



**university of  
 groningen**

**faculty of science  
 and engineering**

# A metadynamics study of simple peptide chains

**Research project of :**  
 Vito Federico Palmisano

**Supervisors :**  
 prof. Siewert-Jan Marrink  
 dr. Ilias Patmanidis

---

**University of Groningen**

**Faculty of Science and Engineering**

**Master in Theoretical Chemistry and Computational Modelling**

January 21, 2021

## Abstract

In nature, proteins possess the remarkable ability to fold spontaneously into precisely determined three-dimensional structures. Understanding of this process requires the characterization of the driving molecular interactions and intermediate states of the system. At a molecular level, the backbone dihedral angles play an important role for the secondary structure formation, providing crucial information about the three-dimensional structure. Protein conformational energy landscapes are complex, high-dimensional surfaces with many local minima and navigating them requires efficient sampling methods. Molecular dynamics simulations of protein folding can provide high-resolution data on the folding process, but intrinsic limitations of atomistic models render the task extremely difficult. This is where enhanced sampling methods become useful to accelerate the dynamics and sample the conformational space of such systems. In this study, an enhanced sampling method, metadynamics, is employed to describe the behaviour of the torsional angles  $\phi$  and  $\psi$  in different tripeptide chains. The torsional angles are also evaluated with different force fields. For all the force fields, local minima are commonly shared with small variations along the free energy profiles. The presence of side chains yields to different maxima in the free energy profiles.

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Protein Folding . . . . .	4
1.2	Computational Methods . . . . .	7
1.3	Molecular Dynamics . . . . .	7
1.4	Enhanced Sampling Methods and Well Tempered Metadynamics . . . . .	9
1.5	PLUMED . . . . .	11
<b>2</b>	<b>Computational Details</b>	<b>12</b>
2.1	Atomistic molecular dynamics simulation details . . . . .	12
2.1.1	Initial Structures . . . . .	12
2.1.2	Molecular dynamics simulation details . . . . .	13
2.1.3	Well Tempered Metadynamics simulation details . . . . .	13
<b>3</b>	<b>Results and Discussion</b>	<b>15</b>
3.1	Convergence of the Collective Variables and the Free Energy . . . . .	15
3.2	Comparison between Force Fields . . . . .	17
3.3	2D Free Energy Profiles and Block Analysis . . . . .	18
<b>4</b>	<b>Conclusions</b>	<b>21</b>
<b>5</b>	<b>Appendix</b>	<b>22</b>
5.1	Tripeptides . . . . .	22
5.2	Pentapeptides . . . . .	26

# 1 Introduction

## 1.1 Protein Folding

The function of biomolecules is determined by both their 3D structure and dynamics. Proteins, for example, are flexible systems which have a large range of dynamics with a timescale from the picoseconds to seconds. With ribosomes, proteins are synthesized in their primary structure and when released, fold to their native state (1). In 1969, the protein folding problem was built on the Levinthal paradox, which essentially states that if the protein folding process is a random search problem, all possible conformations can be equally probable and the solution can only be found by an unbiased random search (2). For such a surface, the time required for folding would be much larger than the actual time required for a protein to fold. The paradox was followed by the idea that a kinetic, nucleation event must occur to allow for a structure formation in biological feasible time (3). This theory was discarded because it was predicting the absence of folding intermediates, which was the central idea on the topic in the 1980s. With the work of R. L. Baldwin to experimentally define kinetic folding intermediates and pathways, the advances in protein folding turned towards a new view of heterogeneous folding establishing the current paradigm of multi-path funneled energy surface (4; 5). In a classical single pathway, the protein folds through distinct intermediates in a distinct pathway which is the only possibility. In a multi-path solution, the protein must fold energetically downhill (the Z axis) and shrink in conformational extent (the XY plane) as shown in Figure 1.

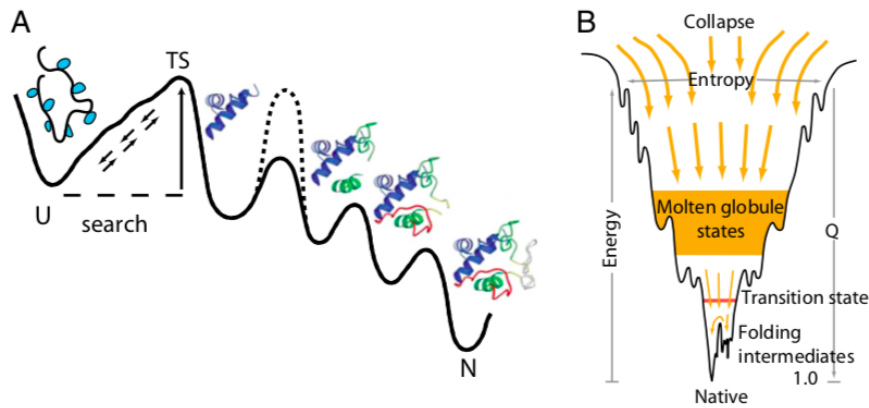


Figure 1: (A) The classical single pathway solution. (B) The multipath route constructed through a funneled landscape (6).

Nowadays, the understanding of protein function at the atomic level has been revolutionized by high resolution X-ray crystallography, but with this technique, the dynamical nature of proteins is not captured (7). Spectroscopic methods such as circular dichroism or infra-red spectroscopy can follow the kinetic folding in real time but are blind to the structure and possibilities of alternative mechanisms (6). Due to its dynamic nature, a detailed description on the intermediate structures in the folding funnel requires a multidimensional energy landscape that defines the relative

probabilities of the conformational states (thermodynamics) and the energy barriers between them (kinetics) (7). This is where a computational approach as atomistic molecular dynamics (MD), could in principle give meaningful insights on the folding mechanism of proteins and ultimately predict the structure of a protein from its sequence.

A protein is a polymer of amino acids with each amino acid residue linked to its neighbor by a specific type of covalent bond. All amino acids have a carbonyl group and an amino group bonded to the same carbon atom ( $C_\alpha$ ). Figure 2 gives a representation of a general protein structure.

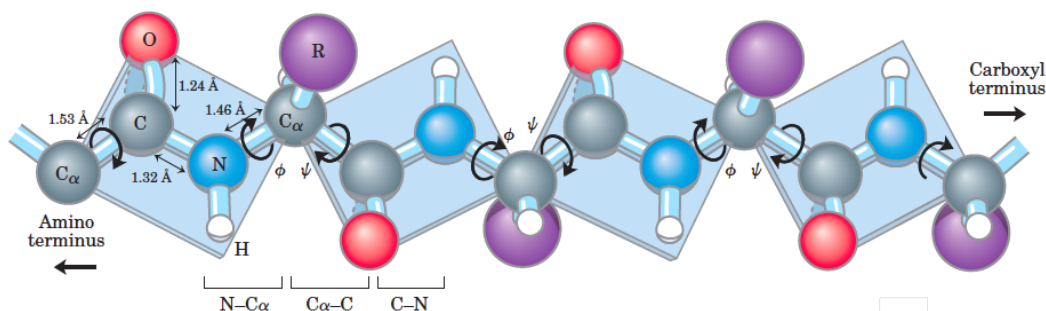


Figure 2: Schematic representation of a polypeptide structure. The dihedral angles of the backbone are represented with small curved arrows.

A side chain (R) is linked to the  $C_\alpha$ , which vary in structure, size and electric charge, giving unique properties for the amino acid. Because of the tetrahedral arrangement, the  $C_\alpha$  is a chiral center and the four different groups can occupy two unique spatial arrangements (L and D). The peptide bond is planar, rigid and shows a double bond character due to resonance between the nitrogen lone pair and the carboxyl oxygen.

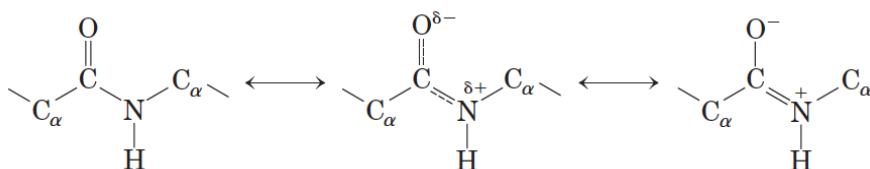


Figure 3: Resonance structure of a peptide bond.

For this reason, the omega ( $\omega$ ) torsional angle is always flat and fixed to almost  $180^\circ$ . Rotation is instead allowed in the  $N-C_\alpha$  and the  $C_\alpha-C$  bonds, therefore the backbone of a protein can be seen as a series of rigid planes with a point of rotation at  $C_\alpha$ . Those two torsional angles, phi ( $\phi$ ) and psi ( $\psi$ ) are free to rotate and therefore provide flexibility to the backbone of the protein. The  $\phi$  and  $\psi$  angles, also called Ramachandran angles (after the Indian physicist who worked on modeling the interactions in polypeptide chains), can vary from  $-180^\circ$  to  $180^\circ$ , with  $180^\circ$  defining the polypeptide in its most extended conformation and all the peptide groups are in the same plane. For the Ramachandran principle, the  $\alpha$ -helices,  $\beta$ -strands and turns are the most likely conformations that a polypeptide chain can adopt, since the other conformations are unfavorable

due to steric collisions between atoms (8). A Ramachandran plot for a polypeptide chain is shown in Figure 4.

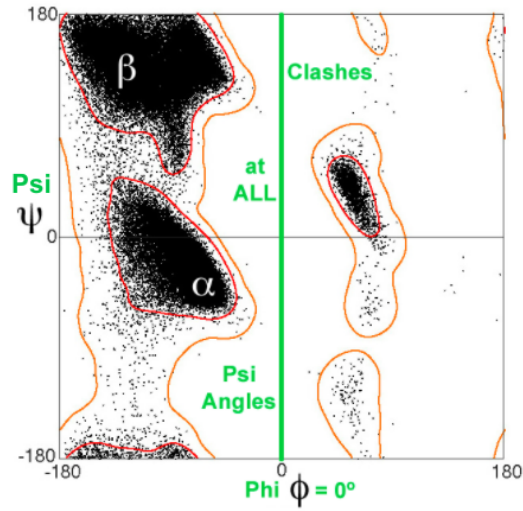


Figure 4: Example of a Ramachandran plot. The small black dots show the  $\phi$  and  $\psi$  angles for each amino acid in a polypeptide.

Each small black dot in the plot represent torsional angles  $\phi$  and  $\psi$  of an amino acid for a polypeptide chain. The cluster of dots named  $\alpha$  represents  $\alpha$ -helices secondary structures while the  $\beta$  cluster represents the  $\beta$ -strand. The green line at  $\phi=0$  shows that for any value of  $\psi$ , there are steric clashes and no amino acids in the chain can adopt those conformations. The  $\alpha$ -helix and  $\beta$ -strand are therefore the most common folding patterns of a polypeptide backbone. The reason is, in part, given by the optimal use of hydrogen bonds where in the  $\alpha$ -helix are formed between the hydrogen atom attached to the electronegative nitrogen atom of a peptide linkage and the electronegative carbonyl oxygen atom of the fourth amino acid on the amino-terminal side of that peptide bond while in the  $\beta$ -strand are formed between adjacent segments of the polypeptide chain.

