgan, Clément, Aurianne, Benjamin, Anne, David.

Finally I also want to thank Prof. Stefan Willitsch who accepted to be chairman for the coming PhD defence, Dr. Georg Funk who is in charge of the excellent CCCS competence centre, and members of the staff of the University of Basel–Departement Chemie.

Many thanks to all of you !

Contents

List of Abbreviations

BO	**B**orn **O**ppenheimer (approximation)
ESP	**E**lectro**S**tatic **P**otential
FES	**F**ree **E**nergy **S**urfaces
FF	**F**orce **F**ield
FFT	**F**ast **F**ourier **T**ransform
FW	**F**itting **W**izard
INS	**I**nfinite **S**wapping
LJ	**L**ennard **J**ones
MC	**M**onte **C**arlo
MCMC	**M**arkov **C**hain **M**onte **C**arlo
MD	**M**olecular **D**ynamics
MM	**M**olecular **M**echanics
MTD	**M**e**T**a **D**ynamics
MTP	**M**ul**T**i**P**ole(s)
NMR	**N**uclear **M**agnetic **R**esonance
P.d.f or **p.d.f**	**P**robability **d**ensity **f**unction
PINS	**P**artial **I**nfinite **S**wapping
PME	**P**article **M**esh **E**wald
PT	**P**arallel **T**empering
QM	**Q**uantum **M**echanics
RE	**R**eplica **E**xchange
SA-MC	**S**patial **A**veraging **M**onte **C**arlo
US	**U**mbrella **S**ampling

Physical Constants

Avogadro constant	$\mathcal{N}_A = 6.022 \times 10^{-23}\,\mathrm{mol}^{-1}$
Boltzmann constant	$k_\mathrm{B} = 1.381 \times 10^{-23}\,\mathrm{J\,K}^{-1}$
Electron mass	$m_e = 9.109 \times 10^{-31}\,\mathrm{kg}$
Elementary charge	$e = 1.602 \times 10^{-19}\,\mathrm{C}$
Ideal gas constant	$R = 8.314\,\mathrm{J\,K}^{-1}\,\mathrm{mol}^{-1}$
Neutron mass	$m_n = 1.675 \times 10^{-27}\,\mathrm{kg}$
Proton mass	$m_p = 1.673 \times 10^{-27}\,\mathrm{kg}$
Reduced Planck constant	$\hbar = 1.054 \times 10^{-34}\,\mathrm{J\,s}$
Vacuum permittivity	$\varepsilon_0 = 8.854 \times 10^{-12}\,\mathrm{C\,V}^{-1}\,\mathrm{m}^{-1}$

Mathematical conventions

$a \cdot b$	Dot product between vectors a and b
$a \times b$	Cross product between vectors a and b
X^T	Transpose of matrix X
X^{-1}	Inverse of matrix X
Id_n	Identity matrix of size n
$\|\|a\|\|$	Norm of vector a
$\|\mathcal{U}\|$	Cardinality (numb. of elements) of a set $\mathcal{U}$
$\mathcal{A} \cup \mathcal{B}$	Union of two sets $\mathcal{A}$ and $\mathcal{B}$
$\mathcal{A} \cap \mathcal{B}$	Intersection of two sets $\mathcal{A}$ and $\mathcal{B}$
$\mathbb{E}(A)$	Expected value of an observable A
$\mathcal{F}(\cdots)$	Fourier Transform of $\cdots$
$\nabla f(x,y) = \frac{\partial f}{\partial x}\vec{i} + \frac{\partial f}{\partial x}\vec{j}$	Gradient of a real function $f(x,y)$
$\{f,g\} = \nabla_q f \ \nabla_p g \ - \ \nabla_p f \ \nabla_q g$	Poisson brackets applied to functions f,g
$f \circ g = (f \circ g)(x) = f\big(g(x)\big)$	Composition of 2 functions f and g

Introduction

Although the modern computational resources are continuously increasing, accurately sampling the conformational space for a system of interest can still be a very challenging task. Significant conformational changes, such as atomic clusters rearrangement, peptide/protein folding, ligand migration, … may involve numerous intermediate configurations separated by significant energy barriers, resulting in a very low probability of observing the transition event. Those are usually referred to as *rare events*, which are sometimes observed only after billions of simulation steps (i.e. μs to ms of total simulation time). For systems in which configuration space is well connected, standard techniques such as Molecular Dynamics (MD) and Monte Carlo methods (MC) (especially using the Metropolis-Hastings algorithm) can still be efficient. However, enhanced sampling methods are usually required in order to obtain a sufficient sampling of low-probability configurations. During my PhD time I had to implement in CHARMM two rare event sampling methods, SA-MC and PINS, and learned how to use MC and MD methods and apply them for studying system of increasing size and complexity, from simple rare gases clusters in vacuum to large Hæmoglobin tetramers in a simulation box containing up to approximately 350,000 atoms ! But let us start the journey from the beginning…

When I arrived in Basel in March 2011 for my six months of master Thesis, I had only a really primitive and limited knowledge of sampling methods, and Prof. Meuwly proposed me to work on the *spatial averaging* algorithm: I first wrote a simple Fortran code for application to Lennard Jones clusters, and results were promising. We thus started thinking about writing a proper implementation in CHARMM and at the end of the six months we had a features limited (only in gas phase) but first valid implementation into CHARMM. Then when Prof. Meuwly gave me the possibility to stay in his group for my PhD in october 2011 I happily accepted.

During the first months I followed lectures and tried to progress in my understanding of molecular mechanics methods: by the end of 2011 I had learned enough about CHARMM for continuing the spatial averaging implementation, and within a few weeks we managed to obtain the first results for which an implicit solvent model was used. At the same time, Nuria Plattner, who wrote the original spatial averaging articles, came back into the group for a few month between two post-doctoral contracts, and she has helped me a lot to understand the mathematics behind spatial averaging; it is also at his time that I heard about *infinite swapping* for the first time.

Then during summer and autumn 2012, I had the opportunity to participate to two workshops, one in July in Austria where the topic was “Free Energy landscapes”, and one at the end of October in USA where the topic was “MC simulations with application to biological systems” (where I also met Jimmie D. Doll, the Expert of rare events sampling methods in the Chemistry department of Brown University), and they again revealed extremely useful for my understanding of the sampling methods. It was also a good opportunity for me to present for the first time my results to a large audience of scientists, and it is at this time that I realised how useful were rare event sampling methods for biologists, physicists, chemist, mathematicians... We continued working on the spatial averaging implementation in CHARMM during the following months (which became the “SA-MC” module), submitted it to the community and finished writing the first version of the article.

Just before final publication of this SA-MC article, I discovered atomic multipoles and the fitting of force field parameters during spring 2014 when I had to start working on Tristan Bereau’s scripts: I never realised before this moment how much time consuming was the fitting of force fields parameters. This project extended up to July 2016 when I published, together with Krystel El Hage, an article were we demonstrated the use of this all-in-one Fitting Wizard toolkit. This project revealed itself really useful because it motivated me to reach a better understanding of standard molecular dynamics and force fields in general. It also allowed be to improve a lot my software development knowledge.

My second rare event sampling experience was the implementation of the above mentioned infinite swapping algorithm, one more method coming from J.D. Doll and Nuria Plattner ! This work started

in March 2015, and the CHARMM re-implementation ("PINS" module) was surprisingly fast, taking not more than 5 or 6 weeks. However we had to intensively test and validate the methods, and it took 8 more months before the first submission of the article. This period of time allowed me to get used to some of the computational biology analysis methods, while reinforcing my knowledge about free energy estimation (and thermodynamics in general).

My last project started in October 2015 when I continued the work of Prashant Gupta who started to investigate the stability of hæmoglobin tetramer in solution. This work is still not finalised but it definitely allowed me to get used to bio-systems, and strengthened my knowledge of molecular dynamics methods, especially for the "performance" because I got used to PME methods, domain decomposition, etc …

The organisation of this thesis will reflect the various methods I got familiar with and mentioned above:

In Chapter 1 the basic principles of molecular mechanics and thermodynamics required for understanding the following Chapters are introduced.

In Chapter 2 MD and MC standard methods are introduced using concepts from Chapter 1, then SA-MC and PINS are introduced as advanced sampling techniques.

In Chapters 3 and 4 the two SA-MC and PINS articles are commented, with results of supplementary analyses provided if available.

In Chapter 5 the Multipoles and the article describing the elaboration of the Fitting Wizard tool are presented.

In Chapter 6 are gathered all the results we obtained when investigating the stability of hæmoglobin tetramer in solution.

Finally Appendix A contain an article from Pierre-André Cazade, a former colleague, to which I contributed as second author.

At the end of this thesis, extensive bibliography and an index are provided (with back reference links for the pdf version that allows one to click on the citation/index in order to find where in the main text it was introduced).

Part I

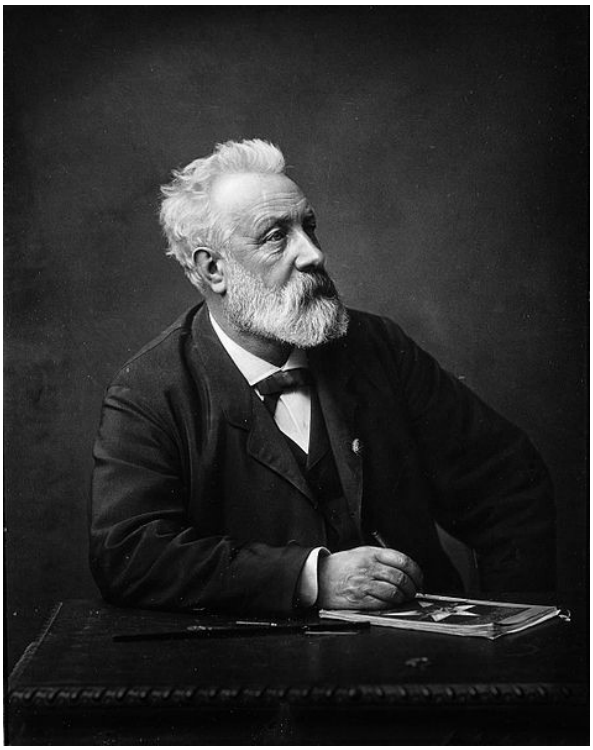
METHODS

Chapter 1

Molecular Simulations: Principles

"*La science, mon garçon, est faite d'erreurs, mais d'erreurs qu'il est bon de commettre, car elles mènent peu à peu à la vérité. "*

"*Science, my lad, has been built upon many errors; but they are errors which it was good to fall into, for they led to the truth "*



Jules Verne, Voyage au Centre de la Terre (Journey to the Center of the Earth) , 1864, Chapt. XXXI

While reading the following Chapters, the user will notice that standard and advanced sampling methods (see Chapter 2) heavily rely on a core of fundamental statistical mechanics concepts: it is thus necessary to clarify directly from the beginning those notions. Hence Section 1.1 focuses on a description of molecular interactions, and on notions of thermodynamics ensembles and averaging. In Section 1.2 the concept of Force Field is presented, with each term of the potential energy function being mathematically introduced and illustrated if necessary. The last Section 1.3 focuses on Free Energy, a thermodynamic property of interest regularly used in Chapters 2 to 5.

The equations and concepts they represent are, whenever possible, detailed using a mathematically rigorous notation: abbreviations, physical constant and mathematical conventions are listed at the beginning of the thesis. Key concepts are *emphasised* the first time they are introduced, and are usually listed in the Index at the end of the thesis.

One can refer to the excellent book "Free Energy Computations: A Mathematical Perspective"[1] from Lelièvre, Stoltz, and Rousset, and especially the Introduction from which the organisation of Chapters 1 and 2 is partially inspired.

1.1 Statistical Mechanics and Thermodynamics principles

For an accurate description of a chemical system at a microscopic level, it is necessary to introduce several key concepts of statistical mechanics: first, the definition of intra and inter atomic *interaction laws* (Subsection 1.1.1), modelled using a Hamiltonian H, are introduced. Then the concept of *thermodynamic ensembles* (Subsection 1.1.2), i.e. measures of probability from which macroscopic observables are estimated from an *averaging*, are presented.

1.1.1 Interactions

Classical Hamiltonian Let us first consider a system, composed of N particles, described by a position vector $q = (q_1, \cdots, q_N) \in D$, and a momenta vector $p = (p_1, \cdots, p_N) \in \mathbb{R}^{3N}$ (momentum being the product mass*velocity of a particle, i.e. $p_i = m_i * v_i$). $D \subseteq \mathbb{R}^{3N}$ is the *configuration space*, populating the whole set of the possible atomic configurations (or a only a part of it, if boundaries or position constraints are defined). The couple (q, p) describes a possible *microscopic state* (or *microstate*) of the system of interest. The set of all the possible microscopic states is the *phase space* Ω, and following the previously introduced definitions it can be written as $\Omega = D \times \mathbb{R}^{3N}$.

Molecular interactions can be described using the *potential energy* function $V(q)$, and the *kinetic energy* function $K(p)$. The total energy of the system is thus given by the Hamiltonian H:

$$H(q, p) = K(p) + V(q) \tag{1.1}$$

The kinetic energy $K(p)$ can be written as: $K(p) = \frac{1}{2} p^T M^{-1} p$ where M is the diagonal mass matrix $M = \text{diag}(m_1 \times Id_3, \cdots, m_N \times Id_3)$ and Id_3 the 3-identity matrix.

Quantum Methods The potential energy function can ideally be obtained by using *ab-initio* computations, relying on non-empirical approaches. It is possible to rewrite Equation 1.1 using quantum operators as following:

$$\hat{H} = \hat{K} + \hat{V} \tag{1.2}$$

In *Quantum Mechanics* (QM) the *Born-Oppenheimer* (BO) approximation [2] is usually used. It assumes that the motion of nuclei and electrons are separable, since nuclei are much heavier than electrons and thus they move more slowly. Hence, kinetic energy contributions can also be split in two parts. When applied to Equation 1.2 the BO approximation allows to simplify to the following:

$$\hat{H} = \hat{K}_{nucl} + \hat{K}_{elec} + \hat{V} \tag{1.3}$$

Where $\hat{V} = V(q) = V(r, R)$ is the potential energy operator, comprising electron-electron, electron-nuclei and nuclei-nuclei interactions, $\hat{K}_{nucl}$ the kinetic contribution of the nuclei, and $\hat{K}_{elec}$ the kinetic contribution of the electrons. r and R respectively denotes the electronic and nuclear coordinates, such as $q = (r, R)$ and $(r, R) \in D$.

If A is the total number of nuclei, m_a the mass of a given nucleus, and $\hbar$ the reduced Planck constant, one can define

$$\hat{K}_{nucl} = \sum_{a=1}^{A} \frac{\hat{p}_a \cdot \hat{p}_a}{2m_a} = -\frac{\hbar^2}{2} \sum_{a=1}^{A} \frac{1}{m_a} \nabla_a^2$$

and similarly if F the total number of electrons and m_f the electron mass :

$$\hat{K}_{elec} = -\frac{\hbar^2}{2} \sum_{f=1}^{F} \frac{1}{m_f} \nabla_f^2$$

As introduced above, $V(r, R)$ can be decomposed as the sum of three terms: nuclei-nuclei, nuclei-electrons and electrons-electrons interactions. Let e be the elementary charge, Z_a the atomic number of nucleus a and ε_0 the vacuum permittivity:

$$V(r, R) = \frac{1}{4\pi\varepsilon_0} \left(\sum_{a=1}^{A} \sum_{b>a}^{A} \frac{Z_a Z_b e^2}{|R_a - R_b|} + \sum_{a=1}^{A} \sum_{f=1}^{F} \frac{-Z_a e^2}{|R_a - r_b|} + \sum_{f=1}^{F} \sum_{g>f}^{F} \frac{e^2}{|r_a - r_b|} \right)$$

By substituting in Equation 1.3 one can write the BO $\hat{H}$ as:

$$\hat{H} = -\frac{\hbar^2}{2} \sum_{a=1}^{A} \frac{1}{m_a} \nabla_a^2 - \frac{\hbar^2}{2} \sum_{f=1}^{F} \frac{1}{m_f} \nabla_f^2 + V(r, R) \tag{1.4}$$

The electronic ground-state energy $\mathcal{E}$ is obtained by minimising the electronic problem over the *Hilbert space* $\mathcal{H}$ of the possible *wave-functions* ψ (*Schrödinger equation*[3]):

$$\mathcal{E} = \inf\{ \langle \psi | \hat{H} | \psi \rangle_{\mathcal{H}} \mid \psi \in \mathcal{H} \} \tag{1.5}$$

where the ψ wave-functions are normalised (L^2 norm) so that

$$||\psi||_{L^2} = 1$$

However QM methods for solving Equations 1.4 and 1.5 are particularly time-consuming, and furthermore limited to time scales much smaller than the ones on which chemical and biological events occur. Thus only small to middle-size systems can be dynamically treated, using methods such as Born-Oppenheimer MD or the *Car-Parrinello* approach.[4]

Empirical Potentials In practice larger systems are often studied using *Empirical potentials* for modelling the classical Hamiltonian $H(q, p)$. The formulae of those functional forms are usually designed and parametrised in order to reproduce accurately an ab-initio energy $\mathcal{E}$, or a set of experimental thermodynamic properties, estimated from simulation using ensemble average (see 1.1.2). A collection of empirical potentials necessary for approximating with accuracy $H(q, p)$ is a *Force Field* (FF), and

although some FFs may be constructed via a highly specific formulation, depending on system and application criteria, they usually estimate the total potential energy $V(q)$ as following:

$$\begin{aligned} V(q) &= V_{bonded} + V_{non-bonded} \\ V_{bonded}(q) &= V_{bonds} + V_{angles} + V_{torsion} \\ V_{non-bonded}(q) &= V_{electrostatic} + V_{van\ der\ Waals} \end{aligned} \tag{1.6}$$

Note that the (q) dependency was omitted for clarity, but all the $V\cdots$ terms depend on the coordinates q of the system.

It is worth mentioning that QM/MM methods, combining both the QM accuracy on a small part of the system and the FFs speed for treating the rest of the system are more and more commonly used. More details about the FFs will be introduced later in Section 1.2.

Now that basic notions concerning atomic interactions and the energy terms have been introduced, the following subsection will focus on the use of the total energy above defined using the Hamiltonian $H(q,p)$, in order to estimate *observables* of interest.

1.1.2 Thermodynamic ensembles and averages

In Subsection 1.1.1 the phase space $\Omega = D \times \mathbb{R}^{3N}$ was introduced as the set of all the microstates a system can exhibit. For each couple $(p,q) \in \Omega$ it is possible to define a probability of observance $\rho(q,p)$, and ρ is usually named *Probability density function (P.d.f or p.d.f).*

On the contrary a *Macroscopic* state (or *macrostate*) is defined using external parameters (observables), usually illustrating the global environment and the conditions of a simulation (Temperature,Volume…). For a given macrostate there is usually an uncountable amount of associated microstates, so a macrostate can be considered as an ensemble of microstates exhibiting a specific property through a probability measure ρ. For instance, the macrostate satisfying a temperature of 300 K can be imagined as the set of all the microstates characterised by a set of $\rho(q,p)$, for which the total *average* kinetic energy obeys the formula $K(p) \approx \frac{3}{2} K_B T$ with T = 300 K.

In the following paragraphs notions of average and *thermodynamic ensembles* are detailed.

Average of an observable More generally, for a given observable A, its expected value $\mathbb{E}_\rho(A)$ is given by:

$$\mathbb{E}_\rho(A) = \int_\Omega A(q,p)\ \rho(q,p)\ d\Omega \tag{1.7}$$

where $d\Omega = d^3q\ d^3p$ is the volume element of the phase space. In order to obtain a converged *ensemble average* using Equation 1.7, it is expected that the two following conditions are fulfilled: (i) the numerical methods of interest can perform an *exhaustive* sampling of the (ideally whole) phase space $\Omega : (q^D, p^{3N})$, and (ii) it is possible to generate *independent* microscopic states following the p.d.f $\rho(q,p)$.

Therefore the ensemble average $\bar{A}_\rho$ is derived from Equation 1.7 as:

$$\bar{A}_\rho = \lim_{N \to +\infty} \sum_{n=1}^{\Omega} A(q^D, p^{3N}) \tag{1.8}$$

In the case of *Markov Chain Monte Carlo* (MCMC or simply MC for *Monte Carlo*) methods, such has the *Metropolis-Hastings* Algorithm (which will be considered later in Chapter 2 Section 2.2.2), Equation 1.8 can be applied directly because the two above mentioned criteria are implicitly satisfied. But for *Molecular Dynamics* (MD) methods, the sequences (q,p) are generated through a time

discrete trajectory, thus two microscopic configurations (q^n, p^n) and $(q^{n+\tau}, p^{n+\tau})$ can be considered independent only if $\tau \gg 1$ (τ can be interpreted as a *decorrelation time*). Equations 1.7 and 1.8 can be modified in order to introduce the notion of *dynamical average*, more appropriate for MD methods:

$$\bar{A}_\tau = \mathbb{E}_\tau(A) = \int_{t=0}^{\tau} A(q(t), p(t))\ \rho(q,p)\ dt \tag{1.9}$$

For an infinitely long trajectory ($\tau \to +\infty$) the dynamical trajectory can be considered as a *stochastic trajectory*, thus following the p.d.f ρ. This is the *ergodic hypothesis*:

$$\lim_{\tau \to +\infty} \bar{A}_\tau = \bar{A}_\rho \tag{1.10}$$

Thermodynamic ensembles The nature of $\rho(q,p)$ has not been described in detail so far. It was introduced at the beginning of Subsection 1.1.2 as a measure of the probability to observe a given microstate (q,p). Considering all the possible microstates populating the phase space Ω, the p.d.f ρ can be generalised as:

$$\int_\Omega \rho(q,p)\ d\Omega = 1 \tag{1.11}$$

A set of (q,p) that fulfils $\rho(\Omega) = 1$ is a *statistical ensemble* (it is the *probability space* for which the p.d.f ρ is valid). A *thermodynamic ensemble* is a subset of the above introduced statistical ensembles: it has reached a *statistical equilibrium*, and thus it can be described by macroscopic observables. In the following, the term *ensemble* will always refer to the concept of *thermodynamic ensemble*, unless if the adjective *statistical* is explicitly used.

The following ensembles are commonly encountered, and where initially introduced by Gibbs:

The Microcanonical ensemble or **NVE** is an *isolated* (no particle exchange allowed thus **N** = constant) system, of fixed volume **V**, and for which the value of the total Hamiltonian is fixed to a value **E**. The set $\mathcal{U}$ of the possible microstates fulfilling the macroscopic definition **NVE** can be written, using a conditional notation, as:

$$\mathcal{U} = \Big\{ (q,p) \in \Omega \mid H(q,p) = E \Big\}$$

which is read as "the set $\mathcal{U}$ of the configurations (q,p) for which the Hamiltonian $H(q,p)$ has a total energy of E". Thus each member of the set $\mathcal{U}$ is equiprobable and ρ is simply defined as

$$\rho\big(\forall (q,p) \in \mathcal{U}\big) = \frac{1}{|\mathcal{U}|} \tag{1.12}$$

where $|\mathcal{U}|$ is the *cardinality* (i.e. the number of elements) of the set $\mathcal{U}$.

The Canonical ensemble or **NVT** is a *closed* system, where the energy is not exactly defined and thus fluctuates, but instead where the temperature **T** is fixed. The number of particles **N** and the volume **V** are also fixed. This ensemble is really appropriate for describing systems in contact with a heat bath (also referred as a "thermostat"). In this case the microstates follow the *canonical measure* distribution $\rho(q,p)$, written as:

$$\rho(q,p) = Z_{NVT}^{-1} \exp\big(-\beta H(q,p)\big) \tag{1.13}$$

where $\beta = 1/k_\mathrm{B}T$ is the inverse temperature, k_B the Boltzmann constant, and Z_{NVT} is the *canonical partition function*, a normalisation constant:

$$Z_{NVT} = \int_\Omega \exp\big(-\beta H(q,p)\big)\ d\Omega \tag{1.14}$$

Since the Hamiltonian H is usually *separable* (energetic contributions of q and p can be treated separately see Equation 1.1), it is possible to rewrite $\rho(q,p)$ as the product of two independent p.d.f. υ and κ (since the exponential of a sum e^{a+b} is the product of two exponentials $e^a e^b$):

$$\rho(q,p) = \upsilon(q)\ \kappa(p)$$

where $\upsilon(q)$ is a p.d.f governed by the potential energy V, normalised by $Z_\upsilon = \int_D \exp(-\beta V(q))\ d^3q$:

$$\upsilon(q) = Z_\upsilon^{-1} \exp(-\beta V(q)) \tag{1.15}$$

and $\kappa(p)$ is governed by the kinetic energy K and is normalised by Z_κ:

$$\kappa(p) = Z_\kappa^{-1} \exp\Big(-\frac{\beta}{2} p^T M^{-1} p \Big) \tag{1.16}$$

If Z_κ is written as:

$$Z_\kappa = \left(\frac{\beta}{2\pi}\right)^{\frac{3N}{2}} \prod_{i=1}^{N} M(i,i)^{-\frac{3}{2}} \int_{\mathbb{R}^{3N}} \exp\Big(-\frac{\beta}{2} p^T M^{-1} p \Big)\ dp$$

Then $\kappa(p)$ in Equation 1.16 follows the *Maxwell-Boltzmann distribution.*[5, 6]

Equation 1.16 can be trivially sampled by generating Gaussian distributed random velocities, and this is usually how initial momenta are generated for initiating MD simulations. The sampling challenge thus only concerns Equation 1.15, and this is where the rare event sampling problem usually arises. While standard and advanced sampling methods will be further described in Section 2.2, it is already important to point out that MC methods using the Metropolis-Hastings acceptance criterion directly generate uncorrelated states respecting Equation 1.15.

Other derived ensembles It is usually possible to deduce the p.d.f and the partition function of a new ensemble through a modification of the NVT Equations 1.13 and 1.14.

Let $\mathcal{X}$ be the set of the possible values that a new observable x can take: x can be considered as an extra *degree of freedom*, and the microstates are now described by a triplet (q,p,x). An extended phase space Γ can be defined as the union between the previously defined space Ω and all the possible x from $\mathcal{X}$, i.e.:

$$\Gamma = \bigcup_{x \in \mathcal{X}} \Omega \times x$$

One can then introduce a set of new p.d.f and partition function related to Equations 1.13 and 1.14:

$$\begin{aligned}
\rho(q,p,x) &= Z_\gamma^{-1} \exp\big(-\beta H(q,p,x) \big) \\
\rho(q,p,x) &= Z_\gamma^{-1} \exp\big(-\beta (H(q,p) + C(x)) \big) \\
\rho(q,p,x) &= Z_\gamma^{-1} \exp\big(-\beta H(q,p) \big) \exp\big(-\beta C(x) \big)
\end{aligned} \tag{1.17}$$

where $C(x)$ is an extra potential, detached from the Hamiltonian, and defined for a given value of x, and where Z_γ is:

$$
\begin{aligned}
Z_\gamma &= \int_\Gamma \exp\big(-\beta H(q,p,x)\big)\ d\Gamma \\
&= \int_\Gamma \exp\big(-\beta H(q,p)\big) \exp\big(-\beta C(x)\big)\ d\Gamma \\
&= \int_\Omega \exp\big(-\beta H(q,p)\big)\ d\Omega \int_{\mathcal{X}} \exp\big(-\beta C(x)\big)\ dx \\
Z_\gamma &= Z_{NVT} \int_{\mathcal{X}} \exp\big(-\beta C(x)\big)\ dx
\end{aligned} \tag{1.18}
$$

Therefore, the new partition function Z_γ is defined as a weighted sum of the canonical partition function Z_{NVT}.

One can briefly mention two other ensembles, elaborated through the Equations 1.17 and 1.18:

The Isobaric-Isothermal ensemble or **NPT** is characterised by a constant pressure P. The system is still considered closed and in contact with a heat reservoir, hence N and T are also fixed at a constant value. The additional degree of freedom is dV, a volume variation, such as $dV \in \mathcal{V}$, where $\mathcal{V}$ is the space of the possible variations. The potential introduced in Equations 1.17 and 1.18 is the product pressure–volume PV, having units of energy. Thus one can write the p.d.f. $\rho(q,p,V)$ and the Z_{NPT} partition function as:

$$
\begin{aligned}
\rho(q,p,V) &= Z_{NPT}^{-1} \exp\big(-\beta(H(q,p)+PV)\big) \\
Z_{NPT} &= \int_{\mathcal{V}} Z_{NVT} \exp\big(-\beta PV\big)\ dV
\end{aligned} \tag{1.19}
$$

where the value of V and dV are easily determined by considering the periodic boundaries defined on the previously introduced domain D, whereas P is usually calculated from the classical *virial theorem* [7].

The Grand Canonical ensemble or μ**VT** is an *opened* system where the number of particles varies, but maintained in a thermodynamic equilibrium state, where the temperature **T** and the *chemical potential* μ are kept constant by usage of a heath bath and a *chemical reservoir*. The additional variable is this time $N \in [0; +\infty]$, and the supplementary term detached from the Hamiltonian is $-\mu N$, where the chemical potential can be defined as the resulting energy variation of a thermodynamic system when the quantity of a given species N varies. Therefore substitutions in Equations 1.17 and 1.18 allow to defined the two following p.d.f and partition function for the (μVT) ensemble:

$$
\begin{aligned}
\rho(q,p,N) &= Z_{\mu VT}^{-1} \exp\big(-\beta(H(q,p)-\mu N)\big) \\
Z_{\mu VT} &= \sum_{N=1}^{+\infty} Z_{NVT} \exp\big(\beta \mu N\big)
\end{aligned} \tag{1.20}
$$

Note that the integral is replaced by a discrete sum, in order to respect the physical fact that the addition or removal of a particle cannot be infinitesimal, and also that a proper estimation of $Z_{\mu VT}$ implies also an accurate sampling of the canonical Z_{NVT}. Hence it is possible to consider the (μVT) ensemble as a *superimposition* of an infinity of (NVT) macrostates.

Now that averages and ensembles were clearly presented, it is time to go back to the definition of the potential energy $V(q)$, using empirical formulae, i.e. the above briefly mentioned (Equation 1.6) Force Fields (FFs).

1.2 Force Fields

Let us now consider more in detail the previously mentioned Equation 1.6:

$$\begin{aligned} V(q) &= V_{bonded} + V_{non-bonded} \\ V_{bonded}(q) &= V_{bonds} + V_{angles} + V_{torsion} \\ V_{non-bonded}(q) &= V_{electrostatic} + V_{van\ der\ Waals} \end{aligned}$$

Each of the five V energy terms represent a possible type of inter-molecular (non-bonded) or intra-molecular (bonded) interaction. In order for *Force Fields* (FFs) based method to be competitive versus ab-initio methods, the lost of accuracy induced by the empirical estimations should be minimised while the computational efficiency should be maximised.

MC methods, further described in Section 2.2.2, only require potential energy calculations. But for MD, as it will be emphasised in Chapter 2 Section 2.2.1, forces are required for integrating equations of motion. The force F is thus determined as $F = -\nabla V(q)$; while only potential energy equations are introduced in the following, a force field code is always structured such as to calculate energy and force contributions at the same time. For this analytical derivatives are coded together with the energy formula, as numerical differentiation would be too demanding. However it is worth to mention one exception, not detailed in the following paragraphs: some molecular mechanics codes may allow the user to specify a custom *tabulated* potential, i.e. a table which associates an energy value to a given distance between two atoms, and for which no formula is defined. In that case, accurate numerical differentiation methods are used, for instance finite difference methods.

In the following mathematical formulations for non-bonded and bonded terms are introduced. The requirements and necessary steps for parametrising the empirical parameters are also briefly mentioned when necessary. Let $r_{ij} = q_j - q_i$ be the distance between two atoms i and j.

1.2.1 Non-bonded terms

Non-bonded terms in FFs usually consist in *Coulombic potential* between point charges (or extended at a higher level using *Multipoles*, see Section 5.1) which models electrostatic interactions, and in the *Lennard-Jones potential* which attempt to reproduce the short-distances *Van der Waals forces*.

Coulomb potential The point charge electrostatic potential between two atoms (i, j) of respective partial charges (z_i, zj) is defined using the Coulomb's Law:

$$V_{electrostatic} = V(q_i, q_j) = \sum_{(i,j)-pairs} \frac{z_i z_j}{4\pi\epsilon_0 r_{ij}} \tag{1.21}$$

See Figure 1.1 for an illustration with $z_i = z_j = 1$ and $z_i = 1; z_j = -1$.

It is important to mention that the notion of charge in this context slightly differs from the physical definition. Indeed, most of the FFs use a *fixed-charge* approach in which each atom is assigned a single possible charge value, prior to the simulation (parametrisation phase). Therefore during the simulation charge is not affected by the local electrostatic environment. Several polarisable FFs, where each charge is influenced by interaction with its neighbours, and thus dynamically evolves during the simulation, have been in development over the last years. Several methods were proposed over the years, including:

Fluctuating charges models (CHEQ [8, 9], available for CHARMM) where charges are still located on each atoms, but where charges partially fluctuate between the atoms of a given molecule during the simulation. Coulomb's law is still used without modifications.

Drude oscillators ([10, 11]), available for CHARMM, GROMACS, OpenMM, and NAMD. Each atom is represented using two charge sites: one is the atomic nucleus, as for standard methods, but the

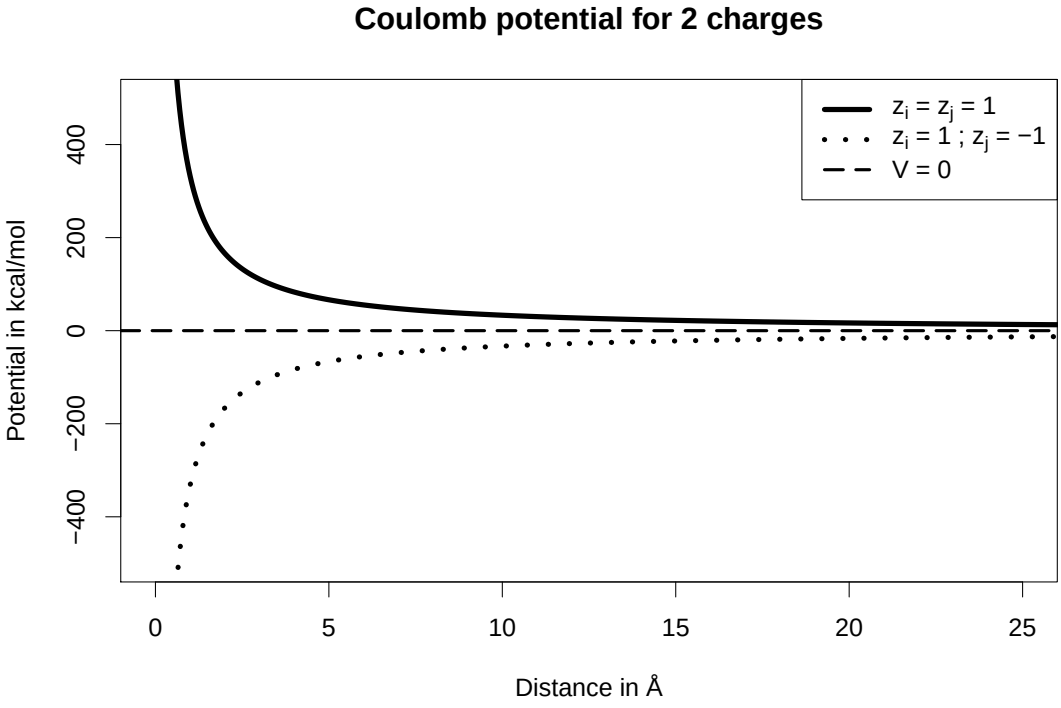


FIGURE 1.1: Illustration of the coulomb potential for cases where $z_i = z_j = 1$ and $z_i = 1; z_j = -1$

second is a massless particle (a Drude particle), linked to the nucleus by a spring. The total charge for a given atom is the sum of those two, and the total charge does not change during simulation. The Drude particle is relatively free to move around the nucleus during simulation (the amplitude of this allowed movement being fixed by the value of the harmonic spring constant), and this move mimics the induced dipoles. Here also the Coulomb's law is used without further modification.

Inducible dipoles methods, where extra sites are placed at specific places like the nuclei, or on bonds between atoms…and where for each site the value of the induced dipole is determined by the total electric field. Thus extra calculations are required, such as charge-dipole or dipole-dipole interactions. This approach is currently available in the AMBER software.

Multipole Electrostatics methods, further described later in Section 5.1, where monopoles (charges), dipoles (vectors), and quadrupoles (tensors) are used for accurately describing the possible charge anisotropy. AMOEBA [12] and CHARMM provide [13–18] such computational methods.

For a comparison of the above mentioned polarisable and multipoles methods, one can refer to the review written by C. M. Baker [19].

Lennard-Jones potential The Lennard-Jones (LJ) potential, introduced by John Lennard-Jones in 1924 [20] is a simple mathematical model for approximating the Van der Waals interaction between two particles. Although it was originally defined for a pure gas or fluid of uncharged atoms, it is nowadays used for modelling short range interactions between all types of atoms and for any material phase.

Two equivalent formulations are usually encountered in Literature:

$$V_{LJ} = V(q_i, q_j) = \sum_{(i,j)-pairs} 4\varepsilon \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right)$$
$$V_{LJ} = \sum_{(i,j)-pairs} \varepsilon \left(\left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{6} \right) \tag{1.22}$$

where ε_{ij} is the depth of the potential well, and σ_{ij} the distance at which the potential is 0. Some FFs (e;g. CHARMM) use the second formulation instead where r_{ij}^{min} correspond to the distance where the potential is at minimum and thus where the resulting force F_{LJ} is null. The relation between the two terms is $r_{ij}^{min} = 2^{\frac{1}{6}} * \sigma_{ij}$.

The rules for defining ε_{ij} and σ_{ij} for a pair (i, j) are called the *Lorentz-Berthelot* [21] combining rules: the ε are combined using a geometric mean $\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}$, and the σ using an arithmetic mean $\sigma_{ij} = (\sigma_i + \sigma_j)/2$. Figure 1.2 illustrates the Lennard-Jones potential, and the effect of the Lorentz-berthelot mixing rules, for two arbitrary atoms (parameters to be read from the figure).

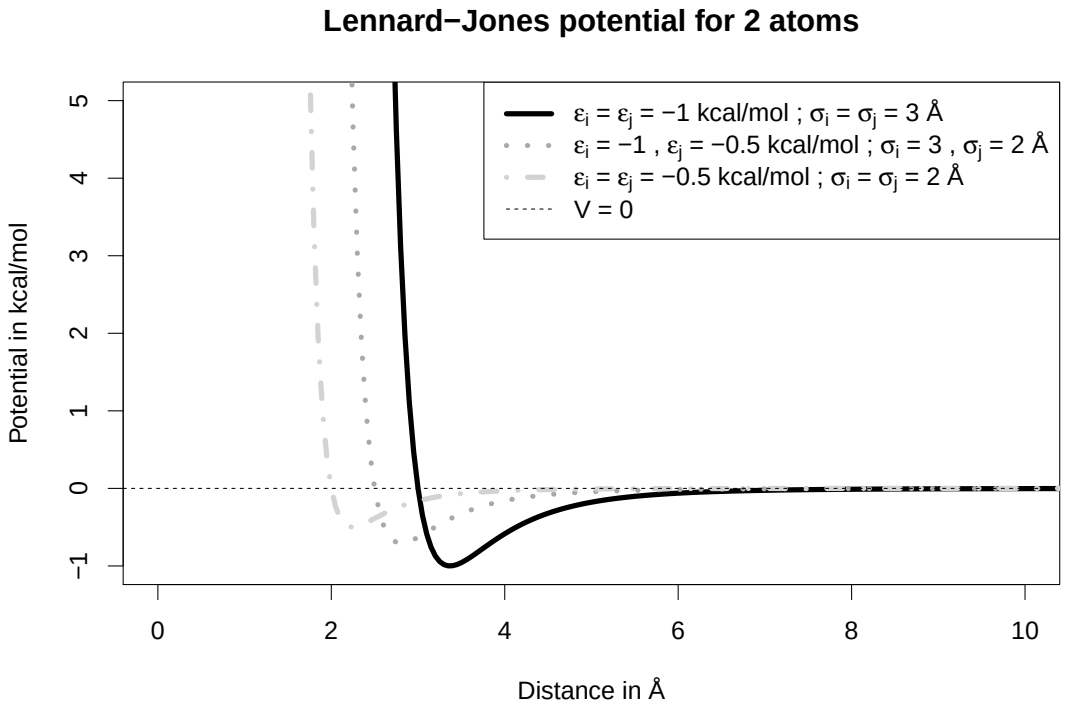


FIGURE 1.2: Illustration of the Lennard-Jones potential and the Lorentz-Berthelot mixing rules, for 2 arbitrary atoms (i, j) with the parameters to be read from the legend.

The power of 6 term is used for modelling dipole-dipole interactions caused by the electron dispersions, as introduced by the *London forces* [22, 23]. The power of 12 term in charge of short length repulsions has no clear physical meaning, and was probably only chosen because it can be trivially calculated by just squaring the power of 6 term.

It is worth to mention that the LJ $(12, 6)$ can be generalised to a LJ (n, m) potential, the combination $(9, 3)$ being probably the second most used after $(12, 6)$, for instance for simulating fluid–solid interactions.[24] Those potentials are in fact par of the family of the *Mie potentials*, already introduced by Gustav Mie in 1903 for describing the kinetic theory of the mono-atomic bodies.[25]

Complexity of the non-bonded calculations The evaluation of the Coulomb and Lennard-Jones potential as described above will induce a sum on all the (i, j) pairs of atoms, thus in result setting the complexity of the algorithm to $\mathcal{O}(n^2)$, n being the number of atoms on which the potentials are evaluated. Two optimisations can reduce the size n such as $n < N$:

First it is possible to build *exclusion list* by analysing the connectivity of the atoms. One can exclude non-bonded interactions for pairs of atoms for which there is already a bond, angle or dihedral term: indeed force constants for those potentials can be tuned in order to already include the non-bonded interaction.

Secondly mathematical analysis of the Coulomb and LJ potentials will show a decay to almost zero for long distances r_{ij} (see Figures 1.1 and 1.2) : hence one can choose a *cutoff* distance r_{ct} at which the interactions stop being calculated. A *Verlet list* [26] is built at the beginning of the simulation, and regularly updated, that will contain for each atom i the list of the j atoms for which there is a non-bonded interaction to calculate.

However if a simple *truncation* scheme is applied, i.e.:

$$V_{non-bonded}^{trunc} = \begin{cases} V_{non-bonded}(r_{ij}) & \text{if } r_{ij} < r_{ct} \\ 0 & \text{if } r_{ij} \geq r_{ct} \end{cases} \tag{1.23}$$

Then a discontinuity appears at $r_{ij} = r_{ct}$, the energy will not be properly conserved, and a sharp variation of the force around r_{ct} is to be expected (see Ref. [27] for a discussion concerning drawbacks of simple truncation).

The simplest approach for avoiding this discontinuity is to use a *shifting* of the potential in order to force it to be 0 at r_{ct} :

$$V_{non-bonded}^{shift} = \begin{cases} V_{non-bonded}(r_{ij}) - V_{non-bonded}(r_{ct}) & \text{if } r_{ij} \leq r_{ct} \\ 0 & \text{if } r_{ij} > r_{ct} \end{cases} \tag{1.24}$$

But this will lead to a modification of the non-bonded potential for all distances. Another approach uses a *switching* function $S(r_{ij}) \in [1; 0]$, defined using an additional *cuton* distance r_{cn}. In the CHARMM software this is for example written as:

$$S(r_{ij}) = \frac{(r_{ct}^2 - r_{ij}^2)^2 (r_{ct}^2 + 2 * r_{ij}^2 - 3 * r_{cn}^2)}{(r_{ct}^2 - r_{cn}^2)^3}$$

And the corresponding switch potential is defined as:

$$V_{non-bonded}^{switch} = \begin{cases} V_{non-bonded}(r_{ij}) & \text{if } r_{ij} < r_{cn} \\ S(r_{ij}) * V_{non-bonded}(r_{ij}) & \text{if } r_{cn} \leq r_{ij} \leq r_{ct} \\ 0 & \text{if } r_{ij} > r_{ct} \end{cases} \tag{1.25}$$

The effect of the shift and switch methods on a Coulomb potential is shown on Figure 1.3: two charges of the same sign are considered (as in Figure 1.1) for the "No cutoff" curve, and the "Shifted" and "Switched" curves correspond to application of Equations 1.24 and 1.25, using a cutoff of 12 Å for the shifted and switched curves, and a cuton of 10 Å for the switched curved.

Another possible approach is to use the *Particle Mesh Ewald* (PME) method [28], originally developed for the Coulomb potential (but also extended to the LJ potential [29]).

The non-bonded potential V_{NB} is assumed to be separable in two parts, a short-range V_{sr} evaluated traditionally on (i, j) pairs (similar to Equations 1.21 and 1.22), and a long-range part V_{sr} evaluated on discrete points k of a 3-dimensional grid (mesh):

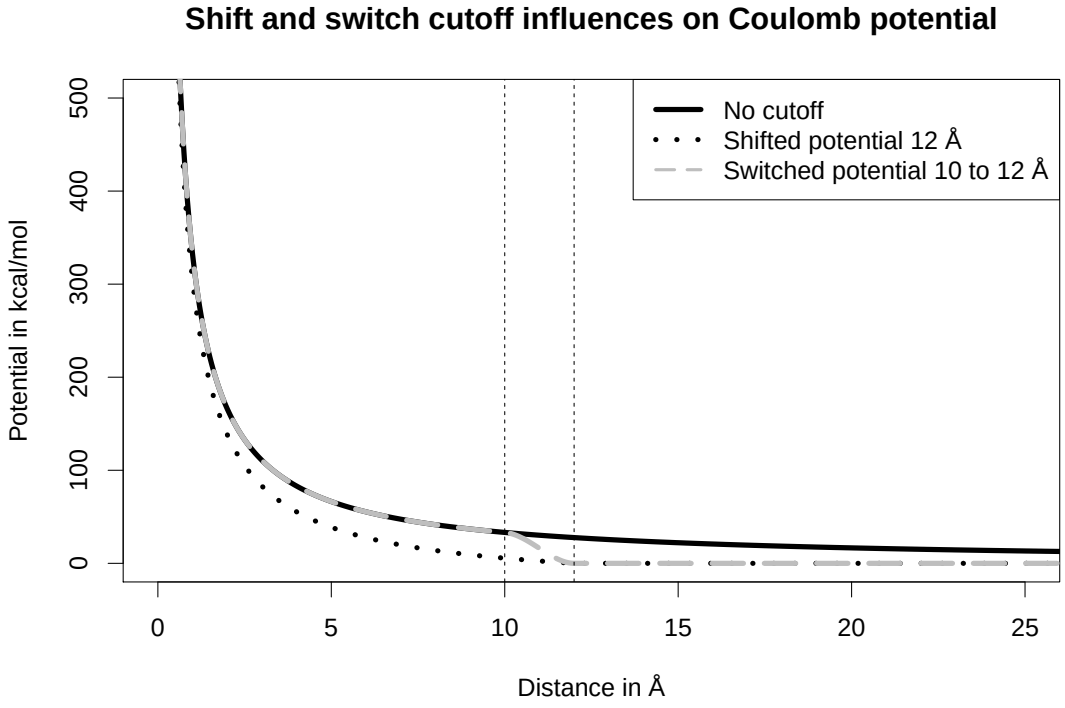


FIGURE 1.3: Effects of shifting and switching on a Coulomb potential. Cutoff at 12 Å for the shifting and switching, and cuton at 10 Å for the switching.

$$
\begin{aligned}
V_{NB} &= V_{sr} + V_{lr} \\
V_{sr} &= \sum_{(i,j)} V_{sr}(r_{ij}) \\
V_{lr} &= \sum_{k} \mathcal{F}\big(\phi_{lr}(k)\big) \; |\mathcal{F}\big(\rho(k)\big)|^2
\end{aligned}
\tag{1.26}
$$

Where $\phi(lr)$ is a modified version of the potential of interest (with addition of an interpolation feature, as potential is now evaluated on a grid, using for example *B-splines*), $\mathcal{F}(\cdots)$ denotes a *Fast Fourier Transform* (FFT), and $\rho(k)$ a density (of charge or particle) at grid point k.

When the FFT are performed with an efficient algorithm (such as ones provided by the FFTW library [30, 31]), PME method allows to reduce the total complexity of the non-bonded calculations to $\mathcal{O}(n \log n)$.

1.2.2 Bonded terms

For molecular systems made of molecules, non-bonded interactions are not enough for an accurate description of the geometry, and bonded potentials are necessary.

Bond potential The *bond potential* V_{bonds} represents the potential energy of a chemical bond between 2 atoms (i, j) at distance r_{ij}. It is usually modelled using a *harmonic potential*:

$$V_{bonds} = V(q_i, q_j) = \sum_{(i,j)-bonds} k_{ij}(r_{ij} - r_0)^2 \tag{1.27}$$

where k_{bond} is the bond potential at equilibrium distance r_0 for a given (i, j) couple.

It should be mentioned that this formula does not allow bond breaking: if such a property is required some Force Fields have the possibility to use a *Morse potential* [32] where the energy tends to the *dissociation energy* D_e for $r_{ij} \gg r_0$:

$$V_{bonds} = \sum_{(i,j)-bonds} D_e(1 - e^{-a(r_{ij}-r_0)})^2$$

and where $a = \sqrt{\frac{k_{ij}}{2D_e}}$ regulates the width of the potential well around r_0. However it should be remembered that the evaluation of an exponential term is still computationally expensive, hence the use of Morse potential is rarer.

Figure 1.4 illustrates both the Harmonic and Morse potentials using as parameters $k = 600$ kcal/mol, $r_0 = 1.23$ Å and $D_e = 120$ kcal/mol, which can be used for modelling dioxygen.

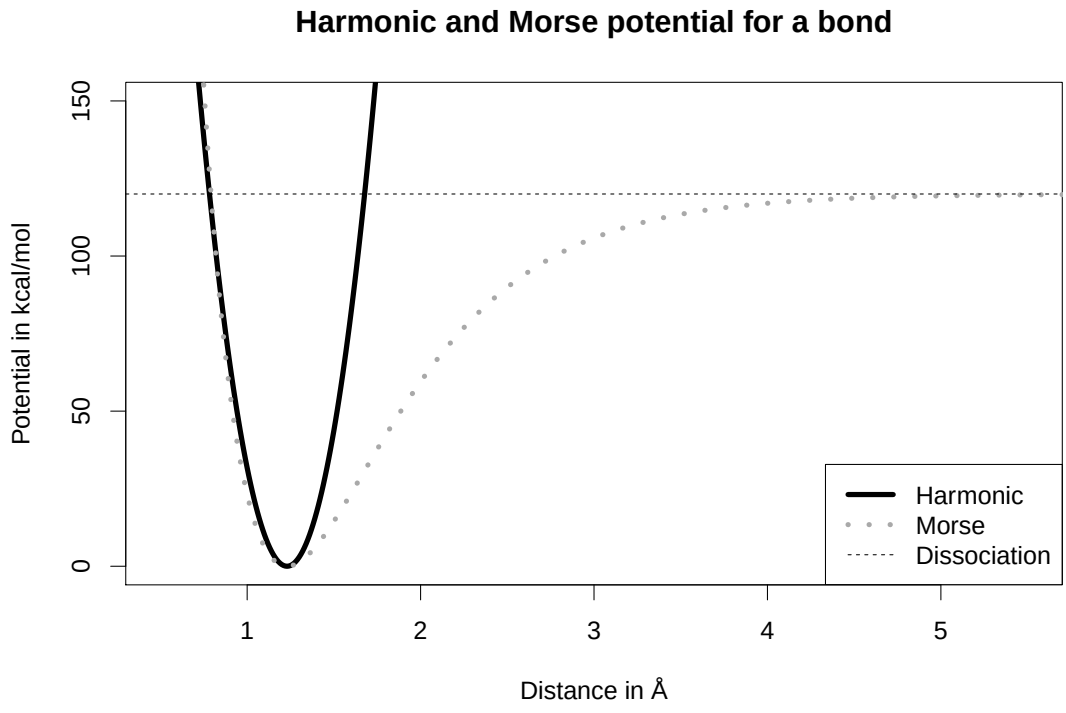


FIGURE 1.4: Illustration of the Harmonic and Morse bond potentials, using $k = 600$ kcal/mol, $r_0 = 1.23$ Å and $D_e = 120$ kcal/mol

Angle potential The *angle potential* V_{angles} is another essential term for reproducing accurately the geometry of a molecule. It is defined for triplets (i, j, k) of atoms, and also modelled using an harmonic potential:

$$V_{angles} = V(q_i, q_j, q_k) = \sum_{(i,j,k)-angles} k_{ijk}(\theta_{ijk} - \theta_0)^2 \tag{1.28}$$

where k_{angles} is the angle potential at equilibrium angle θ_0 between the triplet (i, j, k). The angle θ_{ijk} can be trivially measured using the arc-cosine of the dot product between normalised vectors $\vec{ij}$ and $\vec{jk}$:

$$\theta_{ijk} = \arccos\left(\frac{r_{ij}}{|r_{ij}|} \cdot \frac{r_{jk}}{|r_{jk}|}\right)$$

Some FFs, such as CHARMM, also define the *Urey-Bradley* potential, which can be considered as a "virtual" bond between atoms (i,k) of a given (i,j,k) angle, modelled using a bond-like potential: $V_{UB} = \sum_{(i,k)-UB} k_{UB}(U_{ik} - U_0)^2$. This is useful for adding an extra rigidity to some (i, j, k) angles.

Torsion potentials The *torsion potentials* $V_{torsion} = V_{dihedrals} + V_{impropers}$ are also used in order to constraint the geometry of a molecule. *Dihedral angles* $\phi_{i,j,k,l}$, defined for a set of four atoms (i, j, k, l), correspond to the angle between the two planes $[ijk]$ and $[jkl]$. When following the IUPAC/IUB convention[1] for the definition of planes and the sign of the dihedral, then ϕ_{ijkl} is calculated using:

$$\phi_{ijkl} = -\arccos\left(\frac{r_{ij} \times r_{jk}}{|r_{ij} \times r_{jk}|} \cdot \frac{r_{jk} \times r_{kl}}{|r_{jk} \times r_{kl}|}\right) \tag{1.29}$$

and the *dihedral potential* is written as:

$$V_{dihedrals} = V(q_i, q_j, q_k, q_l) = \sum_{(i,j,k,l)-dihedrals} k_{ijkl}\big(1 + \cos(n_{ijkl}\phi_{ijkl} - \phi_0)\big)$$

where k_{ijkl} is a force constant, $\frac{\phi_0}{n_{ijkl}}$ the angle range between a minimum and a maximum, and the additional *multiplicity* term n_{ijkl} is added which corresponds to the number of energy minima observed when ϕ is rotated over 360°. Hence ϕ_0 is the equilibrium value of the lowest minima. See Figure 1.5 for an illustration with test parameters.

For accurately reproducing the planarity of some molecules, most of the FFs also define a special type of dihedrals, the *impropers* angles. Let us consider the case of the nitrate ion, NO_3: because of the de-localisation of the double bond on the 3 NO bonds it exhibits on average a trigonal planar geometry. An improper for this molecule would be defined by assigning to the central N atom the rank i, and to the three O atoms ranks j, k, l clockwise. With such a definition the above defined dihedral angle would be zero or approximately zero. Therefore FFs usually define the *improper potential* energy $V_{impropers}$ using an harmonic equation, characterised by an improper angle ω_{ijkl}, a force constant k_{ijkl} and an equilibrium value $\omega_0 \approx 0°$:

$$V_{impropers} = V(q_i, q_j, q_k, q_l) = \sum_{(i,j,k,l)-impropers} k_{ijkl}(\omega_{ijkl} - \omega_0)^2$$

Additional stability terms For an accurate treatment of large biomolecules, such as poly-peptides, proteins, enzymes…the above detailed potentials are sometime not enough for assuring the stability of the system on a long timescale. Therefore additional potential were introduced, such as the *CMAP*[33], a grid-based energy correction on the (ϕ, ψ) dihedral terms of all backbone type residues:

$$V_{CMAP} = \sum_{(\phi,\psi)-residue\ 1}^{(\phi,\psi)-residue\ N} f_{CMAP}(\phi, \psi) \tag{1.30}$$

For a definition of $f_{CMAP}(\phi, \psi)$ and details concerning the procedure, refer to [33]. It was shown that CMAP additions minimise the root mean square fluctuations when compared to Nuclear Magnetic Resonance (NMR) experiments,[34] and that their role is essential in order to maintain the folded states of the protein stable over long (several tens of nano-seconds) simulations.

[1] http://www.chem.qmul.ac.uk/iupac/misc/ppep1.html

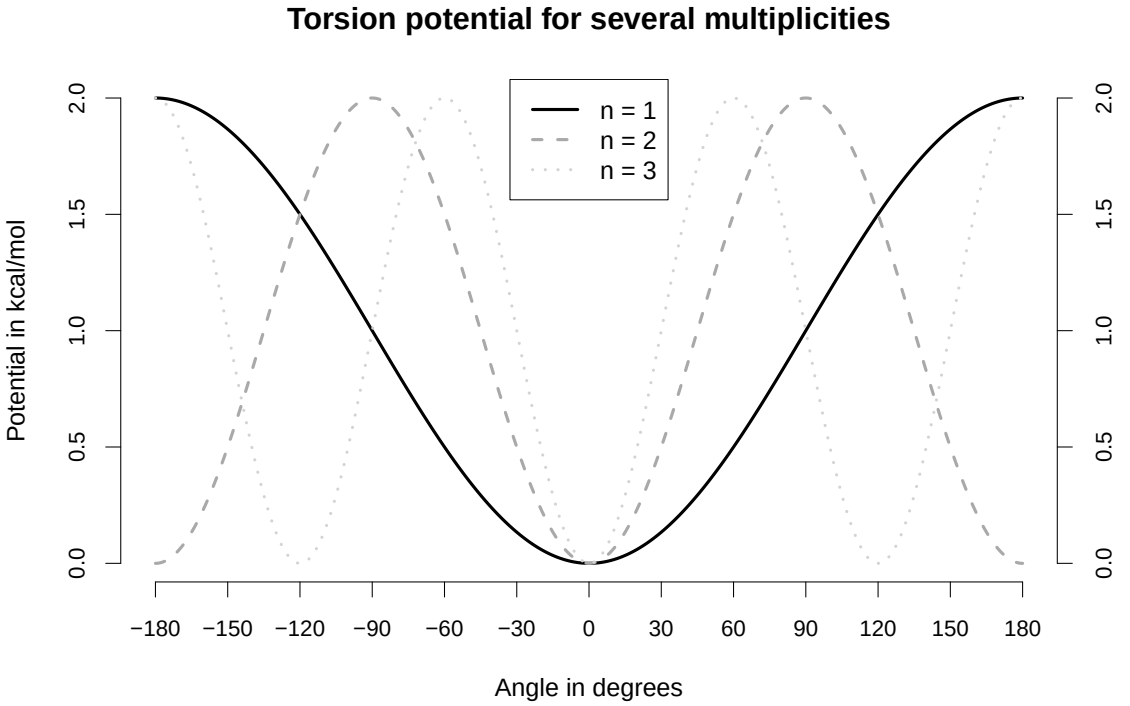


FIGURE 1.5: Illustration of the dihedral potential, built arbitrarily with $k_{ijkl} = 2$, $\phi_0 = 180°$ and $n = \{1, 2, 3\}$

1.2.3 Atom typing and fitting of the parameters

Because of the empirical aspect of formula 1.6, parameters such as force constant, equilibrium length/angle for the bonded terms and charge, equilibrium distance and well depth for the non-bonded potentials have to be accurately defined, and adjusted if necessary: this procedure is called force field fitting.

It is also extremely important, when designing a force field, to define for the same chemical type different *atom types*: this helps taking into account the various connectivities and hybridisation level that a chemical type can exhibit, and also the effect of the local environment and neighbours of a given atom. It is for example common to find in most of the force fields up to ≈ 20 possible types for the carbon C^{12} chemical atom, in order to reproduce the sp^3, sp^2 and sp hybridisations, different partial charges caused by an electronegative neighbour such as N or O, or protein backbone structural types, etc…

The following paragraphs gives a (non-exhaustive) list of sources for fitting parameters.

Bonded parameters The first main source of knowledge comes from experiments: X-ray crystallography, NMR, spectroscopic methods…are an example of the many methods available for determining with precision the structure of a molecule, and thus an accurate value for a bond length, or an angle. A second important source comes from results of ab-initio calculations performed with high accuracy methods, from which equilibrium values of bond stretching and angle bending can be extracted. Using results from experiments and ab-initio calculations, one can also estimate the rotational barriers and thus fit the torsion parameters.

Non-bonded parameters The accurate fitting of the non-bonded parameters is probably one of the most challenging task when elaborating a force field. Charges and Lennard-Jones parameters can be easily extracted from an ab-initio calculation of isolated molecules in gas phase, but when

calculating parameters for compounds usually found in a liquid state, those parameters will most likely not reproduce thermodynamic observables that one can estimate using an ensemble average approach as introduce with Equations 1.8 or 1.9. Thus a serie of parameters optimisation followed by simulations for performing the average is required, until observables of interest are reproduced. This will be detailed in Chapter 5 where the development of a tool assisting the user in the fitting of non-bonded parameters for the CHARMM force field is presented.

1.2.4 The CHARMM Force Field

The CHARMM c36 forcefield implements all the above mentioned (Equation 1.6) bonded and non-bonded terms, including the CMAP terms, crucial for stability of macromolecules over long simulation times. It is distributed with the CHARMM software, but can also be used from other packages such as GROMACS, OpenMM…or can also be downloaded directly.[2] It exists in several versions, with optimised parameters tuned for a given application: proteins, nucleic acids, polymers, carbohydrates, ethers, lipids, small molecules (CGenFF[35, 36])…Thanks to the large community of contributors, the CHARMM FFs supports many extensions: Reactive MD [37], Polarisability [8, 10], atomic Multipoles Expansion (discussed later in Chapter 5)…

Now that Statistical Mechanics principles and Force Fields have been properly introduced, it is possible to detail the key concept of *Free Energy* in Section 1.3.

1.3 The importance of Free Energy Estimation

From a macroscopic thermodynamic point of view, Free Energy is a quantity representing the internal energy of a system available for performing a work. This quantity is of great importance in Computational Chemistry and Biology nowadays, where free energy difference between two states, or free energy change, is usually evaluated. It can for instance be used for plotting free energy profiles (1-dim), surfaces (2-dim) or grids (3-dim) which allow an easy visualisation of conformational changes, or for estimating ligand binding affinities in a given protein.

Because of the rich history of the development of thermodynamics in the nineteenth century, different names were given to the free energy depending on the ensemble: *Helmholtz free energy* in the NVT ensemble, denoted as A or F (F is used in the current work). *Gibbs free energy* in the NPT ensemble, denoted as G. And *Grand Potential* (or also *Landau potential*) for the Grand canonical ensemble μVT, often denoted as Ω_G.

In the following mathematical definition of absolute free energy (Subsection 1.3.1) and free energy difference (Subsection 1.3.2) will be introduced.

1.3.1 Absolute Free Energy

The absolute free energy is defined as the amount of available internal energy for a system, from all the possible microstates ; as the partition function Z_* for a given ensemble is defined for the whole phase space Ω (i.e. $(q,p) \in \Omega$), the absolute free energy is defined as:

$$F = -\frac{1}{\beta} \ln Z_* \tag{1.31}$$

where Z_* can be any of the above defined partition function. In the following we will focus on the canonical ensemble Z_{NVT} (see Equation 1.14):

$$F = -\frac{1}{\beta} \ln \left(\int_\Omega \exp\big(-\beta H(q,p) \big) \, d\Omega \right)$$

[2] http://mackerell.umaryland.edu/charmm_ff.shtml

If one remembers the previously detailed Equations 1.15 and 1.16, Z_{NVT} can be split in two parts, i.e. kinetic and potential contributions, the first one being easily sampled in simulations by assigning random velocities respecting the Maxwell-Boltzmann distribution. Thus the computational challenge in order to estimate the absolute free energy is usually to sample properly the configuration space $\mathcal{D}$ through the sampling of the p.d.f $\rho(q) \propto \exp(-\beta V(q))$.

1.3.2 Free Energy differences

Although the estimation of the absolute free energy of a system might of interest in some fields of research, in Computational Chemistry/Biology one usually investigates free energy difference between two states. Considering two states A and B one can define the free energy difference (or relative free energy) ΔF as:

$$\Delta F = \Delta F_{A \to B} = F(B) - F(A) \tag{1.32}$$

$$\Delta F = -\frac{1}{\beta} \ln \int_{\Omega} \exp\Big(-\beta \Big(H\big(q^B, p^B\big) - H\big(q^A, p^A\big) \Big) \Big) \, d\Omega \tag{1.33}$$

where $H(q^A, p^A) \neq H(q^B, p^B)$ are two distinct states from the phase space Ω.

Once again, by considering the Hamiltonian separable, and by assuming that by use of the Maxwell-Boltzmann distribution it is possible to obtain $p^A = p^B = p$ such as $H(q^A, p) \neq H(q^B, p)$, then Equation 1.33 simplifies to:

$$\Delta F = -\frac{1}{\beta} \ln \int_{D} \exp\Big(-\beta \Big(V\big(q^B\big) - V\big(q^A\big) \Big) \Big) \, dq \tag{1.34}$$

1.3.3 Methods for computing free energy differences

In the following Chapters 3 – 4 – 5 *Free Energy Surfaces* (FES) (or *Free Energy Grids*) are built for visualising either conformational changes (Chapters 3 – 4) or protein-ligand interactions (Chapter 5), usually for validating a newly implemented sampling method. In all cases the free energy differences are obtained using an Histogram Method, and for the case of Chapters 3 – 4 one or two *reaction coordinates* were used: those two concepts are detailed below.

Defining a reaction coordinate A *reaction coordinate* $\xi(q)$ is usually defined on the set $D \subseteq \mathbb{R}^{3N}$ of the possible atomic coordinates. The idea is to find a subset $X \subset D$ of cardinality m satisfying $m << 3N$:

$$\xi(q) : D \to \mathbb{R}^m$$

i.e. to find a new set of coordinates of reduced dimensionality, thus easier to sample, but that still exhibits a free energy difference ΔF^{ξ} close to the original ΔF in order to provide a meaningful free energy estimation. In mathematical terms, by modifying Equation 1.34 one can write:

$$\Delta F \approx \Delta F^{\xi} = -\beta^{-1} \ln \int_{X} \exp\Big(-\beta \Big(V\big(\xi(q^B)\big) - V\big(\xi(q^A)\big) \Big) \Big) \, d\xi \tag{1.35}$$

Example of possible reaction coordinate include: dihedral angle or distance between two groups of atoms of interest, or a mapping of the coordinates to a scoring function for measuring a folding process...See Chapters 3 – 4.

Histogram methods The histogram method consists in a discretisation of Equation 1.35: if assuming that the partition function is accurately sampled, and that the states distribution follows the canonical p.d.f ρ (see Equation 1.13), i.e. $\rho\big(\xi(q)\big) \propto \exp\Big(-\beta V\big(\xi(q)\big)\Big)$, then for N observations:

$$\Delta F_i = F_i - F_0 = -\beta^{-1} \ln \sum_{n=1}^{N} \rho\big(\xi_i(q^n)\big) \tag{1.36}$$

Where ΔF_i is the free energy difference for bin i of the histogram associated to values ξ_i. F_0, the bin with the lowest free energy, is usually substracted in order to provide a free energy difference, i.e. the bin F_0 with a value of 0 for its free energy will represent the most stable discretised ξ_i.

Other advanced methods Several advanced methods have been developed for facilitating estimation of free energy differences, and one can give as a (non-exhaustive) list.

Sampling methods that do not require reaction coordinates and which are "bias free', such as Metropolis-Hastings MC or MD can be used with the simple histogram method, if the sampling task is straightforward enough. Related methods such as Parallel Tempering/Replica Exchange [38–41] may also allow provide an additional sampling boost.

Free energy perturbation,[42] thermodynamic integration,[43] are usually useful for estimating simple properties such as solvation free energy.

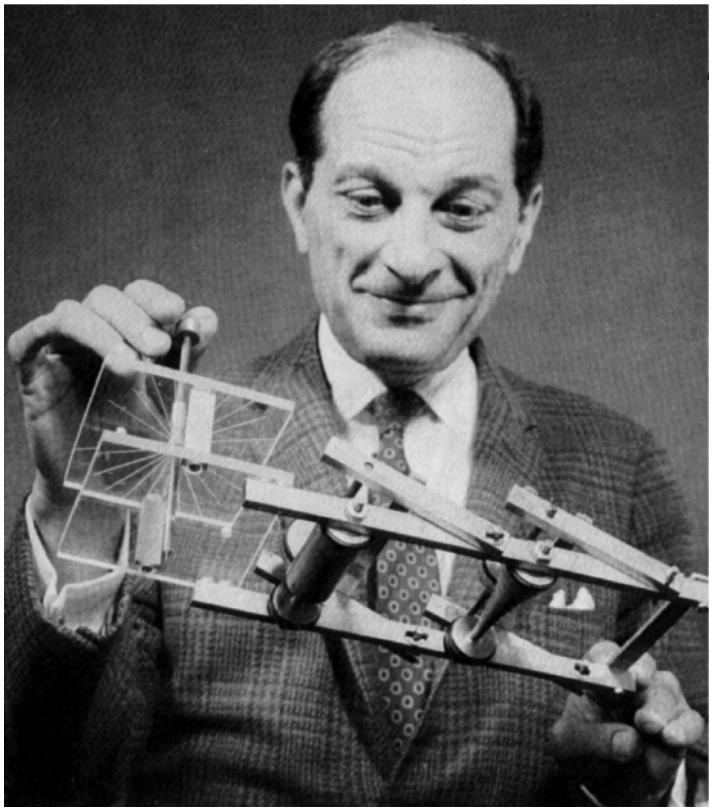
When investigating conformational changes or ligand bindings, methods such as umbrella sampling[44] (combined with the WHAM[45] method), metadynamics,[46–49] adaptive biasing dynamics methods (ABF,[50, 51] Wang-Landau[52] method) …are of interest. Those methods have in common the requirement to define *a priori* one or more reaction coordinate in order to guide the sampling, through the definition of a biasing potential.

In the following Chapter 2, after an introduction to standard sampling methods (Section 2.2) (such as MC and MD methods), and a few words on above mentioned advanced sampling methods (Section 2.3), two *Rare Events sampling methods* are introduced: Spatial Averaging Monte Carlo (SA-MC, Section 2.4) and Partial Infinite Swapping (PINS, Section 2.5). Although the primary motivation for the development of those two methods was not free energy estimation, they were shown to be reliable enough for producing FES using the histogram method.

Chapter 2

Sampling Methods

"*It is not so much whether a theorem is useful that matters, but how elegant it is.* "



Stanisław Marcin Ulam , Adventures of a Mathematician (1991), Chapter 15, Random Reflections on Mathematics and Science, p. 274

2.1 Overview

In this Chapter, MD and MC methods are first introduced in Sections 2.2. Then a brief review of some of the advanced sampling methods is given in Section 2.3.

Spatial averaging MC (SA-MC) sampling belongs to the family of the enhanced MC methods, where a new family of probability density functions are constructed from the introduction of a biasing term.[53] Until now, SA-MC has been applied to model systems and in special applications[53, 54] such as the diffusion of small molecules in condensed phase environments. But no public implementation of SA-MC was available until 2014: indeed, the first part of this thesis work consisted in implementing SA-MC in CHARMM in such way that it could be used by the community for efficiently sampling conformational space of biomolecules. CHARMM implementation will be detailed in Section 2.4, and results commented in Chapter 3.

Another method which has recently been investigated is Partial Infinite Swapping (PINS)[55–59] which is based on the PT/RE algorithms. PINS uses a symmetrisation strategy for combining probability distributions at different temperatures, so that they become more connected and thus easier to sample than the original ones. As for SA-MC, several INS/PINS articles described the algorithms but never provided a "ready-to-use" module in any MD package: this was also an important part of this thesis work, and since February 2016 a MD based PINS algorithm, based on an existing CHARMM ENSEMBLE module, was made available. Implementation details will be provided in Section 2.5, while for validation and investigation, one should refer to Chapter 4.

But first of all the following paragraph introduces a simple test potential used for illustrating the sampling gain effect of some of the below detailed methods.

Study case: the double well potential A double well potential will be used for illustrating some methods of interest. It is defined as:

$$f : x \to (x^2 - \sqrt{\lambda})^2 \tag{2.1}$$

Its derivative is:

$$f' : x \to 4x(x^2 - \sqrt{\lambda}) \tag{2.2}$$

Note that $f' = 0$ if $x = 0$ or $x = \pm\sqrt{\lambda}$.

Table 2.A represents the variation table of the function f: one can see that the Equation 2.1 allows to build a symmetric potential with the following properties: Two minima located at $(x, V(x)) = (\pm\sqrt[4]{\lambda}, 0)$, and one maximum at $(x, V(x)) = (0, \lambda)$.

x	$-\infty$		$-\sqrt[4]{\lambda}$		0		$\sqrt[4]{\lambda}$		$+\infty$
f'		$-$	0	$+$	0	$-$	0	$+$	
f	$+\infty$	↘	0	↗	λ	↘	0	↗	$+\infty$

Table 2.A: Variation table for the double well potential, and its derivative, respectively defined using Equations 2.1 and 2.2.

For the rest of this chapter we chose $\lambda = 1$, thus imposing a barrier height of 1, and two minima located at $\pm$ 1. See Figure 2.1 for an illustration.

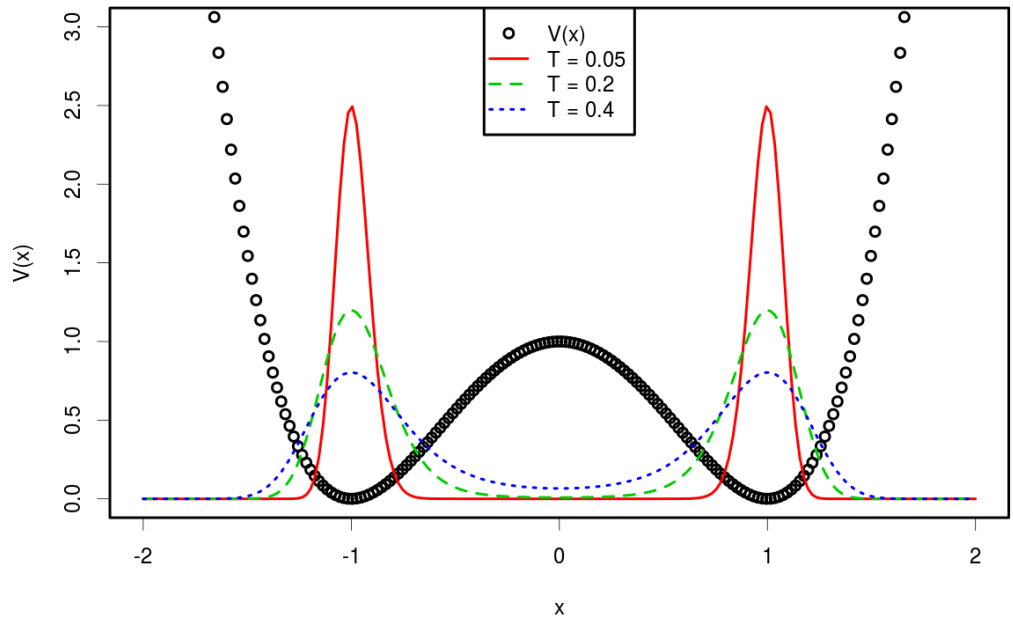


FIGURE 2.1: Plot of the potential $V(x) = (x^2 - 1)^2$ (dots), and of three ideal p.d.f. $\rho(x) = e^{-\beta V(x)}$ (we assume a canonical like distribution) at $T = \{0.05, 0.2, 0.4\}$ (arbitrary units used for $x, V(x), T$ and $\beta = 1/T$)

2.2 Standard Sampling Methods

2.2.1 Molecular Dynamics

In Chapter 1 we introduced the general expression of the potential energy $V(q)$ using Equation 1.6, and mentioned at the beginning of Section 1.2 that in Molecular Dynamics (MD) one can get the force F using the gradient of $V(q)$.

This can be generalised to the Hamiltonian $H(q, p)$ by introducing the concept of Hamiltonian Dynamics.

Hamiltonian Dynamics For a time-dependent Hamiltonian $H\big(q(t), p(t)\big)$, time evolution of an isolated (NVE) system is described by the following first-order differential equations:

$$\begin{cases} \dfrac{dq(t)}{dt} &= +\left(\dfrac{\partial H\big(q(t), p(t)\big)}{\partial p}\right)_q \\ \dfrac{dp(t)}{dt} &= -\left(\dfrac{\partial H\big(q(t), p(t)\big)}{\partial q}\right)_p \end{cases} \tag{2.3}$$

Denoting a partial derivatives with the following formalism: $\nabla_x f(x, y) = \left(\dfrac{\partial f}{\partial x}\right)_y$, and assuming in the following an implicit time dependence, i.e. that $H(q, p) = H\big(q(t), p(t)\big)$ the previous equation is re-written as:

$$\begin{cases} \frac{dq}{dt} & = +\nabla_p H(q,p) \\ \frac{dp}{dt} & = -\nabla_q H(q,p) \end{cases} \tag{2.4}$$

By remembering Equation 1.1 (Hamiltonian is separable) and the fact that the kinetic energy is $K(p) = \frac{1}{2} p^T M^{-1} p \equiv \frac{1}{2} M^{-1} p^2$, one can evaluate the two ∇ terms as :

$$\nabla_p H(q,p) = \nabla_p (K(p) + V(q)) = \frac{1}{2} M^{-1} 2p^{2-1} + 0$$

and

$$\nabla_q H(q,p) = \nabla_q (K(p) + V(q)) = 0 + \nabla V(q)$$

Then Equation 2.4 rewrites as (with time dependence re-introduced for clarity):

$$\begin{cases} \frac{dq(t)}{dt} & = M^{-1} p(t) \\ \frac{dp(t)}{dt} & = -\nabla V\big(q(t)\big) \end{cases} \tag{2.5}$$

This set of two first-order ordinary differential equations can be solved if initial conditions (q^0, p^0) at $t = 0$ are provided.

Finally one can introduce the transformation ω_t, the flow of the Hamiltonian dynamics (or Hamiltonian vector field) that from initial conditions (q^0, p^0) leads to the state of the system at time $t \in \mathbb{R}$ as :

$$\big(q(t), p(t)\big) = \omega_t(q^0, p^0) \tag{2.6}$$

Furthermore the flow ω_t obeys the function composition rules, i.e. $\omega_{t+u} = \omega_t \circ \omega_u = \omega_t(\omega_u)$.

Newtonian notation In MD it is usual to reformulate Equation 2.5 in terms of the positions q and potential $V(q)$ only: in that case one obtains:

$$M \frac{d^2 q(t)}{dt^2} = -\nabla V\big(q(t)\big) \tag{2.7}$$

The term $-\nabla V\big(q(t)\big)$ being the force F, and the second derivative of positions $\frac{d^2 q(t)}{dt^2}$ being the acceleration a, Equation 2.7 reads as Newton's second law:

$$F_i = m_i * a_i$$

for any particle i.

Poisson brackets notation Equation 2.4 can be reformulated using *Poisson brackets* .

Poison brackets are defined, for two functions $f(q,p)$ and $g(q,p)$ as:

$$\{f, g\} = \nabla_q f \ \nabla_p g \ - \ \nabla_p f \ \nabla_q g$$

Introducing the notations $\{q, H\}$ and $\{p, H\}$ such as:

$$\begin{aligned}\{q, H\} &= \nabla_q q \ \nabla_p H \ - \ \nabla_p q \ \nabla_q H \\ &= 1 * \nabla_p H \ - \ 0 * \nabla_q H = \nabla_p H\end{aligned}$$

and

$$\begin{aligned}\{p, H\} &= \nabla_q p \ \nabla_p H \ - \ \nabla_p p \ \nabla_q H \\ &= 0 * \nabla_p H \ - \ 1 * \nabla_q H = -\nabla_q H\end{aligned}$$

One obtains:

$$\begin{cases} \dfrac{dq}{dt} &= \{q, H\} \\ \dfrac{dp}{dt} &= \{p, H\} \end{cases} \tag{2.8}$$

Furthermore Equation 2.8 can be reformulated in one line by introducing the function $\gamma = q$ or $\gamma = p$:

$$\frac{d\gamma\big(q(t), p(t)\big)}{dt} = \{\gamma, H\}\big(q(t), p(t)\big) \tag{2.9}$$

This formulation will reveal itself really convenient for illustrating some key properties of Hamiltonian dynamics in the next paragraph.

Essential properties of Hamiltonian and Newtonian dynamics Molecular dynamics in the NVE ensemble, as governed by Equations 2.5 and 2.7, exhibit the following intrinsic properties:

Energy conservation: Using the Poisson brackets reformulation of Equation 2.9 and choosing $\gamma = H$ one can deduce the following equality:

$$\begin{aligned}\frac{dH\big(q(t), p(t)\big)}{dt} &= \{H, H\} \\ &= \nabla_q H \ \nabla_p H \ - \ \nabla_p H \ \nabla_q H \\ \frac{dH\big(q(t), p(t)\big)}{dt} &= 0\end{aligned} \tag{2.10}$$

which implies that

$$H\big(q(t), p(t)\big) = H\big(q^0, p^0\big) = \text{const} \tag{2.11}$$

Equations 2.10 and 2.11 thus naturally impose the total energy conservation through the constance of the Hamiltonian over the whole simulation time.

Momenta conservation From the second line of Equation 2.5 one writes the variation of the momentum p_i of a particle i as:

$$-\nabla V = F = \frac{dp_i}{dt}$$

According to Newton's third law, the forces between two particles are equal and opposite, so introducing a second particle j one can write:

$$\frac{dp_i}{dt} = -\frac{dp_j}{dt}$$

Generalising to N particles and summing the momenta in P this leads to:

$$P = \sum_{i=1}^{N} p_i = \text{constant}$$

Reversibility: Now let us define a function $R(q,p) = (q,-p)$ that "reverses" the momenta: Equation 2.8 becomes:

$$\begin{cases} \frac{dq}{dt} & = \{q, H\} \quad = +\nabla_p H \\ \frac{dp}{dt} & = \{-p, H\} = +\nabla_q H \end{cases} \tag{2.12}$$

By just inverting the sign of the forces one can, from any time $t > 0$ come back to the initial starting point $H\left(q^0, p^0\right)$.

