Enhanced sampling and force field corrections for RNA oligomers

A thesis submitted for the degree of
Philosophiæ Doctor

(October 2016)

Candidate
Alejandro Gil Ley

Supervisor
Prof. Giovanni Bussi

Molecular and Statistical Biophysics Sector

PhD course in Physics and Chemistry of Biological Systems
Scuola Internazionale Superiore di Studi Avanzati - SISSA
Trieste

Abstract

The computational study of conformational transitions in nucleic acids still faces many challenges. For example, in the case of single stranded RNA tetranucleotides, agreement between simulations and experiments is not satisfactory due to inaccuracies in the force fields commonly used in molecular dynamics. Improvement of force fields is however hindered by the difficulties of decoupling those errors from the statistical errors caused by insufficient sampling. We here tackle both problems by introducing a novel enhancing sampling method and using experimental data to improve RNA force fields.

In this novel method, concurrent well-tempered metadynamics are integrated in a Hamiltonian replica-exchange scheme. The ladder of replicas is built with different strength of the bias potential exploiting the tunability of well-tempered metadynamics. Using this method, free-energy barriers associated to individual collective variables are significantly reduced compared with simple force-field scaling. The introduced methodology is flexible and allows adaptive bias potentials to be self-consistently constructed for a large number of simple collective variables, such as distances and dihedral angles.

Additionally, a modified metadynamics algorithm is used to calculate correcting potentials designed to enforce distributions of backbone torsion angles taken from experimental structures. Replica-exchange simulations of tetranucleotides including these correcting potentials show significantly better agreement with independent solution experiments for the oligonucleotides containing pyrimidine bases. Although the proposed corrections do not seem to be portable to generic RNA systems, the simulations reveal the importance of the α and ζ backbone angles for the modulation of the RNA conformational ensemble. The correction protocol presented here suggests a systematic procedure for force-field refinement.

Contents

Chapter 1

Introduction

RNA is being recognized as a key player on many different functions in the cell [1–3] and as a potential target for therapeutics [4, 5]. Understanding the physical interactions of RNA molecules that are associated with folding, catalysis and other essential molecular recognition processes can provide insight into those functions [6, 7]. In order to understand the structure-function relations governing those processes one should go beyond strict structural aspects and explore RNA dynamical features.

Beside experimental single molecular approaches based on fluorescence [8–10] and force measurements techniques [11, 12], computational techniques like molecular dynamics (MD) simulations [13, 14] have provided a microscopic picture of the mechanism and dynamics of RNA systems [15, 16]. MD simulations as standalone experiments started for proteins in 1975 [14] and for RNA in 1984 [17–20]. But as nucleic acids are highly charged polymers, stable MD simulations of fully solvated systems were only achieved with the introduction of the particle mesh Ewald method [21] for the treatment of long-range electrostatic interaction in 1995 [22, 23]. Since then, MD has gained in robustness and predictive power [24], and the advances in software and hardware have made the simulation of complex conformational transitions like the folding and refolding of RNA tetraloops [25–27], and the analysis of large molecular machines like the ribosome [16, 28, 29], possible.

Simulation times of microseconds are normally accessible nowadays, and even longer timescales can be simulated when specialized hardware is used [30]. However, many molecular processes are rare events that could be seen only a few times, if seen at all, in a microsecond timescale. In order to obtain precise averages from the computational generated ensembles, advanced sampling techniques have been developed to accelerate the exploration of the conformational space and bridge the gap between experiments and simulations. However, in the case of RNA, simulations of short oligonucleotides with parallel tempering (a popular enhanced sampling technique technique) and Hamiltonian replica exchange (another powerful technique) have been shown to generate unconverged ensembles, even for simulations close to a 100 μs [31, 32]. On the other hand, the current empirical functions, force fields, used to represent the energetic interactions in RNAs are not accurate enough to reproduce solution experiments of unstructured

oligonucleotide systems [33, 34] or to predict the correct stability of RNA tetraloops [27]. Therefore, further advances in enhanced sampling and force field refinements are required in the field of computer-simulation experiments of RNA.

The results presented in this thesis are mainly concerned with the development of a new enhanced sampling method for the study of RNA molecules and with the inclusion of empirical corrections into the RNA force field that improve the agreement with solution NMR experiments. The new method introduced here, replica exchange with collective-variable tempering (RECT), greatly improves the conformational sampling of the challenging RNA tetranucleotides, which have become a benchmark to evaluate enhanced sampling techniques and force field accuracy [31–34]. Concerning the force field corrections, a self-consistent procedure based on metadynamics [35, 36] is used to calculate correcting potentials that enforce distributions of dihedral angles taken from experimental structures in the RNA AMBER force field [37–39]. Since the target distributions are multimodal, we use RECT to accelerate the convergence of the correcting potential calculation. The resulting corrections are tested on tetranucleotides where standard force field parameters are known to fail in reproducing NMR data. The new AMBER force field lead to ambiguous results when applied to different tetranucleotide sequences. However, the simulations reveal that by only penalizing a rotameric phosphate-backbone conformation, the quality of the ensemble is significantly improved to levels not reported before.

The material presented in this thesis is organized as follows:

In Chapter 2, an overview of RNA structure is presented, focused on the internal motions that characterize the flexibility of single-stranded RNA structures. Chapter 3 is devoted to a brief summary of the state of the art of RNA force fields and enhanced sampling methods commonly used to aid the exploration of RNA conformational space in MD simulations. Special attention is dedicated to well-tempered metadynamics [40] which is the base of the new enhancing sampling method presented in Chapter 4. In this new method concurrent well-tempered metadynamics simulations are integrated in a replica exchange framework so as to effectively overcome the high free energy barriers of the RNA dihedral angles transitions. Chapter 5 presents the results from the empirical correction of the RNA force field using a combination of metadynamics-based techniques, which suggest a systematic procedure for force field refinement. Finally, the conclusions of the thesis and the perspectives are contained in Chapter 6.

The results discussed in Chapter 4 and 5 are largely based on the following publications:

Gil-Ley, A and Bussi, G. *Enhanced Conformational Sampling using Replica Exchange with Collective-Variable Tempering*. Journal of Chemical Theory and Computation. 2015, 11 (3), 1077-1085. (Cover article for the March 2015 issue of JCTC, Figure 1.1)

Gil-Ley, A.; Bottaro, S.; Bussi, G. *Empirical corrections to the Amber RNA force field with Target Metadynamics*. Journal of Chemical Theory and Computation. 2016, 12 (6), 2790–2798.

In addition, collaboration with other members of Prof. Bussi's group has lead to the following publications, not included in this thesis:

Bottaro, S.; Gil-Ley, A.; Bussi, G. *RNA Folding Pathways in Stop Motion*. Nucleic Acid Research. 2016, 44 (12), 5883-5891.

Cesari, A.; Gil-Ley, A.; Bussi, G. *Combining simulations and solution experiments as a paradigm for RNA force field refinement*. Submitted.

Figure 1.1: Cover art highlighting the RECT method.

Chapter 2

RNA Structure

RNA is a polymeric molecule formed by a combination of 4 different nucleotides [41–43]. Each nucleotide contains a furanose-type sugar (β-D ribose), an aromatic heterocyclic base, and a phosphate group. The nucleotides are linked to one another in a linear manner, by phosphodiester bonds between the sugar of one nucleotide and the phosphate group of the adjacent nucleotide. The most common nucleobase types are: adenine (A), cytosine (C), guanine (G), and uracil (U). Cytosine and uracil are derivatives of the pyrimidine (Py) ring, while adenine and guanine have a purine (Pu) scaffold, a pyrimidine ring fused to an imidazole ring. The phosphate groups have a negative charge each, making RNA a polyanionic molecule. The structure of the ribose ring and the nucleobases is represented in Figure 2.1.
The single stranded RNA flexibility is characterize by the motion of the nucleobase with respect to the sugar (*syn*/*anti* orientation around the torsion angle χ), the pseudo-rotation of the furanose ring, and the conformation of the sugar-phosphate backbone (torsion angles γ, α, β, ϵ and ζ, defined in Fig 2.2). In the following sections each of these internal modes of motion will be discussed in more detail.

2.1 *syn*/*anti* orientation about the glycosyl bond

The glycosidic bond links a ribose sugar and a nucleobase. Structural constrains result in marked preferences for the torsion angle χ around this bond. There are two principal low-energy domains for this angle, corresponding to the *anti* conformation ($\chi = 180 \pm 90^{o}$) and the *syn* conformation ($\chi = 0 \pm 90^{o}$) [44]. In the *anti* conformation the face of the nucleobase is directed away from the sugar ring, while in the *syn* it is over or toward the sugar (Figure 2.3). In general, it is expected the *anti* conformation to be more energetically favorable than the *syn*, as in the latter one the bulky part of the base is located over the sugar, which generates close interatomic contacts. Due to primarily steric hindrances, the barrier to the interconversion between *syn* and *anti* conformations is higher for Py than for Pu [45–50]. Previous ultrasonic relaxation experiments suggested that the barrier height for the base rotation in Pu nucleosides (in a nucleoside no phosphate group is attached to the 5' hydroxyl group) was 6.2

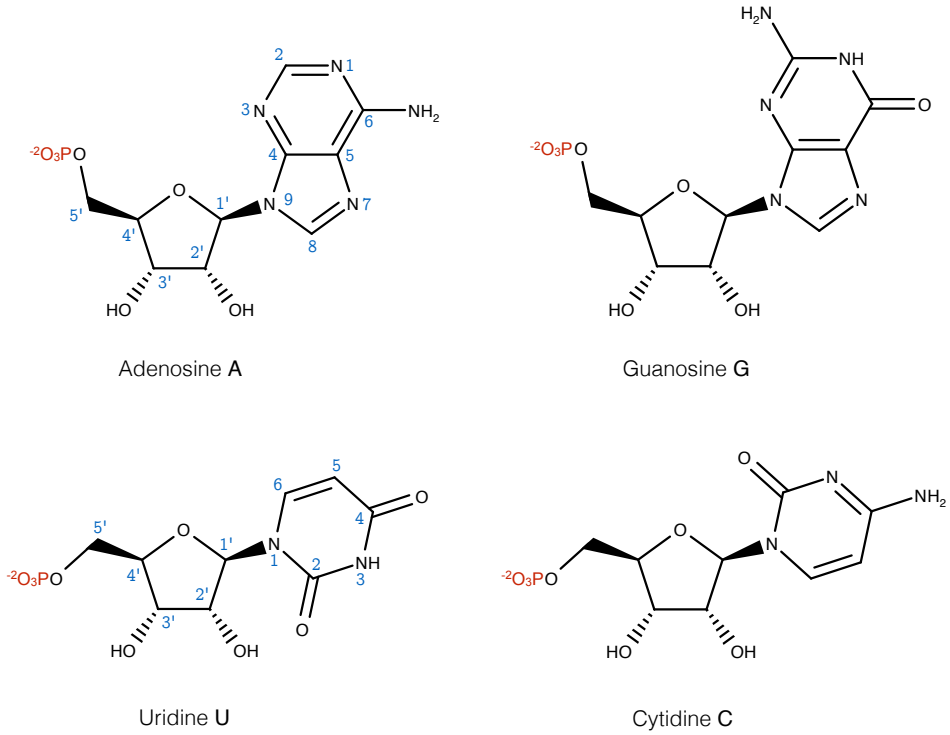


Figure 2.1: Chemical structure and atom notations of the most common purine and pyrimidine ribonucleotides. The substituent atom take the same notation as the immediate carbon or nitrogen in the sugar or base ring (*e.g.* the O2' indicates the oxygen linked to the C2' of the sugar).

kcal/mol, and the transition rate was 0.25 ns^{-1} [51]. Similarly, ultrasonic relaxation with ribonucleotides 5'-AMP and 5'-GMP showed that the activation enthalpy changes for the rotation was 1 and 1.5 kcal/mol [52, 53]. Moreover, in this study the *syn* conformation was found more stable than the *anti*, and its stability was believed to be controlled by entropy rather than enthalpy. In the case of Py nucleosides, recent NMR experiments estimated the free energy difference $\Delta G^{\circ}_{anti \to syn}$ to be 1.07 kcal/mol for Cytosine and $\Delta G^{\circ}_{anti \to syn}$ 1.45 for Uridine [54]. Although in RNA crystal structures the *anti* rotamer is the most common, a recent study have revealed that the majority of the *syn* nucleobases are in regions assigned to function, with many *syn* nucleobases interacting directly with a ligand or ribozyme active site [55].

2.2 Pseudo-rotation of the sugar ring

The five-membered ribose sugar ring is innately nonplanar. This non-planarity is termed puckering [43]. The ring can be puckered in an envelope (E) form with four atoms in a plane and the fifth atom out by approximately 0.5 Å; or in a twist (T) form with two adjacent atoms displaced on opposite sides of a plane passing through the other three atoms [42]. Conventionally, atoms displaced from these three- or four-atom planes and on the same side as C5', are called *endo*; those on the opposite side are called *exo*. The sugar puckering modes are illustrated in Fig. 2.4, with the two most common

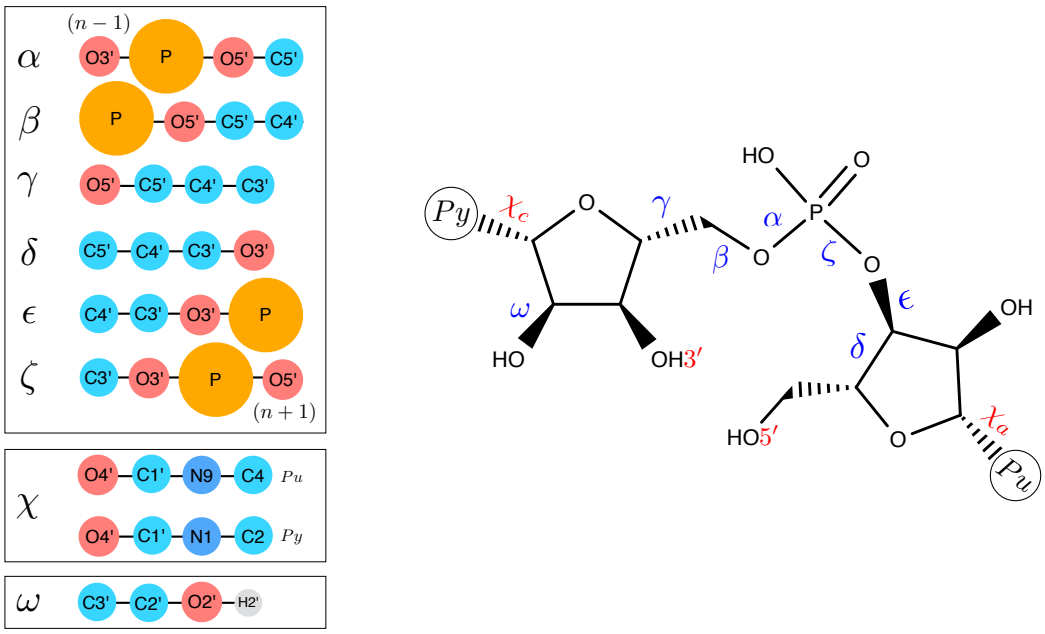


Figure 2.2: Definition of torsion angles in RNA nucleotides. The sizes of the atom circles illustrate their relative van der Waals radii. Atoms designated (n-1) and (n+1) belong to adjacent units.

states of the ribose, the C2'-*endo* (^{2}E) and the C3'-*endo* (^{3}E). Several possible geometric definitions of sugar puckering exist [56–63]. The furanose sugar has five internal sugar torsions $\nu 0 - \nu 4$ but only two torsions angles are needed to define its geometry . In this thesis we adopt the approaches introduced by Hill and Reilly [62] and by Huang *et al.* [63]. The main difference between these approaches is the definition of the two dihedral angles selected as pseudorotation variables (improper in [62] or proper in [63]). The methods simplify the definition of pseudorotation of furanose puckering and allow easy and accurate calculation of the structural quantities (See Fig. 2.5). Changes in sugar pucker are important determinants of oligo- and polynucleotide structure because they can alter the orientation of C1', C3' and C4' substituents, resulting in major changes in backbone conformation and overall structure (e.g. C3'-*endo* in A-DNA or RNA, while C2'-*endo* in B-DNA) [43].

In solution the C2'-*endo* and C3'-*endo* sates are in rapid equilibrium, as shown by NMR investigations and theoretical studies [64–69]. In general terms, free Py nucleotides favor C3'-*endo* puckering while Pu derivatives occur preferentially in the C2'-*endo* mode. The free-energy difference between C2'-*endo* and C3'-*endo* ($\Delta G^{\circ}_{C2' \rightarrow C3'}$) in RNA nucleosides has been estimated by NMR experiments: for Adenosine (0.43 kcal/mol [70]), Guanosine (0.36 [70]), Cytosine (-0.24 [54] / -0.36 [70]) and Uridine (-0.15 [54] / -0.07 [70]). Interconversion between C2'-*endo* and C3'-*endo* states has two principal routes, one with a barrier around O4'-*endo* and another one passing through the O4'-*exo* pucker. The O4'-*exo* route is more energetically unfavorable [63, 71**?**], which can be understood based on steric hindrances: in the O4'-*exo* pucker the base and C5' exocyclic substituents are both in axial position which leads to steric interference, while in the O4'-*endo* mode both are in equatorial orientation which place them

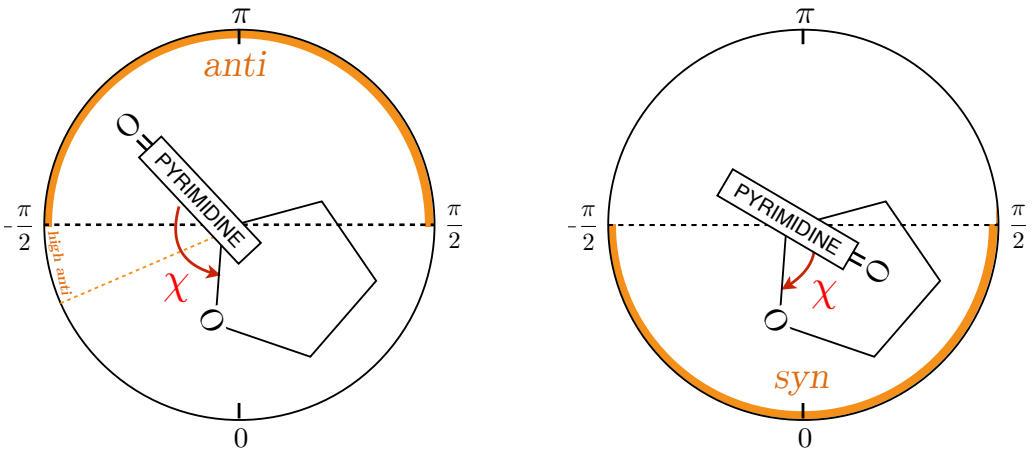


Figure 2.3: Definition of *anti* and *syn* conformational ranges, shown for Pyr residue. The nucleobase is toward the viewer and the base is rotated relative to the sugar. Adapted from ref [42].

farther apart [42].

The interaction between the sugar and the nucleobase can modulate the two-state C2'-*endo* ⇌ C3'-*endo* pseudorotational equilibrium on the basis of various steric and stereoelectronic effects. In terms of steric effect alone, C2'-*endo*-type pseudorotamers are energetically favored in comparison with C3'-*endo*-type counterparts, since the pseudo-equatorially oriented nucleobase in the former exerts less steric repulsions with the other substituents on the pentofuranose moiety than when it is pseudoaxial in the latter [72, 73]. Stereoelectronic forces can also modulate the pseudorotational equilibrium. Some of these forces oppose to the steric ones and stabilize the C3'-*endo*-type conformations in RNA nucleotides [74, 75]. Scrutiny of all nucleoside crystal data suggests that Pu nucleosides with C2'-*endo* pucker adopt both *syn* and *anti* forms in nearly equal distribution but C3'-*endo* puckering shifts the orientation about the glycosyl bond to *anti* [76]. For Py ribonucleosides, the *syn* form is found less frequently, and it occurs with both C2'- and C3'-*endo* sugars, while the dominant *anti* conformation is associated with C3'-*endo* [77–80].

It is important to summarize the distinction between the ribose and deoxyribose puckering cycles, as the only difference between RNA and DNA comes from the presence of the hydroxyl substituent at the 2' position. In polymeric DNA structures, deoxyriboses are primarily in the C2'-*endo* form, while in RNA molecules, ribonucleotides favor C3'-*endo* [42, 43]. Systematic surveys of 2'-substituted adenosine and uridine derivatives indicated that the amount of the C3'-*endo* conformer increases linearly with the electronegativity of the 2'-substituent [75]. Moreover in RNA C3'-*endo* is also stabilized by additional hydrogen bonding opportunities, for example a direct hydrogen bond between O2'H and the O4' of the adjacent nucleotide, as well as a water-mediated hydrogen-bonded bridge between O2'H and the 3'-phosphate, have been advocated as factors of the C3'-*endo* stabilization in RNAs [81–86].

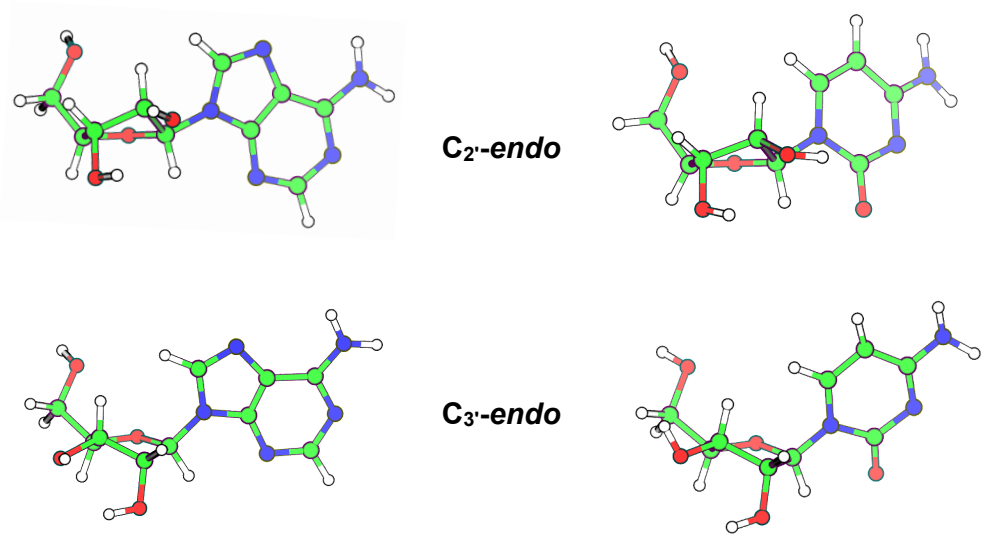


Figure 2.4: C2'-*endo* and C3'-*endo* sugar puckering conformations for an Adenosine nucleoside (right) and a Cytidine nucleoside (left). The conformation of the nucleobase is *anti* in all the structures.

In resume, the conformational equilibrium of the ribose ring is energetically controlled by various competing factors, like stereoelectronic effects, hydration, steric effects, inter and intramolecular hydrogen bonds or by the conformational constraints imposed by the RNA polymorphism [42, 43, 75].

2.3 Conformations of the sugar-phosphate backbone

The sugar-phosphodiester backbone of an oligonucleotide has six different torsion angles, designated α, β, γ, δ, ϵ and ζ in addition to the five internal sugar torsions $\nu 0 - \nu 4$ and the glycosidic angle χ (Figure 2.2). Steric considerations alone dictate that the backbone angles are restricted to discrete ranges [87, 88]. A common convention for describing these ranges is to term values of $\sim 0°$ as *cis*, $\sim 180°$ as *trans* (t), and $\sim \pm 60°$ as *gauche*$^{\pm}$ ($g^{\pm}$). The allowed ranges for these angles is shown in Figure 2.6. Determining the energetic balance among the different allowed conformers is a difficult task considering it is the result of several competing factors, like steric interactions, stereolectronic effects and electrostatic repulsions [42, 79].

The orientation along the exocyclic C4'-C5' bond is controlled by the γ angle. Rotation about this bond plays a crucial role in positioning the 5'-phosphate group relative to the sugar and base. The γ angle has three main rotamers g^+, g^- and t, the classical threefold staggered pattern of ethane. There is, however, a similarity between the χ and γ rotations: *syn* and g^+ position the nucleobase and the O5'PO_3^{-2} over the ribose whereas *anti* and g^- or t direct the base and O5'PO_3^{-2} away from it. The three rotamers are not uniformly populated because their distribution is dependent on sugar pucker and on the base identity [42].

The rotations about C−O ester bonds are determined by the ϵ and β torsion angles. These rotations are the more restricted transitions in the nucleic acid backbone. In

$$Z_x = \frac{\nu_1 + \nu_3}{2\cos(4\pi/5)}$$

$$Z_y = \frac{\nu_1 - \nu_3}{2\sin(4\pi/5)}$$

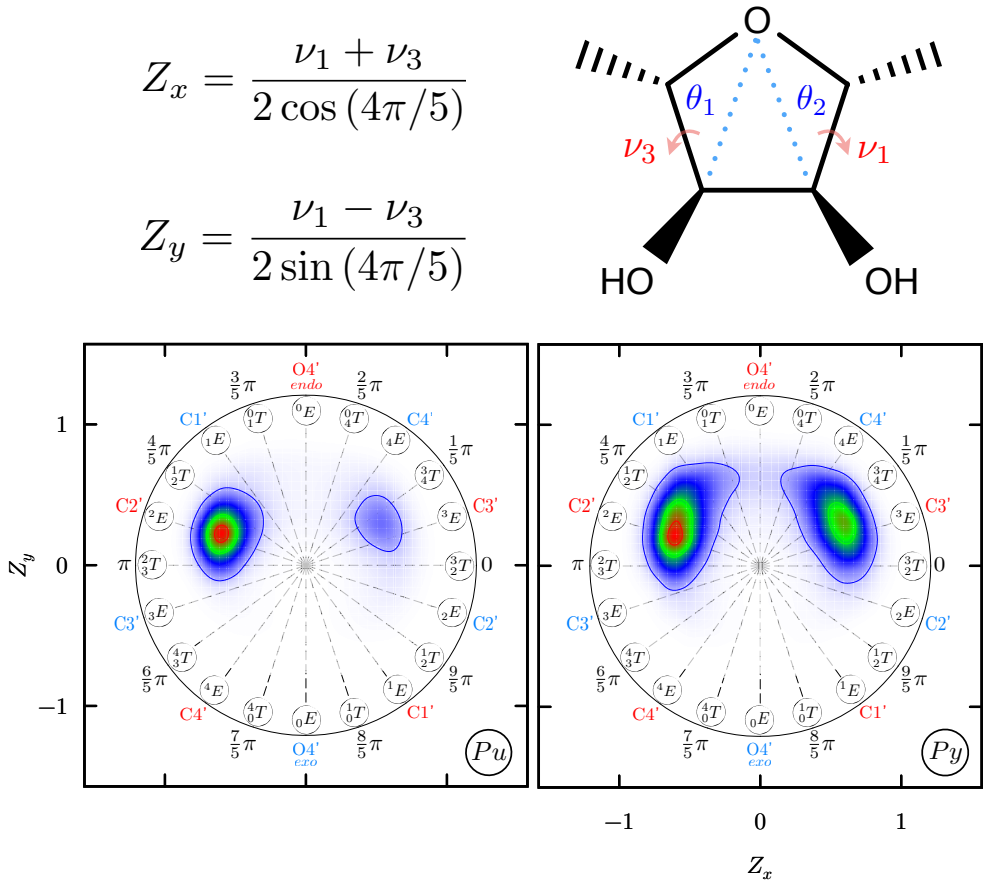


Figure 2.5: Pseudorotation wheel of the ribose sugar. Using the proper ν_1 and ν_3 dihedral angles, or equivalently the improper θ_1 and θ_2, a pair of Cartesian coordinates Z_x/Z_y can be defined to described the conformation of the furanose moiety [63]. The conformational ensemble of the Adenosine and Cytosine nucleosides generated with the ff99-bsc0-χ_{OL3} force field is projected onto the Z_x/Z_y space as an example of the effect of the nucleobase in the pseudorotation equilibrium. Adapted from ref [63].

crystals of mono-, oligo-, and polynucleotides, the torsion angle β defining rotation about the C5'-O5' bond is largely limited to the t range (Figure 2.6). The rotation about the C3'-O3' bond, denoted by ϵ, follows a similar trend yet the main clustering is not t but is shifted slightly to 220^0 in the *trans*$^-$ range ($-\frac{\pi}{2} \leftrightarrow \pi$) [79, 87, 89–92]. Theoretical considerations greatly agree with experimental data, showing that severe steric hindrance between the phosphate group and sugar moiety restricts C−O torsion angles β essentially to t and ϵ to the t and t^- ranges [93–99].

Rotations about P−O ester bonds, controlled by the α and ζ angles, are less restricted than rotations about C−O bonds, thus P−O bonds are the major pivots affecting polynucleotide structure. The stereoelectronic effects favors orientations about P−O ester bonds to (α/ζ) g/g, t/g or g/t. *Ab initio* calculations, using dimethyl phosphate anion as a model of the phosphodiester linkage, have determined that the g^-/t and t/t conformers are 1.45 and 3.66 kcal/mol higher in energy, respectively, than g^-/g^- (the energy profile is symmetric around the t position, so the g^+/g^+ is equally favored). This preference for *gauche* conformations is due to a stabilizing interaction caused by a lone pair located on O5' (or O3') that can partially donate charge to the

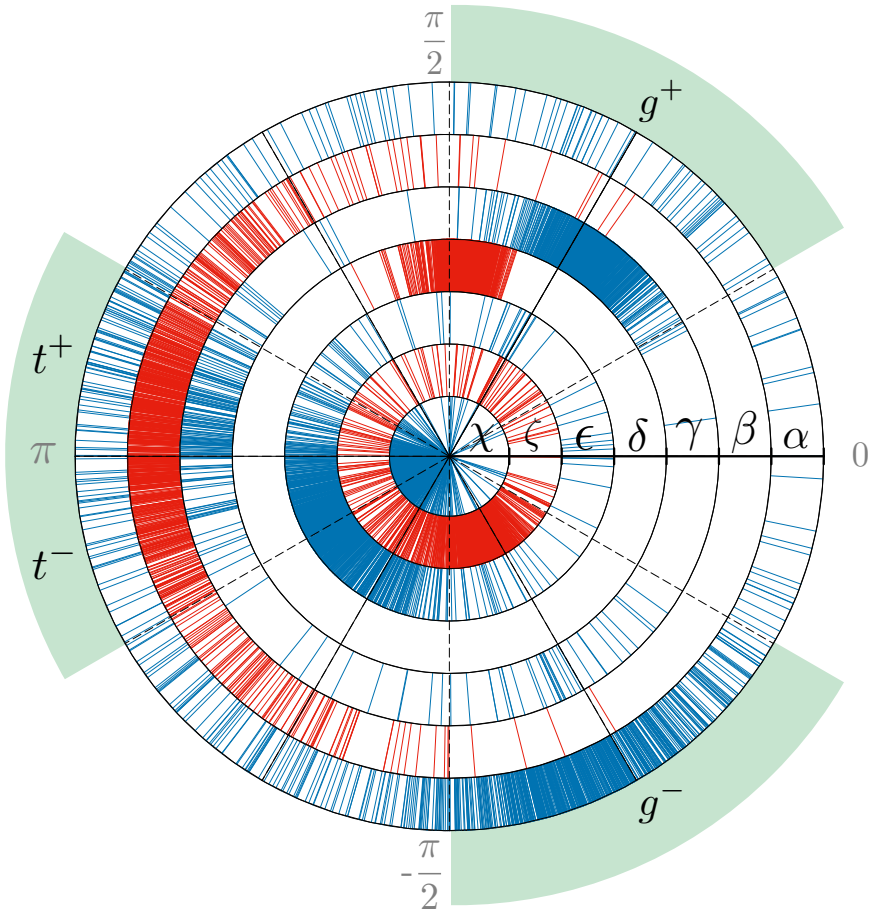


Figure 2.6: Conformational wheel showing the allowed ranges of backbone torsion angles. Values were taken from a X-ray structure of the large ribosomal unit from *D. Radiodurans* (PDB: 3JQ4).

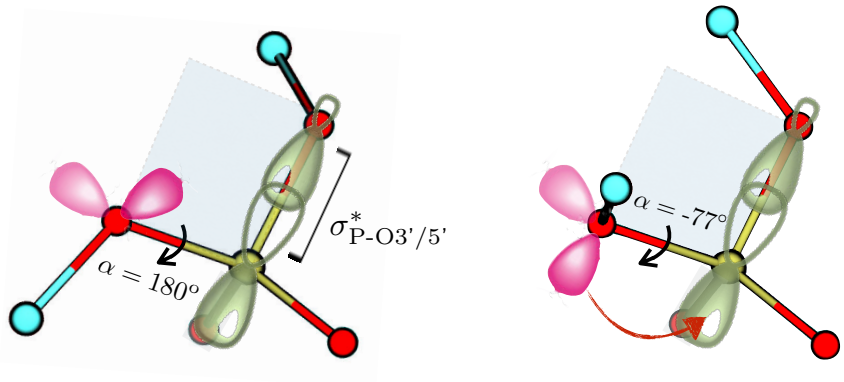


Figure 2.7: Description of the stereoelectronic effect in phosphodiester backbone of nucleic acids. The *gauche* conformation (left) of the C−O−P−O group is favorable because an oxygen electron lone pair is *trans* to the adjacent, polarized P−O bond and can donate electrons. In the *trans* orientation (right) orbitals and adjacent P−O bonds are in $g^{\pm}$ positions and electron transfer is diminished. Adapted from ref [42].

$\sigma^*_{P-O3'/5'}$ antibonding orbital. This type of interaction is illustrated in Figure 2.7. The inclusion of water and cations could affect the stabilization of the gauche conformation. Principally, the complexation with a cation can change the charge distribution of the phosphodiester group and decrease the stereoelectronic effect. Moreover, the symmetry between the g^-/g^- and g^+/g^+ conformations found in α/ζ energy maps calculated with dimethyl phosphates or dinucleoside monphosphates have been found to be broken when the chain is elongated, due to close contacts between second neighbor phosphate groups. In dinucleoside di- or triphosphates the g^+/g^+ orientations are less stable than the g^-/g^- conformations [42]. This is also supported by X-ray crystal studies, which show that the main geometrical arrangement of the α/ζ angles are right-handed helical conformations, with rotamers in the g^-/g^- range around 270^0 [42, 100]. In RNA structures the t rotamers of these angles are found usually in loops and turns motives, but disruption of the common g^-/g^- phosphate conformation is energetically costly (loss of the gauche effect is estimated at 2 kcal/mol) [101]. This energetic loss is partially offset by a hydrogen bond between the nucleobases atoms or sugar and an oxygen of the turning phosphate [101].

2.4 Summary

The unique structure-dynamic function relations in RNA are the result of a cooperative interplay among ribose sugar, nucleobase and phosphodiester moieties [75]. In this chapter we have briefly shown that the conformational equilibria of each of these structural motifs are driven by various internal effects, steric and stereoelectronic, and interactions with counterions and water. Moreover, these internal degrees of freedom are not uncorrelated, sugar conformation influences the orientation of the nucleobase, and vice versa, and conformational transitions are transmitted through the sugar-phosphate backbone to influence the rotameric preferences of the phosphodiester rotamers [74].

This structural complexity imposes a challenge for the interpretation of RNA structural experimental data [102] and for the molecular mechanics potentials used in computer simulations to describe RNA conformational space.

Chapter 3

RNA Simulations

In classical MD simulations, the potential energy is expressed in terms of bond length, angles between bonds, torsion angles, Lennard-Jones and Coulomb pairwise interactions. The energy function is known as force field (FF). There are a great variety of these FF, with slightly different functional forms, and each of them have different set of parameters which have been developed over the years. The most populars ones are AMBER [37], CHARMM [103], OPLS [104] and GROMOS [105]. The success of MD simulation is intrinsically dependent of the quality of these FF parameters, which describe the energetic landscape of the molecule. Despite the advances achieved in the last decades, with the increase of computational power and the refinement of RNA FFs, recent extensive simulations of unstructured oligonucleotides for which converged sampling is affordable have unambiguously shown that current RNA force-field parameters are not accurate enough to reproduce solution experiments. On the other hand, as new efforts in improving those force fields have been made it has been clear the importance of applying enhanced sampling methods to generate well-converged conformational ensembles which could be used to identify problems and validate those updates. The following sections present a discussion of the state of the art in RNA FF and enhanced sampling methods commonly used to aid the exploration of RNA conformational space.

3.1 RNA Force fields

Nucleic acids structures are complex and involve a subtle balance between charge interactions, hydrogen bonding, stacking contacts, backbone conformational flexibility, sugar puckers, and glycosidic torsions, all adding significant difficulty to the FF development. Though modest success has been seen with regard to reevaluating experimental data and qualitatively describing dynamics, the accuracy of RNA force fields is still lagging behind that of FF used in protein simulations [106] and even to the most recent DNA FF [107, 108]. The advantage of describing DNA over RNA could be due to the fact that DNA molecules are mostly stable double helix while RNA molecules are usually single-stranded that fold back upon themselves and show a very rich dy-

namic [109]. Moreover, the O2'H group of RNA ribose, absent in DNA, is a powerful donor and acceptor of hydrogen bonds that is involved in an astonishing repertoire of non-Watson−Crick interactions.

Current atomistic simulations of nucleic acids are still mostly based on second-generation pair-additive force fields derived in the 90s [15, 16, 27]. There have been efforts to improve their performance by partial reparametrizations [26, 38, 39, 54]. Though attempts at improving RNA FF have been dominated by modifications of dihedral parameters [38, 39, 54], recent parameters have addressed also the various problems with the non-bonded components [26, 110]. The most used FF for RNA are descend from the AMBER [37] and CHARMM family [111–113]. For clarity FF refinements have been divided into 2 groups, non-bonded corrections and the changes in the dihedral parameters.

3.1.1 Improvement of dihedral angle rotations

The latest CHARMM force field for RNA is the CHARMM36 [114], where the dihedral parameters of the O2'H group were tuned to improved the agreement with quantum-mechanical energy profiles. Recent test have shown that CHARMM36 suffers from some understabilization of canonical A-RNA helices on the nanosecond time scale [115, 116]. In contrast to the good performance of CHARMM force field in the proteins and B-DNA simulations, FF of the AMBER family are usually preferred for long RNA simulations. Many variants of the AMBER Cornell *et al.* force field [37] are currently in use by the RNA community. Most of them are based on a combination of AMBER99 (ff99) [117] and the bsc0 reparametrization of the α/γ dihedral pair [38], which eliminated spurious flips resulting in a progressive degradation of B-DNA structure during simulations. Although bsc0 was devised for DNA, subsequent simulations have proved that it also improves the RNA description [115, 118]. In 2010, a reparametrization of the glycosidic torsion, denoted as χ_{OL3} [39, 118], was developed to correct the formation of ladder-like structures in the microsecond time scale in RNA canonical A-form helices [119]. The ff99-bsc0-χ_{OL3} variant have been tested on many RNA systems including RNA helices, tetranucleotides and tetraloops [31–33, 115, 120, 121]. The experimental agreement of the ensemble generated with this variant is far from robust, fundamentally in the prediction of the native conformation stability in single-stranded structures and hairpins. Another reparametrization of the χ torsion was independently suggested by I. Yildirim *et al.*, using NMR data of nucleosides to validate the new FF [54]. The ff99-bsc0-χ_{YIL} force field also prevents the spurious ladder-like structures but causes some flattening of the A-form helix by underestimation of inclination and roll parameters [118]. Subsequently, I. Yildirim *et al.* extended their version by including a reparametrization of the ε, ζ, and β torsions [122], but latter tests have shown it caused canonical A-form helices to deteriorate [123].

3.1.2 Modifications of Non-bonded interactions

Though modifications of dihedral parameters have dominated the AMBER corrections, recent works have suggested that changes in non-bonded parameters improve key problems of the ff99 like over-stacking and the imbalance in solute-solvent interactions. D. Case *et al.* reparametrized the van der Waals (vdW) radii of the oxygens in the phosphate group to obtain consistent thermodynamic results with better balanced electrostatic interactions between water and the phosphate oxygens [110]. REMD simulations of tetranucleotides [33] and tetraloops [27] showed that RNA simulations might be improved by the implementation of this new parameters.

An alternative reparametrization of ff99 has been suggested by A. Chen and A. Garcia (ff99-vdW$_{base}$-χ_{CG}) during their efforts to achieve a reversible folding of RNA tetraloops [26]. They proposed a rescaling of the Lennard-Jones potential parameters from the nucleobase heavy atoms, together with a modification of off-diagonal Lennard-Jones terms in nucleobase–water interactions. The non-bonded corrections were accompanied by an adjustment of the ff99 χ dihedral potential. The main aim of the new FF was to eliminate the known overestimation of the stacking interactions [124, 125]. With these parameters it was possible to observe multiple folding events to the folded state with correct signature interactions for two out of three studied RNA tetraloops. A subsequent benchmark study by Cheatham *et al.* confirmed that ff99-vdW$_{base}$-χ_{CG} leads to an improvement over all the current AMBER FF variants. However, ff99-vdW$_{base}$-χ_{CG} may also lead to excessive stabilization of some non-native base-pairs and to an imbalance between modified and unmodified vdW parameters, underlining the difficulty of obtaining a force field that would simultaneously reproduce all properties of RNA molecules [33, 126].

