A toolkit to quantify the sampling quality of molecular dynamics trajectories: Studying highly flexible biomolecules

Inaugural-Dissertation

zur Erlangung des Doktorgrades
Dr. rer. nat.

der Fakultät für
Biologie an der

Universität Duisburg-Essen

vorgelegt von

Mike Nemec

aus Offenbach am Main

April 2017

Die der vorliegenden Arbeit zugrunde liegenden Experimente wurden am Zentrum für Medizinische Biotechnologie (ZMB) in der Abteilung für Bioinformatics and Computational Biophysics der Universität Duisburg-Essen durchgeführt.

1. Gutachter : Prof. Dr. Daniel Hoffmann

2. Gutachter : Prof. Dr. Holger Gohlke

Vorsitzender des Prüfungsausschusses: Prof. Dr. Sanchez-Garcia

Tag der mündlichen Prüfung: 16.11.2017

To all the honest people who keep smiling, enjoying their lives and trying to make the world a better place. Stay patient and strong although it is hard, then others will follow your way.

Abstract

In this computational work, we investigate the sampling of molecular dynamics (MD) simulations of the two highly flexible biomolecules: Methionine-enkephalin (Met-Enkephalin) and the third variable loop (V3) of the glycoprotein 120 (gp120) from the human immunodeficiency virus type-1 (HIV-1).

The conformational dynamics of the three-dimensional (3D) protein structures are of central importance for the biomolecular function. A common possibility to obtain these dynamics at atomic resolution are MD simulations. But reaching a converged MD sampling in adequate time is limited by the huge conformational space of flexible systems. Moreover, an automatic sampling validation is still not established as settled protocol in today's MD studies. Furthermore, existing tools aim primarily to investigate single trajectory convergence which is not always practical for flexible molecules. But in fact, a universal assessment is necessary to classify, whether the sampling is sufficient or not. Otherwise the extracted thermodynamic results are completely meaningless.

The aim of this work is to develop a toolkit to quantitatively assess the MD sampling quality for flexible systems. This toolkit is freely available at https://github.com/MikeN12/PySamplingQuality. We use diverse sets of trajectories with different initial conformations along with enhanced sampling techniques such as accelerated MD (aMD) and scaled MD (sMD). These distort the energy landscape to ease conformational transitions. The sampling is assessed by two new quantities, the conformational O_{conf} and density overlap O_{dens}, including also the cluster number N_C and cluster distribution entropy S_C. These new overlap quantities measure the self-consistency of sampling as a necessary condition for complete sampling.

We use Met-Enkephalin as benchmarking system because of its small size but non-trivial dynamics. Our tool reveals that the MD sampling of already such a small molecule converges in a microsecond regime. Furthermore, we can show that aMD is the most efficient algorithm to assess the convergence and also to detect wrong sampling. However, O_{dens} analysis comparing MD with aMD/sMD reveals that we have not completely corrected the bias from enhanced sampling. Therefore, O_{dens} can also be used to compare different methods. On the other hand, V3 demonstrates that much more resources must be spent to achieve convergence compared to those generally invested today. The results highlight the necessity of a multi-trajectory approach to detect incomplete sampling.

Altogether, we are able to generate a universally and easily applicable toolkit to assess the MD sampling quality of any kinds of multi-trajectory experiments using certain error estimates and decide, whether the extracted thermodynamic properties are correct or not.

Zusammenfassung

In dieser Arbeit wird das Sampling von Molekulardynamik (MD) Simulationen von zwei flexiblen Biomolekülen untersucht: Methionin-Enkephalin (Met-Enkephalin) und dem dritten variablen Loop (V3) des Glykoproteins 120 (gp120) des Humanen Immundefizienz-Virus Typ-1 (HIV-1).

Die Dynamik von drei-dimensionalen (3D) Protein-Strukturen ist von zentraler Bedeutung für die Beschreibung der biomolekularen Funktion. Die Dynamik wird mittels MD Simulationen auf atomarem Level untersucht. Das Erreichen eines konvergierten MD Samplings in adäquater Zeit ist jedoch durch den riesigen Konformationsraum von flexiblen Molekülen begrenzt. Des Weiteren ist eine automatische Validierung des Samplings bisher nicht etabliert in heutigen MD Studien, und existierende Verfahren konzentrieren sich vorwiegend auf die Konvergenzanalyse einzelner Trajektorien. Dies ist für flexible Moleküle problematisch. Dabei ist es notwendig ein ausreichendes Sampling zu quantifizieren, ansonsten sind berechnete thermodynamische Größen bedeutungslos.

Das Ziel dieser Arbeit ist die Entwicklung eines Toolkits, welches die Samplingqualität von MD Simulationen von flexiblen Systemen quantifiziert. Dieses ist frei verfügbar unter https://github.com/MikeN12/PySamplingQuality. Hierzu werden verschiedene Sätze von Trajektorien aus verschiedenen Startkonformationen und sogenannte Enhanced Sampling Algorithmen wie accelerated MD (aMD) und scaled MD (sMD) kombiniert. Diese modifizieren die Energielandschaften um Übergänge zu vereinfachen. Die Samplingqualität wird durch zwei neue Messungen quantifiziert, dem Konformations- O_{conf} und Dichteüberlapp O_{dens}, unter Hinzunahme der Clusteranzahl N_C und der Entropie der Clusterverteilung S_C. Diese neuen Überlappgrößen klassifizieren die Selbstkonsistenz.

Met-Enkephalin wird als Testsystem verwendet, aufgrund dessen geringer Peptidlänge aber dennoch hochflexiblen Verhaltens. Unser Tool zeigt, dass bereits ein so kleines Molekül Simulationen von Mikrosekunden zur Konvergenz des Samplings benötigt. Weiterhin gilt, dass aMD sowohl Konvergenz als auch ungenügendes Sampling am schnellsten erkennt. Dennoch hat der Vergleich von O_{dens} zwischen MD und aMD/sMD gezeigt, dass die Modifikation des Enhanced Samplings nicht vollständig wiederhergestellt werden konnte. Dies kann jedoch mittels O_{dens} untersucht werden. V3 hingegen beweist, dass viel mehr Ressourcen als gewöhnlich notwendig sind, um Konvergenz zu erhalten. Die Ergebnisse unterstreichen die Notwendigkeit eines Multitrajektorien Ansatzes, um ungenügendes Sampling eindeutig zu erkennen.

Zusammenfassend ist es mit dem Toolkit möglich, das Sampling von Multitrajektorie-Experimenten zu validieren, unter der Angabe von Fehlerabschätzungen, und zu entscheiden, ob die berechneten thermodynamischen Größen korrekt beschrieben werden.

Table of Contents

List of Abbreviations

Cl^-	negative chloride ion
N_C	number of clusters
Na^+	positive sodium ion
$O_{\mathbf{conf}}$	conformational overlap measure
$O_{\mathbf{dens}}$	density overlap measure
S_C	cluster distribution entropy
2D	two-dimensional
3D	three-dimensional
ACE	acetyl moïety
aMD	accelerated molecular dynamics
CCR5	C-C chemokine receptor 5
CD4	cluster of differentiation 4 receptor
CDE	cluster distribution entropy
cMD	conventional molecular dynamics
CPU	central processing unit
CXCR4	C-X-C chemokine receptor 4
DNA	deoxyribonucleic acid
DOPE	discrete optimized protein energy modeling score
Env	human immunodeficiency virus envelope protein
Exp	exponential re-weighting for accelerated molecular dynamics

FFT	fast Fourier transformation
gp120	glycoprotein 120
gp41	glycoprotein 41
GPU	graphic processing unit
HIV-1	human immunodeficiency virus type-1
McL	Maclaurin expansion re-weighting for accelerated molecular dynamics
MD	molecular dynamics
Met-Enkephalin	methionine-enkephalin
Met153	second starting conformation of methionine-enkephalin
Met79	first starting conformation of methionine-enkephalin
MF	mean-field re-weighting for accelerated and scaled molecular dynamics
MM	molecular mechanics
NME	N-methylamine moïety
NMR	nuclear magnetic resonance
NPT	isothermic-isobaric ensemble
NVT	isothermic-isochoric ensemble
PBC	periodic boundary condition
PDB	protein data bank
PME	particle mesh Ewald summation
QM	quantum mechanics
RMSD	root mean square deviation
RNA	ribonucleic acid

sMD	scaled molecular dynamics
TIP3P	transferable intermolecular potential three-point water model
V3	third variable loop
V3a	first starting conformation of the V3-loop
V3b	second starting conformation of the V3-loop

List of Figures

List of Tables

1. Introduction

Studying biomolecular systems Studying the function of biomolecules, analyzing their diversity and investigating the evolutionary development are key fields to understand the fundamental principles of life. Describing the chemical and physical properties of various interlocking biological processes, like energy transport in cells, enzymatic reactions, replication, cell diffusion, receptor binding or treatment of diseases, is very difficult. To be intensively studied, these properties require the application of a combination of diverse tools. One relevant part to characterize the biomolecular function is encoded in the three-dimensional (3D) structures and their conformational dynamics describing the flexibility of the proteins [1–4]. Conformational flexibility of the systems is fundamentally required to adopt on different functions like enzymatic catalysis since in practice many processes do not work with the simple rigid model of a lock-and-key analogy [5, 6]. Moreover, different systems, substrates as well as receptors, are likely to change their shapes to ensure functionality described in conceptual models as induced fit [7] and conformational selection [6, 8]. Although experiments like x-ray crystallography are becoming better and better yielding 3D structures in resolutions of few Ångström [9], the dynamics of flexible molecules are hardly accessible by experiments at atomic scale. In the past few years, cryo-electron microscopy has become more popular in structural biology since the resolution is consistently enhanced to near-atomic scale [10, 11]. But it is still an open question, how to reliably analyze the dynamics of proteins or biomolecules in accurate resolution not detecting only background noise if they lack a well-defined structural composition due to their high flexibility. For example, in the entry process of the human immunodeficiency virus type 1 (HIV-1) into the human cell, the corresponding envelope protein gp120 undergoes several conformational changes, where the variable loops determine the co-receptor selection and binding [12–14]. Two 3D structures could only be obtained with intense work and various crystallization techniques [15, 16], but the dynamics are not exhaustively investigated, so far.

One common possibility to analyze the dynamics on molecular level are molecular dynamics (MD) simulations [17]. They have been widely used to simulate the dynamics of biomolecules since the 1970s [18, 19]. With increasing calculation power and enhancements in model descriptions and parameter calculation accuracy [20–24], it is possible to simulate hundreds of nanoseconds of systems of also larger size in explicit water in

adequate calculation time. MD simulations can provide insights into individual atomic motions during the course of the run, yielding important thermodynamic properties, which were not accessible beforehand. The only necessary condition, beside a correct description of the atoms and their interactions [25], is an exhaustive conformational sampling of the underlying energy landscape. This is simultaneously the limiting aspect of MD, since one would like to run simulations about ten times longer than the slowest important timescale in the system which can exceed 1 ms. This is hardly reachable for complex systems in reasonable calculation times. For flexible biomolecules with complex and rugged energy landscapes this is even harder since relevant conformations can be separated by large energetic barriers and the conformational space might be huge [26]. It is therefore a critical task to obtain converged MD simulations or at least assess the sampling quality quantitatively to know, how reliable are the results of thermodynamic observables. In the past few years, several methods have been developed to estimate the convergence of trajectories, reviewed in Refs. [17, 27]. There are well-documented software implementations [28] mainly focused on single trajectory [29–32] or two trajectory [33, 34] convergence assessment, or using a subset like the first two eigenvectors of PCA [29, 33, 35]. However, there is still a sizable portion of actual published MD studies which do not even mention the use of trajectory validations.

Research motivation Validating the sampling of MD trajectories of highly flexible biomolecules like the V3-loop is difficult and reveals problems if one relies solely on the established assessment tools. Furthermore, there is no settled workflow or tool for validating the sampling of highly flexible systems in an automated fashion using a multi-trajectory approach. Our research motivation was therefore to develop a universally applicable tool to quantitatively assess the sampling obtained by MD simulations of highly flexible systems. We incorporate the following points to treat this issue. We use *multiple trajectories* which are used without pre-processing to not suffer from information loss. We develop two simple self-consistency measures, the *conformational* and *density overlap*, which quantify the sampling quality between two up to numerous multiple trajectories. *Enhanced sampling algorithms* which speed up the dynamics are included and tested to ease conformational transitions. A simple *effective clustering* is implemented which handles huge amounts of data from multiple trajectories to analyze the size of the conformational space, yielding a comprehensive assessment along with the overlap measures. All assessment tools together with analysis methods and a possibility for visualization are implemented in a toolkit written in *Python* [36] and is freely available on *github* `https://github.com/MikeN12/PySamplingQuality` . These tools are easily usable and

a *documented tutorial* is attached. Furthermore, we published our tool in the *Journal of Chemical Theory and Computation* [37].

The work is organized as follows. In chapter 2, we introduce the computational methods to simulate the dynamics of biomolecules generating the input for our tool. The aim is to motivate the use of MD simulations and introduce all necessary parameters for the practical calculation like force-fields, thermodynamic observables, periodic boundary conditions, and system preparation. We also go more into the details of the sampling problem, giving a short overview about existing approaches and introduce the two enhanced sampling techniques used in this thesis. Finally, the chapter ends with the introduction of the two flexible biomolecules, specifying their origins and functions, along with homology modeling of unknown structures for starting MD runs.

In chapter 3, the tool for the sampling assessment is presented, starting with the general idea and motivation, defining the self-consistency measures in detail, introducing the effective clustering algorithm and showing the general workflow through the analysis.

Chapter 4 contains all results of the sampling analysis of both biomolecules methionine-enkephalin (Met-Enkephalin) and the third variable loop of HIV-1 gp120 (V3). The general goal is first to validate our tool by the small pentapeptide and discuss the parameters like the resolution and re-weighting of biased ensembles. Then, both molecules are analyzed for their convergence and sampling quality, combining all modules introduced in the previous chapter. Furthermore, we will conclude about the influence of biased sampling and the influence on thermodynamic quantities.

Finally, we will summarize the outcomes and give a brief overview about future applications and open points in chapter 5, which are not addressed in detail in this work.

2. Theory, background and motivation

In this chapter, all necessary theoretical concepts are introduced. This should give a brief and consistent overview to understand the motivation of the present study. We concentrate on the relevant parts of the theoretical background to support the research presented in this thesis. Therefore, we start with a brief introduction to quantum mechanics (QM) with the relevant approximation to make the transition to molecular mechanics (MM) which is treated with classical molecular dynamics (MD) simulations.

Since MD simulations play a central role in this thesis, they will be explained in more detail in the theoretical and also practical part. We want to motivate the necessity of complete exhaustive sampling of MD simulations to describe the dynamics of flexible biomolecules correctly and introduce the validation of MD trajectories. We ask ourselves: How can we quantitatively assess the sampling obtained by MD simulations for flexible biomolecules?

Additionally, we discuss the possibility of sampling enhancements and the two acceleration algorithms used in this study.

Finally, we introduce the two flexible biomolecules alongside with the structure modeling to construct 3D templates for MD runs.

2.1. Computational methods to simulate the dynamics of biomolecules

Biological functions of molecular systems are mainly driven by their dynamics. Protein folding, receptor-ligand binding, and many other processes undergo multiple states within their large conformational space to reach their full functionality [1–4, 8, 38]. For instance, the human immunodeficiency virus 1 (HIV-1) goes through several complex multi-state conformational changes during its replication cycle, which will be discussed in subsection 2.3.2. On the experimental side, it is so far not possible to study the full dynamics of flexible systems in a reasonable resolution (see subsection 2.3.2) to understand the underlying physicochemical processes.

This issue can be addressed with molecular modeling to extract the dynamics on the theoretical side.

2.1.1. Brief introduction to quantum mechanics

The dynamics and therefore any state of a molecular system are exactly described by a multi-dimensional wave function $\Psi(\vec{r},t)$ which obeys the time-dependent Schrödinger equation with the Hamiltonian $\mathcal{H}$

$$i\hbar\frac{\partial}{\partial t}\Psi(\vec{r},t) = \mathcal{H}\Psi(\vec{r},t)\,. \tag{2.1}$$

For a single and non-relativistic particle, the Hamiltonian $\mathcal{H}$ can be explicitly written in coordinate space obtaining

$$i\hbar\frac{\partial}{\partial t}\Psi(\vec{r},t) = \left[-\frac{\hbar^2}{2m}\nabla^2 + V(\vec{r},t)\right]\Psi(\vec{r},t) \tag{2.2}$$

with $\hbar$ is the Planck constant divided by 2π, i means the imaginary unit, $\partial/\partial t$ defines the partial derivative with respect to the time t, m is the corresponding mass of a specific particle, ∇^2 is the Laplace operator, $\vec{r}$ is the position vector, and $V(\vec{r},t)$ is the potential energy function. For n particles, the position vector and Laplace operator will be given as

$$\begin{aligned}\vec{r} &\in \{\vec{r}_1,\vec{r}_2,...,\vec{r}_n\}\\ \nabla^2 &= \sum_j^n \frac{\partial^2}{\partial\vec{r}_j^{\,2}}\end{aligned}\quad .$$

If the potential energy is not time-dependent $V(\vec{r},t) = V(\vec{r})$, the position and time coordinates can be separated, yielding the stationary solution [39]

$$\begin{aligned}\Psi(\vec{r},t) &= \psi(\vec{r})\cdot\phi(t)\\ \Rightarrow \mathcal{H}\psi(\vec{r}) &= E\psi(\vec{r})\end{aligned} \tag{2.3}$$

with E is the total energy and ψ eigenstate of the system. This can be generalized to a many-electron and many-nuclear system described by the Hamiltonian $\mathcal{H}_N$

$$\begin{aligned}\mathcal{H}_N = &-\sum_j^{\text{elec}}\frac{\hbar}{2m_j}\nabla_j^2 - \sum_A^{\text{nucl}}\frac{\hbar}{2m_A}\nabla_A^2 - \sum_j^{\text{elec}}\sum_A^{\text{nucl}}\frac{Z_A e^2}{4\pi\epsilon_0 r_{jA}}\\ &+\sum_{j<k}^{\text{elec}}\sum_{k>j}^{\text{elec}}\frac{e^2}{4\pi\epsilon_0 r_{jk}} + \sum_{A<B}^{\text{nucl}}\sum_{B>A}^{\text{nucl}}\frac{Z_A Z_B e^2}{4\pi\epsilon_0 r_{AB}}\,,\end{aligned} \tag{2.4}$$

with $r_{..}$ being the distance between two particles, $Z_{.}$ is the atomic number, e defines the electron charge, and ϵ_0 is the dielectric constant of vacuum. The Hamilton operator $\mathcal{H}_N$ describes a multi-dimensional problem with extensive degrees of freedom and could not been solved exactly, yet. There are three major approximations amongst others to handle the multi-dimensional problem: the *Born-Oppenheimer*, the *Hartree-Fock* and *Linear Combination of Atomic Orbitals* (LCAO) approximation (see chapter 2 of Ref. [40] and chapter 8 of Ref. [41]).

The *Born-Oppenheimer* approximation assumes fixed nuclei on the timescale of electron vibrations, and thus the electron Schrödinger equation is simplified by neglecting the second and fifth term in Eq. (2.4) (see chapter 2 of Ref. [40] and chapter 8 of Ref. [41]). The solution of this reduced equation yields the electron energy E_{elec}, whereas the total energy is obtained by

$$E_n = E_{\text{elec}} + \sum_{A<B}^{\text{nucl}} \sum_{B>A}^{\text{nucl}} \frac{Z_A Z_B e^2}{4\pi\epsilon_0 r_{AB}} \,.$$

The *Hartree-Fock* approximation assumes that the electrons move independently of each other. The motion of one electron can be calculated by self-consistency equations of an average-field, deduced by all other electrons, surrounding the particle in question. This approach yields a set of coupled differential equations which is presented in detail in Refs.[40, 41].

Lastly, we introduce briefly the *LCAO* approximation, which assumes that each molecular orbital is proportional to the (linear) sum of all atomic orbitals (see chapter 2 of Ref. [40]).

These and other approaches are discussed in detail in Refs.[40] and build the basis for the next sections.

The combination of these approximations allows to solve the effective Schrödinger equation numerically for approximately 50 to 100 atoms in adequate calculation times. This is far away from an automated calculation of large biomolecules and complexes.

2.1.2. Theory of molecular dynamics (MD) for simulating biomolecules

Molecular mechanics (MM) As discussed above, QM might give exact results but is limited to a small amount of atoms due to the complexity of the numerics and extensive amount of degrees of freedom. It is therefore necessary to use a classical approximation

to speed up the calculation for large and complex molecules. This is utilized by molecular mechanics (MM).

In general, MM consists of spherical atoms which are connected by (chemical) bonds. The motions are calculated from the distortion of their optimal geometric lengths due to non-bonded interactions. The latter are described by van der Waals and Coulomb interactions (see chapter 3 of Ref. [40] and chapter 2 of Ref. [42]). The requirements for good results obtained by MM are correct parameters for this classical model. These are not extractable by the model itself in contrast to QM, but they have to be evaluated and optimized by empirical values or come from QM calculations (see chapter 8 of Ref. [41]). The basis builds again the Born-Oppenheimer approximation that electrons of the system are assumed to move instantaneously together with the nuclei (see Ref. [43], chapter 1).

Newtonian dynamics The time evolution of the MM system with N particles is calculated by Newton's equation of motion for a conservative potential energy function $V(\vec{r})$

$$\vec{F}_j = m_j \frac{d^2\vec{r}_j}{dt^2} = -\vec{\nabla}_{\vec{r}_j} V(\vec{r}_j)\,, \tag{2.5}$$
$$\vec{\nabla}_{\vec{r}_j} = \begin{pmatrix} \frac{\partial}{\partial x_j} \\ \frac{\partial}{\partial y_j} \\ \frac{\partial}{\partial z_j} \end{pmatrix}\,, \quad j = 1, 2, ..., N\,,$$

where m_j and $\vec{r}_j$ are the mass and position of particle j, respectively, $\vec{\nabla}_{\vec{r}_j}$ defines the partial derivative vector with respect to the coordinates and $\vec{F}_j$ is the force acting on particle j. One possibility to solve these equations is the introduction of adequate potential energy functions discussed in the next paragraph.

In molecular dynamics (MD) simulations, the dynamics are simulated by integrating the equations Eqs. (2.5) for all particles in an step-wise fashion. In general, there are several possibilities to numerically integrate differential equations with higher order terms (see for example chapter 2 of Ref. [44]). In practice, there are several hundreds of thousand interactions where the forces have to be calculated following the order $\mathcal{O}(N^2)$, where N is the number of particles. Therefore in MD simulations, integrators with a low number of force calculations like the *Verlet* [45] algorithm are preferred to speed up the calculation with sufficient accuracy. Here, we use the *AMBER*14 simulation software [46], which uses

the simple *leapfrog* [47] algorithm similar to the so-called *Velocity-Verlet*

$$\begin{aligned} \vec{v}_j\left(t+\frac{\Delta t}{2}\right) &= \vec{v}_j\left(t-\frac{\Delta t}{2}\right)+\frac{\vec{F}_j(t)}{m_j}\Delta t \\ \vec{r}_j\left(t+\Delta t\right) &= \vec{r}_j(t)+v\left(t+\frac{\Delta t}{2}\right)\Delta t\,, \end{aligned} \tag{2.6}$$

where the velocities and positions are calculated in an alternating way (see Fig. 2.1). The

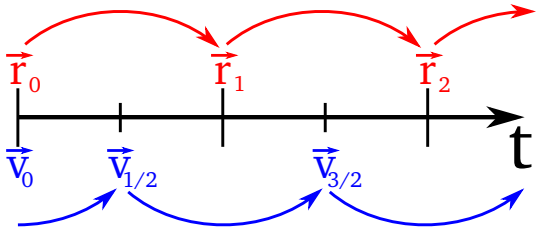

Fig. 2.1.: Schematic illustration of the *leapfrog* integration algorithm. The velocities $\vec{v}$ *leap* over the positions $\vec{r}$, which then *leap* over the velocities, again.

initialization of the velocities is usually done by setting a starting temperature T of the system by the classical relation

$$T = \frac{E_{\text{kin}}}{N_f k_{\text{B}}} = \frac{\sum_j m_j |\vec{v}_j(t=0)|^2}{2N_f k_{\text{B}}} \tag{2.7}$$

$$\sum_j \vec{v}_j = \vec{0} \tag{2.8}$$

with E_{kin} as the kinetic energy, N_f as the number of degrees of freedom and k_{B} being the Boltzmann constant. Eq. (2.8) ensures that there is no overall momentum in the system. Up to numerical uncertainties, this integration method is fully deterministic and reversible in time (see Ref. [44], chapter 3).

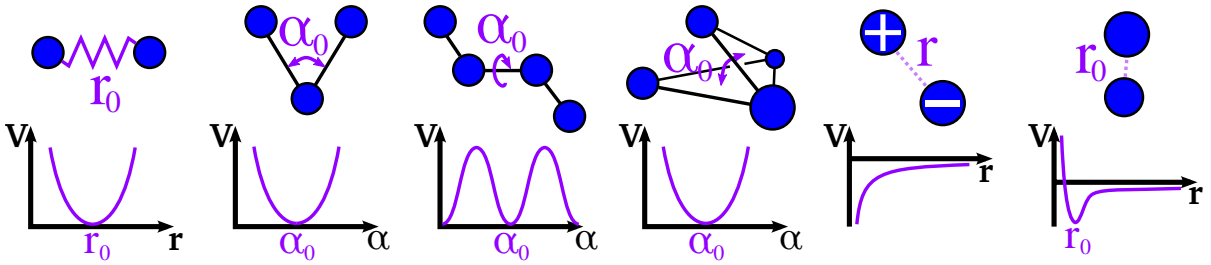

Fig. 2.2.: Schematic illustration of the contributions to the force-field potential function. From left to right: bond, angle, proper dihedral, improper dihedral, Coulomb and Lennard-Jones potentials.

Empirical force-fields The potential energy function $V(\vec{r})$ (Eq. (2.5)) is derived using the assumption that the effective molecular energy is expressed as the sum of potentials coming from physical forces describing the following contributions: The perturbations for bond stretching V_{stretch}, angle bending V_{bend}, and torsion contribution V_{torsion}, together with the non-bonded contributions from the van der Waals and Coulomb interactions shown in Fig. 2.2 (see chapter 3 of Ref. [40] and chapter 9 of Ref. [41]).

The functional form and parameters are obtained from empirical/experimental values and are agglomerated in so-called force-fields [20]. There are different groups of force-field parameterizations like *GROMOS* [48], *AMBER* [49], *CHARMM* [50], etc., which have different parametrization and either use unified or all-atom representations for certain groups of atoms.

In this study, we use the *AMBER* force-field ff99SB-ILDN [21, 51, 52]. It is used in various standard MD studies giving good results, and therefore fits in our research motivation to validate commonly obtained samplings. The force-field has the functional form

$$\begin{aligned} V(\vec{r},\vec{\theta},\vec{\omega}) = &\sum_{j}^{\text{bonds}} K_{r,j}\left(\vec{r}_j - \vec{r}_{j0}\right)^2 + \sum_{j}^{\text{angles}} K_{\theta,j}\left(\vec{\theta}_j - \vec{\theta}_{j0}\right)^2 \\ &+ \sum_{j}^{\text{torsions}} \sum_{u=1}^{3} K_{\omega,u,j}\left[1 + \cos\left(|u\,\vec{\omega}_j - \vec{\gamma}_{u,j}|\right)\right] \\ &+ \sum_{j}^{\text{nonb}} \sum_{k>j}^{\text{nonb}} \left[\frac{A_{jk}}{|\vec{r}_{jk}|^{12}} - \frac{B_{jk}}{|\vec{r}_{jk}|^{6}} + \frac{Q_j Q_k}{\epsilon |\vec{r}_{jk}|}\right], \end{aligned} \tag{2.9}$$

with $K_{.,.,.}$ are the force constants for the bonds, angles or torsional potential terms, $\vec{r}_{.0}, \vec{\theta}_{.0}$ are the ideal/equilibrium bond lengths or angles, $\vec{\gamma}_{.,.}$ is a phase of the dihedral angle, u is the dihedral periodicity, $A_{..}, B_{..}$ are the parameters for the Lennard-Jones potential, $Q_.$ is the parameter for the charges involved in the Coulomb potential, ϵ is the parameter defining the dielectric permittivity and $\vec{r}_.,\vec{\theta}_.,\vec{\omega}_.$ are the instantaneous bond length, angle and dihedral angle, respectively. Finally, $|\vec{r}_{..}|$ is the distance between two particles.

The first three terms of Eq. (2.9) define the bonded interactions involving two, three and four atoms. The covalent bonds (first summation) are treated in MM as simple harmonic oscillators following *Hook's* law. This implies that effects like bond-breaking events or other quantum chemical reactions cannot be treated. In the same way as bond lengths, the changes of angles are calculated. The torsional term handles the proper and improper dihedrals with their inherent periodicity. Proper dihedrals define the angles

involved with four covalently bounded atoms, whereas improper torsions handle planar groups as illustrated in Fig. 2.2.

The last summation models the pair-wise non-bonded terms which involves van der Waals and charge interactions. These are expressed by the well-known 6-12 Lennard-Jones (depending on atomic radii and distances) and the Coulomb potential (depending on partial atomic charges and distances). The two sums are iterated over all pairs of atoms. If they are bound by covalent bonds, only contributions from atom pairs which are at least separated by three covalent bonds are taken into account.

The parametrization is a crucial step to obtain adequate results compared to experiments. A brief review can be found in Ref. [20] about the developments of different force-fields. For the *AMBER* force-field ff99SB-ildn, equilibrium bond lengths, angles and their force constants were taken from crystal structures and fitted to match normal mode frequencies [20]. Charges were possible to be derived by quantum chemistry calculations fitting them to the quantum electrostatic potentials due to increasing computation power. In the preliminary force-field ff94 [49], using a restrained electrostatic potential fit, charges were parametrized as averages of multiple conformations [49]. Van der Waals parameters were obtained from fits to amide crystal data and optimized further with liquid-state simulations [20]. Obtaining reasonable torsional parameters is a difficult task, since they are closely related to the non-bonded potentials. They are in general obtained by matching torsional barriers extracted from experiments or quantum chemistry calculations [20]. The set of dihedral parameters were subsequently improved including long-range effects (ff99 [51]), high order ab initio quantum mechanical calculations (ff99SB [21]) and QM data validated with nuclear magnetic resonance (NMR) results improving side-chain torsion potentials for four amino acids (ff99SB-ILDN [52]).

In general, different force-fields are parametrized to focus on different tasks of protein function and parameters are not equivalent. For instance, the *AMBER* ff99SB-ILDN force-field makes extensive use of ab initio QM data, whereas for instance the *OPLS* (Optimized Potentials for Liquid Simulations) is based mainly on liquid-state thermodynamics [53].

On the basis of one force-field parameter set, transferability of many representative parameters is an advantage of such a treatment. Parameters like bond lengths or bond angles can be transferred from small molecule parametrizations to similar/related and possibly larger molecules, since they adopt similar states under normal conditions (see Ref. [41], chapter 9). Therefore, parameters for unknown compounds can be approximated beforehand without fitting the full complex on experimental values. Nevertheless, this

is not applicable on special cases such as cycloalkanines, where the values may vary significantly from equilibrium values (see Ref. [41], chapter 9).

Remarkably, the simple classical description works generally well for describing many selected molecular structures and processes [41]. For instance even fast protein folding simulations could be directly compared with experimental structures [54, 55]. But still, force-field evaluation is a hard and complicated process [25].

A final note should be mentioned here: In general, MM energies have no physical equivalent meaning due to the simple approaches, but the differences between two or multiple conformations can be compared to experimental values [42]. It is very hard to extract the correct absolute energy values from MD simulations [42].

For further details, we recommend Refs. [40, 42, 43].

Thermodynamic ensembles: Molecular dynamics (MD) simulations as described above, produce trajectories containing all atomic positions $\vec{r}$ and momenta $\vec{p}$ for each timestep (Ref. [56], chapter 1). Thus, one obtains the microscopic behavior of the biomolecular system. To be able to measure relevant thermodynamic observables, the microscopic information is transferred to macroscopic properties following statistical mechanics. The essential foundation builds the *ergodic hypothesis*: The time average over a (optimal) trajectory is equal to the ensemble average of the system (Ref. [56], chapter 10). This allows to extract thermodynamic properties like heat capacity, pressures or energies from MD trajectories. For example the kinetic energy E_{kin} is given by

$$E_{\text{kin}} = \left\langle \sum_j \frac{\vec{p}_j{}^2}{2m_j} \right\rangle \tag{2.10}$$

with $\langle . \rangle$ is the time average, m_j the mass of particle j, and the sum goes over all particles j.

In detail, we will use the isothermic-isochoric NVT ensemble for preparing the system into a thermostat. It is used to describe a closed system which is weakly coupled to a thermal heat bath (see chapter 2 of Ref. [56]). There are different algorithms which keep the temperature at a constant level and exchange energy with the environment: Berendsen [57], Nosé-Hoover [58, 59] and Langevin [60]. The latter is used in this study because it gives a more stable temperature coupling and temperature constancy behavior. The Langevin thermostat adds a random frictional force from a Gaussian distribution with a certain collision frequency to control the temperature by adjusting the kinetic energy of the particles in the system (see Ref. [60] and chapter 7 of Ref. [44]). The

isothermic-isobaric NPT ensemble is often used to mimic laboratory conditions. The Berendsen barostat [57] keeps the pressure constant by a isotropic position scaling of the simulation box. We will use NPT as the production ensemble for the MD runs.

The definition of an optimal trajectory together with the fulfillment of the ergodic hypothesis will be discussed in the sampling section 2.2 and plays the central role of this thesis. The sampling quality is critical for the correctness of the observables obtained by MD simulations (assuming that the force-fields and classical approximations are correct). The validation is crucial to be able to rely on the results.

2.1.3. Requirements to run MD simulations

As already mentioned, we use the *AMBER*14 simulation software with the ff99SB-ILDN [52] force-field. In this subsection, we specify the necessary details to setup the MD simulations used in this study. For the numerical approach, we need to introduce different conditions, which tackle the infinitely large space, the solvent, the energetic minimization of the system and the equilibration.

Periodic Boundary Conditions (PBC) The first condition treats the size of the system. It is clear, the larger the system, the more particles have to be calculated and the longer the simulation will take to finish. On the other hand, if the system is set to a small finite size, it will rarely produce correct behavior of a naturally infinite system. Additionally for a finite system, it is problematic if particles interact with the edges of the box.

These effects can be overcome by introducing periodic boundary conditions (PBC): The molecule is placed in a 3D periodic unit cell considering to have infinitely many copies (called mirror images) in space (see for instance chapter 10 of Ref. [41] and chapter 1 of Ref. [44]). This means, if a particle moves outside the unit cell, it is returned into the box on the opposite side. This is schematically illustrated in 2D in Fig. 2.3. Several periodic boxes exist, like the cubic, truncated octahedron or a hexagonal prism (see bottom of Fig. 2.3). The general idea is to minimize the number of (solvent) atoms for the simulation. Thus longer molecules like Deoxyribonucleic Acid (DNA) chains are typically placed to hexagonal boxes [41]. We will use the truncated octahedron, because our molecules are more centered.

The PBC must fulfill two criteria: Particles are not allowed to have interaction contributions between their mirror images, and the long and short term interactions must be treated adequately. The long-range electrostatic field is proportional to the inverse distance $1/|\vec{r}|$ and the number of terms is proportional to the quadratic number of parti-

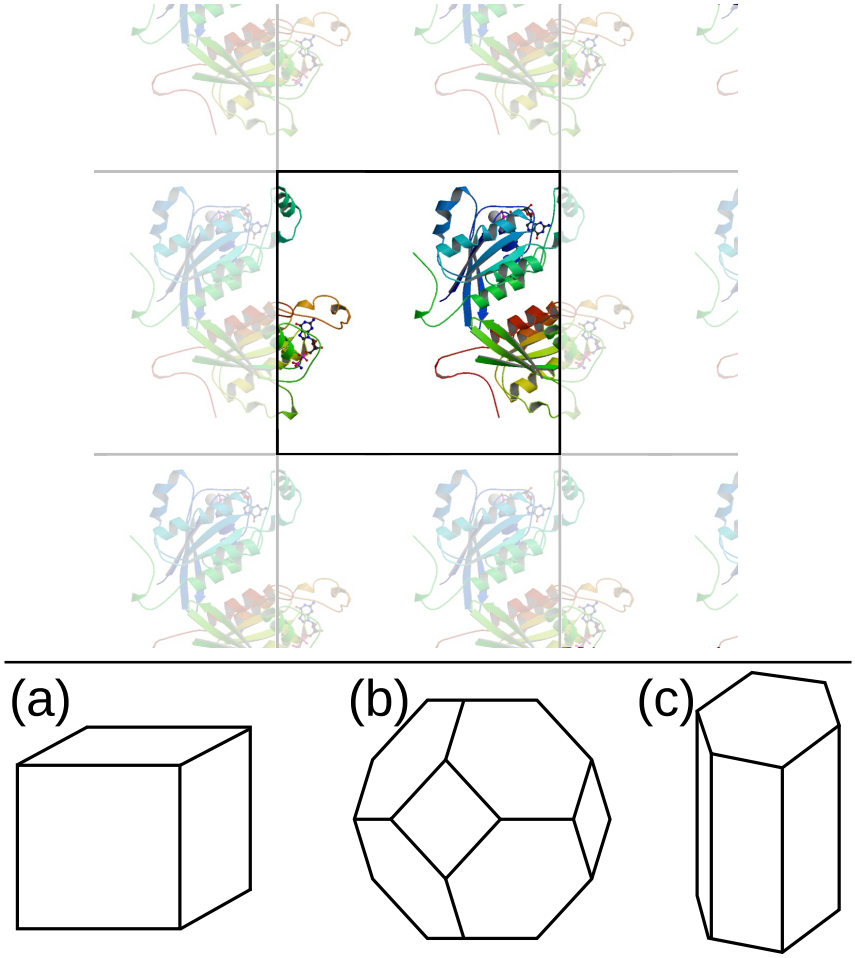

Fig. 2.3.: Periodic boundary condition. Top: 2D illustration of a periodic boundary condition (PBC). Bottom: Examples of periodic domains in 3D: (a) Cubic, (b) truncated octahedron and (c) hexagonal prism boxes.

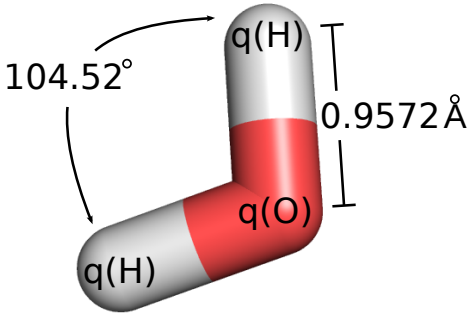

Fig. 2.4.: The Transferable Intermolecular Potential three-point TIP3P water model. The oxygen and hydrogen charges $q(O)$ and $q(H)$, respectively, are given in Eq. (2.13).

cles N^2. Thus in general, the sum of the Coulomb potential does not converge. A slow convergence can be obtained by neutralizing the system with counter ions sodium Na^+ or chloride Cl^- (see for instance the Appendix of Ref. [43]). In practice, we will additionally use the Particle Mesh Ewald (PME) [61] summation decomposing the Coulomb potential into short- and long-range terms

$$V_{\text{Coulomb}}(\vec{r}) = V_{\text{short}}(\vec{r}) + V_{\text{long}}(\vec{r})\,. \tag{2.11}$$

The short-range terms are treated by direct summation of force contributions, and the long-range part is calculated by a convolution on a discrete grid in reciprocal space using 3D Fast Fourier Transformations (FFT). The latter follows asymptotically $\mathcal{O}(N\log(N))$ in computational complexity [61]. The van der Waals terms are proportional to $1/|\vec{r}|^6$ and converge quickly. For numerical treatment, both non-bonded interactions are truncated using a cutoff of 0.8-1.2nm using a minimal-image convention where typically each particle interacts only with the closest periodic image of the other particles (see for instance chapter 10 of Ref. [41]).

Water treatment Physical relevant properties of biomolecular systems will rarely be reproduced using vacuum simulations. One therefore needs to model an aqueous solvent to investigate the hydration influence at biomolecular surfaces on protein-ligand binding or enzymatic function. A benchmark on these effects using different force-fields and explicit water models was done in Ref. [62]. In general, the solvent is modeled by individual water molecules as rigid bodies with partial charges, a certain OH-distance and van der Waals interactions.

There are different explicit water models parametrized which are optimized to match different experimental properties like hydration enthalpies or heat capacities [62]. The most prominent models are the normal and extended single point charge (SPC, SPC/E) models [63, 64] and the transferable intermolecular potential three-, four- and five site models (TIP3P, TIP4P, TIP4Pew, TIP5P) [65–67]. Since the force-field ff99SB-ILDN is parametrized with TIP3P water, it is recommended to use this solvent model. The TIP3P water model illustrated in Fig. 2.4 was parametrized in 1983 [65] reproducing the experimental dimerization energies $E^{\text{dim}}_{m,n}$ between two water molecules m and n with

$$E^{\text{dim}}_{m,n} = \sum_j^m \sum_k^n \frac{q_j q_k e^2}{|\vec{r}_{jk}|} + \frac{A}{|\vec{r}_{OO}|^{12}} - \frac{C}{|\vec{r}_{OO}|^6} \tag{2.12}$$

with $q_.$ is either the hydrogen or oxygen charge, e is the elementary charge, $|\vec{r}_{jk}|$ is the distance between two atoms (hydrogen, oxygen) of two different water molecules, A, C are the Lennard-Jones parameters which were determined, and $|\vec{r}_{OO}|$ is the distance between two oxygen atoms. The fitted parameters [65] are

$$\begin{aligned} d(\mathrm{OH}) &= 0.9572\,\mathring{A} \\ \alpha(\mathrm{HOH}) &= 104.52^\circ \\ A &= 2.435\,\frac{\mathrm{kJ}\mathring{A}^{12}}{\mathrm{mol}} \\ C &= 2489.480\,\frac{\mathrm{kJ}\mathring{A}^{6}}{\mathrm{mol}} \\ q(\mathrm{O}) &= -0.834 \\ q(\mathrm{H}) &= 0.417 \end{aligned} \tag{2.13}$$

(see also Fig. 2.4).

It should be noted that, since the main contributions to MD simulations are the solvent-solvent and protein-solvent interactions (with explicit water), an accurate and adequate description of the water molecules is needed and studies in this field might enhance the overall MD result, which is not the focus of this work.

System preparation A global, energetic minimization of the system is used to reduce bad bond lengths and torsional angles improving the overall system geometry (see chapter 13 of Ref. [41]). Especially, if the structure is built using a homologue (section 2.4), it can be significantly enhanced removing unfavorable interactions or states [41]. Additionally, an energetic minimization can relax the rigid water molecules which are put in a grid-fashion into the system.

There are many possibilities to perform multi-dimensional energy minimization. Here, we will focus on two popular first derivative algorithms, which are fast and memory efficient: *steepest descent* followed by *conjugated gradient*. Both algorithms are described in more detail in Ref. [41]. The idea is to follow the gradient of the energy function with *steepest descent*. This is very fast for large slopes and becomes slow reaching the (local) minimum. Thus, it is usually followed by the *conjugated gradient* algorithm, which becomes more efficient closely to the energy minimum. The latter is computationally more expensive, since it uses two successive gradients to guess the direction toward the (local) minimum.

Obtaining the global minimum of the system is very difficult without knowing the

complete underlying energy landscape. Thus one usually will start the MD simulation in a local minimum [68]. This is not a problem, if the sampling is complete which is discussed in section 2.2. The main purpose of the energy minimization is the optimization of length or steric distortions between particles [68].

The usual practice after minimization is called *equilibration*. In the past, groups often discarded the first part of their simulation based for instance on structure deviations compared to the crystal structure [69]. In fact, Genheden et al. [70] could show that in general all simulated structures are equally important after an appropriate preparation where unphysical interactions might occur due to the arbitrary starting structures of the MD simulation.

Here, we will use a multi-state preparation described in section 4.1 to overcome these unphysical artifacts especially for the structures built by homologues. The steps involve multiple cycles of energy minimization and short simulations trying to kick the ongoing optimized structure out of inappropriate local energy minima (see Fig. 2.5) and relax the system toward their equilibrium behavior.

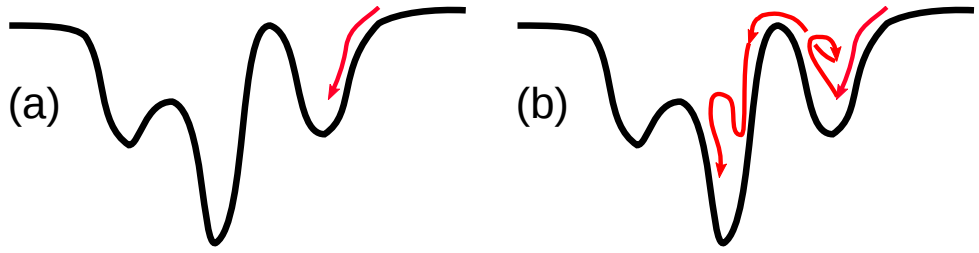

Fig. 2.5.: Schematic multi-state system preparation. (a) A single minimization may end in a local minimum with large energy. (b) Short MD simulations may kick the protein out of local minima, followed by another energy minimization.

Typical timescales for molecule simulations When setting up a MD simulation, one needs to keep in mind that the dynamics are calculated in discrete timesteps. There is a general trade-off between the step size and the length of the simulation. The shorter the step size, the more force interactions must be calculated and the less total simulation time can be obtained in the same calculation time. On the other hand, the longer the timesteps between force calculations, the less accurate certain properties are treated. The typical timesteps/-scales are illustrated in Fig. 2.6.

The timestep Δt for the MD integration is determined by the requirement $\Delta t \ll t_{\text{period}}$ where t_{period} is the period of the highest frequency motion in the system (see for example chapter 13 of Ref. [41]). If the step is set $\Delta t \geq t_{\text{period}}$, these effects are not taken into account and the system might become unstable. A common treatment is to constrain some

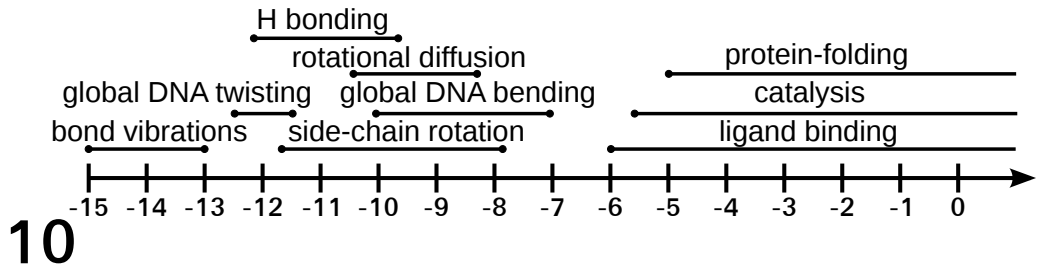

Fig. 2.6.: Typical timescales for molecule dynamics [41, 75, 76]. The scale goes from femtoseconds (10^{-15}) to seconds (10^{0}).

or several bond lengths to their equilibrium lengths because they do not significantly alter the dynamical properties of biomolecular functions in the simulations [71]. This constrain corresponds to the out-averaging of bond vibrational motions with $t_{\text{period}} \approx 3 - 8$ fs [72] and allows the user to set the integration step around three to four times as large when bonds are constrained instead of being treated as harmonic oscillators [73]. The timestep Δt can even be increased by constraining the next fastest motions (e.g. angle vibrations involving hydrogen atoms $t_{\text{period}} \approx 13$ fs [72]). But then one has to ensure to not over- or underestimate important properties of the system. Thus, we will only use the bond constrains involving hydrogen atoms and use an integration timestep of $\Delta t = 2 \cdot 10^{-15}$ s $=$ 2 fs. In *AMBER*14, this is done with the SHAKE [74] algorithm, which modifies the *leapfrog* integration algorithm to constrain the chosen bonds, i.e. the constrained bonds are kept at their constant equilibrium distance.

For such small timesteps $\Delta t \propto$ fs, one has to simulate a massive amount of states to obtain characteristic timescales in biomolecule processes like protein folding ($\propto$ ms to s, see Fig. 2.6). On the other hand, only reaching the necessary timescales does not mean that the sampling is sufficient. The question about sufficient sampling plays the central role of this thesis and will be discussed in the next section.

2.2. Sampling

A complete and exhaustive sampling is the main task which has to be fulfilled by MD simulations, so that the system is ergodic and observables can be extracted by time-averages (see subsection 2.1.2). An exhaustive sampling means that the complete conformational space is visited with its correct probability density distribution $p(\vec{r})$. In a (discrete) clas-

sical system, the density function follows

$$p(\vec{r}) = \frac{e^{-\beta V(\vec{r})}}{Z} = \frac{e^{-\beta V(\vec{r})}}{\sum_j^{\text{states}} e^{-\beta V(\vec{r_j})}}\,, \tag{2.14}$$

where Z is the partition function of the system, $V(\vec{r})$ is the potential energy, and $\beta = 1/(k_\mathrm{B}T)$ is the inverse temperature function with the temperature T and the Boltzmann constant k_B. (Note that in general for a classical system, the states are continuous and the sum is then replaced by an integral over all states. For convenience, this representation should suffice for the general introduction.) But what happens, if there are many rare transitions in the system, where conformations are separated by large energetic barriers? The sampled system may then be partitioned into different energetic minima connected by low transition probabilities [26, 77]. That means, such systems show broken ergodic behavior and even long simulations may stay trapped in an irrelevant local minimum [78]. The result is even worse, if a trajectory shows convergence, because the relaxation time in its energy well is much faster than the transition [26], which is illustrated in Fig. 2.7. Such a trajectory is then energetically trapped. Genheden et al. [70] showed that in-

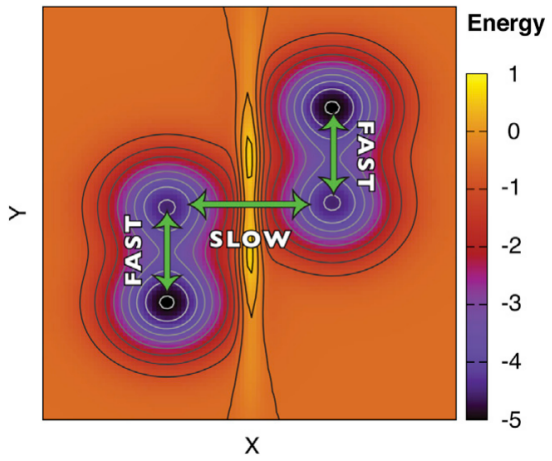

Fig. 2.7.: 2D energy landscape showing different relaxation times. Copyright 2014 from Ref. [26], reused by permission from Elsevier.

complete sampling leads to incorrect thermodynamic observables. In some cases it is possible to classify the sampling quality by direct comparison with experiments: in a fast folding protein, the well-defined conformational clusters from simulations showed good agreement with experimental values [54, 55]. But in general, such experimental values are missing. Furthermore, massively flexible systems of intrinsically disordered proteins [79] have complex and rugged energy landscapes.

It is therefore of crucial importance to quantify and validate the sampling quality of MD trajectories. Only then, one is able to obtain correct thermodynamic results. The central task of this theses will be to shed light into this field.

2.2.1. Validation of MD sampling

State of the art Still today, if one takes a representative look into published MD studies, there is a sizable portion which do not even report the use of sampling validation. There are several validation analyses and implementations which can be separated into three groups: (1) Single trajectory validation, (2) two trajectories validation and (3) multi-trajectory subspace validation. We will give a brief review about the current sampling validation techniques and their limits, which is necessary to make the transition to our own solution of this problem.

Single trajectory validation The first group is based solely on the information of one trajectory. The purpose is the classification of one MD run whether it has successfully converged or it should be discarded due to wrong behavior. On the other hand, the other big advantage is that only a small amount of data must be generated and the convergence prediction is very fast.

Trajectories are taken either as a whole or split into multiple parts. The first quantities are for instance the decorrelation time of the system [29, 30], the number of found clusters as a function of the simulation time [80] and the corresponding cluster distribution entropy as a function of the time [32]. The decorrelation time is error-prone since it can wrongly lead to a converged picture if there are slow relaxing along fast relaxing transitions [26]. The other two quantities are based on a clustering which must be correctly validated. This is a non-trivial problem and strongly depends on the data [81].
The second quantities are for instance the block averaging method [29] (reviewed in Ref. [27]), randomly distributed block population histograms [82], the effective sample size calculation based on a structural decorrelation time [30] and the block covariance overlap method based on the covariance matrix between all structures in one trajectory [83]. They mostly aim to extract standard error estimates of observables for increasing block sizes or compare populations in partitions of different trajectory parts. The disadvantage is that these estimates are specialized for some observables giving no information about the true convergence [27]. On the other hand, a partitioning into disjunct blocks or clusters should be used with care. If one imagines a very flexible system, where there are no clear contiguous conformational clusters, but different structures are lumped

closely together, errors in the population probabilities may occur between different non-representative clusters.
The largest drawback of single trajectory convergence criteria is the question, what happens if the trajectory is trapped without knowing that something is missing? It might always happen that one trajectory shows convergence in its limited space, because there were no transitions to another rare events. This single trajectory validation is a useful pre-filtering method to exclude completely wrong results, but it will not help in answering, whether the sampling is complete. It is therefore favorable to use more trajectories for a quantitative sampling assessment, since they have to separately reproduce the results.

Two trajectory validation There are two validation techniques using two trajectories: the root mean square inner product (RMSIP) [34] and the covariance overlap [33, 84, 85] mentioned previously. Both approaches are based on the construction of the covariance matrix between all structures of one trajectory and extracting the eigenvectors and -values. These are representative for the specific trajectory, and so different trajectories can be compared with the RMSIP and covariance overlap using all modes. The two approaches give then values between 0 and 1 for poor and perfect agreement. The problem is that these approximations are based solely on two trajectories and they are not generalized for a multi-dimensional problem. Additionally, Hess [33] studied that if the single covariance matrix did not converge, the analysis might give wrong results.

Subspace validation In the so-called subspace validation group, one can investigate the sampling quality between numerous trajectories, but this is limited to the subspace of the trajectories. Often, only the first two eigenvector projections from the covariance matrices are used to identify, whether different trajectories show a shared sampled space. Also if higher modes have small eigenvalues, it is in general problematic and alters the results to use only a subset [33]. Furthermore, the projection on the eigenvectors of the covariance matrix may produce artifacts if not done properly. The group of Gerhard Stock [86, 87] showed that using cartesian coordinates produces artifacts because internal and overall motions of the system are not well separated.

Multiple independent simulations It could be shown that combining multiple (shorter) MD trajectories with different initial conditions improves the conformational sampling compared to one or few long trajectories [88]. Genheden and Ryde [89] showed the advantages of three different initialization procedures: Velocity Induced Independent Trajectories (VIIT) using different starting velocities, Solvation Induced Independent Tra-

jectories (SIIT) using different solvation boxes and Conformation Induced Independent Trajectories (CIIT) using different starting conformations for instance different crystal structures. The more different the initial conditions are, the less probable it is to obtain synchronization effects between different runs, and thus the less probable it is that two trajectories are trapped in the same energetic minimum. Only if all different trajectories (which are not wrong due to trapped behavior or similar reasons) do reproduce the sampling, the sampling can be complete. It is therefore puzzling that the use of multiple starting conformations or a multi-trajectory sampling validation is not rigorously established to quantify the sampling assessment. For rigid and simple systems, this might not be necessary. But for highly flexible and complex structures it is questionable whether the results are correct without a proper validation.

2.2.2. Research motivation

The lack of a proper and universal sampling validation scheme for flexible molecules encouraged us to generate a tool to quantitatively assess the sampling quality of MD simulations with the focus on flexible systems. We incorporate the previously mentioned advantageous conditions: universality, a multi-trajectory approach and the use of different starting conformations. We use the full trajectory without information loss, no pre-partitioning and define multiple supporting criteria to obtain a classifier between 0 (poor sampling) and 1 (perfect sampling) alongside the possibility of detecting the reason of poor sampling. This software package written in Python is freely available as source code at `https://github.com/MikeN12/PySamplingQuality` and will be explained in full detail in chapter 3.

The validation is critical to obtain correct thermodynamic properties. But how can one enhance the sampling of conventional MD simulations? There are several possibilities which are briefly introduced in the next subsection.

2.2.3. Sampling enhancements

In subsection 2.1.3, we discussed the typical timescales for biomolecular systems. With conventional MD (cMD) simulations and standard computer hardware, for a long time it was possible to simulate only few nanoseconds (ns) for large proteins. Also, small systems are computationally costly [90].

MD runs can be used as a super-microscope to investigate proteins at atomic level, but it is highly problematic if they do not reach relevant timescales for certain physical

reactions (see Fig. 2.6). There are many studies with focus on the acceleration of MD simulations, i.e. favoring rare transitions. This can be done on the hardware or software level.

Hardware acceleration In the last decades, the computational power increased exponentially allowing the user to explore longer MD runs. 2007, the Shaw group [91] released a specialized super computer for long MD simulations, allowed to access milliseconds timescales and investigate folding-unfolding events [22, 92, 93]. This is supported by other super-computing centers.

Additionally, there are also developers incorporating the power of graphic processing units (GPU) alongside the central processing units (CPU). The largest advantage is the possibility for massive parallel computing. There are different MD simulation packages like *AMBER* [46] or *GROMACS* [94] supporting the use of GPUs, gaining large speed ups in calculation [23, 95–99]. Thus, we will use the GPU power of the *AMBER*14 implementation.

Recently, novel chips are developed called Accelerated Processing Units (APU) which combine CPU with GPU architectures, showing that this might push the parallel computation power even further [100]. All these developments sound very promising for advancing macromolecular simulations and modeling.

Enhanced sampling algorithms On the other hand, the sampling can effectively be accelerated using enhanced sampling techniques without further hardware costs. These allow to sample not only larger regions of the conformational space but also inaccessible rare events of cMD in the same simulation time. They can be summarized into two groups:

1. Algorithms which guide the simulation along certain pathways.

2. Techniques modifying the energy landscape introducing a certain energetic bias.

The first group uses prior knowledge of the system, to define collective variables, sampling along a certain free energy path and/or using history dependent potential modifiers, like umbrella sampling or metadynamics [101–106]. For flexible and unknown biomolecules, it might be problematic to extract or estimate these prior conditions, because appropriate cMD simulations might be necessary.

The second group is instead directly applicable, which incorporates for instance replica exchange molecular dynamics (REMD) [107] running different exchanging simulations with different temperatures, simulated tempering [108, 109] varying the temperature

within a single run or using non-Boltzmann distributions to bias simulations toward rare-events like integrated tempering [110] or accelerated MD (aMD)/scaled MD (sMD) [111–113]. The canonical ensembles for the non-Boltzmann distributions are recovered by re-weighting [114, 115].

Due to the direct applicability, we will use aMD and sMD and integrate them into our analysis tool introduced in chapter 3 to enhance the simulations of flexible biomolecules. The two algorithms are defined in the following.

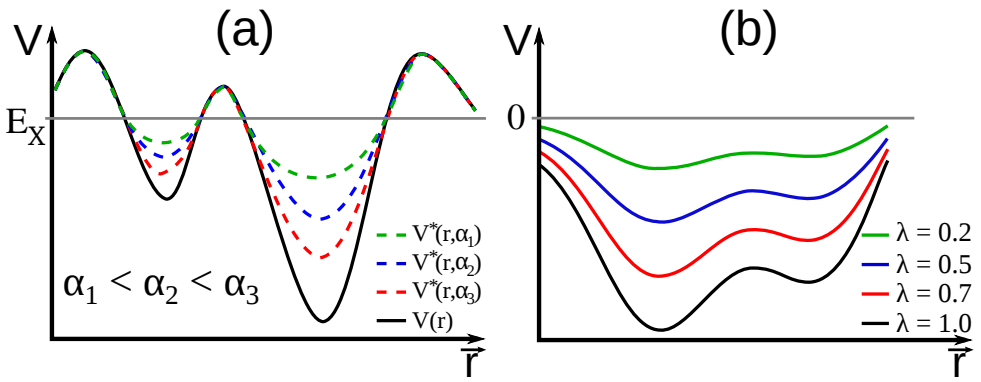

Fig. 2.8.: aMD and sMD potentials. Biased potentials $V^*(\vec{r})$ for different parameters for (a) aMD [116] and (b) sMD [113].

aMD To speed up the dynamics and thus to ease conformational transitions, we use aMD [111, 112] which applies a boost potential $\Delta V(\vec{r})$ lifting potential energies below certain thresholds E_X (see Fig. 2.8 left). Hence, simulations are performed with boosted potentials $V^*(\vec{r})$ instead of the standard force field $V(\vec{r})$:

$$\begin{aligned} V^*(\vec{r}) &= V(\vec{r}) + \Delta V(\vec{r}) \\ &= V(\vec{r}) + \Delta V_P(\vec{r}) + \Delta V_D(\vec{r})\,. \end{aligned} \tag{2.15}$$

Here, we apply a dual boost combination [116, 117] of potentials $\Delta V_P(\vec{r})$ on the total potential energy and an additional $\Delta V_D(\vec{r})$ for dihedral energy terms with

$$\Delta V_X(\vec{r}) = \begin{cases} 0 & \text{for } V_X(\vec{r}) \geq E_X \\ \frac{(E_X - V_X(\vec{r}))^2}{\alpha_X + (E_X - V_X(\vec{r}))} & \text{for } V_X(\vec{r}) < E_X \end{cases} \tag{2.16}$$

with one equation for $X = D$ and one for $X = P$. It is also possible to boost only the dihedral potential energy terms or apply the boost only on the total potential energy [116]. The parameters E_P, E_D are different thresholds and α_P, α_D are inverse strength factors for the total potential and the dihedral potential, respectively (see Fig. 2.8 left).

These parameters directly affect the strength and shape of the acceleration function. For example, too large thresholds E_P, E_D with simultaneous too low α_P, α_D may lead to flat and isoenergetic landscapes, where the statistics are dominated by a few heavily weighted points [111]. In general, it is recommended to use various sets of acceleration parameters for validation and to systematically screen through the conformational space [116, 118]. The latter may allow the user to obtain new conformations to start cMD or another simulations which require prior knowledge like reaction coordinates. Here, we use as first approximate the recommended values from Pierce et al. [117]

$$\begin{aligned} E_P &= \langle E_{P,\mathrm{cMD}} \rangle + \alpha_P, \qquad \alpha_P = 0.16 \frac{\mathrm{kCal}}{\mathrm{mol\ atom}} \cdot N_{\mathrm{atoms}}, \\ E_D &= \langle E_{D,\mathrm{cMD}} \rangle + 5 \cdot \alpha_D, \quad \alpha_D = \frac{4}{5} \frac{\mathrm{kCal}}{\mathrm{mol\ residue}} \cdot N_{\mathrm{res}}, \end{aligned} \tag{2.17}$$

with $\langle E_{P/D,\mathrm{cMD}} \rangle$ averaged energies from corresponding cMD simulations, and N_{atoms}, N_{res} the numbers of atoms and residues, respectively. The acceleration introduces a biased distribution $p^*(\vec{r})$

$$p(\vec{r}) = e^{-\beta V(\vec{r})} \rightarrow p^*(\vec{r}) = e^{-\beta V(\vec{r})} \cdot e^{-\beta \Delta V(\vec{r})} \tag{2.18}$$

where the unbiased distribution $p(\vec{r})$ can be obtained by multiplying the inverse Boltzmann factor, with $\beta = (k_{\mathrm{B}} T)^{-1}$ is the temperature factor defined by the reciprocal Boltzmann constant k_{B} and the temperature T. This is a critical step, because if the re-weighting is not done correctly, all thermodynamic observables will be biased although the trajectory might explore a large conformational space.

