SCUOLA INTERNAZIONALE SUPERIORE DI STUDI AVANZATI (SISSA)
INTERNATIONAL SCHOOL FOR ADVANCED STUDIES

Pre-mRNA Splicing: An Evolutionary Computational Journey from Ribozymes to Spliceosome

MOLECULAR AND STATISTICAL BIOPHYSICS SECTOR

Thesis submitted for the degree of
Philosophiae Doctor
in
PHYSICS AND CHEMISTRY OF BIOLOGICAL SYSTEMS

Candidate:
Lorenzo Casalino

Supervisor:
Dr. Alessandra Magistrato

October 18th, 2017, SISSA - via Bonomea, 265 - 34136 Trieste (Italy)

To my family.

Table of contents

List of abbreviations

3'SS	3'-Splice Site
5'SS	5'-Splice Site
BSSE	Basis Set Superposition Error
BPS	Branch-Point Sequence
CDA	Cationic Dummy Atom
CT	Charge Transfer
CN	Coordination Number
CP	Car-Parrinello
CPs	Coordination Patterns (Chapter 5)
DFT	Density Functional Theory
EBS	Exon-Binding Site
FF/ff	Force Field
G2IR	Group II Intron Ribozyme / Group II Intron
HDV	Hepatitis Delta Virus / Hepatitis Delta Virus Ribozyme
IBS	Intron-Binding Site
IEP	Intron-Encoded Protein
ILS	Intron-Lariat-Spliceosome
KS	Kohn-Sham
LCR	Ligand Charge Rearrangement
LJ	Lennard-Jones
MD	Molecular Dynamics
N-t	N-terminal domain
NBO	Natural Bond Orbital
NTC	NineTeen Complex
ncRNA	noncoding RNA
ORF	Open Reading Frame
pre-mRNA	pre-mature messenger RNA
PDB	Protein Data Bank
QM	Quantum Mechanics
QM/MM	Quantum Mechanics/Molecular Mechanics
RT	Reverse Transcriptase
RNP	Ribonucleoprotein
snRNP	small nuclear Ribonucleoprotein
SPL	Spliceosome
vdW	van der Waals

1 Preface: An evolutionary computational journey

It was in 1978 when Walter Gilbert, an American biochemist, Nobel Prize in chemistry in 1980, manifested all his excitement for the introns revolution in the memorable *Nature* letter "*Why genes in pieces?*" [1]. In this famous script, not only he coined for the first time the terms introns and exons, but also took stock of the recent discoveries, which marked the beginning of a new revolutionary era, the splicing era.

"… a transcription unit containing regions which will be lost from the mature messenger – which I suggest we call introns (for intragenic regions) – alternating with regions which will be expressed – exons".

Walter Gilbert, 1978

The intron–exon organization of the genes is nowadays taken for granted and constitutes a fully established theory. DNA protein-coding sequences (exons) are not contiguous, but rather separated by silent intervening fragments (introns), which must be removed in a process called pre-mRNA splicing. However, this fragmented composition of the eukaryotic genome has ancient origins. It appears that, during the initial stages of eukaryotic evolution, group II introns, i.e. self-splicing catalytic ribozymes, invaded the eukaryotic genome via the endosymbiosis of an *α-proteobacterium* in an archaeal host. Interestingly, *α-proteobacteria* are considered the ancestors of eukaryotic mitochondria. Upon endosymbiosis, group II introns were released and invaded the host genome. Afterwards, they split into the inert spliceosomal introns and the catalytically active small nuclear (sn)RNAs, which, together with additional splicing factors, gave rise to the eukaryotic spliceosome. This

marked the transition from the autocatalytic splicing, mediated by ribozymes (RNA filaments endowed with an intrinsic catalytic activity) to splicing mediated by a protein-RNA machinery, the spliceosome [2].

In my Ph.D. years, I have tried to retrace the evolutionary relationship between group II introns and the spliceosome from a computational perspective, studying the splicing process of these two different – but mechanistically related – large and sophisticated biomolecules. Classical molecular dynamics simulations (MD), quantum mechanics calculations (QM) and combined quantum-classical simulations (QM/MM) have been the "vessels" of this exciting evolutionary computational journey from group II introns to spliceosome.

In **Chapter 2** I introduce the importance of Mg^{2+} ions in the RNA biology, not only as catalytic cofactors, but also as essential structural and functional elements for RNA filaments. Moreover, I present the structural and molecular biology of two main "stops" of this journey, group II intron ribozymes and the spliceosome machinery, also focusing on their evolutionary links.

Chapter 3 consists of a brief review of all the computational techniques that I have used in this thesis, from classical MD to QM and QM/MM simulations and enhanced sampling methods aimed at reconstructing the free energy of a process.

Chapter 4 is entirely dedicated to the splicing mechanism promoted by group II intron ribozymes. It represents the starting point of the evolutionary journey. In this chapter I report a QM/MM study of the molecular mechanism of group II introns first-step hydrolytic splicing. This research was published in 2016 in the *Journal of the American Chemical Society* and was conducted for six months at EPFL (Switzerland) under the co-supervision of Prof. Ursula Röthlisberger.

Chapter 5 is focused on Mg^{2+} ions, which are the natural cofactors of splicing, both in group II introns and the spliceosome. Mg^{2+} is the most abundant intracellular divalent cation, taking part in many chemical reactions involving nucleic acids. In this chapter I present a study, published in 2017 on the *Journal of Chemical Theory and Computation*, in which I focused on Mg^{2+}/RNA interplay, benchmarking the performance of different force fields currently used to describe Mg^{2+} in MD simulations of large RNA molecules. A group II intron is used as prototype of a large RNA molecule binding Mg^{2+}. The non-trivial electronic effects induced by Mg^{2+} on its ligands, such as charge transfer and polarization, are also characterized and discussed. The study offers some guidelines on Mg^{2+} force fields for users and developers.

Chapter 6 represents the final stop of the evolutionary journey. In this chapter I present a MD simulations study based on the first cryo-EM structure of a yeast spliceosome solved at near-atomistic resolution. This offered precious information on the catalytic site as well as on the main proteins and snRNAs involved in the pre-

mRNA splicing in eukaryotes. In particular, I have investigated the structural and dynamical properties of the spliceosome machinery at atomistic level, with a particular emphasis on protein/RNA interplay through the characterization of their principal motions.

2 Biological introduction

2.1 Magnesium ion in biological systems

Magnesium ion (Mg^{2+}) is the dominant divalent cation in biological systems. It is the second most abundant intracellular cation and the fourth in the body [3]. Mg^{2+} acts as a cofactor for about 300 cellular enzymatic reactions, from metabolism of proteins and nucleic acids to specific actions in different organs, mostly affecting the neuromuscular and cardiovascular systems [4-6]. Inside the cell Mg^{2+} is present in three different states: (i) free ionized form, which constitutes only the 0.5–5% of the total cellular magnesium; (ii) bound to anionic compounds such as ATP, ADP, citrate, proteins, RNA and DNA, and (iii) sequestered within mitochondria and endoplasmic reticulum. The free intracellular concentration is maintained in the mM range (~0.5 mM), with heterogeneous distribution in the cytoplasm [7]. Magnesium ion has a key role in many important biological processes such as cellular energy metabolism, cell replication, muscle contraction, maintenance of cellular ionic balance, protein synthesis and RNA processing [8]. When exerting its functions, magnesium can either bind to ligands such as ATP in ATP-requiring enzymes (ATPases [9], ATP synthases [10]) or to the active site of an enzyme, like in enolases, pyruvate kinases, RNases and pyrophosphatases [11]. It is also capable of promoting conformational changes during catalytic/transport processes, as in Na^+/K^+-ATPase [5]. Moreover, and most importantly for this thesis, magnesium ion strikingly affects the biology of nucleic acids.

2.1.1 Magnesium ion in RNA biology

Mg^{2+} is an alkaline earth metal ion, found in the third period of the periodic table. It has an ionic radius of 72 pm with a high charge density and a coordination number (CN) of 6, such determining an octahedral geometry of its ligands, which are placed at a distance of about 204 pm [12]. All these properties affect the strength and the geometry of the Mg^{2+} complexes, especially those involving nucleic acids. RNA molecules are characterized by a highly negative charge of the sugar-phosphate backbone. Thus, their interactions with metals range from general charge shielding, due to a mobile counterion atmosphere, to specific coordination sites, which are formed when they come in close contact with RNA filaments during their folding into a functional tertiary structure. Monovalent cations (Na^+, K^+) can screen the negative charge of RNA chains, but divalent cations (Ca^{2+}, Mg^{2+}) are more effective in reducing the electrostatic repulsion between phosphate crowded regions [13]. Moreover, balancing the RNA negative charge with divalent cations is entropically favored with respect to monovalent ones, since half the number of ions is required with the former. Mg^{2+} performs this task more efficiently than other divalent cations as it can optimally fit into the RNA compact sites, screening the backbone charges, regulating, accelerating and even tuning the free energy landscape of folding processes [14]. Mg^{2+} ions markedly stabilize the RNA folded structure and form important sites for RNA functions [15].

Each Mg^{2+} ion in solution can bind and orients six water molecules with a large enthalpy of hydration and exerting large polarization effects. Then, hydrated Mg^{2+} ions can interact with the electrostatic field of the RNA, either at large distances (diffuse ions) or via second-shell interactions (through H-bonds of the coordinated water molecules with RNA atoms). This latter binding mode is commonly referred to as *outer-sphere* binding (Figure 1.1a). It is believed that *outer-sphere* binding mostly contributes to drive the folding and to stabilize the RNA tertiary structure as there is no dehydration penalty in this case [16]. Instead, upon dehydration, Mg^{2+} ions can directly associate to RNA electron rich groups (imino nitrogens, keto oxygens from nucleobases and phosphate and sugar oxygens) (Figure 1.1b), forming the so-called *inner-sphere* binding sites. Mg^{2+} ions are usually chelated by one to four (rarer) RNA ligands. Highly RNA-chelated Mg^{2+} sites are mostly present in metal-dependent ribozymes because they are necessary to assume specific coordination geometries of reactive groups. In fact, RNA is capable of catalyzing essential biochemical reactions despite possessing significant biophysical handicaps, such as the small repertoire of functional groups, the highly negative charge exhibited by the phosphates and the constrained RNA backbone which forms the active site [17].

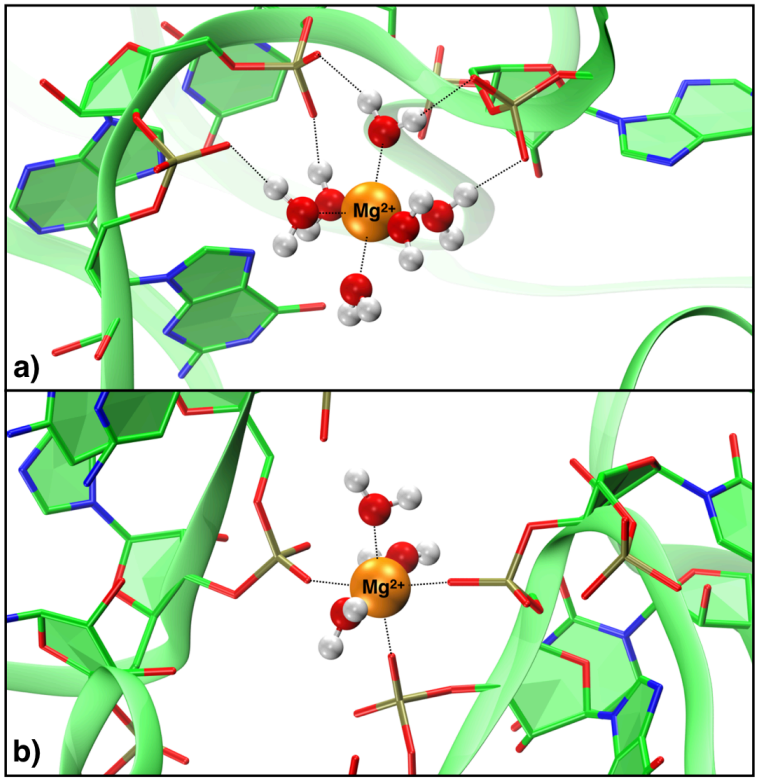


Figure 1.1. *Outer-sphere* (a) and *inner-sphere* (b) Mg^{2+} binding sites. Diffuse ions (not shown) are separated by multiple layers of water molecules, but feel the long-range interactions with RNA, here represented in green new ribbons. Mg^{2+} is depicted with orange spheres, whereas oxygen, phosphorus, nitrogen and hydrogen atoms are colored in red, gold, blue and white, respectively.

2.1.2 Magnesium ion and catalytic RNAs

Ribozymes are RNA molecules that act as chemical catalysts. They participate to some of the most important processes taking place in the cells [17], among which peptide synthesis, splicing of pre-mature messenger RNA, and formation of mature transfer RNA. The above-mentioned events are respectively regulated by the ribosome (Figure 1.2), which catalyzes peptidyl-transfer reactions, and by naturally occurring large ribozymes (self-splicing group I and II introns, and RNase P), which promote phosphoryl-transfer reactions leading to RNA cleavage and ligation. These last processes are usually dependent on a two-Mg^{2+}-ion mechanism (detailed in the next paragraph) in analogy to RNase H, alkaline phosphatase and DNA polymerase enzymes [11, 18-21].

Also small nucleolytic ribozymes exist, including, among others, the hepatitis delta virus (HDV) and the hammerhead (HHR) [22]. The former is present in circular subviral RNAs, while the latter was originally found in vironoids and autonomous

subviral plants pathogens. Both ribozymes are required for processing the unit length viral genome during rolling cycle replication and they are believed to catalyze transesterification reactions supported by Mg^{2+} ions and RNA bases with atypical pK_a values [22].

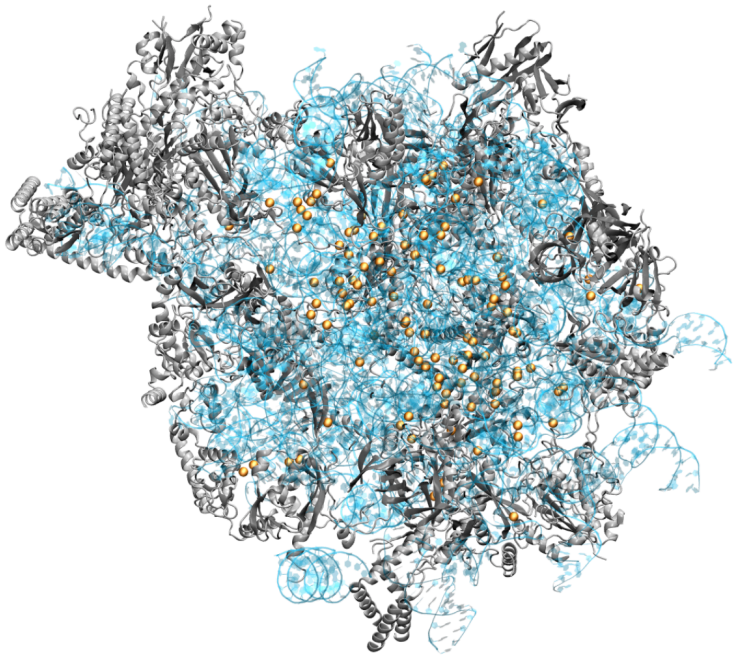


Figure 2.1. Structure of the yeast mitochondrial large ribosomal unit from PDB entry: 3J6B. Mg^{2+} ions are shown as orange spheres. RNA and proteins are shown in light blue and grey, respectively.

Most of the cited reactive RNAs are metal-dependent ribozymes, which achieve their optimal catalytic efficiency only in the presence of Mg^{2+} ions. As such, in crystallographic studies these metals are often replaced by ions which impair catalysis (i.e. Ca^{2+}, Sm^{2+}, Sr^{2+}). For this reason, accurate structural information on the catalytic competent state of Mg^{2+}-dependent ribozymes is often lacking [23]. Moreover, Mg^{2+} is silent to most of the spectroscopic techniques, thus hampering an unambiguous characterization of the coordination sites. In these cases, molecular simulations may fill the experimental voids by providing atomistic detailed data on the catalytic competent form of the ribozyme in the presence of the active metal ions and also on its structural, dynamical and functional properties. While the structural/dynamical interplay between Mg^{2+} and RNA has not been exhaustively investigated by simulations studies mostly due to force fields inaccuracies, the role of Mg^{2+} ions in RNA catalysis has been extensively addressed. In fact, ribozymes attracted a large theoretical interest due to their importance in biology, for their evolutionary

significance and for their potential implications in drug discovery and biomedical technology [24].

2.1.3 The two-Mg^{2+}-ion aided mechanism

The catalytic activity of self-cleaving ribozymes is believed to depend on a S_N2-like two-Mg^{2+}-ion mechanism as proposed by Steitz and Steitz (Figure 1.3) [21]. According to this hypothesis, the two metals act as Lewis' acid, one activating the nucleophile (making the proton of the nucleophilic water or 2'-OH group more acidic), and the other facilitating the cleavage by stabilizing the leaving group. They both orient the substrate to undergo the nucleophilic attack and stabilize a pentacovalent transition state [11, 18].

This mechanism, originally proposed on the basis of the crystal structure of the DNA polymerase I 3',5'-exonuclease domain complexed with single-stranded DNA or product, has been postulated to occur also on group I and II introns ribozymes [25, 26] and ribonuclease P [27] on the basis of their crystal structures. Pure QM studies carried out on small model systems of the active site with one or two metal ion(s) established that the presence of two metals is necessary to reduce the free energy barrier of the transesterification process [28]. However, only a recent study performed by me [29], and reported in the Chapter 4 of this thesis, has unveiled an RNA-adapted two-Mg^{2+}-ion mechanism for the first-step of the hydrolytic splicing catalyzed by group II intron ribozymes.

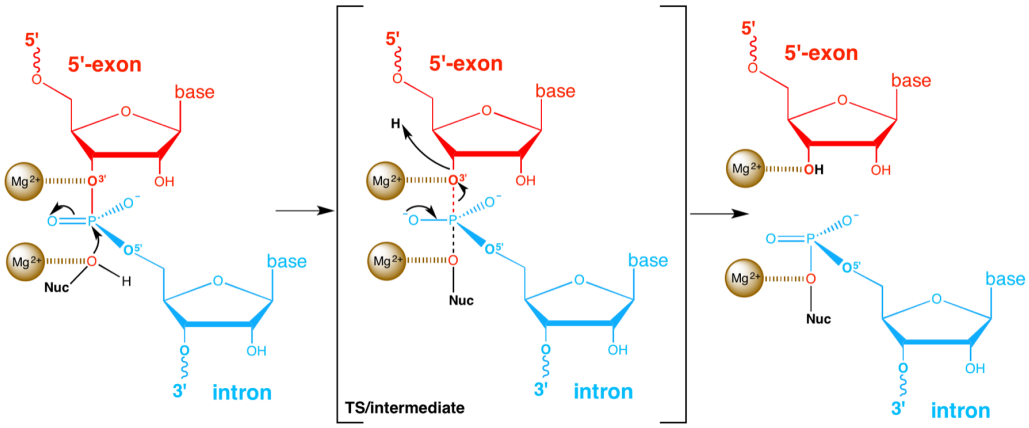


Figure 1.3. The Steitz and Steitz's two-Mg^{2+}-ion S_N2-like mechanism proposed for group I/II intron ribozymes. Intron and 5'-exon are colored in blue and red, respectively.

For group I introns, the role of a third metal ion in catalysis was also speculated [22], while a four metal ions cluster was observed in group II intron (2 Mg^{2+} and 2 K^{+}) [30] and (four Mg^{2+}) in the recently solved yeast and human spliceosome [31-34].

2.2 RNA splicing: group II introns and spliceosome

After the discovery of the DNA, the structure of a gene was assumed to be a contiguous series of base pairs carrying the information for protein synthesis. This assumption, even if might be correct for many prokaryotes, could not explain some observations that emerged for eukaryotic organisms in the following years. First, the common scenario in which the large amount of DNA present in the nucleus could all encode for proteins started to be argued. Then in the '70s, an unusually long RNA found in the nucleus of vertebrate cells was compared with a shorter mRNA that emerged in the cytoplasm: they both were characterized by a cap structure at the 5'-end and a poly-A tract at the 3'-end. A comparison of the sequence of the mRNA and its corresponding nuclear DNA unveiled that some fragments were removed during the processing of the nuclear "pre-mature" filament of RNA. The advent of splicing solved many paradoxes and contributed to explain protein diversity in eukaryotes [35, 36].

Pre-mature messenger RNA (pre-mRNA) splicing is, in fact, a central step of gene expression, occurring between transcription and protein synthesis. It is a crucial biological process whereby large non-coding oligonucleotide sequences (introns) are removed from precursor mRNAs, while coding sequences (exons) are linked together in a functional mature mRNA filament. Introns are non-coding segments of a nascent mRNA transcript that are spliced out before the RNA molecule is translated into a protein by the ribosome [37]. Interestingly, many genes might have several introns such that their alternative splicing, involving a differential use of splice sites, leads to the production of multiple mRNAs from individual genes, which are then translated into different protein isoforms. This significantly contributes to expand the form and function of the eukaryotic proteome [38]. High-throughput RNA sequencing studies have suggested that alternative splicing occurs routinely in human cells, with 90–95% of genes being alternatively spliced.

In eukaryotes, the excision of introns from pre-mRNAs takes place in the nucleus and it is catalyzed by a massive ribonucleoprotein (RNP) complex called spliceosome, comprising five small nuclear RNAs (snRNAs) – U1, U2, U4, U5, and U6 – and approximately 150 proteins [39]. Each pre-mRNA splicing cycle entails two sequential transesterification reactions, in which two ends of an intron are brought in spatial proximity, resulting in a released exon and the formation of an intron lariat (Figure 1.4). In the first step the 2'-OH of an invariant adenosine nucleotide in the branch-point sequence (BPS) of an intron attacks as a nucleophile the phosphate of a guanine at the 5'-end of the 5'-splice site (5'SS), forming an intron lariat–3'-exon intermediate. In the second step, the 3'-OH at the 3'-end of the cleaved 5'-exon attacks as a

nucleophile the phosphorus atom at the 5'-end of the 3'-exon (3'-splicing site, 3'SS), yielding the ligated exons as product and releasing the intron lariat [39].

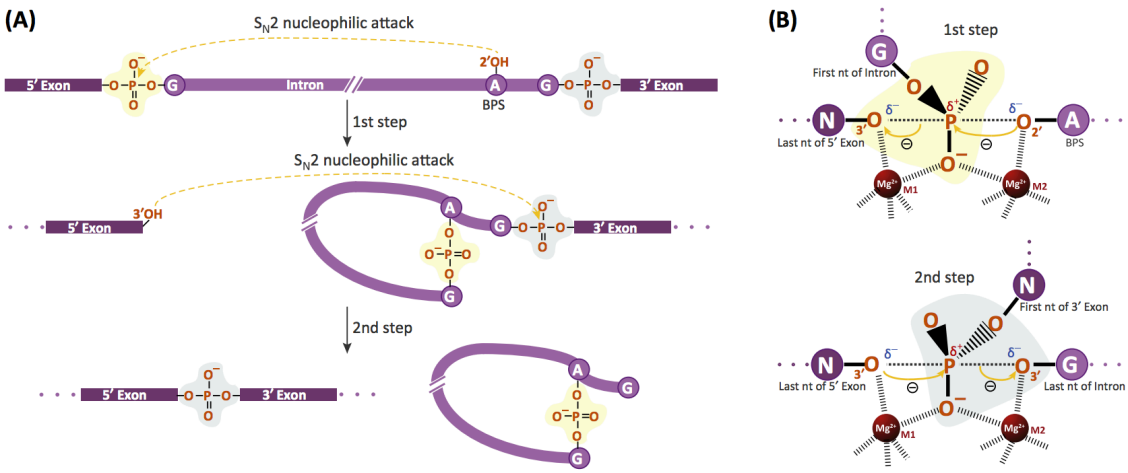


Figure 1.4. (A) The two steps of pre-mRNA splicing catalyzed by the eukaryotic spliceosome. (B) Stabilization of transition states by Mg^{2+} ions proposed for the two phosphoryl-transfer reactions. The figure was adapted from [39].

Over the past decade, remarkable progress has been made to isolate, purify, and characterize the protein composition and the biochemical activity of the spliceosome components along its functional cycle. Splicing requires extreme precision because even a single nucleotide addition or deletion at the site of exon joining will shift the reading frame with adverse consequences to the protein-coding potential. It turns out that more than 200 human diseases are caused by pre-mRNA splicing aberrations [40]. There are two main categories of mutations that can disrupt splicing or alternative splicing, causing defects and misregulation which are at the origin of many pathologies. i) *Cis*-acting mutations of even one single nucleotide on constitutive splice sites or alternative splice sites, which are therefore skipped or not correctly read by the spliceosome. The result is the expression of an unnatural mRNA which might result in diseases like familial isolated growth hormone deficiency type II, Frasier syndrome, frontotemporal dementia, parkinsonism linked to chromosome 17 and atypical cystic fibrosis. Instead, the ii) *trans*-acting splicing mutations affect the function of the basal splicing machinery or factors that regulate alternative splicing. These alterations cause a misregulation of the pre-mRNA splicing of all genes, resulting in diseases like *retinitis pigmentosa*, spinal muscular atrophy, myotonic dystrophy, neoplasia and malignancy (related to genes encoding for regulators of apoptosis, hormones, and receptors mediating cell–cell and cell–matrix interactions) [41].

2.2.1 Group II intron ribozymes

Group II introns (G2IRs) are widespread mobile ribozymes (enzymes made entirely of RNA) capable of self-splicing and retrotransposition (i.e., reverse splicing) reactions. These RNA molecules mostly contribute to RNA metabolism of prokaryotes (bacteria), but are also present in the genome of mitochondria and chloroplasts of lower eukaryotes (fungi, yeasts, plants). Although being absent in the genomes of higher eukaryotes, they are considered to share a common ancestor with the eukaryotic spliceosome machinery [42, 43]. Indeed, recent structural reconstructions and metal rescue experiments have demonstrated that G2IRs share a common catalytic site and mechanism with the U6 snRNA of the spliceosome, thus confirming their close evolutionary origins [32, 44]. These highly reactive retrotransposable RNA molecules have invaded the genomes of most life forms, contributing to their evolution and genomic diversity [42]. G2IRs consist of a catalytically active intron RNA, which performs self-splicing and an intron-encoded protein (IEP), which assists reverse splicing (retrotransposition) into a DNA target (*vide infra*). The IEP is a single-chain multidomain maturase, which promotes intron mobility by synthesizing a complementary DNA copy of the intron RNA (through a reverse transcriptase domain) [45].

2.2.1.1 G2IRs structure and classification

Group II introns fold into a conserved secondary structure, which spans 400-800 nucleotides and is characterized by six interacting domains, DI-VI. These domains interact to form a conserved tertiary structure that brings together distant sequences defining the active site. Domain I is the largest domain that functions as a scaffold and comprises about half of the ribozyme. It contains sequence motifs that base-pair with exon sequences to align them at the active site. The former are denoted as exon-binding sites (EBSs), while the latter as intron-binding sites (IBSs). Domain II and Domain III are smaller domains that have a structural role. Instead, the Open Reading Frame (ORF) encoding for the IEP protrudes from Domain IV, which projects away from the catalytic core, and contains a high-affinity binding site for the IEP. Domain V is the most important domain as it contains the catalytic site, binding the two Mg^{2+} ions to self-catalyze pre-mRNA splicing and reverse splicing reactions. Domain V is the heart of the ribozyme and features some important and highly conserved motifs: the so-called catalytic triad AGC (or CGC for some introns), the two-nucleotide bulge (AY, where Y = C or U) and the J2/3 junction. These motifs form a major-groove triple-helix, which is also found in the spliceosome active site. The two metal ions are coordinated to the phosphate groups of the AGC triad and AY bulge, and are separated by a distance consistent with a two-metal-ion aided catalysis [21]. Finally, Domain VI

contains the branch-point nucleotide, generally a bulged A, i.e. the branch site during the splicing reaction. Domain VI undergoes a conformational change to reposition the branching point between the two steps of splicing and it is therefore highly flexible [42, 45-47].

Three major classes, termed IIA, IIB and IIC, have been identified, displaying diversification of the architectural scaffold, protein interaction networks, and some aspects of reactivity [43]. The most prominent difference among IIA, IIB, and IIC ribozymes is the mechanism of exons recognition, because each class uses a distinct combination of base pairing interactions to recognize the 5' and 3' exons (i.e., different combinations of IBS1-EBS1, IBS2-EBS2, IBS3-EBS3). The group IIA and IIB are approximately 900 nt long and are found in bacteria, archaea, mitochondria and chloroplasts, while the introns belonging to the group IIC are shorter (approximately 400 nt) and they are present exclusively in prokaryotes, representing the most primitive lineage of group II intron ribozymes [48]. The IIC is the smallest family among G2IRs and uses only two recognition sequences (IBS1/EBS1 and IBS3/EBS3) [43].

Due to their relative small size and stability, group IIC introns were the first to be trapped with crystallography experiments, thereby revealing important conserved structural features shared by all G2IRs (Figure 1.5) [37, 42].

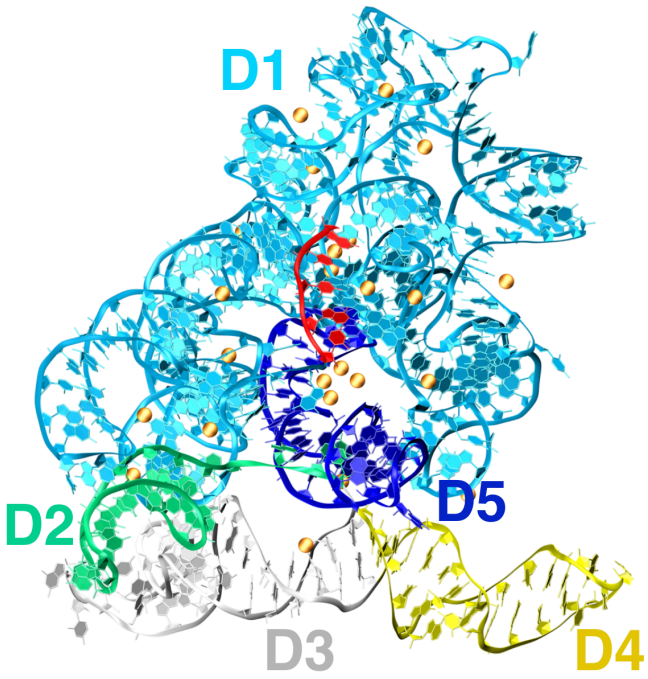


Figure 1.5. Group IIC intron from *Oceanobacillus Iheyensis* (PDB entry: 4FAQ). Mg^{2+} ions are represented with orange spheres. RNA is shown with ribbon representation and the different domains of the large intron are highlighted with different colors as indicated in the picture, while the 5'-exon is colored with red.

The first G2IR to be successfully crystallized and studied using x-ray diffraction methods was a IIC intron from the eubacterium *Oceanobacillus Iheyensis*, which provided a breakthrough information on G2IRs structural domains and reactivity (Figure 1.5). Unfortunately the location of intron Domain VI, which encodes the branching adenosine, was not solved [25], but the crystallized ribozyme construct maintained a catalytic activity.

The IIB family, larger in size, is generally believed to be the most evolved form of G2IR. Although high-resolution structures of the IIB family became available only recently [49], this class is the most studied and is arguably the best understood. Several works focused on the *ai5γ* IIB intron from the mitochondrial genome of the yeast *Saccharomyces Cerevisiae* and the *P.li.LSUI2* IIB intron from the brown alga *Pylaiella Littoralis*. The structure of *P.li.LSUI2* is particularly important because it captures the lariat form of the intron, revealing the position of DVI and the structure of the branch-point nucleotides [49].

A less understood class is represented by IIA, which lacks of a crystal structure, and whose chemical mechanisms have not been examined in detail. Through advances in cryo-EM, a group IIA intron lariat from *Lactococcus lactis* (*L.l.* or *Ll.LtrB*) was solved at 3.8 Å resolution [50]. However, group IIA introns share almost all major structural features with IIB introns and have similar structural and reactivity properties [42, 43, 46].

2.2.1.2 G2IRs reactivity: self-splicing, retrohoming and retrotransposition

Group II intron ribozymes uniquely catalyze their own splicing via two sequential transesterification events that eventually produce ligated exons and an excised intron. This self-splicing occurs via multiple and concatenated stages: (*a*) the precatalytic state, (*b*) the Step 1, (*c*) active-site rearrangement between the two steps, (*d*) the Step 2, and (*e*) the postcatalytic state. These states were characterized through a series of crystal structures of a group IIC intron from *O. Iheyensis* [25, 30]. The G2IRs splicing reaction can follow either a branching pathway (similar to the splicing cycle described for the spliceosome, *vide supra*) or a hydrolytic pathway, depending on the nature of the nucleophile carrying out the first nucleophilic attack during the first step of catalysis (Figure 1.6). In the first step of branching pathway, the 2'-OH of the bulged A in DVI is the nucleophile and the final excised intron is in the lariat form. Alternatively, in the hydrolytic pathway, the nucleophile of the first step is a water molecule, rather than the 2'-OH of the bulged A, and a linear intron is eventually released [25, 37, 43].

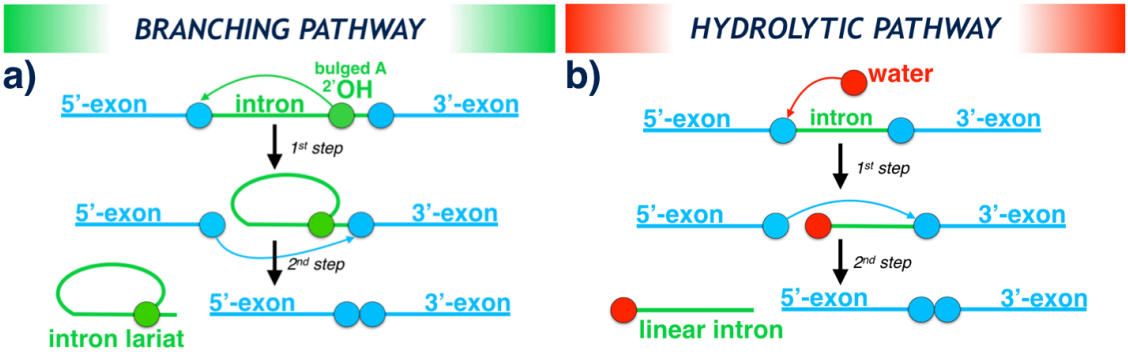


Figure 1.6. Branching pathway (a) vs hydrolytic pathway (b) for group II introns pre-mRNA splicing.

The hydrolytic splicing reaction mechanism self-catalyzed by a group IIC intron from *O. Iheyensis* is extensively discussed in the Chapter 4.

Besides the self-splicing, G2IRs can also reverse splice into DNA targets, a process which is generally called retrotransposition. This reactivity is facilitated by the encoded IEP, which associates to the intron forming a RNP holoenzyme [43]. The high mobility G2IRs as genetic elements is due to their capability of reverse splicing directly into a DNA strand, where they are subsequently reverse transcribed by the IEP. They can either reverse splice into specific DNA target sites in a process named "retrohoming" or into ectopic sites that resemble the normal homing sites ("retrotransposition"), but at low frequencies. Intron-IEP RNP initiates retrohoming by recognizing DNA target sequences (i.e. specific bases or structural features of the DNA target site, which differ for each intron). This interaction helps to separate the DNA strands, enabling the intron RNA to base-pair with the 5' and 3' DNA exons. After base-pairing, the intron reverse-splices into the DNA strand, resulting in its insertion between the two DNA exons. Since G2IRs recognize DNA target sites mostly by base pairing, it is possible to modify them such that they can be inserted into desired DNA sites [51]. This feature, combined with the high efficiency and specificity of the retrohoming reaction, has favored the use of G2IRs into gene targeting vectors ("targetrons") [52]. A targetron based on the *Ll.LtrB* intron is sold commercially through Sigma-Aldrich and widely used for gene targeting in bacteria.

2.2.1.3 G2IRs as ancestors of eukaryotic spliceosome

In relation to their high mobility, G2IRs are thought to have contributed to the eukaryotic genome evolution as ancestors of spliceosomal introns, snRNAs and non-long terminal repeat (non-LTR) retrotransposon. The distribution of group II introns, mostly spread in eubacteria and eukaryotic organelles, while being rare in

archaebacteria, suggests a scenario in which mobile group II introns originated in eubacteria and subsequently were transmitted to eukaryotic genome, possibly via endosymbiotic bacteria that invaded an archaeal host and generated mitochondria and chloroplasts [43]. In eukaryotes, G2IRs are thought to have invaded the nucleus and proliferated to many genomic sites. After that, the cell evolved splicing factors to facilitate and control introns splicing, such that the ribozyme structure degenerated and fragmented into different snRNAs. Then, the snRNAs together with the splicing factors replaced the individual G2IRs as common RNA-based catalytic machineries [53]. The snRNAs continue to recognize the introns via conserved 5' and 3' sequences, promoting a catalysis strikingly similar to that of G2IRs, with the same transesterification reactions and branch-point adenosine. The evolutionary link between G2IRs and eukaryotic spliceosome is, in fact, suggested by their identical splicing pathways, similar boundary sequences, and structural similarities of key regions as shown in Figure 1.7.

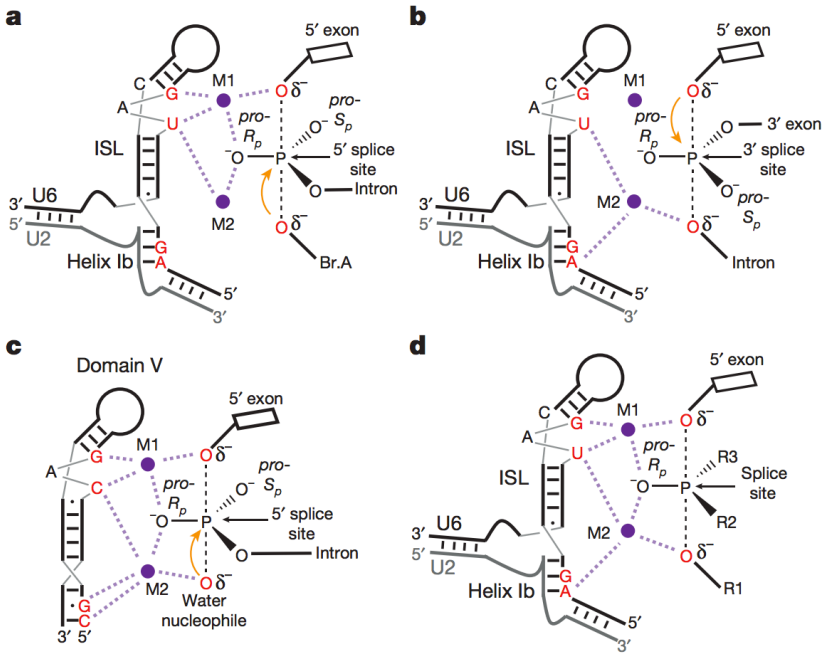


Figure 1.7. Sketch of catalytic metal interactions during branching (a) and exon ligation (b) in the spliceosome. (c) Model of DV during hydrolytic splicing in G2IRs (PDB entry 4FAQ). (d) Two-metal model for the RNA catalytic core of the spliceosome. For the branching reaction, R1 represents the 2'-OH of the branch adenosine; R2, the intron; and R3, the pro-Sp oxygen. For exon ligation (Step 2) R1 represents the 3' oxygen leaving group, R2 the pro-Sp oxygen; and R3 the 3' exon. Throughout the reactive oxygens are colored red, the pre-mRNA scissile phosphate is depicted in a transition state, and interactions between specific ligands and the reactive oxygens mediated by M1 and M2 are shown as light magenta dashed lines. The figure was adapted from reference [44].

These facts, combined with the obvious structural and sequence similarities, support the idea that the spliceosome maintains the features of a ribozyme and uses a catalytic mechanism similar to that of group II introns. Recent experiments using metal rescue strategies on *Saccharomyces Cerevisiae* spliceosome have recently unveiled that the two Mg^{2+} that catalyze splicing and interact with the scissile phosphate are bound by U6 snRNA ligands, thus demonstrating that the RNA moiety of the spliceosome is directly in charge of the catalysis [44]. Remarkably, all of the five U6 snRNA ligands coordinating the two Mg^{2+} correspond to the catalytic metal ligands observed in the DV of G2IRs structures, even with a similar orientation (Figure 1.7).

2.2.2 The eukaryotic spliceosome

The spliceosome, a multimegadalton RNP complex, is one of the most complex and sophisticated machineries of eukaryotic cells. It is an extraordinary RNA-protein apparatus, which removes introns and joins exons from primary pre-mRNA transcripts, giving rise to functional mature mRNA and long noncoding RNAs (ncRNAs). To understand the utmost importance of the spliceosome activity it is enough to highlight that only less than 2% of the human genome is translated into proteins [54, 55]. There are two different types of spliceosome that coexist in most eukaryotes: the U2-dependent spliceosome and the U12-dependent spliceosome. The former (hereafter and before mentioned simply as "spliceosome") catalyzes the removal of the canonical U2-type introns, while the latter is present in a few eukaryotes and splices the very rare and atypical U12-type class of introns [56]. The main differences between U2- and U12-type introns are the 5' splice site and the BPS, which requires different snRNAs and splicing factors for recognition and processing [57].

During spliceosome assembly and catalysis, snRNAs are combined with proteins into RNPs, giving rise to intricate RNA–RNA and RNP networks, which are repeatedly rearranged along the splicing cycle. These are aimed to precisely align the reactive groups of the pre-mRNA for catalysis within this highly dynamic complex (Figure 1.4) [56]. It has been estimated that a typical human gene is composed by 8 exons with an average of 145 nucleotides (nt) in length. Introns are instead usually 10 times larger. The spliceosome machinery must operate with high precision as a single error that adds or removes even 1 nt will disrupt the open reading frame of the resulting mRNA. The first real challenge for the spliceosome is the recognition of the correct splice sites (5'SS and 3'SS) which takes place prior to the "snip-and-stitch" reactions [41].

2.2.2.1 Spliceosome assembly and catalytic cycle

Much of the actual knowledge of the molecular mechanism of pre-mRNA splicing is based on intensive researches using extracts made from the budding yeasts *Saccharomyces Cerevisiae* and *Schizosaccharomyces Pombe*. These provide unique insights to understand the conserved spliceosome mechanism and assembly also in humans [58]. With the advent of high-resolution cryo-EM, several spliceosome structures have been determined in the last few years [31, 32, 59-64], with different structural models corresponding to distinct steps of the splicing reaction [65]. This new era represents a breaking point for the field and, although model organisms such as the above-mentioned fungi will likely continue to furnish important details, human spliceosome structures are starting to be solved [33, 34, 63, 66]. Spliceosome extracts usually contain U1 and U2 snRNPs and U4/U6.U5 tri-snRNPs as major components, providing information on spliceosome catalytic cycle when synthetic pre-mRNA substrates are added. Biochemical data and studies of these systems, combined with genetic approaches, led to a universal consensus for the spliceosome mechanism and assembly, which can be divided into six phases (Figure 1.8) [39, 56, 62, 67, 68].

Phase I (Complex A formation). At the very beginning of this cycle a "commitment complex" is formed, i.e. an ATP-independent complex committing the pre-mRNA to the splicing pathway, where U1 snRNP recruits the 5'SS and U2 snRNP recognizes the 3'SS with the help of associated factors.

Subsequently, in an ATP-dependent process catalyzed by the DExD/H helicases pre-mRNA processing 5 (Prp5) and Sub2, U2 snRNA also recognizes the 5'SS and the branch-point region of the intron, leading to the formation of the pre-spliceosome or "complex A".

Phase II (Complex B formation). The pre-assembled U4/U6.U5 tri-snRNP is recruited to form the fully assembled, but catalytically inactive, "complex B".

Phase III (activation to Complex B*). The transition to a catalytically competent structure is driven by major structural and compositional changes, that eventually result in the formation of the activated B complex (B^{act}) and the catalytically competent "complex B*". The Prp8 protein, part of U5 snRNP (Spp42 in *S. Pombe*), is the principal actor in docking the tri-snRNP to U1 snRNP and in the activation of the B complex, by coordinating the actions of different helicases [69]. Namely U1 snRNP dissociates from the 5'SS in a reaction requiring ATP hydrolysis and mediated by the DExD/H helicase Prp28, while the RNA helicase Brr2 promotes, always through ATP hydrolysis, the unwinding of the extensively base-paired U4/U6 snRNA duplex, leading to U4 snRNA dissociation along with its associated proteins.

In addition, NTC (NineTeen Complex) and NTC-associated (NTC-a) proteins are recruited from the B^{act} complex to stabilize the assembled spliceosome during its

activation and will remain associated to the spliceosome during all the following phases [64].

Finally, one of the most important conformational change involves U2 and U6 snRNAs, which base-pair to generate an active site similar to that of G2IRs.

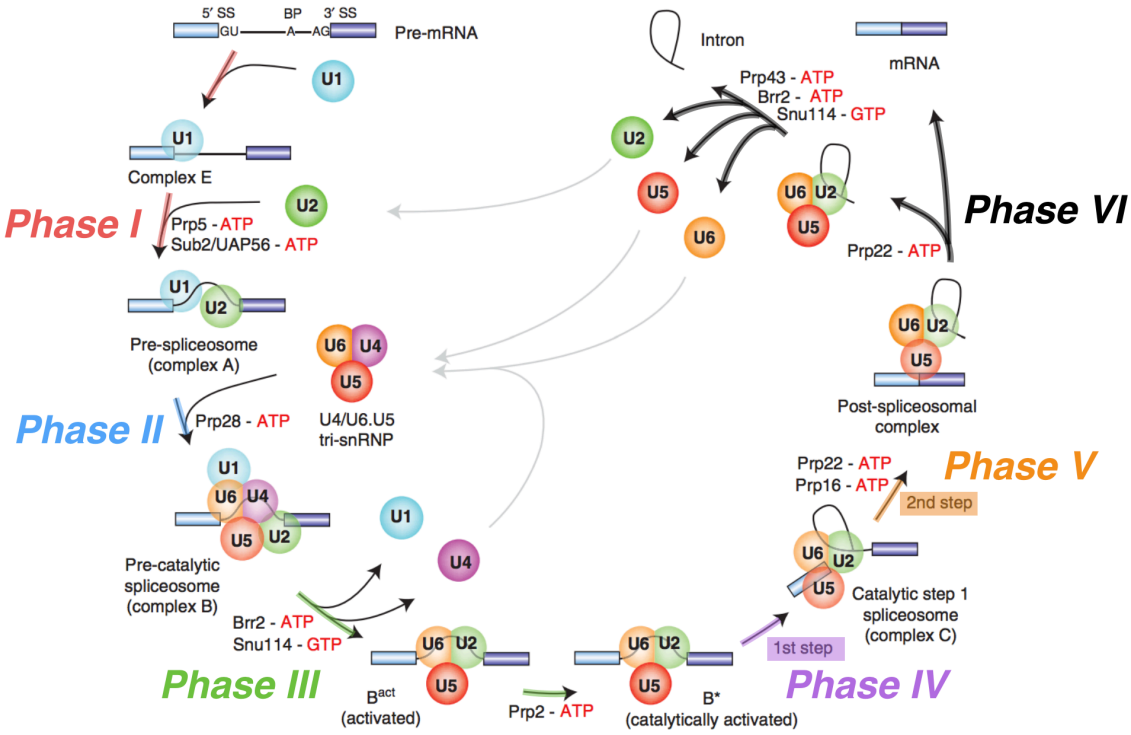


Figure 1.8. Pre-mRNA splicing cycle of the U2-type spliceosome. The canonical cross-intron assembly and disassembly pathway of the U2-dependent spliceosome is represented. For simplicity, the ordered interactions of the snRNPs (indicated by circles), but not those of non-snRNP proteins, are shown. Exon and intron sequences are indicated by boxes and lines, respectively. The stages at which the evolutionarily conserved RNA ATPases/helicases Prp5, Sub2/UAP56, Prp28, Brr2, Prp2, Prp16, Prp22 and Prp43, or the GTPase Snu114, act to facilitate conformational changes are indicated. The figure was adapted from [56].

Phase IV (first splicing step and Complex C formation). All the previous events are aimed at positioning and orienting the 2'-OH group of the branching adenosine to allow its nucleophilic attack on the scissile phosphate at the 5'SS, which yields, as product, the "complex C", containing the free 5'-exon and the intron lariat–3'-exon intermediates.

Phase V (activation to Complex C* and second splicing step). After the first cleavage, the C complex undergoes further remodeling to become "complex C*". The exons are aligned on the conserved loop 1 of U5 snRNA for the second transesterification reaction, which depends on Prp8 (Spp42 in *S. Pombe*), Prp16 and

Slu7. The result of the second step is a post-spliceosomal complex ("P complex"), which contains the intron lariat and the mRNA filament (i.e., the ligated exons).

Phase VI (mRNA Release and ILS Disassembly). The spliced mRNA product (5'-exon–3'-exon) is subsequently released and the residual intron-lariat-spliceosome ("ILS") complex is disassembled. While U2, U6 and U5 snRNPs are recycled for subsequent rounds of splicing, the excised intron lariat is degraded. The release of the spliced product from the spliceosome is catalyzed by helicase Prp22 and the disassembly of the post-catalytic spliceosome is also driven by several RNA helicases (among which Brr2, Snu114, Prp22 and Prp43) in an ATP-dependent manner [56].

2.2.2.2 SnRNP and non-snRNP splicing factors: structure and functions along the splicing cycle

The five snRNPs, composed by several proteins and snRNAs, are involved in the splicing cycle with different roles and conformational transitions. Even if closely related, some difference exists between fungi and humans in terms of composition, size, number and name of the protein components. The spliceosome assembly occurs at the beginning of every cycle and, thereby, each time the pre-mRNA splicing is performed.

Spliceosome proteins are divided into snRNP proteins that are tightly associated with a specific snRNAs, forming specific and functional snRNPs, and non-snRNP splicing factors, which instead are not associated to any snRNAs and are recruited along the cycle. SnRNP and non-snRNP splicing factors are usually named in *S. Cerevisiae* and *S. Pombe* with "Prp#", "Cwf#" or "Cwc#" where Prp stands for "Pre-mRNA processing", Cwf "Complexed with Cdc5" and Cwc for "Complexed with Cef1". Cdc5 (*S. Pombe*) and its *S. Cerevisiae* homolog Cef1 are essential proteins implicated in pre-mRNA splicing and contained within the large NTC multiprotein complex.

Splicing factors might present particular sequence motifs/domains, such as zinc finger, helicase, protein kinase, GTPase or peptidyl/propyl cis-trans isomerase. ATP-dependent helicases containing the DEAD or DEAH box (for example, Prp28 and Brr2) are often involved in the rearrangement of the RNA–RNA network, like helices unwinding. Instead, GTPases (like Snu114, Cwf10 in *S. Pombe*) and peptidyl/prolyl cis-trans isomerases usually play an important role in the conformational changes occurring within the spliceosome [70].

Moreover, there are additional proteins, such as the Sm proteins, which assemble indiscriminately with each snRNA (i.e., U1, U2, U4 and U5 snRNAs). The Sm proteins usually associate around U-rich RNA sequences, forming a stable globular core domain in the snRNPs called the Sm snRNP core particle [71]. The Sm core is

essential for the metabolic stability of the snRNP, preventing the bound snRNA from degradation, but it is also fundamental for the import of the snRNP into the nucleus and for protein recruitment.

U1 snRNP. The very first snRNP involved in the catalytic cycle is the U1 snRNP, which starts the engine of the machinery. Human U1 snRNP is formed by U1 snRNA, the seven Sm proteins, named in order of decreasing molecular mass (SmB/SmB', SmD1, SmD2, SmD3, SmE, SmF and SmG), and three U1-specific proteins (U1-70K, U1-A and U1-C). Yeast U1 snRNP instead contains a larger and more complex U1 snRNA, which is associated with many protein factors (Prp39, Snu71, Prp40, Prp42, Nam8, Snu56, Urn1 and Prp5), which have no counterparts in human U1 snRNP [66].

The main function of snRNP is the recruitment of the 5'SS of the intron through a base pair between a few nucleotides located at the 5'-end of U1 snRNA with a short sequence at the 5'SS of the pre-mRNA. U1 snRNP also promotes an ordered assembly of the four remaining snRNPs, which will form the spliceosome. Despite the differences between human and yeasts, the sequence of the 5'-single stranded region of U1 snRNA (nts 1–10) is invariant, thus making the exactly same contacts with the 5'SS of pre-mRNA [66].

In the following phases of the cycle, the interaction between U1 snRNP and the 5'SS is disrupted by RNA helicase Prp28. Afterwards, the 5'SS intron sequence base-pairs with part of the ACAGAGA box in U6 snRNA [64].

U2 snRNP. This is a crucial component during the splicing cycle. U2 snRNP is composed by the U2 snRNA, the 7 Sm proteins, and approximately 15 U2 snRNP-specific proteins [72, 73].

In the early stages of spliceosome assembly, with an ATP-dependent process, U2 snRNP recognizes the branch-point adenosine by forming a short U2/intron duplex through specific base pair between U2 snRNA and the BPS, which contains the branching adenosine. This step is fundamental to position and bulge out the branching A to perform the nucleophilic attack during the first catalytic step. Prior to splicing activation (i.e., in the transition between B^{act} and B*), U2 snRNA base-pairs with U6 snRNA, forming the typical triple-helix motif of the active site [74].

U4/U6.U5 snRNP. The largest (~1.5 MDa) snRNP participating in the spliceosome assembly is the trimeric U4/U6.U5 snRNP (Figure 1.9), which associates to complex A to form the complex B.

U4/U6.U5 snRNP consists of U5 snRNP and the U4/U6 di-snRNP. It comprises more than 30 proteins including, among others, Prp8 (Spp42 in *S. Pombe*), the GTPase Snu114 (Cwf10 in *S. Pombe*), and the helicase Brr2, and 3 snRNAs, namely U5 snRNA, U4 and U6 snRNAs, which are extensively base-paired. As revealed by recent

cryo-EM structures [62, 75], U4/U6.U5 tri-snRNP complex has a triangular shape (Figure 1.9).

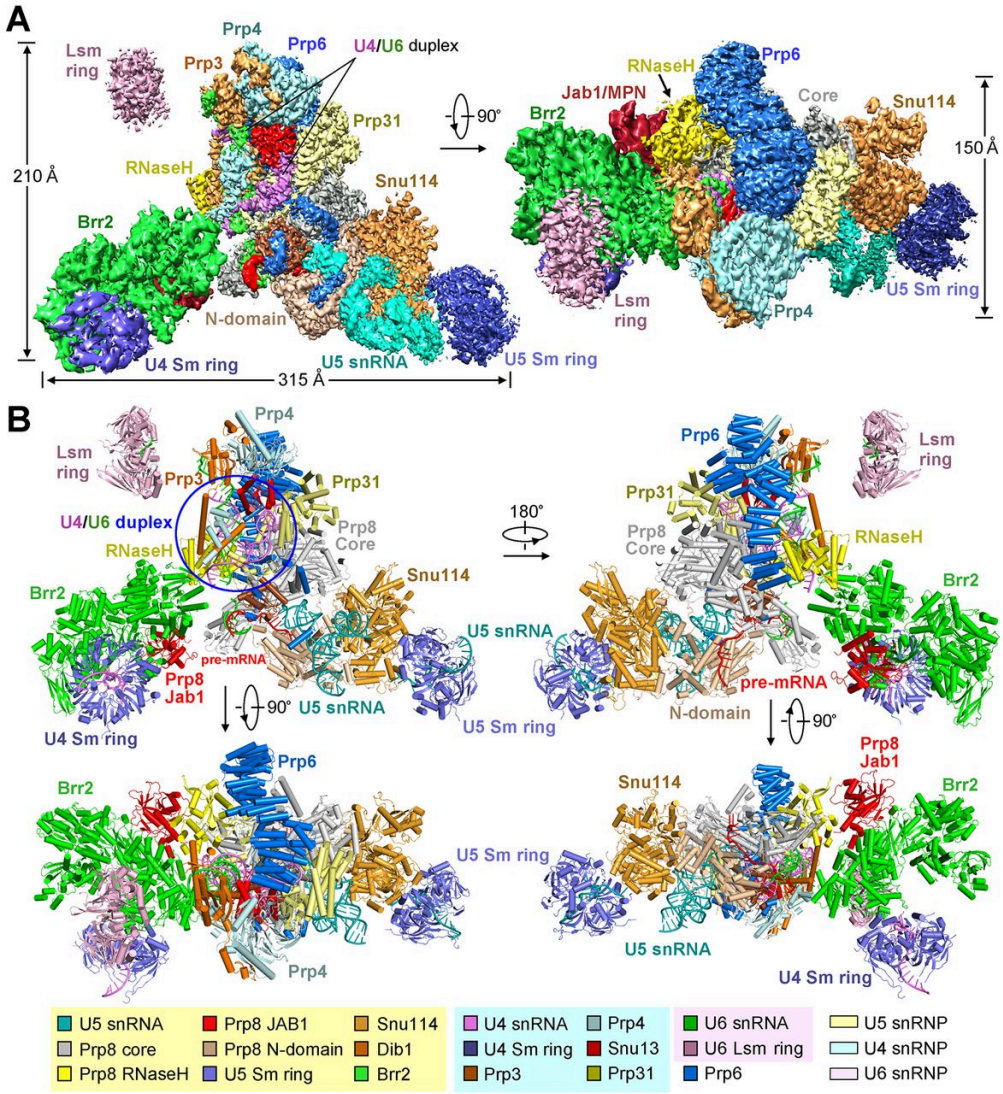


Figure 1.9. The cryo-EM maps of the yeast U4/U6.U5 tri-snRNP at an overall resolution of 3.81 Å. (B) A cartoon of the yeast U4/U6.U5 tri-snRNP complex. The protein and RNA components are color-coded. Four views are shown. This structure includes 30 proteins, three snRNA molecules, and a pre-mRNA molecule, with a combined molecular weight of ~1 MD. The figure was adapted from [62].

Proteins Prp8, Snu114 and Sm ring (belonging to U5 snRNP) occupy the lower part of this triangular assembly. Instead, the Brr2 helicase (still belonging to U5 snRNP) and the U4/U6 di-snRNP fills the upper part. The region containing Brr2 appears to be quite flexible and more difficult to be resolved with respect to the other parts.

Prp8 (Spp42 in *S. Pombe*), the most important protein of U5 snRNP, is located in the core of the spliceosome and it acts as a hub for several crucial conformational rearrangements taking place along the splicing cycle [76]. Remarkably, the organization of some domains of Prp8 resembles that of the proteins encoded by certain G2IRs (IEP), which facilitate RNA folding and ribozyme catalysis. This suggests a relationship between these RNA chaperons [39]. Prp8 is found in all the spliceosome complexes along the splicing cycle, starting from complex B until ILS. The N-terminal domain (N-t domain) of Prp8 is mainly responsible for binding U5 snRNA and the GTPase Snu114, whereas the C-terminal Jab1/MPN domain of Prp8 binds Brr2. The RNaseH-like domain of Prp8 interacts with Prp3 and the tri-snRNP-specific protein Prp6.

Snu114 GTPase (Cwf10 in *S. Pombe*), another specific protein of U5 snRNP, plays an important role in promoting the release of U4 snRNA in a process that requires GTP hydrolysis [77]. Intriguingly, the binding of GTP, and not of GDP, is necessary for the formation of a stable interaction with Prp8 in an early stage of the cycle [78]. Moreover, after the B complex activation, GDP-bound Snu114 appears to function as a regulator of the helicase Brr2, repressing its activity [79]

Finally, the core of the U4/U6 snRNP comprises U4/U6 snRNA, Prp3, Prp4, Snu13, and Prp31. The U4/U6 snRNP closely associates with the core of Prp8 through the ferredoxin-like protein Prp3 and the Nop domain of Prp31, forming a compact structure that stands out from the rest of the U4/U6.U5 tri-snRNP [62]. U2 and U6 snRNAs reach the catalytic core of the spliceosome and forms the active site by coordinating the Mg^{2+} ions which act as cofactors.

Catalytic site. In principle, any –OH and phosphate groups from any pair of nucleotides could undergo a transesterification reaction. This is the reason why the active site has to be generated only transiently and only after the correct reactants have been identified. To avoid the first scenario, and before reaching the B* complex, the U6 snRNA is sequestered by extensive base pairing with U4 snRNA. U6 is, in fact, the only snRNA that is indispensable for splicing. Indeed, *in vitro* studies showed that U6 snRNA in complex with another snRNA can catalyze splicing reaction in the absence of all other spliceosomal factors. The U4/U6 duplex is subsequently unwound by Brr2 and U4 is released by the action of Snu114 and GTP hydrolysis. Afterwards, U6 snRNA base-pairs with U2, with the final result of producing the B* complex exhibiting a catalytic site prone to react [80]. In this crucial step, a conserved structural motif, called triple-helix, is formed between the U6 AGC triad, a complementary UCG sequence in U2 (Helix Ib), and two 5'-downstream and one 3'-upstream nucleotides of U6 snRNA (Figure 1.10). Then, four phosphate residues of U6 snRNA are positioned to coordinate the two Mg^{2+} ions cofactors, forming an active site strikingly similar to

that of G2IRs. The catalytic site docks on a cavity of Prp8, which stabilizes the structure and also impedes that other regions belonging to U2 and U6 snRNAs can interfere [39].