However, exact reversibility and energy conservation are computationally impossible to reach for long enough simulations because of the following computer limits: (i) First, there is an intrinsically limited precision for the floating points numbers used for storing values of (q,p): for double precision floating points, numbers are stored on 64 bits, 1 for the sign, 11 for the exponent and 52 for storing the fractional representation of the number, thus the smallest difference between two real numbers, called the *machine epsilon*, is of the order of $2^{-52} \approx 10^{-16}$. There is nowadays the possibility to improve the precision by using extended double precision (80 bits) or quadruple precision (128 bits) floating point numbers (standard IEEE 754-2008[60]), resulting in a rounding error of respectively 2^{-63} and 2^{-112}, but they are rarely used as they imply extended computation time. (ii) Second, there is also a mathematical error when discretising integrals or gradients when performing numerical integration/differentiation: this error tends to 0 for an infinitely small integration step δt, so this error can be controlled and estimated by a proper choice of δt, but never avoided.

The choice of an appropriate numerical integration is discussed in the following paragraph.

Symplecticity and Liouville theorem: The flow ω_t defined by Equation 2.6 is *symplectic*: for all $t \in \mathbb{R}$ and any subset S of the phase space Ω there is volume preservation, i.e.

$$V = \int_{\omega_t(S)} dq\ dp = \int_S dq\ dp = \text{constant}$$

This idea is generalised by the *Liouville theorem* :

The probability $\rho(q(t), p(t))\ dq\ dp$ to find the system in state (q,p) is constant over time:

$$\frac{d\rho}{dt} = \frac{\partial \rho}{\partial t} + \sum_{i=1}^{3N} \left(\frac{\partial \rho}{\partial q_i} \frac{\partial q_i}{\partial t} + \frac{\partial \rho}{\partial p_i} \frac{\partial p_i}{\partial t} \right) = 0 \tag{2.13}$$

Symplecticity is an important property for stability of long term dynamics: an accurate integrator should preserve as possible the volume V above defined in order to satisfy long term validity of Equation 2.13.

Numerical integration As stated above, the use of an accurate and *symplectic integrator* is essential for the ability to perform long time MD simulations.

Let us introduce the time step δt, which can be seen as the "time-resolution" of the discretisation induced by the numerical integration. δt is also the smallest time difference between two observation of a time dependent property (such as the Hamiltonian), or between two applications of the flow ω defined in Equation 2.6.

Loup Verlet introduced in 1967 [26] the Verlet integration scheme for use in MD simulations (however a similar integration scheme was already used by Carl Størmer [61] at the beginning of the 20th century in astrophysics, thus the algorithm is also named Størmer's method). The integrator is shown to be time-reversible, symplectic and to provide a good numerical stability (discussed later). In the following, the *Velocity Verlet* variant,[62] which became the *de facto* choice for integration in MD, is presented.

Introducing (q^n, p^n) to be the state of the system at time $t_n = t_0 + n\delta t$, it is written:

$$\begin{cases} p^{n+\frac{1}{2}} & = p^n - \frac{\delta t}{2}\nabla\big(V(q^n)\big) \\ q^{n+1} & = q^n + \delta t M^{-1} p^{n+\frac{1}{2}} \\ p^{n+1} & = p^{n+\frac{1}{2}} - \frac{\delta t}{2}\nabla\big(V(q^{n+1})\big) \end{cases} \tag{2.14}$$

It can be shown that the general order of the error is $\mathcal{O}(\delta t^2)$ for an ideal harmonic potential. In MD applications the rule of thumb is usually to choose δt to be a fraction of the fastest bond vibration of the system of interest: C-H bonds are usually the ones with the fastest bond stretching frequency, ≈ 3000 cm^{-1}, which gives a frequency of ≈ 90 THz, therefore a period of ≈ 11 fs. Thus it is common to choose $0.5 \leq \delta t \leq 2$ fs, 1 fs being the most common choice.

Generalisation to other ensembles In the case of NPT or NVT simulation, the energy conservation property is lost but the Hamiltonian equations are still valid. However, supplementary steps are required fore regulating the Temperature (and Pressure if necessary). This is done using Thermostats and Barostats. One should choose with care between the different algorithms available, as in some cases the Canonical distribution (Equation 1.13) or Isobaric-isothermal distribution (Equation 1.19) are not always properly sampled.

Thermostats

The *Andersen thermostat*,[63] by Hans Andersen, was the first proposed extension of MD from the NVE to the NVT ensemble: it couples the system to a heat reservoir and occasional exchange of energy is performed (through a rescaling of the momenta) between both: the number of particles impacted and the frequency of exchange $P(t)$ are randomly chosen but follow a Poisson distribution:

$$P(t) = \nu e^{-\nu t} \quad , \quad \nu \propto \frac{\kappa V^{1/3}}{3 k_B N} \tag{2.15}$$

where V is the volume of the system, and κ the thermal conductivity. The stochastic exchange of energy can be interpreted as "collisions" between the system and the reservoir, during which a part of the kinetic energy is exchange between both. When a collision occurs a random number $n \leq N$ of particle will have their momenta rescaled following the Maxwell-Boltzmann distribution (Equation 1.16).

This model allowing fluctuations around an equilibrium value, it can be considered as sampling properly the NVT ensemble. However it was shown that it is not appropriate for studying time-dependent properties such as diffusion coefficients.

The *Berendsen thermostat*,[64] by Herman Berendsen, is another approach, which directly rescales the momenta in the Hamiltonian at each simulation step: using Equation 2.5 an extra term is added to the $\frac{dp}{dt}$ term:

$$\begin{cases} \frac{dq}{dt} & = M^{-1} p \\ \frac{dp}{dt} & = -\nabla V + \lambda p \end{cases} \tag{2.16}$$

where

$$\lambda = \sqrt{1 + \frac{\delta t}{\tau}\left(\frac{T_0}{\overline{T}} - 1\right)}$$

and where τ is a coupling time constant which determines how quickly the target temperature T_0 is reached. The current average temperature is obtained from the total kinetic energy, $\bar{T} = \frac{p^T M^{-1} p}{3Nk_B}$.

The Berendsen thermostat is probably the most robust one for equilibrating a system at the beginning of a simulation (i.e. in order to obtain a distribution of $\bar{T}$ fluctuating over T_0). However the fact that values of p are directly rescaled inside the Hamiltonian breaks the canonical rules, and the sampled ensemble is close to but not exactly equal to an NVT ensemble. Therefore after production it is recommended to switch either to the Andersen or the below detailed Nosé-Hoover thermostat.

The *Nosé-Hoover thermostat*[65] was initially introduced by Shuichi Nosé and further improved by William Hoover.[66] This is the most rigorous approach as it adds extra degrees of freedom to the Hamiltonian while still maintaining a strict canonicality. A new degree of freedom, ζ is added to the Hamiltonian from Equation 2.5, modified as following:

$$\begin{cases} \frac{dq}{dt} &= M^{-1}p \\ \frac{dp}{dt} &= -\nabla V - \zeta p \\ \frac{d\zeta}{dt} &= \frac{1}{Q}\left(p^T M^{-1} p - (3N+1)k_B T\right) \end{cases} \tag{2.17}$$

where $3N + 1$ emphasises that the Nosé-Hoover thermostat acts on the $3N$ degrees of freedom of p plus one corresponding to ζ, and Q can be interpreted as a fictive mass for the extra degree ζ.

Barostats

In the case of a MD simulation running in an isobaric-isothermal ensemble (NPT), a thermostat keeping the temperature stable is still required, together with an extra algorithm that will keep the pressure stable, a *barostat* . This is important as the NPT ensemble represents experimental conditions in a laboratory where pressure is considered constant.

A *Berendsen barostat* can be defined similarly to the definition of the Berendsen thermostat:

$$L = 1 - \frac{\delta t}{\tau}(P_0 - \bar{P}) \tag{2.18}$$

Where P_0 is the desired pressure, P the system pressure estimated from the *Virial Theorem*, and τ a relaxation constant. Then the measure L is used for rescaling both the periodic box dimension and the coordinates q.

Without further detailing, one can also mention that there exist *stochastic barostats* [67]. The *Parrinello-Rahman barostat* [68–70] is also a commonly encounter method, whereas for the Nosé-Hoover thermostat the Hamiltonian in modified in such a way that the Equation 1.19 p.d.f remains valid. The main advantage of the Parrinello-Rahman barostat is that it does not only allows volume to fluctuate, but it can also modify the shape of the periodic box during simulation, useful for describing phase changes in solids simulations.

2.2.2 Monte Carlo sampling

Monte Carlo (MC) (sometimes generalised as Markov Chain Monte Carlo (MCMC)) methods[71] are widely used in modern computer simulations to study high-dimensional, many-body systems.

They idea is to populate the definition domain of a high dimensional integral (such as the partition function 1.15), in order to be able to apply the ensemble averaging described by Equations 1.7 – 1.8. For that a set of proposal configurations following the p.d.f. of the ensemble of interest is generated randomly.

Monte Carlo molecular simulations are usually based on the Metropolis-Hastings algorithm , based on the construction of a *Markov chain* of states. The two notions are introduced below, starting with Markov chains.

Markov chain Let us consider once again the possibility to separate the Hamiltonian and to sample only the distribution $\nu(q)$ defined on a configurational space D, for the NVT ensemble (Section 1.1.2), and assume its normalisation by application of Equation 1.11. Let q^i and q^j be two atomic configurations, and let $\nu_i = \nu(q^i)$ (and respectively ν_j) be the normalised probability measure to observe such state.

A Markov chain is a succession of states $x = (x^1, \cdots, x^n)$ obeying the following conditional probability:

$$P(x^{n+1} = q^{n+1} | x^1 = q^1, \cdots, x^n = q^n) = P(x^{n+1} = q^{n+1} | x^n = q^n) \tag{2.19}$$

i.e. the system follows a "memoryless" evolution (Markov property), the probability to go from a state n to a state $n+1$ does not depend on the history of all the previously visited ones, but is instead only determined by the current state n.

Let us consider the above mentioned states i and j, and define π_{ij} as the conditional probability

$$\pi_{ij} = P(x^{n+1} = q^j | x^n = q^i)$$

then the Markov Chain imposes the following property of micro-reversibility, also called *detailed balance*:

$$\nu_i \pi_{ij} = \nu_j \pi_{ji} \tag{2.20}$$

The essential condition of detailed balance will be of crucial importance in the application of the Metropolis-Hastings algorithm detailed below.

Metropolis-Hastings algorithm This algorithm was proposed in 1953 by Nicholas Metropolis, Arianna and Marshall Rosenbluth, Augusta and Edward Teller[72] for studying a two dimensional rigid spheres problem, in a form limited to the study of symmetrical probability distributions; W. Keith Hastings extended it to any p.d.f in 1970 [73].

The algorithm uses a Markov Chain as defined by Equation 2.19 for generating a chain of states $(q^1, \cdots, q^n)$ following the canonical distribution 1.15. Let us reconsider the terms q^i, q^j, ν_i, ν_j and π_{ij} introduced in the previous paragraph. Let us define the additional P_{ij} as the probability of observing the $i \to j$ transition, and α_{ij} as the probability to perform a random move leading from i to j, thus one can write:

$$\pi_{ij} = \alpha_{ij} P_{ij}$$

The transition matrix π of the conditional probabilities can only be approximated by counting transitions between all the possible states the system can exhibit, for a simulation time $t \to +\infty$; it is thus convenient to rewrite Equation 2.20 as:

$$\nu_i \alpha_{ij} P_{ij} = \nu_j \alpha_{ji} P_{ji}$$

which rewrites as:

$$\frac{P_{ij}}{P_{ji}} = \frac{\alpha_{ji}}{\alpha_{ij}} \frac{\nu_j}{\nu_i} = \frac{\alpha_{ji}}{\alpha_{ij}} \frac{Z_{NVT}^{-1} e^{-\beta V(q^j)}}{Z_{NVT}^{-1} e^{-\beta V(q^i)}} \tag{2.21}$$

The left hand ratio is referred to as an acceptance distribution: $A(i \to j) = \dfrac{P_{ij}}{P_{ji}}$: if $\Delta V = V(q^j) - V(q^i)$, the difference of potential energy between configurations i and j is defined, Equation 2.21 rewrites to:

$$A(i \to j) \propto e^{-\beta \Delta V} \tag{2.22}$$

From Equation 2.22 one can see that with the Metropolis-Hastings algorithm, states generated by application of the acceptance distribution will follow automatically the NVT p.d.f. without necessitating any explicit estimation of the partition function: this makes the algorithm extremely useful in combination with ensemble averaging methods. The ratio $\frac{\alpha_{ji}}{\alpha_{ij}}$ is usually chosen to be one (classical Metropolis algorithm), the contribution of Hastings was to allow $\alpha_{ji} \neq \alpha_{ij}$ which may enhance the sampling for some cases but will introduce a bias to be corrected.[73]

The distribution $A(i \to j)$ is practically populated following the Metropolis rules:

1. Evaluate $V(q^i)$, the energy of the current state i, before any modification
2. Generate a proposal move $\Theta(i \to j)$, i.e. a stochastic modification of the coordinates representing a transition in the configurational space $i \to j$
3. Evaluate the new energy $V(q^j)$
4. Estimate $\Delta V = V(q^j) - V(q^i)$:
   - If $\Delta V \leq 0$ the proposal $\Theta(i \to j)$ is accepted
   - If $\Delta V > 0$ the proposal $\Theta(i \to j)$ is accepted if:

$$\xi < e^{-\beta * (\Delta V)} \tag{2.23}$$

   where ξ is random number uniformly distributed in $]0; 1[$.
5. Iterate to 1.

The Figure 2.2 is a workflow representation of this algorithm.

In Figure 2.3 the $e^{-\beta \Delta V}$ are plotted for values of $\beta = 1$ or $\beta = 2$ (arbitrary units). The area under the curves (AUC) (solid grey and dashed white lines respectively) is always 1, respecting Equation 1.11. The area $\Delta V \leq 0$ is not shown on the plot as moves respecting this condition are always accepted. For cases where $\Delta V > 0$ the rule $\xi < e^{-\beta * (\Delta V)}$ with $\xi \in]0, 1[$ is motivated by the fact that one wants to sample the AUC, which can be intepreted as the integral of Equation 1.11.

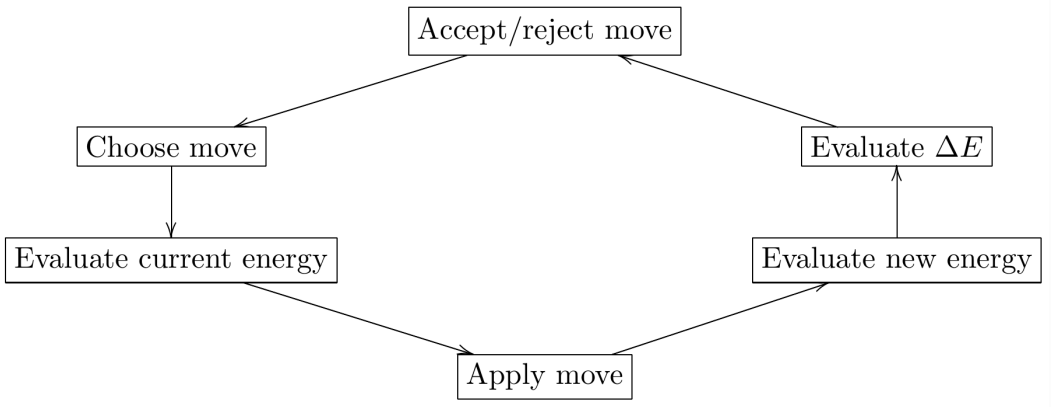


FIGURE 2.2: Diagram detailing the Metropolis-Hastings algorithm.

The critical thing when applying the Metropolis Hastings method to a chemical or biological system is to generate stochastic modifications $\Theta(i \to j)$ of the system significant enough for exploring the configurational space, but that will not make the value $\Delta V >> 0$, otherwise the probability to accept such move $Pr(\Theta(i \to j))$ is close to 0. A common rule is to tune dynamically, during the simulation, the maximum random coordinates modification d_{max} applied to the system, so that on long term the acceptance rate (i.e. the ratio accepted/rejected moves) follows $0.3 \leq Pr(\Theta(i \to j)) \leq 0.7$, with $Pr(\Theta(i \to j)) \approx 0.5$ being usually a good choice.

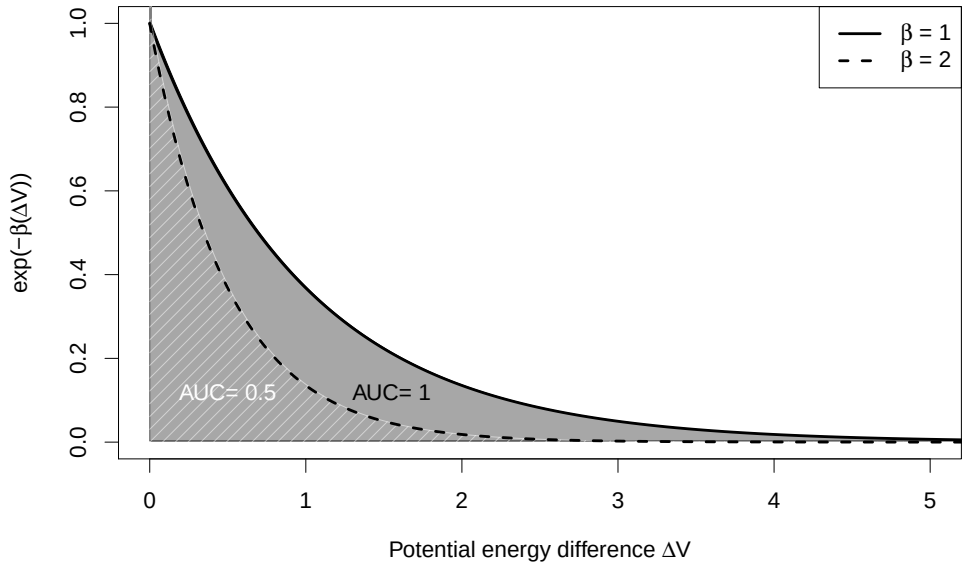


FIGURE 2.3: Illustration of the choice of $\xi < e^{-\beta*(\Delta V)}$ where $\xi \in [0,1]$ for the Metropolis-Hastings algorithm: the idea is to sample the area under the curves (AUC) using the Monte Carlo method.

While for simple applications such as the double-well potential (see next paragraph and Chapter 3) or Lennard-Jones particles clusters (Chapter 3) the random move proposal $\Theta(i \to j)$ simply consists in translating one or several coordinates $q_i \in q$ using random numbers, for the case of real molecules this is usually not sufficient.

Indeed one should remember the expression of the potential energy function introduced by Equation 1.6 and the bond, angle and dihedral terms: it is necessary to introduce Monte Carlo moves that will vary the length of some of the bonds, and the value of some angles and dihedral angles, in order to explore properly the p.d.f. designed from Equation 1.6.

The CHARMM MC module [74] introduces a class of different possible Θ modifications, including: rigid rotations, rigid translations, and torsional modifications for exploring the configurational space induced by the CHARMM Force Field.[75] The possibility to perform *concerted dihedral rotations*, i.e. multiple torsion moves on successive dihedral angles at once, was shown to be a key feature when studying polypeptides and proteins,[76, 77] and is also available in CHARMM.

Application Here a simple application of MC sampling to the double-well potential introduced at the beginning of this Chapter is presented. For more applications please refer to Chapter 3 where MC and SA-MC sampling are discussed.

As already mentioned, reduced units are used: distance, potential and temperature are unit-less, and $k_{\mathrm{B}} = 1$. 10^8 Monte Carlo steps are performed, at three temperatures $T = \{0.05, 0.2, 0.4\}$ (as for Figure 2.1). The initial coordinates were $x_0 = (0.0, 0.0)$ i.e. at the top of the barrier.

The proposal transformation $\Theta(i \to j)$ is defined, for the coordinates variable x as following:

$$\Theta(i \to j) : x_j = x_i + d_{max} * (\xi - 0.5)$$

where $\xi \in]0.0; 1.0[$ is a uniformly distributed random number, and d_{max} the maximal amplitude for a random move: detailed balance is ensured by the fact that the random move is uniformly distributed

in $]-\frac{d_{max}}{2};\frac{d_{max}}{2}[$. The value of d_{max} was adjusted every 100 steps for reaching a total acceptance of $Pr\big(\Theta(i \to j)\big) \approx 50\%$.

Results are shown in Figure 2.4: the sampling of $\rho(x)$ for $T = \{0.2, 0.4\}$ is exactly what was predicted on Figure 2.1, confirming the ability of the MC technique to generate states following perfectly the canonical p.d.f. However for the case $T = 0.05$ only the right well is sampled, which means that the barrier is never crossed at this temperature, even when 10^8 steps were performed. This illustrates one limitation of the MC method: accurate sampling of $\rho(x)$ is feasible, but this may take an almost infinite time.

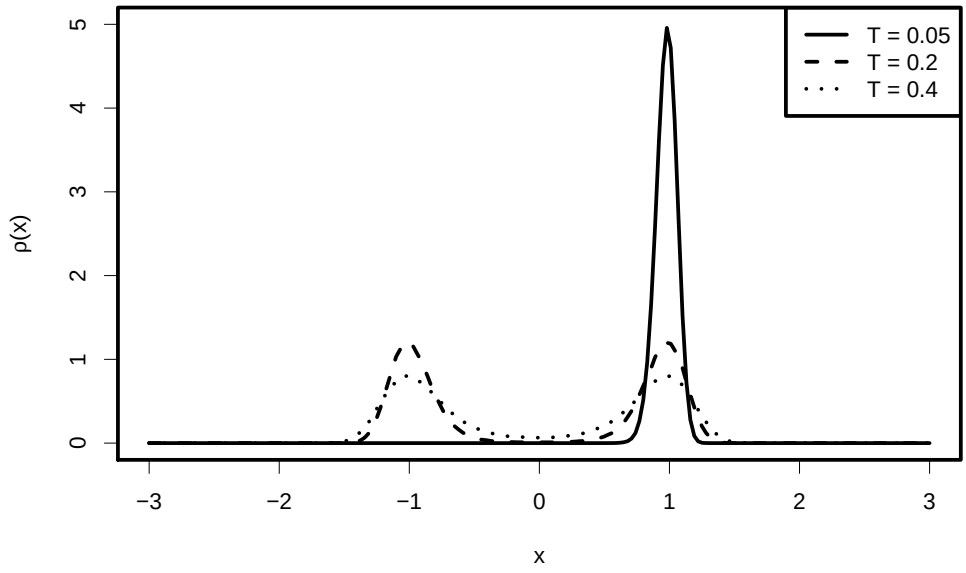


FIGURE 2.4: MC sampling of $V(x) = (x^2 - 1)^2$ at $T = \{0.05, 0.2, 0.4\}$, 10^8 steps, starting point was the top of the barrier at $(0, 0)$.

Limitations For high energy barriers ($\Delta V >> 0$) , the term $e^{-\beta*(\Delta V)}$ is close to zero (see large values of ΔV on Figure 2.3), and the probability of accepting such a move is extremely low: some possible simple solutions for sampling those rare events are: (i) increase the temperature of the simulation, as it will reduce the impact of the high ΔV (see Figure 2.4): but it alters the underlying thermodynamic properties, and furthermore it may impact the stability of the chemical or biological system studied. (ii) Increase the number of steps (random move trials) : statistically it allows the crossing of the barrier by a successive "chain" of states where the ΔV between points of this chain are lower; but the probability of observing this crossing event is also exponentially decaying with ΔV. See Figure 2.4 where even 10^8 steps were not enough for observing a single crossing at low temperature.

For large chemical and biological systems, one may need to use a dedicated rare events sampling techniques. A brief review of the current methods of interest is given in Section 2.3.

2.3 Review of rare event sampling methods

Several rare event sampling methods have been developed in the past. They include parallel tempering (PT),[38, 39, 41] umbrella sampling (US),[78] metadynamics (MTD),[79] or replica exchange (RE),[40] and some optimised MC schemes.[80, 81] The available sampling techniques can usually be classified in one of the three following categories:

(i) MC move optimisations, such as the displacement vector MC technique[80], or more recent studies specifically aiming at simulations of proteins.[81] But this kind of parameters tuning requires some a priori knowledge about the topology of the to be explored potential energy surface.

(ii) Parallel tempering[82], Replica Exchange[38], based on repeated information exchange between replicas of a simulation system, running at different values of an external control parameter (such as temperature). Lower-temperatures replicas are enriched with knowledge coming from higher-temperature ones where high-energy barriers are more easily crossed. Several strategies for defining the tempering ensemble have been discussed[83–86].

(iii) Through the addition of an external bias, such as a supplementary potential for filling basins of energy surface (metadynamics,[46–49] flooding[87, 88]), auxiliary probability density (Tsallis weight sampling),[89–91] energy smoothing methods,[92–95] or constrained geometry (Umbrella Sampling).[44] One should also mention Hamiltonian Exchange (HEX), Hamiltonian Replica Exchange (HREX) or Generalised Ensemble techniques, [96–102] variants of RE simulations where each replica can run with a modified Hamiltonian in order to enhance sampling.

The PT/RE method will be briefly detailed in Section 2.5 as the INS/PINS approaches rely on it. For details concerning other methods, one can read the corresponding references, or some reviews.[103–105]

The following section will introduce the SA-MC, a robust and versatile rare events sampling method.

2.4 The SA-MC method

Spatial Averaging Monte Carlo (SA-MC) approach was introduced by N. Plattner, J. D. Doll, M. Meuwly[53] and other collaborators. The key feature is the construction of a modified p.d.f. with the following properties: (i) The integral of the modified p.d.f. over the whole working space is (ideally) identical to the original one: this is needed if one wants accurate thermodynamic properties through application of Equation 1.8. (ii) The modified p.d.f. is easier to sample than the original one, which leads to a faster sampling of the configurational space in return.

Let us consider an uni-dimensional potential energy curve $V(x)$ on which evolves a coordinates vector $x \in D, D \subseteq \mathbb{R}^N$: its potential energy dependent p.d.f is, according to Equation 1.15:

$$\rho(x) = \frac{1}{Z} \exp(-\beta V(x)) \tag{2.24}$$

In the following the normalisation by the partition function Z is omitted for simplifying the equations, and by an abuse of notation it is assumed that $\rho(x) = \exp(-\beta V(x))$.

The new set of modified p.d.f. was defined[53] as following:

$$\rho(x, \varepsilon) = \int P_\varepsilon(y) \exp\big((-\beta V(x+y)\big)\ dy \tag{2.25}$$

Where $P_\varepsilon(y)$ is a normalised probability distribution of length scale ε. It is possible to rewrite Equation 2.25 using the flow and composition mathematical notation introduced earlier:

$$\rho(x, \varepsilon) = \int_{\forall x_\varepsilon \in D} \rho \circ P_\varepsilon\ dx_\varepsilon \tag{2.26}$$

where $x_\varepsilon = P_\varepsilon(x)$ is a set of coordinates altered by the distribution P_ε and centred around the initial x. Then the flow $\rho \circ P_\varepsilon = (\rho \circ P_\varepsilon)(x) = \rho\big(P_\varepsilon(x)\big) = \rho(x_\varepsilon)$ corresponds to the application of Equation 2.24 to an altered dataset x_ε. Finally $\rho(x, \varepsilon)$ is the resulting averaged p.d.f. built from all the distinct sets x_ε generated through use of P_ε.

P_ε is usually chosen to be a Gaussian distribution of standard deviation ε, but the authors mentioned in Ref. [53] that the method should be robust enough for allowing the use of other distributions, as long as

$$\lim_{\varepsilon \to 0} (\rho \circ P_\varepsilon)(x) = \rho(x)$$

By adjusting the parameter ε one can adapt the biasing distribution to the particular problem of interest, and sample more easily states far from the centre $\rho(x)$.

One should also note that $\rho(x, \varepsilon)$ is centred around $\rho(x)$ so its integral over the whole configuration space is equal to the integral of the unmodified p.d.f.:

$$\int_D \rho(x) = \int_D \rho(x, \varepsilon) \tag{2.27}$$

Equations 2.26 and 2.27 are key to SA-MC, as they imply that thermodynamic properties derived from $\rho(x, \varepsilon)$ should be identical to those estimated from $\rho(x)$ for a given temperature.

Application to higher dimensional systems Now that the mathematical foundations have been introduced (Equations 2.25 and 2.27), one has to adapt the algorithm for a general $\mathbb{R}^{3N}$ dimensional space. In a second publication [54], N. Plattner, J. D. Doll and M. Meuwly proposed the following procedure:

1. Consider an initial configuration q^i (coordinates vector at simulation step i).
2. Around this q^i, generate M_ε sets of N_ε configurations, following a normal law of standard deviation W_ε and centred on q^i: one obtains a set of $M_\varepsilon * N_\varepsilon$ vectors $q^{i,\varepsilon}$.
3. Apply the chosen MC move $\Theta_{i \to j}$ to all of the $M_\varepsilon * N_\varepsilon$ configurations: a second set of vectors $q^{j,\varepsilon}$ is obtained.
4. Compute the $M_\varepsilon * N_\varepsilon$ corresponding potential energies for $q^{i,\varepsilon}$ and $q^{j,\varepsilon}$. Let us denote them simply as $V(q^{i,\varepsilon})$ and $V(q^{j,\varepsilon})$ (it is possible to imagine $q^{i,\varepsilon}$ as a matrix of $3N$ rows by $M_\varepsilon * N_\varepsilon$ columns). Then one defines the pseudo p.d.f.s (once again the Z normalisation is ignored) as:

   $$\rho(q^{i,\varepsilon}) = \exp\big(-\beta V(q^{i,\varepsilon})\big) \quad \text{and} \quad \rho(q^{j,\varepsilon}) = \exp\big(-\beta V(q^{j,\varepsilon})\big)$$

   The $\rho(q^{i,\varepsilon})$ and $\rho(q^{j,\varepsilon})$ are then vectors of size $M_\varepsilon * N_\varepsilon$
5. For each M_ε set, evaluate the in-set sum of the canonical measures:

   $$S_m^i = \sum^{N_\varepsilon} \rho_m(q^{i,\varepsilon}) \quad \text{and} \quad S_m^j = \sum^{N_\varepsilon} \rho_m(q^{j,\varepsilon})$$

   And then take the ln of the per-set ratio between S_m^j and S_m^j:

   $$\delta_m = -\ln\left(\frac{S_m^j}{S_m^i}\right)$$

   δ_m can be seen as a measure of the sampling gain that the SA-MC provides, for each set M_ε
6. Then the block averaged gain δ is introduced:

   $$\delta = \frac{1}{M_\varepsilon} \sum^{M_\varepsilon} \delta_m$$

   together with the variance σ^2 over the M_ε blocks:

$$\sigma^2 = \frac{1}{M_\varepsilon * (M_\varepsilon - 1)} \sum^{M_\varepsilon} (\delta_m - \delta)^2$$

7. Then the SA-MC criterion $\delta + \frac{\sigma^2}{2}$ will replace the ΔV of the Metropolis-Hastings acceptance criterion, and the Equation 2.23 became :

$$\xi < \exp\left(-\beta * \left(\delta + \frac{\sigma^2}{2}\right)\right) \tag{2.28}$$

With this approach the criterion represents a potential energy difference averaged over all the sets. The motivation for generating M_ε blocks of N_ε is that one can correct the δ term with the value of σ^2, a procedure defined as *variance reduction*[106], a method known for improving accuracy of random estimates.

It is thus possible to accept a state with a ΔV significantly higher than with a classical Metropolis because its energy is averaged with the one of the other $M_\varepsilon * N_\varepsilon$ configurations, and so the probability of crossing high energy barriers is increased.

The Figure 2.5 is a workflow representation of this algorithm. In the following a SA-MC simulation will be characterised by a triplet $[W_\varepsilon, M_\varepsilon, N_\varepsilon]$ corresponding respectively to the width of the normal law, to the number of sets, and to the number of points per set.

The slowest part of the algorithm consists in the $M\varepsilon * N\varepsilon$ loop one can see at the bottom of Figure 2.5: each loop iteration requires two energy evaluations, as a classical MC algorithm. Therefore the algorithm is natively $M\varepsilon * N\varepsilon$ times slower than a classical Metropolis-Hastings simulation. Fortunately, as it will be illustrated in Chapter 3 the number of steps N_{SA-MC} required for obtaining converged results with SA-MC is much lower than for MC, i.e. $N_{SA-MC} << N_{MC}$, counterbalancing the time spent in the $M\varepsilon * N\varepsilon$ energy evaluations.

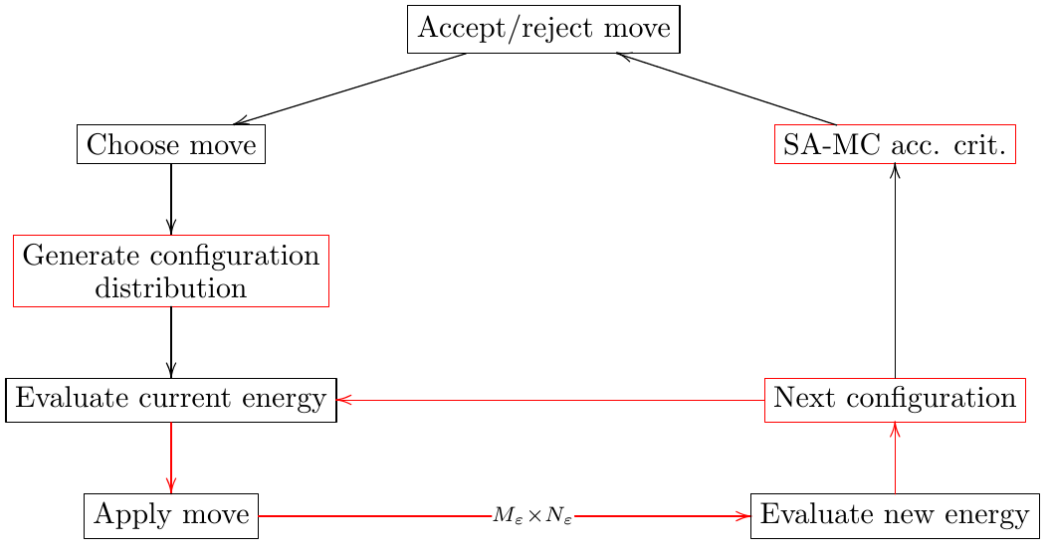


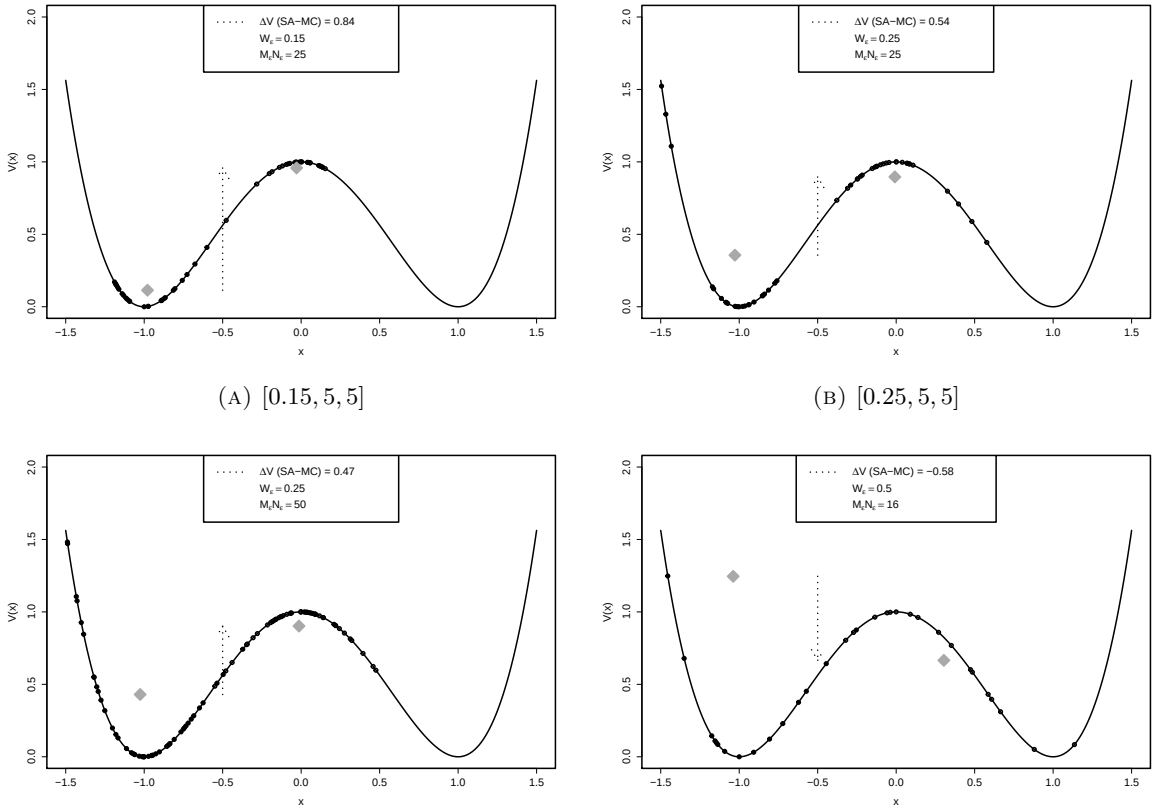
FIGURE 2.5: Diagram detailing the SA-MC algorithm: dashed parts are specific to spatial averaging; dashed arrows represents a loop over $M\varepsilon * N\varepsilon$.

In Figure 2.6 the SA-MC sampling is illustrated for the double-well potential. Two i–j configurations respectively in the left minimum (coordinates $[-1, 0]$) and at the top of the barrier coordinates $[-0, 1]$ are considered. Around each i–j, $M_\varepsilon * N_\varepsilon$ points are distributed following a normal law of width W_ε. Two grey diamonds represent the SA-MC "averaged" potential energies, and the dashed arrow represents the energy difference between them.

The benefits of SA-MC are observed on Subfigures 2.6a – 2.6b – 2.6c, for different combinations of $[W_\varepsilon, M_\varepsilon, N_\varepsilon]$: one can see that increasing values of W_ε logically allows sampling of points farther

from the initial starting configuration, and possibly characterised by a higher energy. Increasing the product $M\varepsilon * N\varepsilon$ also reduces the size of the "virtual" potential energy barrier (values in the legend of each plot). However one should remember the $M\varepsilon * N\varepsilon$ above mentioned time dependence of the algorithm: hence when comparing Subfigures 2.6b – 2.6b one should note that the computational time required doubled for simply lowering the energy by 0.07 arbitrary units, thus making the choice of parameters $[0.25, 5, 10]$ over $[0.25, 5, 5]$ not so clever.

The Subfigure 2.6d illustrates an application of SA-MC where parameters $[0.5, 4, 4]$ were chosen, i.e. a large distribution width combined with a few points. An unexpected behaviour is observed were the relative energy of the minimum and the barrier are "reversed" and where a negative potential energy difference is observed. Although one may consider this effect to be useful for sampling, it completely modifies the underlying thermodynamic properties that one may estimate from the results in a post-processing phase: the simulation is completely *biased*. This is because the equality 2.26 is only mathematically valid for either $W_\varepsilon \to 0$ or $M\varepsilon * N\varepsilon \to +\infty$. Therefore one should find a balance between improved sampling and respect of Equations 2.26 and 1.15.



(A) $[0.15, 5, 5]$ (B) $[0.25, 5, 5]$

(C) $[0.25, 5, 10]$ (D) $[0.5, 4, 4]$

FIGURE 2.6: Comparison of SA-MC sampling with different parameters on the double well potential. Black dots correspond to the $M_\varepsilon * N_\varepsilon$ normal distributed points around each starting configuration. Gray diamonds illustrate the "averaged" SA-MC potential energies and the dashed arrow illustrates the energy difference.

The bias introduced for the case of Subfigure 2.6d is in fact inherent to all the SA-MC simulations, because of the relatively low number of points $M\varepsilon * N\varepsilon$ chosen when discretising Equation 2.26: however the next paragraph introduces a useful method for unbiasing results of a SA-MC simulation.

Data unbiasing for thermodynamic applications Let $\bar{f}_0(x)$ be an unbiased estimate of a thermodynamic property defined by using the densities of states $\rho(x)$: it is expressed as following (using Equation 1.7):

$$\bar{f}_0(x) = \int_{\forall x \in D} f(x)\rho(x)dx \tag{2.29}$$

Introducing $f_\varepsilon(x)$ as a biased measure of $f(x)$, combining Equations 2.29 and 2.26 one gets:

$$\bar{f}_0(x) = \int_{\forall x \in D} \rho(x,\varepsilon) f_\varepsilon(x) \frac{\rho(x)}{\rho(x,\varepsilon)} dx$$

Let $\bar{f}_\varepsilon(x) = \int_{\forall x \in D} f_\varepsilon(x)\rho(x,\varepsilon)dx$ be a biased estimate of the thermodynamic property f: for a simulation with enough sampling (N steps, $N \to \infty$), this is rewritten, using the notion of ensemble average from Equation 1.8:

$$\bar{f}_0(x) = \bar{f}_\varepsilon(x) * \sum_{n=1}^{N} \frac{\rho(x^n)}{\rho(x^n,\varepsilon)} \tag{2.30}$$

Equation 2.30 is extremely important: for any thermodynamic property (for example free energy, as estimated using the Histogram method from Section 1.3.3), one can get an unbiased SA-MC estimate by saving during simulation the ratio $\frac{\rho(x^n)}{\rho(x^n,\varepsilon)}$, and using it appropriately during the post-processing phase when the free energy is estimated. From the simulation this ratio is simply obtained by dividing the Metropolis energy difference $e^{-beta\Delta V}$ by the above introduced SA-MC equivalent $\exp\left(-\beta * \left(\delta + \frac{\sigma^2}{2}\right)\right)$. But once again it should be emphasised that those results assume Equality 2.26 to be valid, and as it will be illustrated in Chapter 3 when the biasing is too strong the unbiasing procedure cannot recover a proper estimate $\bar{f}_0(x)$.

2.5 The INS and PINS methods

This Section introduces the infinite swapping (INS) and partial infinite swapping (PINS) methods.[107–110] As parallel tempering (PT) methods, they use an expanded ensemble built from a number of replicas running at different temperatures. But contrary to PT, INS uses the fully symmetrized distribution of configurations in temperature space, whereas PT just occasionally enriches the local replica with configurational information coming from simulations at a higher temperature.

The first paragraph briefly introduces the PT method, then the INS and PINS algorithms are introduced.

Parallel tempering Parallel Tempering (PT) (also known as Replica Exchange (RE)) methods[38, 82, 111] were introduced in 1986 by Swendsen and Wang.[38] They were shown to be useful for many studies fo chemical and biological systems. In a PT simulation with K replicas, each being an NVT ensemble but with a different temperature, the partition function Z of the combined ensemble built from the K replicas is :

$$Z = \prod_{i=1}^{K} \frac{p^i}{N!} \int_D dq^i \; e^{-\beta_i V(q^i)} \tag{2.31}$$

Where $p^i = \prod_{k=1}^{N} (2\pi M_k \beta_i^{-1})^{3/2}$ is obtained by integrating momenta of the N particles of mass M_k, where $V(q^i)$ is the potential energy of the coordinates set q for replica i, and $\beta_i = \frac{1}{k_B T_i}$ is the reduced temperature for replica i.

Replicas are exchanged between two adjacent temperatures $i - j$ with probability :

$$P_{acc}(i \leftrightarrow j) = \min\left\{1, e^{(\beta_i - \beta_j)\Delta V}\right\} \tag{2.32}$$

The temperatures are usually distributed non-linearly between T_1, which is the desired simulation temperature, and T_K, in order to have a constant value of $P_{acc}(i \leftrightarrow j)$, typically around 20 to 25% (see [82] for a discussion on the choice of temperatures and the impact on P_{acc}).

Infinite Swapping limit for PT simulations INS is based on a mathematical analysis of the convergence rate of PT simulations as a function of the temperature swap trial.[107, 109] It was demonstrated [109] that this convergence rate is a monotonically increasing function of the swap rate, and thus that an optimal sampling is possible in the *infinite swapping* limit.

Therefore, INS provides optimal sampling for a given replica by using information from all other temperatures used in the simulation. This could be achieved using PT allowing exchanges between all replicas at each time step. However, as there are $K!$ exchanges for K replicas this would become an unmanageable number of exchanges for realistic choices, e.g. $K = 20$. Furthermore, many of the exchanges would not be accepted which would further compromise the efficiency of the method. Instead of attempting all $K!$ exchanges and estimating their acceptance P_{acc} following Eq. 2.32 for each permutation, with INS the probability of such a general exchange is estimated according to

$$\rho_k(q^i) = \frac{\pi_k(q^i)}{\sum_{k=1}^{K!} \pi_k(q^i)}. \tag{2.33}$$

Here $\pi_k(q^i)$ is given by

$$\pi_k(q^i) = \prod_{i=1}^{K} e^{-\beta_i V(q^{k,i})} \tag{2.34}$$

and $q^{k,i}$ is the configuration of replica i corresponding to the assignment of configurations to temperatures in permutation k. Therefore, with INS the optimal permutation is found by comparing all the possible $\rho_k(q^i)$ permutation probabilities, and by performing the global swapping corresponding to the ρ_k that maximises the convergence of the simulation.

Let us illustrate once again the algorithm with a simple double-well potential $V(x) = (x^2 - 1)^2$ (see Figure 2.1).

Let us define a multi-variable distribution μ that combines the 3 temperatures $T_x, T_y, T_z = 0.05, 0.20, 0.40$ as :

$$\mu(x, y, z, T_x, T_y, T_z) = \rho(x, T_x)\rho(y, T_y)\rho(z, T_z) \tag{2.35}$$

The resulting plot can be seen on Figure 2.7, left part, where isosurface is plotted for $\mu = 0.05$. The fact that the isosurface is elongated along the z-axis reveals that, logically, the temperature T_z is the one where the densities are the most connected.

The symmetrized approach of INS can be written as $\mu_{x,y,z}$ where the $3! = 6$ permutations are included, where a bold font emphasise a swapped temperature:

$$\begin{aligned}\mu_{x,y,z}(x, y, z, T_x, T_y, T_z) = \mu(x, y, z, T_x, T_y, T_z) + \mu(x, y, z, T_x, \mathbf{T_z}, \mathbf{T_y}) + \\ \mu(x, y, z, \mathbf{T_y}, \mathbf{T_x}, T_z) + \mu(x, y, z, \mathbf{T_y}, \mathbf{T_z}, \mathbf{T_x}) + \\ \mu(x, y, z, \mathbf{T_z}, \mathbf{T_x}, \mathbf{T_y}) + \mu(x, y, z, \mathbf{T_z}, T_y, \mathbf{T_x})\end{aligned} \tag{2.36}$$

Thus one applies Equation 2.35 to the 6 temperatures permutations. The resulting $\mu_{x,y,z}$ is plotted on Figure 2.7, right part. One can see that the resulting isosurface is isotropically distributed in the

temperature dependent configurational space x, y, z: there is no more a correlation between the p.d.f. of a given replica and the temperature at which it was assigned at the beginning.

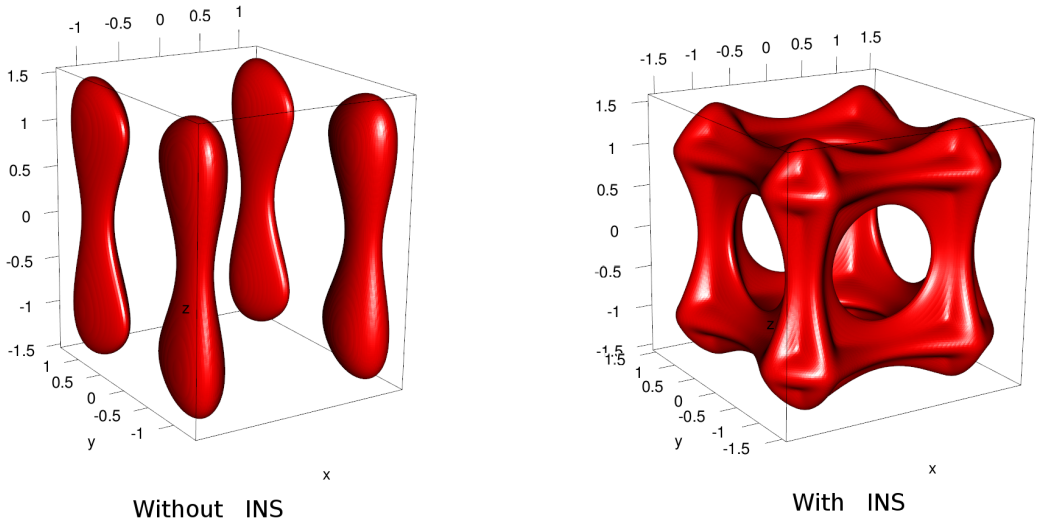


FIGURE 2.7: Plot of the the isosurface defined by the multi-variable density $V(x) = (x^2 - 1)^2$ for 3 temperatures $T_x, T_y, T_z = 0.05, 0.20, 0.40$. Left : without symmetrisation. Right : with full symmetrisation, i.e. INS.

Concretely INS allows an optimal sampling of the configurational space D for a given number of replica K, by using information coming from all the other temperatures used during the simulation. Nevertheless, for large systems that implies a consequent number of temperatures, it is computationally too expensive to calculate the $N!$ probabilities, and to obtain a proper evaluation of Equations 2.33 and 2.34. Therefore a *partial infinite swapping* (PINS) algorithm was introduced.[107–110]

Partial Infinite Swapping With the PINS approach, a partitioning strategy is used: temperature space is divided into blocks, and local (but full) symmetrisation is used within each block. More precisely, the current implementation uses the "dual-chain" approach[108], where the $K-$temperature set is partitioned into blocks in two different ways, one for each chain. The two blocks must have a complementary structure without a boundary between the blocks defined for the two chains. This is required in order to achieve sampling of the overall temperature space for all the replicas.

Let us consider for instance a set of 12 temperatures: a possible partitioning for the two chains $(a|b)$ is $(3, 6, 3|4, 4, 4)$, where the a boundaries are $T_3 - T_4$ and $T_9 - T_{10}$, and for b they are $T_4 - T_5$ and $T_8 - T_9$. On the other hand, the partitioning $(3, 3, 6|6, 3, 3)$ is not valid, as chain a boundaries' are $T_3 - T_4$ and $T_6 - T_7$, and for chain b they are $T_6 - T_7$ and $T_9 - T_{10}$, thus sharing the common boundary $T_6 - T_7$.

The PINS approach can again be illustrated straightforwardly with help of the double well potential: using Equation 2.35, one can define the multi-temperature $\mu_{x,y}$ and $\mu_{y,z}$ combinations of isosurfaces:

$$\mu_{x,y}(x, y, z, T_x, T_y, T_z) = \mu(x, y, z, T_x, T_y, T_z) + \mu(x, y, z, \mathbf{T_y}, \mathbf{T_x}, T_z) \tag{2.37}$$

and

$$\mu_{y,z}(x, y, z, T_x, T_y, T_z) = \mu(x, y, z, T_x, T_y, T_z) + \mu(x, y, z, T_x, \mathbf{T_z}, \mathbf{T_y}) \tag{2.38}$$

Figure 2.8 illustrates Equations 2.37 and 2.38: with the simple partial swapping of only two temperatures, one gets already improved connections (i.e. bridging between well sampled areas), but for a computational cost which is only a fraction of the full Equation 2.36.

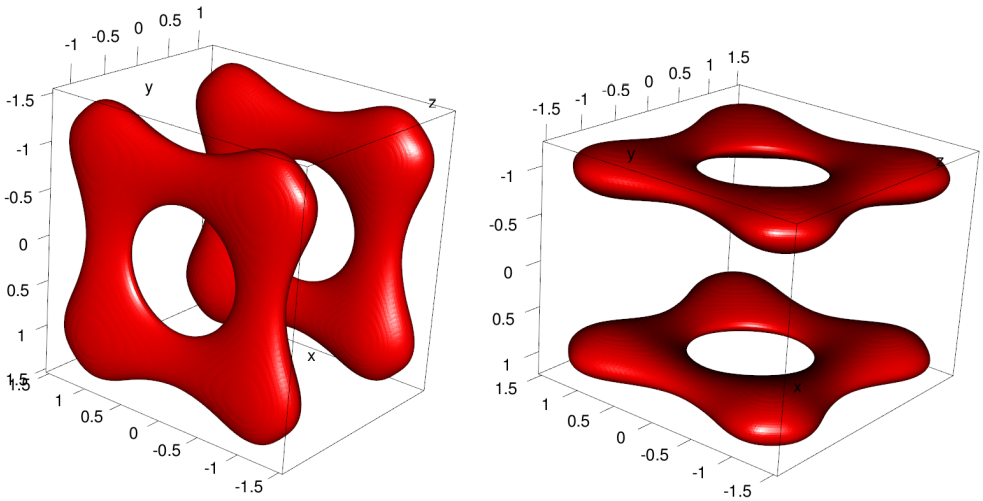


FIGURE 2.8: Plot of the isosurface defined by the multi-variable density $V(x) = (x^2 - 1)^2$ for 3 temperatures $T_x, T_y, T_z = 0.05, 0.20, 0.40$. With PINS partial symmetrisation of $\mu_{x,y}$ (left) or $\mu_{y,z}$ (right), respectively corresponding to Equations 2.37 and 2.38.

Now let us consider that we use the dual chain approach previously mentioned: the $\mu_{x,y}$ swapping is first performed (first chain of swapping), then just after the second swapping $\mu_{x,y}$ is performed: the resulting "combined" surface will connect the whole temperature space, as represented in Figure 2.9, and can be seen as a kind of approximation of the surface defined by the whole permutations (Figure 2.7 (Right)).

Of course here the usefulness of PINS is limited by the small number of 3 temperatures, and combining $\mu_{x,y} + \mu_{y,z}$ should not be done because they have a common boundary y (c.f. previous discussion on the dual-chain method). However, for a higher number of replicas it will be shown in Chapter 4 that PINS can improve considerably the sampling when compared to a PT simulation running with the same K temperatures.

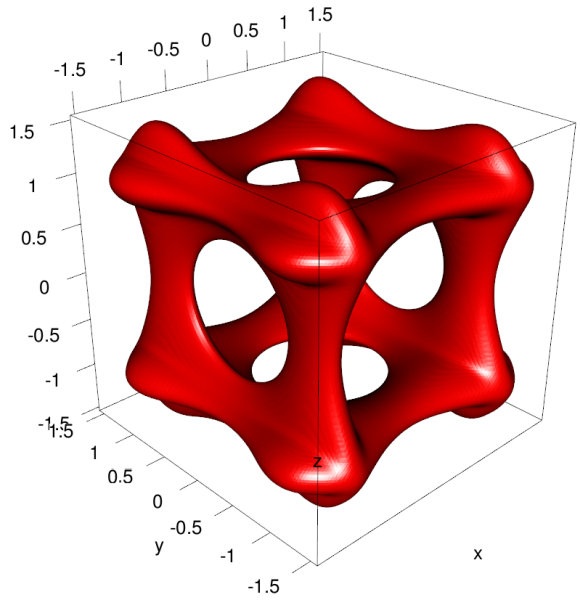


FIGURE 2.9: Plot of the combined isosurface defined by $\mu_(x, y) + \mu_(y, z)$. (right).

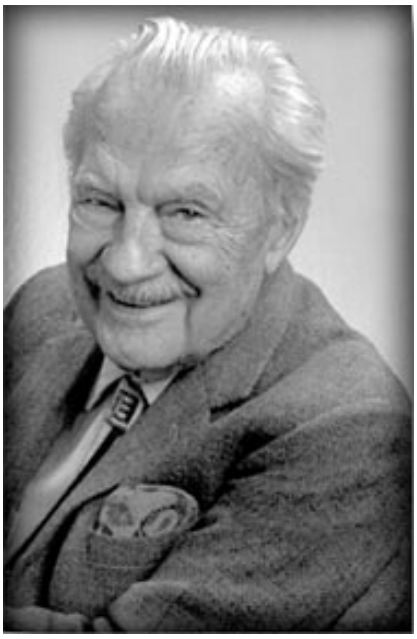
Part II

INVESTIGATIONS

Chapter 3

Validations, applications and results for SA-MC

"*Most of us have grown so blase about computer developments and capabilities — even some that are spectacular — that it is difficult to believe or imagine there was a time when we suffered the noisy, painstakingly slow, electromechanical devices that chomped away on punched cards.* "



Nicholas Constantine Metropolis, The beginning of the Monte Carlo method (1989).

In this Chapter the results obtained during the PhD with SA-MC are presented. For a mathematical and algorithmic background, one can refer to Section 2.4 where the methodology was properly introduced.

As previously mentioned in Chapter 2 SA-MC has been applied to model systems in Ref.[53], and in Ref. [54] a first application to chemical systems was presented, for studying diffusion of small molecules in condensed phase environments. But no public implementation of SA-MC was available after publication of those articles. Therefore the first part of this thesis work consisted in implementing SA-MC in CHARMM, so that it could be used by the CHARMM community, for application to a variety of systems.

An article was written for presenting the new implementation, and validating it. It was published in August 2014, in the *Journal of Chemical Theory and Computation* (JCTC), Vol. 10, Pages 4284–4296, [112] co-written with Nuria Plattner, Jimmie. D. Doll and Markus Meuwly.

This article is appended to the current Chapter and can be found in Section 3.4.

The SA-MC code is written as a Fortran sub-module of CHARMM's MC module. Compiling and use information is available from the official CHARMM documentation, within the source archive, or online.[1]

The stand-alone C code for performing MC and SA-MC simulations on Lennard-Jones cluster is available on GitHub and can be freely accessed (3-clause BSD license).[2]

Some unpublished content and results are also briefly introduced in the following Sections: it mainly consists in Figures and Plots that were not included in the final article but may still be of interest because they are representative of implementation's evolution.

3.1 Double-Well potential

In the article the first validation is done on the double well potential (Article → 3.Applications → 3.1. Double Well Potential). Results are given for a barrier of 2 or 5 units of k_BT. A reconstructed 1-dim free energy profile is given for the barrier of 2 k_BT. Please refer to the corresponding page for more details

3.1.1 Supplementary unpublished content

Asymmetric potential Although not mentioned in the article an asymmetric double well potential was also studied:

$$f : x \to (x^2 - 1)(x - 3)^2 \tag{3.1}$$

Its derivative is:

$$f' : x \to 2(2x^3 - 9x^2 + 8x + 3) \tag{3.2}$$

Note that $f' = 0$ if $x_1 = \frac{1}{4}(3 - \sqrt{17})$ or $x_2 = \frac{1}{4}(3 + \sqrt{17})$ or $x_3 = 3$, and that $f(x_1) \approx -9.9149$; $f(x_2) \approx 3.2274$; $f(x_3) = 0$.

Table 3.A represents the variation table of the function f.

Figure 3.1 presents results for several couples of $(\varepsilon, N_\varepsilon)$ values for this asymmetric potential: same conclusions as for the symmetric potential apply, i.e. increasing those values leads to a better sampling of high energy regions, regions that the MC Metropolis algorithm cannot sample.

[1] https://www.charmm.org/charmm/documentation/by-version/c40b1/params/doc/mc/#SA-MCsimulations
[2] https://github.com/FHedin/mc_LJ

x	$-\infty$	$x_1 = \frac{1}{4}(3-\sqrt{17})$	$x_2 = \frac{1}{4}(3+\sqrt{17})$	$x_3 = 3$	$+\infty$
f'		$-$ 0 $+$	0 $-$	0 $+$	
f	$+\infty$	$\searrow$ $f(x_1)$ $\nearrow$	$f(x_2)$ $\searrow$	0 $\nearrow$	$+\infty$

TABLE 3.A: Variation table for the asymmetric double well potential and its derivative (Equations 3.1) –3.1)

Error estimate An error estimate on the reconstructed free energy surfaces for the symmetric double well potential was also performed during the resubmission of the article. A reviewer wanted to know what was the magnitude of the statistical errors when reconstructing the surface.

Figure 3.2 shows the distribution of the confidence intervals for a barrier of height $1k_BT$ by using a Bootstrapping approach. The number of steps is 10^6, the error is of the same order for all SA-MC simulations. One finds negligible statistical errors (around a hundredth of k_BT) for most of the surface. The conclusion was that differences after FES reconstruction were directly caused by an unwise choice of the SA-MC parameters.

3.2 LJ_N clusters

The second test application was to study Lennard-Jones clusters (LJ_N), which are geometrically stable arrangements of rare gases atoms in vacuum and at low temperature. For an increasing number N of atoms the total number of local minima growths dramatically fast, and locating the lowest possible energy configuration is challenging. It is specially the case for clusters 31 and 38 which are characterised by an irregular geometry. In the article → 3.Applications → 3.2. Global Minima of Lennard-Jones Clusters, results are presented for $\{N = 13, 19, 31, 38, 55, 75\}$.

Figure 3.3 shows the lowest energy minima found for clusters of size $\{N = 7, 13, 19, 31, 38, 55\}$ using the SA-MC algorithm. Trajectories were generated with the above mentioned stand-alone C code. Geometries and energies agree with the database of lowest energy minima established by David Wales et al.[113, 114]

3.3 Alanine Dipeptide

The third test system consisted in a study of the conformational equilibria of the alanine dipeptide. Free energy surfaces (FES) were built for implicit and explicit solvent, based on the histogram method introduced in Section 1.3.3 were the two reaction coordinates are the (ϕ, ψ) dihedral angles. SA-MC sampling is compared to conventional MD and MC sampling. Biased and unbiased (following procedure from Section 2.4) FES are both displayed for SA-MC. From the FES slices and values of the free energy are extracted, an compare favourably to literature.

Please consult the article → 3.Applications → 3.3. and 4.Discussion and outlook, for more details.

3.4 SA-MC article

Published in August 2014, in the *Journal of Chemical Theory and Computation* (JCTC), Vol. 10, Pages 4284–4296, [112] co-written with Nuria Plattner, Jimmie. D. Doll and Markus Meuwly.

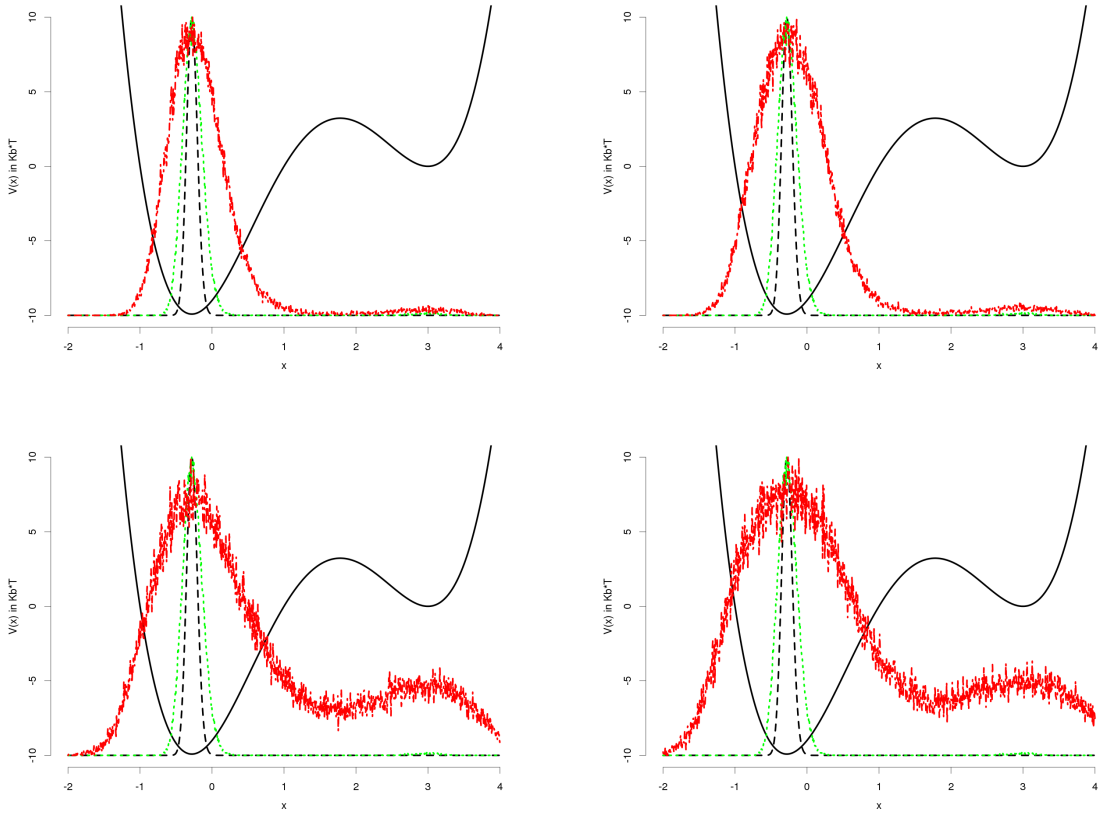


FIGURE 3.1: Representation of the potential $V(x)$ (black), the theoretical density $\rho(x)$ (dashed black), the MC density $\rho(x, MC)$ (dashed green) and the spatial averaging densities $\rho(x, \varepsilon)$ (red), for different sets of parameters $(\varepsilon, N_\varepsilon)$: (a) $(\varepsilon = 0.1, N_\varepsilon = 10)$ and (b) $(\varepsilon = 0.2, N_\varepsilon = 10)$ (c) $(\varepsilon = 0.1, N_\varepsilon = 30)$ and (d) $(\varepsilon = 0.2, N_\varepsilon = 30)$

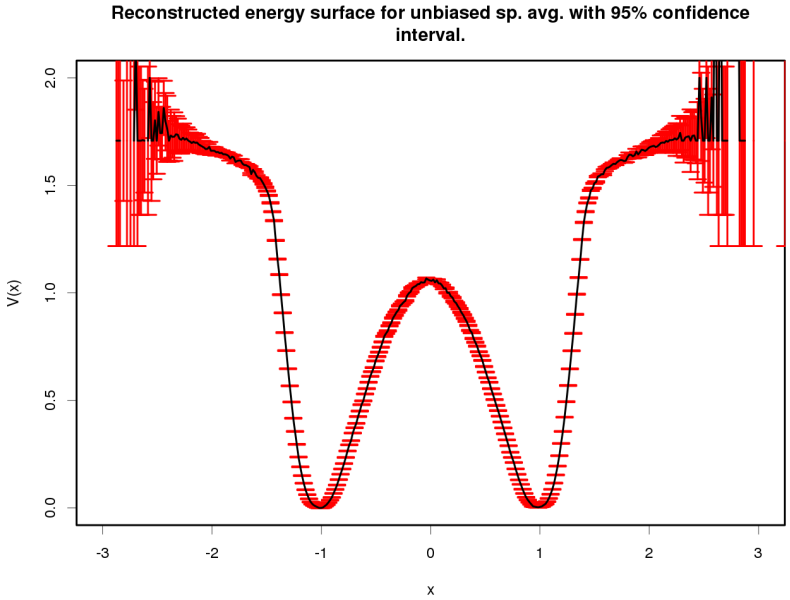
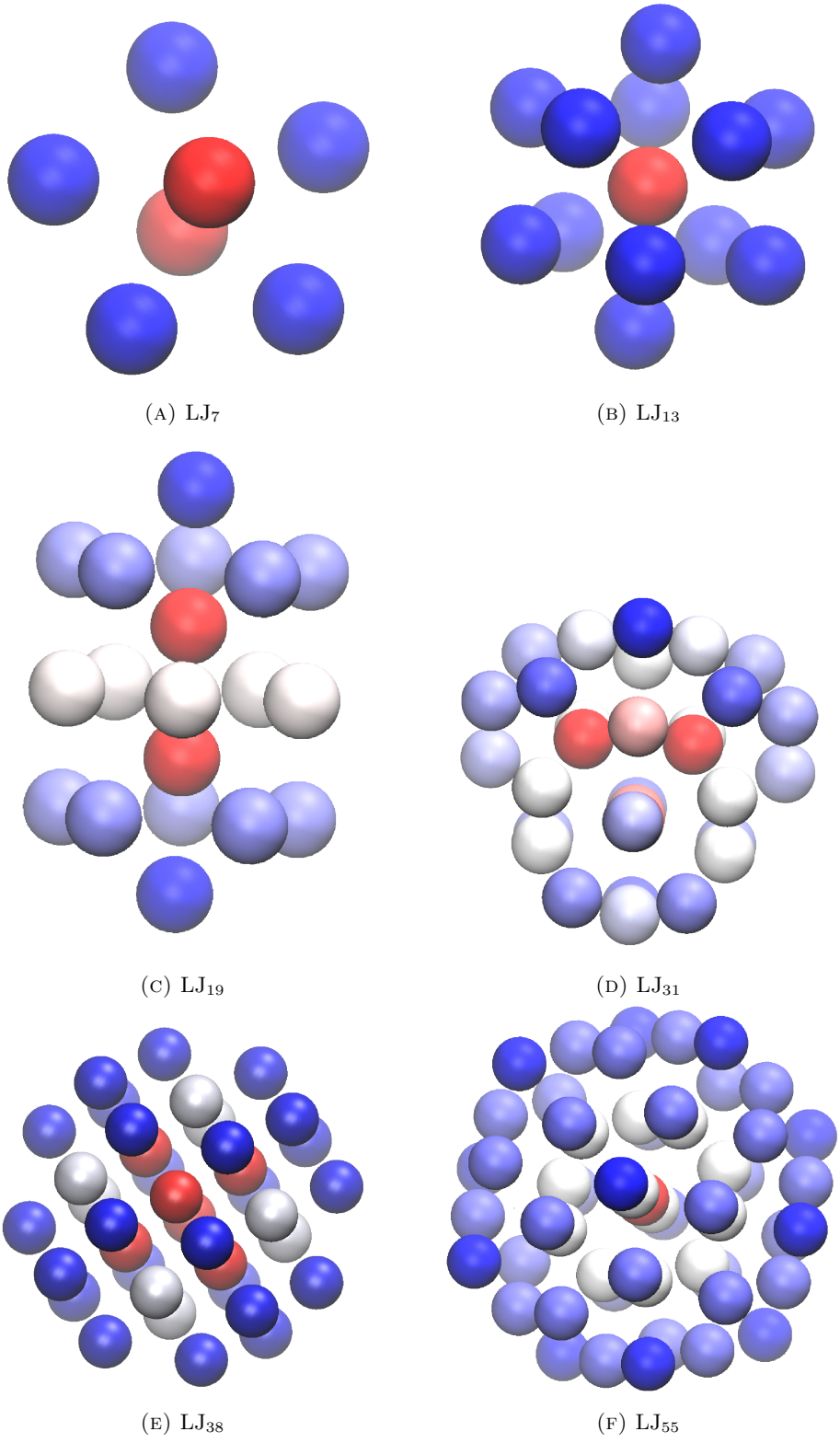


FIGURE 3.2: SA-MC error estimate: the 95% confidence intervals are show using red bars. Estimated from a boostrapping procedure



(A) LJ_7 (B) LJ_{13}

(C) LJ_{19} (D) LJ_{31}

(E) LJ_{38} (F) LJ_{55}

FIGURE 3.3: Lowest energy minima found using the SA-MC method for LJ_N clusters, $\{N = 7, 13, 19, 31, 38, 55\}$. Colour denotes distance from the centre of mass of the cluster, from red for the closest to dark blue for the most distant.

JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Spatial Averaging: Sampling Enhancement for Exploring Configurational Space of Atomic Clusters and Biomolecules

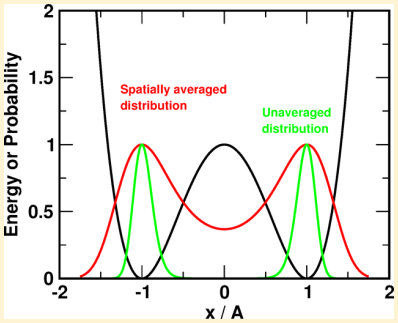
Florent Hédin,[†] Nuria Plattner,[‡] J. D. Doll,[§] and Markus Meuwly*[,†,§]

[†]Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland

[‡]Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, D-14195 Berlin, Germany

[§]Department of Chemistry, Brown University, Providence, Rhode Island 02912, United States

ABSTRACT: Spatial averaging Monte Carlo (SA-MC) is an efficient algorithm dedicated to the study of rare-event problems. At the heart of this method is the realization that from the equilibrium density a related, modified probability density can be constructed through a suitable transformation. This new density is more highly connected than the original density, which increases the probability for transitions between neighboring states, which in turn speeds up the sampling. The first successful investigations included the diffusion of small molecules in condensed phase environments and characterization of the metastable states of the migration of the CO ligand in myoglobin. In the present work, a general and robust implementation including rotational and torsional moves in the CHARMM molecular modeling software is introduced. Also, a procedure to estimate unbiased properties is proposed in order to compute thermodynamic observables. These procedures are suitable to study a range of topical systems including Lennard-Jones clusters of different sizes and the blocked alanine dipeptide $(Ala)_2$ in implicit and explicit solvent. In all cases, SA-MC is found to outperform standard Metropolis simulations in sampling configurational space at little extra computational expense. The results for $(Ala)_2$ in explicit solvent are in good agreement with previous umbrella sampling simulations.



1. INTRODUCTION

Monte Carlo (MC) methods[1] are widely used in modern computer simulations to study high-dimensional, many-body systems.[2] One of their key features is their dimensional tolerance that makes it possible to study large systems with a significant number of degrees of freedom. Furthermore, when applied to atomic systems, and by choosing an appropriate statistical mechanical ensemble, MC simulations are useful in estimating the partition function, from which thermodynamic properties can be determined.

Despite their general utility, MC methods have practical limitations, one of which is related to rare-event sampling, which is a particular challenge.[3] Conventional stochastic methods typically use random walk procedures for generating a statistical sampling of the desired equilibrium probability distribution, useful for obtaining numerical estimates. For systems in which configuration space is well connected, standard techniques such as the Metropolis−Hastings approach[4,5] are efficient. However, often configuration space decomposes into poorly connected subregions, which makes realistic and exhaustive sampling problematic, and sampling needs to be enhanced. Several strategies have been developed in the past to address the rare event sampling problem, including parallel tempering (PT),[6] umbrella sampling (US),[7] metadynamics,[8] or replica exchange (RE).[9] These techniques either use a bias to drive the system from one region in configuration space to another, neighboring region (US, metadynamics), whereas PT and RE—which are related to each other—expand thermodynamic state space. A broader overview of these techniques has been presented recently in the literature.[2,3] Broadly speaking, the available techniques fall in one of the three following categories:

(i) Trial move optimizations, as the displacement vector MC technique,[10] or more recent studies specifically aiming at MC simulations of proteins,[11] but this kind of parameter tuning requires some a priori knowledge about the "shape" of the underlying potential energy surface.

(ii) Parallel tempering,[6] replica exchange,[12] and infinite swapping methods,[13−17] which are based on repeated information exchange between copies of the simulation system, which are run at different values of an external control parameter (such as temperature). Lower-temperature replicas are enriched with knowledge coming from higher-temperature ones where high-energy barriers are more easily crossed. Several strategies for defining the tempering ensemble have been discussed.[18−21]

(iii) Through the addition of an external bias, such as a supplementary potential for filling basins of energy surface (metadynamics,[8] flooding[22,23]), auxiliary probability density (Tsallis weight sampling),[24−26] energy smoothing methods,[27−30] or constrained geometry (umbrella sampling).[7]

Received: June 18, 2014
Published: August 29, 2014

dx.doi.org/10.1021/ct500529w | J. Chem. Theory Comput. 2014, 10, 4284−4296

Spatial averaging MC (SA-MC) sampling belongs to this last category, where a new family of probability density functions are constructed.[31] Until now, SA-MC has been applied to model systems and in special applications[31,32] such as the diffusion of small molecules in condensed phase environments. The aim of the present work is (i) to introduce a general and robust implementation of SA-MC into CHARMM;[33] (ii) to generalize the available move set to include rotations and torsions; (iii) to investigate the possibility of determining unbiased thermodynamic properties from SA-MC in order to extract approximate thermodynamic information from the simulations; (iv) to apply SA-MC to the well-known problem of finding the optimal geometry of Lennard-Jones clusters (it is of particular interest to compare the efficiency in terms of the number of MC-steps compared to Metropolis sampling and the relative CPU requirements of the two approaches); and (v) to apply SA-MC to the conformational sampling of the blocked alanine dipeptide in implicit and explicit solvent.

2. COMPUTATIONAL METHODS

2.1. Spatial Averaging MC. In the canonical (NVT) ensemble, the probability $\rho(\mathbf{X})$ of observing a given system in state $\mathbf{X}$ is related to its energy $V(\mathbf{X})$ through

$$\rho(\mathbf{X}) = \frac{1}{Z} e^{-\beta V(\mathbf{X})} \tag{1}$$

where $\mathbf{X} = X_1,...,X_k$ is a k-dimensional vector of coordinates (where $k = 3$ for MC or $k = 6$ for MD), populating a subset D of the configuration space $\mathbb{R}^{kN}$, Z is the canonical partition function $Z = \int^{D\subset\mathbb{R}^{kN}} e^{-\beta V(\mathbf{X})}\, d\mathbf{X}$, and $\beta = 1/k_B T$ is the inverse temperature and k_B the Boltzmann constant.

Monte Carlo (MC) methods[1] are one powerful way for sampling the high dimensional integral Z which runs over $3N$ degrees of freedom for a general Euclidean 3-space and for an N−particle system. The Metropolis−Hastings approach was specifically designed for addressing this problem when considering the canonical ensemble. Initially proposed for sampling the Boltzmann distribution,[4] it was later extended to nearly all sampling problems.[5] In practice, a system $\mathbf{X}$ is stochastically modified leading to a new configuration $\mathbf{Y}$. Based on the energy difference $\Delta E = V(\mathbf{Y}) - V(\mathbf{X})$ the probability of accepting the new configuration is then

$$P_{acc} = \min\{1, e^{-\beta \Delta E}\} \tag{2}$$

For high energy barriers, the term $e^{-\beta\Delta E}$ is close to zero, and the probability of accepting such a move is extremely low. Previously introduced methods (PT/RE, US, metadynamics) addressed this problem by proposing a physical modification of the system (e.g., a set of temperatures for PT/RE). With SA-MC, increased sampling is achieved by directly modifying the underlying probability density function.[31,32] In a one-dimensional notation, if the density to be sampled is $\rho(x)$, a new set of modified densities is obtained by writing

$$\rho(x, \varepsilon) = \int_D P_\varepsilon(y) \exp(-\beta V(x + y)) dy \tag{3}$$

where $P_\varepsilon(y)$ is a normalized probability distribution with characteristic length scale ε. The parametrization of $P_\varepsilon(y)$ is that of a Gaussian distribution with standard deviation ε. Adjusting this parameter allows to adapt the biasing distribution to the particular problem of interest. In practice, the convolution of the true distribution with $P_\varepsilon(y)$ will decrease the barriers of $V(x)$ and hence accelerate sampling of neighboring minima if ε is appropriately chosen. Furthermore, the Gaussian transform of the potential is centered around $\rho(x)$ so the integrals of the original and the transformed density are equal

$$\int_D \rho(x) dx = \int_D \rho(x, \varepsilon) dx \tag{4}$$

Equation 4 is key to SA-MC, as it implies that thermodynamic properties derived from the modified density are related to those corresponding to the original density $\rho(x)$ for a given temperature. Let $\langle f(x)\rangle_0$ be a thermodynamic property (where the subscript 0 denotes an unbiased value) estimated through an average of the form

$$\langle f(x)\rangle_0 = \frac{\int_D \rho(x) f(x) dx}{\int_D \rho(x) dx} \tag{5}$$

By combining eqs 4 and 5, this average can be expressed by using the modified densities:

$$\langle f(x)\rangle_0 = \frac{\int_D \rho(x, \varepsilon) \left(\frac{\rho(x)}{\rho(x, \varepsilon)} f(x)\right) dx}{\int_D \rho(x, \varepsilon) dx}$$

which can be simplified to

$$\langle f(x)\rangle_0 = \left\langle \left(\frac{\rho(x)}{\rho(x, \varepsilon)} f(x)\right)\right\rangle_\varepsilon \tag{6}$$

Hence, $\langle f(x)\rangle_0$ is expressed as an accumulated average of the instantaneous value $f(x)$ weighted by the ratio between the original and spatially averaged densities. Hence, the unbiased thermodynamic property of interest can be estimated.

As an example the Helmholtz Free Energy F as a thermodynamic function of state (ensemble NVT) is considered. The unbiased value F_0 estimated from a SA-MC simulation is

$$F_0 = F_\varepsilon \frac{\rho(x)}{\rho(x, \varepsilon)} \tag{7}$$

where F_ε is a biased estimate of F. In practice, the value of F for a given configuration x is estimated by counting the number of occurrences n of the configuration over all the sampled configurations N, which yields $\beta F = -\ln(n/N)$. By introducing a reference value F^0 for the free energy (for example the most stable configuration sampled), and by choosing a correct metric or reduced coordinates, it is possible to generate a surface of $\Delta F = F - F^0$: such energy landscapes are a powerful way of visualizing the configuration space and particularly useful for localizing minima regions, barriers, and saddle points.[34]

As usual, when providing an estimate of any property it is important to also quantify the underlying statistical error. According to eq 7 errors in the estimates originate from (i) the error on F when counting the configurations and denoted as $\sigma(0) = -k_B T(\sigma(\rho(x))/\rho(x))$, and (ii) the error in the unbiasing ratio, directly related to the statistical variance on the spatially averaged densities. This variance can be estimated according to[31]

$$\sigma^2(\varepsilon) = \frac{\rho(x)}{\rho(x, \varepsilon)} \left(\frac{\rho(x)}{\rho(x, \varepsilon)} - 1\right) F_\varepsilon \tag{8}$$

dx.doi.org/10.1021/ct500529w | J. Chem. Theory Comput. 2014, 10, 4284−4296

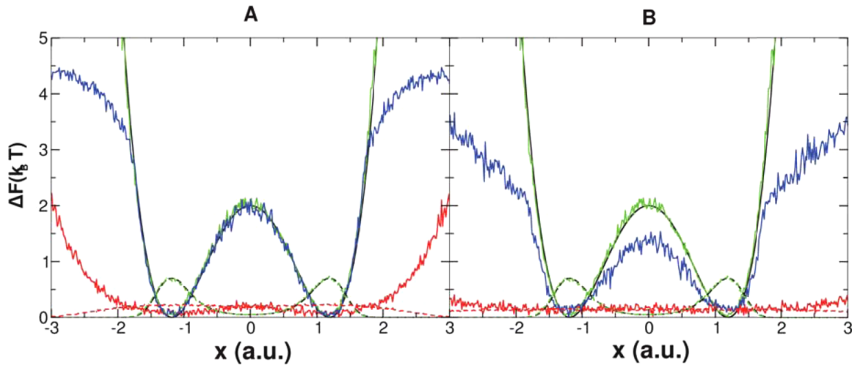


Figure 1. Reconstructed energy surface for the double well potential. Solid lines are for the surfaces ΔF and dashed lines are the corresponding densities $\rho(x)$. Left panel for (W_ε = 0.4 and N_ε = 10), right panel for (W_ε = 0.8 and N_ε = 25). Color code: analytical results (black), Metropolis MC (green), biased SA-MC (red), unbiased SA-MC (blue). a.u. = arbitrary units of distance.

