universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

RNA structuredness of viral genomes

verfasst von / submitted by

Teodora Bucaciuc-Mracica

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2022 / Vienna 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 875
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Bioinformatik
Betreut von / Supervisor:	Univ.-Prof. Dipl.-Phys. Dr. Ivo Hofacker

Acknowledgements

I would like to express my gratitude to Prof. Dr. Ivo Hofacker for giving me the opportunity to conduct my Master's Thesis project in his research group at the Institute for Theoretical Chemistry. The discussions and guidance during the time spent in his group helped me achieve my goals and improved my work, shaping my bioinformatics related skills and challenging me to think outside the box. His input regarding matters outside of my research project meant a lot to me and I am very grateful for the positive impact he had on me, not only as a student, but also as a young adult aiming to find my place in the world. Thank you for taking the time to give me feedback on my manuscript.

I am also very grateful for the opportunity to work in close collaboration with Dr. Michael Wolfinger. His advice and research input along the way helped me become a better researcher and pushed me out of my comfort zone, which contributed a great deal to my learning process and gaining skills which will be very useful further on in my career. Throughout my master's project, he contributed a great deal to building up my confidence and our numerous discussions added a lot of value to my project and also meant a lot to me on a personal level. Thank you for taking the time to give me feedback on my manuscript.

Many thanks to the whole research group at the Institute of Theoretical Chemistry for their willingness to share their work experiences with me and giving me advice whenever I needed it, regardless of topic.

Last but not least, a great deal of gratitude to my group of friends and family for sharing laughs, tears and meaningful discussions during my Masters degree here in Vienna. The pandemic surely didn't make things easier and without you, my Vienna people (A x 2, C, D, P), it would have been way harder to deal with everything. The visits from my family (Francesca, mersi pentru surpriza de ziua mea!) meant a lot to me and helped me recharge my batteries.

R & A, you know that I have been mentioning you since forever in everything important that I achieve and this here is no exception. Pretty cool that I can still do that after all this time, you are great people :)

Contents

1 Summary

The goal of this work is to assess the structuredness of all complete RNA viral genomes available in the NCBI database. To this end, a method was developed to determine the Z-score of all nucleotides in a given genome, as a way to express the overall thermodynamic ability of the viral genome to form stable secondary structures. A sliding window approach was employed that computes a local measure of structuredness for each position by averaging over all enclosing windows. In order to make the measure of structuredness independent of sequence composition, a Z-score was used, comparing the folding energy of a sequence window to randomized sequences of the same composition. Folding energies were determined using the RNAfold program from the ViennaRNA package. Apart from the genome's overall assessment of structuredness using the Z-score, a more in depth analysis was carried out to compare the distribution of Z-scores between the coding regions and non-coding regions in the genome. Two other different methods were also used to assess structuredness, namely the free opening energies and locally stable structure formation in the whole genome. The usage of different methods to investigate the ability of a genome to form secondary structures were employed as a way to understand these contribute to the stability of the genome, thus protecting it from degradation. A Python automated approach was developed to determine the mean Z-scores per organism for a given taxon, by producing analysis files for each virus species and sending each script on a computer cluster, as well as for pre and post processing the output data, ranging from visualization of the annotated genomes together with their mean Z-scores, to integrating modules from the ViennaRNA package in the form of bash scripts and statistical evaluation of the data. A total of 4142 RNA genomes were individually analyzed, annotated, processed and visualized. The modules RNALfold and RNAplfold from the ViennaRNA package were used to better estimate the stability of local secondary structures and the value of the free energy needed to unfold those structures, as a way to determine the quality of expression of a specific protein or non-coding RNA.This thesis provides a global picture of RNA structuredness across different virus groups, as well as relevant knowledge about each species' RNA structure, which has paramount functional implications not only for the viral life-cycle, but also potentially for medical research and biotechnology. We find that structuredness varies significantly between different groups of viruses, as well as between different genomic regions. In particular the single stranded positive sense RNA viruses form more stable structures than expected by chance, with their non-coding region being more structured than the coding region. In contrast, the single stranded negative sense RNA viruses appear to be more structured in the coding regions, with no significant differences between the coding and non-coding region on both forward and backward strand. Double stranded RNA viruses have on average less structured genomes than the other two groups, regardless of genomic region.

2 Kurzfassung

Das Ziel dieser Masterarbeit ist es, die Strukturiertheit von allen vollständigen RNS-Virengenomen aus der NCBI Datenbank zu bestimmen. Dazu wurde eine Methode entwickelt, den Z-Score für alle Nukleotide eines Genoms zu berechnen, mithilfe dessen die gesamten thermodynamischen Fähigkeiten des Genoms, stabile sekundäre Strukturen zu bilden, ausgedrückt werden können. Das RNAfold-Modul des ViennaRNA-Programms wurde in Verbindung mit einer Sliding-Window-Methode verwendet, um ein ganzes Genom zu scannen; in jedem Sliding Window wurde die Minimum Free Energy (MFE) berechnet sowie ein Z-Score bestimmt. Der Z-Score wird kalkuliert, indem die native MFE mit dem Mittelwert der MFEs mehrerer zufälliger Sequenzen verglichen wird, die dieselbe Länge und Nukleotidzusammensetzung haben wie die native MFE; die Signifikanz einer errechneten MFE wird in Form des Z-Scores als die Anzahl der Standardabweichungen vom Mittelwert dargestellt. Folglich weist ein negativer Z-Score auf eine im Vergleich zu zufälligen RNS-Sequenzen stabilere native RNS-Sequenz hin. Für jede Position im Genom wurde ein durchschnittlicher Z-Score berechnet, indem die MFE-Werte aus all jenen Sliding Windows gemittelt wurden, die ebendiese Position enthalten. Ein automatisiertes Python-Programm wurde entwickelt, um den durchschnittlichen Z-Score in jedem Organismus aus einem Taxon zu bestimmen. Für jede Virusspezies wurden automatisch eigene SLURM-Dateien erstellt und an das Computer-Cluster des Instituts für Theoretische Chemie gesendet. Dadurch werden neben der Z-Score-Berechnung für jedes Nukleotid die Output-Daten gleichzeitig vor- und nachbearbeitet; dazu zählen die Visualisierung der Genom-Annotationen zusammen mit ihren durchschnittlichen Z-Scores oder die Integration anderer ViennaRNA-Module in Form von Bash-Skripten und statistischen Datenanalysen.
Insgesamt wurden 4.142 RNS-Genome einzeln analysiert, annotiert, verarbeitet und visualisiert. Zusätzlich zur Strukturanalyse des Genoms anhand der Z-Scores wurde ebenfalls eine ausführliche Vergleichsanalyse zur Verteilung der Z-Scores in den kodierenden und nicht-kodierenden Teilen der RNS durchgeführt. Die RNALfold- und RNAplfold-Module aus dem ViennaRNA-Programmpaket wurden verwendet, um die Stabilität lokaler Sekundärstrukturen zu analysieren sowie den Wert der benötigten freien Energie zu berechnen, um diese Sequenzen zu entfalten; so soll die Ausdrucksqualität eines spezifischen Proteins oder einer nicht-kodierenden RNS bestimmt werden.
Diese Masterarbeit zeigt die globale RNS-Strukturiertheit in verschiedenen RNS-Virengruppen und erfasst wichtige Erkenntnisse zu den Strukturen der jeweiligen RNS-Spezies, welche nicht nur für den Lebenszyklus eines Virus sondern potentiell auch für die medizinische sowie biotechnologische Forschung entscheidende Implikationen aufzeigen.

3 Introduction

3.1 Functional, structural and thermodynamic properties of RNA

3.1.1 The RNA World

RNA is a nucleic acid present in living cells and has structural and functional differences as opposed to DNA. The RNA is usually single stranded, has Uracil (U) as nucleotide, contains a five-carbon sugar called ribose in its backbone and its function is to encode for proteins. On the other hand, the DNA is found in a double-helix conformation, contains Thymine (T) as nucleotide, the sugar found in its backbone is desoxyribose and contains RNA encoding information. It has been hypothesized that at some point in the evolution of life, the genetic continuity was guaranteed by RNA replication, that the Watson-Crick base-paring was the core of the replication and that genetically encoded protein were not involved as catalysts [97]. These assumptions taken together form the 'The RNA World hypothesis', which has been proposed more than forty years ago and has gained a lot of attention from the scientific community. When the catalytic RNA molecules (ribozymes) were discovered in 1982 and were later shown to be able to catalyze diverse chemical reaction, it causes a big interest in the hypothesis than RNA came before the DNA/RNA/Protein world [33]. One known argument for RNA being the first molecule that provided life on Earth is the "The Molecular Biologist' Dream" theory, which states that all components of RNA were made available in a prebiotic pool and they could assemble into replication, evolving polynucleotides without the prior existence of any macromolecules [97]. Still, the RNA World problem is not yet solved, but rather paved the way into an extensive research of this very versatile and complex molecule.

3.1.2 RNA secondary structure prediction

Knowing the structure of the RNA is paramount for gaining information on gene expression, regulation of mRNA and overall RNA function. The primary RNA structure is just the plain sequence of nucleotides in a RNA strand. The RNA can contain both canonical (Watson-Creek base pairs) and non-canonical base pairs, such as a guanine-uracil pair, contributing to the high diversity of RNA's structure. The secondary structures, formed when two parts of the sequence are complementary to each other, are held stable for example by hydrogen bonds, stacking interactions and van der Waals interaction between the nucleotides. There are multiple types of secondary structures that can be achieved, such as stem-loops, hairpins and multiloops, to name a few. An example of the diversity of secondary structures can be seen in Figure1.
The secondary structure folding from a primary structure has been widely investigated and a few proposed algorithms have been proven to accurately match computation predictions with the experimental data. On the search for an energy model that is computationally easy to handle, the Nearest Neighbor Energy Model was proposed [76]. The model computes the free energy of a structure as a sum of the loop energies. The loop decomposition is achieved by attributing for each type of loop a energy value, based on stacking interactions and hydrogen bonds. The energy parameters were experimentally measured for each type of loop. The core of