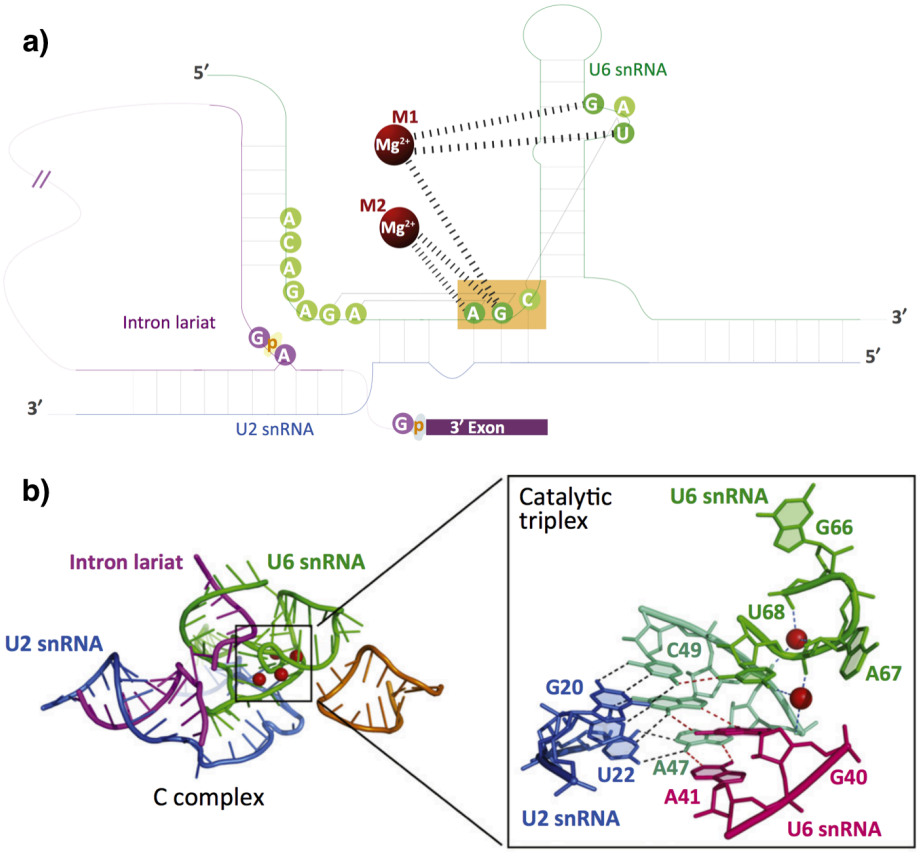


Figure 1.10. (a) Sketch of the arrangement of RNAs around the spliceosomal catalytic pocket after the first catalytic step. The U6 snRNA AGC triad (highlighted in orange) base pairs with U2 snRNA residues as well as with distant regions of U6 snRNA (discontinuous red lines), forming a triplex-helix, which brings in close proximity the nucleotides coordinating the two catalytic Mg^{2+} ions. Abbreviation: ISL, internal stem loop. (b) 3D structure of the triple-helix. The figure was adapted from [39].

NTC, NTC-associated and other splicing factors. As highlighted by a cryo-EM structure resolved immediately after branching, the NTC and NTC-associated proteins, recruited before the first step, seem to act as a multipronged clamp that stabilizes the binding of the U2 snRNP core, the substrate, and also auxiliary splicing factors to the U5 snRNP [64]. The NTC and NTC-associated proteins are found within the B, B^{act}, B*, C, P and ILS complex [81]. The NTC is composed by Prp19 and by at least seven other Prp19-associated proteins and it is also known as Prp19 complex. At least 18 NTC-associated proteins interact with NTC core, among which Cwc2 (Cwf2 in *S. Pombe*) and Cwf19 (*S. Pombe*) are close to the active site [82]. Cwf19 seems to be

involved in the displacement of the intron lariat/U2 duplex before the disassembly of the ILS complex [64]. This role of Cwf19 will be discussed in Chapter 6.

After the first step of catalysis, the DEAH-box helicase Prp16 splicing factor catalyzes the conversion of the C complex to the catalytically activated C* complex, preparing the spliceosome for the second step. The structure of the active C* complex has been recently resolved, suggesting a conformational rearrangement driven by Prp16 and Prp8 [83]. Prp16, the engine of this reorganization, presumably pulls the 3'-end of the lariat and triggers its translocation. The RNaseH-like domain of Prp8 also appears to undergo a translocation, which helps moving the BPS/U2 duplex to its new location in the C* complex. Moreover, several other second step splicing factors associate, such as Slu7, Prp18, and Prp22. The second catalytic step results in the formation of the P complex, which resemble the C* complex, but contains the two ligated exons and the lariat intron.

Once both the transesterifications have been completed, the mature mRNA is released from the spliceosome. The ATPase Prp22 splicing factor promotes this event by unwinding the mRNA/U5 snRNA duplex and favoring the release of the mRNA using the energy coming from ATP hydrolysis. Afterwards, Slu7, Prp18 and Prp22 dissociate from the spliceosome.

When the release of the mRNA filament has occurred, another splicing factor, the helicase Prp43, associates with Ntr1 and Ntr2 to form the NTR complex (NineTeen complex Related) and promotes ILS disassembly in an ATP-dependent manner, together with Brr2. Small and collaborators [79] have proposed that Snu114 de-represses Brr2 activity after the second catalytic step by exchanging the GDP with GTP. Hence, the re-activated Brr2, along with Prp43, would then promote the intron release and spliceosome disassembly [84]. A structure that likely corresponds to the last state of the spliceosome before disassembling, i.e., the ILS state, was recently solved, and was the first to unveil important insights on spliceosome composition at high resolution (3.6 Å) [31]. This structure has been the subject of part of the research that I will present in this thesis (detailed in Chapter 6).

3 Methods

3.1 Molecular dynamics

3.1.1 Statistical mechanics and molecular dynamics

Computational simulations represent the bridge between microscopic lengths and time scales and the macroscopic world of the laboratory. Statistical mechanics, a branch of physical sciences, provides the mathematical and theoretical tools to connect the micro-scales with macroscopic quantities. Considering a system of *N* particles, this is characterized by a set of atomic positions $\left(\vec{R} = \{\vec{R}_1, \dots, \vec{R}_N\}\right)$ and relative momenta $\left(\vec{P} = \{\vec{P}_1, \dots, \vec{P}_N\}\right)$ which define its microscopic state. This can be seen and represented as a single point in a 6*N* multidimensional space, called phase space (Γ). Thus, a single point in the phase space represents a microscopic state of the system, while a collection of points in the Γ defines an ensemble. Molecular Dynamics (MD) simulations are the applicative tool that generates a time sequence of points in the phase space, i.e. a sequence of different positions and momenta of the system belonging to the same ensemble. For each microscopic state of the system in the phase space, it is possible to estimate the observable value of a certain property *A* as a function of Γ, *A*(Γ), as the ensemble average or thermodynamic average:

$$A_{obs} = \langle A \rangle_{ens} = \int A(\Gamma)\rho(\Gamma)d\Gamma \tag{3.01}$$

where $\rho(\Gamma)$ is the probability distribution function of collection of points Γ, and $d\Gamma = d\vec{R}_1 \cdots d\vec{R}_N d\vec{P}_1 \cdots d\vec{P}_N$. The probability distribution function depends on macroscopic parameters, which define the thermodynamic state of a system like the number of

particles, N, volume V, temperature T and pressure P. For example, in the canonical ensemble (NVT), N, V and T are constant, and the probability distribution function has the form of the Boltzmann distribution function:

$$\rho_{NVT} = \frac{e^{-\frac{H(\Gamma)}{K_B T}}}{Z} \tag{3.02}$$

where $H(\Gamma)$ is the classical Hamiltonian of the system defined as:

$$H(\Gamma) = H\left(\{\vec{R}_I\}, \{\vec{P}_I\}\right) = \sum_{I=1}^{N} \frac{\vec{P}_I^{\,2}}{2M_I} + U\left(\{\vec{R}_I\}\right) \tag{3.03}$$

where $\vec{R}_I, \vec{P}_I$ and M_I are the position, momentum and mass of the particle I, U is the potential energy, K_B is the Boltzmann constant and Z is the canonical partition function. MD simulations allow estimating ensemble averages by simply integrating the Newton's equations of motion such that the system is evolved as a function of time starting from its microstate at time 0 until its microstate at time τ. A set of microstates of the system, i.e. a trajectory of points in the phase space Γ(t), is generated. From this trajectory, the time average value of an observable $\langle A \rangle_\tau$ can be calculated and connected to $\langle A \rangle_{obs}$ according to the "ergodic hypothesis" [85]. Indeed, if the system is evolved for an infinitively long time it should be able to visit all the states and thus its behavior averaged over time and over the phase space become the same:

$$\lim_{\tau \to \infty} \langle A(\Gamma) \rangle_\tau = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau A[\Gamma(t)]dt = \langle A(\Gamma) \rangle_\Gamma. \tag{3.04}$$

The longer is the simulation time, better this equality is satisfied [86]. Thus, the application of MD to molecular biological systems is an immediate and usable way to predict their average behavior and estimate macroscopic observables. The time evolution of the system is calculated by integrating the second Newton's equations of motion, with atoms treated as point particles:

$$\vec{F}_i = M_i \vec{a}_i \tag{3.05}$$

where $\vec{F}_i$ is the force acting on particle i, M_i is the mass of the particle i and $\vec{a}_i$ the second derivative of the particle's position with respect to time t, i.e. its acceleration. The atoms will move under the influence of the internal forces acting on them, which are derived from the potential energy of the system ($U(\vec{R})$):

$$\vec{F}_i = -\frac{\partial U(\vec{R})}{\partial \vec{R}_i}. \tag{3.06}$$

As long as the force in equations 3.05 and 3.06 is a function of all the positions of the N particles, eq. 3.05 constitutes a set of dN coupled second-order differential equations, where d is the number of spatial dimensions. Newton's equations completely determine the full set of positions and velocities as functions of time and thus specify the classical state of the system at time t. An analytical solution to the equations of motion 3.05 is obtained from a set of initial conditions of particles positions and velocities. The former are taken from PDB crystal structures, NMR data, cryo-EM or homology modelling coordinates, while the latter are usually randomly generated from the Maxwell-Boltzmann probability distribution at a given temperature T. For this reason, time discretized numerical algorithms are employed to update the positions and velocities of the particles at each time step, Δt. This is chosen at the beginning of the simulation and is usually in the range of 1-2 fs, such that the fastest motion in the system can be integrated stably and accurately. A frequently used integration algorithm is the velocity-Verlet algorithm [87, 88], which makes use of a Taylor expansion truncated beyond the quadratic term for the coordinates:

$$\vec{R}(t+\Delta t) \approx \vec{R}(t) + \vec{v}(t)\Delta t + \frac{\vec{F}(t)}{2m}\Delta t^2 \tag{3.07}$$

with the velocities $\vec{v}$ updated as:

$$\vec{v}(t+\Delta t) \approx \vec{v}(t) + \frac{\vec{F}(t) + \vec{F}(t+\Delta t)}{2m}\Delta t. \tag{3.08}$$

with the relation $\vec{P} = m\vec{v}$. A variation of velocity-Verlet algorithm is the Leap-Frog algorithm [89], which uses velocities at half-integer time steps to determine new particles' positions:

$$\vec{v}\left(t+\frac{\Delta t}{2}\right) = \vec{v}\left(t-\frac{\Delta t}{2}\right) + \frac{\vec{F}(t)}{m}\Delta t + O(\Delta t^3) \tag{3.09}$$

$$\vec{R}(t+\Delta t) = \vec{R}(t) + \vec{v}\left(t+\frac{\Delta t}{2}\right)\Delta t + O(\Delta t^3). \tag{3.10}$$

Leap-Frog calculates positions and forces at interleaved time points. As a consequence, kinetic and potential energy are also not defined at the same time.

3.1.2 Force fields based molecular dynamics

Force fields (FF) based MD has been largely used in atomistic studies of biological systems, including proteins, amino acids, lipid bilayers, and carbohydrates. Starting from 70s', when the first MD simulations of protein started to appear [90], its application has remarkably increased in size, complexity, and length through the years,

allowing simulations of systems with millions of atoms and reaching the microsecond- or even the millisecond-scale [91]. This remarkable improvement is in large part a consequence of the use of high performance computing (HPC), the improvement of supercomputer, the introduction of Graphic Processing Units (GPUs) and the simplicity of the basic MD algorithm [91-93]. In force field-based MD, also called classical MD, the potential energy, U^{FF}, is expressed as an empirical force field (FF) parametrized to reproduce experimental data or *ab-initio* calculations. In fact, in classical MD, atoms are considered as point particles and the electronic degrees of freedom are not taken into account. Thus, instead of solving the Schrödinger equation as in *ab-initio* MD, in classical MD the potential energy is simply approximated by using parametric functions of nuclear coordinates, the force field. Nowadays, several force fields are available for biomolecular applications such as AMBER [94], GROMOS [95], CHARMM [96]. In all the works presented in this thesis I have employed the AMBER force field, which has the following functional form:

$$U^{FF}(r_1,\dots,r_N) = \sum_{bonds} K_r\,(r-r_0)^2 + \sum_{angles} K_\theta\,(\theta-\theta_0)^2 + \sum_{dihedrals} \frac{V_n}{2}[1+cos(n\phi-\gamma)] + \sum_{i<j} \varepsilon_{ij}\left[\left(\frac{R_{min\,i,j}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{min\,i,j}}{r_{ij}}\right)^{6}\right] + \sum_{i<j}\frac{q_i q_j}{\varepsilon r_{ij}}. \quad (3.11)$$

The potential energy U^{FF} is divided in bonded interactions within atoms involved in a chemical bond (1-2), an angle (1-2-3) or a dihedral (1-2-3-4), represented by the first three terms (bonds, angles and dihedrals), and non-bonded ones, describing the interactions between atoms 1-5 and farther. The non-bonded interactions correspond to the last two terms in the FF, the Lennard-Jones (LJ) potential which defines the van der Waals (vdW) interactions and the Coulomb potential for the electrostatic interactions. Interaction 1-4 are also included in non-bonded but scaled with a factor 0.5, while 1-2 and 1-3 are excluded. In particular, K_r, K_θ and V_n are bond, angle and dihedrals force constant; r and r_0 represent bond lengths and their equilibrium values, while θ and θ_0 are angles and their equilibrium values; n, ϕ and γ are the number of barriers, dihedral angle and phase, respectively, which define the torsional potential. The first term of the LJ potential describes the repulsive forces at short interatomic distances due to the Pauli repulsion (modeled with r_{ij}^{-12}), while the second term represents the attractive forces at intermediate distances due to instantaneous dipole-induced dipole interactions (decaying with r_{ij}^{-6}). For this reason, the LJ potential is considered a short-range potential. In particular, r_{ij} is the distance between atoms i and j, while $R_{min\,i,j}$ is their internuclear separation at equilibrium and ε_{ij} is the well depth of the LJ potential. To speed-up the calculation of the van der Waals interactions, a list of particles within a cutoff $R_c + \Delta_{list}$ (Δ_{list} is a small buffer margin defined at the beginning of the

simulation and is usually 1-2 Å) of each particle i is generated (neighbor lists) and updated every N steps such that beyond that distance the interaction is set to 0. Usually the cutoff is set to 10 Å. Instead, the Coulomb potential decrease with r_{ij}^{-1} and therefore is a long-range potential, with q_i and q_j being the partial electric charges of atoms i and j, while ε is the dielectric constant. In terms of computation time, the calculation of non-bonded interactions is the most demanding part and in particular for the long-range Coulomb term.

In actual MD simulations finite systems are used, thus requiring a special treatment as the particles at the boundaries of the system would lie at the edge with the vacuum, generating artifacts due to finite-size effects. In order to avoid this problem, periodic boundary conditions (PBC) are commonly applied. With PBC, the central system, the only one that will be explicitly treated, is surrounded by an infinite number of replicas of itself which should reproduce an infinite solution around the system. When the system is infinitely replicated, long-range interactions like electrostatic ones, whose spatial range may extend beyond the boundaries of the central image, present a challenge for the computation, thus calling for a fast but still accurate treatment. Fortunately, many algorithms have been developed and implemented to overcome this problem and among them the Ewald summation [97] and the particle mesh Ewald method [98] (PME) are widely used. These methods successfully treat long-range interactions of the form $1/r^n$ for $n \leq 3$ and so the Coulomb interactions. The basic idea behind these techniques is to split the relevant part of the potential into a short-range part, ordinarily treated within a cutoff, and a long-range term, in which the remaining interactions are Fourier transformed [99].

3.1.2.1 Accuracy and reliability of RNA and protein force fields

MD simulations represent a useful tool to study the structural dynamics of biomolecules like RNAs, DNAs or proteins and they can provide precious atomistic insights which sometimes are not accessible to experimental techniques. The main drawbacks of MD simulations are related to the short time scales sampled and to the limited accuracy of the force fields which describe the potential energy function. Indeed, force fields flaws may compromise the entire reliability of the simulations, sometimes generating misleading results. While many steps forward have been done in the development of protein force fields [100, 101], RNA force fields are still under the glance for a better refinement [102]. RNA force fields are, in fact, plagued by imbalances in the description of appropriate base-pairing, stacking and base/sugar interactions, sometimes precluding from properly sampling native basins. Achieving a balanced description of the above-mentioned interactions thus represents a serious challenge. The first RNA force field available in the Amber package was the *ff94*

developed by Cornell et al. [94], which was then improved in the following years with corrections to many dihedral parameters (*ff98* [103], *ff99* [104]). The latest and widely used refinements are represented by the *bsc0* [156] (for nucleic acids) and χ*OL3* [157] (for RNA) corrections to the *ff99*. Whereas the *bsc0* addressed the tendency of double helices to convert to extended forms by fixing α and γ backbone torsion angles, the χ*OL3* modifications improved the description of χ (the glycosidic dihedral angle) potential for RNA, which characterizes the relative base/sugar orientation. In particular, this last correction allowed balancing the *anti* and the high-*anti* conformation related to the χ angle. The *bsc0* and the χ*OL3* modifications to *ff99* (*ff99bsc0*χ*OL3*) have been incorporated into *ff10*, *ff12SB* and *ff14SB* of the Amber code. A further development of the RNA force fields, not only relative to dihedral angles, but also regarding van der Waals parameters and the partial charges, still represents an active field of research [105].

Additional FFs-specific problems may arise from protein/RNA simulations, which require a careful preparation of the starting structures. These simulations are affected by numerous factors, including properties of the starting structures (the initially high force field potential energy, resolution limits, conformational averaging, crystal packing, etc.), force field imbalances, and real flexibility of the studied systems [105]. A recurrent drawback is the poor description of the native H-bonds between protein and RNA, especially for dense H-bond network. However, protein/RNA simulations constitute nowadays a routine complement of experiments and the overall performance of the methodology is apparently better than in small RNA systems [106], thus indicating a satisfactory degree of compensation of force fields errors at the interface of protein/RNA complexes [105]. When simulating a protein/RNA system, multiple stable microsecond time scale simulations are usually suggested to obtain converged results. Moreover, non-standard equilibration protocol must be adopted because of the frequently high initial force field potential energy [105]. As a result of recent tests [105, 106], the best force field for protein/RNA simulations is the ff12SB [107] (over the ff14SB) for proteins and the *ff99bsc0*χ*OL3* for RNA, which is also incorporated in the ff12SB in the Amber package.

As a final note of remark, the accuracy of MD simulations does not necessarily decrease as the system size increases [105]. Sometimes more instabilities are obtained out of MD simulations on tetranucleotides with respect to more structured biomolecules. Indeed, the performance of the FFs may improve in larger systems due to a better compensation of errors [105]. Instead, for large-size systems, the main pitfalls are represented by the uncertainties in the starting geometries, which must be adequately evaluated [105].

3.1.3 Temperature and pressure coupling schemes

An ensemble that might be sampled with MD simulations is the microcanonical ensemble, where the number of particles (*N*), the volume (*V*) and the total energy (*E*) are kept constant. However, in order to obtain more meaningful information that can also be related to the experiments, MD simulations are usually coupled with thermostats or barostats. In the first case volume and temperature (*T*) are maintained constant (canonical ensemble, *NVT*), while in the second case volume is allowed to change, whereas the pressure (*P*) and temperature do not vary (*NPT* ensemble). Many thermostat and barostat algorithms are available and commonly used in MD simulations. I will briefly review some of them.

Langevin thermostat. The Langevin thermostat [108, 109] is implemented in the Amber software and mostly used in the *NVT* MD simulations presented in this thesis. In this thermostat, a constant friction, γ_i, which lowers the velocities, and a random force acting on all the particles are introduced, resulting in the following differential equations:

$$\frac{d\vec{R}_i}{dt} = \frac{\vec{P}_i}{M_i} \tag{3.12}$$

$$\frac{d\vec{P}_i}{dt} = \vec{F}_i - \gamma_i \vec{P}_i + \sigma \frac{\vec{R}_i}{\sqrt{dt}} \tag{3.13}$$

where $\vec{F}_i$ is the force deriving from interaction potential and the last term, $\vec{R}_i$, is the random force contribution and $\vec{P}_i$ is the particle momentum. σ is the dispersion of the random force and it is related with the frictional coefficient γ_i through the eq. 3.14:

$$\sigma = \sqrt{2\gamma_i M_i K_B T}. \tag{3.14}$$

K_B is the Boltzmann's constant, M_i is the mass of the particle i, and T is the temperature. The random force is randomly determined from a Gaussian distribution and adds kinetic energy to the particles, thus balancing the negative frictional contribution. The Langevin thermostat is a stochastic ergodic thermostat, which reproduces correctly the canonical ensemble thanks to fluctuation-dissipation relation between σ and γ_i.

Nosé-Hoover thermostat. The Nosé-Hoover thermostat [110, 111], was initially formulated by Nosé and subsequently improved by Hoover. It is a deterministic algorithm which, in its final formulation, modifies the equations of motions by introducing a frictional force proportional to $\xi\vec{P}_i$, where ξ is a thermodynamic friction parameter, and $\vec{P}_i$ is the momentum of each particle. The ξ parameter is a dynamic

quantity with its own momentum $\vec{P}_\xi$ and equation of motion. The equations of motions, including the one for the heat bath parameter ξ, are written as:

$$\dot{\vec{R}}_i = \frac{\vec{P}_i}{M_i} \tag{3.15}$$

$$\dot{\vec{P}}_i = \vec{F}_i - \xi \vec{P}_i \tag{3.16}$$

$$\dot{\xi} = \frac{1}{Q}\left(\sum_i \frac{\vec{P}_i^2}{M_i} - g k_B T\right) = \frac{1}{Q}\,(T(t) - T_0) \tag{3.17}$$

where $(T(t) - T_0)$ is the difference between the actual temperature of the system and the reference one. Q is the thermal inertia parameter, also referred as the "mass" of the oscillator ξ, which determines the rate of the heat transfer, i.e. the strength of the bath coupling:

$$Q = \frac{\tau^2 T_0}{4\pi^2} \tag{3.18}$$

where τ is the period of the oscillations of kinetic energy between the system and the reservoir. This thermostat has been demonstrated to be non-ergodic in some cases. It was then improved and made ergodic by adding a chain of thermostats (Nosé-Hoover chain) [112].

Berendsen barostat. In the NPT ensemble MD simulations, the pressure is controlled by coupling the system with a barostat. Berendsen and coworkers have developed a first-order coupling barostats [113] in conjunction with the temperature control method (not detailed in this work), such that the pressure of the simulated system $P(t)$ is relaxed toward the reference pressure P_0 with a time constant τ_P:

$$\frac{dP(t)}{dt} = \frac{1}{\tau_P}[P_0 - P(t)]. \tag{3.19}$$

The coordinates, x_{new}, and the volume of the box, V_{new}, are rescaled by a scaling factor μ at every step, such that:

$$x_{new} = \mu x_{old} \tag{3.20}$$

$$V_{new} = \mu^3 V_{old} \tag{3.21}$$

$$\mu = \sqrt[3]{1 - \frac{\beta \delta t}{\tau_P}\big(P_0 - P(t)\big)} \tag{3.22}$$

where β is the compressibility of the system. Berendsen barostat is a weak coupling scheme that might be more suitable for pressure equilibration than for MD production run because the length scaling can lead to violent oscillations of the pressure.

Parrinello Rahman barostat. The pressure control due to Parrinello and Rahman [114, 115], developed in the 80s', allows the simulation box to change its shape by considering as new variables in the system the nine component of the unit cell vectors. This is represented by the matrix **h**, whose columns are the three vectors a, b and c, which arbitrarily define its shape. The cell volume can be written as:

$$V = det\, \boldsymbol{h} = \vec{a} \cdot \left(\vec{b} \times \vec{c}\right). \tag{3.23}$$

The position $\vec{R}_i$ of the particle i can be written in terms of **h** and a column vector representing the scaled coordinates, $\vec{S}_i = [\xi_i\ \eta_i\ \zeta_i]$, with $0 \leq \xi_i, \eta_i, \zeta_i \leq 1$.

$$\vec{R}_i = \boldsymbol{h}\vec{S}_i = \xi_i \vec{a} + \eta_i \vec{b} + \zeta_i \vec{c}. \tag{3.24}$$

The squared distance between particles i and j can be therefore rewritten as

$$R_{ij}^2 = S_i^T \boldsymbol{G} S_j \tag{3.25}$$

where **G** is a symmetric matrix, defined as the metric tensor:

$$\boldsymbol{G} = \boldsymbol{h}^T \boldsymbol{h}. \tag{3.26}$$

With the introduction of the scaled coordinates $\vec{S}_i$ for each atom i, the original Lagrangian of $3N$ variables becomes now an extended Lagrangian of $(3N + 9)$ variables, written as:

$$L_{PS} = \frac{1}{2}\sum_i M_i \dot{S}_i^T \boldsymbol{G} \dot{S}_i - \sum_{i<j} U\left(R_{ij}\right) + \frac{1}{2} W Tr\left(\dot{\boldsymbol{h}}^T \dot{\boldsymbol{h}}\right) - p_0 V \tag{3.27}$$

where $U\left(R_{ij}\right)$ is the pair potential, p_0 is the reference external applied pressure, V is the unit cell volume, W is constant of proportionality (with mass dimensionality) of the kinetic term associated with the time variation of **h**. The corresponding equations of motion are then derived for $\vec{S}_i$ and **h**.

3.2 Quantum mechanics

In classical mechanics, given the exact knowledge of the present state of a system it is possible to predict its future state by integrating the Newton's equations of motion. The atoms are treated as sphere particles (i.e., nuclei), which move according to a parametrized empirical FF, while electrons are not taken into account. As such, the

study of chemical reactions is impossible because the electronic degrees of freedom of the atoms are not explicitly considered. In quantum mechanics (QM), the state of a system is described by a function of the particles' coordinates and time, called wavefunction, $\Psi\left(\{\vec{r}_i\},\{\vec{R}_I\};t\right)$, which contains all the possible information about the system. This function explicitly takes into account electrons and nuclei and has the important property that $\Psi^*\Psi dx$ represents the probability of finding a particle in an infinitesimal region between x and $x+dx$ at time t, where dx is an infinitesimal element of length. As such, QM is statistical in nature, meaning that given the present state (i.e., position and velocity) of a system, the result of a position measurement cannot be predicted with certainty, but only with a probability. The probability of finding the state somewhere in the space is 1, which is the normalization condition. However, there exists a fundamental equation that tells how the wavefunction changes with time and provides important information about the state of a system [116]. This equation is called time-dependent Schrödinger equation:

$$H\Psi\left(\{\vec{r}_i\},\{\vec{R}_I\};t\right)=i\hbar\frac{\partial}{\partial t}\Psi\left(\{\vec{r}_i\},\{\vec{R}_I\};t\right) \tag{3.28}$$

where $\{\vec{r}_i\}$ and $\{\vec{R}_I\}$ are the position vectors of the electrons and nuclei of the system, respectively, t is the time, $\hbar$ is the Planck's constant divided by 2π, and i is the square root of -1. Ψ represents the total wavefunction and H the Hamiltonian of the system. Excluding the time dependency and the relativistic effect due to the mass and velocity, the Schrödinger equation can be rewritten in its time-independent form:

$$H\Psi\left(\{\vec{r}_i\},\{\vec{R}_I\}\right)=E\Psi\left(\{\vec{r}_i\},\{\vec{R}_I\}\right) \tag{3.29}$$

where E is the energy of the system and H, the Hamiltonian operator, is defined as:

$$\begin{aligned}H=&-\sum_I\frac{\hbar^2}{2M_I}\nabla_I^2-\sum_i\frac{\hbar^2}{2m_e}\nabla_i^2+\frac{1}{4\pi\varepsilon_0}\sum_{i<j}\frac{e^2}{|\vec{r}_i-\vec{r}_j|}-\frac{1}{4\pi\varepsilon_0}\sum_{I,i}\frac{e^2Z_I}{|\vec{R}_I-\vec{r}_i|}\\&+\frac{1}{4\pi\varepsilon_0}\sum_{I<J}\frac{e^2Z_IZ_J}{|\vec{R}_I-\vec{R}_J|}\end{aligned} \tag{3.30}$$

which, in atomic units becomes:

$$H=-\frac{1}{2}\sum_I\frac{1}{M_I}\nabla_I^2-\frac{1}{2}\sum_i\nabla_i^2+\sum_{i<j}\frac{1}{|\vec{r}_i-\vec{r}_j|}-\sum_{I,i}\frac{Z_I}{|\vec{R}_I-\vec{r}_i|}+\sum_{I<J}\frac{Z_IZ_J}{|\vec{R}_I-\vec{R}_J|} \tag{3.31}$$

In eq. 3.30, the first term represents the operator for the kinetic energy of the nuclei, the second term the operator for the kinetic energy of the electrons and the last three terms the electron-electron repulsive, electron-nuclear attractive and nuclear-nuclear repulsive Coulomb interactions, respectively. M_I and Z_I are the mass and atomic number of the I^{th} nucleus; m_e and $-e$ are the mass and charge of the electron

and ε_0 is the vacuum permittivity. To each classical observable (i.e., energy, position, momentum, etc.), corresponds a QM operator that, acting on the wave function, returns the expectation value of this operator, which can be considered the average value. The Hamiltonian operator is the energy operator that acts on the wavefunction and returns it multiplied by a scalar, E. Thus, Ψ is the eigenfunction of H with the eigenvalue E, and to solve the Schrödinger equation both are needed. Unfortunately, the exact solution can be provided for only a few problems, such as the particle in a box, the harmonic oscillator and the hydrogen atom, containing only one electron. No exact solution exists for system with more than one electron [116]. A one electron wavefunction is called "orbital", and an orbital for an electron in an atom is called "atomic orbital". The wavefunction of an interacting many-body system is not known *a priori*. Therefore, the Schrödinger equation is calculated from a trial wave-function Ψ_T, which is just Ψ expressed as a linear combination of a complete orthonormal set of eigenfunctions of H:

$$\Psi_T = \sum_n c_n \psi_n \tag{3.32}$$

$$H\psi_n = \mathcal{E}_n\psi_n \tag{3.33}$$

where c_n are time-independent coefficients, n indicates the state of the system ($n = 0$ is the ground-state), ψ_n is the set of orthonormal basis functions and $\mathcal{E}_n$ is the energy associate to the state n. Because of the completeness of ψ_n, any wavefunction can be written as in 3.32, but the obtained trial wavefunction represents a state that does not have a definite energy. In order to approximate the true ground-state wave-function, the variational principle is applied. This postulates that the energy expectation value, E_T, of a Hamiltonian, H, calculated with the trial wavefunction, Ψ_T, is always an upper bound to the exact ground-state energy, $\mathcal{E}_0$, calculated with the true ground-state wavefunction ψ_0. This can be expressed, using the bra–ket notation, as:

$$\langle\Psi_T|H|\Psi_T\rangle \geq \mathcal{E}_0. \tag{3.34}$$

In other words, the energy of any approximate wavefunction is always greater than or equal to the exact ground state energy. Therefore, it is possible to find a good approximation of the ground-state wave-function ψ_0 and energy $\mathcal{E}_0$. This explains the strategy of the variational method: since the energy of any approximate trial function is always above the true energy, then any variations in the trial function which lower its energy are necessarily making the approximate energy closer to the true ground state energy. As remarked above, the Schrödinger equation cannot be solved exactly for any atoms or molecules with more than one electron, thus any solutions for polyelectronic atoms or molecules can only be approximations to real ones. Quantum chemistry is

based on the Schrödinger equation within the Born-Oppenheimer (BO) approximation. The key of this approximation lies in the fact that nuclei are much heavier than electrons and so they move slowly. On the other hand, the electrons move much faster than nuclei, adjusting almost instantaneously to any changes in the position of the nuclei. As such, the nuclei can be regarded as fixed while the electrons carry out their motion as a cloud. Under the Born-Oppenheimer approximation the kinetic energy operator for the nuclei can be dropped from the Hamiltonian, which becomes:

$$H_{ele} = -\frac{1}{2}\sum_i \nabla_i^2 + \sum_{i<j} \frac{1}{|\vec{r}_i - \vec{r}_j|} - \sum_{I,i} \frac{Z_I}{|\vec{R}_I - \vec{r}_i|} + \sum_{I<J} \frac{Z_I Z_J}{|\vec{R}_I - \vec{R}_J|}. \tag{3.35}$$

Here, H_{ele} is the electronic Hamiltonian given by the sum of the kinetic energy of the electrons, the electron-electron repulsive and electron-nuclei attractive potentials, and the nuclear energy due to the electrostatic repulsion between nuclei. This last quantity is constant for a given nuclear configuration, hence, by omitting also this term, a pure electronic Schrödinger equation is left, still parametrically dependent on the nuclear coordinates. When the Born-Oppenheimer approximation is used, the Schrödinger equation is solved only for the electrons for each arrangement of the nuclei. When the position of the nuclei changes, it becomes necessary to consider the nuclei electrostatic repulsion energy and re-calculate the total energy.

3.2.1 Density functional theory

One of the principal task of quantum chemists is to find approximate solutions to the Schrödinger equation. There are several approximations which are commonly used to this aim, including the density functional theory (DFT), Hartree-Fock (HF), the full configuration interaction (full CI) and the semiempirical methods. DFT allows to deal with large molecular systems with more computational efficiency than post-HF methods (not detailed in this thesis) and broad applicability for many chemical/physical problems [116]. DFT foundations lie on the idea that the ground-state properties (like the ground-state energy) of a quantum system with *N* particles are uniquely defined by its electron density $\rho(\vec{r})$:

$$\rho(\vec{r}) = N \int |\psi(\vec{r}_1, \dots, \vec{r}_N)|^2 d\vec{r}_1 \cdots d\vec{r}_N. \tag{3.36}$$

The electron density $\rho(\vec{r})$ is a function that depends on 3 variables only, i.e. the spatial coordinates, plus a spin coordinate in the case of spin polarized DFT. As such, with respect to the wavefunction-based methods in which the wavefunction depends on $3N$ ($\psi(\vec{r}_1, \dots, \vec{r}_N)$) spatial (plus N spin) coordinates, in DFT the dimensionality is reduced to 3, also for many-electron systems. A functional is a function that maps

another function into a number. In the case of DFT, the mapped function is the ground-state electron density $\rho(\vec{r})$, and every observable of a quantum system can be written as a functional of it, returning the properties of the systems. They can either have a simple dependency ("local" functionals) upon $\rho(\vec{r})$ or they can depend upon the gradients of $\rho(\vec{r})$ ("non-local" functionals). The two Hohenberg-Kohn theorems [117] constitute the foundations of DFT and describe how the properties of a system can be determined by its electron density function:

i. The first Hohenberg-Kohn theorem states that the external potential, v_{ext}, acting on the electrons and due to the nuclear charges, is a unique functional of the ground-state electron density $\rho(\vec{r})$.

$$v_{ext}(\vec{r}) = -\sum_{I} \frac{Z_I}{|\vec{R}_I - \vec{r}|} + \sum_{I<J} \frac{Z_I Z_J}{|\vec{R}_I - \vec{R}_J|} \tag{3.37}$$

where the first term is electron-nuclear attraction and the last term is the nuclear-nuclear repulsion. Since the Hamiltonian of the system depends on $v_{ext}(\vec{r})$ and on $\rho(\vec{r})$, its full many-electron ground-state is a unique functional of $\rho(\vec{r})$ and the total ground-state energy is therefore written as a functional of $\rho(\vec{r})$. Within the Born-Oppenheimer approximation, the second term of $v_{ext}(\vec{r})$ in 3.37 drops and the energy functional is written as:

$$E_0[\rho(\vec{r})] = T[\rho(\vec{r})] + E_{e-e}[\rho(\vec{r})] + E_{n-e}[\rho(\vec{r})] \tag{3.38}$$

where the kinetic energy of the electrons, $T[\rho(\vec{r})]$, and the electron-electron repulsion, $E_{e-e}[\rho(\vec{r})]$ depend exclusively on the coordinates of the electrons, while $E_{n-e}[\rho(\vec{r})]$ is the nuclei-electrons attraction term. Therefore, the first two terms can be grouped together into the universal Hohenberg-Kohn functional, $F_{H-K}[\rho(\vec{r})]$, independent of the external potential v_{ext}:

$$F_{H-K}[\rho(\vec{r})] = T[\rho(\vec{r})] + E_{e-e}[\rho(\vec{r})]. \tag{3.39}$$

The energy functional can be rewritten as:

$$E_0[\rho(\vec{r})] = F_{H-K}[\rho(\vec{r})] + \int \rho(\vec{r})\, v_{ext}(\vec{r}) d\vec{r}. \tag{3.40}$$

ii. The second Hohenberg-Kohn theorem introduces the variational principle into DFT in order to determine the ground-state density of a system, $\rho_0(\vec{r})$. This theorem states that for a trial electron density, $\tilde{\rho}(\vec{r})$, of a system with an external potential $v_{ext}(\vec{r})$, its energy $E_v[\tilde{\rho}(\vec{r})]$ constitutes an upper bound to the true ground-state energy, E_0:

$$E_v[\tilde{\rho}(\vec{r})] = F_{H-K}[\rho(\vec{r})] + \int \rho(\vec{r})\, v_{ext}(\vec{r}) d\vec{r} \geq E_0. \tag{3.41}$$

This means that in order to find to find the ground state energy E_0, it is necessary to minimize the energy functional $E_v[\tilde{\rho}(\vec{r})]$ by applying the variational principle. Considering the constraint on the number of electrons *N*, this leads to:

$$\delta\left\{E_v[\tilde{\rho}(\vec{r})] - \mu\left[\int \rho(\vec{r}) d\,\vec{r} - N\right]\right\} = 0 \tag{3.42}$$

and the corresponding Euler equation is given by:

$$\mu = \frac{\delta F_{H-K}[\rho(\vec{r})]}{\delta\rho(\vec{r})} + v_{ext}(\vec{r}) \tag{3.43}$$

where μ is the Lagrange multiplier associated with the constant *N*.

The difficulty of equation 3.38 is that the explicit form of $T[\rho(\vec{r})]$ and $E_{e-e}[\rho(\vec{r})]$ is not known and thus also of the functional $F_{H-K}[\rho(\vec{r})]$, which is needed to apply in practice the variational principle. The electron-electron repulsion functional $E_{e-e}[\rho(\vec{r})]$ can be divided into a part that arises from the classical interaction between two charge densities, $J[\rho(\vec{r})]$, which can be derived from the Hartree approach, and a non-classical electron-electron Coulombic energy, $V_q[\rho(\vec{r})]$, which instead remains unknown:

$$E_{e-e}[\rho(\vec{r})] = J[\rho(\vec{r})] + V_q[\rho(\vec{r})] = \frac{1}{2}\iint \frac{\rho(\vec{r})\rho(\vec{r}')}{|\vec{r} - \vec{r}'|} d\vec{r} d\vec{r}' + V_q[\rho(\vec{r})]. \tag{3.44}$$

Hence, the main challenge of DFT is to find an explicit expression for $T[\rho(\vec{r})]$ and $V_q[\rho(\vec{r})]$. In 1965, Kohn and Sham tackled this problem with the so-called Kohn and Sham equations [118], suggesting a practical way to solve the Hohenberg-Kohn theorem for a set of interacting electrons. Kohn and Sham assumed that each physical system of *N* interacting electrons can be mapped into a corresponding reference system of *N* non-interacting electrons (Kohn-Sham electrons described by the Kohn-Sham orbitals ψ_i^{KS}) that feels an effective potential (Kohn-Sham potential, v_{KS}) so that its ground state density is identical to that of the system with interacting electrons. For the reference system, the ground-state density and the kinetic energy, $T_s[\rho(\vec{r})]$, can be written as a summation of one-electron orbital:

$$\rho(\vec{r}) = \sum_{i=1}^{N} \left|\psi_i^{KS}(\vec{r})\right|^2 \tag{3.45}$$

$$T_s[\rho(\vec{r})] = -\frac{1}{2}\sum_{i=1}^{N}\int \psi_i^{KS*}(\vec{r})\nabla^2\psi_i^{KS}(\vec{r})d\vec{r}. \tag{3.46}$$

The kinetic energy of the real system, $T[\rho(\vec{r})]$, can be expressed as a sum of the kinetic energy of the reference system, $T_s[\rho(\vec{r})]$, and the kinetic energy that measure the electron correlation, $T_c[\rho(\vec{r})]$:

$$T[\rho(\vec{r})] = T_s[\rho(\vec{r})] + T_c[\rho(\vec{r})]. \tag{3.47}$$

Consequently, the Hohenberg-Kohn functional in eq. 3.39 can be written in terms of the non-interacting system orbitals:

$$\begin{aligned} F_{H-K}[\rho(\vec{r})] &= T[\rho(\vec{r})] + E_{e-e}[\rho(\vec{r})] \\ &= T_s[\rho(\vec{r})] + T_c[\rho(\vec{r})] + J[\rho(\vec{r})] + V_q[\rho(\vec{r})]. \end{aligned} \tag{3.48}$$

The unknown terms in 3.48, i.e., the kinetic energy that measure electron correlation, $T_c[\rho(\vec{r})]$, and the non-classical electron-electron interaction, $V_q[\rho(\vec{r})]$, can be collected in the exchange correlation functional, $E_{xc}[\rho(\vec{r})]$ as:

$$E_{xc}[\rho(\vec{r})] = T_c[\rho(\vec{r})] + V_q[\rho(\vec{r})] \tag{3.49}$$

$$F_{H-K}[\rho(\vec{r})] = T_s[\rho(\vec{r})] + J[\rho(\vec{r})] + E_{xc}[\rho(\vec{r})]. \tag{3.50}$$

By definition $E_{xc}[\rho(\vec{r})]$ is the only unknown term that has to be approximated. The total electronic energy for the real system can then be rewritten as:

$$\begin{aligned} E_{KS}[\rho(\vec{r})] &= F_{H-K}[\rho(\vec{r})] + \int \rho(\vec{r})\, v_{ext}(\vec{r})d\vec{r} \\ &= -\frac{1}{2}\sum_{i=1}^{N}\int \psi_i^{KS*}(\vec{r})\nabla^2\psi_i^{KS}(\vec{r})d\vec{r} + \frac{1}{2}\iint \frac{\rho(\vec{r})\rho(\vec{r}')}{|\vec{r}-\vec{r}'|}\,d\vec{r}d\vec{r}' \\ &\quad + E_{xc}[\rho(\vec{r})] + \int \rho(\vec{r})\, v_{ext}(\vec{r})d\vec{r}. \end{aligned} \tag{3.51}$$

This equation acts to define the exchange correlation functional $E_{xc}[\rho(\vec{r})]$. By applying the variational principle to the equation 3.51, imposing the wavefunction orthonormality condition and using the Lagrange multiplier method, the previous equation 3.43 turns into:

$$\mu = \frac{\delta T_S[\rho(\vec{r})]}{\delta\rho(\vec{r})} + \frac{\delta J[\rho(\vec{r})]}{\delta\rho(\vec{r})} + \frac{\delta E_{xc}[\rho(\vec{r})]}{\delta\rho(\vec{r})} + v_{ext}(\vec{r}) \tag{3.52}$$

from which, taking together the last three terms, the Kohn-Sham effective potential $v_{KS}(\rho)$ reads as:

$$v_{KS}(\rho) = v_{EFF}(\rho) = \int \frac{\rho(\vec{r}')}{|\vec{r} - \vec{r}'|} d\vec{r}' + v_{xc}(\rho) + v_{ext}(\rho) \tag{3.53}$$

where the first term is the Hartree potential, $v_{ext}(\vec{r})$ is the external potential acting on the interacting system, which within the Born Oppenheimer approximation corresponds to the electron-nuclei interaction, while $v_{xc}(\rho)$ is the exchange-correlation potential defined as the functional derivative of E_{xc} with respect to $\rho(\vec{r})$:

$$v_{xc}(\rho) = \frac{\delta E_{xc}[\rho(\vec{r})]}{\delta \rho(\vec{r})}. \tag{3.54}$$

Equation 3.52 says that it is possible to find the ground-state energy for a system of non-interacting electrons in an effective potential, which is the same as the exact ground-state density. A set of Kohn-Sham equations as functions of the Kohn-Sham orbitals $\{\psi_i^{KS}\}$ is then derived, from which it is possible to obtain the ground-state electron density by iteratively solving them:

$$H^{KS}\psi_i^{KS}(\vec{r}) = \left[-\frac{1}{2}\nabla^2 + v_{KS}(\rho)\right]\psi_i^{KS}(\vec{r}) = \varepsilon_i \psi_i^{KS}(\vec{r}) \qquad i = 1, \dots, N. \tag{3.55}$$

Since the Kohn-Sham potential depends on the density, it is necessary to solve these equations self-consistently i.e. (i) making a guess for the density, (ii) constructing $v_{KS}(\rho)$, (iii) solving the Kohn-Sham equations from where a new set of orbitals $\{\psi_i^{KS}(\vec{r})\}$ is obtained, (iv) computing a new density, and the repeating iteratively the process starting from (ii) until the input and output densities are the same. The solution of these equations involves repeated matrix diagonalizations followed by self-consistent iterations, with a computational effort growing with the size of the problem. As a note of warning, the total energy of the real interacting system will not be the energy obtained from the sum of the one-electron energy eigenvalues ε_i, because this double-counts the Hartree energy and over-counts the exchange-correlation energy.

Exchange correlation functionals. In principle, the Kohn-Sham equations can be solved exactly, because DFT in his KS formulation is an exact theory. In practice the exchange correlation functional $E_{xc}[\rho(\vec{r})]$ is not known, such that approximations have to be introduced. One strenght of DFT is that even simple approximations of E_{xc} can provide good results. Physically, this term can be interpreted as containing the contributions of detailed correlation (i.e., interaction of electrons with opposite spin multiplicity) and exchange (i.e., interaction between electrons with the same spin multiplicity) to the system energy. However, the actual form of E_{xc} is not known. There are two approximations which are commonly used: the "Local Density Approximations" (LDAs), and the "Generalized Gradient Approximations" (GGAs). The LDA was proposed by Kohn and Sham [118, 119], in which the exchange-

correlation energy is assumed to be the same as the one of a uniform electron gas, at every point in the space:

$$E_{xc}^{LDA}[\rho(\vec{r})] = \int \rho(\vec{r})\varepsilon_{xc}^{hom}[\rho(\vec{r})]d\vec{r} = \int \rho(\vec{r})\{\varepsilon_{x}^{hom}[\rho(\vec{r})] + \varepsilon_{c}^{hom}[\rho(\vec{r})]\}d\vec{r} \quad (3.56)$$

where $\varepsilon_{xc}^{hom}[\rho(\vec{r})]$ is the exchange correlation energy per electron as a function of the density $\rho(\vec{r})$ of a uniform gas. For a homogenous electron gas the exchange contribution is exactly known:

$$\varepsilon_{x}^{hom}[\rho(\vec{r})] = -\frac{3}{4}\left(\frac{3\rho(\vec{r})}{\pi}\right)^{\frac{1}{3}}. \quad (3.57)$$

Instead, for the correlation contribution the situation is more complicated as it has only been determined analytically in the high and low density limit [120, 121] and by highly accurate quantum Monte-Carlo calculations [122]. Suitable analytical expressions to calculate ε_{c}^{hom} have been developed by Vosko, Wilk and Nusair [123] and by Perdew and Wang [124]. However, for chemical applications LDA is not a useful model because molecules are in general highly inhomogeneous systems. Better results were obtained with the development of GGAs, in which E_{xc} not only depends on the local density $\rho(\vec{r})$, but also on the local gradient $\nabla\rho(\vec{r})$ of the density:

$$E_{xc}^{GGA}[\rho(\vec{r})] = \int \rho(\vec{r})\varepsilon_{xc}[\rho(\vec{r}), |\nabla\rho(\vec{r})|]d\vec{r} = E_{x}^{GGA}[\rho(\vec{r})] + E_{c}^{GGA}[\rho(\vec{r})]. \quad (3.58)$$

The BLYP exchange correlation functional [125, 126] is one of the most applied GGA functionals in biological systems simulations and it also used in this thesis. It provides good results for all the main bond types without being computationally expensive. However, it sometimes underestimates barrier heights and fails in the description of dispersion interactions. BLYP is given by a combination of the Becke exchange functional and the Lee-Yang-Parr correlation functional (BLYP). The Becke exchange functional is defined as:

$$E_{x}^{Becke}[\rho(\vec{r})] = E_{x}^{LDA}[\rho(\vec{r})] - \beta \int \rho(\vec{r})^{\frac{4}{3}} \frac{x^2}{1 + 6\beta\, sinh^{-1} x} d\vec{r} \quad (3.59)$$

where x gives a measure of the local inhomogeneity of the system:

$$x = \frac{|\nabla\rho(\vec{r})|}{\rho(\vec{r})^{\frac{4}{3}}} \quad (3.60)$$

and β has been fixed by Becke to 0.0042 au upon a fit on exact HF calculations for noble gases from He to Rn. In the Lee-Yang-Parr correlation functional, the correlation energy is computed from HF second order density matrices and it is written as:

$$E_c^{LYP}[\rho(\vec{r})] = -a\int \frac{1}{1+d\rho(\vec{r})^{-\frac{1}{3}}}\Bigg\{\rho(\vec{r}) + b\rho(\vec{r})^{-\frac{2}{3}}\Bigg[C_F\rho(\vec{r})^{\frac{5}{3}} - 2t_w + \left(\frac{1}{9}t_w + \frac{1}{18}\nabla^2\rho(\vec{r})\right)\Bigg]e^{-c\rho(\vec{r})^{-\frac{1}{3}}}\Bigg\}d\vec{r} \tag{3.61}$$

where:

$$C_F = \frac{3}{10}(3\pi^2)^{\frac{2}{3}} \tag{3.61}$$

$$t_w = \frac{1}{8}\left(\frac{|\nabla\rho(\vec{r})|^2}{\rho(\vec{r})} - \nabla^2\rho(\vec{r})\right) \tag{3.62}$$

with a = 0.04918 au, b = 0.132 au, c = 0.2533 au and d = 0.349 au obtaining upon the fitting on HF calculations on He noble gas. In DFT, the correlation effects are incorporated from the beginning as approximation of E_{xc}, thus introducing a self-interaction error that does not automatically cancels as in HF and can generate drawbacks in the calculations employing LDAs or GGAs. Hence, Becke developed the so-called hybrid functionals, in which a certain amount of the exact non-local Hartree-Fock exchange is mixed with the GGA exchange-correlation functional. One of the most popular hybrid exchange correlation functional, also used in this thesis, is the B3LYP [126, 127], which has the form of:

$$E_{xc}^{B3LYP} = E_{xc}^{LDA} + a_0(E_x^{HF} - E_x^{LDA}) + a_x(E_x^{Becke} - E_x^{LDA}) + E_c^{LDA} + a_c(E_c^{LYP} - E_c^{LDA}) \tag{3.63}$$

where the three parameters are fitted to reproduce atomic and spectroscopic data and have value of 0.20, 0.72 and 0.81 au, respectively. The development of hybrid exchange correlation functionals has provided higher accuracy of DFT calculations for molecules. However, its use for *ab-initio* MD simulations is limited by the increased computational cost. More accurate functionals are still under development and they still represent one of the main topics of research for the improvement of DFT. For example, more recent exchange correlation functionals are represented by the Minnesota Functionals (M*yz*), developed by the group of Prof. Donald Truhlar. These functionals, like the M06 family [128, 129], are recommended for application in organometallic and inorganometallic chemistry and for noncovalent interactions. They are meta hybrid GGA functionals that contain a large number of free parameters in the functional form. These parameters are semiempirically fit using broad experimental

data sets that include noncovalent interactions. This approach has been used for the M06 suite of functionals which differ in the amount of the exact exchange included, with M06 including 27% of the HF exchange, M06-2X including 54%, and M06-HF including 100%. Dispersion interactions have historically been difficult to account for in the most widely used density functionals, due to the difficulties in identifying appropriate long-range correlation expressions that take proper account of the delicate balance between exchange and correlation in a functional. The M06 suite has shown a very good response under dispersion forces, representing an improvement with respect to B3LYP. M06 has been used in this thesis along with BLYP and B3LYP.

Basis set. In the actual implementations of Kohn-Sham equations (3.55) for *ab-initio* simulations, the Kohn-Sham orbitals $\{\psi_i^{KS}(\vec{r})\}$ are expanded as linear combination of basis functions. A basis set is a set of functions used to create the molecular orbitals (MO). Expanding a molecular orbital in this way is not an approximation if an infinite number of functions is used. However, this is impossible in real simulations. In practice, a finite number of functions is always used and only the components of the MO along those coordinate axes corresponding to the selected basis can be represented. The smaller the basis, the poorer the representation of the MO and the faster the calculations. MOs are usually expanded as a linear combination of *M* basis functions, χ_α, with well-known behavior, which, however, are not the exact atomic orbitals (AO) as their analytic formulas contain simplifications:

$$\psi_i(\vec{r}) = \sum_{\alpha=1}^{M} c_{i\alpha}\chi_\alpha \ \left(\vec{r};\{\vec{R}_I\}\right) \tag{3.64}$$

where $c_{i\alpha}$ are the orbital expansion coefficients. The general procedure is to pick up a basis set and vary the $c_{i\alpha}$ in order to find the $\psi_i(\vec{r})$ orbitals that yield a density as close as possible to the ground-state density. In this thesis, I used localized Gaussian-type basis functions for ligands parameterizations and also for DFT calculations performed with Gaussian09, whereas I employed plane wave basis set in QM/MM MD simulations.

Localized basis sets. The two most widely used atomic centered basis functions (also called AO) in electronic structure calculations are the Slater-type basis functions (STOs) [130] and the Gaussian-type basis functions (GTOs) [131].

The STOs are described by the function depending on spherical coordinates:

$$\chi^S_{\zeta,n,l,m}(r,\theta,\phi) = NY_{l,m}(\theta,\phi)r^{n-1}e^{-\zeta r} \tag{3.65}$$

where *N* is a normalization constant, $Y_{l,m}$ are the spherical harmonic functions, n, l and m are the quantum numbers, r, θ, ϕ are spherical coordinates with r being the distance

of the electrons from the nucleus, ζ ("the exponent") is a constant related to the effective charge of the nucleus.

Instead, the Gaussian-type basis functions (GTOs), called also cartesian gaussians, have the form of (in cartesian coordinates):

$$\chi^{G}_{\zeta,l_x,l_y,l_z}(x,y,z) = Nx^{l_x}y^{l_y}z^{l_z}e^{-\zeta r^2} \tag{3.66}$$

where N is a normalization constant, l_x, l_y, l_z are not quantum numbers but integral exponents at the cartesian coordinates x, y, z. Their sum determines the type of orbital (for example $l_x + l_y + l_z = 1$ is a p orbital).

STOs are primarily used for atomic and diatomic systems where high accuracy is required. The exponential dependence on the distance between the nucleus and the electron mirrors the exact decay behavior of the orbitals for the hydrogen atom. STOs are very suitable for expanding MOs because they have the correct shape near and far from the nucleus, but they cannot be analytically integrated. The main difference of GTOs is that the variable r in the exponential function is squared, thus influencing the representation of the orbital in the region close to the nucleus and at long distances. Three times as many GTOs as STOs are required to reach a given level of accuracy. GTOs are not really orbitals but rather simpler functions frequently called gaussian primitives (PGTOs). In practice, fixed linear combinations of cartesian PGTOs are commonly used to define a STO. These primitives (with fixed coefficients) are combined into contracted gaussians (CGs), each one thus being a linear combination of a given number of PGTOs. CGs are used to approximate STOs. The simplest kind of CG are the STO-*n*G where *n* is the number of gaussian primitives. STO-1G uses 1 gaussian primitive to form 1 CG per atomic orbital. More complicated basis sets are obtained by including more than one CG per atomic orbital (DZ "double zeta", TZ "triple zeta, QZ "quadruple zeta"). Another improvement is constituted by the Pople-style "*k-nlmG*" basis sets [132], where *k* is the number of PGTOs used for the core orbitals, while *nml* indicates both how many functions the valence orbitals are split into, and how many PGTOs are used for their representation. For instance, 6-31G indicates that the core orbitals are a contraction of 6 PGTOs and the valence orbitals are split into 2 parts (*nm*), an inner and an outer part, represented by the contraction of 3 and 1 PGTOs, respectively. To each one of these basis sets one can add diffuse and/or polarization auxiliary functions. The formers are s- or p- functions indicated with + (if applied only on heavy atoms) or ++ (if also applied on hydrogens) before the G. The latters are indicated after the G, with * (d-type polarization function) or ** (d- and p-type polarization functions).

Plane wave basis set. Plane waves basis set is the most convenient basis set used in simulations where the system is subjected to periodic conditions. Plane waves are defined as:

$$f_G^{PW} = \frac{1}{\sqrt{V}} e^{i\vec{G}\vec{r}}, \tag{3.67}$$

where V is the volume of the periodic cell and $\vec{G}$ is the reciprocal lattice vector. Following the Bloch theorem [133], stating that electronic wavefunctions can be expanded in terms of a discrete plane wave basis set, the Kohn-Sham orbitals can be written as:

$$\psi_i(\vec{r},\vec{k}) = e^{i\vec{k}\vec{r}} u_i(\vec{r},\vec{k}), \tag{3.68}$$

where $\vec{k}$ is a wave vector in the first Brillouin zone (i.e., a uniquely defined primitive cell) of the reciprocal lattice, and $u_i(\vec{r},\vec{k})$ is a cell-periodic function;

$$u_i(\vec{r},\vec{k}) = u_i(\vec{r}+\vec{R},\vec{k}). \tag{3.69}$$

The periodic functions can be expanded as a Fourier series and so $u_i(\vec{r},\vec{k})$:

$$u_i(\vec{r},\vec{k}) = \frac{1}{\sqrt{V}} \sum_{\vec{G}} c_i(\vec{G},\vec{k}) e^{i\vec{G}\vec{r}} \tag{3.70}$$

where c_i is the first Fourier component of the plane wave expansion. Kohn-Sham orbitals become:

$$\psi_i(\vec{r},\vec{k}) = \frac{1}{\sqrt{V}} \sum_{\vec{G}} c_i(\vec{G},\vec{k}) e^{i(\vec{G}+\vec{k})\vec{r}}. \tag{3.71}$$

The expansion is truncated at a given energy cutoff, determining the accuracy of the calculation and the number of included plane waves that have a kinetic energy smaller than the cutoff energy.

Pseudopotentials. One important problem of plane wave basis set is that, near to the atomic nuclei, the valence electron wavefunctions show rapid oscillations because they must be orthogonal to the core electron wavefunctions, making impossible a proper description. These large oscillations determine a large kinetic energy and this problem becomes even bigger for heavy atoms with many core electrons. In order to minimize the size of the plane wave basis set, a further approximation is introduced. The true potential of core electrons (i.e., the electrons close to the nuclei) is replaced by a weaker potential, called pseudopotential [134-136], which takes into account how the nucleus and the core electrons interact with the valence electrons. This approximation is made on the observation that core electrons are not involved explicitly in the description of chemical reactions and are unaffected by the chemical

environment and can be subsumed into the nuclear core. Therefore, core electrons effects are implicitly included in the nuclear potential, thus forming a pseudopotential which takes into account joint effect of nucleus and core electrons on the motion of the valence electrons. The pseudopotential is a potential function by which the wavefunctions have the same shape as the true wavefunctions outside the core region, but with fewer nodes inside the core and thus becoming much smoother. As a consequence, the number of plane waves needed to expand the wavefunctions can be considerably reduced. Pseudopotential can be derived from *ab-initio* calculations of isolated atoms by solving the Kohn-Sham equations. Several procedures have been proposed for generating pseudopotentials to be used in combination with a plane waves basis set. The "norm-conserving" pseudopotentials derived through the Martins-Troullier (MT) scheme [137] are used in this thesis.

