Molecular Simulations to Investigate the Impact of Post-transcriptional Modifications on RNA Structural Dynamics

A THESIS FOR THE DEGREE OF
PHILOSOPHIAE DOCTOR (PH.D.)

Academic year 2022/2023

Student: Valerio PIOMPONI
Supervisor: Prof. Giovanni BUSSI

MOLECULAR
AND STATISTICAL
BIOPHYSICS

PHD COURSE IN PHYSICS AND CHEMISTRY OF
BIOLOGICAL SYSTEMS

SCUOLA INTERNAZIONALE SUPERIORE DI STUDI
AVANZATI - SISSA, TRIESTE

Contents

Abstract

Post-transcriptional modifications of RNA play a crucial role in shaping RNA structure and function, and in recent years have generated significant attention in the RNA research community. However, computational studies aimed at unraveling the structural and dynamic aspects of modified RNA remain relatively scarce. This is partly due to the prevalent use of the AUGC four-letter alphabet in structural prediction models, which neglects the possible occurency of modified nucleotides. Molecular dynamics (MD) simulations offer a powerful tool to explore RNA structural dynamics with virtually unlimited spatial and temporal resolution. Nonetheless, the accuracy of MD simulations is inherently tied to the quality of the employed force-fields, which consist of a list of parameters governing interatomic interactions, and their ability to accurately represent the complex and dynamic behavior of biomolecules at the atomic level. RNA molecules exhibit remarkable flexibility and dynamics, which pose significant challenges in characterizing their conformational behavior, with respect for example to double stranded DNA and globular proteins. Moreover, molecular dynamics (MD) simulations have struggled to accurately predict RNA structural dynamics, particularly for short and partially unstructured oligonucleotides. Consequently, the reliability of RNA force-fields has been a subject of scrutiny over the years, motivating the scientific community to invest significant efforts in optimizing and validating them, particularly with regard to the four standard nucleotides. However, in the context of modified RNAs, where limited validation against experimental data exists, this issue of reliability of the force-fields becomes even more pronounced. To address these challenges, our approach integrates MD simulations with experimental data, employing two distinct strategies: (i) refining force-fields for modified RNAs through fitting against experimental data and (ii) utilizing an ensemble refinement technique (maximum entropy) to guide simulations and enforce agreement with experimental observations. In this study, we apply these strategies to investigate the dynamic implications of specific post-transcriptional modifications within RNA molecules. Specifically, in one of our studies, we focus on investigating the influence of inosine hyper-editing on the structural dynamics of double-stranded RNAs (dsRNAs). We achieve this by improving the precision and accuracy of our predictions through the utilization of an enhanced sampling technique known as replica exchange collective variable tempering. Additionally, we apply the maximum entropy principle to constrain our simulations and incorporate solution experimental data (NMR and SAXS). Our findings illuminate the structural mechanisms through which inosine hyper-editing induces flexibility in dsRNAs, facilitating dynamic and non-canonical pairing as well as uncommon sugar puckering conformations. In our other studies, we investigate the effects of N6-methyladenosine (m^{6}A) on RNA structure and its role in RNA recognition by the YHT reader protein. To achieve this, we develop an alchemical free energy calculations procedure (AFEC), which allow us to quantitatively assess the impact of N6-methylation on the thermodynamic stability of dsRNAs and the free energy associated with RNA-protein complex formation. Additionally, we introduce innovative fitting strategies to fine-tune the m^{6}A force-field using AFEC, ensuring agreement between our simulations and experimental data from denaturation experiments, titration calorimetry, and NMR experiments. The AFEC calculations for the YHT-RNA complex were additionally integrated with metadynamics. This step was necessary to enhance the displacement of water molecules into and out of the protein binding pocket, aiming to improve the precision of the free energy estimates by effectively sampling the metastable hydrated states of the complex. Furthermore, we use m^{6}A-containing RNA systems to test a novel enhanced sampling technique called alchemical metadynamics (AM). This technique enables us to conduct alchemical transformations while simultaneously enhancing the exploration of the conformational space along a degree of free-

dom orthogonal to the alchemical variable. Our tests reveal that a single AM simulation can replicate the results obtained from two separate AFEC simulations for two different isomers. Additionally, it provides the capability of reconstructing the free energy profile along the biased torsional angle.

Chapter 1

Introduction

This thesis presents the culmination of four years of doctoral research conducted under the supervision of Professor Giovanni Bussi at the Scuola Internazionale di Studi Avanzati (SISSA). The primary objective of my work has been to explore the influence of RNA post-transcriptional modifications on RNA structural dynamics and recognition. We accomplished this by employing a combination of computational methodologies and by making use of available experimental data. This introductory chapter serves as an overview of general RNA biology and consolidates existing knowledge concerning RNA post-transcriptional modifications.

Chapter 2 focuses on elucidating the computational techniques employed throughout this study, which were integrated with experimental data obtained from solution experiments to generate the results discussed in this thesis.

Chapter 3 reports on a collaborative project between our computational group and the experimental team led by Professor Michael Sattler (Technische Universitat Munchen). In here, we investigate the conformational ensembles of an adeonisne-to-insoine hyper-edited dsRNA, combining molecular simulations with solution experiments. It is worth noting that this work has not been submitted yet.

Chapter 4 highlights the first publication arising from my doctoral research. In this chapter, we show how we can perform molecular simulations that match denaturation experiments for RNA systems containing N6-methyladenosine (m^6A), the most prevalent RNA post-transcriptional modification found in nature. We refine the m^6A force-field by fine-tuning a torsional potential and adjusting six partial charges of the nucleobase to accurately fit experimental free energies. Our fitting procedure makes use of alchemical free energy calculations (AFEC) to quantify the destabilizing effect of methylation on dsRNAs.

Chapter 5 provides an overview of our contribution to a published work that introduces an innovative enhanced sampling methodology called alchemical metadynamics. The lead author of this work, Wei-Tse Hsu, is a doctoral student in Michal Shirts' laboratory at the University of Colorado, Boulder. During my doctoral studies, I had the privilege of visiting Michal Shirts' lab for two weeks, laying the foundation for our collaboration. In this thesis, I briefly introduce the theory of alchemical metadynamics and exclusively report the results generated by myself, which consist in testing the method on some of the m^6A RNA systems previously investigated in Chapter 4.

Lastly, Chapter 6 details a collaborative project resulted from a two-month visit to Jiri Sponer's laboratory in Brno, Czech Republic. In this study, we investigate the role of m^6A in RNA recognition by a specific reader protein. Our investigation examines the impact of hydration within the protein binding pocket when assessing the influence of N6-methylation on the free energy of binding in the protein-RNA complex. Also for this work, we employed alchemical free energy calculations. Furthermore, we explore the effects of the m^{6}A force-field on the accuracy of these estimations. This extends the fitting procedure introduced in Chapter 4, ultimately resulting in a m^{6}A parametrization that can accurately predict dsRNA destabilizations, isomer populations, and RNA recognition simultaneously. This work has not been published yet.

The results discussed in this Thesis are based on the following articles:

- V. Piomponi, T. Fröhlking, M. Bernetti, G. Bussi, *Molecular Simulations Matching Denaturation Experiments for N^6-Methyladenosine*, ACS Central Science. 2022, 8, 8, 1218-1228 (See Chapter 4)

- Wei-Tse Hsu, V. Piomponi, P. T. Merz, G. Bussi, M. R. Shirts, *Alchemical Metadynamics: Adding Alchemical Variables to Metadynamics To Enhance Sampling in Free-Energy Calculations* , Journal of Chemical Theory and Computaion. 2023, 19, 1805-1817 (See Chapter 5)

- V. Piomponi, M. Bernetti, G. Bussi, *Molecular dynamics simulations of chemically modified ribonucleotides*, Chapter in the Springer Book *RNA Structure and Function.*

- C. Muller, V. Piomponi, G. Bussi, M. Sattler *Combining NMR, SAXS and MD to invesigate effects of A-to-I hyper-edting on RNA double strands*, not published yet (See Chapter 3)

- V. Piomponi , M. Krepl, J. Sponer, G. Bussi, *Molecular simulations to investigate the impact of N6-methyation in RNA Recognition: Improving Accuracy and Precision of free energy of binding estimation*, not published yet (See Chapter 6)

This thesis does not include a relevant portion of my research efforts, which involved supervising a Master's student, Axel Dian, for a duration of five months. Axel conducted an internship in our laboratory, where he admirably advanced a project aimed at developing a method to automatically find optimal pathway of alchemical parameters in alchemical free energy calculations. Given that the outcomes of this study are predominantly the result of Axel's diligent work, they are not presented within this thesis.

1.1 The Ribonulceic Acid (RNA)

RNA, or ribonucleic acid, is widely considered as one of the most important and versatile chemical species in molecular biology, and is present in all living cells. RNA is a single-stranded polymer molecule chemically related to DNA [1], and is mainly ad historically known for its primary function of acting as a key player in the transfer of genetic information from DNA to proteins. However, it can play many other fundamental roles in the cell: Different types of RNA, such as messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA), can perform specific roles in several cellular process. For example, they can store genetic information and catalyze chemical reactions at the same time. In bacteria, RNA controls gene expression in response to physiological stimuli [2]. In eukaryotic organisms, RNA is essential for the maintenance, regulation, and processing of genetic information, such as RNA silencing [3].

From a structural point of view, RNA is a linear polymer composed of nucleotides, which consists of a planar aromatic base attached to a ribose unit, a 5-member sugar ring, which is bound to a phosphate group. The 2'OH group of the sugar ring is a profound difference to DNA,

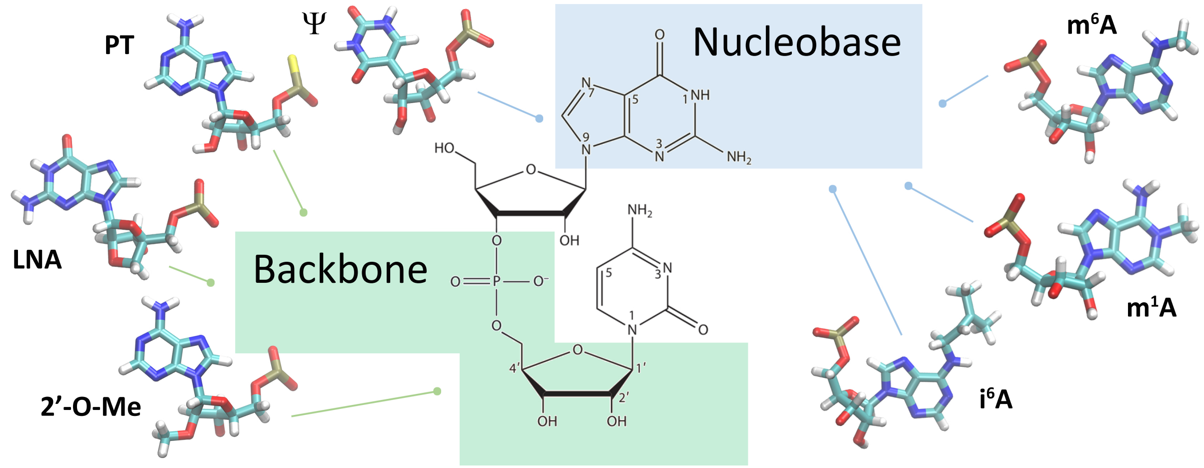


Figure 1.1 Schematic representing some of examples of nucleotide modified in the backbone (green) or in the nucleobase (blue). Readapted from our review [5] [6].

making RNA chemically significantly more versatile because this site can act as hydrogen bond donor and acceptor, enhancing the structural and dynamical complexity of RNA [1].

RNA molecules must adopt complex and functionally competent structures to carry out diverse cellular functions. The process of RNA folding is intricate and not fully understood, but it is evident that, similar to DNA, base-pair interactions play a crucial role in determining the stability of RNA structures [4]. Each nucleobase can be characterized by three edges: Watson–Crick (W), Hoogsteen (H), and Sugar (S). When two bases interact noncovalently, each engaging one of the three edges, they form a base pair with a substantially planar geometry, linked by at least two interbase hydrogen bonds (H-bonds). Canonical base pairs, G:C (3 hydrogen bonds) and A:U (2 hydrogen bonds), interact on the Watson-Crick edges and are primarily found in the double helical regions where two complementary strands are united. Although the G-U wobble base pair is often found in helical regions, it is not considered a canonical base pair. Indeed, the G-U wobble pairing involves a slight distortion of the base pairing geometry, resulting in weaker hydrogen bonding compared to the canonical pairs. Despite this, G-U wobble pairs remain significant in RNA secondary structures and play a crucial role in RNA folding and functional processes. They allow for flexibility and adaptability in RNA structures, contributing to the diversity of RNA's functional capabilities. Unlike double-stranded DNA, single-stranded RNA folds in on itself, giving rise to a wide range of noncanonical base pairs.

1.2 RNA modifications

RNA molecules are arbitrary sequences of four nucleotides that are used as building blocks: adenine (A), uridine (U), cytidine (C), and guanosine (G). These are however just the most commonly observed nucleotides. A large number of different (modified) nucleotides can be incorporated as well (see Fig. 1.1).

RNA modifications are of two types: naturally occurring and artificial. The former are biochemical modifications of nucleotides that are in most cases involving the nucleobase moiety. Many of them are chemical marks on cellular RNA and are regulated by enzymes generally referred as *writers*, and that can be eventually recognized by proteins referred as *readers*. The transformation of standard nucleotides into their modified version occurs in the cell after the transcription process has been completed in the nucleus, and for that reason naturally occurring modifications of RNA are generally called *post-transcriptional* modifications. The first modification was discovered more than 60 years ago [7], and nowadays, more than 100 types of post-transcriptional modifications are known. Historically, RNA modifications were thought to be present exclusively in noncoding RNAs (ncRNAs) and required for their regulatory function. In particular, transfer RNAs (tRNAs) are known to be heavily modified [8], and a wide variety of modifications can be found both in the anticodon region and in the tRNA-body region [9, 10]. The former are crucial to enhance the efficiency of the regulatory mechanism of protein synthesis, whereas the latter have in general a direct impact on structure, tuning the correct folding of the molecule into the well-known cloverleaf structure [11]. Ribosomal RNA (rRNA) is also extensively edited after transcription [12]. However, recent technical advances revealed widespread modifications also on messenger RNAs (mRNAs). A general overview on location, regulation, and function of modifications in the epitranscriptome can be found in Ref. [13]. In general, the roles of post-transcriptional modifications are of two types: (i) they allow correct folding of ncRNAs (e.g., tRNA and ribosomal rRNA) into their functional structure and (ii) they affect the target specificity of RNA-RNA, and RNA-protein interactions. In addition to naturally occurring modifications, a number of artificially modified nucleotides have been studied, usually aimed at increasing hybridization kinetics and stability [14]. Morevoer, it's worth remarking the implication of modified nucleotides in the development of COVID-19 mRNA vaccines, that is primarily related to enhancing the stability, efficiency, and immunogenicity of the mRNA molecules used in these vaccines [15]

Although the research on RNA modifications has been exponentially increasing in the past years, computational studies on modified RNAs are still limited, even in the relatively simpler context of secondary structure prediction [16]. This is due to two factors: first, the majority of models handling secondary structure predictions are limited to the standard 4 letter alphabet (AUGC), and as a consequence, to the standard Watson Crick and Wobble pairings (A-U, G-C, G-U); second, suitable thermodynamic parameters to estimate the effect of post-transcriptional modifications on duplex stability are lacking. Even more complex is the prediction of the impact of modifications on tertiary structure. It is worth recalling that typical models for tertiary structure predictions are trained on available structural datasets [17, 18, 19], and that the amount of RNA systems for which a high resolution structure has been obtained is limited. Needless to say, the statistics available on modified nucleotides is even scarcer. Furthermore, methods trained on static structures give limited access to structural dynamics. In this respect, molecular dynamics (MD) simulations [20, 1] are a very promising tool since (a) they give direct access to dynamics and (b) are grounded in physics-based models, which could possibly be capable to describe systems for which the amount of reference experimental structures is limited. Based on these assumptions, the objective of this thesis is to use molecular dynamics simulations to investigate how modified nucleotides impact RNA structural dynamics. This will be achieved through the integration of computational techniques with experimental data, aimed at improving the reliability of our findings. Additionally, we aim to develop predictive models for the tertiary structure of modified RNA, which can be utilized in future research.

Chapter 2

Methods

In this Chapter we will introduce the main computational methods used in the works described in this thesis. In particular, we will first describe the basic principles of molecular dynamics (MD) simulations, including advanced methods to enhance sampling and compute mutation free energies. Finally, we will show how MD simulations and experimental data can be integrated to improve accuracy of the model.

2.1 Molecular dynamics

Molecular dynamics simulations are a natural tool to characterize RNA structural dynamics [1]. In brief, they consist in solving the Newton's equations of motion for the system under investigation, propagating the coordinates of all the atoms for a number of consecutive steps [21]. Equations of motion are complemented with thermostats and barostats to control temperature and pressure, respectively. Water molecules and ions are usually explicitly represented, greatly increasing the number of simulated atoms. A key ingredient of any molecular dynamics simulations is the employed force-field. A force-field is a function that, given the current coordinates of all the atoms of the system, returns the forces acting on them. The force-field should be evaluated at each step of the MD simulation. Since evaluating the force-field is the computational bottleneck of any MD simulation, its functional form has to be chosen with compromises, so as to be accurate enough to describe the relevant chemistry but not too expensive. The functional form of the commonly used AMBER [22] force-field is the following one:

$$
\begin{aligned}
E = \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2 + \sum_{angles} \frac{1}{2} k_a (a - a_0)^2 + \\
\sum_{torsions} \sum_{n} \frac{V_n}{2} (1 + \cos(n\phi - \delta)) + \\
\sum_{LJ} 4\varepsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right) + \sum_{electrostatics} \frac{q_i q_j}{r_{ij}}
\end{aligned}
\tag{2.1}
$$

Here, k_b, k_a, and V_n control the so-called bonded interactions. Specifically, k_b controls the stiffness of chemical bonds, k_a the stiffness of angles between consecutive chemical bonds, and V_n can be used to provide a Fourier expansion of the energy controlling the rotation around chemical bonds. The remaining terms control non-bonded interactions, and are composed by Lennard-Jones potentials (σ and ε), representing both Van der Waals interactions and short range inter-atomic repulsion, and electrostatics, controlled by charges q.

The parameters of the force-field are heavily system dependent and are derived using a mixture of accurate quantum chemical calculations and of experimental data (see [23] for a recent review). Whereas currently available RNA force-fields are far from perfect, recent progress has allowed to design force-fields able to reproduce correctly the native conformation of small structured RNAs and the conformational ensembles of short disordered oligomers (see, e.g., Refs. [24, 25, 26] for recent works based on the AMBER force-field). The two main families of force-fields used for nucleic acids are AMBER [22] and CHARMM [27], both of which have evolved in multiple revised versions during the past decades. The AMBER family of force-fields offers a well-defined recipe to construct parameters for arbitrary molecules using quantum mechanical calculations [22]. In particular, charges are obtained by fitting the electrostatic potential, and torsional parameters by fitting the energy profiles associated to bond rotations. For the CHARMM force-field, the procedure is more complex and targets both quantum mechanical data on nucleoside and experimental data on nucleosides or oligonucleosides [27].

2.1.1 Force-fields for chemically modified nucleotides

Over the last decades, a lot of effort has been done to parametrize at best RNA force-fields. However, this effort was mostly done just taking into account the 4 standard nucleotides AUCG. In order to simulate RNA molecules containing modified nucleotides, it is necessary to derive specific force-field parameters for each type of modification. Luckily, force-field parameters for approximately 100 naturally occurring modified nucleotides were derived both in the AMBER [28] and in the CHARMM [29] frameworks. Aduri *et al* [28] published in 2007 the *modrna08* force-fields, which provides full parametrization for 107 naturally occurring modification. *Modrna08* was derived using the standard AMBER protocol, fitting torsions and charges to reproduce quantum mechanical calculations. Parameters were validated performing standard MD simulations of a tRNA containing a fraction of the modifications for which parameters were reported. These force-field parameters have been used in several later MD simulations using the AMBER force-fields. However, the parametrization has been shown not to be able to reproduce experimental evidence for a number of modified uridines [30]. Specifically, the parametrization was unable to reproduce conformational characteristics as expected from NMR experiments performed on these nucleotides, revealing the necessity to re-optimize the torsion angles for each individual modified residue and validate with larger RNA structures. The same authors successively re-optimize this force-field for the pseudoridine (Ψ), s^{2}U and s^{4}U by reparametrizing χ torsions (χ_{IDRP}) for all 3 nucleotides [31]. Moreover, they also proposed an alternative parametrization of the Lennard-Jones parameters for oxygen O3 of s^{2}U and s^{4}U (σ_{IDRP}), suggesting an increased σ for this atom so as to shift the population of the C3'-endo conformation of the sugar toward the experimental value. A later work by Dutta et. al [32] confirmed that the χ_{IDRP} parametrization was transferable to other modified uridines (a set of 4 methylated Ψ) Recently, an alternative force-field parametrization was published for Ψ and three different methylayed versions of Ψ [33]. The derivation follows the same stategy of [31], but this time new partial charges were derived using the RESP fitting method [34]. Furthermore, the parameters fitted on the single nucleotides were validated on ssRNA oligonucleotides, obtaining conformational and hydration characteristics in agreement with NMR experiments.

Also the parameters of m^{6}A have been validated quantitatively against experimental data. Hurst *et al*[35] computed the destabilization induced by the presence of the methyl group on RNA duplexes and showed that it can reproduce thermal denaturation experiments [36]. However, we will show in this Thesis that the Aduri force-field fails to reproduce denaturation experiments when considering a more extensive set of systems [36, 37], resulting in a mismatch in the duplex destabilization for sequences that were not tested previously as well as in an incorrect estimate of the relative stability between the two possible conformations of the methyl group. Only a simultaneous reparametrization of charges and a dihedral angle enabled to obtain both preference for the correct conformation of the methyl group and duplex destabilization in quantitative agreement with experiment, as we will show in Chapter 4.

Xu *et al*[29] adopted the CHARMM protocol to derive partial charges and bond potentials, with special attention to the glycosidic torsions. They presented in details 13 modified nucleotides but provided force-fields for 112. Automers and protonation variants have also been included. The parameters were optimized targeting quantum mechanical data and further refined against experimental data when possible. The charge fitting strategy aims at reproducing interactions with water and correct dipole moments. Torsions were fitted computing potential energy surfaces with quantum mechanical methods. Simulations of nucleosides and trinucleotides were compared with NMR data, when available. These force-field parameters have been used in several later MD simulations using the CHARMM force-fields. In addition to these two works, covering the majority of known modified nucleotides, it is relevant to mention that for many of the applications discussed below the authors developed and tested new sets of force-field parameters specific for a single or a few modifications [38, 39, 40, 41, 42, 43, 44, 45]. In this thesis we will take as the reference parametrization for modiefied nucleotides the *modrna08* force-field introduce by Aduri *et al* [28], and we will refer at it simply as Aduri force-field.

2.2 Enhanced sampling methods

RNA molecules are often characterized by conformational ensembles composed of multiple partly heterogeneous structures or substates that are relevant for function [46]. Molecular dynamics simulations can access at most the multi-microsecond timescale with current resources. In order to obtain statistically reliable population of substates, it is necessary to sample transitions between the important substates multiple times. Changes in tertiary interactions and modifications of multiple base pairings cannot thus be directly simulated with MD. To circumvent this problem, enhanced sampling methods can be used.

Enhanced sampling methods can be roughly classified in two categories. One category includes methods based on heating the system so as to accelerate the exploration of the conformational space, typically relying on replica exchange schemes to recover the original properties. Representative of these methods are parallel tempering, also called temperature replica exchange [47], and solute tempering [48]. These methods are typically very expensive and can thus be fruitfully applied only for sampling small oligomers. The other category includes methods based on adding biasing forces on specifically chosen degrees of freedom, or collective variables, representing the energetic barriers that one is willing to cross. Representatives of these methods are umbrella sampling [49], often performed combining multiple windows [50] so as to progressively convert the system from an initial conformation to a final one, and metadynamics [51]. These methods can be used to accelerate relevant events if sufficient prior information about the slow processes of the system is given. Methods which can be interpreted as a combination of the two different classes exposed above also exists, as for example the replica exchange collective variable tempering (RECT) method, which we will introduce below. A systematic review and classification of enhanced sampling methods can be found in Ref. [52], whereas a survey of applications to RNA simulations is presented in Ref. [53]. In the following, we will introduce the enhanced samplings methods that are used in the works described in this thesis.

2.2.1 Well-Tempered Metadynamics

Metadynamics [51] can be easily regarded as one of the most popular advanced sampling methods based on collective variables (CV). In metadynamics, a history-dependent bias potential is added to the system along a predefined CV, allowing the system to overcome energy barriers and explore different conformations more efficiently.

Although the formulation of meta dynamics can be generalized with respect to CV of any dimensionality, in practical applications the method can only be used for a small number of CVs at a time, due to the difficulty in reconstructing a highly dimensional bias with sufficient statistics. The biasing potential depends on the vector of collective variables of interest, denoted as ξ.

Throughout the simulation, biasing potentials are regularly added in the form of small Gaussian potentials, allowing to gradually fill the free energy basins across the relevant reduced-dimensional space defined by the CV, in such a way ensuring balanced sampling with respect to the CV. Let ξ be a d-dimensional CV vector, defined as $\xi = (\xi_1(x), \xi_2(x), ..., \xi_d(x))$. The biasing potential accumulated after a time period t can be expressed as follows:

$$V(\xi,t) = W \sum_{t'=k\tau,k\in N}^{t'<t} \exp\left(-\sum_{i=1}^{d} \frac{(\xi_i - \xi_i(x(k\tau)))^2}{2\sigma_i^2}\right) \tag{2.2}$$

where W is the height of the Gaussian, k is the number of Gaussian depositions, τ is the deposition stride and σ_i is the width of the Gaussian along the i-th dimension.

Well-tempered metadynamics (WT_MetaD) [54] is a variation of the standard metadynamics method. It introduces a "well-tempered" biasing factor to improve the convergence of the simulation. This factor modifies the height of the deposited Gaussian potentials during the simulation, effectively accelerating the convergence of the bias potential. Specifically, the time-dependent Gaussian height $W(k\tau)$ can be written as:

$$W(k\tau) = W_0 \exp\left(-\frac{V(\vec{\xi}(x(k\tau)),k\tau)}{k_B \Delta T}\right) \tag{2.3}$$

where W_0 is the initial Gaussian height and ΔT is a temperature parameter that incorporates a user-defined bias factor $\gamma = (T + \Delta T)/T$ for adjusting the decay rate of the bias. In WT_MetaD, the relation between the accumulated bias potential and underlying free energy surface as a function of the multi-dimensional CV ξ can be estimated as follows:

$$V(\xi,t \to \infty) = -\frac{\Delta T}{T + \Delta T} F(\xi) = -\left(1 - \frac{1}{\gamma}\right) F(\xi) \tag{2.4}$$

In all our applications, we perform WT_MetaD by interfacing GROMACS with the PLUMED package [55]. When setting up WT_MetaD, several considerations must be taken into account when choosing the input parameters γ; τ; W_0; and the σs. The values of these parameters may be crucial to allow proper convergence of the bias potential, and even more importantly, a diffusive behavior of the CV. These choices are system dependent and will be discussed for each application.

When performing metadynamics, it's sometimes considered a good practice to perform a separate simulation using the static bias potential produced by the previous simulations, and compute the weighting factors exclusively from the additional simulation, as it is done, for instance, in metadynamics with umbrella-sampling refinement [56]. Notice that this option is more expensive, as it requires a separate simulation, but in principle removes any potential systematic error due to the history-dependent nature of the metadynamics biasing potential.

In order to obtain reasonable estimates of the free energy difference of interest, the dimensions of the chosen set of CVs must be as low as possible, while still being able to capture the slowest degrees of freedom of the system to an extent which allows to observe multiple transitions, back and forth, along the states which compose the reduced-dimensional space defined by the CV. In the cases in which a low number of CVs is not sufficient to capture the slow degrees of freedom, it becomes prohibitive to get to converge weights, as the space to be explored increases exponentially with the number of CVs. In the next section we will describe a method that allows to tackle the issue, by integrating an high number of independent metadynamics in an Hamiltonian replica exchange scheme.

2.2.2 Replica Exchange Collective Variable Tempering

Replica Exchange Collective Variable Tempering (RECT) is an enhanced sampling method introduced by Gil-Ley *et al* [57], which combines two techniques: Hamiltonian replica-exchange (HREX)[58] [48] and the previously introduced well-tempered metadynamics (WT-MetaD) [54]. HREX is an extension of the replica exchange method, where each replica of the system is simulated using a different Hamiltonian, which represents a different potential energy function. If the Hamiltonian varies only in terms of temperature, the method corresponds to parallel tempering. Alternatively replicas can differ in terms of potential energy parameters. Swaps of coordinates between neighboring replicas are periodically attempted and accepted or rejected with a Metropolis criterion, with the goal of allowing every continuous trajectory to go back and forth in the replica ladder as fast as possible. By simulating replicas at different Hamiltonians, the system can explore a wider range of conformational space and access states that might be difficult to reach using traditional molecular dynamics simulations. Well-tempered metadynamics, as the other methods based on adding a bias along a CV, has the limitation of being applicable only for a small set of CVs, due to the difficulties in building a history dependent potential in a high-dimensional CV space. For many systems it is not possible to find a small number of effective CVs that describe the slow degrees of freedom. The RECT method is based on the idea to perform simultaneously a large number of concurrent metadynamics biasing different local CVs, by integrating them in an HREX scheme. The methods exploits the tunability of WT-MetaD to scale the strength of the bias potential along the ladder of replica, by acting on the bias factor γ. By setting $\gamma = 1$ for the bottom replica, we in practice obtain an unbiased replica which is still able to enhance its sampling thanks to the exchange with the upper replica, along which γ is gradually scaled to bigger values. This 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.

In Chapter 2, we will show our use of this method to enhance the sampling of a 20-bp dsRNA, in such a way as to ensure an exhaustive sampling of all the possible sugar puckering conformations of the dsRNA nucleotides. In this application, we integrate 24 concurrent metadynamics in a 8 replica ladder.

2.2.3 Alchemical Free Energy Calculations

Alchemical methods allow to simulate trajectories where molecular species are mutated to different molecular species, and the free-energy associated to the transformation can be computed [59]. These methods are commonly used to characterize the effect of a given chemical modification on the relative stability of two conformations. Most molecular dynamics code support these methods, but setting up the simulations is usually more complex than for standard MD.

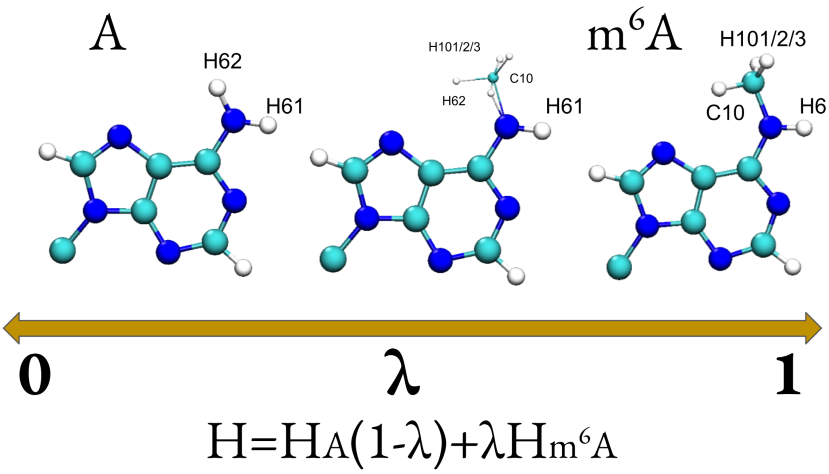


Figure 2.1 Alchemical transformation of the standard adenosine into the N^6-methyladenosine (m^{6}A). The transformation is performed by integrating along an alchemical path λ, by switching *on/off* non-bonded interaction of specifically chosen atoms.

The intermediate states might be simulated independentely of each other or with a more robust replica exchange procedure [60].

For several works presented in this thesis, we set up an alchemical free rnergy calculation (AFEC) protocol which allows integrating along an alchemical path describing the transformation of a standard adenosine (A) into the N6-methyladenosine (m^{6}A) (see Fig. 2.1). The transformation consists in substituting the hydrogen H62 with a methyl group defined by atoms C10, H101, H102 and H103, by gradually switching *on/off* the non-bonded interaction of these atoms. To this extent, we included a hybrid adenosine with double topology in the force-field definition: the first topology corresponding to standard adenosine, and the second one corresponding to m^{6}A. We used 16 replicas in which Lennard-Jones parameters and partial charges were simultaneously interpolated. In our AFEC protocol, we make use of Hamiltonian replica exchange (HREX) scheme, proposing exchanges every 200 fs.

An important choice in setting up AFECs is the number of intermediate replicas and the optimal form of the intermediate Hamiltonian functions. In order to avoid singularities due to electrostatic interaction when the repulsive LJ potential is switched off [59], we used the GROMACS implemented soft core potentials to interpolate Lennard-Jones and Coulomb potentials as follows

$$V_{sc,\lambda}(r) = (1-\lambda)V_A((\alpha\sigma^6\lambda + r^6)^{\frac{1}{6}}) + \lambda V_{m^6A}((\alpha\sigma^6(1-\lambda)^6 + r^6)^{\frac{1}{6}}) \tag{2.5}$$

where $\alpha = 0.5$ and $\sigma = 0.3$ nm. Here, V_A and V_{m^6A} are the Lennard-Jones and Coulomb potential energy functions for unmodified and modified adenine, respectively, $V_{sc,\lambda}$ is the interpolated version of the function, and r the interatomic distance. The energy of the only torsion (η_6) scaled in the A-to-m^{6}A transformation is instead a linear combination of the energy of the two end points with factors $1-\lambda$ and λ, and is defined as

$$V_{\eta_6,\lambda}(x) = (1-\lambda)V_{\eta_6,A} + \lambda V_{\eta_6,m^6A} \tag{2.6}$$

In our case, $\lambda = 0$ denotes the parameters of the unmodified adenine, whereas $\lambda = 1$ those of the modified adenine.

We decided not to scale the bonded interactions that are present in only one of the systems (e.g., torsional parameters associated to the methyl group), but to rather have a single H in one of

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
set1	0.0	0.01	0.03	0.05	0.10	0.2	0.35	0.45	0.55	0.65	0.8	0.9	0.95	0.97	0.99	1.0
set2	0.0	0.02	0.05	0.09	0.14	0.2	0.3	0.43	0.57	0.7	0.8	0.86	0.91	0.95	0.98	1.0

Table 2.1 Sets of lambda coefficients used in AFEC replica exchange simulations for systems with a single methylation (set 1) or with two methylations (set 2).

the topologies and a CH_3 group in the other topology. These groups are present in both systems, though with their nonbonded interactions switched off at one of the end points. This implies that the intermediate topologies contain both C10 and H62 (see Fig. 2.1). Also the torsional potentials controlling the rotation of the amino group in the unmodified nucleotide and of the carbon in the modified nucleotide are not scaled. These potentials are symmetric with respect to *syn*/*anti* rotations (see other Chapters), and thus do not influence the *syn*/*anti* population. We instead scaled the potential acting on η_6 since this torsional potential is not symmetric and its presence would lead to a *syn*/*anti* balance different from zero in the unmodified nucleotide. Readers interested in reproducing this setup are encouraged to inspect the GROMACS topology files provided in the Zenodo archive (link: https://doi.org/10.5281/zenodo.6498021).

We also notice that GROMACS allows setting separate scaling factors for electrostatic, Lennard-Jones, and bonded interactions. We didn't exploit this feature, and rather scaled all interactions with the same λ factor. This still leaves the open issue of placing a sufficient number of λ factors interpolating between 0.0 and 1.0. In a replica-exchange setting, the acceptance rate can be used as a measure of the phase-space overlap between adjacent ensembles. A minimum acceptance is then required to enable mixing of ensembles. At the same time, the spacing in λ required to reach this minimum acceptance might differ in different regions of the λ space, thus leading to an optimal allocation of replicas that is not uniformly spaced in λ.

We decided to use a single system, that is the *stand alone* nucleoside in solution, to optimize this ladder and then reused the same parameters for all systems. Specifically, we empirically adjusted the λ values until we obtained a set of 16 intermediates (set1 in Table 2.1) leading to an approximately uniform acceptance rate each of them greater than 20%. As it can be understood from the table, the density of the chosen λ values is inhomogenous and, in particular, higher close to the boundaries ($\lambda = 0$ or 1). This set of lambdas was then used for all the AFECs presented in our works where a single adenine is methylated, and as expected lead to an acceptance greater than 20% for most replica pairs, and greater than 10% for all replica pairs. In cases where two methylations were included, we found that for some pairs of replicas the acceptance was significantly lower than 20%. We notice that in principle the presence of two simultaneous methylation should lead to a larger number of replicas required to obtain the same acceptance. By reoptimizing the parameters, we obtained a set of 16 λ's (set 2 in Table 2.1) that was able to guarantee an acceptance greater than 20% for all transitions in systems with two methylations. We remark that we are currently working to implement an algorithm yielding an optimized multidimensional ladder using data from a short simulation with the alchemical parameters shown in table 2.1. The method is transferable to any type of alchemical free-energy calculations and can be use to find a pathway of alchemical parameters that enhance replicas mixing. Since this work was predominantly overseen by visiting Master student Axel Dianb, the results are not presented within this thesis.

The HREX scheme allows, at the end of the production phase, to recover 16 independent "demuxed" (i.e., continuous) trajectories, which can then be processed to recompute energies for each of the 16 Hamiltonian functions to compute ΔG via binless weighted histogram analysis method (WHAM) [61, 62, 63]. Specifically, for each trajectory, a weight w is found for each snapshot x that allows computing statistics for the unmodified adenine as a weighted average over the set of concatenated replicas. We consider a set of N trajectories obtained using different value of λ, so that λ_k denotes the value of λ and n_k the number of snapshots in the k-th simulation. The k-th trajectory will thus contain samples from the distribution $P_k(x) \propto e^{-\beta E_{\lambda_k}(x)}$, where $E_\lambda(x)$ is the energy associated conformation x for a given λ and β is the inverse of the thermal energy. We are interested in obtaining weights $w(x)$ that can be used to compute averages corresponding to a reference value of λ ($\lambda = 0$). In other words, for any observable $O(x)$, its average at $\lambda = 0$ is obtained as

$$\langle O \rangle = \sum_x w(x) O(x) \tag{2.7}$$

where the sum runs over the concatenation of the N trajectories. By using the WHAM method in its binless formulation [61, 62, 63], the unnormalized weights can be obtained as:

$$w(x) \propto \frac{1}{\sum_{k=1}^{N} n_k e^{-\beta(E_{\lambda_k}(x) - E_0(x))} Z_k^{-1}} \tag{2.8}$$

and subsequently normalized scaling them by a factor ensuring that $\sum_x w(x) = 1$. The partition function associated with each value of k, Z_k, can be obtained as

$$Z_k = \sum_x w(x) e^{-\beta(E_{\lambda_k}(x) - E_0(x))} \tag{2.9}$$

These two equations should be solved self-consistently. For numerical purposes, it is convenient to initially remove from the computed energies ($E_{\lambda_k}(x)$) their minimum along both k and x so as to avoid numerical overflows in the calculation of the exponential function, and then add the corresponding contributions to the resulting free energies and to the logaritm of the weights. The calculation was performed using the wham tool available in the bussilab python package, which can be obtained at `https://github.com/bussilab/py-bussilab`, version 0.0.36, and that can be used as a reference for the exact numerical procedure used here. Once the weights have been obtained, they can be used to compute the free-energy difference between the two end states using the following equation:

$$\Delta G^{AFEC} = -k_B T \log \left[\frac{\sum_x w(x) e^{-\beta \Delta E(x)}}{\sum_x w(x)} \right] \tag{2.10}$$

where $\Delta E(x) = E_{\lambda=1}(x) - E_{\lambda=0}(x)$ is the difference between the total energy computed with the Hamiltonian energy functions associated to m^{6}A and adenosine, respectively.

Notably, the weights only depend on the conformation (x) and not on the specific value of λ at which the conformation was generated. This implies that trajectories can be concatenated in any order resulting in identical weights. This allows to concatenate “demuxed” (continuous) trajectories, which are virtually independent of each other, being coupled only through the exchange step. By performing a blocked bootstrap with block size identical to the length of each trajectory, one ensures that correlations are minimized [64]. In this case, however, one should explicitly take into account that, for a given bootstrap sample, the number of snapshots generated at each value of λ will differ. To computed statistical error on Free Energies, we used the bootstrapping procedure resampling the 16 continuous trajectories 200 times with replacement. Finally, as a control, we always computed ΔGs also using the standard Bennett-acceptance-rate estimate implemented in GROMACS [65, 66].

The free-energy change associated to an alchemical transformation where the number or type of atoms in the two end states is different has no physical meaning. Once the protocol for mutating one chemical form to another has been established, the simulation should be repeated in different structural contexts. For instance, the conversion between (unmodified) A and (modified) m^{6}A can be performed in a single strand and in a duplex. The difference between the free-energy changes computed in the two simulations corresponds to the stabilization of duplex resulting from the additional methylation (see fig 4.2a). In other words, alchemistry does not allow the calculation of the hybridization free energy, but rather how much the hybridization free energy is affected by the chemical modification. Similarly, the impact of the modification on the affinity between the studied RNA and a protein can be estimated. Once the setup for the alchemical simulation has been prepared, the simulation and analysis steps are relatively simple. However, the results should be judged with care. A particularly problematic case is when slow degrees of freedom are coupled with the alchemical change. One should make sure that sampling of the conformational degrees of freedom is sufficient. Typical cases would be if the rotation around a chemical bond with a high free-energy barrier is coupled with the alchemical change. An even more difficult situation arises when one of the alchemical states, or both, are significantly flexible conformations whose sampling is difficult. Simulations where one of the end states is a flexible single-stranded RNA should then be performed and analyzed with care to rule out potential artifacts. A possible improvement that can alleviate this problem consists in combining alchemical simulations with enhanced sampling methods, as done for instance in alchemical metadynamics, as discussed below [67].

2.2.4 Alchemical Metadynamics

Standard alchemical methods can fail in scenarios where the most important slow degrees of freedom in the configurational space are, for the most part, orthogonal to the alchemical variable, or if the system becomes trapped in a deep basin extending in both the configurational and alchemical space. For example, in the A-to-m^{6}A Alchemical Free Energy Calculations (AFEC) procedure described in the previous section, it would be impossible, within a single simulation on the timescale of ns-µs, to sample the two possible m^{6}A isomers (*syn* and *anti*), as their transition kinetics are expected to occur on the timescale of ms [68]. Alchemical metadynamics (AM), recently proposed in Ref [67], allows overcoming these limitations by performing a 2 dimensional metadynamics. In this approach, one dimension corresponds to the alchemical transformation, while the other dimension enhances the sampling with respect to a collective variable (CV) that describes the slow degrees of freedom of the system. With respect to the Metadynamics formalism introduced in section 2.2.1, the alchemical variable λ is introduced in the generalized CV vector $\xi' = (\lambda, \xi_1(x), \xi_2(x), ..., \xi_d(x))$ such that the joint space of λ and ξ is sampled with the aid of the biasing potential $V(\xi')$. Identically to the AFEC describe in previous section, the alchemical variable is not a function of atomic coordinates and it can take discrete values which govern the interpolation between two different Hamiltonians. In AM the discrete λ ladder is explored through Metropolized-Gibbs algorithm (Monte Carlo sampling), similarly to expanded ensemble method [69], while the coordinate direction is sampled by molecular dynamics as in any other type of metadynamics. As the multi-dimensional biasing potentials can flatten out the free energy landscape in both configurational and alchemical space, one can try to ensure that the system would not get stuck in the phase space.

Theoretically, the free energy estimator for alchemical metadynamics is the same as the one used in any other metadynamics, except that the CV vector is generalized with the introduction of the alchemical variable. Upon the deposition of the biasing potential $V(\xi')$ in alchemical metadynamics, the probability distribution sampled during the simulation is $\tilde{P}(\xi') \propto \exp(-\beta(F(\xi') + V(\xi')))$. One of the possible options to recover the underlying free energy landscape $F(\xi') = -k_\mathrm{B}T \ln P(\xi')$, is to reweight the histogram by assigning an unbiasing weight $w(\xi')$ to each sample with the CV ξ'. [70] Such an unbiasing weight can be expressed as

$$w(\xi') \propto \exp\left(\frac{V(\xi',t_f)}{k_\mathrm{B}T}\right) \tag{2.11}$$

where t_f is the simulation length and $V(\xi',t_f)$ is the total bias accumulated up to t_f. The maximum of $V(\xi',t_f)$ over ξ' is usually subtracted before taking the exponential to avoid overflow, without affecting the normalized weights. More frequently, $V(\xi',t_f)$ is replaced with $\bar{V}(\xi',t_0)$, the total bias averaged over the time period from $t_0 = (1-f_a)t_f$ to t_f [71], where f_a is the fraction over which biases are averaged. Given that $t_f = t_0 + N\tau$, $\bar{V}(\xi',t_0)$ can be written as

$$\bar{V}(\xi',t_0) = \frac{1}{N+1}\sum_{i=0}^{N} V(\xi',t_0+i\tau) \tag{2.12}$$

where N is the number of Gaussians deposited from t_0 to t_f. In Chapter 5 we will show an application of the AM on the A-to-m^{6}A transformation, coupling the alchemical CV to a torsional CV which allows sampling two different isomers within a single simulation.