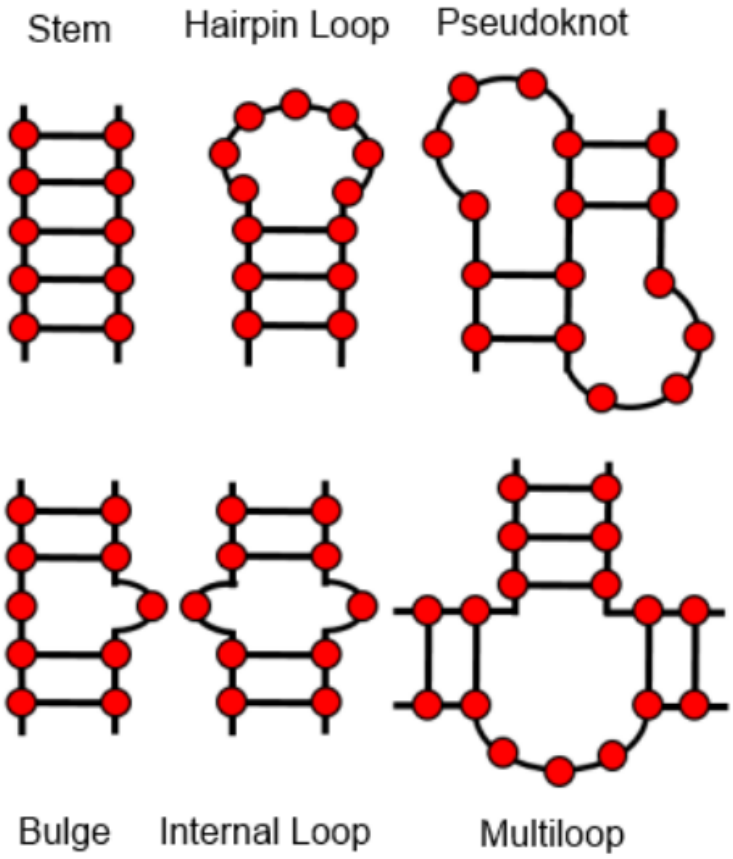

Figure 1: Secondary structures in RNA. Modified after *Cheeda el al.* [22]

the algorithm is the assumption that the unfolded state of a sequence has a higher entropy in contrast to the folded state, due to the fact that the formed base-pairs impose constrains on the flexibility of the sequence. Thus, the secondary structure is comprised of many microstates forming a macrostate, and that macrostate has a free energy that can be calculated. One of the secondary structure prediction algorithms is based on the minimum free energy (MFE). The MFE structure of an RNA sequence is that secondary structure which contributes a minimum of free energy in the energetic ensemble of possible structures. The MFE algorithm uses the Turner energy parameters and dynamic programming recursions to compute the thermodynamically most stable secondary structure [124].
There are a few assumptions that need to be taken into account when computing the MFE. A valid secondary structure requires that each nucleotide interacts with at most one other nucleotide in order to form a base pair. The only base pairs considered are the Watson Crick pairs as well as G-U and U-G. Pseudoknots or crossing-pairs are not allowed, for a more optimal computational effort [72]. Since the number of possible secondary structures grows exponentially with the sequence length [119], dynamic programming algorithms have been employed to avoid listing all possible structures. Dynamic programming is an informatics approach used to solve a variety of combinatorial problems by iteratively breaking them down into smaller problems [13]. Thus, MFE algorithm combines dynamic programming and energy parameters to compute the thermodynamic most probable and stable secondary structure of a RNA sequence. The prediction accuracy for RNAs up to 500 nucleotides can reach 70%, whereas for longer RNA sequences it can fall down to

40% [77, 31]. The reasons for the limited accuracy may be attributed to faulty energy parameters, deviations from standard conditions (such as ion concentrations or temperature), exclusion of RNA interactions with other molecules and the omission of pseudoknots and non-canonical base pairs. It should also be noted that one important limitation of the MFE is that the predicted secondary structure may not be the biological true one [44].
Generating the landscape of all possible RNA secondary structures in thermodynamic equilibrium is a computationally more challenging way to determine if indeed the predicted MFE is accurate. The ensembles of structures are commonly represented as base pair probabilities in a dot-plot. The dot-plot is a n x n matrix, which each square indicating a base pair (Figure 2). Each square indicates a base pair and its area is proportional to the probability of that base pair in the equilibrium ensemble. The upper right half of the plot shows the base pairs of the alternative structures, while the lower left part shows the base pairs of the MFE.

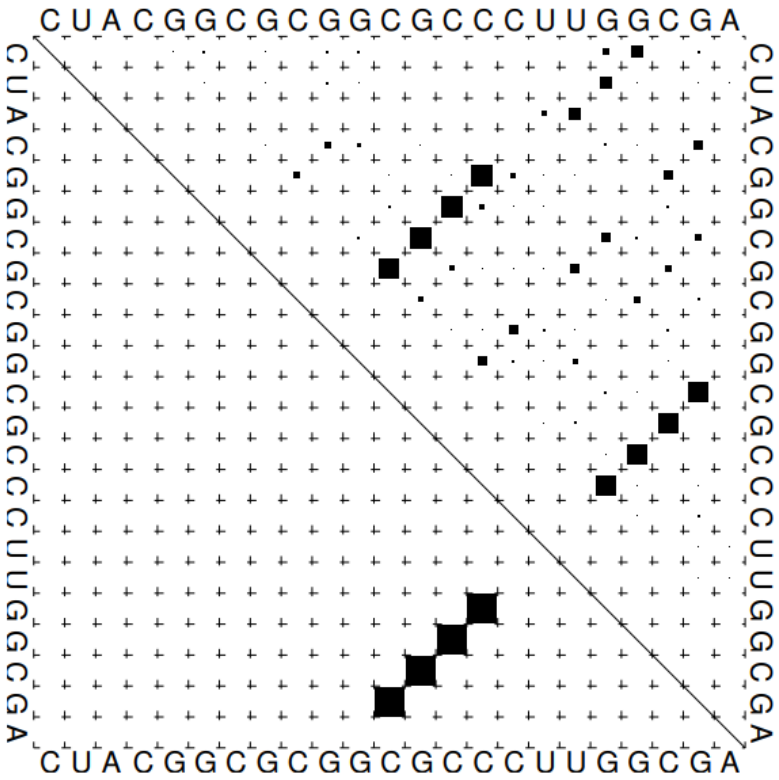

Figure 2: Example of base pair probability dot-plot for visualizing structural alternatives of a RNA sequence

The probability of the base pairs are calculated from the partition function, which describes the statistical properties of a system in thermodynamic equilibrium by summing over all possible states (in this case, secondary structures (s) in that system:

$$Z = \sum_{s} e^{\frac{-E(s)}{RT}}$$

where E(s) is the free energy of the structure, T is the thermodynamic temperature of the system and R is the gas constant. The probability of a secondary structure in equilibrium follows the Boltzmann distribution

$$p(s) \sim e^{\frac{-E(s)}{RT}}$$

By using the partition function for normalizing the equilibrium probability of a secondary structure s, the probability of that structure in the Boltzmann ensemble of all structures in thermodynamic equilibrium becomes

$$p(s) = \frac{1}{Z} \cdot e^{\frac{-E(s)}{RT}}$$

As such, the base pair probabilities and statistical sampling of RNA secondary structures according to their probabilities in equilibrium resulted in the implementation of an algorithm (known as McCaskill's algorithm), that relies on unique decomposing of the ensemble of secondary structures, so that no structure is counted twice, avoiding the exhaustive computation of summing over all possible structures [79]. The MFE secondary prediction method is the core algorithm implemented in the ViennaRNA package within the RNAfold program [71]. It accounts for penalties of different types of loop formation and can also use McCaskill's algorithm to generate the partition function and dot-plot of base pair probabilities. Additionally, the MFE algorithm is highly efficient with a complexity of O(N^3), where N is equal to the sequence length.

3.1.3 Computation of the Z-score

Included in the RNAfold module is also the computation of the Z-score for a given RNA sequence. The Z-score is a method to compare the MFE of the given RNA sequence with the MFEs of random sequences, which have the same length as nucleotide composition as the given sequence. This robust method is a way to determine how stable the secondary structure of the native RNA is compared to random RNAs. The comparison between two MFEs cannot occur in absolute terms, because a single nucleotide can already change the entire conformation of the nucleotide sequence; the concept behind the Z-score is to have a robust and reliable method to assess the structuredness of a RNA sequence, by providing it with a Z-score which is expressed in standard deviations from the mean MFE of the random sequences. Thus, the more negative the Z-score, the more stable the sequence is than expected by change.

3.1.4 Global and local reliability of secondary structure prediction

The reliability of a prediction can be assessed by the base pair probabilities. The global prediction is measured by the Boltzmann probability of the MFE, which is negatively affected by the length of the structure: the longer the structure, the smaller the probability because the prediction accuracy gets lower the longer the nucleotide sequence, as mentioned above. In the ensemble of probable secondary structures, a structure with a high free energy has a smaller probability than a structure with low free energy. In order to determine if the ensemble is comprised of one well defined structure and small variants of it or of a larger number of structures which are very similar to each other, one can measure the dissimilarity of structural alternatives. This well-definedness measure is computed from the pair probabilities

$$< d >= \sum_{i,j} p_{ij} \cdot (1 - p_{ij})$$

where p_{ij}, the second term of the equation, is the probability that a base pair is present in the first structure, while the last term of the equation denotes the probability that the second structure does not include that base pair. This measure computes the expected distance between any two RNA secondary structures in the ensemble.
The local reliability is a measure to define which parts of the prediction are reliable. This has lead to the development of alternative algorithms based on the MFE and partition function, in order to reduce the number of base pairs in a RNA sequence to a certain number. The RNALfold program in the ViennaRNA package is a suitable tool for calculating the local stable structures in RNA sequences, by performing computations for all sequence-windows of a given size in O(N$\cdot$ L^2) time, where L is the number of base pairs allowed along the RNA sequence and N is the length of the sequence [71]. Additionally, the RNAplfold program from the ViennaRNA package can be used to compute the base pairing probabilities averaged over all sequence windows in which that base pair is present. One of the important properties that can be measured using RNAplfold is the accessibility of a RNA region. This means determining how accessible a binding motif in the RNA sequence is, so that it can undergo RNA-RNA interactions. The accessibility is computed as the probability that a given region in single stranded or unpaired sequence, which is equivalent to determining the free energy (opening energy) which is needed to get a RNA region to become unfolded. Since many regulatory functions are possible via RNA interactions between two RNA molecules when they are single-stranded, predicting the opening energies from the accessibility of the RNA sequence plays an important role in determining possible RNA sites that act as binding motifs. Like RNALfold, the RNAplfold also runs in O(N$\cdot$ L^2) time, where L is the number of base pairs allowed along the RNA sequence and N is the length of the sequence.

3.1.5 Methods to determine structuredness in RNA genomes

There is no single or right way in which structuredness can be assessed. In this thesis, multiple ways for structuredness measurement have been used. While the focus lies on the computation of the Z-score for determining structuredness throughout the genome, two other additional methods have been integrated in the analysis pipelines. One of these methods is the calculation of the free opening energies values, by making use of the RNAplfold module included in the ViennaRNA package. The opening energies express the amount of free energy that is needed in order to unfold a given nucleotide sequence. High energy values computed for a nucleotide sequence can be an indicator that the region is particularly structured and therefore more amount of free energy is required to break the structure. This process can be in favor of facilitating an appropriate binding of the ribosome on the RNA sequence, because the interaction between the ribosome and the genome can only occur when there is less structuredness present. The opening energies are calculated by taking the logarithm of the probability of a specific window to be unpaired. In this thesis, the probability of each basepair to be unpaired has been used in order to determine the opening energy value in the whole genome. A second analysis was conducted in order to determine the amount of free energy in the at the start of the coding region as well as inside of the coding region, using a sliding window of length 30.
The second method to assess structuredness is by determining the abundance of locally stable secondary structured predicted for a genome. This approach was

integrated in the analysis pipelines by making use of the RNALfold module from the ViennaRNA package, which computed the locally stable secondary structures in a RNA sequence. In this way, it is possible to determine how structured the RNA is, regardless of genomic region.
Since the goal of the thesis was to determine the global structuredness pattern in all RNA viruses, all the methods were integrated in specific Python pipelines and more so, the visualizations of the data can be automatically generated in order to provide a more insightful understanding of the location of the secondary structures or even to understand correlations between Z-scores, opening energies and genome regions, based on the annotation information found in the Genbank file of each virus.