Pierce et al. [117] divided their systems into N bins, assuming that all data within a bin is in the same microstate. For the concrete example of a discrete 1D biased unnormalized distribution H_a^* with uniformly distributed bins a, one can then obtain the unbiased distribution H_a by

$$H_a^* = \sum_{j=1}^{J} \begin{cases} 1, & j \in \{\mathrm{bin}_a\} \\ 0, & \mathrm{else} \end{cases} \tag{2.19}$$

$$\Rightarrow H_a = \sum_{j=1}^{J} \begin{cases} e^{+\beta \Delta V_j(\vec{r})}, & j \in \{\mathrm{bin}_a\} \\ 0, & \mathrm{else} \end{cases} \tag{2.20}$$

$$= H_a^* \times \begin{cases} \left\langle e^{+\beta \Delta V(\vec{r})} \right\rangle_j, & j \in \{\mathrm{bin}_a\} \\ 0, & \mathrm{else} \end{cases}. \tag{2.21}$$

Here, J means the number of frames defining the distribution and $\Delta V_j(\vec{r})$ is the boost potential energy for the specific conformation j. The unbiased distribution H_a Eq. (2.20) is identical to the biased distribution H_a^* Eq. (2.19) multiplied by the ensemble-averaged Boltzmann factor $\left\langle e^{+\beta\Delta V(\vec{r})}\right\rangle_j$ for simulation frames found in $j \in \{\text{bin}_a\}$ Eq. 2.21. This can be generalized for higher dimensions.

This canonical re-weighting formulation is thermodynamically exact [111, 119], but in practice $\Delta V_j(\vec{r})$ suffers from large energetic fluctuations especially for large acceleration parameters E_P, E_D [111, 114]. It is clear that already small errors in $\Delta V_j(\vec{r})$ will be massively increased by the exponential function of the Boltzmann factor.

One has to deal with two different kinds of errors: the *statistical noise error* and the *statistical mechanical sampling error* [119]. The statistical noise is amplified by the potential energy distortion and therefore has an increased contribution. It is proportional to the size of the system and acceleration [114, 119]. The second error describes the necessity that the biased sampling must also be converged to extract the correct free energy surface [119]. Its magnitude is also proportional to the size of the system. Both errors can be minimized by long and converged aMD runs.

The convenient way was trying to reduce the error of the re-weighting by approximating the exponential function of $\exp(\beta\Delta V_j(\vec{r}))$ from Eq. (2.20). Pierce et al. [117] used Maclaurin series expansion up to 10th order

$$e^{+\beta\Delta V_j(\vec{r})} = \sum_{k=0}^{\infty} \frac{\beta^k}{k!}\Delta V_j(\vec{r}) = \sum_{k=0}^{10} \frac{\beta^k}{k!}\Delta V_j(\vec{r}) + \text{rest} \tag{2.22}$$

which yielded less noisy re-weighting results. Another possibility to approximate the ensemble-averaged re-weighting factor in Eq. (2.21) is to use a cumulant expansion [120, 121]

$$\begin{aligned} \left\langle e^{+\beta\Delta V(\vec{r})}\right\rangle_j &= \exp\left(\sum_{k=1}^{\infty} \frac{\beta^k}{k!} C_{k,j}\right) \\ C_{1,j} &= \langle \Delta V(\vec{r})\rangle_j \\ C_{2,j} &= \left\langle \Delta V^2(\vec{r})\right\rangle_j - \langle \Delta V(\vec{r})\rangle_j^2 \\ &\cdots, \end{aligned} \tag{2.23}$$

with $C_{k,j}$ are the k-th cumulants. Studies revealed that using the cumulant expansion up to the second order was able to greatly suppress the energetic noise from the exponential re-weighting, particularly when the boost potential followed a Gaussian distribution [115,

118]. But Jing et al. [122] showed that the second order cumulant expansion is not a universal recipe for correct re-weighting. If aMD may sample a different energy region compared to the unbiased simulation, the second order cumulant expansion can lead to significant deviations [122].

In chapter 3, we will formulate the re-weighting for our validation tool and discuss the re-weighting scheme used in this thesis.

sMD We discussed the advantages of using aMD to speed up the simulation without necessary prior knowledge, but saw that recovering the canonical ensemble might be tricky. Therefore, Sinko et al. [113] suggested another similar acceleration by flattening the energy landscape $V(\vec{r})$ with a scaling factor $\lambda \in [0, 1]$:

$$V^*(\vec{r}) = \lambda \cdot V(\vec{r})\,, \tag{2.24}$$

whereas $\lambda = 1$ means no re-scaling (see Fig. 2.8 right). The scaling induces also a biased distribution $p^*(\vec{r})$

$$\begin{aligned} p(\vec{r}) = e^{-\beta V(\vec{r})} &\rightarrow p^*(\vec{r}) = e^{-\beta\lambda V(\vec{r})} \\ p(\vec{r}) &= p^*(\vec{r})^{1/\lambda} \end{aligned} \tag{2.25}$$

which can be re-weighted solely based on the population $p^*(\vec{r})$ of conformations instead on energetic terms. Using again the above representation of a 1D biased discrete unnormalized distribution H_a^* with uniformly distributed bins a, the unbiased distribution H_a is obtained by

$$H_a^* = \sum_{j=1}^{J} \begin{cases} 1, & j \in \{\text{bin}_a\} \\ 0, & \text{else} \end{cases} \tag{2.26}$$

$$\Rightarrow H_a = H_a^{*1/\lambda}\,. \tag{2.27}$$

Sinko et al. [113] could show that the Ramachandran plots (2D distribution of backbone dihedral angles ψ against ϕ) of sMD runs of alanine dipeptide can compete with much longer cMD simulations for $\lambda = 0.7$. They additionally recommend for typical biomolecules scaling factors $0.5 \leq \lambda \leq 0.7$ which yield minimal errors. Hence, we use a constant scaling of $\lambda = 0.7$. Again, it might be advantageous to try different scaling factors to optimize thermodynamic observables, which we do not want to focus in this study.

In chapter 3, we discuss the implementation and the re-weighting scheme used to address the presented scientific question in more detail. We will see that it is necessary to slightly modify the re-weighting for our purposes.

2.3. Studied biomolecules

The sampling problem emerges especially for flexible biomolecules, where the energy landscape is rugged and complex, and the multiple degrees of freedom limit the scope of MD [17]. Therefore it is crucial to assess the sampling quality of such flexible systems. The tool we have developed aims to tackle this problem, thus we are interested in studying widely flexible systems, first to validate the method, and second to apply the tool for a scientifically unanswered question. We consider for this purpose the following two flexible biomolecules.

2.3.1. Met-Enkephalin

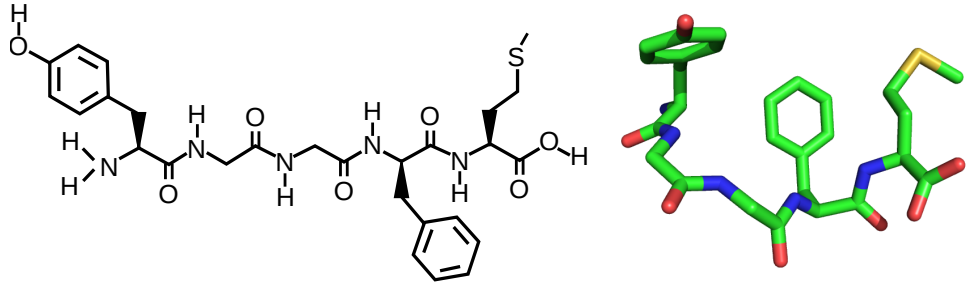

Fig. 2.9.: Chemical [123] and 3D structure of Met-Enkephalin (PDB entry 1plw [124]). Carbons are shown in green, oxygens in red, nitrogens in blue, the sulfur in yellow and hydrogens are not shown.

Enkephalin is an endogenous signaling molecule found in 1975 as a so far unknown substance in the brain [125]. It was found that Enkephalin acts as a neurotransmitter in the central nervous system [125] involved in many regulatory and physiological processes. It binds preferably to specific opioid receptors similar to morphine [126].

Hughes et al. [125] could identify and synthesize two different compositions of five amino acids, namely Met-Enkephalin and Leu-Enkephalin, which share the same sequence *YGGF-M/L* except of the fifth residue, which can either be a *Methionine* or *Leucine*. Here, we will focus on Met-Enkephalin.

In detail, Met-Enkephalin is a pentapeptide composed by 75 atoms with 24 independent backbone and side-chain dihedral angles, see Fig. 2.9. Multiple studies reveal that it

adopts massively different conformations depending on the environment [125, 127] with $\geq 10^{11}$ estimated local minima [128].

The combination of small size but still complex conformational space allows the evaluation of a flexible and complex molecule in reachable calculation times. This has made Met-Enkephalin a popular system to benchmark different molecular methods like new sampling algorithms [107, 129], molecular model validation [130, 131] or analysis techniques of molecular sampling [132, 133].

The biomolecule Met-Enkephalin is therefore an ideal candidate to evaluate our tool for the assessment of the sampling quality of molecular dynamics simulations, which essentially aims to give insight into the difficulties of sampling flexible systems.

2.3.2. V3

The second molecule, which will be studied, is the third variable loop V3 of the envelope protein gp120 of the Human Immunodeficiency Virus 1 (HIV-1). The loop is closed by a disulfide bridge between the two terminal *Cysteines* and is very flexible [134], whereas the sequence contains 31-39 amino acids and is highly variable [135]. To understand and motivate the choice to investigate this molecule, we will first give a brief overview about HIV and the replication cycle, focusing on the host entry process, where V3 is involved. Afterwards, V3 is characterized in full detail, highlighting the problems when investigating such a complex and flexible protein and showing the necessity of assessing the sampling quality.

HIV - history and structure: HIV was first detected in the 1980s, when the virus could be isolated [138, 139]. It causes the Acquired Immunodeficiency Syndrome (AIDS). It occurs in two types, HIV-1 and HIV-2, which are assumed to be evolved from the Simian Immunodeficiency Virus (SIV) infecting non-human primates [140]. According to the UNAIDS report from 2016 [141], there are about 36.7 million people globally living with HIV, where type 1 has spread more significantly than HIV-2. It is of global interest, to investigate the physicochemical properties of the virus and treat the disease.

The structure of an HI virion, schematically represented in Fig. 2.10 (left), is spherical with a diameter of around 120nm [142]. It is composed by three different regions:

1. The core region, enclosed by the capsid, contains the viral genome stored in two single strands of Ribonucleic Acids (RNA), together with important viral enzymes needed for replication, namely *reverse transcriptase, integrase* and *protease*, reviewed in Ref. [143]. The capsid is built by around 2000 capsid-proteins p24.

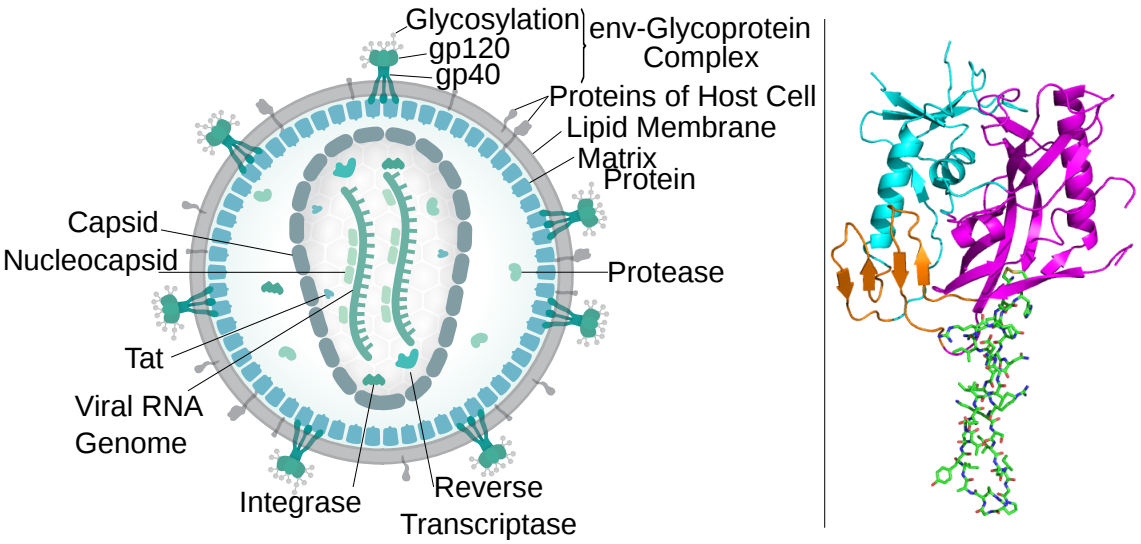

Fig. 2.10.: HI virion [136] and 3D structure of V3 from PDB entry 2qad [16] (right). Left: The fonts of the HI virion scheme are modified manually. Right: The V3-loop is shown in sticks representation, coloring carbons, oxygens, sulfurs and nitrogens in green, red, yellow and blue, respectively, hydrogens are not shown. The inner domain, outer domain and bridging sheets of HIV-1 gp120 are shown in cartoon illustration in cyan, magenta and orange, respectively.

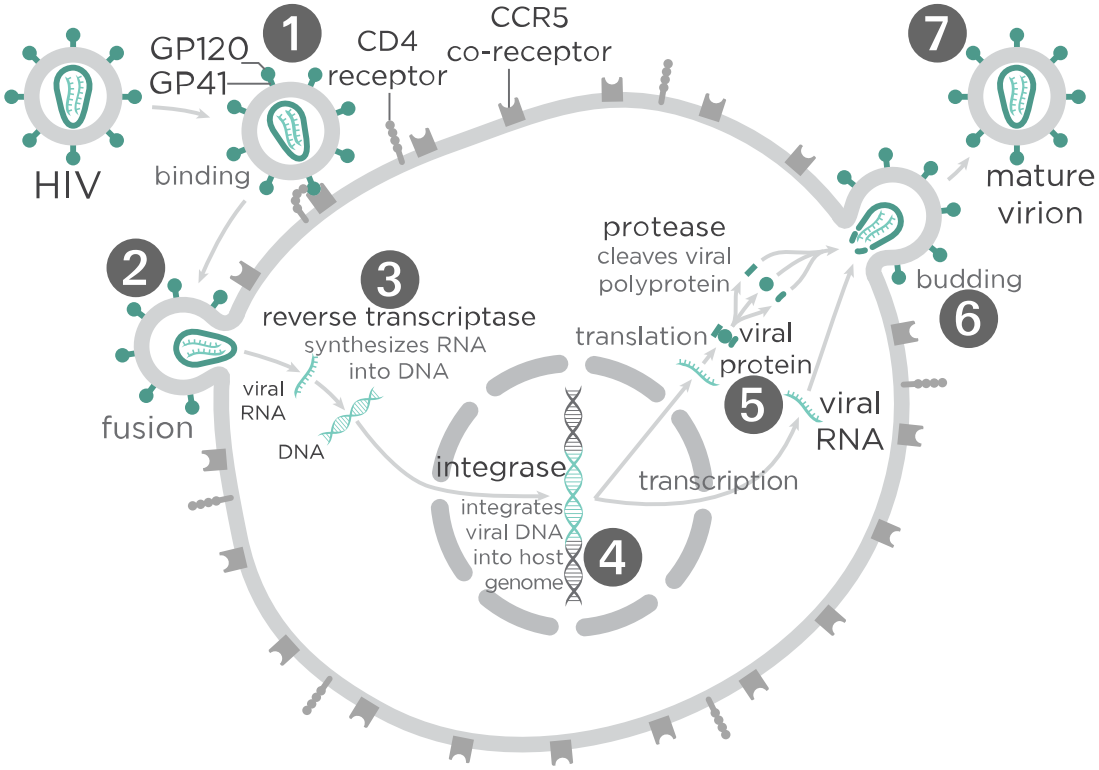

Fig. 2.11.: HIV replication cycle [137]. Fonts and numbers are modified manually.

2. The inner layer between the core and the envelope is formed by units of matrix proteins which stabilizes the envelope protein complexes.

3. The outer area, which is called viral envelope, is composed by a lipid bilayer membrane formed mainly by proteins extracted from the host membrane during replication (see Fig. 2.11 step 6). The host entry function is determined by few viral envelope spikes made of (1) three glycoproteins gp120 forming the exterior part and is heavily glycosylated (parts of gp120 with V3 are illustrated in Fig. 2.10, right) and (2) three glycoproteins gp41 anchoring the structure to the interior of the virion [144–147].

HIV - replication cycle: The replication cycle is schematically represented in Fig. 2.11. The first step is the binding and cell entry, which is mediated by the envelope protein Env involving the third variable loop and will be described in more detail in the next paragraph to outline the function of V3. After the virus-host fusion, the viral capsid content is released into the interior of the infected cell (Fig. 2.11, step 2). The enzyme *reverse transcriptase* translates the single-stranded viral RNA into DNA, which is very error-prone, generating various mutants of the virus (Fig. 2.11, step 3). Subsequently, the next enzyme (*integrase*) integrates the viral DNA into the host genome in the nucleus (Fig. 2.11, step 4) (briefly reviewed in Ref. [148]). The cellular machinery is used to transcript the proviral DNA into RNA, which forms new copies of the virus genome amongst messenger RNA (mRNA). The latter produces first regulatory proteins to support new virus production and the diffusion out of the nucleus (Fig. 2.11, step 5). Second, it produces precursor structure proteins, which are forming together with the viral RNA new immature virus particles after diffusion to the cell membrane [142, 143, 149, 150]. This immature virion starts to bud from the host cell (Fig. 2.11, step 6). Finally, the precursor proteins are cleaved by the viral *protease* into their mature units resulting in a functional HI virion (Fig. 2.11, step 7) [142, 143].

HIV - host entry and tropism: The entry of HIV is driven by the envelope protein Env targeting the Cluster of Differentiation 4 (CD4) receptor of T-cells as well as macrophages [152]. Env is composed by a trimeric formation of three copies of glycoproteins gp120 (N-terminal) and three glycoproteins gp41 (C-terminal), illustrated in Fig. 2.12. gp41 forms the transmembrane of the HI virion, whereas gp120 is the non-covalently bound exterior. The latter is composed by a bridging β sheet, one inner and one outer domain containing five conserved regions (C1-C5) which form the binding sites

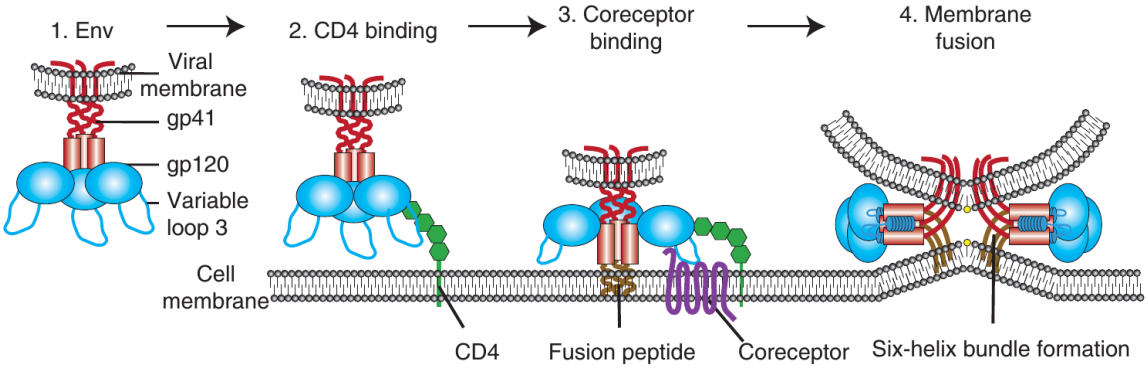

Fig. 2.12.: Schematic illustration of the HIV cell entry. Copyright 2012 from Ref. [151], reused by permission of Cold Spring Harbor Laboratory Press.

to gp41 and five variable surface exposed loops (V1-V5) [15, 16, 153].

The cell entry is separated into many complicated steps (see Fig. 2.12) starting with the recognition and binding of gp120 and the leukocyte glycoprotein CD4 [154, 155]. During this attachment, gp120 undergoes several conformational changes, fusing parts of CD4 and gp120 and bringing both cell membranes close to each other [156–158]. It is assumed that this conformational changes lead to an exposure of the chemokine co-receptor binding site [156, 159–161], which binds to the C-C Chemokine Receptor 5 (CCR5) or C-X-C Chemokine Receptor 4 (CXCR4) [162–165]. One supposes that V3 interacts with the Extracellular Loop 2 (ECL2) of one of the co-receptors, whereas the bridging β sheet interacts with the N-terminal part [15, 16, 159, 166, 167]. Finally, this co-receptor binding results in further conformational changes of the full Env protein, leading to an exposure of the previously inaccessible gp41 regulating the virus-host membrane fusion [152, 168].

The co-receptor binding was identified as a crucial step in the entry process leading to the viral phenotype classification by its tropism [169]: The virus is categorized to be either R5-, X4- or dual tropic, depending if HIV binds to CCR5, CXCR4 or is even capable to bind to both co-receptors. It is assumed that the V3-loop is one major determinant for the co-receptor selection and binding [12–14] acting like a hook to bind to the co-receptor [15]. This makes it highly interesting to study and understand the underlying physicochemical processes of V3.

V3 - state of the art: The field of studying V3 can be split into two areas: First, the investigation of the underlying physicochemical processes during the HIV binding, second, prediction of the tropism. The first field reaches from experiments [15, 16, 134, 166, 170, 171] to theoretical studies involving also molecular dynamics simulations [97, 165, 172–174]. The second field uses mainly sequence information [175–178] to predict

the co-receptor selection, but there are also predictors incorporating structural information [179–181], which makes it crucial to have adequate template structures of V3, ideally in many conformations.

In the past, it was a long time not possible to obtain a complete crystallized 3D structure of V3 attached to gp120 and/or to the co-receptor due to its notorious flexibility [134]. There are some studies like Vranken et al. [182], where V3 were solely investigated by NMR measures in water solution (see Fig. 2.13 (c)) giving first ideas about the conformational spread and flexibility of V3. Kwong and co-workers were able to crystallize gp120 and V3 together in complex with CD4 and an antibody in 2005 [15] and 2007 [16] with different V3 sequences (see Fig. 2.13 (a)-(b)). Commonly, 3D structures of biomolecules are stored in the Protein Data Bank (PDB) (`www.rcsb.org` [183]). Interestingly, both 3D structures of V3 show completely different conformations. But there are no further clear experimental results, which can describe and explain the conformational changing process upon binding of HIV and the host cell.

Here, MD simulations might be a possibility to resolve this problem and shed light into which unique conformations are sampled since this will determine the specific interactions with its receptors. It could be found that V3 moves more or less independent and uncorrelated to the movement of the gp120 core [97, 184], thus it might be reasonable to simulate only V3 as part of the conformational analysis. But the results are disillusioning: It was not possible to detect exhaustively relevant conformations of V3 [173]. Even worse, the sampling of V3 were not validated and proven to be converged and no group used different starting conformations to test, whether both simulations produce the same result.

It is therefore highly interesting, to investigate the sampling of V3 with focus on consistency and convergence.

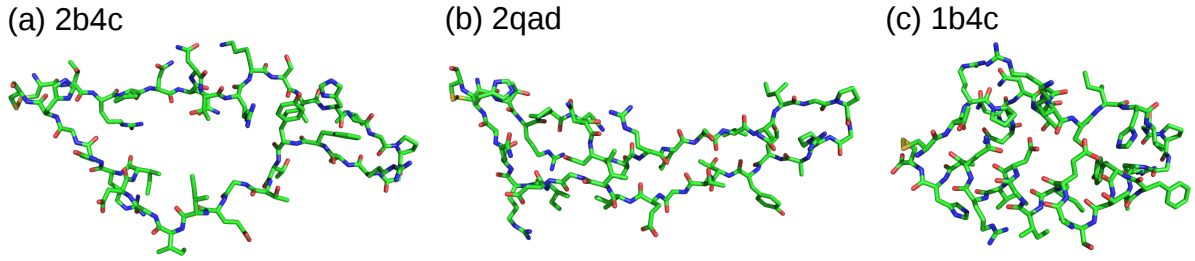

Fig. 2.13.: 3D structures of V3 from HIV-1 gp120. (a) PDB entry 2b4c [15], (b) PDB entry 2qad [16] and (c) PDB entry 1ce4 [182]. They are shown in sticks representation, coloring carbons, oxygens, sulfurs and nitrogens in green, red, yellow and blue, respectively, hydrogens are not shown

2.4. Generating different starting structures for MD

To validate a conformational sampling ensuring that the resulting thermodynamic observables are correct, multiple independent simulations must sample the same conformational space. We learned in section 2.2 that it is recommended not only generating velocity induced independent trajectories (VIIT), as it is commonly done, but also solvation induced independent trajectories (SIIT) and conformation induced independent trajectories (CIIT) [89]. SIIT can be obtained using for all simulations a different placing of water molecules in the same 3D periodic box. In the present study, this is obtained implicitly in the multistage preparation protocol described in section 4.1.3 by relaxing the water molecules around the restrained protein with different velocity seeds. Obtaining the CIIT is a hard task having a massively flexible system, where few or only one crystal structure exist. Additionally, if frequent mutations occur in the sequence, there might not be even one experimentally derived 3D structure for this special sequence. Proper starting structures for a MD simulation can then be obtained by homology modeling. We will use homology modeling to obtain starting structures for same V3 sequences.

Homology modeling Homology modeling, also known as comparative or knowledge-based modeling, describes a method to obtain a model at atomic resolution of an unknown 3D structure (target) from its amino acid sequence on the basis of one or multiple experimentally derived structures (templates) of homologous proteins (a workflow is shown in Fig. 2.14) [185, 188–190]. This is possible, because the structure of homologous proteins is more conserved than its varying sequence equivalents [190, 191].

The modeling is a multi step process, starting with the identification of relevant template structures from protein databases, using for instance BLAST [192]. The necessary criterion here is the sequence identity, which should be at least $> 25\%$ for longer proteins with more than 100 amino acids, and at least $> 30\%$ for smaller lengths [189–191].

The next step is the alignment of the sequence of unknown structure to the template structures, which is usually obtained using BLASTp [192]. Additionally, it is possible for a multi-template modeling to use the information of a structure alignment beforehand [186]. If multiple structure templates are very different with large deviations, using the structure information might result in better models [186]. A correct alignment is crucial, especially for conserved regions, where only one alignment mismatch can result in a residue being wrongly oriented to the protein interior and not to the exterior part [190]. Additionally, further information like active sites, binding pockets or constraint regions should be taken into account [185, 188, 189].

Subsequently, the target structure is built, where the coordinates of atoms in conserved regions are copied matching the alignment, and the backbone atoms are joined fulfilling the requirements for correct bonds and angles of the side-chains [190]. The procedure is different for loops. Insertions and deletions will be annealed to the core structure by local minimizations [190]. The sequence variability and structure flexibility make it difficult to predict the most correct structure regardless of a good alignment. Loops are therefore modeled by optimizing the energy function in their environment [187, 193]. The obtained model is improved further with a global energy minimization [187, 190]

Finally, the generated target models are evaluated checking different parameters, like correct bond-lengths, -angles, backbone torsion angles or non-bonded contacts [187]. These are important for rigid proteins. Additionally, the models are evaluated using statistical potentials [187]. Commonly used scores are the Discrete Optimized Protein Energy (DOPE) [194] based on a probability density function approach, the z-score derived from the DOPE estimate [187] and the GA341 score [195, 196] using information of the z-score, target-template sequence identity and structural compactness. The first two scores evaluate the models as follows: the lower their value, the better the model. The score GA341 ranges from 0.0 (wrong model) to 1.0 (native-like model). The advantage of the z-score is that it gives the possibility to compare different proteins and/or alignments.

We will apply the homology modeling program MODELLER v9.13 [187] to generate two different starting structures for V3. The details are shown in subsection 4.1.2.

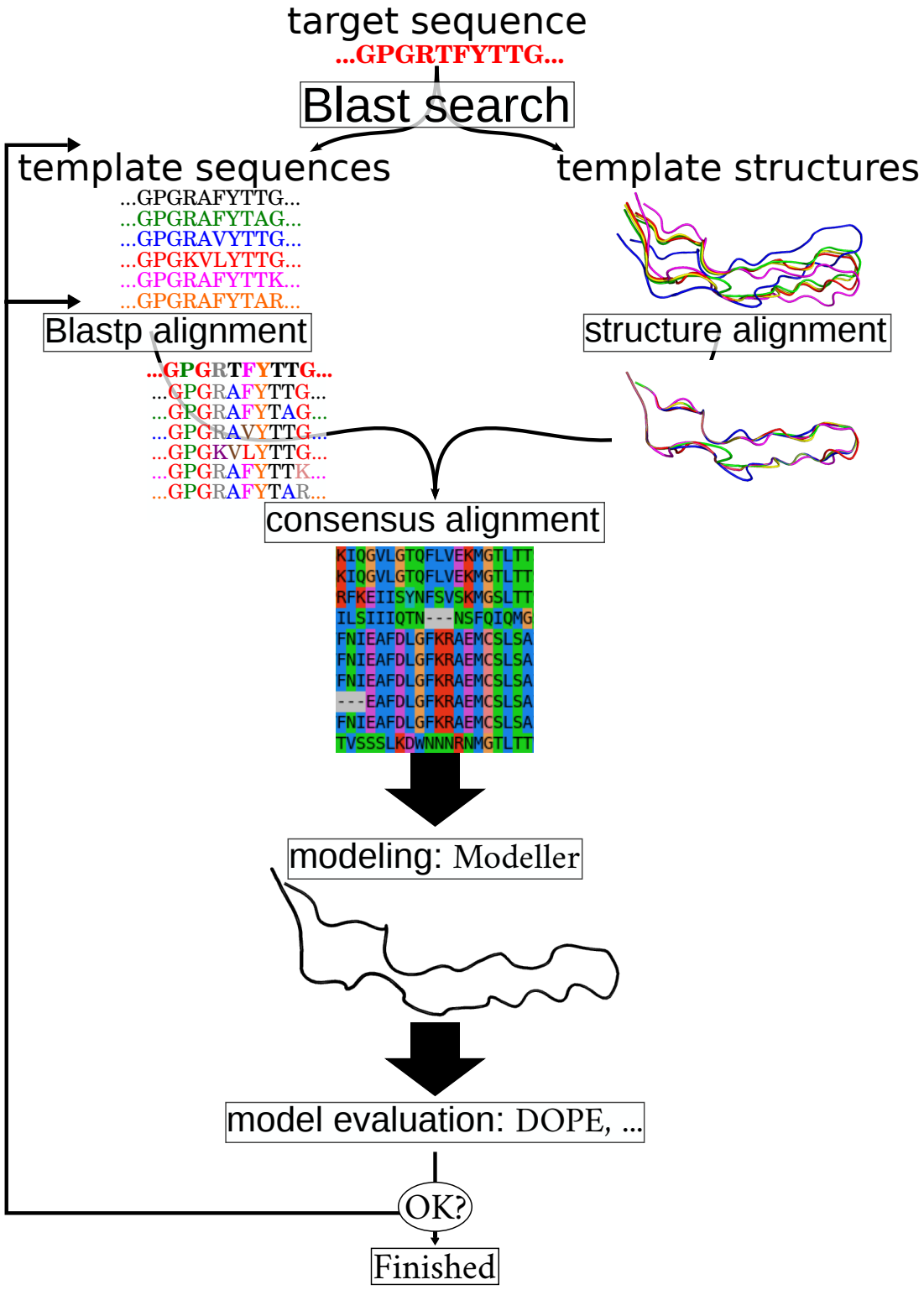

Fig. 2.14.: Workflow of homology modeling. It is not necessary to use multiple templates or incorporate the structure alignment. After generating (several) models, they have to be properly evaluated. Otherwise one has to start again using other templates or enhance the alignment. The scheme is generated manually following Refs. [185–187].

3. Tool - PySamplingQuality

In this chapter, we will introduce our tool *PySamplingQuality.py* which is designed to assess the sampling quality of molecular dynamics simulations of flexible systems using a multi-trajectory approach. In the first section, we start with the conceptual definition of a complete sampling on the level of the potential landscape. Next, we make the transition to our approach, defining a threshold parameter r and corresponding events e_r. The next two sections contain the classifiers for the sampling quality, a self-consistency measure *overlap* and an effective clustering, yielding information about the size of the conformational space. Finally, we explicitly show the workflow, introduce the modules and discuss the usage of the tool *PySamplingQuality.py*.

We published the ideas, definitions and corresponding equations of this chapter in Ref. [37].

3.1. Idea of detecting a good sampling

In theory, one can extract correct thermodynamic averages from MD simulations, if the conformational space is exhaustively sampled including all relevant rare event transitions (see section 2.1). A simple test of convergence is to run a second or more simulations, which must then give the same results. Deviations are a strong indicator that some MD runs miss relevant parts of the conformational space.

3.1.1. Conformational approach

For an exhaustive, complete sampling, different MD trajectories of the same molecule must occupy all conformations with the same density. Low potential wells correspond to high density, high potential energy conformations are occupied with a low density. This leads to the same equilibrium probability distribution $p(\vec{r})$ (see Fig. 3.1 top) for different MD runs.

Strictly speaking, for complete sampling with simulation time $t \to \infty$, the number of identical structures at a given energy level and for a specific conformation must be identical for different trajectories, which is schematically illustrated in the lower panel of Fig. 3.1. This behavior is true for all combinations of potential energy and conformation,

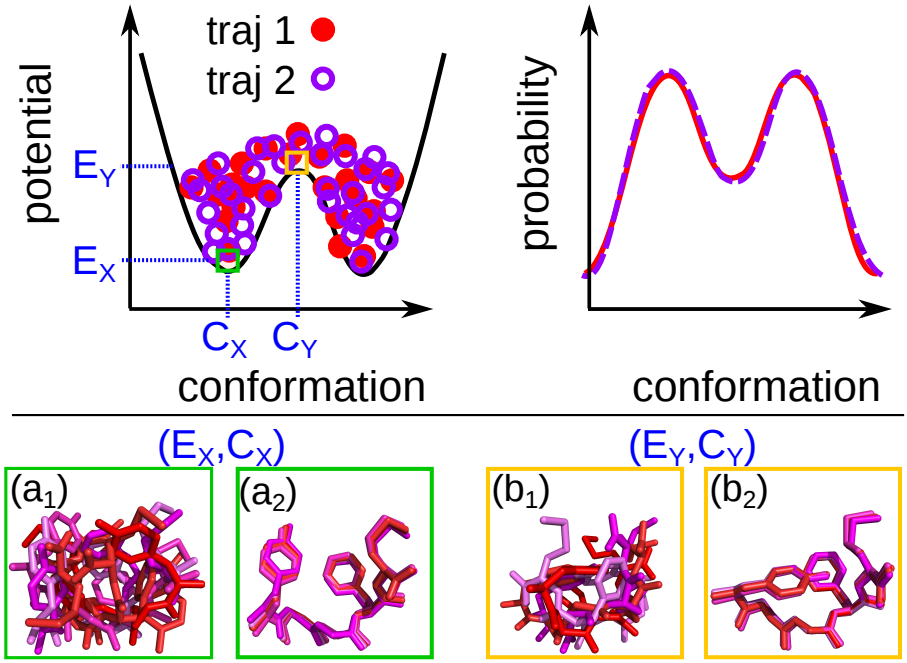

Fig. 3.1.: Schematic illustration of complete sampling of two MD simulations with the same lengths. Both trajectories reproduce the same probability distribution $p(\vec{r})$. The lower panels (a_1) and (b_1) show simulated structures at two energetic levels E_X and E_Y for different conformations C_X and C_Y, with (a_2)/(b_2) showing the corresponding alignments. The number of same/similar structures must be identical for different trajectories in (E_X, C_X) and (E_Y, C_Y).

thus one simply needs to go through every tuples of energy and conformation and count the density of identical structures for different trajectories. If the densities are always identical, we have a perfect sampling, assuming that every conformation was found. If the densities deviate between different trajectories, the sampling is not complete. One can derive a classifier of the sampling using these information.

The problem is that usually the partition function or the conformations of the system are unknown or hardly accessible (see section 2.2). We introduce therefore the trajectory overlap approach.

3.1.2. Trajectory overlap approach

In the last subsection, we learned that in different windows of potential energy and conformation tuples the number of structures must be the same for different trajectories (with the same lengths) (see Fig. 3.1). Now, we introduce a trajectory overlap approach. First, we simply use all simulated structures of all different trajectories, which correspond to the selection of different windows of energy and conformation tuples. These simu-

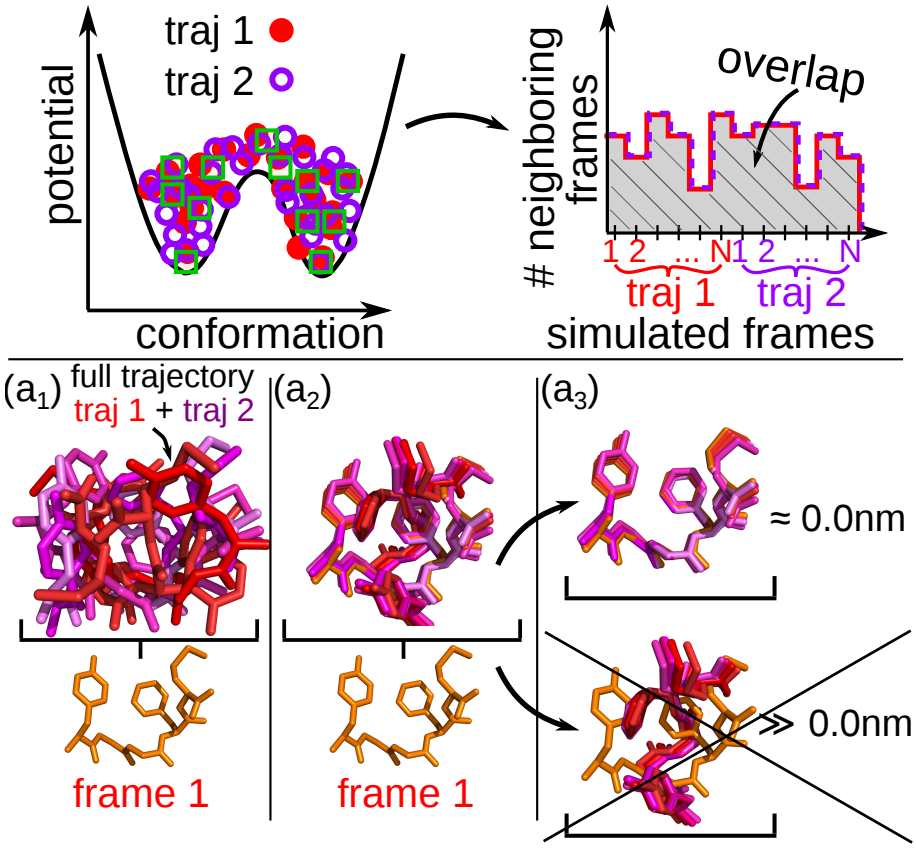

Fig. 3.2.: Trajectory overlap approach of detecting a complete sampling. For every reference frame, the number of identical/similar structures must be the same for all different trajectories with the same lengths in the case of ideal sampling.

lated structures are called reference frames. Second, we count how often we see the same conformation in independent trajectories with respect to these reference frames. Again, assuming complete sampling with $t \to \infty$, for a specific reference frame the number of identical structures (neighboring frames) to this particular reference must be the same for different trajectories with the same lengths (see Fig. 3.2).

All structures in all trajectories are superimposed and aligned to the specific reference frame (see Fig. 3.2 (a_1) and (a_2)). Then the number of identical structures (neighboring frames) are extracted for every trajectory separately as shown in Fig. 3.2 (a_3). These numbers as a function of different reference frames are then compared (see Fig. 3.2 top right): if the curves are identical, this results to a perfect overlap and reproducible sampling. The overlap will be introduced in full detail below in section 3.2. For now, the overlap is schematically represented as the shared area under all curves which are defined by the number of neighboring frames as a function of all simulated frames (see Fig. 3.2 top right). As reference frames, we use all simulated structures of all different MD simulations.

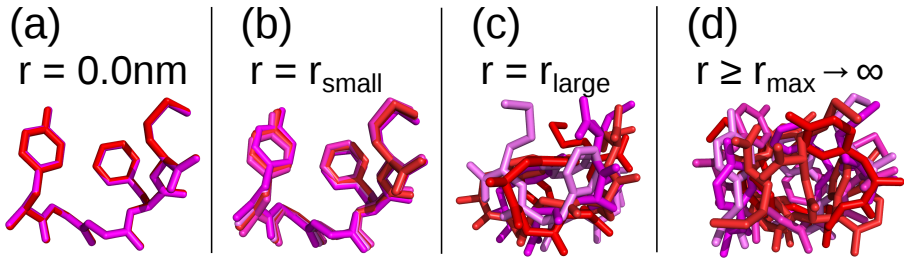

Fig. 3.3.: The effect of the threshold r. If the threshold r is equal to zero, conformations must be identical to be considered the same (a). For a small threshold $r = r_{\text{small}}$ (b), small deviations are tolerated to consider different structures to be identical. For larger r, the criterion for a same conformation is more tolerant (c), whereas for $r \to \infty$, every trajectory is assumed to come from the same conformation.

Threshold parameter r In practice, for independent trajectories of finite length of the same molecule, they will rarely produce the numerically identical conformation, even if both sample the same energy minimum. We therefore define a threshold r, where different conformations are considered the same if they are closer to each other than r and thus lying in their "r-neighborhood". The difference between two conformations a and b with N atoms and masses m_i is measured by their mass weighted root mean square deviation (RMSD) after optimal superposition [80, 82, 197]:

$$\text{RMSD}(a,b) = \sqrt{\frac{\sum_{i=1}^{N} m_i ||\vec{x}_{i,a} - \vec{x}_{i,b}||^2}{\sum_{i=1}^{N} m_i}}, \tag{3.1}$$

with positions $\vec{x}_{i,a}$ and $\vec{x}_{i,b}$ of atom i referring to the heavy atoms of the peptide backbone in the corresponding conformations a or b. These differences are stored in RMSD matrices for each pair of simulated frames for each single trajectory and each trajectory pair, which were generated with *GROMACS*, v4.6.7 [94].

The optimal superposition is in general a crucial step to obtain correct differences and to be able to compare different structures. It can be tricky and time-consuming for large and complex systems [198]. This is not an issue for the presented systems and is shortly discussed in Appendix A.

The threshold r determines if two conformations a and b are considered the same, i.e. if $\text{RMSD}(a,b) \leq r$. Effectively, the threshold r can be understood as a resolution for the overlap: The larger r, the more different structures are considered the same, and the coarser is the resolution for the measurement (see Fig. 3.3). The smaller r, the more identical must be two structures to be counted as similar. One can define a minimal value r_{min}, where at least two structures are considered the same. This also leads to the trivial relation, if r is set larger than the largest deviation between two structures r_{max} in

the trajectory, one assumes that every structure is identical, thus all densities will be the same. The threshold r will be analyzed in detail in section 4.2 focusing on the question, if there is an optimal r and how one can determine a relevant range for this parameter.

The necessary condition to decide, whether two trajectories sample the same conformational space, is to count how often we see the same conformation in independent trajectories. We therefore define each occurrence of $\mathrm{RMSD}(a, b) \leq r$ an *event* e_r.

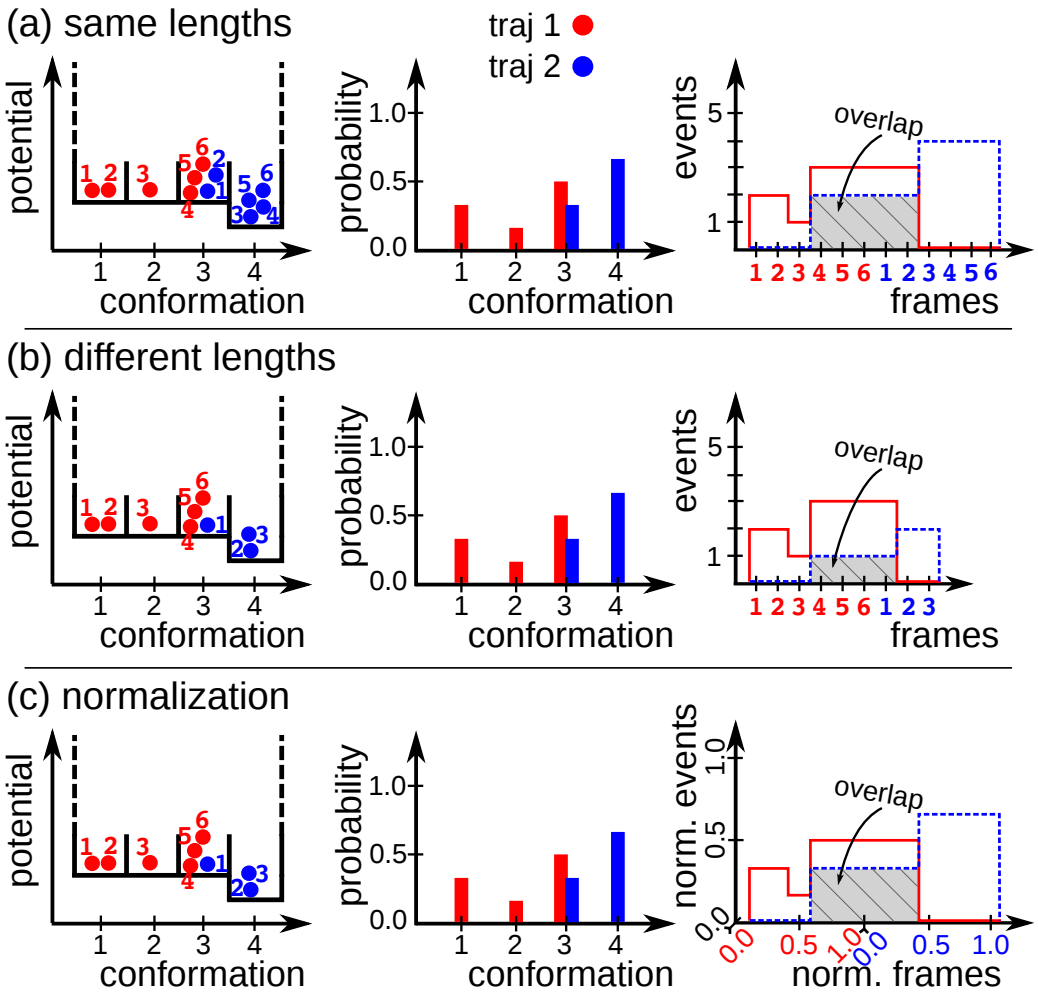

Fig. 3.4.: Definition of the event curves and their normalization. Examples for trajectories with (a) same lengths, (b) different lengths, (c) normalization applied on the events and on the frames. One can see that the overlap would lead to different results for different trajectory lengths, although the probability distributions are the same, i.e. the samplings are identical. This is repaired by the simultaneous normalization of the events and the simulated (reference) frames.

Definition of events e_r Events e_r are the number of conformations which are considered to be the same compared to the specific reference frames. They are defined for each

trajectory separately, whereas the different event curves are compared to classify the sampling. Events e_r are calculated by

$$e_{r,\kappa l} = \sum_{\alpha=1}^{n_l} H\left(r - \mathrm{RMSD}(\kappa, \alpha l)\right), \tag{3.2}$$

with

$$H(x) = \begin{cases} 1 & (x > 0) \\ 0 & (x \leq 0) \end{cases}, \tag{3.3}$$

where $e_{r,\kappa l}$ defines the number of events of trajectory l compared to the reference frame κ, $H(x)$ is the Heaviside step function, r is the threshold parameter, $\mathrm{RMSD}(\kappa, \alpha l)$ means the RMSD defined in Eq.(3.1) and n_l is the number of frames of trajectory l. In the following, indices with Greek symbols will refer to frames and Roman letters to trajectories, except r always means the threshold parameter.

The more similar different event curves e_r are, the better is the sampling. We defined in subsection 3.1.2 that we go through all reference frames of all involved trajectories to monitor the density of events for every energetic and conformational level (compare Fig. 3.2 top). But what happens, if trajectories do not have the same lengths? Then two influences must be considered which are illustrated in Fig. 3.4: First, the number of events in a certain r-neighborhood for different trajectory lengths cannot be the same although this should be the case for perfect sampling (identical $p(\vec{r})$). Thus, the event numbers must be normalized with respect to the trajectory lengths

$$\tilde{e}_{r,\kappa l} = \frac{e_{r,\kappa l}}{n_l} \quad \in [0, 1]. \tag{3.4}$$

Second, the number of reference frames are different for different trajectory lengths, which will result in different shared areas under different event curves, although the underlying probability distributions $p(\vec{r})$ are the same. Thus the (reference) frames must also be normalized, which will be done in the overlap definitions in section 3.2. Then, the event curves and the resulting overlap area are independent of the trajectory lengths and will produce the same results for the same probability distributions $p(\vec{r})$, shown in Fig. 3.4(a) and (c).

3.2. Self-consistency measure

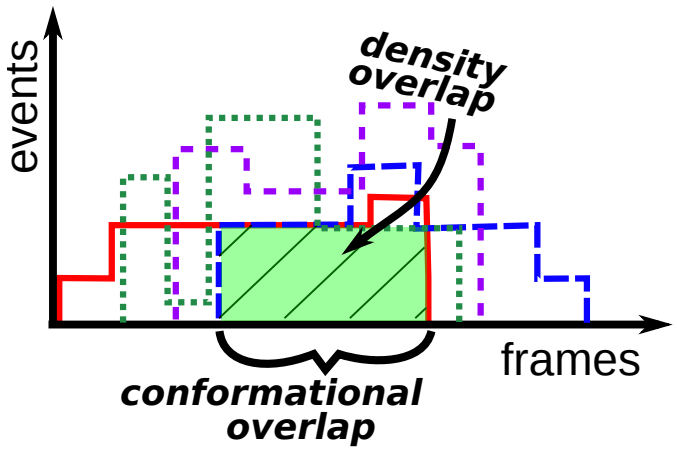

Fig. 3.5.: Schematic representation of the conformational $O_{\mathbf{conf}}$ and the density overlap $O_{\mathbf{dens}}$. The overlaps measure the area/region, where all event curves share the same area/region.

In the previous sections, we introduced events $e_{r,\kappa l}$ as a function of all (reference) frames κ for different trajectories l as density indicator, how often we see the same conformation in independent trajectories in one r-neighborhood. The sampling is now classified by two different overlap measures:

1. The *conformational* overlap O_{conf} answers the question if independent trajectories cover the same conformational space, reaching from zero, i.e. different trajectories sample completely different conformational regions, to one, where all different trajectories cover the same region.

2. The *density* overlap $O_{\text{dens}} \in [0, 1]$ quantifies the sampling criterion, if trajectories cover the same conformational space with the same probability $p(\vec{r})$.

O_{conf} is the more general and necessary criterion which allows a simple differentiation between poor and good sampling. If different MD runs do not meet themselves during the course of the simulation, one can definitely conclude that the sampling is not sufficient and longer runs are necessary. If different trajectories cover the same space, O_{dens} quantifies the quality of the sampling, whether the underlying probability distributions $p(\vec{r})$ do correspond. The two overlap measures are schematically illustrated in Fig. 3.5.

For both overlap definitions, we will always use two sets of trajectories: First, the reference trajectory set K from which the reference frames κ are taken. It can contain one or multiple trajectories, which are concatenated for the latter case. All reference

frames κ are equally important/weighted. Second, the overlap is calculated between the comparison set of trajectories L. The comparison set L can either contain two, multiple or groups of concatenated trajectories. In this study, either $K \subset L$ or $K \in L$ is true. This issue will be addressed in subsection 3.2.3 in more detail.

3.2.1. Conformational overlap O_{conf}

The conformational overlap O_{conf} gives the information, how many reference frames $\kappa \in K$ have at least one r-neighbor in each of the comparison trajectories $l \in L$, normalized by the total number of reference frames n_K. The closer O_{conf} is to one, the more the conformational space is covered by all involved trajectories L. Here, we aim to obtain an estimate whether we miss large parts of the conformational space. Thus, we do not normalize the trajectories to have the same lengths but take them as they are. For an overlap value of 0.5, we obtain the information that 50% of the frames do not cover the same conformational space, with no matter of the single trajectory lengths.

This leads to the following expression fulfilling $O_{\text{conf}} \in [0, 1]$:

$$O_{\text{conf}}(K, L; r) = \frac{1}{n_K} \sum_{\kappa \in K} H\left(\prod_{l \in L} e_{r,\kappa l}\right) \tag{3.5}$$

with $H(x)$ is the Heaviside function defined in Eq. (3.3), $e_{r,\kappa l}$ are the unnormalized events defined in Eq. (3.2) and r is the neighborhood threshold. The product within the sum of Eq. (3.5) together with the Heaviside function detects, whether all involved trajectories L have at least one occurrence in the r-neighborhood of the specific reference frame κ, otherwise it will give a zero contribution. Only if for every reference frame $\kappa \in K$ there is at least one occurrence of all different trajectories $l \in L$, one obtains $O_{\text{conf}}(K, L; r) = 1$. If the conformational overlap is close to one, the sampling may be in the regime where all conformations are found and the densities in different conformations are sampled toward a converged equilibrium. Then, it is necessary to take the probability density functions $p(\vec{r})$ into account, which is done by the density overlap.

3.2.2. Density overlap O_{dens}

The density overlap O_{dens} yields insight whether the same conformational space is covered with the same probability distributions $p(\vec{r})$ for different trajectories $l \in L$ for a given threshold r. This corresponds effectively to the shared area under multiple event curves illustrated in Fig. 3.5. Remembering that (1) every reference frame is equally important,

(2) the reference trajectories $k \in K$ have to be normalized and (3) the sampling is only complete if *all* involved trajectories have the same density/event numbers, we can quantify the sampling by

$$O_{\text{dens}}(K, L; r) = \frac{1}{N_K} \sum_{k \in K} \underbrace{\frac{1}{n_k} \sum_{\kappa \in k} \frac{\min\{\tilde{e}_{r,\kappa l} : l \in L\}}{\max\{\tilde{e}_{r,\kappa l} : l \in L\}}}_{f_{\text{dens}}(k,L;r)} . \tag{3.6}$$

The ratio between the minimal and maximal normalized event number $\tilde{e}_{r,\kappa l}$ (see Eq. (3.4)) of all trajectories $l \in L$ is the classifier for the sampling quality. This ratio is summed over all reference frames κ of one reference trajectory k and normalized by its number of reference frames n_k. This is combined in the expression $f_{\text{dens}}(k, L; r)$ which is the density overlap for only one reference trajectory k. This ensures two things: First, every reference frame κ is equally weighted, and second, every reference trajectory k is normalized to the same length of one. The latter is implicitly defined in $f_{\text{dens}}(k, L; r)$, because the total overlap for all reference trajectories K is calculated by the average of all single trajectory k measures $f_{\text{dens}}(k, L; r)$. Thus every reference trajectory k also contributes equally to $O_{\text{dens}}(K, L; r)$.

For converged trajectories L, the ratio of minimum and maximum $\tilde{e}_{r,\kappa l}$ is close to one for every individual reference frame κ, i.e. the densities and therefore the probability distributions are identical for different trajectories, and one obtains $O_{\text{dens}}(K, L; r) \to 1$. The density overlap O_{dens} will drop to zero, if the minimum to maximum ratio varies between different l for a specific threshold r for multiple reference frames κ.

This ratio defines a strict criterion for the sampling quality classification, because we use the two extremes (minimum and maximum) of densities at a certain κ. Thus, we do not overrate the overlap, but all trajectories must reproducibly give the same results. It is possible with different sets of trajectories K and L to screen through different analysis groups and for example detect outliers or combine different trajectories. This will be discussed in subsection 3.2.3.

Averaged overlap The threshold parameter r can be understood as a resolution, as discussed above. For a high resolution (small r), we are less tolerant in the event counting, because two structures κ, α must be very similar to fulfill $\text{RMSD}(\kappa, \alpha) \leq r$. For a low resolution (large r), the criterion is very tolerant, thus more different structures will be assumed to be similar and counted as an event $e_{r,\kappa l}$ in Eq. (3.2).

Theoretically, perfect sampling should be independent of the chosen threshold r and

always lead to $O_{\text{conf}} = O_{\text{dens}} = 1$. Also for $r = 0$ nm, perfect sampling ($t \to \infty$) should give the same number of identical structures for all reference frames κ for different trajectories l. In practice, this will rarely be the case, but these relations can be used as another criterion *averaged overlap* Ω_{conf} and Ω_{dens} detecting the performance of the sampling

$$\Omega_{\text{conf/dens}}(K, L) = \frac{1}{r_{\max} - r_{\min}} \int_{r_{\min}}^{r_{\max}} O_{\text{conf/dens}}(K, L; r)\, dr\,. \tag{3.7}$$

Integrating the conformational or density overlap (see Eqs. (3.5)-(3.6)) as a function of r between $r_{\min}$ and $r_{\max}$ (see subsection 3.1.2) and normalize the result by the maximally reachable area $r_{\max} - r_{\min}$ will lead to $\Omega_{\text{conf/dens}} \in [0, 1]$. The better and exhaustive the sampling, the faster the overlap as a function of r will converge toward one and we obtain $\Omega_{\text{conf/dens}} \to 1$.

3.2.3. Reference set K and comparison set L

The overlap measure is driven by the trajectory set K and comparison set L. Therefore, it is important to understand different choices and possibilities for these parameters.

One needs to keep in mind that K is only responsible for the reference frames κ. In principle, any arbitrary trajectory could be used, which does not need to be contained in the comparison set L. But in this work, we will always work either with $K \subset L$ or $K \in L$. On the other hand, the overlap is only calculated between trajectories defined in L.

For the references K, the choice of the trajectories will yield different aspects of the measure. For $K = L$, the overlap values will consider all frames of all trajectories. For $K \neq L$, we are investigating the overlap between L trajectories calculated only for a subset of reference trajectories. This makes a significant difference if we investigate two different types of trajectories, e.g. one converged and one unconverged trajectory or trajectories coming from different methods.

For instance, let us assume that we have two trajectories l_1 and l_2, where the first trajectory is complete and converged and the second shows incomplete sampling. This will in general lead to

$$\begin{aligned} O_{\text{conf}}(K = \{l_1\}, L = \{l_1, l_2\}; r) &< O_{\text{conf}}(K = \{l_2\}, L = \{l_1, l_2\}; r) \\ O_{\text{dens}}(K = \{l_1\}, L = \{l_1, l_2\}; r) &< O_{\text{dens}}(K = \{l_2\}, L = \{l_1, l_2\}; r)\,, \end{aligned}$$

because the unconverged trajectory l_2 "sees" in all its reference frames events from the converged trajectory l_1. This is not true for the opposite case, because the converged

trajectory l_1 explores space which is not reached by l_2, thus O_{conf} Eq. (3.5) and also the ratio in O_{dens} Eq. (3.6) will be small. It may even be possible that

$$O_{\text{conf}}(K = \{l_1\}, L = \{l_1, l_2\}; r) \approx 0$$
$$O_{\text{conf}}(K = \{l_2\}, L = \{l_1, l_2\}; r) = 1$$

is true, if the second trajectory is trapped only in few conformational states, whereas l_1 explores thousands of minima. Thus, the choice of K reveals different aspects of the analysis and allows to investigate for instance, how simulations behave coming from different algorithms. One would expect that trajectories from accelerated algorithms should also cover the space of the conventional simulations but not necessarily vice versa.

On the other hand, the comparison set L can either contain at least two, multiple or groups of concatenated trajectories.

Multiple trajectories mean $L = \{l_1, l_2, ..., l_N\}$, whereas every trajectory is treated individually in the overlap measure to extract for instance the minimum to maximum ratio. The only difference between two and multiple trajectories is that for O_{conf} all trajectories must have at least one r-neighbor for the corresponding reference frame κ and for O_{dens} the ratio between minimum to maximum takes the extremes between all submitted trajectories. The more trajectories are taken, the stricter is the overlap criterion, because every trajectory must independently satisfy a complete sampling. For instance, if all trajectories except one are trapped in the same energetic minimum and wrongly yield a large overlap value, then only one trajectory, which samples another unexplored region of the conformational space can make the difference. This means the overlap will drop toward zero and signalizes that the sampling is incomplete because a large conformational space is not covered exhaustively. One has to keep in mind that if only one trajectory behaves differently that it should not be discarded as an outlier. In contrast, it is a strong indicator that something went wrong with the sampling, because the latter MD simulation found another new physically meaningful states. Moreover, the less so called outliers are present, the worse might be the sampling, because only one or few simulations could reveal new states compared to $N - 1$ other runs.

Furthermore, it may also be interesting to investigate the overlap between different groups of concatenated trajectories, which will be called *group-overlap*. The underlying idea is either to merge different trajectories with same properties or to combine different short simulations to one super-trajectory. The first case might be advantageous to investigate the behavior of the sampling between all concatenated trajectories of conventional MD simulations and all concatenated trajectories of accelerated MD simulations.

Then one effectively enhances the simulation time assuming that the trajectories are independent. The second issue might be interesting for guided simulations which explores only a certain area of the full conformational space and are then combined to one super-trajectory. As an example, the group-overlap is given by $L = \{l_1 + l_2 + l_3, l_4 + l_5 + l_6\}$ indicating the concatenation of the first three and the last three trajectories, calculating the overlap between these two super-groups.

Note that, for the reference trajectories K, we simply go through every frame κ of all reference trajectories defined in K, which effectively always corresponds to a concatenation of all involved trajectories.

3.2.4. Re-weighting of biased potential runs

In subsection 2.2.3 we discussed the advantage of using accelerated sampling algorithms to ease the transition between large energetic barriers and thus access the full conformational space faster. We will make use of the introduced techniques aMD and sMD to generate trajectories and then investigate their overlap. This requires a proper re-weighting to recover the unbiased ensemble to be able to compare different trajectories. Otherwise the equilibrium probability distributions $p(\vec{r})$ are biased.

Analogously to the re-weighting of the distributions introduced in the theory chapter using Eqs. (2.19)-(2.21) and (2.26)-(2.27), we also have effectively an one-dimensional problem. Instead of using a disjunct binning as it was done in Refs [113, 117], here we have individual r-neighborhoods: We divide our system in n_K frames of the reference trajectory set K and monitor the presence (O_{conf} Eq. (3.5)) or the density (O_{dens} Eq. (3.6)) of events of different trajectories. This corresponds to a shifting window through every reference frame κ considering the r-neighbors as microstate estimates. So it is possible that different frames will fall simultaneously into multiple r-neighborhoods. Still for the re-weighting, we should only have to apply Eqs. (2.21) and (2.26)-(2.27) with $j = \kappa$. This means for aMD that we have to multiply each r-neighbor of the reference frame κ by the inverse Boltzmann factor $\exp(+\beta\Delta V_{\gamma l})$ to obtain the unbiased density, with β is the temperature factor and $\Delta V_{\gamma l}$ means the boost potential applied to frame γ of the aMD trajectory l. For sMD, the number of events just have to be re-scaled by an exponent of $1/\lambda$.

But it is not trivial, whether such a re-weighting will suffice for the overlap calculation, because multiple trajectories are involved in the measurement: The reference, from where we look into its r-neighborhood, and at least two trajectories for the minimum to maximum ratio determination. To resolve this central question of re-weighting, we will

apply a simple gedankenexperiment with known outcome comparing the overlap between the identical distributions of cMD and accelerated simulations.