2.3 Combining MD and experiments

The accuracy of molecular dynamics, which determines its ability to replicate and predict experiments, is often constrained by the quality of the employed force-fields. Despite recent advancements [1], the quality of RNA force-fields remains a limiting factor for MD simulations. However, substantial efforts have been dedicated over the years to parametrize standard RNA nucleotides. Although these force-fields heavily rely on parameters developed in the 1990s, they have been validated and compared against a significant volume of experimental data. This validation has provided insights into the reliability of MD in various systems while identifying potential artifacts in others [72] [1] [26]. Conversely, as discussed in section 2.1.1, limited progress has been made in parameterizing and validating force-fields for modified nucleotides. Consequently, the utilization of experimental data from literature becomes even more critical to corroborate computational findings and ensure their reliability.

Two primary strategies exist to integrate simulations and experiments. The first, as previously mentioned, involves fitting force-fields against experiments, trying to ensure their transferability to other systems not involved in the training. Although historically applied to small fragments, this approach can be theoretically extended to macro-molecular systems, as also shown in this Thesis. The second strategy involves ensemble refinement, exemplified by methods like Maximum Entropy (ME) or Maximum Parsimony (MP) principles. These techniques enable the reweighting of MD trajectories to enhance their alignment with experimental observables. These methods are not transferable, in the sense that experimental data should be available for exactly the same system that one wishes to model.

Within this thesis, we demonstrate applications of both strategies, by combining MD simulations with solution-based experiments to investigate the impact of RNA modifications on RNA structural dynamics and recognition.

2.3.1 Force-field Fitting

RNA force-fields, much like other force-fields for biomolecules, have traditionally been developed using a 'bottom-up' approach. This involves fitting a combination of reference quantum chemistry data and experimental information obtained from small molecular fragments. In the specific case of the AMBER force-field, parameters governing bonded interactions are derived from experimental data, often obtained through spectroscopy experiments. On the other hand, parameters related to torsional interactions and Lennard-Jones (LJ) parameters, which govern non-bonded interactions, are determined using a mix of quantum mechanical calculations, empirical fitting, and experimental data. Furthermore, the partial charges employed to describe electrostatic interactions are exclusively derived from quantum mechanical calculations. A challenge with these conventional methodologies is ensuring that the derived parameterization are transferable, that means they can be successfully applied to different systems. For instance, torsional parameters and partial charges in the AMBER force-field are initially calculated using quantum chemistry on small fragments containing only a few dozen atoms, often including just a few amino acids. However, these parameters are subsequently used to simulate larger systems like oligopeptides or entire protein domains.

In recent years, the proliferation of solution experimental studies, coupled with the emergence of machine learning techniques capable of achieving transferable outcomes in the fitting process, has paved the way for directly integrating a multitude of molecular dynamics simulations with experimental data on macromolecules. This direct integration allows for the development of transferable parameterizations. Several methods have been introduced to facilitate the direct fitting of force-fields to experimental data obtained from macromolecular systems. The general workflow employed for this fitting process follows the subsequent steps:

- Derive initial parameterization by fitting quantum chemistry data and experimental data on small systems
- Perform MD simulations with the initial reference force-field on macromolecules systems
- Reweight trajectories optimizing force-field parameters in order to maximize the agreement with a set of available data
- Re-perform MD simulations with the new optimized parameters

The process of optimizing force-field parameters involves reassigning new weights to examined conformations, then predicting results based on these slightly adjusted parameters. However, if there's a need to enforce experimental data by exploring parameters significantly different from the reference ones, there's a potential risk of compromising the statistical significance of the calculations. At a certain point, it becomes essential to iterate the procedure, by performing new simulations, at least one time in order to validate the predictions obtained with reweighting.

The advantage of fitting against solution experiments is that they often report results that are averaged over an ensemble of copies of the same molecule, making it possible to enforce ensemble averages rather than instantaneous values. Another crucial aspect to consider when fitting a model to experimental data is the avoidance of overfitting. Fortunately, in recent years, the field of machine learning has introduced several tools that can be employed to incorporate regularization terms into optimization algorithms, thereby mitigating the risk of overfitting. Typically, fitting algorithms operate by minimizing a cost function, which can be augmented with regularization terms guided by hyperparameters. These hyperparameters must be selected through a cross-validation procedure. This procedure involves fitting the model to the training set data and subsequently validating the optimized model against the validation set data, which is distinct from the training data. Alternatively, the dataset can be divided into three subsets: in addition to the standard training and cross-validation subsets, an independent test set is introduced. The training set is utilized to identify the optimal parameter values at fixed hyperparameters. Optimal hyperparameters are then determined using the cross-validation set. Eventually, the performance of the model, defined by the optimal parameters and hyperparameters, is assessed using the test set. For a collection of examples detailing the fitting of RNA force-fields to macromolecular systems, refer to the review by Frohlking *et al* [23].

In Chapter 4 of this thesis, we will present our efforts to refine the force-field for a methylated Adenosine. This involves reparameterizing partial charges and a torsional term to match simulations with experimental data available in the literature, that are denaturation experiments and NMR data pertaining to isomeric populations of both paired and unpaired nucleotides.

2.3.2 Ensemble Refinement - Maximum Entropy

Ensemble refinement methods aim to corroborate simulation with experimental data without modifying the force-fields, but rather by adding minimal bias in order to improve quantitative agreement with experimental values while minimizing the change in the potential energy function. One of these biasing methods build upon the maximum entropy (ME) principle, which in its original formulation [73] states that, given a system described by a number of states, the best probability distribution for these states compatible with a set of observed data is the one maximizing the associated Shannon's entropy. The entropy is computed here relative to a given prior distribution $P_0(\mathbf{q})$ and, is defined as

$$S[P][P_0] = -\int d(\mathbf{q})\, P(\mathbf{q}) ln \frac{P(\mathbf{q})}{P_0(\mathbf{q})} \tag{2.13}$$

Given a set of M experimental observables to be enforced, the Shannon entropy should be maximized based on the following costraints:

$$\begin{cases} P_{ME}(\mathbf{q}) = \arg\max_{P(\mathbf{q})}(S[P][P_0]) \\ \int d\mathbf{q}\, s_i(\mathbf{q})P(\mathbf{q}) = s_i^{exp}; i = 1,..,M \\ \int d\mathbf{q}\, P(\mathbf{q}) = 1 \end{cases} . \tag{2.14}$$

This system of equation can be interpreted as a search for the posterior distribution $P_{ME}(\mathbf{q})$ that is as close as possible to the prior distribution $P_0(\mathbf{q})$ among those which agree with the given experimental observations.

The solution of 2.14 can be obtained using the method of Lagrangian multipliers, namely searching for the stationary points of the Lagrange function

$$\mathscr{L} = S[P][P_0] - \sum_{i=1}^{M} \lambda_i \Big(\int d\mathbf{q}\, s_i(\mathbf{q})P(\mathbf{q}) - s_i^{exp}\Big) - \lambda_0\Big(\int d\mathbf{q}\, P(\mathbf{q}) - 1\Big) \tag{2.15}$$

where the λs values correspond to the Lagrangian multipliers. By setting $\frac{\delta \mathscr{L}}{\delta P(\mathbf{q})} = 0$, we find that the posterior which maximize the Shannon entropy can be expressed, neglecting the normalization factor, as:

$$P_{ME}(\mathbf{q}) \propto e^{-\sum_{i=1}^{M} \lambda_i s_i(\mathbf{q})} P_0(\mathbf{q}) = e^{-\boldsymbol{\lambda} \mathbf{s}(\mathbf{q})} P_0(\mathbf{q}) \tag{2.16}$$

where the vector of Lagrangian multipliers $\boldsymbol{\lambda}$ are those which allows to enforce the experimental averages.

In short, the maximum relative entropy principle gives a recipe to obtain the posterior distribution that is as close as possible to the prior distribution and agrees with some experimental observation. In order to apply this formalism to MD trajectories, we can assume that $P_0(\mathbf{q})$ is the probability distribution predicted by the used force-field through the MD simulation. Reweighting the trajectory through equation 2.16 corresponds to individuate a Boltzmann distribution that is the result of using a potential energy defines as:

$$V_{ME}(\mathbf{q}) = V_0(\mathbf{q}) + k_B T \boldsymbol{\lambda} \mathbf{s}(\mathbf{q}) \tag{2.17}$$

where $V_0(\mathbf{q})$ is the potential energy given by the used force-fields, and the second term correspond to the minimal biasing needed to adjust the predicted ensemble with respect to the M experimental data considered.

In short, the choice to generate an ensemble that is as close as possible to the prior knowledge (the force-field) implies that the correcting potential has a specific functional form, that it is linear in the observables that have been measured. In recent years, the ME method has been widely used to enforce experimental averages on MD simulations of RNA systems. Some of the most common data used to improve the accuracy of the simulations include ^{3}J scalar couplings and NOE signals derived from solution Nuclear Magnetic Resonance (NMR) experiments [74], as well as Small-Angle X-ray Scattering (SAXS) spectra [75]. More recently, Cryo-Electron Microscopy (Cryo-EM) density maps have also been utilized [76]. In all these cases, a so called *forward model* is needed to map the atomic coordinates of the system to the measured quantity, allowing the experimental data to be back-calculated from the simulated structures. It must be noted that the formulas used in standard forward models are often parameterized empirically and may not be extremely accurate, introducing in such a way systematic errors in the ensemble refinement. More generally, as in force-field fitting methods, in the ME refinement, some regularization terms could be introduced in order to strike a balance between trusting the pure MD predictions and placing trust in the forward models and/or the experimental data.

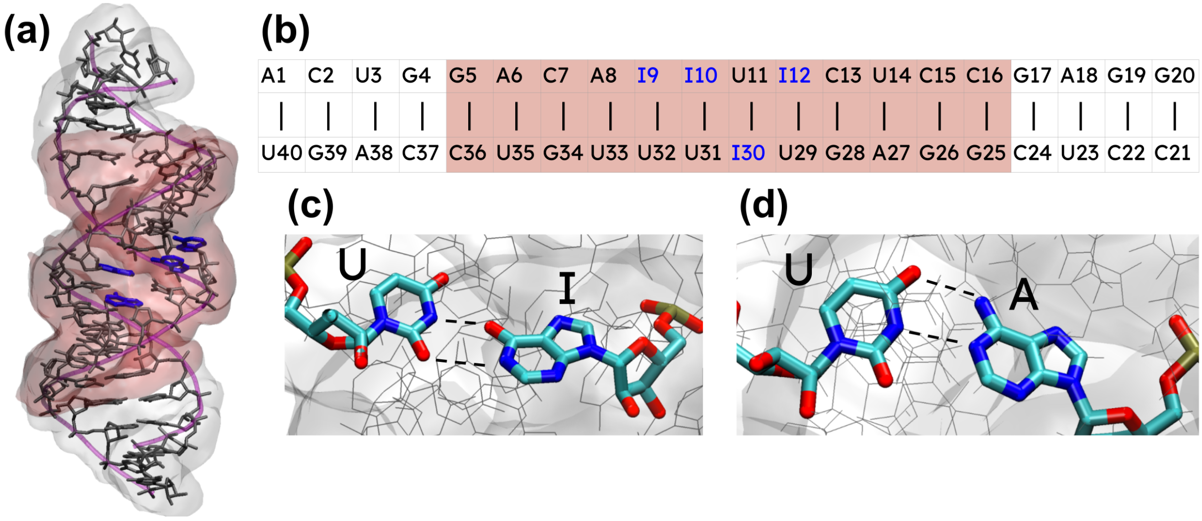
ME has also other limitations. First, only variables for which experimental data are available can be refined. Additionally, it is not possible to modify the functional form of the force-field. In this sense, the corrections derived using the ME principle are not transferables to other systems.

In the upcoming Chapter, we will show an application of the ME principles, used to incorporate both NMR and SAXS data in predicting an ensemble of structures for a double-stranded RNA that undergoes hyper-Adenosine-to-Inosine editing.

Chapter 3

Combining MD and solution experiments to investigate the impact of inosine hyper-editing in dsRNA

Inosine is a naturally modified nucleotide that has been widely studied among the RNA community. It was initially identified in 1965 during the analysis of RNA transferase [77] and plays a crucial role in facilitating the accurate translation of the genetic code in wobble base pairs. Inosine is synthesized through the deamination of adenine, a chemical process that involves the removal of an amino group (-NH2). Consequently, this process converts the amino group of adenine into a keto group, resulting in the formation of a novel nucleobase called hypoxanthine. This transformation is referred to as adenosine-to-inosine (A-to-I) editing [78]. The distinct aspect of inosine lies in its ability to form pairs with various nucleobases during RNA transcription and translation processes [79]. In RNA, base pairing preferentially follows the principles of complementary base pairing. The most stable parings, also called canonical pairings, involve adenine (A) binding with uracil (U) and cytosine (C) pairing with guanine (G). Alternatively, non-canonical interactions within RNA structures can involve U binding with G, and inosine (I) pairing with both U or C. This inosine's remarkable versatility arises from its capacity to establish hydrogen bonds with all these nucleobases, albeit with reduced strength compared to canonical pairings. Inosine's capability to pair with multiple bases renders it particularly significant in scenarios demanding flexibility in codon-anticodon recognition, such as the wobble position of the anticodon loop in transfer RNA (tRNA) during protein synthesis. This adaptive feature allows tRNAs containing inosine to identify multiple codons differing in the third position, thereby enhancing translation efficiency [80]. A-to-I editing in double-stranded RNA (dsRNA) is also known to play a major role in regulating immune response. Indeed, intracellular dsRNAs are perceived as a threat by the cell, including those originating from RNA viruses as well as genomic and transcriptomic elements containing inverted repeat sequences that can form dsRNA regions. This situation activates an immune response that could harm the cell. To mitigate this response, cells employ the A-to-I editing mechanism. In cytosolic dsRNA regions, an enzyme called ADAR1 converts a substantial number of adenosines into inosines through a process called hyper-editing. This hyper-editing inhibits the immune response, as the hyper-edited RNA can no longer be recognized by protein factors. Notably, studies involving mice demonstrated the indispensability of ADAR1, as the absence of this enzyme rendered the animals nonviable [81]. The inability of immune factors (MDA5 and RIG-I) to interact with hyperedited dsRNA suggests the presence of an altered conformation. This is a focal point of our investigation: deciphering the structural modifications occurring in hyperedited RNAs. Ad-



A1	C2	U3	G4	G5	A6	C7	A8	I9	I10	U11	I12	C13	U14	C15	C16	G17	A18	G19	G20
U40	G39	A38	C37	C36	U35	G34	U33	U32	U31	I30	U29	G28	A27	G26	G25	C24	U23	C22	C21

Figure 3.1 (a)-(b) Inosines dsRNA. Inosines nucleobases are shown in blue. The red shaded central part correspond to the 24 nucleotides for which the sampling of the sugar puckering conformation was enhanced (see methods).Sample pairings: (c) a wobble pair between I9 and U32 and (d) a standrd WC pair between A6 and U35.

ditionally, it has been observed that IU base pairs introduce flexibility into RNA structures [82] [79], which is an aspect we also want to explore. A-to-I editing plays a dual role: it serves as an essential cellular mechanism and is increasingly recognized as being deregulated in specific diseases, including certain types of cancer and cardiovascular disease. This underscores the importance of comprehending the molecular implications of A-to-I editing, since such understanding could potentially lead to therapeutic applications.

To explore the impact of hyper A-to-I editing on dsRNA, we investigate the conformational dynamics of a 20-base pair double-stranded RNA (dsRNA) segment, featuring a central region hosting 4 inosines. The schematic representation of the system under study is provided in Figure 3.1. Notably, this system is characterized by an hyper-edited motif (**IIUI**), where inosines engage base pairing with uracils. I-U base pairs resemble wobble G-U pairs (see Fig. 3.1) and are hence predicted to introduce considerable flexibility in dsRNA, when compared with canonical A-U pairs. In order to quantify the extent to which inosine editing introduces flexibility to the system, we use in this study a combination of solution NMR and SAXS experiment, corroborated with MD simulations, to generate an ensemble of structures that are compatible with the experimental data.

The study that we are going to present is the result of a collaboration between the experimental laboratory of Michael Sattler and our group. While the experimental part was undertaken by Sattler's student, Christoph Mueller-Hermes, the computational work was overseen by myself. Consequently, this thesis exclusively showcases the computational facet of our study. Here, we harness experimental data to both restrain MD simulations via the maximum entropy principle (see section 2.3.2) and validate the generated structural ensembles. In the following sections, we first showcase the methods used to combine molecular simulation and solution experiments data in order to generate structural ensembles. After that, we will show the results of our investigation by comparing features of different ensembles. One of these ensembles characterize a dsRNA similar to the one shown in figure 3.1, but with adenosine in the place of inosine. The other ensembles instead, aim to characterize the inosine dsRNA, and are generated by using MD with enhanced sampling, and in some cases by further restraining the simulations in order to match with experimental data. Our findings will show to what extent the A-to-I hyper-editing can induce flexibility in the dsRNA, and at the same time will underscore the limits of MD in predicting accurate ensembles, and how these limitations can be solved by combining simulations with a relatively small amount of experimental information.

3.1 Methods

Starting structures for MD simulations were built using the proto–Nucleic Acid Builder [83]. Since inosines are not implemeted in this tool, the insosines ds-RNA was initially generated with guanosines instead of inosines. Then atom types in the generated PDB were corrected to convert the guanosines into inosines. Simulations were performed with GROMACS 2020 [84], using TIP3P water [85], the AMBER force-field for nucleic acids (AMBER99 + PARMBSC0 + χOL3) [22] [86] [87] plus modrna08 for inosines [28], and ion parameters from Joung and Cheatham [88]. The systems were first energy minimized and subjected to a multi-step equilibration procedure: 100 ps of thermalization to 300 K in the NVT ensemble was conducted through the stochastic dynamics integrator (i.e., Langevin dynamics) [89], and other 100 ps were run in the NPT ensemble simulations using the Parrinello–Rahman barostat [90]. For the productions runs, the stochastic velocity rescaling thermostat [91] was used to keep the system at a temperature of 300 K in combination with the cell-rescale [92] barostat to keep the pressure at 1 bar. Long-range electrostatic interactions were handled by particle-mesh Ewald [93]. The inosines (adenosines) double-strand system had 53868 (53944) atoms, 1274 (1278) of which constitute the solute; the rest were 72 sodium ions, 34 chloride ions and 52488 (52560) water molecules, resulting in a neutralized system with a salt concentration of 0.1 mol/l. An integration step of 0.002 ps was used, and trajectory frames were saved every 5000 steps with full precision. Additionally, coordinates were also saved every 500 steps with a compressed format and without water atoms. These latter coordinates were used for all the analysis shown in this work. MD simulations were performed using a replica exchange with collective-variable tempering (RECT, introduced in 2.2.2) [57]. 8 replica were used, in which well-tempred Metadynamics is performed with a bias factor that is scaled along the replica ladder, in such a way that the first replica is unbiased. Exchanges within replicas are proposed every 100 steps. We collected 3 set of simulations:

- a: Adenosines ds-RNA: 350 ns per replica
- b: Inosines ds-RNA: 366 ns per replica
- c: Insoines ds-RNA: 350 ns per replica + restraints on ^{3}J scalar couplings

For a total of (350+366+350 ns)x8 = 8.528 μs. The same enhanced sampling scheme was used in the 3 simulations. In addition, in case c additive restraints were used, acting on the 9 ^{3}J scalar couplings observables (see next sections).

3.1.1 Enhanced Sampling

Experimental observations conducted by our collaborators have revealed a relatively high occurrence of C2'-endo sugar puckering conformations for nucleotides located in the central region of inosine dsRNA, as opposed to the typical expectation of C3'-endo conformations in standard RNA helices. Consequently, our interest lies in conducting molecular simulations that can exhaustively sample all potential sugar puckering configurations. In this way, we are able

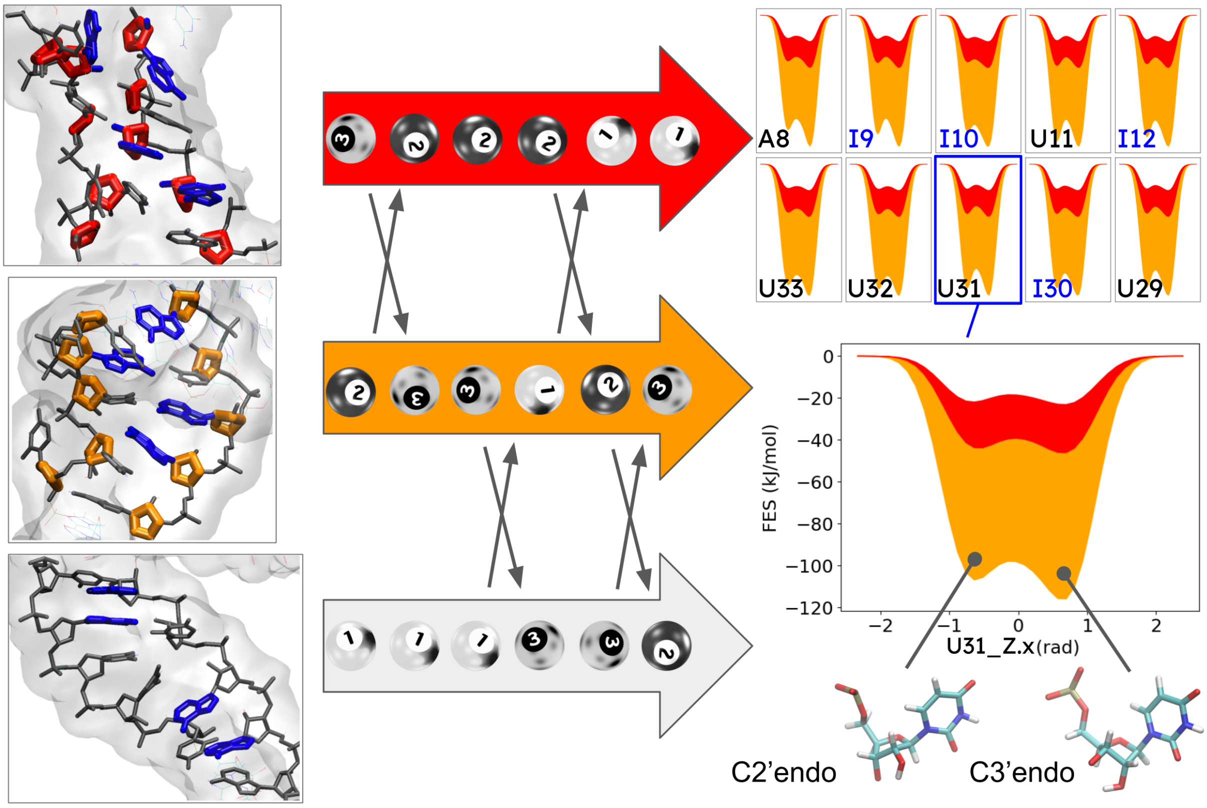


Figure 3.2 Scheme of the Replica Exchange with Collective-Variable Tempering (RECT) method used in this work. Different replicas correspond to scaled strength of the well-tempered Metadynamics on the sugar puckerings. The lower replica (grey) correspond to the unbiased Hamiltonian, but still its sampling is enhanced through the exchanges with the upper replicas (orange and red). For simplicity of the representation, the scheme is shown with only 3 replica instead of 8 (the lower, an intermediate, and the upper replica) and 10 enhanced sugars are shown, where the total is 24. Inosine nucleobases are colored in blue.

to obtain ensembles which contain C2'-endo conformations, so we can subsequently reweight these populations to match experimental ^{3}J scalar couplings with averages back-calculated from the simulation trajectories. To ensure this exhaustive sampling, we performed MD simulations using the replica exchange with collective-variable Tempering (RECT) scheme [57] introduced in 2.2.2. This scheme was applied to the sugar pseudorotation Zx variables of the 24 central nucleotides. The Zx variables, defined in the work of Huang *et al* [94], are expressed as:

$$Zx = \frac{\nu_1 + \nu_3}{2\cos\left(\frac{4\pi}{5}\right)} \tag{3.1}$$

where ν_1 and ν_3 are torsional angles shown in Figure 3.3c. The Zx variables are capable of distinguishing between C3'-endo conformations (positive values) and C2'-endo conformations (negative values) of the sugars. In our scheme, eight replicas were gradually biased to enhance sampling of the 24 selected degrees of freedom. However, we chose not to apply the RECT scheme to the 16 peripheral nucleotides of the RNA double-strand to avoid disruption of the helix.

A schematic representation of the RECT scheme used here is shown in Figure 3.2. The metadynamics simulations were performed using the PLUMED package [55] adding a Gaussian every 500 time steps. The 8 replicas correspond to well-tempered metadynamics with Bias Factor values γ: 1; 1.5; 2.1; 2.7; 3.2; 3.9; 4.5; and 5.0. Other parameters for the well-tempered MetaD are: σ=0.35 *rad*, τ = 5 *ps*.

3.1.2 Maximum Entropy Corrections

We integrate our MD simulations with experiments using the standard maximum-entropy (ME) reweighting procedure, introduced in 2.3.2. This procedure aims to find the probability distribution that closely matches the prior distribution predicted by MD, while also being consistent with experimental averages. It involves determining the set of Lagrange multipliers λ_i that minimizes the functional form:

$$\Gamma = \ln(Z(\lambda)) + \sum_{i}^{m} \lambda_i F_i^{\text{exp}} + \frac{1}{2} K \lambda^2 \tag{3.2}$$

Here, $Z(\lambda)$ is the partition function given by:

$$Z(\lambda) = \sum_{j}^{N} w_0^j e^{-\sum_i^m \lambda_i F_i(x_j)} \tag{3.3}$$

F_i^{exp} represents the experimental observables, while $F_i(x_j)$ denotes the observables back-calculated from the MD trajectories. The regularization hyperparameter K can be adjusted to prevent overfitting to the experimental data or, more in general, to strike a balance between relying on the experiment ($K = 0$) and trusting the MD model ($K = \infty$). In this study, the term "not regularized ensemble" refers to cases where we employed the ME refinement with $K = 0$. Conversely, the term "regularized ensemble" is used for cases where K was properly chosen to achieve a desired discrepancy (reduced $\chi^2 \sim 1$) between the simulation-calculated averages and the experimental averages. The value of the reduced χ^2 is computed as $\chi^2 = \frac{1}{M} \sum_{i=1}^{M} (J_i^{\text{exp}} - J_i^{\text{MD}})^2$, where $M = 9$ represents the number of experimental averages enforced with ME. These averages correspond to 9 J-coupling signals obtained from NMR measurements. The signals pertain to the sugar conformations ($^3J_{H1'H2'}$) of 9 nucleotides in the inosine dsRNA (I9-I10-U11-I12-U29-I30-U31-U32-U23).

Another essential component of the ME procedure is the forward model, which is necessary to back-calculate the experimental averages from the simulation trajectories. For J-coupling signals, Karplus equations are commonly used. These empirical equations establish a relationship between the NMR signal and dihedral angles. Specifically, the $^3J_{H1'H2'}$ signal is related to the torsional angle θ defined by the sugar atoms H1'-C1'-C2'-H2' through the following empirical equation:

$$^3J_{H1'H2'} = A\cos^2(\theta) + B\cos(\theta) + C\cos(\theta)\sin(\theta) + D \tag{3.4}$$

In this work, we considered two possible sets of Karplus parameters from the literature [95] [96] [97]. The first choice was the Condon parameters ($A = 9.67$, $B = -2.03$, $C = D = 0$) [95], which are the default parameters in the Barnaba package [98]. After further investigation, we opted for a more conservative choice, namely the Davies parameters ($A = 10.2$, $B = -0.8$, $C = D = 0$) [97]. The Davies parameters represent an intermediate case between the Condon Karplus equation, where the C2'-endo state of sugar puckering corresponds to a $^3J_{H1'H2'}$ signal of approximately 12 Hz, and the Marino Karplus equation [96], where C2'-endo corresponds to a $^3J_{H1'H2'}$ signal of approximately 7 Hz. The parameterization of Marino curves relative to

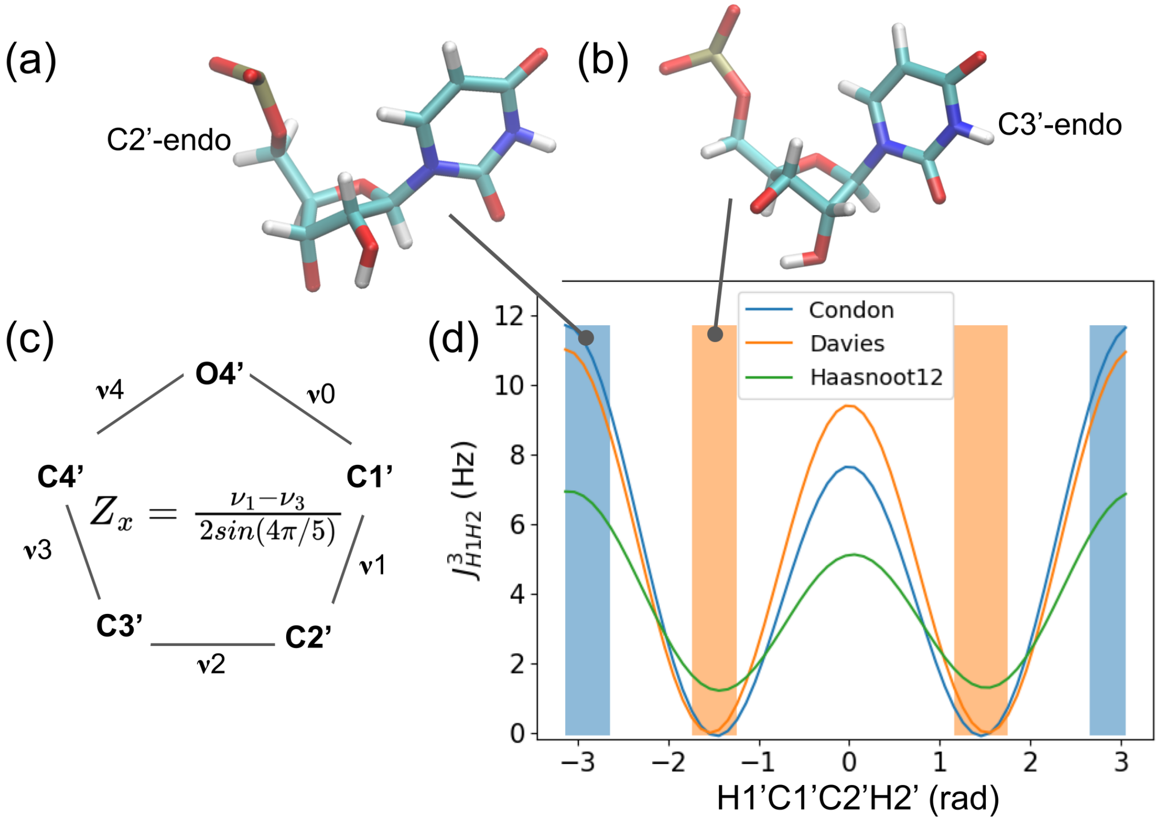


Figure 3.3 Examples of Uracils in C2'-endo(a) and C2'-endo (b) conformations. Scheme of the sugar ring (c). Karplus curves for Condon (blue), Davies (yellow) and Haasnoot12 (green) parameters (d). The HCCH region correspnding to C2'-endo si blue shaded, whereas the region corresponding to C3'-endo is yellow shaded.

equation 3.4 is referred as Haasnoot12 [95]. The Karplus curves for the Condon, Davies and Haasnoot12 parameterizations are shown in Figure 3.3d.

In this work, we generate ensembles in which only the 9 $^3J_{H1'H2'}$ signals are enforced, plus an ensemble in which the $^3J_{H1'H2'}$ are enforced together with the radius of gyration squared (Rg^2) as extrapolated from small angle scattering (SAXS) spectra. The average Rg^2 is extrapolated from SAXS spectra by out collaborator using Guinier fit procedure [99], whereas we computed it from the MD ensembles directly from atom coordinates using PLUMED. It is known that the two measures are inconsistent since solvent contributions influence the experimental estimates, which will be systematically higher with respect to the computational counterparts [100]. To address this issue, we decide to re-calibrate the Rg^2 estimates by aligning the adenosine dsRNA values, whose MD simulations can be considered reliable being the system a stable dsRNA with a A-form helix. The SAXS data available indicated an averaged radius of gyration of 1.83 nm for the adenosine dsRNA and 1.97 nm for inosine dsRNA. Since from our trajectories we could compute an $\sqrt{\langle Rg^2 \rangle}$ =1.81 nm for the adenosine helix, we decide to enforce a $\sqrt{\langle Rg^2 \rangle}$ =1.95 nm (1.95=1.81+1.97-1.83) for the inosine dsRNA through the ME.

3.1.3 ^{3}J coupling restraint

In the second set of MD simulations for the Inosines ds-RNA, additional Bias potentials were applied to increase the sampling of conformers compatible with the NMR ^{3}J scalar couplings. For this purpose, we used Lagrangian multipliers λ_i derived using the regularized ME on the previous set of simulations. The restraints were scaled along with the replica index, by dividing for the γ values. The restraint bias potential acting on the replica γ is:

$$R_\gamma(x) = \sum_{i=1}^{9} \frac{\lambda_i^3 J^i_{H1'H2'}(x)}{\gamma} \tag{3.5}$$

3.1.4 Reweighting

A crucial step in the generation of our ensemble is the evaluation of weights. We consider a set of N trajectories obtained using different values of the bias factor (γ) in the well-tempered metadynamics, so that n_γ the number of snapshots in the γ-th simulation.

The γ-th trajectory will thus contain samples from the distribution $P_\gamma(x) \propto e^{-\beta[E(x)+B_\gamma(x)]}$, where $E(x)$ is the energy associated conformation x, β is the inverse of the thermal energy, and $B_\gamma(x)$ is the bias potential which depend on the replica γ, and is given by the bias potential constructed by the metadynamics plus (possibly) the restraint bias potential defined in eq. 3.5 .

We are interested in obtaining weights $w(x)$ that can be used to compute averages corresponding the unbiased systems , which correspond to γ=1. In other words, for any observable $O(x)$, its average at γ=1 is obtained as

$$\langle O \rangle = \sum_x w(x) O(x) \tag{3.6}$$

where the sum runs over the concatenation of the N trajectories. By using the WHAM method in its binless formulation [61, 62, 63], the unnormalized weights can be obtained as:

$$w(x) \propto \frac{1}{\sum_\gamma n_\gamma e^{-\beta(B_\gamma(x)-B_1(x))} Z_\gamma^{-1}} \tag{3.7}$$

and subsequently normalized scaling them by a factor ensuring that $\sum_x w(x) = 1$. The partition function associated with each value of γ, Z_γ, can be obtained as

$$Z_\gamma = \sum_x w(x) e^{-\beta(B_\gamma(x)-B_1(x))} \tag{3.8}$$

The calculation was performed using the wham tool available in the bussilab python package, which can be obtained at `https://github.com/bussilab/py-bussilab`, version 0.0.36, and that can be used as a reference for the exact numerical procedure used here.

Notably, the weights only depend on the conformation (x) and not on the specific value of γ at which the conformation was generated. This implies that trajectories can be concatenated in any order resulting in identical weights. This allows to concatenate “demuxed” (continuous) trajectories, which are virtually independent of each other, being coupled only through the exchange step. By performing a blocked bootstrap with block size identical to the length of each trajectory, one ensures that correlations are minimized. In this case, however, one should explicitly take into account that, for a given bootstrap sample, the number of snapshots generated at each value of γ will differ.

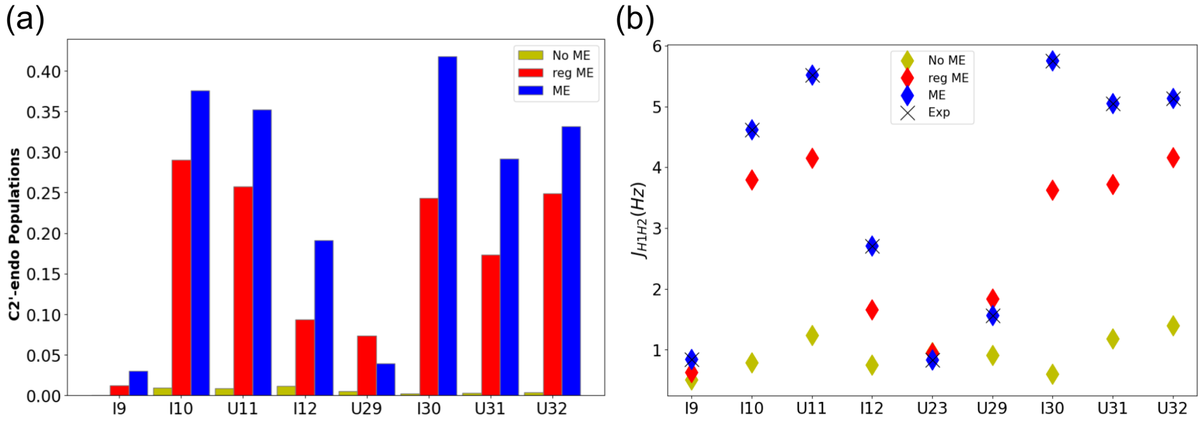


Figure 3.4 Result from preliminary simulations on the inosine dsRNA. (a) Populations of sugar in C2'-endo with respect to different reweighted ensemble. The pure MD (yellow bars) predicts very low populations of the C2'-endo state, below 2% for each of the central nucleotides in the IIUI motif. These populations significatively increase when enforcing the experimental $^3J_{H1'H2'}$ with regularized ME (red bars) and even more with no regularized ME (blue bars). (b) $^3J_{H1'H2'}$ signals measured by NMR experiments (black crosses) and back-calculated from MD trajectories (colored diamonds)

3.2 Preliminary simulations on the Inosines ds-RNA

The first set of simulations for the inosines ds-RNA were performed using the RECT scheme described above, but without using the restraint on the $^3J_{H1'H2'}$ signals. The populations of the C2'-endo conformations of the sugar puckerings as predicted by the pure MD are shown in Fig. 3.4a and result to be significatively lower with respect to the populations obtained when enforcing the NMR $^3J_{H1'H2'}$ data through the ME. Indeed, the $^3J_{H1'H2'}$ back-calculated (with Condon parameters in this case) from the reweighted trajectories are significatively smaller than those measured by NMR (see Fig. 3.4b). When enforcing the experimental data, the populations of the C2'-endo increase by up to a couple of orders of magnitude. The regularized ME ensemble in this case was obtained using $K = 6\ Hz^{-2}$ (see eq. 3.2).

Figure 3.5 shows the probability distributions of the $^3J_{H1'H2'}$ signals for the I30 nucleotide back-calculated from the trajectories considering three possible reweighted ensembles. The first corresponds to the pure MD ensemble, where a large peak is observed for low values of $^3J_{H1'H2'}$, corresponding to C3'-endo conformations of the nucleotide. Also, a small peak (log scale is used to better appreciate it) can be observed for high values of the signal. This small peak more likely correspond to the few conformations of the I30 nucleotide in the C2'-endo state that are predicted by the standard MD. When enforcing the experimental average for the $^3J_{H1'H2'}$ signal with ME, the effect is a reweighting which allows a rebalance of the two peaks populations.

3.3 Results

The results shown in this section are obtained from simulations (a) (see section 3.1) as far as the adenosine dsRNA is concerned, whereas the ensembles of the inosine dsRNA are obtained from simulations (c). We decide to discard (b) since (c) ensures a better sampling of the sugar puckering, thanks to the restraints used. In particular, in (b) the populations of C2'-endo resulted

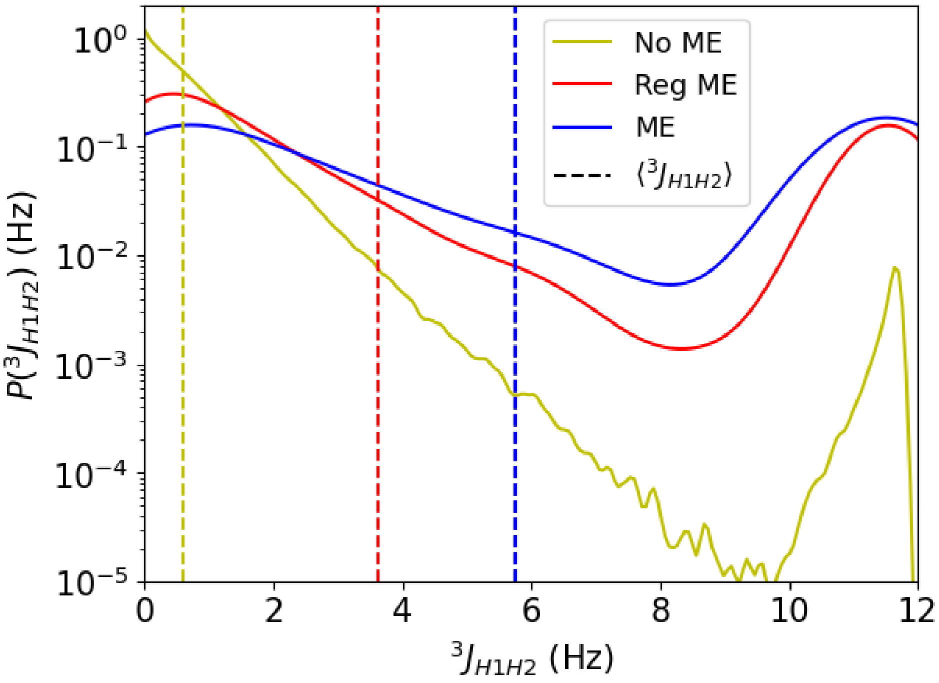


Figure 3.5 Result from preliminary simulations on the inosine dsRNA. Distribution of the $^3J_{H1'H2'}$ signals for the I30 nucleotide, back-calculated from the trajectories for the three different ensembles. The dotted vertical lines indicate the experimental averages signals back-calculated from the trajectories, which in the blue case (not regularized Max. Ent.) correspond to the experimental average.

to be very far from the experiments. In (c) we guide the MD in the productions phase, so it is easier to generate ensembles which are closer to those which are compatible with experimental observables, in this way enabling the successive reweghting to work more efficiently.

If not differently specified, the ME on ${}^3J_{H1'H2'}$ signals were performed considering Davies parameters and without regularization.

3.3.1 Investigating general conformational properties of the dsRNAs

Figure 3.6a displays the probabilities of formation of the 20 canonical pairings in the dsRNAs for four different ensembles:

- A - Adenosine dsRNA as predicted by MD (grey).
- I - Inosine dsRNA as predicted by MD (red)
- I+NMR - Inosine dsRNA with enforced ^{3}J scalar couplings (blue)
- I+NMR+SAXS - Inosine dsRNA with enforced ^{3}J scalar couplings and radius of gyration squared (green).

Pairing populations are counted based on Barnaba annotation [98], considering Watson-Crick (WC and WW) and Wobble (GU) pairings both as canonical. The latter represents the type of pairing expected for the I-U base pair (see fig 3.1). In the adenosine dsRNA, all base pairs are almost 100% populated, except for the periferical A-U pair. In the inosine dsRNA, the populations significantly decrease only in the central part of the helix where the inosines are present, indicating that these modifications may induce a significant increase in the flexibility of the dsRNA in its central part. The pairing populations further decrease when ^{3}J scalar couplings are enforced and even more so when the Rg^2 is enforced. Interestingly, the I12-U29 base pair is the least affected by the ensemble refinement, becoming the most populated base pair among those in the IIUI motif for the ME-refined ensembles. This experimental result aligns with the predictions of NMR exchange-rate experiments, not shown in this thesis, which suggest that the I12-U29 base pair is the most stable in the IIUI motif. This result provides evidence for the validity and transferability of the ensemble refinement performed using only nine ${}^3J_{H1'H2'}$.

Figure 3.6b displays the populations of C2'-endo conformations of sugar puckering in the four distinct ensembles. As previously observed in Section 3.2, the populations of C2'-endo conformations increase significantly, by orders of magnitude, when incorporating NMR data. This increase is a consequence of the inaccuracies of molecular dynamics (MD) simulations with the current force-field, which result in a substantial underestimation of these populations for this system. However, by combining enhanced sampling and ensemble refinement techniques, we can accurately reproduce the populations that align with the predictions from the ^{3}J scalar couplings NMR signals. These results underscore the need for a revision of the inosine force-field, and possibly also for the other nucleotides force-fields, since also C2'-endo populations of the uracils in the central motif result highly underestimated in the I (red) ensemble.