The effect of different water models on the experimental agreement of the RNA ensembles have also been tested [115, 127, 128]. For example, in the simulation of single-stranded RNA oligonucleotides the combination of AMBER parameters and the OPC water model have shown a significant improvement over the commonly used TIP3P water model [128]. However, variations of water models are unlikely to resolve the limited accuracy of the primary solute force field, which originates from its inability to reliably describe inherent conformational preference of nucleic acids.

3.2 RNA Sampling

Molecular dynamics (MD) with accurate force fields can in principle be used as a virtual microscope to investigate motions at atomistic resolution [129]. However, its applicability to problems such as folding or conformational transitions in proteins and RNA is limited by the fact that only short time scales ($\sim \mu$s) are directly accessible by straightforward simulation. Although recently developed *ad hoc* hardware allowed for a three-order-of-magnitude gain in the accessible time scales [106], many relevant conformational transitions are still out of reach for accurate atomistic modeling. Several

different techniques have been developed in the last decades to address this issue [130]. These techniques can be roughly classified in two groups: methods based on Annealing [131] and techniques based on Importance Sampling [132]. The next sections will introduce some basic concepts of these computational techniques.

3.2.1 Annealing-based methods

These class of methods were traditionally based on increasing the temperature of a MD simulation to overcome high energy barriers [130]. This strategy relies on the fact that in an Arrhenius process the logarithm of the relevant performance parameter (*e.g.* the rate at which the barrier-crossing events happen) depends linearly on the reciprocal of the temperature [133, 134]. Thus, for this kind of processes, a procedure where the system is first heated and then cooled allows the quick generation of samples which are largely uncorrelated. The annealing strategy was first translated from metallurgy to combinatorial optimization in the seminal work of Kirkpatrick et al. [131], later extended to spin glasses [135] and finally to biological systems [136].

The annealing procedure in molecular dynamics simulations has been implemented mainly in two ways: Simulated annealing [137] and Parallel tempering [135, 136]. In a typical parallel tempering simulation (also known as temperature replica exchange, T-REMD) there is a ladder of replicas, each at a different temperature T_i. Across the replica ladder the temperatures increase progressively, the lowest replica is simulated normally at a room temperature, and the highest (replica) temperature is chosen so the system can easily cross barriers between minima. The coordinates of the replicas are periodically exchanged between the ensembles and the velocities are appropriately rescaled to the new temperature. Since the replicas do not interact, the partition function of this larger (generalized) ensemble is given by the product of the individual partition function of each (replica) ensemble. If the probability of attempting a swap move (α) is equal for all conditions, exchanges between ensembles i and j are accepted with the probability

$$\alpha = \min\left(1, e^{\Delta_{i,j}}\right) \tag{3.1}$$

with $\Delta_{i,j} = (\beta_j - \beta_i)(E_j - E_i)$, where β is the reciprocal temperature $\frac{1}{T}$ and E_i is the potential energy of the system i (to simplify $k_B = 1$). If we assume the systems have Gaussian energy distributions, with mean $\langle U(\beta)\rangle$ and width $\sigma(\beta)$, the probability distribution of Δ will be also a Gaussian with mean $\Delta_0 = (\beta_j - \beta_i)(\langle E_j\rangle - \langle E_i\rangle)$ and width $\sigma_0^2 = (\beta_j - \beta_i)^2(\sigma_j^2 + \sigma_i^2)$. The min function can be evaluated analytically to obtain the average acceptance ratio $\langle\alpha\rangle$ [138], which is equal to

$$\langle\alpha\rangle = \operatorname{erfc}\left(\frac{\Delta_0}{\sigma_0}\right). \tag{3.2}$$

Equation 3.2 can be used to estimate the acceptance between two replicas given either the average Δ_0 and the variance σ of the Δ term. Assuming that the heat

capacity of the system C_V is constant, then $\langle E_j \rangle - \langle E_i \rangle = C_V(T_j - T_i)$ and $\sigma_j^2 + \sigma_i^2 = C_V(T_j^2 + T_i^2)$ substituting both expressions in equation 3.2 leads to [139]

$$\langle \alpha \rangle = \operatorname{erfc}\left(\sqrt{C_V}\frac{T_j - T_i}{\sqrt{T_i^2 + T_j^2}}\right). \tag{3.3}$$

From this expression it can be seen that in systems where the C_V is constant the density of replicas should decrease as the temperature raises to maintain an uniform acceptance: for a fixed replica spacing (constant $\Delta T_{i,j}$) if the magnitude of the temperatures rises ($\uparrow \sqrt{T_j^2 + T_i^2}$) the $\langle \alpha \rangle$ increases. In this case, a geometric distribution of temperatures (constant $\frac{T_j}{T_i}$) has been found to be optimal to maintain an uniform α across the replica ladder.

Another factor influencing α is the size of the system. The heat capacity is proportional to the number of particles N (or degrees of freedom) which means the acceptance will decrease when N increases [140]. The parallel tempering simulation of large biomolecular systems in water (hundreds of thousand of atoms) is computationally demanding, as it requires a large number of replicas to maintain a moderate acceptance in a range of temperatures that allows the system to ensure transitions over high energy barriers.

In practical cases the specific heat is not a constant, and especially for biomolecular systems in vacuum or implicit solvent, the C_V can change significantly with the temperature. In these cases a geometric distribution of temperatures will not generate a constant acceptance. On the other hand, in simulations of solvated molecules, the acceptance ratio will be dominated by the specific heat of the water model, for which the approximation of constant C_V is more plausible. The selection of an optimal distribution of temperatures is not trivial, but solutions to take into account a more realistic dependence of the heat capacity with temperature in the context of explicit solvent simulations are available [141].

Whereas parallel tempering is a powerful method and it has been applied to practically all biochemical systems with great results, temperature is an intensive quantity and does not allow the selective enhancement of specific degrees of freedom. The method is also ineffective on entropic barriers and in systems with anti-Arrhenius behavior [142–144]. Scaling portions of the Hamiltonian is a common alternative (H-REMD) and could have a better convergence behavior for large systems. A promising technique in this group is replica-exchange with solute tempering, where solute-solvent interactions and the force-field parameters of the solute are modified [145, 146]. T-REMD and H-REMD can also be combined, by integrating both schemes on each replica [147] or in a multidimensional framework [31]. In particular for RNA systems, the multidimensional replica exchange [31] have outperformed one-dimensional T-REMD and H-REMD simulations, for the same conditions and total simulated time [32]. Even for a small RNA system like a tetranucleotide, in order to generate a converged ensemble, a total of 57.6 μs of simulated time in a multidimensional replica exchange framework (24×8 replicas) have shown to be required [31].

3.2.2 Importance-sampling based methods

The second group of enhanced sampling techniques includes methods based on importance sampling. This class has its root in the umbrella sampling method [132], and includes local elevation [148], conformational flooding [149], adaptive biasing force [150], and metadynamics [40, 151], among others. In this kind of methods the canonical Boltzmann weighting is modified by a bias potential designed to cancel the effect of free-energy barriers and increase the frequency of rarely-sampled conformations. The potential is usually defined in a *reduced* set of coordinates, known as collective variables (CVs). These techniques are very effective but require a careful choice of the CVs that must provide a satisfactory description of the reaction coordinate [152, 153]. If important degrees of freedom are not taken into account, it could hinder the exploration of the phase space and generate hysteresis and lack of convergence. Moreover, when more than a few (~ 3) CVs are used, the computational performance rapidly degrades as a function of the number of variables. For many biomolecular systems it is difficult to find a small number of effective CVs that describe all the slow degrees of freedom.

Consider a function s of the system coordinates $s(x_1, x_2.., x_N)$ that allow the projection of the system conformational space in a *reduced* surface and includes some important features of the system dynamic and phase space. For example the minima in the *reduced* space should correspond to the metastable states of the system, and the relevant transition events should be represented there by matching barriers. The probability distribution of the CV is given by

$$P(s) = \frac{1}{Z} \int e^{-\beta U(x)} \delta(s - s(x)) dx \tag{3.4}$$

where Z is the partition function $Z = \int e^{-\beta U(x)} dx$ and the corresponding free energy is estimated as

$$F(s) = -\frac{1}{\beta} \ln \int e^{-\beta U(x)} \delta(s - s(x)) dx. \tag{3.5}$$

In the umbrella sampling framework the normal dynamics of the system is biased by a smartly chosen bias potential $V(s(x))$ that depends on x only via $s(x)$. The bias potential facilitates the exploration of the system conformational space and so the biased probability distribution $\tilde{P}$ will be *easier* to estimate

$$\tilde{P}(s) = \frac{1}{Q} \int e^{-\beta(U(x)+V(s(x))} \delta(s - s(x)) dx \tag{3.6}$$

here Q is the partition function of the biased ensemble. The effect of the bias potential can be reweighed to obtain the unbiased probability distribution

$$P(s) = \tilde{P}(s) e^{\beta(V(s)-f)} \tag{3.7}$$

where $f = \frac{1}{\beta} \ln \frac{Z}{Q}$ is a constant that does not depend on s. Equation 3.7 is the fundamental relation behind the umbrella sampling and related methods. These methods are

very efficient but require large *a priori* information, in order to define a proper CV and choose an efficient bias potential. One solution for the latter problem is the adaptive construction of the bias potential during the MD or MC simulations using kernel functions like Gaussians or splines. In this thesis we focus on well-tempered metadynamics (WT-MetaD), a self-consistent adaptive-bias method introduced in 2008 by A. Barducci, G. Bussi and M. Parrinello [40], which is a variant of the original metadynamics method devised by A. Laio and M. Parrinello in 2002 [151].

3.2.2.1 Well-tempered metadynamics

In well-tempered metadynamics a history dependent potential $V(s,t)$ acting on the collective variable s is introduced and evolved according to the following equation of motion

$$\dot{V}(s,t) = \frac{k_B \Delta T}{\tau_B} e^{-\frac{V(s,t)}{k_B \Delta T}} K(s - s(t)) \tag{3.8}$$

here k_B is the Boltzmann constant, T the temperature, τ_B is the characteristic time for the bias evolution, ΔT is a boosting temperature, and K is a kernel function which is usually defined as a Gaussian. For simplicity we consider the case of a single CV. The variance of the Gaussian provides the binning in CV space and is usually chosen based on CV fluctuations or adjusted on the fly [154]. By assuming that the bias is growing uniformly with time one can show rigorously [40, 155] that in the long time limit the bias potential tends to

$$\lim_{t\to\infty} V(s,t) = -\frac{\Delta T}{T + \Delta T} F(s) + C(t) \tag{3.9}$$

so that the following probability distribution is sampled

$$\lim_{t\to\infty} P(s,t) \propto e^{-\frac{F(s)}{k_B (T+\Delta T)}}. \tag{3.10}$$

The role of ΔT is that of setting the effective temperature for the CV. The explored conformations are thus taken from an ensemble where that CV only is kept at an artificially high temperature, similarly to other methods [156–158], but has the nice feature that it is obtained with a bias that is quasi-static in the long lime limit. The bias is usually grown by adding a Gaussian every N_G steps. As a consequence, to obtain an initial growing rate equal to $\frac{k_B \Delta T}{\tau_B}$, the initial Gaussian height should be chosen equal to $\frac{k_B \Delta T}{\tau_B} N_G \Delta t$ where Δt is the MD time step.

3.2.3 *E pluribus unum*

The advantage of replica exchange methods is that they generally require very little *a priori* knowledge of the system, as opposed to the methods based on importance-sampling. However, the former methods can be rather computationally expensive, especially parallel tempering, and the expansion of the generalized ensemble could

lead to problems of convergence. The combination of the replica exchange framework with umbrella-sampling-type methods like Metadynamics could solve this problems and reduce the own limitations of the importance sampling strategy. Such a combination results in a synergic effect. Parallel tempering metadynamics [159], bias-exchange metadynamics [160] and the well-tempered ensemble [161, 162] are great examples of the integration between these two frameworks.

In the case of nucleic acid systems, compare to proteins, the application of these methods have been limited, partly due to the difficulties of designing *ad hoc* CVs which can correctly describe the conformational transitions. Some applications have circumvented this problem by biasing a large number of local CVs (e.g. dihedral angles). For example, J. Curuksu and M. Zacharias introduced a technique where bias potentials acting on dihedrals were used in a replica exchange framework to specifically promote dihedral transitions in the nucleic acid backbone [163]. The dihedral angle conformational space is discretized, in one or two-dimensions, to identify the position of the metastable basins. For example, in a bidimensional ϵ/ζ space once a minimum is located (ϵ_i, ζ_i) a bias potentials $V(x)$ is settled, where x is the angular distance $\sqrt{(\epsilon - \epsilon_i)^2 + (\zeta - \zeta_i)^2}$. This function is constant when $x \leq r$, and at distances larger than r it decreases continuously on its edges down to zero at a distance R.

$$\begin{aligned} V(r) &= E_{\max} && (x < r) \\ &= \tfrac{E_{\max}}{(r-R)^4}[(x-r)^2 - (r-R)^2]^2 && (r \leq x \leq R) \\ &= 0 && (x > R) \end{aligned}$$

After all the relevant basins in the dihedral angle space are determined a replica exchange simulation is run where the height ($E_{\max}$) of the bias potential is increased along the replica ladder. This technique require very few replicas and it has been upgraded to include a dynamic adjustment of the bias potential height during the simulation to ensure high acceptance rates and a good mixing of sampled structures in the replicas [164]. However, these potentials do not account for the specific identity of each residue and for the cross-talk between correlated dihedrals.

Another approach have been attempted by Roe *et al.* who combined accelerated molecular dynamics (aMD) with replica exchange method to explore the conformational space of a RNA tetranucleotide [32]. In aMD a boosting potential is applied to the torsion energy $E(r)$ when its values drop below a user specific energy cutoff E_{cutoff} [165]. This boosting potential is a function of the torsion energy itself

$$V_{boost}(r) = \frac{(E_{cutoff} - E(r))^2}{\alpha + (E_{cutoff} - E(r))}$$

and across the replica ladder its strength is incremented by reducing the value of α. In this application, the boosting potentials were not able to compensate the free-energy barriers to rotation around many of the biased torsion angles, specially ϵ and χ angles. The reason is that barriers to rotations do not include just steric and electrostatic contributions from the dihedral atoms encoded in the torsion energy, but also long-distance non-bonded contacts and solute-water interactions.

3.3 Summary

In the first section of this chapter we present a survey of the different force fields available for RNA. When compared with solution NMR experiments of RNA single-stranded tetranucleotides none of the ensembles generated with the many AMBER FF variants have given a satisfactory agreement. This could be caused by the inability of pair-additive FFs to accurate reproduce RNA structural features, due to physical approximations, like not considering explicitly polarization effects, or could be the result of simplifications taken during the FF parametrization process, like assuming the sugar-backbone angles are uncorrelated. One solution to this problem could be the inclusion of correcting potentials into the FF, tuned in order to increase the agreement with solution experiments.

In section 3.2, a short introduction to the field of enhanced sampling techniques is given. An especial attention is taken on *ad hoc* sampling methods developed for the simulation of nucleic acid systems. These methods have shown that acceleration of RNA conformational transitions can be achieved effectively by biasing dihedral angles or the dihedral energy. However, the method presented by J. Curuksu and M. Zacharias [163] was only applied to a pair of dihedral angles in DNA backbone and in the second method [32] discussed here only the energetic contributions to the rotameric equilibrium of RNA dihedrals were considered, which led to residual free-energy barriers that hindered some of the torsional angles rotations. In the spirit of these previous methods, a new technique is introduced in the next chapter, that combines concurrent well-tempered metadynamics simulations with replica exchange.

Chapter 4

Replica Exchange with Collective-Variable Tempering

4.1 Overview

As discussed in Chapter 3, section 3.2.3, the combination of the replica exchange framework with importance sampling techniques that biased a large number of local collective variables (*e.g.* dihedral angles), have been employed effectively to promote conformational transitions in nucleic acids. In the present chapter, a new methodology is presented, which uses concurrent well-tempered metadynamics simulations [40] (WT-MetaD) to build bias potentials acting on a large number of local CVs. We then show how to integrate this approach in a Hamiltonian replica exchange (H-REMD) scheme, exploiting the replica ladder to obtain unbiased conformations. In WT-MetaD the compensation of the underlying free-energy landscape is modulated by the boosting temperature ΔT. We here change this parameter across the replica ladder, adjusting the ergodicity of each replica. The final bias can be also used as a static potential so as to completely eliminate any non-equilibrium effect. Since the effect of the bias is that of keeping the chosen CVs at an effectively higher temperature, we refer to the introduced method as replica exchange with collective-variable tempering (RECT). The method is first tested on alanine dipeptide in water and then applied to the conformational sampling of a RNA tetranucleotide where it outperforms dihedral-scaling REMD and plain MD. The chosen tetranucleotide is a very challenging system that has been extensively studied with long MD simulations and different variants of REMD [31–34, 120, 128, 166].

4.2 Methods

In this Section we show how to use WT-MetaD as an effective method to build concurrent bias potentials that allow barriers to be easily crossed. One of the input parameters of well-tempered metadynamics is a boosting temperature $\Delta T = (\gamma - 1)\, T$, where γ is the bias factor and T is the temperature of the system. In the rest of the chapter

we will equivalently use either γ or ΔT so as to simplify the notation. This parameter can be used to smoothly interpolate between unbiased sampling ($\gamma = 1$, $\Delta T = 0$) and flat histogram ($\gamma = \infty$, $\Delta T \to \infty$). One can thus introduce a set of replicas using different values of ΔT, ranging from 0 to a value large enough to allow all the relevant barriers to be crossed. Metadynamics relies on the accumulation of a history dependent potential and cannot be applied straightforwardly to a large number of CVs. In the next subsection we show that this issue can be circumvented by constructing many, low-dimensional, concurrent metadynamics potentials. We then show how to combine many simulations of this kind in a multiple-replica scheme.

4.2.1 Concurrent Well-tempered Metadynamics

We here propose to introduce a separate history-dependent potential on each CV

$$\dot{V}_\alpha(s_\alpha) = \frac{k_B \Delta T}{\tau_B} e^{-\frac{V_\alpha(s_\alpha,t)}{k_B \Delta T}} K(s_\alpha - s_\alpha(t)) \tag{4.1}$$

where $\alpha = 1, \ldots, N_{CV}$ is the index of the CV and N_{CV} is the number of CVs. The growth of each of these bias potentials will depend only on the marginal probability for each CV

$$P(s_\alpha) \propto \int ds_1 ds_2 \ldots ds_{\alpha-1} ds_{\alpha+1} \ldots ds_{N_{CV}} P(s_1, s_2, \ldots, s_{N_{CV}}) \tag{4.2}$$

In the long time limit, this potential will tend to flatten the marginal probabilities for every single CV. In the general case one should consider the fact that whenever a bias is added on a CV also the distribution of the other CVs is affected. In the following we will discuss this issue considering two CVs only, but the argument is straightforwardly generalized to a larger number of CVs.

Two independent variables. If two CVs are independent, the joint probability is just the product of the two marginal probabilities, i.e. $P(s_\alpha, s_\beta) = P_\alpha(s_\alpha) P_\beta(s_\beta)$. Adding a bias potential on a CV will not affect the distribution of the other. As a consequence, in the long time limit the two bias potentials will converge independently to the predicted fraction of the free energy as in Eq. 3.9. The final bias potential will be completely equivalent to that obtained from a two-dimensional well-tempered metadynamics, but will only need the accumulation of two one-dimensional histograms, thus requiring a fraction of the time to converge. A simple example on a model potential is shown in Fig. 4.1.

Two identical variables. We also consider the case of two identical CVs, $s_\alpha = s_\beta$. This can be obtained for instance by biasing twice the same torsional angle. Here the potentials V_α and V_β will grow identically, and the total bias potential acting on s_α will be $V_{tot} = 2V_\alpha$. The total potential will grow as

$$\dot{V}_{tot}(s_\alpha) = \frac{2k_B \Delta T}{\tau_B} e^{-\frac{V_\alpha(s_\alpha,t)}{k_B \Delta T}} K(s_\alpha - s_\alpha(t)) = \frac{2k_B \Delta T}{\tau_B} e^{-\frac{V_{tot}(s_\alpha,t)}{2k_B \Delta T}} K(s_\alpha - s_\alpha(t)) \tag{4.3}$$

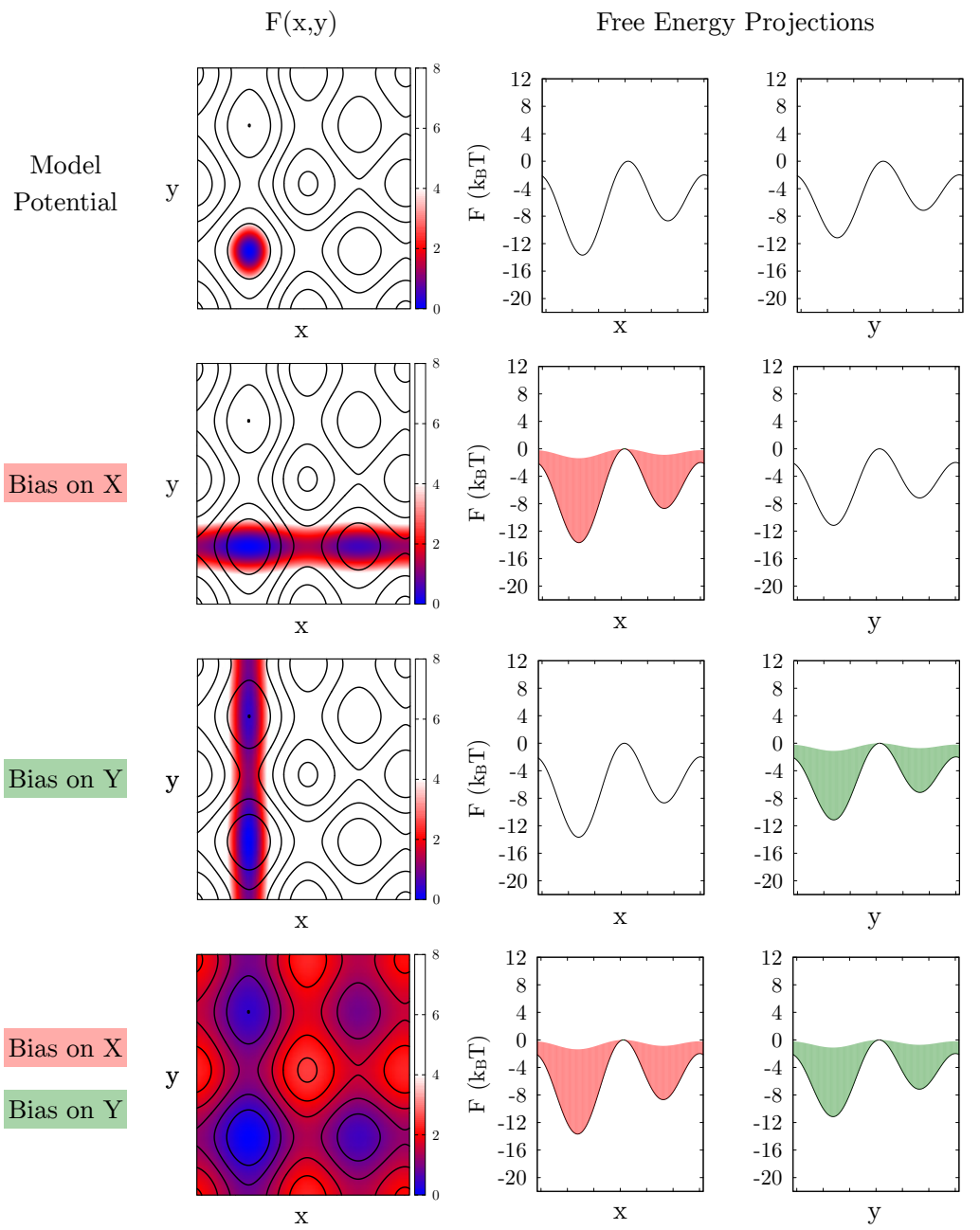


Figure 4.1: A model two-dimensional energy potential where the two variables are independent (left panel). Isolines are spaced by $4k_BT$. The free-energy space accessible to the system at $4k_BT$ is colored according to the canonical probability of each region. Projections on x and y variables are also represented (right panels). Self-consistent bias potentials generated by WTMetaD are shown acting on x (red) and y (green). The potentials calculated with concurrent WTMetaD (fourth row) are identical to the bias potentials produced during the WTMetaD simulation of each variable independently (second and third rows).

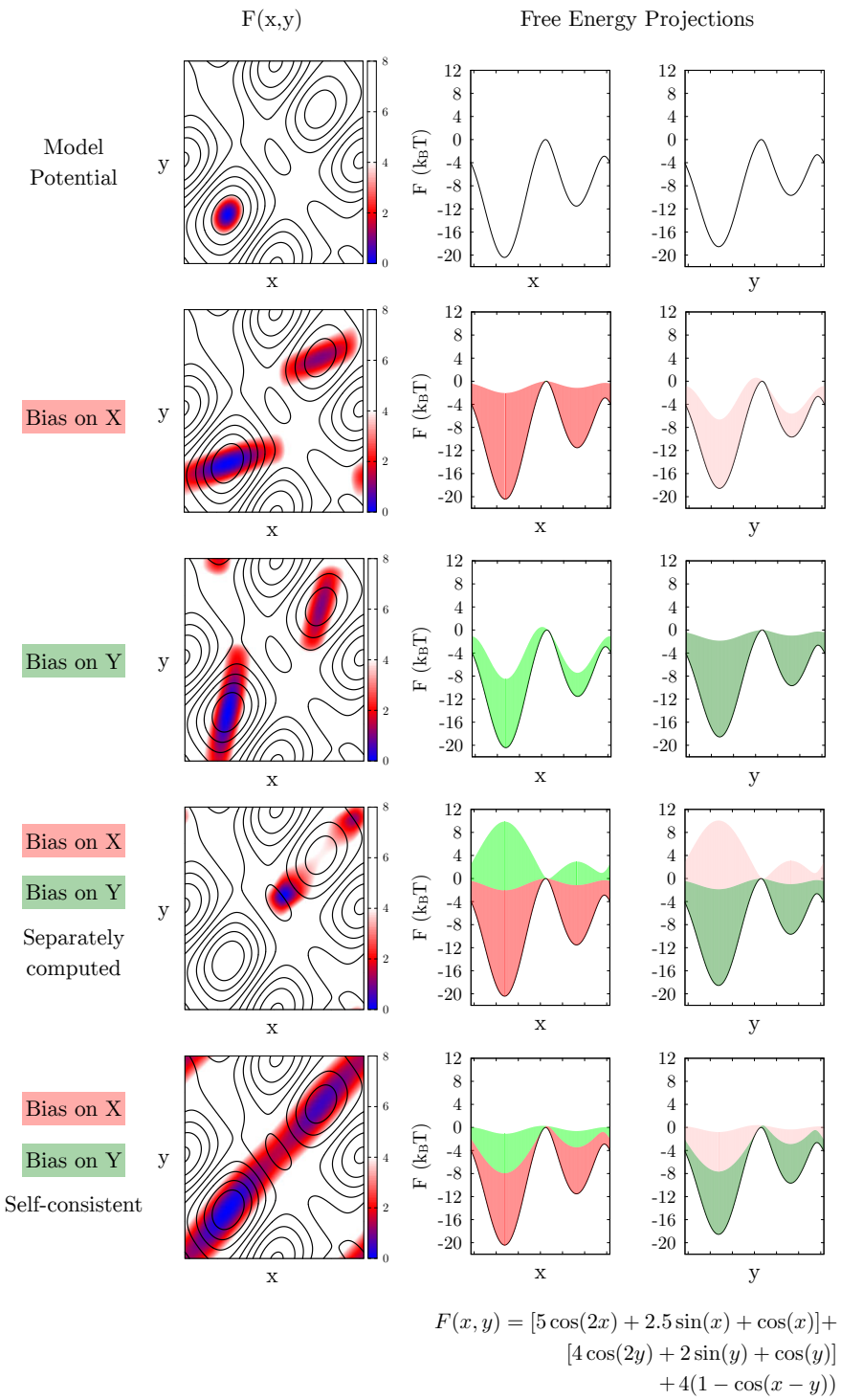


Figure 4.2: A model two-dimensional energy potential where the two variables are correlated. Bias potentials calculated with WTMetaD for one of the two variables (x or y) separately successfully compensate projected free-energy barriers on the respective variable. However, this one-dimensional bias potential has a side effect on the distribution of the other variable due to their correlation. This generates an additional effective bias potential that is shown in light color (second and third rows). When the two one-dimensional bias potentials are applied simultaneously, the action of the WTMetaD potential is superimposed to the action of the effective bias potential. As a result, the space sampled by each variable is greatly restricted (fourth row). The self-consistent construction of the two one-dimensional bias potentials by concurrent WTMetaD eliminates this effect, generating bias potentials capable to flatten the marginal probabilities when applied at the same time (fifth row).

Thus, the net effect will be exactly equivalent to that of choosing a doubled ΔT parameter. In other words, the ΔT parameter acts in an additive way on the selected CVs. A similar effect can be expected if two CVs are linearly correlated.

In realistic applications one can expect the behavior to be somewhere in the middle between these two limiting cases. The most important consideration here is that the bias potentials will tend to flatten all the marginal probabilities, but there will be no guarantee that the joint probability is flattened. Results for a simple functional form can be seen in Appendix A, Fig. 4.1 and 4.2. In figure 4.2 it is possible to appreciate importance of using a self-consistent procedure when CVs are correlated. In ref [162] two metadynamics were applied on top of each other, namely on the potential energy and on selected CVs, in a non self-consistent way. This was possible because the correlation between the potential energy and the selected CVs is small. The need for a self-consistent solution was also pointed out in a recent paper [167] where a generalization of the adaptive biasing force method (ABF) [150] was introduced. In that work independent one-dimensional adaptive forces were applied at the same time to different CVs so as to enhance the sampling of a high multidimensional space.

In short, the novelty of the introduced procedure is that many low-dimensional metadynamics potentials are grown instead of a single multi-dimensional one. This allows the bias to converge very quickly to a flattening potential, with the degree of flatness controlled by the parameter ΔT. The flattening is expected to enhance conformational transitions which are otherwise hindered by free-energy barriers on the biased CVs. When variables are correlated the exact relationship between bias and free energy (Eq. 3.9) could be lost.

4.2.2 Hamiltonian Replica Exchange

The procedure introduced above produces conformations in an ensemble which is in general difficult to predict. However, since the bias potential is known, one can in principle reweight results so as to extract conformations in the canonical ensemble. In the case of static bias potentials acting on the CVs, this can be done by weighting each frame as $e^{\frac{\sum_\alpha V_\alpha(s_\alpha)}{k_B T}}$. This can provide in principle correct results even if the joint probability is not flattened. It must be noticed that such a reweighting can provide statistically meaningful results only for small fluctuations of the total biasing potentials, on the order of $k_B T$ [167]. However, in a typical setup one would be interested in biasing all the torsional angles of a molecule. Even if each of them contributes with a few $k_B T$, the total fluctuation of the bias would grow with the system size. For similar reasons, also the ABF-based scheme introduced in ref [167] is limited to a relatively low number of CVs.

A more robust and scalable procedure can be designed by introducing a ladder of replicas with increasing values of ΔT, ranging from 0 to a value large enough to enhance the relevant conformational transitions. The first replica ($\gamma = 1, \Delta T = 0$) can be used to accumulate unbiased statistics. Replicas other than the first one feel

multiple biasing potentials on all CVs. From time to time an exchange of coordinates between neighboring replicas is proposed and accepted with probability (α) chosen so as to enforce detailed balance with respect to the current biasing potential:

$$\alpha = \min\left(1, e^{\Delta}\right) \tag{4.4}$$

$$\Delta = \frac{\sum_\alpha V_\alpha^{(i)}(s_\alpha^{(i)}) + \sum_\alpha V_\alpha^{(j)}(s_\alpha^{(j)})}{k_B T} - \frac{\sum_\alpha V_\alpha^{(i)}(s_\alpha^{(j)}) + \sum_\alpha V_\alpha^{(j)}(s_\alpha^{(i)})}{k_B T} \tag{4.5}$$

Here the suffix $i = 1, \ldots, N_{rep}$ indicates the replica index, N_{rep} being the number of replicas. The exchanges allow the bias potential of every single replica to grow as close as possible to equilibrium taking advantage of the enhanced ergodicity of the more biased replicas. We notice that to reach a quasi-static distribution it is necessary that all the bias potentials converge for all the replicas. Since the time scale for convergence is related to the parameter τ_B [40], it is convenient to use the same τ_B for all the replicas or, equivalently, to choose the initial deposition rate as proportional to ΔT. The number of replicas required to span a given range in the ΔT parameter is proportional to $\sqrt{N_{CV}}$.

We notice that in principle one could use the bias potentials built with this protocol to perform a replica-exchange umbrella sampling simulation. In this manner the final production run would be performed with an equilibrium replica exchange simulation. However, we observe that well-tempered metadynamics is designed so that the speed at which the bias grows decreases with time and the potential becomes quasi-static. In the practical cases we investigated, this second stage was not necessary.

4.2.3 Model systems

4.2.3.1 Alanine dipeptide

Alanine dipeptide (dALA) was modeled with the AMBER99SB-ILDN [168, 169] force field and solvated in an truncated octahedron box containing 599 TIP3P [170] water molecules. The LINCS [171, 172] algorithm was used to constrain all bonds and equations of motion were integrated with a timestep of 2 fs. For each replica the system temperature was kept at 300 K by the stochastic velocity rescaling thermostat [173]. For all non-bonded interactions the direct space cutoff was set to 0.8 nm and the electrostatic long-range interactions were treated using the default particle-mesh Ewald [21] settings. All the simulations were run using GROMACS 4.6.5 [174] patched with the PLUMED plugin [175], version 2.0. We underline that the possibility of running concurrent metadynamics within the same replica is a novelty introduced in PLUMED 2.0.

The RECT simulation was performed with 6 replicas. The backbones dihedral angles (Ψ and Φ) and the gyration radius (R_g) were selected as CVs. The γ factors were chosen from 1 to 15 following a geometric distribution. We recall that a geometric replica distribution is optimal for constant specific-heat systems. In RECT, this would

be true if the exploration of each of the biased CV were limited to a quasi-parabolic minimum in the free-energy landscape. Whereas this is clearly not true in real cases (e.g. double-well landscapes) we found that a geometric schedule was leading to a reasonable acceptance in the cases investigated here. The possibility of optimizing the replica ladder is left as a subject for further investigation. For the dihedral angles the Gaussian width was set to 0.35 rad and for the R_g to 0.007 nm. The Gaussians were deposited every 500 steps. The initial Gaussian height was adjusted to the ΔT of each replica, according to the relation $h = \frac{k_B \Delta T}{\tau_B} N_G \Delta t$, in order to maintain the same $\tau_B = 12$ ps across the entire replica ladder. The CVs were monitored every 100 steps, and exchanges were attempted with the same frequency. The simulation was run for 20 ns per replica.

A H-REMD simulation where the force-field dihedral terms were scaled (H_{dih}-REMD) was also performed, as implemented in an in-house version of the GROMACS code [176]. The same initial structures, number of replicas and simulated time as in RECT was used. The scaling factor λ for each replica was selected using the relation $\lambda = 1/\gamma$ to allow for a fair comparison of RECT and H-REMD. Finally a conventional MD simulation in the NVT ensemble was run for 120 ns using the same settings.

4.2.3.2 Tetranucleotide

The second system considered was an RNA oligonucleotide, sequence GACC. The initial coordinates were taken from a ribosome crystal structure (PDB: 3G6E), residue 2623 to 2626. Simulations were performed using the ff99-bsc0-χ_{OL3} force field [38, 39, 168]. The system was solvated in a box containing 2502 TIP3P [170] water molecules and the system charge was neutralized by adding 3 Na^+ counterions, consistently with previous simulations [120, 166]. A RECT simulation was performed using 16 replicas simulated for 300 ns each. The γ ladder was chosen in the range from 1 to 4 following again a geometric distribution. The initial structures for the H-REMD were taken from a 500 ps MD at 600 K, to avoid correlations of the bias during the initial deposition stage of the WT-MetaD. Other details of the simulation protocol were chosen as for the previous system. As depicted in Fig. 4.3, for each residue the dihedrals of the nucleic acid backbone (α, β, ϵ, γ, ς), together with the pseudo-dihedrals angles of the ribose ring (θ_1 and θ_2) and the glycosidic torsion angle (χ) were chosen as CVs. To help the free rotation of the nucleotide heterocyclic base around the glycosidic bond, the minimum distance between the center of mass of each base with the other three bases was also biased. For the WT-MetaD we used the same parameters as in the previous system. Gaussian width for the minimum distance between bases was chosen equal to 0.05 nm.

For this system a H_{dih}-REMD, a T-REMD and a plain MD simulation were performed in addition to the RECT. In the case of H_{dih}-REMD we used 24 replicas with scaling factors λ ranging from 1 to 0.25, so as to cover the same range of the γ values chosen for the RECT. In the T-TREMD 24 replicas were used to cover a temperature

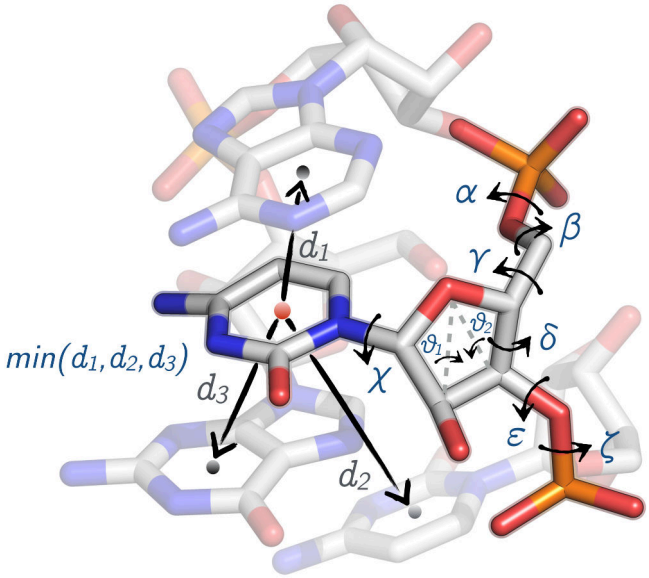


Figure 4.3: Schematic representation of the collective variables used for the tetranucleotide simulation. For each nucleotide, the labeled dihedral angles and the minimum distance between the nucleobase center of mass and the other three nucleobases were biased.

range between 300 K and 400 K with a geometric distribution. For both methods, T-REMD and H_{dih}-REMD, the simulation length was 200 ns per replica. Exchanges were attempted every 120 steps. The conventional MD simulation was run for 4.8 μs. All the simulations (RECT, H_{dih}-REMD, T-REMD and conventional MD) correspond to the same total simulated time.

4.2.4 Analysis

4.2.4.1 Dihedral entropy

As the bias compensates the underlying free energy the probability distribution of the biased CVs is partially flattened. The main CVs used in our method are dihedrals angles. To quantify the effect of the Hamiltonian modifications on the angle distributions one-dimensional entropies (S_{1d}) were estimated. The calculation procedure was equivalent to the one used in ref [177] to evaluate the configurational entropy associated with soft degrees of freedom in proteins. We employed wrapped Gaussian kernels to estimate the histogram profile of each dihedral. Histograms were calculated with PLUMED 2.0. For all the distributions the bandwidth for the kernel density estimation was set to 0.017 rad. We underline that using this definition we only evaluate the flatness of the individual one-dimensional distributions, and cross-correlation between CVs is ignored.

4.2.4.2 RNA conformations

RNA conformations were classified according to the combination of the nucleotides χ angles rotameric states. Torsions orientations in the range of -0.26 to 2.01 rad were consider as *syn*, while the remaining ones were classified as *anti*. The limiting values were chosen according to the position of the barriers in the χ free-energy profiles of all the residues. The result of this clustering procedure gave $2^4 = 16$ different states that are kinetically well separated by the high torsional barriers. We observe that the population of these states does not depend only on the torsional potential associated to the χ dihedrals but include contributions from base-base stacking, hydrogen bonds, solvation of bases, etc.

4.3 Results

In this section we first test our methodology on a standard model system, dALA in water. Then we present results for the more challenging case of the conformational sampling of a tetranucleotide. For all the applications we benchmark against plain MD and a H-REMD where the dihedral potentials are scaled. All the comparisons are made using the same total simulated time.