The total error on the estimate of F_0 is

$$\sigma_{F_0} = \frac{1}{\sqrt{N}}(\sigma(0) + \sqrt{\sigma^2(\varepsilon)}) \tag{9}$$

The $N^{-0.5}$ dependency in eq 9 is inherent to stochastic sampling methods.[35] However, by dividing the data in k data sets of a given size M (with $N = kM$) and by averaging over such blocks, the error can be reduced. More precisely, bootstrapping[36−38] whereby only part of the data—randomly chosen from the overall distribution sampled by these four simulations—will be employed to estimate the error in the free energy profiles.

2.2. Algorithm and Implementation into CHARMM.

The extension of SA-MC to multidimensional molecular systems has been successfully applied to the diffusion of small molecules (H_2 and CO) in condensed phase environments.[32] This first algorithmic implementation was limited to translational and rotational moves, which makes possible to study diffusion processes. This will be generalized in the present work to also allow treatment of the configurational space of systems such as peptides and proteins.

The MC module[39] in CHARMM[33] is suitable for such an implementation as it allows the user to define an arbitrary set of moves for optimizing the sampling of a given molecular system. The main types of moves are (i) rigid translations of one or more atoms (RTRN), (ii) rigid rotations of one or more atoms around a center of rotation: this center may be another set consisting of one or more atoms, or the center of mass of the rotating atoms (RROT), (iii) dihedral angles torsions (TORS), and (iv) concerted rotations of dihedral angles (CROT). The current implementation handles (i−iii) in the NVT ensemble in explicit or implicit solvent. The present simulations were carried out with both the Analytical Continuum Electrostatics (ACE)[40,41] implicit water model and the TIP3P[42] explicit water model.

Starting from a trial configuration $\vec{x}_0$ of the system, a Gaussian distribution for M_ε sets of N_ε configurations with standard deviation W_ε, centered around $\vec{x}_0$ is generated in SA-MC.[32,43] The chosen MC move—such as translation or rotation—is then applied to all M_ε*N_ε configurations and the corresponding energies $E^{(m,n)}_{\text{new}}$ are determined. Two sets of Boltzmann weights are then computed, one for the old and one for the new configurations: $E^{(m,n)}_{\text{old,Boltz}} = e^{-\beta E^{(m,n)}_{\text{old}}}$ and $E^{(m,n)}_{\text{new,Boltz}} = e^{-\beta E^{(m,n)}_{\text{new}}}$. For each set M_ε, the difference between the aggregated old and new weights is determined: $\delta_m = \ln(S^m_{\text{new}}/S^m_{\text{old}})$ where $S^m = \sum^{N_\varepsilon} E^{(m,n)}_{\text{Boltz}}$. Adding up all the δ_m yields $\delta = (1/M_\varepsilon)\sum^{M_\varepsilon}\delta_m$ from which also a variance $\sigma^2 = (1/M_\varepsilon(M_\varepsilon - 1))\sum^{M_\varepsilon}(\delta_m - \delta)^2$ can be computed. These quantities are then used for a modified acceptance/rejection criterion $\xi < \exp(-\beta(\delta + (\sigma^2/2)))$ (see eq 2).

For each MC move type, the new configurations have to be generated in the corresponding configurational space, such as for angle moves in the angular space around the initial $\vec{x}_0$. This is accomplished as follows. (i) For a rigid translation, the procedure consists of adding a Gaussian distributed random number with zero mean and standard deviation W_ε to the coordinates of the atoms that were selected for a particular move. (ii) For a rotation of a group of several atoms with coordinates $\mathbf{X}$, a random angle θ, normally distributed between $-\theta_{\max}$ and $+\theta_{\max}$, is generated and the corresponding new coordinates are $\mathbf{X}' = \mathbf{RX}$ where $\mathbf{R}$ is a rotation matrix. (iii) Dihedral angles, defined as the intersection of two planes formed by four atoms, are also altered by drawing from a normal distribution and again by finding the Euler rotation matrix for the set of all atoms which are involved in the dihedral angle.

The ratio $\rho(x)/\rho(x,\varepsilon)$, as used in eqs 6 and 7, is required for determining unbiased thermodynamic properties and is optionally stored for each frame of the trajectory in a dedicated file. This data can then be used in postprocessing from which the unbiased free energy and other observables can be estimated.

3. APPLICATIONS

In the following sections, SA-MC is applied to a range of three typical rare-event sampling problems, and its efficiency is compared to reference simulations, including standard Metropolis sampling. First, the current implementation together with the unbiasing procedure is tested on the double well potential to obtain thermodynamic properties.[31] In the second example, the minimum energy structures of Lennard-Jones clusters are considered with particular focus on how to rapidly find the lowest energy configuration of such systems. The third and final example is the study of the free energy landscape of the blocked alanine dipeptide, which highlights the efficiency of SA-MC. For the first two examples, the simulations are performed with a dedicated code specifically written for the application whereas the third system is studied with the generalized CHARMM implementation described above.

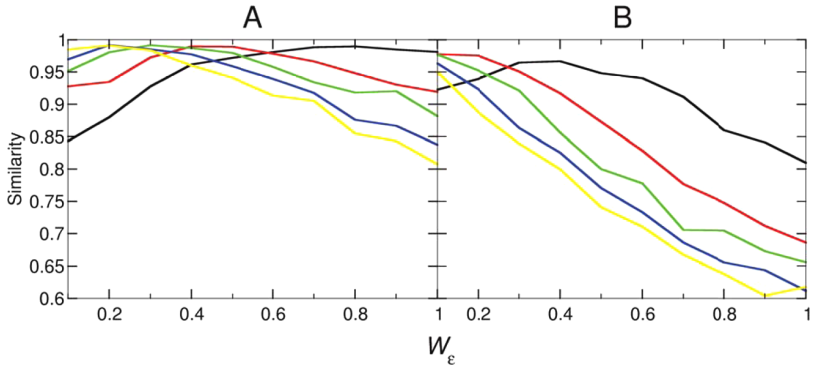


Figure 2. Similarity between reconstructed and theoretical surface for (a) barrier of $2k_BT$ or (b) barrier of $5k_BT$. Color code: $N_\varepsilon = 5$ (black), $N_\varepsilon = 10$ (red), $N_\varepsilon = 15$ (green), $N_\varepsilon = 20$ (blue), $N_\varepsilon = 25$ (yellow).

3.1. Double Well Potential. To illustrate the efficiency of SA-MC but also the need for unbiasing when estimating thermodynamic properties, first a one-dimensional problem involving a simple double-well potential is studied. Explicitly, $V(x) = (x^2 - \sqrt{\lambda})^2$, where λ is the height of the barrier separating the two minima, which are located at $\pm(\lambda)^{1/4}$. Reduced units are used throughout which makes temperature dimensionless and energies are given in units of k_BT

For a given temperature, the probability density of sampling x is $\rho(x) \propto \exp(-\beta V(x))$. Sampling $V(x)$ is sufficiently straightforward for low barriers that conventional MC yields the correct free energy profile. Therefore, the sensitivity of SA-MC to various choices of W_ε and N_ε ($M_\varepsilon = 1$ in the present application) can be tested. For a reduced temperature of $\beta = 0.75$, 10^6 MC steps, and barrier heights between 1 and 10 simulations were carried out by using conventional MC and SA-MC. For the latter, $0.1 \leq W_\varepsilon \leq 1.0$ in increments of 0.1 and $5 \leq N_\varepsilon \leq 25$ in increments of 5.

The free energy curves are reconstructed and unbiased as explained in the computational methods part. For quantifying the similarity between the sampled density $\rho^\alpha(x)$ and the true normalized density $\rho(x)$ a score S^α is introduced:

$$S^\alpha = \frac{\int_{-\infty}^{+\infty} \rho(x)\rho^\alpha(x)\mathrm{d}x}{\int_{-\infty}^{+\infty} \rho(x)\rho(x)\mathrm{d}x} \tag{10}$$

where α = MC or SA-MC, respectively, and $\rho(x)$ is the true, normalized density. Hence, S^α measures the overlap between the sampled densities and the true Boltzmann density. For perfect sampling one should find $S = 1$.

Figure 1a is an example of reconstructing the FES for a barrier height of $\Delta F = 2k_BT$, where the theoretical surface $V(x)$ and the results of the Metropolis sampling (which overlaps ideally) are presented both with results from SA-MC with $W_\varepsilon = 0.4$ and $N_\varepsilon = 10$. Although SA-MC itself only poorly samples the reference FES, unbiasing as discussed in the Methods yields a very realistic FES (compare black and blue traces). Changing the parameters to $W_\varepsilon = 0.8$ and $N_\varepsilon = 25$ (Figure 1b) leads to almost uniform sampling with SA-MC (red trace). Despite this, the reconstructed, unbiased FES can capture the shape of the true FES although the free energy barrier is underestimated. This already highlights that SA-MC can be effectively used—even with unoptimized parameters W_ε and N_ε—to characterize the true *shape* of the free energy surface although barrier heights may only be qualitatively correct.

In a next step, the reconstructed (unbiased) FESs from SA-MC are further characterized, in particular with regards to the parameters W_ε and N_ε. For example, if the width W_ε is too large, all information about the existence of local minima is washed out. Such considerations are of particular importance when applying SA-MC to a problem for which the underlying FES is incompletely or poorly characterized, that is, in cases where the positions and relative stabilizations of the minima are largely unknown. Figure 2a reports the similarity (estimated by using eq 10) between the reference and the unbiased SA-MC FES for barrier height $\lambda = 2$, $0.2 \leq W_\varepsilon \leq 1.0$ and $5 \leq N_\varepsilon \leq 25$. For the present case, increasing W_ε improves the results initially for most N_ε. However, beyond $W_\varepsilon = 0.4$, the overlap between the reference and the SA-MC FES deteriorates. Hence, the sampling becomes less reliable. This is even more so for a larger barrier ($\lambda = 5$, panel b) for which small values of W_ε give the best results.

This finding can be interpreted as follows. W_ε is the width of the Gaussian distribution, that is, how far from the original configuration a new one will be generated, whereas N_ε is the number of those additional configurations. Increasing both parameters increases the number of configurations generated, which are more and more distant from the original one, resulting in a large variance which causes an inaccurate estimate of the free energy. This is illustrated in Figure 3, where the free

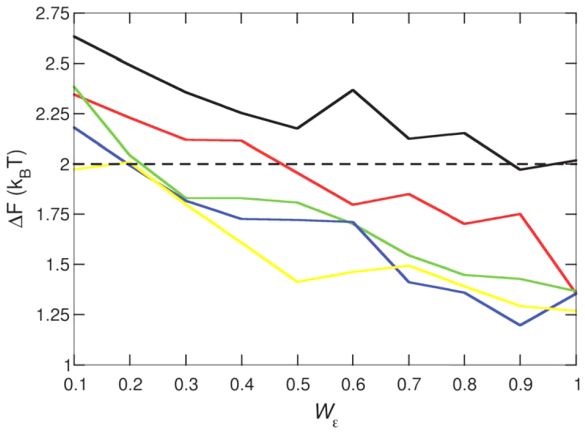


Figure 3. Unbiased barrier energy ΔF (reference value is $\Delta F = 2k_BT$, dashed black line) as a function of W_ε. Systematic errors are of the order of $k_BT/100$. Color code (plain lines): $N_\varepsilon = 5$ (black), $N_\varepsilon = 10$ (red), $N_\varepsilon = 15$ (green), $N_\varepsilon = 20$ (blue), $N_\varepsilon = 25$ (yellow).

dx.doi.org/10.1021/ct500529w | J. Chem. Theory Comput. 2014, 10, 4284−4296

energy at the top of the barrier ($\Delta F(x = 0) = 2k_BT$) is reported for different sets of parameters N_ε and W_ε. For small values of N_ε it is necessary to increase W_ε for obtaining the correct value of $\Delta F = 2k_BT$. With larger N_ε, a value of $W_\varepsilon = 0.2$ is sufficient, and further increasing the Gaussian width will result in a less accurate value for $\Delta F(x = 0)$.

In this first application, it is found that the bias introduced by SA-MC is a powerful feature that can more readily connect densities in local minima, separated by a barrier which is difficult to overcome with standard MC sampling. Furthermore, it is shown that the bias can be accounted for over a certain system parameter space (W_ε and N_ε) to faithfully reconstruct the true, underlying free energy profile. The degree to which this is possible depends on the system and the parameters chosen.

3.2. Global Minima of Lennard-Jones Clusters.

Lennard-Jones (LJ) clusters are an ideal class of systems to which MC-based sampling approaches can be applied. Some of the problems can be exhaustively sampled whereas others are computationally too demanding for this. Here, SA-MC is applied to determine low-energy configurations of LJ clusters of different sizes. The particular focus for this example is (i) whether or not the global minimum as known from the literature is found at all and (ii) the speed with which the global minimum is found. This also motivates the comparison of conventional MC and SA-MC in the present context. However, it should be mentioned that more established algorithms exist for global optimization.[44] LJ clusters are an ensemble of nonreactive atoms in vacuum (for example noble gases), interacting only through Lennard-Jones[45] potentials

$$V^{\mathrm{LJ}} = 4\epsilon \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] \tag{11}$$

Here, r_{ij} is the distance between atoms i and j, ϵ is the depth of the potential well, and σ the distance at which $V^{\mathrm{LJ}} = 0$. Again, reduced units are employed, that is, $\epsilon = \sigma = 1$, and the energy will be reported in units of ϵ. Extensive previous literature on these systems is available, and a Web site[46] provides a collection of known structures, lowest minima, and symmetry groups for clusters ranging from 2 to 1610 atoms: several MC methods,[47−51] quantum calculations,[52] MD simulations,[53−55] parallel tempering,[56,57] or others such as discrete path sampling[58,59] were used for characterizing the systems, and LJ_N (with N the number of atoms) clusters became reference systems for methods dedicated to finding global minima. The number of local minima grows exponentially as a function of N, and hence, determining the global minimum of such clusters is a computationally challenging problem. As an example, between $N = 2$ and $N = 33$ the number of known minima increases from 1 to $\approx 4 \times 10^{14}$. Nevertheless, some recent studies were able to treat the broken ergodicity and then provide the correct number of minima for the LJ_{31} and LJ_{75} clusters.[60]

The low energy minima for various LJ_N clusters were investigated by both, conventional MC and SA-MC. Specifically, the systems included $N = 13$, 19, 31, 38, 55, and 75. Some of the systems are relatively "easy" while others—such as LJ_{38}, see Figure 4—are known to be very challenging (see below). The methodology applied for all coming examples is as follows: (i) 10^4 independent runs are started from the same initial (random) configuration. (ii) At the end of each step, if the energy difference relative to the reference configuration is less than 5ϵ the system is minimized, and if the known lowest energy minimum structure is obtained (tolerance of $10^{-4} \times \epsilon$) the calculation is stopped and considered as converged; otherwise, the simulation continues. (iii) If the global minimum is not reached after a given number of steps (depending of cluster size) the simulation is considered to be not converged.

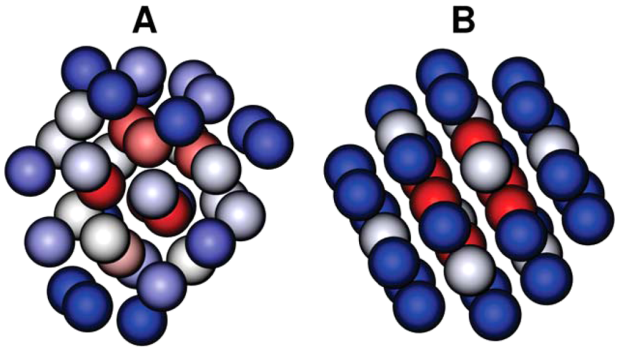


Figure 4. Lowest energy configurations found for the LJ_{38} atoms, obtained with SA-MC. The structure in panel a has an energy of $E = -171.357\epsilon$ (see Table 1) whereas the structure in panel b is that of the absolute minimum ($E = -173.928\epsilon$) found when starting from the cluster LJ_{37} and randomly adding an atom. Red atoms are closer to the center of mass of the cluster than gray and blue ones.

For LJ_{13}, the global minimum has an energy of $E = -44.327\epsilon$ (see Table 1). Figure 5 shows a cumulative distribution of the required number of steps before reaching the global minimum for conventional MC and SA-MC with different parameters $[W_\varepsilon;M_\varepsilon;N_\varepsilon]$. After the 10^6 MC steps considered here, only 24% of the MC simulations are able to locate the global minimum energy structure. This compares with between 50% and 98% for SA-MC, depending on the choice of M_ε and N_ε. In general, SA-MC outperforms conventional MC considerably. For $W_\varepsilon = 0.25$ and 0.5 a clear improvement is observed, as almost all simulations converge before 2.5×10^5 steps. Results are particularly noteworthy with $W_\varepsilon = 0.5$ for which 98% of the simulations converged during the first 10^5 steps. For $W_\varepsilon = 1.0$ the convergence speed slows down. One explanation is that depending on the value for W_ε—typically the larger W_ε the flatter the FES—the SA-MC-modified densities become too connected which changes the topology of the FES such as to slow down convergence. Nevertheless, this may be corrected by using increased values of M_ε and N_ε, which leads to variance reduction. However, the computational time would also increase.

The previous conclusions are supported by an analysis of the median of the number of steps for converged simulations, that is, the value for which 50% of the calculations reach the minimum energy structure. For conventional MC, this value is 3.3×10^5 compared to 6×10^4, 2×10^4, and 3×10^5 for SA-MC with [0.25;5;5], [0.5;5;5], and [1.0;5;5], respectively. Hence, for the best performing SA-MC simulation, the average number of steps required to reach the global minimum is smaller by a factor of 30 compared to conventional MC. It is also possible to determine the rate at which the various simulations converge by fitting the cumulative successful runs to an empirical relationship $y = d \tanh((ax + b)/d)$ where d describes the asymptotic convergence (plateau of the number of converged simulations, ideally 10000) and a describes the growth of the first part of the curve (i.e., how rapidly the

Table 1. Minimum Energy Configurations (in Units of ϵ) for All LJ Clusters Studied, and Best Convergence Rates Observed, for MC and SA-MC[a]

LJ_N	E_{ref}[47]	E_{MC}	$E_{SA\text{-}MC}$	steps	conv. MC (%)	conv. SA-MC (%)
13	−44.326	−44.326	−44.326	10^6	24	98
19	−75.659	−75.659	−75.659	10^6	22	97
31	−133.586	**−126.081**	−133.586	10^8		26
38(a)	−173.928	**−160.556**	**−171.357**	10^8		
38(b)	−173.928	−173.928	−173.928	10^9	3	35
37 + 1	−173.928	**−170.807**	−173.928	5×10^7		6
55	−279.248	−279.248	−279.248	10^8	16	65
75	−397.492	**−381.173**	−397.492	10^8		28

[a]Reference values are from the literature.[47] Numbers in bold face are unconverged values. The "steps" column indicates how long were both MC and SA-MC simulations. What differs between 38(a) and 38(b) is the number of steps. 37 + 1 means that the starting point is the lowest energy geometry of LJ_{37} to which a 38th atom is included.

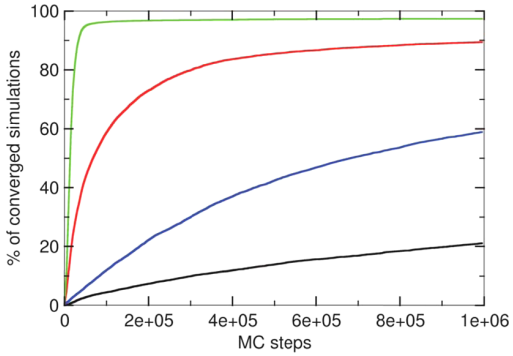


Figure 5. Convergence analysis for MC (black), and SA-MC simulations with different sets of parameters $[W_\epsilon;M_\epsilon;N_\epsilon]$: [0.25;5;5] (red), [0.5;5;5] (green), and [1.0;5;5] (blue).

plateau is reached). The parameter b ensures that the fit passes through the origin. For conventional MC, a = 1.2 compared with 5.6, 48.2, and 1.6 from SA-MC, which quantifies the above observations about the median. For parameter d, the fit yields 2400, 8700, 9700, and 6200 for the four simulations. This, together with the observations for parameter a suggests that the *rate* of successful runs for the worst SA-MC simulation is still better than that of conventional MC whereas the *number* of successful runs is larger by a factor of 3. On the other hand, the best performing SA-MC simulation is about 30 times as efficient while finding at the same time the global minimum in almost all simulations (98%).

The computational overhead in using SA-MC is in the $M_\epsilon \times N_\epsilon$ additional energy evaluations which, however, can be easily parallelized. In the present case, a factor of 5 × 5 = 25 is expected for a given number of MC steps (here 10^6). If all 10^4 simulations are run for 10^6 MC steps, SA-MC with [0.5;5;5] is 23-times slower than conventional MC. However, if simulations are terminated when the lowest minimum is found, this reduces to a factor of 1.2. Hence, in cases where suitable termination criteria can be found, the computational overhead of SA-MC is well below an order of magnitude compared to conventional MC with the added value of the much increased likelihood for locating the correct lowest energy configuration.

For the larger LJ clusters, Metropolis MC simulations have difficulties in successfully locating the known minima at all. In order to assess the performance of SA-MC for such cases, additional simulations were carried out for LJ_{19}, LJ_{31}, LJ_{38}, LJ_{55}, and LJ_{75}. The same procedure as before is used except for the total number of MC or SA-MC steps, which was increased to 10^9 for larger clusters. For LJ_{19}, the convergence speed analysis gives results similar to those presented in Figure 5; that is, SA-MC reaches the global minimum (energy -75.659ϵ, see Table 1) in much fewer steps than regular Metropolis, when using 10^6 steps. The LJ_{55} and LJ_{75} are much larger systems and Metropolis sampling is extremely slow to obtain converged results. Table 1 shows that for LJ_{55} MC and SA-MC converge to the reference value from the literature, the convergence rate for MC is 15.8% with a median number step required of 7.6×10^7 (the maximal number of steps being 10^8), and for SA-MC the numbers are 65% and 5.4×10^6, respectively, that is, one order of magnitude faster when just considering the number of steps. For LJ_{75}, conventional MC sampling is unable to locate the global minimum within 10^8 steps. On the contrary, SA-MC does find this minimum for 28% of the simulations within a median number of steps of 5.1×10^7.

The LJ_{31} and LJ_{38} clusters are known for their funneled energy landscape.[47,49,58,61,62] LJ_{38} is a particularly interesting system as it has a double-funnel landscape, one ending in the global minimum, the other in the second minimum. Doye et al. showed with disconnectivity graphs[49,50] that 446 minima are related to the second funnel but only 28 to the first one, making the transition from one funnel to the other extremely rare. With 10^8 MC steps, our implementation of the Metropolis algorithm was unable to converge to the lowest known minimum for both clusters, which are at -133.586ϵ and -173.928ϵ, respectively, see Table 1. The best configurations sampled in this set of simulations are still 6ϵ and 13ϵ higher in energy than the known minima. Contrary to that, SA-MC successfully converged for the LJ_{31} cluster (see Table 1) with similar sets of parameters as for LJ_{13}, but for LJ_{38} (Figure 4a) the best energy obtained is still 2.5ϵ too high (-171.357ϵ compared to -173.928ϵ, cf. Table 1). A second set of 10 000 simulations for the LJ_{38} cluster was carried out with 10 times more Monte Carlo steps (10^9 instead of 10^8, see Table 1 line 38(b)). In this case, both MC and SA-MC find the known minimum energy structure[47,49,61] for 3 and 35% of the simulations, respectively.

A final set of 10 000 simulations for the LJ_{38} cluster was carried out using a slightly different approach: instead of starting from a fully random initial configuration, the lowest minimum of the LJ_{37} cluster (which was successfully found by SA-MC) was employed and randomly a 38th atom was added to the system. Then, simulations were run for 5×10^7 steps. The lowest energy obtained from the Metropolis algorithm is then -170.807ϵ, which is considerably closer to the best minimum with fewer MC steps (see 38(a) versus 37 + 1 in Table 1), and -173.928ϵ for SA-MC, which is the correct

minimum energy structure (Figure 4b). Overall, it is found that SA-MC successfully converges to the global minimum for all the studied LJ_N clusters, including both funneled clusters (LJ_{38}) and larger clusters such as LJ_{75}.

3.3. Blocked Alanine Dipeptide in Implicit and Explicit Water. The blocked alanine dipeptide (Ac−Ala−N−H−Me, Figure 6) has been used as a test system for computational studies[63−79] of conformational equilibria and free energy landscape reconstruction and analysis. This dipeptide contains many of the structural features of proteins, including the two (ϕ, ψ) dihedrals angles, NH and CO groups capable of H-bond formation, and a methyl group attached to the C_α atom. Successful studies used quantum chemistry, MD and MC simulations, and several conformations were identified:[63,64,67,68,70,79] (i) β, also called C_5, for $(\phi, \psi) \sim (-140°, 150°)$, (ii) C_{7eq} for $(\phi, \psi \sim -90°, 80°)$, (iii) α_R (right-handed α helix) for $(\phi, \psi) \sim (-80°, -60°)$, (iv) α_L (left-handed α helix) for $(\phi, \psi) \sim (60°, 60°)$ and (v) C_{7ax} for $(\phi, \psi) \sim (60°, -60°)$. One suitable way to visualize the free energy landscape for the conformations and the transitions between them is to report an energy surface as a Ramachandran plot.[80] Simulations for the blocked alanine dipeptide were carried out both in implicit (Analytical Continuum Electrostatics (ACE)[40,41]) and explicit solvent (TIP3P[42] water). ACE is known for providing a meaningful description of solvation effects for peptides.[17,67,81,82] In the following, results from simulations with ACE are first described. Next, the simulations in explicit water are summarized.

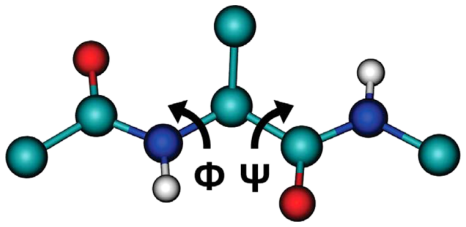


Figure 6. Blocked alanine dipeptide (Ac−Ala−N−H−Me), and the two dihedral angles of interest Φ (C−N−C_α−C) and Ψ (N−C_α−C−N).

Initially, two reference simulations were carried out. They included an MD and a Metropolis MC simulation and served as benchmarks with which to compare the SA-MC simulations. For the latter simulations with a range of parameters $[W_\varepsilon;M_\varepsilon;N_\varepsilon]$ were carried out. In all cases, the blocked alanine dipeptide is treated in a united atom representation (12 atoms, see Figure 6), nonbonded interactions are fully calculated, and

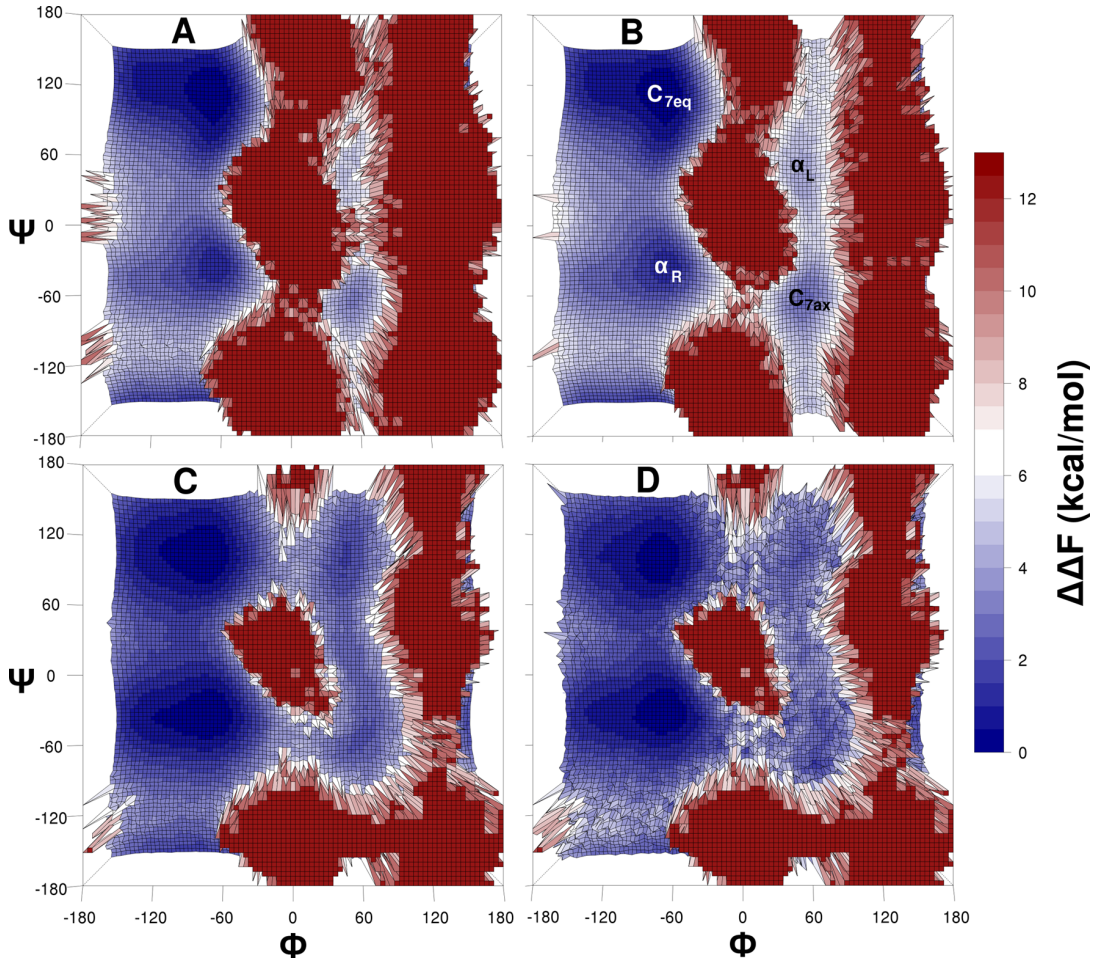


Figure 7. FES of alanine dipeptide: (A) MD, 1.5 μs, 300 K; (B) Metropolis MC, 100 × 10^6 steps, 300 K; (C) biased SA-MC with parameters [0.5;5;5]; (D) unbiased SA-MC (same parameters), 5 × 10^6 steps, 300 K. All free energies are reported relative to the C_{7eq} minimum.

dx.doi.org/10.1021/ct500529w | J. Chem. Theory Comput. 2014, 10, 4284−4296

the temperature is 300 K in the NVT ensemble. For simulations with the ACE implicit solvent, default parameters, such as Born solvation radii, dielectric constants, and atomic volumes, are taken from the literature.[40,41] The MD simulations use the velocity Verlet integrator with the Nosé−Hoover thermostat for a simulation time of 1.5 μs, a cutoff of 12 Å and a time step of Δt = 0.5 fs. The MC simulation was run for 10^8 steps. For SA-MC, simulations with several parameter sets were carried out: (i) $W_\varepsilon \in \{0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0\}$, (ii) $M_\varepsilon \in \{5, 10, 15, 20\}$, and (iii) $N_\varepsilon \in \{5, 10, 15, 20\}$. M_ε and N_ε = 5 or 10 proved to be sufficient for the present purpose (there is no gain with larger parameters justifying the overhead). Here, results for N_ε = 5 are presented.

Simulations in Implicit Solvent. Figure 7 shows the Helmholtz Free Energy Surfaces (FES) for both MD (A) and the Metropolis (B) simulations. It is first observed that both surfaces are quite similar to each other and closely resemble those obtained previously in the literature using the same computational setup.[67,79] The data reported in Figures 7A and B already indicate that the barrier regions between the basins are not well sampled. This is true in particular for the MD simulations. The four following regions (see labels in Figure 7) are sampled sufficiently for providing an estimate of the associated free energy differences: (i) C_{7eq} (top left basin of lowest energy), (ii) α_R (bottom left), (iii) C_{7ax} (bottom right), and (iv) α_L (top right). Positions and estimates for the free energy for those four minima are summarized in Table 2. A

Table 2. Relative Free Energies (kcal/mol) and Minima Locations for the Blocked Alanine Dipeptide at 300 K, for MD, MC, SA-MC Simulations, and Three External References,[64,67,79] All Using the ACE Implicit Solvent Model[a]

	methods			
basin	ΔF MD	ΔF MC	ΔF SA-MC	position (Φ, Ψ)
C_{7eq}	0.00 ± 0.01	0.00 ± 0.02	0.00 ± 0.07	(−83°, 136)
α_R	1.10 ± 0.03	1.16 ± 0.04	0.21 ± 0.06	(−79°, −42°)
C_{7ax}	3.26 ± 0.16	2.91 ± 0.12	3.11 ± 0.08	(67°, −75°)
α_L	4.62 ± 0.32	4.86 ± 0.36	4.12 ± 0.09	(47°, 55°)

	references		
basin	ΔF ref 64	ΔF ref 67	ΔF ref 79
C_{7eq}	0.00	0.00	0.00
α_R	0.71	1.5	0.93
C_{7ax}	4.34	4.1	2.94
α_L	4.35	5.0	4.27

[a]The statistical error was estimated using bootstrapping described previously, and ± values represent a 95% confidence interval. All free energies are shifted relative to the C_{7eq} structure which is the reference energy.

95% statistical confidence interval is provided (see previous description of the bootstrapping procedure) for MC, MD, and SA-MC simulation. The fact that this error is somewhat larger for SA-MC than for MC is caused by the additional error introduced by the unbiasing step of SA-MC (eq 8). Nevertheless, when considering higher energy minima as C_{7ax} and α_L, this value is several times lower than the error estimated for the MD case, where the poor sampling causes an error of 0.32 kcal/mol. Furthermore, the highest error estimated for SA-MC is only 0.09 kcal/mol. It is also of interest to briefly comment on the effect of using bootstrapping for error estimation. For example, directly using eq 9 without bootstrapping leads to an error of 0.18 kcal/mol for the α_L structure with SA-MC, which is reduced to 0.09 kcal/mol when using bootstrapping.

Figure 7C shows the FES from simulations with the SA-MC algorithm, with parameters W_ε = 0.5 and $M_\varepsilon = N_\varepsilon$ = 5 whereas panel D reports the unbiased FES from the same data. Compared to the MD and conventional MC simulations, SA-MC leads to a much improved sampling of the valley around Φ = 75°, and more specifically the two saddle points connecting the left and right parts of the FES. Such transitions are typically rare events in Metropolis MC but rather well sampled within SA-MC. Differences between the biased and unbiased SA-MC FESs are minor. On the biased FES (Figure 7 C), SA-MC lowers barriers by ≈1.2 kcal/mol, which means that the corresponding states are better sampled.

Numerical values for the relative free energy values of the minima and at the top of the barriers are summarized in Table 3. First, it is observed that the current MD and MC simulations

Table 3. Comparison of ΔF from (a) MD, (b) Unbiased Targeted MD Simulations,[64] (c) MC, and (d) SA-MC[a]

	(a) MD				(b) ref 64			
	C_{7eq}	α_R	C_{7ax}	α_L	C_{7eq}	α_R	C_{7ax}	α_L
C_{7eq}	0.0	3.25			0.0	2.61		6.47
α_R	3.25	1.10			2.61	0.71	6.88	
C_{7ax}			3.26			6.88	4.34	5.98
α_L				4.62	6.47		5.98	4.35
	(c) MC				(d) SA-MC			
	C_{7eq}	α_R	C_{7ax}	α_L	C_{7eq}	α_R	C_{7ax}	α_L
C_{7eq}	0.0	3.37			0.0	1.96		5.08
α_R	3.37	1.16			1.96	0.21	4.91	
C_{7ax}			2.91	5.66		4.91	3.11	4.20
α_L			5.66	4.86	5.08		4.20	4.12

[a]Diagonal entries from Table 2 are stabilization energies relative to the global minimum C_{7eq}, whereas off-diagonal entries refer to the barriers between the minima. Empty cells indicate that the direct transition was not observed or is not possible. All free energies are reported relative to the C_{7eq} structure, which is the reference energy.

only partially sample the FES compared to previous targeted MD simulations[64] and the SA-MC simulations. On the other hand, SA-MC and the reference simulations[64] sample similar amounts of the available configuration space. In general, the location of the minima and their energy is similar to that found from previous work.[63,65,67,68] The C_{7eq} minimum is the most stable state on the FES for all types of simulations, followed by α_R. Its relative stabilization energy compared to the global minimum is ≈0.2 kcal/mol from unbiased SA-MC, which compares with 0.7 kcal/mol[64] and above 1 kcal/mol from MD and MC simulations.

The relative stability of the C_{7ax} and α_L structures from unbiased SA-MC are close to the MD simulations and differ by about 1 kcal/mol from reference simulations in the literature.[64,67] This suggests, that SA-MC in the present case is a suitable method to locate stable and metastable states on the FES with high confidence but that the quality of the unbiasing depends somewhat on the state considered.

It is also interesting to consider the energy at the top of the barriers separating two stable conformations. This information is summarized in Table 3. In general, the unbiased SA-MC data follow those from previous simulations.[64] Typically, the

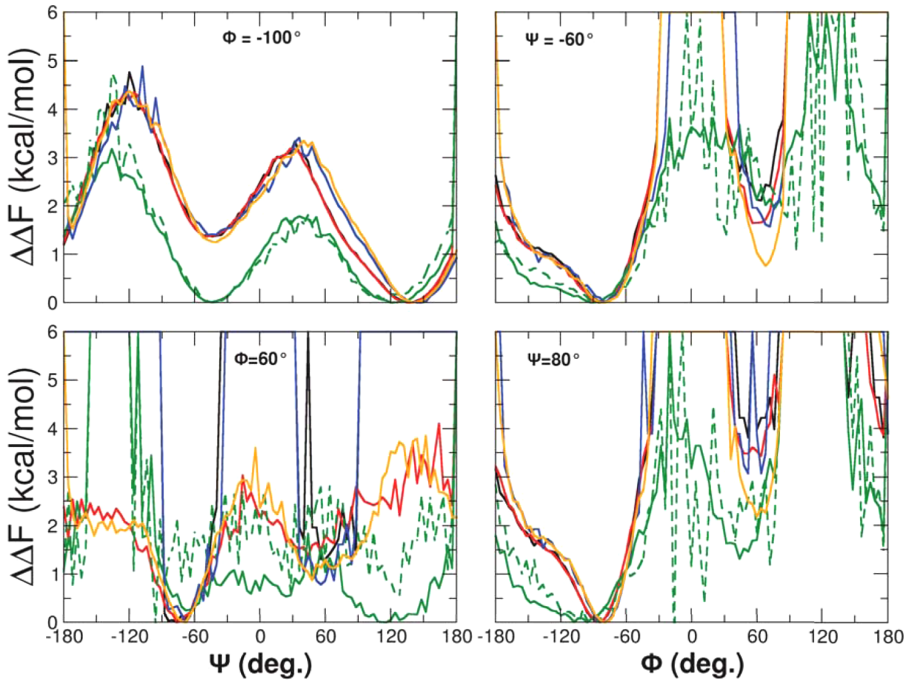


Figure 8. Slices through the FES from Figure 7 for MD (black), MC (red), SA-MC biased (dashed green), SA-MC unbiased (green), parallel tempering[17] (blue), and infinite swapping[17] (orange). SA-MC parameters are [0.5;5;5].

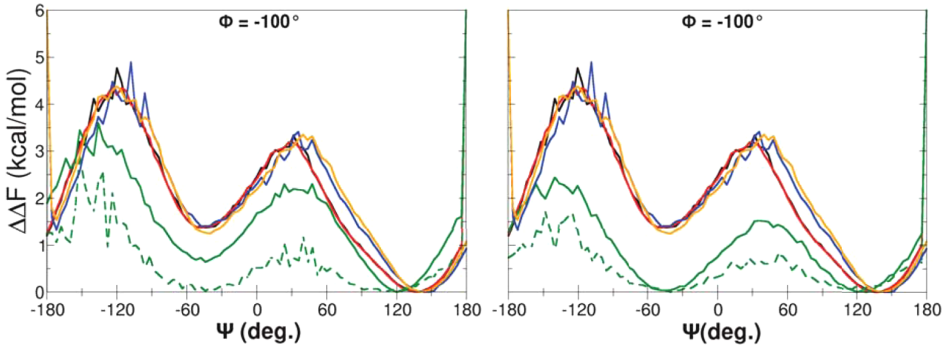


Figure 9. Slices through the FES, showing the influence of SA-MC parameters $[W_\varepsilon;M_\varepsilon;N_\varepsilon]$ on the biased and unbiased SA-MC energy profiles. SA-MC parameters are [0.1;10;10] (left) and [0.1;10;15] (right). MD (black), MC (red), SA-MC biased (dashed green), SA-MC unbiased (green), parallel tempering[17] (blue), and infinite swapping[17] (orange) are shown as separate curves. Free energy is shifted in order to have a value of 0.0 for the C_{7eq} minimum.

transition barriers are lower by about 1 kcal/mol but all orderings of the barriers agree with the data from the literature.

Figure 8 shows slices through the FESs from Figure 7 together with cuts from parallel tempering and infinite swapping simulations using the same setup of the systems,[17] for $\Phi = -100°$ and $\Phi = -60°$, (left panels) and $\Psi = -60°$ and $\Psi = 80°$ (right panels). For the cut at $\phi = -100°$, the topography of the FES from SA-MC is similar to all four other methods although quantitatively differences can be up to 1 kcal/mol for barriers and more for the secondary minimum. For the other three cuts, it is noted that there are much fewer unsampled regions (spikes) when using SA-MC than compared to any other method. Again, the unbiased SA-MC results underestimate the barriers and overstabilize the metastable states. However, from a sampling perspective SA-MC is clearly superior to MC: with 20 times fewer steps (5×10^6 for SA-MC against 100×10^6), for a similar CPU time usage, and transition regions are considerably more sampled with SA-MC than for MD, MC, and PT simulations.

Figure 9 shows slices through the same FES as in Figure 8, that is, for $\phi = -100°$, but reports results from simulations with different sets of SA-MC parameters: [0.1;10;10] (left) and [0.1;10;15] (right). It is apparent that the choice of SA-MC parameters influences the results. The data reported in Figure 9a better reproduces the reference simulations than the data in Figure 9b.

Simulations in Explicit Solvent. Sampling the free energy landscape of blocked alanine dipeptide in explicit water is computationally much more challenging.[83−86] The present system consists of 462 water molecules to which SHAKE constraints[87] are applied and one blocked $(Ala)_2$. The nonbonded cutoff parameter is 12 Å. Figure 10 reports the 2-dimensional FES obtained from 10^8 steps of SA-MC simulations with parameters [0.1;5;5] and Figure 11 reports

dx.doi.org/10.1021/ct500529w | J. Chem. Theory Comput. 2014, 10, 4284−4296

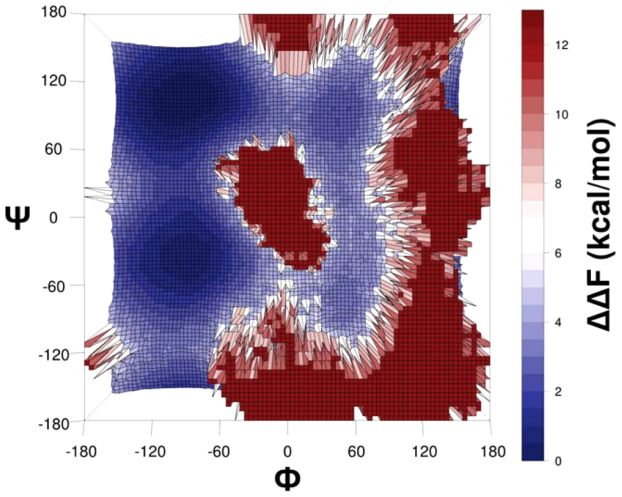


Figure 10. Unbiased FES for SA-MC from simulations of blocked alanine-dipeptide in explicit TIP3P water and with SA-MC parameters [0.1;5;5]. The number of steps is 100×10^6 steps. All free energies are reported relative to the C_{7eq} minimum.

slices (for $\Phi = -100°$ and $\Phi = -60°$, (left panels) and $\Psi = -60°$ and $\Psi = 80°$ (right panels)) through the FES of Figure 10 for SA-MC. The overall topology of the FES in implicit and explicit solvent is similar. However, it is noted that the transition between C_{7eq} and α_R is considerably wider in explicit water. Table 4 shows the energy estimated for the four known minima of the FES of Figure 10. Data from simulations with ACE are also included for comparison. The reference data is from refs 64 and 88, which was determined from MD simulations with both explicit solvent and a Generalized Born implicit solvent, and error bars represent a 95% confidence interval.

Table 4. Free Energies (in kcal/mol) Relative to the C_{7eq} Minimum from SA-MC (ACE), SA-MC (TIP3P) Simulations Compared to Reference Data from the Literature, Which Employed Umbrella Sampling in Explicit Solvent[64,88a]

	methods and references				
basin	SA-MC (ACE)	SA-MC (TIP3P)	ref 88	ref 64	position (Φ, Ψ)
C_{7eq}	0.00 ± 0.07	0.00 ± 0.08	0.00	0.0	(−83°, 136°)
α_R	0.21 ± 0.06	1.20 ± 0.07	1.30	1.41	(−79°, −42°)
C_{7ax}	3.11 ± 0.08	2.99 ± 0.08	NA	3.85	(−67°, −75°)
α_L	4.12 ± 0.09	3.60 ± 0.10	3.80	4.38	(47°, 55°)

[a]For C_{7ax} no data are available from ref 88. The statistical error in the present work was estimated from bootstrapping and ± values represent a 95% confidence interval.

It is found that the C_{7eq} minimum is still the most stable state, followed by α_R, C_{7ax}, and α_L. However, the relative stabilizations are somewhat altered in that α_R is destabilized relative to C_{7eq} whereas C_{7ax} and α_L are somewhat stabilized. Comparison with literature data shows that for the simulations in explicit solvent the present results agree favorably for C_{7eq}, α_R, α_L, and for C_{7ax} when available (see ref 64). It should be noted that no numerical values are provided in ref 88, and the numbers reported here have been inferred from the graphical illustrations (not possible for C_{7ax}). Comparison with ref 64, which also provided values for simulations with the ACE model, shows good agreement with results obtained with SA-MC; however, once again, it appears that the α_R minimum is somewhat overstabilized when using SA-MC.

4. CONCLUSIONS AND OUTLOOK

In the present work, a practical and comprehensive implementation for spatial averaging MC (SA-MC) simulations into the CHARMM general purpose atomistic simulation

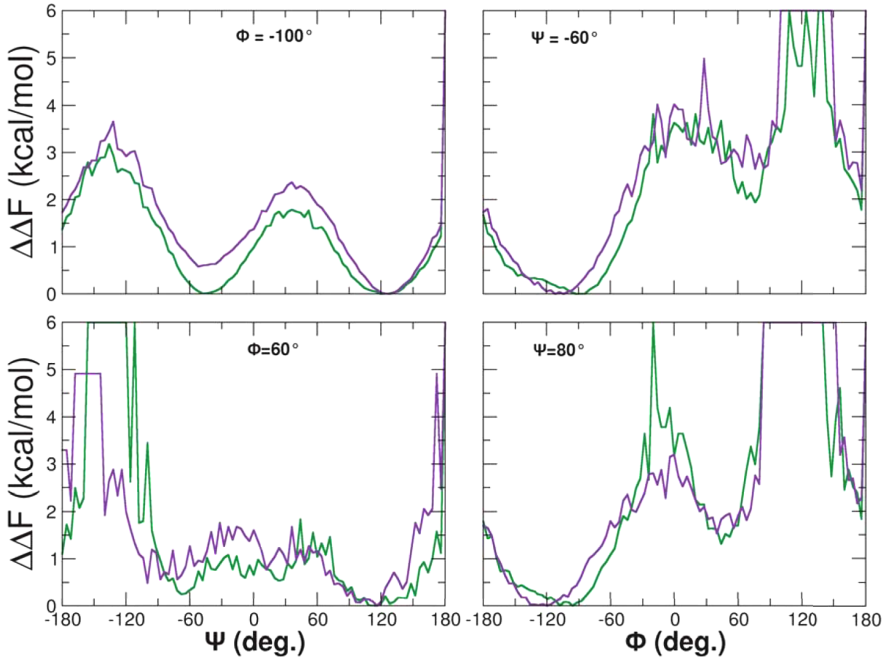


Figure 11. Slices through the FES from Figures 7 and 10: unbiased SA-MC with ACE (green), unbiased SA-MC with TIP3P (violet). All values are reported relative to the C_{7eq} minimum.

program has been described. Also, an unbiasing procedure is discussed which allows to estimate thermodynamic observables. The implementation and unbiasing strategy are validated for model and topical systems including the double well potential, Lennard-Jones clusters and the blocked alanine dipeptide in implicit and explicit solvent. The considerably increased efficiency for exploring configuration space has been demonstrated for all three applications. However, the degree to which this is possible depends on the properties and connectivity of the systems' conformational space, which is usually a priori unknown. The central asset of SA-MC is that it generates a more highly connected ensemble, which makes exploration of the underlying free energy surface more readily possible.

It is expected that SA-MC can be beneficial for a range of future applications. As already indicated, SA-MC can be used to efficiently explore configurational space, based on which unbiased free energy surfaces can be obtained from the spatially averaged distribution. Furthermore, SA-MC is well suited to approximately locate transition states and to characterize the transition state ensemble.[89,90] This is the starting point for enhanced exploration of barrier-crossing problems in more complex systems (such as small solvated peptides or proteins), which is typically difficult to achieve from standard MC or MD simulations. Given that SA-MC primarily connects neighboring metastable states, which are usually separated by barriers of a few k_BT, we expect SA-MC to perform well for such problems as was already demonstrated for the solvated dipeptide in the present work. Also, SA-MC can be employed to find approximate reaction coordinates, which is useful for subsequent umbrella sampling simulations.[7] Finally, SA-MC could be employed together with Hamiltonian replica exchange molecular dynamics simulations (H-REMD).[91] Hamiltonian replica exchange can be used for studying several types of problems, but in practice, its performance depends substantially on the details of the biased Hamiltonian. Similar to combining umbrella sampling simulations with H-REMD,[92] employing SA-MC together with H-REMD could be potentially beneficial and provide a systematic way to generate biased Hamiltonians.

A common characteristic of all MC methods is that simulation parameters such as the move range, the acceptance ratio, or the swapping rate need to be optimized to some extent to obtain computational performance. This is also the case for SA-MC. One future improvement for the SA-MC algorithm is therefore to facilitate finding optimized sets of parameters $[W_\varepsilon;M_\varepsilon;N_\varepsilon]$ during the simulation. It is not necessary to use the same values for each of the MC steps because of the Markovianity of the procedure. The examples investigated here in more detail emphasize that larger values of the system parameters enhance the sampling of barriers and transition states at the cost of extra computational time. Hence, another possible improvement concerns the decrease of those parameters for regions well sampled by the MC algorithm for speeding up the sampling, and to increase them for poorly sampled regions, possibly during the simulation by using an "on-the-fly" optimization technique. This will be important for applying efficiently the SA-MC algorithm to larger systems.

AUTHOR INFORMATION

Corresponding Author

*Email: m.meuwly@unibas.ch.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation through grants 200021-117810 and the NCCR MUST (to M.M.). J.D.D. acknowledges support from the National Science Foundation award DMS-1317199.

REFERENCES

(1) Metropolis, N.; Ulam, S. The Monte Carlo Method. *J. Am. Stat. Assoc.* **1949**, *44*, 335−341.
(2) Kroese, D.; Taimre, T.; Botev, Z. *Handbook of Monte Carlo Methods*; John Wiley & Sons, Inc.: Hoboken, NJ, 2011.
(3) Rubino, G.; Tuffin, B. *Rare Event Simulation Using Monte Carlo Methods*; John Wiley & Sons, Inc.: Hoboken, NJ, 2009.
(4) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087.
(5) Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **1970**, *57*, 97−109.
(6) Earl, D. J.; Deem, M. W. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910.
(7) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Chem. Phys.* **1977**, *23*, 187−199.
(8) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562−12566.
(9) Hukushima, K.; Nemoto, K. Exchange Monte Carlo Method and application to spin glass simulations. *J. Phys. Soc. Japan* **1996**, *65*, 1604−1608.
(10) Voter, A. F. A Monte Carlo method for determining free-energy differences and transition state theory rate constants. *J. Chem. Phys.* **1985**, *82*, 1890.
(11) Betancourt, M. R. Optimization of Monte Carlo trial moves for protein simulations. *J. Chem. Phys.* **2011**, *134*, 14104.
(12) Swendsen, R. H.; Wang, J.-S. Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.* **1986**, *57*, 2607.
(13) Plattner, N.; Doll, J. D.; Dupuis, P.; Wang, H.; Liu, Y.; Gubernatis, J. E. An infinite swapping approach to the rare-event sampling problem. *J. Chem. Phys.* **2011**, *135*, 134111.
(14) Dupuis, P.; Liu, Y.; Plattner, N.; Doll, J. D. On the infinite swapping limit for parallel tempering. *Multiscale Model. Simul.* **2012**, *10*, 986−1022.
(15) Doll, J. D.; Plattner, N.; Freeman, D. L.; Liu, Y.; Dupuis, P. Rare-event sampling: Occupation-based performance measures for parallel tempering and infinite swapping Monte Carlo methods. *J. Chem. Phys.* **2012**, *137*, 204112.
(16) Lu, J.; Vanden-Eijnden, E. Infinite swapping replica exchange molecular dynamics leads to a simple simulation patch using mixture potentials. *J. Chem. Phys.* **2013**, *138*, 84105.
(17) Plattner, N.; Doll, J. D.; Meuwly, M. Overcoming the rare event sampling problem in biological systems with infinite swapping. *J. Chem. Theory Comput.* **2013**, *9*, 4215−4224.
(18) Kofke, D. A. On the acceptance probability of replica-exchange Monte Carlo trials. *J. Chem. Phys.* **2002**, *117*, 6911.
(19) Predescu, C.; Predescu, M.; Ciobanu, C. V. The incomplete beta function law for parallel tempering sampling of classical canonical systems. *J. Chem. Phys.* **2004**, *120*, 4119−4128.
(20) Katzgraber, H. G.; Trebst, S.; Huse, D. A.; Troyer, M. Feedback-optimized parallel tempering Monte Carlo. *J. Stat. Mech.: Theory Exp.* **2006**, *2006*, 3018.
(21) Sabo, D.; Meuwly, M.; Freeman, D. L.; Doll, J. D. A constant entropy increase model for the selection of parallel tempering ensembles. *J. Chem. Phys.* **2008**, *128*, 174109.
(22) Grubmüller, H. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Phys. Rev. E* **1995**, *52*, 2893−2906.
(23) Müller, E. M.; de Meijere, A.; Grubmüller, H. Predicting unimolecular chemical reactions: Chemical flooding. *J. Chem. Phys.* **2002**, *116*, 897.

(24) Tsallis, C. Possible generalization of Boltzmann−Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479−487.
(25) Fukuda, I.; Nakamura, H. Deterministic generation of the Boltzmann−Gibbs distribution and the free energy calculation from the Tsallis distribution. *Chem. Phys. Lett.* **2003**, *382*, 367−373.
(26) Kim, J. G.; Fukunishi, Y.; Nakamura, H. Dynamical origin of enhanced conformational searches of Tsallis statistics sampling. *J. Chem. Phys.* **2004**, *121*, 1626−1635.
(27) Li, Z.; Scheraga, H. A. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 6611−6615.
(28) Piela, L.; Kostrowicki, J.; Scheraga, H. A. On the multiple-minima problem in the conformational analysis of molecules: Deformation of the potential energy hypersurface by the diffusion equation method. *J. Phys. Chem.* **1989**, *93*, 3339−3346.
(29) Ma, J.; Straub, J. E. Simulated annealing using the classical density distribution. *J. Chem. Phys.* **1994**, *101*, 533.
(30) Pappu, R. V.; Hart, R. K.; Ponder, J. W. Analysis and application of potential energy smoothing and search methods for global optimization. *J. Phys. Chem. B* **1998**, *102*, 9725−9742.
(31) Doll, J. D.; Gubernatis, J. E.; Plattner, N.; Meuwly, M.; Dupuis, P.; Wang, H. A spatial averaging approach to rare-event sampling. *J. Chem. Phys.* **2009**, *131*.
(32) Plattner, N.; Doll, J. D.; Meuwly, M. Spatial averaging for small molecule diffusion in condensed phase environments. *J. Chem. Phys.* **2010**, *133*.
(33) Brooks, C. L.; Mackerell, A. D.; Karplus, M. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545−1614.
(34) Wales, D. J. *Energy Landscapes: Applications to Clusters, Biomolecules, and Glasses*; Cambridge University Press: Cambridge, U.K., 2003.
(35) Eleftheriou, M.; Kim, D.; Doll, J. D.; Freeman, D. L. Information theory and the optimization of Monte Carlo simulations. *Chem. Phys. Lett.* **1997**, *276*, 353−360.
(36) Efron, B. Bootstrap Methods—Another look at the jackknife. *Ann. Stat.* **1979**, *7*, 1−26.
(37) Nangia, S.; Jasper, A. W.; Miller, T. F., III; Truhlar, D. G. Army ants algorithm for rare event sampling of delocalized nonadiabatic transitions by trajectory surface hopping and the estimation of sampling errors by the bootstrap method. *J. Chem. Phys.* **2004**, *120*, 3586−3597.
(38) Nutt, D.; Meuwly, M. Studying reactive processes with classical dynamics: Rebinding dynamics in MbNO. *Biophys. J.* **2006**, *90*, 1191−1201.
(39) Hu, J.; Ma, A.; Dinner, A. R. Monte Carlo simulations of biomolecules: The MC module in CHARMM. *J. Comput. Chem.* **2006**, *27*, 203−216.
(40) Schaefer, M.; Karplus, M. A comprehensive analytical treatment of continuum electrostatics. *J. Phys. Chem.* **1996**, *100*, 1578−1599.
(41) Schaefer, M.; Bartels, C.; Karplus, M. Solution conformations and thermodynamics of structured peptides: Molecular dynamics simulation with an implicit solvation model. *J. Mol. Biol.* **1998**, *284*, 835−848.
(42) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(43) Ceperley, D. M.; Dewing, M. The penalty method for random walks with uncertain energies. *J. Chem. Phys.* **1999**, *110*, 9812−9820.
(44) Goedecker, S. Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.* **2004**, *120*, 9911−9917.
(45) Jones, J. E. On the determination of molecular fields. II. From the equation of state of a gas. *Proc. R. Soc. London* **1924**, *106*, 463−477.
(46) Wales, D. J.; Doye, J. P. K.; Dullweber, A.; Hodges, M. P.; Naumkin, F. Y.; Calvo, F.; Hernández-Rojas, J.; Middleton, T. F. The Cambridge Cluster Database: http://www-wales.ch.cam.ac.uk/CCD.html (accessed Aug. 13, 2014).
(47) Wales, D. J.; Doye, J. P. K. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Phys. Chem. A* **1997**, *101*, 5111−5116.
(48) Schelstraete, S.; Verschelde, H. Finding minimum-energy configurations of Lennard-Jones clusters using an effective potential. *J. Phys. Chem. A* **1997**, *101*, 310−315.
(49) Doye, J. P. K.; Miller, M.; Wales, D. The double-funnel energy landscape of the 38-atom Lennard-Jones cluster. *J. Chem. Phys.* **1999**, *110*, 6896−6906.
(50) Doye, J. P. K.; Miller, M.; Wales, D. Evolution of the potential energy surface with size for Lennard-Jones clusters. *J. Chem. Phys.* **1999**, *111*, 8417−8428.
(51) Xiang, Y.; Cheng, L.; Cai, W.; Shao, X. Structural distribution of Lennard-Jones clusters containing 562 to 1000 atoms. *J. Phys. Chem. A* **2004**, *108*, 9516−9520.
(52) Calvo, F.; Doye, J. P. K.; Wales, D. J. Quantum partition functions from classical distributions: Application to rare-gas clusters. *J. Chem. Phys.* **2001**, *114*, 7312−7329.
(53) Honeycutt, J. D.; Andersen, H. C. Molecular dynamics study of melting and freezing of small Lennard-Jones clusters. *J. Phys. Chem.* **1987**, *91*, 4950−4963.
(54) Neirotti, J. P.; Calvo, F.; Freeman, D. L.; Doll, J. D. Phase changes in 38-atom Lennard-Jones clusters. I. A parallel tempering study in the canonical ensemble. *J. Chem. Phys.* **2000**, *112*, 10340−10349.
(55) Calvo, F.; Neirotti, J. P.; Freeman, D. L.; Doll, J. D. Phase changes in 38-atom Lennard-Jones clusters. II. A parallel tempering study of equilibrium and dynamic properties in the molecular dynamics and microcanonical ensembles. *J. Chem. Phys.* **2000**, *112*, 10350−10357.
(56) Sharapov, V.; Mandelshtam, V. Solid−solid structural transformations in Lennard-Jones clusters: Accurate simulations versus the harmonic superposition approximation. *J. Phys. Chem. A* **2007**, *111*, 10284−10291.
(57) Sharapov, V.; Meluzzi, D.; Mandelshtam, V. Low-temperature structural transitions: Circumventing the broken-ergodicity problem. *Phys. Rev. Lett.* **2007**, *98*, 105701.
(58) Wales, D. J. Discrete path sampling. *Mol. Phys.* **2002**, *100*, 3285−3305.
(59) Wales, D. J. Some further applications of discrete path sampling to cluster isomerization. *Mol. Phys.* **2004**, *102*, 891−908.
(60) Wales, D. J. Surveying a complex potential energy landscape: Overcoming broken ergodicity using basin-sampling. *Chem. Phys. Lett.* **2013**, *584*, 1−9.
(61) Adjanor, G.; Athènes, M.; Calvo, F. Free energy landscape from path-sampling: Application to the structural transition in LJ38. *Eur. Phys. J. B* **2006**, *53*, 47−60.
(62) Oakley, M. T.; Johnston, R. L.; Wales, D. J. Symmetrization schemes for global optimization of atomic clusters. *Phys. Chem. Chem. Phys.* **2013**, *15*, 3965−3976.
(63) Tobias, D. J.; Brooks, C. L. Conformational equilibrium in the alanine dipeptide in the gas phase and aqueous solution: A comparison of theoretical results. *J. Phys. Chem.* **1992**, *96*, 3864−3870.
(64) Apostolakis, J.; Ferrara, P.; Caflisch, A. Calculation of conformational transitions and barriers in solvated systems: Application to the alanine dipeptide in water. *J. Chem. Phys.* **1999**, *110*, 2099.
(65) Chekmarev, D. S.; Ishida, T.; Levy, R. M. Long-time conformational transitions of alanine dipeptide in aqueous solution: Continuous and discrete-state kinetic models. *J. Phys. Chem. B* **2004**, *108*, 19487−19495.
(66) Ma, A.; Dinner, A. R. Automatic method for identifying reaction coordinates in complex systems. *J. Phys. Chem. B* **2005**, *109*, 6769−6779.
(67) Gfeller, D.; De Los Rios, P.; Caflisch, A.; Rao, F. Complex network analysis of free-energy landscapes. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1817−1822.
(68) Yang, L.; Yi, Q. G. A selective integrated tempering method. *J. Chem. Phys.* **2009**, *131*, 214109.

(69) Multiple state transition interface sampling of alanine dipeptide in explicit solvent. *J. Chem. Phys.* **2011**, *135*, 145102.
(70) García-Prieto, F. F.; Fdez Galván, I.; Aguilar, M. A.; Martn, M. E. Study on the conformational equilibrium of the alanine dipeptide in water solution by using the averaged solvent electrostatic potential from molecular dynamics methodology. *J. Chem. Phys.* **2011**, *135*, 194502.
(71) Lee, I. H. Free-energy profile along an isomerization pathway: Conformational isomerization in alanine dipeptide. *J. Korean Phys. Soc.* **2013**, *62*, 384−392.
(72) Morishita, T.; Itoh, S. G.; Okumura, H.; Mikami, M. On-the-fly reconstruction of free-energy profiles using logarithmic mean-force dynamics. *J. Comput. Chem.* **2013**, *34*, 1375−1384.
(73) Kondo, H. X.; Taiji, M. Enhanced exchange algorithm without detailed balance condition for replica exchange method. *J. Chem. Phys.* **2013**, *138*, 244113.
(74) Lankau, T.; Yu, C.-H. A constrained reduced-dimensionality search algorithm to follow chemical reactions on potential energy surfaces. *J. Chem. Phys.* **2013**, *138*, 214102.
(75) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From A to B in free energy space. *J. Chem. Phys.* **2007**, *126*, 54103.
(76) Barducci, A.; Bussi, G.; Parrinello, M. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **2008**, *100*, 20603.
(77) Bonomi, M.; Barducci, A.; Parrinello, M. Reconstructing the equilibrium Boltzmann distribution from well-tempered metadynamics. *J. Comput. Chem.* **2009**, *30*, 1615−1621.
(78) Branduardi, D.; Bussi, G.; Parrinello, M. Metadynamics with adaptive Gaussians. *J. Chem. Theory Comput.* **2012**, *8*, 2247−2254.
(79) Strodel, B.; Wales, D. J. Free energy surfaces from an extended harmonic superposition approach and kinetics for alanine dipeptide. *Chem. Phys. Lett.* **2008**, *466*, 105−115.
(80) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **1963**, *7*, 95−99.
(81) Schaefer, M.; Bartels, C.; Leclerc, F.; Karplus, M. Effective atom volumes for implicit solvent models: Comparison between Voronoi volumes and minimum fluctuation volumes. *J. Comput. Chem.* **2001**, *22*, 1857−1879.
(82) Calimet, N.; Schaefer, M.; Simonson, T. Protein molecular dynamics with the generalized Born/ACE solvent model. *Proteins* **2001**, *45*, 144−158.
(83) Henin, J.; Fiorin, G.; Chipot, C.; Klein, M. L. Exploring multidimensional free energy landscapes using time-dependent biases on collective variables. *J. Chem. Theory Comput.* **2010**, *6*, 35−47.
(84) Sutto, L.; DÁbramo, M.; Gervasio, F. L. Comparing the efficiency of biased and unbiased molecular dynamics in reconstructing the free energy landscape of Met-Enkephalin. *J. Chem. Theory Comput.* **2010**, *6*, 3640−3646.
(85) Zhou, T.; Caflisch, A. Free energy guided sampling. *J. Chem. Theory Comput.* **2012**, *8*, 2134−2140.
(86) Wojtas-Niziurski, W.; Meng, Y.; Roux, B.; Bernèche, S. Self-learning adaptive umbrella sampling method for the determination of free energy landscapes in multiple dimensions. *J. Chem. Theory Comput.* **2013**, *9*, 1885−1895.
(87) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of *n*-alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.
(88) Scarsi, M.; Apostolakis, J.; Caflisch, A. Comparison of a GB solvation model with explicit solvent simulations: Potentials of mean force and conformational preferences of alanine dipeptide and 1,2-dichloroethane. *J. Phys. Chem. B* **1998**, *102*, 3637−3641.
(89) *Advances in Chemical Physics*; Prigogine, I., Rice, S. A., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, 2002; Vol. *123*.
(90) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291−318.
(91) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.* **2002**, *116*, 9058−9067.
(92) Jiang, W.; Luo, Y.; Maragliano, L.; Roux, B. Calculation of free energy landscape in multi-dimensions with hamiltonian-exchange umbrella sampling on petascale supercomputer. *J. Chem. Theory Comput.* **2012**, *8*, 4672−4680.

Chapter 4

Validations, applications and results for PINS

"*You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage.*"

John von Neumann, Suggesting to Claude Shannon a name for his new uncertainty function.

In this Chapter the results obtained during the PhD with PINS are presented. For a mathematical and algorithmic background, one can refer to Chapter 2 Section 2.5 where the methodology was properly introduced.

A previous CHARMM implementation was already described and validated[110], using the available ENSEMBLE module of CHARMM. However, the ENSEMBLE code underwent profound modifications in its architecture for improved compatibility with modern Fortran standards and use of novel computational methods. Hence, it was necessary to reimplement PINS with a dual-chain approach into the most recent CHARMM c41 release.

An article was written for presenting the new implementation, and validating it. It is currently ready to be published and should be resubmitted soon in *Journal of Chemical Theory and Computation* (after a first submission, reviewers pointed the necessity of several required revisions before acceptance). It was co-written with Nuria Plattner, Jimmie. D. Doll and Markus Meuwly. PINS implementation and algorithmic details a first given (less detailed version of Section 2.5), then the validation is illustrated with: (i) study of the conformational equilibria of the alanine decapeptide (article *to* Section 3.), and (ii) a study of the Xenon atoms migration in the Myoglobin protein (article → Section 4.)

The manuscript of this article is appended to the current Chapter and can be found in Section 4.3, together with some supplementary material.

The PINS code is written as a Fortran sub-module of CHARMM's ENSEMBLE module. Compiling and use information is available from the official CHARMM documentation, within the source archive (for version numbers >= c41a1).

In the following, two supplementary sections not included in the current article manuscript are presented.

Section 4.1 validates the PINS implementation by investigating the FES of Alanine Dipeptide, in a similar way than the SA-MC validation from Chapter 3. It was originally part of the manuscript appended in Section 4.3 however it was decided to focus on more challenging validation systems in order to publish the article in JCTC. Nevertheless Section 4.1 can probably, after a few modifications, be published independently as a letter or communication.

Section 4.2 presents an interesting set of performance measures for evaluating the sampling boost obtain by PINS over PT, for the case of deca-alanine. This was initially performed as a reply to a reviewer, and part of it may be inserted in the manuscript before resubmission.

4.1 Validation for Alanine Dipeptide

The blocked alanine dipeptide (Ac-Ala-N-H-Me, Figure 4.1) has been used as a validation system for computational studies of conformational equilibria, and free energy landscape reconstruction and analysis. [47–49, 115–128] The dipeptide contains several notable structural features, including the two (ϕ, ψ) dihedral angles, NH- and CO-groups capable of H-bond formation, and a methyl group attached to the C_α atom. Successful computational studies used quantum chemistry, MD and MC simulations, and several conformations were identified [115, 116, 119, 120, 122, 128]: (i) β, also called C_5, for $(\phi, \psi) \sim (-140°, 150°)$, (ii) $C_{7\mathrm{eq}}$ for $(\phi, \psi) \sim (-90°, 80°)$, (iii) α_R (Right-handed α helix) for $(\phi, \psi) \sim (-80°, -60°)$, (iv) α_L (Left-handed α helix) for $(\phi, \psi) \sim (60°, 60°)$ and (v) $C_{7\mathrm{ax}}$ for $(\phi, \psi) \sim (60°, -60°)$. One suitable way to visualize the free energy landscape for the conformations and the transitions between them is to report an energy surface as a Ramachandran plot.[129] Simulations were carried out using two implicit solvent models available within CHARMM: ACE (Analytical Continuum Electrostatics[130, 131]) and GENBORN (GENeralized BORN Model).[132] Both are known for providing a meaningful description of solvation effects for peptides.[59, 119, 133, 134] and[132, 135–137]

Validation: In the following, results obtained with PT and PINS are systematically compared to each other and with previously published results from MC, MD and SA-MC simulations (see SI Section 1 for more details concerning SA-MC). The particular points of interest are : (i) whether the methods

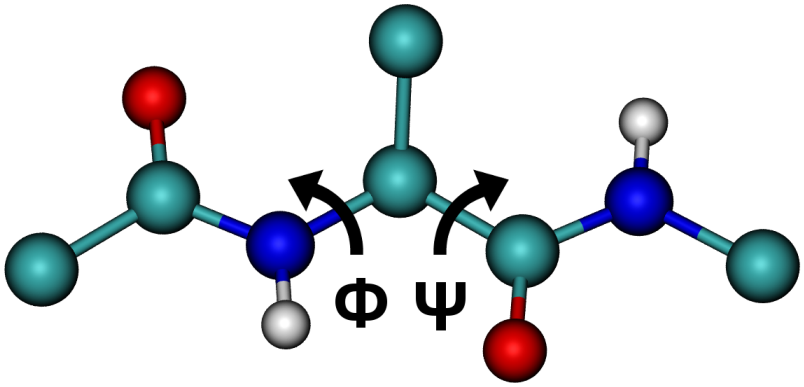


FIGURE 4.1: The blocked alanine dipeptide (Ac-Ala-N-H-Me), and the 2 dihedral angles of interest Φ (C-N-C_α-C) and Ψ (N-C_α-C-N)

adequately sample the 4 minima, and (ii) whether the Helmholtz free energy ΔF (simulations performed in the Canonical NVT ensemble) at each minimum (relative stabilization) and at the top of the transition barriers (transition state energies) agree with previous work.

The PT and PINS simulations were carried out as follows: the ENSEMBLE module from CHARMM c40 was used, modified as described above for dual-chain PINS. Twelve replicas were used, running at temperatures ranging from 300 K to 1395 K with a constant ratio of $\frac{T_{i+1}}{T_i} = 1.15$, for PT and PINS methods. For PINS the 12 temperatures are divided in two INS chains using a 3-block structure $(3, 6, 3|4, 4, 4)$.

The structures were first equilibrated using PT, at the corresponding T, for 10^4 steps using a timestep of 1 fs and no exchange was performed during the equilibration phase. Next, each replica is propagated for $5 \cdot 10^7$ steps using a timestep of 1 fs, and exchanges between adjacent replicas (for PT) or between symmetrized blocks (for PINS) are attempted every 500 steps. The total aggregated simulation time is then 600 ns for both PT and PINS.

Results obtained with PT and PINS are first compared, using both the ACE and GENBORN implicit solvent models. Figures 4.2A and C show FESs obtained with standard PT whereas Figures 4.2B and D were obtained with PINS. The methodology for building the surfaces was the following: (i) for each configuration the two Φ and Ψ dihedral angles were extracted (see Figure 4.1), (ii) a two dimensional, normalised histogram was built on a grid of 90 by 90 points, from $-180°$ to $180°$, and (iii) the free energy is approximated using the normalised probability density $\rho(\Phi_i, \Psi_j)$ available for each grid point of the 2D histogram, $\Delta F(\Phi_i, \Psi_j) = -RT \ln(\rho(\Phi_i, \Psi_j))$.

A first visual analysis shows that PINS provides enhanced sampling around $\Phi = 60°$ which corresponds to the α_{L} and $C_{7\mathrm{ax}}$ minima (respectively top-right and bottom-right). The transition barriers $C_{7\mathrm{eq}} \leftrightarrow \alpha_{\mathrm{L}}$ and $\alpha_{\mathrm{R}} \leftrightarrow C_{7\mathrm{ax}}$, not well sampled with PT, are also sufficiently sampled with PINS.

Table 4.A reports the stabilization of the local minima relative to $C_{7\mathrm{eq}}$ on the diagonal, and the barrier heights between the local minima as the off-diagonal entries. The data is reported for simulations with PT and PINS using the ACE ([a] and [b]), or the GENBORN ([c] and [d]) implicit solvent models, see Figure 4.2. For a given solvent model the free energies for states sampled by both PT and PINS are similar, and also agree with previously published results.[112] This validates the implementation and post-processing of PINS. ACE and GENBORN being both Generalized Born Methods, close free energy values are to be expected. Table 4.A confirms this: free energy differences between minima and barriers are within 0.5 kcal/mol, for either PT or PINS.

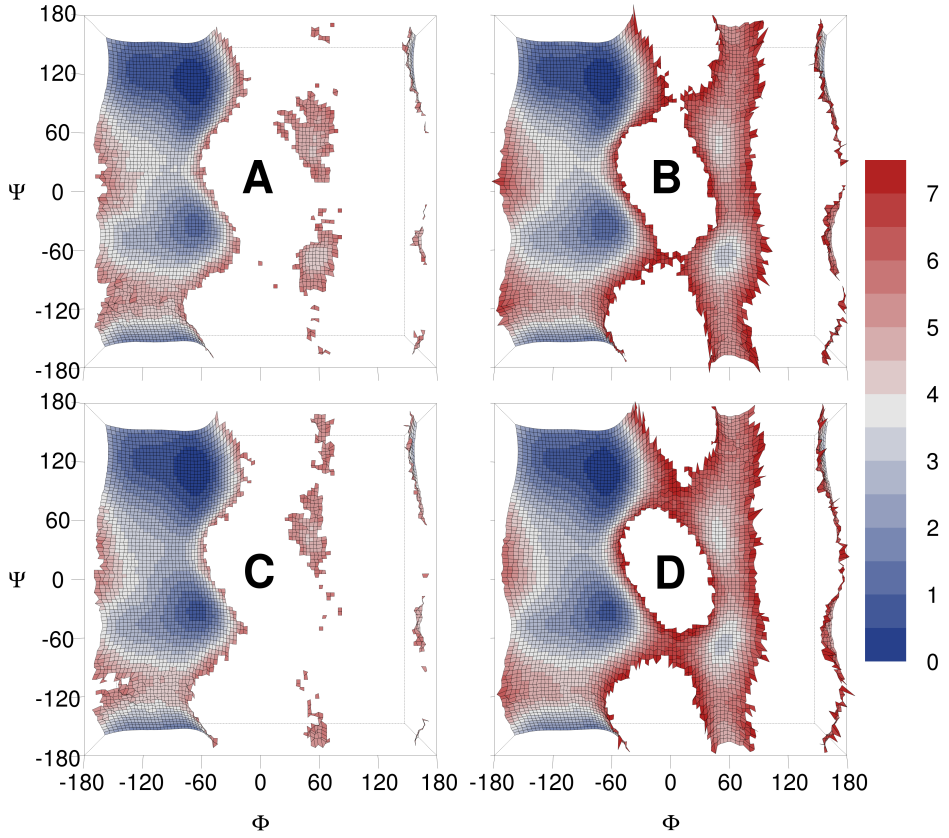


FIGURE 4.2: 2D free energy surfaces (ΔF in kcal/mol) obtained with PT (Left) or PINS (Right), using either the ACE (panels A and B) or GENBORN (panels C and D) models. Only results for the replica at 300 K are presented. Twelve replicas running for 50 ns each were used, so the aggregated simulation time is 600 ns. PINS simulations show a systematic higher sampling efficiency for the right valley where $\Phi = 60°$, and transitions between local minima (bridging saddle points for $\Phi \approx 0°$, see panels C and D) are sampled with PINS but not with PT.

	[a] PT (ACE)				[b] PINS (ACE)			
	C_{7eq}	α_R	C_{7ax}	α_L	C_{7eq}	α_R	C_{7ax}	α_L
C_{7eq}	0.00	3.50			0.00	3.55		7.0*
α_R	3.50	1.20			3.55	1.10	6.5*	
C_{7ax}			4.00			6.5*	3.20	5.50
α_L				4.50	7.0*		5.50	4.15
	[c] PT (GENBORN)				[d] PINS (GENBORN)			
	C_{7eq}	α_R	C_{7ax}	α_L	C_{7eq}	α_R	C_{7ax}	α_L
C_{7eq}	0.00	3.10			0.00	3.10		6.50
α_R	3.10	0.80			3.10	0.80	6.10	
C_{7ax}						6.10	3.25	5.25
α_L				4.2	6.50		5.25	4.30

TABLE 4.A: Comparison of ΔF (kcal/mol) from simulations using ACE ([a] PT or [b] PINS) or GENBORN ([c] PT and [d] PINS). Diagonal entries are stabilization energies relative to the global minimum C_{7eq}, whereas off-diagonal entries refer to the barriers between the minima. Empty cells indicate that the direct transition was not observed or is not possible. A star (*) indicates poor sampling of the corresponding minimum or transition (some of the neighbouring grid points were unsampled), and such values should be considered with care.

Table 4.B compares results obtained for PINS [b] (B in Figure 4.2), standard MD [a], MC [c], and SA-MC [d] with the ACE implicit solvent model.[112] The corresponding FES for [a] [c] [d] are reported

	[a] MD Ref. [112]				[b] PINS (ACE)			
	C_{7eq}	α_R	C_{7ax}	α_L	C_{7eq}	α_R	C_{7ax}	α_L
C_{7eq}	0.0	3.25			0.00	3.55		7.0*
α_R	3.25	1.10			3.55	1.10	6.5*	
C_{7ax}			3.26			6.5*	3.20	5.50
α_L				4.62	7.0*		5.50	4.15
	[c] MC Ref. [112]				[d] SA-MC Ref. [112]			
	C_{7eq}	α_R	C_{7ax}	α_L	C_{7eq}	α_R	C_{7ax}	α_L
C_{7eq}	0.0	3.37			0.0	1.96		5.08
α_R	3.37	1.16			1.96	0.21	4.91	
C_{7ax}			2.91	5.66		4.91	3.11	4.20
α_L			5.66	4.86	5.08		4.20	4.12

TABLE 4.B: Comparison of ΔF (kcal/mol) from [a] MD, [b] PINS, [c] MC, and [d] SA-MC. All simulations used the ACE implicit solvent model. Values for [a] [c] [d] are taken from Ref. [112]. Diagonal entries are stabilization energies relative to the global minimum C_{7eq}, whereas off-diagonal entries refer to the barriers between the minima. Empty cells indicate that the direct transition was not observed or is not possible. A star (*) indicates poor sampling of the corresponding minimum or transition, and such values should be considered with care.

in Figure 7 of Ref. [112]. Free energies from PINS are remarkably close to results from MD and MC simulations, validating the accuracy of PINS. When comparing to SA-MC, differences in free energies range from 0.5 to 2.0 kcal/mol. The fact that SA-MC yields slightly lower values compared to MC or MD was already discussed in Ref. [112] and is caused by the averaging process (see Eq. 3 from the SI). However, the relative stabilities of each of the minima and transitions is maintained. PINS produces values close to what is observed with SA-MC, especially for the transition barrier heights, which were poorly (or not at all) sampled using MC and MD methods.