The SAXS spectra, obtained by our collaborator, for both adenosine and inosine double-stranded RNAs (dsRNAs) indicate a significant increase in the averaged radius of gyration ($\sqrt{\langle Rg^2 \rangle}$) in the hyper-edited dsRNA compared to its adenosine counterpart. Figure 3.7a illustrates the distributions of the radius of gyration in the four ensembles we derived, while Figure 3.7b displays the corresponding relative averages, $\sqrt{\langle Rg^2 \rangle}$. For the inosine cases, we also provide the experimental reference calibrated with respect to the adenosine measurement, as discussed at the end of Section 3.1.2. The results depicted by these findings indicate that the

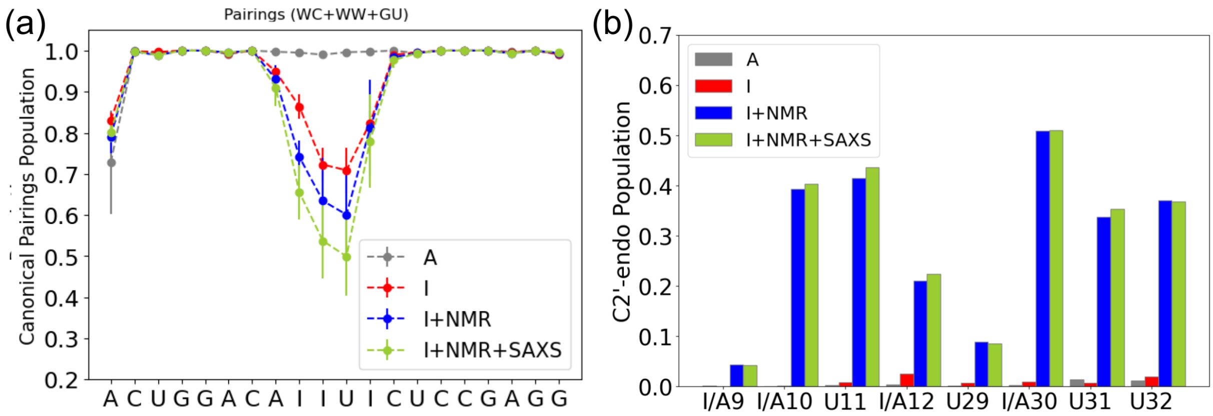


Figure 3.6 (a) Probabilities of canonical pairings in the dsRNAs for 4 different ensembles. Adenosines dsRNA as predicted by MD (Grey); Inosines dsRNA as predicted by MD (Red); Inosines dsRNA where ^{3}J scalar couplings are enforced (Blue); Inosines dsRNA where ^{3}J scalar couplings and radius of gyration squared are enforced (Green). Pairings are counted based on Barnaba annotation [98], considering as canonical Watson-Cricks (WC and WW) and Wobble (GU) pairings. For the same four ensembles, in panel (b) we refer populations of C2'-endo conformations of the sugar puckering for the central nucleotide of the dsRNAs.

I ensemble cannot reproduce the substantial increase in $\sqrt{\langle Rg^2 \rangle}$ compared to the A ensemble, as expected from the SAXS data. Interestingly, when we incorporate NMR data, $\sqrt{\langle Rg^2 \rangle}$ increases in a manner that significantly reduces the discrepancy with the SAXS experiment. Full agreement is only achieved in the I+NMR+SAXS ensemble, where the experimental $\langle Rg^2 \rangle$ was enforced through maximum entropy. The results obtained in the I+NMR ensemble for the radius of gyration serve as validation for the ensemble refinement carried out by enforcing NMR data. In fact, this ensemble also exhibits improved accuracy in reproducing an observable, such as the Rg, for which information was not explicitly included in the ensemble refinement.

In order to qualitatively show how inosine editing improves the flexibility of the dsRNA, a *bouquet* representation of the dsRNA ensembles is proposed in Figure 3.8. The figures were obtained by randomly extracting 100 structures from the reweighted trajectories and aligning them with respect to residues G20-G19-A18-G17-C21-C22-U23-C24 in the bottom part of the helix. From the figure, it is noticeable that the adenosine dsRNA (grey) cloud appears more regular and less chaotic compared to the inosine counterpart. However, differences between the different inosine dsRNA ensembles are not appreciable from this representation. To assess the rigidity of these ensembles, we calculated their root mean square deviation (RMSD) relative to the same alignments used for the bouquet representations. For each ensemble, we identified a centroid from a pool of 1000 structures extracted. The centroid is the structure that minimizes the average squared RMSD $\sqrt{\langle RMSD^2 \rangle}$ when compared to all the other 1000 structures. Subsequently, we computed the ensemble RMSD values displayed in the table of Figure 3.8. This was done by considering the entire reweighted trajectories and determining the $\sqrt{\langle RMSD^2 \rangle}$ with respect to the previously identified centroid.

To provide deeper insights on the general conformations of the dsRNAs, we propose a quantitative and detailed analysis of the helical parameters of the different ensembles. Figure 3.9 shows 2D density plots with kink and twist angles for 1000 structures extracted from each ensemble. These angles correspond respectively to the total bending of the helix and to degree of twisting of the two strands, and were computed using the Curves+ software [101]. The

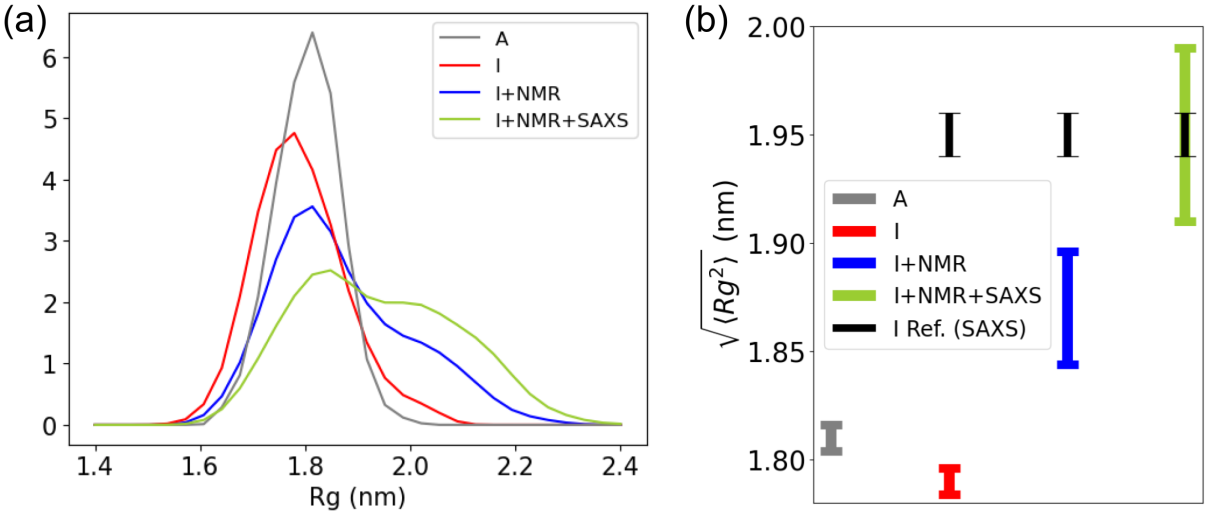


Figure 3.7 (a) distributions of the radius of gyration (Rg) of the dsRNAs for the 4 different ensembles. (b) Square root of averaged radius of gyration squared ($\sqrt{\langle Rg^2 \rangle}$) for the 4 different ensembles and experimental reference values extrapolated from SAXS data (black).

Ensemble RMSD			
A	I	I+NMR	I+NMR+SAXS
0.24 nm	2.01 nm	2.08 nm	2.08 nm

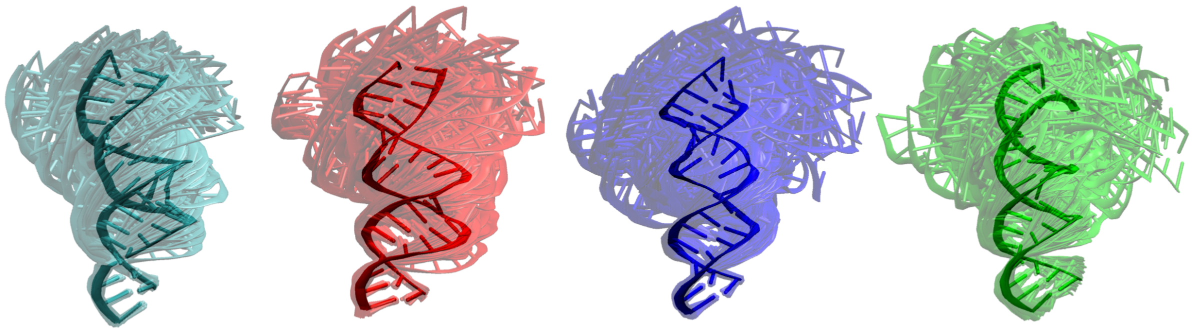


Figure 3.8 *Bouquet* representation of the structural ensembles. 100 structures randomly extracted are aligned with respect to the 8 residues at the bottom of the helix. The centroids of the ensembles are opaque colored. In the top table the ensemble RMSDs are given, computed on the whole reweighted trajectories and with respect to the same alignment of the bouquets, using as a reference the centroids.

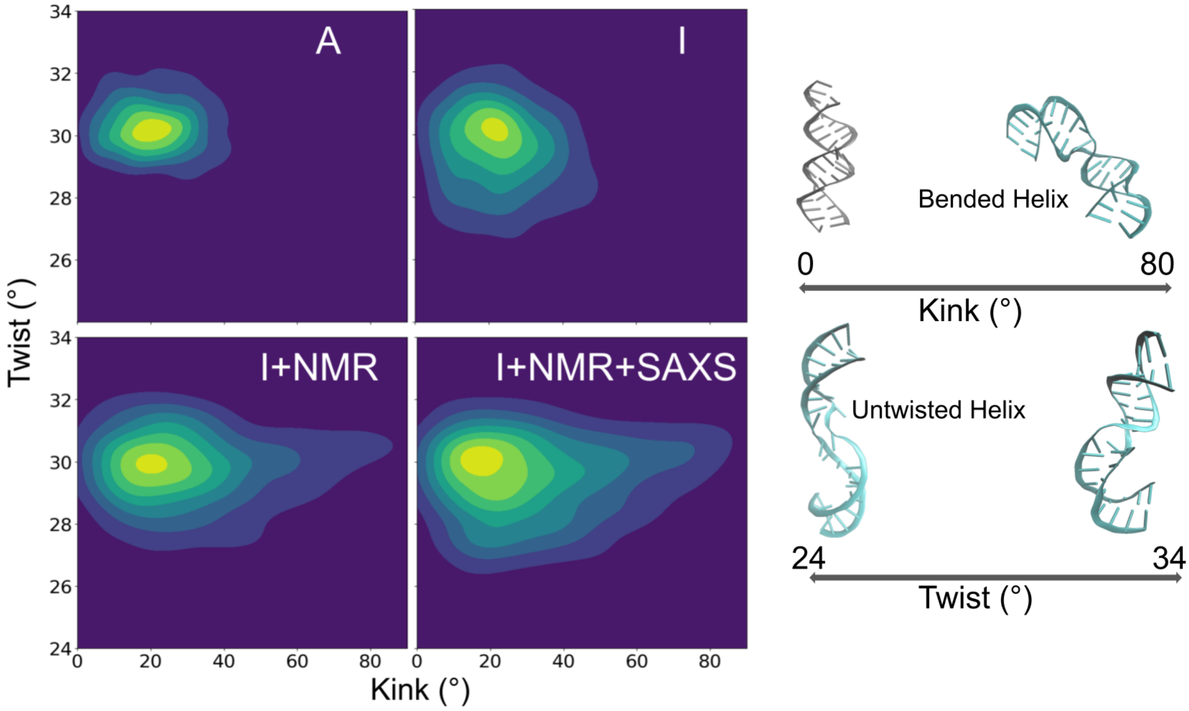


Figure 3.9 Kink and twist angles density distributions for the 4 different ensembles. The distributions are computed out of 1000 structures extracted for each ensembles. Kink and twist angles where computed using Curves+ package [101] and correspond respectively to the total bending of the helix and to degree of twisting of the two strands. On the right, 4 limiting cases of the Kink and twist angles among the extracted structures are shown.

region with an averaged twist of approximately 30° and a total bend (kink) of less than 40°, corresponds to the standard A-form helix of a dsRNA. The four structures shown on the right represent the limiting cases for both twist and kink angles.

A comparison between the density plots interestingly shows that the I ensemble closely resembles the adenosine case as far as the kink angle is concerned. However, when enforcing ^{3}J scalar coupling signals, the density at higher kink angles increases. This suggests that the higher propensity of the sugars to be in the C2'-endo conformation causes the dsRNA to bend more, allowing for increased flexibility in the central part of the helix corresponding to the IIUI motif. Regarding the twist angles, the densities of the three inosine ensembles are quite similar, resulting in an averaged twist angle reduced compared to the adenosine system. This is mainly due to the I-U wobble pairings occurring in the central motif, which cause a shift of the strands to allow these non canonical pairings. On the other hand, enforcing higher C2'-endo populations through the ^{3}J scalar couplings does not significantly affect this parameter.

3.3.2 Analysis of Conformers

In the previous section, we explored the overall features of dsRNAs by examining ensemble averages. To identify specific structures that may represent a significant portion of the ensembles, we attempted clustering using Principal Component Analysis (PCA) [102]. The PCA analysis was performed using the PyEMMA package [103], with torsional angles and *G*-vectors as input data derived from 4000 structure (1000 structures from each of the four ensembles). Torsional angles and *G*-vectors were computed using Barnaba package [98], focusing on the 8 central residues in the dsRNAs. We considered both sine and cosine values for 6 backbone torsional angles (α, β, γ, δ, ε, ζ), the glycosidic angle (χ), and 5 sugar angles ($\nu_1, \nu_2, \nu_3, \nu_4, \nu_5$). This resulted in a total of 192 components from torsional angles and 256 from *G*-vectors, for a total of 448 components. We remark that *G*-vectors and torsional angles have different units, and that we aribitrary choose to use a scaling factor of 1 to combine them. In Figure 3.10, we present 2D density plots based on the first and second principal components obtained through PCA for each of the four ensembles. These plots reveal that most of the variance arises from the I+NMR and I+NMR+SAXS ensembles, while the A and I ensembles exhibit greater homogeneity. Additionally, we depict a scatter plot showing all the extracted structures, with colors indicating the number of nucleotides in the C2'-endo conformation. From these plots, we can distinguish a primary basin present in all four cases, corresponding to a standard A-form helix conformation. In this basin, the majority of structures have no C2'-endo nucleotide populations. In contrast, the I+NMR or I+NMR+SAXS cases display distinguishable basins with higher C2'-endo populations. Structures with the highest number of C2'-endo populations are scattered throughout the density plots, making it challenging to associate them with a specific representative conformer.

In Figure 3.11, we present five conformers corresponding to the five primary basins identified in the PCA density plot for the I+NMR+SAXS ensemble. We represent these conformers using dynamic secondary structure representations generated by Barnaba [98], obtained by manually selecting structures from each of the basins. Conformers B1 and B2 exhibit canonical pairings, while conformers B3, B4, and B5 demonstrate dynamic pairing along with the formation of non-canonical pairings. Notably, conformer B5 exhibits a non-canonical pairing between residues I9 and I30. This specific contact is corroborated by NOE data, as we will discuss in the following sections.

3.3.3 Quantifying Cooperativity of the Sugar Puckerings

In the ensemble of inosine dsRNA, where the population of C2'-endo sugar puckering conformations is high, it is intriguing to understand whether there is any cooperativity between sugar conformations of different nucleotides. In particular, we seek to understand whether the sugars switch to C2'-endo collectively, in specific combinations, or independently of each other. To investigate this, we conducted cluster analysis by extracting 1000 structures from the reweighted ensembles and categorizing all possible combinations of sugars in the C2'-endo conformation. As an example, Figure 3.12 displays the 15 most populated clusters for the I+NMR ensemble. Populations for these 15 cases are indicated also for the I ensemble (red) and the I+NMR+SAXS ensemble (green). The clusters are shown only for the 12 central nucleotides using the dynamic secondary structure representations implemented in Barnaba [98]. Nucleotides in the C2'-endo conformation are highlighted in orange. Figure 3.13 aims to summarize the results of the clustering analysis on the I+NMR ensemble using a rectangular pie chart representation. Each column corresponds to a cluster, with nucleotides in the C2'-endo conformation displayed in orange. The width of each column represents the population of the clusters, with less populated clusters merged into the far-right column. From this representation, we can observe that all possible combinations of nucleotides in the C2'-endo conformation are feasible, but their populations may vary, ranging from significant to insignificant. This scenario is distinct from one featuring a few large clusters with numerous nucleotides in C2'-endo conformations, as one might have expected in the case of strong cooperativity between sugar puckering conformations.

In order to assess quantitatively if any cooperativity between C2'-endo conformations of the sugars is present, we computed energetic cooperativities for the 28 pairs of nucleotides related

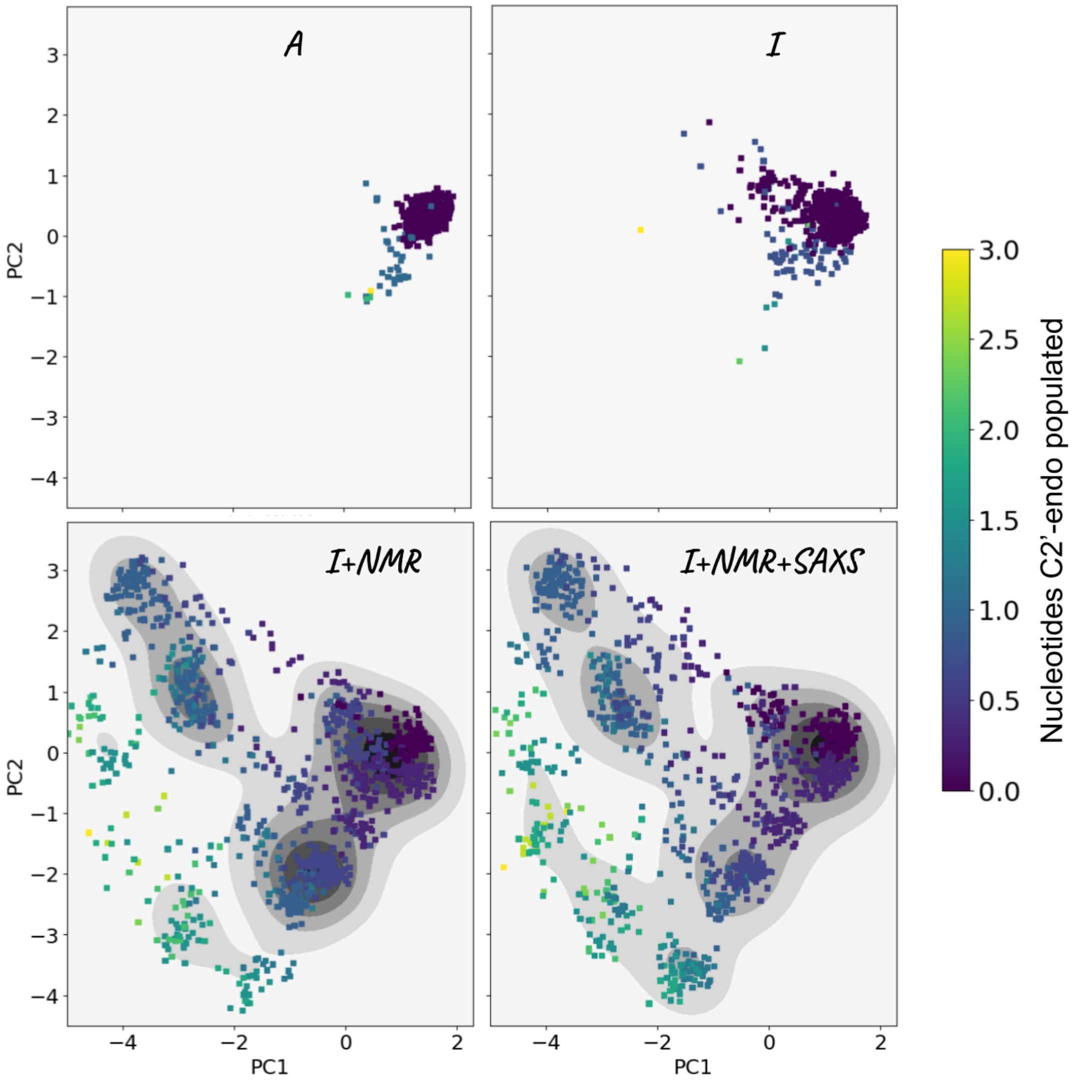


Figure 3.10 Results of Principal Component Analysis (PCA) performed using as input backbone torsional angles and G-vectors [98] for the 8 central nucleotides, from 4000 structures extracted from the ensembles (1000 each). For each ensemble, we show 2D density plots with respect to the first and second component. Each point is colored based of number of nucleoitides in C2'-endo conformation

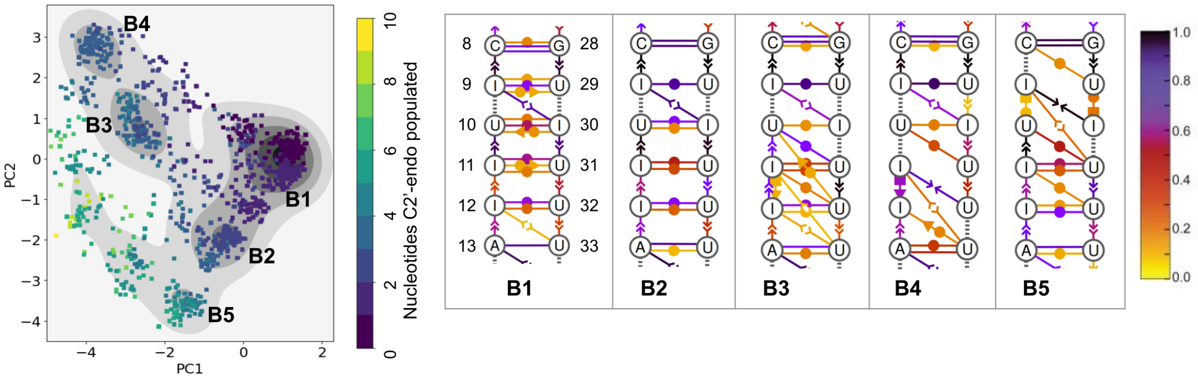


Figure 3.11 Individuation of conformers for the I+NMR+SAXS ensemble. 5 conformers were recognized collecting by hand strctures from the 5 evident basins individuated by the PCA (left panel) and are represented through dynamic secondary structures representations [98] (right panel).

the IIUI motif (see figure 3.14). Cooperativities are computed as follow:

$$\Delta\Delta G = k_B T \log(\frac{P_{11}P_{00}}{P_{10}P_{01}}) \tag{3.9}$$

P_{11} counts how many times both sugars are in C2'-endo; P_{00} counts how many times both sugars are not in C2'-endo; and P_{01} (P_{10}) counts when only the first (second) is in C2'-endo. Negative (positive) values of $\Delta\Delta G$ indicate cooperativity (anti-cooperativity). However, since these values may be dominated by the statistical error, we computed them over 1000 iterations of boostrapping, counting how many times the $\Delta\Delta G$s are negative for each pair. Since we deal with 28 hypotheses simultaneously, we rely on the Benjamini-Hochberg procedure [104] to keep the false discovery rate of our estimates at a significance level of $p = 0.05$, similarly to [105]. Figure 3.14b shows results concerning the I+NMR ensemble. Although negatives $\Delta\Delta G$s are observed for the majority of neighbouring nucleotides, the Benjamini-Hochberg analysis indicates only 5 significant cooperativities, corresponding to the 5 dots below the black dotted-line in the bottom-right plot in the figure. Surprisingly, cooperativity is observed for the I9-I30 pair, which doesn't correspond to neighbouring nucleotides. Interestingly, the I9-I30 pair is found to form non canonical pairing in a representative conformer of the I+NMR+SAXS ensemble, as shown in previous section. The same analysis is shown in Figure 3.14a for the pure MD ensemble (I). In this case, the Benjamini-Hochberg analysis indicates 13 significantly cooperative pairs. This result demonstrates that the cooperativity comes from the simulations and not from the experimental data enforced. Furthermore, the ensemble refinement has the effects of reducing the statistical significance of the cooperativities. This could simply be caused by the statistical error introduced by the reweighting, and additionally by multi-bodies effects that the simple model characterized by the equation 3.9 does not take into account.

3.3.4 Validation against NOEs

Nuclear Overhauser effect (NOE) NMR data were used in this work to validate the ensembles that were refined over ^{3}J scalar couplings and SAXS data. We used 197 NOEs signals corresponding to 197 protons pairs. In order to back-calculate NOEs from the simulations ensembles we used the standard relation:

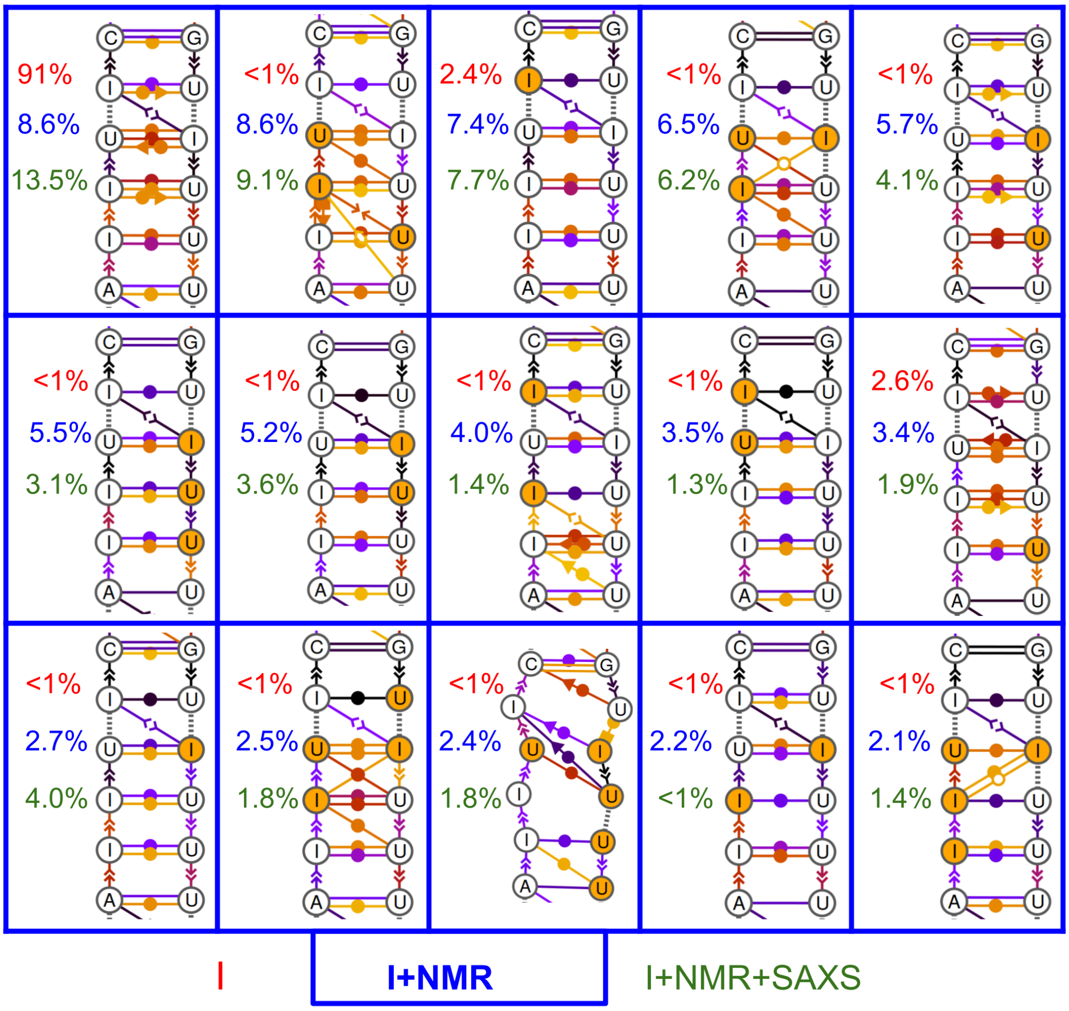


Figure 3.12 15 most populated clusters individuated among 1000 structures extracted from the I+NMR ensemble (populations reported in blue). This clustering simply differentiates between possible combinations of sugars in C2'-endo coformations, which are colored in orange. Populations for these 15 cases are shown also for the I ensemble (red) and the I+NMR+SAXS ensemble (green). The dsRNAs are shown only for the 12 central nucleotides using the dynamic secondary structure representation derived from the I+NMR ensemble.

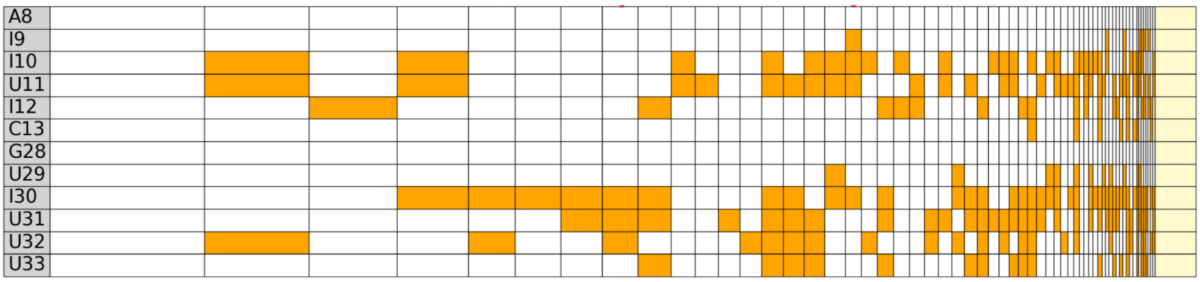


Figure 3.13 Rectangular pie chart representing the clusterization of the I+NMR ensemble based on C2'-endo conformations of the sugars. Each column correspond to a cluster for which nucleotides in C2'-endo are orange colored. The width of the column reflect the population of the clusters. Less populated clusters are merged in the far right column.

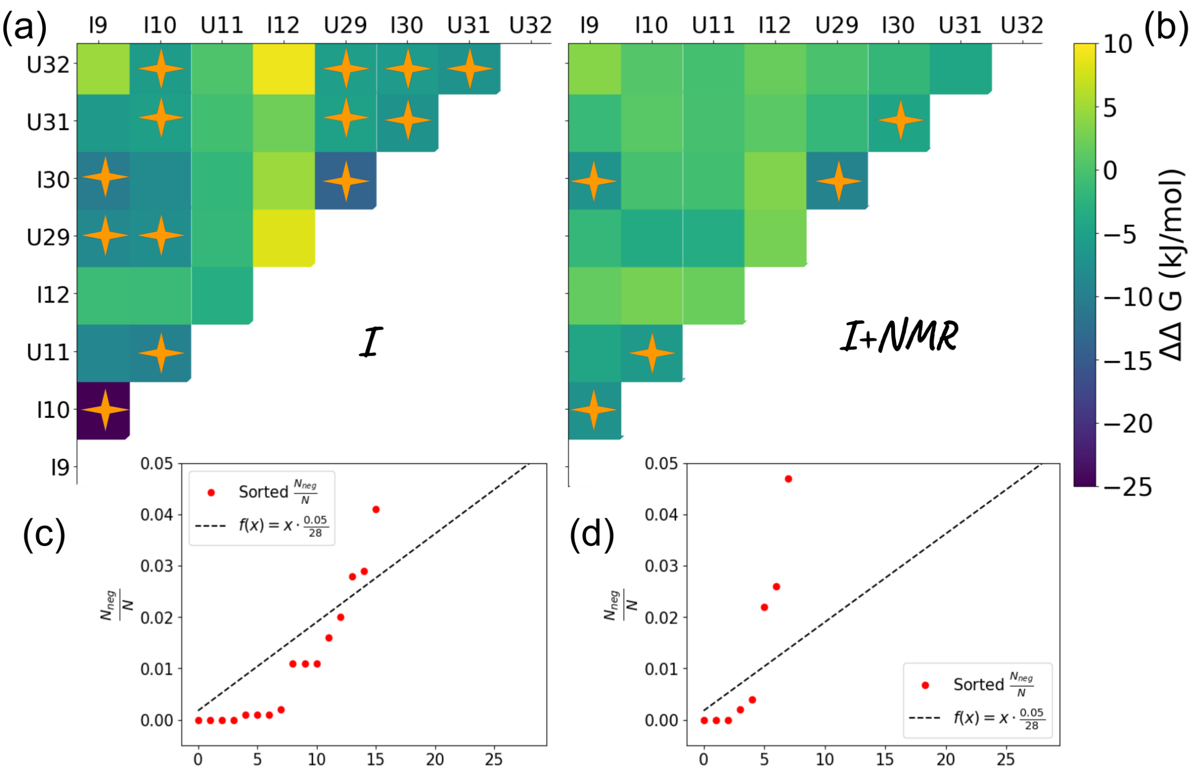


Figure 3.14 C2'-endo Cooperativity matrix $\Delta\Delta G$ for the I ensemble (a) and for the I+NMR ensemble (b). Pairs of nucleotides for which the cooperativity has a statistical significance level greater than 0.05 are marked with an orange star. 13 Significant cooperativities are individuated through the Benjamini-Hochberg procedure (points below the dotted line in the bottom-right plot) for the I ensemble (c), whereas only 5 for the I+NMR ensemble (d). Significant anti-cooperativity is not observed.

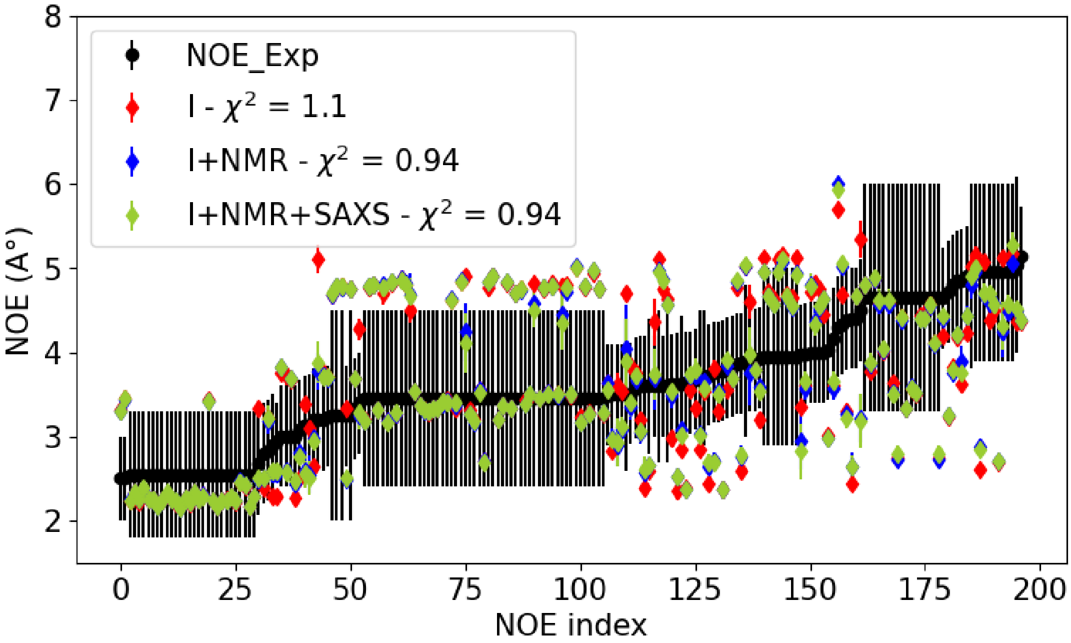


Figure 3.15 NOEs data sorted with respect to experimental values (black). Diamonds correspond to signals back-calculated from the trajectory with respect to the I ensembles (red), the I+NMR ensemble (blue), I+NMR+SAXS ensemble (green). Averaged χ^2 computed for the three ensembles are indicated in the legend.

$$NOE_{sim} = [\sum_{i}^{N} w_i r_i^{-1/6}]^{-6} \tag{3.10}$$

where index i runs over all the frames of the trajecotry and w_i are the weights which depend on the ensemble considered. The statistical errors were computed using a bootstrapping procedure, that is by resampling the 8 continuos trajectories generated in the RECT simulations 200 times by replacement [64]. At each iteration of the bootstrapping, the weights related to the bias used in the simulations are recomputed using WHAM, whereas the lagrangian multipliers to the ME restraint are kept as those individuated performing ME on the complete set of demuxed trajectories, without the need to reperform ME at every iteration of the bootstrapping. The 197 NOEs values are shown in Figure 3.15 sorted with respect to the experimental value. Simulation NOEs computed from the I ensemble (red) falls within the experimental bar 120/197 times and led to a $\chi^2_{NOE} = 1.1$, which is computed as follows:

$$\chi^2_{NOE} = \frac{1}{197} \sum_{m=1}^{197} \frac{(NOE_m^{exp} - NOE_m^{sim})^2}{\sigma^2_{exp,m} + \sigma^2_{sim,m}} \tag{3.11}$$

NOEs computed respectively from the I+NMR and I+NMR+SAXS lay in the experimental bar 137/197 and 136/197 times, both giving $\chi^2_{NOE} = 0.95$. These reduced χ^2_{NOE}s mean that the ensemble refinement performed by enforcing the $^3J_{H1'H2'}$ through ME provides ensembles which are better in agreement also with independent observables as the NOEs signals, proving the transferability of the predicted ensemble. We remark that, although the decrease of χ^2_{NOE} is moderate, it originates from the inclusion of completely independent experimental data ($^3J_{H1'H2'}$ and SAXS).

Table 3.1 shows the list of protons pairs which correspond to NOE signals that are consistent with experimental NOEs for the I ensemble but not for the I+NMR+SAXS ensemble, and *viceversa*. Most of these pairs involve intra-nucleotide or intra-strand distances between a sugar proton and a nucleobase proton. This observation is not surprising, as the differences in the ensembles are characterized by discrepancies in sugar puckering conformations, induced by inducing higher populations of the C2'-endo conformations in the I+NMR+SAXS ensemble, through the imposition of $^3J_{H1'H2'}$ data. The fact that this latter ensemble also better matches NOE experiments suggests that the higher C2'-endo populations in the central part of the inosine dsRNA are supported by the NOE data. In Table 3.1, the only inter-strand distance that appears in the list involves atoms I30-H1' and I12-H2. Interestingly, inosines I12 and I30 are those for which non-canonical pairings are observed in conformer B5, as discussed in Section 3.3.2. Furthermore, statistical analysis in the previous section has shown that there is significant cooperativity in sugar puckering conformations for these residues. These findings suggest that the interaction between I30 and I12, predicted by our combination of MD, NMR $^3J_{H1'H2'}$, and SAXS data, is validated by NOE data and is correlated with both residues simultaneously transitioning to the C2'-endo sugar puckering conformation.

3.3.5 Robustness of the results with respect to forward model and regularization strength

In this section, we aim to compare the results obtained by performing ME with different forward models, namely Condon and Davies, and examine the effect of ensemble refinement with or without regularization. Figures 3.16 depict the ratio of nucleotides in the C2'-endo conformation and the $^3J_{H1'H2'}$ signals for nucleotides in the IIUI motif, respectively. These plots show how in the ensembles without regularization, there is a higher tendency for nucleotides with higher experimental $^3J_{H1'H2'}$ signals to adopt the C2'-endo conformation. Additionally, it can be observed that the predicted populations of C2'-endo conformations are generally higher in the Davies ensembles compared to the Condon ensembles. This difference arises from the lower values of the Davies Karplus curve, as shown in Figure 3.3, in the θ region corresponding to the C2'-endo conformation: because of this lower values, higher C2'-endo populations are needed to enforce experimental $^3J_{H1'H2'}$s. Figure 3.17 shows the clusters already shown in Figure 3.14, but this time also indicating the population for the other ensembles that would have been obtained for the same combinations of nucleotides in C2'-endo. In general, the clusters that are obtained with respect to different reweightings are the same but with different populations, in spite of the fact that the structures are extracted from the same trajectory. As expected, the first cluster corresponding to the case of no nucleotides in C2'-endo, is much more populated for the regularized ensembles.

Furthermore, the 4 different ensembles are analyzed using NOE data, as shown in Figure 3.18a. In the legend, the χ^2_{NOE}s are shown computed as in 3.11. Interestingly, the Davies ensemble, which is the one allowing for highest f C2'-endo populations, is the ensemble that minimizes the discrepancy with the experiments. Moreover, in Figure 3.18b the square root of the averaged radius of gyration squared ($\sqrt{\langle Rg^2 \rangle}$) is plotted for each ensemble. Although the results exhibit comparable statistical errors computed using the bootstrapping procedure, it is notable that the Davies ensemble without regularization yields the highest value of $\sqrt{\langle Rg^2 \rangle}$, thereby enhancing the agreement with the $\sqrt{\langle Rg^2 \rangle}$ extrapolated from SAXS spectra. The regularization ensembles were derived with the scope of avoiding overfitting on experimental $^3J_{H1'H2'}$ signals. However, the validations performed on NOEs and $\sqrt{\langle Rg^2 \rangle}$ suggest to consider the Davies ensemble as the most valuable, without the need of the regularization.

Agreement in I and not in I+NMR+SAXS		
pair type	protons pair	
Intra-nt	I10-H8	I10-H2’2
Intra-nt	I10-H2’1	I10-H2
Intra-nt	U11-H6	U11-H2’1
Intra-nt	U33-H6	U33-H2’1
Agreement in I+NMR+SAXS and not in I		
pair type	protons pair	
Intra-nt	IU11-H2’1	U11- H6
Intra-nt	I12-H8	I12-H3’
Intra-nt	I30-H8	I30-H2’1
Intra-nt	I30-H8	I30-H3’
Intra-nt	I30-H1’	I30-H5’2
Intra-nt	U31-H1’	U31-H5’2
Intra-nt	U32-H1’	U31-H5’2
Intra-nt	U32-H6	U32-H4’
Intra-strand	I9-H2	A8-H2
Intra-strand	U11-H2’1	I12-H8
Intra-strand	U11-H1’	I12-H8
Intra-strand	C13-H1’	U14-H6
Intra-strand	I30-H8	U29-H2’1
Intra-strand	I30-H1’	U31-H6
Intra-strand	U31-H1’	U32-H6
Intra-strand	U33-H1’	G34-H8
Intra-strand	U35-H1’	C36-H6
Intra-strand	G34-H1’	U35-H6
Inter-strand	I30-H1’	I12-H2

Table 3.1 List of protons pairs which correspond to NOE signals that are in agreement with experimental NOEs for the I ensemble but not for the I+NMR+SAXS ensemble, and *viceversa*.

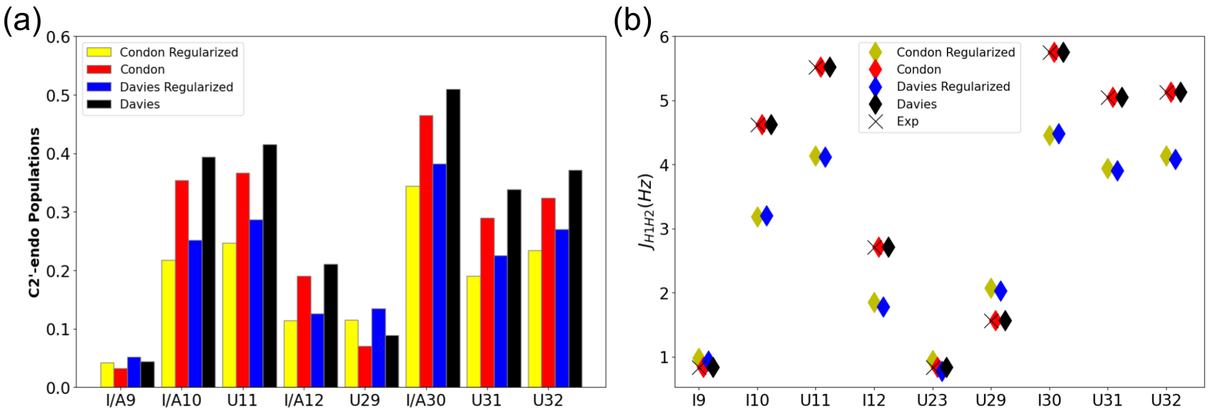


Figure 3.16 (a) Ratio of sugar in C2'-endo with respect to different reweighted ensembles, obtained using Davies Karplus equations and not regularizing the fit (Davies) or regularizing the fit (Davies Reg), or by using Condon Karplus equations for in the same ways (Condon and Condon Reg). (b) $^3J_{H1'H2'}$ signals measured by NMR experiments (black crosses) and back-calculated from the reweighted trajectories (colored rhombuses).

3.4 Discussion

This chapter focuses on the computational aspect of a collaborative study between Michael Sattler's experimental laboratory and our research group, in which I oversaw the computational work. Within this collaborative framework, we made use of experimental data for two key purposes: guiding MD simulations using the principles of maximum entropy (detailed in Section 2.3.2) and validating the resulting generated ensembles of structures. Our study centers on a dsRNA system comprising 20 base pairs, with four inosines situated in the central portion of the helix, each paired with uracils (refer to Figure 3.1). Since the available experimental results were indicating interesting and unexpected sugar puckering conformations in the central part of the helix, we conducted simulations using the replica exchange collective variable tempering (RECT) approach to enhance sampling. This method allowed for an exhaustive exploration of all possible configurations of the dsRNA with respect to sugar puckering conformations. Subsequently, we applied the maximum entropy (ME) principle to reweight the trajectories. This reweighting process aimed to generate ensembles of structures that align with nine NMR signals ($^3J_{H1'H2'}$) associated with the torsional angles of the nucleotide sugars. Additionally, we enforced the averaged radius of gyration as predicted by SAXS experiments.