4.3.1 Alanine Dipeptide

The goal of the introduced method is to enhance conformational sampling in the unbiased replica. The possibility to explore different metastable conformations in this replica relies on the fact that probability distributions in the biased replicas are flattened and that conformations can travel across the replica ladder. These conditions can be verified by monitoring the exchange rate and the flatness of the distributions.

The acceptance rate is in the range 65-72% for RECT and in the range 43-53% for H_{dih}-REMD, indicating that the former method requires less replicas. This is likely due to the fact that the total number of scaled dihedrals in H_{dih}-REMD is larger than the number of biased CVs in RECT. For both REMD methods we also verified that all the trajectories in the generalized ensemble sampled the same conformational ensemble (see Fig. 4.4).

A quantitative measure of the flatness of the distribution in the biased replicas can be obtained from the dihedral entropy, shown in Fig. 4.5 as a function of scaling factors (γ and λ for RECT and H_{dih}-REMD, respectively). The limiting value corresponding to a flat distribution is also indicated. Entropy grows faster as a function of the scaling factor when using RECT, indicating that free-energy barriers on the dALA isomerization transition are more effectively compensated by the bias potentials. With H_{dih}-REMD entropy of Ψ angle saturates and apparently the distribution cannot be further flattened by decreasing λ. In the case of the Φ angle, the dihedral entropy does not grow monotonically when is decreased. This behavior indicates that the relevant free-energy barriers are not only originating from the dihedral force-field terms. The

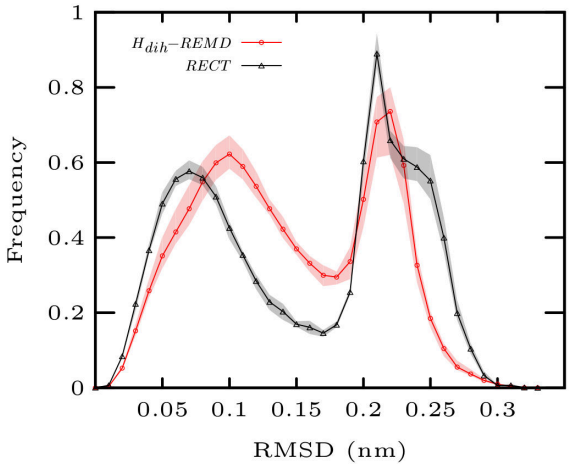


Figure 4.4: Empirical probability distribution of heavy atom RMSD from the alanine dipeptide in water as computed from the trajectories across the replica ladder, shown for both REMD methods. Average probability is shown in solid line and range between maximum and minimum probability among all trajectories is shaded. It can be appreciated that, for each method, all the trajectories span the same conformational distribution.

conformational transitions involve indeed also changes in water coordination, reorganization of hydrogen bonds, non-bonded interactions, etc. On the contrary, RECT achieves an almost flat distribution for both dihedral angles at the highest value of the γ factor. Backbone dihedral distributions for all the replicas are shown in Figs A.1-A.2. The conformations sampled on each replica are shown projected on the Φ,Ψ free-energy landscape in Fig. A.3, where it can be appreciated that all the relevant basins (α, β, and α_R) are explored and connected by points close to the minimum-action pathways. (see refs [167, 178]).

To assess the efficiency and the accuracy of the introduced enhanced sampling technique the free energy difference ΔF between the states $\phi \in [-\pi, 0]$ and $\phi \in [0, \frac{\pi}{2}]$ was calculated from the distribution of the unbiased replica. Results are shown as a function of time in Fig. 4.6, for the two REMD schemes and for the reference conventional MD. Both H-REMD methods converge to the right value with a similar behavior, whereas plain MD needs several tens of ns for the first transition to be observed. The similarity in the convergence of RECT and H_{dih}-REMD indicates that for this system the moderate flattening of the distribution induced by H_{dih}-REMD is sufficient to achieve ergodicity on this time scale. In order to better evaluate differences between the performance of RECT and H_{dih}-REMD we applied this methodologies to a more complex system. Results are shown in the next section.

4.3.2 Tetranucleotide

Also in this case we monitor the average exchange ratio (76-83% for H_{dih}-REMD, 25-32% for T-REMD, and 60-80% for RECT). In Fig. 4.7 the variation of the exchange ratios in time is shown for the exchanges between the first and the last 2 replicas of

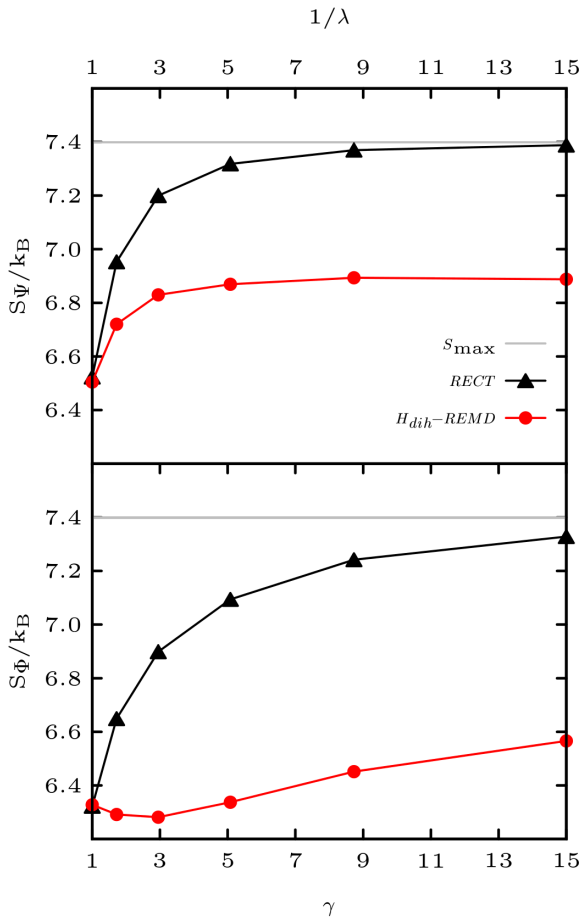


Figure 4.5: Entropy for Ψ (top) and Φ (bottom) dihedral angles in alanine dipeptide. Entropies are shown as a function of $1/\lambda$ and γ for H_{dih}-REMD and RECT respectively. As the entropies values increase the dihedral distributions become more flat. The maximum entropy value corresponding to a flat distribution is represented with a straight line.

each method. We also checked the consistency of trajectories along the replica ladder. As it can be appreciated in Fig. 4.8, for RECT the trajectories in the generalized ensemble are more consistent than those obtained with the other methods. On the contrary, in the case of T-REMD, agreement among the distributions of RMSD is very poor. During this simulation trajectories across the temperature space remain trapped on different metastable conformations. The same behavior was obtained in ref 26 were several T-REMD simulations were performed on the same system, with the same number of replicas and a similar temperature range. In that work divergence among the obtained generalized ensembles was observed even for a simulated time as long as 2 μs per replica. For H_{dih}-REMD and RECT round-trip times are shown on Fig. A.4. The average round-trip time is $\approx$ 0.5 ns for H_{dih}-REMD, $\approx$ 1.8 ns for T-REMD, and $\approx$ 1.2 ns for RECT.

In Fig. 4.9 we show the sum of the entropies for the 32 dihedrals used as CVs. In this respect, RECT is clearly more effective than H_{dih}-REMD in flattening the dihedral distributions, consistently with what was observed for dALA. Notably, the entropic increment observed in RECT is close to the one observed in T-REMD when

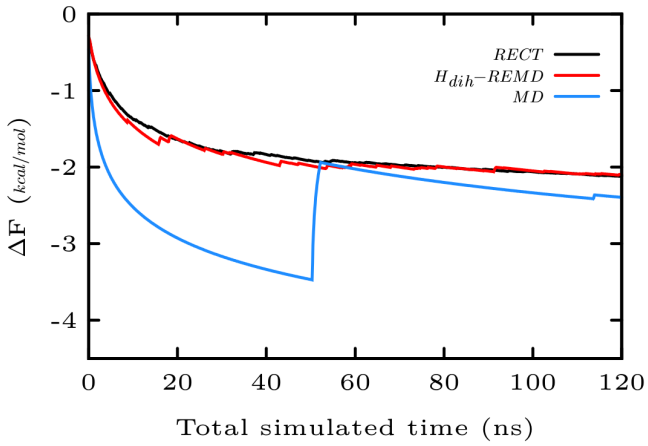


Figure 4.6: Estimate of the free-energy difference between the two metastable minima in alanine dipeptide. Data are shown for both replica exchange methods (H_{dih}-REMD and RECT) and for conventional MD as a function of the total simulation time.

using an equivalent temperature This confirms that RECT has an effect comparable to that of raising the temperature of the biased CVs by a factor γ.

The significancy of this entropic values could be appreciated on the time series and related histograms for all the dihedral angles shown in Figs. A.5-A.8 for the most and least ergodic replica of H_{dih}-REMD and RECT. It is clear that for RECT, at the most ergodic replica, all the accessible torsional range is sampled. On the contrary, in the highest replica of H_{dih}-REMD the distributions of some torsions are not flattened.

The transition around the glycosidic bond, from *anti* to *syn*, is among the slowest relaxation times in RNA dynamics [26]. To evaluate the convergence of the unbiased replica we analyzed the population of the *anti* rotamer for each nucleotide χ angle. Populations are shown in Fig. 4.10 as a function of the total simulated time. For all the nucleotides the *anti* conformations are preferred. The guanosine is the nucleotide with the highest *syn* proportion, and the cytidines the ones with the smallest ($<$ 2%), as correspond to their rotameric preferences [43]. Values from both H-REMD approaches seem well consistent, except for the population of the first nucleotide. From the time behavior of these populations, it is clear that for all the REMD approaches the guanosine proportion of *anti* is the most difficult to converge. Here RECT can reach values close to a longer reservoir-REMD simulation [120] while both H_{dih}-REMD and T-REMD show results closer to those obtained from conventional MD, with a higher occupation of the *anti* conformer.

We observe that our method is enforcing the exploration of both *anti* and *syn* conformations in the biased replicas for each nucleotide independently. This however does not guarantee that all the 16 combinations of *anti* and *syn* conformations are explored. Fig. 4.11 shows the free energy of the RNA structures grouped by the combination of the χ angle *anti(a)/syn(s)* rotamers. All 16 combinations, except for *ssss* and *asss*, are sampled in the unbiased replica from RECT. On the contrary, the unbiased replica from T-REMD and H_{dih}-REMD explores respectively 13 and 8 of the states, and plain MD only 5 of them. The most populated cluster corresponds to an

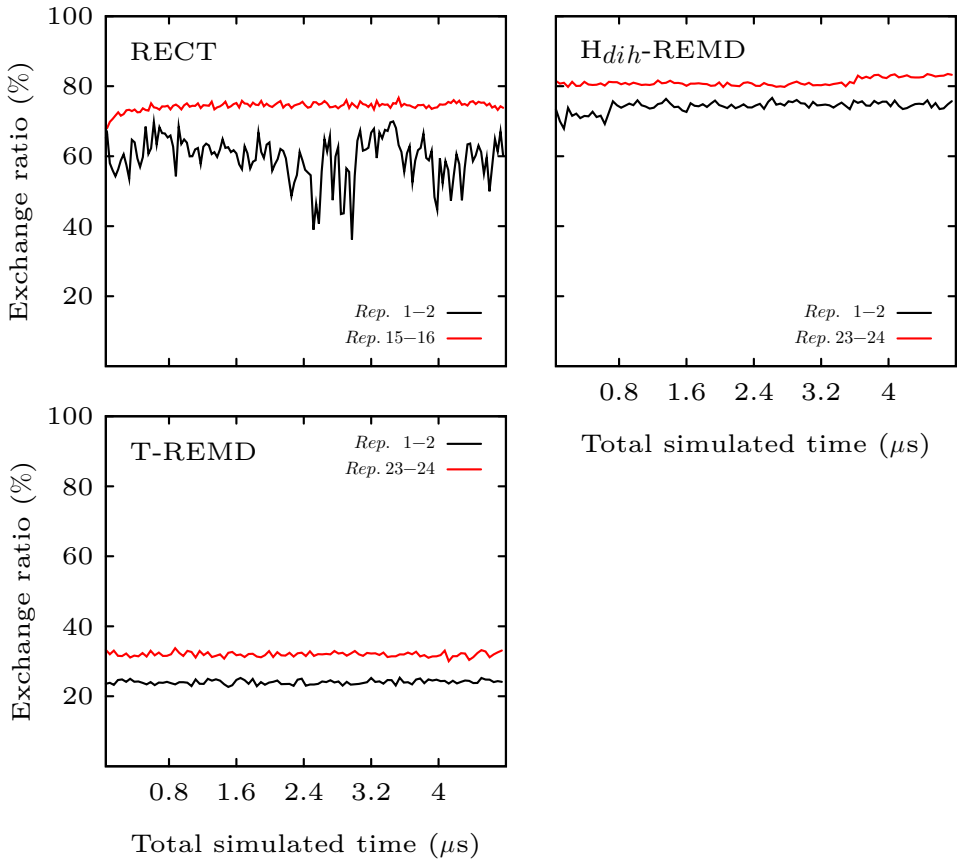


Figure 4.7: Average exchange ratio in subsequent blocks of 2 ns for the exchange between the first two and the last two replicas. In the case of RECT, since the bias potential is changing during the simulation, the acceptance ratio has a higher fluctuation.

all-*anti* conformation, followed by the *saaa*. Then, the three clusters *asaa, ssaa* and *sasa* appear with similar population.

In the same figure the free energy values for the ergodic replica show that all the 16 combinations are populated in RECT within a range of $6k_BT$. In the case of H$_{dih}$-REMD the most ergodic replica visits only 9 combinations with a population that is very close to that of the unbiased replica.The most ergodic replica in T-REMD explores 14 clusters, but their populations have a large statistical errors. We highlight the fact that results from T-REMD could be affected by the lack of convergence of trajectories across the temperature space (see Fig. 4.8). This could lead to an underestimation of the errors as evaluated from block analysis

4.4 Discussion

The introduced method allows to build bias potentials for a Hamiltonian replica-exchange scheme using concurrent well-tempered metadynamics on several CVs at the same time. Replicas are simulated using a ladder of well-tempered bias factors γ. When CVs are correlated, the self-consistency among the bias potentials is crucial to achieve flat sampling in each individual CV, as illustrated in Fig. 4.2. In this case the exact

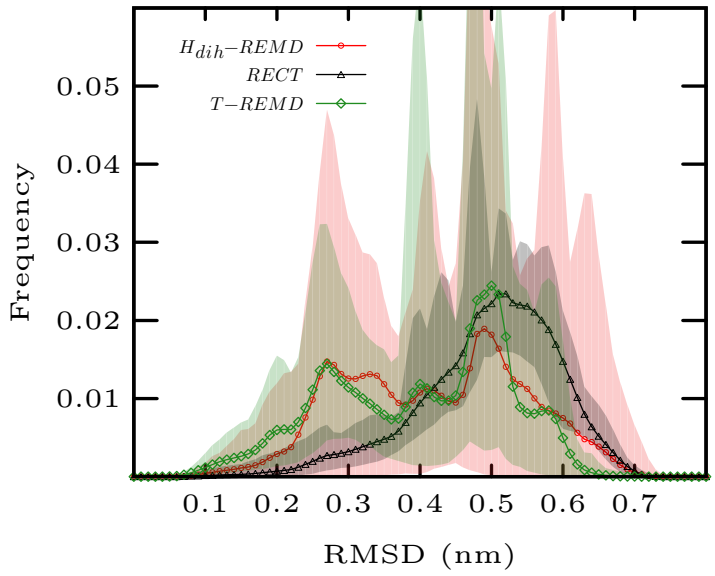


Figure 4.8: Empirical probability distribution of heavy atom RMSD from the canonical A-form as computed from the trajectories across the replica ladder, shown for all REMD methods. Average probability is shown in solid line, and range between maximum and minimum probability among all trajectories is shaded. It can be appreciated that the agreement among the conformational distributions of trajectories from T-REMD and H_{dih}-REMD is poorer than the one of those obtained with RECT. We notice that RECT samples a very different generalized ensemble from those of T-REMD and H_{dih}-REMD schemes.

relationship between bias and free energy is lost. We also remark that here flattening is not complete but modulated by the value of γ. This is useful since it avoids sampling very high energy states (e.g. with steric clashes) that would have a very low chance of being accepted in the unbiased replica. The method compares favorably with both conventional MD and H_{dih}-REMD. The method slightly outperforms T-REMD, where the entire system is heated, indicating that for these small systems there is not a substantial advantage in schemes where part of the system is biased. However, RECT can be straightforwardly generalized to large systems since the acceptance only depends on the size of the biased portion.

Results from both dALA and tetranucleotide simulations show that the bias potentials constructed with concurrent WT-MetaD are able to gradually scale the free-energy barriers. We notice that only barriers in the one-dimensional free-energy profiles are compensated, which means that some regions in the multidimensional space of all the CVs might not be explored. In principle this could hide some important minima that would never be observed. We did not observe this problem in the applications presented here.

The second application on which we tested RECT, namely conformational sampling of a tetranucleotide, is particularly challenging. The conformational space of these small RNA molecules is not constrained by Watson-Crick pairings and ergodic sampling is out of reach of conventional MD simulations [120, 166]. So far, converged ensembles have been obtained only trough highly expensive multidimensional REMD simulations,

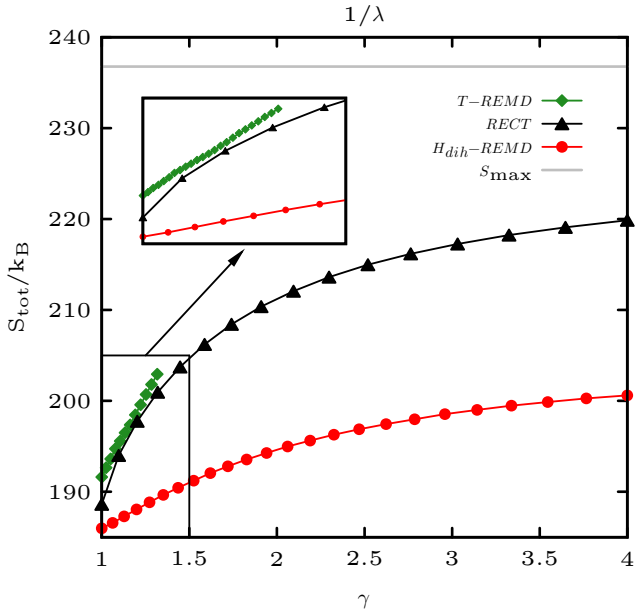


Figure 4.9: Total entropy of backbone, puckering and glycosidic dihedral angles in the tetranucleotide for both replica exchange methods. Entropies are shown as a function of $1/\lambda$ and γ for H_{dih}-REMD and RECT respectively. For T-REMD, temperature is chosen as $T = \gamma \cdot 300$K. As the entropy increases the dihedral distributions become more flat. The maximum entropy value corresponding to a flat distribution is represented with a straight line. Entropies obtained for the unbiased replicas in the three methods are consistent within their error bars (error not shown).

corresponding to a total simulated time of several tens of μs [31, 32]. One of the reasons for this difficult convergence is the long relaxation time for the *anti* to *syn* transitions, which could be additionally hindered by an incorrect force-field description of base-base stacking and base-solvent interactions [26].

Fig. 4.11 illustrates the ability of RECT to accelerate conformational transitions among the χ angle *anti/syn* rotamers. Although the conformational space of the more biased replicas is highly expanded, the convergence in the unbiased replica is not affected. On the contrary the method facilitates the sampling of glycosidic rotamer conformations that otherwise would not be explored by MD simulations of the same overall length. We finally remark that our procedure can be combined with weighted histogram [179] so as to include the statistics of the biased replicas.

4.4.1 Comparison with related state-of-the-art methods

RECT is based on the idea of building a replica ladder where a large set of selected CVs is progressively heated. CVs are heated by flattening their distribution with concurrent well-tempered metadynamics. We first discuss the possibility of using methods other than well-tempered metadynamics to build the replica ladder. Possible alternatives here include ABF [150] or a recently proposed variational approach [180]. These methods could be used in a RECT scheme provided they are suitably extended so as to sample a partially flattened distribution. We also observe that other methods aimed

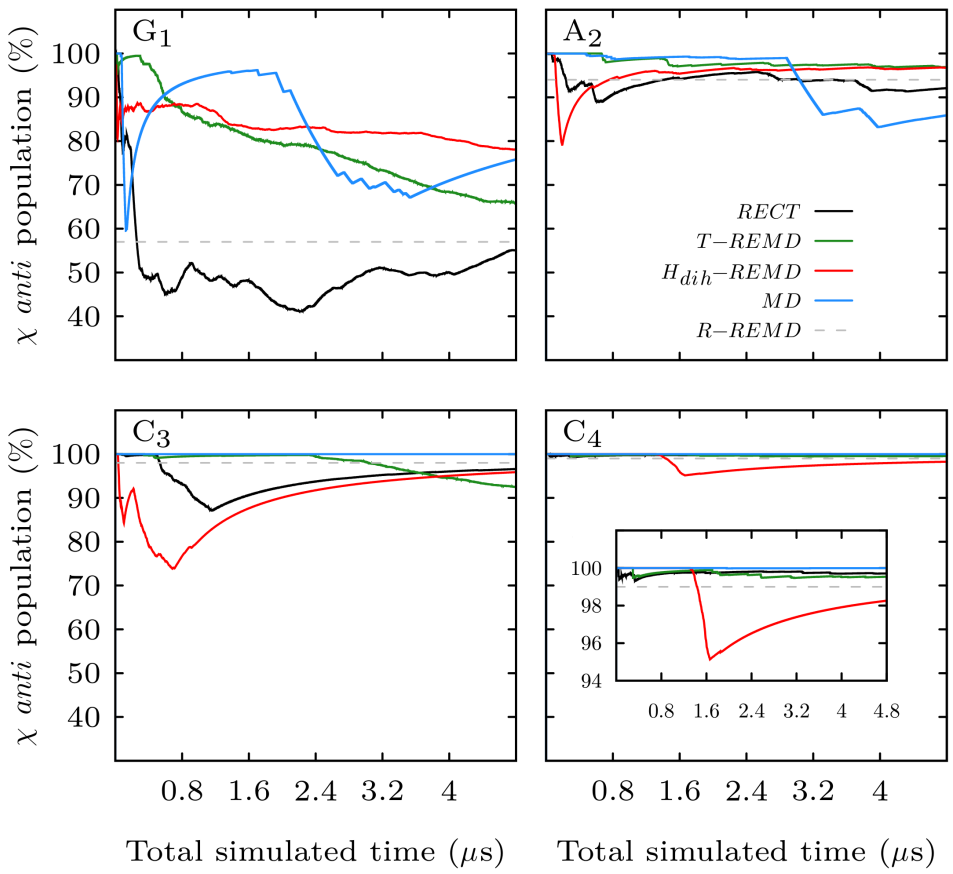


Figure 4.10: Estimated glycosidic angle *anti* population for each nucleotide as a function of the total simulation time. Data are shown for H_{dih}-REMD and RECT unbiased replicas and for conventional MD. Reference values taken from ref [120] are shown as dashed lines.

at keeping selected CVs at a given temperature have been proposed based on coupling thermostats to CVs directly [156–158]. These techniques have been mostly used in the past with an exploration purpose relying on additional calculations so as to provide free energies (see ref. [181]) but it is not clear if they can be integrated in a RECT scheme.

In the following we discuss the comparison of RECT with related methods that are not based on CV tempering.

Comparison with H-REMD of Curuksu and Zacharias. Our method is closely related to the one introduced in ref [163] (see also section 3.2.3). There, a bias potential aimed at disfavoring the most probably rotamers is manually constructed and applied on several replicas using a scaling factor. This bias disfavors the major minima but does not ensure a proper compensation of the free-energy barriers, as their positions and magnitudes are not *a priori* known. The main advantage of RECT is that several low-dimensional bias potentials are built with a self-consistent procedure so that the technique can be straightforwardly applied to a large number of degrees of freedom.

Comparison with bias-exchange metadynamics. In bias-exchange metadynamics every replica performs an independent metadynamics simulation so that one CV at a time is feeling the flattening potential. Thus, it is typically used with a relatively

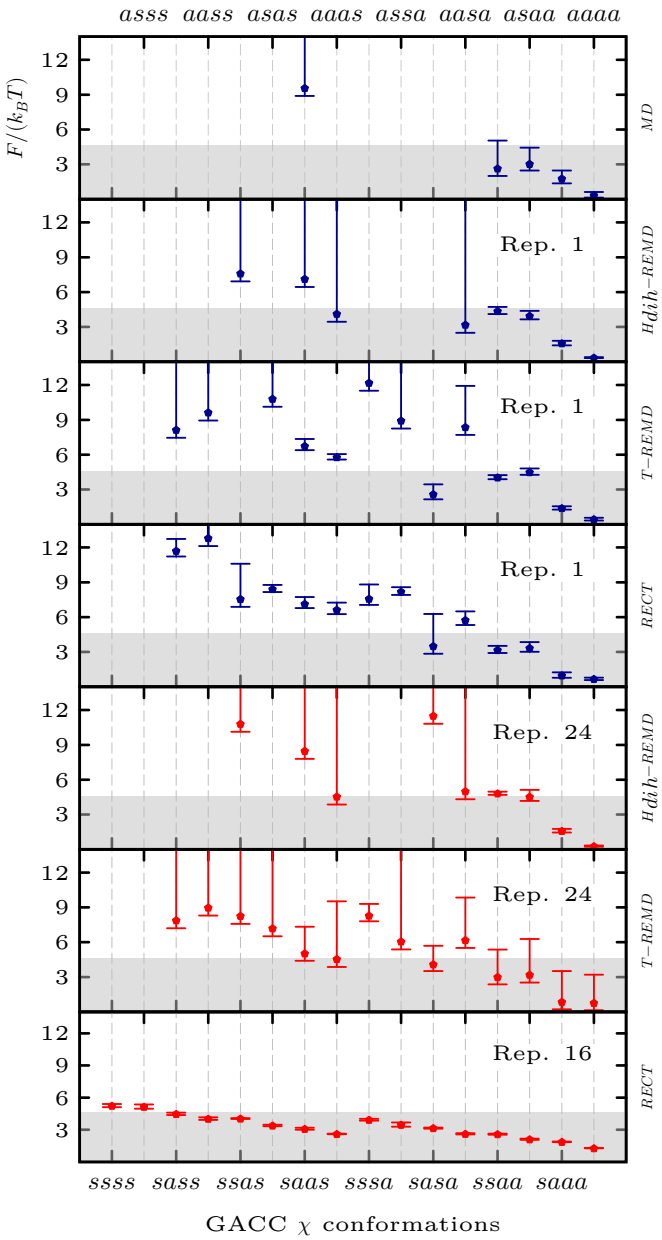


Figure 4.11: Estimated free energies for the tetranucleotide conformations clustered according to the χ angle *anti/syn* rotameric combinations (circles). Free energies are computed as $-k_BT \log P_i$, where P_i is the normalized population of each cluster in the unbiased replica. Grey boxes represent relative populations higher than 1%. Confidence intervals are shown as bars and span the range $[-k_BT \log(P_i+\Delta P_i), -k_BT \log(P_i-\Delta P_i)]$, where ΔP_i is the standard deviation of the average P_i as obtained from four blocks. Clusters which are observed in only one of the four blocks has an infinite upper bound.

small number of *ad hoc* designed CVs capable to describe the relevant conformational transitions. On the other hand, RECT is designed to be used with a very large number of dummy CVs with little *a priori* information and to bias them concurrently to exploit their cooperation in enhancing conformational sampling. For this reason, the two approaches are complementary and could even be combined in a multidimensional replica exchange suitable for a massively parallel environment.

Comparison with solute tempering and related methods. In replica exchange solute tempering the solute Hamiltonian is scaled so as to obtain an effect equivalent to a rise in the simulation temperature [145, 146]. Any set of atoms can be identified as solute, giving the opportunity to enhance sampling in a region localized in space [176, 182].This requires modifying charges of the enhanced region, with long range effects and sometime affecting fundamental properties such as hydrophobicity. In our method, the bias potentials act on precisely selected degrees of freedom without perturbing their coupling with the rest of the system. Moreover, the bias is adaptively built so as to compensate the free energy and not the potential energy, so that with properly chosen CVs it could be used to compensate entropic barriers.

Comparison with hyperdynamics and accelerated MD. In these methods the potential energy of the system is modified so as to decrease the probability to sample minima on the potential energy [32, 165, 183] (see section 3.2.3). On the contrary, RECT employs a bias which is related with the free energy so as to achieve a flatter histogram on the selected CVs.

We finally remark that RECT, although formally based on the *a priori* choice of a set of CVs, typically requires the same amount of information as methods not based on CVs. Indeed, as we have shown, the method can be easily applied to a very large number of CVs, virtually including by construction all the slow degrees of freedom of the system. Additionally, when a few relevant CVs can be identified based on chemical intuition, RECT can be straightforwardly combined with standard metadynamics similarly to parallel tempering [159] or solute tempering [184].

4.5 Conclusion

Replica exchange with collective-variable tempering (RECT) has been here proposed as a novel and flexible enhanced-sampling method. RECT takes advantage of the adaptive nature of well-tempered metadynamics to build bias potentials that compensate free-energy barriers. The flattening of the barriers is modulated by the well-tempered factor γ, and the chosen collective variables (CVs) are effectively kept at a higher temperature. The biasing potentials are built combining concurrent low-dimensional metadynamics protocols so as to be usable on a very large number of CVs. Multiple replicas are then used so as to smoothly interpolate between a highly biased, ergodic simulation and an unbiased one ($\gamma = 1$). The number of required replicas scales with the square root of the number of chosen CVs for a fixed range of γ factors. This allows a very large number of CVs to be biased, so that virtually all the relevant transitions

can be accelerated. The CVs used here were mostly dihedral angles, which exhibit relevant barriers in many biomolecular conformational transitions, but the method can be used with any CV. The application of this technique to the dALA in water shows that the CV probability distributions are effectively flattened by the action of the bias potentials and unbiased statistics is correctly recovered. In the case of the tetranucleotide conformational sampling is greatly enhanced since RECT effectively overcome the high free energy barriers of the χ angle transitions that hindered the conformational sampling at room temperature. RECT is a promising tool to enhance the exploration of the conformational space in highly flexible biomolecular systems such as RNA, proteins, or RNA/protein complexes.

Chapter 5

Empirical corrections to the Amber RNA force field

5.1 Overview

Recent tests [33, 34] have shown that state-of-the-art force fields for RNA are still not accurate enough to produce ensembles compatible with NMR data in solution in the case of single stranded oligonucleotides. Similar issues have been reported for DNA and RNA dinucleosides [185, 186]. Previous studies have shown that the distribution of structures sampled from the protein data bank (PDB) may approximate the Boltzmann distribution to a reasonable extent [187–190] and could even highlight features in the conformational landscape that are not reproduced by state-of-the-art force fields [191, 192]. This has been exploited in the parametrization of protein force fields. For example, a significant improvement of the force fields of the CHARMM family has been obtained by including empirical corrections commonly known as CMAPs based on distributions from the PDB [193, 194].

In this work, we apply these ideas to the RNA field and show how it is possible to derive force-field corrections using an ensemble of X-ray structures. At variance with the CMAP approach [195], we here correct the force field using a self-consistent procedure where metadynamics is used to enforce a given target distribution [35, 36]. Correcting potentials are obtained for multiple dihedral angles using the metadynamics algorithm in a concurrent fashion. Since the target distributions are multimodal, we also use the enhanced sampling technique introduced in Chapter 4, replica exchange with collective-variable tempering (RECT), to accelerate the convergence of the algorithm. The correcting potentials are obtained by matching the torsion distributions for a set of dinucleoside monophosphates. The resulting corrections are then tested on tetranucleotides where standard force field parameters are known to fail in reproducing NMR data.

5.2 Methods

In this Section we briefly describe the target metadynamics approach and discuss the details of the performed simulations.

5.2.1 Targeting Distributions with Metadynamics

Metadynamics (MetaD) has been traditionally used to enforce an uniform distribution for a properly chosen set of collective variables (CV) that are expected to describe the slow dynamics of a system [151]. However, it has been recently shown that the algorithm can be modified so as to target a preassigned distribution which is not uniform [35, 36]. In this way a distribution taken from experiments, such as pulsed electron paramagnetic resonance, or from an X-ray ensemble, can be enforced to improve the agreement of simulations with empirical data. We refer to the method as target metadynamics (T-MetaD), following the name introduced in ref [35]. For completeness, we here briefly derive the equations. It is also important to notice that the same goal could be achieved using a recently proposed variational approach [180, 196].

In our implementation of T-MetaD a history dependent potential $V(s,t)$ acting on the collective variable s at time t is introduced and evolved according to the following equation of motion

$$\dot{V}(s,t) = \omega e^{\beta(\tilde{F}(s(t))-\tilde{F}_{\max})} e^{-\beta(\frac{V_{\max}}{D})} e^{-\frac{(s-s(t))^2}{2\sigma^2}} \tag{5.1}$$

Here $\beta = 1/k_B T$, k_B is the Boltzmann constant, T the temperature, ω is the initial deposition rate of the kernel function which is here defined as a Gaussian with width σ, $\tilde{F}$(s) is the free energy landscape associated to the target distribution, $\tilde{F}_{\max}$ indicates the maximum value of the function $\tilde{F}$, and D is a constant damping factor. The target distribution is thus proportional to $e^{-\beta\tilde{F}(s)}$. We define $\omega = \frac{Dk_BT}{\tau}$ where τ is the characteristic time of bias deposition. The term $e^{\beta(\tilde{F}(s)-\tilde{F}_{\max})}$ adjusts the height of the bias potential, making Gaussians higher at the target free-energy maximum and lower at its minimum. This forces the system to spend more time on regions where the targeted free-energy is lower. We notice that a similar argument has been used in the past to derive the stationary distribution of both well-tempered metadynamics, where Gaussian height depends on already deposited potential [40], and of adaptive-Gaussian metadynamics, where Gaussian shape and volume is changed during the simulation [154]. The subtraction of $\tilde{F}_{\max}$ sets an intrinsic upper limit for the height of each Gaussian, thus avoiding the addition of large forces on the system. We notice that other authors used terms such as the minimum of F or the partition function to set an intrinsic lower limit for the height of each Gaussian [35, 36]. At the same time, the term $e^{-\beta(\frac{V_{\max}}{D})}$ acts as a global tempering factor [155] and makes the Gaussian height decrease with the simulation time so as to make the bias potential converge instead of fluctuating. As observed in ref [35], the tempering approach used in well-tempered MetaD in this case would lead to a final distribution that is a mixture of the target

one with the one from the original force field. For this reason, we prefer to use here a global tempering approach [155].

In the long time limit (quasi-stationary condition) the bias potential will on average grow as [40, 155]

$$\left\langle \dot{V}(s) \right\rangle = \int ds' \omega e^{\beta(\tilde{F}(s')-\tilde{F}_{\max})} e^{-\beta(\frac{V_{\max}}{D})} e^{-\frac{(s'-s)^2}{2\sigma^2}} P(s') \tag{5.2}$$

where $P(s)$ is the probability distribution of the biased ensemble. Defining the function $g(s') = \omega e^{\beta(\tilde{F}(s')-\tilde{F}_{\max})} e^{-\beta(\frac{V_{\max}}{D})}$ we can see this equation is a convolution of a Gaussian and a positive definite function.

$$\left\langle \dot{V}(s) \right\rangle = \int ds' e^{-\frac{(s'-s)^2}{2\sigma^2}} g(s') P(s'). \tag{5.3}$$

As shown in ref [40, 155] this average should be independent of s in stationary conditions, so that the function $g(s')P(s')$ should be also independent of s', though still dependent on time

$$h_0 e^{\beta(\tilde{F}(s(t))-\tilde{F}_{\max})} e^{-\beta(\frac{V_{\max}}{D})} P(s) = C(t). \tag{5.4}$$

By recognizing that $\tilde{F}_{\max}$ and $V_{\max}$ do not depend on s, one can transform the last equation to

$$e^{\beta\tilde{F}(s)} P(s) = C'(t) \tag{5.5}$$

which implies that

$$P(s) \propto e^{-\beta\tilde{F}(s)}. \tag{5.6}$$

Thus, the system will sample a stationary distribution of s which is identical to the enforced one.

Whereas the equations are here only described for a single CV, this method can be straightforwardly applied to multiple CVs in a concurrent manner. In this case, the total bias potential is the sum of the one-dimensional bias potentials applied to each degree of freedom. Indeed, similarly to the concurrent metadynamics used in RECT [197] (see also Chapter 4), all the distributions are self-consistently enforced [36]. This is particularly important when biasing backbone torsion angles in nucleic acids since they are highly correlated [42, 198]. In this situation it is also convenient to use a biasing method that converges to a stationary potential through a tempering approach, to include in the self-consistent procedure of MetaD an additional effective potential associated to the correlation between the dihedral angles that is as close as possible to convergence.

5.2.2 Model systems

5.2.2.1 RNA dinucleoside monophosphates

Fragments of dinucleoside monophosphate with the sequence CC, AA, CA, and AC were extracted from the PDB database of RNA X-ray structures at medium and high resolution (resolution < 3 Å). The selected structures were protonated using *pdb2gmx* tool from GROMACS 4.6.5 [174] Free-energy profiles along the backbone dihedral angles were calculated with the *driver* utility of PLUMED 2.1 [175].

Molecular dynamics simulations of the chosen RNA dinucleoside monophosphate sequences were performed using the ff99-bsc0-χ_{OL3} force field (named here Amber14) [37–39]. The systems were solvated in an octahedron box of TIP3P water molecules [170] with a distance between the solute and the box wall of 1 nm. The system charge was neutralized by adding 1 Na^+ counterion. The LINCS [171] algorithm was used to constrain all bonds containing hydrogens and equations of motion were integrated with a timestep of 2 fs. All the systems were coupled to a thermostat through the stochastic velocity rescaling algorithm [173]. For all non-bonded interactions the direct space cutoff was set to 0.8 nm and the electrostatic long-range interactions were treated using the default particle-mesh Ewald [21] settings. An initial equilibration in the NPT ensemble was done for 2 ns, using the Parrinello-Rahman barostat [199]. Production simulations were run in the NVT ensemble. All the simulations were performed using GROMACS 4.6.7 [174] patched with a modified version of the PLUMED 2.1 plugin [175].

T-MetaD simulations were run to enforce the probability distributions of the angles ϵ_1, ζ_1, α_2 and β_2, which were calculated from the X-ray fragments. The target free-energy profiles were calculated with PLUMED 2.1. Distributions were estimated as combination of Gaussian kernels, with a bandwidth of 0.15 rad, and written on a grid with 200 bins spanning the $(-\pi, \pi)$ range. The bias potential used for the T-MetaD was grown using a characteristic time $\tau = 200$ ps and a dampfactor $D = 100$. Gaussians with a width of 0.15 rad were deposited every $N_G = 500$ steps.

We underline that simulations performed using T-MetaD could be non ergodic for two reasons. First, there could be significant barriers acting on CVs that are not targeted and thus not biased at this stage (e.g. χ dihedral angles). Second, if the enforced distribution of a CV is bimodal it will be necessary to help the system in exploring both modes with the correct relative probability. It is thus necessary to combine the T-MetaD approach with an independent enhanced-sampling scheme. Here we used RECT, a replica exchange method, introduced in Chapter 4, where a group of CVs is biased concurrently using a different bias factor for each replica and one reference replica is used to accumulate statistics [197]. When T-MetaD and RECT are combined, in each replica a T-MetaD is run with the same settings, including the reference replica. The T-MetaD/RECT simulation was run with 4 replicas for 1 μs each. For each residue the dihedrals of the nucleic acid backbone (α, β, γ, ϵ, ζ), together with one of the Cartesian coordinates of the ring puckering [63] (Zx) and the

glycosidic torsion angle (χ) were chosen as accelerated CVs. To help the free rotation of the nucleotide heterocyclic base around the glycosidic bond, the distance between the center of mass of nucleobases was also biased. For the dihedral angles the Gaussian width was set to 0.25 rad and for the distance it was set to 0.05 nm. The Gaussians were deposited every $N_G = 500$ steps. The initial Gaussian height was adjusted to the bias factor γ of each replica, according to the relation $h = \frac{k_B T(\gamma-1)}{\tau_B} N_G \Delta t$, in order to maintain the same $\tau_B = 12$ ps across the entire replica ladder. The bias factor γ ladder was chosen in the range from 1 to 2, following a geometric distribution. In replicas with $\gamma \neq 1$ the target free energy was scaled by a factor $1/\gamma$. Exchanges were attempted every 200 steps. Statistic was collected from the unbiased replica.

Finally, a new RECT simulation was run for each dinucleoside with the bias potentials obtained from the T-MetaD applied statically on each replica. These calculations represent the results obtained with a force field that includes the corrections from the PDB distributions and are thus labeled as Amber$_{pdb}$. Statistics from these simulations were collected to evaluate the effects of the corrections. The simulation time was 1 μs per replica.