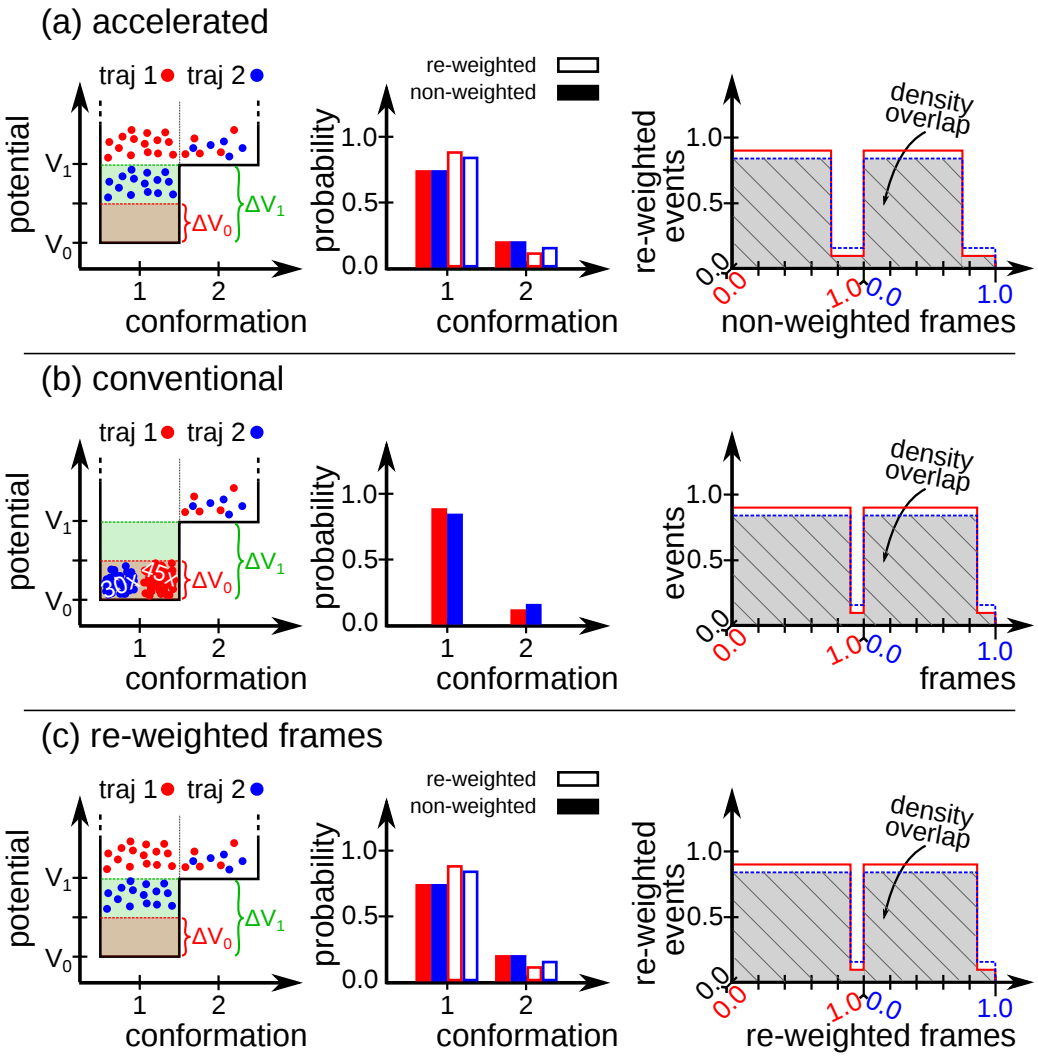

Fig. 3.6.: Correct re-weighting of the overlap measures. (a) Two trajectories with enhanced potentials ΔV_0 and ΔV_1 with only re-weighting the events. (b) Conventional, non-weighted trajectories corresponding to the re-weighted distributions of the upper panel. (c) Re-weighting the (reference) frames correct the overlap. One can see that for accelerated trajectories, one needs to re-weight the events and also the corresponding reference frames.

Gedankenexperiment to re-weight the self-consistency measure O_{dens} The central point of this gedankenexperiment is the fact that identical probability distributions $p(\vec{r})$ of different measurements must lead to the same overlap.

The gedankenexperiment is illustrated in Fig. 3.6. Let us assume a potential $V(x)$ with two conformations $x = 1, x = 2$ at the potential energies V_0, V_1 and two accelerated levels

at $V_0 + \Delta V_0$ and $V_0 + \Delta V_1 = V_1$:

$$V(x) = \begin{cases} V_0 & x = 1 \\ V_1 & x = 2 \\ \infty & \text{else} \end{cases} \tag{3.8}$$

$$V_0 = \frac{\ln(2)}{\beta}, \quad V_1 = \frac{\ln(6)}{\beta}, \quad \Delta V_0 = \frac{\ln(2)}{\beta}, \quad \Delta V_1 = \frac{\ln(3)}{\beta}. \tag{3.9}$$

Now, we sample two trajectories blue and red with 20 frames in total with two accelerated potentials ΔV_0 and ΔV_1, respectively. Furthermore, we assume to obtain 15 frames of both accelerated trajectories in conformation 1, although it does not represent the correct underlying distributions.

With these simple relations, every sampled frame on the energetic level $V_0 + \Delta V_0$ accelerated by the boost potential ΔV_0 is multiplied by $\exp(\beta \ln(2)/\beta) = 2$ and every sampled frame on the energetic level $V_0 + \Delta V_1$ by $\exp(\beta \ln(3)/\beta) = 3$ to obtain the unbiased distribution of a cMD analogue. This is illustrated in Fig. 3.6 (a) and (b) in the first two columns. But, if now only the events e_r are re-weighted according to the description above, the density overlap differs as shown in Fig. 3.6 (a) and (b) in the last column. The density overlap is only then recovered identically to cMD if also the reference frames are re-weighted accordingly, as shown in Fig. 3.6 (c).

The simple gedankenexperiment shows the necessity to appropriately re-weight the events from Eqs. (3.2)-(3.4) and also the reference frames κ of the density overlap measure Eq. (3.6). Note that the conformational overlap is not re-weighted because we are not interested in the correct densities but in the presence or absence of at least one r-neighbor of different trajectories in reference frames κ. If we know the weights for every single frame α of a trajectory l, the events are then changed to

$$e_{r,\kappa l} = \sum_{\alpha=1}^{n_l} w_{r,\alpha l} \cdot H(r - \mathrm{RMSD}(\kappa, \alpha l)), \quad \tilde{e}_{r,\kappa l} = \frac{e_{r,\kappa l}}{\sum_{\alpha=1}^{n_l} w_{r,\alpha l}} \tag{3.10}$$

and the density overlap to

$$O_{\mathrm{dens}}(K, L; r) = \frac{1}{N_K} \sum_{k \in K} \underbrace{\frac{1}{\sum_{\kappa \in k} w_{r,\kappa}} \sum_{\kappa \in k} w_{r,\kappa} \cdot \frac{\min\{\tilde{e}_{r,\kappa l} : l \in L\}}{\max\{\tilde{e}_{r,\kappa l} : l \in L\}}}_{f_{\mathrm{dens}}(k,L;r)}, \tag{3.11}$$

where the events and reference frames are both re-weighted according to Eqs. (2.19)-(2.21)

and Eqs. (2.26)-(2.27). Again, f_{dens} defines the overlap for only one reference trajectory k, $H(x)$ is the Heaviside step function Eq. (3.3), n_l is the number of frames of trajectory l, RMSD$(\kappa, \alpha l)$ defines the deviation between two structures Eq. (3.1), N_K is the number of reference trajectories K and the minimum to maximum ratio is calculated between all comparison trajectories L. The weight $w_{r,\alpha l}$ is applied on the frame α of a specific trajectory l, whereas $w_{r,\kappa}$ just re-weights all frames $\kappa \in k$, thus the trajectory index is omitted. Note that in general, the weights are defined for specific thresholds r, which will be shown below.

The introduced weights $w_{r,\alpha l}$ of a frame α of a trajectory l must correct the perturbations of the potential in aMD or sMD simulations to not overestimate the frequency of higher potential energy conformations.

cMD weights For cMD trajectories, Eqs. (3.10)-(3.11) must not distort the results to be universally applicable. This leads to the trivial weight definition for cMD simulations

$$w_{r,\alpha l}^{(\text{cMD})} = 1 \tag{3.12}$$

which yields the old definitions of Eqs. (3.2), (3.4) and (3.6).

aMD weights For conformations α of an aMD trajectory l, we have implemented three re-weighting variants:

1. Exponential re-weighting (Exp) which refers to the simple multiplication of the reciprocal Boltzmann factor for the specific frame α [111].

2. The approximation of the exponential term of the Boltzmann factor done by Maclaurin expansion (McL) up to order m which could reduce the energetic noise from the exponential term [115, 117].

3. A mean-field approximation (MF) which is inspired by the cumulant expansion up to first order (see subsection 2.2.3). We will use the averaged boost potential $\langle \Delta V(\vec{r}) \rangle_{r,\alpha}$ (see Eqs. (2.21) and (2.23)) of all r-neighbors of reference frame α to approximate the MF weight $w_{r,\alpha l}^{(\text{aMD})}$. Because different reference frames α can include the same frames in their r-neighborhood, one can extract a self-consistent mean-field approach which is defined in the following.

One can now derive the aMD weights as

$$w_{r,\alpha l}^{(\mathrm{aMD})} = \begin{cases} \exp\left(+\beta \Delta V_{\alpha l}\right) , & (\mathrm{Exp}) \\ \sum_{j=0}^{m} \frac{\beta^j}{j!} \Delta V_{\alpha l}^{j} , & (\mathrm{McL}) \\ \exp\left(+\beta \left\langle \Delta V^{(n)} \right\rangle_{r,\alpha l}\right) , & (\mathrm{MF}) \end{cases} \tag{3.13}$$

with thermodynamic temperature factor β and boost potential $\Delta V_{\alpha l}$ applied on frame j of trajectory l containing n_l frames. The n-th iteration mean-field average of the boost potential discussed above is given by

$$\left\langle \Delta V^{(n+1)} \right\rangle_{r,\alpha l} = \frac{\sum_{\gamma=1}^{n_l} \left\langle \Delta V^{(n)} \right\rangle_{r,\gamma l} \cdot H(r - \mathrm{RMSD}(\alpha, \gamma))}{\sum_{\gamma=1}^{n_l} H(r - \mathrm{RMSD}(\alpha, \gamma))} \tag{3.14}$$

with $\left\langle \Delta V^{(0)} \right\rangle_{r,\gamma l} = \Delta V_{\gamma l}$ defines the starting point of the MF iteration. The denominator is the number of frames of trajectory l in the r-neighborhood of frame α. The MF weights depend on the threshold r assuming that r-neighbors estimates the corresponding microstate, similar to the binning approach of Ref. [117].

sMD weights The weights for trajectories l with n_l frames coming from sMD runs need a different treatment. It would be possible, just to apply the relation $p(\vec{r}) = p^*(\vec{r})^{1/\lambda}$ with the scaling factor λ (compare Eq. (2.27)) to every reference frame κ to obtain the corrected number of events. The problem is that we have no knowledge about the corrected total number of events, because in general the sum of events over all reference frames κ of one trajectory l is not equal to the total number of frames n_l. Thus, we are not able to normalize the events, which is necessary for the r-neighborhood approach. We need to extract the weights $w_{r,\alpha l}$ for single frames α. For the binned distribution used in Ref. [113], single weights for all N frames falling in one bin are just the average of the number of frames re-scaled by $N^{1/\lambda}$ [113]. Since we do not have a disjunct binning but reference frames κ, where multiple frames can be in multiple r-neighborhoods, we estimate the single weights $w_{r,\alpha l}$ by averaging the re-scaled number of events in the r-neighborhood of

α:

$$w_{r,\alpha l}^{(\mathrm{sMD})} = \frac{\left[\sum_{\gamma=1}^{n_l} H(r - \mathrm{RMSD}(\alpha,\gamma))\right]^{1/\lambda}}{\sum_{\gamma=1}^{n_l} H(r - \mathrm{RMSD}(\alpha,\gamma))} \tag{3.15}$$

$$= \left[\sum_{\gamma=1}^{n_l} H(r - \mathrm{RMSD}(\alpha,\gamma))\right]^{\frac{1}{\lambda}-1} . \tag{3.16}$$

It is clear that some frames, which are in multiple r-neighborhoods, will contribute multiple times, which leads to self-consistency equations. Hence, we can now formulate another MF approach for sMD re-weighting trying to minimize the error induced that multiple frames might influence different reference frames κ by

$$w_{r,\alpha l}^{(\mathrm{sMD})} \equiv w_{r,\alpha l}^{(n+1)} = \left[\sum_{\gamma=1}^{n_l} w_{r,\gamma l}^{(n)} \cdot H(r - \mathrm{RMSD}(\alpha,\gamma))\right]^{\frac{1}{\lambda}-1} , \qquad \text{(MF)} \tag{3.17}$$

with $w_{r,\alpha l}^{(0)} = 1$ as starting point. This equation (3.17) starts with the averaged re-scaled number of events. In the next and following steps, the weights for every single frame γ are taken into account. Note that applying the weights on the specific frames changes the total number of events in a certain r-neighborhood of α, hence the denominator of the average must also contain the weighted number. This is already incorporated by the exponent $(\frac{1}{\lambda} - 1)$.

The MF approach can be iterated until convergence is reached, smoothing the edges of the neighboring windows defined by the reference frames κ, because some simulated frames can be included in r-neighborhoods of different reference frames κ. We discussed in subsection 2.2.3 that the re-weighting can be very tricky, and we do not want to focus on the validation of re-weighting procedures. We will therefore use the first iteration step of MF$^{(1)}$ for aMD and sMD which are equivalent to the first order cumulant expansion [115] and population re-weighting [113], respectively. The reason is that these procedures could already be shown to produce good results [113, 115]. The possible deviations will be discussed and investigated in section 4.7. As outlook, we will also briefly analyze the comparison between the first step MF$^{(1)}$ and the converged MF$^{(\infty)}$ results in sections 4.3 and 4.7. But we want to emphasize for a fair evaluation of the MF re-weighting, multiple acceleration parameters, systems and also an extensive study of the contribution of different r to the weights should be validated.

3.2.5. Overlap error estimates

An error estimation is necessary to validate the confidence of the results. The density overlap O_{dens} (Eq. (3.11) is defined by the average over single reference trajectory values f_{dens}.

The error of single f_{dens} can be estimated by the variation of the minimum to maximum ratios of independent reference frames κ. Remember that the ratio obtained for every reference frame is an individual estimate of the overlap value of f_{dens}: Each reference frame has to independently give a large event ratio for complete sampling. Thus, the more reference frames are used, the better is the statistic for the resulting overlap value of f_{dens}. For instance, if one half of the reference frames have a ratio of zero and the ratio of the other half is one, you will obtain $f_{\text{dens}} = 0.5$, but with a large variance compared to the same result, where all ratios are equal to 0.5. Thus, the overlap calculation is implemented in *PySamplingQuality.py* [37] to generate for every f_{dens} its standard deviation between their values.

The error of O_{dens} can be estimated by the distribution of f_{dens} of different reference trajectories. Only if every reference trajectory yields the same overlap result, there is no variation in O_{dens}.

The error of the conformational overlap O_{conf} Eq. (3.5) can be estimated in a similar way, where it is valuable to calculate $O_{\text{conf}}(k, L; r)$ for each different reference trajectory $k \in K$ and evaluate the distribution of different reference trajectories k.

This allows to plot asymmetric error bars for both overlap measures using the distributions of single reference trajectory results, where for instance the first (lower error bar), second (median) and third quartile (upper error bar) are visualized (see Fig. B.3 of Appendix B). The corresponding averaged overlaps $\Omega_{\text{conf}}, \Omega_{\text{dens}}$ are then estimated from the integrals over all lower error bar and upper error bar values.

In the following, we will use the first, second and third quartiles for error estimates unless specified otherwise.

3.2.6. Limits of $O_{\text{conf}}, O_{\text{dens}}$

So far, we argued that complete, exhaustive sampling of MD must be reproducible. Therefore, multiple simulations have to describe the same probability distributions $p(\vec{r})$, which give $O_{\text{conf}} = O_{\text{dens}} = 1$. This fulfills the criteria that independent trajectories cover the same conformational space ($O_{\text{conf}} = 1$) with the same probability $p(\vec{r})$ ($O_{\text{dens}} = 1$). But what happens, if the covered area is not the complete accessible conformational space?

The quantity, which we did not address by both overlap measures and which is hardly accessible, is the size of the sampled conformational space. With this size, one might be able to identify trajectories, which are trapped in one or few energetic minima. On the other hand, the size reached during independent runs is another criterion to classify the sampling. For instance, if the sampled conformational space has the same size for independent runs with $O_{\text{conf}} = O_{\text{dens}} = 0$, then the complete space is probably very large.

Still, the question of "unknown unknowns" is really hard to address [26]: Did we miss parts of the conformational space during MD sampling? Imagine that all trajectories are trapped in the same few conformations, which would yield large overlap values and also the same sampled size. It is clear that the more independent trajectories with independent starting conformations are used, the less likely they are all trapped in the same minimum, but this is no guarantee.

So, we try to tackle the weaker question, if there is evidence that still new areas in conformational space are discovered. As indicator, we use the convergence of conformational cluster count N_C [80] and the evaluation of the corresponding cluster distribution entropy S_C [32].

3.3. Analysing the size of the conformational space

We obtain the size of the sampled conformational space by another measure, which is a simple clustering of the sampled space. Since we store each pair of simulated frames for each single trajectory and each trajectory pair in an RMSD matrix, we run fast into memory problems by using standard clustering procedures like hierarchical clustering with complete linkage *hClust* [199, 200] or partitioning around medoids *pamk* [201, 202]. Thus, we developed an own clustering algorithm to ensure two things: (1) The clustering should be able to deal with very large RMSD matrices in appropriate time and (2) should yield the closest packed partitioning. Note, we are not interested in grouping same conformations or structures with similar properties, but only obtain a measure for the sampled size.

In the next subsections, we characterize our clustering algorithm, show the specifications and extract the cluster number N_C and cluster distribution entropy S_C as additional quantifiers for the sampling quality.

It has to be mentioned that the simple clustering is used as an additional classifier investigating whether single trajectories might sample the full conformational space or/and are trapped in some conformations. In subsection 2.2.1, we underline the necessity to properly validate clustering results. But here, we do not want to compare single con-

formations with each other by the clustering. This comparison is done with the overlap measures. Moreover, we just want to know whether trajectories discover new space or are trapped in few energetic minima. Hence, we do not necessarily have to validate the clustering with focus on the question, whether we partition similar conformations correctly together.

3.3.1. Clustering algorithm

In this subsection, the clustering algorithm is described, which allows a complete partitioning of the sampled conformational space at an approximately homogeneous resolution. To be comparable with the overlap measures, we use again the threshold parameter r as minimal distance between different cluster centroids. Then, we construct a contiguous, disjunct partitioning in chunks of RMSD-radius R with $\frac{r}{2} \lesssim R \leq r$, where next centroids are chosen to be the closest to the previous centroids. Additionally, for comparison reasons between different clusterings, we select the starting centroid according to a reference structure, which can be for example a starting conformation of a MD run.

The clustering algorithm is schematically illustrated in Fig. 3.7. Initially, the first cluster centroid C_1 is determined as the sampled conformation that has the lowest RMSD to a given reference structure (Fig. 3.7 (1)-(3)). Then, the next centroid C_2 is the closest frame outside the RMSD radius r of C_1, whereas all frames within the r-neighborhoods of C_1 and C_2 are discarded (Fig. 3.7 (4)-(7)). All other centroids C_{j+2} (with $j = 1, 2, ..., N_C - 1$ and N_C the total number of clusters found) are obtained by iterating over three steps: First, we generate an auxiliary center A_j as the coordinate average between C_{j+1} and A_{j-1} (in the first iteration $A_0 = C_1$, Fig. 3.7 (8)). Second, the next centroid C_{j+2} is the closest structure compared to A_j (Fig. 3.7 (9)). Third, frames within the r-neighborhood of C_{j+2} are discarded from the list of potential remaining centroids. The iteration is finished after no potential centroid is left (Fig. 3.7 (11)). Finally, each sampled frame is assigned to its closest cluster centroid (Fig. 3.7 (12)).

Code-wise, the centroid generation is done in three steps. The advantage is that one only needs to store one array with the RMSD values of the potentially remaining centroids, the corresponding sorted indices and keep track of the indices defining the cluster centroids. In the 1st step (Fig. 3.8), one row of the full RMSD matrix is loaded, which contains the RMSD values with respect to the first centroid C_1, which was previously generated as the closest structure to a reference. Then, the RMSD is sorted according to the threshold r, and all values smaller than r are discarded, whereas the remaining RMSD

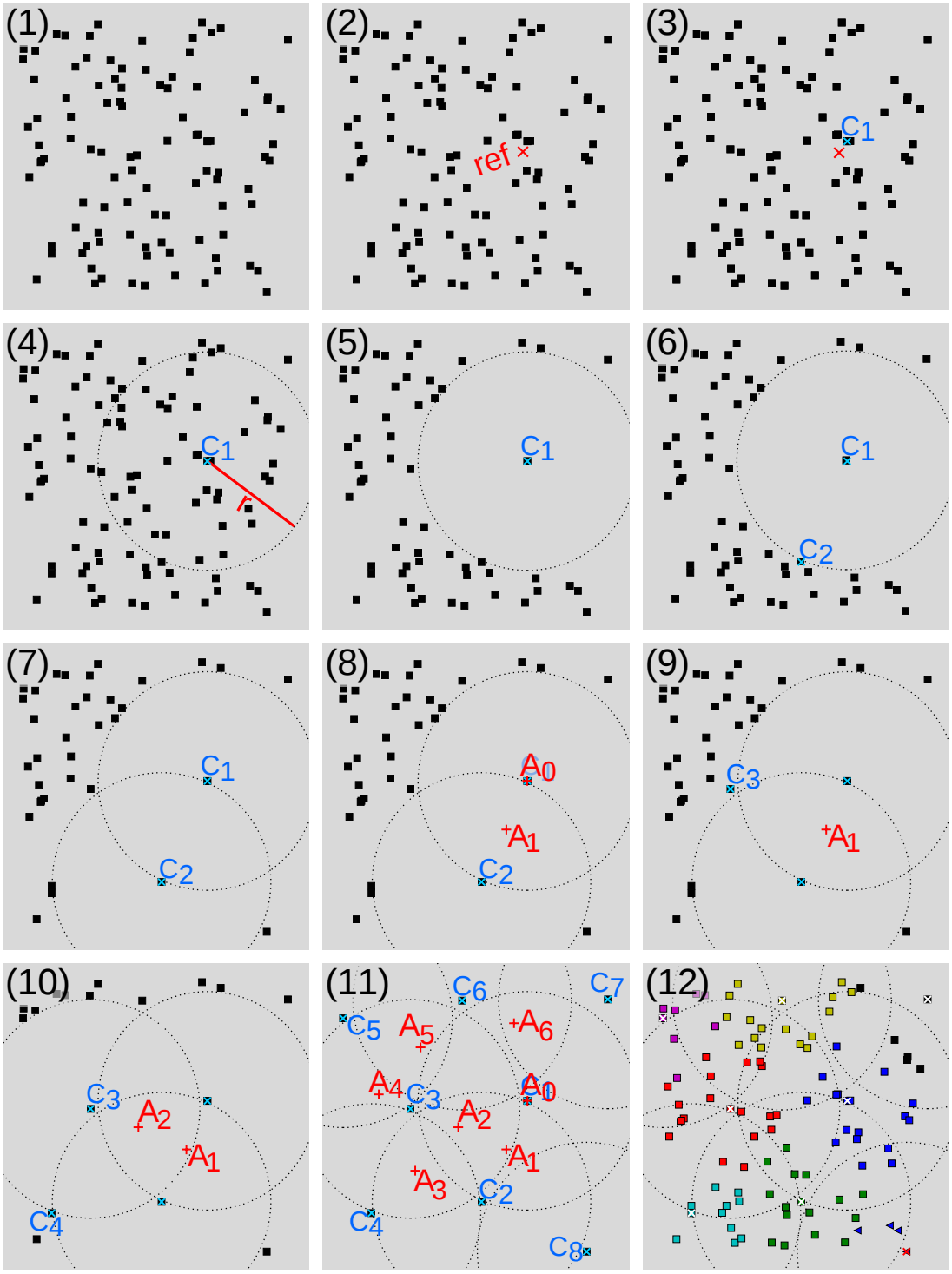

Fig. 3.7.: Step-wise representation of the effective clustering. An arbitrary 2D distribution is partitioned (1), starting with a reference frame (2), showing the formation of new clusters (3)-(7) and the definition of auxiliary centers (8)-(11). The final partitioning (12) is obtained by assigning all non-centroids to their closest centers.

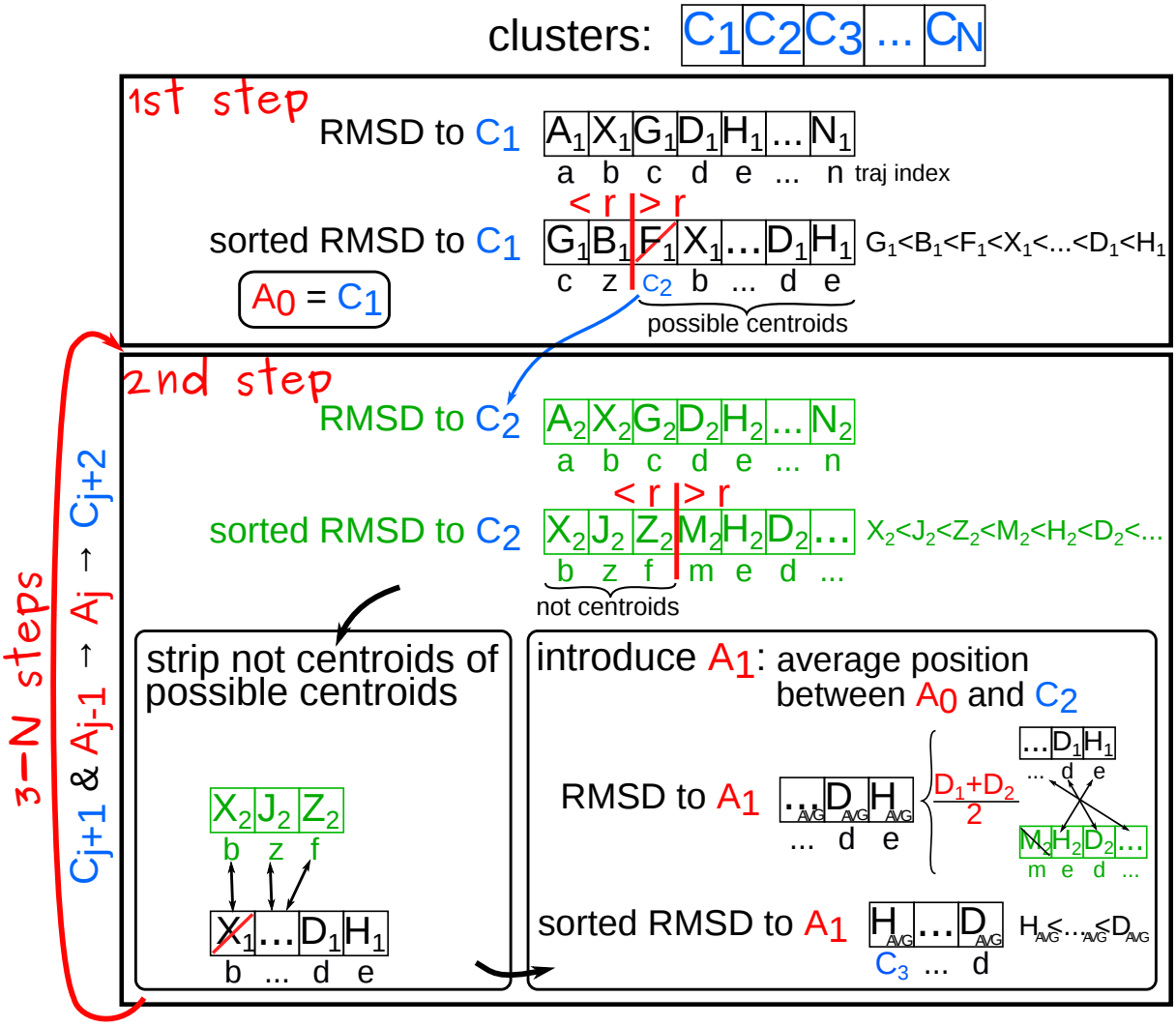

Fig. 3.8.: Schematic workflow representation of the clustering algorithm to define all centroids. One only needs to store one array containing the possible centroids, keeping track of their indices and load for every new centroid only one array which corresponds to the C_x-th line of the RMSD matrix.
1st step corresponds to steps (4)-(6) of Fig. 3.7, 2nd step corresponds to steps (7)-(8) of Fig. 3.7, and 3-N steps correspond to steps (9)-(11) of Fig. 3.7.

values are stored as possible centroids. The index with the lowest RMSD in this array is the next centroid C_2 and is also discarded from this array. In the 2nd step (Fig. 3.8), the RMSD row with respect to C_2 is loaded and sorted in ascending order. Then first (Fig. 3.8 lower left), all indices, which correspond to RMSD values smaller than r, are deleted from the possible centroids array. And second (Fig. 3.8 lower right), the RMSD values of the remaining possible centroids array together with the RMSD values of the same corresponding indices of the loaded RMSD row with respect to C_2 are averaged. This refers implicitly to a calculation of an auxiliary center, which would be the coordi-

nate average between the formerly defined centroids. Finally, in steps 3 to N, the second step is repeated until the possible centroids array is empty. Then, each sampled frame is assigned to its closest centroid as mentioned above.

The clustering yields the full clustering profile, which frame belongs to which cluster, the sizes of each cluster and the total number of found clusters N_C. Fig. 3.9 shows the necessary calculation time for hierarchical clustering with complete linkage [199, 200], partitioning around medoids *pamk* [201, 202] and our effective clustering algorithm. Our implementation outperforms both standard possibilities of data partitioning, whereas we were able to cluster structures within a half an hour with 16GB RAM which would need 80GB of memory using the standard clustering algorithms.

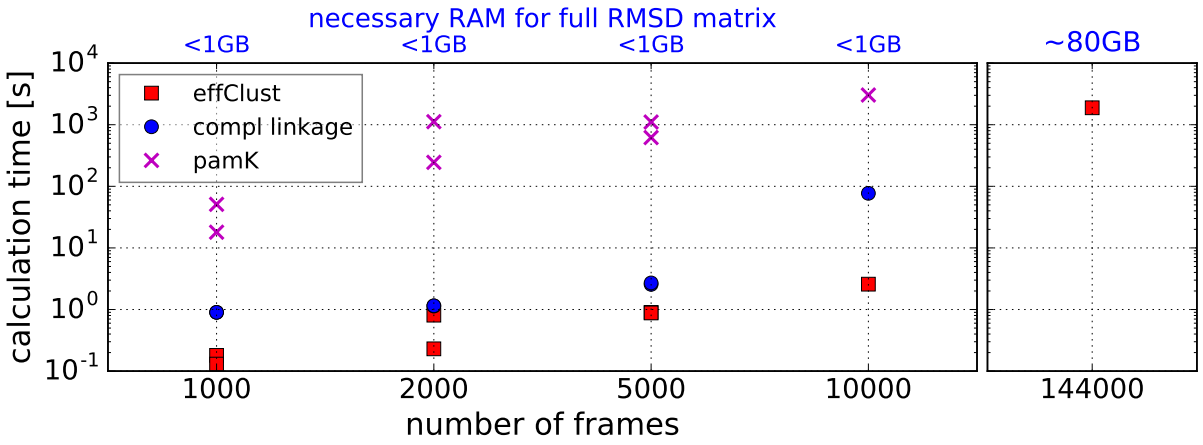

Fig. 3.9.: Benchmark comparing the effective clustering, hierarchical clustering *hClust* [199, 200] and partitioning around medoids *pamk* [201, 202]. For 1000, 2000 and 5000 structures, there are two calculations, otherwise one point gives the necessary calculation time in seconds for the amount of clustered structures, along with the size of the full RMSD matrix.

Next, we discuss different applications of the clustering, then investigating the time development over the course of the simulation and introduce the cluster distribution entropy (CDE) [32].

3.3.2. Application

The clustering can be applied to answer different questions. If one trajectory is partitioned individually, one obtains the best packed clustering for this specific trajectory and it is possible to investigate the time evolution and changes in the entropy without other perturbations getting the estimate of the sampled size by the total number of clusters. This will be referred to *local clustering*.

On the other hand, the clustering can be done once for all concatenated trajectories involved, obtaining one complete partitioning. Afterwards, one extracts which and how many clusters are reached by one specific single trajectory. The significant difference is that due to the complete partitioning one can compare the results from individual trajectories one-by-one, without deviations coming from slightly different clusterings. The total number of clusters N_C is then a good criterion to detect differences in the size of the sampled space. This will be referred to *global clustering*. The disadvantage is that the partitioning might have gaps between structures because the centroids are not constructed to have the closest distance to each other of the individual single trajectory.

It might be advantageous to compare the results from the *local* and *global* clustering to benefit from both approaches: the first to investigate single trajectories, the second to compare different trajectories.

3.3.3. Cluster number N_C and cluster distribution entropy S_C

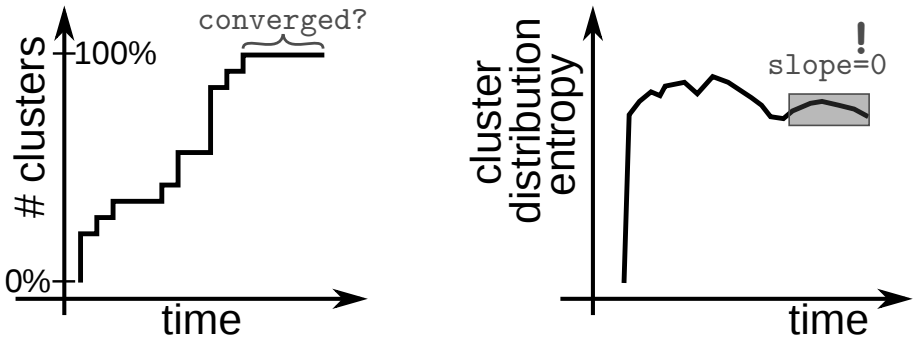

Fig. 3.10.: Schematic illustration detecting convergence by the development of the cluster number N_C (left) and cluster distribution entropy S_C (right).

To support the analysis of the sampling quality by O_{conf} and O_{dens}, we add another measures to tackle aspects of trajectory sampling convergence which are not treated by the overlap. Both can be extracted from the clustering and are discussed in the following. These two following measures were developed to treat single trajectory convergence estimates. We will introduce them in the same fashion considering the development of single trajectories but we will also enhance this picture to multiple trajectories. The relation to the overlap measures will be done afterwards.

Cluster number N_C The development of the number of clusters/centroids as a function of time $N_C(t)$ is a further indicator of the sampling convergence (see Fig. 3.10 left). We define two measures N_C^{local} and N_C^{global} to indicate, if the number of clusters for an

individual trajectory originates from the *local* or *global* clustering. For a completely converged set of trajectories, the curves $N_C^{\text{local}}(t)$ must show convergence and the final value N_C^{global} has to be the same for each trajectory. Consequently, N_C^{global} is the value specifying the size of the accessed conformational space.

The convergence of $N_C^{\text{local}}(t)$ is evaluated by the numerical derivatives dN_C/dt for the relevant last parts of the simulation time. The longer $dN_C/dt = 0$ is true, the more probable it is that no further clusters are found (see Fig. 3.10 left). These slopes are calculated by least squares regression over the last time interval of appropriately chosen sizes Δt. On the other hand, if almost every new timestep finds a new cluster and $dN_C/dt \gg 0$, one can be sure that the trajectory still explores new regions of the conformational space.

Additionally in the region of the simulation time, where the slopes are zero, one can investigate the sampling by the distribution over the found clusters. The reason is that probably all clusters are found, thus we are in the regime where the trajectory only equilibrates density between the clusters. This is treated by the cluster distribution entropy.

Cluster distribution entropy S_C If one considers the convergence estimate of a single trajectory, one should not rely solely on the size of the conformational space estimated by the number of found clusters N_C to define, whether a single trajectory could be trapped in few clusters or tends to discover new conformational space. As indicated previously, this analysis completely lacks the information of the underlying distribution. Imagine that during the simulation the trajectory quickly finds a large number N_C of different clusters, but samples 90% of the time only one conformation. This is not detected by the number of clusters.

Recently, this issue was addressed by the cluster distribution entropy S_C [32]:

$$S_C(t) = -\sum_{i=1}^{N_C(t)} p_i(t) \cdot \log\left(p_i(t)\right) \ , \tag{3.18}$$

where p_i is the probability that the i-th cluster is sampled. The simulation time-dependence t defines the current state of the simulation, i.e. the frames up to time t are assigned to $N_C(t)$ clusters and the current distribution $p_i(t)$ is calculated by the number of frames assigned to cluster i divided by the total number of frames collected up to the specific simulation time t. Sawle and Ghosh [32] argue that the curve of $S_C(t)$ should remain constant to ensure that the correct underlying probability distribution $p(\vec{r})$ is sampled. This means, the trajectory equilibrates density between the clusters, while a continuous

decreasing of the entropy signalizes a biased sampling of one energetic minimum and could therefore indicate conformational trapping.

We again define two different quantities S_C^{local} and S_C^{global} for the different underlying *local* or *global* clustering of the specific trajectory. We investigate the constancy by the numerical derivatives dS_C/dt for the last time interval of appropriately chosen sizes Δt. If the sampling of the trajectories converged, $S_C^{\text{local}}(t)$ should have horizontal slopes for constant $N_C^{\text{local}}(t)$ regions and simultaneously all converged S_C^{global} values should be the same for different trajectories (see Fig. 3.10 right).

With the defined values, we are able to detect by $dS_C/dt \ll 0$ that the sampled cluster distribution is biased toward few energetic minima. On the other hand, $dS_C/dt \gg 0$ can either mean that the distribution starts to converge toward the true conformational probability distribution, or new clusters are probable to be detected.

All in all, a necessary condition to fulfill the completeness of sampling is that the number of conformational clusters and the underlying cluster distribution entropy are converged. But, these criteria are not a guarantee because apparent convergence can also result from conformational trapping.

Robustness of the effective clustering The main purpose of the effective clustering implementation is the efficiency, handling huge RMSD matrices and get a simple partitioning to estimate the sampled conformational space, as mentioned above. The question which arises is whether the results are compatible with standard clustering approaches or whether we introduce significant differences or even artifacts. We investigate the development of N_C and S_C as a function of the simulation time t and the corresponding slopes of their linear regressions for different clustering methods to evaluate the robustness of the effective clustering. The results for an arbitrary chosen Met-Enkephalin trajectory are illustrated in Fig. 3.11. There are only minor differences between *pamk* and our effective clustering algorithm if the same amount of clusters are found. The slopes also correspond to each other. To obtain a corresponding partitioning with *hClust*, we identify clusters with a partitioning height of 0.19 nm, getting one more cluster than the effective clustering. Then, the developments of N_C and S_C give similar results, especially for the last 50 ns but have slight differences in the first 50 ns. The outcome is very similar using another trajectory also from V3, which is not shown. The results let us conclude that the results of the effective clustering are robust within the comparability to other clustering methods, since all have different criteria to partition the data.

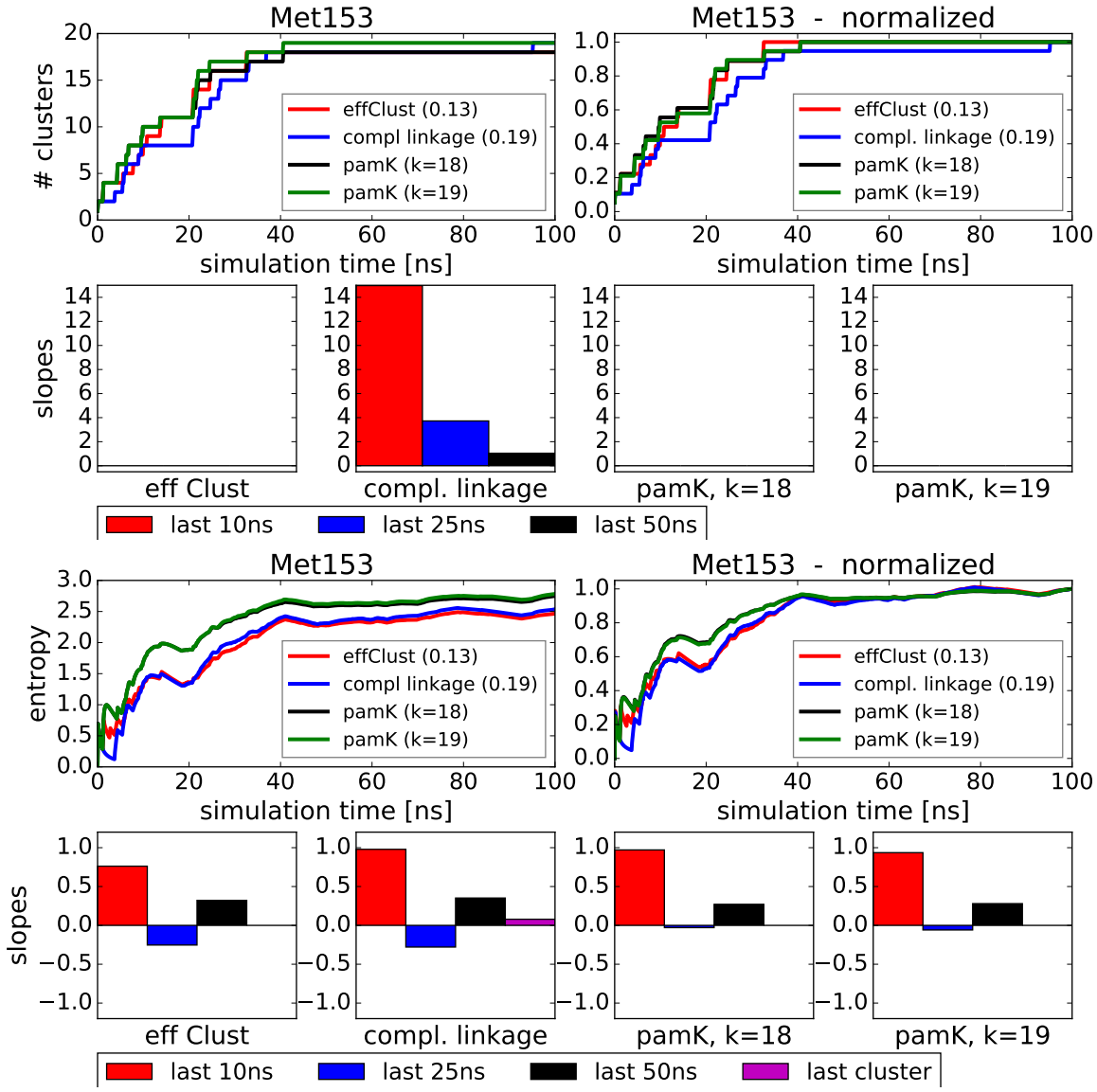

Fig. 3.11.: Comparing results of the effective clustering, *hClust* [199, 200] and *pamk* [201, 202] for N_C (top) and S_C (bottom). An arbitrary trajectory of Met-Enkephalin starting from *Met153* is chosen. The developments of N_C and S_C are shown as a function of the simulation time along with the slopes from linear regressions of the last 10, 25 and 50 ns. "last cluster" refers to the time interval after the last cluster was found. Slopes mean that the corresponding value is approximately changed by the slope value within the next 100 ns. The effective clustering was done at $r = 0.13$ nm, corresponding *hClust* at height $= 0.19$ nm and *pamk* with 18 or 19 clusters, respectively.

3.4. Combination of overlap and clustering

The overlap measures and the clustering results tackle both different aspects of the sampling quality. The first classifies the self-consistency and reproducibility of the sampling,

giving the answer whether trajectories cover the same conformational space with the same probability density distribution. The second investigates the sampled size and the underlying distributions. All four quantities $O_{\text{conf}}, O_{\text{dens}}, N_C$ and S_C must give a converged result that the sampling may be complete. Nevertheless, they are not completely independent.

One can only obtain a large $O_{\text{dens}} \approx 1$ if the conformational overlap O_{conf} is close to one. The latter will result consequently in very similar values of N_C^{global} for different trajectories. In the same way, an increasing number of clusters N_C will also increase the entropy S_C, but improved sampling at converged N_C is only detectable by S_C or the density overlap O_{dens}.

On the other hand, a converged $O_{\text{dens}} \approx 1$ will automatically yield $dS_C/dt \approx 0$ with similar S_C values. But this is not necessarily true for the opposite case, where different trajectories have the same cluster distribution entropy but may be converged in separated energy wells, i.e. low O_{dens}. This reveals also the disadvantage of using single trajectory convergence criteria, which do not give the information, whether for instance different starting conformations stay separately trapped but seem to be converged individually.

One always needs to complementary use both conditions, the overlap and the clustering, to comprehensively quantify the sampling for consistency. High values of O_{dens} together with convergence of N_C are necessary criteria for good sampling. On the other hand, for poor to moderate O_{dens}, O_{conf} and S_C yield insight if trajectories sample different conformational regions, show trapped behavior or the simulation time is just too short to equilibrate the density. Remember that the overlap measures quantify the sampling between different trajectories, thus S_C might give insight into one single trajectory, how the corresponding sampling behaves during the course of the simulation.

3.5. PySamplingQuality

All previously defined quantities ($O_{\text{conf}}, O_{\text{dens}}, N_C$ and S_C) are implemented in a package written in Python called *PySamplingQuality.py* [37] (version v05.04.17-1). It also includes the re-weighting variants defined in Eqs (3.13)-(3.17). The package allows to quantify the sampling quality for multi-trajectory experiments using molecular dynamics simulations cMD, aMD or sMD. It is freely available as source code at `https://github.com/MikeN12/PySamplingQuality`. There is also uploaded a simple tutorial to run the analysis.

It contains different modules which are grouped in three different categories: *Overlap*,

Clustering and *Visualization* (see Fig. 3.12). The necessary (and tested) versions of different programs are given in Table 3.1. There are two different possibilities to run one

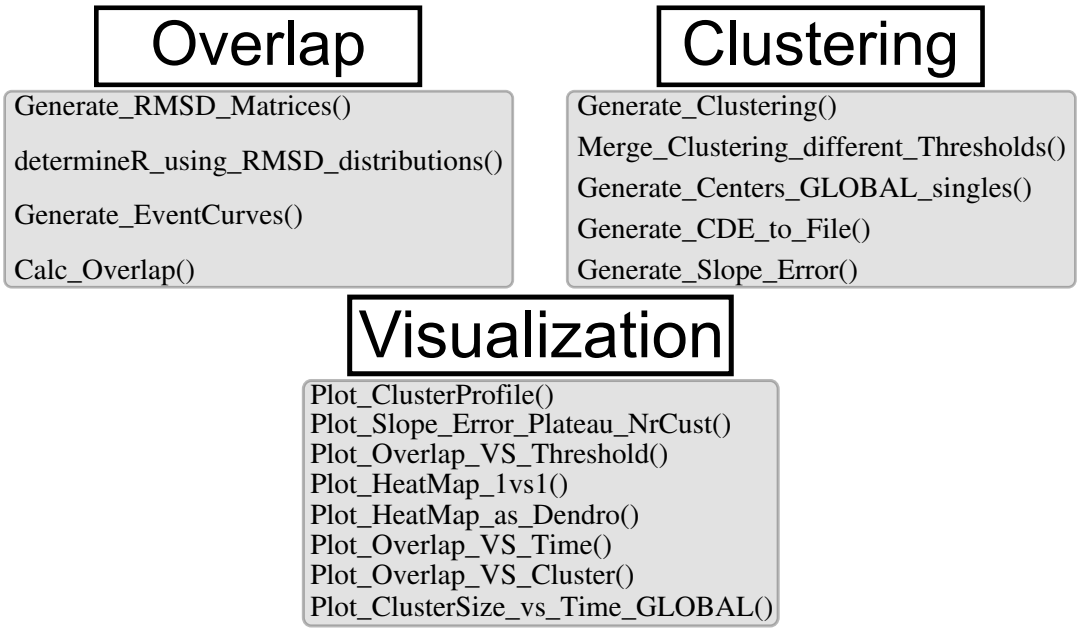

Fig. 3.12.: Modules of *PySamplingQuality.py*.

Table 3.1.: Required programs and versions to run *PySamplingQuality.py*.

program	version
Python	2.7.12 [36]
Anaconda	2.4.1 (64-bit) [203]
Matplotlib	1.5.1 [204]
scipy	0.17.0 [205]
numpy	1.10.4 [206]
Gromacs	v4.6; v5.1 [94]
Amber	AmberTools14 [46]

of the modules: Either one generates configuration files, which are then submitted to the module with all parameters

```
python PySamplingQuality.py -module GenerateIn -in MODULE -out CONFIG.in
python PySamplingQuality.py -module MODULE     -in CONFIG.in
```

or directly in *IPython* [207] or a corresponding *jupyter notebook* [208] by first importing the specific module and submit the necessary *options*:

```
from PySamplingQuality import MODULE
MODULE(options)
```

The functionality and the specific modules will briefly be discussed in Appendix C and can be accessed in more detail in the tutorial which can be found in https://github.com/MikeN12/PySamplingQuality/blob/master/PySamplingQuality_Tutorial.ipynb. Every module has its own description page called *doc-string* in Python, containing examples, descriptions and default values. A schematic workflow is illustrated in Fig. 3.13. Starting from 3D structures of a system, one has to generate two to multiple trajectories. These are then submitted into *PySamplingQuality.py*, where first the RMSD matrices can be generated. These are the standard input for the overlap and clustering measures, where the results can either be visualized independently or one can use both outcomes for comprehensive studies. All results in chapter 4 are done with our tool, showing the application and impact for quantitative assessment of MD sampling quality for flexible molecules.

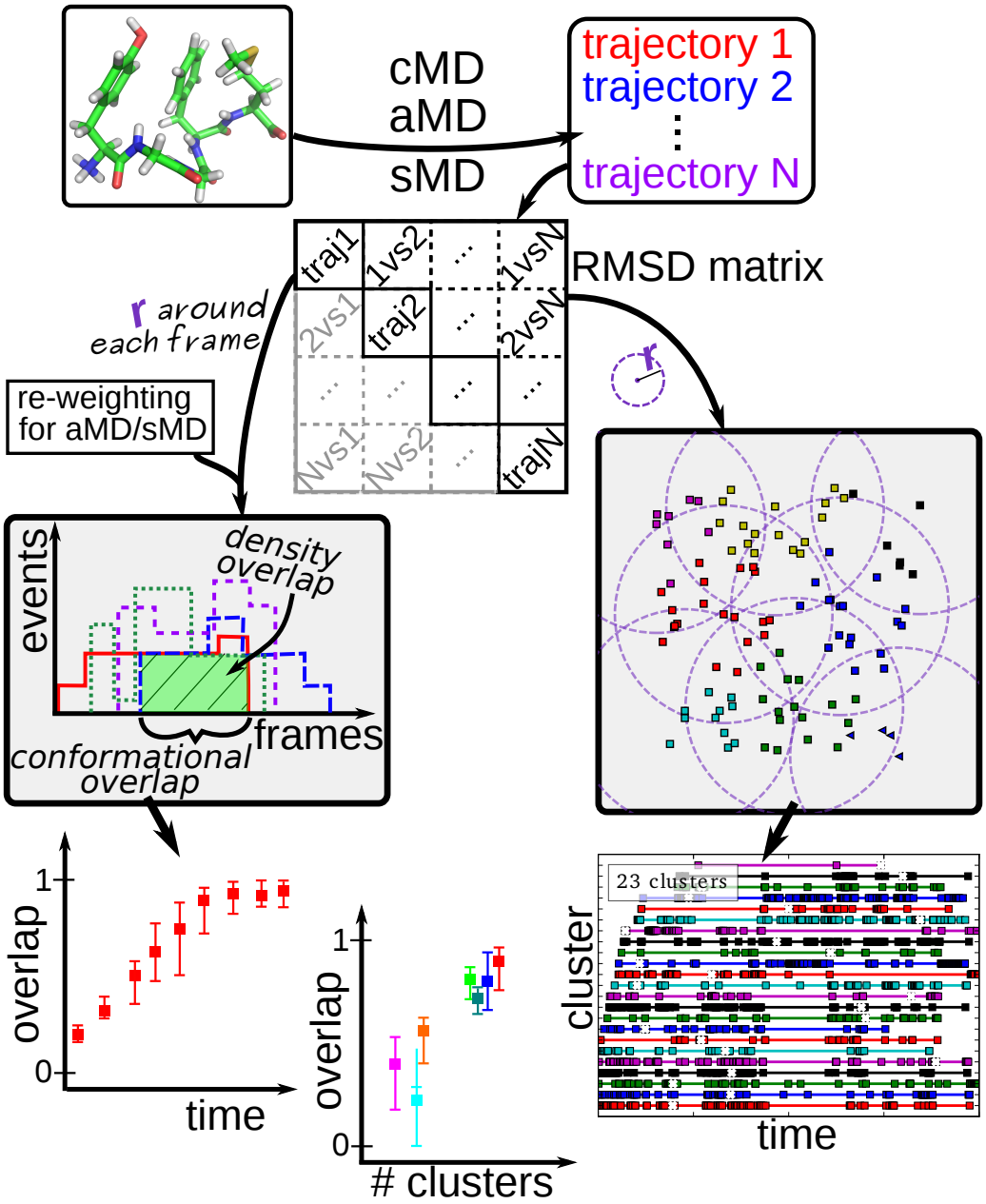

Fig. 3.13.: Workflow analyzing multi-trajectory convergence with *PySamplingQuality.py*. Multiple trajectories are translated into several partial RMSD matrices, which are used as input for the partitioning and overlap measures. The outcomes can be visualized.

4. Results and discussion

The central topic of this thesis is the quantification of the sampling quality of MD simulations for flexible biomolecules. What results can be obtained using a multi-trajectory approach and enhanced sampling techniques (aMD and sMD)? This will be answered in a comprehensive study using extensive molecular dynamics simulations with different conditions and analyses.

We start with reporting the starting structures for the two studied systems: the small pentapeptide Met-Enkephalin and the large flexible V3-loop. Then, we discuss the parameters and the setup for the different simulations and finally investigate the influence of the different starting conformations.

Furthermore, since our analysis depends on the threshold r, we ask ourselves, whether there is an optimal value for this resolution parameter.

Then, the sampling quality is assessed by the overlap analysis O_{conf} and O_{dens}, the size of the (sampled) conformational space N_C and the cluster distribution entropy S_C, also combining all criteria for different conditions. Additionally, we investigate the effects of re-weighting and enhanced sampling algorithms, discussing also the effects of the overlap on thermodynamically relevant observables.

Several results will be shown as boxplots which are defined in Appendix B.

4.1. Starting structures and setup

As discussed in section 2.2, the sampling quality from MD runs benefits from multiple independent simulations with different starting conditions. Only if the simulations are independent from the starting conditions, one can obtain a complete sampling. One way to test this issue is to use totally different starting conformations. Then, it is less probable that the corresponding trajectories coincidentally sample the same conformational space just because they are trapped in the same local minimum. Moreover, both trajectories must at least cross the energy barrier between the two starting conformations and sufficiently sample both potential wells to give a reproduced picture (see Fig. 2.7). Thus, we aim first to generate two independent starting structures for each of the two studied systems.

4.1.1. Starting structures of Met-Enkephalin

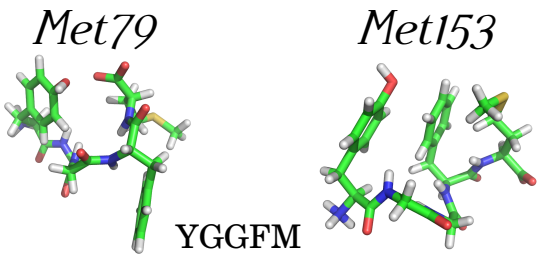

Fig. 4.1.: The two starting structures of Met-Enkephalin *Met79* (PDB entry *1plw* [124]) and *Met153* (PDB entry *1plx* [124]) with their amino acid sequence. The structures are shown in sticks representation, coloring carbons, oxygens, hydrogens and nitrogens in green, red, white and blue, respectively

The first studied system is Met-Enkephalin (see subsection 2.3.1). There are two NMR model ensembles with the PDB entries *1plw* and *1plx* [124]. Both contain 80 different models each. As starting structures, we select the two with the largest RMSD = 0.23 nm using Eq. (3.1) after optimal superposition. We call these two starting structures *Met79* and *Met153* respectively (see Fig. 4.1) [37]. The N- and C-terminal are capped with acetyl (ACE) and N-methylamine (NME) moïeties added by PyMol [209]. Uncharged ACE and NME termini are often used to cap the truncated peptide bonds at the terminal ends of a protein or peptide to help to stabilize the structure during MD simulations [210, 211]. The starting structures are not further optimized since this step will be done in the preparation stage of the MD simulation.

4.1.2. Starting structures of V3

The second studied system is the third variable loop V3 of the glycoprotein gp120 coming from HIV-1 introduced in subsection 2.3.2. We discussed the conformational flexibility and sequence variability, which made it difficult to obtain a various set of crystal structures for the same sequence. In fact, there are two crystal structures of gp120 with the full V3-loop with the PDB entries *2b4c* [15] and *2qad* [16] with completely different conformations but different sequences. To be able to investigate the MD sampling from different conformations of the same molecule, we generated starting structures from homologous modeling using MODELLER v9.13 [187]. The general workflow is described in section 2.4.

We want to emphasize that the main goal is to generate different starting conformations. In general, loop modeling is very difficult (for further reading, see for example chapter 13 of Ref. [41]) and needs a step-by-step optimization to give a good physical model. Such an optimization is done in the preparation stage of the MD simulation discussed in the

next subsection. Moreover, if the starting models are not wrong regarding the modeling scores, the more different the two starting structures are, the clearer is the message if both reproduce the same results. This is true, because the trajectories must probably overcome multiple energetic barriers, which is commonly the case for a large rugged flexible system, to sample the conformational space with the same probability density $p(\vec{r})$.

We selected a V3-loop sequence (R5-tropic) from the Los Alamos HIV database with the GenBank entry AF112548 (`http://www.hiv.lanl.gov/`) and amino acid sequence

```
CTRPNNNTRKGIHIGPGRTFYTTGEIIGDIRQAHC .
```

As templates, we considered three different 3D structures using a Blastp [192] alignment search in the protein database [183] reported in Table. 4.1: *2qad* [16], *2b4c* [15] and *1ce4* [182]. The details for these templates are given in subsection 2.3.2 and Fig. 2.13. All these templates have an adequately large sequence similarity compared to our chosen molecule, which is necessary to obtain a good homologous model.

Table 4.1.: Template specifications for V3 showing the sequence similarity, length and the E-value after Blast [192] searching. The latter describes the random background noise to find a similar score simply by chance in the protein database.

PDB entry	chain	length	similarity [%]	E-value
2qad	A	35	89	$4 \cdot 10^{-16}$
1ce4	A	35	94	$2 \cdot 10^{-18}$
2b4c	G	35	91	$5 \cdot 10^{-17}$

We decided to do two different modelings to obtain two different starting structures: First, we applied a single template modeling using the latest crystal structure *2qad* [16]. The first resulting starting structure from the modeling process is called *V3a*. The structure of *2qad* has a special narrow form compared to the other templates and we aimed to retain this shape. Second, we did a multi-template modeling considering all three different template structures. The second resulting starting structure from the modeling process is called *V3b*. The reason was to obtain a completely different starting conformation based on the structural flexibility of different templates. For both modelings we generated five different candidates and selected the two starting structures *V3a* and *V3b* with the best modeling scores.

The single template model *V3a* was obtained by a Blastp [192] alignment and the automodel function of MODELLER [187]. For the multi-template model *V3b*, we compared the scores of two different modeling stages: Again, we first obtained a Blastp [192]

alignment generating models with the automodel function. Second, we made models using additionally structure alignments before the automodel function (see the workflow in Fig. 2.14). These structure alignments superimpose the templates based on RMSD differences using the default input of the *salign* function [186] introduced in section 2.4. The latter procedure leads to better scores of the models, which were classified by the DOPE, GA341 and z-score (see section 2.4), shown in Table 4.2.

The two final starting structures *V3a* and *V3b* are shown in Fig. 4.2 [37]. The loop is closed by a disulfide bridge between the two terminal *Cysteines*, the termini are again capped with ACE and NME groups to stabilize the truncated protein and the *Histidines* were protonated on the second epsilon nitrogen $N_{\epsilon 2}$.

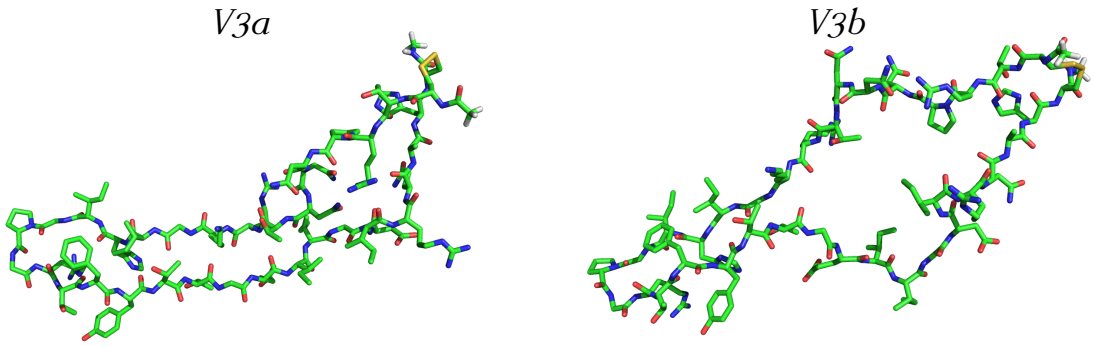

Fig. 4.2.: The two starting structures of V3 *V3a* and *V3b* with their amino acid sequence. The left structure was generated by single template modeling, the right with multiple template modeling using MODELLER v9.13 [187]. The V3-loops are shown in sticks representation, coloring carbons, oxygens, sulfurs and nitrogens in green, red, yellow and blue, respectively, hydrogens are not shown.

Table 4.2.: Modeling scores for the single-template model *V3a* and multi-template model *V3b*. Bold numbers correspond to the final models taken as starting structures. GA341 ranges from $[0, 1]$, whereas models should only be considered for values > 0.6. The DOPE and z-score marks better models the lower their score is.

	single-template *V3a*			multi-template *V3b*		
Model	DOPE	GA341	z-score	DOPE	GA341	z-score
1	-1336.309	0.659	0.550	-1021.500	0.979	1.187
2	-1313.284	0.728	0.596	-1080.004	0.989	1.068
3	-1202.314	0.455	0.821	-1061.942	0.977	1.105
4	**-1343.775**	**0.836**	**0.535**	-1192.885	0.993	0.840
5	-1260.741	0.868	0.703	**-1223.715**	**0.954**	**0.778**

4.1.3. Simulation setup

All preparations and simulations of Met-Enkephalin and V3 were done with the *AMBER*14 [46] software and the ff99SB-ILDN force field [52], whereas the production phase was accelerated using GPUs with the CUDA implementation [212]. All theoretical details are introduced in chapter 2. For the simulation steps, lengths of bonds involving hydrogen atoms were constrained with the SHAKE algorithm [74], allowing to use an integration step of 2 fs. A 1 nm cut-off was applied to the non-bonded interactions, whereas long-range electrostatics were computed with PME [61]. Using the *AMBER*14 program TLEAP, hydrogens were added to the experimental structures according to the ff99SB-ILDN force field. We used a multistage preparation protocol comparable to Ref. [213] to refine the homologous starting structures on the one hand and also optimize possible unfavored contacts introduced in the crystallization process of the experimental structures. The following results refer to investigating every 100 ps of each trajectory as frames to keep the size of RMSD matrices and event curves on a moderate level. We tested the choice of different frequencies between 10 to 300 ps for arbitrarily chosen trajectories of V3 and Met-Enkephalin, which produced comparable overlap values. For convenience, we will use every 100 ps as intermediate frequency because the generated trajectory lengths are a multiple of this value.