3.2.2 *Ab-initio* MD

Force field-based MD simulations are a great tool to estimate equilibrium and non-equilibrium properties of condensed systems. However, one drawback of classical MD simulations is the impossibility of studying chemical reactions and describing bond breaking and formation as they are based on empirical interatomic potentials without taking into account the electronic degrees of freedom. On the other hand, *ab-initio* MD (AIMD) simulations overcome this issue, as they provide an accurate electronic description of the chemical bond. They allow the derivation of the forces acting on the investigated system directly from electronic structure calculations, which are performed *"on-the-fly"* as the MD trajectory is generated. In this way, AIMD are used to investigate the time-evolution of a chemical state of a system and to study chemical reactions observing bonds breaking and formation.

Born-Oppenheimer MD (BOMD). BOMD is based on the Born-Oppenheimer approximation which separates the electrons and nuclei motion on the basis of their different masses. As the electrons are three orders of magnitude lighter than nuclei they are supposed to follow instantaneously the nuclear motion like an electronic cloud. Within this approximation, the nuclei are stationary and their positions become just parameters in the wavefunction for the electrons such that a pure electronic Schrödinger equation can be solved for each fixed nuclear configuration:

$$H_e \Psi_e = E_e \Psi_e \tag{3.72}$$

where E_e is the adiabatic contribution of the electrons to the energy of the system and H_e is the pure electronic Hamiltonian:

$$H_e = -\frac{1}{2}\sum_i \nabla_i^2 + \sum_{i<j}\frac{1}{|\vec{r}_i - \vec{r}_j|} - \sum_{I,i}\frac{Z_I}{|\vec{R}_I - \vec{r}_i|}. \tag{3.73}$$

In BOMD, the electronic problem is solved at each time step and forces are computed by minimizing the KS energy functional at the present nuclear configuration. The electronic contribution to the forces acting on the ions is calculated via the Hellmann-Feynman theorem and the nuclei are moved according to the laws of classical mechanics as:

$$H_e\Psi_0 = E_0\Psi_0 \tag{3.74}$$

$$M_I\ddot{\vec{R}}_I(t) = -\nabla_I \min_{\Psi_0}\langle\Psi_0|H_e|\Psi_0\rangle \tag{3.75}$$

for the electronic ground-state. The minimum of $\langle H_e\rangle$ has to be reached in each BOMD step.

Car-Parrinello MD (CPMD). In 1985 Roberto Car and Michele Parrinello presented a new scheme which combined molecular dynamics and density functional theory, the Car-Parrinello (CP) Molecular Dynamics [138]. This method allowed the application of DFT to much larger systems performing *ab-initio* MD simulations. The only assumptions were the validity of classical mechanics to describe the ionic motions, and the BO approximation to separate the nuclear and electronic degrees of freedom. Their revolutionary idea was to introduce a Newtonian fictitious dynamics also for the electrons in addition to ions (nuclei) dynamics. According to this method, the fast electronic and the slow nuclear degrees of freedom evolve simultaneously following a set of classical equations of motion and are maintained adiabatically separated. In order to achieve this adiabatic separation, a fictitious electron mass μ is assigned to the electronic Kohn-Sham orbitals ψ_i, which are thus treated as fictitious classical particles. Forces are evaluated both on the nuclei and the electrons applying the KS equations. The basic equation of CP method is given by the Lagrangian as a functional of the wavefunction (KS orbitals, ψ_i) and the atomic positions $\vec{R}_I$, which reads:

$$\begin{aligned} L_{CP} = \sum_I \frac{1}{2}M_I\dot{\vec{R}}_I^2 + \sum_i \frac{1}{2}\mu_i\int d^3\vec{r}\left|\dot{\psi}_i\right|^2 - E\left[\{\psi_i\},\{\vec{R}_I\}\right] \\ + \sum_{ij}\lambda_{ij}\left(\int d^3\vec{r}\left(\psi_i^*(\vec{r},t)\psi_j(\vec{r},t)\right) - \delta_{ij}\right). \end{aligned} \tag{3.76}$$

In equation 3.76 the first term is the kinetic energy of the nuclei with mass M_I, the second term is the fictitious kinetic energy of the KS orbitals with mass μ_i, the third term is the Kohn-Sham energy functional, and the last term is the orthogonality

constraint, where δ_{ij} is the Kronecker delta. This Lagrangian coincides with the true Lagrangian in the limit of the fictitious mass going to zero, where the dynamics approach the BOMD. The corresponding Newtonian equations of motion for ions dynamics are:

$$M_I \ddot{\vec{R}}_I(t) = -\nabla_{R_I} E. \tag{3.77}$$

Instead, for the fictitious electrons dynamics:

$$\mu \ddot{\psi}_\iota(\vec{r}, t) = -\frac{\delta E}{\delta \psi_i^*} + \sum_j \Lambda_{ij} \psi_j(\vec{r}, t) = -H^{KS} \psi_i(\vec{r}, t) + \sum_j \Lambda_{ij} \psi_j(\vec{r}, t) \tag{3.78}$$

where Λ_{ij} are the Lagrange multipliers to ensure the orthonormality between the KS orbitals. The effect of the nuclei on the electrons is included in the Kohn-Sham Hamiltonian H^{KS}, whereas the nuclear motion depends also on the electronic degree of freedom through $E\left[\{\psi_i\}, \{\vec{R}_I\}\right]$. While the nuclei move at physical temperature, this is not true for the electrons, which move at a fictitious temperature. By tuning the fictitious mass is possible to ensure the adiabatic separation of the nuclei from the electrons. A small value of μ allows the electronic degree of freedom to be decoupled from the nuclear one and the wavefunctions to oscillate very rapidly around the Born-Oppenheimer energy surface, instantaneously adjusting to the nuclear coordinates during the dynamics. Higher the fictitious kinetic energy, more electrons will be far from the minimum energy configuration. In particular, a ground state wavefunction optimized at time t_0 will stay close to the ground state if it is kept at sufficiently low temperature, i.e. if the adiabatic separation is maintained. On the other hand, smaller μ means smaller integration step. Therefore, a compromise between accuracy and efficiency in the simulations has to be found when choosing the value of μ.

3.2.3 Hybrid quantum mechanics/molecular mechanics (QM/MM) MD

Hybrid quantum mechanics/molecular mechanics molecular dynamics simulations (QM/MM MD) were introduced in 1976 from Warshel and Levitt [139], signing the beginning of a new era. QM/MM MD filled the gap between pure QM calculations and classical MD studies of biological systems, becoming the method of choice for modelling chemical reactions. In fact, while the former are very accurate but computationally prohibitive to be applied on large biomolecular systems, the latter are based on empirical force fields, neglecting the quantum nature of chemical bonds. Since the reactive part of a system is usually restricted to a small (few hundreds) subset of atoms defining the active site or the catalytic center, the two approaches were combined such that the chemically active region was treated at QM level and the surrounding environment with the classical approach or an alternative lower theory.

Thus, with the combination of these two techniques and the development of different coupling schemes, studying chemical reactions in large biomolecules became computationally affordable. QM/MM MD represent an improvement over QM calculations *in vacuo*, as long as the effects of the environment are explicitly taken into account directly influencing the reactive atoms. As a general sketch of hybrid QM/MM approach, the system is partitioned in two parts, one treated at quantum level (QM region) and the other described by a force field (MM region) [140-143]. These two regions can accommodate almost any combination of QM and MM methods. For the QM part DFT and semiempirical methods are often used, while the MM part can be modeled with any available force field (AMBER, CHARMM, GROMOS, etc.). In the present thesis, I used the Car-Parrinello MD scheme for the quantum part, at DFT level of theory, and AMBER FF for the MM part. The two regions are interfaced and, as such, the total energy of the system is not simply a summation of the energies of the two subsystems but it contains contributions coming from the interaction between them. Moreover, a particular care must be devoted to the boundaries between QM and MM region, especially if the borders cut through chemical bonds. Several QM/MM coupling schemes have been developed to compute the energy of the entire system among which the additive QM/MM scheme, in which the total Hamiltonian of the system reads:

$$H = H_{MM} + H_{QM} + H_{QM/MM}. \tag{3.79}$$

H_{MM} is the classical Hamiltonian, including interactions between atoms in the classical part:

$$\begin{aligned} H_{MM} = & \sum_{bonds} K_r\,(r - r_0)^2 + \sum_{angles} K_\theta\,(\theta - \theta_0)^2 \\ & + \sum_{dihedrals} \frac{V_n}{2}[1 + cos(n\phi - \gamma)] \\ & + \sum_{i<j} \varepsilon_{ij}\left[\left(\frac{R_{min\,i,j}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{min\,i,j}}{r_{ij}}\right)^{6}\right] + \sum_{i<j}\frac{q_i q_j}{\varepsilon r_{ij}} \end{aligned} \tag{3.80}$$

while H_{QM} is the Car-Parrinello Lagrangian presented above, describing the interactions between QM atoms:

$$\begin{aligned} H_{QM} = L_{CP} = & \sum_I \frac{1}{2} M_I \dot{\vec{R}}_I^2 + \sum_i \frac{1}{2}\mu_i \int d^3\vec{r}\left|\dot{\psi}_i\right|^2 - E\left[\{\psi_i\},\{\vec{R}_I\}\right] \\ & + \sum_{ij} \lambda_{ij}\left(\int d^3\vec{r}\left(\psi_i^*(\vec{r},t)\psi_j(\vec{r},t)\right) - \delta_{ij}\right). \end{aligned} \tag{3.81}$$

The most critical term is $H_{QM/MM}$, which collects the interactions between QM and MM atoms. The definition of this term has been challenging and its exact form defines a particular QM/MM method. A typical scenario in QM/MM simulations is represented by a repartition in which the boundary between QM and MM cross at least one chemical bond. In such situation, $H_{QM/MM}$ includes bonded, van der Waals and electrostatic interactions between QM and MM atoms. In a different case, where the MM region is the solvent and QM part is the solute, the bonded interactions are not included. In particular, I used the QM/MM method developed by Laio, VandeVondele and Röthlisberger [144, 145]. The coupling term can be expanded as:

$$H_{QM/MM} = H_{QM/MM}^{bonded} + H_{QM/MM}^{non-bonded} \tag{3.82}$$

QM/MM bonded interactions. The first term of eq. 3.82 refers to the bonded interactions between QM and MM atoms, meaning stretching, bending and torsional terms between atoms belonging to QM and MM region. All these terms are described at classical level, by the empirical force field terms. The problem of dealing with bonded interactions arises when the QM/MM boundary cuts through a covalent bond. In this situation one or more QM atoms are left with unsaturated valence orbitals. There are several solutions to saturate these valence orbitals [146, 147], among which the use of a "capping" dummy-hydrogen atom acting as a linker between the QM and the MM atom. QM calculations are then performed on an electronically saturated system and the bond Q—M is described at the MM level. However, the link atom is not a part of the real system and its interaction with MM part are not included. Another solution is to use a link-atom pseudopotential. In this case, it is required to constrain the distance between the link atom and the QM neighbor atom to the QM equilibrium distance, to preserve the electronic structure in the center of the QM subsystem.

QM/MM non-bonded interactions. The second term in eq. 3.82 refers to non-bonded interactions between QM system and the MM atoms. This is composed by the van der Waals term accounting for dispersion attractive and Pauli repulsion interactions, and by the electrostatic Coulomb interactions between the QM electron density and the MM point charges:

$$E_{QM/MM}^{non-bonded} = E_{QM/MM}^{vdW} + E_{QM/MM}^{ele} \tag{3.83}$$

$$E_{QM/MM}^{vdW} = \sum_{\substack{i \in MM \\ j \in QM}} v_{vdW}(\vec{r}_i, \vec{r}_j). \tag{3.84}$$

While the van der Waals interactions (3.84) are treated using the classical force field as in the MM region, the description of the QM/MM electrostatic coupling is very

subtle as it often constitutes the main environmental effect on QM region. According to the Laio et al. scheme [144, 145], the second term in 3.83 reads as:

$$E_{QM/MM}^{ele} = \sum_{i \in MM} q_i \int \varrho(\vec{r}) v_i(|\vec{r} - \vec{r}_i|) d\vec{r} \tag{3.85}$$

where q_i is the partial charges of the classical atoms i at $\vec{r}_i$, $\varrho(\vec{r})$ is the total charge density of the QM atoms (electronic plus ionic), and $v_i(|\vec{r} - \vec{r}_i|)$ is a modified suitable Coulomb potential. The choice of this potential becomes crucial for the description of the electrostatic coupling both in terms of theoretical problems and of computational cost. A recurrent theoretical problem of QM/MM simulations is the so-called electron spill-out phenomenon, i.e. the unphysical accumulation of quantum charge density at the boundary of the QM region due to the presence of classical positive charges of the nearby MM region that act as electron traps. In order to avoid the incurrence of the spill-out, this potential has been modified in its short-range behavior:

$$v_i = \frac{r_{ci}^n - r^n}{r_{ci}^{n+1} - r^{n+1}}, \tag{3.86}$$

where n is usually fixed to 4 and r_{ci} is the covalent radius of atom i. Therefore, $v_i(\vec{r})$ behaves as r^{-1} for large distances, while for r values shorter than r_{ci} it goes smoothly to a finite value. Another important technical problem is that the full evaluation of the electrostatic interactions, i.e. the calculation of the integrals in equation 3.85, is computationally very demanding in a delocalized plane wave basis set. Therefore, the long-range electrostatic interactions are described according to a hierarchical scheme in which the MM region is divided in three shells surrounding the QM region and the calculations become less accurate when moving from the first closest shell to the farthest one. According to this partition scheme two cut-off radii (r_1 and r_2) from the QM boundary have to be set to define the three regions. The electrostatic interactions between the QM region and the MM atoms comprised in the first shell ($r < r_1$) are explicitly computed as in eq. 3.85. Instead, in the second shell ($r_1 < r < r_2$) the interactions are calculated between the classical point charges and the QM D-RESP point charges [148], computed on the fly and derived upon a fit on the electrostatic potential. Finally, in the third shell ($r > r_2$), the electrostatic interactions are calculated between the classical point charges of the MM atoms included in the shell and a multipolar expansion of QM charge density.

3.3 Enhanced sampling techniques and chemical reactions

Understanding and predicting the evolution of a complex biological systems, like protein or RNA macromolecules, through computational simulations has become more affordable in the latest years thanks to the remarkable growth and improvement of supercomputers, thus allowing computational biochemists to reach the micro- and even the millisecond time scale with their MD simulations. Moreover, the advent of QM/MM MD has offered the possibility of deciphering complex chemical reactions occurring in these large biomolecules, unveiling complicated mechanisms and the free energy change associated with them. However, chemical reactions are rare events and simple *ab-initio* molecular dynamics simulations would be insufficient to capture such special events as they would require a much too long simulation time without imposing any bias. The origin of this drawback is the inability to sample the entire accessible phase space within any reasonable amount of simulation time, even for rather simple systems and for the latest powerful supercomputers. To circumvent this problem, enhanced sampling methods have been developed to accelerate rare events with notable success. These techniques, by enhancing the sampling as a function of one or a few predefined collective variables relevant for the event, offer a strategy to study such events and reconstruct the free energy profiles, which would not be possible with brute-force simulations in a limited amount of time. Among these methods, in my thesis I used the thermodynamic integration and Blue-Moon ensemble simulations to study the mechanism of a hydrolysis reaction catalyzed by group II intron ribozymes and estimate the free energy profile (chapter 4). Hereafter I will briefly review this approach.

3.3.1 Thermodynamic integration and Blue-Moon ensemble

The method of thermodynamic integration (TI) [149-151] has become very popular among chemists and physicists and it is used to obtain free energy differences in a wide variety of systems [152]. TI is based on the idea of estimating the free energy difference between an initial and a final state as a function of a reaction coordinated (RC) by computing the integral of its derivatives along a reaction path defined by varying the RC. The RC, $\xi(\vec{r})$, is a function of the coordinates $\vec{r}$ that determines the state of the system and which can be monitored to look at the investigated rare event. The free energy of the system as a function of ξ can be written as:

$$F(\xi) = -k_b T \, ln \, P_\xi(\xi). \tag{3.87}$$

$P_\xi(\xi)$ is the canonical probability distribution of ξ, which reads:

$$P_{\xi}(\xi') = \langle \delta(\xi(\vec{r}) - \xi') \rangle \tag{3.88}$$

where ξ' indicates a given value of ξ and the brackets denote a standard statistical average over an equilibrium ensemble. Differentiating the eq. 3.87 with respect to ξ gives [151]:

$$\frac{\partial F(\xi)}{\partial \xi} = \left\langle \frac{\partial \mathcal{H}}{\partial \xi} \right\rangle_{\xi'}. \tag{3.89}$$

Then, according to the TI method, the free-energy difference between two states can be calculated using:

$$F(\xi_2) - F(\xi_1) = \int_{\xi_1}^{\xi_2} d\xi' \left\langle \frac{\partial \mathcal{H}}{\partial \xi} \right\rangle_{\xi'} \tag{3.90}$$

where the brackets denote a conditional average evaluated at $\xi = \xi'$ of the mechanical quantity $\partial \mathcal{H} / \partial \xi$ over the equilibrium ensemble of a system with Hamiltonian $\mathcal{H}(\vec{r}, \vec{p}, \xi)$. The negative value of the integrand in eq. 3.90 is what is generally called the "mean force". Thus, the free energy can be considered as the potential of mean force using this terminology. This ensemble average is readily obtained from any direct sampling scheme such as MD or Monte Carlo (MC). Since molecular simulations are performed with discrete steps, this integral has to be evaluated as a sum of ensemble averages [151]:

$$F(\xi_2) - F(\xi_1) = \sum_i \left\langle \frac{\partial \mathcal{H}}{\partial \xi} \right\rangle_{\xi_i} \Delta\xi_i \tag{3.91}$$

where i counts over the number of different values of ξ, and $\Delta\xi_i$ is the difference between successive values of ξ. However, a particular value of the RC can have a low probability such that cannot be directly sampled because the spontaneous occurrence is indeed a rare event. A convenient method to circumvent this problem is the so-called Blue-Moon ensemble method proposed by Carter and Ciccotti [149]. Their idea was introduced in 1989 to sample rare events that occur "once in a blue moon". According to this approach, the conditional average of equation 3.90 is estimated by a time average over a constrained trajectory in which the RC is fixed at a specific value such that:

$$\xi(\vec{r}) = \xi', \quad \dot{\xi}\left(\vec{r}, \dot{\vec{r}}\right) = 0. \tag{3.92}$$

Following the blue-moon ensemble method, the pathway along the RC starting from the reactant state value ($\xi = \xi_1$) until the product state value ($\xi = \xi_2$) can be divided into a number of intervals ($\Delta\xi_i$) in which ξ is progressively varied and kept fixed within each interval. Then the free energy difference between the two states is

derived as in 3.89 and 3.90. In order to ensure that the reaction coordinate ξ is constant and equal to ξ', such that the system is constrained to remain on the reaction path during the MD simulation, a modified Lagrangian with the Lagrange multiplier λ associated with the reaction coordinate is used:

$$L^*(\vec{r}, \vec{p}, \xi) = L(\vec{r}, \vec{p}) + \lambda(\xi(\vec{r}) - \xi'). \tag{3.93}$$

The value of the Lagrange multiplier is adjusted at each step of the simulation so that $\xi = \xi'$. The SHAKE algorithm can be used to determine the Lagrange multipliers. It can be shown that the free energy gradients $\partial\mathcal{H}/\partial\xi$ can be calculated using the Lagrange multipliers by comparing equation 3.89 to the Hamiltonian equation of motion of the reaction coordinate ξ [153]:

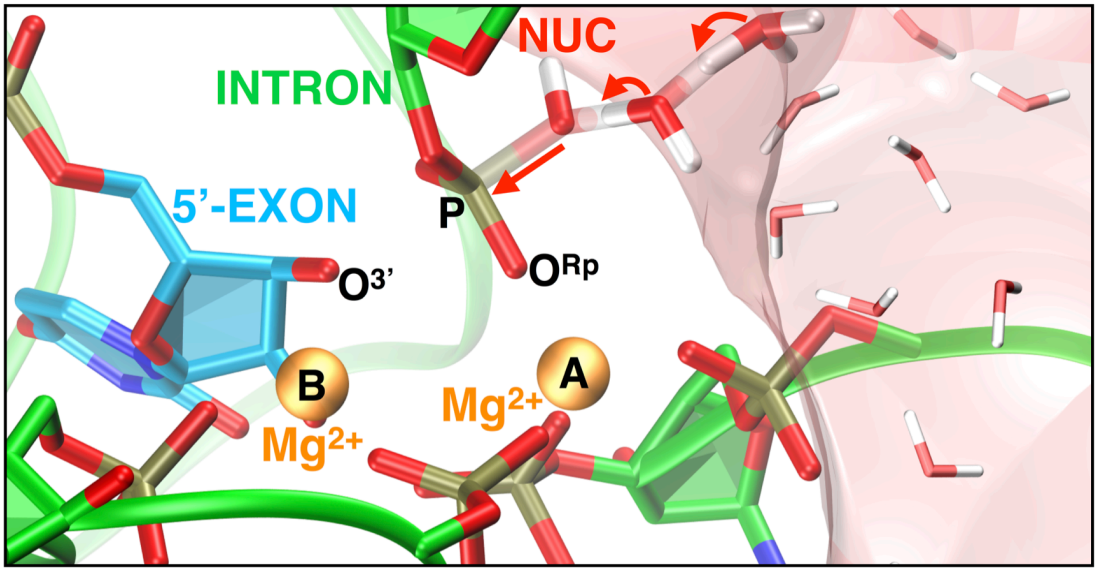
$$\dot{p}^{\xi} = -\frac{\partial \mathcal{H}}{\partial \xi} - \lambda \tag{3.94}$$

which identifies the strength of the constrained force with the Lagrangian multiplier λ. Demanding that $\ddot{\xi} = \dot{p}^{\xi} = 0$ in a constrained simulation, 3.94 turns into:

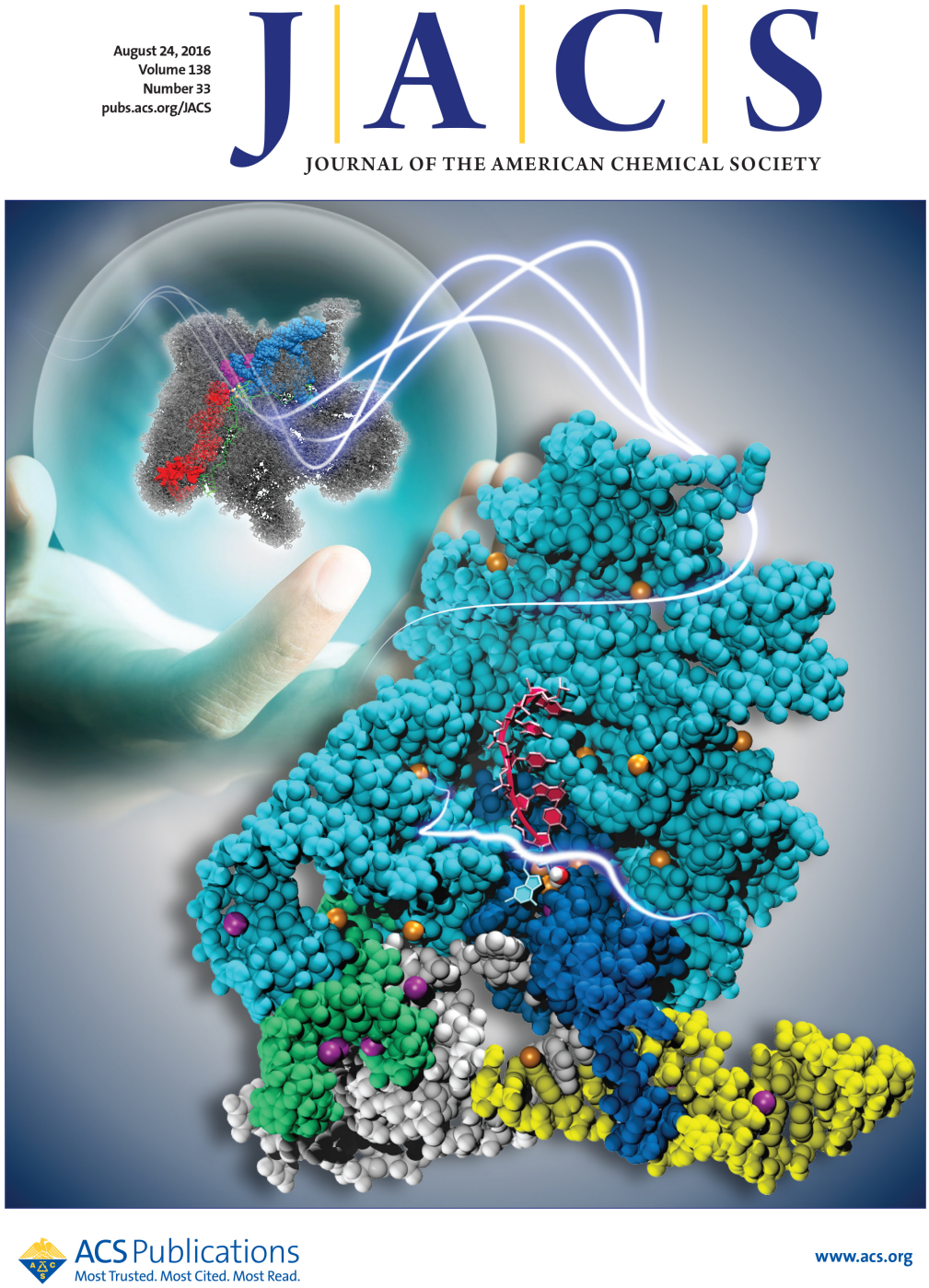
$$\frac{\partial \mathcal{H}}{\partial \xi} = -\lambda \tag{3.95}$$

from where we can calculate the free energy difference using the equation 3.90. The choice of a proper order parameter which reasonably approximates the reaction coordinate is a crucial point in the blue-moon ensemble technique.

4 Molecular mechanism of splicing in group II introns



Reference paper: Lorenzo Casalino, Giulia Palermo, Ursula Röthlisberger, and Alessandra Magistrato, A. *Who Activates the Nucleophile in Ribozyme Catalysis? An Answer from the Splicing Mechanism of Group II Introns,* J. Am. Chem. Soc. **2016**, *138*, 10374–10377.



Group II Introns, A Glimpse into the Spliceosome. Group II introns (close-up) are Mg^{2+}-dependent ribozymes capable of self-splicing and considered to share a common ancestor with the eukaryotic spliceosome machinery (in the crystal ball). By using combined quantum-classical molecular dynamics simulations their reaction mechanism is unprecedentedly unveiled, providing a glimpse into the splicing mechanism in higher eukaryotes.

This study on group II introns splicing was on the cover of the *Journal of the American Chemical Society*, August 24, 2016, Volume 138, Issue 33.

4.1 Abstract

Group II introns are Mg^{2+}-dependent ribozymes that are considered to be the evolutionary ancestors of the eukaryotic spliceosome, thus representing an ideal model system to understand the mechanism of conversion of premature messenger RNA (mRNA) into mature mRNA. Neither in splicing nor for self-cleaving ribozymes has the role of the two Mg^{2+} ions been established, and even the way the nucleophile is activated is still controversial. Here we employed hybrid quantum–classical QM(Car–Parrinello)/MM molecular dynamics simulations in combination with thermodynamic integration to characterize the molecular mechanism of the first and rate-determining step of the splicing process (i.e., the cleavage of the 5'-exon) catalyzed by group II intron ribozymes. Remarkably, our results show a new RNA-specific dissociative mechanism in which the bulk water accepts the nucleophile's proton during its attack on the scissile phosphate. The process occurs in a single step with no Mg^{2+} ion activating the nucleophile, at odds with nucleases enzymes. We suggest that the novel reaction path elucidated here might be an evolutionary ancestor of the more efficient two-metal-ion mechanism found in enzymes.

4.2 Introduction

Group II intron ribozymes (G2IRs) are self-splicing RNAs mostly found in bacteria and lower eukaryotes. They play an essential role for gene expression, being in charge of converting pre-mature messenger RNA (mRNA) into mature mRNA [43]. These RNA macromolecules can autocatalyze their excision from an RNA strand via two distinct transesterification events (i.e., self-splicing), yielding ligated exons and the intron in a lariat/linear form, or can also undergo a reverse splicing reaction into RNA/DNA filaments, contributing to gene diversification. Remarkably, they are believed to share common evolutionary origins with the eukaryotic spliceosome [32, 39, 44]. Indeed, it has recently been demonstrated that the RNA moiety of the spliceosome is exclusively in charge of catalysis, with the active site and a two-Mg^{2+}-dependent mechanism similar to G2IRs [32, 44]. Pre-mRNA splicing is a key biological process *per se*, turning into critical considering that its aberrations in humans are responsible for 13 % of genetic diseases and other complex pathologies (cancer and neurodegeneration) [39]. Hence, deciphering the mechanism of splicing at the atomistic level is of utmost importance as it may result in revolutionary gene modulation tools and novel therapeutic approaches [39]. A breakthrough in the

chemistry of G2IRs splicing was provided by a series of crystal structures, capturing a group IIC intron at sequential stages of the catalytic process [25, 30]. These revealed an active site containing a four-metal ion cluster made of two Mg^{2+} and two K^{+} ions, the former being catalytically active, the latter likely playing a structural role (Figure 4.1).

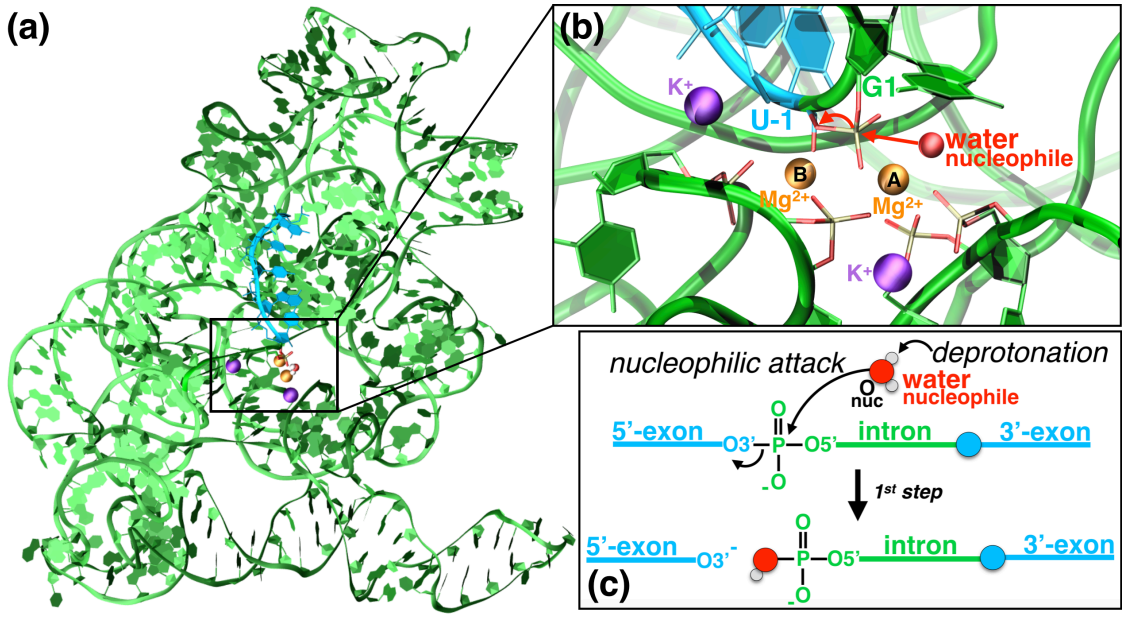


Figure 4.1. (a) Architecture of the *O. Iheyensis* group IIC intron before splicing (pdb: 4FAQ [25]); the intron and the exon are depicted in green and blue ribbons, respectively. (b) Inset of the catalytic site with Mg^{2+}, K^{+} and the nucleophilic water shown as orange, violet and red spheres, respectively. The phosphate atoms of the active site are depicted in sticks. (c) Scheme of the first hydrolysis step investigated here.

Mg^{2+}-dependent phosphodiester bond hydrolysis is believed to occur via a two-Mg^{2+}-ion mechanism, proposed by Steitz and Steitz on the basis of polymerases crystal structures and assumed to be representative also for G2IRs [21, 154, 155]. In their proposal the two-Mg^{2+}-ions act concertedly as Lewis' acids by properly orienting and activating the nucleophile, and by stabilizing the leaving group and the phosphorane transition state/intermediate. This role, confirmed in several computational studies of nucleases [11, 18, 20, 156], has not been assessed for two-Mg^{2+}-dependent ribozymes. Moreover, for self-cleaving ribozymes the function and the identity of general acid/base contributions are controversial [157-160], mostly due to the lack of RNA residues with unperturbed pK_as around 7 [161]. Thus, in computational studies aimed at characterizing their reaction mechanism, a pre-activated nucleophile is used [157, 158], or the role of general base is either assigned to hydroxyl ions located in the

vicinity of the nucleophile [159] or to nucleobases, sometimes, in rare protonation states [160].

In this work, we focus on the mechanism of the first and rate-determining step of the splicing reaction of G2IRs [25, 30], in which the 5'-exon cleavage is mediated by a nucleophilic water molecule. By using classical and quantum-classical (QM(Car-Parrinello)/MM) molecular dynamics (MD) [144, 145], with the QM part treated at DFT-BLYP [125, 126] level of theory, in combination with thermodynamic integration, we reveal a novel dissociative two-Mg^{2+}-ion mechanism in which the bulk water acts as general base.

4.3 Methods

Structural models. All the calculations have been based on the X-ray structure in which the group IIC intron ribozyme from *O. Iheyensis* was trapped at its precatalytic state in the presence of catalytic inhibitory Ca^{2+} ions (pdb code: 4FAQ [25]). The putative nucleophile is a water molecule captured in proximity (3.17 Å) of the scissile phosphate (P@G1) and coordinated by the divalent cation in position A. In the model, Ca^{2+} ions have been replaced with the catalytically active Mg^{2+} ions and the remaining negative charge has been neutralized by means of Na^{+} counterions (leading to 24 Mg^{2+}, 20 K^{+} and 327 Na^{+}). In order to reproduce the experimental concentration of divalent cations (Mg^{2+}) and monovalent cations (Na^{+}, K^{+}) at which the ribozyme is operative, namely 10 mM and 150 mM [25], respectively, the system was solvated with a 18 Å layer of TIP3P water molecules [162], reaching a total of ~330 000 atoms (including ~105 000 waters) and a periodic box size of 176 · 132 · 161 $Å^3$. The validation of the model is described in the Appendix A1.

Classical molecular dynamics simulations. Classical molecular dynamics (MD) simulations were carried out on the initial crystal structure to relax its geometry and obtain a reactive adduct suitable to undergo QM/MM MD calculations. 50 ns of MD simulations were run using the ff12SB (ff99+bsc0+χOL3) force field [94, 104, 163, 164], and the Åqvist [165] parameters for Mg^{2+} ions. However, we have checked that the stability of the system and the integrity and structural features of the active sites were retained for additional 150 ns (200 ns in total). Moreover, we have tested four additional Mg^{2+} ions parametrizations [166-169] to further check the consistency of our findings, performing overall ~1 μs of MD simulations for the reactant state. The calculations were done with the Amber12 PMEMD [170] program using a timestep of 2 fs. The temperature control (300K) was performed by Langevin dynamics [109]. The pressure control (1 atm) was accomplished by Berendsen barostat [113]. During the

heating phase (~1.1 ns), the catalytic core was subjected to positional restraints to prevent a distortion of the reactive residues. Subsequently, in the productive phase, all the restraints were released.

Two additional models with a nucleobase in a rare protonation state were also constructed: one in which A287 was deprotonated in N6 and the other in which G1 was deprotonated in N1. These models underwent ~50 ns of classical MD simulations without producing a configuration in which the nucleobase could have triggered the deprotonation of the nucleophilic water. Moreover, a fourth model was built up by replacing the nucleophilic water with a hydroxyl ion (OH^-). The RESP charges of the modified nucleobases and the OH^- were obtained by performing a minimization with Gaussian09 [171] followed by a fitting on the generated ESP with antechamber module of ambertools12 [170].

QM/MM MD. After the initial model (i.e, the one produced from the crystal structure with the water as nucleophile) underwent ~50 ns of classical MD simulations, the resulting structure was relaxed by means of ~10 ps of QM(Car-Parrinello)/MM MD simulations, using the fully Hamiltonian coupling scheme [144], which is part of the CPMD code [172]. The QM region comprised the phosphate backbone of the nucleobases U-1, G1, C358, G359, U375, C377, and 5 water molecules (90 atoms), while the rest was treated at MM level with the same force field mentioned above. The number of water molecules in the QM region has been extended up to 9 after the equilibration, in order to check the proton delivery path going from the nucleophile to the bulk water. Valence of the terminal atoms has been capped with hydrogen atoms. The interactions between valence electrons and ionic cores were described with norm-conserving Martins-Troullier pseudopotentials [137, 173]. For Mg^{2+} ions a semicore pseudopotential was employed. QM/MM MD were carried out with a time step of 0.12 fs and a fictitious electronic mass of 600 au; constant temperature simulations were achieved by coupling the system with a Nosé-Hoover thermostat [110, 111, 174] at 500 cm^{-1} frequency. The QM region was contained in an orthorhombic box of 22.6 · 20.5 · 25.3 $Å^3$. We treated the QM region at DFT level of theory using the BLYP [125, 126] and B3LYP [126, 127] exchange correlation functionals. A plane wave basis set has been used with a cutoff of 70 Ry. For the QM region periodic images were decoupled with the Martyna-Tuckerman scheme [175]. Dynamical-RESP [145] charges were obtained according to ref. [145] and were averaged over the equilibrated part of the trajectories.

Free energy calculations. The free energy profile was obtained by performing blue moon ensemble [149] calculations using as reaction coordinate (RC) the difference between the distance of the breaking bond ($O^{3'}$@U-1–P@G1) and the distance of the forming bond (O_{nuc}–P@G1). Starting from a value of the RC = -1.5 Å

we sampled in 20 sequential windows along the RC, with a resolution of 0.2 Å (0.1 Å in proximity of the transition state). A growth rate of 0.1 Å/ps has been used to move from a window to the following one and then each point has been simulated for ~2.5-5.0 ps for a total QM/MM MD simulation time of ~80 ps. Considering the forward and backward reactions we performed ~100 ps of QM/MM MD, further complemented with additional runs with B3LYP [126, 127] functional. The average constraint force within each window was taken from the points in which the force reached the convergence along the equilibrated trajectory. The statistical error on each point of the free energy profiles (forward-BLYP [125, 126], backward-BLYP [125, 126] and forward-B3LYP [126, 127] pathways) have been calculated by error propagation analysis. The overall error on the free energy barrier has been estimated as the sum of the statistical error and the error due to hysteresis between the forward and backward pathway. In order to check if hysteresis was present we selected the point at RC = 0.9 Å and then we followed the backward pathway, sampling in five consecutive points RC = 0.7 Å, 0.6 Å, 0.5 Å, 0.4 Å, 0.3 Å. We then released the constraint on the RC to assess that transition state was crossed in the backward pathway with the system falling into the reactant state. For a selected region of the free energy profile the B3LYP [126, 127] exchange correlation functional was also used, performing constrained QM/MM MD simulations within the points corresponding to RC = 0.7 Å, 0.9 Å, 1.1 Å for ~0.5-1 ps. Due to the extreme computational cost of these calculations we have considered only the RC points corresponding to the deprotonation event (i.e., 0.9-1.1 Å) and a point within the transition state window (i.e, 0.7 Å). These simulations were meant to investigate if an extra energy barrier due to the deprotonation event was present and for this reason we selected only the window of the RC in which the deprotonation event occurred.

As a further check of the free energy barrier we performed a geometry optimization and vibrational frequency analysis on reduced model systems of the catalytic site (93 QM atoms) for both reactant and transition state. These simulations were carried out with Gaussian09 program using the 6-31g* basis set and PCM [176] as implicit solvent model. Both for the reactant and transition state models the calculations were performed at DFT-BLYP [125, 126] and DFT-B3LYP [126, 127] level of theory in order to verify the influence of the exchange correlation functionals on the free energy barrier.

4.4 Results and discussion

Our simulations are based on the Ca^{2+}-inhibited structure upon replacement of Ca^{2+} with Mg^{2+} ions. The reliability of the initial adduct has been assessed before studying the reaction (see Appendix A1, Figures A1.1-1.3). The classical (maintained for ~200 ns) and QM/MM (~10 ps) MD equilibrated reactant shows the putative catalytic water coordinated to Mg^{2+}-A (MgA) [177] and a structural arrangement that is perfectly consistent with catalysis (i.e. the nucleophilic oxygen is in line with the scissile phosphate). However, there is no stable H-bond network heading from the nucleophile to a putative general base that could activate it [11, 18, 20, 156], re-confirming the non-trivial assignment of the general base in ribozyme catalysis. We remark that, although the X-ray structure of this G2IR lacks domain VI, which might include a general base, this reduced ribozyme construct is fully active for the hydrolytic pathway investigated here [25, 30]. To address alternative viable mechanisms, we have considered two additional model systems bearing nucleobases in rare protonation states at H-bond distance from the nucleophile: a G1 deprotonated at N1 and A287 deprotonated at N6. However, ~50 ns of classical MD simulations for each model did not lead to a H-bond network potentially capable of promoting catalysis (Figure A1.4a, b). Instead, the bases of the D5 bulge (A376-C377) and those for which a pK_a close to 7 has been reported (here A364-U384) [161], were not within H-bond distance from the nucleophile. These may be implicated in the conformational switch occurring after the first step [25]. Finally, replacing the nucleophilic water with a hydroxyl ion resulted in a complete distortion of the catalytic site (Figure A1.4c), suggesting that a reaction path with a pre-activated nucleophile can be ruled out either via canonical or alternative channels [178].

We then performed QM/MM blue moon ensemble simulations (~100 ps) on the model with the nucleophilic water and nucleobases in standard protonation states, choosing as reaction coordinate (RC) the difference between the breaking bond ($O^{3'}$@U-1–P@G1; d1) and the forming bond distances (O_{nuc}–P@G1; d2), with O_{nuc} being the oxygen of the crystallized nucleophilic water (Fig. 4.2a). These simulations unveiled the following mechanism.

(i) from RC = -0.5 to -0.1 Å a large increase of the distances between O^{Rp}@G1 and Mg^{2+}-B (MgB) and between O_{nuc} and MgA occurs (d6 and d4, respectively, in Fig.s 4.2b, 4.3). This corresponds to the transfer of the O^{Rp} of the scissile phosphate to MgA and to the simultaneous detachment of the nucleophile from MgA, respectively. Thus, a first difference from the protein two-Mg^{2+}-ion mechanism emerges: MgA does not act as Lewis' acid in activating the nucleophilic water.

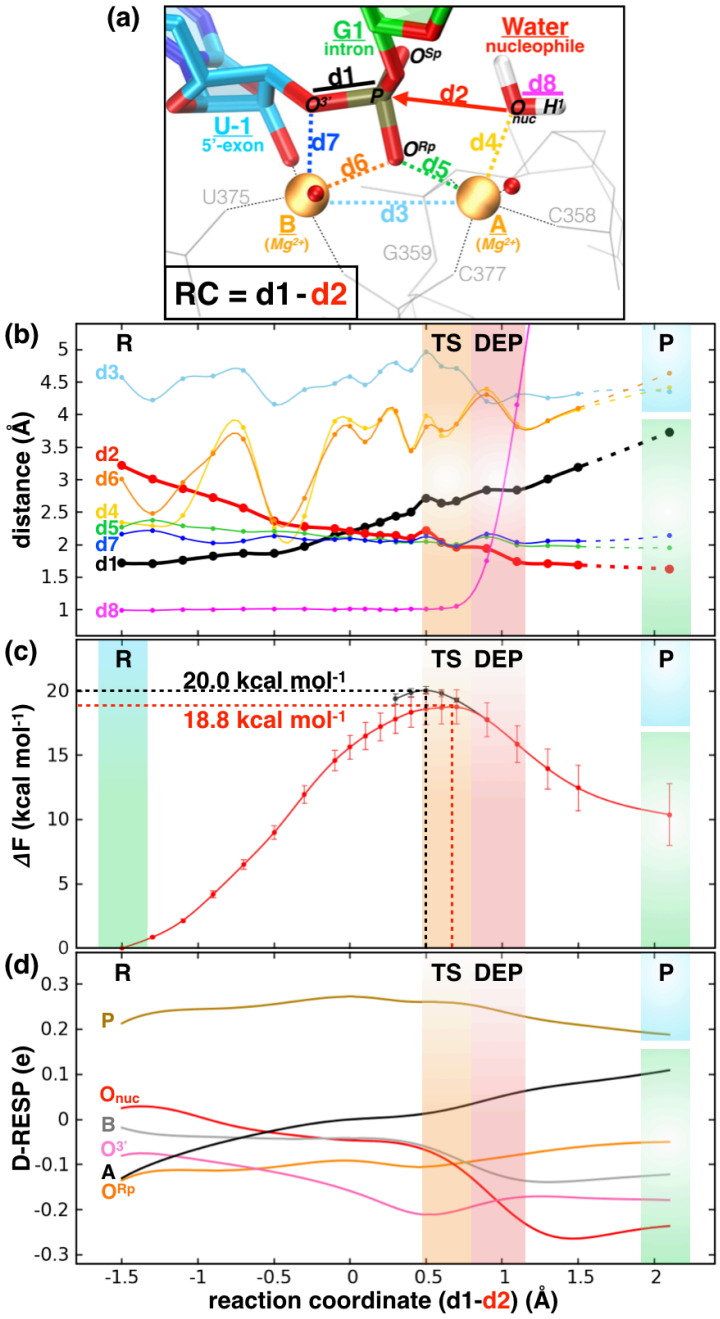


Figure 4.2 (a) Snapshot of the reactant state showing the reaction coordinate (RC) and the bond distances monitored along the RC in (b); (c) free energy profiles (ΔF, kcal/mol): red and black lines refer to ΔF calculated using BLYP [125, 126] for the direct and reverse process, respectively; (d) D-RESP (Dynamical-RESP) [145] charges monitored along the RC. The RC portions corresponding to reactant state (R), transition state (TS), water deprotonation (DEP) and product (P) formation regions are highlighted with colored areas.

(ii) At RC = 0.0 Å, before the transition state (TS) is reached, the dissociation of the leaving group ($O^{3'}$-@U-1) from the scissile phosphate (P@G1) takes place (Fig. 4.2b), in line with a dissociative mechanism. Here, MgB facilitates the cleavage by

stabilizing the oxyanion leaving group while MgA stabilizes the metaphosphate TS. Although never observed in nucleases [11], the dissociative cleavage of the phosphodi(mono)ester bond has been reported to occur in other enzymes, like alkaline phosphatases [179, 180], sometimes promoted by a non-Mg^{2+}-coordinated water as in actin or EcoRV [181, 182].

(iii) At the TS (RC = 0.5-0.7 Å) the bond between O_{nuc} and P@G1 starts forming (Fig.s 4.2b and 4.3). Differently to what is typically found in RNase enzymes [20],[156], at the TS the water is still protonated.

(iv) At RC = 0.9-1.5 Å, after O^{Rp}@G1 moved towards MgA and the proton of the nucleophile (H^1) is released to the bulk water (d8 in Fig. 4.2b), the O_{nuc}–P@G1 bond completely forms. By changing the QM/MM partitioning in favor of a more hydrated catalytic site, we observed a proton transfer to the bulk water involving up to 5 water molecules.

(v) At RC = 2.1 Å the formation of the product occurs with structural features resembling those of the corresponding crystal structure (pdb: 4FAR [25], Fig. 4.3). Note that the leaving group ($O^{3'-}$@U-1 of the 5'-exon) retains a negative charge (stabilized by coordination to MgB) to act as nucleophile in the second step of splicing.

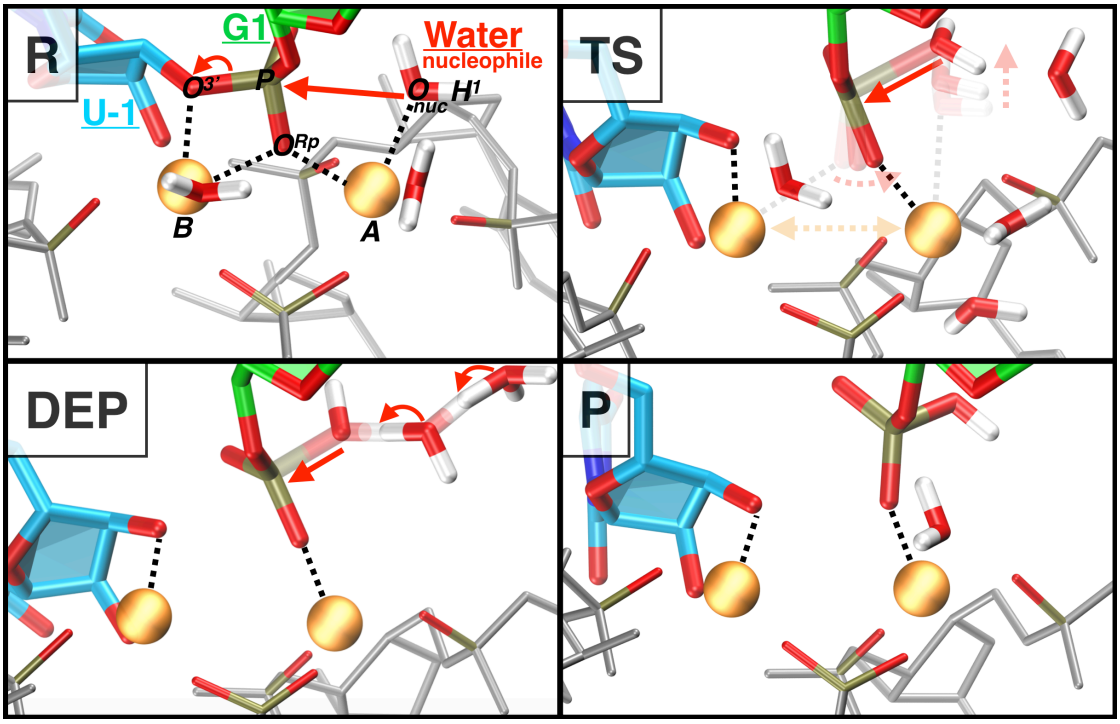


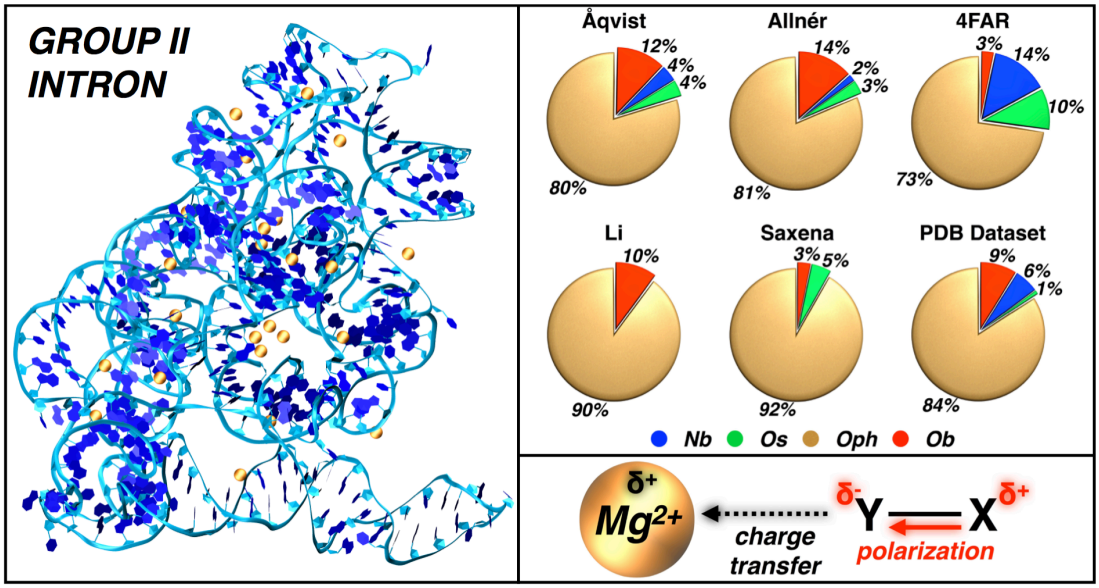
Figure 4.3. Snapshots of the reactant (R), transition state (TS), deprotonation event (DEP) and the product (P). Reactive atoms are depicted in sticks with intron and exon carbon atoms colored in green and blue, respectively; Mg^{2+} ions are shown in orange, the rest of the catalytic site is represented in light sticks.

This process occurs with a Helmholtz's free energy ($F^{\#}$) barrier of 18.8 ± 1.5 kcal/mol (Figure 4.2c; Appendix A1 and Figure A1.5), in line with the experimental catalytic rate of 0.011 min^{-1} [25], corresponding to a $\Delta G^{\#}$ of ~23 kcal/mol. We remark that the measured rate constant includes also the slow Mg^{2+}-dependent folding of G2IR, thus providing an upper limit to the splicing kinetic. Moreover, the calculated barrier is just slightly higher than the one calculated for RNase H with the same computational approach[20]. Additional control calculations performed on small model systems of the catalytic site for Reactant (R) and Transition State (TS), using the B3LYP [126, 127] and BLYP [125, 126] functionals and a localized basis set (see Appendix A1 for further details), led to a $\Delta G^{\#}$ of ~25 and ~19 kcal/mol, respectively, suggesting that BLYP slightly underestimates the barrier. To further check the trend of the free energy profile within the water deprotonation window (DEP), we have also performed constrained QM/MM MD simulations with B3LYP for selected RC points. This confirmed that no additional barrier is associated to the proton transfer event, with the TS lying at the same RC value (Appendix A1, Fig. A1.5). To establish the functional role of the observed structural rearrangements, we have plotted the Dynamical-RESP [145] (D-RESP) charges along the RC (Fig. 4.2d). These were dynamically calculated during our QM/MM MD simulations, allowing us to monitor the changes in the chemical state of the system along the reaction. This analysis reveals an increase of the P@G1 and a decrease of $O^{3'}$@U-1 charges just before the TS is reached, in line with a dissociative mechanism. At the TS, the O_{nuc}–P@G1 bond (d2) starts forming and, consequently, the charge on P@G1 decreases again, while O^{Rp}@G1 maintains an increasing trend, as it progressively moves towards MgA and detaches from MgB. Finally, after the TS, the nucleophilic water becomes a hydroxyl group by releasing its proton (H^{1}) to the bulk and completely forming the O_{nuc}–P@G1 bond. This dissociative mechanism, clearly points to a distinct catalytic process from the canonical two-Mg^{2+}-ion mechanism observed in enzymes [11, 18, 20, 156]. Here, MgA activates the electrophile rather than the nucleophile, strongly contributing also to the stabilization of the metaphosphate at the TS, while the water is converted in an OH^{-} group only after the formation of the O_{nuc}–P@G1 bond has started. In this manner, the proton release to the bulk occurs with no barrier and without the need of a specific base. To the best of our knowledge this is the first mechanistic study suggesting the bulk water as possible proton acceptor, offering a proposal for the controversial assignment of the general base in ribozymes.

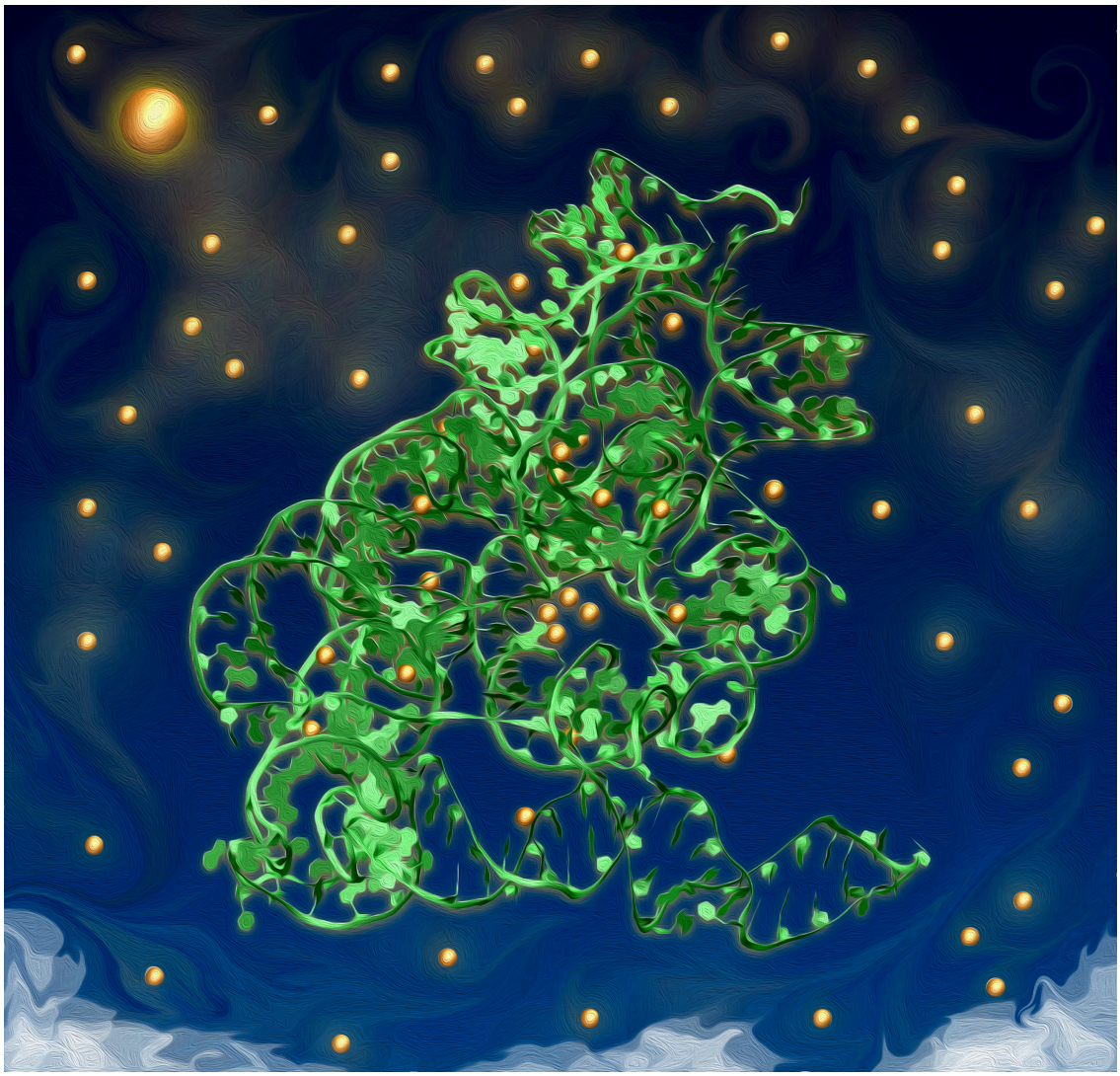
4.5 Conclusions

In summary, our study unveils a dissociative mechanism in which the role of Mg^{2+} ions remarkably differs from that reported for most enzymes. We believe that this novel mechanism is specific for the ribozyme scaffold. In fact, in enzymes the active site is shaped to host a specific substrate and each residue has a functional role to enhance catalysis. Conversely, in ribozymes the catalytic site is formed by the RNA sugar–phosphate backbone, whose specificity for the reaction is lower than that of amino acids. We have shown here that ribozymes, considered the evolutionary ancestors of enzymes, can perform phosphodiester hydrolysis almost as effectively as the latter by adapting the mechanism to their less specific RNA scaffold. This mechanism opens a new scenario concerning the identity of general acid/base in ribozymes, which will stimulate further experimental and computational studies.

5 Mg^{2+}/RNA interplay: a computational perspective



Reference paper: Lorenzo Casalino, Giulia Palermo, Nodira Abdurakhmonova, Ursula Röthlisberger, and Alessandra Magistrato. *Development of Site-specific Mg^{2+}–RNA Force Field Parameters: A Dream or Reality? Guidelines from Combined Molecular Dynamics and Quantum Mechanics Simulations*, J. Chem. Theory Comput. **2017**, *13(1)*, 340–352.



The Starry Night of RNA. As the stars light up the sky at night, Mg^{2+} ions highlight RNA functions by enhancing catalysis, tuning its folding and conferring stability to the tertiary structure. Using combined classical molecular dynamics and quantum mechanics simulations we have characterized the performance of five different Mg^{2+} force fields models in two prototypical ribozymes (group II introns (close-up) and hepatitis delta virus), also disclosing how charge transfer and polarization affect the binding of Mg^{2+} ions (the stars) to RNA.