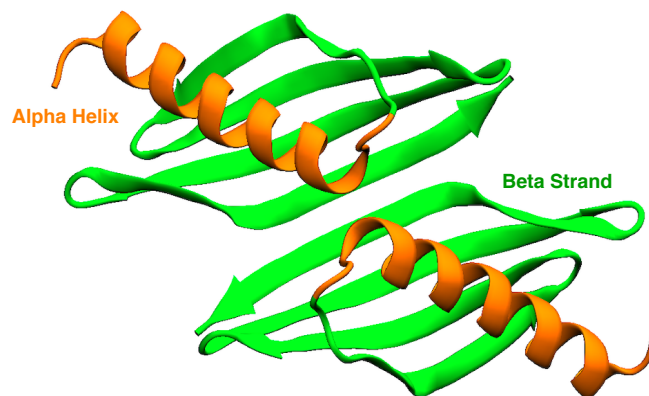


Figure 5: New cartoon representation of an  $\alpha$  helix and  $\beta$  sheet (PDB ID : 4R80)

Therefore the position of the side chains due to steric effects and the hydrogen bonds in the

backbone are the main driving forces for a secondary structure formation. In the following study, atomistic MD in combination with an enhanced sampling method, metadynamics, is employed to study the behaviour and compare the  $\phi$  and  $\psi$  dihedral angles of small tripeptide chains and compare three different force fields.

## 1.2 Computational Methods

### 1.3 Molecular Dynamics

Computational methods based on molecular models have been playing an essential role in biology and biophysics. The large growth of computer power and efficient algorithms have finally made possible to compare experimental data and theoretical predictions of biomolecular systems (9). Simulations can provide details concerning individual particle motion as a function of time and often used to address specific questions about the properties of a model system (10). The first MD simulation of a biomolecule, performed in the late 1970s, was less than 10 ps and with only 500 atoms (11). Current simulations are often  $\sim 1000$  times longer and contain systems of  $10^4$  to  $10^6$  atoms including membranes and explicit solvent.

In MD, an atomistic based description of a system is constructed and propagated deterministically or stochastically to generate a series of frames (trajectory) which describes the evolution of the particles over the simulation (12). Requirements to run a MD simulation are a model Hamiltonian which describes the thermodynamics of a system and a sampling algorithm able to sample relevant conformations from a chosen ensemble (13). In atomistic MD, a classical Hamiltonian that incorporates fixed point charges, is split into a kinetic and potential contribution,

$$H(p, r) = K(p) + V(r) \quad (1)$$

where  $p$  and  $r$  are the momentum and position of a particle, respectively, and  $K(p)$  and  $V(r)$  are the kinetic and potential term of the Hamiltonian  $H(p, r)$ . The kinetic term  $K(p)$  is written as

$$K(p) = \sum_{i=1}^N \frac{p_i^2}{2m_i} \quad (2)$$

where  $m_i$  is the mass of particle  $i$  and  $K(p)$  is independent of the position of the other particles. The potential energy function  $V(r)$  describes the energy landscape of the system with respect to atomic coordinates. Also defined as force field (FF), the energy function  $V(r)$  usually consists of a sum of parametrized terms which describe the energy of interaction of a given configuration (14). For example, the functional form of the gromos force field (15) can be written as,

$$V(r; s) = V^b(r; s) + V^{nb}(r; s) \quad (3)$$

where  $s$  represents the force field parameters,  $V(r; s)$  is the potential describing physical interactions split into bonded interactions and non-bonded pairwise interactions. The bonded and non-bonded

interactions can be further split,

$$V^b(r; s) = V^{bond}(r; s) + V^{angle}(r; s) + V^{har}(r; s) + V^{trig}(r; s) \quad (4)$$

$$V^{nb}(r; s) = V^{LJ}(r; s) + V^C(r; s) \quad (5)$$

with  $V^b(r; s)$  being the sum of bond, angle and dihedral angle interactions and  $V^{nb}(r; s)$  the sum of van der Walls and Coulomb interactions. With such model Hamiltonian, the Newtonian equation of motion can be solved where the net force exerted on the atom  $i$ ,  $F_i$ , is given by the negative gradient of the potential energy function  $V(r)$ ,

$$F_i = -\frac{dV}{dr_i} \quad (6)$$

and Newton's second law describing the acceleration is given as,

$$\frac{d^2r(t)}{dt^2} = \frac{F_i}{m} \quad (7)$$

Since there is no solution to the equation of motion for more than three interacting bodies, the dynamics is approximated numerically,

$$r(t + \Delta t) = r(t) + \frac{dr(t)}{dt} \Delta t + \frac{1}{2} \frac{d^2r(t)}{dt^2} \Delta t^2 + \dots \quad (8)$$

with a truncation of this Taylor series after the second term. The velocity-Verlet algorithm, for example, enables to find the trajectory of an object influenced by a force field by considering the expansion of coordinates forward and backward in time (16). It requires the atomic positions and accelerations at time  $t$  and the positions from the prior step,  $r(t - \Delta t)$ , to determine the new position at  $t + \Delta t$  (16).

$$r(t + \Delta t) = r(t) + \frac{dr(t)}{dt} \Delta t + \frac{1}{2} \frac{d^2r(t)}{dt^2} \Delta t^2 \quad (9)$$

$$r(t - \Delta t) = r(t) - \frac{dr(t)}{dt} \Delta t + \frac{1}{2} \frac{d^2r(t)}{dt^2} \Delta t^2 \quad (10)$$

$$r(t + \Delta t) = 2r(t) - r(t - \Delta t) + \frac{d^2r(t)}{dt^2} \Delta t^2 \quad (11)$$

where eq.(9) is a step forward in time and eq.(10) is a step backward in time. A popular algorithm based on the Verlet's equation is the leapfrog algorithm which uses the positions at time  $t$  and the velocities at time  $t - (\Delta t/2)$  to update both position and velocities via the calculated forces,  $F(t)$  as show in eq.(12) and (13).

$$r(t + \Delta t) = r(t) + \frac{dr(t)}{dt} \left( t + \frac{\Delta t}{2} \right) \Delta t \quad (12)$$

$$\frac{dr(t)}{dt} \left( t + \frac{\Delta t}{2} \right) = \frac{dr(t)}{dt} \left( t - \frac{\Delta t}{2} \right) + \frac{d^2r(t)}{dt^2} \Delta t \quad (13)$$



Additionally to a model Hamiltonian and an integrator, the system must estimate/predict the thermodynamic behaviour of a real system. Simulations occurs in a cubic (or other geometrical shapes) box in which periodic boundary conditions (PBC) are applied to remove hard edges and virtually replace the box in every direction, to give better estimation of bulk properties, as shown in Figure 6.

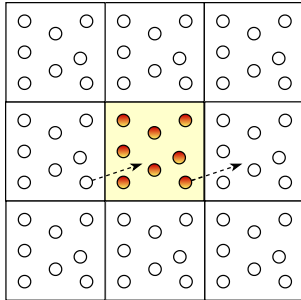


Figure 6: Periodic boundary conditions for a cubic box. The dashed arrow shows a particle exiting the box from the right side and entering the box from the opposite side.

The particle exiting the box from one side, enters the box from the opposite side. MD is conventionally performed under conditions of constant number of particles, volume and energy, i.e., microcanonical ensemble (NVE). However, experiments are usually performed at constant temperature and volume (NVT), or constant pressure and temperature (NPT), hence those conditions are simulated by the addition of external factors, thermostats and barostats (17).

#### 1.4 Enhanced Sampling Methods and Well Tempered Metadynamics

A system with a large number of degrees of freedom (DoF) may not have a single global minimal energy configuration but can only be described by a statistical mechanical ensemble of configurations in which the weight of each configuration  $r$  is given by the Boltzmann factor

$$P(r) \sim \exp(-V(r)/k_bT) = \exp(-\beta V(r)) \quad (14)$$

where  $P(r)$  is the probability of having a given configuration  $r$  with potential  $V(r)$ ,  $k_b$  is the Boltzmann constant and  $T$  is the temperature (9). The exponential factor implies that high energy systems are less relevant to the state of the system and only regions of high probability are visited. Those regions are usually separated by high energy barriers (much larger than  $k_bT$ ), so it is not likely that all important configurations are visited in a limited amount of time; this is known as the problem of quasi-nonergodicity (13; 18). The enhanced sampling methods have been developed to minimize the problem of quasi-nonergodicity. Those methods can be split in two categories depending on the prior knowledge on the regions which are not well sampled. If these important regions are not known, several approaches have been developed shown in Figure 7.

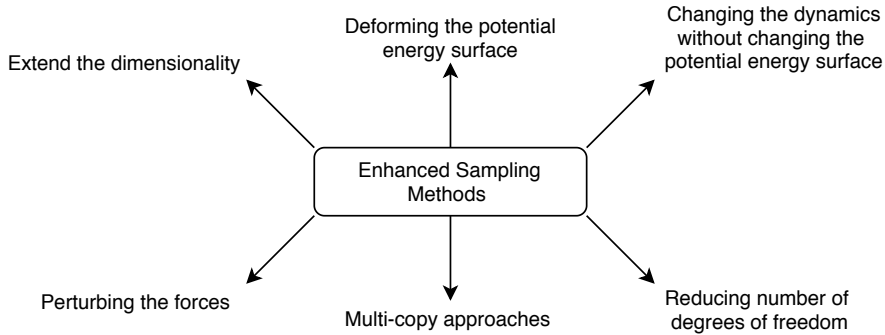


Figure 7: Different approaches to enhance the sampling of the phase space.

For the purpose of this study, only the *Deforming the potential energy surface* will be further explored. This category modifies the potential energy surface,  $V(r)$ , by adding a bias potential to the Hamiltonian, decreasing the energy barriers and increasing the sampling transitions (19). To effectively guide the simulations, some of these methods (20; 21) uses predefined reaction coordinates, also called collective variables (CVs). CVs are low dimensional functions of the atomistic coordinates  $r$  which should capture the slowest motion occurring in the reaction. In CVs based sampling, a bias potential  $V(s)$  is introduced in the system to jump over the energy barriers that separate different minima in that configurational space (19). A recently developed method is metadynamics which falls into the class of enhanced sampling methods which bias the potential that acts on a selected number of DoF previously called CVs (21). Precisely, in metadynamics, an external history dependent bias potential which is a function of the CVs is added to the Hamiltonian of the system (22). This potential is a sum of Gaussians which are deposited along the system's trajectory in the CVs space to discourage the already visited conformations. The bias at time  $t$  is written as an integral on the past trajectory  $r(t)$ ,

$$V_G(s, t) = \int_0^t dt' \omega \exp\left(-\sum_{i=1}^d \frac{s_i(r) - s_i(r(t'))^2}{2\sigma_i^2}\right) \quad (15)$$

where  $V(s, t)$  is the bias potential at time  $t$  in the space of the CV,  $\omega$  is the energy rate at which the bias grows,  $\sigma_i$  is the Gaussian width corresponding to the  $i$ th CV. The energy rate is expressed in terms of the Gaussian height  $W$  and a deposition stride  $\tau_G$  as,

$$\omega = \frac{W}{\tau_G} \quad (16)$$

Figure 8 shows an example of a one dimensional potential in which three local minima are present to understand the effect of  $V_G$  in time.

