SISSA
Scuola
Internazionale
Superiore di
Studi Avanzati

Physics Area — PhD course in
Molecular and Statistical Biophysics

Asymmetric Response of Selected RNAs to Forced Nanopore Translocation

Candidate:
Matteo TAJANA

Advisor:
Prof. Cristian MICHELETTI

Academic Year 2025-26



Commission of the Final Examination:

External Member:
Prof. Enzo ORLANDINI (University of Padua)

External Member:
Prof. Riccardo CAPELLI (University of Milan)

Final Examination:

September 23rd, 2026
SISSA, Trieste (TS), Italy

Questa (forse piccola) Tesi è dedicata tutti coloro che, come me, hanno pensato di non farcela.
Ma soprattutto, è dedicata a chi non ce l'ha fatta: a Sam.

Contents

1 Overview and Motivation **1**
- 1.1 RNA: the Factotum of Life . . . 1
- 1.2 Translocation of Polymers . . . 1
  - 1.2.1 Experimental Works . . . 2
  - 1.2.2 Theoretical Studies and MD Simulations . . . 3
- 1.3 Viruses and Viral RNA . . . 3
  - 1.3.1 xrRNAs . . . 4
- 1.4 Motivation . . . 5

2 Selection of RNAs of Viral Origin **7**
- 2.1 (Non-redundant) Databases of RNAs . . . 7
- 2.2 Construction of the List of Viral RNAs . . . 8
  - 2.2.1 Viral RNAs . . . 9
  - 2.2.2 Gapless RNAs . . . 10
  - 2.2.3 Compact RNAs . . . 10
  - 2.2.4 Structured RNAs . . . 12
  - 2.2.5 Removal of RNAs with Intercalated Ligands . . . 13
  - 2.2.6 Actually Non-Redundant RNAs . . . 13
  - 2.2.7 RNAs with Exposed Ends . . . 16
  - 2.2.8 Grouping RNAs by Topological Complexity . . . 18
- 2.3 The Final List . . . 22
  - 2.3.1 A Note on IRESs and xrRNAs . . . 26
- 2.4 Summary . . . 29

3 Simulations Setup **31**
- 3.1 Force Field and Software . . . 31
  - 3.1.1 The SMOG Force Field . . . 31
  - 3.1.2 LAMMPS . . . 33
- 3.2 Priming the RNAs . . . 35
- 3.3 The Temperature . . . 35
  - 3.3.1 Preparation . . . 35
  - 3.3.2 Temperature Calibration . . . 39
- 3.4 The Nanopore . . . 40
  - 3.4.1 Size Calibration . . . 41
- 3.5 Initialization Protocol . . . 44
- 3.6 The Force . . . 44
  - 3.6.1 Setup Details . . . 44
  - 3.6.2 Force Calibration . . . 47
- 3.7 Observables . . . 47

3.8 Summary 47

4 Activation Times Asymmetry **49**
4.1 Static Force Protocol 49
4.1.1 Translocation Traces 49
4.1.2 Activation Times 52
4.2 Ramping Force Protocol 54
4.2.1 Translocation Traces 54
4.2.2 Activation Times 56
4.3 Correlations Between Static and Ramping Force Protocols 57
4.4 Framing The Activation Times 60
4.4.1 Static Force Protocol 60
4.4.2 Ramping force protocol 61
4.5 Summary 61

5 Simulations Analyses **63**
5.1 Contacts 63
5.1.1 Primary Contacts 64
5.2 Structural Description of the “Hard Cases” 68
5.2.1 Lasso and Resisting End 69
5.2.2 xrRNAs 71
5.2.3 The last RNA 73
5.3 Summary 74

Thesis Motivation

Polymer translocation is a ubiquitous phenomenon in biology. From protein synthesis to cellular transport and viral infections, several biological processes involve the passage of a polymer (be it DNA, RNA or proteins) through some kind of narrow constriction. The *"structure makes function"* paradigm of molecular biology extends naturally into this dynamic setting: not only the fold of a molecule determines the biological task it can perform, but when the constriction is narrow enough, that same fold shapes how the molecule translocates through it. An untangled rope and a tangled ball of yarn each pass through the eye of a needle differently, and polymer chains are no exception to this. For instance, messenger RNA (mRNA) is designed by evolution to be scanned by ribosomes, and its structure accounts for this; other RNA molecules like pseudoknotted domains present a fold selected to halt the translocation at precise spots of the genome during the *programmed ribosomal frameshifting*, or xrRNA are designed to withstand the translocation in a specific direction. In other words, the structure of a molecule actively drives the dynamics of its translocation through a constriction.

Over the past few years, polymer translocation through nanopores has successfully emerged as a fruitful area of scientific investigation, both under the experimental and the theoretical point of view. Initially designed as a plausible sequencing tool, recent advancement in the fields lead nanopore translocation to even probe the secondary and tertiary structures of molecules. The translocation of structured RNA molecules is not straightforward, given the heterogeneous set of nucleotides and interactions which dictate the fold at equilibrium and also their response to the forced translocation. Viral genomes offer a particularly rich setting in which to study this heterogeneity, as they fold into an enormous diversity of secondary and tertiary structures packed into genomes of widely different lengths. This diversity makes viral RNA an ideal substrate for inspecting the relation between structure and translocation dynamics.

In this Thesis, we will perform a systematic study on the different responses of the two ends of a set of selected RNAs of viral origin encompassing a heterogeneous ensemble of lengths and conformations.

Chapter 1

Overview and Motivation

1.1 RNA: the Factotum of Life

RNA is a multipotent molecule which is essential in life as it plays a plethora of roles in cells. Longtime thought only as information carrier from DNA to the ribosome between transcription and translation of the genome [1], RNA is an extremely versatile molecule. To name a few examples, *messenger RNA* (mRNA) was discovered in the 1960s as the product of the RNA polymerase from a DNA template during the transcription of a gene that must be expressed to make proteins [2, 3]. The synthesis of proteins is carried out by *ribosomes*, molecular complexes made of proteins and RNA [4]. The single amino acids are delivered to the ribosome and then covalently bonded to the growing polypeptide by means of *transfer RNA* (tRNA). Each tRNA is loaded with its associated amino acid and recognizes the corresponding triplet (codon) on the mRNA by pairing its anticodon to the corresponding three-nucleotide sequence of the mRNA [5]. tRNA molecules work synergistically with the ribosome, which temporary aligns tRNA molecules on the mRNA template in order to grow the forming peptide chain: the transfer of amino acids from tRNA units to the growing chain is catalyzed by *ribosomal RNA* (rRNA), which is one of the primary components of ribosomes [6]. The catalytic nature of some RNA molecules is even more accentuated in *ribozymes*: discovered at the beginning of the 1980s, ribozymes are catalyzers for some biochemical reactions like the self-excision of introns [7], that is the removal of portions of genes which are not expressed into proteins. The gene expression is controlled also by *riboswitches*, molecular structures formed within the mRNA upon binding small ligands [8]. *Micro RNA* (miRNA) and *small interfering RNA* (siRNA) are two related classes of small regulatory RNAs that repress gene expression by interacting directly with the mRNA and inducing its cleavage. Both molecules consist of short duplex RNA molecules of approximately 20 nucleotides per strand that exert gene silencing by targeting and binding mRNA in different ways. siRNA is highly specific with precise, single mRNA targets [9], while miRNA typically does not pair up perfectly with the multiple targets it has [10, 11]. Either way, the mRNA is recognized as abnormal after binding with siRNA/miRNA and the gets degraded, repressing the translation [12].

1.2 Translocation of Polymers

RNA is a nucleic acid that plays a central role in gene expression and protein synthesis. Typically, RNA is actively threaded through processive enzymes. For example, mRNA is scanned by ribosomes during the translation of RNA [13], or can translocate among cells through nuclear pores [14]. Similarly, the genomes of several RNA viruses like the SARS-CoV-2 virus are duplicated by means of RNA-dependent RNA polymerases, which scan the RNAs of these viruses

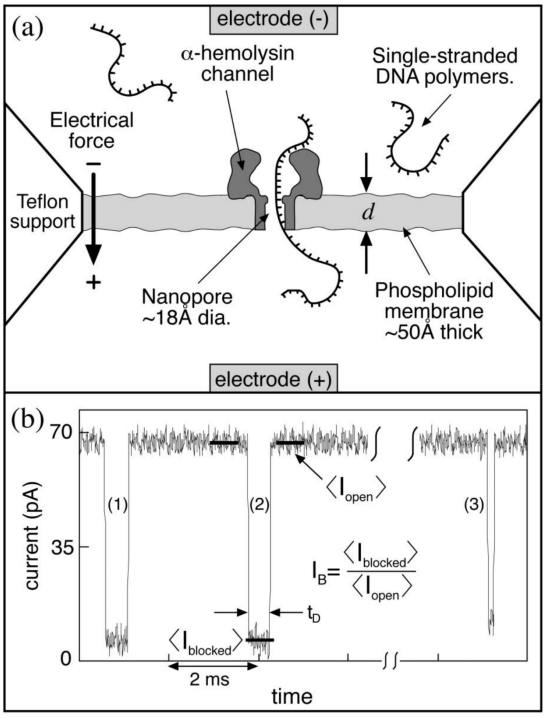


Figure 1.1 In panel (a) above, the schematics of a typical biological nanopore translocation setup involving the α-HL channel; below in panel (b), the current blockades recorded as the translocations of different polymers occur. Image taken from Ref. [17].

catalyzing the synthesis of the complementary strand [15].

Regardless of their nature or biological role of the translocating chain, all forms of translocation through a narrow constriction can be conveniently modeled as the passage of the reference molecule through a cylindrical pore of proper dimensions. Experimental and theoretical works on nanopore translocations of polymers and biopolymers started in the 1990s [16] and their applications went far beyond the pure mimicry and study of *in vivo* translocation, showing how translocation is a versatile and productive area of scientific investigation.

1.2.1 Experimental Works

The first idea of an experimental application of nanopore translocation sprouted in the 1980s, with the aim of sequencing long strands of nucleic acids [16]. The idea is straightforward: a biological pore embedded in a lipid bilayer separates two chambers of the setup, one of which contains the polymer mixture in solution. As a voltage is applied across the membrane, ions and polymers move through the nanopore. As the ions transverse the channel and exit the aperture, a ionic current is recorded externally. Given the narrow constriction of the pore, the blockage of such ionic current is associated to the passage of a strand of nucleotides, see Fig. 1.1.

The first work with this framework was carried out by Kasianowich and coworkers in their seminal work [18] where poly-U chains were shown to give rise to current blockades due to the obstruction to the passage of ions through the pore. By the end of the 1990s, it was possible to discriminate poly-C from poly-A strands [19], and in 2005 single-nucleotide mutations in the translocating strand were identified accurately [20]. With the advancement of technology and the integration of state-of-the-art algorithms for the analysis of the high-throughput generated by nanopore translocation, in the past 10 years it has been possible to reach accuracies well above 90% of the high-throughput [21].

The first experimental works used biological pores like α-hemolysin (α-HL) or *Mycobacterium smegmatis* porin A (MspA) for the translocation of polymers, since they have precise geometries and can be appropriately modified through genetic modifications for specific cases. However, their constriction is fixed and does not exceed 20 Å in diameter, plus their stability depends on the temperature and pH of the experimental setup. Solid state nanopores were developed as a solid alternative to biological pores, allowing the control over the diameter and length of the channel, with the possibility of tuning specific surface properties and potential integrations into devices or arrays of pores [22]. As the experimental setups improved over time, nanopore translocations showed a wider range of applications not expected before, from measuring the translocation velocity of DNA strands [17], to recognizing and discriminating the translocating end of the chains [23], to unzipping hairpin molecules during the translocation [24], to quantification of mRNA and tRNA [25, 26]. Polymer translocation experiments were also successfully able to identify RNA isoforms and conformations [27–30], hinting how this kind of experimental setup can be used to probe the three-dimensional structure of molecules, a notoriously daunting task in biology. Finally, the aforementioned dreams about sequencing long transcripts of nucleic acids are becoming a reality, providing a quick and cheap tool for nucleic acid sequencing [31–33].

1.2.2 Theoretical Studies and MD Simulations

Alongside with experimental advances, theory and molecular dynamics (MD) simulations integrated the endeavors of experimental research on nanopore translocation.

Theoretical studies on the translocation of polymers through a nanopore showed how nontrivial scaling exponents emerge between the translocation time of a single-stranded homopolymer chain through a pore and the chain length [34–36]. In Ref. [37] was developed a theoretical framework which keeps into account the relative orientation of the nucleotides translocating through the pore to extract the effective mobility of a translocating chain, showing a nontrivial interplay between chain length, driving force and friction.

MD simulations of nanopore translocations have become a fundamental tool to integrate both the experimental and the theoretical investigations, making it possible the observation of the various stages of the RNA translocation with atomistic detail as a powerful microscope. The process was shown to be non-universal and strongly dependent on the assumptions and specifications made on the observed model [38]. For example, different regimes were observed for the distribution of blobs in a pulled polymer [39], and the propagation of tension was shown to play a crucial role in the time each residue spends at the pore entrance [40]. MD simulations were successfully employed to observe the different conformations assumed by an RNA hairpin when squeezed through different pores and to map them to the recorded ion current [41], or to see the effect of knots on the translocation properties of polymers [42–44]. Finally, simulations were also applied on the translocation involving RNA of biological systems: for instance, it was shown in Refs. [45–47] the mechanism responsible for the resistance to translocation of a portion of the genome of Zika virus (ZIKV) in a specific direction, and in Ref. [41] it was shown how current signals can be mapped to different conformations of an RNA strand, discriminating pseudoknot from melting and molten globule.

1.3 Viruses and Viral RNA

Viruses are biological infectious parasites which replicate only inside the living cells of an organism after the transfer of genetic material [48, 49]. The largest class of viruses has a single-stranded RNA genome [50]. Most viruses are rather compact in size, ranging from 20 nm in diameter up to few hundreds nanometers [51]. Evolution designed optimally the viral genomes

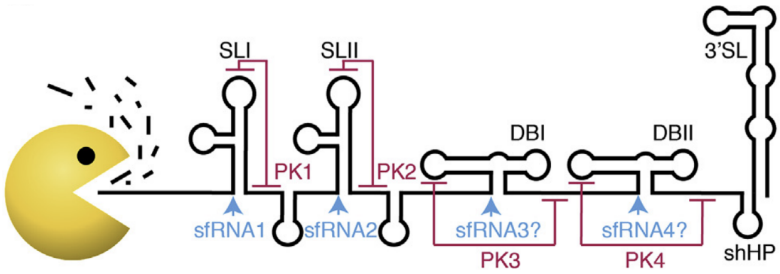


Figure 1.2 Schematics of the sfRNA biogenesis via exoribonuclease partial digestion of the DENV genome. As the exoribonuclease (pictured as the yellow "Pac-Man" cartoon) can halt in correspondence of specific sites of the viral genome (stem-loops SLI-II, or dumb bells DBI-II), giving birth to different sfRNAs. Image taken from Fig. 1(b) of Ref. [59].

to fit within such small-sized capsids, comprising both regulatory sequences and protein coding regions. The wide variety of functions that the viral genome is able to perform often culminates in topologically fascinating genomes decorated with pseudoknots, bulges, circular motifs and multiple-strand junctions, making RNA molecules of viruses a stimulating area of research [52, 53].

As aforementioned, viruses cannot replicate by themselves but rely on the replication machinery of the hijacked cells in order to produce copies of the original virus. The cellular translation if often manipulated by viruses using in many cases elegant RNA-centered strategies: some viruses bypass the need for initiation signals formed in mRNAs, others recruit the translation machinery of the hijacked cell by means of specific RNA structures (like the *internal ribosome entry site*, IRES) [53, 54], others assume tRNA-like structures to engage the ribosome for the synthesis of viral proteins [55].

Even more interestingly, in order to optimize the genome size inside the capsid, some virus genomes induce *programmed ribosomal frameshifting*: as the name suggests, the process consists of possible changes of the reading frame the ribosome is scanning during the synthesis of viral proteins. The presence of multiple reading frames within the genome of viruses effectively increases the size of the coding regions within their genome to double or triple the initial length [56–58], maximizing the information contained within the limited length of the packaged RNA inside the capsids. Frameshifting occurs as the result of interactions with pseudoknots or stem-loops in correspondence of specific slippery sites, even though the exact mechanism behind is still a matter of debate [54].

Among viruses, the family of *Flaviviruses* is of particular interest. Flaviviruses form a family of enveloped positive-strand RNA viruses which are primarily spread mostly through ticks and mosquitoes [59]. Flaviviral infections occur by accumulating noncoding RNA (*subgenomic flaviviral RNA*, sfRNA) within the host cell by means of partial degradation of viral genomic RNA by *exoribonucleases*, enzymes recruited by the immune system of the host cell to degrade RNA by processing it in the $5' \to 3'$ direction, see Fig. 1.2. Interestingly, sfRNAs cannot be degraded by nucleases, still they can be processed by polymerases and replicases, contributing to the spread of several diseases. The resistance to degradation is traced back to specific modules of the viral genome called *exoribonuclease resistant RNAs* (xrRNAs), intricate folds decorated with multiple pseudoknots located in series in the $3'$ untraslated region ($3'$-UTR) of the genome [60].

1.3.1 xrRNAs

xrRNAs are structured RNA modules located at the $3'$ untranslated region (UTR) in the genome of different viruses from the *Flaviviridae* family (notable examples containing such modules are

the *Zika virus* ZIKV, the *Dengue virus* DENV, the *Murray Valley encephalitis virus* MVEV, the *Tamana Bat virus* TABV, and the *Powassan virus* POWV), which confer an exceptional resistance to the RNA degradation by means of exoribonucleases [61–65]. Such resistance is general, that is a single xrRNA module is able to halt diverse nucleases, revealing a non-specific response to degradation that must be traced back solely to the structure of this module [66].

A fascinating characteristic of the architecture of xrRNAs is the conserved ring-like structure that encircles the 5′-end with a single [62–64] or double loop [65]. Interestingly, this topological feature of xrRNAs was observed also in plant-infecting viruses, which are phylogenetically unrelated to flaviviruses [67]. The circular motif seems to acts as a molecular brace towards the translocation in the $5' \to 3'$ direction.

Two recent works by our group explained the rationale behind the mechanical resistance of the xrRNA of the ZIKV. In these studies, the ZIKV xrRNA was translocated through a pore whose geometry resembles the lumen of the exoribonuclease XRN-1. When the molecule was translocated in the $3' \to 5'$ direction (as done by replicases), the molecule straightforwardly unzips the terminal helix and the translocation progresses; however, pulling in the $5' \to 3'$ direction (as done by the degrading exonucleases), the 5′-end runs into a counterintuitive molecular tightening which favors the redistribution of the stress generated by the pulling force over an intricate interaction network of bonded nucleotides, delaying the translocation by orders of magnitude [45, 47]. In particular, it was shown in Ref. [47] how the activation of the translocation in the $5' \to 3'$ direction is delayed with respect to that in the $3' \to 5'$ one over a broad range of driving forces. Furthermore, the activation times of the two ends can differ by several orders of magnitude as the driving force decreases.

1.4 Motivation

The results showed in the previous works from our group on the ZIKV xrRNAs raise the question of whether such directional resistance is a peculiarity of this specific class of resisting RNAs only, or if the asymmetry can be extended to other RNA folds. The heterogeneous set of structures reported in the genomes of viruses, often comprising pseudoknotted domains, bulged stems, and multiple-stranded junctions which increase the topological complexity of the strands, provides a fertile ground base for the study of the effect of a polymer structure on its translocation through a narrow constriction. To what extent is the polymer translocation affected by the structure of the molecules? Do molecules with similar secondary and/or tertiary structures behave similarly during the translocation? What is the effect of an increasing structural complexity on the activation of such process? To our knowledge, a systematic study on the response to translocation of viral RNA through a nanopore in general is still missing. The goal of this Thesis is thus to perform a comparative study of the response to the forced translocation of RNA molecule through a nanopore, selecting the nucleic acids belonging to viral genomes.

The Thesis is organized as follows. Chapter 2 describes the construction of a non-redundant list of 16 RNA structures of viral origin employed throughout this work, obtained from a large public database through a sequence of taxonomic, structural, and topological filtering criteria. In Chapter 3 we will present the details of the simulation setup, including the structure-based force field, the calibration of the simulation temperature, the nanopore geometry, the pulling force protocols, and the priming procedure used to thread each RNA into the pore. Chapter 4 presents the central result of this Thesis, namely a systematic asymmetry between the activation of the $5' \to 3'$ and $3' \to 5'$ translocation directions, observed consistently across all selected RNAs, independently of the pulling protocol employed. Finally, in Chapter 5 we investigate the structural origin of this asymmetry. We find that the density of resisting contacts at the chain termini correlates only mildly with the observed asymmetry in the activation, indicating that

this energetic imbalance bias is not sufficient *per se* to account for the systematic asymmetry described in the preceding Chapter. We report that the ZIKV xrRNA exerts the maximal resistance to the forced translocation in the $5' \rightarrow 3'$ direction, corroborating previous results obtained in Ref. [45, 47]. Focusing on the full subset of RNAs that failed the activation from the $5'$-end under static force within the allocated computational resources, we observe that all of these molecules present pseudoknotted folds at the $5'$-terminal, and that some of them share specific topological features which we describe. The structural description of these termini can be used as a starting point for future works concerning the mechanism behind the strong directional resistance, in the spirit of Ref. [45].

Chapter 2

Selection of RNAs of Viral Origin

2.1 (Non-redundant) Databases of RNAs

The Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) is currently the largest accessible collection of biomolecules, consisting of almost 250 000 publicly available structures of proteins, nucleic acids, and complex assemblies. Such database is constantly fueled with thousands of biomolecular structures that are deposited every year, see `https://www.rcsb.org/stats/growth/growth-released-structures`. The vast majority of the newly added structure consists of proteins or complexes between amino acids and proteins, as only around few hundreds RNA molecules have been uploaded every year on the RCSB PDB over the last decade, see `https://www.rcsb.org/stats/growth/growth-rna`. As of 2026, among the around 250 000 entries of the RCSB PDB, only approximately 2000 entries consist of PDB files containing only RNA molecules, accounting for less than 1% of the total number of structures available.

This large discrepancy between proteins and nucleic acids deposited is due to the intrinsic difficulty in obtaining experimentally the structure of RNA molecules. In fact, RNA molecules are particularly flexible and can form several tertiary contacts, making their crystallization particularly challenging [68–70]. The advent of technologies like Nuclear Magnetic Resonance (NMR) [71–73] and Cryo-Electron Microscopy (Cryo-EM) [74, 75] has allowed structural biologists to solve an increasing number of RNA structures, contributing to the number of deposited structures.

The ongoing determination of new structures inevitably leads to the deposition of multiple conformers of the same biomolecule in the RCSB PDB, thereby linking several entries to a single RNA strand. This can happen for several reasons. For example, various experimental works may solve the structure of the same biomolecule, depositing different conformations of such molecules under different entries of the RCSB PDB. Alternatively, nucleotide-wise mutations can be investigated to go beyond the wild-type molecule, resulting in RNAs of the same organism with similar sequence and structure but which constitute separate entries in the RCSB. Thus, the need of classifying similar structures has recently emerged from the various forms of redundancy occurring in the available structures. Over the years, to address this need, several databases have been developed, each clustering *similar* structures according to distinct and rigorously defined criteria.

To name a few notable examples, DARTS [76] was one of the first databases clustering RNA molecules while keeping into account the conservation of RNA at the primary, secondary, and tertiary structural level. The RNA clusters display a hierarchical classification that reflects structural relationships *inter*- and *intra*-clusters. The representative structure of each cluster consists of the RNA that maximizes the sequence or structure similarity among the chains within

the cluster with similar length.

RNA3DHub is an actively maintained non-redundant dataset of RNA 3D structures, being updated on a regular basis since its release fifteen years ago [77]. The classification of the RNAs is performed on chains sharing similar sequence, geometry, and species. For each redundant group, the so-called 'Integrated Functional Element' (IFE) is chosen as the molecule or set of molecules which optimize the combination of resolution, percentage of nucleotides with steric clashes, basepairing consistence and other scores with the intent to select the structure (or structures) with the best experimental evidence for the reported coordinates. In this regard, more than one chain can be chosen as representative of a single redundant cluster.

Rfam is the largest database of non-coding RNA (ncRNA) families for genomic annotations [78]. Each family is represented by multiple sequence alignments and statistical models annotated with consensus secondary structures. Since the first release of Rfam in 2003, the database coverage was expanded to include also metagenomic, miRNA and viral families like *Flaviviridae* and *Coronaviridae*.

RNA-NRD is one of the most recent non-redundant databases released publicly [79]. The database consists of clusters of similar structures, where a representative chain is selected for each group. The similarity is assessed both under the sequence and the structural point of view. The representative is chosen according to a quality score that keeps into account structural features, experimental method of investigation and release date. There are two versions available of RNA-NRD, namely *with* and *without* organism of origin. In the former, redundant clusters are initially formed by separating RNAs coming from different organisms, labeling them as non-redundant *a priori* and facilitating the identification of the biological origin (for our purposes, viruses) of the nucleic acids. In case of RNAs originating from unidentified organisms, the chains are clustered together in the same group of "other" organisms. In the latter, the clusters are formed starting from the redundant clusters of the first database, merging redundant RNA chains from separate organisms and grouping them together.

2.2 Construction of the List of Viral RNAs

The goal of this Thesis is to perform a systematic study on the translocation properties of RNAs of viral origin. In the following Sections, we will describe the various steps performed to select such viral RNAs. Our *desiderata* is to construct a size-manageable collection of biomolecules encompassing the widest possible range of sequences, structures, and biological functions, in order to probe the systematicity of their effects on selected viral RNAs.

We constructed the set of RNAs of viral origin starting from the database RNA-NRD by picking the version in which the representative clusters are divided by source organism. In this way, the identification of viral RNA is facilitated. As aforementioned, it is possible to run into some form of redundancy across the selected structures because of the organism division, as RNAs chains belonging to different organisms are labeled as non-redundant by construction in RNA-NRD [79]. Thus, similar structures should be removed *a posteriori* to obtain a set of truly non-redundant molecules, as we will show in Sec. 2.2.6.

Firstly, we used the RNA-NRD on February 19, 2024. The dataset consists of a collection of 9705 RNAs with at least 20 nucleotide, clustered into 3157 non-redundant groups. A representative chain is picked for each cluster, being chosen as the chain with the largest quality score, a scalar number that keeps into account several factors ranging from the resolution of the structure, to the sequence length, number of base pairs, chain degree, experimental method to finally the release date [79]. Each chain within RNA-NRD is reported in the form `XXXX_Y`, where `XXXX` is the PDB identifier, and `Y` is the alphanumerical code of the reference chain in the PDB entry. We will adopt the same notation throughout this Thesis.

ID	Representative	Cluster	Organism	Molecule Name(s)	Rfam Name
1	5KK5_B	5KK5_B	Acidaminococcus	RNA (40-MER)	undefined
2	5B43_B	5B43_B, 5XH6_B, 5XH7_B	Acidaminococcus	crRNA, RNA (43-MER)	undefined
3	7D82_A	7D81_A, 7D82_A	Acidobacteriaceae bacterium	832GAAA (50-MER)	undefined
...	...	...	...	...	...
3155	2LI4_A	2LI4_A	Yersinia	RNA (32-MER)	undefined
3156	5AJ0_Bv	5AJ0_Bv, 5AJ0_Bw	Yersinia	tRNA	tRNA
3157	5TPY_A	5TPY_A	Zika virus	RNA (71-MER)	undefined

Table 2.1 Head and tail entries of RNA-NRD database downloaded on February 19, 2024. The representative chain is extracted from the redundant cluster as the chain with the highest quality score (value not reported).

The RNA-NRD database is provided as a comma-separated text file in which each entry contains the following fields: cluster identifier, representative chain, list of RNAs belonging to the cluster, source organism, macromolecule name, and Rfam family name. A few sample entries of the original database are reported in Tab. 2.1.

2.2.1 Viral RNAs

The first step of this work consists of the selection of the RNAs of actually viral origin from the dataset, since it comprises *all* RNA structures of the RCSB PDB sourced from different organisms.

In order to do so, we selected the unique organism descriptors available in the "Organism" field of the database, retaining 188 distinct values. Given the limited number of occurrences, we inspected each of these entries labeling it as "viral", "non-viral" or "unclear". The latter category was used for RNAs derived from unknown organisms, which therefore could not be unquestionably assigned *a priori* to either the "viral" or "non-viral" groups. The various entries were probed with the National Center for Biotechnology Information (NCBI) Taxonomy Browser [80]. RNAs whose "Organism" field was evidently a virus, *e.g.* phages, *Hepatoviruses* or *Coronaviruses*, were labeled accordingly (43 unique entries, associated to 125 representative chains). Similarly, clearly non-viral RNAs whose "Organism" field in the database contained labels like eukaryotes, bacteria and *Archaea* were labeled accordingly (138 unique entries, associated to 3016 representative chains). The remaining 16 representative structures were labeled as "unclear" and later inspected.

The starting core of the list of RNAs of viral origin consisted of the clearly viral entries identified above. In addition to that, we checked the other two groups, as they might have been mislabeled and contain putative viral RNAs in the same way. For instance, a cursory inspection of the "non-viral" RNAs revealed that many instances of nucleic acids that infect humans, such as AIDS, were listed under *Homo Sapiens* organism.

Then, we checked the 16 "unclear" entries by downloading the corresponding PDB entry and manually checking the source organism. We retrieved seven entries within the "unclear" group as being of viral origin.

Next, we inspected the larger group of "non-viral" entries in order to spot actually viral entries erroneously included within this set. For example, the xrRNA of the *Potato leafroll virus*

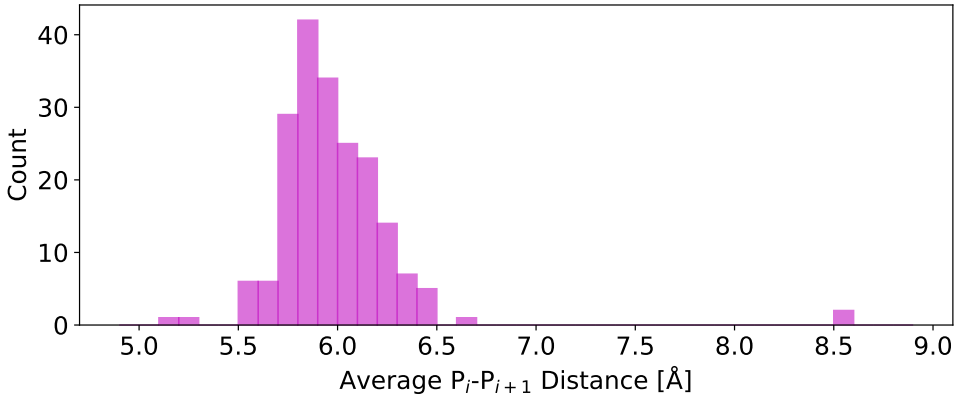


Figure 2.1 Distribution of the average distance between consecutive phosphates for the 196 RNAs selected up to Sec. 2.2.2. The data are peaked around 6 Å with few outliers with large average phosphates distance.

(PLRV), labeled `7JJU_A`, was misclassified as "Potato" in RNA-NRD. Thus, we searched for virus-related keywords making use of the regular expressions `vir` (for viral, Viral, virus, etc.) and `hage` (for phage, Phage) across the corresponding crystallographic files. In this way, the 3016-entry set of "non-viral" RNAs was reduced to a more manageable set of 331 entries of putative viral origin. Similarly to the "unclear" set, we manually inspected the entries of such set to retrieve in the end 237 molecules of actually viral origin.

Finally, we collected from RNA-NRD all the selected RNAs of actually viral origin, reducing the initial list of 3157 molecules to 369 candidates.

2.2.2 Gapless RNAs

For the RNA simulations, we require that the chains involved used to be gapless, meaning that the coordinates of *all* residues and their heavy atoms (*i.e.*, all non-hydrogen atoms) must be determined experimentally. Such information is included in the initial remarks (from `REMARK 465` to `REMARK 480` in each file) outlined in the header of any PDB files, which report the list of possible missing residues and non-hydrogen atoms in the corresponding PDB entry. In this way, we retrieved 196 gapless RNA molecules from the previous set of 369 candidates.

Then, we analyzed the distances between consecutive phosphate atoms, to ensure not to include any gapped structure. We observed that for most of the RNAs collected, the average distance of consecutive backbone phosphate atoms ranged from 5.5 to 7 Å, see Fig. 2.1. This range of distances is comparable with the distance distribution for consecutive phosphate atoms calculated in Ref. [81]. Moreover, we noticed that nine molecules have a maximum distance between consecutive phosphate atoms $d(\mathrm{P}_i, \mathrm{P}_{i+1}) > 8.0$ Å, presenting a particularly large distortion between these atoms.

We thus select only the RNAs with average distance between consecutive atoms smaller than 7 Å, and whose maximum distance between consecutive phosphates is smaller than 8 Å. We thus discard them, collecting in the end 187 molecules with all atom positions defined, and with reasonable distances between consecutive phosphate atoms.

2.2.3 Compact RNAs

The selected molecules encompass pseudoknotted structures, stem-loops, and multiple-domain structures, with shapes ranging from globular to corkscrew-like. Some of these conformations

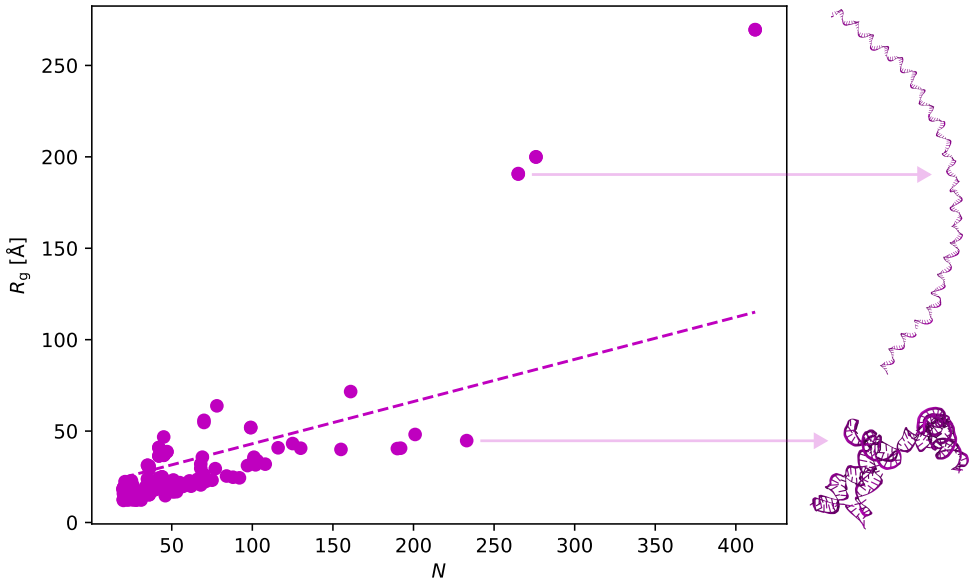


Figure 2.2 Radius of gyration R_g as a function of the chain length. The dashed line was drawn to identify clear outliers in the plot: RNAs with a radius of gyration R_g larger than such threshold should be discarded, as their structure is exceptionally elongated. Two illustrative RNAs of similar length are represented (they are not scaled relatively to each other). Both molecules have similar lengths ($N \approx 250$ nucleotides), but their three-dimensional structures are very different: one is compact (RNA `4C4Q_N`, below) while the other assumes a corkscrew-like conformation (RNA `1H1K_J`, above). Both representations of the RNAs were realized with VMD [82].

are strongly elongated and therefore are not suitable for our simulations, as such stretched structures are often stabilized by *inter*-molecular interactions or by pairing with other RNA strands. We thus want to remove these molecules from the selected list of RNAs.