System Preparation The system preparation is done in the following seven stages [37]. All energy minimizations are achieved by 15000 steps of *steepest descent* followed by 15000 steps *conjugate gradient* setting the convergence criterion to $\approx 0.02\ \frac{\text{kCal}}{\text{mol}\mathring{A}}$. All heating steps are done in the NVT ensemble from an initial temperature of 0 K to 300 K over a period of 1 ns using the Langevin thermostat [60] option with a collision frequency of 2 ps^{-1}. A constant pressure of 1 atm is obtained using the NPT ensemble over 1 ns with the Berendsen barostat [57] and the same Langevin thermostat.

1. The molecule is firstly energetically minimized in vacuum after the ACE,NME attachment. In the case of the homologous models, this is the first step to optimize possible unfavored configurations.

2. Afterwards, periodic boundary conditions are applied with a truncated octahedron box, where the minimal distance between the box boundary and the molecule is set to 1.1 nm. Then, TIP3P water molecules [65] are inserted using TLEAP. In the case of V3, the system is neutralized with three chlorine ions Cl^- replacing water molecules.

3. For the full system, only the water molecules are energetically minimized with position restrained molecule atoms, at first. This shall resolve large forces between the molecule and the rigid water bodies which are placed into the box around the molecule.

4. Then, the full system is energetically minimized with released molecule atoms to allow the system to come to its favored (local) energy minimum.

5. Now, the water and side-chain atoms of the molecule are relaxed in 1 ns NVT heating and 1 ns NPT constant pressure simulations with harmonically position constrained peptide heavy atoms using a restraining weight of $10 \frac{\text{kCal}}{\text{mol}\mathring{A}^2}$.

6. This is followed by another 1 ns NVT heating and 1 ns NPT constant pressure runs without position constraints. The system should now be able to proceed from the local state introduced by the starting point. This stage is now used as the starting point of the simulation.

7. Hence, the system is now finally energetically minimized, followed by a heating and equilibration to the desired values of 300 K at 1 atm over a 1 ns NVT and 1 ns NPT simulation.

The MD productions of cMD, aMD or sMD runs are simply continuations in the NPT ensemble, whereas every 10 ps are stored. We generated several trajectories for the combination of the three sampling algorithms and two starting structures. For Met-Enkephalin, we simulated 1×100 ns, 4×200 ns and 3×1000 ns for each of the six combinations, obtaining in total 48 trajectories. For V3, we generated in total 60 different trajectories, i.e. for each combination 3×100 ns and 7×200 ns.

For aMD simulations, we applied the dual boost potential following Eq. (2.16). The parameters for E_P, E_D, α_P and α_D are given in Tables. 4.3-4.4.

The sMD simulations are all done with a scaling factor of $\lambda = 0.7$ following Eq. (2.24).

For all simulations, we use different velocity seeds, also in the preparation steps, to avoid synchronization effects between trajectories and generate independent results.

4.1.4. Conformational analysis after MD preparation

We use a multistage preparation of the starting structures for refinement as described before. It is therefore interesting to investigate the impact of this preparation. This is done by monitoring the RMSD values involving all atoms of the corresponding protein

Table 4.3.: Parameters for the aMD simulations of Met-Enkephalin. They follow Eqs. (2.17) for the eight different velocity seeds and two different starting structures *Met79* and *Met153*. All parameters are given in kCal/mol.

	Met79 [kCal/mol]				*Met153* [kCal/mol]			
Seed	E_P	E_D	α_P	α_D	E_P	E_D	α_P	α_D
1	-14012.610	71.991	753.6	5.6	-14365.963	72.310	772.32	5.6
2	-14012.674	72.126	753.6	5.6	-14366.651	72.098	772.32	5.6
3	-14013.114	72.382	753.6	5.6	-14366.175	72.120	772.32	5.6
4	-14012.920	72.256	753.6	5.6	-14366.655	72.055	772.32	5.6
5	-14012.852	72.308	753.6	5.6	-14365.911	72.074	772.32	5.6
6	-14013.239	72.211	753.6	5.6	-14366.255	72.310	772.32	5.6
7	-14012.854	72.322	753.6	5.6	-14366.620	72.203	772.32	5.6
8	-14012.777	72.246	753.6	5.6	-14366.206	71.853	772.32	5.6

Table 4.4.: Parameters for the aMD simulations of V3. They follow Eqs. (2.17) for the ten different velocity seeds and two different starting structures *V3a* and *V3b*. All parameters are given in kCal/mol.

	V3a [kCal/mol]				*V3b* [kCal/mol]			
Seed	E_P	E_D	α_P	α_D	E_P	E_D	α_P	α_D
1	-63915.873	531.359	3395.52	29.6	-64985.331	535.238	3452.16	29.6
2	-63910.738	532.779	3395.52	29.6	-64982.185	533.748	3452.16	29.6
3	-63912.494	535.902	3395.52	29.6	-64978.640	532.316	3452.16	29.6
4	-63915.873	531.359	3395.52	29.6	-64985.331	535.238	3452.16	29.6
5	-63915.873	531.359	3395.52	29.6	-64985.331	535.238	3452.16	29.6
6	-63915.873	531.359	3395.52	29.6	-64985.331	535.238	3452.16	29.6
7	-63912.823	534.132	3395.52	29.6	-64978.720	535.256	3452.16	29.6
8	-63910.143	531.269	3395.52	29.6	-64982.629	534.914	3452.16	29.6
9	-63916.140	533.627	3395.52	29.6	-64980.097	534.822	3452.16	29.6
10	-63916.858	533.254	3395.52	29.6	-64986.086	529.319	3452.16	29.6

after optimal superposition to the backbone atoms between the structures of the different trajectories in the different preparation steps. For Met-Enkephalin, we obtain 16 different structures, which are used to generate 16 cMD, 16 aMD and 16 sMD trajectories. For V3, we have 20 different structures for the 20 cMD, 20 aMD and 20 sMD trajectories.

It is clear that in the beginning, there are only two different starting structures by definition for both molecules separated by a certain RMSD value. Hence, we monitor the following steps for Met-Enkephalin and V3, respectively, in Figs. 4.3-4.4: (1) After the vacuum, water with restrained protein and complete minimization, (2) after the first

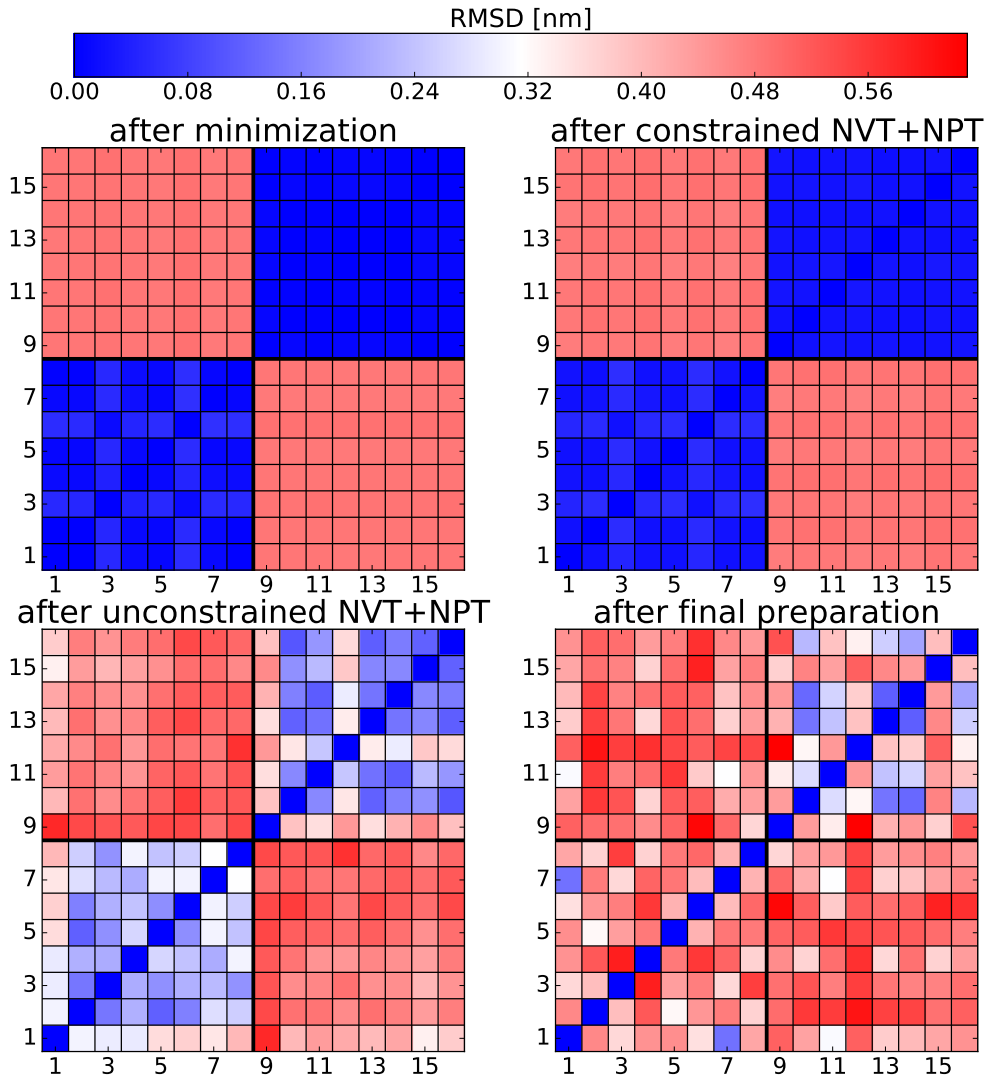

Fig. 4.3.: RMSD values between structures after MD preparation of Met-Enkephalin. All different 16 (two starting structures and 8 velocity seeds) trajectories of Met-Enkephalin are prepared for MD production using the following steps: (Top left) After the fourth step, where the system is multiple times energetically minimized. (Top right) after the first 1 ns heating and 1 ns constant pressure simulations with position constrained protein. (Bottom left) After the 1 ns heating and 1 ns constant pressure simulations without position constraints. (Bottom right) After the final preparation just before the production phase. The minimal to maximal RMSD values ranges from $[0, 0.63]$ nm.

heating and constant pressure equilibration with position constrained backbone heavy atoms, (3) after the unconstrained NVT and NPT simulation and (4) after the final minimization and preparation to 300 K and 1 atm.

It is expectable that after the first two steps, there are only minor changes in the structures compared to each other for both molecules. Afterwards, there should be a

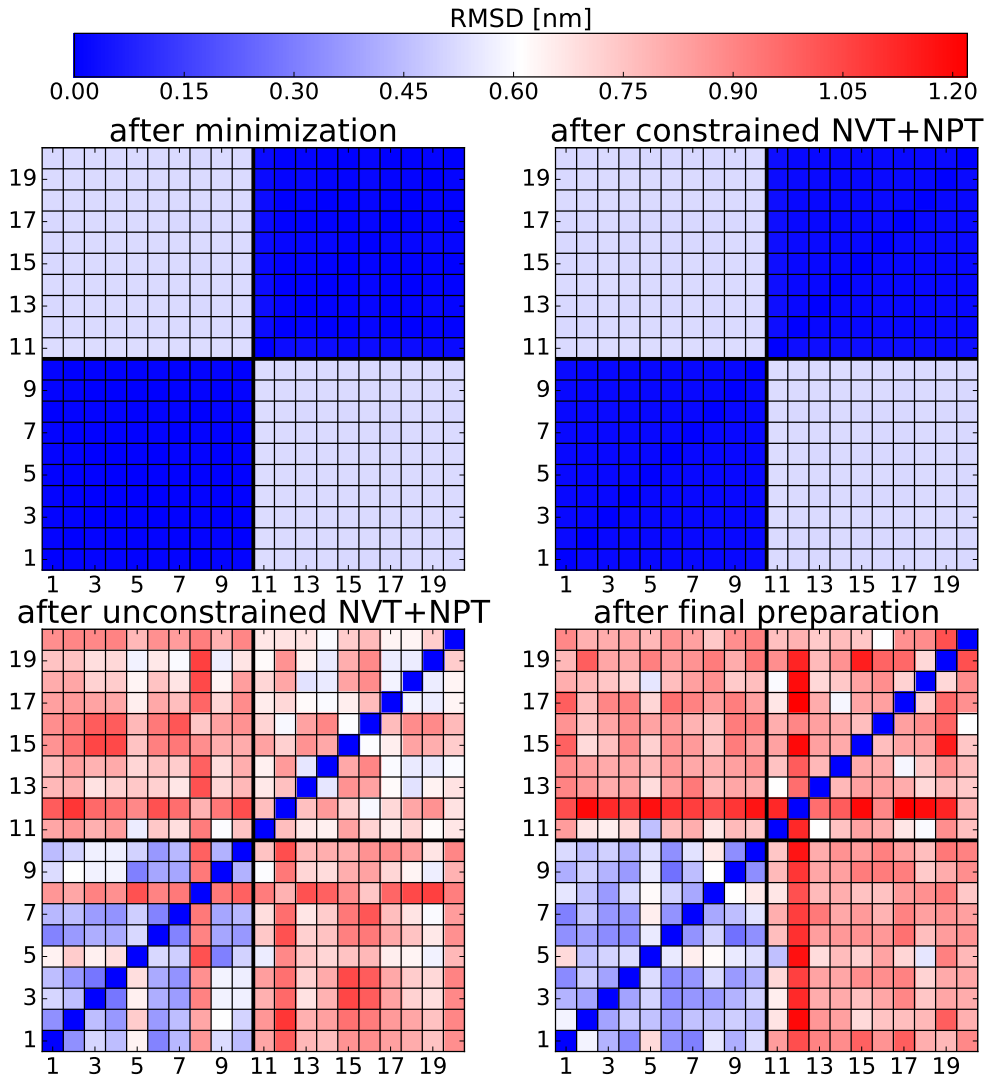

Fig. 4.4.: RMSD values between structures after MD preparation of V3. All different 20 (two starting structures and 10 velocity seeds) trajectories of V3 are prepared for MD production using the following steps: (Top left) After the fourth step, where the system is multiple times energetically minimized. (Top right) after the first 1 ns heating and 1 ns constant pressure simulations with position constrained protein. (Bottom left) After the 1 ns heating and 1 ns constant pressure simulations without position constraints. (Bottom right) After the final preparation just before the production phase. The minimal to maximal RMSD values ranges from $[0, 1.22]$ nm.

significant change between the different structures.

For Met-Enkephalin, the first unconstrained NVT and NPT simulations generate deviations between the configurations of the same initial structures. But the RMSD ≈ 0.5 nm is maintained between the configurations of the different initial structures (see Fig. 4.3 bottom left). After the final minimization and the two equilibration processes (see Fig. 4.3

bottom right), the structures also lose the similarity if they came from the same initial structure. Thus, we obtain almost 16 totally different configurations for the subsequent MD productions where they might have lost the bias from the two starting structures.

For V3, the behavior is different after the unconstrained equilibration step, illustrated in Fig. 4.4. First of all, the two starting structures *V3a* and *V3b* have an initial deviation of RMSD ≈ 0.5 nm. This deviation is increased to a value of RMSD > 0.6 nm during the preparation steps between the structures coming from the one and structures coming from the other initial structure. The same deviation can be detected between structures 11 to 20, which originate from *V3b*. In contrast, the deviations between structures 1 to 10 (coming from *V3a*) are lower, but are also significantly increased compared to the initial state (Fig. 4.4 bottom left). The overall behavior is not changed after the final minimization and equilibration (see Fig. 4.4 bottom right): The final structures coming from *V3a* are more related to each other with an average RMSD ≈ 0.5 nm. Thus, the structure *V3a* seems to be more conserved and the corresponding configurations stay closer to this initial model after the full preparation. The other structures coming from *V3b* have large deviations of up to RMSD ≈ 1 nm between all other structures, losing the information about their origin. It will be interesting, if this behavior will be detectable in the production step.

4.2. Threshold parameter r

In subsection 3.1.2, we introduced the threshold r which is used across all analyses, $O_{\text{conf}}, O_{\text{dens}}, N_C$ and S_C. It can be understood as a resolution: the smaller r, the more similar must be two different structures to be considered the same. Therefore, the following two questions arise: How can we detect reasonable values for r? And is there an optimal choice for r? Since r is based on RMSD values between two (superimposed) structures, it is a good strategy to investigate the distributions of the RMSD values of different single and trajectory combinations. Second, we will analyze the number of found clusters N_C as a function of the threshold r trying to identify an optimal value for the threshold.

4.2.1. RMSD distributions

The RMSD distributions of every single trajectory and all pairs of trajectories give insight about the range of relevant threshold parameters r. Furthermore, one can extract the critical points $r_{\text{min}}, r_{\text{max}}$ (see subsection 3.1.2) between which the overlap measures $O_{\text{conf}}, O_{\text{dens}}$ range from zero to one. On the other hand, the RMSD distributions of single

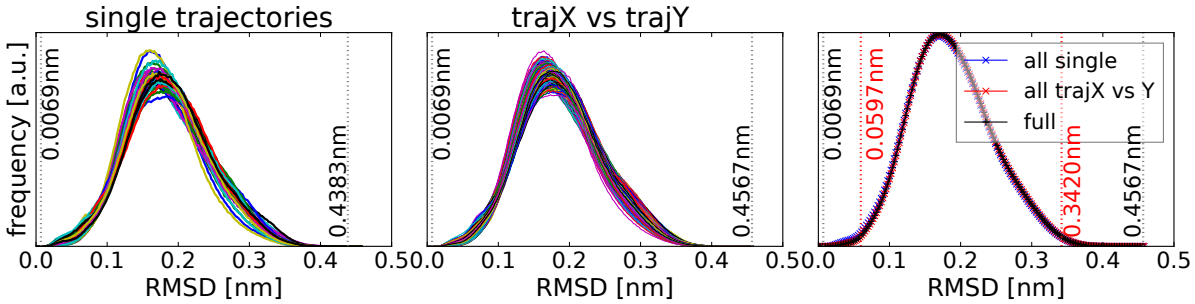

Fig. 4.5.: RMSD distributions of Met-Enkephalin trajectories. Curves refer to all 42×200 ns trajectories from cMD, aMD and sMD sampling, showing the minimal and maximal RMSD values obtained from all histograms of 200 bins. aMD and sMD results are not re-weighted. Left: RMSD values between all pairs of frames in each single trajectory. Middle: RMSD values between all pairs of frames from each two-trajectory combinations. Right: The RMSD distributions of all combined single trajectories (blue), all combined pairs of trajectories (red) and all combined trajectories (black); red vertical lines enclose 99% of the area below the distribution of all combined trajectories.

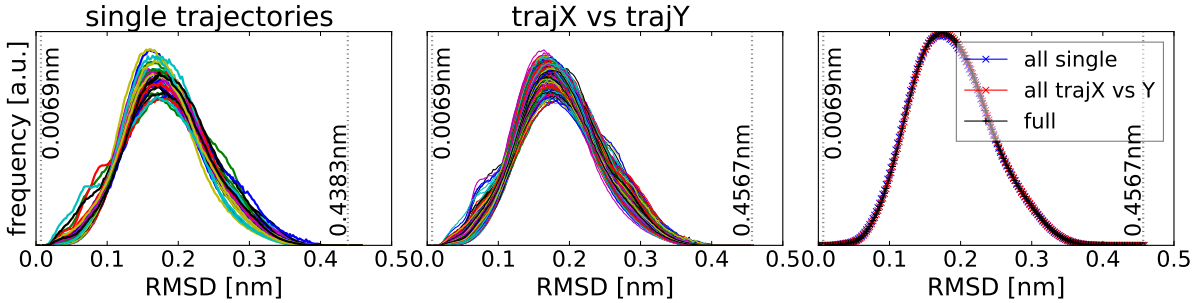

Fig. 4.6.: Re-weighted RMSD distributions of Met-Enkephalin trajectories. Curves refer to all 42×200 ns trajectories from cMD, aMD and sMD sampling, showing the minimal and maximal RMSD values obtained from all histograms of 200 bins. aMD and sMD results are re-weighted using $\mathrm{MF}^{(1)}$ at $r = 0.11$ nm. Left: RMSD values between all pairs of frames in each single trajectory. Middle: RMSD values between all pairs of frames from all single two-trajectory combinations. Right: The RMSD distributions of all combined single trajectories (blue), all combined pairs of trajectories (red) and all combined trajectories (black).

and concatenated trajectories can already reveal first tendencies of the underlying sampling. For instance, if two independent trajectories sample the same free energy minimum, they will have a monomodal RMSD probability distribution, which will lead to a peak at low RMSD values according to the small structure deviations in the potential well. If two independent MD runs result in sampling of two distinct free energy minima, one can expect a bimodal RMSD probability distribution, where the two peaks refer on the one hand to the small structure deviations in the corresponding different energy minima at small RMSD values and on the other hand to the deviations coming from the structure

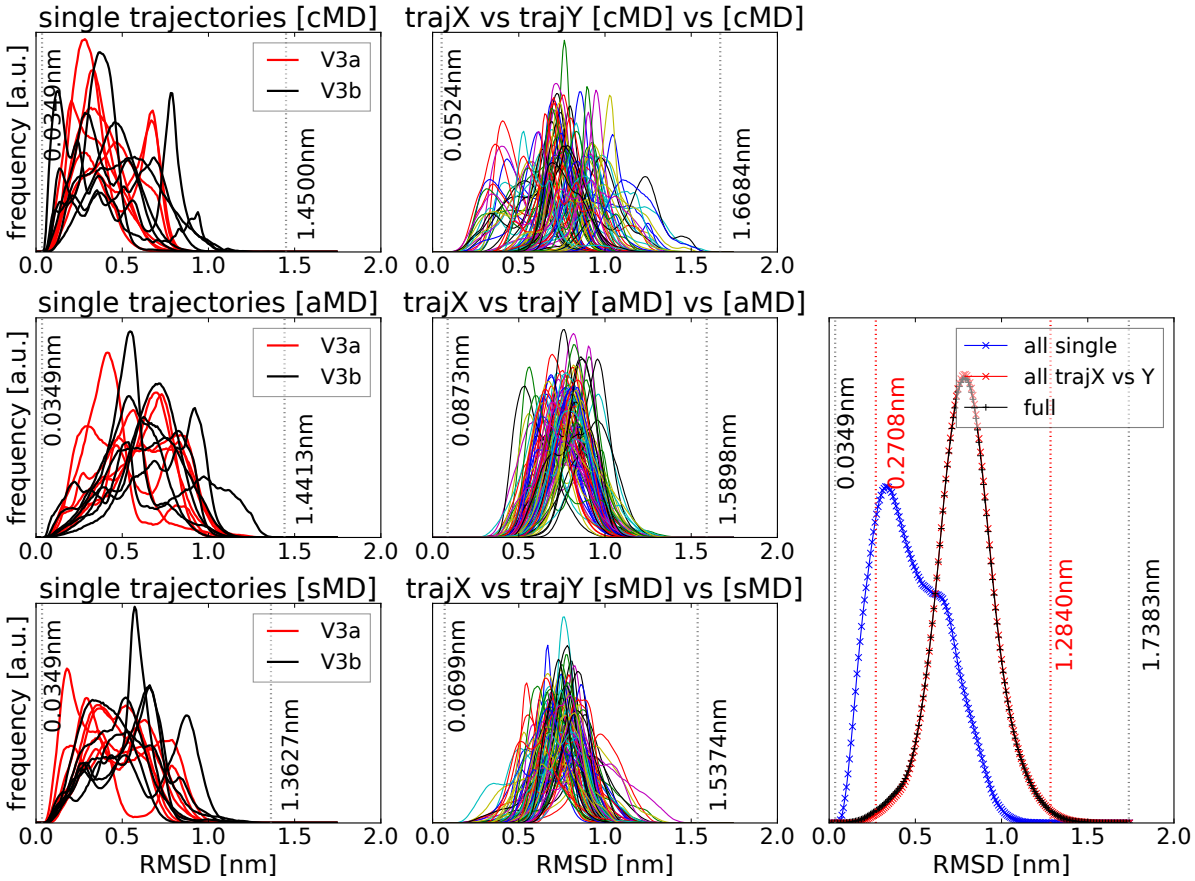

Fig. 4.7.: RMSD distributions of V3 trajectories. Curves refer to all 42 × 200 ns trajectories from cMD, aMD and sMD sampling, showing the minimal and maximal RMSD values obtained from all histograms of 200 bins. aMD and sMD results are not re-weighted. Left column: RMSD values between all pairs of frames in each single trajectory of all cMD (top), aMD (middle) and sMD (bottom) runs. Middle column: RMSD values between all pairs of frames from each two-trajectory combinations of all cMD (top), aMD (middle) and sMD (bottom) runs. Right: The RMSD distributions of all combined single trajectories (blue), all combined pairs of trajectories (red) and all combined trajectories (black); red vertical lines enclose 99% of the area below the distribution of all combined trajectories.

deviations between both energy wells at large RMSD values.

Hence, the RMSD distributions give a first information about the quality of the sampling and are used as a first classification in the tool *PySamplingQuality.py* [37]. Here, we will use 200 bins to generate the discrete RMSD distributions.

Since we use two enhanced sampling algorithms aMD and sMD, which distort the energy landscapes and lead to biased probability distributions $p(\vec{r})$, the RMSD distributions of these trajectories will lead to wrong frequencies and must be re-weighted to yield the correct ensembles. In detail, since the RMSD values are defined by two simulated frames,

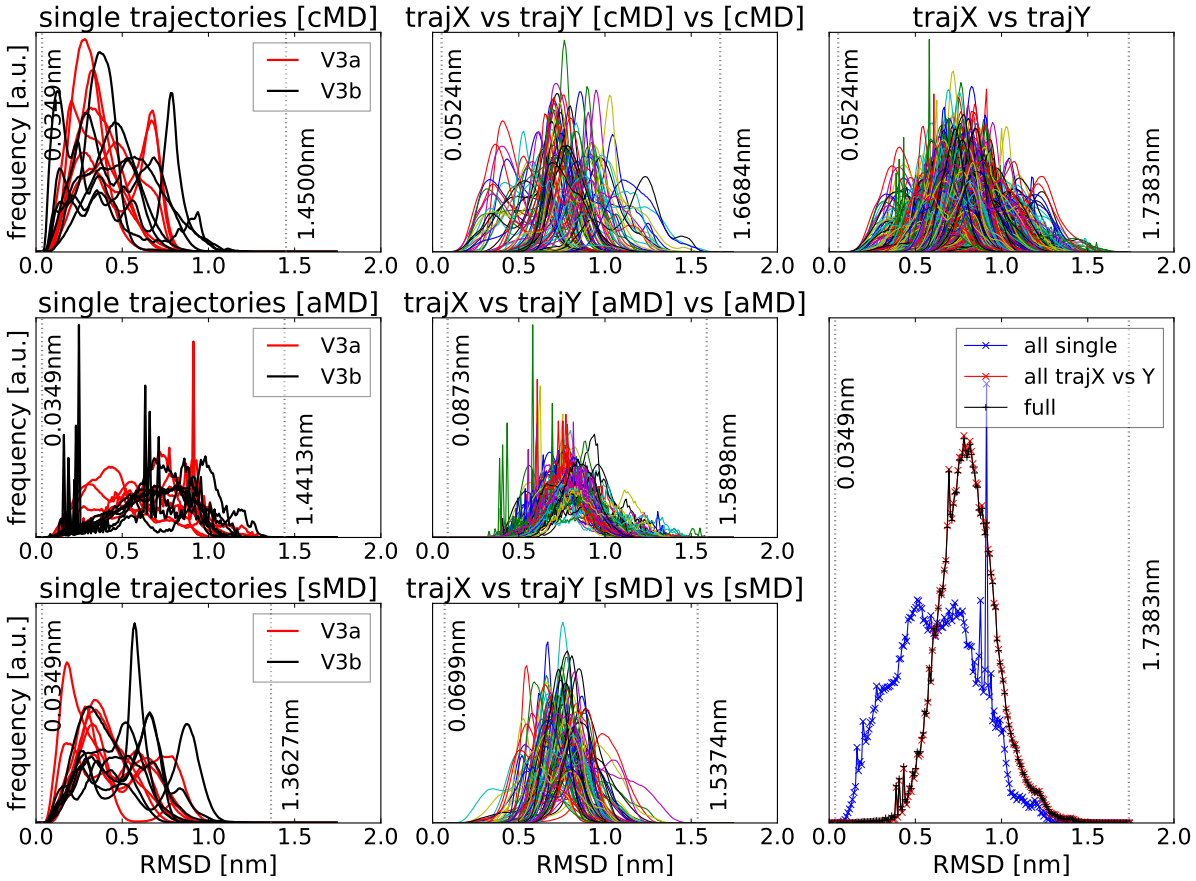

Fig. 4.8.: Re-weighted RMSD distributions of V3 trajectories. Curves refer to all 42×200 ns trajectories from cMD, aMD and sMD sampling, showing the minimal and maximal RMSD values obtained from all histograms of 200 bins. aMD and sMD results are re-weighted using $\mathrm{MF}^{(1)}$ at $r = 0.35$ nm. Left column: RMSD values between all pairs of frames in each single trajectory of all cMD (top), aMD (middle) and sMD (bottom) runs. Middle column: RMSD values between all pairs of frames from each two-trajectory combinations of all cMD (top), aMD (middle) and sMD (bottom) runs. Right top: RMSD values between all pairs of frames from each two-trajectory combinations. Right bottom: The RMSD distributions of all combined single trajectories (blue), all combined pairs of trajectories (red) and all combined trajectories (black).

they must be multiplied by the weights from both corresponding frames. But these weights are generated in later stages of the calculation, based on the event curves as a function of the simulation time, see subsection 3.2.4. This is unproblematic for the extraction of relevant r-values together with $r_{\min}$ and $r_{\max}$, but the comparison of biased and unbiased distributions must be done with care. More relevant is the comparison between trajectories with the same/similar acceleration. Thus, we will discuss the non-weighted case first. As brief outlook, we show the re-weighted distributions for the first mean-field step $\mathrm{MF}^{(1)}$

for relevant thresholds r discussed in later sections.

For Met-Enkephalin, the (non-weighted) RMSD distributions are shown in Fig. 4.5. All single trajectories, all pairs of trajectories and the combined case of all concatenated trajectories show (almost) the same monomodal RMSD distributions in the range of about 0.01 nm to 0.45 nm with the maximum at 0.17 nm. One possible explanation for this is a good sampling already within single 200 ns cMD trajectories, which seems not to be improved by accelerated sampling methods. It is very interesting that the biased aMD and sMD trajectories show the same bell shaped distributions. As arbitrary choice, we use the range of 99% of the area below the RMSD distribution of all combined trajectories to extract values for $r_{\text{min}}^{(\text{Met})} = 0.0597$ nm and $r_{\text{max}}^{(\text{Met})} = 0.3420$ nm as minimum and maximum for the threshold values. The reason to limit the range to 99% is to obtain an adequate region for the integral of the averaged overlap from Eq. (3.7), because too large r values trivially lead to an overlap of one and therefore will overestimate the average overlap Ω_{conf} and Ω_{dens}. Interestingly, the re-weighted distributions of aMD and sMD trajectories ($\text{MF}^{(1)}$ at $r = 0.11$ nm following Eq. (3.13) and (3.16)), illustrated in Fig. 4.6, show almost no difference to the non-weighted case. This underlines the indication of good sampling of cMD trajectories, whereas the detailed analysis of the sampling will be done in the next sections.

For the non-weighted V3 experiments shown in Fig. 4.7, the results are completely different. In contrast to Met-Enkephalin, V3 shows a set of highly diverse RMSD distributions with multimodal shape already within single cMD trajectories (see Fig. 4.7 top left). Remarkably, the combination of two trajectories leads in all cases to a significant shift of RMSD values by about 0.2 nm to larger values (see Fig. 4.7 middle column). This is also visible in the combined case of all single, all pairs and all trajectories in the right panel of Fig. 4.7. One explanation of this behavior might be that many single trajectories sample different regions of the conformational space in contrast to Met-Enkephalin. This shift between the single trajectory and pair RMSD distributions already tells a lot about a possible threshold value r: For instance, for a reference frame κ coming from trajectory l and a value of $r = 0.35$ nm, the normalized number of events $\tilde{e}_{r,\kappa l}$ Eq. (3.10) might contain a large amount of structures of trajectory l, but a small amount of structures of trajectory $j \neq l$ with $\tilde{e}_{r,\kappa l} \gtrsim \tilde{e}_{r,\kappa j}$, comparing the left and middle columns of Fig. 4.7. Again, we use the arbitrary 99% of the area below the RMSD distribution of all combined trajectories to set $r_{\text{min}}^{(\text{V3})} = 0.2708$ nm and $r_{\text{max}}^{(\text{V3})} = 1.2840$ nm as minimal and maximal values. Comparing the RMSD distributions of cMD with aMD or sMD, the two latter have smaller densities at low RMSD and larger at higher values, which can

be expected from sampling methods that drive the system out of local minima. This is modified for the re-weighted case illustrated in Fig. 4.8 ($MF^{(1)}$ at $r = 0.35$ nm following Eq. (3.13) and (3.16)): For aMD, some trajectories from the second starting structure have steep peaks at small RMSDs showing the sampling of multiple structures in few minima and also large steep peaks at large RMSD values, which might originate from different energy minima. Remarkably, the steep peak at large RMSD comes from only one trajectory starting from *V3a*, thus both starting conformations behave differently. Such peaks may indicate unconverged aMD sampling which then leads to errors in the re-weighting discussed in subsection 2.2.3. On the other hand, the applied re-weighting for these distributions assumes that one discrete RMSD value is sampled with a higher probability. This assumes that (two) structures are multiple times identically reproduced which will rarely be the case and is an approximation. Such a discrete assumption will automatically lead to very large peaks in the RMSD distribution instead of a smooth bell shaped curve. For sMD sampling, there is a minor effect coming from the re-weighting which shifts the RMSD distributions toward lower RMSDs similar to cMD. Because sMD is based on population re-weighting, which re-scales the distributions by an exponent of $1/\lambda$ discussed in subsection 2.2.3, these large irregular peaks do not appear.

The effect of re-weighting and the quantitative assessment of the sampling quality will be investigated in the next sections.

4.2.2. Is there an optimal r?

It is reasonable trying to extract an optimal neighboring threshold r. In theory, ideal ergodic sampling ($t \to \infty$) will give the same overlap values for all r, because the probability density functions $p(\vec{r})$ are the same for different trajectories. On the other hand, r sets the resolution of the analysis. This means if r is set to too small values, the number of found clusters tends toward the number of single frames because every structure defines its own cluster. The other extreme of a too large r means, almost everything will end up in one single cluster because every structure is considered the same.

For this reason, we analyze the number of found clusters N_C^{global} as a function of different relevant threshold parameters r [37]. Since we want to compare different results of different trajectories, we use here the global clustering defined previously using all 200 ns trajectories of the corresponding molecule. The results of both molecules are illustrated in Fig. 4.9 in log-log plots. One can see that the functions follow a power-law distribution $N_C^{\text{global}}(r) \propto r^{-\beta}$, which is a characteristic property of a scale-free system. This is a strong indication that there is no optimal choice for r but the clustering follows a same random walk at different resolutions [214].

There are the following relevant outcomes which should be mentioned. For Met-Enkephalin, all fits are in agreement with an exponent of $\beta \approx -4.7$ where all values lie in their confidence intervals of 95%. This is true for the concatenated case as well as for the single sampling algorithms. Additionally, there is almost no deviation in the cluster numbers between different trajectories and both starting structures behave the same (see Fig. 4.9 left). For V3, there is a clear difference between the concatenated and the other cases. Whereas for the combined case of 42 trajectories, the exponent $\beta \approx -4.26$ is compatible with Met-Enkephalin, the numbers of clusters N_C^{global} of single trajectories are much smaller and behave differently. This leads to a possible conclusion that different groups of trajectories explore different parts of the conformational space. For cMD and sMD (Fig. 4.9 (a),(c) right), the deviations are similar, leading to two different fits for the two starting conformations. The behavior between both initial structures of aMD (Fig. 4.9 (b) right) almost leads to the same exponent. Nevertheless, all exponents lie in their 95% confidence intervals for the single trajectory numbers, although the numbers of found clusters are different.

It is very interesting that regarding the clustering, the choice of the neighborhood threshold r seems to have no preference, which is also true for the optimal case of ergodic sampling for the overlap measures. Therefore, one needs to select a broader range of r-values to screen through different resolutions or set the threshold r to a value reflecting

the system of interest. With smaller r, one investigates different aspects of a system compared to larger r values.

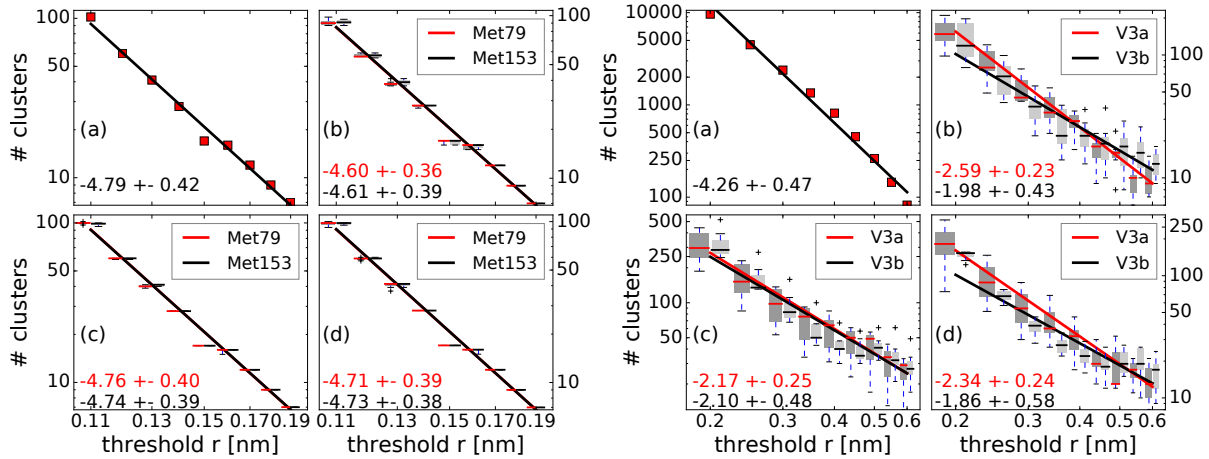

Fig. 4.9.: The number of found clusters N_C^{global} as a function of the threshold r for Met-Enkephalin (left four panels) and V3 (right four panels) in log-log plots. (a) N_C^{global} of all 42×200 ns concatenated trajectories of the respective molecule for the combination of both starting structures and three sampling methods. The other panels show boxplots for single trajectory N_C^{global} of both starting structures (red and black) and sampling method cMD (b), aMD (c), and sMD (d). The lines are fits of $N_C^{\text{global}}(r) = \alpha r^{-\beta}$ with fit parameters α, β, whereas the exponents β are given with their 95% confidence intervals. The figure is taken from Ref. [37].

4.3. Insert: Weights for the correction of enhanced sampling

In subsection 2.2.3, we introduced the necessary re-weighting for the biased ensembles sampled by aMD and sMD. We also discussed the possible sources of error and the upcoming difficulties. Still today, it is subject of active research [115, 118, 122, 215] and not solved for general cases. Nevertheless, it is critically needed to obtain the unbiased results. For our purpose, the re-weighting is slightly different from the standard procedure of a discrete projection onto a N-dimensional space by binning the results into disjunct partitions. For this reason, we defined the mean-field treatment in subsection 3.2.4.

In this insert, we want to discuss and show the influences of different re-weighting schemes to correct the events and overlap obtained from aMD/sMD runs. First of all, the weights are explicitly r-dependent, i.e. each microstate is linked to the chosen resolution, similar to the binning approach of the McCammon group [112, 116, 117]. For a threshold of $r \approx 0.0$ nm, the mean-field solution will converge in one step to the result of the

standard exponential re-weighting Exp, because the average $\left\langle \Delta V^{(n)} \right\rangle_{r,\alpha l}$ of Eq. (3.14) will contain only the frame α itself. This assumes that every frame α is a separate microstate and possible errors due to energy fluctuations cannot be decreased by averaging frames within a microstate. A threshold of $r \geq r_{\max}$ will lead to a uniform $\left\langle \Delta V^{(n)} \right\rangle_{r,\alpha l}$ for all α. Hence, the resolution will be very bad.

The first point of interest is the convergence behavior, whether our mean-field approach converges for the given trajectories coming from the two different flexible molecules. For both molecules, we chose arbitrarily four 200 ns-trajectories of different starting structures and both combinations of enhanced sampling methods, together with four different thresholds r. In Table 4.5, there are the necessary steps to reach the convergence criterion that the difference of each weight is $< 10^{-6}$ for the next iteration step. In our algorithm, weights are considered unitless for aMD ($\beta \cdot \Delta V_{\cdot}$, Eq. (3.13)) and sMD ($N_{\cdot}^{1/\lambda}$, Eq. (3.17)), whereas the exponential function for aMD weights is applied after the mean-field iteration. One can see that convergence is consistently reached much faster for sMD trajectories, independent of the starting conformation or even the molecule. There are two reasons for this behavior: First, sMD might sample the conformational space less aggressive compared to aMD, because the potential is simply scaled down. Second, there are no errors coming from energetic fluctuations, thus one does not rely on the precision of the measured boost potentials. This is different for the aMD weights. aMD weights suffer from the two sources of errors discussed in subsection 2.2.3. Thus, it is expectable that, for the smaller molecule, the aMD trajectories are much closer to the convergence regime than in the case of aMD runs of V3 for the same simulation time. This was already indicated in the RMSD distributions in subsection 4.2.1. Nevertheless, all weights converge fast, which is shown in Table 4.5.

The weights for the 2000 frames of the 200 ns-trajectories for both molecules and aMD/sMD sampling are illustrated in Figs. 4.10-4.13. For the aMD weights, we show $w_{r,\alpha l}^{(\mathrm{aMD})}$ following Eq. (3.13) for different re-weighting schemes Exp, McL up to 10th order and MF for different steps, introduced in subsection 3.2.4.

For Met-Enkephalin at $r = 0.11$ nm (Fig. 4.10), there is a clear difference between Exp, McL and MF. With Exp re-weighting, there are only very few frames which hold almost the full weight of the system, which is only slightly changed using the Maclaurin approximation. The weights of the mean-field approach are much smaller and distributed between many frames, due to the averaged boost potential across one microstate or the r-neighborhood, respectively . Interestingly, further steps seem to represent quickly the same behavior but the amplitudes are changed, which is especially true for *Met79*

Table 4.5.: The number of steps needed to reach the convergence criterion that the difference of each weight is $< 10^{-6}$ for the next step. The steps are shown for different thresholds r for (arbitrary chosen) two aMD and two sMD 200 ns-trajectories of Met-Enkephalin and V3 from different starting structures *Met79*,*Met153* and *V3a*, *V3b*, respectively. The last two lines show the calculation time on a single state of the art CPU for the weight generation of all weights on their corresponding columns. The times for the first step $\mathrm{MF}^{(1)}$ and the converged $\mathrm{MF}^{(\infty)}$ are reported.

	aMD		sMD			aMD		sMD	
r [nm]	*Met79*	*Met153*	*Met79*	*Met153*	r [nm]	*V3a*	*V3b*	*V3a*	*V3b*
0.08	680	652	21	20	0.15	4617	2545	22	21
0.10	254	201	22	21	0.25	26634	7067	23	22
0.11	171	89	22	21	0.35	13393	1696	23	22
0.13	90	83	22	22	0.45	1402	340	23	22
$\mathrm{MF}^{(1)}$ [s]	0.13	0.13	0.05	0.05		0.05	0.14	0.07	0.05
$\mathrm{MF}^{(\infty)}$ [s]	13.9	12.0	1.11	1.04		570.3	141.9	1.54	1.15

(Fig. 4.10 left). It is remarkable that the average over all weights stay almost constant, which is not shown. The weights for the other thresholds $r = 0.1$ nm and $r = 0.13$ nm show approximately the same result, whereas $r = 0.08$ nm seem to be so small that the weights go toward the regime of Exp.

The aMD weights of V3 for $r = 0.35$ nm, illustrated in Fig. 4.11, show a completely different behavior. The results of Exp and McL are almost the same as for Met-Enkephalin. But the first mean-field steps $\mathrm{MF}^{(1)}$ of *V3a* and *V3b* have also very large peaks with similar order of magnitudes as the Maclaurin expansion. For more mean-field steps, the weights show irregular shapes with different monotonous behavior for different starting conformations for increasing frames. The frames correspond to increasing simulation times (Fig. 4.11). Both weight curves seem to converge to a straight line after a certain amount of simulation time. But this is not the case, only the magnitudes of weights of later frames are much smaller than in the beginning (see Fig. 4.12 bottom panels). Thus, the mean-field approximation does not only smooth the weights, but if there are r-neighborhoods with few frames and large boost potential energies, this shape is still represented by the weights. In contrast, the McL weights show the hierarchical behavior, where few frames have much larger weights than the rest, if subparts of the trajectory are inspected (Fig. 4.12 top panels). This is different for the mean-field approach (Fig. 4.12 bottom panels), where different frames have relatively different weights representing the underlying conformational landscape. It has to be mentioned that the large peaks in

$MF^{(1)}$ are unproblematic in the overall overlap analysis, because they will be present only in a small amount of r-neighborhoods, and hence the relative difference between smaller weights will have larger impact. This is not the case for McL and Exp since also a corresponding "zoom" into smaller parts of the trajectory frames reveal the peak behavior that few weights are dominating the rest.

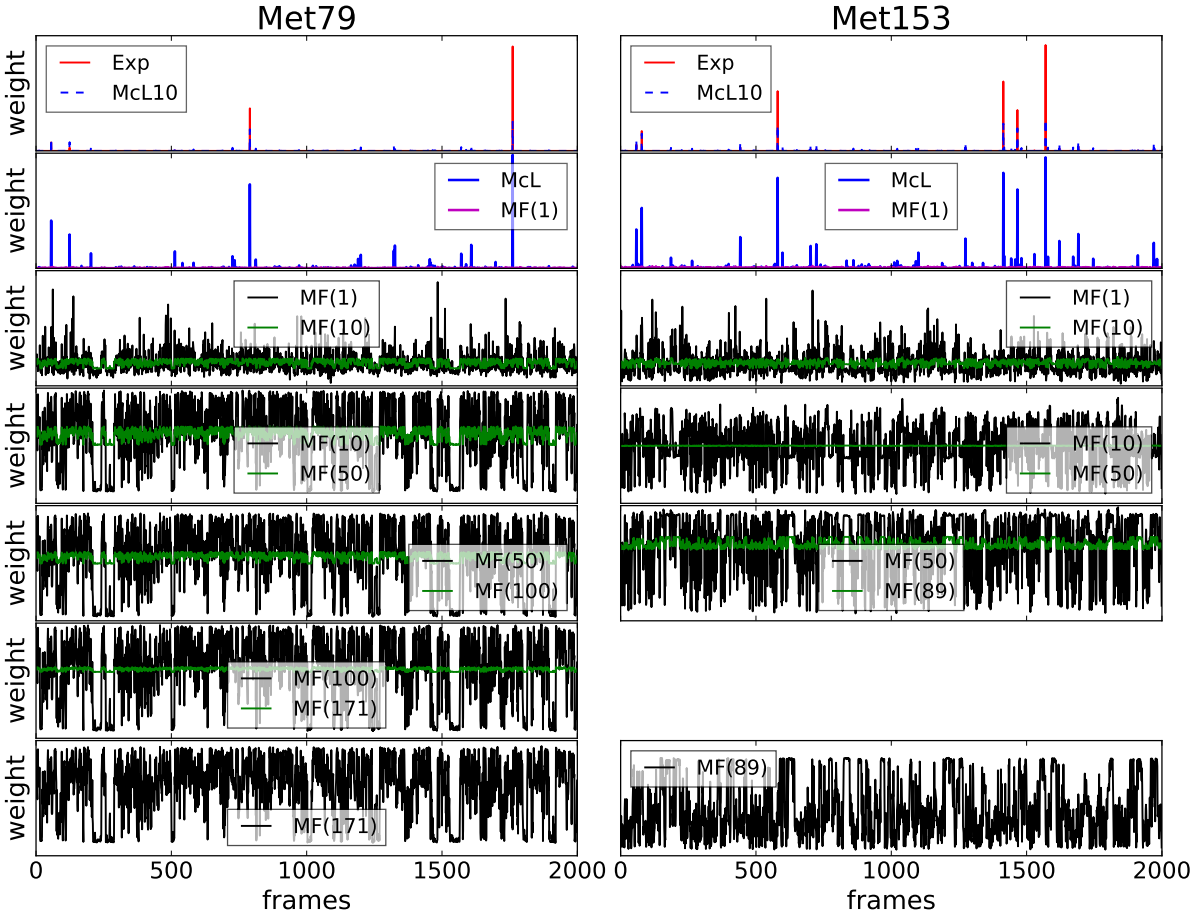

Fig. 4.10.: Weights of two arbitrary chosen aMD trajectories of Met-Enkephalin. Weights following Eq. (3.13) of different re-weighting schemes for all 2000 frames of two 200 ns aMD trajectories at $r = 0.11$ nm, starting from *Met79* (left) and *Met153* (right). Weights are normalized to a sum of one whereas two different re-weighting schemes are compared. From top to bottom: Exp, McL (10th order), $MF^{(1)}$, $MF^{(10)}$, $MF^{(50)}$, $MF^{(100)}$ and the converged $MF^{(171)}$ and $MF^{(89)}$, respectively.

For sMD sampling, the mean-field approach following Eq. (3.17) converges fast to similar values and the same order of magnitude compared to the first step $MF^{(1)}$. Remarkably, the range between minimal and maximal weight of the converged result $MF^{(\infty)}$ is consistently larger than the first step for all different thresholds r. For Met-Enkephalin at $r = 0.11$ nm (Fig. 4.13 top panels), both starting conformations show different behavior.

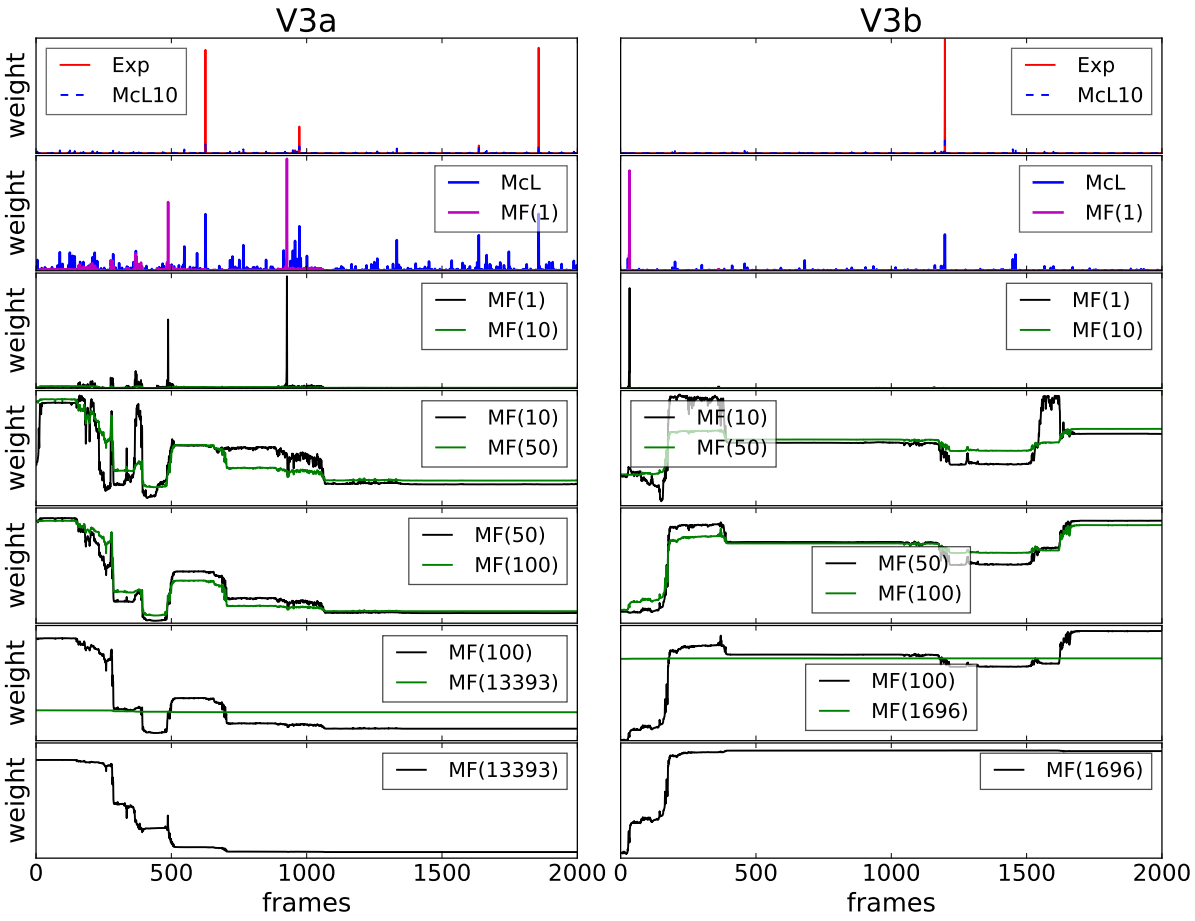

Fig. 4.11.: Weights of two arbitrary chosen aMD trajectories of V3. Weights following Eq. (3.13) of different re-weighting schemes for all 2000 frames of two 200 ns aMD trajectories at $r = 0.35$ nm, starting from *V3a* (left) and *V3b* (right). Weights are normalized to a sum of one whereas two different re-weighting schemes are compared. From top to bottom: Exp, McL (10th order), $\mathrm{MF}^{(1)}$, $\mathrm{MF}^{(10)}$, $\mathrm{MF}^{(50)}$, $\mathrm{MF}^{(100)}$ and the converged $\mathrm{MF}^{(13393)}$ and $\mathrm{MF}^{(1696)}$, respectively.

For *Met79*, there seem to be fewer transitions between low and large weights forming several plateaus. For *Met153*, neighboring frames have fast changing transitions between the minimum and maximum weight. Since the weights are based on the population of different r-neighborhoods, it will be interesting to investigate the underlying overlap result. For the two V3 trajectories at $r = 0.35$ nm, the behavior between both starting structures is also different (Fig. 4.13 bottom panels). There seem to be two regimes: low weights and large weights, where there are only few transitions between both states for both trajectories. All other investigated thresholds r yield similar behavior, only for V3 at $r = 0.15$ nm there are large fluctuations because only very few frames fall into the same r-neighborhoods.

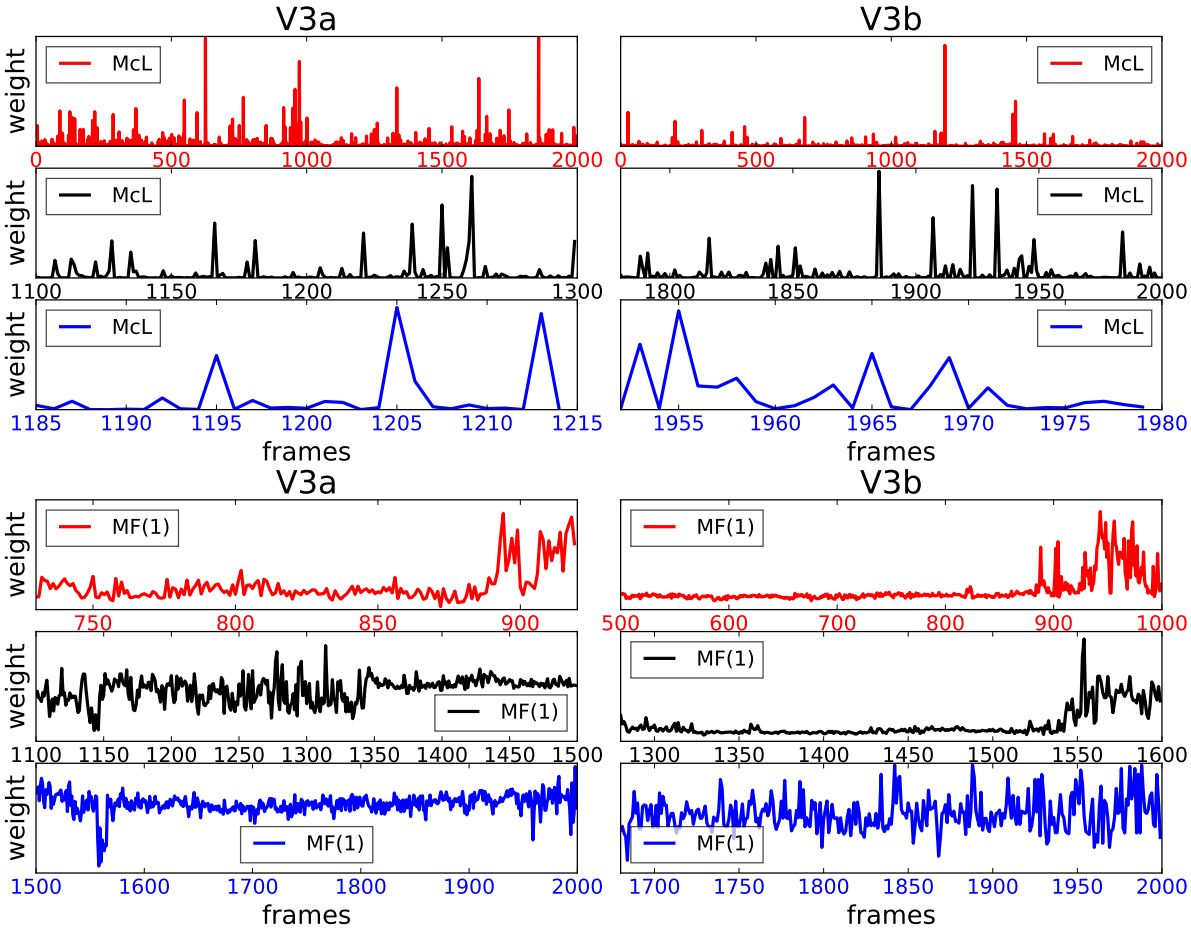

Fig. 4.12.: Comparison between McL (top) and MF$^{(1)}$ at $r = 0.35$ nm (bottom) weights for V3. *V3a* (left) and *V3b* (right) of two arbitrary chosen 200 ns aMD trajectories. Weights are normalized to a sum of one. Focus on different windows of frames to investigate the relative frequencies.

Now, we are interested in the resulting (density) overlap measure using different re-weighting schemes. For this purpose, we use for both molecules six different trajectories with the combination of cMD, aMD and sMD sampling and both starting conformations. We always evaluate the overlap between pairs of trajectories with at least one accelerated trajectory using $K = L$ for the reference and comparison set of trajectories of Eq. (3.11).

In Fig. 4.14, the results are illustrated for Met-Enkephalin. Considering the overlap results of the aMD sampling (Fig. 4.14 top), all MF steps outperform the exponential or Maclaurin re-weighting. Additionally, more steps for the mean-field iteration of aMD weights enhance consistently the overlap, whereas the overlap between both aMD trajectories benefits the most. For sMD (Fig. 4.14 center), more iteration steps lead to a slight decrease of the pair-overlaps.

The overlap values for different MF iterations of V3 are shown in Fig. 4.15. The

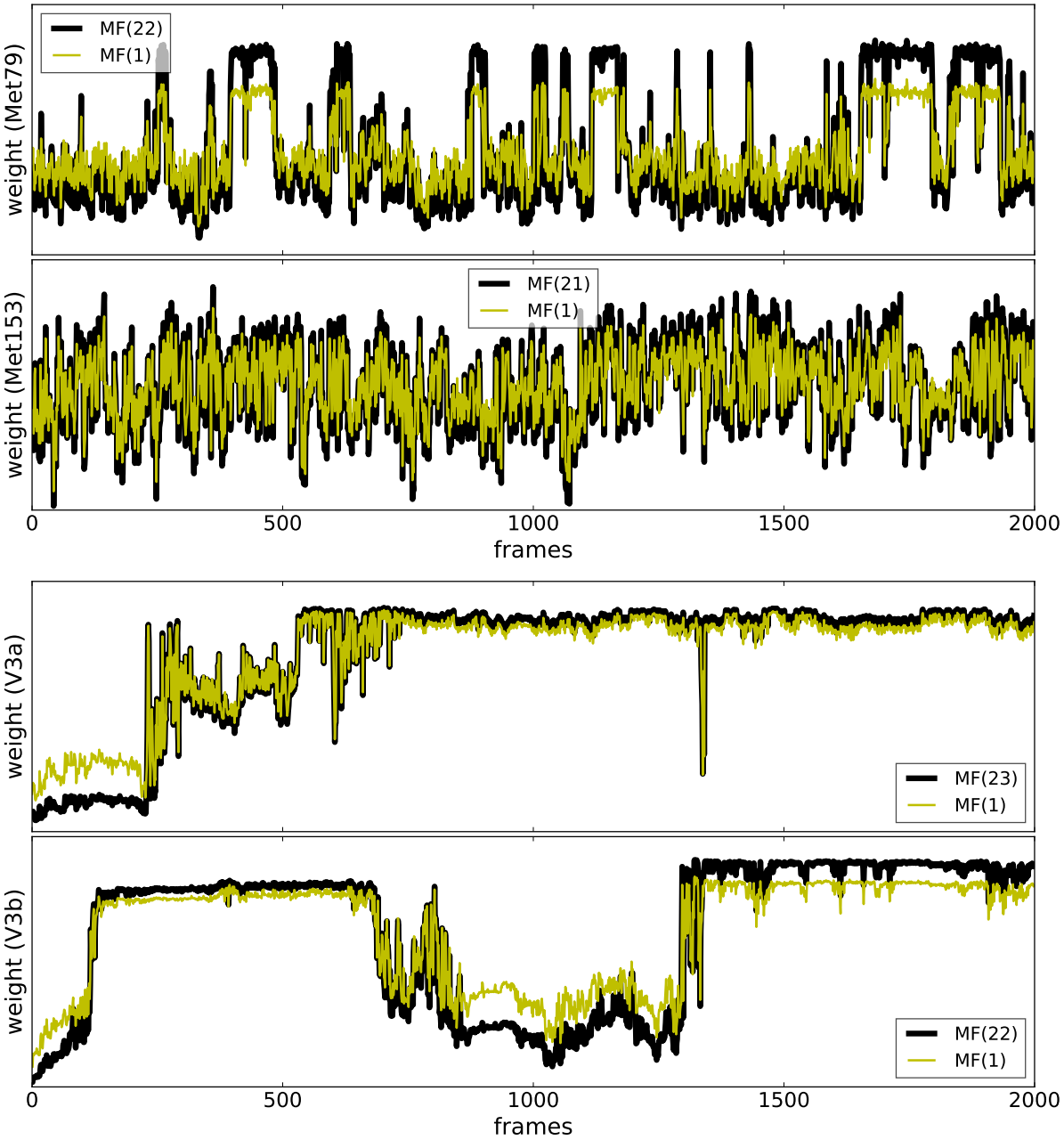

Fig. 4.13.: Weights of arbitrary chosen sMD trajectories. The weights follow Eq. (3.17) for all 2000 frames of two 200 ns sMD trajectories of Met-Enkephalin at $r = 0.11$ nm (upper panels) and V3 (lower panels). They are normalized to a sum of one. The first $\mathrm{MF}^{(1)}$ and converged $\mathrm{MF}^{(22)}$, $\mathrm{MF}^{(21)}$, $\mathrm{MF}^{(23)}$, $\mathrm{MF}^{(22)}$ mean-field steps are compared, respectively.

results are in agreement with the RMSD distributions (Figures. 4.5-4.6) and the formerly investigated weights. There is a minor impact on the pair-overlap between a cMD and aMD trajectory from the same starting structure using more mean-field iterations, but

the first and the converged steps lie within the error of the Maclaurin result. All overlap values are very low or even zero and thus not representative for the weight convergence analysis.