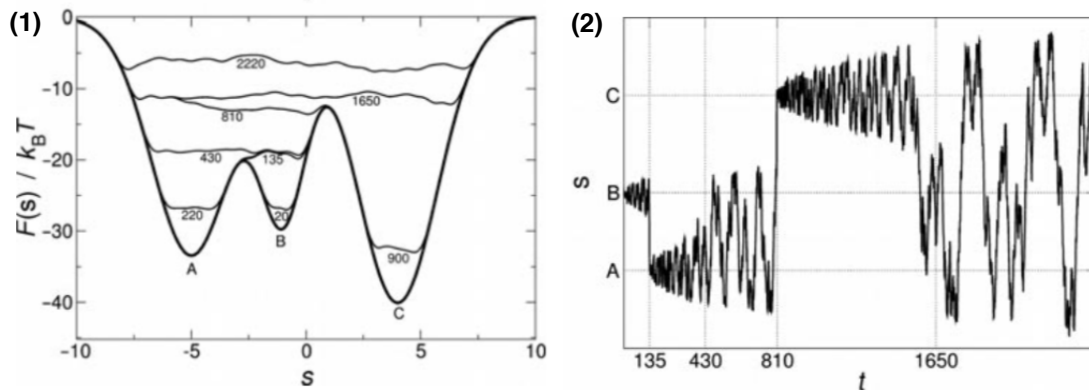


Figure 8: Example of a 1D metadynamics model potential. (1) Schematic representation of the progressive filling of the potential by the gaussians deposited along the trajectory. (2) Time evolution of the CVs during the simulation.

The simulation starts at B, and in a normal atomistic MD simulation, the system would remain in that minimum for a long time because the barriers are much larger than the thermal fluctuations,  $k_b T$ . Instead in this metadynamics simulation, Gaussian functions are deposited with time until at  $t = 135$  the system is pushed out of B into a new local minimum. Now the system is trapped in A until at  $t = 810$ , it can access also region C. With this method, the sampling of rare events can be accelerated by pushing the system away from local free energy minima, explore new reaction pathways as the system escapes from the minima and no *a priori* knowledge of the landscape is required. At the same time, the bias potential overfills the underlying FES and pushes the system towards high energy regions of the CV space. Consequently, it is not trivial to know when to stop the simulation. Also, as the system becomes more complex, it is also not trivial to choose an appropriate CV to bias.

## 1.5 PLUMED

PLUMED is a flexible complement MD code which can be interfaced with most of the MD softwares available (21). It is used both to analyze MD trajectories and also as a metadynamics code used for enhanced sampling methods. Figure 8 below shows the way PLUMED is integrated into a MD simulation.

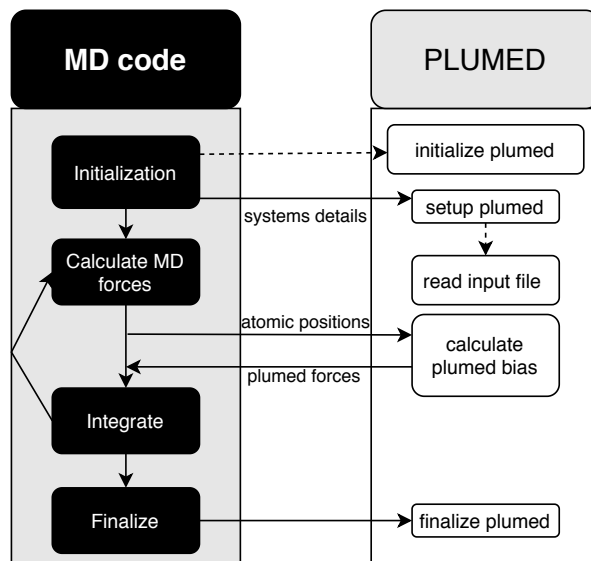


Figure 9: Schematic representation of the interface between plumed and an MD engine.

PLUMED is called during the initialization and its input file is read. It is then called after each run, after the forces that describe the interactions between the atoms are calculated. At this point, PLUMED allows the plugin to add the bias potential to the MD code so that they can be integrated with the equations of motion. Also, the selected post-process analysis are performed and printed in a file.

## 2 Computational Details

### 2.1 Atomistic molecular dynamics simulation details

All the modifications to the topology and trajectory files were performed with the MPI version of gmx tool of GROMACS, tleap and cpptraj of Ambergtools and PyTraj, a python library which exposes cpptraj's functions to the python environment (23; 24; 25). Visual Molecular Dynamics (VMD) was used for visual inspection of the trajectories and build visual representations and Python3.7 was used to produce plots (26).

#### 2.1.1 Initial Structures

The structures of alanine tripeptide (3-p) (AAA) and pentapeptide (5-p) (AAAAA), glycine 3-p (GGG) and 5-p (GGGGG), tryptophan 3-p (WWW) and 5-p (WWWWW), glutamic acid 3-p (EEE) and 5-p (EEEEEE) were built using tleap and Avogadro molecular editor (27). Those peptides were chosen to have a sample of each of the amino acid classes. All of the peptides were capped at the N-terminus with  $\text{NH}_2$  and at the C-terminus with  $\text{CH}_3$  to avoid artificial interaction between the termini. The structures of the 3-p and 5-p were arranged in a periodic cubic box of length 3nm and 5nm respectively and solvated with water molecules. For the EEE and EEEEE systems, the solution was neutralized with addition of 3 and 5  $\text{Na}^+$  ions. The boxes were composed

of approximately 2600 and 12000 for all the 3-p and 5-p systems.

### 2.1.2 Molecular dynamics simulation details

All MD simulation were performed using the MPI version of GROMACS with PLUMED plugin. For the atomistic simulations, the potential parameters for the proteins were taken from the Amber99SB-ILDN (28), Charmm27 (29) and Gromos54a7 (23) force field and the TIP3P model (30) was employed for the water molecules. Initially the energy was minimized for 1000 steps with the steepest descent method with 2 fs timestep, to find a local energy minimum and avoid large atom's distance shift in a single timestep. Electrostatic interactions were evaluated with reaction field (31) with a short range Coulomb cutoff of 1.4 nm and vdW cutoff was set to 1.4 nm (32). Subsequent simulation were run in a canonical ensemble (NVT) at 300 K over 10 ps with a modified Berendsen thermostat followed by an isothermal-isobaric ensemble (NPT) at 300 K for 100 ps with the addition of a Berendsen barostat at 1.00 bar and an isotropic pressure coupling (33). Finally the production run was performed for 100 ns and snapshots of the simulations were recorded every 10 ps.

### 2.1.3 Well Tempered Metadynamics simulation details

Additionally, in the production run, the PLUMED plugin is used to add the biasing force to the potential energy and to post-process the generated MD trajectories. The plumed algorithm is activated with an additional `-plumed` flag in the mdrun engine. The peptides' index numbers are specified to allow PLUMED to apply the correct distances as demonstrated in Figure 10.

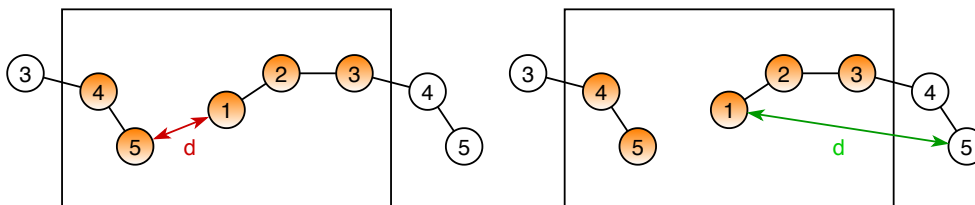


Figure 10: Illustration showing how molecules can split by the periodic boundary conditions and how this can cause a problem when computing collective variables. In the left figure, the distance between atom 1 and 5 is computed from the periodic boundary condition. In the right figure, the distance between atom 1 and 5 is computed ignoring the pbc condition.

A metadynamics bias is applied in the two CVs represented by the torsional angles  $\phi$  and  $\psi$  of all the 3-p and 5-p systems. Figure 11 shows an example of PLUMED file where the two torsional angles  $\phi$  and  $\psi$  are biased.

```
# treat each molecule as whole
WHOLEMOLECULES ENTITY0=1-39
```

```
# define atoms for dihedrals
dih1: TORSION ATOMS=15,17,19,25
dih2: TORSION ATOMS=17,19,25,27
```

```
# metadynamics argument
metad: METAD
ARG=dih1,dih2
PACE=500
HEIGHT=1.0
SIGMA=0.3,0.3
FILE=HILLS
BIASFACTOR=10
GRID_MIN=-pi,-pi
GRID_MAX=pi,pi TEMP=300
GRID_BIN=300,300
```

```
# monitor dihedral angles and print the calculated distance in COLVAR every 500 steps
PRINT STRIDE=5000
ARG=dih1,dih2,metad.bias
FILE=COLVAR
```

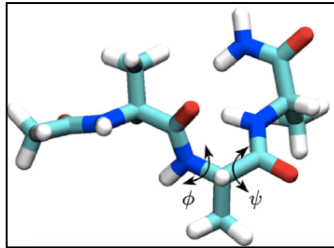


Figure 11: torsional angles  $\phi$  and  $\psi$  in the alanine tripeptide molecule.

The metadynamics arguments were applied depositing a Gaussian every 500 steps (1 ps), with an height equal to 1.0 kcal/mol and a width of 0.3 for both the CVs in a grid going from  $-\pi$  to  $\pi$  with 300 bins in each grid. The dihedral angles  $\phi$  and  $\psi$  values and the bias were printed every 5000 steps.

### 3 Results and Discussion

#### 3.1 Convergence of the Collective Variables and the Free Energy

As explained above, all the solvated tripeptides' and pentapeptides systems were simulated for 100 ns with the bias potential applied. To understand whether the calculation has converged, the system should diffuse along the full CVs space during the simulation as shown in Figure 12.

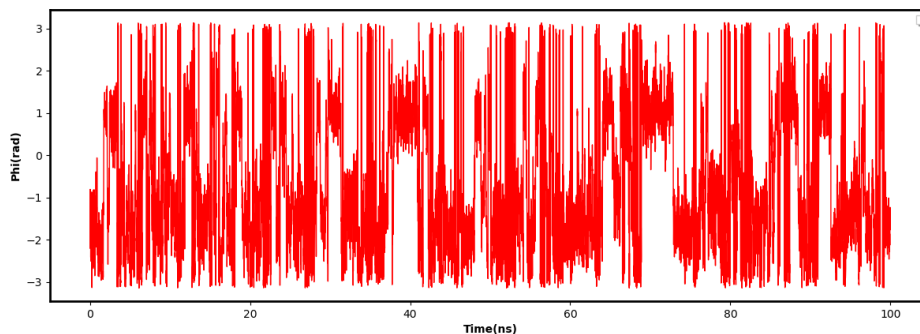


Figure 12: Time evolution of the metadynamics  $\phi$  CV for glutamic acid tripeptide in water.

Along the 100 ns, the system has well sampled the CVs space and all the local minima for this variable have been visited. Figure 13 shows the first 4 ns of simulation for EEE system,

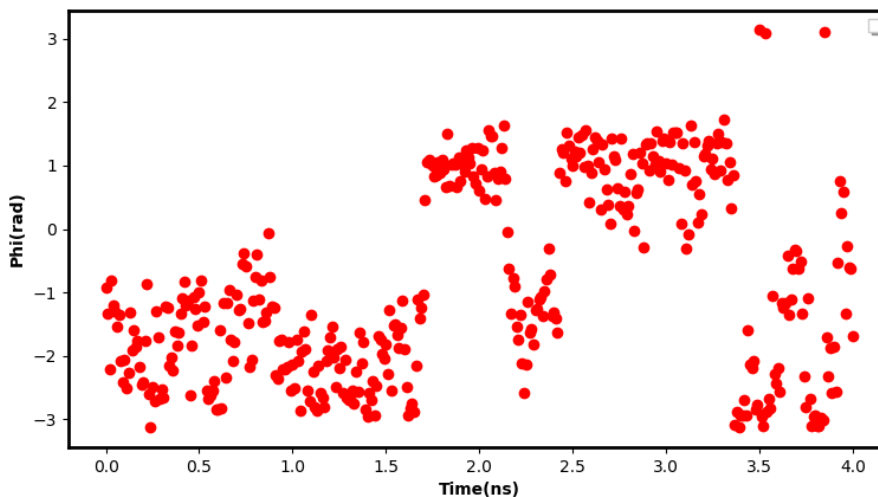


Figure 13: Time evolution of the metadynamics  $\phi$  CV for glutamic acid tripeptide in water.