5.1 Abstract

The vital contribution of Mg^{2+} ions to RNA biology is challenging to dissect at the experimental level. This calls for the integrative support of atomistic simulations, which at the classical level are plagued by limited accuracy. Indeed, force fields intrinsically neglect nontrivial electronic effects that Mg^{2+} exerts on its surrounding ligands in varying RNA coordination environments. Here, we present a combined computational study based on classical molecular dynamics (MD) and Density Functional Theory (DFT) calculations, aimed at characterizing (i) the performance of five Mg^{2+} force field (FF) models in RNA systems and (ii) how charge transfer and polarization affect the binding of Mg^{2+} ions in different coordination motifs. As a result, a total of ~2.5 μs MD simulations (100/200 ns for each run) for two prototypical Mg^{2+}-dependent ribozymes showed remarkable differences in terms of populations of *inner-sphere* coordination site types. Most importantly, complementary DFT calculations unveiled that differences in charge transfer and polarization among recurrent Mg^{2+}–RNA coordination motifs are surprisingly small. In particular, the charge of the Mg^{2+} ions substantially remains constant through different coordination sites, suggesting that the common philosophy of developing site-specific Mg^{2+} ion parameters is not in line with the physical origin of the Mg^{2+}–RNA MD simulations inaccuracies. Overall, this study constitutes a guideline for an adept use of current Mg^{2+} models and provides novel insights for the rational development of next-generation Mg^{2+} FFs to be employed for atomistic simulations of RNA.

5.2 Introduction

Mg^{2+} is the most abundant alkali-earth metal in the biosphere with a high concentration in living cells [13]. Besides being operative in ATPases, ATP synthases and Mg^{2+}-dependent enzymes processing nucleic acids [5, 11, 18], this ion is ubiquitously present in RNA, playing a key function in both tuning RNA folding and catalysis. Due to their high charge density, Mg^{2+} ions afford the unique capability of effectively screening the negative charge of the RNA phosphate backbone, contributing to shape its folding landscape and conferring stability to RNA tertiary structures [183]. Their crucial role is emphasized by the fact that RNA filaments can reach their native folded conformation only in the presence of Mg^{2+} ions [184].

Moreover, Mg^{2+} ions can enhance catalysis in ribozymes by stabilizing active sites and by properly orienting reactants and/or polarizing reactive groups [29, 155, 185].

Mg^{2+} ions accomplish this wide range of functions by binding to RNA in different manners. While diffuse Mg^{2+} ions interact with RNA via long-range electrostatic interactions mediated by several layers of water molecules, specific Mg^{2+}-RNA binding sites can form either when Mg^{2+} ions come in direct contact with RNA atoms (hereafter referred to as "*inner-sphere"* coordination sites), or when Mg^{2+}-RNA interactions are mediated by one shell of water molecules (hereafter referred to as "*outer-sphere"* coordination sites) [186]. In both cases, when embedded in a RNA filament, Mg^{2+} ions directly/indirectly interact with several different combinations of RNA atoms (i.e., *coordination patterns,* CPs*)*, which can constitute specific and recurrent structural architectures (i.e., *binding motifs*). These are specific structural arrangements provided by RNA for Mg^{2+} binding that are found in multiple RNA molecules [187, 188].

Recently, a comprehensive study considering all RNA crystal structures deposited in the protein data bank (PDB) classified the Mg^{2+}-RNA CPs and binding motifs, assaying also the statistical binding preferences of this metal toward different RNA ligands [187]. This analysis pointed out 41 and 95 distinct types for *"inner-sphere"* and *"outer-sphere"* CPs, respectively, among which 13 recurrent binding motifs were observed [187].

In spite of the critical role of Mg^{2+} ions for RNA functions, the experimental characterization of Mg^{2+}-RNA binding sites is currently limited for several reasons: (i) Mg^{2+} ions are difficult to detect via X-ray crystallography given their anomalous X-ray scattering and their isoelectronicity with water molecules and Na^{+} ions [187]; (ii) these ions are silent to most spectroscopic techniques, such complicating the dissection of their binding site composition on the basis of the ligands spectroscopic signature; (iii) catalytic RNA is most often crystallized in the presence of metals inhibiting catalysis, thus biasing the identification of catalytic Mg^{2+} ions in nucleic acids [25]; (iv) a detailed understanding of the dynamical interplay between Mg^{2+} and RNA requires an atomistic resolution, which may be difficult to access in most wet-lab experiments [189].

In this respect, molecular simulations can constitute a valuable support in the characterization of Mg^{2+}-RNA binding sites. Force field (FF)-based molecular dynamics (MD) simulations are largely employed for the study of biological macromolecules, including systems containing Mg^{2+} ions [11, 18, 29, 155, 190, 191]. In current FFs, these ions are typically represented by fixed-point charges, bearing a doubly positive charge that electrostatically interacts with RNA macromolecules. However, Mg^{2+} can also exert a significant amount of non-trivial interactions such as

polarization and charge transfer (CT) effects [183, 192, 193], which are difficult to capture with simplified empirical FFs. Although some recently developed polarizable FFs may partially overcome these issues, their application has been so far mostly limited to model Mg^{2+}-water interactions [194-196]. Instead, an *ad-hoc* bonded parametrization of Mg^{2+}, in which the bonds between Mg^{2+} and its ligands are explicitly defined, constitutes an impractical solution. Indeed, at variance with transition metal ions, Mg^{2+} ions are ubiquitous in RNA, bind in a wide range of CPs [13] and do not form coordination bonds with their d orbitals [188]. An alternative model is represented by the multisite ion approach (hereafter referred to as *"cationic dummy atom"* (CDA)), in which the Mg^{2+} ion is replaced by a central atom covalently bound to six dummy sites placed in the direction of coordinating atoms (i.e., at the vertexes of an octahedron) and parameterized to reproduce both the geometrical and energetic features of Mg^{2+}. The 2+ charge is differently (i.e., depending on the CDA type) distributed among the central atom and the dummies [197-199]. Finally, another recent solution has been proposed on the basis of a modified "12–6–4" van der Waals (vdW) potential for divalent metal ions [200, 201].

An accurate parameter-free description of Mg^{2+} binding to RNA can be achieved via mixed quantum mechanics/molecular mechanics (QM/MM) methods, which allow taking explicitly into account the electronic structure of a small QM part, while the rest of the system is described at a classical level [202]. These methods have been extensively employed to study ribozyme catalysis [28, 29, 155, 159, 203, 204]. However, QM/MM studies are restricted in the size of the QM region and in the accessible time scales (*pico*-seconds). This hampers the characterization of many Mg^{2+} binding sites (due to the large size limit), as well as of their influence in the structural and folding properties (due to the time-scale limit). Evidently, FF-based MD still represents the most useful tool to gain insights into the long-time scale conformational and compositional changes of the Mg^{2+} coordination sites, as well as on the associated RNA folding properties. As such, the development of reliable FFs for Mg^{2+} ions is urgently needed.

Here we report a comparative study relying on extensive (a total of ~2.5 μs, 100/200 ns for each run) classical MD simulations, based on the AMBER-ff12SB FF [163, 164], in combination with five different non-polarizable Mg^{2+} FFs in order to benchmark their relative performances. These simulations are carried out on two prototypical ribozymes, namely the group IIC intron (G2IR) [25] (Figure 5.1) and the hepatitis delta virus (HDV) [205], at two different Mg^{2+} concentrations. Density Functional Theory (DFT) calculations on small models of representative Mg^{2+} coordination sites (both of *inner-* and *outer-sphere*) have been further performed, providing a systematic characterization of electronic effects occurring at *inner* and

outer-sphere Mg^{2+} coordination sites. These calculations unveil for the first time the general principles driving the binding of Mg^{2+} ions to varying RNA motifs and pin down possible sources of errors in the current FFs. This information can help for an adept use of currently available Mg^{2+} FFs and in the rational development of next-generation Mg^{2+} FF models [201, 206].

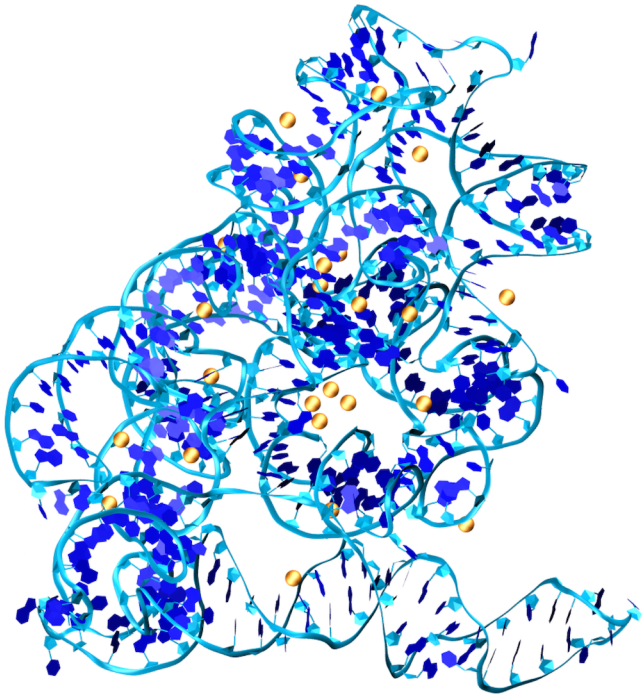


Figure 5.1. Group IIC intron (G2IR) ribozyme from *Oceanobacillus Iheyensis* (PDB entry 4FAQ [25], including 24 Mg^{2+} binding sites. The ribozyme is depicted using blue ribbons, while Mg^{2+} ions are represented as orange spheres.

5.3 Methods

Model systems. MD simulations of G2IR from *Oceanobacillus Iheyensis* were done starting from the X-ray structure of the reactive adduct (PDB code 4FAQ, solved at 3.11 Å resolution) [25]. We built two model systems of G2IR with different Mg^{2+} concentrations, namely [Mg^{2+}] = 10 mM and 25 mM. The former is the experimental concentration at which the self-catalyzed splicing reaction occurs [29], while the latter was selected to estimate the effect of Mg^{2+} concentration on its coordination properties. Both the simulations were carried out at 150 mM for monovalent ions. The first system was solvated with a 18 Å layer of TIP3P water molecules [162], reaching a total of ~330 000 atoms and containing 24 Mg^{2+}, 20 K^+ (originally present in the X-ray structure) and 327 Na^+ ions. The second system was solvated with a 14 Å layer of

TIP3P waters [162], thus leading to ~270 000 atoms, with 48 Mg^{2+}, 20 K^+ and 279 Na^+ ions. The X-ray structure of the reactive adduct was solved in the presence of the catalytically inactive Ca^{2+} ions that in our simulations were substituted with the biologically active Mg^{2+} ions. We decided to focus on the reactive adduct, such allowing the characterization of the structural and electronic features of a two-Mg^{2+}-ion catalytic site, which is rare in RNA, but extremely important for ribozymes catalysis. To assess the reliability of our model, we verified that the occupancy and the distribution of the divalent cation sites in the reactant state were retained in the X-ray structure of the first-step-splicing product (PDB entry 4FAR) [25], which instead was solved in the presence of Mg^{2+} ions (Figure A2.1 in the Appendix A2).

MD simulations of the HDV ribozyme were based on the PDB entry 1VC7 (2.45 Å resolution) [205], in which the crystallized Sr^{2+} ions were replaced by Mg^{2+} ions. In conformity with G2IR, we built two models considering [Mg^{2+}] = 10 mM and 25 mM. The first system (10 mM of Mg^{2+}, including 6 Mg^{2+} and 61 Na^+) was solvated with an 18 Å layer of TIP3P waters, resulting in a total of ~83 000 atoms. The second system (25 mM of Mg^{2+}, including 20 Mg^{2+} and 33 Na^+) was solvated with a 24 Å layer of TIP3P waters, corresponding to a total of ~110 000 atoms. The data harvested out of MD simulations on HDV were used as complement of the simulations of G2IR and, as such, they are mostly reported in the Appendix A2.

2.2. Classical MD Simulations. MD was used to equilibrate the systems at physiological conditions. The AMBER-ff12SB (ff99+bsc0+χOL3) [163, 164] was adopted for the RNA. In order to compare different Mg^{2+} ions parametrizations, we considered non-bonded fixed point charge and CDA models. Among the former, we selected the parametrization due to Åqvist [207], Allnér et al. [208], and Li et al. [209]. Among the CDA models, we have considered the ones from Oelschlaeger et al. [198] and Saxena et al. [197]. Monovalent ion parameters were taken from Joung et al. [210]. MD simulations were carried out using an integration time step of 2 fs, keeping all bonds with hydrogen atoms fixed with SHAKE [211]. Temperature control (300 K) was performed by Langevin dynamics [109], with a collision frequency $\gamma = 1$. Pressure control was accomplished by coupling the system to a Berendsen barostat [113] at a reference pressure of 1 atm and with a relaxation time of 2 ps. All the simulations were carried out with the following protocol. First, the systems were subjected to energy minimization to relax the water molecules and the Na^+ counter ions, keeping the RNA, Mg^{2+} and K^+ ions fixed with harmonic position restraints of 300 kcal/mol · $Å^2$. Then, the systems were heated up from 0 to 100 K in the canonical ensemble (NVT), by running two NVT simulations of 5 ps each, imposing position restraints of 300 kcal/mol · $Å^2$ on the key elements of the catalytic site (Mg^{2+}, K^+, RNA ligands and waters coordinating the two Mg^{2+} ions) and of 100 kcal/mol · $Å^2$ on the remaining Mg^{2+}

and K^+ ions. The temperature was further increased up to 200 K in 100 ps of MD in the isothermal-isobaric ensemble (NPT), in which the restraint on the catalytic site was reduced to 25 kcal/mol · $Å^2$. Subsequently, all the restraints were released and the temperature of the system was ultimately raised up to 300 K in a single NPT simulation of 1 ns. After ~1.1 ns of equilibration, ~10 ns of NPT production was carried out allowing the density of the system to stabilize around 1.01 g/cm^{-3}. Finally, production runs were carried out in the NVT ensemble, collecting ~100/200 ns, depending on the system. A list of all the simulations performed as well as their length is provided in Table A2.1 in the Appendix A2. Overall, a total of ~2.5 μs of classical MD simulations has been done by using the Amber12 [170] code in its GPU CUDA accelerated PMEMD version. Although several parametrizations have been recently proposed to overcome some of the identified problems [164, 212-214], it is well known that RNA FFs experience problems for long MD simulations. Hence, in order to increase the statistics of the Mg^{2+} binding sites, avoiding the incurrence of long-time RNA FF instabilities, we decided to perform several simulations of limited length (~100/200 ns) with different starting conditions (assigning different random velocities). We remark that the time scale of Mg^{2+} association/dissociation to/from RNA is of the order of millisecond [189], which is currently not accessible to standard MD simulations. As such, we would not have been able to directly observe such events even by extending the lengths of the MD runs.

Mg^{2+} force fields. Among the non-bonded fixed point charge, we selected the parametrization due to Åqvist [207], which has been originally developed to reproduce the solvation free energy and radial distribution function *(g(r))* of aqueous Mg^{2+} ions, the recent model developed by Allnér et al. [208], in which Mg^{2+} parameters have been tuned to reproduce the experimental first shell water/phosphate exchange rate and the parameters developed by Li et al. [209], which are specific for Particle Mesh Ewald-based simulations. In particular, for the latter we chose the parameter set better accounting for Mg^{2+} coordination number and hydration free energy (i.e., the CN set for TIP3P water) [209]. Among the cationic dummy atom (CDA) models, we have considered the one from Oelschlaeger et al. [198] and Saxena et al. [197]. The former assigns a charge -1 to the central atom, while the surrounding dummy atoms, disposed as octahedron, have a fractional positive charge accounting for +3. The Saxena model distributes the +2 charge entirely among the dummies, which are given only a repulsive "A coefficient" for the vdW potential. FF parameters for fixed point charge and CDA models are reported in Table 5.1.

In the Amber FF, vdW interactions are calculated according to the simplified Lennard-Jones potential:

$$V_{L-J} = \frac{A_{ij}}{r^{12}} - \frac{B_{ij}}{r^{6}} \tag{5.1}$$

Where $A_{ij} = \varepsilon_{ij}(R_{min,ij})^{12}$ is the repulsive term and $B_{ij} = 2\varepsilon_{ij}(R_{min,ij})^{6}$ is the attractive term. ε_{ij} and $R_{min,ij}$ correspond to the energy minimum (i.e., the depth of the potential well) and the internuclear separation between species i and j at the energy minimum, respectively. ε_{ij} and $R_{min,ij}$ are calculated by means of Lorentz-Berthelot mixing rules in which $\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}$ and $R_{\min,ij} = \left(R_{min,ii} + R_{min,jj}\right)/2$. The self-interaction values for $A_{Mg\text{-}Mg}$ and $B_{Mg\text{-}Mg}$ for each parametrization, as calculated in the Amber FF, are reported in the Table 5.1 ($A_{d\text{-}d}$ and $B_{d\text{-}d}$ for dummy atom species), together with the assigned point charge for Mg^{2+} ion/central atom (i.e., q_{Mg}) and, where necessary, for the dummy atoms (i.e., q_d).

NAME	$\boldsymbol{A_{Mg\text{-}Mg}}$ (Å^{12} · kcal/mol)	$\boldsymbol{B_{Mg\text{-}Mg}}$ (Å^{6} · kcal/mol)	$\boldsymbol{A_{d\text{-}d}}$ (Å^{12} · kcal/mol)	$\boldsymbol{B_{d\text{-}d}}$ (Å^{6} · kcal/mol)	$\boldsymbol{q_{Mg}}$ *(e)*	$\boldsymbol{q_d}$ *(e)*
Åqvist	225.26	28.39	-	-	+2	-
Allnér	2405.86	5.32	-	-	+2	-
Li	1673.12	8.26	-	-	+2	-
Saxena	729.00	278.89	0.0025	0.000	0	+0.333
Oelschlaeger	4901.75	1681.29	2.126	0.326	-1	+0.5

Table 5.1. Force field (FF) vdW parameters (*A* and *B*, Å^{12} · kcal/mol and Å^{6} · kcal/mol) and charges (q, e) for fixed point charge (i.e., Åqvist, Allnér and Li) and CDA (i.e., Saxena and Oelschlaeger) models employed in this work.

Analysis of structural data. Analysis of the Root Mean Square Deviation (RMSD), Root Mean Square Fluctuation (RMSF), Radius of gyration (R_g) and Radial Distribution Function ($g(r)$) have been performed with the cpptraj module of Amber12 [170]. The coordination number (CN) analysis of all Mg^{2+} ions was performed with Plumed 2.0 [215] based on the switching function reported in Equation 5.2:

$$s = \frac{1 - \left(\frac{r-d_0}{r_0}\right)^{n}}{1 - \left(\frac{r-d_0}{r_0}\right)^{m}} \tag{5.2}$$

where r_0 is the cutoff distance of the first coordination shell, corresponding to the first peak of the calculated $g(r)$ of Mg^{2+} *versus* oxygen and nitrogen coordinating atoms, while d_0 is assigned to zero; n and m are set to 50 and 100, respectively. In this analysis, we have divided the interacting RNA atoms in 4 groups (Figure 5.2): the

RNA phosphate OP1 and OP2 oxygen atoms (labeled as O_{ph}), the O2', O3' O4', O5' atoms of the ribose sugar (labeled as O_s), the oxygen and the nitrogen atoms of the bases (labeled as O_b and N_b, respectively). Oxygen atoms of water molecules are labeled as O_w. In order to make our analysis independent from the reference structure, we have also calculated the normalized interaction frequency of Mg^{2+}-ligand contacts, *F(X)*, as defined in Zheng et al. [187] and reported in Equation 5.3:

$$F(X) = \frac{p(Mg—X)}{p(X)}. \tag{5.3}$$

Here, *p(Mg–X)* is the fraction of a given coordination type, namely the number of the Mg^{2+}–X interactions (where X is one atom type among O_{ph}, N_b, O_b, O_s and O_w) with respect to the total number of Mg^{2+} interactions with all ligands; *p(X)* represents the fraction of atoms, i.e. the number of atoms of X type with respect to the total number of atoms in the data set (i.e., the system in analysis). Hence, *F(X)* measures the frequency of a particular Mg^{2+}–ligand interaction normalized by the frequency of the considered atom type in the overall structure.

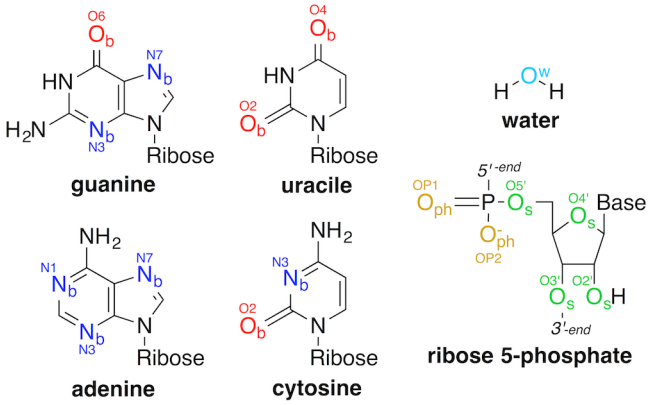


Figure 5.2. Donor atoms of RNA and water interacting with Mg^{2+} ions. Nitrogens (N_b) and oxygens (O_b) from nucleobases are shown in blue and red, respectively. Ribose oxygens (O_s) are shown in green, while phosphate oxygens (O_{ph}) in gold. Water oxygens (O_w) are depicted in light blue.

Density Functional Theory (DFT) calculations. To perform a systematic analysis of the charge transfer and polarization effects taking place between Mg^{2+} and its first and second shell ligands, DFT calculations have been done on a set of small model systems accounting for most of the Mg^{2+} binding motifs reported by Zheng et al. [187]. In particular, from our MD simulations we identified and extracted 16 models including from 19 to 72 atoms, in which the number of RNA ligands in the *inner-* and in the *outer-sphere* range from zero to four, with the remaining ligands being water

molecules. We remark that for *inner-sphere* coordination sites a "ligand" refers to an RNA moiety or water molecule directly interacting with Mg^{2+} and the "donor atom" is the atom actually involved in the contact, while for *outer-sphere* coordination sites a "ligand" is intended as a functional group of a nucleotide (i.e. phosphate or base) interacting with a hexa-hydrated Mg^{2+} ion.

All the models were subjected to geometry optimizations and vibrational frequencies analysis at room temperature using the Gaussian09 code [171]. For the *inner-sphere* models, we initially imposed a positional constraint on terminal carbon atoms of RNA ligands. However, in some cases the RNA–Mg^{2+} coordination distances of the models extracted from MD simulations were too short/long being affected by the FFs inaccuracies (Table A2.2), potentially biasing the analysis of the electronic effects. Thus, we treated all the *inner-sphere* sites without any constraint. Instead, for the *outer-sphere* models, since the RNA ligands are not directly interacting with Mg^{2+} ions, we run a first optimization cycle imposing a positional constraint on terminal carbons followed by an additional unconstrained optimization (starting from the obtained minima). DFT calculations were performed with the M06 [128] and B3LYP [126, 127] exchange correlation functionals and using the 6-311++G** basis set. The Polarizable Continuum Model (PCM) [176] was used with water as a solvent (dielectric constant, $\varepsilon = 78.355$). For the sake of completeness, for the *inner-sphere* systems we have also investigated the effect of a dielectric constant of 4 (i.e., representing the RNA environment) [216]. For the geometry optimization the convergence criterion of the RMS force was set to $1*10^{-5}$ [217]. Natural Bond Orbital (NBO) charges [218] have been calculated using the NBO 6.0 program [219]. For comparison, Bader charge analysis were also performed using the program developed by Tang et al. [220]. To check the dependence of NBO charges on the basis set employed we also run single point calculations with the 3-21G basis set on the geometries optimized with the 6-311++G** basis set.

Charge transfer (CT) to Mg^{2+} ion from first/second shell ligands was quantified on the basis of the effective (NBO/Bader) charge of the Mg^{2+} ion with respect to its formal 2+ charge. For the *outer-sphere* sites the CT occurring between the second and the first shell of coordination is determined as the difference between the formal charge of the second shell ligands and their NBO/Bader charge when interacting with the hexa-hydrated Mg^{2+}. In our schematic model, when Mg^{2+} ion coordinates its ligands it exerts a polarization effect on them, triggering a general ligand charge rearrangement (LCR) that eventually includes also the charge transfer towards the metal. In this scenario, it is difficult to clearly dissect the polarization from CT effects. Thus, hereafter, we will discuss the polarization by estimating the LCR (Δq). This is calculated by taking the difference between the NBO/Bader charge of the isolated ligands and the charge they

assume when bound to Mg^{2+}. This operation is done considering the coordinating and non-coordinating atoms separately. In this manner Δq provides a qualitative and simplified picture of the polarization occurring throughout the ligands.

For each model, we have also calculated its free energy of formation ΔG_{form}. For the *inner-sphere* coordination sites, $\Delta G_{form\text{-}is}$ is calculated as the RNA ligand/water exchange free energy accordingly to Equation 5.4:

$$\Delta G_{form-is} = \left[G_{Mg-motif-is} + nG_{wat}\right] - \left[G_{RNA-is} + G_{Mg-wat}\right] \quad (5.4)$$

where $G_{Mg\text{-}motif\text{-}is}$ is the total free energy of the specific Mg^{2+} *inner-sphere* coordination site, $G_{Mg\text{-}wat}$ is the free energy of Mg^{2+} hydrated by six explicit water molecules, $G_{RNA\text{-}is}$ is the free energy of the same RNA motif without the Mg^{2+} ions bound, while nG_{wat} is the free energy of the n water molecules that are released during the formation of the Mg^{2+} coordination site. Considering, for example, the chemical equality (Equation 5.5) for the formation of $[RNA_nMg(H_2O)_{6-n}]^{q+2}$ coordination site:

$$[Mg(H_2O)_6]^{2+} + nRNA^q \rightarrow [RNA_nMg(H_2O)_{6-n}]^{q+2} + nH_2O \quad (5.5)$$

where $G_{Mg\text{-}motif\text{-}is}$ is the free energy of the $[RNA_nMg(H_2O)_{6-n}]^{q+2}$ *inner-sphere* coordination site, $G_{Mg\text{-}wat}$ is the free energy of the $[Mg(H_2O)_6]^{2+}$ complex, $G_{RNA\text{-}is}$ is the free energy of $nRNA^q$, and nG_{wat} is the energy of nH_2O. In Eq. 4, q is the charge of RNA ligands.

For *outer-sphere* coordination sites, $\Delta G_{form\text{-}os}$ is equal to (Equation 5.6):

$$\Delta G_{form-os} = \left[G_{Mg-motif-os}\right] - \left[G_{RNA-os} + G_{Mg-wat}\right] \quad (5.6)$$

where $G_{Mg\text{-}motif\text{-}os}$ is the total free energy of the specific Mg^{2+} *outer-sphere* coordination site, $G_{Mg\text{-}wat}$ is the free energy of Mg^{2+} hydrated by six explicit water molecules, $G_{RNA\text{-}os}$ is the free energy of the same RNA motif without the Mg^{2+} ions bound. This equation is consistent with the following chemical equality (Equation 5.7):

$$[Mg(H_2O)_6]^{2+} + nRNA^q \rightarrow [RNA_nMg(H_2O)_6]^{q+2}. \quad (5.7)$$

Corrections for basis set superposition errors (BSSE) and zero point energies have been applied and we have also considered the entropic contribution (translational, rotational and vibrational) to the free energy of formation.

5.4 Results and discussion

5.4.1 Classical force field models

Mg^{2+}-coordination sites reproduced by FFs models. Classical MD simulations have been initially performed on G2IR considering five different Mg^{2+} FF parametrizations (i.e., Åqvist, Allnér, Li, Saxena and Oelschlaeger). Besides its biological relevance as model system of the eukaryotic spliceosome, we selected G2IR as prototypical system to study Mg^{2+}-RNA interactions because: (i) it is among the largest RNA macromolecules of known structure [184]; (ii) it presents a large number of Mg^{2+} binding sites (i.e., 24), including a catalytic bimetal site rarely resolved in ribozymes [155, 184]. The transferability of the statistical analysis obtained for G2IR was tested by performing MD simulations also on HDV [205], a well characterized Mg^{2+}-dependent ribozyme. During the production runs, the structural stability of the two ribozymes has been evaluated by calculating the RMSD, the RMSF and the R_g. Except for few cases, G2IR (Figure A2.2) and HDV (Figure A2.11) remain stable throughout all simulations, showing that different Mg^{2+} parametrizations and ionic strengths do not affect their overall structural stability.

Analysis of $g_{Mg\text{-}X}(r)$ (Figure A2.3, Table A2.2) clearly shows that the distances between Mg^{2+} and its ligands are underestimated in all cases with respect to the corresponding DFT values obtained in this study, with the only exception of the Mg^{2+}–N_b distance, which is instead overestimated. For the Oelschlaeger parametrization, a different $g(r)$ is observed, respect to all other models. This corresponds to unrealistically large distances between Mg^{2+} and its ligands (see Appendix A2 for additional details).

Examination of the coordination numbers (CN) at both [Mg^{2+}] = 10 and 25 mM discloses that all the models reproduce the typical octahedral coordination sphere (i.e., CN = 6, data not shown), which is assumed by Mg^{2+} in water solution and constitutes its dominant coordination mode when bound to biological systems. An exception is again represented by the Oelschlaeger model, which often leads to an expanded coordination spheres (i.e., CN = 7/8). However, this model has been parametrized to reproduce the Mg^{2+} coordination sphere in protein enzymes (i.e., DNA polymerase β) [155, 198]. For this reason, this DCA was not further considered here. For all the employed Mg^{2+} FFs, the CN stabilizes within the first ~70 ns of MD simulation, highlighting the convergence of our MD simulations with respect to Mg^{2+} coordination properties. However, a peculiar oscillatory behavior of the CN was overall observed for the Mg^{2+}–N_b contacts.

The impact of the FF parametrization on the Mg^{2+} capability to account for multiple Mg^{2+}-RNA binding modes was investigated by performing a statistical analysis of the ligands composition of the Mg^{2+} *inner-sphere* binding sites, obtained from the MD simulations of G2IR at [Mg^{2+}] = 10 mM (Figure 5.3). As a result, the Åqvist, Allnér and Li FFs most frequently reproduce two CPs: the most populated CP is constituted by two O_{ph} donor atoms and four water ligands ($2O_{ph}$:$4O_w$), while the second most populated one is composed by three O_{ph} donor atoms and three water ligands ($3O_{ph}$:$3O_w$). These CPs are also well represented by the Saxena model, although their relative population is inverted.

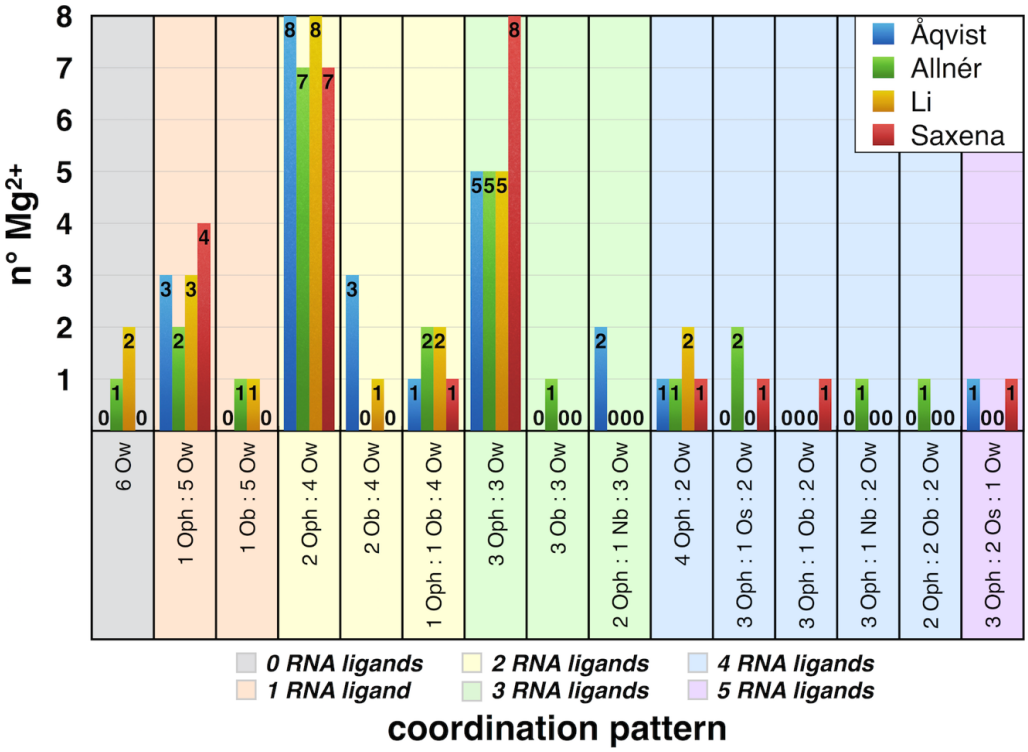


Figure 5.3. Histograms showing the population of each Mg^{2+}–RNA CP observed during MD of G2IR performed with the Åqvist, Allnér, Li and Saxena parametrizations at [Mg^{2+}] = 10 mM. Bars of different colors are used to identify the FF model, as specified in the top right corner legend. The *x-axis* reports the CPs identified from MD simulations. The number of RNA ligands (from 0 to 5) is highlighted with different colors, as indicated in the bottom legend. The *y-axis* reports the number of Mg^{2+} ions (i.e., population) having a specific CP.

Remarkably, the Allnér model shows the widest range of possible CPs, although most of them with low population. Interestingly, in all simulations the most recurrent CPs are characterized by the presence of only O_{ph} donor atoms as non-water ligands, while the CPs with at least one O_b or one O_s donor atom are less frequent and the CPs with one N_b are rare.

A comparison with the ligand composition of Mg^{2+} sites in the X-ray structure of G2IR in the first-step-splicing product state (PDB entry 4FAR) [25] (Figure A2.4 vs Figure A2.6), shows that the crystallographic Mg^{2+} sites are only partially reproduced by all the MD simulations. Sites 1 and 2, belonging to the catalytic binuclear site, are well accounted only by the Åqvist and Saxena FFs. Notably, the crystallographic sites in which Mg^{2+} is bound to N_b (i.e., sites 4, 16, 17 and 21) are not reproduced by any parametrization. Unfortunately, the limited resolution of the 4FAR X-ray structure (i.e., 2.86 Å), which affects a complete characterization of the Mg^{2+} coordination spheres (Figure A2.4), hampers a well-defined structural match to assess the reliability of the Mg^{2+} FFs employed. However, an interesting comparison can be done between the MD CPs populations (Figure 5.3) and the histogram of the CPs identified in the whole PDB dataset analysis [187] (Figure A2.5). This unveils that only five CPs (i.e., $1O_{ph}$:$5O_w$, $2O_{ph}$:$4O_w$, $1O_{ph}$:$1O_b$:$4O_w$, $3O_{ph}$:$3O_w$, $4O_{ph}$:$2O_w$) among those identified by Zheng et al. [187], are reproduced by the Mg^{2+} FFs here employed. Remarkably, the composition of the coordination sphere of the 24 Mg^{2+} sites is conserved only for two sites (i.e., site 3 and 11) among all the Mg^{2+} parametrizations (Figure A2.6).

Finally, to establish if the Mg^{2+} models herein considered can reproduce the ligands distribution observed in RNA crystal structures, we performed an extensive statistical analysis comparing the percentages of the Mg^{2+} *inner-sphere* donor atoms, as obtained from our MD simulations (Figure 5.4a), with the data extracted by Zheng et al. [187] and with the crystal structure (4FAR) [25] (Figure 5.4b). From this analysis emerged that: (i) for all the Mg^{2+} FFs the majority of the donor atoms is represented by O_{ph}, in agreement with Zheng et al. [187]; (ii) the Åqvist, Allnér and Li models overestimate the number of interactions with O_b atoms; (iii) the Li and Saxena FFs do not account for the interactions with O_s (Li) and N_b (Li and Saxena). Interestingly, although in the crystal structure the number of N_b donor atoms (14 %) is higher than in the PDB dataset (6 %), this is markedly reduced in all the simulated systems and completely disappears in the Li and Saxena models, clearly pinpointing an underestimation of the Mg^{2+}–N_b contacts by all the employed FFs. Overall, the Åqvist and Allnér parameters similarly account for the Mg^{2+}–RNA ligands contacts, showing a distribution that is in better compliance with the PDB dataset [187]. To further check that these results were not biased by an arbitrary kinetic trapping of Mg^{2+} to the nearest close by minimum, we performed two additional MD simulations with different starting conditions for both Allnér and Åqvist models, confirming the trend discussed above (Figure A2.7).

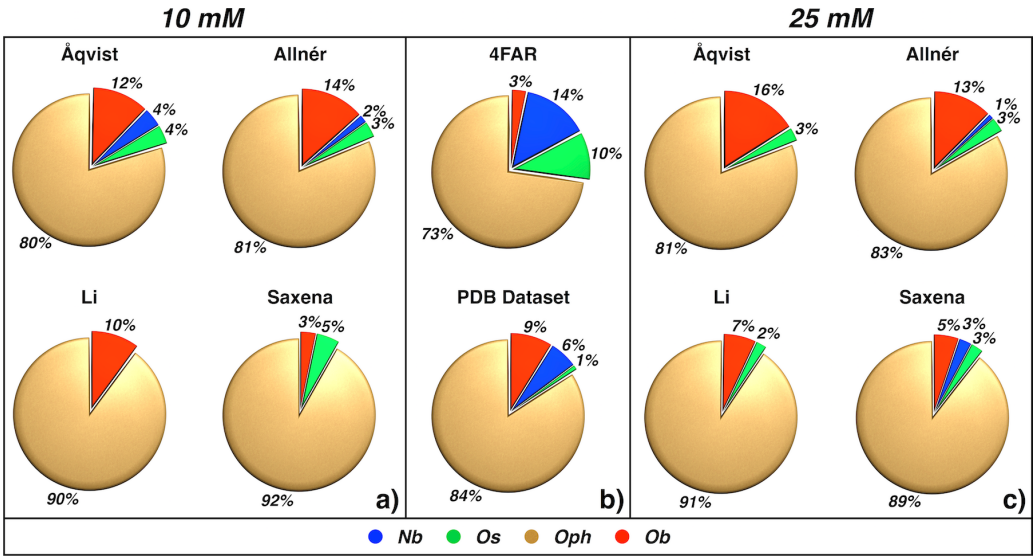


Figure 5.4. Statistical distribution of RNA donor atoms in the Mg^{2+} *inner-sphere* (expressed as percentages) as obtained from MD simulations of G2IR performed with the Åqvist, Allnér, Li and Saxena Mg^{2+} models at [Mg^{2+}] = 10 mM (a) and 25 mM (c). Data for the 4FAR X-ray structure of G2IR and for the entire PDB dataset relative to the *inner-sphere* sites are also reported (b) [25, 187].

By extending this analysis to the water ligands (O_w, Figure A2.8), which are not easily captured by X-ray crystallography, we found that water constitute the majority of the *inner-sphere* ligands with the Åqvist and Allnér parametrizations exhibiting an almost identical statistical distribution. Since these models appeared to be the most reliable in reproducing the relative abundance of Mg^{2+}–RNA ligand contacts present in the PDB dataset, we considered only them in the following tests.

In a similar study [221], the performance of the Åqvist, Allnér, Li FFs and two additional models based on a modified 12-6-4 LJ potential developed by Li et al. [200] and by Panteva et al. [201] have been tested on the RNA stem-loop, which exhibits a Mg^{2+} ion dependent conformational shift. All the FFs promoted the transition towards a Mg^{2+}-bound conformation. However, the formation of likely artificial chelated interactions in the 12-6-4 models interfered with folding of the RNA stem-loop. Interestingly, these 12-6-4 potentials showed the highest occupancy of the four key Mg^{2+}–RNA binding sites with a better description of Mg^{2+}–N_b interactions. Instead, consistently with our data, the point charge models exhibited a comparable occupancy centered on phosphate atoms throughout all the simulations, with a general underestimation of Mg^{2+}–N_b contacts.

Effect of Mg^{2+} ions concentration. Specific Mg^{2+} concentrations are important to modulate the activity of self-splicing ribozymes and to affect the stability of RNA molecules [25, 222], since many-body effects arise in the close presence of Mg^{2+} ions

[223]. However, given the difficulty to reproduce the experimental metal ions concentrations in MD simulations, the majority of the computational studies on biomolecules are usually done at high ionic strengths [11, 19]. Simulations performed at [Mg^{2+}] = 25 mM show the appearance of novel Mg^{2+} CPs, and the disappearance of some sites lowly populated at [Mg^{2+}] = 10 mM (Figure A2.9). In particular, novel low populated sites (i.e., not identified by Zheng et al. [187]) come out by using Allnér (5 sites) and Åqvist (2 sites) models (Figure A2.9). Additionally, the relative abundance of the most populated patterns is altered. The Allnér model favors the $1O_{ph}$:$5O_w$ CP, which is the most abundant also in X-ray studies [187], whereas the Åqvist model prefers the $2O_{ph}$:$4O_w$ pattern (Figure A2.9). Notably, the Mg^{2+} concentration can also affect the composition of the original 24 coordination sites with the Allnér model being the less sensitive to the change in concentration (Figure A2.10).

The RNA donor atoms distribution in the *inner-sphere* sites (Figure 5.4c) also shows remarkable differences at the higher Mg^{2+} ions concentration: (i) the N_b interacting atoms are no longer accounted by the Åqvist model, and are substituted by an increased percentage of O_b donor atoms; (ii) the Saxena and Li models recover N_b and O_s contacts, respectively, but their relative abundances are rather far from what can be expected from the PDB dataset; (iii) the statistical distribution of RNA ligands for the Allnér model does not change significantly, confirming its minor sensitivity to the ionic strength; (iv) by further including the water ligands in this analysis, some new *outer-sphere* sites appear and the number of Mg^{2+}–O_w contacts increases (Figures A2.8, A2.10).

MD simulations on HDV. To complement our statistical analysis, we performed MD simulations on the HDV ribozyme with the Åqvist and the Allnér FFs. At [Mg^{2+}] = 10 mM the number of sites was too low to be statistically relevant and the simulation has been discussed only as a complement of the result of G2IR (Figure A2.15). Instead, at [Mg^{+2}] = 25 mM the same relative abundance of the most populated CPs, which emerged for G2IR, is encountered: (i) for the Allnér model the most recurrent pattern is $1O_{ph}$:$5O_w$ followed by $2O_{ph}$:$4O_w$; (ii) for the Åqvist model is $2O_{ph}$:$4O_w$ followed by $1O_{ph}$:$5O_w$ (Figure A2.12). Consistently with what observed for G2IR, the composition of the binding sites is different with the two parametrizations employed, enlightening that our results are not system specific. Additionally, the simulations on the HDV ribozyme assess again the sensitivity of the *inner-sphere* coordination sites to varying Mg^{2+} concentrations (Figures A2.13 and A.214) and the general difficulty of the currently available force fields in accounting for Mg^{2+}–N_b contacts (Figures A2.13 and A2.14). Obviously, the same holds true combining the statistics of both RNA systems investigated here (Figure A2.15).

In order to make our analysis independent from the structures investigated, we have also calculated the normalized interaction frequency *F(X)* between Mg^{2+} and RNA ligands (Table 5.2), as defined by Zheng et al. [187] and detailed in the methods section. Table 5.2, which displays the results for both G2IR and HDV ribozymes, shows that contacts with O_{ph} and O_s are reasonably reproduced by both Åqvist and Allnér parametrizations, while binding to O_b donor atoms is overestimated (i.e., the Mg^{2+}–O6@G contacts are overestimated in both models and the Mg^{2+}–O2@C contacts in Allnér, irrespective of the Mg^{2+} concentration).

RNA donor atoms			**F(X)**				
			Åqvist 10 mM	**Allnér 10 mM**	**Åqvist 25 mM**	**Allnér 25 mM**	**PDB dataset**
O_{ph}		OP1	3.99	4.14	3.94	3.90	4.19
		OP2	4.60	4.75	5.02	5.22	4.99
O_s		O2’	0.31	0.31	0.31	0.26	0.07
		O3’	0.15	-	-	-	0.04
		O4’	-	-	-	-	0.004
		O5’	-	-	-	-	0.04
A	N_b	N1	-	-	-	-	0.04
		N3	-	-	-	-	0.003
		N7	0.58	0.58	-	-	0.74
G	O_b	O6	2.95	3.44	2.97	2.55	1.46
	N_b	N3	-	-	-	-	0.002
		N7	0.49	-	0.25	0.28	1.35
C	O_b	O2	0.70	0.70	0.71	0.81	0.14
	N_b	N3	-	-	-	-	0.01
U	O_b	O2	0.76	0.76	0.38	0.44	0.08
		O4	1.53	0.76	1.92	1.76	2.33

Table 5.2. Normalized interaction frequency *F(X)* [187] calculated for *inner-sphere* Mg^{2+}–RNA contacts for the whole PDB dataset[25] and obtained from our simulations performed with the Åqvist and Allnér parametrizations at $[Mg^{2+}]$ = 10 mM and 25 mM on G2IR and HDV. Reported data are comprehensive of the two studied systems and grouped according to the Mg^{2+} model and the concentration conditions. RNA donor atoms, including the phosphate oxygens (O_{ph}), the O3’, O5’ O2’, O5’ atoms of the sugar (O_s), the nitrogen (N_b) and oxygen (O_b) atoms of the bases are specified in the first column.

Coherently with this observation, we remark that the Allnér model accounts for four CPs, characterized by the presence of at least one coordination to O_b atoms, that are not observed in the PDB dataset [187] (Figure A2.9). In addition, in the simulations with the Åqvist FF, one unprecedented site involving $4O_b$ is noticed (Figure A2.9). These results may be related to lowly populated CPs that have not yet been identified in the

PDB dataset or, more likely, to inaccuracies in the Mg^{2+} or RNA vdW parameters, as already pointed out in other studies [214].

5.4.2 The two-Mg^{2+}-ion catalytic site

Group II intron ribozyme presents a peculiar catalytic site which enables the intron self-splicing reaction [29]. The importance of this site is associated to a two-Mg^{2+}-aided mechanism that is also exploited by the spliceosome and is therefore critical in regulating gene expression in humans. The peculiar G2IR active site, characterized by the presence of two Mg^{2+} at a ~4 Å distance [29] (Figure 5.5), is not reported among the CPs classified by Zheng et al. [187] because of its low statistical abundance, which is most likely associated to its very specific biological function [28, 187, 224].

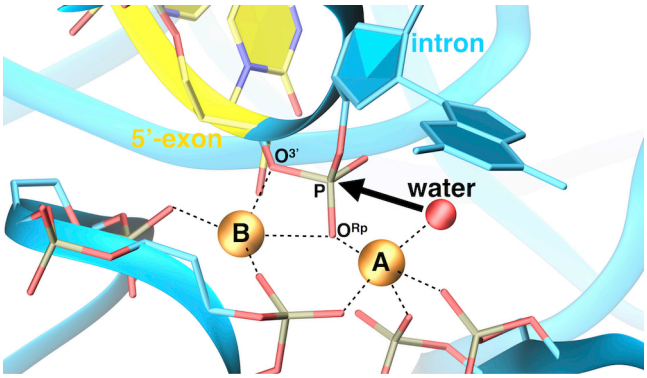


Figure 5.5. Active site from the crystal structure of the reactive adduct (PDB entry: 4FAQ) [25]. The oxygen atom of the nucleophilic water molecule and the Mg^{2+} ions are shown as red and orange spheres, respectively. Phosphates are shown in licorice and are colored according to element. An arrow indicates the direction of the nucleophilic attack on the scissile phosphate.

Our molecular simulations reveal that only the Åqvist and the Saxena FFs are capable of reproducing a catalytically competent conformation of the two-metal-aided active site (Figure 5.6), consistent with the X-ray structure. Surprisingly, by employing the Allnér parameters, the active site distorts after a few nanoseconds of MD (i.e., ~10 ns, Figure A2.16). The Allnér model has been tuned to reproduce the kinetics of Mg^{2+}-water/phosphate exchange reactions and presents a lower free energy barrier for water exchange than the Åqvist model [208]. This may determine the structural instability of the catalytic site, which is largely solvent exposed. Additionally, the vicinity of the two catalytic Mg^{2+} ions (~4 Å) leads to a strong electrostatic repulsion, which may favor Mg^{2+}-water/phosphate exchange and the structural distortion of the active site. A CDA model as the one due to Saxena succeeds in limiting these instabilities by spreading the

2+ charge among the dummy atoms in the direction of the coordinating donor atoms. At increased Mg^{2+} ions concentration (i.e., 25 mM) (Figure 5.6) one of the metal ions (i.e., Mg^{2+}-A) loses its original coordination geometry even with the Åqvist parameters, although no additional ions locate in the vicinity of the binding site, while the geometry is retained with the Saxena model. Overall, these data indicate that the choice of the classical parameters for Mg^{2+} ions in metal-dependent active sites can be influenced by the structural characteristics of the site itself and, importantly, by the experimental working conditions relative to Mg^{2+} concentration.

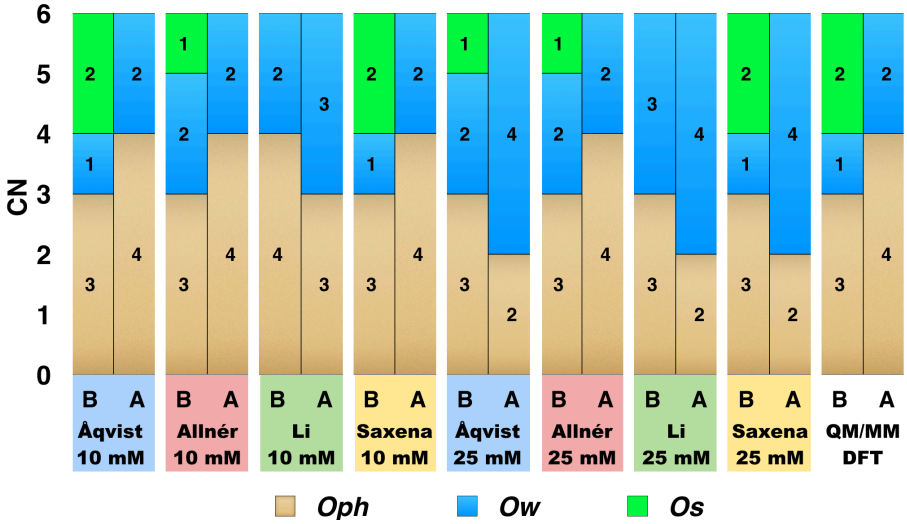


Figure 5.6. Histogram showing the composition of the two Mg^{2+} ions (A and B) of the catalytic site of G2IR ribozyme as reproduced the different Mg^{2+} force field models and by QM/MM MD simulations [197, 198, 207-209]. The QM/MM MD simulations were detailed in a previous study performed by us [29].

5.4.3 *Ab-initio* models

A systematic analysis of the electronic effects taking place between Mg^{2+} and its first and second shell ligands was done by performing DFT calculations on a set of cluster models representative of the recurrent Mg^{2+}–RNA binding architectures (Figure 5.7). In the following we have quantified the CT effect and the LCR as an estimate of the polarization effect, as detailed in the methods section.

Electronic signature of Mg^{2+}-RNA *inner-sphere* coordination sites. The DFT-NBO analysis of the *inner-sphere* models shown in Figure 5.7a reveals that no coordination bonds are formed between Mg^{2+} and first shell ligands, confirming that Mg^{2+} interactions with the surrounding ligands is based exclusively on electrostatics, CT and polarization effects.

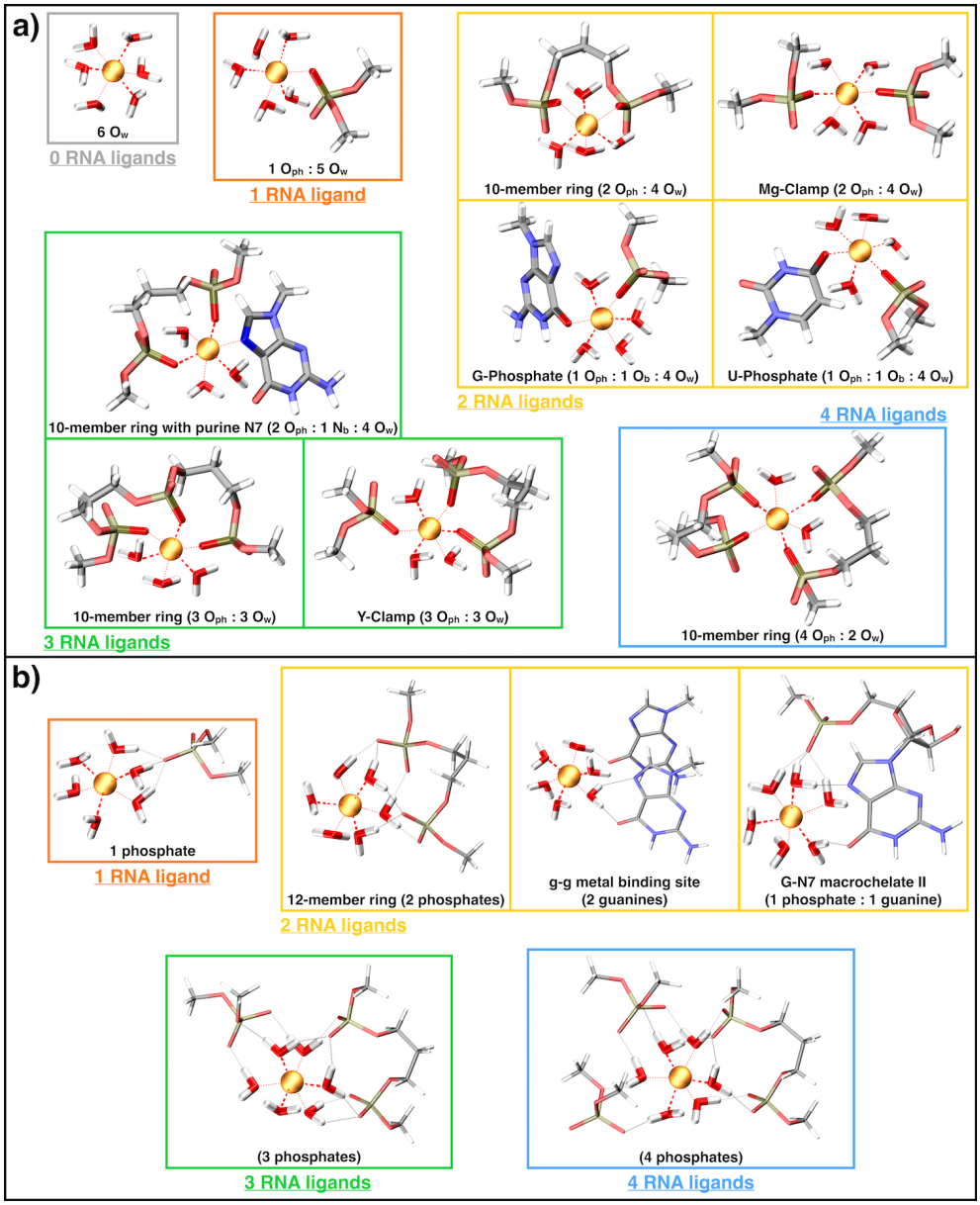


Figure 5.7. Model systems of Mg^{2+}-RNA binding architectures. 16 models have been selected, including both *inner-sphere* (a) and *outer-sphere* (b) coordination sites, in which the number of RNA ligands increases from 0 to 4. Boxes of different colors are used to identify the model systems characterized by 0 (gray), 1 (orange), 2 (yellow), 3 (green) and 4 (blue) RNA ligands. RNA ligands and water molecules are shown as sticks, while Mg^{2+} ions are shown as orange spheres. For each model, the motif name (if present) and the specific CP (in parenthesis) are reported.

In particular, when Mg^{2+} ion coordinates its six ligands, it induces a charge movement from the more distant atoms toward those directly coordinated to Mg^{2+}, resulting in their polarization. Simultaneously, part of the charge accumulated on the coordinating atoms is transferred to Mg^{2+} ion (CT) (Figure 5.8d). As a consequence of

this LCR, the more distant ligand atoms become more positive (i.e., display a positive Δq, Figure 5.8a), while the directly coordinating donor atoms become more negative (i.e., show a negative Δq, Figure 5.8b). Here, we have monitored in detail the amount of CT and of LCR for each selected coordination site.

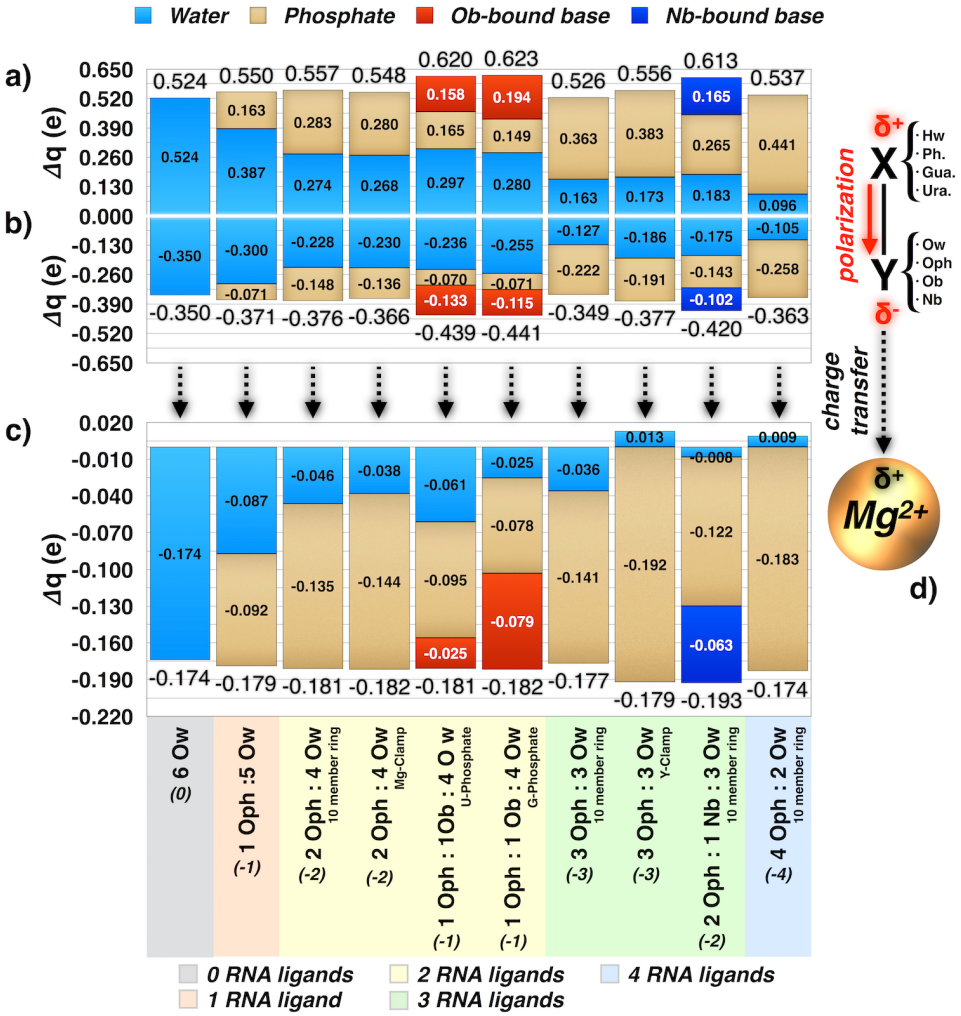


Figure 5.8. Charge rearrangements (Δq, e) of the (a) non-Mg^{2+}-coordinated and (b) Mg^{2+}-coordinated ligands atoms in the *inner-sphere* coordination sites; (c) amount of charge transferred (Δq, e) from the ligands towards Mg^{2+} ion calculated from the NBO charge distribution and the M06 functional with 6-311++G** basis set. Each contribution is dissected by atom type with light blue, gold, red and dark blue referring to water (O_w), phosphate (O_{ph}), and to nucleobases coordinated via O_b or N_b atoms, respectively. (d) Schematic picture of the polarization and charge transfer effects exerted by a Mg^{2+} ion. The formal charge of each CP is reported in parenthesis. Despite the charge distribution accuracy is significantly lower, we report here fractional charges with three decimal digits in order to have a complete balance among the charges received by Mg^{2+} ions and the ones distributed over the ligands.

The LCR of the non-coordinating atoms (Figure 5.8a) is remarkable and similar in all models, ranging from +0.52 e to +0.62 e. Obviously, the average variation of the partial charge of each atom is significantly smaller.

By looking at the contribution of each ligand type, it arises that the charge rearrangement of the water hydrogen atoms (H_w) conspicuously decreases when a non-water ligand (phosphate, base or both) is introduced in the coordination sphere (a decrease of 45% is registered for each H_w in the $4O_{ph}$:$2O_w$ 10-member ring [187] motif with respect to the $6O_w$ model). This enlightens the dominant role played by RNA ligands, with the water molecules largely adapting their charge distribution to each specific binding site. The adaptive behavior of water ligands stands out again as their strength as charge donors rapidly decreases to zero when more than two non-water ligands are present in the coordination site. Interestingly, in the Y-Clamp ($3O_{ph}$:$3O_w$) and 10-member ring ($4O_{ph}$:$2O_w$) motifs [187], the oxygens of the water molecules even accept from Mg^{2+} part of the charge donated by phosphate ligands, displaying a more negative Δq. Indeed, the coordinating atoms of the phosphates and the nucleobases (i.e., O_{ph}, O_b and N_b) of these models transfer a significant amount of charge to the Mg^{2+} ion (Figure 5.8b). As expected, the largest LCR is observed for the motifs containing aromatic nucleobases (i.e. guanine or uracil in the G-Phosphate, U-Phosphate and $2O_{ph}$:$1N_b$:$3O_w$ motifs) due to the high delocalization of the electrons over the aromatic ring.