In Table 4.6, there are the density overlap values for different mean-field steps between the pairs of trajectories cMD, aMD and sMD. The values are the average between overlaps of single reference trajectories $f_{\text{dens}}(k, L; r)$, the errors correspond to the range between both values. It is remarkable that for the aMD results the converged mean-field $\text{MF}^{(\infty)}$ re-weighting yield almost identical values as for the non-weighted case. This does not need to be an error or indication that $\text{MF}^{(\infty)}$ yields generally wrong results, because it is theoretically possible that they re-weight the densities correctly, but for the r-neighborhoods of the reference frames κ, this leads still to similar densities for Met-Enkephalin. These densities are a comprehension of different frames with their different weights. Nevertheless, we cannot ensure that $\text{MF}^{(\infty)}$ with the selected r values does not only equalize the weights by smoothing. Thus, they should be used with care for aMD runs without proper validation. The results are significantly different for sMD re-weighting as already indicated by the weights in Fig. 4.13. The weights converge to realistic magnitudes just taking the influence of shared frames between different r-neighborhoods into account (see subsection 3.2.4). There are two interesting outcomes: First, O_{dens} is consistently lower for $\text{MF}^{(\infty)}$ than for $\text{MF}^{(1)}$ if sMD runs are involved. Second, the non-weighted sMD trajectories produce always larger density overlaps O_{dens} for the evaluated trajectory combinations. The reason can have different origins and cannot be clearly detected. The density overlap between both cMD trajectories is also below 60% (Table 4.6), which shows that both do not sample the converged equilibrium density. It might be that sMD samples better the underlying energy landscape due to decreased energy barriers but are not converged, yet. This results in a more uniform density because different energy minima are easier reachable, which would then lead to an increase of non-weighted overlap between sMD trajectories. On the other hand, maybe the sampled density of one cMD is more concentrated on one part of the conformational space, the other cMD on another part. This behavior leads then to an increase of the overlap between the non-weighted sMD and the corresponding cMD which favors one part that is intensively sampled by the distorted sMD run. But the corrected/re-weighted density does not overlap due to unconverged cMD.

As we already discussed, the re-weighting is a difficult task. On the one hand, one tries to be the most accurate, i.e. taking the weights as unchanged as possible but then also the errors have their full impact in the results. On the other hand, one wants to

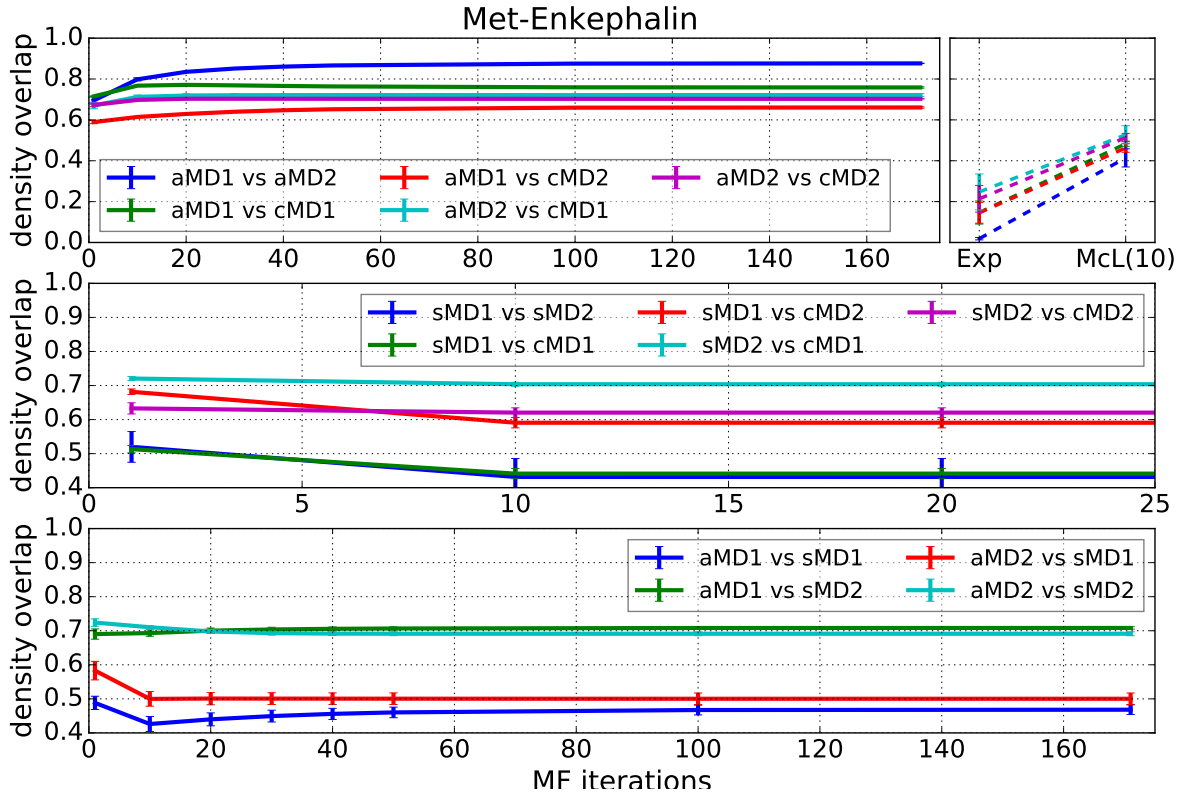

Fig. 4.14.: Density overlap O_{dens} for different re-weighting schemes of Met-Enkephalin. Values refer to all pairs of six 200 ns-trajectories combining cMD, aMD and sMD with different starting structures *Met79* and *Met153* for different MF iteration steps, Exp and McL up to order 10. cMD1/aMD1/sMD1 refers to the first and cMD2/aMD2/sMD2 to the second starting conformation. The MF weights correspond to $r = 0.11$ nm.

keep the errors as small as possible, i.e. approximating microstates or expanding the exponential function with possible smoothing the weights which might lead to biased potentials. Both extremes yield wrong results, thus one has to keep the balance between both ways. We want to be as critical and conservative as possible, to not overestimate the quality of sampling. If one wrongly concludes that the sampling is good, one will totally disqualify the assessment tool, because all following results will be based on wrong assumptions. We saw that the first mean-field step already has an impact in the relative weights between frames but also are compatible with the large peaks of Exp or McL. The converged $\text{MF}^{(\infty)}$ weights for aMD successively decrease the amplitudes of single weights. Hence, there is the risk of underestimating the weights of several frames and overestimating the overlap. In fact, for the investigated trajectory combinations, there was (almost) no difference between the non-weighted and converged $\text{MF}^{(\infty)}$. For this reason, we use $\text{MF}^{(1)}$ at the same threshold r for the re-weighting of aMD and sMD

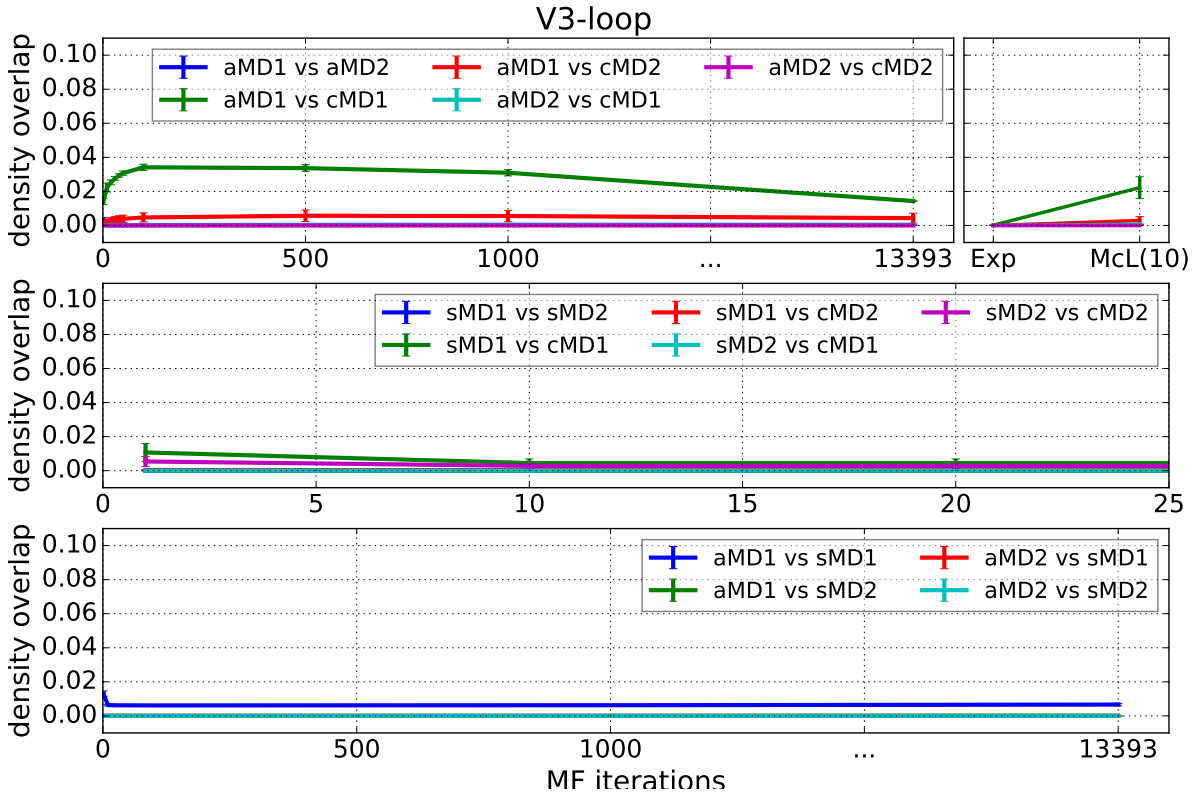

Fig. 4.15.: Density overlap $O_{\mathbf{dens}}$ for different re-weighting schemes of V3. Values refer to all pairs of six 200 ns-trajectories combining cMD, aMD and sMD with different starting structures *V3a* and *V3b* for different MF iteration steps, Exp and McL up to order 10. cMD1/aMD1/sMD1 refers to the first and cMD2/aMD2/sMD2 to the second starting conformation. The MF weights correspond to $r = 0.35$ nm.

trajectories as for the overlap calculation, if not specified otherwise. This shall serve as first approximate for the overlap measures and highlight that the converged mean-field weights are an interesting point to be rigorously and completely validated in the course of studying the re-weighting of biased MD runs. But this shall not be the main focus of this thesis. Moreover, we want to apply accelerated sampling to reveal, whether the sampling obtained by conventional MD simulations is sufficient, because enhanced techniques will show the uncertainties very quickly if they find more undetected conformations.

The influence of converged $\mathrm{MF}^{(\infty)}$ on the overlap analysis will be briefly discussed later in section 4.7.

Table 4.6.: O_{dens} values for different re-weighting schemes of Met-Enkephalin. Values refer to the following configurations: after first step $\text{MF}^{(1)}$, after ten steps $\text{MF}^{(10)}$, after mean-field convergence $\text{MF}^{(\infty)}$ and without re-weighting for trajectory pair combinations aMD with aMD, aMD with cMD, sMD with cMD and cMD with cMD. The values are the average between overlaps of single reference trajectories $f_{\text{dens}}(k, L; r)$, the errors correspond to the range between both values.

	$\text{MF}^{(1)}$	$\text{MF}^{(10)}$	$\text{MF}^{(\infty)}$	non-weighted
aMD1 vs. aMD2	0.6951 ± 0.0177	0.7986 ± 0.0079	0.8764 ± 0.0011	0.8765 ± 0.0011
aMD1 vs. cMD1	0.7152 ± 0.0018	0.7675 ± 0.0016	0.7587 ± 0.0035	0.7587 ± 0.0035
aMD1 vs. cMD2	0.5892 ± 0.0077	0.6143 ± 0.0012	0.6608 ± 0.0033	0.6608 ± 0.0032
aMD2 vs. cMD1	0.6703 ± 0.0160	0.7126 ± 0.0092	0.7214 ± 0.0080	0.7215 ± 0.0081
aMD2 vs. cMD2	0.6731 ± 0.0087	0.6977 ± 0.0010	0.7029 ± 0.0011	0.7029 ± 0.0011
sMD1 vs. sMD2	0.5200 ± 0.0453	0.4317 ± 0.0541	0.4316 ± 0.0541	0.6554 ± 0.0267
sMD1 vs. cMD1	0.5133 ± 0.0108	0.4414 ± 0.0150	0.4414 ± 0.0150	0.6069 ± 0.0043
sMD1 vs. cMD2	0.6816 ± 0.0083	0.5908 ± 0.0152	0.5908 ± 0.0152	0.7831 ± 0.0016
sMD2 vs. cMD1	0.7206 ± 0.0062	0.7038 ± 0.0058	0.7038 ± 0.0058	0.7284 ± 0.0032
sMD2 vs. cMD2	0.7206 ± 0.0062	0.7038 ± 0.0058	0.7038 ± 0.0058	0.7284 ± 0.0032
cMD1 vs. cMD2				0.5977 ± 0.0090

4.4. Overlap measures

We will focus on the overlap measures in this section. The large advantage of our approach is the possibility to analyze the overlap between two up to theoretically infinite trajectories at once. Additionally, it is possible to group different trajectories together, i.e. multiple independent trajectories are concatenated to a super-trajectory which can then be compared with others to enlarge the sampled conformational space. This will be reflected in the comparison set of trajectories L as discussed in subsection 3.2.3. For re-weighting, we will consistently use the first mean-field iteration $\text{MF}^{(1)}$, as discussed previously.

Different groups of trajectories are interesting to be investigated: The overlap between all single trajectories at once, denoted as "ALL", will represent the hardest criterion which has to be fulfilled for complete sampling ($L = \{l_1, l_2, ..., l_n\}$). Remember that only if one single trajectory samples a different space than the other runs, the overlap will be zero. On the other hand, the influence of the sampling algorithm and also the starting conformation can reveal important information about the sampling. Hence, we investigate the overlap between all trajectories coming from one sampling method (denoted as "cMD", "aMD" and "sMD") and dividing these groups further to contain only trajectories from one

starting structure. Finally, we look into the different pair-overlaps between combinations of two trajectories.

4.4.1. Influence of r on the overlap measure

The first interesting point is the overlap as a function of the threshold r [37]. As already introduced in subsection 3.1.2, the overlap should quickly converge between r_{min} and r_{max} describing a convex curve. As minimal and maximal values, we use the values defined by the 99% of the RMSD distributions in subsection 4.2.1. To have a representative set of trajectories for both molecules, we will analyze all 42 $\times$ 200 ns trajectories, 7 for each combination of starting conformation and conventional or enhanced sampling. Furthermore, we use the same set of trajectories for the comparison and the references $K = L$.

The overlaps (conformational and density) as a function of the threshold r are illustrated in Fig. 4.16 for both molecules investigating the overlap between all trajectories and sampling algorithm subgroups. The conformational overlap O_{conf} shows a clear deviation between both molecules. In section 3.4, we discussed that O_{conf} is the necessary criterion for complete sampling. For Met-Enkephalin, the curves show almost perfect convergence for this parameter, where O_{conf} is constantly 1 for $r \geq 0.1$ nm (blue traces in left of Fig. 4.16). Remarkably, this is also true using all trajectories. Thus, one can assume that the trajectories are in the regime where the density equilibrates between conformations, because already all conformational clusters are found. For V3, the behavior is completely different, showing concave curves starting to give non-zero overlap values at around $r = 0.6$ nm (red traces in left of Fig. 4.16). Considering the RMSD distributions, $r = 0.6$ nm is already a quite large value where one cannot distinguish whether the overlap is not trivially increased due to a too tolerant threshold. The same behavior is true for V3 considering O_{dens}, where the overlap is increased firstly for $r \gtrsim 0.8$ nm. Met-Enkephalin has still convex curves for O_{dens} for the given groups, but for instance at $r = 0.11$ nm the overlap is only between 20 to 40%. Only at very coarse resolutions around $r \approx 0.2$ nm the overlap reaches $0.7 < O_{\text{dens}} < 0.9$. Hence, the density is far away from reaching convergence at an acceptable resolution. Nevertheless, one can see that aMD and sMD consistently perform better except for $r < 0.1$ nm where the mean-field re-weighting is in the regime of Exp for aMD and single frames in r-neighborhoods for sMD (compare subsection 4.3). The averaged overlaps $\Omega_{\text{conf}}, \Omega_{\text{dens}}$ evaluate the area under the curves, and can immediately show tendencies between different groups, like the aMD performance. But still, a value of $\Omega_{\text{dens}} \approx 0.7$ does not correspond to a satisfactory sampling.

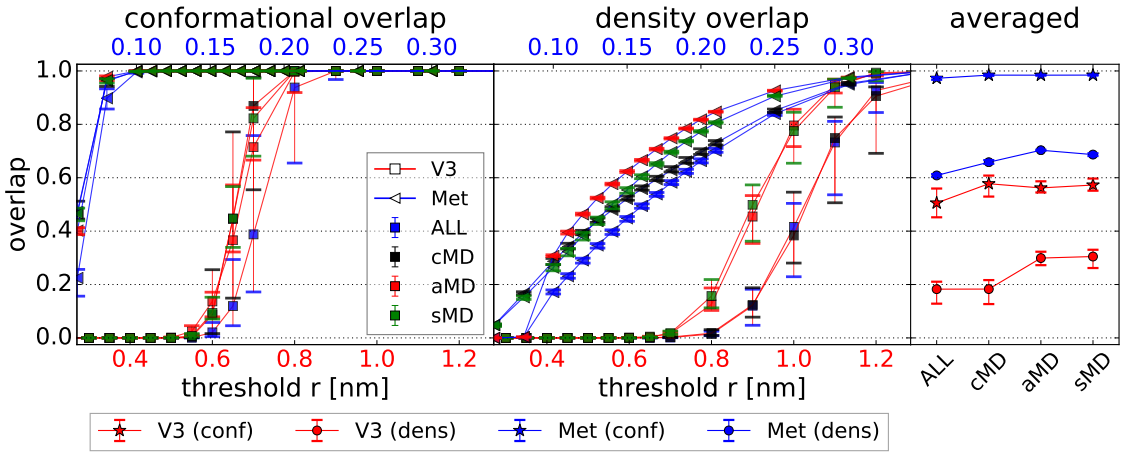

Fig. 4.16.: Overlap measures $O_{\mathbf{conf}}$ (left), $O_{\mathbf{dens}}$ (middle) as a function of the threshold r, and averages $\Omega_{\mathbf{conf}}, \Omega_{\mathbf{dens}}$ (right). Met-Enkephalin (blue and triangle markers) and V3 (red and square markers). Different analysis groups are applied: "ALL" refers to single overlaps between all 42 × 200 ns trajectories, whereas for each method (cMD, aMD, sMD) 14×trajectories were evaluated, 7 of each of the two starting structures. The figure is taken from Ref. [37].

Why is this the case for a relatively small molecule? Do single trajectories alter the results of the chosen analysis groups, because some runs failed? To answer these questions, it is reasonable to look into the pair-overlap between all trajectory combinations, to detect which sampling methods or trajectories are responsible for this result. This is done by using an asymmetric heatmap of all pair-overlap combinations with $L = \{l_X, l_Y\}$, with X, Y are one of the 42 × 200 ns trajectories. We do not use the same reference set but $K = \{l_X\}$ for the lower and $K = \{l_Y\}$ for the upper triangular of the heatmaps (Figs. 4.17-4.18). This choice allows us to investigate the deviation of the sampled conformational space between sampling methods or single trajectories. It is expectable that using cMD trajectories as reference K, the calculated overlap is larger than if K corresponds to trajectories of enhanced sampling methods, because the latter should sample a larger space in the same simulation time.

Let us first consider the pair-overlaps of Met-Enkephalin at a reasonable resolution $r = 0.11$ nm (Fig. 4.17 left). The conformational overlap O_{conf} ranges between 0.99 to 1.00, which was expectable from the overlap curves in Fig. 4.16. But one can also see that the overall density pair-overlap O_{dens} is much larger than using all trajectories from the groups defined above. The main reason seems to be the pair-overlap between trajectories $29 - 35$ (sMD starting from *79*) and runs originating from other sampling methods. Remarkably, the pair-overlap between trajectories of cMD starting from *Met153* and sMD starting

from *Met79* contain the lowest values 0.48 and 0.50, respectively (Table 4.7). These sMD trajectories seem to sample similar probability densities but consistently different compared to others. Neither the number of found clusters nor the other overlap values can explain this behavior. On the other hand, there might be one outlier in the group of cMD starting from *Met153*, because others behave well. The pair-overlap of aMD trajectories ranges from 0.61 to 0.87 which is good for such a resolution, but there, one can see the impact of the harder criterion calculating the overlap of multiple trajectories, where the full group of aMD does not cross a value of $O_{\text{dens}} = 0.4$. It will be interesting to investigate the overlap as a function of r for other groups, whether the results from the pair-overlap can be reproduced.

The pair-overlaps for V3 at high ($r = 0.35$ nm) and lower resolutions ($r = 0.5$ nm, $r = 0.7$ nm) are illustrated in Figs. 4.17-4.18. The first threshold $r = 0.35$ nm might be the critical point where, for given reference frames $\kappa \in k$ of trajectory k, the normalized events of trajectories $l \neq k$ are non-zero and the normalized events of the trajectory k are not trivially 1, considering the RMSD distributions in Figs. 4.7-4.8. At this resolution, the analysis groups yield zero overlap for O_{conf} and O_{dens}. The reason is illustrated in the pair-overlap heatmaps. Only cMD trajectories starting from *V3a* cover the same areas of conformational space, which is even better visible for different reference K:

$$O_{\text{conf}}\left(K \in \{l_X\}, L \in \{l_X, l_Y\}; r = 0.35\text{nm}\right) \approx 1\,, \qquad X \in \text{cMD}$$
$$O_{\text{conf}}\left(K \in \{l_Y\}, L \in \{l_X, l_Y\}; r = 0.35\text{nm}\right) \approx 0\,, \qquad Y \in \text{aMD}$$

It is highly probable that these trajectories are trapped in few conformational clusters, because other trajectories seem to sample completely different areas. This is supported by the numbers of found clusters for each 200 ns trajectory of a global clustering N_C^{global} which are illustrated below the heatmaps in Figs. 4.7-4.8. N_C^{global} refers to a global clustering involving all shown trajectories at once. $N_{C,\text{cMD}}^{\text{global}}$ is two to three times smaller than $N_{C,\text{aMD}}^{\text{global}}$. Interestingly, the number of reached clusters by single sMD trajectories are compatible with cMD, but still, the sampling is not comparable. Increasing the threshold to $r = 0.5$ nm and $r = 0.7$ nm (Fig. 4.18) increases O_{conf}, but the result is not comparable to O_{conf} of Met-Enkephalin at high resolution. The number of clusters N_C^{global} and O_{conf} show the underlying behavior which still lead to very low overlap values except for cMD starting from the first conformation. This might be an indicator for a huge conformational space of V3 which is far from being sampled exhaustively using 200 ns trajectories, although a simulation time of about 100 ns is a typical timescale in current MD simulations.

As mentioned, we investigate further analysis groups shown in Fig. 4.19 starting again

with Met-Enkephalin. Here, the overlap between trajectories of one sampling method and one starting conformation is monitored, together with the combination of cMD and the trajectories of each enhanced sampling method. From the pair-overlap, it was expected for Met-Enkephalin that O_{dens} of aMD trajectories of each starting structure will outperform the other sampling methods. This is true for all thresholds except the discussed regime $r \leq 0.1$ nm (Fig. 4.19 top center). The influence of the bad pair-overlaps of cMD from *Met153* and sMD from *Met79* are shown in bottom center of Fig. 4.19, where the combination of cMD and aMD (or sMD) give comparably worse results than for the combination of cMD and aMD trajectories starting only from *Met79*. It is remarkable that, still for such a small molecule, the sampling of the 200 ns trajectories reveals some sort of dependence from their starting conformation.

For V3 trajectories, we see the drastic difference between cMD trajectories starting from *V3a* and the rest (see Fig. 4.19). O_{conf} and O_{dens} are significantly increased for cMD from *V3a* but still the curves indicate incomplete sampling by their shape. Only the overlap results of sMD trajectories starting from the same conformation *V3a* might also be increased. The most remarkable thing is that trajectories from *V3a* give consistently significantly better results than *V3b* which is also visible in the corresponding average overlap. This is an indication for trapped behavior. Interestingly, this can be linked to the picture of RMSD values obtained after the MD preparation (Fig. 4.4, see also subsection 4.1.4), which are much more conserved than for the second starting conformation. The main reasons for the large errorbars for V3, especially for O_{conf}, are the different results of reference sets K of different sampling methods.

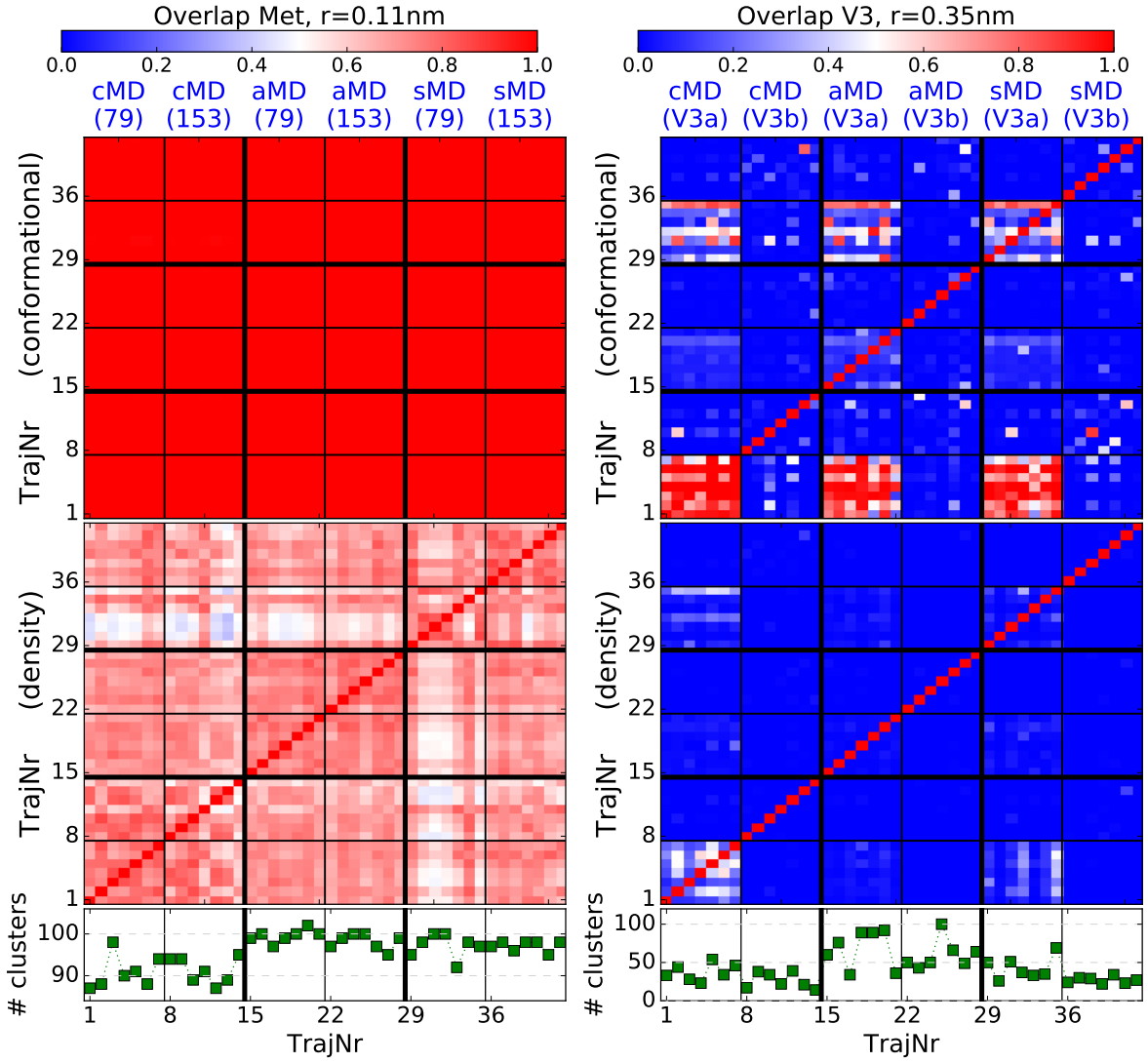

Fig. 4.17.: Pair-overlap heatmaps for Met-Enkephalin at $r = 0.11$ nm (left) and V3 at $r = 0.35$ nm (right). $O_{\text{conf}}, O_{\text{dens}}$ between all pairs of 42×200 ns trajectories and N_C^{global} for each trajectory. They are split into blocks of 7 for each sampling method and starting conformation indicated as blue labels. The heatmaps are asymmetric, whereas the lower triangular matrices correspond to $K = \{l_X\}, L = \{l_X, l_Y\}$, and the upper triangular to $K = \{l_Y\}, L = \{l_X, l_Y\}$. The figure is reproduced from Ref. [37].

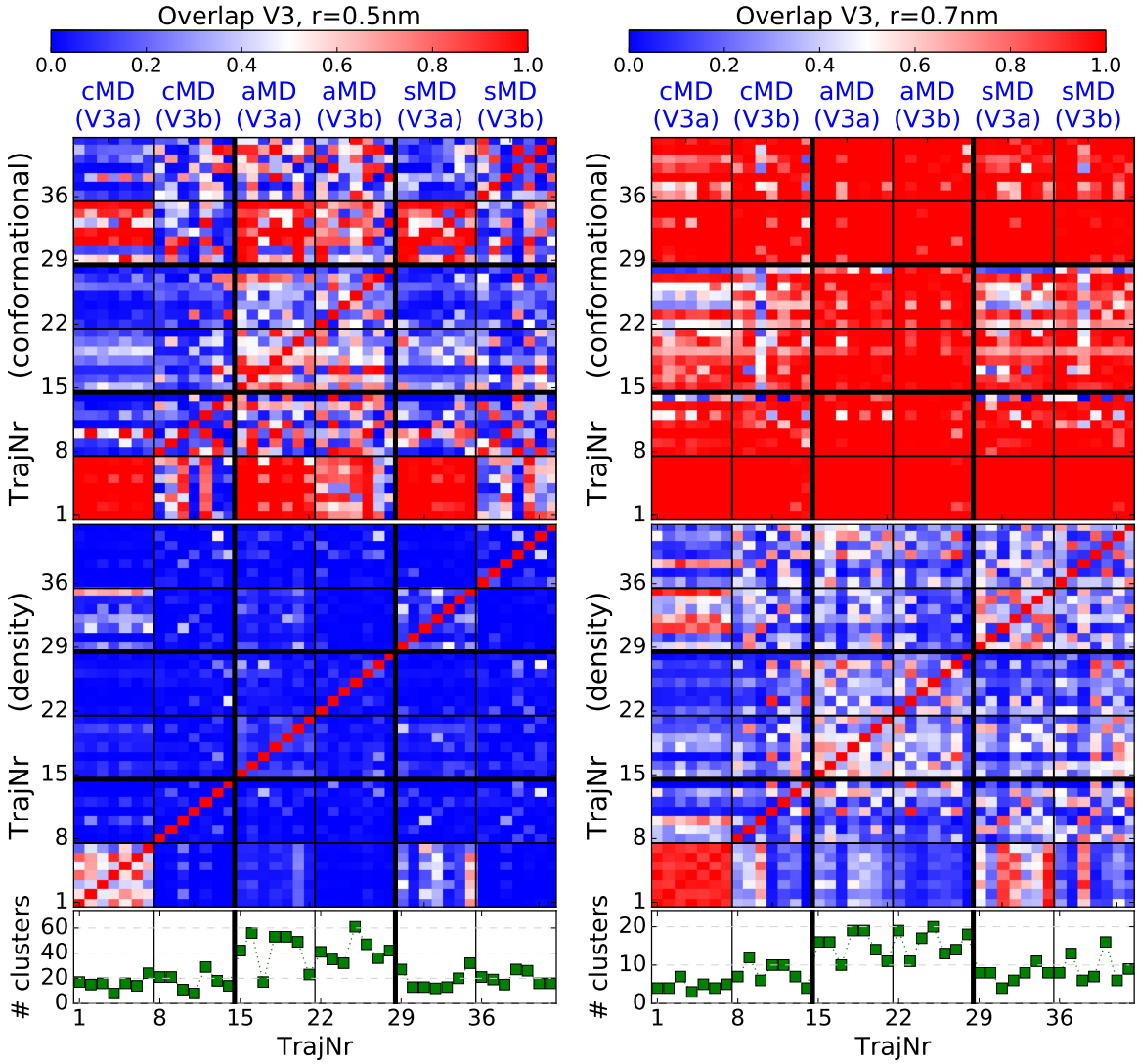

Fig. 4.18.: Conformational and density overlap $O_{\mathbf{conf}}, O_{\mathbf{dens}}$ between all pairs of 42×200 ns trajectories for V3 at $r = 0.5$ nm (left) and $r = 0.7$ nm (right) and the corresponding number of clusters $N_C^{\mathbf{global}}$ for each trajectory. The trajectories are split into blocks of 7 for each sampling method (cMD, aMD, sMD) and starting conformation indicated as blue labels. The heatmaps are asymmetric, whereas the lower triangular matrices correspond to the overlap between trajectories l_X and l_Y with $K = \{l_X\}, L = \{l_X, l_Y\}$, and the upper triangular vice versa with $K = \{l_Y\}, L = \{l_X, l_Y\}$. The figure is reproduced from Ref. [37].

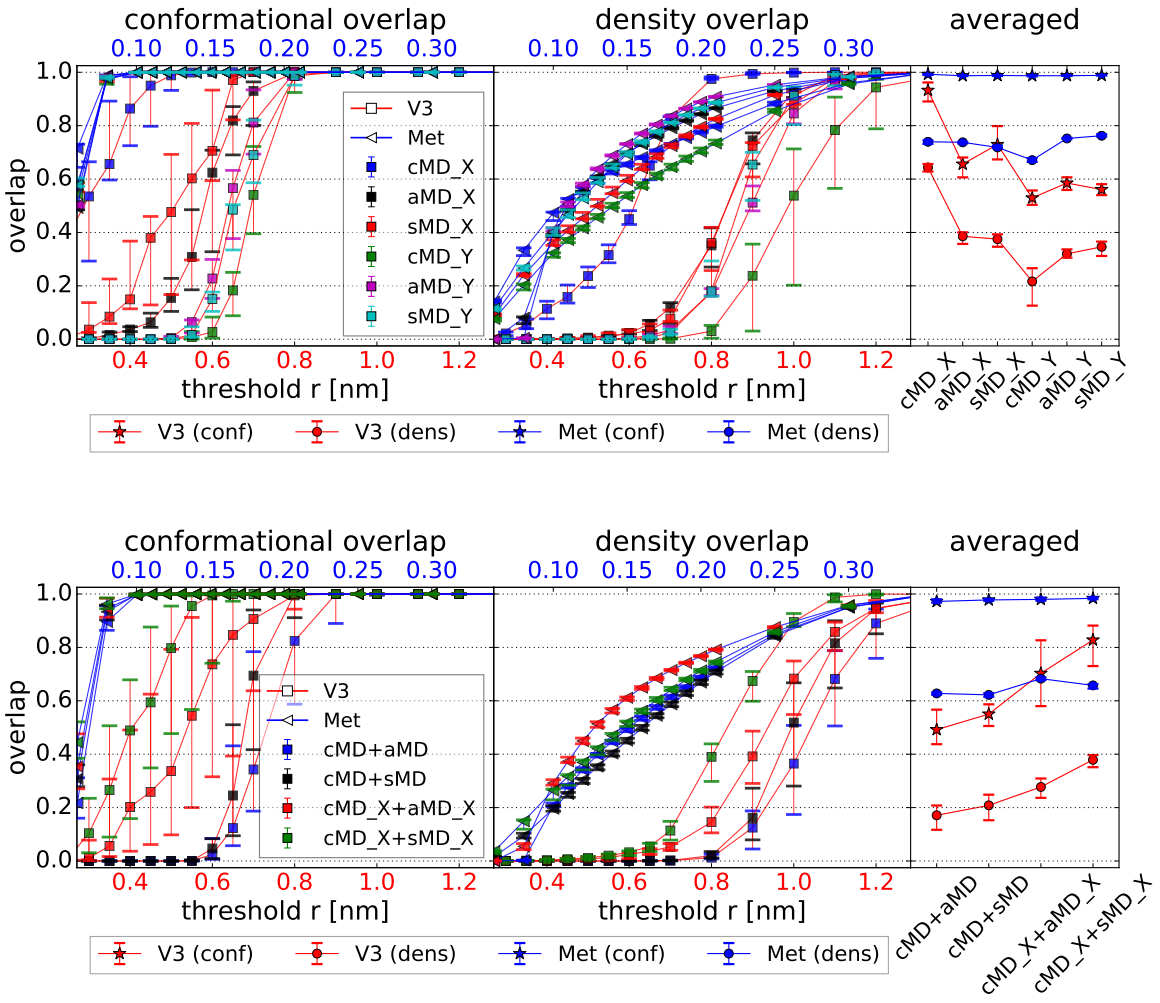

Fig. 4.19.: Overlap measures $O_{\mathbf{conf}}$ (left), $O_{\mathbf{dens}}$ (middle) as a function of the threshold r, and averages $\Omega_{\mathbf{conf}}, \Omega_{\mathbf{dens}}$ (right) for different groups. Met-Enkephalin (blue and triangle markers) and V3 (red and square markers). Different analysis groups are applied: (top row) seven 200 ns trajectories originating from each sampling method and starting conformation (X = *Met79* or *V3a*, Y = *Met153* or *V3b*). The figure is reproduced from Ref. [37]; (bottom row) 28 trajectories combining cMD and aMD or sMD, and the same groups with trajectories coming only from the first starting conformation.

Table 4.7.: Minimal and maximal pair-overlap values $O_{\mathbf{dens}}^{(\mathbf{min})}, O_{\mathbf{dens}}^{(\mathbf{max})}$. The certain groups correspond to Fig. 4.17 (left). Pair-overlap between same trajectories is not taken into account.

	cMD79	cMD153	aMD79	aMD153	sMD79	sMD153	cMD + aMD	cMD + sMD
min	0.66	0.48	0.68	0.61	0.50	0.63	0.54	0.37
max	0.85	0.83	0.82	0.82	0.87	0.87	0.80	0.82

4.4.2. Influence of the simulation time t on the overlap behavior

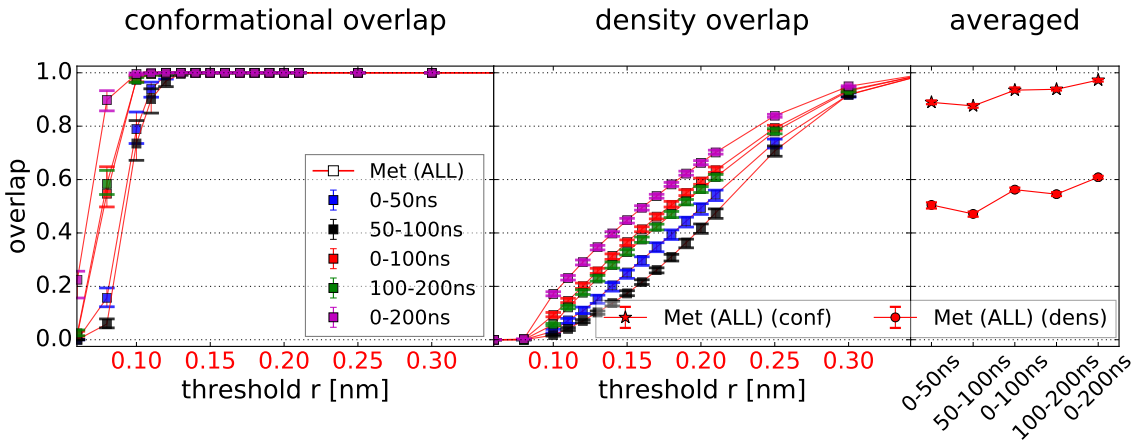

Fig. 4.20.: Overlap measures $O_{\mathbf{conf}}$ (left), $O_{\mathbf{dens}}$ (middle) as a function of the threshold r, and averages $\Omega_{\mathbf{conf}}, \Omega_{\mathbf{dens}}$ (right) for Met-Enkephalin for different simulation times t. "ALL" is used as analysis group, referring to single overlaps between all 42×trajectories.

One central parameter of MD runs is the simulation time t and the estimation of the necessary time to reach convergence of trajectories. On the other hand, it is interesting to extract the behavior of different time-windows of the simulation. For instance, do first parts of the trajectory behave differently compared to last parts? This might also be useful in preparation processes of very complex or large systems, where one might detect the simulation time necessary to overcome physically meaningless interactions introduced by artificial starting conditions [70]. In such cases, one might be able to detect significantly low overlap for first parts of the trajectories compared to later simulation parts, where the system changes to equilibrium states. Disregarding these first part of the simulation

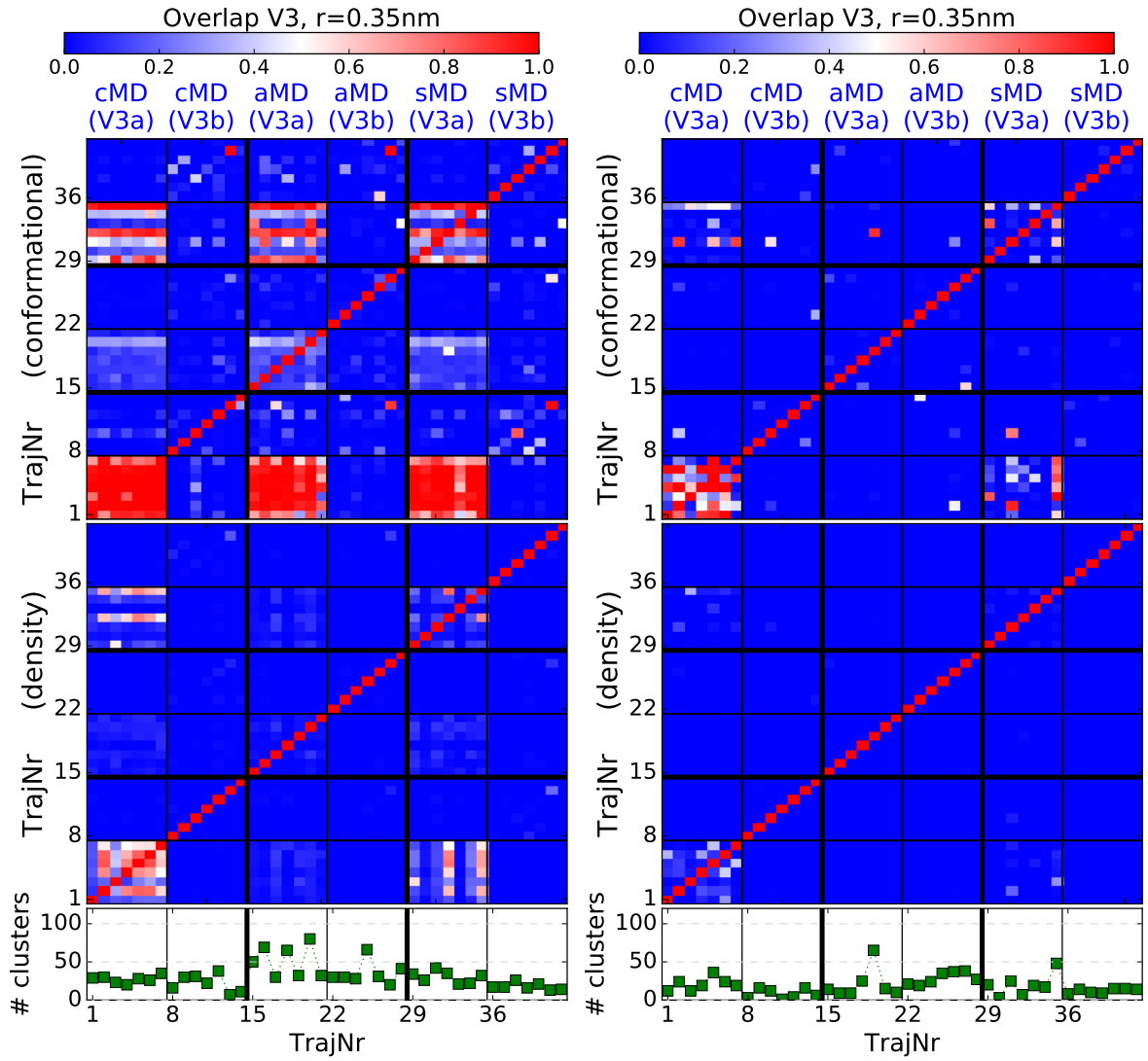

Fig. 4.21.: Conformational and density overlap $O_{\mathbf{conf}}, O_{\mathbf{dens}}$ between all pairs of $42\times$trajectories for V3 at $r = 0.35$ nm for simulation times $0 - 100$ ns (left) and $100 - 200$ ns (right) and the corresponding number of clusters $N_C^{\mathbf{global}}$ for each trajectory. The trajectories are split into blocks of 7 for each sampling method (cMD, aMD, sMD) and starting conformation indicated as blue labels. The heatmaps are asymmetric, whereas the lower triangular matrices correspond to the overlap between trajectories l_X and l_Y with $K = \{l_X\}, L = \{l_X, l_Y\}$, and the upper triangular vice versa with $K = \{l_Y\}, L = \{l_X, l_Y\}$.

time should give significantly better results, because the unphysical configurations will rarely be reproduced by independent MD runs.

We will first investigate the overlap as a function of threshold r for different simulation times t of the overlap between all 42 trajectories of Met-Enkephalin. In Fig. 4.20, one can see that the overlap is consistently better for first parts of the trajectories but are

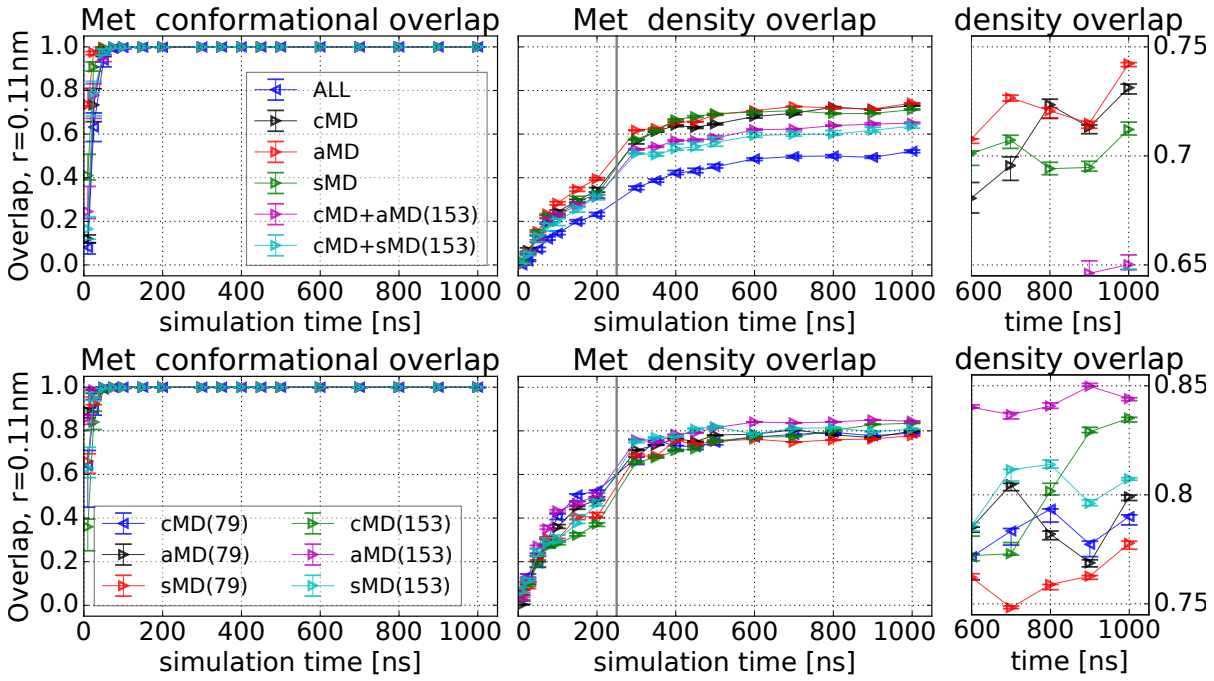

Fig. 4.22.: Conformational and density overlap as a function of different simulation times t for Met-Enkephalin at $r = 0.11$ nm. Up to 200 ns, there are seven trajectories for each combination of sampling method (cMD, aMD, sMD) and starting conformation (*Met79*, *Met153*). For simulation times larger than 200 ns, only three trajectories for each combination are evaluated. Top row: Analysis group "ALL" with five another groups with the same number of trajectories for each sampling method alongside with the combination of trajectories from two sampling methods starting from *Met153*. Bottom row: Overlap between all single trajectories for each combination of starting structure and sampling method.

outperformed by the full 200 ns. The combination of different windows is necessary to obtain the overlap values for the full 200 ns. Hence, also first parts are important and produce relevant overlap values although the starting structures after MD preparation (Fig. 4.3) are (almost all) totally different. Still, one would say that different parts behave very similar, which is shown by the two time-windows $0-100$ ns and $100-200$ ns. This is totally different for V3. We saw in the previous section that the evaluation of the overlap between all single 42 trajectories is zero, thus we show the pair-overlaps between different time-windows for V3 at the high resolution $r = 0.35$ nm in Fig. 4.21. For the 200 ns trajectories (Fig. 4.17), $O_{\text{conf}}, O_{\text{dens}}$ were negligibly small except for cMD starting from *V3a*, but the number of clusters found of the aMD trajectories of the global clustering were on average 2 to 3 times larger than for the other sampling. Interestingly, considering simulation times between $0-100$ ns, all non-zero overlap values are increased compared to $0-200$ ns. For the second halves of the trajectories $100-200$ ns, the number of

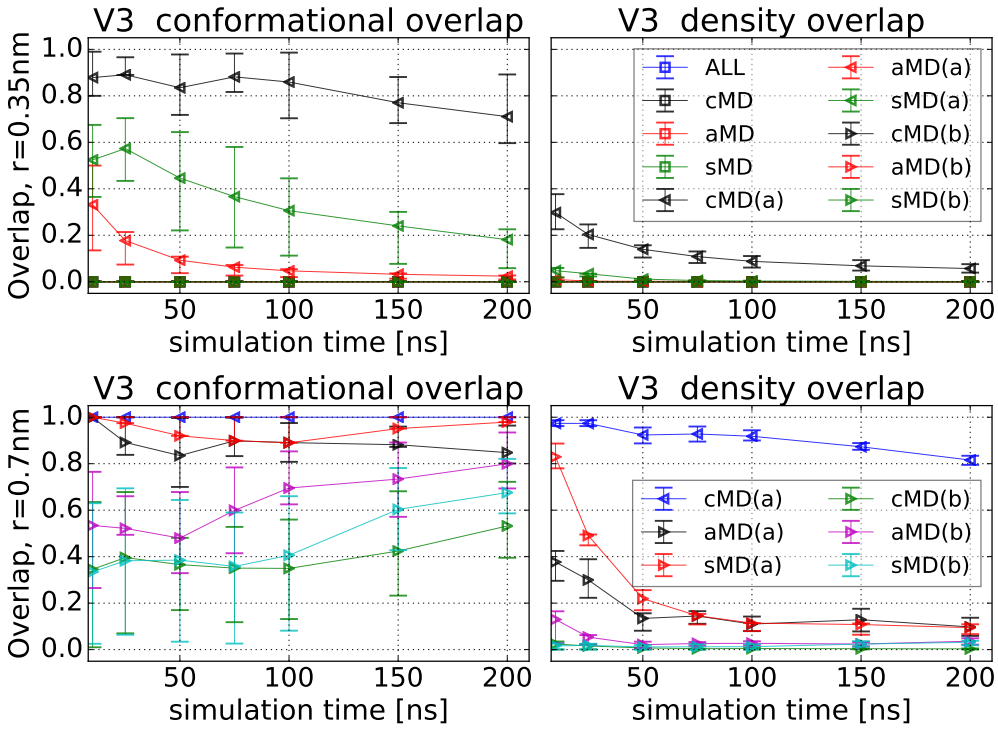

Fig. 4.23.: Conformational and density overlap as a function of different simulation times t for V3 at $r = 0.35$ nm (top row) and $r = 0.7$ nm (bottom row). For each combination of starting conformation (*V3a*, *V3b*) and sampling method (cMD, aMD, sMD) seven trajectories are evaluated. Analysis groups "cMD", "aMD", "sMD" hold 14 trajectories and "ALL" is the combination of all 42×trajectories. The figure corresponds to Ref. [37].

clusters are comparable for all sampling methods, and the overlap is much smaller for the non-zero values compared to $0 - 200$ ns. The reason might be that in the first part of the simulation, the conformations are similar due to the more conserved starting structure after preparation for *V3a* (Fig. 4.4). After a certain critical point t_{crit}, these runs lose the information about their origin and sample into different more distinct regions due to the huge conformational space. The second halves of trajectories from *V3a* behave similar to the other simulations according to their overlap, only in O_{conf} of cMD the starting influence is still present. This reveals that one needs multiple hundreds of nanoseconds simulation time just to be sure to lose the influence from the starting structure. Thus, it is very dangerous to rely on simulations of about 100 ns that only start from one initial conformation describing a certain very flexible molecule.

In Fig. 4.22, the overlaps $O_{\text{conf}}, O_{\text{dens}}$ are shown as a function of the simulation time t for Met-Enkephalin [37]. O_{conf} immediately reaches a value of one within the first 100 ns. Thus, the conformational space can be reached well at the high resolution of $r = 0.11$ nm in

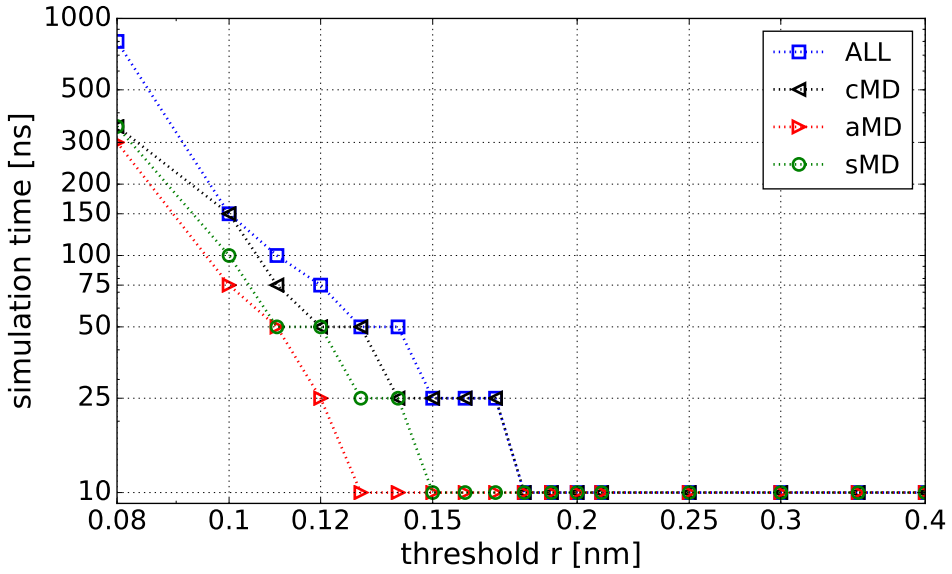

Fig. 4.24.: Simulation time t in [ns] as a function of the threshold r in [nm] which was necessary to obtain $O_{\mathbf{conf}} \geq 0.99$ for Met-Enkephalin. For simulation times up to 200 ns, 42 trajectories are evaluated ("ALL") which are further split into 14 trajectories of each sampling method cMD, aMD, sMD. For simulation times above 200 ns, we use $6 \times 1\mu s$ trajectories for each method. Scaling is log-log.

timescales which are typical of current MD simulations. However, the sampled equilibrium density still needs simulation lengths of the order of several μs to converge toward one. The curve denoted as "ALL" incorporates all trajectories of all sampling methods and starting conformations. It reaches only a value of $O_{\text{dens}} \approx 0.5$ for $1\mu s$. We discussed that this is the hardest quality criterion. To compare the different sampling methods, we use five different analysis groups ("cMD", "aMD", "sMD", "cMD+aMD(153)", "cMD+sMD(153)") which all incorporate the same number of trajectories, in order to be comparable. The latter two combine trajectories of two sampling methods, both starting from *Met153*, because they have the largest overlap of different combinations of algorithms.

Remarkably, the overlaps between trajectories of the same sampling method, cMD and aMD, are comparable. Both sampling methods give larger overlap values compared to sMD. But the combinations of cMD with the other sampling methods lead to significantly lower density overlap values. It will be interesting to investigate the reached clusters in the next section to determine the reason for this uncertainties. Furthermore, we will investigate the results between different sampling methods in section 4.7 regarding the question, whether the enhanced algorithms are properly re-weighted or yield biased ensembles. Two important things have to be mentioned. First, up to 200 ns, there are 42 independent trajectories, and above 200 ns, the overlaps correspond to the evaluation of

18 independent $1\mu s$-trajectories for the combination of three runs per sampling method and starting conformation. This explains the jump in the overlap between 200 and 300 ns, because the criterion for the overlap is less strict with less trajectories. Second, the larger O_{dens}, the slower is the convergence behavior, because it is harder to obtain the correct probability density function $p(\vec{r})$ in the course of the MD simulation. Remember that for instance 18 different independent trajectories have to sample in the same simulation time window the same energy wells with the same probability to further increase the strict density overlap criterion if already large overlap values are reached. This is the reason, why some overlap values go down for some times t and rise again later on (Fig. 4.22).

For V3, the simulation time of 200 ns is far away from even reaching converged values of O_{conf} [37]. In Fig. 4.23, the overlaps as a function of the simulation time t are illustrated for high resolution $r = 0.35$ nm and low resolution $r = 0.7$ nm for different combinations of 42×200 ns trajectories. It is remarkable that O_{dens} is monotonously decreasing for the non-zero overlaps even for $r = 0.7$ nm. Hence in these simulation time regimes, the simulations are still finding more conformations than revisiting clusters which were already sampled before. For the lower threshold, again only trajectories from *V3a* have non-zero overlaps, which might be explained by trapped behavior. The longer the simulation, the more these trajectories find new conformational clusters and leave the conserved starting point. Another information which can be extracted from the blue curve of O_{dens} for $r = 0.7$ nm in Fig. 4.23 is the following: the sampled conformational space of cMD trajectories originating from *V3a* must be much smaller because a radius around the reference frames κ of $r = 0.7$ nm yields a huge density overlap in comparison to the other MD runs. The cluster analysis (next section) will shed light on this issue.

Finally, it is reasonable to extract the behavior between the simulation time t and the necessary threshold r, while the overlap is kept constant. We already know that higher resolutions (small r) require a strongly increasing sampling effort to visit all conformational clusters with the same relative frequency in a set of independent trajectories. We find that for Met-Enkephalin the simulation time required to achieve convergence of $O_{\text{conf}} \geq 0.99$ as a function of r follows approximately a power-law function (Fig. 4.24), similar to the number of clusters N_C^{global} as a function of r (Fig. 4.9). This means, for an exponentially decreasing threshold r, an exponentially longer simulation is needed to obtain convergence in the conformational overlap. It is therefore highly advisable to determine, which spatial resolution is necessary for the underlying system of interest, to approximate the necessary simulation time for convergence.

4.5. Clustering analysis

We use the clustering as complementary tool together with the overlap measures to quantify the sampling quality. The overlap measures give the information about whether different trajectories sample the same conformational space with similar densities. Only then, independent trajectories will reproduce the experiments. The problem is that if multiple trajectories are trapped within the same potential minima and only rarely cross the energetic barriers, the overlap will yield large values without detecting this issue. Thus, we use the clustering to detect the size of the conformational space and monitor, if trajectories sample only few conformations and are therefore trapped for certain simulation times. For this, we first investigate the development of the cluster number N_C as further indicator for convergence.

All clusterings are done using the second starting conformation as reference. Additionally, in the analysis of the cluster number as a function of the threshold r (Fig. 4.9), we found out that there is no unique clustering radius. Hence, we use the high resolutions, for Met-Enkephalin $r = 0.11$ nm and V3 $r = 0.35$ nm, if not specified otherwise.

4.5.1. Development of the cluster number N_C

The development of the cluster number N_C can give insights about the convergence of single trajectories, introduced in subsection 3.3.3. But first, we are interested in the size of the conformational space sampled by each trajectory, to compare the sampling and classify whether different trajectories might sample different regions.

The total number of reached clusters of each trajectory can be best analyzed with the global clustering (subsection 3.3.2). Therefore, we generate one global partitioning for Met-Enkephalin including all 18×1 μs, 24×200 ns and 6×100 ns trajectories from different starting conformations and sampling methods. All total numbers of found clusters of each trajectory $N_C^{\text{global,Met}}$ at given simulation time t are extracted from this global partitioning by detecting, how many clusters are reached in the given time of the specific MD run. Trajectories, which are shorter than the certain simulation time, are omitted. This workflow allows us to explicitly compare the reached size of each trajectory over the course of increasing simulation time t. The same approach is used for V3 including all 42×200 ns and 6×100 ns trajectories extracting $N_C^{\text{global,V3}}$ for each trajectory for given time t.

