SISSA
Scuola
Internazionale
Superiore di
Studi Avanzati

Physics Area – PhD course in

Physics and Chemistry of Biological Systems

Unraveling the Molecular Mechanism of Splicing through Molecular Dynamics Simulation

Candidate:

Andrea Saltalamacchia

Advisor:

Alessandra Magistrato

Academic Year 2021-22

Table of Contents

Nomenclature

3'SS	3'-Splice Site
5'SS	5'-Splice Site
BP	Branch-Point
CN	Coordination Number
FF/ff	Force Field
G2IR	Group II Intron Ribozyme / Group II Intron
ILS	Intron-Lariat-Spliceosome
LJ	Lennard-Jones
MD	Molecular Dynamics
NTC	NineTeen Complex
pre-mRNA	pre-mature messenger RNA
PDB	Protein Data Bank
RNP	Ribonucleoprotein
snRNP	small nuclear Ribonucleoprotein
SPL	Spliceosome
vdW	van der Waals
FES	Free Energy Surface
Py-tract	Polypyrimidine tract

1 Preface: Summary

The genes architecture as made of intron and exons is now a widely accepted fact and a well-established hypothesis. Indeed, the exons regions of a DNA molecule that code for proteins are not a continuous unitary sequence, but silent intervening segments (introns) that must be eliminated in a process known as pre-mRNA splicing. However, the splicing process is far from being fully understood, as the subtle regulatory mechanisms underlying the generation of premature mRNA, hide a large number of unresolved biological questions. The main keeper of this enigma is represented by the spliceosome, a highly dynamic molecular machinery and the main actor of the splicing process. With recent technological advancement in structural biology, crystallographic techniques together with a remarkable and continuous improvement of computational tools, we are witnessing a breakthrough era that allows us to study and to understand at the atomic-scale key functional aspects of the working mechanisms of large biological macromolecules such as the spliceosome.
In my Ph.D years, I have tackled mechanistic aspects of splicing process, trying to address to three main questions: (i) discovery of small molecules to target specific splicing factors for treatment of splicing-related diseases; (ii) unraveling the molecular mechanism at the basis of pre-mRNAs recognition and splicing fidelity; (iii) elucidating the structural and dynamical properties of the spliceosome and its allosteric regulatory networks. This have been achieved by the use of classical molecular dynamics simulations (MD), Metadynamics and Virtual screening simulations.

In **Chapter 2** I introduce you to the biological significance and conservation of the splicing and alternative splicing, how it is used in normal eukaryotic cells, as well as in cancer cells. The impact of deregulated splicing on a plethora of human diseases is also discussed as well. Finally, I will present the structural and molecular biology of the spliceosome machinery, also explaining how it can precisely process different pre-mRNA sequences.

Chapter 3 reports a review of all the computational techniques that I have used in this thesis. Namely, a brief introduction to classical molecular dynamics simulations, virtual screening, enhanced sampling methods and network theory analysis is reported.

Chapter 4 is entirely dedicated to identifying small molecules inhibitors for a particular kinase that is involved in pre-mRNA splicing and that contributes to the migration propensity of Triple Negative Breast Cancer, one of the most aggressive breast cancer types.

Chapter 5 focuses on the early recognition mechanism of specific pre-mRNA sequences by the splicing cofactor U2AF2, that is also involved in the first steps of the spliceosome assembly. The recognition of these sequences represents one of the first pre-mRNA recognition events, and it underlies the alternative splicing of pre-mRNA, directing the spliceosome to generate one specific transcript rather than another, which result into distinct protein isoforms. The effect of cancer-associated mutations on this delicate recognition step is also investigated.

Chapter 6 represents the first attempt at understanding the reshaping properties of the spliceosome and the allosteric signaling underlying it. In this chapter I report a MD simulations study based on the cryo-EM structure of a yeast spliceosome solved at near-atomic-level resolution. In particular, I have investigated the structural and dynamical properties of the spliceosome machinery, making use of network theory in order to trace the information exchange pathways at the basis of the characterized functional motions.

2 Biological Introduction

2.1 Splicing

The genome within a cell is the complete set of genetic material of an organism, comprising the DNA that carries the information to produce different proteins. This information is included into genes and is subjected to several modifications at each step. At the beginning the gene's structure was assumed to be a clean sequence of base pairs exactly and only coding for the proteins to produce. However, even though being correct for many prokaryotes, after having compared the sequence of a mRNA strand and its corresponding nuclear DNA sequence, in the '70s was postulated the theory that some gene fragments have to be removed during the processing of the nuclear "pre-mature" filament of RNA. This theory is now a fact, and the advent of splicing (the process of removal of the non-coding region from genes) contributed to explain protein diversity in eukaryotes [1], [2]. Pre-mature messenger RNA (pre-mRNA) splicing occurs between transcription and protein synthesis is at the crossroad of gene expression, whereby large non-coding nucleotide sequences within a nascent mRNA transcript, called introns, are removed, or spliced out, whereas coding sequences called exons are joined together in a functional mature mRNA filament before the translation into protein occurs [3].

In eukaryotes, the main director of splicing is a complex ribonucleoprotein (RNP) machinery called spliceosome that catalyzes the excision of introns from pre-mRNAs in the nucleus. The spliceosome is made of five small nuclear RNAs (snRNAs) – U1, U2, U4, U5, and U6 – and approximately 150 proteins [4]. Each pre-mRNA splicing cycle comprehends two sequential transesterification reactions, in which the 2'-OH of an invariant adenosine nucleotide in the branch-point (BP) sequence of an intron is brought in spatial proximity to the guanine at the 5'-end of the 5'-splice site (5'SS) extremity, attacking as a nucleophile the phosphate of this guanine to form 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) resulting in a joined exons sequence and releasing the intron lariat (Figure 2.1) The whole catalytic reaction is coadjuvated by two magnesium ions that coordinate the reagents and stabilize the partial negative charges of the intermediate states. [4]

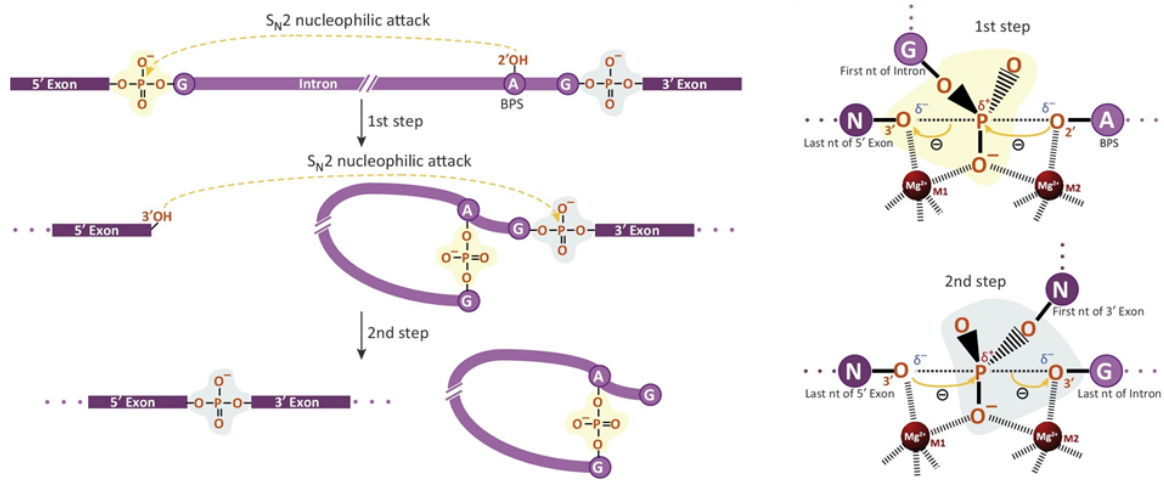


Figure 2.1 The two steps splicing reaction of pre-mRNA and the stabilization of transition states by magnesium (Mg^{2+}) ions proposed for the two phosphoryl-transfer reactions. The figure was adapted from [4], Copyright 2021, Elsevier.

2.1.1 Alternative Splicing and its regulation

In higher eukaryotes, different combinations of exons can be included within a mature mRNA cutting out introns and exons in alternative ways through a process known as alternative splicing [5] (Figure 2.2) obtaining thus mRNAs and different protein products from a single gene. This has provided eukaryotes an evolutionary advantage over other organisms, adding another layer of complexity for gene regulation. This yields a more various protein production from the same number of genes and expands the information content [6]. A nascent mRNA sequence can thus undergo constitutive splicing when all exons are included in the mature transcripts and all introns are spliced out (A), or alternative splicing, when not all splice sites are used, leading to different outcomes e.g. skipping of exons (when also one or more exons are spliced out), introns retention (when an intron is not

splice but it is retained in mRNA) in the mature transcripts or mutually exclusion of exons (B). Combinations of the different types of events are also possible, giving rise to complex events. Alternatively spliced mRNAs (isoforms) can affect either at the protein level, as translated proteins have key functional differences, e.g. the presence or absence of protein domains as well as unstructured polypeptide regions important for protein-protein interactions [7]–[9], but also at expression level, affecting the translation efficiency through differential binding of RNA-binding proteins (RBPs) [10].

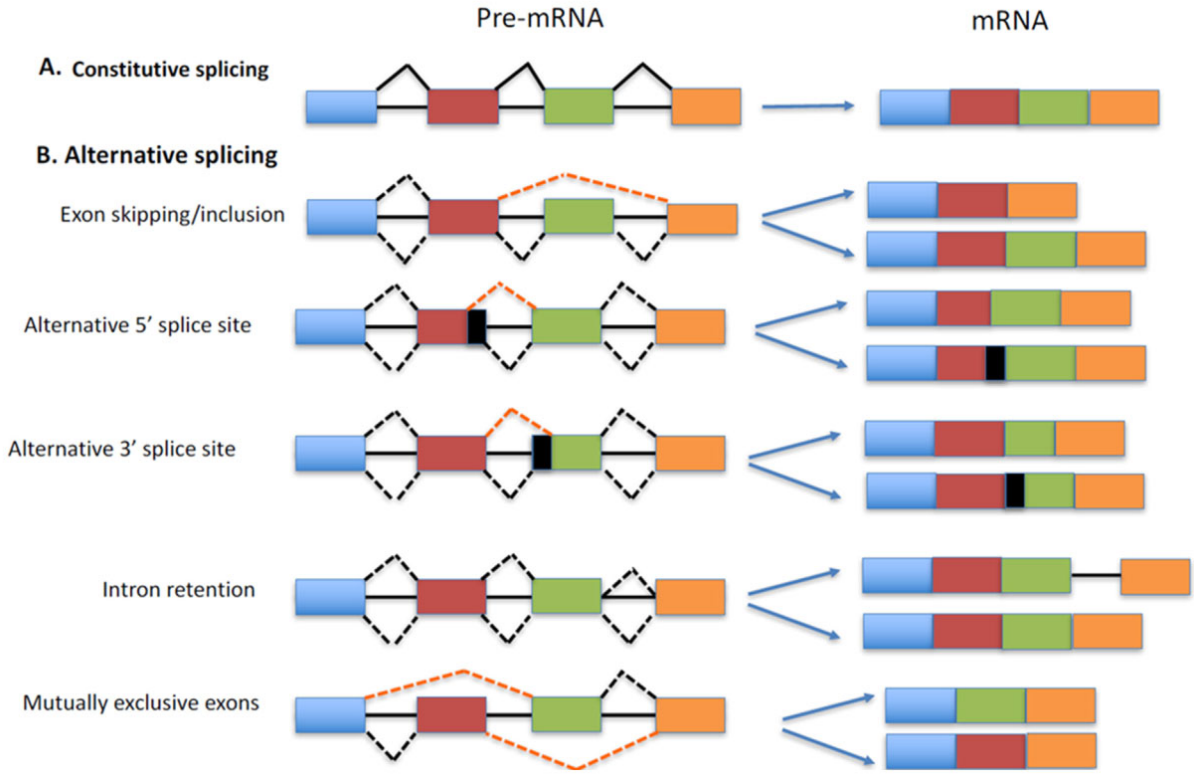


Figure 2.2 Different types of splicing. The constitutive splicing (A) occurs when all introns are spliced and all exons are joined together, while in B different alternative splicing events are shown. The figure was adapted from [11] .

Thus, these examples show how alternative splicing constitutes a sophisticated step of gene regulation, making possible to meet the different gene regulation needs of cells that belong to distinct tissues or different development stages but that share the same genome, through changes within the transcriptome and proteome by transcriptional and post-transcriptional mechanisms[12], [13].

2.1.2 A conservative process

The splicing process has been shown to be highly conserved among very different organisms from viral sequences [1], [14], to eukaryotes, demonstrating its paramount importance. Indeed, spliceosomal introns of eukaryotes' genomes, that make use of the spliceosome complex to be removed, evolved from group II self-splicing introns of bacteria, mitochondria and chloroplasts, where introns removal is catalyzed by the RNA molecule itself [1], [14]. The splicesome and the group II intron ribozymes share the same chemical steps and similar arrangement of RNAs and magnesium ions at their catalytic cores [4], [15], [16]).
Within an organism, tissue diversity is characterized by different gene expression patterns for genes belonging to different tissue cells. Interestingly, when comparing transcriptomic profiles of organs' tissues among different organisms, it emerges that alternative splicing triggers the highest complexity in primates, pointing out its role during evolution and highlighting its importance in complex eukaryotes [17], [18].
Consistently, even though human and *C. Elegans* worms genomes have quite similar numbers of genes (around 20'000), in human the alternative splicing is intensively used compared to *C. Elegans*, explaining in part the higher human complexity [19]. Indeed, more than 95% of human genes can undergo alternative splicing [20], [21]. As such, alternative splicing is the key to understand why evolution exploits introns to increase the complexity of gene expression and generating a fascinating and wide variety of different features among organisms.

2.1.3 The spliceosome and the splicing cycle

The spliceosome is the protein-directed molecular machinery responsible for RNA splicing [22], [23]. It is a complex and dynamic RNA-protein apparatus that assemble over pre-mRNA on each intron prior to splicing, removing introns and joining exons together. During its assembly and catalysis, snRNAs and proteins are combined together into small nuclear ribonucleoproteins (snRNP), and large number of protein-protein, protein-RNA and RNA-RNA interactions are orchestrated to facilitate compositional and conformational rearrangements along the splicing cycle, which are instrumental to bring in close proximity

the reactive groups of the pre-mRNA for catalysis (Figure 1.4) [24]. This composite protein-RNA enzyme has to correctly recognize the splicing sites (5'SS and 3'SS) with high precision to assemble itself and cut the intron-exon boundary at the right place. Indeed, even a single error of one base would disrupt the open reading frame of the pre-mRNA, leading to the production of a functionally altered mRNA. This is a challenging task since a typical human gene is composed by 8 exons with an average of 145 nucleotides (nt) in length, whereas the introns are instead usually 10 times larger [25] . The spliceosome is composed by five snRNPs (U1, U2, U4, U5 and U6), each containing one functional RNA molecule (U1, U2, U4, U5 and U6 snRNAs) and several proteins. The spliceosome assembly begins with the recognition of 5'SS sequences by the U1 snRNP through a base-pairing of U1 snRNA and the 5' SS GURAGU consensus sequence at the intron/exon boundary, while the other important initiation sites, the branch point (BP) sequence with an invariant adenine base, the poly pyrimidine (Py-tract) tract and the AG dinucleotide at 3'SS, are recognized and bound by the proteins SF1, U2AF65 (or U2AF2) and U2AF35 (or U2AF1), respectively, leading to the formation of the early spliceosomal E complex (Figure 2.3) [4], [26].
Subsequently, the initial binding helps the recruitment of U2 snRNP, which through base pairing of its U2 snRNA with the BP sequence [27], displaces the SF1 forming the complex A. These steps are crucial for the regulation of alternative splicing [28], [29]. The pre-assembled U4/U6.U5 tri-snRNP is then recruited to form the fully assembled, but still catalytically inactive, "B complex". The spliceosome activation goes through drastic conformational and compositional changes leading to U1 snRNP dissociation from the 5'SS by the ATP-dependent action of helicase Prp28, while the helicase Brr2 promotes the unwinding of the U4/U6 snRNA duplex, leading to the dissociation of U4 snRNP. Prp8 protein, part of U5 snRNP is the principal actor in coordinating the helicases' actions and activation of spliceosome [30] leading to the activated B complex (Bact). Finally, one of the most important conformational change involves U2 and U6 snRNAs, which remodel to base-pair together, thus generating a well-formed active site, stabilized by two other groups of proteins known as NTC (NineTeen Complex) and NTC-related (NTCR) proteins.

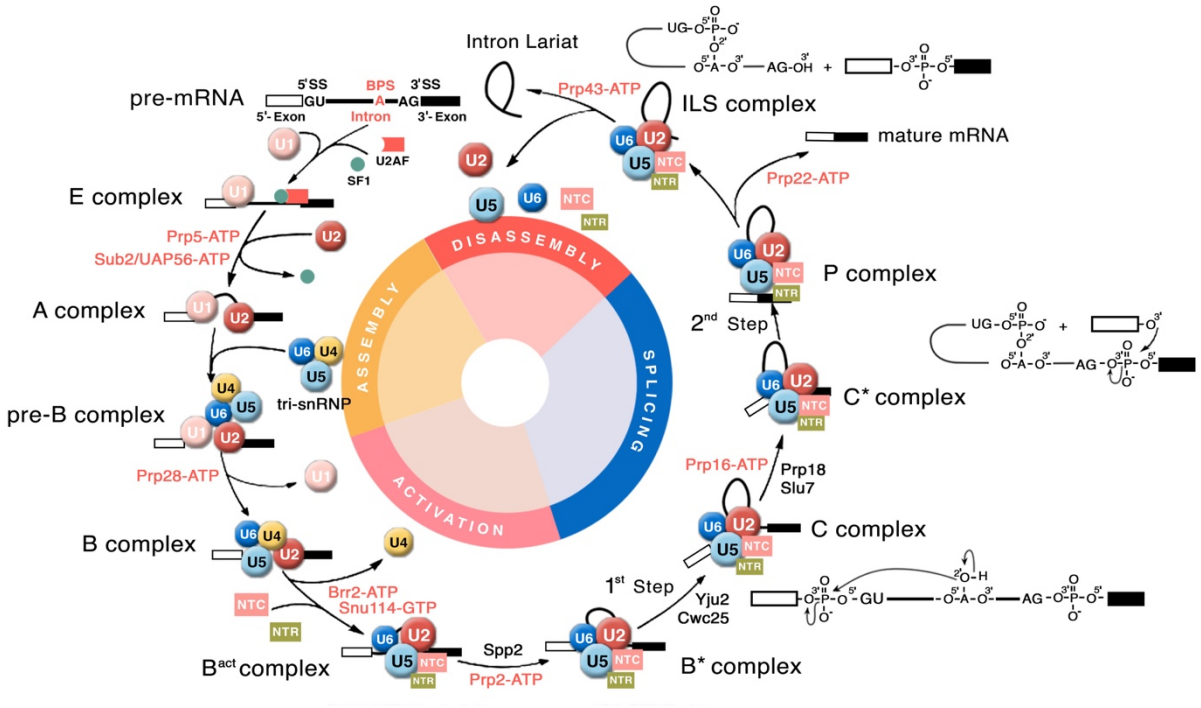


Figure 2.3 The pre-mRNA splicing cycle. The ten complexes are subdivided in four splicing phases, assembly, activation, splicing and disassembly. The snRNPs are displayed in circles while exon and intron sequences are indicated by boxes and lines, respectively. The stages at which ATPases/helicases act to facilitate conformational changes are indicated in red. The figure was adapted from [31]*, Copyright 2021, American Chemical Society.*

This lead to the formation of a catalytically competent B* complex [32]. When the 2'-OH group of the branching adenosine is positioned and oriented in such a way to allow its nucleophilic attack on the scissile phosphate at the 5'SS, the first step reaction occurs, yielding the C complex as product. The latter contains the free 5'-exon and the intron lariat–3'-exon intermediates. After further remodeling, the exons are aligned for the second transesterification reaction in the C* complex, which then results in the ligated exon sequence (mRNA filament) and the intron lariat when the post-catalytic spliceosome (P complex) form at the end of the two-steps reaction [4], [26].

The last steps of the cycle include the release of the spliced mRNA product and the disassembly of the residual intron-lariat-spliceosome ("ILS") complex where the U2, U6 and U5 snRNPs are recycled and the excised intron lariat is degraded [24]. In the last few years, several spliceosome structures have been determined thanks to the recent advancements of

high-resolution cryo-EM technologies [16], [32]–[39], with different structural models corresponding to distinct steps of the splicing reaction.

2.1.3.1 Recognition of Splicing Sites

In the E complex, U1 snRNP recognizes the 5' SS through the base pairing of its U1 snRNA, while the protein complex of splicing factor 1 (SF1) and the large 65 kDa subunit of U2 small nuclear ribonucleoprotein auxiliary factor (U2AF65) binds to the BP sequence and recognizes the Py-tract near the 3' SS, respectively [40]. Instead, the small 35 kDa subunit of U2AF (U2AF35), that also tightly binds to U2AF65, recognizes the AG dinucleotide at the 3' SS [41]–[44]. After this first recognition, the U2 snRNP binds the intronic 3'SS region leading to the pre-spliceosomal complex A. This snRNP includes U2 snRNA and different other proteins, including the complexes Splicing Factor 3A and 3B (SF3A and SF3B). SF3A is composed of three subunits (SF3A1, SF3A2 and SF3A3), while SF3B comprehends at least eight subunits (SF3B1, SF3B2, SF3B3, SF3B4, SF3B5, SF3B6, SF3B7 and SF3B8). Albeit the precise function of most of these proteins is still unclear, [24], [26] SF3B1 is highly conserved among eukaryotes, and it is fundamental for stabilizing U2 snRNP recruitment [26], [45]. The correct recognition of the BP is a key step for the splicing process. This occurs via the branching adenosine (BPA) bulging out of the U2/intron helix [27] so that the SF3B1 (also known as SF3b155) can recruit it [46]. Furthermore, SF3B1 was found to crosslink upstream and downstream of the BPA and to interact with U2AF65 and U2AF35 [47]. Indeed, the BP sequence downstream is used only if it is not too close to the Py-tract [47], meaning that, for an efficient assembly of complex A, the distance between the BP sequence and the Py-tract has an important role.
U2AF65 is required for adaptive recognition of Py-tracts with variations in length and sequence composition, even though the mechanism by which these sequence variations are recognized by U2AF65 remains unclear. The Py-tract, in fact, serves as an important sequence signal to enable both constitutive and alternative pre-mRNA splicing [48]–[50] since it alters 3' SS selection by promoting alternative BP selection [51].

As a matter of fact, progressive deletions leading to shortening of the Py-tract result in the abolishment of lariat formation, spliceosome assembly and splicing [52]–[55]. In spite of its important role in splicing, the Py-tract sequence shows great variability (i.e., different base composition). Indeed, for certain substrates, the Py-tract can accommodate purines, but it might be deleterious to splicing if the length of the pyrimidine tract reduces to nine nucleotides with fewer than five consecutive uridines [54], [56], [57]. In this framework, Py-tracts containing 11 continuous uridines were found to be the strongest binding sequences to U2AF2, making their location relative to the BP sequence and 3' SS not sequential [58]. In contrast, a Py-tract made of five or six continuous uridines requires to be located adjacent to the 3' SS for optimal efficiency of competitive splicing [58].
Along with the recognition sequences described above, additional cis-acting signals can guide the spliceosome to recognize the correct splice sites. These elements can either stimulate (exon splicing enhancers, ESEs and intron splicing enhancers, ISEs) or repress (exon splicing silencers, ESSs and introns splicing silencers, ISSs) the expression of spicing isoforms, by recruiting trans-acting splicing factors that activate or suppress splice site recognition [59], [60]. Among these cis-regulatory elements, ESEs are the most prevalent [61]. These are characterized by purine-rich sequences and serve as binding sites for specific serine/arginine-rich (SR) proteins [62], [63]. The SR proteins regulate splicing by binding RNA sequences through their N-terminal RNA Recognition Motif (RRM) domains and mediating protein-protein interactions to facilitate spliceosome assembly through C-terminal RS domains [64].

2.1.4 Diseases and Cancer

More than 200 human diseases are caused by pre-mRNA splicing aberrations. Indeed, a single nucleotide addition or deletion at the exon joining site may have severe consequences to the protein-coding potential of the resulting mRNA [65]. As a result, splicing defects and mis-regulation are at the origin of many pathologies and these can occur because of either *(i)* Cis-acting mutations on splicing sites which are therefore skipped or not correctly read by the spliceosome, or *(ii)* trans-acting splicing mutations affecting factors that regulate alternative splicing. The first result in a localized effect on the single mRNA, while the latter cause a misregulation of the pre-mRNA splicing of all genes [25].

Spinal Muscular Atrophy (SMA), one splicing-related disease, is a degenerative pathology due to the loss-of-function of *SMN1* gene, coding for SMN protein, which is critical for snRNPs assembly. The homologous *SMN2* gene is not able to compensate for pathological mutations in SMN1 because of a nucleotide difference within SMN2 exon 7 that induces skipping of this exon and leads to the production of a truncated and unstable version of the SMN protein [66], [67].

Another splicing-related disease is the retinitis pigmentosa, a genetic disease consisting in progressive retina degeneration that is caused by mutations within splicing factors genes, mainly PRPF8 [29], the largest and most conserved spliceosomal protein [4], [26], [33], [68].

An example of cis-acting mutation is at the base of the Duchenne Muscular Dystrophy (DMD). This disease is induced by frame-shifting mutations within the dystrophin gene, leading to truncated proteins that are not able to maintain the integrity of muscular fibers and giving rise to the pathological condition [69].

Tumor cells are able to take advantage of the splicing process, hence they often display an altered balance of alternative isoforms that play relevant roles in preventing apoptosis or promoting proliferation and invasion [70]. FAS exon 6 skipping is among the most studied alternative splicing events. Full-length FAS is a membrane protein able to activate the apoptotic cascade when bound from the FAS ligand, FASL, and exon 6 skipping leads to the production of FAS isoforms lacking the transmembrane domain. Thus, not being able to stick into the membrane, this is secreted outside the cell, where it will be able to sequester FASL and therefore inhibit the apoptotic process. Several tumors switch FAS splicing towards the antiapoptotic isoform in order to escape apoptotic stimuli [71]. Many other examples have been reported so far, including the impact of signaling pathways activated in cancer cells on splicing aberrations, and how these splicing aberrations can control a variety of mechanisms, including metastasis, angiogenesis and cell cycle [70], [72]. Also, most of the mutations are heterozygous and mutually exclusive, suggesting that different factors could have some common consequences and that complete loss of the wild-type versions of these factors is deleterious [73]. Remarkably, mutations of the U2AF35, ZRSR2, SRSF2 and SF3B1 splicing factors were found to be associated to myelodysplastic disorders [74]. Strikingly, SF3B1 mutations highly correlate with the presence of ring sideroblasts (RS), i.e. erythroblasts

(erythrocyte precursors) with mitochondrial iron deposits forming perinuclear granules [75]. Mutations on this factor were then described for chronic lymphocytic leukemia (CLL) [76] and solid tumors, including uveal melanoma [77].

Finally many cancer-related mutations were found localized on U2AF65 [78] and showed to cause cystic fibrosis [79], tumors [80]–[82] and myotonic dystrophy [83], and we will focus on some of these U2AF2 mutations in the Chapter 0 of this thesis.

Collectively, these observations show that the RNA affinity of mutated splicing factors is altered and that differential RNA binding can cause sequence-dependent alternative splicing changes, some of which can partially explain the pathologic phenotype [84]. Although research in the field is progressing very fast, many mechanistic questions remain to be answered in order to be able to translate mechanistic knowledge into therapy.

3 Computational Methods

3.1.1 Molecular Dynamics and Integrators

Molecular Dynamics (MD) is a computational tool that permits to compute equilibrium properties of a system, and it represents an atomic-level resolution microscope within a laboratory. The mathematical and theoretical bases come from statistical mechanics and classical physics to connect the micro-scales with macroscopic quantities. In an MD simulation the time evolution of atoms and molecules is calculated by computing the forces acting on each atom and integrating the Newton's equations of motion after having defined starting conditions as temperature, pressure, initial configuration and velocities of atoms. These steps are repeated iteratively until the desired simulation length is reached and averages of observables of interest can be calculated. The particular microscopic state of a set of N atoms can be thus described using the positions $\vec{R} = \{\vec{R}_1, \dots, \vec{R}_N\}$ and momenta $\vec{P} = \{\vec{P}_1, \dots, \vec{P}_N\}$ in a 6N multidimensional phase space (Γ). MD generates a time-evolution trajectory of points in the phase space, leading then to an ensemble, or collection, of points on which it is possible to estimates ensemble average of an observable value for a certain property A as a function of Γ, $A(\Gamma)$:

$$< \boldsymbol{A} >_{\boldsymbol{ens}} = \int \boldsymbol{A}(\boldsymbol{\Gamma})\boldsymbol{\rho}(\boldsymbol{\Gamma})\boldsymbol{d\Gamma} \qquad 3.1$$

where $\rho(\Gamma)$ is the probability distribution function of the collection of points that depends on macroscopic parameters like the number of particles, N, volume V, temperature T and pressure P, defining the thermodynamic state of a system. In the canonical ensemble (NVT), N, V and T are constant, and the probability distribution function has the form of the Boltzmann distribution function.

MD simulations makes an ensemble of points evolve as a function of time integrating the Newton's equations of motion and generating a trajectory of points in the phase space Γ(t).

According to the ergodic hypothesis [85], if the system is evolved for an infinitively long time so that to visit all the states, it is possible to estimate a macroscopic observable as an ensemble average since it will be equal to the average over time. For this reason, the longer is the simulation length, the most this equality is satisfied.

$$\lim_{\tau\to\infty}\langle A(\Gamma)\rangle_\tau = \lim_{\tau\to\infty}\frac{1}{\tau}\int_0^\tau A[\Gamma(t)]dt\,\langle A(\Gamma)\rangle_\tau \qquad 3.2$$

The second Newton's equations of motion treats atoms as point particles:

$$\vec{F}_i = m_i\vec{a}_i \qquad 3.3$$

where F is the force acting on particle *i,* with *m* and a its mass and acceleration i.e., the second derivative of the particle's position with respect to time *t*. These forces are derived from the system's potential energy U($\vec{R}$) and influence the atoms' motion:

$$\overrightarrow{F_i} = -\frac{\partial U(\vec{R})}{\partial\overrightarrow{R_i}} \qquad 3.4$$

Newton's equation defines then the state of a system at time *t* with a set of positions and velocities, where the first are initially taken from experimental data (NMR, crystal structures, cryo-EM) and the second from the Maxwell-Boltzmann probability distribution at a given temperature *T.* The solution to the second order differential equation is given by time discretized numerical algorithms which, after choosing a proper time step δt at the beginning of the simulation (usually in the range of 1-2fs or lower than the fastest atomic vibrational frequency of the system), update the positions and velocities of the particles. One of the simplest integration algorithms, which is also the algorithm used in this thesis, is the so-called Leap Frog algorithm [86], which uses velocities at half-integer time steps to determine the new particles' positions:

$$\vec{v}\left(t+\frac{\Delta t}{2}\right)=\vec{v}\left(t-\frac{\Delta t}{2}\right)+\frac{\vec{F}(t)}{m}\Delta t+O(\Delta t^3) \quad 3.5$$

$$\vec{R}(t+\Delta t)=\vec{R}(t)+\vec{v}\left(t+\frac{\Delta t}{2}\right)\Delta t+O(\Delta t^3) \quad 3.6$$

This algorithm calculates positions and forces at time t + δt, while velocities at time t +1/2δt. As a consequence, kinetic and potential energy are not defined at the same time.

3.1.2 Force Fields

A functional form of the potential, from which the forces acting on the atoms are derived, has been developed and improved through the years to provide a good approximation of the physics and chemistry of the system. This functional form and the set of empirical parameters expressing the potential energy, U_{FF}, in a way to reproduce the properties of molecules, is called force field (FF). The FF describes the atoms as charged point particles, without considering electrons explicitly. Thus, the potential energy of the system can be written as a sum of bonded and non-bonded molecular forces for bonds, angles, dihedrals and interaction of particles:

$$U(r)=\sum_{bonds}U^{str}+\sum_{angles}U^{bend}+\sum_{torsions}U^{dihe}+\sum_{imp-tors}U^{im-dihe}+\sum_{i<j}(U^{LJ}+U^{Coul}) \quad 3.7$$

More in detail the FF can be written as:

$$U^{FF}(r_1,\dots,r_N)=\sum_{bonds}K_r(r-r_0)^2+\sum_{angles}K_\theta(\theta-\theta_0)^2+\sum_{dihedrals}\frac{V_n}{2}[1+\cos(n\phi-\gamma)]+\sum_{i<j}\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}-2\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]+\sum_{i<j}\frac{q_iq_j}{\varepsilon r_{ij}} \quad 3.8$$

Where the bond (1-2) stretching and angle (1-2-3) bending are described by harmonic potentials with k_r and k_θ as the force constants and where the energy increases as the bond

length r or the angle θ deviate from the equilibrium reference values, r_{eq} and θ_{eq}. These equilibrium values are usually derived from structural databases, while the force constants are derived from infrared spectroscopy [87] or quantum chemical calculations. The dihedral term (1-2-3-4), describe the torsion potential energy for the rotation of a bond, where V_n is the dihedral force constant and n, ϕ and γ are the number of barriers, dihedral angle and phase, respectively. Pairs of atoms that are separated by at least three bonds share non-bonded interactions that are described by the short-range Lennard-Jones term and a long-range Coulomb term. The former models the van der Waals (vdW) interactions, with a repulsive term for short interatomic distances due to the Pauli repulsion, and an attractive term to represent the attractive forces of dipole interactions, where ε_{ij} is the depth of the potential well, σ_{ij} is the equilibrium distance at which the potential is zero and r_{ij} the distance between the two particles. The latter term account for the electrostatic interactions tacking place within the system with the long-range Coulomb potential that decay as r_{ij}^{-1}, with q_i and q_j being the partial electric charges of atoms i and j, while ε is the dielectric constant.

Finally, to prevent the system from generating artifacts at the edges of the simulation box (finite size effects) and to mimic the bulk water molecules, periodic boundary conditions (PBC) are adopted, and the system is replicated in all directions as an infinite periodic lattice (periodic images), reproducing an infinite solution around the system, so that when a molecule exits from the simulation box, it enters from the opposite side.
The calculation of the non-bonded interactions is the most expensive and computational demanding part of the MD simulation, and the periodic images complicate this task as these interactions may extend beyond the boundaries of the central cell. For this reason a cutoff distance is applied to limit the number of interactions for the short distance term, while for Coulomb term, that decays much less rapidly with the distance, Particle Ewald Summation method (PME) [88] is used, splitting this interaction in a short term part (that will be included in the cutoff) and a long-range term that will be calculated by Fourier transformation.

3.1.3 Thermostats and Barostats

3.1.3.1 Velocity-Rescaling

Among the possible thermostats that have been developed to perform MD simulation in this thesis we used the velocity rescaling method [89]. This aims at finding a rescaling factor by which the velocities of all the particles are multiplied, obtaining in this manner a total kinetic energy that is equal to the average kinetic energy at the target temperature:

$$\alpha = \sqrt{\frac{\overline{\overline{K}}}{K}} \qquad \overline{K} = \frac{N_f}{2\beta} \tag{3.9}$$

where N_f is the number of degrees of freedom and β is the inverse temperature. This operation is usually performed at a predetermined frequency during equilibration, or when the kinetic energy exceeds the limits of an interval centered around the target value. This choice can lead to a bad sampling of the fluctuations and quantities depending on this may have large errors. To enforce a canonical distribution for the kinetic energy, the target value K_t is selected with a stochastic procedure aimed at obtaining the desired ensemble.
The scaling factor will be then:

$$\alpha = \sqrt{\frac{K_t}{K}} \tag{3.10}$$

where K_t is drawn from the canonical equilibrium distribution for the kinetic energy:

$$\bar{P}(K_t)\, dK_t \propto K_t^{\left(\frac{N_f}{2}-1\right)} e^{-\beta K_t}\, dK_t \tag{3.11}$$

The stochastic dynamics for the kinetic energy K has been chosen in such a way to leave the canonical distribution above invariant and imposing a first order stochastic differential equation in K. Hence the following equation has been proposed to describe this stochastic kinetic dynamics and to generate the correct canonical distribution:

$$dK = (\overline{K} - K)\frac{d\tau}{\tau} + 2\sqrt{\frac{K\overline{K}}{N_f}}\frac{dW}{\sqrt{\tau}} \qquad 3.12$$

Where the first term of the right side of the equation is deterministic, while the second is stochastic and *dW* denotes a Weiner process. The coupling τ parameter has the dimension of a time and determines the time-scale of the thermostat. When a system is far from equilibrium, the deterministic part of the equation dominates and the algorithm leads to fast temperature equilibration.

3.1.3.2 Berendsen Barostat

To simulate an atmospheric pressure in the NPT ensemble, we need to make use of barostats, where, in the same spirit as the temperature coupling, the system can also be coupled to a pressure bath. The Berendsen algorithm [90] is one of the possible algorithms used to this purpose and it has been used in this thesis. This method rescales the coordinates and box vectors at every step, or every *n* steps, with the effect of a first-order kinetic relaxation of the pressure towards a given reference pressure P_0 according to:

$$\frac{dP}{dt} = \frac{P_0 - P}{\tau_p} \qquad 3.13$$

where P is the instantaneous pressure, P_0 is the desired pressure, and τ_P is the barostat relaxation time constant.

This leads to cell size variations, where the MD cell volume is scaled by a factor η, and the coordinates and cell vectors by $\eta^{1/3}$:

$$\eta(t) = 1 - \frac{\Delta t}{\tau_P}\gamma(P_0 - P(t)) \qquad 3.14$$

where γ is the isothermal compressibility of the system.

Berendsen barostat is more suitable for equilibration phases rather than simulation production, as it allows to quickly reach the requested pressure even though it can lead to violent oscillations.

3.1.3.3 Parrinello-Rahman Barostat

The Parrinello-Rahman barostat [91] makes possible to rescale the volume considering the simulation cell as described by the matrix H composed by the three vectors a, b, c as the sides of the cell. Then the volume is given by:

$$V = \det \boldsymbol{H} = \vec{a} \cdot (\vec{b} \times \vec{c}) \tag{3.15}$$

The coordinates of the molecules are now introduced in terms of scaled variables, s_i:

$$r_i = \boldsymbol{H} S_i \tag{3.16}$$

The box vectors as represented by the matrix H obey the matrix equation of motion:

$$\frac{dH^2}{dt^2} = VW^{-1}\boldsymbol{H}'^{-1}(P - P_{ref}) \tag{3.17}$$

where V is the volume of the box, W a matrix parameter determining the strength of the coupling and the matrices P and P_{ref} are the current and the reference pressures, respectively. This method has been used in this thesis after the Berendsen barostat as it maintains canonical ensemble.

3.1.4 Pearson Cross-Correlation

From a MD simulation a trajectory of atomic positions id collected, from which it is possible to analyze the atomic fluctuations and in turn to capture the linear coupling of the motions

between amino acid residues over the trajectory. This is done by calculating the dynamic correlation between the atoms i.e., the degree to which they move together.
The correlation can be quantified starting from the computation of the covariance between the fluctuation of the atoms:

$$c(i,j) =< \Delta r_i \cdot \Delta r_j > \quad 3.18$$

in which Δr_i is the displacement vector of atom i and the angle brackets denote an ensemble average. The cross-correlation coefficient, or normalized covariance, is calculated by:

$$CC(i,j) = \frac{< \Delta r_i \cdot \Delta r_j >}{< \Delta r_i^2 >^{1/2} < \Delta r_j^2 >^{1/2}} \quad 3.19$$

where at the denominator the product between the standard deviations of the two position vectors is used as the normalization factor. The cross-correlation defined in this manner is also known as Pearson's correlation, and it ranges from a value of -1, which indicates a totally anti-correlated motion between two atoms, and a value of +1, which implies positively correlated movement whereby the two atoms moved in the same direction.

3.1.5 Principal Component Analysis

Molecular dynamics trajectories show local and global motion of atoms that makes hard for the human eye to evaluate the protein functional movements. Principal component analysis is a popular instrument employed to filter out this noise provoked by local movements, separating it from the essential and most interesting functional dynamics.
Starting from a 3Nx3N covariance matrix where N is the number of atoms, the analysis performs a diagonalization to identify a set of 3N eigenvectors (Principal Components or PCs) and the associated eigenvalues which describe the modes of the collective motion and their respective amplitudes, i.e., the main directions on which the atoms fluctuate, and the variance of the fluctuation associated to that eigenvector. The eigenvectors are then sorted according to their eigenvalues, giving a rationale to identify the protein movements (PC) that account for most of the variance. Typically, if the first two eigenvalues represent more than

sixty per cent of the variance, we may project the trajectory along these first two eigenvectors and look at the configurations described by this new essential subspace. Projecting the trajectory onto one of these PCs, it is possible to generate a pseudo trajectory as an interpolation between two frames corresponding to the extreme values reached on this component, called essential dynamics, which is a valuable tool to reveal important features of the functional movements of the protein.

3.1.6 Allosteric modulation and network-based algorithms

Enzymes have an active site where the main chemical reactions occur. The main modulation of enzyme's activity goes through the use of substrates/inhibitors that bind to the active site, activating or inhibiting the enzyme function. However, a second and more subtle mechanism can induce the same kind of effect through the binding of a so-called effector molecule to a secondary (non-active) site of the enzyme. This results in a positive or negative allosteric mechanism which modulates the enzyme either by inducing a conformational shift within the enzyme itself, or by changing the dynamic fluctuation around the mean conformation i.e. solely through changes in the protein dynamics [92]. Therefore, a detailed understanding of allosteric signaling and pathway is needed to understand this type of regulatory mechanism, but it can be harsh to obtain at experimental level. Different algorithms have been proposed to explain and predict such subtle changes of protein structure/dynamics resulting from the binding of an allosteric modulator. In the next paragraph I will explain one of these algorithms that has been used in this thesis.

3.1.6.1 WISP Algorithm

The Weighted Implementation of Suboptimal Paths (WISP) [93] method attempts to elucidate the fine allosteric regulation describing the protein as a weighted network of nodes (residues) connected by edges that are built taking into account the contacts and correlations between residues. After having removed rotational and translational movements from a trajectory, this method at first builds a contact map M for the system of N residues. If the residues *i* and *j* have been close (or under a certain cut-off) more than 75% of the trajectory,

the coefficient w_{ij} is set to 1 and an edge will be created in the network, whereas if the two residues are not in contact, the matrix element will be set to 0. Subsequently, atomic cross-correlations (C_{ij}) are calculated using the equation number and are used to weight every edge of the network as the following:

$$W_{ij} = -\log\left(\left|C_{ij}\right|\right) \quad 3.20$$

Through multiplication of the weight and contact matrices, *W* and *M*, we obtain the weighted edges matrix W_{simp} describing a well-defined weighted graph.

The second step is to find a route between the active site and the effector molecule on which the signaling is propagated. A path in this graph is a sequence of linked nodes with a path-length corresponding to the sum of the weights w_{ij} along the route. Short pathways correspond to highly correlated motions of the residues, while long paths correspond to poorly correlated motion. It is possible then to compute the shortest (optimal) path and the first *n* shortest paths (sub-optimal paths) fixing a starting (source) node and an ending node (sink). This calculation is very computationally demanding to obtain suboptimal paths as we have to compute all the path, representing a problem of that scales as *N!,* where *N* is the number of residues. WISP performs the calculation of all Forced Node Paths (FNPs) between the source and sink nodes, i.e. compute the optimal path between the two residues forcing it to pass through a third node n_i. [94], [95] and then repeat this procedure for all nodes n_i in the graph. After fixing an arbitrarily cut-off d_{cutoff} on the maximum length for the desired suboptimal path, if the FNP for node n_i is larger than d_{cutoff}, the node is removed from the graph, setting all the elements in W_{simp} associated to it to 0 to lighten the matrix and reduce the computational problem. Then, bidirectional research of the sub-optimal paths is started from the source and the sink residues, where each path is interrupted as soon as the length exceed the d_{cutoff}. At each step, paths sharing a common node are joined together.

The research ends with a list of all sub-optimal paths and the respective length, from which it is possible to analyze the path length distribution and, most importantly, the node degeneracy associated to a residue, representing the number of time that residue is found in the sub-optimal paths, as these residues are shown to be relevant for biological functions [93].

This method may then be useful to identify the key residues in the protein activity and to explain the differential behavior of distinct residues' motions before and after the allosteric modulation.

3.1.6.2 Community Network Analysis

The Pearson's correlation coefficient is lacking the non-linear contributions to pair correlations, and it is orientation-dependent, i.e. orthogonal correlated motions are completely neglected. A more accurate calculation of motion correlations is, thus, desirable for weighting protein networks [96] based on correlated motions. To this scope one can rely on the mutual information (MI) measure to obtain the generalized correlation coefficients [97]. The MI between two variables (such as the $\vec{r}_i$ and $\vec{r}_j$ position vectors) is defined as:

$$MI\left[\vec{r}_i, \vec{r}_j\right] = H[\vec{r}_i] + H\left[\vec{r}_j\right] - H\left[\vec{r}_i, \vec{r}_j\right] \quad 3.21$$

where $H\left[\vec{r}_i, \vec{r}_j\right]$ is the joint Shannon entropy of the variables and $H[\vec{r}_i]$, $H\left[\vec{r}_j\right]$ are their marginal entropies, providing a direct link between motion correlations and information content. The MI can be conveniently converted into an orientation-independent generalized correlation coefficient ($^{MI}CC_{ij}$, defined between 0 and 1, i.e., from uncorrelated to correlated motions) by:

$$^{MI}CC_{ij} = \left(1 - exp\left(-\frac{2}{d} MI\left[\vec{r}_i, \vec{r}_j\right]\right)\right)^{-\frac{1}{2}} \quad 3.22$$

where d is the dimensionality of the r_i and r_j variables. The calculation of the $^{MI}CC_{ij}$ coefficients for an extremely large system, such as that studied here, is computationally very demanding. Therefore, here we limited the calculation to the linearized version of $^{MI}CC_{ij}$, i.e. $^{LMI}CC_{ij}$, based on the linear mutual information (LMI) measure. This relies on a Gaussian approximation (i.e., the quasi-harmonic approximation to the density of the atomic fluctuations) to reduce the computational cost. This computationally-efficient version of MI

developed by Lange et al.[97], while neglecting the non-linear contributions to the correlation, yet still does not depend on the relative orientation of the atomic fluctuations and it provides an excellent approximation to the generalized correlation coefficients.
These coefficients are then used to weight a communication network of a protein complex [96], [98], [99]. A protein network based on the information exchange between amino acid residues (represented by their Cα atoms) can be constructed considering residues as nodes that are connected by edges, whose lengths is related to their motion correlations [100]. Here, the edge lengths are weighted using the $^{LMI}CC_{ij}$, with the weight, w_{ij}, of the edge connecting nodes *i* and *j*, being calculated as:

$$w_{ij} = -\log {}^{LMI}CC_{ij} \qquad 3.23$$

so that highly correlated pairs of residues are associated to efficient links for information exchange and thus lie at close distances within the (protein) communication graph. In such protein graph, two nodes are considered connected when the distance between any heavy atoms of two residues is lower than 5.0 Å (distance cutoff) for at least 75% of the frames (percentage cutoff) analyzed. These values are chosen according to previous studies on protein RNA complexes [99]. The resulting weighted graph is, then, partitioned into communities using the Edge Betweenness (EB) criterion and the modularity measure [100], [101]. The EB and the node betweenness (NB) are defined as the number of shortest (and thus more relevant) paths passing through that edge (or that node for NB). Namely, the EB (or NB) accounts for the number of times an edge (or a node) acts as a bridge in the communication flow between any pair of nodes of the network. The shortest paths used to determine EB and NB values are computed using the Floyd-Warshall algorithm [102], [103]. The EB is used to partition the network (starting from a single community for the whole system) into multiple communities, using the Girvan-Newmann algorithm. The modularity parameter, defined (between 0 and 1) as the difference in probability of intra- and inter-community connections for a given network division, is adopted to select the optimal division, i.e., the optimum community structure. Such definition of the network provides a coarse-grained and intuitive picture of the complex internal communication network within

the macromolecular system studied here, and it allows the dissection of critical nodes and communication channels.

3.1.7 Metadynamics

Although MD simulations can account for protein motion considering explicit water molecules, they are computationally costly. As such, events of biochemical interest occur on a time scale longer that those spanned in typical MD simulations (10-1000 micros). As a result, enhanced sampling approaches such as metadynamics [104] have become increasingly popular in the last decades for simulating rare events. Metadynamics is a method that adds time dependent bias potential, $V_G(s,t)$, to the real energy landscape of the system enabling to overcome obstacles in the free energy environment as well as to accelerate the sampling of the system. To this end, specific reaction coordinates (so-called collective variables (CVs)) have to be selected, along which the bias potential is added [105].
The bias potential is added with a chosen pace rate τ_G as a sum of gaussians of height *w* and width δ [106]:

$$V_G(S(r),t) = w \sum_{t'=\tau_G,2\tau_G,..} exp\left(-\sum_{\alpha=1}^{d} \frac{(S_\alpha(r) - s_\alpha(t'))^2}{2\delta s_\alpha^2}\right) \quad 3.24$$

Where s(t) is the value of the CV at time *t.* Filling the relevant minima, the average of the bias will converge to the negative of the free energy, i.e., the optimal bias to increase the number of rare events:

$$-F(s) \sim \overline{V_G}(s) = \frac{1}{t_{tot} - t_{eq}} \int_{t_{eq}}^{t_{tot}} dt V_G(s,t') \quad 3.25$$

where $\bar{V}_G(s)$ is the time average of the bias, t_{tot} the total simulation time, t_{eq} the time needed to fill all the relevant minima.
A system at temperature T samples conformations from the canonical ensemble:

$$P(q) \propto e^{-\frac{U(q)}{k_B T}} \quad 3.26$$

Here q is the microscopic coordinates and kB is the Boltzmann constant. The probability distribution for a CV s is:

$$P(s) \propto \int dq e^{-\frac{U(q)}{k_B T}} \delta(s - s(q)) \quad 3.27$$

This probability is inversely proportional to the free energy landscape F(s) as:

$$F(s) = -k_B T \log P(s) \quad 3.28$$

Thus, the biased distribution of the CV will be:

$$P'^{(s)} \propto \int dq e^{-\frac{U(q)+V(s(q))}{k_B T}} \delta\big(s - s(q)\big) \propto e^{-\frac{V(s(q))}{k_B T}} P(s) \quad 3.29$$

and the biased free energy landscape:

$$F'(s) = -k_B T \log P'(s) = F(s) + V(s) + C \quad 3.30$$

Where C is an undetermined constant. Thus, the effect of a bias potential on the free energy is additive [107], making possible to retrieve the original, unbiased free-energy landscape from the biased one just subtracting the bias potential:

$$F(s) = F'(s) - V(s) + C \quad 3.31$$

From the exploration of the resulting free energy surface, it is possible to assess the thermodynamic and kinetic properties of the system such as for example in this thesis the binding free energy of ligands and the free energy barriers associated to their dissociation.