5.2.2.2 RNA Tetranucleotides

To test the force field corrections derived on dinucleoside monophosphates, temperature replica-exchange molecular dynamics (T-REMD) simulations [136] were performed on different tetranucleotide systems with sequence CCCC, GACC and AAAA. The correcting potentials calculated for the AA and CC dinucleosides were applied to all the backbone angles of AAAA and CCCC tetranucleotides, respectively. For the GACC tetranucleotide we combined the correcting potentials from the T-MetaD simulations of AA, AC and CC, assuming a similarity between purines A and G.

The T-REMD data related to the Amber14 force field and the protocol for the new simulations performed using the Amber$_{pdb}$ force field were taken from ref [192]. The systems were solvated with TIP3P waters and neutral ionic conditions. We used 24 replicas with a geometric distribution of temperatures from 300 to 400 K. Exchanges were attempted every 200 steps. The simulation length was 2.2 μs per replica.

5.2.3 Analysis

5.2.3.1 Comparison with experimental data

The result of the molecular dynamics simulations was compared to NMR experimental data of dinucleosides [185, 200–202] and tetranucleotides [34, 166, 203]. We used ^3J scalar couplings and NOE distances from those experiments to evaluate the quality of the FF ensembles.

The vicinal nuclear spin-spin ^3J scalar couplings constants were calculated according to the conformation of the related torsion angles using the Karplus relationship in the form $J(\phi) = A\cos^2(\phi+\varphi) + B\cos(\phi+\varphi) + C$ [204, 205]. Several sets of coefficients

are available for each specific observable/torsion equation (see ref [206]) and there has been no clear consensus on which of them is to be preferred. We took into account the analysis made in refs [185, 207, 208] to select the most precise sets of parameters. For ${}^3J_{H_{4'}H_{5'}}$ and ${}^3J_{H_{4'}H_{5''}}$ we derived a simplified expression for the generalized Karplus equation in the form $J(\phi) = A\cos^2(\phi) + B\cos(\phi) + \tilde{B}\sin(\phi)\cos(\phi) + C$. The parameters used in this study are listed in Table 5.1.

Coupling	Angle	A	B	$\tilde{B}$	C	φ	Ref
${}^3J_{H_{1'}H_{2'}}$,${}^3J_{H_{2'}H_{3'}}$,${}^3J_{H_{3'}H_{4'}}$	$\nu_{1,2,3}$	9.67	-2.03	0	0	0	[34]
${}^3J_{H_{4'}H_{5'}}$	γ	8.31	-0.99	0.27	1.37	-120°	[209]
${}^3J_{H_{4'}H_{5''}}$	γ	8.31	-0.99	-4.72	1.37	0	[209]
${}^3J_{H_{5'}P}$	β	18.1	-4.8	0	0	-120°	[210]
${}^3J_{H_{5''}P}$	β	18.1	-4.8	0	0	120°	[210]
${}^3J_{H_{3'}P}$	ϵ	15.3	-6.1	0	1.6	120°	[211]
${}^3J_{C2'P}$	ϵ	6.9	-3.4	0	0.7	-120°	[211]
${}^3J_{C4'P}$	β/ϵ	6.9	-3.4	0	0.7	0	[211]
${}^3J_{C2H_{1'}}$	χ	3.9	1.7	0	0.3	-70.4°	[212]
${}^3J_{C4H1'}$	χ	3.6	1.8	0	0.4	-68.6°	[212]
${}^3J_{C6H_{1'}}$	χ'	4.8	0.7	0	0.3	-66.9°	[212]
${}^3J_{C8H1'}$	χ'	4.2	-0.5	0	0.3	-68.9°	[212]

Table 5.1: Karplus parameters for the dihedral angles considered in this study. χ' indicates the H1'-C1'-N1/9-C6/8 torsion along with a phase shift of 60°, which in the special case of base planar at N1/9 is equal to χ. Actually, the relations of ${}^3J_{C-H}$ with the χ angle have been shown to depend non-trivially on the sugar pucker and on the nonplanarity of nucleobases [208, 213].

The ${}^3J$ scalar couplings from the simulations were calculated as the ensemble average over the sampled conformational space, using the following equation

$$\langle J \rangle = \sum_{\theta=-\pi}^{\pi} f(\theta)\, J(\theta)\, \delta\theta \tag{5.7}$$

where $J(\theta)$ represent the Karplus relation between the vicinal coupling and the dihedral angle and $f(\theta)$ is the probability density of the dihedral angle bin. To calculate the torsion angle histograms we employed wrapped Gaussian kernels with a bandwidth of 0.017 rad. Histograms were calculated with PLUMED 2.1 [175].

The overall agreement between the NMR data and the average values calculated in this study was measured using the root mean square error (RMSE):

$$RMSE = \sqrt{N^{-1}\sum_{i=1}^{N}(\langle J_i \rangle_{calc} - J_{i,exp})^2} \tag{5.8}$$

The consistency of the error measurements was analyzed by blocking the trajectory in 4 blocks of equal length and calculating the standard deviation of the different error estimations.

2D NOESY experiments of different tetranucleotide sequences have provided rigorous benchmarks for force-fields modifications [34, 166, 203, 214]. NOE distances

were calculated by averaging pairwise proton-proton distances over all the structures within the ensemble. The deviation of MD distances from experimental NOE derived distances is calculated as [175]

$$RMSE = \sqrt{N^{-1} \sum_{i}^{noes} ((\frac{1}{N} \sum_{j=1}^{N} (\frac{1}{r_j^6}))^{-\frac{1}{6}} - d_i^{exp})^2} \tag{5.9}$$

A very important indicator of the ensemble agreement with the experiment is the number of proton-proton contacts with an MD averaged distance of ≤ 5 Å which are not visible in the NOESY spectra [34]. Calculations were performed using the software tool baRNAba [215].

5.2.3.2 Thermodynamics

To calculate the free-energy of stacking we used the definition similar to the one of ref [34] to define the stacked and unstacked states. In particular, we calculate the distance between the center of mass of the nucleobases using only the heteroatoms (with a cutoff of 5 Å), the angle between the vectors normal to the planes of the bases (from 0º to 45º and from 135º to 180º) to separate the parallel to the T-shaped complexes, and the angle between the distance vector between the bases and the 5'-nucleobase normal vector (< 50º). This definition is very similar to the one used on ref [34].

5.2.3.3 Mutual Information and Jensen-Shannon divergence

The correlation between the dihedral angles in the tetranucleotide T-REMD simulations was estimated with the Mutual Information (MI) [216]. We used the *driver* command of PLUMED 2.1 [175] to calculate the MI as an average along the trajectory.

$$MI_{xy} = \left\langle \ln \left[\frac{p(x,y)}{p(x)p(y)} \right] \right\rangle \tag{5.10}$$

The difference between the probability distributions from the A-form and Non-A-form sub-ensembles was measured using the Jensen–Shannon divergence (JS) [217, 218]. The JS is zero for identical distributions and reaches it maximum ($\ln 2$) for non-overlapping ones. The probability distributions used to estimated JS are shown in Figs. 5.6, B.1 and 5.10.

$$JS_{AB} = \frac{1}{2} \left\langle \ln \left[\frac{2p_A(x,y)}{p_A(x,y) + p_B(x,y)} \right] \right\rangle_A + \frac{1}{2} \left\langle \ln \left[\frac{2p_B(x,y)}{p_A(x,y) + p_B(x,y)} \right] \right\rangle_B \tag{5.11}$$

5.3 Results

As a first step we identified the dihedral angles whose correction could benefit the most the experimental agreement of the whole conformational ensemble. Then, we used our approach to enforce for those dihedrals the distributions from the X-ray

fragments on monophosphate dinucleosides AA, AC, CA, and CC. Finally, we show that the corrections are partly transferable and could improve agreement with solution experiments for tetranucleotides.

5.3.1 Selection of the target collective variables

The prevalence of compact intercalated and inverted conformations in the ensembles of RNA tetranucleotides generated with AMBER force fields is a known problem (see Fig 5.1 for a representation of typical structures) [33, 34]. This can be due to an over-stabilization of stacking interactions, poor water models, and/or incorrect dihedral parameters. Changing the non-bonded interactions in a force field to improve stacking is a difficult task, as the classification of stacked (closed) and non-stacked (open) structures in a molecular dynamic simulation is largely arbitrary and slight changes can lead to very different values of the open-closed population ratio [219, 220]. On the other hand, dihedrals terms are more flexible and small corrections in the free-energy profiles of a minimal number of angles can have huge impact on the whole nucleic acids ensemble [16]. Therefore, we decided to correct the free energy landscape of an essential group of dihedral angles in the RNA, in order to improve the state-of-the-art AMBER force field agreement with solution NMR data.

For this analysis we used the T-REMD simulations of AAAA, GACC, and CCCC tetranucleotides performed on ref [192] using the Amber14 force field. We divided each of the Amber14 ensembles into two groups, in order to identify the structural features that differentiate the structures compatible with the NMR data (A-form-like conformations) from the non-compatible compact structures that overpopulate the ensembles. The A-form sub-ensemble was defined as the set of conformations with a distance-RMSD < 2.5 Å form the canonical A-form, while the Non-A-form group comprises the rest of the frames. The ratio between the population of the Non-A-form sub-ensemble over the A-form one is different for each tetranucleotide: $\sim$5.5 for AAAA, $\sim$1.4 for GACC and $\sim$21.7 for CCCC.

Differences among the collective variable (CV) distributions of the sub-ensembles were measured using the Jensen-Shannon divergence between the two-dimensional probability distributions of dihedrals (α to χ), puckering coordinates (Z_x) and the nucleobase-nucleobase coordination number (S): $\alpha\beta$, $\beta\gamma$, γZ_x, $Z_x\chi$, $S\chi$, $Z_x\epsilon$, $\epsilon\zeta$, $\zeta\alpha$. Coordination numbers were estimated using a switching function with form $\frac{1}{1+(r/r_0)^6}$ with $r_0 = 0.3$nm. In Fig. 5.2 it can be appreciated the JS divergences for each pair of CVs, while the probability distribution maps employed in the JS calculation are shown in Fig 5.3, B.1 and B.2. The pair of CVs with the highest JS values are the ones containing χ, ζ and α. It should be noticed that JS values in AAAA for the $Z_x\chi$ and $S\chi$ pairs are in general higher than the ones corresponding to the GACC and CCCC. Analysis of the probability distributions shows that in the case of AAAA the χ angle in the A-form sub-ensemble favors the high-*anti* and *syn* conformations instead of the canonical A-form all *anti* rotamer, while in the Non-A-form group the

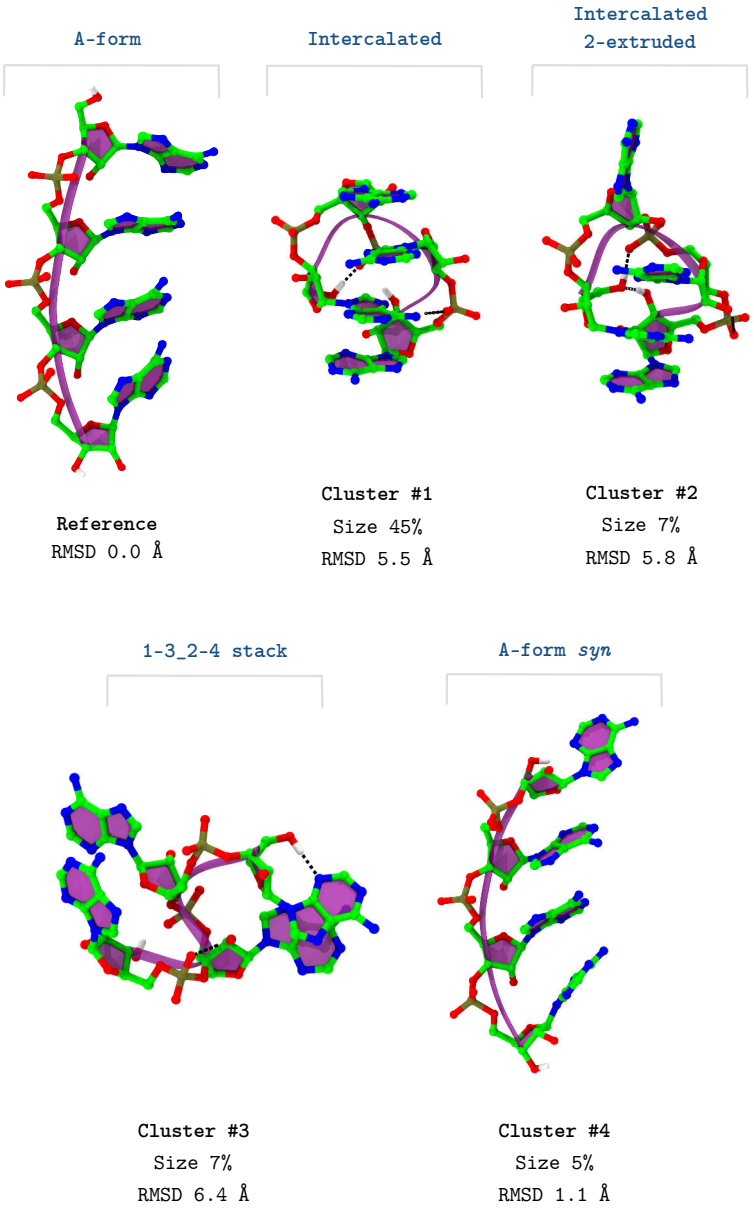


Figure 5.1: Representative clusters of the tetranucleotide Amber14 ensemble of AAAA. The clusters were calculated with the *gromos* [221] algorithm implemented in the *g_cluster* tool of Gromacs [174]. Representative structures for the CCCC and GACC can be appreciated in ref [33], for a highly sampled Amber14 ensemble.

minima are shifted to the *anti* state. For the tetranucleotides containing pyrimidines the χ angles mainly populates the *anti* state in both sub-ensembles. This result could suggest some problematic behavior of the χ angle in adenosine that should be further investigated (see ref [33] for a discussion of problematic behavior of χ angle in RNA tetraloops). A consistent trend among all tetranucleotide is related to the high JS values for $\alpha\beta$, $\epsilon\zeta$ and $\zeta\alpha$ dihedral pairs. Looking into the probability distributions it is clear that the high JS divergence values are due to a shift of the $\zeta\alpha$ minimum from the $\zeta(g^-)/\alpha(g^-)$ (the one corresponding to the right-handed helix) in the A-form sub ensemble, to the $\zeta(g^+)/\alpha(g^+)$ conformation in the Non-A-form. Each ζ/α minimum seems to be characteristic of each sub-ensemble independently of the RNA sequence,

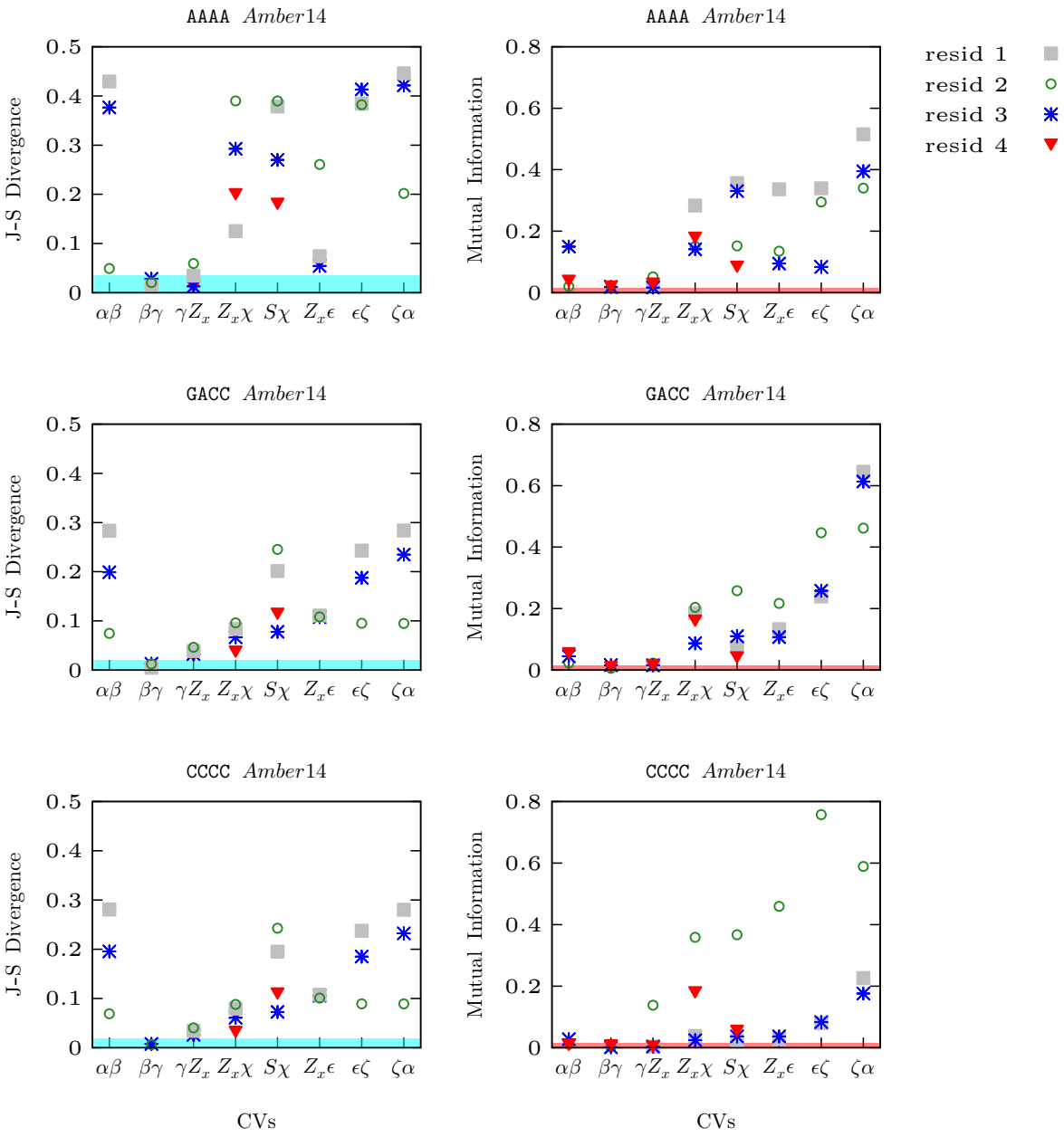


Figure 5.2: Jensen-Shannon (JS) divergence and Mutual Information (MI) calculated from the joint probability distributions of the CV_1CV_2 pairs indicated in the labels. The distributions were estimated from the tetranucleotides Amber14 ensembles taken from ref [192]. The Amber14 ensemble was divided into two groups, containing the A-form-like and Non-A-form structures respectively. The Jensen-Shannon divergence measures the difference between the bidimensional CV_1CV_2 probability distributions from the A-form-like and Non-A-form structures. The Mutual Information was calculated for the full Amber14 ensemble. MI indicates the correlation between the CV_1 with respect to the CV_2. The shaded areas represent the JS and MI values obtained for a random generated set of data of the same size of the Amber14 ensemble. Those values differ from zero due to the finite size of the datasets. The significance of the calculated JS and MI values is proportional to their distance from the shaded area.

while χ and stacking are very system dependent. Taking these results into account we hypothesize that changing the stability of the *gauche* minima in ζ/α can improve the experimental agreement of the force field, as most conformations compatible with the canonical A-form extended structure will be favored, despite the properties of the

sugar-base domain not been changed.

The ζ and α torsion angles are highly correlated between each other and with the ϵ and β angles respectively, as appreciated in Fig 5.2. Therefore, we assume that any modification on the phosphodiester backbone should include also the adjacent torsions.

The probability distributions of the RNA backbone angles obtained from the PDB can be a good reference to correct the Amber14 force field, as long as those distributions are compatible with the solution RNA ensemble at room temperature. In order to analyze the suitability of the PDB distributions, we used solution NMR data of RNA dinucleosides as a reference. RNA dinucleoside monophosphates can be considered as the smallest structural unit of the RNA that includes all the major conformational degrees of freedom. Thanks to their small size, converged ensembles are easily generated using enhanced sampling simulations. Moreover, taking fragments of dinucleotides from the RNA X-ray structures, instead of tetranucleotides, improves considerably the statistics. In Fig B.3 the agreement between experimental and calculated ^3J scalar couplings for the dinucleosides is shown. For the X-ray ensemble in general the agreement with the scalar couplings of the backbone angles (ϵ and β) is better than that of angles of the ribose-nucleobase region (χ, Z_x and γ). The disagreement in the last region is expected considering the X-ray ensemble is biased to the *anti* and $C_{3'}$-*endo* states, which predominate on the double helical structures. Compared to the force field performance, the ^3J RMSD of the PDB fragments is at least 0.5 *Hz* lower for the backbone angles. All the calculated and experimental ^3J scalar couplings used are presented in Table B.1.

We decided then to enforce the X-ray distributions of α, β, ϵ, and ς dihedral angles in the Amber14 force field, using concurrent Target Metadynamics simulations.

5.3.2 Calculation of correcting potentials

The Amber14 force field is considered to be one of the most accurate ones for RNA, though it is failing to reproduce solution experiments for short flexible oligomers. Recent benchmarks of different AMBER force field modifications based on reparametrization of the torsion angles and non-bonded terms have shown that these changes did not lead to a satisfactory agreement with solution experiments for tetranucleotides [33, 34]. On the other hand, ensembles of tetranucleotides taken from the PDB have a very good agreement with NMR data [192]. We thus decided to add correcting potentials to the dihedral angle terms of Amber14, based on information recovered from high-resolution X-ray structures of RNA deposited in the PDB. The probability distributions obtained from fragments of X-ray structures were enforced on the backbone dihedrals with T-MetaD. RNA dinucleoside monophosphates were chosen as model systems to obtain the correcting potentials. As the corrections are sequence dependent, for each nucleobase combination we generated an ensemble of experimental conformations from the PDB database that had the same sequence as the dinucleoside monophosphates.

In Fig. 5.4 we show the free energy profiles of AA and CC dinucleosides projected on

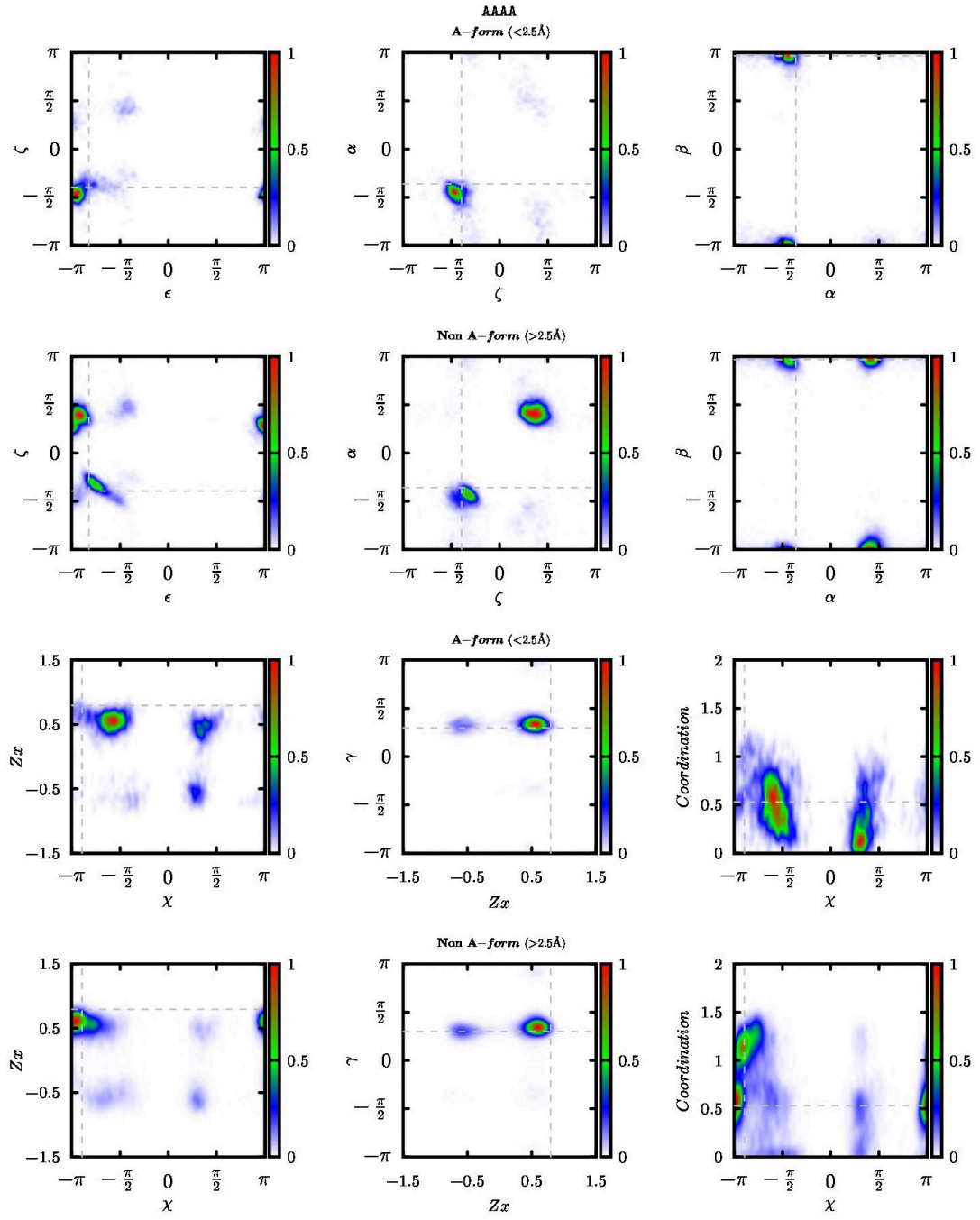


Figure 5.3: Probability distributions of dihedral angles (backbone, puckering and glycosidic angle) and coordination number of the nucleobases center-of-mass distance. These CVs values corresponding to the RNA canonical A-form are marked with a gray dashed line. The probability distributions were calculated from the AAAA T-REMD simulations.[192] The distributions marked as "A-form" includes the structures with a distance RMSD $\leq$ 2.5 Å to the canonical A-form conformation in the Amber14 ensemble, while the "Non A-form" group contains the rest, mostly compact and highly stacked structures.

the ϵ, ζ, α and β angles. Amber14, $\text{Amber}_{\text{pdb}}$, as well as the target PDB ensembles are represented. The profiles of AC and CA are shown in Fig B.3. The similarity between the PDB and $\text{Amber}_{\text{pdb}}$ profiles makes it clear that the corrections efficiently enforce the distributions taken from the X-ray ensemble. Although some differences are visible

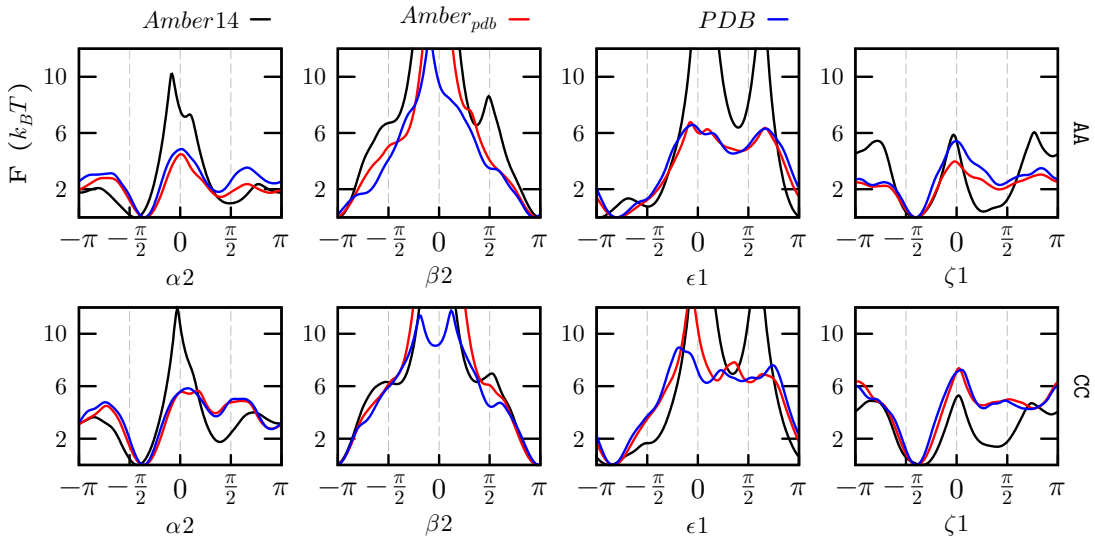


Figure 5.4: Free-energy profiles of backbone dihedral angles for the AA and CC dinucleosides monophosphates from the X-ray ensemble (PDB) and the RECT simulations with the standard force-field (Amber14) and the correcting potential (Amber$_{\text{pdb}}$).

around the free-energy barriers, they are expected not to be relevant for room temperature properties at equilibrium. Nevertheless, the transition times and the behavior of the Amber$_{\text{pdb}}$ potential at high temperatures could be affected by these barriers. In general, barriers in the experimental ensemble are several k_bT lower than those from the Amber14 force field. In the corrected ensemble the multimodal character of the force field probability distributions for the angles ϵ, ζ and α is reduced, to favor the conformations corresponding to the canonical A-form. The observed agreement between the PDB and Amber$_{\text{pdb}}$ one-dimensional probability distributions for the selected angles is not necessarily translated into equivalence of the respective ensembles. This is seen for example in the two-dimensional distributions shown in Figs B.4-B.7.

Correcting potentials might in principle also affect the distribution of non-biased degrees of freedom if the latter ones are correlated with the former ones. The distribution of non-biased degrees of freedom, such as the angles γ, χ and puckering coordinate Z_x, is shown in Fig. B.8. Overall, no difference is observed between the Amber14 and Amber$_{\text{pdb}}$ free-energy profiles, with the exception of the ratio between the C3'-*endo* and C2'-*endo* conformations in CC. This is a consequence of the significant correlation between the backbone angle ϵ and the puckering.

To asses the validity of the corrections, we compared all the ensembles against NMR experimental data [185] (Fig 5.5). Individual ^3J vicinal coupling values from the experiments and the simulations are reported in Table B.1. In the case of AA, AC and CA dinucleosides the agreement of Amber$_{\text{pdb}}$ with the experimental data is better than that of Amber14 and of the X-ray ensemble. This can be explained noticing that Amber$_{\text{pdb}}$ combines the good agreement with NMR experiments of Amber14 for angles in the nucleoside (dihedrals γ, ν_3 and χ) with that of the PDB distribution for angles in the backbone (dihedrals ϵ and β), as shown in Fig B.9. A notable exception is the CC

dinucleoside, where the correlation of backbone angles with puckering mentioned above leads to slightly larger deviation in $\text{Amber}_{\text{pdb}}$ with respect to Amber14. It should be noticed that the NMR observables analyzed here cannot be used to directly determine the conformation around the phosphodiester backbone (α/ζ), so the comparison with the NMR ^3J vicinal coupling dataset does not take intoaccount the distribution of these angles.

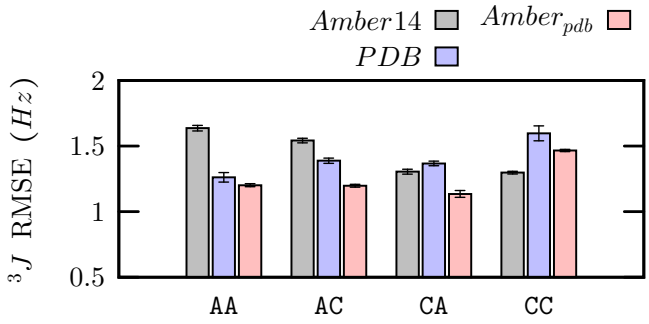


Figure 5.5: Agreement with the NMR ^3J vicinal coupling dataset of dinucleosides, measured using the root mean square error (RMSE), for the ensembles of X-ray structures (PDB), the AMBER force field (Amber14) and the corrected force field ($\text{Amber}_{\text{pdb}}$). Statistical errors were calculated using block averaging.

We noticed that, whereas the NMR data was measured at 293 K (AA, CA and AC) and 320 K (CC), simulations were performed at 300 K. However, the agreement between the data for CC obtained at 320K and similar NMR data obtained for a smaller number of couplings at 280K [202] shows that deviations induced by temperature changes are expected to be much smaller than the typical deviations between molecular dynamics and experiment observed here. It is also important to mention that these RMSE values do not take into account systematic errors in the Karplus formulas employed in this study.

It is also interesting to measure the effect of the proposed backbone corrections on the stacking interactions. Stacking free energies computed according to the definition used in a recent paper [34] show that the correcting potential have barely no effect on stacking (Fig B.10). These numbers can also be compared with experimental values [201, 202, 222], and indicate that AMBER force field is likely overestimating stacking interactions as suggested by several authors [26, 220]. This comparison is however affected by the definition of stacked conformation, which introduces a large arbitrariness in the estimation of stacking free energies from MD.

5.3.3 Validation of $\text{Amber}_{\text{pdb}}$ potential on RNA tetranucleotides

The correcting potentials discussed above are designed so as to enforce the PDB distribution on dinucleosides monophosphates. We here used these corrections to perform simulations on larger oligonucleotides. In particular, we performed extensive simulations of tetranucleotides, which are considered as good benchmarks for force-field testing, as their small size makes the generation of converged ensembles accessible to

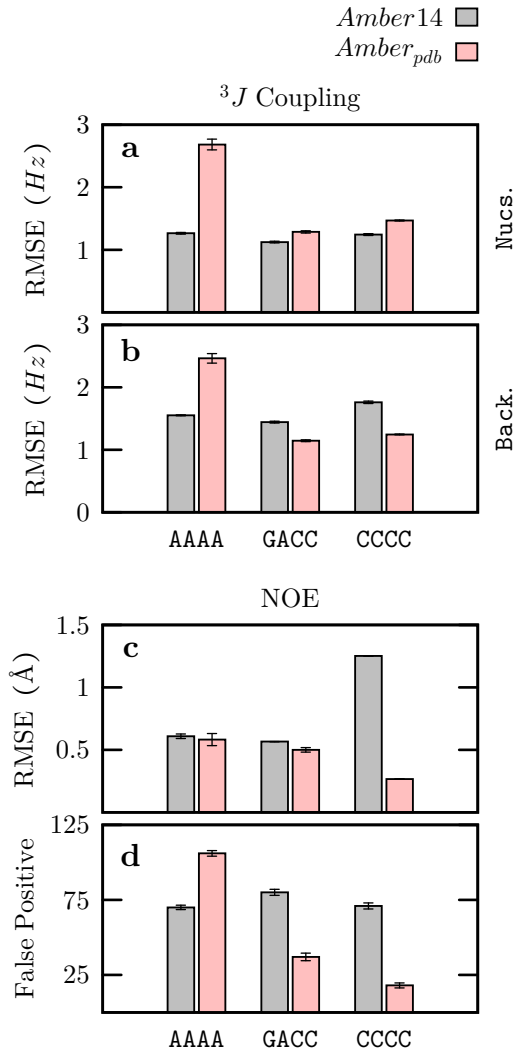


Figure 5.6: Agreement with the experimental ^3J vicinal couplings and NOE distances of tetranucleotides. For the calculation of the ^3J RMSE the RNA torsion angles were divided in two groups: **a**) the dihedral angles in the ribose-ring region (χ, ν and γ) and **b**) the phosphate-backbone angles (ϵ, ζ, α and β). In **c**) the RMSE between calculated and predicted average NOE distances is presented and in **d**) it is shown the number of false positives, i.e. the predicted distances below 5 Å not observed in the experimental data.

modern enhanced sampling techniques. We performed three T-REMD simulations with the $\text{Amber}_{\text{pdb}}$ potential for the tetranucleotide sequences AAAA, GACC and CCCC. These systems have been used before in very long (hundred of μs) simulations [31–33, 120, 128] and NMR experimental data is available [34, 166, 203]. The Amber14 T-REMD data were taken from ref [192].

The ^3J coupling RMSE, the NOE-distance RMSE, and the number of distance false positives, i.e. the MD predicted NOEs not observed in the experiment, are presented in Fig 5.6. For these systems the number of false positives is one of the most important parameters to assess the quality of the MD ensembles [34]. In the case of tetranucleotides containing pyrimidines (GACC and CCCC), the correcting potential improves significantly the agreement with the experimental data, mostly for the NOEs

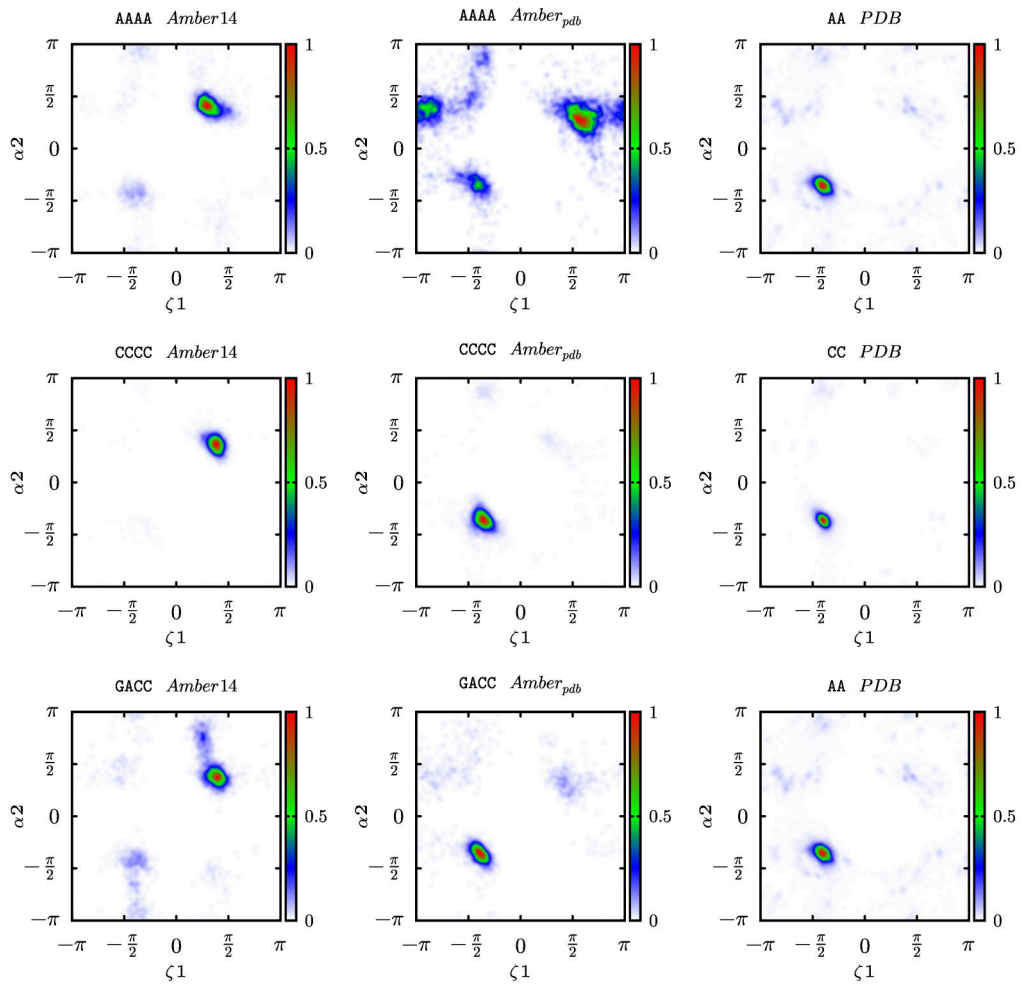


Figure 5.7: Probability distributions of the backbone dihedral angles of AAAA and CCCC tetranucleotides, in the region between residue 1 and 2. Results from the RECT simulations with the standard force-field (Amber14), the correcting potential (Amber$_{\mathrm{pdb}}$) and the dinucleoside X-ray ensembles (PDB) used to generate the correcting potentials.

(see Fig B.11). This is confirmed by the root-mean-square deviation (RMSD) distribution shown in Figure 5.8 where it can be appreciated that for these two sequences the corrections lead to an overall improvement of the ensemble by disfavoring the intercalated and inverted structures with a large RMSD from native. A completely different scenario is found for the Amber$_{\mathrm{pdb}}$ ensemble of AAAA, where the corrections surprisingly diminish the agreement with experiments. This can be also appreciated in a shift of the Amber$_{\mathrm{pdb}}$ RMSD distribution peaks to higher RMSD values due to an increased population of compact structures (Fig 5.8). It should be noticed that the effect of the correcting potentials in purines and pyrimidines depends strongly on the sequence length. Whereas the AAAA tetranucleotide is negatively affected by the corrections, the AA dinucleoside is the one that benefits the most from them.