The system is initialized in one of the two metastable states of EEE and after 1600 ps (1.6 ns), is pushed out of the basin by the metadynamics bias potential to visit another local minimum. Another way to estimate convergence of the system can be observed from the free energy as a function of the CV. Figure 14 shows the reweighted free energy profile of the system as a function of the  $\phi$  torsional angle for the Amber99SB-ILDN force field.

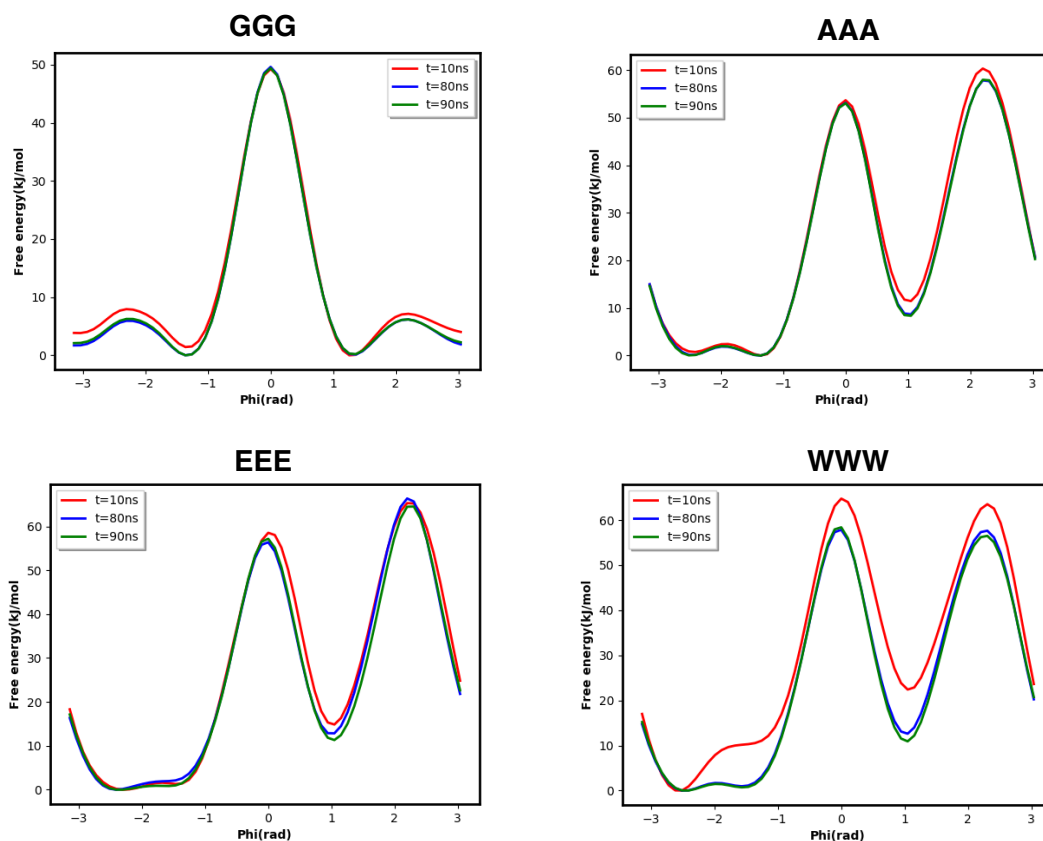


Figure 14: Estimate of the free energy as a function of the dihedral  $\phi$  from a 100ns of glycine, alanine, glutamic acid and tryptophan at different simulation times.

The free energy should not significantly change along the time of the trajectory. For the GGG, AAA and EEE systems, at  $t = 10$  ns the free energy profile has already reached a good convergence and this is why the surface does not change significantly compared to  $t = 80$  ns and  $t = 90$  ns. For the WWW system, there is a large change in free energy profile from  $t = 10$  ns to  $t = 80$  ns while the difference between  $t = 80$  ns and  $t = 90$  ns is not large. This means that the simulation has started at the minimum with  $\phi = -2.5$  and in the first 10 ns of simulation, the other basins were not well sampled while at  $t = 80$  ns the other two basins at  $\phi = -1.5$  and  $\phi = 1.0$  were sufficiently sampled. For this reason, 10 ns are not sufficient to have an appropriate sampling, therefore 100 ns of simulation was the time selected for all the simulations, to also have the same simulation time for all the peptides.



### 3.2 Comparison between Force Fields

Different force fields were employed for the simulations of the peptide chains. Figure 15 shows the reweighted free energy profiles of  $\phi$  for the tripeptides with three different force fields.

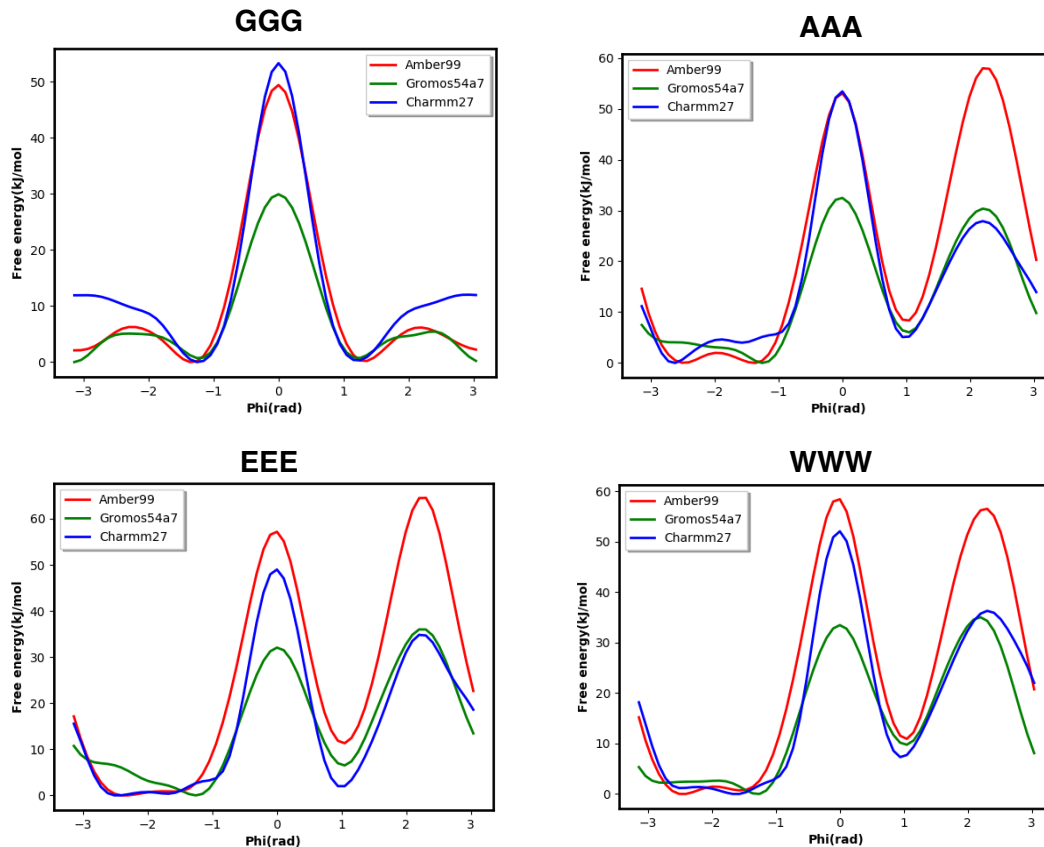


Figure 15: Estimate of the free energy as a function of the dihedral  $\phi$  from a 100ns for glycine, alanine, glutamic acid and tryptophan tripeptide for Amber99SB-ILDN, Gromos54a7 and Charmm27 force fields.

In the glycine system, the minima at  $\phi = -1$  and  $\phi = 1$  and the maximum at  $\phi = 0$  are commonly shared between the force fields. In the alanine, glutamic acid and tryptophan systems, the minima at  $\phi = -3$ ,  $\phi = -1$  and  $\phi = 1$  are also commonly shared between the three force fields. Overall, the force fields share the same common features, with small variations along the free energy profiles.

### 3.3 2D Free Energy Profiles and Block Analysis

The 2D reweighted free energy profiles are calculated for all the peptides. The block analysis is then applied to calculate, for different block sizes, the average free energy and the error associated with each block. The error will increase with block size, until it reaches a plateau when there is no more correlation between data points. Figure 16 shows the 2D-heat maps and the associated block analysis for the tripeptides with the Amber99SB-ILDN force field.

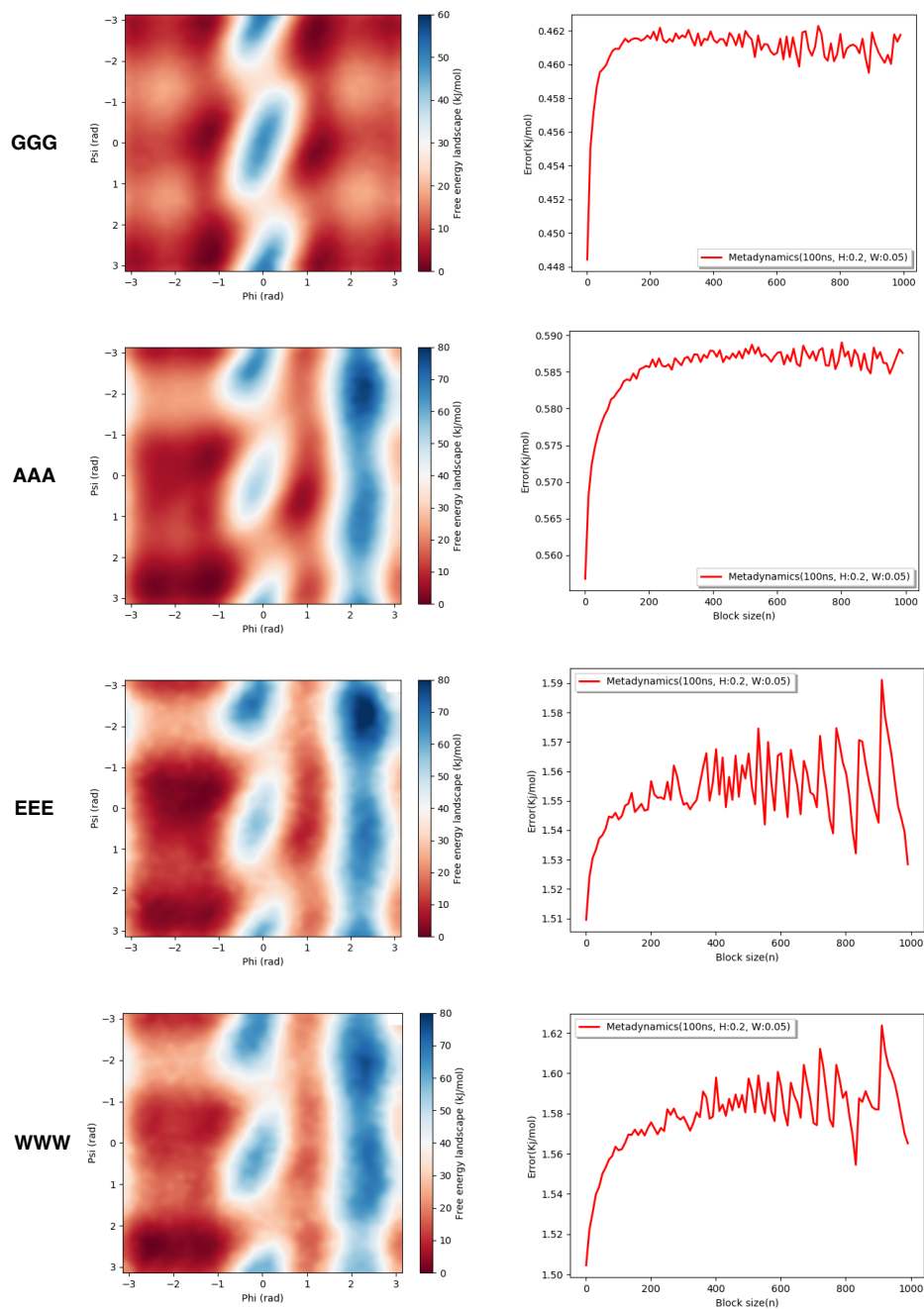


Figure 16: 2D heat maps and block analysis for glycine, alanine, glutamic acid and tryptophan tripeptide for the Amber99SB-ILDN force field.