3.2 Docking and Virtual Screening

The identification of new drug candidates is adjuvated by computers that allow to predict the function and design of new molecules at various stages of drug development [108] pipeline. Two techniques are popular for lead discovery: the virtual screening (VS) and de novo drug design [109]. In particular, VS has proven to be complementary to experimental high throughput screening (HTS) and to be able to handle millions of molecules in a much shorter time compared to experimental techniques/approaches, saving time and costs [110]–[112]. VS is used to identify an effective lead molecules by looking for a ligand, that binds to a target biomacromolecule of interest, among large databases of virtual compounds [113]–[117]. However, rational design methods are entirely dependent on the availability of the biological target's three-dimensional structure. There are two types of computer-aided drug discovery techniques: ligand-based and receptor-based methods [118], [119]. The first is often used when the receptor target's structure is missing and it aims at finding similarities within the known molecules databases respect to a known active ligand structure [120]. When the 3D structure of the target is known, the Structure-Based Virtual Screening (SBVS) method is used to dock the compound library and, through the implementation of energy scoring function, to compute the score of each compound. However, depending on various factors, e.g. receptor's structure, algorithm used, sampled conformations, scoring functions, compounds' library preparation, these approaches can give limited results, reaching a rate of false positive up to 99% [121]. Since the target structure of the system is analyzed in SBVS and it may be difficult to find optimum solvation and force field parameters, this method is computationally more expensive and complex. It is worth to mention that the majority of SBVS assume the target pocket to be rigid or doesn't allow for significant conformational changes upon ligand binding, which in part explains the larger number of false negatives than the ligand-based methods [122].

3.2.1 Structure-based Virtual screening

The SB methods aim at identifying hit compounds that may be further developed into drug-candidates, building on the knowledge of the target's 3D structure. The SB methods that can lead to the identification of new hit molecules can be classified in two main groups *(i)* de

novo techniques and *(ii)* screening techniques (structure-based virtual screening). The de-novo design is based on the assembly of molecular entities (fragments) known to be able to bind the protein target with a low affinity. Differently, the screening techniques use docking methodologies to select molecules with a good affinity for the target among those included in large databases of existing molecules. Docking is a widely used computational approach in SBVS and its goal is to identify the best structural and electrostatic complementarity between a ligand and a target pocket [111].The docking software needs an atomic resolution structures of a target and ligand(s) e.g., crystal structures, structures derived by NMR, or homology models, as well as an understanding of the location of the binding site where to dock the compounds [119]. The basic premise of receptor-based design is that effective inhibitors must be structurally and chemically complementary with their target receptor [123]. In cases where the target binding site is unknown, binding pocket prediction methods like PASS [124], PocketPicker [125] or LIGSITE [126] can be employed. Hence, the software explores the different configurations of the ligand that best fit into the receptor cavity, it calculates the strength of the interaction energies to estimate a binding free energy and it scores the molecules accordingly, giving as a result a ranking of compounds from the highest to the lowest score.

Different types of scoring functions are available in docking programs: Force-field based, empirical and knowledge-based scoring functions. The force field functions are molecular-mechanics energy functions that, depending on the forcefield used, describe the ligand-receptor interactions with non-bonding terms such as van der Waals (vdW) interactions and electrostatic interactions and bonded interactions such as stretching/bending/torsional forces. The Glide program from Schrodinger [127]–[129] suite, the docking software used on this thesis, falls into this category and it uses the OPLS forcefield to score a pose. Empirical scoring functions are models trained on dataset of diverse protein–ligand complexes structures associated with the corresponding experimental affinity data, that aim at recognizing interactions such as vdW, hydrogen bond (H-bond), hydrophobicity, electrostatics, desolvation, entropy, etc [130]. Finally, knowledge-based scoring is a statistical potential of ligand-target complexes from structural information of experimentally determined structures [131].

The calculated free energy of binding is defined by the Gibbs-Helmholtz equation:

$$\Delta G = \Delta H - T\Delta S \quad 3.32$$

ΔG, T and ΔH are the free energy of the binding, the enthalpy and the temperature in Kelvin, respectively, and ΔS is the entropy.
The relation between the ΔG and the affinity between a ligand and a target (Ki) is described by:

$$\Delta G = -RTlnK_i \quad 3.33$$

Where R is the gas constant, T is the temperature in Kelvin and Ki is the equilibrium constant.

3.2.2 Ligand-based Virtual Screening

Ligand-based virtual screening (LSVS) relies on finding structural similarity between a known ligand of the target and new small molecules potentially able to bind to the target. By analyzing and collecting information on the structural properties of known active ligands, this method searches for new molecules intended to bind to the target pocket, assuming that compounds with similar structures are likely to have similar binding mode and activity [120]. LSVS includes a variety of methods like pharmacophore methods, machine learning methods and similarity search methods. Pharmacophore methods identify common structural features among active ligands used to interact with the target. Quantitative structure-activity relationship (QSAR) methods are models trained on known compounds to predict the relationship between chemical structure and pharmacological activity and use that information to find new compounds in a database. Finally, similarity methods calculate the structural similarity between an active compound and the molecules in a database looking for the most similar molecules. This is done by employing different criteria, one of which is the 2D fingerprints that encode the presence of 2D substructural fragments of a molecule into a string allowing to efficiently perform a similarity comparison. To measure the degree of

similarity between two molecules, we have employed the Tanimoto similarity metric [132], [133].

The general form of Tanimoto similarity is:

$$Tanimoto = \frac{c}{(a + b - c)} \quad 3.34$$

Where a and b are the number of structural features in the first and second molecules and c is the number of identical fragments in common between the two [132].

4 Inhibition of Prp4 protein

4.1 Abstract

Pre-mRNA processing 4 (Prp4) is a kinase protein, which plays different roles in regulating pre-mRNA splicing and spliceosome assembly. Albeit its interactome is still not fully uncovered, it has been recently indicated to be a promising target candidate for breast cancer treatment as its expression affects cell survival, apoptosis and migration. Different Prp4 crystal structures and few inhibitors have been characterized so far, yielding the structural basis for computational studies aimed to find possible drug-candidates targeting Prp4. In this study, we exploited a virtual screening protocol, complemented by molecular dynamics simulations, to screen and predict the activity of commercially available compounds as inhibitors of Prp4. As a result, we obtained 11 compounds binding to the Prp4 active site with low μM affinity, as confirmed by SPR and molecular simulations.

4.2 Introduction

Most of the human genes are composed of multiple coding exon sequences interspersed with non-coding sequences called introns that have be spliced to form mature mRNA [20], [21]. Splicing relies on a variety of trans-acting proteins and numerous protein–protein and protein–RNA interactions that are carefully regulated in order to operate in the correct manner. This variety of interactions and regulatory stages opens up to a plethora of opportunities to exploit the splicing cascade for therapeutic purposes. Most of the trans-acting proteins that regulate splicing through the binding of specific nucleotide sequences are serine/arginine-rich (SR) [134] proteins or belongs to heterogeneous nuclear ribonucleoprotein (hnRNP) [135] families. These can be differentially expressed in tissues in

order to regulate tissue-specific splicing patterns. Serine/arginine rich factors play then a key role at many different stages of the splicing process, including splicing regulation, spliceosome formation [136], genomic stability, mRNA export, mRNA stability and translation [137], [138].

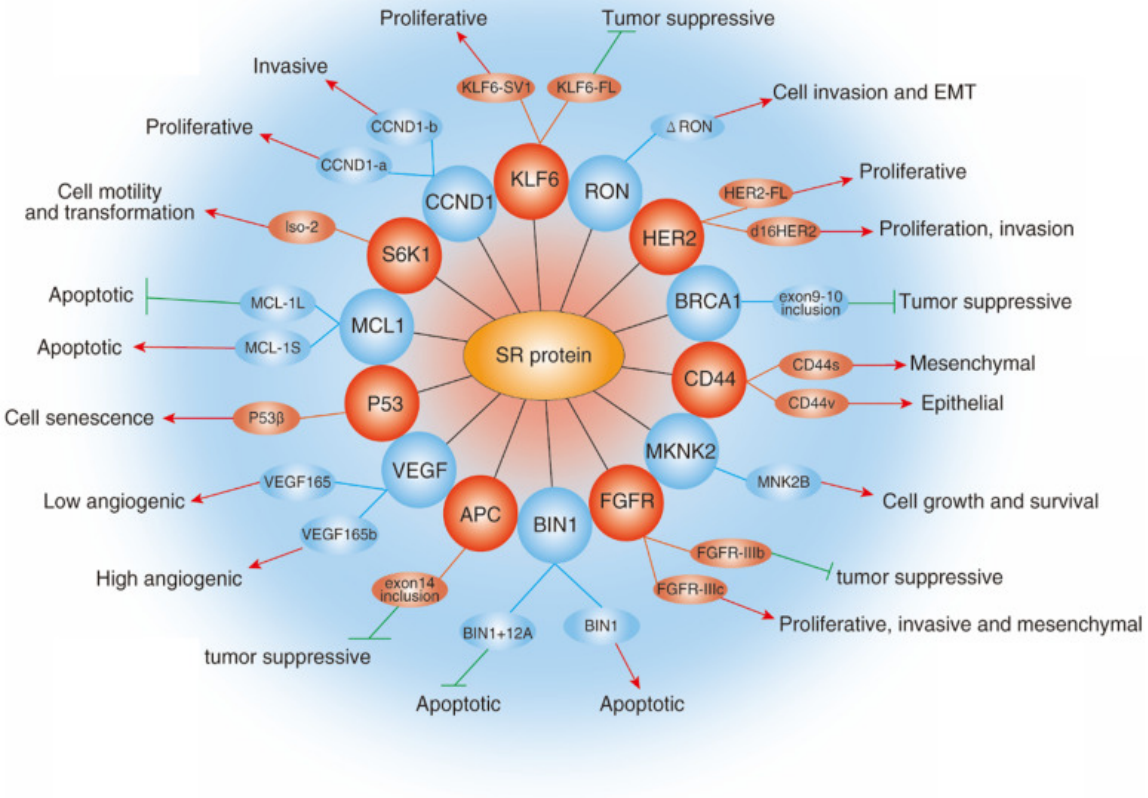


Figure 4.1 Role of Serine Rich (SR) protein family in cancer. The diagram shows selected examples of genes with cancer-related alternatively spliced isoforms that are influenced by the activity of specific SRs. Red arrowheads represent promotion and green lines represent inhibition of the phenotype. The figure was adapted with permission from [139].

As a result, SR rich proteins assume a critical role during transcription and translation (Figure 4.1). SR could indeed act as oncoproteins since their dysregulation affect the normal protein expression patterns, resulting in improper biological activity in various tumor types such as lung, colon, breast, pancreas cancer and leukemia [140]–[142]. Pre-mRNA splicing has been recently shown to be altered in different pathological ways, which results in enhancing the development of cancer. This includes mutations in oncogenes that affect splicing [143],

[144], alterations of gene expression and mutations affecting the spliceosome components [74]–[76], [140], [142], [145]–[147].
SRs undergo phosphorylation by SRPK1 and SRPK2 (SRs kinases) and the U4/U6 small ribonucleoprotein Prp4 [148]. As such, the latter proteins are promising candidates for anti-tumor treatment through the use of small-molecule inhibitors [149].
Prp4 kinase is involved in cell mitosis and splicing, and it has been reported to be a possible therapeutic target in a variety of cancers. Indeed, it has been shown that its inhibition, by using short hairpin (sh)RNAs, reverses paclitaxel resistance in human ovarian cancer by 5-10-fold [150] and enhances paclitaxel activity in breast cancer cells [151]. Triple negative breast cancer (TNBC) [152] is the most aggressive subtype of breast cancer with a high migratory propensity and metastatic rate [153]–[155]. Together with the BUD31 splicing factor, Prp4 has been identified as one of the essential genes for TNCB cell migration. Remarkably, the knockdown of either splicing factor decreased the development of focal adhesions and the interactions with the extracellular matrix, reducing the formation of TNBC metastasis in vivo [152]. This finding also highlights a critical role of splicing deregulation in cancer metastasis. Indeed, a removal of Prp4 leads to incomplete splicing with consequent suppression of cell proliferation, migration, invasion, and induction of apoptosis [156].
Breast cancer is a major cause of death in women, underlining the necessity of identifying innovative therapeutic targets to treat this disease [157]. In this chapter we sought to identify commercially available small molecules inhibitors targeting Prp4 to be used as potential treatment for breast cancer.

4.3 Methods

4.3.1 Virtual screening and Docking

Different libraries were selected to perform the structure based virtual screenings. Namely we used the Enamine, Chembridge databases as well as more specific libraries from the Molport database, i.e., the Kinase Inhibitors and NCCR54K (a collection of around 54k compounds

intended to represent the commercially available chemical space). Instead, the complete Molport library was used for the ligand-based virtual screening.

In order to take into account ionization and tautomerization states of the compounds, the Epik tool of Schrodinger Suite 2017 was employed [158], [159], considering a maximum number of 4 structures for each molecule. Ligands violating the Lipinski's rule of five were filtered out, thus eliminating the compounds possessing poor absorption and permeation [160]. As well, compounds with more than 10 rotatable bonds were discarded since high ligand flexibility implies higher entropic contributions and reduced oral availability [161]. Next, the Qikprop tool [162] was employed to predict LogP values of the compounds in order to further select the molecules.

The protein structure (PDB: 4IJP), containing the catalytic domain of Prp4 (residue 670 to 840) was prepared for docking using Protein Preparation Wizard in Schrödinger suite [162] using the following general scheme: (i) Hydrogens and missing amino acid side chains were added; (ii) Protonation states were predicted at pH 7.0 using PROPKA tool, (iii) hydrogen bond network was optimized; (iv) Missing residues were modelled using Swiss modeller webserver [163].

The docking simulations were performed by using the Glide application of Schrödinger suite. A van der Waals (vdW) radius scaling factor of 0.80 Å for protein and ligands atoms having a partial charge less than 0.15 was used in order to account for protein flexibility. A Virtual Screening (VS) workflow, based on three-steps of docking with increasing level of accuracy, was adopted [128]. Namely, (i) a fast high throughput virtual screening (HTVS) was initially performed in order to efficiently select promising ligands among millions of compounds-large libraries; (ii) 5% of the best ranked ligands were retained and used to perform a single precision (SP) docking calculation; (iii) the top 5% of the resulting compounds were finally screened using the extra precision (XP) protocol. This latter should eliminate false positives by using more accurate scoring functions. The resulting molecules were sorted according to GlideScore scoring function, and, after visual inspection, the poses of the top-ranked compounds were refined by performing classical MD simulations.

In addition to the SBVS we also performed a ligand-based virtual screening (LBVS) by 2-D fingerprints using with the Schrodinger [164] program Canvas. The Rebastinib was used as a

query molecule against the Molport library. Briefly, 2-D molecular fingerprints were generated using [132], [133] for the library and query molecule and Tanimoto similarities were computed between the fingerprints of screening ligands and the query ligand. The fingerprint types were Dendritic with Daylight atom types.
A different LBVS by shape was also implemented using the default parameters of the Schrodinger suite's PHASE [162] tool to screen the Molport database. The 3-D coordinates of Compound A from the crystalized structure reported in [165] were extracted and used as query molecule against the library. For volume scoring, pharmacophore types were used, and the overlapping volume was computed only between atoms of the same type. A maximum number of 10 conformers for each molecule was generated. Molecules with similarity below 0.5 were discarded.

4.3.2 Molecular Dynamics (MD) Simulations

MD simulations were done for each protein-ligand complex, using the FF14SB AMBER force field (FF) [166] for the protein. The ligands were instead described using the general Amber FF (GAFF) [167]. ESP charges [168] were calculated by performing geometry optimization of the substrates at Hartree-Fock level of theory using a 6-31G* basis set with the Gaussian 09 software [169] and were later transformed in RESP charges with the Antechamber module of ambertools 18 [170]. The system was solvated with 10 Å of water molecules (TIP3P model) [171], leading to a total of 48242 atoms. The topology of each protein/small-molecule adduct was built with the ambertools 18 and later converted in a GROMACS format using the software acpype [172]. MD simulations were performed with GROMACS 2020.1 [173]. An integration time step of 2 fs was used and all covalent bonds involving hydrogen atoms constrained with the LINCS algorithm. Particle Mesh Ewald algorithm [174] was used in order to account electrostatic interactions. MD simulations were performed in the isothermal-isobaric NPT ensemble, at a temperature of 300 K, under control of a velocity-rescaling thermostat [89].
In all cases, a preliminary energy minimization was done by employing a steepest descent algorithm. Next, the system was gradually heated to 300 K keeping the entire structure highly

restrained, with exception of the solvent and solute hydrogens. Then, an NPT simulation at 1 bar was done using first the Berendsen barostat and then the Parrinello-Rahman barostat, leaving the side chains free of constraints. Finally, we gradually decreased the restraints of the backbone for a total of 80 ns of equilibration phase.
In order to assess the stability of the docking poses obtained from the SBVS of the Chembridge and Enamine databases the top ranked ligands were inserted into the equilibrated protein structure and each complex underwent 200 ns classical MD simulations.
Molecular Mechanics Generalized Born Surface Area (MM-GBSA) free energy calculation were performed with the MM_PBSA.py tool of Amber 12 [175]program on the equilibrated part of the trajectories, following a procedure successfully applied in previous studies [176], [177] keeping parameters at the default values.

4.3.3 Binding experiments with SPR

The interaction between Prp4 and the selected molecules was investigated at 25 ºC using a Biacore 8K instrument (GE Healthcare) and PBS (D8537, Sigma Aldrich, Saint Louis, MO, USA), 0.005% Tween-20, 2% DMSO as running buffer. The target protein (200 nM in acetate buffer at pH 4.5) was immobilized on the surface of CM5 chips (Cytiva) through standard amine coupling. Five increasing concentrations of the small molecules (7.8, 15.6, 31.25, 62.5, 125 μM or 3.9, 7.8, 15.6, 31.25, 62.5 μM, depending on the small molecules solubility) were injected in a single-cycle kinetics setting. An analogous injection procedure was performed injecting only the running buffer to generate the reference blank curve. Curve fitting and data analysis was performed with Biacore Insight Evaluation Software.
Recombinant human Prpr4 (code P87-35G), expressed by baculovirus in Sf9 insect cells using an N-terminal GST tag, was acquired by Signalchem. The molecules were purchased from different vendors and dissolved in pure DMSO at a concentration of 25 mM.

4.4 Results and Discussion

4.4.1 Structural characteristics and native interactions of the substrate and inhibitors at the binding site

All protein kinase domains consist of a small, mostly β-stranded N-lobe, connected by a short hinge region to a larger α-helical C-lobe (Figure 4.2 A). The ATP molecule binds in the cleft between the N- and C-terminal lobes of the kinase domain with the adenine moiety of ATP being sandwiched between hydrophobic residues and making hydrogen bonds with the hinge region (Figure 4.2 B) [178]–[180]. The N-lobe contains a five-stranded β-sheet (β1–β5) with a single α-helix (the C-helix, αC). The Gly-rich loop (also known as P-loop ort G-loop) lies between the β1 and β2 strands and contains important hydrophobic residues at its tip, which contributes to coordinate the phosphates of ATP [178], [179], [181]. This is the most flexible part of the N-lobe, which folds over the nucleotide, suitably positioning the γ-phosphate of ATP for catalysis. The activation loop (A-loop), is a common trait among all the kinases and its open ATP-bound or closed conformation underlies the kinase active or inactive state, occluding in this latter case the access of the substrate to the active site of the kinase [178], [181]. The N-terminus of the C-helix has to be positioned correctly for efficient catalysis facilitating the interaction between the active site Lys717 (on the β3-strand) and the Glu732 from the C-helix (‘C-helix-in’), whereas a suboptimal position for catalysis e.g. rotating the N-terminus of the C-helix in (‘C-helix-out’), results in an inactive state of the kinase [179], [181]–[183]. At the hinge between the two lobes, deep in the ATP pocket, there is a so called ‘gatekeeper’ residue, which controls the access to the ‘back-pocket’ of the kinase and which is often mutated in kinases developing resistance to inhibitors (Figure 4.2 B) [178]–[180], [183], [184]. The larger or C-terminal lobe of the kinase domain is mostly helical, and it is made of four β-strands in the active state (β6-9). These β-strands exhibit two other important structural components: (i) the catalytic loop, which contains most of the catalytic residues (Y/HRD or Tyr/His-Arg-Asp), and (ii) the DFG-motif where the Asp recognizes one of the ATP-bound Mg^{2+} ion. Recently, a number of Prp4 structures have been characterized in complex with its native ATP/ADP ligand as well as with small-molecules inhibitors. Overall, the general binding mode to Prp4 is similar to that observed in other kinases, with the Prp4

hinge residues Glu768 and Leu770 forming two canonical hydrogen bonds with the adenosine ring.

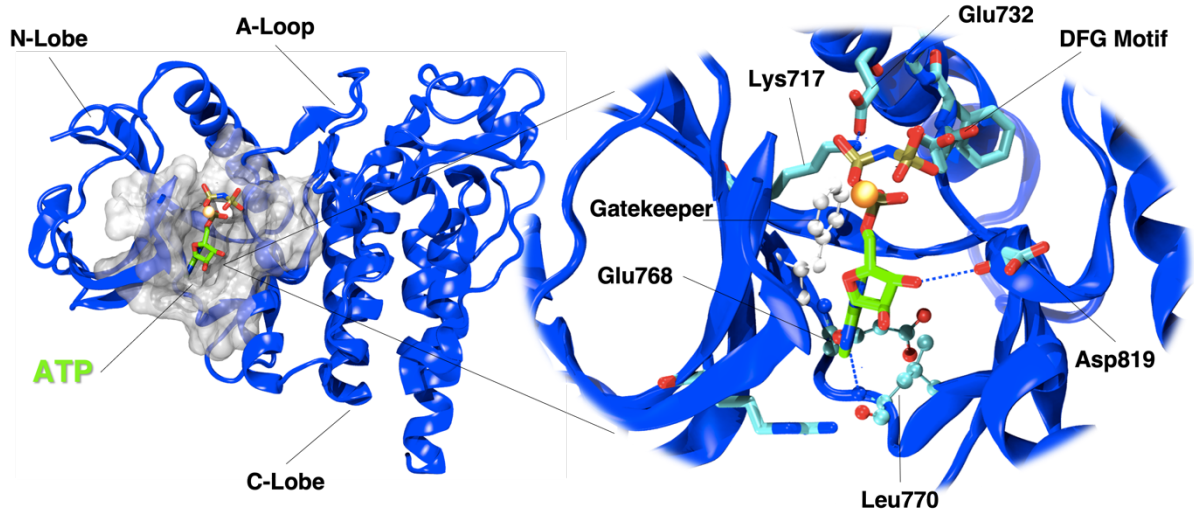


Figure 4.2 Structure and active site of Prp4 catalytic domain. On the left the active binding site is highlighted in white surface, with an ATP molecule shown in licorice and colored by atom name. On the right an inset displays the hydrogen bonding network, the DFG motif (cyan), the phenylalanine gatekeeper residue (white) and a Mg^{2+} metal cation (orange). Prp4 protein is displayed in blue new cartoons.

In our study we used two crystal structures (PDB 6CNH, 4IJP) of Prp4 catalytic domain in complex with either Rebastinib, a potential first-in-class inhibitor, which has also shown activity against other different kinases e.g. Abl1, FLT3 and KDR with an IC50 up to 0.8 nM [185], or a second inhibitor referred as Compound A, identified from a high throughput compound screening and shown to be active in Mass-spectroscopy (MS)-based biochemical assay with an IC50 = 0.016 µM. Nevertheless the latter exhibits low cellular permeability, [165].

As illustrated in Figure 4.3, this co-crystal structure (at 2.25-Å resolution) shows the Prp4 kinase domain in its active conformation with an intact salt bridge between the conserved Lys717 and Glu732 residues. In addition, the DFG motif adopts an “in” conformation with Asp834 pointing towards the active site, while the catalytic Lys717 and the end of the hinge are bridged by Compound A, which exploits the oxygen of the carboxamide of the benzothiophene scaffold as an H-bond acceptor for Lys717, and the amide moiety of the

carboxamide to form a water-mediated H-bond with the main chain carbonyl oxygen of Asp819.

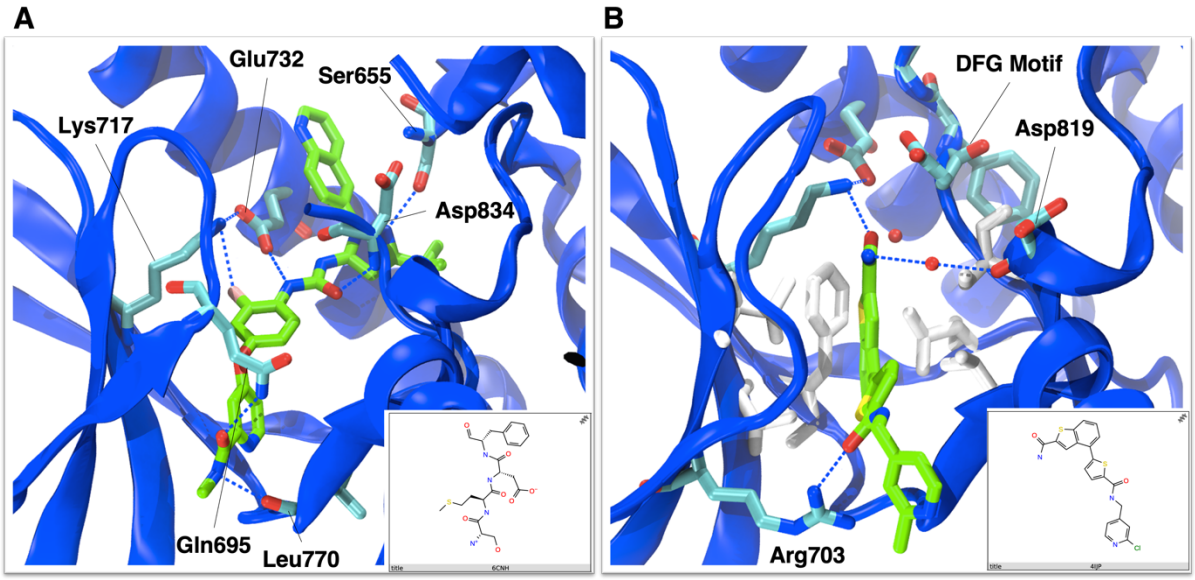


Figure 4.3 View of the binding mode of the inhibitors into the active site of Prp4. The Rebastinib (A) and Compound A (B) are shown in licorice and highlighted by green carbon atoms. Interacting residues of the binding site are shown in licorice with cyan carbon atoms. A 2D sketch of the inhibitors chemical structure is also shown in the inset.

Compound A also forms a H-bond network with the Gly-rich loop. Namely, the carbonyl oxygen of its amide linker acts as an H-bond acceptor for the guanidinium group of the Arg703 side chain. Instead, the benzothiophene ring is placed within the hydrophobic part of the ATP binding pocket, being sandwiched by the side chains of Val701, Ala715, Leu751, Leu822, Cys833, the aliphatic portion of the catalytic Lys717, and the phenyl ring of the Phe767 residue.

4.4.2 Virtual Screening and Molecular Dynamics

Virtual screening (VS) simulations were performed to identify drug-like molecules binding to the Prp4 catalytic pocket. After the docking protocol detailed in the methods section, among the best 100 best ranked compounds for each database, we selected 24 molecules. The resulting molecule/Prp4 adduct underwent 200 ns-long Molecular Dynamics simulations to

assess the stability of the docking pose and to take into account the structural flexibility of the kinase. After the MD simulation, we re-ranked these molecules according to the binding free energy score calculated by the Molecular Mechanics Generalized Born Surface Area (MMGBSA) [186]. As a result, 13 molecules were stable and deeply inserted into the catalytic pocket (Table 4.1, Fig. A.1, Fig. A.2). Subsequently, Single-Cycle Surface Plasmon Resonance (SPR) was used to investigate ligand binding and, when possible, to estimate the ligand dissociation constant (Kd) (Fig. A.4, Fig. A.5, Fig. A.6). However, even though five of them resulted to bind to the kinase, the signal was very weak, suggesting that the molecules have a very high Kd.

A closer inspection of other experimental X-ray structures (6CNH, 4IAN, 4IFC,4IIR, 6PJJ, 6PK6) [165] in the presence of the natural ATP cofactor, revealed the presence of some structural waters, mediating the binding of the ATP molecule. As such, we decided to perform additional SBVS simulations, taking also into account the water molecules. In this case we used all possible water-kinase combinations models (resulting in three systems, besides the protein without any water molecule) as target structures for the SBVS simulations. Moreover, for the VS we used libraries containing already known kinases inhibitors and a collection of around 54k compounds intended to represent the commercially available chemical space (see Methods section). After visual inspection of the resulting best ranked ligands (Fig. A.3, Table 4.2), 7 molecules were selected for their ability to interact with critical residues of the Prp4 active site and in order to maximize their structural diversity. Remarkably, all of them displayed higher docking scores than Compound A and similar results to Rebastinib, even though only this last reached the highest scoring value. Consequently, these molecules were experimentally tested as potential Prp4 inhibitors by SPR assay, to estimate the ligand dissociation rate Kd. Noticeably 5 of them resulted to bind the Prp4 target (VS2-NCCR-3, VS2-KI-5, VS2-KI-4, VS2-KI-2, VS2-KI-1), 2 of them (VS2-KI-1, VS2-NCCR-3) showed high binding affinity toward Prp4: VS2-KI-1 showed a Kd of 6.1μM, while VS2-NCCR-3 instead shows to bind very well to the pocket even in the nanomolar range. Nevertheless, additional experiments at lower concentration are required to precisely estimate its Kd.

As a final attempt to enrich the dataset of molecules to be tested, two LBVS were done. In this case we searched within the full MOLPORT database other ligands with similar structural characteristics, either similar scaffold shape or similar interacting functional groups (fingerprints), to Compound A and Rebastinib, respectively. To assess the reliability of the SPR protocol, the Kd of Rebastinib was also measured and resulted of 12.0μM Consistently with its high inhibitory potency.

From this screening we selected 7 additional molecules whose docking score are reported in . From this set of molecules EXP-a-1 and EXP-a-2, structurally similar to Compound A, bound Prp4, with EXP-a-1 showing a Kd comparable to that of Rebastinib (i.e., 12.6 μM). Additionally, 4 molecules similar to Rebastinib (EXP-b-2, EXP-b-3, EXP-b-4, EXP-b-5) resulted to bind Prp4, with one of them (EXP-b-2) that has to be tested at lower concentration, to exactly determine its Kd which most likely lies in the sub μM range.

As such, in this second round of SBVS and of the LBVS we selected a total of 14 molecules, with nine molecules being able to bind to Prp4.

VS_ID	Glide score	MM-GBSA	Kd
Compound A	**-6,007**	**-34,354 ± 0,38**	IC50 = 0.016 μM
VS1-ENM-1	-10,483	-34,384 ± 0,46	Not Soluble
VS1-ENM-2	-10,765	-30,179 ± 0,45	No Interaction
VS1-ENM-3	-12,609	-29,993 ± 0,40	No Interaction
VS1-ENM-4	-11,582	-31,719 ± 0,33	Not Soluble
VS1-ENM-5	-11,245	-29,815 ± 0,34	No Interaction
VS1-ENM-6	-11,679	-37,377± 0,42	77.0 μM
VS1-CHB-1	-11,268	-45,14 ± 0,48	116.4μM
VS1-CHB-2	-10,728	-40,138 ± 0,38	79.0 μM
VS1-CHB-3	-12,671	-35,573 ± 0,35	No Interaction
VS1-CHB-4	-12,472	-35,069 ± 0,32	No Interaction
VS1-CHB-5	-11,829	-34,118 ± 0,31	52.6 μM
VS1-CHB-6	-12,081	-32,696 ± 0,33	80.0 μM
VS1-CHB-7	-10,723	-30,004 ± 0,27	No Interaction

Table 4.1 Results from the first structure-based virtual screening using the Enamine and Chembridge databases. The docking score, MM-GBSA score with standard deviation (kcal/mol) and Kd from SPR experiments are reported. The nomenclature is defined as follow: VS1 stand for the first round of Virtual Screening, CHB and ENM stand for the libraries used (Chembridge and Enamine). Each molecule is sequentially numbered.

VS_ID	Glide Score	Kd
Rebastinib	**-15.077**	**12.0 µM**
EXP-a-1	-7.960	12.6µM
EXP-a-3	-8.428	>100 µM
EXP-b-1	-9.827	Not Soluble
EXP-b-2	-8.468	<1.0 µM*
EXP-b-3	-8.177	99.8µM
EXP-b-4	-8.007	31.8µM
EXP-b-5	-7.617	16.2µM
VS2-KI-1	-12.284	6.1µM
VS2-KI-2	-10.970	>100µM
VS2-KI-3	-9.748	Not Soluble
VS2-KI-4	-9.688	>100µM
VS2-KI-5	-9.563	>100 µM
VS2-NCCR-2	-8.828	Not Soluble
VS2-NCCR-3	-8.586	<1.0 µM*

*Table 4.2 Results from the second structure-based virtual screening (VS2) using the Kinase-Inhibitor (KI), NCCR databases. Results from the ligand-based virtual screening using CompoundA and Rebastinib as templates (EXP-a/b). The docking score and the Kd from SPR experiments are reported. * refers to currently ongoing experiments to determine the Kd at a lower concentration range.*

4.4.3 Discussion and Conclusions

In this study, we proposed a computational protocol that combines molecular docking or ligand based virtual screening, based on fingerprint and shape similarity, with MD simulations and free energy calculations. Different molecules' databases and datasets have been explored to find inhibitors that could have bind and inhibit Prp4 kinase with high affinity. Prp4 has been in fact recently proposed to be a valuable target to stop the migration and invasiveness of TNBC, even though the proper mechanism of action has not been unveiled yet. After the screening of molecules within the selected databases of commercially available compounds, molecular dynamics and binding free energy estimate were performed to assess the protein flexibility and re-rank the MD-relaxed binding poses before experimentally testing their binding affinity with Surface Plasmon Resonance studies. As a result of all the screening attempted with distinct strategies 59% of tested molecules resulted

to bind the target. Nevertheless, the first SBVS generated a high rate of false positives: although the molecules remained stably bound during the MD simulations and exhibited good binding free energies, they showed poor or no binding affinity for the Prp4 target.
This result was likely due to the neglection of water molecules laying in the vicinity of the ligand, in addition to other well-known limitations of the docking protocols, such as the scoring functions and the static representation of the target structure.
In contrast, the LBVS approach remarkably increased the rate of success of experimentally tested molecules, leading to candidates with very high binding affinity (up to 6.1 μM). This result can be easily explained by the fact that a ligand similarity search, based on molecules known to already bind to the kinase, and followed by docking, appears to be a more robust approach, which is less biased by the particular conformation of Prp4 used. In this case, in fact, the source of error is mainly due to the ligand conformational isomer used as query template for the shape screening and from the kind of fingerprint used. Nevertheless, in this study the crystallographic solved binding pose of known inhibitors in complex with Prp4 was used as template to perform this study, limiting the possible source of errors.

4.4.4 Future Perspectives

Besides testing the compounds with kd < 1 μM at lower concentration range, cytotoxicity studies will be done on a panel of different cell lines such as the TNBC cell lines MDA-MB-231 and the healthy mammalian breast cells MCF10A, in order to assess the cytotoxicity of the compounds at different ranges of concentrations.
Instead in order to monitor the ability of the selected compounds to block the migration of MDA-MB-231, cell migration scratch and Boyden Chamber assays will be done.
The molecules resulting to be able to block the migration of TNBC will be subjected to Structure Activity Relationship studies. On the basis of these findings, the molecules will be optimized through modification of the chemical functional groups by performing free energy simulations. Additional studies might be also done to test other scoring functions and use a consensus score.

5 Molecular Mechanism of py-Tract Recognition

5.1 Abstract

The recognition of poly-pyrimidine (Py) tract splice site signal occurs at the early steps of spliceosome assembly. This is promoted by the heterodimeric U2 auxiliary factor (U2AF), that is made by a large (U2AF65 also referred as U2AF2) and a small (U2AF35 also referred as U2AF1) subunit. The U2AF65, object of study in this chapter, is composed by two RNA Recognition Motifs, RRM1 and RRM2, that can adopt an open or a closed conformation, with only the first being able to effectively bind the Py-tract. U2AF2 has to discriminate among very heterogeneous Py-tract sequences in order to differentiate between strong and weak splicing site. Moreover, this splicing factor is object of frequent cancer-associated mutation, that alter the poly-py recognition and results in deregulated splicing. In this work, by performing biased and unbiased molecular dynamics simulations, we sought to understand and unravel the molecular recognition mechanism of U2AF2 and the impact of selected cancer-associated mutations on the RNA binding affinity and internal protein signaling recognition.

5.2 Introduction

Eukaryotic genes are expressed in the form of premature messenger RNA (pre-mRNA) and must then be converted into mature messenger RNA (mRNA) through a process called splicing. In this process the non-coding regions, called introns, are removed, while the coding regions, called exons, are joined together [187] leading to functional protein-coding mRNA transcripts. In higher eukaryotes, the different combinations of spliced exons, also referred as alternative splicing (AS), lead to distinct forms of mRNA from a single pre-mRNA filament,

increasing the number of proteins produced from a single gene, and thus raising the complexity of the organisms. Otherwise exons can also be constitutively spliced, meaning they are present in every mRNA derived from a particular pre-mRNA transcript [19].
For this reason, the splicing process is a crucial step for the regulation and diversification of genes. Understanding its regulatory mechanism at molecular level is of paramount medical relevance since deregulated alternative splicing (or aberrant splicing) is at the root of or contributes to many human illnesses [188], [189].
In order to correctly perform splicing, the spliceosome, a huge and complex ribonucleoprotein (RNP) machine [26], [190], [191] composed by dozens of proteins and 5 small nuclear (sn) RNA, must recognize critical sites of pre-mRNA at single nucleotide precision. Critical recognition sites are the 5' and 3' splice sites (SS), the breach point (BP) site, and a polypyrimidine tract (Py-tract) located upstream the 3'SS and consisting of 15-20 polypyrimidines bases, whose composition in uridines and cytidines endows the sequence with different properties [192], [193] (Figure 5.1).

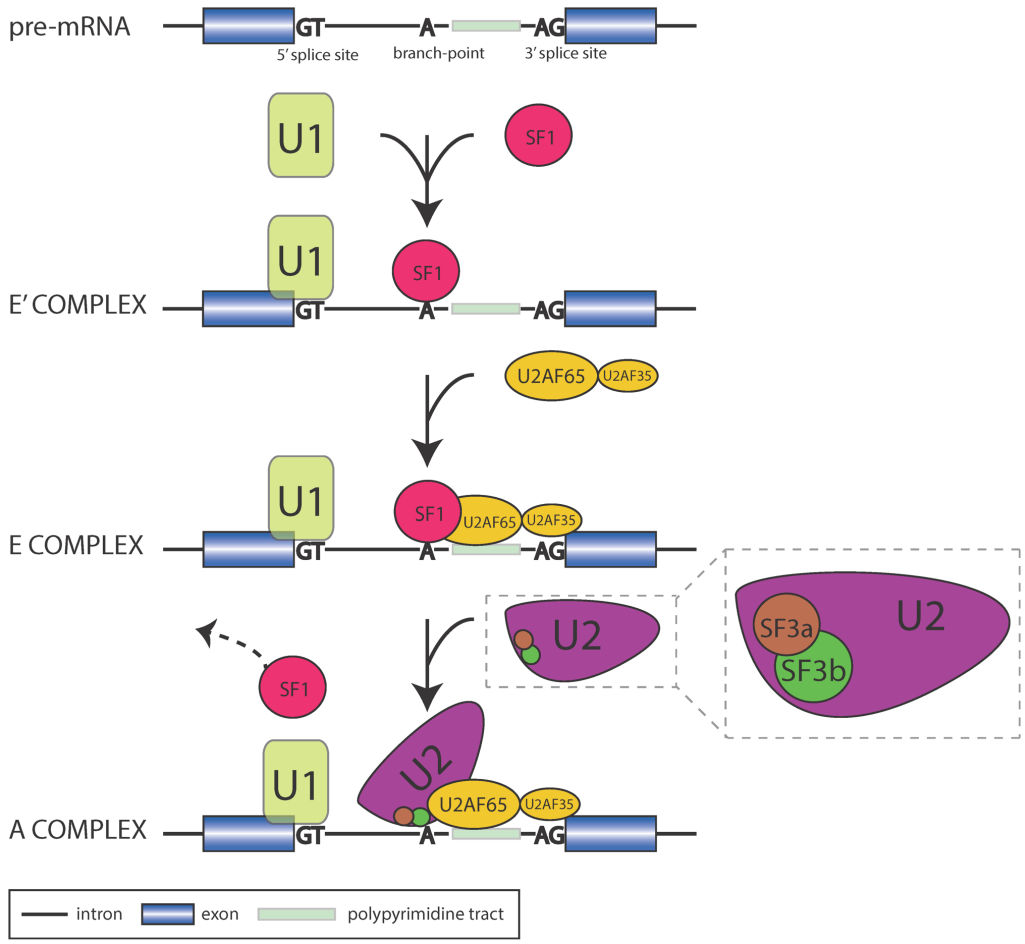


Figure 5.1 Stepwise assembly of the early spliceosome highlighting the known splicing factors that bind to the substrate. The figure was adapted with permission?? from reference [194]

Among these recognition events, the Py-tract, that is adjacent to and strongly entwined with the recognition of the 3SS, is one of the very first events taking place during the spliceosome assembly [195].

The 3'SS and its neighbor Py-tract are recognized by the U2AF heterodimer, consisting of a large 65kDa subunit, U2AF65 (or U2AF2), that binds the pre-mRNA at the poly-Py tract, and a small 35kDa subunit called U2AF35 (or U2AF1), which, instead, recognizes the strictly conserved AG dinucleotide placed at the 3'SS [196]–[198]. U2AF2 is composed by two central tandem RNA recognition motifs (RRM1 and RRM2) held together by a 32-residues long linker. The U2AF2 can adopt two conformations: an open-state, exhibiting the two RRMs side-by-side that is able to bind the Py-tract, or a closed-conformation in the absence

of the Py-tract in which the two RRM domains are found in a back-to-back position (Figure 5.2) [192], [198], [199].

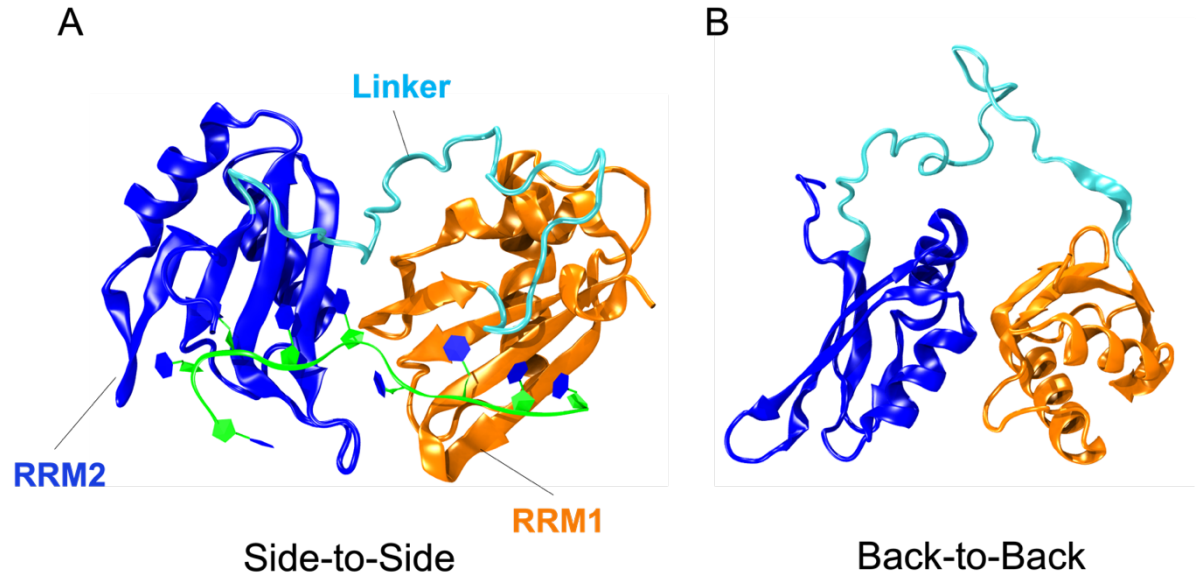


Figure 5.2 Conformational changes of U2AF protein between open (side-to-side, left panel) and closed (back-to-back, right panel) conformations. RRM1, RRM2 and linker are depicted in orange, blue and cyan, respectively. A poly-U RNA of 9 nt is shown in green cartoons.

The Py-tract can have different characteristics in length and composition i.e. it can be interspersed with cytosines (C) or uridines (U), which result in a different pre-mRNA fate [200]. Indeed, U2AF2 can recognize Py-tracts with different purine combinations [196], [201], but a larger number of uridines generally increases the use of an adjacent 3'SS [55], [58], [202], while a Py-tract rich in consecutive cytidines completely abolishes detectable splicing [58], [203]. As such, while poly-uracil (U) leads to 3'SS selection, the poly-cytosine (C) triggers the skipping of the downstream exon [204], [205].

By means of high resolution Nuclear Magnetic Resonance (NMR) structures and Surface Plasmon Resonance (SPR) as well as X-ray crystallography experiments, it was possible to determine the structure of the two RRMs and the loop-linker joining them, and then to measure the different affinity of the RRMs towards poly-C or Poly-U Py-tracts [192], [196], [206]. These experiments allowed to establish that RRM2 binds the 5' region of the Py-tract, displaying a marked preference for poly-U sequences, while having difficulties in selecting a

poly-C tract. Conversely, RRM1 binds the 3' region of the Py-tract, promiscuously recognizing both types of U or C nucleobases [197], [207].

The correct recognition of the 3'SS on the pre-mRNA by U2AF2 is critical to obtain splicing fidelity. Indeed, a number of human diseases owe to 3'SS incorrect selection [208]. U2AF2 subunit is fundamental for vertebrate development [209]. Moreover, defects associated with this splicing factor has been shown to cause cystic fibrosis [79], tumors [80]–[82] and myotonic dystrophy [83].

Remarkably, the two U2AF2 RRMs domains as well as the N-terminal area for heterodimerization with U2AF1 are object of frequent cancer related mutations. Some of these are localized along the RNA/RRM1 or RRM2 interfaces in the open and closed conformations, respectively [78] (Figure 5.3).

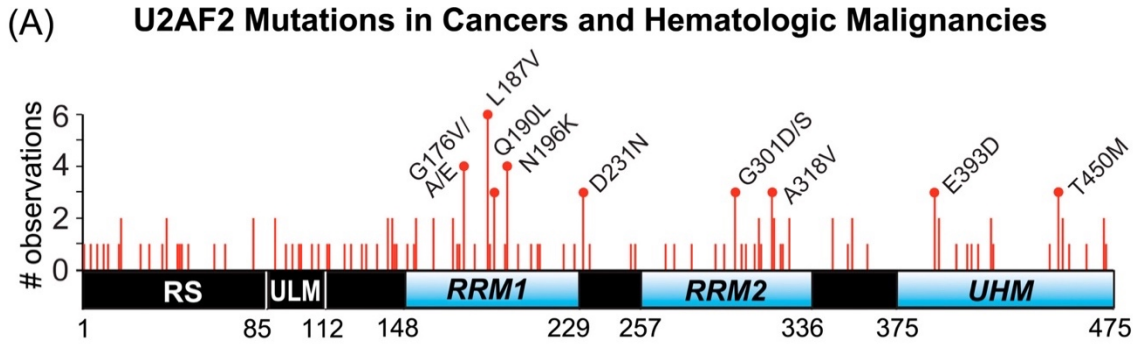


Figure 5.3 Number of independent observations of U2AF2 mutations in cancer among independent patient samples. Amino acids that are changed in three or more samples are labeled. RNA recognition motifs (RRM1 and RRM2) are colored blue while RS (arginine-serine-rich), ULM (U2AF ligand motif) and UHM (U2AF homology motif) domains are in black. The image has been adapted with permission from [78]*, Copyright 2021 American Chemical Society.*

This suggests that these RRM mutations may either modify the Py-tract binding affinity, resulting into altered (aberrant) pre-mRNA splicing or affect the conformational, and most likely functional, equilibrium existing between the closed and open conformation. In spite of their implication in human diseases onset, the structural and functional implications of these mutations on recognition/selection of the pre-mRNA poly-Py tract have yet to be extensively explored.

The purpose of this chapter is to study the pre-mRNA recognition mechanism of U2AF2 and

to assess the impact of its most frequent cancer-related mutations on pre-mRNA's Py-tract recognition mechanism through the use of standard and enhanced sampling (metadynamics) molecular dynamics (MD) simulations.

5.3 Methods

5.3.1 Model Building and MD protocol

The models of the U2AF2 were based on the X-Ray structure (PDB code: 5ev4) where the protein is in complex with a 9-nucleotides long poly-U pre-mRNA strand with a resolution of 1.5Å. Protonation state of ionizable residues have been determined with the PropKa software [210]. The studied mutations were inserted by using leap module of AmberTools 18 [211]. Topologies for all the systems were built using the FF14SB AMBER force field (FF) for protein and whereas ff99+bsc0+χOL3 FF was used for RNA [166]. The charge was neutralized with the addition of 9 Na+ ions, using the parameters of Joung and Cheatham [212]. All systems were embedded in a 12 X 12 X 12Å layer of TIP3P water molecules [171] leading to a box size of 94 x 76 x 77 $Å^3$, for a total of 46223 atoms. Subsequently, the topologies were converted to the GROMACS format using the acpype software [172]. Classical MD simulations were run with GROMACS 2020.1 code [173]. An integration time step of 2 fs was used and all covalent bonds involving H atoms were constrained with the LINCS algorithm [213]. Particle Mesh Ewald scheme was employed in order to account for electrostatic interactions [174], using a real space cut-off of 10 Å. MD simulations were performed in the isothermal-isobaric NPT ensemble, at a temperature of 300 K, under control of a velocity-rescaling thermostat [89] and a Parrinello-Rahman barostat [91]. In all cases, a preliminary energy minimization was performed by employing a steepest descent algorithm. Subsequently, we gradually heated the system to 300 K with an increase of 50 K every 2 ns for a total of 12 ns, keeping the entire system highly restrained except for the solvent and solute hydrogens. Then, we switched to the NPT ensemble, scaling the pressure to 1 bar and using two different barostats: (i) the Berendsen barostat was used for 20 ns with the same restraints on the atoms and (ii) the Parrinello-Rahman barostat for 30 additional ns, while

leaving the side chains free of constraint. Next, we gradually decreased the restraints in 20 ns. Finally, the wild type and each mutated system was relaxed for 2 μs of classical MD, the last 1900 ns were used for the analysis of the simulation trajectories.