The analysis of the CT (Figure 5.8c and 5.9a) shows a relevant decrease of the Mg^{2+} NBO charge when a hexa-coordinated Mg^{2+} complex is formed ($6O_w$ model). The Mg^{2+} ion is indeed capable of withdrawing -0.17 e from its six ligands, reaching a net charge of +1.83 (Figure 5.9a). Intriguingly, there are no differences in the amount of CT between the $6O_w$ and $4O_{ph}$:$2O_w$ models, and only very subtle variations exist in the other coordination sites. This clearly pinpoints that the amount of CT from the first shell ligands towards the Mg^{2+} ion does not depend on the number of non-water ligands introduced, but occurs in a limited and saturated manner (i.e., ~0.18 e transferred), consistently with a similar NBO study [188]. We remark that in a different study [28] on a two-Mg^{2+} ion system the withdrawn charge, calculated with the Mulliken scheme, was ~0.43 e. Figure 5.8c and 5.9a strikingly show that the effective charge of Mg^{2+} is essentially constant, while that of the ligands is strongly adapting to the coordination environment.

The O_{ph} donors are responsible for the largest amount of CT, prevailing over the O_w, O_b and N_b (with the size of the base playing a noteworthy role in the last two cases, i.e. the largest CT occurs for guanine with respect to uracil). The trend observed for the NBO charges is independent from the exchange correlation functional employed, i.e. M06 vs. B3LYP (Figures A2.17 and A2.18). Although a small dependence on the basis

set (i.e. 6-311++G** vs. 3-21G) is observed for the amount of CT and LCR (Figure A2.19), the general trend is maintained. For the sake of completeness, the charge distribution analysis was done also with the Bader partitioning scheme [220]. This shows a peculiar behavior for the hexa-hydrated Mg^{2+} ion motif ($6O_w$), in which, at variance with all the other models, only a small CT takes place (Figure A2.20). However, all NBO results discussed here and obtained with an extended basis set are consistent with those observed from the Bader analysis.

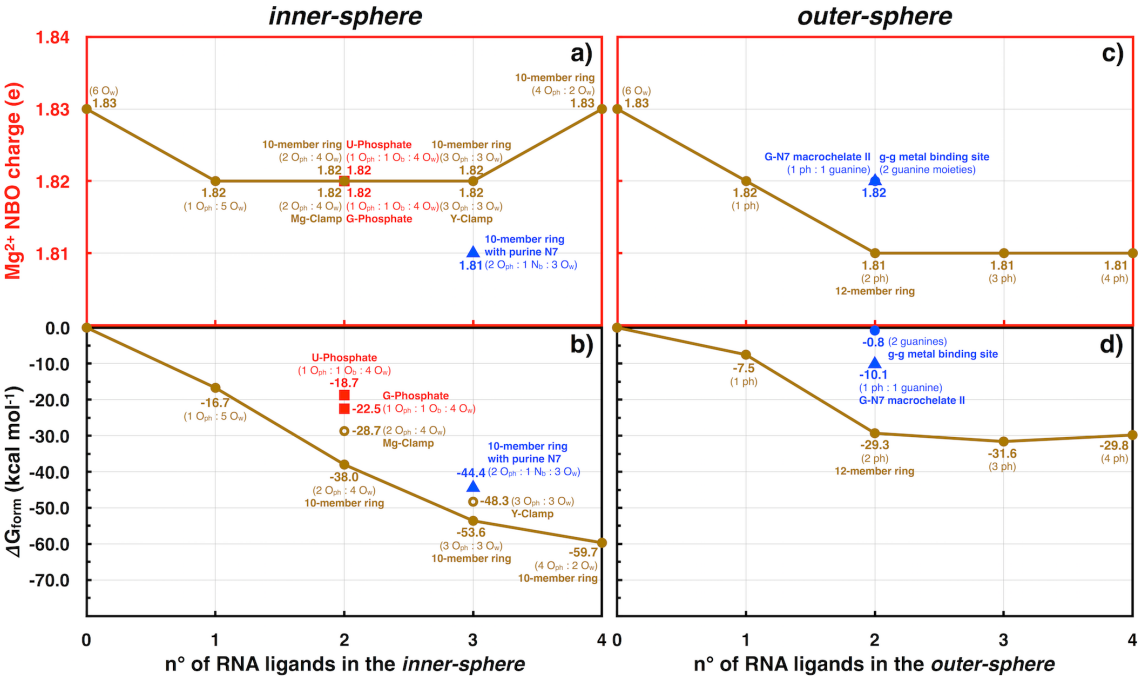


Figure 5.9. Mg^{2+} charge (e) and free energy of formation (ΔG_{form}, kcal/mol) of *inner-sphere* Mg^{2+} coordination sites, (a) and (b) respectively, and *outer-sphere* Mg^{2+} coordination sites, (c) and (d) respectively, plotted as a function of the number of RNA ligands, and calculated at the DFT/M06/6-311++G** level for the models shown in Figure 5.7. The Natural Bond Orbital (NBO) charge is used to estimate the charge. Gold circles, red squares and blue triangles refer to model systems characterized by the presence of O_{ph}-only, at least one O_b or one N_b as non-water ligands, respectively. Model systems characterized by O_{ph} ligands only but corresponding to a different geometrical isomer are indicated with golden empty circles. For each model system, the CP is reported in parenthesis.

For each model, we have also calculated the free energy of formation considering the water/non-water ligand exchange reaction ($\Delta G_{form\text{-}is}$, Figure 5.9b) (see methods section). This gives a measure of how each type of binding site contributes to the stability of a folded RNA macromolecule. $\Delta G_{form\text{-}is}$ shows an almost linear trend until three non-water ligands interact with magnesium, while slightly deviating from the linearity when a fourth O_{ph} ligand is introduced. This is perfectly in line with the small

decrease of both polarization and charge transfer occurring in this motif. If we consider the stabilization energy per RNA ligand in the coordination sphere (calculated as $\Delta G_{form\text{-}is}$ divided by the number of RNA ligands), we clearly see that the contribution of the first phosphate is the largest, while all the subsequent ligands similarly contribute to the stability of the coordination site (Figure A2.22). Interestingly, remarkable differences among the models displaying the same number of RNA ligands (two or three) are observed. These are mostly due: i) to the electrostatic contribution associated with the formal charge of the ligands (for example between $3O_{ph}$:$3O_w$ and $2O_{ph}$:$1N_b$:$3O_w$ exhibiting a -3 and -2 charge, respectively); ii) to the entropic contribution in case of geometrical isomers like for instance 10-member ring (2 consecutive phosphates) and Mg-Clamp (2 separated phosphates) with the same $2O_{ph}$:$4O_w$ CP. Indeed, the negative entropic contribution upon Mg^{2+} binding is clearly smaller for the former model, resulting in a more negative ΔG. By changing the solvent dielectric constant to 4, thus simulating the RNA environment, we note a similar trend of the $\Delta G_{form\text{-}is}$, even if in a larger extent, and an almost equal CT (~0.1 % variation) towards Mg^{2+} ions (Figure A2.21).

Electronic signature of Mg^{2+}–RNA *outer-sphere* coordination sites. A DFT-NBO analysis has been performed also for the *outer-sphere* models (Figure 5.7b). The polarization and CT effects occurring in each system are extended also to the second coordination shell (Figure 5.10e). Indeed, upon interacting with the hexa-hydrated Mg^{2+} ion, the second shell ligands get polarized and transfer part of their charge to the first shell water molecules, resulting in a peculiar LCR of H_w and O_w. By virtue of this contribution from the *outer-sphere*, the Δq of the water hydrogens (H_w) decreases (becomes less positive) with respect to the $6O_w$ motif, while the O_w atoms accumulate a large amount of negative charge. For the sake of clarity, while the CT taking place between the second and the first shell of coordination is evaluated for the *inner-sphere* sites, the polarization effect exerted on the *outer-sphere* RNA moieties is not detailed.

By looking at the CT taking place from the *outer-* to the *inner-sphere* ligands (Figure 5.10a) an almost linear contribution of each phosphate is observed, with the nucleobases participating to a lesser extent. We remark that the amount of CT might depend on the number of atoms in the *outer-sphere* ligands directly interacting with the *inner-sphere* water molecules. The effect of introducing a second shell of ligands is clearly perceptible in the LCR of the first shell (Figure 5.10b and 5.10c). In fact, in the model with 4 phosphates in the outer shell, the Δq for each H_w decreases by 13% (less positive), while the Δq for each O_w becomes two fold more negative than in the free $6O_w$ motif. Interestingly, the amount of charge transferred to the Mg^{2+} ion (Figure 5.10d and 5.9c) is similar to that observed for the *inner-sphere* models. Also in this case, a constant NBO charge ranging from -0.18 e to -0.19 e is withdrawn by the metal,

independently on the number of the RNA moieties in the *outer-sphere*. This saturated amount of CT makes the LCR of the water molecules extremely relevant.

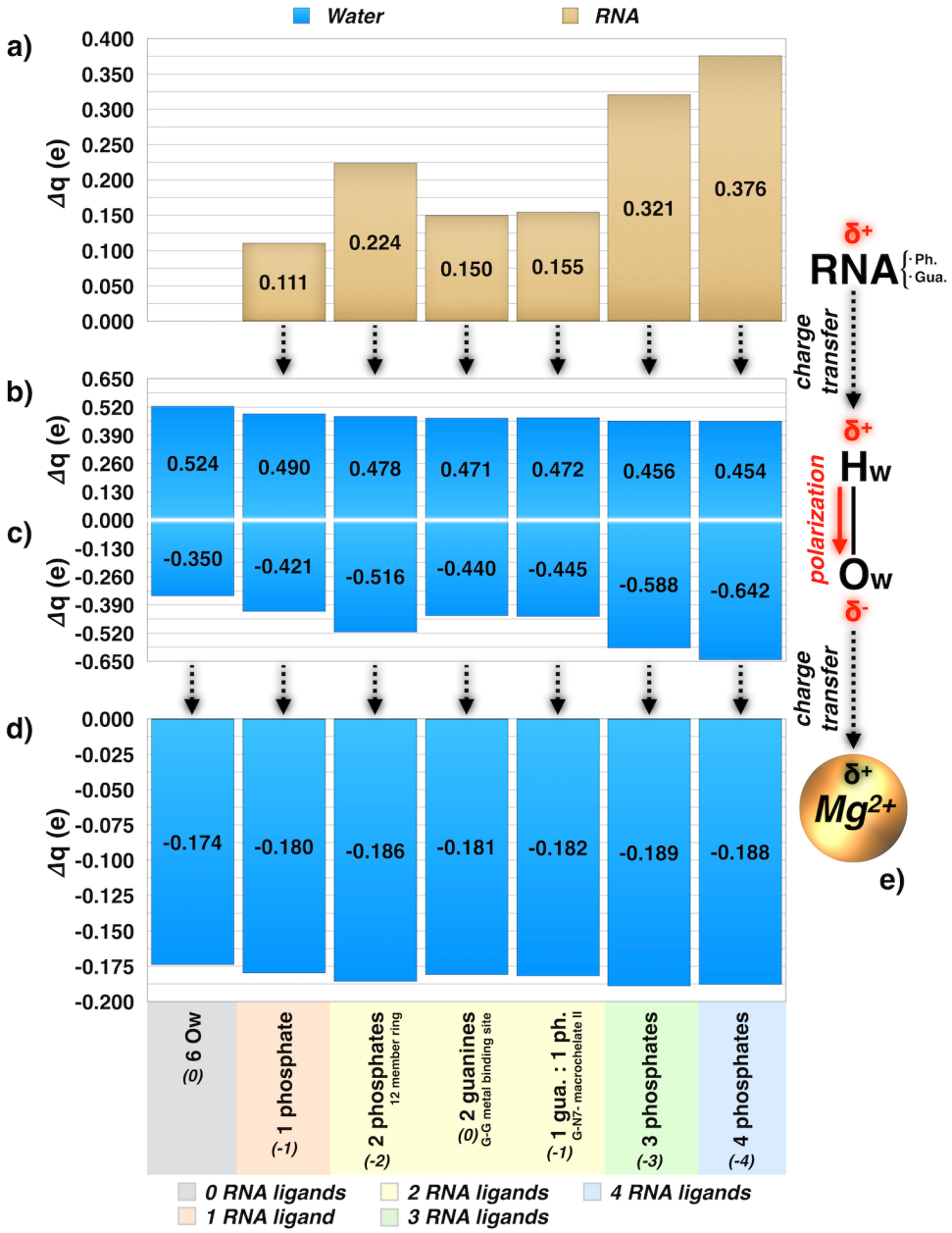


Figure 5.10. (a) Charge transfer (Δq, e) from the second shell ligands (phosphates or guanines) in the *outer-sphere* coordination sites; charge rearrangement (Δq, e) of H_w (b) and of O_w (c) of first shell water molecules; (d) amount of charge (Δq, e) transferred from the waters towards Mg^{2+} ion calculated from the NBO charge distribution and the M06 functional with 6-311++G** basis set; (e) schematic picture of the polarization and charge transfer effects exerted by a Mg^{2+} ion. The formal charge of each CP is reported in parenthesis.

As expected, the $\Delta G_{form\text{-}os}$ of the *outer-sphere* sites is smaller than the corresponding *inner-sphere* models (Figure 5.9d). After a second RNA ligand is

introduced in the second shell of coordination, the free energy of formation reaches a plateau, suggesting that more than two second shell RNA ligands do not contribute to an extra stabilization, conversely to what observed for the *inner-sphere* sites. However, our results highlight the importance of *outer-sphere* coordination sites in the stabilization of RNA structures, even if in a lesser extent with respect to the *inner-sphere* ones. Finally, the $\Delta G_{form\text{-}os}$ contribution per RNA ligand (Figure A2.22) reaches its maximum when two phosphates are present, then notably decreasing in the models with three and four RNA ligands.

5.5 Conclusions

The extensive benchmarks performed from MD simulations of G2IR and HDV ribozymes have shown remarkable differences in the performances of distinct Mg^{2+} FF models. This indicates that a conscious use of the classical FF parameters for Mg^{2+} ions is essential when running MD simulations of RNA filaments. This choice rigorously depends on the structural features of the specific Mg^{2+} site in RNA and on the experimental working conditions relative to Mg^{2+} concentration.

By performing DFT calculations on a representative set of model systems of Mg^{2+}–RNA binding architectures we have strikingly disclosed for the first time that Mg^{2+} ion exhibits similar electronic properties in varying Mg^{2+}–RNA binding sites, while remarkable differences are observed for its surrounding ligands (Figures 5.8, 5.9a, 5.9c, 5.10). As such, these results stunningly point out that the development of Mg^{2+} site-specific FFs is not in line with the physical origin of the inaccuracies in the Mg^{2+}–RNA MD simulations [201].

We report below a series of practical FFs user guidelines for an adept use of current Mg^{2+} models and some key insights from electronic structure calculations, constituting a rationale for the development of next-generation Mg^{2+} FFs.

5.5.1 Practical FFs user guidelines

As a general feature our extensive benchmark among the available Mg^{2+} ion models has revealed that all the FFs tend to underestimate the Mg^{2+}–N_b contacts. Both Åqvist and Allnér parametrizations better account for the experimental distribution of RNA ligands [187], although overestimating the Mg^{2+}–O_b contacts [214]. Additionally, while the Allnér model reproduces more diverse and hydrated Mg^{2+} binding sites, the Åqvist model better accounts for highly RNA-coordinated sites (Figures 5.3 and 5.4). As such, our findings can summarily suggest:

(i) The Allnér parameters are less sensitive to variations in Mg^{2+} ion concentration.

(ii) The Allnér parameters tend to reproduce more hydrated sites and are thus indicated for *outer-sphere* binding sites or for small RNA molecules in which few highly chelated phosphate sites are present.

(iii) The Åqvist parameters reproduce highly phosphate coordinated sites and may be the best choice in large RNA macromolecules.

(iv) The Saxena parameters reproduce the active site properties of a two-Mg^{2+}-ion motif irrespective of the Mg^{2+} concentration employed and are thus recommended for these highly buried coordination sites in which other divalent ions are close by.

(v) Careful attention should be given to the experimental working conditions relative to Mg^{2+} concentration, given the remarkable differences shown by the Åqvist and Allnér FFs in the reproduction of the Mg^{2+}–RNA coordination sphere.

5.5.2 Mg^{2+} force fields developer guidelines

Surprisingly, this study reveals that the electronic properties of Mg^{2+} ion remain constant in all these varying RNA environments. On the basis of these findings, a set of developer guidelines is suggested below:

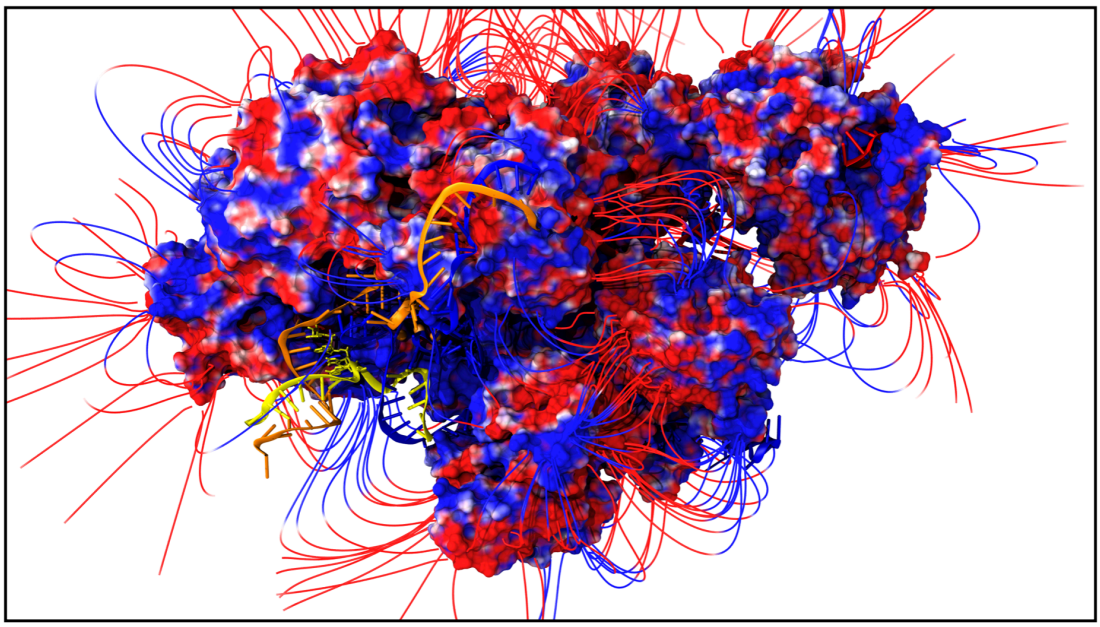
(i) The net amount of charge transferred to Mg^{2+} ion is roughly constant after a first RNA ligand is coordinated. This suggests that also the vdW parameters of Mg^{2+} should be maintained in different RNA environments.

(ii) Charge rearrangements on first shell RNA ligands are quite similar for the different binding motifs investigated, with the largest contribution coming from the nucleobases.

(iii) Special care must be devoted to water molecules, which demonstrated the unique capability to act as a buffer, adapting the amount of charge transferred to Mg^{2+} ion and the extent of polarization to the specific coordination site and for which the largest differences in LCR among the different binding sites are observed.

(iv) The common view/practice of adapting Mg^{2+} ions to the specific binding site seems to be in contrast with the physical origin of inaccuracies in the Mg^{2+}–RNA MD simulations. In this scenario, the development of site specific FF parameters for the different binding sites appears to be an immediate, practical, but an unphysical manner to indirectly account for the different electronic effects induced by Mg^{2+} on the surrounding ligands.

6 Atomistic characterization of spliceosome dynamics



Reference paper: *in preparation*.

6.1 Abstract

The spliceosome (SPL), a protein-directed ribozyme composed by 100s of proteins and four small nuclear (sn)RNA, removes non-coding intronic sequences from premature messenger (m)RNA transcripts with a single nucleotide precision, and ligates coding exons. This machinery produces functional mRNA, being, therefore, a key regulator of gene expression. The advent of high-resolution cryo-electron microscopy has revolutionized the SPL structural biology so far hampered by its exceptional conformational and compositional plasticity taking place along mRNA maturation cycle.

By exploiting the first SPL structure solved at near-atomistic resolution, here, we have characterized for the first time, through extensive force field-based MD simulations (> 2 microseconds), the dynamical behavior of the central core regions (i.e. 16 proteins, 3 snRNA and the intron lariat in explicit water) of the post-splicing Intron-Lariat-Spliceosome complex (ILS), which forms upon the release of the spliced mRNA. The essential dynamics of the spliceosome components strikingly reveals the unwinding motion of the IL/U2 branch helix cooperatively promoted by Cwf19 and Spp42 proteins. In this scenario, Spp42, the central scaffold of the SPL assembly, assumes a unique role in orchestrating the SPL genomic symphony, governing the motion of many different key proteins/snRNAs components. Thus, our study provides unprecedented atomistic insights on the spliceosome functional plasticity. Our study, dispensing an additional fundamental small piece of knowledge, contributes to move a step forward towards a detailed understanding of pre-mRNA splicing, with a potentially fundamental impact in biology, medicine and biotechnology.

6.2 Introduction

Precursor messenger RNA (pre-mRNA) splicing is a process of paramount importance for living cells, spanning all the domains of life. Splicing is a majestic "snip and stitch" editing of primary RNA transcripts emerging from genes transcription. These are composed by intervening non-coding sequences that must be removed, the introns, and by neighboring protein-coding tracts that instead must be joined together, the exons. This process yields to functional mRNA filaments, which are then translated into proteins [39]. Each splicing cycle entails two subsequent S_N2 transesterification reactions, namely the first branching step and the second exon ligation step. During the first step, the 2'-OH of a bulged adenosine of the branching

point sequence (BPS) near to the 3'-end of the intron is recruited to carry out a nucleophilic attack on the phosphate of a guanine at the 5'-end of the intron (intron 5'-splice site, 5'SS), resulting in the formation of a lariat intron–3'-exon intermediate and a free 5'-exon. In the second step, the free 3'-OH group at the 3'-end of the 5'-exon triggers the nucleophilic attack on the phosphorus at the 5'-end of the 3'-exon (intron 3'-splice site, 3'SS), leading to the release of the lariat intron and the two ligated exons. Both the phosphoryl-transfer reactions are facilitated by two Mg^{2+} ions, which act as natural cofactors and allow a two-metal-ion catalysis [21]. In eukaryotes, the main actor of pre-mRNA splicing is an extraordinary multi-megadalton machinery, called the spliceosome (SPL). Due to the complexity of the eukaryotic genome, the spliceosome has to recognize introns which are extremely different in sequences and lengths (ranging from 100 to 100'000 nucleotides). Correct splicing is mandatory for genome fidelity and its dysfunction are associated to 33% of genetic diseases. This astonishing ribonucleoprotein (RNP), composed by five small nuclear RNAs (the U1, U2, U4, U5, and U6 snRNAs) and hundreds of different proteins endowed with diverse functions (~100 in yeasts and more than 300 in humans) [225], is the tailor of the splicing process, capable of snipping and stitching pre-mRNA with single nucleotide precision. In particular, snRNAs and proteins are assembled to form spliceosomal snRNPs subunits, which, together with other protein complexes like NTC and NTC-a (NineTeen complex and NineTeen Complex-associated, respectively), enzymes and cofactors, tune the spliceosome catalytic, structural and dynamical properties along its functional cycle [226]. The spliceosome assembly begins with the recognition of the 5'SS and 3'SS of the pre-mRNA precursor by U1 snRNP and U2 snRNP, respectively, forming the complex A. The recruitment of the preassembled U4/U5.U6 tri-snRNP gives rise to the complex B, which is subsequently activated into B^{act} upon the ejection of U1 and U4 snRNPs and recruitment of NTC and NTC-a. A subsequent rearrangement converts B^{act} into B^{*}, which undergoes the first transesterification reaction. The RNA components are in charge of the catalysis, with an active site architecture and a chemical mechanism strikingly resembling those of its evolutionary ancestors, i.e. group II introns ribozymes (G2IRs) [32, 44], for which we recently unveiled the mechanism of the first step of splicing [29]. The product of the branching step is the complex C. This is rapidly converted into the activated C* complex who carry out the second step, resulting into the final post-catalytic P complex, containing the ligated exons and the intron lariat. Afterwards, the joined exons are released, while the intron lariat is maintained in the last complex of the cycle, namely the intron-lariat-spliceosomal (ILS) complex. Finally, the intron lariat is released, followed by the spliceosome subunits disassembly and recycling.

The SPL is a stunningly dynamic machine, which undergoes a continuous conformational and compositional remodeling during its cycle [226]. Hence, its highly dynamical nature, along with its large size prevented its characterization by the X-ray crystallography. The advent of high-resolution cryo-electron microscopy (cryo-EM) has prompted a transformative era in the structural biology of the SPL. Indeed, an increasing number of structures from yeast and humans has been released in the past two years [31-34, 50, 59-62, 64-67, 75, 83]. This burst of spliceosome structural biology has contributed to provide fundamental insights at near-atomistic level on its assembly, reactivity and composition.

Among all the latest high-resolution cryo-EM models, here we focus on the first structure solved from the *Schizosaccharomyces Pombe* reconstructed at an average resolution of 3.6 Å [31, 32] and deposited in Protein Data Bank with id 3JB9, which embodies the landmark of the SPL structural biology. This structure exhibits a good definition of the intron lariat (IL), while showing a weak EM density for the exon, which was either already released as pre-mRNA, or lost during the purification. Not only the 5'-exon absence, but also some essential regulatory factors characterizing the C and C* complex (i.e., Slu7, Prp18, Prp22) are missing, suggesting that this structure most likely corresponds to the post-splicing ILS complex [31, 67]. Most importantly, this model provided for the first time precious near-atomistic details (exceeding 3.2 Å for some proteins in the core region) on the intact catalytic site architecture and on four multicomponent subcomplexes, U5 snRNP, U2 snRNP, NTC and NTC-a, comprising a total of 37 proteins, 3 snRNAs and the IL. In particular, nearly complete atomic models for some crucial U5 snRNPs proteins like the central Spp42 (Prp8 in *S. Cerevisiae*) and Cwf10 (Snu114 in *S. Cerevisiae*) were elegantly defined along with a first glimpse into some NTC and NTC-a proteins.

In this study, we unveil unprecedented atomistic insights on spliceosome structure and dynamics by means of extensive microsecond long molecular dynamics (MD) simulations of a ~1 million atoms model system exploiting the spliceosomal ILS complex firstly reconstructed by Yan and coworkers [31, 32]. Our MD simulations are focused on the central core region of the spliceosome, solved at higher resolution and representing the RNP scaffold conserved in many steps of the spliceosome cycle.

Our study strikingly elucidates the central role of Spp42, which acts as an orchestra conductor, directing the slow functional motions of key spliceosome components. Remarkably, we disclose an essential involvement of the Cwf19 protein, which, jointly with Spp42, unwinds the IL/U2 double helix, inducing the release the IL, before the spliceosome final disassembly occurs. Altogether, our findings represent an unprecedented contribution - at atomistic resolution - to the current understanding

of splicing, providing general mechanistic details on the principal functional motions of the core spliceosomal subunits.

6.3 Methods

Model systems. MD simulations of *S. Pombe* spliceosome were based on the cryo-EM model reconstructed at 3.6 Å resolution (PDB id 3JB9) [31, 32]. This structure shows an asymmetric morphology which exceeds 300 Å in its longest dimension. While the core proteins and RNAs have a resolution ranging from 2.9 Å to 3.6 Å (up to 5 Å in some cases), the most peripheral regions exhibit a poor EM-density (with a resolution larger than 5 Å) and present large gaps, hampering their accurate modelling. In our work, we aimed to study the dynamics and the properties of the most important spliceosomal components. For this reason, in addition to the above-mentioned issues, we considered only the central core region of the structure (Figure A3.1 in the Appendix A3). In particular, we built two model systems of different sizes, namely a "test model" (model-1, Figure 6.1) of 721'089 atoms and an "extended model" (model-2) counting 914'099 atoms, whose data are not showed in this thesis as only preliminary.

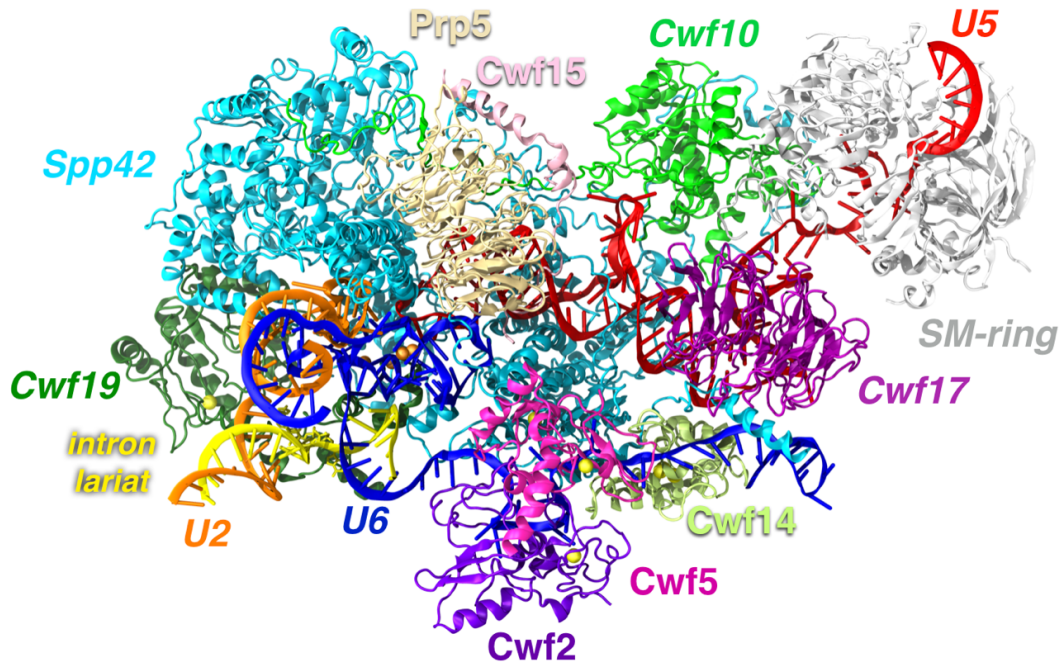


Figure 6.1. Spliceosome model-1 based on the yeast *S. Pombe* ILS complex, reconstructed at 3.6 Å. Proteins, snRNA and the intron lariat are shown with new cartoon representation and highlighted with different colors. Mg^{2+} and Zn^{2+} are depicted with orange and yellow spheres, respectively. Na^{+} ions and water layer are omitted for the sake of clarity.

Model-1 consists of 16 proteins, 3 snRNAs (U5, U6 and U2 snRNA) and the intron lariat. Among the proteins present in the deposited structure, we considered (i) Spp42, Cwf10, Cwf17 and the 7 Sm-ring chains of U5 snRNP, (ii) Cwf2 and Cwf15 of the NTC core, and (iii) Prp5, Cwf5, Cwf19, Cwf14 from NTC-a. In model-2 we included additional domains of Spp42 (Endonuclease and RNaseH-like domains, 498 aa in total) and Cwf10 (domain I, II, III, IV and V, 571 aa in total) and two extra proteins (for a total of 18 proteins), i.e. Prp45 of NTC-a and Prp17, which increased the size of the system by ~11'000 heavy atoms with respect to model-1. Both systems were embedded in a 14 Å layer of TIP3P water molecules [162], thus leading to a periodic box size of $168 \cdot 193 \cdot 249$ Å^3 (model-1) and $212 \cdot 189 \cdot 256$ Å^3 (model-2), containing also the 4 catalytic Mg^{2+} ions, 7 Zn^{2+} and 202/194 (model-1/model-2) Na^+ that were added as counter ions.

The final atomic systems were generated using the coordinates provided in the original PDB entry. Chain A (Spp42), E (Sm-B), G (Sm-D2) and L (Cwf17) contained small gaps due to unresolved residues (from 1 up to 12) in the fragments effectively included in our study. *De novo* model building as implemented in Modeller 9v16 [227] was used to reconstruct the missing loops, which were further refined through the loop refinement procedure [228, 229]. We remark that Modeller has been shown to be very accurate for small loops modeling [230]. The generated loops were first selected among 50 models according to the DOPE score [231] and subsequently evaluated through an accurate visual inspection. The complete information about spliceosome model-1 and model-2 are provided in Figure A3.2 and A3.3 in the Appendix A3.

The simulations on the model-2 were still ongoing at time of writing of the thesis and, as such, these results are not discussed here.

Molecular dynamics simulations. The two models were subjected to extensive classical MD simulations carried out with Gromacs5 software package [232]. The AMBER-ff12SB force field was adopted for proteins [107], with ff99+bsc0+χOL3 for RNAs [163, 164], since these are the most validated and recommended force field for protein/RNA systems [105]. Mg^{2+} ions were described with the non-bonded fixed point charge parametrization due to Åqvist [207] as it was shown to properly describe binuclear sites [233]. Na^+ ions parameters were taken from Joung et al. [210] while Zn^{2+} ions were modelled with the cationic dummy atoms approach developed by Pang [234]. The RESP charges of the branching adenosine of the intron lariat (A501) were derived upon a minimization of the A501-G100 dinucleotide (named GA2) performed with Gaussian 09 [171] at Hartree-Fock level of theory with 6-31g* basis set, followed by a fitting on the electrostatic potential with the antechamber module of ambertools12 [170].

MD simulations were performed on the isothermal-isobaric ensemble using periodic boundary conditions. Temperature control at 300 K was achieved by stochastic velocity rescaling thermostat [235], while pressure control was accomplished by coupling the systems to a Parrinello-Rahman barostat with a reference pressure of 1 bar [114, 115]. LINCS algorithm [236] was used to constrain the bonds involving hydrogen atoms and the particle mesh Ewald method to account for long-range electrostatic interactions with a cutoff of 12 Å [98]. Five replicas (three for model-1 and two for model-2) of ~750 ns each were run using an integration time step of 2 fs, reaching an overall simulation time of ~4 μs (2.25 μs for model-1 and 1.5 μs for model-2).

In all the simulations, we have used a very careful and slow equilibration protocol. Namely, the systems were initially put through a soft minimization using a steepest descent algorithm with a force convergence criterion set to 1000 kJ mol^{-1} nm^{-1}. Then, the models were smoothly annealed from 0 to 300 K with a temperature gradient of 50 K each 2 ns and for a total of 12 ns. In this phase, only water molecules and Na^+ ions were allowed to freely move while the rest was subjected to harmonic position restraints with a force constant of 1000 kJ/mol · nm^2. Once the temperature was raised up to 300 K, 20 ns of NPT simulations were conducted to stabilize the pressure to 1 bar by coupling the systems to a Berendsen barostat [113] and imposing the same restraints used in the heating phase. Subsequently, the barostat was switched to Parrinello-Rahman and the position restraints on proteins and RNAs were restricted only to the backbone atoms. These were gradually decreased in three consecutive steps of 30, 10, 10 ns each, during which the force constant was set to 1000, 250, 50 kJ/mol · nm^2, respectively. Finally, after an attentive equilibration protocol of ~80 ns, all the restraints were released and the production runs were performed for ~670 ns, for a total of ~750 ns, saving the snapshots each 50 ps. The trajectories were visualized with Visual Molecular Dynamics (VMD) software [237]. Analysis were done on trajectories with a frame stride of 2 (i.e., 1 frame each 100 ps). In particular, analyses of the Root Mean Square Deviation (RMSD), Root Mean Square Fluctuation (RMSF) and Radius of gyration (R_G) have been performed with the cpptraj module of Amber16 [238] and Gromacs5 suite [232].

Principal Component Analysis (PCA). PCA was applied to the sampled trajectories to extract the essential dynamics of the spliceosome. PCA was performed with cpptraj module of ambertools16 [238].

PCA can report on the large-scale, collective motions occurring in biological macromolecules undergoing MD simulations. Thus, PCA can provide valuable information on major conformational changes taking place along the trajectory [239, 240]. In fact, through this statistical technique it is possible to reduce the large number

of degrees of freedom to an essential subspace set, which captures large-amplitude motions of the system [241]. As such, the essential dynamics is a widely used tool to disclose biological functions of biomolecules, which are related to their principal motions.

Here, the essential motions of proteins and snRNAs have been captured starting from the mass-weighted covariance matrix of the Cα and P atoms, respectively. The covariance matrices were constructed from the atoms position vectors upon an RMS-fit to the reference starting configuration of the production run MD in order to remove the rotational and translational motions. Each element of the covariance matrix is the covariance between atoms *i* and *j,* defining the *i,j* position of the matrix. This is defined as:

$$C_{ij} = \langle (\vec{r}_i - \langle \vec{r}_i \rangle)(\vec{r}_j - \langle \vec{r}_j \rangle) \rangle \tag{6.1}$$

where $\vec{r}_i$ and $\vec{r}_j$ are the position vectors of atoms *i* and *j*, and the brackets denote an average over the sampled time period. In the present work (model-1), the matrix was calculated on 3833 C-alpha and 255 P atoms over 6700 frames, corresponding to last 670 ns of the MD simulations. The two terms in equation 6.1 represent the displacement vectors for atoms *i* and *j*. A positive sign of this product indicates that the two atoms move in a correlated manner, otherwise, if it is negative, it means that they are anti-correlated. If the product is zero, then it evinces that the atoms displacements are independent of each other. The covariance matrix was then diagonalized, leading to a complete set of orthogonal collective modes (eigenvectors), each with a corresponding eigenvalue (variance). The eigenvalues denote how much each eigenvector is representative of the system dynamics. Indeed, the eigenvectors with the largest eigenvalues correspond to the most relevant motions. By projecting the displacements vectors of each atom along the trajectory onto the eigenvectors (i.e., by taking the dot product between the two vectors at each frame), the Principal Components (PC) were then obtained. A total of ~6700 frames were used for PCA with the maximum number of eigenvalues given by *min*(n°-of-atoms, n°-of-frames) = 6700 PCs, i.e. one for each frame. A plot of the cumulative variance accounted by all the PCs was calculated with Gromacs5 suite [232]. However, only the motion of the system along the first PC (PC1) usually defines its "essential dynamics", even if also the second (PC2) and third (PC3) might be still significant. Thus, PC1 was plotted against PC2 to generate the scatter plot displaying how the conformational space defined by the first two modes is sampled through the MD simulations. This graph allows disclosing different states of the system.

Cross-correlation score. The cross-correlation matrices (or normalized covariance matrices) based on the Pearson's correlation coefficients (CC_{ij}) were calculated with the cpptraj module of ambertools16 [170] from the covariance matrices

previously obtained out of each simulation. The cross-correlation analysis offers the possibility to capture the linear coupling of the motions between two residues over the entire trajectory, without, however, bearing any information about the magnitude of the motions. Each element of the cross-correlation matrix in the *i*,*j* position corresponds to a Pearson's CC_{ij}, i.e. the normalized covariance between atoms *i* and *j*, calculated with the formula:

$$CC_{ij} = \frac{\langle (\vec{r}_i - \langle \vec{r}_i \rangle)(\vec{r}_j - \langle \vec{r}_j \rangle) \rangle}{\left[(\langle \vec{r}_i^{\,2} \rangle - \langle \vec{r}_i \rangle^2)(\langle \vec{r}_j^{\,2} \rangle - \langle \vec{r}_j \rangle^2) \right]^{1/2}} \tag{6.2}$$

where the normalization factor at the denominator is the product between the standard deviations of the two position vectors. Cross-correlation coefficients range from a value of -1, which indicates a totally anti-correlated motion between two atoms, and a value of +1, which instead means a linearly correlated lockstep motion.

In order to make the cross-correlation matrices more readable we have calculated the *intra-* and the *inter*-correlation scores for each of the 16 proteins and the 4 RNAs included in the model [242]. The *intra*-correlation score was obtained upon the summation of all the CC_{ij} within a protein/RNA, while the *inter*-correlation score was calculated by taking the sum of the CC_{ij} between residues belonging to the two different macromolecules (protein or RNA) considered. As such, for each macromolecule one *intra-* and $(m - 1)$ *inter*-correlation scores were computed, where *m* is the number of macromolecules included in the model. Importantly, the values $-0.6 < CC_{ij} < +0.6$ were discarded in the reckoning of the scores in order to eliminate the noise due uncorrelated motions. Indeed, our aim was to spotlight from the rough cross-correlation matrix only the most relevant correlated and anti-correlated motions between two proteins/RNAs to further inspect a possible biological function linked with their dynamics. All the scores (*intra* and *inter*) obtained for each macromolecule have been normalized by the highest score registered for that specific macromolecule. This reduced all the scores to values ranging from -1 to +1, getting rid of the bias due to the different sizes of the proteins considered (i.e., larger macromolecule, higher score). Subsequently the normalized scores were plotted in a histogram showing the correlation/anti-correlation motions between each pair of macromolecules.

Electrostatic calculations. Electrostatic calculations were accomplished by means of the Adaptive Poisson-Boltzmann Solver (APBS1.4) software [243]. The calculations were performed on the proteins included in model-1 considering the cryo-EM model and configurations harvested at different times along the simulations. APBS evaluates the electrostatic properties of large biomolecules by efficiently solving the Poisson-Boltzmann electrostatic equation (PBE) [243]. This analysis allows the unraveling of the electrostatics of a system, pinpointing the interactions between

positively and negatively charged residues that might be at the origin of crucial conformational changes or transitions. The selected geometries were first converted to the pqr format with pdb2pqr software [244, 245] with unvaried protonation state and by using the same force field employed in the MD simulations. Subsequently, following previous applications [246], APBS calculations were carried out using the Linearized Poisson-Boltzmann Equation (LPBE) with a grid spacing of ~0.7 Å, at 298 K and 150 mM as ionic strength for monovalent ions. The external dielectric constant was set to 78.0 to reproduce the aqueous medium, while the internal dielectric constant was fixed at 2.0 to mimic the non-polar environment of the solute.

6.4 Results and discussion

Conformational spliceosome dynamics. We performed ~2.3 μs (3 replicas) of MD simulations in explicit water on the core region of the ILS spliceosome complex. The first model discussed in this thesis was initially set as a test model to evaluate if this kind of macromolecular aggregates, solved at near-atomistic resolution, could be investigated by atomistic simulations. As such, the stability of the structures was assessed by monitoring the RMSD and the R_G along the MD (Figure 6.2 for replica #1.1, Figure A3.4 for the other two replicas, #1.2 and #1.3).

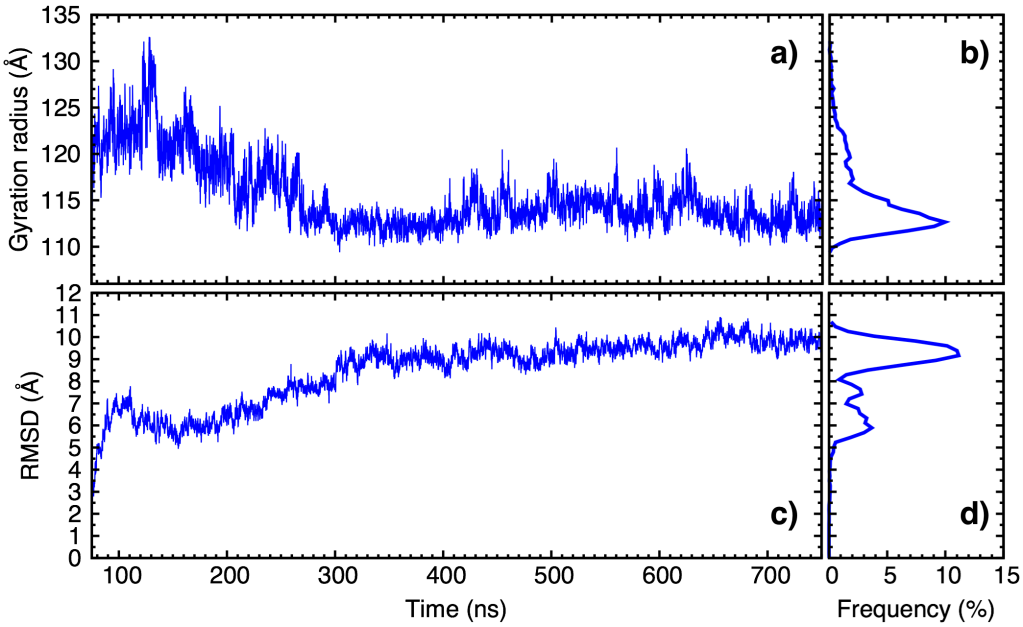


Figure 6.2. Time evolution (ns) of Gyration Radius (a) and RMSD (c) and their relative frequencies (%) shown in (b) and (d), respectively, obtained out of MD simulations of replica #1.1 MD. The profiles are obtained including all the proteins and RNAs in the analyses.

The histogram of the RMSD values in Figure 6.2d suggests the presence of an initial "open" state and final "compact" state. The R_G profile (Figure 6.2a and 2.b) exhibits a decreasing trend that underlines this smooth transition. In fact, during the first ~300 ns the system progressively evolves toward a more compact structure, visiting minor states before reaching the final conformation. Consistently, the distribution of RMSD and R_G values shows that the structure undergoes structural rearrangements. This general tendency is overall confirmed in the others replicas (Figure A3.4), even if it is more emphasized in replica #1.1 and #1.3.

As a note of warning, we remark that system evolution towards a compact structure may be at least in part due to the removal of some peripheral proteins solved at low resolution. For this reason, the results obtained from model-1 will be validated with the extended model system (model-2) (not discussed here).

In order to dissect the essential dynamics of our system, we have applied PCA analysis on the productive phase of the MD simulations with the purpose of ascertaining relevant motions possibly associated to spliceosome biological functions. Interestingly, the first principal component (PC1) accounts for more than 50% of the variance of the SPL motion, while the first 3 PCs for almost the 70% (Figure 6.3).

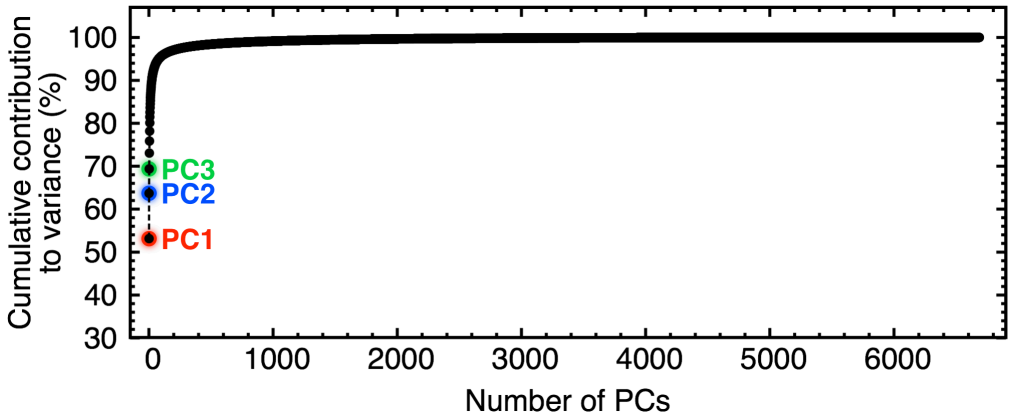


Figure 6.3. Cumulative contribution (%, y-axis) of all the principal components (PCs, x-axis) to the variance of the spliceosome motion. The contribution from the first three PCs are highlighted in red, blue and green, respectively.

To better decode the motion of the spliceosome in the conformational space we plotted the first two PCs one against the other, generating the scatter plot depicted in Figure 6.4. This clearly reveals that PC1 is mostly responsible of the conformational space sampling, with different of states visited along the trajectory.

In line with RMSD and R_G trend, the system evolves from an initial state, highlighted in yellow in Figure 6.4, to a final compact state (in red), which exhibits a narrower distribution. In the transition from the initial to the final state, a subset of less

defined clusters is observed (colored with lighter yellow and red in Figure 6.4). This indicates the pronounced dynamical behavior of the spliceosome assembly, in which the subunits progressively interact with each other giving rise to different configurations. Replicas #1.2 and #1.3 confirmed the outcomes of replica #1.1 with strikingly similar scatter plots (Figure A3.5 and A3.6).

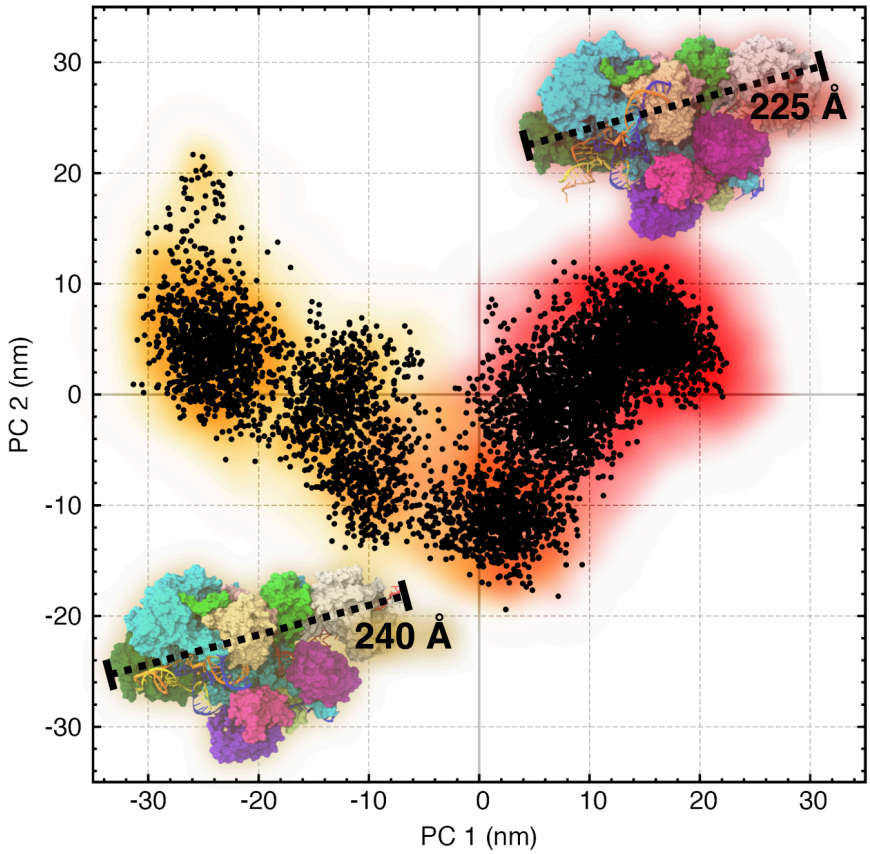


Figure 6.4. Scatter plot representing the projections of the C-alpha and P displacements along the trajectory onto the first principal eigenvector, PC1 (x-axis) vs the projections onto the second principal eigenvector, PC2 (y-axis) as derived from MD simulation of model-1, replica #1.1. The initial states are highlighted in yellow, while the final states in red.

The PCA analysis was done starting from the covariance matrix, which was subsequently normalized to construct the Pearson's coefficients cross-correlation matrix (Figure 6.5 for replica #1.1 and Figure A3.7, A3.8 for replicas #1.2 and #1.3, respectively) and its simplified *intra*-/*inter*-cross correlation histogram (Figure 6.6). In the histogram, the scores are plotted for each macromolecule in separated columns after being normalized by the highest score of each column. By doing this, it is possible to unmask the correlated and anti-correlated motions for all the components of the spliceosome, circumventing the bias of their different sizes.

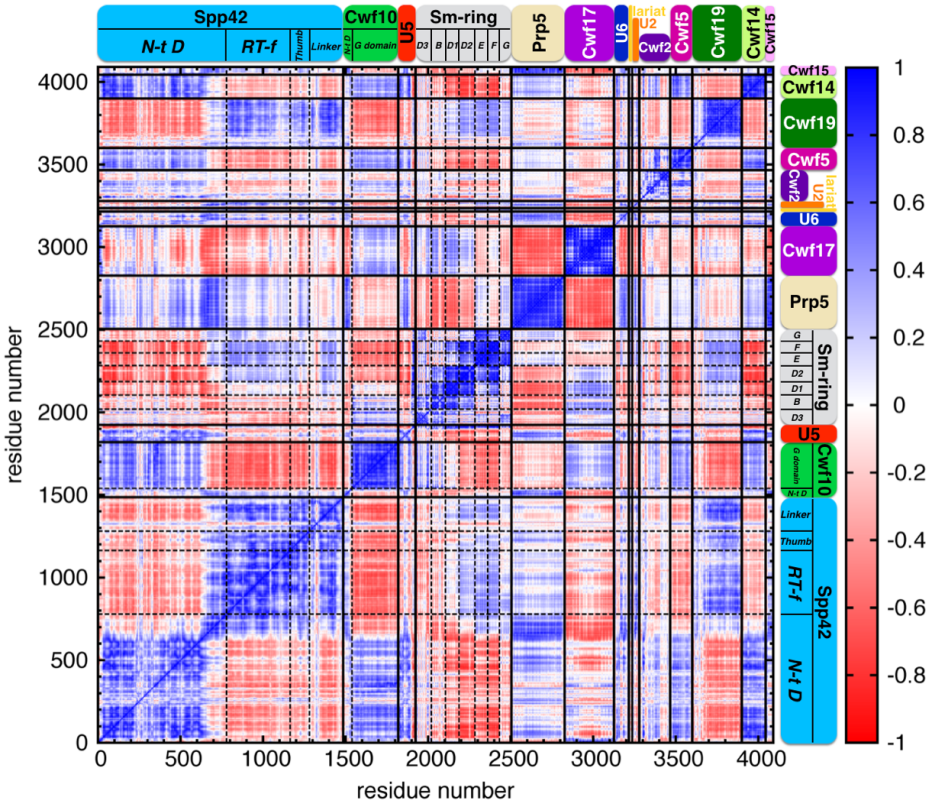


Figure 6.5. Pearson's coefficients cross-correlation matrix derived from the mass-weighted covariance matrix constructed over the last 670 ns of MD simulations of replica #1.1 for C-alpha and P atoms. The Pearson's CC_{ij} are comprised between -1 (anti-correlation, red) and +1 (correlation, blue). Macromolecules and domains names are highlighted with different colors.

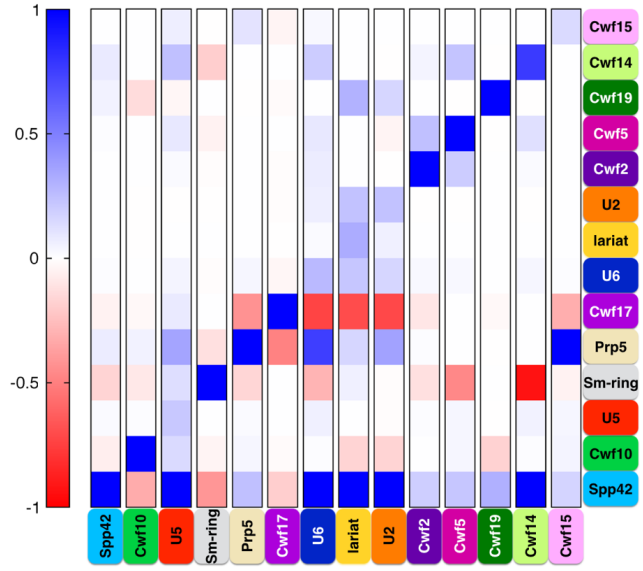


Figure 6.6. Histogram reporting the normalized *intra-* and *inter*-correlation scores between the macromolecules of the system listed at the bottom and on the right and highlighted with different colors. The histogram is normalized per-column, therefore it is not symmetric, and must be read per-column.

Figure 6.6 shows that (i) all the snRNAs and the IL are highly coupled with Spp42, which constitutes the protein central scaffold of the macromolecular aggregate. Spp42 belongs to U5snRNP, which is the only invariant snRNP for both splicing reactions, thus representing the hub of the spliceosome. (ii) U5 snRNA moves concertedly with the proteins of the U5 snRNP (Spp42, Cwf10, Cwf17 and Sm-ring). (iii) The proteins composing U5 snRNP exhibit an anti-correlated dynamics among each other, modulated by the different domains of Spp42. (iv) The NTC (Cwf2, Cwf15) and the NTC-a (Prp5, Cwf5, Cwf19, Cwf14) proteins shows an opposite behavior, all moving lockstep (blue scores in the top-right corner of Figure 6.6) among each other, but in opposite direction with respect the Sm-ring and Cwf17 of U5 snRNP. These anti-correlated motions may be at the origin of the gradual stabilization of the spliceosome towards a more compact structure, revealed by the R_G and PCA analyses. (v) Strikingly, the intron lariat and U2 snRNA are the only RNA filaments whose motions are correlated both with Cwf19 and Spp42 proteins. In particular, the Pearson's coefficients matrices (Figure 6.5, A3.7 and A3.8) show that Cwf19 is positively correlated only with the Reverse-Transcriptase palm/finger domain (RT) of Spp42 and they concertedly interact with the double helix formed by the intron lariat and U2 snRNA, in proximity of the catalytic site. These motions are of utmost relevance as they might be linked with the unwinding of the IL/U2 double helix as discussed in the next paragraph.

IL/U2 unwinding. After the second step of catalysis, the final mRNA product is released, while the intron lariat remains transiently bound to the spliceosome in the ILS complex [67]. The structure investigated here captures a spliceosome state right after the splicing reaction, where the IL still forms a double helix with U2 snRNA. It is known that the intron lariat becomes accessible for linearization only upon the displacement and unwinding of the IL/U2 helix (branch helix) [247]. As such, the displacement of the branch helix is a critical event which is supposed to be marked also by the Cwf19 debranching factor (CWF19L2 in humans), the *S. Pombe* paralog of *S. Cerevisiae* Drn1 [67, 248, 249]. This latter is a splicing factor that enhances the action of the debranching enzyme Dbr1 in *S. Cerevisiae* [249]. According to Garrey and coworkers [249], since *S. Pombe* and humans lack the evolutionary homolog of Dbr1, Drn1 paralogs like Cwf19 and CWF19L2 may have been endowed with auxiliary roles, exceeding the simple regulation of Dbr1 function. The cryo-EM spliceosome model subject of the present study [31, 32] is the first providing structural information about Cwf19, whose functionality remains still unclear.

Our MD simulations, complemented by PCA and electrostatic analyses strikingly disclose a dual participation of Spp42 and Cwf19 in unwinding and displacing of IL/U2 helix, which might unprecedentedly shed lights on the uncertain role of Cwf19.

Indeed, APBS analysis on the cryo-EM structure reveals the presence of a positively charged cavity rich in arginines and lysines formed by the RT domain of Spp42 and Cwf19 (Figure 6.7). Namely, among many others reported Figure 6.7, residues R878 and R884 from Spp42 anchor U2 snRNA, whereas R388, K360, K364 from Cwf19 interact with the branch point adenosine and its flanking 3-downstream nucleotides. This pocket acts as an electrostatic trap for the negatively charged RNA nucleotides of U2 and IL.