One simple way to probe the overall size of an RNA molecule is to examine its radius of gyration R_g. Given an N-nucleotide RNA, we define R_g as

$$R_g^2 = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_{i,C_{3'}} - \langle \mathbf{r} \rangle_{C_{3'}})^2 \tag{2.1}$$

that is, the squared radius of gyration R_g^2 is the average squared distance of the $C_{3'}$ atoms of the sugar moieties from their center of mass

$$\langle \mathbf{r} \rangle_{C_{3'}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{r}_{i,C_{3'}} \tag{2.2}$$

The radius of gyration R_g alone is insufficient to fully characterize molecular shape of an RNA chain; however, when compared to the chain length N for a set of heterogeneous molecules, it can be effectively used to discriminate compact from particularly elongated shapes, which are identified as outliers in the R_g vs N distribution. We computed the radius of gyration for all RNA structures in the dataset, which we reported in Fig. 2.2.

Along with the increase of R_g with the chain length, we observe that a few RNA molecules display a radius of gyration which greatly exceeds the global scaling with the molecule size [83]. In order to define and identify such outliers, we set a cutoff R_g as a function of the chain length

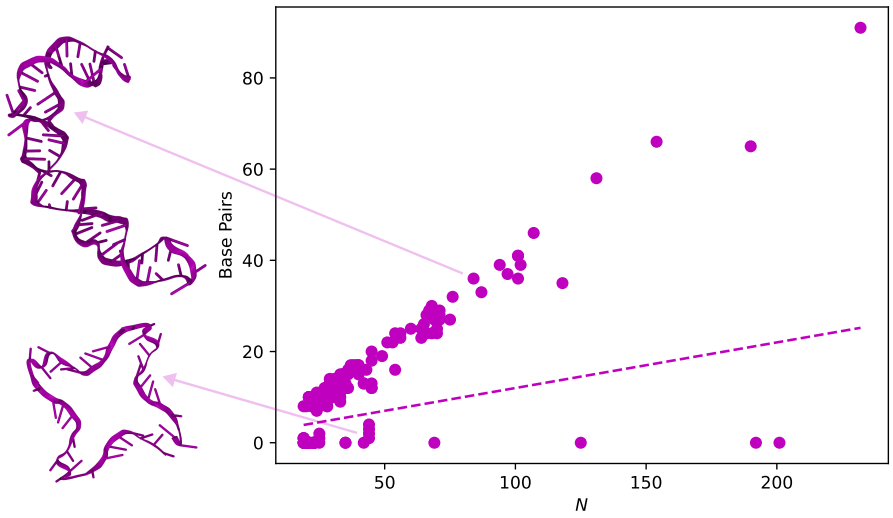


Figure 2.3 Number of base pairings as a function of the chain length. The dashed line was drawn to help identify the outliers in the plot: RNAs with a particularly small number of base pairings as a function of the chain length should be discarded as they are not properly structured. Two illustrative RNAs are represented (the molecules are not scaled to each other): one which is fully structured (RNA `1P5P_A`, above), and the other which makes no base pairings between its nucleotides (RNA `3ZLA_J`, below). The representations of both RNAs were realized with VMD [82].

N, which allows us to discriminate the excessively elongated structures from the more compact ones, as observed in Fig. 2.2. We exclude from our list these structures as they correspond to outliers with anomalously large R_g values.

Selecting the chains in such a way, we reduced the list to 155 viral, gapless and globular RNAs.

2.2.4 Structured RNAs

Then, we evaluated to what extent the collected chains have a self-contained secondary structure by examining the formed base pairs. In particular, we annotated for each PDB entry both the number of nucleotides and the number of base pairs made by the representative chain of the entry using X3DNA/DSSR [84, 85], a program for the analysis and modeling of nucleic acids. Then, we plotted the number of base pairs as a function of the chain length for the selected 155 molecules. The results are reported reported in Fig. 2.3, where we observe two clearly distinct groups of molecules with different behaviors. In the first one, we observe an approximately linear growth of the number of base pairs formed with the chain length; the second group, on the other hand, consists of the set of molecules which form very few base pairings or no pairings at all, which we discard from our set. The former set of molecules, which we retain, contains 119 molecules, and we discard the 36 molecules belonging to the latter.

After excluding RNAs clearly without a proper secondary structure, we analyzed the number of base pairs formed within the representative chain of each PDB entry (*intra*-chain base pairing) and those formed between the representative chain and other molecules present in the same entry (*inter*-chain base pairing). In this way, we can identify molecules that establish contacts with other conformers of the same molecule within the crystal structure reported in the PDB file. Of

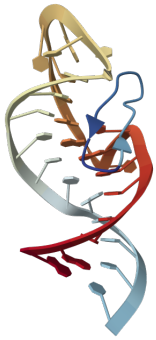


Figure 2.4 Representation of the RNA `2A9X_2`, which was removed among others in Sec. 2.2.5. The 5′- and 3′-end are drawn in red and blue, respectively. It is clear how the structures are affected by the presence of molecules inserted in their grooves. The chains were represented with the *Mol* 3D Viewer* of RCSB PDB [87].

the 119 RNA molecules collected insofar, five molecules showed a mismatching number of *intra*- and *inter*-chain base pairing. After discarding these molecules, we trimmed our set to 114 RNA molecules of viral origin, gapless, relatively compact and with a proper secondary structure.

2.2.5 Removal of RNAs with Intercalated Ligands

We excluded from the selected set all RNA molecules containing peptides or ligands intercalated within their structure, since the presence of such external molecules may significantly alter the native RNA conformation. The primary objective of this work is to investigate how the structural properties of RNAs influence their translocation behavior. Consequently, the presence of bound peptides or ligands could affect the translocation process itself, shadowing the role of the architecture of the RNA molecule alone.

Given the limited size of the dataset, we inspected all the 114 molecular structures of the RNAs, listing which RNAs are bound to external molecules. For example, the structure of the *Bovine immunodeficiency virus* (BIV) trans-activation response (TAR) element, deposited in the RCSB PDB under the entry `2A9X_2` and represented in Fig. 2.4, has a cyclic peptide inserted in the major groove. In this way, multiple contacts are formed between the nucleotides spanning the major groove and the wrapped peptide [86].

After scanning all the molecules selected so far, we identified 71 RNAs without intercalated ligands, peptides or proteins, which we retained.

2.2.6 Actually Non-Redundant RNAs

Up to this point, we have retained the 71 RNA chains according to taxonomic and structure-based criteria, keeping gapless and compact molecules of the initial database RNA-NRD with a self-contained secondary structure. However, these selected chains can be partially redundant, since RNAs originating from different organism are classified as non-redundant by construction in the version with organism division of RNA-NRD we used as a ground base for the selection of our list [79].

Thus, the objective of the next step in the selection of the RNA molecules is to eliminate redundant structures from the candidate set by reassessing possible redundancies through a systematic analysis of both sequence homology and structural similarity across the RNA chains. The former similarity across the chains was analyzed using ClustalOmega [88, 89], which is

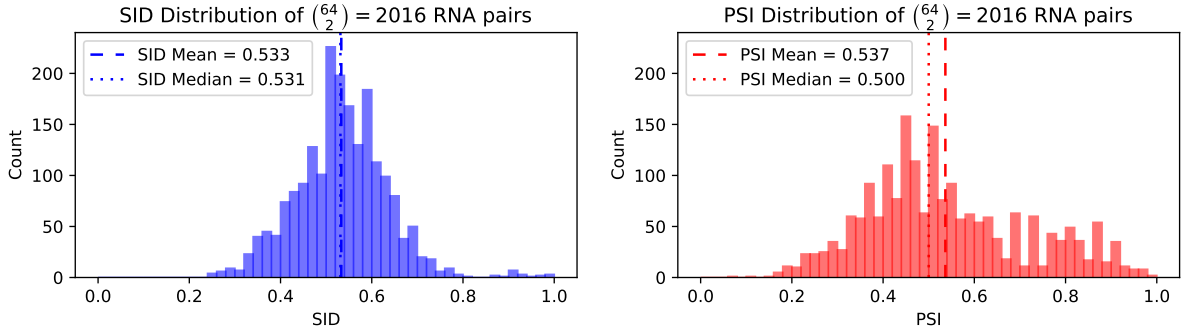


Figure 2.5 On the left, the histograms with the occurrences of SID scores for the ensemble of 64 RNAs collected; on the right, the histograms with the PSI occurrences. The mean and median values of each distribution are reported as a dashed and dotted line, respectively.

arguably the most used tool for sequence alignments of nucleic acids. The structural alignments were performed using SARA [90], a computer program for pairwise alignment of RNA structures. The statistical significance of the structural alignment scores associated to the secondary and tertiary structures of each RNA pair is assessed by different p-values computed over a benchmark set of RNAs with known distribution of scores.

In order to perform such alignments, we collected first all the sequences of the 71 RNAs by retrieving for each RNA their FASTA text file from the RCSB PDB. FASTA text files are commonly used to represent nucleotide (or amino acid, in case of proteins) sequences of biomolecules as strings of characters, using single-letter codes to identify single residues. We noticed that seven of these sequences contain *non-standard* nucleotides (that is, nucleotides which are not A, C, G or U). Thus, we removed *a priori* these sequences from the list of RNAs to align, as the sequence and structure alignment tools handle the non-standard nucleotides[1] with different, non-compatible approaches. In particular, ClustalOmega substitutes non-standard residues with a dummy nucleotide labeled "X" and then performs the alignment, whereas SARA simply removes the nucleotide from the structure, leaving a hole in correspondence of the altered nucleotides the chain, and then aligns the two molecules. Thus, the similarity assessment was performed on a reduced set of 64 selected RNA molecules containing only standard residues.

Sequence Homology

Thus, we listed first all the $\binom{64}{2} = 2016$ sequence pairs and aligned them with ClustalOmega. Then, to quantify the similarity in sequence, we computed for each aligned pair the Sequence Identity (SID) in the same manner of Ref. [76], that is as the fraction of aligned nucleotides of the shortest chain

$$\mathrm{SID} = N_{\mathrm{aligned}}/N_{\mathrm{short}} \tag{2.3}$$

where N_{aligned} and N_{short} are the number of aligned nucleotides with ClustalOmega and the length of the shortest chain of each pair, respectively. This score ranges from 0 for totally mismatching sequences to 1, when the shortest sequence is perfectly aligned with the longest one.

The distribution of SID values is reported on the left of Fig. 2.5. The minimum and maximum values of SID observed read 0.250 and 1.000, respectively, with a mean value of 0.537 and a median equal to 0.531.

[1]For example, the residues RY (5′-3,6-dihydrocytidylic acid) and GTP (guanosine-5′-triphosphate) occurred in different nucleotide sequences of our set.

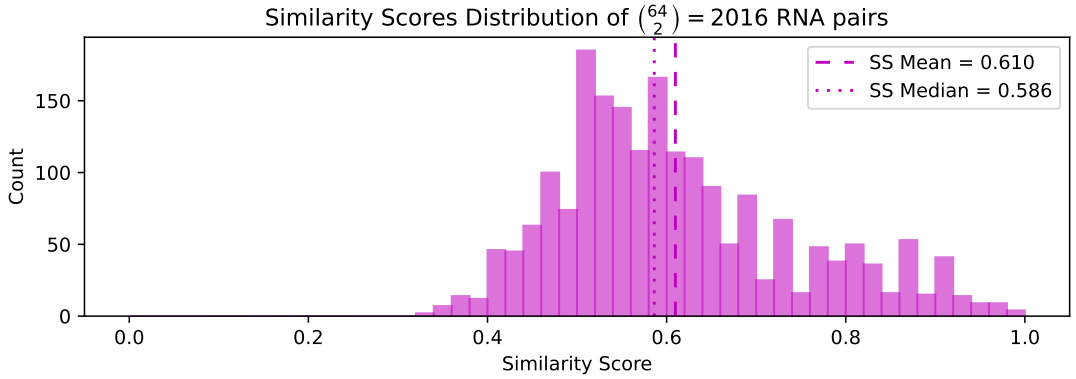


Figure 2.6 Histogram with the distribution of the Similarity Score for the ensemble of 64 RNAs collected. The mean and median values are reported as a dashed and dotted line, respectively.

Structural Similarity

The same 2016 RNA pairs were then structurally aligned with SARA using the default parameters, that is aligning the chains on the $C_{3'}$ atoms. We retrieved from the output of each alignment with SARA the Percentage of Structural Identity (PSI), which is defined in Eqn. (1) of Ref. [90] as the percentage of the aligned nucleotides of the shortest chain in the pair

$$\text{PSI} = 100 \times N_{\text{aligned}}/N_{\text{short}} \tag{2.4}$$

where N_{aligned} and N_{short} are the number of aligned nucleotides and length of the shortest chain of each RNA pair, respectively. Basically, this score corresponds to the structural counterpart of the sequence identity SID Eqn. (2.3), reported as a percentage.

To conform the data of the structural alignment with those of the sequence alignment, we collected the various percentage of structural identity values as normalized scores from 0 to 1 rather than percentages from 0% to 100%. The minimum and maximum values of PSI observed read 0.077 and 1.000 (when the shortest chain is perfectly aligned on the longest one), respectively, with a mean value of 0.541 and a median score equal to 0.500. The distribution of PSI is reported alongside that of SID on the right of Fig. 2.5.

Bringing It All Together

To assess the similarity between two RNAs, we merged the two sequence and structure similarity scores into the more general Similarity Score similarly as done in [76]:

$$\text{Similarity Score} = \max(\text{SID}, \text{PSI}) \tag{2.5}$$

The Similarity Score spectrum ranges from a minimum value of 0.324 to a maximum score equal to 1.000, where a perfect alignment in either structure or sequence occurs. The distribution of the scores is reported in Fig. 2.6: the mean and median values read 0.610 and 0.586, respectively.

In order to remove highly similar structures, we set a threshold Similarity Score = 0.900 following Ref. [76], where such score was used to identify pairs of non-redundant RNAs filtering out those with larger sequence or structural identity.

Thus, we ordered all RNA pairs by increasing length of their elements. For example, let the RNAs labeled i, j, k and l. Suppose the first three ordered pairs are $p_1 = (i, j)$, $p_2 = (i, k)$, and $p_3 = (i, l)$, where the chain lengths satisfy $N_i \leq N_j \leq N_k \leq N_l$. For molecules of equal length,

we applied alphabetical ordering (*i.e.*, i precedes j, k and l, j precedes k and l, and so on). Then, we looped over all the sorted RNA pairs, and for each RNA pair with Similarity Score ≥ 0.900 we removed the shortest chain (or the one whose identifier comes alphabetically before in case of equally long molecules) from the list of selected viral structures. After the removal of such chain and all the corresponding pairs in which it is involved, we start over again from the new first pair. In this way, we retained 41 non-redundant RNA molecules.

After the filtering step over the threshold 0.900, we identified a pair of RNA chains, namely `2NBY_A` (39 nucleotides) and `2NBX_A` (108 nucleotides), whose Similarity Score is very close to the threshold set. In fact, these two chains exhibit Sequence and Structural Similarity values of SID = 0.872 and PSI = 0.897, respectively. Both these values lie below the Similarity Score cutoff, yet they clearly suggest a substantial degree of similarity (hence, redundancy) between the chains. We observed that the PDB descriptions of the two structures read "Solution structure of the J domain of EMCV IRES" and "Solution structure of the J-K region of EMCV IRES", respectively, meaning that there is clearly an overlapping domain between the two structures [91]. Thus, we manually removed *a posteriori* the shortest RNA chain `2NBY_A` from the selected list, reducing the set down to 40 compact and non-redundant nucleic acids of viral origin.

2.2.7 RNAs with Exposed Ends

As aforementioned, the goal of this Thesis is to study how the translocation of viral RNA through a nanopore is affected by the structure of the biomolecule. Experimentally, the translocation of structured biomolecules through a nanopore is carried out by an electric field that drives the RNAs through the pore. Occasionally, the translocation of folded biomolecules is facilitated by adding an unstructured leading strand at the pulled end of the RNA, so that it can enter the pore with minor hindrance. For example, a poly-dA_{50} tail was attached to different hairpins to study in [24] the unfolding properties of hairpins when threaded though a α-hemolysin pore; likewise, leading strands can be used to discriminate the orientation of engineered ssDNA strands inside α-hemolysin pore [23].

As we will show in Chapter 3, the force that drives the RNAs through the nanopore is applied solely inside the pore and in the near proximity of its entrance in our simulations of RNA forced translocations. Without a guiding leading strand as mentioned above, the translocation can occur when the termini of the chain are exposed as they need to enter in the nanopore. By analogy, threading a ball of yarn through the eye of a needle is straightforward only when the free end of the yarn is accessible. Thus, the next filter consists of retaining the structures with exposed termini.

The task of assessing whether a terminal is exposed is equivalent to finding a plane passing through such terminal such that all the nucleotides are included in one of the two half spaces identified by the plane itself, see Fig. 2.7. We opted for a simplified representation of the RNA chains using one atom per nucleotide in order to maximize the chances of finding the separating plane. The most natural choice for the target atoms is the phosphate atoms, as they are the ones on which the driving force inside the nanopore is applied.

Finding the separating plane means finding a vector $\mathbf{w}$ that points in the opposite direction of all the distance vectors between the terminal (phosphate) atom and the other (phosphate) atoms of the chain. If such plane exists, the whole space is divided into two parts, with one half containing all such atoms. The existence (or non-existence) of $\mathbf{w}$ can be proved using a linear binary classifier algorithm, that is, a *perceptron* [93], which we report in Alg. 1 for one of the termini (the one indexed $i = 0$). The procedure described in Alg. 1 can be applied also to the other terminal (with index $i = N - 1$) to assess if both chain ends are separable.

After running the perceptron algorithm on the RNA chains selected so far, we report that one RNA had both ends buried, another one had only the $5'$-terminal revealed, and that six

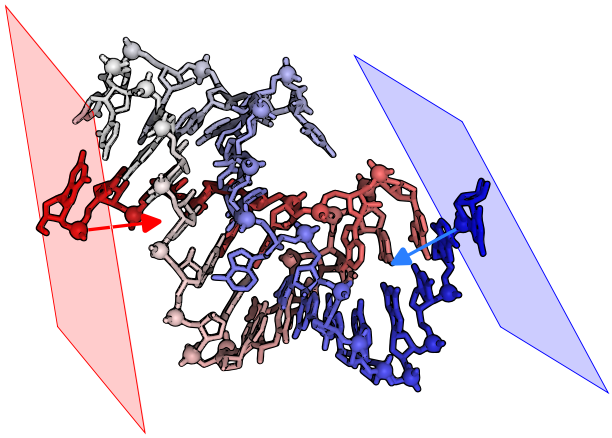


Figure 2.7 The pseudoknot of SRV-1 RNA (PDB code and representative chain: `1E95_A`) has exposed ends: for each terminal there exists a plane, passing through the terminal, such that *all* the phosphate atoms are included in one of the two half-spaces identified by the plane. The 5′- and 3′-end are colored in red and blue, respectively; the represented planes as well as the corresponding normal vectors are indicative. Phosphate atoms are represented as beads. The RNA picture was generated with VMD [82] and is adapted from Ref. [92].

Algorithm 1: Assess the exposure of the termini of an RNA molecule

Data: $\mathbf{x}_i$, phosphate coordinates of an N-residue RNA (index $i = 0, 1, \ldots, N-1$);
$\epsilon \ll 1$, learning rate; $N_{\text{threshold}}$, limit to the number of iterations

Result: $\mathbf{w}$, vector normal to the separating plane

Without loss of generality, we analyze only the first end with index $i = 0$;

begin
  $\mathbf{w} \leftarrow$ random unit vector; `/* Initialize w to a random unit vector */`
  $d_k \leftarrow -1$; `/* Initialize` d_k `to a negative number */`
  while $d_k \leq 0$ *and* $N_{iterations} \leq N_{threshold}$ **do**
    for $i = 1, \ldots, N-1$ **do**
      $d_i \leftarrow \mathbf{w} \cdot (\mathbf{x}_i - \mathbf{x}_0)$; `/* Compute projections of all distances from terminal P */`
    end
    $d_k \leftarrow \max(d_i)_{i=1,\ldots,N-1}$; `/* Find largest projection */`
    $\mathbf{w} \leftarrow \mathbf{w} - \epsilon(\mathbf{x}_k - \mathbf{x}_0)$; `/* Slightly tilt the plane */`
    $\mathbf{w} \leftarrow \mathbf{w}/|\mathbf{w}|$; `/* Normalize the vector plane */`
  end
end

structures had only the 3′-end exposed. Thus, we discarded these eight RNAs, retaining 32 RNAs of viral origin with a relatively compact and gapless structure, and with both termini exposed.

2.2.8 Grouping RNAs by Topological Complexity

Some of the 32 RNA molecules selected so far encompass a similar secondary structure: for example, we annotate among the RNAs 19 stem-loops, and four H-type pseudoknots. We wish to know if there are forms of structural redundancy across molecules with similar fold: for example, assessing the degree of similarity of the H-type pseudoknots selected so far.

Thus, we used x3DNA/DSSR to annotate the secondary structures of the selected chains and retrieve the number of pseudoknots of each molecule. Then, we arranged the RNAs into different groups of increasing topological complexity, which we quantified with the number of pseudoknots of each structure. Recall pseudoknots are RNA secondary structures containing at least two helical segments with an intercalated single strand or loop between them. They fold into knot-shaped three-dimensional conformations but are not true topological knots [52]. These groups comprise RNAs with zero to four pseudoknots, as the latter is the largest number of pseudoknots detected.

The underlying idea is to merge structures with same topological complexity into "similarity clusters"; after that, we retrieve only one representative RNA for each cluster, analogously to the procedure used to build RNA-NRD [79]. Thus, we aligned all RNAs within the same topological group using RNAHugs [94], an alignment program that allows multiple structures to be aligned simultaneously. Once the alignments were performed, we computed the Alignment Score for each RNA pair as

$$\text{Alignment Score} = N_{\text{aligned}}/N_{\text{long}} \tag{2.6}$$

where N_{aligned} and N_{long} are the number of aligned nucleotides and the length of the longest chain of each pair, respectively. The Alignment Score is now different from the one employed in Sec. 2.2.6. There, we aimed to detect the extent of the structural inclusion between two RNAs, like a fragment of one RNA and the other. Here, instead, the objective is to quantify structural redundancy within groups of RNAs sharing the same topological complexity in order to define representative structures regardless of their size. In this setting, normalizing by the longer chain length N_{long} provides a more conservative measure of similarity, reflecting better the overall structural similarity of two RNA chains, namely both the aligned nucleotides, the relative size of the two chains and their topological complexity.

Then, the similarity clusters are constructed as follows. For each group of RNAs with k pseudoknots, we treat each RNA as a node of a graph $\mathcal{G}_k$. First, we select a similarity threshold which tells whether two RNAs are structurally similar. Then, we build the adjacency matrix A_{ij} of the graph $\mathcal{G}_k$ of k-pseudoknot RNAs from the matrix of Alignment Scores S_{ij}, setting

$$\begin{cases} A_{ij} = 1 & \text{if } S_{ij} > \text{threshold} \\ A_{ij} = 0 & \text{otherwise} \end{cases} \tag{2.7}$$

Once $\mathcal{G}_k$ is built, we analyze its connected components (that is, its similarity clusters). For each connected component, we pick as representative RNA the molecule with the largest average similarity across *all* of its members. For clusters of two, we pick as representative the longer molecule; in case of equally long RNAs, we pick the representative based on functional considerations, or quality of the structure, if known. Finally, the remaining isolated nodes (single-molecule clusters) are automatically included in the list of selected RNAs. Notice for threshold = 0 one obtains a complete (fully-connected) graph, while for threshold = 1 no edge can occur and all nodes are disconnected.

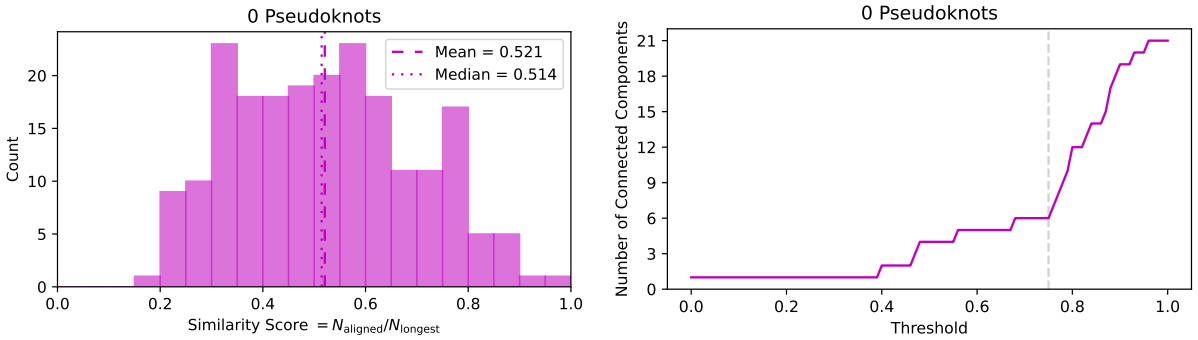


Figure 2.8 On the left, the distribution of the Similarity Score $= N_{\text{aligned}}/N_{\text{longest}}$ values for the group of non-pseudoknotted structures. The mean and median values are reported as a dashed and a dotted line, respectively. On the right is reported the number of connected components (both clusters and isolated nodes) in the similarity graph $\mathcal{G}_0$ of non-pseudoknotted molecules. The threshold value 0.75 at which the elbow point occurs is marked as with a dashed gray line.

Non-pseudoknotted Structures

We first analyzed the non-pseudoknotted structures. This is the largest group of molecules of our set, comprising 21 RNAs which account for almost two thirds of the selected RNAs. After the alignment of all RNA molecules with RNAHugs, we collected the similarity scores of this group and computed the distribution, which is reported on the left of Fig. 2.8. We observe a distribution of the similarity scores $N_{\text{aligned}}/N_{\text{long}}$ where the mean and median value read 0.521 and 0.514, respectively.

Next, we scanned all thresholds ranging from 0 to 1. For each threshold value we constructed the graph $\mathcal{G}_0$ with the procedure previously described, and then we computed the number of connected component (that is, both clusters and isolated nodes) as a function of the threshold. The results are plotted on the right of Fig. 2.8. The curve has a sigmoid-like behavior: when threshold $\lesssim 0.40$, $\mathcal{G}_0$ consists of a single giant cluster spanning all RNAs/nodes. Then, it increases slowly until threshold $= 0.75$, where an elbow point is found. After the elbow, the number of clusters increases steeply until threshold ≈ 0.90, after which the curve slowly saturates to 21 (the number of non-pseudoknotted structure) as the similarity threshold approaches the limit 1. We thus set the threshold to the elbow value threshold $= 0.75$ in order to construct $\mathcal{G}_0$. We will use the same threshold to construct also $\mathcal{G}_{k>0}$, see the following Sections.

In Fig. 2.9 it is sketched the representation of $\mathcal{G}_0$. We observe that there emerge five isolated nodes, corresponding to the RNAs `1P5P_A`, `2MQT_A`, `2NBX_A`, `3R4F_A` and `6VAR_A` (reported in alphabetical order), and a single giant cluster comprising the remaining 16 molecules, for a total of 6 connected components.

Finally, we need to extract the representative RNA for the giant similarity cluster of $\mathcal{G}_0$. As aforementioned, we select across the RNAs constituting such cluster the molecule with the largest average similarity across *all* members of the cluster. The RNA labeled `1KP7_A` is selected as representative of the cluster, since it displays the largest average Alignment Similarity, equal to 0.715. Notice such average Alignment Similarity is smaller than the selected threshold $= 0.75$, because the mean value is computed across *all* elements of the clusters, also between nodes which do not have an edge connecting them (because in this case $A_{ij} = 0$ does not imply $S_{ij} = 0$ but simply $S_{ij} < 0.75$, see Eqn. (2.7)).

To summarize, we retained from the initial 21 non-pseudoknotted structures the following six RNAs, reported in alphabetical order: `1KP7_A`, `1P5P_A`, `2MQT_A`, `2NBX_A`, `3R4F_A` and `6VAR_A`.

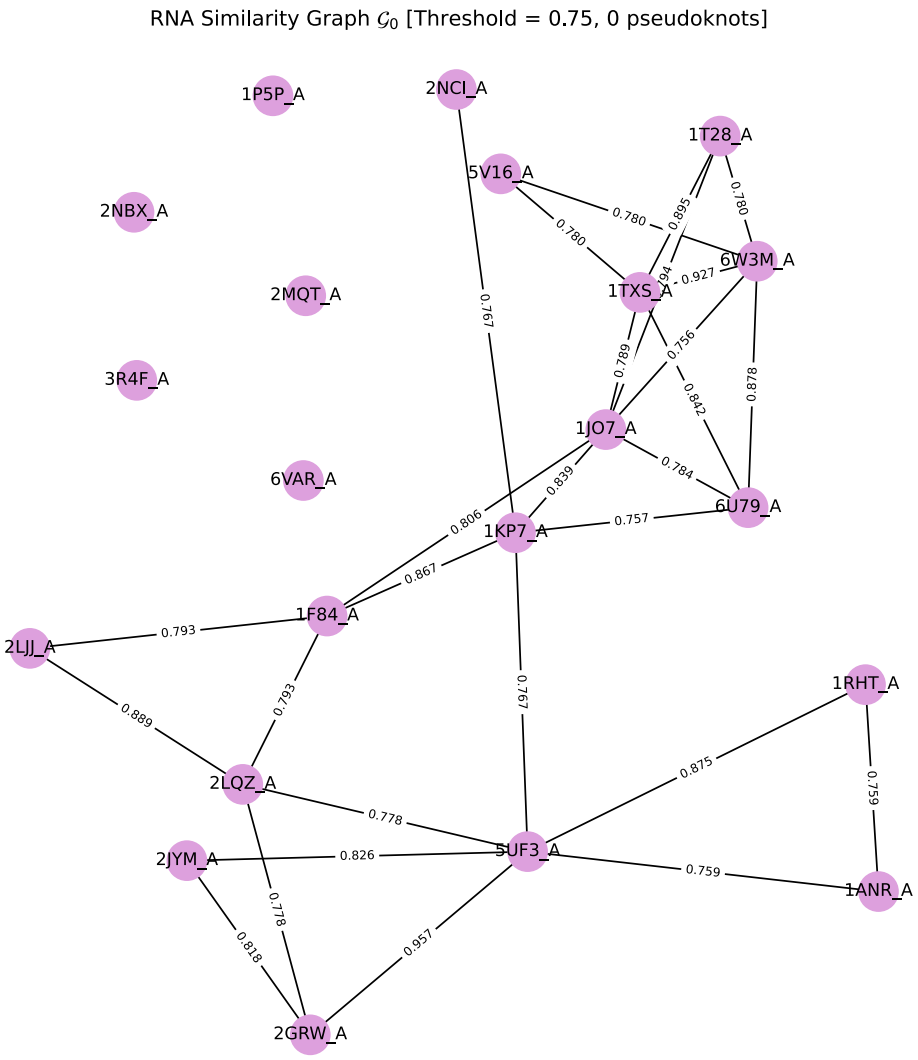


Figure 2.9 Visual representation of the similarity graph $\mathcal{G}_0$. Notice the six connected components: five isolated nodes and a giant cluster comprising the remaining 16 nodes. Edges are drawn only if threshold > 0.75 for the pair of RNA nodes. The image was generated with NetworkX [95].

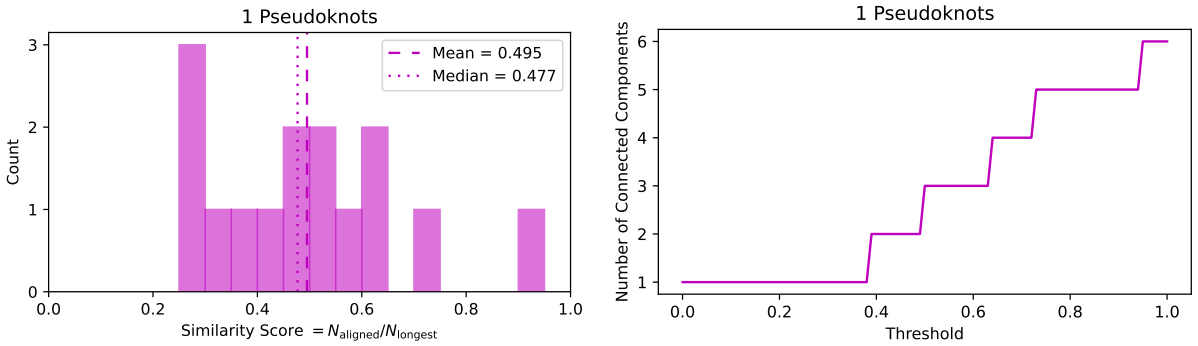


Figure 2.10 On the left, the distribution of the Similarity Score $= N_{\text{aligned}}/N_{\text{longest}}$ values for the group of structures with a single pseudoknot. The mean is represented as a dashed line, while a dotted line is drawn in correspondence of the median Alignment Score. On the right is reported the number of connected components (both clusters and isolated nodes) in the similarity graph $\mathcal{G}_1$ of molecules with a single pseudoknot.

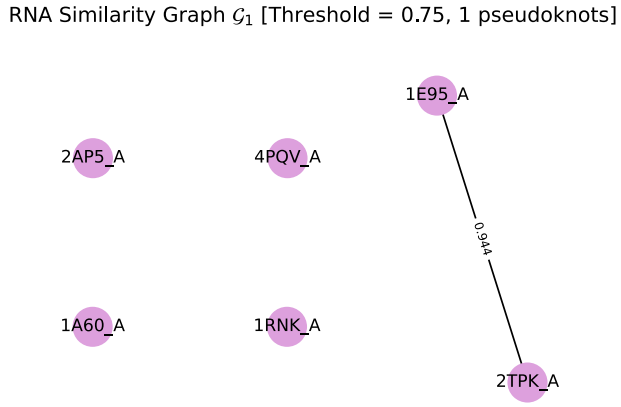


Figure 2.11 Visual representation of the similarity graph $\mathcal{G}_1$, constituted by four isolated nodes and a single pair of nodes. Edges are drawn only if threshold > 0.75 for the pair of RNA nodes. The image was generated with NetworkX [95].

1-pseudoknot Structures

We iterated the procedure for the group of six RNAs with a single pseudoknot. Given the more constrained size of the group with respect to the one of non-pseudoknotted structures, the distribution of similarity scores has a less clear trend compared to the one of Fig. 2.8. We report the histogram of the distribution on the left of Fig. 2.10, where the mean and median value of the alignment similarity scores read 0.495 and 0.477, respectively.

Then, we scanned again the thresholds ranging from 0 to 1, and constructed the similarity graph $\mathcal{G}_1$. For the selected threshold $= 0.75$, $\mathcal{G}_1$ consists of four isolated nodes and a single cluster of two RNAs, as sketched in Fig. 2.11.

This pair of similar molecules is formed by the RNAs labeled `1E95_A` and `2TPK_A`. We need to extract the representative for such duplex. The two pseudoknotted RNAs have the same length, thus we pick as representative the former due to its functional importance: such RNA is in fact the pseudoknot of the *Simian retrovirus* SRV-1, and it is involved in ribosomal frameshifting, a crucial biological phenomenon which allows viruses to compact more genetic information into their genome [56, 57, 96].