5.3.2 Metadynamics Simulation

In order to study the recognition mechanism of a poly-U tract, we used metadynamics simulation to ensure exploring the binding/dissociation process of the linear RNA strand. To this aim we selected three collective variables (CVs): CV1 is defined as the distance between the glycosidic carbon atom of base U2 at one extremity of the strand and G265@N backbone atom; CV2 accounts for the distance between U4 glycosidic carbon atom and A303@N backbone atom at the center of RRM2 and CV3 is defined as a coordination number (CN) considering the atoms able to form hydrogen (H)- bonds (Oxygen and Nitrogen) of bases U1-U5 and those of RRM2 domain atoms laying within 5 Å from the selected nucleobase atoms. After an MD simulation of 200 ns, well-tempered (WT) metadynamics simulations [106] were performed using the plumed 2.4.3 plugin [214]. In the simulation we used gaussian hills of heights of 1.2 kJ/mol and sigma of 0.1 Å, 0.1 Å and 1.0 for CV1, CV2, and CV3 respectively with a gaussian deposition rate of one every 500 steps and a biasfactor of 20. The error of the free energy profile was calculated as the standard deviation of different time averages of CVs from different simulation blocks. The number of blocks has been chosen using the block analysis technique in order to obtain uncorrelated data, i.e., looking at the average error as function of the block sizes until reaching a plateau.

5.3.3 Molecular Mechanics Generalized Born Surface AREA (MM-GBSA)

The binding free energies (ΔGb) and their components were calculated with the Molecular Mechanics Generalized Born Surface Area (MM-GBSA) method by using MMPBSA.py [175] program on 200 frames taken from the equilibrated part of the trajectories (100 - 2000 ns). In our calculations we have not taken into account the entropic contribution of the free

energy, as it was reported that this term may not substantially improve the quality of the results [186], [215] . In addition, a per-residue decomposition analysis has been done to assess the role of each residue and nucleobase in the U2AF2 recognition process.

5.3.4 Network Analysis

The Weighted Implementation of Suboptimal Paths (WISP) program [93] was used to trace the signaling pathways tacking place with the U2AF2 protein, employing the NetWork Analysis (NWA) to discover cooperation between protein residues.
NWA allows for the identification of optimum and suboptimal communication channels, as well as allosteric cross-talk and information on the signaling route's efficiency and strength. In network theory the protein is represented as a correlation-based weighted network, in which the nodes (center of mass of each residue), are connected by edges weighted by the pairwise Pearson's Correlation-Coefficient (CC) as -log CC, which accounts for the amount of correlations between each pair of residues. Cross-correlations scores is defined then as the normalized mass-weighted covariance matrix of Cα atoms' atomic fluctuations, and it has been computed using the cpptraj module of AmberTools18 [211] after removing the global rotational and translational motions by fitting the structure to the initial frame of the equilibrated trajectory between residues along an MD trajectory. As a result, low/high weights into the network imply highly/poorly correlated and anticorrelated movements, respectively, thus NWA calculates the optimal (i.e., shortest) distance across the weighted edges, as well as the slightly longer suboptimal communication pathways, after computing. In these calculations, it is necessary to specify a source and a sink residue that define the start and the end of each route. In the network space, the lengths of the resultant pathways are inversely proportional to the level of correlation between their constituent nodes. For each trajectory, 10000 frames (every 200 ps) from the equilibrated phase of the MD simulation were collected. The WISP algorithm [102]then uses these frames to reconstruct the correlated network and the best path of connected nodes, as well as 100 suboptimal pathways.

5.4 Results and Discussion

5.4.1 Recognition process by RRM2 domain

The RRM2 domain has been demonstrated in experimental studies to exert a higher selectivity for the poly-U pre-mRNA Py-tract as compared to the RRM1 domain [192]. As such, U2AF2 plays a key role in determining whether a Py-tract sequence is a strong or weak binding sequence. In this study we initially sought to investigate the recognition mechanism of the poly-U tract by the RRM2 domain of U2AF2. Distinct structures are available in the Protein Data Bank (PDB) for this system, nevertheless, only few of them trapped the U2AF2 protein in an open-conformation, while binding 9-nt long poly-U strand (PDB 2YH1, 5EV1/2/3/4). Hence, after relaxing an U2AF2 structure binding the poly-U strand using a careful equilibration protocol (see methods), we performed a 2 μs-long MD simulation. Subsequently, we analyzed the H-bonds network (Table 5.1) and the difference in binding free energy (ΔGb) for both the protein and RNA counterparts with the Molecular Mechanics Generalized Born Surface Area (MM-GBSA) method and we decomposed the resulting ΔGb for each base and residue (Figure 5.4) to assess the cardinal points for the binding/recognition of the poly-U sequence.

As a result, the bases U3, U4, U5, U6, U7 and U8 display the largest ΔGb. Most of these bases interact and are stabilized by the linker residues and surprisingly by the least selective RRM1 domain. In particular U7 and U8 are stabilized by aromatic residues Tyr152, Phe197 and Phe199 through π-stacking interactions with an additional H-bond with Arg150, leading to a ΔGb contribution up to -6.13 (±0.01) kcal/mol.

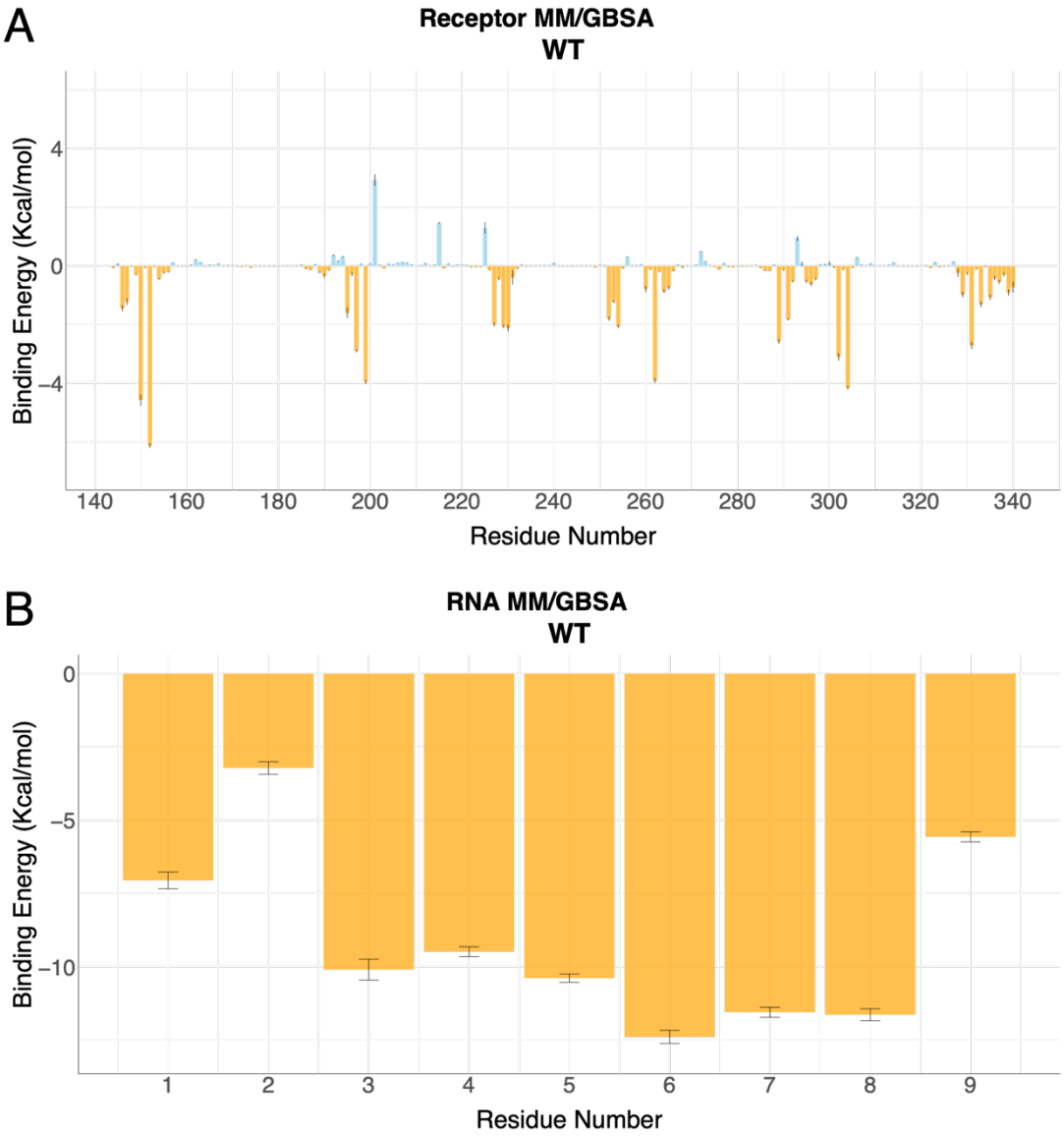


Figure 5.4 Per-residue binding free energy (ΔGb, kcal/mol), calculated using the MM-GBSA method for each U2AF2 residue (A) and each nucleotide of the poly-U tract (B).

RNA	U2AF2	% Persistency
U5_1@N3	Lys_329@O	0.54
U_3@O4	Gln_333@NE2	0.55
U_3@N3	Ala_335@O	0.45
U_4@O2	Asn_289@ND2	0.83
U_5@N3	Thr_252@O	0.88
U_5@O4	Val_254@N	0.43
U_5@O2	Lys_225@NZ	0.77
U_6@OP2	Tyr_152@OH	0.97
U_6@O2'	Lys_195@O	0.74
U_8@O2	Arg_150@NH2	0.73

Table 5.1 Table of Hydrogen bonds persistency (calculated as the number of frames the H-bond is present with respect to the total number of frames) established between the U2AF2 protein and poly-U RNA tract. Protein residues belonging to RRM1, RRM2 and linker are colored in blue, orange and cyan, respectively.

Similarly, the RRM2 domain stabilizes U3 and U4 by engaging π-stacking interactions with Phe262, Phe304, Tyr302, with the addition of the H-bond by Asn289, Als335 and Gln333. In this case the ΔGb of each of these residues is up to -4.16 (±0.05) Kcal/mol. The U5 base, located instead at the interface of the two domains is stabilized by RRM2's Lys225 and the linker residues Thr252 and Val254, while U6 is tightly bound by the two β-loop at the interface of the two RRMs. The set of interactions observed is fully consisted with those trapped in the X-ray structure. The overall ΔGb of the poly-U tract is of -137.79 ± 0.99 kcal/mol.

To enhance the exploration of the possible states visited by the polyU tract (bases U1 to U4) during its recognition/binding by/to RRM2 we made a step further and performed a well-tempered metadynamics simulation using three CVs, two distances (CV1 and CV2) between two RNA points and the closest protein backbone atoms and one coordination number (CV3) at the interface with the first five bases. The distances have been set to be able to effectively sample several unbinding and binding events taking into account the linear shape and flexibility of the RNA (Figure 5.5).

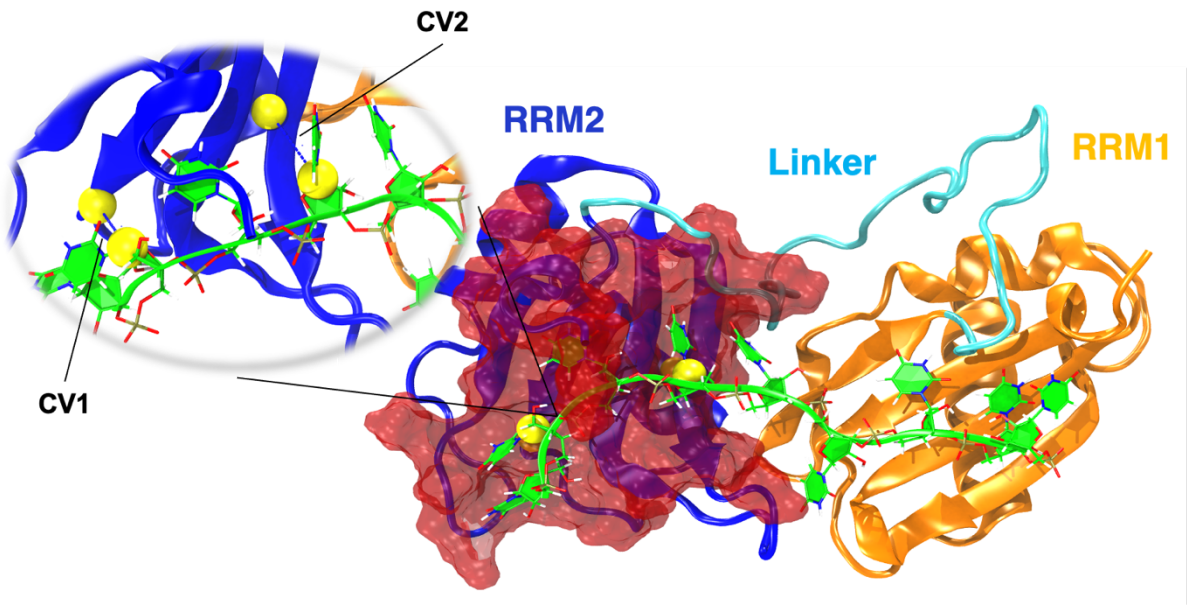


Figure 5.5 Definition of Collective variables. In yellow van der Waals spheres depict the atoms defining the distances of Collective variable (CV)1 and CV2, whereas in red surface the residues considered to define the coordination number for CV3 are shown.

With this set of CVs, we performed 1μs-long metadynamics simulation. The resulting Free energy surface (FES) was projected onto two CVs, generating as a result 2d FES for the 3 combinations of collective variables (Figure 5.6), after having monitored the time-evolution of the difference between local energy minima and the main energy basin as a measure of

convergence of the simulation (Figure 5.7).

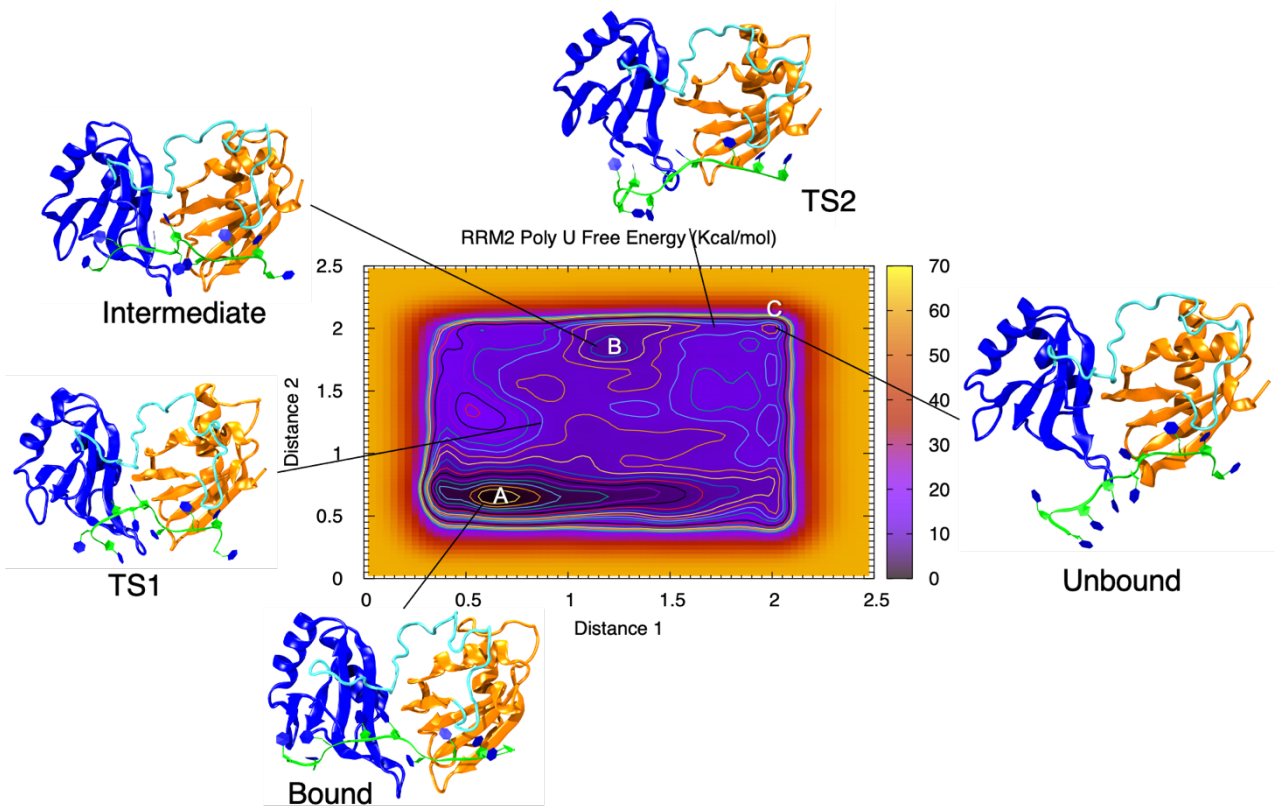


Figure 5.6 Free energy surface (FES, kcal/mol) for the dissociation of the U1-U4 pre-mRNA poly-U tract from the RNA recognition motif 1 (RRM1) of U2AF2 plotted as a function of the two distances CV1 and CV2 defined as the distance between the glycosidic carbon atom of base U2 at one extremity of the strand and its closest protein backbone atom, whereas CV2 accounts for the distance between U4 glycosidic carbon atom and its closest protein backbone atom at the center of RRM2. The FES ranges from dark purple to yellow and contour plots are reported every 1 kcal/mol. A, TS1, B, TS2 and C indicate the main minimum, the first transition state leading to the intermediate state B, the second transition state leading to the fully dissociated state C. Representative structure are shown.

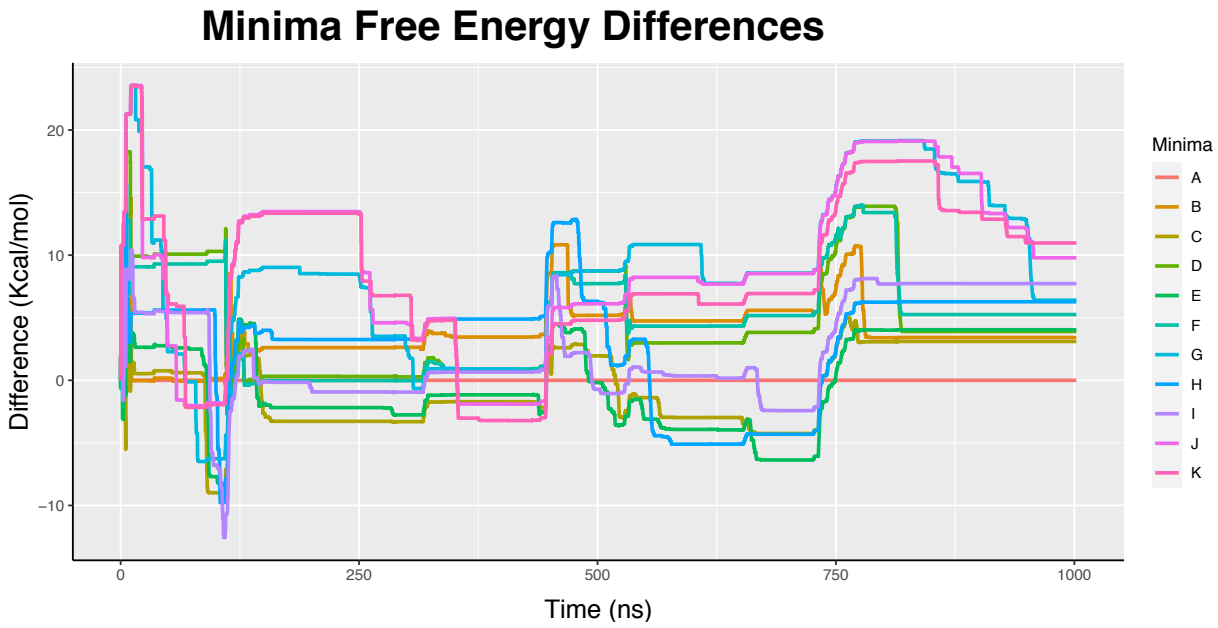


Figure 5.7 Free energy convergence. Difference in time of the free energy (kcal/mol) of each local minimum as compared to the main minimum. Each minimum has been labeled in alphabetic order, starting from the main minimum A to the dissociated minimum K.

Interestingly, on the resulting FES it is possible to identify the free energy minima among which the system is likely to move during the dissociation/binding process. These are metastable states that contribute to the selective recognition of the poly-U tract by U2AF2 RRM2 domain. In particular, three states have been observed, ranging from the RNA-protein complex (state A), where most of the interactions reported in the starting crystal structure are maintained, to the dissociated state (state C) where the U1-U5 pre-mRNA portion is completely dissociated from the RRM2 domain. A third intermediate state is visited along the dissociation (state B), where the RNA bases interact with the protein counterpart in a disordered manner (Figure 5.8). State B is reached passing through a first transition state (TS1) by surmounting a free energy barrier ($\Delta Gd^{\#}$) of ~10 ± 0.56 kcal/mol in an endothermic process (DG = ~7 ± 0.45). From here the complete dissociation of the poly-U requires

overcoming an additional $\Delta Gd^{\#}$ of 3 Kcal/mol to overcome the second transition state (TS2).

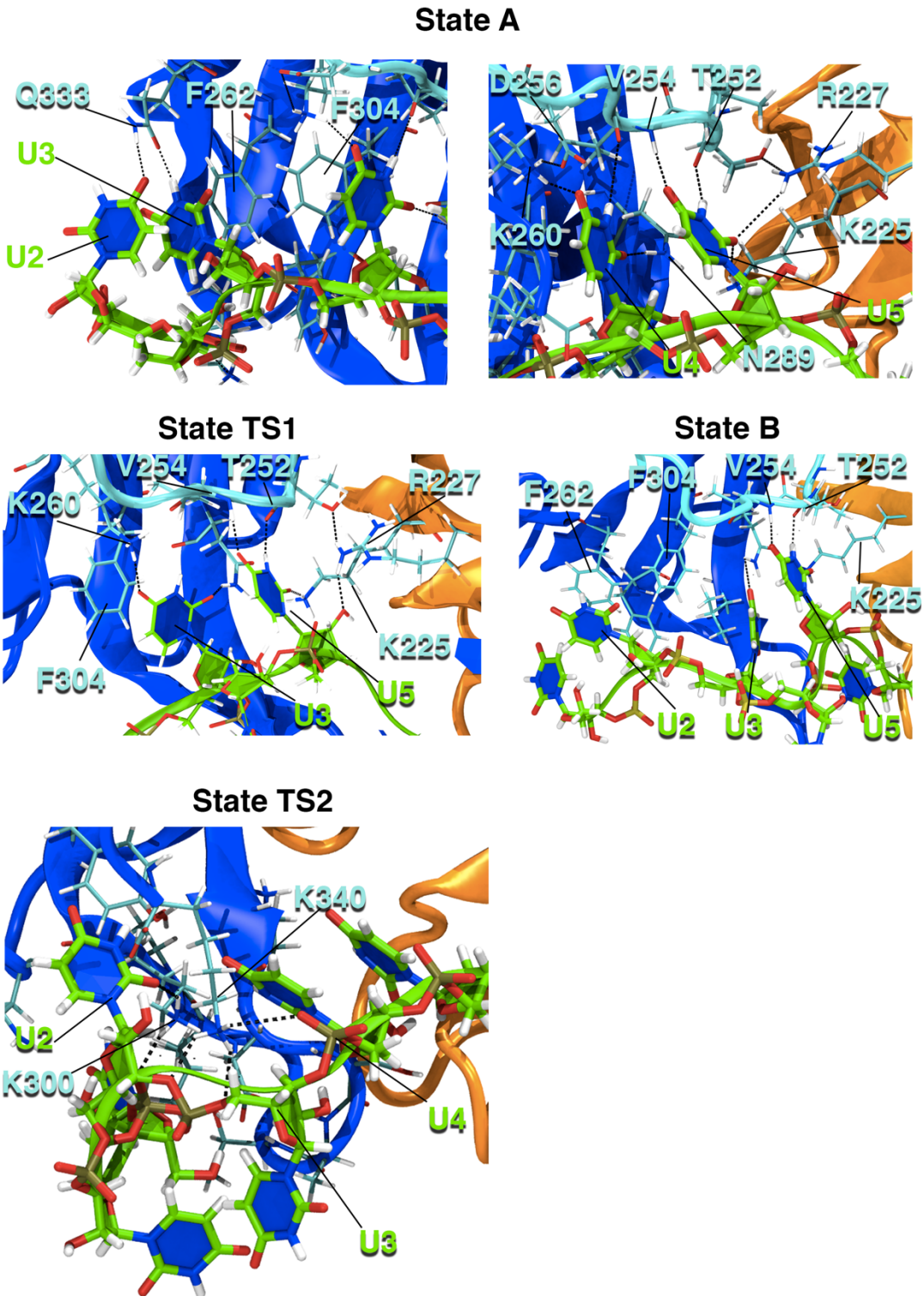


Figure 5.8 Interactions of main minima and transition states visited during poly-U recognition process by U2AF2 RRM2 domain. State A, TS1, B, TS2 display the bound state, the transition state toward state B, the intermediate state and the transition state toward the unbound state, respectively.

The bound state shows all of the native interactions found in the starting crystal structure, namely the backbone amide of the linker T252@O and V254@N residues hydrogen (H)-bonds to the U5@N3H and O4 atoms. In the C-terminal β-strand of RRM1, the side chains of Lys225 and Arg227 donate additional H-bonds to the U5@O2 lone pairs. The C-terminal region of the inter-RRMs linker also participates in the binding of the U4 base, with the Val254 backbone carbonyl and the Lys260 side chain H-bonding to U4@N3 and O4 (Figure 5.8 A). Furthermore, the U4 base engages a π-stacking interaction with Phe304 of RRM2 and an additional H-bond between U4@O2 and Asn289@N. Conversely, U3 π-stacks to Phe262 side chain and U3@N3 H-bonds with Q333@O. Overall, these interactions ensure a stable binding of the polyU-tract. At TS1 (Figure 5.8 B) the interactions between U4@N3 and Val254 backbone and between Q333@O and U3@N3 are broken, whereas the interaction of residues Lys260 and Asn289 with U4 base is replaced with the U3 base, until the intermediate state B is reached. In this latter state (Figure 5.8 C) only U5 strongly binds to RRM2 retaining the same interactions detailed above, while U2 establishes π-stacking interactions with the side chains of Phe262 and Phe304 residues. Interestingly, the transition state between states B and C (TS2) is mediated by H-bonding interactions with Lys300 and Lys340, protruding from the RRM2 and establishing salt-bridges with the phosphate groups of U2, U3 and U4 bases (Figure 5.8 D).

A	TS1	B	TS2
T252@O-U5@N3H			
V254@N-U5@O4			
K225@NZ-U5@O2			
R227@NH2-U5@O2			
V254@O-U4@N3			
K260@NZ-U4@O4	K260-U3@O4		
N289@N - U4@O2	N289@N–U3@O2	N289@N–U3@O2	
Q333@O - U3@N3			
F304-U4	**F304-U3**	**F304-U2**	
F262-U3		**F262-U2**	
			K300-U2@OP1
			K340-U3@O5’
			K340-U4@OP2

Table 5.2 Hydrogen bonds and π-stacking interactions of different recognition states, namely the bound state (A), the Transition State 1 (TS1), the Intermediate state (B) and Transition State 2 (TS2). The red and green cells represent the broken and maintained interactions, respectively, while the blue cells stand for new interactions established. In bold are highlighted π-stacking interactions.

Being Lys300 and Lys340 the last residues to dissociate from pre-mRNA, they are most likely involved also in the early selection/recognition mechanism of the poly-U tract during its binding process. Namely, Lys300 and Lys340, acting as RRM2 fingers protrude from the protein to early (non-selectively) recognize the pre-mRNA filament by establishing salt bridges with its backbone. Conversely, the hydrophobic and π-stacking interactions and linker H-bonds trap the pre-mRNA filament by executing the final selective recognition of the poly-U tract. This leads to the set of sequence specific interactions observed in bound A state. Intriguingly, the number of Lysines/Arginines differs drastically between the two RRMs. While RRM1 at the RNA binding interface is populated by five Arginines (Arg146, Arg149, Arg150 Arg227, Arg228) and has few Lysines (Lys195, Lys225), the RRM2

exhibits the opposite distribution of basic amino acids, namely it presents one Arg (Arg334) and six Lysines (Lys260, Lys292, Lys300, Lys328, Lys329, Lys340) (Figure 5.9).

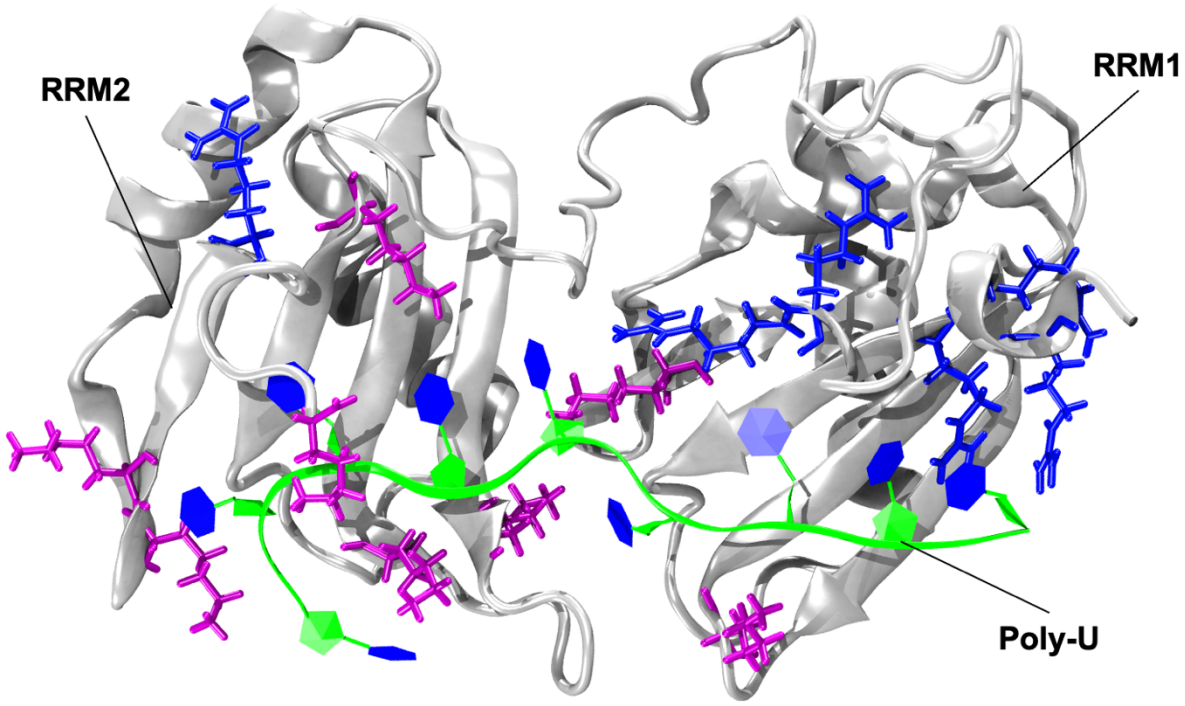


Figure 5.9 Distribution of Arigines and Lysines on the binding interface with the Poly-U Py-tract. Lysines and Arginines are shown in magenta and blue, respectively.

This may explain the difference of stability exerted by the two domains, where the highly flexible Lysines residues on RRM2 early traps the poly-U tract, and establish salt bridges interactions, whereas Arginines on RRM1 more tightly bind both the uridines and cytidines bases, being able to both establish salt bridges and stacking interactions to the nucleobases.

5.4.2 Impact of cancer associated mutations on RNA Recognition

Mutations on the two U2AF2 motifs RRM1 and RRM2 and of the loop-linker connecting them trigger rare genetic diseases and cancer by provoking aberrant pre-mRNA splicing, which leads to the formation of non-functional proteins (Figure 5.3). It has been observed that the mutations present on this splicing factor often occur at the interface between the RRMs

and the pre-mRNA (Figure 5.10) or between the two RRMs domains when in closed conformation [78].

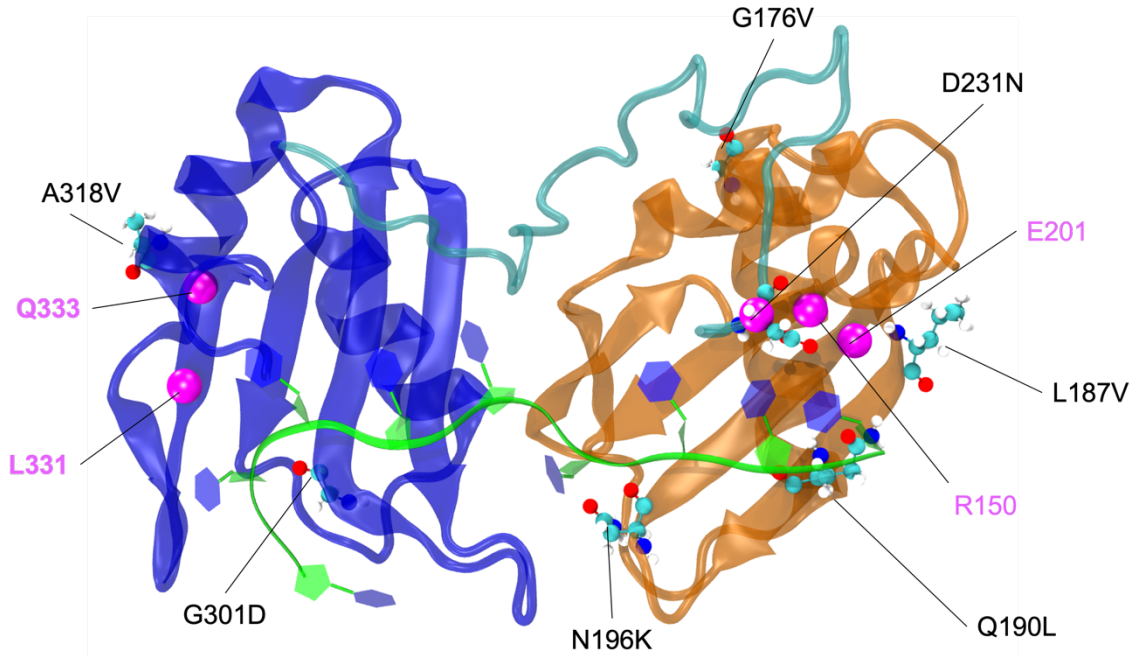


Figure 5.10 Structure of U2AF2 with the RRM1 and RRM2 domains shown in orange and blue new cartoons and the RNA in green ribbons. In Ball and sticks are shown the residues object of pathogenic mutation, while in magenta is highlighted the position of the residues R150, D231, Q333 and E201 whose interaction with RNA is recurrently affected in all investigated mutants.

Among these mutations, N196K is associated to the onset of Acute Myeloid Leukemia (AML), Q190L to Chronic Lymphocytic Leukemia (i.e. accumulation of lymphocytes in the bone marrow, blood and lymphatics), while lung cancer adenoma has been found in at least three patients [78], [216] .

Other two frequent mutations, observed to be linked to cancer onset, are located on the RRM1 domain, namely G176V and L187V. While the former induces pulmonary adenocarcinoma, the latter, being also the most frequent, induces the formation of liquid tumors such as AML and the myelodysplastic syndrome (MSD) [217]. As well, other frequent mutations, G301D and A318V, are observed on RRM2. These induce the onset of colorectal or prostate cancer and stomach cancer, respectively. Finally, the D231N mutation is instead located on the linker connecting the two domains, while being positioned also

nearby the RRM1 domain. This mutation was observed in three patients affected by squamous cell carcinoma, the high-grade pediatric glioma and Wilms' tumor [78].
To investigate how the selected most frequent mutations alter the recognition mechanism of the poly-Py tract, we mutated each residue on the WT structure accordingly and performed 2μs-long simulations for each system, reaching a cumulative simulation time of 16 μs.
When analyzing the differences in the total ΔGb (Table 5.3) induced by the selected mutants as compared to the WT, we unexpectedly observed that mutations located on the RRM2 predominantly destabilize the RNA binding. This is particularly evident in case of G301D mutation, having a positive difference in the total ΔGb (+10 ± 2 kcal/mol) with respect to WT U2AF2. Conversely, the mutations located on the RRM1 domain do not markedly affect or stabilize pre-mRNA binding. This is particularly evident for L187V which increases the ΔGb by -8.7 ± 1.87 kcal/mol. Furthermore, for a better comparison between WT and the distinct mutants, the relative binding free energy (ΔΔGb) per residues has been reported (Figure 5.11).
Aiming to pinpoint the existence of possible general trends emerging from this set of MD simulations we attempted at capturing few cardinal points of U2AF2 which retain a stabilizing/destabilizing effects on RNA binding in all mutants investigated. As a result, we noticed that Glu201 and Asp231 on RRM1 always exert a stabilizing effect (inducing a more negative ΔGb) on the U8 and U9 bases in presence of almost all the investigated mutations. This occurs via the formation of a more persistent H-bonds with the U8 and U9 (See tables in Appendix B) or it is mediated by the formation of a more stable network of H-bonding interactions between Glu201 and the nearby arginines (Arg146, Arg150) resulting in a more efficient screening of its glutammate negative charge, thus ameliorating its repulsion for the RNA strand. As a result of this interaction network, the rearrangement Arg150@RRM1 destabilizes the pre m-RNA binding in all mutants investigated.

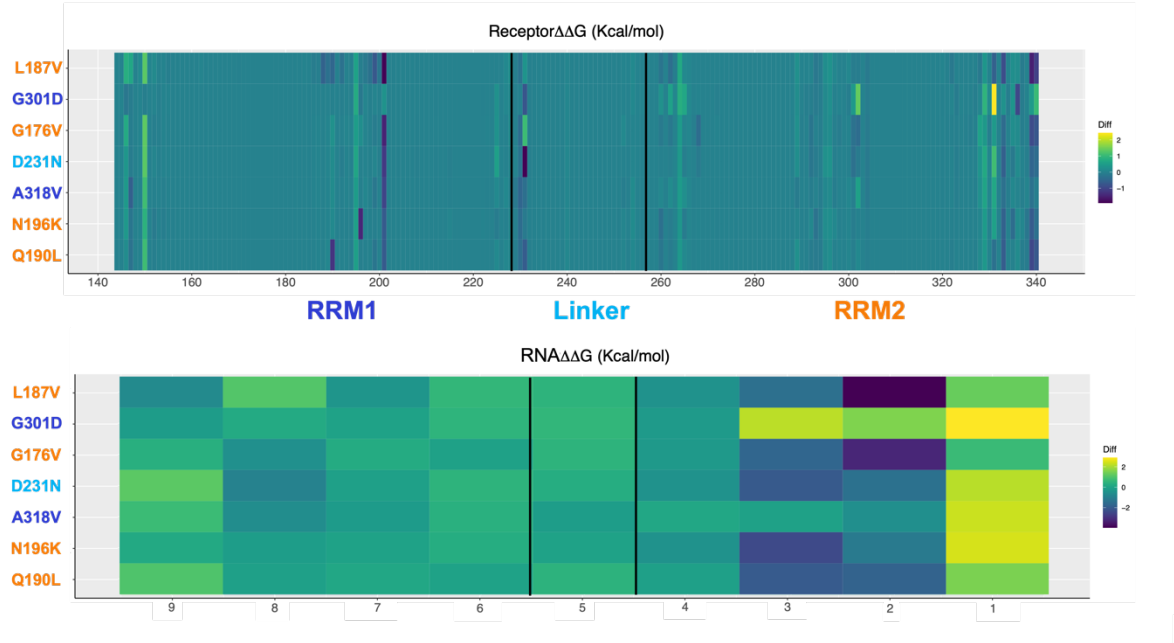


Figure 5.11 Matrix of the difference in the binding free energy for each mutant of the U2AF2 protein as compared to its wild type counterpart. Examined mutants vs free energy difference (ΔΔGb) of each protein residues are reported on y and x axis, respectively. ΔΔGb is reported as color bar ranging from dark blue to yellow in order of decreasing ΔΔGb.

Conversely, Gln333 and Leu331 remodeling result into a negative and positive ΔΔGb, respectively, in all mutants. In this case Leu331, laying in the vicinity of U1 and U2 remodels to increase the steric hindrance with the U1 and U2 bases, thus destabilizing their binding. This is nevertheless counterbalanced by Gln333, which instead form more stable H-bond interactions with U2 and U3, thus contributing to stabilize the pre-mRNA. As such Gln333 and Leu331, and Glu201, Asp231 and Arg150 represent the key hop spot regions responding to the mutations-induced perturbation for poly-U recognition. Also other residues present a trend, but the difference in energy is more modest (< than 1 kcal/mol) and therefore are not analyzed and discussed in the following.

System	Total Energy
WT	-137.79 ± 0.99
N196K	-145.52 ± 0.78
L187V	-146.53 ± 0.88
Q190L	-141.22 ± 0.69
G176V	-139.69 ± 0.84
A318V	-129.94 ± 0.92
G301D	-127.05 ± 1.01
D231N	-140.27 ± 0.86

Table 5.3 Binding free energies (kcal/mol) of a 9-nucletode long poly-U RNA tract to wild type and mutant U2AF2. Mutations highlighted in orange blue and light blue are placed on RRM1, RRM2 and the linker-loop, respectively

5.4.3 Optimal and Suboptimal signaling pathways across RRM domains

Since almost all mutations contribute to lower the ΔGb of residues Glu201, Asp231@RRM1 and Gln333@RRM2 and to increase that of Arg150@RRM1 and of Leu331@RRM2 we sought to trace the existence of common signaling routes heading from the mutation site towards some of these residues. As such, we applied NetWork theory Analysis (NWA) [93] to decrypt the information-exchange pathways underlying the observed ΔGb alterations and to decode whether and how the mutant residues exploit common mechanism and routes to enhance the inter- and intra-RRM allosteric cross-talk. In NWA, the protein is represented as a correlation-based weighted network, having nodes (the residues' center of mass) connected by edges weighted with the Pearson correlation coefficient between the residue pairs (i.e., small/large weights reflect highly/poorly correlated and anticorrelated motions). By computing cross-correlations between residues along an MD trajectory, NWA can find the optimal and suboptimal signaling-paths between two user-selected source (mutation site residue) and sink (Arg150, Gln333 and Asp231) residues. These sink residues were chosen as they were observed *(i)* to affect in a consistent and relevant manner the pre-mRNA binding in all the investigated mutants, *(ii)* to be in direct H-bond contact with the RNA strand and *(iii)*

to belong to the three different structural elements of U2AF2 (RRM1, RRM2 and the linker loop, respectively). The outcoming path-lengths should be thus inversely proportional to the signaling-strength and to the amount of correlation existing among their tracing nodes. Thus, by performing NWA on the WT and all mutant systems, we examined and compared the resulting path distribution of the first 100 shortest paths, dividing them in intra-domain or inter-domain (i.e. when the source and sink residues are in the same or in two different RRMs). Since the linker residue Asp231 lies in contact with RRM1 we consider the paths originating from mutants belonging to RRM1 as intra-domain path. Conversely those originated from RRM2 mutants as interdomain.

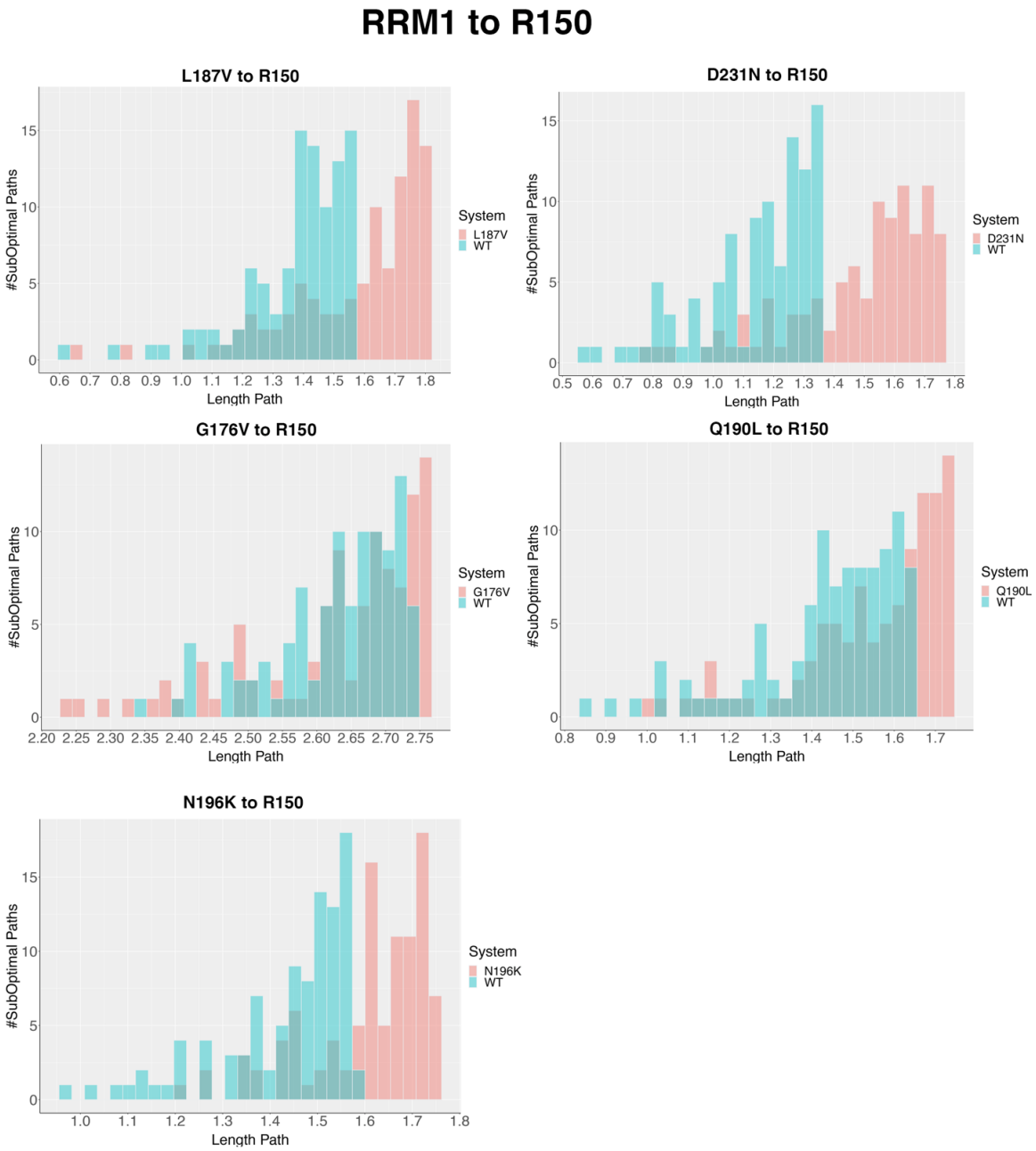


Figure 5.12 Distribution of RRM1 intradomain signaling-path lengths from all investigated mutations toward Arg150@RRM1 compared to the WT distribution. In pink the distribution of paths within the mutant system is shown. In cyan the WT distribution is displayed.

As a result, the intra RRM1 paths or the inter RRM2-to-RRM1 paths heading to Arg150 are longer in the presence of all the investigated mutants (Figure 5.12, Figure 5.13).

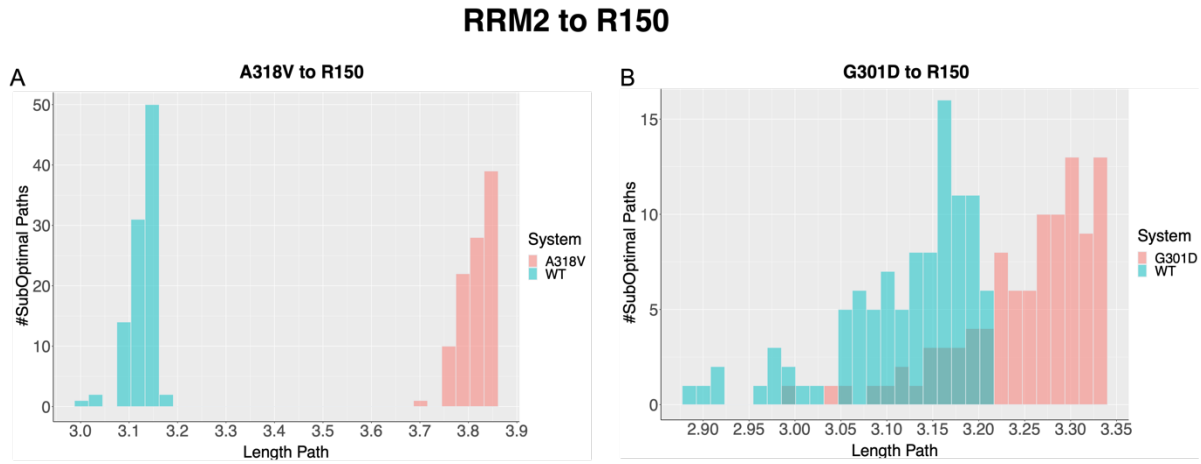


Figure 5.13 Distribution of interdomain signaling-path lengths from all investigated mutations on RRM2 toward Arg150@RRM1 compared to the WT distribution. In pink the distribution of paths within the mutant system is shown. In cyan the WT distribution is displayed.

The same trend is also observed for the intra RRM1 and inter RRM2-to-RRM1 paths heading to Asp231 (Figure 5.14, Figure 5.15).