3.2 RNA viruses

Viruses are obligate intracellular parasites, whose genome is either DNA or RNA, surrounded by virus-coded protein coat, known as capsid. Since they cannot grow or divide without infecting a host cell, nor can they generate energy, they lie in the gray area of biology between living cells and plasmids [23]. The proteins associated with the genetic material (nucleoprotein), together with the genomes, form the nucleocapsid. The totality of components needed for viral infection is known as the virion, or complete virus [1, 18]. The purpose of a virus is to deliver its genome into the host cell, in order to allow transcription and translation, as they are unable to survive outside the host. Based on their genetic material, they can be categorized as either DNA or RNA viruses. Viruses exhibit a huge diversity of genome structures, such as positive or negative sense viruses, linear or circular, single or double-stranded, segmented or unsegmented. Due to the relatively small genome size and high mutation rate, RNA viruses are said to be more abundant than DNA viruses [92]. The higher mutation rate can be partially explained by the RNA-dependent RNA polymerase (RdRp), which replicates their genomes and is virally encoded. Apart from the *Nidovirales* family, RdRp does not have proofreading ability and cannot correct mistakes during replication, in contrast to the majority of DNA polymerases [38]. The small genome size (on average 10.000 nucleotides) was hypothesized to be linked to the maximum genomic material size that can be contained in the virion, but newest research has shown that actually the high mutation rate has a bigger impact on the size of the genome [48].
The abundance of both RNA and DNA viruses, together with their highly versatile genome structures, has determined the research community to create the The International Committee on Taxonomy of Viruses (ICTV). The ICTV is responsible for the correct classification and taxonomic architecture of viruses. One of the tasks of the committee is to develop internationally agreed upon names for virus taxa, including species. Thus, throughout this thesis, the nomenclature of the viruses and virus species follows the ICTV regulations.
It is possible to classify the RNA viruses based on the type of RNA in the genome, as such it is differentiated between single stranded positive sense viruses (ssRNA(+)), single stranded negative sense viruses (ssRNA(-)) and double stranded RNA viruses (dsRNA). This classification, also known as the Baltimore classification [10], is used throughout the thesis to differentiate between the RNA virus groups.

3.2.1 Single stranded positive sense RNA viruses

For the ssRNA(+), the genome serves as mRNA and is infectious upon entering the host cell. The genome also serves as template for synthesis of additional viral RNAs, with the first event after entering a host cell being the ribosomal assembly for the synthesis of viral proteins [91]. The replication in ssRNA(+) starts with translation of mRNA, to produce essential proteins, such as RdRp. After translation, the genome serves as template for copy RNA, which is then used as template for synthesis of additional genomes mRNAs [3]. In general, a vast majority of the genomes in the ssRNA(+) group fold into stable secondary structures, whose role is vital for translation, transcription and replication. In this thesis, the global structuredness of the genome has been investigated, without emphasis on particular regions. SsRNA(+) viruses include both eukaryotic and prokaryotic pathogens. Virus families such as *Flaviviridae*, *Picornaviridae* or *Coronaviridae* are known for containing health threatening species such as Hepatitis C virus (HCV), Poliovirus and SARS-CoV-2, to name a few. HCV causes acute and chronic hepatitis in humans and mammals. If left untreated, it can progress to cirrhosis and hepatocellular carcinoma [57]. Also known as the "silent epidemic" due to being generally asymptomatic for many years after infection, it became an important public health problem and disease burden all over the globe [110]. The RNA in HCV containes a 5'-UTR (non-conding region), a long open reading frame (ORF) and a 3'-UTR. The 5'-UTR and 3'-UTR contain stable secondary structures, which were shown to play an important role in the replication and translation off this virus [67]. The ORF encoded polyprotein is processed co- and posttranslationally by the host into more than ten different viral proteins, involved in processes such as hepatocarcinogenesis, insulin resistance [66], viral entry or completion of the viral replication cycle [67].
The Dengue virus is a mosquito borne virus belonging to the Flavivirus genus, that spread from the tropical climates to the subtropics and temperate climates in the last 30 years. It is the fastest spreading virus transmitted by mosquitoes in many parts of the world, causing dengue hemorrhagic fever which can ultimately lead to death [7]. Cis-acting elements, which are non-coding regions regulating the gene expression in the genome, contain folded structures, conserved across all mosquito-borne species in the *Flaviviridae* family. Two stem-loops in the 5'-UTR play important roles in mRNA stability, localization and translational efficiency as well as gene expression control [70]. The 3'-UTR contains three distinct domains in which conserved sequences and structures are present. The stable secondary structures in these domains are crucial for the replication process and mediation of long RNA-RNA interactions with the 5'-UTR [42]. It was also reported that conserved structures in the 3'-UTR of flaviviruses appear in multiple copies to better sustain the functional relevance of the elements. One of the conserved structural elements in the 3'-UTR are the xrRNA (exoribonuclease-resistant RNA elements), which stall the exoribonuclease Xrn1. The Xrn1 stalling process results in the production of viral lncRNAs, also known as subgenomic flaviviral RNAs. The xrRNAs protect the RNA from degradation, and thus the stalling of the Xrn1 leads to dysregulation of cellular function, promoting viral infections [87].

3.2.2 SARS-CoV-2 genome in the context of the 2019 pandemic

The global pandemic from 2019 has put the focus on the viruses belonging to the *Coronaviridae* family, especially on different strains of the SARS-CoV-2 virus, which have caused more than 6.4 million deaths so far, according to the World Health Organization (WHO). The rapid spread of the virus starting from December 2019 manifested in pneumonia like respiratory infections and high fever. What sets this virus apart is the presence of the spike protein, which is enhancing the infectivity of SARS-CoV-2 and is not found in any other member of the *Coronaviridae* [95]. The single stranded positive sense genome has 14 ORFs and encodes for 27 proteins, with a total length of approximately 30.000 nucleotides [122]. There are multiple conserved secondary structures in the 5'-UTR of coronaviruses, which have central roles in viral replication and subgenomic RNA formation [63].
Especially the frameshifting stimulation element, which is a pseudoknot structure found in the coding region of the genome, was shown to adopt a highly stable structure in multiple studies [93, 98, 30]. This element plays a key role in the gene expression pattern in all *Coronaviridae*. The ORF1a gene, located at the 5' end, encodes for the proteins required for securing the ribosomes, dysregulation the cellular innate immune response of the host and cleaving polyproteins into different proteins [55]. The ORF1b gene encodes for the RdRp and different proteins that are involved in the RNA syntesis of the negative strand, which is used as template for building new positive strands, facilitating the viral replication cycle [5]. The partial overlap of ORF1a and ORF1b causes the ORF1b to be the -1 reading frame relative to ORF1a. The frameshifting element ensures the ribosomal translation of the ORF1b proteins, which would otherwise not be translated [107]. The pseudoknot has been extensively researched as a potential candidate for drug target in an effort to develop therapies for stopping the pandemic [21].
The rapid spread of the SARS-CoV-2 virus in different geographical areas, combined with the high mutation rate in RNA viruses, caused the accumulation of different mutations in the spike protein as well as throughout the genome. These new strains (CoVariants) had different virulence patterns and posed a public health threat at least as high as SARS-CoV-2. In an effort to provide the scientific community with access to sequencing data from different patients across the world infected with different SARS-CoV-2 strains, the GisAid database collected genome sequences from research facilities and made them freely available for research purposes [56]. Thus, millions of genomes could be retrieved and analysed in order to advance the scientific knowledge of the genetic structure(s) in the CoVariants. In line with the recent availability of sequencing data from multiple SARS-CoV-2 genomes across the world, an in depth analysis into the folding pattern of multiple SARS-CoV-2 CoVariants has been pursued in this thesis. Using the background information found at https://covariants.org/ in May 2022, six different CoVariants were taken into account: 20I (Alpha), 20H (Beta), 20J (Gamma), 21A (Delta), 21K (Omicron), 21L (Omicron). For each CoVariant, twenty fully sequenced genomes from different continents were retrieved and characterized from a thermodynamic point of view, with respect to the SARS-CoV-2 reference genome.

3.2.3 Single stranded negative sense RNA viruses

In contrast to the ssRNA(+) viruses, the ssRNA(-)'s genomes are composed of negative sense RNA (genome or 3'-5' strand). Thus, the negative RNA is not infectious upon entering the host cells and needs to be transcribed in order to start the infection. All the ssRNA(-) viruses carry within their genome and viral particles all the components needed for transcription [96]. After entering the host cell, the first event of ssRNA(-) is to build the antigenome strand (5'-3' strand) from the genome, so that to use both strands as templates for manufacturing more viral genomes: the 5'-3' strand acts as mRNA, while the 3'-5' strand is used to synthesize more copies of the 5'-3' strand [23]. Both vertebrates and invertebrates serve as hosts for the ssRNA(-) viruses.

Among the most well-known human pathogens are the viruses in the Ebolavirus genus, members of the *Filoviridae* taxonomic family. They are unsegmented viruses, approximately 19.000 nucleotides long and cause highly lethal hemorrhagic fever, severe bleeding, lived damage and hypotensive shock [108]. The highest lethality rate is attributed to the Zaire ebolavirus (<90%), which occurred for the first time in the Democratic Republic of Congo in 1976, former known as Zaire [58]. It soon spread in different parts of Africa, including Gabon, Liberia and Mali, according to the Centers for Disease Control and Prevention. According to the WHO, two vaccines against the Zaire ebolavirus species are currently licensed, which appeared just recently in 2019 and 2020. The linear genome in all ebolaviruses encodes for a total of seven proteins, in the order 3'-NP-VP35-VP40-GP-VP30-VP24-L-5' [50]. The RNA genome is encapsulated by the nucleoprotein (NP) and further forms a ribonucleoprotein complex with the RdRp, similar to other ssRNA(-) viruses [11]. Each gene contains its respective ORF, which is flanked by long nontranslated regions, a feature only known to occur in filoviruses and some henipaviruses [83]. The role of these long UTRs may be for mRNA stabilization and translation, as it was shown that the 5'-UTR of the L gene (which encodes for the RdRp) controls the translation by diminishing ribosome initiation at the L start codon [104]. The highly conserved transcription start and stop signals regulate the mRNA transcription: the polymerase binds to the genome at the 3' end and scans for the transcription start signal of the first gene, further proceeding along the RNA by stopping and re-initiating transcription at each gene signal [100, 116]. Potential stem-loop secondary structures have been predicted at the 5'-UTRs on both the genomic and antigenomic RNA [84]. Also, probing experiments have revealed hairpin formation at the 3'-end on both strands [120]. It was also recently shown that the length and the stability of hairpin structures in the 3'-end influences the replication and transcription in ebolaviruses [9].

The diversity of ssRNA(-) has only been acknowledged in the last decades and thanks to RNA sequencing, many arthropod, insects and centipedes negative RNA viruses have been discovered. The majority of these viruses seem to be ancestors of disease-causing viruses such as influenza and ebolaviruses [65]. Numerous plant viral pathogens may have been overlooked because of the extreme divergence from known viruses and their characterization using high-throughput sequencing (HTS) enabled to identify new virus strains in several plant hosts as well as completion of genome sequences where only genome fragments were known [12].

The *Bunyaviridae* family contains both animal and plant infecting viruses. The species in this family have negative sense, segmented genomes with lengths ranging

between 2.000 and 12.300 nucleotides. Among the most threatening human and cattle pathogens is the Rift Valley Fever virus, which became an important zoonotic threat to the public and veterinary health, due to its impressive capacity to spread across geographic barriers and re-emerge after long periods of silence [90]. One of the most devastating plant viruses worldwide are the viruses in the *Tospoviridae* family, whose host range exceeds 1000 different plant species. Especially problematic is the infection of species of economic interest, such as tomato, potato and pepper, as well as many weeds [61]. The plant viriome landscape, due to advances in the HTS, will uncover the evolutionary history of plant viruses and increase the knowledge about the complexity and phylogenetic relationships in the plant virome landscape [12].