We illustrated our findings by comparing 4 different derived ensembles, corresponding to the adenosine dsRNA (A), inosine dsRNA as predicted by MD (I), the inosine dsRNA with enforced ^{3}J scalar couplings (I+NMR), and finally the inosine dsRNA with enforced ^{3}J scalar couplings and radius of gyration squared as predicted by SAXS experiments (I+NMR+SAXS). Our findings show to which extent A-to-I hyper-editing can induce flexibility in dsRNA. Simultaneously, they underscore the limitations of MD in predicting accurate ensembles . Indeed, although all inosine ensembles are able to predict dynamic and non canonical pairings in the central part of the dsRNA, the I ensemble is not able to reproduce the sugar puckering populations predicted by the NMR experiments, which suggest high populations of C2'-endo conformations. However, by combining enhanced sampling and ensemble refinement techniques, we can accurately reproduce the populations that align with the predictions from the ^{3}J scalar couplings NMR signals. These results underscore the need for a revision of the inosine

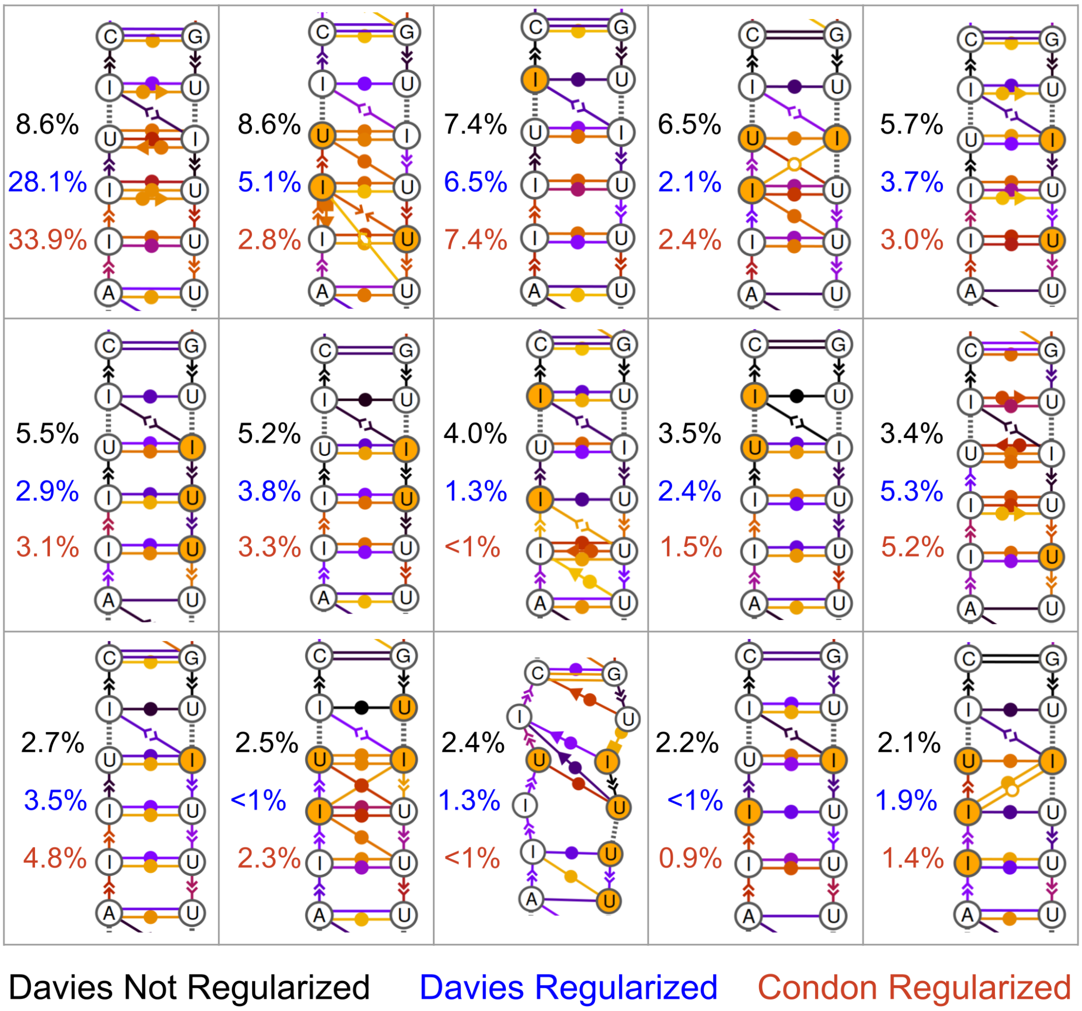


Figure 3.17 15 most populated clusters for the I+NMR ensemble, noted here as Davies non-Regularized. This clustering simply differentiates between possible combinations of sugars in C2'-endo coformation, which are colored in orange. The population of these clusters are given also for the Davies Regularized ensemble (Blue) and Condon regularized ensemble (Red). The dsRNAs are shown only for the 12 central nucleotides using the dynamic secondary structure representation generated for the I+NMR ensemble.

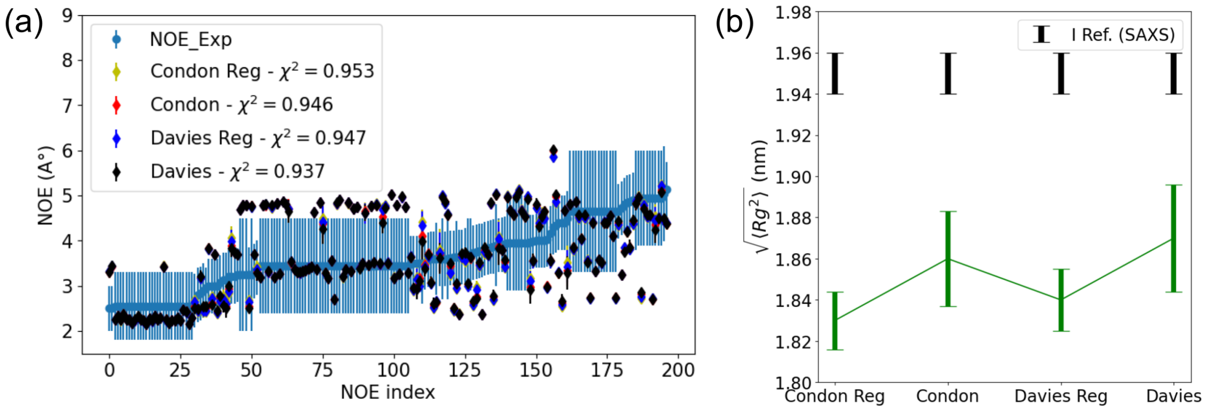


Figure 3.18 Comparision of ensembles obtained using Davies Karplus equations and not regularizing the fit (Davies) or regularizing the fit (Davies Reg), or by using Condon Karplus equations for in the same ways (Condon and Condon Reg). (a) NOEs data sorted with respect to experimental values (purple). Diamonds correspond to signals back-calculated from the trajectory with respect to different reweighted ensembles. (b) Square root of averaged radius of gyration squared ($\sqrt{\langle Rg^2 \rangle}$) for the 4 different ensembles . Regularized ensembles highly underestimate the $\sqrt{\langle Rg^2 \rangle}$ with respect to the experimental reference values extrapolated from SAXS data (black line). Not regularized ensembles slightly increase the agreement with experiments. Statistical errors are computed through bootstrapping.

force-field. Moreover, it would be interesting to investigate if also the force-fields of the other standard nucleotides may overstabilize C3'-endo conformations, based on the observation that also C2'-endo conformations of uracils result underestimated in the I ensemble. However, it is not trivial to dissect if the incorrect populations of uracils sugar puckering in the IIUI motif are a consequence of the overstabilization of I-U base pairing, or of the uracil AMBER force-field.

Further investigations on general structural features of the dsRNA, indicate how the inosine dsRNA is much more flexible, manifesting an ensemble RMSD larger by a factor 10 with respect to the adenosine counterpart. Moreover, the inosine dsRNA is characterized by having uncommon helical parameters with respect to standard A-form helix, allowing for relevant populations of conformers with the two strands being partially untwisted or having increased tendency for bending. This latter features is not found in the I ensemble, but only in the I+NMR and I+NMR+SAXS ensembles, suggesting that the bending is induced by the higher population of C2'-endo conformations. Finally the ensemble generated by enforcing ^{3}J scalar couplings (NMR) signals and averaged radius of gyration (SAXS), were validated against alternative experimental data. In particular, we showed how 197 NOE signals back-calculated from I+NMR and I+NMR+SAXS ensembles result to have increased agreement with the experimental values with respect to the I ensemble. Another validation not shown in this thesis was performed by our collaborator, and consisted in fitting of the entire SAXS spectra on a pool of structures extracted from the MD generated ensemble. Also in this case, the I+NMR+SAXS show significant increased accuracy with respect to the other ensembles. As the ensemble refinement process relied on the averaged radius of gyration as the only information obtained from SAXS, the improved reproduction of the entire SAXS spectra serves as further validation for the predicted ensemble.

Chapter 4

Fitting the N6-methyladenosine force field on denaturation experiments

N6-methyladenosine (m^6A) is the most common post-transcriptional modification found in nature, and is widely spread in both coding and noncoding RNAs [106, 107, 108, 109]. This modification consists in the methylation of the standard adenosine in position N6, and as all other modifications, it can impact RNA stability and structural dynamics. However, the most important effect of the N6-methylation appears to be the regulation of the interaction of RNA with specific proteins known as m^6A readers [41, 110, 111, 112, 42]. An important feature of m^6A is that it can exist in two possible conformations depending on the orientation of the methyl group with respect to the rest of the nucleobase. These two possible isomers are determined by the value of the torsional angle $\eta 6$ defined by the atoms N1-C6-N6-C10, and are called *syn* and *anti* (see Fig. 4.1a). According to multiple experimental evidences [36, 68], the *syn* conformation is expected to be the most stable for unpaired m^6A, whereas the *anti* conformation is the one expected for the m^6A when Watson-Crick paired with uracil in an internal position of a dsRNA. In order to use molecular dynamics simulations to investigate the impact of N6-methylation on RNA structural dynamics and recognition, we first need to ensure a parameterization which is able to reproduce some fundamental features of the m^6A nucleobase, as for example the correct *syn*/*anti* populations as predicted by the experiments. Interestingly, the modrna08 (Aduri) force-field [28], which is the most common AMBER parameterization for modified nucleotides, is not able to reproduce the correct *syn*/*anti* balance for the unpaired m^6A, as we will see in the following.

In this Chapter we show the results published in our work [113], where we refined the Aduri parameterization by fitting a subset of parameters against experimental free energies differences available in the literature, by using alchemical free-energy calculations (AFECs) [59]. To this end, we extend a recently-introduced force-field fitting strategy [114] to be usable in the context of alchemical simulations. The introduced approach allows training six charges and a dihedral potential so as to quantitatively reproduce methylation effects in denaturation experiments. The resulting force-field can be used to properly describe paired and unpaired m^6A in both *syn* and *anti* conformation.

4.1 Methods

A preliminar simulation to estimate $\Delta G_{syn/anti}$ of the m^6A nucleobase was performed using well-tempered metadynamics calculation [115], having $\eta 6$ as a collective variable. Metadynamics was performed using the PLUMED package [55], with a simulation length of 100 *ns*,

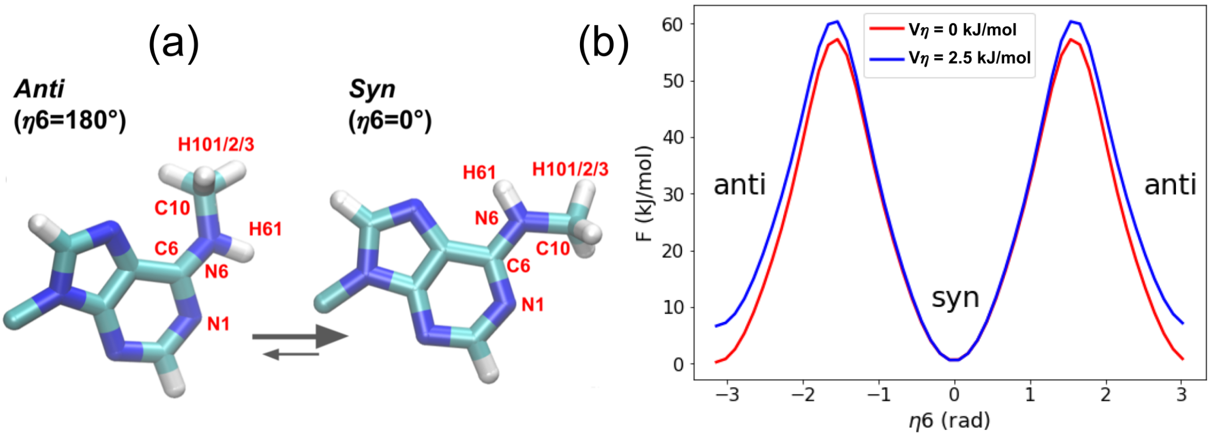
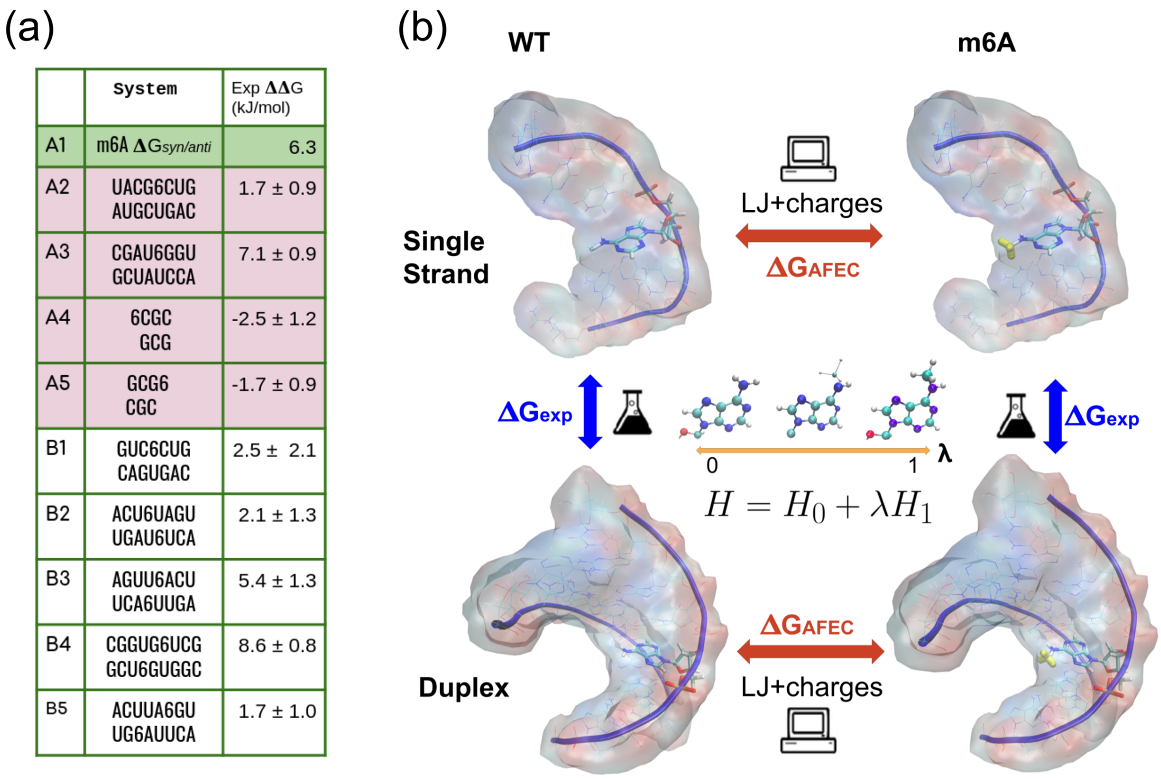


Figure 4.1 (a) N^6-methyladenosine (m^{6}A) nucleobase in *anti* (less stable) and *syn* (more stable) conformations [36, 68]. Atom names in red correspond to charges reparameterized in this work. (b) Free energy profiles along $\eta 6$ reconstructed looking at the bias potential produced through metadynamics along the collective variable $\eta 6$. The red line corresponds to the profile obtained with the standard Aduri parameterization ($V_\eta = 0$ kJ/mol), whereas the blue line corresponds to the one obtained using $V_\eta = 2.5$ kJ/mol.

depositing a Gaussian every 500 time steps, with initial height equal to 1.2 kJ/mol and width $\sigma = 0.35$. The bias factor was set to 10. We then used the free energy profile computed along $\eta 6$ (see Fig. 4.1b) to estimate the $\Delta G_{syn/anti}$ by integrating over the two corresponding minima. The list of experimental free energies fitted in this work are listed in panel (a) of Figure 4.2. To compare MD with these experiments, we had to perform the AFEC on m^{6}A, as described in section 2.2.3, in different contexts. We simulated the isolated m^{6}A nucleoside, 9 m^{6}A-methylated duplexes for which denaturation experiments are available in literature [36, 37] (see Table 4.2a), and the corresponding single-stranded RNAs. For the isolated m^{6}A nucleoside, we computed the $\Delta G_{syn/anti}$ by taking the difference in the ΔGs obtained with AFEC by methylating the adenosine in *syn* or *anti* conformations, for which experimental data are reported in Ref. 36. For systems A4 and A5, where m^{6}A is present as a dangling end and thus unpaired, we only performed AFEC corresponding to the *syn* conformation. For the other systems, we performed AFEC in the expected *anti* conformation. For the A2 and A3 systems we additionally performed AFEC in the unexpected *syn* conformation as a validation (population reported in Ref. 68 is $\approx 1\%$). In addition, we chose 5 more systems from Ref. 37, with the following criterion: they have a single methylation per strand and the methylation occurs in an internal position of the duplexes. For all these systems, we performed AFEC in the expected *anti* conformation. Simulation boxes consist of rhombic dodecahedrons containing RNA, water, Na$^+$ and Cl$^-$ ions with an excess salt concentration of 0.1 M. For a subset of the systems, further simulations were performed for a salt concentration of 1 M. The systems were energy minimized and subjected to a multi-step equilibration procedure for each replica: 100 ps of thermalization to 300 K in the NVT ensemble was conducted through the stochastic dynamics integrator (i.e., Langevin dynamics) [89], and other 100 ps were run in the NPT ensemble simulations using the Parrinello–Rahman barostat [90]. In production runs, the stochastic dynamics integrator was used in combination with the stochastic cell rescaling barostat [92] to keep the pressure at 1 bar. Equations of motion were integrated with a time-step of 2 fs. Long-range electrostatic interactions were handled by particle-mesh Ewald [93]. Each replica was simulated for 10 ns,



(a)

	System	Exp ΔΔG (kJ/mol)
A1	m6A $\Delta G_{syn/anti}$	6.3
A2	UACG6CUG AUGCUGAC	1.7 ± 0.9
A3	CGAU6GGU GCUAUCCA	7.1 ± 0.9
A4	6CGC GCG	-2.5 ± 1.2
A5	GCG6 CGC	-1.7 ± 0.9
B1	GUC6CUG CAGUGAC	2.5 ± 2.1
B2	ACU6UAGU UGAU6UCA	2.1 ± 1.3
B3	AGUU6ACU UCA6UUGA	5.4 ± 1.3
B4	CGGUG6UCG GCU6GUGGC	8.6 ± 0.8
B5	ACUUA6GU UG6AUUCA	1.7 ± 1.0

Figure 4.2 (a) List of experiments used for the fitting. Experiment A1 correspond m^{6}A single nucleotide $\Delta G_{syn/anti}$, whereas the other 9 $\Delta\Delta G$ correspond to destabilization induced by the methylation on dsRNAs, as measured by denaturation experiments. A1-A5 data are taken from [36], B1-B5 data are taken from [37]. In B2–B5 systems, the methylation occurs in both strands, however, the $\Delta\Delta G$s reported are intended per methylation. (b) Thermodynamic cycle used to compute m^{6}A induced destabilization on dsRNA. The relative free-energy change due to the modification can be estimated as the $\Delta\Delta G$ between AFECs performed on a duplex and on the corresponding single strand. This quantity can be directly compared to the difference in thermodynamic stability of duplexes with or without the modification, which can be measured experimentally through denaturation experiments.

for a total of 16×10 ns $= 160$ ns for each system. $\Delta\Delta G$s were obtained taking the difference between ΔGs computed by methylating the adenosine in *anti* or *syn* conformation on the duplex or dangling end, respectively, and the ΔG obtained methylating in *syn* conformation on the relative single strand. Transitions between *syn* and *anti* states were never detected during the alchemical simulations. In this way, the contribution to the free energy given by the *syn* (*anti*) conformation in the duplex (single strand or dangling end) was ignored. Indeed, we expect these contributions to be negligible based on the experimental evidences [36, 68], which show a *syn/anti* isomer preference when paired (≈1:100) versus unpaired (≈10:1). This was additionally verified with supplementary simulations performed on the A2 and A3 systems (see Table 2 in the Appendix). Moreover, we computed $\Delta G_{syn/anti}$ by performing the alchemical transformations on the isolated nucleoside in solution for the two isomers and computing their difference (see Table 2 in the Appendix)

Starting structures for MD simulations were built using the proto–Nucleic Acid Builder [83]. Single strands were generated by deleting one of the chains from duplex structures. All the MD simulations were performed using a modified version of GROMACS 2020.3 [84] which also implements the stochastic cell rescaling barostat [92]. The AMBER force-field was used for RNA [22, 86, 87], TIP3 model for water [85], and Joung and Cheatham parameters for ions [88]. As a starting parameterization for m^{6}A, we used AMBER adenosine parameters combined with modrna08 [28] charges for the nucleobase, adjusted to preserve the total charge of the nucleoside (as already described above). We refer to this parameterization as the Aduri force-field.

4.1.1 List of simulations

We hereby report the list of simulations performed in this work:

- We simulated a total of 22 systems reported in Table 2. For all of them, alchemical simulations were performed using Aduri, fit_A and fit_AB force-field parameters.
- For a subset of 7 systems, control alchemical simulations were performed at a higher salt concentration.
- For a subset of 4 systems, control alchemical simulations were performed at a higher temperature

This resulted in a total of $22 \times 3 + 7 + 4 = 77$ simulations. Each simulation was run with 16 replicas for 10 ns per replica, for a total simulated time of $77 \times 16 \times 10\text{ns} = 12.32\mu\text{s}$.

The size of the simulated systems depended on the number of simulated nucleotides. For the smallest A1 system (one nucleoside), the setup included ≈ 1500 water molecules, 3 Na$^+$ and 3 Cl$^-$ ions. Double stranded RNAs were simulated in boxes typically containing ≈ 6000 water molecules, the largest system being B4 with 7082 water molecules, 32 Na$^+$ and 14 Cl$^-$ ions. Single stranded RNAs were simulated using slightly smaller boxes typically containing ≈ 4500 water molecules. The smallest systems were A4 and A5, which were solvated in less than 3000 water molecules.

4.1.2 Fitting Procedure

We employ a fitting strategy based on reweighting [114] where a subset of the partial charges and a dihedral potential are adjusted to match experimental data. Specifically, we decided to fit charges of the atoms that are closer to the methyl group (N6, C6, H61, C10, H101/2/3, and N1, see Fig. 4.1). The total charge was maintained, leading to 5 free parameters associated with the partial charges. A single cosine was added to the η_6 torsional angle identified by atoms N1–C6–N6–C10: $U(x) = V_\eta[1 + cos(\eta_6(x) - \pi)]$. This angle controls the *syn/anti* relative populations, leading to a total of 6 parameters, and the shift is chosen so that a positive value of V_η favors *syn* configurations over *anti*.

To optimize the calculation of the total energy of the system at every iteration of our fitting procedure, where up to 6 charges were possibly modified, we notice that the total energy of the system is a quadratic function of the charge perturbations ΔQ_i. Without loss of generality, one can write the energy change associated to charges and torsion perturbation as

$$\Delta U(x) = \sum_{i=1}^{5} K_i(x)\Delta Q_i + \sum_{i=1}^{5}\sum_{j=i}^{5} K_{ij}(x)\Delta Q_i \Delta Q_j + V_\eta[1 + cos(\eta_6(x_i) - \pi)] \tag{4.1}$$

In total, for every analyzed snapshot (x), 20 coefficients (K_i and K_{ij}) can be precomputed that allow obtaining the energy change for arbitrary choices of ΔQ with simple linear algebra operations, without the need to recompute electrostatic interactions explicitly. The coefficients were obtained by using GROMACS in rerun mode for 20 sets of test charge perturbation, which were extracted from a Gaussian with zero average and standard deviation set to 1 e. The perturbations were constructed to maintain constant the total charge. Importantly, this approach correctly takes into account the effect of charge perturbations on 1–4 interactions, where electrostatics is scaled with a force-field-dependent fudge factor, as well as on 1–2 and 1–3 interactions, for which it is discarded, and interaction with all the periodic images. The second order expansion above is exact if one neglects roundoff errors. The magnitude of charge perturbations was chosen to minimize such errors. Eq. 2.10 should then be suitably modified replacing ΔE with $\Delta E + \Delta U$. Its derivatives with respect to the free parameters (charge and dihedral potential coefficient) can be computed as well.

Our fitting is based on the minimization of an L2-regularized cost function defined as follows:

$$C = \chi^2 + \alpha \sum_{i=0}^{5} \Delta Q_i^2 + \beta V_\eta^2 = \chi^2 + \alpha [\sum_{i=1}^{5} \Delta Q_i^2 + (\sum_{i=1}^{5} \Delta Q_i)^2] + \beta V_\eta^2 \tag{4.2}$$

where the regularization terms on the charges and the torsional η_6 are governed by the hyperparameters α and β and are needed to avoid overfitting on the training set.

Here we assumed that $\Delta Q_0 = -\sum_{i=1}^{5} \Delta Q_i$, to preserve the total charge. The χ^2 measures the discrepancy between computations and experiments:

$$\chi^2 = \frac{1}{N_{exp}} \sum_{i=1}^{N_{exp}} \frac{(\Delta G_{2i-1} - \Delta G_{2i} - \Delta\Delta G_i^{exp})^2}{\sigma_i^2} \tag{4.3}$$

Here, σ_i corresponds to the experimental error, and $\Delta G_{2i-1} - \Delta G_{2i} = \Delta\Delta G_i^{AFEC}$ is difference in free energy differences computed respectively on the dsRNA $(2i-1)$ and the ssRNA $(2i)$.

Alchemical ΔGs are computed through a reweighting procedure via the equation:

$$\Delta G^{AFEC} = -k_B T \log(\frac{\sum_i^{N_{frame}} w_i e^{-\beta[\Delta E(x_i) + \Delta U(x_i)]}}{\sum_i^{N_{frame}} w_i}) \tag{4.4}$$

where w_i are the weight derived by the binless WHAM on the original set of energies (see section 2.2.3). ΔU is the perturbed potential energy (4.1) and ΔE is:

$$\Delta E(x_i) = E_{\lambda=1}(x_i) - E_{\lambda=0}(x_i) \tag{4.5}$$

This function is minimized using the L-BFGS-B method [116] as implemented in SciPy [117], for which the derivative of the cost function with respect to the fitted parameters is needed. This should be computed for charges as follows:

$$\frac{\partial C}{\partial \Delta \mathbf{Q}} = \frac{\partial C}{\partial \Delta \mathbf{G}} \frac{\partial \Delta \mathbf{G}}{\partial \mathbf{L}} \frac{\partial \mathbf{L}}{\partial \Delta \mathbf{Q}} \tag{4.6}$$

Here we introduced the 20-components vector

$$\mathbf{L} = (\Delta Q_1, \Delta Q_2, ..., \Delta Q_1 \Delta Q_1, \Delta Q_1 \Delta Q_2, ..., \Delta Q_5 \Delta Q_5) \tag{4.7}$$

For the torsional parameter instead we have:

$$\frac{\partial C}{\partial V_\eta} = \frac{\partial C}{\partial \Delta \mathbf{G}} \frac{\partial \Delta \mathbf{G}}{\partial V_\eta} \tag{4.8}$$

The derivative of the free-energy change with respect to **L** components can be computed as

$$\frac{\partial \Delta G_k}{\partial L_l} = \langle K_l \rangle_k = \sum_i^{N_{frame}} w_i K_l^i e^{-\beta[\Delta E(x_i)+\Delta U(x_i)+V_\eta[1+cos(\eta_6(x_i)-\pi)]]} \tag{4.9}$$

The derivative of the free-energy change with respect to the torsional parameter can be computed as

$$\frac{\partial \Delta G_k}{\partial V_\eta} = \langle [1+cos(\eta_6(x_i)-\pi)] \rangle_k = \sum_i^{N_{frame}} w_i[1+cos(\eta_6(x_i)-\pi)]e^{-\beta[\Delta E(x_i)+\Delta U(x_i)+V_\eta[1+cos(\eta_6(x_i)-\pi)]]} \tag{4.10}$$

The result crucially depends on the choice of the hyperparameters α and β. Lower values for the hyperparameters imply that larger corrections are allowed, with the risk of overfitting, and thus lower transferability to new experiments. Higher values for the hyperparameters imply that lower corrections are allowed, with the risk of underfitting, and thus lower accuracy in reproducing experimental data. The sweet point could be in principle found with a cross-validation (CV) procedure and a scan over possible values for α and β [114, 23]. For the smallest dataset (set A1-A5 in Fig. 4.2), we used a leave-one-out CV strategy, i.e., we trained the parameters on all systems except one. For the largest dataset (set AB in Fig. 4.2), we used a leave-3-out strategy, iteratively training the parameters on 7 randomly chosen experiments and validating on the 3 left-out experiments. In both cases, we then assessed the transferability of the model by evaluating its average χ^2 on the system (or the subset of systems) that was left out.

4.1.3 Statistical Significance

When recomputing energies through a reweighting procedure, particular attention must be taken towards the statistical significance that may be lost during the computation, by reducing the effective sample size of the data set. This is usually monitored by computing the Kish effective sample size [118, 119]. In our case, the most affected ensemble is the one corresponding to m^{6}A ($\lambda = 1$). We thus monitor the Kish size computed using weights corresponding to the $\lambda = 1$ ensemble, defined as

$$KS_{\lambda=1} = \frac{[\sum_x w(x)e^{-\beta(\Delta E(x)+\Delta U(x))}]^2}{\sum_x [w(x)e^{-\beta(\Delta E(x)+\Delta U(x))}]^2} \tag{4.11}$$

We then compare it with the Kish size obtained with the original force-field, defined as

$$KS^0_{\lambda=1} = \frac{[\sum_x w(x)e^{-\beta\Delta E(x)}]^2}{\sum_x [w(x)e^{-\beta\Delta E(x)}]^2} \tag{4.12}$$

To quantify how much statistical efficiency is lost due to the reweighting to a modified set of parameters we use the Kish size ratio (KSR), that we define as

$$KSR = \frac{KS_{\lambda=1}}{KS^0_{\lambda=1}} \tag{4.13}$$

4.2 Preliminary Estimation of m^{6}A *syn/anti* populations with Aduri Force-Field

Before going through the results of our fitting, we report here the results of a preliminary full-atom biased simulation performed for the stand-alone m^{6}A nucleoside, which we used to estimate the $\Delta G_{syn/anti}$ through the free energy profile reconstructed using WT-MetaD [115]. The torsional angle $\eta 6$ was used as collective variable for the metadynamics, and its relative free energy profile is shown in fig 4.1b. We computed the $\Delta G_{syn/anti}$ by integrating over the two corresponding minima, and it results in $\Delta G_{syn/anti} = 1.5$ kJ/mol, which is an underestimation with respect to the experimental value 6.3 kJ/mol. In principle, the correct $\Delta G_{syn/anti}$ could be recover by simply adding the potential of a single torsional term in this form

$$U(x_i) = V_\eta [1 + cos(\eta_6(x_i) - \pi)] \tag{4.14}$$

For positive values of the parameter V_η this correction penalizes the *anti* conformations. We then used a reweighting approach to tune the parameter V_η in order to enforce the experimental value of $\Delta G_{syn/anti}$. Specifically, we assigned a weight $w(x)$ to each frame, computed as

$$w(x) \propto e^{\beta B(\eta_6(x))} e^{-\beta V_\eta [1+cos(\eta_6(x)-\pi)]} \tag{4.15}$$

Here $B(\eta_6)$ is the bias potential constructed during MetaD simulation.

The $\Delta G_{syn/anti}$ was then obtained as

$$\Delta G_{syn/anti} = -\frac{1}{\beta} \log(\frac{\sum_{x \in syn} w(x)}{\sum_{x \in anti} w(x)}) \tag{4.16}$$

We iteratively adjusted V_η until we found that $V_\eta = 2.5$ kJ/mol results in a $\Delta G_{syn/anti} = 6.4 \pm 0.3$ kJ/mol, which is compatible with experiment. Statistical error was computed using block analysis [115]. Figure 4.1b shows the free ernergy profiles reconstructed along $\eta 6$ for $V_\eta = 0$ (reference) and $V_\eta = 2.5$ kJ/mol.

Although we are able to enforce the correct $\Delta G_{syn/anti}$ for the unpaired nucleotide by simply modifying a torsional parameter, the Aduri force-field is still not capable to reproduce some other important experimental evidences, as we will see in the following. In particular, we will show that the torsional term needed to enforce the $\Delta G_{syn/anti}$ would cause an incorrect estimation of other observables.

4.3 Fitting Results

In this work, we fit point charges and a single torsional potential correction for an m^{6}A RNA residue using alchemical MD simulations and a set of experimental data, following the scheme shown in Fig. 4.2b. In all the fittings, charges and torsional potential were subject to L2 regularization with hyperparameters α and β, respectively. We initially employed only the first 5 experimental data points of Table in panel (a) of Fig. 4.2, namely (A1) $\Delta G_{syn/anti}$ for a nucleobase and (A2–A5) $\Delta\Delta G$ in melting experiments [36]. Thus, we first report the results obtained with such a set of charges, including a validation done on a more recent set of melting experiments (B1–B5) [37]. We then report results obtained with charges that were fitted on the entire dataset (A1–A5 and B1–B5). As a reference, results obtained with the Aduri *et al* [28] modifications (modrna08) for the commonly used AMBER force-field are also reported, either *as is* or complemented with a custom torsional correction that results in a $\Delta G_{syn/anti}$ matching experiment A1. All the calculated ΔGs are reported in Table 2 in the Appendix.

4.3.1 Fitting on the smaller dataset

For this first fitting, we only employed data set A1–A5 (see Fig. 4.2). χ^2 errors were computed using Eq. (4.3) and setting the experimental error of each data point (σ_i) to be equal to each other and to 1 kJ/mol.

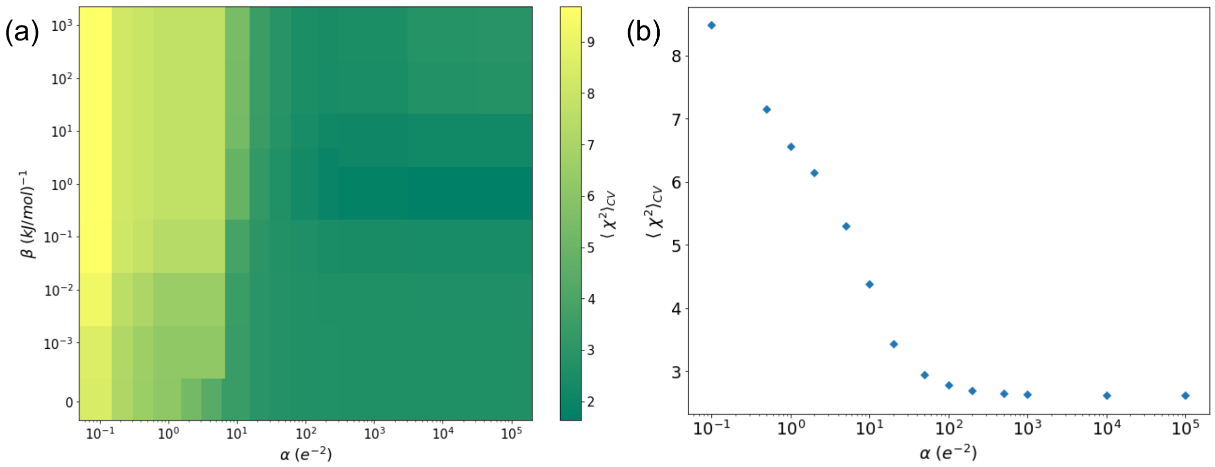


Figure 4.3 Cross validation error obtained for the fitting on the initial dataset A1–A5, with a leave-one-out-procedure, shown as a function of the two regularization hyperparameters α, for charges, and β, for the torsional potential (panel a). Darker green colors correspond to lower values of the average χ^2 computed on the systems left out iteratively from the fitting. (b) Projection of data along $\beta = 0$

Figure 4.3a reports the results of a cross-validation test performed with a leave-one-out procedure. Namely, we fit the whole experimental dataset leaving out one experimental data point at a time, and report the average error on the left-out experiment. In this leave-one-out procedure, we decided not to iterate on the $\Delta G_{syn/anti}$ experiment (A1), since this is expected to be crucial to correctly reproduce the conformation of non-Watson–Crick-paired residues (mostly *syn*). From this map, we can hardly appreciate any variation of the χ^2 along the vertical axis corresponding to the β hyperparameter. This suggests that β could be set to zero, thus simplifying all subsequent hyperparameter scans. Conversely, the χ^2 grows significantly for low α values. This implies that regularization of charges is required to avoid overfitting. In general, one should expect a minimum to be observed in this type of hyperparameter scan [114, 23]. This is not the case here for the α scan, as it can be appreciated in Figure 4.3b, showing projection on α for $\beta = 0$, implying that the performance of the parameters on a given system is not improved when excluding that system from the training set. This is likely due to the small data set employed.

Figure 4.4a shows the optimized parameters (charge and torsional corrections) as a function of the regularization hyperparameter α while fixing $\beta = 0$. A transition can be seen at $\alpha \approx 10$. Namely, when $\alpha > 10$, parameters have a smooth dependence on α, whereas when $\alpha < 10$, both the charges and the torsional potential change suddenly. In the limit $\alpha \to \infty$, it can be seen that charge corrections tend to zero with an inverse law dependence, which is expected for L2 regularization, and the torsional correction tends to $V_\eta \approx 1.5$ kJ/mol, which corresponds to the amplitude of the torsional potential that optimizes the χ^2 without modifying the charges of the reference Aduri *et al* model. We notice that $\Delta G_{syn/anti}$ obtained when using the Aduri *et al.* force-field is ≈ 1.7 kJ/mol, and thus this correction results in $\Delta G_{syn/anti} \approx 1.7 + 2 \times 1.5 = 4.7$ kJ/mol, which is still smaller than the experimental reference ≈ 6 kJ/mol. The obtained parameters indeed strike a balance between favoring the *syn* state in the isolated nucleoside and not favoring it too much in the single-stranded calculations used to predict the $\Delta\Delta G$ from melting experiments, which would lead to too large destabilizations associated with the methylation. When α is decreased, the optimal torsional correction changes, since all the parameters are

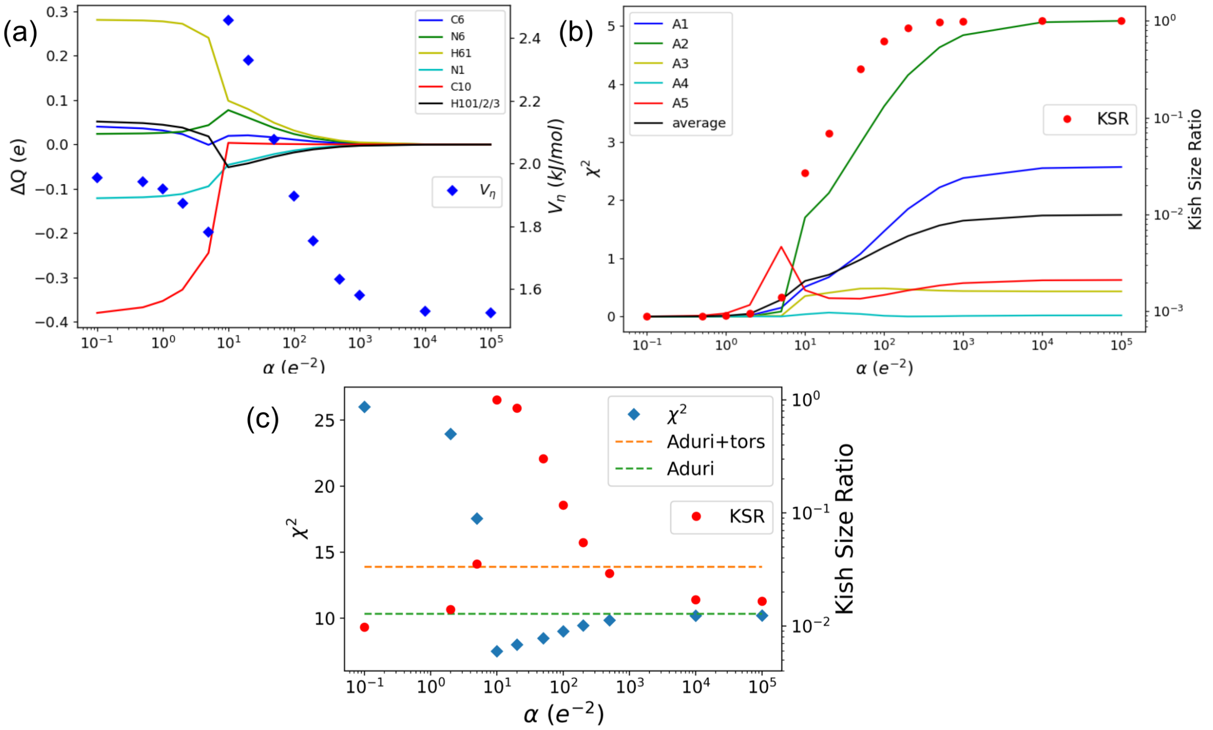


Figure 4.4 Parameters (ΔQ and V_η) obtained from the entire initial dataset as a function of α, with $\beta = 0$ (panel a). χ^2 errors for individual experiments and Kish size ratio (KSR, see text for definition) obtained using parameters fitted on the entire initial dataset as a function of α, with $\beta = 0$ (panel b). Validation on the second dataset (B1–B5) of the parameters obtained on the first dataset (panel c). Results using Aduri parameters are shown as horizontal lines, either as reported in the original paper (green) or including a single torsional correction to obtain the correct *syn/anti* population (data point A1)

coupled. This confirms that charges and torsional parameters should be fitted simultaneously.

Figure 4.4b shows the individual χ^2 associated with the same hyperparameter scan. The average χ^2 error is, by construction, monotonically increasing with α, and most of the individual errors follow the same trend. Figure 4.4b also shows the statistical efficiency of the analysis, quantified by the relative reduction of the Kish effective sample size associated with reweighting. A low number here indicates that the tested charges are so different from those employed in the simulation to make the result statistically not significant. The Kish size displays a significant drop for $\alpha < 10$, indicating that results in this regime might be not significant. This is a likely explanation for the discontinuous behavior observed in Fig. 4.4a.

We then tested the charges obtained with this reduced training set on the newer data set B1–B5, see Fig. 4.2a, which was not included in the training phase. This set of data involves 5 recently published melting experiments [37], 4 of which have m^{6}A occurring in both chains of the duplex. We notice that double methylations are expected to lead to an even lower statistical efficiency of the reweighting procedure. We thus performed this analysis by reweighting simulations that were generated using the set of parameters derived fitting on systems A1-A5 for $\alpha = 10$ and $\beta = 0$. Since this parameterization is closer to the right solution of the fitting when compared with the Aduri one, it obtains higher Kish size values in the relevant α range (see Fig. 4.4c). The χ^2 computed on the second data set shows that an optimal result can be obtained by setting $\alpha \approx 10$. We also compared with results obtained using the original Aduri charges and optionally including a torsional correction to fix the *syn/anti* balance. These results are obtained with direct simulation, that is without reweighting. It can be seen that the results with the parameters trained on systems A1–A5 largely outperform those obtained with Aduri parameters on systems B1–B5, thus confirming the transferability of the parameters. Aduri+tors parameterization corresponds to setting V_η =2.35 kJ/mol in such a way to perfectly fit experiment A1 (single nucleoside) without modifying charges. The χ^2 computed for Aduri+tors demonstrates that acting exclusively on the torsional is not sufficient to reproduce both $\Delta G_{syn/anti}$ and melting experiments. It is also important to note that the improvement in reproducing experiments is obtained by changes in the partial charges that are small when compared to differences between charges derived with the standard restrained electrostatic potential protocol [120] in different conformations, as we will discuss in the following sections.

4.3.2 Fitting on the full data set

Next, we perform a fitting using the full data set reported in Fig 4.2a. Since the variability of error in this data set is larger, we here computed χ^2 using the experimental errors reported in Table 4.2a. For the $\Delta G_{syn/anti}$ experiment, for which an experimental error is not reported, we used a nominal $\sigma = 0.5$ kJ/mol to assign to this experiment a larger weight when compared to the other data points corresponding to melting experiments.

Figure 4.5a reports the results of a cross-validation test performed with a leave-three-out procedure. Namely, we randomly select seven systems to be used in training and we report the average χ^2 error obtained for the remaining three systems. This time also system A1 was allowed to be left out of the training set. Results are qualitatively consistent with those obtained with the smaller data set (see Fig. 4.3). It is difficult to appreciate any variation of the χ^2 along the vertical axis corresponding to the β hyperparameter, suggesting that we can safely set $\beta = 0$. We also do not find any clear minimum when scanning over α, as it can be appreciated by Figure 4.5b, showing projection on α for $\beta = 0$. Figure 4.6a shows the parameters as a function of the regularization hyperparameter α while fixing β. A clear transition can be seen at $\alpha \approx 20$. The average χ^2 error is monotonically increasing with α, but some of the systems have a non-trivial behavior (Fig. 4.6b). The Kish size shows a significant drop for $\alpha < 50$, showing that results in this regime might be not statistically reliable. We thus select the parameters obtained with $\alpha = 50$ as the optimal ones trained on the entire data set.