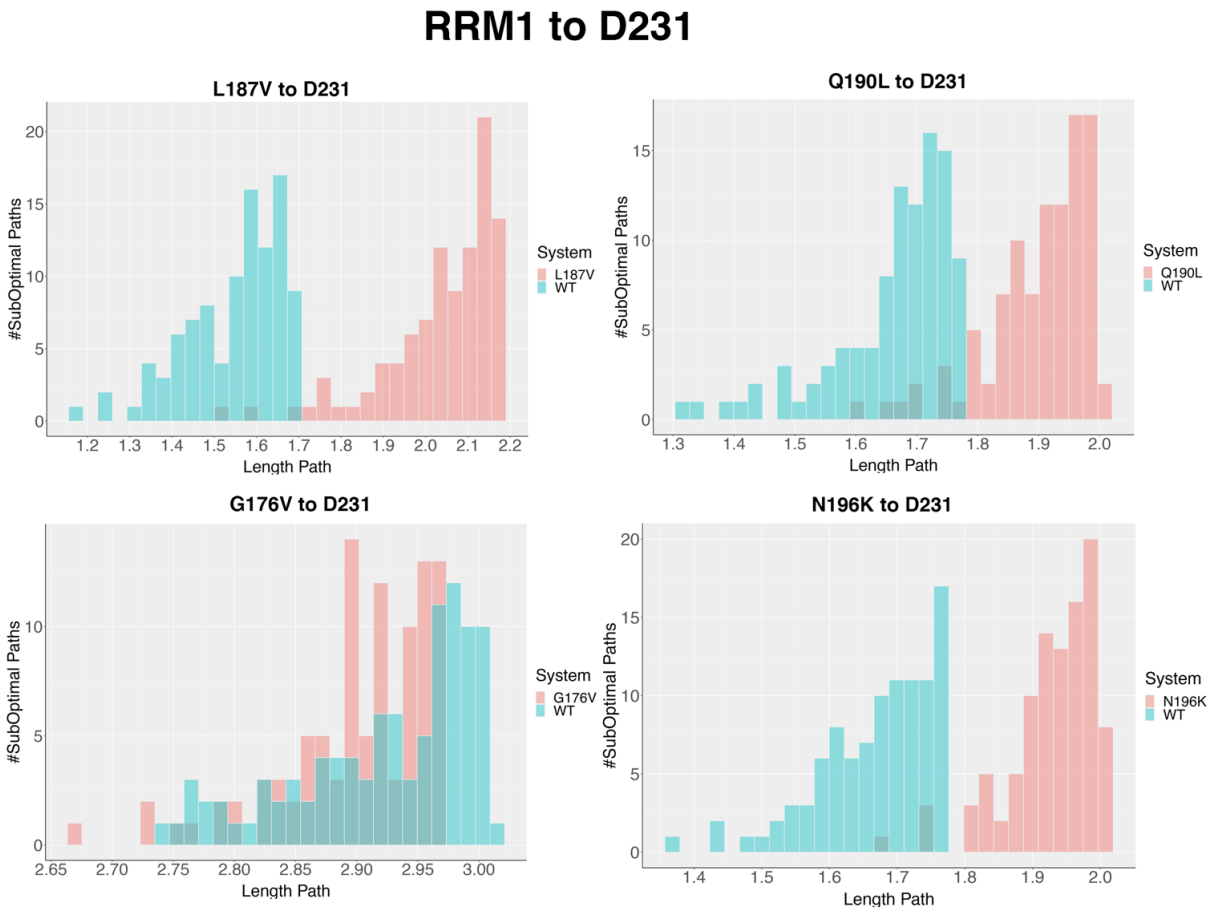


Figure 5.14 Distribution of RRM1 intradomain signaling-path lengths heading from the mutations site on RRM1 toward Asp231@RRM1 compared to the WT distribution. In pink the distribution of paths within the mutant system is shown. In cyan the WT distribution is displayed.

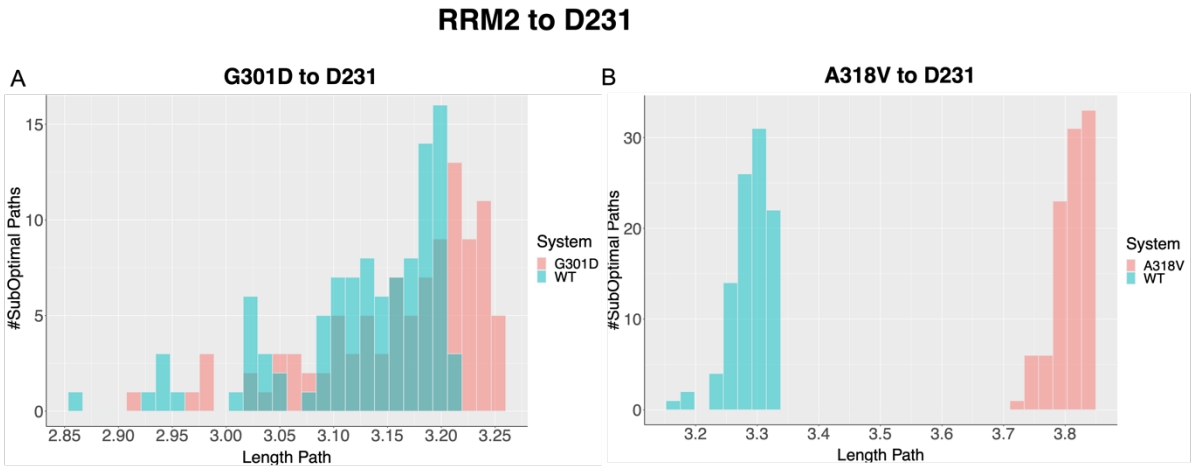


Figure 5.15 Distribution of interdomain signaling-path lengths from all investigated mutations on RRM2 toward Asp231@RRM1 compared to the WT distribution. In pink the distribution of paths within the mutant system is shown. In cyan the WT distribution is displayed.

As well, we also attempted at identifying the effect of all mutations on Gln333@RRM2. Interestingly, both the intra RRM2 and RRM1 to RRM2 signaling routes in most mutants systems are again longer or poorly affected as compared to the WT system (Figure 5.16, Figure 5.17).

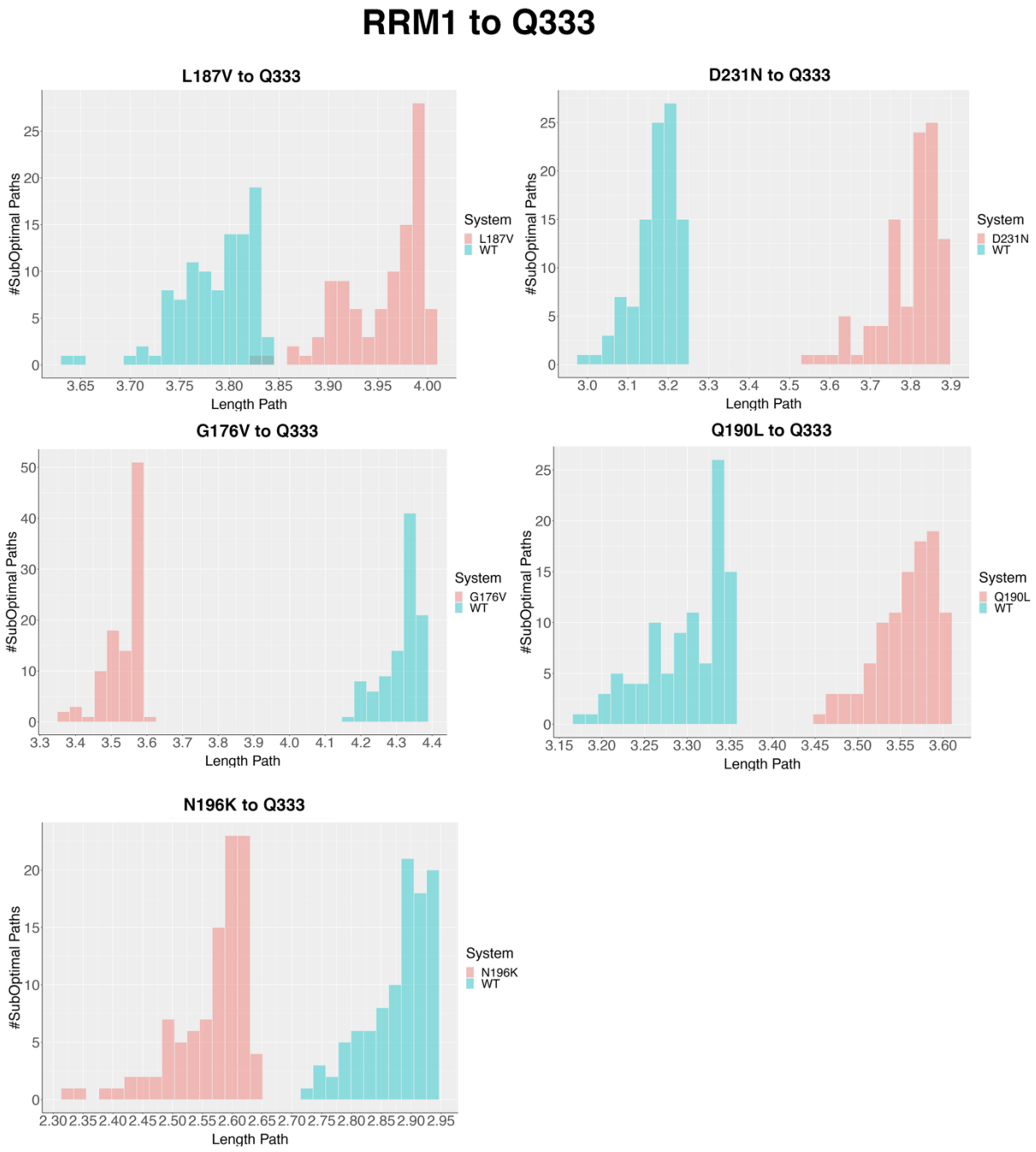


Figure 5.16 Distribution of interdomain signaling-path lengths heading from the mutations site on RRM1 toward Gln333@RRM2 compared to the WT distribution. In pink the distribution of paths within the mutant system is shown. In cyan the WT distribution is displayed.

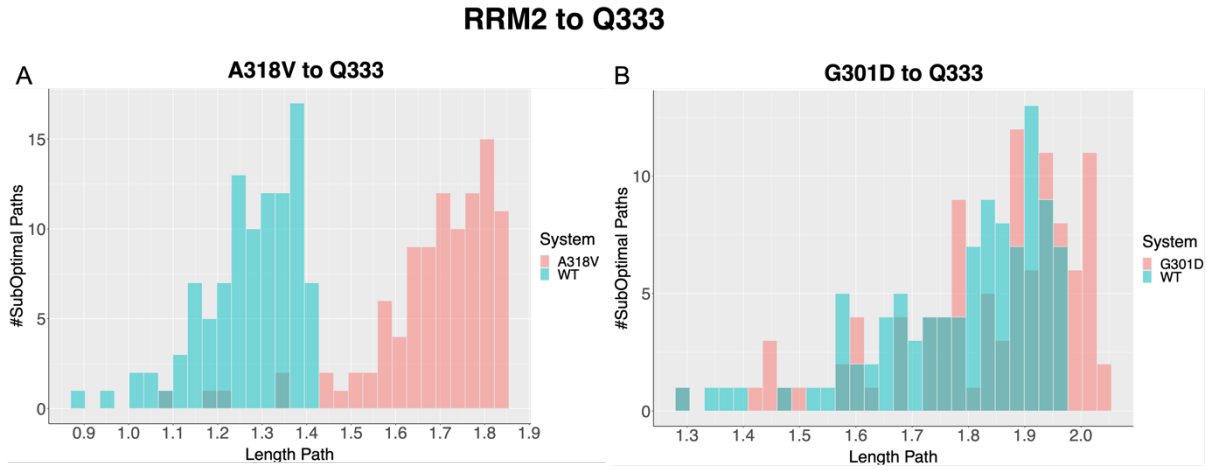


Figure 5.17 Distribution of intradomain signaling-path lengths heading from the mutations site on RRM2 toward Gln333@RRM2 compared to the WT distribution. In pink the distribution of paths within the mutant system is shown. In cyan the WT distribution is displayed.

As a result, it is clear the mutations perturb the internal signaling of the U2AF2 protein, and it reduces the correlation coupling between the mutation sites at cardinal RNA recognition hotspots of all domains.

Hence, in order to pinpoint the most relevant residues mediating the communication between the two domains, we computed the node degeneracy (i.e., number of times a node is present in the first 100 calculated signaling paths). Namely, in analyzing this salient trait along the routes heading from the RRM2 mutation sites to Arg150@RRM1, we observed that the pathways pass through common nodes residues (Figure 5.18 A) involving Thr252, Arg227 and Leu151. Moreover, comparing these routes with the WT (Fig. B.10), Thr252 and Arg227 show an increase in degeneration for both A318V and G301D mutations.

Then, we next assessed the node degeneracy of the paths heading from RRM2 mutations to Asp231@RRM1 (Figure 5.18 B). In addition to Arg228 and Pro229, Arg227 and Thr252 were again the most degenerated residues along the computed pathways. Some of these residues appear to be similarly degenerated also in the WT system (Arg227, Arg228, Peo229), however Thr252's degeneration increases again in the G301D mutation (Fig. B.10).

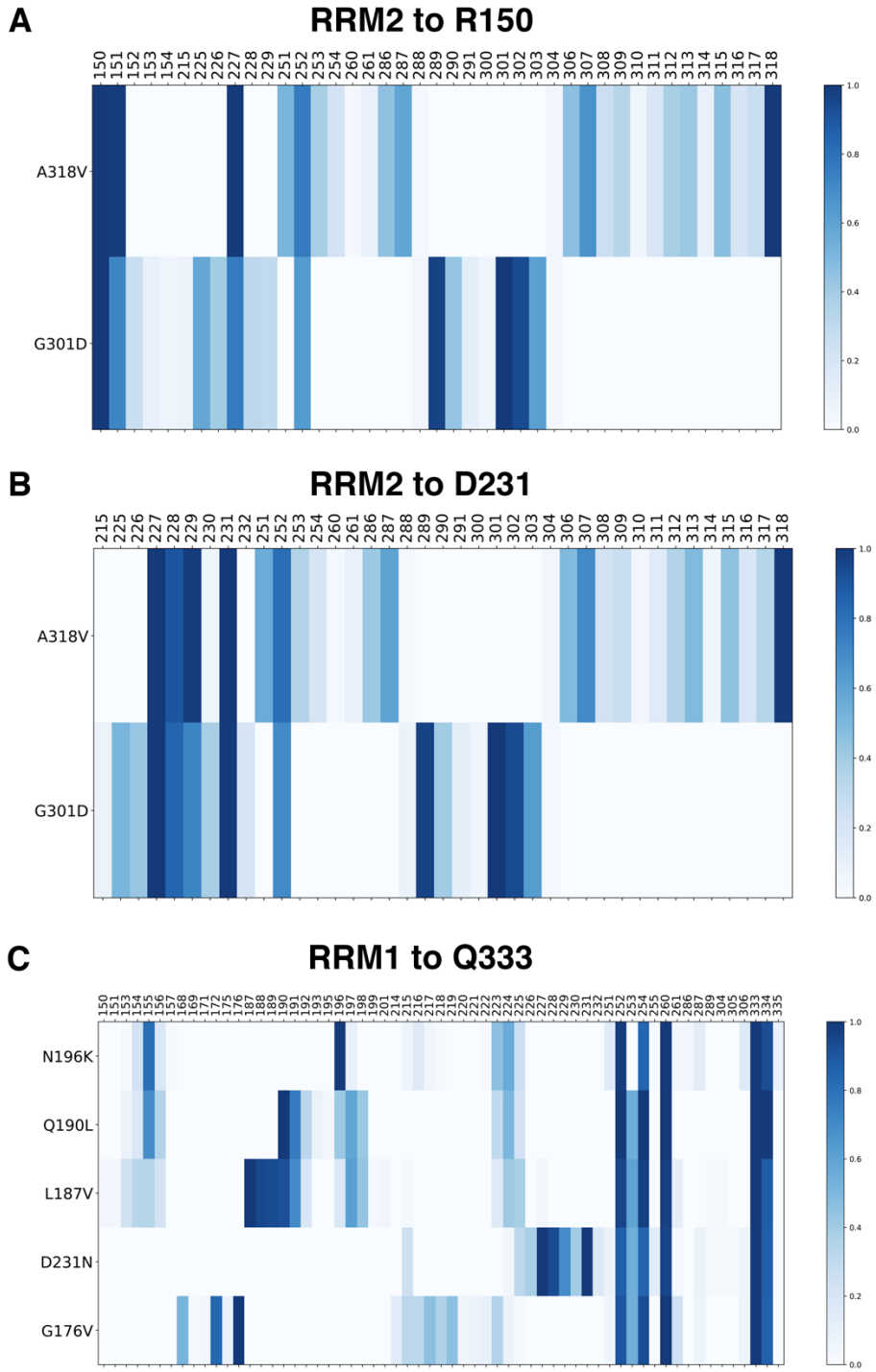


Figure 5.18 Matrices of common degenerated residues for interdomain pathways from (A, B) all mutations on RRM2 toward residues Arg150 and Asp231 on RRM1, (C) and from all mutations on RRM1 to Gln333 on RRM2, ranging from 0 to 1.

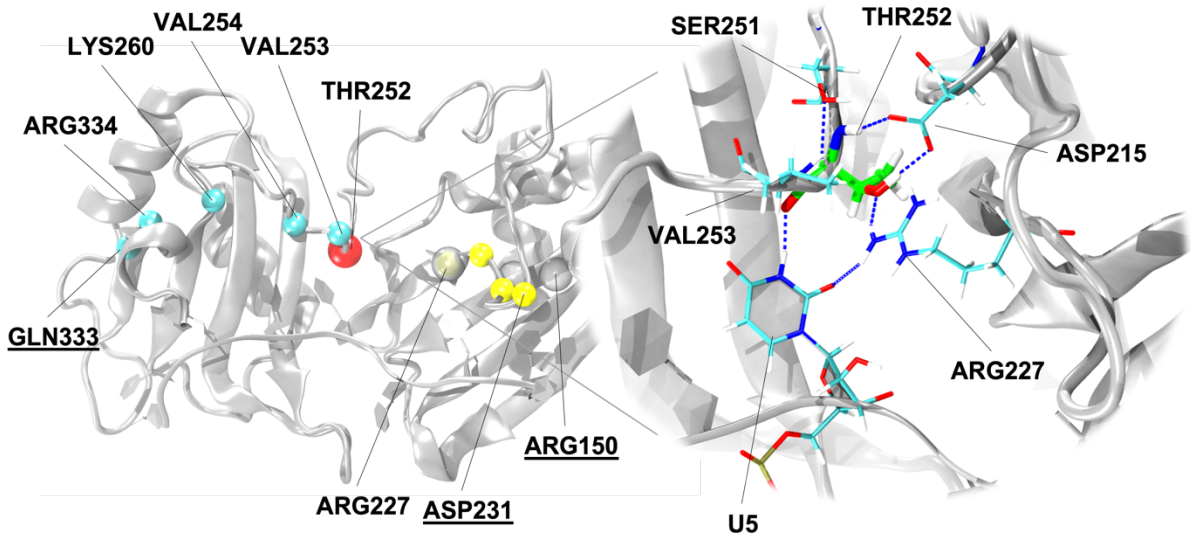


Figure 5.19 Common residues among degenerated interdomain paths between RRM1 and RRM2. In van deer Waals (vdw) are depicted the degenerated residues among the paths from RRM1 to Gln333 (Cyan), from RRM2 to Arg150 (transparent white) and from RRM2 to Asp231 (yellow). In red vdw is shown Tht252, as it found in all of the three cases, and an inset represent its H-bond interactions with the surrounding residues in licorice.

Finally, we computed the node degeneracy of the opposite signaling paths, i.e. those heading from RRM1 mutations sites to Gln333@RRM2 (Figure 5.18 C), and we pinpointed as recurrent residues along the signaling-path again Thr252, Val253, Val254, Lys260, all placed on the linker-loop, and Arg334@RRM1, all residues that again markedly show an increase of degeneration respect to the WT system (Fig. B.10), strongly suggesting this route to be the most affected by mutations.

Remarkably, the interdomain communication for different mutant systems always flow through Thr252, placed on the linker loop, that is the only residue fully degenerate in all the three interdomain pathways. As such this represents a critical crossroad at the interface of the two domains and strikingly mediates the interactions and the communications. This residue is of pivotal importance due to its engagement in persistent H-bond interactions with the surrounding Arg227, Asp215, Ser251 residues and U5 (Figure 5.19). As well, Arg227, as another degenerate node residue, establishes H-bonds with Thr252 and U5 nucleotide. Finally, Val254 is also a degenerate residue that establishes stable H-bonds with the U5 base through its backbone. Remarkably, its importance is confirmed by experimental studies

showing that Val254 stabilizes pre-mRNA binding and plays a key role for the cooperative pre-mRNA recognition by the tandem U2AF2 RRMs [197] .

5.4.4 Summary

Aiming to dissect the molecular basis for the selectivity of RRM2 domain of the U2AF2 protein towards poly-U sequences, we investigated the binding/dissociation mechanism of the poly-U facing the most selective RRM2 domain via plain and enhanced sampling simulations. As a first result we unexpectedly observed that the RRM1, known as to be less discriminant between the poly-U vs poly-C sequences, is instead the domain which more strongly contributes to pre-mRNA binding.

Noticeably, studying the dissociation of the poly-U strand, we identified an early and non-selective recruitment of the pre-mRNA where the Lysines (K300 and K340) residues early recruit the pre-mRNA strand by engaging salt bridge interactions with its backbone. Conversely the selectivity is most likely assessed in the final bound state when the RNA bases establish specific H-bonding and hydrophobic interactions with the protein residues. As such, while highly flexible and positively charged Lysines are used to grasp and catch phosphate groups of RNA backbone, π-stacking is used to stabilize the bases into position for hydrogen bonding with the linker. Remarkably, a marked difference is observed in the distribution of Lys and Arg residues in the two domains. While Arginines are largely found on RRM1, Lysines are more abundant on RRM2. As such Args appear to establish the strongest and diverse interactions (H-bonding and stacking) which may be responsible of the increased promiscuity of the RRM1 domain, which can recognize and bind both poly-U and poly-C Py-tracts, decreasing its selectivity. Additionally, a detailed analysis of the binding free energy pinpoints critical π-stacking interactions engaged by the residues of both RRMs with the RNA bases, demonstrating this type of interaction to stabilize the most pre-mRNA binding.

Remarkably, when investigating the impact of cancer-associated U2AF2 mutations, we interestingly identified some intriguing commonalities. Indeed, all mutations located on the RRM1 domain stabilize the pre-mRNA due to direct and indirect H-bonding and electrostatic

interactions of the Asp231 and Glu201 residues. Asp231 in fact interacts with the U8 base and Tyr152, which in many mutant systems establishes a π-stacking interaction with the U7 base of the pre-mRNA. Instead Glu201 salt bridges to two Arginines (Arg150 and Arg147) close to the RNA extremity. Conversely, the mutations placed on RRM2 domain destabilize the binding of the pre-mRNA, inducing a remodeling of Leu331 which in turn affects the U1 and U2 of the pre-mRNA via a steric hindrance. Conversely, Glu333 in the presence of all mutants is consistently object of stabilization.

Remarkably, all mutations influence pre-mRNA binding free energy by affecting three cardinal points of the U2AF2/polyU interactions which are Asp231, Glu201 and Arg150 and, partially Gln333 and Leu331 on linker-loop, RRM1 and RRM2, respectively. This suggests that a communication between the different U2AF2 domains must take place. As such, we attempted at finding the optimal and suboptimal communication pathways between the mutations and these cardinal points, selecting the residues that directly H-bonds with the pre-mRNA filament. Surprisingly, although the overall path does not change between the WT and the mutant systems, in most mutants signaling paths are less effective (longer), perturbing as a result the interplay of RRM1/RRM2 towards pre-mRNA binding. Moreover, we also dissected the cardinal residues involved in the communication between the two recognition domains by computing their degeneracy among the computed paths.

Our findings highlight the importance of Thr252 placed at the crossroad of the two domains as a central and conserved node of the interdomain signaling. This node mediates the signaling perturbation between the distinct U2AF2 domains in the presence of the WT system and in all cancer-associated mutations. Thr252@RRM2 is involved in strong H-bonding interactions with Arg227@RRM1 and U5@pre-RNA, thus being a critical crossroad point in the communication of the two domain and contributes to stabilize the open (RNA binding prone) conformation of U2AF2 and RNA binding.

Overall, the molecular dynamics simulations reported in this chapter have helped to clarify at atomic-level key functional aspects of U2AF2, providing important information for further experimental studies.

5.4.5 Further Perspectives

This chapter was a first attempt in unraveling the recognition mechanism underlining the U2AF2 mode of action toward the Py-tract. Additional enhanced sampling simulations will have to be carried out on the poly-C strand to enable a direct comparison of the mechanism underlying the selective recognition of the poly-U over the poly-C strand. Complementarily, MTD simulations will have to be devoted to study the mechanism of selection of the pre-mRNA also for the RRM1 counterpart, which is the least selective domain, even though our results on multiple mutant systems have surprisingly revealed that this domain more strongly contribute to stabilize the poly-U binding as compared to the RRM2 one. These additional simulations would provide a complete picture of the recognition mechanism. Finally other plain MD simulation of the mutant systems in complex with a poly-C strand would clarify the differential impact that these mutations would have on a weak or strong Py-tract.

6 Communication Network within the Spliceosome

6.1 Abstract

Intron splicing of a nascent messenger RNA transcript by spliceosome (SPL) is a hallmark of gene regulation in eukaryotes. SPL is a majestic molecular machine composed of an entangled network of proteins and RNAs that meticulously promotes intron splicing through the formation of eight intermediate complexes. Cross-communication among the critical distal proteins of the SPL assembly is pivotal for fast and accurately directing the compositional and conformational readjustments necessary to achieve high splicing fidelity. Here, Molecular Dynamics (MD) simulations of an 800,000 atoms model of SPL C complex from yeast *S. cerevisiae* and community network analysis enabled us to decrypt the complexity of this huge molecular machine, by identifying the key channels of information transfer across the long distances separating key protein-components. The study reported in this chapter represents an unprecedented attempt in dissecting cross-communication pathways within one of the most complex machines of eukaryotic cells, supporting the critical role of Clf1 and Cwc2 splicing cofactors and specific domains of the Prp8 protein as signal conveyors for pre-mRNA maturation. Our outcomes provide fundamental advances into mechanistic aspects of SPL, providing a conceptual basis for controlling the SPL via small-molecules modulators able to tackle splicing-associated diseases by altering/obstructing these information-exchange paths.

6.2 Introduction

Complex and sophisticated conformational remodeling underlies the function of many types of biological systems [218]. Conformational changes are critically entwined with promotion

and regulation of information-exchange within biomolecules and bimolecular aggregates. Nevertheless, decrypting the pathways and the mechanisms modulating their cross-communication at atomic-level remains as of yet challenging[96], [98], [99], [219]–[221], especially when tackling large macromolecular machines. Here, we face this challenge on the spliceosome (SPL), which, in eukaryotes, promotes premature-messenger (pre-m) RNA splicing, hence being a key modulator of gene expression and diversification. The SPL is a multi-megadalton machine composed of an intricate network of hundreds of proteins and five small nuclear RNAs (snRNAs) (U1, U2, U4, U5 and U6) organized into small nuclear ribonucleoproteins subunits (snRNPs). The splicing cycle proceeds via the formation of at least eight intermediate SPL states (i.e., A, B, B^{act}, B*, C, C*, P, ILS), leading to the release of mature (m)RNAs upon excision of the non-coding regions (introns) from primary RNA transcripts and ligation of the protein coding segments (exons). The SPL conducts this pivotal step of gene expression by recognizing three key intronic sequences, the 5' and 3' splicing sites (5'SS and 3'SS, respectively), which delimit the intron boundaries, and the branch-point adenosine (BPA) lying within the branch point (BP) site. Splicing is accomplished via two subsequent trans-esterification reactions co-adjuvated by two catalytically active Mg^{2+} ions[222]. In the first step, a free upstream exon and an intron-lariat (IL), named hereafter as intron-lariat exon intermediate (ILE), is formed. While, in the second step, the exons ligation and IL release takes place. An idiosyncratic trait of the SPL is its marked structural plasticity, which promotes splicing thanks to a relentless conformational and compositional reshaping of its snRNPs. This latter is mediated by a sophisticate and precise signaling networks. The mechanistic understanding of SPL function is burgeoning due to the cryogenic electron microscopy (cryo-EM) structures solved at near-atomic-level resolution from both human and yeast strains[223]. All-atom molecular dynamics (MD) have supported and amplified the impact of cryo-EM data by dissecting the functional dynamics encoded into its distinct proteins/RNA components[224]–[228]. In this chapter, we provide a ground-breaking advance in the field by unprecedentedly addressing the molecular origin of signal transfer within the SPL machinery, which underlays its complex functional transitions. To this end we performed all-atom MD simulations of the C complex from yeast S. Cerevisiae, as a prototypical example of the SPL assembly, for a cumulative statistic of 6 μs, complemented

by correlation and community network analyses. Our approach allowed decrypting the crosstalk channels for the information transfer between the distal (160 Å) Clf1 and RNAse-H domain of Prp8, functional for the transition from the investigated state (C complex) towards the subsequent intermediate of the cycle (C* complex) and necessary to accurately promote gene expression. Our results are conducive to resolve the puzzling scenario underlying signal communication within the SPL and assert the critical role of computer simulations to dissect the mechanisms of complex molecular machines at atomic level. Harnessing this knowledge may open the tantalizing perspective of identifying small-molecules modulators able to interfere with the SPL's signaling pathways as a novel strategy to fight the nearly 200 human diseases associated to splicing deregulation.

6.3 Methods

6.3.1 Model construction

The simulations were based on the S. cerevisiae C complex cryo-EM structure at a resolution of 3.8 Å in average (PDB ID: 5lj3), with components reaching a resolution of 3.4 Å. This model is composed by three functional snRNAs (U2, U5, U6), 5' exon filament and the intron lariat-exon (ILE) junction intermediate where O2' of the BPA has already reacted with the phosphate group of the first intron base. The model also comprised 15 proteins. In detail, the included proteins are Prp8 and Snu114 (from U5 snRNP), Cef1, Isy1, Clf1, and the splicing factors Yju2, Cwc25, Cwc21, Cwc22, Prp45, Prp46, Cwc15, Bud31, Cwc2, Emc2. Four Mg2+ ions were originally present in the structure. However, since the structure of the B^{act} complex (PDB id: 5gm6)[34] shares an almost identical active site in which five ions are present, we recovered a fifth ion from the B^{act} structure. The presence of a five-metal ion motif was later confirmed in other steps of the SPL cycle[229]–[231] and only in the presence of this additional metal ion we were able to achieve a stable active site architecture. Overall, also considering 5 Mg^{2+} ions and 7 Zn^{2+} ions originally present, the counterions and the explicit water molecules, our SPL model consists of 772,682 atoms. In order to find a compromise between system size and accuracy, we discarded all the peripheral proteins due

to their incomplete chains, their low resolution, and the presence of multiple gaps. Small gaps (about 14 residues long, besides one exception of 46 residues long) in the loops within the retained regions were instead modelled with *De novo* model building, as implemented in Modeler 9, version 16 [232]. The loops were first selected among 50 models according to their DOPE (Discrete Optimized Protein Energy) score and subsequently evaluated through an accurate visual inspection.

6.3.2 MD Simulations

MD simulations were carried out with the Gromacs 5.0.7 suite [233] using the most tested force field (FF) for proteins/RNA complexes. Namely the AMBER-ff12SB [166] was used for proteins, whereas ff99+bsc0+χOL3 FF was used for RNAs [234], [235]. This protocol has been validated in other protein/RNA macromolecular complexes [98], [99], [224], [225], [236]. For Mg^{2+} ions we used dummy cation parameters, developed by Saxena et al [237], since according to our benchmarks this parametrization best reproduces the structural features of sites hosting several Mg^{2+} ions in close proximity within RNA structures [238]. Na^{+} ions parameters were taken from Joung and Cheatham [212], while Zn^{2+} ions were modeled with the cationic dummy atoms approach developed by Pang et al [239]. The system was embedded in a 12 Å layer of TIP3P water molecules leading to a box size of 196 x 220 x 200 $Å^3$, and 201 Na^{+} counter ions leading to 772,679 atoms. The topology was built with the tleap module of AmberTools 16 and later converted into the GROMACS format by using the acpype program [240]. We carefully equilibrated the system to maintain unaltered coordination of the active site. We initially performed a minimization step with the steepest descent method of 1000 steps, up to a convergence criteria of 1000 kJ/mol nm of maximum force. Next, we gradually heated the system to 300 K with an increase of 50 K every 2 ns for a total of 12 ns, keeping the entire system highly restrained (1000 kJ/mol nm^2) except for the solvent and solute hydrogens. Then, we switched to the NPT ensemble, scaling the pressure to 1 bar and using two different barostats: (i) the Berendsen barostat was used for 20 ns with the same restraints on the atoms and (ii) the Parrinello-Rahman barostat for 30 additional ns, while leaving the side chains free of constraint. Next, we gradually decreased the restraints in 20 ns. Finally, we performed the simulations of five replicas for 1 μs each. One of the

replicas was later extended to 2 μs to inspect and assess convergence issues of the PCA. In each replica, we used the same starting structure, while the velocities were differently initialized. The root mean square deviation, (RMSD), root mean square fluctuations (RMSF) and radius of gyration (Rg), hydrogen (H)-bond analysis, as well as covariance matrix were computed with cpptraj module of AmberTools16 [241].

6.3.3 Principal Component Analysis

This statistical technique is used to filter out the vibrational noise and redundant/non-relevant conformational transition in MD simulations, while capturing the essential motions hidden behind an MD trajectory. These correspond to the lowest frequency motions, usually responsible of the large conformational transitions, which modulate biological functions. PCA relies on the calculation of the covariance matrix. This is calculated from the atoms' position vectors after an RMS-fit on the first frame of the MD trajectory to remove translational and rotational motions. One average structure over the aligned trajectory is computed and a covariance score with respect to this average is assigned to each mass-weighted Cα and P atoms, obtaining a matrix where each element represents the covariance between each pair of atoms, *i* and *j*, defining the *i*, *j* position of the matrix. The covariance is defined as:

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

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. The matrix is then diagonalized to find the eigenvectors, or Principal Components (PC), and their corresponding eigenvalues. These represent the directions of the motions and their associated amplitude (i.e., the eigenvalues represent the extension of the fluctuations around the average structure along the eigenvector direction). As a result, the projection of Cartesian coordinates vectors onto the eigenvectors (i.e., by taking the dot product between the two vectors at each frame), allows reducing the dimensionality and the noise hidden behind an MD trajectory, to capture and visualize the most relevant

motions sampled during the simulations. The PCA has been extensively and successfully applied in many distinct applications of biological systems, however, it is well known that limited sampling may reduce the confidence in most representative motions identified in the sampled trajectories [242], [243]. At variance with previous simulations of the SPL in which the multi-replica approach was employed to validate at qualitative level the reproducibility of the results[224], [225], here PCA has been conducted on single trajectories as well as on merged trajectories (Fig. C.1, Fig. C.2), generated by concatenating three or five replicas[244], [245] in order to increase the sampling of this large system. Nevertheless, due to their limited sampling overlap, we have also verified that the observed essential dynamics projected on both PC1 and PC2 were independent from the manipulation of the trajectory (i.e., if this pseudo-trajectory was qualitatively alike to that decrypted from the single replicas). Since the main results for single replicas, the 5- and 3-replicas trajectories were similar, we discussed only the results obtained from the 3-replicas pseudo-trajectory due to its more portable format in post-processing analyses. Projecting the coordinate onto the PCs and plotting PC1 and PC2 generate a scatter plot displaying how the conformational space defined by the first two modes is sampled through the MD simulations (Fig. C.3). The scatter plot of the 5-replica trajectory shows that each of them samples different points of the free energy landscape. Hence, the reference structure to which all trajectories were aligned was the starting structure of the simulation, which is common to all replicas. For each replica, the matrix was calculated on 4804 Cα, 270 P atoms, and considering 15,000 frames, corresponding to last 750 ns of the MD simulations. Here, we discuss the essential dynamics obtained from PC1 and PC2 representative of most of the variance (30-50 % in all single and combined replicas trajectories) (Fig. C.4) .The Normal Mode Wizard plugin in VMD [246] was used to visualize the PC1 and PC2 along the principal eigenvectors and to draw the arrows highlighting their direction.

6.3.4 Cross Correlation Matrix

A straightforward way to normalize the covariance matrix is by using the Pearson's coefficient, giving as a result a cross-correlation matrix based on the Pearson correlation coefficient (CC_{ij}).

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

These were calculated with the cpptraj module of AmberTools 16 [241] for the combined replicas trajectories as well as for each single trajectory, after verifying that the results obtained from the combined version were qualitatively similar to the single replicas. This matrix allows us to qualitatively interpret the inter-residue pair correlations by measuring the linear correlations of atomic motions. These coefficients span from a value of -1, which corresponds to an anti-correlation motion between two residues, to a value of +1, which instead corresponds to a fully linearly correlated lockstep motion. Zero values indicate uncorrelated motions. In complex macromolecular systems, this matrix can be reduced into a coarse and simplified version, for clarity reasons, where each pair of proteins (matrix blocks) and domains considered are averaged over the number of residues in order to find a 'correlation density', allowing to easily decrypt the principal correlations. Due to the large size of Prp8 and Clf1 proteins, to better pinpoint their functional role, we separately considered each domain and HAT repeat, respectively.

6.3.5 Weighted protein network and community analysis

Pearson correlation coefficients is still limited to linear correlation and orientation dependent movements. Using the position vectors of Cα atoms along the MD trajectories previously described, we then computed the $^{LMI}CC_{ij}$ as implemented in the the GROMACS v4.6.4 package[233] and described in more details in Chapter 0. The CCs based on the LMI are hereafter referred as ^{LMI}CCs. In this case, before summing ^{LMI}CCs to generate the correlation scores (^{LMI}CSs) in the coarse matrix, a threshold was applied to filter the noise, retaining only ^{LMI}CCs values larger than 0.6 (Fig. C.5). The RMSD matrix calculated between CCM and

LMI matrices (Fig. C.6) displays graphically how much the LMI matrix differs from the more standard Pearson-based version. In this case LMI fills the voids of undetected orthogonal motions, but still retains the correlation captured by the CCM, proving to be a complementary approach to Pearson coefficients. Whereas the former is able to more quantitatively determine the correlations, the latter adds a qualitative picture of the directions of parallel correlated motions. Remarkably, the mean RMSD value between the two matrices is 0.31. The ^{LMI}CCs, are more reliable than the Pearson's CCs and are linked to the information content retained in the protein motions. Subsequently, we employed the communication network analysis described in the methods section of this thesis (Chapter 3) to obtain an intuitive picture of the complex internal communication network. This exploit a correlation-weighted protein communication graph that is partitioned in clusters of nodes (communities) by using the edge betweenness criterion and modularity measure. The optimum community structure obtained for the SPL has a modularity of ca. 0.8, in line with the common range observed for 3D structure (i.e., 0.4–0.7).

6.3.6 Electrostatic calculations

Electrostatic calculations were performed with the Adaptive Poisson-Boltzmann Solver (APBS) software [248] on selected frames of the C model as extracted from the cluster analysis of the MD trajectory. APBS calculations were carried out using the Linearized Poisson-Boltzmann Equation (LPBE) in the VMD software with the following settings: surface density of 10.0 points/Å^2, solvent radius of 1.4 Å, system temperature of 298.15 K, solute dielectric constant of 2.0, and solvent dielectric constant of 78.54 with smoothed molecular surface.

6.4 Results

6.4.1 Structural and dynamical properties of the SPL C Complex

The system investigated here is based on the cryo-EM structure of the C complex SPL from S. cerevisiae, solved at an average resolution of 3.8 Å (PDB ID: 5LJ3)[32]. This SPL model (Figure 6.1, Tab. C.1) encompasses 15 proteins, 3 snRNAs (U2, U5, U6), the ILE intermediate and the 5'-exon, as well as 5 Mg^{2+} ions and 7 Zn^{2+} ions. Thus, in the presence of explicit water molecules, our SPL model consists of 772,682 atoms. Five µs-long all-atom MD simulations in explicit solvent have been performed in order to trace the signaling pathways present within the SPL. The structural convergence of the simulation was achieved in each replica within 120 ns, as shown by the analysis of the RMSD, the gyration radius and the active site architecture (Fig. C.7, Fig. C.8, Fig. C.9). The SPL structure explored here catches the ILE intermediate immediately after the first splicing step has occurred (Fig. C.10) and the ILE is stabilized by the formation of an intricate H-bond network to U6 and U2 snRNA, respectively.

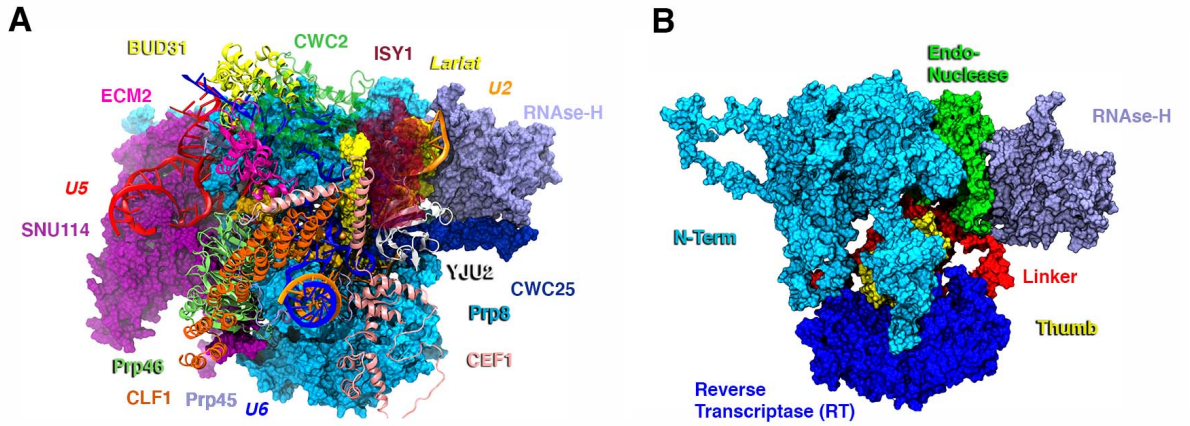


Figure 6.1 (A) Model of the C complex spliceosome from the yeast S. cerevisiae cryo-EM structure (PDB entry: 5lj3). Clf1, Prp46, Cwc2, Ecm2, Prp45, Bud31 proteins are shown as orange, light green, green, dark pink, lilac and yellow new cartoons respectively. Cef1, Yju2, Isy1, Cwc25, contributing to the intron lariat exon (ILE)'s stabilization, are depicted as pink, white, dark red (transparency), dark blue, respectively. ILE is shown as yellow surface, while Mg^{2+} ions are depicted as orange van der Waals spheres. U5, U2, U6 snRNA are shown as red, orange and blue ribbons, respectively. Prp8, its RNAse-H domain and Snu114 are displayed in light blue, lilac and magenta surface respectively. (B) Domain subdivision of Prp8, with RNase-H, Endonuclease, N-terminal (Nterm), reverse transcriptase (RT), thumb, linker domains shown in lilac, green, light blue, blue, yellow, red surface, respectively. Reprinted with permission from [249]*, Copyright 2021, American Chemical Society.*

To better extricate the complexity and disclose the critical proteins underlying the C complex functional dynamics, we have initially computed the cross-correlation matrix (CCM) based

on Pearson's correlation coefficient (CCs) from the combined-replicas trajectories along with its coarse-grained variant (see Methods and Fig. C.1, Fig. C.2), to more easily identify the dynamically coupled regions [98], [224], [225], [250], [251]. Next, we performed the Principal Component Analysis (PCA) (Fig. C.3, Fig. C.4) to extract the essential dynamics of the SPL C complex from the MD trajectory. This analysis allowed us to draw out the functional motions associated to the CCM and to visualize which protein component/domain collectively contributed to it. The essential dynamics obtained from PC1 reveals that (i) Clf1 and RNAse-H move lock-step in a hammer-like motion by contracting the SPL core and enabling, as a consequence, the movement of Cwc2, which acts as a mediating factor (Fig. C.1). (ii) The essential dynamics related to PC2 underlines a second cooperative movement of Clf1 and RNAse-H domain, which undergo a twist of the α-helices and a marked rotation (Figure 6.2 B), respectively. (iii) Additionally, by inserting its β-finger motif into the U2/IL helix (in PC1), the RNAse-H domain promotes the wrapping of the U2/IL branch helix (Figure 6.2). This movement is regulated by electrostatic interactions, namely N1869 from

the RNAse-H β-finger, and K22, K26, K30 from Yju2 α-helix, which interacts with the IL bases and the phosphate backbone of U2 (Fig. C.11), respectively.

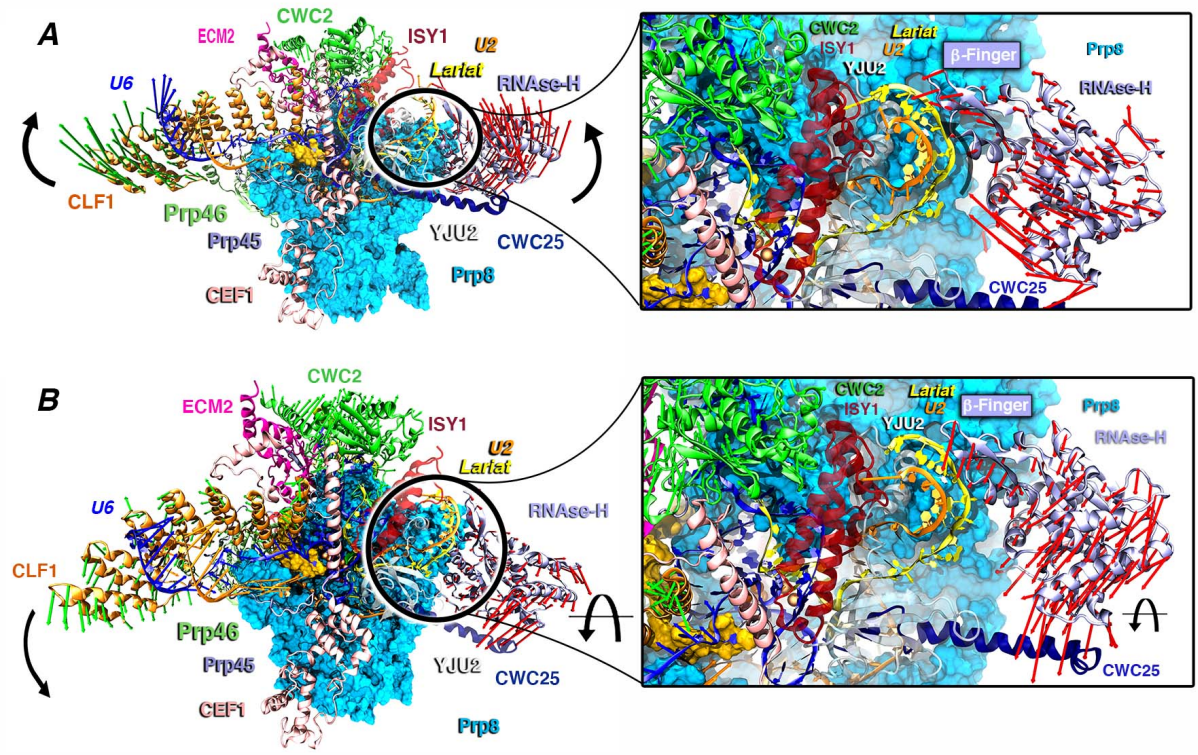


Figure 6.2 Essential dynamics as extracted from the principal component analysis (PCA) of the combined 3-replicas pseudo-trajectory. Red and green arrows depict the type and the direction of the motions. (A) Principal Component 1. Clf1 (orange and RNAse-H domain (lilac) are the arms of the hammer-like movement towards Cwc2 (green). The inset captures the wrapping of the Intron-lariat (IL)/U2 double helix promoted by the β-finger motif of RNAse-H domain. (B) Principal Component 2. Clf1 and RNAse-H domain perform a downward rotation. U2 and U6 snRNA are shown as orange and blue new cartoon, respectively. The inset focuses on the rotation of the RNAse-H domain toward Cwc25 (dark blue). Prp8 and its RNAse-H are depicted as light blue surface and lilac new cartoons, respectively. IL, Cwc2, Cwc25, Cef1, Clf1, Prp46, U2 and U6 snRNA are shown as yellow, green, dark blue, pink, orange, light green, orange and blue new cartoons, respectively. Reprinted with permission from [249]*, Copyright 2021, American Chemical Society.*

Consistent with the critical importance of β-finger pinpointed by our simulations, biochemical studies disclosed that four missense mutations of this motif (V1860D, T1865K, A1871E and T1872E) affect the transition between first and second splicing step [252]. Among these, Val1860, Ala1871, and Thr1872 lie nearby the negatively charged RNA backbone of the U2/IL helix. Ostensibly, our simulations suggest that these mutations most

likely impair the second step of SPL catalysis by altering the U2/IL wrapping and displacement. As observed in PC1 and PC2, the structural superposition of the cryo-EM structure of the C and C* complexes (Fig. C.12) solved by cryo-EM (PDB id: 5WSG),[229] reveals that the rotation of RNase-H (PC2) and the wrapping of IL/U2 branch helix (PC1) are clearly in line with the positions they adopt in the C* aggregate. Although the complete rotation of the β-finger motif and of the RNAse-H domain is hindered in our MD simulations by the presence of the C-complex stabilizing factors (i.e., the Yju2, Cwc25 and Isy1 proteins), the β-finger motif appears to rearrange the U2/IL helix. This latter is, indeed, expected to remodel, creating the room necessary to load the second reactant (the 3'-exon) for the subsequent exon ligation step [230]. Remarkably, from this structural comparison it also clearly appears that the Clf1 helices, along with those of the Syf1 protein, create an arch connecting the large portion of Prp8 (N-term and RT domains) to U2snRNP (Lea1, Msl1 and the Sm-ring). This latter has to detach from the RNAse-H domain's surface to enable the transition from C to C* complex [230]. Hence, Clf1 may act as a protruding arm connecting the SPL core to the most peripheral proteins, possibly contributing to displace the U2snRNP from the RNAse-H surface via a rotation around the its own pivot located at the HAT-repeat H2-H3, as enlightened by the CCM analysis (Fig. C.1, Fig. C.13). As such, the hammer-like motion exerted concertedly by Clf1 and the RNAse-H domain appears to be instrumental for the progression of the SPL's cycle.

6.4.2 Dissecting the pathways of signal transfer.

SPL is a large and highly plastic machine vitally regulated by signal transfer between the central scaffold of the snRNPs and the distal proteins. In order to decrypt the mechanism of information exchange in charge of the functional motions detailed above, we have employed protein network methods by performing a community network analysis (CNA) on our MD simulations. This approach enabled to trace the signaling routes responsible for the communication between the critical regions of the C complex assembly (i.e., the Clf1 proteins and the RNAse-H domain). The CNA methodology [96] relies on a protein-based weighted network where the nodes, representing the Cα atoms of amino acids, are connected by edges, whose weighs depend on the correlations of residues' pairs. Since CCM based on

Pearson coefficients lacks a fraction of correlation (see Methods), we exploited the mutual information approach [97] to accurately compute the CNA. The aforementioned communication network, built on the basis of the ^{LMI}CCs, can be then exploited to trace the most likely communication pathways connecting the regions critically entailed within the functional movements of the system. In fact, by identifying the protein communities, i.e. the groups of strongly correlated residues [101], the CNA provides a coarse-grained picture of the intercommunications happening among the distinct regions of the SPL machinery. As shown in Figure 6.3A, the protein communities can be graphically displayed as groups of correlated residues(Figure 6.3 C). The links connecting each pair of communities stand for the corresponding *inter-communities' edge betweenness* (IEB), i.e. the sum of the EBs of the pairs of residues connecting two adjacent communities, hence indicating the strength of the communication flow between two communities. Strikingly, CNA (Figure 6.3 C) reveals that Prp8 is involved in half of the totality of the SPL C complex's communities and that five of them, i.e. communities #2, #5, #7 #10, #11, with #2 and #11 almost fully coinciding with Prp8's endonuclease and RT domains, are the largest and the most connected communities in the whole C complex network. This depicts Prp8 as a signals conveying platform within the SPL proteins/RNA network. To exploit the insights provided by the CNA, we tackled the communication taking place between the distinct SPL components, with a focus on the information flow between Clf1 (community #19) and the RNAse-H (community #15), which are separated by 160 Å. As shown in Figure 6.3 A, several pathways might be involved in the communication between communities #19 and #15.