Hence, the representative structures collected consist of the following five RNA chains, re-

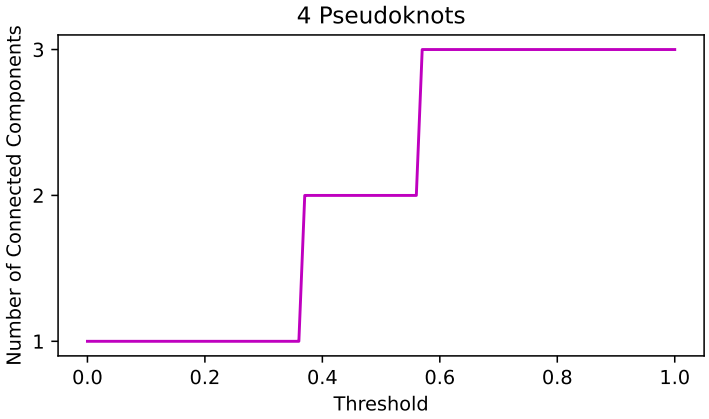


Figure 2.12 Plot of the number of connected components (both clusters and isolated nodes) in the similarity graph $\mathcal{G}_4$ of molecules with four pseudoknots.

ported in alphabetical order: `1A60_A`, `1E95_A`, `1RNK_A`, `2AP5_A` and `4PQV_A`.

2-pseudoknot Structures

We report a single structure displaying two pseudoknots, namely `2N8V_X`. We collect it directly inside our selected list of viral RNAs.

3-pseudoknot Structures

Only one structure comprises three pseudoknots, namely `4C4Q_N`. As previously done, we include it directly within our chosen set of viral RNAs.

4-pseudoknot Structures

The group of most topologically complex structures comprises three RNAs. Iterating the procedure reported in the previous Sections, we report in Fig. 2.12 the number of connected components of $\mathcal{G}_4$ as a function of the threshold employed. Again, for threshold $\lesssim 0.40$, the similarity graph $\mathcal{G}_4$ consists of a single cluster containing all RNAs/nodes. The mean value of the alignments is equal to 0.420, and the median is equal to 0.361.

The largest similarity is shared across the RNAs `5TPY_A` and `7K16_P`, with a score equal to 0.563, which means that for the selected threshold $= 0.75$ no cluster is formed. We thus retain all three RNAs with four pseudoknot crossings, reported in alphabetical order: `3T4B_A`, `5TPY_A` and `7K16_P`.

2.3 The Final List

In this Section we report in alphabetical order the complete list of the identifiers of the 16 viral RNA chains which were selected in this Chapter: `1A60_A`, `1E95_A`, `1KP7_A`, `1P5P_A`, `1RNK_A`, `2AP5_A`, `2MQT_A`, `2N8V_X`, `2NBX_A`, `3R4F_A`, `3T4B_A`, `4C4Q_N`, `4PQV_A`, `5TPY_A`, `6VAR_A` and `7K16_P`. In Tab. 2.2 are summarized the basic pieces of information about the structures: PDB ID identifier, virus of origin of the RNA, functional role of the RNA, number of atoms and nucleotides, radius of gyration R_g and reference to the experimental work in which the structure was solved. Below, we provide a brief description of all the RNAs selected, outlined in alphabetical order. The molecular structures represented with VMD [82] are reported in Fig. 2.13.

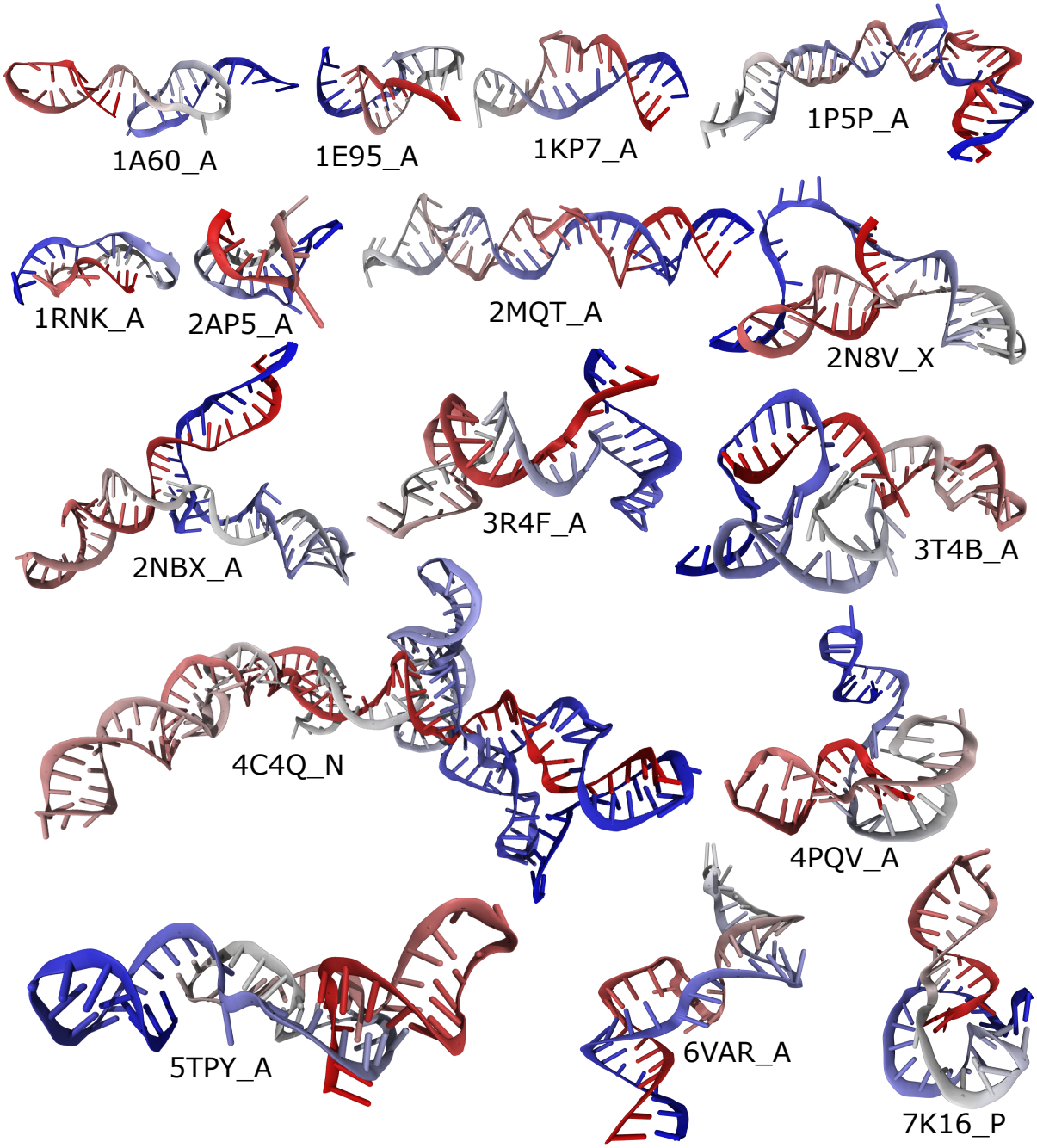


Figure 2.13 Molecular structures of the selected RNAs sketched with the ribbon representation of VMD [82]. The 5′- and 3′-terminal are colored in red and blue, respectively. The relative sizes of the molecules are not to scale. In Tab. 2.2 are reported the basic information of these RNAs.

PDB ID	Virus	Function	N_{atoms}	N_{ntd}	R_{g} [Å]	Ref.
1A60_A	TYMV	tRNA Mimicry	923	44	21.042	[97]
1E95_A	SRV-1	Ribosomal Frameshifting	767	36	14.310	[96]
1KP7_A	HCV	IRES Binding Site Loop	634	30	16.316	[98]
1P5P_A	HCV	IRES Domain II	1638	77	29.249	[99]
1RNK_A	MMTV	Ribosomal Frameshifting	726	34	15.702	[100]
2AP5_A	ScYLV	Ribosomal Frameshifting	601	28	11.862	[101]
2MQT_A	MLV	(U5-PBS) Domain	1449	68	29.592	[102]
2N8V_X	IAPV	IRES Domain PK-I	1473	69	25.945	[103]
2NBX_A	EMCV	IRES J-K Region	2304	108	31.451	[91]
3R4F_A	ϕ29	pRNA	1400	66	23.429	[104]
3T4B_A	HCV	IRES Central Domain	1771	83	24.623	[105]
4C4Q_N	CSFV	IRES Domain III	4986	233	44.454	[106]
4PQV_A	MVEV	xrRNA	1466	68	20.412	[63]
5TPY_A	ZIKV	xrRNA	1527	71	22.094	[62]
6VAR_A	HBV	εRNA	1290	61	22.557	[107]
7K16_P	TABV	xrRNA	1088	51	16.366	[64]

Table 2.2 Summary of the selected RNA with virus of origin, function/role of the RNA, number of atoms and nucleotides, radius of gyration R_{g} and reference to the experimental study. The "PDB ID" field comprises the PDB code, given by the first four characters, and the representative chain, given by the last character.

1A60_A The structure is a 44-nucleotide pseudoknotted fragment of the genome of *Turnip yellow mosaic virus* (TYMV) [97]. The molecule consists of an H-type pseudoknot stabilized by a hairpin located at the 5′-end of the RNA. The hairpin resembles the complete T-arm and valine-acceptor arm of the tRNA-like structure of the YUMV genomic RNA, while the acceptor arm is represented by the pseudoknotted domain of the RNA. Furthermore, the structural resemblance to the tRNA is testified by the similar reactivity to tRNA-specific enzymes [108].

1E95_A This RNA is a 36-nucleotide H-type pseudoknot, which is found in the *Simian retrovirus* type-1 (SRV-1) [96]. Together with a heptanucleotide slippery sequence it is responsible for programmed ribosomal frameshifting, which is exploited as a mechanism for regulating translation of its mRNA. The pseudoknot is made of two stacked stems connected by as many loops. All nucleotides except for the bulged C24 are paired, forming a very compact molecule.

1KP7_A This 30-nucleotide stem-loop constitutes an internal loop structure of the *Hepatitis C virus* (HCV) IRES [98], which contributes significantly to the IRES activity, providing most of the binding site affinity for eIF3 translation initiation factor. The stem of the 5′-end contains two bulges, while the other one is more regular with all the nucleotides towards the core of the helix. The structure contains a reverse wobble pair due to two inverted nucleotides (C183 and A215), and also two mismatched cytosines (C186 and C211) which stack onto each other inside the helix. The stems are capped by a stabilizing UUCG tetraloop.

1RNK_A This molecule consists of a 34-nucleotide H-type pseudoknot of the *Mouse mammary tumor virus* (MMTV), that works as frameshift stimulator similarly to other pseudoknots [100]. The structure is made of two GC-rich, A-form helices whose stems are not coaxial because of the intercalated A14 nucleotide, which is stacked between the two stems. This adenosine confers

the two axes a twist of approximately 112°. Furthermore, the 5′-terminal base pair is protected from cleaving by the first nucleotide of the second loop, that stacks upon the terminal G1.

`1P5P_A` This 77-nucleotide long stem-loop constitutes the domain II of the HCV IRES [99]. The molecule assumes a distorted L-shape which is crucial for the IRES activity, forming several contacts with the 40S subunit during protein translation. A stretch of noncanonical and Watson-Crick base pairs creates a continuous helix with the apical part of domain II. The hinge bulge (which confer the overall L-shape to the molecule) between the two helices is not particularly rigid according to the NMR data. The formation of such bulge is to be traced back to five single-stranded nucleotides. A second, less pronounced bend is due to loop E motif formed by noncanonical base pairs and a S-turn.

`2AP5_A` This structure is a 28-nucleotide pseudoknotted RNA, which comes from the genome of *Sugarcane yellow leaf virus* (ScYLV) [101]. The molecule presents a G25A mutation (indexing the nucleotides from 1, differently from the original paper [101]), but such mutation has a negligible effect on the three-dimensional structure, which is basically indistinguishable from the wild type. The molecule is very compact and presents three bulged nucleotides, namely G7, A11 and C23.

`2MQT_A` This 68-nucleotide stem-loop is the U5 primer binding site (U5-PBS) of the *Murine leukemia virus* (MLV) [102]. U5-PBS is a largely linear molecule capped by a structured UCAG tetraloop. and contains one single-nucleotide bulge After the first turn of the helix, the molecule presents a distortion in the major groove containing an extruded uracile. This major blge is necessary for binding of viral nucleocapsid proteins with the molecule.

`2N8V_X` This RNA consists of the PK-I domain of the IRES of the *Israeli Acute Paralysis* virus (IAPV) [103]. This domain forms multiple pseudoknots, mimicking the typical L-shape of tRNAs and allowing the RNA to perform frameshifting, switching from the 0 to the +1 reading frame. The structure comprises three helices, which stem from a three-way junction. Furthermore, two adenosine bulges in proximity of such junction were shown to be crucial for the reading frame selection. The 3′-terminal consists of a long strand that stretches from A58 to the terminal G69, partly superimposing with the variable loop and the T-arm of the mimicked tRNA. We cleaved the first nucleotide to facilitate the priming process, see Sec. 3.2. By removing such nucleotide, the global structure of the RNA is not particularly affected, as very few contacts are made by this base—thus, we do not expect it to play a major role in the activation process. In this way the domain reduces to 69 nucleotides.

`2NBX_A` This molecule consists of the 108-nucleotide J and K regions of the IRES of *Encephalomyocarditis virus* (EMCV) [91]. The structure has a global Y-shape, which is composed of the two J and K domains that bifurcate from the base helix, joined by a six-adenosine linker. Several bulged nucleotides decorate each domain, exposing them outside the core of the molecule.

`3R4F_A` This RNA is one of the 66-nucleotide multimer of the prohead RNA (pRNA), an essential component of the bacteriophage $\phi 29$ packaging motor [104]. pRNA is one of the units that form the ring which connects the virus capsid with the ATPase, the actual motor of $\phi 29$. The structure consists of three helices, two of which stem from a U-rich three-way junction near the RNA ends. The global shape of the molecules resembles a “U”, due to the fact that the helices are misaligned.

`3T4B_A` This RNA is the central domain of the HCV IRES, an 83-nucleotide pseudoknotted molecule [105]. The principal helix of the molecule is located below an engineered tetraloop and tetraloop receptor that were placed to facilitate the interaction with neighboring molecules. Another helix runs alongside the principal one. The global fold of the RNA is unusual, as a four-way junction is part of a classic H-type pseudoknot that comprises the two chain ends, and another pseudoknot is formed within one of the domain of the molecule.

`4C4Q_N` The structure is the Domain III of the *Classical swine fever virus* (CSFV) IRES [106]. It is a large 233-nucleotide RNA, the largest of our list, with multiple junctions from where several helices of different length depart. Multiple bulges decorate the helices of the RNA, and a pseudoknot is formed within the two subdomains of the chain, adding further topological complexity to the molecule.

`4PQV_A` The *Murray valley encephalitis virus* (MVEV) accumulates its nonoding RNA through the incomplete degradation of its RNA by the XRN1 exonuclease. The degradation resistance is conferred to the virus genome by a series of modules called xrRNAs. In the case of MVEV, these modules consist of 68-nucleotide chunks of the genome with a complex topological structure, filled with pseudoknots and a ring-like structure protecting the 5′-end [63]. The xrRNA consists of four helices which spring from a three-way junction and that crumple to form a pseudoknot. The global L-shape resembles that of the other xrRNAs present in this work, namely `5TPY_A` and `7K16_P`.

`5TPY_A` The *Zika virus* (ZIKV), being a member of the *Flaviviridae* family, accumulates its noncoding RNA in the infected cell thanks to the partial degradation of its genome. The fraying of the RNA, usually by means of exonucleases, is halted by xrRNAs, submodules of tens of nucleotides which are found in series in the 5′-end of the UTR. Similarly to other xrRNAs, the ones present in the ZIKV UTR comprise four helices that form multiple pseudoknots. The whole module consists of 71 nucleotides, and shows also the important ring-like structure that protects the 5′-end from degradation [62]. The global L-shape of the ZIKV xrRNA is similar to that of the other xrRNAs showed in this study, `4PQV_A` and `7K16_P`.

`6VAR_A` The εRNA of the HBV is a 61-nucleotide helical nucleic acid which is crucial to the replication of the virus [107]. The helix is divided into two sub-helices, the lower and the upper helix which contains the bulged nucleotides C36 and U43, respectively. The two helices are joined together by a 6-nucleotide bulge, which confer an overall Z-shape to the molecule.

`7K16_P` The *Tamana Bat* virus (TABV) was shown to contain several xrRNA modules in the 3′-UTR of its genome [64]. The structure of each of these xrRNA modules consists of three helices which arise from a three-way junction. Several pseudoknots make the molecule rather compact, bridging together arms of the various helices. Similarly to other xrRNAs, a protective ring embraces the 5′-end to protect it from degradation of nucleases. However, the ring-like structure is more compact than that observed in MVEV and ZIKV xrRNAs. Still, the overall exhibited L-shaped configuration is similar to that observed in other xrRNAs examined in this work.

2.3.1 A Note on IRESs and xrRNAs

We observe from Tab. 2.2 that several RNAs are different domains of the same type of biomolecule, like in the case of the Internal Ribosome Entry Site (IRES), or modules with notorious same

RNA_1	RNA_2	N_1	N_2	SID	PSI	Similarity Score
1KP7_A	1P5P_A	30	77	0.600	0.433	0.600
1KP7_A	2NBX_A	30	108	0.533	0.700	0.700
1P5P_A	2NBX_A	77	108	0.519	0.221	0.519

Table 2.3 Non-pseudoknotted IRES domain pairs retrieved from the selected list of viral RNAs. For each RNA pair we report the chain identifiers, lengths, Sequence Identity SID, Percentage of Structural Identity PSI and Similarity Score = max(SID, PSI).

function, like xrRNAs. In particular, we report six domains of IRES modules of different viruses (RNAs 1KP7_A, 1P5P_A, 2N8V_X, 2NBX_A, 3T4B_A and 4C4Q_N) and three different xrRNA modules (4PQV_A, 5TPY_A and 7K16_P) from as many viruses.

IRESs

Viruses require the host cell to synthesize viral proteins. Among the strategies employed to hijack the host cellular translation, *Internal Ribosome Entry Sites* (shortened IRESs) are often used by viruses to ensure that viral translation is active when host translation is inhibited. These motifs are employed by several viruses to bypass the cap-dependent translation mechanisms involving external molecular complexes [109–111]. In this way, IRES-driven translation has fewer constraints for the canonical translation initiation factors. Due to their crucial role in infection, IRESs are desirable targets for therapeutic drugs [112].

The first discovery of the presence of such motif in the genome of some viruses occurred in 1988 in the *poliovirus* (PV) [113]. After that, tens of viruses were discovered to employ the same mechanism. To name a few notable examples, the *Hepatitis A* and *Hepatitis C* viruses HAV/HCV, the *Rhinovirus* and the *Classical swine fever virus* CSFV all employ IRES modules to initiate the viral translation [114–117].

Within the selected RNAs, we retained different domains of the IRESs of the *Hepatitis C virus* HCV (RNAs 1KP7_A, 1P5P_A and 3T4B_A), a pseudoknotted domain of the IRES of the *Israeli acute paralysis virus* IAPV (RNA 2N8V_X), a portion of the IRES of the *Encephalomyocarditis virus* EMCV 2NBX_A, and finally a large IRES domain of the *Classic swine fever virus* CSFV (RNA 4C4Q_N). Three of these RNAs are not pseudoknotted, namely RNAs 1KP7_A, 1P5P_A and 2NBX_A while the remaining three structures have all a different number of pseudoknots.

We will present now the similarities between the non-pseudoknotted IRES domains, which are in principle (topologically speaking) the closest molecules of this group. We observe that the Similarity Scores of the three non-pseudoknotted RNA pairs are compatible with the range reported in Fig. 2.6, see Tab. 2.3. The largest similarity is found between the RNAs labeled 1KP7_A (30 nucleotides) and 2NBX_A (108 nucleotides), where the Sequence Identity SID = 0.533 and the Percentage of Structure Identity PSI = 0.700, for a resulting Similarity Score = 0.700. The other alignments yield a Similarity Score closer to the mean and median values observed in Sec. 2.2.6, where these values read 0.610 and 0.586, respectively.

xrRNAs

Exoribonuclease resistant RNAs (shorthanded xrRNAs) are structured RNA modules located at the $3'$ untranslated region (UTR) in the genome of different viruses from the *Flaviviridae* family. These viruses contain positive-sense single stranded RNA (+ssRNA) consisting of a single genomic RNA. Notable examples are *Zika virus* ZIKV, *Dengue virus* DENV, *Murray Valley encephalitis virus* MVEV, *Tamana Bat virus* TABV, and *Powassan virus* POWV, which are all major human health threats [118–123]. As aforementioned, these RNA modules flank

RNA_1	RNA_2	N_1	N_2	SID	PSI	Similarity Score
4PQV_A	5TPY_A	68	71	0.676	0.647	0.676
4PQV_A	7K16_P	68	51	0.549	0.647	0.647
5TPY_A	7K16_P	71	51	0.588	0.549	0.588

Table 2.4 xrRNAs pairs retrieved from the selected list of viral RNAs. For each RNA pair we report the chain identifiers, lengths, Sequence Identity SID, Percentage of Structural Identity PSI and Similarity Score = max(SID, PSI).

the reading frame of the viral RNA at the 3′-end, conferring an exceptional resistance to the RNA degradation by means of *exoribonucleases*, enzymes that degrade RNA by cleaving the terminal nucleotides from either ends [61–65]. Such resistance is general, that is a single xrRNA module is able to halt diverse nucleases, revealing a non-specific response to degradation that must be traced back solely to the structure of this module [66]. A fascinating characteristic of the architecture of xrRNAs is a conserved ring-like structure that encircles the 5′-end with a single [62–64] or double loop [65]. The circular motif acts as a molecular brace towards the translocation in the $5' \rightarrow 3'$ direction. Interestingly, this topological feature of xrRNAs was observed also in plant-infecting viruses, which are phylogenetically unrelated to flaviviruses [67].

We retrieved in our list of RNAs three different xrRNA molecules, namely RNAs 4PQV_A (from MVEV), 5TPY_A (from ZIKV) and 7K16_P (from TABV). Although the biological function is the same across the structures, their folds present notable differences. The MVEV xrRNA 4PQV_A comprises in fact only one fully formed pseudoknot, while the other two xrRNAs display a more complex fold with multiple putative pseudoknots.

The first pair, 4PQV_A (68 nucleotides) and 5TPY_A (71 nucleotides) has the largest sequence similarity SID = 0.676 and structural similarity PSI = 0.647 across the alignments among xrRNAs. However, the two structures have different topological complexity: this observation is corroborated by the experimental work made to solve the structure of 5TPY_A, where these two structures were compared to each other [62]. In particular, it was observed in the 4PQV_A a significant change in the position of the terminal hairpin at the 3′-end, which occurs also in the alignment performed with SARA. Furthermore, it is explicitly stated in [62] that an important pseudoknot in the MVEV xrRNA 4PQV_A is not formed, contrary to the fully pseudoknotted ZIKV xrRNA. This is linked to the localization of these xrRNA modules. In fact, xrRNA are often displayed in series in the UTR of flaviviruses; in particular, the former MVEV xrRNA is *downstream* of two in series, while the latter ZIKV xrRNA is *upstream* of two. Upstream xrRNAs are predicted to have a more stable fold, whose loss has a greater effect on viral infection [124, 125]. Thus, these structures topological differences are reported also experimentally: we retain both the molecules within our list.

The second pair of xrRNAs is made of the MVEV and TABV xrRNAs 4PQV_A and 7K16_P (51 nucleotides), respectively. The Sequence Identity SID = 0.549, while the Percentage of Structure Identity reads PSI = 0.647. The SID is smaller compared to that of the previous pair, while the PSI is comparable. However, similarly to the previous case, the two xrRNAs have different topological complexity and their overall similarity does not exceed the selected threshold. Hence, we retain both structures within our list.

Finally, the third xrRNA pair comprises the ZIKV and TABV xrRNAs 5TPY_A and 7K16_P, respectively. This is the most diversified RNA pair, with Sequence Identity SID = 0.588 and Percentage of Structural Identity PSI = 0.549, despite the fact that the chains belong to the same topological group. Again, their overall similarity does not exceed the threshold set in the previous Sections and we thus retain both structures within our list.

2.4 Summary

In this Chapter, we selected a set of 16 non-redundant structures to perform a systematic study on the translocation properties of viral RNA. The 16-chain set was constructed by analyzing and filtering for our purposes a non-redundant database of viral RNAs recently released [79]. We selected actually viral, relatively compact and with a self-contained secondary structure. Sequence and structure redundancies were filtered out to encompass a wider set of sequences and architectures, respectively.

The source organism of the chains was first assessed with a text-based analysis on the RNA-NRD database used as ground base for the construction of the list. The molecules were then analyzed structurally, evaluating their compactness by monitoring their R_g and then verifying that the phosphate atoms were displaced along the chain at constrained distances from each other. The secondary structures were analyzed with x3DNA/DSSR to keep molecules with self-contained secondary structure, and RNAs with intercalated molecules were filtered out. The classification of non-redundant chains made by RNA-NRD did not keep into account redundancies across RNAs sourced from different organisms. Thus, in order to obtain a heterogeneous set of structures and sequences, we reassessed sequence and structural homologies with ClustalOmega [88, 89] and SARA [90], respectively. Then, we retained only molecules with exposed ends, in order to facilitate their translocation through the nanopore. Finally, we clustered together the RNAs of different topological complexity grouping them according to their number of pseudoknots. After grouping the molecules, we aligned batches of RNAs with similar topology with RNAHugs [94] and then clustered similar RNAs to select a representative for each similarity cluster formed.

The formed set encompasses a minimal and heterogeneous ensemble of sequences and structures of diverse topological complexity, with RNAs comprising up to four pseudoknots. In this way, we can perform a systematic study on the translocation properties of viral RNA.

Chapter 3

Simulations Setup

In this Chapter we will report the setup of the simulations and how the various parameters like temperature, pulling force and size of the nanopore were calibrated. The setup used resembles that of Refs. [45, 47] where the ZIKV xrRNA (present within our list as well) was translocated through a nanopore whose size resembles that of the lumen of XRN-1 exoribonuclease, an enzyme that degrades nucleic acids like xrRNAs by cleaving the end nucleotides processing the biomolecules in the $5' \to 3'$ direction. In order to account for the heterogeneity of the viral RNAs selected in Chapter 2, we tuned the nanopore parameters as the optimal compromise between computational overhead and overall setup hindrance, while maintaining the sizes in the same range of previous works from our group to facilitate the comparison between different setups. Likewise, the forces used in the simulations cover the same range of forces employed in those works.

3.1 Force Field and Software

3.1.1 The SMOG Force Field

Interactions

The intramolecular interactions between the atoms of each RNA were based on SMOG, an implicit-solvent force field which is native-centric, retaining atomistic structural details of each structure [126]. The potential energy is given by

$$\begin{aligned} V = &\sum_{\text{bonds}} \epsilon_r (r - r_0)^2 + \sum_{\text{angles}} \epsilon_\theta (\theta - \theta_0)^2 + \sum_{\text{impropers/planar}} \epsilon_\chi (\chi - \chi_0)^2 \\ &+ \sum_{\text{backbone}} \epsilon_{\text{backbone}} F_{\text{D}}(\phi) + \sum_{\text{sidechain}} \epsilon_{\text{sidechain}} F_{\text{D}}(\phi) \\ &+ \sum_{\text{contacts}} \epsilon_{\text{contacts}} \left[\left(\frac{\sigma_{ij}}{r} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r} \right)^{6} \right] + \sum_{\text{non-contacts}} \epsilon_{\text{non-contacts}} \left(\frac{\sigma_{\text{non-contacts}}}{r} \right)^{12} \end{aligned} \tag{3.1}$$

where $F_{\text{D}}(\phi) = (1 - \cos(\phi - \phi_0)) + \frac{1}{2}(1 - \cos(3(\phi - \phi_0)))$ accounts for sidechains rotations. The functional form of Eqn. (3.1) accounts for bonded and non-bonded interactions, angular and dihedral terms that stabilize the structure of each molecule. The structural parameters r_0, θ_0, χ_0, ϕ_0 and the distances σ_{ij} between contacting atoms i and j are extracted from the native conformation, which minimizes the potential V. The repulsive non-contact parameter is set to $\sigma_{\text{non-contacts}} = 1.7\,\text{Å}$ by default. Notice that in the SMOG framework only heavy atoms (that is, all atoms but the hydrogens) are considered.

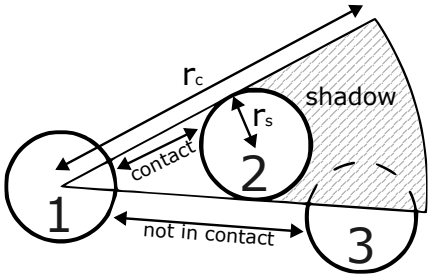


Figure 3.1 Schematics of the contact map of SMOG. If the objective is to find the contacts made by atom 1 and their distances σ_{1j} within the cutoff bubble with radius r_c, an analogy with a light source placed on such atom can clarify the mechanism behind the "Shadow" contact map. Let then a light bulb in correspondence of atom 1. All other atoms have a shadowing radius r_s and are opaque, that is they cast a shadow behind them. All atoms which are illuminated by atom 1 and that are not shadowed by any surrounding atom are considered *in contact* with atom 1. In this case, atoms 1 and 2 are in contact with contact distance $\sigma_{12} < r_c$, contrary to atoms 1 and 3. Image adapted from Ref. [127].

We used the "Shadow" contact map of SMOG, as it includes relevant contacts at constrained distances σ_{ij} without introducing nonphysical contacts between spatial next-nearest nucleotides, that is *second* closest residues relative to the reference nucleotide. The various σ_{ij} values are retrieved from the molecule structures as follows: given the cutoff radius r_c, the atoms in contact with atom i are determined as all the atoms within the cutoff distance which do not have any intervening atom of radius r_s between i and themselves. The parameter r_s scales the effect of the shadowing of the contacting atoms within the cutoff r_c, see Fig. 3.1.

This type of contact map yields a better account for the base stacking interaction compared to the simpler all-atom, cutoff-based contact map, which tends to disproportionally emphasize them when compared to Watson-Crick pairs and tertiary contacts [127]. In particular, it was shown in Ref. [128] that for MD simulations of of the mRNA SAM-I riboswitch folding, in order to achieve an appropriate balance between secondary and tertiary structure energetics it was necessary to rescale the base-stacking interactions by a factor of 3 if a simpler cutoff-based contact map during the MD simulations. On the other hand, this 1/3 scaling in the stacking interactions occurs naturally when employing the "Shadow" contact map, with no need of further tuning the interaction parameters *a posteriori* [127].

The cutoff and shadowing radii were set to $r_c = 5\,\text{Å}$ and $r_s = 1\,\text{Å}$, respectively. The various interaction strengths ϵ's (for bonds, angles, dihedrals, backbone, sidechain, contacts and non-contacts) reported in Eqn. (3.1) are not fixed *a priori*, but they are the result of an internal balance of contacts, backbone and sidechain energies computed by SMOG, namely

$$\begin{cases} \epsilon_{\text{backbone}} / \epsilon_{\text{sidechain}} = R_{\text{BB/SC}} \\ \sum \epsilon_{\text{contacts}} \Big/ \left(\sum \epsilon_{\text{backbone}} + \sum \epsilon_{\text{sidechain}} \right) = R_{\text{C/D}} \\ \sum \epsilon_{\text{contacts}} + \sum \epsilon_{\text{backbone}} + \sum \epsilon_{\text{sidechain}} = \epsilon N_{\text{atoms}} \end{cases} \tag{3.2}$$

where ratios $R_{\text{BB/SC}} = 1$ for RNAs and $R_{\text{C/D}} = 2$, and ϵ is the reduced energy unit. In particular, the first constraint in Eqn. (3.2) fixes the scaling between backbone and sidechain contacts; the second one fixes the ratio of total contact energy to total dihedral energy; finally, the third one sets the total stabilizing energy. In this way, when the cutoff parameters of the contact map are changed, the total energy contribution of the contacts remains constant at $\frac{2}{3}\epsilon N_{\text{atoms}}$ even though the number of contacts may change, varying the energy per contact accordingly.

Then, the other interaction strengths are set as follows: $\epsilon_r = 100\epsilon$, $\epsilon_\theta = 20\epsilon$, $\epsilon_\chi = 10\epsilon$, and $\epsilon_{\text{non-contacts}} = \epsilon$. Basically, once the energy scale ϵ is set for the system, the constraints reported in Eqn. (3.2) scale the various interactions accordingly for each molecule in order to guarantee both the extensivity of total energy and the correct balance between bonded and non-bonded terms.

Finally, a truncated-and-shifted Lennard-Jones potential was used for the excluded volume interactions with the pore and slab walls.

Pros and Cons of Implicit Solvent Simulations and SMOG Force Field

SMOG is a suitable choice to model the RNA interactions and stability for the translocation simulations, since any deviation from the native structure is energetically penalized by the structure-base Hamiltonian, see Eqn. (3.1): the energy minimum is associated to the native conformation of the molecule. This provides an appropriate proxy for translocation processes, during which RNA chains are distorted when threaded through the nanopore, departing from their native fold.

As a native-centric force field, various types of interactions like interaction with ions and transient tertiary contacts between bases and atoms are effectively absorbed and accounted in the Hamiltonian of Eqn. (3.1). Furthermore, SMOG ignores chemistry- and sequence-specific energetics like the GC-content, which were shown to affect the translation and stability of mRNA molecules [129]. Also, the force field accuracy hinges on the quality of the structure, and unstructured/flexible portions of the RNA chains may be overstabilized by the structure-based Hamiltonian.

3.1.2 LAMMPS

We performed implicit solvent Langevin simulations using the LAMMPS simulation software [130]. The simulations were carried out in the overdamped regime at constant temperature T. For each of the N particles, whose position is described by the coordinate vector $\mathbf{x} = \mathbf{x}(t)$, the dynamics is governed by the Langevin equation

$$\gamma \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = -\boldsymbol{\nabla} V + \mathbf{R}(t) \tag{3.3}$$

where V is the potential, $-\boldsymbol{\nabla} V$ is the corresponding force, γ is the damping constant and $\mathbf{R}(t)$ is a Gaussian white noise where

$$\begin{cases} \langle R_i(t) \rangle = 0 \\ \langle R_i(t)\, R_j(t') \rangle = 2\gamma k_\mathrm{B} T \delta_{ij} \delta(t - t') \end{cases} \tag{3.4}$$

where $R_i(t)$ is the i-th Cartesian component of the vector $\mathbf{R}(t)$ at time t, k_B is the Boltmzann constant and T is the temperature of the system.