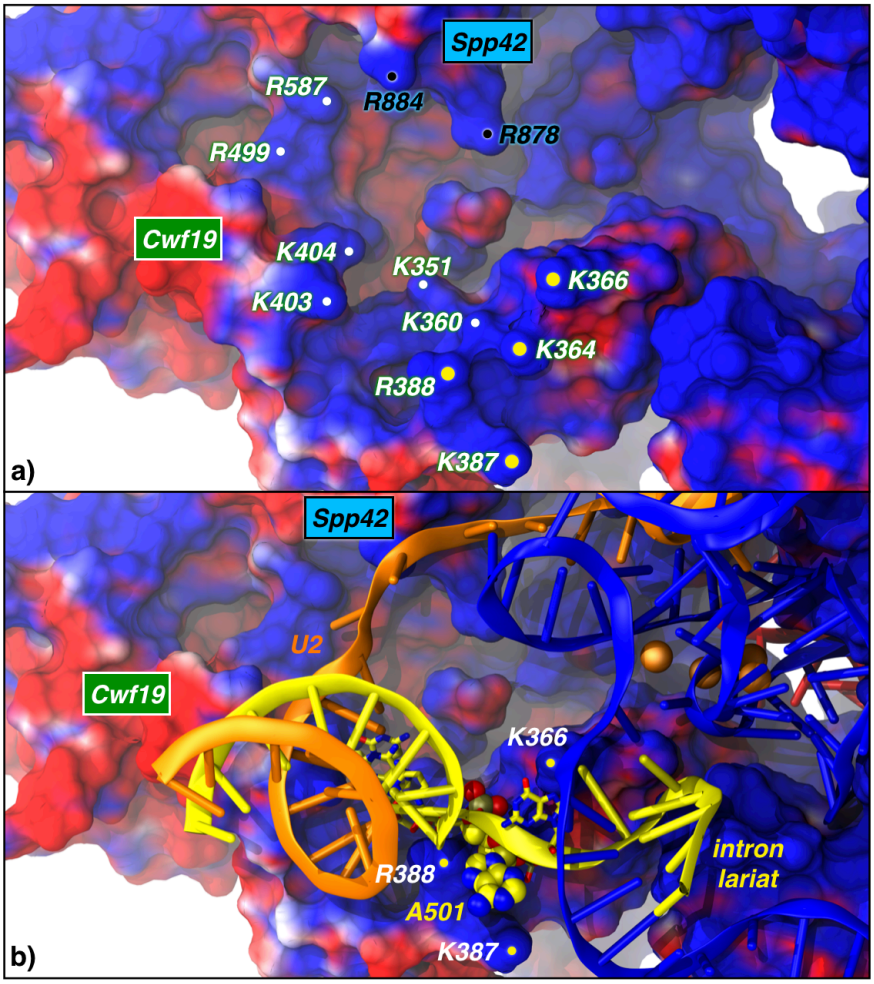


Figure 6.7. (a) Positively charged pocket formed by Cwf19 and Spp42 as in the cryo-EM structure, represented with the electrostatic surface where blue indicates a positive charge, and red the negative one. The most important positively charged residues are indicated with white (Cwf19) and black (Spp42) labels. K364, K366, K387, R388 are highlighted with a yellow dot as they are in proximity of the branching adenosine. (b) U2, U6 snRNAs (orange and blue), and IL (yellow) are also represented, with the branching A501 depicted with vdW spheres.

Our simulations show that two key events modulate the IL/U2 unwinding: i) the electrostatic recruitment of IL/U2 in the vicinity of the positive pocket at ~130 ns

(Figure 6.8a, b); and ii) the electrostatic lock of the BP adenosine modulated by R388, K387, K364 and K366 of Cwf19 (Figure 6.9a, b), which becomes stable at ~200 ns. These last residues form a "polar tweezers", tightly holding the branching adenosine and the IL 3'-termini into registry.

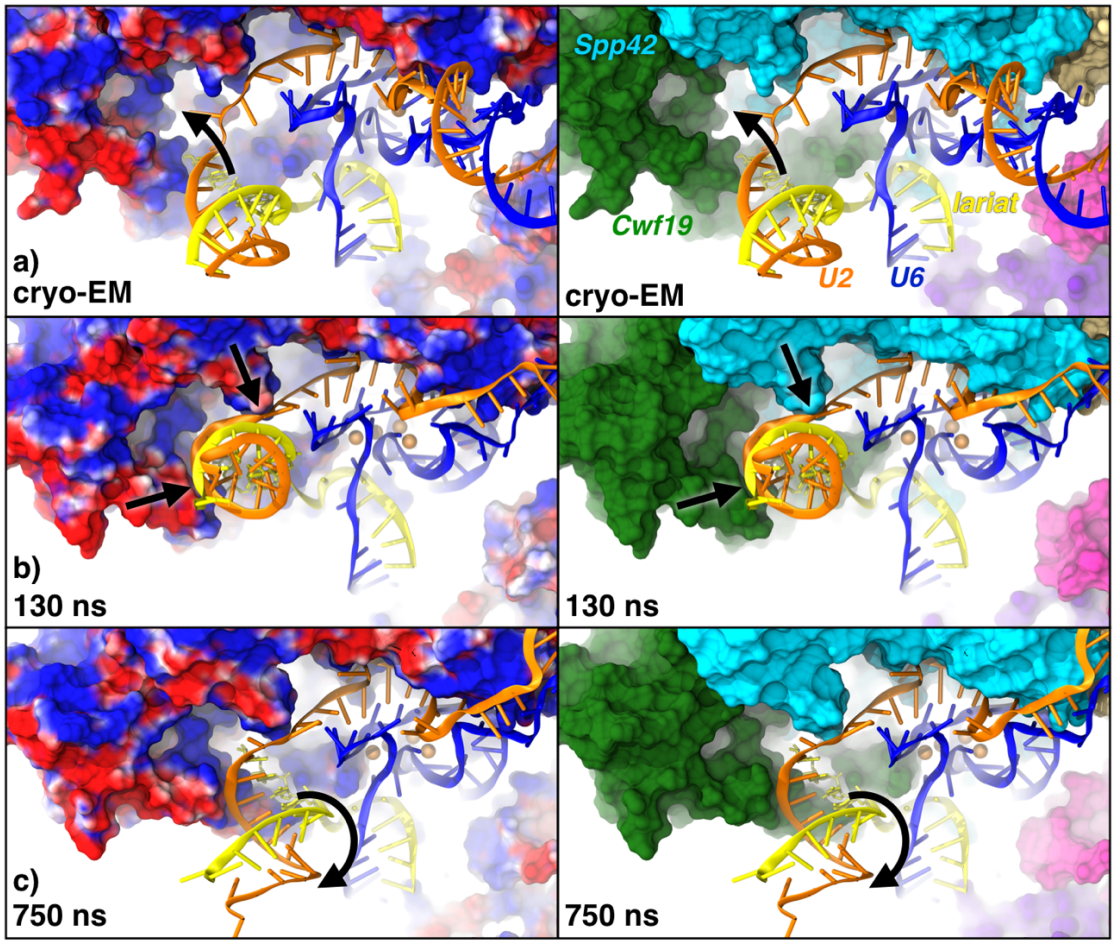


Figure 6.8. Snapshots from MD simulations taken at 0 (a), 130 (b) and 750 (c) ns, representing the electrostatically-driven unwinding of the IL/U2 helix, whose dynamics was unveiled by PCA analysis (Figure 6.11). On the left panels Cwf19 and Spp42 are represented with the electrostatic surface generated, where blue and red indicate positive and negative charges, respectively. On the right panels Cwf19 is depicted in green, Spp42 in light blue, U2 in orange, IL in yellow and U6 in blue. (a) The positive pocket formed by Cwf19 and Spp42 progressively recruits the IL/U2 helix due to the strong electrostatic interaction with the negative sugar-phosphate backbone of RNAs. (b) At ~130 ns the double helix tightly interacts with Cwf19 and Spp42 and at ~200 (shown in Figure 6.9) the 3'-termini of the IL is locked by an electrostatic tweezers formed by K364, K366, K387, and R388. (c) At ~750 the double helix is partially displaced and unwound due to the concerted action of Cwf19 and Spp42, with the branching adenosine and 3'-termini of IL firmly anchored.

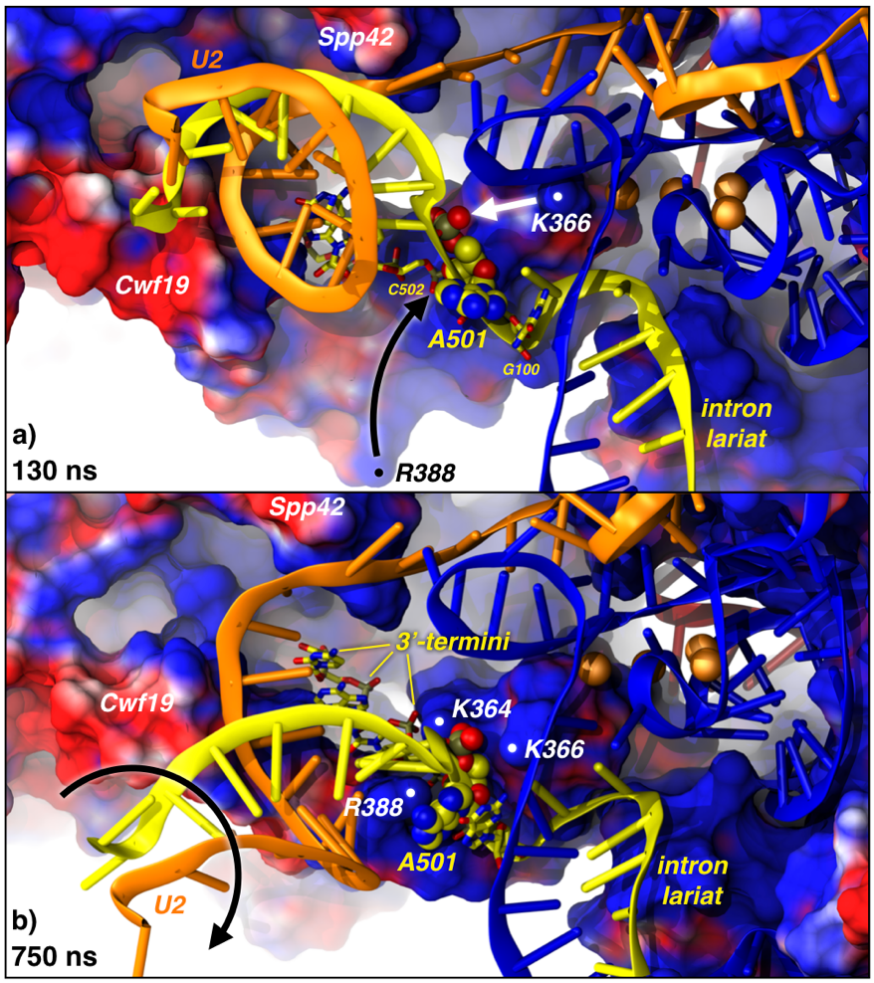


Figure 6.9. Snapshots from MD simulations taken at 130 (a) and 750 (b) ns. Cwf19 and Spp42 are represented with the electrostatic surface (blue positive charge, red negative charge). U2, U6 snRNAs and IL are depicted in orange, blue and yellow new cartoon. The branching adenosine, A501 is shown with VDW spheres, while the 3'-termini of the IL with licorice representation. (a) At ~130 ns R388 and K366 are separated from the BP due to the mobility of this region. At ~200 ns they are electrostatically recruited and progressively get closer to the BP. This motion is highlighted by the black (R388) and white (K366) arrows. (b) After this rearrangement, a polar tweezer firmly hold the branching adenosine (A501) and the 3-termini, constituting the pivot of the IL/U2 unwinding.

The formation of the electrostatic tweezers made by K364, K366, K387, and R388 around the branching adenosine (A501), involving also the C502 and G100 nucleotides of the IL stably anchor the 3'-termini of the IL, favoring the 5'-downstream unwinding of the double helix promoted by the concerted action of Spp42 and Cwf19.

Indeed, the positive hinge around the 3'-termini of the IL constitutes the pivot of the IL/U2 unwinding assisted by Cwf19 and Spp42. Interestingly, in the cryo-EM structure K364, K366, K387, and R388 residues are in the vicinity of A501. During the

MD simulations this region exhibits a remarkable mobility, which initially determines a departure of R388 from the BP (as shown in Figure 6.10). After ~200 ns R388 and K366 are strongly drawn by the negative charge exposed by the nucleotides in the vicinity of the BP and establishing an interaction with C502 and A501, respectively (Figure 6.10). In this manner, the 3'-termini of the IL is tightly locked in the positive hinge, while Cwf19 and Spp42 pull the branch helix, triggering its displacement and unwinding (Figure 6.8 and 6.9).

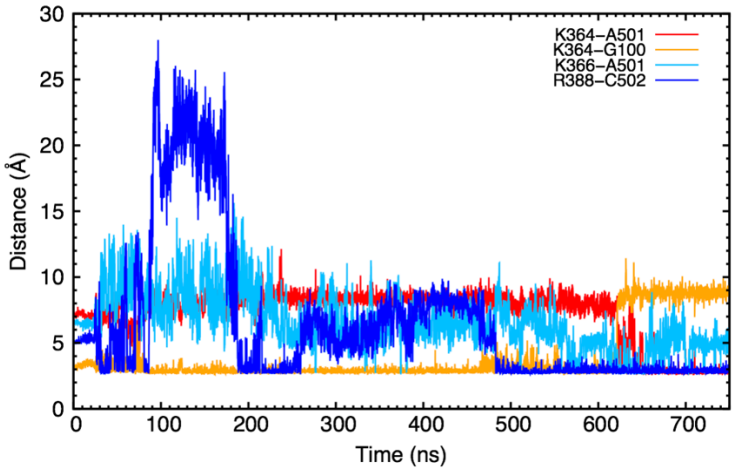


Figure 6.10. Evolution of selected distances (Å) between R388@NH2, K364/K366@NZ nitrogens and RNA phosphate oxygens of C502 and A501/G100 vs simulation time (ns) for replica #1.1, represented with different colors as in the top-right legend.

Consistently, the essential dynamics (i.e., the motion along the first principal component, PC1) reveals that Cwf19 and Spp42 concertedly push the IL/U2 helix (Figure 6.11), inducing its unwinding.

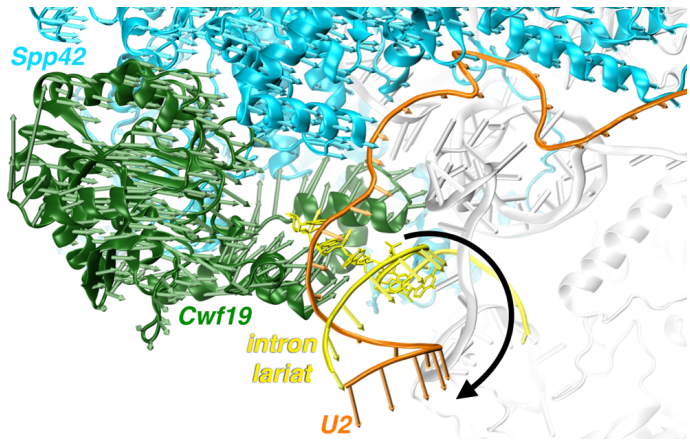


Figure 6.11. Essential dynamics (PC1) of Cwf19, Spp42, IL and U2 snRNA represented in green, light blue cartoon and yellow, orange tubes, respectively. The eigenvectors are shown on C-alpha and P atoms with colored arrows. The black arrow displays the unwinding motion.

Spp42: the orchestral director. Spp42 protein (Prp8 in *S. Cerevisiae*) is the largest protein of the spliceosome and the principal constituent of U5 snRNP (Figure 6.12), playing a key role in pre-mRNA splicing. In the model discussed in this thesis, we have considered 1486 residues, comprising the N-terminal (N-t), RT-palm/finger (RT), thumb/X (thumb) and finger domains spanning for more than 150 Å in its longest dimension.

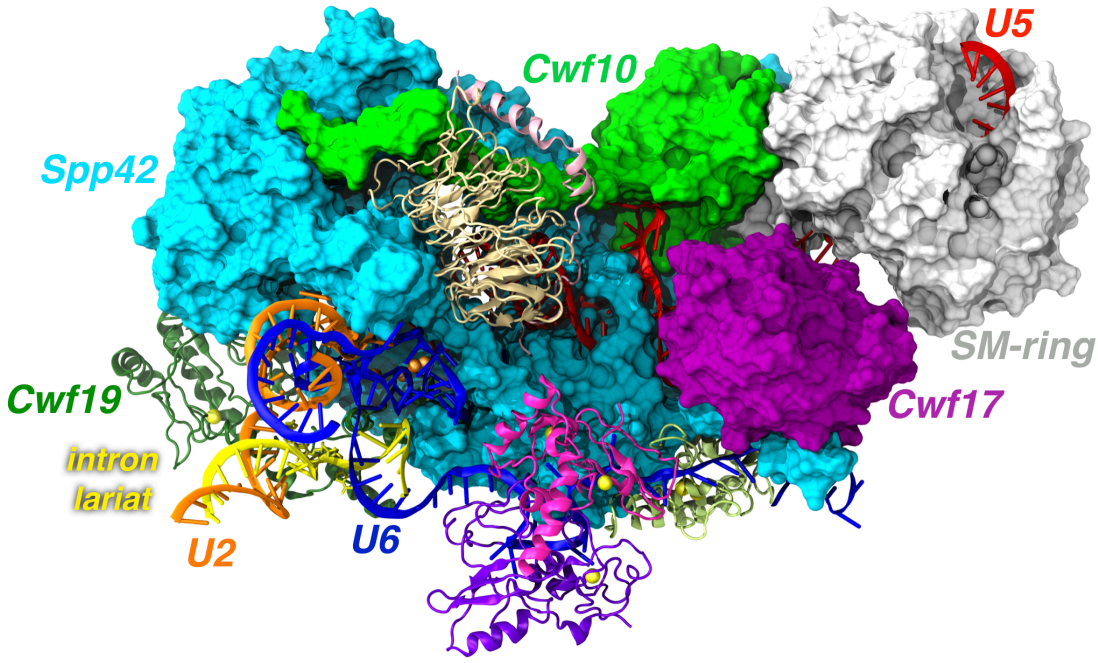


Figure 6.12. Spliceosome structure of model-1 investigated in the present study. Proteins are highlighted with different colors. U5 snRNP proteins are represented with solid surface, while the other components with new cartoon (proteins) and new ribbons (RNAs).

Spp42 is a huge protein, which embraces all other proteins of the system studied. As shown by the cross-correlation scores calculated also for its different domains (Figure 6.13), Spp42 modulates the movements of the surrounding proteins, while its domains are often characterized by internal anti-correlated motions. One example is constituted by the linker and RT domains with respect to the lasso and 26-313 sub-regions of the N-t domain. These parts of Spp42 are in fact in direct contact with two different proteins, Cwf19 and Cwf10, respectively. In particular, 26-313 and lasso move lockstep with the Cwf10 protein, while linker and RT with Cwf19.

Spp42 fulfills this task also through a peculiar architecture, which allows an exquisite connection with specific proteins. In particular, the N-terminal alpha helix (N-t Binding to Cwf17, hereafter named B-Cwf17) constitutes a protruding arm, which, binding to Cwf17, a U5 snRNP protein, regulates its motion. The interactions of the Spp42 “arm” with Cwf17 and of the Spp42 “lasso” with Cwf10, another

component of U5 snRNP, are maintained along the MD simulations, leading to the cooperative motions highlighted in our analysis.

Another key component of U5 snRNP is the U5 snRNA, the longest RNA included in our model, which spans the whole spliceosome core with its 105 nucleotides. Interestingly, its motion is remarkably coupled with the central regions of the Spp42 N-t domain, represented by residues 26-313 and 375-649. These portions entangle and anchor the U5 snRNA through an extended loop, thus explaining their lockstep dynamics.

Instead, U2 snRNA and IL motions are strikingly correlated with the RT domains and the B-RT (Binding to RT) of N-t. This confirms the role of Spp42 in the displacement of the IL/U2 branch helix in cooperation with Cwf19.

From this analysis appears that Spp42 plays the role of an orchestral director at the core of the spliceosome.

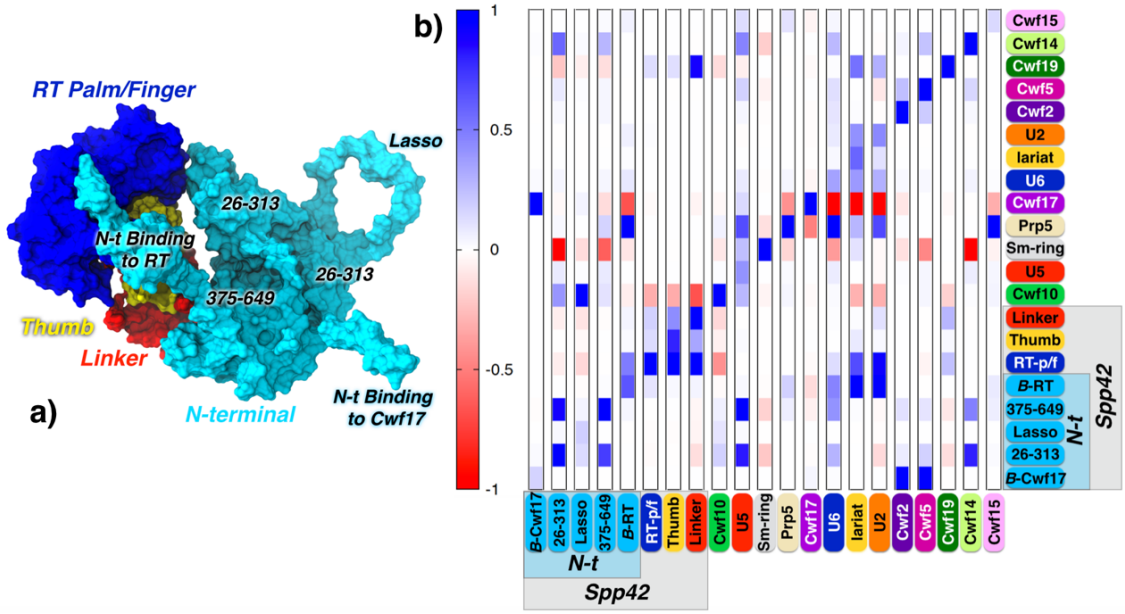


Figure 6.13. (a) Domains of Spp42 depicted with different colors with solid surface. (b) Spp42-detailed histogram reporting the normalized *intra-* and *inter*-correlation scores between the macromolecules of the system listed at the bottom and on the right and highlighted with different colors. The histogram is normalized per-column, therefore it is not symmetric, and must be read per-column.

Catalytic site. The spliceosome is a protein-directed ribozyme in which the catalytic site docks into a positively charged cavity formed by the N-t and thumb domains of Spp42. U6 and U2 snRNAs give rise to a peculiar triple-helix in which the catalytic triad of U6 (A47-G48-C49) base-pairs with U22-C21-G20 nucleotides of U2 and with A41-G40-U68 from U6 (Figure 6.14). The triplex is embedded into the

Spp42 cavity along with the two catalytic Mg^{2+} ions, M1 and M2, coordinated by the phosphate groups of A47 and G48 of the catalytic triad, and of U68 and G66. This structural rearrangement is strikingly conserved in the group II introns active site [32].

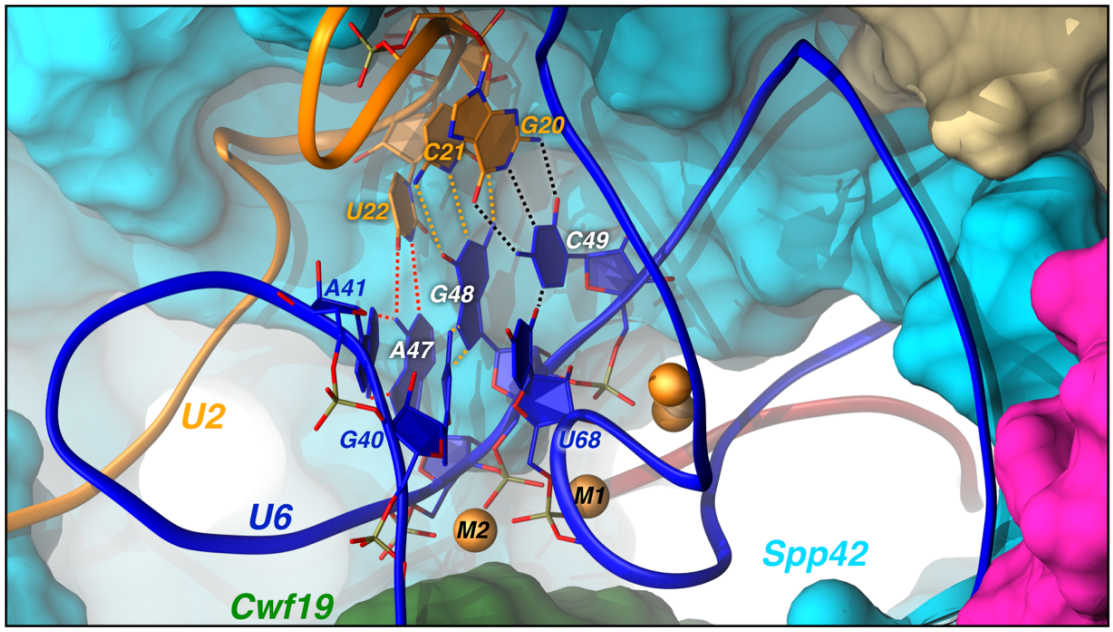


Figure 6.14. Snapshot representing the catalytic site of the spliceosome after 750 ns of MD simulations. U2 and U6 snRNAs are represented with orange and blue tubes, respectively. The nucleotides involved in the triple-helix are depicted with licorice representation. Magnesium ions are represented with orange spheres. The proteins forming the catalytic cavity are instead shown with a surface representation and highlighted with different colors.

The structural features of the triple-helix base pairs and the coordination geometry of the two metal ions are maintained in the MD simulations, as shown by the persistency of hydrogen bonds between the bases of the triple helix (Figures 6.14 and 6.15).

The observed structural stability of the triple helix also extends to the coordination geometry of the active site (Figure 6.16), thus confirming that the catalytic cavity defined by Spp42 is shaped *ad-hoc* to preserve a stable reactive center along the splicing cycle and that our model system provides a realistic picture of the central core of the spliceosome machinery. In group II introns this role is accomplished by the huge intron fragment embracing the two magnesium ions. Two additional Mg^{2+} ions in proximity of M1 and M2 were resolved in the cryo-EM structure [31, 32], possibly contributing to the stabilization of the phosphate-crowded region, and likely playing the role of the two K^{+} ions which form the metal cluster in the active site of the group IIC intron from *O. Iheyensis* [25].

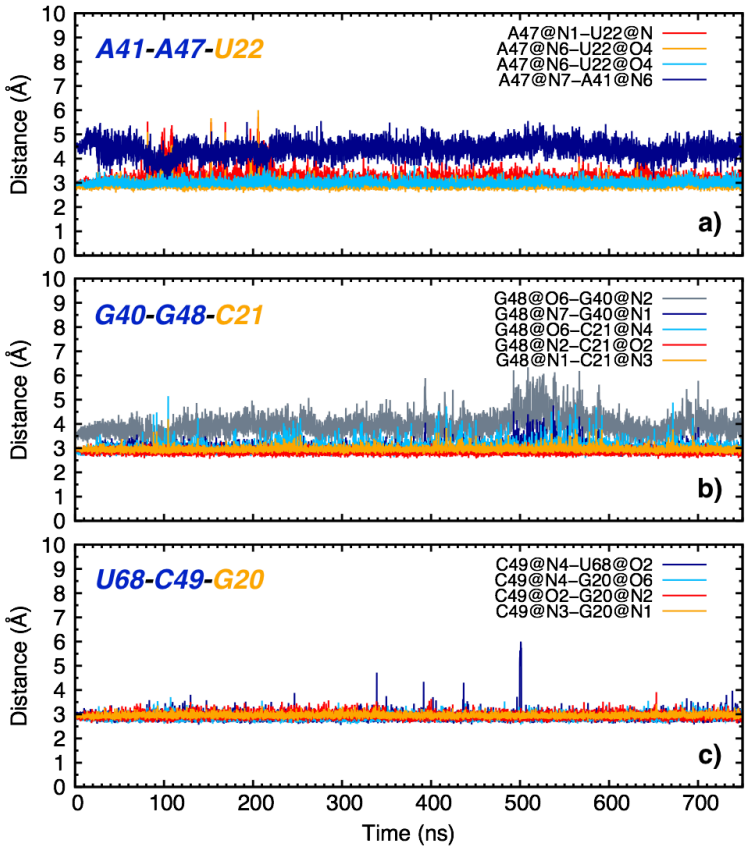


Figure 6.15. Time (ns) evolution distances (Å) between the heavy atoms involved in the hydrogen bonds between the bases of the triple helix, along the MD replica #1.1. Panels (a), (b) and (c) monitor the base pairs between A41-A47-U22, G40-G48-C21 and U68-C49-G20, respectively.

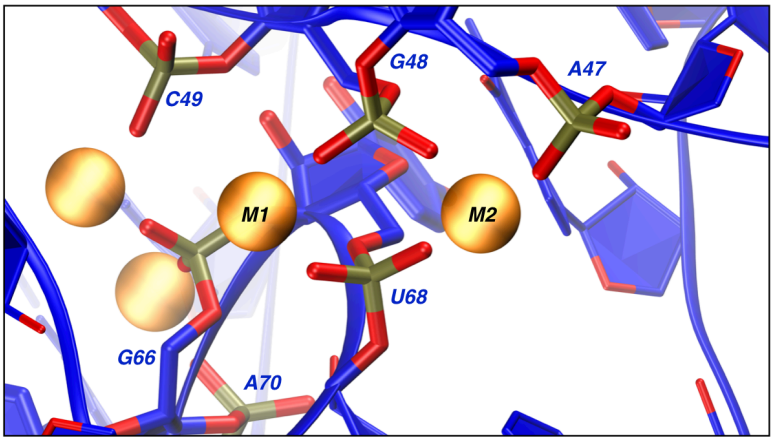


Figure 6.16. Snapshot representing the active site at 750 ns of MD simulations. Magnesium ions are depicted with orange sphere and the ones involved in the splicing reaction are indicated with M1 and M2. U6 snRNA is shown as blue ribbons, while the phosphate group directly involved in the coordination of the magnesium ions are highlighted with licorice representation.

6.5 Conclusions

Spliceosome is an exceptional dynamical machinery made of proteins and snRNAs, continuously undergoing conformational and compositional variations along the pre-mRNA splicing cycle. Here, through extensive MD simulations (more than 2 microseconds overall) we have investigated the dynamical behavior of the central core regions of the ILS complex, providing unprecedented insights on the spliceosome functional plasticity at atomistic level. Our work represents the first attempt to study the spliceosome structural, dynamical and functional properties with force field-based MD simulations. The essential dynamics of the spliceosome components strikingly revealed an electrostatically-driven unwinding motion of the IL/U2 branch helix, cooperatively promoted by Cwf19 and the RT domain of Spp42. In particular, it is tempting to speculate that the 3'-termini of the IL is locked in a positively charged tweezers defined by the guanidinium group of R388 and the ammonium groups of K364 and K366 (Cwf19). Once these residues electrostatically anchor the branching adenosine and the 3'-termini of the IL, the branch helix is unwound around this hinge point and displaced. In the present study, we have also disclosed how Spp42, the central protein of the spliceosome, accomplishes a central role directing the motion of many different spliceosome components. Its huge size and multipronged domains architecture allow a simultaneous interaction with different proteins, like Cwf10, Cwf17, Cwf19, regulating their functional motions. Moreover, we have confirmed the strong coupling between the dynamics of Spp42 and U5 snRNA, which is tightly anchored to the N-t domain.

Overall our work provides for the first time unique information on spliceosome dynamics, supplying atomistic details to the near-atomistic structural biology models. Our effort in elucidating the ILS complex dynamics dispenses an additional small piece of knowledge that may be relevant for an in-depth understanding of the pre-mRNA splicing. A detailed comprehension of the eukaryotic splicing mechanism has breakthrough implications for medicine and biotechnology, from gene functions prediction to revolutionary gene modulation tools to fight complex diseases such as cancer and neurodegeneration.

7 Conclusions and perspectives

Pre-mRNA splicing is a foremost process in gene expression, during which the coding fragments of the primary RNA transcript – the exons – are joined, whereas the intervening silent pieces – the introns – are removed. In eukaryotes, splicing is governed by is a sophisticated ribonucleoprotein machinery, the spliceosome, which mediates two subsequent transesterification reactions necessary to obtain a mature mRNA filament. In the latest years, crystallographic and cryo-EM structural models of both group II introns and spliceosome provided a striking near-atomistic characterization of these biomolecules at different stages of the catalytic cycle, offering a unique possibility to disentangle hidden mysteries – not fully decipherable at experimental level – by means of computer simulations.

In this thesis, I have traced the evolutionary link between group II introns and the spliceosome from a computational point of view. I first investigated the pre-mRNA splicing mechanism self-catalyzed by group II introns, by using classical MD and first-principles QM/MM methods (Chapter 4). The results unveiled an RNA-adapted Steitz and Steitz's two-metal-ion mechanism for the first step of hydrolytic splicing, proceeding by a dissociative S_N1-like catalysis. At odds with protein enzymes undergoing a two-metal-aided catalysis, like RNase H, the active site of group II introns is entirely composed by the negatively charged phosphate groups of the RNA sugar-phosphate backbone, which constitute a more rigid and less specialized scaffold with respect to the more flexible and highly functionalized side chains of the amino acids. As such, group II introns have developed a mechanism which apparently differs from the canonical Steitz and Steitz's proposal. While one Mg^{2+} promotes the cleavage by stabilizing the negative charge on the leaving group, the other contributes to the catalysis by proper orienting the nucleophile and activating the electrophile, i.e. the

phosphorus atom of the scissile phosphate, rather than the nucleophile. The attacking water, even if a poor nucleophile, can then attack the metaphosphate species and easily release the proton to the bulk water, which is proposed as final acceptor, also due to the absence of nucleobases with unperturbed pK_as around 7. However, here, only the first step of the catalysis has been investigated as the X-ray model corresponding to second step pre-catalytic adduct has not been characterized yet. Moreover, my study depicts an RNA-adapted Steitz and Steitz's two-Mg^{2+}-ion mechanism only for the hydrolytic pathway. If this mechanism applies also to the branching pathway and to different group II introns (or the spliceosome) needs still to be verified. As such, future work in the field strictly depends on the availability of new G2IRs crystal structures. In particular, a pre-catalytic model containing D6 and the branching adenosine would finally allow a QM/MM study of the branching pathway, which may confirm our mechanistic hypothesis. Another challenge is represented by the characterization of a conformational change which is supposed to take place between the first and second step of catalysis. Marcia et al. [25] have hypothesized that a large structural rearrangement involving the catalytic core between the two steps should prepare the ribozyme to undergo the second step of catalysis. Microsecond time scale MD simulations as well as enhanced sampling techniques may shed light on this critical event.

Magnesium ions represent the connection between group II introns and the spliceosome as they are the fundamental cofactors of pre-mRNA splicing. Mg^{2+} ions screen the negative charge of nucleic acids and contribute to the stability and folding of RNAs. As such, unravelling Mg^{2+}/RNA interplay at atomistic level is of utmost relevance. Many efforts have been done in the development of reliable force fields (FFs) to describe Mg^{2+} in MD simulations. In Chapter 5, using a group II intron ribozyme as a prototype model of large RNA molecules, I have benchmarked the performance of five different Mg^{2+} parametrizations in multiple MD replicas under different ionic concentrations. I was able to show that all the Mg^{2+} models are plagued by an inadequate description of Mg^{2+}–N_b contacts, being overall underestimated. The FFs due to Åqvist and Allnér, even if affected by a general overestimation of Mg^{2+}–O_b contacts, better accounted for the experimental RNA ligands distribution observed in the PDB. Remarkably, the dummy cation model due to Saxena produced good results in the modeling of binuclear sites, such as the catalytic site of group II introns. In order to dissect the source of force fields failures and categorize the non-trivial electronic effects taking place between Mg^{2+} and its ligands, such as charge transfer and polarization, I have studied at DFT level of theory 16 recurrent Mg^{2+}–RNA binding motifs, comprising *inner-sphere* and *outer-sphere* sites. These last calculations have surprisingly disclosed that the electronic properties of Mg^{2+} remain roughly unvaried in

distinct RNA coordination environment. Overall, this study on one side provides useful guidelines for a conscious use of available Mg^{2+} force fields for MD simulations along with a list of major sources of drawbacks which may help in the development of future Mg^{2+} parametrizations. This still represents a challenging task. In fact, even a flawless Mg^{2+} force field would be always plagued by the inaccuracies arising from the RNA FFs. On the other hand, DFT calculations have revealed that *inner-sphere* water molecules have a peculiar buffering role as charge donors. The description of this scenario at classical level appears to be difficult to reproduce with non-polarizable FFs. As such, a possible solution is represented by a site-specific re-parametrization of the RNA ligands FFs rather than Mg^{2+} FFs, since the electronic properties of Mg^{2+} are invariant, whereas those of the ligands vary according to coordination site type.

The final part of this computational evolutionary journey has been fueled by the burst of spliceosome structural biology in the last two years. In 2015, the first near-atomistic cryo-EM reconstruction of the *S. Pombe* spliceosome offered an exquisite model of the ILS spliceosomal complex at 3.6 Å resolution, that I used for a force field-based computational study discussed in Chapter 6. Up to now, this and all the other tens of cryo-EM structures solved in the following two years either did not capture the proper conformational state to study the chemical mechanism of splicing or were affected by resolution limits. Therefore, my investigation was limited to unveil the essential dynamics of the spliceosome by performing multiple microsecond time scale MD simulations, with a particular focus on Spp42 protein and on the catalytic site. PCA and APBS analyses showed an electrostatically-driven concerted action of Cwf19 and Spp42 in the displacement of the branch helix formed by U2 snRNA and the intron lariat. The so far poorly characterized Cwf19 splicing factor (*S. Pombe*), which was tentatively hypothesized to play a role in the IL/U2 branch helix displacing, it is here found to be effectively involved in this operation together with the RT domain of Spp42. This represents the most important finding of this final study, which also assesses the stability of the active site even at a final stage of splicing. The highly conserved catalytic architecture formed by the U6/U2 snRNA triple-helix motif and the two Mg^{2+} ions is tightly hold in a positive cavity formed by Spp42 and prevented from any conformational rearrangement. This appears to be in line with the hypothesis that a compact and rigid catalytic scaffold should adapt to host and process introns of very different lengths and sequences. Overall, this study represents the first attempt to provide atomistic details on spliceosome dynamics, unveiling the principal motions of the ILS complex components. More recent cryo-EM structures can already be used to investigate, with classical simulations, different spliceosome complexes at earlier stages of the splicing cycle, complementing the mechanistic details provided by my study. The spliceosome structural biology revolution is just at the beginning and future

directions are focused on a better characterization of the human spliceosome. This scenario is tantalizing as it lays the foundation for the addition of a further stop to this incredible evolutionary computational journey: unraveling the atomistic details of human splicing.

A1 Appendix 1

Validation of the starting Ca^{2+}-inhibited structure as reactant. As mentioned in the paragraph 4.3, our simulations were started from the crystal structure corresponding to the pre-catalytic state inhibited by the presence of Ca^{2+} ions (pdb id: 4FAQ), which were replaced with Mg^{2+} ions to build the initial adduct. However, a superposition between this structure and that of the product crystallized in the presence of the reactive Mg^{2+} ions (pdb id: 4FAR) demonstrates that not only the overall fold of the ribozyme is conserved, but also the local active site geometry does not undergo relevant structural modifications (Figure A1.1).

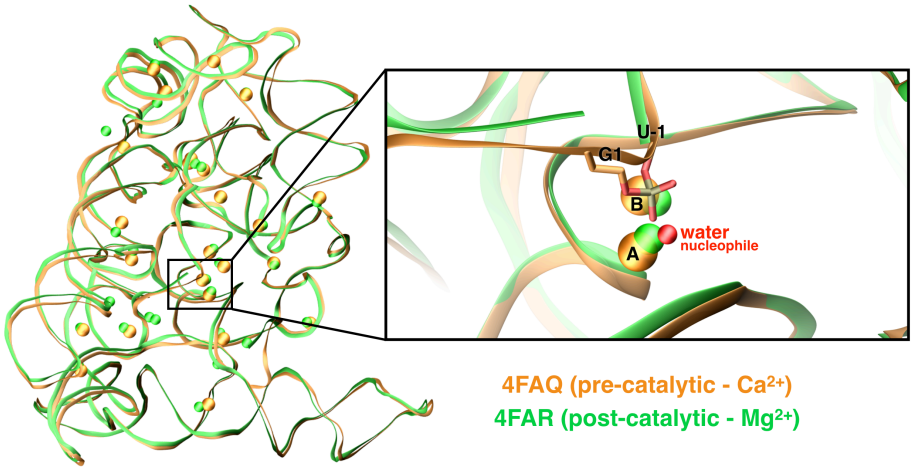


Figure A.1.1. Superposition of the Ca^{2+}-inhibited pre-catalytic crystal structure (4FAQ) and the first-step product solved in the presence of Mg^{2+} ions (4FAR) (green). The overall fold and the position of the Ca^{2+} and Mg^{2+} ions are conserved.

Moreover, the superposition between the reactant adduct obtained after classical and QM/MM MD equilibration and the Ca^{2+}-inhibited crystallized structure (pdb id: 4FAQ) shows how the active site is fully conserved (Figure A1.2). Most importantly, these structural features are conserved also for the product that we have obtained from the Blue-moon ensemble QM/MM MD simulations as it is highlighted by the comparison with the product crystallized in the presence of Mg^{2+} (Figure A1.3).

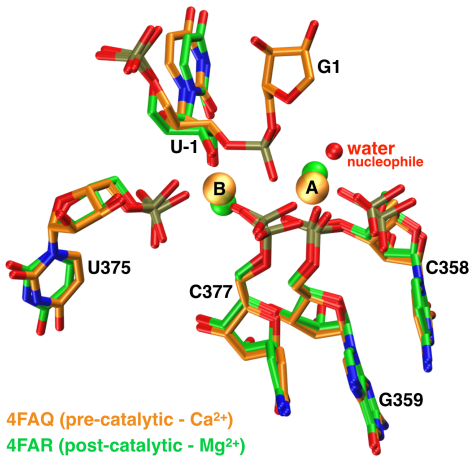


Figure A1.2. Superposition of the active site residues as in the Ca^{2+}-inhibited pre-catalytic crystal structure (4FAQ) and the first step product solved in the presence of Mg^{2+} ions (4FAR) (green). Please note that the nucleobases of G1 was not solved in the 4FAQ X-ray structure.

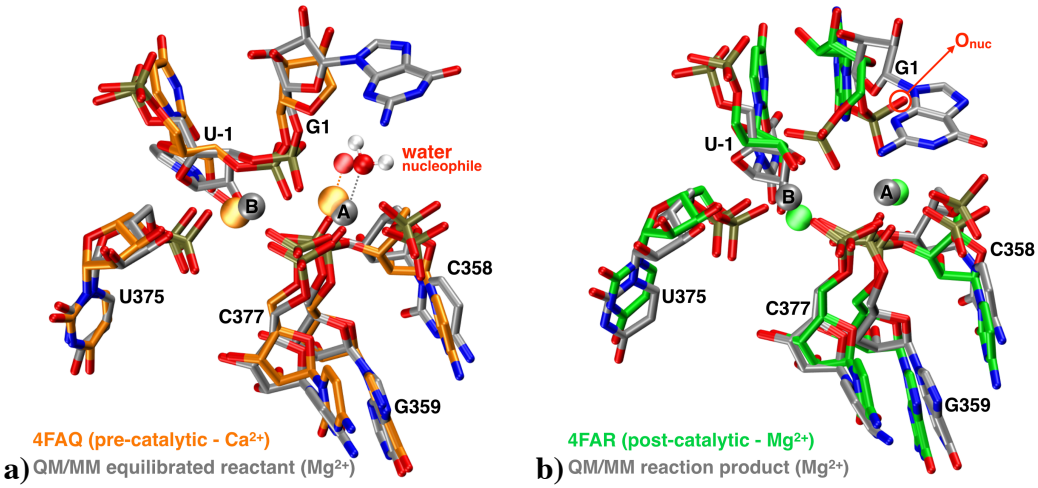


Figure A1.3. Superposition of the active site residues as in the Ca^{2+}-inhibited pre-catalytic crystal structure (4FAQ, orange, a) and post-catalytic crystal (4FAR, green, b) with the structure of the QM/MM MD equilibrated reactant adduct (a) and QM/MM MD product (b) (grey), respectively. Please note that the nucleobase of G1 was not solved in the 4FAQ X-ray.

Alternative models.

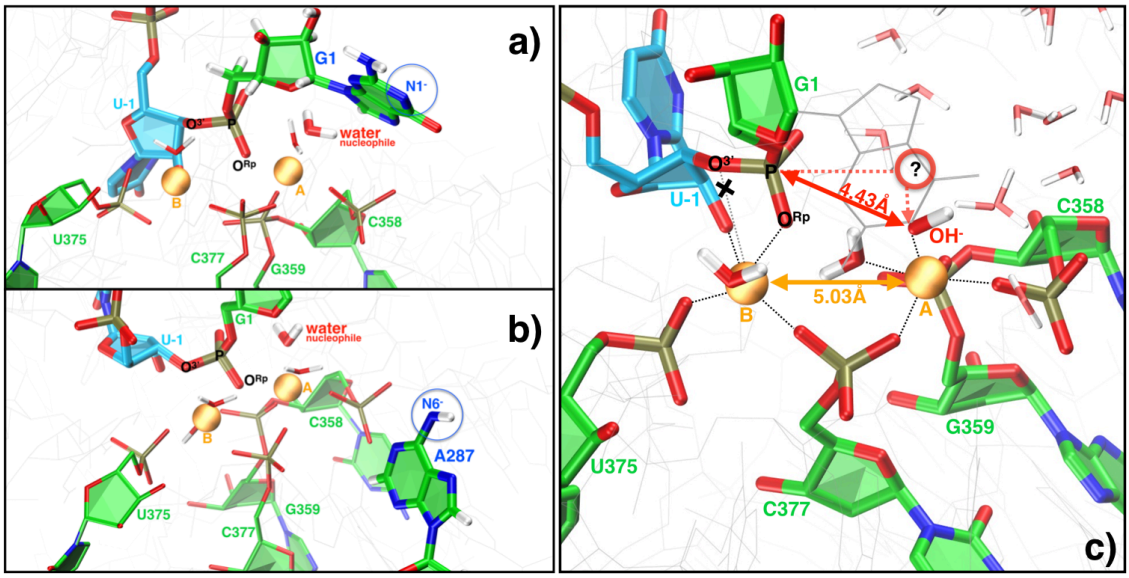


Figure A1.4. (a, b) Representative snapshots of the (a) N1-@G1 and (b) N6-@A287 models after 50 ns of classical MD simulation. Mg^{2+} ions (A and B) are represented with orange spheres, intron RNA residues of the catalytic site are depicted with green sticks, 5'-exon RNA residue (U-1) with blue sticks. Both the (a) N1-@G1 and (b) N6-@A287 models did not produce any H-bond network capable of promoting the deprotonation of the nucleophilic water.

(c) Representative snapshot from QM/MM MD simulations on the OH^- model (i.e., the fourth model). A pre-activated nucleophile (OH^-) has been considered in place of the originally crystallized water molecule. Mg^{2+} ions (A and B) are represented with orange spheres, intron RNA residues of the catalytic site are depicted with green sticks, 5'-exon RNA residue (U-1) with blue sticks. The question mark represents the optimal position of the nucleophile for the attack on the scissile phosphate as occupied by the water molecule in the original model. The hydroxyl ion leads to an unstable reactant as it falls apart (red arrow) from the scissile phosphate (P@G1) as a result of a stronger coordination by MgA. Moreover, the distance between the two Mg^{2+} ions increases up to ~5 Å, distorting the catalytic site (orange arrow). It is also remarkable as MgB loses the coordination of the leaving group ($O^{3'}$@U-1), as indicated by a black cross over the dashed coordination line.

Free energy calculations and transition state validation. We remark that the choice of a mono-dimensional RC may have a small impact on the calculated free energy barrier, but the same reaction coordinate has been successfully used in very similar reactions [20]. Moreover, the use of a simple inappropriate reaction coordinate typically results in overestimation of free energy barriers and large hysteresis. Instead, in our simulations no large hysteresis is observed between the forward and the

backward processes and the barrier appears to be slightly underestimated due to the BLYP exchange correlation functional.

In Figure A1.5, a magnified view of the reaction free energy profile shows the red and the black line corresponding to the forward and backward pathway, respectively, as investigated with BLYP exchange correlation functional. The blue line represents the curve obtained integrating the average forces resulting from the sampling in the 0.7 Å, 0.9 Å, 1.1 Å windows along the RC with B3LYP exchange correlation functional. We remark that the free energy profile shown for B3LYP has been obtained by integrating the average constraint forces extracted from BLYP QM/MM MD simulations within RC = -1.5 Å to 0.6 Å and the forces from B3LYP in the RC = 0.7 Å, 0.9 Å, 1.1 Å. Hence, it is not possible to estimate the value of the free energy barrier for the B3LYP QM/MM MD simulations, but only to have a picture of its shape at and after the transition state.

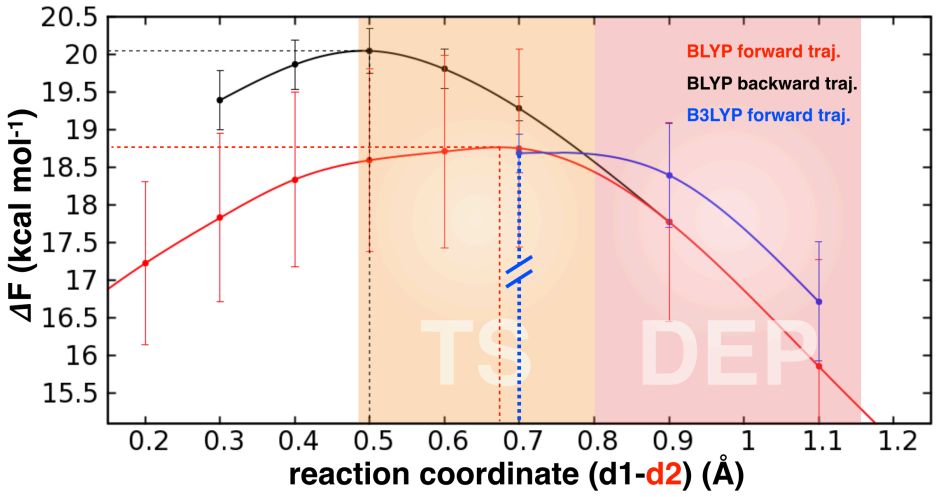


Figure A1.5. Magnified view of the free energy profile showing the transition state (TS) and deprotonation (DEP) regions. The BLYP forward profile is represented with a red line, the BLYP backward path with a black line, while the B3LYP curve obtained for the points at reaction coordinate RC = (0.7 Å, 0.9 Å, 1.1 Å) is depicted with a blue line. Discontinuous dotted blue line remarks that no information can be given on the height of the free energy barrier from B3LYP calculations. The RC is the difference between the breaking bond (d1) and the forming bond (d2). Transition state region is highlighted in orange, while the deprotonation window in red.

With respect to the BLYP forward trajectory, the overall free energy barrier is 18.8 ± 1.2 kcal/mol and the TS falls at RC ~0.7 Å, while for the BLYP backward trajectory is 20 ± 0.3 kcal/mol with the peak of the curve at RC = 0.5 Å. Combining

the forward and the backward pathways the total Helmholtz's free energy ($\Delta F^{\#}$) for the reaction becomes 18.8 ± 1.5 kcal/mol.

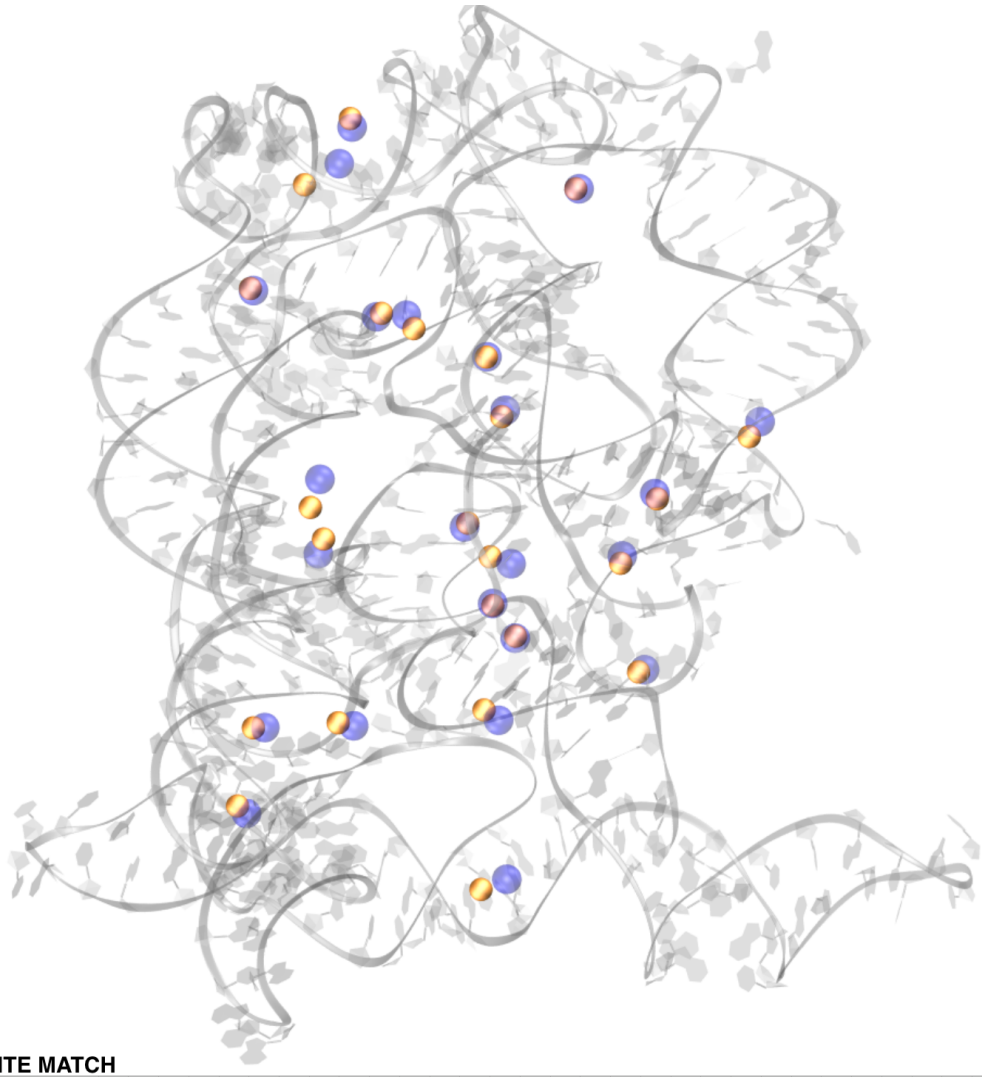
Interestingly, the curve resulting from the B3LYP QM/MM MD simulations confirms that no extra energy barrier is due to the proton transfer. In particular, at RC = 0.7 Å the nucleophilic water is still protonated and the average force is at zero (corresponding to the TS) with a flat free energy profile. At RC = (0.9 Å, 1.1 Å) the deprotonation takes place and the average forces exhibit the same trend as in the BLYP QM/MM MD simulations, with a decreasing free energy. This clearly suggests that even with the B3LYP exchange correlation functional no additional energy barrier is associated with the proton abstraction. It is rather plausible to identify an extended transition state window ranging from ~0.5 Å to ~0.8 Å and highlighted with orange color in Figure A1.5. In the deprotonation window, highlighted with red color in Figure A1.5, the water molecule loses its proton without any additional contribution to the energy barrier because of the dissociative nature of the reaction mechanism. Indeed, the cleavage of the $O^{3'}$@U-1–P@G1 bond takes place before the formation of the new O_{nuc}–P@G1 bond, thus leading to the activation of the electrophile (the phosphorus atom of the scissile phosphate, i.e. P@G1). As a consequence of this event the nucleophilic water can attack the activated phosphorus without being previously deprotonated. When the new bond (i.e., O_{nuc}–P@G1) is almost formed, the nucleophilic water easily releases its proton to the bulk (RC = 0.9-1.1 Å), leading to the complete formation of the O_{nuc}–P@G1 bond.

As a further check of the free energy barrier we performed a geometry optimization and vibrational frequency analysis on reduced model systems of the catalytic site (93 QM atoms) for both reactant and transition state. These simulations were carried out with Gaussian09 program using the 6-31g* basis set and PCM as implicit solvent model. Both for the reactant and transition state models the calculations were performed at DFT-BLYP and DFT-B3LYP level of theory in order to verify the influence of the exchange correlation functionals on the free energy barriers. The $\Delta G^{\#}$ obtained from these calculations were of ~19.0 and ~25.0 kcal/mol for BLYP and B3LYP, respectively. With these additional calculations, we verified the influence of the exchange correlation functional on the barrier. The value obtained on the reduced model with the B3LYP functional suggests that the barrier is probably underestimated by using BLYP. We remark that both barriers are however consistent with the available experimental kinetic data reporting a catalytic rate of 0.011 min^{-1} at the same ionic strength reproduced in our simulations, corresponding to a $\Delta G^{\#}$ of ~23 kcal/mol [25].

A2 Appendix 2

Radial distribution function. Analysis of $g_{Mg\text{-}X}(r)$ (Figure A2.3 and Table S2) displays a contraction of the Mg^{2+}-X bonds for the Åqvist parametrization in which the first peak is split in two: one corresponding to the Mg^{2+}-O_{ph} coordination distance, and the other to the remaining coordinating ligands lying at slightly higher *r* value. The Saxena model reveals a similar pattern of the *g(r)*, shifted toward larger values of the radial distances. In the Allnér and Li models the *g(r)* profile exhibits a single well defined first peak, lying at slightly larger values than the one obtained with the Åqvist FF. Also for these parametrizations, the distances between Mg^{2+} and its ligands are underestimated in all cases, while the Mg^{2+}-N_b distance is again overestimated in Allnér. For the Oelschlaeger parametrization, a different *g(r)* is observed, showing a broad first peak shifted toward larger values of *r* with respect to all other models. This corresponds to unrealistically large distances between Mg^{2+} and its ligands.

Additional Figures.



SITE MATCH

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
PDB 4FAQ (Ca^{2+})	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	-						
PDB 4FAR (Mg^{2+})	431	401	405	407	408	409	410	411	412	413	414	416	425	419	420	421	422	423	406	427	428	429	417	-	-	-	-	-	-	-

Figure A2.1. Superposition between the 4FAQ and 4FAR X-ray structures of group II intron ribozyme (G2IR) [25], which have been crystallized in the presence of Ca^{2+} and Mg^{2+} ions, respectively. Ca^{2+} (4FAQ) and Mg^{2+} (4FAR) ions are shown as blue and orange van der Waals spheres. The overall architecture of G2IR is represented as gray ribbons. In the table, the ions (i.e., Ca^{2+} and Mg^{2+}) which have a site match among the two structures are reported with the numbering scheme as in the two X-ray structures (blue and orange lines). The sign "-" refers to the ions which do not have a match.

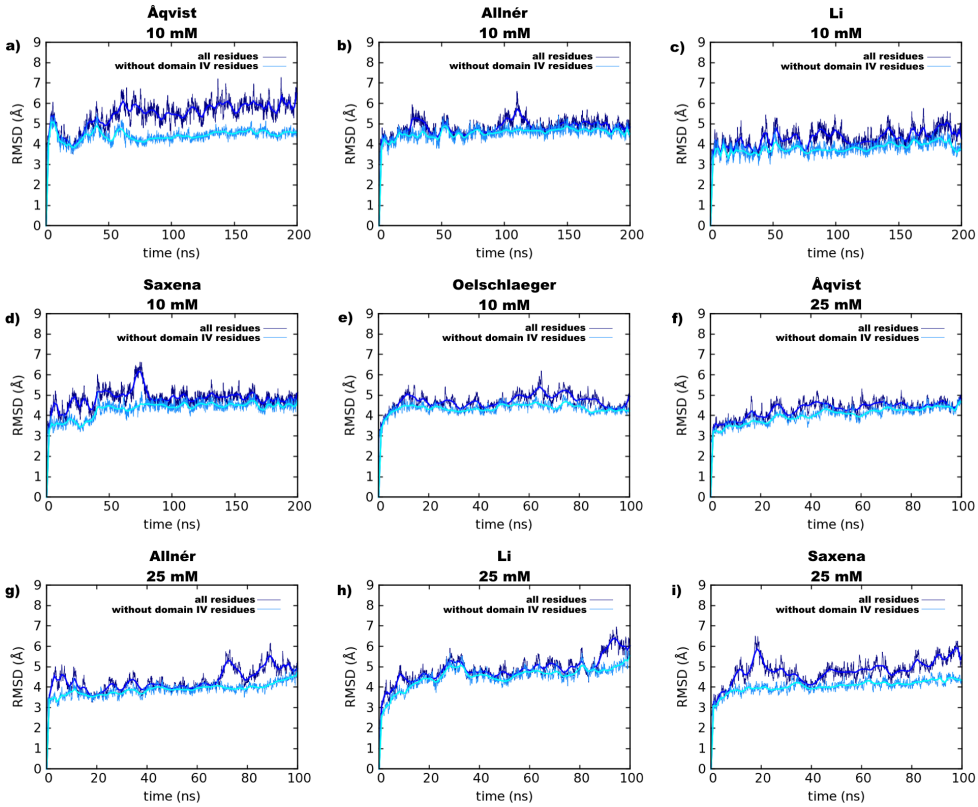


Figure A2.2. RMSD (Å) vs simulation time (ns) of the catalytically competent form of G2IR with empirical Mg^{2+} force field parameters according to the Åqvist, Allnér, Li, Saxena and Oelschlaeger models in (a), (b), (c), (d) and (e), respectively, at [Mg^{2+}] = 10 mM, and to the Åqvist, Allnér, Li and Saxena models in (f), (g), (h) and (i), respectively, at [Mg^{2+}] = 25 mM. The RMSD without residues of D-IV is reported with the light blue line. This part of the ribozyme is highly solvent exposed and is responsible in some case of the oscillatory behavior.