This validation of PINS for the Alanine dipeptide establishes the efficiency of the algorithm for sampling rare-events for free energy simulations such as transitions between neighbouring minima. The improved sampling is evident when comparing to results from MD, MC, PT, and SA-MC simulations. Concerning the computational requirements afforded by PINS relative to PT, PINS was 9 % and 3 % slower for simulations with ACE and GENBORN, respectively. Simulations with GENBORN are usually 20 to 25 % slower than the corresponding ACE simulation (for MD, PT and PINS simulations). So when the energy evaluations are the most time consuming part of the simulations, as for GENBORN, the computational overhead of PINS versus PT appears to be only a few percent which is remarkable when considering the improved sampling for the same number of steps in the simulations compared to PT (see Figure 4.2).

The topology of the FES at high temperatures: While the previous section validated the accuracy of PINS by comparing to standard methods, and its sampling superiority over PT, it is also interesting to consider another possibility offered by this method: simultaneous sampling of several thermodynamic states (here distinguished by the temperature) at the same time.

Post-processing (Eqns. 7 and 8 from the article) allows to study any thermodynamic property of interest at all temperatures. Above, the FESs were reported for $T = 300$ K, and the influence of the 11 higher temperatures used in this example was only indirectly considered, through the sampling gain they offered. The same analysis can be carried out for any of the 11 other temperatures. In the following the investigation focuses on: (i) tracking the position of the four stable minima (by order of stability : C_{7eq} α_R C_{7ax} α_L), i.e. to check if the higher temperatures cause a noticeable displacement of those minima, and (ii) to assess the global influence of the temperature on the topology of the surface by considering values of ΔF for the 4 Minimum Energy Paths (MEPs) $C_{7eq} \leftrightarrow \alpha_R$, $\alpha_R \leftrightarrow C_{7ax}$, $C_{7ax} \leftrightarrow \alpha_L$ and $\alpha_L \leftrightarrow C_{7eq}$.

Previously published studies already investigated the location of the paths connecting minima of the alanine dipeptide. Apostolakis *et al.* [116] provided their location, and characterized some optimal free energy paths (OFEPs) connecting them. Jang *et al.* [138] defined a set of transition pathways, using the dynamic importance sampling (DIMS) method, between the same four minima previously introduced. They also counted the observed transitions, and assigned to each possible pathway a given probability. The methodology used for finding the MEPs (Dijkstra's algorithm, see SI Section 4) is related to the OFEPs from Ref.[116], as it finds the path between two points for which (i) the number of grid cells crossed by the path is minimal, and (ii) the change of free energy ($\Delta\Delta F$) when moving from one cell to another is as small as possible.

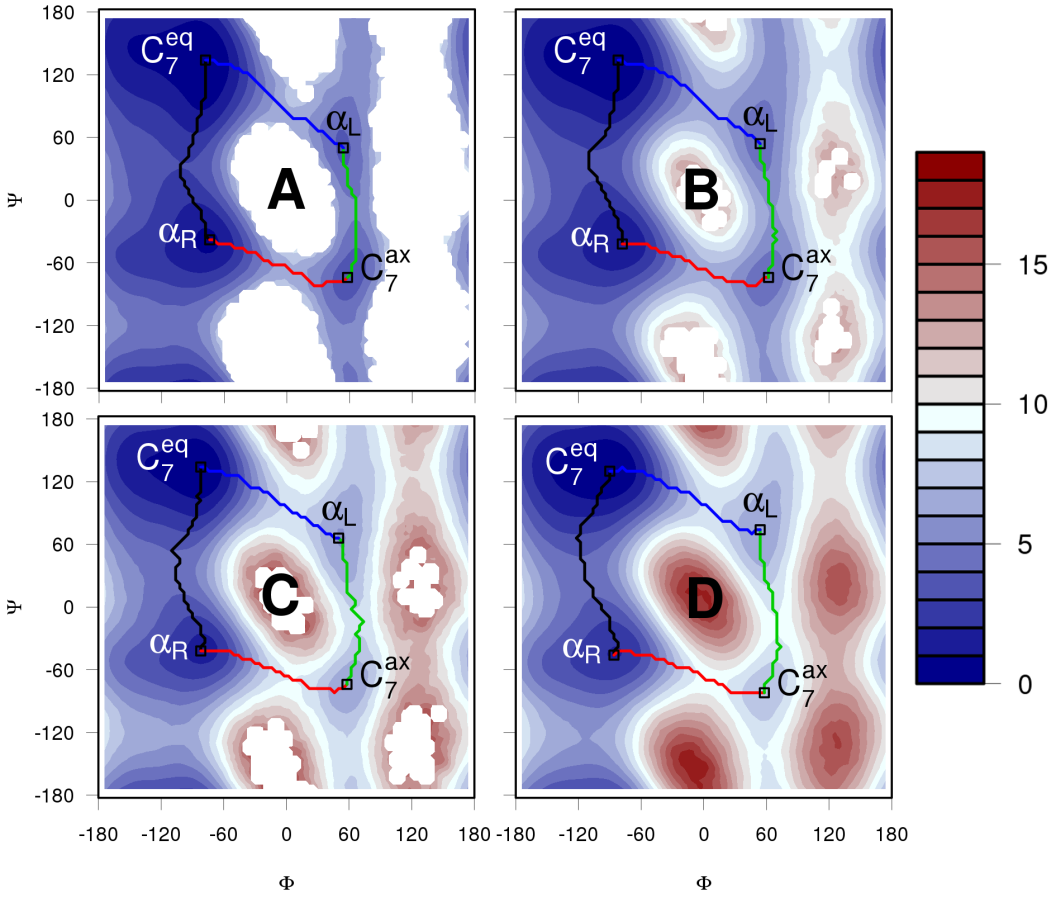


FIGURE 4.3: 2D projections of the FES (ΔF in kcal/mol) obtained with PINS, at $T = \{300.0, 603.40, 917.71, 1395.71\}$ K in panels A, B, C, and D. The 4 Minimum Energy Paths (MEPs) between the 4 stable minima are also indicated. See SI Section 4 for the MEP finding methodology. White areas at lower temperatures correspond to un-sampled areas.

Figure 4.3 shows the 2-dimensional FESs from PINS at four of the twelve simulation temperatures. Starting from $C_{7\text{eq}}$ and following the paths in Figure 4.3 in a counterclockwise direction yields the 1-dimensional minimum energy paths in Figure 4.4. Table 4.C reports the location of the four minima for each of the four plots from Figure 4.3.

The four minima are located at similar coordinates even when the temperature increases to $T =$ 1395.71 K. Up to $T = 603.4$ K (see Figure 4.3B and Table 4.C) the location of each of the minima is still within 4° of the minimum position at $T = 300$ K. A relative displacement of ±6° to ±8° of the (Φ, Ψ) values is found for higher temperatures, see Figures 4.3C and D; the corresponding values are displayed in bold face in Table 4.C. Hence, the four minima $C_{7\text{eq}}$ α_{R} $C_{7\text{ax}}$ α_{L} are well defined over a wide temperature range.

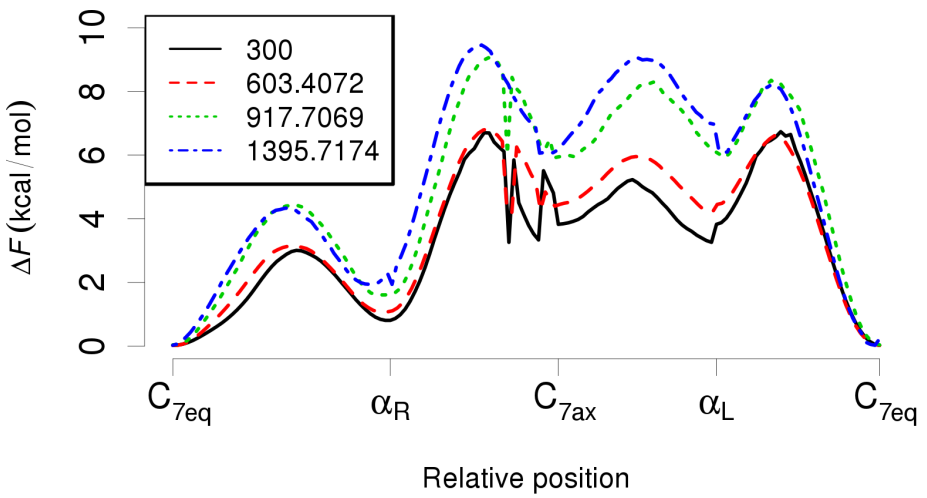


FIGURE 4.4: Free energy profiles of the 4 MEPs defined in Figure 4.3, at 4 different temperatures. The $x-$axis labels are the local minima. The spikes along the path $\alpha_{\mathrm{R}} \leftrightarrow C_{7\mathrm{ax}}$ are caused by insufficient sampling of the top of the barrier.

	$T = 300$ K		$T = 603.4$ K		$T = 917.7$ K		$T = 1395.7$ K	
	Φ	Ψ	Φ	Ψ	Φ	Ψ	Φ	Ψ
$C_{7\mathrm{eq}}$	–78	134	–82	134	–82	134	**–90**	130
α_{L}	54	50	54	54	50	**66**	54	**74**
α_{R}	–74	–38	–78	–42	**–82**	–42	**–86**	**–46**
$C_{7\mathrm{ax}}$	58	–74	62	–74	58	–74	58	**–82**

TABLE 4.C: Location of the 4 stable minima for Alanine Dipeptide (see Figure 4.3) at 4 temperatures. In bold face, values of Φ or Ψ which are more than 4° from the 300 K reference.

A visual comparison with Figures 2 to 4 from Ref. [116], where the authors sampled the system using targeted MD, shows that the MEPs closely match. This further establishes that after post-processing, results from PINS are in favourable agreement with previous studies. The results are also in agreement with Figures 2 and 3 from Ref. [138] showing transition pathways between the four minima. Note that the authors emphasised that they obtained different pathways depending on the starting points, i.e. that the path $C_{7\mathrm{eq}} \rightarrow \alpha_{\mathrm{R}}$ was not following the same points than the path $C_{7\mathrm{eq}} \leftarrow \alpha_{\mathrm{R}}$, but this is probably an artifact of the optimal free energy paths (OFEP) method. Here, the MEPs are determined from Dijkstra's algorithm [139] and are completely reversible, i.e. $C_{7\mathrm{eq}} \leftrightarrow \alpha_{\mathrm{R}}$.

The relative stability of the minima changes as a function of temperature. For the α_{R} minimum the free energy increases from 1.1 to $\sim$ 2 kcal/mol when increasing the temperature by $\sim$ 1100 K. As the barrier between $C_{7\mathrm{eq}}$ and α_{R} also increases by1 kcal/mol, α_{R} is a metastable state also at higher temperatures. For the $C_{7\mathrm{ax}}$ and α_{L} minima the increase of free energy is between 2 and 3 kcal/mol for higher temperatures. The associated barrier increases from 5.5 kcal/mol at 300 K to 9 kcal/mol at 1395 K. The "Left $\leftrightarrow$ Right" transition barriers ($\alpha_{\mathrm{R}} \leftrightarrow C_{7\mathrm{ax}}$ and $\alpha_{\mathrm{L}} \leftrightarrow C_{7\mathrm{eq}}$ transitions) involve free energy barriers of $\sim$ 5 kcal/mol at 300 K which increase to $\approx$ 8 kcal/mol at higher temperatures.

The previous observations can also be considered from an entropy-perspective. The free energy of a conformation i is $F_i = U_i - T_i \times S_i$, where U_i is the internal energy, T_i the temperature and S_i the entropy. For two identical conformations evolving at two different temperatures $T_i \neq T_j$ the potential energies are equal, thus $U_i = U_j$. Hence, the ratio between the free energies is $\frac{F_i}{F_j} = -\frac{T_i}{T_j} \times \frac{S_i}{S_j}$. As for given temperatures the ratio $-\frac{T_i}{T_j}$ is constant, any variation in $\frac{F_i}{F_j}$ must originate from the entropic contribution $\frac{S_i}{S_j}$. This can be seen in Figure 4.4. For the black and red curves ($T_i = 300$ K and $T_j = 603.4$ K) their profiles only differ significantly for the transition $C_{7\mathrm{ax}} \leftrightarrow \alpha_{\mathrm{L}}$ and around α_{R}. The temperature ratio $\frac{T_i}{T_j}$ is ≈ 0.5. The ratios of the free energies $\frac{F_i}{F_j}$ for four different states are as follows:

(i) 1.94 for $C_{7\mathrm{eq}}$, (ii) 1.51 for α_{R}, (iii) 1.73 for $C_{7\mathrm{ax}}$ and (iv) 1.55 for α_{L}. In other words, for $C_{7\mathrm{eq}}$ a value of almost 2 is found which suggests that entropy increases proportional to temperature T. On the other hand, for the $C_{7\mathrm{ax}}$ minimum the ratio of ~ 1.7 points towards an increased unfavourable entropic effect for increasing temperature which is further increased for the two α structures for which the ratio is ~ 1.5. This suggests that hydrogen bonding interactions are entropically destabilized at higher temperatures.

From the interpretation of Figs. 4.3 and 4.4, combined with Table 4.C, it was shown that PINS can also be used for investigating properties of a system of interest at several thermodynamic states. For alanine dipeptide, an extended and accurate conformational study was performed, from which it was deduced that the four minima keep their relative order of stability ($C_{7\mathrm{eq}}, \alpha_{\mathrm{R}}, C_{7\mathrm{ax}}, \alpha_{\mathrm{L}}$) with increasing temperature, and that the transitions between them follow similar pathways. It was also emphasised that the increase of free energy is higher for the two less stable minima than for the more stable ones: although the high temperatures allow an increased sampling of the positive Φ minima ($C_{7\mathrm{ax}}$ and α_{L}) of Figure 4.3, the negative Φ ones ($C_{7\mathrm{eq}}$ and α_{R}) still remain the most stable and most sampled minima for a broad range of temperatures. This observation can be explained by an evolution of the entropic contribution to the free energy.

4.2 Supplementary investigations for the deca-alanine

In the following a set of performance measures for evaluating the sampling boost obtained with PINS over PT is introduced. The system of interest is the Alanine deca-peptide, also referred as deca-alanine. PT and PINS simulations were running for 100 ns per replica with a time step of 1 fs, replica exchange was attempted every 100 steps (i.e. 0.1 ps), same for the I/O frequency. CHARMM with the GENBORN implicit solvent model were used. 16 temperatures are considered, ranging from 300 K to 1139.03 K, in order to keep a constant ratio $\frac{T_{i+1}}{T_i}$ (see the following Figure 4.5).

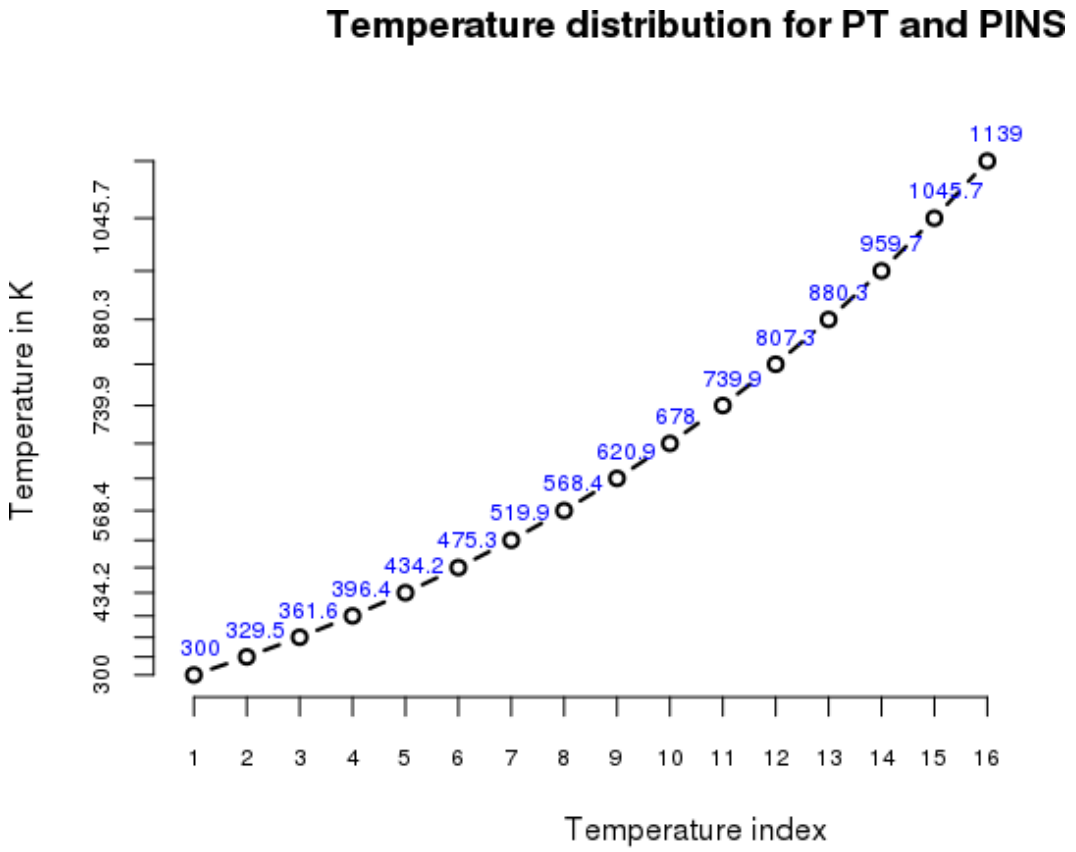


FIGURE 4.5: The 16 temperatures used when studying alanine deca-peptide with the implicit GENBORN solvent model, following a geometric progression

Definitions of round-trip time The number of sampling moves required to traverse the computational temperatures ensemble, usually defined as the round-trip time, was shown to be a convenient measure of sampling performance.[140–142]. It was already successfully used for evaluating the efficiency of PINS over PT.[58]

The round-trip time between temperatures T_i and T_j is formally defined as the simulation time required for bringing a replica:

1. first from T_i to T_j
2. and then back from T_j to T_i

If N is the total number of replicas, and δt the time interval between 2 replicas exchange, then the shortest possible round-trip time r_t^{ideal} is :

$$r_t^{ideal} = 2 * \delta t * N \tag{4.1}$$

I.e. we assume that all $p_{i->j}$ transitions have a probability of 1, and are all accepted. With the current parameters this would be $r_t^{ideal} = 3.2$ ps.

But in practice it is observed that $r_t >> r_t^{ideal}$ because the δt between 2 accepted transitions is much more larger than replica exchange attempt frequency, because of rejected exchanges.

Occupation traces for replica 1 The following Figure 4.6 shows the occupation traces for PT (left) and PINS (right), for 100 ns long simulations (per replica). The replica considered is the first one, i.e. initially starting at 300 K.

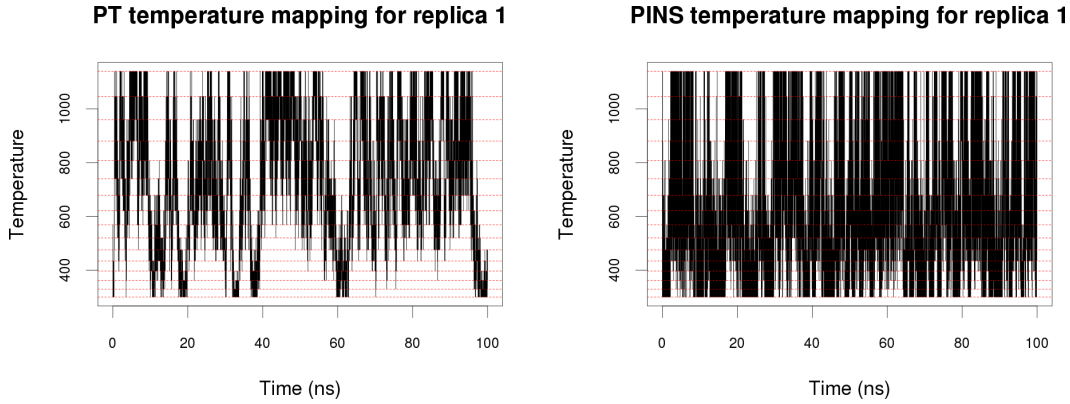


FIGURE 4.6: Traces for replica 1 (T = 300 K initially) for PT (Left) and PINS (right). Red dashed lines correspond to the temperatures defined in Figure 4.5.

One can immediately see that for PT the first round-trip time is of the order of a few nano-seconds and that a few events are observed over 100 ns. For PINS much more round-trip events are observed.

Statistical analysis of the round-trip time r_t A script was written for counting all the round-trip events. From that a list of round-trip time is built, and statistical analysis is performed, and summarized in the following Table 4.D. Only the replica 1 (initially T = 300 K) is considered, thus the round trip $T_1 \rightarrow T_{16} \rightarrow T_1$ is estimated.

It can be seen that the average round-trip time is 10.746 ns for PT vs. 0.998 ns for PINS, i.e. PINS can propagate T_1 to T_{16} one order of magnitude faster. The standard deviation is also one order of magnitude larger, which provides a good 95% confidence interval for PINS, but clearly not for PT.

Distribution validation with histograms and a non-linear fit:

	PT	PINS
Observations	9	100
Mean r_t (ns)	10.746	0.998
Std. dev. on r_t (ns)	8.596	0.889
Conf. interval (95%) on r_t (ns)	[4.138;17.353]	[0.822;1.175]

TABLE 4.D: Statistical analysis for the round-trip time r_t defined for $T_1 \rightarrow T_{16} \rightarrow T_1$.

The following Figures 4.7 and 4.8 show the Histogram distributions of r_t for PT and PINS. The red lines represent the population (number of observations) for each bin of the histogram.

They confirm the previous observation for PINS, i.e. there seem to be an underlying probability distribution function describing r_t. For PT the sample is clearly too small for extracting any valid hypothesis.

For the PINS results it is possible to fit the observations (red line) to a non-linear model:

$$y \sim a * \exp(-b * x) \tag{4.2}$$

The blue dashed line represents the result of the fit. Fit parameters and the residual sum of squares (R.S.S.) are also displayed on the plot and available in the following Table 4.E. The value of **b** can be interpreted as a decay constant in units of ns^{-1}.

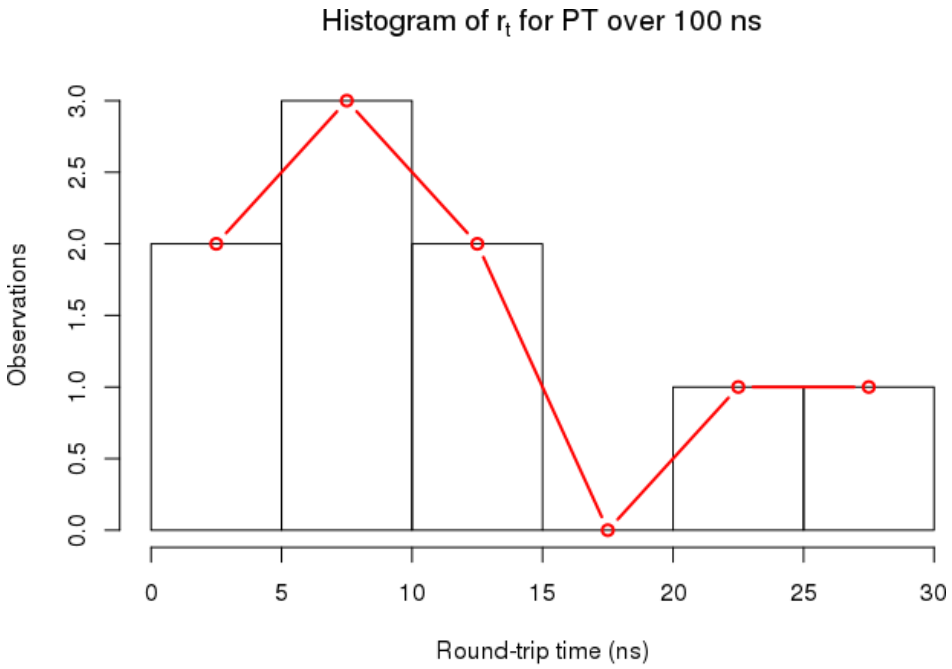


FIGURE 4.7: Histogram build from PT r_t observations. No clear probability distribution function (red line) could be defined.

Autocorrelation analysis of the temperature traces Another meaningful performance measure that one can have a look at is the autocorrelation function of the occupation trace, as pointed by Plattner et al.[58]

Definition and implementation:

Let f_i^α be the temperature at step i for the trace of replica α (i.e. the replica initially running at $T = \alpha$ when $i = 0$). The autocorrelation function $C^\alpha(s)$ is defined as :

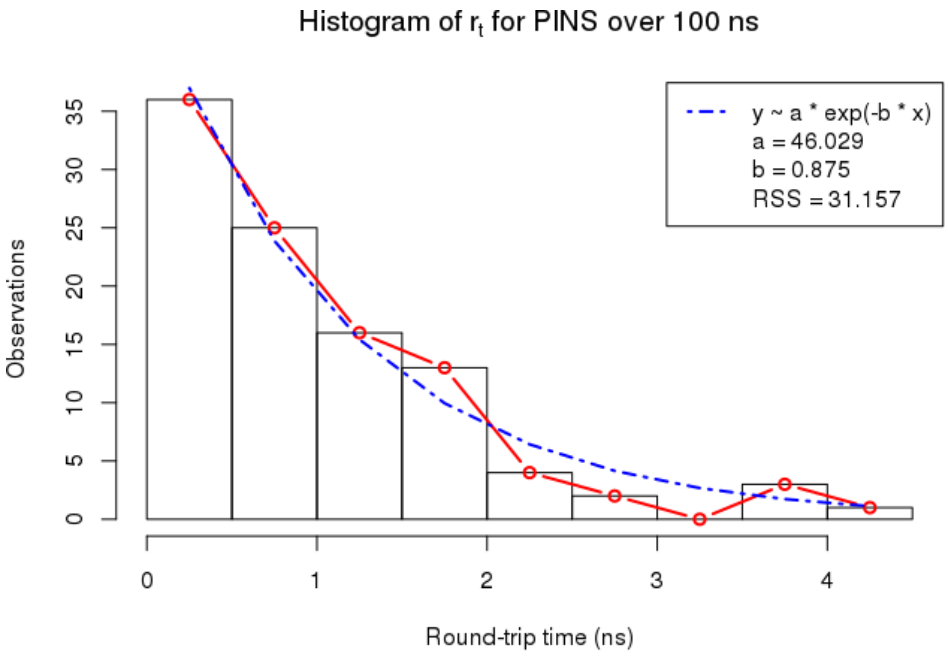


FIGURE 4.8: Histogram build from PINS r_t observations. The red line shows an underlying probability distribution function, and the blue line a possible non-linear regression using Equation 4.2.

Parameters	Values
a	46.029
b (ns^{-1})	0.875
R.S.S.	31.157

TABLE 4.E: Fit parameters for Equation 4.2, and their least-square optimised values. The residual sum of squares (R.S.S) is also provided. Parameter **b** has an inverse unit of time (ns^{-1}) and can be interpreted as a decay constant.

$$C^{\alpha}(s) = \frac{\mathbb{E}\left[(f_i^{\alpha} - \mu) * (f_{i+s}^{\alpha} - \mu)\right]}{\sigma^2} \tag{4.3}$$

where s is the lag time at which to estimate the autocorrelation, μ and σ^2 are respectively the mean and variance estimated over all the f^{α} observations, and $\mathbb{E}[\cdots]$ denotes an expected value.

The estimation of C^{α} using a naive algorithm (Equation 4.3) is of complexity $\mathcal{O}(n^2)$, thus a scan over the whole 100 ns of simulation would be considerably time-demanding. However it is possible to use the Wiener–Khintchine theorem [143, 144], which relates autocorrelation and power spectrum, in order to estimate C^{α} over the whole time interval with a reduced complexity of $\mathcal{O}(n \log n)$:

$$C^{\alpha} = \frac{1}{\sigma^2}\mathcal{F}^{-1}\left(|\mathcal{F}(f^{\alpha} - \mu)|^2\right) \tag{4.4}$$

where $\mathcal{F}(\cdots)$ is the Fast Fourier Transform (FFT).

Application:

The following Figures 4.9 4.10 4.11 show the autocorrelation function estimated for PT and PINS for the case where $\alpha = \{1, 8, 16\}$, i.e. for the replica initially at $T = \{300, 568, 1139\}$ K.

Equation 4.4 is used in order estimate C^{α} over the whole 100 ns of trajectory, although the x-axis (lag time) is adjusted in order to display only the first 1.5 or 2.5 ns which is enough for PINS to reach

values of $C(s) \approx 0$.

One can see that PINS always reaches quasi-uncorrelated temperature distribution for each of the trace much faster than PT does. An autocorrelation value of $C(s) \leq 0.2$ is usually reached within 0.5 to 1 ns for PINS, where it usually takes 2 to 6 ns for PT (extended traces not shown here). Similarly, autocorrelation values $C(s) \leq 0.1$ appear within 2 ns for PINS, against 4 to 10 ns for PT.

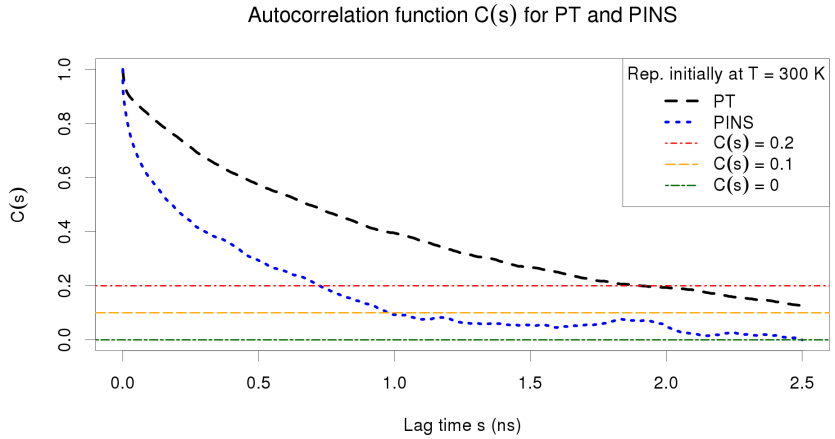


FIGURE 4.9: Autocorrelation for PT and PINS for replica initially at T = 300 K

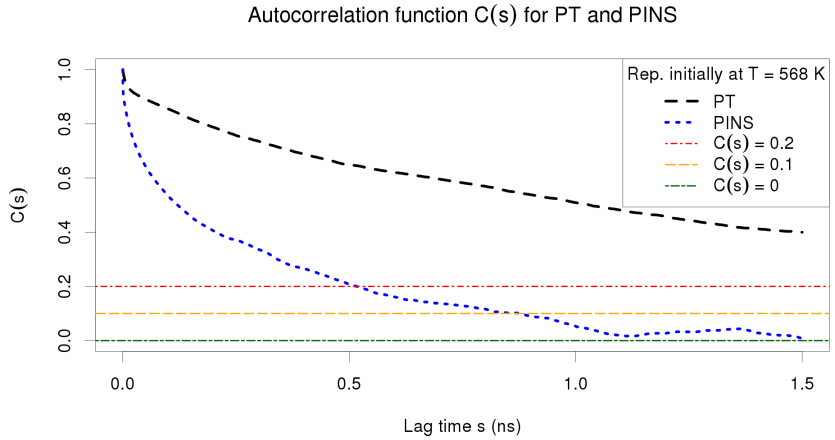


FIGURE 4.10: Autocorrelation for PT and PINS for replica initially at T = 568 K

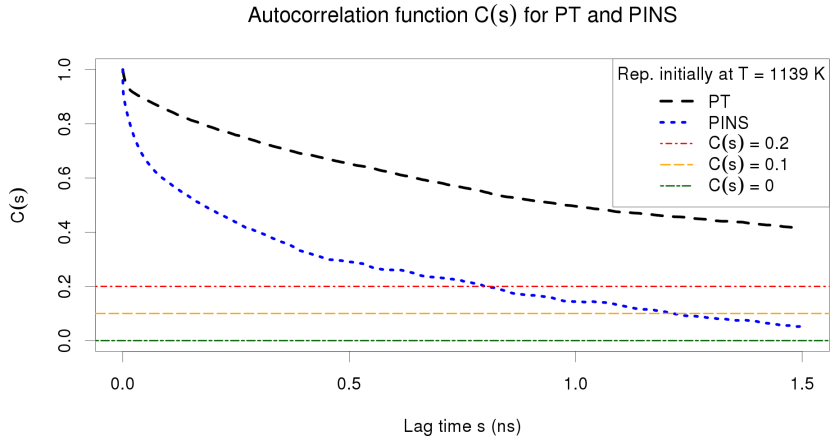


FIGURE 4.11: Autocorrelation for PT and PINS for replica initially at T =1139 K

4.3 PINS article

This is the ready to be published article illustrating implementation and validation of the PINS algorithm. It should be resubmitted soon in *Journal of Chemical Theory and Computation*. It was co-written with Nuria Plattner, Jimmie. D. Doll and Markus Meuwly. Several pages of Supplementary Information are also included.

Partial Infinite Swapping: Implementation and Application to alanine-decapeptide and Myoglobin in the Gas Phase and in Solution

Florent Hédin,[†] Nuria Plattner,[‡] J. D. Doll,[¶] and Markus Meuwly[*,†,¶]

Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland., Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, D-14195 Berlin., and Department of Chemistry, Brown University, Providence, Rhode Island 02912, USA.

E-mail: m.meuwly@unibas.ch

Abstract

Partial infinite swapping (PINS) is a powerful enhanced sampling method for complex systems. PINS is based on infinite swapping (INS) which constructs an expanded ensemble from K replicas at different simulation temperatures. Contrary to parallel tempering (PT), INS uses the fully symmetrized distribution of configurations in temperature space. Due to the factorial growth of the number of permutations of the K replicas, applications employ PINS which uses a block structure whereby full symmetrization is used in each block. Thermodynamic observables are determined in a post-processing step. PINS is applied to problems of different complexity: the

[*]To whom correspondence should be addressed

[†]Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland.

[‡]Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, D-14195 Berlin.

[¶]Department of Chemistry, Brown University, Providence, Rhode Island 02912, USA.

conformational space and free energy landscape for folding of alanine-decapeptide in implicit solvent and in solution, and Xenon migration in Myoglobin. In every case conformational free energy surfaces are determined. The efficiency of PINS is evaluated by comparing with Molecular Dynamics, Parallel Tempering and Umbrella Sampling methods and is often found to be more efficient for the problems considered.

1 Introduction

Molecular Dynamics (MD) and Monte Carlo methods (MC) are widely used for characterizing biological processes using computer simulations. Although the computational resources are continuously increasing, sampling the large conformational space available for proteins is very challenging. However, in order to provide an atomistically refined picture for processes such as protein folding or large conformational changes with functional relevance, such rare events must be sampled. As they usually occur on time scales of the order of microseconds (or longer), directly sampling them with unbiased MD or MC simulations is difficult. Hence, enhanced sampling methods are a possible way forward as such approaches increase the probability for accessing low-probability configurations.

An important aspect which is subject to continuous improvement efforts is the ability to sample rare-events, a particular challenge for complex systems. For systems in which configuration space is well connected, standard techniques (e.g. Metropolis-Hastings[1–3]) are efficient. However, for situations in which configuration space decomposes into poorly connected subregions or where barriers between neighbouring states are high, enhanced sampling is required. Several such methods for rare event sampling have been developed in the past. They include parallel tempering (PT),[4–6] umbrella sampling (US),[7] metadynamics,[8] or replica exchange (RE).[9] The methods either use a bias to steer the system between regions in configuration space (US, metadynamics) or they expand thermodynamic state space as is done for PT or RE. This contrasts with conventional stochastic methods which typically

use random walks for generating a statistical sampling of the desired equilibrium probability distribution.

Another method which has recently been investigated is Partial Infinite Swapping (PINS) [10–13] which is based on the PT/RE algorithms. PINS uses a symmetrisation strategy for combining probability distributions at different temperatures, so that they become more connected and thus easier to sample than the original ones. The present work discusses the statistical reweighting to extract thermodynamic information from PINS simulations [11,13] and applications to two systems: the alanine decapeptide (or deca-alanine) which becomes a challenging system particularly in explicit solvent, [14–21] and Xenon migration in Myoglobin, [22–29] a system requiring extensive direct sampling of the free energy surface.

2 Computational Methods

Considering a Canonical (NVT) ensemble, the probability $\rho(\mathbf{X})$ of observing a system in state $\mathbf{X}$ is related to its potential energy $V(\mathbf{X})$ through

$$\rho(\mathbf{X}) = \frac{1}{Z} e^{-\beta V(\mathbf{X})} \tag{1}$$

where $\mathbf{X} = X_1, ..., X_k$ is a $k-$dimensional vector (where $k = 3$ for MC or $k = 6$ for MD), populating a subset D of the configuration space $\mathbb{R}^{kN}$, Z is the canonical partition function $Z \sim \int^{D \subset \mathbb{R}^{kN}} e^{-\beta V(\mathbf{X})} d\mathbf{X}$, and $\beta = 1/k_{\mathrm{B}}T$ is the inverse temperature and k_{B} the Boltzmann constant.

Parallel Tempering (PT) (also known as Replica Exchange (RE)) methods [4–6] were successfully applied to investigating a wide range of chemical and biological systems. In PT K replicas are considered and the partition function Z of the overall ensemble is:

$$Z = \prod_{i=1}^{K} \frac{q_i}{M!} \int d\mathbf{X}_i \; e^{-\beta_i V(\mathbf{X}_i)} \tag{2}$$

where $q_i = \prod_{k=1}^{M} (2\pi m_k k_B T_i)^{3/2}$ is obtained by integrating out the momenta of the M particles with mass m_k, $V(\mathbf{X}_i)$ is the potential energy for the coordinates $\mathbf{X}_i$, and $\beta_i = 1/k_{\mathrm{B}}T_i$ is the reduced temperature for replica i. In the simulations, replicas are exchanged between two adjacent temperatures $T_i \leftrightarrow T_j$ with probability

$$P_{acc}(i \leftrightarrow j) = \min\{1, e^{(\beta_i - \beta_j)(V(\mathbf{X}_i) - V(\mathbf{X}_j))}\} \tag{3}$$

The K temperatures are usually distributed non-linearly between T_1, which is the desired simulation temperature, and T_K, in order to have a constant value of $P_{acc}(i \leftrightarrow j)$, typically around 20 to 25% (see Ref.[6] for a discussion on the choice of temperatures and the impact on P_{acc}).

2.1 Infinite Swapping limit for Parallel Tempering simulations

The infinite swapping (INS) method[10–13] also uses an expanded ensemble built from a number of replicas at different temperatures. Contrary to PT, INS uses the fully symmetrized distribution of configurations in temperature space, whereas PT just occasionally enriches the local temperature with configurational information coming from simulations at a higher temperature. Formally, INS is based on a mathematical analysis of the convergence rate of PT simulations as a function of the temperature swap attempt frequencies.[10,12] It was proven[12] that this convergence rate is a monotonically increasing function of the swap rate, and thus optimal sampling is reached in the *infinite swapping* limit.

In other words, INS provides optimal sampling for a given replica by using information from all other temperatures used in the simulation. This could be achieved by allowing exchanges

between all replicas at each time step. However, as there are $K!$ exchanges for K replicas this would become an unmanageable number of exchanges for realistic choices, e.g. $K = 20$. Furthermore, many of the exchanges would not be accepted which would further compromise the efficiency of the method. Instead of attempting all $K!$ exchanges and estimating their acceptance P_{acc} following Eq. 3 for each permutation, the probability of such a general exchange is estimated according to

$$\rho_k(\mathbf{X_i}) = \frac{p_k(\mathbf{X_i})}{\sum_{k=1}^{K!} p_k(\mathbf{X_i})}. \tag{4}$$

Here $p_k(\mathbf{X_i})$ is given by

$$p_k(\mathbf{X_i}) = \prod_{i=1}^{K} e^{-\beta_i V(\mathbf{X}_{k,i})} \tag{5}$$

and $\mathbf{X}_{k,i}$ is the configuration of replica i corresponding to the assignment of configurations to temperatures in permutation k. Therefore, with INS the optimal permutation is found by comparing the $\rho_k(\mathbf{X_i})$ permutation probabilities. Nevertheless, for large systems this includes a large number of permutations, and it is computationally too expensive to calculate all $K!$ probabilities.

For putting INS to practical use, the partial infinite swapping (PINS) algorithm was introduced.[10–13] PINS uses a partitioning strategy whereby temperature space is divided into blocks, and local (but full) symmetrisation is used within each block. More precisely, the current implementation uses the "dual-chain" approach[11], where the $K-$temperature set is partitioned into blocks in two different ways, one for each chain. The two blocks must have a complementary structure without a boundary between the blocks defined for the two chains. This is required in order to achieve sampling of the overall temperature space for all the replicas. For a set of 12 temperatures, a possible partitioning for the two chains $(a|b)$ is $(3, 6, 3|4, 4, 4)$, where the boundaries for chain a are between T_3 and T_4, T_9 and T_{10}, and for

chain b they are between T_4 and T_5 and T_8 and T_9, respectively. On the other hand, the partitioning $(3, 3, 6|6, 3, 3)$ is not valid, as chains a and b share a common boundary between T_6 and T_7.

2.2 Implementation of PINS

A previous implementation into CHARMM was already described and validated[13], using the available ENSEMBLE module. However, the CHARMM code underwent profound changes in its architecture for improved compatibility with modern Fortran standards and use of novel computational algorithms such as domain decomposition[30] combined with GPU-based calculations. Hence, it was decided to implement PINS with a dual-chain approach into the most recent CHARMM c41 release.

Similar to standard PT simulation, PINS requires K replicas, and the temperatures $\{T_1, \cdots, T_K\}$ at which they are run. The user also provides a frequency of attempted exchanges between replicas. The sampling efficiency of PT simulations is optimal with attempted exchanges at each MD step, see above. However, this requires communication of the coordinate vectors, using technologies such as message passing (MPI) which is a bottleneck for the simulation of large systems as inter-node communication is usually slower than computation. It is thus required to attempt exchanges as often as possible, but not too often to avoid inter-node communication saturation. For a concrete application the best choice depends on (i) the system size because the smaller the system, the larger the ratio of communication time/calculation time, and (ii) hardware/software considerations, mainly the maximum communication speed possible between two replicas running on different compute nodes.

2.3 Statistical reweighting phase as a post-processing step

PINS provides data at all thermodynamic states (characterized by the T_i) which can be used for computing properties at a given state. The sampling convergence is significantly improved by this step which requires reweighting of the data collected at different T_i. This step can either be performed during the simulation ("on the fly" reweighting), or at the end of the simulation, as a post-processing step. For the current PINS implementation it was decided to employ post-processing.

The first step is to obtain a list of the permutations at a given step of the simulation. For this the following simulation parameters are required: (i) the total number of simulation steps, and the swapping frequency at which the PINS algorithm was applied, (ii) the number of temperatures K, (iii) the dual chain parameters, i.e. the number of temperature blocks, and the number of temperatures within each block (see above). It is also necessary to save the potential energy $V(\mathbf{X}_i)$ at each of the atomic configurations along the trajectory. With this information it is possible to calculate the overall probability of all the attempted permutations at a given simulation step i using Eq. (6)

$$\rho(\mathbf{X_i}) = \sum_{p=1}^{P} \rho_p(\mathbf{X_i}) \tag{6}$$

where P is the total number of permutations allowed by the dual-chain structure, and the right hand side is given in Eq. 4. For PINS to be computationally efficient it is essential that $P << K!$ is fulfilled.

The next step is the reweighting of the estimated thermodynamic property which depends on the observable of interest. For increased flexibility it is advantageous to separate the sampling and the post-processing. Consider a 2D free energy surface (FES) built from two variables A and B (illustrated later with an application to the alanine-decapeptide). Values

of A and B are first estimated for each configuration from each trajectory. From a normalised 2D histogram the free energy is therefore

$$\Delta F^s(A_i, B_j) = -RT^s \ln(\rho(A_i, B_j)) \tag{7}$$

where the s superscript denotes a single thermodynamic state estimate, and where $\rho(A_i, B_j)$ is the probability density for cell (i, j) on the 2D grid. An estimate based on data generated at various thermodynamic states (m superscript) of Eq. (7) is obtained from

$$\Delta F^m(A_i, B_j) = \Delta F^s(A_i, B_j) \cdot \sum_m \rho(\mathbf{X_m}) \tag{8}$$

where (i, j) run over the discretised cells, and m over all the thermodynamic states. This procedure is easily adapted and extended to other thermodynamic property which can be estimated from a discretised grid.

3 Ala_{10} in Implicit and Explicit Solvent

Alanine decapeptide $(Ala)_{10}$ is a chain of 10 residues folding to a regular α-chain structure in solvent. Its stability is the result of several favourable hydrogen bonding interactions. $(Ala)_{10}$ has been used as a test system in the recent development of several optimised sampling MD techniques such as unconstrained MD (explicit solvent),[14] adaptive steered molecular dynamics (in vacuum),[15] Multi-Replica and Multiple-Walker Adaptive Biasing Force (in vacuum),[16] or Simulated Tempering (explicit solvent),[17] and thus is a suitable benchmark system. The thermodynamic stability of the α-chain and folding or unfolding pathways were also investigated in vacuo[19] and in explicit TIP3P solvent[21,31], and free enthalpy differences between the α-chain and two less stable π- and 3_{10}- chains were also recently investigated (coarse-grained water model).[20] However a recent study[21] showed that folding/unfolding is much more complex than the previously reported "accordion-like scheme"[32] in explicit sol-

vent, and indeed involves an extended set of non-helical and compact states. This system was investigated with the new PINS CHARMM implementation, and results compared to Molecular Dynamics (MD) and Parallel Tempering (PT) simulations running with (when possible) identical simulation parameters for direct comparison.

3.1 Simulations in Implicit solvent

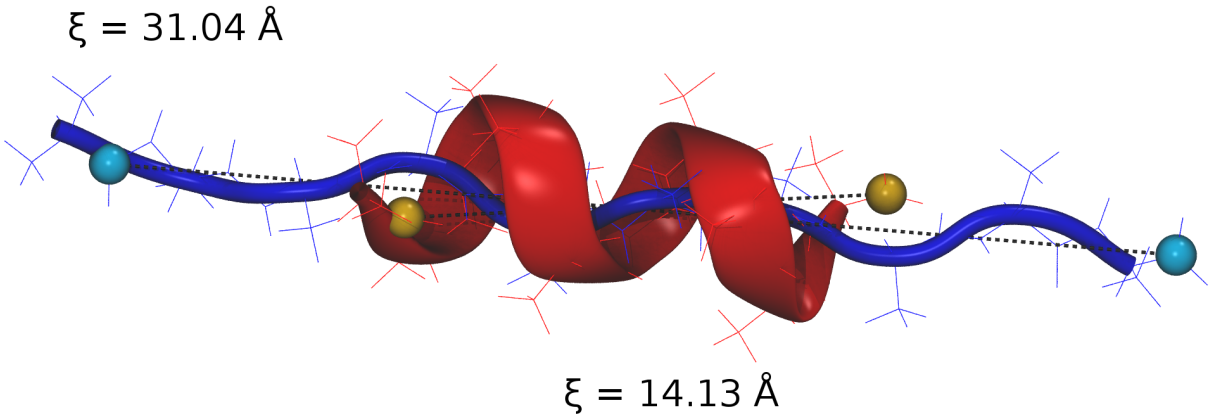


Figure 1: (Ala_{10}): extended starting structure (blue), and folded structure (red) obtained after 100 ns of MD with GENBORN implicit solvent. In cyan and orange, the carbonyl-carbon atoms define the end-to-end distance ξ in Å, used for following folding and building ΔF surfaces. The extended structure has $\xi = 31.04$ Å, and the α-helical structure is characterised by a $\xi = 14.13$ Å.

All simulations were run with CHARMM c41, used the CHARMM Force Field version 36, including CMAP corrections[33,34] the GENBORN implicit solvent model and started from an extended structure (Figure 1, in blue). A time step of 1 fs was used, and three independent simulations of 100 ns each were first performed with each method (MD, PT and PINS). For all structural comparisons the folded structure (red in Figure 1) was the reference for

computing the RMSD.

The temperature was 300 K for the MD simulation, and for the PT and PINS simulations an ensemble of 16 replicas at the following temperatures was used: 300.00, 329.52, 361.58, 396.42, 434.24, 475.31, 519.92, 568.35, 620.92, 677.99, 739.94, 807.32, 880.32, 959.70, 1045.71, 1139.03 K. Those temperatures were chosen using a temperature prediction algorithm[35] available as a free web-service, which generates a temperature set optimised for obtaining a desired exchange acceptance ratio, which was chosen to be 40% in the present case. The dual-chain PINS approach with two chains of 3 blocks $(6, 6, 4|4, 6, 6)$ was used.

Figure 2 shows the RMSD(t) for simulations with GENBORN for MD (top), PT (middle) and PINS (bottom). The top panel (MD) corresponds to the first 20 ns of Figure S1 from the SI. The red label indicates the simulation time required to reach the first structure with RMSD < 2 Å which occurs by 11.99 ns for MD, whereas for PT and PINS this threshold is reached within 0.40 ns and 0.12 ns, respectively. Thus for this case, PINS reaches a compact state two orders of magnitude more rapidly than MD, and approximately three times faster than PT.

Secondly, it is of interest to assess the gain in term of diversity of sampling provided by PINS compared to PT. A simple visual analysis of the middle and bottom plots of Figure 2 shows that the RMSD fluctuations from PINS (bottom) are usually larger compared to PT (middle), compatible with a more exhaustive sampling in RMSD space. In particular, taking RMSD < 2 Å as the threshold, many more recrossings are found with PINS compared to PT or MD which leads to a more diverse set of structures.

RMSD analysis and k-means clustering: The classification of recurring structural motifs can be based on different measures. In the following RMSD is used together with $k-$means.[36–38]

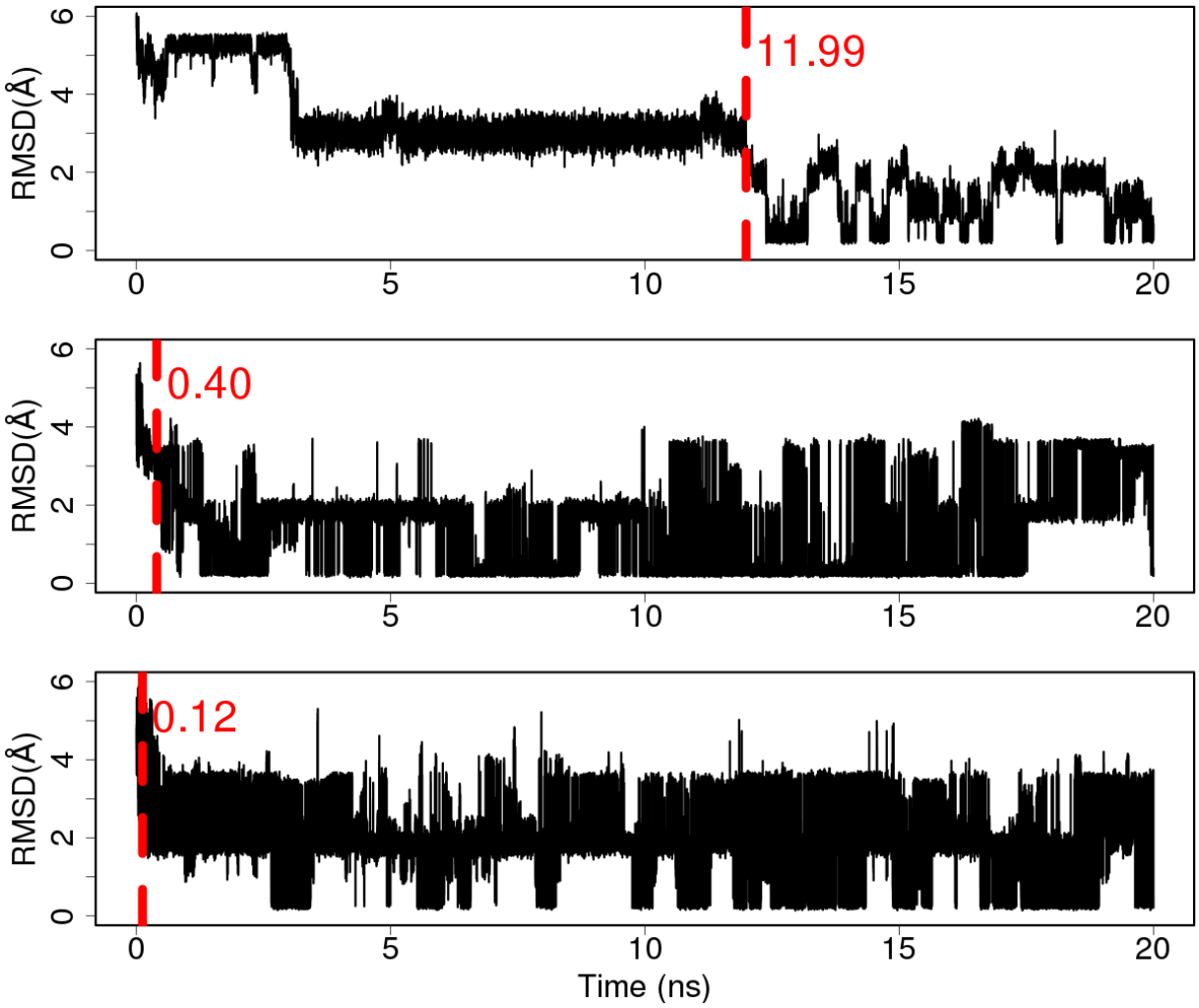


Figure 2: RMSD fluctuations for folding of (Ala_{10}). 20 ns (from a total of 100 ns) of MD (top), PT (middle) and PINS (bottom), with GENBORN solvent. The red vertical lines indicate the time at which RMSD$\leq$ 2 Å (partially helical compact state). The reference structure is the α-helix from Figure 1.

Section 2 from the SI contains details on the methodology followed for performing the clustering, and justifies the choice of $k = 6$ clusters, used for all the MD PT and PINS analysis.

Table 1 summarizes the results of a clustering of the data shown in Figure 2, where the centers are sorted by increasing RMSD, and the columns contain the relative population of each center. As observed above, the first occurrence of a compact (RMSD $<$ 2 Å) state happens on a much longer timescale for MD than for PT or PINS ($\approx$ 12, 0.4 and 0.2 ns

Table 1: $K-$means clustering with $k = 6$ centers applied to RMSD fluctuations from Figure 2. Clusters are sorted by increasing RMSD. PT (81 %) and PINS (71 %) both show an increased sampling of the low RMSD centers (RMSD < 2 Å) compared to MD (65 %). See SI Section 2 for justification of $k = 6$.

MD		PT		PINS	
centers (Å)	pop. (%)	centers (Å)	pop. (%)	centers (Å)	pop. (%)
0.4	15.8	0.3	38.7	0.3	22.8
1.1	9.2	1.2	3.7	1.2	4.2
1.8	39.7	1.9	40.7	1.9	45.3
2.2	19.4	2.8	2.8	2.2	10.5
3.0	12.8	3.5	8.9	3.0	3.6
5.1	3.0	4.2	5.2	3.6	13.6

respectively), so the length of the clustered dataset should be sufficiently large in order to include enough transitions between compact and intermediate states, for the three methods of interest. As PT and PINS quickly converge to compact structures, choice of the dataset length was based on MD results. It was chosen to perform clustering on the first 20 ns of the trajectories, which is also the timescale reported on Fig. 2. Indeed, with a lower value none or a few of the transitions would have been observed, and with a longer timescale the population of the lower clusters for MD would have been overestimated relatively to the population of the other transition centers, because of a poor sampling compared to PT and PINS.

It is found that PT (39 %) and PINS (23 %) lead to a larger population of the lowest RMSD cluster around 0.3–0.4 Å compared to MD (16 %). It is also interesting to note that the cluster center with the largest RMSD is centered at 5.1, 4.2 and 3.6 Å for MD, PT, and PINS, respectively, which confirms that PT and PINS lead to an enrichment of compact configurations. Furthermore, PT and PINS lead to very similar cluster centers for the three most compact states whereas the next three cluster centers are more compact for PINS compared to PT, highlighting that PINS favours compact states. This seems to confirm that PINS samples more stable and metastable, partially folded state, than PT or MD simulations. This is supported by visual inspection of Figure 2 (Center and Bottom) where the

bottom plot (PINS) shows a larger amplitude in the fluctuations than the top and middle ones, corresponding to the MD and PT RMSD analyses, respectively.

Next, the ensemble of MD, PT and PINS structures from Figure 2 was clustered which leads to only one set of cluster centers. Then, the structures from each method were projected onto the cluster centers which yields their population for each method (see Table S1 from the SI Section 2). Again, PT and PINS yield a larger population of the most compact state. Furthermore, the transition between compact (RMSD < 2.0 Å) and extended (RMSD > 3.0 Å) structures is more frequently sampled, as is also evident from Figure 2.

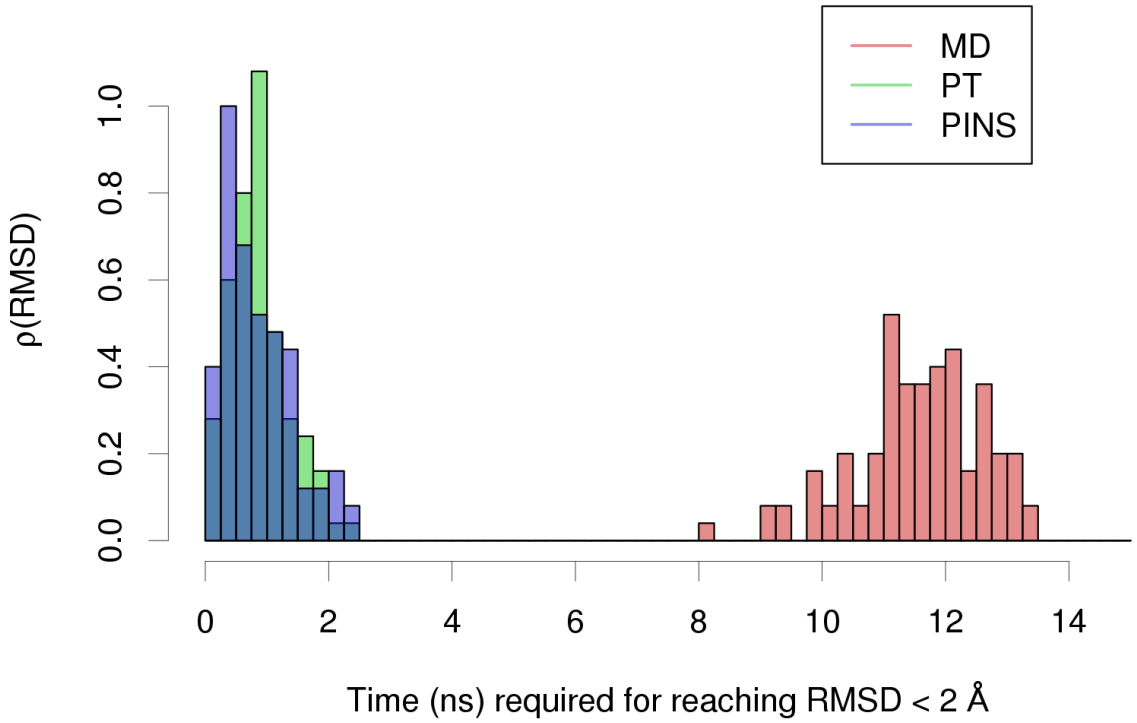


Figure 3: Histograms of the required time before reaching a RMSD of less than 2 Å, for 100 simulations of MD (red), PT (green), and PINS (blue). Simulations of 20 ns maximum for MD, and 5 ns for PT-PINS, all using the implicit GENBORN solvent model.

By repeating such simulations one hundred times for each of the three methods (for a maximum of 20 ns for MD, and 5 ns for PT-PINS), the distribution of times required before

sampling a structure with RMSD < 2 Å ("folding time") is obtained (see Figure 3). This confirms the results from a single run (Figure 2). For MD (red) most of the folding events center around 12 ns with a broader distribution, whereas for PT (green) the folding times range from 0.5 to 2.0 ns, with a peak around 1.0 ns, and for PINS (blue) from 0.25 to 2.0 ns with two peaks at 0.5 and 1.0 ns. This validates that PINS converges to a folded structure faster than PT, albeit only slightly more rapidly.

End-to-end distance and ΔF profiles: In a next step the end-to-end distance ξ between the carbonyl carbon atoms of the first and last residue was analysed (see Figure 1). This coordinate was already used previously for monitoring the progress of folding.[21,32] The α-helical structure was assigned to $\xi = 15.2$ Å or $\xi = 14.2$ Å, and linear structures were associated with $\xi = 33.0$ Å or $\xi = 32.0$ Å, respectively.[21,32] Structures shown in Figure 1 correspond to $\xi = 14.1$ Å (red α-helix) and $\xi = 31.0$ Å (blue). Compact structures are usually defined as configurations with $\xi \leq 16.75$ Å.[21]

Figure 4 shows free energy profiles from MD (black), PT (red) and PINS (blue) simulations. They were generated from the 100 ns simulations by extracting and binning the end-to-end distance ξ from which the Helmholtz Free Energy was estimated according to $\Delta F(\xi) = -\mathrm{RT}\ln(\rho(\xi))$, where $\rho(\xi)$ is the normalized density. The error bars correspond to the statistical 95% confidence interval. From the present simulations, minima were found at $\xi = 14.7$ Å (MD) and $\xi = 14.5$ Å (PT and PINS), respectively. The extended states ($16.0 \leq \xi \leq 28.0$ Å) are associated with free energies ranging from 1.0 to 6.0 kcal/mol.

PINS simulations can be used for investigating the stability of Ala_{10} in implicit solvent at higher temperatures. Figure 5 reports ΔF at five different temperatures. First, ΔF curves for 300 K and 329 K are fairly similar. This suggests that the decapeptide is stable at ambient temperatures. The α-helix structure, for $\xi = 14.5$ Å is still found to be the most

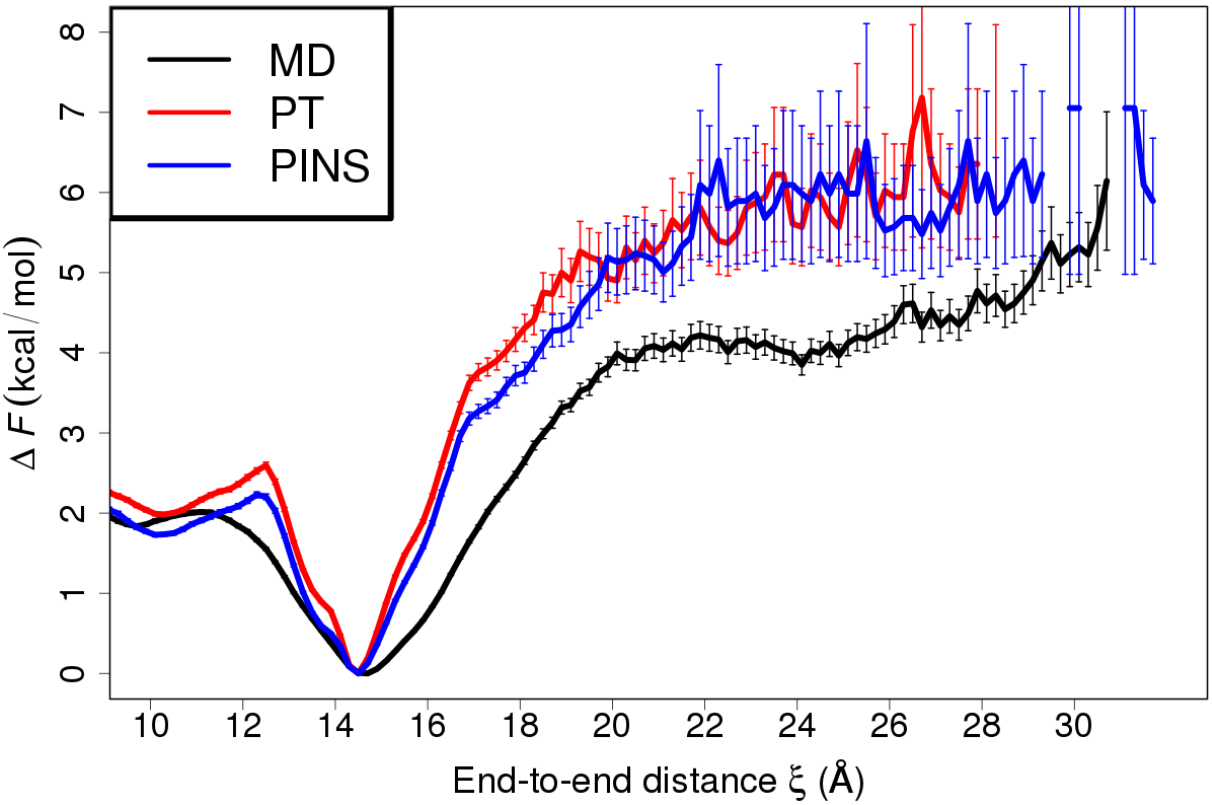


Figure 4: Free energy profile (ΔF in kcal/mol) built using the end-to-end distance ξ between carbonyls' carbon from first and last residue of Ala_{10} (see Figure 1) in implicit GENBORN model. Estimated for a total simulation time of 100 ns of MD (black), PT (red) and PINS (blue). The error bar is the statistical 95% confidence interval.

stable state at 329 K. At $T = 396$ K (green curve on Fig 5), this is still the case, but it is observed that extended states ($\xi > 15$ Å) start being more sampled and thus more stable. At $T = 568$ K (blue curve of Fig 5), the funnel-like structure centered around the α-helix minimum disappears. The lowest value of ΔF is still found at $\xi = 15$ Å, but the free energy curve flattens considerably. All configurations characterised by ξ between 10 and 27 Å are within 2 kcal/mol of the minimum, so many frequent conformational changes will be observed along this range of end-to-end distances. When considering even higher temperatures such as 807 K (cyan curve of Fig 5), it is observed that the most stable configuration is found around $\xi = 21$ Å, representing a clearly non-helical, extended structure, and that many other configurations in the range 15 to 25 Å for the end-to-end distance show a value

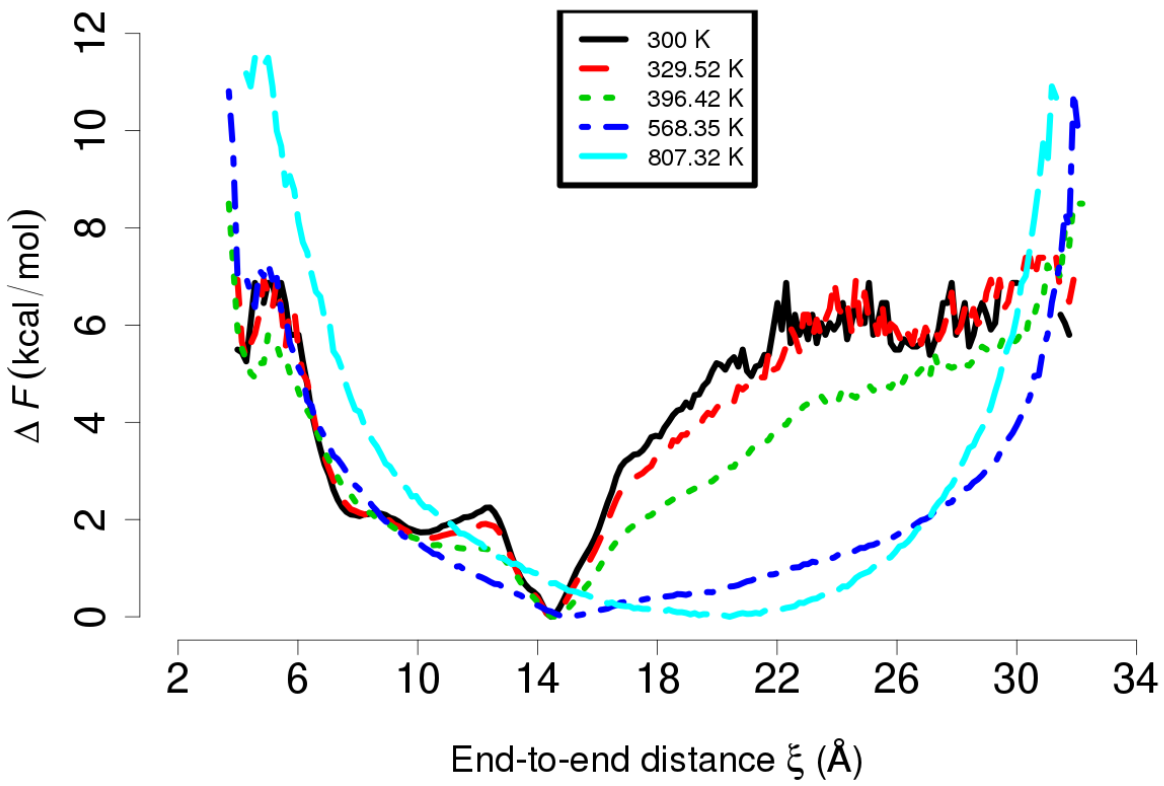


Figure 5: Free energy profile (ΔF in kcal/mol) built using the end-to-end distance between carbonyls' carbon from first and last residue of Ala_{10} (see Figure 1) in implicit GENBORN model. Estimated for 5 temperatures from the PINS simulation of 100 ns long.

of ΔF within 1 kcal/mol.

2D ΔF surfaces: Using 2D FESs it is possible to further characterize the relative stabilities of native and intermediate states. For this it is necessary to introduce two meaningful progression coordinates describing the process of interest (folding in the current case). They were chosen as the end-to-end distance ξ (compactness of a structure) and the degree of α-helical content α. These coordinates were already used previously[21] which allows direct comparison. α quantifies both, the number of hydrogen bonds and the angle between three successive $\alpha-$carbons (see SI for more details).

From the 100 ns MD simulations in implicit solvent a 2D histogram was first built along ξ

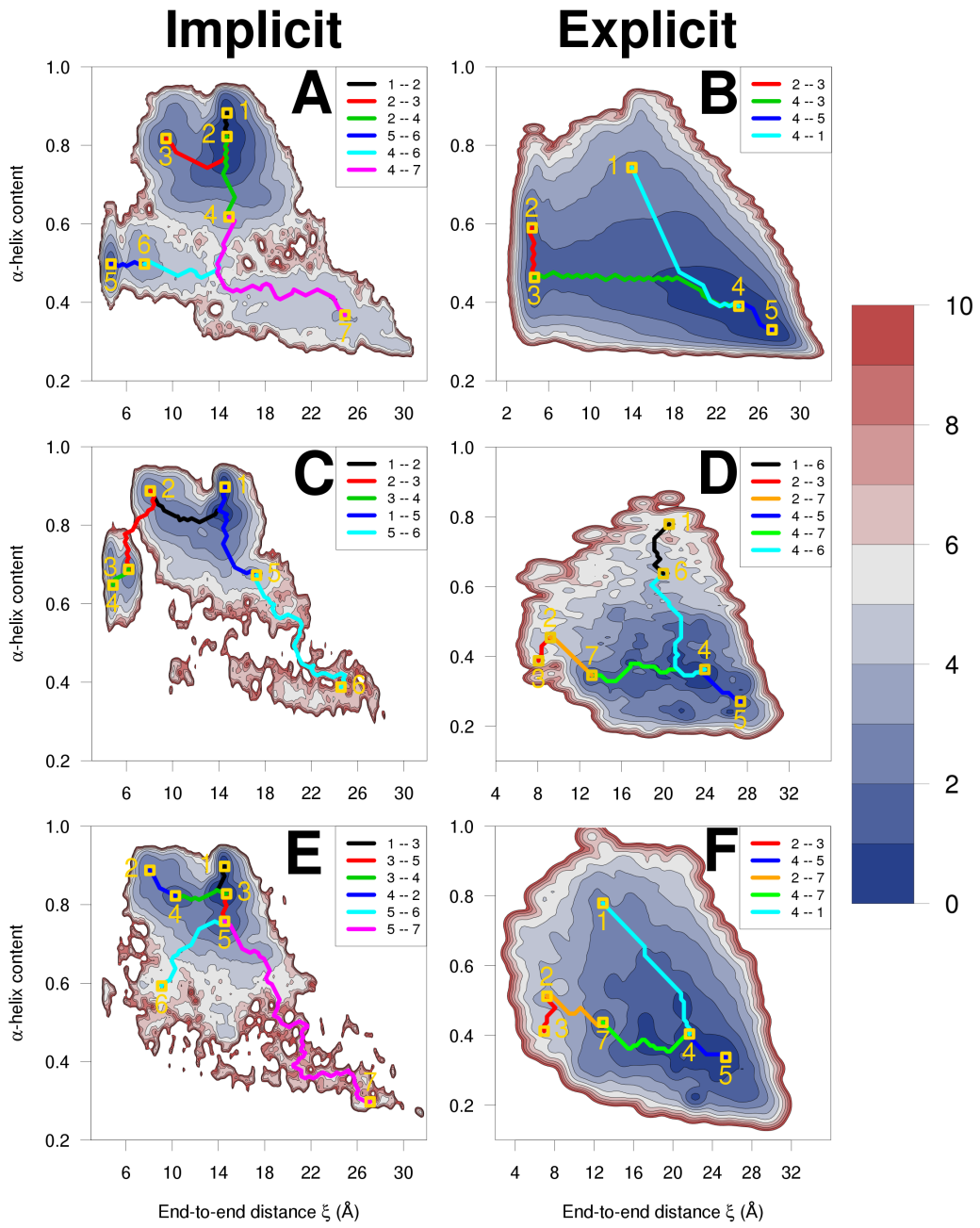


Figure 6: Ala_{10} FES (ΔF in kcal/mol, colour coded) from sampling built along ξ ($x-$axis) and α ($y-$axis) from simulations at 300 K in implicit GENBORN solvent (left column) and in explicit TIP3P water (Right column). Panels A and B for MD, C and D for PT, E and F for PINS. The 2D FES were built using Kernel Density Estimation (KDE), which provides an intrinsic interpolation step. When compared to a standard 2D histogram (see Figure S4 from SI), KDE yields a smoother surface with more connected areas, while the data is still faithfully reproduced.

and α as progression coordinates, see Figure S4 in the SI. However, as the 2d FES is sparsely sampled in the transition regions a multi-variate Kernel Density Estimation (KDE)[39,40] was used for estimating the density matrix, resulting in Figure 6 A. KDE methods provide an accurate density estimation, combined with an intrinsic interpolation step, compensating the poor sampling of some of the bridging regions and higher energy areas (see SI Section 3 for details). Figures 6 C and E are also KDE-interpolated FESs from 100 ns PT and PINS simulations in implicit solvent. Again, the MEP finding method was used to determine paths between important (local) minima.

A 2D projection reveals additional local minima characterized by similar $\xi-$values but differing in α, see for example minima 1 and 2 from Figures 6 A and E found in MD and PINS simulations. For small values of $\xi \sim 5-6$ Å and $\alpha \sim 0.5$, stable configurations appear in Figures 6 A (MD, points 5 and 6), C (PT, points 3 and 4) and E (PINS, point 6). They correspond to β-hairpin states for which most of the hydrogen bonds are formed, but the structure is not helical. The $\alpha-$score contains information about hydrogen bonds and helicity (see SI Section 4), with equal weighting: half of the $\alpha-$score originates from the hydrogen bond term (**hbf**), and the other half from the helix counting (**angf**) term. This explains the amplitude of the score $\alpha = 0.5$ for $\beta-$hairpin states. The most stable extended states are those with ξ between 25 and 28 Å and values of α between 0.3 and 0.4 (point 7, 6, and 7 for MD, PT, PINS, respectively). They are characteristic of fully extended structures without hydrogen bonds and with poor helical content.

An intermediate is found for values of ξ between 14 and 17 Å and for values of α between 0.6 and 0.8 (point 4, 5, and 5 for MD, PT, PINS, respectively). They represent an intermediate between extended and helical states. For MD and PINS this is also an intermediate bridging point to the β-hairpin states (cyan paths on Figs. 6 A and E).

These results can be compared with Figure 6 (bottom) from Ref.[21] where the authors focused on values of $\xi > 12$ Å using US and Adapting Biasing Forces (ABF) MD methods. Here, paths between minima 1 and 7 (MD and PINS) or 1 and 6 (PT) connect the α-helical state to extended ones (green line). The path in Ref.[21] starts from $(\xi, \alpha) = (14, 0.9)$ (α-helix). This agrees with points 1 on the MD, PT, and PINS FESs. The path in Ref.[21] passes through three points defined by coordinates $(16, 0.7)$, $(18, 0.6)$, and $(20, 0.5)$ which agrees with the present PT and PINS (Figures 6C and E) simulations, but not with the MD simulations (Figure 6A). Then the path from Ref.[21] proceeds to an $\alpha-$value of 0.2 between $20 < \xi < 22$ Å and finishes at an extended structure characterised by $(28, 0.15)$. PT and PINS surfaces show this rapid decrease of the α value between $20 < \xi < 22$ Å, but once again not so for MD. The final point found in the three FESs is at significantly larger values of α than the one from Ref.[21] ($0.25 - 0.3$ vs. 0.15). However, the ABF and US simulations used the α and ξ variables in their biasing potential which explains the numerical differences compared to results from unbiased MD, PT and PINS simulations used here.

Secondary structure evolution: Finally, the evolution of the secondary structure of Ala_{10} is considered. For this, the DSSP software[41,42] was used, which classifies each residue of Ala_{10} according to one of the following categories: (i) helices which are divided in three sub categories, **H** for α helices, **G** for 3_{10} helices and **I** for π helices, (ii) $\beta-$structures, divided in 2 subcategories, B for $\beta-$bridges and E for $\beta-$bulges, (iii) strand and turn structures denoted as **S** and **T**, and (iv) no secondary structure for which no letter is assigned. The most recent version of DSSP (2.2.1) was used. It was modified in order to allow direct analysis of CHARMM trajectory files.

Figure 7 shows the secondary structures (the bar counting residues without secondary structure is not displayed), found when analysing the first 100 ns of MD (white), PT (red) and PINS (blue) simulations. Similar distributions are found for each of the methods, confirming

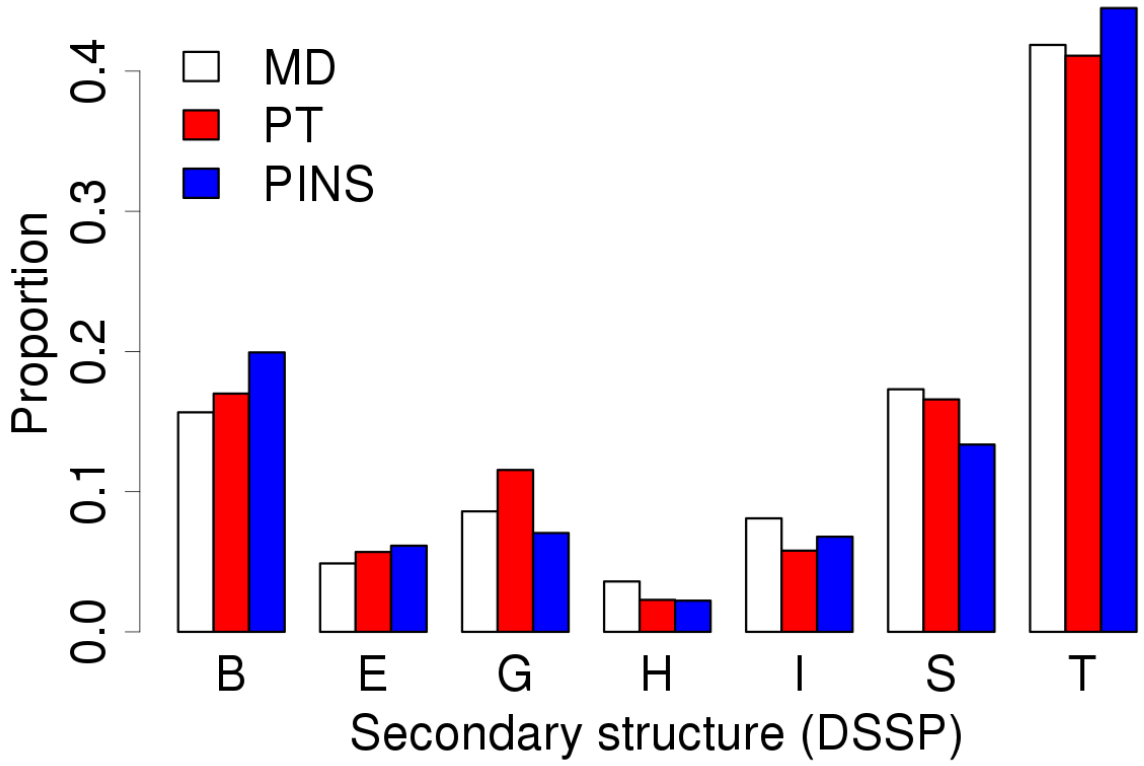


Figure 7: Secondary structure distribution for MD (white) PT (red) and PINS (blue) simulations, for 100 ns in implicit solvent. The structure analysis was carried out with DSSP.[41,42] **H,G,I** are α, 3-10, and $\pi-$helices ; **B,E** for $\beta-$bridges or bulges, and **S,T** for strand and turn structures.

that PINS does not oversample secondary structures and shows differences from MD similar to PT. The influence of the large number of partially folded states from Figure 4 corresponds to $16.0 \leq \xi \leq 28.0$ Å, and is also observed, as they generate many non-helical configurations **B, S, T**.