We then compare the performance of several different sets of parameters in reproducing all the available experimental data points. Namely, we compare (a) the original Aduri parameters (Aduri), (b) the Aduri parameters augmented with a torsional correction to enforce the correct *syn/anti* balance in a nucleobase (Aduri+tors), (c) the parameters obtained fitting on the initial dataset (A1–A5), with hyperparameter $\alpha = 10$ (fit_A), and (d) the parameters obtained fitting on the full dataset (A1–A5 and B1–B5), with hyperparameter $\alpha = 50$ (fit_AB). Free energies are computed directly from the alchemical simulations, that is without reweighting. Results are reported in Fig. 4.7. The quality of the fit is also summarized in the reported χ^2 values. The addition of a simple torsional correction to the Aduri parameters results in a decrease in the overall χ^2 from 15.23 to 9.17. However, this decrease is dominated by the χ^2 of the A1 datapoint, which is reduced from $\chi^2 = 84.64$ to zero. Conversely, the χ^2 averaged on all the other experiments increases from $\chi^2 = 7.57$ to $\chi^2 = 10.19$. This indicates that including in the fitting the single A1 datapoint makes the agreement with denaturation experiments worse. On the other hand, the two sets of parameters obtained in this work (fit_A and fit_AB) display a significantly better agreement with experimental data. Note that fit_A, surprisingly, performs moderately better than fit_AB. The reason is that fit_AB, based on systems with double methylation and

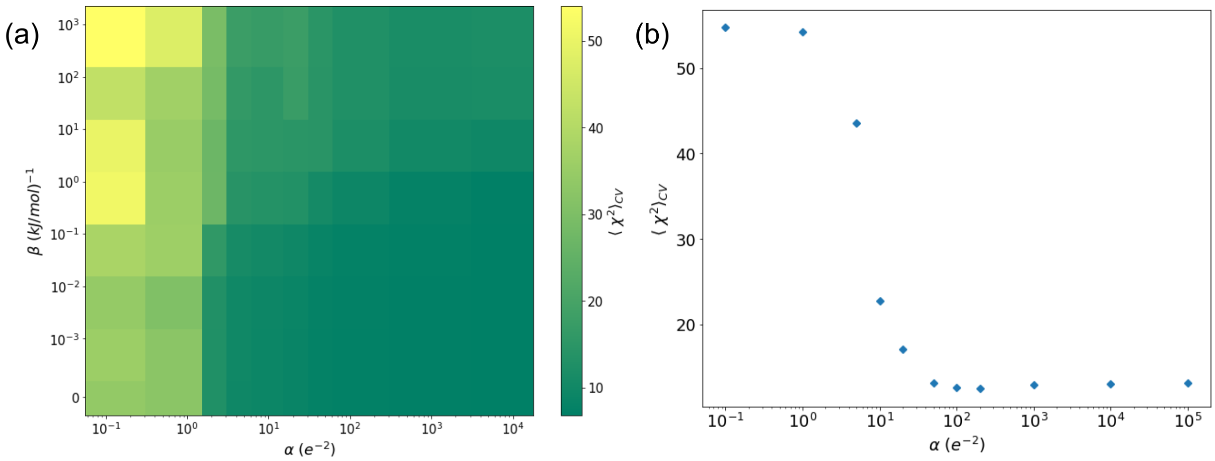


Figure 4.5 Cross-validation error obtained fitting on the entire data set with a leave-three-out-procedure, shown as a function of the two regularization hyperparameters α, for charges, and β, for the torsional potential (panel a). Darker green colors correspond to lower values of the average χ^2 computed on the systems left out iteratively from the fitting. Projection of data along $\beta = 0$

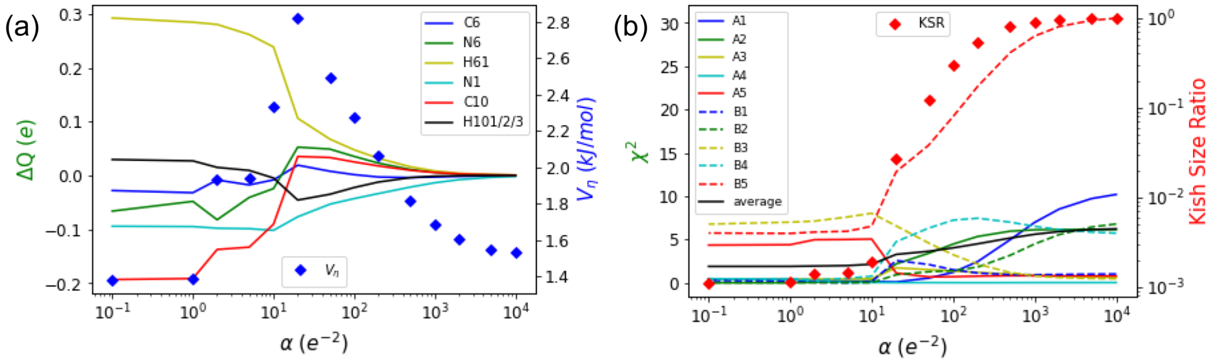


Figure 4.6 Parameters (ΔQ and V_η) obtained from the entire dataset as a function of α, with $\beta = 0$ (panel a). χ^2 errors for individual experiments and Kish size ratio (KSR, see text for definition) using parameters fitted on the entire initial dataset as a function of α, with $\beta = 0$ (panel b).

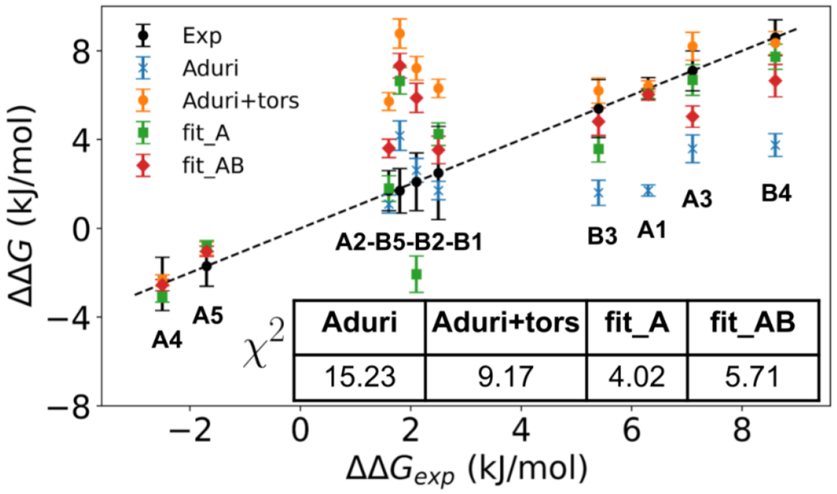


Figure 4.7 $\Delta\Delta G$ computed for each of the ten analyzed systems with 4 different sets of parameters. fit_A are parameters obtained fitting on the first data set (A1–A5) with regularization $\alpha = 10$. fit_AB are derived fitting on the entire data set (A1–A5 and B1–B5) for $\alpha = 50$. χ^2 obtained for each force-field set of parameters are shown in the table.

	C6 (e)	**N6** (e)	**H61** (e)	**N1** (e)	**C10** (e)	**H100** (e)	V_η (kJ/mol)
fit_A	0.019	0.077	0.099	-0.046	0.004	-0.051	2.46
fit_AB	0.009	0.049	0.067	-0.053	0.033	-0.035	2.49

Table 4.1 Charge modifications (ΔQs) and torsional potential (V_η) for the fitting performed on the smaller dataset (fit_A, $\alpha = 10$) and for the fitting performed on the larger dataset (fit_AB, $\alpha = 50$). For future simulations, we recommend using fit_A, which leads to a lower error on the larger set of available experiments.

thus lower statistical efficiency, was performed with a higher regularization hyper parameter and thus parameterization closer to the reference one. The fitted parameters are summarized in Table 4.1.

4.3.3 Sets of charges

As observed in the previous subsections, we note that the improvement in reproducing experiments through the fitting is obtained by relatively small changes in the partial charges, as it can be understood comparing different sets of charges derived for m^{6}A in different ways.

Table 1 in the Appendix shows all the sets of charges considered for the m^{6}A nucleobase. The first column corresponds to the charges from Aduri *et al.* [28], adjusted to be compatible with the current AMBER force-field [22, 86, 87]. The following columns represent the charges obtained in our fittings, using: the regularized fitting (with $\alpha = 10$) on set A, fit_A; the regularized fitting (with $\alpha = 50$) on set AB, fit_AB. In addition, we show charges that we derived following the standard procedure on a nucleobase. We here considered the geometry of both isomers (*syn* and *anti*), computed the electrostatic potentials of the N6-methylated adenine base by Gaussian 09 [121] using the HF/6-31G* level of theory, subsequently deriving the partial charges via the RESP method [120]. For these calculations we replaced the sugar with a closing

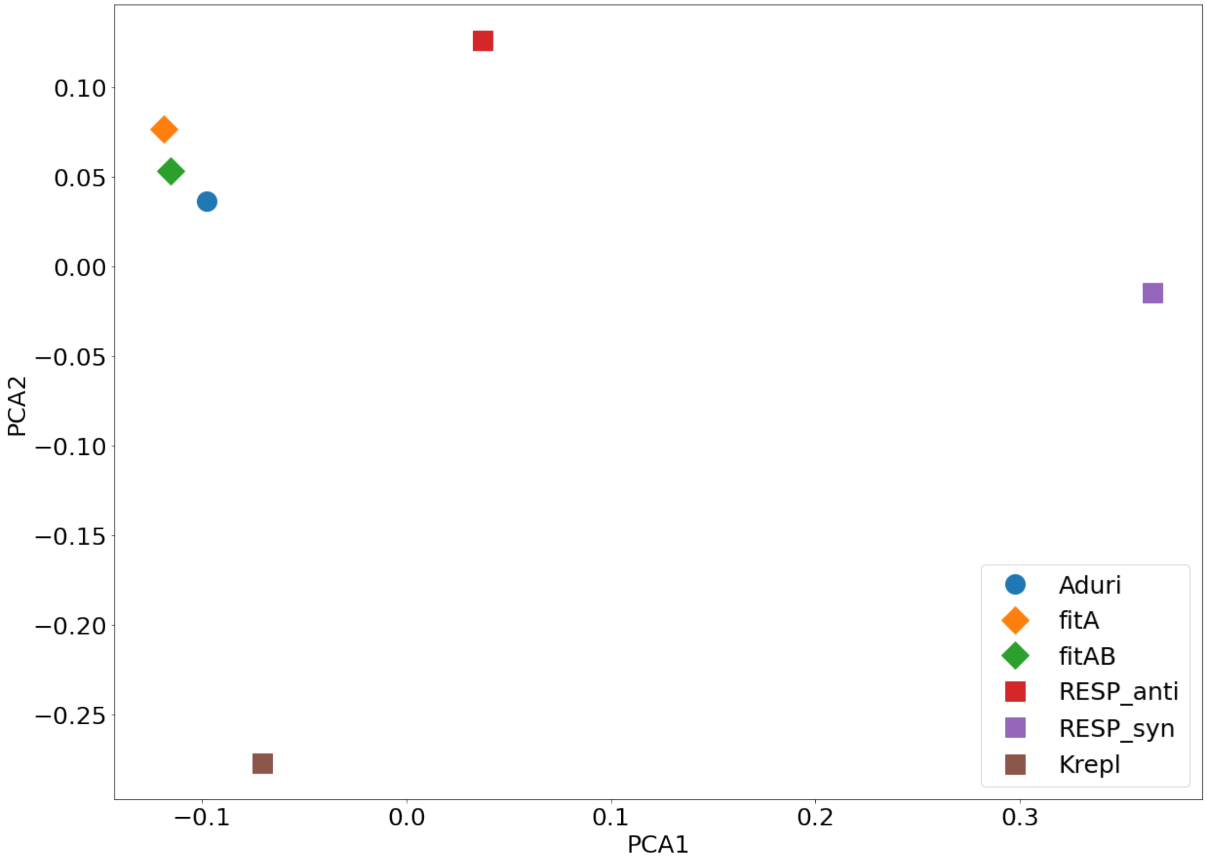


Figure 4.8 PCA performed giving as input all the charges of the nucleobase.

methyl group, as done by Aduri *et al.* We notice however that Aduri *et al.* does not report the chosen isomer, which is likely a *syn*, the most populated one for an isolated nucleobase. The last set of parameters (Krepl) have been used in Ref. [112] and were kindly shared by Miroslav Krepl. In order to visualize these sets of charges in a space of reduced dimensionality, we perform principle components analysis (PCA) on the charges data sets, considering the entire nucleobase (Fig. 4.8). As it can be appreciated in the PCA analysis, the difference by the charges resulting from our fitting procedures and those reported by Aduri is very small, and significantly lower than the typical variability between different sets of charges obtained with the same standard QM methods, but with slightly different procedure. Based on these qualitative observations, we can consider our fitting a delicate refinement of a reference force-field (Aduri in this case), which despite this small modification on partial charges, is still able to have a significant impact on experimental observables, adjusting the correct chemistry and physics of the nucleobase and recovering agreement between very sensible experimental and computational free energy measurements.

4.3.4 Relative stability of *syn* and *anti* conformations

One piece of the experimental information that we implicitly used in our fitting procedure is the relative stability of *syn* and *anti* conformations in a nucleotide. We indeed assumed a predominant population of *syn* conformation for the unpaired nucleotides used in the reference single-stranded systems. We also assumed that m^{6}A adopts exclusively its *anti* conformation when paired, in agreement with experiments [36, 68]. In particular, Ref. 68 reports that, for the most common G6C sequence, m^{6}A forms a Watson–Crick base pair with uridine that transiently

	Aduri	**Aduri+tors**	**fit_A**	**fit_AB**	**Exp**
A1. $\Delta G_{syn/anti}$	1.71 ± 0.25	6.33 ± 0.25	6.07 ± 0.21	6.04 ± 0.26	6.3
A2. $\Delta G^{dup}_{syn/anti}$	- 7.7 ± 0.5	- 3.1 ± 0.4	- 10.4 ± 0.6	- 7.8 ± 0.4	∼ -11
A3. $\Delta G^{dup}_{syn/anti}$	- 5.4 ± 0.5	- 0.8 ± 0.4	- 4.9 ± 0.6	- 5.8 ± 0.5	–

Table 4.2 Free-energy differences between *syn* and *anti* isomer states in systems A1–A3. The last column corresponds to experimental estimates, whereas the other columns correspond to computed $\Delta\Delta G$ for different parameterization. Energies are given in kJ/mol units.

exchanges on the millisecond time-scale between the main substate (*anti*) and a low populated (1%), singly hydrogen-bonded and mismatch-like conformation through isomerization of the methylamino group to the *syn* conformation. This population corresponds to a $\Delta G^{duplex}_{syn/anti} \approx -11$ kJ/mol. We *a posteriori* validated this population by performing alchemical transformations on the duplex systems enforcing the *syn* conformation. The predicted $\Delta G_{syn/anti}$ for a nucleotide and two of the tested duplexes are reported in Table 4.2, where the corresponding experimental values are also included. For the A1 experiment, as expected, the proposed sets of parameters closely match the experimental value that was used during training. The Aduri *et al.* force-field underestimates the $\Delta G_{syn/anti}$, resulting in a relatively high population of the unexpected *anti* conformation in a nucleoside. This difference can be directly corrected with a torsional potential applied on the η torsion (Aduri+tors). However, when analyzing duplexes A2 and A3 with the Aduri+tors parameters, we found that the predicted $\Delta G_{syn/anti}$ would be close to zero, in fact resulting in the assumption of neglecting the *syn* conformation in duplexes in our alchemical calculations to be difficult to justify, and in disagreement with experimental findings. In other words, the original Aduri charges allow reproduction of the relative stability of *syn* and *anti* conformations either in the paired state (Aduri parameters) or the unpaired state (with torsional correction), but not in both simultaneously. Remarkably, the sets of parameters proposed here, which also contain a torsional term penalizing the *anti* conformation, result in a significantly higher value for $\Delta G^{dup}_{syn/anti}$, much closer to a qualitative agreement with the experiment. This suggests that the proposed parameters better describe the interactions of the m^{6}A nucleobase with the surrounding environment and are thus more transferable. We notice that the relative stability of *syn* and *anti* conformations is predicted to be sequence dependent, being different for system A3 (sequence U6G).

To gain insight into how the m^{6}A–U pairings occur in the duplexes, we analyzed snapshots of system A2, both for m^{6}A in *syn* and *anti*, together with histograms of distances between atoms belonging to the two nucleobases (Fig. 4.9). The reported histograms are unimodal and with an increased average associated with the distortion of the A-U Watson–Crick pairings due to the steric clash induced by the methylation. However, the hydrogen bond between A-N1 and U-H3 is present, in agreement with what has been suggested previously [68].

4.3.5 Interpretation of the fitted parameters

To provide an interpretation for the obtained parameters, we performed a few additional fittings. In particular, we investigated which charges have a major impact on enforcing agreement with experiments. First, we notice that Aduri charges for N1 and H61, which are involved in Watson Crick parings with the paired uridine, have partial charge absolute value significatively lower compared to the standard adenine parameters (0.28948 vs 0.41150 for H61, -0.675968 vs.

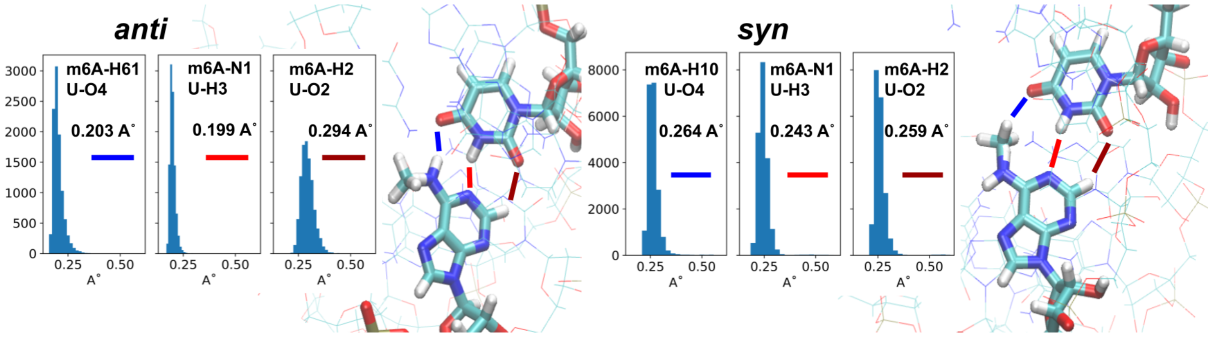


Figure 4.9 Interfacing atom distances for m^{6}A-U pairing in system A2 in Fig. 4.2a, for *anti* conformation (left) and *syn* (right). Histograms show unimodal distributions, and the averaged values are indicated in the box. Distances are sampled from the alchemical trajectories considering only the $\lambda = 1$ replica. In the *syn* conformation, m^{6}A-H10 correspond to the hydrogen of the methyl group closest to the uracil oxygen O4.

	ΔQ (e)	V_η (kJ/mol)	χ^2	**KSR**
H61-N1	0.0652	2.37	4.42	0.18
N6-N1	0.0802	1.98	4.52	0.31
H61-H100	0.04932	1.72	5.92	0.54
N6-H100	0.0648	1.49	5.93	0.74

Table 4.3 Result for fitting 2 charges plus the torsional with hyperparameters set to 0. Only the ΔQ associated to the first atom is shown (H61 or N6). The ΔQ associated to N1 has the same absolute value and opposite sign. The H100 charge is equally distributed on the 3 hydrogens of the methyl group, so that the charge on each hydrogen has 1/3 absolute value and opposite sign when compared with the reported ΔQ.

-0.76150 for N1). This may lead to a weakening of hydrogen bonds which may cause an overestimation of the destabilization induced on duplexes, as we observed in Aduri+tors cases (see Fig. 4.7c). The results of our fitting systematically increase the absolute value of H61 ad N1 partial charges, hence resulting in a stronger Watson Crick pairing. At the same time, the torsional term allows to reproduce the correct *anti* isomer penalty. Parameters are coupled, so that it is necessary to fit them simultaneously so as to avoid double counting effects. To demonstrate that N1 and H61 are the most important charges to tune in order to reproduce experiments by strengthening hydrogen bonds, we performed 4 further fittings on the entire data set (AB) by tuning only the torsional plus 2 charges at time, respectively for the pairs N1-H61; N1-N6; H61-H101/2/3; N6-H101/2/3, which are taking into account atoms that have systematic positive and negative ΔQ both in fit_A and fit_AB. Results are summarized in Figure 4.10.

Interestingly, when fitting only 2 charges, the results are converging for α going to zero. Furthermore, the Kish Size Ratio obtained for $\alpha = 0$ is always greater than 0.18, demonstrating that statistical significance is always maintained when fitting only two charges. For α and β set to zero, results of the fitting are summarized in Table 4.3.

The lowest χ^2 is obtained in H61-N1 case with a value of 4.42 (for comparsion, the fitted χ^2 obtained in fit_AB is 3.61), confirming the hypothesis that tuning these two charges is crucial to reproduce experiments. A slightly larger χ^2 is obtained in N6-N1 case. Figure 4.10 also shows that the correction on the torsional angle is highly coupled with modifications of N1

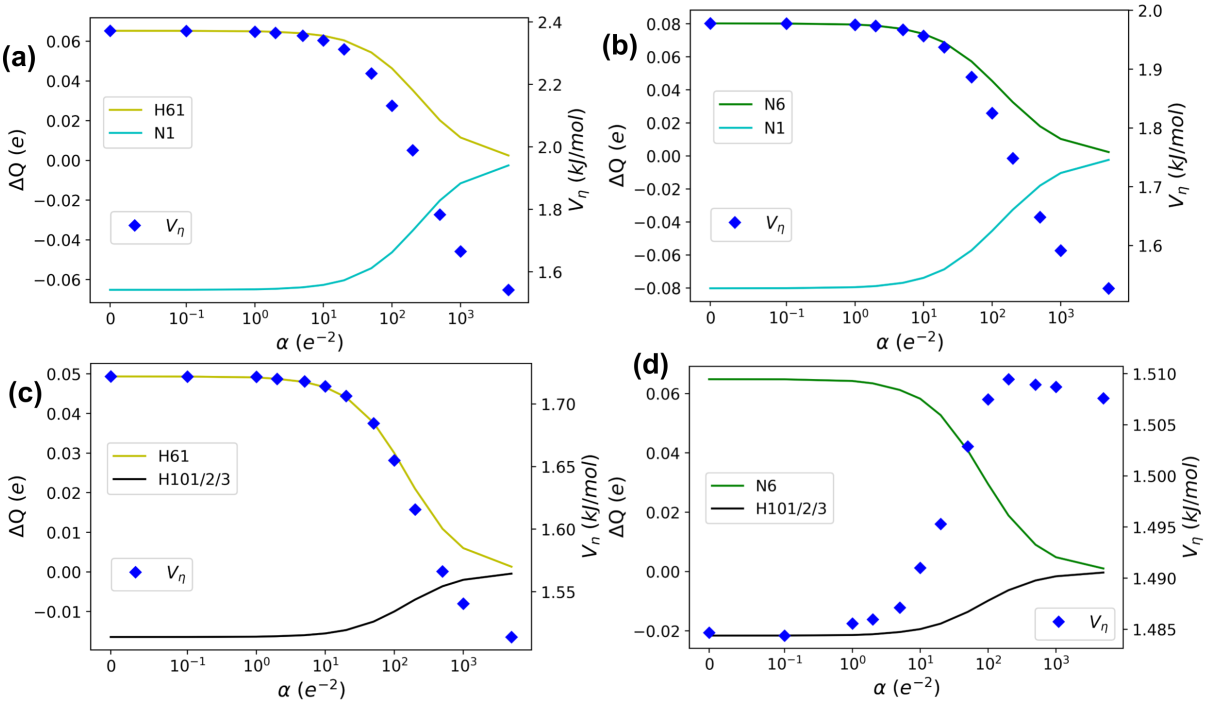


Figure 4.10 Fitted charges and torsional term V_η as a function of α ($\beta = 0$). Horizontal axes are in log scale except for 0–0.1 sections which are linear.

charge. In the two tested cases where N1 was not fitted, the torsional parameter V_η has a smaller dependence on the charges.

Overall, the results suggest that the main contribution of the fitted correction is to increase the stability of Watson–Crick hydrogen bonds by making N1 and H61 more polar and at the same time using the η torsional potential to control the *syn*/*anti* relative population.

4.3.6 Dependence on ionic strength and temperature

In this work, the simulations used in the fitting procedure were performed at a ionic concentration of 0.1 M NaCl, which is a value commonly used in MD simulations. However, the standard condition in which denaturation experiments, including those analyzed in this work, were performed is 1 M NaCl. In order to quantify the dependence of the computed $\Delta\Delta G$s on the ion concentration, we performed further control simulations for a subset of systems at 1 M NaCl. Systems A1, A2 and A4 were chosen in order involve in this checking all possible environments for the methyl group: a nucleoside, where the methyl group is isolated; a duplex with internal m^{6}A, where the methyl group is partly hidden from interactions with ions; a duplex with m^{6}A as a dangling end, where the methyl group is more exposed to interactions with ions; and the corresponding single stranded RNAs, so as to be able to obtain the $\Delta\Delta G$s. For the nucleoside, the methylation was added in both *syn* and *anti* conformations. For the other systems, the methylation was added in the expected conformation, as we did for all other systems (see main text).

Results are summarized in Tables 4.4 and 4.5. The A1 system, that is the single nucletoide in solution, reports a shift in the ΔGs of about 1.2 kJ/mol with respect to 0.1 M cases, for both *syn* and *anti*. As a result, the relative $\Delta\Delta G$ is not affected. For all other system, the ΔGs for the two ionic concentration are in agreement within their statistical error. These results indicate that the fitting is not affected by the discrepancy between the ionic concentration used in computations

	fit_AB						Aduri
[NaCl]	A1 syn	A1 anti	A2 dup	A2 ss	A4 dup	A4 ss	A2 ss
0.1 M	211.23 ± 0.18	217.27 ± 0.19	214.01 ± 0.35	210.40 ± 0.22	208.07 ± 0.17	210.62 ± 0.19	257.54 ± 0.27
1 M	212.33 ± 0.36	218.46 ± 0.25	213.99 ± 0.25	210.53 ± 0.42	208.30 ± 0.33	210.91 ± 0.26	257.56 ± 0.26

Table 4.4 ΔGs computed through alchemical computations and binless WHAM method. Each row corresponds to a different NaCl ionic concentrations used in the simulation. Each column correspond so a different system. In the last column, results obtained with the original Aduri parameters are shown as well for one of the systems, confirming that the mild dependence on ion parameters is independent of the precise partial charges used in the simulation.

[NaCl]	A1	A2	A4
0.1 M	6.04 ± 0.26	3.6 ± 0.4	-2.55 ± 0.25
1 M	6.1 ± 0.4	3.5 ± 0.5	-2.6 ± 0.4
Exp	6.3	1.7 ± 0.9	-2.5 ± 1.2

Table 4.5 $\Delta\Delta G$s computed through alchemical computations and binless WHAM method. The first two rows correspond to different NaCl ionic concentrations used in the simulation, and last row corresponds to the reference experimental values.

and experiments.

In addition, the simulations used in the fitting were performed at a temperature of 300 K, which is a value commonly used in molecular dynamics simulations. However, the experimental free energy differences used in the fitting refer to a temperature of 310 K. In order to quantify the dependence of the computed $\Delta\Delta Gs$ on temperature, we performed further control simulations for systems A1 and A2 at 310 K. Table 4.6 compares ΔGs computed at 300 K or 310 K, whereas Table 4.7 compares the $\Delta\Delta G$s. Differences are compatible within their statistical error. These results suggest that the fitting is not affected by the choice of performing the simulations at 300 K rather than 310 K.

It is also possible to extrapolate experimental and computational ΔGs from 300 K to 310 K by making use of following thermodynamics relationship:

$$\Delta G_{310} = \Delta G_{300} - (310-300)\Delta S \tag{4.17}$$

which can be applied to the calculation of $\Delta\Delta G$s, resulting in:

$$\Delta\Delta G_{310} = \Delta\Delta G_{300} - (310-300)\Delta\Delta S \tag{4.18}$$

or

$$\Delta\Delta\Delta G = -\Delta T \Delta\Delta S \tag{4.19}$$

T	A1 syn	A1 anti	A2 dup	A2 ss
300 K	211.23 ± 0.18	217.27 ± 0.19	214.01 ± 0.35	210.40 ± 0.22
310 K	211.87 ± 0.18	217.65 ± 0.15	214.05 ± 0.24	210.87 ± 0.36

Table 4.6 ΔGs computed through alchemical computations and binless WHAM method, using the fit_AB parametrization. The rows correspond to different temperatures used in the simulations.

T	A1	A2
300 K	6.04 ± 0.26	3.6 ± 0.4
310 K	5.78 ± 0.23	3.18 ± 0.4
Exp	6.3	1.7 ± 0.9

Table 4.7 $\Delta\Delta G$s computed through alchemical computations and binless WHAM method, using fit_AB parametrization. The first two rows correspond to different temperatures used in the simulation, and last row corresponds to the reference experimental values.

	B1	B2	B3	B4	B5
$-\Delta T\Delta\Delta S$ (kJ/mol)	0.6 ± 0.9	-1.0 ± 0.5	-0.5 ± 0.4	0.33 ± 0.20	0.3 ± 0.5

Table 4.8 $\Delta\Delta\Delta G$s for ΔT=10 K computed from experimental entropies

By making use of this relation, we can investigate how $\Delta\Delta G$ would be affected for a change in temperature of 10 K, both for experimental and computational values. As far as the experimental values are concerned, we compute $\Delta T\Delta\Delta S$ for systems B1-B5 by taking the difference between the ΔS measured for the methyated systems [37] and those for the unmethylated systems [122] [123]. Results are reported in Table 4.8. Since we didn't find the experimental errors for the unmethylated systems, we assumed them to be identical to those obtained in the methylated systems. The changes in $\Delta\Delta G$ are small and dominated by their experimental error.

We then computed ΔSs from our simulations by making use of the relationship:

$$\Delta S = -\frac{\Delta G - \Delta U}{T} = -\frac{\Delta G - (\langle U\rangle_{\lambda=1} - \langle U\rangle_{\lambda=0})}{300} \tag{4.20}$$

$\Delta\Delta\Delta G$s for ΔT=10 K are shown in Table 4.9. Statistical errors were computed with blocked bootstrap. Also in this case, changes in $\Delta\Delta G$ are small and dominated by their statistical errors. A recalculation of the χ^2 using $\Delta\Delta G$ extrapolated at 310K returns a value $\chi^2 = 6.03$, which is equivalent to the value reported in Fig. 4d using 300K results ($\chi^2 = 5.71$), confirming that changing temperature does not affect the comparison between simulation and experiment.

4.4 Discussion

In this work, we proposed a protocol to parametrize charges in modified nucleobases using available melting experiments. The approach is applied to m^{6}A and leads to a set of charges that can reproduce a set of 10 independent experimental values. The approach is based on the force-field fitting strategies introduced in earlier works [124, 125, 114], which are here extended with several technical improvements.

	A1	A2	A3	A4	A5
$-\Delta T\Delta\Delta S$ (kJ/mol)	0.1 ± 1.3	-0.57 ± 0.33	0.37 ± 0.36	-0.19 ± 0.36	-0.36 ± 0.22
	B1	**B2**	**B3**	**B4**	**B5**
$-\Delta T\Delta\Delta S$ (kJ/mol)	-0.75 ± 0.33	0.5 ± 0.4	-0.1 ± 0.6	0.0 ± 0.4	0.38 ± 0.39

Table 4.9 $\Delta\Delta\Delta G$s for ΔT=10 K estimated from computations

A first methodological contribution is a formalism that allows alchemical calculations to be used as a reference. Previous works were only using observables computed with a single set of force-field parameters [124, 125, 114, 26]. The method introduced here allows free-energy differences between different sets of parameters to be evaluated and compared with the experiment. This opens the way to the optimization of parameters based on experimentally measured $\Delta\Delta G$s. We based our analysis on optical melting experiments, which are commonly employed in the nucleic acids community [126], but other types of experiments might be considered. In our specific application, only the parameters of one of the two end states were refined, but one could similarly fit parameters for both adenosine and m^{6}A, at the price of increasing the number of parameters and thus the risk of overfitting. A second improvement is that we developed a way to efficiently recompute the total energy of the system using test charges. This is achieved by precomputing the total electrostatic energy of the system with a set of randomly perturbed charges. Given the high cost of electrostatic calculations, this makes the cost of each of the iterations performed during force-field fitting significantly faster and implicitly takes into account combination rules, non-bonded exclusions, and periodicity. These two improvements can be readily integrated into other MD-based force-field optimization strategies.

A limitation of optimizing charges with the introduced procedure is that the statistical efficiency of reweighting is significantly decreased even by small charge perturbations. This implies that simultaneously parametrizing many copies of the same nucleotide, or parametrizing a larger number of charges for the same nucleotide, would be more difficult. In our case, we had to include at most two m^{6}A residues in the same simulation. If more copies of the same reparametrized nucleotide are present in the same system, one might have to design strategies where only a few copies at a time are reparametrized, or follow an iterative procedure where modifications are included in consecutive steps [114]. In this application, this was not necessary.

Overfitting was avoided by using a standard L2 regularization term on the charge increments. This penalty does not depend on the charge location. Importantly, the regularization hyperparameters tune the relative weight of the experimental data and of the reference charges, here taken from Ref. 28, thus allowing to achieve a meaningful set of parameters also in regimes where the number of data points is very limited. It is worth noting that the standard restrained ESP fitting is performed including a restraint that acts as a hyperbolic regularization term [120], which is introduced to keep the absolute values of the obtained charges as small as possible. Our regularization, instead, keeps the resulting charges as close as possible to the initial guess obtained with the restrained ESP procedure [28]. This allows to implicitly include in the fitting the result of the corresponding quantum-mechanical calculation. More effective regularization strategies might be designed based on the molecular dipole, as done in Ref. 127, to minimize the perturbation of the electrostatic potential at a large distance from the molecule. Alternatively, one might directly use as a regularization term the deviation from the quantum-mechanical electrostatic potential at a short distance. In the limit of a large regularization hyperparameter, this would lead to ESP charges [128]. Finally, other regularization criteria might be used [26]. When comparing our procedure with standard ESP charge fitting, it is important to realize that we are aiming to reproduce experimentally observed $\Delta\Delta G$, which are non-linear functions of the energy of each configuration, which in turn is a quadratic function of the charges. These non-linearities make it possible for multiple local minima of the cost function to exist, and could thus make the minimization not reproducible. However, when sufficiently regularized, the fitting procedure results in reproducible charges that depend smoothly on the control parameters. In standard ESP fitting, instead, the electrostatic potential is fitted, thus resulting in a linear fit with a unique solution.

We notice that the parameters of the unmethylated force-field were not modified. This was based on the assumption that the employed set of force-field parameters is already capable to reproduce $\Delta\Delta G$ experiments associated with mutations between non-modified nucleobases [129]. The m^{6}A charge optimization could be easily repeated using another set of initial parameters, and the parameters of non-modified nucleobases might be adjusted as well, although with the caveat discussed above.

Another possible limitation of the employed alchemical simulations is the sufficient sampling of the end states. The duplex is expected to be stable and well structured, so sampling multiple structures should not be necessary. For selected cases, we also explored the possibility to include the unlikely *syn* paired state, which, as expected, gives a negligible contribution to the stability of the duplex. For single strands, instead, we only sampled the *syn* state. More importantly, our simulations were short enough to avoid any significant reconformation of the single strand. Sampling the conformations of flexible, single-stranded RNAs is notoriously difficult [1]. In addition, the generated ensemble might contain artificially stabilized intercalated structures, whose population is known to be overestimated by the RNA and water force-fields adopted here [95, 130]. This would make the correct sampling of the single-stranded state unfeasible. We also notice that the experimental results that we aimed to reproduce were performed on systems designed to have the isolated strands unstructured, to capture the effect of methylation on hybridization. Putting everything together, we conclude that the approximation of a single strand ensemble that does not depart too much from the initial A-form helix is a sensible choice for this specific application.

An important finding of this work is that the parameters of Aduri *et al.* cannot reproduce the *syn*/*anti* balance expected for m^{6}A residues. This balance is extremely important and is related to the mechanism by which m^{6}A modifications modulate duplex stability [36]. This could not be rectified with a straightforward correction of the single torsion involved. The optimized charges, instead, allow the correct *syn*/*anti* balance to be recovered both in paired and unpaired nucleobases, as well as a heterogeneous set of optical melting experiments to be reproduced. Interestingly, the Aduri *et al.* parameters were tested in a recent work [35], with results for system A2 in Table 4.2 consistent with ours and with experiments. However, systems A1 and A3 were not tested, and thus the problem that we observed here could not be identified. Another interesting finding is that the $\Delta\Delta G$ associated with N6 methylation are here predicted to be independent of ion concentration. We are not aware of any experimental validation of this finding, which could be obtained by comparing melting experiments at different ion concentrations. Finally, our results suggest that the relative population of the *syn* excited state in duplexes [68] might significantly depend on the identity of the neighboring nucleotides. The precise hybridization kinetics could thus be quantitatively different for RNAs with different sequences.

A convenient property of our approach is that it does not require changing the functional form of the interaction potential so that new parameters can be readily incorporated into existing MD software. This is not the case if *ad hoc* corrections are employed [25, 26]. In addition, it is worth noting that the charge modifications obtained are very small, and in particular they are smaller than the typical difference between sets of charges derived with slightly different procedures or using different reference conformations. Despite this small difference, the effect on experimental observables is significant. These observations imply that there is still significant space to improve the performance of current force-fields without necessarily modifying the functional form if experimental information is used during training [23].

Using our approach it is possible to dissect the individual contribution of the modified force-field parameters. The main factors playing a role in the change of duplex stability induced by m^{6}A methylation are (a) the penalty for switching to the unfavored *anti* isomer [36], (b) the stabilization induced by hydrophobic shielding of the methyl group against surrounding bases

[131, 132], (c) the impact of partial charges on stacking interactions [131], and (d) the impact on the strength of Watson–Crick hydrogen bonds. Since, on average, experimental $\Delta\Delta G$ for denaturation experiments performed on duplexes are smaller than the *anti* isomer penalty, we could expect that the sum of the other factors has a stabilizing effect on the majority of the considered duplexes. We notice that Aduri charges for N1 and H61, which are involved in Watson–Crick pairings with the complementary uridine, have partial charge absolute values significantly lower compared to the standard adenine parameters (0.28948 vs 0.41150 for H61, -0.675968 vs. -0.76150 for N1). This may lead to a weakening of hydrogen bonds which may cause an overestimation of destabilization induced on duplexes, as we observed in Aduri+tors cases (see Fig. 4.7). The results of our fitting systematically increase the absolute value of H61 and N1 partial charges, hence resulting in a stronger Watson–Crick pairing. At the same time, the torsional term allows reproducing the correct *anti* isomer penalty. Parameters are coupled so that it is necessary to fit them simultaneously to avoid double counting effects.

To the best of our knowledge, this is the first attempt to tune partial charges of a biomolecular force-field based on experiments performed on macromolecular complexes. We expect that this approach could be used in the future to improve the capability of biomolecular force-fields to match experimental observations by exploiting a part of the functional form that has been traditionally derived in a bottom-up fashion. Remarkably, the parameters derived here for m^{6}A allow to properly describe paired and unpaired m^{6}A in both *syn* and *anti* conformation, and thus open the way to the use of molecular simulations to quantitatively investigate the effects of N6 methylations on RNA structural dynamics.

Chapter 5

Alchemical Metadynamics for the N6-methyladenosine

As discussed in the previous Chapter, N6-methyladenosine (m^6A) can exist in two different conformations: *syn* and *anti*. These conformations can be easily distinguished by examining the torsional angle η_6, which is defined by the atoms N1-C6-N6-C10. The barriers associated with the η_6 angle have a significant impact on the kinetics of hybridization [68], and they are expected to be so high that observing a switching between the two m^6A isomers in an unbiased molecular dynamics (MD) simulation within the time scale of nanoseconds to microseconds would be nearly impossible. Therefore, in our previous work described in Chapter 4, we studied this system by separately simulating the *syn* and *anti* conformations, using the alchemical free energy calculation (AFEC) approach with standard Hamiltonian replica exchange (HREX) simulations. Recently, a new method called alchemical metadynamics (AM) was introduced by Hsu *et al.* [67], and its theory is discussed in Section 2.2.4. This method provides an opportunity to obtain the same information as HREX but in a single simulation. Additionally, AM allows us to gather additional information about the isomerization barrier of m^6A, enabling the reconstruction of the free energy profile along the torsional angle η_6.

In this Chapter, we validate the implementation of alchemical metadynamics by Hsu *et al* on the systems A1 and A2 introduced in Chapter 4. These systems involve the alchemical transformation of a standard adenosine (A) into N6-methyladenosine (m^6A) for a single nucleoside in solution and an 8-bp double-stranded RNA, as illustrated in Figure 5.1.

5.1 Methods

All the setups have been described extensively in Chapter 4 and are available on Zenodo (https://zenodo.org/record/6498021). The GROMACS input files are identical to those used in our previous work, except that here the λ ladder is sampled with the Metropolized-Gibbs algorithm with attempted moves spaced with 100 integration steps. For the simulations reported in this work, we used the fit_A parametrization for m^6A discussed in previous Chapter.

We used alchemical metadynamics to flatten the sampling along both the alchemical λ state and along a physical collective variable. For this system, we tested a modified setup where we apply two concurrent metadynamics [57]. The first metadynamics process is one-dimensional and acts only along the alchemical variable. Since the free energy differences along this non-physical variable can be huge, we use a large bias factor ($\gamma = 100$). The second, concurrent, metadynamics process is two-dimensional and acts both on the alchemical variable and on η_{avg}, an averaged torsional angle elaborated in the next section. Since the barriers along η_{avg} are

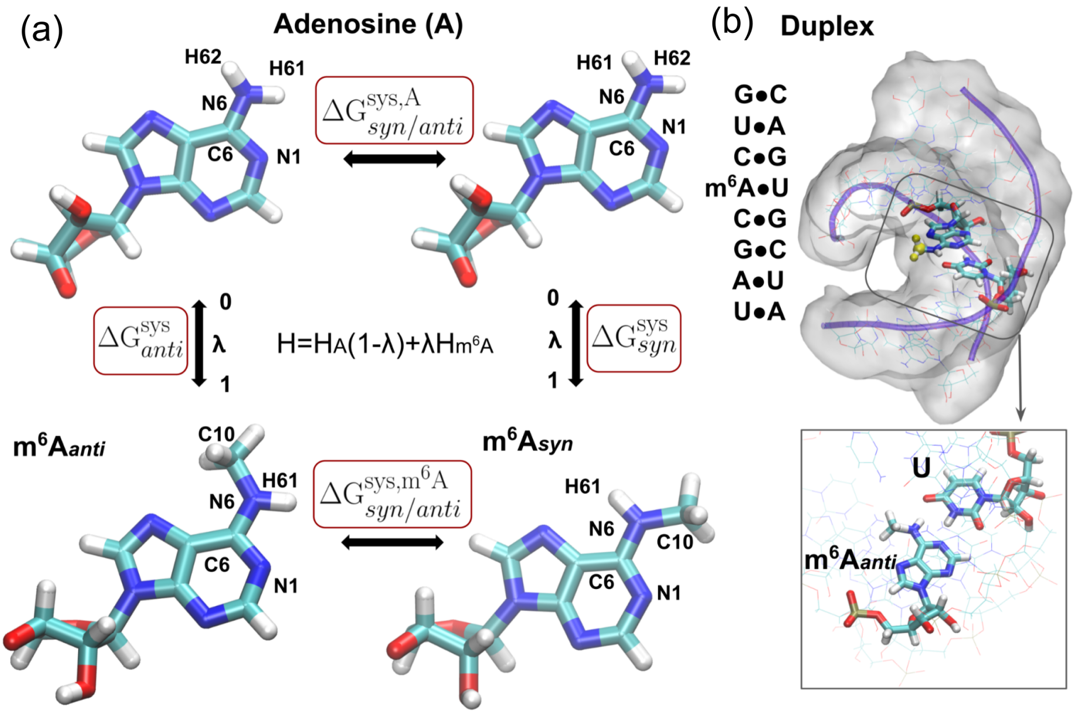

(a)
Adenosine (A)
H62
H61
N6
C6
N1
$\Delta G^{\mathrm{sys,A}}_{syn/anti}$
H61
H62
N6
N1
C6
0
λ
1
$\Delta G^{\mathrm{sys}}_{anti}$
H=HA(1-λ)+λHm6A
$\Delta G^{\mathrm{sys}}_{syn}$
m6Aanti
C10
H61
N6
C6
N1
$\Delta G^{\mathrm{sys,m^6A}}_{syn/anti}$
m6Asyn
H61
N6
C10
C6
N1
(b)
Duplex
G•C
U•A
C•G
m6A•U
C•G
G•C
A•U
U•A
U
m6Aanti

Figure 5.1 (A) The 4 considered states of the alchemical transformation of A into m^{6}A. Isomers are characterized by the value of the torsional angle defined by atoms N1-C6-N6-H62 or N1-C6-N6-C10. The isomers are indistinguishable in the adenine case, so $\Delta G^{\mathrm{sys,A}}_{syn/anti} = 0$. On the other hand, in m^{6}A the position of the methyl group defines the states *anti* and *syn*. The former is the most favored for the paired m^{6}A in a duplex, while the latter is the most favored for the isolated nucleoside. (B) The 8 base-pairs duplex considered in this work, shown in the case of methylated adenosine in *anti* state.

smaller, this second metadynamics is performed with a lower bias factor ($\gamma = 10$). The overall bias potential acting on the system can thus be written as

$$V_{\text{tot}}(\eta_{\text{avg}}, \lambda) = V_1(\lambda) + V_2(\eta_{\text{avg}}, \lambda) \tag{5.1}$$

where V_1 and V_2 are the Gaussian biases added during the one-dimensional metadynamics and the two-dimensional metadynamics, respectively. This combined bias potential can be directly used for reweighting, as discussed above. Notably, by using only two collective variables, a direct reweighting is sufficient in our case. In other words, it is not necessary to include multiple replicas to generate unbiased results, as done in [57] and in Chapter 3.