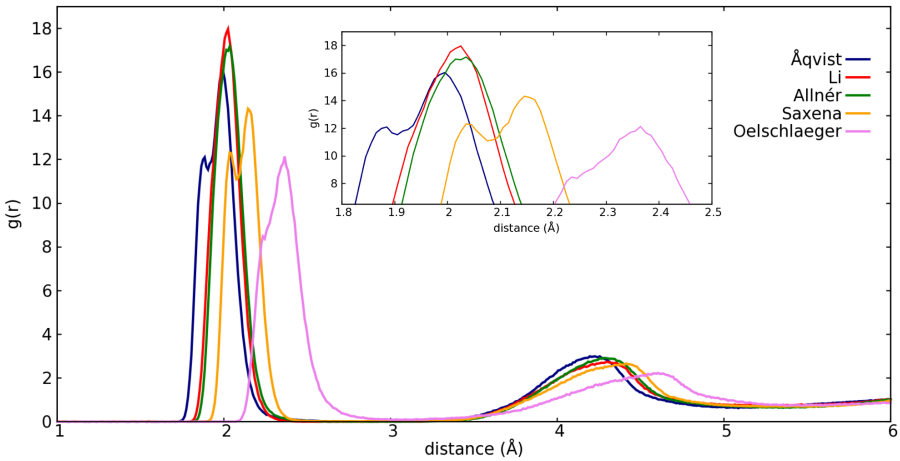


Figure A2.3. Radial distribution function (g(r)) for Mg2+ ions within 6 Å of their coordination sphere (including the Oph, Nb, Ob, Os and Ow ligands), calculated over MD simulations of G2IR ribozyme, performed with the Åqvist (blue), Allnér (green), Li (red), Saxena (orange) and Oelschlaeger (magenta) parameters at [Mg2+] = 10 mM.

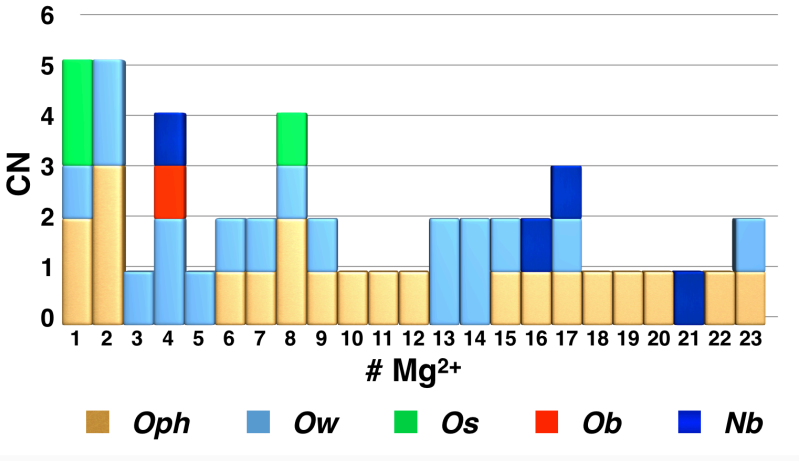


Figure A2.4. Histograms showing the ligand composition of the coordination sphere of each Mg^{2+} site in G2IR as found in the 4FAR [25] X-ray structure. In the *x-axis*, binding sites are renumbered according to the numbering of our simulated structure (4FAQ). Namely only the Mg^{2+} sites of 4FAR matching with Ca^{2+} sites of 4FAQ structure are reported, with internal numeration from #1 to #23 according to Figure S1. The *y-axis* reports the coordination number (*CN*) for each Mg^{2+} binding site. The limited resolution of the X-ray structure (i.e., 2.8 Å) impedes the complete dissection of the coordination sites.

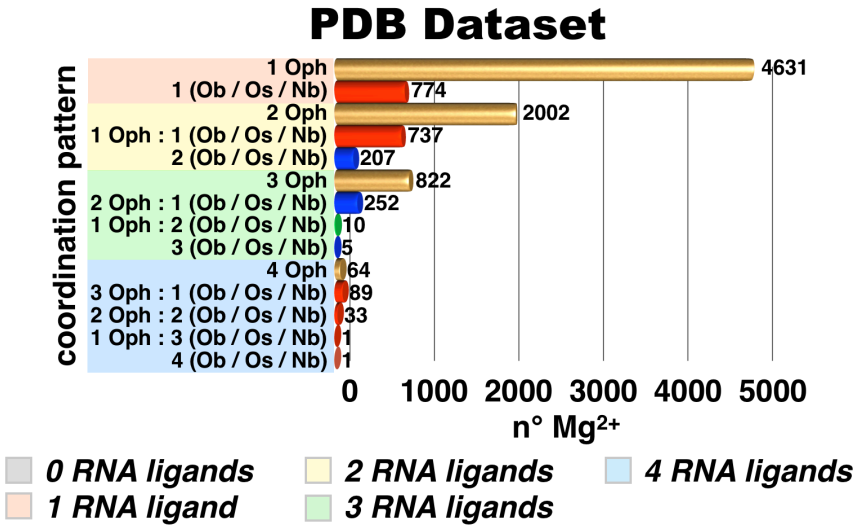


Figure A2.5. Histogram showing the population of Mg^{2+}-RNA coordination patterns extracted from Zheng et al. [187] for the whole PDB dataset (*y-axis*). The number of RNA ligands (from 0 to 4) is highlighted with different colors, as specified in the bottom legend. The *x-axis* reports the number of Mg^{2+} ions (i.e., population) having a specific coordination pattern. Bars of different colors are used to identify coordination patterns characterized by the presence of different ligand types: only O_{ph} (gold) or at least one O_b (red, if in majority with respect to sites containing at least one O_s or N_b), O_s (green, if in majority) or N_b (blue, if in majority) ligand.

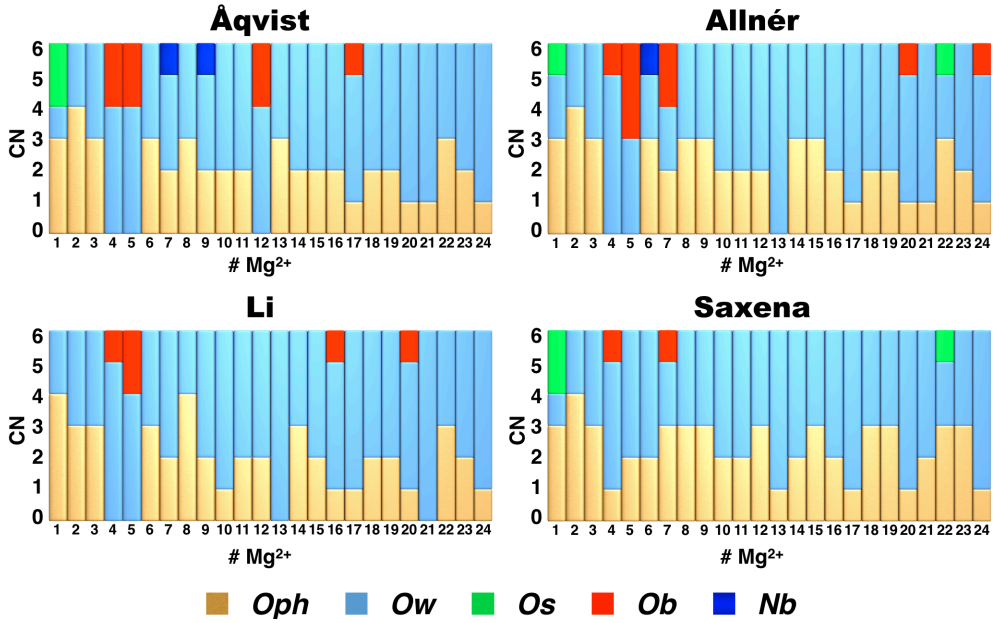


Figure A2.6. Histograms showing the composition of the first coordination sphere of each of the 24 Mg^{2+} sites in G2IR as determined from MD performed with the Åqvist, Allnér, Li and Saxena parametrizations at [Mg^{2+}] = 10 mM. The *x-axis* reports the 24 Mg^{2+} sites in G2IR, including sites 1 and 2, which are constitutive of the catalytic binuclear site. The *y-axis* reports the *CN* for the Mg^{2+} coordination sphere. Donor atom types are identified with different colors as specified in the bottom legend.

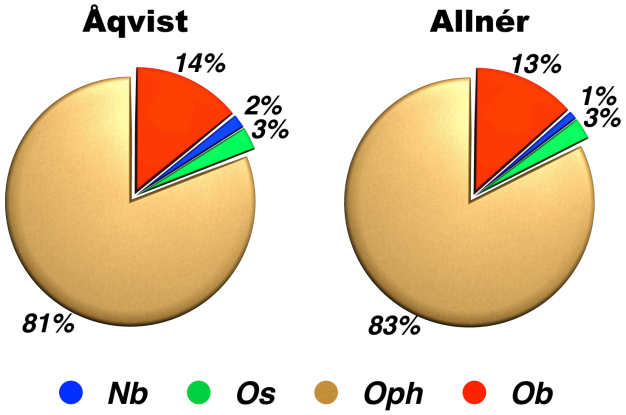


Figure A2.7. Statistical distribution of the Mg^{2+}–X ligands (where X = O_{ph}, N_b, O_b and O_s) in the Mg^{2+} *inner-sphere* (expressed as percentages), obtained from the combination of three different replicas of G2IR, performed with the Åqvist and Allnér parametrizations at [Mg^{2+}] = 10 mM. Donor atom types are identified with different colors as specified in the bottom legend.

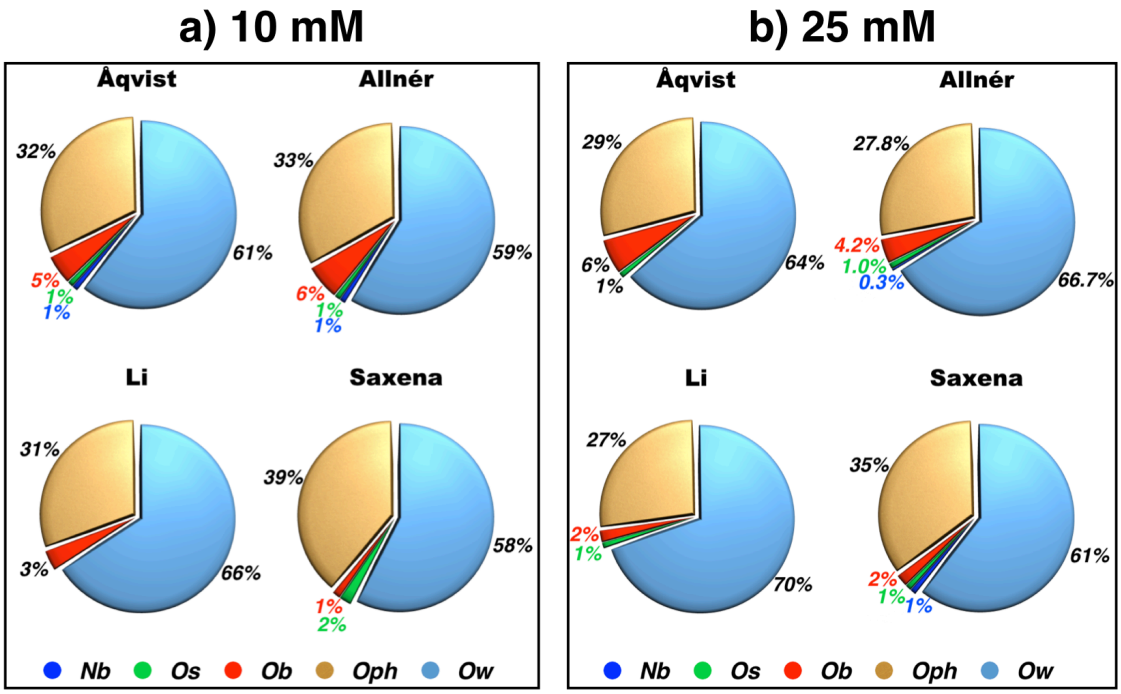


Figure A2.8. a) Statistical distribution of the Mg^{2+}–X ligands (where X = O_{ph}, N_b, O_b, O_s and O_w) in the Mg^{2+} *inner-sphere* (expressed as percentages) as calculated from MD simulations of G2IR performed with the Åqvist, Allnér, Li and Saxena parametrizations at [Mg^{2+}] = 10 mM. Donor atom types are identified with different colors as specified in the bottom legend. Remarkably, the Mg^{2+}–X statistical distribution is almost identical in the simulations performed with the Åqvist and Allnér parametrizations, while the occurrence of water ligands increases in Li and decreases in the Saxena model.

b) Statistical distribution of the Mg^{2+}–X ligands (where X = O_{ph}, N_b, O_b, O_s and O_w) in the Mg^{2+} *inner-sphere* (expressed as percentages) as calculated from MD simulations of G2IR performed with the Åqvist, Allnér, Li and Saxena parametrizations at [Mg^{2+}] = 25 mM. Donor atom types are identified with different colors as specified in the bottom legend.

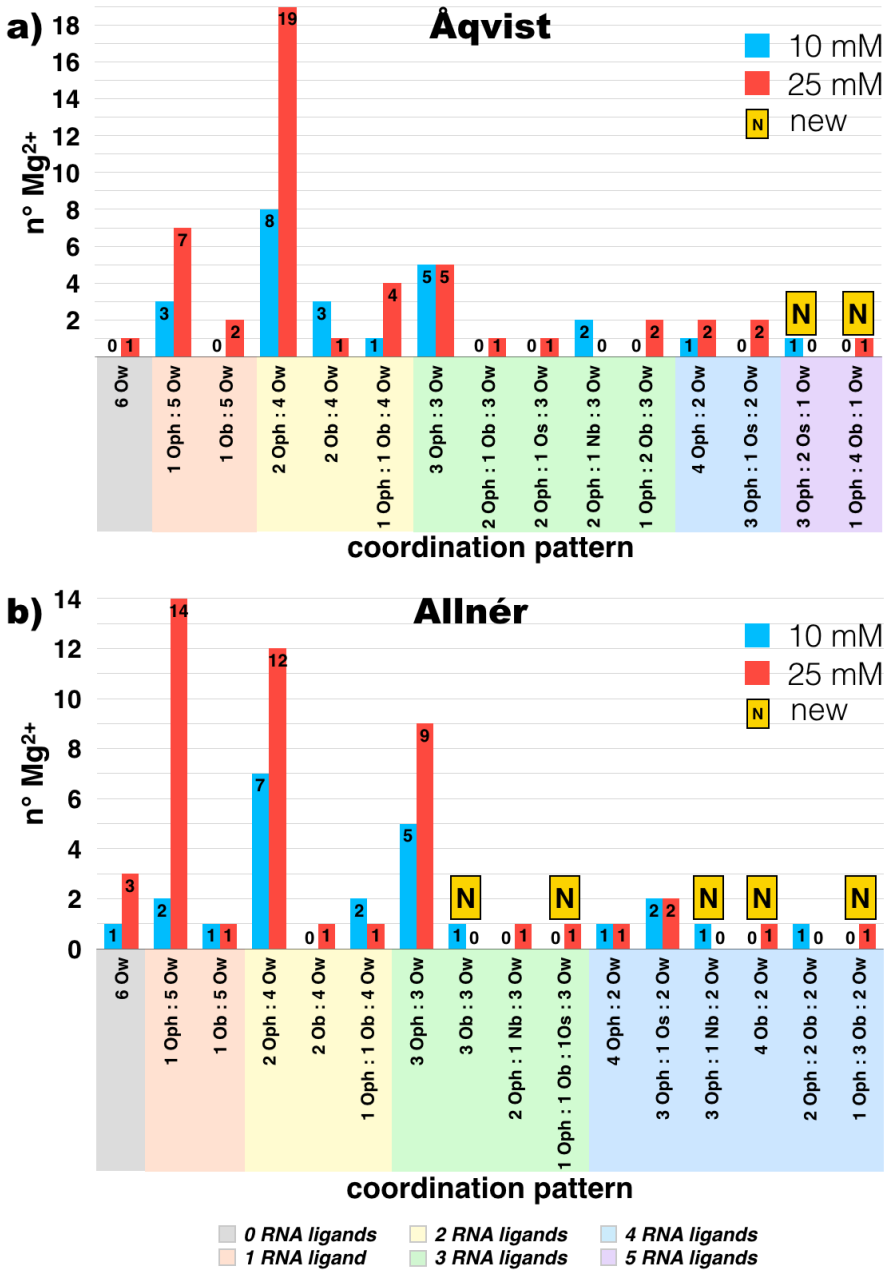


Figure A2.9. Histograms determining the population of the Mg^{2+}–RNA CPs, observed during MD simulations of G2IR, performed with the Åqvist (a) and Allnér (b) models at [Mg^{2+}] = 10 mM (blue bars) and 25 mM (red bars). The *x-axis* reports the Mg^{2+} CPs identified from MD simulations. The number of RNA ligands (from 0 to 5) is highlighted with different colors as shown in the bottom legend. The *y-axis* reports the number of Mg^{2+} ions (i.e., population) having a specific CP. The label N refers to binding sites not observed in the PDB dataset.

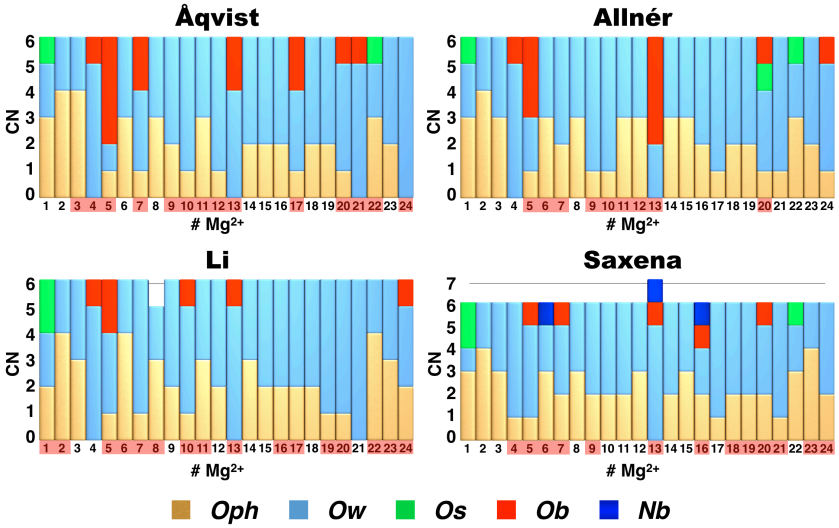


Figure A2.10. Histograms showing the composition of the coordination sphere of each of the first 24 out of 48 Mg^{2+} sites in G2IR from MD performed with the Åqvist, Allnér, Li and Saxena ff at [Mg^{2+}] = 25 mM. The *x-axis* reports the 24 Mg^{2+} sites in G2IR, including sites 1 and 2, which are constitutive of the catalytic site. The sites changing their ligand composition with respect to the simulation at [Mg^{2+}] = 10 mM are highlighted in red. Donor atom types are identified with different colors (bottom legend). The *y-axis* reports the *CN* for the Mg^{2+} coordination sphere. 63% of the sites exhibits the same composition at the two Mg^{2+} concentrations with the Allnér model, followed by Saxena (46 %), Åqvist (42%) and Li (33%).

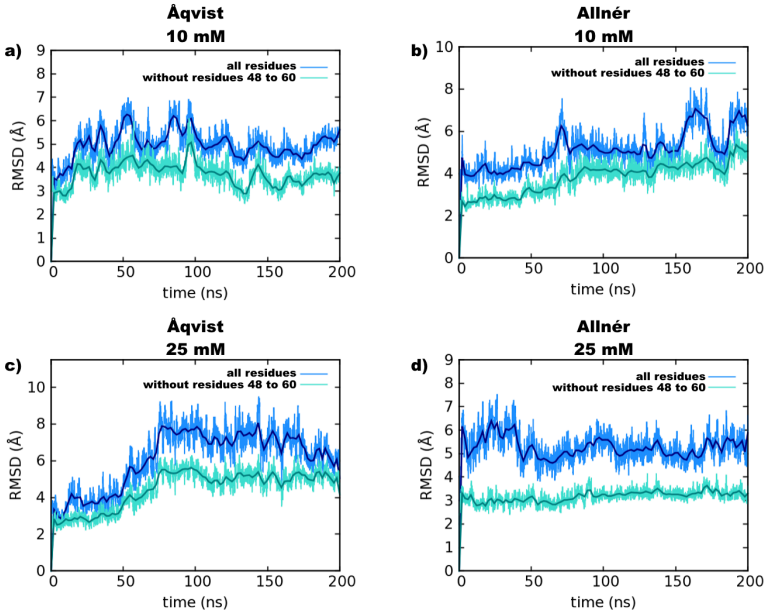


Figure A2.11. RMSD (Å) vs simulation time (ns) of the hepatitis delta virus (HDV) ribozyme with empirical Mg^{2+} ff parameters according to the Åqvist and Allnér models in (a) and (b), respectively, at [Mg^{2+}] = 10 mM and to the Åqvist and Allnér in (c) and (d), respectively, at [Mg^{2+}] = 25 mM. The RMSD without residues 48 to 60 (responsible in some case of the oscillatory behavior of the RMSD) is reported with the green line.

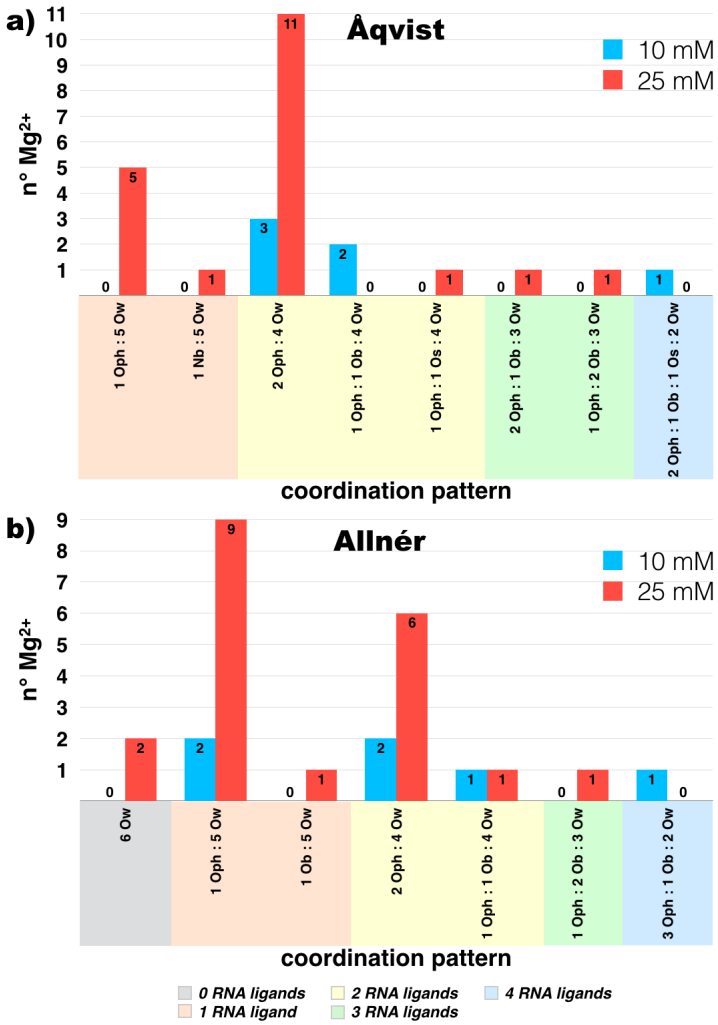


Figure A2.12. Histograms determining the population of the Mg^{2+}–RNA coordination patterns observed during MD simulations of the HDV ribozyme performed with the Åqvist (a) and Allnér (b) Mg^{2+} ffs at [Mg^{2+}] = 10 mM (blue bars) and 25 mM (red bars). The *x-axis* reports the Mg^{2+} CPs identified from MD simulations. The number of RNA ligands (from 0 to 4) is highlighted with different colors. The *y-axis* reports the number of Mg^{2+} ions (i.e., population) having a specific CP. The most populated Mg^{2+}–RNA coordination patterns are $2O_{ph}$:$4O_w$ (Åqvist) and $1O_{ph}$:$5O_w$ (Allnér), with the Allnér parameters favoring more hydrated sites.

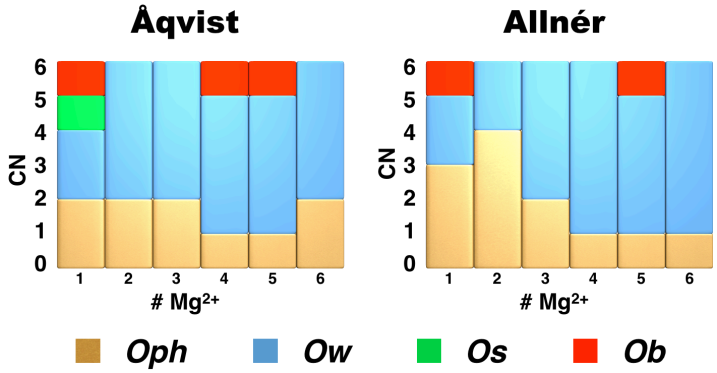


Figure A2.13. Histograms showing the composition of the coordination sphere of the Mg^{2+} *inner-sphere* sites in HDV as determined from MD performed with the Åqvist and Allnér parametrizations at [Mg^{2+}] = 10 mM. The *x-axis* reports the Mg^{2+} sites in HDV. The *y-axis* reports the *CN* for the Mg^{2+} coordination sphere. Donor atom types are identified with different colors as specified in the bottom legend.

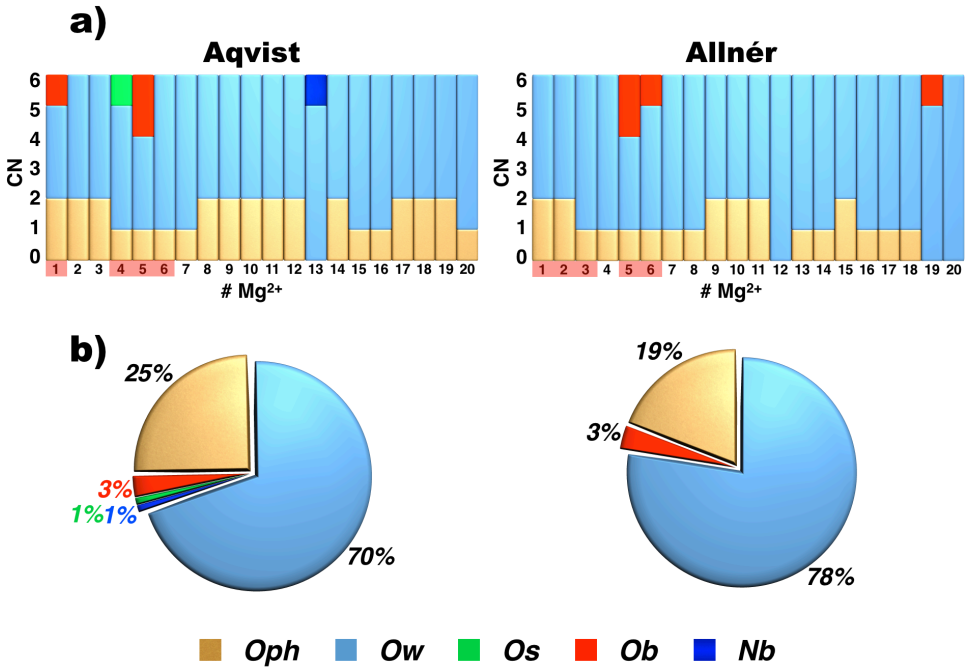


Figure A2.14. (a) Histograms showing the composition of the coordination sphere of the Mg^{2+} sites in HDV from MD performed with the Åqvist and Allnér ffs at [Mg^{2+}] = 25 mM. The *x-axis* reports the Mg^{2+} sites in HDV. The *y-axis* reports the *CN* for the Mg^{2+} coordination sphere. The sites that change their ligand composition with respect to the simulation at [Mg^{2+}] = 10 mM are highlighted in red. Remarkably, for both Mg^{2+} models, only two of the six original sites preserve the same ligand composition at the two Mg^{2+} concentrations (see Figure S18). (b) Statistical distribution of different RNA ligands in the Mg^{2+} *inner-sphere* (expressed as percentages) as calculated from MD simulations of HDV performed with the Åqvist and Allnér ffs at [Mg^{2+}] = 25 mM. Donor atom types are identified with different colors as specified in the bottom legend.

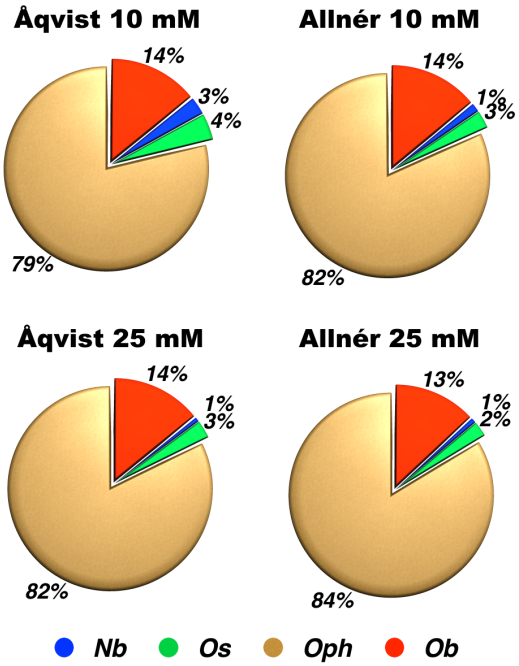


Figure A2.15. Statistical distribution of different RNA ligands in the Mg^{2+} *inner-sphere* (expressed as percentages) as calculated from MD simulations grouping the results of both the G2IR and the HDV according to the parametrization (Åqvist and Allnér) and the Mg^{2+} concentration conditions employed.

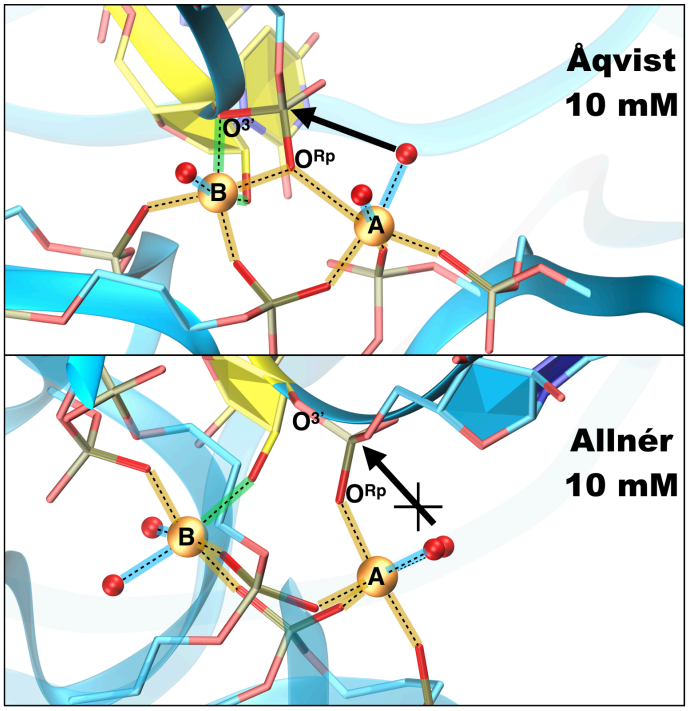


Figure A2.16. Snapshot of the catalytic site of G2IR ribozyme as obtained after 200 ns of classical MD simulations with the Åqvist and the Allnér model at $[Mg^{2+}]$ = 10 mM. A complete distortion of Mg^{2+}-B coordination sphere is visible in the simulations performed with the Allnér model.

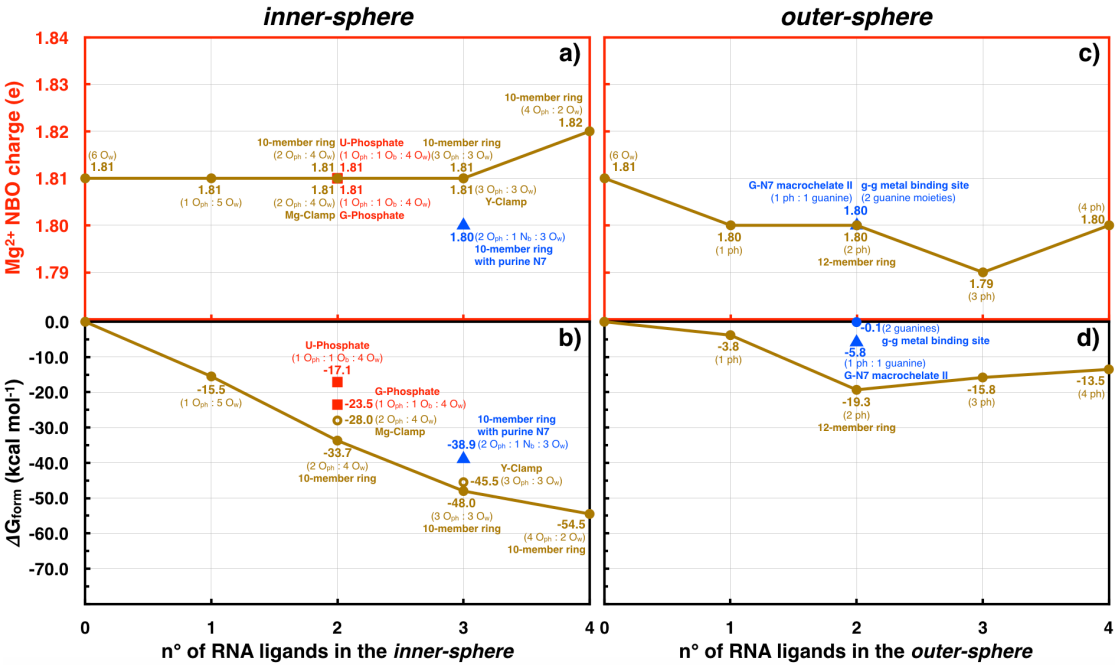


Figure A2.17. Mg^{2+} charge (e) and free energy of formation (ΔG_{form}, kcal/mol) of *inner-sphere* Mg^{2+} coordination sites, (a) and (b) respectively, and of *outer-sphere* Mg^{2+} coordination sites, (c) and (d) respectively, plotted as a function of the number of RNA ligands, and calculated at the DFT/B3LYP/6-311++G** level for the models shown in Figure 5.7a. The Natural Bond Orbital (NBO) charge is used to estimate the charge. Gold circles, red squares and blue triangles refer to model systems characterized by the presence of O_{ph}-only, at least one O_b or one N_b as non-water ligands, respectively. Model systems characterized by O_{ph} ligands only but corresponding to a different geometrical isomer are indicated with golden empty circles. For each model system the CP is reported in parenthesis.

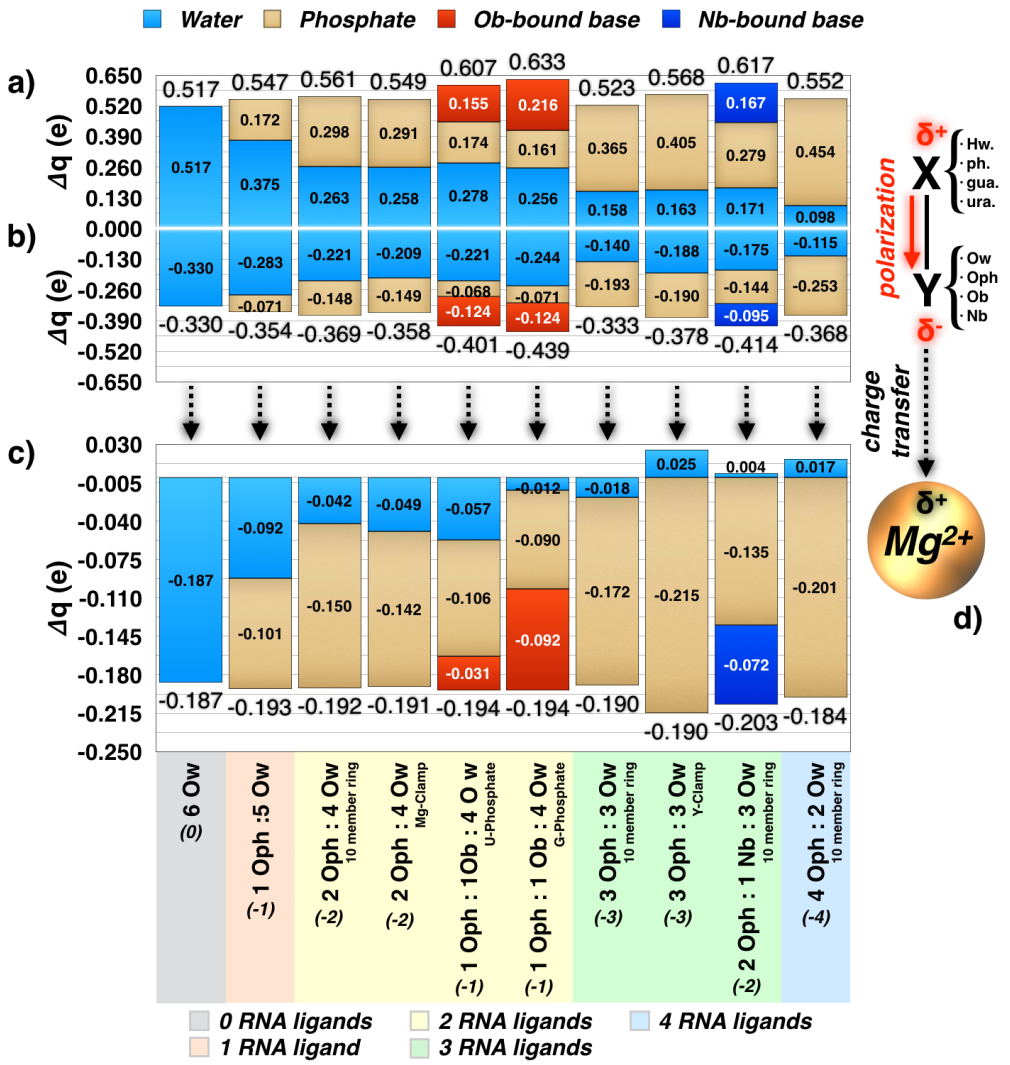


Figure A2.18. Charge rearrangements (Δq, e) of the (a) non-Mg^{2+}-coordinated and (b) Mg^{2+}-coordinated ligands atoms in the *inner-sphere* coordination sites; (c) amount of charge transferred (Δq, e) from the ligands towards Mg^{2+} ion calculated from the NBO charge distribution and the B3LYP functional with 6-311++G** basis set. Each contribution is dissected by atom type with light blue, gold, red and dark blue referring to water (O_w), phosphate (O_{ph}), nucleobases coordinated via O_b and N_b atoms, respectively. (d) Schematic picture of the polarization and charge transfer effects exerted by a Mg^{2+} ion. The formal charge of each CP is reported in parenthesis.

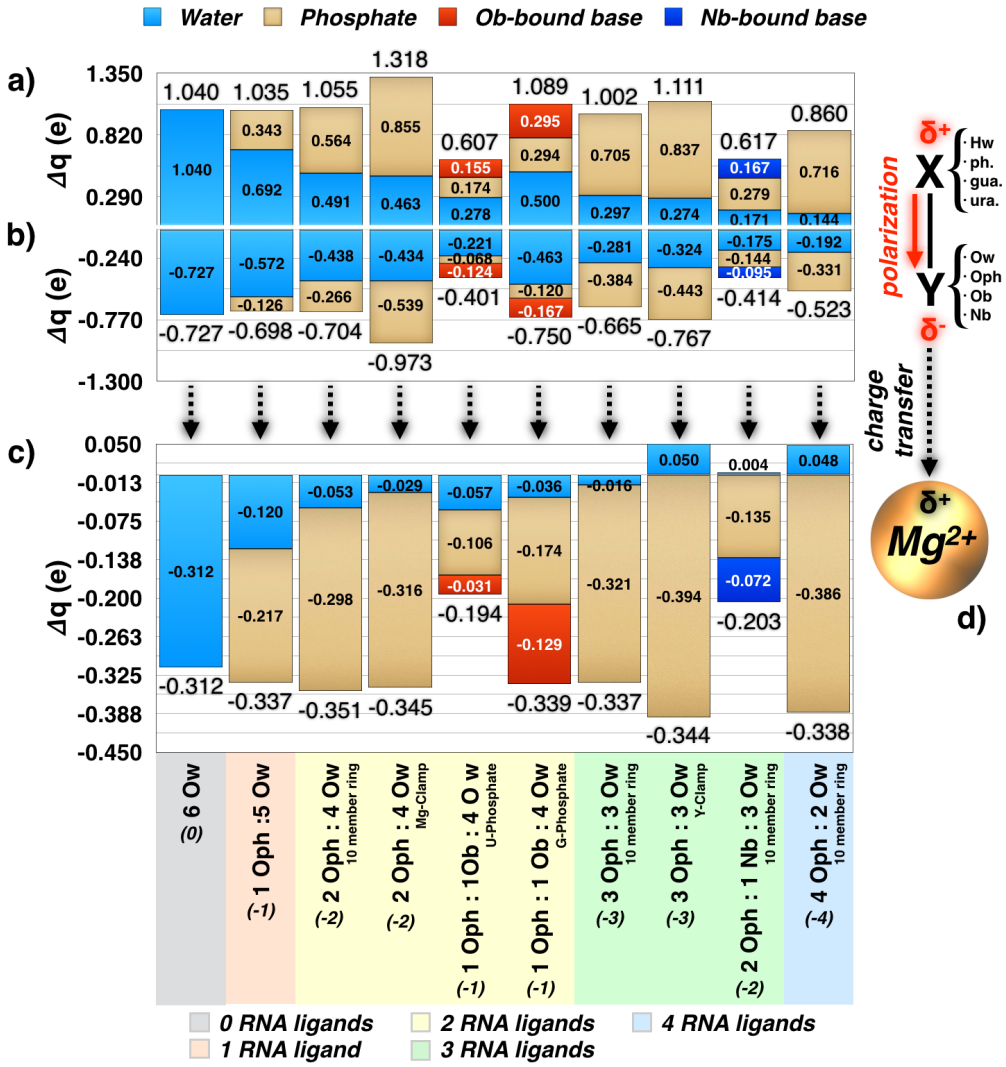


Figure A2.19. Charge rearrangements (Δq, e) of the (a) non-Mg^{2+}-coordinated and (b) Mg^{2+}-coordinated ligands atoms in the *inner-sphere* coordination sites; (c) amount of charge transferred (Δq, e) from the ligands towards Mg^{2+} ion calculated from the NBO charge distribution and the M06 functional with 3-21G basis set. Each contribution is dissected by atom type with light blue, gold, red and dark blue referring to water (O_w), phosphate (O_{ph}), nucleobases coordinated via O_b and N_b atoms, respectively. (d) Schematic picture of the polarization and charge transfer effects exerted by a Mg^{2+} ion. The formal charge of each CP is reported in parenthesis.

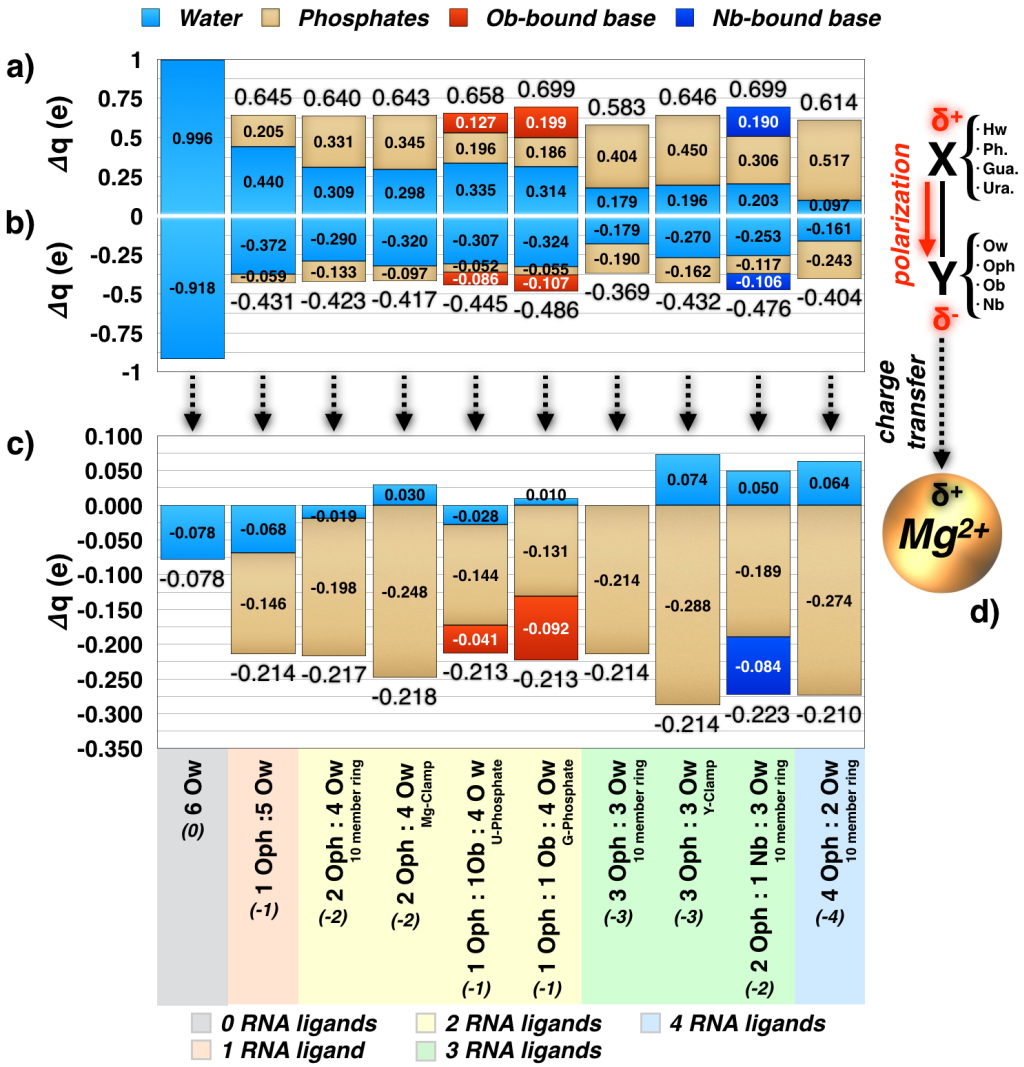


Figure A2.20. Charge rearrangements (Δq, e) of the (a) non-Mg^{2+}-coordinated and (b) Mg^{2+}-coordinated ligands atoms in the *inner-sphere* coordination sites; (c) amount of charge transferred (Δq, e) from the ligands towards Mg^{2+} ion calculated from the Bader charge distribution and the M06 functional with 6-311++G** basis set. Each contribution is dissected by atom type with light blue, gold, red and dark blue referring to water (O_w), phosphate (O_{ph}), nucleobases coordinated via O_b and N_b atoms, respectively. (d) Schematic picture of the polarization and charge transfer effects exerted by a Mg^{2+} ion. The formal charge of each CP is reported in parenthesis.

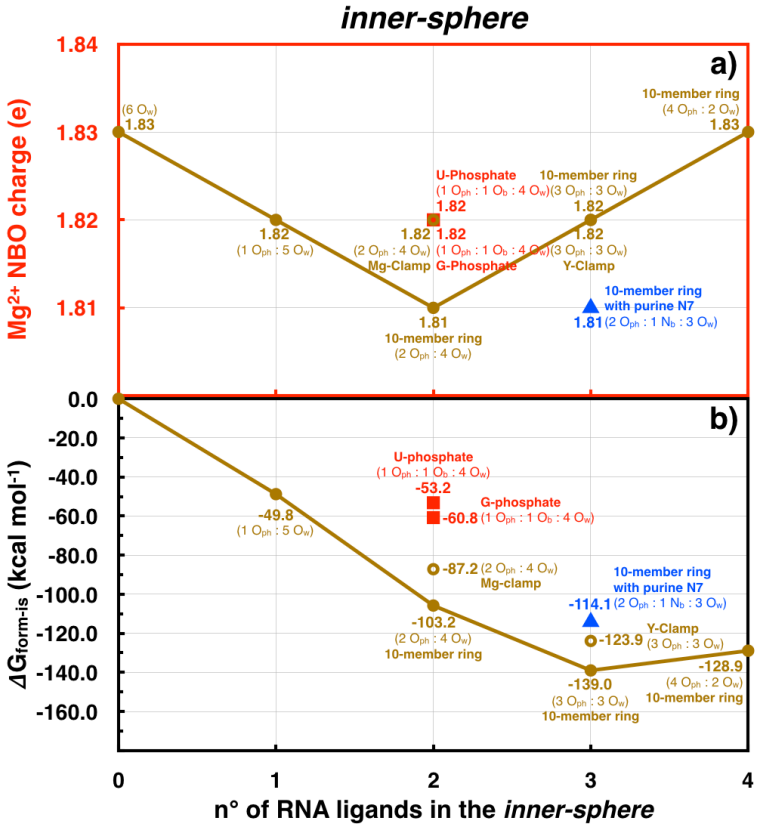


Figure A.21. Mg^{2+} charge (e) and free energy of formation (ΔG_{form}, kcal/mol) of *inner-sphere* Mg^{2+} coordination sites, (a) and (b) respectively, plotted as a function of the number of RNA ligands, and calculated at the DFT/B3LYP/6-311++G** level for the models shown in Figure 5.7a. The Polarizable Continuum Model (PCM) [176] was used with a dielectric constant of 4. The Natural Bond Orbital (NBO) charge is used to estimate the charge. Gold circles, red squares and blue triangles refer to model systems characterized by the presence of O_{ph}-only, at least one O_b or one N_b as non-water ligands, respectively. Model systems characterized by O_{ph} ligands only but corresponding to a different geometrical isomer are indicated with golden empty circles. For each model system the CP is reported in parenthesis.

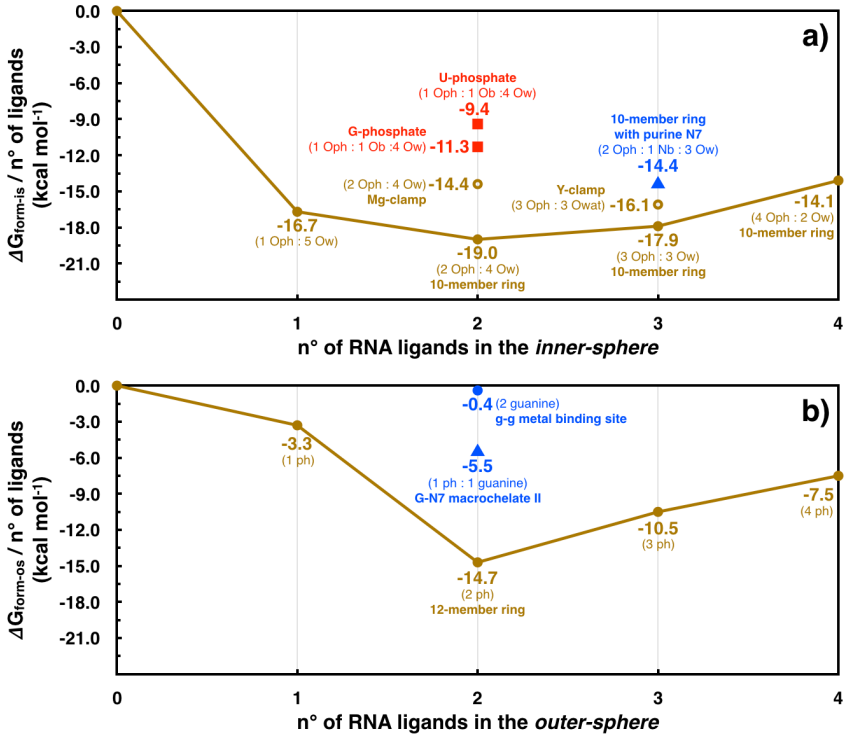


Figure A2.22. Average contribution to $\Delta G_{form\text{-}is}$ and $\Delta G_{form\text{-}os}$ per ligand in (a) and (b), respectively. Data come from simulations at the DFT/M06/6-311++G** level of theory.

Additional Tables.

MODEL	G2IR Ribozyme		HDV Ribozyme	
	10 mM	25 mM	10 mM	25 mM
Åqvist	200 ns 100 ns 100 ns	100 ns	200 ns	200 ns
Allnér	200 ns 100 ns 100 ns	100 ns	200 ns	200 ns
Li	200 ns	100 ns	-	-
Saxena	200 ns	100 ns	-	-
Oelschlaeger	100 ns	-	-	-
SUBTOTAL	1300 ns	400 ns	400 ns	400 ns
TOTAL	2.5 μs			

Table A2.1. Summary of the simulations length performed for the different systems investigated.

R_{max} (Å)					
Åqvist	**Oph**	**Ow**	**Os (O2' O3')**	**Ob (O6 O4)**	**N7**
	1.865	1.995	2.075	1.975	2.265
Allnér	**Oph**	**Ow**	**Os (O2')**	**Ob (O6 O4 O2)**	**N1 N7**
	1.965	2.055	2.135	2.045	2.295
Li	**Oph**	**Ow**	**Os**	**Ob (O6 O4 O2)**	**N7***
	1.935	2.035		2.015	2.195
Saxena	**Oph**	**Ow**	**Os (O2' O3')**	**Ob (O6 O4)**	**N1 N7**
	2.035	2.165	2.175	2.085	
Oelschlaeger	**Oph**	**Ow**	**Os (O2' O3')**	**Ob (O6 O4 O2)**	**N1 N3 N7**
	2.235	2.365	2.325	2.285	2.465
DFT/M06	**Oph**	**Ow**	**Os (O2')**	**Ob (O6 O4 O2)**	**N1 N3 N7**
	2.03	2.115	2.12	2.07	2.225

Table A2.2. R_{max} (Å) of the radial distribution function (*g(r)*) for different Mg^{2+} ligand types, as reproduced by the Åqvist, Allnér, Li, Saxena and Oelschlaeger models. The Mg^{2+}–ligands coordination distance (Å), as calculated at the DFT/M06 level using the using the 6-311++g** basis set, is also reported (last row).

A3 Appendix 3

Additional figures.

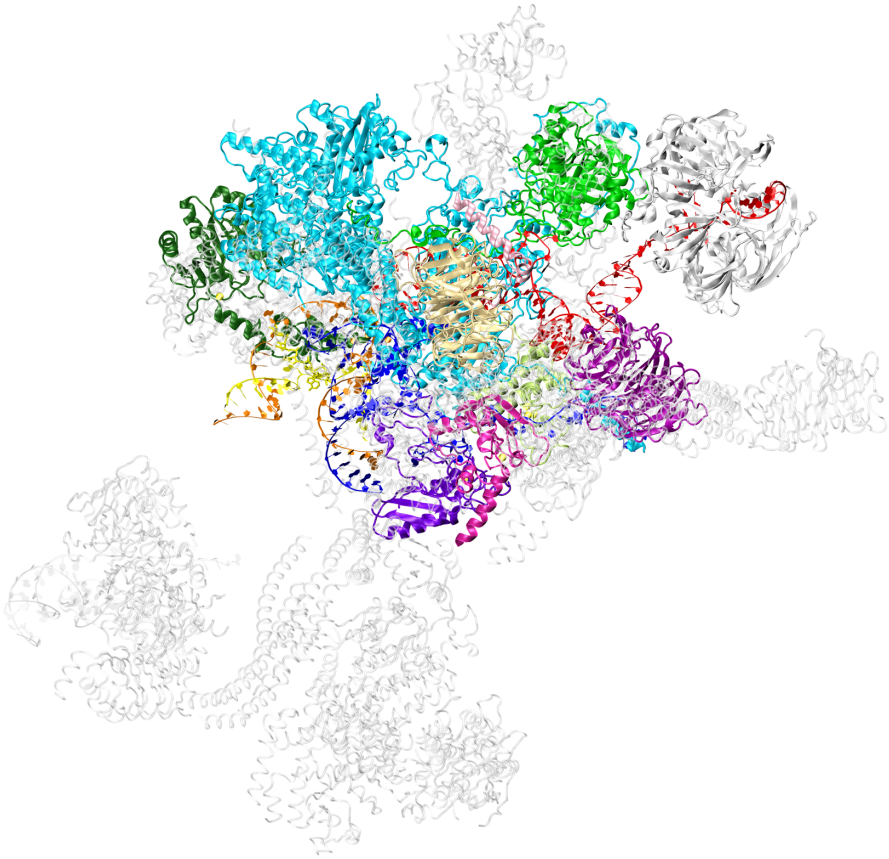


Figure A3.1. Only the core region of the spliceosome from the PDB 3JB9 has been considered for this study. Here, our model (model-1) is shown with a colored opaque new cartoon (proteins) and ribbons (RNAs) representation, while the discarded low-resolution portions of 3JB9 are depicted with a grey transparent representation.

SPLICEOSOME TEST MODEL ("model-1")					
Total number of atoms (water included) = 721089		**Cryo-EM 3.6 A (3jb9)**		**Organism:** *Schizosaccharomyces Pombe*	
Solute atoms: 70190 atoms, 36501 heavy atoms					
CHAIN	**PROTEIN NAME (*S. Cerevisiae*)**	**CONSIDERED**	**MODELLED by US**	**RESOLUTION**	**VMD RESIDUE**
A	**Spp42 (*Prp8*)**	47 - 1532	303 to 313	2.9 ~ 3.6	0 to 1485
				~ 4.0	
B	**Cwf10 (*Snu114*)**	68 - 400	/	2.9 ~ 3.8	1486 to 1818
C	**U5 snRNA**	7 - 111	/	2.9 ~ 3.6	1819 to 1923
D	**SM-D3**	2 - 97	/	3.3 ~ 4.0	1924 to 2019
E	**SM-B**	2 - 86	48 to 59	3.3 ~ 4.0	2020 to 2104
F	**SM-D1**	1 - 82	/	3.3 ~ 4.0	2105 to 2186
G	**SM-D2**	19 - 115	85 to 86	3.3 ~ 4.0	2187 to 2283
H	**SM-E**	9 - 84	/	3.3 ~ 4.0	2284 to 2359
I	**SM-F**	4 - 75	/	3.3 ~ 4.0	2360 to 2431
J	**SM-G**	3 - 75	/	3.3 ~ 4.0	2432 to 2504
K	**Prp5 (*Cwf1*)**	149 - 470	/	~ 3.4	2505 to 2826
L	**Cwf17**	42 - 340	81, 147, 250 to 253	3.3 ~ 4.0	2827 to 3125
N	**U6 snRNA**	1 - 90	/	2.9 ~ 4.5	3126 to 3215
O + Q	**lariat intron**	100 - 107 + 492 - 504	"GA2" = A501 bonded to G100	/	3216 to 3235
P	**U2 snRNA**	1 - 43	/	2.9 ~ 4.5	3236 to 3278
Y	**Cwf2 (*Prp3*)**	49 - 235	/	3.3 ~ 5.0	3279 to 3465
a	**Cwf5**	18 - 151	/	3.3 ~ 4.0	3466 to 3599
c	**Cwf19**	334 - 633	/	3.4 ~ 4.0	3600 to 3899
e	**Cwf14**	3 - 146	/	~ 3.4	3900 to 4043
h	**Cwf15**	24 - 70	/	3.0 ~ 4.0	4044 to 4090
	Mg+	# 4	Aqvist ff	/	4091 to 4094
	ZNB (Zn2+)	# 7 (35 atoms)	Pang dummy cations ff	/	4095 to 4101
	Na+	# 202	Joung & Cheatham ff	/	4102 to 4303
	Wat	# 216886	TIP3P	/	4304 to 221189

Figure A3.2. Details of the spliceosome test model "model-1".