3.2.4 Double stranded RNA viruses

In the dsRNA group, it appears that some families have evolved from different lineages of ssRNA(+) viruses, while other have roots in the dsRNA bacteriophages [60]. DsRNA viruses represent a numerous group of fungi, plants and animal pathogens whose genome can be segmented or unsegmented. The number of segments can vary from one to even twelve, packed together in a virion. The majoritiy of dsRNA viruses have segmented genomes though, providing a mechanism that conveniently breaks up the total translation products from the genome into several distinct proteins [81]. Each segment usually encodes for up to two proteins, amongst which the RdRp is present. Depending on the size of the segments, different mRNAs are generated in different molar amounts. This offers an important differential control over the expression levels of distinct viral genes [81]. However, there are cases in which dsRNA viruses have developed mechanisms to generate multiple proteins from single mRNA, mostly for viruses with one, two or three segments [81]. The dsRNA genomes cannot be translated directly, thus they serve as templates for building the positive strand mRNA molecules. A subviral particle retains the dsRNA in the cytoplasm, as to protect it from the detection of the innate immune system of the host cell [86]. The largest dsRNA family is *Reoviridae*. The viruses in this family have linear segmented genomes, isolated from various organisms, such as mammals, birds, reptiles and fungi, to name a few. Even though the functional core of the RdRp has similar motifs across the entire family, phylogenetic analysis have shown that the distinct genera in this family exhibit less than 30% amino acid identities in the sequence of the RdRp [2]. The segments contain conserved terminal sequences in the 5' and 3' ends, which may have involved as recognition signals to facilitate transcription and replication. Sequences near the termini share extensive complementary, interrupted by short sequences which were predicted to form secondary structures, such as stem-loops [2]. The vast majority of the genome segments contain only one ORF encoding one gene and short non-coding regions.
The Rotavirus genus (RVA) belongs to the subfamily *Sedoreovirinae* in the *Reoviridae* family. The rotavirus' genomes consist of eleven segments and cause acute gatroenteritis in young mammals and birds [109]. In 2013, 215.000 children under the age of five have died globally as a result of infection with RVA and thus the WHO recommended introducing the rotavirus vaccine into the routine immunization programs for children [111]. The RVA genome encodes for six structural and six non-structural proteins and its virion is triple-layered [103]. Based on their nucleotide sequence differences in the gene encoding the inner capsid protein VP6, they are categorized in nine serogroups, designated as A-D and F, G, N [78]. Viruses in

group A,B,C and H infect both humans and animals, while the rest infect predominately animals, but zoonotic transmission of members in group A has been observed to occur from animals to humans in Morocco [28, 4]. Segments 1-10 encode for one protein, whereas segment 11 encodes for two proteins [59]. Many studies have confirmed the that the terminal ends in each segment contain packaging signals required for incorporation of the segments into budding virions during the replication process [114].
One of the newest families in the dsRNA taxonomy is *Amalgaviridae*. The viruses in this family are plant pathogens with small linear genomes, containing two overlapping ORFs [50]. Interestingly, phylogenetic and bioinformatic analysis have demonstrated that, even though their genome architecture is characteristic to the fungi infecting viruses in the *Totiviridae* family, the RdRps in amalgaviruses form a sister clade to the proteins in *Partitiviridae* family, whose genomes are segmented [99]. Thus, it has been proposed that the amalgaviruses constitute an evolutionary link between partitiviruses and totiviruses [99]. These viruses are an outstanding case of transfer of viral hallmark genes between different RNA viruses, underlining the modularity of the RNA virus world [62].

3.2.5 Structuredness in coding and non-coding RNA and statistical comparison methods

Coding RNA refers to RNA that encodes for proteins (mRNA) and was for a very long time the only focus in the RNA research field. In RNA viruses, depending on the genome type, mRNAs represent either directly the genome, as it is the case for ssRNA(+) viruses, or need to be synthesized from the genome, which serves as transcription template, in the case of ssRNA(-) and dsRNA.
When the Human Genome Project revealed in 2005 that the genome contains vast regions of non-coding RNA (ncRNA) [26], it became clear the ncRNAs should not be neglected, as they could be a functionally important part of the genome. As such, the focus has shifted in the last decade from coding RNA to ncRNA, determining its role in many biological processes, such as tumor suppressor and oncogenic driver in some cancer types or key regulator in stress response in plants [124]. The non-CDS (non-coding regions) were shown to play an important role in counteracting the defense of the host organism against the viral infection. Further, it was proved how advantageous it is to synthesize ncRNA rather than proteins, in order to escape the radar of the immune system of the host cell [113]. There are many studies which have shown that the both the CDS and non-CDS regions across different ssRNA(+) families contained structured regions. For example, conserved RNA structures found in the UTRs of the RNA viruses in the *Flaviviridae* family are known to mediate viral life cycle by either promoting or enhancing replication [87]. It was shown that the Hepatitis C virus (HCV) contains structured elements both in the 5' and 3'-UTR, whose role is vital for the translation and replication of the viral RNA [36]. There are multiple studies which demonstrated the importance of the secondary structures in the UTRs of the HCV. Even though there are major differences in the nucleotide sequence of the six HCV genotypes, the sequences in the 5'- and 3'-UTR are well conserved, very likely due to the evolutionary pressure to maintain those RNA elements that are needed for translation initiation of the viral polyprotein, for example [123]. The 3'UTR contains three different domains, each forming multiple stem-loop structures, the most stable one being the terminal SL3 element with a

length of 46 nucleotides [37]. The 3'X-tail, which is an untranslated region of the 3'UTR, plays an important role in the regulation of key processes in the life-cycle, replication and infectivity of the viral genome, and is involved in long range RNA-RNA interactions [32]. There are multiple structural organizations proposed for the 3'X region and is has been hypothesised that it can adopt more than one structural forms, which could be favoured by the district genotypes. This would explain the different drug resistance as well as virulence of the HCV [19]. The 5'-UTR also was reported to contain secondary structures across its four domains; in the first domain there is one single stem-loop, and the rest of the domains constitute an internal ribosomal entry site [49].

Using local windows alongside the genome of different Dengue virus serotypes, a study was able to determine specific regions in the CDS that maintain a strong local folding across different viral variants [39]. The 3'-UTR in the ssRNA(+) possesses higher order structures (such as dumbbell or hairpin loops) that are generally conserved across species, even though their nucleotide sequence of the 3'-UTR differs [69]. Long ncRNAs (lncRNA) are usually 5'-capped, spliced and polyadenylated to form structures similar to mRNA, with a length of more than 200 nucleotides [125]. Their long nucleotide sequence can fold into complex secondary structures with roles in viral replication. Indeed, a few studies have determined that viruses such as SARS-CoV or Hepatitis B virus dysregulate the expression of host lncRNAs, which results in the progression of infection [125]. Abnormalities in the expression of lncRNAs also affect the innate immune response of the host, which are caused by aberrant modulations of lncRNAs during viral infections [112]. LncRNAs can exhibit different patterns of conservation to protein-coding genes, which are under the pressure of functional constrains; lncRNAs exhibit shorter conserved sequences needed to maintain functional domains and structures [80]. Since the importance of the ncRNA has surfaced, bioinformatics tools have been developed in order to identify functional ncRNAs in the nucleotide sequence, based on their ability to form thermodynamic stable secondary structures [118]. These examples confirm that there there are regions in the ssRNA(+) genomes that are under evolutionary pressure to exhibit stable secondary structure elements in order to be able to proper replicate and adapt to the host environment.

Of the ssRNA(-) viruses, perhaps the best known is the *Filoviridae* family, which contains non-segmented viruses such as ebolaviruses and marburgviruses. There is evidence for stemloop structures in the transcription start site of the viruses, whose functionality is still under research [83]. Little is known about the conserved structures and secondary folding in the ssRNA(-) viruses, however one feature in all filoviruses is that their genes are flanked by transcription start and stop signals, which determines the beginning and the end of the transcribed mRNA [51].

Even though very little is known about the folding properties in dsRNA coding and non-coding regions, the important role of non-coding RNA surfaced in the latest research, attributed to functions such as viral replication, persistence, pathogenesis and even regulation of anti-viral response proves that the importance of the structuredness in non-CDS of the dsRNA group should not be underestimated [25]. The ability of the non-coding RNA to form secondary structures is the foundation of executing its function, thus the work presented in this thesis may provide a starting point for further research in the dsRNAs, to get a more detailed picture of the structuredness of the underlying genomes in this group.

Different statistical approaches can be used to assess differences of thermodynamic properties, such as the Z-score, between the coding and non-coding regions in the RNA. Distribution and empirical cumulative distribution function plots are amongst the most widely used visualization methods to observe ranges and extract meaningful summary statistics from the underlying data. Thus, they can be suitable for comparing the distribution of the Z-scores across CDS and non-CDS in the different taxonomic categories. Statistical tests are also frequently used in data science to decide whether the available data provides enough evidence to reject or support a particular hypothesis. This method of statistical inference can be used to determine, for example, if two distributions behave similarly or if the underlying data in the distribution comes from a normal, Gaussian distribution. Hypothesis testing is a common way to make decisions using the accessible data, not only in the field of statistics, but also in medicine and research overall.

3.3 Earlier work

Previous research has demonstrated that the thermodynamic properties of RNA in terms of MFE and Z-score computation have proven to be reliable methods for determining the structure of RNA genomes. A few published research articles have used this method on ssRNA(+) [106] or on SARS-Cov-2 [105, 6]. In [106], the secondary structure prediction was computed using the MFOLD program developed by Zucker *et al.*, which also relies on predicting the MFE value of a given genome using sliding windows and the Z-score is used as a way to compare the native MFE with the mean MFE of random sequences, maintaining the trinucleotide base frequencies. In their work, the sliding-window used to scan the genomes is 498 nucleotides long, which is quite a large number, when considering that the MFE prediction accuracy worsens with the length of the RNA sequence. The data in the study included only genomes of some families belonging to ssRNA(+) and ssRNA(-) and showed that there is an extensive variability between the taxonomic genera in terms of possession of genome scale ordered RNA structures, as well as significant differences in structuredness between closely related taxonomic families. Another publication used a 120 nucleotide sliding-window to determine the structuredness of the SARS-Cov-2 genome, using the RNAfold program from ViennaRNA package [6]. They determined that, in terms of free energy and Z-score, the genome of the SARS-CoV-2 genome is highly structured, more so than the genomes of the Zika virus or human immunodeficiency virus (HIV). In [105], the same approach was used with the tools from the ViennaRNA package to investigate the secondary structure formation in SARS-CoV-2, however the chosen size of the sliding-window of 350 nucleotides may be, as in [106], somewhat larger than the usual range of the MFE prediction accuracy of a RNA sequence. Nevertheless, the research published this far using the Z-score model and the corroboration of the thermodynamic *in silico* predictions with the *in vivo* experiments prove the reliability of the method and thus provide a good support for the research advances in the field of RNA bioinformatics and Virology.

3.4 Motivation

There are many studies proving the existence of stable secondary structures in specific regions of the viral RNAs across different taxonomic families and species. However, a global overview of the overall structuredness across different viral genomes is, at the point of writing this thesis, not available. To overcome this shortcoming, a total of 4142 RNA segmented and unsegmented virus genomes have been analyzed, belonging to ssRNA(+), ssRNA(-) and dsRNA virus groups. The main property analyzed in all viruses was the ability of their genomes to form stable secondary structures, both locally and globally, regardless of genomic region, and to perform comparative analyses between the viral groups. Thus, for each taxonomic group, the mean Z-score was calculated by averaging the obtained Z-scores per all species in that taxon. The differentiation between structuredness (in terms of Z-score) for the 5'-3' and 3'-5' strand was done for the ssRNA(-) and dsRNA viruses for different reasons. In the case of ssRNA(-) viruses, the first event after entering a host cell is to synthesize the corresponding 5'-3' RNA positive strand in order to be able to produce proteins, because their 3'-5' RNA is complementary to the mRNA [23]. Hence, it was important to assess the thermodynamic properties on both strands, as to detect if there are any significant differences in how the secondary structures are formed and how stable they are. In the case of dsRNA viruses, since their genome is comprised of two strands of RNA, the analysis of only one of the strands would be incomplete without having the same information for the other strand. Opening energy values as well as abundance of RNALfold predicted structuress were additional methods by which structuredness was measures throughout the NCBI retrieved genomes.