3.2 Ala_{10} in Explicit Solvent

Next, the performance of PINS was assessed for studying the folding of deca-alanine in explicit solvent using the TIP3P water model.[43] The same starting unfolded configuration, Figure 1 (blue), was used. It was solvated in a cubic box of size 40.5 Å, heated and equili-

brated to a temperature of 300 K for MD (or to the target temperature for PT and PINS) for 100 ps. The timestep was always 1 fs and SHAKE[44] was used for bonds involving H-atoms. For MD, 100 independent simulations were performed, each 40 ns in length, by restarting every 10 ns. Thus the total simulation time is 4 μs (to be compared to 2.5 μs from Ref.[21]). The PT and PINS simulations used 32 temperatures, between 300.00 and 380.87 K and 50 independent simulations, 1 ns each were carried out which yields a total aggregated simulation time of 1.6 μs. The PINS dual-chain structure used 6 temperature blocks $(6,6,6,6,5,3|3,5,6,6,6,6)$. The Particle Mesh Ewald method[45,46] was used, combined with domain decomposition (DOMDEC)[30] for all simulations. The non-bonded energy cutoff-cuton were set to, respectively, 9 Å and 7.5 Å, and the non-bonded lists were built using a heuristic algorithm with a buffer of 11 Å. Those are the values from the official documentation of CHARMM and the DOMDEC module.

ξ-based ΔF profile for MD: Figure 8 shows a ΔF profile obtained after 4 μs of MD simulations in explicit water simulations and can be compared with Figure 3 from Ref.[21] (top and bottom solid lines corresponding to unconstrained ABF and US simulations). It exhibits the same flat profile for $12.0 \leq \xi \leq 28.0$ Å where the free energy difference is always within 1 kcal/mol of the global minimum for $20.0 \leq \xi \leq 25.0$ (22.75 kcal/mol (cf. Ref.[21]) for Fig. 8 but the curve is so flat that this number should be considered with caution).

Figure 8 can also be compared with Figures 6 and 7 (bottom) from Ref.[21], showing 2D PMF and histograms for simulations in explicit water. They show that for simulations starting from linear configurations the disordered states ($12.0 \leq \xi \leq 28.0$ Å) are usually sampled extensively before the system folds to an α-helix.

The 4 configurations shown in Figure 8 are examples of typically and frequently sampled conformations ($5.0 \leq \xi \leq 25.0$) during the simulations. Their free energy is within

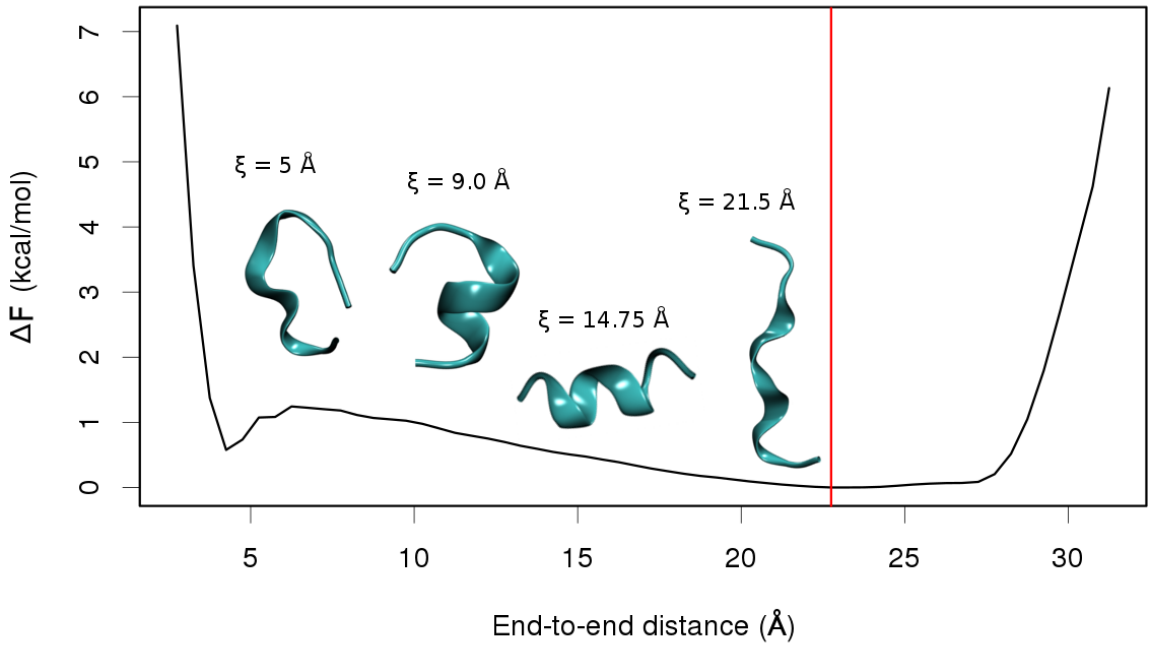


Figure 8: Free Energy as a function of ξ for Ala_{10} in explicit TIP3P water from 4 μs MD simulation. The red vertical line marks the point where $\Delta F = 0$ kcal/mol for $\xi = 22.75$ Å, i.e. the most sampled extended (non-helical) state. The 4 displayed configurations are examples of structures for which $\Delta F \leq 1$ kcal/mol.

$\Delta F \leq 1$ kcal/mol of the global minimum, illustrating the large number of metastable configurations observed in solvent for the Ala_{10} when using standard MD techniques.

2D ΔF surfaces for MD PT and PINS in explicit water: 2d FESs can also be constructed using ξ and α as progression coordinates. Figure 6B uses data from 4 μs of MD simulations. The MEPs detection algorithm was also applied, and five minima were identified: point 1 corresponds to a well-formed α-helix, points 2 and 3 are β-hairpin structures (see Figure 8, structure $\xi = 5$ Å), and points 4 and 5 correspond to two extended structures, characterized by the flat region between $22-28$ Å in Figure 8. Figure S5 from SI Section 5, shows the free energy along the four MEPs highlighted in Figure 6 B. Figures 6D and F are corresponding FESs from 1.6 μs of PT and PINS simulations.

Results for the PT simulations (Figure 6D) are first compared to those obtained using MD (Figure 6B), as the total simulation time of 4 μs for MD is assumed to provide a representative surface. Several differences are noted: Point 1 from PT corresponds to the highest $\alpha-$score for $\xi = 20$ Å, indicative of an incomplete α-helix. Point 6 occurs for PT at a similar $\xi-$value but with a lower $\alpha-$score. Both may correspond to a partially formed helix with strong hydrogen bonding, thus explaining the high $\alpha-$score value. Points 2 and 3 from PT are close to the previously mentioned $\beta-$hairpin MD states although their $\xi-$value is somewhat too high. Point 7 from PT is an intermediate, bridging hairpins states to the extended ones. Points 4 and 5 from PT, representing the extended configurations, are well located and correspond to the MD reference. This most probably means that more sampling would be required for producing meaningful results with PT.

Figure 6F reports results from simulations using PINS. Comparison with the MD results shows that these two methods yield similar FESs. The extended minima (points 4 and 5 of PINS) occur at similar (ξ, α) values, points 2 and 3 for PINS corresponding to the hairpin structures still have a $\xi-$value slightly over-estimated, the point 7 from PINS also connects the hairpin region with extended structures. Point 1, with coordinates (14,0.8), corresponds to an α-helix, a state not clearly observed with PT.

Comparison with previous simulations shows that the path $(1-4-5)$ in Figure 6B is similar to that in Figure 6 (bottom) from Ref.[21]. The estimated free energy is around 3 kcal/mol, compared with 2 kcal/mol from combining Figures S5b and S5c from the SI. The results in Figure 6F (the PINS FES) closely match those from Hazel et al. (Figure 6 (bottom)[21]). Their least free-energy path corresponds to the $(1--4--5)$ MEP from Figure 6F.

In summary, PINS (Figure 6F) provides results similar to what is obtained from MD (Figure 6B), but with a total simulation time of 4 μs for MD, compared to an aggregated 1.6 μs

from PINS. This amounts to a speedup of 2.5 compared to MD, in terms of total simulation time. When comparing PT to PINS, it is shown that for the deca-alanine in explicit water PINS provides converged and accurate results, whereas PT did not.

4 Xe Migration in Myoglobin

Myoglobin (see Figure 10, Left part) is one of the best characterized proteins, both experimentally and by using various types of simulation techniques, and serves as a model system for studying ligand binding, unbinding, and migration.[47] While the pockets accessible to guest atoms (Xenon) and small molecules (O_2, NO, CO) are well characterized from experiment[22–25] and theory/computer simulation[26–29], the stabilization energies in these pockets, the pathways between them and the energy barriers separating them are more debatable. A full characterization of these properties requires direct sampling of the entire free energy surface. A considerable step towards this goal has been the analysis of several trajectories of 90 ns (with 8 CO molecules each) to identify ligand entry pathways from the solvent. Despite such a serious effort no free energy profiles were presented because most transitions between pockets are still rare events and occur only once per trajectory.[27] Since such extensive sampling is computationally expensive, application of enhanced sampling methods is of great interest. As a recent example, Xenon migration in Cytochrome ba_3 oxidase has been found to involve rate coefficients for echange between neighboring sites on the order of 1 s^{-1}.[48] In myoglobin, CO-migration was followed using Laue diffraction and the integrated electron content of the CO-associated features were found to extend over 6 orders of magnitude in time between 10^{-9} and 10^{-3} s with signal decay only starting after 10^{-5} s.[49] These two examples highlight the potentially slow dynamics of guest molecules through the internal cavity network in heme-containing proteins.

The use of Xenon as a guest molecule is motivated by the fact that it diffracts well (54 electrons) in X-ray studies. Furthermore, Xenon only interacts via Van der Waals interactions with its environment which - in addition to its large mass - further slows down diffusion inside and between the cavities which makes the use of rare event sampling techniques mandatory. Hence, Xe diffusion in Mb is a typical example of a topical and complex system for which PINS offers potential advantages over other sampling techniques.

Computational setup: The CHARMM-GUI[50,51] interface was used for generating an initial structure, based on the Protein Data Bank Record 4NXA[52]. The CHARMM c36 FF[34] was used together with CHARMM version c41. A cubic box of 67 Å^3, containing 8596 TIP3P[43] water molecules was used for solvating the system. The non-bonded parameters for the Xe atom were $\varepsilon = -0.423$ kcal/mol and $R_{\mathrm{min,Xe}}/2 = 2.05$ Å, which are comparable to those used in previous work ($\varepsilon = -0.494$ kcal/mol and $R_{\mathrm{min,Xe}}/2 = 2.24$ Å).[53]. The Particle Mesh Ewald[45,46] algorithm is used for treating the non-bonded interactions (cuton–cutoff of respectively 10–12 Å), bonds involving hydrogen were constrained using SHAKE[44], and a timestep of 2 fs was used. The system was heated and equilibrated for 100 ps for MD at a temperature of 300 K. The same heating–equilibrating protocol is used for PINS, which uses 32 replicas, and to each replica a temperature is assigned, distributed between 300.00 and 393.95 K. Simulations were started using a set of configurations in which one Xe atom is initially assigned to one of the 4 experimentally known pockets (Xe1 to Xe4, see Figure 10 right part). For each of the 4 systems MD simulations 100 ns long were carried out. For PINS each replica was simulated for 3.0 ns resulting in a total aggregated simulation time of 96 ns, in order to compare similar amount of data from MD and PINS.

Results: For both, MD and PINS, trajectories were aligned relative to the crystal structure. In order to ascertain the long-time stability of the system, the RMSD relative to the crystal structure was determined and was found to be below 2 Å throughout, confirming the

observed stability and structural integrity of the protein. For the analysis, the distance between the Xe atom and the center of each pocket (Xe1 to Xe4 as found in the X-ray reference structure[22]) was determined for all configurations and all trajectories which provides a first clustering. Then the Xe atom in a particular snapshot was assigned to the pocket for which the distance between the current location and the center of each pocket is lowest. This yields a discretized trajectory. From this it is possible to estimate the relative occupation (q_{PINS} or q_{MD}) of each pocket Xei along the trajectory of interest. In order to compare the relative efficiency of sampling, a boost factor R of PINS over MD is defined as $R = \frac{q_{\mathrm{PINS}}}{q_{\mathrm{MD}}}$. Figure 9 shows R for the Xe atom in any of the 4 different starting positions. As sampling of the protein interior is of concern here, events in which the Xe atom remains in the initial pocket are not considered. Furthermore, situations in which Xe escapes to the solvent are also discarded. Red bars with a "∞ symbol" correspond to transitions which are not sampled at all using conventional MD (i.e. $q_{\mathrm{MD}} = 0$), and for which PINS finds transitions. The results show that PINS increases the sampling efficiency by a factor of 2 to 10. Furthermore, for 3 of the 4 simulations one transition which was not sampled using MD is sampled with PINS (pockets Xe4, Xe2, Xe3 when starting from Xe2, Xe3, Xe4, respectively). Hence, overall PINS samples transitions more effectively.

As R not only reports on the number of transitions but also on the actual occupation of particular pockets, the transition count matrices were also determined (Table 2). These matrices were built using the full data from the MD and PINS simulations with snapshots taken every 1 ps. With PINS more frequent transitions from or to pocket Xe3 are found, which is rarely and poorly sampled with MD. Another interesting observation is that PINS simulations allow direct Xe1$\leftrightarrow$Xe4 transitions. Analysis of the 300 K to 350 K replicas supports that this only occurs for replicas run at higher temperatures. Finally, it is also noticed that the transition matrices are near-symmetric.

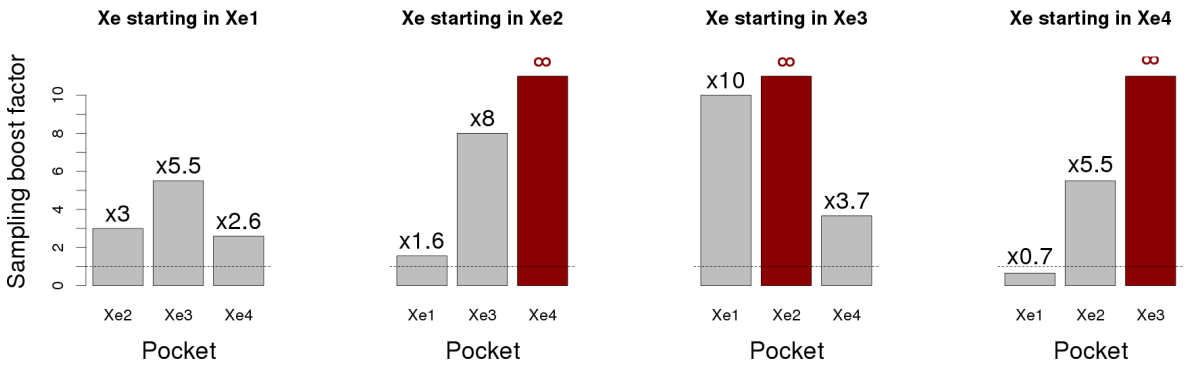


Figure 9: Ratio between the relative occupation of each pocket ($y-$axis) defined as $R = \frac{q_{\text{PINS}}}{q_{\text{MD}}}$. Values of $R > 1$ mean that PINS is more efficient than MD. $R = \infty$ denotes situations where the corresponding pocket was not sampled by MD simulations starting from a given initial pocket.

Table 2: Transition matrix estimated for MD (left) and PINS (right), using 4 simulations of respectively 100 and 96 ns, each starting in one of the four pockets. The transition boost provided by PINS is evident, and the effect of high temperature replicas allows for example direct jumps Xe1↔Xe4, unobserved with MD.

	MD				PINS			
	Xe1	Xe2	Xe3	Xe4	Xe1	Xe2	Xe3	Xe4
Xe1	·	66	8	0	·	240	30	610
Xe2	68	·	38	66	234	·	40	112
Xe3	12	40	·	0	34	40	·	2
Xe4	0	54	0	·	620	94	4	·

The efficiency of PINS as reported in Figure 9, and especially its capability of sampling low probability (transition) regions connecting the pockets, can also be directly visualised. For that, coordinates of the Xe atom are extracted, and the normalized probability distribution $\rho(x, y, z)$ at a given point (x, y, z) is evaluated on a 3D grid with resolution 0.5 Å, see Figure 10 for simulations with Xe initially in Xe4. The densities shown are for $\rho(x, y, z) = 10^{-5}$. This analysis confirms the results from Figure 9 and Table 2, i.e. that the Xe3 pocket is poorly sampled by MD, whereas PINS explores this region of the protein. It is also demonstrated that PINS samples the transition region more extensively, e.g. the Xe4↔Xe2 and Xe1↔Xe2 transitions (see upper- and bottom-right parts of the isosurfaces).

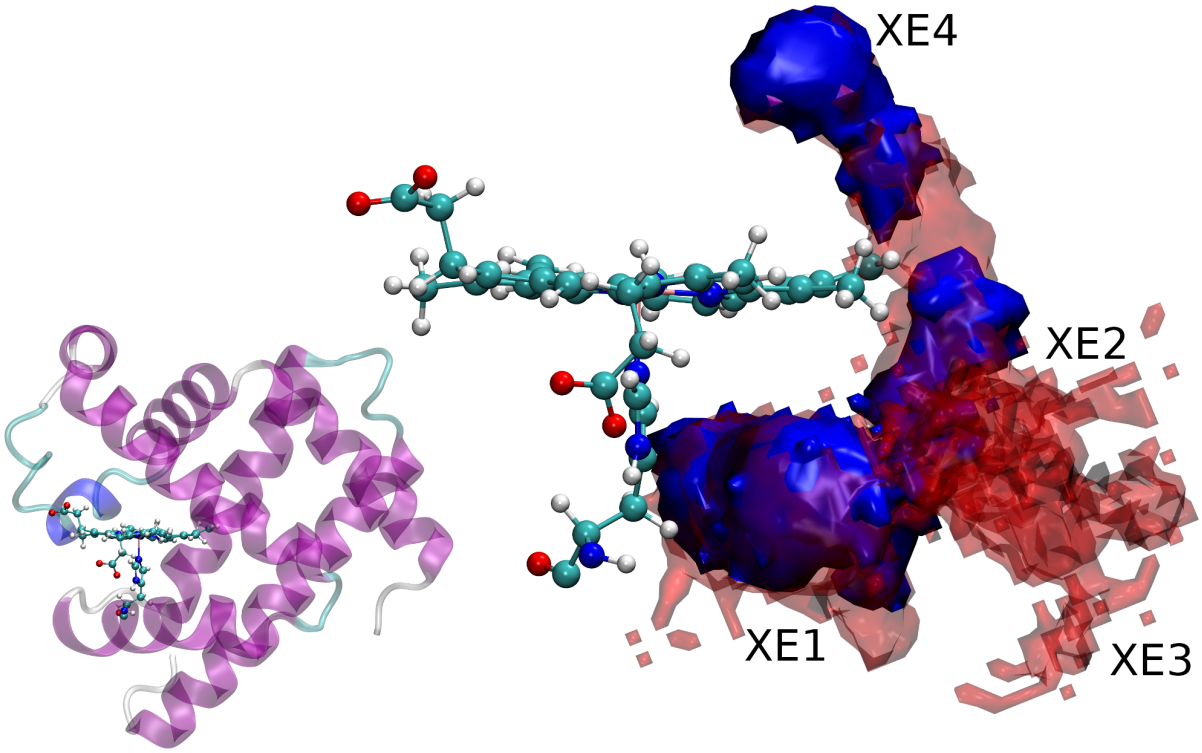


Figure 10: Left: Myoglobin with its heme functional group. Color code for the protein secondary structure is purple and blue for α and 3-10 helices, cyan and white for turn and coil, respectively. Right: Isosurface of normalised probabilities ($\rho = 10^{-5}$) to find the Xe atom at a given grid point, and definition of the 4 Xe pockets. Blue surface for MD, red for PINS. Built using the 100 ns and 96 ns long simulations. For simulations starting in pocket Xe4. PINS samples pocket Xe3 not explored with conventional MD. The transition channels Xe4 $\leftrightarrow$Xe2 and Xe1$\leftrightarrow$Xe2 are also more widely sampled when using PINS than with standard MD.

Table 3: Stabilization free energy ΔF_{stab} (kcal/mol) for the 4 Xe pockets, estimated for the MD and PINS simulations, and compared with the Implicit Ligand Sampling results from Ref.[53]. The 95 % confidence interval was estimated using bootstrapping, dividing data in 10 sets.

	ΔF_{stab} (kcal/mol)			
	MD	PINS	Cohen et al.[53]	Exp.[53]
Xe1	–4.41 ± 0.10	–6.19 ± 0.09	–6.4	–5.1
Xe2	–2.56 ± 0.35	–4.58 ± 0.34	–5.2	–4.5
Xe3	–3.69 ± 0.28	–5.58 ± 0.07	–5.1	–4.6
Xe4	–4.31 ± 0.25	–5.59 ± 0.24	–5.5	–4.4

From the probability distribution functions $P(x, y, z)$ the relative stabilization energies of Xe in the 4 pockets can be determined and are summarized in Table 3. The MD results are obtained from inverting $P(x, y, z) \propto \exp(-\beta \Delta F_{\text{stab}}(x, y, z))$ and for PINS the post-processing procedure, as described in the Methods section, was applied (see Equation 8). The PINS results compare quite favourably with those from an earlier study[53] (values also reported in Table 3 for comparison) based on a 5 ns simulation, with relative absolute differences ranging between 0.1 and 0.6 kcal/mol. Binding free energies from the present MD simulations are somewhat too low which may be related to under-sampling in the MD simulations although the aggregate simulation time is 400 ns (i.e. 100 ns per initial Xe placement). It should be recalled that implicit ligand sampling[53] carries out simulations without the guest molecule present (i.e. the empty protein) and coupling between protein and ligand dynamics is absent. Also, there is little guarantee that large energy barriers will be sampled accurately which leads to overestimated energy barriers. Given the considerably larger amount of data from the present simulations (aggregate of 400 ns for PINS) compared to the previous study[53] (5 ns of MD with Implicit Ligand Sampling), it is expected that the present stabilization energies ΔF_{stab} are more representative.

Finally by extracting the free energy along the path connecting two pockets, it is also possible to estimate the transition barrier free energies from the PINS simulations. For the Xe1↔Xe2 transition the barrier height is estimated to be 4.4 kcal/mol and for the Xe2↔Xe4 transition it is 3.9 kcal/mol, corresponding to typical transition times on the sub-nanosecond time scale according to transition state theory. This is confirmed when considering the MD transition matrix from Table 2 (left), where 120 to 138 transitions are observed for Xe1↔Xe2 and Xe2↔Xe4 during 100 ns. These results were confirmed for the Xe2↔Xe4 transition by using umbrella sampling simulations[7]. The progression coordinate for this transition was the distance between the center of gravity of the Phe138 carbon atoms and the Xe-atom. This coordinate was found to be useful in previous simulations of transition paths for CO between

these two pockets for which a barrier height of 6.0 kcal/mol or larger was found depending on the initial protein structure.[54] For Xe, which is expected to interact less strongly with the protein environment, a barrier height of 4.5 ± 0.4 kcal/mol was obtained using WHAM[55]. This agrees favourably with the estimate of 3.9 kcal/mol from PINS simulations, which further validates the implementation and analysis protocol.

5 Summary

The present work describes the implementation, analysis and application of PINS to two systems of different complexity: the folding of deca-alanine in implicit and explicit solvent and Xenon migration in Myoglobin. For deca-alanine, folding to the $\alpha-$helical structure in implicit solvent was found to occur more rapidly by one order of magnitude with PINS and PT compared to MD simulations. For this system, PT and PINS perform almost equally. The analysis of the folded state is consistent with previous work[21,32] which yielded values for the end-to-end distance of $\xi = 14.5 - 14.7$ Å.

A more challenging application was the study of the folding process in explicit solvent. The 2D FESs from MD, PINS and PT simulations suggest that PINS is capable of characterizing the important states and transitions between them from an aggregate of 1.6 μs compared to 4 μs required for MD simulations. The overall topography of the FESs confirms previously published observations,[21] where an extended flat region was found for $12.0 \leq \xi \leq 28.0$ Å, caused by numerous non-compact structures.

The third application considered Xenon atom migration in the internal cavities of Mb. PINS extensively samples the 4 experimentally known Xe pockets and the transition regions between them. This contrasts with MD simulations which provide little information about

barrier crossings for comparable simulation times. PINS yields estimates for Xe-binding free energy comparable to alternative methods such as implicit ligand sampling.[53] The height of the Xe4↔Xe2 transition barrier was estimated to be ≈ 3.9 kcal/mol from the PINS-unbiasing procedure and was confirmed using umbrella sampling simulations.

Finally, it is important to point out that PINS could be further generalised. In Equations 4 and 5 the probability density $\rho(\mathbf{X})$ and the partition function Z were defined by only considering the potential energy $V(\mathbf{X})$. Instead of $V = E_{\text{pot}}$ it is possible to use a classical Hamiltonian $\mathcal{H} = E_{\text{kin}} + E_{\text{pot}}$ where E_{kin} and E_{pot} are the kinetic and potential energy, respectively. Then it is possible to define K Hamiltonians instead of K temperatures for the replicas, each Hamiltonian thus containing e.g. different biasing potentials. The two Equations 4 - 5 are still valid, so from the algorithmic point of view the only necessary modification is to broadcast the total Hamiltonian between the replicas instead of the potential energy.

Acknowledgement

This work was supported by the Swiss National Science Foundation through grants 200021-7117810, the NCCR MUST (to MM). JDD wishes to acknowledge support from the National Science Foundation award DMS-1317199, from the DARPA EQUiPS award W911NF-15-2-0122, and for continuing discussions with P. Dupuis.

The authors declare no competing financial interest.

Associated Content

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.xxxxxxx :

- Supplementary Figures and Tables for Sections 3 and 4
- Details on analysis methodology with extra references

References

(1) Metropolis, N.; Ulam, S. *J. Am. Stat. Assoc.* **1949**, *44*, 335–341.

(2) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

(3) Hastings, W. K. *Biometrika* **1970**, *57*, 97–109.

(4) Swendsen, R. H.; Wang, J.-S. *Phys. Rev. Lett.* **1986**, *57*, 2607–2609.

(5) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604–1608.

(6) Earl, D. J.; Deem, M. W. *PCCP* **2005**, *7*, 3910–3916.

(7) Torrie, G. M.; Valleau, J. P. *J. Chem. Phys.* **1977**, *23*, 187–199.

(8) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci.* **2002**, *99*, 12562–12566.

(9) Kofke, D. A. *J. Chem. Phys.* **2002**, *117*, 6911–6914.

(10) Plattner, N.; Doll, J. D.; Dupuis, P.; Wang, H.; Liu, Y.; Gubernatis, J. E. *J. Chem. Phys.* **2011**, *135*, 134111.

(11) Doll, J. D.; Plattner, N.; Freeman, D. L.; Liu, Y.; Dupuis, P. *J. Chem. Phys.* **2012**, *137*, 204112.

(12) Dupuis, P.; Liu, Y.; Plattner, N.; Doll, J. D. *Multiscale Model. Simul.* **2012**, *10*, 986–1022.

(13) Plattner, N.; Doll, J. D.; Meuwly, M. *J. Chem. Theory Comput.* **2013**, *9*, 4215–4224.

(14) Hénin, J.; Chipot, C. *J. Chem. Phys.* **2004**, *121*, 2904–2914.

(15) Ozer, G.; Keyes, T.; Quirk, S.; Hernandez, R. *J. Chem. Phys.* **2014**, *141*, 064101.

(16) Comer, J.; Phillips, J. C.; Schulten, K.; Chipot, C. *J. Chem. Theory Comput.* **2014**, *10*, 5276–5285.

(17) Zhang, T.; Nguyen, P. H.; Nasica-Labouze, J.; Mu, Y.; Derreumaux, P. *J. Phys. Chem. B* **2015**, *119*, 6941–6951.

(18) Apostolakis, J.; Ferrara, P.; Caflisch, A. *J. Chem. Phys.* **1999**, *110*, 2099–2108.

(19) Uribe, L.; Gauss, J.; Diezemann, G. *J. Phys. Chem. B* **2015**, *119*, 8313–8320.

(20) Lin, Z.; Riniker, S.; Gunsteren, W. F. v. *J. Chem. Theory Comput.* **2013**, *9*, 1328–1333.

(21) Hazel, A.; Chipot, C.; Gumbart, J. C. *J. Chem. Theory Comput.* **2014**, *10*, 2836–2844.

(22) Tilton, R.; Kuntz, I. D.; Petsko, G. A. *Biochem.* **1984**, *23*, 2849–2857.

(23) Olson, J. S.; Phillips, G. N. *J. Biol. Chem.* **1996**, *271*, 17593–17596.

(24) Scott, E. E.; Gibson, Q. H.; Olson, J. S. *J. Biol. Chem.* **2001**, *276*, 5177–5188.

(25) Schotte, F.; Lim, M.; Jackson, A.; Smirnov, V.; Soman, J.; Olson, J.; Phillips, G.; Wulff, M.; P., A. *Science* **2003**, *300*, 1944–1947.

(26) Elber, R.; Karplus, M. *J. Am. Chem. Soc.* **1990**, *112*, 9161–9175.

(27) Ruscio, J. Z.; Kumar, D.; Shukla, M.; Prisant, M. G.; Murali, T. M.; Onufriev, A. V. *Proc. Natl. Acad. Sci.* **2008**, *105*, 9204–9209.

(28) Bossa, C.; Anselmi, M.; Roccatano, D.; Amadei, A.; Vallone, B.; Brunori, M.; Di Nola, A. *Biophysical J.* **2004**, *86*, 3855–3862.

(29) Plattner, N.; Doll, J. D.; Meuwly, M. *J. Chem. Phys.* **2010**, *133*, 044506.

(30) Hynninen, A.-P.; Crowley, M. F. *J. Comput. Chem.* **2014**, *35*, 406–413.

(31) Esque, J.; Cecchini, M. *J. Phys. Chem. B* **2015**, *119*, 5194–5207.

(32) Park, S.; Khalili-Araghi, F.; Tajkhorshid, E.; Schulten, K. *J. Chem. Phys.* **2003**, *119*, 3559–3566.

(33) Brooks, B.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545–1614.

(34) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D., Jr. *J. Chem. Theo. Comp.* **2012**, *8*, 3257–3273.

(35) Patriksson, A.; van der Spoel, D. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2073–2077.

(36) MacQueen, J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. Berkeley, Calif., 1967; pp 281–297.

(37) Lloyd, S. *IEEE T. Inform. Theory* **1982**, *28*, 129–137.

(38) J. A. Hartigan, M. A. W. *J. Roy. Stat. Soc. C-App.* **1979**, *28*, 100–108.

(39) Rosenblatt, M. *Ann. Math. Statist.* **1956**, *27*, 832–837.

(40) Parzen, E. *Ann. Math. Statist.* **1962**, *33*, 1065–1076.

(41) Kabsch, W.; Sander, C. *Biopol.* **1983**, *22*, 2577–2637.

(42) Touw, W. G.; Baakman, C.; Black, J.; te Beek, T. A. H.; Krieger, E.; Joosten, R. P.; Vriend, G. *Nucleic Acids Res.* **2014**, D364–D368.

(43) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(44) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327 – 341.

(45) Hockney, R.; Eastwood, J. *Computer Simulation Using Particles*; CRC Press, 1988.

(46) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(47) Frauenfelder, H.; McMahon, B. H.; Fenimore, P. W. *Proc. Natl. Acad. Sci.* **2003**, *100*, 8615–8617.

(48) Luna, V. M.; Fee, J. A.; Deniz, A. A.; Stout, C. D. *Biochem.* **2012**, *51*, 4669–4676.

(49) Srajer, V.; Ren, Z.; Teng, T.; Schmidt, M.; Ursby, T.; Bourgeois, D.; Pradervand, C.; Schildkamp, W.; Wulff, M.; Moffat, K. *Biochem.* **2001**, *40*, 13802–13815.

(50) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. *J. Comput. Chem.* **2008**, *29*, 1859–1865.

(51) Lee, J.; Cheng, X.; Swails, J. M.; Yeom, M. S.; Eastman, P. K.; Lemkul, J. A.; Wei, S.; Buckner, J.; Jeong, J. C.; Qi, Y.; Jo, S.; Pande, V. S.; Case, D. A.; Brooks, C. L.; MacKerell, A. D.; Klauda, J. B.; Im, W. *J. Chem. Theory Comput.* **2016**, *12*, 405–413.

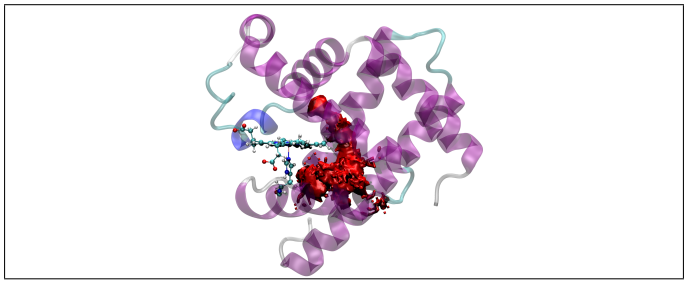
(52) Abraini, J. H.; Marassio, G.; David, H. N.; Vallone, B.; Prangé, T.; Colloc'h, N. *Anesthesiology* **2014**, *121*, 1018–1027.

(53) Cohen, J.; Arkhipov, A.; Braun, R.; Schulten, K. *Biophysical J.* **2006**, *91*, 1844–1857.

(54) Plattner, N.; Meuwly, M. *Biophysical J.* **2012**, *102*, 333–341.

(55) Grossfield, Alan, WHAM: the weighted histogram analysis method, version 2.0.9. Accessed on 8th June 2016; `http://membrane.urmc.rochester.edu/content/wham`.

Graphical TOC Entry



Supporting information for: Partial Infinite Swapping: Implementation and Application to alanine-decapeptide and Myoglobin in the Gas Phase and in Solution

Florent Hédin,[†] Nuria Plattner,[‡] J. D. Doll,[¶] and Markus Meuwly[*,†,¶]

Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland., Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, D-14195 Berlin., and Department of Chemistry, Brown University, Providence, Rhode Island 02912, USA.

E-mail: m.meuwly@unibas.ch

1 RMSD analysis for $(Ala)_{10}$

Figure S1 illustrates the RMSD fluctuations during folding of Ala_{10}, for a 100 ns long MD simulation. The GENBORN implicit solvent model was used. The reference structure with RMSD = 0 is the folded $\alpha-$helix, cf. Figure 1 from the main text. After 15 ns one can observe a majority of quasi folded states, with a RMSD of ≈ 2 Å.

[*]To whom correspondence should be addressed

[†]Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland.

[‡]Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, D-14195 Berlin.

[¶]Department of Chemistry, Brown University, Providence, Rhode Island 02912, USA.

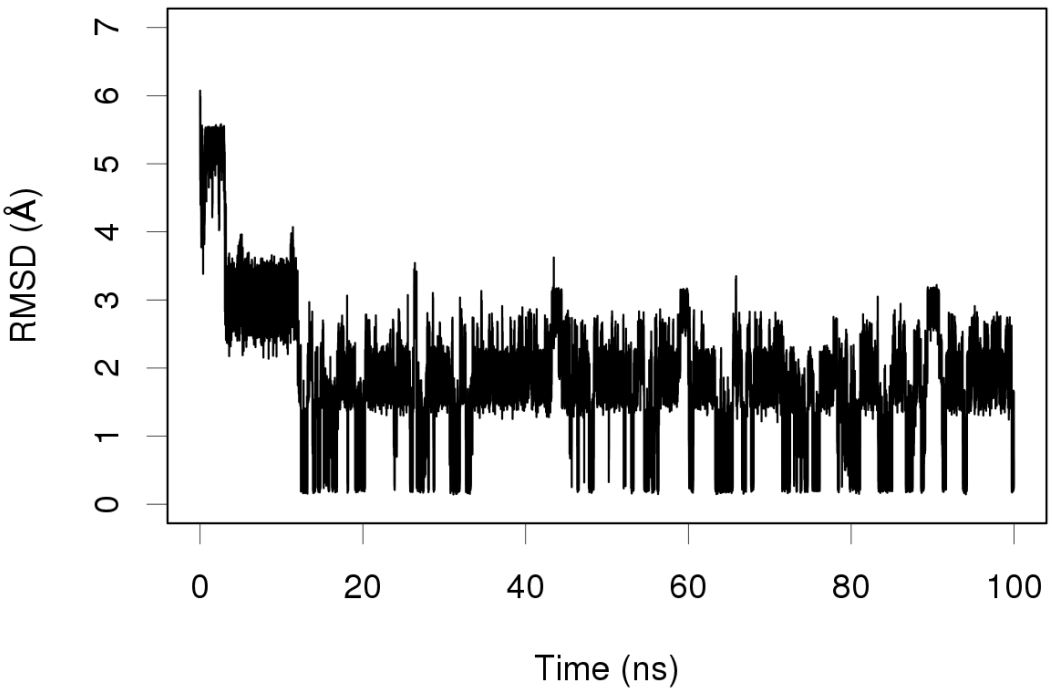


Figure S1: RMSD fluctuations observed for folding of (Ala_{10}), 100 ns of MD with GENBORN solvent. The reference structure is the α-helix from Figure 1 from the main text.

2 K-means clustering on the RMSD distributions

Table S1: RMSD clustering of the combined MD, PT, and PINS structures from Figure 2 from the main text, around 6 centres. After obtaining the cluster centres, each structure is assigned to the closest of the 6 centres.

centres (Å)	pop. MD (%)	pop. PT (%)	pop. PINS (%)
0.3	15.5	38.7	22.8
1.2	10.1	3.7	4.3
1.9	53.7	40.5	52.9
2.7	15.0	3.0	5.7
3.5	2.7	11.8	14.1
4.8	3.0	2.3	0.2

K-means clustering[S1–S3] is a straightforward method for characterizing the diversity of sampling based on a progression coordinate, which is the RMSD in the present case. Figures S2 – S3 illustrate the procedure used for choosing the number of clusters for performing the clustering.

Figure S2 estimates which proportion $\frac{\sum_{i=1}^{K}\sigma^2(X_i)}{\sigma^2(X)}$ of the total variance of the RMSD dataset (denoted as X in the following equations) is reproduced when considering K clusters: for $K \to +\infty$ the total variance of the dataset is described. Here, the sum of the variance around each cluster is $\sigma^2(X_i)$ and the total variance of the original dataset is $\sigma^2(X)$. It is commonly observed that at some point increasing the number of clusters does not appreciably improve the variance description, and the value of K after this point is considered an acceptable value of k for the k-means clustering. The detection of such an inflexion angle, is referred to as "The Elbow Method".[S4] Although this inflexion point may be challenging to locate in some cases[S5] for the data analyzed here those points are easily found as $k = 6$ for MD and $k = 4$ for PT and PINS.

Figure S3 counts the sum of squares of the RMSD X within each group defined around

a cluster (WSS): this time for $K \to +\infty$ this WSS tends to 0. It is estimated according to

$$WSS = \sum_{n=1}^{K} \sum_{p=1}^{P} (X_p - X_n)^2 \tag{1}$$

where K is the number of clusters allowed for the k-means clustering, P is the total number of X points around a cluster k, X_p the RMSD value of point p and X_k is the RMSD value of the centre of the cluster k. The results from the previous Figure S2 are confirmed by Figure S3, i.e. values of $k = 6$ and $k = 4$ seem to be a reasonable choice for performing the k-means.

For those reasons it was decided to use $k = 6$ in all k-means analyses performed for the present study (see Table 4 from the main text and Table S1). Indeed this value of 6 appears to be required for describing well the RMSD distribution of the MD dataset, to which PT and PINS are compared, so it is practical to use the same k for the three methods.

But as the previous plots suggested $k = 4$ for PT and PINS, one could argue that providing $k = 6$ for those two methods adds an unnecessary number of clusters which may reduce the statistical significance of the results. Table S2 shows results of a k-means clustering with $k = 4$: when compared to Table 4 from the main text it is noticed that the 4 most populated centres are close in RMSD and then it could be concluded that using $k = 6$ instead of $k = 4$ for allowing a precise comparison with MD does not invalidate the discussion from the Applications section.

Table S2: Results of a k-means clustering of the RMSD data from Figure 2 from the main text, for PT and PINS with $k = 4$ as suggested by Figures S2 – S3. Results are similar to those with $k = 6$ in Table 4 from the main text.

centres (Å)	pop. (%)	centres (Å)	pop. (%)
0.3	40.0	0.3	22.8
1.9	43.7	1.2	4.4
3.4	10.7	1.9	55.6
4.2	5.6	3.4	17.2

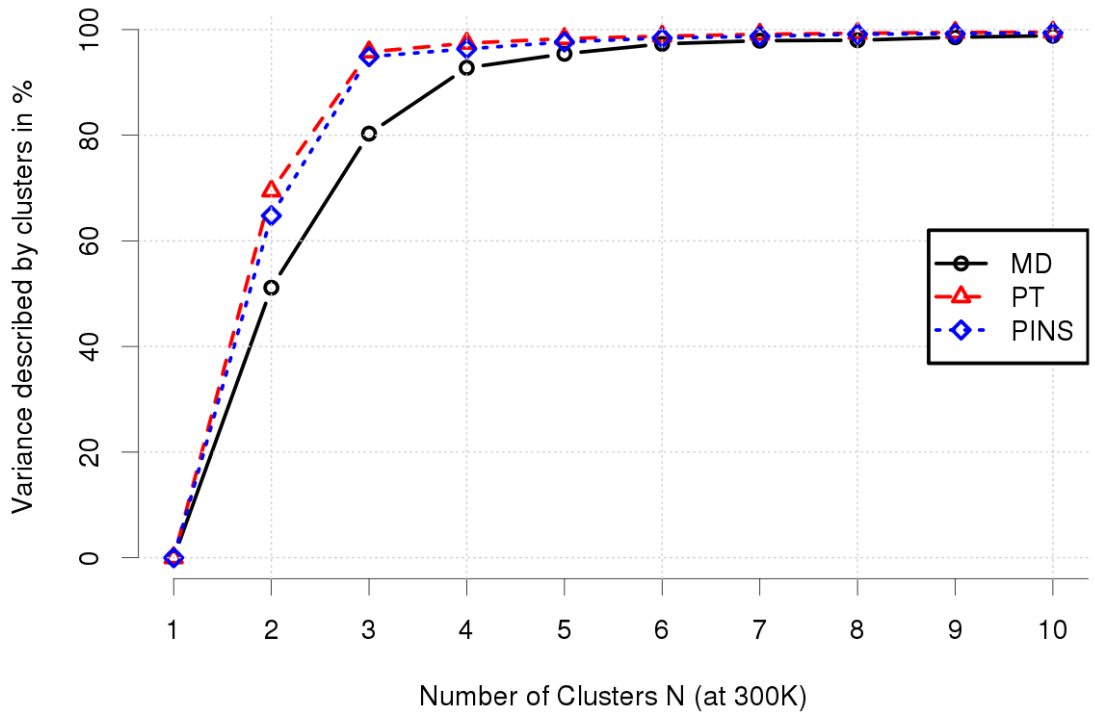


Figure S2: Proportion of the variance of the RMSD dataset described by N clusters, for the 20 ns long simulations from Figure 2 from the main text. The asymptotic behaviour for $K \geq k$ indicates that k clusters are apparently enough for describing accurately the RMSD, with $k = 6$ clusters for MD, and $k = 4$ clusters for PT/PINS.

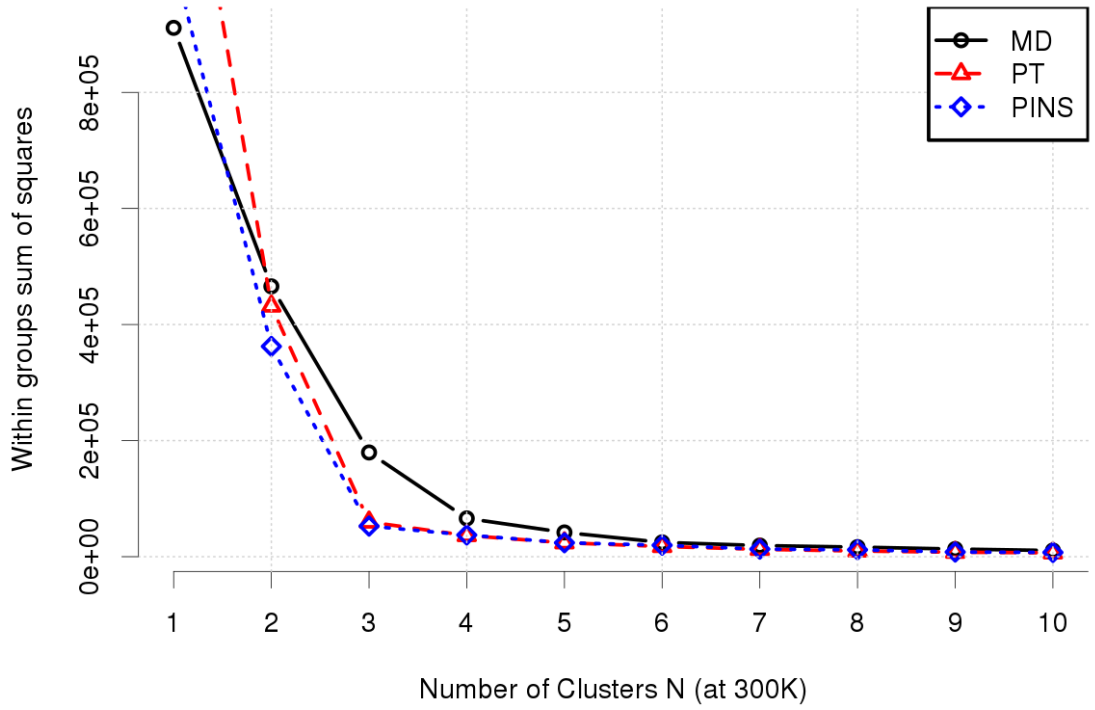


Figure S3: Evolution of the total WSS (Within groups Sum of Squares) when increasing the number of clusters k for the k-means clustering, applied to the 100 ns long implicit solvent simulations from Figure 2 from the main text. Values of $k = 6$ and $k = 4$, for respectively MD and PT/PINS, look reasonable, as adding more clusters does not reduce the overall WSS.

3 2D density estimation using KDE, and MEPs finding method

The R[S6] package **gdistance**[S7] provides classes and functions to calculate various distance measures and routes in heterogeneous geographic spaces represented as grids, but it is possible to apply the algorithm to any surface. The `shortestPath()` function was used for finding the Minimum Energy Path (MEP), based on the Dijkstra[S8] algorithm.

The Dijkstra algorithm expects no discontinuity on the grid when searching for a path: when building a surface using a standard 2D Histogram ($\Delta F(\xi, \alpha) = -RT \ln(\rho(\xi, \alpha))$, see Figure S4 in red for an example with deca-alanine) the transition areas are sometimes sampled poorly, and the application of the path finding algorithm may be challenging. For this reason, Kernel Density Estimation methods[S9,S10] were used for providing a trustful interpolation of the ΔF values at poorly sampled grid areas (see Figure S4 in black). Figure 6 A to F from the main text are examples of such interpolated KDE surfaces.

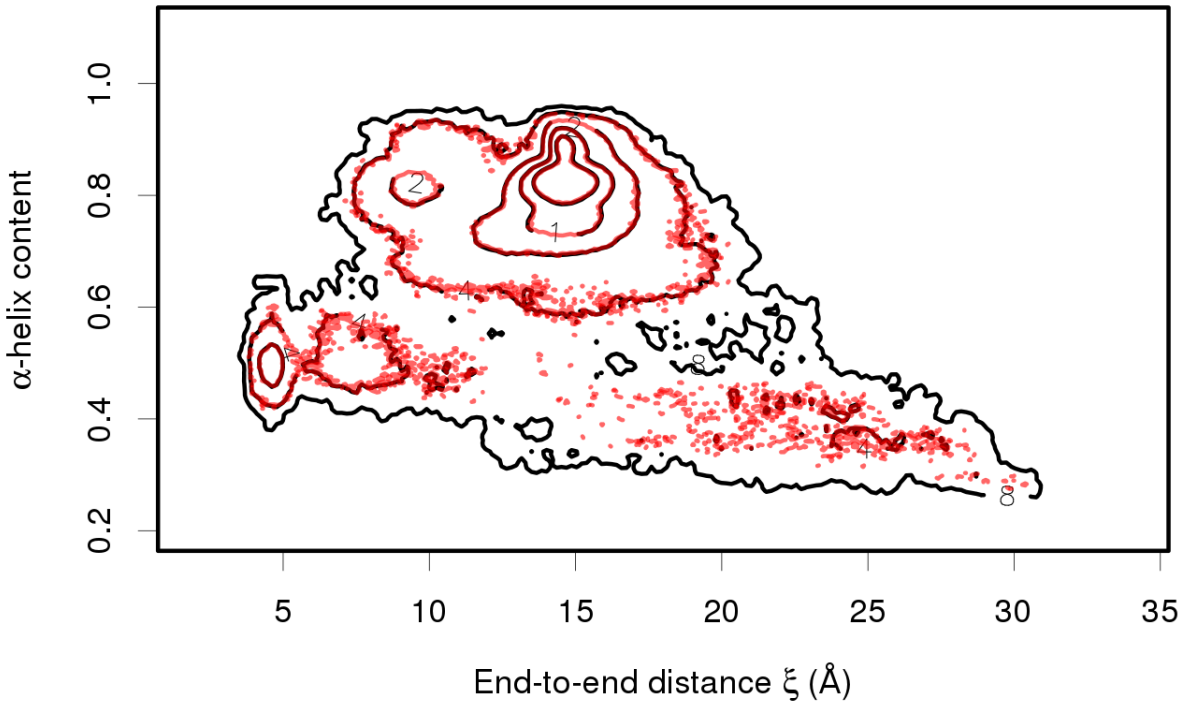


Figure S4: Free Energy contour plots built using a 2D Histogram (red) or on a 2Dim Kernel Density Estimation (black), using the end-to-end distance ($x-$axis) and the α-helix content ($y-$axis), for Ala_{10} MD simulations at 300K in implicit GENBORN solvent. The sparsity of the contour when using standard histograms (red) justifies the use of the KDE method for interpolating results (black), as one can see on Figure 6 A from the main text.

4 Calculation of the α-helical content

In order to build meaningful 2D free energy surfaces for deca-alanine, it is required to use as coordinates two properties which are easy to map to real numbers. It was decided to use the end-to-end distance ξ between carbonyls' carbons from the first and last residue (see Figure 1 from the main text), and a helicity score α detailled below. Those two coordinates were already successfully used for investigating the folding of the deca-alanine by Hénin et al.[S11] and implemented in the **colvars** package.

The α-helical content for the $N+1$ residues N_0 to N_0+N is calculated using the formula:

$$\alpha = \frac{1}{2(N-2)} \sum_{n=N_0}^{N_0+N-2} \text{angf}\left(\text{C}_\alpha^{(n)}, \text{C}_\alpha^{(n+1)}, \text{C}_\alpha^{(n+2)}\right) + \frac{1}{2(N-4)} \sum_{n=N_0}^{N_0+N-4} \text{hbf}\left(\text{O}^{(n)}, \text{N}^{(n+4)}\right) \tag{2}$$

where the scoring function angf(...) for the $\text{C}_\alpha - \text{C}_\alpha - \text{C}_\alpha$ angle is defined as:

$$\text{angf}\left(\text{C}_\alpha^{(n)}, \text{C}_\alpha^{(n+1)}, \text{C}_\alpha^{(n+2)}\right) = \frac{1 - \left(\theta(\text{C}_\alpha^{(n)}, \text{C}_\alpha^{(n+1)}, \text{C}_\alpha^{(n+2)}) - \theta_0\right)^2 / \left(\Delta\theta_{\text{tol}}\right)^2}{1 - \left(\theta(\text{C}_\alpha^{(n)}, \text{C}_\alpha^{(n+1)}, \text{C}_\alpha^{(n+2)}) - \theta_0\right)^4 / \left(\Delta\theta_{\text{tol}}\right)^4} \tag{3}$$

and the scoring function for the hydrogen bonding, hbf(...), is defined using:

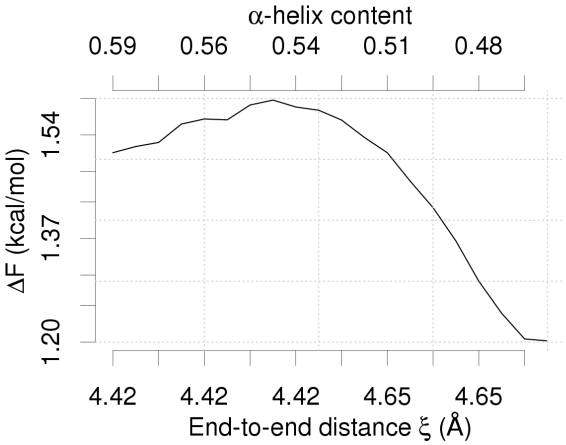
$$\text{hbf}\left(\text{O}^{(n)}, \text{N}^{(n+4)}\right) = \sum_{i \in \text{O}^{(n)}} \sum_{j \in \text{N}^{(n+4)}} \frac{1 - \left(|\mathbf{x}_i - \mathbf{x}_j| \ / \ hb_{cut}\right)^6}{1 - \left(|\mathbf{x}_i - \mathbf{x}_j| \ / \ hb_{cut}\right)^8} \tag{4}$$

where $\theta_0 = 88°$ and $\Delta\theta_{\text{tol}} = 15°$ are respectively reference and tolerance values of the $\text{C}_\alpha - \text{C}_\alpha - \text{C}_\alpha$ angle ; and $hb_{cut} = 3.3$ Å is the cutoff value under which a hydrogen bond is defined.

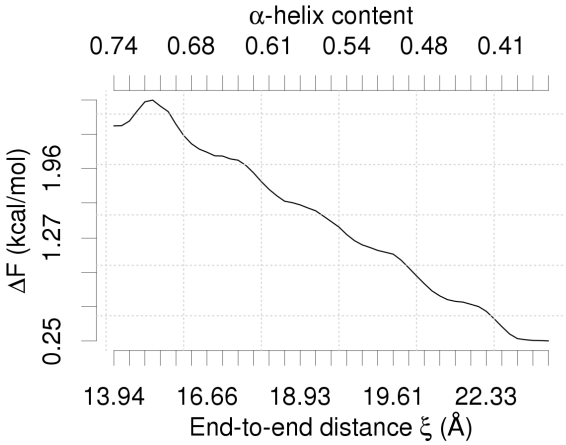
The final value of α maps to a real number between 0 and 1. When combined to the ξ end-to-end distance, one can build meaningful 2D surfaces, as seen in Figure 6 from the main text.

5 ΔF along the MEPs in explicit solvent Ala_{10} simulations

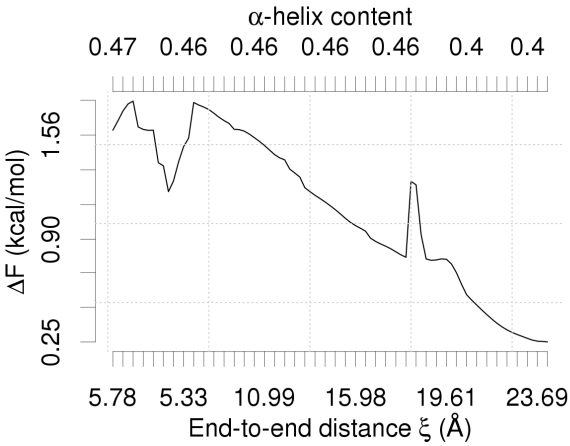
Figure S5 shows the free energy extracted along the four MEPs represented as coloured lines in Figure 6 B from the main text. The barriers between points 2–3 and 4–5 are approximately of 0.4 – 0.5 kcal/mol, making transitions between those points highly probable. The free energy profile for paths 4–1 and 4–3, respectively connecting extended states to the β-hairpin and α-helix conformations, are shown on Figures S5b and S5c. The free energy change ($\Delta\Delta F$) is respectively of 2 and 1.25 kcal/mol emphasising again the easy conformational changes during the simulation.



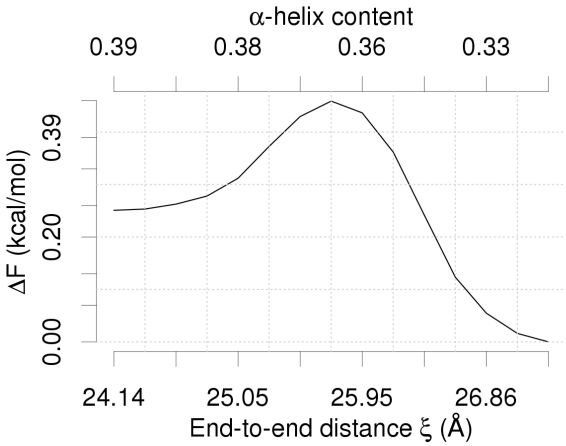
(a) ΔF between points 2 and 3 from Figure 6 B from the main text.



(b) ΔF between points 4 and 1 from Figure 6 B from the main text.



(c) ΔF between points 4 and 3 from Figure 6 B from the main text.



(d) ΔF between points 4 and 5 from Figure 6 B from the main text.

Figure S5: Free energy of the paths (ΔF in kcal/mol) displayed on Figure 6 B from the main text. The two dramatic changes in panel (c) are most probably errors either from the KDE smoothing of the MEP finding algorithm and should not be considered during analysis.

References

(S1) MacQueen, J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. Berkeley, Calif., 1967; pp 281–297.

(S2) Lloyd, S. *IEEE T. Inform. Theory* **1982**, *28*, 129–137.

(S3) J. A. Hartigan, M. A. W. *J. Roy. Stat. Soc. C-App.* **1979**, *28*, 100–108.

(S4) Thorndike, R. *Psychometrika* **1953**, *18*, 267–276.

(S5) Ketchen, D. J.; Shook, C. L. *Strateg. Manag. J.* **1996**, *17*, 441–458.

(S6) R Core Team, *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2015.

(S7) van Etten, J. *gdistance: distances and routes on geographical grids. R package version 1.1-9*; 2015.

(S8) Dijkstra, E. *Numer. Math.* **1959**, *1*, 269–271.

(S9) Rosenblatt, M. *Ann. Math. Statist.* **1956**, *27*, 832–837.

(S10) Parzen, E. *Ann. Math. Statist.* **1962**, *33*, 1065–1076.

(S11) Fiorin, G.; Klein, M. L.; Hénin, J. *Mol. Phys.* **2013**, *111*, 3345–3362.

Chapter 5

MTPs Fitting Wizard

"Never trust a computer you can't throw out a window."

Steve "Woz" Wozniak, co-founder of Apple Inc.

In this Chapter another important project of this thesis is presented: the development of an automated workflow for fitting Lennard-Jones (Equation 1.22), point charges (PC, Equation 1.22) and Multipoles (MTP) parameters (briefly mentioned in Section1.2.1): the Fitting Wizard (FW).

An article was written for presenting the FW, and validating results it produced. It was published in July 2016, in the *Journal of Chemical Information and Modeling* (JCIM), Vol. 56, Pages 1479–1489, [145] co-written with Krystel El Hage and Markus Meuwly.

This software is available free of charge, and under the 3-clause BSD license, from a github repository. [1].

The Fitting Wizard (FW) is a graphics-based, versatile, and modular fitting environment for PC and MTP-based force fields for condensed-phase simulations. It is demonstrated in the article [145] that accurate parametrisations can be obtained for molecules in gas and solvent phase, and that the fitted parameters are usually transferable. Three thermodynamic properties of interest were considered for fitting the LJ FF parameters: density ρ, enthalpy of vaporisation ΔH_{vap} and free energy of solvation ΔG_{solv}. Thermodynamic integration, already mentioned in Section 1.3.3, was used for estimating the enthalpy and free energy, and more details on the method are available in the article (Section 5.3).

In Section 5.1 the theory behind the electrostatic *multipoles expansion* is introduced, together with a few notions concerning FFs parameters fitting.

Section 5.2 contains supplementary details concerning the FW software development.

This article is appended to the current Chapter, together with supplementary information, and can be found in Section 5.3.

5.1 Atomic Multipoles

Multipoles (MTPs) are usually determined from and fitted to an ab-initio Electrostatic Potential (ESP).

Let us assume that $\phi(q)$ represents a continuous electronic density function of the molecular coordinates q, and that one can discretise the values of $\phi(q)$ on a 3-dim grid where the points are defined with coordinates (r_k), then one can write:

$$\phi(q_i) \approx \Phi(r_k)$$

where $k = \{k^x, k^y, k^z\}$ represents the coordinates of the point of the grid which is the closest to the position of atom q_i.

Then one can estimate the ESP for all points r as:

$$\begin{aligned}
\Phi(r) &= \sum_i \sum_j Q_j^{(i)} f_j^{(i)}(r) \\
&\approx \sum_i Q_{00}^{(i)} \mathcal{R}^{-1} + Q_{10}^{(i)} \mathcal{R}^{-2} \hat{r}_z + Q_{11c}^{(i)} \mathcal{R}^{-2} \hat{r}_x + Q_{11s}^{(i)} \mathcal{R}^{-2} \hat{r}_y \\
&+ Q_{20}^{(i)} \mathcal{R}^{-3} (3\hat{r}_z^2 - 1)/2 + Q_{21c}^{(i)} \mathcal{R}^{-3} \sqrt{3} \hat{r}_x \hat{r}_z \\
&+ Q_{21s}^{(i)} \mathcal{R}^{-3} \sqrt{3} \hat{r}_y \hat{r}_z + Q_{22c}^{(i)} \mathcal{R}^{-3} \sqrt{3} (\hat{r}_x^2 - \hat{r}_y^2)/2 \\
&+ Q_{22s}^{(i)} \mathcal{R}^{-3} \sqrt{3} \hat{r}_x \hat{r}_y
\end{aligned} \tag{5.1}$$

where i iterates over all atoms and j over all MTP coefficients, $\mathcal{R} = ||r||$ is the norm of vector r, $\hat{r}_x = r \times \dfrac{\hat{a}}{\mathcal{R}}$ is the norm of the projection of vector r onto one of the three vector defining the basis

[1] https://github.com/MMunibas/FittingWizard

$[\hat{x}, \hat{y}, \hat{z}]$, $Q_{kl}^{(i)}$ is the lth MTP moment of rank k in spherical coordinates, and $f_j^{(i)}(r)$ are geometrical factors, including distance- and angular-dependent terms for the MTP moment $Q_j^{(i)}$ at point r.

The goal is to minimise the difference between $\Phi_{AB}(r)$ (defined using an-initio estimated multipoles moments $Q_{AB}(r)$ and $\Phi_{MTP}(r)$ which contains terms Q_{MTP} used within CHARMM's MTPL module.[15–18]

For that a linear least-square fitting method is applied, where the following target function χ is iteratively minimised:

$$\chi = \min \left\{ \sum_k \left(\Phi_{AB}(r_k) - \Phi_{MTP}(r_k) \right)^2 \right\} \tag{5.2}$$

For more details concerning the procedure, one can refer to the following references: [15–18] and to the article in Section 5.3.

5.2 Overview of the FW workflow procedure

The FW wizard consists in a Java GUI built on top of a set of Python scripts, originally written by C. Kramer and T. Bereau.[16], and extended during this PhD for adding the capability to fit LJ parameters.

A first version of this GUI was initially programmed by an external company, Super Computing Systems Zürich [2], for performing the left part of the workflow presented on Figure 5.1,i.e. corresponding to the above mentioned[16] scripts.

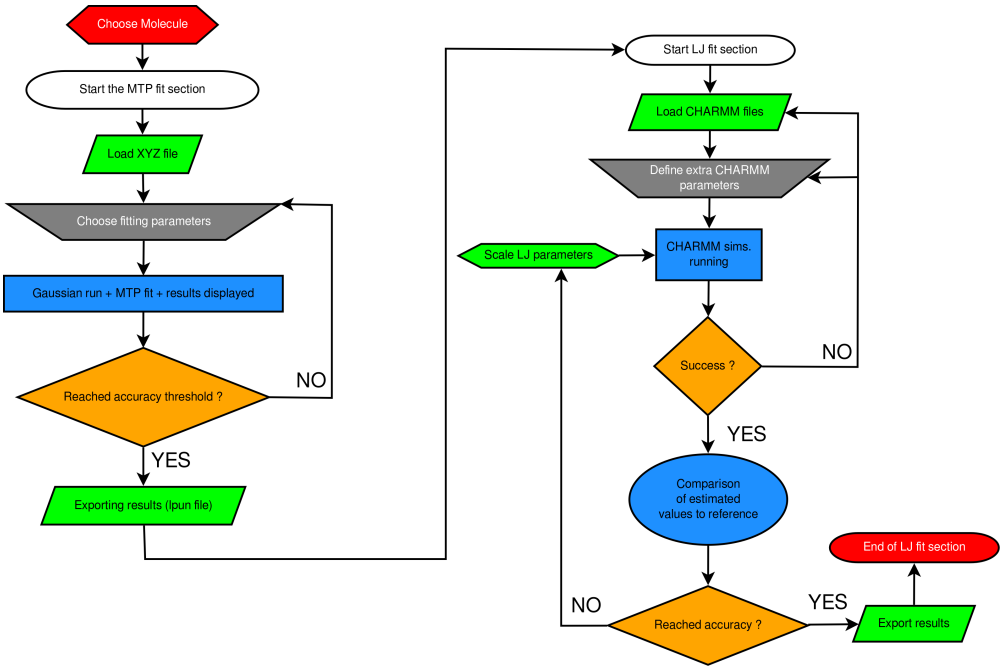


FIGURE 5.1: Operational workflow of the Fitting Wizard toolkit. The left part corresponds to the fit of the MTP parameters to an ab-initio simulation. the right part describes the fit of the LJ parameters for accurately reproducing thermodynamic properties. Refer to the article (Section 5.3)for more details.

The right part of the workflow (Figure 5.1) corresponds to the extra programming work that involved the publication of the article. The most challenging part of the work was to obtain a smooth integration of all the software pieces together: indeed one has to manage, from a Java graphical interface, python and bash scripts, that will submit calculations to a distant computers cluster or run CHARMM locally;

[2] https://www.scs.ch/en/home.htm

i.e. three to four levels of different languages/scripts have to "collaborate", and if one of the level encounters an error there should be enough communication between the concurrent threads or scripts running, in order to provide meaningful information to the user.

Another important feature, detailed in the SI (Section 5.3) was the integration of an embedded *SQL* database of properties of chemical compounds, in order to provide to the user a place where to find experimental reference for thermodynamic properties of interest when fitting the Lennard-Jones parameters.

5.3 MTPs article

The following is the article published in July 2016, in the *Journal of Chemical Information and Modeling* (JCIM), Vol. 56, Pages 1479–1489, [145] co-written with Krystel El Hage and Markus Meuwly. Supplementary information is also provided.

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

A Toolkit to Fit Nonbonded Parameters from and for Condensed Phase Simulations

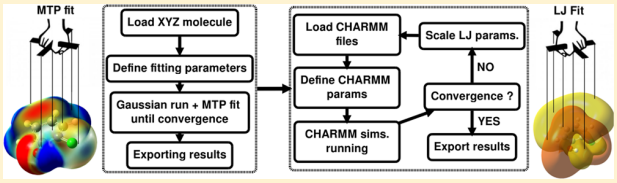
Florent Hédin,[†,§] Krystel El Hage,[†,§] and Markus Meuwly*[,†,‡]

[†]Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland

[‡]Department of Chemistry, Brown University, Providence, Rhode Island 02912, United States

Supporting Information

ABSTRACT: The quality of atomistic simulations depends decisively on the accuracy of the underlying energy function (force field). Of particular importance for condensed-phase properties are nonbonded interactions, including the electrostatic and Lennard-Jones terms. Permanent atomic multipoles (MTPs) are an extension to common point-charge (PC) representations in atomistic simulations. MTPs are commonly determined from and fitted to an *ab initio* Electrostatic Potential (ESP), and Lennard-Jones (LJ) parameters are obtained from comparison of experimental and computed observables using molecular dynamics (MD) simulations. For this a set of thermodynamic observables such as density, heat of vaporization, and hydration free energy is chosen, to which the parametrization is fitted. The current work introduces a comprehensive computing environment (Fitting Wizard (FW)) for optimizing nonbonded interactions for atomistic force fields of different qualities. The FW supports fitting of standard PC-based force fields and more physically motivated multipolar (MTP) force fields. A broader study including 20 molecules ranging from *N*-methylacetamide and benzene to halogenated benzenes, phenols, anilines, and pyridines yields a root mean squared deviation for hydration free energies of 0.36 kcal/mol over a range of 8 kcal/mol. It is furthermore shown that PC-based force fields are not necessarily inferior compared to MTP parametrizations depending on the molecule considered.



INTRODUCTION

Consistent and convenient force field parametrization remains one of the main challenges for more widespread use and high quality atomistic simulations of complex systems. Although considerable progress has been made in implementing advanced treatments of intermolecular interactions, such as multipolar[1−3] and/or polarizable[4−6] force fields, their parametrization still presents a major impediment. Typically, force fields need to be fitted to a heterogeneous set of reference data originating from electronic structure calculations and experiment.[7−9] While fitting to reference energies from *ab initio* calculations is standard and only requires individual energy evaluations, using condensed-phase data such as diffusion coefficients or hydration free energies necessitates entire molecular dynamics (MD) runs.[10] This makes such parametrizations also computationally demanding.

Due to the fundamental importance of accurate descriptions of the inter- and intramolecular energetics, several tools have been developed which make force field parametrizations more amenable. Often, these approaches rely on databases and employ analogies between molecules or functional groups to minimize computational effort. Such tools include ParamChem[11] and MATCH[12] for the CHARMM force field, and the Automated Topology Builder[13] web server for the GROMOS force field. The SwissParam[14] initiative assigns vdW terms by analogy to existing CHARMM atom types while all other parameters (charges, bonds, angles, dihedrals, impropers) are assigned by analogy from the Merck Molecular Force Field[15,16] and translated into the CHARMM format.

Significantly fewer tools are available for developing parameters directly from electronic structure calculations or from fitting to experimental data or both. One of them is Antechamber,[17] which is used to generate parameters for the AMBER and associated general Amber force field (GAFF) force field.[18] and another one is the Force Field Toolkit (FFTK),[19] which is linked to VMD (Visual Molecular Dynamics), provided limited functionality to derive CHARMM parameters from quantum mechanical (QM) calculations. The release of CGenFF, along with a set of procedures for parametrization made possible the development of a comprehensive tool capable of yielding a complete set of CHARMM-compatible parameters.[11,20] To the contrary, recent software solutions (e.g., CGenFF, MATCH) have focused on parameter assignment based on analogy only, although GAAMP (General Automated Atomic Model Parametrization)[21] does derive charge and dihedral parameters based on QM calculations.

With the advent of more advanced multipolar implementations, the need for robust parametrization tools has even increased. Here, we describe a versatile fitting environment which allows determining high-quality multipole-based force fields together with suitable Lennard-Jones parameters for condensed phase simulations. The environment is based on a graphical user interface (GUI) which handles computations and subsequently analyses data from electronic structure and

Received: May 18, 2016
Published: July 20, 2016

DOI: 10.1021/acs.jcim.6b00280
J. Chem. Inf. Model. 2016, 56, 1479−1489

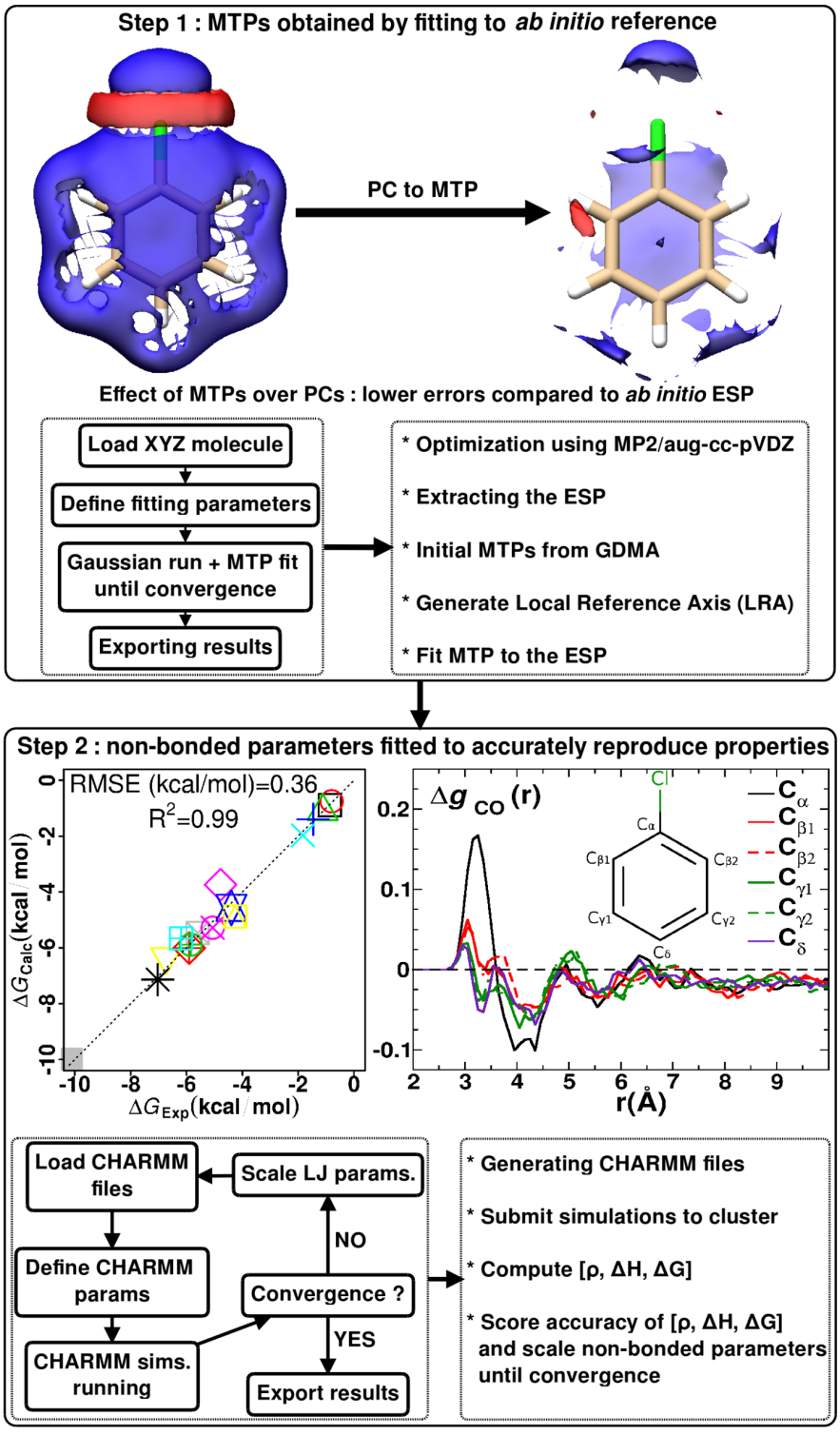


Figure 1. Flowchart illustrating the Fitting Wizard. (top) Fitting the MTPs to the ESP as obtained from the electronic structure calculations. (bottom) Refinement of LJ parameters for optimal reproduction of selected thermodynamic observables. (left) Comparison of experimental and computed ΔG_{hyd}. (right) Atom-specific differences in the radial distribution function $\Delta g(r)$ between a PC and a MTP parametrization for chlorobenzene.

molecular dynamics codes. In the present case this is output from Gaussian09[22] and input to/output from CHARMM.[23] The reference data consists of electronic structure information and thermodynamic properties from experiment.

METHODS AND IMPLEMENTATION

A stand-alone, convenient, and accurate force field fitting environment involves handling several tasks. First, for the procedure pursued here, the electron density $\rho(\vec{x})$ is determined

DOI: 10.1021/acs.jcim.6b00280
J. Chem. Inf. Model. 2016, 56, 1479−1489

from electronic structure calculations for an optimized structure at a given level of theory. Next, local reference axis (LRAs) systems need to be defined for calculating multipolar interactions. Then, atomic multipole coefficients (MTPs) are fitted to best reproduce the electrostatic potential (ESP). Next, atom types and bonded force field terms (bonds, angles, dihedrals) are assigned and LJ parameters for the particular atom types are required for the molecular dynamics (MD) simulations. Finally, MD simulations are run and analyzed from which the necessary thermodynamic observables are determined, compared with experiment and provide information about how to adjust the LJ parameters. These steps together with some formal background are described next.

Electronic Structure Calculations. All *ab initio* calculations in the present work were carried out with the Gaussian09 suite of codes,[22] using second-order Møller−Plesset (MP2)[24] theory and the aug-cc-pVDZ[25−27] basis set. This level of theory is a good compromise between accuracy and speed. These are parameters that are easily changed in the protocol. After optimization of the molecular structure the electron density is extracted with the cubegen utility on a rectangular grid. Grid spacings ranging from 0.1 to 0.4 Å yield almost identical results. The initial atomic multipole moments are obtained from a Distributed Multipole Analysis[28] using the GDMA code. This corresponds to the first three steps in the top panel of Figure 1.

Determine LRAs. Local reference axes are required to define the static multipoles assigned to an atom relative to the global coordinate system. LRAs need to be assigned to each atom of the molecule which are treated with MTPs. The assignment has been described in detail previously.[29] Briefly, the procedure (see fourth step of the top panel of Figure 1) starts from the chemical atom type and determines the number and connectivity of the nearest neighbor atoms. From this information the "full atom type" is generated as a list of the atom type itself and its nearest and second nearest neighbors. From this, the LRA for each atom can be determined.[30]

Fitting MTPs. To ensure consistency between the CGenFF nonbonded parameters[11,20] (PCs and LJ) and the fitted MTPs, each monopole was constrained to deviate at most by an amount λ_{PC} from the reference value (i.e., provided by CGenFF). Effectively, larger values of λ_{PC} will provide more flexibility—and thus better fits—at the expense of consistency with the reference PCs. Such an approach considers higher order multipoles as corrections to a zeroth-order PC force field.

The ESP can be approximated using MTPs (up to quadrupoles), at any grid point $\mathbf{r}^{(p)}$, from[31−34]

$$\begin{aligned}\Phi(\mathbf{r}^{(p)}) &= \sum_i \sum_j Q_j^{(i)} f_j^{(i)}(\mathbf{r}^{(p)}) \\ &\approx \sum_i Q_{00}^{(i)} r^{-1} + Q_{10}^{(i)} r^{-2} \hat{r}_z + Q_{11c}^{(i)} r^{-2} \hat{r}_x + Q_{11s}^{(i)} r^{-2} \hat{r}_y \\ &+ Q_{20}^{(i)} r^{-3} (3\hat{r}_z^{\,2} - 1)/2 + Q_{21c}^{(i)} r^{-3} \sqrt{3}\, \hat{r}_x \hat{r}_z \\ &+ Q_{21s}^{(i)} r^{-3} \sqrt{3}\, \hat{r}_y \hat{r}_z + Q_{22c}^{(i)} r^{-3} \sqrt{3}\, (\hat{r}_x^{\,2} - \hat{r}_y^{\,2})/2 \\ &+ Q_{22s}^{(i)} r^{-3} \sqrt{3}\, \hat{r}_x \hat{r}_y \end{aligned} \tag{1}$$

where i iterates over all atoms and j over all MTP coefficients, $\mathbf{r}$ is the vector from atom i to $\mathbf{r}^{(p)}$, $r = \|\mathbf{r}\|$ is the norm of $\mathbf{r}$, and $\hat{r}_a = \mathbf{r}\cdot\hat{a}/r$ is normalized using one of the three unit vectors $\mathbf{x}$, $\mathbf{y}$, or $\mathbf{z}$. $Q_{kl}^{(i)}$ is the lth MTP moment of rank k in spherical coordinates, and $f_j^{(i)}(\mathbf{r}^{(p)})$ are geometrical factors, including distance- and angular-dependent terms for the MTP moment $Q_j^{(i)}$ at point $\mathbf{r}^{(p)}$.

MTP coefficients $Q_j^{(t)}$ are fitted (last step of the top panel of Figure 1) to the collection of ESP grid points $\mathbf{r}^{(p)}$ by optimizing the target function

$$\chi^2 = \min \sum_p (\Phi_{ab\ initio}(\mathbf{r}^{(p)}) - \Phi_{\mathrm{MTP}}(\mathbf{r}^{(p)})) \tag{2}$$

which minimizes the error between the *ab initio* and MTP-derived ESPs.[29] Because the problem is linear, we can rewrite the problem as $\mathbf{Xb} = \mathbf{y}$, and because of the sparsity of $\mathbf{X}$ we instead solve

$$\mathbf{X}^{\mathrm{T}}\mathrm{Xb} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{3}$$

where $\mathbf{X}^{\mathrm{T}}$ is the transpose of $\mathbf{X}$.

Assignment of Atom Types for MD Simulations. Next, atom types are required for assigning bonded terms between atoms and Lennard-Jones parameters. This step is automated and the methodology is related to the one used by the SwissParams web-portal.[14] Based on the connectivity of the atoms, a hybridization state (e.g., sp^2, sp^3) is assigned to each of the heavy (i.e., not hydrogen) atoms. Then, based on its hybridization and the hybridization of its neighbor atoms, a CGenFF FF atom type is assigned to each atom, e.g. "CT3" for an sp^3 carbon with four explicit substitutes. From this a PDB file compatible with CHARMM can be generated, together with a topology and a structure file.

With chemical atom types assigned, the force field for the compound (including bonds, angles, dihedrals, partial charges and Lennard-Jones parameters) is generated according to the CGenFF force field.[20] However, because the electrostatic interactions are modified (i.e., switching from PCs to MTPs), reparametrization of the LJ coefficients is necessary as they were optimized for use with PCs. This is another reason why keeping PCs in the fitting close to the CGenFF values, namely that the CGenFF LJ parameters can be used as a consistent starting point in their refinement. This is part of the first step of the bottom panel of Figure 1.

MD Simulations. Atomistic simulations (bottom panel of Figure 1) are carried out in order to determine the necessary thermodynamic data (see below). The CHARMM-input files are assembled from the `Java` GUI. These are then submitted to a computing pipeline through the GUI, relying on a `Python` scripts engine, in order to allow users to easily customize the procedure.