SPLICEOSOME EXTENDED MODEL ("model-2")						
Total number of atoms (water included) = 914099		**Cryo-EM 3.6 A (3jb9)**		***Organism:*** *Schizosaccharomyces Pombe*		
Solute atoms: 92276 protein and RNA atoms, 47510 heavy atoms						
CHAIN	**PROTEIN NAME (*S. Cerevisiae*)**	**CONSIDERED**	**MODELLED by US**	**RESOLUTION**	**VMD RESIDUE**	**N° of RES**
A	**Spp42 (*Prp8*)**	47 - 2030	303-313/1533-1538/1781-1783	2.9 ~ 3.6	0 to 1983	1984
				~ 4.0		
B	**Cwf10 (*Snu114*)**	68 - 971	/	2.9 ~ 3.8	1984 to 2887	904
C	**U5 snRNA**	7 - 111	/	2.9 ~ 3.6	2888 to 2992	105
D	**SM-D3**	2 - 97	/	3.3 ~ 4.0	2993 to 3088	581
E	**SM-B**	2 - 86	48 to 59	3.3 ~ 4.0	3089 to 3173	
F	**SM-D1**	1 - 82	/	3.3 ~ 4.0	3174 to 3255	
G	**SM-D2**	19 - 115	85 to 86	3.3 ~ 4.0	3256 to 3352	
H	**SM-E**	9 - 84	/	3.3 ~ 4.0	3353 to 3428	
I	**SM-F**	4 - 75	/	3.3 ~ 4.0	3429 to 3500	
J	**SM-G**	3 - 75	/	3.3 ~ 4.0	3501 to 3573	
K	**Prp5 (*Cwf1*)**	149 - 470	/	~ 3.4	3574 to 3895	322
L	**Cwf17**	42 - 340	81, 147, 250 to 253	3.3 ~ 4.0	3896 to 4194	299
N	**U6 snRNA**	1 - 90	/	2.9 ~ 4.5	4195 to 4284	90
O + Q	**lariat intron**	100 - 107 + 492 - 504	"GA2" = A501 bonded to G100	/	4285 to 4304	20
P	**U2 snRNA**	1 - 43	/	2.9 ~ 4.5	4305 to 4347	43
Y	**Cwf2 (*Prp3*)**	49 - 235	/	3.3 ~ 5.0	4348 to 4534	187
a	**Cwf5**	18 - 151	/	3.3 ~ 4.0	4535 to 4668	134
c	**Cwf19**	334 - 633	/	3.4 ~ 4.0	4669 to 4968	300
e	**Cwf14**	3 - 146	/	~ 3.4	4969 to 5112	144
h	**Cwf15**	24 - 70	/	3.0 ~ 4.0	5113 to 5159	47
M	**Prp45**	100 - 271	/	3.0 ~ 4.5	5160 to 5331	172
g	**Prp17**	29 - 161	/	3.3 ~ 4.5	5332 to 5464	133
	Mg+	# 4	Aqvist ff	/	5465 to 5468	4
	ZNB (Zn2+)	# 7 (35 atoms)	Pang dummy cations ff	/	5469 to 5475	7
	GDP	#1 (40 atoms)	Meagher KL ff	/	5476	1
	Na+	# 194	Joung & Cheatham ff	/	5477 to 5670	194
	Wat	# 273850	TIP3P	/	5671 to 279520	273850

Figure A3.3. Details of the spliceosome extended model "model-2". Results are not discussed in this thesis work.

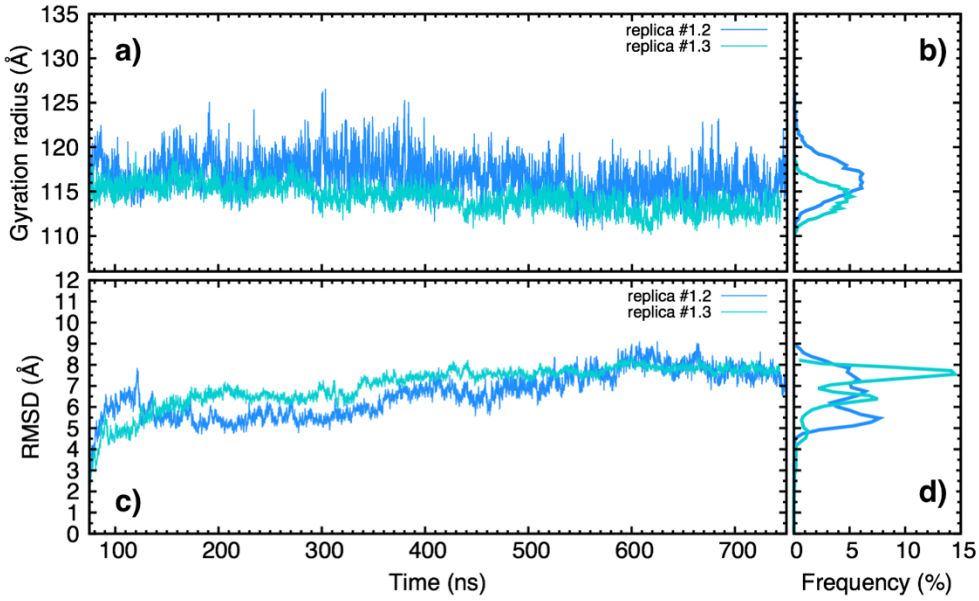


Figure A3.4. Time evolution (ns) of Gyration Radius (a) and RMSD (c) and their relative frequencies (%) shown in (b) and (d), respectively, obtained from MD replicas #1.2 and #1.3 MD (blue and light blue, respectively). The profiles are obtained including all the proteins and RNAs in the analyses.

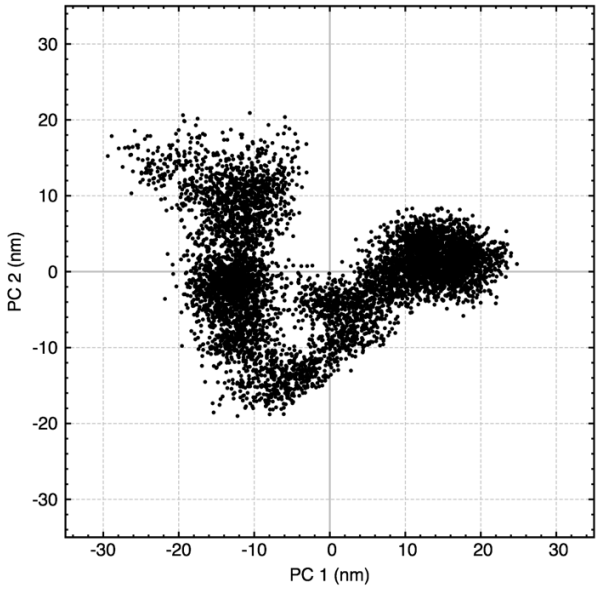


Figure A3.5. Scatter plot representing the projections of the C-alpha and P displacements along the trajectory onto the first principal eigenvector, PC1 (x-axis) vs the projections onto the second principal eigenvector, PC2 (y-axis) as derived from MD simulation of model-1, replica #1.2.

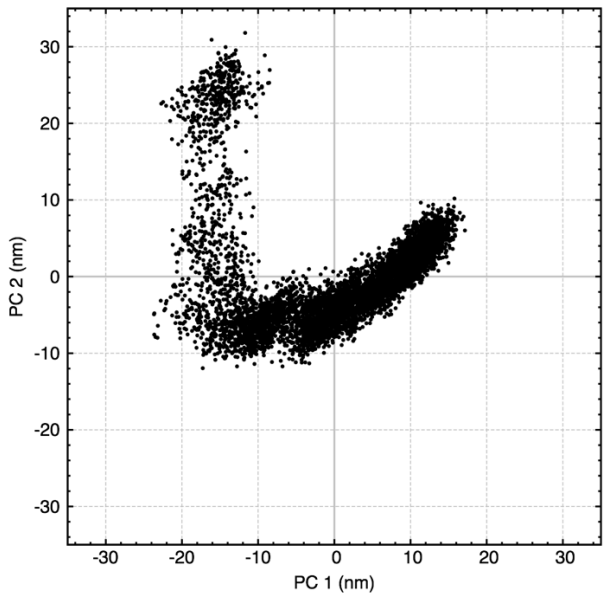


Figure A3.6. Scatter plot representing the projections of the C-alpha and P displacements along the trajectory onto the first principal eigenvector, PC1 (x-axis) vs the projections onto the second principal eigenvector, PC2 (y-axiss) as derived from MD simulation of model-1, replica #1.3.

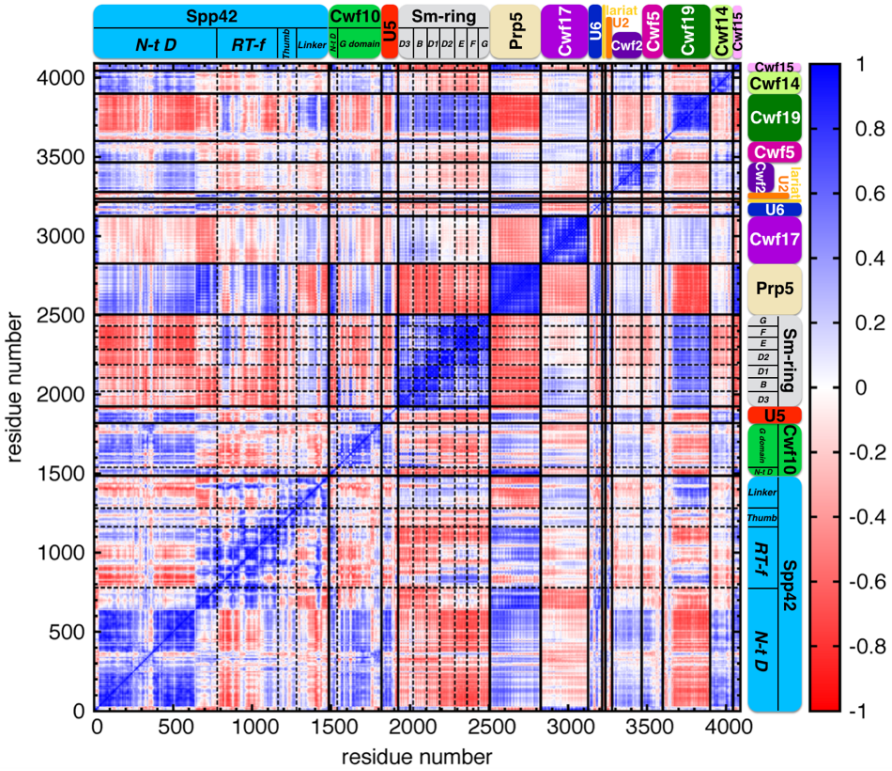


Figure A3.7. Pearson's cross-correlation matrix derived from the mass-weighted covariance matrix constructed over the last 670 ns of MD simulations of replica #1.2 for C-alpha and P atoms. The Pearson's coefficients are comprised between -1 (anti-correlation, red) and +1 (correlation, blue). Macromolecules and domains names are highlighted with different colors.

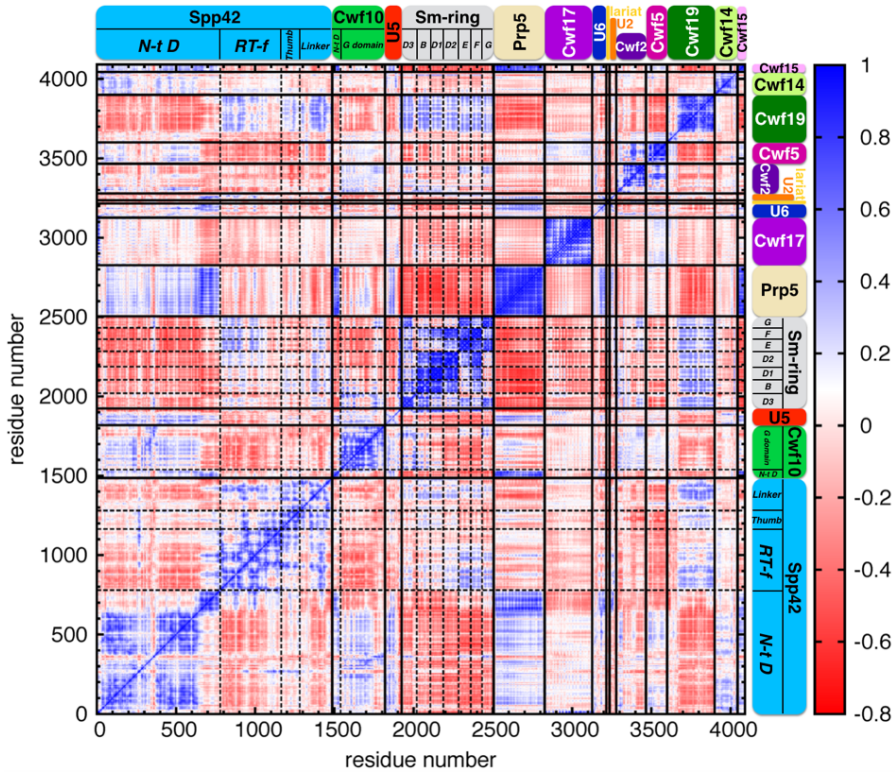


Figure A3.8. Pearson's cross-correlation matrix derived from the mass-weighted covariance matrix constructed over the last 670 ns of MD simulations of replica #1.3 for C-alpha and P atoms. The Pearson's coefficients are comprised between -1 (anti-correlation, red) and +1 (correlation, blue). Macromolecules and domains names are highlighted with different colors.

8 Bibliography

[1] Gilbert, W. Why genes in pieces? *Nature* **1978**, *271*, 501.

[2] Rogozin, I.B., Carmel, L., Csuros, M. and Koonin, E.V. Origin and evolution of spliceosomal introns. *Biol. Direct* **2012**, *7*, 11.

[3] Swaminathan, R. Magnesium metabolism and its disorders. *Clin. Biochem. Rev.* **2003**, *24*, 47-66.

[4] Elin, R.J. Magnesium: the fifth but forgotten electrolyte. *Am. J. Clin. Pathol.* **1994**, *102*, 616-622.

[5] Ryan, M.F. The role of magnesium in clinical biochemistry: an overview. *Ann. Clin. Biochem.* **1991**, *28 (Pt 1)*, 19-26.

[6] Wacker, W.E. The biochemistry of magnesium. *Ann. N. Y. Acad. Sci.* **1969**, *162*, 717-726.

[7] Quamme, G.A. and Rabkin, S.W. Cytosolic free magnesium in cardiac myocytes: identification of a Mg2+ influx pathway. *Biochem. Biophys. Res. Commun.* **1990**, *167*, 1406-1412.

[8] Gums, J.G. Magnesium in cardiovascular and other disorders. *Am. J. Health Syst. Pharm.* **2004**, *61*, 1569-1576.

[9] Apell, H.J., Hitzler, T. and Schreiber, G. Modulation of the Na,K-ATPase by Magnesium Ions. *Biochemistry* **2017**, *56*, 1005-1016.

[10] Ko, Y.H., Hong, S. and Pedersen, P.L. Chemical mechanism of ATP synthase. Magnesium plays a pivotal role in formation of the transition state where ATP is synthesized from ADP and inorganic phosphate. *J. Biol. Chem.* **1999**, *274*, 28853-28856.

[11] Palermo, G., Cavalli, A., Klein, M.L., Alfonso-Prieto, M., Dal Peraro, M. and De Vivo, M. Catalytic metal ions and enzymatic processing of DNA and RNA. *Acc. Chem. Res.* **2015**, *48*, 220-228.

[12] Erat, M.C. and Sigel, R.K. Divalent metal ions tune the self-splicing reaction of the yeast mitochondrial group II intron Sc.ai5gamma. *J. Biol. Inorg. Chem.* **2008**, *13*, 1025-1036.

[13] Bowman, J.C., Lenz, T.K., Hud, N.V. and Williams, L.D. Cations in charge: magnesium ions in RNA folding and catalysis. *Curr. Opin. Struct. Biol.* **2012**, *22*, 262-272.

[14] Woodson, S.A. Metal ions and RNA folding: a highly charged topic with a dynamic future. *Curr. Opin. Chem. Biol.* **2005**, *9*, 104-109.

[15] Cunha, R.A. and Bussi, G. Unraveling Mg2+-RNA binding with atomistic molecular dynamics. *RNA* **2017**, *23*, 628-638.

[16] Hayes, R.L., Noel, J.K., Mohanty, U., Whitford, P.C., Hennelly, S.P., Onuchic, J.N. and Sanbonmatsu, K.Y. Magnesium fluctuations modulate RNA dynamics in the SAM-I riboswitch. *J. Am. Chem. Soc.* **2012**, *134*, 12043-12053.

[17] Strobel, S.A. and Cochrane, J.C. RNA catalysis: ribozymes, ribosomes, and riboswitches. *Curr. Opin. Chem. Biol.* **2007**, *11*, 636-643.

[18] Sgrignani, J. and Magistrato, A. QM/MM MD Simulations on the Enzymatic Pathway of the Human Flap Endonuclease (hFEN1) Elucidating Common Cleavage Pathways to RNase H Enzymes. *ACS Catal.* **2015**, *5*, 3864–3875.

[19] Ho, M.H., De Vivo, M., Dal Peraro, M. and Klein, M.L. Understanding the effect of magnesium ion concentration on the catalytic activity of ribonuclease H through computation: does a third metal binding site modulate endonuclease catalysis? *J. Am. Chem. Soc.* **2010**, *132*, 13702-13712.

[20] De Vivo, M., Dal Peraro, M. and Klein, M.L. Phosphodiester cleavage in ribonuclease H occurs via an associative two-metal-aided catalytic mechanism. *J. Am. Chem. Soc.* **2008**, *130*, 10955-10962.

[21] Steitz, T.A. and Steitz, J.A. A general two-metal-ion mechanism for catalytic RNA. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 6498-6502.

[22] Ward, W.L., Plakos, K. and DeRose, V.J. Nucleic Acid Catalysis: Metals, Nucleobases, and Other Cofactors. *Chem. Rev.* **2014**, *114*, 4318-4342.

[23] Schnabl, J. and Sigel, R.K. Controlling ribozyme activity by metal ions. *Curr. Opin. Chem. Biol.* **2010**, *14*, 269-275.

[24] Lee, T.S., Lopez, C.S., Giambasu, G.M., Martick, M., Scott, W.G. and York, D.M. Role of Mg2+ in hammerhead ribozyme catalysis from molecular simulation. *J. Am. Chem. Soc.* **2008**, *130*, 3053-3064.

[25] Marcia, M. and Pyle, A.M. Visualizing group II intron catalysis through the stages of splicing. *Cell* **2012**, *151*, 497-507.

[26] Stahley, M.R. and Strobel, S.A. Structural evidence for a two-metal-ion mechanism of group I intron splicing. *Science* **2005**, *309*, 1587-1590.

[27] Reiter, N.J., Osterman, A., Torres-Larios, A., Swinger, K.K., Pan, T. and Mondragon, A. Structure of a bacterial ribonuclease P holoenzyme in complex with tRNA. *Nature* **2010**, *468*, 784-789.

[28] Boero, M., Tateno, M., Terakura, K. and Oshiyama, A. Double-metal-ion/single-metal-ion mechanisms of the cleavage reaction of ribozymes: First-principles molecular dynamics simulations of a fully hydrated model system. *J. Chem. Theory Comput.* **2005**, *1*, 925-934.

[29] Casalino, L., Palermo, G., Rothlisberger, U. and Magistrato, A. Who Activates the Nucleophile in Ribozyme Catalysis? An Answer from the Splicing Mechanism of Group II Introns. *J. Am. Chem. Soc.* **2016**, *138*, 10374-10377.

[30] Marcia, M., Somarowthu, S. and Pyle, A.M. Now on display: a gallery of group II intron structures at different stages of catalysis. *Mob. DNA* **2013**, *4*, 14.

[31] Yan, C., Hang, J., Wan, R., Huang, M., Wong, C.C.L. and Shi, Y. Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science* **2015**, *349*, 1182-1191.

[32] Hang, J., Wan, R., Yan, C. and Shi, Y. Structural basis of pre-mRNA splicing. *Science* **2015**, *349*, 1191-1198.

[33] Bertram, K., Agafonov, D.E., Liu, W.T., Dybkov, O., Will, C.L., Hartmuth, K., Urlaub, H., Kastner, B., Stark, H. and Luhrmann, R. Cryo-EM structure of a human spliceosome activated for step 2 of splicing. *Nature* **2017**, *542*, 318-323.

[34] Zhang, X., Yan, C., Hang, J., Finci, L.I., Lei, J. and Shi, Y. An Atomic Structure of the Human Spliceosome. *Cell* **2017**, *169*, 918-929 e914.

[35] Sharp, P.A. The discovery of split genes and RNA splicing. *Trends Biochem. Sci* **2005**, *30*, 279-281.

[36] Berget, S.M., Moore, C. and Sharp, P.A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. U. S. A.* **1977**, *74*, 3171-3175.

[37] Toor, N., Keating, K.S. and Pyle, A.M. Structural insights into RNA splicing. *Curr. Opin. Struct. Biol.* **2009**, *19*, 260-266.

[38] Lee, Y. and Rio, D.C. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu. Rev. Biochem* **2015**, *84*, 291-323.

[39] Papasaikas, P. and Valcarcel, J. The Spliceosome: The Ultimate RNA Chaperone and Sculptor. *Trends Biochem. Sci.* **2016**, *41*, 33-45.

[40] Chabot, B. and Shkreta, L. Defective control of pre-messenger RNA splicing in human disease. *J. Cell Biol.* **2016**, *212*, 13-27.

[41] Faustino, N.A. and Cooper, T.A. Pre-mRNA splicing and human disease. *Genes Dev.* **2003**, *17*, 419-437.

[42] Pyle, A.M. Group II Intron Self-Splicing. *Annu. Rev. Biophys.* **2016**, *45*, 183-205.

[43] Lambowitz, A.M. and Zimmerly, S. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb. Perspect. Biol.* **2011**, *3*, a003616.

[44] Fica, S.M., Tuttle, N., Novak, T., Li, N.S., Lu, J., Koodathingal, P., Dai, Q., Staley, J.P. and Piccirilli, J.A. RNA catalyses nuclear pre-mRNA splicing. *Nature* **2013**, *503*, 229-234.

[45] Lambowitz, A.M. and Belfort, M. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiol Spectr* **2015**, *3*, MDNA3-0050-2014.

[46] Lambowitz, A.M. and Zimmerly, S. Mobile group II introns. *Annu. Rev. Genet.* **2004**, *38*, 1-35.

[47] Pyle, A.M. The tertiary structure of group II introns: implications for biological function and evolution. *Crit. Rev. Biochem. Mol. Biol.* **2010**, *45*, 215-232.

[48] Rest, J.S. and Mindell, D.P. Retroids in archaea: phylogeny and lateral origins. *Mol. Biol. Evol.* **2003**, *20*, 1134-1142.

[49] Robart, A.R., Chan, R.T., Peters, J.K., Rajashankar, K.R. and Toor, N. Crystal structure of a eukaryotic group II intron lariat. *Nature* **2014**, *514*, 193-197.

[50] Qu, G., Kaushal, P.S., Wang, J., Shigematsu, H., Piazza, C.L., Agrawal, R.K., Belfort, M. and Wang, H.W. Structure of a group II intron in complex with its reverse transcriptase. *Nat. Struct. Mol. Biol.* **2016**, *23*, 549-557.

[51] Karberg, M., Guo, H., Zhong, J., Coon, R., Perutka, J. and Lambowitz, A.M. Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. *Nat. Biotechnol.* **2001**, *19*, 1162-1167.

[52] Perutka, J., Wang, W., Goerlitz, D. and Lambowitz, A.M. Use of computer-designed group II introns to disrupt Escherichia coli DExH/D-box protein and DNA helicase genes. *J. Mol. Biol.* **2004**, *336*, 421-439.

[53] Zimmerly, S. and Semper, C. Evolution of group II introns. *Mob. DNA* **2015**, *6*, 7.

[54] Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **2005**, *308*, 1149-1154.

[55] Matera, A.G., Terns, R.M. and Terns, M.P. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 209-220.

[56] Will, C.L. and Luhrmann, R. Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.* **2011**, *3*.

[57] Dietrich, R.C., Incorvaia, R. and Padgett, R.A. Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. *Mol. Cell* **1997**, *1*, 151-160.

[58] Lin, R.J., Newman, A.J., Cheng, S.C. and Abelson, J. Yeast mRNA splicing in vitro. *J. Biol. Chem.* **1985**, *260*, 14780-14792.

[59] Yan, C., Wan, R., Bai, R., Huang, G. and Shi, Y. Structure of a yeast activated spliceosome at 3.5 A resolution. *Science* **2016**, *353*, 904-911.

[60] Wan, R., Yan, C., Bai, R., Huang, G. and Shi, Y. Structure of a yeast catalytic step I spliceosome at 3.4 A resolution. *Science* **2016**, *353*, 895-904.

[61] Nguyen, T.H.D., Galej, W.P., Bai, X.C., Oubridge, C., Newman, A.J., Scheres, S.H.W. and Nagai, K. Cryo-EM structure of the yeast U4/U6.U5 tri-snRNP at 3.7 A resolution. *Nature* **2016**, *530*, 298-302.

[62] Wan, R., Yan, C., Bai, R., Wang, L., Huang, M., Wong, C.C. and Shi, Y. The 3.8 A structure of the U4/U6.U5 tri-snRNP: Insights into spliceosome assembly and catalysis. *Science* **2016**, *351*, 466-475.

[63] Agafonov, D.E., Kastner, B., Dybkov, O., Hofele, R.V., Liu, W.T., Urlaub, H., Luhrmann, R. and Stark, H. Molecular architecture of the human U4/U6.U5 tri-snRNP. *Science* **2016**, *351*, 1416-1420.

[64] Galej, W.P., Wilkinson, M.E., Fica, S.M., Oubridge, C., Newman, A.J. and Nagai, K. Cryo-EM structure of the spliceosome immediately after branching. *Nature* **2016**, *537*, 197-201.

[65] Cate, J.H. A Big Bang in spliceosome structural biology. *Science* **2016**, *351*, 1390-1392.

[66] Kondo, Y., Oubridge, C., van Roon, A.M. and Nagai, K. Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. *Elife* **2015**, *4*.

[67] Nguyen, T.H., Galej, W.P., Fica, S.M., Lin, P.C., Newman, A.J. and Nagai, K. CryoEM structures of two spliceosomal complexes: starter and dessert at the spliceosome feast. *Curr. Opin. Struct. Biol.* **2016**, *36*, 48-57.

[68] Wachtel, C. and Manley, J.L. Splicing of mRNA precursors: the role of RNAs and proteins in catalysis. *Mol. Biosyst.* **2009**, *5*, 311-316.

[69] Fromont-Racine, M., Mayes, A.E., Brunet-Simon, A., Rain, J.C., Colley, A., Dix, I., Decourty, L., Joly, N., Ricard, F., Beggs, J.D., et al. Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. *Yeast* **2000**, *17*, 95-110.

[70] Nagai, K., Muto, Y., Pomeranz Krummel, D.A., Kambach, C., Ignjatovic, T., Walke, S. and Kuglstatter, A. Structure and assembly of the spliceosomal snRNPs. Novartis Medal Lecture. *Biochem. Soc. Trans.* **2001**, *29*, 15-26.

[71] Urlaub, H., Raker, V.A., Kostka, S. and Luhrmann, R. Sm protein-Sm site RNA interactions within the inner ring of the spliceosomal snRNP core structure. *EMBO J.* **2001**, *20*, 187-196.

[72] Stark, H. and Luhrmann, R. Cryo-electron microscopy of spliceosomal components. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 435-457.

[73] Kramer, A., Gruter, P., Groning, K. and Kastner, B. Combined biochemical and electron microscopic analyses reveal the architecture of the mammalian U2 snRNP. *J. Cell Biol.* **1999**, *145*, 1355-1368.

[74] Sashital, D.G., Venditti, V., Angers, C.G., Cornilescu, G. and Butcher, S.E. Structure and thermodynamics of a conserved U2 snRNA domain from yeast and human. *RNA* **2007**, *13*, 328-338.

[75] Nguyen, T.H., Galej, W.P., Bai, X.C., Savva, C.G., Newman, A.J., Scheres, S.H. and Nagai, K. The architecture of the spliceosomal U4/U6.U5 tri-snRNP. *Nature* **2015**, *523*, 47-52.

[76] Grainger, R.J. and Beggs, J.D. Prp8 protein: at the heart of the spliceosome. *RNA* **2005**, *11*, 533-557.

[77] Bartels, C., Urlaub, H., Luhrmann, R. and Fabrizio, P. Mutagenesis suggests several roles of Snu114p in pre-mRNA splicing. *J. Biol. Chem.* **2003**, *278*, 28324-28334.

[78] Brenner, T.J. and Guthrie, C. Assembly of Snu114 into U5 snRNP requires Prp8 and a functional GTPase domain. *RNA* **2006**, *12*, 862-871.

[79] Small, E.C., Leggett, S.R., Winans, A.A. and Staley, J.P. The EF-G-like GTPase Snu114p regulates spliceosome dynamics mediated by Brr2p, a DExD/H box ATPase. *Mol. Cell* **2006**, *23*, 389-399.

[80] Yean, S.L. and Lin, R.J. U4 small nuclear RNA dissociates from a yeast spliceosome and does not participate in the subsequent splicing reaction. *Mol. Cell. Biol.* **1991**, *11*, 5571-5577.

[81] De Almeida, R.A. and O'Keefe, R.T. The NineTeen Complex (NTC) and NTC-associated proteins as targets for spliceosomal ATPase action during pre-mRNA splicing. *Rna Biol.* **2015**, *12*, 109-114.

[82] Hogg, R., McGrail, J.C. and O'Keefe, R.T. The function of the NineTeen Complex (NTC) in regulating spliceosome conformations and fidelity during pre-mRNA splicing. *Biochem. Soc. Trans.* **2010**, *38*, 1110-1115.

[83] Yan, C., Wan, R., Bai, R., Huang, G. and Shi, Y. Structure of a yeast step II catalytically activated spliceosome. *Science* **2017**, *355*, 149-155.

[84] Chen, H.C., Tseng, C.K., Tsai, R.T., Chung, C.S. and Cheng, S.C. Link of NTR-mediated spliceosome disassembly with DEAH-box ATPases Prp2, Prp16, and Prp22. *Mol. Cell. Biol.* **2013**, *33*, 514-525.

[85] Frenkel, D. and Smit, B. *Chapter 4 - Molecular Dynamics Simulations*, in *Understanding Molecular Simulation (Second Edition)*. 2002, San Diego: Academic Press. p. 63-107.

[86] Allen, M.P. and Tildesley, D.J. *Computer simulation of liquids*. 1989, USA: Oxford University Press.

[87] Swope, W.C., Andersen, H.C., Berens, P.H. and Wilson, K.R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* **1982**, *76*, 637-649.

[88] Andersen, H.C. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *J. Comput. Phys.* **1983**, *52*, 24-34.

[89] Van Gunsteren, W.F. and Berendsen, H.J.C. A Leap-frog Algorithm for Stochastic Dynamics. *Mol. Simul.* **1988**, *1*, 173-185.

[90] McCammon, J.A., Gelin, B.R. and Karplus, M. Dynamics of folded proteins. *Nature* **1977**, *267*, 585-590.

[91] Dror, R.O., Dirks, R.M., Grossman, J.P., Xu, H. and Shaw, D.E. Biomolecular simulation: a computational microscope for molecular biology. *Annu. Rev. Biophys.* **2012**, *41*, 429-452.

[92] Salomon-Ferrer, R., Gotz, A.W., Poole, D., Le Grand, S. and Walker, R.C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **2013**, *9*, 3878-3888.

[93] Hospital, A., Goni, J.R., Orozco, M. and Gelpi, J.L. Molecular dynamics simulations: advances and applications. *Adv. Appl. Bioinform. Chem.* **2015**, *8*, 37-47.

[94] Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W. and Kollman, P.A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179-5197.

[95] Van Gunsteren, W.F. *Biomolecular simulation: the GROMOS96 manual and user guide*. 1996, Zürich: Hochschulverlag AG der ETH Zürich.

[96] MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586-3616.

[97] Ewald, P.P. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik* **1921**, *369*, 253-287.

[98] Darden, T., York, D. and Pedersen, L. Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089-10092.

[99] Tuckerman, M.E. and Martyna, G.J. Understanding modern molecular dynamics: Techniques and applications (vol 105B, pg 159, 2000). *J. Phys. Chem. B* **2001**, *105*, 7598-7598.

[100] Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J.L., Dror, R.O. and Shaw, D.E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct. Funct. Bioinform.* **2010**, *78*, 1950-1958.

[101] Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M.P., Dror, R.O. and Shaw, D.E. Systematic Validation of Protein Force Fields against Experimental Data. *Plos One* **2012**, *7*.

[102] Sponer, J., Banas, P., Jurecka, P., Zgarbova, M., Kuhrova, P., Havrila, M., Krepl, M., Stadlbauer, P. and Otyepka, M. Molecular Dynamics Simulations of Nucleic Acids. From Tetranucleotides to the Ribosome. *J. Phys. Chem. Lett.* **2014**, *5*, 1771-1782.

[103] Cheatham, T.E., Cieplak, P. and Kollman, P.A. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845-862.

[104] Wang, J., Cieplak, P. and Kollman, P.A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* **2000**, *21*, 1049-1074.

[105] Sponer, J., Krepl, M., Banas, P., Kuhrova, P., Zgarbova, M., Jurecka, P., Havrila, M. and Otyepka, M. How to understand atomistic molecular dynamics simulations of RNA and protein-RNA complexes? *WIREs RNA* **2017**, *8*.

[106] Krepl, M., Havrila, M., Stadlbauer, P., Banas, P., Otyepka, M., Pasulka, J., Stefl, R. and Sponer, J. Can We Execute Stable Microsecond-Scale Atomistic Simulations of Protein-RNA Complexes? *J. Chem. Theory Comput.* **2015**, *11*, 1220-1243.

[107] Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E. and Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696-3713.

[108] Adelman, S.A. and Doll, J.D. Generalized Langevin Equation Approach for Atom-Solid-Surface Scattering - General Formulation for Classical Scattering Off Harmonic Solids. *J. Chem. Phys.* **1976**, *64*, 2375-2388.

[109] Turq, P., Lantelme, F. and Friedman, H.L. Brownian dynamics: its application to ionic solutions. *J. Chem. Phys.* **1977**, *66*, 3039-3044.

[110] Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **1984**, *81*, 511-519.

[111] Hoover, W.G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* **1985**, *31*, 1695-1697.

[112] Martyna, G.J., Klein, M.L. and Tuckerman, M. Nose-Hoover Chains - the Canonical Ensemble Via Continuous Dynamics. *J. Chem. Phys.* **1992**, *97*, 2635-2643.

[113] Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A. and Haak, J.R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684-3690.

[114] Parrinello, M. and Rahman, A. Crystal Structure and Pair Potentials: a Molecular-Dynamics Study. *Phys. Rev. Lett.* **1980**, *45*, 1196-1199.

[115] Parrinello, M. and Rahman, A. Polymorphic Transitions in Single-Crystals - a New Molecular-Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182-7190.

[116] Leach, A. *Molecular modelling: principles and applications*. 2009: Pearson Prentice Hall.

[117] Hohenberg, P. and Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev. B: Condens. Matter* **1964**, *136*, B864.

[118] Kohn, W. and Sham, L.J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, 1133.

[119] Parr, R.G. and Yang, W. *Density-functional theory of atoms and molecules*. 1989, US: Oxford University Press.

[120] Carr, W.J. Energy, Specific Heat, and Magnetic Properties of Low-Density Electron Gas. *Phys. Rev.* **1961**, *122*, 1437.

[121] Carr, W.J. and Maradudin, A.A. Ground-State Energy of High-Density Electron Gas. *Phys. Rev.* **1964**, *133*, A371.

[122] Ceperley, D.M. and Alder, B.J. Ground-State of the Electron-Gas by a Stochastic Method. *Phys. Rev. Lett.* **1980**, *45*, 566-569.

[123] Vosko, S.H., Wilk, L. and Nusair, M. Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin-Density Calculations - a Critical Analysis. *Can. J. Phys.* **1980**, *58*, 1200-1211.

[124] Perdew, J.P. and Wang, Y. Accurate and Simple Analytic Representation of the Electron-Gas Correlation-Energy. *Phys. Rev. B: Condens. Matter* **1992**, *45*, 13244-13249.

[125] Becke, A.D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098-3100.

[126] Lee, C., Yang, W. and Parr, R.G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B: Condens. Matter* **1988**, *37*, 785-789.

[127] Becke, A.D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648-5652.

[128] Zhao, Y. and Truhlar, D.G. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *J. Chem. Phys.* **2006**, *125*, 194101.

[129] Zhao, Y. and Truhlar, D.G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120*, 215-241.

[130] Slater, J.C. Atomic shielding constants. *Phys. Rev.* **1930**, *36*, 0057-0064.

[131] Boys, S.F. Electronic Wave Functions .1. A General Method of Calculation for the Stationary States of Any Molecular System. *Proc. R. Soc. Lond. A Math. Phys. Sci.* **1950**, *200*, 542-554.

[132] Hehre, W.J., Stewart, R.F. and Pople, J.A. Self-Consistent Molecular-Orbital Methods .I. Use of Gaussian Expansions of Slater-Type Atomic Orbitals. *J. Chem. Phys.* **1969**, *51*, 2657.

[133] Ashcroft, N.W. and Mermin, N.D. *Solid state physics*. 1976, Philadelphia: Saunders College Publishing.

[134] Heine, V. *The pseudopotential concept*, in *Solid State Physics*. F.S. Henry Ehrenreich, and David, T. Vol. Volume 24. 1970: Academic Press. p. 1-36.

[135] Ihm, J. Total Energy Calculations in Solid-State Physics. *Rep. Prog. Phys.* **1988**, *51*, 105-142.

[136] Pickett, W.E. Pseudopotential Methods in Condensed Matter Applications. *Comput. Phys. Rep.* **1989**, *9*, 115-197.

[137] Troullier, N. and Martins, J.L. Efficient pseudopotentials for plane-wave calculations. *Phys. Rev. B: Condens. Matter* **1991**, *43*, 1993-2006.

[138] Car, R. and Parrinello, M. Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.* **1985**, *55*, 2471-2474.

[139] Warshel, A. and Levitt, M. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **1976**, *103*, 227-249.

[140] Dal Peraro, M., Ruggerone, P., Raugei, S., Gervasio, F.L. and Carloni, P. Investigating biological systems using first principles Car-Parrinello molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2007**, *17*, 149-156.

[141] Röthlisberger, U. and Carloni, P. *Simulations of enzymatics systems: perspectives from Car-Parrinello molecular dynamics simulations*, in *Theoretical and computational chemistry*. L.A. Erikson, Politzer, P., and Maksić, Z. Vol. 9. 2001, Amsterdam: Elservier. p. 215-251.

[142] Senn, H.M. and Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem. Int. Ed.* **2009**, *48*, 1198-1229.

[143] Rovira, C. The description of electronic processes inside proteins from Car-Parrinello molecular dynamics: chemical transformations. *WIREs Comput. Mol. Sci.* **2013**, *3*, 393-407.

[144] Laio, A., VandeVondele, J. and Rothlisberger, U. A Hamiltonian electrostatic coupling scheme for hybrid Car–Parrinello molecular dynamics simulations. *J. Chem. Phys.* **2002**, *116*, 6941-6947.

[145] Laio, A., VandeVondele, J. and Rothlisberger, U. D-RESP: Dynamically Generated Electrostatic Potential Derived Charges from Quantum Mechanics/Molecular Mechanics Simulations. *J. Phys. Chem. B* **2002**, *106*, 7300-7307.

[146] Singh, U.C. and Kollman, P.A. A Combined Abinitio Quantum-Mechanical and Molecular Mechanical Method for Carrying out Simulations on Complex Molecular-Systems - Applications to the Ch3cl + Cl- Exchange-Reaction and Gas-Phase Protonation of Polyethers. *J. Comput. Chem.* **1986**, *7*, 718-730.

[147] Zhang, Y.K., Lee, T.S. and Yang, W.T. A pseudobond approach to combining quantum mechanical and molecular mechanical methods. *J. Chem. Phys.* **1999**, *110*, 46-54.

[148] Jacquot, Y., Cleeren, A., Laios, I., Yan, M., Boulahdour, A., Bermont, L., Refouvelet, B., Adessi, G., Leclercq, G. and Xicluna, A. Pharmacological profile of 6,12-dihydro-3-methoxy-1-benzopyrano[3,4-b] [1,4]benzothiazin-6-one, a novel human estrogen receptor agonist. *Biol. Pharm. Bull.* **2002**, *25*, 335-341.

[149] Carter, E.A., Ciccotti, G., Hynes, J.T. and Kapral, R. Constrained reaction coordinate dynamics for the simulation of rare events. *Chem. Phys. Lett.* **1989**, *156*, 472-477.

[150] Sprik, M. and Ciccotti, G. Free energy from constrained molecular dynamics. *J. Chem. Phys.* **1998**, *109*, 7737-7744.

[151] Straatsma, T.P. and Mccammon, J.A. Multiconfiguration Thermodynamic Integration. *J. Chem. Phys.* **1991**, *95*, 1175-1188.

[152] Bucko, T. Ab initio calculations of free-energy reaction barriers. *J. Phys.: Condens. Matter* **2008**, *20*, 064211.

[153] Mulders, T., Kruger, P., Swegat, W. and Schlitter, J. Free energy as the potential of mean constraint force. *J. Chem. Phys.* **1996**, *104*, 4869-4870.

[154] Boero, M., Tateno, M., Terakura, K. and Oshiyama, A. Double-Metal-Ion/Single-Metal-Ion Mechanisms of the Cleavage Reaction of Ribozymes: First-Principles Molecular Dynamics Simulations of a Fully Hydrated Model System. *J. Chem. Theory Comput.* **2005**, *1*, 925-934.

[155] Sgrignani, J. and Magistrato, A. The structural role of Mg2+ ions in a class I RNA polymerase ribozyme: a molecular simulation study. *J. Phys. Chem. B* **2012**, *116*, 2259-2268.

[156] Rosta, E., Nowotny, M., Yang, W. and Hummer, G. Catalytic mechanism of RNA backbone cleavage by ribonuclease H from quantum mechanics/molecular mechanics simulations. *J. Am. Chem. Soc.* **2011**, *133*, 8934-8941.

[157] Ganguly, A., Thaplyal, P., Rosta, E., Bevilacqua, P.C. and Hammes-Schiffer, S. Quantum mechanical/molecular mechanical free energy simulations of the self-cleavage reaction in the hepatitis delta virus ribozyme. *J. Am. Chem. Soc.* **2014**, *136*, 1483-1496.

[158] Wong, K.Y., Lee, T.S. and York, D.M. Active participation of Mg ion in the reaction coordinate of RNA self-cleavage catalyzed by the hammerhead ribozyme. *J. Chem. Theory Comput.* **2011**, *7*, 1-3.

[159] Mlynsky, V., Walter, N.G., Sponer, J., Otyepka, M. and Banas, P. The role of an active site Mg(2+) in HDV ribozyme self-cleavage: insights from QM/MM calculations. *Phys. Chem. Chem. Phys.* **2015**, *17*, 670-679.

[160] Zhang, S., Ganguly, A., Goyal, P., Bingaman, J.L., Bevilacqua, P.C. and Hammes-Schiffer, S. Role of the active site guanine in the glmS ribozyme self-cleavage mechanism: quantum mechanical/molecular mechanical free energy simulations. *J. Am. Chem. Soc.* **2015**, *137*, 784-798.

[161] Pechlaner, M., Donghi, D., Zelenay, V. and Sigel, R.K. Protonation-Dependent Base Flipping at Neutral pH in the Catalytic Triad of a Self-Splicing Bacterial Group II Intron. *Angew. Chem. Int. Ed. Engl.* **2015**, *54*, 9687-9690.

[162] Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. and Klein, M.L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926-935.

[163] Perez, A., Marchan, I., Svozil, D., Sponer, J., Cheatham, T.E., III, Laughton, C.A. and Orozco, M. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* **2007**, *92*, 3817-3829.

[164] Zgarbova, M., Otyepka, M., Sponer, J., Mladek, A., Banas, P., Cheatham, T.E., 3rd and Jurecka, P. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.* **2011**, *7*, 2886-2902.

[165] Aqvist, J. Ion-water interaction potentials derived from free energy perturbation simulations. *J. Phys. Chem.* **1990**, *94*, 8021–8024.

[166] Saxena, A. and Sept, D. Multisite Ion Models That Improve Coordination and Free Energy Calculations in Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2013**, *9*, 3538-3542.

[167] Allner, O., Nilsson, L. and Villa, A. Magnesium Ion-Water Coordination and Exchange in Biomolecular Simulations. *J. Chem. Theory Comput.* **2012**, *8*, 1493-1502.

[168] Li, P., Roberts, B.P., Chakravorty, D.K. and Merz, K.M., Jr. Rational Design of Particle Mesh Ewald Compatible Lennard-Jones Parameters for +2 Metal Cations in Explicit Solvent. *J. Chem. Theory Comput.* **2013**, *9*, 2733-2748.

[169] Oelschlaeger, P., Klahn, M., Beard, W.A., Wilson, S.H. and Warshel, A. Magnesium-cationic dummy atom molecules enhance representation of DNA polymerase beta in molecular dynamics simulations: improved accuracy in studies of structural features and mutational effects. *J. Mol. Biol.* **2007**, *366*, 687-701.

[170] D.A. Case, T.A.D., T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R., Luo, R.C.W., W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra,, J. Swails, A.W.G., I. Kolossváry, K.F.Wong, F. Paesani, J. Vanicek, R.M.Wolf, J. Liu,, X. Wu, S.R.B., T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G., Cui, D.R.R., D.H. Mathews, M.G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. and Luchko, S.G., A. Kovalenko, and P.A. Kollman, *AMBER 12*. 2012: University of California, San Francisco.

[171] Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G.A., et al., *Gaussian 09*. 2009, Gaussian, Inc.: Wallingford, CT, USA.

[172] *CPMD*. Copyright für Festkörperforschung Stuttgart 1997-2001, Copyright IBM Corp 1990-2015, http://www.cpmd.org/

[173] Kleinman, L. and Bylander, D.M. Efficacious Form for Model Pseudopotentials. *Phys. Rev. Lett.* **1982**, *48*, 1425-1428.

[174] Nosé, S. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **1984**, *52*, 255-268.

[175] Martyna, G.J. and Tuckerman, M.E. A reciprocal space based method for treating long range interactions in ab initio and force-field-based calculations in clusters. *J. Chem. Phys.* **1999**, *110*, 2810-2821.

[176] Improta, R., Barone, V., Scalmani, G. and Frisch, M.J. A state-specific polarizable continuum model time dependent density functional theory method for excited state calculations in solution. *J. Chem. Phys.* **2006**, *125*, 054103.

[177] Boero, M., Ikeda, T., Ito, E. and Terakura, K. Hsc70 ATPase: an insight into water dissociation and joint catalytic role of K+ and Mg2+ metal cations in the hydrolysis reaction. *J. Am. Chem. Soc.* **2006**, *128*, 16798-16807.

[178] Boero, M., Terakura, K. and Tateno, M. Catalytic role of metal ion in the selection of competing reaction paths: a first principles molecular dynamics study of the enzymatic reaction in ribozyme. *J. Am. Chem. Soc.* **2002**, *124*, 8949-8957.

[179] Lopez-Canut, V., Roca, M., Bertran, J., Moliner, V. and Tunon, I. Theoretical study of phosphodiester hydrolysis in nucleotide pyrophosphatase/phosphodiesterase. Environmental effects on the reaction mechanism. *J. Am. Chem. Soc.* **2010**, *132*, 6955-6963.

[180] Nishino, T. and Morikawa, K. Structure and function of nucleases in DNA repair: shape, grip and blade of the DNA scissors. *Oncogene* **2002**, *21*, 9022-9032.

[181] Akola, J. and Jones, R.O. Density functional calculations of ATP systems. 2. ATP hydrolysis at the active site of actin. *J. Phys. Chem. B* **2006**, *110*, 8121-8129.

[182] Imhof, P., Fischer, S. and Smith, J.C. Catalytic mechanism of DNA backbone cleavage by the restriction enzyme EcoRV: a quantum mechanical/molecular mechanical analysis. *Biochemistry* **2009**, *48*, 9061-9075.

[183] Ditzler, M.A., Otyepka, M., Sponer, J. and Walter, N.G. Molecular Dynamics and Quantum Mechanics of RNA: Conformational and Chemical Change We Can Believe In. *Acc. Chem. Res.* **2010**, *43*, 40-47.

[184] Marcia, M. and Pyle, A.M. Principles of ion recognition in RNA: insights from the group II intron structures. *RNA* **2014**, *20*, 516-527.

[185] Hermann, T., Auffinger, P., Scott, W.G. and Westhof, E. Evidence for a hydroxide ion bridging two magnesium ions at the active site of the hammerhead ribozyme. *Nucleic Acids Res.* **1997**, *25*, 3421-3427.

[186] Casalino, L. and Magistrato, A. Structural, dynamical and catalytic interplay between Mg2+ ions and RNA. Vices and virtues of atomistic simulations. *Inorg. Chim. Acta* **2016**, *452*, 73-81.

[187] Zheng, H.P., Shabalin, I.G., Handing, K.B., Bujnicki, J.M. and Minor, W. Magnesium-binding architectures in RNA crystal structures: validation, binding preferences, classification and motif detection. *Nucleic Acids Res.* **2015**, *43*, 3789-3801.

[188] Petrov, A.S., Bowman, J.C., Harvey, S.C. and Williams, L.D. Bidentate RNA-magnesium clamps: on the origin of the special role of magnesium in RNA folding. *RNA* **2011**, *17*, 291-297.

[189] Kowerko, D., Konig, S.L., Skilandat, M., Kruschel, D., Hadzic, M.C., Cardo, L. and Sigel, R.K. Cation-induced kinetic heterogeneity of the intron-exon recognition in single group II introns. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 3403-3408.

[190] Perez, A., Luque, F.J. and Orozco, M. Frontiers in molecular dynamics simulations of DNA. *Acc. Chem. Res.* **2012**, *45*, 196-205.

[191] Vidossich, P. and Magistrato, A. QM/MM molecular dynamics studies of metal binding proteins. *Biomolecules* **2014**, *4*, 616-645.

[192] Gresh, N., Sponer, J.E., Spackova, N., Leszczynski, J. and Sponer, J. Theoretical study of binding of hydrated Zn(II) and Mg(II) cations to 5 '-guanosine monophosphate. Toward polarizable molecular mechanics for DNA and RNA. *J. Phys. Chem. B* **2003**, *107*, 8669-8681.

[193] Sponer, J.E., Sobalik, Z., Leszczynski, J. and Wichterlova, B. Effect of metal coordination on the charge distribution over the cation binding sites of zeolites. A combined experimental and theoretical study. *J. Phys. Chem. B* **2001**, *105*, 8285-8290.

[194] Yu, H.B., Whitfield, T.W., Harder, E., Lamoureux, G., Vorobyov, I., Anisimov, V.M., MacKerell, A.D. and Roux, B. Simulating Monovalent and Divalent Ions in Aqueous Solution Using a Drude Polarizable Force Field. *J. Chem. Theory Comput.* **2010**, *6*, 774-786.

[195] Kumar, M., Simonson, T., Ohanessian, G. and Clavaguera, C. Structure and Thermodynamics of Mg: Phosphate Interactions in Water: A Simulation Study. *Chemphyschem* **2015**, *16*, 658-665.

[196] Panteva, M.T., Giambasu, G.M. and York, D.M. Comparison of Structural, Thermodynamic, Kinetic and Mass Transport Properties of Mg2+ Ion Models Commonly used in Biomolecular Simulations. *J. Comput. Chem.* **2015**, *36*, 970-982.

[197] Saxena, A. and Sept, D. Multisite Ion Models That Improve Coordination and Free Energy Calculations in Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2013**, *9*, 3538-3542.

[198] Oelschlaeger, P., Klahn, M., Beard, W.A., Wilson, S.H. and Warshel, A. Magnesium-cationic dummy atom molecules enhance representation of DNA polymerase beta in molecular dynamics simulations: Improved accuracy in studies of structural features and mutational effects. *J. Mol. Biol.* **2007**, *366*, 687-701.

[199] Duarte, F., Bauer, P., Barrozo, A., Amrein, B.A., Purg, M., Aqvist, J. and Kamerlin, S.C.L. Force Field Independent Metal Parameters Using a Nonbonded Dummy Model. *J. Phys. Chem. B* **2014**, *118*, 4351-4362.

[200] Li, P.F. and Merz, K.M. Taking into Account the Ion-Induced Dipole Interaction in the Nonbonded Model of Ions. *J. Chem. Theory Comput.* **2014**, *10*, 289-297.

[201] Panteva, M.T., Giambasu, G.M. and York, D.M. Force Field for Mg(2+), Mn(2+), Zn(2+), and Cd(2+) Ions That Have Balanced Interactions with Nucleic Acids. *J. Phys. Chem. B* **2015**, *119*, 15460-15470.

[202] Carloni, P., Rothlisberger, U. and Parrinello, M. The role and perspective of ab initio molecular dynamics in the study of biological systems. *Acc. Chem. Res.* **2002**, *35*, 455-464.

[203] Palermo, G., Stenta, M., Cavalli, A., Dal Peraro, M. and De Vivo, M. Molecular Simulations Highlight the Role of Metals in Catalysis and Inhibition of Type II Topoisomerase. *J. Chem. Theory Comput.* **2013**, *9*, 857-862.

[204] Lee, T.S., Silva-Lopez, C., Martick, M., Scott, W.G. and York, D.M. Insight into the role of Mg2+ in hammerhead ribozyme catalysis from x-ray crystallography and molecular dynamics simulation. *J. Chem. Theory Comput.* **2007**, *3*, 325-327.

[205] Kapral, G.J., Jain, S., Noeske, J., Doudna, J.A., Richardson, D.C. and Richardson, J.S. New tools provide a second look at HDV ribozyme structure, dynamics and cleavage. *Nucleic Acids Res.* **2014**, *42*, 12833-12846.

[206] Maurer, P., Laio, A., Hugosson, H.W., Colombo, M.C. and Rothlisberger, U. Automated parametrization of biomolecular force fields from quantum mechanics/molecular mechanics (QM/MM) simulations through force matching. *J. Chem. Theory Comput.* **2007**, *3*, 628-639.

[207] Aqvist, J. Ion Water Interaction Potentials Derived from Free-Energy Perturbation Simulations. *J. Phys. Chem.* **1990**, *94*, 8021-8024.

[208] Allner, O., Nilsson, L. and Villa, A. Magnesium Ion-Water Coordination and Exchange in Biomolecular Simulations. *J. Chem. Theory Comput.* **2012**, *8*, 1493-1502.

[209] Li, P.F., Roberts, B.P., Chakravorty, D.K. and Merz, K.M. Rational Design of Particle Mesh Ewald Compatible Lennard-Jones Parameters for +2 Metal Cations in Explicit Solvent. *J. Chem. Theory Comput.* **2013**, *9*, 2733-2748.

[210] Joung, I.S. and Cheatham, T.E., III. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **2008**, *112*, 9020-9041.

[211] Ryckaert, J.-P., Ciccotti, G. and Berendsen, H.J.C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327-341.

[212] Bergonzo, C., Henriksen, N.M., Roe, D.R. and Cheatham, T.E., III. Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields. *RNA* **2015**, *21*, 1578-1590.

[213] Zgarbova, M., Sponer, J., Otyepka, M., Cheatham, T.E., III, Galindo-Murillo, R. and Jurecka, P. Refinement of the Sugar-Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *J. Chem. Theory Comput.* **2015**, *11*, 5723-5736.

[214] Chen, A.A. and Garcia, A.E. High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 16820-16825.

[215] Bonomi, M., Branduardi, D., Bussi, G., Camilloni, C., Provasi, D., Raiteri, P., Donadio, D., Marinelli, F., Pietrucci, F., Broglia, R.A., et al. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* **2009**, *180*, 1961-1972.

[216] Misra, V.K., Hecht, J.L., Yang, A.S. and Honig, B. Electrostatic contributions to the binding free energy of the lambdacI repressor to DNA. *Biophys. J.* **1998**, *75*, 2262-2273.

[217] Li, X. and Frisch, M.J. Energy-Represented Direct Inversion in the Iterative Subspace within a Hybrid Geometry Optimization Method. *J. Chem. Theory Comput.* **2006**, *2*, 835-839.

[218] Reed, A.E., Curtiss, L.A. and Weinhold, F. Intermolecular Interactions from a Natural Bond Orbital, Donor-Acceptor Viewpoint. *Chem. Rev.* **1988**, *88*, 899-926.

[219] E. D. Glendening, J.K.B., A. E. Reed, J. E. Carpenter, J. A. Bohmann, C. M. Morales, C. R. Landis, and F. Weinhold, *NBO 6.0*. 2013: Theoretical Chemistry Institute, University of Wisconsin, Madison.

[220] Tang, W., Sanville, E. and Henkelman, G. A grid-based Bader analysis algorithm without lattice bias. *J. Phys.: Condens. Matter* **2009**, *21*, 084204.

[221] Bergonzo, C., Hall, K.B. and Cheatham, T.E., III. Divalent Ion Dependent Conformational Changes in an RNA Stem-Loop Observed by Molecular Dynamics. *J. Chem. Theory Comput.* **2016**, *12*, 3382-3389.

[222] Nowotny, M., Gaidamakov, S.A., Crouch, R.J. and Yang, W. Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. *Cell* **2005**, *121*, 1005-1016.

[223] Zhu, Y. and Chen, S.J. Many-body effect in ion binding to RNA. *J. Chem. Phys.* **2014**, *141*, 055101.

[224] Boero, M., Park, J.M., Hagiwara, Y. and Tateno, M. First principles molecular dynamics study of catalytic reactions of biological macromolecular systems: toward analyses with QM/MM hybrid molecular simulations. *J. Phys.: Condens. Matter* **2007**, *19*, 365217.

[225] Hoskins, A.A. and Moore, M.J. The spliceosome: a flexible, reversible macromolecular machine. *Trends Biochem. Sci.* **2012**, *37*, 179-188.

[226] Matera, A.G. and Wang, Z. A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.* **2014**, *15*, 108-121.

[227] Sali, A. and Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234*, 779-815.

[228] Fiser, A., Do, R.K. and Sali, A. Modeling of loops in protein structures. *Protein Sci.* **2000**, *9*, 1753-1773.

[229] Fiser, A. and Sali, A. ModLoop: automated modeling of loops in protein structures. *Bioinformatics* **2003**, *19*, 2500-2501.

[230] Jamroz, M. and Kolinski, A. Modeling of loops in proteins: a multi-method approach. *BMC Struct. Biol.* **2010**, *10*, 5.

[231] Shen, M.Y. and Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **2006**, *15*, 2507-2524.

[232] Van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E. and Berendsen, H.J.C. GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701-1718.

[233] Casalino, L., Palermo, G., Abdurakhmonova, N., Rothlisberger, U. and Magistrato, A. Development of Site-Specific Mg2+-RNA Force Field Parameters: A Dream or Reality? Guidelines from Combined Molecular Dynamics and Quantum Mechanics Simulations. *J. Chem. Theory Comput.* **2017**, *13*, 340-352.

[234] Pang, Y.P. Novel zinc protein molecular dynamics simulations: Steps toward antiangiogenesis for cancer treatment. *J. Mol. Model.* **1999**, *5*, 196-202.

[235] Bussi, G., Donadio, D. and Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*.

[236] Hess, B., Bekker, H., Berendsen, H.J.C. and Fraaije, J.G.E.M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463-1472.

[237] Humphrey, W., Dalke, A. and Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph. Model.* **1996**, *14*, 33-38.

[238] D.A. Case, R.M.B., D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke,, A.W. Goetz, N.H., S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C., Lin, T.L., R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I., Omelyan, A.O., D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, and R.C. Walker, J.W., R.M. Wolf, X. Wu, L. Xiao and P.A. Kollman, *AMBER 2016*. 2016: University of California, San Francisco.

[239] David, C.C. and Jacobs, D.J. Principal component analysis: a method for determining the essential dynamics of proteins. *Methods Mol. Biol.* **2014**, *1084*, 193-226.

[240] Amadei, A., Linssen, A.B.M. and Berendsen, H.J.C. Essential Dynamics of Proteins. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 412-425.

[241] Lange, O.F. and Grubmuller, H. Can principal components yield a dimension reduced description of protein dynamics on long time scales? *J. Phys. Chem. B* **2006**, *110*, 22842-22852.

[242] Palermo, G., Miao, Y., Walker, R.C., Jinek, M. and McCammon, J.A. Striking Plasticity of CRISPR-Cas9 and Key Role of Non-target DNA, as Revealed by Molecular Simulations. *ACS Cent. Sci.* **2016**, *2*, 756-763.

[243] Baker, N.A., Sept, D., Joseph, S., Holst, M.J. and McCammon, J.A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037-10041.

[244] Dolinsky, T.J., Czodrowski, P., Li, H., Nielsen, J.E., Jensen, J.H., Klebe, G. and Baker, N.A. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **2007**, *35*, W522-W525.

[245] Dolinsky, T.J., Nielsen, J.E., McCammon, J.A. and Baker, N.A. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004**, *32*, W665-W667.

[246] Palermo, G., Ricci, C.G., Fernando, A., Basak, R., Jinek, M., Rivalta, I., Batista, V.S. and McCammon, J.A. Protospacer Adjacent Motif-Induced Allostery Activates CRISPR-Cas9. *J. Am. Chem. Soc.* **2017**.

[247] Martin, A., Schneider, S. and Schwer, B. Prp43 is an essential RNA-dependent ATPase required for release of lariat-intron from the spliceosome. *J. Biol. Chem.* **2002**, *277*, 17743-17750.

[248] Fica, S.M., Oubridge, C., Galej, W.P., Wilkinson, M.E., Bai, X.C., Newman, A.J. and Nagai, K. Structure of a spliceosome remodelled for exon ligation. *Nature* **2017**, *542*, 377-380.

[249] Garrey, S.M., Katolik, A., Prekeris, M., Li, X.N., York, K., Bernards, S., Fields, S., Zhao, R., Damha, M.J. and Hesselberth, J.R. A homolog of lariat-debranching enzyme modulates turnover of branched RNA. *RNA* **2014**, *20*, 1337-1348.