Metadynamics simulations were run for 60 ns, with Gaussians of initial height 12 kJ/mol, for V_1, and 1.2 kJ/mol, for V_2, deposited every 500 steps. The Gaussian width along the η_{avg} variable was chosen to be 0.35 rad. The 2D free energy surface was computed directly from the bias potentials, while the 1D profile along η was reconstructed using reweighting. Free energy differences and their statistical errors were computed by reweighting a second 160 ns-long simulation where the bias potentials were kept constant. In the case of this calculation, as has also been observed anecdotally in other cases [133], using a static bias resulted in slightly more statistically robust free energy differences.

5.1.0.1 Free energy calculations

For this system, we are interested in calculating the following three relative free energy differences: $\Delta\Delta G^{\text{ns}}_{syn/anti}$, $\Delta\Delta G^{\text{dup}}_{syn/anti}$, and $\Delta\Delta G^{\text{dup/ns}}_{syn+anti}$, where the first two denote the difference in the methylation free energy between the transformation processes that lead to a *syn* or *anti* m^{6}A, in the isolated nucleoside (ns) and in the duplex (dup), respectively. They can be calculated by taking the difference between the free energy differences of interest, namely,

$$\Delta\Delta G^{\text{ns}}_{syn/anti} = \Delta G^{\text{ns}}_{anti} - \Delta G^{\text{ns}}_{syn} \tag{5.2}$$

$$\Delta\Delta G^{\text{dup}}_{syn/anti} = \Delta G^{\text{dup}}_{anti} - \Delta G^{\text{dup}}_{syn} \tag{5.3}$$

The same set of free energy differences (ΔG's in Equation 5.2 and 5.3) can be used to calculate $\Delta\Delta G^{\text{dup/ns}}_{syn+anti}$, the relative methylation free energy between the nucleoside and the duplex systems considering both *syn* and *anti* conformations:

$$\Delta\Delta G^{\text{dup/ns}}_{syn+anti} = \Delta G^{\text{dup}}_{syn+anti} - \Delta G^{\text{ns}}_{syn+anti} \tag{5.4}$$

with

$$\Delta G^{\text{ns}}_{syn+anti} = -\frac{1}{\beta}\ln(\exp(-\Delta G^{\text{ns}}_{syn}) + \exp(-\Delta G^{\text{ns}}_{anti})) \tag{5.5}$$

$$\Delta G^{\text{dup}}_{syn+anti} = -\frac{1}{\beta}\ln(\exp(-\Delta G^{\text{dup}}_{syn}) + \exp(-\Delta G^{\text{dup}}_{anti})) \tag{5.6}$$

In Equations 5.2, 5.3, 5.5 and 5.6, $\Delta G^{\text{ns}}_{syn}$, $\Delta G^{\text{ns}}_{anti}$, $\Delta G^{\text{duplex}}_{syn}$, and $\Delta G^{\text{duplex}}_{anti}$ are the free energy differences of converting adenosine into a *syn* m^{6}A or *anti* m^{6}A in either the isolated form or the duplex, each of which can be calculated from a separate alchemical simulation at fixed rotameric state. For example, in our previous work shown in Chapter 4, four independent Hamiltonian replica exchange simulations were performed, each estimating one of these four values, which combined to give estimates of the three relative free energy differences of interest ($\Delta\Delta G^{\text{ns}}_{syn/anti}$, $\Delta\Delta G^{\text{dup}}_{syn/anti}$, and $\Delta\Delta G^{\text{dup/ns}}_{syn+anti}$).

However, using alchemical metadynamics, we can sample both rotamers in a single simulation methylating the adenosine. Thus, $\Delta G_{syn}^{\mathrm{sys}}$, $\Delta G_{anti}^{\mathrm{sys}}$, with $^{\mathrm{sys}}$ being either $^{\mathrm{ns}}$ or $^{\mathrm{dup}}$, can be directly obtained from a single alchemical metadynamics simulation. Given the access to all metastable states in the alchemical and configurational space, we can calculate free energy differences with more flexibility by considering ratios of partition functions corresponding to different states. For example, with alchemical metadynamics, we can calculate $\Delta\Delta G_{syn/anti}^{\mathrm{ns}}$ and $\Delta\Delta G_{syn/anti}^{\mathrm{dup}}$ as follows, instead of using Equations 5.2 and 5.3:

$$\Delta\Delta G_{syn/anti}^{\mathrm{ns}} = \Delta G_{syn/anti}^{\mathrm{ns,\ m^6A}} = -\frac{1}{\beta}\ln\left(\frac{\sum_{i\in anti} e^{\beta V_{\mathrm{tot}}^{\mathrm{ns}}(\eta_i,\lambda=1)}}{\sum_{i\in syn} e^{\beta V_{\mathrm{tot}}^{\mathrm{ns}}(\eta_i,\lambda=1)}}\right) \tag{5.7}$$

$$\Delta\Delta G_{syn/anti}^{\mathrm{dup}} = \Delta G_{syn/anti}^{\mathrm{dup,\ m^6A}} = -\frac{1}{\beta}\ln\left(\frac{\sum_{i\in anti} e^{\beta V_{\mathrm{tot}}^{\mathrm{dup}}(\eta_i,\lambda=1)}}{\sum_{i\in syn} e^{\beta V_{\mathrm{tot}}^{\mathrm{dup}}(\eta_i,\lambda=1)}}\right) \tag{5.8}$$

$\Delta G_{syn/anti}^{\mathrm{ns,\ m^6A}}$ and $\Delta G_{syn/anti}^{\mathrm{dup,\ m^6A}}$, which are the free energy differences between the two rotamers in the nucleoside and in the duplex, respectively, are not available in Hamiltonian replica exchange but in alchemical metadynamics. Similarly, $\Delta G_{syn+anti}^{\mathrm{ns}}$ and $\Delta G_{syn+anti}^{\mathrm{ns}}$ can be calculated as follows:

$$\Delta G_{syn+anti}^{\mathrm{ns}} = -\frac{1}{\beta}\ln\left(\frac{\sum_{i\in syn+anti} e^{\beta V_{\mathrm{tot}}^{\mathrm{ns}}(\eta_i,\lambda=1)}}{\sum_{i\in syn+anti} e^{\beta V_{\mathrm{tot}}^{\mathrm{ns}}(\eta_i,\lambda=0)}}\right) \tag{5.9}$$

$$\Delta G_{syn+anti}^{\mathrm{dup}} = -\frac{1}{\beta}\ln\left(\frac{\sum_{i\in syn+anti} e^{\beta V_{\mathrm{tot}}^{\mathrm{dup}}(\eta_i,\lambda=1)}}{\sum_{i\in syn+anti} e^{\beta V_{\mathrm{tot}}^{\mathrm{dup}}(\eta_i,\lambda=0)}}\right) \tag{5.10}$$

so that $\Delta\Delta G_{syn+anti}^{\mathrm{dup/ns}}$ can be calculated using Equation 5.4. The goal of our application of the alchemical metadynamics, is to compare the three obtained relative free energy differences ($\Delta\Delta G_{syn/anti}^{\mathrm{ns}}$, $\Delta\Delta G_{syn/anti}^{\mathrm{dup}}$, and $\Delta\Delta G_{syn+anti}^{\mathrm{dup/ns}}$) with the values recovered from Hamiltonian replica exchange reported in Chapter 4.

5.2 Results

5.2.1 Individuating the Optimal Collective Variable

One critical issue in this system is the proper choice of the configurational collective variable. In the first attempt, we used the torsional angle defined as the torsion identified by atoms N1-C6-N6-C10 (see Figure 5.1). This choice was found to be suboptimal. In the production runs, we used as a biased variable a mean torsion obtained by averaging the three torsions identified by atoms N1-C6-N6-C10, N1-C6-N6-H61, and N1-C6-N6-H62. We remind that atoms C10 and H62 may have non-bonded interaction turned *on* or *off* depending on the λ state, but their bonded interactions are always turned *on*. The average was computed as the arctangent of the sine and cosine averages. These three torsions are coupled by an improper torsion present in the original force-field parametrization, which maintains the group C10, N6, H61, and H62 planar, but this torsion is insufficiently stiff to maintain the consistency between the three torsions when enforcing the barrier crossing. When biasing the average, a diffusive behavior of the biased CV was obtained (Figure 5.2). Specifically, with the torsions N1-C6-N6-C10 (η_6 or η_{C10}), N1-C6-N6-H61 (η_{H61}), and N1-C6-N6-H62 (η_{H62}), the average is computed as

$$\eta_{\mathrm{avg}} = \mathrm{atan2}\left(\frac{\sin(\eta_{\mathrm{C10}})+\sin(\eta_{\mathrm{H61}}+\pi)+\sin(\eta_{\mathrm{H62}})}{3}, \frac{\cos(\eta_{\mathrm{C10}})+\cos(\eta_{\mathrm{H61}}+\pi)+\cos(\eta_{\mathrm{H62}})}{3}\right) \tag{5.11}$$

Where atan2 is the two-argument arctan function, defined as the angle between the positive x-axis and the vector (x,y); it is equal to $\arctan(y/x)$ when $x > 0$, but involves corrections of $\pm\pi$ when $x \leq 0$. We also note that η_{H61} must be shifted by π rad when taking the average. Fig. 5.2 aims to summarize the issues encountered when using N1-C6-N6-C10 as a CV and how they can be solved by switching to the averaged torsional η_{avg}. In panel (a) it is shown the value of torsional angle N1-C6-N6-C10 as a function of time, when the same torsion was used as CV, in a simulation performed at dynamic bias potential. In the 160 ns of simulations, the system only switched once from *syn* to the *anti* state after about 8 ns and then back to *syn* after about 60 ns. When using the averaged torsion as a CV instead Fig. 5.2b), the system became diffusive on N1-C6-N6-C10 after a few ns. Fig. 5.2c shows N1-C6-N6-C10 versus N1-C6-N6-H62 when N1-C6-N6-C10 was used as CV, while Fig. 5.2d shows the same but in the case of averaged torsion used as CV. The results shown here demonstrate that the improper torsion is not sufficiently stiff to maintain the consistency between the three torsions when enforcing the barrier crossing. As a consequence, the single N1-C6-N6-C10 torsion is not an optimal CV to allow a proper sampling of the torsional space.

5.2.2 Free Energy Estimates

The free energy profile along the λ state index is computed using reweighting and reported in Figure 5.3a. The significant difference observed is non-physical and depends on the relative definitions of the A and m^{6}A force-field parameters. Figure 5.3b shows the 2D surface as a function of the λ state index and the averaged torsional angle η_{avg}, which is computed using the usual relationship between bias and free energy [54], and then subtracting the Boltzmann-averaged free energy along the λ state index. We notice that the residual dependence of the free energy on λ depends on the fact that barriers on η_{avg} change when λ is changed. The profiles along η_{avg} were computed using the relationship between the bias and the free energy and are shown in Figure 5.3c. Notably, this approach allows free energy profiles along the biased variable to be obtained simultaneously with alchemical differences. These profiles show that the *syn* conformation (central basin) is favored in the m^{6}A nucleoside, whereas the *anti* conformation (lateral basins) is favored in the duplex. The final $\Delta\Delta G$'s, which represent the amount by which the methylation disfavors the duplex, are consistent with those reported in Chapter 3 within the respective statistical errors (Figure 5.3d). Importantly, even though the exploration of λ is guaranteed by the one-dimensional metadynamics, the inclusion of λ in the two-dimensional metadynamics allows to effectively reconstruct free energies along η_{avg} that are depending on λ.

5.2.3 Comparison of methylation free energy calculations with dynamic and static biases

In order to shed lights on the possible limitations of metadynamics, we show in this section the free energy calculations with a dynamic bias for the nucleoside and duplex systems. These calculations are compared to the free energy differences computed with static bias presented in the previous section. Specifically, simulations at dynamic bias were elongated up to 160 ns. For

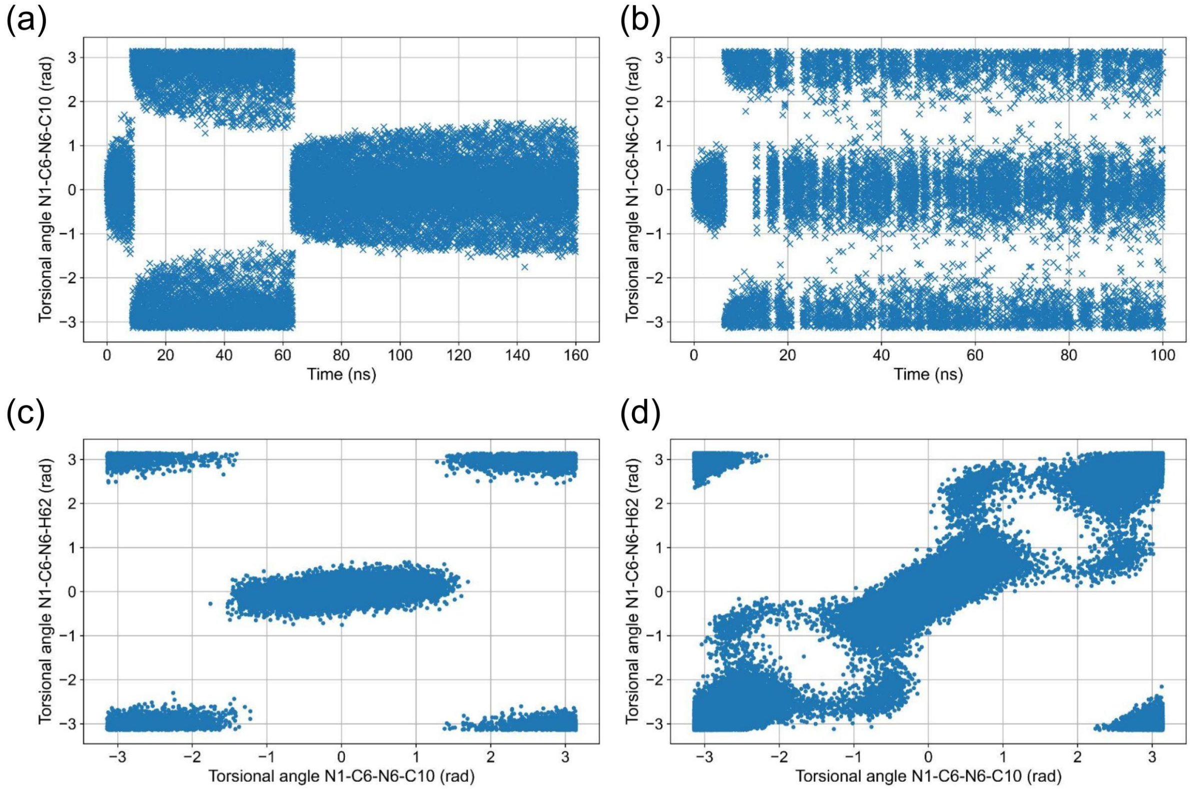


Figure 5.2 (a) The torsional angle N1-C6-N6-C10 as a function of time when the same torsion was used as CV. (b) The torsional angle N1-C6-N6-C10 as a function of time when an averaged torsion between N1-C6-N6-C10, N1-C6-N6-H62, and N1-C6-N6-H61 ($+\pi$) was used as biasing collective variable. (c) N1-C6-N6-C10 versus N1-C6-N6-H62 when N1-C6-N6-C10 was used as CV (d) N1-C6-N6-C10 versus N1-C6-N6-H62 when the averaged torsion was used as CV

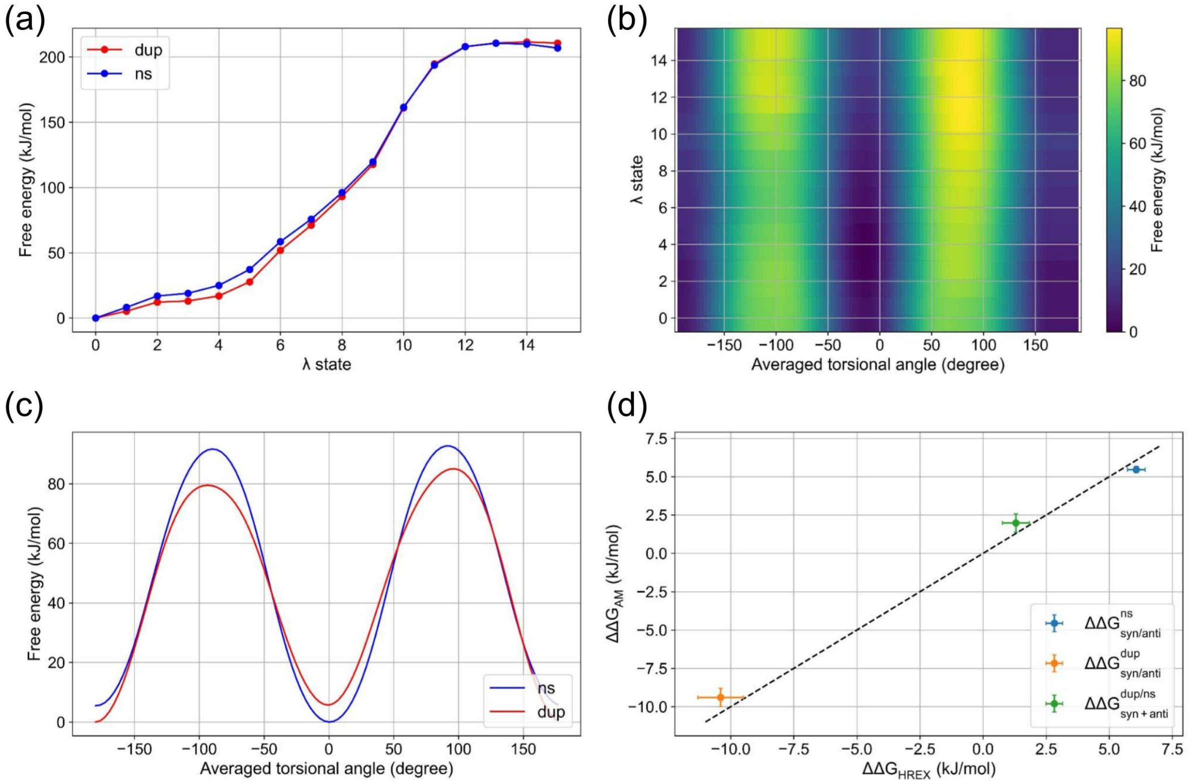


Figure 5.3 (a) The free energy profile along the state index for the RNA duplex and for the isolated nucleoside. (b) Residual free energy surface along the state index and the averaged torsional angle for the RNA duplex (c) The free energy computed as a function of η_{avg} at fixed $\lambda = 1$, both for the RNA duplex (red) and for the m^{6}A nucleoside (blue) (d) Comparison of $\Delta\Delta G$ obtained with alchemical metadynamics (AM) and with Hamiltonian replica exchange (HREX) from Piomponi *et al.*, 2022 [113], with their respective statistical errors.

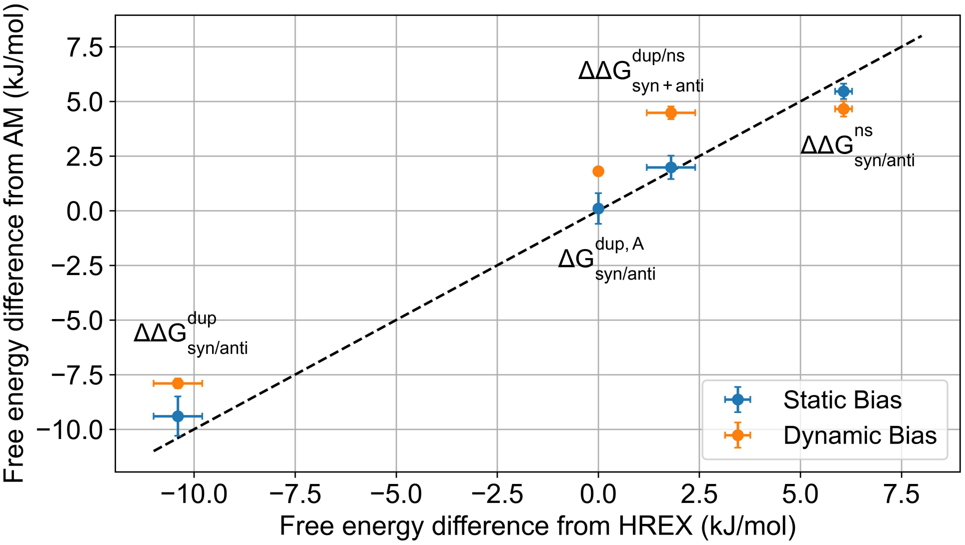


Figure 5.4 Comparison of free energy differences computed in Chapter 3 with Hamiltonian replica exchange (HREX) and $\Delta\Delta G$ computed with alchemical metadynamics (AM) in this work, for two cases: (1) static bias and (2) dynamic bias.

analysis, the first 60 ns were discarded, and the bias averaged over the remaining 100 ns was used to compute weights [71] [70] . Different numbers of blocks ranging from 2 to 1000 were used to construct histograms in block bootstrapping (200 iterations) and the largest uncertainty is reported.

Fig. 5.4 shows that with dynamic bias, the free energy estimates are more precise (lower statistical errors). This is most likely attributable to the fact that the sampling in the CV space is more diffusive in these systems with dynamically updated weights. However, free energy estimates computed with dynamic bias are less accurate, i.e., they differ more from those obtained with Hamiltonian replica exchange (HREX).

To further demonstrate the lower accuracy of the dynamic bias computation, the free energy difference ($\Delta G^{\text{dup, A}}_{syn/anti}$) between the two conformations of adenosine shown in Figure 5.1a is calculated. In our previous work [113] this value was assumed to be 0 because of the symmetry of the hydrogen atoms H61 and H62. Also, the HREX used does not have access to the free energy landscape along the biased torsion, so the relative error for $\Delta G^{\text{dup, A}}_{syn/anti}$ is not given for the HREX case in fig 5.4. In alchemical metadynamics, $\Delta G^{\text{dup, A}}_{anti/syn}$ is calculated as follows:

$$\Delta G^{\text{dup, A}}_{syn/anti} = -\frac{1}{\beta}\ln\left(\frac{\sum_{i\in anti} e^{\beta V^{\text{dup}}_{\text{tot}}(\eta_i,\lambda=0)}}{\sum_{i\in syn} e^{\beta V^{\text{dup}}_{\text{tot}}(\eta_i,\lambda=0)}}\right) \tag{5.12}$$

For most systems, the general understanding is that using plain metadynamics, i.e., analyzing the dynamically biased simulation, is better than doing the two step procedure used here [115]. The result is likely system-dependent and related to the fact that even without a dynamic bias we can see many transitions, thus a reasonable statistical error. In this way, we are clearly in a regime where fewer transitions at equilibrium are a safer estimate.

5.3 Conclusion

The alchemical metadynamics methods expands the configurationally-defined sampling space allowed in traditional metadynamics with an additional alchemical sampling direction. With the configurational bias, alchemical metadynamics encourages the system to escape from configurational metastable subspaces that could have easily trapped the system. It retains the advantages of traditional alchemical free energy methods, but also enables higher flexibility in sampling rough free energy surfaces. Applying the methods on our m^{6}A systems, we demonstrated that 2D alchemical metadynamics eliminated the need to perform multiple Hamiltonian replica exchange simulations to estimate the relative methylation free energy, and simultaneously allow to reconstruct the free energy profile along the biased torsional angle. The alchemical simulation of conversion from A to m^{6}A is an interesting physical example because it shows that alchemical metadynamics gives simultaneous access to free energy barriers for both the two end systems. While this result could have been obtained by performing two separate metadynamics simulations, being able to use a single simulation has substantial advantages. First, it ensures that other possibly slow degrees of freedom are sampled consistently in the two end states, making differential results more reliable. For instance, if the isomerization barrier were affected by binding with another molecule present in the simulation box, the dynamics of λ would have ensured binding to be equally represented in the A and m^{6}A states. Second, in cases where the conformational transitions are better described by the physical CVs in one of the states with respect to the other state, thus resulting in more transitions in one of the end states when compared to the other, having a single simulation would enable the ensemble of the slower state to benefit from the enhanced ergodicity in the faster state. These benefits could also be obtained by combining metadynamics with Hamiltonian replica exchange along the alchemical variable, however, at the price of higher computational cost and less flexibility in the setup. The combination of one-dimensional and two-dimensional bias potentials allows for simultaneous (a) flattening of the large artificial free energy difference along the alchemical variable and (b) effective compensation of the torsional barriers, considering the fact that the precise profiles depend on the alchemical variable. The two potentials can be constructed using different bias factor coefficients so as to optimize their capability to explore the two profiles. This idea might also be exploited in different contexts, whenever one wants to simultaneously facilitate transitions over a large free energy barrier (e.g., a chemical reaction) and, at the same time, smooth residual barriers on softer degrees of freedom. This is at variance with the RECT method by Gil-Ley *et al*, [57] discussed in section 2.2.2 , where a large number of collective variables were concurrently biased, thus requiring a replica ladder to obtain unbiased populations. A similar issue occurs when simultaneously biasing the total energy of a solvated system and solute-dependent CVs. In this case, indeed, two separate metadynamics, possibly with different bias factors, can be applied fruitfully. This was done, for instance, in the work by Deighan *et al.* [134], though in a sequential rather than self-consistent procedure. The protocol is also related to the one proposed by Chipot and Lelièvre [135], although it is here applied in (a) metadynamics context and (b) combining potentials in 1D and 2D with the alchemical CV shared among the two biases. This problem might also be tackled using global tempering methods, where a flat histogram is reached on all the biased CV [136].

Chapter 6

Molecular simulations to investigate the impact of N6-methyation in RNA Recognition: Improving Accuracy and Precision of free energy of binding estimation

In the preceding Chapters, our investigation revolved around understanding the influence of N6-methylation of adenosine on RNA structure, with particular focus on the destabilization of double strands. We refined the m^{6}A force-fields to enhance the capability of MD simulations in replicating denaturation experiments and accurately representing the populations of *syn/anti* isomers. Such a force-field can be considered much more reliable than alternative parametrizations that have not been validated against experiments yet, at least for reproducing impact of N6-methylation on RNA structural dynamics. Indeed, even if the destabilization induced on duplexes is generally low, it has been proven that m^{6}A can have significant impact on the structural dynamic of specific systems. For example, Jones *et al* showed how the N6-methylation of adenosine favors rearrangement of nucleotides of an RNA hairpin tetraloop [137]. Despite this, a substantial body of literature suggests that the primary role of m^{6}A in nature does not seem to be centered on its impact on RNA structural dynamics, but rather on augmenting RNA recognition by proteins known as reader proteins. As a consequence, it is crucial to validate, and if necessary to further refine, the m^{6}A parametrization against experiments which report the impact of N6-methylation on the free energy of bindings (FEB) between RNA and reader proteins. Among these m^{6}A readers, the YT521-B (YTH) family of proteins stands out as the most prominent and extensively examined [138] [139] [140]. In this Chapter, we will initially demonstrate that alchemical free energy calculations, in conjunction with the fit_A force-field developed in Chapter 4, are incapable of replicating the stabilization induced by N6-methylation on the FEB between RNA and a YHT m^{6}A reader protein, as predicted by titration calorimetry experiments [141]. Subsequently, we will illustrate how we can achieve better agreement with experimental results by further refining the m^{6}A force-field. This refinement involves expanding the dataset considered in Chapter 4 and improving the precision of the alchemical calculations by enhancing the exploration of various hydration states within the YHT-RNA binding pocket.

6.1 Effects of N6-methylation on RNA recognition in the YHT domain - A background overview

The role of m^{6}A in recognition has been extensively examined in recent years for the YTH domain of the YTHDC1 protein, for which several structures have been solved and deposited in the PDB, for different types of oligonucleotides [141] [110] [41]. All these structures show how m^{6}A is recognized by being captured in an aromatic cage, with the flanking nucleotides laying in the RNA-protein surface. In recent years, Molecular Dynamics have been used in several works to investigate the binding mechanism by which m^{6}A is recruited by the protein [112] [41] [110] [111] [42]. All these works show how m^{6}A and the amino acids residues forming the aromatic cage favor the formation of a stable hydrogen-bond networks which is maintained all over the simulations. The first attempt to perform MD on a RNA-protein complex with m^{6}A is the work of Li *et al* [41], where simulations were performed using the CHARMM force-field including the m^{6}A parameters developed by Xu *et al* [29]. In this work, unbiased MD starting from crystal structures was perfomed on the complex for a 5 nucleotides RNA singe strand (5$'$-GGm^6ACU-3$'$). Results show the importance of two tryptophan residues in the binding pocket, one displaying stacking interaction with m^{6}A and the other stabilizing the m^{6}A methyl group. Unbiased simulations also demonstrated high flexibility of the guanosine residues located at the 5$'$ direction of m^{6}A, when compared with the more rigid residues in the 3$'$ direction, in accordance with experiments. The role of the flanking nucleotides in binding was investigated by performing simulations and experiment for the variants 5$'$-Gm^6AUC-3$'$ and 5$'$-GGm^6AC-3$'$. Simulations for the single strands in bulk water demonstrate as the presence of the first guanosine and the methylation of the adenosine increases the probability to switch the strand conformation to the bound-like conformation in which m^{6}A is solvent-exposed. A continuation of this work is reported in Ref. [110]. Here, alchemical transformations were performed in order to compute the difference in free energy of binding for methylated and unmethylated cases. To this purpose, a unidirectional thermodynamic integration was performed, starting from m^{6}A and transforming it into standard A. The comparison of results obtained in the complex with those in an isolated nucleoside resulted in the prediction of a stabilization of the complex induced by the methylation slightly overestimated with respect to reference experimental data [141]. This work also addressed the role of water molecules in the binding site, reporting alchemical annihilation of a molecule involved in a water-bridge interaction that stabilizes the bound complex. A similar work was published by Krepl *et al* [112] for a complex with a 6 nucleotides strand (5$'$-CGm^6ACAC-3$'$). Here, the AMBER force-field was used, and parameters for m^{6}A were derived specifically for this work using the standard protocol. In this work the authors accurately investigated the role of hydration in the binding mechanism. They first noticed that in the unbiased simulation of the complex with unmethylated adenosine, a water molecule was sometimes entering and leaving the binding pocket, occupying a position that would be occupied by the methyl group in the methylated simulation. The stabilization of the complex associated with N6 methylation was attributed not only to the hydrophobic interactions but also to the capability of m^{6}A to displace this water molecule. Thermodynamic calculations were performed similarly to the previously discussed work [110], and showed that the slightly different results are obtained depending on the details of the simulation protocol. Similarly to Ref. [110], the stabilization of the complex was overestimated when compared to experiment. In their interpretation, the binding in the unmethylated case is disfavored not only because the bound state is energetically disfavored with respect to the methylated case, but also because for the unmethylated adenosine it is more difficult to displace water molecules from the binding pocket and allow binding. This interpretation was corroborated by free energy of solvation estimation which gave higher absolute value for adenine compared to m^{6}A. Based on these assumptions, the authors also tried to performed more accurate thermodynamic integration to compute $\Delta\Delta Gs$ in the free energy of binding, with respect to the previous work by Li *et al* [110]. In here, they performed both a forward transformation starting from m^{6}A bound in the complex, and a backward transformation starting from the unmethylated state with a water molecule present in the pocket.

The work presented in this chapter aims to build upon the research conducted by Krepl *et al.* [112]. Our first objective is to further investigate the influence of water displacement in/out of the aromatic cage on the estimation of the binding free energies. Furthermore, we propose a combination of our alchemical free energy calculation protocol for m^{6}A (refer to Section 2.2.3) with metadynamics, which aims to give more accurate estimations of free energies difference by sampling a variety of possible conformations of the binding pocket with respect to hydration. This metadynamics approach will enhance the displacement of water molecules both into and out of the YHT binding pocket. In addition, we explore the effects of the m^{6}A force-field on FEB estimation. This involves expanding upon the fitting procedure discussed in Chapter 4 by incorporating an updated training dataset. This dataset comprehensively accounts for the stabilization induced by m^{6}A on the YHT-RNA binding free energy.

6.2 Methods

Starting structures for MD simulations were taken from [141] (PDB ID: 2MTV) and equilibrated using the pmemd.MPI implementation of AMBER, following the identical procedure used in [112]. To be consistent with [112], we used their same parametrization including the SPC\E water model [142], with the only difference being the m^{6}A force-field. Indeed, we decided to use the fit_A force-field derived in Chapter 4, being the only available parametrization for m^{6}A that have been validated against a large experimental data set. As in [112], we also made use of HBfix potentials to increase the stability of the native A5(OP1)/LYS18(NZ) and C6(OP1)/LYS129(NZ) interactions in all YHT-RNA complex simulations.

For the production runs, we used a modified version GROMACS 2020.3 [84] which also implements the stochastic cell rescaling barostat [92]. We prepared a total of 11 systems:

- YHT-RNA (5′- CGACAC-3′) - Used to test metadynamics
- YHT-RNA (5′- CGm^{6}ACAC-3′) - Used for HREX on charges
- YHT-hybridRNA (5′- CG(A-to-m^{6}A)CAC-3′) - Used for AFEC
- ssRNA (5′- CGm^{6}ACAC-3′) - Used for HREX on charges
- ss-hybridRNA (5′- CG(A-to-m^{6}A)CAC-3′) - Used for AFEC
- 3xYHT-RNA (5′- CGACAC-3′) + alchemical water (SPC\E, TIP3P and OPC)
- 3xBulk water + alchemical water (SPC\E, TIP3P and OPC)

For the A-to-m^{6}A alchemical transformations, the procedure explained in 2.2.3 was used. For each replica, the systems were once again energy minimized and subjected to a multi-step equilibration procedure: 100 ps of thermalization to 300 K in the NVT ensemble was conducted through the stochastic dynamics integrator (i.e., Langevin dynamics) [89], and other 100 ps were run in the NPT ensemble simulations using the Parrinello–Rahman barostat [90].

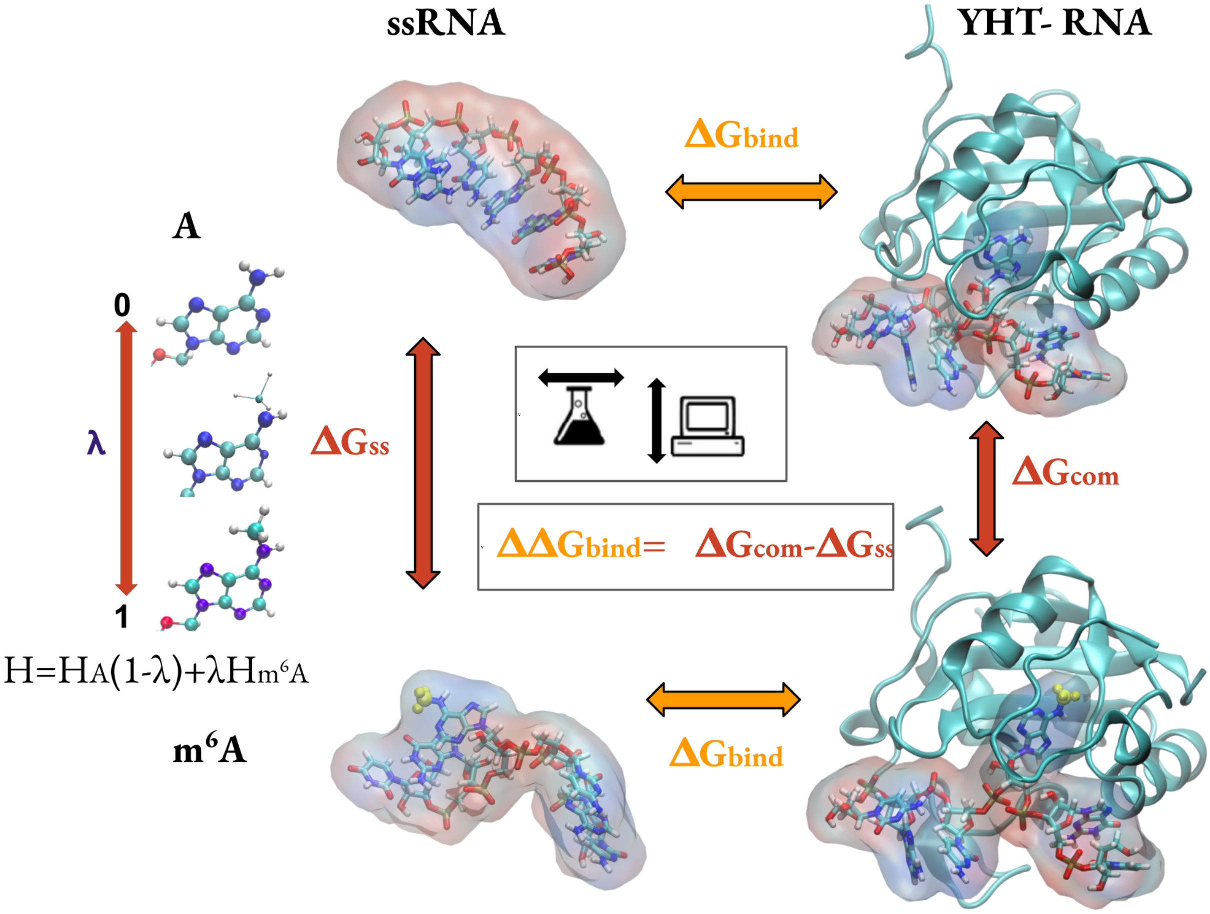
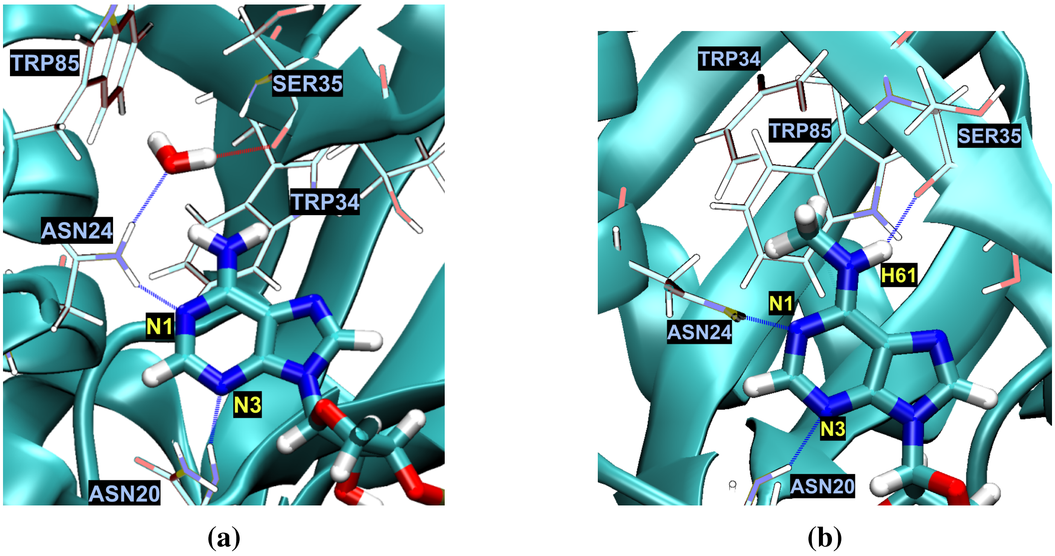


Figure 6.1 Thermodynamic cycle used to compute impact of m^{6}A methylation on the FEB of the YHT-RNA complex. The relative free-energy change due to the modification can be estimated as the $\Delta\Delta G$ between AFECs performed on the complex and on the single strand RNA in solution. This quantity can be directly compared to the difference in FEB ($\Delta\Delta G_{bind}$), which was measured experimentally by Theler *et al* [141].



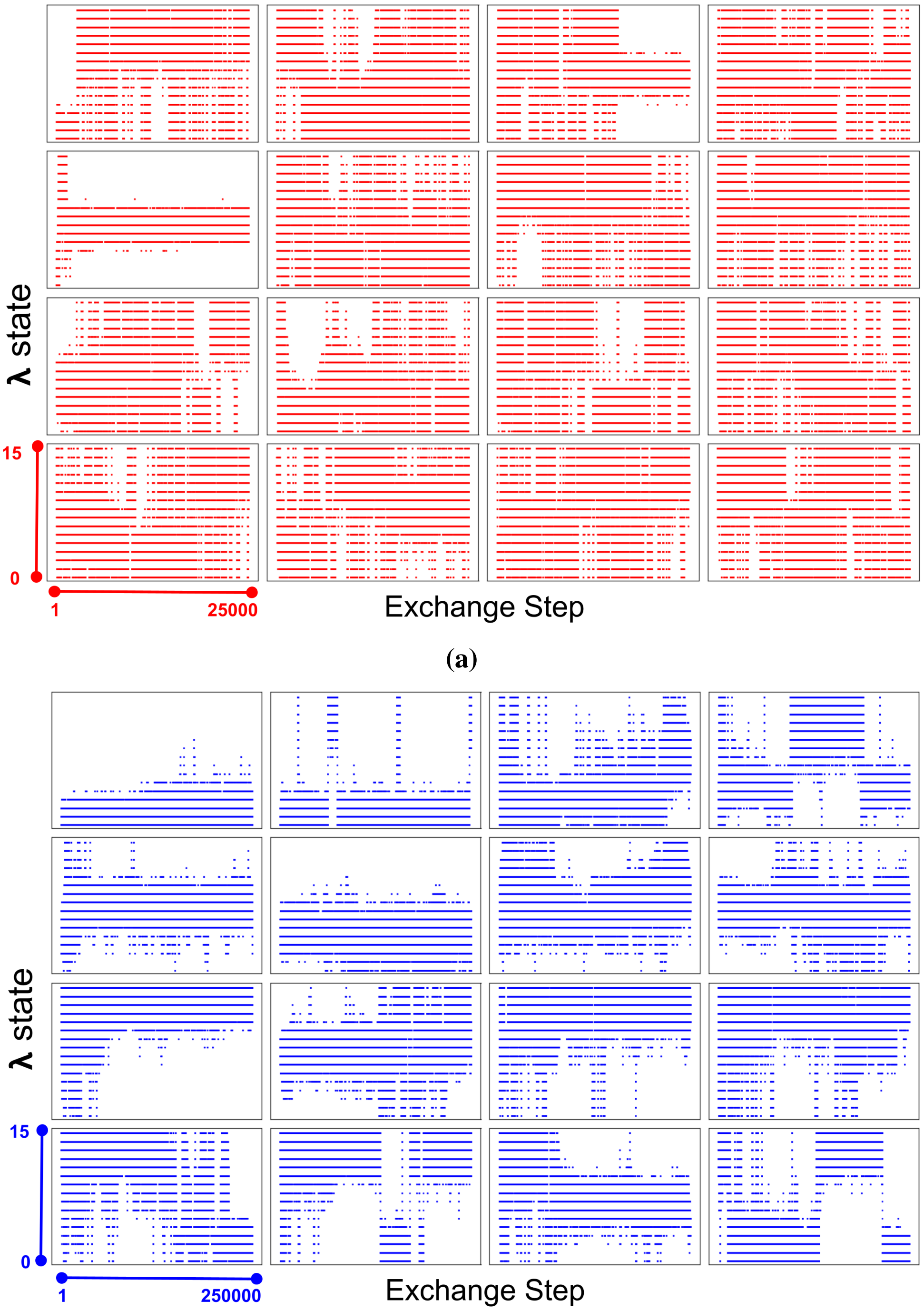
(a)

(b)

Figure 6.2 Snapshots from AFEC+WT-metaD simulation representing the unmethylated adenosine coordinated with a water molecule (a), or the methylated adenosine in the de-hydrated binding pocket (b).

6.3 Results

Alchemical free energy computation (AFEC) for the A-to-m^{6}A transformation (see 2.2.3) were performed for the YHT-RNA system in order to compute $\Delta\Delta G_{bind}$, that is the impact of N6-methylation on the YHT-RNA free energy of binding (see fig 6.1). For the protein-RNA complex, we performed simulations of 10 ns per replica. For the ssRNA double strands, we performed instead simulations of 20 ns per replica. Interestingly, in the HREX scheme we could observe transition probabilities on average significatively higher in the YHT-RNA complex with respect the ssRNA case. This is probably due to the fact that in this transformation the methyl group is appearing/disappearing in the binding pocket where there is no water molecules causing steric clashes. Figures 6.3a shows exploration of the λ ladder for each of the 16 independent demuxed trajectories. All of them can explore the whole ladder, with only the fifth trajectory remaining stuck in the intermediate λ region for most of the time. The ΔGs obtained are shown in table 6.1. The $\Delta\Delta G_{bind}$ result to be 22.1 ± 0.8 kJ/mol (5.3 ± 0.2 kcal/mol), corresponding to significant overestimation of the experimental value (9.9 kJ/mol or 2.3 kcal/mol), even larger than the estimation reported in [112] (18.0 kJ/mol or 4.3 kcal/mol), where a different parametrization for m^{6}A was used. Based on our knowledge, we expect that the overestimation of $\Delta\Delta G_{bind}$ is mainly caused by two factors: inaccuracy of the force-fields and limited sampling. In Krepl parametrization, the negative partial charges of nitrogens N1 and N3 have smaller absolute values with respect to the fit_A counterparts (see Table 1 in Appendix). Both N1 and N3 forms hydrogen bonds in the binding pocket (see Figure 6.2), which is stronger in the fit_A case, possibly explaining why we observe a larger stabilization of the FEB induced by the N6-methylation. As for the limited sampling issue, as suggested previously by [112], a factor which could impact the precision of $\Delta\Delta G_{bind}$ calculation is the role of hydration in the binding pocket. In our AFEC computation we never observe water molecules entering the binding pocket, but a variety of different hydrated state of the binding pocket may exist. In the following we investigate how the role of hydration in the binding pocket could affect the $\Delta\Delta G_{bind}$.