Fitting the Lennard-Jones Parameters. For refining the LJ parameters, thermodynamic properties are often used as a reference. Here, they include pure liquid density ρ, heat of vaporization ΔH_{vap} and hydration free energies ΔG_{hyd}. Ideally, one would proceed by fitting the LJ radius of each atom type independently. However, this is neither practical (because for each combination of parameters an independent MD simulation is needed) nor desirable, as it would require a high-dimensional parametrization for an undetermined problem (typically considerably more parameters than observables). Furthermore, established LJ parameters from a validated force field often have already a certain balance which would be compromised if arbitrary scaling would be allowed and retaining this balance may be advantageous. Hence, LJ parameters are rescaled by a parameter l according to $\varepsilon^* = l\varepsilon$ and $R_{\min}^*/2 = lR_{\min}/2$. It is possible to use a separate scaling for ε and $R_{\min}/2$. However, for a full grid evaluation this considerably increases the number of

DOI: 10.1021/acs.jcim.6b00280
J. Chem. Inf. Model. 2016, 56, 1479−1489

simulations to perform. This is part of the last step of the bottom panel of Figure 1.

For determining the pure liquid density, multipole−multipole interactions are needed. This requires the definition of local reference axes (see discussion above).[29] Since all coefficients are expressed in the atom's local frame, they are independent of orientation. The geometry of two atoms a and b relative to the orientation of their MTP sites is then determined by incorporating the unit vectors of their local axis systems $\{\mathbf{w}^a\} = \{\mathbf{x}^a, \mathbf{y}^a, \mathbf{z}^a\}$ for atom a and likewise for b. The set of $\{\mathbf{w}^a\}$ and $\{\mathbf{w}^b\}$ combined with the intersite unit vector $\hat{\mathbf{R}}$ defines the direction cosines $\mathbf{q} = \{R, \mathbf{w}^a\cdot\hat{\mathbf{R}}, \mathbf{w}^b\cdot\hat{\mathbf{R}}, \mathbf{w}^a\cdot\mathbf{w}^b\}$ that provides a geometric description of the two MTP sites. From the interaction functions $T^{ab}_{tu}(\mathbf{q})$ for two MTP moments Q^a_t and Q^b_u of order t and u on atomic sites a and b, respectively, the interaction energy is

$$U^{ab}_{tu}(\mathbf{q}) = Q^{\ a}_{t}\cdot Q^{\ b}_{u}\cdot T^{ab}_{tu}(\mathbf{q}) \tag{4}$$

This is the MTP implementation pursued in the MTPL module.[2]

The bottom panel of Figure 1 also reports concrete results from fitting studies. The left-hand panel highlights the accuracy of ΔG_{hyd} for 20 compounds studied for which calculated and experimental ΔG_{hyd} agree very favorably. The right-hand panel reports differences in the radial distribution functions $\Delta g(r)$ between the C atoms of PhCl and the water-oxygen atoms for PC and MTP parametrization with optimized LJ parameters.

Additional Remarks. While the GUI runs on the local machine, *ab initio* and MD calculations are carried out on a distributed computing environment, and data files are retrieved using the `ssh` transmission protocol. This approach allows to use any computing cluster and no dedicated installation procedure is required on the server side. For the LJ fit (Figure1 (bottom)), all MD simulations for estimating the thermodynamic observables are submitted at once, in order to exploit as much as possible the distributed architecture of the computing cluster. The above-mentioned set of scripts currently supports the `qsub` jobs submissions, but extending the workflow for supporting other systems such as `sbatch` should be straightforward.

COMPUTING THERMODYNAMIC OBSERVABLES

The thermodynamic observables considered here (ρ, ΔH_{vap}, ΔG_{hyd}) require entire MD simulations to be run.[2] For automating this step, a suitable set of core input files for the MD engine used (here CHARMM) is set up. All MD simulations use a time step of Δt = 1 fs, solvent simulations are carried out with periodic boundary conditions (PBC) with a nonbonded cutoff of 12 Å and using Particle Mesh Ewald summation[35] for the PCs, with a width of the Gaussian distribution κ = 0.34, a B-spline interpolation of fifth degree, and 32 grid points along each spatial dimension. The box size is adapted to the probe molecule's size and usually of dimension 20^3 to 25^3 Å^3, corresponding to a total number of ∼270 to 520 water molecules. For calculating solvation free energies, the TIP3P[36] water model is used, although this is easily modified to other available water models. All simulations are carried out in the *NPT* ensemble, using the Leap-Frog integrator, and the Hoover algorithm is used for constant pressure and constant temperature simulations. Bonds involving hydrogens were constrained with SHAKE.[37] Further details are given below in the sections which discuss individual observables.

Heat of Vaporization. Molecular dynamics simulations provide a convenient way to compute the heat of vaporization[38]

$$\Delta H_{\mathrm{vap}}(T) = E_{\mathrm{gas}}(T) - E_{\mathrm{liq}}(T) + RT \tag{5}$$

where E_{gas} and E_{liq} are the potential energies of one molecule in the gas and liquid (i.e., *NPT*) phases, respectively, and R is the gas constant. The gas-phase energy is computed from the minimized energy and the number of atoms, N, and constrained degrees of freedom, N_{cons}, in the molecule, according to

$$E_{\mathrm{gas}}(T) = E^{\mathrm{minimized}}_{\mathrm{gas}} + \frac{1}{2}RT(3N - 6 - N_{\mathrm{cons}}) \tag{6}$$

Thermodynamic Integration. Free energies of hydration (i.e., solvation in water) are computed using thermodynamic integration (TI). TI gradually couples/decouples chemical groups from the system by applying a scaling parameter λ to the nonbonded interactions (i.e., electrostatics and LJ). The total Hamiltonian is written as a function of λ

$$\Delta G_{\mathrm{A}\to\mathrm{B}} = \int_0^1 \mathrm{d}\lambda \left\langle \frac{\partial \mathcal{H}}{\partial \lambda} \right\rangle_{\lambda} \approx \sum_i (\lambda_{i+1} - \lambda_i) \left\langle \frac{\partial \mathcal{H}}{\partial \lambda} \right\rangle_{\lambda_m} \tag{7}$$

where A → B refers to the alchemical transformation between compounds A and B. The canonical average $\langle\cdot\rangle_\lambda$ is performed over the phase space generated by the Hamiltonian $\mathcal{H}(\lambda)$, and $\lambda_m = (\lambda_i + \lambda_{i+1})/2$. For the LJ and PC derivatives CHARMM's PERT module using soft-core potentials for the LJ interactions[39,40] is used. No long-range corrections were applied to the LJ-interactions, as no noticeable change was found when increasing the nonbonded cutoff beyond r_c = 12 Å.

The LJ and electrostatic interactions are turned on separately.[2] First, the LJ interactions with soft-core potentials are fully grown, followed by the electrostatics in the presence of the full van der Waals interactions, thereby avoiding the need for soft-core electrostatic potentials. The change in free energy due to MTP electrostatics with coupling λ_m was computed by first performing a simulation where all MTP energies (see eq 4), forces, and torques were linearly scaled by λ_m. In a postprocessing step the energies with the original Hamiltonian (unscaled, λ = 1) are extracted and averaged over the solute−solute and solute−solvent energies (i.e., solvent−solvent interactions are not affected by λ_m) in such a way that its derivative with respect to λ yields the original energy (unscaled λ = 1).

Using a thermodynamic cycle, the hydration free energy is computed according to $\Delta G_{\mathrm{hyd}} = \Delta G_{\mathrm{sol}} - \Delta G_{\mathrm{vac}}$, where ΔG_{sol} and ΔG_{vac} correspond to the free energy of insertion of the compound in a box of water and vacuum, respectively. For the simulations in water the solute was placed in a box of ≈500 solvent molecules.

The grid of λ points is chosen in different ways. The most accurate, automatic procedure starts from 20 evenly spaced λ windows between 0 and 1. For further refinement, windows at the two ends of the λ interval (typically $\lambda \in [0, 0.1]$ and $\lambda \in [0.9, 1]$) are further partitioned to retain accuracy. However, introduction of additional partitions is inconvenient in the present context as it requires to run an a priori unknown number of simulations in a sequential manner. Hence, further different strategies were explored for this step. The best performance was found for grid spacing $\Delta\lambda$ = 0.025 for $\lambda \in [0.0, 0.1]$; $\Delta\lambda$ = 0.100 for $\lambda \in [0.1, 0.9]$ and $\Delta\lambda$ = 0.025 for $\lambda \in [0.9, 1.0]$. Such a procedure allows to submit all λ windows at once which considerably speeds up turnover times for individual fitting cycles. However, for accuracy checks the interface also allows

DOI: 10.1021/acs.jcim.6b00280
J. Chem. Inf. Model. 2016, 56, 1479−1489

automatic subdivision of the windows for particularly relevant parametrization problems.

Database of Compounds. Fitting force fields for condensed-phase simulations requires reference data for adjusting the parameters, as described above. In the present approach, the atomic multipoles are fit to best reproduce the electrostatic potential from electronic structure calculations whereas adjustment of van der Waals parameters requires solution-phase data. For this, a database containing experimental values from the literature has been built. The current version of the database includes mass, density, enthalpy of vaporization and the hydration free energy (where available) as reference data. Mass, density, and ΔH_{vap} are those from PubChem,[41] and solvation free energies were taken from the FreeSolv[42,43] database built by Mobley et al., which contains values collected from the literature.

The database is searchable by name, chemical formula or SMILES[44,45] and uses the SQL language.[46] It was decided to provide, as an embedded feature within the Fitting Wizard (FW), the access to a database of chemical compounds. The database was built according to the following procedure: (i) the version v0.31 of the database was downloaded, containing values for ΔG_{hyd} for 643 compounds, together with their PubChem ID, SMILES notation, IUPAC name and a DOI literature reference. (ii) a MySQL database was created using the database content. (iii) the PubChem ID was used for automatically retrieving (using the provided Application Programming Interface (API)[47]) the previously mentioned properties (m, ρ, ΔH_{vap}, ΔG_{hyd}). However, missing values or inconsistencies may remain for some of the compounds even after data curation: thus the database is editable, and then provided as a starting set the user can use and improve. See SI section I and Figure S1 for further details concerning the database.

VALIDATION AND RESULTS

For validating the Fitting Wizard several problems are considered. First, the parametrization of *N*-methylacetamide (NMA) is reconsidered and extended as it serves as a model for peptides and proteins. Second, the parametrization of substituted benzenes is presented as a case where—particularly for the case of halogen-substituted benzenes—MTPs have been found to be essential for an accurate description of solvent properties.[2,48]

***N*-Methyl-acetamide.** As the central building block for peptides and proteins, NMA is a meaningful test system. Experimental data is available for all three observables considered. Starting from an optimized MTP model, the LJ parameters are adjusted to best reproduce the experimentally measured ρ, ΔH_{vap}, and ΔG_{hyd}.[48] The influence of scaling the LJ parameters is summarized in Table 1 where results for $l \in [0.9; 1.1]$ are presented, meaning that LJ parameters were changed by up to 10% around their reference CGenFF-values. In order to determine the best-performing model, a simple weighted score $S = \sum_{i=1}^{3} w_i(\mathrm{Obs}_i - \mathrm{Calc}_i)^2$ with $w_\rho = 1$, $w_{\Delta H} = 3$ and $w_{\Delta G} = 5$ is introduced to differently weight the three observables. Such a weighting puts more emphasis on hydration free energies but alternative choices are possible for particular purposes and applications. The model with $l = 0.95$ yields the lowest score ($S = 0.1$) and is therefore the preferred one. Both, ΔH_{vap} and ΔG_{hyd} are reproduced to within less than 1% compared to the reference data whereas the density differs by 6%. Obvious extensions involve separate scaling factors for σ and ϵ which, however, further increases computational demands. Nevertheless, other models yield competitive scores well below $S = 1.0$. It should be noted that the experimental ΔH_{vap} used in force field parametrizations has been studied recently and it was found that $\Delta H_{vap} = 13.0 \pm 0.1$ kcal/mol at 410 K is the preferred value.[49] As ΔH_{vap} increases with decreasing temperature, the value used in the present work ($\Delta H_{vap} = 14.2$ at 300 K) should be qualitatively correct. However, refinement of this based on the detailed study in ref 49 may be desirable.

Table 1. Dependence of ρ (g/cm^3), ΔH_{vap} (kcal/mol), and ΔG_{hyd} (kcal/mol) when Scaling the Lennard-Jones Parameters[a]

scaling l	ρ	ΔH_{vap}	ΔG_{hyd}	score S
0.9	1.13	14.24	−9.82	0.4
0.925	1.08	13.95	−9.89	0.4
0.95	**1**	**14.11**	**−9.99**	**0.1**
0.975	0.99	13.84	−10.22	0.5
1	0.95	13.82	−9.88	0.6
1.025	0.92	13.68	−9.06	6.0
1.05	0.88	13.57	−8.75	10.0
1.075	0.84	13.29	−8.38	16.9
1.1	0.81	13.47	−8.07	21.8
exp	0.94[41,50]	14.2[41,51]	−10.08[52]	

[a]Bold text shows the value of l minimizing the score S.

Performance of a Predefined λ Grid. As mentioned in the Methods and Implementation section (see Thermodynamic Integration), automated refinement of the λ grid on either side of the interval $\lambda \in [0, 1]$ is computationally inconvenient as each subdivision can only be made once the updated hydration free energy is available. Ideally, one would work with a predefined grid of λ values which allows to submit all necessary simulations at the same time. This improves turnover times, and the total time for an entire optimization (a few hours for a molecule such as NMA) can therefore be estimated a priori. The choice of this subdivision is flexibly handled in the fitting wizard. Here, it is merely illustrated that such a predefined grid can yield good-quality parametrizations, but the subdivision is likely to depend on the particular molecule or class of molecules considered. Three possibilities I–III were explored in the following.

(I) [$\lambda \in [0, 0.1]$ with $\Delta\lambda = 0.010$; $\lambda \in [0.1, 0.9]$ with $\Delta\lambda = 0.100$; $\lambda \in [0.9, 1.0]$ with $\Delta\lambda = 0.025$].
(II) [$\lambda \in [0, 0.1]$ with $\Delta\lambda = 0.020$; $\lambda \in [0.1, 0.9]$ with $\Delta\lambda = 0.100$; $\lambda \in [0.9, 1.0]$ with $\Delta\lambda = 0.020$].
(III) [$\lambda \in [0, 0.1]$ with $\Delta\lambda = 0.025$; $\lambda \in [0.1, 0.9]$ with $\Delta\lambda = 0.100$; $\lambda \in [0.9, 1.0]$ with $\Delta\lambda = 0.025$].

The results for MTP/LJ optimizations for trans-NMA with different subdivisions of the λ windows are summarized in Table 2. It is found that the hydration free energy changes by about 5% depending on the subdivision of the λ interval. On the other hand it is possible to find a subdivision (here case I) which provides an accurate estimate and is computationally efficient.

ΔG_{hyd} for Cis- and Trans-NMA from a Polarizable Drude Model. In previous and also more recent computational studies it was observed that calculated ΔG_{hyd} values differ for the cis- and trans-isomers for NMA.[55,56] Experimentally, the direct determination of ΔG_{hyd}^{cis} is difficult due to the low population of this isomer (<2%) in solution[53] although it is generally believed that the differential hydration free energy $\Delta\Delta G_{hyd} = \Delta G_{hyd}^{trans} - \Delta G_{hyd}^{cis} \approx 0$.[57,58] Hence, experimental values for ΔG_{hyd}^{cis} are indirect and also have been put into question for different reasons.[59] As a comparison with a recently published generic and polarizable force field, hydration free energies were also determined for cis- and trans-NMA using the Drude force field.[6] As recommended, simulations are carried out with the SWM4-DP[5] water model

DOI: 10.1021/acs.jcim.6b00280
J. Chem. Inf. Model. 2016, 56, 1479−1489

Table 2. Comparison of ΔG_{hyd} for Different λ Subdivisions (sets I–III) or Heuristically Decided by CHARMM's PERT Module (Based on the Fluctuation of the Average)[a]

method	ΔG_{hyd}
expt[52−54]	−10.08
simulation[6]	−9.90
automated division	−9.99
case I	−9.61
case II	−9.38
case III	−9.29

[a]Automated is the default mode. Predefined λ windows speed up the process through a pre-determined number of simulations.

instead of TIP3P, the (default) automatic λ−division procedure is used for TI, and all other simulation parameters are identical to those in the MTP-simulations. For the two isomers ΔG^{cis}_{hyd} = −8.67 kcal/mol and ΔG^{trans}_{hyd} = −9.81 kcal/mol were found. The value for ΔG^{trans}_{hyd} agrees to within 0.09 kcal/mol with the reference value[6] which validates the present protocol. Despite using a polarizable model, $\Delta\Delta G_{hyd}$ = 1.1 kcal/mol between the two isomers, which differs from the assumed value of close to zero from experiment. Compared to this, the present nonpolarizable simulations yield $\Delta\Delta G_{hyd}$ = 1.8 kcal/mol. As the Drude simulations do not employ multipoles and the present MTP simulations are nonpolarizable it is possible that combining the two will yield satisfactory agreement with experiment. As another comparison, a recent parametrization study based on electron density partitioning found $\Delta\Delta G_{hyd}$ = −1.0 kcal/mol with the cis-isomer more stable than trans-NMA.[56]

HALOGENATED AND SUBSTITUTED BENZENES

Next, a validation study was performed for halogenated and substituted benzenes. They constitute important building blocks in medicinal chemistry and pharmaceutically active substances.[60−63] Also, halogenated amino acid side chains have recently found to be useful modifications in protein biochemistry, such as in insulins.[64] Besides the accuracy of such a parametrization it is also of interest to test the transferability of the final parameters. This is important in situations when the chemical environment of a group changes and the accuracy of the original parametrization should be retained.

Halogenated Phenols. As a first example, 4-BrPhOH is considered. Table 3 reports the calculated hydration free energy

Table 3. Hydration Free Energies Calculated for 4-BrPhOH Depending on the Electrostatic and LJ Parameter Treatment Used[a]

Treatment	ΔG_{hyd}	$\lvert\Delta\Delta G_{hyd}\rvert$
CGenFF parameters (unoptimized PC and LJ)	−10.07	4.22
optimized MTP/CGenFF LJ	−6.37	0.52
optimized MTP/LJ transferred from the work of Bereau et al.[2]	−5.56	0.29
optimized MTP/LJ transferred from the work of Bereau et al.[2] for C, H, Br and optimization of (ϵ, σ) for −OH	−5.89	0.04

[a]$\lvert\Delta\Delta G_{hyd}\rvert$ represents the absolute error relative to the experimental value (−5.85 kcal/mol).[65]

depending on the parametrization and level of optimization used for 4-BrPhOH. The calculated ΔG_{hyd} with PCs and LJ parameters transferred from CGenFF (i.e., unoptimized PC and LJ) overestimates the solvation energy by 4.22 kcal/mol. A considerable improvement of the calculated ΔG_{hyd} is obtained by including MTP electrostatics whereby ΔG_{hyd} drops from −10.07 kcal/mol (CGenFF parameters) to −6.37 kcal/mol (optimized MTP and CGenFF LJ parameters) that differs from the experimental value by only 0.52 kcal/mol. This can be explained by the fact that a simple unoptimized PC electrostatic model cannot describe the large electronic reorganization around the Bromobenzene ring when −OH is introduced in the para position. Moreover, when using previously optimized LJ parameters for bromobenzene (PhBr)[2] instead of standard-CGenFF parameters, the error in ΔG_{hyd} further reduces to 0.29 kcal/mol. Finally, with a slight optimization of the "−OH" group LJ parameters (scaling of σ and ϵ, see above) the calculated ΔG_{hyd} reproduces the experimental value with a difference of 0.04 kcal/mol, which falls within the statistical error typically found on computed values (around 0.05 kcal/mol).

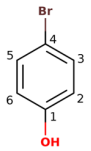
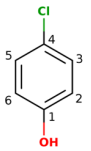
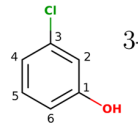
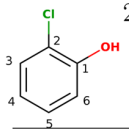
Transferability. One essential aspect in modern force field development and practical applications is the transferability of parametrizations for a chemical building block (e.g., an amino acid side chain) between two different chemical environments which can considerably speed up parametrization tasks and is also conceptually appealing. To assess transferability within the given fitting methodology the hydration free energy of different parametrizations was computed for Br and ClPhOH. Here, the differential solvation free energy $\Delta\Delta G_{hyd}$ in transferring LJ parameters (ϵ and σ) for common atom types (aromatic C, H, Cl, and Br) obtained from previous parametrizations[2] of PhBr and PhCl to 4-BrPhOH and 4-ClPhOH and from the current parametrization of 4-BrPhOH's polar −OH group (see Table 3) to 2,3,4-ClPhOH, is considered. The effect of reoptimizing the LJ parameters on the −OH group in positions (2-, 3-, and 4-) is also evaluated. For all molecules considered (2-, 3-, 4-ClPhOH, and 4-BrPhOH), MTP electrostatics was first fitted individually and not transferred since the impact of the −OH group insertion on the electron distribution varies depending on its position (2-, 3-, or 4-) and the type of the halogen present (Cl or Br).

Table 4 reports calculated hydration free energies for 2-, 3-, 4-ClPhOH, and 4-BrPhOH with LJ parameters transferred for aromatic C, H, Cl, and Br from previous parametrizations of PhBr and PhCl,[2] and with LJ parameters for the −OH group (a) taken from CGenFF or (b) optimized for 4-BrPhOH and transferred to 2,3,4-ClPhOH and compares them to experimentally determined values.[65,66] For the transferred parametrizations, the difference between computed and experimentally determined ΔG_{hyd} is 0.3 and 0.8 kcal/mol for 4-BrPhOH and 2-ClPhOH, respectively. For 3- and 4-ClPhOH they are below 0.4 kcal/mol. Hence, results with transferred parameters are well within 1 kcal/mol which points toward a good degree of transferability. Furthermore, improved hydration free energies after reoptimizing the −OH LJ parameters (ϵ and σ) are also reported for 4-BrPhOH and 2-ClPhOH in Table 4. They decrease to 0.04 and 0.34 kcal/mol for 4-BrPhOH and 2-ClPhOH, respectively.

As an illustration of the effect of different LJ parameters, the water structure around 2-ClPhOH is considered. For this, the radial distribution function $g(r)$ of water around the solute is determined. As an example for a recent application, it has been shown for fluoro-acetonitrile solvated in water that the combination of optical spectroscopy and atomistic simulations is able to detect incipient halogen bond formation.[67] Figure 2 reports the radial distribution function between 2-ClPhOH(O) and water(O) from 2 ns of *NPT* simulations using two different sets of −OH LJ parameters. The first set (ΔG_{hyd} = −5.32 kcal/mol) uses the LJ parameters transferred from 4-BrPhOH to which they

DOI: 10.1021/acs.jcim.6b00280
J. Chem. Inf. Model. 2016, 56, 1479−1489

Table 4. Hydration Free Energies Calculated Depending on the Radical's Position (Here −OH) Relative to the Halogen and the LJ Parameters Used[a]

Molecule		ΔG_{hyd}^{Exp} [65,66]	ΔG_{hyd}^{Calc}	$\lvert\Delta\Delta G_{hyd}\rvert$	Transferred	Optimized
	4-BrPhOH	−5.85	−5.56	0.29	C, H, Br Bereau et *al.* -OH CGenFF	
			−5.89	0.04	C, H, Br Bereau et *al.*	-OH (ϵ, σ)
	4-ClPhOH	−7.03	−6.65	0.38	C, H, Cl Bereau et *al.* -OH LJ optimized for 4-BrPhOH	
	3-ClPhOH	−6.62	−6.34	0.28	C, H, Cl Bereau et *al.* -OH LJ optimized for 4-BrPhOH	
	2-ClPhOH	−4.55	−5.32	0.77	C, H, Cl Bereau et *al.* -OH LJ optimized for 4-BrPhOH	
			−4.89	0.34	C, H, Cl Bereau et *al.*	-Cl (ϵ) -OH (ϵ, σ)

[a] $\lvert\Delta\Delta G_{hyd}\rvert = \lvert\Delta G_{hyd}^{Exp} - \Delta G_{hyd}^{Calc}\rvert$. All MTPs optimized individually.

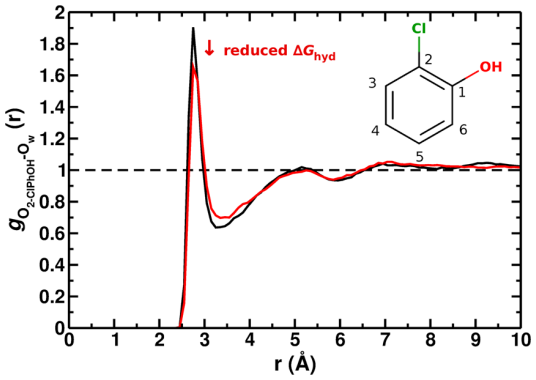


Figure 2. Radial distribution function $g(r)$ for (2-ClPhOH)O−O(water). The black and red traces represent the distribution functions before (first set) and after (second set) optimizing the −OH LJ parameters, respectively. The inset represents the chemical structure of 2-ClPhOH.

were optimized (Table 4 second row). The second set (ΔG_{hyd} = −4.89 kcal/mol; Table 4 last row) uses the −OH LJ parameters (σ and ϵ) optimized specifically for 2-ClPhOH with respect to ρ, ΔH_{vap} and ΔG_{hyd}, starting from the parameters of the first set. The first set was optimized for an −OH group in position 4- (opposite to the halogen atom) whereas the second set was optimized for an −OH group in position 2- (adjacent to the halogen). The O−O_W pair distribution function obtained with both LJ parameter sets (Figure 2, black and red lines) peaks at ∼2.8 Å. However, the amplitude of the peak is smaller for the second set (Figure 2, red line) and the first minimum is less pronounced. The reduced amplitude of $g(r)$ also decreases the occupation number $N(r_s) \propto \int_0^{r_s} g(r) r^2\, dr$ of water molecules within a distance r_s around the −OH group, which also reduces the hydration free energy by 0.4 kcal/mol.

DEGREES OF PARAMETER OPTIMIZATION

To further illustrate the effect of LJ reparametrization, Table 5 reports the three thermodynamic observables (ρ, ΔH_{vap}, ΔG_{hyd}) for different optimization levels for NMA and 4-ClPhOH. While ρ varies little throughout the range studied, ΔH_{vap} and ΔG_{hyd} strongly change.

The results in Table 5 establish that depending on the system studied (NMA or 4-ClPhOH), an optimized PC/LJ model (S = 0.3) can perform very well compared to an optimized MTP/LJ parametrization (S = 0.1). This is the case for NMA. Contrary to that, the halogenated system 4-ClPhOH evidently requires optimized MTP electrostatics and optimized LJ parameters. It is also important to note that LJ parameters can be transferred

Table 5. Computed (ρ, ΔH_{vap}, ΔG_{hyd}) Values for Force Fields of Different Optimization Levels for NMA and 4-ClPhOH[a]

	ρ	ΔH_{vap}	ΔG_{hyd}	score S
	NMA			
CGenFF (PC and LJ)	0.98	15.09	−11.03	6.9
opt PC/CGenFF LJ	0.99	14.49	−10.11	0.3
opt PC/opt LJ	0.99	14.49	−10.11	0.3
opt MTP/CGenFF LJ	0.95	13.82	−9.88	0.6
opt MTP/opt LJ	0.95	14.11	−9.99	0.1
exp	0.94[41,50]	14.20[41,51]	−10.08[52]	
	4-ClPhOH			
CGenFF (PC and LJ)	1.28	15.74	−5.47	72.9
opt PC/CGenFF LJ	1.27	10.78	−5.44	13.2
opt PC/opt LJ	1.25	11.85	−5.61	11.2
opt MTP/CGenFF LJ	1.27	12.86	−5.91	14.2
opt MTP/opt LJ	1.25	11.46	−7.14	0.2
exp	1.22[41,68]	11.24[41,69]	−7.03[66]	

[a] The score is used to differentiate between different levels of optimization. Units are g·cm^{-3} for ρ and kcal/mol for ΔH_{vap} and ΔG_{hyd}.

DOI: 10.1021/acs.jcim.6b00280
J. Chem. Inf. Model. 2016, 56, 1479−1489

Table 6. ΔH_{vap}, ΔG_{hyd} (kcal/mol), and ρ (g/cm^3) as Calculated Using the FW (Calc) with Optimized MPT and LJ Parameters Compared to Experimental References (Exp)[a]

	ρ			ΔH_{vap}			ΔG_{hyd}			
	exp[41]	calc	|devl	exp[41]	calc	|devl	exp[42,43]	calc	|devl	score S
benzene	0.88	0.9	0.02	7.89	7.88	0.01	−0.86	−0.89	0.03	0.01
fluorobenzene	1.02	1.05	0.03	8.26	8.6	0.34	−0.80	−0.75	0.05	0.36
chlorobenzene	1.11	1.14	0.03	9.97	10.13	0.16	−1.12	−1.11	0.01	0.08
bromobenzene	1.5	1.47	0.03	10.65	11.98	1.33	−1.46	−1.40	0.06	5.33
iodobenzene	1.83	1.84	0.01	11.85	12.43	0.58	−1.83	−1.97	0.14	1.11
1h-pyrrole	0.97	0.99	0.02	10.78	10.87	0.09	−4.78	−3.74	1.04	5.43
6-chloropyridin-3-ol	1.39	1.36	0.03	14.81	15.36	0.55	−6.73	−6.32	0.41	1.75
6-chloropyridin-3-amine	1.33	1.29	0.04	12.71	12.44	0.27	−5.60	−5.47	0.13	0.30
4-chlorophenol	1.22	1.25	0.03	11.24	10.46	0.78	−7.03	−7.14	0.11	1.89
4-chloroaniline	1.17	1.19	0.02	11.2	10.51	0.69	−5.90	−6.01	0.11	1.49
4-bromophenol	1.84	1.83	0.01	14.04	14.1	0.06	−5.85	−5.89	0.04	0.02
2-chloropyridine	1.2	1.21	0.01	10.18	9.93	0.25	−4.39	−4.57	0.18	0.35
4-fluorophenol	1.31	1.32	0.01	10.43	10.77	0.34	−6.19	−5.66	0.53	1.75
4-fluoroaniline	1.17	1.18	0.01	10.16	9.43	0.73	−5.06	−5.28	0.22	1.84
4-fluoro-*n*-methylaniline	1.04	1.08	0.04	9.98	10.06	0.08	−4.26	−4.88	0.62	1.94
n-methylacetamide	0.94	1	0.06	14.2	14.11	0.09	−10.08	−9.99	0.09	0.07
average deviation			0.03			0.40			0.24	

[a]The absolute deviation is also reported (|devl). Experimental values of ρ and ΔH_{vap} were taken from Pubchem,[41] and values of ΔG_{hyd} from the FreeSolv database.[42,43] See Table S1 from the SI for an extended version.

from a previous optimization of a similar compound as in Table 4 (second row), where the latter were transferred from previous optimizations of PhCl and 4-BrPhOH to 4-ClPhOH and yield a ΔG_{hyd} of −6.65 kcal/mol that only differs by 0.38 kcal/mol from the experimental value. The score for the plain CGenFF parametrization reduces by a factor of 6 upon optimization of the PC model but essentially remains unchanged in the next few refinements. Only when both, MTP and LJ parameters, are optimized the score improves by almost 2 orders of magnitude and excellent agreement with experiment is obtained. This highlights that not all chemical building blocks may need the same level of parameter optimization and for some systems good and computationally inexpensive PC-based parametrizations can be obtained.

BROADER PARAMETRIZATION STUDY

Additional halogenated and substituted benzenes were parametrized along the same protocol and all results for ρ, ΔH_{vap} and ΔG_{hyd} are summarized and discussed in the following. Table 6 and Figure 3 compares the free energy of hydration (ΔG_{hyd}^{Calc}) as calculated using the FW and compares them to experimental data (ΔG_{hyd}^{Exp}).

The agreement between computed and observed ΔG_{hyd} is excellent. Over a range of 8 kcal/mol, the RMSE is 0.36 kcal/mol and R^2 = 0.99, see Figure 3. As a comparison, in a study of the solvation free energies of amino acid side chains the RMSE for ΔG_{hyd} using TIP3P water and the OPLS-AA force field was 0.79 kcal/mol with an R^2 = 0.93 which changed to 0.51 kcal/mol and R^2 = 0.94 upon modification of the LJ parameters of the TIP3P water model.[70] In a more recent study focusing on 40 small organic molecules and charges from atoms-in-molecules electron density partitioning, using environment-specific charges and LJ parameters from quantum chemical calculations, the mean unsigned errors relative to experiment are 0.014 g/cm^3 for the density ρ, 0.65 kcal/mol for the heat of vaporization ΔH_{vap} and 1.03 kcal/mol over a range of 12 kcal/mol for ΔG_{hyd}.[56] In yet another, broader study of 239 molecules, the mean unsigned error for ΔG_{hyd} was 1.93 kcal/mol (CHARMM), 1.17 kcal/mol (GAFF) and 0.73 kcal/mol (OPLS2.1).[71]

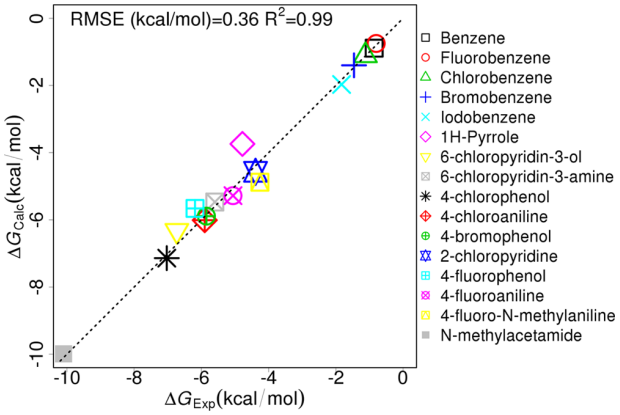


Figure 3. Correlation between experimental and computed solvation free energies ΔG_{hyd} (kcal/mol, respectively, *x*- and *y*-axis) for a range of compounds of interest. Computed values are obtained after optimization of the LJ parameters.

For ΔH_{vap} the RMSE (estimated for the same family of compounds than for ΔG_{hyd}) is 0.53 kcal/mol with an R^2 = 0.97, see Figure 4. This compares with 0.65 kcal/mol from a recent study on a different set of small molecules.[56] For the pure liquid density (see Figure S2 from the SI) the current study yields an RMSE of 0.02 g/cm^3 with an R^2 = 0.99, compared with an RMSE of 0.01 g/cm^3 of the same recent parametrization work.[56] Hence, for a range of compounds the fitting environment presented here yields comparable if not superior performance based on a user-friendly interface.

For an extended version of Table 6, including also compounds for which one or more of the experimental references are missing, see SI section II Table S1.

OUTLOOK AND PERSPECTIVES

The present work introduces a graphics-based, versatile, and extensible fitting environment for PC- and MTP-based force fields for condensed-phase simulations. It is demonstrated that

DOI: 10.1021/acs.jcim.6b00280
J. Chem. Inf. Model. 2016, 56, 1479−1489

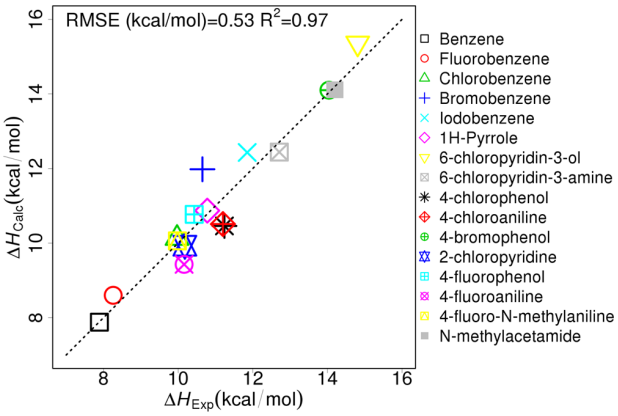


Figure 4. Correlation between experimental and computed enthalpy of vaporization ΔH_{vap} (kcal/mol, respectively, *x*- and *y*-axis) for a range of compounds of interest. Both, MTP and LJ parameters were optimized.

accurate parametrizations for solution-phase simulations can be obtained and that medium-scale (several 10 molecules) parametrization tasks can be routinely undertaken as a typical parametrization takes a few hours for a molecule the size of NMA. Within the chemical space covered, the transferability of parametrizations yields results well within chemical accuracy.

The fitting environment can be easily adapted to different and higher levels of theory for the reference data from electronic structure calculations. Also, extension to other molecular dynamics codes (AMBER, GROMACS, TINKER) is possible because of the modular architecture of the software provided that multipolar interactions can be computed. For molecules exhibiting two or more linked ring systems (e.g., biphenyl) it will be important to consider refitting dihedral parameters because of multipole-multipole interactions between atoms on different ring systems.

Currently, thermodynamic properties (ρ, ΔH_{vap}, ΔG_{hyd}) are used to improve the force field. This can be easily extended to additional interesting (and experimentally accessible) quantities such as diffusion coefficients D, or heat capacities C_p. Also, infrared and NMR spectroscopic data may be of interest in the future.[72−75]

A valuable extension will be the computation of derivatives $\frac{d}{dp}\langle A\rangle_p$ of observables A with respect to the LJ parameters p from suitable ensemble averages. This has recently been done for the parametrization of the iAMOEBA force field for water.[76] It will be of interest to assess whether a grid-based search as proposed here or a gradient-based approach to improve parameter values converges more rapidly in concrete applications.

Furthermore, it was found in recent work that averaging over a number of conformations can yield meaningful parametrizations of conformationally dependent multipoles.[30] Including such effects should further improve transferability of the parametrizations. A final asset is the storage and retrieval of particular parametrizations for validated simulation and parametrization conditions, in particular for chemically and pharmaceutically important molecular fragments. If the transferability of the parametrizations can be ascertained, this will allow simple assembly of larger molecules from well-parametrized building blocks (molecules) and considerably speed up future parametrizations.

In summary, the present work describes a user-friendly, graphics-based interface for the parametrization of multipolar force fields for quantitative atomistic simulations of small molecular building blocks.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00280.

Details concerning the database design, construction and curation; supplementary correlation plot for density ρ (similar to Figures 3 and 4); extended version of Table 6 including compounds for which experimental data is incomplete (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: m.meuwly@unibas.ch.

Author Contributions

§F.H. and K.E.H. contributed equally to this work.

Funding

We acknowledge financial support from the Swiss National Science Foundation (Grant 200020-132406) and the NCCR MUST (to M.M.).

Notes

The authors declare no competing financial interest.

This software is available free of charge, and under the 3-clause BSD license, from the following github repository: https://github.com/MMunibas/FittingWizard. Archived releases can be downloaded from https://github.com/MMunibas/FittingWizard/releases, and documentation and instructions are available from https://github.com/MMunibas/FittingWizard/wiki.

■ REFERENCES

(1) Ponder, J. W.; Case, D. A. Force Fields for Protein Simulations. *Adv. Protein Chem.* **2003**, *66*, 27−85.

(2) Bereau, T.; Kramer, C.; Meuwly, M. Leveraging Symmetries of Static Atomic Multipole Electrostatics in Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2013**, *9*, 5450−5459.

(3) Simmonett, A. C.; Pickard, F. C., IV; Schaefer, H. F., III; Brooks, B. R. An Efficient Algorithm for Multipole Energies and Derivatives Based on Spherical Harmonics and Extensions to Particle Mesh Ewald. *J. Chem. Phys.* **2014**, *140*, 184101.

(4) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A., Jr; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **2010**, *114*, 2549−2564.

(5) Lamoureux, G.; MacKerell, A. D., Jr; Roux, B. A Simple Polarizable Model of Water Based on Classical Drude Oscillators. *J. Chem. Phys.* **2003**, *119*, 5185−5197.

(6) Lin, B.; Lopes, P. E. M.; Roux, B.; MacKerell, A. D., Jr. Kirkwood-Buff Analysis of Aqueous N-methylacetamide and Acetamide Solutions Modeled by the CHARMM Additive and Drude Polarizable Force Fields. *J. Chem. Phys.* **2013**, *139*, 084509.

(7) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(8) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(9) MacKerell, A. D., Jr; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe,

DOI: 10.1021/acs.jcim.6b00280
J. Chem. Inf. Model. 2016, 56, 1479−1489

M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(10) MacKerell, A. D. Empirical Force Fields for Biological Macromolecules: Overview and Issues. *J. Comput. Chem.* **2004**, *25*, 1584−1604.
(11) Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D., Jr. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.* **2012**, *52*, 3155−3168.
(12) Yesselman, J. D.; Price, D. J.; Knight, J. L.; Brooks, C. L., III MATCH: An Atom-Typing Toolset for Molecular Mechanics Force Fields. *J. Comput. Chem.* **2012**, *33*, 189−202.
(13) Malde, A. K.; Zuo, L.; Breeze, M.; Stroet, M.; Poger, D.; Nair, P. C.; Oostenbrink, C.; Mark, A. E. An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0. *J. Chem. Theory Comput.* **2011**, *7*, 4026−4037.
(14) Zoete, V.; Cuendet, M. A.; Grosdidier, A.; Michielin, O. SwissParam: A Fast Force Field Generation Tool for Small Organic Molecules. *J. Comput. Chem.* **2011**, *32*, 2359−2368.
(15) Halgren, T. A. Merck molecular force field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(16) Halgren, T. A. Merck molecular force field. II. MMFF94 van der Waals and Electrostatic Parameters for Intermolecular Interactions. *J. Comput. Chem.* **1996**, *17*, 520−552.
(17) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247−260.
(18) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(19) Mayne, C. G.; Saam, J.; Schulten, K.; Tajkhorshid, E.; Gumbart, J. C. Rapid Parameterization of Small Molecules Using the Force Field Toolkit. *J. Comput. Chem.* **2013**, *34*, 2757−2770.
(20) Vanommeslaeghe, K.; MacKerell, A. D., Jr. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.* **2012**, *52*, 3144−3154.
(21) Huang, L.; Roux, B. Automated Force Field Parameterization for Nonpolarizable and Polarizable Atomic Models Based on Ab Initio Target Data. *J. Chem. Theory Comput.* **2013**, *9*, 3543−3556.
(22) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; J, A. M., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision B.01; 2010.
(23) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545−1614.
(24) Head-Gordon, M.; Pople, J. A.; Frisch, M. MP2 Energy Evaluation by Direct Methods. *Chem. Phys. Lett.* **1988**, *153*, 503.
(25) Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron Through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007−1023.
(26) Woon, D. E.; Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. III. The Atoms Aluminum Through Argon. *J. Chem. Phys.* **1993**, *98*, 1358−1371.
(27) Wilson, A. K.; Woon, D. E.; Peterson, K. A.; Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. IX. The Atoms Gallium Through Krypton. *J. Chem. Phys.* **1999**, *110*, 7667−7676.
(28) Stone, A. J. Distributed Multipole Analysis: Stability for Large Basis Sets. *J. Chem. Theory Comput.* **2005**, *1*, 1128−1132.
(29) Kramer, C.; Bereau, T.; Spinn, A.; Liedl, K. R.; Gedeck, P.; Meuwly, M. Deriving Static Atomic Multipoles from the Electrostatic Potential. *J. Chem. Inf. Model.* **2013**, *53*, 3410−3417.
(30) Kramer, C.; Gedeck, P.; Meuwly, M. Atomic Multipoles: Electrostatic Potential Fit, Local Reference Axis Systems and Conformational Dependence. *J. Comput. Chem.* **2012**, *33*, 1673−1688.
(31) Stone, A. J. The Description of Bimolecular Potentials, Forces and Torques: The S and V Function Expansions. *Mol. Phys.* **1978**, *36*, 241−256.
(32) Price, S. L.; Stone, A. J.; Alderton, M. Explicit Formulae for the Electrostatic Energy, Forces and Torques Between a Pair of Molecules of Arbitrary Symmetry. *Mol. Phys.* **1984**, *52*, 987−1001.
(33) Koch, U.; Popelier, P. L. A.; Stone, A. J. Conformational Dependence of Atomic Multipole Moments. *Chem. Phys. Lett.* **1995**, *238*, 253−260.
(34) Stone, A. J. *The Theory of Intermolecular Forces*; Clarendon Press Oxford: Oxford, UK, 1996; Vol. *32*.
(35) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An Nlog(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(36) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(37) van Gunsteren, W.; Berendsen, H. Algorithms for Macromolecular Dynamics and Constraint Dynamics. *Mol. Phys.* **1977**, *34*, 1311−1327.
(38) Wang, J.; Hou, T. Application of Molecular Dynamics Simulations in Molecular Property Prediction. 1. Density and Heat of Vaporization. *J. Chem. Theory Comput.* **2011**, *7*, 2151−2165.
(39) Zacharias, M.; Straatsma, T. P.; McCammon, J. A. Separation-Shifted Scaling, a New Scaling Method for Lennard-Jones Interactions in Thermodynamic Integration. *J. Chem. Phys.* **1994**, *100*, 9025−9031.
(40) Boresch, S. The Role of Bonded Energy Terms in Free Energy Simulations - Insights from Analytical Results. *Mol. Simul.* **2002**, *28*, 13−37.
(41) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44*, D1202−D1213.
(42) Mobley, D. L. Experimental and Calculated Small Molecule Hydration Free Energies. http://escholarship.org/uc/item/6sd403pz (accessed May 18, 2016).
(43) Mobley, D. L.; Guthrie, J. P. FreeSolv: A Database of Experimental and Calculated Hydration Free Energies, with Input Files. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 711−720.
(44) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.
(45) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Model.* **1989**, *29*, 97−101.
(46) Chamberlin, D. D.; Boyce, R. F. SEQUEL: A Structured English Query Language. *Proceedings of the 1974 ACM SIGFIDET (Now SIGMOD) Workshop on Data Description, Access and Control*; New York, 1974; pp 249−264.
(47) PubChem REST API. https://pubchem.ncbi.nlm.nih.gov/pug_rest/PUG_REST.html (accessed: May 18, 2016).
(48) Cazade, P.-A.; Bereau, T.; Meuwly, M. Computational Two-Dimensional Infrared Spectroscopy without Maps: N-Methylacetamide in Water. *J. Phys. Chem. B* **2014**, *118*, 8135−8147.

DOI: 10.1021/acs.jcim.6b00280
J. Chem. Inf. Model. 2016, 56, 1479−1489

(49) MacKerell, A. D., Jr.; Shim, J. H.; Anisimov, V. M. Re-evaluation of the Reported Experimental Values of the Heat of Vaporization of N-methylacetamide. *J. Chem. Theory Comput.* **2008**, *4*, 1307−1312.
(50) *CRC Handbook of Chemistry and Physics*, 96th ed.; CRC press: Boca Raton, FL, 2015.
(51) Riddick, J. A.; Bunger, W. B.; Sakano, T. K. *Physical Properties and Methods of Purification*, 4th ed.; John Wiley & Sons: New York, NY, 1985; Vol. *II* Organic Solvents; p 660.
(52) Abraham, M. H.; Andonian-Haftvan, J.; Whiting, G. S.; Leo, A.; Taft, R. S. Hydrogen Bonding. Part 34. The Factors that Influence the Solubility of Gases and Vapours in Water at 298 K, and a New Method for its Determination. *J. Chem. Soc., Perkin Trans. 2* **1994**, *2*, 1777−1791.
(53) Wolfenden, R. Interaction of the Peptide Bond with Solvent Water: a Vapor Phase Analysis. *Biochemistry* **1978**, *17*, 201−204.
(54) Radzicka, A.; Pedersen, L.; Wolfenden, R. Influences of solvent water on protein folding: free energies of solvation of cis and trans peptides are nearly identical. *Biochemistry* **1988**, *27*, 4538−4541.
(55) Jorgensen, W. L.; Gao, J. Cis trans energy difference for the peptide-bond in the gas-phase and in aqueous-solution. *J. Am. Chem. Soc.* **1988**, *110*, 4212−4216.
(56) Cole, D. J.; Vilseck, J. Z.; Tirado-Rives, J.; Payne, M. C.; Jorgensen, W. L. Biomolecular Force Field Parameterization via Atoms-in-Molecule Electron Density Partitioning. *J. Chem. Theory Comput.* **2016**, *12*, 2312−2323.
(57) Ding, Y.; Bernardo, D. N.; Krogh-Jespersen, K.; Levy, R. M. Solvation Free Energies of Small Amides and Amines from Molecular Dynamics/Free Energy Perturbation Simulations Using Pairwise Additive and Many-Body Polarizable Potentials. *J. Phys. Chem.* **1995**, *99*, 11575−11583.
(58) Rick, S. W.; Berne, B. J. Dynamical Fluctuating Charge Force Fields: The Aqueous Solvation of Amides. *J. Am. Chem. Soc.* **1996**, *118*, 672−679.
(59) Spector, T. I.; Kollman, P. A. Investigation of the Anomalous Solvation Free Energies of Amides and Amines: FEP Calculations in Cyclohexane and PS-GVB Calculations on Amide-Water Complexes. *J. Phys. Chem. B* **1998**, *102*, 4004−4010.
(60) Lu, Y.; Shi, T.; Wang, Y.; Yang, H.; Yan, X.; Luo, X.; Jiang, H.; Zhu, W. Halogen Bonding − A Novel Interaction for Rational Drug Design? *J. Med. Chem.* **2009**, *52*, 2854−2862.
(61) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Joerger, A. C.; Boeckler, F. M. Principles and Applications of Halogen Bonding in Medicinal Chemistry and Chemical Biology. *J. Med. Chem.* **2013**, *56*, 1363−1388.
(62) Hernandes, M. Z.; Cavalcanti, S. M. T.; Moreira, D. R. M.; de Azevedo, W.; Filgueira, W.; Leite, A. C. L. Halogen Atoms in the Modern Medicinal Chemistry: Hints for the Drug Design. *Curr. Drug Targets* **2010**, *11*, 303−314.
(63) Müller, K.; Faeh, C.; Diederich, F. Fluorine in Pharmaceuticals: Looking Beyond Intuition. *Science* **2007**, *317*, 1881−1886.
(64) Pandyarajan, V.; Phillips, N. B.; Cox, G. P.; Yang, Y.; Whittaker, J.; Ismail-Beigi, F.; Weiss, M. A. Biophysical Optimization of a Therapeutic Protein by Nonstandard Mutagenesis: Studies of an iodo-insulin derivative. *J. Biol. Chem.* **2014**, *289*, 23367−23381.
(65) Parsons, G. H.; Rochester, C. H.; Wood, C. E. C. Effect of 4-substitution on the Thermodynamics of Hydration of Phenol and the Phenoxide Anion. *J. Chem. Soc. B* **1971**, 533−536.
(66) Rizzo, R. C.; Aynechi, T.; Case, D. A.; Kuntz, I. D. Estimation of Absolute Free Energies of Hydration Using Continuum Methods: Accuracy of Partial Charge Models and Optimization of Nonpolar Contributions. *J. Chem. Theory Comput.* **2006**, *2*, 128−139.
(67) Cazade, P.-A.; Tran, H.; Bereau, T.; Das, A. K.; Kläsi, F.; Hamm, P.; Meuwly, M. Solvation of Fluoro-Acetonitrile in Water by 2D-IR Spectroscopy: A Combined Experimental-Computational Study. *J. Chem. Phys.* **2015**, *142*, 212415.
(68) Williams, M. The Merck Index: an Encyclopedia of Chemicals, Drugs, and Biologicals. 14th Edition. *Drug Dev. Res.* **2006**, *67*, 870−870.
(69) Yaws, C. L. *Chemical Properties Handbook: Physical, Thermodynamic, Environmental, Transport, Safety, and Health Related Properties for Organic and Inorganic Chemicals*; McGraw-Hill: New-York, NY, 1999.
(70) Shirts, M. R.; Pande, V. S. Solvation Free Energies of Amino Acid Side Chain Analogs for Common Molecular Mechanics Water Models. *J. Chem. Phys.* **2005**, *122*, 134508.
(71) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137*, 2695−2703.
(72) Lee, M. W.; Meuwly, M. On the Role of Nonbonded Interactions in Vibrational Energy Relaxation of Cyanide in Water. *J. Phys. Chem. A* **2011**, *115*, 5053−5061.
(73) Schmid, F. F.-F.; Meuwly, M. Direct Comparison of Experimental and Calculated NMR Scalar Coupling Constants for Force Field Validation and Adaptation. *J. Chem. Theory Comput.* **2008**, *4*, 1949−1958.
(74) Huang, J.; Meuwly, M. Explicit Hydrogen-Bond Potentials and Their Application to NMR Scalar Couplings in Proteins. *J. Chem. Theory Comput.* **2010**, *6*, 467−476.
(75) Huang, J.; MacKerell, A. D. CHARMM36 All-Atom Additive Protein Force Field: Validation Based on Comparison to NMR Data. *J. Comput. Chem.* **2013**, *34*, 2135−2145.
(76) Wang, L.-P.; Head-Gordon, T.; Ponder, J. W.; Ren, P.; Chodera, J. D.; Eastman, P. K.; Martinez, T. J.; Pande, V. S. Systematic Improvement of a Classical Molecular Model of Water. *J. Phys. Chem. B* **2013**, *117*, 9956−9972.

NOTE ADDED AFTER ASAP PUBLICATION

Three additional citations to various references were added to the text in the version published ASAP August 5, 2016. The corrected paper was published ASAP on August 10, 2016.

DOI: 10.1021/acs.jcim.6b00280
J. Chem. Inf. Model. 2016, 56, 1479−1489

Supporting information:

A Toolkit to Fit Nonbonded Parameters from and for Condensed Phase Simulations

Florent Hédin,[†,¶] Krystel El Hage,[†,¶] and Markus Meuwly[*,†,‡]

†*Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland*

‡*Department of Chemistry, Brown University, Providence, Rhode Island 02912, USA.*

¶*Contributed equally to this work*

E-mail: m.meuwly@unibas.ch

SI for subsection "Database of compounds"

The database of compounds, described in the article, was implemented using the SQL language, and the **MariaDB** (https://mariadb.org/) open source software, version **10.0.23**. The Fitting Wizard software provides an access to such a database, hosted on a server. As explained in the article, properties of interest either come from **PubChem** (https://pubchem.ncbi.nlm.nih.gov/search/)[S1] or the FreeSolv database[S2,S3].

Figure S1 shows the different data tables, their fields, and the relations between them:

- Table **compounds** is the master table, the integer field *id* is the primary key used for retrieving data from all the other linked tables. The integer field *idPubchem* is another

index, which corresponds to the identifier of a given compound in the external PubChem database. The other fields *name*, *added*, *lastUpdate* respectively correspond to the IUPAC name of the compound, date of addition of the compound to the database, and date of last modification of the compound.

- Table **structure** contains structural information about a compound, identified by an *id* field which is linked to the *id* field of **compounds**. Fields *formula*, *inchi* and *smiles* respectively correspond to the molecular formula, IUPAC International Chemical Identifier (INChI), and Simplified Molecular-Input Line-Entry System (SMILES).

- Table **prop** (for properties) contains molecular properties of interest, identified by an *id* field which is linked to the *id* field of **compounds**. Current version of the database includes fields *mass*, *density*, *Hvap* (ΔH_{vap}) and *Gsolv* (ΔG_{hyd}).

- Table **ref** (for references) contains literature references for experimental measurements of ΔH_{vap} and ΔG_{hyd}, identified by an *id* field which is linked to the *id* field of **compounds**. The fields *ref_dg* and *ref_dh* provide a Digital object identifier (DOI) reference pointing to the publication of interest, when available.

- Tables **ff** (for forcefield), **par** (for parameters) and **top** for topology, are designed for linking each compound to an optimal Molecular Dynamics forcefield. This feature, not fully implemented for the moment, would allow the user to save in the database, after optimisation using the Wizard, a version of the optimised Force Field parameters. With such a collaborative behaviour, a large set of optimised parameters for fragments of interest would be available for the community, avoiding the need to re-parametrise a fragment if another user already published convincing results.

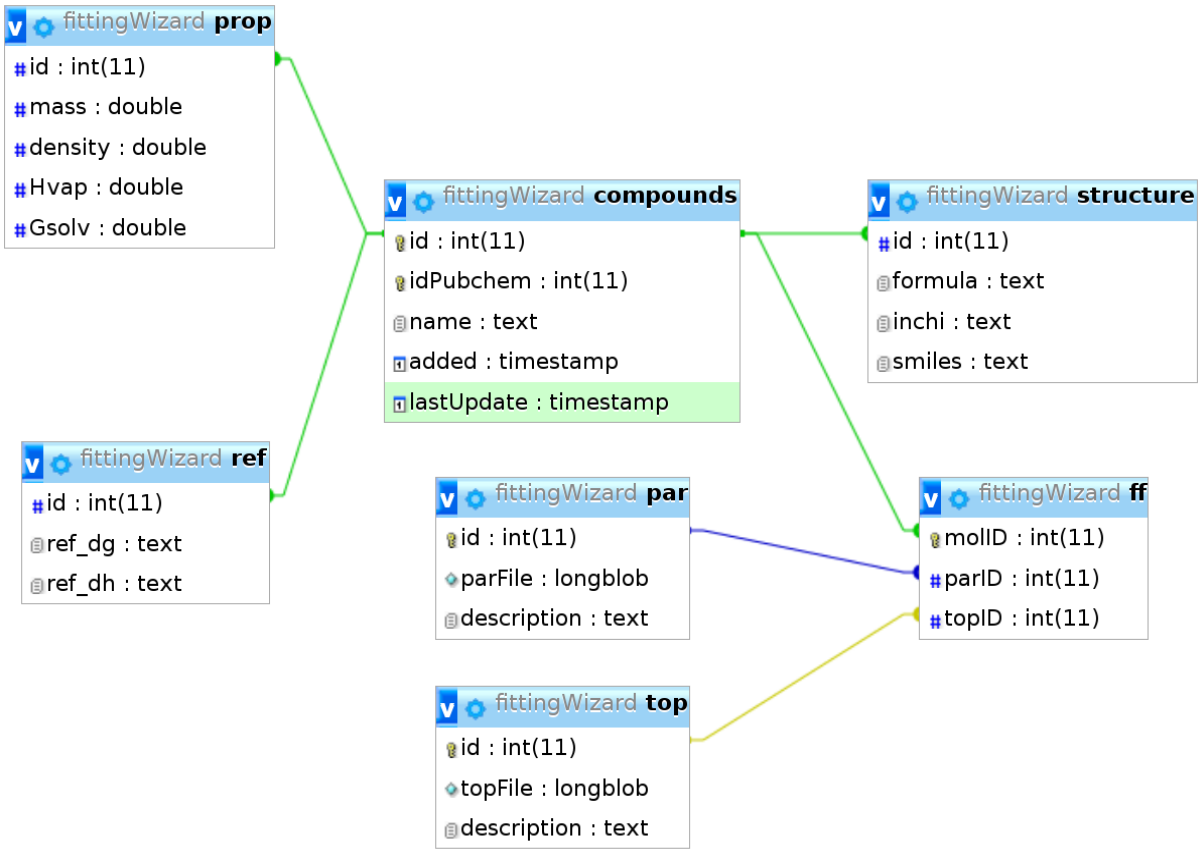


Figure S1: Description of the SQL tables in the compounds database, and the relationships between them. The key and # pictograms describe a primary key or an index. Coloured arrows correspond to relationships between the tables (e.g. foreign key constraints).

SI for subsection "A Broader Parametrization Study"

Table S1 provides values for ΔH_{vap}, ΔG_{hyd} (kcal/mol) and ρ (g/cm^3) as calculated using the FW (Calc.) with optimized MPT and LJ parameters compared to experimental references (Exp.), when available. The absolute deviation is also reported (|Dev.|). Experimental values of ρ and ΔH_{vap} were taken from Pubchem[S1], and values of ΔG_{hyd} from the FreeSolv database[S2,S3]. The scoring function S was calculated using $S = \sum_{i=1}^{3} w_i(\mathrm{Obs_i} - \mathrm{Calc_i})^2$ with $w_\rho = 1$, $w_{\Delta H} = 3$ and $w_{\Delta G} = 5$.

Figure S2 shows the correlation between experimentally measured and calculated (this study) values of the density ρ (g/cm^3), for compounds from Table S1 for which an experimental value is available. An excellent correlation is obtained, with an RMSE of 0.02 g/cm^3 and with a correlation coefficient $R^2 = 0.99$.

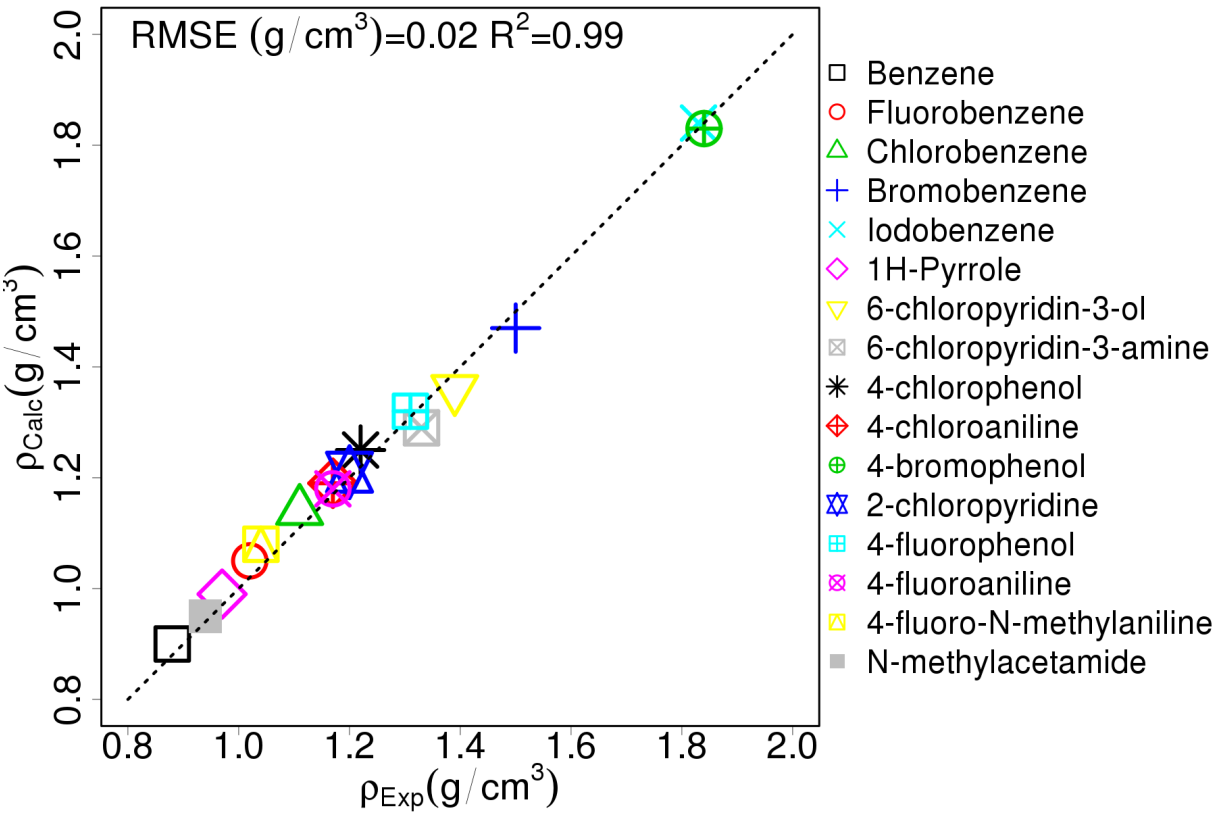


Figure S2: Correlation between experimental and computed density ρ (g/cm^3, respectively, $x-$axis and $y-$axis) for a range of compounds of interest. Both, MTP and LJ parameters were optimized.

Table S1: ΔH_{vap}, ΔG_{hyd} (kcal/mol) and ρ (g/cm^3) as calculated using the wizard (Calc.) compared to experimental references (Exp.) when available. The absolute deviation is also reported (|Dev.|)

	ρ			ΔH_{vap}			ΔG_{hyd}			Score S
	Exp.[S1]	Calc.	|Dev.|	Exp.[S1]	Calc.	|Dev.|	Exp.[S2,S3]	Calc.	|Dev.|	
Toluene	0.86	1.03	0.17	—	7.99	—	–0.9	–0.86	0.04	—
trifluoromethylbenzene	1.19	1.08	0.11	—	10.35	—	–0.25	–0.34	0.09	—
2-iodophenol	—	—	—	—	—	—	–6.2	–6.5	0.3	—
Benzene	0.88	0.9	0.02	7.89	7.88	0.01	–0.86	–0.89	0.03	0.01
Fluorobenzene	1.02	1.05	0.03	8.26	8.6	0.34	–0.8	–0.75	0.05	0.36
Chlorobenzene	1.11	1.14	0.03	9.97	10.13	0.16	–1.12	–1.11	0.01	0.08
Bromobenzene	1.5	1.47	0.03	10.65	11.98	1.33	–1.46	–1.4	0.06	5.33
Iododbenzene	1.83	1.84	0.01	11.85	12.43	0.58	–1.83	–1.97	0.14	1.11
1H-Pyrrole	0.97	0.99	0.02	10.78	10.87	0.09	–4.78	–3.74	1.04	5.43
6-bromopyridin-3-ol	1.79	1.81	0.02	13.05	13.72	0.67	—	–6.85	—	—
6-bromopyridin-3-amine	1.71	1.68	0.03	12.67	12.13	0.54	—	–6.17	—	—
6-bromo-N-methylpyridin-3-amine	1.58	1.57	0.01	12.17	11.6	0.57	—	–5.68	—	—
6-chloropyridin-3-ol	1.39	1.36	0.03	14.81	15.36	0.55	–6.73	–6.32	0.41	1.75
6-chloropyridin-3-amine	1.33	1.29	0.04	12.71	12.44	0.27	–5.6	–5.47	0.13	0.30
6-chloro-N-methylpyridin-3-amine	1.25	1.21	0.04	12.28	11.6	0.68	—	–5.61	—	—
4-chlorophenol	1.22	1.25	0.03	11.24	10.46	0.78	–7.03	–7.14	0.11	1.89
4-chloroaniline	1.17	1.19	0.02	11.2	10.51	0.69	–5.9	–6.01	0.11	1.49
N-methyl-(4-chlorophenyl)amine	1.2	1.2	0	11.4	12.08	0.68	—	–5.1	—	—
4-bromo-N-methylaniline	1.66	1.64	0.02	11.22	11.5	0.28	—	–6.29	—	—
4-bromoaniline	1.59	1.55	0.04	11.74	12.38	0.64	—	–5.68	—	—
4-bromophenol	1.84	1.83	0.01	14.04	14.1	0.06	–5.85	–5.89	0.04	0.02
2-fluoropyridine	1.13	1.17	0.04	8.22	8.67	0.45	—	–4.08	—	—
2-chloropyridine	1.2	1.21	0.01	10.18	9.93	0.25	–4.39	–4.57	0.18	0.35
2-bromopyridine	1.66	1.64	0.02	9.95	9.53	0.42	—	–4.72	—	—
4-fluorophenol	1.31	1.32	0.01	10.43	10.77	0.34	–6.19	–5.66	0.53	1.75
4-fluoroaniline	1.17	1.18	0.01	10.16	9.43	0.73	–5.06	–5.28	0.22	1.84
4-fluoro-N-methylaniline	1.04	1.08	0.04	9.98	10.06	0.08	–4.26	–4.88	0.62	1.94
N-methylacetamide	0.94	1	0.06	14.2	14.11	0.09	–10.08	–9.99	0.09	0.07

References

(S1) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Research* **2016**, *44*, D1202–D1213.

(S2) Mobley, D. L. Experimental and Calculated Small Molecule Hydration Free Energies. 2013; `http://escholarship.org/uc/item/6sd403pz`, Accessed: 2016-05-18.

(S3) Mobley, D. L.; Guthrie, J. P. FreeSolv: A Database of Experimental and Calculated Hydration Free Energies, with Input Files. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 711–720.

Chapter 6

Stability of solvated Hæmoglobin Tetramers

"I originally implemented PME to prove that you didn't need it... "

Erik Lindahl, GROMACS head author and project leader

Hæmoglobin is a *metalloprotein* (containing iron), in charge of oxygen transport in red blood cells of animals. It transports oxygen from the lungs to the rest of the body where it releases the oxygen for cell use. Human hæmoglobin is a tetrameric protein (see Fig. 6.1) consisting of two α and two β subunits.

The α and β subunits are structurally identical, consisting of 141 and 146 amino acids residues, respectively. Each subunit contains a heme group at the center to which molecular oxygen or other ligands bind: the ligand-bound (*oxy*) state is identified as *R-state* or *2DN3*, and the ligand-free (*deoxy*) state as *T-state* or *2DN2*.

The distance between the two terminal Histidines of the β chains (see Fig. 6.1) is characteristic of each state: this distance fluctuates between 10 – 15 Å for the compact oxy (2DN3) state, and 30 – 35 Å for the more extended deoxy (2DN2) state.

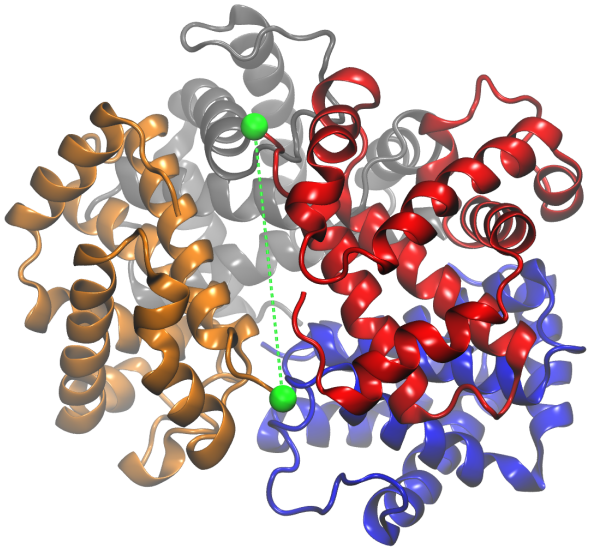


FIGURE 6.1: Tetrameric hæmoglobin. In green, distance between two C-α atoms of the 2 Histidine residues present at the terminal of 2 β chains. The initial structure is characterised by a distance of $\sim$ 30 Å for the deoxy 2DN2 state.

The structural stability and dynamics of oxy and deoxy tetrameters can be investigated using Molecular Dynamics (MD) Methods (see Section 2.2.1). CHARMM [146] and GROMACS [147] implement modern molecular mechanics algorithms, such as the Particle-Mesh Ewald [28] summation, and Domain Decomposition approach for parallelisation, appropriate when considering long simulations of a large biological system.

While the T state is expected to be stable in water, it was already reported [148, 149] that it is not the case for Molecular Dynamics simulations, where a $T \rightarrow R$ quaternary structure rearrangement is observed. As this result was obtained from different Force Fields, it is highly improbable that this comes from an error in the FF parametrisation. Preliminary studies were performed in the M. Meuwly group by Prashant Gupta (unpublished work), using CHARMM, for up to 100 ns of simulation. From those simulations, it appeared that one of the possible cause of the observed instability might be a lack of water around the tetramer for simulation boxes not built large enough. This causes larger *water density fluctuations* at the protein–solvent interface than what is usually observed. Such fluctuations were already investigated by the introduction of a *coarse grained density* measure.[150, 151]

This Chapter is organised as following: in Section 6.1 the computational parameters of the simulations are detailed. In Section 6.2 several conformational analyses are performed: measure of distance or angles between the four units, measure of hydrogen bonds at protein-protein interfaces,…In Section 6.3, a coarse grained water density is introduced, and the fluctuations of this property are investigated.

Therefore, although the content of this Chapter is not yet ready for publication as open questions still remain, it can be seen as the basis of a future manuscript.

6.1 Setup

In all the following the GROMACS MD software was used, version 5.1.1 and newer. However the Force Field parameters were taken from the CHARMM FF version c36, and downloaded from A. Mackerell Lab's website.[1]

The oxy crystal structure pdb files were provided by a collaborator of Martin Karplus, and were solvated in 3 cubic boxes of increasing size (see Fig.6.2): 75 Å, 120 Å and 150 Å : the number of atoms is respectively 39432, 163480 and 318911 for each box size. Na^+ and Cl^- ions were added for assuring a physiological concentration of approximately 0.15 mol/L, like in blood/cells. Solvated system was minimised, equilibrated in the NVT ensemble for 0.5 ns, and finally in the NPT for also 0.5 ns.

The LINCS[152] algorithm was used for constraining bonds involving hydrogen, in order to choose a timestep of $\delta t = 2\ fs$. Simulations run in the NPT ensemble at 1 bar and 300 K: the thermostat is a modified version of Berendsen[153] called "V-rescale" which is guaranteed to properly sample the NVT or NPT ensembles, and the Parrinello-Rahman barostat is used.[68–70]

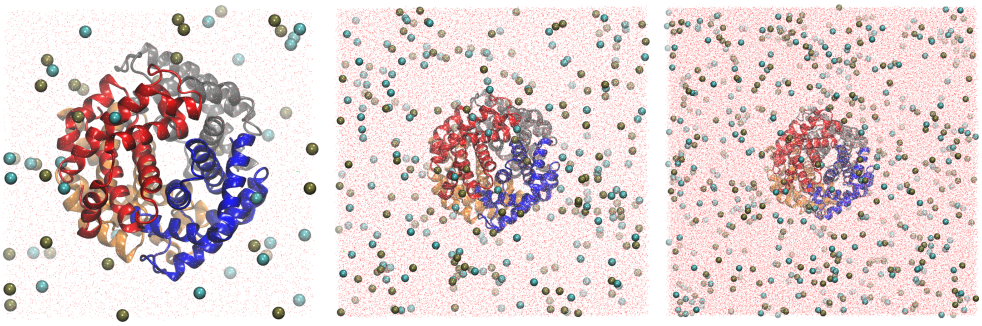


FIGURE 6.2: Tetrameric hæmoglobin solvated in boxes of size: (Left) 75 Å, (Center) 120 Å and (Right) 150 Å. Spheres correspond to Na+ and Cl- ions, added for neutralising the system and reaching a biological salt concentration of 0.15 mol/L.

6.2 Conformational analyses

Histidine distances First a simple distance analysis was performed, where the distance between two C-α atoms of the Histidine residues at the end of the 2 β chains is measured (see Fig. 6.1).

For the smaller size box (75 Å, see Fig. 6.4 bottom, black) it was found that after 20 to 30 ns the deoxy states decays to a compact structure. For the medium size box (120 Å, see Fig. 6.4 bottom, red), the tetramer remains stable for approximately 50 to 70 ns and then reaches a lower-distance structure, but apparently not as compact as before. For the larger size box (150 Å, see Fig. 6.4 bottom, green), no large conformational change is observed i.e. the distance is just reduced by 3 to 4 Å on average.

All those observations lead to the following opened questions: (i) Are the instability of the ligand-free state an artifact caused by the size of the solvent box ? (ii) If yes, what is the minimum box size required for suppressing this artifact ? (iii) Can other parameters (such as the geometry or the box) have an influence on this effect ? (iv) Is there a real correlation between the density fluctuations at the protein–solvent interface and the size of the box ?

The following paragraphs try to clarify points (i) and (ii) of the previous opened questions.

Angle between $\alpha_1\beta_1$ and $\alpha_2\beta_2$ sub-units groups The 4 sub-units of Hæmoglobin tetramer can be observed on Fig. 6.3 : α_1 (cyan), β_1 (red), α_2 (green) and β_2 (orange). It was shown [154–156] that during the $\mathbf{T} \rightarrow \mathbf{R}$ transition the angle between the 2 blocks $\alpha_1\beta_1$ and $\alpha_2\beta_2$ changes by a value of approximately 15°.

[1] http://mackerell.umaryland.edu/charmm_ff.shtml#gromacs

In order to measure this effect for the MD simulations, the centre of mass of each chain is calculated. Then 2 vectors $\overrightarrow{\alpha_1\beta_1}$ and $\overrightarrow{\alpha_2\beta_2}$ are defined, and the angle θ between them is calculated using:

$$\theta = \frac{\overrightarrow{\alpha_1\beta_1} \cdot \overrightarrow{\alpha_2\beta_2}}{\|\overrightarrow{\alpha_1\beta_1}\| \times \|\overrightarrow{\alpha_2\beta_2}\|} \tag{6.1}$$

Fig. 6.4 (Top) shows the value of the θ angle measured for the 3 boxes size. Results confirm previous observations, i.e. that for the 75 Å box there is a clear $\mathbf{T} \rightarrow \mathbf{R}$ transition, and that only the 150 Å box appears to be stable on long term.

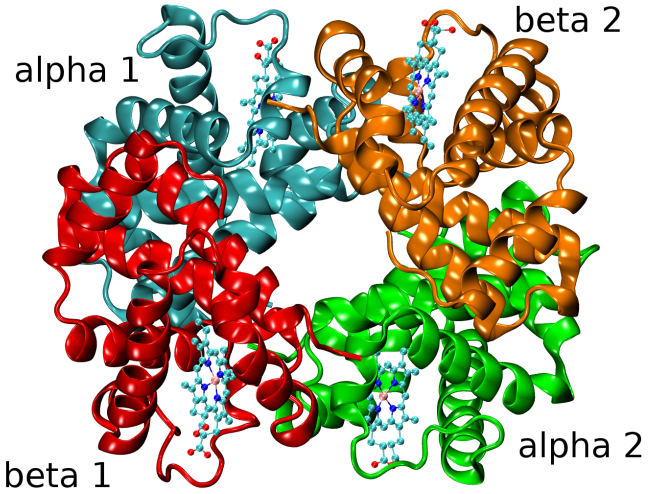


FIGURE 6.3: View of the 4 units of Hæmoglobin tetramer : α_1 (cyan), β_1 (red), α_2 (green) and β_2 (orange). The angle θ (see Eqn. 6.1) between 2 vectors connecting 2 of the centre of masses of the sub-units is characteristic from a given **T** or **R** state, as the $\mathbf{T} \rightarrow \mathbf{R}$ transition shows a variation of 10° to 15°

***Hinge* contacts at the $\alpha_1\beta_2$ interface** A study from Jones et al. [157] suggested the existence of special hydrogen bonds contacts, called "hinge" contacts, at the $\alpha_1\beta_2$ interface. The evolution of those *hinge* bonds is related to the $\mathbf{T} \rightarrow \mathbf{R}$ transition.

The following atoms are involved in the *hinge* contacts (see Figure 6.5): (i) atom HH from residue 42 and atom OD1 from residue 99, and (ii) atom HE1 from residue 37 and atom OD1 from residue 94.

In order to observe if a strengthening of those bonds may increase the time required before observing the $\mathbf{T} \rightarrow \mathbf{R}$ transition, or even prevent it, in the following the charges of the involved atoms are increased by a given amount, for simulations in a box of 75 Å. See Table 6.A for the value of the modified charges.

Hinge bond	q	q ± 0.07(10%)	q ± 0.15(20%)
42-HH	0.43	0.50	0.58
99-OD1	-0.76	-0.83	-0.91
94-OD1	-0.76	-0.83	-0.91
37-HE1	0.37	0.44	0.52

TABLE 6.A: Value of the CHARMM FF 36 charges, and the modified values, used for the hinge atoms.

Effect on distances:

Figs. 6.6,6.7,6.8 show the evolution of the hinge h-bonds over 250 ns of simulation, for the cases where: (i) charges are untouched (Fig. 6.6 corresponding to simulation from Fig. 6.4), (ii) for an increase of | **q** | by $0.07e$ (Fig. 6.7), and (iii) for an increase of | **q** | by $0.15e$ (Fig. 6.8). Fig. 6.9

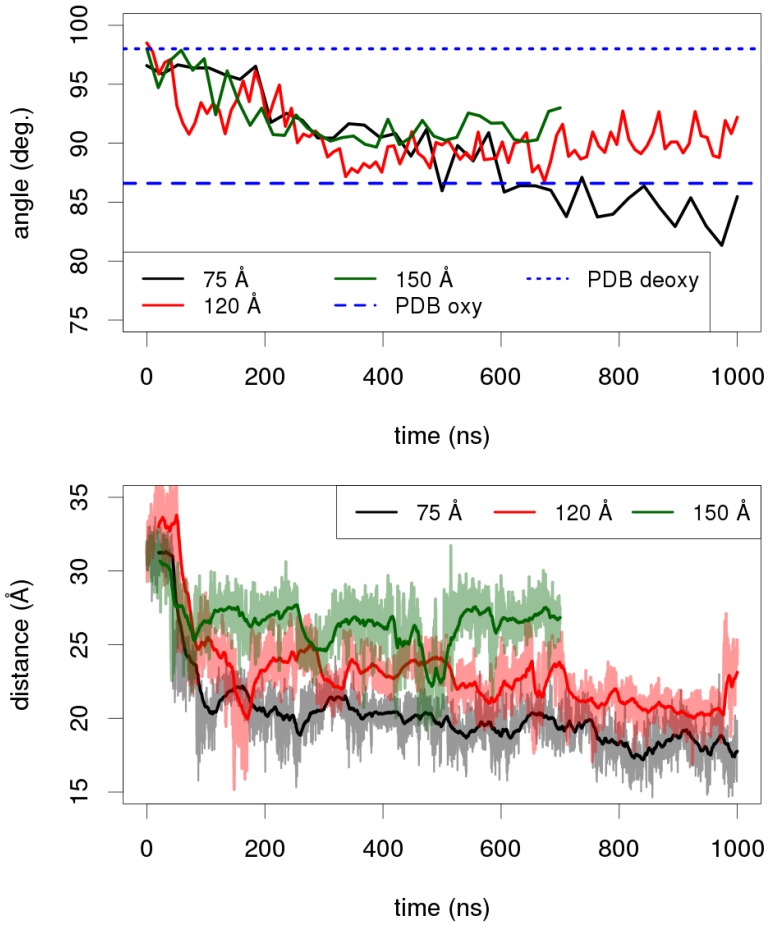


FIGURE 6.4: **Top:** Evolution of the angle between 2 vectors connecting the centre of masses of chains $\alpha_1\beta_1$ and $\alpha_2\beta_2$. In dashed blue lines is shown the angle as measured in the crystal pdb for both the oxy and deoxy structures, where the angle shift is measured to be $\approx 12°$. **Bottom:** Evolution of the distance between two C-α atoms of the 2 Histidine residues present at the terminal of the 2 β chains (see Fig. 6.1) for the deoxy state. The smoother line represents an exponentially weighted moving average on 200 steps. Results are shown for simulations performed in 3 water boxes of increasing size: 75 Å, 120 Å and 150 Å (see Fig. 6.2).

shows the terminal histidines C_α distances as defined in Fig. 6.1, for standard charges (corresponds to simulation from Fig. 6.4) , and scaled charges.