As discussed in the section 5.3.1, the conformation along the phosphodiester backbone is very different between compact and extended tetranucleotide structures. The

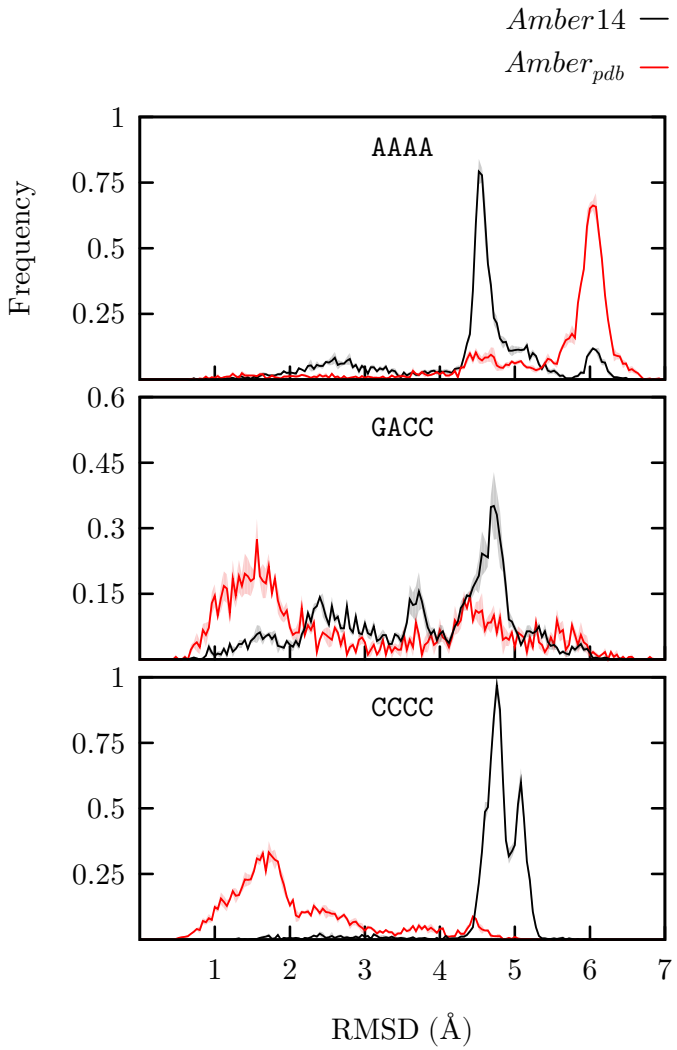


Figure 5.8: Empirical probability distribution of heavy atom RMSD from the canonical A-form as computed for the reference replica trajectory. Distributions are shown for the REMD simulations with Amber14 (black) and $\text{Amber}_{\text{pdb}}$(red). The total probability is shown in solid line and the above and below limits determined by the blocking error are shaded. It can be appreciated that the correcting potentials increase the population of extended structures (RMSD $\sim$ 1-2 Å) for the CCCC and GACC tetranucleotides, while for AAAA the $\text{Amber}_{\text{pdb}}$ ensemble is farther from the canonical A-form.

probability distribution maps of the α_2/ζ_1 backbone dihedral angles from the tetranucleotides T-REMD simulations and the dinucleosides X-ray ensembles used to generate the corrections are depicted in Fig B.1. Only phosphodiester backbone torsion angles are shown, because they are the ones mostly affected by the correction. The other backbone angles maps are shown in the Appendix B (Figs B.12-B.20). In the PDB ensembles the distributions are always unimodal, independently of the sequence, with a peak at the $\alpha(g^-)/\zeta(g^-)$ conformation, whereas in the Amber14 ensemble the $\alpha(g^+)/\zeta(g^+)$ and $\alpha(g^-)/\zeta(g^-)$ conformations are both significantly populated. The effects of the corrections, as seen before, are highly sequence dependent. In case of GACC and CCCC, the $\alpha(g^-)/\zeta(g^-)$ rotamer is stabilized in the $\text{Amber}_{\text{pdb}}$ distributions, with

the population of $\alpha(g^+)/\zeta(g^+)$ significantly decreased with respect to Amber14. On the contrary, for AAAA the $\alpha(g^+)/\zeta(g^+)$ conformation is not unfavored by the correcting potentials, despite not being significantly present in the PDB ensemble. This could be due to the fact that the one dimensional target free-energy profile for dihedrals α and ζ for the AA (Fig 5.4) exhibits barriers which are approximately 4 k_bT smaller with respect to the ones from the Amber14 force field. The effect of the decreased barrier height can be appreciated in the α_2/ζ_1 probability distribution of AAAA, where the amount of torsional space explored is increased by the corrections.

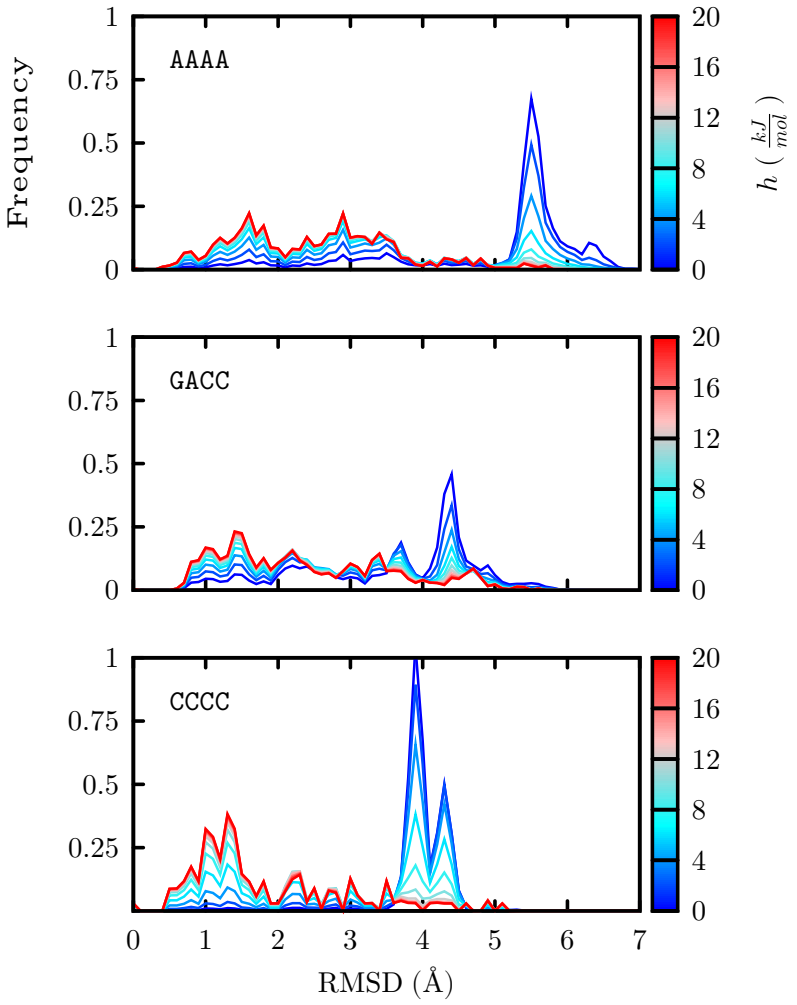


Figure 5.9: Empirical probability distribution of heavy atom RMSD from the canonical A-form computed for the reweighted Amber14 ensemble as a function of the Gaussian potential height For all sequences the increase of the penalty potential shifts the distributions closer to the A-form structure.

5.3.4 Consequences on future force field refinements

The good agreement of the $\text{Amber}_{\text{pdb}}$ ensembles with the NMR observables, in the case of CCCC and GACC tetranucleotides, suggests that the RNA conformational space sampled by state-of-the-art force field could be modified to better match experimental

solution data by penalizing rotamers of the α and ζ angles. As a further test, we reweighted the T-REMD Amber14 ensembles with an additional two-dimensional penalizing Gaussian potential centered on the $\alpha(g^+)/\zeta(g^+)$ conformation (See Fig 5.9). Results are shown in Fig 5.10 for different Gaussian heights. Overall, the agreement with the NMR experimental data improves considerably with respect to the original force field as the Gaussian height increases. The relative population of the α/ζ conformations has an important impact on the number of false positive NOE contacts which indicates the presence of intercalated structures. This improvement is achieved without changing the non bonded interactions as it has also been proposed [26]. It is however important to observe that these results are obtained by performing a reweighting, and that corrections should be validated by performing separate simulations with this bias potential.

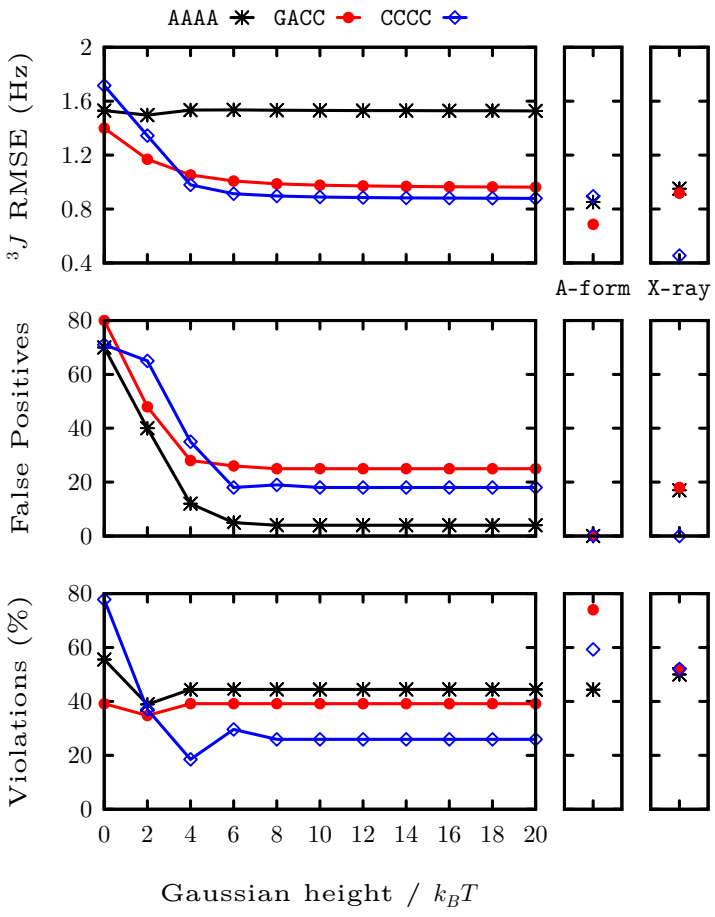


Figure 5.10: Agreement with the experimental data for the Amber14 reweighted ensemble as a function of the Gaussian potential height. The bias potential was centered on $\alpha(g^+)/\zeta(g^+)$ conformation $(\frac{\pi}{2}, \frac{\pi}{2})$ with a sigma per angle of 0.7 rad. "A-form" represent a canonical A-form structure and "X-ray" an ensemble of tetranucleotide fragments, with the same sequence, from the PDB (all taken from ref [192]).

5.4 Discussion

In this chapter we apply targeted metadynamics to sample preassigned distributions taken from experimental data [35, 36]. At variance with the original applications, we here combine T-MetaD with enhanced sampling showing that these protocols can also be used when the investigated ensembles have non-trivial energy landscapes separated by significant barriers .

We apply the method to RNA oligonucleotides, for which the Amber14 force field was proven to be in significant disagreement with solution NMR data [31, 33, 34, 120, 128, 166, 203, 214]. Since tetranucleotide fragments extracted from high resolution structures in the PDB were shown to match NMR experiments better than Amber14 force field [192], we here used X-ray structures to build reference distributions of backbone dihedral angles that are then used to devise correcting potentials. More precisely, we use T-MetaD to enforce the empirical distribution of the dihedral angles in the phosphate backbone (ϵ, α, ζ and β) on four dinucleoside monophosphates.

We calculated the correcting potentials concurrently for all the four angles in order to change the distribution of these consecutive dihedrals along the backbone chain taking into account their correlation. The method successfully enforced the distributions taken from the PDB on all the angles. The new ensemble generated by the corrected force field ($\text{Amber}_{\text{pdb}}$) was independently validated against solution NMR data that was not used in the fitting of the corrections. For three of the four dinucleosides studied, $\text{Amber}_{\text{pdb}}$ showed a better agreement with the NMR data compared with Amber14 and with the X-ray ensemble.

We then tested the portability of the correcting potentials by simulating three tetranucleotides, GACC, CCCC and AAAA. In the case of GACC and CCCC the agreement with NMR data is significantly improved by the corrections. Surprisingly, for AAAA the corrections have the opposite effect and increase the probability of visiting compact structures making the simulated ensemble less compatible with solution experiments. It should be noticed here that this is a non obvious result since the PDB database is expected to have an intrinsic bias towards A-form structures and should thus in principle increase the agreement with solution experiments in this specific case. This indicates that porting the corrections from dinucleosides to tetranucleotides is not straightforward because the coupling between the multiple corrected dihedrals could affect the resulting ensemble in an non-trivial way. Additionally, corrections applied to dihedral angles alone might be not sufficient to compensate errors arising from inexact parametrization of van der Waals or electrostatic interactions [26]. Overall, the tests we performed indicate that the corrections derived here should not be considered as portable corrections for the simulation of generic RNA sequences.

Nevertheless, by comparing the backbone angle distributions on the different RNA simulations and the X-ray ensembles, we were able to find possible hints pointing at where refinement of dihedral potentials could lead to an advancement in RNA force fields. In this respect, the results for GACC an CCCC show the significant improvement

observed in the $\text{Amber}_{\text{pdb}}$ simulations for those systems could be reproduced by simply penalizing the $\alpha(g^+)/\zeta(g^+)$ conformation, which is overpopulated in Amber14. By a straightforward reweighting procedure, we showed that simple Gaussian potentials that disfavor this conformation significantly improved the experimental agreement with solution experiments for all the three tetranucleotides. Recent modifications of the Lennard-Jones parameters for phosphate oxygens [110] and different water models [128] were shown to affect the conformational ensemble of RNA tetranucleotides [33, 128]. It might be interesting to combine these modified parameters for non-bonded interactions with the here introduced procedure for dihedral angle refinement.

The nature of the correction methodology discussed in this chapter is very different from the classical approach to force field parametrization, as it aims to correct the free energy of the system, instead of fitting the potential energy landscape of the dihedral angles while constraining the other degrees of freedom. It is important to notice that the dihedral angle distributions taken from the fragments of the PDB structures do not necessarily represent the conformational ensembles of dinucleosides or tetranucleotides in solution. Indeed, some of the interaction patterns that are present in large structures crystallized in the PDB do not exist in short oligonucleotides. For this reason, in this work the distributions were validated against independent solutions NMR experiments. This allowed the dihedral angles from the PDB distributions that performed better than the force field to be identified. We also recall that in our procedure the force-field torsion energy function is not refitted, but a bias potential is added to the total energy of the system in order to match the free-energy profile of the torsion angles with target ones. Thus, a major advantage of this approach is that it takes explicitly into account the entropic contributions, the cross correlations between torsional angles, and inaccuracies in the non-bonded interactions, among other effects.

5.5 Conclusion

In conclusion, in this work we applied the target metadynamics protocol to modify dihedral distributions in dinucleosides. The procedure successfully enforces reference distributions taken from the PDB without affecting the distribution of the dihedral angles that were not biased. However, the attempt to port these corrections to tetranucleotides lead to ambiguous results when applied to different sequences. This could be partly due to the fact that distribution form the PDB are not necessarily a good reference for refinement.

Nevertheless, the simulations revealed the importance of the α/ζ angles rotamers on the modulation of the conformational ensemble, and that by only penalizing the $\alpha(g^+)/\zeta(g^+)$ rotamer the quality of the ensemble is significantly improved to levels not reported before.

Chapter 6

Conclusions and Perspectives

In this thesis the problem of conformational sampling in MD simulations of RNA systems and the low agreement of current RNA force fields is addressed with the introduction of a novel and flexible enhancing sampling method, replica exchange with collective-variable tempering (RECT), and the calculation of correcting potentials that enforce distributions of dihedral angles taken from experimental structures. RECT takes advantage of the adaptive nature of well-tempered metadynamics to build bias potentials that compensate free-energy barriers. The results from a simulation of a single-stranded RNA tetranucleotide show this new method is a promising tool to accelerate the exploration of RNA conformational space. On the other hand, the introduction of the corrected potentials on the AMBER force field lead to a better agreement with independent solution experiments for the oligonucleotides containing pyrimidine bases, but failed for the oligomer containing only Adenosine. However, the simulations reveal that by only penalizing the $\alpha(g^+)/\zeta(g^+)$ rotamer the experimental agreement of the ensemble is significantly improved for all RNA tetranucleotide sequences.

Perspectives of this PhD thesis will be presented now. An issue with the current formulation of RECT is that the convergence of the bias potentials could take some time (tens of nanoseconds or more). To alleviate this, RECT could be modified to allow each replicas feel the bias potentials of the other replicas, in the way of Multiple Walker [223] and Altruistic Metadynamics [224]. Additionally, the geometric replica distribution used here is merely heuristic: In the replicas with high γ the Gaussian approximation for the distribution of bias potentials is satisfactory, but the behavior in the lower replicas diverges from the prediction. Finding a distribution of replicas that maintain a constant acceptance ratio across the replica ladder is not a simple task, but for the systems studied here a geometric distribution resulted in high acceptance ratios and low round-trip times. Even if the number of CV is very large, the density of replicas in RECT should not be higher than in other popular H-REMD methods, as just moderated γ factors are needed in the highest replicas to accelerate dihedral transitions. Applying RECT in larger systems could be negatively affected by the computational cost of building hundreds of WT-MetaD potentials at the same time. If that is the case, the use of multiple-time step to integrate the biasing forces [225] could

provide the necessary speed up. Moreover, if there is an *a priori* knowledge of the system, knowledge-based CVs can be included in RECT, like CVs based on εRMSD, which in nucleic acids are particularly well-suited to distinguish among conformational states [215]. RECT can be also integrated with other replica exchange methods, like parallel tempering [136] or solute tempering [146]. This combination could make a difference in the difficult task of generating converged ensembles of RNA oligomers, which demands hundreds of microseconds of simulated time and is fundamental for the evaluation of new FF parameters or *ad hoc* corrections [33].

As it was mentioned before, the empirical corrections to the AMBER FF calculated here led to ambiguous results when applied to different tetranucleotide sequences. We recognize the dihedral free-energy profiles estimated from X-ray ensembles are not completely reliable. One solution to this problem could be to apply quality filters on the X-ray ensembles to eliminate conformational errors from the experimental structures [102]. Moreover, free-energy profiles from QM/MM calculations in solution could be used to generate accurate correcting potentials, in the spirit of the QM/MM force matching approach [226]. The penalty potential suggested here for a rotameric phosphate-backbone conformation has recently been tested in RNA tetraloops, resulting in a significant improvement over the Amber14 force field [227].

Appendix A

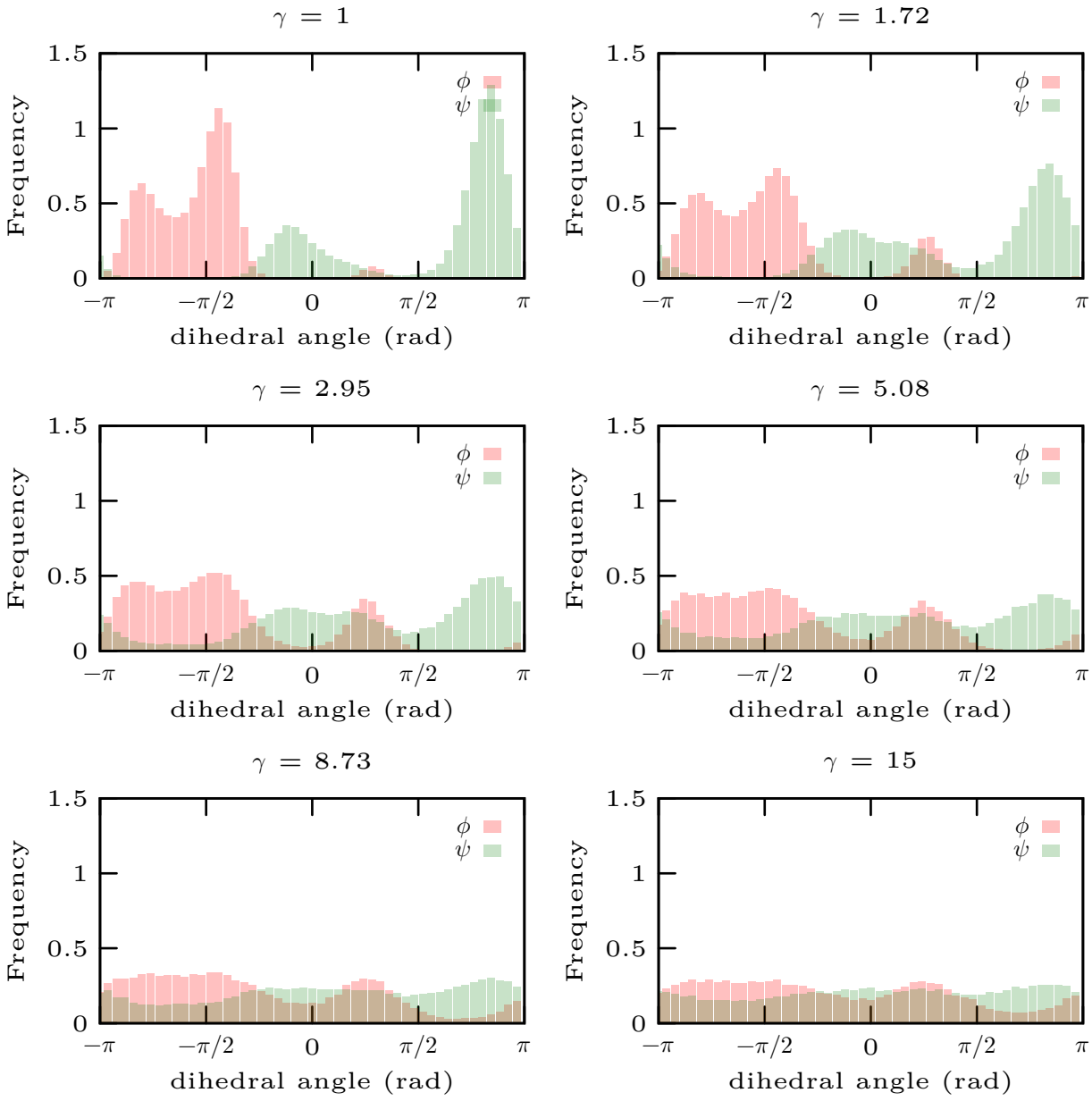


Figure A.1: Histograms of dALA Ψ and Φ dihedral angles for the RECT simulation at each replica.

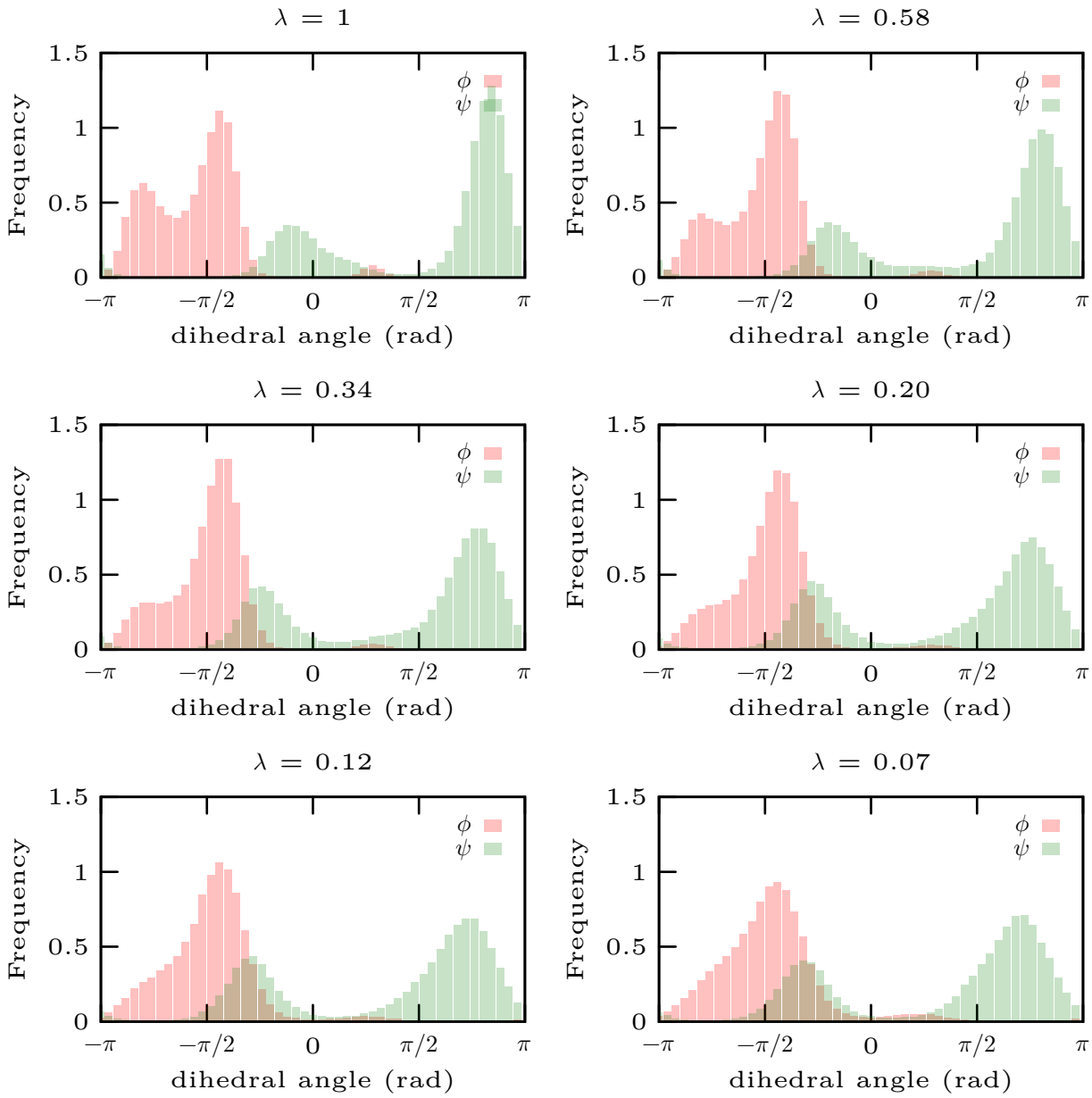
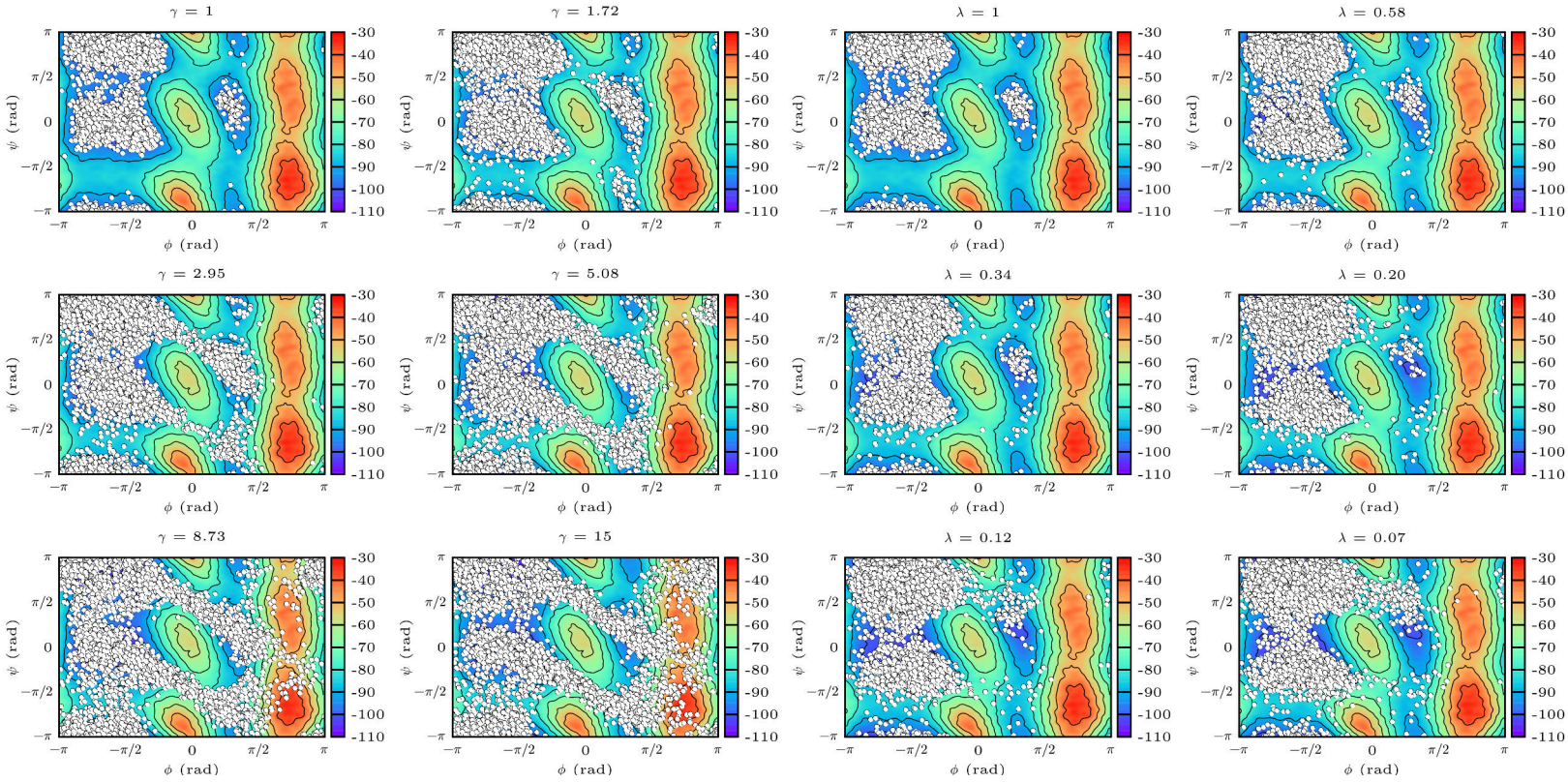
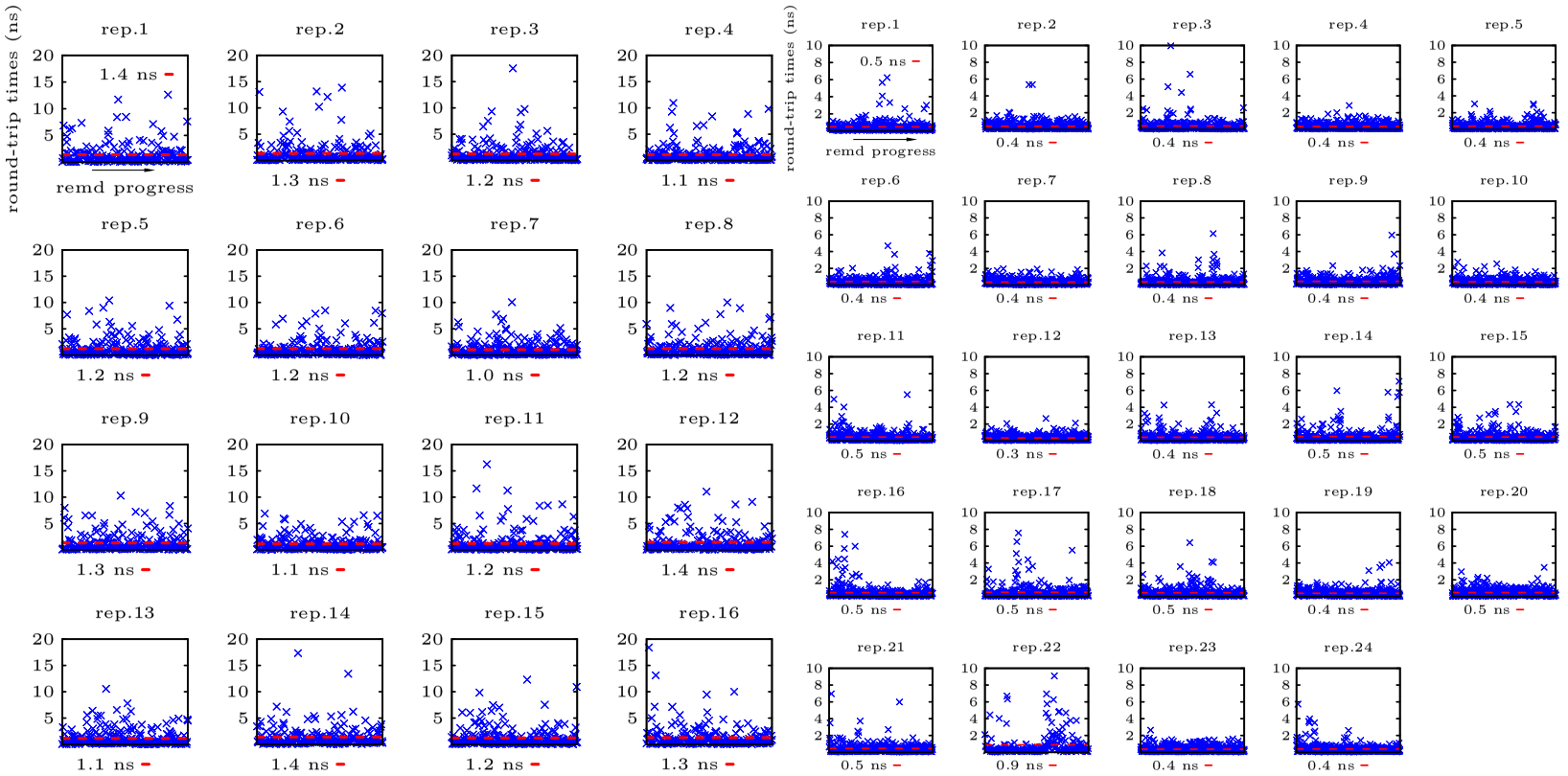


Figure A.2: Histograms of dALA Ψ and Φ dihedral angles for the H_{dih}-REMD simulation at each replica.



(a) RECT (b) H_{dih}-REMD

Figure A.3: The projection of each replica trajectory on the dihedral free-energy landscape $F(\Psi, \Phi)$ for both H-REMD methods. Although uniform exploration of $F(\Psi, \Phi)$ is not achieved, each angle is uniformly sampled.



(a) RECT (b) H_{dih}-REMD

Figure A.4: Round-trip times (rtt) of each trajectory in the generalized ensemble during the H-REMD simulations of the RNA tetranucleotide. Red lines represent the averages rtt for each set of data.

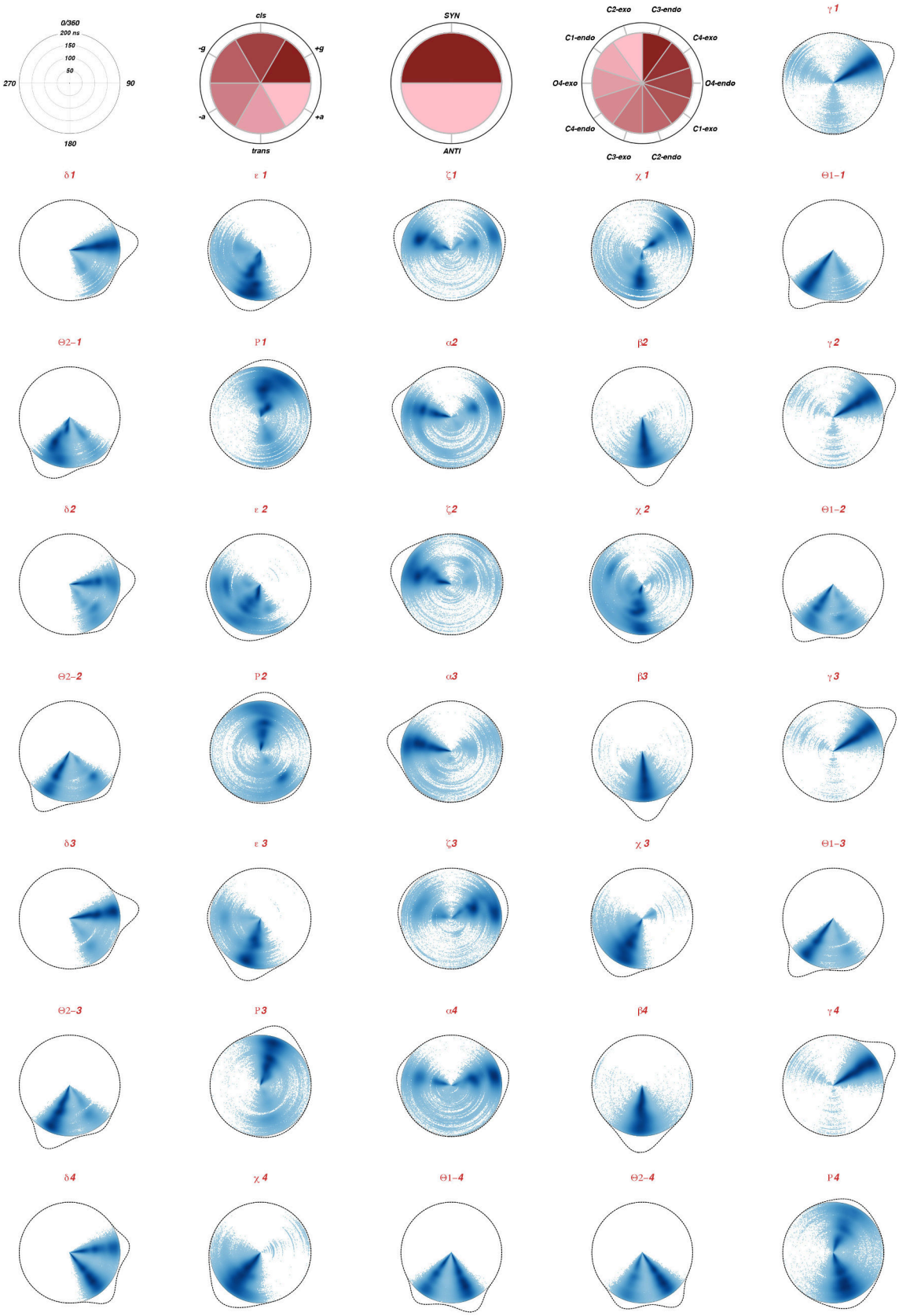


Figure A.5: Time series and histograms of the 32 torsion angles biased during the RECT simulation for the unbiased replica.

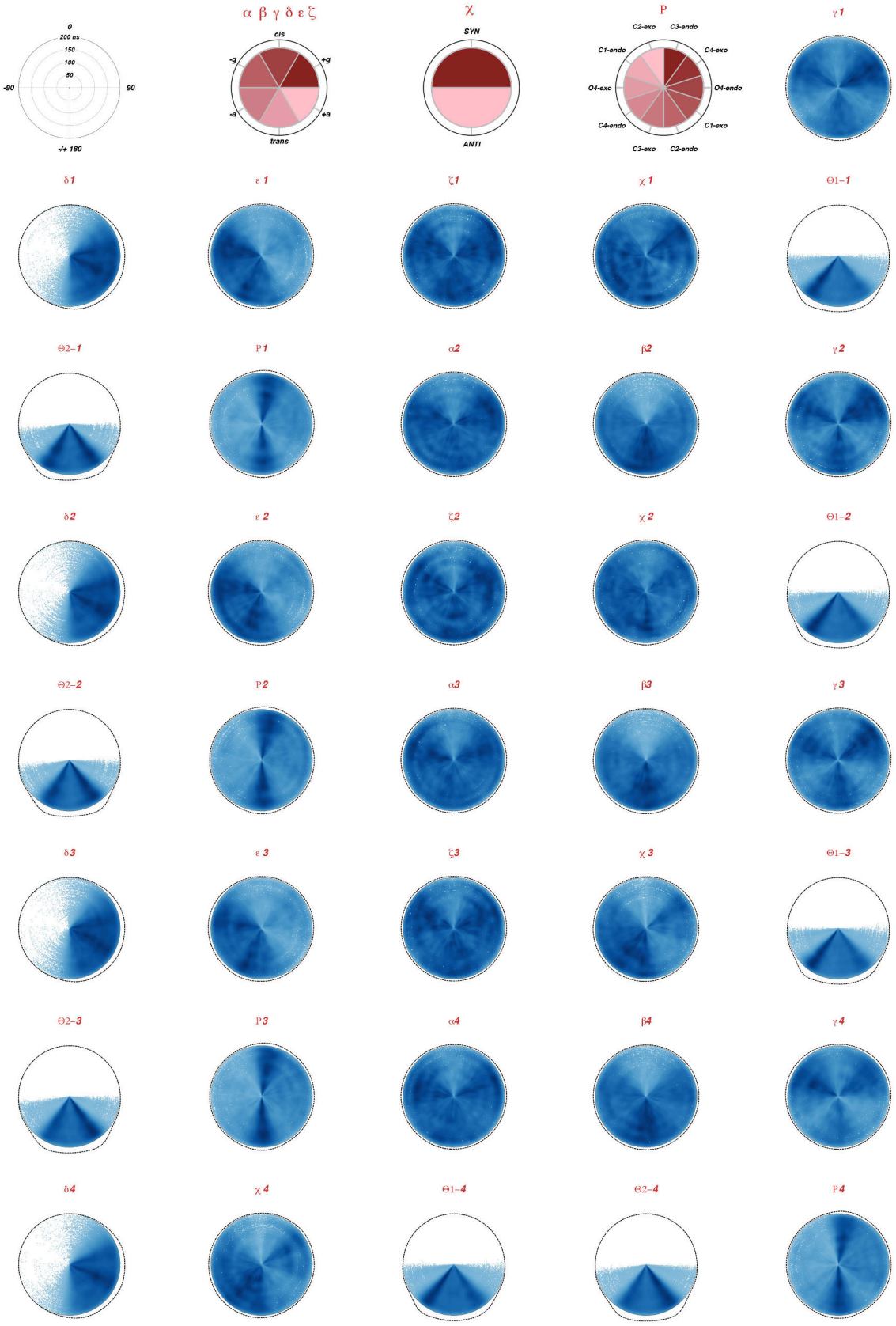


Figure A.6: Time series and histograms of the 32 torsion angles biased during the RECT simulation for replica 16.