We set uniform atomic masses ($m = 16\,\mathrm{g/mol}$) to all atoms in our system. In the overdamped regime considered here, the choice of the atomic masses is not expected to significantly affect the overall behavior nor the stability of the RNA folds. Notice that for the simulations, the energetic properties of the system containing the polymer and the surrounding fluid are effectively accounted into the stochastic term. In this way, the implicit-solvent framework gives a great computational advantage by replacing the large number of degree of freedom of the solvent with the stochastic term $\mathbf{R}(t)$, reducing notably the computational overhead compared to their explicit-solvent counterpart. The characteristic simulation time in LAMMPS/SMOG units is equal to

$$\tau_\mathrm{MD} = \sigma_\mathrm{LJ} \sqrt{\frac{m}{\epsilon_\mathrm{LJ}}} = 1.8\,\mathrm{ps} \tag{3.5}$$

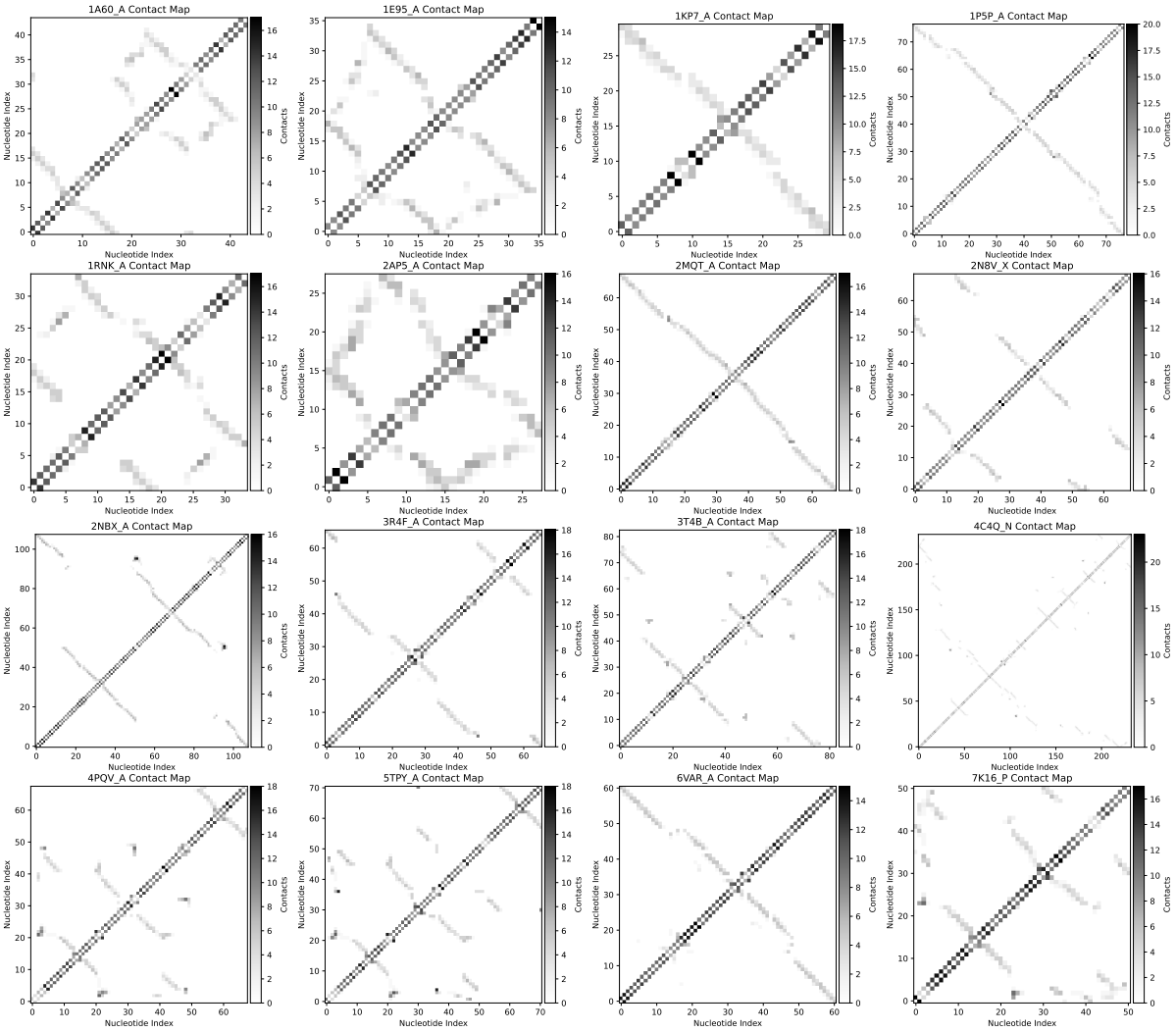


Figure 3.2 The SMOG "Shadow" contact maps for the set of 16 non-redundant RNAs.

where $\sigma_{\mathrm{LJ}} = 5\,\text{Å}$ is the cutoff distance for the LJ interaction between atoms, $m = 16\,\mathrm{g/mol}$ is the atomic mass, and $\epsilon_{\mathrm{LJ}} = 0.30\,\mathrm{kcal/mol}$ represents the characteristic strength of the LJ interaction for non-stacked atoms (in simulation units). These are the values we obtained for non-stacked atoms using the "Shadow" contact map with $r_{\mathrm{c}} = 5\,\text{Å}$ and $r_{\mathrm{s}} = 1\,\text{Å}$ for the set of 16 selected RNAs. The integration step was set to a fraction of this characteristic time τ_{MD}, namely

$$\Delta t = 2\,\mathrm{fs} = 1.11 \times 10^{-3} \tau_{\mathrm{MD}} \tag{3.6}$$

The damping coefficient of Eqn. (3.3) was set to the default value $\gamma = 5 \times 10^{-4}\,\mathrm{g/(mol\,fs)} = 9.09 \times 10^{-7}\,\mathrm{g/(mol\,\tau_{MD})}$, while the Boltzmann constant $k_{\mathrm{B}} = 1.987 \times 10^{-4}\,\mathrm{kcal/(mol\,K)}$.

3.2 Priming the RNAs

For each simulation, the RNAs were primed by the terminal phosphate near the pore entrance so that each chain can make its way through the pore once pulled. The priming process was initially performed by positioning the leading phosphate of each RNA at the pore entrance in such a way that the pore axis and the separating vector $\mathbf{w}$ obtained in Sec. 2.2.7 coincide. In this way, the distances between the phosphate atoms and the nanopore are maximized.

However, steric clashes between the so-positioned RNAs and the simulation nanopore setup occurred often, as the perceptron algorithm was run considering only the phosphate atoms for simplicity. Non-phosphate atoms of the chains may in fact fall inside the slab once the priming process is done, leading to non-physical starting configurations. Thus, in order to avoid steric clashes with the pore and to facilitate the threading of each RNA inside the pore, each biomolecule was translated in the vertical z-direction in such a way that the leading phosphate is confined within a small cylindrical area (namely, the capture region, see Sec. 3.6) above the pore and coaxial with it and with a height $z \leq 5\,\text{Å}$, in a way that no steric clash occurs.

We noticed that several molecules enter the pore in a hook-like conformation when primed by the P atom at the 5′-end, and this conformation occurs for every RNA in the selected list during the 3′-terminal priming. This happens because the end phosphate connecting the terminal nucleotide and the following nucleotide is not located at the topological end of the strand but it connects those two nucleotides. Thus, in order to treat the two termini on equal footing, we treated the terminal $O_{3'}$ atom of the terminal sugar moiety effectively as a dummy phosphate, so that the pulled atom at the end of the chains is actually located at the topological end of them, see Fig. 3.3.

Finally, for each simulation run, the primed configuration was randomly rotated of an angle $\vartheta \in (0, 2\pi)$ around a randomly chosen axis passing through the primed dummy phosphate, discarding rotations and axes leading to steric clashes with the nanopore. This procedure ensures that no *a priori* bias is introduced due to the initial orientation between the RNA terminal segment and the pore axis. We report in Fig. 3.4 a sample of an accepted primed configuration for the pseudoknotted RNA labeled `2AP5_A`. Notice that in this case the leading phosphate at the 5′-end is located at the actual physical end of the chain, not requiring the priming of the $O_{3'}$ near the nanopore.

3.3 The Temperature

3.3.1 Preparation

In order to find a suitable reference temperature for the simulations, we computed the melting curves for five RNAs of the selected list. These five molecules have a linear length comparable to the average RNA length in our list ($N \approx 70$, see Tab. 2.2) and display quite diverse

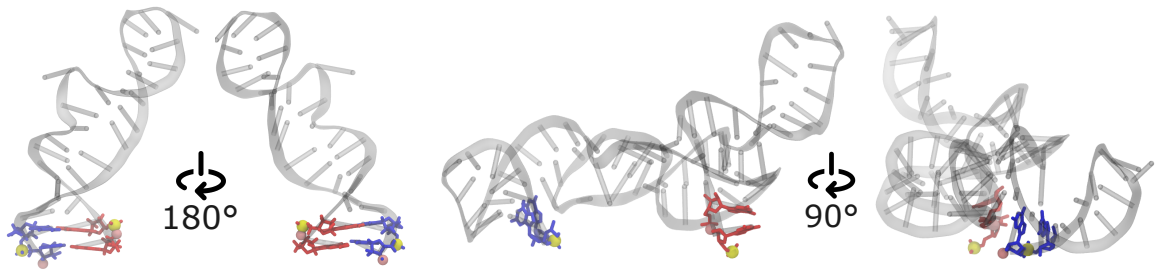


Figure 3.3 Schematic representation of the possible configuration of terminal P and $O_{3'}$, represented as yellow and pink solid beads, respectively. The 5′- and 3′-end are reported respectively in red and blue, following the color scheme of the previous figures. On the left it is represented the stem-loop labeled `1KP7_A`, where the two phosphate atoms connect the two end nucleotides symmetrically, leading to a hook-like conformation once the translocation occurs: the RNA must be primed by the terminal $O_{3'}$ both ends. On the right is reported the ZIKV xrRNA labeled `5TPY_A`, which has a dangling phosphate at the 5′-end, while the 3′ terminal phosphate connects the two terminal nucleotides, creating an asymmetric configuration in the of the phosphate atoms positioning: only the 3′-end is primed using the $O_{3'}$ as leading dummy phosphate: in this way the two pulled ends are equivalently pulled. The RNA renderings were made with VMD [82].

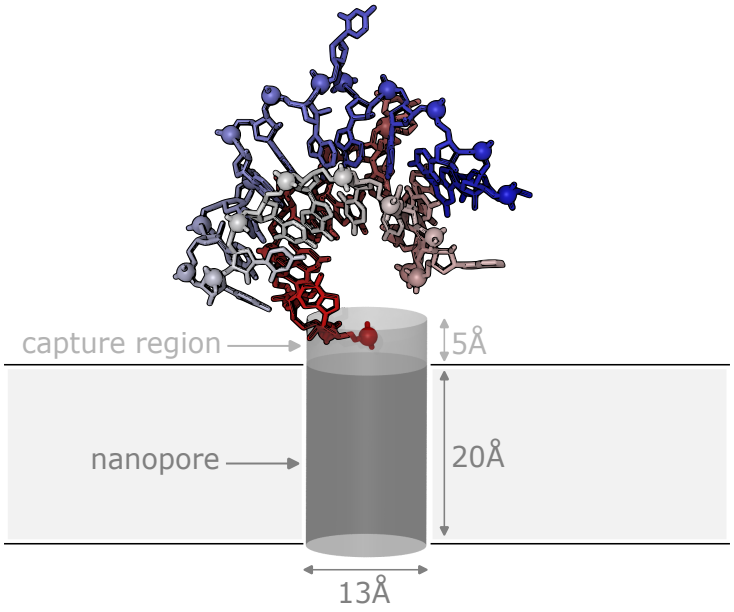


Figure 3.4 Representation of a valid configuration for the primed molecule labeled `2AP5_A` inside a pore of radius $\varrho = 6.5\,\text{Å}$ and length $h = 20\,\text{Å}$. The pore is embedded in an infinite slab of height equal to the pore length. The 5′ and 3′ ends are colored in red and blue, respectively. The phosphate atoms are drawn as solid beads. The capture region, sketched in light gray, extends the pore in the positive z-direction by 5 Å, avoiding steric clashes between the RNA and the nanopore and enabling the capturing of the first (eventually dummy) phosphate. The renderings of the RNA and the nanopore were made with VMD [82].

topologies. The RNAs selected are labeled `1A60_A` (44 nucleotides, made of an H-type pseudoknot in series with a hairpin), `2MQT_A` (a 68-nucleotide long stem-loop), `3R4F_A` (a 66-nucleotide non-pseudoknotted structure with an overall U-shape), `5TPY_A` (a 71-nucleotide xrRNA with multiple pseudoknots) and `6VAR_A` (a 61-nucleotide non-pseudoknotted structure with an overall Z-shape).

The melting curves were obtained by scanning several temperatures from $T = 50\,\mathrm{K}$ to $T = 145\,\mathrm{K}$ with a 5-degree step (in LAMMPS/SMOG units), running 10 independent trajectories per temperature for each RNA. In order to set the duration of the melting trajectories, we inspected first the root-mean square deviation (RMSD) of each molecule in bulk at $T = 120\,\mathrm{K}$ in LAMMPS/SMOG units. In this way, it is possible to estimate the minimum time necessary to relax such observable. The RMSD curves are reported in Fig. 3.5. The RMSD of a molecule is a measure of its deviation from a reference conformation. For our purposes, the RMSD was computed at each frame as the deviation of the $C_{3'}$ atom positions of each RNA residue with respect to the last frame of each trajectory. The last frame was used as reference for the calculation because for sufficiently long trajectories the reported conformation is equilibrated, while the structure reported in the crystal file (reported in the first frame of the trajectory) is out of equilibrium. Thus, given for an N-residue molecule the set of $C_{3'}$ atoms coordinates at time t, $\mathbf{X}(t) = \{\mathbf{x}_1(t), \dots, \mathbf{x}_N(t)\}$, the RMSD with respect to the reference set of coordinates $\mathbf{X}^{\mathrm{ref}} = \{\mathbf{x}_1^{\mathrm{ref}}, \dots, \mathbf{x}_N^{\mathrm{ref}}\}$ is

$$\mathrm{RMSD}\left(\mathbf{X}(t), \mathbf{X}^{\mathrm{ref}}\right) = \min_{\text{roto-translation}} \left(\sqrt{\frac{1}{N} \sum_{i=1}^{N} \left|\mathbf{x}_i(t) - \mathbf{x}_i^{\mathrm{ref}}\right|^2} \right) \tag{3.7}$$

where the minimization term $\min_{\text{roto-translation}}(\cdots)$ refers to the fact that for each calculation of the RMSD, the molecule is superimposed to the reference structure with the roto-translation that minimizes its value. This ensures that the RMSD accurately reflects the actual structural similarity of the two conformations regardless of the positioning of the molecules, reflected in the absolute value of their coordinates.

The autocorrelation time of the slowest variable in a physical system dictates the lower bound for the duration of its thermalization: for our purposes, in order to have a set of sound thermalized melting curves, the trajectories should be much longer than the autocorrelation time, so that the slowest observables (and thus also the faster ones) in the system are properly relaxed. We observe from the trajectories reported in Fig. 3.5 that the RMSD presents detectable fluctuations on a timescale of $10^5 \tau_{\mathrm{MD}}$ for all the selected RNAs. The RMSD was computed from the final configuration of each simulation as the first ones are the most out of equilibrium, being directly inherited from the structure deposited in each PDB entry. Moving from the tail of each plot towards the left, we notice that the equilibrium value is approached after a timescale of around $10^5 \tau_{\mathrm{MD}}$ for each RNA.

For practical reasons, we could not afford a set of trajectories at various temperatures longer than this for all RNAs, thus the duration of the melting trajectories was set to at least $10^5 \tau_{\mathrm{MD}}$ for all RNAs, which we are aware being inferior to the real equilibration time of the system. Then, we performed 10 bulk simulations for each melting temperature scanned. We computed both the average fraction of contacts $\langle Q \rangle$ and the average root mean square displacement $\langle \mathrm{RMSD} \rangle$ for these runs, obtaining the melting curves which are reported in Fig. 3.6. Again, the RMSD was computed with respect to the last configuration of each RNA along each trajectory. The contact fraction of two atoms (i, j) at distance σ_{ij} was computed using the COLVARS module [131] using the sigmoidal weight previously employed in Ref. [45]:

$$q(\sigma_{ij}) = \frac{1 - (\sigma_{ij}/(1.5\,\sigma_{ij}^0))^6}{1 - (\sigma_{ij}/(1.5\,\sigma_{ij}^0))^{12}} = \frac{1}{1 + (\sigma_{ij}/(1.5\,\sigma_{ij}^0))^6} \tag{3.8}$$

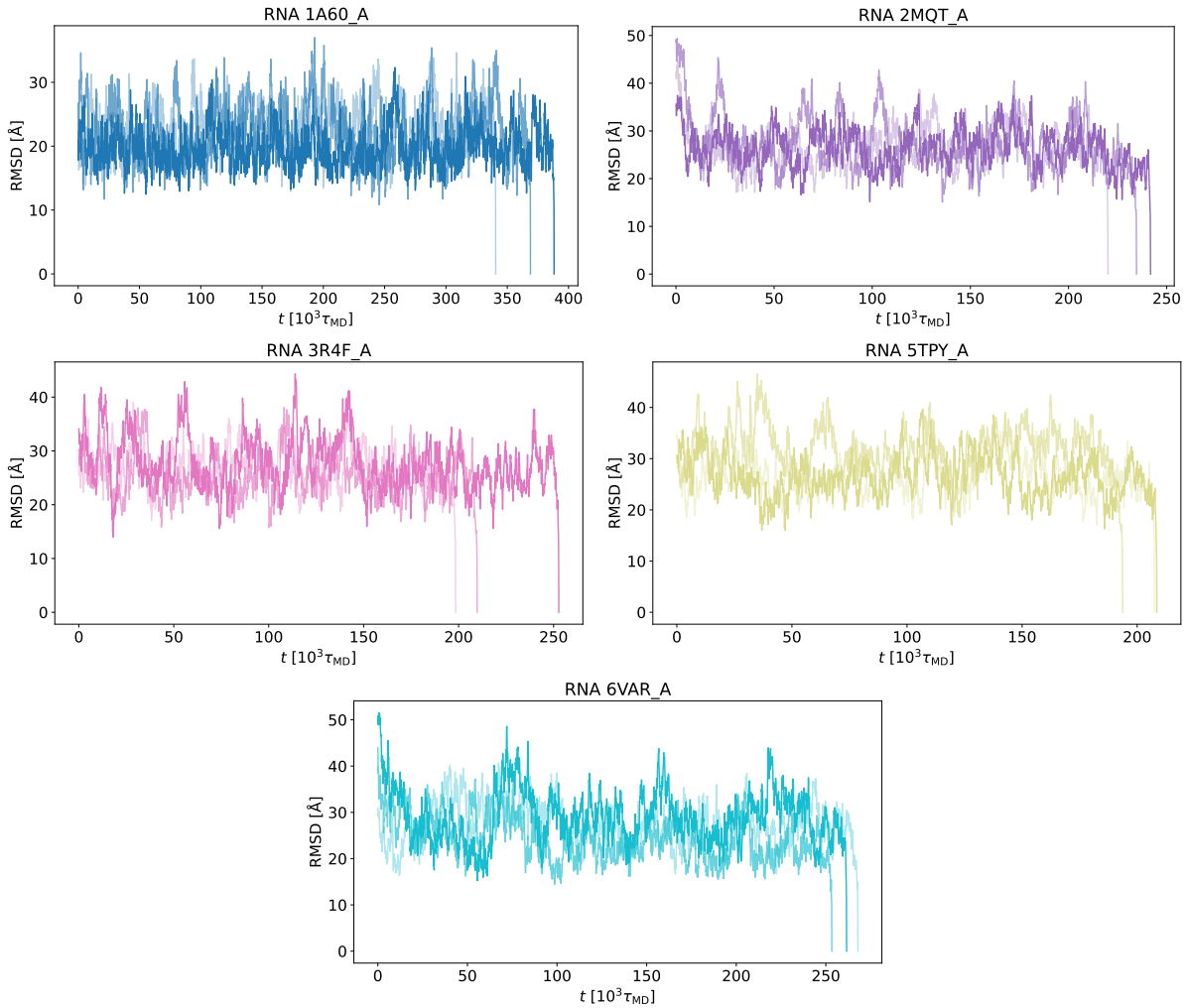


Figure 3.5 Bulk simulations of the five selected RNAs at $T = 120\,\mathrm{K}$ in LAMMPS/SMOG units. We report for each RNA only three trajectories out of the 10 simulated to make the plot less crowded and more clear. From the plots we observe how the fluctuations of the RMSD occur on a timescale of order $10^5\tau_{\mathrm{MD}}$.

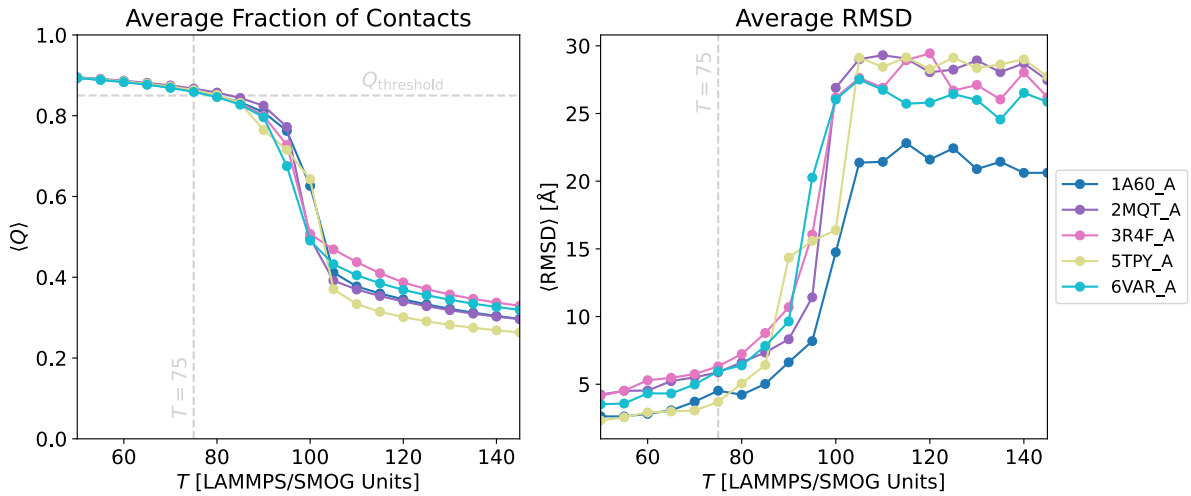


Figure 3.6 On the left, the melting curves of five selected RNAs from our minimal list, labeled `1A60_A`, `2MQT_A`, `3R4F_A`, `5TPY_A` and `6VAR_A`. On the right, their ⟨RMSD⟩ as a function of temperature computed for the bulk simulations at different temperatures. The selected simulation temperature $T = 75\,\mathrm{K}$ (in LAMMPS/SMOG units) is reported as a vertical line in the two plots.

where σ_{ij}^0 is the distance between contacting atoms i and j, obtained from the crystal structure of each RNA as in Eqn. (3.1) by the SMOG force field software with the "Shadow" contact map. Then, the total fraction of contacts Q was calculated by averaging $q(\sigma_{ij})$ over all the N contact pairs retrieved from the SMOG contact map:

$$Q = \frac{1}{N} \sum_{i<j} q(\sigma_{ij}) \tag{3.9}$$

Finally, the average fraction of contacts, that is $\langle Q \rangle$, was obtained by averaging the values of Q described above over the collected trajectories.

3.3.2 Temperature Calibration

In order to find the appropriate temperature for the simulations, we first set a threshold on the minimum fraction of contacts to be preserved for the RNAs. The threshold was set in such a way that Q approaches the typical plateau observed in the melting curves. This heuristic choice was dictated by the fact that the molecules are expected to be compact and folded when simulated in bulk water at the reference simulation (room) temperature if no external force is applied on them. Given the melting curves in Fig. 3.6, we set $Q_{\mathrm{threshold}} = 0.85$, so that $T < T_{\mathrm{melt}}$ and the fraction of contacts approaches the plateau value; after that, we picked as room temperature the *largest* simulation temperature which satisfies $\langle Q \rangle > Q_{\mathrm{threshold}}$ for all RNAs, to avoid excessive suppression of thermal fluctuations. From the melting curves in Fig. 3.6 we observe that the largest temperature satisfying this constraint is $T = 75\,\mathrm{K}$.

Thus, the simulation temperature is a *purely operational* choice, dictated by the requirement of compact molecules and reasonable thermal fluctuations (as the simulated system should not be frozen), rather than a physics-motivated one. The actual temperature must not be assumed as a realistic choice, since a (linear) mapping to the real room temperature $T = 300\,\mathrm{K}$ would make the melting temperatures non-realistic and not comparable with the predictions made by DINAMELT [132], a two-state model prediction tool for nucleic acids, which we report in Tab. 3.1. Likewise, matching the obtained melting temperatures to the DINAMELT predictions

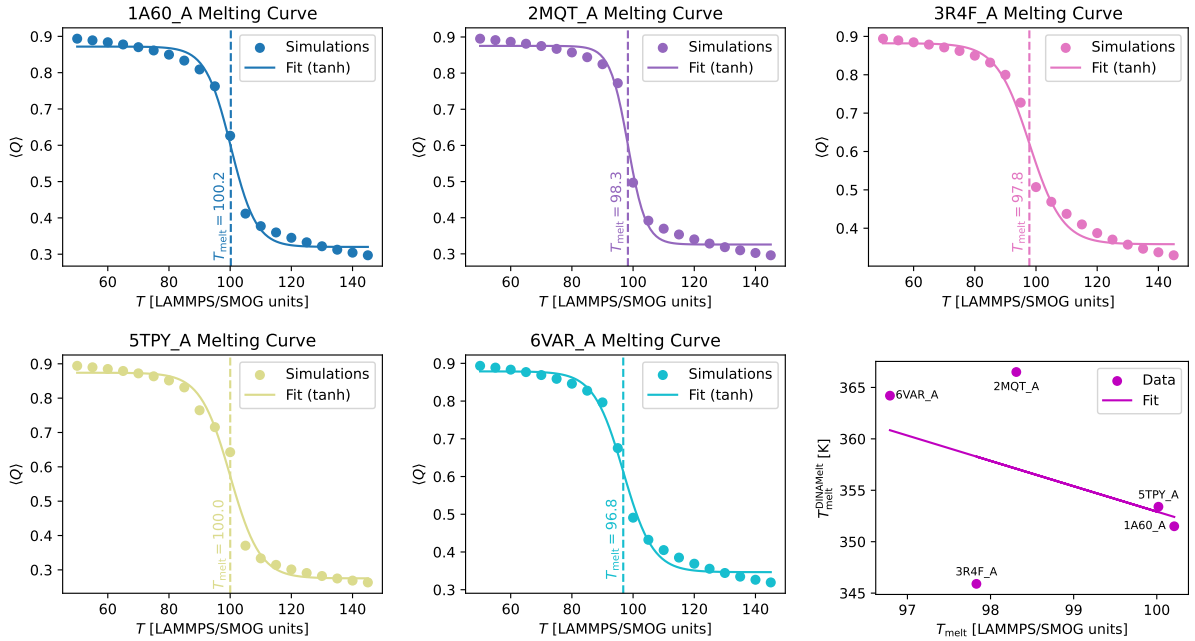


Figure 3.7 The top rank and the bottom, center-left panels contain the sigmoid melting curves for the five RNAs used to calibrate the simulation temperatures. The data are the same of the right panel of Fig. 3.6 and the datapoints of each RNA were fitted via a hyperbolic tangent function to extract the melting temperature. In the bottom-right panel are reported the computed melting temperatures with the predictions made by the DINAMELT server, showing an anticorrelation between such temperatures.

RNA	1A60_A	2MQT_A	3R4F_A	5TPY_A	6VAR_A
T_{melt}	100.21	98.31	97.83	100.02	96.79
DINAMELT [K]	351.5	366.5	345.9	353.4	364.2

Table 3.1 Melting temperatures obtained by fitting the melting curves of the five sample RNAs and the corresponding predictions made with DINAMELT.

would result in non-realistic room temperatures. From the panels of Fig. 3.7 for the sample RNAs we observe how the melting temperatures obtained from the simulations and from the DINAMELT predictions are anticorrelated: the SMOG model is not easily transferable to a heterogeneous set of RNA molecules maintaining the same parameters.

We report that in the simulations at $T = 75\,\text{K}$ the RMSD of the five RNAs is constrained within 6.5 Å, see the bottom-right panel of Fig. 3.6. Adopting less stringent thresholds for the fraction of native contacts would have led to larger simulation temperatures. Since the temperature calibration was performed using a subset of the RNAs introduced in Chapter 2 (specifically, 5 out of 16 RNAs simulated in bulk), employing a relatively low simulation temperature (*i.e.*, $T = 75\,\text{K}$ instead of $T = 80\,\text{K}$ or $T = 85\,\text{K}$) guarantees the generation of more structurally compact ensemble, avoiding excessively flexible molecules.

3.4 The Nanopore

Each RNA was translocated through a 13 Å wide and 20 Å deep nanopore that runs perpendicularly in the z-direction of a 20 Å thick infinite slab in the (x, y)-plane, which is sketched in

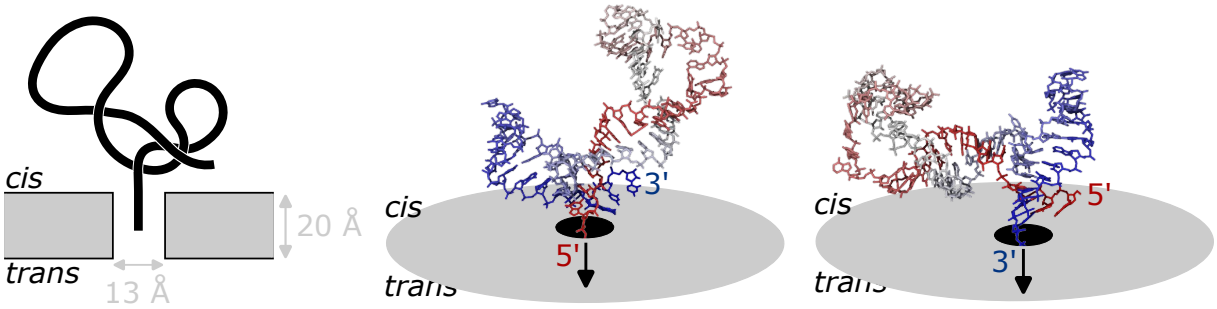


Figure 3.8 On the left, a sketch of the system setup. An infinite slab parallel to the (x, y) plane contains a 13 Å wide and 20 Å deep nanopore through which threaded RNAs translocate by means of an external force. In the center and on the right are reported two typical configurations for the simulations when pulling the RNA by the 5′- and 3′-end, respectively. The depicted molecule is the RNA labeled `3R4F_A`, one of the multimers of the pRNA of ϕ29 packaging motor.

Fig. 3.8.

The width of the pore is comparable to that at the narrowest point of biological pores which are usually employed in translocation experiments. For example, the narrowest region of α-HL has a width of about 15 Å, whereas the constriction in *Mycobacterium smegmatis* porin A (MspA) measures roughly 12 Å. Both channels exhibit a transverse length of approximately 100 Å [133, 134]. Either case, the fact that the diameter is similar to the size of a single nucleotide makes the two biological pores (and consequently also the chosen nanopore size for our simulations) good candidates for single-stranded chains translocation. In case of structures with larger size like double-stranded RNA molecules (notice dsRNA A-helix structures display a width of around 24 Å [135]), the forced translocation compels the molecule to unzip in order to translocate through the pore, given that the applied force is sufficiently large to overcome the energetic barrier associated with the bonds breaking.

To facilitate the comparison with previous works on our group, we set the pore height to $h = 20$ Å, a value that is similar to the one of the setup inherited from Ref. [45, 47].

3.4.1 Size Calibration

To select the actual pore dimensions, we scanned three different cylindrical geometries with height $h = 20$ Å and radii $r = 6.0$ Å, $r = 6.5$ Å and $r = 7.0$ Å. All these pore radii make the total constriction of the setup comparable to that of the aforementioned biological pores, ensuring that in principle only single-stranded RNAs translocate through the pore.

Then, in order to observe the effect of the pore radius on the forced translocation of the RNAs, we selected three RNAs to translocate through the pore. We selected two RNAs of length comparable to the average length of the selected molecules ($N \approx 70$, see Tab. 2.2) and a long RNA, that is the RNAs labeled `6VAR_A` (a non-pseudoknotted stem-loop molecule of length $N = 61$), `5TPY_A` (an xrRNA decorated with multiple pseudoknots and with linear size $N = 71$) and `4C4Q_N` (the longest RNA in our list, containing $N = 233$ nucleotides). We ran five simulations for each RNA to explore the qualitative effects of the radius on the whole translocation process. Initially, we translocated the three RNAs by both termini with the reference force $F_s = 20F_0$. However, the RNAs `5TPY_A` and `4C4Q_N` did not activate the translocation within a computationally feasible timescale of a few hundreds of hours of CPU time when pulled in the $5' \to 3'$ direction at the reference force (we simulated more than 500 CPU hours, equivalent to 3 weeks of simulations). Thus, in order to collect also the missing data

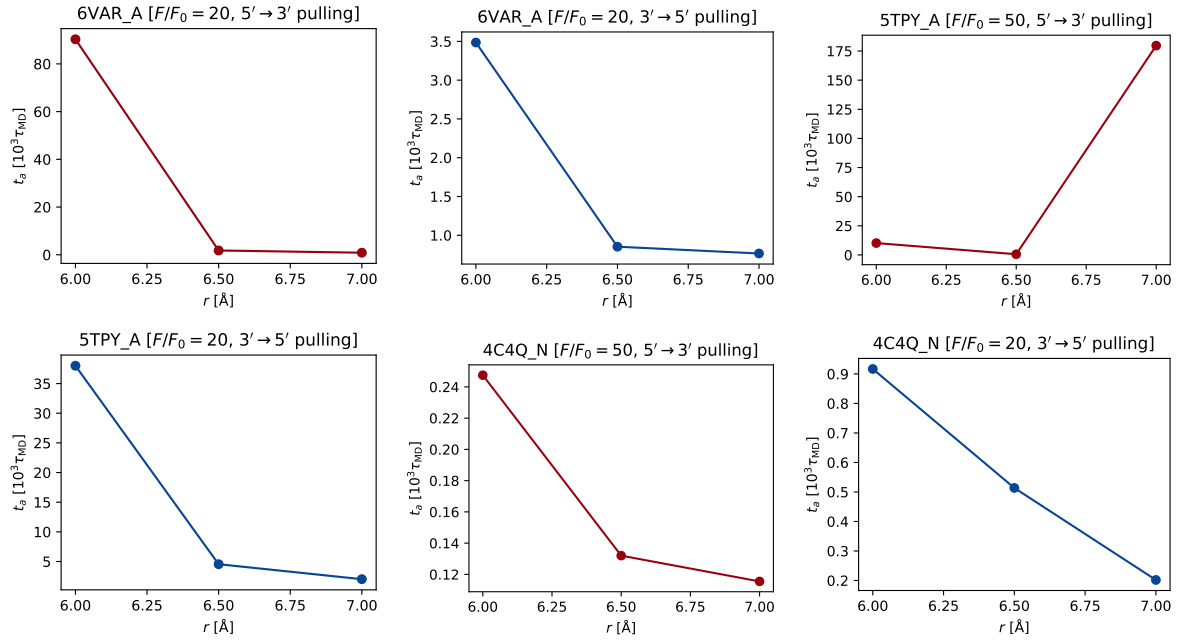


Figure 3.9 In red and blue are reported the median activation times for the $5' \to 3'$ pulling and the $3' \to 5'$ pulling, respectively. Some of the pulling forces were increased with respect to the reference force $F_s = 20F_0$ to make the activation process faster. Notice the largest activation time from the 5′ of RNA `5TPY_A` occurs when the radius is larger, $r = 7.0$ Å. The smaller hindrance of the pore makes it easier for the chains to have multiple portions of the strand within the pore simultaneously, halting the translocation of the chain.

from the 5′-end within few days of simulations, we pulled the RNA molecules labeled `5TPY_A` and `4C4Q_N` with larger forces in the $5' \to 3'$ translocation direction, using $F_s = 50F_0$.