In all four simulations, the error increases until it reaches a plateau showing the convergence of the simulation. The 2D-heat maps of the CVs were compared also between the different tripeptides. The same pattern is observed for the AAA, EEE and WWW system which is not present in the GGG system. The glycine has only an H atom as a  $C\alpha$  substituent while the alanine, glutamic acid and tryptophan have a methyl, a carboxylic acid and an indole group respectively. Figure 17 compares two points in the 2D heat maps of AAA and GGG.

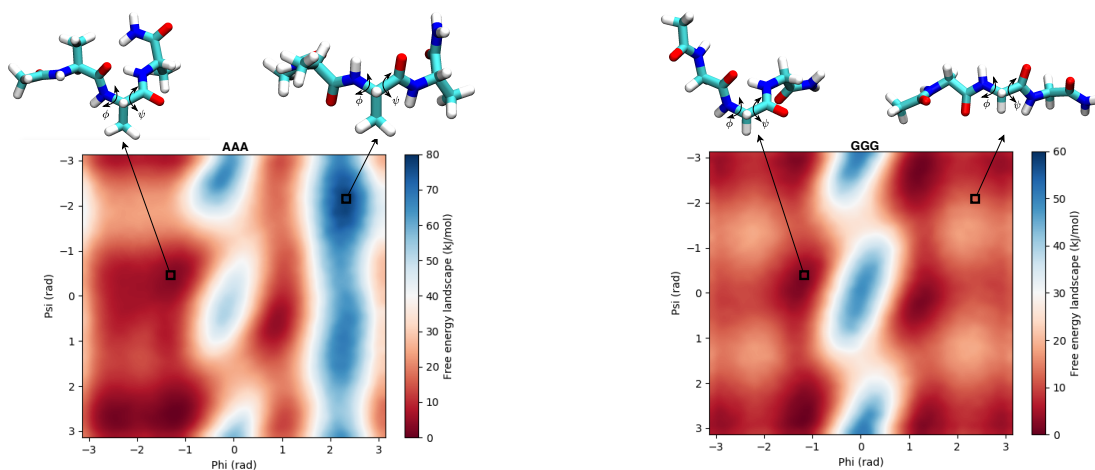


Figure 17: 2D heat maps for alanine and glycine tripeptides. Snapshot of the simulation at  $\phi, \psi = (-1.18, -0.40)$  and  $(2.45, -2.20)$  were extracted for both systems and compared.

The presence of a side chain on the  $C\alpha$  gives an additional maximum in the free energy profiles of alanine, glutamic acid and tryptophan. Glycine is different from the other amino acids since it lacks the presence of a side chain. In particular, it does not have a  $C\beta$  which induces many steric clashes in a Ramachandran plot. For example, Figure 17 shows that the presence of the methyl groups as a side chain in the alanine tripeptide creates a maximum for  $\phi = 2.2$  and all values of  $\psi$  which is not present in the glycine tripeptide system. Alanine dipeptide's  $\phi$  and  $\psi$  torsional angles were previously studied with metadynamics and an unconstrained enhanced sampling method, Gaussian accelerated MD (34; 35). In the metadynamics case, the free energy profile was compared to umbrella sampling showing a significant level of similarity with overall shapes and topologically important points (minima, barriers, and transition pathways) located at the same place.

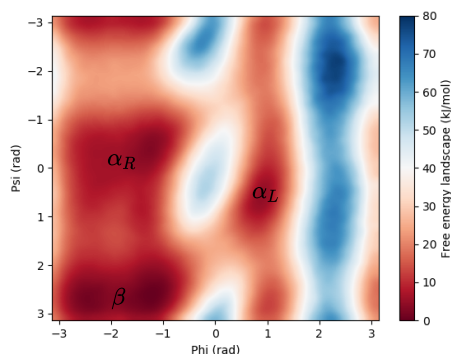


Figure 18: Free energy profile of alanine tripeptide with Amber99SB-ILDN. The three most favorable regions are shown as  $\alpha_R$ ,  $\alpha_L$  and  $\beta$ .

The three most favorable regions  $\alpha_R$ ,  $\alpha_L$  and  $\beta$  are found to be the same between the previously studied alanine dipeptide in both the metadynamics and Gaussian accelerated MD and the currently studied alanine tripeptide providing the same energetic order for the minima. The 2D-heat maps were compared also between the tripeptides and pentapeptides of glycine and alanine as shown in Figure 19 below.

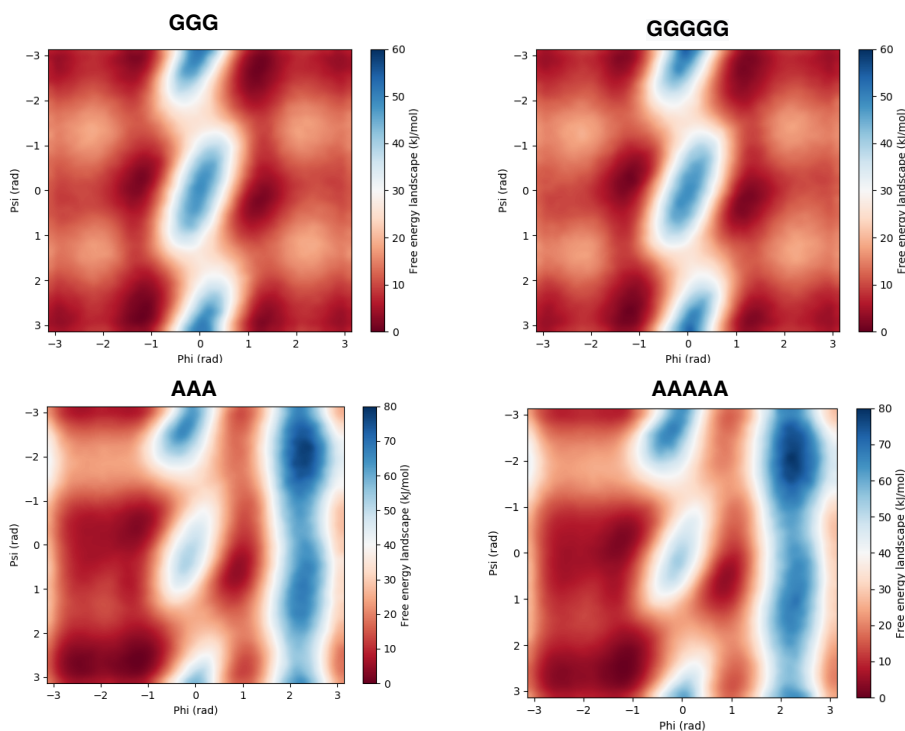


Figure 19: 2D heat maps for alanine and glycine tripeptides and pentapeptides.

In both glycine and alanine systems, no difference is observed in the central torsional angles  $\phi$  and  $\psi$  when the length of the peptide is increased from three to five amino acids. No secondary structure formation is observed with a variation from 3 to 5 amino acids chain.

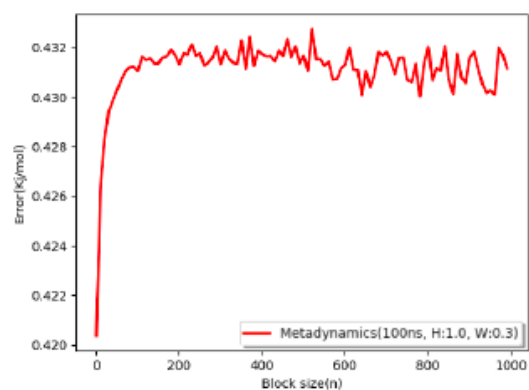
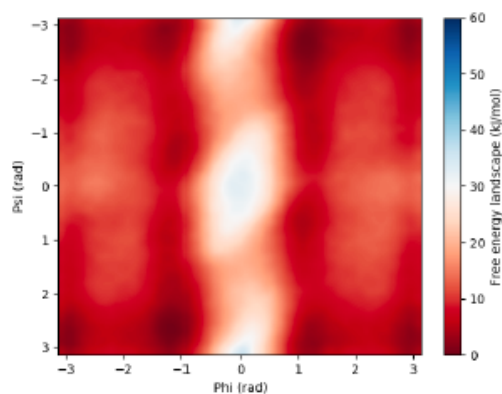
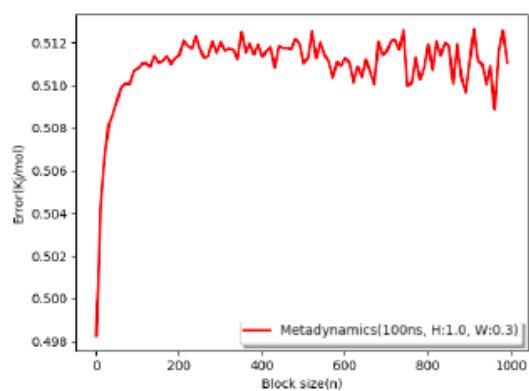
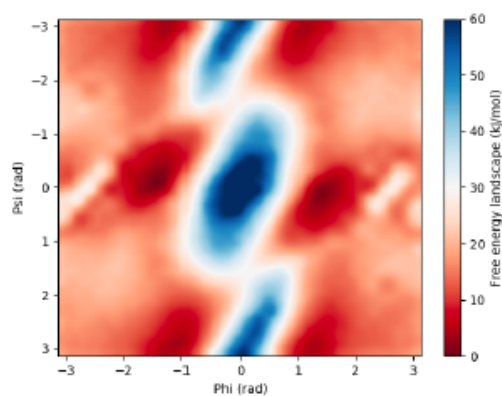
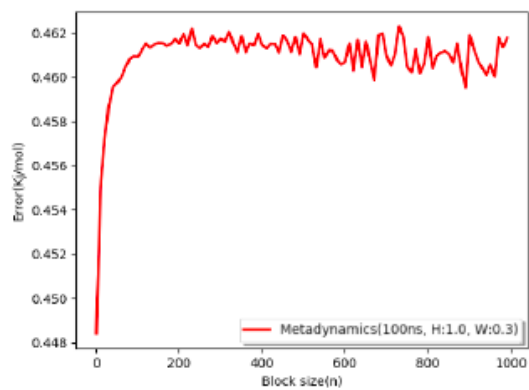
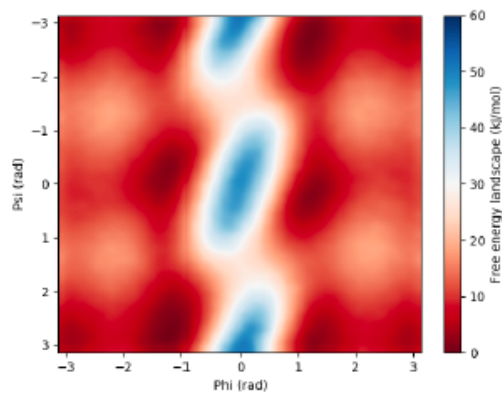
## 4 Conclusions

Protein folding plays a fundamental role in nature for the understanding of the functionality of polypeptides and proteins. Rotation around the torsional angles  $\phi$  and  $\psi$  is limited by steric collisions which together with hydrogen bonding between the amide hydrogens and carbonyl oxygens of the backbone are the main driving forces for a secondary structure formation. In this study, the  $\phi$  and  $\psi$  torsional angles of four different tripeptides are compared and assessed with Amber99SB-ILDN, Charmm27 and Gromos54a7 force fields. For the tested force fields, the local minima are commonly shared with small variations along the free energy profiles. The presence of different side chains in alanine, glutamic acid and tryptophan give an additional maximum in the energy profile due to steric clashes between side chains. Increase in peptide length from three to five amino acids in glycine and alanine do not change the energy profile for the torsional angles.