By considering the IEB links of community #19, the largest communication signal flows either via community #9 (comprising Ecm2 and part of Cef1) or community #10 (involving Prp46 and part of the large N-term of Prp8). Following the pathway via community #10 (Path I), the IEBs indicate that the information can easily flow through community #11 (comprising the RT domain and part of Cef1), finally reaching #15 via community #14 (corresponding to Yju2 and part of Cwc25). Alternatively, considering the pathway via community #9 (Path II), the IEB values indicate a strong communication with community #17 (located on Cwc2), from which the information flow could either remains on the same path II via communities #1 (involving Bud31), #2 (that is part of Prp8's N-term) and #3 (corresponding to Endo

domain) or heads towards Path III via communities #16 (associated to the Isy1 protein and part of Cef1) and then #14 to reach community #15. Of note, the physically shortest pathway (Path IV), i.e. the communication path along the shortest physical distance between Clf1 and RNAse-H, involves just communities #16 and #14. Nevertheless, the communication flow along this path is expected to be limited by the poor IEB between communities #9 and #16, and therefore it is unlikely. In order to extricate in more detail the communication between distinct communities, we analyzed the nodes characterized by the highest NB (see Methods) that represent the cardinal residues through which the majority of the communication travels, forming, therefore, the principal channel of information flow across the SPL components (Figure 6.3 D). Remarkably, most of the nodes characterized by the largest NB (Figure 6.3 D) belong to the most important communication pathways (Path I and Path II) as suggested by the CNA (Figure 6.3 A).

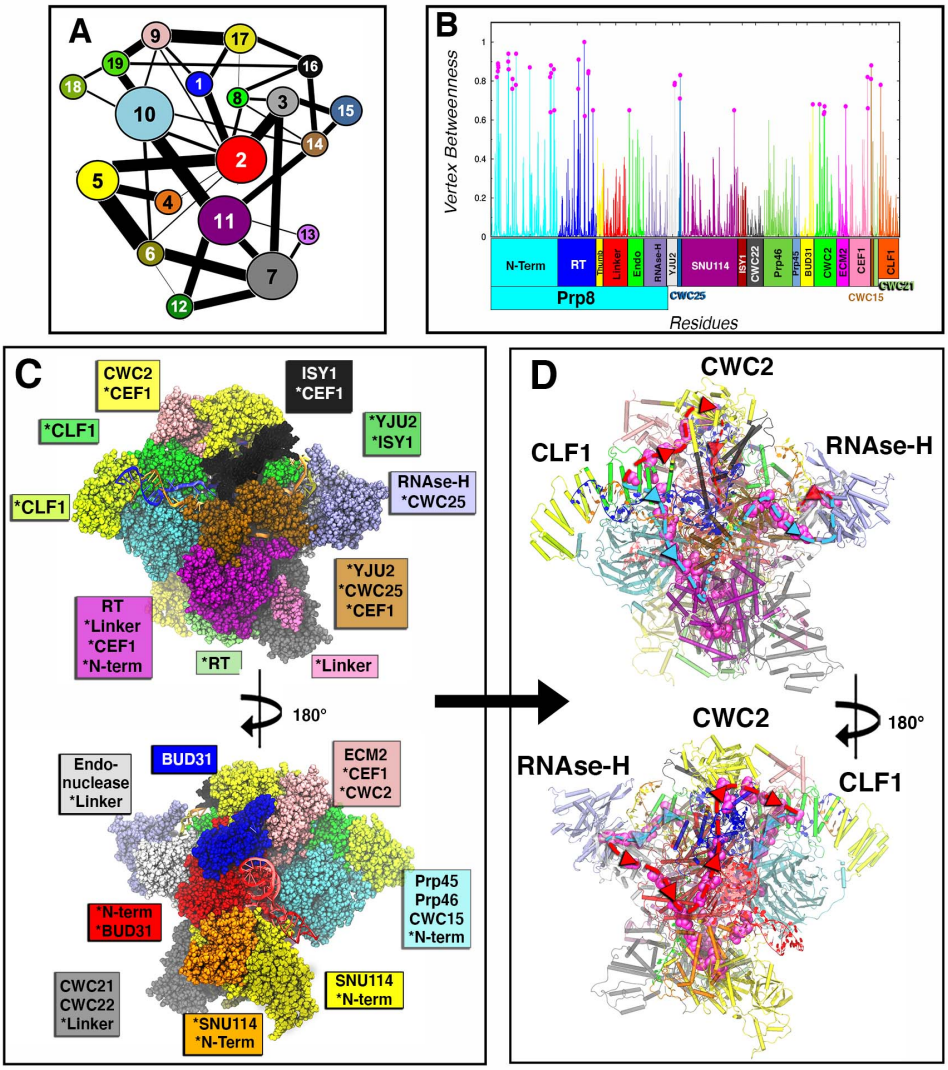


Figure 6.3 Community Network Analysis of the spliceosome. (A) 2D representation of the community network. The connecting links have a width proportional to the sum of all edges betweenness connecting two communities, thus measuring the corresponding inter-communities' communication flux. (B) Normalized per-residue node betweenness (line) color coded by protein domains as in Figure 1, and points with betweenness more than 0.6 in magenta. (C) 3D-structure (Front and Back) of the community network, color coded with the same color of the 2D graph. Asterisks indicate domains or proteins when spread in various communities. (D) Spliceosome communication highway (back and front). The residues with node betweenness higher than 0.6 are highlighted in magenta, thus displaying the two principal routes (path I in light blue and path II in red) for the communication flux through which most signaling occur. In cartoon are depicted the communities with the community color-code. Reprinted with permission from [249]*, Copyright 2021, American Chemical Society.*

By this analysis we could observe that a key point of the communication between the endonuclease and RNAse-H domains (in Path II) is the specific interaction between residues Lys1912 of RNAse-H and Asp1664 of Endo, located at the surface between the communities #3 and #15. A list of residues along the communication pathways (path I and II) is provided in , including the amino acids Ser13, Cys792, Asn1099, Gln558, Asn203, Thr205, Arg207, Ile209, Leu318 which are characterized by very high node betweenness, representing good candidates for point mutation experimental studies. The communication pathways characterized across the SPL also allow to propose potential binding sites for small molecules that could modulate the information exchange and, hence, be the target for virtual-screening studies. In particular, we have localized a possible binding pocket lying on the communication path II that is found in either open or closed state during the 'hammer-like' motion described by PC1 (Fig. C.16).
Both the CNA and the analysis of NB values indicate that the endonuclease and RNAse-H domains strongly communicate between each other, consistently with correlation pattern observed in the CCM. A similar strong communication has been detected between Asp216 of Cef1 and Arg62 of Clf1, acting as a possible signal bridge between these two elongated proteins composed of α-helices.
Noteworthy, also some residues at the core of Cwc2, i.e. Phe71, Leu106, Lys116, are characterized by very high EB values, asserting the importance of Cwc2 in the information flow. Overall, this information opens the avenues to future computational and experimental studies aiming at exploiting signal information exchange channels for an allosteric regulation of the spliceosome.

6.5 Conclusions

The spliceosome is a huge metallo-ribozyme composed of an entangled network of proteins and RNA filaments. Protein communication over the long distances of the SPL congregate is a vital requirement to enable the precise functional movements and the meticulous information flow essential for faithful splicing. The characterization at the molecular level of

the fundamental interactions, establishing communication channels in an immensely complex macromolecular assembly, such as SPL, is challenging and has not been previously attempted. In this chapter, we combined all-atom MD simulations with the community network analysis to detect the specific functional movements underlying the internal communication of the C complex, as a prototypical case of SPL. The reported analysis unravels how information exchange is facilitated by the inner SPL conformational plasticity. In particular, the essential dynamics describes a 'hammer-like' motion of two 160 Å distal proteins which most likely displaces the unneeded splicing factors for proceeding along the SPL cycle (Figure 6.2). The motion underlays the twisting/repositioning of the IL/U2 branching helix, possibly triggering the beginning of its conformational readjustment towards the position occupied in the subsequent SPL (C*) complex [230]. The experimental evidence provided by a comparison of the Cryo-EM maps of the C and C* intermediate states fully support our findings (Fig. C.11, Fig. C.12, Fig. C.13, Fig. C.14, Fig. C.15) [230]. Therefore, it is natural to conjecture that Prp8 is a key component in remodeling the substrate in the C complex, due to its β-finger motif in the RNAse-H domain. Most importantly, our network analysis of MD simulations discerned the communication channels underlying these functional movements of Clf1 and the RNAse-H. By connecting the nodes mostly involved in communication among different communities, we identified two most relevant paths - Clf1, Ecm2 Cwc2, Prp8's Endo and finally the RNase-H domain (Path I), or Clf1, Prp46, the large N-term/RT domain of Prp8, Cef1, Yju2, Cwc25 heading towards the RNase-H (Path II) - (i) disclosing a critical participation of Prp8 to many strongly correlated communities, and (ii) outlining the key role of Prp8's RNAse-H, along with Clf1 and Cwc2, in conveying signals towards the functional RNAse-H domain. The reported findings provide fundamental advances in dissecting pivotal mechanistic aspects of this amazing gene maturation machinery from an atomic-level perspective. Conceivably, the observed information exchange routes may be exploited to devise small-molecules modulator able to hinder the functional SPL's dynamics by interfering/blocking this signaling path. Given the mounting evidence that splicing defects are an increasingly appreciated hallmark of tumorigenesis, our outcomes may be harnessed to intervene against distinct cancer types deriving from splicing alterations.

6.6 Future Perspectives

Atomic-level computer simulations, although limited by the always increasing size of the system under study, have advanced our understanding of the splicing mechanism and have provided the conceptual foundation for the development of novel splicing modulators. Nevertheless, longer and more precise atomic-level investigations of splicing will be facilitated by new technical improvements in computer hardware for simulations and new methods for modeling and analytic tools.

The computational approach proposed in this chapter of the thesis proven to be a valid way of analyzing the plastic reshaping and structural rearrangements happening within the Spliceosome through the different stages of the splicing cycle. In the near future, a broader application of this study to the rest of the available experimental structures, along the splicing cycle, would yield useful insights for the structural and dynamical understanding of this highly complex process. Moreover, coarse grained (CG) models, where multiple atoms are described as single beads, may add another layer of details to the exploration of the phase space by the system, allowing simulations for a longer time scale, and bridging the gap between experiments and all-atom simulations. Finally, a thorough investigation of cryptic and potentially druggable binding pockets would help to identify small-molecule modulators to target cancer cells, which have a very precarious balance of correctly spliced protein isoforms, and to tune the expression levels of alternatively spliced mRNA.

7 Conclusions

Pre-mRNA splicing is a crucial step in gene expression that involves joining the coding regions of the main RNA transcript – the exons – and removing the silent sections – the introns. Splicing in eukaryotes is mediated by the spliceosome, a complex ribonucleoprotein machinery that facilitates two consecutive transesterification steps to produce a mature mRNA filament. Crystallographic and cryo-EM structures have recently characterized many splicesome complexes along the catalytic cycle, providing a unique opportunity for a deeper level of understanding through computer simulations.

In this thesis, we have studied the different steps of the splicing cycle from a computational perspective to advance our understanding of the sophisticated mechanism underlying splicing regulation.

By using virtual screening methodologies and classical MD simulations we first investigated strategies for the inhibition of Prp4 kinase (Chapter 4), a promising target for breast cancer treatment that has been identified as one of the essential genes for Triple Negative Breast cancer cell migration and splicing. This kinase was recently crystallized yielding the first basis for computational studies. We used Structure Based and Ligand Based Virtual Screening to screen large databases of molecules as potential inhibitors of the kinase ATP binding pocket, leading to a set of compounds that were subsequently tested through Surface Plasmon Resonance. This study resulted into a set of potential inhibitors that binds to the kinase with low Kd. However cellular tests are needed to understand the cytotoxicity of the compounds and the activity of these in the more complex environment of the cell.

We also focused on the early steps performed by the spliceosome machinery in the recognition of those sequences that signs the border of the pre-mRNA introns and that must be cut off. One of these sequences, the Py-tract, is a poly pyrimidine tract that, depending on its composition of uracils/cytosines, increases or decreases the splicing of nearby segments. Here we investigated the recognition mechanism of the splicing factor U2AF of such Py-tract, employing enhanced sampling simulations to assess the molecular mechanism

responsible for the affinity toward the RNA tract. As a first result, we observed that the RRM1, one of the two U2AF domains known to be less discriminant between the poly-U vs poly-C sequences, is instead the domain which more strongly contributes to pre-mRNA binding.

Using metadynamics simulations to study the dissociation of the poly-U strand, we identified highly flexible and positively charged Lysines residues as early recruitments fragments toward the pre-mRNA strand, that are used to grasp and catch phosphate groups of RNA backbone, while π-stacking is used to stabilize the bases into position for hydrogen bonding with the linker. Moreover, Arginines residues appear to establish the strongest and diverse interactions which, recognizing and binding both poly-U and poly-C Py-tracts, might be the responsible of the increased promiscuity of the RRM1 domain. Additionally, a detailed analysis of the binding free energy pinpoints critical π-stacking interactions engaged by the residues of both RRMs with the RNA bases, demonstrating this type of interaction to stabilize the most pre-mRNA binding.

Complementarily, we also investigated the impact of cancer-associated U2AF2 mutations and identified intriguing commonalities. Indeed, all mutations located on the RRM1 domain stabilize the pre-mRNA binding, while those placed on RRM2 destabilize it. All mutations influenced pre-mRNA binding free energy by affecting three cardinal points of the U2AF2/polyU interactions which are Asp231, Glu201 and Arg150 and, partially Gln333 and Leu331 on linker-loop, RRM1 and RRM2, respectively, regardless of where the mutations were placed, suggesting that a communication between the different U2AF2 domains must take place. As such, we sought to investigating the communication pathways between the mutations and these cardinal points, surprisingly resulting in less effective (longer) signaling path in most of the mutants, perturbing as a result the interplay of RRM1/RRM2 towards pre-mRNA recognition and binding. Interestingly, we also dissected the cardinal residues involved in the communication between the two recognition domains by computing their degeneracy among the computed paths. This analysis remarked the pivotal role of Thr252 placed at the crossroad of the two domains as a central and conserved node of the interdomain signaling. Overall, the molecular dynamics simulations reported in this chapter have helped to clarify at atomic-level key functional aspects of U2AF2, providing important

information for further experimental studies. However, this chapter has just began to scratch the surface of this mechanism, as more extensive simulations are required to assess the binding of the RNA Recognition Motifs toward different composition of Py-tracts.

The final part of this thesis (Chapter 6) has taken advantage by the increasing number of spliceosome structures in the last years. A near-atomistic cryo-EM reconstruction of the *Saccharomices C.* spliceosome offered the opportunity to study the C complex structure, the spliceosome captured immediately after the first branch reaction, via molecular dynamics simulations and to understand it dynamical and signaling behavior. Thanks to the recent advance of computational power, we were able to extensively simulate this structure by performing multiple microsecond time scale MD simulations to unveil the crosstalk occurring among the spliceosome's protein components and its essential dynamics. Statistical analysis of correlations and the Principal Component Analysis showed a concerted action of Prp8's RNAse-H domain and Clf1 protein in the displacement of the branch helix formed by U2 snRNA and the intron lariat, possibly triggering the beginning of its conformational readjustment towards the position occupied in the subsequent SPL (C*) complex along the cycle.

In this last chapter, we suggested that Prp8 is a key component in remodeling the C complex, using its β-finger motif on the RNAse-H domain. Finally, because of the large distance separating Clf1 and the RNAse-H domain, we investigated the signaling routes and communication channels underlying these functional movements of Clf1 and the RNAse-H. By our network analysis of MD simulations, we discerned the connecting the nodes mostly involved in the communication among different communities, identifying two most relevant path and outlining the key role of Prp8, Clf1 and Cwc2 in conveying signals towards the functional RNAse-H domain.

Overall, this study represents the first attempt to provide atomistic details on spliceosome inner signaling network and dynamics, providing hints for small molecule modulators that could exploit the signaling path. This same computational approach could be also applied to other cryo-EM structures to investigate, with classical simulations, different spliceosome complexes along the splicing cycle, complementing the mechanistic details provided by this study. The spliceosome structural biology revolution is just at the beginning and a better

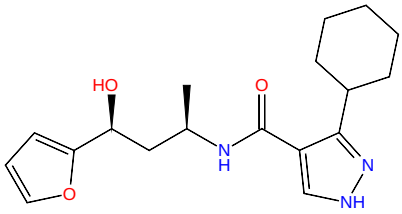
characterization of the human spliceosome has already populated the Protein Data Bank with H*omo Sapiens*' spliceosome structures, yielding the foundation for additional studies of the human splicing.

Taken together the results reported in this thesis have provided examples of how molecular simulations, thanks to recent development in computational methods, and to the ever increasing computational power available, can give important contribution and useful insights to characterize the mechanism of highly complex biological systems.

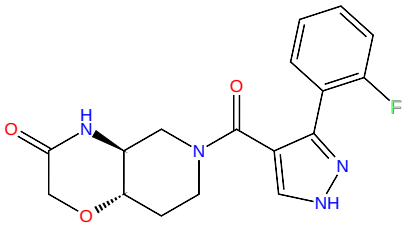
Appendices

Appendix A

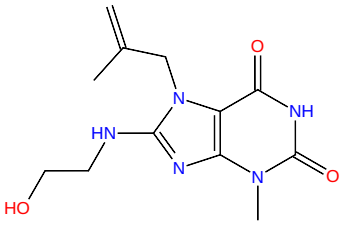
A.1 Additional Figures



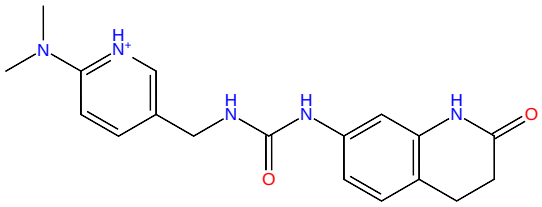
idnumber: Z1726990952



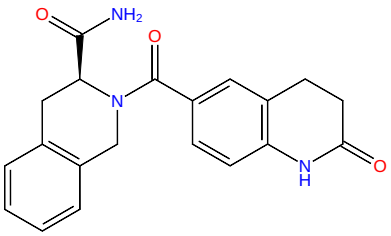
idnumber: Z1973215093



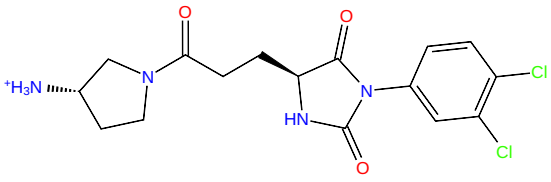
idnumber: Z238920900



idnumber: Z1262386126



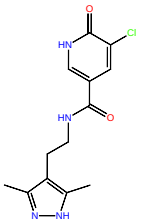
idnumber: Z168475560



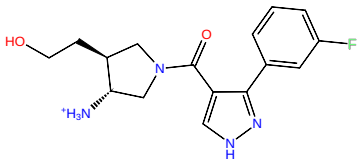
idnumber: Z2242907282

Fig. A.1 2D structures of the molecules resulting from the first virtual screening using the Enamine database.

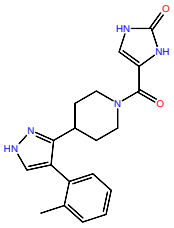
4 Appendix A



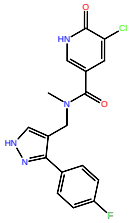
ID: 45346544



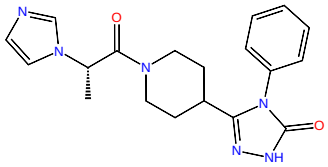
ID: 61703018



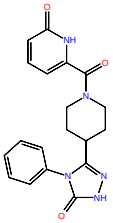
ID: 18835258



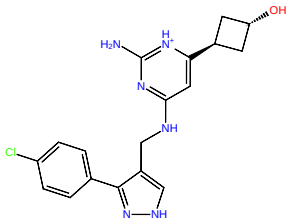
ID: 11475121



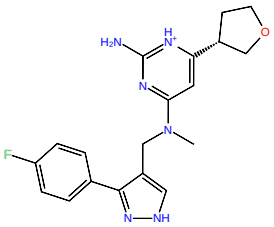
ID: 86075168



ID: 40508753



ID: 79841352



ID: 67068575

Fig. A.2 2D structures of the molecules resulting from the first virtual screening using the Chembridge database.

5 Appendix A

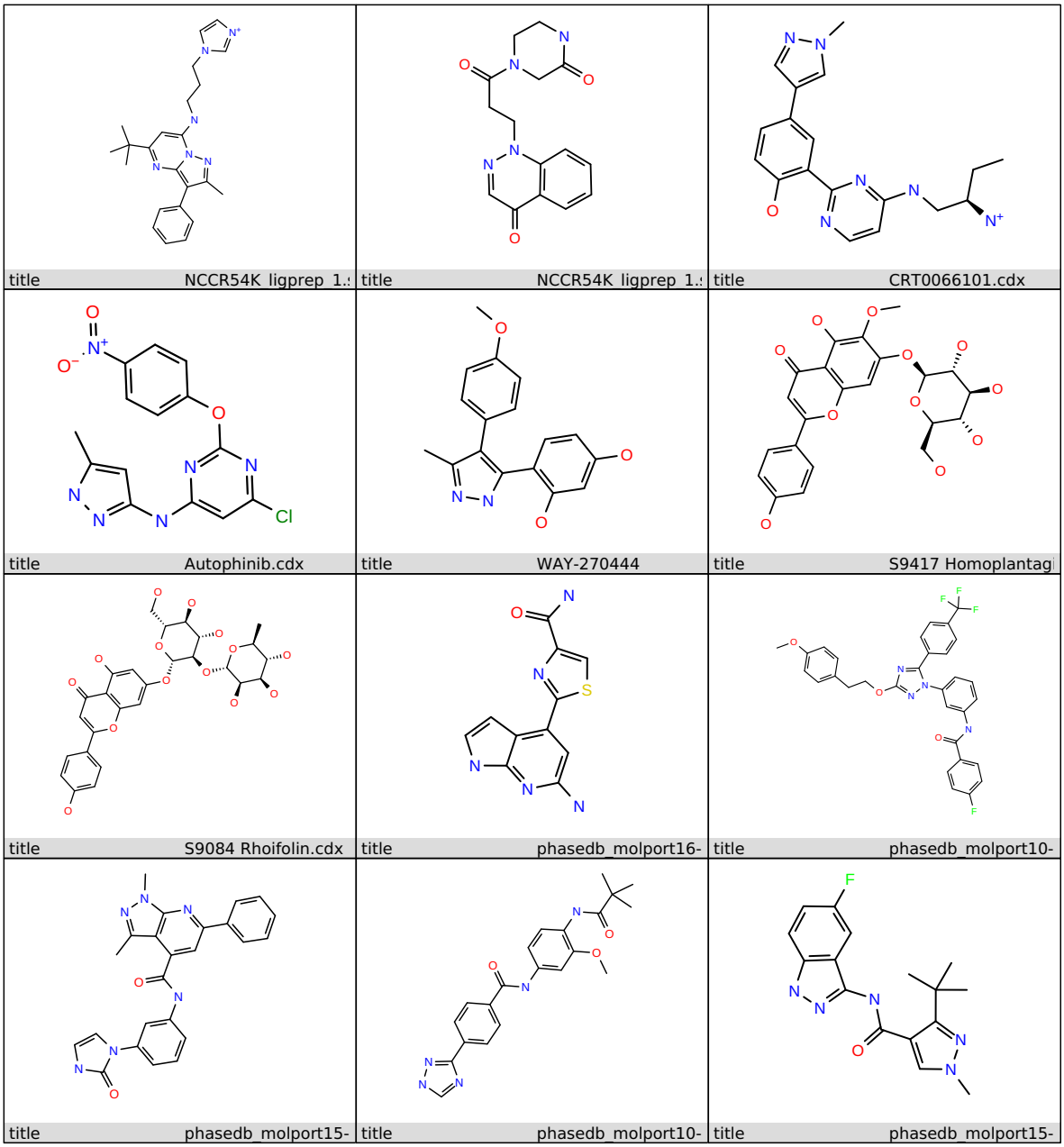


Fig. A.3 2D structures of the molecules resulting from the ligand-based virtual screening on the Rebastinib and CompoundA templates and from the structure-based virtual screening considering 2 water molecules.

6 Appendix A

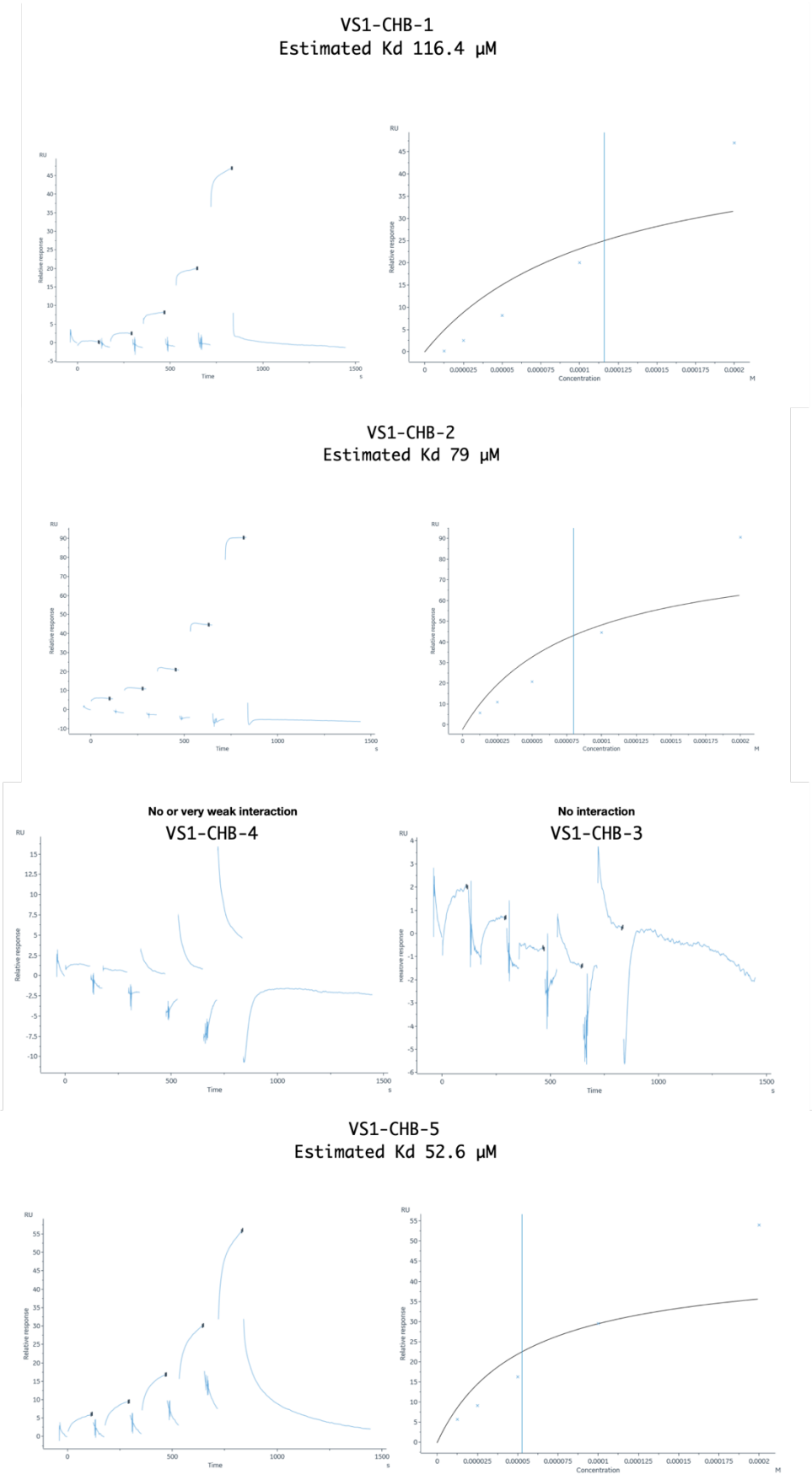


Fig. A.4 Single-Cycle SPR experiment results for estimated Kd using 5 analyte concentrations of the first set of screened molecules.

7 Appendix A

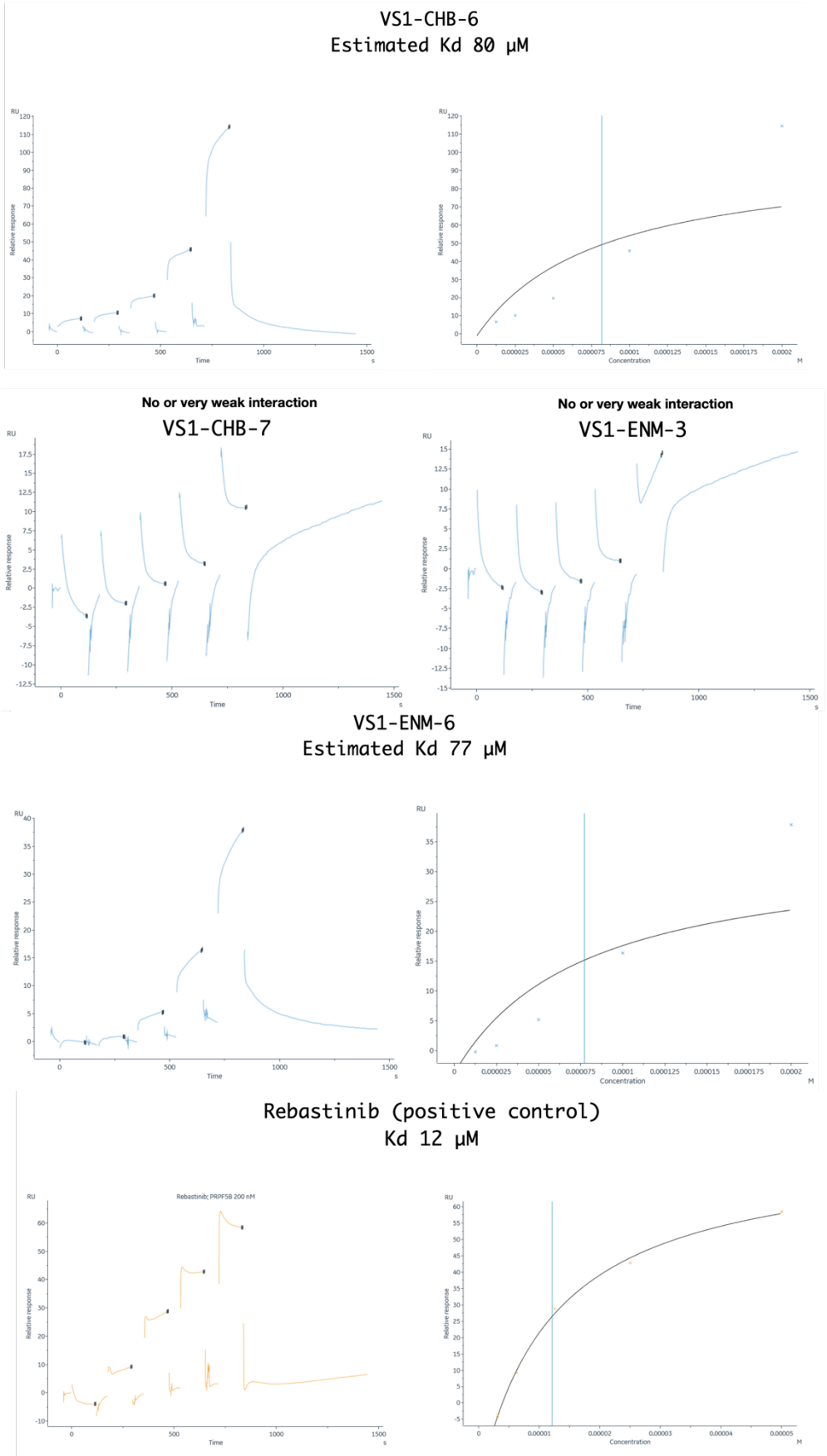


Fig. A.5 ***Single-Cycle SPR experiment results for estimated Kd using 5 analyte concentrations of the first set of screened molecules.***

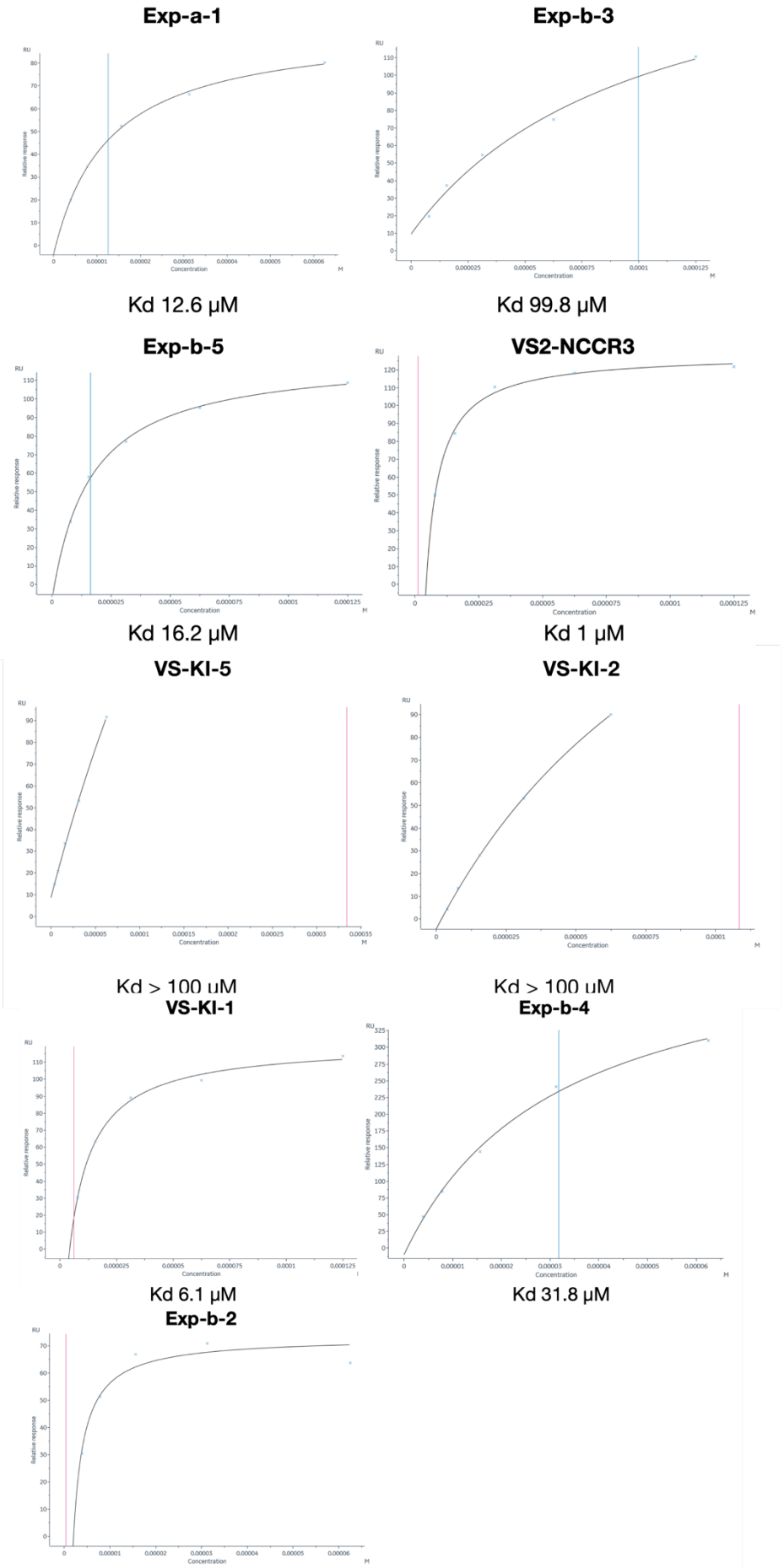


Fig. A.6 Single-Cycle SPR experiment results for estimated Kd using 5

Appendix B

B.1 Additional Figures

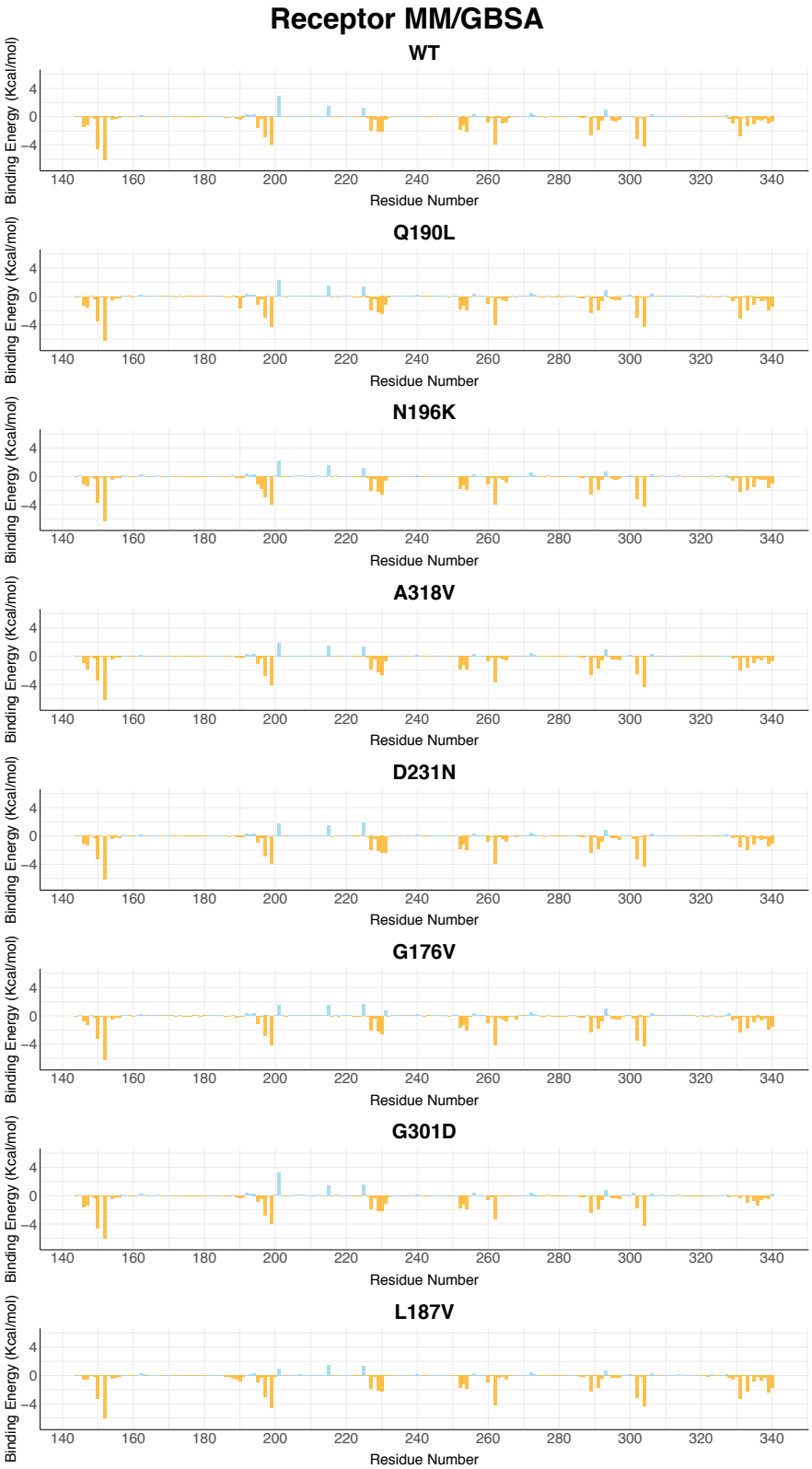


Fig. B.1 Per-residue MM-GBSA binding free energy (ΔGb, kcal/mol) of U2AF on WT and each mutant system.

11 Appendix B

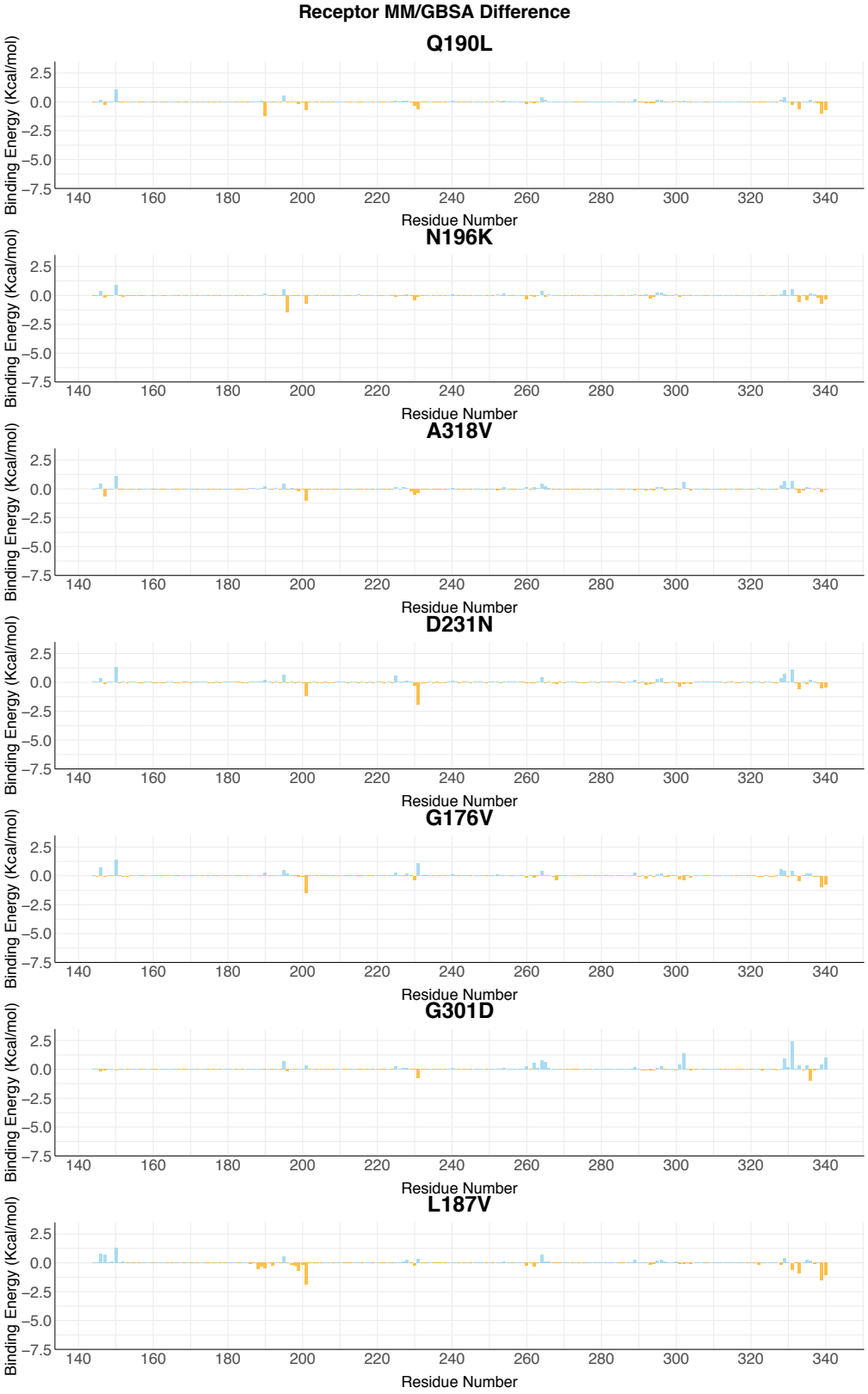


Fig. B.2 Per-residue MM-GBSA binding free energy difference (ΔΔGb, kcal/mol) between WT and each mutant system of U2AF.

13 Appendix B

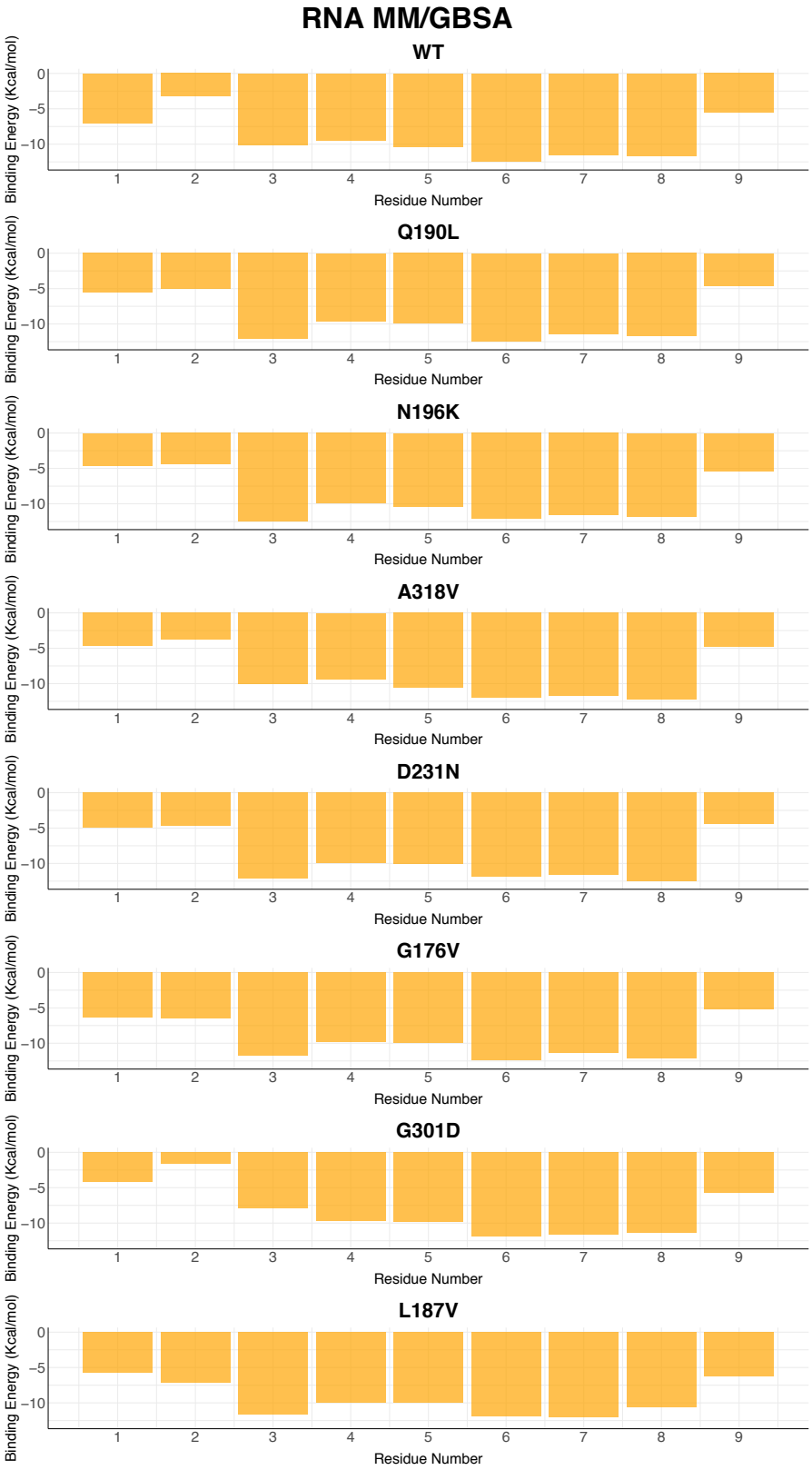


Fig. B.3 *Per-residue MM-GBSA binding free energy (ΔGb, kcal/mol) of RNA on WT and each mutant system.*

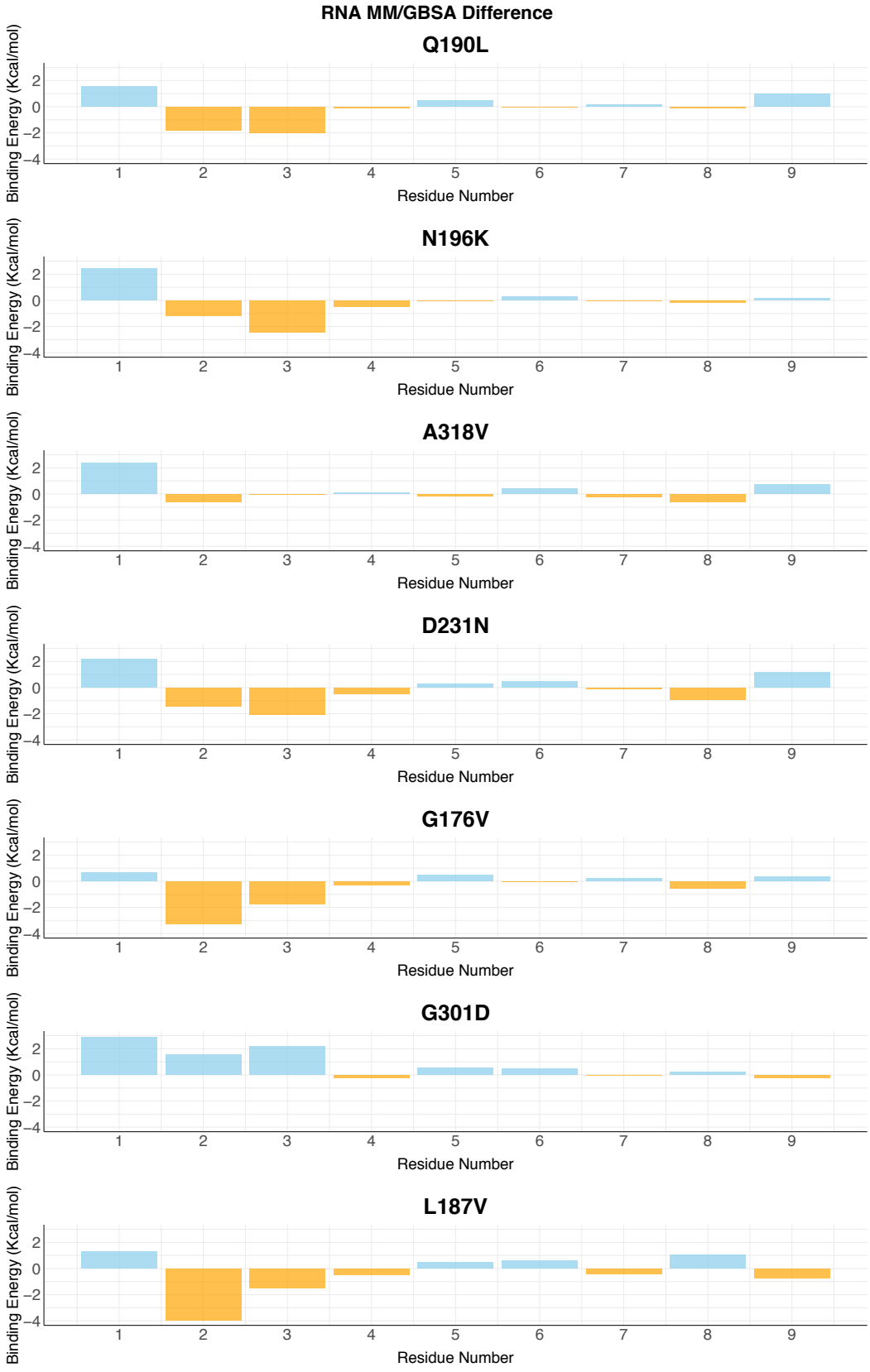

RNA MM/GBSA Difference
Q190L
N196K
A318V
D231N
G176V
G301D
L187V
Binding Energy (Kcal/mol)
2
0
−2
−4
1
2
3
4
5
6
7
8
9
Residue Number

Fig. B.4 Per-residue MM-GBSA binding free energy difference (ΔΔGb, kcal/mol) between WT and each mutant system of RNA.