(a)

(b)

Figure 6.3 Each plot correspond to different continuos demuxed trajectories for the AFEC (a), and for the AFEC+WT-MetaD (b) simulations of the YHT-RNA complex. The Y axis reports the λ state at any exchange step in the HREX scheme. Simulations in (b) are 10 times longer than in (a).

	fit_A		**fit5_AC**	
	ΔG	$\Delta\Delta G$	ΔG	$\Delta\Delta G$
ssRNA	205.2 ± 0.7	0	237.0 ± 0.6	0
YHT-RNA	183.1 ± 0.4	22.1 ± 0.8	-	-
YHT-RNA + dynamic bias	186.6 ± 0.8	18.6 ± 1.1	221.0 ± 1.1	16.0 ± 1.3
YHT-RNA + static bias	185.4 ± 1.3	19.7 ± 1.5	224.6 ± 1.0	12.4 ± 1.2

Table 6.1 ΔGs and $\Delta\Delta G$s ($\Delta G_{ss} - \Delta G_i$) computed through alchemical computations, with different parametrizations, reported in kJ/mol.

6.3.1 Alchemical free energies for water insertion

As suggested by Krepl *et al* al [112], one of the factors contributing to the overestimation of $\Delta\Delta G_{bind}$ could be the omission of scenarios where a water molecule is situated inside the binding pocket and coordinates with atom H62. Such a configuration is plausible at $\lambda = 0$ but becomes improbable at $\lambda = 1$ due to steric hindrance with the methyl group occupying that space. However, in the plain MD simulations conducted by Krepl *et al*, it was observed that a water molecule stays inside the binding pocket only 10% of the time in the absence of methylation. The fact that this hydrated state of the unmethylated complex is energetically disfavored compared to the reference state, suggests that hydration plays a minor role in $\Delta\Delta G_{bind}$. To investigate this further, we conducted AFEC simulations involving the annihilation of a water molecule within the binding pocket. This computation aims to assess:

$$\Delta\Delta G^{alc-H2O} = \Delta G^{alc-H2O}_{bulk} - \Delta G^{alc-H2O}_{YHT-RNA} \tag{6.1}$$

where $\Delta G^{alc-H2O}_{bulk}$ is associated to the annihilation of a water molecule (alc-H2O) in the bulk and $\Delta G^{alc-H2O}_{YHT-RNA}$ to the annihilation of alc-H2O in the binding pocket, more precisely in the H62-coordinated position. $\Delta\Delta G^{alc-H2O}$ correspond to the free energy difference between the hydrated and non-hydrated unmethylated YHT-RNA complex, and its impact on ΔG_{com} can be written as follows:

$$\Delta G_{com} = -k_B T ln(e^{-\beta\Delta G^{no-H2O}_{com}} + e^{-\beta\Delta G^{H2O}_{com}}) = \Delta G^{no-H2O}_{com} - k_B T ln(1 + e^{-\beta\Delta\Delta G^{alc-H2O}}) \tag{6.2}$$

assuming that

$$\Delta G^{H2O}_{com} = \Delta G^{no-H2O}_{com} + \Delta\Delta G^{alc-H2O} \tag{6.3}$$

This assumption is based on the fact that the hydrated state is not negligible only for the $\lambda = 0$ state. In these calculations we also estimated the impact of the water model in the result.

6.3.1.1 Methods

We computed $\Delta\Delta G^{alc-H2O}$ for three different water models: SPC\E [142], TIP3P [85], and OPC [143]. As a consequence, we performed a total of 6 AFEC simulations involving the annihilation of a single water molecule, which we will refer to as alc-H2O. For three different water parametrizations we performed the alchemical computation both in the YHT-RNA complex and in bulk. In the bulk simulations, all water molecules were parametrized based on the chosen model. In the YHT-RNA instead, we reparametrized only the alc-H2O, whereas the rest of the solvent was maintained with the SPC\E model used in the rest of the work. This choice

	$\Delta G^{alc-H2O}_{bulk}$	$\Delta G^{alc-H2O}_{com}$	$\Delta\Delta G^{alc-H2O}$	$\Delta\Delta\Delta G_{bind}$
SPC\E	29.51 ± 0.08	21.5 ± 0.7	8.0 ± 0.8	-0.10
TIP3P	25.48 ± 0.07	15.1 ± 1.2	10.4 ± 1.2	-0.038
OPC	33.7 ± 0.5	22.8 ± 1.0	10.9 ± 1.2	-0.031

Table 6.2 Free energies differences computed through alchemical computations, for different water models, reported in kJ/mol.

was done in order to avoid re-preparation and re-equilibration of the YHT-RNA system. Since the alc-H2O interacts exclusively with the RNA and the YHT protein, the parametrization of the solvent is not expected to impact these calculations. The simulations are 10 ns per replica long. We used 16 replica with λ spacing: [0.00 0.01 0.03 0.05 0.10 0.20 0.35 0.45 0.55 0.65 0.80 0.90 0.95 0.97 0.99 1.00], except for ΔG_{com} with OPC and TIP3P water models where we used 8 replica, with λ spacing [0.00 0.03 0.06 0.13 0.30 0.50 0.75 1.00]. In $\lambda = 0$, alc-H2O interactions are switched on, *viceversa* switched off for $\lambda = 1$. The potential interpolation is the same described in 2.2.3.

During the alc-H2O AFEC in the binding pocket, a restraint was used to avoid the alc-H2O leaves the coordination spot, in the form:

$$R(x) = K\theta(x-0.2)(x-0.2)^2 \tag{6.4}$$

where $K = 400$ kJ/mol and θ is the step function. This restraint was applied on a RMSD computed on the coordinates of alc-H2O and A3 nucleobase with respect to a structure extracted from MD simulations (the frame was taken from biased simulations we will introduce in next section). Free energies were computed using WHAM, including in the energies the bias due to the restraint on the alchemical water.

6.3.1.2 Results

All computed free energies for the alchemical transformation of water are detailed in Table 6.2. Notably, all $\Delta G^{alc-H2O}$s values are positive, indicating a disfavoring of the hydrated state, which aligns with our expectations. Intriguingly, when using the TIP3P and OPC water models, the hydrated state becomes even more disfavored, resulting in a further marginal impact on $\Delta\Delta G^{bind}$. We can quantify the correction to $\Delta\Delta G^{bind}$, in relation to the estimates obtained in the previous section that did not account for hydration effects, as $\Delta\Delta\Delta G_{bind} = -k_B T ln(1 + e^{-\beta\Delta\Delta G^{alc-H2O}})$. These corrections are presented in the fourth column of Table 6.2, and they are found to be very small in comparison to the differences between the experimental values and those estimated in computational studies. We conclude that this hydrated configuration has a minor impact on the FEB and cannot explain for the mismatch between experimental and computational $\Delta\Delta G^{bind}$ estimations.

6.3.2 Enhancing binding pocket water exchange in alchemical simulation

The hydrated state considered in previous section is only one possible metastable state, individuated from plain MD simulations performed in [112], but in principle different hydrated state of the binding pocket may occur, giving their contribution the free energy differences between methylated and unmethylated state of YHT-RNA complex. Although the hydrated state investigated previously seems to have a minor impact on these free energies, we still aim to improve the precision of our AFEC in the binding pocket, by allowing more exhaustive sampling with respect to water displacement in and out of the aromatic cage, all along the alchemical integration. In order to accelerate this process, we implemented the A-to-m^{6}A AFEC with a WT-MetaD (2.2.1) acting on a collective variable (CV) which is able to quantify the amount of water molecules approaching the binding pocket. To address this task, we made use of the coordination switching function implemented in PLUMED [55], which is able to calculate the coordination number of two groups of atoms (A and B), and is defined as:

$$CN = \sum_{i \in A} \sum_{j \in B} \frac{1}{1 + (\frac{r_{ij}}{r_0})^6} \tag{6.5}$$

In our implementation, we define group A as a single point at the center of the atoms N6; H61; and C10, whereas B is all water oxygens in the system. r_0 was set to 0.45 nm. A WT-MetaD on CN without any restraints could cause multiple water molecules entering the binding pocket in the same time, likely causing the RNA to detach from the binding. To avoid this, we implemented this AFEC setting up an upper harmonic wall potential, defined as follows:

$$\begin{cases} V_{walls}(x_i) = K(CN(x_i) - UW)^2 \ ; \ \ CN(x_i) > UW \\ V_{walls}(x_i) = 0 \ ; \ \ CN(x_i) \leq UW \end{cases} \tag{6.6}$$

where we set $K = 200$ kJ/mol and $UW = 2.5$.

The metadynamics was performed using the PLUMED package [55], depositing a Gaussian every 500 time steps, with initial height equal to 5 kJ/mol and width $\sigma = 0.05$. The bias factor was set to 3. The calculation of CN was accelerated making use of a neighbor list, which makes it that only a relevant subset of the pairwise distance are calculated at every step. We used a neighbor list cut-off of 0.8 nm, updating the lists every 10 steps.

We first performed the AFEC computation with WT-MetaD on CN running for 20 ns per replica. We then performed another AFEC with 100 ns per replica with a static bias, by restarting the previous AFEC with WT-Metad without further updating the bias.

Figure 6.4 shows the values of CN and a control variable d along the demuxed continuous trajectories. d is defined as a distance between the center of mass of m^{6}A nucleobase and the center of the residues forming the binding pocket. This variable can be monitored in order to check that the RNA does not displace from is binding pose.

During the static bias simulations, the hydrated state described in previous section, that is a water molecule coordinated with atom H62 of the hybrid adenosine, appears only in two circumstances: In the first demuxed trajectory it is always present, as a consequence this trajectory is not able to explore the full λ ladder (see top left corner of Figure 6.3), because of the steric clash occuring between the water molecule and the appearing methyl group. The second case corresponds to trajectory 13, where initially no water is present inside a binding pocket. However, after more than 80 ns, a water molecule enters in the aromatic cage and coordinates with H62. Also in this case, once the hydrated state forms, the trajectory is not able anymore to get to high λ values. All other cases in which CN goes to high values correspond to multiple water molecules approaching the binding pocket, but remaining stuck on the other side of the amino group, coordinating with atom H61 and residue SER35.

Another limit of the Metadynamics performed is that transition in the CV are here observed only in one direction. Although this enhanced sampling attempt shows many limitations, the obtained sampling is certainly more exhaustive of the one obtained previously without biasing CN.

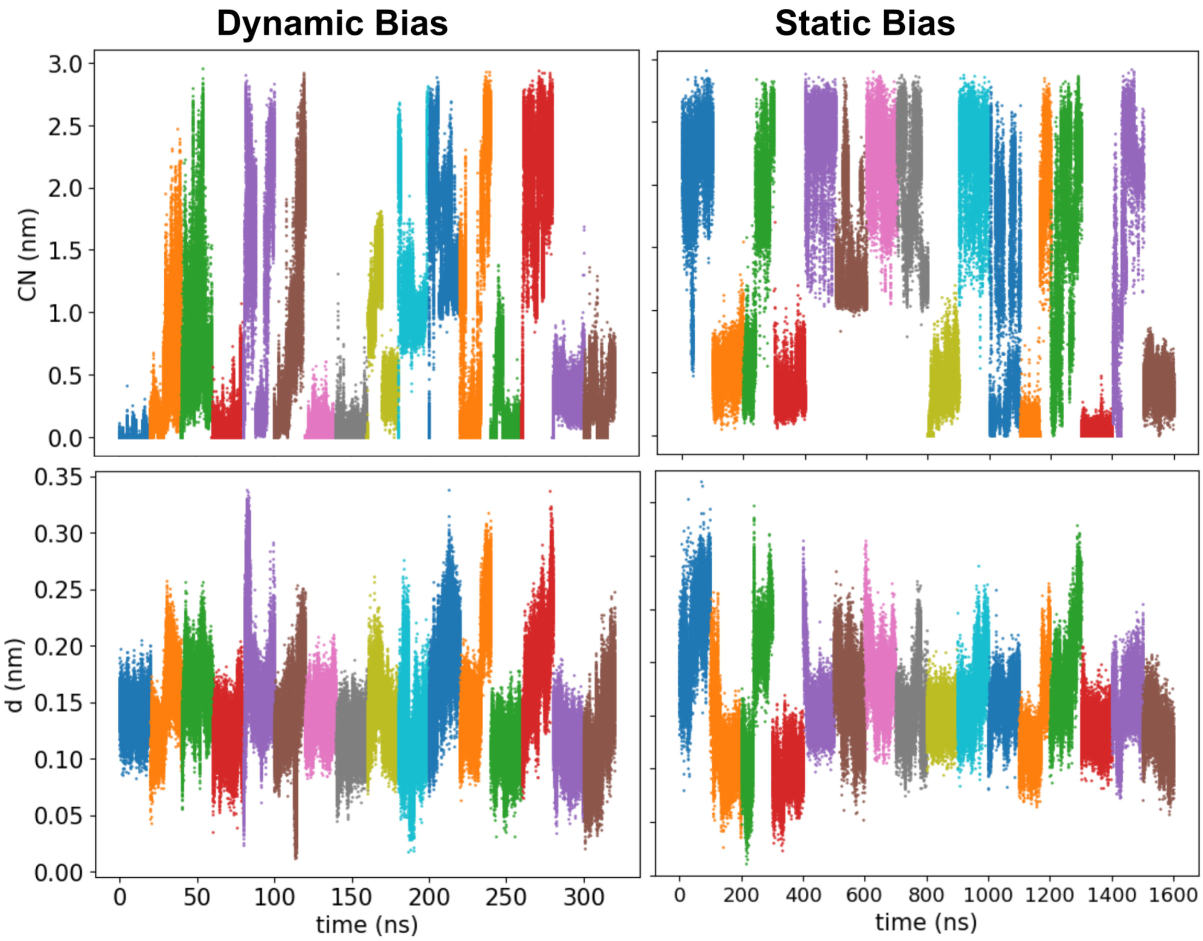


Figure 6.4 *CN* and *d* values along the demuxed trajectories, where each independent trajectory is represented with a different color. Plots on the left correspond to simulations performed with a dynamic bias. Plots on the right correspond to longer simulations performed with a static bias.

The ΔG computed with WHAM from the static bias simulation results in 185.4 ± 1.3 kJ/mol (see Table 6.1), resulting in a $\Delta\Delta G$ of 19.7 ± 1.5 kJ/mol (4.7 ± 0.4 kcal/mol, which is slightly reduced compared to the estimation done without enhancing the water displacement (5.3 ± 0.2 kcal/mol), but still highly overestimated compared to the experimental reference (2.3 kcal/mol).

6.3.3 Exploring m^{6}A force-fields perturbation effects on FEB

Since the influence of hydration appears to have a limited impact on the accuracy and precision of free energy estimations, we tested the hypothesis that the primary reasons for the discrepancies between experimental and computational results can result from the inaccuracies in the force-field parameters. We have used here the fit_A force-fields for m^{6}A, that we derived in our previous work (see Chapter 4) to better match *syn/anti* populations and denaturation experiments. This refinement involved adjusting a subset of partial charges that play a significant role in the stability of duplexes, particularly with respect to hydrogen bond strength involving atoms in the WC edges of the nucleobase. As far as the m^{6}A recognition by the YHT protein is concerned, there are other parameters which may play significant role in the stabilization. For instance, in the YHT complex m^{6}A performs hydrogen bonding with the protein residues also on its sugar edge, so it could be useful to refine the partial charge of nitrogen atom N3,

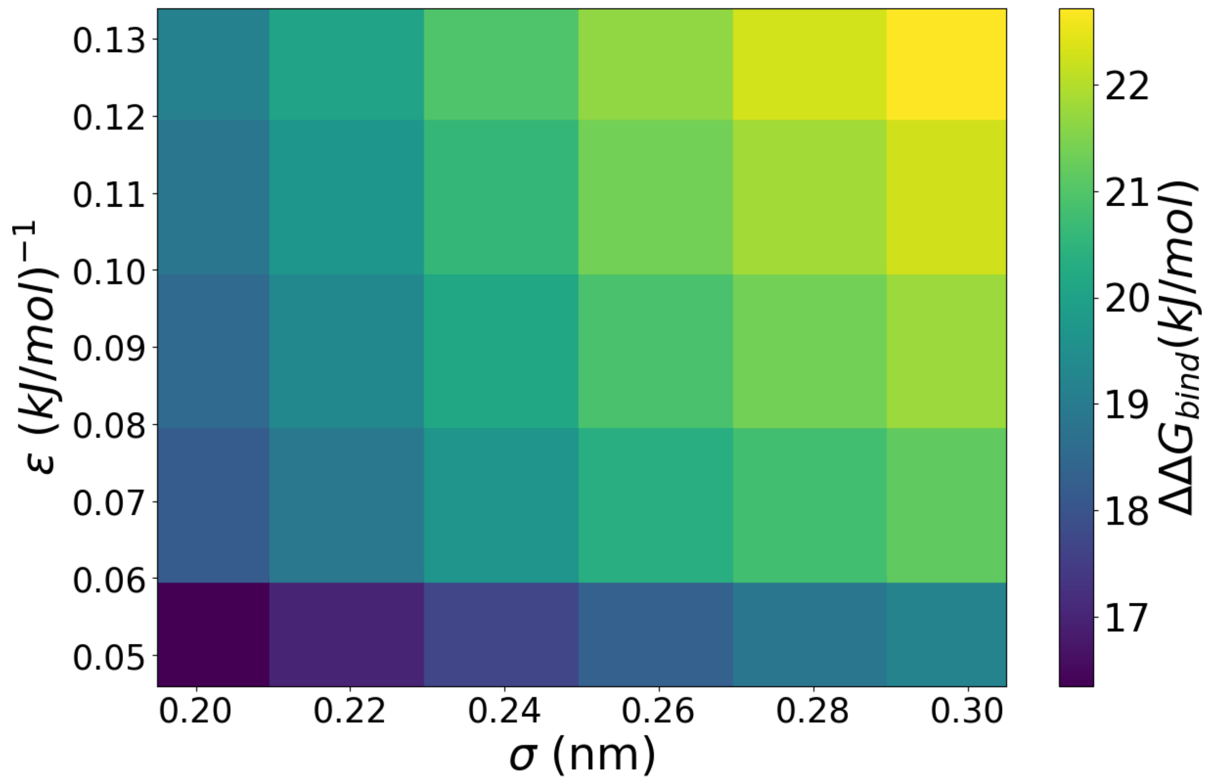


Figure 6.5 $\Delta\Delta G_{bind}$ computed through reweighting scanning over LJ parameters of the m^{6}A methyl hydrogens.

which was not considered in the fit_A fitting. Moreover, we would like to investigate if perturbations in the LJ parameters of the methyl group may have significant impact on the FEB. In the following, we investigate further into which parameters of the m^{6}A force-field have the most significant impact on fine-tuning the estimation of $\Delta\Delta G_{bind}$.

6.3.3.1 Lennard-Jones perturbations

As a first step, we intend to explore a range of reasonable values for the Lennard-Jones (LJ) parameters associated with the hydrogen atoms in the methyl group. In the Amber force-field, methyl group hydrogen atoms are characterized by LJ parameters ε and σ values of 0.0657 kJ/mol and 0.2471 nm, respectively. However, for other types of hydrogen atoms, these parameters can fall within the intervals of 0.05-0.13 kJ/mol and 0.2-0.3 nm. We computed $\Delta\Delta G_{bind}$ for parameter values within these defined intervals using a reweighting procedure. The results are summarized in Figure 6.5. It's worth noting that all the computed free energies presented in this study exhibit a Kish Size Ratio (KSR, refer to Section 4.1.3) greater than 0.1, indicating their statistical significance. We observe lower values of $\Delta\Delta G_{bind}$ when both ε and σ are set to low values. However, these perturbations are still insufficient to approach the experimental $\Delta\Delta G_{bind}$, which is 9.9 kJ/mol.

6.3.3.2 Alternative Charges Parametrizations

In Chapter 4 we have demonstrated how small variations in the partial charges can have significant impact when one aim to achieve good accuracy on computing free energy differences

	$\Delta G_{ss}^{\Delta Q}$	$\Delta G_{com}^{\Delta Q}$	$\Delta\Delta G_{bind}^{\Delta Q}$	$\Delta\Delta G_{bind}^{metad}$	$\Delta\Delta G_{bind}^{nometad}$
fit_A-to-Aduri	50.71 ± 0.10	62.37 ± 0.25	-11.7 ± 0.4	8.8 ± 1.6	10.2 ± 1.0
fit_A-to-Krepl	-19.78 ± 0.13	-11.12 ± 0.28	-8.7 ± 0.3	11.8 ± 1.6	13.2 ± 0.9

Table 6.3 First two columns report computed ΔGs with respect to different m^{6}A partial charges parametrization. Third column ($\Delta\Delta G_{bind}^{\Delta Q}$) report difference in $\Delta\Delta G_{bind}$ with respect to performed simulation with fit_A force-field. Fourth and fifth columns report $\Delta\Delta G_{bind}$ estimated with respect to AFEC with or without WT-MetaD on water displacement. All free energies are reported in kJ/mol.

induced by small modifications as the methylation of a nucleotide. Based on that knowledge, we can expect that variation in the m^{6}A partial charges can have a major impact on the $\Delta\Delta G_{bind}$ with respect to modification on LJ parameters. To investigate this, we first compute $\Delta\Delta G_{bind}$ for m^{6}A force-fields alternative to fit_A, as the Aduri force-field [28] and the parametrization used by Krepl *et al* [112]. Instead of performing new AFEC or used reweighting, we implemented simulations using an Hamiltonian Replica Exchange (HREX) scheme similar to the one used in AFEC (see 2.2.3), but where initial and final state in the integration correspond to different parametrization of the m^{6}A charges. $\lambda = 0$ would correspond to methylated state with fit_A charges, whereas $\lambda = 1$ would correspond to the methylated state with Aduri or Krepl parametrization. By performing this transformation on the YHT-RNA complex and on the relative ssRNA, and computing respectively $\Delta G_{com}^{\Delta Q}$ and $\Delta G_{ss}^{\Delta Q}$, we can compute the $\Delta\Delta G_{bind}^{ff}$ for the two different force-fields as:

$$\Delta\Delta G_{bind}^{ff} = \Delta\Delta G_{bind}^{fit\_A} + \Delta\Delta G_{bind}^{\Delta Q} \tag{6.7}$$

where

$$\Delta\Delta G_{bind}^{\Delta Q} = \Delta G_{ss}^{\Delta Q} - \Delta G_{com}^{\Delta Q} \tag{6.8}$$

We performed simulations starting from the YHT-RNA and the ssRNA (5′- CGm^{6}ACAC-3′) , using the HREX scheme with only 2 replica for the fit_A-to-Aduri integration, and 4 replica for the fit_A-to-Krepl integration. This choice of number of replica allows ensuring averaged transition probabilities over 20%. Simulations were 10 ns per replica long. Free energies difference were computed with BAR method implemented in GROMACS and are listed in Table 6.3.

Not surprisingly, Krepl and Aduri force-field destabilize the YHT-RNA complex with respect to Fit_A. Indeed, the latter parametrization is characterized by atom H61, N3 and N1 being more polar than in the other cases. All these atoms form hydrogen bonds in the aromatic cage, respectively with SER35, ASN20 and ASN24 residues, as shown in Fig.6.2. Based on our estimations of $\Delta\Delta G_{bind}$s, the Aduri force-field seems to be the most compatible with the experimental values, as it can be seen in the plot in fig 6.7a. However, Aduri force-field is not able to reproduce denaturation experiments, as shown in Chapter 4. As a consequence, none of the so far explored m^{6}A force-fields are able to reproduce all together isomers populations; denaturation experiments of duplexes, calorimetry experiments on the YHT-RNA complex.

6.3.4 Force-field refinement

As previously demonstrated, the m^{6}A fit_A force-field fails to accurately reproduce the results of ITC experiments concerning the stabilization induced by N6-methylation on the YHT-RNA

Training Set			**Validation Set**		
System	$\Delta\Delta G$ (kJ/mol)	Exp	System	$\Delta\Delta G$ (kJ/mol)	Exp
$A1_{syn/anti}$	6.3 ± 0.5	NMR [36]	$A2_{syn/anti}$	-11 ± 2	NMR [68]
A2	1.7 ± 0.9	DE [36]	B1	2.5 ± 2.1	DE [37]
A3	7.1 ± 0.9	DE [36]	B2	2.1 ± 1.3	DE [37]
A4	-2.5 ± 1.2	DE [36]	B3	5.4 ± 1.3	DE [37]
A5	-1.7 ± 0.9	DE [36]	B4	8.6 ± 0.8	DE [37]
C1	9.9 ± 0.5	ITC [141]	B5	1.7 ± 1.0	DE [37]

Table 6.4 List of Systems and relative Experimental $\Delta\Delta G$ considered in the fitting. These values and relative error are derived from Nuclear Magneti Resonance (NMR) experiments, optical melting Denaturation Experiments (DE) and Isothermal titration calorimetry (ITC) measurements.

complex. Although Aduri m^{6}A parametrizations provide a better match for this experimental observation, it has been proven to be inadequate in reproducing denaturation experiments and the *syn/anti* populations, as detailed in Chapter 4.

Consequently, we recognize the need to refine the m^{6}A parametrization further by extending the fitting procedure outlined in Chapter 4. This extension involves incorporating an expanded experimental dataset, which includes the YHT-RNA $\Delta\Delta G_{bind}$. The list of experiments considered for this fitting is provided in Table 6.4 and is divided into a training dataset and a validation dataset.

The fitting procedure described in 4.1.2 was re-adapted to work over the simulations performed with fit_A parametrization on systems A1-A2-A3-A4-A5, along with the YHT-RNA $\Delta\Delta G_{bind}$, which we will refer to as the C1 system. Systems B1-B2-B3-B4-B5, in addition to the $\Delta G_{syn/anti}$ for system A2 ($A2_{syn/anti}$), will be used to validate the parametrization derived from the fitting process.

We have chosen to refine once again the torsional parameter V_η and, in conjunction with it, we aim to optimize two distinct subsets of partial charges independently, resulting in finding two separate parametrizations, namely:

- fit6_AC: fitting 6 partial charges (C6-N6-H61-N1-C10-H101)
- fit5_AC: fitting 5 partial charges (N6-H61-N1-N3-C4)

While fit6_AC aims to explore the same charges space explored by the fittings illustrated in Chapter 4, fit5_AC is designed to investigate a smaller multidimensional space that includes atoms N3 and C4. In particular, the polarity of N3 may play a significant role in stabilizing the binding pocket in C1, as this atom forms hydrogen bonds with the ASN20 residue within the aromatic cage. Additionally, we have retained the charges of N1 and H61 atoms, which are involved in hydrogen bonding both in the dsRNA (A2–A4 and B1–B5) and in the YHT binding pocket (C1). Therefore, we expect that the fitting process will be highly sensitive to these charges. On the other hand, atoms N6 and C4 are primarily intended to absorb the perturbations introduced by the fitting process to the other three charges.

We remind that our fitting strategy, more deeply explained in 4.1.2, consists in the mini-

	C6 (e)	**N6** (e)	**H61** (e)	**N1** (e)	**C10** (e)	**H100** (e)	**N3** (e)	**C4** (e)	V_η (kJ/mol)
fit5_AC	0	-0.0363	-0.0595	0.0086	0	0	0.0657	0.0215	2.18
fit6_AC	0.0644	-0.0550	-0.0720	0.0687	-0.0272	0.0211	0	0	2.35

Table 6.5 Charge modifications (ΔQs) and torsional potential (V_η) for the fitting performed on the training data set AC.

mization of a Cost function defined as:

$$C = \chi^2 + \alpha \sum_{i=0}^{N} \Delta Q_i^2 + \beta V_\eta^2 = \chi^2 + \alpha [\sum_{i=1}^{N} \Delta Q_i^2 + (\sum_{i=1}^{N} \Delta Q_i)^2] + \beta V_\eta^2 \tag{6.9}$$

Here, in the context of fit6_AC, we have N=5, while in fit5_AC, N=4. The results of the two fittings are shown in Figure 6.6. Based on the insights learned in Chapter 4, we discarded regularization on V_η setting $\beta = 0$. Panels 6.6a and 6.6b display the optimized parameters at different α values, while panels 6.6c and 6.6d depict the corresponding χ^2 values and KSR values for each parameter set obtained at different α values. In both cases, at lower α values, the fitting effectively enforces experiment C1, but at the expense of yielding very low KSR values, making the free energy estimation statistically insignificant. As α values increase, the χ^2 values for C1 rise significantly, while the χ^2 values for other experiments remain relatively stable and sometimes even improve. This outcome is not unexpected, as higher values of α constrain the parametrization to the fit_A force-field, which was designed to match the A1–A5 experimental data and is intended to perform well for them. The minimum α values that ensure a KSR above 0.1 are respectively $\alpha = 1000$ e^{-2} and $\alpha = 2000$ e^{-2} for fit6_AC and fit5_AC. The charge values obtained by minimizing the cost function for these α values were selected as the results of the two fittings. This choice is further validated by the estimation of the χ^2 on the validation dataset, as shown in panels 6.6e and 6.6f. ΔQs values with respect to fit_A parametrizations are shown in Table 6.5, along with the V_η values. It's worth noting that both fittings result in a decrease in the polarity of atoms N1, H61, and also N3 in the fit5_AC case, as expected.

Table 6.6 provides the averaged χ^2 values computed separately for the training and validation datasets, as well as the overall average. The columns labeled fit6_AC (rew) and fit5_AC (rew) represent the results obtained through reweighting. While fit6_AC demonstrates better performance in the training dataset compared to fit5_AC, it exhibits poor performance in the validation dataset, resulting in a total χ^2 score that is even worse than the initial state of the fitting (fit_A). On the other hand, Fit5_AC performs quite well on the validation dataset, making it the better candidate to serve as the best parametrization to align with the entire dataset. Based on this observation, we conducted new simulations of the complete dataset shown in Table 6.4 using the fit5_AC m^{6}A force-field. This also included a new simulation of the YHT-RNA complex employing the same AFEC+WT-MetaD procedure as previously utilized.

Panel 6.7a presents a summary of all the $\Delta\Delta G_{bind}$ values computed in this study for different parametrizations, alongside the experimental value. It is evident that the newly fitted parameters do not replicate $\Delta\Delta G_{bind}$ as effectively as fit_A does. Instead, they represent a balanced compromise between matching C1 $\Delta\Delta G$ and denaturation experiments, as demonstrated in panel 6.7b. Specifically, while A2-A3 and B1–B5 demand an enhancement in the polarity of N1 and H61 atoms to strengthen hydrogen bonds and stabilize the duplexes, the C1 experiment necessitates the opposite effect to reduce the overestimation of $\Delta\Delta G_{bind}$. In comparison to fit6_AC, fit5_AC exhibits greater flexibility by allowing adjustments to the partial charge of atom N3, which is

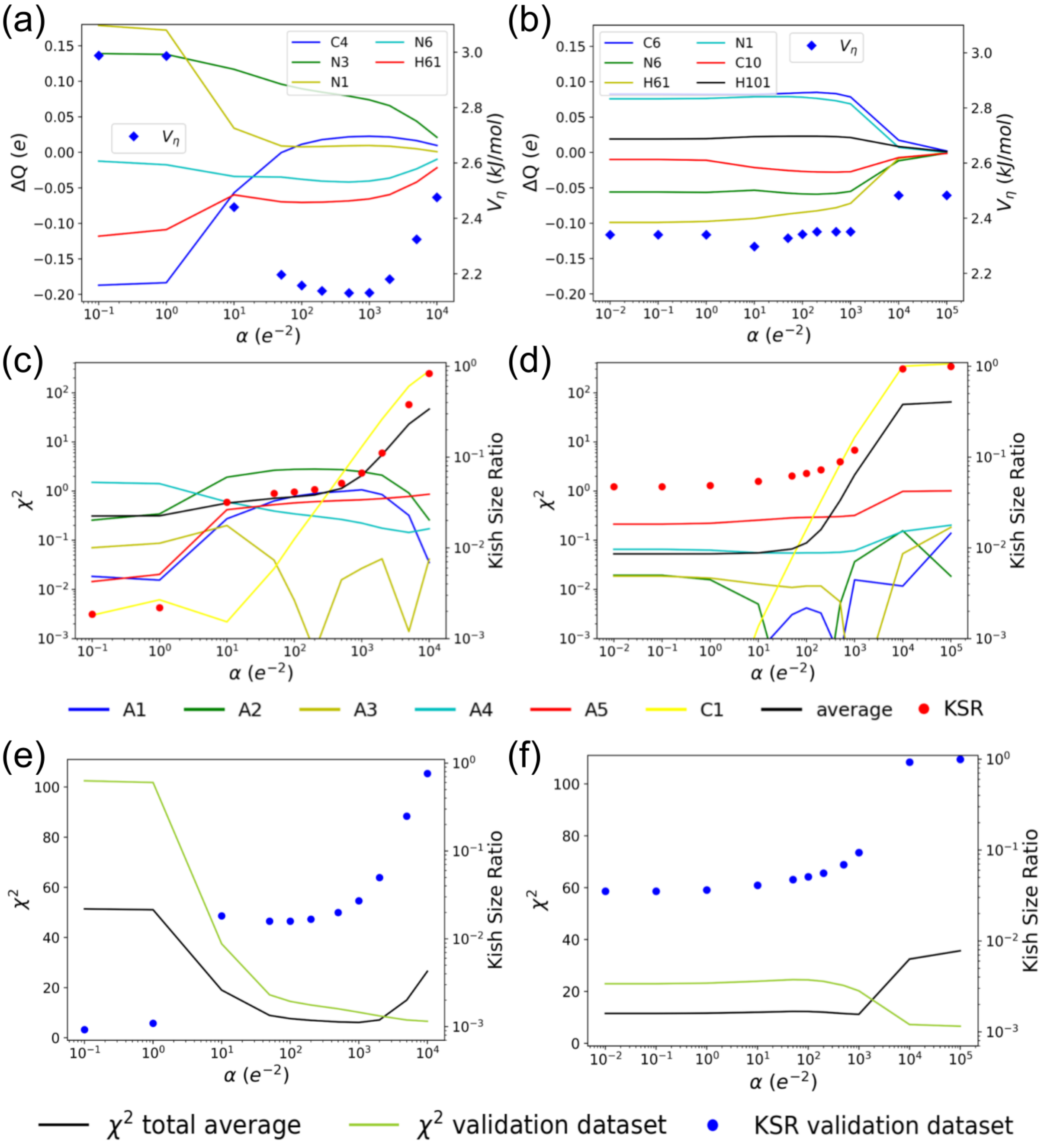


Figure 6.6 Parameters (ΔQ and V_η) obtained fitting on the traning dataset as a function of α, with $\beta = 0$, for fit5_AC (panel a) and fit6_AC (panel b). χ^2 errors for individual experiments of the training dataset and Kish size ratio (KSR) as a function of α, with $\beta = 0$, for fit5_AC (panel c) and fit6_AC (panel d). Averaged χ^2 obtained for the total dataset (black line) and on the validation dataset (yellow line), for fit5_AC (panel e) and fit6_AC (panel f). The KSR computed on the validation dataset is also shown (blue dots).

χ^2	Aduri	Aduri+tors	fit_A	fit6_AC (rew)	fit5_AC (rew)	fit5_AC
Training Set	16	3.8	4.5	0.33	0.9	2.2
Validation Set	9.7	14	6.5	18	7.5	6.7
Total	12.9	8.9	5.5	9	4.2	4.5

Table 6.6 χ^2 computed for the training data set AC (second row) the validation data set B+A2$_{syn/anti}$ (third row), and the total average (fourth row) for different m^{6}A force-fields. fit6_AC (rew) and fit5_AC (rew) are χ^2 values obtained trough the fitting by reweighting.

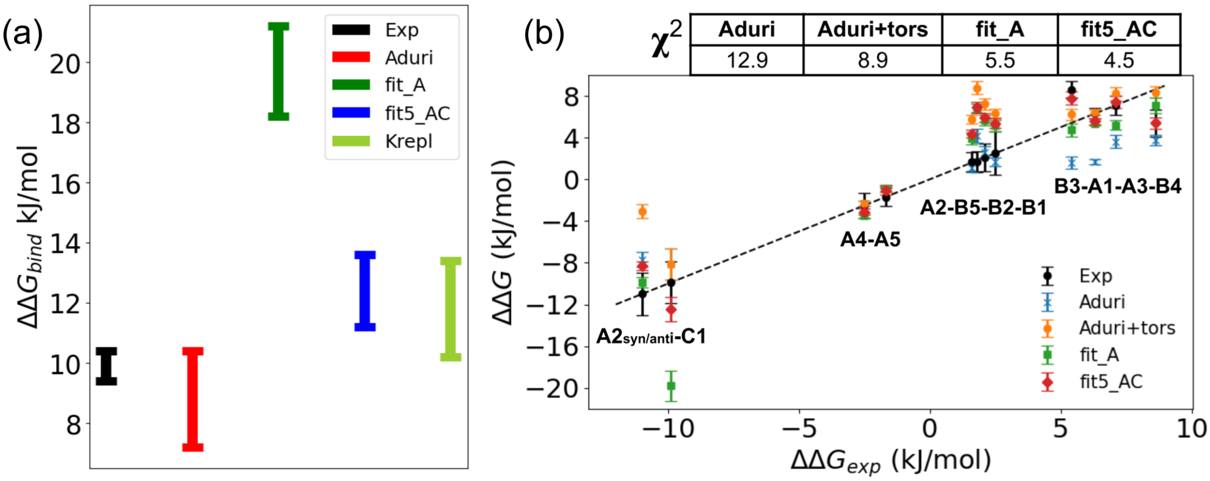


Figure 6.7 (a) $\Delta\Delta G_{bind}$ values and relative experimental or statistical error (b) $\Delta\Delta G$ computed for each of the 12 analyzed systems with 4 different sets of parameters. $\Delta\Delta G$ for system C1 is shown as the inverse of $\Delta\Delta G_{bind}$. χ^2 obtained for each force-field set of parameters are shown in the table.

believed to be more sensitive to experiment C1 than in the duplex systems, where N3 does not form hydrogen bonds.

6.4 Conclusions

In this chapter we investigated the role of m^{6}A in RNA recognition, in the context of the YTH domain of the YTHDC1 protein. Firstly, we gave a general picture of what is known about this RNA-protein complex, going through several structural and computational studies which revealed how m^{6}A is recognized within an aromatic cage. MD simulations have already been used to investigate the binding mechanism in this system [41] [110] [112], but they failed to reproduce the m^{6}A-induced stabilization ($\Delta\Delta G_{bind}$) in YHT-RNA binding as expected from experiments. In our work, we aimed to repeat the alchemical calculation needed to estimate $\Delta\Delta G_{bind}$, by using our A-to-m^{6}A AFEC procedure described in 2.2.3 and by making use of the m^{6}A force-field derived in Chapter 4 (fit_A). Since our estimation resulted in a large overestimation of $\Delta\Delta G_{bind}$, we started investigating the possible factors causing this inaccuracy, and in general, affecting the computational estimation of $\Delta\Delta G_{bind}$. We attributed it to two main factors: inaccuracies in the force-fields and limited sampling. Specifically, the role of hydration in the binding pocket is considered a potential source of limited sampling, impacting the precision of $\Delta\Delta G_{bind}$ calculations. In particular, we investigated the impact on $\Delta\Delta G_{bind}$ of a specific hydrated state already observed by Krepl *et al* [112], consisting in a water molecule coordinating with hydrogen H62 in the unmethylated state. We first conducted alchemical calculation on this water molecule, and our results suggest that this metastable state of the binding pocket has a minor impact on the free energy of binding and cannot explain the discrepancies between experimental and computational $\Delta\Delta G_{bind}$ estimations. After that, we implemented a WT-MetaD in the A-to-m^{6}A AFEC, in order to improve the precision of our calculations by enhancing the water displacement in the binding pocket. This biasing acted on a collective variable (CV) that quantified the approach of water molecules to the binding pocket. With this method we could compute a $\Delta\Delta G_{bind}$ which is slightly reduced with respect to the one estimated through unbiased AFEC, and as a consequence still highly overestimated with respect to the experimental reference, confirming the fact that alternative hydrated states of the binding pocket have minor impact on $\Delta\Delta G_{bind}$. In Chapter 4 we demonstrated how small variations in partial charges can impact the accuracy of computing free energy differences in the case of adenine N6-methylation. This knowledge led to the expectation that variations in m^{6}A partial charges could substantially affect $\Delta\Delta G_{bind}$ compared, for example, to modifications in LJ parameters. However, we also decided to investigate the perturbation on $\Delta\Delta G_{bind}$ induced by variations on the LJ parameters of the methyl hydrogens. Our results demonstrated that this perturbation can be significant, but still inferior to the correction obtainable with alternative parametrizations for the charges, especially if considering perturbations in LJ parameters for the methyl hydrogens that are compatible with alternative hydrogens LJ parameters in the Amber force-field. To investigate the impact of partial charges on binding, we computed $\Delta\Delta G_{bind}$ for alternative m^{6}A force-fields, including the Aduri force-field and the parametrization used by Krepl *et al.* The results showed that the Krepl and Aduri force-fields destabilized the YHT-RNA complex compared to Fit_A. We attributed this to Fit_A having more polar H61, N3 and N1 atoms, which formed hydrogen bonds in the aromatic cage. Although Aduri force-field can reproduce experimental $\Delta\Delta G_{bind}$, none of the so far considered m^6A force-fields could fully replicate isomer populations, duplex denaturation experiments, and calorimetry experiments on the YHT-RNA complex simultaneously. To address these issues, we extended the fitting procedure used in our previous work [113], using an expanded experimental dataset, that includes the YHT-RNA $\Delta\Delta G_{bind}$. The newly fitted parameters (fit5_AC) do not replicate $\Delta\Delta G_{bind}$ as effectively as fit_A but offer a balanced compromise between matching $\Delta\Delta G_{bind}$ and denaturation experiments. This is due to the adjustments made to the polarity of N1, H61, and N3 atoms, which impact hydrogen bonding and stability in stability in different systems.

Chapter 7

Conclusions

This thesis presents our research into the influence of post-transcriptional modifications on RNA structural dynamics. To achieve this, we employed molecular dynamics simulations (MD) in conjunction with advanced enhanced sampling techniques. Recognizing the inherent limitations of MD in accurately predicting RNA dynamics, particularly for modified nucleotides for which MD force-fields have not yet been extensively validated, we consistently incorporated data from solution experiments into our computational methodologies.

In chapter 3, we discuss about a collaborative research project focusing on the study of a 20 base pairs double-stranded RNA (dsRNA) containing four inosines in the central part of the helix, each paired with uracils. In this project we use experimental data generated by our collaborator for two main purposes: guiding MD simulations using maximum entropy principles and validating the resulting ensembles of structures. In our simulations we employed the replica exchange collective variable tempering (RECT) method to enhance sampling and explore different configurations of the dsRNA, specifically focusing on sugar puckering conformations. The maximum entropy (ME) principle was then applied to reweight the simulation trajectories, aiming to generate ensembles of structures that match NMR ^{3}J coupling signals and enforce averaged radius of gyration as predicted by SAXS experiments. In this work we highlighted the limitations of MD in predicting accurate ensembles, particularly regarding sugar puckering populations. However, we also show how the combination of enhanced sampling and ensemble refinement techniques which integrate experimental data can be used to generate accurate ensembles. Indeed, the accuracy of our refined ensembles was further validated against various experimental data not included in refinement, as NOE signals and the full SAXS spectra. Our results revealed the impact of A-to-I hyper-editing on the dsRNA conformational ensemble, which with respect to the unmodified counterpart shows increased flexibility, uncommon helical parameters, and increased populations of unexpected C2'-endo sugar puckering conformations.