4 Materials and Methods

4.1 Viral genome data sets

All the viral genomes analyzed in this work were automatically retrieved from the NCBI database [101]. Only RefSeq approved, complete genomes were considered for analysis [88]. Table 11 gives on overview of the total genomic data used here, where the data is categorized based on the Baltimore classification system [10]. The Baltimore classification is used throughout this thesis to differentiate between RNA or DNA virus genomes and compare different properties between these groups. A list of all the viruses and their accession is provided on the GitHub account where all the analysis pipelines are deposited. The viral genome sequences retrieved from the NCBI database provide the nucleotide sequence solely for the 5'-3' strand. Hence, the reverse complement sequence has been computed from the 5'-3' nucleotide sequence to obtain the 3'-5' strand of the genome, for the ssRNA(-) as well as dsRNA viruses. For these organisms, it was possible to have the 5'-3' and 3'-5' strands analysed separately, in order to get a complete overview of their thermodynamic properties as well as a thorough comparison between their structural behaviour. The focus of this work lies on RNA viruses, hence these genomes will be discussed in detail further on. As most of the analysis was done at species level in the taxonomic hierarchy, the meaning of a data set refers here to the genome of one particular virus or virus species, accoring to the ICTV nomenclature. To distinguish between coding and non-coding regions in each data set, the genome annotation data was retrieved from the according GenBank record using the library Biopython, version 1.76 [24]. A Python program was implemented to directly access the genomic intervals containing coding regions, by parsing those features from the GenBank files that contained the keyword 'CDS'. The non-coding intervals were obtained by extracting those regions that are not included in either overlapping or non-overlapping coding regions.

In the case of the segmented viruses, the lack of rightful annotation of one segment resulted in dropping that specific virus from the analysis. As such, for the Orthohantavirus taxonomic genus, where each virus is comprised of three segments, only 26 viruses were considered in the analysis from the total 39 viruses with RefSeq approved genomes, due to lacking segment(s) information in 13 viruses. For the viruses whose genomes contain more than seven segments, the lack of rightful annotation of more than a half of the segments resulted in dropping that specific virus from the analysis. Hence, the following viruses in the dsRNA group where not considered: the Rotavirus B virus from the Rotavirus genus and the Idnoreovirus from the Idnoreovirus genus. The following viruses in the ssRNA(-) were not considered: Groundnut chlorotic fan-spot virus and Peanut bud necrosis virus from the Orthotospovirus genus.

Baltimore classification group	**Unsegmented genomes**	**Segmented genomes**	**Total number of genomes**
ssRNA(+)	1333	373	1706
ssRNA(-)	355	1118	1473
dsRNA	73	890	963
dsDNA	714	0	714

Table 1: Overview of analyzed data

4.2 Analysis pipeline

The analysis pipeline was written in Python version 3.9.13 and incorporated the computation of the Z-scores and/or MFED values, as well as the further processing, statistical analysis and visualization of each data set, as well as .bash scripts. All the scripts are deposited on a public GitHub account repository belonging to the Institute of Theoretical Chemistry research group, found at https://krios.tbi.univie.ac.at/.

4.3 Z-Score and MFED computation

To calculate the Z-score per each position in the genome, a sliding-window approach was first used to scan the whole genome, as shown in Figure 3. Using step size value equal to one, starting from each nucleotide in the genome, for each window of length 100 the MFE was computed. The computation of the secondary structure prediction and the minimum free energy for each window were obtained from the ViennaRNA package version 2.5.0, using the RNAfold Python interface and MFE algorithms described here [71]. The MFE is dependent on the length and base composition of the sequence, which makes it difficult to be interpreted straightforwardly; thus, the Z-score method is used to compare the MFE values of different sequences. The Z-score is calculated as

$$Z = \frac{(m - \mu)}{\sigma}$$

where m is the MFE of the native RNA, μ is the mean and σ is the standard deviation of the MFEs of the shuffled sequences. The computation of μ and σ is usually achieved by generating a large (e.g. N=1000) sample of shuffled sequences and computing the MFE for each of them, which makes computation of z-scores fairly expensive. However, μ and σ are simply functions of the sequence length and composition. Washietl et al [118] showed that a huge speedup can be gained by training a support vector machine (SVM) to estimate μ and σ. An implementation is provided in the ViennaRNA package and achieves accuracy as good as a sample of size 1000.

As the structures that are thermodynamically stable have a low energy, the meaning of a negative Z-score is that the native RNA is more stable than random RNAs, measured independently from the base composition. Thus, the Z-score expresses the significance in standard deviations that the native RNA sequence is more stable that expected by chance. The calculation of the Z-score for each position in the genome was achieved by computing the average Z-score of all sliding-windows in which a particular position was found. This robust approach makes it possible to assess the structuredness of a full genome and provides a reliable method to compare this property across various organisms.

MFED (minimum free energy difference) values were obtained by subtracting the mean MFE value of randomly shuffled sequences (of the same length and base composition as the native RNA) from the native MFE, for each sliding-window, as shown below.

$$MFED = m - \mu$$

The window-length size used in this work is 100 nucleotides. This size was consid-

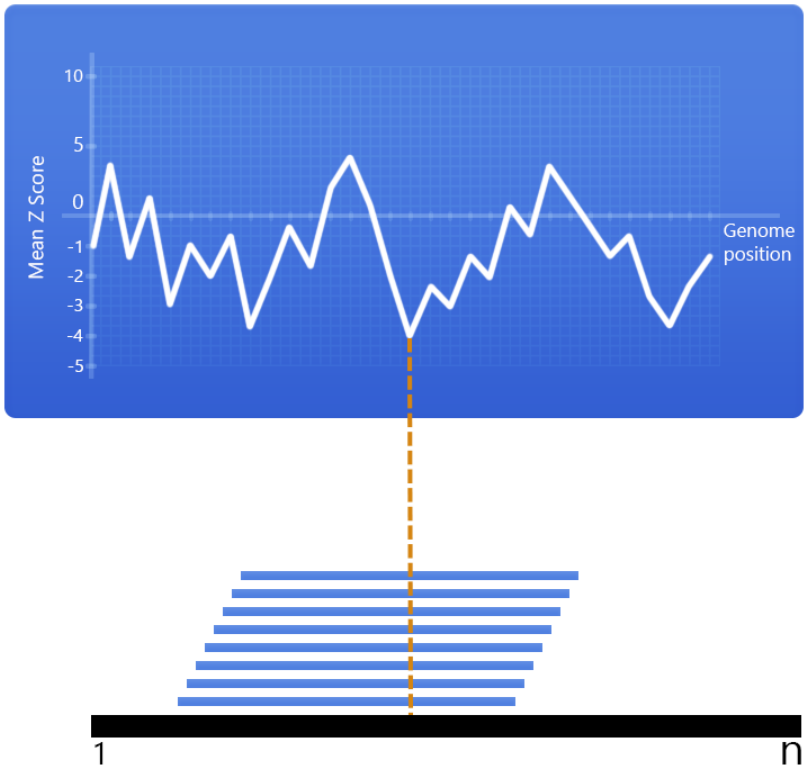

Figure 3: Graphic illustration of the sliding window approach used to scan an entire genome

ered because the average secondary structures in RNA span between 50 and 200 nucleotides [47]. Z-score and MFED values were obtained by automatizing the calculations of the subsequent Python 3 scripts via the internal computer cluster at the Institute of Theoretical Chemistry at the University of Vienna, using Slurm Workload Manager.

4.4 RNALfold and RNAplfold computations

The module RNALfold from the ViennaRNA package was used with the parameter -L equal to the sliding-window length size chosen for the analysis. The function call for RNALfold was incorporated as .bash script in the Python 3 analysis pipeline. The visualization of the RNALfold predictions for the locally stable structures in the genome were designed using the DNA Features Viewer Python library version 3.1.0 [126], with some adaptations to make the visualization of the RNALfold output more insightful with respect to this work. Only the predicted structures with a RNALfold Z-score lower than -2 were considered in the analysis.
To calculate the opening energies, the module RNAplfold was used with the parameters -W 200, -L 100 and -u 30. The parameter -u defines the probability of u consecutive nucleotides to be unpaired. The chosen value of 30 was justified by the fact that the length of binding sites of the ribosome on the RNA are on average around 30 nucleotides long [73, 35]. The -O option allows the output of all the opening energies (logarithm of probability) between a span of u nucleotides. The opening energies are calculated as

$$\Delta G^0 = -RTln(p_u)$$

where p_u is the probability that u consecutive nucleotides are unpaired. The opening energies were calculated for the 30 nucleotide window for which the start position of the first CDS is in the middle and also for the next 30 nucleotide window after that, as seen in Figure 4.

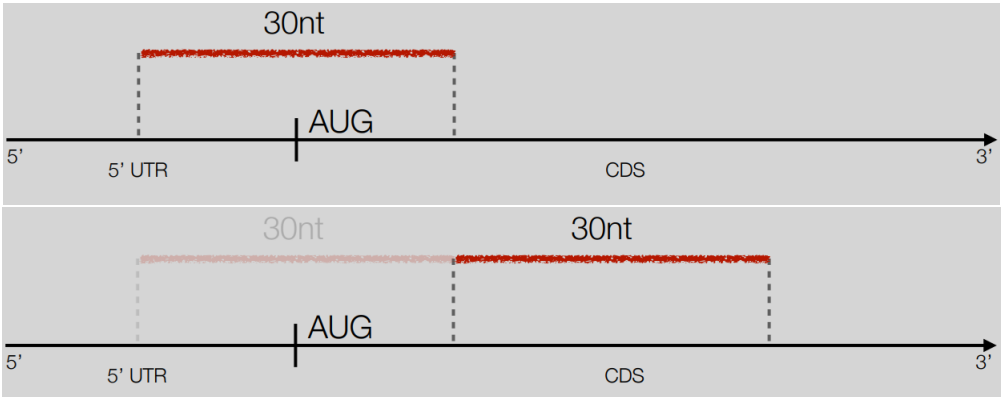

Figure 4: Windows of length 30 spanning the start codon in CDS for the analysis of mean opening energies

4.5 Statistical analysis

The Python library Scipy, version 1.6.2 [115] was used for hypothesis testing. The statannot package version 0.2.3, found at https://github.com/webermarcolivier/statannot, was used to plot and determine the statistical differences between data. The statannot package uses the Scipy library for calculation of the test statistics and p-values. A p-value of 0.05 was considered suitable for the interpretation of the hypothesis tests. To determine if the data in a distribution is normally distributed, the Shapiro-Wilk test has been used from the Python Scipy library. The H_0 hypothesis of the test is that the data is taken from a normal distribution and it is rejected if the p-value for the test statistics (W) is less than 0.05. The two sample Mann-Whitney U test (MWU) was used to test is the data from two groups is sampled from the same distribution. The H_0 hypothesis of the test is that the data has been sampled from two groups with the same distributions and it is rejected if the p-value for the test statistics is less than 0.05. The MWU is suitable for the cases when the distributions set for comparison are unlikely to follow a normal distribution. For the comparison of the distributions of the mean Z-scores in the CDS and non-CDS regions, one of the assumptions of the test is that the data is drawn independently. The classification of the mean Z-score based on the genomic location cannot be done without the annotations provided in the Genbank files from each virus and as such, this approach was considered to be in agreement with the independence requirements of the test.