We report in Fig. 3.9 the collected median activation times for the three selected RNAs. We find that, in nearly all instances, the activation time decreases with the pore size, because the reduced hindrance at the pore entrance allows the RNAs to enter more easily. However, in the case of the ZIKV xrRNA `5TPY_A`, we notice that at pulling force $F_s = 50F_0$ the largest activation time is found when the pore radius is $r = 7.0$ Å. This is due to the fact that larger pores enhance the probability of multiple strands to enter simultaneously inside the pore, halting and delaying the translocation of the RNAs. Furthermore, as the force increases (as in this case of `5TPY_A` in the $5' \to 3'$ direction, compared to the translocations in the other direction), once multiple strands are captured inside the pore, the pulling force overcomes the entropic recoil due to the large hindrance in the narrow pore, and the RNA struggles to worm its way inside it.

In fact, as the pore radius increases, the chance of capturing a portion of the same strand trailing the ones which are inside the pore increases as well, altering the natural head-to-tail sequence of (consecutive) nucleotides entering the pore. This is especially true for RNAs which display a complex fold with an intertwined strand (*e.g.*, multiple pseudoknotted structures): due to thermal fluctuations, it is more likely that a lasso-like segment wraps around the pulled portion of the RNA, sucking into the pore multiple strands of the same chain, halting the translocation due to the larger hindrance exerted by the molecule inside it, see *e.g.* Fig. 3.10 for a simulation snapshot of the RNA `4C4Q_N` with multiple portions of the strand inside the pore simultaneously.

Thus, the constriction of a pore of radius $r = 6.5$ Å emerges as the best compromise between the duration of the activation stage and the hindrance exerted by the constriction, reducing the

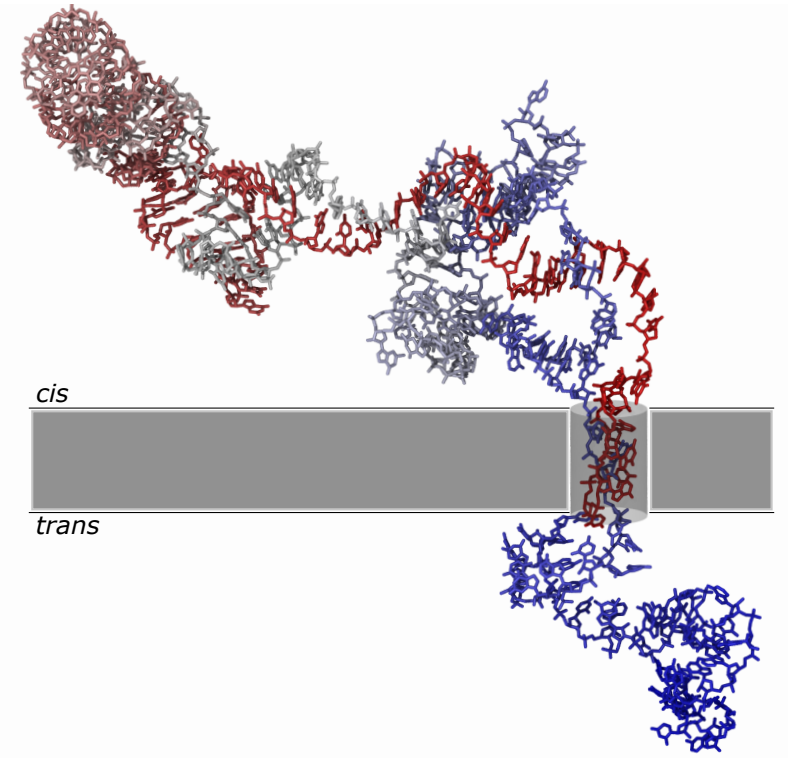


Figure 3.10 Example of a conformation in which multiple portions of the RNA labeled `4C4Q_N` enter simultaneously inside the pore of radius $r = 7.0$ Å. The molecule was translocated in the $3' \rightarrow 5'$ direction, where the $5'$ and $3'$ termini are colored in red and blue, respectively. During the translocation, as the molecule worms its way through the pore, a portion of the RNA around the $5'$-end enters the pore beforehand. The larger the pore, the larger the probability that this type of conformation occurs. The *cis* and *trans* sections are sketched to display the relative location of the pore with respect to the slab. The RNA and nanopore renderings were made with VMD [82].

chances of multiple strands of the same chain to be inside the pore simultaneously.

3.5 Initialization Protocol

As discussed in Sec. 3.3.1, a sound thermalization of the selected RNAs would require trajectories of duration longer than the order of $10^5\tau_{\mathrm{MD}}$. However, this timescale is computationally demanding, given the full set of 16 RNAs and the many combinations of trajectories and pulling end to explore. This heavy computational load is confirmed also by the self-diffusion times reported in Tab. 3.2, which are of the order of several thousands τ_{MD}. Even simulations of duration longer than these autocorrelation times, as already observed from the RMSD curves, are very onerous to carry out.

Thus, in order to *initialize* the tens of initial configurations for the translocation simulations, we based the initialization on another observable faster to relax, that is the autocorrelation time of the gyration radius $\tau_{R_{\mathrm{g}}}$. For all simulations, an initialization run of duration equal to twice the autocorrelation time of the radius of gyration (computed as of Eqn. (2.1)) was performed. Such autocorrelation time was obtained from ten trajectories in bulk water at the selected simulation temperature $T = 75\,\mathrm{K}$ (SMOG/LAMMPS units) for each RNA by integrating the autocorrelation function over the time lags τ as follows:

$$\tau_{R_{\mathrm{g}}} = \int_0^{\tau^\star} \mathrm{d}\tau\, C(\tau) = \int_0^{\tau^\star} \mathrm{d}\tau\, \frac{\left\langle \left(R_{\mathrm{g}}(t) - \langle R_{\mathrm{g}}\rangle\right)\left(R_{\mathrm{g}}(t+\tau) - \langle R_{\mathrm{g}}\rangle\right)\right\rangle}{\langle R_{\mathrm{g}}^2\rangle - \langle R_{\mathrm{g}}\rangle^2} \tag{3.10}$$

where $\tau^\star$ is the time lag such that $C(\tau^\star) = 0.01$. The radius of gyration $R_{\mathrm{g}}(t)$ was computed with COLVARS module of LAMMPS [131], and $\langle R_{\mathrm{g}}\rangle$ is the average radius of gyration, computed over the each trajectory. The various autocorrelation times are reported in Tab. 3.2, and the autocorrelation functions $C(\tau)$ of each RNA are sketched in Fig. 3.11. We observe from Tab. 3.2 that $\tau_{R_{\mathrm{g}}} < \tau_{\text{self-diffusion}}$ by orders of magnitude, meaning that $\tau_{R_{\mathrm{g}}}$ is a much shorter relaxation timescale.

During the initialization runs, a ratchet-like restraint is applied on the leading phosphate to avoid that the molecule fluctuates back in bulk, remaining around the pore entry. The restraint confines the leading phosphate in the same aforementioned cylindrical region coaxial with the nanopore (same radius $\varrho = \sqrt{x^2 + y^2} \leq 6.5\,\text{Å}$ and height $z \leq 5\,\text{Å}$).

3.6 The Force

3.6.1 Setup Details

We considered two different driving protocols, namely one with a static force $F_{\mathrm{s}}(t) = F_{\mathrm{s}}$, and the other with a force that increases linearly with time $F_{\mathrm{r}}(t) = rt$, where r is the constant pulling rate. The results of the simulations for these two pulling protocols are reported in Sec. 4.1 and Sec. 4.2, respectively.

The forced translocation of the biomolecules across the nanopore was performed by applying a force on the phosphate atoms located inside the pore. This is a very simplified mimic of electrophoretic experiments, in which an electric field drives the molecular translocation by acting on the negatively charged phosphate groups [18]. As an approximation, higher-level effects like electro-osmotic flow [136, 137] or leakages of the electric field beyond the pore itself [17, 138] are neglected. Additionally, to facilitate the RNA threading into the nanopore, the leading dummy phosphate is guided into the pore using a capture force equal to the pulling force. This *ad hoc* force is applied only on the leading (dummy) phosphate inside a small cylindrical region positioned above and coaxial with the pore (radius $\varrho = \sqrt{x^2 + y^2} \leq 6.5\,\text{Å}$ and $z \leq 5\,\text{Å}$),

RNA	R_g [Å]	N_{atom}	N_{ntd}	$\tau_{self\text{-}diffusion}$ [τ_{MD}]	τ_{R_g} [τ_{MD}]
1A60_A	21.042	923	44	6119.7	42.0
1E95_A	14.310	767	36	2352.0	18.3
1KP7_A	16.316	634	30	2527.4	89.6
1P5P_A	29.249	1638	77	20984.1	1070.0
1RNK_A	15.702	726	34	2680.4	20.7
2AP5_A	11.862	601	28	1266.3	5.3
2MQT_A	29.592	1449	68	19000.7	264.0
2N8V_X	25.945	1473	69	14847.9	193.0
2NBX_A	31.451	2304	108	34127.5	511.0
3R4F_A	23.429	1400	66	11507.7	413.0
3T4B_A	24.623	1771	83	16078.8	225.0
4C4Q_N	44.454	4986	233	147546.0	676.0
4PQV_A	20.412	1466	68	9146.6	1080.0
5TPY_A	22.094	1527	71	11162.0	218.0
6VAR_A	22.557	1290	61	9828.9	254.0
7K16_P	16.366	1088	51	4363.8	66.2

Table 3.2 Summary of few structural and dynamical observables for each RNA: radius of gyration R_g, number of nucleotides, self-diffusion time of the molecule and autocorrelation time of the radius of gyration. The self-diffusion time is reported for completeness and was calculated as $\tau_{self\text{-}diffusion} = R_g^2/6D_{RNA}$, where the diffusion coefficient of an N-atom RNA reads by the Flory theory $D_{RNA} = D_{16}/N_{atoms}$. The diffusion coefficient of an atom with mass $m = 16\,g/mol$ is $D_{16} = 11.13\,Å^2/\tau_{MD}$, which we computed over 20 simulations in bulk of a single atom with such mass at the selected simulation temperature. We notice the self-diffusion time is tens to hundreds times larger than the autocorrelation time of the radius of gyration.

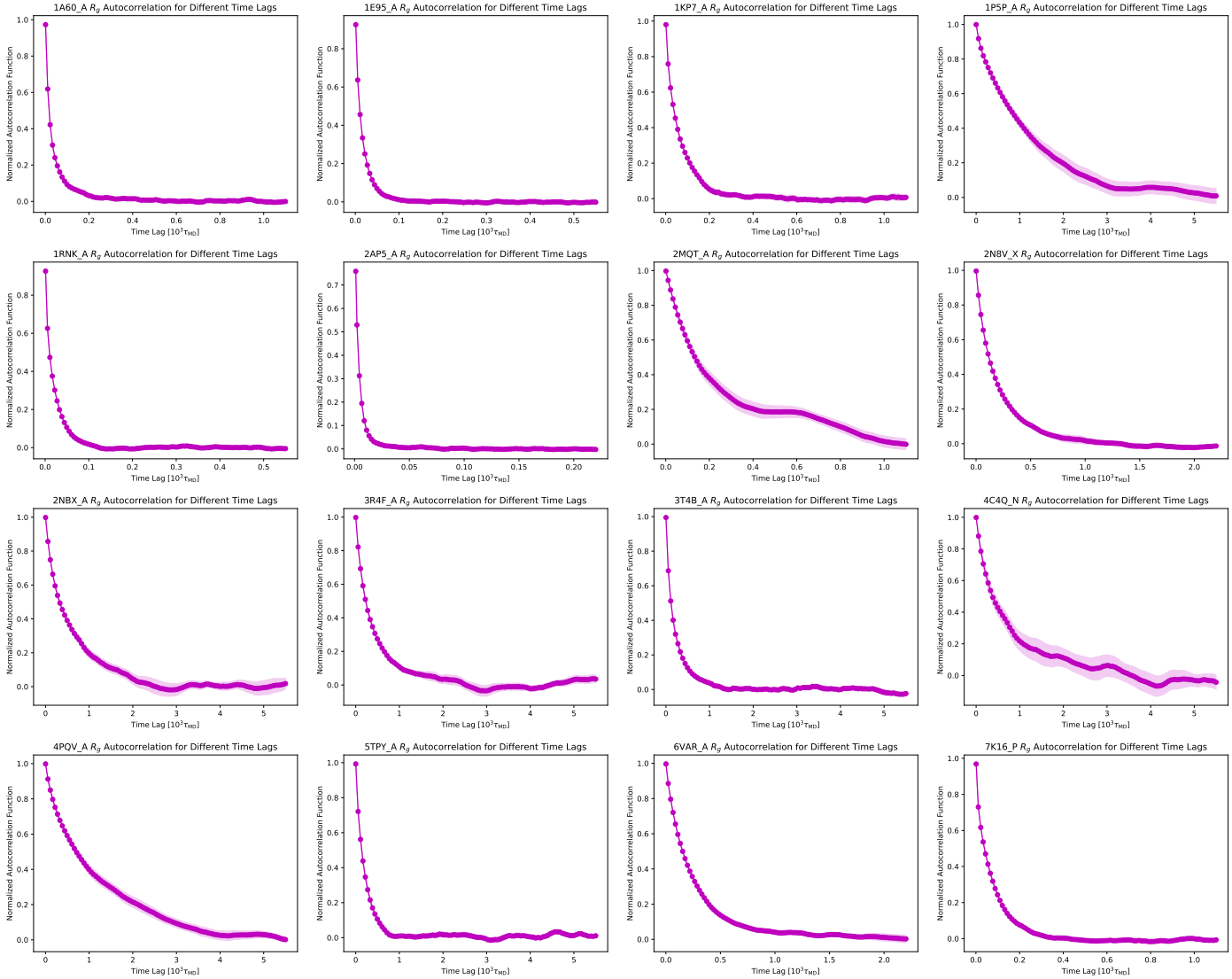


Figure 3.11 Normalized autocorrelation function $C(\tau)$ of the radius of gyration R_{g} as a function of the time lag τ for each RNA. Each shaded area corresponds to the errorbars of $C(\tau)$.

see Fig. 3.4. A ratchet-like restraint is applied continuously throughout the simulations to the leading (dummy) phosphate, preventing its backward retraction into the bulk from the capture region.

However, the number of phosphate atoms inside the pore fluctuates during the translocation process. During the activation stage, the nanopore is progressively filled from none to four nucleotides and as many phosphate atoms. Then, during the translocation stage, such number oscillates between three and five, depending on the extent of the chain stretch. In order to keep the pulling force constant on the translocating molecules, the force $F(t)$ was equally distributed among the phosphates inside the nanopore at every instant t for both pulling protocols.

3.6.2 Force Calibration

The combination of the thermal energy divided by the cutoff distance for the interaction between atoms has units of a force, which we can use as ruler to calibrate the pulling force. Let then the force scale $F_0 = k_\mathrm{B}T/\sigma$ such combination of the thermal energy at room temperature ($T = 75$ in LAMMPS/SMOG simulation units), divided by the cutoff distance for the Lennard-Jones interaction between atoms ($\sigma = 5\,\text{Å}$).

Static Force Protocol

We set the reference static pulling force to $F_\mathrm{s} = 20F_0$, since it allows us to obtain a set of complete translocations that are typically carried out in a range on a simulation scale of at most few hundreds CPU hours, see Sec. 4.1.1. In particular, most of the translocations were concluded within 500 hours of CPU time, equivalent to three weeks of simulations.

Ramping Force Protocol

Regarding the ramping force protocol $F_\mathrm{r}(t) = rt$, we scanned two orders of magnitude for the pulling rates, namely $r = \frac{50F_0}{a\times 10^6}\frac{1}{\text{timestep}}$, where $a = 500,\ 200,\ 100,\ 50,\ 20,\ 10,\ 5,\ 1$ in LAMMPS/SMOG units. In proper MD units, these ramping rates can be recast in the form $\frac{F_0}{22\times a}\frac{1}{\tau_\mathrm{MD}}$, where $a = 500,\ 200,\ 100,\ 50,\ 20,\ 10,\ 5,\ 1$ again.

3.7 Observables

Each translocation run can be divided into three stages, namely the thermalization, the activation and translocation. During the activation stage, the leading atom is brought to the *trans* region of the simulation box after translocating through the nanopore, while during the translocation stage the rest of the chain passes through the pore to reach the *trans* region. The duration of the activation stage is computed through the *activation time* $t_\mathrm{a} = t_\mathrm{start} - t_0$, where t_start is the time elapsed to bring the leading phosphate through the nanopore to the *trans* region, and t_0 is the duration of the relaxation run, which begins at $t = 0$. Notice that given the pore geometry and the typical stretching of the RNAs threaded through it, the first nucleotide reaches the *trans* region approximately as the fifth nucleotide enters the nanopore. For this reason, the nucleotide index $n(t)$ of the residue straddling the pore entrance is constantly recorded throughout each run.

3.8 Summary

In this Chapter, we showed how the overall simulation setup was prepared. We performed implicit-solvent simulations, where the SMOG force field was used to treat the molecular in-

teractions between atoms. This structure-based model is a good proxy for the translocation simulations because it captures the energetic barriers that must be overcome for the RNAs to thread their way into the nanopore. Each molecule was primed near the pore entrance by first placing it close to the pore so as to maximize the distance between the slab and the phosphate atoms, and then randomly rotating it around the leading phosphate to eliminate any potential bias due to the initial entry angle. In order to treat on equal footing the two termini of each RNA and to avoid hooked conformations inside the pore, each molecule was primed near the pore entrance using the $O_{3'}$ atom as dummy phosphate if necessary. The simulation temperature was empirically calibrated on a subset of the selected RNAs by setting a minimum value for their fraction of contacts in order to guarantee folding stability in bulk at the selected temperature. The calibration was done by looking at the melting curves assuring that the selected value is below the melting temperature T_{melt} and the fraction of contacts is in the range of the plateau associated with folded structures. The pore size was dictated as the best compromise between computational feasibility to obtain computationally accessible activation times and minimizing the changes of introducing multiple strands of RNAs simultaneously inside the pore, resulting in a cylindrical geometry with radius $\varrho = 6.5\,\text{Å}$ and height $h = 20\,\text{Å}$. This constriction resembles that of biological pores usually employed in translocation experiments. The initialization protocol was based on the autocorrelation time of the radius of gyration R_g of each molecule, because a thermalization based on slower observables like the RMSD could not be carried out on the whole set of selected RNAs due to the heavy computational load. Two pulling protocols were considered, namely constant and ramping force: their outcomes will be presented in the next Chapters. The static force was set in order to be able to finish the simulations within the allocated computational resources of few hundred hours of CPU time for most RNAs, while two orders of magnitude were scanned for the ramping protocol.

Chapter 4

Activation Times Asymmetry

In this Chapter we report the results of the nanopore simulations. We will show that the activation of the translocation in the $3' \to 5'$ is *faster* than that in the $5' \to 3'$ direction. This asymmetry holds for different pulling protocols, namely constant or linearly-increasing force.

Under conditions of applied constant force, a subset of the selected RNAs exhibits such pronounced mechanical resistance that their translocation is initiated exclusively from the $3'$-terminus. By switching pulling protocol and employing a ramping force which increases linearly over time, every RNA eventually activates their translocation with both termini, but for each RNA the activation of the $5'$-end is always delayed with respect to that of the $3'$-end. We will show how the data collected under these two types of driving forces are consistent between the two pulling protocols, possibly suggesting that the mechanical response to the forced translocation may be independent of the pulling protocol.

4.1 Static Force Protocol

In this Section we will report the results obtained for the static force pulling protocol. We recall that we performed 21 simulations for each end of the selected RNAs. Prior to each simulation, the RNAs are randomly rotated avoiding steric clashes with the slab and the nanopore, in order to remove possible biases due to the initial orientation of the molecule. Next, the leading dummy phosphate is pulled through the capture region and enters the pore, where each RNA are pulled with the reference force $F_{\mathrm{s}} = 20F_0$ equally divided among the phosphates inside the nanopore, in order to keep the total force acting on the RNA constant.

We will present in this Section and the following ones the translocation traces and their subsequent analyses of four RNAs only instead of those of all the 16 molecules selected in Chapter 2 to keep the discussion more concise and clear. These four representative molecules are a stem-loop of the HCV IRES eIF3 Binding Site (labeled `1KP7_A`), one of the modules of the pRNA of ϕ29 (labeled `3R4F_A`), the H-type pseudoknot of the MMTV mRNA (labeled `1RNK_A`) and one of the ZIKV xrRNA modules (labeled `5TPY_A`). They were selected as they reflect the full spectrum of responses that were observed throughout the simulations. The activation times will be reported for all the molecules.

4.1.1 Translocation Traces

We report the translocation traces in Fig. 4.1, where we monitored $n(t)$, that is the index of the nucleotide straddling the pore entrance. From the translocation traces $n(t)$ it is possible to distinguish two distinct processes, namely the *activation* and the *translocation*. The former corresponds to the stage during which the leading phosphate atom reaches the *trans* region

without retracting from it, effectively corresponding to the onset of irreversible translocations. During this stage, the chain jams around the pore entrance with partial retractions, until the process indeed activates so that the RNA can make its way through the pore. Given the nanopore size, the duration of the activation stage, that is the activation time t_a, corresponds to the time required to bring the fifth nucleotide from the chain end inside the pore. The latter begins as the activation occurs and finishes when the last nucleotide exits the pore.

While analyzing the whole set of translocation traces for each molecule pulled at $F_s = 20F_0$, we observed a continuum of activation times from both ends, with the $5'$-terminal spanning a wider range of times compared to the $3'$-one and encompassing a few orders of magnitude. The single translocation traces in the $3' \to 5'$ are generally activated and completed on a shorter timespan compared to the ones in the other direction. Some RNAs activated their translocation pulling both ends, globally showing a continuous spectrum of activation times: we group them into three broad categories corresponding to *minor*, *typical* and *major* difference in the duration of the activation of the two ends. Interestingly, a few molecules did not activate at all when pulled in the $5' \to 3'$ direction with force $F_s = 20F_0$. Thus, we conveniently subdivide this continuous spectrum of activation times into four categories, each represented by one of the RNAs mentioned above.

(i) No major difference in the response to the forced translocation from the two ends: the activation occurs on a similar timescale between the two termini, where the activation of the $5'$-end is delayed with respect to the $3'$-one, differing by a factor smaller than 5; the representative RNA for this group is the stem-loop of the HCV IRES eIF3 Binding Site, labeled `1KP7_A`.

(ii) The characteristic response consists of a delay of the activation from the $5'$-end with respect to the $3'$-end; we group here the RNAs whose activation times differ by one order of magnitude, and we selected as representative for this group the RNA `3R4F_A`, one of the modules of the pRNA of ϕ29.

(iii) One RNA, namely the H-type pseudoknot of the MMTV mRNA (labeled `1RNK_A` in the PDB), shows a severe delay in the activation from the $5'$-terminal with respect to that from the other end, with a 500-fold difference between the activation times of the two RNA ends. This is the only molecule showing a major difference between the two ends.

(iv) A few RNAs did not activate the translocation from the $5'$-end at all within the set simulation time,[1] with prolonged jamming of the $5'$-end at the pore entrance. We notice how the first nucleotides of the $5'$-end straddle the pore entrance for a notable amount of time with absolutely no sign of advancement, while the $3'$-end manages to penetrate into the pore in a significantly smaller number of integration steps. We select as representative of this group the ZIKV xrRNA `5TPY_A`, since it was extensively studied in Ref. [45, 47].

From the nucleotide traces in Fig. 4.1 we noticed that after the initial stalling, the translocation of all RNAs halts in a more or less pronounced manner when specific nucleotides enter the pore. The halting regions differ for every molecule, and we observe that for each RNA the interruptions occur systematically at every simulation run in the same location.

For example, in case (i) RNA `1KP7_A`, during the $5' \to 3'$ pulling direction there is a consistent jamming when the first four nucleotides are entering the pore entrance. This is observed

[1] We considered as upper bound for the allocated time of the simulations approximately three weeks of simulation time for a single trajectory. Given that some molecules were simulated for more than 10^9 timesteps without activating, an average performance of 500 timesteps/s (that is, the approximate value found for an RNA of average length within our set of molecules) would result in more than three weeks of simulations per trajectory.

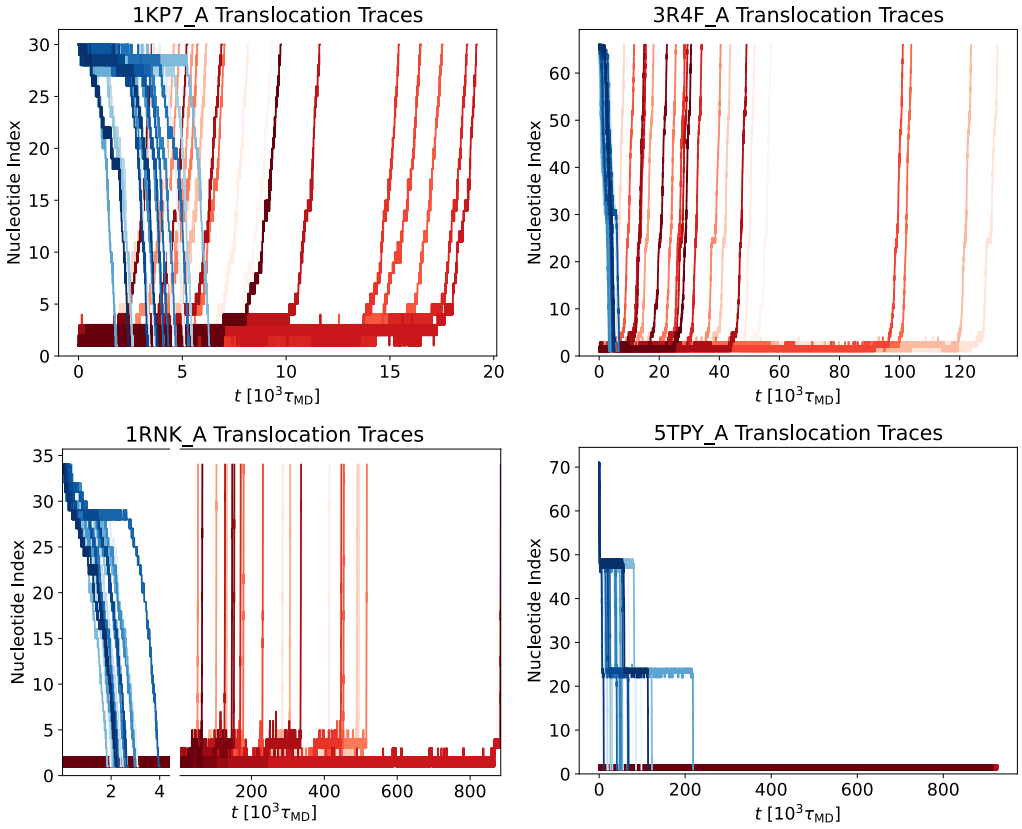


Figure 4.1 Translocation curves at the reference constant driving force $F_s = 20F_0$ of the four sample RNAs `1KP7_A`, `3R4F_A`, `1RNK_A` and `5TPY_A`. The translocation $5' \to 3'$ and $3' \to 5'$ directions are pictured with red and blue shading palettes, respectively. RNA `1KP7_A` shows a similar response to the translocation for both ends, with the $3'$-terminal which activates generally first; RNA `3R4F_A` manifests the typical response with a delayed activation when pulling the molecule from the $5'$-end; RNA `1RNK_A` shows a pronounced difference in the activation by the two ends, retarding the activation of the $5'$-end by two orders of magnitude with respect to the $3'$-terminus (notice the different scales in the broken time axis); finally, the ZIKV xrRNA `5TPY_A` activates only from the $3'$-end within the allocated simulation time and resources.

throughout all the 21 trajectories performed. When pulling in the $3' \to 5'$ direction, the interruptions occur at the pulled terminal, that is when nucleotides 27–30 straddle the pore entrance. RNA `3R4F_A`, representative of case (ii), displays a less marked but consistent stop as nucleotides 23–25 and 30, 31 enter the pore in the $5' \to 3'$ and $3' \to 5'$ translocation direction, respectively. For case (iii) of large difference in the response of the termini, the pseudoknotted RNA `1RNK_A` has a major stalling during the translocation in the $5' \to 3'$ direction for the five terminal nucleotides, and then it precipitously proceeds with the translocation. However, in the opposite direction we observe translocation jamming as nucleotides 27 and 28 begin their way inside the pore. Finally, the ZIKV xrRNA `5TPY_A`, which activated only in the $3' \to 5'$ direction, shows important jamming halfway the translocation around nucleotides 22, 23 and 47–49. The $5'$-end is stuck at the pore entrance with the first nucleotides jamming around the slab. It was shown in Ref. [47] how for the case of the ZIKV xrRNA these stops can be rationalized in terms of specific secondary and tertiary structure elements which drive the response to the forced translocation as the RNA worms its way inside the pore.

4.1.2 Activation Times

From Fig. 4.1, a consistent pattern emerges from the translocation traces all RNAs: the two ends have very distinct activation stages in terms of duration, with the $5'$-end that requires more time to initiate the translocation through the pore. Furthermore, for all molecules but the ZIKV xrRNA in the $3' \to 5'$ direction, the activation and translocation times occur on different timescales, leading us to focus specifically on the activation time, that is generally the dominant term of the simulations in terms of duration.

Thus, we computed the *median* activation time for each set of 21 trajectory per RNA to quantify better these qualitative differences, given the constrained number of trajectories per RNA and the presence of outliers in the single activation time. The statistical uncertainties were computed by block-averaging the activation time values into three blocks of seven datapoints each.

The median activation times of the two pulled ends were then plotted against each other in Fig. 4.2 for the RNAs which activated the translocation in both directions. It is evident from the plot that even considering the statistical errors of the simulations, every viral RNA selected presents *consistently* the same asymmetry in the activation times, namely $t_{\mathrm{a}}^{5'\to3'} > t_{\mathrm{a}}^{3'\to5'}$. For the seven RNAs which activated only from the $3'$-end we reported the median activation time without uncertainty to make the plot clearer.

We will report now the activation instances of all the RNAs, grouping them as aforementioned at the beginning of this Chapter. The RNAs are reported in alphabetical order for each group: the enumeration ordering does not reflect the ordering of the activation times. In Tab. 4.1 are summarized the activation times of the two pulled ends and their respective ratios. We recall that all the data were collected pulling with a constant force $F_{\mathrm{s}} = 20F_0$ only.

(i) Five RNAs report only a *minor* difference between the activation times from the two ends (RNAs labeled `1KP7_A`, `1P5P_A`, `2MQT_A`, `2NBX_A` and `6VAR_A`). The activation times of the two ends show at most a 4-fold increase, with the $5'$-terminal that activates after the $3'$ one, that is $t_{\mathrm{a}}^{5'\to3'} \gtrsim t_{\mathrm{a}}^{3'\to5'}$.

(ii) Three RNAs show the *characteristic* response of viral RNA to the forced translocation, namely that the $5'$-end activates consistently after the $3'$-end (RNAs labeled `1A60_A`, `3R4F_A` and `3T4B_A`), $t_{\mathrm{a}}^{5'\to3'} > t_{\mathrm{a}}^{3'\to5'}$. We observe a 20–30-fold increase in the activation time when pulling these RNAs by the $5'$-end. It is clear from the traces of the $5'$-end how the activation process spans a wider range of times compared to the other terminal.

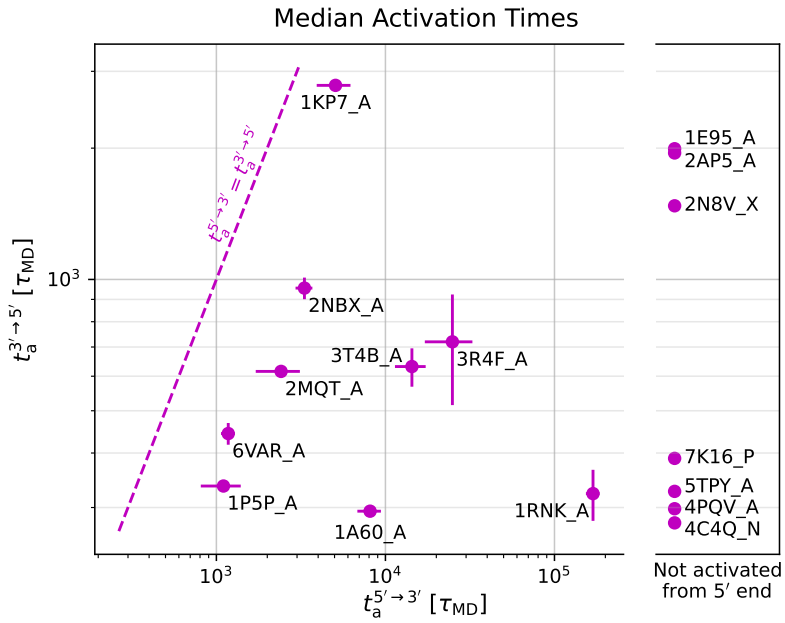


Figure 4.2 Median activation times of the RNAs when pulled from the two ends with the static force $F_s = 20F_0$. The dashed diagonal corresponds to the isochrone line where the activation times of the two ends are equal. Since all datapoints lie on the right side of the diagonal, the activation of the forced translocation is consistently quicker when the molecules are pulled by the 3′-end. On the right of the plot are reported the activation times (without statistical uncertainties for major clarity) of the RNAs which activated only in the $3' \to 5'$ translocation.

(iii) The asymmetric response of the two ends is *pronounced* in the case of the pseudoknotted RNA `1RNK_A`, which displays an approximately 500-fold increase in the activation time from the 5′-end, $t_a^{5'\to3'} \gg t_a^{3'\to5'}$. The 5′-end is engaged at the pore for a much wider time span than the other terminal before the chain starts to translocate.

(iv) Seven RNAs (RNAs labeled `1E95_A`, `2AP5_A`, `2N8V_X`, `4C4Q_N`, `4PQV_A`, `5TPY_A` and `7K16_P`) activated only when pulled by the 3′-end, as their 5′-terminal is locked up with the pore entrance.

The smallest ratio between the median activation times $t_a^{5'\to3'} / t_a^{3'\to5'}$ occurs for the RNA `1KP7_A`, where the activation from the 5′-end takes approximately twice as much as the 3′-terminal: the corresponding datapoint of RNA `1KP7_A` lies the closest to the diagonal in Fig. 4.2. Incidentally, this RNA is the molecule which shows the *largest* activation time from the 3′-end. On the other hand, the smallest activation time in the $3' \to 5'$ pulling direction is found for `4C4Q_N`, the largest RNA of our sample, while the quickest activation of the 5′-end occurs for the stem-loop RNA `1P5P_A`.

We recall that these activation time values were obtained for a single static pulling force only, namely $F_s = 20F_0$. Thus, it is not possible to extrapolate scalings between the activation times or their ratios and the driving force. Intuitively, one expects an increase in the activation times as the force decreases, as shown in Ref. [47] for the case of the ZIKV xrRNA. Of course, the ratios between the activation times of the two ends are likely to change when employing different forces. From panel (a) of Fig. 3 of Ref. [47] we can infer how for the case of the ZIKV xrRNA the ratio $t_a^{5'\to3'} / t_a^{3'\to5'}$ increases as the pulling force decreases, since the logarithms of the activation times have different slopes with respect to the driving force.