The results of $N_C^{\text{global,Met}}$ for trajectory sets from different sampling methods (cMD, aMD, sMD) for discrete timesteps t are shown in Fig. 4.25 as boxplots for single trajecto-

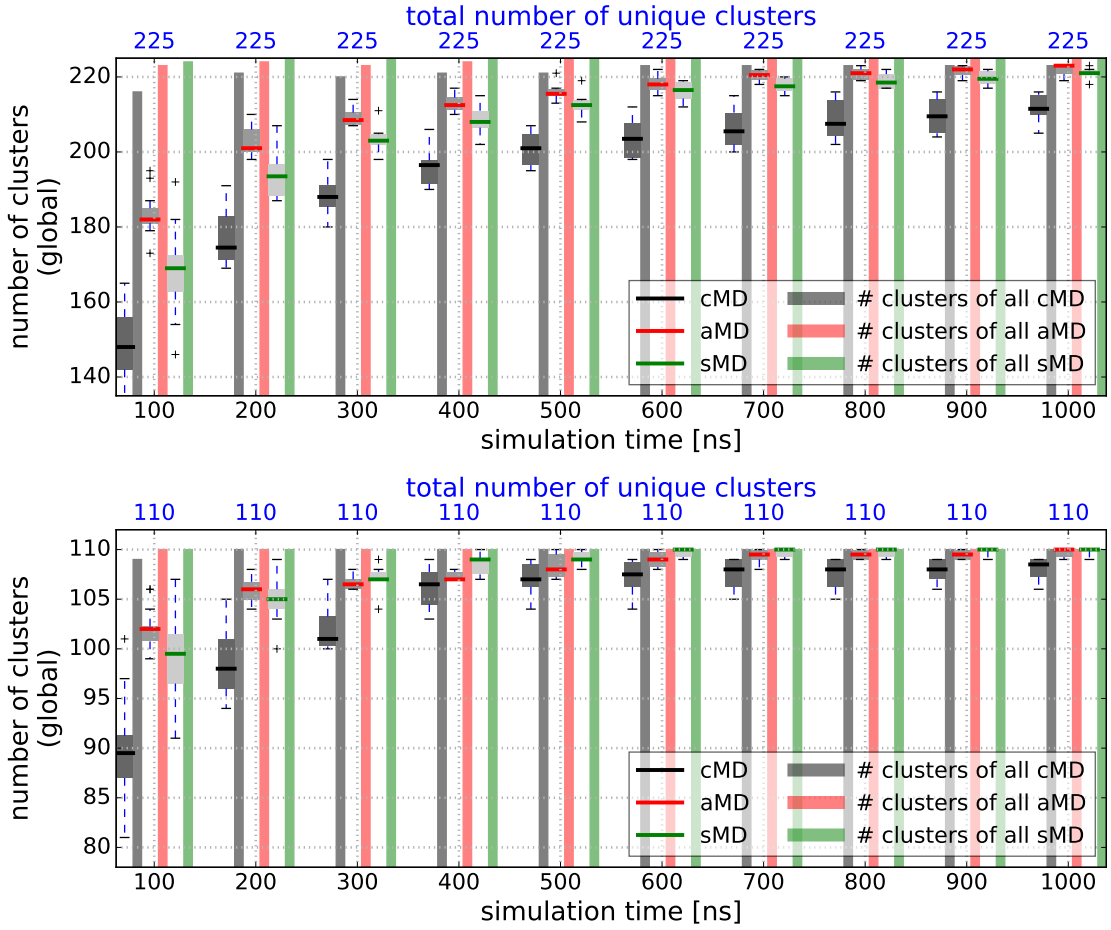

Fig. 4.25.: Number of clusters $N_C^{\text{global,Met}}$ for different simulation times. Trajectories of Met-Enkephalin are shown as boxplots for the three groups of sampling methods together with the number of clusters of the combined groups shown as bars. The second x-axis gives the total number of unique clusters combining all trajectories. Top: Clustering at $r = 0.1$ nm. Bottom: Clustering at $r = 0.11$ nm.

ries and bars for the combination of multiple trajectories at high resolutions $r = 0.1$ nm (top) and $r = 0.11$ nm (bottom). One can see the clear impact of the enhanced sampling techniques, where single trajectories consistently find more clusters compared to cMD. Remarkably, the unique number of clusters found by the combination of all cMD trajectories (black bars) is compatible with the other sampling methods, showing that each single cMD run did not converge yet for the high resolution of $r = 0.1$ nm. For $r = 0.11$ nm, the outcomes are similar, but the cluster numbers of cMD are more compatible to the others. For a threshold of $r = 0.11$ nm there is no clear evidence that, at a simulation time $t = 1$ μs, cMD trajectories should sample much worse than aMD or sMD because there is a deviation in the cluster number of ± 3 for single trajectories and the corresponding conformational overlap O_{conf} is one (Fig. 4.22). Nevertheless, there is a

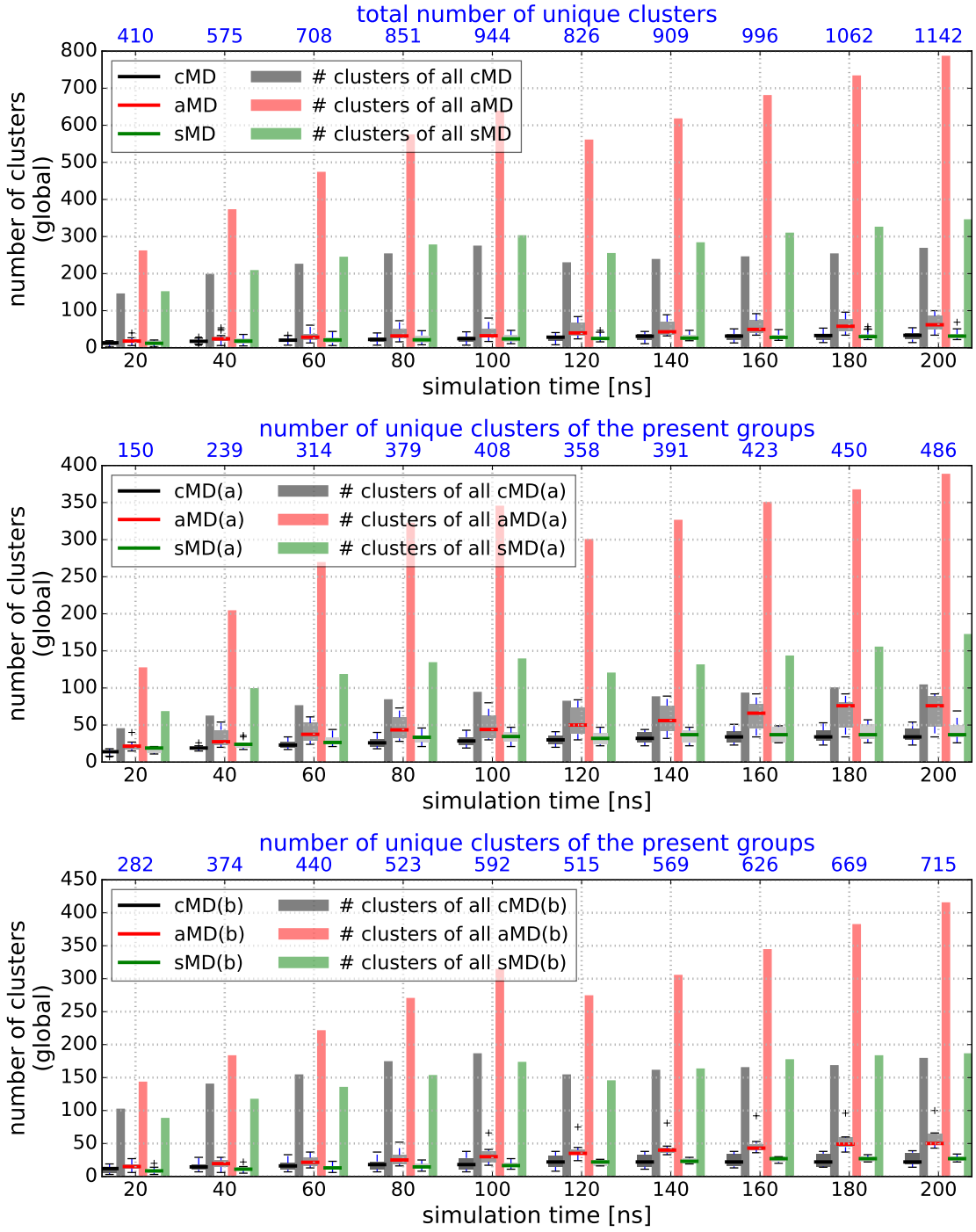

Fig. 4.26.: Number of clusters $N_C^{\mathbf{global,V3}}$ for different simulation times. Trajectories of V3 are shown as boxplots for the three groups of sampling methods (top) further split into starting structure *V3a* (center) and *V3b* (bottom). The number of clusters of the combined groups are shown as bars. The second x-axis gives the total number of unique clusters combining all trajectories of the given groups. The clustering was performed at $r = 0.35$ nm.

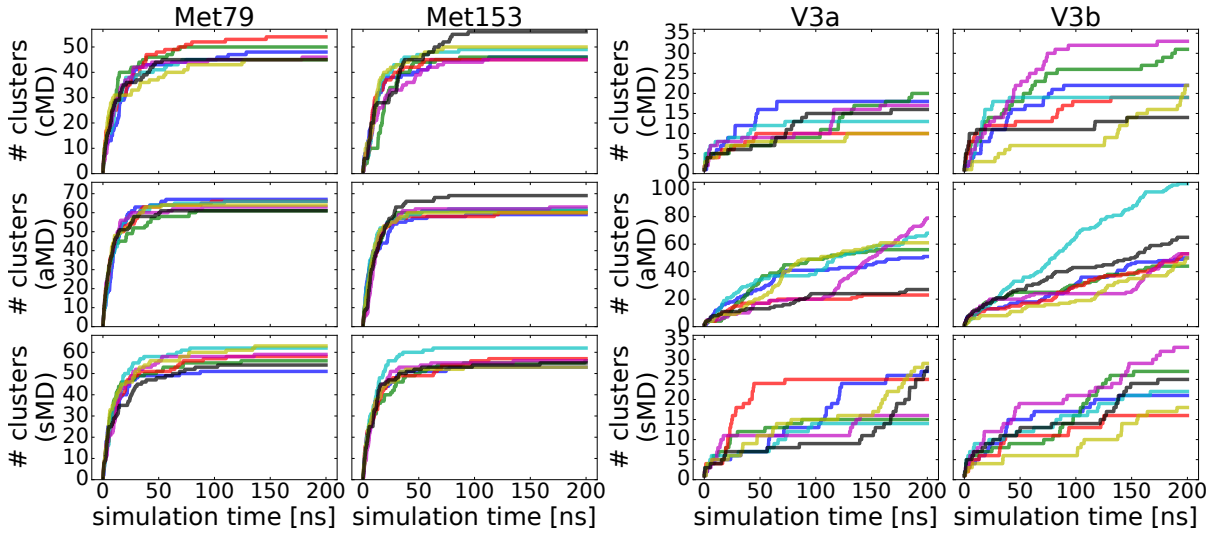

Fig. 4.27.: Development of the number of clusters N_C^{local} for single trajectories as a function of the simulation time t. Met-Enkephalin (left) and V3 (right). For each of the combination of sampling method and starting structure, seven 200 ns trajectories (distinguished by colors) were clustered separately at $r = 0.11$ nm for Met-Enkephalin and $r = 0.35$ nm for V3. The figure is taken from Ref. [37].

clear difference between the overlap of one sampling method and the combination of different sampling methods using the same amount of trajectories, where the first is clearly larger than the second case. This issue has to be resolved in further analysis below.

In Fig. 4.26, $N_C^{\text{global,V3}}$ is shown for discrete timesteps t at $r = 0.35$ nm for the trajectories of cMD, aMD and sMD, which are then split into subgroups involving only one of the two starting conformations. The outcome shows the huge conformational space and explains why the overlap measures are negligibly small. Although single trajectories of the same sampling method find a similar amount of clusters, these are almost completely different because the combinations of trajectories yield much more unique clusters. Here, we can see the biggest impact of aMD on the sampling, since it finds two to four times more clusters and clearly illustrates the failure of sampling convergence. Still, the total number of unique clusters is larger than the amount found only by all aMD trajectories, thus the other sampling methods also sample conformational space completely undetected by aMD. The reason might be that due to the lifted potentials and huge conformational space, some states may be skipped. However, sMD produces only more clusters for the first starting conformation *V3a* compared to cMD, which is a remarkable result because we expected these cMD trajectories to be trapped. For the second starting conformation *V3b*, cMD and sMD produce similar numbers with small benefit toward cMD, as well as for the single trajectories as for the combination. Finally, we want to mention that the

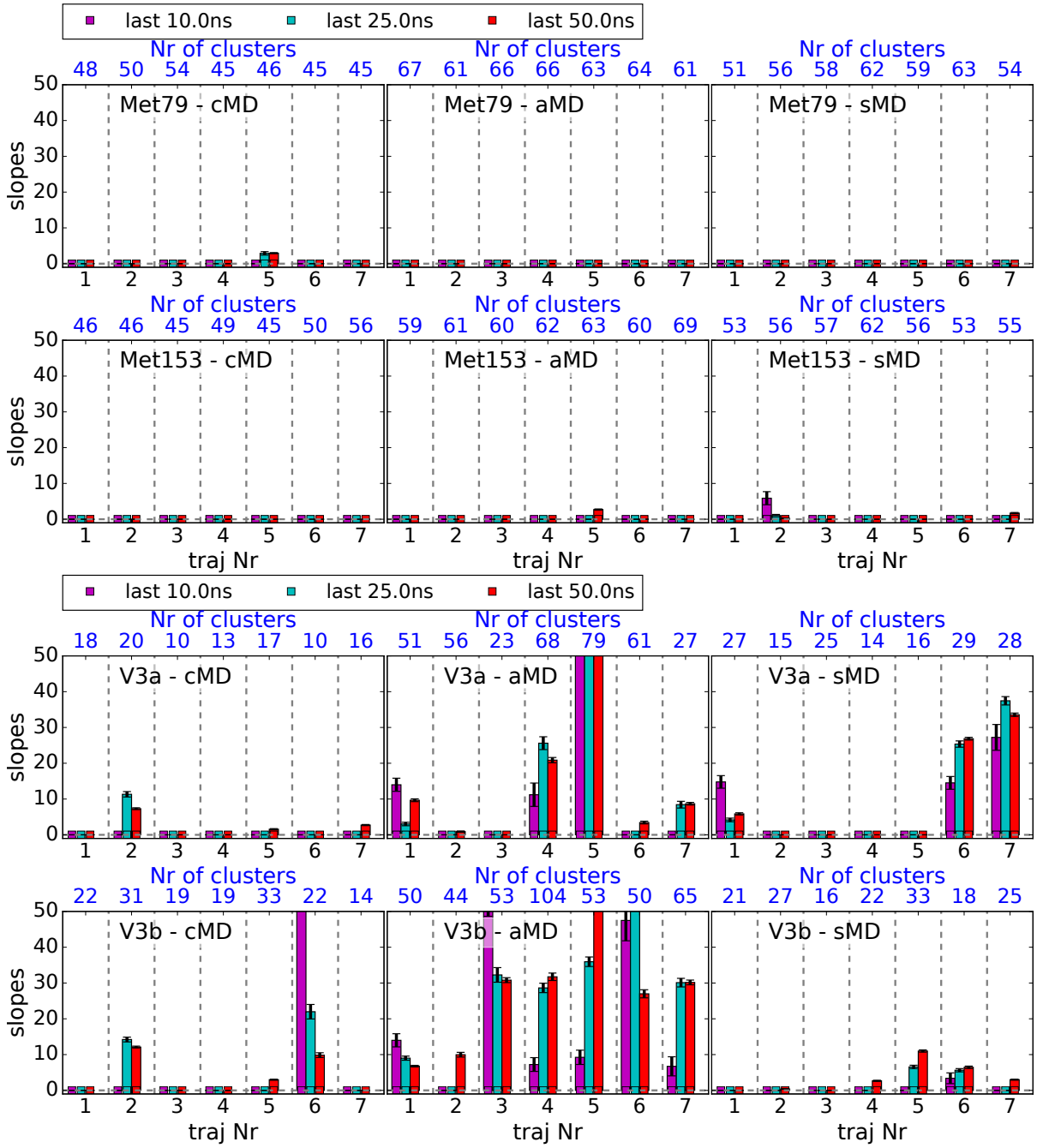

Fig. 4.28.: Slopes dN_C^{local}/dt of the last 10 ns, 25 ns and 50 ns for all different 200 ns trajectories of Met-Enkephalin at $r = 0.11$ nm (top) and V3 at $r = 0.35$ nm (bottom). The slopes refer to an increase by the certain value for the next 100 ns. Error bars correspond to 95% confidence intervals. Rows and columns refer to starting structure and sampling method. The second x-axis gives the number of clusters found in total by the local clustering. The figure is taken from Ref. [37].

slight decrease in total numbers of clusters above 100 ns is the result of the decreasing number of trajectories, since the six 100 ns trajectories for each of the combination of

starting conformation and sampling method are omitted.

Another measure of convergence is the development of the cluster number N_C^{local} of each single trajectory to detect, whether the number of found clusters converge to a stable plateau. Here, we use the local clustering to obtain the best unique clustering for every single trajectory, since we want to analyze the slopes of the single curves and not necessarily the comparability of absolute numbers. The developments of all 200 ns trajectories for Met-Enkephalin for a local clustering at $r = 0.11$ nm and V3 at $r = 0.35$ nm are shown in Fig. 4.27 [37]. The slopes dN_C^{local}/dt of the last 10 ns, 25 ns and 50 ns are evaluated and illustrated in Fig. 4.28. For Met-Enkephalin, almost all trajectories converge to a long plateau which result in slopes of zero. Again, aMD trajectories show the most robust results, because they have on average the largest plateaus. Hence, the number of found clusters stabilize the earliest. V3 shows much more deviations in the development of $N_C^{\text{local, V3}}$, consistent with $O_{\text{conf}}, O_{\text{dens}}$, analyzed previously. cMD, aMD and sMD behave differently, which is consistent to $N_C^{\text{global,V3}}$: cMD runs produce the most stable $N_C^{\text{local, V3}}$ curves except for four or five trajectories, which could be misinterpreted as converged trajectories with slopes $dN_C^{\text{local}}/dt \approx 0$, although we already knew from the multi-trajectory approach that this is not correct. The situation is similar for sMD trajectories, where much less trajectories could be interpreted as stable (Figs. 4.27 left and 4.28). Only the trajectories generated by aMD correctly indicate the unconverged state with increasing $N_C^{\text{local, V3}}$ and large slopes, except for two trajectories coming from *V3a*. These results emphasize that one cannot rely solely on single trajectory convergence of the cluster number, also because the underlying distributions are not taken into account. These distributions will be treated with the cluster distribution entropy S_C^{local}.

4.5.2. Constancy of the cluster distribution entropy S_C

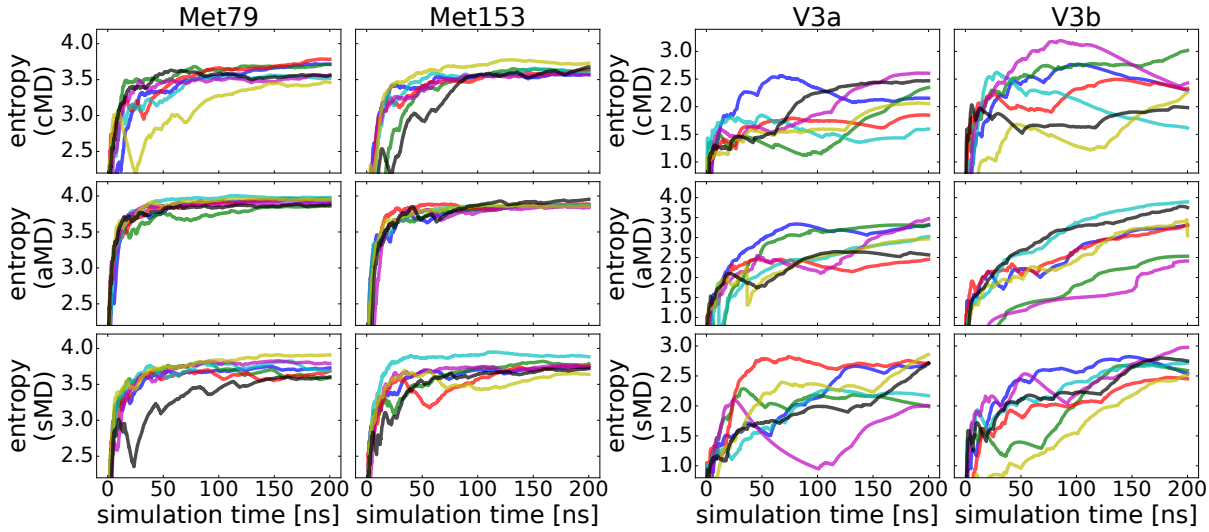

Fig. 4.29.: Development of the cluster distribution entropy S_C^{local} for single trajectories as a function of the simulation time t. Met-Enkephalin (left) and V3 (right). For each of the combination of sampling method and starting structure, seven 200 ns trajectories (distinguished by colors) were clustered separately at $r = 0.11$ nm for Met-Enkephalin and $r = 0.35$ nm for V3. The figure is taken from Ref. [37].

As discussed above, the last measure for the convergence is the cluster distribution entropy S_C^{local} following Eq. (3.18). This quantity allows a complementary measure to a converged number of clusters N_C^{local} from the previous subsection to investigate, whether also the underlying distribution converged. The initial idea of Sawle and Ghosh [32] was to detect constant regions in the curves of S_C^{local}. We will again evaluate the slopes of the curves for the last 10 ns, 20 ns, 50 ns and for the time interval after the last cluster was found. A value of $dS_C^{\text{local}}/dt \approx 0$ indicates correct sampling of the underlying energy landscape if the cluster numbers are stable.

Since S_C^{local} is calculated by the number of frames in certain clusters, we will re-weight the number of frames if they originate from aMD or sMD runs according to the weights calculated for Met-Enkephalin at $r = 0.11$ nm and V3 at $r = 0.35$ nm [37]. It has to be mentioned that a global partitioning and also the non-weighted S_C^{local} curves are very similar (not shown).

We will again start with the analysis of Met-Enkephalin (left panels of Fig. 4.29 and top rows of Fig. 4.30). All trajectories show indications of convergence by stable $S_C^{\text{local,Met}}$ for $t \gtrsim 100$ ns. Remarkable, again aMD gives the best results obtaining slopes $dS_C^{\text{local,Met}}/dt$ closest to zero. The entropy development is closely related to the curves of $N_C^{\text{local,Met}}$,

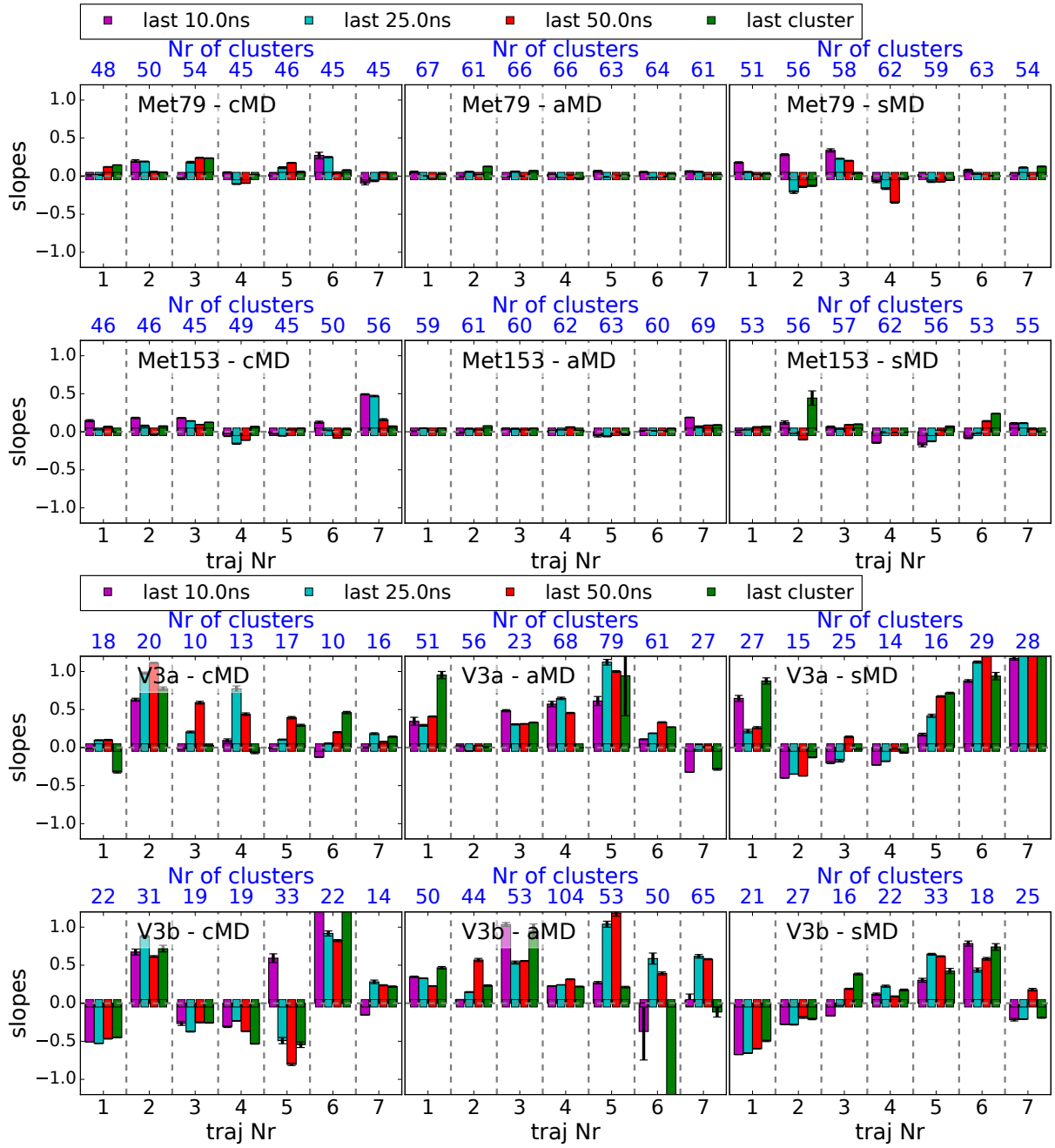

Fig. 4.30.: Slopes of the cluster distribution entropy dS_C^{local}/dt of the last 10 ns, 25 ns and 50 ns for all different 200 ns trajectories of Met-Enkephalin at $r = 0.11$ (top) and V3 at $r = 0.35$ nm (bottom). "last cluster" means the time interval between addition of last cluster and end. The slopes refer to an increase or decrease by the certain value for the next 100 ns. Error bars correspond to 95% confidence intervals. Rows and columns refer to starting structure and sampling method. The second x-axis gives the number of clusters found in total by the local clustering. The figure is taken from Ref. [37].

i.e. $S_C^{\text{local,Met}}$ stabilizes in the same region where the last cluster was found. The overall conclusion for the Met-Enkephalin runs is that no run seems to show trapped behavior.

This is different for V3 (right panels of Fig. 4.29 and bottom rows of Fig. 4.30). As expected from previous analyses, the distributions $S_C^{\text{local,V3}}$ are consistently very unstable for all sampling methods and starting conformations. In contrast to the slopes $dN_C^{\text{local,V3}}/dt$, one is now able to detect the incomplete sampling for cMD and sMD. aMD gives (almost) always $dS_C^{\text{local,V3}}/dt > 0$ which is related to the increase of cluster number, i.e. related to sampling new conformational space. On the other hand, $dS_C^{\text{local,V3}}/dt < 0$ refers to states, where sampling is distributed only between few clusters. Comparing to the number of clusters, these regions can be linked to simulation times where the cluster number stays constant.

The overall summary is that the number of clusters N_C and cluster distribution entropy S_C may be pre-criteria to investigate single trajectories for instable behavior and thus unconverged simulations if used in combination. But only a multi-trajectory approach is an effective way to prevent a wrong classification of convergence.

4.6. Combined assessment of convergence

In the last two sections about the overlap and clustering analyses, we learned that it is necessary to combine both approaches, the overlap and clustering, to comprehensively assess the convergence quantitatively. Only if both show consistent results of convergence, the sampling can be complete, assuming that the full conformational space is sampled. On the one hand, the probability density functions $p(\vec{r})$ of independent experiments must correspond to each other, which is fulfilled by $O_{\text{conf}} = 1$ and a large O_{dens} value. On the other hand, the sampling is only then complete, if the size of the sampled conformational space converges, and different trajectories explore the same number of clusters N_C^{global}. Therefore, we will investigate the combination of O_{dens} and N_C^{global} to evaluate the convergence of different sets of MD trajectories and/or compare different sampling methods [37]. Here, N_C^{global} is the number of unique clusters found by all trajectories involved in the corresponding overlap value O_{dens}. The clustering was again done globally using all different trajectories to obtain one partitioning due to comparison reasons.

We evaluate different simulation times t for Met-Enkephalin (Fig. 4.31), whereas we investigate the 1 μs and 200 ns trajectories separately. For the 1 μs trajectories, analyzing the combined assessment of convergence at different time points t reveals that the combination of sMD runs seems to explore the conformational space as fast as aMD, finding at

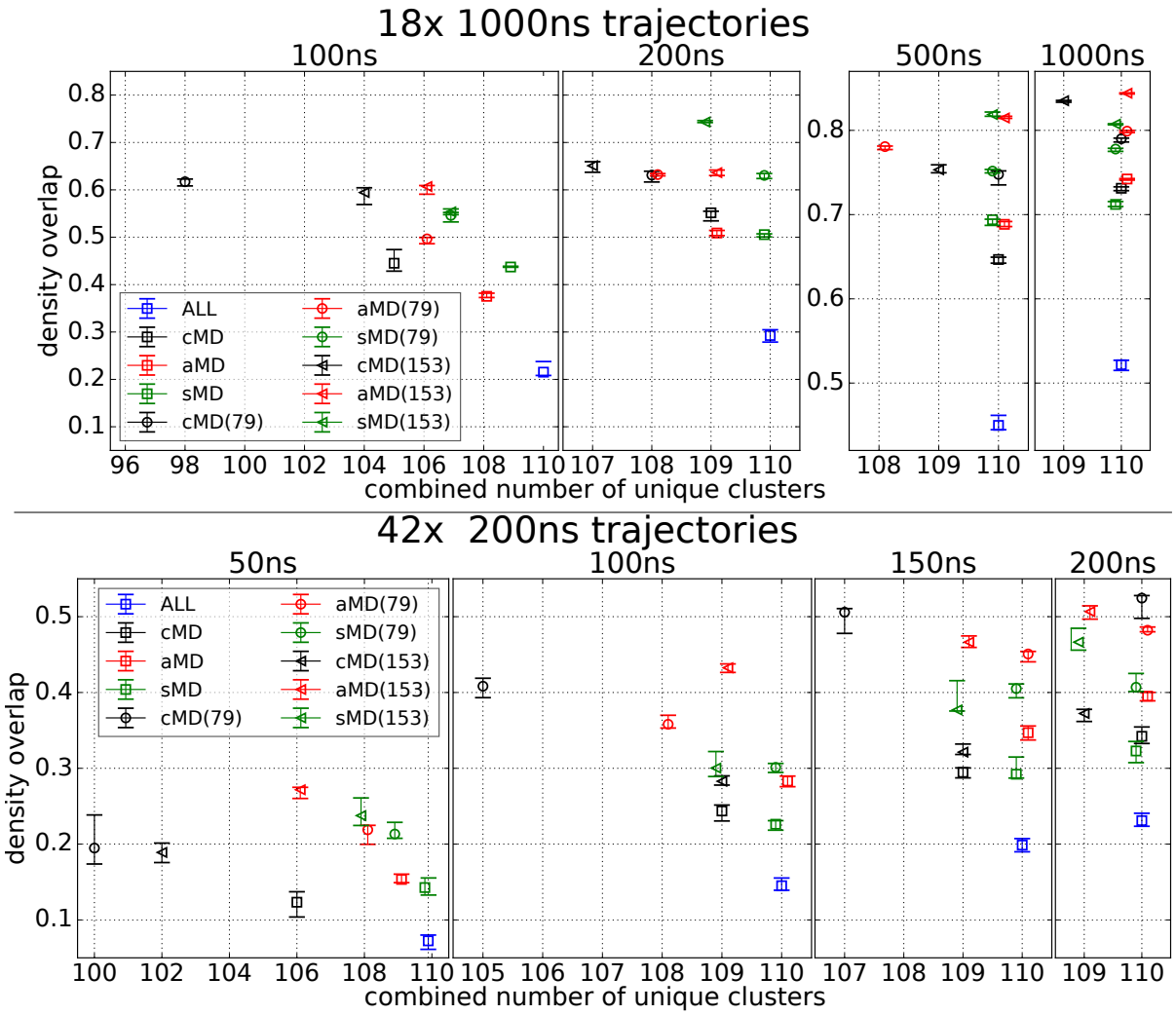

Fig. 4.31.: Density overlap O_{dens} vs. number of clusters N_C^{global} for different groups at different simulation times t for Met-Enkephalin. Top: 18×1 μs trajectories. Bottom: 42×200 ns trajectories. The clustering and overlap measures are done at $r = 0.11$ nm. The total number of trajectories ("ALL") are divided into subgroups by a factor of three for different sampling methods and by another factor of two for different starting structures. Cluster numbers for aMD and sMD are slightly shifted with $\ll 1$ for visibility reason.

some points even one more cluster compared to aMD. This is surprising if one considers the outcome of V3: There, combinations of aMD trajectories reach much more clusters much faster, detecting the huge conformational space of V3 (Fig. 4.26). On the other hand, the density overlap for Met-Enkepphalin of aMD and sMD are comparable up to 500 ns (compare also Fig. 4.22), but are then outperformed by aMD. One reason might be that aMD finds the last clusters more quickly, and then the equilibrium sampling is faster. Remarkably, this is also the case for cMD, after it finally detects (almost) all clusters, the

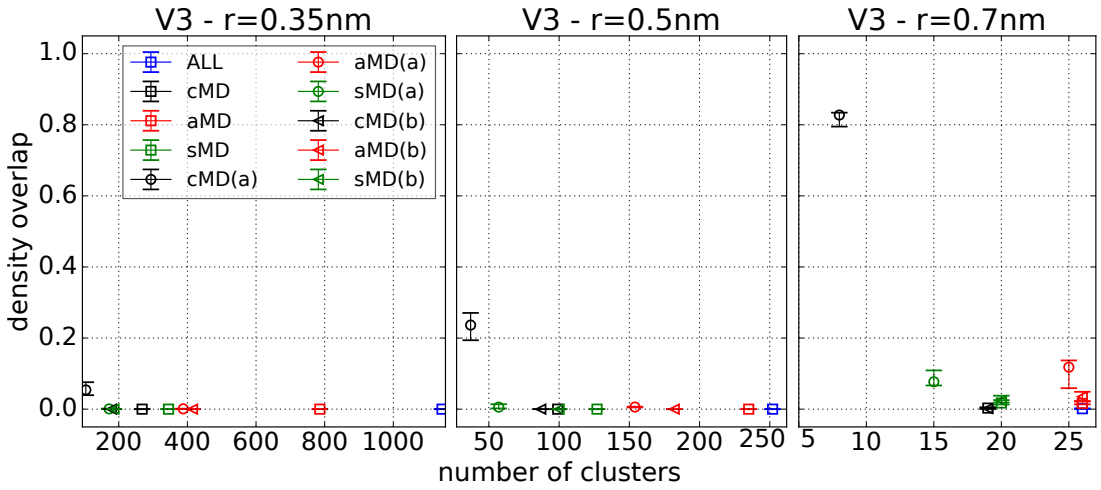

Fig. 4.32.: Density overlap O_{dens} vs. number of clusters N_C^{global} for different groups at different threshold parameters r for V3. Forty-two 200 ns trajectories are shown. The total number of trajectories ("ALL") are divided into subgroups by a factor of three for different sampling methods and by another factor of two for different starting structures. The figure corresponds to Ref. [37].

overlaps seem to pass the values of sMD. For the 42×200 ns trajectories (bottom row of Fig. 4.31), there are much more trajectories. Thus, a certain cluster number is easier to be reached and a certain overlap is harder to be achieved compared to the cases above. The development of the point corresponding to cMD trajectories from *Met79* is outstanding and unexpected, because the overlap value grows very fast and the maximal number of clusters is reached at $t = 200$ ns (see Fig. 4.31 bottom). This value is not comparable to the trajectories from the other starting structures but there is also no indication that the behavior could be explained by an artifact like trapped trajectories. But one has to keep in mind that the overlap value of $O_{\text{dens}} \approx 0.5$ is still far from being converged.

The results for V3 at different resolutions are displayed in Fig. 4.32. We already know that $O_{\text{dens}} \approx 0$ up to $r \leq 0.7$ nm for most combinations but trajectories from *V3a*. For a high resolution at $r = 0.35$ nm and an intermediate resolution $r = 0.5$ nm, only trajectories from cMD and *V3a* have non-zero density overlaps. Because trajectories are trapped in few states with $N_C^{\text{global, cMD(a)}}$ much smaller than the maximally reachable cluster numbers, they are able to sample these states more intensely. If trajectories from both starting structures are combined, the cluster number is about four times larger and the overlap drops to zero.

One can see the importance of validating the sampling with different sets of starting conformations for these two flexible biomolecules, Met-Enkephalin and V3, for a comprehensive study.

4.7. Bias analysis of enhanced sampling methods

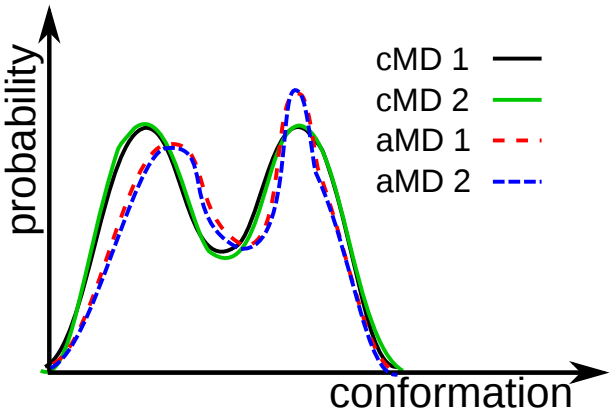

Fig. 4.33.: Schematic illustration of a deviation between cMD and biased aMD probability distributions. $O_{\text{dens}} = 1$ between cMD and between aMD trajectories, but $O_{\text{dens}} < 1$ between cMD and aMD trajectories.

So far, we focused on the assessment of the sampling quality of different MD simulations characterized by the overlap measures $O_{\text{conf}}, O_{\text{dens}}$ incorporating the development of the number of clusters N_C and cluster distribution entropy S_C. This allows to quantitatively investigate the underlying sampling for two or multiple trajectories. But, it is also possible to classify different sets or groups of trajectories. We also presented and used two enhanced sampling algorithms which distort the energy landscape to ease conformational transitions. To re-obtain the correct ensemble, we implemented different re-weighting schemes and developed a mean-field treatment specialized for our r-neighborhood approximation. Nevertheless, it is well-known that the re-weighting can lead to deviations from Boltzmann distributions [114, 115, 119, 122]. We have not investigated this issue, yet, but detected that aMD itself behaves the best by obtaining the fastest large overlap values O_{dens} along with converged cluster numbers N_C (Figs. 4.27 and 4.31). Additionally, only aMD was able to clearly identify the sampling failure for V3 finding 2 to 4 times more clusters than the other two sampling methods. But indeed, it is necessary to identify, whether the sampling of the two enhanced sampling methods is really correct and describes the unbiased ensemble after re-weighting.

It is possible to test wrong or biased distributions with O_{dens} in the following way [37]: Imagine two sets of trajectories $L_A = \{l_{A1}, l_{A2}, ..., l_{An}\}$ and $L_B = \{l_{B1}, l_{B2}, ..., l_{Bn}\}$, where both sample their underlying energy landscapes A and B completely and correctly. This will lead to density overlaps of 1 for the corresponding sets

$$O_{\text{dens}}(L_A, L_A; r) = 1 = O_{\text{dens}}(L_B, L_B; r)\,.$$

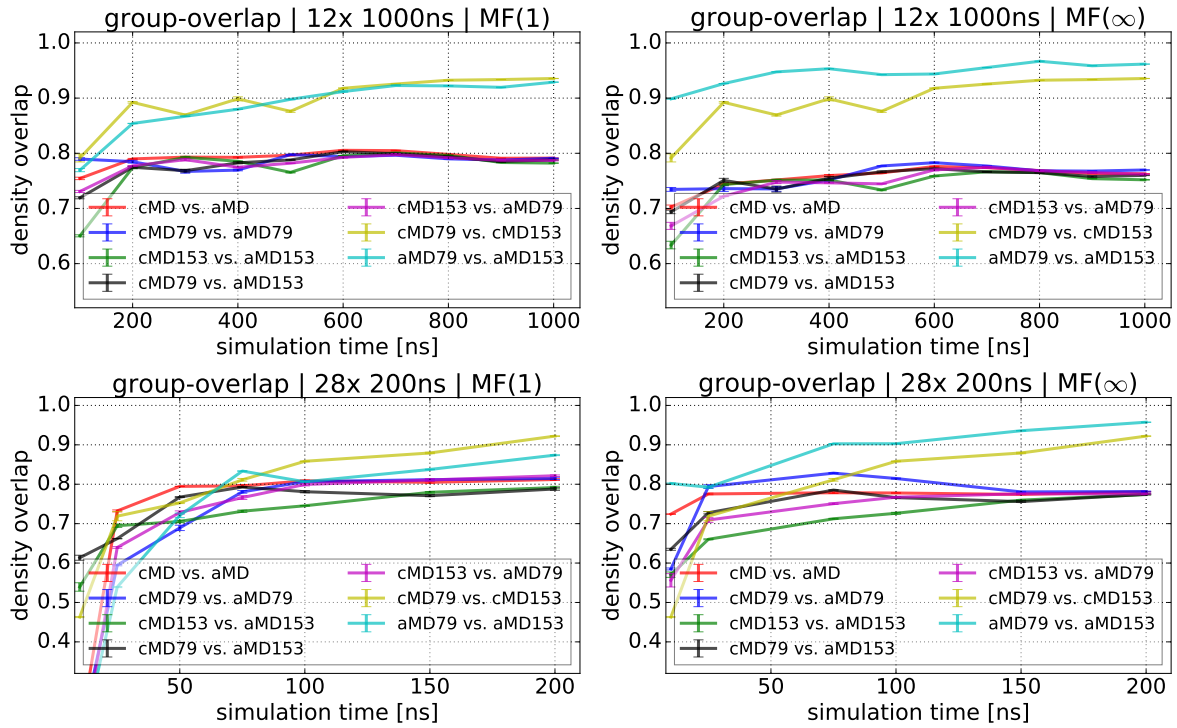

Fig. 4.34.: Group-overlap O_{dens} between different sets of concatenated trajectories of cMD and aMD runs as a function of different simulation times. For example, "cMD vs. aMD" refers to the density overlap between all concatenated cMD vs. all concatenated aMD trajectories. Top row: Three 1 μs trajectories per start structure (*Met79*, *Met153*) and sampling method (cMD,aMD) were used. Bottom row: Seven 200 ns trajectories per start structure and sampling method were used. Columns refer to mean-field re-weighting after first step $\text{MF}^{(1)}$ (left) and after converged weights $\text{MF}^{(\infty)}$ (right).

Automatically, it has to follow that the density overlap of the combined set $\{L_A, L_B\}$ must also be 1 if both energy landscapes A and B are identical, but < 1 if one potential B is biased. The resulting probability densities are schematically shown in Fig. 4.33. Thus, we have a criterion to test, whether (re-weighted) distributions are still biased compared to conventional MD, assuming that cMD sampling is correct. It must follow

$$\left.\begin{array}{ll} O_{\text{dens}}\left(L_{\text{cMD}}, L_{\text{cMD}}; r\right) = O_{\text{dens}}\left(L_{\text{xMD}}, L_{\text{xMD}}; r\right) & \rightarrow 1 \\ O_{\text{dens}}\left(\{L_{\text{cMD}}, L_{\text{xMD}}\}, \{L_{\text{cMD}}, L_{\text{xMD}}\}; r\right) & \rightarrow 1 \end{array}\right\} \Rightarrow \text{correct} \tag{4.1}$$

$$\left.\begin{array}{ll} O_{\text{dens}}\left(L_{\text{cMD}}, L_{\text{cMD}}; r\right) = O_{\text{dens}}\left(L_{\text{xMD}}, L_{\text{xMD}}; r\right) & \rightarrow 1 \\ O_{\text{dens}}\left(\{L_{\text{cMD}}, L_{\text{xMD}}\}, \{L_{\text{cMD}}, L_{\text{xMD}}\}; r\right) & < 1 \end{array}\right\} \Rightarrow \text{biased}\,, \tag{4.2}$$

where "xMD" can stand for aMD, sMD or a completely different method. With this crite-

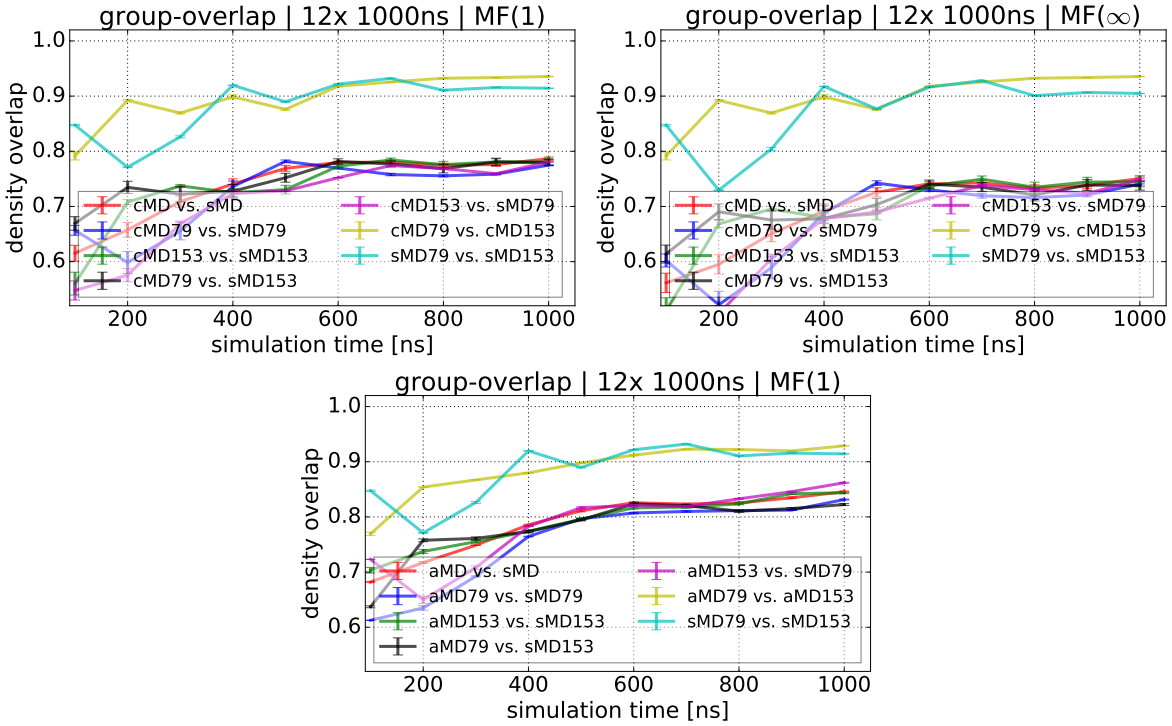

Fig. 4.35.: Group-overlap O_{dens} between different sets of concatenated trajectories of cMD and sMD (top) and aMD and sMD (bottom) runs as a function of different simulation times. For example, "cMD vs. sMD" refers to the density overlap between all concatenated cMD vs. all concatenated sMD trajectories. Three 1 μs trajectories per start structure (*Met79*, *Met153*) and sampling method (cMD,aMD,sMD) were used. Panels top left and bottom refer to mean-field re-weighting after first step $\text{MF}^{(1)}$, top right refers to converged weights $\text{MF}^{(\infty)}$.

rion, we test the sampling of aMD and sMD for Met-Enkephalin to detect a possible bias after re-weighting, because we assume that the sampling of Met-Enkephalin trajectories of μs-lengths are approximately exhaustive.

We use six 1 μs trajectories per sampling algorithm (three per starting structure) and concatenate different independent trajectories to investigate O_{dens} between combined groups of trajectories, the group-overlap. This approximately enlarges the trajectories of interest to $3-6$ μs, assuming that the combination of independent 1 μs trajectories incorporates also conformational regions which are weaker sampled in single runs and therefore the overall sampling is enhanced. For these combinations, we already saw that the number of found clusters between cMD, aMD and sMD are closely related to each other for $t \geq 200$ ns (Fig. 4.31), thus the same conformational space was visited. The group-overlap comparing cMD sampling with aMD is illustrated in top of Fig. 4.34. All corresponding conformational overlap O_{conf} is practically equal to 1. The hypothesis that aMD runs are still biased after re-weighting can be impressively evaluated. Group-overlap involving only

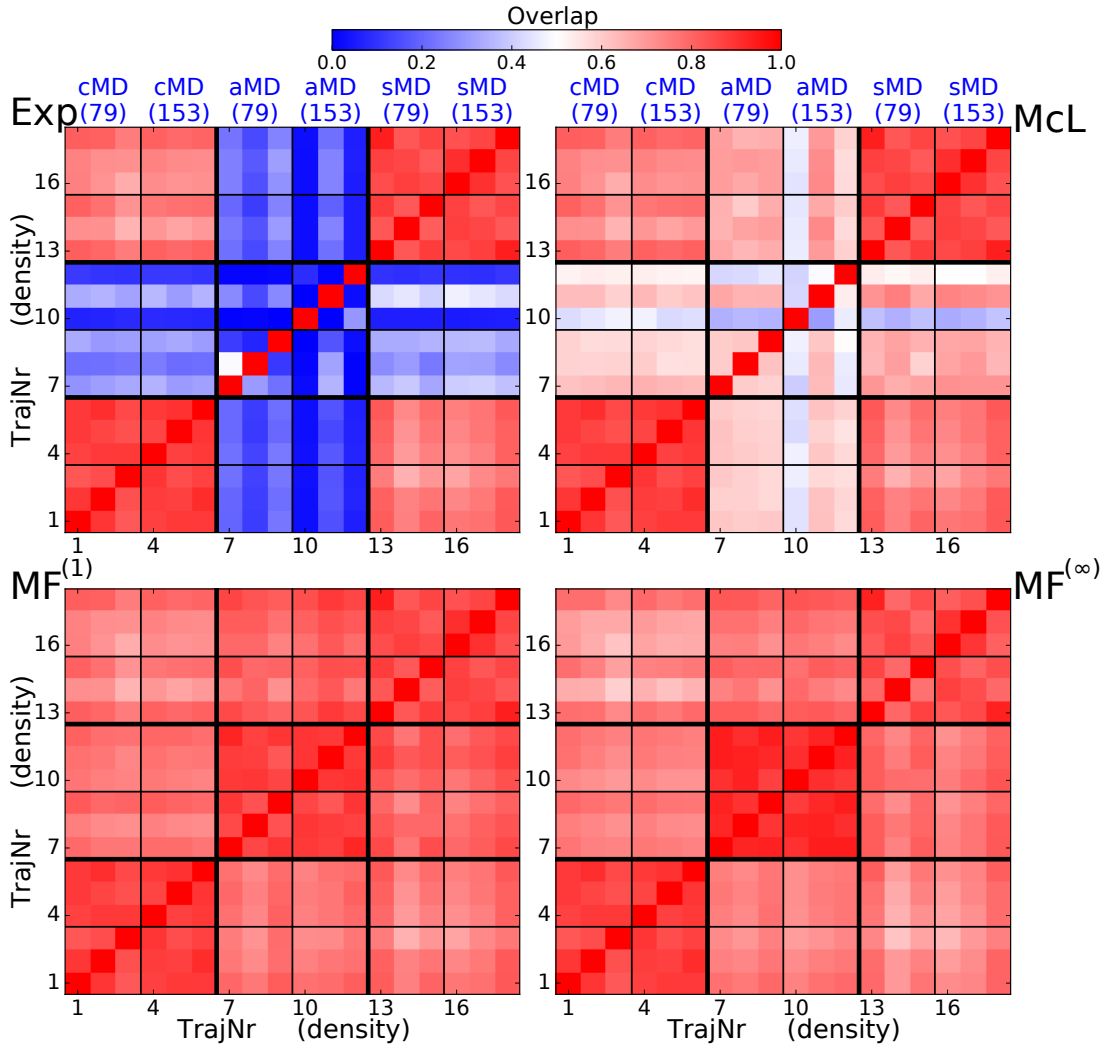

Fig. 4.36.: Density overlap O_{dens} between all pairs of 18×1 μs trajectories for Met-Enkephalin at $r = 0.11$ nm. Different re-weighting schemes were used for aMD (Exp, 10th order McL, $\text{MF}^{(1)}$, $\text{MF}^{(\infty)}$). For sMD, $\text{MF}^{(1)}$ was always used except bottom right where $\text{MF}^{(\infty)}$ was used for both accelerated methods. The re-weighting refers to Eqs. (3.13) and (3.17).

cMD trajectories (yellow curves) and only aMD trajectories (cyan curves) tend toward a value of 1 (> 0.9 for full trajectories), whereas all overlap cross-combinations of cMD and aMD simulations end very quickly at a constant value of ≈ 0.8 without indication of a further increase. For $\text{MF}^{(\infty)}$ (top right of Fig. 4.34), which is almost identical to the non-weighted case, the results are the same, but the cross-combinations have even lower overlap < 0.8, which must be expected since these trajectories sampled a biased potential by definition. We asked ourselves, whether this is also true, if more trajectories are involved. For the seven 200 ns per sampling method and starting conformation (bottom row of Fig. 4.34), the outcome is the same for the group-overlap, although more than

twice the number of trajectories are combined and for $t < 200$ ns the number of found clusters for cMD starting from *Met79* is smaller compared to the other cases (Fig. 4.31). Thus, the result seems to be robust.

We investigated the same for combinations of cMD and sMD shown in top of Fig. 4.35 for the same number of 1 μs trajectory combinations as before. Remarkably, the same bias behavior can be identified for the sMD trajectories, but the overlap values are not as constant as for the aMD cases. The overlap values increase and decrease between chunks of 100 ns. The converged $\mathrm{MF}^{(\infty)}$ weights yield the same (Fig. 4.35 bottom), although the mean-field iteration has significant influence on the sMD weights shown in Fig. 4.13. The most interesting detail is that the comparison between aMD and sMD does also result in a deviation of the group-overlap values between same and different sampling methods, but all curves show a monotonously increase toward a possible final O_{dens} of 1. Hence, both energy distortions, the lifting for aMD and down-scaling for sMD, produce more compatible results compared to cMD.

Finally, we want to analyze the influence of all four different re-weighting schemes (Exp, 10th order McL, $\mathrm{MF}^{(1)}$ and $\mathrm{MF}^{(\infty)}$ following Eqs. (3.13) and (3.17)). The influence can be inspected comprehensively in the density pair-overlap shown as heatmaps (Fig. 4.36). One can see that for the straight forward exponential or Maclaurin re-weighting, the overlap is much smaller compared to the mean-field steps, because the distributions of the first re-weighting schemes are dominated by very few and very large weights. This issue can be resolved with the mean-field approximation, but it could not resolve the bias toward a correct Boltzmann distribution.

The bottom line of this subsection is the difficulty of a proper re-weighting for distorted energy sampling. This issue is a well-known problem [114, 119], is actively investigated [115, 118, 122, 215] and the problems are still unresolved. For aMD trajectories, the straight forward re-weighting applying the inverse Boltzmann factors either directly or by approximating the exponential function with a series expansion is dominated by few frames with 99% of all weights, yielding low overlap and poor results. Cumulant expansion and our mean-field approach MF can resolve this issue but still lead to biased distributions. One has to invest a lot of effort to calculate correct weights for the converged microstates. It might be possible to obtain much better re-weighting with an exhaustive validation of different neighborhood thresholds r. Maybe, re-weighting with a varying r for different reference frames κ could also lead to an enhancement for the weights. So far, the mean-field iteration yield for aMD a too large suppression until convergence is reached that the re-weighting in the converged case is almost negligible. Interestingly,

although sMD does not suffer from energetic fluctuations [113], we could detect a bias using the re-scaled population in each r-neighborhood, which could not be resolved by the mean-field convergence. At least, the density overlap O_{dens} is able to compare different sampling methods and re-weighting schemes to find a possible bias.

In the next section, we will investigate the influence of the overlap measures on thermodynamic averages and will also briefly discuss the influence of different re-weighting schemes.

4.8. Influence of O_{conf} and O_{dens} on thermodynamic observables

In the last sections, the focus was to detect the convergence of trajectories, whether the conformational space was appropriately sampled. This is the necessary condition that all further results are reliable and that extracted thermodynamic observables are correct (assuming that no conformational space is missed). In the end, one is interested in thermodynamic averages to draw conclusions about systems. Thus, in the following we will investigate the influence of changing overlap values on different thermodynamic quantities.

4.8.1. Convergence of thermodynamic averages

Table 4.8.: Density overlap O_{dens} for different pairs of arbitrary chosen Met-Enkephalin trajectories of different time lengths.

	100 ns	500 ns	1000 ns
Met79 (cMD) vs. Met153 (cMD)	0.642	0.741	0.883
Met79 (aMD) vs. Met153 (aMD)	0.541	0.870	0.889
Met79 (cMD) vs. Met79 (aMD)	0.615	0.747	0.765
Met79 (cMD) vs. Met153 (aMD)	0.613	0.745	0.751
Met153 (cMD) vs. Met79 (aMD)	0.622	0.790	0.784
Met153 (cMD) vs. Met153 (aMD)	0.504	0.766	0.756

The backbone angle distributions (ϕ,ψ) are often used to compare different results of MD simulations. Therefore, we compare the distributions of different trajectories (two cMD, two aMD) for increasing density overlap O_{dens} between 0.5 and 0.9 (Table 4.8) of Met-Enkephalin. Interestingly, single trajectories show increasingly smooth probability

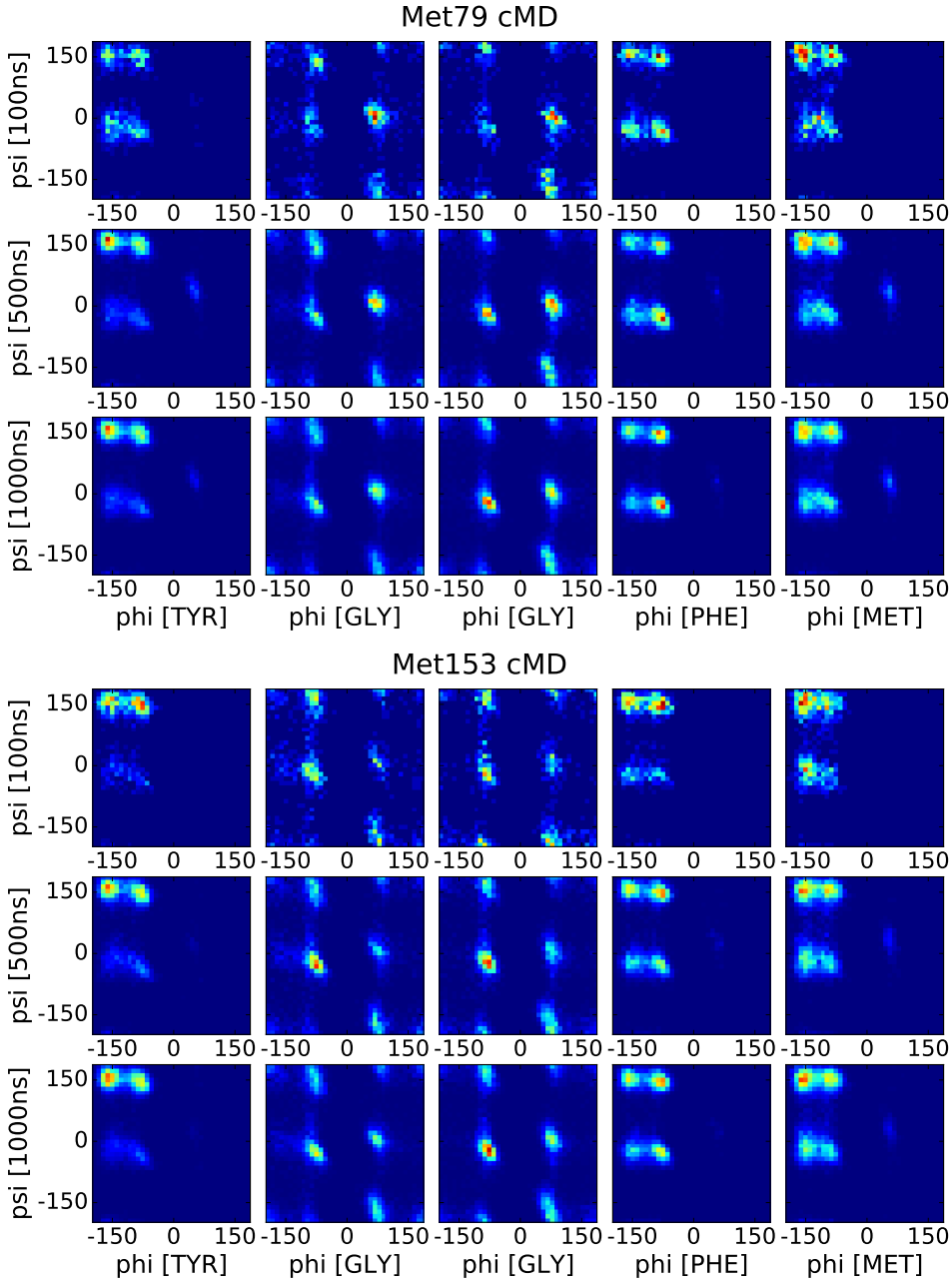

Fig. 4.37.: Normalized distributions of backbone dihedral angles ϕ and ψ of two arbitrarily chosen cMD trajectories of Met-Enkephalin starting from *Met79* (top block) and *Met153* (bottom block). Rows represent different time states (100 ns, 500 ns, 1000 ns), and columns refer to different residues (*Tyrosine, Glycine, Phenylalanine, Methionine*). The distributions correspond to a binning with a resolution of 10°, whereas the probability is shown as colorcode from blue to red. Dark blue means 0 probability, red changes between columns (0.03, 0.02, 0.02, 0.026, 0.02), respectively.

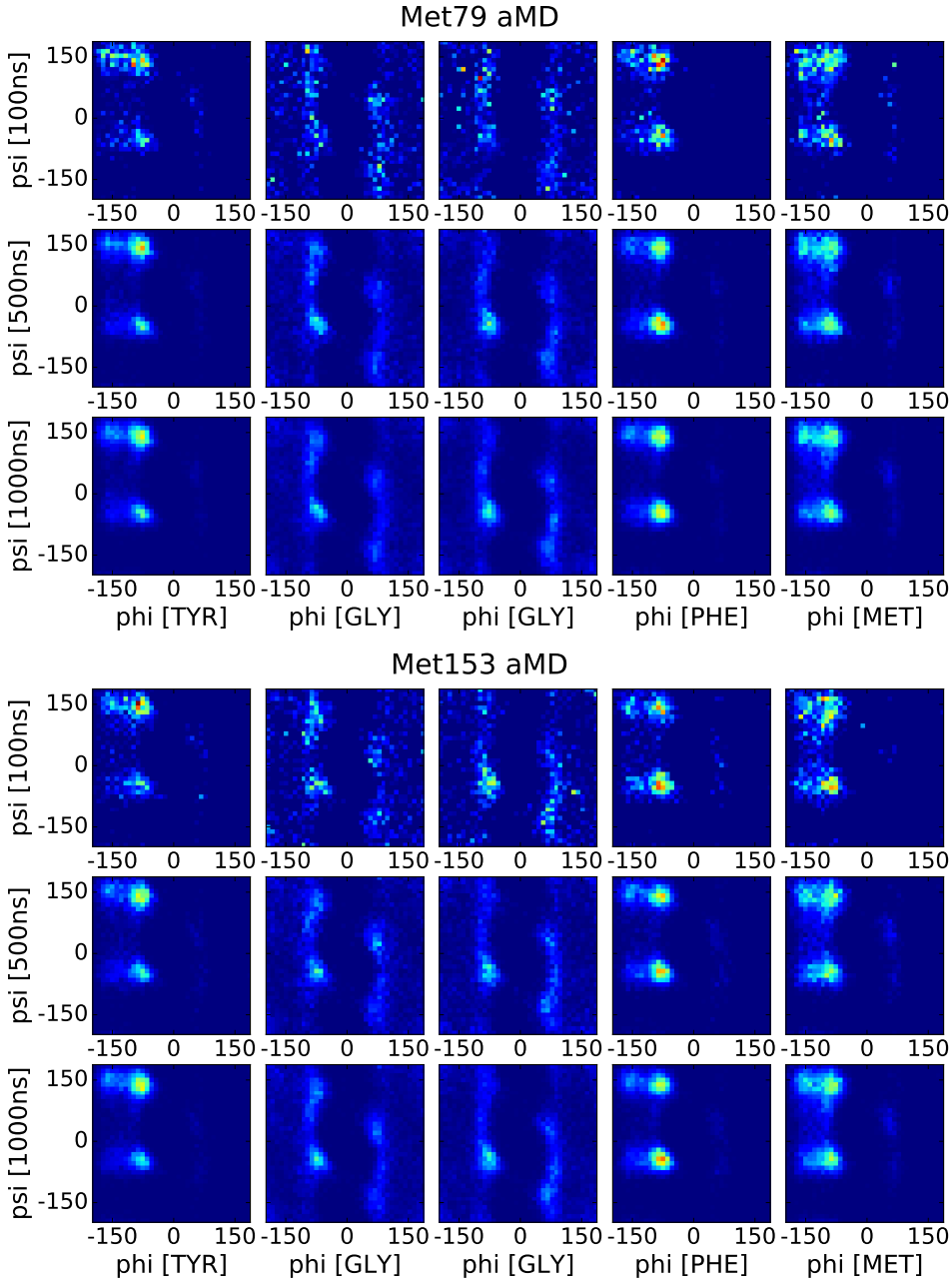

Fig. 4.38.: Re-weighted and normalized distributions of backbone dihedral angles ϕ and ψ of two arbitrarily chosen aMD trajectories of Met-Enkephalin starting from *Met79* (top block) and *Met153* (bottom block). The re-weighting was done using $\mathrm{MF}^{(1)}$ at $r = 0.11$ nm. Rows represent different time states (100 ns, 500 ns, 1000 ns), and columns refer to different residues (*Tyrosine, Glycine, Phenylalanine, Methionine*). The distributions correspond to a binning with a resolution of 10°, whereas the probability is shown as colorcode from blue to red. Dark blue means 0 probability, red changes between columns (0.03, 0.02, 0.02, 0.026, 0.02), respectively.