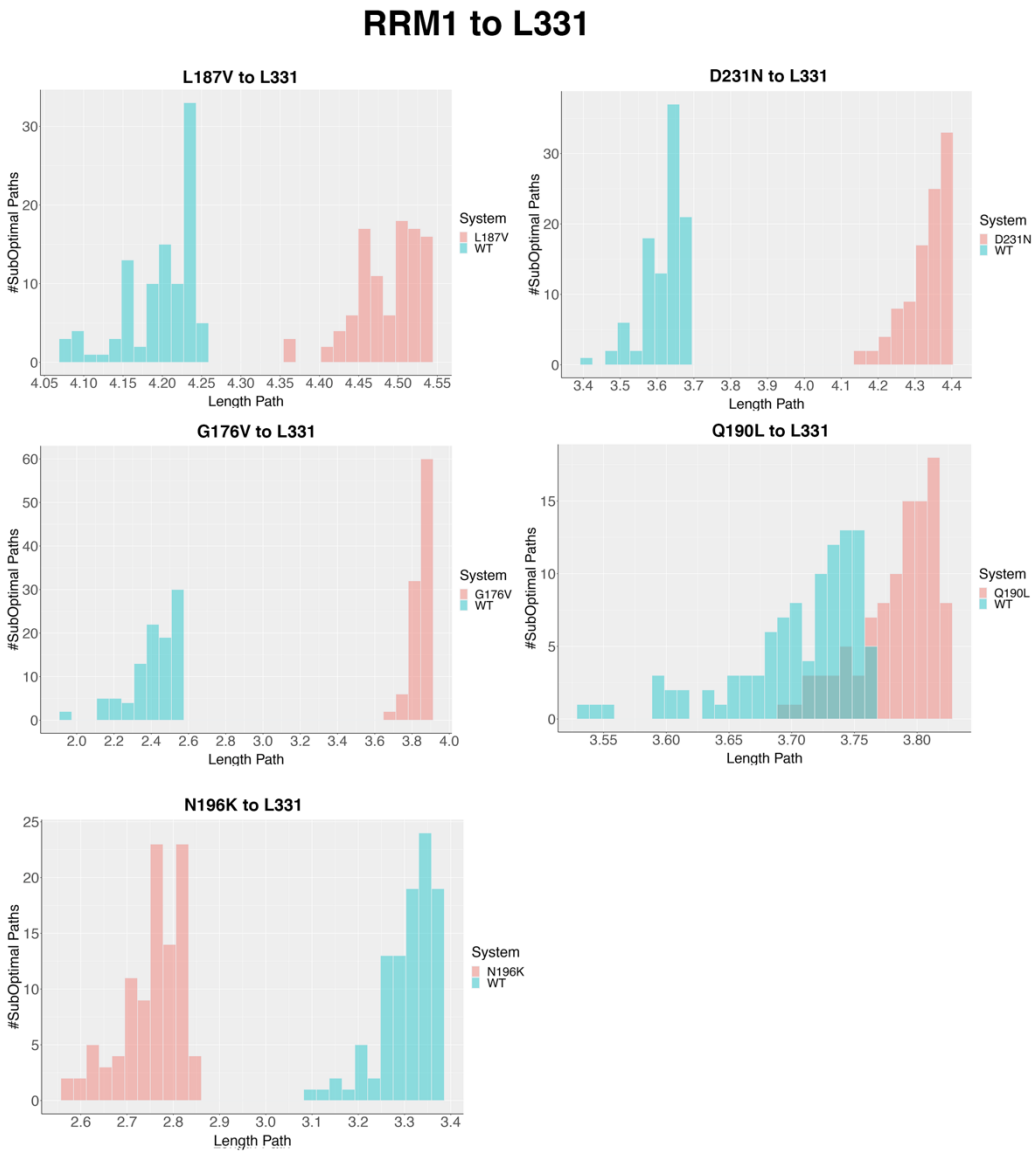


Fig. B.5 Distribution of interdomain signaling-path lengths heading from the mutations site on RRM1 toward Leu331@RRM2 compared to the WT distribution. In pink the distribution of paths within the mutant system is shown. In cyan the WT distribution is displayed.

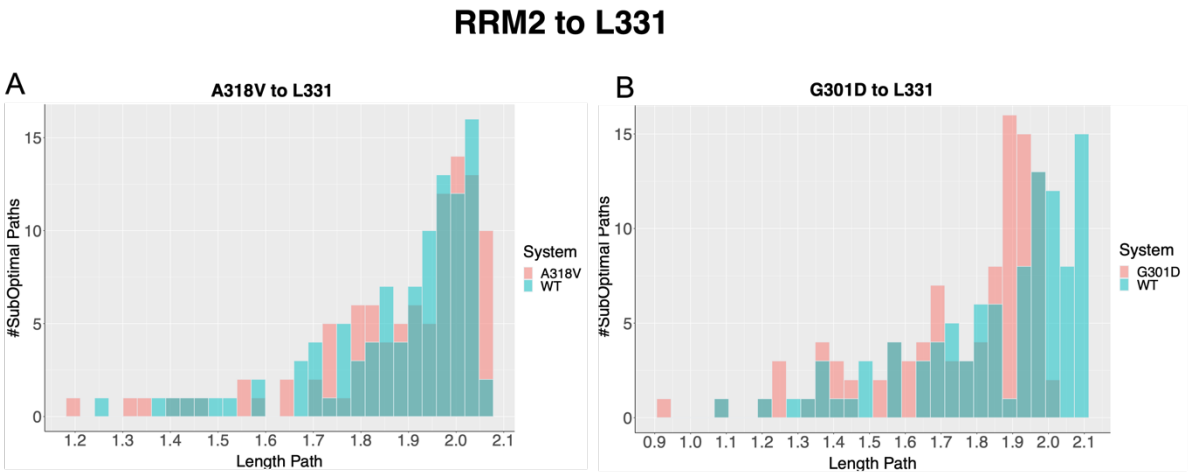


Fig. B.6 Distribution of intradomain signaling-path lengths heading from the mutations site on RRM2 toward Leu331@RRM1 compared to the WT distribution. In pink the distribution of paths within the mutant system is shown. In cyan the WT distribution is displayed.

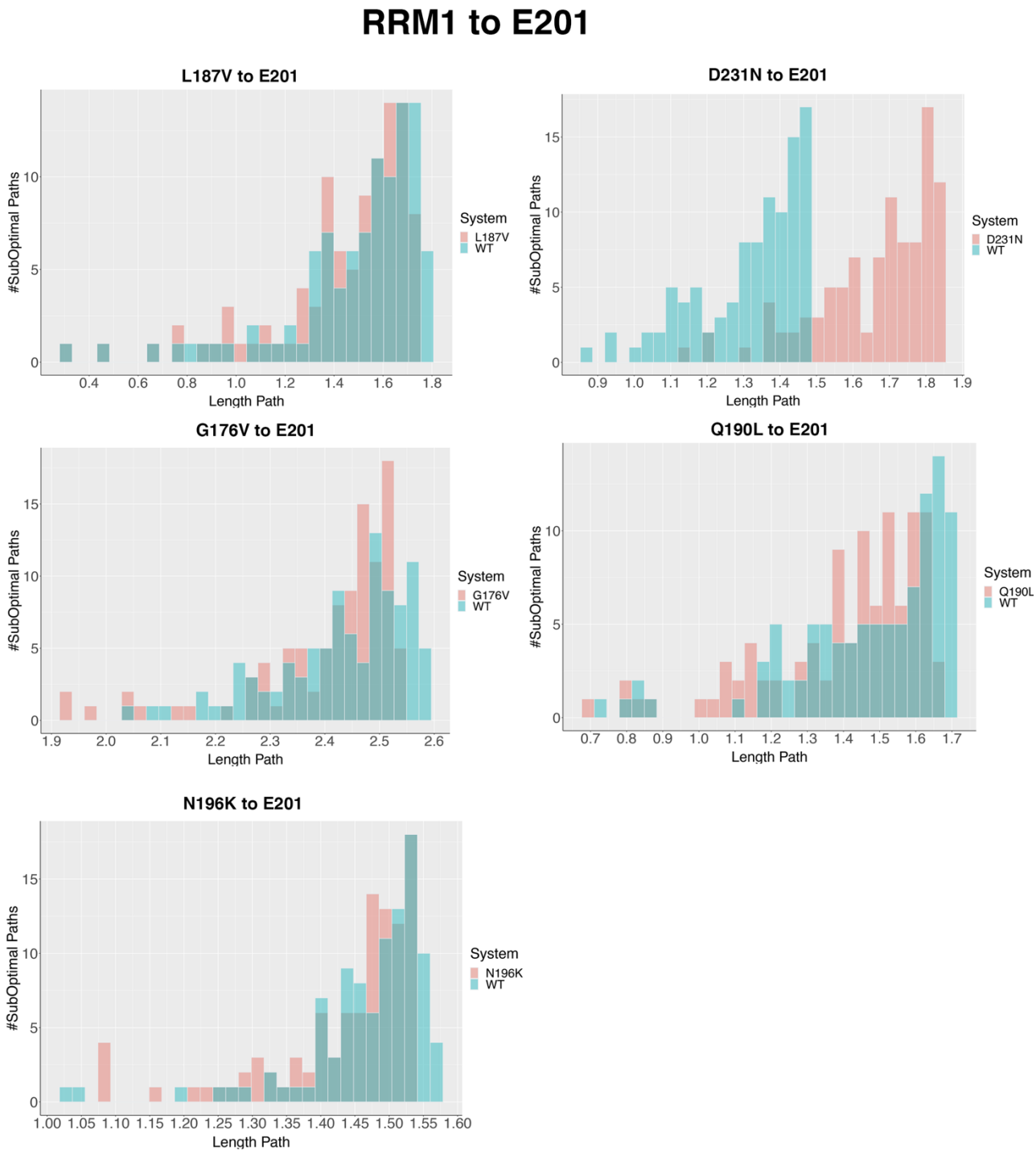


Fig. B.7 Distribution of intradomain signaling-path lengths heading from the mutations site on RRM1 toward Glu201@RRM1 compared to the WT distribution. In pink the distribution of paths within the mutant system is shown. In cyan the WT distribution is displayed.

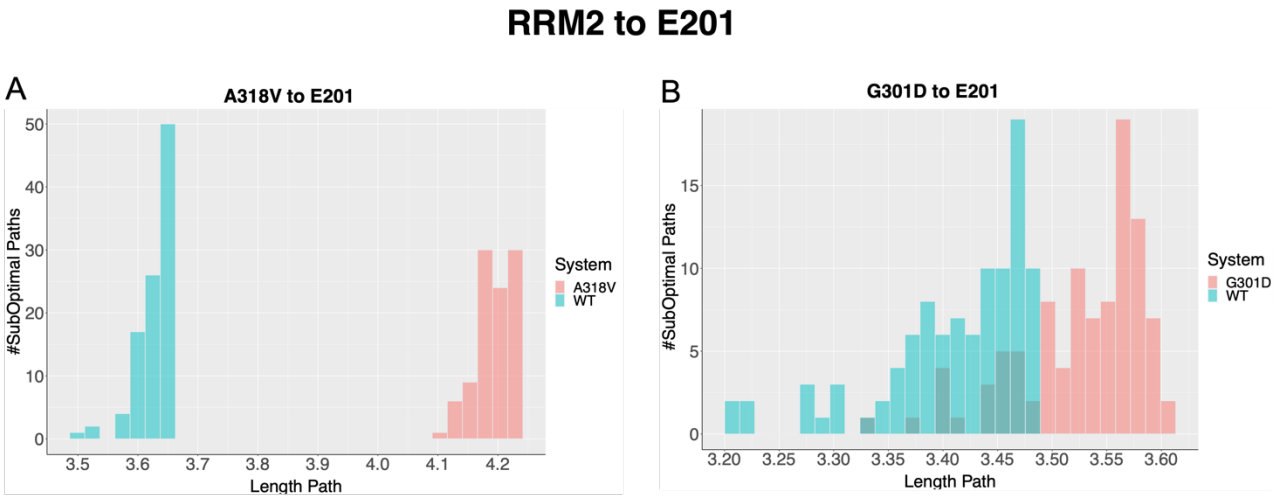


Fig. B.8 Distribution of interdomain signaling-path lengths heading from the mutations site on RRM2 toward Glu201@RRM1 compared to the WT distribution. In pink the distribution of paths within the mutant system is shown. In cyan the WT distribution is displayed.

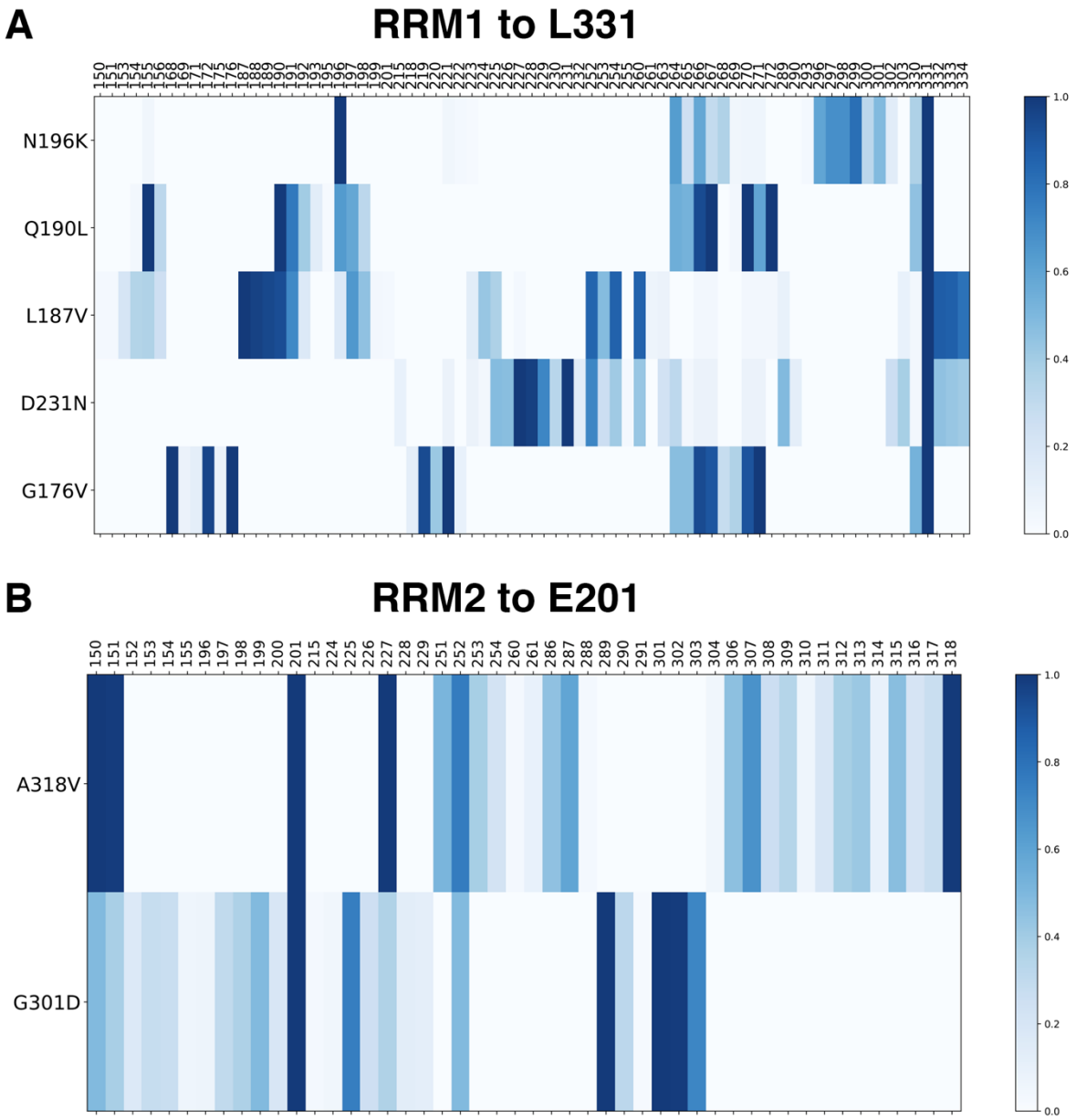


Fig. B.9 Matrices of common degenerated residues for interdomain pathways from (A) all mutations on RRM1 toward residue Leu331 on RRM2, and (B) from all mutations on RRM2 to Glu201 on RRM1, ranging from 0 to 1.

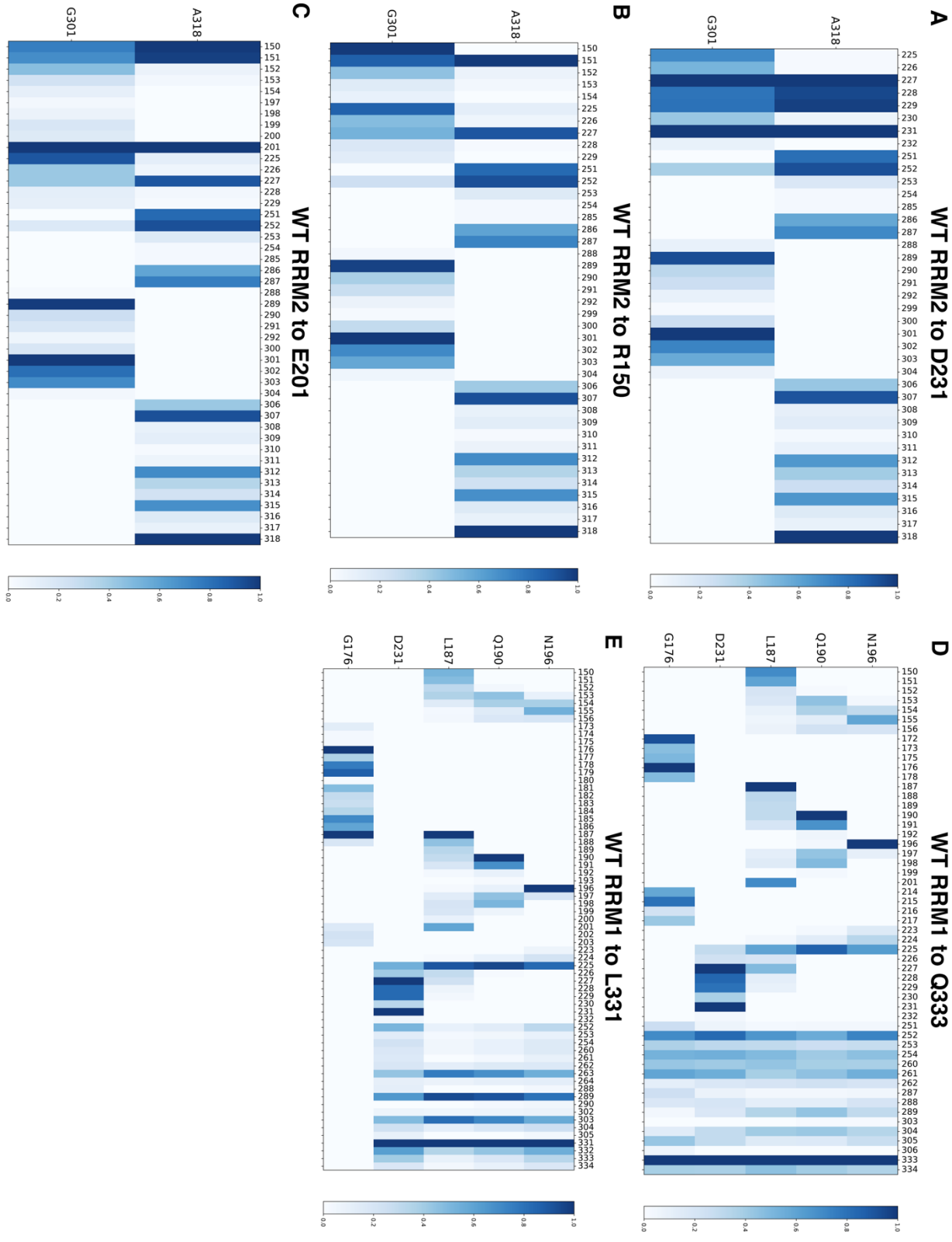

A
WT RRM2 to D231
G301
A318
B
WT RRM2 to R150
G301
A318
C
WT RRM2 to E201
G301
A318
D
WT RRM1 to Q333
G176
D231
L187
Q190
N196
E
WT RRM1 to L331
G176
D231
L187
Q190
N196

Fig. B.10 Matrices of common degenerated residues for interdomain pathways in WT from (A, B, C) all mutations on RRM2 toward residues Arg150, Asp231, Glu201 on RRM1, (D, E) and from all mutations on RRM1 to Gln333 and Leu331 on RRM2, ranging from 0 to 1.

B.2 Additional Tables

L187V		% Half-Life	
Protein	**RNA**	**L187V**	**WT**
Arg_150	U_8	107%	107%
Tyr_152	U_6	95%	97%
Asp_231	U_8	64%	57%
Lys_195	U_7	62%	48%
Gln_333	U_3	84%	53%
Lys_329	U_1	73%	65%
Gln_147	U_8	39%	58%
Ala_335	U_3	55%	45%
Asn_289	U_4	81%	83%
Thr_252	U_5	87%	88%
Val_254	U_5	38%	42%
His_230	U_7	59%	63%

Tab. B.1 Persistence (%) of intermolecular hydrogen bonds in the mutated system and in the WT system. The residues that have higher hydrogen bond differences in the two systems are highlighted in red. Hydrogen bonds persistence might exceeds 100%, since they have been considered as the sum of all the atoms within the same residue or base that form hydrogen bonds.

Q190L		% Half -Life	
Protein	**RNA**	**Q190L**	**WT**
Arg_150	U_8	115%	107%
Lys_195	U_6	72%	73%
Lys_195	U_7	60%	47%
Gln_333	U_3	78%	53%
Val_254	U_5	40%	42%
Lys_329	U_1	32%	65%
Gln_147	U_8	58%	48%
His_230	U_7	49%	62%
Ala_335	U_3	66%	45%
Asp_231	U_8	58%	57%
Asn_289	U_4	81%	83%
Thr_252	U_4	86%	88%
Gln_147	U_8	58%	58%

Tab. B.2 Persistence (%) of intermolecular hydrogen bonds in the mutated system and in the WT system. The residues that have higher hydrogen bond differences in the two systems are highlighted in red. Hydrogen bonds persistence might exceeds 100%, since they have been considered as the sum of all the atoms within the same residue or base that form hydrogen bonds.

G176V		% Half -Life	
Protein	**RNA**	**G176V**	**WT**
Arg_150	U_8	110%	107%
Lys_195	U_6	69%	73%
Lys_195	U_7	61%	47%
Gln_333	U_3	71%	53%
Lys_329	U_1	41%	65%
Gln_147	U_8	52%	48%
Ala_335	U_3	56%	45%
Thr_252	U_5	86%	88%
Asn_289	U_4	80%	83%
Asn_196	U_6	32%	12%
Asp_231	U_8	55%	57%
Val_254	U_5	40%	42%
Tyr_152	U_6	97%	97%

Tab. B.3 Persistence (%) of intermolecular hydrogen bonds in the mutated system and in the WT system. The residues that have higher hydrogen bond differences in the two systems are highlighted in red. Hydrogen bonds persistence might exceeds 100%, since they have been considered as the sum of all the atoms within the same residue or base that form hydrogen bonds.

A318V		% Half -life	
Protein	**RNA**	**A318V**	**WT**
Gln_147	U_8	62%	48%
Arg_150	U_8	110%	107%
Lys_329	U_1	14%	54%
Gln_333	U_3	34%	53%
Ala_335	U_3	27%	45%
Lys_195	U_6	66%	73%
Lys_195	U_7	68%	47%
Tyr_152	U_6	97%	97%
Thr_252	U_5	89%	88%
Asn_289	U_4	83%	83%
Asp_231	U_9	31%	0%
Val_111	U_5	40%	42%
Gln_333	U_3	34%	53%
Lys_225	U_5	78%	76%

Tab. B.4 Persistence (%) of intermolecular hydrogen bonds in the mutated system and in the WT system. The residues that have higher hydrogen bond differences in the two systems are highlighted in red. Hydrogen bonds persistence might exceeds 100%, since they have been considered as the sum of all the atoms within the same residue or base that form hydrogen bonds.

G301D		% Half-Life	
Protein	**RNA**	**G301D**	**WT**
Tyr_152	U_6	93%	97%
Thr_252	U_6	87%	88%
Asn_289	U_4	83%	83%
Arg_150	U_8	112%	107%
Asp_231	U_9	22%	0%
Lys_195	U_6	57%	73%
Lys_195	U_7	63%	47%
Gln_147	U_8	54%	48%
Val_254	U_4	42%	42%
Gln_333	U_3	48%	53%
His_230	U_7	55%	63%
Ala_335	U_3	26%	45%
Lys_225	U_5	67%	76%

Tab. B.5 Persistence (%) of intermolecular hydrogen bonds in the mutated system and in the WT system. The residues that have higher hydrogen bond differences in the two systems are highlighted in red. Hydrogen bonds persistence might exceeds 100%, since they have been considered as the sum of all the atoms within the same residue or base that form hydrogen bonds.

N196K		% Half-Life	
Protein	**RNA**	**N196K**	**WT**
Tyr_152	U_6	94%	97%
Thr_252	U_5	88%	88%
Gln_333	U_3	83%	53%
Asn_289	U_4	79%	83%
Arg_150	U_8	111%	107%
Lys_195	U_6	66%	73%
Lys_195	U_7	64%	47%
Ala_335	U_3	61%	45%
Asp_231	U_8	59%	57%
Gln_147	U_8	59%	48%
Val_254	U_5	38%	42%
His_230	U_7	52%	63%
Lys_329	U_1	30%	65%
Lys_225	U_5	77%	76%

Tab. B.6 Persistence (%) of intermolecular hydrogen bonds in the mutated system and in the WT system. The residues that have higher hydrogen bond differences in the two systems are highlighted in red. Hydrogen bonds persistence might exceeds 100%, since they have been considered as the sum of all the atoms within the same residue or base that form hydrogen bonds.

D231N		% Half-Life	
Protein	**RNA**	**D231N**	**WT**
Tyr_152	U_6	97%	97%
Thr_252	U_4	87%	88%
Asn_289	U_4	80%	83%
Gln_333	U_3	80%	53%
Arg_150	U_8	117%	107%
Lys_195	U_6	65%	73%
Lys_195	U_7	62%	47%
Ala_335	U_3	65%	45%
Gln_147	U_8	61%	58%
Asp/Asn_231	U_8	80%	57%
His_230	U_7	56%	63%
Val_254	U_5	34%	42%
Lys_225	U_5	76%	76%
Lys_260	U_4	69%	70%

Tab. B.7 Persistence (%) of intermolecular hydrogen bonds in the mutated system and in the WT system. The residues that have higher hydrogen bond differences in the two systems are highlighted in red. Hydrogen bonds persistence might exceeds 100%, since they have been considered as the sum of all the atoms within the same residue or base that form hydrogen bonds.

Appendix C

C.1 Additional Figures

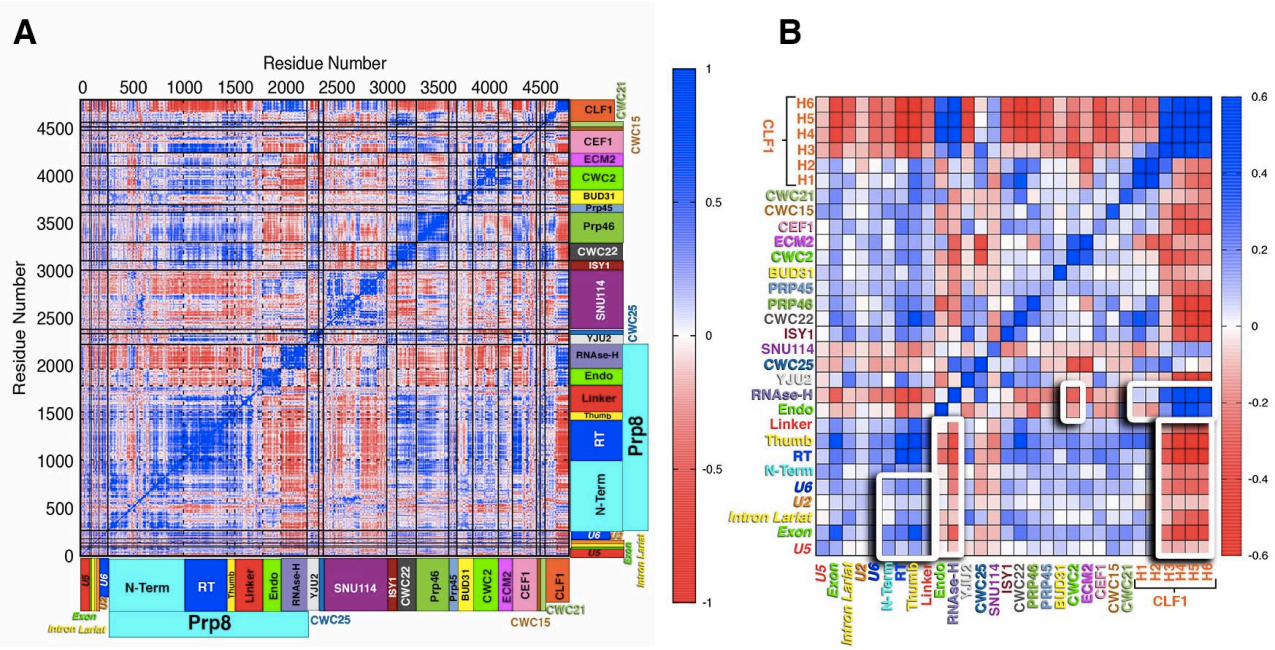


Fig. C.1 (A) Cross-correlation matrix based on per-residue Pearson's correlation coefficients (CCs) as derived from the mass-weighted covariance matrix calculated over 3-replicas of classical molecular dynamics trajectories. CCs values range from -1 (red, anti-correlated motions) to +1 (blue, correlated motions). (B) Coarse matrix summing the correlation scores and normalizing over the products of the residues of the considered SPL proteins/domains. Pairwise correlation scores (CSs) are reported in the range from -0.6 to 0.6 for clarity reasons. In green are encircled regions relevant to explain the functional movements captured by the principal component analysis. Protein names and their domains are labeled with the same color code of Figure 1.

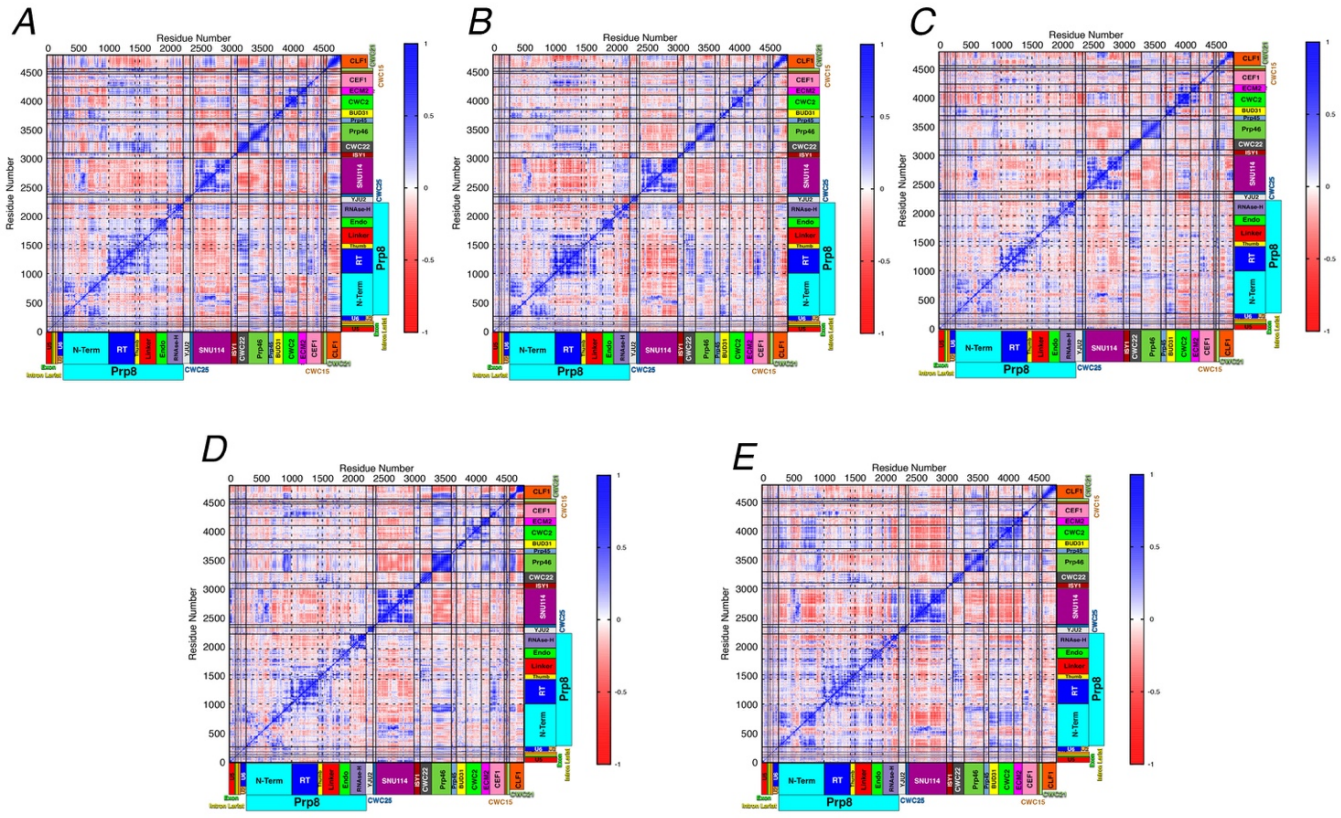


Fig. C.2 Per-residue Pearson's cross-correlation coefficients (CCs) derived from the mass-weighted covariance matrix calculated over the last 700 ns of MD trajectories for the 5 replicas of the C model in (A)-(E), respectively. CCs values range from -1 (read, anti-correlated motions) to +1 (blue, correlated motions). The protein names and their domains are reported on the bottom and on the left side, highlighted with boxes of different colors.

C-3 **Appendix C**

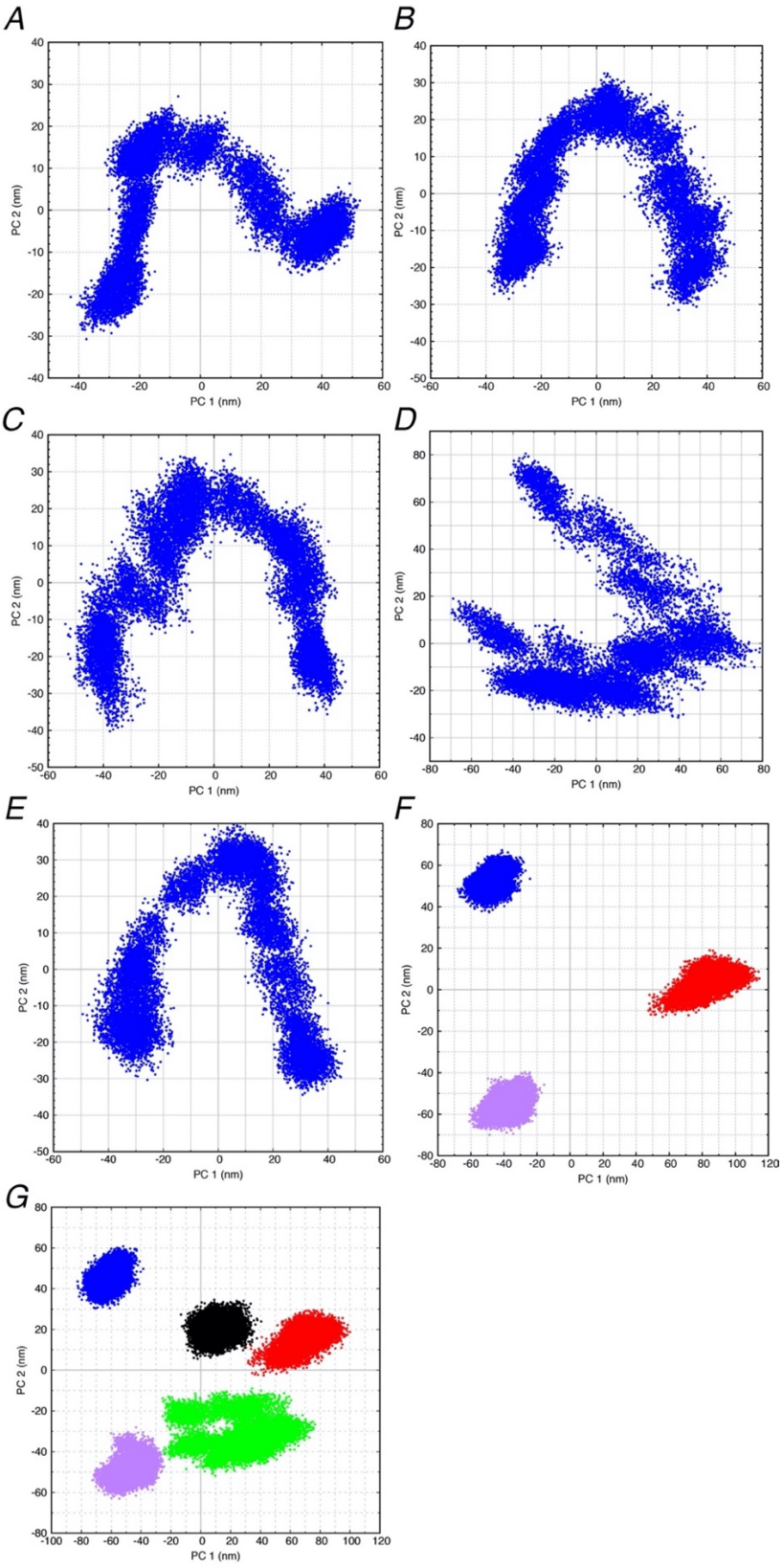


Fig. C.3 Scatter plot of Conformational subspace of PC1 vs PC2 for replica 1 to 5 in (A-E), respectively, merged replicas 1-3 (F), and merged replicas 1-5 (G).

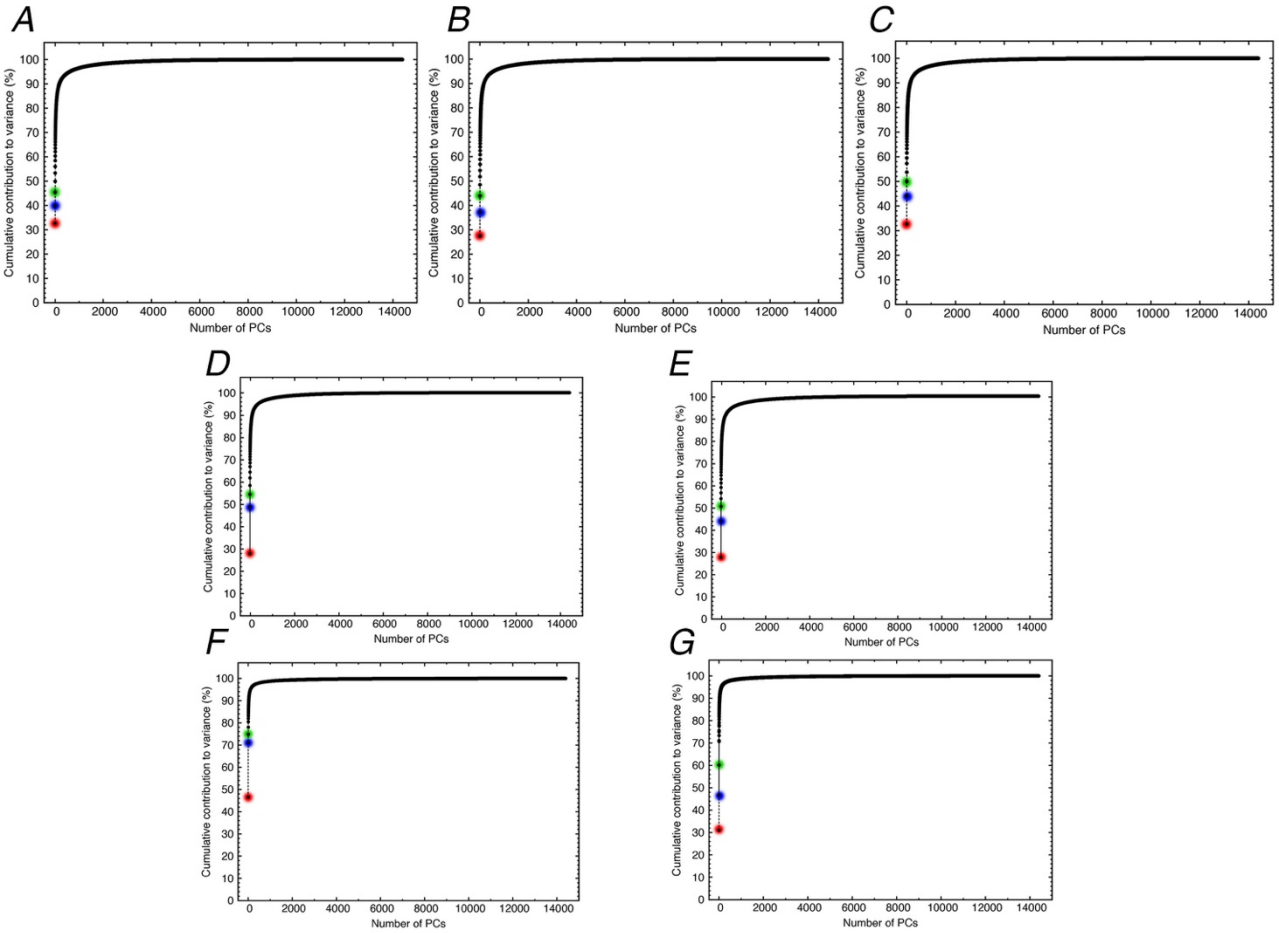


Fig. C.4 Principal components (PCs) cumulative contribution to variance for (A) replica 1, (B) replica 2, (C) replica 3 and (D) replica 4, (E) replica 5, (F) 3 replica combined trajectory, (G) 5 replicas combined trajectory. On y-axis is depicted Cumulative contribution of PCs (x-axis) to the variance of the overall motion calculated upon Principal Component Analysis. The contributions from the first three PCs are highlighted in red, blue and green, respectively.

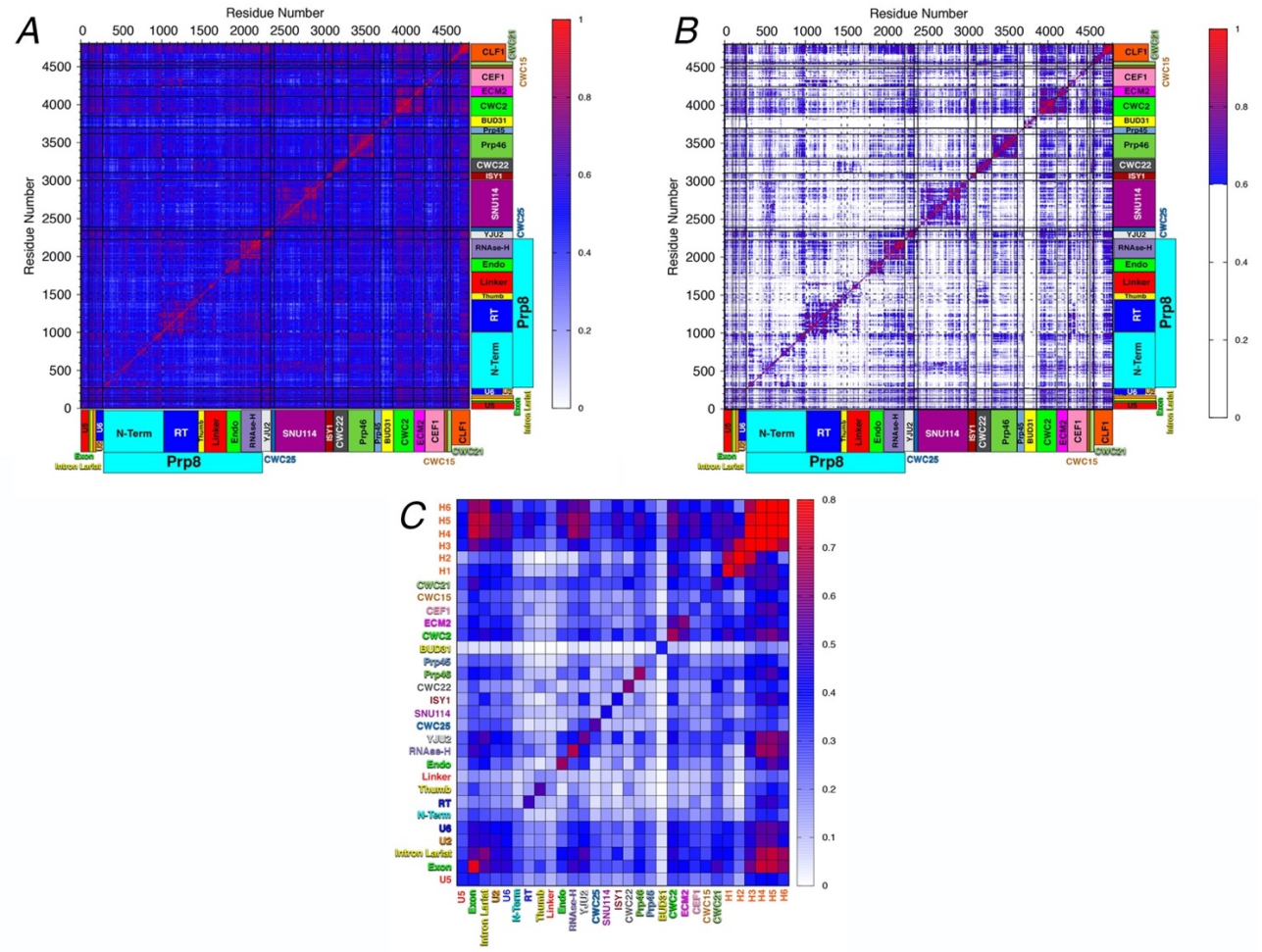


Fig. C.5 (A) Linear Mutual Information correlation coefficients (^{LMI}CCs) for the combined 3 replicas trajectory. These values range from 0 (white, no correlated motions) to +1 (red, correlated motions). (B) Linear Mutual Information matrix after filtering the correlation coefficients below 0.6. (C) Coarse grained matrix of pairwise correlation scores (^{LMI}CSs) given by summing ^{LMI}CCs of each pair of protein/domain and averaging by the corresponding number of residues, after filtering all values below 0.6. ^{LMI}CSs are reported in the range from 0 to 0.8 for clarity reasons. Protein names and their domains are labeled with the same color code of Figure 1 of the main text. Multiple strong correlations are clearly visible in this matrix confirming the pivotal role of Prp8 in establishing the intricate correlation network among its domains, which direct the SPL motion. As well, Clf1 shares many strong correlations, in particular, with the RNAse-H, Endo domains of Prp8 and Cwc2.

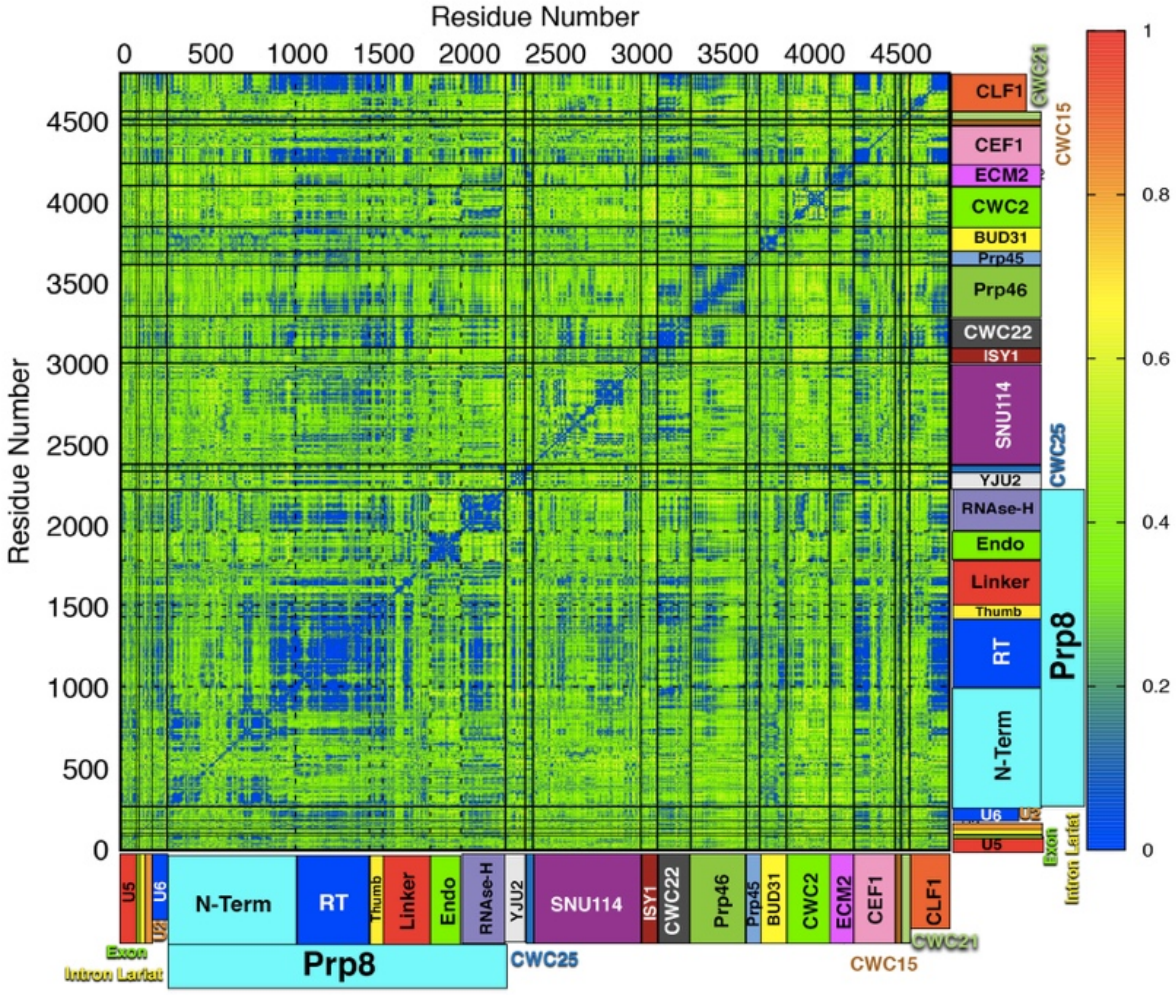


Fig. C.6 Root Mean Square Deviation matrix between [LMI]CCs and Pearson CCs for the combined 3 replicas trajectory. These values range from 0 (blue, low RMSD i.e., same values) to +1 (red, high RMSD i.e. completely different values). Protein names and their domains are labeled with the usual color code of Figure 1 of the Main Text.