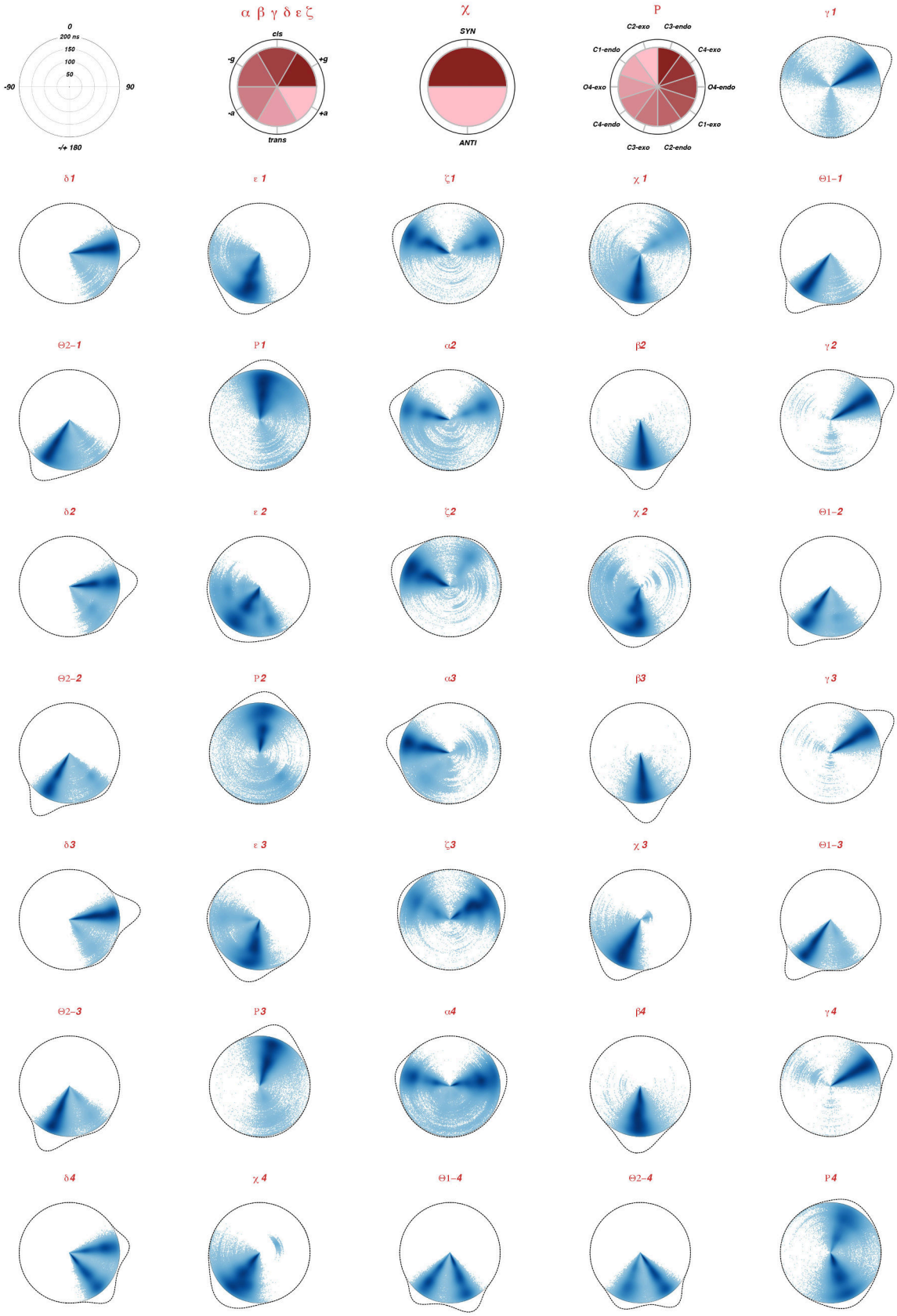


Figure A.7: Time series and histograms of the 32 torsion angles with energies scaled during the H_{dih}-REMD simulation for the unbiased replica.

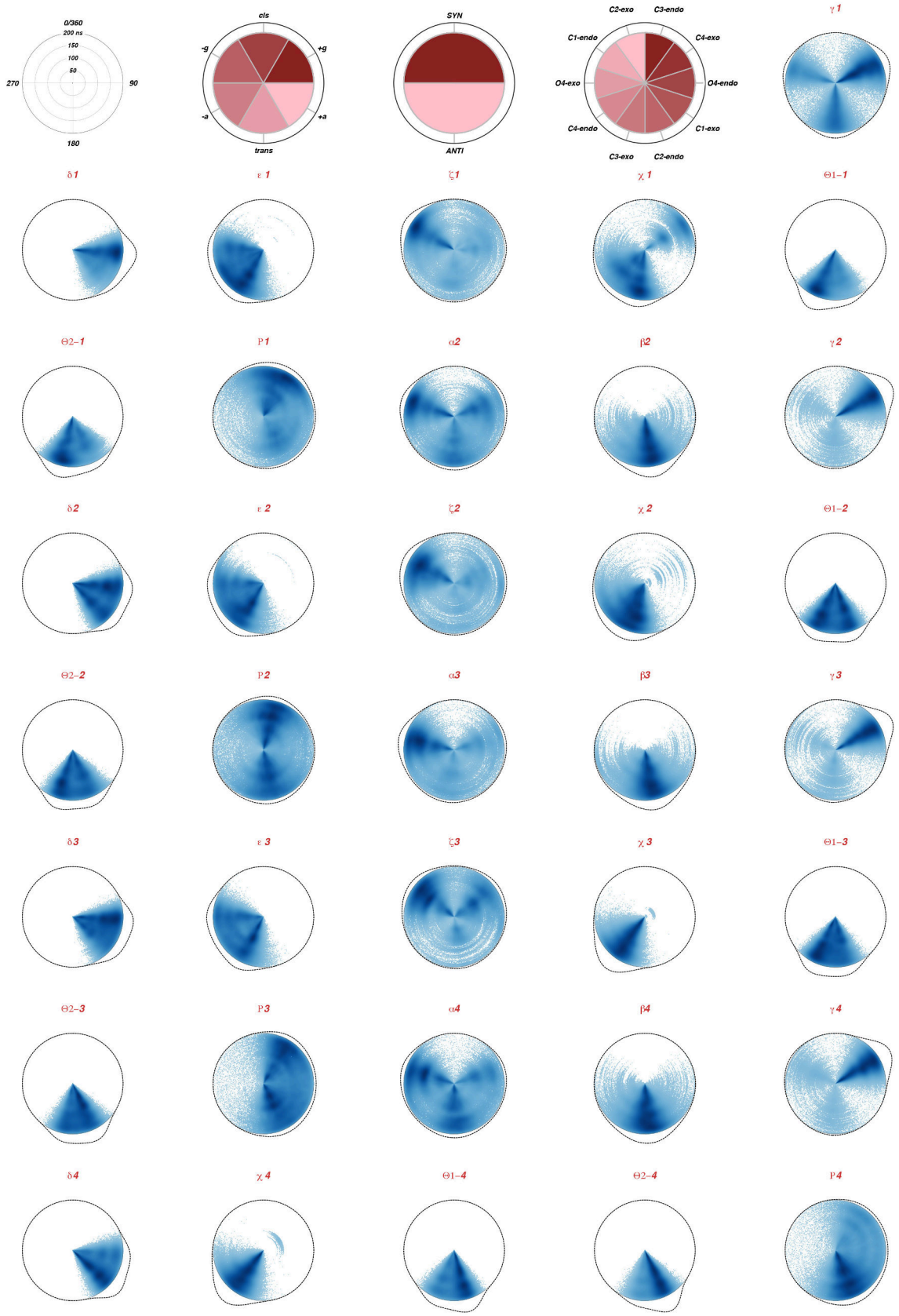


Figure A.8: Time series and histograms of the 32 torsion angles with energies scaled during the H_{dih}-REMD simulation for replica 24.

Appendix B

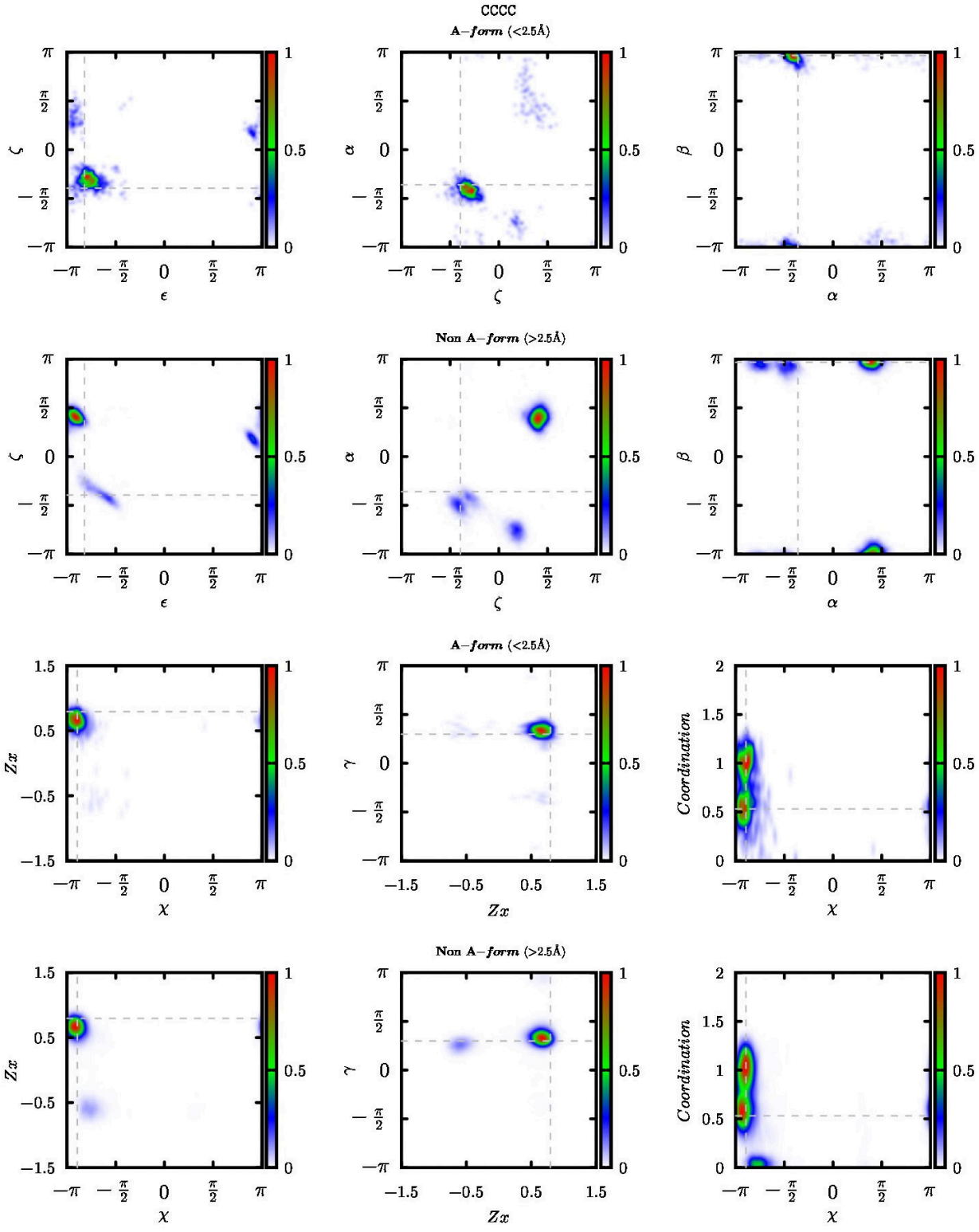


Figure B.1: Same as Fig. 5.6 but for the CCCC Amber14 ensemble.

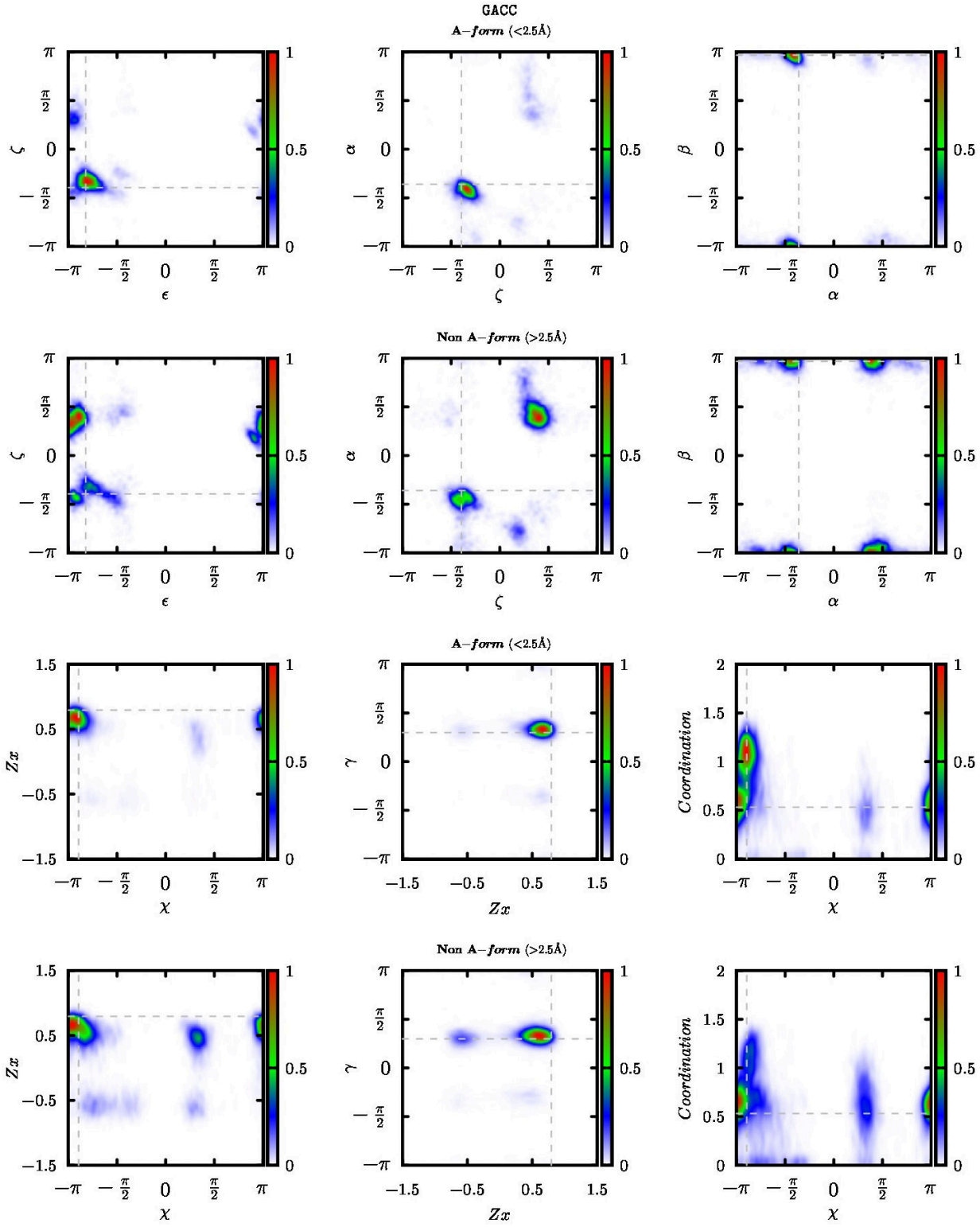


Figure B.2: Same as Fig. 5.6 but for the GACC Amber14 ensemble.

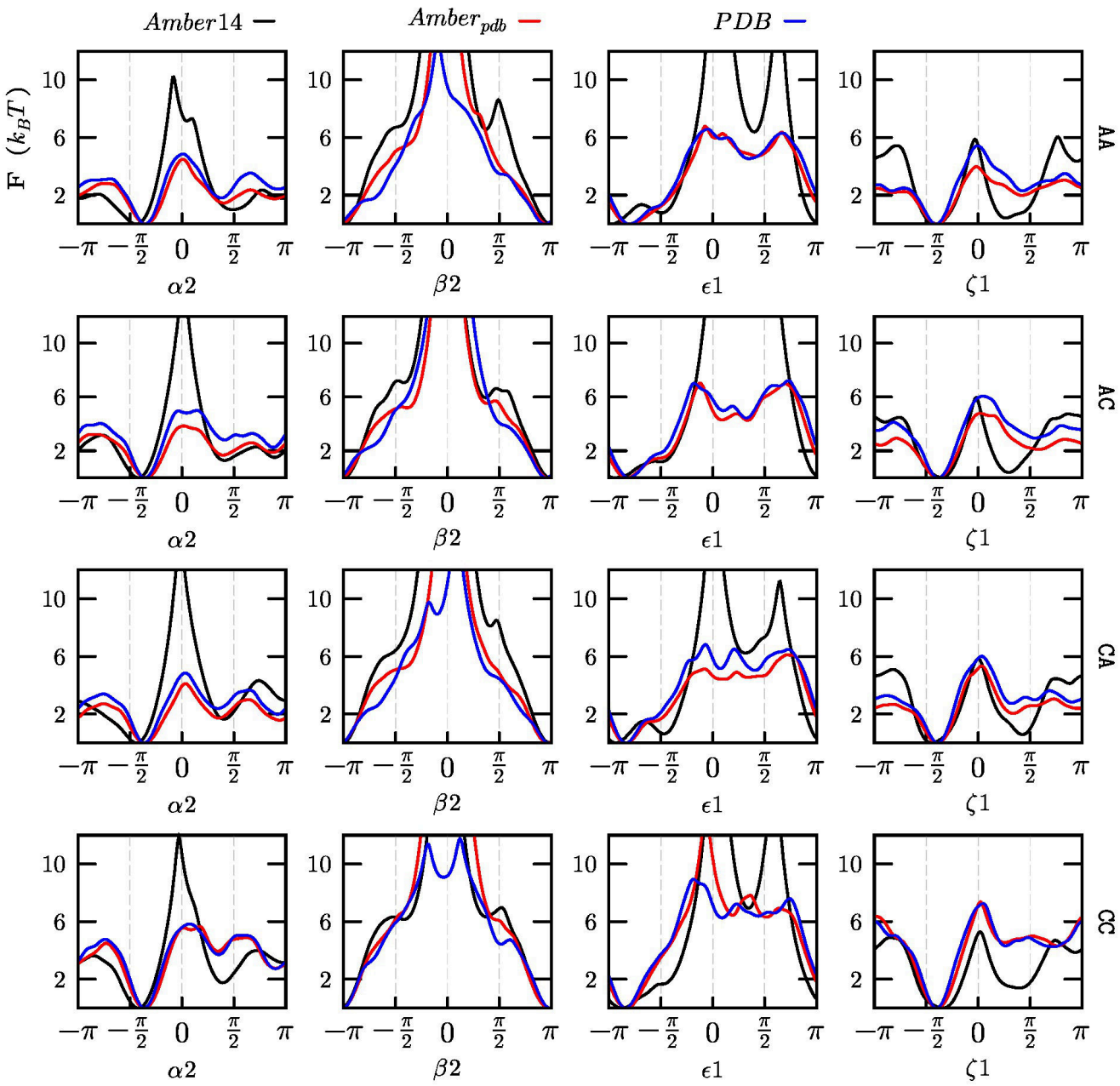


Figure B.3: Free-energy profiles of backbone dihedral angles for all the dinucleosides monophosphates studied here, from the X-ray ensemble (PDB) and the RECT simulations with the standard force-field (Amber14) and the correcting potential (Amber$_{\mathrm{pdb}}$).

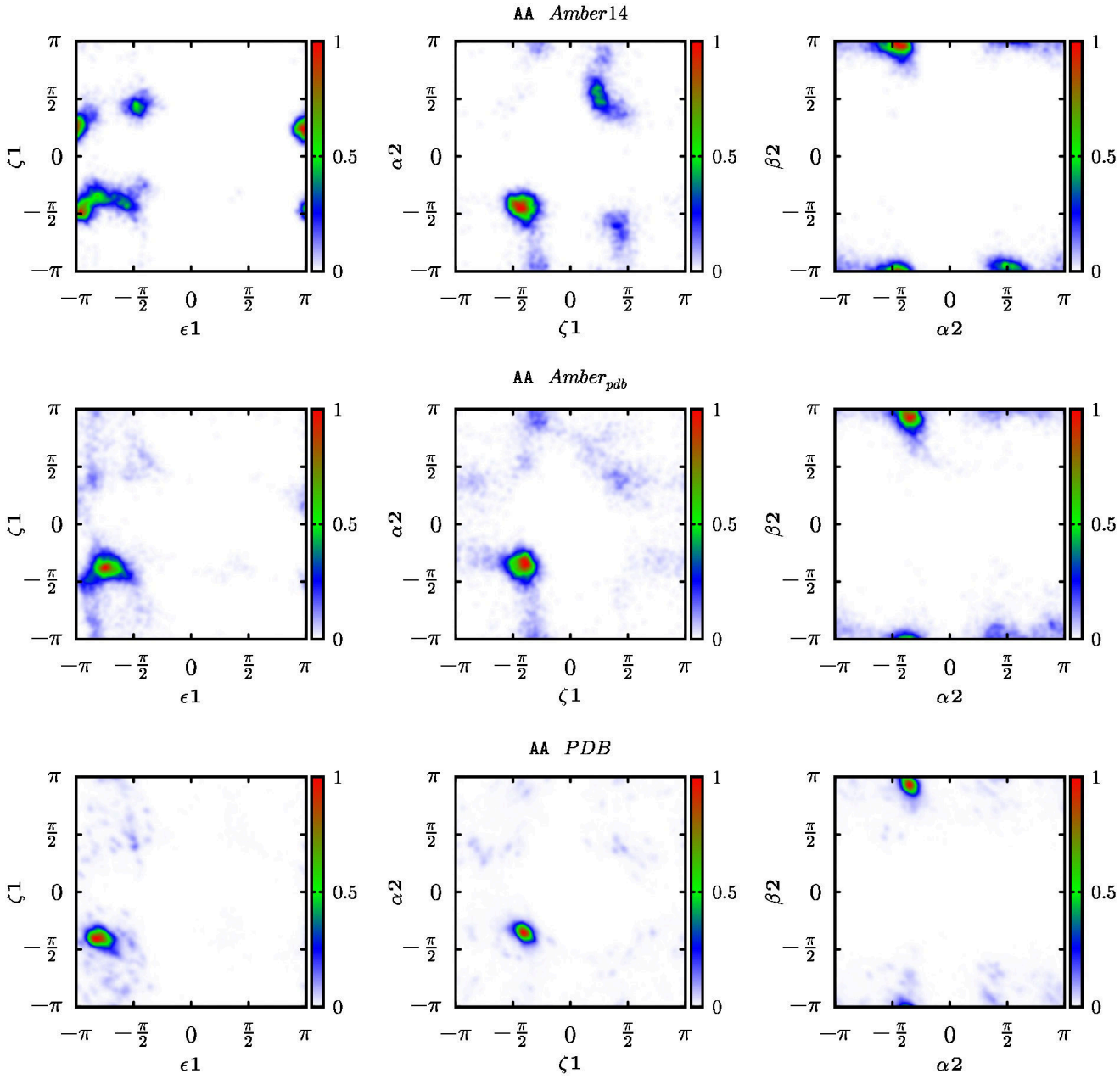


Figure B.4: Probability distributions of the backbone dihedral angles of AA dinucleoside monophosphate, from the X-ray ensemble (PDB) and the RECT simulations with the standard force-field (Amber14) and the correcting potential ($\text{Amber}_{\text{pdb}}$).

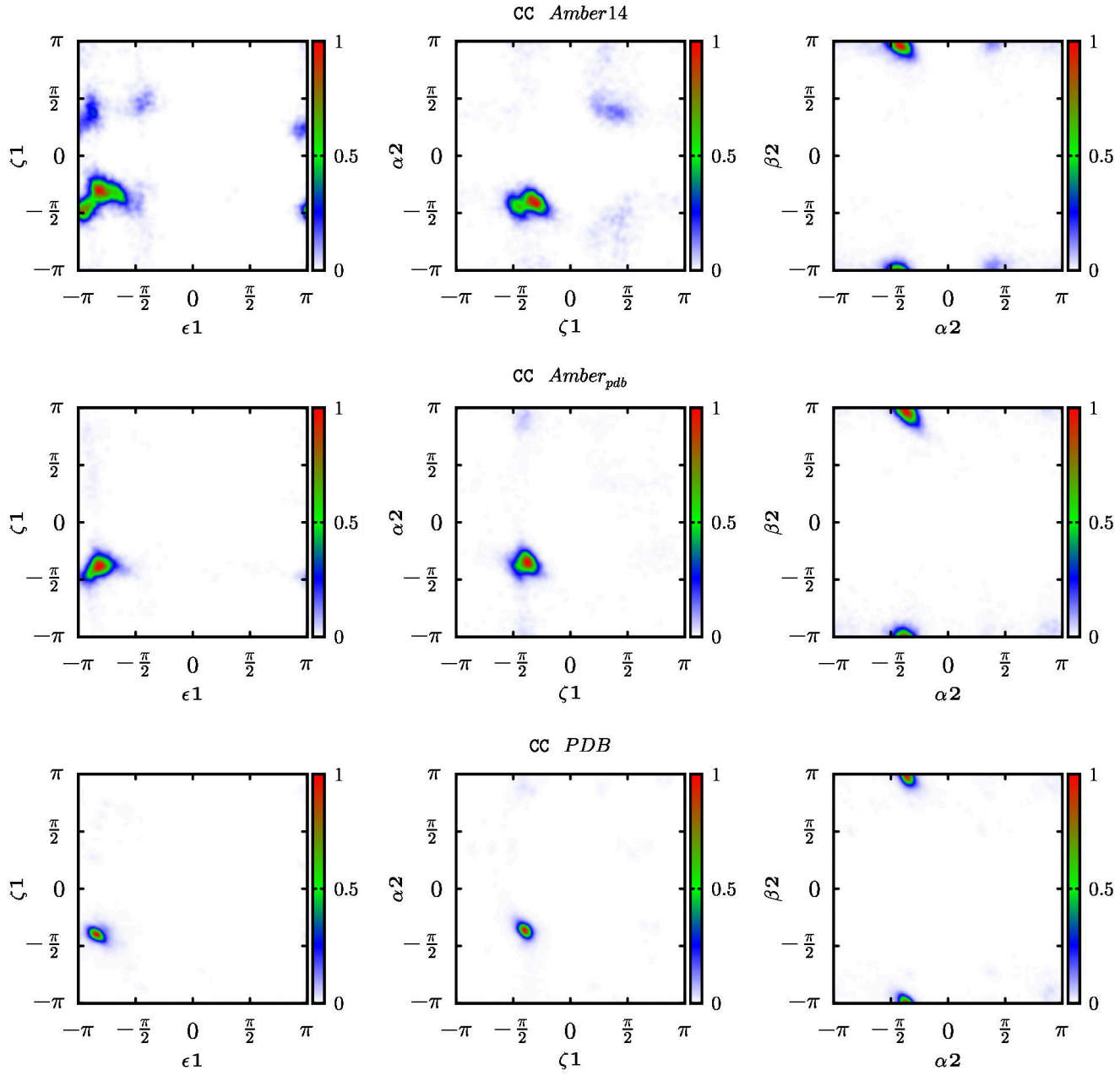


Figure B.5: Probability distributions of the backbone dihedral angles of CC dinucleoside monophosphate, from the X-ray ensemble (PDB) and the RECT simulations with the standard force-field (Amber14) and the correcting potential ($\text{Amber}_{\text{pdb}}$).

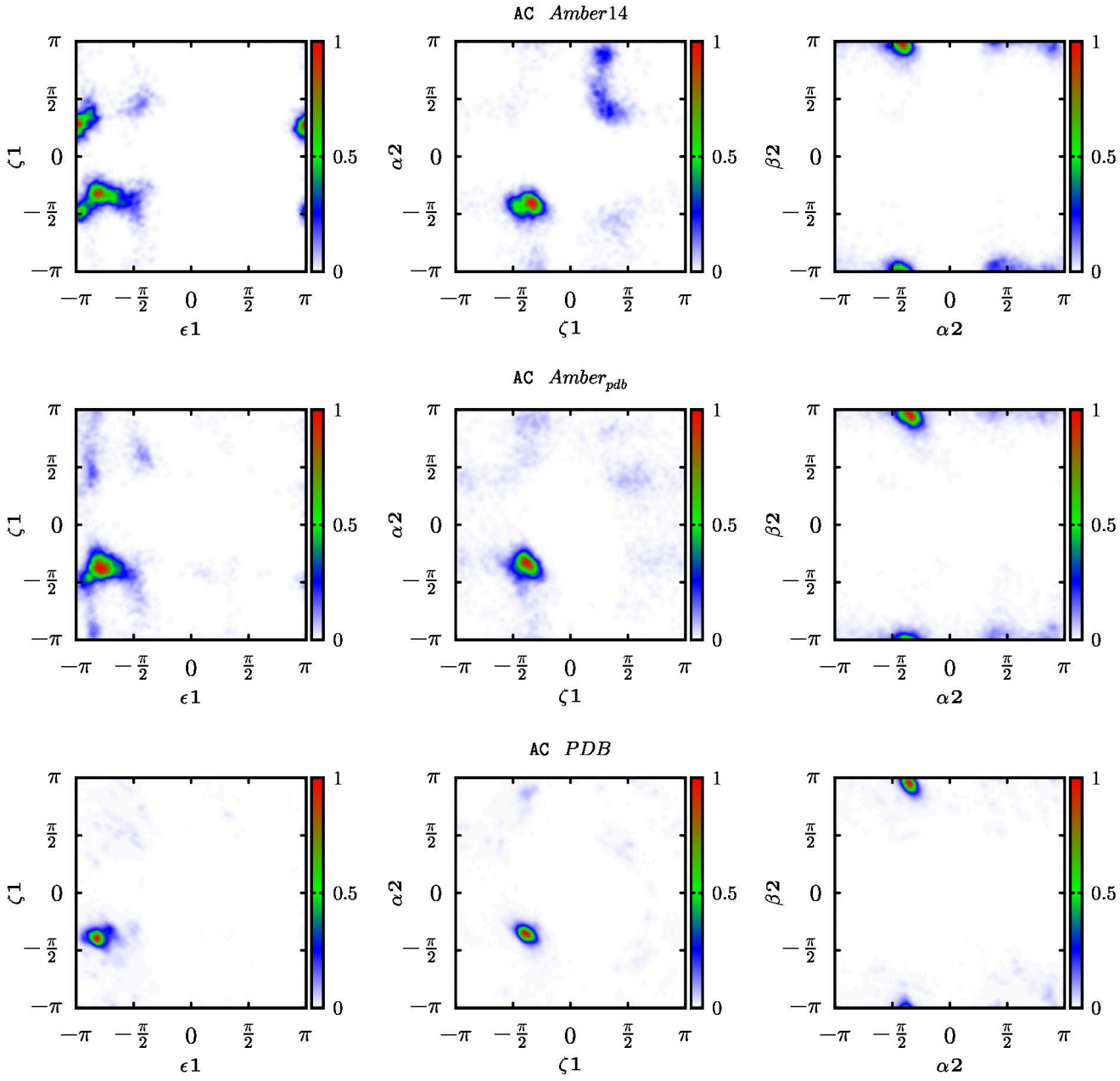


Figure B.6: Probability distributions of the backbone dihedral angles of AC dinucleoside monophosphate, from the X-ray ensemble (PDB) and the RECT simulations with the standard force-field (Amber14) and the correcting potential ($\text{Amber}_{\text{pdb}}$).

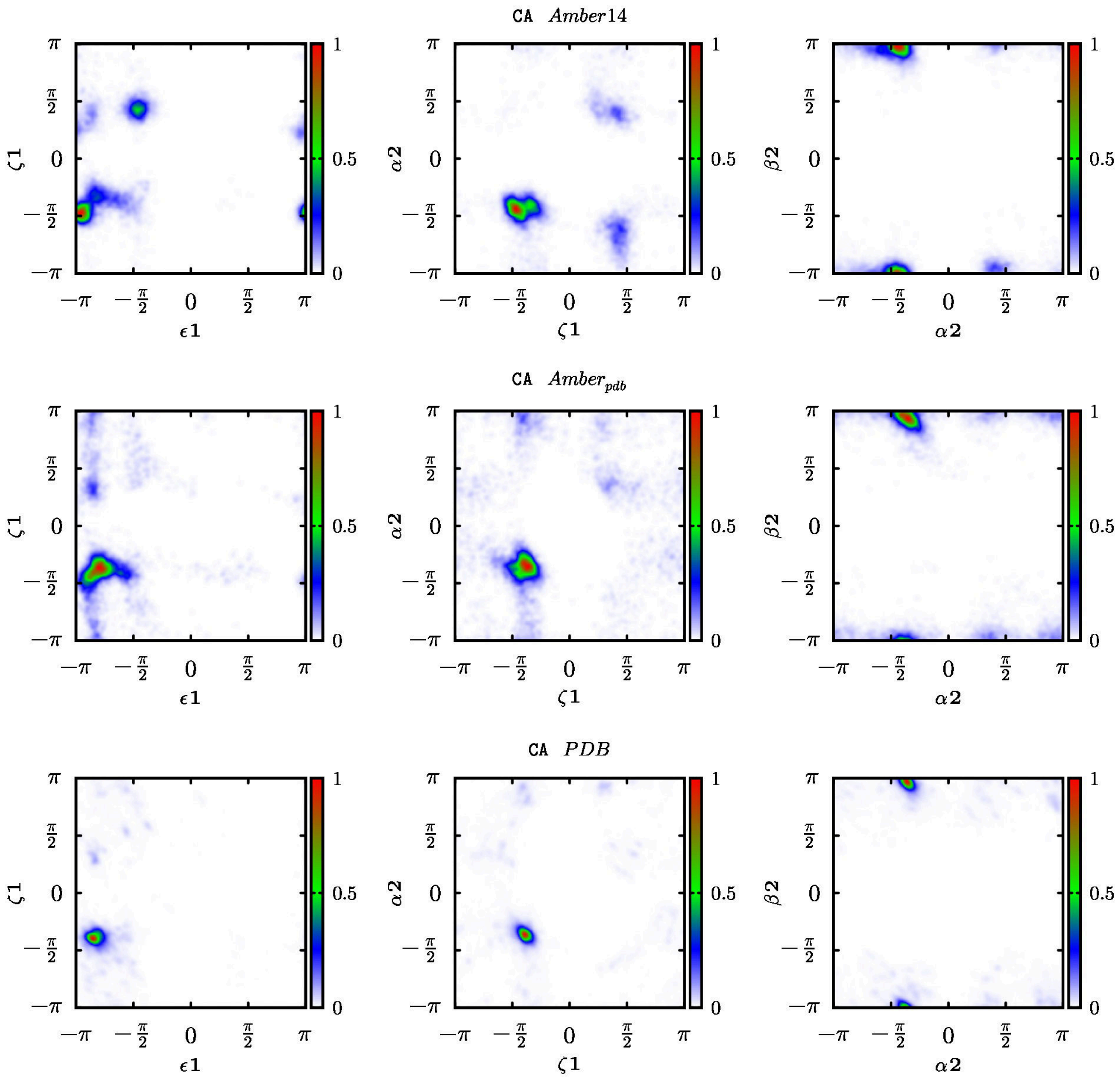


Figure B.7: Probability distributions of the backbone dihedral angles of CA dinucleoside monophosphate, from the X-ray ensemble (PDB) and the RECT simulations with the standard force-field (Amber14) and the correcting potential ($\text{Amber}_{\text{pdb}}$).

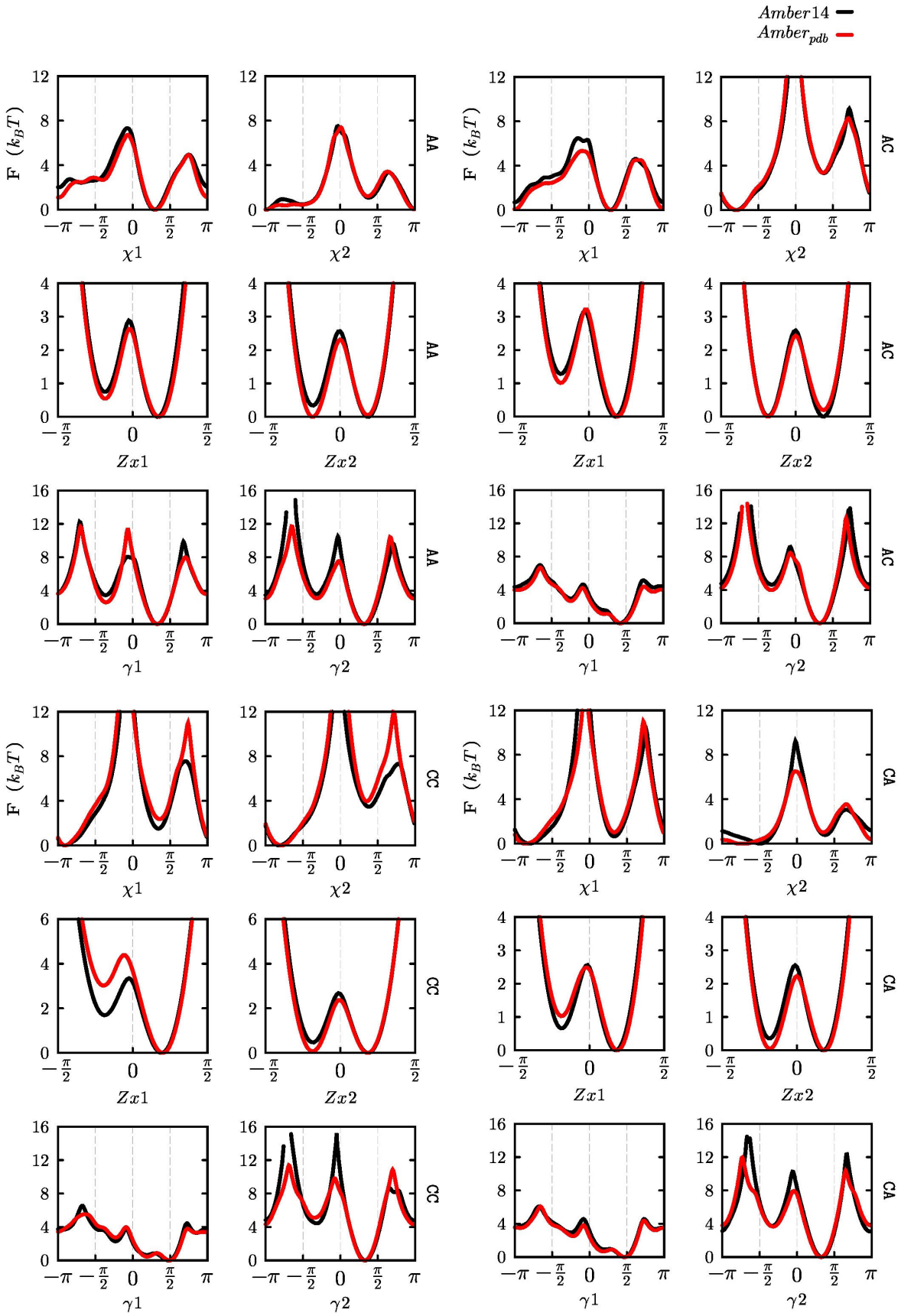


Figure B.8: Free-energy profiles of non-corrected degrees of freedom (χ, γ and puckering Z_x) from the RECT simulations of the standard force-field (Amber14) and the correcting potential (Amber$_{\mathrm{pdb}}$).