4.6 GisAid database

For the analysis of the CoVariants sequencing data, the differentiation between the different SARS-CoV-2 strains was made based on the information provided at https://covariants.org/ [46]. The different strains are differentiated based on the accumulated mutations with respect to the reference genome, (especially in the spike protein), as well as their geographic origin. Nucleotide sequences from the different CoVariants were manually curated and downloaded from the GisAid database [56], in

May 2022. The GisAid database collects genetic sequences and epidemiological data from laboratories around the world and makes them available for research purposes, overcoming some of the restrictions discouraging the disclosure of virological data prior to publication, for example. The selected genomes for each CoVariant were chosen arbitrarily to be from patients in different geographic locations and to have less that 1% missing nucleotides. A multi Fasta file was downloaded containing all genomes for each CoVariant. For the Z-score computation, a Python (version 3.9.13) script was developed to obtain the individual nucleotide sequences and process the data for the further analysis.

5 Results

5.1 Mean Z-scores in the RNA virus groups

Z-scores were computed for the ssRNA(+), ssRNA(-) and dsRNA viruses using a sliding window approach of length 100, as described in the Materials and Methods part. In Figure 5, the violinplots for each Baltimore group are shown with differentiation between the forward and backward strand for the ssRNA(-) and dsRNA viruses.

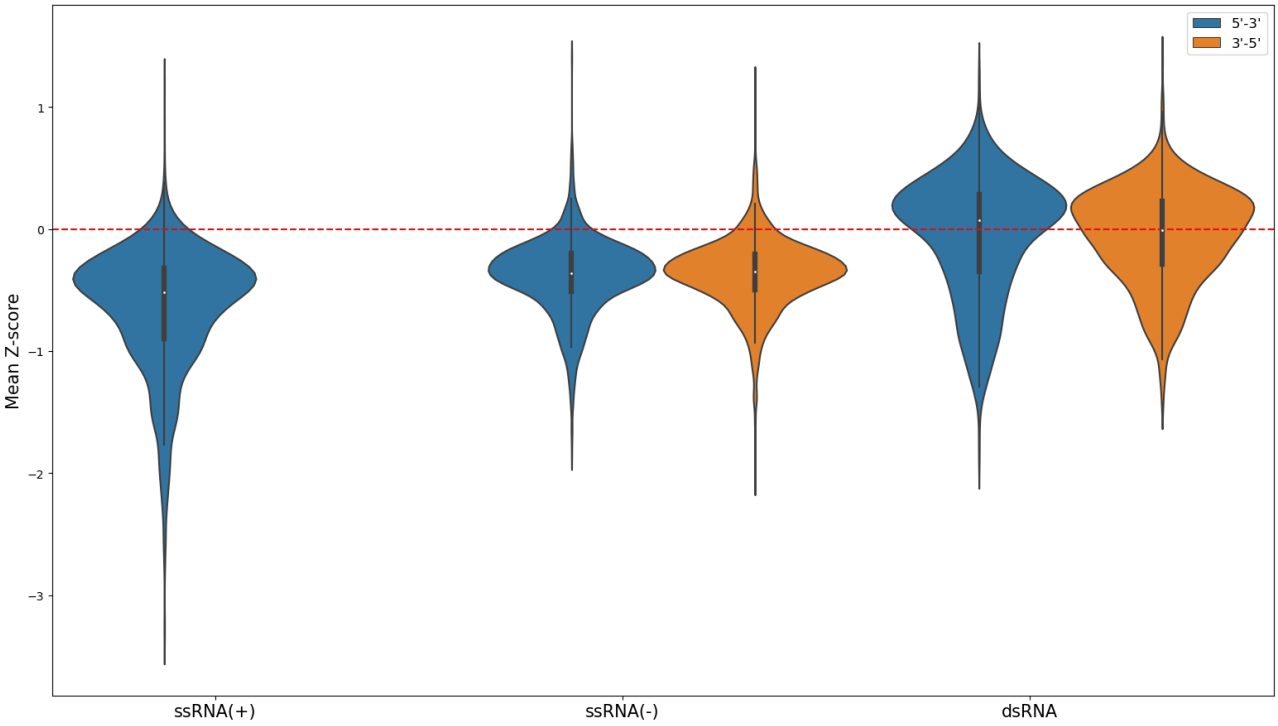

Figure 5: Violinplot of mean Z-scores for all RNA virus groups

For the ssRNA(+) and ssRNA(-), the majority of the mean Z-scores per virus are below 0, which suggests that there is a trend of structuredness in most genomes. All virus groups have mean Z-scores that are below 0: ssRNA(+) -0.65, ssRNA(-), 5'-3' -0.386, ssRNA(-), 3'-5' -0.368, dsRNA, 5'-3' -0.058, dsRNA,3'-5' -0.057. To be noted that in the case of dsRNA, the mean Z-scores for either strands are more positive than for the ssRNA(-) and ssRNA(+), which is to be expected, since their genomes exists only as a double helix and thus offers the stability and structure the genomes need to fulfil their functions. In contrast, the ssRNA(+) viruses seem to have the most structured genomes of all groups, with a mean Z-score of -0.65, which implies a certain genome stability and structuredness that may be needed for the ssRNA(+) to survive inside of their hosts and replicate their genomes successfully. The ssRNA(-) viruses have similar mean Z-score values for both the 5'-3' and 3'-5' strand, with the antigenome strand have a slightly higher mean Z-score that the genome strand. The distribution plots in ssRNA(-) show a similar behaviour, in that the majority of the Z-scores are highly centered around the median, as opposed

to the other two groups. In the case of the dsRNA, on the 5'-3' it seems like the distribution of the Z-scores is multimodal, but the most predominant peak is located in the proximity of the median value.

5.2 Mean Z-scores in coding and non-coding regions

To further assess the structuredness of RNA viruses, it was interesting to determine whether there is a difference in Z-scores between coding (CDS) and non-coding (non-CDS) regions in the genome. This differentiation has been made based on the annotation found in the Genbank file of each virus. All the entries labeled as CDS in the Genbank file have been considered as coding regions, while remaining positions have been annotated as non-CDS. All Z-scores in either CDS or non-CDS have been calculated in each virus group in order to assess the differences in structuredness between them and are shown as violinplots (Figure 6).

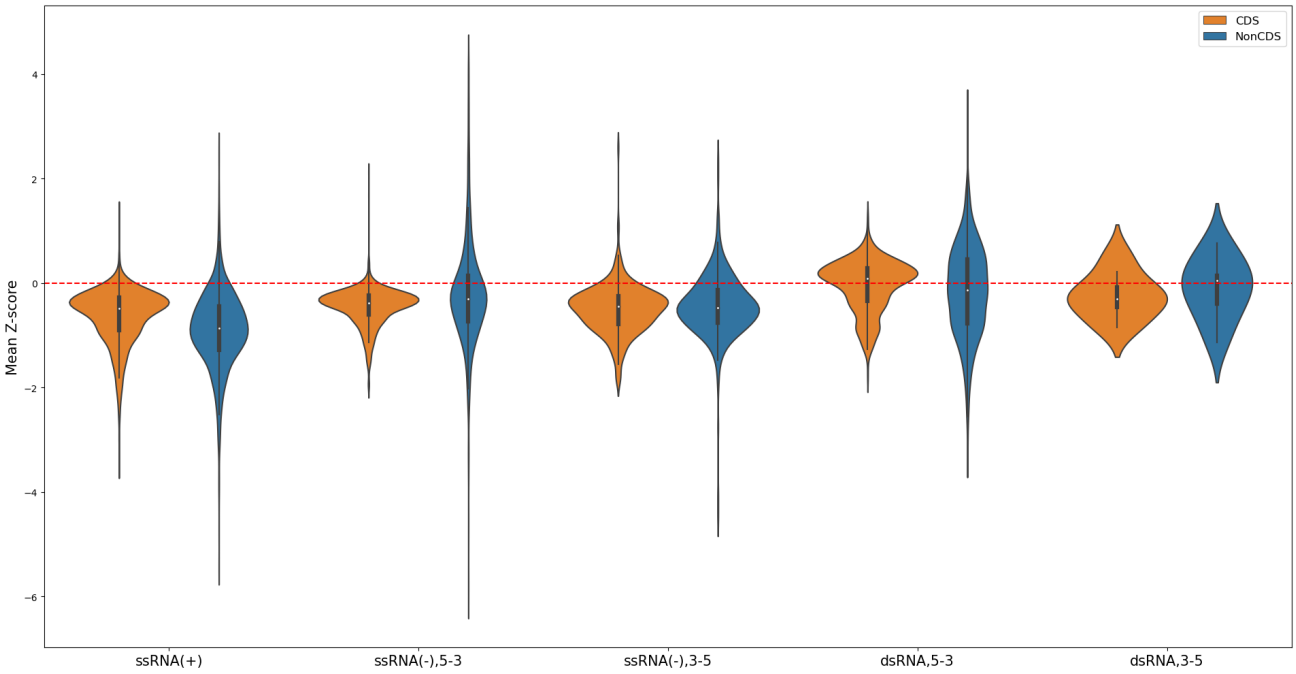

Figure 6: Violinplot of mean Z-scores in CDS and non-CDS per taxonomic group

For the ssRNA(-) and dsRNA viruses it was differentiated again between the Z-scores per region for the 5'-3' and 3'-5' strands, as there are CDS found on either strands and the annotation in the Genbank file provided the coding regions for both the forward and backward strand. In the case of ssRNA(+) viruses, since their genome is acting as the mRNA and there are no coding regions on the backward strand [82], only the 5'-3' strand has been considered. Figure 6 shows, for each Baltimore group, the distribution of the Z-score in the CDS and non-CDS region. In the ssRNA(+), the average Z-score in the CDS (-0.64) is more positive than in the non-CDS (-0.9), which suggests that the non coding regions for these viruses are more structured than the coding regions, as opposed to the ssRNA(-), where the mean Z-score in the non-CDS is higher, in both strands (-0.46 vs -0.23 on the 5'-3' and -0.5 vs 0.45 on the 3'-5'). The distribution of the Z-scores in the non-CDS on the 5'-3' of the ssRNA(-) is skinny on the ends and wide in the middle, indicating that the Z-scores in the non-CDS are highly concentrated around the median. For the

dsRNA viruses, there seems to be more structuredness in the non-CDS of the 5'-3' strand, however for the 3'-5' strand, the CDS regions tend to be more structured. The results for the dsRNA viruses may not be as robust as for the other two groups, as the genome annotation in these viruses is very often missing with regards to the delimitation of coding regions inside the 3'-5' genome, therefore only 9 viruses could be used to characterize that strand.
Additionally, Figure 7 shows the mean Z-score values in CDS and non-CDS as box plots for all the viral groups. This overview makes it possible to observe some trends in how different regions of the genome, which undergo different biological functions, behave from a thermodynamic point of view.