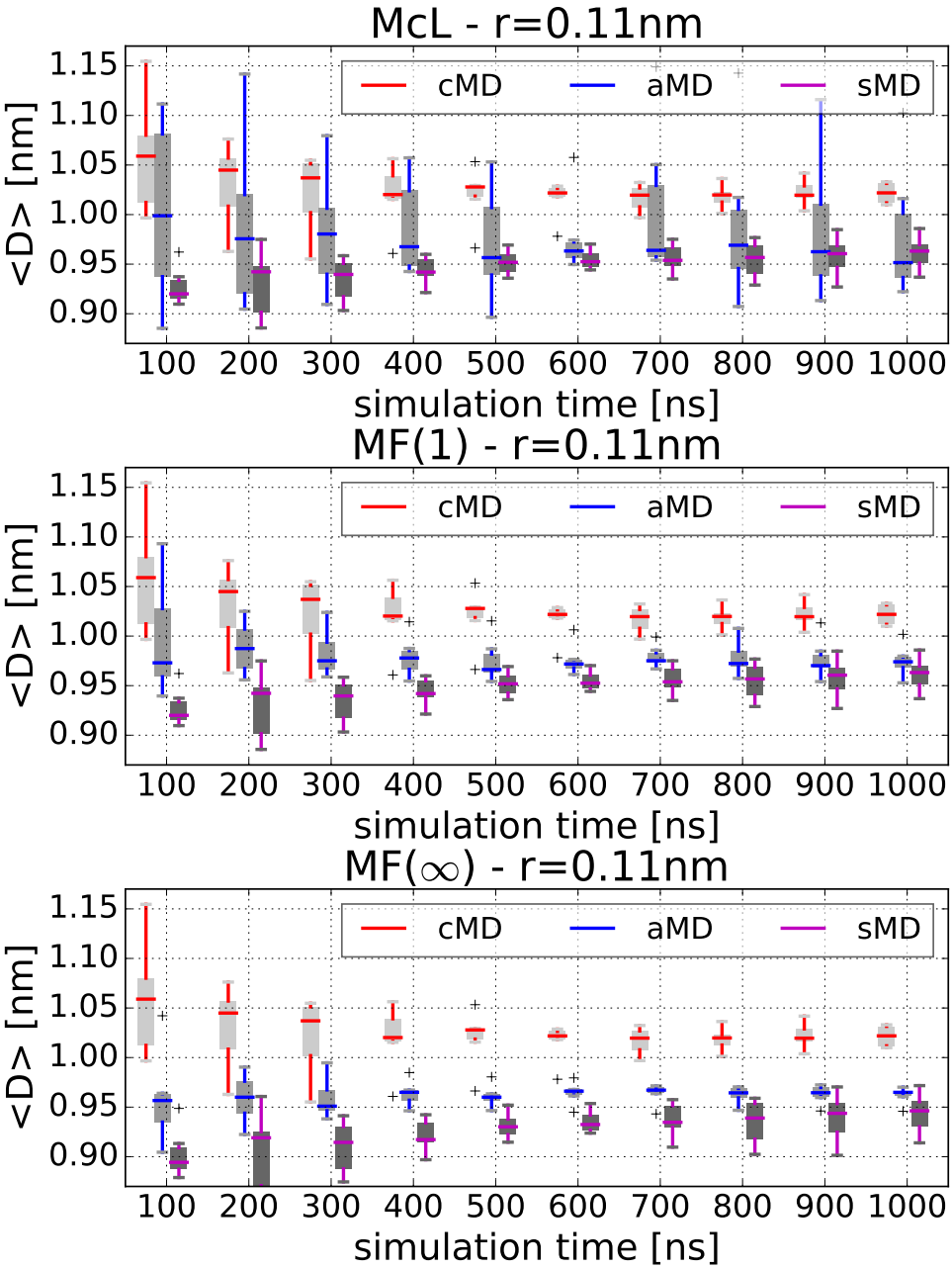

Fig. 4.39.: Averaged end-end distances $\langle D \rangle$ of Met-Enkephalin. Trajectories are sampled with cMD, aMD and sMD for different simulation times t and different reweighting schemes: Maclaurin expansion up to 10th order for aMD and first step mean-field $\mathrm{MF}^{(1)}$ for sMD at $r = 0.11$ nm (top), $\mathrm{MF}^{(1)}$ for aMD and sMD at $r = 0.11$ nm (center) and converged $\mathrm{MF}^{(\infty)}$ for aMD and sMD trajectories at $r = 0.11$ nm (bottom) following Eqs. (3.13)-(3.17). Each boxplot corresponds to six trajectories, three for *Met79* and *Met153*.

densities for different angles, but no major structural differences between time states with different O_{dens} [37]. Hence, they seem to converge more quickly than quantities which

require global convergence. On the other hand, both cMD trajectories and both aMD trajectories show separately very similar results, whereas the two glycines behave differently for cMD and aMD. There are regions at $\phi \lesssim 80°$ and $\phi \gtrsim 80°$ with zero probability for cMD, which have non-zero values for aMD. This is remarkable, although we could show that the re-weighted aMD trajectories are still biased. One possible explanation could be that these regions have a very low probability to occur, which is overestimated by the re-weighted aMD trajectories but still missed within cMD. Remember that also O_{dens} of the cMD trajectories is below 1 and some runs did not found all clusters (see Fig. 4.31).

As a global measure of Met-Enkephalin, we investigate the end-end distance distributions and the averaged distance $\langle D \rangle$ between the terminal nitrogen of *Tyrosine* and the terminal carbon of *Methionine* [37]. The latter $\langle D \rangle$ is shown in Fig. 4.39 as a function of different simulation time lengths for different sampling methods and re-weighting schemes. Each sampling method incorporates six different 1 μs trajectories, which showed consistently increasing density overlap O_{dens} up to around 0.7 with convergence of the number of found clusters (see Figs. 4.27 left and 4.31). This implies that a clear convergence behavior should be visible in $\langle D \rangle$ as a function of the simulated time, which is indeed the case between $\approx$ 100 to 600 ns for all sampling methods (cMD, aMD, sMD). The distribution of values of different trajectories show continuously decreasing spread and stay nearly constant for $t \geq 600$ ns with an error of the order of 0.01 nm. Nevertheless, the outcome of cMD is significantly different compared to the other two sampling methods, where the values do not lie within the error intervals. The end-end distances $\langle D \rangle$ show also the large error introduced by the Maclaurin expansion for aMD re-weighting, which is even more visible for the distance distributions in Fig. 4.40. On the other hand, the first mean-field step $\text{MF}^{(1)}$ gives consistent results for $\langle D \rangle$ and converges the fastest (Fig. 4.40 third panel). Remarkably, the converged $\text{MF}^{(\infty)}$ trajectories for aMD show almost no error, although they result in almost no re-weighting due to smoothing the weights (see section 4.3 and Fig. 4.10). Hence, the effect of accelerating conformational transitions by aMD is clearly visible, because low potential energies are lifted and the biased sampling can much quicker sample reproducibly the underlying potential. In contrast, the re-weighted sMD results have the largest error values for $t > 500$ ns, although these trajectories should also benefit from the acceleration of conformational transitions. The non-weighted sMD distance averages have comparable errorbars as the cMD result and converge even to slightly larger values than aMD (not shown). Again, the converged $\text{MF}^{(\infty)}$ weights have a certain impact on sMD runs, but $\langle D_{\text{sMD}} \rangle$ is shifted to even smaller

values compared to cMD. Remarkably, the results of sMD and aMD are compatible which is in agreement with the previous bias analysis (Fig. 4.35).

In summary, we could show that the convergence of O_{dens} and N_C are linked to the error of the thermodynamic average of the end-end distance $\langle D \rangle$ of Met-Enkephalin. Thus the quantities have a clear impact to the results. It is also interesting to see that $\langle D \rangle$ does not strictly require $O_{\text{dens}} = 1$ to show convergence and a realistic value with a small error estimate. Thus, depending on the system and quantity of interest, one has to decide, whether it is necessary to invest much more calculation time to drive $O_{\text{dens}} \gg 0.8$ for the overlap between a group of single trajectories.

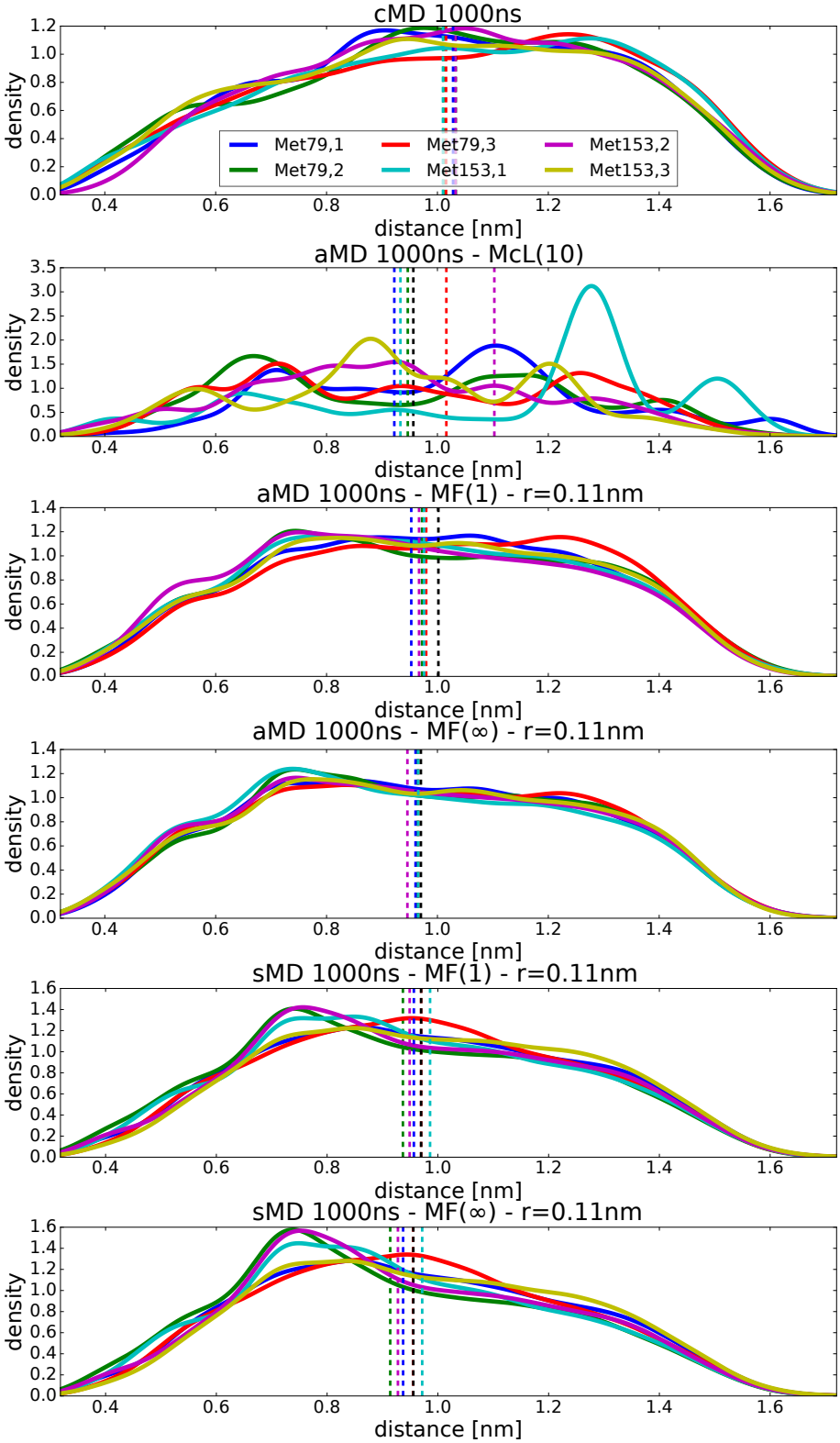

Fig. 4.40.: Probability distribution approximates of end-end distances D of Met-Enkephalin after 1 μs for different sampling methods and re-weightings. For each panel, six trajectories were evaluated, three from *Met79* and *Met153*. Vertical dashed lines mark the means of $\langle D \rangle$ for each trajectory. Different re-weighting schemes are shown: MF(1) and MF(∞) at $r = 0.11$ nm, and McL(10).

4.8.2. Effect of the threshold r on thermodynamic averages

Table 4.9.: The density overlap O_{dens} of the six 1 μs long cMD trajectories of Met-Enkephalin ($K = L$) for different combinations of threshold r and simulation time t. The asymmetric error estimate ΔO_{dens} corresponds to the first and third quartile of the six overlap values calculated for the six individual reference trajectories defined in subsection 3.2.5.

O_{dens}	0.810	0.796	0.805	0.797	0.790	0.794	0.814	0.790
ΔO_{dens}	$^{+0.001}_{-0.002}$	$^{+0.004}_{-0.004}$	$^{+0.003}_{-0.002}$	$^{+0.004}_{-0.000}$	$^{+0.004}_{-0.000}$	$^{+0.004}_{-0.004}$	$^{+0.004}_{-0.006}$	$^{+0.006}_{-0.007}$
r [nm]	0.13	0.14	0.15	0.16	0.17	0.19	0.20	0.21
t [ns]	1000	700	600	500	400	300	200	100

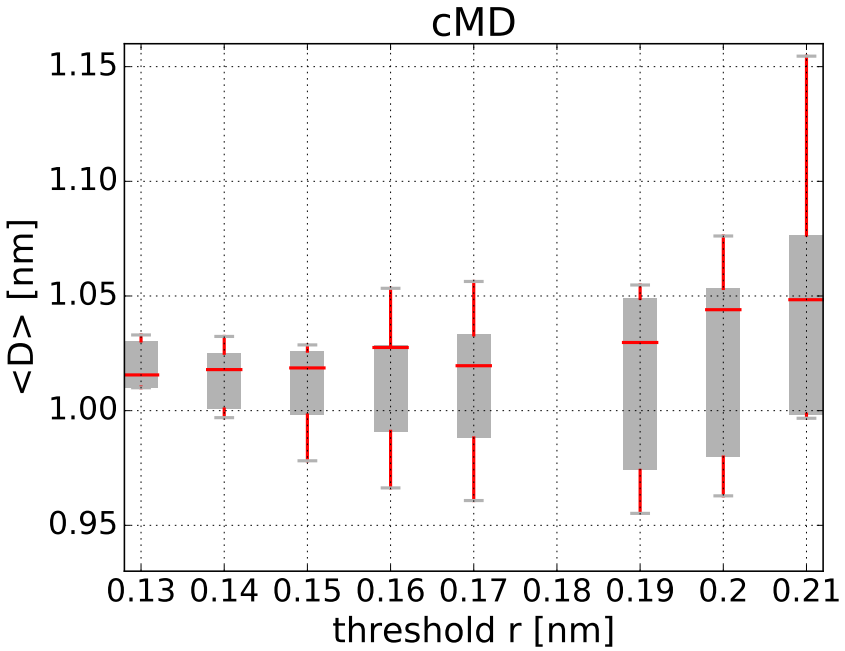

Fig. 4.41.: Accuracy of the end-end distance estimate $\langle D \rangle$ of Met-Enkephalin as a function of the threshold r maintaining the same density overlap $O_{\text{dens}} \approx 0.8$. Results are shown as boxplots with medians. For low resolutions (large r), less simulation time is needed to achieve a given O_{dens} value, but the estimated value of $\langle D \rangle$ becomes inaccurate. For small r, the estimate gains accuracy. The corresponding overlap, threshold and simulation time values are given in Table 4.9. The figure is taken from Ref. [37].

So far, we argued that the threshold r can be understood as a resolution: For small r, one obtains a very detailed view of the conformational space, i.e. even small deviations

between two structures will be counted as different conformations. For a large r, the view is very coarse, where even large deviations between two structures are tolerated and these are assumed to represent the same conformation. The effect of different r values could be detected in different analyses (see for example Figures. 4.16-4.19), where in general, a larger r results in higher overlap. For instance, the simulation time which is required to obtain a conformational overlap $O_{\text{conf}} \geq 0.99$ as a function of r decreases following approximately a power-law (Fig. 4.24). This could naively be understood that the threshold r can be adjusted freely to larger values to achieve large overlap values and thus convergence.

Of course, this is not the case. To demonstrate this issue, we analyze again the end-end distance estimate $\langle D \rangle$ of six cMD trajectories of Met-Enkephalin with 1 μs lengths, this time for different setups maintaining a constant $O_{\text{dens}} \approx 0.8$ [37]. All six cMD trajectories are individually taken for the reference and comparison trajectories $K = L = \{l_1, l_2, l_3, l_4, l_5, l_6\}$. This density overlap O_{dens} is obtained for different tuples of the threshold r and the simulation time t, which are given in Table 4.9. Now, it is possible to plot the end-end distance $\langle D \rangle$ as a function of r illustrated in Fig. 4.41. Larger errors of $\langle D \rangle$ correspond to coarser r. This result can be explained by the trivial relation, where small thresholds r also correspond to long simulation times t and vice versa. It is clear that a low resolution (large r) will lead to inaccurate estimates of observables because short simulation times will usually lead to incomplete sampling.

For every system of interest, it is necessary to think about and choose a resolution which covers the scientific question and is acceptable for the relevant observables. If small deviations are irrelevant, for instance in systems where end states are separated by a large distance and only the density in these different regions should be measured, it might be appropriate to choose a low resolution (large r). This will be briefly discussed in the outlook in chapter 5.

4.9. Conclusion

In this work, the sampling of MD simulations of flexible biomolecules was studied and evaluated. We have developed and implemented two new overlap measures, the conformational O_{conf} and density overlap O_{dens}. For a comprehensive assessment of the sampling, we also used the development of the number of clusters N_C [80] and cluster distribution entropy S_C [32].

In general, we could show the impact and necessity of a multi-trajectory approach for

highly flexible systems. We could evaluate that the MD sampling of the small pentapeptide Met-Enkephalin converges in the order of microsecond trajectories. Furthermore, the MD sampling of V3 considering trajectories of about 200 ns length is far from being converged. This could be shown without misinterpretations.

It could also be shown that enhanced sampling algorithms (like aMD and sMD) can significantly accelerate the sampling and yield good indications whether conformational space was missed (Fig. 4.26). This can especially make a difference, where conformational transitions are suppressed by large energetic barriers and conventional MD runs stay trapped in few energetic minima. Nevertheless, the re-weighting of such biased ensembles produced by aMD or sMD is still unresolved and requires a lot of effort to minimize the errors (Figs. 4.34-4.35). On the other hand, these accelerated trajectories can be used to generate (multiple) independent starting structures to initialize new cMD runs, which significantly increase the sampling quality and show whether parts of the conformational space are still undetected.

The two overlap measures (O_{conf}, O_{dens}) can be applied to different data as well as discontinuous samples. The only condition is that the data must be comparable, similar to the RMSD (or distance) for different structures. Then, it is possible to first detect, whether the data cover the same (conformational) space with O_{conf}, and if this is true, analyze the probability density functions with O_{dens} for self-consistency. If we reach also $O_{\text{dens}} = 1$ (for high resolution, small r), then all trajectories are equivalent and it does not matter which one is used for the extraction of thermodynamic properties. The impact of increasing O_{dens} could clearly be shown in the decreasing error of the end-end distance averages (Fig. 4.39). In practice, it might not be necessary to reach $O_{\text{dens}} = 1$ for all thermodynamic observables, but multiple trajectories reaching $0.8 \leq O_{\text{dens}} \lesssim 1$ can be used as replicates in the evaluation and error treatment of thermodynamic averages.

The (density) overlap is a very strict quantity to assess the convergence of the sampling, especially if multiple trajectories are submitted individually. The ratio in Eq. (3.11) between the minimum and maximum will drastically drop if only one trajectory samples completely different parts of the conformational space. But this is exactly what we want to obtain to not overestimating the sampling quality: if only one MD run (which provides physically meaningful results) shows a totally different behavior than other trajectories, a large conformational space is missed, and indeed the sampling should be questioned. On the other hand, due to simple stochastic reasons, O_{dens} will usually decrease with increasing number of trajectories, if convergence is not reached, yet. This is true, because different trajectories will slightly produce different probability density functions and the

minima to maxima will deviate in Eq. (3.11). There are a lot of aspects which can have an effect on O_{dens}. We did not address the dependence of overlaps as a function of the amount of trajectories in detail. But, in the regime where all conformational space was detected and the probability density functions of different experiments relate to each other, O_{dens} will be increased with longer simulation lengths.

Furthermore, we could show that a comprehensive assessment of trajectories is necessary to include different aspects of convergence. The overlaps alone lack information about the size of the (sampled) conformational space, the development of the number of clusters N_C misses information about the underlying distribution and the constancy of the cluster distribution entropy S_C are almost not able to compare different trajectories. Only the combination can yield a complete picture and enables the conclusion about the sampling quality. For instance, low O_{dens} may be caused by detecting new conformational space or by insufficient long equilibrium sampling. This can be clarified by either a constant or increasing N_C discovery. But one has to keep in mind that sampling convergence indeed will result in $O_{\text{dens}} \to 1$ and converged N_C, but the opposite does not need to be equivalent. Large density overlaps and converged N_C might also result from trapped trajectories in low energy minima, whereas parts of the conformational space separated by large activation barriers could be still missed. Nevertheless, the use of more trajectories lowers the probability to miss parts of the conformational space and makes results much more reliable. The additional combination of enhanced sampling to decrease energetic barriers makes the outcomes even clearer.

Finally, our tool worked unproblematic for both studied molecules. Met-Enkephalin yielded the expected good results and convergence, although it has a non-trivial flexible behavior [125, 127]. On the other hand, V3 is about 7-fold larger in sequence than Met-Enkephalin and did not show any sign of convergence. In fact, V3 lacks a classical description of a rigid structure due to its flexibility, but in comparison to the huge complexes routinely simulated today, it is still a small molecule. However, we can conclude from our results that MD simulations of such flexible systems are still severely limited by the available calculation resources and the conformational space grows exponentially with their complexity.

5. Summary and future directions

Studying flexible biomolecules and describing their functions and physical properties are fundamental not only to understand the functional principles of life [1–4, 8] but also be able to treat and generate inhibitors for viral diseases such as the HIV infection mediated by the host entry [142, 143, 152]. The dynamics of such systems undergo complex conformational changes and molecular dynamics (MD) simulations are a good candidate to shed light into this field at atomic resolution. But still, there are a lot of studies extracting thermodynamic properties of systems from single trajectories and/or MD runs of about 100 ns without proper validation of the underlying sampling. On the other hand, sampling assessment is also often based only on single trajectories, dimensional subsets or pre-clustering without validation [27, 29, 30, 32–34, 80]. There is the question, whether this is valid for highly flexible biomolecules with rugged energy landscapes. Therefore, we have studied the validation and quality assessment of molecular dynamics (MD) sampling for flexible biomolecules. In this work, we could show that for highly flexible systems, it is crucial to assess the convergence of the sampling as precondition. Additionally, we could see that single trajectory conclusions can easily be misinterpreted.

We aimed to develop a universal tool using a multi-trajectory approach to assess the sampling quality. We implemented two different overlap measures, namely, the conformational overlap O_{conf} and density overlap O_{dens} along two established quantities [32, 80] to investigate the convergence of a diverse set of multiple trajectories, simultaneously. The two overlap measures quantify the self-consistency of sampling of two or multiple trajectories ranging from 0 (no overlap) to 1 (perfect overlap and reproducibility), and do not require any pre-processing which could be part of information loss. Our tool is freely available as source code at `https://github.com/MikeN12/PySamplingQuality` [37] and is applicable to different systems and datasets as long as one can extract distance based measures between experiments. Here, we use the root mean square deviation (RMSD) as the distance measure for different structures obtained in the course of the simulation. Depending on the similarity between structures, which is defined by a neighboring threshold r, the conformational overlap O_{conf} counts, whether there is at least one r-neighbor of all trajectories for all corresponding simulated (reference) frames. If this is the case, all trajectories cover the same conformational space. Then, the density overlap O_{dens} counts the density of structures coming from different trajectories in each neighborhood

of all simulated (reference) frames. Only if these densities are the same, the probability density function $p(\vec{r})$ of all trajectories are the same and the sampling is sufficient, assuming that no conformational space is missed. The neighboring threshold r is used as a resolution measure: the smaller r, the similar must be two structures to be considered the same. Thus, the resolution is very high. The larger r, the more tolerant is the measure and also large deviations between structures are considered to originate from one conformation; low resolution. Amongst the overlap quantities, we monitor the size of the (sampled) conformational space, with the development of the number of found clusters N_C [80] and constancy of the cluster distribution entropy S_C [32]. For these measures, we also implemented a simple clustering algorithm to partition the conformational space into disjunct chunks with a radius $r/2 \lesssim R \leq r$ with focus on efficient applicability to huge RMSD matrices. The development of N_C and S_C allow us to conclude, if we are still in the time regime of detecting new conformational clusters or already sampling equilibrium probability.

Furthermore, we included two different enhanced sampling methods as additional possibilities to investigate the sampling quality, namely accelerated MD (aMD) [111, 112] and scaled MD (sMD) [113]. For this purpose, we also implemented three different re-weighting schemes: exponential, Maclaurin expansion and a mean-field based re-weighting.

Two different biomolecules were investigated, the small pentapeptide Met-Enkephalin and the highly flexible V3-loop of gp120 coming from HIV-1. The first molecule yielded very good results in the sampling quality assessment, as we used it as benchmarking system to validate our tool. We found that convergence can be obtained within a timescale of microseconds with conventional MD simulations, which is larger than simulations of about 100 ns applied in typical MD studies, today. The enhanced method aMD can accelerate the sampling but a proper and correct re-weighting to diminish the bias is still an unresolved issue [114, 115, 118, 119, 122, 215]. But with O_{dens}, we were able to develop a criterion to compare different sampling methods and successfully detect bias in distributions. On the other hand, an accurate calculation of thermodynamic averages like the end-end distance average $\langle D \rangle$ do not necessarily need $O_{\text{dens}} = 1$, but it converges already after ≈ 600 ns with a small error estimate.

The results of V3 did not show any reasonable sign of convergence for 200 ns trajectories. Although it is 7-fold larger than Met-Enkephalin, it represents still a small system compared to the complexes simulated today in standard MD studies. The conclusion would be that for such flexible molecules, which in fact lack of a well-defined rigid structure, MD is still limited by the available resources, since hundreds of microseconds or

milliseconds are necessary in multiple experiments to achieve convergence. Here, aMD impressively showed the poor sampling of the V3 trajectories, which all independently found new conformational clusters without visiting conformations from different trajectories. Hence, although the re-weighting is difficult and then yield wrong thermodynamic observables, we encourage to use aMD to quickly explore a huge conformational space. This gives the possibility to delimit the size and generate meaningful independent starting conformations, to be able not to miss relevant and important parts in the conventional MD sampling.

Finally, we want to underline the importance to use multiple independent trajectories starting from different initial conditions to be able to make substantial conclusions about the sampling. Only a comprehensive study of the sampling which involves a combination of the overlap measures and the clustering, can give a complete picture of the sampling result. Flexible systems have a huge conformational space with a lot of degrees of freedom, which were underestimated in the past. A development of guided strategies to overcome these issues and be able to representatively simulate such flexible complexes is needed. At least, it is necessary to detect insufficient sampling and perform the validation as established pre-condition, otherwise extracted thermodynamic properties are or may be completely meaningless. This can now be done with our tool.

Protein-ligand systems: In this thesis, we focused on single peptide MD simulations to classify their sampling and analyze the convergence of multiple trajectories. The underlying idea is to detect identical/similar structures within a certain r-neighborhood (resolution) and compare the densities of different trajectories in these r-neighborhoods. If the densities of structures of each individual trajectory do correspond to each other, the density overlap O_{dens} will be equal to one, and the sampling is complete, assuming that no conformational space is missing. This principle must also be true for more complicated systems or coherent complexes, because only if multiple experiments are able to reproduce same conformations or bound states, the sampling can be exhaustive.

For protein-ligand systems, which are another field of MD studies, a different spectrum of application might be interesting, because the RMSD values might be dominated by the large, probably rigid receptor.

First, optimally superimposed ligand structures without receptor can be used to obtain an overlap measure of ligand conformations. On the one hand, perfect sampling must reproduce the probability of different binders. On the other hand, this can be used to investigate different binding conformations if different binding sites yield different ligand structures. This approach is straight-forward using our tool *PySamplingQuality.py*, one

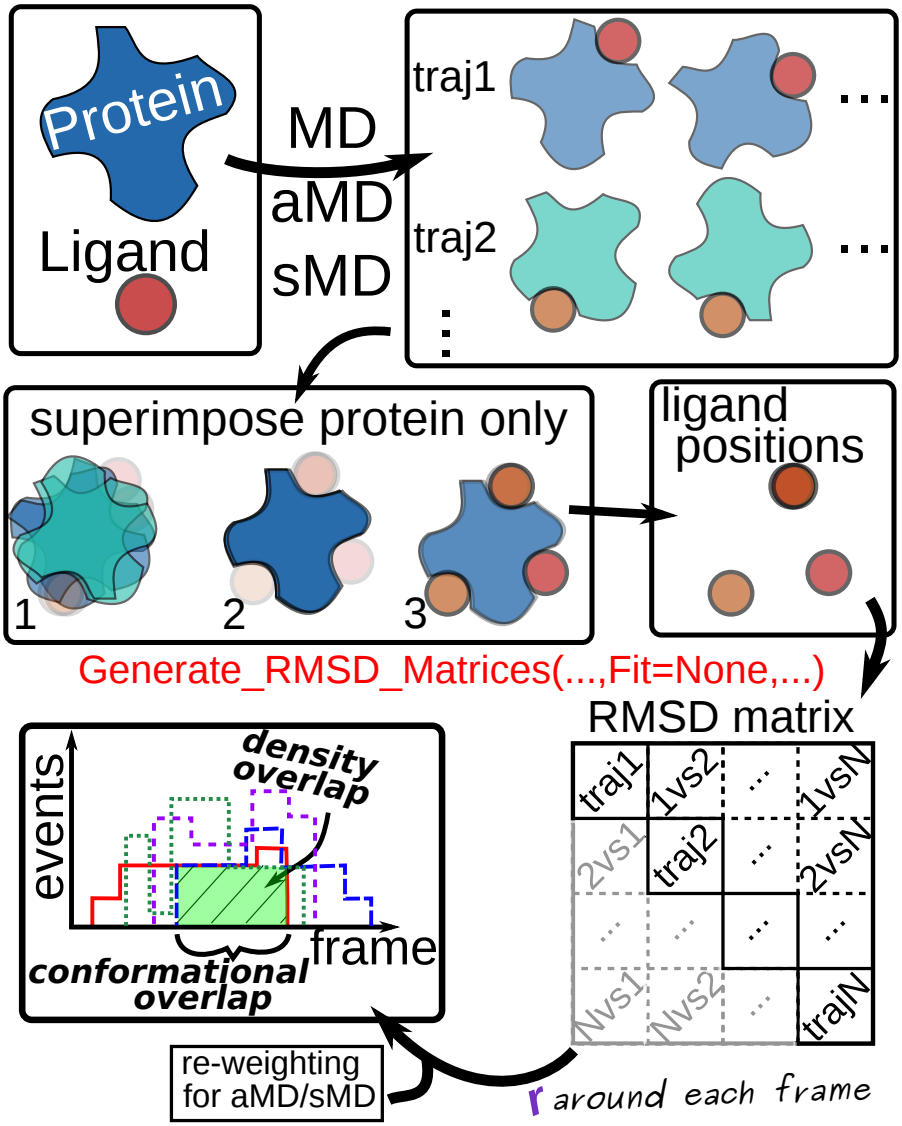

Fig. 5.1.: Workflow for protein-ligand systems. The only difference compared to the single peptide workflow is the superposition step of the ligand.

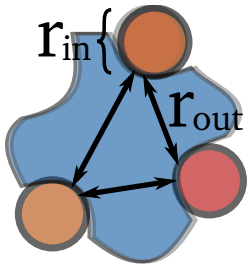

Fig. 5.2.: Different threshold r regimes for protein-ligand systems. Large r_{out} gives information about sampling the positions, only. Small r_{in} treats the binding conformations and relative positions.

only has to strip the protein atoms (along with water and ions) from the system, and submit the ligand trajectory for RMSD matrix generation, event curve and finally overlap calculation without special options or treatment. But this approach will not contain the information about the position of binding pockets, thus it is possible that O_{dens} is large but different trajectories sample different binding pockets.

Second, only the ligand dynamics can be investigated to quantify the sampling quality, assuming that the protein receptor forms a rigid core which remains almost stable without contributions to the overall dynamics. To do so, one needs to maintain the relative positions of the ligand in the system around the protein and simultaneously strip the protein atoms. The overlap measures O_{conf} and O_{dens} contain then also the relative positions of the ligand to the receptor, thus every binding position must be sampled equally well by different experiments to give large overlaps.

Third, imagine that the dynamics of the ligand is the following: Starting from an arbitrary position, being then energetically attracted by the binding pocket and guided into the bound state remaining there forever. It is conceivable that the first part of this ligand dynamics will never be reproduced by another experiments, because they both started from different but arbitrary positions, which are physically irrelevant. Only the last parts of the MD runs, where the ligand samples closely to the binding pocket, are important and should be reproducible. Thus, only last parts of the trajectories should be taken into account by setting `StartFrame` and `EndingFrame` (see Appendix C.1) as options in the tool. Additionally, different ligands or different starting positions may address different binding pockets if more than one are present for the receptor. It is unlikely that in the course of one typical MD simulation, one ligand will sample multiple or all pockets equally well. This is another application for the group-overlap concatenating multiple trajectories to consider all binding states.

A representative workflow for the second and third application possibility is illustrated in Fig. 5.1. There are two major differences compared to the general workflow of *PySamplingQuality.py* (Fig. 3.13): The optimal superposition is only done for the protein core without the ligand as intermediate step before the actual RMSD matrix generation, and one has to set `Fit=''None‘‘` (see Appendix C.1) for the RMSD matrix generation. The latter ensures that no further fitting/superposition is done and the absolute distances of the ligand are kept.

Finally, we want to mention the role of the threshold r and therefore the resolution of these analyses. Effects like induced fit [7], which might change the distances or absolute positions of the binding pockets, might alter the overlaps, because they can contribute

to the relative positions between bound ligands. These effects can be reduced with an appropriately set threshold r, which resolves these deviations. Furthermore, more tolerant r values are conceivable, because ligand pockets might be separated by their relative distances in space. The latter idea can be exploited for another application, if the end bound states of the protein-ligand systems are extracted due to a guided MD simulation. Then, one will obtain multiple bound states, which can be first investigated with a large threshold $r_{\text{in}} < r \leq r_{\text{out}}$ yielding the information, whether the binding pockets are exhaustively sampled, followed by a low threshold $r \leq r_{\text{in}}$ giving insights about the sampling of different binding conformations (Fig. 5.2). Surely, a small r includes also the relative distances in space, but if the overlaps are small for small r, the other regime $r_{\text{in}} < r \leq r_{\text{out}}$ might explain the reason.

Note that symmetry is not explicitly treated by our tool. This means that if ligands are symmetric and bind in different but symmetric ways yielding identical binding affinities, this is not considered in the RMSD metric and therefore not detected by our tool. Our tool would expect the same amount of same and symmetric counterparts in the sampling, although this might not be needed in the binding experiments. For simple (complete) sampling analysis, it is still true that both configurations should give the same probability. Hence, for such a specialized case, one has to keep that in mind and/or modify such occurrences by hand. It might be resolved by flipping the atom numbers for symmetric cases to one representative orientation. But such special cases are not implemented.

Other scientific studies: As already mentioned, our tool is universally applicable to various datasets. For an extensive experiment producing a huge amount of comparison data, it is conceivable that the overlap measures can give insight into the behavior of the data. This could be a large sequencing dataset, or large samples of patients or results from even other scientific fields. The only necessary condition is a definition of a comparability matrix, which is then transformed into different event curves per experiment, where the conformational and density overlap are calculated from, using the tool. This allows the comparison of large sets from different experiments to investigate the similarity.

Appendix

A. RMSD: fitting and superposition

We use the root mean square deviation (RMSD) Eq. (3.1) as the difference measure between different structures/conformations to classify the overlap and therefore the sampling quality of different trajectories. It is essential that these values are correct and precise. On the other hand, we want to use established tools for simplicity and to quickly transform one or multiple trajectories into RMSD matrices. Hence, we provide the usage of **g_rms** from *GROMACS* (v4.6 and v5.1 are tested) [94] and **rms2d** from *AmberTools*14 [46], two well-established simulation softwares, for the matrix generation. This matrix generation is split into two parts, first an optimal super-positioning of structures, second the RMSD calculation. The (least-squares) superposition can be difficult and time-consuming depending on the system [198]. Furthermore, in general the RMSD does not follow the triangle inequality, which means that RMSD(A, B) between two structures A and B can be significantly different if another structure C is used as reference for the super-position instead of superimposing A and B directly [216].

We test these influences by constructing RMSD matrices for our two molecules V3 and Met-Enkephalin between structures of one arbitrarily chosen trajectory (Fig. A.1) and between structures of two trajectories (Fig. A.2). First, the matrices are constructed with **g_rms** from *GROMACS* choosing one arbitrary reference structure, and second by superimposing pair-wise two structures and construct the full matrix by hand. The difference between the pair-wise fit and the **g_rms** from *GROMACS* construction is $< 10^{-4}$ which is equal to the precision of the chosen trajectory files. Thus we conclude that for our molecules it is sufficient to use the standard techniques from *GROMACS* or *AmberTools*14, which both give deviations of $< 10^{-4}$ (results for *AmberTools*14 not shown).

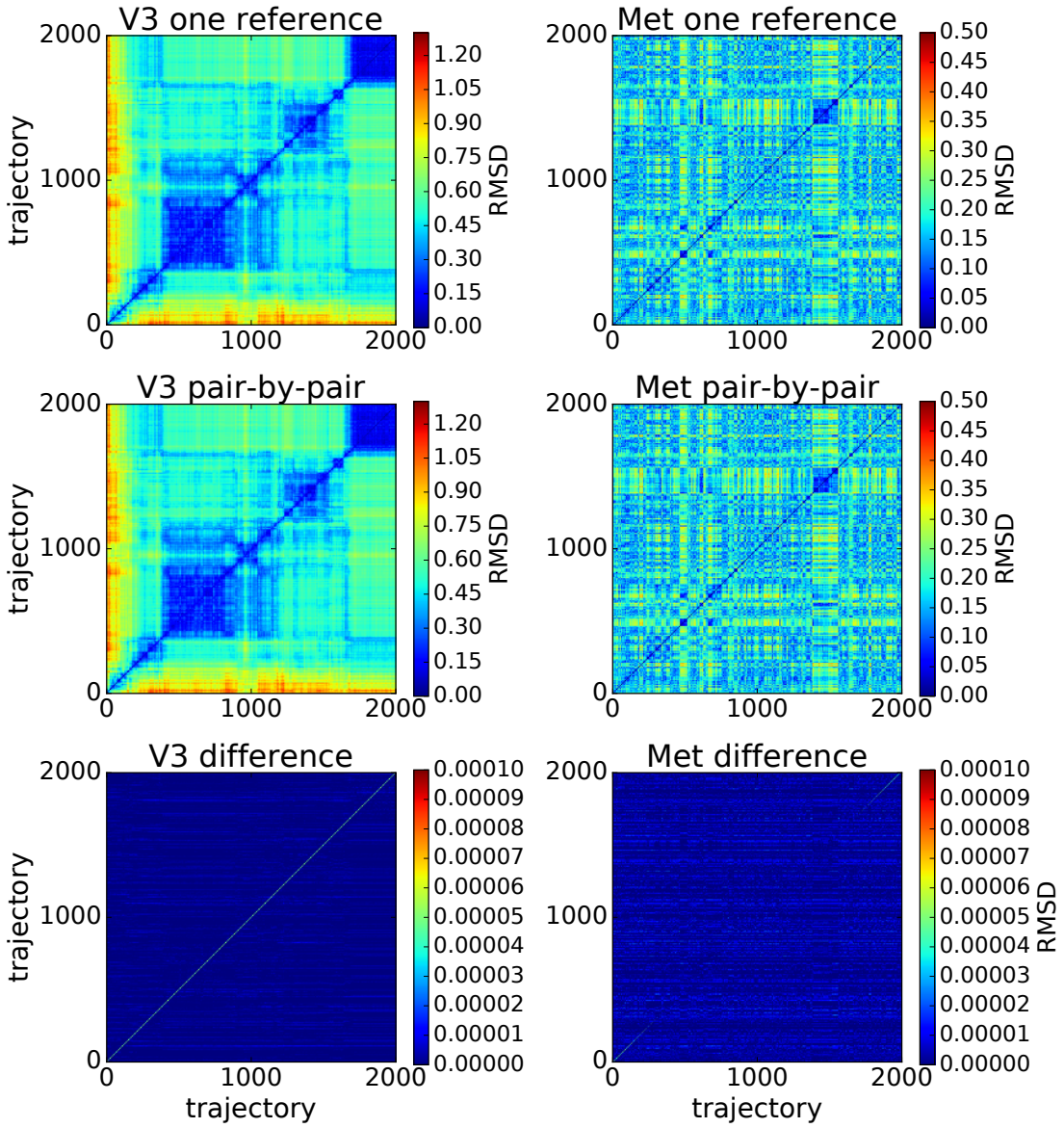

Fig. A.1.: RMSD [nm] values for explicit pair fit vs. *GROMACS* RMSD matrix generation of single trajectories. Left: V3. Right: Met-Enkephalin. The RMSD values refer to all pair structures within one arbitrary chosen 200 ns trajectory. The upper panel shows the RMSD matrix generated by *GROMACS* using an arbitrary chosen reference frame. The middle panel shows the RMSD matrix obtained by explicit pair-by-pair fit and value calculation. The lower panel shows the difference between the upper two panels. The figure is taken from Ref. [37].

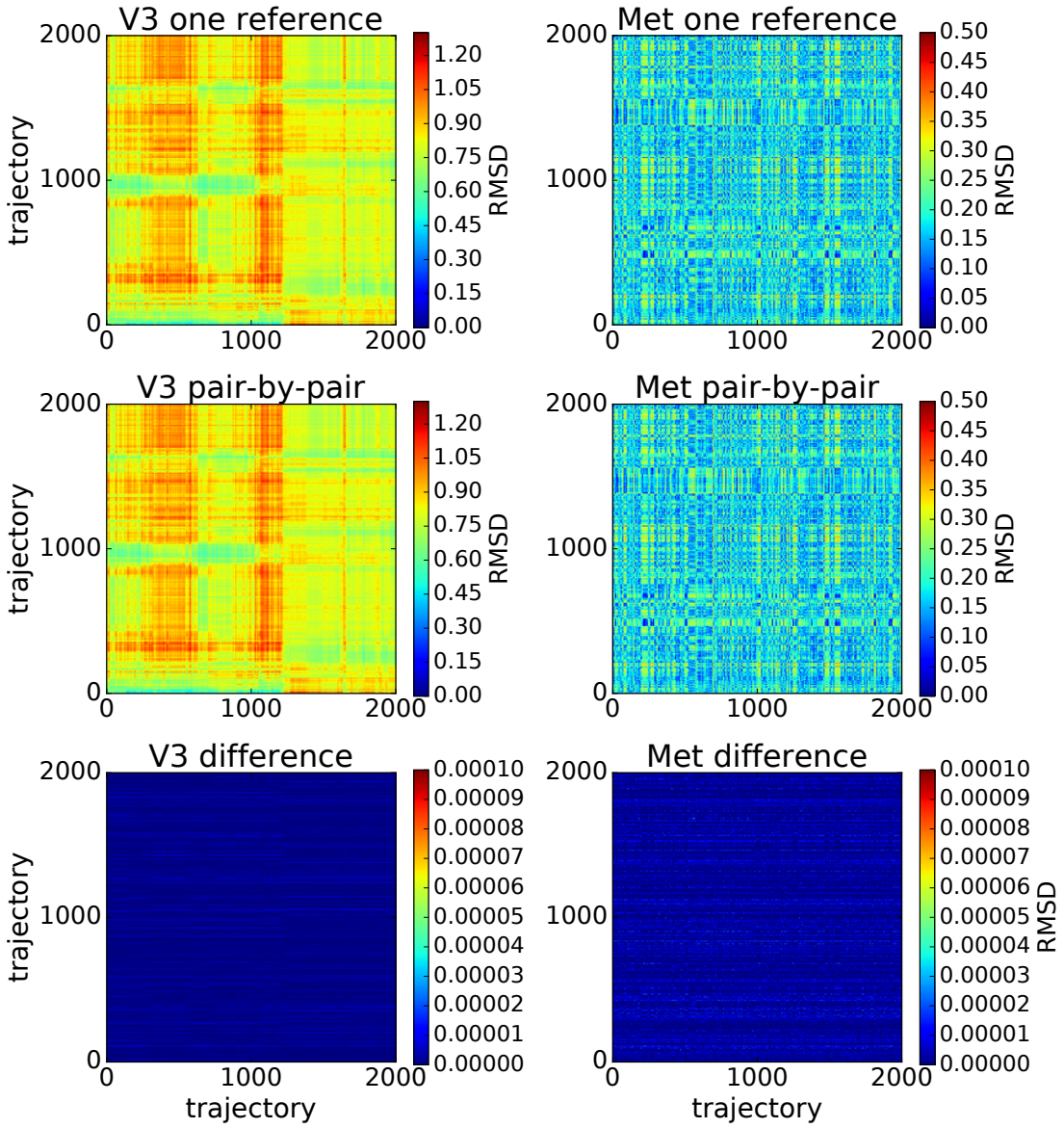

Fig. A.2.: RMSD [nm] values for explicit pair fit vs. *GROMACS* RMSD matrix generation between two different trajectories. Left: V3. Right: Met-Enkephalin. The RMSD values refer to all pair structures between two arbitrary chosen 200 ns trajectories. The upper panel shows the RMSD matrix generated by *GROMACS* using an arbitrary chosen reference frame. The middle panel shows the RMSD matrix obtained by explicit pair-by-pair fit and value calculation. The lower panel shows the difference between the upper two panels. The figure is taken from Ref. [37].

B. Boxplot representation

Boxplots are used to agglomerate data of multiple trajectories. If not specified otherwise, the boxplots show the median as line, the first and third quartile as box, the whiskers as maximal/minimal values in the data (but not extending 1.5 times the box size) and outliers outside the whiskers as single points (see Fig. B.3).

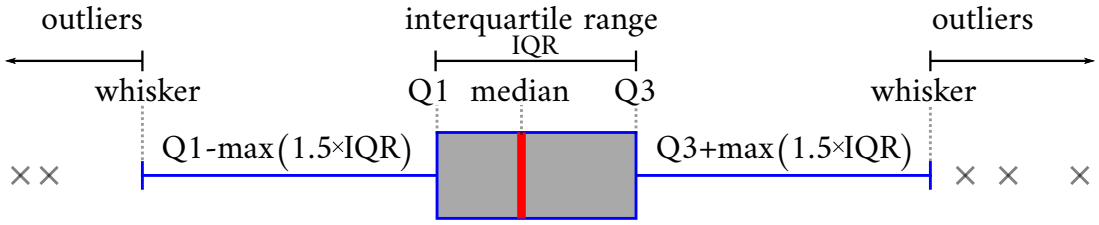

Fig. B.3.: Boxplot representation. Data distribution is shown as median (red), first and third quartiles (Q1, Q3, blue box), whiskers and possible outliers, if values extend 1.5 times the box size.

C. PySamplingQuality: modules, parameters and examples

All analyses presented in chapter 4 can be done using our tool *PySamplingQuality.py* [37]. The modules are separated into three groups: **Overlap**, **Clustering** and **Visualization**. As already mentioned, one is able to run the modules in two different ways, either using configuration files or directly in Python. Both possibilities provide the same descriptions in a *header* (Fig. C.4) and *input parameters* (Fig. C.5). The descriptions are either accessible by calling

```
from PySamplingQuality import Calc_Overlap
??Calc_Overlap()
```

directly in Python or are located in the configuration file generated by

```
python PySamplingQuality.py -module GenerateIn -in Calc_Overlap
                                                -out Calc_Overlap.in
```

The *header* (Fig. C.4) contains the current version, a short guideline and all necessary information about the specific module. The *parameter description* (Fig. C.5) contains all parameters which have to be submitted in double quotes. Additionally, there are descriptions of every parameter alongside with an example. Default parameters are automatically

set. In the following, we will show the modules and briefly discuss the corresponding arguments and their functions. All further details can be found in the specific module and the

```
#### #### ####
    # Config-File for the module <Generate_Centers_GLOBAL_singles()>.
    # Ensure, that every parameter is set with the certain format given
    # as an example (WITHIN ""). All optional parameters are initialized
    # with their default parameters.
    #### #### ####
##  ##  ##  ##  ##
### DESCRIPTION: ###
#
# v07.11.16
# - this function generates Centers_GLOBAL_singles.txt containing
#
#        TrajNr | Threshold | Nr of Clusters | Centers (1 to NrofClusters)
#
# - splitting the GLOBAL (all trajectories are concatenated) clustering
#   into different trajectories and extracting which clusters are
#   occupied by which trajectory number, assigning a single trajectory
#   clustering from the global partition
# - detecting also the Size (=Nr of clusters)
# - this allows to use a GLOBAL clustering and extract, how many
#   clusters are reached by single trajectories
```

Fig. C.4.: Header of the configuration file. It contains the version, general information and the specific descriptions about the module.

```
##  ##  ##  ##  ##
#----------------------------------------
# Directory, where effective Clustering output is located,
# e.g. 'effectiveClustering/'
ClusterDir = ""     # <STRING> format example: (non-directory) "TEXT" |
#                                              (directory) "TEXT/"
#----------------------------------------
# Clustering Name of GLOBAL effective clustering, which stores the clustering
# profile, where all trajs are concatenated e.g. 'Cluster_R0.2-0.7_GLOBAL.txt'
GlobalName = ""     # <STRING> format example: (non-directory) "TEXT" |
#                                              (directory) "TEXT/"
#----------------------------------------
# Clustering ThresholdList [nm], e.g. [0.2, 0.25, 0.3, 0.35, 0.4]
ThresholdList = ""  # <FLOAT-LIST> format example: "0.1 0.2 0.3 0.4"
#----------------------------------------
# <default None>, save directory, e.g. 'effectiveClustering/',
# if None, SaveDir = ClusterDir
SaveDir = "None"    # <STRING> format example: (non-directory) "TEXT"
#                                              (directory) "TEXT/"
```

Fig. C.5.: Parameter input of the configuration file. Here, the parameters are listed, default values are set and short descriptions are given along with a format example.

certain documentation of *PySamplingQuality.py*. We will focus on relevant parameters and will skip trivial arguments like names to store results (`SaveName`) or directories (`Dir` suffix), which have to be properly set. Additionally, same arguments, if not explicitly mentioned, are the same for following modules.

C.1. Overlap modules

RMSD matrix generation

```
Generate_RMSD_Matrices(TrajDir, TopologyDir, TrajNameList, TopologyName,
                       DistSaveDir, MatrixSaveDir ,TimeStep, Select1,
                       Select2=None, AmberHome='', GromacsHome='',
                       Fit='rot+trans', Program_Suffix='', PartList=None)
```

This module generates the necessary huge RMSD matrix between all pairs of simulated structures of all involved trajectories. To be memory and time efficient, we split the calculation into block matrices, whereas we calculate only the non-redundant upper triangular of the huge RMSD matrix together with the diagonal (shown in Fig. 3.13). The calculation is done in parallel, i.e. different blocks are calculated simultaneously on different cpu cores, using **g_rms** from *GROMACS* [94] or **rms2d** from *AmberTools*14 [46]. The necessary input parameters are first a list of trajectory names (`TrajNameList`) with the ending for the corresponding format (.trr or .xtc for *GROMACS*, .nc or .netcdf for *AMBER*). Second, `TimeStep` selects the frequency which frames of the trajectories are used. This is differently defined for the two programs: In *GROMACS*, `TimeStep` is given in nanoseconds to specify, that every `TimeStep-th` time is taken into account starting from the first frame. In *AMBER*, `TimeStep` really is a frequency, i.e. a value of 1 uses all frames, a value of 2 uses every second frame, and so on. Third, the two selections `Select1` and `Select2` define which atoms are used first for the super-position and then for the RMSD calculation. The arguments `AmberHome`, `GromacsHome` and `Program_Suffix` can be set to generate a link to the necessary programs **cpptraj** or **g_rms**, respectively, where the last argument treats possible installation suffices which were used in the *GROMACS* installation (see `www.gromacs.org` for further reading).

There are two other important arguments, namely `Fit` and `PartList`. The first is relevant if the user does not want to superimpose the structures before the RMSD calculation which can be useful for ligand systems discussed in the outlook in chapter 5. The second describes the feature that every RMSD block matrix can be split into any arbitrary size

to fit into the memory of the working machine. Every trajectory can be split by hand in multiple smaller pieces following the name convention

```
MD1.xtc -> MD1_part1.xtc, MD1_part2.xtc, ...
```

where `PartList` is a list of integers defining in how many parts the trajectories are split. Note that the trajectory names defined in `TrajNameList` must be submitted without partitions (`_partX`) to keep the strings small.

Finally, it is also possible to skip this step and supply block RMSD matrices of an own calculation (see also the outlook in chapter 5). The only requirement is that they have to match the naming convention: For trajectories called

```
ExampleName1.ending
ExampleName2.ending
...
```

the file names of the block matrices must be

```
ExampleName1_bin.dat
ExampleName2_bin.dat
ExampleName1_ExampleName2_bin.dat
...
```

to be correctly detected by the other modules. `ExampleName1_ExampleName2_bin.dat` means the RMSD values between all pairs of structures from the first trajectory `ExampleName1.ending` vs. the second `ExampleName2.ending`, whereas rows are defined by the first and columns by the second trajectory.

RMSD distribution analysis

```
determineR_using_RMSD_distributions(TrajNameList, SaveName, SaveNamePdf,
                                    SaveDir, MatrixDir, RMSD_dist_Dir = '',
                                    BinFile_precision=numpy.float32,
                                    Bins=200, Percent=1)
```

The RMSD distributions are generated using all pairs of RMSD values between all structures of all trajectories for a binned distribution of 200 bins by default. This module uses the generated RMSD matrices to extract the RMSD values, which are loaded based on the names submitted in `TrajNameList`. This list is identical to the list defined in the previous module, but the entries do not contain the ending (for example `.xtc`). Since

GROMACS (v4.6; v5.1) store RMSD matrices in a binary and *AMBER* in ascii format, `BinFile_precision` is either a float32 or float64 for single or double precision installation of *GROMACS*, or can be set to `None` for *AMBER*. Finally, `Percent` ($[0, 1]$) produces horizontal lines in the resulting figure which enclose the certain amount of probability in percent.

Event curve generation

```
Generate_EventCurves(TrajNameList, TrajLengthList, MatrixDir, SaveDir,
                     SaveName, ThresholdList, MaxNumberLines,
                     ROW_TrajNrList=None, COL_TrajNrList=None,
                     StartFrame=0, EndingFrame=numpy.infty, PartList=None,
                     BinFile_precision=numpy.float32,
                     aMD_Nrs=[], aMD_reweight='MF', aMDlogDir=None,
                     aMDlogName=None, AmberVersion='Amber14', WeightStep=1,
                     Temp=300, sMD_Nrs=[], Lambda=1, Order=10,
                     Iterations=1)
```

This function is one core module and refers to the calculation of Eq. (3.10). It generates two files, one containing the number of events per trajectory for different thresholds stored in a Python binary format for each reference frame. The other file contains descriptions in the header and the number of frames for each trajectory in one row stored in a text file, which is used to normalize the events. Here, it is necessary to submit the lengths of the trajectories (in number of frames) in a list (`TrajLengthList`) in the same order as `TrajNameList`. With `ThresholdList` it is possible to calculate events for different thresholds r and store them in one file, and `StartFrame`, `EndingFrame` select only certain parts of the trajectories. An important argument is `MaxNumberLines` which defines the number of rows loaded at once from an RMSD matrix block and therefore directly effects the memory usage of the working machine. The more rows are loaded, the faster is the calculation, but the more memory is necessary.

Another feature is the possibility to select only certain trajectories for the reference frames κ (`ROW_TrajNrList`) and for the trajectories l (`COL_TrajNrList`) of Eq. (3.10) to count the number of r-neighbors. This is done selecting certain trajectories by the position stored in `TrajNameList` starting from 1. For example, the entries `ROW_TrajNrList=[1,2]` and `COL_TrajNrList=[5,10]` lead to the calculation of the number of r-neighbors of trajectory $l = 5$ and $l = 10$ for all (reference) frames κ which come from trajectories 1 and 2.

The last eleven arguments are only necessary for re-weighting aMD and/or sMD trajectories. With `aMD_Nrs` and `sMD_Nrs` certain trajectories are marked as aMD or sMD by the position stored in `TrajNameList` starting from 1, `aMD_reweight` selects the re-weighting scheme MF, Exp or McL up to the certain order specified by `Order`. `Temp` defines the temperature of the simulation. `WeightStep` is responsible for the correct selection of weights from the generated *AMBER* files, i.e. for instance `WeightStep=2` selects every second weight, which must correspond to the frame selection determined by `TimeStep` of the previous RMSD matrix generation. If MF re-weighting is used, `Iterations` determine the number of MF iterations, whereas a value of -1 iterates the weights until convergence is reached or 100000 steps are passed. For aMD re-weighting, *AMBER* produces a special weight-file for each trajectory which stores the boost potential ΔV (see subsection 2.2.3). These have to be submitted in the same trajectory order as `aMD_Nrs` by two lists, `aMDlogDir` and `aMDlogName`.

Overlap calculation

```
Calc_Overlap(EventDir, EventNames, SaveDir, SaveName, CompareList,
             WeightDir=None, aMD_Nrs=[], sMD_Nrs=[], SameTraj=None)
```

This function is another core module and refers to the calculation of Eqs. (3.5) and (3.11). It generates only one file, containing the overlap values for different reference trajectories and different thresholds defined in the event curve file. The main input is the list of files produced in the event curve generation `EventNames`. Here, different event files are automatically merged together from different reference sets K and comparison sets L. If event files with different start and ending frames are submitted, the argument `SameTraj` allows to calculate the overlap between different simulation times of one trajectory which is defined by the position (=`SameTraj`) stored in the previous `TrajNameList` starting from 1. `CompareList` is the most tricky argument, it is a list of tuples of lists defining the comparison set L of trajectories for the overlap calculations. The inner list concatenates all trajectories similar to the group-overlap defined in subsection 3.2.3, the tuples define the different (groups of) trajectories for which the overlap is calculated and the outer list gives the possibility to store multiple overlap values in one file. For example, `CompareList = [([1],[2,3])]` leads to an overlap calculation between the trajectory 1 and the concatenated trajectories 2 and 3. Again, trajectories are defined by their positions. The other arguments are only necessary for aMD or sMD trajectories, whereas `WeightDir` has to point to the directory where the weights are stored from the event curve generation.

C.2. Clustering modules

Clustering

```
Generate_Clustering(MatrixDir, SaveDir, TrajNameList, TrajLengthList,
                    Threshold, SaveName, MaxNumberLines, TimeStep=None,
                    StartFrame=0, EndingFrame=numpy.infty, PartList=None,
                    GLOBAL=True, BinFile_precision=numpy.float32,
                    RMSDdir=None, TrajDir=None, TopologyDir=None,
                    TopologyName=None, Ending='.xtc', Select1=None,
                    Select2=None, AmberHome='', GromacsHome='',
                    Program_Suffix='', ReferencePDB=None,
                    RefFrame=None)

Merge_Clustering_different_Thresholds(SingleClustDir, SaveDir, SaveName,
                                      ThresholdList, StartFrame,
                                      EndingFrame, GLOBAL)

Generate_Centers_GLOBAL_singles(ClusterDir, GlobalName, ThresholdList,
                                SaveDir=None)
```

These three modules generate the clustering files using the calculated RMSD matrices. In the first (main) module for the clustering, two files are generated, first the clustering profile containing all frames and the corresponding clusters, second the centroid file containing the number of found clusters per trajectory with the corresponding cluster centers. Furthermore, the last twelve arguments are necessary, if an explicit reference structure is used (see subsection 3.3.1), otherwise they can be left untouched and automatically the first frame of a trajectory is used as reference point. For an explicit reference, one has to either submit the topology, the reference frame in PDB format and the ending of the trajectory (e.g. `.xtc`), or submit a specific frame as integer (=`RefFrame`). The latter specifies the corresponding frame of the submitted trajectory list as reference. Since it is difficult to parallelize cluster calculations with different thresholds efficiently, it is possible to generate multiple files per hand submitting only one threshold and merge then clusterings with different thresholds into one collected file. This can be done with the second module presented here.

Another important argument is `GLOBAL`, which will be necessary for other modules, too. It defines whether all trajectories are concatenated and one single (global) partitioning is

done at once. The great advantage is that one can extract the number of clusters which are reached by single trajectories with the third module by maintaining the comparability of a global partition. One only needs to submit the file which stores the globally partitioned profiles as `GlobalName`, then a corresponding centroid file is generated containing the number of found clusters per trajectory from the global clustering with the corresponding cluster centers.

To run the partitioning, it is necessary that at least one row of the full RMSD matrix of the involved trajectories fits into the memory of the system (see subsection 3.3.1). In the case of the global clustering, this really means the combination of all block matrices, thus it might consume a lot of memory, whereas for local clustering only one row of one trajectory must be loaded.

Cluster number N_C and entropy S_C

```
Generate_CDE_to_File(ClusterDir, ClusterFile, ThresholdList, Case,
                     SaveDir=None, SaveName=None, WeightDir=None,
                     aMD_Nrs=[], sMD_Nrs=[], aMD_reweight='MF',
                     Iterations=1, Lambda=1, Order=10)
```

The cluster distribution entropy $S_C(t)$ and the number of clusters $N_C(t)$ as a function of the simulation time t (see subsection 3.3.3) for each trajectory are stored in a one file using the clustering profile (single or global partitioning) as input `ClusterFile`. The important argument is defined by `Case`, which distinguishes between single trajectory clustering (`Case='LOCAL'`), global clustering (`Case='GLOBAL'`) and global clustering but extracting the results for single trajectories (`Case='GLOBAL_singles'`). This has to correspond to the submitted clustering file. Additionally, the normalized versions $\tilde{N}_C(t)$, $\tilde{S}_C(t)$

$$\tilde{N}_C(t) = \frac{N_C(t)}{N_C(t_{\text{end}})} \qquad \in [0,1]$$

$$\tilde{S}_C(t) = \frac{-\sum_{i=1}^{N_C(t)} p_i(t) \cdot \log\left(p_i(t)\right)}{\log\left(N_C(t_{\text{end}})\right)} \qquad \in [0,1]$$

are calculated, with t_{end} means the end of the simulation.