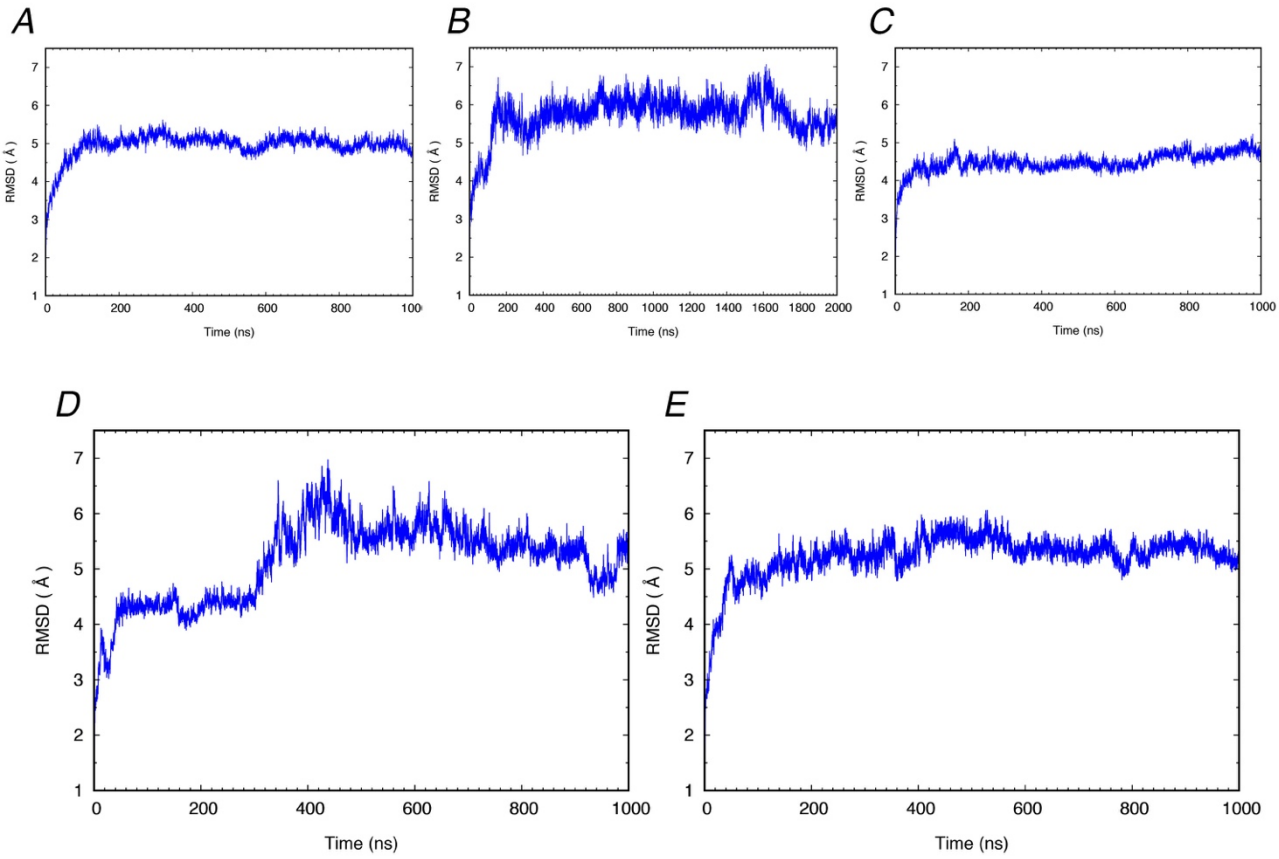


Fig. C.7 Root Mean Square Deviations (RMSD) vs. simulation time (ns) calculated on the production phase of molecular dynamics trajectories for the four 1-μs-long replicas of C model and for replica 2 prolonged up to 2-μs.

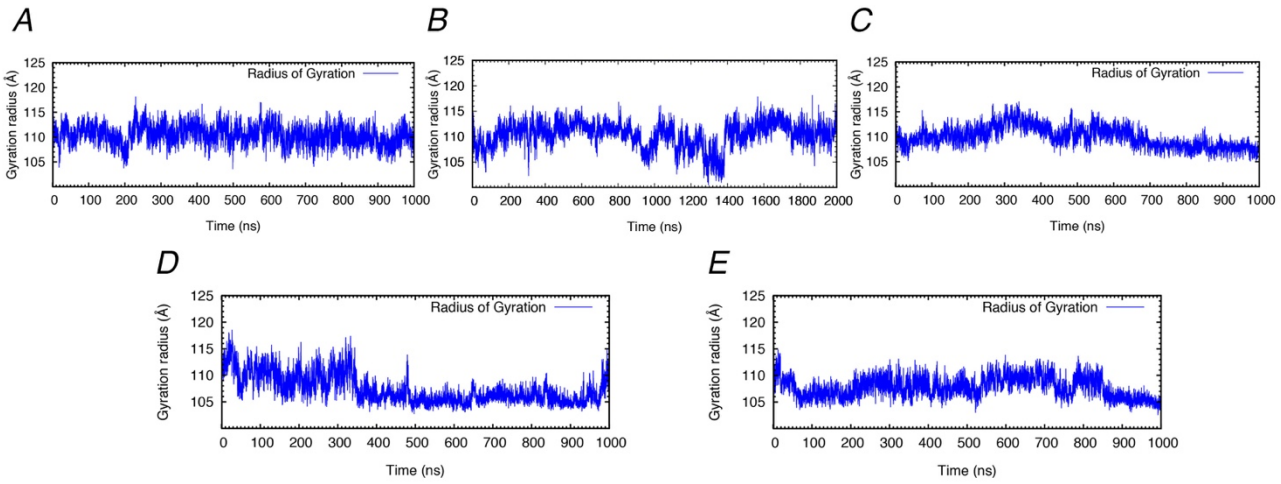


Fig. C.8 Radius of gyration vs. simulation time (ns) calculated on the production phase of molecular dynamics trajectories for the four 1-μs-long replicas of C model and of replica 2 prolonged up to 2-μs.

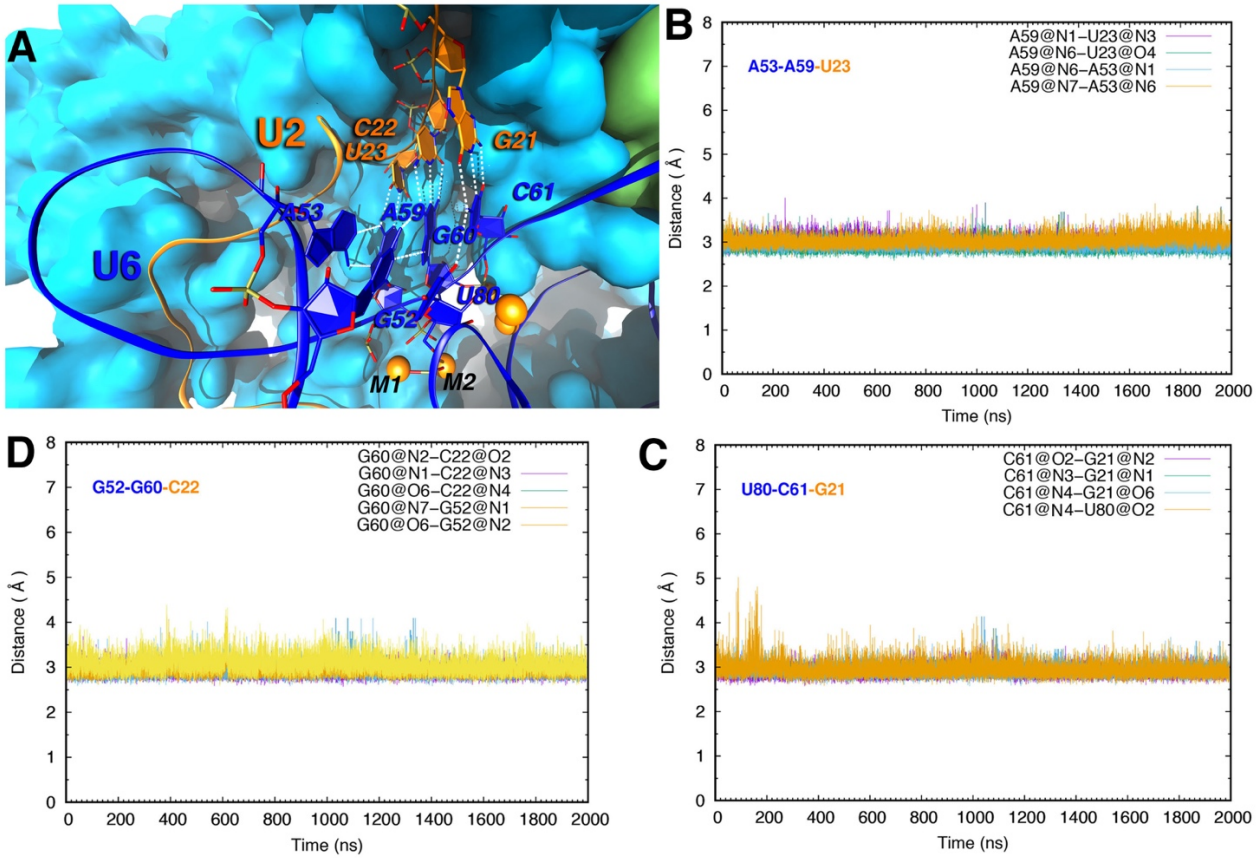


Fig. C.9 (A) Representative snapshot of the catalytic site grafting the triple-helix made by U2/U6 snRNAs. U2 and U6 snRNAs are represented as orange and blue tubes, respectively. Mg^{2+} ions are depicted as orange spheres with the catalytic Mg^{2+} ion labeled as M2. The nucleotides involved in the triple-helix are pictured in licorice. The Prp8 protein is shown in cyan surface. Hydrogen bonds (H-bonds) between RNA base-pairs are depicted as white dash lines. (B) Time evolution (ns) of the H-bonds distances (Å) between base-pairs of the nucleotides involved in the triple-helix in the longest replica, showing that the structural integrity is maintained during all simulation time. In all simulations, the triple helix architecture of the active site remained well preserved (), with the 5 Mg^{2+} ions engaging strong interactions with the phosphate groups of U6 snRNA and being nested within a positively charged pocket formed by Prp8.

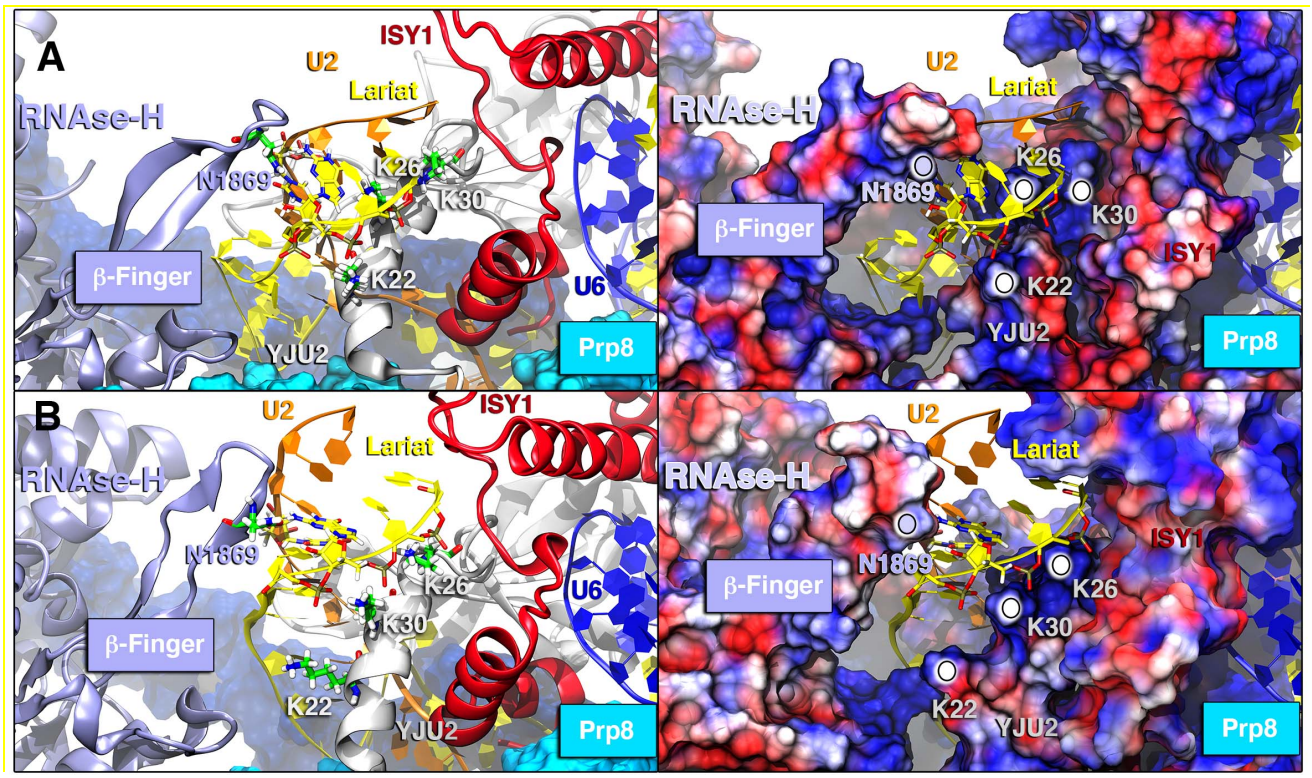


Fig. C.10 Structural rearrangement of the U2/intron lariat helix during PC1 (A and B), and electrostatic potential of representative frames extracted from the essential dynamics trajectory underlying the RNAse-H movement and the consequent IL/U2 helix wrapping. This is mediated by the Asn1869 of the ß-finger and by the Lys22, 26 and 30-forkzof Yju2.

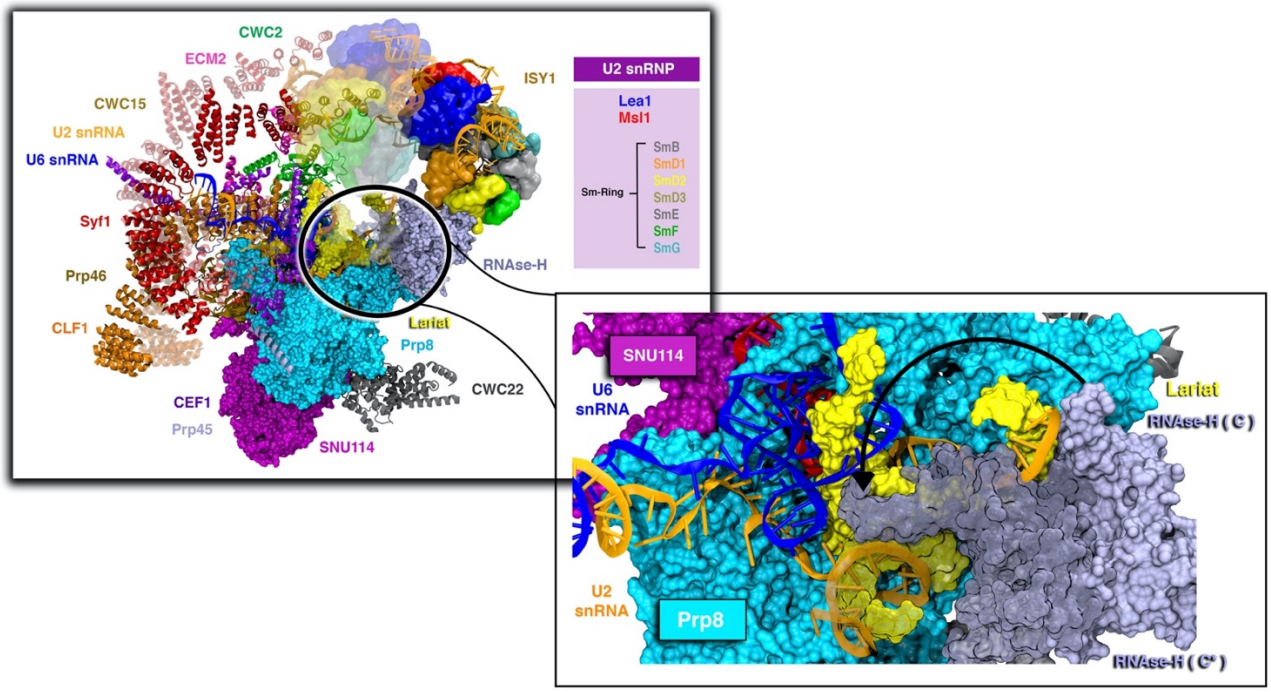


Fig. C.11 Experimental structures alignment of C and C complex focusing on the components affected by conformational changes from one intermediate to the other. In transparency are represented Syf1, Clf1 and U2snRNP of C* components. Darker objects are representing C proteins. CWC2 in green, Clf1 in orange (on the left), Syf1 in red and U2snRNP in plastic-surface, lariat in yellow and RNAseH in iceblue (on the right). U2snRNP Sm-Ring gets away from the RNAse-H domain by a rotation of Clf1 and Syf1. The insect shows the C* conformation of lariat and RNAse-H represented in transparent darker Surf, where the ß-finger of the RNAseH domain embraces the intron/U2 helix, interacting with its minor groove.*

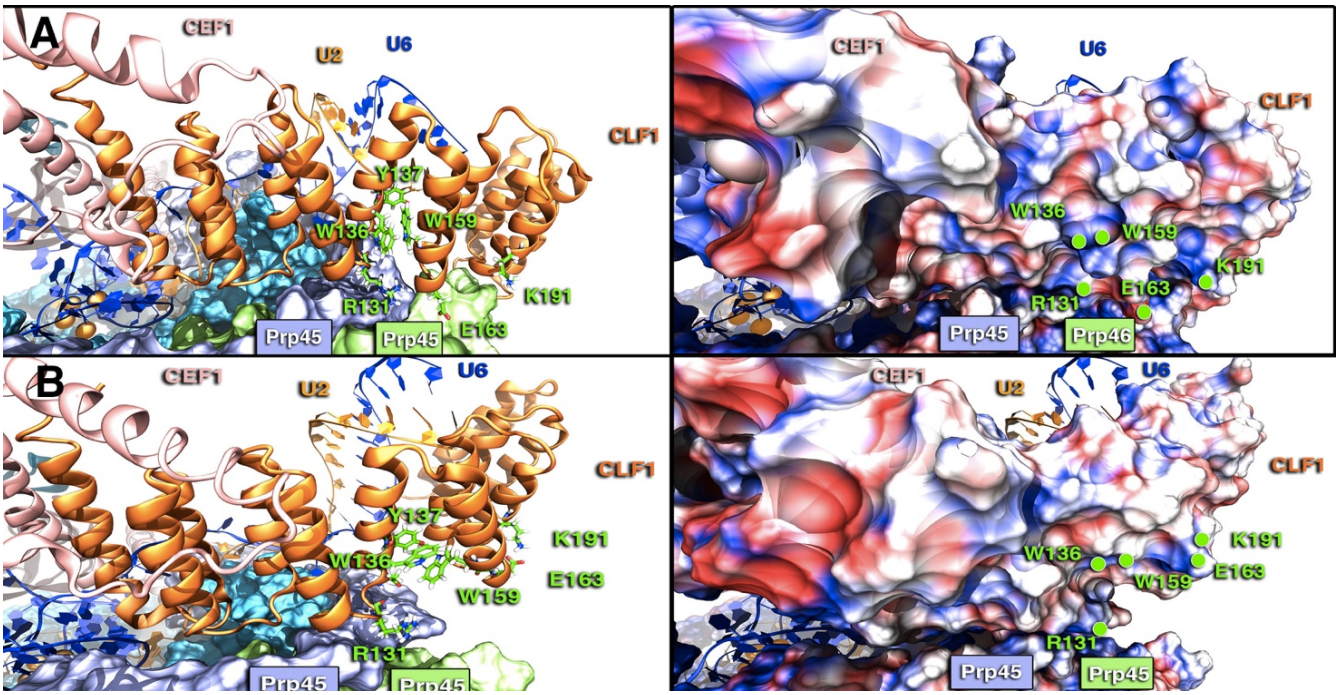


Fig. C.12 The electrostatic hinge of Clf1 and the rearrangements of hydrophobic interactions. (Left)Prp45 (light green), Prp46 (lilac) and Prp8 (cyan) are represented with surface; U2 (orange) and U6 (blue) are shown as New Ribbon. (Right) Proteins are shown with electrostatic surface (blue/red colors for positive/negative charges, respectively). The motions of Clf1 appears to be modulated by a rearrangement of salt-bridge interactions between E163 and K191, while the plasticity of the hinge located at H2-3 is associated to a reorganization of extended π-stacking interactions involving Y137, W159, W136. (A) Namely, Y137 establishes a T-shaped stacking with W159, which π-stacks with W136 through a parallel-displaced conformation. (B) After the functional rearrangement, W136 forms a T-shape stacking with Y137, inducing a downstream displacement of the nearby HAT-repeats.

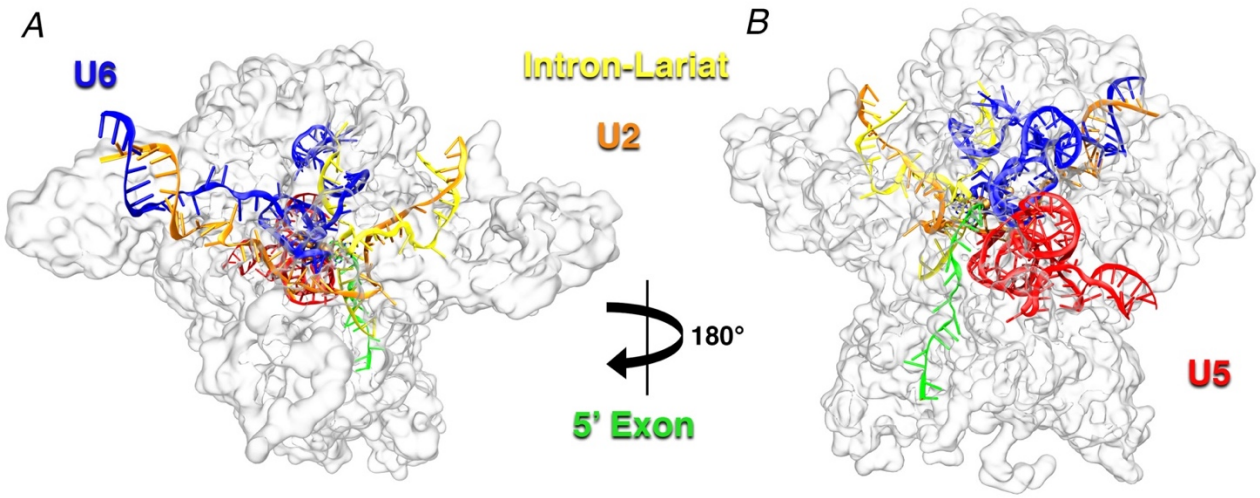


Fig. C.13 Small nuclear (sn)RNAs elements within the spliceosome. Front (A) and back (B) views of RNAs positions. In transparent surface is depicted the protein counterpart of SPL and 5'Exon, Intron-Lariat, U6 snRNA, U2 snRNA, U5 snRNA are displayed in green, yellow, blue, orange and red, respectively.

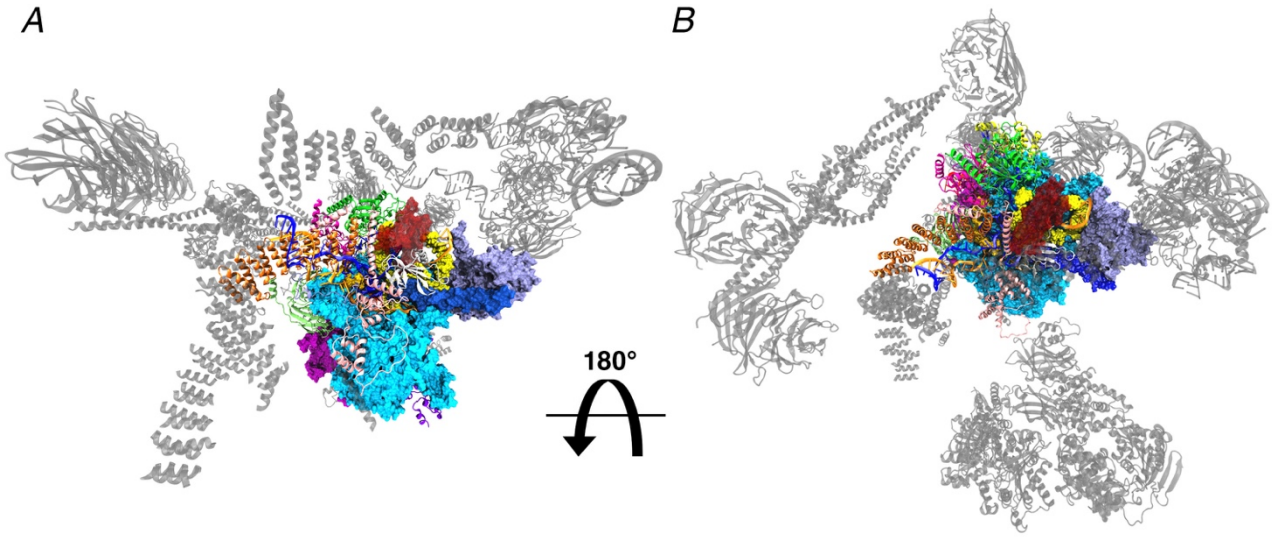


Fig. C.14 Overall structure of C complex (PDB ID: 5LJ5[32] *) front (A) and top (B). In grey are displayed all the proteins that were not included into the model for the MD simulation.*

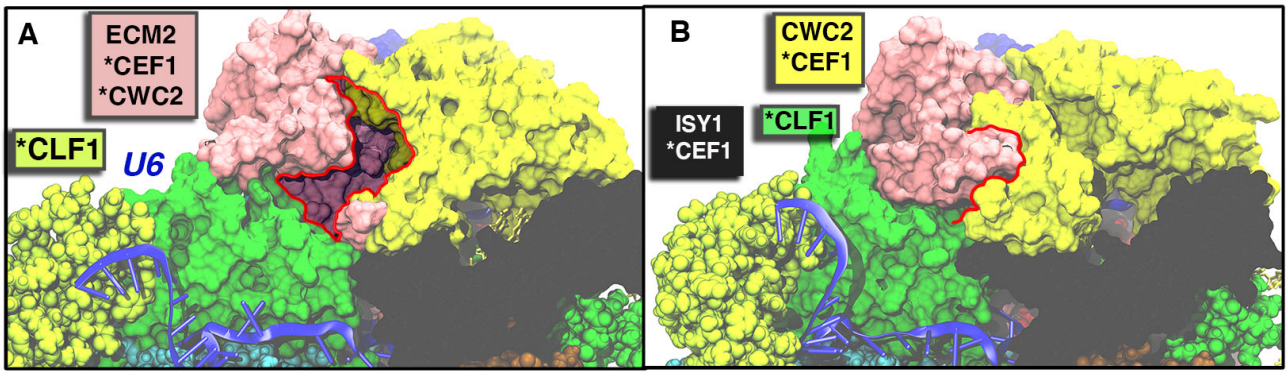


Fig. C.15 Possible binding pocket lying on the communication path II and its open (A) or closed (B) states of the 'hammer-like' motion described by PC1. Small molecules targeting this region could critically interfere with the internal communication network underlying the spliceosome dynamics. Asterisks indicate domains or proteins spread in distinct communities. Proteins are represented with the same color code of Figure 3C.

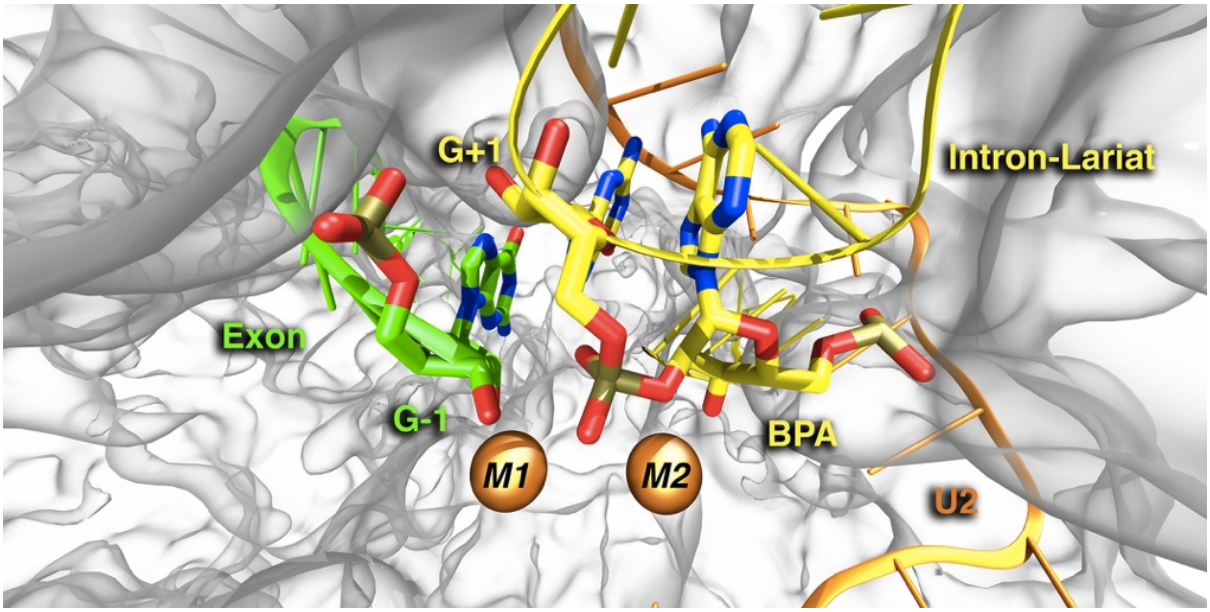


Fig. C.16 5'Splicing-Site showing the truncated 5'exon in green and the Branching Point Adenosine bound to the first intronic base (G+1) via the non-canonical 5'phosphate-O2'oxygen bond. The two Mg^{2+} ions are in orange van der Waals spheres. During the MD simulations, the O3' extremity of the cleaved 5'-exon remains at an average distance of 3.2 Å from the 5'-phosphate of the intron nucleotide G(+1).

C.2 Additional Tables

SPL C Complex Model				
PROTEIN NAME	**CONSIDERED**	**MODELLED**	**RESOLUTION**	**LENGTH**
U5 snRNA	28-53 + 62-125	/	3.8 – 7.6	90
EXON	-16 - -1	/	3.4 – 6.4	16
LARIAT INTRON	1-10 + 54-76	/	3.4 – 7.2	32
U2 snRNA	3-47	/	3.8 - 6.0	45
U6 snRNA	16-102	/	3.6 – 6.4	87
Prp8	128-2085	429-457	3.4 – 5.8	1958
YJU2 (CWC16)	2-115	/	3.8 – 5.4	114
CWC25	3-48	/	3.8 – 7.0	46
SNU114	71-693	516-533	3.8 – 7.2	623
ISY1	1-97	/	3.8 – 6.2	97
CWC22	289-481	400-413	4.6 – 8.2	193
Prp46	111-428	/	3.4 – 6.6	318
Prp45	104-184	151-158	4 - 8.4	81
BUD31	2-156	/	3.6 – 6.8	155
CWC2	3-254	/	3.6 – 6.0	252
ECM2 (SLT11)	6-144	93-100	4.0 – 7.0	139
CEF1	12-247	101-147	3.8 – 6.2	236
CWC15	7-42	/	3.6 – 7.6	36
CWC21	2-50	/	3.8 – 7.4	49
CLF1	37-273	/	3.8 – 6.4	237
Mg+	#5	Saxena Force Field		
Zn(2+)	#7	Pang Force Field		
NA+	#201	Joung & Cheatham FF		
Wat mol	#229850	TIP3P		
Total number of atoms (System) : 772679		**Protein Force Field:** ff12SB	**Cryo-EM:** 3.8 Å (5lj3)	
Solute atoms (SPL): 83129		**RNA Force Field:** ff99+bsc0+χOL3FF	**Organism:** *Schizosaccharomyces Cerevisiae*	

Tab. C.1 Protein/RNA composition of the SPL. CONSIDERED column stand for the residues included in the simulation. MODELLED refers to residues modelled using MODELLER 9v16 [232]. *RESOLUTION indicates the cryo-EM resolution for each protein/RNA.*

PATH I		PATH II	
Protein	Residue	Protein	Residue
CLF1	ARG62	CLF1	GLU89
CWC15	**SER13**	CEF1	ASP216
CWC15	ALA11	CEF1	TYR213
Prp8 (N-Ter)	LEU783	ECM2 (SLT11)	GLU103
Prp8 (N-Ter)	GLU788	CWC2	ALA118
Prp8 (N-Ter)	**CYS792**	CWC2	LYS116
Prp8 (N-Ter)	ALA795	CWC2	LEU109
Prp8 (RT)	MET1095	CWC2	ARG63
Prp8 (RT)	HIS1097	BUD31	PHE142
Prp8 (RT)	**ASN1099**	Prp8 (N-Ter)	**GLN558**
YJU2 (CWC16)	PHE97	Prp8 (N-Ter)	LEU192
YJU2 (CWC16)	ARG83	Prp8 (N-Ter)	**ASN203**
YJU2 (CWC16)	ILE81	Prp8 (N-Ter)	**THR205**
YJU2 (CWC16)	ILE79	Prp8 (N-Ter)	**ARG207**
CWC25	THR26	Prp8 (N-Ter)	**ILE209**
CWC25	LEU30	Prp8 (N-Ter)	**LEU318**
		Prp8 (N-Ter)	ASP651
		Prp8 (N-Ter)	ARG236
		Prp8 (Endo)	PHE1756
		Prp8 (Endo)	VAL1662
		Prp8 Endonuclease	ASP1664
		Prp8 (RNAse-H)	LYS1912

Tab. C.2 List of residues lying along the communication pathways I and II (node betweenness >0.6). In bold are reported the key residues with node betweenness >0.85. Table cells are colored with the same color code of Figure 1 of the main text.

Acknowledgments

This thesis represents for me the end of one of the most difficult periods of my life, which has given to me so much but has also taken the same. For this reason I couldn't make it (especially mentally) without some people in particular.

The first person I have to thank and to whom I owe so much of what I have become today is Caterina, my ever-present, faithful and supportive companion, who celebrated my goals with me but who also saw me in the darkest moments, helping me to rationalize, to grow but above all to believe in myself, allowing me to reach the end of this path with still a little sanity (albeit little).

Next thanks goes to my supervisor, Alessandra Magistrato, always ready with an original and alternative solution to problems, whose experience helped me in my academic career.

To Stefano, the most unlikely person who I could ever have been friend with, but who has proven to be a precious shoulder to count on in these years, someone to confront with (but be careful when talking about politics and taxes!) and with whom to let off steam. But above all, thanks for the fundamental help when I first discovered what statistical mechanics was.

To the other two crazy colleagues, Claudio and Matteo, whose contagious light-heartedness has lightened the burden of the exams.

To Angelo and Lorenzo, who taught me all the tricks of the trade and helped whenever I needed it.

Thanks to my parents, who taught me to fend for myself (sometimes even in the hard way) and who gave - and hoped - so much for my education. To my sisters, who have helped me silently but with flawless timing until today. I hope I have paid off at least a small part of all your efforts.

To Maurizio, Gloria and Vittoria, a new acquired family that has always cheered for me and whose warmth and affection has supported my efforts, giving me precious suggestions in times of need.

Finally, thanks to Thorben, Lucia, Papale, Martina, Matteo and Giulia with whom I spent magical evenings between endless board games, dinners with improbable recipes and secret Santa Clauses. To all these friends, now partially around the world, I wish you a bright and happy life.

8 Bibliography

[1] S. M. Berget, C. Moore, and P. A. Sharp, “Spliced segments at the 5’ terminus of adenovirus 2 late mRNA,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 8, pp. 3171–3175, Aug. 1977, doi: 10.1073/pnas.74.8.3171.

[2] P. A. Sharp, “The discovery of split genes and RNA splicing,” *Trends Biochem. Sci.*, vol. 30, no. 6, pp. 279–281, Jun. 2005, doi: 10.1016/j.tibs.2005.04.002.

[3] N. Toor, K. S. Keating, and A. M. Pyle, “Structural insights into RNA splicing,” *Curr. Opin. Struct. Biol.*, vol. 19, no. 3, pp. 260–266, Jun. 2009, doi: 10.1016/j.sbi.2009.04.002.

[4] P. Papasaikas and J. Valcárcel, “The Spliceosome: The Ultimate RNA Chaperone and Sculptor,” *Trends Biochem. Sci.*, vol. 41, no. 1, pp. 33–45, Jan. 2016, doi: 10.1016/j.tibs.2015.11.003.

[5] P. Early *et al.*, “Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways,” *Cell*, vol. 20, no. 2, pp. 313–319, Jun. 1980, doi: 10.1016/0092-8674(80)90617-0.

[6] T. W. Nilsen and B. R. Graveley, “Expansion of the eukaryotic proteome by alternative splicing,” *Nature*, vol. 463, no. 7280, pp. 457–463, Jan. 2010, doi: 10.1038/nature08909.

[7] M. Buljan *et al.*, “Alternative splicing of intrinsically disordered regions and rewiring of protein interactions,” *Curr. Opin. Struct. Biol.*, vol. 23, no. 3, pp. 443–450, Jun. 2013, doi: 10.1016/j.sbi.2013.03.006.

[8] J. D. Ellis *et al.*, “Tissue-Specific Alternative Splicing Remodels Protein-Protein Interaction Networks,” *Mol. Cell*, vol. 46, no. 6, pp. 884–892, Jun. 2012, doi: 10.1016/j.molcel.2012.05.037.

[9] X. Yang *et al.*, “Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing,” *Cell*, vol. 164, no. 4, pp. 805–817, 2016, doi: 10.1016/j.cell.2016.01.029.

[10] U. Braunschweig, S. Gueroussov, A. M. Plocik, B. R. Graveley, and B. J. Blencowe,

"Dynamic integration of splicing within gene regulatory pathways," *Cell*, vol. 152, no. 6, pp. 1252–1269, Mar. 2013, doi: 10.1016/j.cell.2013.02.034.

[11] S. Sen, "Aberrant pre-mRNA splicing regulation in the development of hepatocellular carcinoma," *Hepatoma Res.*, vol. 4, no. 7, p. 37, Jul. 2018, doi: 10.20517/2394-5079.2018.39.

[12] A. Kalsotra and T. A. Cooper, "Functional consequences of developmentally regulated alternative splicing," *Nat. Rev. Genet.*, vol. 12, no. 10, pp. 715–729, Oct. 2011, doi: 10.1038/nrg3052.

[13] J. Ye and R. Blelloch, "Regulation of pluripotency by RNA binding proteins," *Cell Stem Cell*, vol. 15, no. 3, pp. 271–280, Sep. 2014, doi: 10.1016/j.stem.2014.08.010.

[14] L. T. Chow, R. E. Gelinas, T. R. Broker, and R. J. Roberts, "An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA," *Cell*, vol. 12, no. 1, pp. 1–8, 1977, doi: 10.1016/0092-8674(77)90180-5.

[15] A. R. Robart, R. T. Chan, J. K. Peters, K. R. Rajashankar, and N. Toor, "Crystal structure of a eukaryotic group II intron lariat," *Nature*, vol. 514, no. 7521, pp. 193–197, Sep. 2014, doi: 10.1038/nature13790.

[16] C. Yan, J. Hang, R. Wan, M. Huang, C. C. L. Wong, and Y. Shi, "Structure of a yeast spliceosome at 3.6-angstrom resolution," *Science (80-.).*, vol. 349, no. 6253, pp. 1182–1191, Sep. 2015, doi: 10.1126/science.aac7629.

[17] J. Merkin, C. Russell, P. Chen, and C. B. Burge, "Evolutionary dynamics of gene and isoform regulation in mammalian tissues," *Science (80-.).*, vol. 338, no. 6114, pp. 1593–1599, Dec. 2012, doi: 10.1126/science.1228186.

[18] N. L. Barbosa-Morais *et al.*, "The evolutionary landscape of alternative splicing in vertebrate species," *Science (80-.).*, vol. 338, no. 6114, pp. 1587–1593, Dec. 2012, doi: 10.1126/science.1230612.

[19] T. W. Nilsen and B. R. Graveley, "Expansion of the eukaryotic proteome by alternative splicing," *Nature*, vol. 463, no. 7280, pp. 457–463, Jan. 2010, doi: 10.1038/nature08909.

[20] E. T. Wang *et al.*, "Alternative isoform regulation in human tissue transcriptomes," *Nature*, vol. 456, no. 7221, pp. 470–476, Nov. 2008, doi: 10.1038/nature07509.

[21] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nat. Genet.*, vol. 40, no. 12, pp. 1413–1415, Nov. 2008, doi: 10.1038/ng.259.

[22] E. Brody and J. Abelson, "The 'spliceosome': Yeast pre-messenger RNA associates with a 40s complex in a splicing-dependent reaction," *Science (80-.).*, vol. 228, no. 4702, pp. 963–967, 1985, doi: 10.1126/science.3890181.

[23] P. J. Grabowski, S. R. Seiler, and P. A. Sharp, "A multicomponent complex is involved in the splicing of messenger RNA precursors," *Cell*, vol. 42, no. 1, pp. 345–353, Aug. 1985, doi: 10.1016/S0092-8674(85)80130-6.

[24] C. L. Will and R. Lührmann, "Spliceosome structure and function," *Cold Spring Harb. Perspect. Biol.*, vol. 3, no. 7, pp. 1–2, Jul. 2011, doi: 10.1101/cshperspect.a003707.

[25] N. A. Faustino and T. A. Cooper, "Pre-mRNA splicing and human disease," *Genes Dev.*, vol. 17, no. 4, pp. 419–437, Feb. 2003, doi: 10.1101/gad.1048803.

[26] M. C. Wahl, C. L. Will, and R. Lührmann, "The Spliceosome: Design Principles of a Dynamic RNP Machine," *Cell*, vol. 136, no. 4, pp. 701–718, 2009, doi: 10.1016/j.cell.2009.02.009.

[27] R. Parker, P. G. Siliciano, and C. Guthrie, "Recognition of the TACTAAC box during mRNA splicing in yeast involves base pairing to the U2-like snRNA," *Cell*, vol. 49, no. 2, pp. 229–239, Apr. 1987, doi: 10.1016/0092-8674(87)90564-2.

[28] U. Braunschweig, S. Gueroussov, A. M. Plocik, B. R. Graveley, and B. J. Blencowe, "Dynamic integration of splicing within gene regulatory pathways," *Cell*, vol. 152, no. 6, pp. 1252–1269, Mar. 2013, doi: 10.1016/j.cell.2013.02.034.

[29] M. M. Scotti and M. S. Swanson, "RNA mis-splicing in disease," *Nat. Rev. Genet.*, vol. 17, no. 1, pp. 19–32, Jan. 2016, doi: 10.1038/nrg.2015.3.

[30] M. Fromont-Racine *et al.*, "Genome-Wide Protein Interaction Screens Reveal Functional Networks Involving Sm-Like Proteins," *Yeast*, vol. 1, no. 2, pp. 95–110, Jan. 2000, doi: 10.1155/2000/919260.

[31] J. Boriš, L. Casalino, A. Saltalamacchia, S. G. Mays, L. Malcovati, and A. Magistrato, "Atomic-Level Mechanism of Pre-mRNA Splicing in Health and Disease," *Acc.*

Chem. Res, vol. 54, p. 16, 2021, doi: 10.1021/acs.accounts.0c00578.

[32] W. P. Galej, M. E. Wilkinson, S. M. Fica, C. Oubridge, A. J. Newman, and K. Nagai, "Cryo-EM structure of the spliceosome immediately after branching," *Nature*, vol. 537, no. 7619, pp. 197–201, 2016, doi: 10.1038/nature19316.

[33] J. Hang, R. Wan, C. Yan, and Y. Shi, "Structural basis of pre-mRNA splicing," *Science (80-.).*, vol. 349, no. 6253, pp. 1191–1198, Sep. 2015, doi: 10.1126/science.aac8159.

[34] C. Yan, R. Wan, R. Bai, G. Huang, and Y. Shi, "Structure of a yeast activated spliceosome at 3.5 Å resolution," *Science (80-.).*, vol. 353, no. 6302, pp. 904–912, Aug. 2016, doi: 10.1126/science.aag0291.

[35] R. Wan *et al.*, "The 3.8 Å structure of the U4/U6.U5 tri-snRNP: Insights into spliceosome assembly and catalysis," *Science (80-.).*, vol. 351, no. 6272, pp. 466–475, Jan. 2016, doi: 10.1126/science.aad6466.

[36] T. H. D. Nguyen *et al.*, "Cryo-EM structure of the yeast U4/U6.U5 tri-snRNP at 3.7 Å resolution," *Nature*, vol. 530, no. 7590, pp. 298–302, Feb. 2016, doi: 10.1038/nature16940.

[37] D. E. Agafonov *et al.*, "Molecular architecture of the human U4/U6.U5 tri-snRNP," *Science (80-.).*, vol. 351, no. 6280, Mar. 2016, doi: 10.1126/science.aad2085.

[38] J. H. D. Cate, "A Big Bang in spliceosome structural biology," *Science (80-.).*, vol. 351, no. 6280, pp. 1390–1392, Mar. 2016, doi: 10.1126/science.aaf4465.

[39] R. Rauhut *et al.*, "Molecular architecture of the Saccharomyces cerevisiae activated spliceosome," *Science (80-.).*, vol. 353, no. 6306, pp. 1399–1405, 2016, doi: 10.1126/science.aag1906.

[40] P. D. Zamore, J. G. Patton, and M. R. Green, "Cloning and domain structure of the mammalian splicing factor U2AF," *Nature*, vol. 355, no. 6361, pp. 609–614, 1992, doi: 10.1038/355609a0.

[41] L. Merendino, S. Guth, D. Bilbao, C. Martínez, and J. Valcárcel, "Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3' splice site AG," *Nature*, vol. 402, no. 6763, pp. 838–841, Dec. 1999, doi: 10.1038/45602.

[42] S. Wu, C. M. Romfo, T. W. Nilsen, and M. R. Green, "Functional recognition 3' splice

site AG by the splicing factor U2AF35," *Nature*, vol. 402, no. 6763, pp. 832–835, 1999, doi: 10.1038/45590.

[43] P. D. Zamore and M. R. Green, "Identification, purification, and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 86, no. 23, pp. 9243–9247, Dec. 1989, doi: 10.1073/pnas.86.23.9243.

[44] D. A. R. Zorio and T. Blumenthal, "U2AF35 is encoded by an essential gene clustered in an operan with RRM/cyclophilin in Caenorhabditis elegans," *Rna*, vol. 5, no. 4, pp. 487–494, Apr. 1999, doi: 10.1017/S1355838299982225.

[45] A. Corrionero, B. Miñana, and J. Valcárcel, "Reduced fidelity of branch point recognition and alternative splicing induced by the anti-tumor drug spliceostatin A," *Genes Dev.*, vol. 25, no. 5, pp. 445–459, Mar. 2011, doi: 10.1101/gad.2014311.

[46] M. J. Schellenberg, E. L. Dul, and A. M. MacMillan, "Structural model of the p14/SF3b155·branch duplex complex," *Rna*, vol. 17, no. 1, pp. 155–165, Jan. 2011, doi: 10.1261/rna.2224411.

[47] O. Gozani, J. Potashkin, and R. Reed, "A Potential Role for U2AF-SAP 155 Interactions in Recruiting U2 snRNP to the Branch Site," *Mol. Cell. Biol.*, vol. 18, no. 8, pp. 4752–4760, Aug. 1998, doi: 10.1128/mcb.18.8.4752.

[48] M. R. Green, "Biochemical mechanisms of constitutive and regulated Pre-mRNA splicing," *Annu. Rev. Cell Biol.*, vol. 7, pp. 559–599, Nov. 1991, doi: 10.1146/annurev.cb.07.110191.003015.

[49] M. McKeown, "Alternative mRNA splicing," *Annu. Rev. Cell Biol.*, vol. 8, pp. 133–155, 1992, doi: 10.1146/annurev.cb.08.110192.001025.

[50] B. Nadal-Ginard, C. W. J. Smith, J. G. Patton, and R. E. Breitbart, "Alternative splicing is an efficient mechanism for the generation of protein diversity: Contractile protein genes as a model system," *Adv. Enzyme Regul.*, vol. 31, no. C, pp. 261–286, Jan. 1991, doi: 10.1016/0065-2571(91)90017-G.

[51] M. P. Mullen, C. W. J. Smith, J. G. Patton, and B. Nadal-Ginard, "α-Tropomyosin mutually exclusive exon selection: Competition between branchpoint/polypyrimidine tracts determines default exon choice," *Genes Dev.*, vol. 5, no. 4, pp. 642–655, 1991,

doi: 10.1101/gad.5.4.642.

[52] A. Bindereif and M. R. Green, “Ribonucleoprotein complex formation during pre-mRNA splicing in vitro,” *Mol. Cell. Biol.*, vol. 6, no. 7, pp. 2582–2592, Jul. 1986, doi: 10.1128/mcb.6.7.2582-2592.1986.

[53] D. Frendewey and W. Keller, “Stepwise assembly of a pre-mRNA splicing complex requires U-snRNPs and specific intron sequences,” *Cell*, vol. 42, no. 1, pp. 355–367, 1985, doi: 10.1016/S0092-8674(85)80131-8.

[54] R. Reed, “The organization of 3’ splice-site sequences in mammalian introns,” *Genes Dev.*, vol. 3, no. 12 B, pp. 2113–2123, 1989, doi: 10.1101/gad.3.12b.2113.

[55] B. Ruskin and M. R. Green, “Role of the 3′ splice site consensus sequence in mammalian pre-mRNA splicing,” *Nature*, vol. 317, no. 6039, pp. 732–734, 1985, doi: 10.1038/317732a0.

[56] P. A. Norton, “Alternative pre-mRNA splicing: Factors involved in splice site selection,” *J. Cell Sci.*, vol. 107, no. 1, pp. 1–7, Jan. 1994, doi: 10.1242/jcs.107.1.1.

[57] R. F. Roscigno, M. Weiner, and M. A. Garcia-Blanco, “A mutational analysis of the polypyrimidine tract of introns. Effects of sequence differences in pyrimidine tracts on splicing,” *J. Biol. Chem.*, vol. 268, no. 15, pp. 11222–11229, May 1993, doi: 10.1016/s0021-9258(18)82114-7.