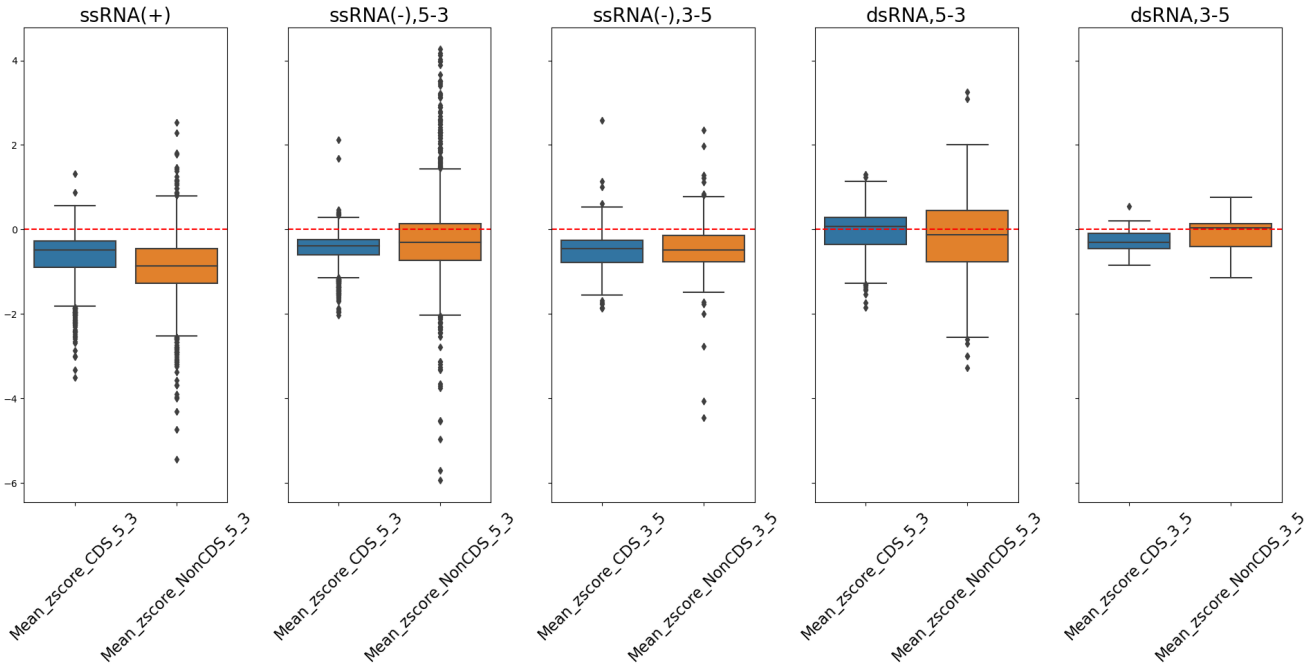

Figure 7: Boxplots of mean Z-scores in CDS and non-CDS per group

5.3 Mean Z-scores and GC content

The mean Z-scores for all viral families belonging to each Baltimore group were calculated and plotted together with the mean GC content per family. The GC content was particularly of interest, as it is known that GC-rich sequences throughout the genome offer more stability, due to the thermodynamic superior stability of a G-C base pair compared to G-U or A-U pair and also by their low mutability rate, which is the reason why the G-C base pairs are overall preferred in the helical regions, for example [45, 20]. It was also determined that an elevated GC content enhances transcription and tunes the overall mRNA expression levels [85]. Since the Z-score is a marker that the MFE structure of a sequence has a significantly smaller value that the MFE of random sequences of the same length and base composition [40], it is interesting to determine how the GC content and mean Z-score values integrate in the structuredness analysis of the viral RNA genomes.

5.3.1 Mean Z-scores and GC content per virus family in ssRNA(+)

Figure 8 shows the relationship between the mean Z-score and GC content for each family in the ssRNA(+) viruses. The top three families with the most negative mean Z-scores values of -2.22, -2.03 and -1.96 are *Steitzviridae*, *Narnaviridae* and *Fiersviridae*, respectively. However, it seems that the genomes of these highly-likely structured viruses are not the richest in the GC content, as the mean value of their GC content is around 0.5. The *Steitzviridae* family contains only one virus, the Caulobacter phage phiCb5, which has been used as a model to study translational control and virus evolution in molecular biology. The 3762 nucleotide long genome forms a putative stemloop structure at the 5'-end, which is characteristic in all RNA phages, as well as additional three stemloops near the RP (replicase subunit) AUG start codon [54]. This denotes that a high level of structuredness can be achieved even with an equal AT/GC content in the nucleotide sequence.
Despite the comparatively much smaller mean Z-score (-0.22), the *Matonaviridae* family has the biggest GC content (almost 70%) in the ssRNA(+) families. In this family, only the Rubella virus is included, which causes stillbirth or spontaneous abortion if the infection takes place during pregnancy in humans [1]. The Rubella virus has a length of 9762 nucleotides and until 2018 was classified in the *Togaviridae* family, however now it belongs to the *Matonaviridae* family [117]. The mean Z-score value for this virus is -0.22, which is the fourth less negative value in ssRNA(+). On the other hand, the *Mononiviridae* family has the lowest GC content (0.27) and also the highest mean Z-score value (-0.03). It also contains only one virus, the Planarian secretory cell nidovirus, whose genome length is 41.178 nucleotides.
A number of conserved secondary structure elements, such as stemloops and hairpins, have been previously described for the viruses in the *Picornaviridae* family, both in the UTR regions as well as in CDS. Using MFE and sequence alignments of the viruses in six different genera of the *Picornaviridae*, it was shown that, in particular, the CDS contains one structural feature across genera, known as CRE (cis-acting replication element) and other secondary elements that are conserved within a genus, especially for Apthovirus [121]. Another study was able to determine the stable secondary structures in the 3'-UTR of the swine pasivirus, a species belonging to the Pasivirus genus [16]. These results support the findings in this thesis, as the 144 analysed genomes in the *Picornaviridae* family have a mean Z-score of -0.53, denoting a high structuredness tendency. It seems that there is not a strict trend for those families with a low mean Z-score to also have a low GC content throughout their genomes and vice versa. This overview plot shows that the property of being structured as a whole is not necessarily linked to having preponderantly GC-rich regions in the RNA genome, at least at family taxonomic level.

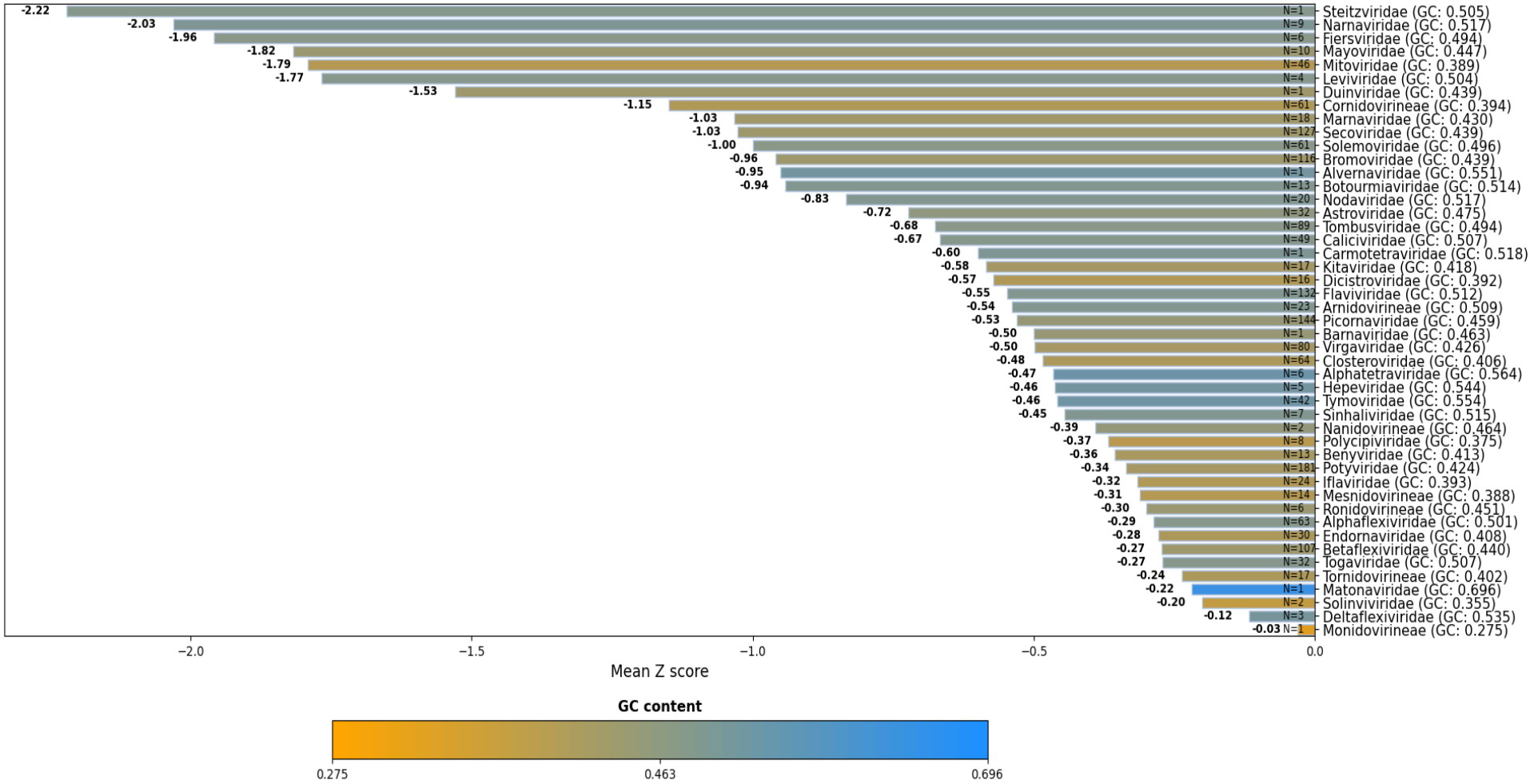

Figure 8: Mean Z-scores and GC content per family in ssRNA(+). The mean value of the GC content per family is plotted on each bar, and the total range is shown below the plot. The number of species in each family is displayed on the right side of the bars. The mean Z-score values are shown in descending order.

5.3.2 Mean Z-scores and GC content per virus family in ssRNA(-)

Figure 9 shows the relationship between the mean Z-score and GC content for each family in the ssRNA(-) viruses for the 5'-3' strand. The family with the most negative mean Z-scores on the forward strand is *Arenaviridae* and consists of 100 viruses, whose mean GC content is 0.41. As seen previously in the case of the analysed ssRNA(+) viruses, the families with the lowest Z-scores do not have the highest GC content in their group. The sixteen viruses from the *Qinviridae* family have the highest GC content of all ssRNA(-) viruses, namely 0.508. The viruses in this family have segmented genomes and the typical hosts are crustaceans and insects [52]. The mean Z-score of the *Qinviridae* family lies at around -0.16, which is close the the mean Z-score value of the *Fimoviridae* family (-0.19), whose GC content is, with a value of 0.317, the lowest of all families for the 5'-3' strand in the ssRNA(-) group. The *Filoviridae* family, which has been well-studied in the last years, has a GC content of 0.42, while its mean Z-score has a value of around -0.2 on both strands. Apart from the Ebola and Marburgvirus genera, which are known to infect humans, this family also contains fish filoviruses, whose genomes are considerably smaller (approximately 16 kb vs 19 kb in Ebolaviruses).
The virus family with the lowest mean Z-score is *Arenaviridae*, whose members are segmented viruses which infect mammals, snakes and rodents [102]. It was suggested that the intergenic region between the proteins that are translated on each segment is highly structured and plays a role in mRNA transcription termination [89]. These results show that is it possible to have a high GC content even if the structuredness measure in term of the Z-score is not as low as one would expect.
Further, Figure10 shows the mean Z-scores and the GC content for the 3'-5' strand in the families of the ssRNA(-) viruses. In the previous section, in Figure 5, it was shown that the value of the mean Z-scores for the forward and backward strand in the ssRNA(-) viruses do not vary a lot, hence it is expected that the behaviour of the mean Z-scores per family in both strands is similar. In Figure 10, the same families as in Figure 9 populate the top of the plot and their mean Z-score values are very similar, while the GC content remains the same for all families, since the 3'-5' strand is the reverse complement of the 5'-3' and the nucleotide composition does not change. Nevertheless, there are some bigger differences in the mean Z-scores, for example in the *Cruliviridae* family (mean Z-score on the 5'-3' strand is -0.29 and on the 3'-5' strand -0.39) or *Yueviridae* family (mean Z-score on the 5'-3' strand is -0.13 and on the 3'-5' strand -0.22).
The virus family with the highest mean Z-score value (-0.13) on the forward strand is the *Xinmoviridae* family, which contains six viruses. Its GC content (0.382) is also one of the lowest in this group. As depicted in Figure 9, on the 5'-3' strand, the mean Z-score of these viruses lies around -0.12 and is the second highest mean Z-score value on the backward strand. As mentioned before, the behaviour of the mean Z-scores per family in ssRNA(-) was expected to be similar, since the mean Z-scores values of all viruses on the 5'-3' and 3'-5' strand are close (see Figure 5).