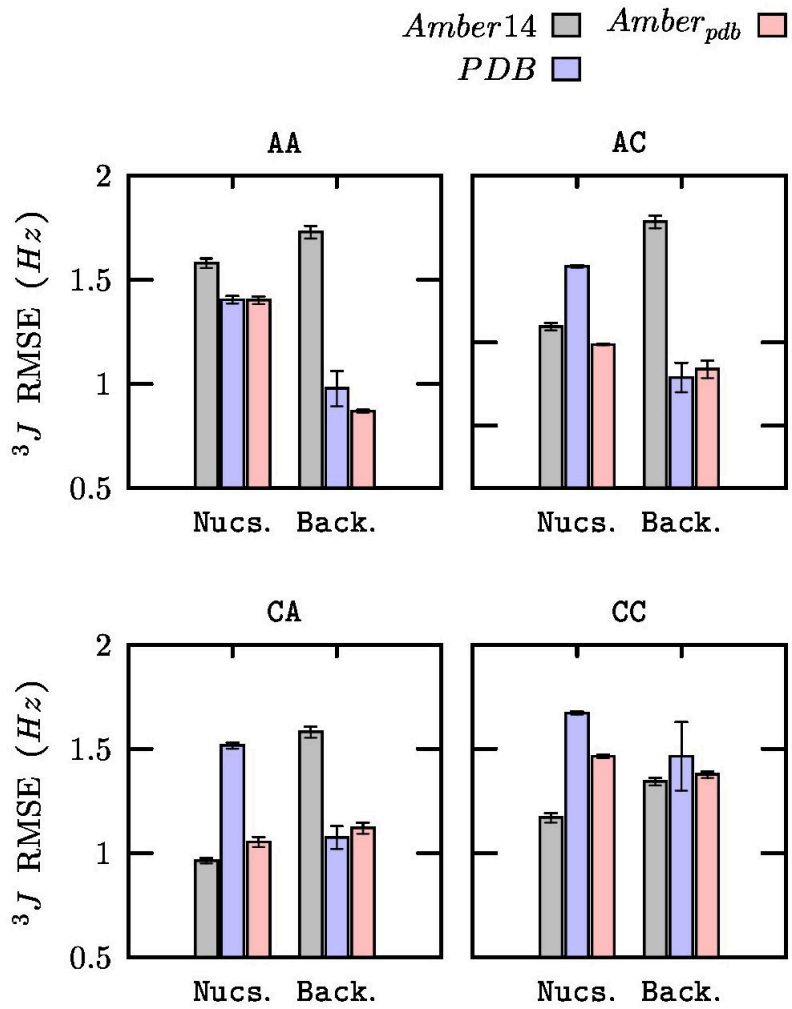


Figure B.9: RMSE between experimental and calculated ^3J scalar couplings for 2 different subset of dihedral angles: Nucs.) containing the nucleoside-unit angles (χ,γ and ν_3) and Back.) including the angles from the monophosphate backbone (ϵ, ζ, α and β).

			J_{HH}			J_{HP}			J_{CP}		J_{HC}	
			$H3'H4'$	$H4'H5'$	$H4'H5''$	$H3'P$	$H5'P$	$H5''P$	$C2'P$	$C4'P$	$H1'C2/4$	$H1'C6/8$
seq.	source	unit	ν_{3-1}/ν_{3-2}	γ_1/γ_2	γ_1/γ_2	ϵ_1	β_2	β_2	ϵ_1	ϵ_1/β_2	χ_1/χ_2	χ_1/χ_2
CpC	NMR*	1	7.3/7.0	2.5/2.5	3.8/3.3	8.9/8.3	-	-	3.1	6.0	1.4	4.5
		2	7.2/6.6	2.4/2.2	2.4/2.6	-	4.3/4.0	3.2/3.6	-	9.5	1.4	4.6
	Amber14	1	9.4	3.8	2.9	6.8	-	-	2.3	7.6	1.7	2.9
		2	7.1	3.3	1.3	-	3.4	1.9	-	10.5	1.8	3.3
	PDB	1	6.8	4.4	2.4	7.6	-	-	1.2	8.3	0.8	1.9
		2	7.1	4.5	2.4	-	3.3	2.6	-	10.3	0.9	1.9
	$\text{Amber}_{\text{pdb}}$	1	10.4	3.8	2.9	7.6	-	-	1.4	8.2	1.3	2.5
		2	6.1	3.4	1.3	-	3.6	2.2	-	10.4	1.7	3.3
ApA	NMR†	1	∼5.0/5.5	2.5/2.5	3.6/3.5	∼9.0/7.6	-	-	3.7	5.3	1.9	4.2
		2	5.5/5.8	-/2.8	∼3.8/3.7	-	-/3.0	∼3.8/3.3	-	9.4	2.5	3.1
	Amber14	1	7.6	3.4	1.5	5.6	-	-	3.1	7.3	4.3	3.5
		2	6.2	3.6	1.6	-	3.1	2.5	-	10.4	1.7	2.7
	PDB	1	6.8	4.8	3.3	8.4	-	-	2.9	6.3	1.2	2.7
		2	7.1	4.5	2.8	-	4.9	3.5	-	9.2	1.3	2.8
	$\text{Amber}_{\text{pdb}}$	1	7.2	3.4	1.7	8.1	-	-	3.1	6.3	3.8	3.1
		2	5.4	3.9	1.9	-	4.2	2.9	-	9.8	1.6	2.8
ApC	NMR†	1	6.1/6.1	2.4/3.3	3.5/2.2	8.7/8.7	-	-	3.3	4.6	2.1	2.8
		2	-/7.1	∼1.7/2.1	∼2.0/2.6	-	∼4.0/3.6	3.4/3.6	-	9.5	1.4	4.5
	Amber14	1	8.4	2.9	1.9	6.1	-	-	2.5	7.7	3.4	2.8
		2	5.9	3.4	1.4	-	3.1	2.4	-	10.5	1.6	3.1
	PDB	1	7.7	4.5	2.7	8.5	-	-	2.4	6.8	1.0	2.4
		2	8.2	4.5	2.6	-	4.5	2.8	-	9.7	1.0	2.1
	$\text{Amber}_{\text{pdb}}$	1	8.0	3.3	2.3	7.9	-	-	2.7	6.7	2.8	2.6
		2	5.6	3.6	1.5	-	3.8	2.4	-	10.2	1.6	3.2
CpA	NMR†	1	6.8/6.8	2.6/3.7	4.0/2.3	8.7/8.4	-	-	3.4	5.4	1.6	4.6
		2	5.5/5.4	2.6/2.0	3.0/3.0	-	4.3/3.8	3.8/3.8	-	9.3	1.8	4.3
	Amber4	1	7.9	3.4	2.5	5.9	-	-	3.4	6.9	2.5	3.6
		2	6.1	3.6	1.7	-	3.2	2.2	-	10.5	2.1	3.5
	PDB	1	8.7	4.6	2.7	7.9	-	-	2.1	7.4	1.0	2.2
		2	8.5	4.5	2.8	-	4.3	3.1	-	9.6	1.1	2.6
	$\text{Amber}_{\text{pdb}}$	1	8.5	3.7	2.9	7.8	-	-	2.4	7.1	2.1	3.2
		2	5.6	3.6	1.6	-	4.1	2.3	-	10.1	1.8	3.3

Table B.1: Scalar couplings for the monophosphate dinucleosides. Experimental values were taken from refs ([185]/[200–202]). *(Experiments temperature: 320/293 K). †(280/293 K).

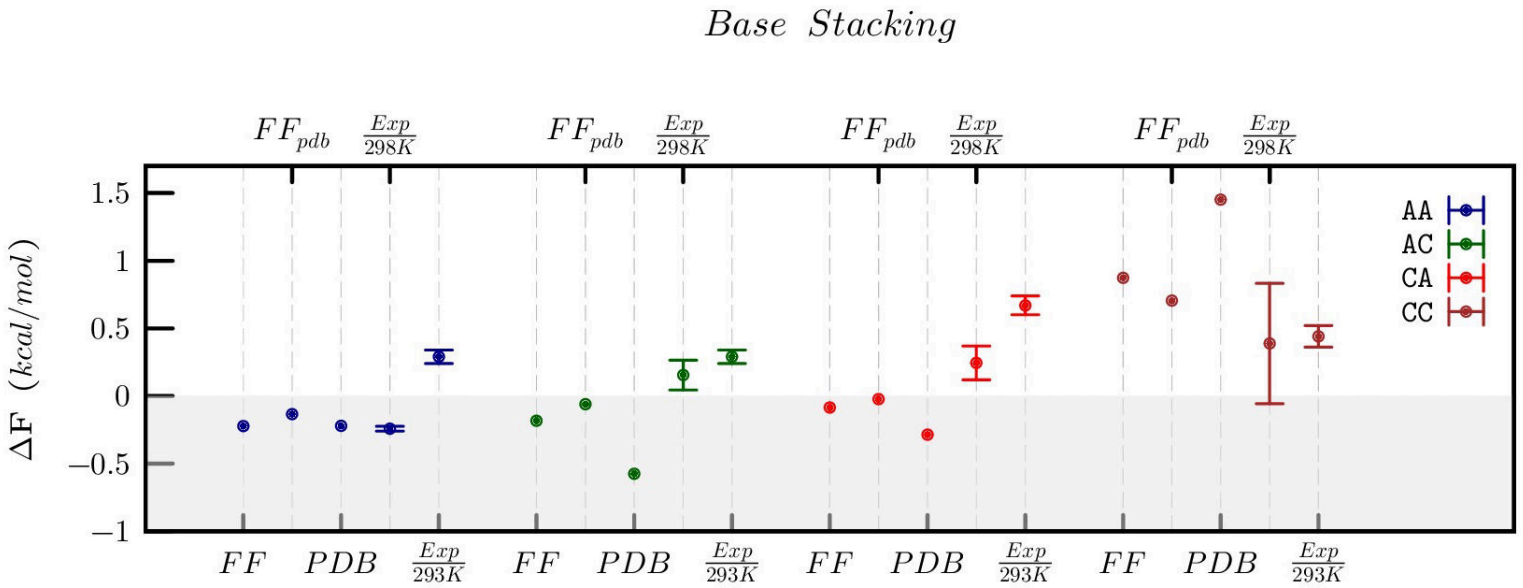


Figure B.10: Free Energy of stacking. Experimental values were taken from refs [201, 202, 222]. The stacking thermodynamics is only slightly affected by the correcting potentials.

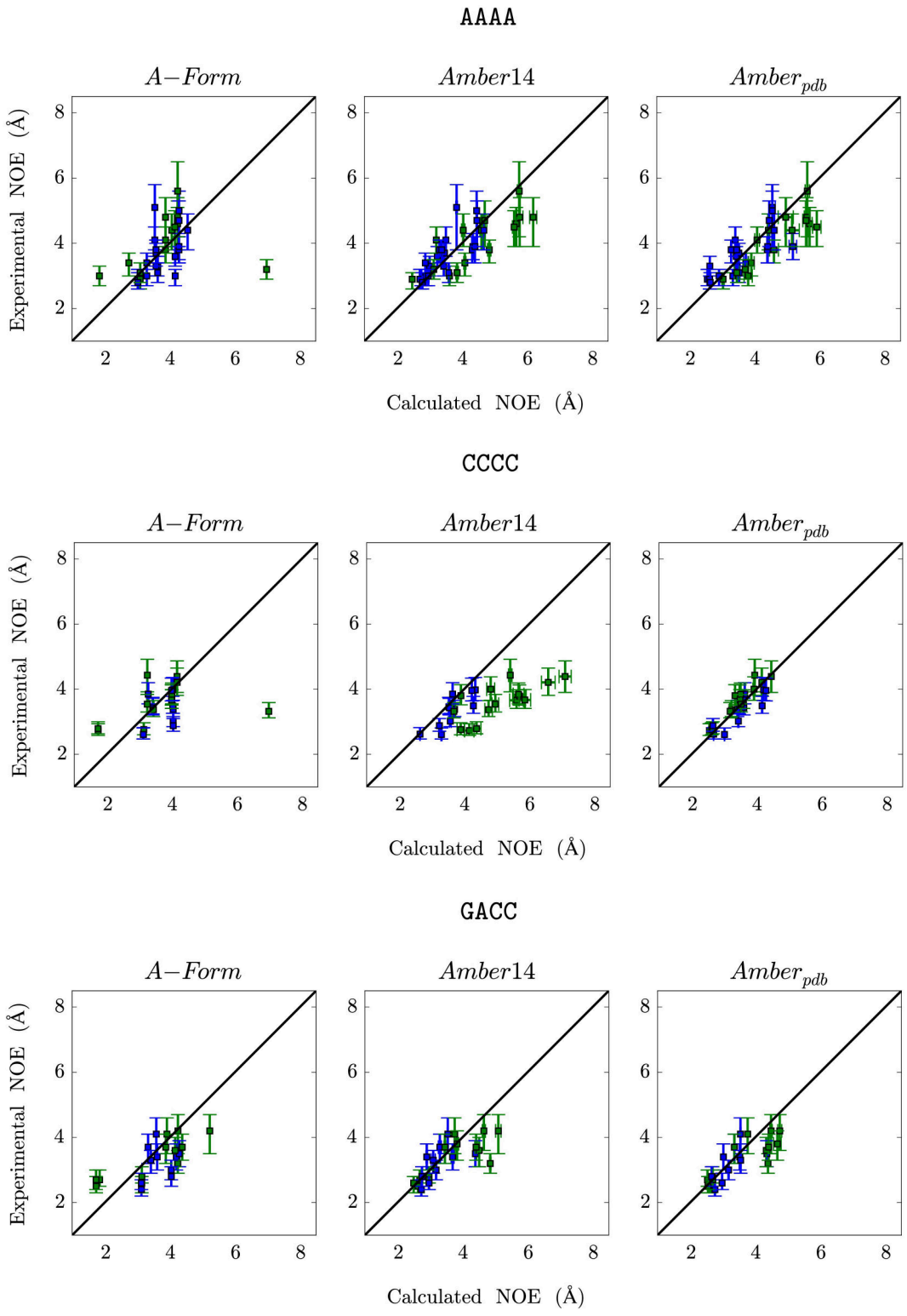


Figure B.11: Predicted versus experimental NOE distance for all ensembles and for all 5 systems are shown below. Bars on the y-axis show experimentally determined minimum and maximum range, while error bars on the predicted values represent statistical errors and were calculated with a blocking procedure. Intra-nucleotide and inter-nucleotide proton-proton distances are shown in blue and green, respectively. Calculations were performed using the software tool baRNAba.[215]

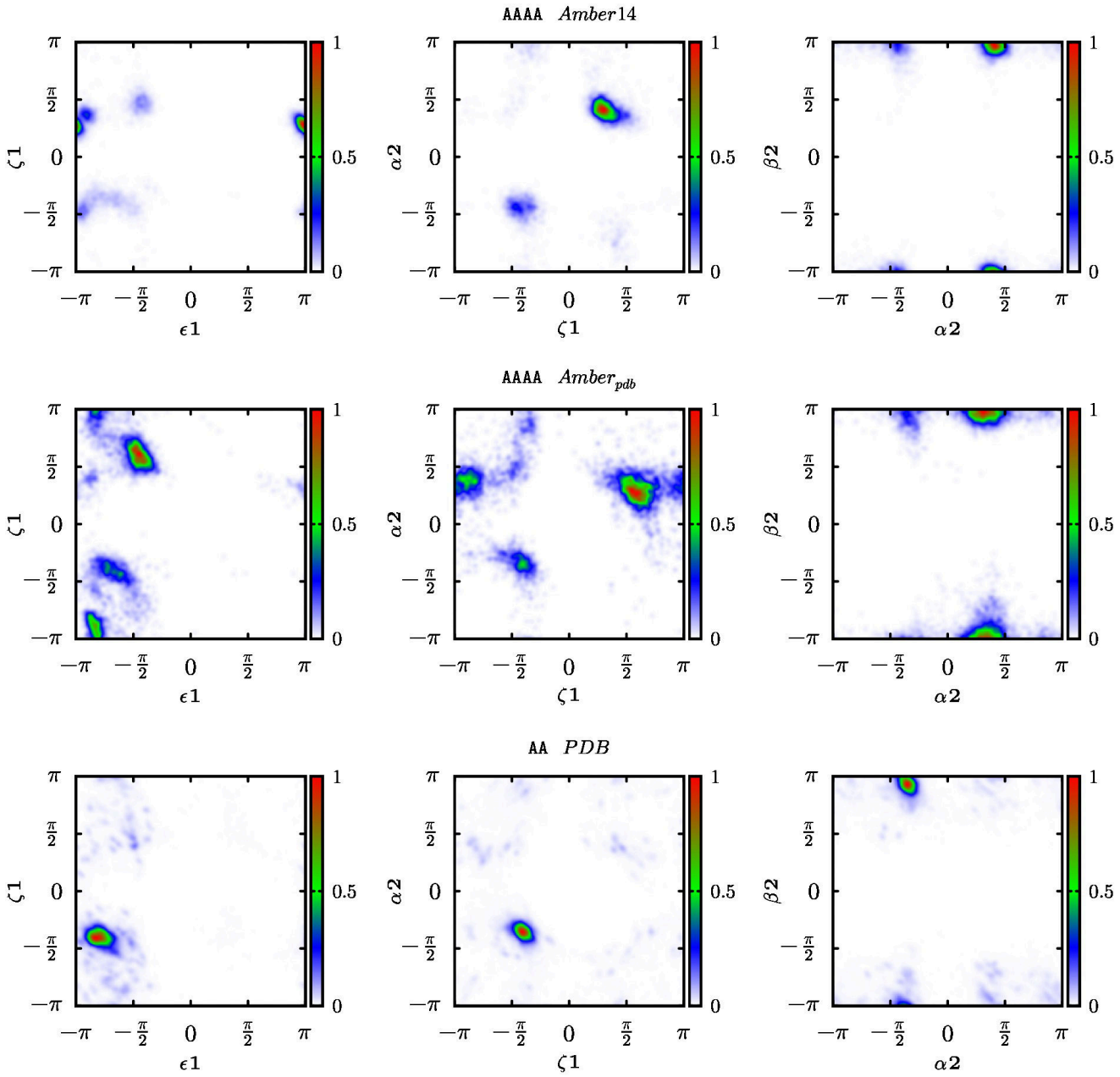


Figure B.12: Probability distributions of the backbone dihedral angles of AAAA tetranucleotide, in the region between residue 1 and 2. First are shown the dihedral distributions from the RECT simulations with the standard force-field (Amber14) and in second the ones performed with the correcting potential ($\text{Amber}_{\text{pdb}}$). The dihedral distributions of AA dinucleoside taken from the PDB are presented in the last row.

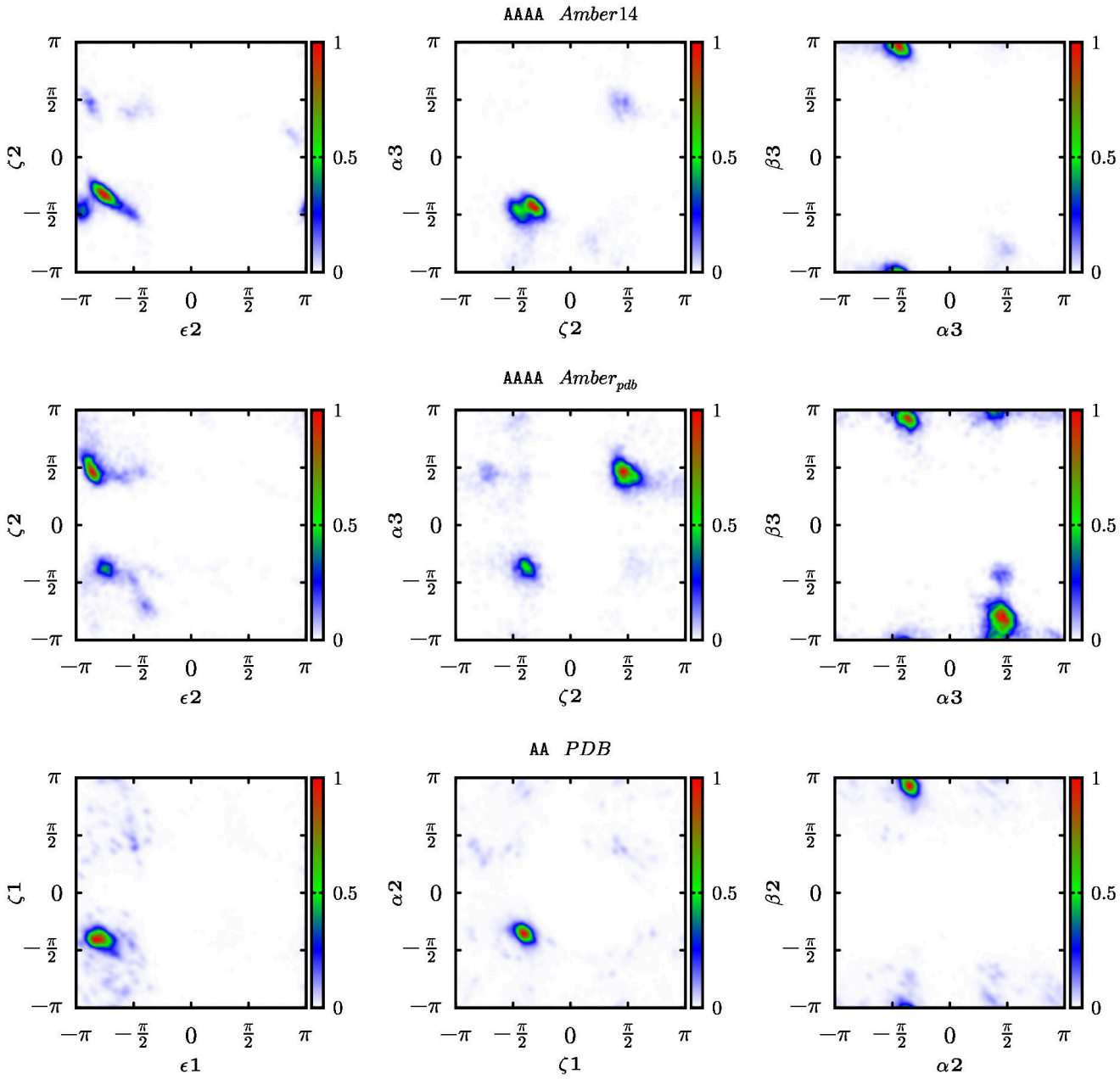


Figure B.13: Probability distributions of the backbone dihedral angles of AAAA tetranucleotide, in the region between residue 2 and 3. First are shown the dihedral distributions from the RECT simulations with the standard force-field (Amber14) and in second the ones performed with the correcting potential (Amber$_{\text{pdb}}$). The dihedral distributions of AA dinucleoside taken from the PDB are presented in the last row.

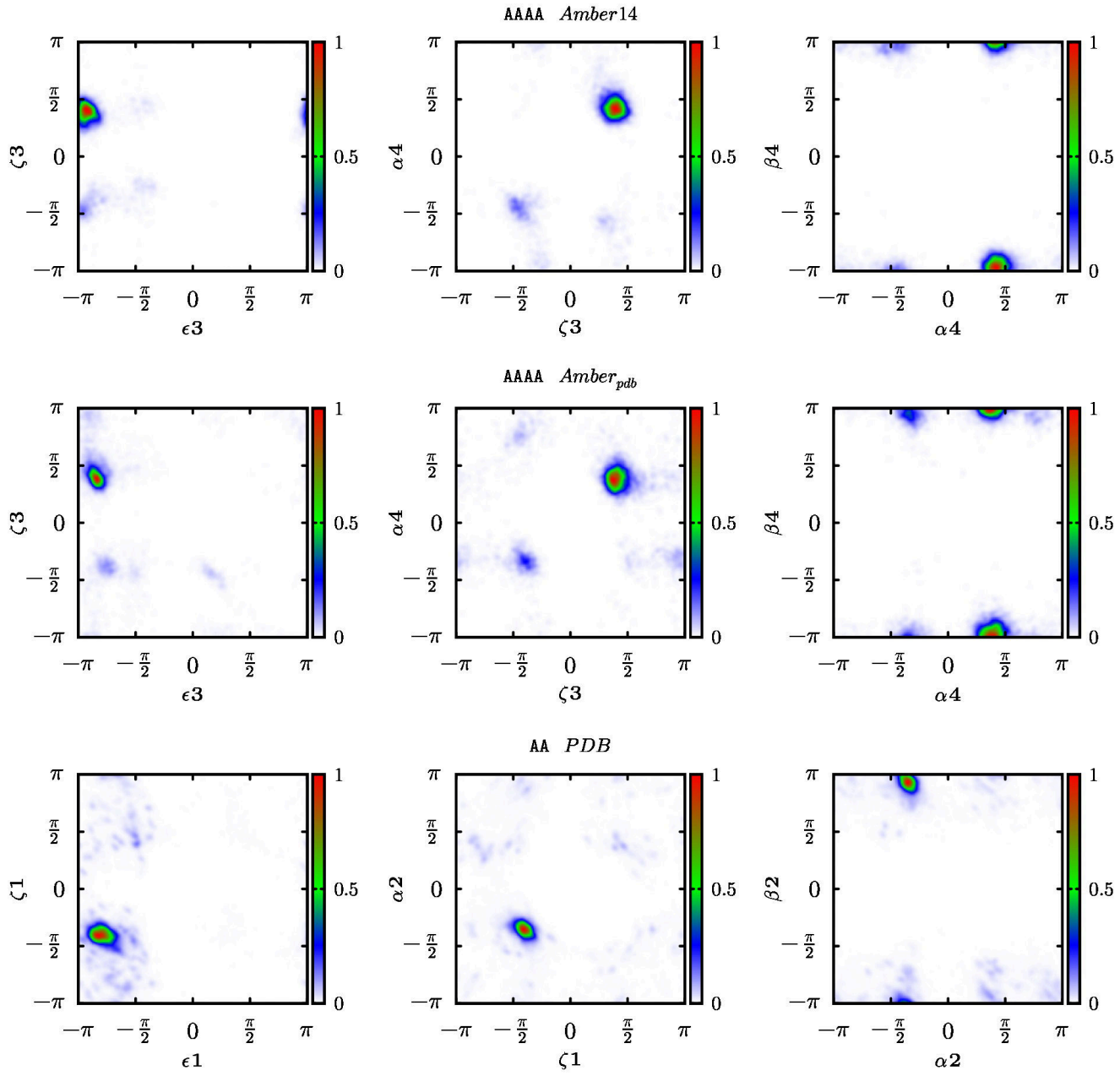


Figure B.14: Probability distributions of the backbone dihedral angles of AAAA tetranucleotide, in the region between residue 3 and 4. First are shown the dihedral distributions from the RECT simulations with the standard force-field (Amber14) and in second the ones performed with the correcting potential ($\text{Amber}_{\text{pdb}}$). The dihedral distributions of AA dinucleoside taken from the PDB are presented in the last row.

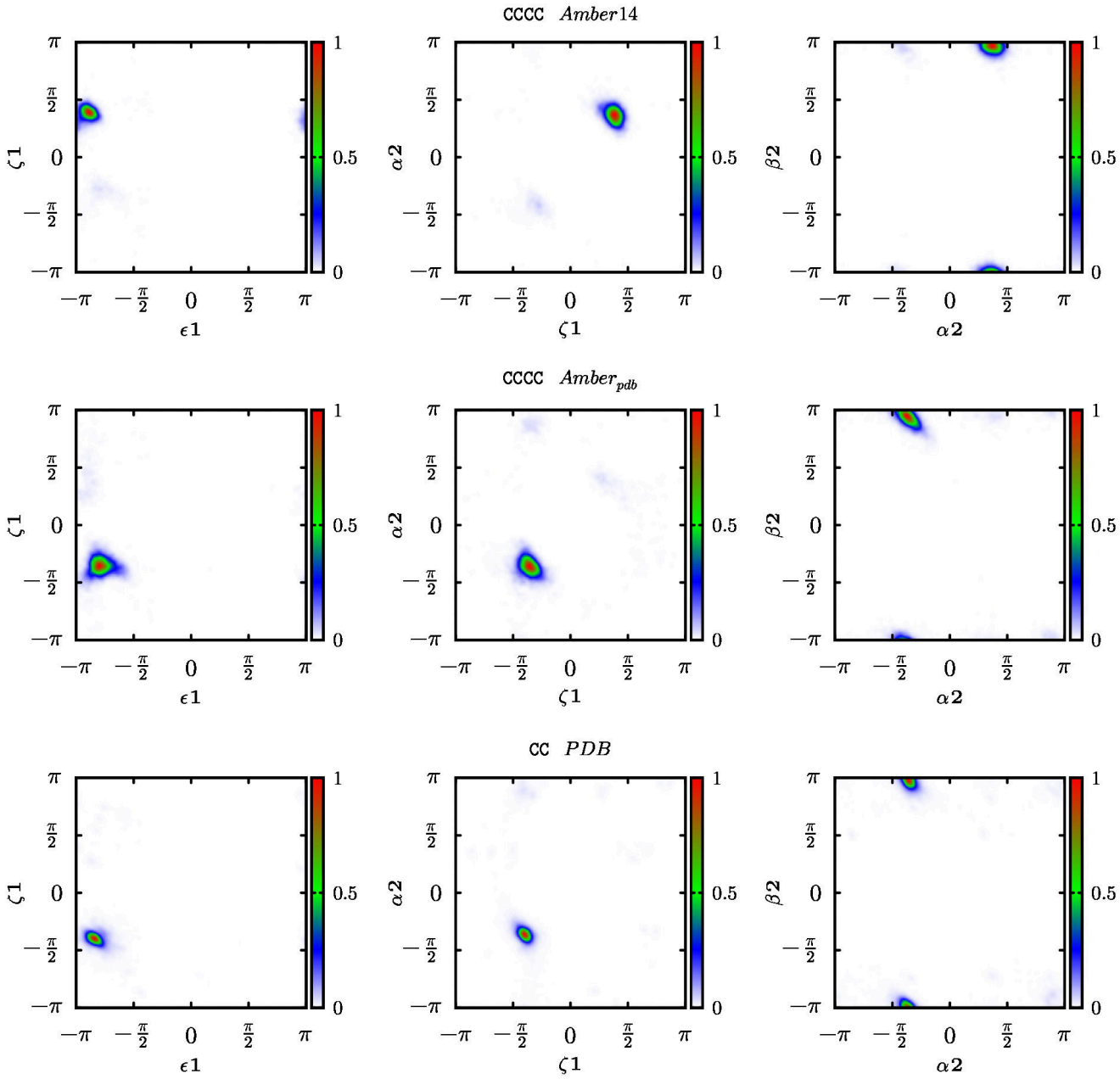


Figure B.15: Probability distributions of the backbone dihedral angles of CCCC tetranucleotide, in the region between residue 1 and 2. First are shown the dihedral distributions from the RECT simulations with the standard force-field (Amber14) and in second the ones performed with the correcting potential (Amber$_{\mathrm{pdb}}$). The dihedral distributions of CC dinucleoside taken from the PDB are presented in the last row.

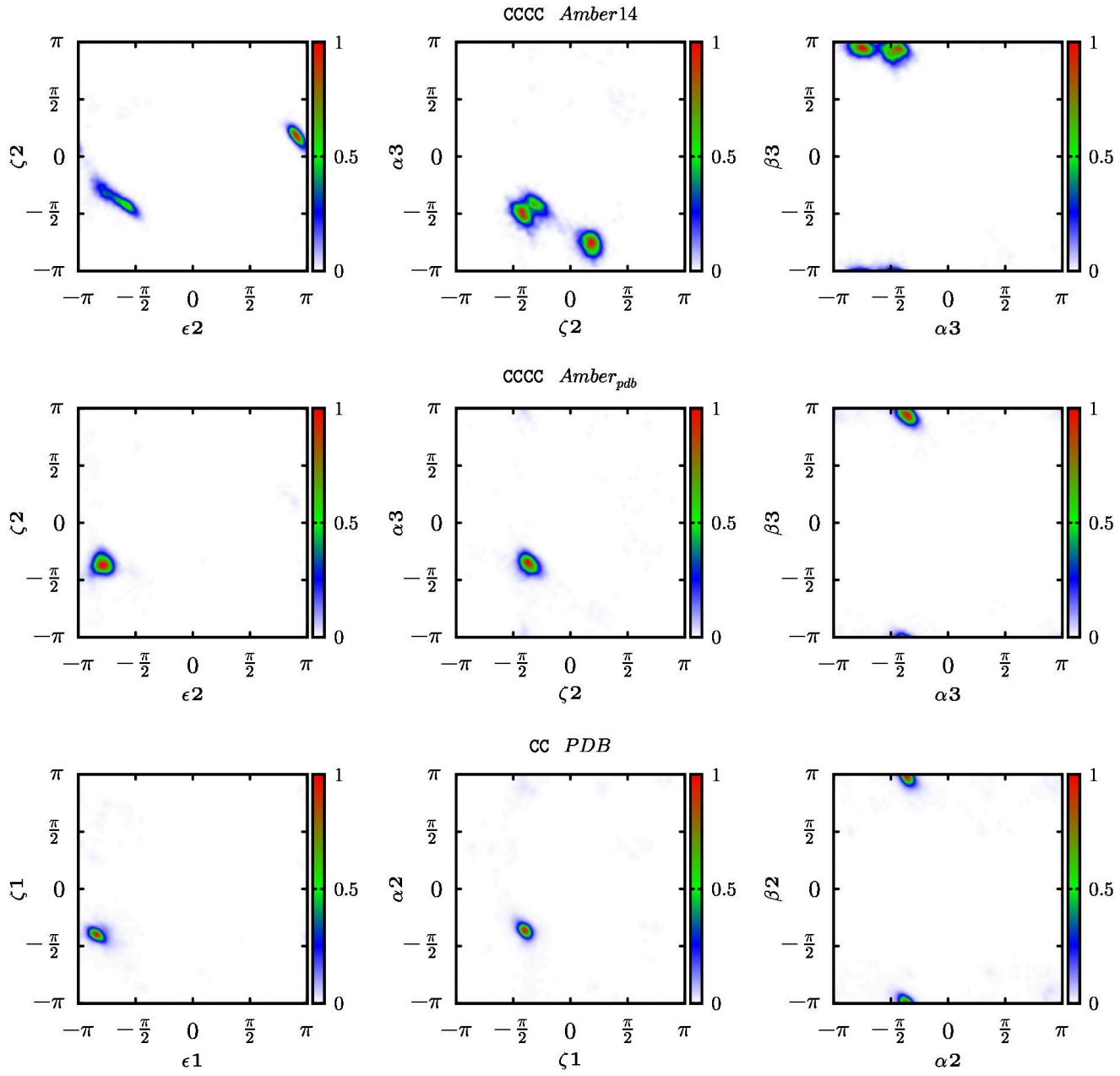


Figure B.16: Probability distributions of the backbone dihedral angles of CCCC tetranucleotide, in the region between residue 2 and 3. First are shown the dihedral distributions from the RECT simulations with the standard force-field (Amber14) and in second the ones performed with the correcting potential ($\text{Amber}_{\text{pdb}}$). The dihedral distributions of CC dinucleoside taken from the PDB are presented in the last row.

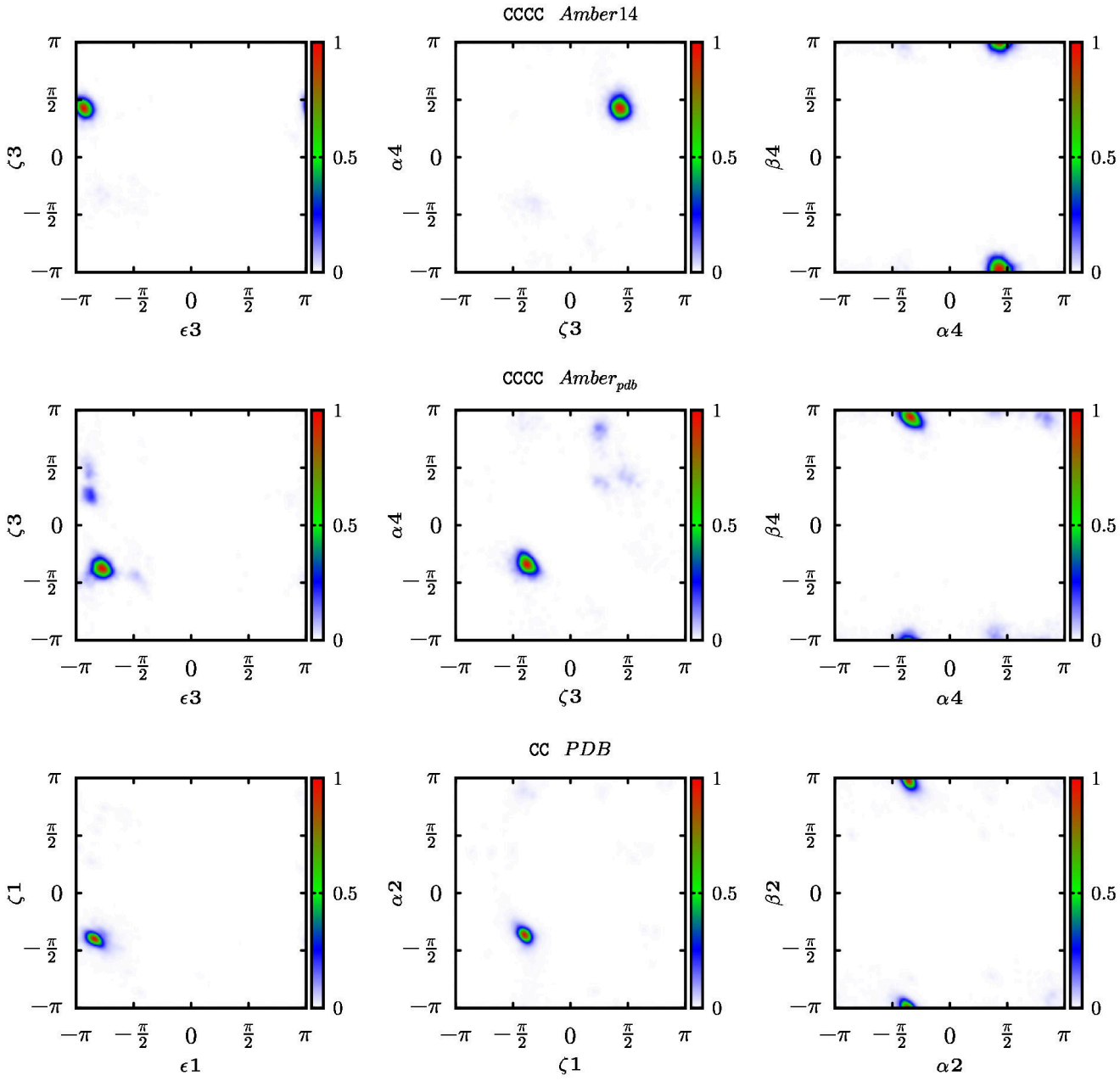


Figure B.17: Probability distributions of the backbone dihedral angles of CCCC tetranucleotide, in the region between residue 3 and 4. First are shown the dihedral distributions from the RECT simulations with the standard force-field (Amber14) and in second the ones performed with the correcting potential (Amber$_{\text{pdb}}$). The dihedral distributions of CC dinucleoside taken from the PDB are presented in the last row.

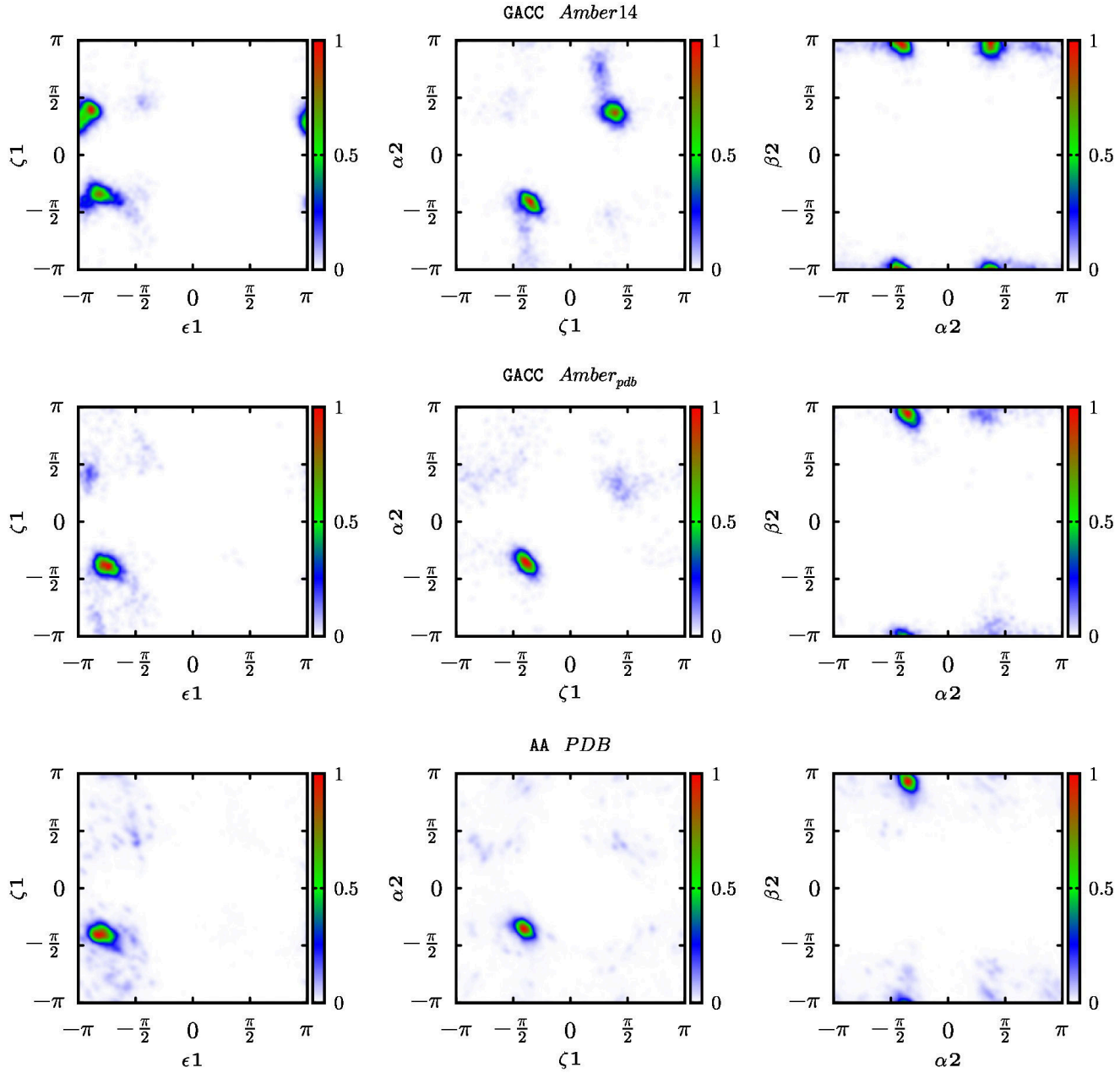


Figure B.18: Probability distributions of the backbone dihedral angles of GACC tetranucleotide, in the region between residue 1 and 2. First are shown the dihedral distributions from the RECT simulations with the standard force-field (Amber14) and in second the ones performed with the correcting potential ($\text{Amber}_{\text{pdb}}$). The dihedral distributions of AA, AC and CC dinucleosides, taken from the PDB, are presented in the last row.