From Fig. 6.9 it is clear that increasing the charges of the atoms by $0.07e$ also increases the stability of the **T** structure, as the moving average of the histidine-histidine C_α distance is stabilised around 25–26 Å after 50 ns, where for the standard charges, as discussed above, this distance is close to 20 Å after 50 to 100 ns, i.e. the value of the **R** conformation. When comparing to Fig. 6.4 this means that this simple modification of the charges brings the same level of stability as box sizes between 120 and 150 Å. When increasing further the absolute values of the charges by $0.15e$, one sees on the contrary that this stabilising effect is lost: indeed this modification destabilises the tetramer which switches even faster to a compact **R** structure.

When having a look at Figs. 6.6,6.7,6.8 one can see that this increase of $|\,\mathbf{q}\,|$ by 0.07 mainly stabilises the hinge h-bond between atom HH from residue 42 and atom OD1 from residue 99, but that there is no clear effect on the other h-bond. The increase by 0.15 apparently stabilises both hinge h-bonds.

Water distribution around the h-bonds Radial Distribution Function (RDF) were estimated around the hinge atoms, for standard and modified charges, to see how they can be related to the decreasing histidine-histidine distance, seen in Fig. 6.6.

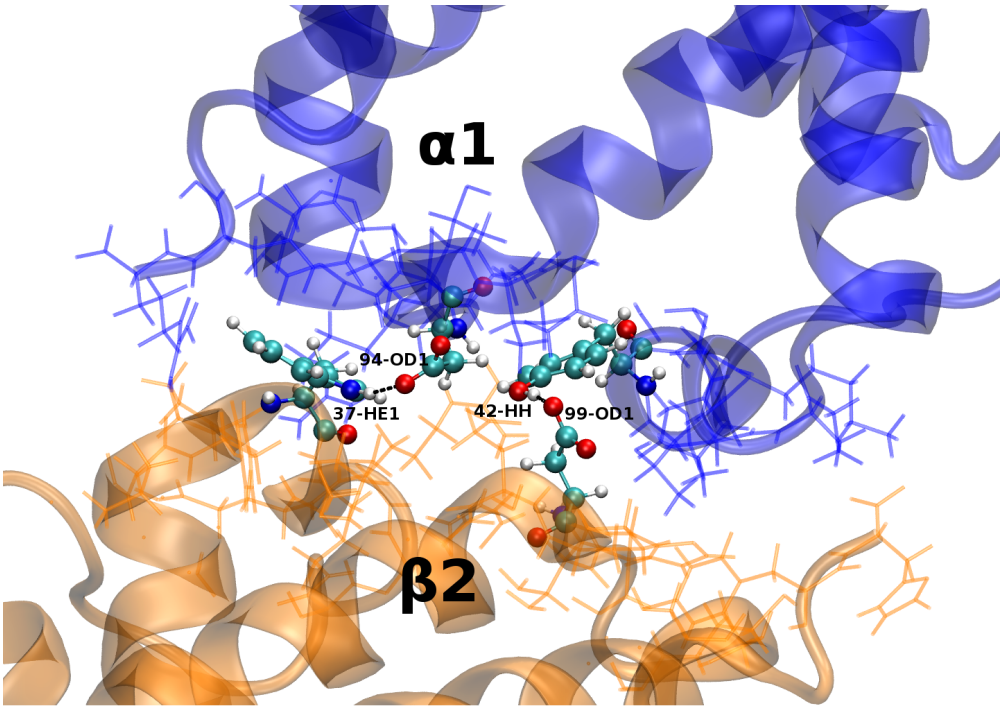


FIGURE 6.5: Definition of the *hinge* h-bonds. They involve chains α_1 and β_2, and are located at their interface. The first one involves atom HH from residue 42 and atom OD1 from residue 99. The second involves atom HE1 from residue 37 and atom OD1 from residue 94.

Fig. 6.10 shows RDF plots (g(r)) between the centre of mass of water molecules and: atom 42-HH (Fig. 6.10a), atom 99-OD1 (Fig. 6.10b), atom 94-OD1 (Fig. 6.10c) and atom 37-HE1 (Fig. 6.10d).

From Fig. 6.10 it seems that the organisation of water around the hinge atoms only shows a noticeable difference for atom 42-HH.

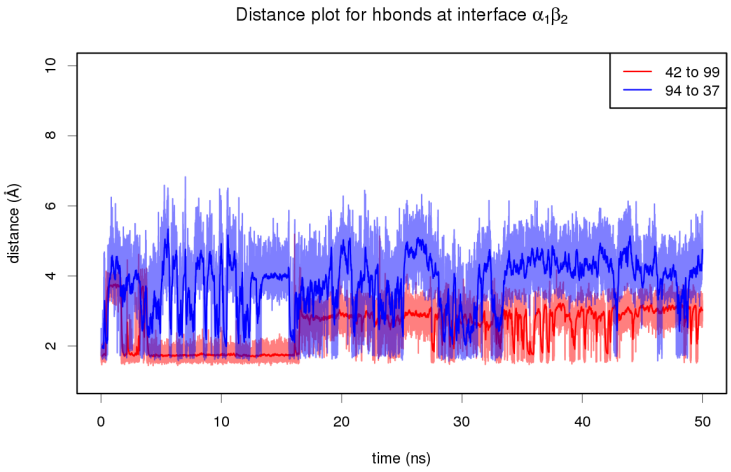


FIGURE 6.6: Hinge distances as defined in Fig. 6.5, for 250 ns long simulations for box of 75 Å. Standard charges from the CHARMM36 FF.

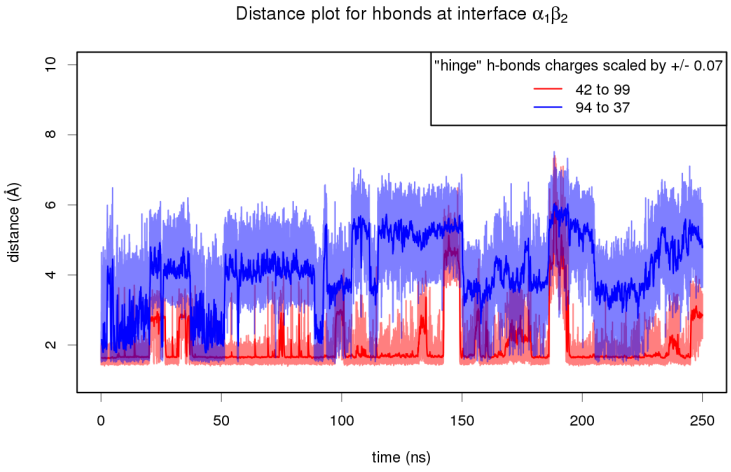


FIGURE 6.7: Hinge distances as defined in Fig. 6.5, for 250 ns long simulations for box of 75 Å. Hinge atoms' charges were increased by 0.07.

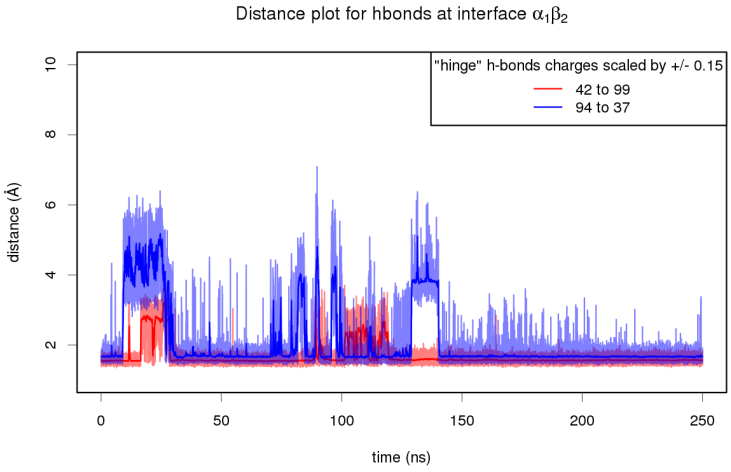


FIGURE 6.8: Hinge distances as defined in Fig. 6.5, for 250 ns long simulations for box of 75 Å. Hinge atoms' charges were increased by 0.15.

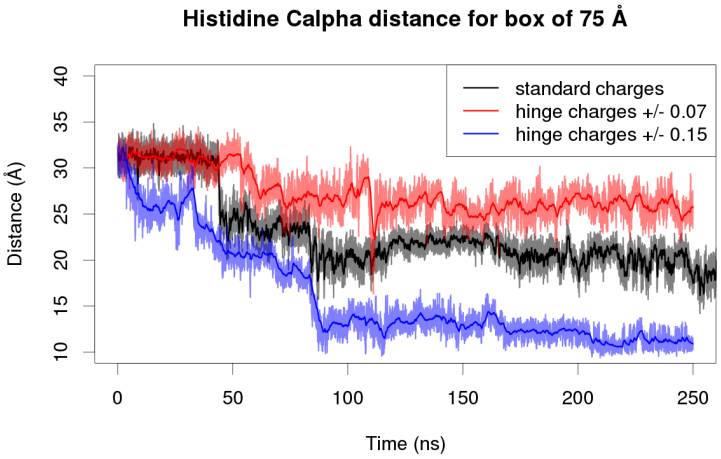
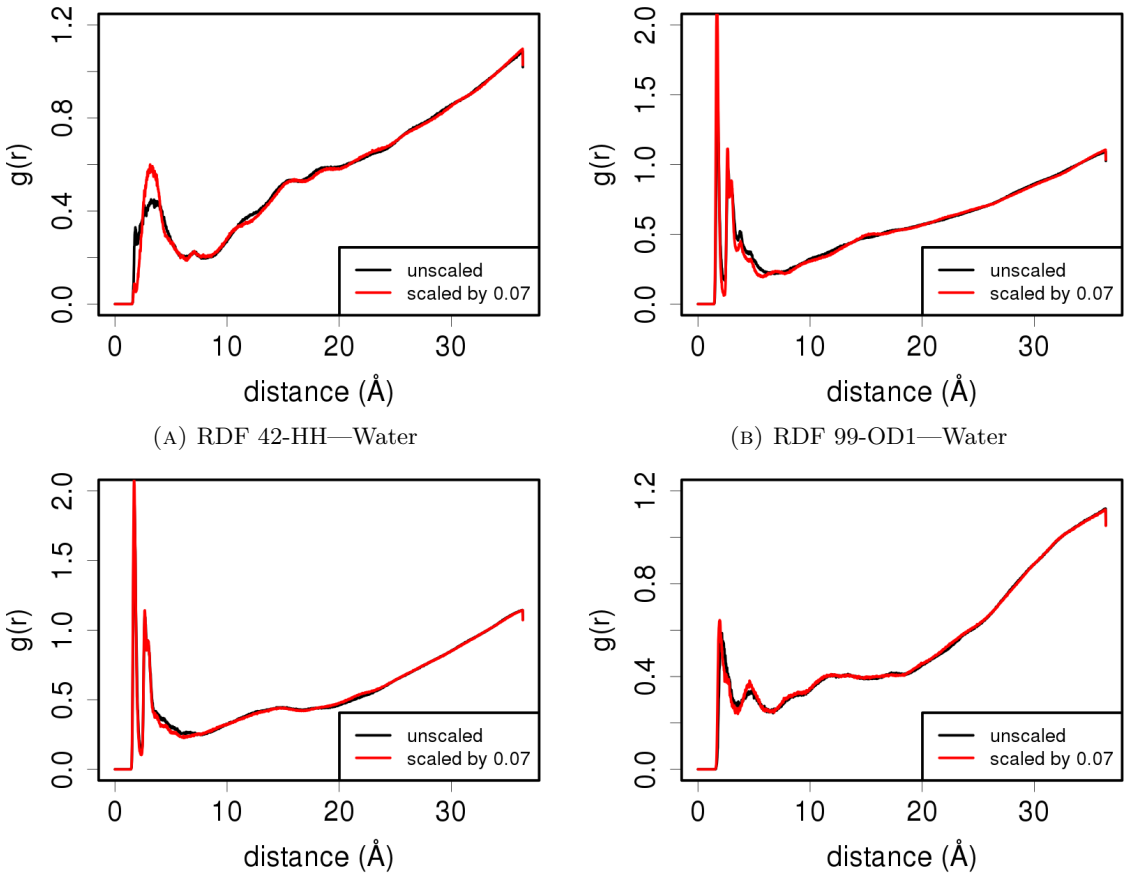


FIGURE 6.9: Terminal Histidines C_α distances as defined in Fig. 6.1. In black charges untouched, data taken from Fig. 6.4. In red the *hinge* charges were modified by adding 0.07 in absolute value. In blue, by 0.15. Increasing charges to 0.07 keeps the tetramer more stable (**T** structures) over 250 ns of simulation in a 75 Å box, but charges increased by 0.15 have the opposite effect, destabilising even more the **T** structure. Darker lines show an exponentially weighted moving average.



(A) RDF 42-HH—Water

(B) RDF 99-OD1—Water

(C) RDF 94-OD1—Water

(D) RDF 37-HE1—Water

FIGURE 6.10: Evolution of the RDF-water for the 4 hinge atoms, for unscaled (black) or scaled (by 0.07) charges (red) in a 75 Å box, over 250 ns of simulation. See Fig. 6.5 for definition of hinge contacts.

6.3 Coarse Grained density analysis

Willard and Chandler introduced in Ref.[150, 151] the following coarse grained density $\rho(x;\xi = cst)$:

$$\rho(x;\xi = cst) = (2 * \pi * \xi^2)^{-1.5} \exp\left(\frac{-x^2}{2 * \xi^2}\right)$$

The value $\xi = 2.4$ Å is used by default.

In the following we investigate the evolution of $\rho(x;\xi = cst)$ for values of ξ between 2.4 and 4.0 Å. This is evaluated for distances x between 0 and 10 Å. See Figure 6.11.

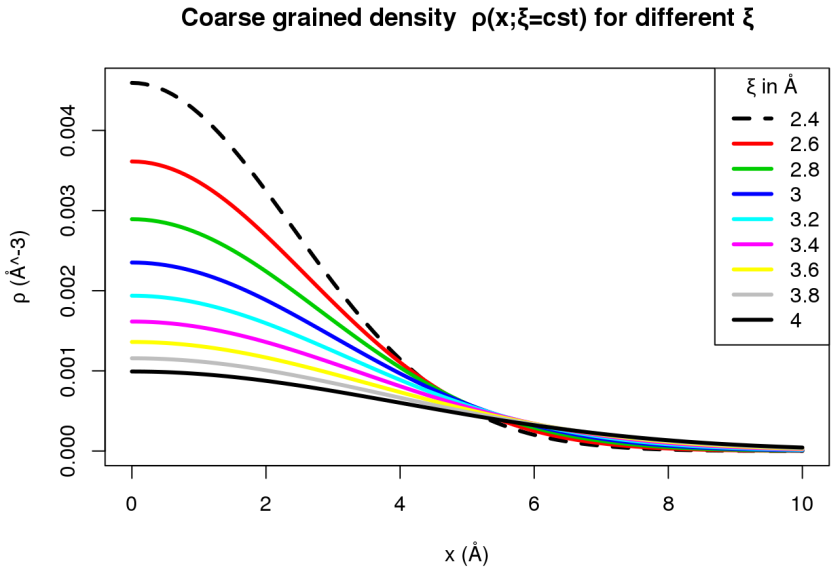


FIGURE 6.11: Evolution of $\rho(x;\xi)$ for several values of the coarse graining parameter $\xi \in [2.4; 4.0]$, and distance $x \in [0; 10]$

6.3.1 Implementation and validation

Influence of the value of ξ on the isosurface On the following Figure 6.12, one can see the influence of changing values of ξ. Isosurfaces were produced for a coarse grained density of 0.016 $Å^{-3}$ (corresponding to approximately half the bulk water density). Density was estimated on a 3D grid of resolution 0.1 Åfor each axis. The system of interest is the Haemoglobin tetramer in water, box of edge length 75 Å.

The black wireframe was rendered for a value of $\xi = 2.4$ Å; the red surface was rendered for a value of $\xi = 3.0$ Å; and the yellow surface was rendered for a value of $\xi = 4.0$ Å

Increasing ξ smooths the surface, justifying the coarse graining term: during dynamics this should allow to visualise more efficiently relevant structural changes.

Effect of a cutoff distance when counting water molecules From Figure 6.11 it is clear that when using the coarse graining approach the value of $\rho(x;\xi)$ tends to 0.0 at a given distance. Thus when estimating the coarse grained density at a given grid point, one can restrict to water molecules within a given cutoff distance.

Figure 6.13 shows the isosurface, for a coarse grained density $\rho(x;\xi = 2.4) = 0.016$ $Å^{-3}$, in two cases: where no cutoff is applied (black wireframe) and with a cutoff of 8 Å (solid yellow surface).

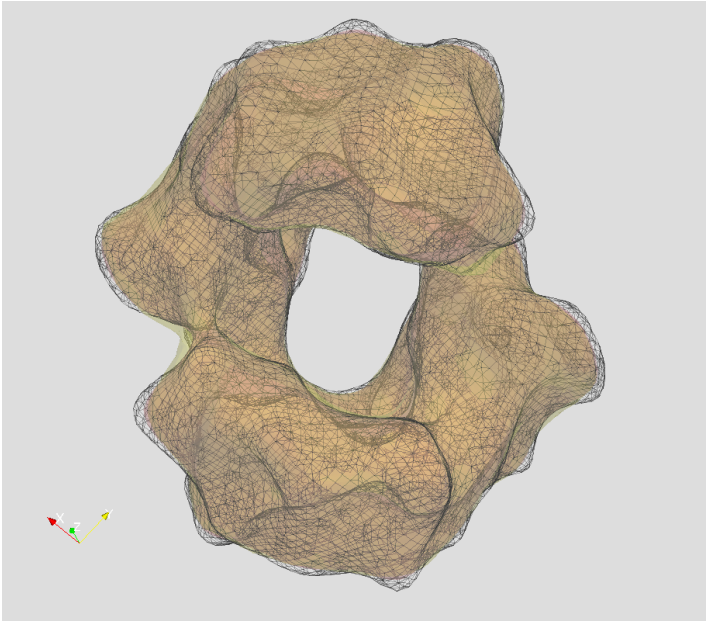


FIGURE 6.12: Effect of higher ξ values : (1) black wireframe = 2.4 Å; (2) red surface = 3.0 Å; (3) yellow surface = 4.0 Å

No difference can be visually observed, but the computation time is logically several orders of magnitude higher without a cutoff.

Visual comparison of the isosurfaces Figure 6.14 compares the coarse grained isosurface (ξ = 3.0 Å) for the box of 75 Å, at : (a) just after equilibration (t = 0), (b) t = 20ns, (c) t = 50ns, (d) t = 100 ns, (e) t = 400 ns and (f) t = 1000 ns. The C-α His146 distance from Figure 6.4 (bottom) is also shown in red.

6.3.2 Using isosurface's normal vectors for extracting density

Methodology From the considered coarse grained density grid (75, 120 or 150 Å) an isosurface is extracted, for half the bulk density. Then 20 normal vectors to this surface are generated: the surface being rendered as many triangular meshes, for each vertex one can easily define a normal vector using the cross product of the 2 edges. Vectors recrossing the surface at some point (like the ones that we could get inside the cavity) are excluded from the analysis for the moment. A straight line follows each vector, and coarse grained density is extracted from the grid following this line. Tri-linear intrapolation is used for extracting values of the density .

In the following 5 vectors from the 20 random generated ones were chosen, then we will track density fluctuations for the 3 boxes size and for various trajectories, using **exactly the same vectors**. The chosen vectors can be visualised from the following Figure 6.15.

The coarse graining analysis was performed on 50 ns of trajectory, every 50 ps, and the corresponding 3D grid stored in a binary file. Then analysis scripts are used in order to define the surface and extract normal vectors, as shown before (Figure 6.15) 5 of them are kept, which explore different parts of the box around the centred tetramer.

Once vectors have been extracted and the corresponding intrapolated density is known, an averaging stage is performed. Indeed instant fluctuations are not so meaningful because of the expected anisotropy observed on the density grid, especially because of the presence of ions.

Box of 75 Å The Figure 6.16 shows the Histidine C_α distance fluctuation over 50 ns in order to identify properly the transition windows. Vertical lines correspond to the time windows detailed below.

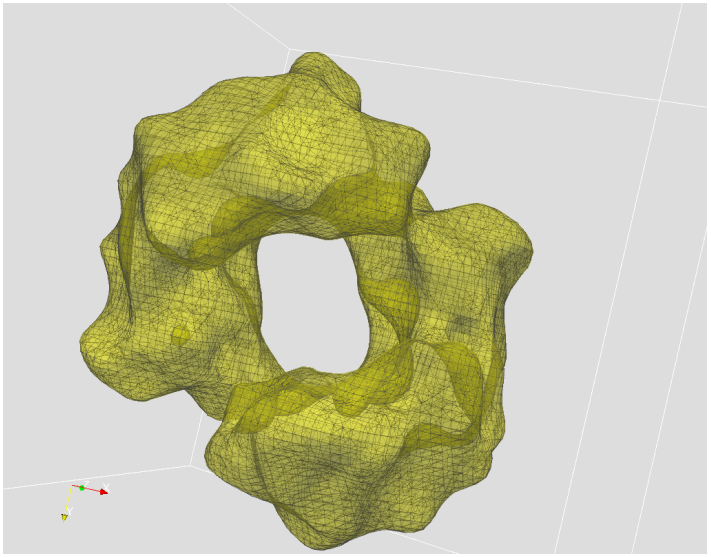


FIGURE 6.13: Effect of cutoff distance on surface estimation, for the same $\xi = 2.4$ Å

In the Figure 6.17 results for the box of 75 Å are presented. For each vector, the profile is extracted for several time windows.

The following time windows were considered:

- 0.0 to 0.5 ns and 10.0 to 10.5 ns, before Histidine C_α distance collapsing
- 35.0 to 35.5 ns, during Histidine C_α distance collapsing
- 49.5 to 50.0 ns, after Histidine C_α distance collapsing

The Figure 6.18 presents the same data but this time each sub-figure tracks one time-window, and the 5 vectors are plotted for each time-window.

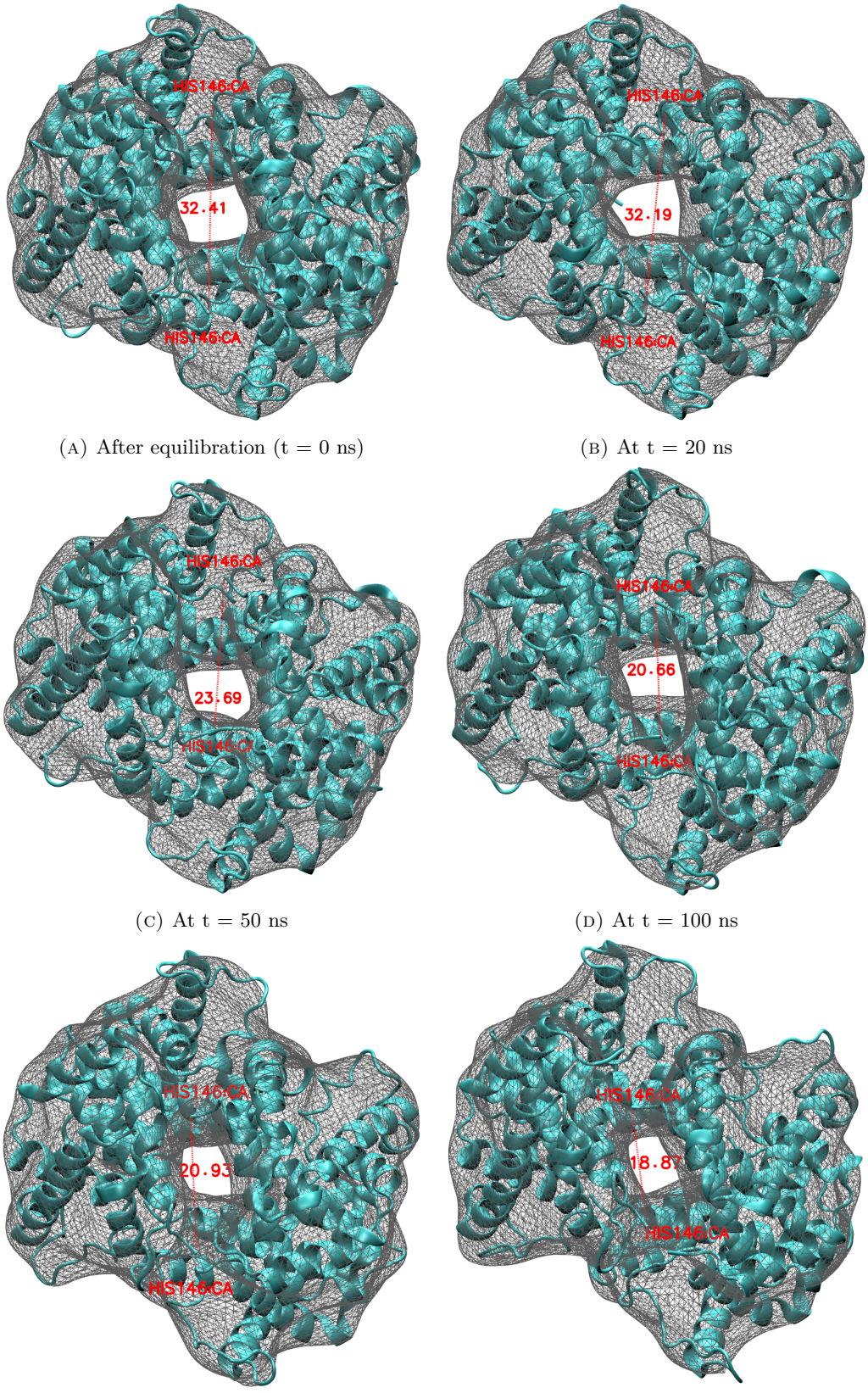
From Figures 6.16 – 6.17 – 6.18 it appears that the density along some of the vectors during the simulation evolves : for example from Figure 6.18d one can see that after ≈ 50 ns of simulation the 5 vectors cluster in two different groups, one for which the density reaches the bulk average after 5 Åand one for which this requires 10 Å. Vectors 13 and 19 (Figures 6.17c and 6.17c) seem to be particularly sensitive to this effect, and should be could candidates.

3 boxes comparison In the following Figure 6.19 similar time window averaged plots are produced , but this time for comparing the 3 box sizes (for one time wondow only here, i.e. 9.5 to 10 ns, but this can easily be plotted for more windows).

It confirms the previous observation for Vectors 13 6.19c to be a good candidate, as it shows fluctuations both for different time windows (as concluded from previous paragraph), but also for different box sizes.

From this Section 6.3 it seems that the analysis of coarse grained density fluctuations through the use of probe vectors distributed on the normal surface may allow one to probe different parts of the cubic box, in order to try to investigate the intensity of the fluctuations.

Therefore the development of a systematic “analysis flow” based on the coarse grained density, combined with the conformational measures presented in Section 6.2 may be an interesting way of investigating the instability of this tetramer in water.



(A) After equilibration (t = 0 ns)

(B) At t = 20 ns

(C) At t = 50 ns

(D) At t = 100 ns

(E) At t = 400 ns

(F) At t = 1000 ns

FIGURE 6.14: Comparison between isosurfaces obtained for the box of 75 Å box, using $\xi = 3.0$ Å, at different increasing simulation times.

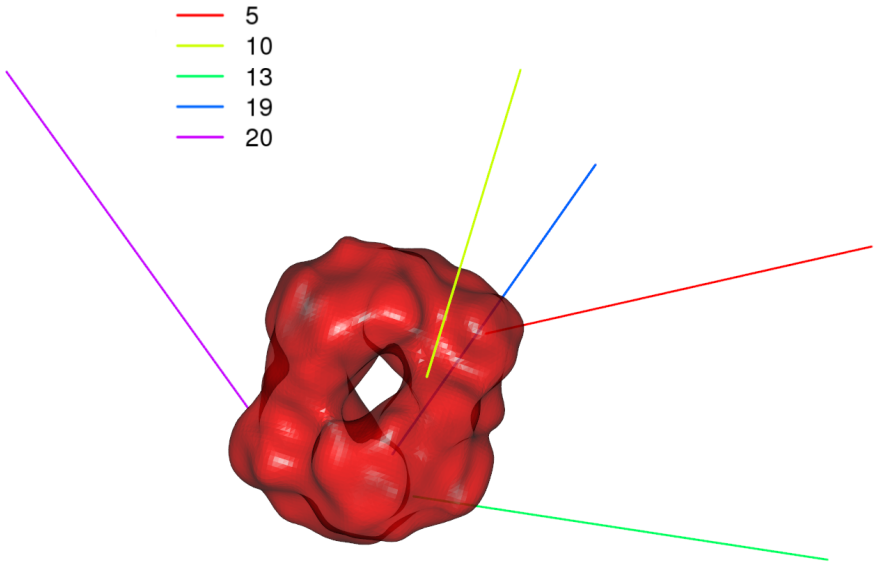


FIGURE 6.15: Definition of the 5 tracked vectors, for a coarse grained density isosurface of $\xi = 3.0$.

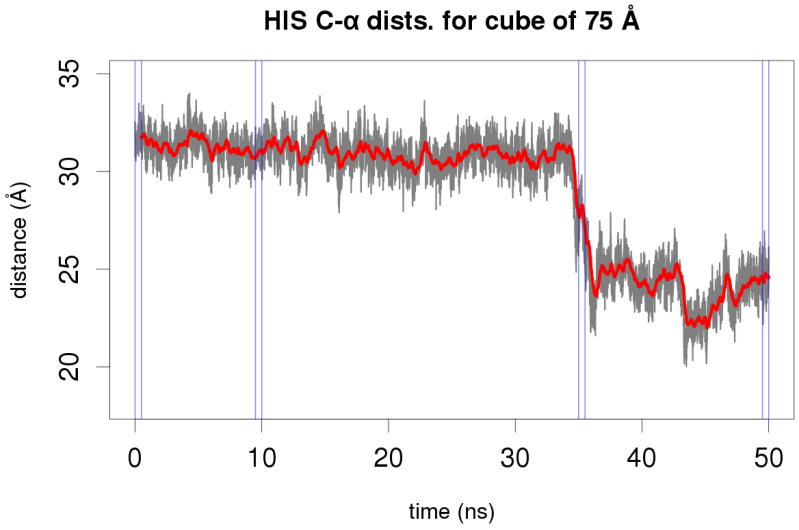
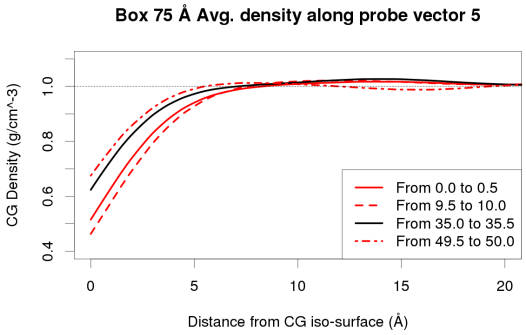
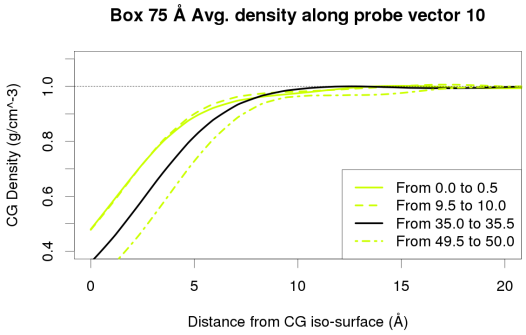


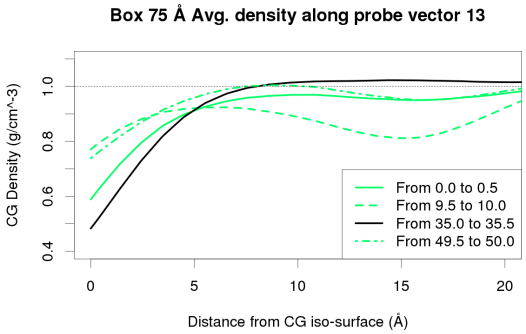
FIGURE 6.16: Histidine C_α distance (in black fluctuations, in red an exponentially smoothed rolling average), the 0.5 ns time windows used in Figure 6.17 are denoted by blue vertical lines



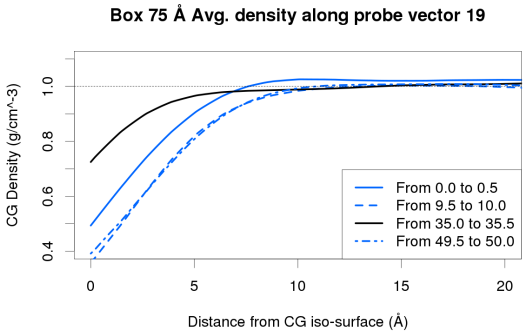
(A) vector 5



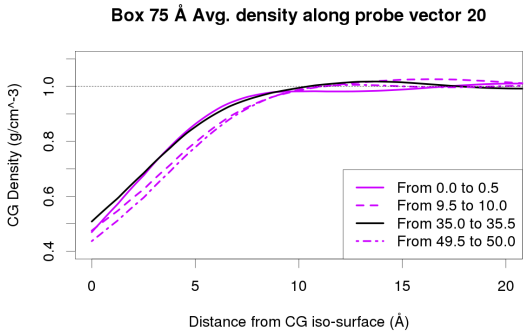
(B) vector 10



(C) vector 13

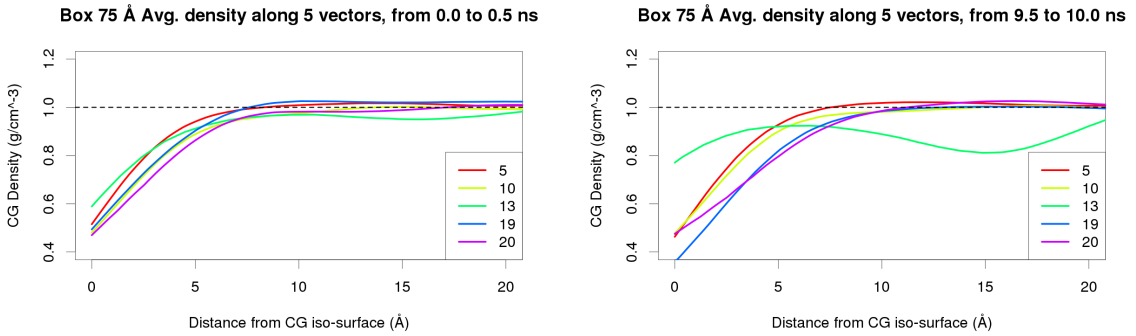


(D) vector 19



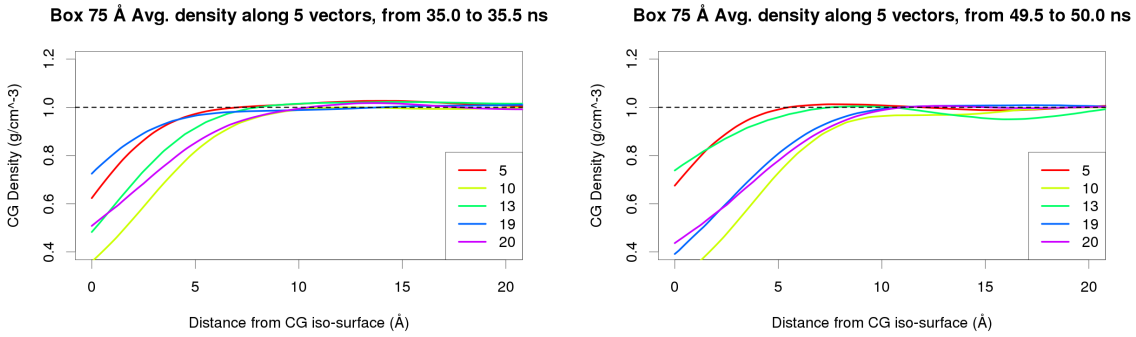
(E) vector 20

FIGURE 6.17: Averaged vectors (over 500 ps) for several time windows before, during, and after the histidine C_α distance collapsing. Box of 75 Å. **T** state for windows 0.0 to 0.5 ns and 10.0 to 10.5 ns ; **R** state for windows 49.5 to 50.0 ns ; somewhere between **T** and **R** for window 35.0 to 35.5 ns, displayed as a dashed black line in all graphs. See Figure 6.16 where the 500 ps windows are identified by vertical bars.



(A) 0 to 0.5 ns

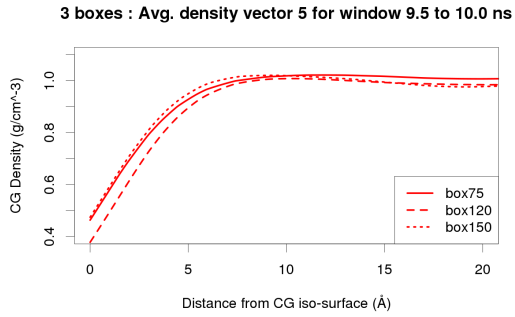
(B) 9.5 to 10 ns



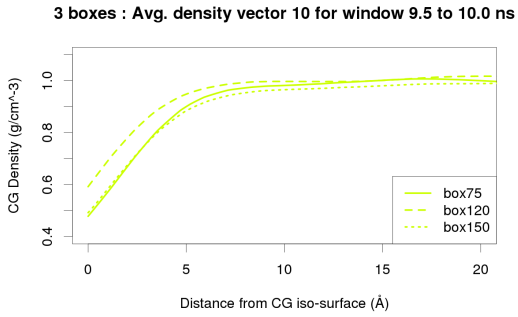
(C) 35 to 35.5 ns

(D) 49.5 to 50 ns

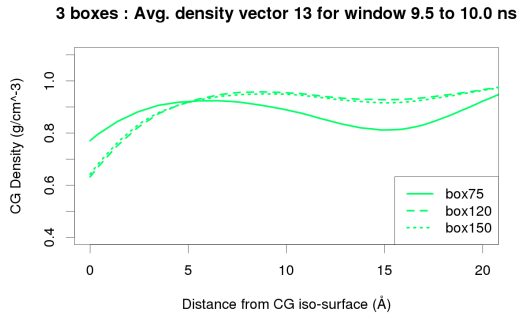
FIGURE 6.18: Averaged vectors (over 500 ps) for several time windows before, during, and after the histidine C_α distance collapsing. Box of 75 Å. See Figure 6.16 where the 500 ps windows are identified by vertical bars. Each panel represents a time-window, and for each panel the 5 vectors are represented.



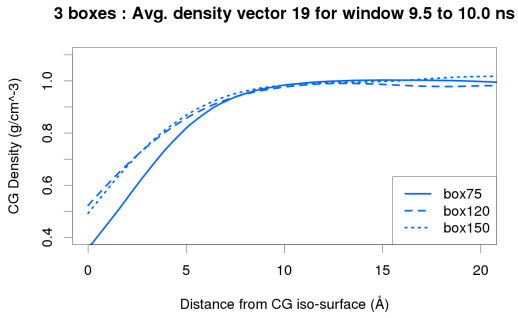
(A) vector 5



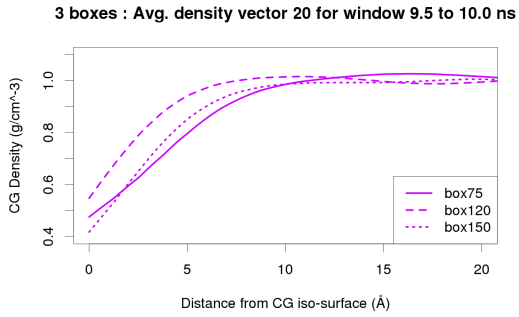
(B) vector 10



(C) vector 13



(D) vector 19



(E) vector 20

FIGURE 6.19: Averaged vectors (over 500 ps) for time window 9.5 to 10 ns for above defined 5 vectors and different box sizes. For all boxes tetrameter are still in the **T** state.

Conclusion

Continuing briefly the personal *"petite digression"* that I started in the Introduction, I will just repeat once again how many things I have learned during this PhD, either from the scientific or methodology point of view, or from the programming/computing side. I rarely spent more than a few days without learning new concepts and experimenting new analyses, and for this I would like to thank once again Prof. Markus Mewuly for the freedom he gave me in the choice of the simulation tools and analyses to perform. I clearly never or almost sampled the "boringness rare event", and that is a good thing.

What will remain from this PhD are two algorithms implemented in CHARMM, plus the Fitting Wizard toolkit, and I hope that they will reveal useful for more people in the future. As already mentioned in the conclusion of each respective article, those methods/tools are far from being perfect and there is still clearly a huge potential of work: for the two rare events algorithms let us mention : (i) smarter MC moves and automatic parameters tuning for SA-MC, (ii) generalisation of PINS for allowing the use of biased Hamiltonians instead (or on top) of multiple temperatures …

For the Fitting Wizard tool there is clearly a huge amount of fixes and improvements one can think about, especially because since August 2016 we now have an evolution of the original FW tool that is now running on a web sever and accessed through a web browser, instead of the previous JAVA client implementation. One can think about extending the FW with the possibility to use other MD/QM codes, for example CP2K or OpenMM.

Finally the article on Hæmoglobin tetramer still has to be published, and I think that the set of all the scripts written during those months for performing analysis and generating all those nice figures should be made available to the community.

My last words will be extra acknowledgments: first to Prof. Markus Meuwly and Prof. Anatole von Lilienfeld for being respectively Examiner and co-Examiner of this thesis, they deserve congratulations for reading this thesis up to this point !

And then once again I would like to thank my family, friends and colleagues for al the possible types of supports they provided during those years.

Florent Henri René Hédin, Sept. 2016, Basel, Switzerland

APPENDICES

Appendix A

NMA work

This appendix contains an article where the vibrational relaxation of the N-Methylacetamide is investigated. The idea was to tried to understand how the energy gained by a $C = O$ bond, after spectroscopic excitation, diffuses in water, using MD methods. My contribution consisted in a statistical analysis of the motion of the three water molecules which are the closest to the $C = O$ bond just after excitation: for that we used the useful Quantile-Quantile (Q-Q) (Figure 5) plots, and observed a non-Gaussian distribution, of type "Log-normal" (Figure 6). This distribution of the points was fitted to a robust model.

A.1 Vibrational Relaxation of N-Methylacetamide

THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

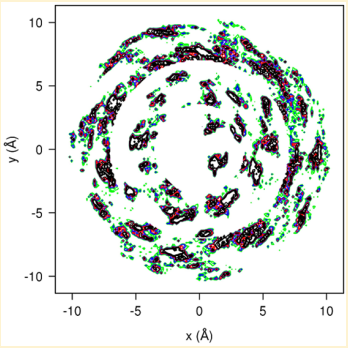
Vibrational Relaxation and Energy Migration of *N*-Methylacetamide in Water: The Role of Nonbonded Interactions

Pierre-André Cazade, Florent Hédin, Zhen-Hao Xu, and Markus Meuwly*

Department of Chemistry, University of Basel, Klingelbergstrasse 80, 4056 Basel, Switzerland

Supporting Information

ABSTRACT: Nonequilibrium molecular dynamics (MD) simulations together with physics-based force fields are used to follow energy flow between vibrationally excited *N*-methylacetamide (NMA) and water. The simulations are carried out with a previously validated force field for NMA, based on a multipolar representation of the electrostatics, and with a new fluctuating point charge model. For the water solvent, a flexible and a rigid model was employed to distinguish between the role of inter- and intramolecular degrees of freedom. On a 10 ps time scale about 90% of the available energy goes into the solvent. The remaining energy resides within internal NMA-degrees of freedom from where energy flow takes place on longer time scales. The total amount of energy transferred to the solvent on the 10 ps time scale does not depend on whether the water molecules are rigid or flexible during the simulations. Vibrational energy relaxation time scales include two regimes: one on the several 100 fs time scale and a longer one, ranging from 6 to 10 ps. This longer time scale agrees with previous simulations but overestimates the experimentally determined relaxation time by a factor of 2, which can be explained by the classical treatment of the vibrations. Including a previously determined quantum correction factor brings the long time scale into quite favorable agreement with experiment. Coupling to the bending vibration of the water molecules in H-bonding contact with the excited C═O chromophore is substantial. The equilibrium and nonequilibrium distribution of the bending angles of the water molecules in contact with the local oscillator are non-Gaussian, and one approaches the other on the subpicosecond time scale. Analysis of the water velocity distribution suggests that the C═O vibrational energy relaxes into the solvent water shells in an impulsive fashion on a picosecond time scale.



1. INTRODUCTION

The exchange of vibrational energy between molecules is important in understanding condensed phase phenomena because reaction rates and pathways are affected by energy exchange between solvent and solute.[1,2] Experimental and computational studies have been carried out for a variety of systems and situations, ranging from di- and triatomic molecules in different solvents[3−5] to CO in metal carbonyl compounds[6] and heme-bound CO in proteins.[7−9] Current progress in ultrafast time-resolved infrared and Raman spectroscopy applied to study intramolecular vibrational redistribution can provide details of the transient energy content of individual vibrations in solvated polyatomic molecules.[10−12] Such studies are complemented and their interpretation is aided by theoretical and computational work which allows one to follow the pathways of vibrational energy relaxation and redistribution.[1,13,14]

A rich dynamical picture has emerged from such investigations. The vibrational relaxation times were found to be typically on the few picosecond time scale and extending to the nanosecond or longer time regimes in exceptional cases, depending on the solvent, ligation state (e.g., bound versus unbound diatomic), and chemical environment of the spectroscopic probe.[2,5,15] Specifically, molecular dynamics (MD) simulations have provided considerable insight into relaxation pathways and the role of intermolecular, in particular, electrostatic interactions. For example, the importance of electrostatic interactions has been investigated in quite some detail for a model dipolar molecule in a polar solvent.[16]

Compounds with peptide bonds have attracted particular interest because of their fundamental role in biological systems. The amide I mode, primarily associated with the peptide carbonyl stretch,[17] has been frequently used in experiments because its strong transition dipole makes it possible to identify its contributions.[10,18−22] This mode is of particular interest as it can be used to probe the topology and hydrogen-bond network through the intensity, spectral shift, and shape of this band.

The molecular system representing a peptide carbonyl most closely is *N*-methylacetamide (NMA; Figure 1). The vibrational relaxation of deuterated NMA (NMAD) in solution has been previously investigated in several experimental and theoretical studies,[10,11,23−25] which have provided a qualitative picture of the main relaxation pathways. However, differences in the interpretation of the vibrational energy decay and uncertainties in the experimental values of the relaxation lifetimes of the amide I mode show that a concise description of the vibrational

Received: November 22, 2014
Revised: January 11, 2015
Published: January 12, 2015

DOI: 10.1021/jp511701z
J. Phys. Chem. B 2015, 119, 3112−3122

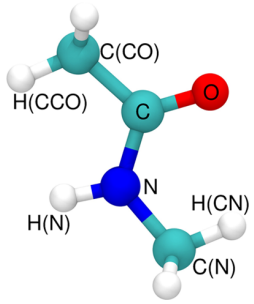


Figure 1. *N*-Methylacetamide molecule with atoms labeled.

relaxation of NMAD is still missing.[10,26] Recently, the heat transfer from NMA to various solvents has been investigated from temperature jump simulations.[27] This work employed a united-atom representation of NMA and found cooling times between 6 ps (water) and 28 ps (CCl_4). In addition, the sensitivity to nonbonded interactions (electrostatics and van der Waals) was investigated by scaling the strength of these contributions. However, the solvent structure and pathways for energy migration were not explicitly considered. Also, any anisotropy in the electrostatic interactions was absent because standard fixed-point charge models were employed.

Earlier work on the vibrational relaxation of NMA in D_2O using a united atom force field and nonequilibrium MD simulations found qualitative agreement with experiment for the rapid time scale of vibrational relaxation (1.5 ps versus 0.5 ps).[23] Using Fermi's golden rule, the vibrational relaxation time is 160 ps, which is 2 orders of magnitude longer than the experimental and nonequilibrium results.[23] A more recent study based on a multistep reaction and using non-Markovian time-dependent perturbation theory with the Neumann−Liouville equation and third-order Fermi resonance parameters to determine the mode-to-mode energy flow rate constants yields good agreement with the experimental results.[10,28] However, such an approach is limited to the subpicosecond time scale dynamics and does not provide a complete description of the relaxation process. Quantum effects on the amide I relaxation in NMA have also been investigated, and it was found that nonequilibrium MD simulations typically overestimate the picosecond relaxation time scale by a factor of 2 to 2.5, depending on how the classical MD simulations were carried out.[29] Yet another computational study used nonequilibrium dynamics of NMAD based on MD with quantum transitions in which the amide I mode is treated quantum mechanically while the remaining degrees of freedom are treated classically.[30] The instantaneous normal modes[31] of the initially excited NMAD molecule are used as internal coordinates. The time evolution of the energy stored in each individual normal mode is subsequently quantified using the hybrid quantum-classical instantaneous normal modes. Such an approach finds that amide I relaxation is dominated by intramolecular vibrational redistribution with little contribution from the solvent.

In the present study we aim at following the redistribution of the excitation energy for vibrational relaxation of the NMA−amide I mode in D_2O. Of particular interest are the modes and time scales for energy transfer into the surrounding solvent. On the basis of previous success for following vibrational relaxation of CN^- in D_2O from nonequilibrium MD simulations,[32] the present study uses high-quality force fields to investigate energy migration between vibrationally excited NMAD and the surrounding D_2O solvent. The influence of the electrostatic model on the results is scrutinized, and the energy migration process is followed at atomic resolution. As NMA is a typical building block of polypeptides and proteins, the general insights regarding the relationship between force field accuracy, simulation strategy, and physicochemical observables is of great interest also in a wider context.

2. COMPUTATIONAL METHODS

2.1. Intermolecular Interactions. Force field parameters for NMA are based on CGenFF,[33] except for the C═O bonded term and electrostatics which are detailed below. For water, two models were considered: (i) the TIP3P model,[34] which is usually used for simulations in which the internal degrees of freedom are constrained, and (ii) the flexible Kumagai, Kawamura, and Yokokawa (KKY) model[32,35] to characterize the role of internal water degrees of freedom in the vibrational relaxation. The KKY potential correctly describes the harmonic frequencies of the water monomer and the infrared spectrum of liquid water[36] and has also been used successfully in characterizing the vibrational relaxation of solvated cyanide.[32]

Previous studies of vibrational relaxation in the condensed phase have shown that the nonbonded interactions are of fundamental importance for realistic simulations.[16,27,32,37] Here we decided to treat the electrostatics of the optically active C═O stretching mode with atom-centered fluctuating point charges (FPC). Such a model has already proved useful for investigations of the Stark effect of photodissociated ligands in Mb.[38−41] Alternatively, to more accurately describe electrostatic interactions, a multipolar (MTP) expansion can be used.[42,43] As the decomposition of the electron density into point charges and higher-order multipoles is not unique, the actual magnitudes of the multipoles can vary significantly. In previous work this was explored by designing three models with weak, medium, and strong multipoles (MTPW, MTPM, and MTPS, respectively) each of which is equally suitable to describe the electrostatics.[44] In the present work MTPW is used as it allows the most direct comparison with FPC because the vdW parameters are identical and the dipole and quadrupole moments are a perturbation compared to the CGenFF model.

All quantum chemical calculations in the present work were carried out with the Gaussian software.[45] For the fluctuating point charge model, the structure of NMA was first optimized at the B3LYP/aug-cc-pVQZ level of theory. Subsequently, the structure of NMA was distorted along the C═O bond length (r) and the partial atomic charges were determined by means of fitting to the electrostatic potential (ESP).[46,47] The r-grid included 21 distances between r = 1.1 Å and r = 1.35 Å. This choice was motivated by the amount of fluctuations of this coordinate in the MD simulations (∼10% around the equilibrium). For each conformation, the charges obtained from the ESP analysis are represented as a linear expansion of the C═O bond length

$$q_X(r) = a_{X,0} + a_{X,1}r \quad (1)$$

where X stands for any atom of NMA. Using the B3LYP/aug-cc-pVQZ level of theory to determine the necessary electrostatic parameters was found to be reliable in previous work for NO, CN^-, O_2, or CO_2.[32,48−51] The results of the fitting procedure are shown in Table I in Supporting Information.

DOI: 10.1021/jp511701z
J. Phys. Chem. B 2015, 119, 3112−3122

All bonded interaction terms are those of the CHARMM22 force field,[52] except for the C═O stretching potential, which was optimized in two steps. First, the parameters of the Morse potential $V(r) = D_e[1 - \exp(-\beta(r - r_e))]^2$ are fitted to electronic structure data (B3LYP/aug-cc-pVQZ). The parameters are then refined by reproducing the experimental gas phase C═O stretch frequency of 1731 cm^{-1} by adjusting D_e.[53] The final parameters are then D_e = 120.47 kcal/mol, β = 2.174 $Å^{-1}$, and r_e = 1.294 Å.

2.2. Molecular Dynamics Simulations. All MD simulations are carried out with CHARMM[54] using a time step of Δt = 0.5 fs when a flexible water model is employed and Δt = 1 fs when the water is constrained using SHAKE.[55] The equations of motion were propagated with the leapfrog algorithm. Periodic boundary conditions are applied in the three spatial directions, and a cutoff of 12 Å is used for the nonbonded interactions. A shifting and switching function is applied to electrostatic and van der Waals (vdW) interactions, respectively. Equilibration for the reference (equilibrium) trajectories is performed in the *NVT* ensemble at 300 K within the weak-coupling[56] limit with a damping constant of τ = 1.0 ps for the thermostat. The reference and the relaxation simulations are run in the *NVE* ensemble. The reference trajectories are 2.5 ns long whereby data is stored every 10 ps for a total of 250 sets of **x** and **v**. Positions (**x**) and velocities (**v**) are required to prepare the nonequilibrium state of the system (see below) and to follow the relaxation trajectories. The latter are carried out in two phases: during 2 ps after the excitation, the data is stored every 1 fs, and for the remaining 23 ps, the data is stored every 10 fs. This provides sufficient resolution for the early events after excitation and also allows one to follow further relaxation while keeping the amount of stored data manageable.

2.3. Vibrational Excitation. From the 250 sets (**x** and **v**), the amide I IR mode of NMA is excited by depositing the corresponding energy (1725 cm^{-1} or 4.92 kcal/mol) as kinetic energy. This is achieved by suitably displacing the molecule along the C═O normal mode and scaling the velocity vector appropriately.[23] NMA is first reoriented in the (*x*,*y*) plane, providing the reorientation matrix **B**. In this frame, the inertia tensor of NMA is calculated. In what follows, vectors and matrices are written in boldface while scalar numbers are in standard italic font. The scaled velocities ($\tilde{\mathbf{v}}$) are determined following the procedure outlined in eqs 2−5:

$$\tilde{\tilde{\mathbf{v}}} = \gamma(\mathbf{I}^{\dagger}\mathbf{B}^{\dagger}\mathbf{v} + \lambda\mathbf{L}(k)) \tag{2}$$

where **v** and **L** are the velocities of the current MD snapshot and the Cartesian displacements along the normal mode of interest, respectively. **B** is the rotation matrix in the NCO plane, and **I** is the matrix of the eigenvectors of the inertia matrix. $\tilde{\tilde{\mathbf{v}}}$ are the scaled velocities in the reoriented frame. The scaling factor λ is determined from

$$\lambda = \sqrt{\frac{2hc\nu}{\sum_i m_i(\mathbf{L}(k,i))^2}} \tag{3}$$

with ν is the wavenumber, m_i the mass of the atom *i*, and *k* the normal mode of interest. The scaling factor λ ensures that the energy deposited corresponds to excitation along the normal mode of interest. Adding the signed components of the normal mode vector and the instantaneous velocity **v** in general leads to a change of the total energy which does not correspond to the desired excitation, which is 4.92 kcal/mol in the present case. Therefore, it is necessary to rescale the resulting velocity by a second factor γ

$$\gamma = \sqrt{\frac{2hc\nu + E^0_{kin}}{E^{\lambda}_{kin}}} \tag{4}$$

where E^0_{kin} is the kinetic energy due to the velocities **v** and E^{λ}_{kin} the kinetic energy due to the increased velocities by the scaled normal mode, $\mathbf{I}^{\dagger}\mathbf{B}^{\dagger}\mathbf{v} + \lambda\mathbf{L}(k)$

$$\tilde{\mathbf{v}} = \mathbf{B}\mathbf{I}\tilde{\tilde{\mathbf{v}}} \tag{5}$$

$\tilde{\mathbf{v}}$ are the modified velocities along the normal mode used for IR excitation.

2.4. Data Analysis. For the present work it is of particular interest to follow the temporal evolution of various energy components. Their change relative to the initial state is obtained from

$$\Delta E(t) = \frac{1}{N}\sum_i^N \sum_j^M (E^*_{ij}(t) - E^0_{ij}) \tag{6}$$

where N = 250 corresponds to the number of trajectories and M = 882 to the number of water molecules. The superscripts * and 0 refer to the excited and the reference trajectory, respectively.

3. RESULTS

In the following, results from atomistic simulations using several force field parametrizations for NMAD are discussed. One of them is a previously validated multipolar force field which correctly reproduces the pure liquid density and heat of vaporization along with the hydration free energy and the 2D infrared spectroscopy.[44] This is contrasted with simulations using a general-purpose (CGenFF) parametrization[33] and one which employs an FPC model (see Methods) to capture effects due to bond-polarizability.

3.1. Solvent Structure around the Chromophore. The local structure of water around NMA using the same force fields has been previously investigated.[44] The pair-correlation function between the oxygen atom of the carbonyl group of NMA and the water-O atoms exhibits a first peak at 2.8 Å, characteristic of a first water H-bonded shell. A second, weaker peak is found at about 5 Å. Such a one-dimensional description averages out several important features which become more prominent when considering two-dimensional water densities. They establish that three main regions can be distinguished: (i) a relatively high-density region close to the molecule (within the first 3.5 Å) corresponding to the first solvation shell; (ii) a region of moderate density at distances 5−8 Å away, corresponding to a second solvation shell; and finally (iii) beyond 8 Å more uniformly distributed water corresponding to the bulk. Within region (i), there are localized and well-defined areas of high density corresponding to H-bonding sites: 2 around the O atom of NMA and one around the NH group. The water-bonding sites around the C═O-group are of particular interest for the present work.

Inspection of the local solvation of the C═O-group shows that at the time of excitation the number (n_w) of water molecules H-bonded to the oscillator can vary between 0 and 4. The criterion for water proximity to the C═O group is that a water-oxygen atom (OW) was within 3 Å of the oxygen atom of NMA. For simulations with MTPW/KKY, the analysis of the 250 excitation trajectories yields n_w = 0 for 1% of the cases

DOI: 10.1021/jp511701z
J. Phys. Chem. B 2015, 119, 3112−3122

compared to 37% in which n_w = 1, 52% with n_w = 2, and 10% with n_w = 3. This compares with 2% for n_w = 0, 35% in which n_w = 1, 55% with n_w = 2, 8% with n_w = 3, and below 1% for 4 water molecules when FPC/TIP3P is used. The lifetimes and energy relaxation characteristics for these different occupation states will be discussed further below.

3.2. Vibrational Relaxation Times. Vibrational energy relaxation of the excited NMA was monitored by following various energy components of the system.[27,32] This was done for simulations carried out with rigid (shaked TIP3P) and flexible (KKY) water molecules as the solvent and the MTPW and FPC models for NMAD.

3.2.1. Simulations with the MTPW Model. The MTPW model for NMAD has been previously validated in condensed-phase simulations of solvation free energies and spectroscopic properties.[44] A set of point dipole and quadrupole moments were attributed to each atom of NMAD. Ewald summation is used for the point charge electrostatics,[42,57] and all bonds involving H atoms are constrained with SHAKE. Excitation is carried out along the C═O-normal mode.

First, energy relaxation was followed by monitoring various energy components averaged over 250 independent trajectories from MTPW/TIP3P simulations, which are reported in Figure 2. For this, the total energy difference $\Delta E_{tot}^{NMA} = \Delta E_{kin}^{NMA} + \Delta E_{pot}^{NMA}$ of NMAD is considered and fit to a biexponential form

$$E(t) = a \exp\left(-\frac{t}{t_1}\right) + b \exp\left(-\frac{t}{t_2}\right) + c \tag{7}$$

However, it should be noted that sometimes biexponential fits have been found to be unstable or to sensitively depend on the initial guesses for the parameters.[27] Hence, single exponential fits of the long time scale have also been carried out separately. The vibrational relaxation time from analyzing the NMA total energy ΔE_{tot}^{NMA} is 8.7 ps from MTPW/TIP3P for the picosecond component, whereas the rapid component is 260 fs. The total energy of the system is almost constant, and rigorous energy conservation could be achieved by using a smaller time step in the *NVE*−MD simulations; see the discussion in ref 42. However, over the relevant time scale for vibrational relaxation (≈ 10 ps), the total energy is well-conserved. Over the first 25 ps of the relaxation dynamics, the average kinetic energy of the water ΔE_{kin}^{WAT} increases by about 2 kcal/mol, which suggests that heating of the solvent following vibrational relaxation is not complete on this time scale.

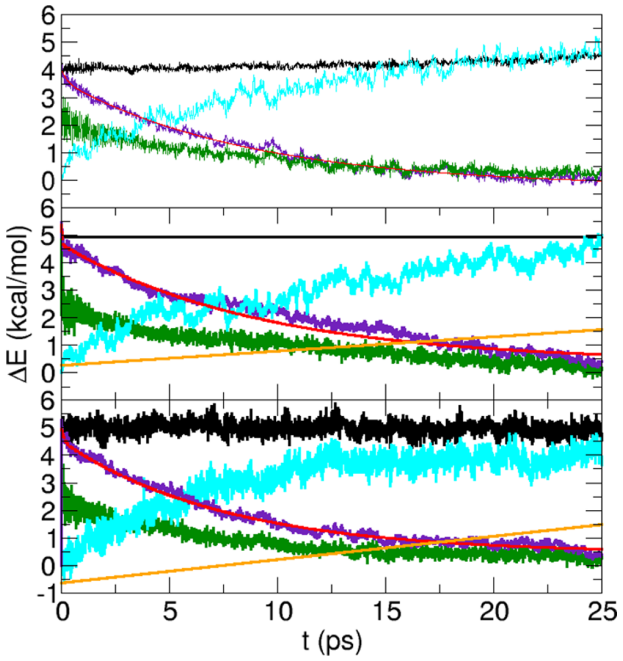


Figure 2. Averaged energy difference components from 250 individual nonequilibrium simulations for MTPW/TIP3P (top panel), FPC/TIP3P (middle panel), and FPC/KKY (bottom panel). Color code: ΔE_{tot} (black), ΔE_{tot}^{WAT} (cyan), ΔE_{tot}^{NMAD} (indigo) together with a two-time scales fit (red dashed), ΔE_{kin}^{WAT} (orange), and ΔE_{kin}^{NMAD} (green). For ΔE_{kin}^{WAT}, a linear fit (orange) is shown.

3.2.2. Simulations with the FPC Model. Next, simulations with the FPC model are discussed. Here, a more detailed analysis is carried out. This is motivated by the fact that the results differ little compared to those of the more elaborate MTPW/TIP3P simulations (see below) which require dedicated parametrizations and coding and because it has been argued that polarizability could contribute to changes in the relaxation times[27] and a fluctuating multipole model was found to perform well for CN^-.[32] A fluctuating point charge model is a first step toward such a polarizable model without, however, capturing effects of external polarization.

Figure 2 reports averaged (over the 250 trajectories) energy differences $\Delta E(t)$ between snapshots at the time of excitation $E(t = 0)$ and during the subsequent dynamics. For simulations with KKY, the total energy slightly fluctuates in the *NVE* ensemble. This fluctuation could again be reduced by using a smaller time step. However, we note that no drift in the total energy occurs, which underlines that the simulations are meaningful.

The total energy of water (cyan), it is found to increase as NMAD relaxes (indigo). The kinetic (orange) and potential (not shown) energy of the water molecules is, however, not equal as it was found for NMAD. This suggests that only part of the available energy is used to heat the solvent and the remaining energy goes into the configurational degrees of freedom of NMAD from where it relaxes on longer time scales. The interaction energy between NMAD and water, discussed further below, remains largely unchanged between the time before and after excitation.

Depending on the water model used in the simulations, the time scale on which water heating (i.e., water kinetic energy increase) occurs differs. This can be seen by comparing the orange traces in the middle and bottom panels of Figure 2 where on average ΔE_{kin}^{WAT} increases by 1.25 and 2.0 kcal/mol over 25 ps, respectively, when using a rigid TIP3P or a flexible KKY model. This amounts to a difference of about 30%. The time scale on which NMAD relaxes (red dashed line) and energy transport to the water occurs also differs to some extent. The longer time scale for this process is 9.1 ps for (FPC/TIP3P) and 7.6 ps for (FPC/KKY), see Table 1, which is a difference of close to 20%.

The decay times and amplitudes for excitation along the normal mode (NM) for the various models investigated in the present work are summarized in Table 1. In all cases, two time scales describe the vibrational relaxation. They include a rapid, subpicosecond time scale, which has an amplitude of approximately 10% of the total energy deposited, and a longer time scale with relaxation times ranging from 6 to 9 ps (except for (FPC/TIP3P-2), discussed further below) characterizing

DOI: 10.1021/jp511701z
J. Phys. Chem. B 2015, 119, 3112−3122

Table 1. Decay Times (Picoseconds) Exponential Fit of Potential and Kinetic Energies of NMAD in D_2O[a]

	a (kcal/mol)	t_1 (ps)	b (kcal/mol)	t_2 (ps)	$t_2/2.5$ (ps)	c (kcal/mol)
			NM			
CGenFF/KKY	0.68	0.08	4.06	6.79	2.72	0.49
FPC/KKY	0.43	0.13	4.07	7.58	3.03	0.44
FPC/TIP3P	0.71	0.03	4.27	9.14	3.66	0.39
FPC/TIP3P-1	0.61	0.06	3.88	6.52	2.61	0.57
FPC/TIP3P-2	0.76	0.04	5.09	18.07	7.23	−0.56
FPC(+5%)/KKY[b]	0.68	0.04	4.00	5.63	2.25	0.52
FPC(+7.5%)/KKY[b]	0.39	0.46	4.21	7.60	3.04	0.23
MTPW/TIP3P	0.40	0.26	3.84	8.72	3.49	−0.24
			CO			
CGenFF/KKY	–	–	4.67	10.55	4.22	0.30
FPC/KKY	0.25	1.36	4.82	8.08	3.23	0.26
FPC/TIP3P	–	–	5.11	10.68	4.27	0.00
FPC(+5%)/KKY[b]	–	–	4.72	7.81	3.12	0.48
FPC(+7.5%)/KKY[b]	–	–	5.33	10.41	4.16	−0.25

[a]Relaxation following normal mode (NM) or CO-bond (CO) excitation are separately reported. TIP3P-1 and TIP3P-2 analyze the data from the FPC/TIP3P trajectories but distinguish two subsets: one when a single water molecule is H-bonded to the carbonyl moiety (TIP3P-1), the second one for which two water molecules are H-bonded to the carbonyl moiety. [b]See Supporting Information; the column $t_2/2.5$ reports the long time scale accounted for quantum corrections.[29]

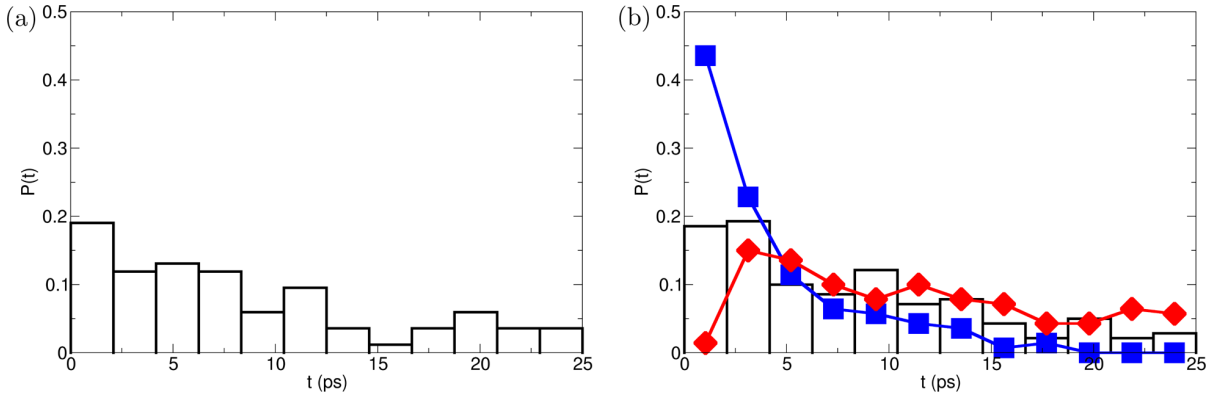


Figure 3. Probability distribution of the postexcitation lifetime of the H-bonded water molecules on the carbonyl group of NMA. (a) 1 H-bonded molecule; (b) 2 H-bonded molecules: (blue) shorter lifetime, (red) longer lifetime, and (black) distribution of the lifetime of the remaining molecule after the first left.

the genuine energy transfer process and is characterized by 90% of the amplitude.[58]

If vibrational excitation is induced by modification of the velocity vector along the C═O bond, energy relaxation is somewhat slower. Nevertheless, the time scale for relaxation is still around 10 ps and hence comparable to excitation along the normal mode.

3.2.3. Comparison with Previous Work. In experimental work[10,26] and previous atomistic simulations,[59] NMAD relaxation was found to involve up to 2 processes with different time scales when the normal mode is excited. Quantum and classical simulations of NMA relaxation by a nonequilibrium approach yield a relaxation time of 2.44 ps for the classical simulations when no zero-point correction is applied and 1.07 ps for the quantum simulations from which an empirical quantum correction factor was inferred.[29] In the present study, the number of time scales differed somewhat depending on which energy component was analyzed. The total energy of NMAD (potential plus kinetic energy) exhibits two relaxation time scales, whereas up to three time scales can be identified for the kinetic energy relaxation of NMA. In the former, the rapid component is on the femtosecond time scale (τ_1 = 40 to 460 fs), which compares favorably with the work by Cho and co-workers[26] and with Hochstrasser and co-workers[10] who find τ_1 = 450 fs comparable to the simulations by Cho and Jeon[59] who report a value of 620 fs. In both experiments and simulations, a second relaxation process on the picosecond time scale was reported, which is 4.0 ps[10] or 0.98 ps.[26] This long time scale is also found in the present work (see Table 1) and is somewhat larger than that of previous studies.[10,26] However, it agrees with results obtained by Jeon and Cho[59] where the longest time scale is 6.9 ps. In this study, the local CO mode (and not the normal mode) is excited and up to 3 time scales are found when fitting the NMA kinetic energy: 0.37, 2.3, and 6.9 ps. The latter component is found only for solvated NMA but not for NMA in the gas phase. This suggests that the short time scales involve relaxation processes within NMA while the long one involves transfer for NMA to the solvent. The present results also support the general observation that classical nonequilibrium MD simulations without zero-point energy overestimate the picosecond component of the vibrational relaxation time by a factor of 2−3 compared to experiment.[29]

3.2.4. Dynamics of Singly and Doubly H-Bonded C═O. As was mentioned above, at the moment of excitation, between

DOI: 10.1021/jp511701z
J. Phys. Chem. B 2015, 119, 3112−3122

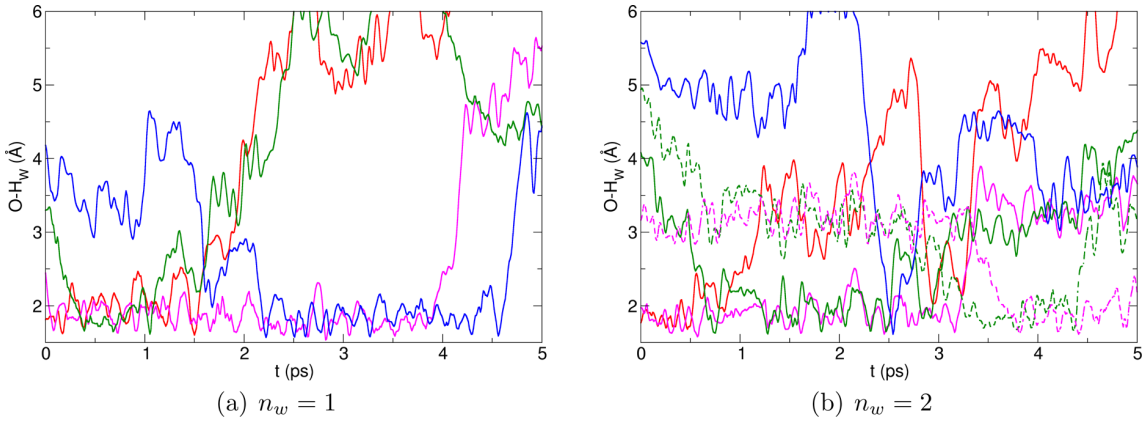


Figure 4. O_{NMA}−H_w separation of the water molecules closest to the NMA oxygen atom at $t = 0$. (left-hand side) $n_w = 1$; (right-hand side) $n_w = 2$. Dashed lines report the distance between O_{NMA} and the second H_w when a rotation of the water molecule is involved.

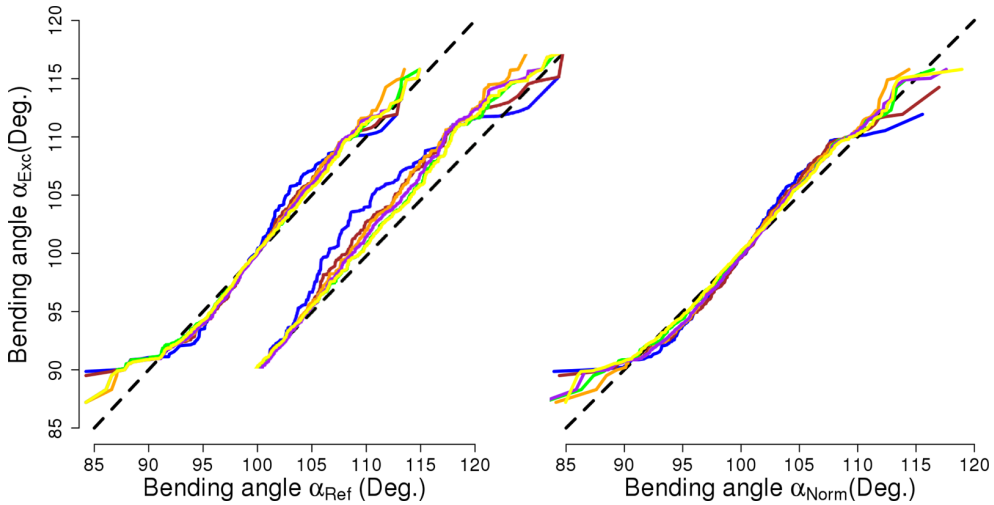


Figure 5. Q−Q plot for the water-bend-angle distribution of the closest three water molecules around the C═O group at the moment of excitation from one nonequilibrium simulation versus (left panel, α_{ref}) the three closest water molecules from an equilibrium simulation and (right panel, α_{Norm}) a Gaussian distribution. The color code refers to distributions from different times after vibrational excitation: 2.5 ps (blue), 5.0 ps (brown), 7.5 ps (orange), 10.0 ps (green), 12.5 ps (purple), and 15.0 ps (yellow). The data illustrates the nonequilibrium (left panel) and non-Gaussian (right panel) character of $p(\alpha)$ at early times (blue, brown). The black dashed line is for an ideal correlation between the nonequilibrium and the two reference distributions. The inset in the left panel shows a close-up for $\alpha_{ref} > 100°$.

$n_w = 0$ and 3 water molecules surround the C═O group, with $n_w = 1$ and $n_w = 2$ being the most probable situations. This allows the more detailed investigation of the subsequent dynamics depending on the initial configuration. First, the average lifetime of the water molecules in both situations is determined. For $n_w = 1$, the average water residence time is 9.6 ps, from which an approximate H-bond energy of 2.4 kcal/mol is inferred from $\Delta G_{AB} = -k_B T \ln (h/(k_B T \tau_{AB}))$. When two water molecules are coordinated initially, the time for the first one to leave is 4.0 ps while on average the second water molecules remains H-bonded for 12.0 ps. This corresponds to an estimated H-bond energy of 1.9 and 2.6 kcal/mol, respectively.

Figure 3 reports the distribution of the postexcitation lifetime of the H-bonded water molecules to the CO group of NMA. The left-hand panel shows the distribution for $n_w = 1$. On the right-hand side, the case for $n_w = 2$ is shown: one trace is the distribution function for the first water to leave (blue), one for the molecules that survives the longest (red), and finally one for the time the remaining molecule survives after the first left (black). The distributions between the two H-bonded molecules are different, illustrating the sequential renewal of the first solvation shell. Moreover, both distributions differ from the one obtained for a single bound molecule. Nevertheless, starting from $n_w = 2$ initially, the distribution of the lifetime of the second water molecule after the departure of the first one is similar to that for $n_w = 1$.

3.3. Solvent−Solute Energy Redistribution. The total energy change for NMAD ($\Delta E_{kin}^{NMA} + \Delta E_{pot}^{NMA}$) was analyzed in two ways. First, as an (unspecific) ensemble average over all 250 trajectories (i.e., the curve that would be obtained from experiment) and second, separately for those with $n_w = 1$ and $n_w = 2$ at the moment of excitation. Energy relaxation is more rapid for $n_w = 1$ compared to the unspecific ensemble average, whereas for $n_w = 2$ it is slower. As the proportion is approximately 1/3 and 2/3 for $n_w = 1$ and $n_w = 2$, respectively, weighting of the two relaxation curves should yield a decay close to that described by the ensemble average, which is also

DOI: 10.1021/jp511701z
J. Phys. Chem. B 2015, 119, 3112−3122

what is found. Hence, depending on whether the C═O group is solvated by one or two water molecules at the moment of excitation, the relaxation behavior differs significantly.

It is also of interest to analyze the difference in the interaction energy between NMAD and the TIP3P water molecules between t = 0 and after vibrational excitation. Analysis of all 250 trajectories finds that this energy difference is close to 0 but slightly positive (see Figure S1 in Supporting Information), i.e., destabilizing. On the other hand, for the situation with n_w = 1, at the moment of excitation a stabilization of the system by ≈1.9 kcal/mol is found, whereas for n_w = 2 at the moment of excitation, the total interaction energy becomes slightly positive (≈0.5 kcal/mol), or destabilizing. Investigation of typical structures suggests that when starting from n_w = 1 (see red trace in Figure 4a), vibrational relaxation leads to rapid replacement of the H-bonded water molecule and a second water molecule forms an H-bond, which yields an overall stabilization because one strong H-bond is replaced by two somewhat weaker ones. On the other hand, when starting from n_w = 2, continuous exchange of water molecules leads to a more or less constant occupation by two water molecules; hence, the differential energy before and after vibrational excitation remains essentially unchanged.

3.3.1. Coupling between the C═O Stretch and the Water Bend. The only energetically feasible pathway between vibrationally excited NMAD and the internal solvent degrees of freedom is energy transfer to the water-bending mode. This is due to the proximity of the wavenumbers of the two modes which are at 1731 and 1595 cm^{-1} for NMA[53] and water, respectively. Such effects can be observed only when the water molecules in the simulations are flexible, which is the case for the KKY model. Involvement of internal solvent degrees of freedom have been mentioned but not quantified in previous work.[27] Nonequilibrium relaxation of the C═O local oscillator to a nearby (H-bonded) solvent molecule will lead to a bend-excited water molecule which subsequently relaxes. One characteristic of a nonequilibrium simulation is the fact that the distribution of a particular coordinate, for example, the bending angle $p(\alpha)$ of the water molecules, deviates from its equilibrium distribution. One convenient way to represent this are quantile−quantile (Q−Q) plots, which measure the deviation of a given distribution from a reference distribution, which is the equilibrium distribution of water-bending angles in the present case.[60] Such a Q−Q plot is shown in the left panel of Figure 5. It is evident that with increasing time (t) after vibrational excitation (represented by the color coding in Figure 5) the deviation from the reference distribution decreases and therefore approaches equilibrium. The same is true if the nonequilibrium distribution is compared with a Gaussian reference distribution, which is illustrated in the right-hand panel of Figure 5.

To better quantify and illustrate energy transfer to the water bending mode, the $p(\alpha)$ for the three closest water molecules around the C═O group at the moment of excitation was determined for 2.5, 5.0, 7.5, 10.0, 12.5, and 15.0 ps after excitation (Figure 6). As a reference, the three closest water molecules in an equilibrium simulation were analyzed in the same fashion. Both averaged distributions were fitted to a log-normal distribution

$$p(x) = \frac{k}{x}\exp\left(\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \quad (8)$$

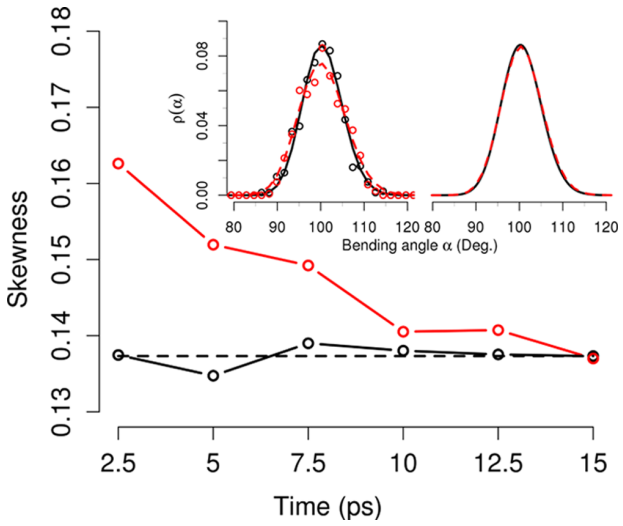


Figure 6. Insets report fitted probability distributions $p(\alpha)$ for the water-bending angle for the three closest water molecules at the time of excitation (red) and from equilibrium simulations (black). The $p(\alpha)$ are determined over increasing time intervals after the time of excitation, with the left panel for 2.5 ps and the right panel for 10.0 ps after excitation. The fits are log-normal distributions with residual sums of squares of 3.8×10^{-4} and 5.1×10^{-4} for the black and red curves in the left panel, respectively, and 1.1×10^{-4} and 6.7×10^{-5} in the right panel, respectively. The main figure reports the skewness as a function of time after excitation and confirms that the skewness in the vibrationally excited trajectories decreases as a function of time to a level corresponding to an equilibrium simulation.

where μ and σ are the mean and standard deviation of $\ln(\alpha)$, respectively; k is an overall scaling parameter, and the skewness parameter $(e^{\sigma^2} + 2)(e^{\sigma^2} - 1)^{1/2}$ is followed as a function of time after excitation. Both distributions have finite skewness, which implies that they are non-Gaussian. The nonequilibrium distribution has a larger skewness which decays toward the equilibrium value on the 10 ps time scale. Fitting the data to a reference Gaussian distribution also yields a satisfactory fit around the maximum of the data but deviates significantly in the wings.