RNA	$t_a^{5'\to3'}$ [τ_{MD}]	$t_a^{3'\to5'}$ [τ_{MD}]	$t_a^{5'\to3'}/\,t_a^{3'\to5'}$	Group
1A60_A	$(8.11 \pm 1.29) \times 10^3$	$(2.94 \pm 0.09) \times 10^2$	27.58 ± 4.46	(ii)
1E95_A	N/A	$(1.99 \pm 1.08) \times 10^3$	N/A	(iv)
1KP7_A	$(5.07 \pm 1.15) \times 10^3$	$(2.79 \pm 0.04) \times 10^3$	1.82 ± 0.41	(i)
1P5P_A	$(1.10 \pm 0.29) \times 10^3$	$(3.36 \pm 0.09) \times 10^2$	3.28 ± 0.87	(i)
1RNK_A	$(1.69 \pm 0.16) \times 10^5$	$(3.23 \pm 0.43) \times 10^2$	523.57 ± 85.11	(iii)
2AP5_A	N/A	$(1.95 \pm 0.32) \times 10^3$	N/A	(iv)
2MQT_A	$(2.41 \pm 0.71) \times 10^3$	$(6.15 \pm 0.14) \times 10^2$	3.92 ± 1.15	(i)
2N8V_X	N/A	$(1.48 \pm 0.10) \times 10^3$	N/A	(iv)
2NBX_A	$(3.32 \pm 0.38) \times 10^3$	$(9.55 \pm 0.54) \times 10^2$	3.47 ± 0.44	(i)
3R4F_A	$(2.49 \pm 0.78) \times 10^4$	$(7.19 \pm 2.04) \times 10^2$	34.61 ± 14.63	(ii)
3T4B_A	$(1.43 \pm 0.30) \times 10^4$	$(6.31 \pm 0.64) \times 10^2$	22.72 ± 5.22	(ii)
4C4Q_N	N/A	$(2.77 \pm 0.12) \times 10^2$	N/A	(iv)
4PQV_A	N/A	$(2.98 \pm 0.44) \times 10^2$	N/A	(iv)
5TPY_A	N/A	$(3.27 \pm 0.05) \times 10^2$	N/A	(iv)
6VAR_A	$(1.18 \pm 0.11) \times 10^3$	$(4.43 \pm 0.25) \times 10^2$	2.65 ± 0.29	(ii)
7K16_P	N/A	$(3.89 \pm 0.21) \times 10^2$	N/A	(iv)

Table 4.1 Summary of the median activation times and their ratio obtained for the set of 16 selected RNAs in the static force protocol with driving force $F_s = 20F_0$.

4.2 Ramping Force Protocol

The chosen reference force $F_s = 20F_0$ is not sufficient to trigger the translocation in the $5' \to 3'$ direction for all RNAs within the allocated simulation time. Thus, in order to observe the response and activation of both termini for *all* RNAs, we switched to a ramping force protocol where the driving force increases in time. In this way, the critical force necessary to activate the translocation will be eventually reached as the simulation progresses. In particular, we changed the pulling force to a linearly increasing one, where $F_r(t) = rt$ given the pulling rate r. This kind of protocol was already successfully exploited in Ref. [45] to unveil the underlying mechanism behind the resistance to degradation of the ZIKV xrRNA.

We scanned various ramping rates spanning two orders of magnitude, namely the pulling rates $r = \frac{50F_0}{a\times10^6}\frac{1}{\text{timestep}}$, where $a = 500, 200, 100, 50, 20, 10, 5, 1$ (rates are reported in simulation units). Expressing the various ratios as functions of the characteristic time τ_{MD} in a more convenient physical format, these ramping rates correspond approximately to the expression $\frac{F_0}{22\times a}\frac{1}{\tau_{MD}}$, where a assumes the previous values. Thus, we explored the rates $\frac{F_0}{22\times500}\frac{1}{\tau_{MD}}$ to $\frac{F_0}{22\times1}\frac{1}{\tau_{MD}}$. Recall $F_0 = k_B T/\sigma$, where $T = 75\,\text{K}$ is the simulation temperature and $\sigma = 5\,\text{Å}$ is the cutoff interaction range for the LJ pairs in the SMOG force field.

Similarly to the static force protocol, we performed 21 simulations for each combination of RNA, pulled end and ramping rate. Also in this case, the priming process was performed like the one of the constant force protocol, that is randomly rotating the molecule around the pore entrance and then letting it relax for a time interval equal to twice the autocorrelation time of the radius of gyration.

4.2.1 Translocation Traces

We report in Fig. 4.3 the translocation traces for the four RNAs selected from our list. Differently from the static force protocol, the activation times for fixed pulling rate are less scattered. Duration-wise, the activation stage dominates the whole translocation process for all RNA cases,

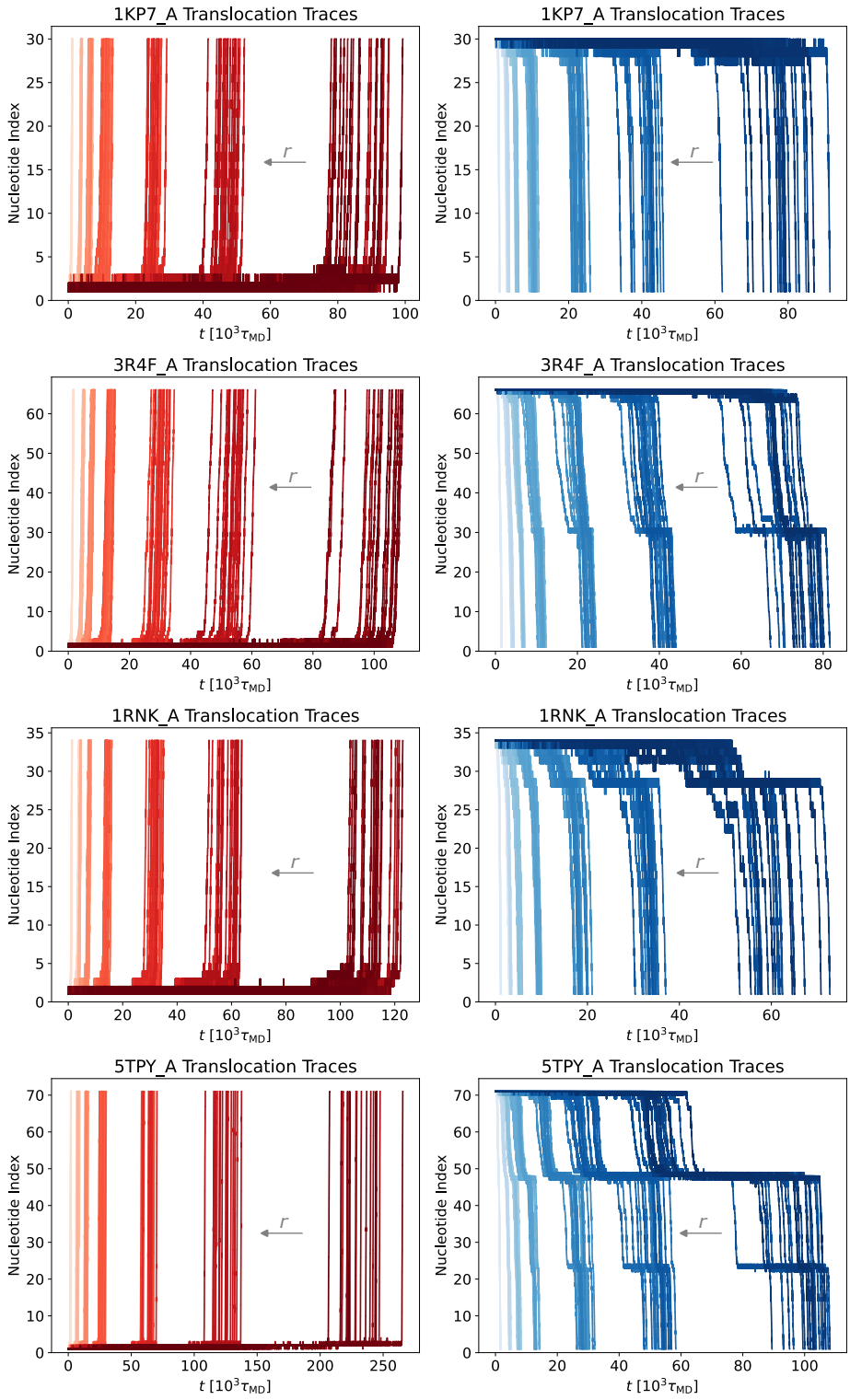


Figure 4.3 Translocation curves with driving ramping force F_{r} for the four sample RNAs `1KP7_A`, `3R4F_A`, `1RNK_A` and `5TPY_A` at different pulling rates. The distinct lumps of trajectories correspond to different simulations performed with the same pulling rate, where larger rates activate the translocation faster. The reported rates are $r = \frac{F_0}{22\times a}\frac{1}{\tau_{\mathrm{MD}}}$, where $a = 200, 100, 50, 20, 10, 5, 1$ for increasing value of the rate r. Larger rates (corresponding to smaller values for a) result in shorter activation times.

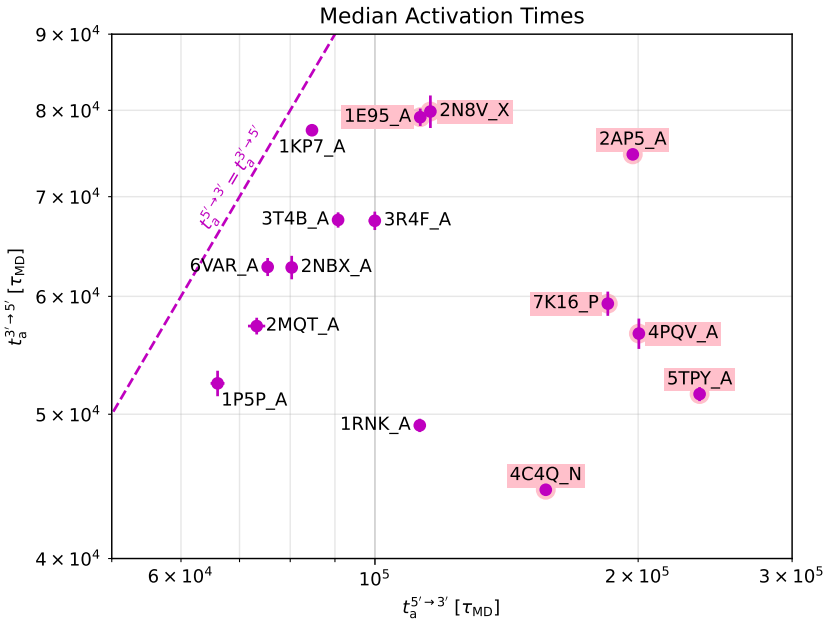


Figure 4.4 Median activation times for the ramping force protocol with pulling rate $r = \frac{F_0}{22\times200}\frac{1}{\tau_{MD}}$. The datapoints corresponding to the RNAs that activated only from the 3′-end in the static case are highlighted in light pink. It is evident that also in the ramping force protocol the activation is faster when pulling from the 3′-end. The 5′-terminal, which shows larger activation times for all the RNAs, comprises with this pulling protocol all the aforementioned "hard" cases.

especially at lower rates where it takes more time to reach the critical force that activates the translocation. We observe also within this ramping pulling framework the same specific stalling regions in correspondence of the nucleotides described in Sec. 4.1, see Fig. 4.1. This behavior is observed throughout *all* pulling rates, indicating a coherent response to the forced translocation which is independent of the force protocol.

For example, the stem-loop `1KP7_A` consistently halts when the four terminal nucleotides enter the pore in both pulling directions, as observed with the static force pulling protocol. Likewise, the ZIKV xrRNA nucleotides 22, 23 and 47–49 straddle the pore entrance for a consistent fraction of the translocation segment. Moreover, lower rates yield larger activation times, halting periods and overall translocation duration.

4.2.2 Activation Times

To establish a reference pulling rate for the subsequent analysis, we chose the lowest rate which ensures the translocation completion of *all* 16 RNAs from both termini while remaining compatible with the allocated computational time, that is the pulling rate $r = \frac{F_0}{22\times200}\frac{1}{\tau_{MD}}$. In analogy with Fig. 4.2, we report in Fig. 4.4 the median activation times by the two ends for the reference pulling rate $r = \frac{F_0}{22\times200}\frac{1}{\tau_{MD}}$. The errorbars were computed also in this case by block-averaging the 21 datapoints for each RNA terminus into three blocks of seven datapoints each.

The results show the systematic differences of the two chain ends also in the ramping force protocol, with the activation at the 5′-end being delayed from that at the 3′ one. Nevertheless, in this case the distribution of activation times is considerably narrower than under a static force, constraining the median activation times within one order of magnitude for both ends for the simulations at the selected ramping rate. We find that the RNAs that show the strongest

resistance under a constant pulling force, namely the RNA chains `1E95_A`, `2AP5_A`, `2N8V_X`, `4C4Q_N`, `4PQV_A`, `5TPY_A` and `7K16_P`, which did not activate at all during translocation in the $5' \to 3'$ direction, together with the strongly delayed RNA `1RNK_A`, also remain the slowest to initiate translocation when the pulling force is ramped.

Extending the results and considering all the pulling rates tested, we report in Fig. 4.5 the computed the median activation time for each combination of RNA, pulled terminal and rate r for groups of RNAs with different topological complexity, which we quantified with their number of pseudoknots as previously done in Chapter 2. The activation times of the various RNAs are more and more constrained as the pulling rate increases, also reducing notably the errorbars at the largest pulling rates. The overall $t_{\mathrm{a}}^{5'\to3'} > t_{\mathrm{a}}^{3'\to5'}$ behavior is preserved across all groups, but we notice that pseudoknotted structures, which encompass more complex topologies, exhibit a broader dispersion of datapoints and larger absolute median activation times. Furthermore, we report that not all the RNAs activated the translocation within the allocated simulation time at the lowest pulling rate $a = 500$, especially as the size or the complexity of the RNA increases. In this case, the primary limitation in the collection of the data is the allocated simulation time, as smaller pulling rates require more time to reach the critical force that triggers the translocation at t_{a}.

4.3 Correlations Between Static and Ramping Force Protocols

The activation times reported in Fig. 4.4 and Fig. 4.5 show consistent differences between the two ends also in the ramping force protocol: in presence of an increasing force, the activation from the $5'$-terminal is delayed with respect to the $3'$-one.

To assess the systematic nature of this asymmetry across the two different pulling protocols, we compared the median activation times of each RNA with the values obtained at constant pulling force when the translocation was actually activated. The scatterplots containing the activation times under different pulling methods are reported in Fig. 4.6. We remark that the outlined plots contain 25 datapoints only: 18 datapoints corresponding to the nine RNAs which activated both ends, and the remaining seven originated by the RNAs which activated only when pulled by the $3'$-terminal.

It is evident that the data are correlated and that a monotonic trend occurs at every pulling rate explored: RNAs with hard activation from the $5'$-end display such delay with both pulling rates, hinting that the key factor in the activation process does not lies in the pulling protocol, but could be encoded in the molecule folds. At the range of pulling force and rates employed, the static force protocol yields generally a broader dispersion of the activation times over several orders of magnitude.

We evaluated the correlation between the activation times under different pulling protocols using the Kendall-τ correlation test. This non-parametric and rank-based method is appropriate for assessing ordinal, monotonic relationships including non-linear ones, as it focuses on the ordinal ranking of the data [139, 140]. In particular, the test is computed from the distinct pairs of N datapoints, establishing the probability that two observation have higher/lower rankings [141], providing a sensitive assessment of correlation even when the sample size is modest.

In Tab. 4.2 we report the results of the correlation tests performed between the static force activation times and those obtained with different pulling rates in the ramping force protocol. The correlation is very strong: Kendall's correlation coefficient $\tau \approx 0.85$, with minor variations depending on the pulling rate. The significance of this correlation is corroborated by the vanishingly small p-value $< 10^{-10}$ found consistently at each pulling rate r. Thus, the mechanical responses at different pulling rates r is consistent with the same set of data collected when translocating the molecules with a constant force. This could suggest that the response to the

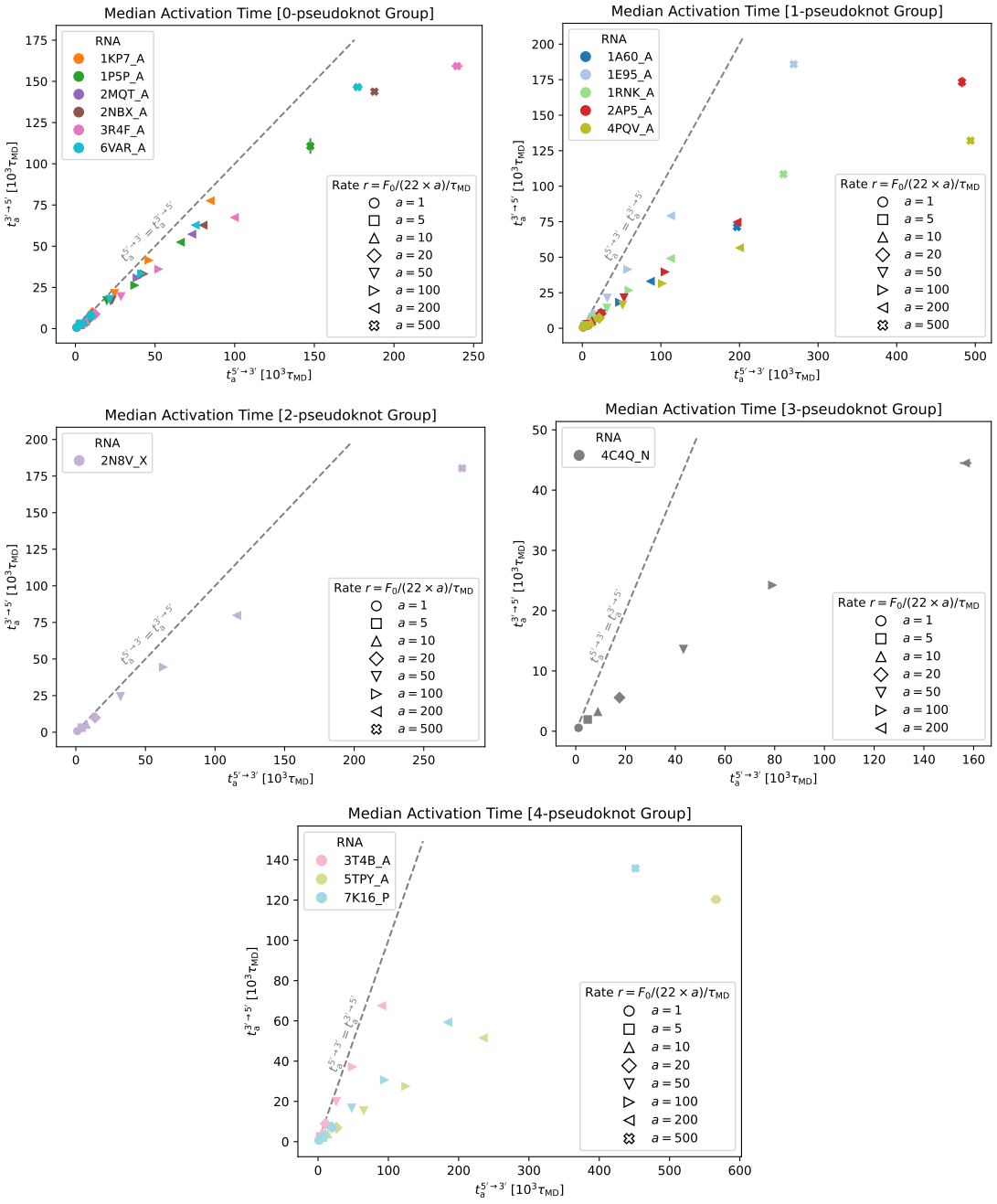


Figure 4.5 Activation times for topologically distinct RNAs under several pulling rates. As the pulling rate decreases, the activation times increase and get more scattered, always in accordance with the global observation that $t_{\mathrm{a}}^{5'\to3'} > t_{\mathrm{a}}^{3'\to5'}$.

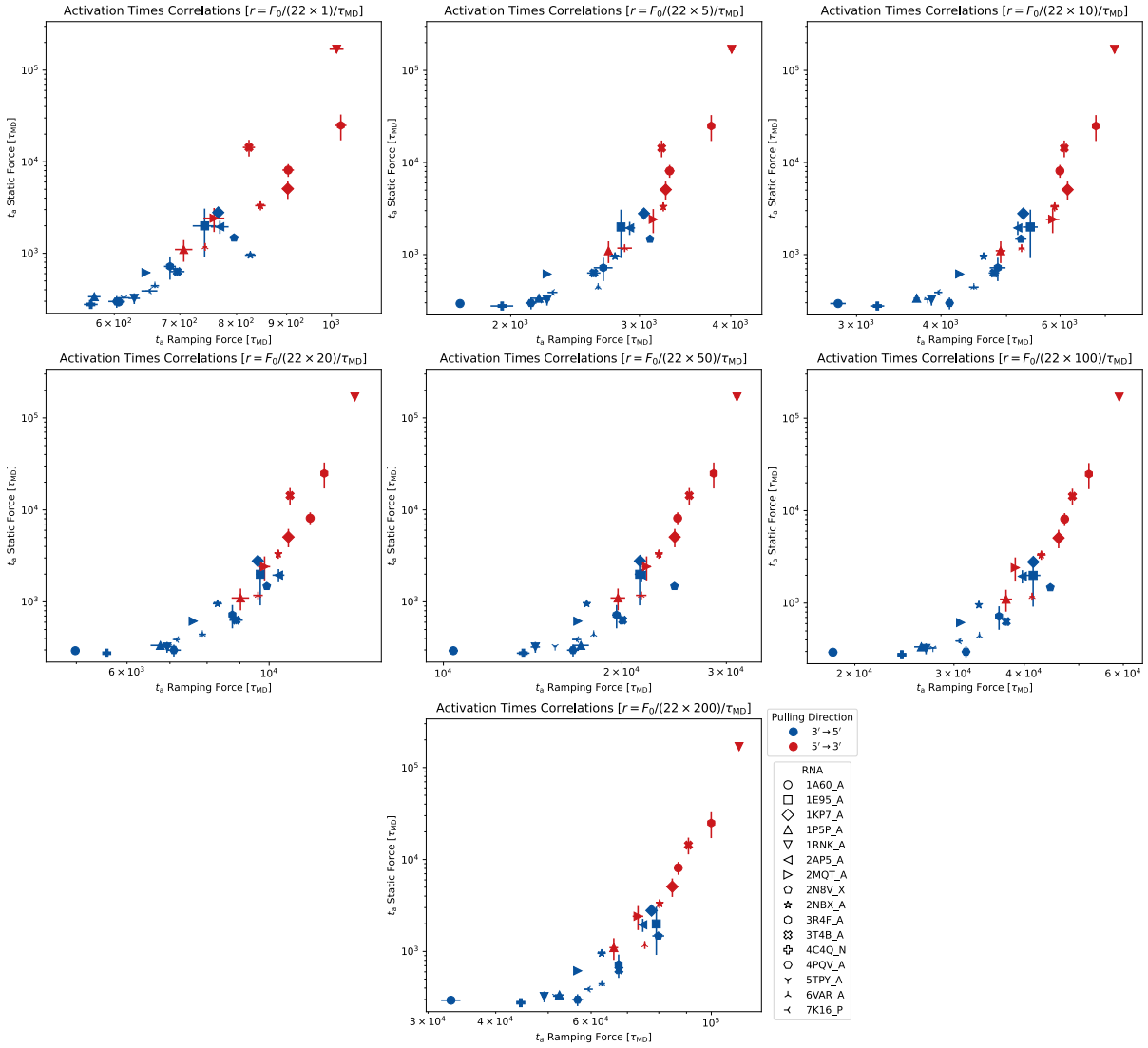


Figure 4.6 Scatterplots of the activation times under static and ramping force for different pulling rates. Notice the larger timescale spanned by the activation times under constant pulling force (vertical axis) compared to the one with the ramping force protocol. Different markers are associated with different RNAs, and the data for $5' \to 3'$ and $3' \to 5'$ translocations are represented in red and blue, respectively. The qualitative response is the same throughout all the tested pulling rates r: the activation times are always strongly correlated between constant and ramping force.

Rate r	$\frac{F_0}{22\times1}\frac{1}{\tau_{\mathrm{MD}}}$	$\frac{F_0}{22\times5}\frac{1}{\tau_{\mathrm{MD}}}$	$\frac{F_0}{22\times10}\frac{1}{\tau_{\mathrm{MD}}}$	$\frac{F_0}{22\times20}\frac{1}{\tau_{\mathrm{MD}}}$	$\frac{F_0}{22\times50}\frac{1}{\tau_{\mathrm{MD}}}$	$\frac{F_0}{22\times100}\frac{1}{\tau_{\mathrm{MD}}}$	$\frac{F_0}{22\times200}\frac{1}{\tau_{\mathrm{MD}}}$
τ	0.81	0.89	0.88	0.86	0.86	0.85	0.86
p-value	4.89×10^{-11}	1.46×10^{-14}	3.31×10^{-14}	3.20×10^{-13}	3.20×10^{-13}	1.28×10^{-12}	3.20×10^{-13}

Table 4.2 Kendall-τ correlation coefficients between activation times obtained under constant force and ramping force pulling. For each loading rate r, the activation times measured with the corresponding ramping protocol $F_{\mathrm{r}}(t) = rt$ were correlated with those obtained under a constant pulling force at $F_{\mathrm{s}} = 20F_0$. The consistently values of $\tau \approx 0.85$ indicate a strong correlation between the two pulling protocols over the entire range of ramping rates sampled.

forced translocation of the selected viral RNAs may be *independent* of the force: as shown by the ramping force protocol, the magnitude of the force does not dictate the directional response of the RNAs, as the condition $t_{\mathrm{a}}^{5'\to3'} > t_{\mathrm{a}}^{3'\to5'}$ is found at every pulling rate. The origin of the distinct behavior exhibited by the two termini must be attributed to the molecular structures of the viral RNAs, which will be described and analyzed in detail in the next Chapter.

4.4 Framing The Activation Times

The results of the previous sections showed that the asymmetry between the two translocation directions is preserved regardless of the pulling protocol employed. Plus, the median activation times obtained under static and ramping force are strongly correlated (Sec. 4.3). This suggests that the activation process can be understood as an escape over a single, force-dependent energy barrier whose properties are set by the molecular architecture of each RNA, rather than by the specific driving protocol. There are different interpretative models which describe activation processes in the literature. Given that the activation time is the time necessary to translocate the leading phosphate through the nanopore, there is a clear mapping between activation times and first passage times through the energetic barrier that models the translocation activation.

4.4.1 Static Force Protocol

As aforementioned in Sec. 4.1.2, a single force is not sufficient to obtain information on the width and height of the energetic barrier associated with the activation process. In Ref. [47] it was shown how the activation times of the ZIKV xrRNA follow an exponential decay with the static pulling force

$$t_{\mathrm{a}} \propto \exp(-F_{\mathrm{s}}\Delta/k_{\mathrm{B}}T) \tag{4.1}$$

where Δ is the effective width of energetic barrier associated to the activation of translocation, modeled as a two-state process. The associated energetic barriers at the $5'$- and $3'$-end revealed a barrier width of $\Delta_{5'} = 5.6 \pm 0.2\,\text{Å}$ and $\Delta_{3'} = 1.69 \pm 0.04\,\text{Å}$, respectively [47]. Notice the force field employed in this work and in Ref. [47] differ, thus the following arguments will be strictly qualitative.

Given that the same asymmetry is observed throughout our set of selected molecules, with activation of the $5'$-end delayed to some extent with respect to the $3'$-one, we expect to observe an asymmetry in the barriers $\Delta_{5'} > \Delta_{3'}$ for all the 16 RNAs. The ZIKV xrRNA shows that the two energetic barrier widths differ by a few angstroms, as $\Delta_{5'} - \Delta_{3'} = 3.91 \pm 0.20\,\text{Å}$. RNAs with less pronounced asymmetry in the activation times by the two ends (groups (i)-(ii) in Tab. 4.1), are likely to show comparable widths of the barriers at the two ends, that is, small differences $\Delta_{5'} - \Delta_{3'}$). On the other hand, the most resisting RNAs (groups (iii)-(iv) in Tab. 4.1) should show a pronounced asymmetry in such widths.

4.4.2 Ramping force protocol

In the ramping force protocol, the pulling force changes over time. The scattering of the activation times, which can be appreciated for the lumps of trajectories reported in Fig. 4.3, results in a scattering of the pulling force that triggers the activation. The analysis in force-spectroscopy contexts of rupture forces is grounded in the Bell-Evans theory [142]. Within the Bell-Evans framework, it is possible to extract the most probable rupture force $F^\star$ as

$$F^\star = \frac{k_\mathrm{B}T}{\Delta} \ln\left(\frac{r\,\Delta}{\nu\,k_B T}\right) + \frac{E_T}{\Delta} \tag{4.2}$$

where Δ and E_T are respectively the thermodynamic barrier width and height, and ν is a dynamic coefficient associated with the bond breaking attempts. The nontrivial aspect of Eqn. (4.2) is that the most probable rupture force has a logarithmic dependence on the pulling rate r.

In the previous study on the translocation properties of the ZIKV xrRNA, Ref. [45], the barriers heights at the two ends read (at room temperature $T = 300\,\mathrm{K}$) $E_T^{5'} = 60 k_\mathrm{B}T$ and $E_T^{3'} = 41 k_\mathrm{B}T$, while the barrier widths are $\Delta_{5'} = 4.4 \pm 0.4\,\text{Å}$ and $\Delta_{3'} = 2.0 \pm 0.1\,\text{Å}$. The widths are comparable with the values found in the static case force protocol, see Ref. [47] and the previous Section 4.4.1.

In order to obtain the kinetic coefficient ν, which is required to compute the barrier height associated to the activation, it is necessary to scan several combinations of simulation temperatures T and pulling rates r. Then, one could plot $F^\star$ as a function of the logarithm of the pulling rate as done in Fig. 3(a) of Ref. [45] to extract from the fit the width and height of the thermodynamic barrier for any translocated RNA. Notice that in order for the Bell-Evans theory to hold, the pulling rates should be small enough to guarantee that the pulling process is quasi-adiabatic, condition necessary to Eqn. (4.2) to be valid, see Ref. [143].

Qualitatively speaking, even though our computational setup differs from the one employed in previous studies of our group on the ZIKV xrRNA, we expect the barrier heights at the $5'$-terminal of the selected RNAs to be larger than the ones at the other end. Similarly to the barrier widths, the barrier heights of RNAs with mild and moderate asymmetric response are expected be different but comparable, $E_T^{5'} \gtrsim E_T^{3'}$, while a major difference is expected to be found in the pseudoknotted structures which showed at least a 500-fold asymmetry in the activation times.

4.5 Summary

In this Chapter, we presented a systematic computational investigation of the nanopore forced translocation of the selected viral RNAs. We showed that the RNAs present an asymmetric response to this process: the duration of the activation stage, that is the activation time t_a, occurs on different time scales between the two ends. The translocation in the $5' \to 3'$ direction activates systematically *after* that in the $3' \to 5'$ sense.

Pulling the molecules with a constant force, we observed a wide spectrum of delays of the activation in the $5' \to 3'$ translocation direction. In particular, at the employed constant driving force $F_\mathrm{s} = 20F_0$, the activation times from the $5'$-end spans a continuum of activation times which encompasses three orders of magnitude considering only the RNAs that activated their translocation from the $5'$-terminus, whereas the $3'$-end are constrained within one order of magnitude Seven of the 16 structures selected in Chapter 2 did not activate with the employed pulling force within the allocated simulation time when pulled from the $5'$-end, indicating that the energetic barrier at the $5'$-end of these molecules is larger than the $3'$ one and not comparable with it.

Thus, in order to the observe the activation of *all* molecules, we switched pulling protocol to a ramping one where the force increases linearly with time. This pulling protocol eases the activation, as a growing force threads the RNA through the nanopore more easily by lowering more and more the energetic barrier to activate the translocation. Two orders of magnitude of ramping rates were scanned, and regardless of the rate employed in the simulations we observed the same qualitative differences in the activation times of the two ends of each RNA: the activation of the $5'$-terminal is delayed compared to the $3'$-end. However, for the selected range of pulling rates r, the activation times are less scattered compared to the static force protocol (the fluctuations are less evident), yielding for both ends a 10-fold range of activation times for each end at the lowest pulling rate $r = \frac{F_0}{22\times200}\frac{1}{\tau_{\mathrm{MD}}}$.

Independently of the pulling rate, the median activation times obtained with the ramping-force protocol are in strong agreement with those measured under a constant pulling force. This is supported by a high Kendall-τ coefficient and a statistically significant p-value, suggesting that the response to forced translocation could be independent of the pulling protocol, reflecting instead the molecular architecture of the viral RNAs.

In this regard, the molecular architectures of the seven RNAs that did not activate in the static force protocol are the responsible for the intrinsic stability and resistance of these RNA molecules to the forced translocation. These folds are structurally fascinating as they comprise several pseudoknots and intertwined strands. Thus, these folds will be analyzed in the following Chapter.

Chapter 5

Simulations Analyses

We showed in the previous Chapter that the selected viral RNAs respond differently to the forced translocation through a nanopore in a consistent manner: the $5'$-end is delayed with respect to the $3'$ one, independently of the pulling protocol employed. Furthermore, the topological complexity of the RNAs seems to play a major role in the activation stage, since pseudoknotted molecules display a major difference in magnitude between the activation times of the two ends, which can exceed more than two orders of magnitude. Since the response appears to be independent of the pulling protocol and force, the responsibility of the asymmetric response should be traced back to the folds of the RNAs. We thus inspect the various molecular architectures of the RNAs to shed a light on the possible causes of this asymmetric response.

5.1 Contacts

Differently from *homopolymers*, which are made of one single type of repeating unit (that is, the monomers), RNA is a *heteropolymer* composed of diverse constituents, that is the nucleotides adenine, cytosine, guanine and uracil. Each nucleotide contributes to the RNA energetics with its own pattern of hydrogen bonds, stacking and electrostatic interactions. The resulting network of interactions between residues dictates the RNA fold and gives rise to energetic barriers that must be overcome in order to activate the translocation: while pulling heteropolymers inside a nanopore, some of the bonds formed inevitably break because of the molecule squeezing within the nanopore constriction. This differs from the theory of forced homopolymer translocation, where a homopolymer is typically idealized as a chain of identical phantom monomers (that is, non-interacting) or interacting at most only through generic excluded-volume and connectivity terms,[1] in order to have a simpler view on the physics of forced translocation [34, 35, 38]. Such chains enjoy larger conformational freedom than heteropolymers like RNA, and their translocation is activated as soon as the pulling force overcomes the entropic recoil. This simple picture made it possible to develop a grounded theory about nanopore translocation [34, 35, 37, 38, 145].

Thus, terminal contacts are an intuitive observable which can be used to characterize the asymmetric response to the forced translocation, since they are responsible for the barriers that must be overcome first in order to activate the translocation. It is reasonable to expect that the larger the number of bonds to break, the larger the onset of the polymer translocation.

[1] In this regards, poly(A) chains, often occurring during mRNA metabolism, should not be considered as homopolymers as they are properly structured thanks to the network of interactions formed by the adenines. In fact, this polymer assumes in physiological conditions the structure of a right-handed helix, stabilized by stacking interactions between the bases [144].