[58] C. J. Coolidge, R. J. Seely, and J. G. Patton, “Functional analysis of the polypyrimidine tract in pre-mRNA splicing,” *Nucleic Acids Res.*, vol. 25, no. 4, pp. 888–895, 1997, doi: 10.1093/nar/25.4.888.

[59] A. J. Matlin, F. Clark, and C. W. J. Smith, “Understanding alternative splicing: Towards a cellular code,” *Nat. Rev. Mol. Cell Biol.*, vol. 6, no. 5, pp. 386–398, May 2005, doi: 10.1038/nrm1645.

[60] C. W. J. Smith and J. Valcárcel, “Alternative pre-mRNA splicing: the logic of combinatorial control,” *Trends Biochem. Sci.*, vol. 25, no. 8, pp. 381–388, Aug. 2000, doi: 10.1016/S0968-0004(00)01604-2.

[61] H. X. Liu, M. Zhang, and A. R. Krainer, “Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins,” *Genes Dev.*, vol. 12, no. 13, pp. 1998–2012, Jul. 1998, doi: 10.1101/gad.12.13.1998.

[62] B. R. Graveley, “Sorting out the complexity of SR protein functions,” *Rna*, vol. 6, no. 9, pp. 1197–1211, 2000, doi: 10.1017/S1355838200000960.

[63] J. Valcárcel, R. K. Gaur, R. Singh, and M. R. Green, “Interaction of U2AF65 RS region with pre-mRNA of branch point and promotion base pairing with U2 snRNA,” *Science (80-.).*, vol. 273, no. 5282, pp. 1706–1709, Sep. 1996, doi: 10.1126/science.273.5282.1706.

[64] B. R. Graveley and T. Maniatis, “Arginine/serine-rich domains of SR proteins can function as activators of pre-mRNA splicing,” *Mol. Cell*, vol. 1, no. 5, pp. 765–771, Apr. 1998, doi: 10.1016/S1097-2765(00)80076-3.

[65] B. Chabot and L. Shkreta, “Defective control of pre-messenger RNA splicing in human disease,” *J. Cell Biol.*, vol. 212, no. 1, pp. 13–27, Jan. 2016, doi: 10.1083/jcb.201510032.

[66] S. Lefebvre *et al.*, “Identification and characterization of a spinal muscular atrophy-determining gene,” *Cell*, vol. 80, no. 1, pp. 155–165, Jan. 1995, doi: 10.1016/0092-8674(95)90460-3.

[67] C. L. Lorson, E. Hahnen, E. J. Androphy, and B. Wirth, “A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 11, pp. 6307–6311, May 1999, doi: 10.1073/pnas.96.11.6307.

[68] E. Daguenet, G. Dujardin, and J. Valcárcel, “ The pathogenicity of splicing defects: mechanistic insights into pre- mRNA processing inform novel therapeutic approaches ,” *EMBO Rep.*, vol. 16, no. 12, pp. 1640–1655, Dec. 2015, doi: 10.15252/embr.201541116.

[69] R. Kole, A. R. Krainer, and S. Altman, “RNA therapeutics: Beyond RNA interference and antisense oligonucleotides,” *Nat. Rev. Drug Discov.*, vol. 11, no. 2, pp. 125–140, Jan. 2012, doi: 10.1038/nrd3625.

[70] C. J. David and J. L. Manley, “Alternative pre-mRNA splicing regulation in cancer: Pathways and programs unhinged,” *Genes Dev.*, vol. 24, no. 21, pp. 2343–2364, Nov. 2010, doi: 10.1101/gad.1973010.

[71] J. Cheng *et al.*, “Protection from Fas-mediated apoptosis by a soluble form of the Fas

molecule," *Science (80-.).*, vol. 263, no. 5144, pp. 1759–1762, 1994, doi: 10.1126/science.7510905.

[72] S. Bonnal, L. Vigevani, and J. Valcárcel, "The spliceosome as a target of novel antitumour drugs," *Nat. Rev. Drug Discov.*, vol. 11, no. 11, pp. 847–859, Nov. 2012, doi: 10.1038/nrd3823.

[73] K. Yoshida and S. Ogawa, "Splicing factor mutations and cancer," *Wiley Interdiscip. Rev. RNA*, vol. 5, no. 4, pp. 445–459, Jul. 2014, doi: 10.1002/wrna.1222.

[74] K. Yoshida *et al.*, "Frequent pathway mutations of splicing machinery in myelodysplasia," *Nature*, vol. 478, no. 7367, pp. 64–69, Sep. 2011, doi: 10.1038/nature10496.

[75] E. Papaemmanuil *et al.*, " Somatic SF3B1 Mutation in Myelodysplasia with Ring Sideroblasts ," *N. Engl. J. Med.*, vol. 365, no. 15, pp. 1384–1395, Oct. 2011, doi: 10.1056/nejmoa1103283.

[76] V. Quesada *et al.*, "Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia," *Nat. Genet.*, vol. 44, no. 1, pp. 47–52, Jan. 2012, doi: 10.1038/ng.1032.

[77] S. J. Furney *et al.*, "SF3B1 mutations are associated with alternative splicing in uveal melanoma," *Cancer Discov.*, vol. 3, no. 10, pp. 1122–1129, Oct. 2013, doi: 10.1158/2159-8290.CD-13-0330.

[78] E. Glasser, A. A. Agrawal, J. L. Jenkins, and C. L. Kielkopf, "Cancer-Associated Mutations Mapped on High-Resolution Structures of the U2AF2 RNA Recognition Motifs," *Biochemistry*, vol. 56, no. 36, pp. 4757–4761, 2017, doi: 10.1021/acs.biochem.7b00551.

[79] E. Zuccato, E. Buratti, C. Stuani, F. E. Baralle, and F. Pagani, "An Intronic Polypyrimidine-rich Element Downstream of the Donor Site Modulates Cystic Fibrosis Transmembrane Conductance Regulator Exon 9 Alternative Splicing," *J. Biol. Chem.*, vol. 279, no. 17, pp. 16980–16988, Apr. 2004, doi: 10.1074/jbc.M313439200.

[80] R. C. Davies, C. Calvio, E. Bratt, S. H. Larsson, A. I. Lamond, and N. D. Hastie, "WT1 interacts with the splicing factor U2AF65 in an isoform-dependent manner and can be incorporated into spliceosomes," *Genes Dev.*, vol. 12, no. 20, pp. 3217–3225,

Oct. 1998, doi: 10.1101/gad.12.20.3217.

[81] J. Fay, P. Kelehan, H. Lambkin, and S. Schwartz, "Increased expression of cellular RNA-binding proteins in HPV-induced neoplasia and cervical cancer," *J. Med. Virol.*, vol. 81, no. 5, pp. 897–907, May 2009, doi: 10.1002/jmv.21406.

[82] Y. Chen, L. Zhang, and K. A. Jones, "SKIP counteracts p53-mediated apoptosis via selective regulation of p21 Cip1 mRNA splicing," *Genes Dev.*, vol. 25, no. 7, pp. 701–716, Apr. 2011, doi: 10.1101/gad.2002611.

[83] G. Tiscornia and M. S. Mahadevan, "Myotonic dystrophy: The role of the CUG triplet repeats in splicing of a novel DMPK exon and altered cytoplasmic DMPK mRNA isoform ratios," *Mol. Cell*, vol. 5, no. 6, pp. 959–967, 2000, doi: 10.1016/S1097-2765(00)80261-0.

[84] H. Dvinge, E. Kim, O. Abdel-Wahab, and R. K. Bradley, "RNA splicing factors as oncoproteins and tumour suppressors," *Nat. Rev. Cancer*, vol. 16, no. 7, pp. 413–430, Jun. 2016, doi: 10.1038/nrc.2016.51.

[85] D. Frenkel and B. Smit, *Understanding molecular simulation: From algorithms to applications*. 1996.

[86] W. F. Van Gunsteren and H. J. C. Berendsen, "A Leap-Frog Algorithm for Stochastic Dynamics," *Mol. Simul.*, vol. 1, no. 3, pp. 173–185, 1988, doi: 10.1080/08927028808080941.

[87] "Molecular Modelling: Principles and Applications - Andrew R. Leach, Leach AR." https://books.google.it/books?hl=it&lr=&id=kB7jsbV-uhkC&oi=fnd&pg=PR9&ots=--tqc2mjKU&sig=7_8Tj9gx8W3AbPa41AK3JAINK7g#v=onepage&q&f=false (accessed Oct. 09, 2021).

[88] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems," *J. Chem. Phys.*, vol. 98, no. 12, pp. 10089–10092, Aug. 1993, doi: 10.1063/1.464397.

[89] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," *J. Chem. Phys.*, vol. 126, no. 1, p. 014101, Jan. 2007, doi: 10.1063/1.2408420.

[90] H. J. C. Berendsen, J. P. M. Postma, W. F. Van Gunsteren, A. Dinola, and J. R. Haak,

"Molecular dynamics with coupling to an external bath," *J. Chem. Phys.*, vol. 81, no. 8, pp. 3684–3690, Aug. 1984, doi: 10.1063/1.448118.

[91] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," *J. Appl. Phys.*, vol. 52, no. 12, pp. 7182–7190, Aug. 1981, doi: 10.1063/1.328693.

[92] C. J. Tsai, A. del Sol, and R. Nussinov, "Allostery: Absence of a Change in Shape Does Not Imply that Allostery Is Not at Play," *J. Mol. Biol.*, vol. 378, no. 1, pp. 1–11, Apr. 2008, doi: 10.1016/j.jmb.2008.02.034.

[93] A. T. Van Wart, J. Durrant, L. Votapka, and R. E. Amaro, "Weighted implementation of suboptimal paths (WISP): An optimized algorithm and tool for dynamical network analysis," *J. Chem. Theory Comput.*, vol. 10, no. 2, pp. 511–517, Feb. 2014, doi: 10.1021/ct4008603.

[94] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numer. Math.*, vol. 1, no. 1, pp. 269–271, Dec. 1959, doi: 10.1007/BF01386390.

[95] K. Mehlhorn and P. Sanders, "Algorithms and data structures: The basic toolbox," *Algorithms Data Struct. Basic Toolbox*, pp. 1–300, 2008, doi: 10.1007/978-3-540-77978-0.

[96] I. Rivalta, M. M. Sultan, N. S. Lee, G. A. Manley, J. P. Loria, and V. S. Batista, "Allosteric pathways in imidazole glycerol phosphate synthase," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 22, 2012, doi: 10.1073/pnas.1120536109.

[97] O. F. Lange and H. Grubmüller, "Generalized correlation for biomolecular dynamics," *Proteins Struct. Funct. Genet.*, vol. 62, no. 4, pp. 1053–1061, 2006, doi: 10.1002/prot.20784.

[98] G. Palermo, Y. Miao, R. C. Walker, M. Jinek, and J. A. McCammon, "Striking plasticity of CRISPR-Cas9 and key role of non-target DNA, as revealed by molecular simulations," *ACS Cent. Sci.*, vol. 2, no. 10, pp. 756–763, Oct. 2016, doi: 10.1021/acscentsci.6b00218.

[99] G. Palermo *et al.*, "Protospacer Adjacent Motif-Induced Allostery Activates CRISPR-Cas9," *J. Am. Chem. Soc.*, vol. 139, no. 45, pp. 16028–16031, 2017, doi: 10.1021/jacs.7b05313.

[100] A. Sethi, J. Eargle, A. A. Black, and Z. Luthey-Schulten, “Dynamical networks in tRNA: Protein complexes,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 16, pp. 6620–6625, 2009, doi: 10.1073/pnas.0810961106.

[101] M. E. J. Newman and M. Girvan, “Finding and evaluating community structure in networks,” *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, vol. 69, no. 2 2, pp. 1–15, 2004, doi: 10.1103/PhysRevE.69.026113.

[102] R. W. Floyd, “Algorithm 97: Shortest path,” *Commun. ACM*, vol. 5, no. 6, p. 345, Jun. 1962, doi: 10.1145/367766.368168.

[103] S. Warshall, “A Theorem on Boolean Matrices,” *J. ACM*, vol. 9, no. 1, pp. 11–12, 1962, doi: 10.1145/321105.321107.

[104] A. Laio and M. Parrinello, “Escaping free-energy minima,” Accessed: Oct. 09, 2021. [Online]. Available: www.pnas.orgcgidoi10.1073pnas.202427399.

[105] A. Laio and F. L. Gervasio, “Metadynamics: A method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science,” *Reports Prog. Phys.*, vol. 71, no. 12, p. 126601, Nov. 2008, doi: 10.1088/0034-4885/71/12/126601.

[106] A. Barducci, G. Bussi, and M. Parrinello, “Well-tempered metadynamics: A smoothly converging and tunable free-energy method,” *Phys. Rev. Lett.*, vol. 100, no. 2, Jan. 2008, doi: 10.1103/PhysRevLett.100.020603.

[107] D. Branduardi, G. Bussi, and M. Parrinello, “Metadynamics with adaptive gaussians,” *J. Chem. Theory Comput.*, vol. 8, no. 7, pp. 2247–2254, Jul. 2012, doi: 10.1021/ct3002464.

[108] J. B. Jasper, L. Humbeck, T. Brinkjost, and O. Koch, “A novel interaction fingerprint derived from per atom score contributions: exhaustive evaluation of interaction fingerprint performance in docking based virtual screening,” *J. Cheminform.*, vol. 10, no. 1, pp. 1–13, Mar. 2018, doi: 10.1186/s13321-018-0264-0.

[109] W. Zhou *et al.*, “Discovery of Novel Androgen Receptor Ligands by Structure-based Virtual Screening and Bioassays,” *Genomics, Proteomics Bioinforma.*, vol. 16, no. 6, pp. 416–427, Dec. 2018, doi: 10.1016/j.gpb.2018.03.007.

[110] J. Lyu *et al.*, “Ultra-large library docking for discovering new chemotypes,” *Nature*,

vol. 566, no. 7743, pp. 224–229, Feb. 2019, doi: 10.1038/s41586-019-0917-9.

[111] L. G. Ferreira, R. N. Dos Santos, G. Oliva, and A. D. Andricopulo, "Molecular docking and structure-based drug design strategies," *Molecules*, vol. 20, no. 7, pp. 13384–13421, Jul. 2015, doi: 10.3390/molecules200713384.

[112] Y. Zhang *et al.*, "An integrated virtual screening approach for VEGFR-2 inhibitors," *J. Chem. Inf. Model.*, vol. 53, no. 12, pp. 3163–3177, Dec. 2013, doi: 10.1021/ci400429g.

[113] A. P. A. Janssen *et al.*, "Drug Discovery Maps, a Machine Learning Model That Visualizes and Predicts Kinome-Inhibitor Interaction Landscapes," *J. Chem. Inf. Model.*, vol. 59, no. 3, pp. 1221–1229, Mar. 2019, doi: 10.1021/acs.jcim.8b00640.

[114] D. Bajusz, G. G. Ferenczy, and G. M. Keseru, "Structure-based Virtual Screening Approaches in Kinase-directed Drug Discovery," *Curr. Top. Med. Chem.*, vol. 17, no. 20, Feb. 2017, doi: 10.2174/1568026617666170224121313.

[115] A. Kumar and K. Y. J. Zhang, "Hierarchical virtual screening approaches in small molecule drug discovery," *Methods*, vol. 71, no. C, pp. 26–37, 2015, doi: 10.1016/j.ymeth.2014.07.007.

[116] V. Ruiz-Torres *et al.*, "An updated review on marine anticancer compounds: The use of virtual screening for the discovery of small-molecule cancer drugs," *Molecules*, vol. 22, no. 7, p. 1037, Jun. 2017, doi: 10.3390/molecules22071037.

[117] D. C. Young, "Computational Drug Design: A Guide for Computational and Medicinal Chemists," *Comput. Drug Des. A Guid. Comput. Med. Chem.*, pp. 1–307, 2009, doi: 10.1002/9780470451854.

[118] T. I. Oprea and H. Matter, "Integrating virtual screening in lead discovery," *Curr. Opin. Chem. Biol.*, vol. 8, no. 4, pp. 349–358, Aug. 2004, doi: 10.1016/j.cbpa.2004.06.008.

[119] P. D. Lyne, "Structure-based virtual screening: An overview," *Drug Discov. Today*, vol. 7, no. 20, pp. 1047–1055, Oct. 2002, doi: 10.1016/S1359-6446(02)02483-2.

[120] P. Willett, "Similarity-based virtual screening using 2D fingerprints," *Drug Discov. Today*, vol. 11, no. 23–24, pp. 1046–1053, Dec. 2006, doi: 10.1016/j.drudis.2006.10.005.

[121] R. S. Ferreira *et al.*, "Complementarity between a docking and a high-throughput

screen in discovering new cruzain inhibitors," *J. Med. Chem.*, vol. 53, no. 13, pp. 4891–4905, Jul. 2010, doi: 10.1021/jm100488w.

[122] D. Huanga and A. Caflischa, "Library screening by fragment-based docking," *J. Mol. Recognit.*, vol. 23, no. 2, pp. 183–193, Mar. 2010, doi: 10.1002/jmr.981.

[123] I. D. Kuntz, "Structure-based strategies for drug design and discovery," *Science (80-.).*, vol. 257, no. 5073, pp. 1078–1082, 1992, doi: 10.1126/science.257.5073.1078.

[124] G. P. Brady and P. F. W. Stouten, "Fast prediction and visualization of protein binding pockets with PASS," *J. Comput. Aided. Mol. Des.*, vol. 14, no. 4, pp. 383–401, 2000, doi: 10.1023/A:1008124202956.

[125] M. Weisel, E. Proschak, and G. Schneider, "PocketPicker: Analysis of ligand binding-sites with shape descriptors," *Chem. Cent. J.*, vol. 1, no. 1, 2007, doi: 10.1186/1752-153X-1-7.

[126] M. Hendlich, F. Rippmann, and G. Barnickel, "LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins," *J. Mol. Graph. Model.*, vol. 15, no. 6, pp. 359–363, Dec. 1997, doi: 10.1016/S1093-3263(98)00002-3.

[127] R. A. Friesner *et al.*, "Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes," *J. Med. Chem.*, vol. 49, no. 21, pp. 6177–6196, Oct. 2006, doi: 10.1021/jm051256o.

[128] R. A. Friesner *et al.*, "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy," *J. Med. Chem.*, vol. 47, no. 7, pp. 1739–1749, Mar. 2004, doi: 10.1021/jm0306430.

[129] T. A. Halgren *et al.*, "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening," *J. Med. Chem.*, vol. 47, no. 7, pp. 1750–1759, Mar. 2004, doi: 10.1021/jm030644s.

[130] I. A. Guedes, F. S. S. Pereira, and L. E. Dardenne, "Empirical scoring functions for structure-based virtual screening: Applications, critical aspects, and challenges," *Front. Pharmacol.*, vol. 9, no. SEP, p. 1089, Sep. 2018, doi: 10.3389/fphar.2018.01089.

[131] S. Y. Huang, S. Z. Grinter, and X. Zou, "Scoring functions and their evaluation methods for protein-ligand docking: Recent advances and future directions," *Phys.*

Chem. Chem. Phys., vol. 12, no. 40, pp. 12899–12908, Oct. 2010, doi: 10.1039/c0cp00151a.

[132] J. Duan, S. L. Dixon, J. F. Lowrie, and W. Sherman, "Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods," *J. Mol. Graph. Model.*, vol. 29, no. 2, pp. 157–170, 2010, doi: 10.1016/j.jmgm.2010.05.008.

[133] M. Sastry, J. F. Lowrie, S. L. Dixon, and W. Sherman, "Large-Scale Systematic Analysis of 2D Fingerprint Methods and Parameters to Improve Virtual Screening Enrichments," *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 771–784, May 2010, doi: 10.1021/CI100062N.

[134] P. J. Shepard and K. J. Hertel, "The SR protein family.," *Genome Biol.*, vol. 10, no. 10, p. 242, Oct. 2009, doi: 10.1186/gb-2009-10-10-242.

[135] A. Busch and K. J. Hertel, "Evolution of SR protein and hnRNP splicing regulatory factors," *Wiley Interdiscip. Rev. RNA*, vol. 3, no. 1, pp. 1–12, Jan. 2012, doi: 10.1002/wrna.100.

[136] S. Jeong, "SR proteins: Binders, regulators, and connectors of RNA," *Mol. Cells*, vol. 40, no. 1, pp. 1–9, Jan. 2017, doi: 10.14348/molcells.2017.2319.

[137] Y. Huang, R. Gattoni, J. Stévenin, and J. A. Steitz, "SR splicing factors serve as adapter proteins for TAP-dependent mRNA export," *Mol. Cell*, vol. 11, no. 3, pp. 837–843, Mar. 2003, doi: 10.1016/S1097-2765(03)00089-3.

[138] X. Li and J. L. Manley, "Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability," *Cell*, vol. 122, no. 3, pp. 365–378, Aug. 2005, doi: 10.1016/j.cell.2005.06.008.

[139] X. Zheng *et al.*, "Serine/arginine-rich splicing factors: The bridge linking alternative splicing and cancer," *International Journal of Biological Sciences*, vol. 16, no. 13. Ivyspring International Publisher, pp. 2442–2453, 2020, doi: 10.7150/ijbs.46751.

[140] R. Karni, E. De Stanchina, S. W. Lowe, R. Sinha, D. Mu, and A. R. Krainer, "The gene encoding the splicing factor SF2/ASF is a proto-oncogene," *Nat. Struct. Mol. Biol.*, vol. 14, no. 3, pp. 185–193, Mar. 2007, doi: 10.1038/nsmb1209.

[141] J. Liu *et al.*, "Aberrant expression of splicing factors in newly diagnosed acute myeloid

leukemia," *Onkologie*, vol. 35, no. 6, pp. 335–340, Jun. 2012, doi: 10.1159/000338941.

[142] O. Anczuków *et al.*, "SRSF1-Regulated Alternative Splicing in Breast Cancer," *Mol. Cell*, vol. 60, no. 1, pp. 105–117, 2015, doi: 10.1016/j.molcel.2015.09.005.

[143] F. Supek, B. Miñana, J. Valcárcel, T. Gabaldón, and B. Lehner, "Synonymous mutations frequently act as driver mutations in human cancers," *Cell*, vol. 156, no. 6, pp. 1324–1335, Mar. 2014, doi: 10.1016/j.cell.2014.01.051.

[144] H. Jung *et al.*, "Intron retention is a widespread mechanism of tumor-suppressor inactivation," *Nat. Genet.*, vol. 47, no. 11, pp. 1242–1248, Nov. 2015, doi: 10.1038/ng.3414.

[145] L. Wang *et al.*, " SF3B1 and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia ," *N. Engl. J. Med.*, vol. 365, no. 26, pp. 2497–2506, Dec. 2011, doi: 10.1056/nejmoa1109016.

[146] J. W. Harbour, E. D. O. Roberson, H. Anbunathan, M. D. Onken, L. A. Worley, and A. M. Bowcock, "Recurrent mutations at codon 625 of the splicing factor SF3B1 in uveal melanoma," *Nat. Genet.*, vol. 45, no. 2, pp. 133–135, Feb. 2013, doi: 10.1038/ng.2523.

[147] M. A. Jensen, J. E. Wilkinson, and A. R. Krainer, "Splicing factor SRSF6 promotes hyperplasia of sensitized skin," *Nat. Struct. Mol. Biol.*, vol. 21, no. 2, pp. 189–197, Feb. 2014, doi: 10.1038/nsmb.2756.

[148] T. Giannakouros, E. Nikolakaki, I. Mylonis, and E. Georgatsou, "Serine-arginine protein kinases: A small protein kinase family with a large cellular presence," *FEBS J.*, vol. 278, no. 4, pp. 570–586, Feb. 2011, doi: 10.1111/j.1742-4658.2010.07987.x.

[149] T. Fukuhara *et al.*, "Utilization of host SR protein kinases and RNA-splicing machinery during viral replication," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 30, pp. 11329–11333, Jul. 2006, doi: 10.1073/pnas.0604616103.

[150] Z. Duan *et al.*, "Lentiviral short hairpin RNA screen of genes associated with multidrug resistance identifies PRP-4 as a new regulator of chemoresistance in human ovarian cancer," *Mol. Cancer Ther.*, vol. 7, no. 8, pp. 2377–2385, Aug. 2008, doi: 10.1158/1535-7163.MCT-08-0316.

[151] J. A. Bauer *et al.*, "RNA interference (RNAi) screening approach identifies agents that

enhance paclitaxel activity in breast cancer cells," *Breast Cancer Res.*, vol. 12, no. 3, p. R41, Jun. 2010, doi: 10.1186/bcr2595.

[152] E. Koedoot *et al.*, "Uncovering the signaling landscape controlling breast cancer cell migration identifies novel metastasis driver genes," *Nat. Commun.*, vol. 10, no. 1, pp. 1–16, Jul. 2019, doi: 10.1038/s41467-019-11020-3.

[153] W. Van Roosmalen *et al.*, "Tumor cell migration screen identifies SRPK1 as breast cancer metastasis determinant," *J. Clin. Invest.*, vol. 125, no. 4, pp. 1648–1664, Apr. 2015, doi: 10.1172/JCI74440.

[154] A. S. Harney *et al.*, "Real-time imaging reveals local, transient vascular permeability, and tumor cell intravasation stimulated by TIE2hi macrophage–derived VEGFA," *Cancer Discov.*, vol. 5, no. 9, pp. 932–943, Sep. 2015, doi: 10.1158/2159-8290.CD-15-0012.

[155] E. Beerling *et al.*, "Plasticity between Epithelial and Mesenchymal States Unlinks EMT from Metastasis-Enhancing Stem Cell Capacity," *Cell Rep.*, vol. 14, no. 10, pp. 2281–2288, Mar. 2016, doi: 10.1016/j.celrep.2016.02.034.

[156] Y. Wang *et al.*, "Prpf4 Is Essential for Cell Survival and Posterior Lateral Line Primordium Migration in Zebrafish," *J. Genet. Genomics*, vol. 45, no. 8, pp. 443–453, Aug. 2018, doi: 10.1016/j.jgg.2018.05.008.

[157] M. J. Higgins and J. Baselga, "Targeted therapies for breast cancer," *J. Clin. Invest.*, vol. 121, no. 10, pp. 3797–3803, Oct. 2011, doi: 10.1172/JCI57152.

[158] J. R. Greenwood, D. Calkins, A. P. Sullivan, and J. C. Shelley, "Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution," *J. Comput. Aided. Mol. Des.*, vol. 24, no. 6–7, pp. 591–604, Mar. 2010, doi: 10.1007/s10822-010-9349-1.

[159] J. C. Shelley, A. Cholleti, L. L. Frye, J. R. Greenwood, M. R. Timlin, and M. Uchimaya, "Epik: A software program for pKa prediction and protonation state generation for drug-like molecules," *J. Comput. Aided. Mol. Des.*, vol. 21, no. 12, pp. 681–691, Sep. 2007, doi: 10.1007/s10822-007-9133-z.

[160] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery

and development settings," *Adv. Drug Deliv. Rev.*, vol. 46, no. 1–3, pp. 3–26, Mar. 2001, doi: 10.1016/S0169-409X(00)00129-0.

[161] D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward, and K. D. Kopple, "Molecular properties that influence the oral bioavailability of drug candidates," *J. Med. Chem.*, vol. 45, no. 12, pp. 2615–2623, Jun. 2002, doi: 10.1021/jm020017n.

[162] "Schrödinger Suite 2017-3 Protein Preparation Wizard; Epik, Schrödinger, LLC, New York, NY, 2017; Impact, Schrödinger, LLC, New York, NY, 2017; LigPrep, Schrödinger, LLC, New York, NY, 2017; Prime, Schrödinger, LLC, New York, NY." Schrödinger Release 2017-3, New York.

[163] A. Waterhouse *et al.*, "SWISS-MODEL: Homology modelling of protein structures and complexes," *Nucleic Acids Res.*, vol. 46, no. W1, pp. W296–W303, Jul. 2018, doi: 10.1093/nar/gky427.

[164] G. Madhavi Sastry, M. Adzhigirey, T. Day, R. Annabhimoju, and W. Sherman, "Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments," *J. Comput. Aided. Mol. Des.*, vol. 27, no. 3, pp. 221–234, Mar. 2013, doi: 10.1007/s10822-013-9644-8.

[165] Q. Gao *et al.*, "Evaluation of cancer dependence and druggability of PRP4 kinase using cellular, biochemical, and structural approaches," *J. Biol. Chem.*, vol. 288, no. 42, pp. 30125–30138, 2013, doi: 10.1074/jbc.M113.473348.

[166] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB," *J. Chem. Theory Comput.*, vol. 11, no. 8, pp. 3696–3713, 2015, doi: 10.1021/acs.jctc.5b00255.

[167] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general Amber force field," *J. Comput. Chem.*, vol. 25, no. 9, pp. 1157–1174, Jul. 2004, doi: 10.1002/jcc.20035.

[168] C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman, "A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model," *J. Phys. Chem.*, vol. 97, no. 40, pp. 10269–10280, 1993, doi: 10.1021/j100142a004.

[169] P. W. Abegg and T. K. Ha, “Ab initio calculation of the spin-orbit coupling constant from gaussian lobe SCF molecular wavefunctions,” *Mol. Phys.*, vol. 27, no. 3, pp. 763–767, 1974, doi: 10.1080/00268977400100661.

[170] J. Wang, W. Wang, P. A. Kollman, and D. A. Case, “Automatic atom type and bond type perception in molecular mechanical calculations,” *J. Mol. Graph. Model.*, vol. 25, no. 2, pp. 247–260, Oct. 2006, doi: 10.1016/j.jmgm.2005.12.005.

[171] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, “Comparison of simple potential functions for simulating liquid water,” *J. Chem. Phys.*, vol. 79, no. 2, pp. 926–935, Aug. 1983, doi: 10.1063/1.445869.

[172] A. W. Sousa Da Silva and W. F. Vranken, “ACPYPE - AnteChamber PYthon Parser interfacE,” *BMC Res. Notes*, vol. 5, no. 1, p. 367, Jul. 2012, doi: 10.1186/1756-0500-5-367.

[173] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, “GROMACS: Fast, flexible, and free,” *J. Comput. Chem.*, vol. 26, no. 16, pp. 1701–1718, Dec. 2005, doi: 10.1002/jcc.20291.

[174] T. Darden, D. York, and L. Pedersen, “Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems,” *J. Chem. Phys.*, vol. 98, no. 12, pp. 10089–10092, Jun. 1993, doi: 10.1063/1.464397.

[175] B. R. Miller, T. D. McGee, J. M. Swails, N. Homeyer, H. Gohlke, and A. E. Roitberg, “MMPBSA.py: An efficient program for end-state free energy calculations,” *J. Chem. Theory Comput.*, vol. 8, no. 9, pp. 3314–3321, Sep. 2012, doi: 10.1021/ct300418h.

[176] J. Sgrignani, M. Bon, G. Colombo, and A. Magistrato, “Computational approaches elucidate the allosteric mechanism of human aromatase inhibition: A novel possible route to small-molecule regulation of cyp450s activities?,” *J. Chem. Inf. Model.*, vol. 54, no. 10, pp. 2856–2868, Oct. 2014, doi: 10.1021/ci500425y.

[177] A. Spinello, E. Vecile, A. Abbate, A. Dobrina, and A. Magistrato, “How Can Interleukin-1 Receptor Antagonist Modulate Distinct Cell Death Pathways?,” *J. Chem. Inf. Model.*, vol. 59, no. 1, pp. 351–359, Jan. 2019, doi: 10.1021/acs.jcim.8b00565.

[178] B. Nolen, S. Taylor, and G. Ghosh, “Regulation of protein kinases: Controlling activity through activation segment conformation,” *Mol. Cell*, vol. 15, no. 5, pp. 661–675, Sep.

2004, doi: 10.1016/j.molcel.2004.08.024.

[179] S. S. Taylor and A. P. Kornev, “Protein kinases: Evolution of dynamic regulatory proteins,” *Trends Biochem. Sci.*, vol. 36, no. 2, pp. 65–77, Feb. 2011, doi: 10.1016/j.tibs.2010.09.006.

[180] S. W. Cowan-Jacob, H. Möbitz, and D. Fabbro, “Structural biology contributions to tyrosine kinase drug discovery,” *Curr. Opin. Cell Biol.*, vol. 21, no. 2, pp. 280–287, Apr. 2009, doi: 10.1016/j.ceb.2009.01.012.

[181] S. W. Cowan-Jacob, “Structural biology of protein tyrosine kinases,” *Cell. Mol. Life Sci.*, vol. 63, no. 22, pp. 2608–2625, Nov. 2006, doi: 10.1007/s00018-006-6202-8.

[182] N. Kannan, S. S. Taylor, Y. Zhai, J. C. Venter, and G. Manning, “Structural and functional diversity of the microbial kinome,” *PLoS Biol.*, vol. 5, no. 3, pp. 0467–0478, Mar. 2007, doi: 10.1371/journal.pbio.0050017.

[183] H. Möbitz and D. Fabbro, “Conformational Bias: A key concept for protein kinase inhibition - European Pharmaceutical Review,” *European Pharmaceutlical Review*, 2012. https://www.europeanpharmaceuticalreview.com/article/11289/conformational-bias-a-key-concept-for-protein-kinase-inhibition/ (accessed Oct. 09, 2021).

[184] A. P. Kornev, N. M. Haste, S. S. Taylor, and L. F. Ten Eyck, “Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 47, pp. 17783–17788, Nov. 2006, doi: 10.1073/pnas.0607656103.

[185] W. W. Chan *et al.*, “Conformational control inhibition of the BCR-ABL1 tyrosine kinase, including the gatekeeper T315I mutant, by the switch-control inhibitor DCC-2036,” *Cancer Cell*, vol. 19, no. 4, p. 556, 2011, doi: 10.1016/J.CCR.2011.03.003.

[186] S. Genheden and U. Ryde, “The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities,” *Expert Opin. Drug Discov.*, vol. 10, no. 5, pp. 449–461, May 2015, doi: 10.1517/17460441.2015.1032936.

[187] M. Chorev and L. Carmel, “The function of introns,” *Front. Genet.*, vol. 3, no. APR, p. 55, 2012, doi: 10.3389/fgene.2012.00055.

[188] T. Novoyatleva, Y. Tang, I. Rafalska, and S. Stamm, “Pre-mRNA missplicing as a cause of human disease.,” *Prog. Mol. Subcell. Biol.*, vol. 44, pp. 27–46, 2006, doi:

10.1007/978-3-540-34449-0_2.

[189] A. J. Ward and T. A. Cooper, "The pathobiology of splicing," *J. Pathol.*, vol. 220, no. 2, pp. 152–163, Jan. 2010, doi: 10.1002/path.2649.

[190] P. Papasaikas and J. Valcárcel, "The Spliceosome: The Ultimate RNA Chaperone and Sculptor," *Trends Biochem. Sci.*, vol. 41, no. 1, pp. 33–45, Jan. 2016, doi: 10.1016/j.tibs.2015.11.003.

[191] Y. Shi, "Mechanistic insights into precursor messenger RNA splicing by the spliceosome," *Nat. Rev. Mol. Cell Biol.*, vol. 18, no. 11, pp. 655–670, 2017, doi: 10.1038/nrm.2017.86.

[192] J. L. Jenkins, A. A. Agrawal, A. Gupta, M. R. Green, and C. L. Kielkopf, "U2AF65 adapts to diverse pre-mRNA splice sites through conformational selection of specific and promiscuous RNA recognition motifs," *Nucleic Acids Res.*, vol. 41, no. 6, pp. 3859–3873, 2013, doi: 10.1093/nar/gkt046.

[193] M. Montes, B. L. Sanford, D. F. Comiskey, and D. S. Chandler, "RNA Splicing and Disease: Animal Models to Therapies," *Trends Genet.*, vol. 35, no. 1, pp. 68–87, Jan. 2019, doi: 10.1016/j.tig.2018.10.002.

[194] A. M. Fredericks, K. J. Cygan, B. A. Brown, and W. G. Fairbrother, "RNA-Binding Proteins: Splicing Factors and Disease," *Biomolecules*, vol. 5, no. 2, p. 893, May 2015, doi: 10.3390/BIOM5020893.

[195] O. A. Kent and A. M. MacMillan, "Early organization of pre-mRNA during spliceosome assembly," *Nat. Struct. Biol.*, vol. 9, no. 8, pp. 576–581, 2002, doi: 10.1038/nsb822.

[196] E. A. Sickmier, K. E. Frato, H. Shen, S. R. Paranawithana, M. R. Green, and C. L. Kielkopf, "Structural Basis for Polypyrimidine Tract Recognition by the Essential Pre-mRNA Splicing Factor U2AF65," *Mol. Cell*, vol. 23, no. 1, pp. 49–59, 2006, doi: 10.1016/j.molcel.2006.05.025.

[197] A. A. Agrawal *et al.*, "An extended U2AF65-RNA-binding domain recognizes the 3′ splice site signal," *Nat. Commun.*, vol. 7, 2016, doi: 10.1038/ncomms10950.

[198] L. V. Von Voithenberg *et al.*, "Recognition of the 3′ splice site RNA by the U2AF heterodimer involves a dynamic population shift," *Proc. Natl. Acad. Sci. U. S. A.*, vol.

113, no. 46, pp. E7169–E7175, 2016, doi: 10.1073/pnas.1605873113.

[199] C. D. MacKereth *et al.*, "Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF," *Nature*, vol. 475, no. 7356, pp. 408–413, Jul. 2011, doi: 10.1038/nature10171.

[200] P. Senapathy, M. B. Shapiro, and N. L. Harris, "Splice junctions, branch point sites, and exons: Sequence statistics, identification, and applications to genome project," *Methods Enzymol.*, vol. 183, no. C, pp. 252–278, Jan. 1990, doi: 10.1016/0076-6879(90)83018-5.

[201] R. Singh, J. Valcárcel, and M. R. Green, "Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins," *Science (80-.).*, vol. 268, no. 5214, pp. 1173–1176, 1995, doi: 10.1126/science.7761834.

[202] J. Bouck, S. Litwin, A. M. Skalka, and R. A. Katz, "In vivo selection for intronic splicing signals from a randomized pool," *Nucleic Acids Res.*, vol. 26, no. 19, pp. 4516–4523, Oct. 1998, doi: 10.1093/nar/26.19.4516.

[203] R. F. Roscigno, M. Weiner, and M. A. Garcia-Blanco, "A mutational analysis of the polypyrimidine tract of introns. Effects of sequence differences in pyrimidine tracts on splicing," *J. Biol. Chem.*, vol. 268, no. 15, pp. 11222–11229, May 1993, doi: 10.1016/s0021-9258(18)82114-7.

[204] Z. Krchňáková *et al.*, "Splicing of long non-coding RNAs primarily depends on polypyrimidine tract and 5′ splice-site sequences due to weak interactions with SR proteins," *Nucleic Acids Res.*, vol. 47, no. 2, pp. 911–928, 2019, doi: 10.1093/nar/gky1147.

[205] S. Cho *et al.*, "Splicing inhibition of U2AF65 leads to alternative exon skipping," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 32, pp. 9926–9931, 2015, doi: 10.1073/pnas.1500639112.

[206] M. M. De Araújo, S. Bonnal, M. L. Hastings, A. R. Krainer, and J. Valcárcel, "Differential 3′ splice site recognition of SMN1 and SMN2 transcripts by U2AF and U2 snRNP," *Rna*, vol. 15, no. 4, pp. 515–523, 2009, doi: 10.1261/rna.1273209.

[207] A. A. Agrawal, K. J. McLaughlin, J. L. Jenkins, and C. L. Kielkopf, "Structure-guided U2AF65 variant improves recognition and splicing of a defective pre-mRNA," *Proc.*

Natl. Acad. Sci. U. S. A., vol. 111, no. 49, pp. 17420–17425, 2014, doi: 10.1073/pnas.1412743111.

[208] G.-B. MA, B. AP, and L. EL, "Alternative splicing in disease and therapy," *Nat. Biotechnol.*, vol. 22, no. 5, pp. 535–546, May 2004, doi: 10.1038/NBT964.

[209] G. Golling *et al.*, "Insertional mutagenesis in zebrafish rapidly identifies genes essential for early vertebrate development," *Nat. Genet.*, vol. 31, no. 2, pp. 135–140, 2002, doi: 10.1038/ng896.

[210] M. H. M. Olsson, C. R. SØndergaard, M. Rostkowski, and J. H. Jensen, "PROPKA3: Consistent treatment of internal and surface residues in empirical p K a predictions," *J. Chem. Theory Comput.*, vol. 7, no. 2, pp. 525–537, Feb. 2011, doi: 10.1021/ct100578z.

[211] D. A. Case *et al.*, "AMBER 2018, University of California, San Francisco."

[212] I. S. Joung and T. E. Cheatham, "Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations," *J. Phys. Chem. B*, vol. 112, no. 30, pp. 9020–9041, Jul. 2008, doi: 10.1021/jp8001614.

[213] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, "LINCS: A Linear Constraint Solver for molecular simulations," *J. Comput. Chem.*, vol. 18, no. 12, pp. 1463–1472, Sep. 1997, doi: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

[214] M. Bonomi *et al.*, "PLUMED: A portable plugin for free-energy calculations with molecular dynamics," *Comput. Phys. Commun.*, vol. 180, no. 10, pp. 1961–1972, Oct. 2009, doi: 10.1016/J.CPC.2009.05.011.

[215] T. Yang *et al.*, "Virtual screening using molecular simulations," *Proteins Struct. Funct. Bioinforma.*, vol. 79, no. 6, pp. 1940–1951, Jun. 2011, doi: 10.1002/prot.23018.

[216] D. Maji *et al.*, "Representative cancer-associated U2AF2 mutations alter RNA interactions and splicing," *J. Biol. Chem.*, vol. 295, no. 50, pp. 17148–17157, Dec. 2020, doi: 10.1074/jbc.RA120.015339.

[217] M. Imielinski *et al.*, "Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing," *Cell*, vol. 150, no. 6, pp. 1107–1120, Sep. 2012, doi: 10.1016/j.cell.2012.08.029.

[218] V. A. Feher, J. D. Durrant, A. T. Van Wart, and R. E. Amaro, "Computational

approaches to mapping allosteric pathways," *Current Opinion in Structural Biology*, vol. 25. Elsevier Ltd, pp. 98–103, 2014, doi: 10.1016/j.sbi.2014.02.004.

[219] A. Spinello *et al.*, "Rational design of allosteric modulators of the aromatase enzyme: An unprecedented therapeutic strategy to fight breast cancer," *Eur. J. Med. Chem.*, vol. 168, pp. 253–262, 2019, doi: 10.1016/j.ejmech.2019.02.045.

[220] J. R. Wagner, C. T. Lee, J. D. Durrant, R. D. Malmstrom, V. A. Feher, and R. E. Amaro, "Emerging Computational Methods for the Rational Discovery of Allosteric Drugs," *Chemical Reviews*, vol. 116, no. 11. American Chemical Society, pp. 6370–6390, Jun. 08, 2016, doi: 10.1021/acs.chemrev.5b00631.

[221] A. Gheeraert, L. Pacini, V. S. Batista, L. Vuillon, C. Lesieur, and I. Rivalta, "Exploring Allosteric Pathways of a V-Type Enzyme with Dynamical Perturbation Networks," *J. Phys. Chem. B*, vol. 123, no. 16, pp. 3452–3461, Apr. 2019, doi: 10.1021/acs.jpcb.9b01294.

[222] L. Casalino, G. Palermo, U. Rothlisberger, and A. Magistrato, "Who Activates the Nucleophile in Ribozyme Catalysis? An Answer from the Splicing Mechanism of Group II Introns," *J. Am. Chem. Soc.*, vol. 138, no. 33, pp. 10374–10377, 2016, doi: 10.1021/jacs.6b01363.

[223] B. Kastner, C. L. Will, H. Stark, and R. Lührmann, "Structural insights into nuclear pre-mRNA splicing in higher eukaryotes," *Cold Spring Harb. Perspect. Biol.*, vol. 11, no. 11, 2019, doi: 10.1101/cshperspect.a032417.

[224] L. Casalino, G. Palermo, A. Spinello, U. Rothlisberger, and A. Magistrato, "All-atom simulations disentangle the functional dynamics underlying gene maturation in the intron lariat spliceosome," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 115, no. 26, pp. 6584–6589, 2018, doi: 10.1073/pnas.1802963115.

[225] J. Borišek *et al.*, "Disclosing the impact of carcinogenic SF3b mutations on pre-mRNA recognition via all-atom simulations," *Biomolecules*, vol. 9, no. 10, Oct. 2019, doi: 10.3390/biom9100633.

[226] J. Borišek, A. Saltalamacchia, A. Spinello, and A. Magistrato, "Exploiting Cryo-EM Structural Information and All-Atom Simulations to Decrypt the Molecular Mechanism of Splicing Modulators," *J. Chem. Inf. Model.*, vol. 60, no. 5, pp. 2510–

2521, Oct. 2020, doi: 10.1021/acs.jcim.9b00635.

[227] L. Casalino and A. Magistrato, "Unraveling the Molecular Mechanism of Pre-mRNA Splicing From Multi-Scale Simulations," *Front. Mol. Biosci.*, vol. 6, Aug. 2019, doi: 10.3389/fmolb.2019.00062.

[228] G. Palermo, L. Casalino, A. Magistrato, and J. Andrew McCammon, "Understanding the mechanistic basis of non-coding RNA through molecular dynamics simulations," *J. Struct. Biol.*, vol. 206, no. 3, pp. 267–279, 2019, doi: 10.1016/j.jsb.2019.03.004.

[229] C. Yan, R. Wan, R. Bai, G. Huang, and Y. Shi, "Structure of a yeast step II catalytically activated spliceosome," *Science (80-.).*, vol. 355, no. 6321, pp. 149–155, 2017, doi: 10.1126/science.aak9979.

[230] S. M. Fica *et al.*, "Structure of a spliceosome remodelled for exon ligation," *Nature*, vol. 542, no. 7641, pp. 377–380, 2017, doi: 10.1038/nature21078.

[231] R. Wan, C. Yan, R. Bai, G. Huang, and Y. Shi, "Structure of a yeast catalytic step i spliceosome at 3.4 Å resolution," *Science (80-.).*, vol. 353, no. 6302, pp. 895–904, Aug. 2016, doi: 10.1126/science.aag2235.

[232] A. Šali and T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints," *Journal of Molecular Biology*, vol. 234, no. 3. pp. 779–815, 1993, doi: 10.1006/jmbi.1993.1626.

[233] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, "GROMACS: Fast, flexible, and free," *Journal of Computational Chemistry*, vol. 26, no. 16. pp. 1701–1718, Dec. 2005, doi: 10.1002/jcc.20291.

[234] M. Zgarbová *et al.*, "Refinement of the Cornell et al. Nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles," *J. Chem. Theory Comput.*, vol. 7, no. 9, pp. 2886–2902, 2011, doi: 10.1021/ct200162x.

[235] A. Pérez *et al.*, "Refinement of the AMBER force field for nucleic acids: Improving the description of α/γ conformers," *Biophys. J.*, vol. 92, no. 11, pp. 3817–3829, 2007, doi: 10.1529/biophysj.106.097782.

[236] J. Šponer *et al.*, "How to understand atomistic molecular dynamics simulations of RNA and protein–RNA complexes?," *Wiley Interdiscip. Rev. RNA*, vol. 8, no. 3, pp. 1–17, 2017, doi: 10.1002/wrna.1405.

[237] A. Saxena and D. Sept, “Multisite ion models that improve coordination and free energy calculations in molecular dynamics simulations,” *J. Chem. Theory Comput.*, vol. 9, no. 8, pp. 3538–3542, 2013, doi: 10.1021/ct400177g.

[238] L. Casalino, G. Palermo, N. Abdurakhmonova, U. Rothlisberger, and A. Magistrato, “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.*, vol. 13, no. 1, pp. 340–352, 2017, doi: 10.1021/acs.jctc.6b00905.

[239] Y.-P. Pang, “Novel Zinc Protein Molecular Dynamics Simulations: Steps Toward Antiangiogenesis for Cancer Treatment,” *J. Mol. Model.*, vol. 5, no. 10, pp. 196–202, Oct. 1999, doi: 10.1007/s008940050119.

[240] A. W. Sousa Da Silva and W. F. Vranken, “ACPYPE - AnteChamber PYthon Parser interfacE,” *BMC Res. Notes*, vol. 5, no. 1, pp. 1–8, Jul. 2012, doi: 10.1186/1756-0500-5-367.

[241] D. Case *et al.*, “AMBER 2016, University of California, San Francisco,” 2016.

[242] B. Hess, “Convergence of sampling in protein simulations,” *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.*, vol. 65, no. 3, pp. 1–10, 2002, doi: 10.1103/PhysRevE.65.031910.

[243] B. Hess, “Similarities between principal components of protein dynamics and random diffusion,” *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.*, vol. 62, no. 6 B, pp. 8438–8448, 2000, doi: 10.1103/PhysRevE.62.8438.

[244] R. Cossio-Pérez, J. Palma, and G. Pierdominici-Sottile, “Consistent Principal Component Modes from Molecular Dynamics Simulations of Proteins,” *J. Chem. Inf. Model.*, vol. 57, no. 4, pp. 826–834, 2017, doi: 10.1021/acs.jcim.6b00646.

[245] G. Pierdominici-Sottile and J. Palma, “New insights into the meaning and usefulness of principal component analysis of concatenated trajectories,” *J. Comput. Chem.*, vol. 36, no. 7, pp. 424–432, 2015, doi: 10.1002/jcc.23811.

[246] W. Humphrey, A. Dalke, and K. Schulten, “VMD: Visual molecular dynamics,” *J. Mol. Graph.*, vol. 14, no. 1, pp. 33–38, 1996, doi: 10.1016/0263-7855(96)00018-5.

[247] M. Ltzelberger and N. F., “The Prp4 Kinase: Its Substrates, Function and Regulation in

Pre-mRNA Splicing," in *Protein Phosphorylation in Human Health*, InTech, 2012.

[248] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon, "Electrostatics of nanosystems: Application to microtubules and the ribosome," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 18, pp. 10037–10041, Aug. 2001, doi: 10.1073/pnas.181342398.

[249] A. Saltalamacchia, L. Casalino, J. Borišek, V. S. Batista, I. Rivalta, and A. Magistrato, "Decrypting the Information Exchange Pathways across the Spliceosome Machinery," *J. Am. Chem. Soc.*, vol. 142, no. 18, pp. 8403–8411, May 2020, doi: 10.1021/jacs.0c02036.

[250] M. Pavlin *et al.*, "A Computational Assay of Estrogen Receptor α Antagonists Reveals the Key Common Structural Traits of Drugs Effectively Fighting Refractory Breast Cancers," *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, doi: 10.1038/s41598-017-17364-4.

[251] C. G. Ricci, R. L. Silveira, I. Rivalta, V. S. Batista, and M. S. Skaf, "Allosteric Pathways in the PPARγ 3-RXRα nuclear receptor complex," *Sci. Rep.*, vol. 6, Jan. 2016, doi: 10.1038/srep19940.

[252] K. Yang, L. Zhang, T. Xu, A. Heroux, and R. Zhao, "Crystal structure of the β-finger domain of Prp8 reveals analogy to ribosomal proteins," 2008. doi: 10.1073/pnas.0805960105.