Such an analysis does not provide detailed information about the time scale on which energy is transferred from the vibrationally excited NMAD to the surrounding water molecules. All that can be said is that it must be shorter than 2.5 ps. As a certain amount of data is required for reliable statistics on $p(\alpha)$, extending this analysis to shorter time scales is usually not meaningful because of increased uncertainties due to the small data sets available at short times after excitation.

3.4. Energy Migration. Atomistic simulations are particularly useful for characterizing quantities that are not directly observable experimentally. One such property is energy transfer in an energized system. For the present case a direct measure for energy migration can be obtained from analyzing the water velocity distribution $\rho_v(\mathbf{r},t)$ before and after excitation. In the following (eq 9) averages of the velocity vector amplitude density are discussed.

$$\rho_v(\mathbf{r}, t) = \langle \| \mathbf{v}(\mathbf{r}, t) \| \rangle = \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbf{v}(\mathbf{r}, i) \quad (9)$$

Here, N_t is the number of frames over which the absolute value of the velocity vector is averaged (i.e., averaging over Δt =

DOI: 10.1021/jp511701z
J. Phys. Chem. B 2015, 119, 3112−3122

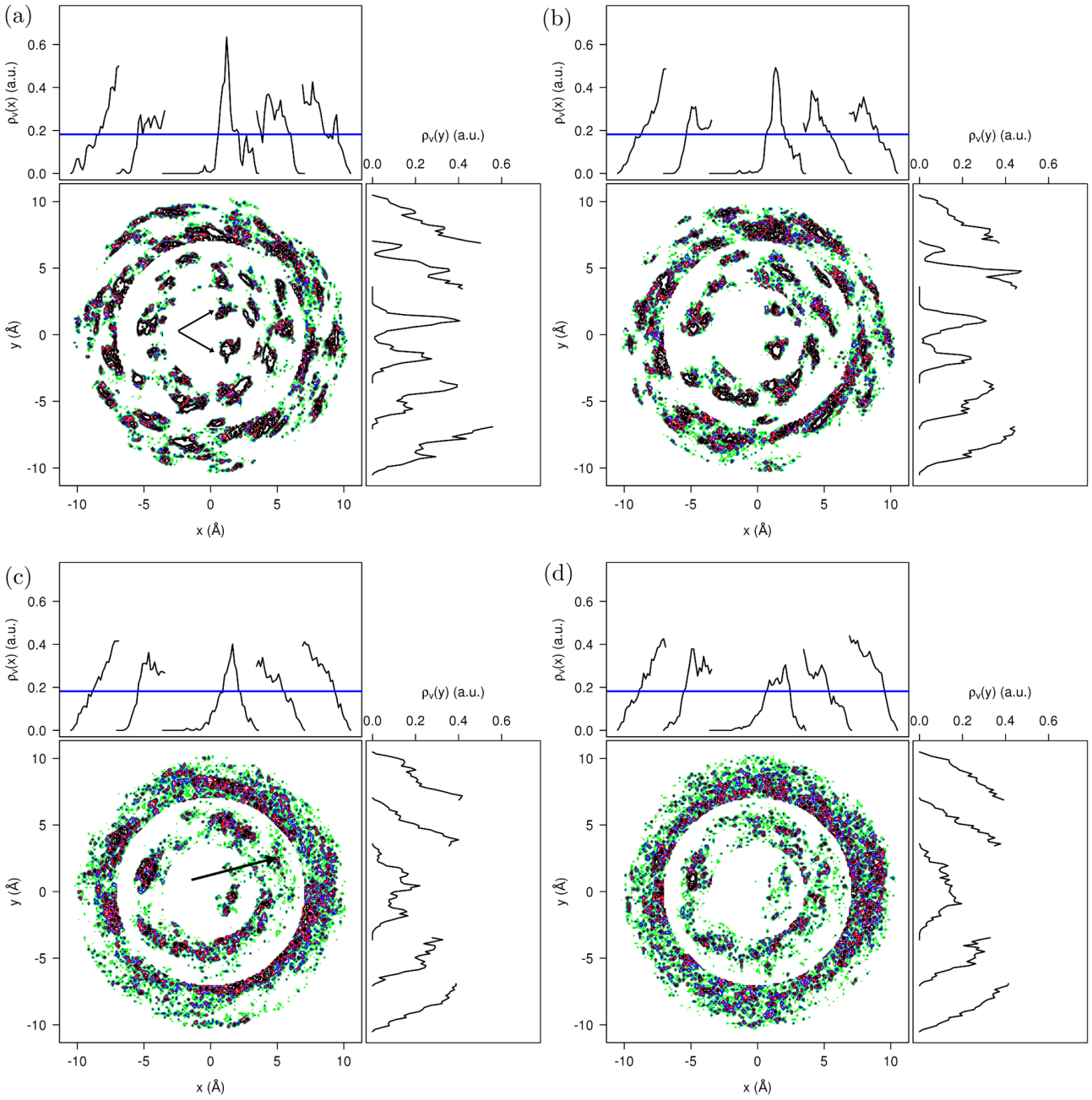


Figure 7. Isocontours of water velocity distribution amplitude as a function of simulation time from the nonequilibrium part of the trajectory. The color code is as follows: white (0.0−0.3), green (0.3−0.35), blue (0.4−0.45), red (0.5−0.55), and black (0.6−1.0). The black curves are the projection of the velocity density amplitude. The concentric circles are a consequence of analyzing water molecules in their respective solvent shells (see text). The horizontal blue line is to guide the eye, and the arrows point toward features discussed in the text.

0.25 ps); the index v in ρ_v refers to the distribution of velocity vector amplitudes, and the quantity ρ_v contains information about the total solvent kinetic energy.

Figure 7 and Figure S2 in Supporting Information report amplitude densities of the water velocity distributions from nonequilibrium and equilibrium simulations, respectively. For this, the NMA-oxygen atom is the origin (0/0) of the coordinate system with the C═O group pointing along the +x-axis. The water shells, determined in recent work from $g(r)$,[44] are defined as follows: first shell within 3.5 Å of the NMA-oxygen atom; second shell between 3.5 and 7.0 Å of the NMA-oxygen atom; third shell between 7.0 and 10.5 Å of the NMA-oxygen atom. Such a procedure leads to the concentric circles in Figure S2 and Figure 7 instead of homogeneous distributions but allow the representation of relative changes between magnitudes and granularity of the velocity amplitude distributions. The spatial resolution of the grid is 0.15 Å, and the density is averaged over 25 consecutive frames (0.25 ps in total) to provide sufficient time resolution but at the expense of a relatively low spatial resolution. NMA is drawn (see Figure S2) to indicate its orientation. On the top and the right-hand side of each contour, the projection of the density on the

DOI: 10.1021/jp511701z
J. Phys. Chem. B 2015, 119, 3112−3122

corresponding axis is shown. The black curves correspond to x- and y-projections of the velocity amplitude density. The distributions within each shell are quite homogeneous and stationary $\rho_v(\mathbf{r},t)$ (eq 9) as a function of time.

The situation changes considerably after vibrational excitation at $t = 0$. The nonequilibrium relaxation of the water velocity density amplitude (eq 9) at times $t = 0.25$, 0.50, 0.75, and 1.0 ps after excitation is reported in Figure 7. During vibrational relaxation, energy migration within the water shells is obvious, as can be seen from comparing Figure S2 in Supporting Information and Figure 7d with Figure 7a–c. Energy migration is evident by the localized high-velocity peaks (black and dark areas in the contour plots) at early time which diminish as a function of time. The projections clearly show the propagation of the excess energy from NMA to the water molecules H-bonded to the carbonyl group (located at $x \approx 0$ and indicated by an arrow in Figure 7a) and then further toward the rest of the first solvation shell. As an example, the projection of the velocity amplitude distribution function in the first solvation shell during equilibrium dynamics is $\rho_v \approx 0.4$, whereas at 0.25 ps after vibrational excitation it has increased to $\rho_v = 0.7$. Over the course of the ensuing relaxation dynamics this peak decreases to the equilibrium value on the 1 ps time scale (Figure 7d).

Detailed consideration of the maximum amplitude in the first and second solvation shells along the positive x-axis (i.e., along the C═O bond) in Figure 7 also suggests that local heating takes place. This can be seen in the decrease of the maximum value of ρ_v in the first solvation shell as a function of time with concomitant increase in the same quantity for the second solvation shell. These features are equally pronounced when considering the projections along the y-direction. Also, contrary to the equilibrium situation, the concentric rings of the velocity density are distorted and a density maximum emerges at $x \approx$ 5 Å. Moreover, the fact that the density of the averaged velocity vector $\rho_v(\mathbf{r},t)$ (not shown) is smaller than the vector amplitude, ρ_v $(\mathbf{r},t)$, indicates that shortly after excitation water molecules have disordered velocity vectors. This differs from the equilibrium trajectory where because of the slowly varying H-bonding pattern the averaged velocity vectors and their amplitude are similar. Within 1 ps most of the local energy redistribution around the C═O group is completed; however, the dynamics and energy exchange continues.

Integrating the velocity density over the first and second solvation shells (as indicated by the blue dashed lines in Figure S2 in Supporting Information) as a function of time after excitation suggests that energy migrates in a shock-wave-like fashion outward. This is illustrated in Figure 8. During the first picosecond, the integrated velocity density $S(\rho_v(\mathbf{r},t))$ in the first solvation shell decreases (black–red–green–blue), whereas in the second solvation shell $S(\rho_v(\mathbf{r},t))$ typically increases (largest for 0.50 and 0.75 ps after excitation and smaller for 0.25 and 1.0 ps). Hence, energy transfer between the two shells considered occurs on the 500 fs time scale. The fact that the velocity density decreases again after 1 ps suggests that energy is further transported into subsequent solvent layers.

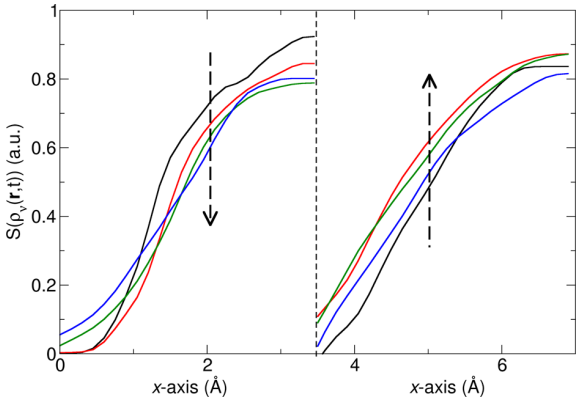


Figure 8. Integrated projections $S(\rho_v(\mathbf{r}, t))$ of $\rho_v(\mathbf{r}, t)$ onto the x-axis (see Figure 7) within the first (left panel) and second shell (right panel). For definition, see caption to Figure 7. The four times after excitation [0.25 ps (black), 0.50 ps (red), 0.75 ps (green), and 1.00 ps (blue)] are shown.

4. CONCLUSIONS

The vibrational relaxation of NMA in the amide I region was investigated from atomistic simulations with validated force fields. Excitation along both the normal and the local C═O mode were considered. The two relaxation time scales are typically on the subpicosecond time scale and between 6 and 10 ps. Depending on the water model used in the simulations (flexible KKY or rigid/shaked TIP3P), the time scale on which water heating (i.e., water kinetic energy increase) occurs differs by about 30% with the faster pathway being the one which allows inter- and intramolecular relaxation. Furthermore, the fact that bifurcation (inter- versus intramolecular) in the relaxation pathway can occur is reminiscent of bifurcating pathways in reactivity as has been previously observed.[61] Energy transfer to the solvent has been followed by considering the two-dimensional velocity distribution of the surrounding water molecules. A nonequilibrium distribution can be clearly identified at early times (0.25 ps) after vibrational excitation which decays toward an equilibrium state on the picosecond time scale. The ensuing nonequilibrium distribution in the bending angles of the water molecules H-bonded to the chromophore at the time of excitation decays on a picosecond time scale. Depending on the H-bonding pattern of the solvent molecules closest to the C═O group, different kinetics is found for the water dynamics after vibrational excitation. For $n_w = 1$, which makes up ≈33% of the population, the average lifetime is 9.6 ps, whereas for $n_w = 2$ (56% of the population), two lifetimes (4.0 and 12.0 ps) were found. Experimentally, a superposition of the relaxation dynamics of these two states is observed. The results for the different force fields of the present work, the sensitivity analysis (see Supporting Information), and the results from previous work which considered vibrational energy relaxation in NMA and different solvents[27] suggest that vibrational relaxation times are not sufficiently sensitive for detailed force field refinements.

In summary, the present work provides an atomistically resolved picture of the vibrational relaxation of NMA and subsequent solvent dynamics. When quantum corrections[29] for the long (picosecond) time scale are accounted for, the simulation results quantitatively agree with experiment. Analysis of energy migration pathways shows that vibrational relaxation of NMA exhibits subpicosecond dynamics resulting in impulsive propagation of the excess energy into the surrounding solvent followed by excitation of the water-bending mode on the picosecond time scale.

DOI: 10.1021/jp511701z
J. Phys. Chem. B 2015, 119, 3112–3122

■ ASSOCIATED CONTENT

Supporting Information

Electrostatic parameters for NMA; averaged water−NMA interaction energy from 250 nonequilibrium runs from FPC/TIP3P and FPC/KKY simulations; isocontours of water velocity distribution amplitude as a function of simulation time from the equilibrium part of the trajectory; and total energy difference for NMA and water for different modified force field parametrizations. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Fruitful discussions with Peter Hamm and Gerhard Stock are gratefully acknowledged. The authors gratefully acknowledge financial support from the Swiss National Science Foundation through Grant 200021-117810 and to the NCCR-MUST.

■ REFERENCES

(1) Oxtoby, D. *Vibrational Relaxation in Liquids.* **1981**, *32*, 77−101.
(2) Stratt, R.; Maroncelli, M. Nonreactive dynamics in solution: The emerging molecular view of solvation dynamics and vibrational relaxation. *J. Phys. Chem.* **1996**, *100*, 12981−12996.
(3) Heilweil, E. J.; Doany, F. E.; Moore, R.; Hochstrasser, R. M. Vibrational Energy Relaxation of the Cyanide Ion in Aqueous Solution. *J. Chem. Phys.* **1982**, *76*, 5632−5634.
(4) Li, M.; Owrutsky, J.; Sarisky, M.; Culver, J. P.; Yodh, A.; Hochstrasser, R. M. Vibrational and rotational relaxation times of solvated molecular ions. *J. Chem. Phys.* **1993**, *98*, 5499−5507.
(5) Egorov, S.; Skinner, J. Vibrational energy relaxation of polyatomic solutes in simple liquids and supercritical fluids. *J. Chem. Phys.* **2000**, *112*, 275−281.
(6) Heilweil, E. J.; Cavanagh, R. R.; Stephenson, J. C. Population Relaxation of Co(v = 1) Vibrations in Solution Phase Metal Carbonyl Complexes. *Chem. Phys. Lett.* **1987**, *134*, 181−188.
(7) Sagnella, D. E.; Straub, J. E.; Jackson, T. A.; Lim, M.; Anfinrud, P. A. Vibrational Population Relaxation of Carbon Monoxide in the Heme Pocket of Photolyzed Carbonmonoxy Myoglobin: Comparison of Time-Resolved mid-IR Absorbance Experiments and Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 14324−14329.
(8) Owrutsky, J. C.; Li, M.; Locke, B.; Hochstrasser, R. M. Vibrational Relaxation of the CO Stretch Vibration in Hemoglobin-Co, Myoglobin-Co, and Protoheme-Co. *J. Phys. Chem.* **1995**, *99*, 4842−4846.
(9) Mizutani, Y.; Kitagawa, T. Ultrafast dynamics of myoglobin probed by time-resolved resonance Raman spectroscopy. *Chem. Rec.* **2001**, *1*, 258−275.
(10) Hamm, P.; Lim, M.; Hochstrasser, R. M. Structure of the Amide I Band of Peptides Measured by Femtosecond Nonlinear-Infrared Spectroscopy. *J. Phys. Chem. B* **1998**, *5647*, 6123−6138.
(11) Zanni, M. T.; Asplund, M. C.; Hochstrasser, R. Two-Dimensional Heterodyned and Stimulated Infrared Photon Echoes of N-Methylacetamide-D. *J. Chem. Phys.* **2001**, *114*, 4579−4590.
(12) Hamm, P.; Helbing, J.; Bredenbeck, J. Two-Dimensional Infrared Spectroscopy of Photoswitchable Peptides. *Annu. Rev. Phys. Chem.* **2008**, *59*, 291−317.
(13) Owrutsky, J. C.; Raftery, D.; Hochstrasser, R. M. Vibrational-Relaxation Dynamics in Solutions. *Annu. Rev. Phys. Chem.* **1994**, *45*, 519−555.
(14) Lawrence, C. P.; Skinner, J. L. Vibrational spectroscopy of HOD in liquid D_2O. III. Spectral diffusion, and hydrogen-bonding and rotational dynamics. *J. Chem. Phys.* **2003**, *118*, 264−272.
(15) Devereux, M.; Meuwly, M. Force Field Optimization Using Dynamics and Ensemble Averaged Data: Vibrational Spectra and Relaxation in Bound MbCO. *J. Chem. Inf. Model.* **2010**, *50*, 349−357.
(16) Rey, R.; Hynes, J. T. Vibrational Phase and Energy Relaxation of CN^- in Water. *J. Chem. Phys.* **1998**, *108*, 142−153.
(17) Krimm, S.; Bandekar, J. Vibrational Spectroscopy and Conformation of Peptides, Polypeptides, and Proteins. *Adv. Protein Chem.* **1986**, *38*, 181.
(18) Peterson, K. A.; Rella, C. W.; Engholm, J. R.; Schwettman, H. A. Ultrafast Vibrational Dynamics of the Myoglobin Amide I Band. *J. Phys. Chem. B* **1999**, *103*, 557−561.
(19) Moran, A.; Park, S.-M.; Mukamel, S. Infrared Photon Echo Signatures of Hydrogen Bond Connectivity in the Cyclic Decapeptide Antamanide. *J. Chem. Phys.* **2003**, *118*, 9971−9980.
(20) Moran, A.; Mukamel, S. The Origin of Vibrational Mode Couplings in Various Secondary Structural Motifs of Polypeptides. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 506−510.
(21) Xie, A.; van der Meer, L.; Hoff, W.; Austin, R. H. Long-Lived Amide I Vibrational Modes in Myoglobin. *Phys. Rev. Lett.* **2000**, *84*, 5435−5438.
(22) Austin, R. H.; Xie, A.; van der Meer, L.; Redlich, B.; Lindgard, P. A.; Frauenfelder, H.; Fu, D. Picosecond Thermometer in the Amide I Band of Myoglobin. *Phys. Rev. Lett.* **2005**, *94*, 128101.
(23) Nguyen, P. H.; Stock, G. Nonequilibrium molecular-dynamics study of the vibrational energy relaxation of peptides in water. *J. Chem. Phys.* **2003**, *119*, 11350−11358.
(24) Corcelli, S. A.; Lawrence, C. P.; Skinner, J. L. Combined Electronic Structure/Molecular Dynamics Approach for Ultrafast Infrared Spectroscopy of Dilute HOD in Liquid H_2O and D_2O. *J. Chem. Phys.* **2004**, *120*, 8107−8117.
(25) Hayashi, T.; Zhuang, W.; Mukamel, S. Electrostatic DFT Map for the Complete Vibrational Amide Band of NMA. *J. Phys. Chem. A* **2005**, *109*, 9747−9759.
(26) Decamp, M. F.; Deflores, L.; Mccracken, J. M.; Tokmakoff, A.; Kwac, K.; Cho, M. Amide I Vibrational Dynamics of *N*-Methylacetamide in Polar Solvents: The Role of Electrostatic Interactions. *J. Phys. Chem. B* **2005**, *109*, 11016−11026.
(27) Park, S.-M.; Nguyen, P. H.; Stock, G. Molecular dynamics simulation of cooling: Heat transfer from a photoexcited peptide to the solvent. *J. Chem. Phys.* **2009**, *131*.
(28) Zhang, Y.; Fujisaki, H.; Straub, J. E. Mode-Specific Vibrational Energy Relaxation of Amide I′ and II″ Modes in *N*-Methylacetamide/Water Clusters: Intra- and Intermolecular Energy Transfer Mechanisms. *J. Phys. Chem. A* **2009**, *113*, 3051−3060.
(29) Stock, G. Classical Simulation of Quantum Energy Flow in Biomolecules. *Phys. Rev. Lett.* **2009**, *102*, 118301.
(30) Bastida, A.; Soler, M. A.; Zuniga, J.; Requena, A.; Kalstein, A.; Fernandez-Alberti, S. Hybrid Quantum/Classical Simulations of the Vibrational Relaxation of the Amide I Mode of *N*-Methylacetamide in D_2O Solution. *J. Phys. Chem. B* **2012**, *116*, 2969−2980.
(31) David, E. F.; Stratt, R. M. The anharmonic features of the short-time dynamics of fluids: The time evolution and mixing of instantaneous normal modes. *J. Chem. Phys.* **1998**, *109*, 1375−1390.
(32) Lee, M. W.; Meuwly, M. On the Role of Nonbonded Interactions in Vibrational Energy Relaxation of Cyanide in Water. *J. Phys. Chem. A* **2011**, *115*, 5053−5061.
(33) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D., Jr. CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010**, *31*, 671−690.
(34) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(35) Kumagai, N.; Kawamura, K.; Yokokawa, T. An Interatomic Potential Model for H_2O: Applications to Water and Ice Polymorphs. *Mol. Simul.* **1994**, *12*, 177−186.

DOI: 10.1021/jp511701z
J. Phys. Chem. B 2015, 119, 3112−3122

(36) Gupta, P. K.; Meuwly, M. Dynamics and vibrational spectroscopy of water at hydroxylated silica surfaces. *Faraday Discuss.* **2013**, *167*, 329−346.
(37) Danielsson, J.; Meuwly, M. Energetics and Dynamics in MbCN: CN^--Vibrational Relaxation from Molecular Dynamics Simulations. *J. Phys. Chem. B* **2007**, *111*, 218−226.
(38) Nutt, D. R.; Meuwly, M. Theoretical Investigation of Infrared Spectra and Pocket Dynamics of Photodissociated Carbonmonoxy Myoglobin. *Biophys. J.* **2003**, *85*, 3612−3623.
(39) Plattner, N.; Meuwly, M. The Role of Higher CO-Multipole Moments in Understanding the Dynamics of Photodissociated Carbonmonoxide in Myoglobin. *Biophys. J.* **2008**, *94*, 2505−2515.
(40) Lutz, S.; Nienhaus, K.; Nienhaus, G. U.; Meuwly, M. Ligand Migration Between Internal Docking Sites in Photodissociated Carbonmonoxy Neuroglobin. *J. Phys. Chem. B* **2009**, *113*, 15334−15343.
(41) Plattner, N.; Meuwly, M. Quantifying the Importance of Protein Conformation on Ligand Migration in Myoglobin. *Biophys. J.* **2012**, *102*, 333−341.
(42) Bereau, T.; Kramer, C.; Meuwly, M. Leveraging Symmetries of Static Atomic Multipole Electrostatics in Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2013**, *9*, 5450−5459.
(43) Kramer, C.; Gedeck, P.; Meuwly, M. Atomic Multipoles: Electrostatic Potential Fit, Local Reference Axis Systems and Conformational Dependence. *J. Comput. Chem.* **2012**, *33*, 1673−1688.
(44) Cazade, P.-A.; Bereau, T.; Meuwly, M. Computational Two-Dimensional Infrared Spectroscopy without Maps: *N*-Methylacetamide in Water. *J. Phys. Chem. B* **2014**, *118*, 8135−8147.
(45) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B. et al., *Gaussian 03*, revision B.01; Gaussian, Inc.: Wallingford, CT, 2003.
(46) Singh, U. C.; Kollman, P. A. An Approach to Computing Electrostatic Charges for Molecules. *J. Comput. Chem.* **1984**, *5*, 129−145.
(47) Besler, B. H.; Merz, K. M.; Kollman, P. A. Atomic Charges Derived from Semiempirical Methods. *J. Comput. Chem.* **1990**, *11*, 431−439.
(48) Nutt, D. R.; Karplus, M.; Meuwly, M. Potential Energy Surface and Molecular Dynamics of MbNO: Existence of an Unsuspected FeON Minimum. *J. Phys. Chem. B* **2005**, *109*, 21118−21125.
(49) Nutt, D. R.; Meuwly, M. Ferric and Ferrous Iron in Nitroso-Myoglobin: Computer Simulations of Stable and Metastable States and their Infrared Spectra. *ChemPhysChem* **2007**, *8*, 527−536.
(50) Mishra, S.; Meuwly, M. Atomistic Simulation of NO Dioxygenation in Group I Truncated Hemoglobin. *J. Am. Chem. Soc.* **2010**, *132*, 2968−82.
(51) Cazade, P.-A.; Meuwly, M. Oxygen Migration Pathways in NO-bound Truncated Hemoglobin. *ChemPhysChem* **2012**, 4276−4286.
(52) A. D. Mackerell, J.; Bashford, D.; Bellott, M.; R. L. Dunbrack, J.; Evan-Seck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-Mccarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; W. E. Reiher, I.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(53) Kubelka, J.; Keiderling, T. A. Ab Initio Calculation of Amide Carbonyl Stretch Vibrational Frequencies in Solution with Modified Basis Sets. 1. *N*-Methyl Acetamide. *J. Phys. Chem. A* **2001**, *105*, 10922−10928.
(54) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545−1614.
(55) van Gunsteren, W.; Berendsen, H. Algorithms for macromolecular dynamics and constraint dynamics. *Mol. Phys.* **1977**, 1311−1327.
(56) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(57) Sagui, C.; Pedersen, L. G.; Darden, T. A. Towards an accurate representation of electrostatics in classical force fields: Efficient implementation of multipolar interactions in biomolecular simulations. *J. Chem. Phys.* **2004**, *120*, 73.
(58) Owrutsky, J. C.; Raftery, D.; Hochstrasser, R. M. Vibrational Relaxation Dynamics in Solutions. *Annu. Rev. Phys. Chem.* **1994**, *45*, 519−555.
(59) Jeon, J.; Cho, M. Redistribution of carbonyl stretch mode energy in isolated and solvated N-methylacetamide: Kinetic energy spectral density analyses. *J. Chem. Phys.* **2011**, *135*, 214504.
(60) Wilk, M. B.; Gnanadesikan, R. Probability plotting methods for the analysis for the analysis of data. *Biometrika* **1968**, *55*, 1−17.
(61) Nienhaus, K.; Lutz, S.; Meuwly, M.; Nienhaus, G. U. Reaction-Pathway Selection in the Structural Dynamics of a Heme Protein. *Chem.—Eur. J.* **2013**, *19*, 3558−3562.

DOI: 10.1021/jp511701z
J. Phys. Chem. B 2015, 119, 3112−3122

Supporting Information: Vibrational Relaxation and Energy Migration of N-methylacetamide in Water: The Role of Nonbonded Interactions

Pierre-André Cazade, Florent Hédin, Zhen-Hao Xu, and Markus Meuwly

Department of Chemistry, University of Basel,

Klingelbergstrasse 80, 4056 Basel, Switzerland

	C	O	C(CO)	N	C(N)	H(CCO)	H(N)	H(CN)
q_i	0.714	-0.567	-0.707	-0.430	-0.032	0.187	0.278	0.061
a_0	1.899	-0.469	-1.468	-1.255	0.437	0.265	0.301	-0.080
a_1	-0.980	-0.079	0.636	0.681	-0.383	-0.067	-0.022	0.116

TABLE I: Electrostatic parameters for NMA in units of e.

This file contains one table and 3 additional figures.

The parameters obtained from the linear fit are reported on Table I.

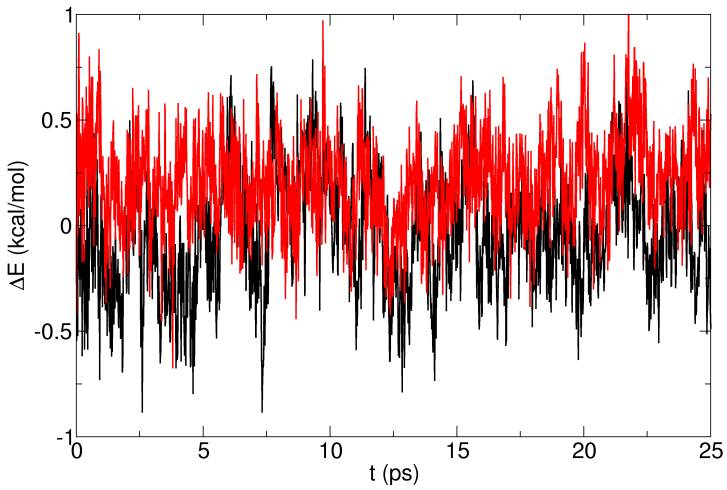
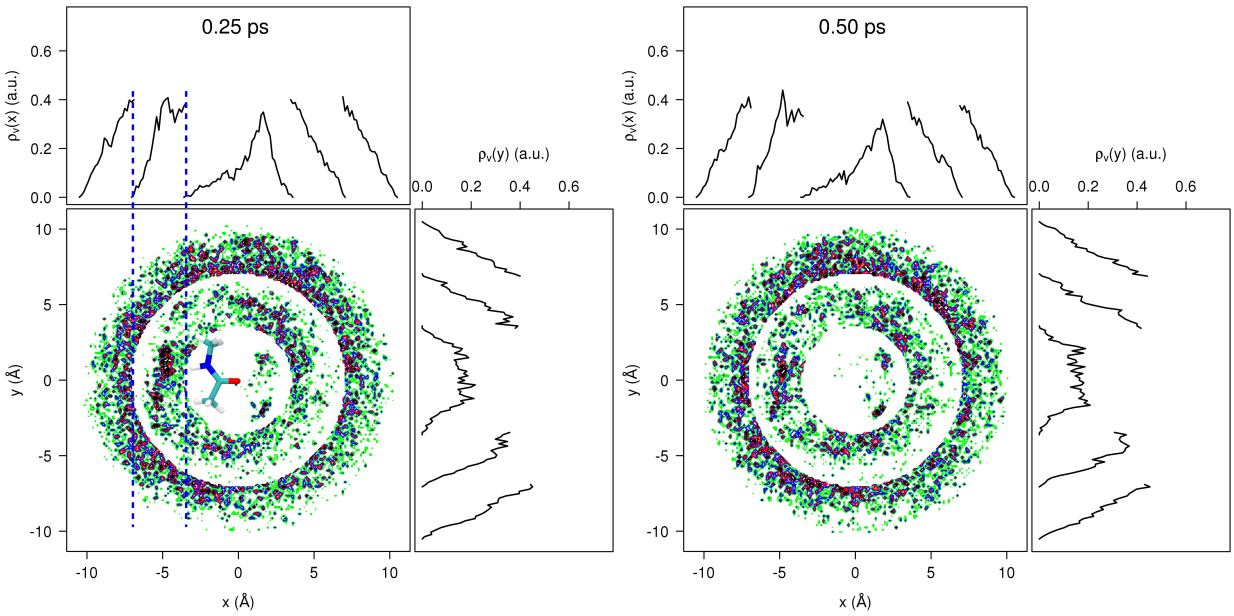


FIG. 1: Averaged water-NMA interaction energy from 250 non-equilibrium runs from FPC/TIP3P (black) and FPC/KKY (red) simulations.



(a) (b) (c) (d)

FIG. 2: Isocontours of water velocity distribution amplitude as a function of simulation time from the equilibrium part of the trajectory. The color code is as follows: white (0.0-0.3), green (0.3-0.35), blue (0.4-0.45), red (0.5-0.55), black (0.6-1.0). The black curves are the projection of the velocity density amplitude. The concentric circles are a consequence of analyzing water molecules in their respective solvent shells. Dashed blue lines define the solvent shells. For more detail see text.

Sensitivity to the Force Field Parametrization

The details of the force field parametrization have been found to affect the vibrational relaxation times of small molecules in solution, such as for CN^- in water.[1,2] In this particular case, the importance of electrostatic interactions has been established for some time[1] and confirmed in more recent work using multipolar force fields.[2] In addition to the electrostatic interactions it was also found that the van der Waals parameters sensitively affect vibrational relaxation times and the solvation free energy.[3] This contrasts with other properties, such as the 1d- or 2d-infrared spectra which are mostly sensitive to the electrostatics and less affected by van der Waals interactions.[4]

Hence, the sensitivity of the present results with respect to modifications of the van der Waals radii was also studied. As in previous work, the vdW radii of the C- and O-atom of the carbonyl group were increased by 5 and 7.5%, respectively, in order to probe the dependence of the physical observables on van der Waals ranges. Again, 250 non-equilibrium trajectories were run and analyzed for each case, with KKY as the force field for water.

Figure **S3** reports the energy difference relative to the last frame of the equilibrium simulation for the total energy of NMAD and for all water molecules from which relaxation times are determined as was done before. Analysis of the decay times suggests that changes in the van der Waals ranges indeed influence the relaxation times which first decrease to 5.6 ps (for a 5 % increase in the radii) and then increase to 7.6 ps (for a 7.5% increase). These changes should be contrasted with a factor of about 5 in lengthening the vibrational relaxation time when scaling the van der Waals radii by 7.5 % on CN^-.[2] Hence, for NMA the dependence of the vibrational relaxation time on the van der Walls radii appears to be small. The variations of relaxation times may also be affected by the fact that they quite sensitively depend on the number of water molecules n_W bound to the CO at the time of excitation. Contrary to CN^- in solution, NMAD can vibrationally relax even if the water solvent is described by a rigid TIP3P water model. However, the energy can relax into low-frequency solvent degrees of freedom and into internal NMAD degrees of freedom. The latter is not possible for CN^- where coupling to the low-frequency solvent modes is weak and hence relaxation is very slow for simulations with rigid water.

---

[1] Rey, R.; Hynes, J. T. *J. Chem. Phys.* **1998**, *108*, 142–153.

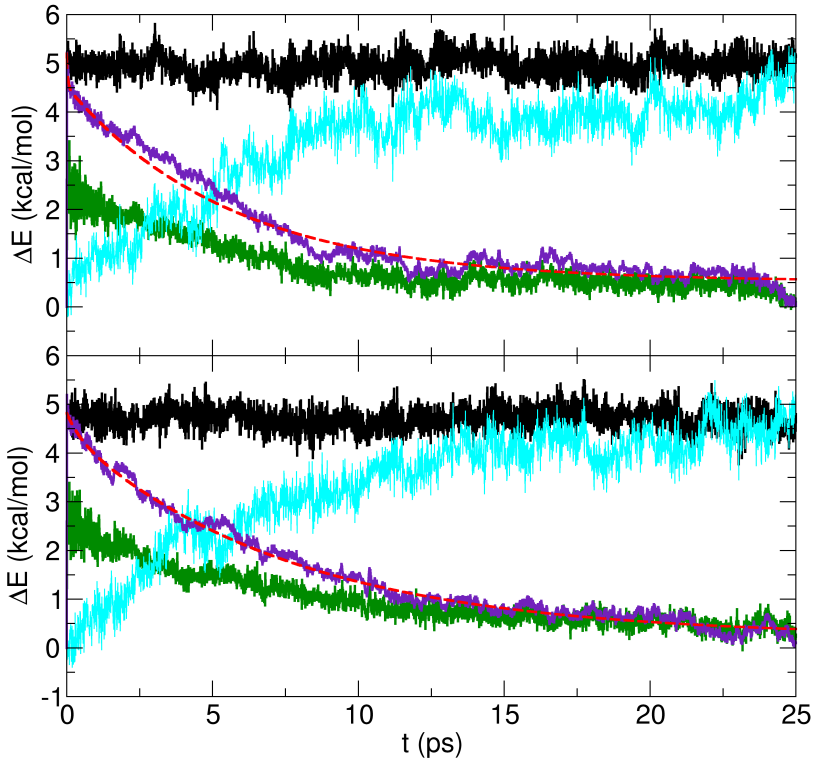


FIG. 3: Total energy difference for NMA (indigo) and water (cyan) for different modified force field parametrizations. All results are averaged over 250 individual nonequilibrium trajectories. Simulations with FPC/KKY and vdW radii on the C- and O-atom of the CO-moiety increased by 5% (top) and 7.5% (bottom).

[2] Lee, M. W.; Meuwly, M. *J. Phys. Chem. A* **2011**, *115*, 5053–5061.

[3] Lee, M. W.; Meuwly, M. *Phys. Chem. Chem. Phys.* **2013**, *15*, 20303–20312.

[4] Lee, M. W.; Carr, J. K.; Göllner, M.; Hamm, P.; Meuwly, M. *J. Chem. Phys.* **2013**, *139*, 054506.

BIBLIOGRAPHY & INDEX

Bibliography

[1] T. Lelièvre, G. Stoltz, and M. Rousset. *Free Energy Computations: A Mathematical Perspective*. Imperial College Press, 2010 (cit. on p. 5).

[2] Paul (de Chemnitz) Auteur du texte Spindler, Georg (1857-1950) Auteur du texte Meyer, and Jacob Hendrik Auteur du texte Meerburg. *Annalen der Physik*. Issue. 1927 (cit. on p. 5).

[3] E. Schrödinger. "An Undulatory Theory of the Mechanics of Atoms and Molecules". In: *Phys. Rev.* 28.6 (1926), pp. 1049–1070. DOI: 10.1103/PhysRev.28.1049 (cit. on p. 6).

[4] Thomas D. Kühne et al. "Efficient and Accurate Car-Parrinello-like Approach to Born-Oppenheimer Molecular Dynamics". In: *Phys. Rev. Lett.* 98.6 (2007), p. 066401. DOI: 10.1103/PhysRevLett.98.066401 (cit. on p. 6).

[5] Maxwell, J.C. "Illustrations of the dynamical theory of gases. Part I. On the motions and collisions of perfectly elastic spheres". en. In: *Philosophical Magazine*. Vol. 19. 4th. Taylor & Francis., 1860, pp. 19–32 (cit. on p. 9).

[6] Maxwell, J.C. "Illustrations of the dynamical theory of gases. Part II. On the process of diffusion of two or more kinds of moving particles among one another". en. In: *Philosophical Magazine*. Vol. 20. 4th. Google-Books-ID: DIc7AQAAMAAJ. Taylor & Francis., 1860, pp. 21–37 (cit. on p. 9).

[7] R. Clausius. "XVI. On a mechanical theorem applicable to heat". In: *Philosophical Magazine Series 4* 40.265 (1870), pp. 122–127. DOI: 10.1080/14786447008640370 (cit. on p. 10).

[8] Sandeep Patel and Charles L. Brooks. "CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations". en. In: *J. Comput. Chem.* 25.1 (2004), pp. 1–16. DOI: 10.1002/jcc.10355 (cit. on pp. 11, 19).

[9] Sandeep Patel, Alexander D. Mackerell, and Charles L. Brooks. "CHARMM fluctuating charge force field for proteins: II Protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model". en. In: *J. Comput. Chem.* 25.12 (2004), pp. 1504–1514. DOI: 10.1002/jcc.20077 (cit. on p. 11).

[10] Guillaume Lamoureux and Benoît Roux. "Modeling induced polarization with classical Drude oscillators: Theory and molecular dynamics simulation algorithm". In: *The Journal of Chemical Physics* 119.6 (2003), pp. 3025–3039. DOI: 10.1063/1.1589749 (cit. on pp. 11, 19).

[11] Guillaume Lamoureux et al. "A polarizable model of water for molecular dynamics simulations of biomolecules". In: *Chemical Physics Letters* 418.1–3 (2006), pp. 245–249. DOI: 10.1016/j.cplett.2005.10.135 (cit. on p. 11).

[12] Yue Shi et al. "Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins". In: *J. Chem. Theory Comput.* 9.9 (2013), pp. 4046–4063. DOI: 10.1021/ct4003702 (cit. on p. 12).

[13] Nuria Plattner and Markus Meuwly. "The Role of Higher CO-Multipole Moments in Understanding the Dynamics of Photodissociated Carbonmonoxide in Myoglobin". In: *Biophys. J.* 94.7 (2008), pp. 2505–2515 (cit. on p. 12).

[14] Nuria Plattner and Markus Meuwly. "Higher order multipole moments for molecular dynamics simulations". In: *J. Mol. Model.* 15.6 (2009), pp. 687–694 (cit. on p. 12).

[15] C. Kramer, P. Gedeck, and M. Meuwly. "Atomic Multipoles: Electrostatic Potential Fit, Local Reference Axis Systems, and Conformational Dependence". In: *J. Comput. Chem.* 33 (20 2012), pp. 1673–1688 (cit. on pp. 12, 134).

[16] C. Kramer et al. "Deriving Static Atomic Multipoles from the Electrostatic Potential". In: *J. Comput. Inf. Model.* 53.12 (2013), pp. 3410–3417 (cit. on pp. 12, 134).

[17] Tristan Bereau et al. "Scoring Multipole Electrostatics in Condensed-Phase Atomistic Simulations". In: *J. Phys. Chem. B.* 117.18 (2013), pp. 5460–5471 (cit. on pp. 12, 134).

[18] Tristan Bereau, Christian Kramer, and Markus Meuwly. "Leveraging Symmetries of Static Atomic Multipole Electrostatics in Molecular Dynamics Simulations". In: *J. Chem. Theory Comput.* 9.12 (2013), pp. 5450–5459 (cit. on pp. 12, 134).

[19] Christopher M. Baker. "Polarizable force fields for molecular dynamics simulations of biomolecules". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 5.2 (2015), pp. 241–254. DOI: 10.1002/wcms.1215 (cit. on p. 12).

[20] J. E. Jones. "On the Determination of Molecular Fields. II. From the Equation of State of a Gas". In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 106.738 (1924), pp. 463–477. DOI: 10.1098/rspa.1924.0082 (cit. on p. 12).

[21] H. A. Lorentz. "Ueber die Anwendung des Satzes vom Virial in der kinetischen Theorie der Gase". In: *Annalen der Physik* 248.1 (1881), pp. 127–136. DOI: 10.1002/andp.18812480110 (cit. on p. 13).

[22] R. Eisenschitz and F. London. "Über das Verhältnis der van der Waalsschen Kräfte zu den homöopolaren Bindungskräften". de. In: *Zeitschrift für Physik* 60.7-8 (1929), pp. 491–527. DOI: 10.1007/BF01341258 (cit. on p. 13).

[23] F. London. "The general theory of molecular forces". en. In: *Transactions of the Faraday Society* 33.0 (1937), 8b–26. DOI: 10.1039/TF937330008B (cit. on p. 13).

[24] Farid F. Abraham and Y. Singh. "The structure of a hard-sphere fluid in contact with a soft repulsive wall". In: *The Journal of Chemical Physics* 67.5 (1977), pp. 2384–2385. DOI: http://dx.doi.org/10.1063/1.435080 (cit. on p. 13).

[25] Gustav Mie. "Zur kinetischen Theorie der einatomigen Körper". In: *Annalen der Physik* 316.8 (1903), pp. 657–697. DOI: 10.1002/andp.19033160802 (cit. on p. 13).

[26] Loup Verlet. "Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules". In: *Physical Review* 159.1 (1967), pp. 98–103. DOI: 10.1103/PhysRev.159.98 (cit. on pp. 14, 30).

[27] Peter J. Steinbach and Bernard R. Brooks. "New spherical-cutoff methods for long-range forces in macromolecular simulation". en. In: *Journal of Computational Chemistry* 15.7 (1994), pp. 667–683. DOI: 10.1002/jcc.540150702 (cit. on p. 14).

[28] Tom Darden, Darrin York, and Lee Pedersen. "Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems". In: *The Journal of Chemical Physics* 98.12 (1993), pp. 10089–10092. DOI: 10.1063/1.464397 (cit. on pp. 14, 157).

[29] Michele Di Pierro, Ron Elber, and Benedict Leimkuhler. "A Stochastic Algorithm for the Isobaric–Isothermal Ensemble with Ewald Summations for All Long Range Forces". In: *Journal of Chemical Theory and Computation* 11.12 (2015), pp. 5624–5637. DOI: 10.1021/acs.jctc.5b00648 (cit. on p. 14).

[30] M. Frigo and S. G. Johnson. "FFTW: an adaptive software architecture for the FFT". In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 1998*. Vol. 3. 1998, 1381–1384 vol.3. DOI: 10.1109/ICASSP.1998.681704 (cit. on p. 15).

[31] M. Frigo and S. G. Johnson. "The Design and Implementation of FFTW3". In: *Proceedings of the IEEE* 93.2 (2005), pp. 216–231. DOI: 10.1109/JPROC.2004.840301 (cit. on p. 15).

[32] Philip M. Morse. "Diatomic Molecules According to the Wave Mechanics. II. Vibrational Levels". In: *Phys. Rev.* 34.1 (1929), pp. 57–64. DOI: 10.1103/PhysRev.34.57 (cit. on p. 16).

[33] Alexander D. Mackerell, Michael Feig, and Charles L. Brooks. "Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations". In: *Journal of Computational Chemistry* 25.11 (2004), pp. 1400–1415. DOI: 10.1002/jcc.20065 (cit. on p. 17).

[34] Matthias Buck et al. "Importance of the CMAP Correction to the CHARMM22 Protein Force Field: Dynamics of Hen Lysozyme". In: *Biophysical Journal* 90.4 (2006), pp. L36–L38. DOI: 10.1529/biophysj.105.078154 (cit. on p. 17).

[35] K. Vanommeslaeghe and A. D. MacKerell Jr. "Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing". In: *J. Comput. Inf. Model.* 52.12 (2012), pp. 3144–3154 (cit. on p. 19).

[36] K. Vanommeslaeghe, E. Prabhu Raman, and A. D. MacKerell Jr. "Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges". In: *J. Comput. Inf. Model.* 52.12 (2012), pp. 3155–3168 (cit. on p. 19).

[37] Tibor Nagy, Juvenal Yosa Reyes, and Markus Meuwly. "Multisurface Adiabatic Reactive Molecular Dynamics". In: *J. Chem. Theory Comput.* 10.4 (2014), pp. 1366–1375. DOI: 10.1021/ct400953f (cit. on p. 19).

[38] R. H. Swendsen and J-S. Wang. "Replica Monte Carlo Simulation of Spin-Glasses". In: *Phys. Rev. Lett.* 57.21 (Nov. 1986), p. 2607. DOI: 10.1103/PhysRevLett.57.2607 (cit. on pp. 21, 35–36, 40).

[39] K. Hukushima and K. Nemoto. "Exchange Monte Carlo Method and Application to Spin Glass Simulations". In: *J. Phys. Soc. Jpn.* 65.6 (1996), pp. 1604–1608. DOI: 10.1143/JPSJ.65.1604 (cit. on pp. 21, 35).

[40] D. A. Kofke. "On the acceptance probability of replica-exchange Monte Carlo trials". en. In: *J. Chem. Phys.* 117.15 (2002), pp. 6911–6914. DOI: 10.1063/1.1507776 (cit. on pp. 21, 35).

[41] D. J. Earl and M. W. Deem. "Parallel tempering: Theory, applications, and new perspectives". In: *PCCP* 7.23 (2005), pp. 3910–3916. DOI: 10.1039/b509983h (cit. on pp. 21, 35).

[42] R W Zwanzig. "HIGH-TEMPERATURE EQUATION OF STATE BY A PERTURBATION METHOD .1. NONPOLAR GASES". In: *J. Chem. Phys.* 22 (1954), pp. 1420–1426 (cit. on p. 21).

[43] J G Kirkwood. "Statistical mechanics of fluid mixtures". In: *J. Chem. Phys.* 3 (1935), pp. 300–313 (cit. on p. 21).

[44] G. M. Torrie and J. P. Valleau. "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling". In: *J. Chem. Phys.* 23.2 (1977), pp. 187–199. DOI: 10.1016/0021-9991(77)90121-8 (cit. on pp. 21, 36).

[45] Grossfield, Alan. *WHAM: the weighted histogram analysis method, version 2.0.9*. Accessed on 8th June 2016 (cit. on p. 21).

[46] Alessandro Laio and Michele Parrinello. "Escaping free-energy minima." In: *Proc. Natl. Acad. Sci.* 99.20 (Oct. 2002), pp. 12562–12566. DOI: 10.1073/pnas.202427399 (cit. on pp. 21, 36).

[47] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. "Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method". In: *Phys. Rev. Lett.* 100.2 (2008), pp. 20603–20607. DOI: 10.1103/PhysRevLett.100.020603 (cit. on pp. 21, 36, 69).

[48] M Bonomi, A Barducci, and M Parrinello. "Reconstructing the equilibrium Boltzmann distribution from well-tempered metadynamics." In: *J. Comput. Chem.* 30.11 (2009), pp. 1615–1621. DOI: 10.1002/jcc.21305 (cit. on pp. 21, 36, 69).

[49] Davide Branduardi, Giovanni Bussi, and Michele Parrinello. "Metadynamics with Adaptive Gaussians". In: *J. Chem. Theory Comput.* 8.7 (2012), pp. 2247–2254. DOI: 10.1021/ct3002464 (cit. on pp. 21, 36, 69).

[50] Gungor Ozer et al. "Multiple branched adaptive steered molecular dynamics". In: *J. Chem. Phys.* 141.6 (2014), p. 064101. DOI: 10.1063/1.4891807 (cit. on p. 21).

[51] Jeffrey Comer et al. "Multiple-Replica Strategies for Free-Energy Calculations in NAMD: Multiple-Walker Adaptive Biasing Force and Walker Selection Rules". In: *J. Chem. Theory Comput.* 10.12 (2014), pp. 5276–5285. DOI: 10.1021/ct500874p (cit. on p. 21).

[52] Fugao Wang and D. P. Landau. "Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States". In: *Phys. Rev. Lett.* 86.10 (2001), pp. 2050–2053. DOI: 10.1103/PhysRevLett.86.2050 (cit. on p. 21).

[53] J. D. Doll et al. "A spatial averaging approach to rare-event sampling". In: *J. Chem. Phys.* 131.10 (Sept. 2009). DOI: 10.1063/1.3220629 (cit. on pp. 25, 36–37, 49).

[54] N. Plattner, J. D. Doll, and M. Meuwly. "Spatial averaging for small molecule diffusion in condensed phase environments". In: *J. Chem. Phys.* 133.4 (July 2010). DOI: 10.1063/1.3458639 (cit. on pp. 25, 37, 49).

[55] N. Plattner et al. "An infinite swapping approach to the rare-event sampling problem." en. In: *J. Chem. Phys.* 135.13 (2011), p. 134111. DOI: 10.1063/1.3643325 (cit. on p. 25).

[56] P. Dupuis et al. "On the Infinite Swapping Limit for Parallel Tempering". en. In: *Multiscale Model. Simul.* 10.3 (2012), pp. 986–1022. DOI: 10.1137/110853145 (cit. on p. 25).

[57] J. D. Doll et al. "Rare-event sampling: occupation-based performance measures for parallel tempering and infinite swapping Monte Carlo methods." In: *J. Chem. Phys.* 137.20 (2012), p. 204112. DOI: 10.1063/1.4765060 (cit. on p. 25).

[58] J. D. Doll et al. "Rare-event sampling: Occupation-based performance measures for parallel tempering and infinite swapping Monte Carlo methods". In: *The Journal of Chemical Physics* 137.20, 204112 (2012). DOI: http://dx.doi.org/10.1063/1.4765060 (cit. on pp. 25, 76–77).

[59] N. Plattner, J. D. Doll, and M. Meuwly. "Overcoming the Rare Event Sampling Problem in Biological Systems with Infinite Swapping". In: *J. Chem. Theory Comput.* 9.9 (2013), pp. 4215–4224. DOI: 10.1021/ct400355g (cit. on pp. 25, 69).

[60] "IEEE Standard for Floating-Point Arithmetic". In: *IEEE Std 754-2008* (2008), pp. 1–70. DOI: 10.1109/IEEESTD.2008.4610935 (cit. on p. 29).

[61] S. Chapman. "Fredrik Carl Mulertz Størmer. 1874-1957". en. In: *Biographical Memoirs of Fellows of the Royal Society* 4.0 (1958), pp. 257–279. DOI: 10.1098/rsbm.1958.0021 (cit. on p. 30).

[62] William C. Swope et al. "A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters". In: *The Journal of Chemical Physics* 76.1 (1982), pp. 637–649. DOI: 10.1063/1.442716 (cit. on p. 30).

[63] Hans C. Andersen. "Molecular dynamics simulations at constant pressure and/or temperature". In: *The Journal of Chemical Physics* 72.4 (1980), pp. 2384–2393. DOI: 10.1063/1.439486 (cit. on p. 30).

[64] H. J. C. Berendsen et al. "Molecular dynamics with coupling to an external bath". In: *The Journal of Chemical Physics* 81.8 (1984), pp. 3684–3690. DOI: 10.1063/1.448118 (cit. on p. 30).

[65] Shuichi Nosé. "A unified formulation of the constant temperature molecular dynamics methods". In: *The Journal of Chemical Physics* 81.1 (1984), pp. 511–519. DOI: 10.1063/1.447334 (cit. on p. 31).

[66] William G. Hoover. "Canonical dynamics: Equilibrium phase-space distributions". In: *Phys. Rev. A* 31.3 (1985), pp. 1695–1697. DOI: 10.1103/PhysRevA.31.1695 (cit. on p. 31).

[67] Giovanni Bussi, Tatyana Zykova-Timan, and Michele Parrinello. "Isothermal-isobaric molecular dynamics using stochastic velocity rescaling". In: *The Journal of Chemical Physics* 130.7 (2009). arXiv: 0901.0779, p. 074101. DOI: 10.1063/1.3073889 (cit. on p. 31).

[68] M. Parrinello and A. Rahman. "Crystal Structure and Pair Potentials: A Molecular-Dynamics Study". In: *Phys. Rev. Lett.* 45.14 (1980), pp. 1196–1199. DOI: 10.1103/PhysRevLett.45.1196 (cit. on pp. 31, 158).

[69] M. Parrinello and A. Rahman. "Polymorphic transitions in single crystals: A new molecular dynamics method". In: *Journal of Applied Physics* 52.12 (1981). DOI: 10.1063/1.328693 (cit. on pp. 31, 158).

[70] M. Parrinello and A. Rahman. "Strain fluctuations and elastic constants". In: *The Journal of Chemical Physics* 76.5 (1982), pp. 2662–2666. DOI: 10.1063/1.443248 (cit. on pp. 31, 158).

[71] N. Metropolis and S. Ulam. "The Monte Carlo Method". In: *J. Am. Stat. Assoc.* 44.247 (1949), pp. 335–341. DOI: 10.2307/2280232 (cit. on p. 31).

[72] N. Metropolis et al. "Equation of State Calculations by Fast Computing Machines". In: *J. Chem. Phys.* 21.6 (1953), p. 1087. DOI: 10.1063/1.1699114 (cit. on p. 32).

[73] W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1 (Apr. 1970), pp. 97–109. DOI: 10.1093/biomet/57.1.97 (cit. on pp. 32–33).

[74] J. Hu, A. Ma, and A. R. Dinner. "Monte Carlo simulations of biomolecules: The MC module in CHARMM". In: *J. Comput. Chem.* 27.2 (Jan. 2006), pp. 203–216. DOI: 10.1002/jcc.20327 (cit. on p. 34).

[75] C. L. Brooks., A. D. Mackerell, and M. Karplus. "CHARMM: The Biomolecular Simulation Program". English. In: *J. Comput. Chem.* 30.10, Sp. Iss. SI (JUL 30 2009), 1545–1614. DOI: {10.1002/jcc.21287} (cit. on p. 34).

[76] Jakob P. Ulmschneider and William L. Jorgensen. "Polypeptide folding using Monte Carlo sampling, concerted rotation, and continuum solvation". eng. In: *J. Am. Chem. Soc.* 126.6 (2004), pp. 1849–1857. DOI: 10.1021/ja0378862 (cit. on p. 34).

[77] Jakob P. Ulmschneider, Martin B. Ulmschneider, and Alfredo Di Nola. "Monte Carlo vs molecular dynamics for all-atom polypeptide folding simulations". eng. In: *J Phys Chem B* 110.33 (2006), pp. 16733–16742. DOI: 10.1021/jp061619b (cit. on p. 34).

[78] G. M. Torrie and J. P. Valleau. "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling". In: *J. Chem. Phys.* 23.2 (1977), pp. 187–199. DOI: 10.1016/0021-9991(77)90121-8 (cit. on p. 35).

[79] Alessandro Laio and Michele Parrinello. "Escaping free-energy minima." In: *Proc. Natl. Acad. Sci.* 99.20 (2002), pp. 12562–12566. DOI: 10.1073/pnas.202427399 (cit. on p. 35).

[80] A. F. Voter. "A Monte Carlo method for determining free-energy differences and transition state theory rate constants". en. In: *J. Chem. Phys.* 82.4 (Feb. 1985), p. 1890. DOI: 10.1063/1.448373 (cit. on pp. 35–36).

[81] M. R. Betancourt. "Optimization of Monte Carlo trial moves for protein simulations." en. In: *J. Chem. Phys.* 134.1 (Jan. 2011), p. 14104. DOI: 10.1063/1.3515960 (cit. on pp. 35–36).

[82] D. J. Earl and M. W. Deem. "Parallel tempering: Theory, applications, and new perspectives". In: *PCCP* 7.23 (2005), p. 3910. DOI: 10.1039/b509983h (cit. on pp. 36, 40–41).

[83] D. A. Kofke. "On the acceptance probability of replica-exchange Monte Carlo trials". en. In: *J. Chem. Phys.* 117.15 (Oct. 2002), p. 6911. DOI: 10.1063/1.1507776 (cit. on p. 36).

[84] C. Predescu, M. Predescu, and C. V. Ciobanu. "The incomplete beta function law for parallel tempering sampling of classical canonical systems." en. In: *J. Chem. Phys.* 120.9 (Mar. 2004), pp. 4119–4128. DOI: 10.1063/1.1644093 (cit. on p. 36).

[85] H. G. Katzgraber et al. "Feedback-optimized parallel tempering Monte Carlo". In: *J. Stat. Mech.: Theory Exp.* 2006.03 (2006), p. 3018 (cit. on p. 36).

[86] D. Sabo et al. "A constant entropy increase model for the selection of parallel tempering ensembles." en. In: *J. Chem. Phys.* 128.17 (May 2008), p. 174109. DOI: 10.1063/1.2907846 (cit. on p. 36).

[87] Helmut Grubmüller. "Predicting slow structural transitions in macromolecular systems: Conformational flooding". In: *Phys. Rev. E* 52.3 (Sept. 1995), pp. 2893–2906. DOI: 10.1103/PhysRevE.52.2893 (cit. on p. 36).

[88] E. Matthias Müller, Armin de Meijere, and Helmut Grubmüller. "Predicting unimolecular chemical reactions: Chemical flooding". In: *J. Chem. Phys.* 116.3 (Jan. 2002), p. 897. DOI: 10.1063/1.1427722 (cit. on p. 36).

[89] C. Tsallis. "Possible generalization of Boltzmann-Gibbs statistics". In: *J. Stat. Phys.* 52.1-2 (July 1988), pp. 479–487. DOI: 10.1007/BF01016429 (cit. on p. 36).

[90] "Deterministic generation of the Boltzmann-Gibbs distribution and the free energy calculation from the Tsallis distribution". In: *Chem. Phys. Lett.* 382.3-4 (Dec. 2003), pp. 367–373. DOI: 10.1016/j.cplett.2003.10.077 (cit. on p. 36).

[91] J. G. Kim, Y. Fukunishi, and H. Nakamura. "Dynamical origin of enhanced conformational searches of Tsallis statistics sampling." In: *J. Chem. Phys.* 121.3 (July 2004), pp. 1626–1635. DOI: 10.1063/1.1763841 (cit. on p. 36).

[92] Z. Li and H. A. Scheraga. "Monte Carlo-minimization approach to the multiple-minima problem in protein folding." In: *Proc. Natl. Acad. Sci.* 84.19 (Oct. 1987), pp. 6611–6615 (cit. on p. 36).

[93] L. Piela, J. Kostrowicki, and H. A. Scheraga. "On the multiple-minima problem in the conformational analysis of molecules: deformation of the potential energy hypersurface by the diffusion equation method". In: *J. Phys. Chem.* 93.8 (1989), pp. 3339–3346. DOI: 10.1021/j100345a090 (cit. on p. 36).

[94] J. Ma and J. E. Straub. "Simulated annealing using the classical density distribution". In: *J. Chem. Phys.* 101.1 (1994), p. 533. DOI: 10.1063/1.468163 (cit. on p. 36).

[95] R. V. Pappu, R. K. Hart, and J. W. Ponder. "Analysis and Application of Potential Energy Smoothing and Search Methods for Global Optimization". In: *J. Phys. Chem. B* 102.48 (1998), pp. 9725–9742. DOI: 10.1021/jp982255t (cit. on p. 36).

[96] Shinichi Banba, Zhuyan Guo, and Charles. "Efficient sampling of ligand orientations and conformations in free energy calculations using the lambda-dynamics method". In: *Journal of Physical Chemistry B* 104.29 (2000) (cit. on p. 36).

[97] Ryan {Bitetti-Putzer}, Wei Yang, and Martin Karplus. "Generalized ensembles serve to improve the convergence of free energy simulations". In: *Chemical Physics Letters* 377 (2003) (cit. on p. 36).

[98] Christopher Woods, Jonathan Essex, and Michael King. "Enhanced Configurational Sampling in Binding Free Energy Calculations". In: *Journal of Physical Chemistry B* 107 (2003) (cit. on p. 36).

[99] Yuko Okamoto. "Generalized-ensemble algorithms: Enhanced sampling techniques for Monte Carlo and molecular dynamics simulations". In: *Journal of Molecular Graphics and Modelling* 22 (2004) (cit. on p. 36).

[100] Benoît Roux and José {Faraldo-Gómez}. "Characterization of conformational equilibria through Hamiltonian and temperature replica-exchange simulations: Assessing entropic and environmental effects". In: *Journal of Computational Chemistry* 28.10 (2007). DOI: 10.1002/jcc.20652 (cit. on p. 36).

[101] Jozef Hritz and Chris Oostenbrink. "Hamiltonian replica exchange molecular dynamics using soft-core interactions". In: *Journal of Chemical Physics* 128.14 (2008), p. 144121. DOI: 10.1063/1.2888998 (cit. on p. 36).

[102] Giovanni Bussi. "Hamiltonian replica-exchange in GROMACS: a flexible implementation". In: *Molecular Physics* 112.3-4 (2014). arXiv: 1307.5144, pp. 379–384. DOI: 10.1080/00268976.2013.824126 (cit. on p. 36).

[103] S. Juneja and P. Shahabuddin. "Chapter 11 Rare-Event Simulation Techniques: An Introduction and Recent Advances". In: *Handbooks in Operations Research and Management Science*. Ed. by Shane G. Henderson and Barry L. Nelson. Vol. 13. Simulation. Elsevier, 2006, pp. 291–350 (cit. on p. 36).

[104] Werner Sandmann. "Rare Event Simulation Methodologies in Systems Biology". en. In: *Rare Event Simulation using Monte Carlo Methods*. Ed. by Gerardo Rubino and Bruno Tuffin. John Wiley & Sons, Ltd, 2009, pp. 243–265 (cit. on p. 36).

[105] Jérôme Morio et al. "A survey of rare event simulation methods for static input–output models". In: *Simulation Modelling Practice and Theory* 49 (2014), pp. 287–304. DOI: 10.1016/j.simpat.2014.10.007 (cit. on p. 36).

[106] H. Kahn and A. W. Marshall. "Methods of Reducing Sample Size in Monte Carlo Computations". In: *Operations Research* 1.5 (1953), pp. 263–278. DOI: 10.1287/opre.1.5.263 (cit. on p. 38).

[107] N. Plattner et al. "An infinite swapping approach to the rare-event sampling problem." en. In: *J. Chem. Phys.* 135.13 (Oct. 2011), p. 134111. DOI: 10.1063/1.3643325 (cit. on pp. 40–42).

[108] J. D. Doll et al. "Rare-event sampling: occupation-based performance measures for parallel tempering and infinite swapping Monte Carlo methods." In: *J. Chem. Phys.* 137.20 (Nov. 2012), p. 204112. DOI: 10.1063/1.4765060 (cit. on pp. 40, 42).

[109] P. Dupuis et al. "On the Infinite Swapping Limit for Parallel Tempering". en. In: *Multiscale Model. Simul.* 10.3 (Sept. 2012), pp. 986–1022. DOI: 10.1137/110853145 (cit. on pp. 40–42).

[110] N. Plattner, J. D. Doll, and M. Meuwly. "Overcoming the Rare Event Sampling Problem in Biological Systems with Infinite Swapping". In: *J. Chem. Theory Comput.* 9.9 (Sept. 2013), pp. 4215–4224. DOI: 10.1021/ct400355g (cit. on pp. 40, 42, 69).

[111] K. Hukushima and K. Nemoto. "Exchange Monte Carlo Method and Application to Spin Glass Simulations". In: *J. Phys. Soc. Jpn.* 65.6 (June 1996), pp. 1604–1608. DOI: 10.1143/JPSJ.65.1604 (cit. on p. 40).

[112] F. Hédin et al. "Spatial Averaging: Sampling Enhancement for Exploring Configurational Space of Atomic Clusters and Biomolecules". In: *J. Chem. Theory Comput.* 10.10 (2014), pp. 4284–4296. DOI: 10.1021/ct500529w (cit. on pp. 49–50, 70–72).

[113] D. J. Wales and J. P. K. Doye. "Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms". In: *J. Phys. Chem. A* 101.28 (July 1997), pp. 5111–5116. DOI: 10.1021/jp970984n (cit. on p. 50).

[114] D. J. Wales et al. *The Cambridge Cluster Database: http://www-wales.ch.cam.ac.uk/CCD.html (accessed on Aug 13, 2014)* (cit. on p. 50).

[115] D. J. Tobias and C. L. Brooks. "Conformational equilibrium in the alanine dipeptide in the gas phase and aqueous solution: a comparison of theoretical results". In: *J. Phys. Chem.* 96.9 (Apr. 1992), pp. 3864–3870. DOI: 10.1021/j100188a054 (cit. on p. 69).

[116] J. Apostolakis, P. Ferrara, and A. Caflisch. "Calculation of conformational transitions and barriers in solvated systems: Application to the alanine dipeptide in water". en. In: *J. Chem. Phys.* 110.4 (1999), pp. 2099–2108. DOI: 10.1063/1.477819 (cit. on pp. 69, 73–74).

[117] D. S. Chekmarev, T. Ishida, and R. M. Levy. "Long-Time Conformational Transitions of Alanine Dipeptide in Aqueous Solution: Continuous and Discrete-State Kinetic Models". In: *J. Phys. Chem. B* 108.50 (Dec. 2004), pp. 19487–19495. DOI: 10.1021/jp048540w (cit. on p. 69).

[118] A. Ma and A. R. Dinner. "Automatic Method for Identifying Reaction Coordinates in Complex Systems". In: *J. Phys. Chem. B* 109.14 (Apr. 2005), pp. 6769–6779. DOI: 10.1021/jp045546c (cit. on p. 69).

[119] D. Gfeller et al. "Complex network analysis of free-energy landscapes." In: *Proc. Natl. Acad. Sci.* 104.6 (2007), pp. 1817–1822. DOI: 10.1073/pnas.0608099104 (cit. on p. 69).

[120] L. Yang and Q. G. Yi. "A selective integrated tempering method." In: *J. Chem. Phys.* 131.21 (2009), p. 214109. DOI: 10.1063/1.3266563 (cit. on p. 69).

[121] Weina Du and Peter G. Bolhuis. "Equilibrium Kinetic Network of the Villin Headpiece in Implicit Solvent". In: *Biophys. J.* 108.2 (2015), pp. 368–378. DOI: 10.1016/j.bpj.2014.11.3476 (cit. on p. 69).

[122] F. F. García-Prieto et al. "Study on the conformational equilibrium of the alanine dipeptide in water solution by using the averaged solvent electrostatic potential from molecular dynamics

methodology." In: *J. Chem. Phys.* 135.19 (2011), p. 194502. DOI: 10.1063/1.3658857 (cit. on p. 69).

[123] I. H. Lee. "Free-energy profile along an isomerization pathway: Conformational isomerization in alanine dipeptide". In: *J. Korean Phys. Soc.* 62.3 (2013), pp. 384–392. DOI: 10.3938/jkps.62.384 (cit. on p. 69).

[124] T. Morishita et al. "On-the-fly reconstruction of free-energy profiles using logarithmic mean-force dynamics." In: *J. Comput. Chem.* 34.16 (2013), pp. 1375–1384. DOI: 10.1002/jcc.23267 (cit. on p. 69).

[125] H. X. Kondo and M. Taiji. "Enhanced exchange algorithm without detailed balance condition for replica exchange method." In: *J. Chem. Phys.* 138.24 (2013), p. 244113. DOI: 10.1063/1.4811711 (cit. on p. 69).

[126] T. Lankau and C.-H. Yu. "A constrained reduced-dimensionality search algorithm to follow chemical reactions on potential energy surfaces." In: *J. Chem. Phys.* 138.21 (2013), p. 214102. DOI: 10.1063/1.4807743 (cit. on p. 69).

[127] Davide Branduardi, Francesco Luigi Gervasio, and Michele Parrinello. "From A to B in free energy space." In: *J. Chem. Phys.* 126.5 (2007), p. 54103. DOI: 10.1063/1.2432340 (cit. on p. 69).

[128] Birgit Strodel and David J. Wales. "Free energy surfaces from an extended harmonic superposition approach and kinetics for alanine dipeptide". In: *Chem. Phys. Lett.* 466.4-6 (2008), pp. 105–115. DOI: 10.1016/j.cplett.2008.10.085 (cit. on p. 69).

[129] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. "Stereochemistry of polypeptide chain configurations". In: *J. Mol. Biol.* 7.1 (1963), pp. 95–99. DOI: 10.1016/S0022-2836(63)80023-6 (cit. on p. 69).

[130] M. Schaefer and M. Karplus. "A Comprehensive Analytical Treatment of Continuum Electrostatics". In: *J. Phys. Chem.* 100.5 (1996), pp. 1578–1599. DOI: 10.1021/jp9521621 (cit. on p. 69).

[131] M. Schaefer, C. Bartels, and M. Karplus. "Solution conformations and thermodynamics of structured peptides: molecular dynamics simulation with an implicit solvation model". In: *J. Mol. Biol.* 284.3 (1998), pp. 835–848. DOI: 10.1006/jmbi.1998.2172 (cit. on p. 69).

[132] B. N. Dominy and C. L. Brooks III. "Identifying native-like protein structures using physics-based potentials". In: *J. Comput. Chem.* 23.1 (2002), pp. 147–160. DOI: 10.1002/jcc.10018 (cit. on p. 69).

[133] M. Schaefer et al. "Effective atom volumes for implicit solvent models: comparison between Voronoi volumes and minimum fluctuation volumes". In: *J. Comput. Chem.* 22.15 (2001), pp. 1857–1879. DOI: 10.1002/jcc.1137 (cit. on p. 69).

[134] N. Calimet, M. Schaefer, and T. Simonson. "Protein molecular dynamics with the generalized Born/ACE solvent model." In: *Proteins* 45.2 (2001), pp. 144–158 (cit. on p. 69).

[135] B. D. Bursulaya and C. L. Brooks III. "Comparative Study of the Folding Free Energy Landscape of a Three-Stranded beta-Sheet Protein with Explicit and Implicit Solvent Models". In: *The Journal of Physical Chemistry B* 104.51 (2000), pp. 12378–12383. DOI: 10.1021/jp0027602 (cit. on p. 69).

[136] Y. Pak, S. Jang, and S. Shin. "Prediction of helical peptide folding in an implicit water by a new molecular dynamics scheme with generalized effective potential". In: *J. Chem. Phys.* 116.15 (2002), pp. 6831–6835. DOI: 10.1063/1.1464120 (cit. on p. 69).

[137] M. Feig, W. Im, and C. L. Brooks III. "Implicit solvation based on generalized Born theory in different dielectric environments". In: *J. Chem. Phys.* 120.2 (2004), pp. 903–911. DOI: 10.1063/1.1631258 (cit. on p. 69).

[138] Hyunbum Jang and Thomas B. Woolf. "Multiple pathways in conformational transitions of the alanine dipeptide: An application of dynamic importance sampling". In: *J. Comput. Chem.* 27.11 (2006), pp. 1136–1141. DOI: 10.1002/jcc.20444 (cit. on pp. 73–74).

[139] E.W. Dijkstra. "A note on two problems in connexion with graphs". English. In: *Numer. Math.* 1.1 (1959), pp. 269–271. DOI: 10.1007/BF01386390 (cit. on p. 74).

[140] Simon Trebst, Matthias Troyer, and Ulrich H. E. Hansmann. "Optimized parallel tempering simulations of proteins". In: *The Journal of Chemical Physics* 124.17, 174903 (2006). DOI: http://dx.doi.org/10.1063/1.2186639 (cit. on p. 76).

[141] Helmut G Katzgraber et al. "Feedback-optimized parallel tempering Monte Carlo". In: *Journal of Statistical Mechanics: Theory and Experiment* 2006.03 (2006), P03018 (cit. on p. 76).

[142] Maksim Kouza and Ulrich H. E. Hansmann. "Velocity scaling for optimizing replica exchange molecular dynamics". In: *The Journal of Chemical Physics* 134.4, 044124 (2011). DOI: http://dx.doi.org/10.1063/1.3533236 (cit. on p. 76).

[143] Norbert Wiener. "Generalized harmonic analysis". In: *Acta Mathematica* 55.1 (1930), pp. 117–258. DOI: 10.1007/BF02546511 (cit. on p. 78).

[144] A. Khintchine. "Korrelationstheorie der stationären stochastischen Prozesse". In: *Mathematische Annalen* 109.1 (1934), pp. 604–615. DOI: 10.1007/BF01449156 (cit. on p. 78).

[145] F. Hédin, K. El Hage, and M. Meuwly. "A Toolkit to Fit Nonbonded Parameters from and for Condensed Phase Simulations". In: *J. Chem. Inf. Model.* 56.8 (2016), pp. 1479–1489. DOI: 10.1021/acs.jcim.6b00280 (cit. on pp. 133, 135).

[146] A.D. MacKerel Jr. et al. "CHARMM: The Energy Function and Its Parameterization with an Overview of the Program". In: ed. by P. v. R. Schleyer et al. Vol. 1. The Encyclopedia of Computational Chemistry. John Wiley & Sons: Chichester, 1998, pp. 271–277 (cit. on p. 157).

[147] Mark James Abraham et al. "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers". In: *SoftwareX* 1–2 (2015), pp. 19–25. DOI: http://dx.doi.org/10.1016/j.softx.2015.06.001 (cit. on p. 157).

[148] J. S. Hub, Marcus B. Kubitzki, and Bert L. de Groot. "Spontaneous Quaternary and Tertiary T-R Transitions of Human Hemoglobin in Molecular Dynamics Simulation". In: *PLoS Comput. Biol.* 6.5 (May 2010), e1000774. DOI: 10.1371/journal.pcbi.1000774 (cit. on p. 157).

[149] K. Y. Olaniyi et al. "Role of the Subunit Interactions in the Conformational Transitions in Adult Human Hemoglobin: An Explicit Solvent Molecular Dynamics Study". In: *J. Phys. Chem. B* 116.36 (2012). PMID: 22838506, pp. 11004–11009. DOI: 10.1021/jp3022908 (cit. on p. 157).

[150] A. P. Willard and D. Chandler. "Instantaneous Liquid Interfaces". In: *J. Phys. Chem. B* 114.5 (2010), pp. 1954–1958. DOI: 10.1021/jp909219k (cit. on pp. 157, 164).

[151] A. P. Willard and D. Chandler. "The molecular structure of the interface between water and a hydrophobic substrate is liquid-vapor like". In: *J. Chem. Phys.* 141.18 (2014), p. 18C519. DOI: 10.1063/1.4897249 (cit. on pp. 157, 164).

[152] Berk Hess et al. "LINCS: A linear constraint solver for molecular simulations". en. In: *J. Comput. Chem.* 18.12 (1997), pp. 1463–1472. DOI: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H (cit. on p. 158).

[153] Giovanni Bussi, Davide Donadio, and Michele Parrinello. "Canonical sampling through velocity rescaling". In: *The Journal of Chemical Physics* 126.1 (2007), p. 014101. DOI: 10.1063/1.2408420 (cit. on p. 158).

[154] M. F. Perutz. "Stereochemistry of Cooperative Effects in Haemoglobin: Haem–Haem Interaction and the Problem of Allostery". en. In: *Nature* 228.5273 (1970), pp. 726–734. DOI: 10.1038/228726a0 (cit. on p. 158).

[155] M. M. Silva, P. H. Rogers, and A. Arnone. "A third quaternary structure of human hemoglobin A at 1.7-A resolution." en. In: *J. Biol. Chem.* 267.24 (1992), pp. 17248–17256 (cit. on p. 158).

[156] Chunyan Xu, Dror Tobi, and I. Bahar. "Allosteric Changes in Protein Structure Computed by a Simple Mechanical Model: Hemoglobin TR2 Transition". In: *Journal of Molecular Biology* 333.1 (2003), pp. 153–168. DOI: 10.1016/j.jmb.2003.08.027 (cit. on p. 158).

[157] Eric M. Jones et al. "Differential Control of Heme Reactivity in Alpha and Beta Subunits of Hemoglobin: A Combined Raman Spectroscopic and Computational Study". In: *J. Am. Chem. Soc.* 136.29 (2014), pp. 10325–10339. DOI: 10.1021/ja503328a (cit. on p. 159).

Index

Florent Henri René HÉDIN

Date of birth: 28 Sept. 1988
Web: https://fhedin.com

Experience

Oct. 2011 – Oct. 2016 ***PhD Student*, Chemistry department, University of Basel, Switzerland**
"Development and Application of Accurate Molecular Mechanics Sampling Methods: From Atomic Clusters to Protein Tetramers"
Supervisor: Prof. Markus Meuwly.

Jul. 2010 – Aug. 2010 ***2 months employment*, University of Strasbourg (France), MSM : Modélisation et Simulations Moléculaires**
"Molecular Dynamics and Simulations of Uranyl Complexes in ionic liquids, with the AMBER Molecular Dynamics Package."
Supervisor: Prof. Georges Wipff.

Education

Oct. 2011 – Sept. 2016 ***PhD Thesis*, University of Basel (Switzerland)**
"Development and Application of Accurate Molecular Mechanics Sampling Methods: From Atomic Clusters to Protein Tetramers"

Mar. 2011 – Aug. 2011 ***Master Thesis*, University of Basel (Switzerland) and University of Strasbourg (France)**
"Spatial Averaging: a new Monte Carlo approach for sampling rare-event problems."

Sept. 2009 – Sept. 2011 ***Master degree Cursus*, University of Strasbourg (France)**
"Master degree in Chemoinformatics."

Sept. 2006 – Jun. 2009 ***Bachelor degree Cursus*, University of Picardie Jules Verne (France)**
"Bachelor degree in Chemistry."

Publications

Jul. 2016 **Article, *J. Chem. Inf. Model.*, 10.1021/acs.jcim.6b00280**
"A Toolkit to Fit Nonbonded Parameters from and for Condensed Phase Simulations"
Florent Hédin, Krystel El Hage, and Markus Meuwly

Jan. 2015 **Article, *J. Phys. Chem. B*, 10.1021/jp511701z**
"Vibrational Relaxation and Energy Migration of N-methylacetamide in Water: The Role of Nonbonded Interactions."
Pierre-André Cazade, Florent Hédin, Zhen-Hao Xu, and Markus Meuwly

Aug. 2014 **Article, *J. Chem. Theory Comput.*, 10.1021/ct500529w**
"Spatial Averaging: Sampling Enhancement for Exploring Configurational Space of Atomic Clusters and Biomolecules."
Florent Hédin, Nuria Plattner, J. D. Doll, and Markus Meuwly

Participation to conferences: posters, presentations

Aug. 2016 **Poster: Theory and applications of Computational Chemistry (TACC 2016), Seattle, USA**
"Partial Infinite Swapping: Implementation and Application to peptides and proteins in the Gas Phase and in Solution"

Jan. 2016 **Poster: 6^{th} Annual Meeting of the NCCR MUST Engelberg, CH**
"A new toolkit for fitting forcefield parameters used for Permanent Multipoles molecular simulations"

Sept. 2015 **Talk: Swiss Chemical Society Fall Meeting 2015, Lausanne, CH**
"Addressing the Rare Event Sampling problem with the PINS and SA-MC Methods : studying Structure and Dynamics of the Myoglobin protein"

Jan. 2015 **Talk and Poster: 5^{th} Annual Meeting of the NCCR MUST Engelberg, CH**
"A new toolkit for fitting forcefield parameters used for Permanent Multipoles molecular simulations"

Sept. 2014 **Poster: Swiss Chemical Society Fall Meeting 2014, Zürich, CH**
"A new toolkit for fitting forcefield parameters used for Permanent Multipoles molecular simulations"

Sept. 2013 **Talk: Swiss Chemical Society Fall Meeting 2013, Lausanne, CH**
"Spatial averaging : enhancement of the sampling of the configuration space for atomic clusters and biomolecules"

Oct. 2012 **Posters (2): Workshop: Monte Carlo Methods in the Physical and Biological Sciences, organised by Brown University, Providence, Rhode Island, U.S.A.**
"Sampling rare events with spatial averaging: theory and applications" and "Ligand uptake in truncated hemoglobin: a Monte Carlo study"

Sept. 2012 **Poster: Swiss Chemical Society Fall Meeting 2012, Zürich, CH**
"Sampling rare events with spatial averaging: theory and applications"

July 2012 **Poster: Energy Landscape Conference, organised by European Science Foundation, Obergurgl, Austria.**
"Sampling rare events with spatial averaging: theory and applications."