5.1.1 Primary Contacts

Contacts Retrieval

We then retrieved the contacts involving the five terminal nucleotides at the two ends of each RNA, as they are the nucleotides which have to translocate inside the pore in order to activate the whole molecule translocation. These contacts are reported in the contact maps generated by the SMOG force field for all RNAs, which are represented in Fig. 3.2. The contact map is encoded by SMOG in a topology file (`.top` extension of the SMOG output) containing information on the atoms of each molecule (type, mass, residue of origin, ...), the various types of interactions expressed in Eqn. (3.1), and other parameters required to run the MD simulations.

However, the contacts reported in such files are computed on the crystal structures retrieved from the RCSB PDB. During the simulations, some of these contacts eventually break, either due to thermal fluctuations, the insertion inside the pore, or by effect of the driving force (independently of the fact that it can be constant or increasing).

To see the effects of the aforementioned causes during the activation stage of each RNA, we retrieved the number of *effective intact primary contacts* during the activation as follows. First, for each trajectory we extracted the atomistic configuration halfway the activation time, namely at the timestamp $t = t_a/2$. For the RNAs which did not activate in the static case, we fetched the configuration at half of the elapsed trajectory. Second, we measured the distances of all atom pairs that form non-covalent and LJ interactions, which are reported in the `[ pairs ]` section of the topology file of each RNA. Then, we computed the atomistic $N_{\text{atom}} \times N_{\text{atom}}$ contact map A_{ij}^{atom} of all the non-covalent and LJ interactions between atoms, where

$$A_{ij}^{\text{atom}} = \begin{cases} q(\sigma_{ij}) & \text{if } i,j \text{ are native contacts} \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

where "contact fraction" $q(\sigma_{ij})$ was computed for each pair of native contacting atoms i and j at distance σ_{ij} at the selected timestamp with the sigmoid function Eqn. (3.8). In order to get a binary contact map of zeros and ones which tells whether atoms i and j are actually in contact, we rounded A_{ij}^{atom} by constructing the binary matrix B_{ij}^{atom} as follow:

$$B_{ij}^{\text{atom}} = \begin{cases} 1 & \text{if } A_{ij}^{\text{atom}} \geq 0.85 \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

The threshold value $q = 0.85$ to define whether a contact is broken or present derives from the fact that at the native distance between atoms σ_{ij}^0 it holds $q(\sigma_{ij}^0) \approx 0.92$ and that in this way we can still consider atoms i and j in contact if the distance between them increases at most by approximately 10% (always maintaining the cutoff distance of 5 Å for each contact). In this way we have a binary contact matrix at the selected frame for each RNA. Finally, we coarse-grained each atomic contact matrix to the nucleotide level by computing a new $N_{\text{ntd}} \times N_{\text{ntd}}$ contact map A_{ij}^{ntd} for the nucleotides, where the matrix entries A_{ij}^{ntd} account for all the non-covalent contacts between the atoms belonging to nucleotides i and j reported in B_{ij}^{atom}, basically making blocks out of it. To avoid interactions between atoms of the same nucleotide (*intra*-nucleotide interactions), we set to zero the diagonal elements of the matrix, $A_{ii}^{\text{ntd}} = 0$, thus effectively retrieving only *inter*-nucleotide interactions.

Finally, given how the activation time was extracted from the translocation traces, we fetched the total number of contacts involving the terminal five nucleotides of each RNA end. Therefore,

RNA	1A60_A	1E95_A	1KP7_A	1P5P_A	1RNK_A	2AP5_A	2MQT_A	2N8V_X
5′-end	106	104	102	96	126	163	104	100
3′-end	61	84	110	88	86	115	99	81
RNA	2NBX_A	3R4F_A	3T4B_A	4C4Q_N	4PQV_A	5TPY_A	6VAR_A	7K16_P
5′-end	106	111	106	128	115	148	98	153
3′-end	104	104	93	90	91	113	96	103

Table 5.1 Summary of the number of contacts formed by the terminal five nucleotides at each end of the selected RNAs in their native conformation retrieved from the structure file. In almost all cases, the 5′-end forms more contacts than the 3′-end. The only exception is the stem-loop RNA 1KP7_A, which displays more contacts at the 3′-end (110) than at the 5′-end (102).

for the 5′- and 3′-termini of an N-nucleotide RNA we respectively get

$$\begin{aligned} n_{\text{contacts}}^{5'\text{-end}} &= \sum_{i=1}^{5} \sum_{j>i} A_{ij}^{\text{ntd}} \\ n_{\text{contacts}}^{3'\text{-end}} &= \sum_{i=N-4}^{N} \sum_{j<i} A_{ij}^{\text{ntd}} \end{aligned} \tag{5.3}$$

Correlations Between Activation Times and Terminal Contacts

For each combination of RNA, pulled terminal and trajectory we retrieved the observable n_{contacts} using Eqn. (5.3) for both the static and the ramping driving protocol at the reference pulling force and rate, respectively. Then, for each RNA we computed the average value of n_{contacts} over the different trajectories. We report in Tab. 5.1 the contacts at the two termini of each molecule in the fold reported in the corresponding PDB entry, while the data for the average number of resisting contacts in the static and ramping pulling protocols are reported in Tab. 5.2-5.3. The median values are very similar to the mean ones, differing at most by approximately two contacts (data not shown). The RNAs display generally a larger density of contacts at the 5′-terminal: across all the retained RNAs, only the RNA labeled 1KP7_A has a larger number of contacts in correspondence of the 3′-end in the crystal structure and in the ramping force protocol, where we report an average of 69.81 contacts at the 5′-terminal against 79.38 at the 3′-one. Also the RNA 6VAR_A shows a similar asymmetry in the ramping force protocol, with 67.05 and 80.62 at the 5′- and 3′-terminal, respectively. During the static force protocol, the RNAs consistently share a larger density of resisting contacts at the 5′-end.

Regardless of the pulling protocol employed, we observe that for both termini approximately 40% of the terminal primary contacts are broken halfway of the activation stage (except for the seven resisting RNAs, which obviously break to a minor extent since they penetrated less inside the pore), with the 3′-end showing a larger increase of this value. The 5′-termini usually display a similar average number of primary contacts regardless of the pulling protocol, with a difference of around 10% between the static and the ramping force protocol; on the other hand, the 3′-termini show larger variability between the different pulling.

Then, we plotted the activation times versus the corresponding number of contacts for the two ends of each RNA for both the static and the ramping pulling protocol to verify the presence of some form of correlation between these observables. The data are reported in the scatterplots of Fig. 5.1. We observe a global trend in which the activation time increases with the number of contacts, even though the correlation is present but not particularly strong. In fact, assessing it

RNA	1A60_A	1E95_A	1KP7_A	1P5P_A	1RNK_A	2AP5_A	2MQT_A	2N8V_X
5′-end	60.67	61.10	58.29	53.52	85.43	113.48	71.48	66.00
3′-end	20.14	46.62	57.86	38.48	49.19	59.19	61.52	42.81
RNA	2NBX_A	3R4F_A	3T4B_A	4C4Q_N	4PQV_A	5TPY_A	6VAR_A	7K16_P
5′-end	63.19	71.52	60.57	79.95	79.24	94.05	59.29	92.33
3′-end	52.33	57.71	41.48	45.81	52.10	62.71	49.14	54.95

Table 5.2 Summary of the number of resisting contacts involving the terminal five nucleotides at each end of the selected RNAs in the static case pulling force.

RNA	1A60_A	1E95_A	1KP7_A	1P5P_A	1RNK_A	2AP5_A	2MQT_A	2N8V_X
5′-end	66.43	71.38	69.81	60.38	89.71	114.81	80.24	68.81
3′-end	36.52	52.86	79.38	48.57	58.24	75.33	79.67	57.33
RNA	2NBX_A	3R4F_A	3T4B_A	4C4Q_N	4PQV_A	5TPY_A	6VAR_A	7K16_P
5′-end	72.43	76.52	69.48	80.43	77.95	92.62	67.05	92.62
3′-end	72.29	74.38	58.48	56.71	56.90	71.14	68.62	68.43

Table 5.3 Summary of the number of resisting contacts involving the terminal five nucleotides at each end of the selected RNAs in the ramping case pulling force.

by means of the Kendall-τ correlation test, we report for the static case a correlation coefficient $\tau = 0.49$, with a p-value $= 4.53 \times 10^{-4}$. We recall that in the static case there were seven RNAs (labeled `1E95_A`, `2N8V_X`, `2AP5_A`, `4C4Q_N`, `4PQV_A`, `5TPY_A` and `7K16_P`) that activated only from the 3′-end within the allocated computational time. For the ramping force case, in which *all* RNAs activated their translocation, the Kendall correlation coefficient decreases slightly with respect to the static case, $\tau = 0.47$. The statistical significance slightly strengthens, as the p-value decreases to p-value $= 1.38 \times 10^{-4}$. Nevertheless, both cases exhibit a mild form of correlation with a statistically significant p-value < 0.001.

The Kendall correlation coefficient in the static force protocol is therefore slightly stronger than that in the ramping case. The larger correlation observed with the constant force is driven by the fact that some of the RNAs did not activate at all, and that the ones that managed to activate display an increasing number of contacts in the RNAs with the largest activation time. In particular, we observe how the five most resisting RNAs in the static force protocol maintain an ordinal increasing trend between the median activation time and the resisting contacts at the pulled end: larger in terminal contacts, larger in t_{a}. On the other hand, considering the full spectrum of translocated RNAs as in the ramping case, this trend weakens due to the presence of molecules with a modest number of contacts which still display a large activation time. Moreover, the Kendall-τ test is based on the *ordinal association* of datapoints [140]: in this case, larger in terminal contacts, larger in activation time. The overall trend is clear by looking at the whole spectrum of activation in the ramping case, but the RNAs with large activation times do not follow an ordinal trend which this non-parametric test is thought to capture.

The Outliers

The seven cases of RNAs which did not activate by the 5′-end in the static force protocol simulations are of particular interest. All these molecules are pseudoknotted and consist of the RNAs `1E95_A`, `2AP5_A`, `2N8V_X`, `4C4Q_N`, `4PQV_A`, `5TPY_A`, and `7K16_P`. All these RNAs display also the largest activation times for the $5' \to 3'$ translocation with pulling rate $r = \frac{F_0}{22\times 200}\frac{1}{\tau_{\mathrm{MD}}}$ in Fig. 5.1, but quite differ among themselves in the number of contacts at the 5′-terminal; we

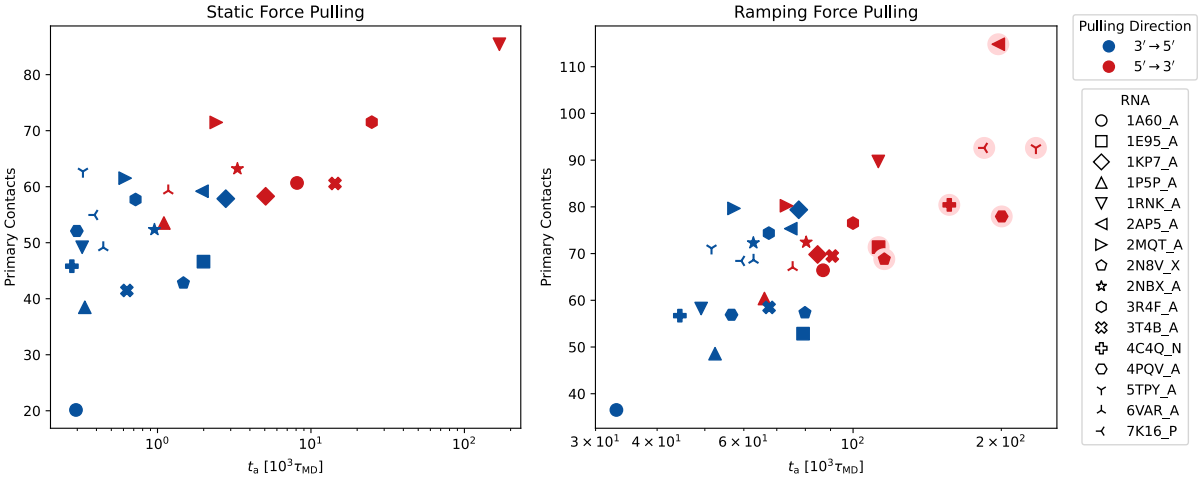


Figure 5.1 On the left, the scatterplot of median activation times versus primary contacts at the ends of the RNAs for the static force pulling protocol, $F_s = 20F_0$. On the right, the same type of scatterplot for the ramping case with pulling rate $r = \frac{F_0}{22\times200}\frac{1}{\tau_{MD}}$. The RNAs which activated only from the 3′-end in the static force protocol are highlighted with a pink circle around their marker. Notice for these RNAs the similar range of activation times in the horizontal axis and the larger spread in the primary contacts at the 5′-end in vertical axis.

can cluster them into two groups depending on the number of contacts at their 5′-terminal.

In particular, the pseudoknotted RNAs `1E95_A` and `2N8V_X` exhibit both a number of resisting contacts around 70, lying in the same range of contacts displayed by several RNAs at the 3′-terminal. We observe in fact from the right panel of Fig. 5.1 that these RNAs lie in the center-right of the plot, that is the area corresponding to "large" activation time and "moderate" number of contacts.

The remaining four RNAs that activated only from the 3′-end, that is the RNAs `2AP5_A`, `4C4Q_N`, `4PQV_A`, `5TPY_A` and `7K16_P`, are represented by the datapoints with largest t_a and contacts count in the top-right corner, extending the trend observed in the static case. All these molecules present more than approximately 80 resisting contacts at the 5′-terminal, a value larger than most of the observed values at the 3′-end across the whole set of selected RNAs. Notice two RNAs belonging to this group share the largest values for both the activation times and the number of contacts: the largest activation time is found for the ZIKV xrRNA (label `5TPY_A`), while the *H*-type pseudoknot labeled `2AP5_A` makes the largest number of contacts at the 5′-end. All these four RNAs share a number of contacts at the 5′-terminal which exceeds the typical value of contacts at the 5′-end.

Analysis

A simple metric based solely on the number of terminal contacts has several limitations. First, for the sake of simplicity, we treated all contacts equally, regardless of their nature. This allowed us to investigate easily whether an imbalance in the number of contacts made by the atoms of the terminal nucleotides could account for the observed asymmetry in the activation times. Given the heterogeneous set of interactions present in RNA molecules, we did not distinguish between different interaction types but simply counted the contacts involving the two chain termini during the translocation process. Second, we know that the also the *coupling* between the molecular structures and the force plays a role during the activation process. For example, we recall that it was shown in Ref. [45] the presence of a different mechanism response upon

external force inside the pore. In particular, the structure of the ZIKV xrRNA responds by triggering a counterintuitive molecular tightening when pulled in the $5' \to 3'$ direction, while the helix at the $3'$-end unzips more straightforwardly.

Our objective, however, was to isolate and assess the contribution to the activation process of the fold of the viral RNAs and their network of interaction between nucleotides at the ends. Within this simplified framework, we do observe a tangible (though moderate) correlation between the activation times and the number of terminal contacts, indicating that the terminal connectivity explains a part of the observed directional asymmetry. Molecules with termini particularly rich in contacts, such as the pseudoknotted RNAs `2AP5_A`, `4C4Q_N`, `4PQV_A`, `5TPY_A`, and `7K16_P`, confirm the correlation between these two quantities. In particular, the largest activation time is found for the ZIKV xrRNA, showing how the intricate topology of the xrRNA dense with actually resisting terminal contacts corroborates the results of previous works from our group, highlighting the exceptional resistance to the translocation of the ZIKV xrRNA [45, 47]. However, other molecules like the RNAs `1E95_A` and `2N8V_X` display a number of resisting contacts at the $5'$-end which is similar to the typical value observed by the selected RNAs at the other terminal; still, they activate with delay when pulled by the resisting $5'$-end.

This suggests that not only the number of contacts, but also their *spatial organization* is crucial for the explanation of the observed asymmetry in the activation times. As each molecule is driven through the pore, the interaction network is distorted, bonds break, and the deformation involves the whole chain as the translocation proceeds. From the partial results of this Section and the previous results of Ref. [45] we can infer that the mechanical response likely depends in general on how stresses propagate through this evolving network, rather than simply on the stationary connectivity of the terminal nucleotides.

Several extensions to this analysis are possible. First, the analysis of Ref. [45] could be addressed to the whole set of RNAs to verify which kind of RNA architectures respond similarly when translocated through a nanopore by means of a driving force. Moreover, focusing on the graph-based approach of Sec. 5.1.1 where an RNA is modeled as a graph where nodes are nucleotides and edges are bonds between them, one could consider not only the primary contacts involving the terminal nucleotides but also secondary contacts, *i.e.*, accounting also for the interactions formed by nucleotides connected to the five terminal residues hitherto considered. In this way, one would incorporate a larger portion of the interaction network responsible for the transmission of the pulling force to the whole structure. Furthermore, interactions with the pore, the slab, and the applied pulling force all contribute to both activation and translocation, as illustrated by the effects of pore constrictions discussed in Sec. 3.4. Also, with the representation of each RNA as a graph it is possible to make use of the full arsenal of graph-based observables like various flavors of connectivity and centrality (betweenness centrality, closeness centrality, harmonic centrality, ...) of nucleotides to identify topological key hubs of nucleotides. For example, a description based on directed graphs could address the directional response to the forced translocation to discriminate the two possible translocation directions. Finally, elastic network models (ENMs), which were already employed to successfully describe the internal dynamics of proteins [146], could be used in principle to analyze also RNA molecules.

5.2 Structural Description of the "Hard Cases"

We have shown in the previous Section how the sole number of contacts is not sufficient to explain extensively the asymmetric response to the forced translocation for the heterogeneous set of RNAs selected, as proved by the mild correlation between contacts located at the chain ends and the corresponding activation times. In particular, among the most delayed RNAs in the $5' \to 3'$ direction of translocation, the RNAs `1E95_A` and `2N8V_X` have a number of contacts at the $5'$-end

which is comparable to the average number of contacts at the 3′-end, yet they cannot activate the translocation when pulled by the 5′-end with the selected force. Also the RNAs labeled `2AP5_A`, `4C4Q_N`, `4PQV_A`, `5TPY_A` and `7K16_P`, which are the ones that show the largest number of contacts at the critical terminal, did not activate at all when pulled by the 5′-end unless the driving force increases sufficiently over time. Thus, we provide here a structural description of all these "hard" cases around the resisting 5′-terminal to show common characteristics and systematicities across different structures with analogous translocation resistance.

We sort these molecules into three groups, where the RNAs within each group share some structural similarities. The first group comprises the two H-type pseudoknots `1E95_A` and `2AP5_A`, together with the IRES domain RNA `4C4Q_N`, and it consists of the RNAs for which the 5′-end is paired to a loop, forming different base triples between the terminal nucleotides and the U-shaped strand. Then, there are three xrRNAs, RNAs `4PQV_A`, `5TPY_A` and `7K16_P`, all of which exhibit the typical ring-like structure which encircles the 5′-terminus. Finally, the remaining RNA `2N8V_X`, not belonging to any of the previous groups is self-standing and will be analyzed separately.

5.2.1 Lasso and Resisting End

The molecules belonging to this group display a similar topological feature around the 5′-end: the molecule terminus is paired with a lasso-like arrangement of the remaining nucleotides. Among the 16 selected RNAs, this aspect is common to these three RNAs only, namely the two H-type pseudoknots `1E95_A` and `2AP5_A`, and the IRES domain RNA `4C4Q_N`, see also Fig. 2.13 which contains the cartoon representation of all the viral RNAs selected.

The presence of the pairing between the 5′-end and a lasso-shaped strand results in the formation of triple strand contacts. In particular, the terminal nucleotides at the 5′-end are inserted into the groove of an already-formed helical stem, so that three strands come into simultaneous contact at a localized helical junction. The threaded 5′-end runs antiparallel to one of the strands of the noose, still forming contacts with the other half of the lasso. Non-Watson-Crick contacts form in these local arrangements of nucleotides multiple base triplets rather than nucleotide pairs.

Notice this does not hold in general for the selected pseudoknotted molecules. For example, also the RNA `1RNK_A`, another example of H-type pseudoknot, has the 5′-end running antiparallel to the paired strand; however, it does not form a triple helix as in these cases, as the nucleotides in the 3′-end sides of the lasso are exposed to the exterior of the nucleic acid, preventing the formation of base triples, see Fig. 2.13. We recall that although the median activation time $t_{\mathrm{a}}^{5'\to3'}$ is among the largest values observed, this molecule activates its translocation at $F_{\mathrm{s}} = 20F_0$ in the static force pulling protocol, contrary to the three RNAs presented here.

The presence of base triples arising from triple helices was well documented in the original references of the two H-type pseudoknots [96, 101]. Regarding the pseudoknot of the mRNA of SRV-1 `1E95_A`, the structure shows a compact H-type pseudoknot fold where the interface between the 5′-terminus and the lasso conformation involving nucleotides C20-U27 is highly structured. Two base triples involving the terminal 5′-terminal nucleotides were probed by the NMR experiment, see Fig. 5(b)-(d) of Ref. [96]. The first triplet consists of nucleotides G4, C16 and A26, while the other documented base triple involves nucleotides C5, G15 and U27. Furthermore, the terminal C2 residue contributes to the triple helix stability by binding to G18 and A21. This configuration dense of contacts has no counterpart at the 3′-end, where the terminal residues C36 and A35 are not involved in base pairings and initiate the molecule unzipping when pulled in the $3' \to 5'$ direction. We represented the molecule in the top panel of Fig. 5.2, highlighting the lasso and the terminal 5′-strand with colors, and shading the rest of the molecule.

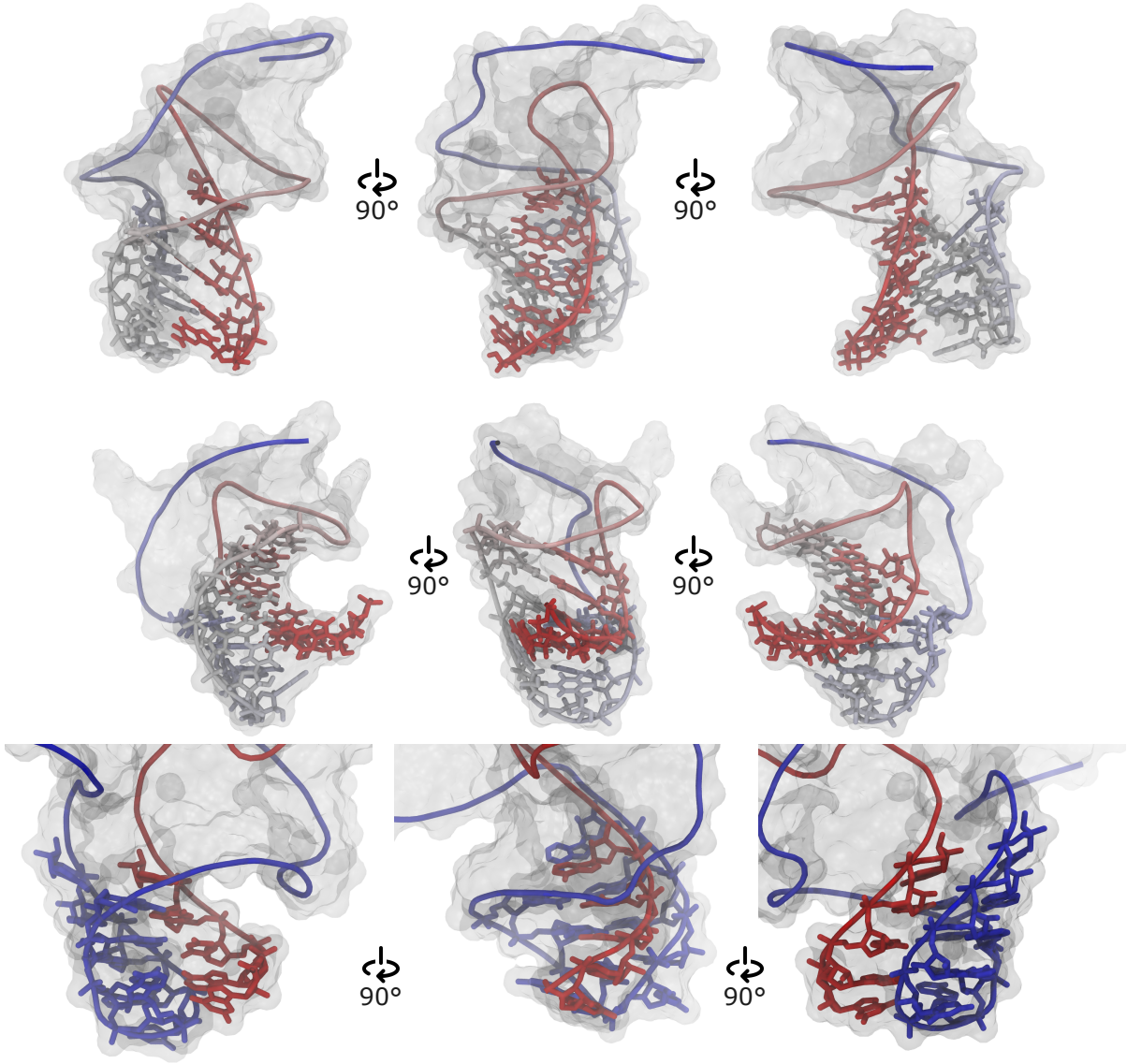


Figure 5.2 Representation of the 5′-ends paired with a lasso-like structure from three different RNAs: `1E95_A` (top), `2AP5_A` (middle), and `4C4Q_N` (bottom). The RNA `4C4Q_N` (233 nucleotides) was represented only near the 5′-end to highlight that portion of the structure, whereas the other RNAs were represented fully due to their smaller size. The terminal nucleotides and the lasso-like structure were represented with color gradient from red (5′-end) to blue (3′-end), and a global shade of the molecule surfaces is provided. The tubes in the various panels represent the backbone of each RNA and run through all the phosphate atoms. The RNAs are not to scale with each other. All images were rendered with VMD [82].

A similar configuration occurs at the 5′-terminal of the 28-nucleotide ScYLV mRNA pseudoknot `2AP5_A`, see Fig. 4(a) of Ref. [101]. In this case, the triple helix is nearly continuous and formed by the terminal A3-G7 stretch of nucleotides which bind to the loop spanned by the C14-U18 and A21-A27 strands. However, the overall architecture is more elaborated, as the triple helix is capped by a base quadruple due to the C8 residue forming bonds with G12, A27 and C28. In this way, the 3′-end is entirely docked onto the 5′-stem residue-by-residue [101]. This coaxial stacking confers a pronounced global twist to the overall architecture of the pseudoknot, which can be visualized in the middle panel of Fig. 5.2.

The global structure of the RNA belonging to the CSFV IRES `4C4Q_N` is much larger than the two previous pseudoknots (233 nucleotides versus approximately 30 nucleotides). Nevertheless, the terminal 5′-strand shares some similarities with the aforementioned pseudoknots. In particular, during the $5' \to 3'$ translocation, the pulled end (namely, the domain II of the whole molecule, see Fig. 3(a) of Ref. [106]) is basically a pseudoknot which is topologically similar to the small isolated pseudoknots above, just embedded in a much larger RNA. We can observe the similarities with the previous descriptions from the bottom panel of Fig. 5.2. However, in this case the lasso is broader and more deformed compared to the other two pseudoknots where the lassos wrap the corresponding 5′-termini. As for the RNA `1E95_A`, also in this case the 3′-end is "free" and not involved in such characteristic conformation, as shown be the bottom-right panel in Fig. 5.2. In fact, as shown in the previous Section, the activation by the 3′-end is triggered more easily, possibly due to the simpler unzipping mechanism of such end which has no counterpart at the 5′-terminal.

5.2.2 xrRNAs

The second group comprises the RNAs `4PQV_A`, `5TPY_A`, and `7K16_P`, which are xrRNA modules derived from different viruses, specifically MVEV, ZIKV, and TABV, respectively.

xrRNAs are in general structured RNA modules located at the 3′ untranslated region (UTR) in the genome of different viruses from the *Flaviviridae* family. These viruses contain positive-sense single stranded RNA (+ssRNA) consisting of a single genomic RNA. These RNA modules flank the reading frame of the viral RNA at the 3′-end, conferring an exceptional resistance to the RNA degradation by means of *exoribonucleases*, enzymes recruited by the host immune system to eliminate viral RNA by cleaving the terminal nucleotides from its termini [61–65]. Such resistance is general, that is a single xrRNA module is able to halt diverse nucleases, revealing a non-specific response to degradation that must be traced back solely to the structure of this module [66].

The fascinating characteristic of the architecture of xrRNAs is a conserved ring-like structure that encircles the 5′-end with a single [62–64] or even double loop [65]. Interestingly, this topological feature of xrRNAs was observed also in plant-infecting viruses, which are phylogenetically unrelated to flaviviruses [67]. The resistance to the forced translocation in a specific direction, common feature of all xrRNAs, must be traced back to this common structural feature of xrRNAs. The mechanical response of the ZIKV xrRNA (molecule labeled `5TPY_A`) was addressed computationally in Ref. [45], where it was shown how this circular motif wraps around the pulled 5′-end during the forced $5' \to 3'$ translocation, tightening the RNA around such terminal and easing the distribution of the pulling force over a denser network of contacts.

At first glance, the three molecules grouped here seem to share a global similarity with such circular element around the 5′-end, see Fig. 5.3. However, it is possible to detect intrinsic subtle differences. For example, their topological complexities differ, as shown also by the grouping performed in Sec. 2.2.8: in particular, the MVEV xrRNA `4PQV_A` comprises only one pseudoknot, while the other two xrRNAs display a more complex fold with four putative pseudoknots. In fact, in the experimental work which solved the fold of the ZIKV xrRNA `5TPY_A`, these two

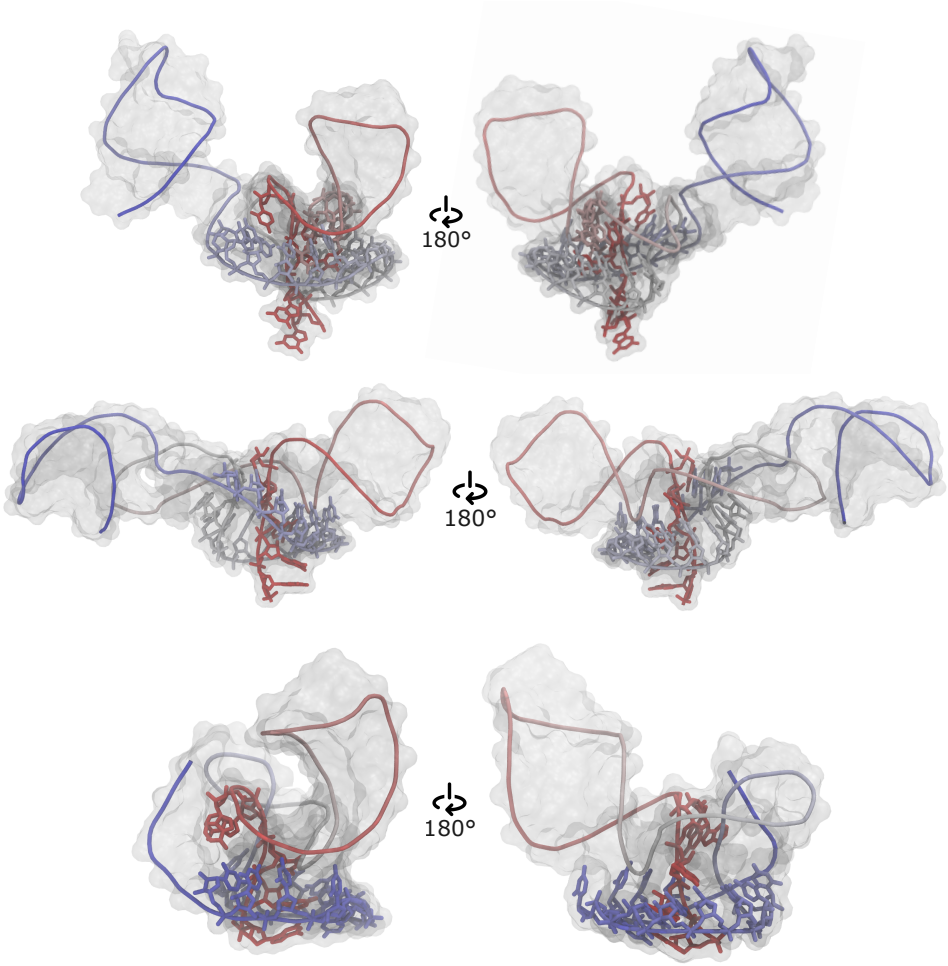


Figure 5.3 Structural representation the three xrRNAs present in the selected list of RNAs: `4PQV_A` (above), `5TPY_A` (middle), `7K16_P` (bottom). The terminal five nucleotides at the 5′-end and the nucleotides involved in the typical ring-like structure encircling the 5′-end are represented with atomistic detail. The tube in the figure represents the backbone of each RNA and runs through the phosphate atoms from the 5′-end (red) to the 3′-one (blue). A solid shading is provided to give an idea of the molecules surface. The molecules are not to scale with each other. Images generated with VMD [82].

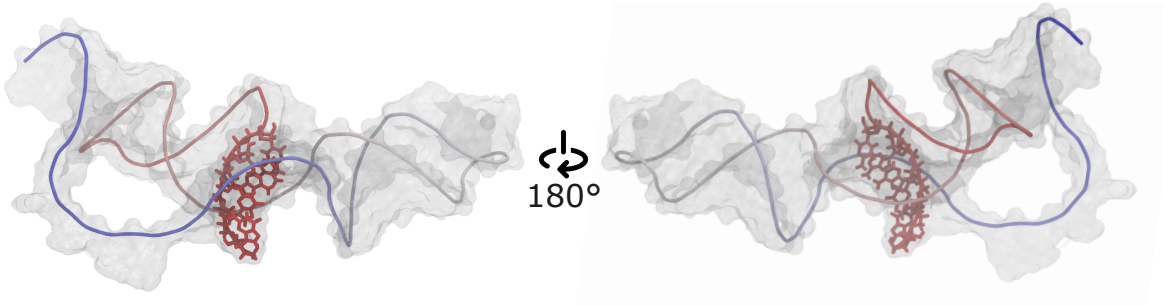


Figure 5.4 Structural representation the *Israeli acute paralysis virus* (IAPV) IRES pseudoknot domain I, labeled `2N8V_X`. The terminal five nucleotides at the 5′-end and the backbone are highlighted and colored with a palette shading from red (5′-end) to blue (3′-end). A shading of the three-dimensional surface is provided to give an idea of the molecules shape. Images generated with VMD [82].

structures were compared to each other [62] and it was observed a significant change in the position of the terminal hairpin at the 3′-end of the xrRNA `4PQV_A`, which we also verified with the structural alignment performed with SARA (see Sec. 2.2.6). Furthermore, it is explicitly stated in Ref. [62] that an important pseudoknot in the MVEV xrRNA `4PQV_A` is *not* formed, contrary to the fully pseudoknotted ZIKV xrRNA. This is likely to be linked to the localization of these xrRNA modules in their respective viral genomes: in fact, xrRNA are often displayed in series in the UTR of flaviviruses. It was shown that the MVEV xrRNA is *downstream* of two in series, while the latter ZIKV xrRNA is *upstream* of two. Notice upstream xrRNAs are predicted to have a more stable fold, whose loss has a greater effect on viral infection [124, 125].