In Chapter 4 we show our attempt to refine the force-field of the N6-methyladenosine (m^{6}A) in order to produce molecular simulations that match denaturation experiments. Our approach resulted in a novel formalims, building upon previous force-field fitting strategies, allowing alchemical free energy calculations (AFEC) to serve as a reference for reparametrizing partial charges and a torsional potential. Within this context, we also proposed a novel efficient method for recomputing the total energy of the system using test charges, making force-field fitting iterations significantly faster. This work represents, to our knowledge, the first attempt to tune partial charges of a biomolecular force-field based on experiments performed on macromolecular complexes. Our fitting procedure enables the use of MD to accurately reproduce nine denaturation experiments and to correctly capture the *syn/anti* populations for both paired and unpaired m^{6}A, as predicted by NMR experiments. It's noteworthy that no data regarding the *syn/anti* populations of paired m^{6}A was utilized in the fitting process. Consequently, the improved performance in reproducing these features by the refined force-field serves as a validation, affirming that the new parameterization is indeed more adept at describing the interaction of m^{6}A with its surrounding environment. This advancement opens the door to the use of MD for investigating the impact of m^{6}A on the structural dynamics of other RNA systems. Additionally, it's worth emphasizing that we achieved a significant improvement in agreement with experimental free energies with relatively minor adjustments to a subset of partial charges in the m^{6}A nucleobase. For example, these adjustments are notably smaller than the variations in charges obtained through standard methods, such as quantum mechanical calculations followed by restrained electrostatic potential (RESP) fitting, but with slightly different procedures. This highlights two key points: (*i*) Utilizing MD to robustly predict free energy differences on the order of a few kJ/mol is a challenging endeavor, as even slight variations in force-field parameters can have a substantial impact on the estimations. (*ii*) Fitting partial charges to macroscopic experiments is a powerful approach for improving the quality of force-fields.

In Chapter 5 we discussed the application of alchemical metadynamics [67], which extends the traditional metadynamics approach by introducing an additional alchemical dimension for sampling. We applied the method to a couple of m^{6}A systems already investigated in Chapter 4, showing how the methods can be used to efficiently reproduced the same results already obtained with our AFEC procedure, which makes use of an Hamiltionian replica exchange mechanism. The advantage of using alchemical metadynamics in our applications, arises from the fact that both the *syn* and *anti* m^{6}A isomers can be sampled within a single alchemical simulation. Additionally, alchemical metadynamics also enables the reconstruction of the free energy profile along the biased torsional angle, giving an estimation of the free energy barriers, and as a consequence, deeper insights into the m^{6}A isomers kinetics.

Finally, in Chapter 6 we investigate the role of m^{6}A in RNA recognition, particularly in the context of the YTH domain of the YTHDC1 protein. Previous MD simulations had attempted to investigate the binding mechanism in this system but struggled to reproduce the m^{6}A-induced stabilization ($\Delta\Delta G_{bind}$) observed in experimental findings. To address this, we aimed to reevaluate the alchemical calculations needed to estimate $\Delta\Delta G_{bind}$ using our A-to-m^{6}A AFEC procedure, and by using the refined m^{6}A force-field derived in Chapter 4. However, our initial estimations significantly overestimated $\Delta\Delta G_{bind}$, prompting us to investigate the factors contributing to this inaccuracy and, more broadly, influencing the computational estimation of $\Delta\Delta G_{bind}$. We identified two main factors: inaccuracies in the force-fields and limitations in sampling. One potential source of limited sampling precision was the role of hydration in the binding pocket, which could impact $\Delta\Delta G_{bind}$ calculations. We specifically examined the impact of a particular hydrated state previously observed by our collaborator, but our alchemical calculations suggested that this hydrated state had a minor impact on binding free energy. In an effort to enhance the precision of our calculations, we implemented a WT-MetaD approach in the A-to-m^{6}A AFEC, focusing on improving water displacement within the binding pocket. This method acted on a collective variable (CV) measuring the approach of water molecules to the binding pocket, and led to a slightly reduced estimation of $\Delta\Delta G_{bind}$ compared to unbiased AFEC. However, this implementation was still not able to account for the discrepancy between simulation of experiments. This further confirmed that alternative hydrated states of the binding pocket had minimal impact on $\Delta\Delta G_{bind}$. Lastly, we expanded the fitting procedure introduced in Chapter 4 by incorporating experimental data for the YHT-RNA $\Delta\Delta G_{bind}$ into our dataset. The newly fitted parameters significantly increase the accuracy of $\Delta\Delta G_{bind}$, by strucking a balanced compromise between matching $\Delta\Delta G_{bind}$ and denaturation experiments. This equilibrium was mostly guided through adjustments to the polarity of N1, H61, and N3 atoms, influencing hydrogen bonding and stability in dsRNAs and in the aromatic cage of the YHT-RNA complex. This new m^{6}A parametrization, referred to as fit5_AC, emerges as the most suitable force-field among those explored in our investigations. It minimizes the discrepancy between simulations and experiments across a diverse dataset, which includes: denaturation experiments (optical melting) reflecting m^{6}A-induced destabilization in dsRNAs; NMR experiments reporting *syn/anti* population ratios in both paired and unpaired m^{6}A; isothermal titration calorimetry experiments quantifying the stabilization induced by m^{6}A on the free energy of binding for the YHT-RNA complex. Consequently, we recommend employing this parametrization for future applications.

I would like to conclude this Thesis saying that working on these projects has been an enlightening journey. Above all, it has unveiled to me the fascinating and intricate world of RNA, and molecular biology more broadly. Furthermore, it has allowed me to gain a profound awareness of the advantages, as well as the limitations, that computational methods like molecular dynamics simulations can have when studying complex molecules such as RNA sysyems. In particular, one of the most important awareness gained relates to the significance of combining molecular dynamics with solution experiments, and therefore, of creating strong collaborations between experimental and computational laboratories. Such collaborations bring together highly complementary expertises, providing both the high-resolution insights afforded by simulations and the accuracy and reliability of solution experiments. I hope that in the future, I will have the chance again to collaborate in broad projects combining my computational skills with other orthoganal expertises, with the goal of tackling the fascinating challenges that structural biology can offer.

Appendices

	Aduri	fit_A	fit_AB	fit5_AC	RESP_anti	RESP_syn	Krepl
N9	-0.07829	-0.07829	-0.07829	-0.07829	-0.0564	-0.1834	-0.1719
C8	0.13844	0.13844	0.13844	0.13844	0.0815	0.2573	0.0631
H8	0.16681	0.16681	0.16681	0.16681	0.1726	0.1329	0.1973
N7	-0.59080	-0.59080	-0.59080	-0.59080	-0.5250	-0.5854	-0.5652
C5	0.03544	0.03544	0.03544	0.03544	0.0226	-0.2346	0.0152
C6	0.44911	0.46801	0.45811	0.46801	0.5880	0.7140	0.5597
N6	-0.30623	-0.22923	-0.25723	-0.26603	-0.3756	-0.4189	-0.4756
H61	0.28948	0.38888	0.35648	0.32888	0.3306	0.3392	0.3232
C10	-0.28897	-0.28467	-0.25597	-0.28467	-0.3009	-0.3239	-0.0774
H101	0.12596	0.07536	0.09096	0.07536	0.1299	0.1400	0.0774
H102	0.12596	0.07536	0.09096	0.07536	0.1299	0.1400	0.0774
H103	0.12596	0.07536	0.09096	0.07536	0.1299	0.1400	0.0774
N1	-0.67597	-0.72167	-0.72897	-0.71357	-0.8746	-0.7617	-0.6604
C2	0.55132	0.55132	0.55132	0.55132	0.6898	0.5688	0.4636
H2	0.05539	0.05539	0.05539	0.05539	0.0485	0.0692	0.0865
N3	-0.73497	-0.73497	-0.73497	-0.66927	-0.8037	-0.7900	-0.7027
C4	0.48723	0.48723	0.48723	0.508732	0.4807	0.6559	0.4589

Table 1 Charges for all atoms of the m6A nucleobase for different parametrizations.

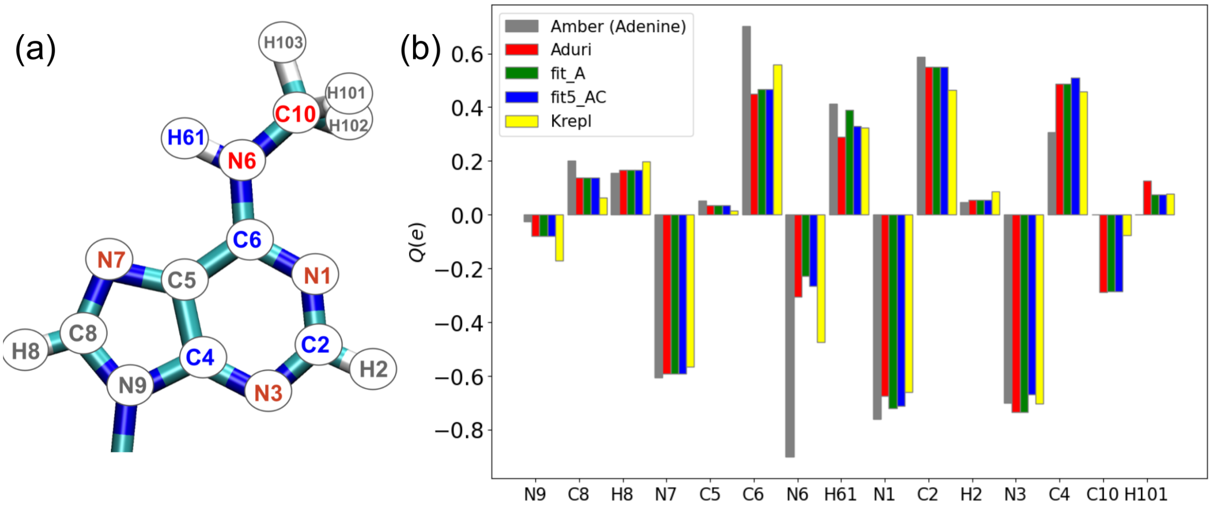


Figure 1 (a)m^{6}A nucleobase scheme. Atoms are colored based on partial charges tendency. (b) Partial charges for ifferent parametrization examined in this thesis.

	Aduri		**Aduri+tors**	**fit_A**		**fit_AB**		**fit5_AC**	
method	BAR	WHAM	WHAM+tors	BAR	WHAM	BAR	WHAM	BAR	WHAM
A1 *syn*	258.24 ± 0.22	258.24 ± 0.21	258.28 ± 0.21	207.19 ± 0.16	207.22 ± 0.16	211.42 ± 0.21	211.23 ± 0.18	237.10 ± 0.20	237.00 ± 0.19
A1 *anti*	260.12 ± 0.12	259.95 ± 0.15	264.61 ± 0.15	213.12 ± 0.20	213.29 ± 0.14	217.03 ± 0.13	217.27 ± 0.19	242.58 ± 0.12	242.57 ± 0.23
A2 dup *anti*	258.85 ± 0.70	258.63 ± 0.33	263.29 ± 0.33	208.32 ± 0.17	208.3 ± 0.5	214.3 ± 0.7	214.01 ± 0.35	241.0 ± 0.6	240.93 ± 0.27
A2 dup *syn*	266.44 ± 0.42	266.4 ± 0.4	266.4 ± 0.4	218.69 ± 0.29	218.74 ± 0.31	223.0 ± 0.5	222.98 ± 0.34	249.2 ± 0.4	249.2 ± 0.3
A2 ss *syn*	257.52 ± 0.31	257.54 ± 0.27	257.58 ± 0.27	206.45 ± 0.28	206.50 ± 0.29	210.34 ± 0.34	210.40 ± 0.22	236.6 ± 0.3	236.6 ± 0.4
A3 dup *anti*	261.56 ± 0.35	261.39 ± 0.32	266.15 ± 0.32	213.10 ± 0.44	213.0 ± 0.6	216.38 ± 0.43	216.21 ± 0.38	244.0 ± 0.3	244.1 ± 0.4
A3 dup *syn*	267.77 ± 0.35	267.75 ± 0.30	267.79 ± 0.30	217.78 ± 0.26	217.93 ± 0.24	221.96 ± 0.23	222.11 ± 0.27		
A3 ss *syn*	257.75 ± 0.24	257.80 ± 0.32	257.84 ± 0.32	206.37 ± 0.29	206.31 ± 0.35	211.9 ± 0.5	211.17 ± 0.29	236.81 ± 0.21	236.6 ± 0.4
A4 dup *syn*	255.06 ± 0.19	255.07 ± 0.17	255.11 ± 0.17	203.76 ± 0.14	203.64 ± 0.19	208.11 ± 0.15	208.07 ± 0.17	234.2 ± 0.5	234.19 ± 0.19
A4 ss *syn*	257.40 ± 0.19	257.42 ± 0.18	257.46 ± 0.18	206.76 ± 0.18	206.70 ± 0.17	210.56 ± 0.13	210.62 ± 0.19	237.30 ± 0.23	237.29 ± 0.25
A5 dup *syn*	256.89 ± 0.11	256.80 ± 0.19	261.76 ± 0.19	206.06 ± 0.25	205.89 ± 0.16	209.88 ± 0.16	209.93 ± 0.16	235.43 ± 0.28	235.64 ± 0.23
A5 ss *syn*	257.56 ± 0.15	257.70 ± 0.23	257.74 ± 0.23	206.65 ± 0.15	206.68 ± 0.19	211.01 ± 0.18	210.96 ± 0.16	236.69 ± 0.15	236.6 ± 0.4
B1 dup *anti*	259.09 ± 0.30	259.18 ± 0.21	263.94 ± 0.21	209.67 ± 0.30	209.62 ± 0.38	213.79 ± 0.35	213.81 ± 0.22	241.5 ± 0.4	241.5 ± 0.4
B1 ss *syn*	257.73 ± 0.25	257.60 ± 0.36	257.64 ± 0.36	205.46 ± 0.14	205.37 ± 0.34	210.27 ± 0.39	210.3 ± 0.6	236.21 ± 0.29	236.14 ± 0.28
B2 dup *anti*	521.6 ± 0.9	521.6 ± 0.9	530.9 ± 0.9	425.24 ± 1.9	425.26 ± 1.3	434.2 ± 1.0	434.1 ± 1.2	486.5 ± 0.9	487.0 ± 1.0
B2 ss *syn*	258.34 ± 0.31	258.20 ± 0.35	258.24 ± 0.35	207.3 ± 0.5	206.3 ± 0.5	211.13 ± 0.40	211.19 ± 0.29	237.6 ± 0.5	237.58 ± 0.19
B3 dup *anti*	518.5 ± 1.0	518.6 ± 0.9	527.9 ± 0.9	420.7 ± 1.0	420.7 ± 0.9	430.6 ± 1.4	430.5 ± 0.5	484.34 ± 0.8	484.5 ± 0.9
B3 ss *syn*	257.74 ± 0.16	257.72 ± 0.27	257.76 ± 0.27	206.77 ± 0.23	206.75 ± 0.39	210.38 ± 0.35	210.45 ± 0.22	236.83 ± 0.10	236.85 ± 0.25
B4 dup *anti*	523.2 ± 1.2	523.1 ± 0.5	532.4 ± 0.5	428.8 ± 1.2	428.9 ± 0.9	433.6 ± 0.9	434.8 ± 1.1	489.7 ± 0.4	489.4 ± 1.1
B4 ss *syn*	257.84 ± 0.45	257.85 ± 0.41	257.89 ± 0.41	206.53 ± 0.29	206.72 ± 0.35	210.87 ± 0.35	210.7 ± 0.5	236.90 ± 0.21	236.9 ± 0.4
B5 dup *anti*	521.9 ± 0.8	522.5 ± 0.7	531.8 ± 0.7	424.0 ± 0.7	424.0 ± 1.1	434.4 ± 0.9	434.4 ± 1.1	487.1 ± 1.0	487.0 ± 1.0
B5 ss *syn*	257.04 ± 0.34	257.10 ± 0.43	257.14 ± 0.43	205.42 ± 0.25	205.37 ± 0.23	209.82 ± 0.35	209.85 ± 0.19	236.63 ± 0.18	236.63 ± 0.17

Table 2 ΔGs computed through alchemical computations, with different parametrizations and Free Energy methods, reported in kJ/mol. We note that, in addition to the systems required to compute the *syn/anti* balance in the nucleoside (A1) and the effect of methylation in hybridization energies (A2–A5 and B1–B5), this table also reports control results for systems A2 and A3 where the duplex simulation was performed in the unexpected *syn* conformation.

References

[1] Jiri Šponer, Giovanni Bussi, Miroslav Krepl, Pavel Banáš, Sandro Bottaro, Richard A Cunha, Alejandro Gil-Ley, Giovanni Pinamonti, Simón Poblete, Petr Jurečka, Nils G. Walter, and Michal Otyepka. RNA structural dynamics as captured by molecular simulations: a comprehensive overview. *Chem. Rev.*, 118(8):4177–4338, 2018.

[2] Evgeny Nudler and Alexander S. Mironov. The riboswitch control of bacterial metabolism. *Trends in Biochemical Sciences*, 29(1):11–17, 2004.

[3] Svetlana A. Shabalina and Eugene V. Koonin. Origins and evolution of eukaryotic rna interference. *Trends in Ecology & Evolution*, 23(10):578–587, October 2008.

[4] Jörg Fallmann, Sebastian Will, Jan Engelhardt, Björn Grüning, Rolf Backofen, and Peter F. Stadler. Recent advances in rna folding. *Journal of Biotechnology*, 261:97–104, November 10 2017.

[5] Valerio Piomponi, Mattia Bernetti, and Giovanni Bussi. Molecular dynamics simulations of chemically modified ribonucleotides. *arXiv*, 2022. Submitted as a chapter for the book "RNA Structure and Function", series "RNA Technologies", published by Springer.

[6] Jan Barciszewski, editor. *RNA Structure and Function*. RNA Technologies. Springer Cham, 1 edition, 2023.

[7] Frank F. Davis and Frank Worthington Allen. Ribonucleic acids from yeast which contain a fifth nucleotide. *J. Biol. Chem.*, 227:907–915, 1957.

[8] Sigrid Nachtergaele and Chuan He. The emerging biology of RNA post-transcriptional modifications. *RNA Biol.*, 14:156–163, 2017.

[9] Jillian Ramos and Dragony Fu. The emerging impact of tRNA modifications in the brain and nervous system. *Biochim. Biophys. Acta Gene. Regul. Mech.*, 1862:412–428, 2019. mRNA modifications in gene expression control.

[10] Tsutomu Suzuki. The expanding world of tRNA modifications and their disease relevance. *Nat. Rev. Mol. Cell Biol.*, 22:375–392, 2021.

[11] Felix Voigts-Hoffmann, Martin Hengesbach, Andrei Yu. Kobitski, et al. A methyl group controls conformational equilibrium in human mitochondrial tRNA$^{\mathrm{Lys}}$. *J. Am. Chem. Soc.*, 129:13382–13383, 2007.

[12] Jun Jiang, Hyosuk Seo, and Christine S Chow. Post-transcriptional modifications modulate rRNA structure and ligand interactions. *Acc. Chem. Res.*, 49:893–901, 2016.

[13] Wendy V. Gilbert, Tristan A. Bell, and Cassandra Schaening. Messenger RNA modifications: Form, distribution, and function. *Science*, 352:1408–1412, 2016.

[14] W Brad Wan and Punit P Seth. The medicinal chemistry of therapeutic oligonucleotides. *J. Med. Chem.*, 59:9645–9667, 2016.

[15] Abid Hussain, Haiyin Yang, Mengjie Zhang, Qing Liu, Ghallab Alotaibi, Muhammad Irfan, Huining He, Jin Chang, Xing-Jie Liang, Yuhua Weng, and Yuanyu Huang. mrna vaccines for covid-19 and diverse diseases. *Journal of Control Release*, 345:314–333, May 2022.

[16] Andrea Tanzer, Ivo L. Hofacker, and Ronny Lorenz. RNA modifications in structure prediction – status quo and future challenges. *Methods*, 156:32–39, 2019.

[17] Marc Parisien and Francois Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452:51–55, 2008.

[18] Rhiju Das, John Karanicolas, and David Baker. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods*, 7:291–294, 2010.

[19] Raphael JL Townshend, Stephan Eismann, Andrew M Watkins, et al. Geometric deep learning of RNA structure. *Science*, 373:1047–1051, 2021.

[20] Ron O Dror, Robert M Dirks, JP Grossman, et al. Biomolecular simulation: a computational microscope for molecular biology. *Annu. Rev. Biophys.*, 41:429–452, 2012.

[21] Martin Karplus and Gregory A Petsko. Molecular dynamics simulations in biology. *Nature*, 347:631–639, 1990.

[22] Wendy D Cornell, Piotr Cieplak, Christopher I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, David C Spellmeyer, Thomas Fox, James W Caldwell, and Peter A Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117(19):5179–5197, 1995.

[23] Thorben Fröhlking, Mattia Bernetti, Nicola Calonaci, and Giovanni Bussi. Toward empirical force fields that match experimental observables. *J. Chem. Phys.*, 152(23):230902, 2020.

[24] Dazhi Tan, Stefano Piana, Robert M Dirks, and David E Shaw. RNA force field with accuracy comparable to state-of-the-art protein force fields. *Proc. Natl. Acad. Sci. U.S.A.*, 115(7):E1346–E1355, 2018.

[25] Petra Kührová, Vojtěch Mlỳnskỳ, Marie Zgarbová, Miroslav Krepl, Giovanni Bussi, Robert B Best, Michal Otyepka, Jiri Šponer, and Pavel Banáš. Improving the performance of the AMBER RNA force field by tuning the hydrogen-bonding interactions. *J. Chem. Theory Comput.*, 15(5):3288–3305, 2019.

[26] Thorben Fröhlking, Vojtěch Mlỳnskỳ, Michal Janeček, Petra Kührová, Miroslav Krepl, Pavel Banáš, Jiří Šponer, and Giovanni Bussi. Automatic learning of hydrogen-bond fixes in an AMBER RNA force field. *J. Chem. Theory Comput.*, 18(7):4490–4502, 2022.

[27] Alexander D MacKerell Jr, Joanna Wiorkiewicz-Kuczera, and Martin Karplus. An all-atom empirical energy function for the simulation of nucleic acids. *J. Am. Chem. Soc.*, 117:11946–11975, 1995.

[28] Raviprasad Aduri, Brian T Psciuk, Pirro Saro, Hariprakash Taniga, H Bernhard Schlegel, and John SantaLucia. AMBER force field parameters for the naturally occurring modified nucleosides in RNA. *J. Chem. Theory Comput.*, 3(4):1464–1475, 2007.

[29] You Xu, Kenno Vanommeslaeghe, Alexey Aleksandrov, Alexander D MacKerell Jr, and Lennart Nilsson. Additive CHARMM force field for naturally occurring modified ribonucleotides. *J. Comput. Chem.*, 37(10):896–912, 2016.

[30] Indrajit Deb, Joanna Sarzynska, Lennart Nilsson, and Ansuman Lahiri. Conformational preferences of modified uridines: comparison of AMBER derived force fields. *J. Chem. Inf. Model.*, 54:1129–1142, 2014.

[31] Indrajit Deb, Rupak Pal, Joanna Sarzynska, and Ansuman Lahiri. Reparameterizations of the χ torsion and Lennard-Jones σ parameters improve the conformational characteristics of modified uridines. *J. Comput. Chem.*, 37:1576–1588, 2016.

[32] Nivedita Dutta, Joanna Sarzynska, and Ansuman Lahiri. Molecular dynamics simulation of the conformational preferences of pseudouridine derivatives: improving the distribution in the glycosidic torsion space. *J. Chem. Inf. Model.*, 60:4995–5002, 2020.

[33] Nivedita Dutta, Indrajit Deb, Joanna Sarzynska, and Ansuman Lahiri. Data-informed reparameterization of modified RNA and the effect of explicit water models: application to pseudouridine and derivatives. *J. Comput. Aided Mol. Des.*, 36:205–224, 2022.

[34] R.J. Woods and R. Chappelle. Restrained electrostatic potential atomic partial charges for condensed-phase simulations of carbohydrates. *Theochem*, 527(1-3):149–156, 2000. Author manuscript; available in PMC 2014 Oct 9. Published in final edited form as: Theochem. 2000 Aug; 527(1-3): 149–156.

[35] Travis Hurst and Shi-Jie Chen. Deciphering nucleotide modification-induced structure and stability changes. *RNA Biol.*, 18(11):1920–1930, 2021.

[36] Caroline Roost, Stephen R Lynch, Pedro J Batista, Kun Qu, Howard Y Chang, and Eric T Kool. Structure and thermodynamics of N6-methyladenosine in RNA: a spring-loaded base modification. *J. Am. Chem. Soc.*, 137(5):2107–2115, 2015.

[37] Elzbieta Kierzek, Xiaoju Zhang, Richard M Watson, Scott D Kennedy, Marta Szabat, Ryszard Kierzek, and David H Mathews. Secondary structure prediction for RNA sequences including N6-methyladenosine. *Nat. Commun.*, 13(1):1–10, 2022.

[38] Anna Pavlova, Jerry M Parks, Adegboyega K Oyelere, and James C Gumbart. Toward the rational design of macrolide antibiotics to combat resistance. *Chem. Biol. Drug. Des.*, 90:641–652, 2017.

[39] Artem Babaian, Katharina Rothe, Dylan Girodat, et al. Loss of m1acp3ψ ribosomal RNA modification is a major feature of cancer. *Cell Rep.*, 31:107611, 2020.

[40] Manuel Bañó-Polo, Carlos Baeza-Delgado, Silvia Tamborero, et al. Transmembrane but not soluble helices fold inside the ribosome tunnel. *Nat. Commun.*, 9:1–9, 2018.

[41] Yaozong Li, Rajiv Kumar Bedi, Lars Wiedmer, Danzhi Huang, Paweł Sledz, and Amedeo Caflisch. Flexible binding of m6A reader protein YTHDC1 to its preferred RNA motif. *J. Chem. Theory Comput.*, 15(12):7004–7014, 2019.

[42] Wenxue Zhou, Zhongjie Han, Zhixiang Wu, et al. Specific recognition between YTHDF3 and m^{6}A-modified RNA: An all-atom molecular dynamics simulation study. *Proteins: Struct. Funct. Bioinf.*, 90:1965–1972, 2022.

[43] Gayathri Govindaraju, Rajashekar Varma Kadumuri, Devadathan Valiyamangalath Sethumadhavan, et al. N6-adenosine methylation on mrna is recognized by yth2 domain protein of human malaria parasite plasmodium falciparum. *Epigenetics Chromatin*, 13:1–15, 2020.

[44] Juan C Gonzalez-Rivera, Asuka A Orr, Sean M Engels, et al. Computational evolution of an RNA-binding protein towards enhanced oxidized-RNA binding. *Comput. Struct. Biotechnol. J.*, 18:137–152, 2020.

[45] Pierre-Yves Colin and Paul A Dalby. Functional and computational identification of a rescue mutation near the active site of an mRNA methyltransferase. *Sci. Rep.*, 10:1–13, 2020.

[46] Laura R Ganser, Megan L Kelly, Daniel Herschlag, and Hashim M Al-Hashimi. The roles of structural dynamics in the cellular functions of RNAs. *Nat. Rev. Mol. Cell Biol.*, 20:474–489, 2019.

[47] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314:141–151, 1999.

[48] Lingle Wang, Richard A Friesner, and BJ Berne. Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). *J. Phys. Chem. B*, 115:9431–9438, 2011.

[49] Glenn M Torrie and John P Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys,*, 23:187–199, 1977.

[50] Shankar Kumar, John M Rosenberg, Djamal Bouzida, et al. The weighted histogram analysis method for free-energy calculations on biomolecules. I. the method. *J. Comput. Chem.*, 13:1011–1021, 1992.

[51] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.*, 99:12562–12566, 2002.

[52] Jérôme Hénin, Tony Lelièvre, Michael R Shirts, et al. Enhanced sampling methods for molecular dynamics simulations [article v1.0]. *Living J. Comp. Mol. Sci.*, 4:1583, 2022.

[53] Vojtěch Mlỳnskỳ and Giovanni Bussi. Exploring RNA structure and dynamics through enhanced sampling simulations. *Curr. Opin. Struct. Biol*, 49:63–71, 2018.

[54] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.*, 100:020603, Jan 2008.

[55] Gareth A Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. Plumed 2: New feathers for an old bird. *Comput. Phys. Commun.*, 185(2):604–613, 2014.

[56] Volodymyr Babin, Christopher Roland, Thomas A Darden, and Celeste Sagui. The free energy landscape of small peptides as obtained from metadynamics with umbrella sampling corrections. *J. Chem. Phys.*, 125(20):204909, 2006.

[57] Alejandro Gil-Ley and Giovanni Bussi. Enhanced conformational sampling using replica exchange with collective-variable tempering. *Journal of Chemical Theory and Computation*, 11(3):1077–1085, 2015. PMID: 25838811.

[58] Hiroaki Fukunishi, Osamu Watanabe, and Shoji Takada. On the hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *The Journal of Chemical Physics*, 116:9058–9067, 2002.

[59] Antonia SJS Mey, Bryce K Allen, Hannah E Bruce Macdonald, John D Chodera, David F Hahn, Maximilian Kuhn, Julien Michel, David L Mobley, Levi N Naden, Samarjeet Prasad, et al. Best practices for alchemical free energy calculations [article v1.0]. *Living J. Comp. Mol. Sci.*, 2(1):18378, 2020.

[60] Yilin Meng, Danial Sabri Dashti, and Adrian E Roitberg. Computing alchemical free energy differences with hamiltonian replica exchange molecular dynamics (H-REMD) simulations. *J. Chem. Theory Comput.*, 7:2721–2727, 2011.

[61] Marc Souaille and Benoit Roux. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Comput. Phys. Commun.*, 135(1):40–57, 2001.

[62] Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.*, 129(12):124105, 2008.

[63] Zhiqiang Tan, Emilio Gallicchio, Mauro Lapelosa, and Ronald M. Levy. Theory of binless multi-state free energy estimation with applications to protein-ligand binding. *J. Chem. Phys.*, 136(14):144102, 2012.

[64] Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.*, pages 54–75, 1986.

[65] Charles H Bennett. Efficient estimation of free energy differences from Monte Carlo data. *J. Comp. Phys.*, 22(2):245–268, 1976.

[66] Di Wu and A. David Kofke. Phase-space overlap measures. ii. design and implementation of staging methods for free-energy calculations. *J. Chem. Phys.*, 123(8):084109, 2005.

[67] Wei-Tse Hsu, Valerio Piomponi, Pascal T. Merz, Giovanni Bussi, and Michael R. Shirts. Alchemical metadynamics: Adding alchemical variables to metadynamics to enhance sampling in free energy calculations. *Journal of Chemical Theory and Computation*, 19(6):1805–1817, 2023. PMID: 36853624.

[68] Bei Liu, Honglue Shi, Atul Rangadurai, Felix Nussbaumer, Chia-Chieh Chu, Kevin Andreas Erharter, David A Case, Christoph Kreutz, and Hashim M Al-Hashimi. A quantitative model predicts how m6A reshapes the kinetic landscape of nucleic acid hybridization and conformational transitions. *Nat. Commun.*, 12:5201, 2021.

[69] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P.N Vorontsov-Velyaminov. New approach to monte carlo calculation of the free energy: Method of expanded ensembles. *J. Chem. Phys.*, 96:1776–1783, 1992.

[70] Davide Branduardi, Giovanni Bussi, and Michele Parrinello. Metadynamics with adaptive gaussians. *J. Chem. Theory Comput.*, 8(7):2247–2254, 2012.

[71] Cristian Micheletti, Alessandro Laio, and Michele Parrinello. Reconstructing the density of states by history-dependent metadynamics. *Phys. Rev. Lett.*, 92(17):170601, 2004.

[72] Klaudia Mráziková, Vojtěch Mlýnský*, Petra Kührová, Pavlína Pokorná, Holger Kruse, Miroslav Krepl, Michal Otyepka, Pavel Banáš*, and Jiří Šponer*. Uucg RNA tetraloop as a formidable force-field challenge for MD simulations. *J. Chem. Theory Comput.*, 16(12):7601–7617, 2020.

[73] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957.

[74] Sabine Reißer, Silvia Zucchelli, Stefano Gustincich, and Giovanni Bussi. Conformational ensembles of an rna hairpin using molecular dynamics and sparse nmr data. *Nucleic Acids Research*, 48(3):1164–1174, 2020.

[75] Mattia Bernetti, Kathleen B Hall, and Giovanni Bussi. Reweighting of molecular simulations with explicit-solvent saxs restraints elucidates ion-dependent rna ensembles. *Nucleic Acids Research*, 49(14):e84, 2021.

[76] Massimiliano Bonomi, Riccardo Pellarin, and Michele Vendruscolo. Simultaneous determination of protein structure and dynamics using cryo-electron microscopy. *Biophysical Journal*, 114(7):1604–1613, 2018.

[77] R.W. Holley, J. Apgar, G.A. Everett, J.T. Madison, M. Marquisee, S.H. Merrill, J.R. Penswick, and A. Zamir. Structure of a ribonucleic acid. *Science*, 1965.

[78] Kazuko Nishikura. A-to-i editing of coding and non-coding rnas by adars. *Nature Reviews Molecular Cell Biology*, 17(2):83–96, 2016.

[79] Sundaramoorthy Srinivasan, Adrian Gabriel Torres, and Lluís Ribas de Pouplana. Inosine in biology and disease. *Genes*, 12:600, 2021.

[80] Adrian Gabriel Torres, Marta Rodríguez-Escribà, Marina Marcet-Houben, Helaine Graziele Santos Vieira, Noelia Camacho, Helena Catena, Marina Murillo Recio, Àlbert Rafels-Ybern, Oscar Reina, and Francisco Miguel Torres. Human trnas with inosine 34 are essential to efficiently translate eukarya-specific low-complexity proteins. *Nucleic Acids Research*, 49(12):7011–7034, 2021.

[81] Yusuke Shiromoto, Masayuki Sakurai, Moeko Minakuchi, Kentaro Ariyoshi, and Kazuko Nishikura. Adar1 rna editing enzyme regulates r-loop formation and genome stability at telomeres in cancer cells. *Nature Communications*, 12(Article number):1654, 2021.

[82] Nad’a Špačková and Kamila Réblová. Role of inosine–uracil base pairs in the canonical rna duplexes. *Genes (Basel)*, 9:324, 2018.

[83] Asem Alenaizan, Joshua L Barnett, Nicholas V Hud, C David Sherrill, and Anton S Petrov. The proto-nucleic acid builder: a software tool for constructing nucleic acid analogs. *Nucleic Acids Res.*, 49(1):79–89, 2021.

[84] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

[85] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[86] Alberto Pérez, Iván Marchán, Daniel Svozil, Jiri Šponer, Thomas E Cheatham III, Charles A Laughton, and Modesto Orozco. Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophys. J.*, 92(11):3817–3829, 2007.

[87] Marie Zgarbová, Michal Otyepka, Jiří Šponer, Arnošt Mládek, Pavel Banáš, Thomas E. Cheatham, and Petr Jurečka. Refinement of the cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theory Comput.*, 7(9):2886–2902, 2011.

[88] In Suk Joung and Thomas E Cheatham III. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B*, 112(30):9020–9041, 2008.

[89] N Goga, AJ Rzepiela, AH De Vries, SJ Marrink, and HJC Berendsen. Efficient algorithms for Langevin and DPD dynamics. *J. Chem. Theory Comput.*, 8(10):3637–3649, 2012.

[90] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52(12):7182–7190, 1981.

[91] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

[92] Mattia Bernetti and Giovanni Bussi. Pressure control using stochastic cell rescaling. *J. Chem. Phys.*, 153(11):114107, 2020.

[93] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[94] Ming Huang, Timothy J. Giese, Tai-Sung Lee, and Darrin M. York. Improvement of dna and rna sugar pucker profiles from semiempirical quantum methods. *Journal of Chemical Theory and Computation*, 2021.

[95] David E Condon, Scott D Kennedy, Brendan C Mort, Ryszard Kierzek, Ilyas Yildirim, and Douglas H Turner. Stacking in RNA: NMR of four tetramers benchmark molecular dynamics. *J. Chem. Theory Comput.*, 11(6):2729–2742, 2015.

[96] John P. Marino, Harald Schwalbe, and Christian Griesinger. J-coupling restraints in rna structure determination. *Accounts of Chemical Research*, 32(7):614–623, 1999.

[97] David B. Davies. Conformations of nucleosides and nucleotides. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 12(3):135–225, 1978.

[98] Sandro Bottaro, Giovanni Bussi, Giovanni Pinamonti, Sabine Reiß er, Wouter Boomsma, and Kresten Lindorff-Larsen. Barnaba: software for analysis of nucleic acid structures and trajectories. *RNA*, 2(25):219–231, 2018.

[99] A. Guinier and G. Fourmet. *Small-Angle Scattering of X-rays*. John Wiley and Sons, New York, 1955.

[100] Mattia Bernetti and Giovanni Bussi. Comparing state-of-the-art approaches to back-calculate saxs spectra from atomistic molecular dynamics simulations. *Eur. Phys. J. B.*, 94(180), 2021.

[101] R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute, and K. Zakrzewska. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Research*, 37(17):5917–5929, 07 2009.

[102] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[103] Martin K. Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Pérez-Hernández, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noé. Pyemma 2: A software package for estimation, validation, and analysis of markov models. *Journal of Chemical Theory and Computation*, 11(11):5525–5542, 2015. PMID: 26574340.

[104] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.*, (57):289–300, 1995.

[105] Nicola Calonaci, Mattia Bernetti, Alisha Jones, Michael Sattler, and Giovanni Bussi. Molecular dynamics simulations with grand-canonical reweighting suggest cooperativity effects in rna structure probing experiments. *Journal of Chemical Theory and Computation*, 19(12):3672–3685, 2023. PMID: 37288967.

[106] Wendy V Gilbert, Tristan A Bell, and Cassandra Schaening. Messenger RNA modifications: form, distribution, and function. *Science*, 352(6292):1408–1412, 2016.

[107] Emily M Harcourt, Anna M Kietrys, and Eric T Kool. Chemical and structural effects of base modifications in messenger RNA. *Nature*, 541(7637):339–346, 2017.

[108] Deepak P Patil, Chun-Kan Chen, Brian F Pickering, Amy Chow, Constanza Jackson, Mitchell Guttman, and Samie R Jaffrey. m6A RNA methylation promotes XIST-mediated transcriptional repression. *Nature*, 537(7620):369–373, 2016.

[109] P Cody He and Chuan He. m6A RNA methylation: from mechanisms to therapeutic potential. *EMBO J.*, 40(3):e105977, 2021.

[110] Yaozong Li, Rajiv Kumar Bedi, Lars Wiedmer, Xianqiang Sun, Danzhi Huang, and Amedeo Caflisch. Atomistic and thermodynamic analysis of N6-methyladenosine (m6a) recognition by the reader domain of YTHDC1. *J. Chem. Theory Comput.*, 17(2):1240–1249, 2021.

[111] Yaozong Li, Rajiv Kumar Bedi, Francesco Nai, et al. Structure-based design of ligands of the m^{6}A-RNA reader YTHDC1. *Eur. J. Med. Chem.*, 5:100057, 2022.

[112] Miroslav Krepl, Fred Franz Damberger, Christine von Schroetter, Dominik Theler, P. Pokorná, F. H.-T. Allain, and J. Šponer. Recognition of N6-methyladenosine by the YTHDC1 YTH domain studied by molecular dynamics and NMR spectroscopy: The role of hydration. *J. Phys. Chem. B*, 125(28):7691–7705, 2021.

[113] Valerio Piomponi, Thorben Fröhlking, Mattia Bernetti, and Giovanni Bussi. Molecular simulations matching denaturation experiments for N6-methyladenosine. *ACS Cent. Sci.*, 8:1218–1228, 2022.

[114] Andrea Cesari, Sandro Bottaro, Kresten Lindorff-Larsen, Pavel Banáš, Jiri Šponer, and Giovanni Bussi. Fitting corrections to an RNA force field using experimental data. *J. Chem. Theory Comput.*, 15(6):3425–3431, 2019.

[115] Giovanni Bussi and Alessandro Laio. Using metadynamics to explore complex free-energy landscapes. *Nature Rev. Phys.*, 2(4):200–212, 2020.

[116] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23(4):550–560, 1997.

[117] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods*, 17(3):261–272, 2020.

[118] Leslie Kish. *Survey sampling*. John Wiley & Sons, New York, 1965.

[119] Ramya Rangan, Massimiliano Bonomi, Gabriella T Heller, Andrea Cesari, Giovanni Bussi, and Michele Vendruscolo. Determination of structural ensembles of proteins: restraining vs reweighting. *J. Chem. Theory Comput.*, 14(12):6632–6641, 2018.

[120] Piotr Cieplak, Wendy D Cornell, Christopher Bayly, and Peter A Kollman. Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins. *J. Comput. Chem.*, 16(11):1357–1377, 1995.

[121] M.J.T. Frisch and et al. Gaussian 09, revision d01. *Gaussian Inc., Wallingford, CT*, 2009.

[122] Tianbing Xia, John SantaLucia, Mark E. Burkard, Ryszard Kierzek, Susan J. Schroeder, Xiaoqi Jiao, Christopher Cox, and Douglas H. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of rna duplexes with watson-crick base pairs. *Biochemistry*, 37(42):14719–14735, 1998.

[123] Jonathan L. Chen, Abigael L. Dishler, Scott D. Kennedy, Ilyas Yildirim, Biao Liu, Douglas H. Turner, and Martin J. Serra. Testing the nearest neighbor model for canonical RNA base pairs: Revision of GU parameters. *Biochemistry*, 51(16):3508–3522, 2012.

[124] Anders B Norgaard, Jesper Ferkinghoff-Borg, and Kresten Lindorff-Larsen. Experimental parameterization of an energy function for the simulation of unfolded proteins. *Biophys. J.*, 94(1):182–192, 2008.

[125] Da-Wei Li and Rafael Brüschweiler. Iterative optimization of molecular mechanics force fields from NMR data of full-length proteins. *J. Chem. Theory Comput.*, 7(6):1773–1782, 2011.

[126] Susan J Schroeder and Douglas H Turner. Optical melting measurements of nucleic acid thermodynamics. In *Meth. Enzymol.*, volume 468, pages 371–387. Elsevier, 2009.

[127] Michal Janeček, Petra Kührová, Vojtěch Mlýnský, Michal Otyepka, Jiri Šponer, and Pavel Banáš. W-RESP: Well-restrained electrostatic potential-derived charges. revisiting the charge derivation model. *J. Chem. Theory Comput.*, 17(6):3495–3509, 2021.

[128] U Chandra Singh and Peter A Kollman. An approach to computing electrostatic charges for molecules. *J. Comput. Chem.*, 5(2):129–145, 1984.

[129] Shun Sakuraba, Kiyoshi Asai, and Tomoshi Kameda. Predicting RNA duplex dimerization free-energy changes upon mutations using molecular dynamics simulations. *J. Phys. Chem. Lett.*, 6(21):4348–4351, 2015.

[130] Sandro Bottaro, Giovanni Bussi, Scott D Kennedy, Douglas H Turner, and Kresten Lindorff-Larsen. Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations. *Sci. Adv.*, 4(5):eaar8521, 2018.

[131] Kevin M Guckian, Barbara A Schweitzer, Rex X-F Ren, Charles J Sheils, Deborah C Tahmassebi, and Eric T Kool. Factors contributing to aromatic stacking in water: evaluation in the context of DNA. *J. Am. Chem. Soc.*, 122(10):2213–2222, 2000.

[132] Johan Isaksson and Jyoti Chattopadhyaya. A uniform mechanism correlating dangling-end stabilization and stacking geometry. *Biochemistry*, 44(14):5390–5401, 2005.

[133] Michele Invernizzi and Michele Parrinello. Exploration vs convergence speed in adaptive-bias enhanced sampling. *Journal of Chemical Theory and Computation*, 2022.

[134] Michael Deighan, Massimiliano Bonomi, and Jim Pfaendtner. Efficient simulation of explicitly solvated proteins in the well-tempered ensemble. *J. Chem. Theory Comput.*, 8(7):2189–2192, 2012.

[135] Christophe Chipot and Tony Lelièvre. Enhanced sampling of multidimensional free-energy landscapes using adaptive biasing forces. *SIAM J. Appl. Math.*, 71(5):1673–1695, 2011.

[136] James F Dama, Grant Rotskoff, Michele Parrinello, and Gregory A Voth. Transition-tempered metadynamics: robust, convergent metadynamics via on-the-fly transition barrier estimation. *J. Chem. Theory Comput.*, 10(9):3626–3633, 2014.

[137] Alisha N Jones, Ekaterina Tikhaia, André Mourão, and Michael Sattler. Structural effects of m6A modification of the Xist A-repeat AUCG tetraloop and its recognition by YTHDC1. *Nucleic Acids Res.*, 50(4):2350–2362, 2022.

[138] X. Wang, Z. Lu, A. Gomez, G. C. Hon, Y. Yue, D. Han, Y. Fu, M. Parisien, Q. Dai, G. Jia, et al. N6-methyladenosine-dependent regulation of messenger rna stability. *Nature*, 505:117, 2014.

[139] Z. Zhang, D. Theler, K. H. Kaminska, M. Hiller, P. de la Grange, R. Pudimat, I. Rafalska, B. Heinrich, J. M. Bujnicki, F. H.-T. Allain, et al. The yth domain is a novel rna binding domain. *Journal of Biological Chemistry*, 285:14701–14710, 2010.

[140] S. Luo and L. Tong. Molecular basis for the recognition of methylated adenines in rna by the eukaryotic yth domain. *Proceedings of the National Academy of Sciences of the United States of America*, 111:13834–13839, 2014.

[141] David Theler, Cyril Dominguez, Markus Blatter, Julien Boudet, and Frédéric H.-T. Allain. Solution structure of the yth domain in complex with n6-methyladenosine rna: A reader of methylated rna. *Nucleic Acids Research*, 42:13911–13919, 2014.

[142] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. The missing term in effective pair potentials. *The Journal of Physical Chemistry A*, 91:6269–6271, 1987.

[143] Saeed Izadi, Ramu Anandakrishnan, and Alexey V. Onufriev. Building water models: A different approach. *Journal of Physical Chemistry Letters*, 5(21):3863–3871, 2014.