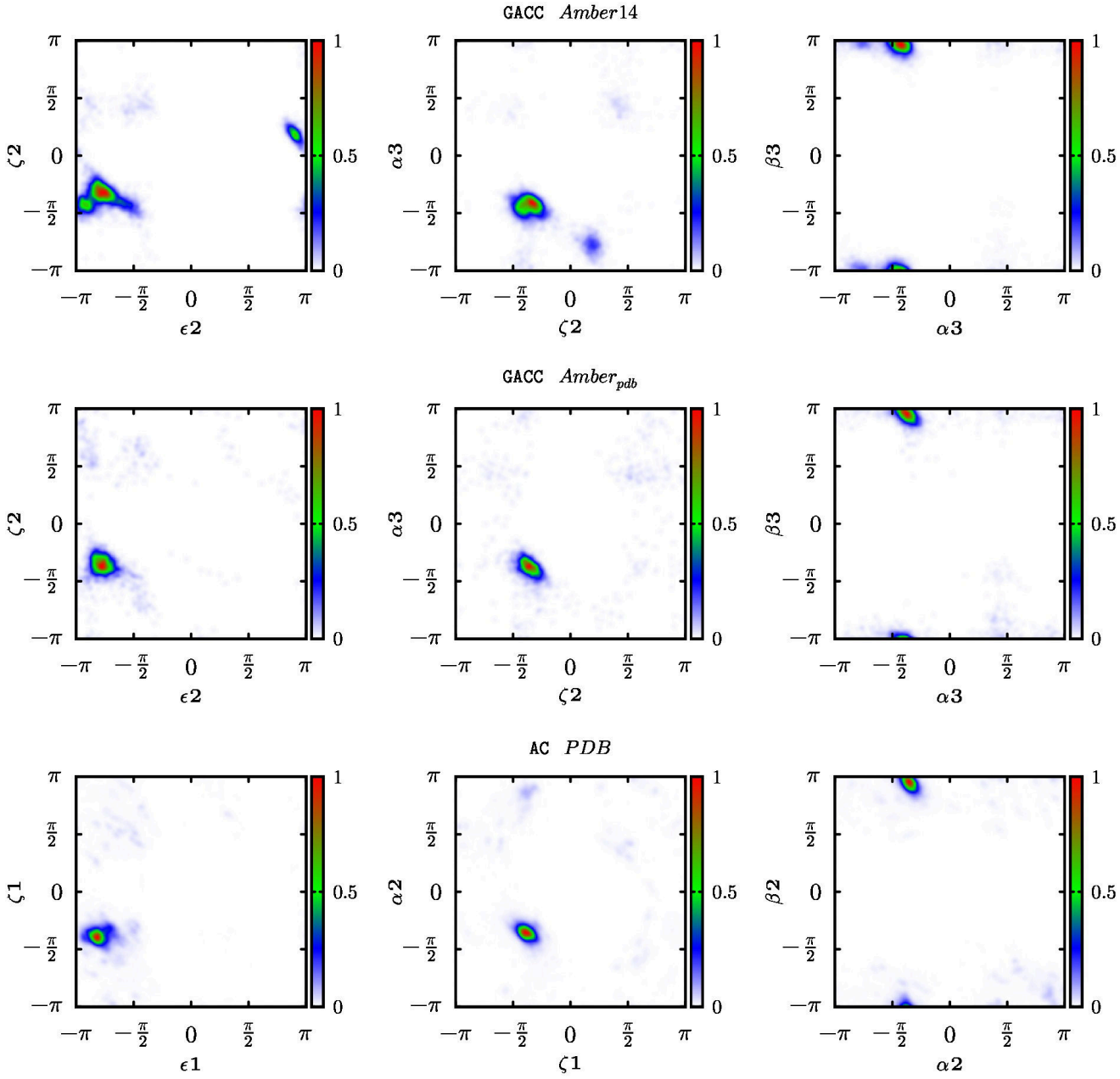


Figure B.19: Probability distributions of the backbone dihedral angles of GACC tetranucleotide, in the region between residue 2 and 3. First are shown the dihedral distributions from the RECT simulations with the standard force-field (Amber14) and in second the ones performed with the correcting potential ($\text{Amber}_{\text{pdb}}$). The dihedral distributions of AA, AC and CC dinucleosides, taken from the PDB, are presented in the last row.

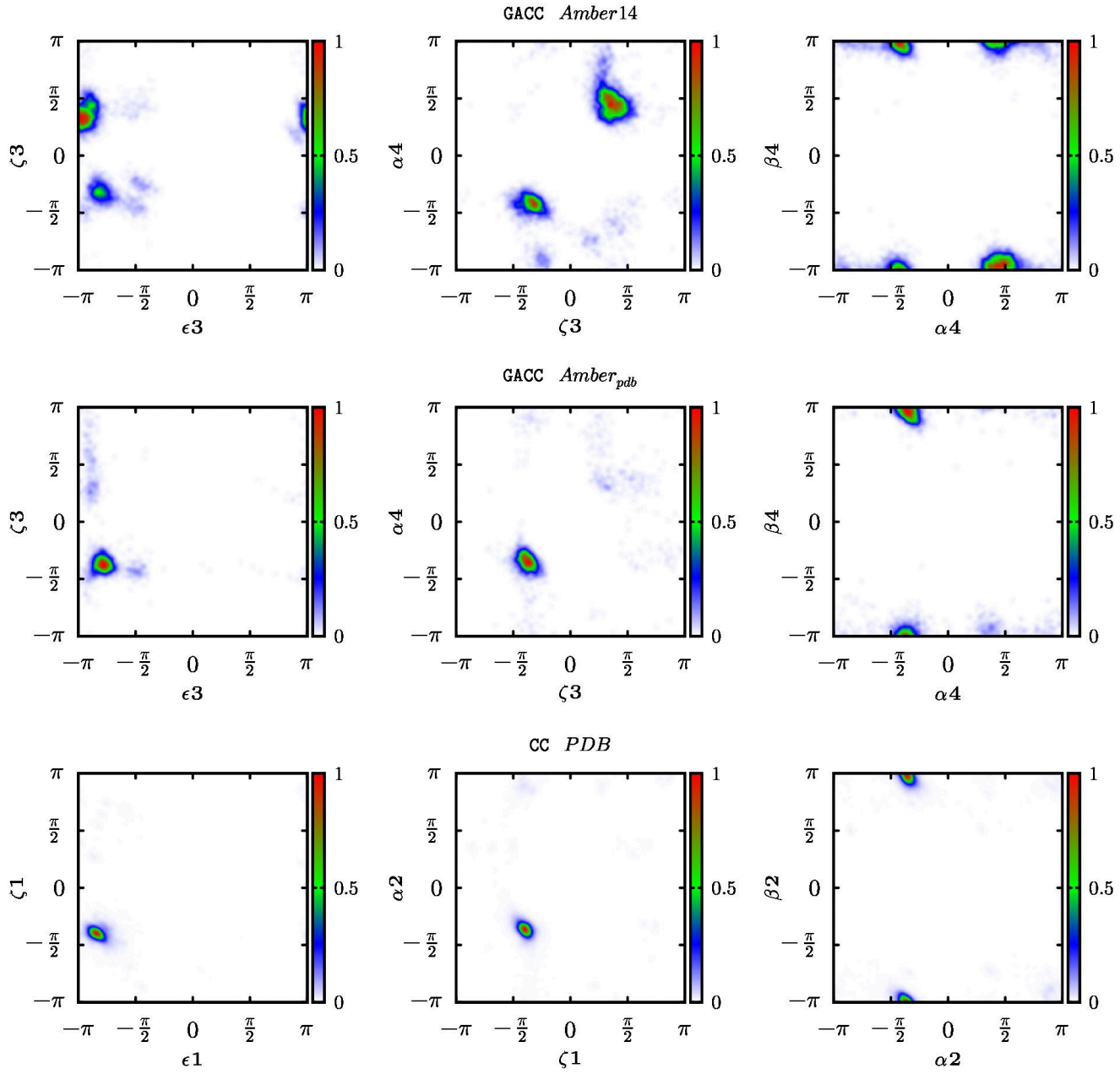


Figure B.20: Probability distributions of the backbone dihedral angles of GACC tetranucleotide, in the region between residue 3 and 4. First are shown the dihedral distributions from the RECT simulations with the standard force-field (Amber14) and in second the ones performed with the correcting potential ($\text{Amber}_{\text{pdb}}$). The dihedral distributions of AA, AC and CC dinucleosides, taken from the PDB, are presented in the last row.

References

[1] Darnell, J. E. *RNA: life's indispensable molecule*; Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY, 2011.

[2] Atkins, J. F.; Gesteland, R. F.; Cech, T. *RNA worlds: from life's origins to diversity in gene regulation*; Cold Spring Harbor Laboratory Press New York;, 2011.

[3] Elliott, D.; Ladomery, M. *Molecular biology of RNA*; Oxford University Press, 2015.

[4] Esteller, M. *Nat. Rev. Genet.* **2011**, *12*, 861–874.

[5] Qureshi, I. A.; Mehler, M. F. *Nat. Rev. Neurosci.* **2012**, *13*, 528–541.

[6] Al-Hashimi, H. M.; Walter, N. G. *Curr. Opin. Struct. Biol.* **2008**, *18*, 321–329.

[7] Dethoff, E. A.; Chugh, J.; Mustoe, A. M.; Al-Hashimi, H. M. *Nature* **2012**, *482*, 322–330.

[8] Ha, T. *Curr. Opin. Struct. Biol.* **2001**, *11*, 287–292.

[9] Zhuang, X. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 399–414.

[10] Zhao, R.; Rueda, D. *Methods* **2009**, *49*, 112–117.

[11] Tinoco, I.; Bustamante, C. *J. Mol. Biol.* **1999**, *293*, 271–281.

[12] Williams, M. C.; Rouzina, I. *Curr. Opin. Struct. Biol.* **2002**, *12*, 330–336.

[13] Verlet, L. *Phys. Rev.* **1967**, *159*, 98.

[14] McCammon, J. A., J Andrew; Gelin, B. R.; Karplus, M. *Nature* **1977**, *267*, 16.

[15] Cheatham, T. E.; Case, D. A. *Biopolymers* **2013**, *99*, 969–977.

[16] Sponer, J.; Banáš, P.; Jurecka, P.; Zgarbova, M.; Kührová, P.; Havrila, M.; Krepl, M.; Stadlbauer, P.; Otyepka, M. *J. Phys. Chem. Lett.* **2014**.

[17] Harvey, S. C.; Prabhakaran, M.; Mao, B.; McCammon, J. A. *Science* **1984**, *223*, 1189–1191.

[18] Prabhakaran, M.; Harvey, S. C.; McCammon, J. A. *Biopolymers* **1985**, *24*, 1189–1204.

[19] Nilsson, L.; Karplus, M. In *Structure and Dynamics of RNA*; Springer, 1986; pp 151–159.

[20] Nilsson, L.; Aahgren-Staalhandske, A.; Sjoegren, A. S.; Hahne, S.; Sjoeberg, B. M. *Biochemistry (Mosc.)* **1990**, *29*, 10317–10322.

[21] Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

[22] Cheatham, T. I.; Miller, J.; Fox, T.; Darden, T.; Kollman, P. *J. Am. Chem. Soc.* **1995**, *117*, 4193–4194.

[23] Weerasinghe, S.; Smith, P. E.; Mohan, V.; Cheng, Y.-K.; Pettitt, B. M. *J. Am. Chem. Soc.* **1995**, *117*, 2147–2158.

[24] Pérez, A.; Luque, F. J.; Orozco, M. *Acc. Chem. Res.* **2012**, *45*, 196–205.

[25] Kührová, P.; Banáš, P.; Best, R. B.; Šponer, J.; Otyepka, M. *J. Chem. Theory Comput.* **2013**, *9*, 2115–2125.

[26] Chen, A. A.; García, A. E. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 16820–16825.

[27] Kührová, P.; Best, R. B.; Bottaro, S.; Bussi, G.; Šponer, J.; Otyepka, M.; Banáš, P. *J. Chem. Theory Comput.* **2016**, *12*, 4534–4548; PMID: 27438572.

[28] Bock, L. V.; Blau, C.; Schröder, G. F.; Davydov, I. I.; Fischer, N.; Stark, H.; Rodnina, M. V.; Vaiana, A. C.; Grubmüller, H. *Nat. Struct. Mol. Biol.* **2013**, *20*, 1390–1396.

[29] Kim, H.; Abeysirigunawarden, S. C.; Chen, K.; Mayerle, M.; Ragunathan, K.; Luthey-Schulten, Z.; Ha, T.; Woodson, S. A. *Nature* **2014**, *506*, 334–338.

[30] Shaw, D. E.; Dror, R. O.; Salmon, J. K.; Grossman, J.; Mackenzie, K. M.; Bank, J. A.; Young, C.; Deneroff, M. M.; Batson, B.; Bowers, K. J.; et al. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*; IEEE; pp 1–11.

[31] Bergonzo, C.; Henriksen, N. M.; Roe, D. R.; Swails, J. M.; Roitberg, A. E.; Cheatham III, T. E. *J. Chem. Theory Comput.* **2013**, *10*, 492–499.

[32] Roe, D. R.; Bergonzo, C.; Cheatham III, T. E. *J. Phys. Chem. B* **2014**, *118*, 3543–3552.

[33] Bergonzo, C.; Henriksen, N. M.; Roe, D. R.; Cheatham, T. E. *RNA* **2015**, *21*, 1578–1590.

[34] Condon, D. E.; Kennedy, S. D.; Mort, B. C.; Kierzek, R.; Yildirim, I.; Turner, D. H. *J. Chem. Theory Comput.* **2015**, *11*, 2729–2742.

[35] White, A.; Dama, J.; Voth, G. A. *J. Chem. Theory Comput.* **2015**, *11*, 2451–2460.

[36] Marinelli, F.; Faraldo-Gómez, J. D. *Biophys. J.* **2015**, *108*, 2779–2782.

[37] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

[38] Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham III, T. E.; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817–3829.

[39] Zgarbová, M.; Otyepka, M.; Šponer, J.; Mládek, A.; Banáš, P.; Cheatham III, T. E.; Jurecka, P. *J. Chem. Theory Comput.* **2011**, *7*, 2886–2902.

[40] Barducci, A.; Bussi, G.; Parrinello, M. *Phys. Rev. Lett.* **2008**, *100*, 020603.

[41] Ts'o, P. O. P.; *Basic Principles in Nucleic Acid Chemistry*; 1974.

[42] Saenger, W. *Principles of nucleic acid structure*; Springer advanced texts in chemistry; Springer-Verlag, 1984.

[43] Neidle, S. *Principles of nucleic acid structure*; Academic Press, 2010.

[44] Haschemeyer, A.; Rich, A. *J. Mol. Biol.* **1967**, *27*, 369–384.

[45] Sasisekharan, V.; Lakshminarayanan, A.; Ramachandran, G. *Conformation of Biopolymers. GN Ramachandran, editor. Academic Press, New York* **1967**, 641–654.

[46] Jordan, F.; Pullman, B. *Theor. Chim. Acta* **1968**, *9*, 242–252.

[47] Tinoco Jr, I.; Davis, R.; Jaskunas, S.; Pullman, B.; *Molecular Associations in Biology*; 1968.

[48] Lakshminarayanan, A.; Sasisekharan, V. *Biochimica et Biophysica Acta (BBA)-Nucleic Acids and Protein Synthesis* **1970**, *204*, 49–59.

[49] Wilson, H.; Rahman, A. *J. Mol. Biol.* **1971**, *56*, 129–142.

[50] Renugopalakrishnan, V.; Lakshminarayanan, A.; Sasisekharan, V. *Biopolymers* **1971**, *10*, 1159–1167.

[51] Rhodes, L. M.; Schimmel, P. R. *Biochemistry (Mosc.)* **1971**, *10*, 4426–4433.

[52] Nishikawa, S.; Kuramoto, N.; Huang, H.; Jordan, F. *J. Phys. Chem. B* **1999**, *103*, 3754–3757.

[53] Nishikawa, S.; Huang, H.; Jordan, F. *J. Phys. Chem. B* **2000**, *104*, 1391–1394.

[54] Yildirim, I.; Stern, H. A.; Kennedy, S. D.; Tubbs, J. D.; Turner, D. H. *J. Chem. Theory Comput.* **2010**, *6*, 1520–1531.

[55] Sokoloski, J. E.; Godfrey, S. A.; Dombrowski, S. E.; Bevilacqua, P. C. *RNA* **2011**, *17*, 1775–1787.

[56] Kilpatrick, J. E.; Pitzer, K. S.; Spitzer, R. *J. Am. Chem. Soc.* **1947**, *69*, 2483–2488.

[57] Altona, C.; Geise, H. t.; Romers, C. *Tetrahedron* **1968**, *24*, 13–32.

[58] Altona, C. t.; Sundaralingam, M. *J. Am. Chem. Soc.* **1972**, *94*, 8205–8212.

[59] Cremer, D. t.; Pople, J. A. *J. Am. Chem. Soc.* **1975**, *97*, 1354–1358.

[60] Rao, S.; Westhof, E. t.; Sundaralingam, M. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* **1981**, *37*, 421–425.

[61] Sato, T. *Nucleic Acids Res.* **1983**, *11*, 4933–4938.

[62] Hill, A. D.; Reilly, P. J. *J. Chem. Inf. Model.* **2007**, *47*, 1031–1035.

[63] Huang, M.; Giese, T. J.; Lee, T.-S.; York, D. M. *J. Chem. Theory Comput.* **2014**, *10*, 1538–1545.

[64] Altona, C.; Sundaralingam, M. *J. Am. Chem. Soc.* **1973**, *95*, 2333–2344.

[65] Olson, W. K.; Sussman, J. L. *J. Am. Chem. Soc.* **1982**, *104*, 270–278.

[66] Westhof, E.; Sundaralingam, M. *J. Am. Chem. Soc.* **1983**, *105*, 970–976.

[67] van Wijk, J.; Huckriede, B. D.; Ippel, J. H.; Altona, C. *Methods Enzymol.* **1992**, *211*, 286–306.

[68] Tonelli, M.; James, T. L. *Biochemistry (Mosc.)* **1998**, *37*, 11478–11487.

[69] Duchardt, E.; Nilsson, L.; Schleucher, J. *Nucleic Acids Res.* **2008**, *36*, 4211–4219.

[70] Thibaudeau, C.; Plavec, J.; Chattopadhyaya, J. *J. Org. Chem.* **1996**, *61*, 266–286.

[71] Foloppe, N.; Nilsson, L.; MacKerell, A. D. *Biopolymers* **2001**, *61*, 61–76.

[72] Rao, S. T.; Sundaralingam, M. *J. Am. Chem. Soc.* **1970**, *92*, 4963–4970.

[73] Suck, D.; Saenger, W. *J. Am. Chem. Soc.* **1972**, *94*, 6520–6526.

[74] Acharya, P.; Ph.D. thesis; Uppsala University; 2003.

[75] Thibaudeau, C.; Acharya, P.; Chattopadhyaya, J. *Stereoelectronic effects in nucleosides and nucleotides and their structural implications*, 2nd ed.; Uppsala University Press: Uppsala, Sweden, 2005.

[76] De Leeuw, H. P. M.; Haasnoot, C. A. G.; Altona, C. *Isr. J. Chem.* **1980**, *20*, 108–126.

[77] Sundaralingam, M. *Conformation of biological molecules and polymers* **1973**, *5*, 417–456.

[78] Wilson, H. In *Conformation of biological molecules and polymers: proceedings of an international symposium held in Jerusalem, 3-9 April 1972*; Academic Press; p 261.

[79] Saenger, W. *Angewandte Chemie International Edition in English* **1973**, *12*, 591–601.

[80] Sundaralingam, M. *Ann. N.Y. Acad. Sci.* **1975**, *255*, 3–42.

[81] Kölkenbeck, K.; Zundel, G. *Eur. Biophys. J.* **1975**, *1*, 203–219.

[82] Jack, A.; Ladner, J.; Klug, A. *J. Mol. Biol.* **1976**, *108*, 619–649.

[83] Young, P. R.; Kallenbach, N. R. *J. Mol. Biol.* **1978**, *126*, 467–479.

[84] Bolton, P.; Kearns, D. *Biochimica et Biophysica Acta (BBA)-Nucleic Acids and Protein Synthesis* **1978**, *517*, 329–337.

[85] Bolton, P. H.; Kearns, D. R. *J. Am. Chem. Soc.* **1979**, *101*, 479–484.

[86] Plavec, J.; Thibaudeau, C.; Chattopadhyaya, J. *J. Am. Chem. Soc.* **1994**, *116*, 6558–6560.

[87] Sundaralingam, M. *Biopolymers* **1969**, *7*, 821–860.

[88] Olson, W. K. In *Topics in nucleic acid structure*; Springer, 1982; pp 1–79.

[89] Davies, D. B. *Prog. Nucl. Magn. Reson. Spectrosc.* **1978**, *12*, 135–225.

[90] Danyluk, S. S. In *Nucleoside Analogues*; Springer, 1979; pp 15–34.

[91] Dhingra, M.; Sarma, R. In *Stereodynamics of Molecular Systems*; Pergamon Press New York, NY, 1979; pp 3–38.

[92] Emerson, J.; Sundaralingam, M. *Acta Crystallographica Section B: Structural Crystallography and Crystal Chemistry* **1980**, *36*, 537–543.

[93] Lakshminarayanan, A.; Sasisekharan, V. *Biopolymers* **1969**, *8*, 475–488.

[94] Saran, A.; Govil, G. *J. Theor. Biol.* **1971**, *33*, 407–418.

[95] Pullman, B.; Perahia, D.; Saran, A. *Biochimica et Biophysica Acta (BBA)-Nucleic Acids and Protein Synthesis* **1972**, *269*, 1–14.

[96] Broyde, S. B.; Wartell, R. M.; Stellman, S. D.; Hingerty, B.; Langridge, R. *Biopolymers* **1975**, *14*, 1597–1613.

[97] Thornton, J. M.; Bayley, P. M. *Biochem. J.* **1975**, *149*, 585–596.

[98] Pullman, B.; Saran, A. *Prog. Nucleic Acid Res. Mol. Biol.* **1976**, *18*, 215–325.

[99] Murthy, V. L.; Srinivasan, R.; Draper, D. E.; Rose, G. D. *J. Mol. Biol.* **1999**, *291*, 313–327.

[100] Ts'o, P. O.; Kondo, N. S.; Schweizer, M. P. *Biochemistry (Mosc.)* **1969**, *8*, 997–1029.

[101] Varani, G. *Annu. Rev. Biophys. Biomol. Struct.* **1995**, *24*, 379–404.

[102] Murray, L. J.; Arendall, W. B.; Richardson, D. C.; Richardson, J. S. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13904–13909.

[103] MacKerell Jr, A. D.; Bashford, D.; Bellott, M.; Dunbrack Jr, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

[104] Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

[105] Schuler, L. D.; Daura, X.; Van Gunsteren, W. F. *J. Comput. Chem.* **2001**, *22*, 1205–1218.

[106] Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341–346.

[107] Ivani, I.; Dans, P. D.; Noy, A.; Pérez, A.; Faustino, I.; Hospital, A.; Walther, J.; Andrio, P.; Goñi, R.; Balaceanu, A.; et al. *Nat. Methods* **2016**, *13*, 55–58.

[108] Galindo-Murillo, R.; Robertson, J. C.; Zgarbová, M.; Šponer, J.; Otyepka, M.; Jurečka, P.; Cheatham III, T. E. *J. Chem. Theory Comput.* **2016**.

[109] Zhao, B.; Zhang, Q. *Curr. Opin. Struct. Biol.* **2015**, *30*, 134–146.

[110] Steinbrecher, T.; Latzer, J.; Case, D. *J. Chem. Theory Comput.* **2012**, *8*, 4405–4412.

[111] Foloppe, N.; MacKerell Jr, A. D. *J. Comput. Chem.* **2000**, *21*, 86–104.

[112] Mackerell, A. D.; Banavali, N. K. *J. Comput. Chem.* **2000**, *21*, 105–120.

[113] MacKerell, A. D.; Banavali, N.; Foloppe, N. *Biopolymers* **2000**, *56*, 257–265.

[114] Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; MacKerell Jr, A. D. *J. Chem. Theory Comput.* **2012**, *8*, 3257–3273.

[115] Beššeová, I.; Banáš, P.; Kührová, P.; Košinová, P.; Otyepka, M.; Šponer, J. *J. Phys. Chem. B* **2012**, *116*, 9899.

[116] Banáš, P.; Sklenovskỳ, P.; Wedekind, J. E.; Šponer, J.; Otyepka, M. *J. Phys. Chem. B* **2012**, *116*, 12721.

[117] Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

[118] Banáš, P.; Hollas, D.; Zgarbová, M.; Jurečka, P.; Orozco, M.; Cheatham III, T. E.; Šponer, J.; Otyepka, M. *J. Chem. Theory Comput.* **2010**, *6*, 3836–3849.

[119] Mlynsky, V.; Banas, P.; Hollas, D.; Reblova, K.; Walter, N.; Sponer, J.; Otyepka, M.; et al. *J. Phys. Chem. B* **2010**, *114*, 6642–6652.

[120] Henriksen, N. M.; Roe, D. R.; Cheatham III, T. E. *J. Phys. Chem. B* **2013**, *117*, 4014–4027.

[121] Kührová, P.; Otyepka, M.; Šponer, J.; Banáš, P. *J. Chem. Theory Comput.* **2014**, *10*, 401.

[122] Yildirim, I.; Kennedy, S. D.; Stern, H. A.; Hart, J. M.; Kierzek, R.; Turner, D. H. *J. Chem. Theory Comput.* **2011**, *8*, 172–181.

[123] Šponer, J.; Mládek, A.; Šponer, J. E.; Svozil, D.; Zgarbová, M.; Banáš, P.; Jurečka, P.; Otyepka, M. *Phys. Chem. Chem. Phys.* **2012**, *14*, 15257–15277.

[124] Banáš, P.; Mládek, A.; Otyepka, M.; Zgarbová, M.; Jurečka, P.; Svozil, D.; Lankaš, F.; Šponer, J. *J. Chem. Theory Comput.* **2012**, *8*, 2448–2460.

[125] Schrodt, M. V.; Andrews, C. T.; Elcock, A. H. *J. Chem. Theory Comput.* **2015**, *11*, 5906–5917.

[126] Havrila, M.; Zgarbová, M.; Jurečka, P.; Banáš, P.; Krepl, M.; Otyepka, M.; Sponer, J. *J. Phys. Chem. B* **2015**, *119*, 15176–15190.

[127] Beššeová, I.; Otyepka, M.; Réblová, K.; Šponer, J. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10701–10711.

[128] Bergonzo, C.; III, T. E. C. *J. Chem. Theory Comput.* **2015**, *11*, 3969–3972.

[129] Dror, R. O.; Dirks, R. M.; Grossman, J.; Xu, H.; Shaw, D. E. *Annu. Rev. Biophys.* **2012**, *41*, 429–452.

[130] Abrams, C.; Bussi, G. *Entropy* **2014**, *16*, 163–199.

[131] Kirkpatrick, S.; Gelatt Jr., C. D.; Vecchi, M. P. *Science* **1983**, *220*, 671–680.

[132] Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187–199.

[133] Arrhenius, S. *Über die Dissociationswärme und den Einfluss der Temperatur auf den Dissociationsgrad der Elektrolyte*; Wilhelm Engelmann, 1889.

[134] Arrhenius, S. *Zeitschrift für physikalische Chemie* **1889**, *4*, 226–248.

[135] Swendsen, R. H.; Wang, J.-S. *Phys. Rev. Lett.* **1986**, *57*, 2607–2609.

[136] Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

[137] Marinari, E.; Parisi, G. *Europhys. Lett.* **1992**, *19*, 451.

[138] Sabri Dashti, D.; Roitberg, A. E. *J. Chem. Theory Comput.* **2013**, *9*, 4692–4699.

[139] Kofke, D. A. *J. Chem. Phys.* **2002**, *117*, 6911–6914.

[140] Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604–1608.

[141] Prakash, M. K.; Barducci, A.; Parrinello, M. *J. Chem. Theory Comput.* **2011**, *7*, 2025–2027.

[142] Zheng, W.; Andrec, M.; Gallicchio, E.; Levy, R. M. *PNAS* **2007**, *104*, 15340–15345.

[143] Nymeyer, H. *J. Chem. Theory Comput.* **2008**, *4*, 626–636.

[144] Denschlag, R.; Lingenheil, M.; Tavan, P. *Chem. Phys. Lett.* **2008**, *458*, 244–248.

[145] Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 13749–13754.

[146] Wang, L.; Friesner, R. A.; Berne, B. *J. Phys. Chem. B* **2011**, *115*, 9431–9438.

[147] Laghaei, R.; Mousseau, N.; Wei, G. *J. Phys. Chem. B* **2010**, *114*, 7071–7077.

[148] Huber, T.; Torda, A. E.; van Gunsteren, W. F. *J. Comput. Aided Mol. Des.* **1994**, *8*, 695–708.

[149] Grubmüller, H. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **1995**, *52*, 2893.

[150] Darve, E.; Pohorille, A. *J. Chem. Phys.* **2001**, *115*, 9169–9183.

[151] Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562–12566.

[152] Laio, A.; Gervasio, F. L. *Rep. Prog. Phys.* **2008**, *71*, 126601.

[153] Bussi, G.; Branduardi, D. *Rev. Comput. Chem.* **2015**, *28*, 1–49.

[154] Branduardi, D.; Bussi, G.; Parrinello, M. *J. Chem. Theory Comput.* **2012**, *8*, 2247–2254.

[155] Dama, J. F.; Parrinello, M.; Voth, G. A. *Phys. Rev. Lett.* **2014**, *112*, 240602.

[156] Rosso, L.; Tuckerman, M. E. *Mol. Simul.* **2002**, *28*, 91–112.

[157] VandeVondele, J.; Rothlisberger, U. *J. Phys. Chem. B* **2002**, *106*, 203–208.

[158] Maragliano, L.; Vanden-Eijnden, E. *Chem. Phys. Lett.* **2006**, *426*, 168–175.

[159] Bussi, G.; Gervasio, F. L.; Laio, A.; Parrinello, M. *J. Am. Chem. Soc.* **2006**, *128*, 13435–13441.

[160] Piana, S.; Laio, A. *J. Phys. Chem. B* **2007**, *111*, 4553–4559.

[161] Bonomi, M.; Parrinello, M. *Phys. Rev. Lett.* **2010**, *104*, 190601.

[162] Deighan, M.; Bonomi, M.; Pfaendtner, J. *J. Chem. Theory Comput.* **2012**, *8*, 2189–2192.

[163] Curuksu, J.; Zacharias, M. *J. Chem. Phys.* **2009**, *130*, 104110.

[164] Ostermeir, K.; Zacharias, M. *J. Comput. Chem.* **2014**, *35*, 150–158.

[165] Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919–11929.

[166] Yildirim, I.; Stern, H. A.; Tubbs, J. D.; Kennedy, S. D.; Turner, D. H. *J. Phys. Chem. B* **2011**, *115*, 9261–9270.

[167] Chipot, C.; Lelièvre, T. *SIAM J. Appl. Math.* **2011**, *71*, 1673–1695.

[168] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712–725.

[169] Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J.; Dror, R.; Shaw, D. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950–1958; cited By (since 1996)291.

[170] Jorgensen, W. L. *J. Am. Chem. Soc.* **1981**, *103*, 335–340.

[171] Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

[172] Hess, B. *J. Chem. Theory Comput.* **2008**, *4*, 116–122.

[173] Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101

[174] Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

[175] Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. *Comput. Phys. Commun.* **2014**, *185*, 604–613.

[176] Bussi, G. *Mol. Phys.* **2013**, *112*, 379–384.

[177] Li, D.-W.; Brüschweiler, R. *Phys. Rev. Lett.* **2009**, *102*, 118108.

[178] Apostolakis, J.; Ferrara, P.; Caflisch, A. *J. Chem. Phys.* **1999**, *110*, 2099–2108.

[179] Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. *J. Chem. Theory Comput.* **2007**, *3*, 26–41.

[180] Valsson, O.; Parrinello, M. *Phys. Rev. Lett.* **2014**, *113*, 090601.

[181] Maragliano, L.; Vanden-Eijnden, E. *J. Chem. Phys.* **2008**, *128*, 184110.

[182] Affentranger, R.; Tavernelli, I.; Di Iorio, E. E. *J. Chem. Theory Comput.* **2006**, *2*, 217–228.

[183] Voter, A. F. *Phys. Rev. Lett.* **1997**, *78*, 3908.

[184] Camilloni, C.; Provasi, D.; Tiana, G.; Broglia, R. A. *Proteins: Struct., Funct., Bioinf.* **2008**, *71*, 1647–1654.

[185] Vokáčová, Z.; Budesinsky, M.; Rosenberg, I.; Schneider, B.; Šponer, J.; Sychrovský, V. *J. Phys. Chem. B* **2009**, *113*, 1182–1191.

[186] Nganou, C.; Kennedy, S. D.; McCamant, D. W. *J. Phys. Chem. B* **2016**, *120*, 1250–1258.

[187] Butterfoss, G. L.; Hermans, J. *Protein Sci.* **2003**, *12*, 2719–2731.

[188] MacKerell, A. D.; Feig, M.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1400–1415.

[189] Morozov, A. V.; Kortemme, T.; Tsemekhman, K.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6946–6951.

[190] Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J.; Dror, R.; Shaw, D. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950–1958.

[191] Brereton, A. E.; Karplus, P. A. *Sci. Adv.* **2015**, *1*, e1501188.

[192] Bottaro, S.; Gil-Ley, A.; Bussi, G. *Nucleic Acids Res.* **2016**, *44*, 5883–5891.

[193] MacKerell, A. D.; Feig, M.; Brooks, C. L. *J. Am. Chem. Soc.* **2004**, *126*, 698–699.

[194] Buck, M.; Bouguet-Bonnet, S.; Pastor, R. W.; MacKerell, A. D. *Biophys. J.* **2006**, *90*, L36–L38.

[195] Guvench, O.; MacKerell Jr, A. D. *J. Mol. Model.* **2008**, *14*, 667–679.

[196] Shaffer, P.; Valsson, O.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 1150–1155.

[197] Gil-Ley, A.; Bussi, G. *J. Chem. Theory Comput.* **2015**, *11*, 1077–1085.

[198] Richardson, J. S.; Schneider, B.; Murray, L. W.; Kapral, G. J.; Immormino, R. M.; Headd, J. J.; Richardson, D. C.; Ham, D.; Hershkovits, E.; Williams, L. D.; Keating, K. S.; Pyle, A. M.; Micallef, D.; Westbrook, J.; Berman, H. M. *RNA* **2008**, *14*, 465–481.

[199] Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

[200] Olsthoorn, C. S.; Doornbos, J.; Leeuw, H. P.; Altona, C. *Eur. J. Biochem.* **1982**, *125*, 367–382.

[201] Ezra, F. S.; Lee, C.-H.; Kondo, N. S.; Danyluk, S. S.; Sarma, R. H. *Biochemistry (Mosc.)* **1977**, *16*, 1977–1987.

[202] Lee, C.-H.; Ezra, F. S.; Kondo, N. S.; Sarma, R. H.; Danyluk, S. S. *Biochemistry (Mosc.)* **1976**, *15*, 3627–3639.

[203] Tubbs, J. D.; Condon, D. E.; Kennedy, S. D.; Hauser, M.; Bevilacqua, P. C.; Turner, D. H. *Biochemistry (Mosc.)* **2013**, *52*, 996–1010.

[204] Karplus, M. *J. Chem. Phys.* **1959**, *30*, 11–15.

[205] Karplus, M. *J. Am. Chem. Soc.* **1963**, *85*, 2870–2871.

[206] Fürtig, B.; Richter, C.; Wöhnert, J.; Schwalbe, H. *Chembiochem* **2003**, *4*, 936–962.

[207] Sychrovskỳ, V.; Vokácová, Z.; Šponer, J.; Špacková, N.; Schneider, B. *J. Phys. Chem. B* **2006**, *110*, 22894–22902.

[208] Vokáčová, Z.; Bickelhaupt, F. M.; Šponer, J.; Sychrovský, V. *J. Phys. Chem. A* **2009**, *113*, 8379–8386.

[209] Haasnoot, C.; de Leeuw, F. A.; Altona, C. *Tetrahedron* **1980**, *36*, 2783–2792.

[210] Lee, C.-H.; Sarma, R. H. *J. Am. Chem. Soc.* **1976**, *98*, 3541–3548.

[211] Lankhorst, P. P.; Haasnoot, C. A.; Erkelens, C.; Altona, C. *Nucleic Acids Res.* **1984**, *12*, 5419–5428.

[212] Munzarová, M. L.; Sklenár, V. *J. Am. Chem. Soc.* **2003**, *125*, 3649–3658.

[213] Vokáčová, Z.; Trantírek, L.; Sychrovský, V. *J. Phys. Chem. A* **2010**, *114*, 10202–10208.

[214] Condon, D. E.; Yildirim, I.; Kennedy, S. D.; Mort, B. C.; Kierzek, R.; Turner, D. H. *J. Phys. Chem. B* **2014**, *118*, 1216–1228.

[215] Bottaro, S.; Di Palma, F.; Bussi, G. *Nucleic Acids Res.* **2014**, *42*, 13306–14.

[216] Cover, T. M.; Thomas, J. A. *Elements of Information Theory* **1991**, 279–335.

[217] Lin, J. *Information Theory, IEEE Transactions on* **1991**, *37*, 145–151.

[218] Endres, D. M.; Schindelin, J. E. *IEEE Trans. Inf. Theory* **2003**.

[219] Jafilan, S.; Klein, L.; Hyun, C.; Florián, J. *J. Phys. Chem. B* **2012**, *116*, 3613–3618.

[220] Brown, R. F.; Andrews, C. T.; Elcock, A. H. *J. Chem. Theory Comput.* **2015**, *11*, 2315–2328.

[221] Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *Angew. Chem. Int. Ed.* **1999**, *38*, 236–240.

[222] Frechet, D.; Ehrlich, R.; Remy, P.; Gabarro-Arpa, J. *Nucleic Acids Res.* **1979**, *7*, 1981–2001.

[223] Raiteri, P.; Laio, A.; Gervasio, F. L.; Micheletti, C.; Parrinello, M. *J. Phys. Chem. B* **2006**, *110*, 3533–3539.

[224] Hosek, P.; Toulcová, D.; Bortolato, A.; Spiwok, V. *J. Phys. Chem. B* **2016**, *120*, 2209–2215.

[225] Ferrarotti, M. J.; Bottaro, S.; Pérez-Villa, A.; Bussi, G. *J. Chem. Theory Comput.* **2014**, *11*, 139–146.

[226] Doemer, M.; Maurer, P.; Campomanes, P.; Tavernelli, I.; Rothlisberger, U. *J. Chem. Theory Comput.* **2013**, *10*, 412–422.

[227] Bottaro, S.; Banáš, P.; Šponer, J.; Bussi, G. *Submitted* **2016**.