The xrRNAs `4PQV_A` and `5TPY_A` display both two helical domains which sprout from the protective ring around the 5′-end, one of which comprises the other end of the molecule (see the blue part of the tube representation in Fig. 5.3). In the third molecule `7K16_P` only one helix stems from the circular motif around the 5′-end: the terminal helical domain is not present, but a free end caps the circular element around the 5′-end. The protective ring of xrRNA `7K16_P` has a diameter of approximately 28 Å, while the ZIKV ring spans a larger diameter of 31 Å, according to the respective experimental studies [62, 64]. The smaller constriction changes the composition of nucleotides and the interactions that are formed across this circle. Nevertheless, all these different circular structures shield in the same manner the xrRNAs from the translocation in the $5' \to 3'$, showing how this structural feature is a key defensive mechanism of viruses to avoid degradation.

5.2.3 The last RNA

Among the seven RNAs that did not activate from the 5′-end in the static force pulling protocol, the RNA `2N8V_X` could not be classified together with any of the other molecules, as it is not an xrRNA nor its 5′-terminus is involved in the strand-and-lasso architecture we highlighted in Sec. 5.2.1. The fold of this pseudoknotted molecule is quite different from all the other molecules selected in this work, as it can be clearly observed by the whole set of structures represented in Fig. 2.13. The structure assumed by this viral RNA mimics the typical L-shape molecule of tRNAs to facilitate reading frame selection and engaging the host ribosome [103]. In particular, the 5′-end is located in the "core" of the imitated tRNA, whereas the 3′-terminus is in correspondence of the anticodon arm.

The terminal five nucleotides at 5′-end are capped by a three-way junction involving three helices. Two unpaired bulged adenosine nucleotides at the junction were shown to be important

for mediating the overall shape of the pseudoknot [103]. This junction was shown in Ref. [103] to be crucial for the reading frame selection, allowing to switch between the 0 and +1 reading frames during the translation. We recall that the nanopore translocation can be used as a convenient proxy for the processing of an RNA by means of other enzymes. Thus, the crucial role of this junction during the translocation should affect the nanopore translocation of RNA `2N8V_X` as well. Moreover, we observe that the $5'$-end is flanked by two single strands which increase the obstruction inside the pore. The $3'$-end, on the other hand, is more flexible and presents a loose strand which makes few contacts with the rest of the RNA.

5.3 Summary

In this Chapter we have analyzed the RNA structures to identify what could be the key factor that rules the asymmetry of the activation process of the two chain ends. Focusing on the terminal five nucleotides of each RNA, we verified that the number of resisting contacts plays only a partial role in the delay of the activation of the $5'$-end. The selected RNAs display generally a larger density of resisting contacts (retained from the MD trajectories during the activation stage) at the $5'$-end, but the number of contacts by itself is not sufficient to fully explain the asymmetry observed. As shown by the mild correlation evaluated with the Kendall-τ test ($\tau = 0.49$ and $\tau = 0.47$ for the static and ramping case, respectively), and the small p-values obtained (p-value $= 4.53 \times 10^{-4}$ for the static case and p-value $= 1.38 \times 10^{-4}$ for the ramping pulling protocol), the total resistance strength provided by the contacts play only a relative role in the translocation resistance. The different correlation values are due to the fact that in the ramping protocol *all* RNA molecules activate the translocation within the allocated computational time at the reference pulling rate $r = \frac{F_0}{22\times200}\frac{1}{\tau_{\mathrm{MD}}}$.

Then, we provided a structural description of the seven RNAs which activated only in the $3' \to 5'$ direction when pulled with the static force $F_{\mathrm{s}} = 20F_0$, namely the molecules labeled `1E95_A`, `2AP5_A`, `2N8V_X`, `4C4Q_N`, `4PQV_A`, `5TPY_A` and `7K16_P`. All these structures present pseudoknotted folds, contrary to most of the RNAs which activated faster and in both pulling directions. Some common topological features were identified, and in all cases the $5'$-terminal is involved in a series of multiple interactions with the rest of the molecule. In the case of RNAs `1E95_A`, `2AP5_A`, and `4C4Q_N`, the $5'$-end is coupled with a lasso-like strand forming a compact segment comprising several base triples, which could be the cause of the delay in the $5' \to 3'$ direction. The three molecules `4PQV_A`, `5TPY_A` and `7K16_P` show the typical feature of xrRNAs, namely a circular stretch of nucleotides which wraps around the $5'$-end. This protective ring was shown to act as a molecular brace during the $5' \to 3'$ translocation, delaying the activation by orders of magnitude with respect to the $3'$-end activation. Finally, in the tRNA-like RNA `2N8V_X` the $5'$-end is capped by a three-way junction governed by two key adenosines. It was shown experimentally that this junction does affect the translocation of the RNA, allowing the switch to different reading frames.

We wish our analyses and descriptions can shed new light on possible key players driving the pronounced asymmetry in these RNAs, serving as a ground base for further studies on this subject.

Conclusions

In this Thesis, we performed MD simulations of translocation of viral RNA through a nanopore to observe the impact that the structure of nucleic acids has on the translocation. Polymer translocation is a ubiquitous process in biology, involved in several processes from cellular nuclear transport, to protein synthesis, to viral infection. In recent years, polymer translocation showed a wide range of applications in Science, from nucleic acid sequencing to being a useful probe for the structure of translocating polymers.

The structures of molecules affect their translocation properties through nanopores. For this reason, viruses are the ideal substrate for the study of nanopore translocation, given their genomes with interesting and nontrivial topologies comprising pseudoknots, bulges, multiple-strand junctions and hairpins. These three-dimensional domains are due to their limited size and their ability to perform multiple tasks.

In Chapter 2 we selected the set of RNA chains of viral origin to simulate and translocate through the nanopore. The selection was based on taxonomical and structural criteria avoiding redundancy in structures and sequences, resulting in the retrieval of 16 nucleic acids of viral origin. The selected molecules display a heterogeneous set of structures with increasing topological complexity, which we quantified with the number of putative pseudoknots.

In Chapter 3 we showed how the computational setup for the nanopore translocation was prepared by calibrating the temperature on heuristic observations on the melting curves. The the nanopore size and the magnitude of the pulling force were tuned as the optimal compromise between pulling intensity and the allocated computational resources. We investigated two pulling protocols, namely a constant force and a linearly-increasing one, for which several pulling rates were selected.

In Chapter 4 we presented the main results of this Thesis. Regardless of the pulling protocol employed, for the set of employed forces the selected RNAs show an asymmetric response to the forced translocation. In particular, the onset of the translocation (that is, the activation time t_{a}) is consistently delayed throughout all the molecules when translocating them in the $5' \to 3'$ direction. In the constant pulling force case, seven of the selected 16 RNAs did not activate at all when pulled by the $5'$-end within the allocated simulation time. We observed that an increasing topological complexity of the translocating chains affects the onset of their translocation by delaying the activation by orders of magnitude at the selected range of forces. Moreover, the activation times under the two tested protocols are strongly correlated with each other, hinting that the response to the forced translocation could be traced back solely to the RNA fold.

Finally, we explored in Chapter 5 the possible structural rationale behind the asymmetric response of the chains to the forced translocation. We investigated the correlation between contact density at the two ends and the duration activation stage, using the number resisting contacts measured during the activation stage as a simple proxy for the activation barriers at the RNA termini, which dictate the activation time t_{a}. Our analyses show that only a mild correlation emerges between the two observables, suggesting that the structure by itself

is not sufficient to fully explain the dichotomic response to the translocation. Previous works from our group showed for the nanopore translocation of the ZIKV xrRNA how the coupling with the pulling force should be considered, as the distortion of the interaction network of the nucleotides at the $5'$-end is responsible for the larger offset of that end. Then, we inspected the molecular structures of the seven most resisting RNAs, that is the ones which activated their translocation only in the $3' \rightarrow 5'$ direction within the allocated CPU time at the selected pulling force. The folds of these molecules present nontrivial topologies comprising pseudoknots. Some common traits were identified across these seven RNAs. Triple helices or junctions were found in correspondence of the $5'$-end, hinting that these conformations are particularly resistant to the stress exerted by the pulling force.

In conclusion, in this Thesis we performed a systematic study on the effects of the topology of RNA on their translocation. Viruses are a fertile ground of investigation for this topic, given the topological variety and complexity of their genomes. We observed an intrinsic asymmetry in the activation of the translocation process, which we tried to trace back to some structure-related observable. The density of contacts at one of the ends correlate only partially with the onset of the translocation, hinting that the structures by themselves are not sufficient to fully characterize the asymmetry. A visual inspection on the most resisting RNAs revealed that some common denominators can be found. In particular, triple-helices and junctions are located at the $5'$-termini of these molecules, suggesting that these kinds of nucleotides arrangements could be a convenient resource to withstand and block the translocation in specific directions.

Bibliography

[1] Francis H Crick. "On protein synthesis". In: *Symp soc exp biol.* Vol. 12. 138-63. 1958, p. 8.

[2] Sydney Brenner, François Jacob, and Matthew Meselson. "An unstable intermediate carrying information from genes to ribosomes for protein synthesis". In: *Nature* 190.4776 (1961), pp. 576–581.

[3] Francois Gros et al. "Unstable Ribonucleic Acid Revealed by Pulse Labelling of Escherichia Coli". In: *Nature* 190.4776 (1961), pp. 581–585. DOI: 10.1038/190581a0. URL: https://doi.org/10.1038/190581a0.

[4] George E Palade. "A small particulate component of the cytoplasm". In: *The Journal of biophysical and biochemical cytology* 1.1 (1955), p. 59.

[5] Mahlon B Hoagland et al. "A soluble ribonucleic acid intermediate in protein synthesis". In: *Journal of Biological Chemistry* 231.1 (1958), pp. 241–257.

[6] Harry F Noller, Vernita Hoffarth, and Ludwika Zimniak. "Unusual resistance of peptidyl transferase to protein extraction procedures". In: *Science* 256.5062 (1992), pp. 1416–1419.

[7] Kelly Kruger et al. "Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena". In: *cell* 31.1 (1982), pp. 147–157.

[8] Alexey G Vitreschak et al. "Riboswitches: the oldest mechanism for the regulation of gene expression?" In: *TRENDS in Genetics* 20.1 (2004), pp. 44–50.

[9] Kevin V Morris et al. "Small interfering RNA-induced transcriptional gene silencing in human cells". In: *Science* 305.5688 (2004), pp. 1289–1292.

[10] David P Bartel. "MicroRNAs: genomics, biogenesis, mechanism, and function". In: *cell* 116.2 (2004), pp. 281–297.

[11] Stefanie Jonas and Elisa Izaurralde. "Towards a molecular understanding of microRNA-mediated gene silencing". In: *Nature reviews genetics* 16.7 (2015), pp. 421–433.

[12] Jenny KW Lam et al. "siRNA versus miRNA as therapeutics for gene silencing". In: *Molecular therapy Nucleic acids* 4 (2015).

[13] Bruce Alberts et al. *Molecular biology of the cell.* Vol. 3. Garland New York, 1994.

[14] Sara Nakielny et al. "RNA transport". In: *Annual review of neuroscience* 20.1 (1997), pp. 269–301.

[15] Yi Jiang, Wanchao Yin, and H Eric Xu. "RNA-dependent RNA polymerase: Structure, mechanism, and drug discovery for COVID-19". In: *Biochemical and biophysical research communications* 538 (2021), pp. 47–53.

[16] David Deamer, Mark Akeson, and Daniel Branton. "Three decades of nanopore sequencing". In: *Nature biotechnology* 34.5 (2016), pp. 518–524.

[17] Amit Meller, Lucas Nivon, and Daniel Branton. "Voltage-driven DNA translocations through a nanopore". In: *Physical review letters* 86.15 (2001), p. 3435.

[18] John J Kasianowicz et al. "Characterization of individual polynucleotide molecules using a membrane channel". In: *Proceedings of the National Academy of Sciences* 93.24 (1996), pp. 13770–13773.

[19] Mark Akeson et al. "Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single RNA molecules". In: *Biophysical journal* 77.6 (1999), pp. 3227–3233.

[20] Nurit Ashkenasy et al. "Recognizing a single base in an individual DNA strand: A step toward nanopore DNA sequencing". In: *Angewandte Chemie (International ed. in English)* 44.9 (2005), p. 1401.

[21] Yunhao Wang et al. "Nanopore sequencing technology, bioinformatics and applications". In: *Nature biotechnology* 39.11 (2021), pp. 1348–1365.

[22] Cees Dekker. "Solid-state nanopores". In: *Nature nanotechnology* 2.4 (2007), pp. 209–215.

[23] Jérôme Mathé et al. "Orientation discrimination of single-stranded DNA inside the α-hemolysin membrane channel". In: *Proceedings of the National Academy of Sciences* 102.35 (2005), pp. 12377–12382.

[24] Jérôme Mathé et al. "Nanopore unzipping of individual DNA hairpin molecules". In: *Biophysical Journal* 87.5 (2004), pp. 3205–3212.

[25] Andrew M Smith et al. "Capture, unfolding, and detection of individual tRNA molecules using a nanopore device". In: *Frontiers in bioengineering and biotechnology* 3 (2015), p. 91.

[26] Yana Rozevsky et al. "Quantification of mRNA expression using single-molecule nanopore sensing". In: *ACS nano* 14.10 (2020), pp. 13964–13974.

[27] Gary M Skinner et al. "Distinguishing single-and double-stranded nucleic acid molecules using solid-state nanopores". In: *Nano Letters* 9.8 (2009), pp. 2953–2960.

[28] Prabhat Tripathi et al. "Discrimination of RNA fiber structures using solid-state nanopores". In: *Nanoscale* 14.18 (2022), pp. 6866–6875.

[29] Filip Bošković and Ulrich Felix Keyser. "Nanopore microscope identifies RNA isoforms with structural colours". In: *Nature Chemistry* 14.11 (2022), pp. 1258–1264.

[30] Chalmers Chau et al. "Probing RNA conformations using a polymer–electrolyte solid-state nanopore". In: *ACS nano* 16.12 (2022), pp. 20075–20085.

[31] Michael Zwolak and Massimiliano Di Ventra. "Colloquium: Physical approaches to DNA sequencing and detection". In: *Reviews of Modern Physics* 80.1 (2008), pp. 141–165.

[32] Daniel R Garalde et al. "Highly parallel direct RNA sequencing on an array of nanopores". In: *Nature methods* 15.3 (2018), pp. 201–206.

[33] Andrew M Smith et al. "Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing". In: *PloS one* 14.5 (2019), e0216709.

[34] Jeffrey Chuang, Yacov Kantor, and Mehran Kardar. "Anomalous dynamics of translocation". In: *Physical Review E* 65.1 (2001), p. 011802.

[35] Yacov Kantor and Mehran Kardar. "Anomalous dynamics of forced translocation". In: *Physical Review E* 69.2 (2004), p. 021806.

[36] Ulrich Gerland, Ralf Bundschuh, and Terence Hwa. "Translocation of structured polynucleotides through nanopores". In: *Physical biology* 1.1 (2004), p. 19.

[37] David K Lubensky and David R Nelson. "Driven polymer translocation through a narrow pore". In: *Biophysical journal* 77.4 (1999), pp. 1824–1838.

[38] Vladimir V Palyulin, Tapio Ala-Nissila, and Ralf Metzler. "Polymer translocation: the first two decades and the recent diversification". In: *Soft matter* 10.45 (2014), pp. 9016–9037.

[39] Takuya Saito and Takahiro Sakaue. "Dynamical diagram and scaling in polymer driven translocation". In: *The European Physical Journal E* 34.12 (2011), p. 135.

[40] Timo Ikonen et al. "Unifying model of driven polymer translocation". In: *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 85.5 (2012), p. 051803.

[41] Minglun Li and Murugappan Muthukumar. "RNA Translocation through Protein Nanopores: Interlude of the Molten RNA Globule". In: *Journal of the American Chemical Society* 147.2 (2024), pp. 1553–1562.

[42] Angelo Rosa, M Di Ventra, and Cristian Micheletti. "Topological jamming of spontaneously knotted polyelectrolyte chains driven through a nanopore". In: *Physical review letters* 109.11 (2012), p. 118301.

[43] Antonio Suma, Angelo Rosa, and Cristian Micheletti. "Pore translocation of knotted polymer chains: How friction depends on knot complexity". In: *ACS Macro Letters* 4.12 (2015), pp. 1420–1424.

[44] Antonio Suma and Cristian Micheletti. "Pore translocation of knotted DNA rings". In: *Proceedings of the National Academy of Sciences* 114.15 (2017), E2991–E2997.

[45] Antonio Suma et al. "Directional translocation resistance of Zika xrRNA". In: *Nature communications* 11.1 (2020), p. 3749.

[46] Xiaolin Niu et al. "Molecular mechanisms underlying the extreme mechanical anisotropy of the flaviviral exoribonuclease-resistant RNAs (xrRNAs)". In: *Nature Communications* 11.1 (2020), p. 5496.

[47] Matteo Becchi et al. "RNA Pore Translocation with Static and Periodic Forces: Effect of Secondary and Tertiary Elements on Process Activation and Duration". In: *The Journal of Physical Chemistry B* 125.4 (2021), pp. 1098–1106.

[48] Stephen Harrison. "Principles of virus structure". In: *Virology*. Raven Press New York, 1985, pp. 27–44.

[49] Milton W Taylor. "What is a virus?" In: *Viruses and man: a history of interactions*. Springer, 2014, pp. 23–40.

[50] Charles M Knobler and William M Gelbart. "How and why RNA genomes are (partially) ordered in viral capsids". In: *Current Opinion in Virology* 52 (2022), pp. 203–210.

[51] Jennifer Louten. "Virus structure and classification". In: *Essential human virology* (2016), p. 19.

[52] David W Staple and Samuel E Butcher. "Pseudoknots: RNA structures with diverse functions". In: *PLoS biology* 3.6 (2005), e213.

[53] Beth L Nicholson and K Andrew White. "Functional long-range RNA–RNA interactions in positive-strand RNA viruses". In: *Nature Reviews Microbiology* 12.7 (2014), pp. 493–504.

[54] Zane A Jaafar and Jeffrey S Kieft. "Viral RNA structure-based strategies to manipulate translation". In: *Nature Reviews Microbiology* 17.2 (2019), pp. 110–123.

[55] David A Costantino et al. “tRNA–mRNA mimicry drives translation initiation from a viral IRES”. In: *Nature structural & molecular biology* 15.1 (2008), pp. 57–64.

[56] Ian Brierley. “Ribosomal frameshifting on viral RNAs”. In: *Journal of General Virology* 76.8 (1995), pp. 1885–1892.

[57] David P Giedroc and Peter V Cornish. “Frameshifting RNA pseudoknots: structure and mechanism”. In: *Virus research* 139.2 (2009), pp. 193–208.

[58] Jonathan D Dinman. “Mechanisms and implications of programmed translational frameshifting”. In: *Wiley Interdisciplinary Reviews: RNA* 3.5 (2012), pp. 661–673.

[59] Andrii Slonchak and Alexander A Khromykh. “Subgenomic flaviviral RNAs: What do we know after the first decade of research”. In: *Antiviral research* 159 (2018), pp. 13–25.

[60] Quentin Vicens and Jeffrey S Kieft. “Shared properties and singularities of exoribonuclease-resistant RNAs in viruses”. In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 4373–4380. DOI: 10.1016/j.csbj.2021.07.024. URL: https://spj.science.org/doi/abs/10.1016/j.csbj.2021.07.024.

[61] Erich G Chapman et al. “RNA structures that resist degradation by Xrn1 produce a pathogenic Dengue virus RNA”. In: *elife* 3 (2014), e01892.

[62] Benjamin M Akiyama et al. “Zika virus produces noncoding RNAs using a multi-pseudoknot structure that confounds a cellular exonuclease”. In: *Science* 354.6316 (2016), pp. 1148–1152.

[63] Erich G Chapman et al. “The structural basis of pathogenic subgenomic flavivirus RNA (sfRNA) production”. In: *Science* 344.6181 (2014), pp. 307–310.

[64] X De Lamballerie et al. “Genome sequence analysis of Tamana bat virus and its relationship with the genus Flavivirus”. In: *Journal of General Virology* 83.10 (2002), pp. 2443–2454.

[65] Conner J Langeberg et al. “Tick-borne flavivirus exoribonuclease-resistant RNAs contain a double loop structure”. In: *Nature Communications* 16.1 (2025), p. 4515.

[66] Andrea MacFadden et al. “Mechanism and structural diversity of exoribonuclease-resistant RNA structures in flaviviral RNAs”. In: *Nature Communications* 9.1 (2018), p. 119.

[67] Anna-Lena Steckelberg et al. “A folded viral noncoding RNA blocks host cell exoribonucleases through a conformationally dynamic RNA structure”. In: *Proceedings of the National Academy of Sciences* 115.25 (2018), pp. 6404–6409.

[68] Ailong Ke and Jennifer A Doudna. “Crystallization of RNA and RNA–protein complexes”. In: *Methods* 34.3 (2004), pp. 408–414.

[69] Jinwei Zhang and Adrian R Ferré-D’Amaré. “Dramatic improvement of crystals of large RNAs by cation replacement and dehydration”. In: *Structure* 22.9 (2014), pp. 1363–1371.

[70] HaJeung Park et al. “Controlled dehydration improves the diffraction quality of two RNA crystals”. In: *BMC Structural Biology* 16.1 (2016), p. 19.

[71] Gabriele Varani and Ignacio Tinoco Jr. “RNA structure and NMR spectroscopy”. In: *Quarterly reviews of biophysics* 24.4 (1991), pp. 479–532.

[72] Gabriele Varani, Fareed Aboul-Ela, and Frédéric H-T Allain. “NMR investigation of RNA structure”. In: *Progress in Nuclear Magnetic Resonance Spectroscopy* 29.1-2 (1996), pp. 51–127.

[73] Boris Fürtig et al. “NMR spectroscopy of RNA”. In: *ChemBioChem* 4.10 (2003), pp. 936–962.

[74] Haiyun Ma et al. “Cryo-EM advances in RNA structure determination”. In: *Signal Transduction and Targeted Therapy* 7.1 (2022), p. 58.

[75] Steve L Bonilla and Jeffrey S Kieft. “The promise of cryo-EM to explore RNA structural dynamics”. In: *Journal of Molecular Biology* 434.18 (2022), p. 167802.

[76] Mira Abraham et al. “Analysis and classification of RNA tertiary structures”. In: *RNA* 14.11 (2008), pp. 2274–2289.

[77] Neocles B Leontis and Craig L Zirbel. “Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking”. In: *RNA 3D structure analysis and prediction* (2012), pp. 281–298.

[78] Nancy Ontiveros-Palacios et al. “Rfam 15: RNA families database in 2025”. In: *Nucleic acids research* 53.D1 (2025), pp. D258–D267.

[79] Nabila Shahnaz Khan et al. “RNA-NRD: a non-redundant RNA structural dataset for benchmarking and functional analysis”. In: *NAR Genomics and Bioinformatics* 5.2 (2023), lqad040.

[80] Scott Federhen. “The NCBI taxonomy database”. In: *Nucleic acids research* 40.D1 (2012), pp. D136–D143.

[81] Emidio Capriotti and Marc A Marti-Renom. “RNA structure alignment by a unit-vector approach”. In: *Bioinformatics* 24.16 (2008), pp. i112–i118.

[82] William Humphrey, Andrew Dalke, and Klaus Schulten. “VMD: visual molecular dynamics”. In: *Journal of molecular graphics* 14.1 (1996), pp. 33–38.

[83] Changbong Hyeon, Ruxandra I Dima, and D Thirumalai. “Size, shape, and flexibility of RNA structures”. In: *The Journal of chemical physics* 125.19 (2006).

[84] Xiang-Jun Lu and Wilma K Olson. “3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures”. In: *Nature protocols* 3.7 (2008), pp. 1213–1227.

[85] Xiang-Jun Lu, Harmen J Bussemaker, and Wilma K Olson. “DSSR: an integrated software tool for dissecting the spatial structure of RNA”. In: *Nucleic acids research* 43.21 (2015), e142–e142.

[86] Thomas C Leeper et al. “TAR RNA recognition by a cyclic peptidomimetic of Tat protein”. In: *Biochemistry* 44.37 (2005), pp. 12362–12372.

[87] David Sehnal et al. “Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures”. In: *Nucleic acids research* 49.W1 (2021), W431–W437.

[88] Fabian Sievers et al. “Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega”. In: *Molecular systems biology* 7.1 (2011), p. 539.

[89] Fabian Sievers and Desmond G Higgins. “Clustal Omega for making accurate alignments of many protein sequences”. In: *Protein Science* 27.1 (2018), pp. 135–145.

[90] Emidio Capriotti and Marc A Marti-Renom. “SARA: a server for function annotation of RNA structures”. In: *Nucleic acids research* 37.suppl_2 (2009), W260–W265.

[91] Shunsuke Imai et al. “An accurately preorganized IRES RNA structure enables eIF4G capture for initiation of viral translation”. In: *Nature structural & molecular biology* 23.9 (2016), pp. 859–864.

[92] A Suma. “Static and dynamic properties of knotted biopolymers: from bulk to nanochannels and nanopores”. Doctoral Dissertation. SISSA, 2017. URL: `https://hdl.handle.net/20.500.14242/168323`.

[93] Werner Krauth and Marc Mézard. "Learning algorithms with optimal stability in neural networks". In: *Journal of Physics A: Mathematical and General* 20.11 (1987), p. L745.

[94] Michal Zurkowski et al. "RNAhugs web server for customized 3D RNA structure alignment". In: *Nucleic Acids Research* (2024), gkae259.

[95] Aric Hagberg, Pieter J Swart, and Daniel A Schult. *Exploring network structure, dynamics, and function using NetworkX*. Tech. rep. Los Alamos National Laboratory (LANL), 2007.

[96] Paul JA Michiels et al. "Solution structure of the pseudoknot of SRV-1 RNA, involved in ribosomal frameshifting". In: *Journal of molecular biology* 310.5 (2001), pp. 1109–1123.

[97] Michael H Kolk et al. "NMR structure of a classical pseudoknot: interplay of single-and double-stranded RNA". In: *Science* 280.5362 (1998), pp. 434–438.

[98] Adam J Collier et al. "A conserved RNA structure within the HCV IRES eIF3-binding site". In: *Nature structural biology* 9.5 (2002), pp. 375–380.

[99] Peter J Lukavsky et al. "Structure of HCV IRES domain II determined by NMR". In: *Nature Structural & Molecular Biology* 10.12 (2003), pp. 1033–1038.

[100] Ling X Shen and Ignacio Tinoco Jr. "The structure of an RNA pseudoknot that causes efficient frameshifting in mouse mammary tumor virus". In: *Journal of molecular biology* 247.5 (1995), pp. 963–978.

[101] Peter V Cornish, Suzanne N Stammler, and David P Giedroc. "The global structures of a wild-type and poorly functional plant luteoviral mRNA pseudoknot are essentially identical". In: *RNA* 12.11 (2006), pp. 1959–1969.

[102] Sarah B Miller et al. "A structure-based mechanism for tRNA and retroviral RNA remodelling during primer annealing". In: *Nature* 515.7528 (2014), pp. 591–595.

[103] Hilda H Au et al. "Global shape mimicry of tRNA within a viral internal ribosome entry site mediates translational reading frame selection". In: *Proceedings of the National Academy of Sciences* 112.47 (2015), E6446–E6455.

[104] Fang Ding et al. "Structure and assembly of the essential RNA ring component of a viral DNA packaging motor". In: *Proceedings of the National Academy of Sciences* 108.18 (2011), pp. 7357–7362.

[105] Katherine E Berry et al. "Crystal structure of the HCV IRES central domain reveals strategy for start-codon positioning". In: *Structure* 19.10 (2011), pp. 1456–1466.

[106] Yaser Hashem et al. "Hepatitis-C-virus-like internal ribosome entry sites displace eIF3 to gain access to the 40S subunit". In: *Nature* 503.7477 (2013), pp. 539–543.

[107] Regan M LeBlanc et al. "Structural insights of the conserved "priming loop" of hepatitis B virus pre-genomic RNA". In: *Journal of Biomolecular Structure and Dynamics* 40.20 (2022), pp. 9761–9773.

[108] Timothy C Hall. "Transfer RNA-like structures in viral genomes". In: *International review of cytology* 60 (1979), pp. 1–26.

[109] Encarnación Martınez-Salas. "Internal ribosome entry site biology and its use in expression vectors". In: *Current opinion in biotechnology* 10.5 (1999), pp. 458–464.

[110] Anton A Komar and Maria Hatzoglou. "Internal ribosome entry sites in cellular mRNAs: mystery of their existence". In: *Journal of Biological Chemistry* 280.25 (2005), pp. 23425–23428.

[111] Justine Mailliot and Franck Martin. “Viral internal ribosomal entry sites: four classes for one goal”. In: *Wiley Interdisciplinary Reviews: RNA* 9.2 (2018), e1458.

[112] Anton A Komar and Maria Hatzoglou. “Exploring internal ribosome entry sites as therapeutic targets”. In: *Frontiers in oncology* 5 (2015), p. 233.

[113] Jerry Pelletier and Nahum Sonenberg. “Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA”. In: *Nature* 334.6180 (1988), pp. 320–325.

[114] Edwin A Brown, Allan J Zajac, and Stanley M Lemon. “In vitro characterization of an internal ribosomal entry site (IRES) present within the 5’nontranslated region of hepatitis A virus RNA: comparison with the IRES of encephalomyocarditis virus”. In: *Journal of virology* 68.2 (1994), pp. 1066–1074.

[115] Lucy P Beales, David J Rowlands, and Andreas Holzenburg. “The internal ribosome entry site (IRES) of hepatitis C virus visualized by electron microscopy”. In: *Rna* 7.5 (2001), pp. 661–670.

[116] Olga Fernández-Miragall, Sonia López de Quinto, and Encarnación Martínez-Salas. “Relevance of RNA structure for the activity of picornavirus IRES elements”. In: *Virus research* 139.2 (2009), pp. 172–182.

[117] Victoria G Kolupaeva, Tatyana V Pestova, and Christopher Ut Hellen. “Ribosomal binding to the internal ribosomal entry site of classical swine fever virus”. In: *Rna* 6.12 (2000), pp. 1791–1807.

[118] Justin Lessler et al. “Assessing the global threat from Zika virus”. In: *Science* 353.6300 (2016), aaf8160.

[119] Maria G Guzman et al. “Dengue: a continuing global threat”. In: *Nature reviews microbiology* 8.Suppl 12 (2010), S7–S16.

[120] James Knox et al. “Murray Valley encephalitis: a review of clinical features, diagnosis and treatment”. In: *Medical Journal of Australia* 196.5 (2012), pp. 322–326.

[121] X De Lamballerie et al. “Genome sequence analysis of Tamana bat virus and its relationship with the genus Flavivirus”. In: *Journal of General Virology* 83.10 (2002), pp. 2443–2454.

[122] Charles H Calisher et al. “Bats: important reservoir hosts of emerging viruses”. In: *Clinical microbiology reviews* 19.3 (2006), pp. 531–545.

[123] Syed Soheb Fatmi, Rija Zehra, and David O Carpenter. “Powassan virus—a new reemerging tick-borne disease”. In: *Frontiers in public health* 5 (2017), p. 342.

[124] Gorben P Pijlman et al. “A highly structured, nuclease-resistant, noncoding RNA produced by flaviviruses is required for pathogenicity”. In: *Cell host & microbe* 4.6 (2008), pp. 579–591.

[125] Ruey-Yi Chang et al. “Japanese encephalitis virus non-coding RNA inhibits activation of interferon by blocking nuclear translocation of interferon regulatory factor 3”. In: *Veterinary microbiology* 166.1-2 (2013), pp. 11–21.

[126] Jeffrey K Noel et al. “SMOG 2: a versatile software package for generating structure-based models”. In: *PLoS computational biology* 12.3 (2016), e1004794.

[127] Jeffrey K Noel, Paul C Whitford, and José N Onuchic. “The shadow map: a general contact definition for capturing the dynamics of biomolecular folding and function”. In: *The Journal of Physical Chemistry B* 116.29 (2012), pp. 8692–8702.

[128] Paul C Whitford et al. “Nonlocal helix formation is key to understanding S-adenosylmethionine-1 riboswitch function”. In: *Biophysical journal* 96.2 (2009), pp. L7–L9.

[129] Maïté Courel et al. “GC content shapes mRNA storage and decay in human cells”. In: *elife* 8 (2019), e49708.

[130] Steve Plimpton. “Fast parallel algorithms for short-range molecular dynamics”. In: *Journal of computational physics* 117.1 (1995), pp. 1–19.

[131] Giacomo Fiorin, Michael L Klein, and Jérôme Hénin. “Using collective variables to drive molecular dynamics simulations”. In: *Molecular Physics* 111.22-23 (2013), pp. 3345–3362.

[132] Nicholas R Markham and Michael Zuker. “DINAMelt web server for nucleic acid melting prediction”. In: *Nucleic acids research* 33.suppl_2 (2005), W577–W581.

[133] Langzhou Song et al. “Structure of staphylococcal α-hemolysin, a heptameric transmembrane pore”. In: *Science* 274.5294 (1996), pp. 1859–1865.

[134] Christian Heinz, Harald Engelhardt, and Michael Niederweis. “The core of the tetrameric mycobacterial porin MspA is an extremely stable β-sheet domain”. In: *Journal of Biological Chemistry* 278.10 (2003), pp. 8678–8685.

[135] Jan Lipfert et al. “Double-stranded RNA under force and torque: similarities to and striking differences from double-stranded DNA”. In: *Proceedings of the National Academy of Sciences* 111.43 (2014), pp. 15408–15413.

[136] Chiu Tai Andrew Wong and M Muthukumar. “Polymer capture by electro-osmotic flow of oppositely charged nanopores”. In: *The Journal of chemical physics* 126.16 (2007).

[137] Minglun Li and Murugappan Muthukumar. “Electro-osmotic flow in nanoconfinement: Solid-state and protein nanopores”. In: *The Journal of Chemical Physics* 160.8 (2024).

[138] Ceming Wang et al. “Slowing down DNA translocation through solid-state nanopores by edge-field leakage”. In: *Nature communications* 12.1 (2021), p. 140.

[139] Maurice G Kendall. “A new measure of rank correlation”. In: *Biometrika* 30.1-2 (1938), pp. 81–93.

[140] William H Press. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[141] Gottfried E Noether et al. “Why kendall tau”. In: *Teaching Statistics* 3.2 (1981), pp. 41–43.

[142] Evan Evans. “Probing the relation between force—lifetime—and chemistry in single molecular bonds”. In: *Annual review of biophysics and biomolecular structure* 30.1 (2001), pp. 105–128.

[143] Olga K Dudko, Gerhard Hummer, and Attila Szabo. “Theory, analysis, and interpretation of single-molecule force spectroscopy experiments”. In: *Proceedings of the National Academy of Sciences* 105.41 (2008), pp. 15755–15760.

[144] Margarita I Zarudnaya et al. “Structural transitions in poly (A), poly (C), poly (U), and poly (G) and their possible biological roles”. In: *Journal of Biomolecular Structure and Dynamics* 37.11 (2019), pp. 2837–2866.

[145] Payam Rowghanian and Alexander Y Grosberg. “Propagation of tension along a polymer chain”. In: *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 86.1 (2012), p. 011803.

[146] Cristian Micheletti. “Comparing proteins by their internal dynamics: exploring structure–function relationships beyond static structural alignments”. In: *Physics of life reviews* 10.1 (2013), pp. 1–26.