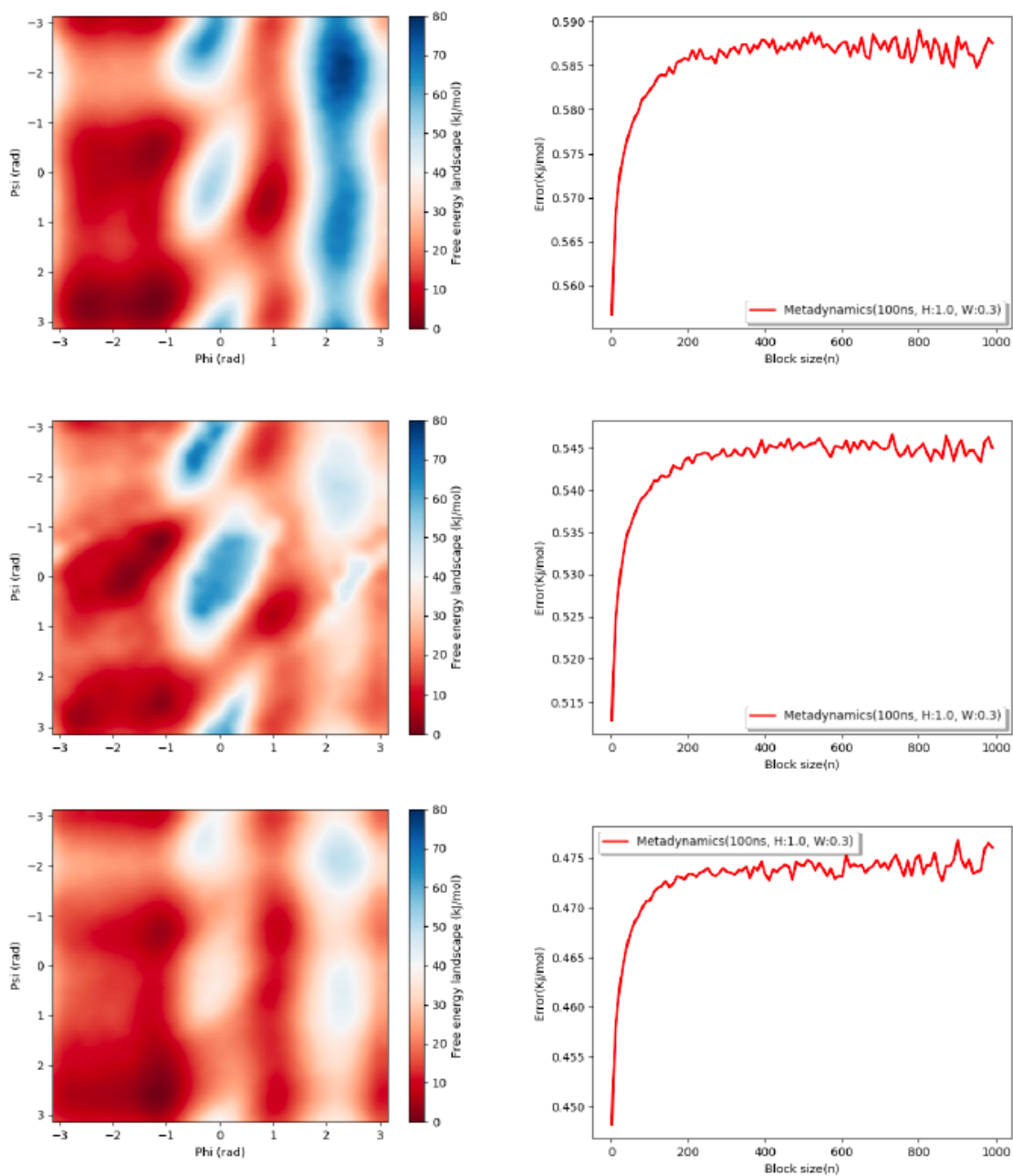
## 5 Appendix

### 5.1 Tripeptides

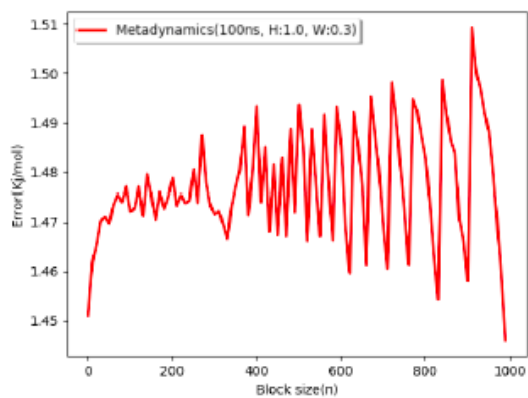
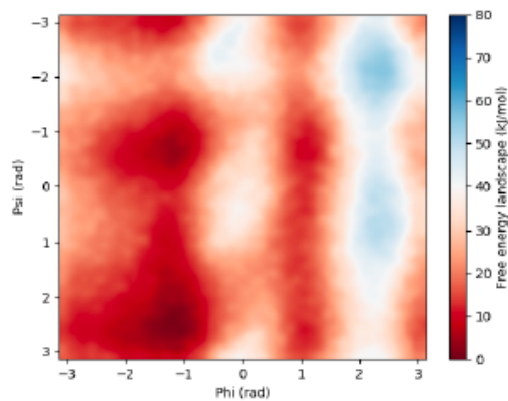
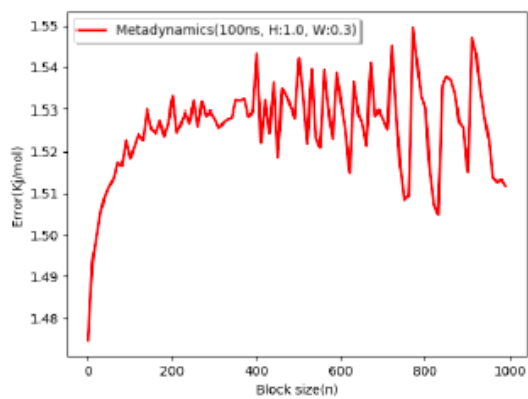
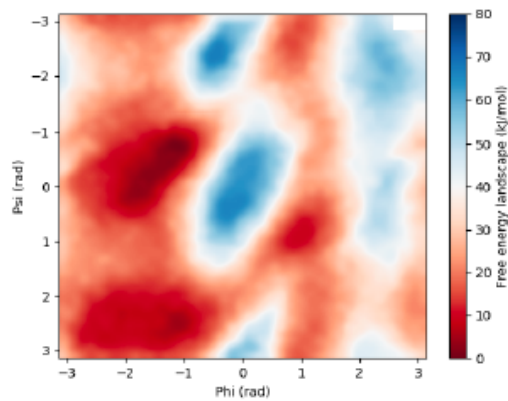
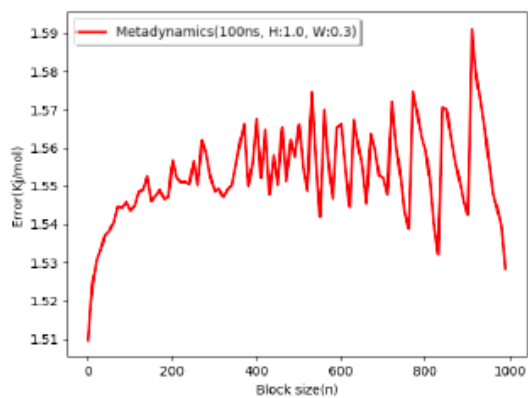
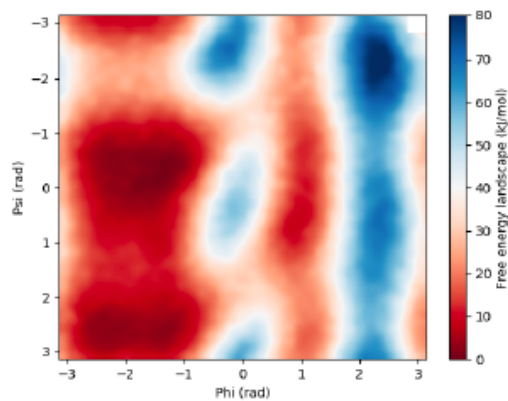
Glycine tripeptide 2D maps and block analysis for Amber99SB-ILDN, Charmm27 and Gromos54a7 respectively.



Alanine tripeptide 2D maps and block analysis for Amber99SB-ILDN, Charmm27 and Gro-mos54a7 respectively.