Slopes defined by dN_C/dt and dS_C/dt

```
Generate_Slope_Error(EntropyDir, EntropyName, SaveDir=None, SaveName=None,
                     SlopeTimeArray=[100,250,500], X_NormFactor=1000)
```

The slopes and the corresponding error estimates of the linear models are calculated by submitting the `EntropyName` file generated from the previous module. It generates one file containing the slopes per trajectory for different thresholds which corresponds to the file generated by the previous function. One has to specify three different frame values (`SlopeTimeArray`), where the slopes are then calculated for the corresponding last frames of the trajectory. The `X_NormFactor` is used to normalize the x-axis, i.e. the number of clusters or entropy is approximately changed in the next steps defined by this value by the corresponding slope.

C.3. Visualization modules

Plot clustering results

```
Plot_ClusterProfile(ClusterDir, ClusterFile, TimeStep, Threshold,
                    TrjLenList, GLOBAL, SaveDir=None, SavePDF=None,
                    Names=[], FigSize=[16,8])
```

The profile of the clustering can be visualized using the clustering profile submitted as `ClusterFile` (single trajectory or global partitioning). It monitors which cluster is occupied during the course of the simulation. The more transitions between different clusters, the lower are the energetic barriers in between. Furthermore, one can detect the development of finding new clusters. The argument `GLOBAL` (True or False) switches between cases, whether trajectories are treated as concatenation or separately. It is possible to generate profiles for more than only one trajectory, which is another quantity to assess the sampling quality: Only if the same clusters are present in multiple trajectories with similar densities and transition frequencies, the sampling can be complete. Depending on the total number of found clusters and trajectory lengths, this can produce huge files, where the figure size can be modified by integers referring to inches in x- and y-directions.

```
Plot_Slope_Error_Plateau_NrClust(SlopeDir, SlopeName, Threshold, Case,
                                 TimeStep, SaveDir=None, Confidence=0.95,
                                 YMAX=50, Splitter=None, SupGrid=None,
                                 TrajExcept=[], FigText=None)
```

This module plots the slopes (`SlopeName`) of dN_C/dt and dS_C/dt with the 95% confidence intervals on default calculated by `Generate_Slope_Error()` and defined in subsection 3.3.3 (see for example Fig. 4.30). One frame of the trajectory refers to the floating point value of `TimeStep` in nanosecond. In this module, `Case` switches between 'Entropy',

'Cluster' and 'Plateau', where the last choice gives the time Δt, between finding the last cluster and the end of the simulation. This is another indicator, whether $N_C(t)$ might be converged. The smaller Δt, the more probable it is that more clusters will be found.

The last four arguments are tricky. `SupGrid` and `Splitter` manipulates the layout, how many rows and columns are shown with how many trajectories in each subplot. If these options are used, one can name each panel from left to right and top to bottom by the list `FigText`. Finally, with `TrajExcept` one can select the trajectories which are discarded from the list of `TrajNameList` used to generate `SlopeName`, starting from 1.

```
Plot_ClusterSize_vs_Time_GLOBAL(ClusterDir, ClusterFile, Threshold,
                                StartEndList, TrajGrpList, SaveDir=None,
                                SaveName=None, SndAxis=2, LegendList=None,
                                YLIM=None, FigSize=(12,5))
```

The number of clusters as a function of the simulation time is a good indicator, how different trajectories behave and whether simulations explore different parts of the conformational space. This requires the comparability of the clustering, thus the module uses the global partitioning of all full length concatenated trajectories and then extracts the number of unique clusters reached by single runs. The plot shows the groups of trajectories as boxplots, where the groups are defined by lists of lists called `TrajGrpList`. Furthermore, the unique clusters found by all combined trajectories of one list are shown as bars, and `SndAxis > 1` generates a second x-axis giving the number of clusters found by all runs. For example, `TrajGrpList =[[1,2], [3,4]]` plots the results for trajectories 1 and 2 and for trajectories 3 and 4 as separate boxplots, and additionally the barplots show the number of unique clusters of the combination of 1 and 2 and the number unique clusters of the combination of 3 and 4. The argument `StartEndList` defines in tuples the starting and ending frame for which the number of clusters are evaluated. Finally, the list called `LegendList` can be used to submit names starting first for all boxplots followed by all barplots.

Plot overlap results

```
Plot_Overlap_VS_Threshold(OverlapDir, OverlapList1, Percentile1=25,
                          Percentile2=75, Median=False,
                          Interpolation='linear', OverlapList2=None,
                          XLIM1=[None,None], XLIM2=[None,None],
                          MolName1='', MolName2='', LegendList=[None],
                          SaveDir=None, SaveName=None)
```

`Plot_Overlap_VS_Threshold()` generates plots of the kind like Fig. 4.16. The red curves and first x-axis correspond to overlap files in list `OverlapList1` and can be manipulated by `XLIM1` setting the limites for the x-axis and by `MolName1` setting the name of the first molecule/system. The legend referring to the same order as `OverlapList1` and `OverlapList2` can be set by `LegendList`. One special feature is that if one overlap file contains multiple overlap values for instance for different analysis groups, they are plotted automatically in ascending order before using the next file in the list `OverlapList1` or `OverlapList2`. The arguments `Percentile1`, `Percentile2`, `Median` and `Interpolation` modify the shown error bars.

```
Plot_HeatMap_1vs1(OverlapDir, OverlapFile, Threshold, StartFrame,
                  EndingFrame, YLIM=None, ClusterDir=None,
                  ClusterFile=None, AllPrject=True, TrajExcept=[],
                  Title='', Grid=[], CaseTitles=[], SaveDir=None,
                  SaveName=None)
```

The heatmap is a good possibility to illustrate the pair-overlap (conformational or density) between a massive amount of trajectories l_X and l_Y as shown in Fig. 4.17. The file specified by `OverlapFile` must contain all pair overlap values in one row, thus the overlap calculation must be done properly specifying all pairs in the following order: Starting calculating all pairs with respect to the first trajectory, then with respect to the second, then with respect to the third, and so on without redundant or multiple same entries. With `AllPrject`, one is able to specify whether the heatmap is symmetric ($K = L$) or asymmetric ($K \neq L$) as done in subsection 4.4.1. Additionally, it is possible to show the number of clusters found by single trajectories below the heatmap, either submitting the centroid file from local clustering, or the clustering profile from global clustering to choose between both approaches. For visual reason, you might sort different trajectory groups together and separate them by a grid, where solid lines are shown after the i-th trajectory specified in the list `Grid`. These groups can be named by `CaseTitles`, where this list has to obviously have one more entry than the grid list.

```
Plot_HeatMap_as_Dendro(OverlapDir, OverlapFile, Threshold, Case='density',
                       TrajExcept=[], Labels=None, Colors=None,
                       SaveDir=None, SaveName=None)
```

Additionally to the heatmap representation, one can generate a dendrogram using hierarchical clustering with average linkage with the same input as previously defined. The argument `Case` switches between the density and conformational overlap. With

`Labels` and `Colors`, the labeling of the leaves can be modified, whereas the latter is represented as a dictionary, where the keys must match the first part of the labels separated by spaces defined by `Labels`. For example `Labels = ['traj 1', 'traj 2']` and `Colors = {'traj':'g'}` will lead to green colored leaves called *traj 1* and *traj 2*.

```
Plot_Overlap_VS_Time(OverlapDir, OverlapList, Threshold, StartEndList,
                     TimeStep, Percentile1=25, Percentile2=75, Median=False,
                     Interpolation='linear', LegendList=[], Title='',
                     LegendNcols=1, SaveDir=None, SaveName=None,
                     logX=False, LegendDens=True)
```

This module generates figures like Fig. 4.22, plotting the overlap as a function of the simulation time defined in `StartEndList`. Moreover, it is possible to plot the time logarithmic by `logX`, setting the number of columns for the legend by `LegendNcols` and choose at which panel (conformational or density) the legend will appear by `LegendDens`. As an example, `StartEndList=[(0,100), (0,200)]` will result in a plot showing the overlap for the first 100 and the first 200 frames.

```
Plot_Overlap_VS_Cluster(OverlapDir, OverlapList, Threshold, ClusterDir,
                        ClusterFile, Case='density', XLIM=None, YLIM=None,
                        LegendList=None, LegendNcols=1, Percentile1=25,
                        Percentile2=75, Median=False,
                        Interpolation='linear', SaveDir=None,
                        SaveName=None, Title='', FigSize=(7,6), Combi=True,
                        Symbols=['bs', 'ks', 'rs', 'gs', 'ko', 'ro', 'go',
                                 'k<', 'r<', 'g<', 'g<', 'm<', 'c<', 'y<'])
```

Finally, the last plotting function generates a figure combining the overlap and clustering result (Fig. 4.31). The input files are the same as before, where `OverlapList` and `ClusterFile` must correspond to each other, i.e. the trajectories and other properties must be the same for both files. The special feature is that one can either illustrate N_C as the number of unique clusters found by all combinations of trajectories in the analysis group (`Combi` = True), or show the distribution of number of clusters for each trajectory separately by the average or median with the specified percentiles (`Combi` = False). Therefore, the outcomes of both settings are different. Furthermore, the user can specify the color and marker of single points following the matplotlib [204] logic, which contains two characters: The first character modifies the color and the second the marker, which can be seen in the default option.

Bibliography

[1] Anfinsen, C. B., Haber, E., Sela, M., and White, F. H. *The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain.* Proc. Natl. Acad. Sci. U. S. A. (1961) 47:1309–14. 1, 4, 134

[2] Perutz, M. F. *Stereochemistry of Cooperative Effects in Haemoglobin: HaemHaem Interaction and the Problem of Allostery.* Nature (1970) 228(5273):726–734. doi:10.1038/228726a0. 1, 4, 134

[3] Boyer, P. D. *The ATP Synthase - A Splendid Molecular Machine.* Annu. Rev. Biochem. (1997) 66(1):717–749. doi:10.1146/annurev.biochem.66.1.717. 1, 4, 134

[4] Abrahams, J. P., Leslie, A. G. W., Lutter, R., and Walker, J. E. *Structure at 2.8 Â resolution of F1-ATPase from bovine heart mitochondria.* Nature (1994) 370(6491):621–628. doi:10.1038/370621a0. 1, 4, 134

[5] Fischer, E. *Einfluss der Configuration auf die Wirkung der Enzyme.* Berichte der Dtsch. Chem. Gesellschaft (1894) 27(3):2985–2993. doi:10.1002/cber.18940270364. 1

[6] Bosshard, H. R. *Molecular recognition by induced fit: how fit is the concept?* News Physiol. Sci. (2001) 16:171–3. 1

[7] Koshland, D. E. *Application of a Theory of Enzyme Specificity to Protein Synthesis.* Proc. Natl. Acad. Sci. U. S. A. (1958) 44(2):98–104. 1, 138

[8] Tsai, C. J., Ma, B., and Nussinov, R. *Folding and binding cascades: shifts in energy landscapes.* Proc. Natl. Acad. Sci. U. S. A. (1999) 96(18):9970–2. 1, 4, 134

[9] Blakeley, M. P., Hasnain, S. S., and Antonyuk, S. V. *Sub-atomic resolution X-ray crystallography and neutron crystallography: promise, challenges and potential.* IUCrJ (2015) 2(4):464–474. doi:10.1107/S2052252515011239. 1

[10] Dellisanti, C. *A barrier-breaking resolution.* Nat. Struct. Mol. Biol. (2015) 22(5):361–361. doi:10.1038/nsmb.3025. 1

[11] Czarnocki-Cieciura, M. and Nowotny, M. *Introduction to high-resolution cryo-electron microscopy.* Postepy Biochem. (2016) 62(3):383–394. 1

[12] Hwang, S. S., Boyle, T. J., Lyerly, H. K., and Cullen, B. R. *Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1.* Science (1991) 253(5015):71–4. 1, 31

[13] D'Souza, M. P. and Harden, V. A. *Chemokines and HIV-1 second receptors. Confluence of two fields generates optimism in AIDS research.* Nat. Med. (1996) 2(12):1293–300. 1, 31

[14] Dittmar, M. T., McKnight, Á., Simmons, G., Clapham, P. R., Weiss, R. A., and Simmonds, P. *HIV-1 tropism and co-receptor use.* Nature (1997) 385(6616):495–496. doi:10.1038/385495a0. 1, 31

[15] Huang, C.-c., Tang, M., Zhang, M.-Y., Majeed, S., et al. *Structure of a V3-containing HIV-1 gp120 core.* Science (2005) 310(5750):1025–8. doi:10.1126/science.1118398. 1, 31, 32, 68, 69

[16] Huang, C.-C., Lam, S. N., Acharya, P., Tang, M., et al. *Structures of the CCR5 N terminus and of a tyrosine-sulfated antibody with HIV-1 gp120 and CD4.* Science (2007) 317(5846):1930–4. doi:10.1126/science.1145373. 1, 29, 31, 32, 68, 69

[17] Zuckerman, D. M. *Equilibrium Sampling in Biomolecular Simulations.* Annu. Rev. Biophys. (2011) 40(1):41–62. doi:10.1146/annurev-biophys-042910-155255. 1, 2, 27

[18] McCammon, J. A., Gelin, B. R., and Karplus, M. *Dynamics of folded proteins.* Nature (1977) 267(5612):585–590. doi:10.1038/267585a0. 1

[19] Karplus, M. and Kuriyan, J. *Molecular dynamics and protein function.* Proc. Natl. Acad. Sci. (2005) 102(19):6679–6685. doi:10.1073/pnas.0408930102. 1

[20] Ponder, J. W. and Case, D. A. *Force fields for protein simulations.* Adv. Protein Chem. (2003) 66:27–85. 1, 9, 10

[21] Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. *Comparison of multiple Amber force fields and development of improved protein backbone parameters.* Proteins (2006) 65(3):712–25. doi:10.1002/prot.21123. 1, 9, 10

[22] Shaw, D. E., Bowers, K. J., Chow, E., Eastwood, M. P., et al. *Millisecond-scale molecular dynamics simulations on Anton.* Proc. Conf. High Perform. Comput. Networking, Storage Anal. - SC '09. ACM Press, New York, New York, USA. ISBN 9781605587448 (2009) 1. doi:10.1145/1654059.1654099. 1, 22

[23] Stone, J. E., Hardy, D. J., Ufimtsev, I. S., and Schulten, K. *GPU-accelerated molecular modeling coming of age.* J. Mol. Graph. Model. (2010) 29(2):116–25. doi:10.1016/j.jmgm.2010.06.010. 1, 22

[24] Beauchamp, K. A., Lin, Y.-S., Das, R., and Pande, V. S. *Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements.* J. Chem. Theory Comput. (2012) 8(4):1409–1414. doi:10.1021/ct2007814. 1

[25] Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M. P., Dror, R. O., and Shaw, D. E. *Systematic Validation of Protein Force Fields against Experimental Data.* PLoS One (2012) 7(2):e32131. doi:10.1371/journal.pone.0032131. 2, 11

[26] Romo, T. D. and Grossfield, A. *Unknown Unknowns: the Challenge of Systematic and Statistical Error in Molecular Dynamics Simulations.* Biophys. J. (2014) 106(8):1553–1554. doi:10.1016/j.bpj.2014.03.007. 2, 18, 19, 54

[27] Grossfield, A. and Zuckerman, D. M. *Quantifying Uncertainty and Sampling Quality in Biomolecular Simulations.* Annu. Rep. Comput. Chem., chapter 2, 23–48. Elsevier (2009) doi:10.1016/S1574-1400(09)00502-7. 2, 19, 134

[28] Romo, T. D., Leioatts, N., and Grossfield, A. *Lightweight object oriented structure analysis: Tools for building tools to analyze molecular dynamics simulations.* J. Comput. Chem. (2014) 35(32):2305–2318. doi:10.1002/jcc.23753. 2

[29] Flyvbjerg, H. and Petersen, H. G. *Error estimates on averages of correlated data.* J. Chem. Phys. (1989) 91(1):461–466. doi:10.1063/1.457480. 2, 19, 134

[30] Lyman, E. and Zuckerman, D. M. *On the Structural Convergence of Biomolecular Simulations by Determination of the Effective Sample Size.* J. Phys. Chem. B (2007) 111(44):12876–12882. doi:10.1021/jp073061t. 2, 19, 134

[31] Zhang, X., Bhatt, D., and Zuckerman, D. M. *Automated Sampling Assessment for Molecular Simulations Using the Effective Sample Size.* J. Chem. Theory Comput. (2010) 6(10):3048–3057. doi:10.1021/ct1002384. 2

[32] Sawle, L. and Ghosh, K. *Convergence of Molecular Dynamics Simulation of Protein Native States: Feasibility vs Self-Consistency Dilemma.* J. Chem. Theory Comput. (2016) 12(2):861–869. doi:10.1021/acs.jctc.5b00999. 2, 19, 54, 58, 60, 113, 131, 134, 135

[33] Hess, B. *Convergence of sampling in protein simulations.* Phys. Rev. E (2002) 65(3):031910. doi:10.1103/PhysRevE.65.031910. 2, 20, 134

[34] Fuglebakk, E., Echave, J., and Reuter, N. *Measuring and comparing structural fluctuation patterns in large protein datasets.* Bioinformatics (2012) 28(19):2431–40. doi:10.1093/bioinformatics/bts445. 2, 20, 134

[35] Hess, B. *Similarities between principal components of protein dynamics and random diffusion.* Phys. Rev. E (2000) 62(6):8438–8448. doi:10.1103/PhysRevE.62.8438. 2

[36] Python Software Foundation. *Python Language Reference.* Version 2.7, http://www.python.org. 2, 64

[37] Nemec, M. and Hoffmann, D. *Quantitative Assessment of Molecular Dynamics Sampling for Flexible Systems.* J. Chem. Theory Comput. (2017) 13(2):400–414. doi:10.1021/acs.jctc.6b00823. http://pubs.acs.org/doi/abs/10.1021/acs.jctc.6b00823. 3, 36, 53, 63, 68, 70, 71, 78, 82, 83, 94, 95, 98, 99, 100, 104, 106, 110, 111, 112, 113, 114, 115, 117, 118, 126, 127, 130, 131, 134, 141, 142, 143

[38] Wang, W., Donini, O., Reyes, C. M., and Kollman, P. A. *Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions.* Annu. Rev. Biophys. Biomol. Struct. (2001) 30:211–43. doi:10.1146/annurev.biophys.30.1.211. 4

[39] Ivanov, S. Theoretical and Quantum Mechanics. Springer Netherlands (2006). ISBN 978-1-4020-3365-0. doi:10.1007/1-4020-3688-4. Chapter 11. 5

[40] Hehre, W. J. A guide to molecular mechanics and quantum chemical calculations. Irvine, CA : Wavefunction, ©2003. 6, 7, 9, 11

[41] Schlick, T. Molecular Modeling and Simulation: An Interdisciplinary Guide, *Interdisciplinary Applied Mathematics*, volume 21. Springer New York (2010). ISBN 978-1-4419-6350-5. doi:10.1007/978-1-4419-6351-2. 6, 7, 9, 10, 11, 12, 14, 15, 16, 17, 68

[42] Burkert, U. and Allinger, N. L. Molecular mechanics. ACS monograph. American Chemical Society (1982). ISBN 9780841205840. 7, 11

[43] Leimkuhler, B. and Matthews, C. Molecular Dynamics, *Interdisciplinary Applied Mathematics*, volume 39. Springer International Publishing, Cham (2015). ISBN 978-3-319-16374-1. doi:10.1007/978-3-319-16375-8. 7, 11, 14

[44] Allen, M. P. and Tildesley, D. J. Computer Simulation of Liquids. Oxford Science Publ. Clarendon Press (1989). ISBN 9780198556459. 7, 8, 11, 12

[45] Verlet, L. *Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules*. Phys. Rev. (1967) 159(1):98–103. doi:10.1103/PhysRev.159.98. 7

[46] Case, D. A., Babin, V., Berryman, J. T., Betz, R. M., et al. *Amber 14* (2014). University of California, San Francisco, ambermd.org. 7, 22, 64, 71, 140, 145

[47] Hockney, R., Goel, S., and Eastwood, J. *Quiet high-resolution computer models of a plasma*. J. Comput. Phys. (1974) 14(2):148–158. doi:10.1016/0021-9991(74)90010-2. 8

[48] Scott, W. R. P., Hünenberger, P. H., Tironi, I. G., Mark, A. E., et al. *The GROMOS Biomolecular Simulation Program Package*. J. Phys. Chem. A (1999) 103(19):3596–3607. doi:10.1021/jp984217f. 9

[49] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., et al. *A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules*. J. Am. Chem. Soc. (1995) 117(19):5179–5197. doi:10.1021/ja00124a002. 9, 10

[50] MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., et al. *All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins* . J. Phys. Chem. B (1998) 102(18):3586–3616. doi:10.1021/jp973084f. 9

[51] Wang, J., Cieplak, P., and Kollman, P. A. *How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?* J. Comput. Chem. (2000) 21(12):1049–1074. doi:10.1002/1096-987X(200009)21:12⟨1049::AID-JCC3⟩3.0.CO;2-F. 9, 10

[52] Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., et al. *Improved side-chain torsion potentials for the Amber ff99SB protein force field.* Proteins Struct. Funct. Bioinforma. (2010) NA–NA. doi:10.1002/prot.22711. 9, 10, 12, 71

[53] Jorgensen, W. L. and Tirado-Rives, J. *The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin.* J. Am. Chem. Soc. (1988) 110(6):1657–1666. doi:10.1021/ja00214a001. 10

[54] Duan, Y. and Kollman, P. A. *Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution.* Science (1998) 282(5389):740–4. doi:9784131. 11, 18

[55] Lindorff-Larsen, K., Piana, S., Dror, R. O., and Shaw, D. E. *How Fast-Folding Proteins Fold.* Science (80-.). (2011) 334(6055):517–520. doi:10.1126/science.1208351. 11, 18

[56] Schwabl, F. Statistische Mechanik. Springer-Lehrbuch. Springer-Verlag, Berlin/Heidelberg (2006). ISBN 3-540-31095-9. doi:10.1007/3-540-31097-5. 11

[57] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. *Molecular dynamics with coupling to an external bath.* J. Chem. Phys. (1984) 81(8):3684–3690. doi:10.1063/1.448118. 11, 12, 71

[58] Nosé, S. *A unified formulation of the constant temperature molecular dynamics methods.* J. Chem. Phys. (1984) 81(1):511–519. doi:10.1063/1.447334. 11

[59] Hoover, W. G. *Canonical dynamics: Equilibrium phase-space distributions.* Phys. Rev. A (1985) 31(3):1695–1697. doi:10.1103/PhysRevA.31.1695. 11

[60] Pastor, R. W., Brooks, B. R., and Szabo, A. *An analysis of the accuracy of Langevin and molecular dynamics algorithms.* Mol. Phys. (1988) 65(6):1409–1419. doi:10.1080/00268978800101881. 11, 71

[61] Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. *A smooth particle mesh Ewald method.* J. Chem. Phys. (1995) 103(19):8577–8593. doi:10.1063/1.470117. 14, 71

[62] Hess, B. and van der Vegt, N. F. A. *Hydration Thermodynamic Properties of Amino Acid Analogues: A Systematic Comparison of Biomolecular Force Fields and Water Models.* J. Phys. Chem. B (2006) 110(35):17616–17626. doi:10.1021/jp0641029. 14

[63] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., and Hermans, J. *Interaction Models for Water in Relation to Protein Hydration.* Intermol. Forces, 331–342. Reidel (1981) doi:10.1007/978-94-015-7658-1_21. 14

[64] Berendsen, H. J. C., Grigera, J. R., and Straatsma, T. P. *The missing term in effective pair potentials.* J. Phys. Chem. (1987) 91(24):6269–6271. doi:10.1021/j100308a038. 14

[65] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. *Comparison of simple potential functions for simulating liquid water.* J. Chem. Phys. (1983) 79(2):926–935. doi:10.1063/1.445869. 14, 15, 71

[66] Horn, H. W., Swope, W. C., Pitera, J. W., Madura, J. D., et al. *Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew.* J. Chem. Phys. (2004) 120(20):9665–9678. doi:10.1063/1.1683075. 14

[67] Mahoney, M. W. and Jorgensen, W. L. *A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions.* J. Chem. Phys. (2000) 112(20):8910–8922. doi:10.1063/1.481505. 14

[68] Adcock, S. A. and McCammon, J. A. *Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins.* Chem. Rev. (2006) 106(5):1589–1615. doi:10.1021/cr040426m. 16

[69] Knapp, B., Frantal, S., Cibena, M., Schreiner, W., and Bauer, P. *Is an Intuitive Convergence Definition of Molecular Dynamics Simulations Solely Based on the Root Mean Square Deviation Possible?* J. Comput. Biol. (2011) 18(8):997–1005. doi:10.1089/cmb.2010.0237. 16

[70] Genheden, S. and Ryde, U. *Will molecular dynamics simulations of proteins ever reach equilibrium?* Phys. Chem. Chem. Phys. (2012) 14(24):8662. doi:10.1039/c2cp23961b. 16, 18, 101

[71] Van Gunsteren, W. F. and Karplus, M. *Effect of constraints on the dynamics of macromolecules.* Macromolecules (1982) 15(6):1528–1544. doi:10.1021/ma00234a015. 17

[72] Schlick, T., Barth, E., and Mandziuk, M. *Biomolecular dynamics at long timesteps: bridging the timescale gap between simulation and experimentation.* Annu. Rev. Biophys. Biomol. Struct. (1997) 26:181–222. doi:10.1146/annurev.biophys.26.1.181. 17

[73] van Gunsteren, W. and Berendsen, H. *Algorithms for macromolecular dynamics and constraint dynamics.* Mol. Phys. (1977) 34(5):1311–1327. doi:10.1080/00268977700102571. 17

[74] Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.* J. Comput. Phys. (1977) 23(3):327–341. doi:10.1016/0021-9991(77)90098-5. 17, 71

[75] Fisette, O., Lagüe, P., Gagné, S., and Morin, S. *Synergistic Applications of MD and NMR for the Study of Biological Systems.* J. Biomed. Biotechnol. (2012) 2012:1–12. doi:10.1155/2012/254208. 17

[76] Milo, R. and Phillips, R. Cell Biology by the Numbers. Garland Science (2015). ISBN 9781317230694. 17

[77] Palmer, R. *Broken ergodicity.* Adv. Phys. (1982) 31(6):669–735. doi:10.1080/00018738200101438. 18

[78] Marsili, S., Signorini, G. F., Chelli, R., Marchi, M., and Procacci, P. *ORAC: A molecular dynamics simulation program to explore free energy surfaces in biomolecular systems at the atomistic level.* J. Comput. Chem. (2009) NA–NA. doi:10.1002/jcc.21388. 18

[79] Dunker, A. K., Silman, I., Uversky, V. N., and Sussman, J. L. *Function and structure of inherently disordered proteins.* Curr. Opin. Struct. Biol. (2008) 18(6):756–764. doi:10.1016/j.sbi.2008.10.002. 18

[80] Smith, L. J., Daura, X., and van Gunsteren, W. F. *Assessing equilibration and convergence in biomolecular simulations.* Proteins (2002) 48(3):487–96. doi:10.1002/prot.10144. 19, 39, 54, 131, 134, 135

[81] Halkidi, M., Batistakis, Y., and Vazirgiannis, M. *On Clustering Validation Techniques.* J. Intell. Inf. Syst. (2001) 17(2/3):107–145. doi:10.1023/A:1012801612483. 19

[82] Lyman, E. and Zuckerman, D. M. *Ensemble-based convergence analysis of biomolecular trajectories.* Biophys. J. (2006) 91(1):164–72. doi:10.1529/biophysj.106.082941. 19, 39

[83] Romo, T. D. and Grossfield, A. *Block Covariance Overlap Method and Convergence in Molecular Dynamics Simulation.* J. Chem. Theory Comput. (2011) 7(8):2464–2472. doi:10.1021/ct2002754. 19

[84] Faraldo-Gómez, J. D., Forrest, L. R., Baaden, M., Bond, P. J., et al. *Conformational sampling and dynamics of membrane proteins from 10-nanosecond computer simulations.* Proteins Struct. Funct. Bioinforma. (2004) 57(4):783–791. doi:10.1002/prot.20257. 20

[85] Grossfield, A., Feller, S. E., and Pitman, M. C. *Convergence of molecular dynamics simulations of membrane proteins.* Proteins (2007) 67(1):31–40. doi:10.1002/prot.21308. 20

[86] Mu, Y., Nguyen, P. H., and Stock, G. *Energy landscape of a small peptide revealed by dihedral angle principal component analysis.* Proteins (2005) 58(1):45–52. doi:10.1002/prot.20310. 20

[87] Altis, A., Nguyen, P. H., Hegger, R., and Stock, G. *Dihedral angle principal component analysis of molecular dynamics simulations.* J. Chem. Phys. (2007) 126(24):244111. doi:10.1063/1.2746330. 20

[88] Caves, L. S. D., Evanseck, J. D., and Karplus, M. *Locally accessible conformations of proteins: Multiple molecular dynamics simulations of crambin.* Protein Sci. (1998) 7(3):649–666. doi:10.1002/pro.5560070314. 20

[89] Genheden, S. and Ryde, U. *A comparison of different initialization protocols to obtain statistically independent molecular dynamics simulations.* J. Comput. Chem. (2011) 32(2):187–95. doi:10.1002/jcc.21546. 20, 33

[90] Durrant, J. D. and McCammon, J. A. *Molecular dynamics simulations and drug discovery.* BMC Biol. (2011) 9:71. doi:10.1186/1741-7007-9-71. 21

[91] Shaw, D. E., Chao, J. C., Eastwood, M. P., Gagliardo, J., et al. *Anton, a special-purpose machine for molecular dynamics simulation.* ACM SIGARCH Comput. Archit. News (2007) 35(2):1. doi:10.1145/1273440.1250664. 22

[92] Lindorff-Larsen, K., Trbovic, N., Maragakis, P., Piana, S., and Shaw, D. E. *Structure and Dynamics of an Unfolded Protein Examined by Molecular Dynamics Simulation.* J. Am. Chem. Soc. (2012) 134(8):3787–3791. doi:10.1021/ja209931w. 22

[93] Piana, S., Lindorff-Larsen, K., and Shaw, D. E. *Protein folding kinetics and thermodynamics from atomistic simulation.* Proc. Natl. Acad. Sci. (2012) 109(44):17845–17850. doi:10.1073/pnas.1201811109. 22

[94] Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. *GROMACS: Fast, flexible, and free.* J. Comput. Chem. (2005) 26(16):1701–1718. doi:10.1002/jcc.20291. 22, 39, 64, 140, 145

[95] Yang, J., Wang, Y., and Chen, Y. *GPU accelerated molecular dynamics simulation of thermal conductivities.* J. Comput. Phys. (2007) 221(2):799–804. doi:10.1016/j.jcp.2006.06.039. 22

[96] Stone, J. E., Phillips, J. C., Freddolino, P. L., Hardy, D. J., Trabuco, L. G., and Schulten, K. *Accelerating molecular modeling applications with graphics processors.* J. Comput. Chem. (2007) 28(16):2618–40. doi:10.1002/jcc.20829. 22

[97] Liu, S.-Q., Liu, S.-X., and Fu, Y.-X. *Molecular motions of human HIV-1 gp120 envelope glycoproteins.* J. Mol. Model. (2008) 14(9):857–70. doi:10.1007/s00894-008-0327-7. 22, 31, 32

[98] Friedrichs, M. S., Eastman, P., Vaidyanathan, V., Houston, M., et al. *Accelerating molecular dynamic simulation on graphics processing units.* J. Comput. Chem. (2009) 30(6):864–872. doi:10.1002/jcc.21209. 22

[99] Götz, A. W., Williamson, M. J., Xu, D., Poole, D., Le Grand, S., and Walker, R. C. *Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born.* J. Chem. Theory Comput. (2012) 8(5):1542–1555. doi:10.1021/ct200909j. 22

[100] Daga, M., Aji, A. M., and Feng, W.-c. *On the Efficacy of a Fused CPU+GPU Processor (or APU) for Parallel Computing.* 2011 Symp. Appl. Accel. High-Performance Comput. IEEE. ISBN 978-1-4577-0635-6 (2011) 141–149. doi:10.1109/SAAHPC.2011.29. 22

[101] Carter, E., Ciccotti, G., Hynes, J. T., and Kapral, R. *Constrained reaction coordinate dynamics for the simulation of rare events.* Chem. Phys. Lett. (1989) 156(5):472–477. doi:10.1016/S0009-2614(89)87314-2. 22

[102] Grubmüller, H. *Predicting slow structural transitions in macromolecular systems: Conformational flooding.* Phys. Rev. E (1995) 52(3):2893–2906. doi:10.1103/PhysRevE.52.2893. 22

[103] Huber, T., Torda, A. E., and van Gunsteren, W. F. *Local elevation: a method for improving the searching properties of molecular dynamics simulation.* J. Comput. Aided. Mol. Des. (1994) 8(6):695–708. 22

[104] Rosso, L., Minárý, P., Zhu, Z., and Tuckerman, M. E. *On the use of the adiabatic molecular dynamics technique in the calculation of free energy profiles.* J. Chem. Phys. (2002) 116(11):4389–4402. doi:10.1063/1.1448491. 22

[105] Torrie, G. and Valleau, J. *Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling.* J. Comput. Phys. (1977) 23(2):187–199. doi:10.1016/0021-9991(77)90121-8. 22

[106] Laio, A. and Parrinello, M. *Escaping free-energy minima.* Proc. Natl. Acad. Sci. U. S. A. (2002) 99(20):12562–6. doi:10.1073/pnas.202427399. 22

[107] Sugita, Y. and Okamoto, Y. *Replica-exchange molecular dynamics method for protein folding.* Chem. Phys. Lett. (1999) 314(1-2):141–151. doi:10.1016/S0009-2614(99)01123-9. 22, 28

[108] Lyubartsev, A. P., Martsinovski, A. A., Shevkunov, S. V., and VorontsovVelyaminov, P. N. *New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles.* J. Chem. Phys. (1992) 96(3):1776–1783. doi:10.1063/1.462133. 22

[109] Marinari, E. and Parisi, G. *Simulated Tempering: A New Monte Carlo Scheme.* Europhys. Lett. (1992) 19(6):451–458. doi:10.1209/0295-5075/19/6/002. 22

[110] Gao, Y. Q. *An integrate-over-temperature approach for enhanced sampling.* J. Chem. Phys. (2008) 128(6):064105. doi:10.1063/1.2825614. 23

[111] Hamelberg, D., Mongan, J., and McCammon, J. A. *Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules.* J. Chem. Phys. (2004) 120(24):11919–29. doi:10.1063/1.1755656. 23, 24, 25, 50, 135

[112] Sinko, W., de Oliveira, C. A. F., Pierce, L. C. T., and McCammon, J. A. *Protecting High Energy Barriers: A New Equation to Regulate Boost Energy in Accelerated Molecular Dynamics Simulations.* J. Chem. Theory Comput. (2012) 8(1):17–23. doi:10.1021/ct200615k. 23, 83, 135

[113] Sinko, W., Miao, Y., de Oliveira, C. A. F., and McCammon, J. A. *Population Based Reweighting of Scaled Molecular Dynamics.* J. Phys. Chem. B (2013) 117(42):12759–12768. doi:10.1021/jp401587e. 23, 26, 47, 51, 52, 123, 135

[114] Shen, T. and Hamelberg, D. *A statistical analysis of the precision of reweighting-based simulations.* J. Chem. Phys. (2008) 129(3):034103. doi:10.1063/1.2944250. 23, 25, 118, 122, 135

[115] Miao, Y., Sinko, W., Pierce, L., Bucher, D., Walker, R. C., and McCammon, J. A. *Improved Reweighting of Accelerated Molecular Dynamics Simulations for Free Energy Calculation.* J. Chem. Theory Comput. (2014) 10(7):2677–2689. doi:10.1021/ct500090q. 23, 25, 50, 52, 83, 118, 122, 135

[116] Hamelberg, D., de Oliveira, C. A. F., and McCammon, J. A. *Sampling of slow diffusive conformational transitions with accelerated molecular dynamics.* J. Chem. Phys. (2007) 127(15):155102. doi:10.1063/1.2789432. 23, 24, 83

[117] Pierce, L. C., Salomon-Ferrer, R., Augusto F. de Oliveira, C., McCammon, J. A., and Walker, R. C. *Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics.* J. Chem. Theory Comput. (2012) 8(9):2997–3002. doi:10.1021/ct300284c. 23, 24, 25, 47, 50, 51, 83

[118] Miao, Y., Feixas, F., Eun, C., and McCammon, J. A. *Accelerated molecular dynamics simulations of protein folding.* J. Comput. Chem. (2015) 36(20):1536–1549. doi:10.1002/jcc.23964. 24, 26, 83, 122, 135

[119] Markwick, P. R. L. and McCammon, J. A. *Studying functional dynamics in bio-molecules using accelerated molecular dynamics.* Phys. Chem. Chem. Phys. (2011) 13(45):20053. doi:10.1039/c1cp22100k. 25, 118, 122, 135

[120] Hummer, G. *Fast-growth thermodynamic integration: Error and efficiency analysis.* J. Chem. Phys. (2001) 114(17):7330–7337. doi:10.1063/1.1363668. 25

[121] Eastwood, M. P., Hardin, C., Luthey-Schulten, Z., and Wolynes, P. G. *Statistical mechanical refinement of protein structure prediction schemes: Cumulant expansion approach.* J. Chem. Phys. (2002) 117(9):4602–4615. doi:10.1063/1.1494417. 25

[122] Jing, Z. and Sun, H. *A Comment on the Reweighting Method for Accelerated Molecular Dynamics Simulations.* J. Chem. Theory Comput. (2015) 11(6):2395–2397. doi:10.1021/acs.jctc.5b00236. 26, 83, 118, 122, 135

[123] Montcalm, T., Cui, W., Zhao, H., Guarnieri, F., and Wilson, S. R. *Simulated annealing of met-enkephalin: low energy states and their relevance to membrane-bound, solution and solid-state conformations.* J. Mol. Struct. Theochem (1994) 308:37–51. doi:10.1016/0166-1280(94)80093-6. 27

[124] Marcotte, I., Separovic, F., Auger, M., and Gagné, S. M. *A Multidimensional 1H NMR Investigation of the Conformation of Methionine-Enkephalin in Fast-Tumbling Bicelles.* Biophys. J. (2004) 86(3):1587–1600. doi:10.1016/S0006-3495(04)74226-5. 27, 68

[125] Hughes, J., Smith, T. W., Kosterlitz, H. W., Fothergill, L. A., Morgan, B. A., and Morris, H. R. *Identification of two related pentapeptides from the brain with potent opiate agonist activity.* Nature (1975) 258(5536):577–80. 27, 28, 133

[126] Schwyzer, R. *Molecular mechanism of opioid receptor selection.* Biochemistry (1986) 25(20):6335–42. 27

[127] Graham, W. H., Carter, E. S., and Hicks, R. P. *Conformational analysis of met-enkephalin in both aqueous solution and in the presence of sodium dodecyl sulfate micelles using multidimensional NMR and molecular modeling.* Biopolymers (1992) 32(12):1755–1764. doi:10.1002/bip.360321216. 28, 133

[128] Li, Z. and Scheraga, H. A. *Structure and free energy of complex thermodynamic systems.* J. Mol. Struct. Theochem (1988) 179(1):333–352. doi:10.1016/0166-1280(88)80133-7. 28

[129] Kostov, K. S. and Freed, K. F. *Long-Time Dynamics of Met-Enkephalin: Comparison of Theory with Brownian Dynamics Simulations.* Biophys. J. (1999) 76(1):149–163. doi:10.1016/S0006-3495(99)77185-7. 28

[130] Zaman, M. H., Shen, M.-Y., Berry, R. S., and Freed, K. F. *Computer Simulation of Met-Enkephalin Using Explicit Atom and United Atom Potentials: Similarities, Differences, and Suggestions for Improvement.* J. Phys. Chem. B (2003) 107(7):1685–1691. doi:10.1021/jp026994s. 28

[131] Berg, B. A. *Metropolis importance sampling for rugged dynamical variables.* Phys. Rev. Lett. (2003) 90(18):180601. doi:10.1103/PhysRevLett.90.180601. 28

[132] Frickenhaus, S., Kannan, S., and Zacharias, M. *Efficient evaluation of sampling quality of molecular dynamics simulations by clustering of dihedral torsion angles and Sammon mapping.* J. Comput. Chem. (2009) 30(3):479–492. doi:10.1002/jcc.21076. 28

[133] Malevanets, A. and Wodak, S. J. *Multiple Replica Repulsion Technique for Efficient Conformational Sampling of Biological Systems.* Biophys. J. (2011) 101(4):951–960. doi:10.1016/j.bpj.2011.06.043. 28

[134] Stanfield, R., Cabezas, E., Satterthwait, A., Stura, E., Profy, A., and Wilson, I. *Dual conformations for the HIV-1 gp120 V3 loop in complexes with different neutralizing fabs.* Structure (1999) 7(2):131–42. doi:10.1016/S0969-2126(99)80020-3show. 28, 31, 32

[135] Korber, B. T., Farber, R. M., Wolpert, D. H., and Lapedes, A. S. *Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis.* Proc. Natl. Acad. Sci. U. S. A. (1993) 90(15):7176–80. 28

[136] Splettstoesser, T. *Diagram of HIV virion* (2014). HIV Wikipedia, enlarged font size, www.scistyle.com, license creative commons CC BY-SA 3.0. 29

[137] Splettstoesser, T. *Schematic description of the HIV replication cycle* (2014). HIV Wikipedia, enlarged font size and added step numbers, www.scistyle.com, license creative commons CC BY-SA 4.0. 29

[138] Barré-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., et al. *Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS).* Science (1983) 220(4599):868–71. 28

[139] Gallo, R. C., Salahuddin, S. Z., Popovic, M., Shearer, G. M., et al. *Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS.* Science (1984) 224(4648):500–3. 28

[140] Sharp, P. M. and Hahn, B. H. *Origins of HIV and the AIDS Pandemic.* Cold Spring Harb. Perspect. Med. (2011) 1(1):a006841–a006841. doi:10.1101/cshperspect.a006841. 28

[141] UNAIDS. *World AIDS Day Report 2015: On the fast-track to end AIDS by 2030: Focus on Location and Population* (2016). 28

[142] Briggs, J. A. *Structural organization of authentic, mature HIV-1 virions and cores.* EMBO J. (2003) 22(7):1707–1715. doi:10.1093/emboj/cdg143. 28, 30, 134

[143] Freed, E. O. *HIV-1 assembly, release and maturation.* Nat. Rev. Microbiol. (2015) 13(8):484–496. doi:10.1038/nrmicro3490. 28, 30, 134

[144] Wyatt, R. *The HIV-1 Envelope Glycoproteins: Fusogens, Antigens, and Immunogens.* Science (80-.). (1998) 280(5371):1884–1888. doi:10.1126/science.280.5371.1884. 30

[145] Chan, D. C., Fass, D., Berger, J. M., and Kim, P. S. *Core Structure of gp41 from the HIV Envelope Glycoprotein.* Cell (1997) 89(2):263–273. doi:10.1016/S0092-8674(00)80205-6. 30

[146] Zhu, P., Liu, J., Bess, J., Chertova, E., et al. *Distribution and three-dimensional structure of AIDS virus envelope spikes.* Nature (2006) 441(7095):847–52. doi:10.1038/nature04817. 30

[147] Liu, J., Bartesaghi, A., Borgnia, M. J., Sapiro, G., and Subramaniam, S. *Molecular architecture of native HIV-1 gp120 trimers.* Nature (2008) 455(7209):109–113. doi:10.1038/nature07159. 30

[148] Zheng, Y.-H., Lovsin, N., and Peterlin, B. M. *Newly identified host factors modulate HIV replication.* Immunol. Lett. (2005) 97(2):225–234. doi:10.1016/j.imlet.2004.11.026. 30

[149] Pollard, V. W. and Malim, M. H. *THE HIV-1 REV PROTEIN.* Annu. Rev. Microbiol. (1998) 52(1):491–532. doi:10.1146/annurev.micro.52.1.491. 30

[150] Butsch, M. and Boris-Lawrie, K. *Destiny of Unspliced Retroviral RNA: Ribosome and/or Virion?* J. Virol. (2002) 76(7):3089–3094. doi:10.1128/JVI.76.7.3089-3094.2002. 30

[151] Wilen, C. B., Tilton, J. C., and Doms, R. W. *HIV: Cell Binding and Entry.* Cold Spring Harb. Perspect. Med. (2012) 2(8):a006866–a006866. doi:10.1101/cshperspect.a006866. 31

[152] Berger, E. A., Murphy, P. M., and Farber, J. M. *Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease.* Annu. Rev. Immunol. (1999) 17:657–700. doi:10.1146/annurev.immunol.17.1.657. 30, 31, 134

[153] Kwong, P. D., Wyatt, R., Robinson, J., Sweet, R. W., Sodroski, J., and Hendrickson, W. A. *Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody.* Nature (1998) 393(6686):648–59. doi:10.1038/31405. 31

[154] Chan, D. C. and Kim, P. S. *HIV entry and its inhibition.* Cell (1998) 93(5):681–4. 31

[155] Pierson, T. C. and Doms, R. W. *HIV-1 Entry and Its Inhibition.* Cell. Factors Involv. Early Steps Retroviral Replication, 1–27. Springer Berlin Heidelberg (2003) doi:10.1007/978-3-642-19012-4_1. 31

[156] Sattentau, Q. J. *Conformational changes induced in the human immunodeficiency virus envelope glycoprotein by soluble CD4 binding.* J. Exp. Med. (1991) 174(2):407–415. doi:10.1084/jem.174.2.407. 31

[157] Thali, M., Moore, J. P., Furman, C., Charles, M., et al. *Characterization of conserved human immunodeficiency virus type 1 gp120 neutralization epitopes exposed upon gp120-CD4 binding.* J. Virol. (1993) 67(7):3978–88. 31

[158] Sattentau, Q. J. *HIV gp120: double lock strategy foils host defences.* Structure (1998) 6(8):945–949. doi:10.1016/S0969-2126(98)00096-3. 31

[159] Wu, L., Gerard, N. P., Wyatt, R., Choe, H., et al. *CD4-induced interaction of primary HIV-1 gp120 glycoproteins with the chemokine receptor CCR-5.* Nature (1996) 384(6605):179–183. doi:10.1038/384179a0. 31

[160] Wyatt, R., Moore, J., Accola, M., Desjardin, E., Robinson, J., and Sodroski, J. *Involvement of the V1/V2 variable loop structure in the exposure of human immunodeficiency virus type 1 gp120 epitopes induced by receptor binding.* J. Virol. (1995) 69(9):5723–33. 31

[161] Rizzuto, C. D. *A Conserved HIV gp120 Glycoprotein Structure Involved in Chemokine Receptor Binding.* Science (80-.). (1998) 280(5371):1949–1953. doi:10.1126/science.280.5371.1949. 31

[162] Dragic, T., Litwin, V., Allaway, G. P., Martin, S. R., et al. *HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5.* Nature (1996) 381(6584):667–73. doi:10.1038/381667a0. 31

[163] Deng, H., Liu, R., Ellmeier, W., Choe, S., et al. *Identification of a major co-receptor for primary isolates of HIV-1.* Nature (1996) 381(6584):661–6. doi:10.1038/381661a0. 31

[164] Wu, B., Chien, E. Y. T., Mol, C. D., Fenalti, G., et al. *Structures of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists.* Science (80-.). (2010) 330(6007):1066–1071. doi:10.1126/science.1194396. 31

[165] Tamamis, P. and Floudas, C. A. *Molecular Recognition of CXCR4 by a Dual Tropic HIV-1 gp120 V3 Loop.* Biophys. J. (2013) 105(6):1502–1514. doi:10.1016/j.bpj.2013.07.049. 31

[166] Brelot, A., Heveker, N., Adema, K., Hosie, M. J., Willett, B., and Alizon, M. *Effect of mutations in the second extracellular loop of CXCR4 on its utilization by human and feline immunodeficiency viruses.* J. Virol. (1999) 73(4):2576–86. 31

[167] Farzan, M., Mirzabekov, T., Kolchinsky, P., Wyatt, R., et al. *Tyrosine sulfation of the amino terminus of CCR5 facilitates HIV-1 entry.* Cell (1999) 96(5):667–76. 31

[168] Hughes, A. and Nelson, M. *HIV entry: new insights and implications for patient management.* Curr. Opin. Infect. Dis. (2009) 22(1):35–42. doi:10.1097/QCO.0b013e3283213093. 31

[169] Berger, E. A., Doms, R. W., Fenyö, E. M., Korber, B. T., et al. *A new classification for HIV-1.* Nature (1998) 391(6664):240. doi:10.1038/34571. 31

[170] Tugarinov, V., Zvi, A., Levy, R., Hayek, Y., Matsushita, S., and Anglister, J. *NMR structure of an anti-gp120 antibody complex with a V3 peptide reveals a surface important for co-receptor binding.* Structure (2000) 8(4):385–395. doi:10.1016/S0969-2126(00)00119-2. 31

[171] Tian, S., Choi, W.-T., Liu, D., Pesavento, J., et al. *Distinct Functional Sites for Human Immunodeficiency Virus Type 1 and Stromal Cell-Derived Factor 1 on CXCR4 Transmembrane Helical Domains.* J. Virol. (2005) 79(20):12667–12673. doi:10.1128/JVI.79.20.12667-12673.2005. 31

[172] Napier, K. B., Wang, Z.-x., Peiper, S. C., and Trent, J. O. *CCR5 interactions with the variable 3 loop of gp120.* J. Mol. Model. (2006) 13(1):29–41. doi:10.1007/s00894-006-0117-z. 31

[173] López de Victoria, A., Tamamis, P., Kieslich, C. A., and Morikis, D. *Insights into the Structure, Correlated Motions, and Electrostatic Properties of Two HIV-1 gp120 V3 Loops.* PLoS One (2012) 7(11):e49925. doi:10.1371/journal.pone.0049925. 31, 32

[174] Wood, N. T., Fadda, E., Davis, R., Grant, O. C., et al. *The Influence of N-Linked Glycans on the Molecular Dynamics of the HIV-1 gp120 V3 Loop.* PLoS One (2013) 8(11):e80301. doi:10.1371/journal.pone.0080301. 31

[175] Pollakis, G., Kang, S., Kliphuis, A., Chalaby, M. I., Goudsmit, J., and Paxton, W. A. *N-linked glycosylation of the HIV type-1 gp120 envelope glycoprotein as a major determinant of CCR5 and CXCR4 coreceptor utilization.* J. Biol. Chem. (2001) 276(16):13433–41. doi:10.1074/jbc.M009779200. 31

[176] Resch, W., Hoffman, N., and Swanstrom, R. *Improved Success of Phenotype Prediction of the Human Immunodeficiency Virus Type 1 from Envelope Variable Loop 3 Sequence Using Neural Networks.* Virology (2001) 288(1):51–62. doi:10.1006/viro.2001.1087. 31

[177] Jensen, M. A., Coetzer, M., van 't Wout, A. B., Morris, L., and Mullins, J. I. *A Reliable Phenotype Predictor for Human Immunodeficiency Virus Type 1 Subtype C Based on Envelope V3 Sequences.* J. Virol. (2006) 80(10):4698–4704. doi:10.1128/JVI.80.10.4698-4704.2006. 31

[178] Pfeifer, N. and Lengauer, T. *Improving HIV coreceptor usage prediction in the clinic using hints from next-generation sequencing data.* Bioinformatics (2012) 28(18):i589–i595. doi:10.1093/bioinformatics/bts373. 31

[179] Sander, O., Sing, T., Sommer, I., Low, A. J., et al. *Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage.* PLoS Comput. Biol. (2007) 3(3):e58. doi:10.1371/journal.pcbi.0030058. 32

[180] Masso, M. and Vaisman, I. I. *Accurate and efficient gp120 V3 loop structure based models for the determination of HIV-1 co-receptor usage.* BMC Bioinformatics (2010) 11(1):494. doi:10.1186/1471-2105-11-494. 32

[181] Dybowski, J. N., Heider, D., and Hoffmann, D. *Prediction of Co-Receptor Usage of HIV-1 from Genotype.* PLoS Comput. Biol. (2010) 6(4):e1000743. doi:10.1371/journal.pcbi.1000743. 32

[182] Vranken, W. F., Budesinsky, M., Fant, F., Boulez, K., and Borremans, F. A. *The complete Consensus V3 loop peptide of the envelope protein gp120 of HIV-1 shows pronounced helical character in solution.* FEBS Lett. (1995) 374(1):117–21. 32, 69

[183] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., et al. *The Protein Data Bank.* Nucleic Acids Res. (2000) 28(1):235–42. 32, 69

[184] Hsu, S.-T. D. and Bonvin, A. M. J. J. *Atomic insight into the CD4 binding-induced conformational changes in HIV-1 gp120.* Proteins (2004) 55(3):582–93. doi:10.1002/prot.20061. 32

[185] Sali, A. and Blundell, T. L. *Comparative protein modelling by satisfaction of spatial restraints.* J. Mol. Biol. (1993) 234(3):779–815. doi:10.1006/jmbi.1993.1626. 33, 35

[186] Marti-Renom, M. A. *Alignment of protein sequences by their profiles.* Protein Sci. (2004) 13(4):1071–1087. doi:10.1110/ps.03379804. 33, 35, 70

[187] Eswar, N., Eramian, D., Webb, B., Shen, M.-Y., and Sali, A. *Protein structure modeling with MODELLER.* Methods Mol. Biol. (2008) 426:145–59. doi: 10.1007/978-1-60327-058-8_8. 34, 35, 68, 69, 70

[188] Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E., and Thornton, J. M. *Knowledge-based prediction of protein structures and the design of novel molecules.* Nature (1987) 326(6111):347–352. doi:10.1038/326347a0. 33

[189] Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F., and Sali, A. *Comparative protein structure modeling of genes and genomes.* Annu. Rev. Biophys. Biomol. Struct. (2000) 29:291–325. doi:10.1146/annurev.biophys.29.1.291. 33

[190] Zvelebil, M. and Baum, J. Understanding Bioinformatics. Taylor & Francis Group (2007). ISBN 9781136976964. Chapter 13. 33, 34

[191] Chothia, C. and Lesk, A. M. *The relation between the divergence of sequence and structure in proteins.* EMBO J. (1986) 5(4):823–6. 33

[192] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., et al. *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res. (1997) 25(17):3389–402. 33, 69

[193] Fiser, A., Do, R. K., and Sali, A. *Modeling of loops in protein structures.* Protein Sci. (2000) 9(9):1753–73. doi:10.1110/ps.9.9.1753. 34

[194] Shen, M.-y. and Sali, A. *Statistical potential for assessment and prediction of protein structures.* Protein Sci. (2006) 15(11):2507–2524. doi:10.1110/ps.062416606. 34

[195] Melo, F., Sánchez, R., and Sali, A. *Statistical potentials for fold assessment.* Protein Sci. (2002) 11(2):430–48. doi:10.1002/pro.110430. 34

[196] John, B. and Sali, A. *Comparative protein structure modeling by iterative alignment, model building and model assessment.* Nucleic Acids Res. (2003) 31(14):3982–92. 34

[197] Kabsch, W. *A solution for the best rotation to relate two sets of vectors.* Acta Crystallogr. Sect. A (1976) 32(5):922–923. doi:10.1107/S0567739476001873. 39

[198] Snyder, D. A. and Montelione, G. T. *Clustering algorithms for identifying core atom sets and for assessing the precision of protein structure ensembles.* Proteins (2005) 59(4):673–86. doi:10.1002/prot.20402. 39, 140

[199] Murtagh, F. Multidimensional clustering algorithms. Physika Verlag (1985). 54, 58, 62

[200] R Development Core Team. *R: A Language and Environment for Statistical Computing* (2008). Http://www.r-project.org. 54, 58, 62

[201] Kaufman, L. and Rousseeuw, P. J. (editors). Finding Groups in Data. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA (1990). ISBN 9780470316801. doi:10.1002/9780470316801. 54, 58, 62

[202] Hennig, C. *fpc: Flexible procedures for clustering* (2014). R package version 2.1-7, https://cran.r-project.org/package=fpc. 54, 58, 62

[203] Continuum-Analytics. *Anaconda Software Distribution* (2016). Version 2.4.1. 64

[204] Hunter, J. D. *Matplotlib: A 2D Graphics Environment.* Comput. Sci. Eng. (2007) 9(3):90–95. doi:10.1109/MCSE.2007.55. 64, 154

[205] Jones, E., Oliphant, T., Peterson, P., and Others. {*SciPy*}*: Open source scientific tools for* {*Python*} (2001). 64

[206] van der Walt, S., Colbert, S. C., and Varoquaux, G. *The NumPy Array: A Structure for Efficient Numerical Computation.* Comput. Sci. Eng. (2011) 13(2):22–30. doi:10.1109/MCSE.2011.37. 64

[207] Pérez, F. and Granger, B. E. {*IP*}*ython: a System for Interactive Scientific Computing.* Comput. Sci. Eng. (2007) 9(3):21–29. doi:10.1109/MCSE.2007.53. 64

[208] Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., et al. *Jupyter Notebooks - a publishing format for reproducible computational workflows* (2016). V4.0.6. 64

[209] Delano, W. L. *The PyMOL Molecular Graphics System* (2002). Version 1.7.2.1, http://www.pymol.org. 68

[210] Kier, B. L., Shu, I., Eidenschink, L. A., and Andersen, N. H. *Stabilizing capping motif for -hairpins and sheets.* Proc. Natl. Acad. Sci. (2010) 107(23):10466–10471. doi:10.1073/pnas.0913534107. 68

[211] Zuo, Z.-L., Guo, L., and Mancera, R. L. *Free Energy of Binding of Coiled-Coil Complexes with Different Electrostatic Environments: The Influence of Force Field Polarisation and Capping.* Nat. Products Bioprospect. (2014) 4(5):285–295. doi:10.1007/s13659-014-0036-0. 68

[212] Nickolls, J., Buck, I., Garland, M., and Skadron, K. *Scalable parallel programming with CUDA.* Queue (2008) 6(2):40. doi:10.1145/1365490.1365500. 71

[213] Wallnoefer, H. G., Liedl, K. R., and Fox, T. *A challenging system: Free energy prediction for factor Xa.* J. Comput. Chem. (2011) 32(8):1743–1752. doi:10.1002/jcc.21758. 71

[214] Sethna, J. Statistical Mechanics: Entropy, Order Parameters and Complexity. Oxford Master Series in Physics. OUP Oxford (2006). ISBN 9780198566779. 82

[215] Miao, Y., Feher, V. A., and McCammon, J. A. *Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation.* J. Chem. Theory Comput. (2015) 11(8):3584–3595. doi:10.1021/acs.jctc.5b00436. 83, 122, 135

[216] Crippen, G. M. and Ohkubo, Y. Z. *Statistical mechanics of protein folding by exhaustive enumeration.* Proteins (1998) 32(4):425–37. 140

List of Publications

Journal Articles

M. Nemec, D. Hoffmann. Quantitative Assessment of Molecular Dynamics Sampling for Flexible Systems. *Journal of Chemical Theory and Computation*, 13: 400-414, 2017. doi:10.1021/acs.jctc.6b00823.

M. Nemec, Gregor R. Foltin, Kai P. Schmidt. Microscopic mechanism for the 1/8 magnetization plateau in $SrCu_2(BO_3)_2$. *Physical Review B*, 86: 174425, 2012. doi:10.1103/PhysRevB.86.174425

Conference poster

M. Nemec, D. Hoffmann. 2016. Frühjahrstagung DPG. Regensburg, Germany. *Quantitative assessment of sampling quality of molecular dynamics simulations of biomolecular systems.*

Conference talk

M. Nemec, Gregor R. Foltin, Kai P. Schmidt. 2012. Frühjahrstagung DPG. Berlin, Germany. *Microscopic mechanism for the 1/8 magnetization plateau in* $SrCu_2(BO_3)_2$.

Acknowledgements

First of all, I need and want to thank my beloved friends, who helped me through a lot of trouble, especially in the last two years. **Anita Böhm** gave and gives me her full support regardless on the kind of problem. Thank you **Gregor Foltin** giving me always a smile, funny conversations and great video game evenings. I also greatly appreciate the colorful and deep conversations, "mad laughs" and all the help from **Ludwig Ohl**. Furthermore, I want to greatly thank **Izabela and Tereza Hejlová** for an incredible friendship and reminding me that I can count on you although we did not meet in the last ten years. Thank you. The same is true for **Kuba Urban** and both families. Finally I thank **Eva Zapletalová**. I met you recently, but it already feels like we would be friends for years. Thank you for the invitations for beer, hours of deep conversations and a lots of laughs.

I thank my competent and supporting supervisor **Daniel Hoffmann** for giving me the opportunity to work in a biophysical group and successfully enhance my competences and scientific point of view. He could encourage me in difficult times, pushing me hard to being successful and always stood behind me whenever necessary. I learned a lot, also about myself and do not want to miss these years.

Special thanks shall also be granted to the whole group of **Bioinformatics and Computational Biophysics**. I enjoyed many helpful conversations and discussions with **Karsten Sewczyk**, **Olli Kuhn**, **Ludwig Ohl**, and **Anja Lange**. The latter two were always a friendly help for any administrative problem. Thanks also to **Claudia Wilmes** for a lot of organization help.

Finally I want to thank all the people around me, who are friendly and supportive but I did not mention their names.

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.

Declarations

Erklärung:
Hiermit erkläre ich, gem. §6 Abs. (2) f) der Promotionsordnung der Fakultäten für Biologie, Chemie und Mathematik der Universität Duisburg-Essen vom 04. Februar 2010 zur Erlangung des Dr. rer. nat., dass ich das Arbeitsgebiet, dem das Thema "A toolkit to quantify the sampling quality of molecular dynamics trajectories: Studying highly flexible biomolecules" zuzuordnen ist, in Forschung und Lehre vertrete und den Antrag von Mike Nemec befürworte und die Betreuung auch im Falle eines Weggangs, wenn nicht wichtige Gründe dem entgegenstehen, weiterführen werde.

Essen, den _____ _____

Unterschrift eines Mitglieds der Universität Duisburg-Essen

Erklärung:
Hiermit erkläre ich, gem. §7 Abs. (2) c)+e) der Promotionsordnung der Fakultäten für Biologie, Chemie und Mathematik der Universität Duisburg-Essen vom 04. Februar 2010 zur Erlangung des Dr. rer. nat., dass ich die vorliegende Dissertation selbständig verfasst und mich keiner anderen als der angegebenen Hilfsmittel bedient habe.

Essen, den _____ _____

Unterschrift des Doktoranden

Erklärung:
Hiermit erkläre ich, gem. §7 Abs. (2) d)+f) der Promotionsordnung der Fakultäten für Biologie, Chemie und Mathematik der Universität Duisburg-Essen vom 04. Februar 2010 zur Erlangung des Dr. rer. nat., dass ich keine anderen Promotionen bzw. Promotionsversuche in der Vergangenheit durchgeführt habe und dass diese Arbeit von keiner anderen Fakultät/Fachbereich abgelehnt worden ist.

Essen, den _____ _____

Unterschrift des Doktoranden