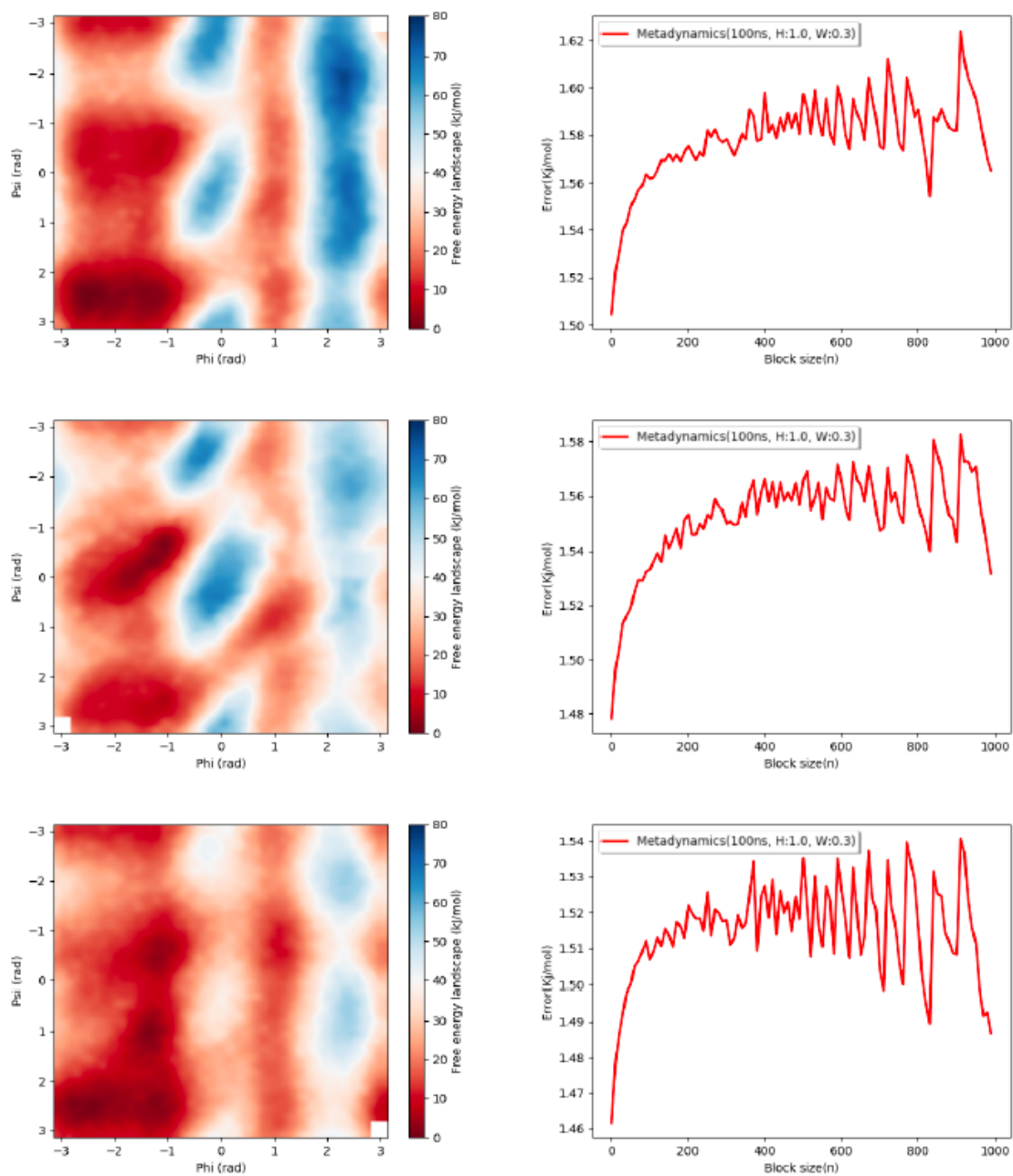


Glutamic acid tripeptide 2D maps and block analysis for Amber99SB-ILDN, Charmm27 and Gromos54a7 respectively.



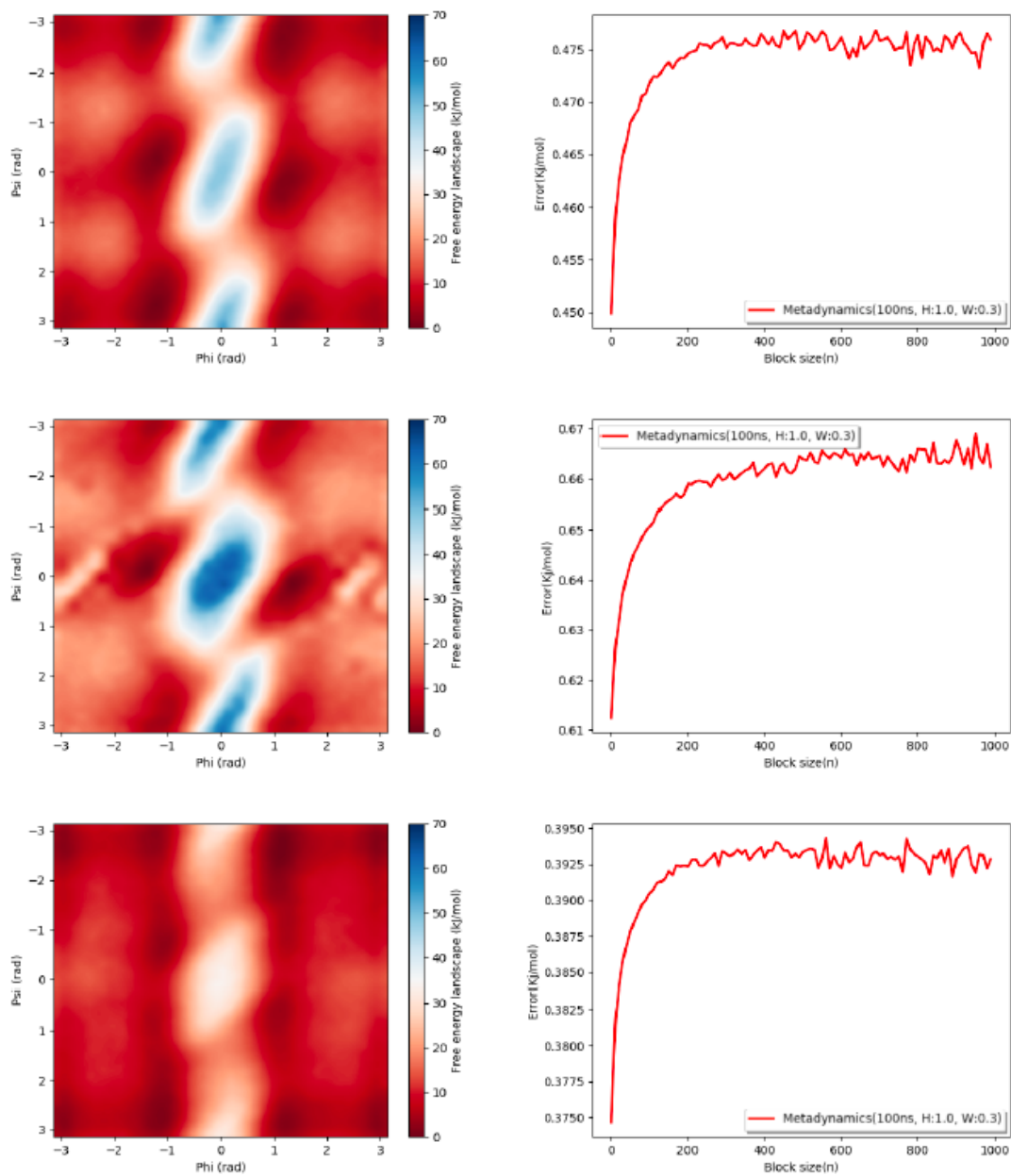


Tryptophan tripeptide 2D maps and block analysis for Amber99SB-ILDN, Charmm27 and Gro-mos54a7 respectively.

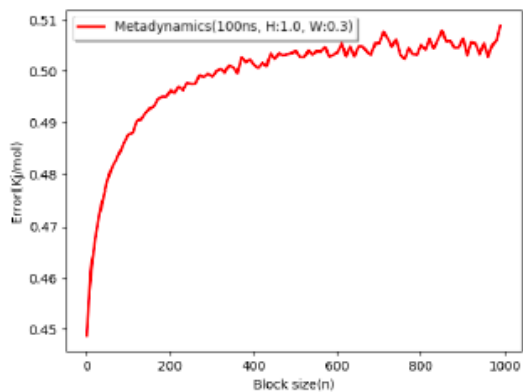
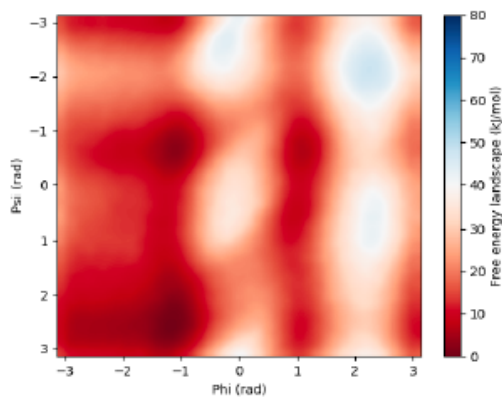
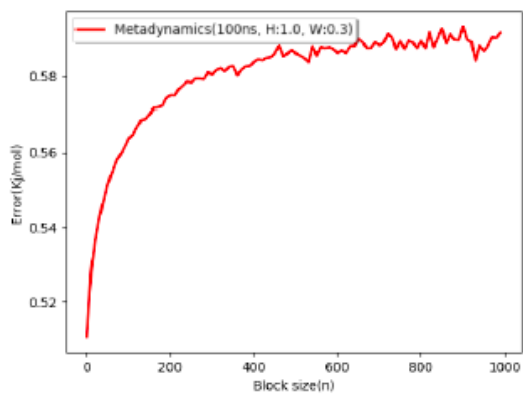
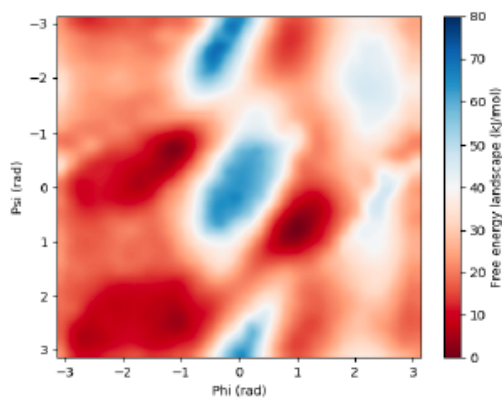
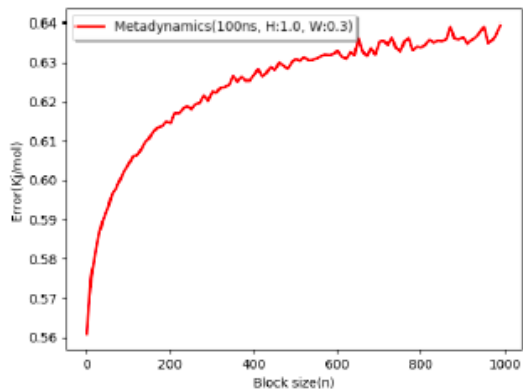
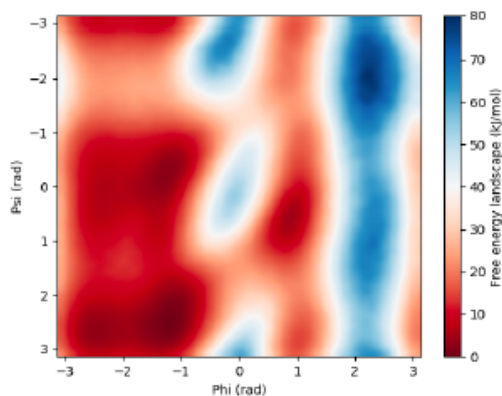


## 5.2 Pentapeptides

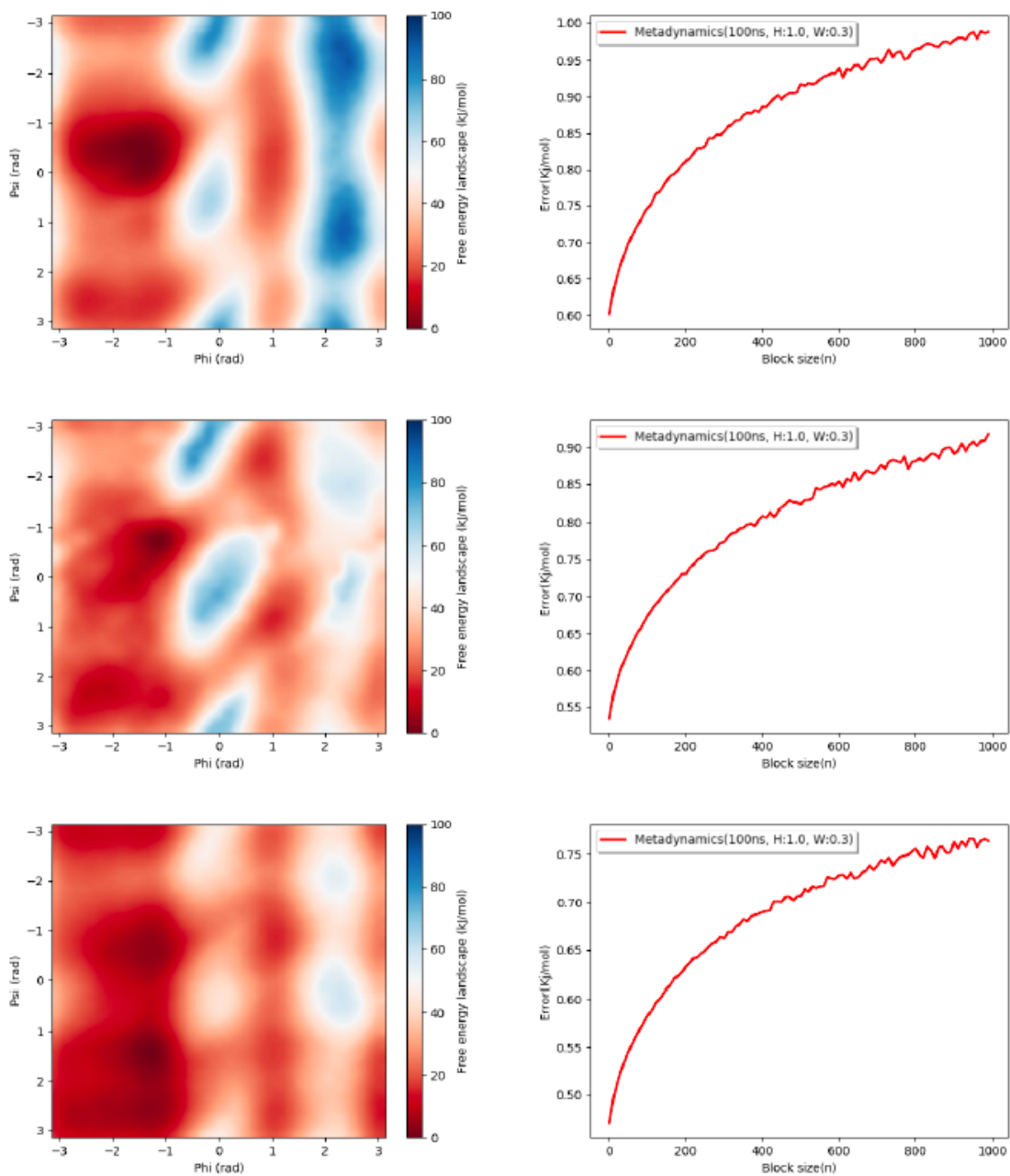
Glycine pentapeptide 2D maps and block analysis for Amber99SB-ILDN, Charmm27 and Gro-mos54a7 respectively.



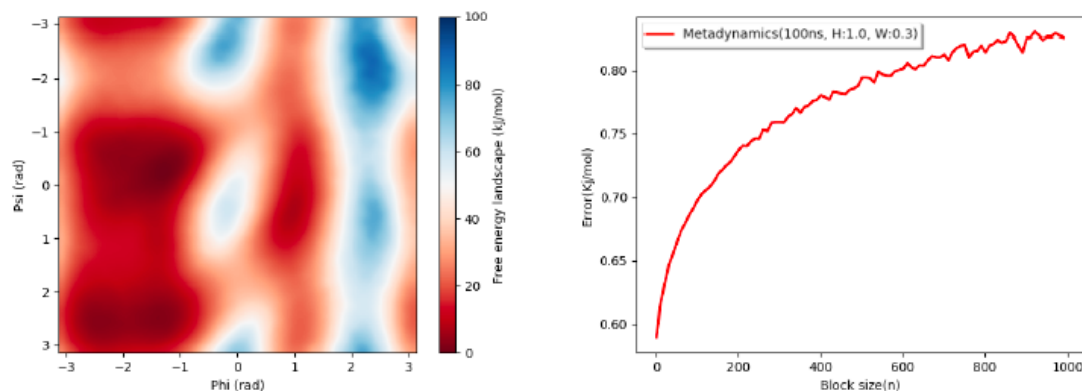
Alanine pentapeptide 2D maps and block analysis for Amber99SB-ILDN, Charmm27 and Gro-  
moms54a7 respectively. For alanine, tryptophan and glutamic acid pentapeptides the free energy  
calculations did not reach convergence.



Tryptophan pentapeptide 2D maps and block analysis for Amber99SB-ILDN, Charmm27 and Gromos54a7 respectively.



Glutamic acid pentapeptide 2D maps and block analysis for Amber99SB-ILDN, Charmm27 and Gromos54a7 respectively.



## References

- [1] Y. Bai and S. W. Englander, “Future directions in folding: The multi-state nature of protein structure,” *Proteins: Structure, Function, and Bioinformatics*, vol. 24, no. 2, pp. 145–151, 1996.
- [2] C. Levinthal, “How to fold graciously,” in *Mossbauer spectroscopy in biological systems* (P. DeBrunner, J. Tsibris, and E. Munck, eds.), (Urbana, IL), University of Illinois Press, 1969.
- [3] D. B. Wetlaufer, “Nucleation, rapid folding, and globular intrachain regions in proteins,” *PNAS*, vol. 70, pp. 697–701, 03 1973.
- [4] P. S. Kim and R. L. Baldwin, “Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding,” vol. 51, no. 1, pp. 459–489, 1982.
- [5] P. S. Kim and R. L. Baldwin, “Intermediates in the folding reactions of small proteins,” vol. 59, no. 1, pp. 631–660, 1990.
- [6] S. W. Englander and L. Mayne, “The nature of protein folding pathways,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 45, pp. 15873–15880, 2014.
- [7] K. Henzler-Wildman and D. Kern, “Dynamic personalities of proteins,” *Nature*, vol. 450, no. 7172, pp. 964–972, 2007.
- [8] G. Ramachandran, C. Ramakrishnan, and V. Sasisenharan, “Stereochemistry of polypeptide chain configurations,” *Journal of molecular biology*, vol. 7, pp. 95–99, 1963.
- [9] van Gunsteren WF, B. D, B. R, C. I, C. M, D. X, G. P, G. DP, G. A, H. PH, K. MA, O. C, S. M, T. D, van der Vegt NF, and Y. HB, “Biomolecular modeling: Goals, problems, perspectives.,” vol. 45, no. 25, pp. 4064–92, 2006.

- [10] M. Karplus and J. A. McCammon, “Molecular dynamics simulations of biomolecules,” *Nature Structural Biology*, vol. 9, no. 9, pp. 646–652, 2002.
- [11] J. A. McCammon, B. R. Gelin, and M. Karplus, “Dynamics of folded proteins,” *Nature*, vol. 267, no. 5612, pp. 585–590, 1977.
- [12] S. A. Adcock and J. A. McCammon, “Molecular dynamics: survey of methods for simulating the activity of proteins,” *Chemical Reviews*, vol. 106, no. 5, pp. 1589–1615, 2006.
- [13] C. D. Christ, A. E. Mark, and W. F. van Gunsteren, “Basic ingredients of free energy calculations: A review,” *Journal of Computational Chemistry*, vol. 31, no. 8, pp. 1569–1582, 2010.
- [14] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, “A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos force-field parameter sets 53a5 and 53a6,” *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1656–1676, 2004.
- [15] Z. Lin and W. F. van Gunsteren, “Refinement of the application of the gromos 54a7 force field to -peptides,” *Journal of Computational Chemistry*, vol. 34, no. 32, pp. 2796–2805, 2013.
- [16] L. Verlet, “Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules,” *Phys. Rev.*, vol. 159, pp. 98–103, Jul 1967.
- [17] A. Leach, *Molecular Modelling: Principles and Applications*. Prentice Hall, 2001.
- [18] V. Spiwok, Z. Sucer, and P. Hosek, “Enhanced sampling techniques in biomolecular simulations,” vol. 33, no. 6 Part 2, pp. 1130–1140, 2015.
- [19] Y. I. Yang, Q. Shao, J. Zhang, L. Yang, and Y. Q. Gao, “Enhanced sampling in molecular dynamics,” *The Journal of Chemical Physics*, vol. 151, no. 7, p. 070902, 2019.
- [20] G. Torrie and J. Valleau, “Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling,” *Journal of Computational Physics*, vol. 23, no. 2, pp. 187 – 199, 1977.
- [21] B. A. B. G, and P. M, “Well-tempered metadynamics: a smoothly converging and tunable free-energy method,” vol. 100, no. 2, p. 020603, 2008.
- [22] A. Barducci, M. Bonomi, and M. Parrinello, “Metadynamics,” *WIREs Computational Molecular Science*, vol. 1, no. 5, pp. 826–843, 2011.
- [23] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, “Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers,” *SoftwareX*, vol. 1-2, pp. 19 – 25, 2015.
- [24] R. Salomon-Ferrer, D. A. Case, and R. C. Walker, “An overview of the amber biomolecular simulation package,” *WIREs Computational Molecular Science*, vol. 3, no. 2, pp. 198–210, 2013.

- [25] D. R. Roe and T. E. Cheatham, "Ptraaj and cpptraj: Software for processing and analysis of molecular dynamics trajectory data," *Journal of Chemical Theory and Computation*, vol. 9, no. 7, pp. 3084–3095, 2013.
- [26] W. Humphrey, A. Dalke, and K. Schulten, "Vmd: Visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33 – 38, 1996.
- [27] M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek, and G. R. Hutchison, "Avogadro: an advanced semantic chemical editor, visualization, and analysis platform," *Journal of Cheminformatics*, vol. 4, no. 1, p. 17, 2012.
- [28] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the amber ff99sb protein force field," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 8, pp. 1950–1958, 2010.
- [29] D. Yin and A. D. MacKerell Jr., "Combined ab initio/empirical approach for optimization of lennard–jones parameters," *Journal of Computational Chemistry*, vol. 19, no. 3, pp. 334–348, 1998.
- [30] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, no. 2, pp. 926–935, 1983.
- [31] I. G. Tironi, R. Sperb, P. E. Smith, and W. F. van Gunsteren, "A generalized reaction field method for molecular dynamics simulations," *The Journal of Chemical Physics*, vol. 102, no. 13, pp. 5451–5459, 1995.
- [32] J. Barker and R. Watts, "Monte carlo studies of the dielectric properties of water-like models," *Molecular Physics*, vol. 26, no. 3, pp. 789–792, 1973.
- [33] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *The Journal of Chemical Physics*, vol. 81, no. 8, pp. 3684–3690, 1984.
- [34] J. Vymětal and J. Vondrášek, "Metadynamics as a tool for mapping the conformational and free-energy space of peptides — the alanine dipeptide case study," *The Journal of Physical Chemistry B*, vol. 114, no. 16, pp. 5632–5642, 2010.
- [35] Y. Miao, V. A. Feher, and J. A. McCammon, "Gaussian accelerated molecular dynamics: Unconstrained enhanced sampling and free energy calculation," *Journal of Chemical Theory and Computation*, vol. 11, no. 8, pp. 3584–3595, 2015.