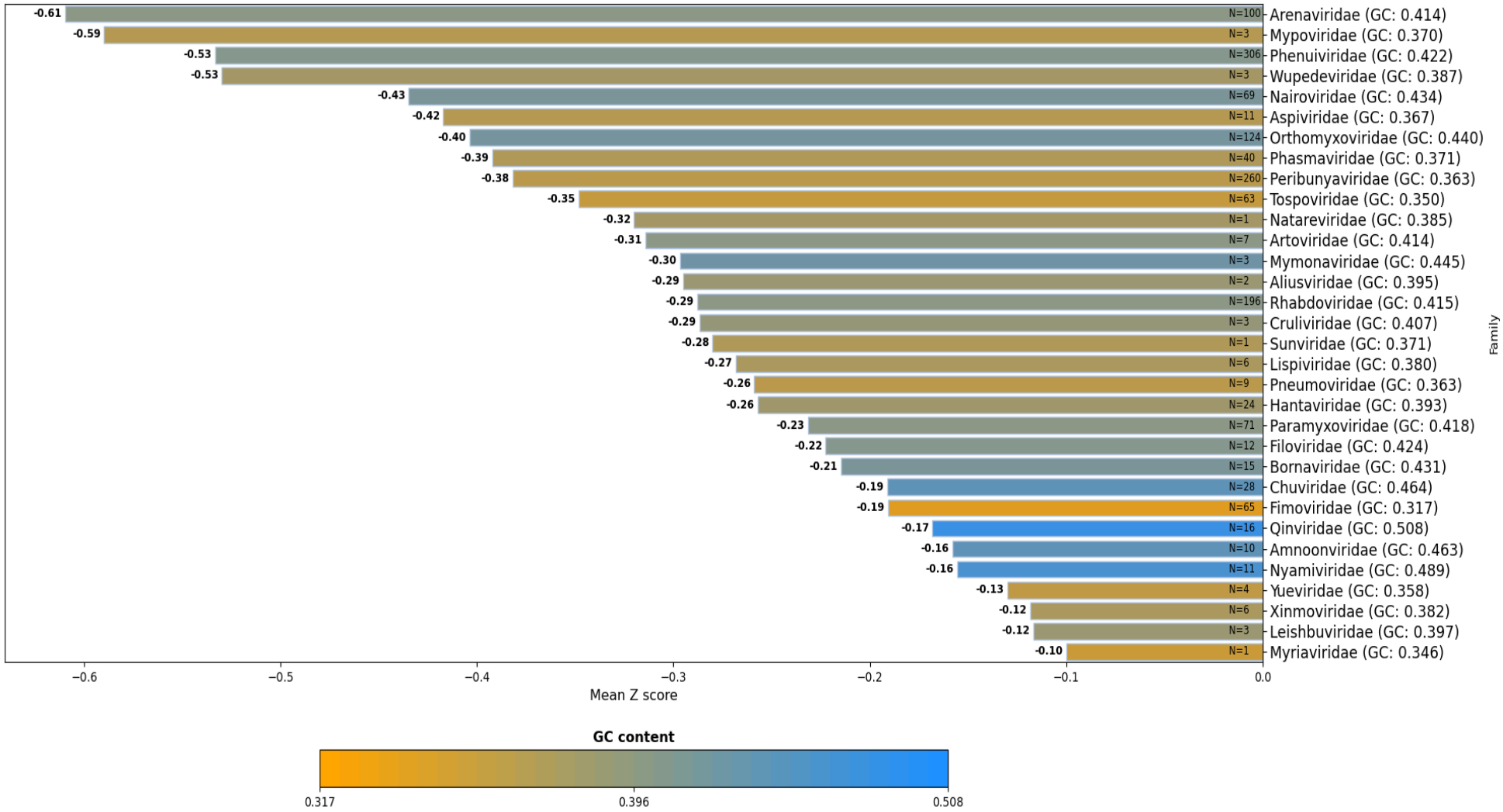

Figure 9: Mean Z-scores and GC content per family in ssRNA(-), 5'-3' strand. The mean value of the GC content per family is plotted on each bar, and the total range is shown below the plot. The number of species in each family is displayed on the right side of the bars. The mean Z-score values are shown in descending order.

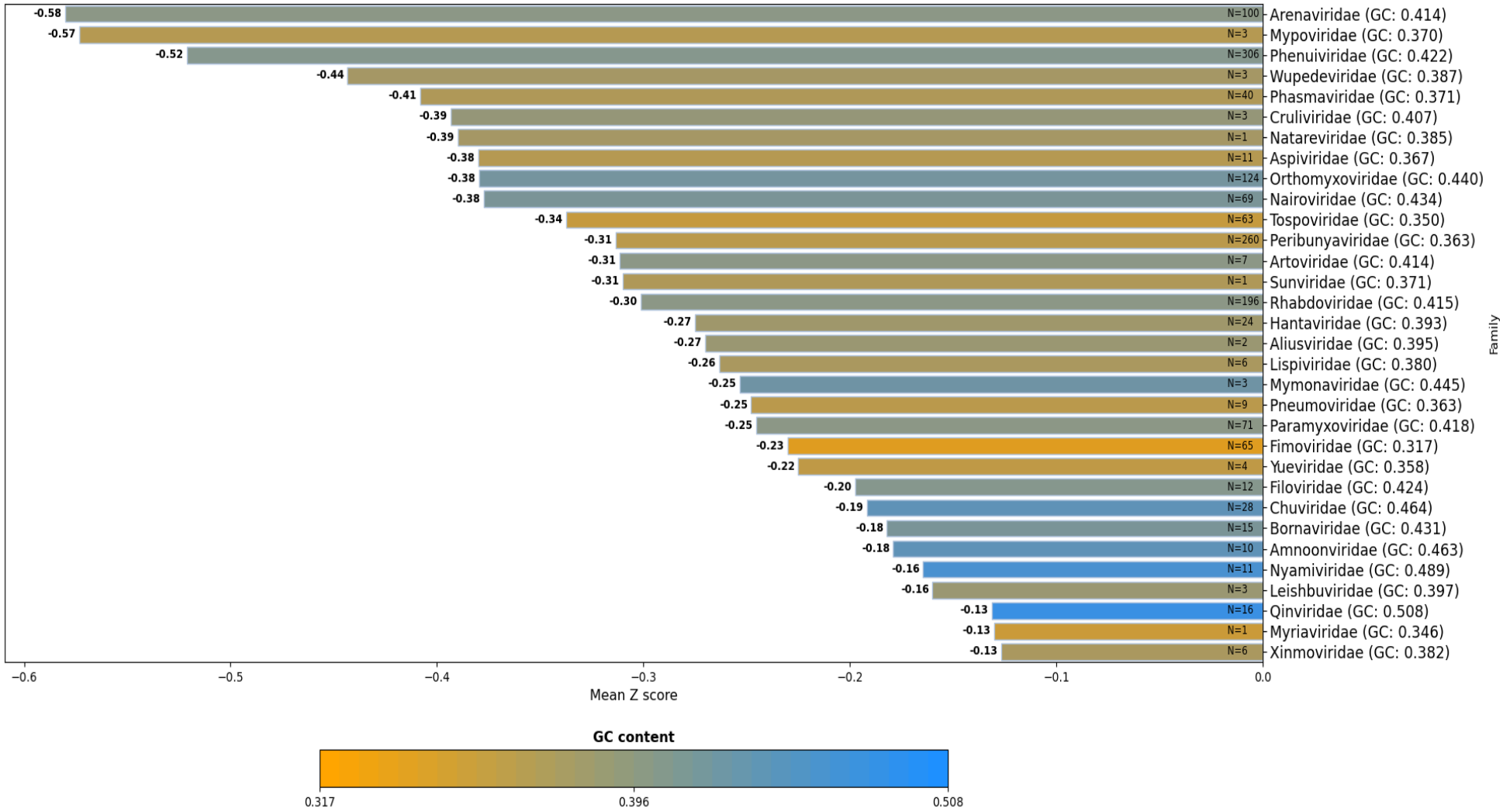

Figure 10: Mean Z-scores and GC content per family in ssRNA(-), 3'-5' strand. The mean value of the GC content per family is plotted on each bar, and the total range is shown below the plot. The number of species in each family is displayed on the right side of the bars. The mean Z-score values are shown in descending order.

5.3.3 Mean Z-scores and GC content per virus family in dsRNA

The same analysis was conducted for the dsRNA viruses, which were also divided by family and shown with their mean Z-scores and GC content on the 5'-3' strand (Figure 11) and 3'-5' strand (Figure 12). Interestingly, the dsRNA virus family that has the most species also has a positive mean Z-score. The *Reoviridae* family contains segmented viruses and are known to infect fungi, plants, invertebrates and vertebrates [91]; it is comprised of 720 genomes with a mean Z-score of 0.10, which makes it the only family in all Baltimore groups to have a positive measure of structuredness. The segmented genomes of the viruses in the Reoviridae family range from 400 to 4000 nucleotides in length and their mean GC content lies at around 0.4, the lowest in all dsRNA families. The Rotavirus genus, which belongs to the *Reoviridae* family, was subject to extensive research and it was shown that the majority of conserved secondary structures identified in the eleven segments of rotaviruses have a medium to low-average base pairing probability [68], which is in line with the mean Z-score result (0.1) presented in this thesis , as the tendency of having small base pairing probabilities is related to the property of a particular region not to be very structured. Nonetheless, the value of the GC content (0.4) is considerably high when comparing it to the GC content of some families in the ssRNA(+) and ssRNA(-) viruses, while the mean Z-score implies a rather unstructured architecture of their genomes.
Also with a low GC content inside the dsRNA group are the viruses of the *Picobirnaviridae* family; they exhibit the most negative mean Z-score in this group (-0.95) on the 5'-3' strand, but their GC content is around 0.40. *Picobirnaviridae* contains small, segmented viruses which cause gastroenteric and respiratory infections in mammals [74]. Research in the structure of this family is still very limited, but the results of the thermodynamic predictions in this work may be a starting point in providing more information about the genome structure of these viruses. Their genomes are the most structured ones in the dsRNA group, with a mean Z-score of -0.95 on the positive and -0.8 on the negative strand. Together with the *Cystoviridae*, *Fiersviridae* and *Leviviridae* families, the latter two belonging to the ssRNA(+) group, they are prokaryotic viruses. The *Cytoviridae* family has the highest GC content (0.55) and also the second most negative mean Z-score (-0.87). This family is comprised of 21 segmented genomes whose lengths range from 2000 to 4700 nucleotides. Interestingly, the mean Z-scores in these prokarytic viruses denote that they have one of the most structured genomes in their groups (mean Z-score *Cystoviridae* -0.87, mean Z-score *Fiersviridae* -1.96, mean Z-score *Leviviridae* -1.77). These prokaryotic RNA viruses have been recently added in their own taxonomic families and therefore have not yet been vastly investigated by the research community.
Figure 12 shows the analysis results for the 3'-5' strand of the dsRNA viruses. Not surprisingly, the *Reoviridae* family mean Z-scores are consistent on both strands, as their mean Z-score on the 3'-5' strand is also positive (0.06). The mean Z-score of the *Cystoviridae* family is slightly lower on this strand, with a value of -0.61 compared to -0.87 on the 5'-3' strand. Overall, the mean Z-scores on the 3'-5' strand are higher than on the 5'-3'. It should be noted that the GC contents in dsRNA viral families range only from 0.43 to 0.55, making the difference between the lowest and highest GC content less than 0.15.

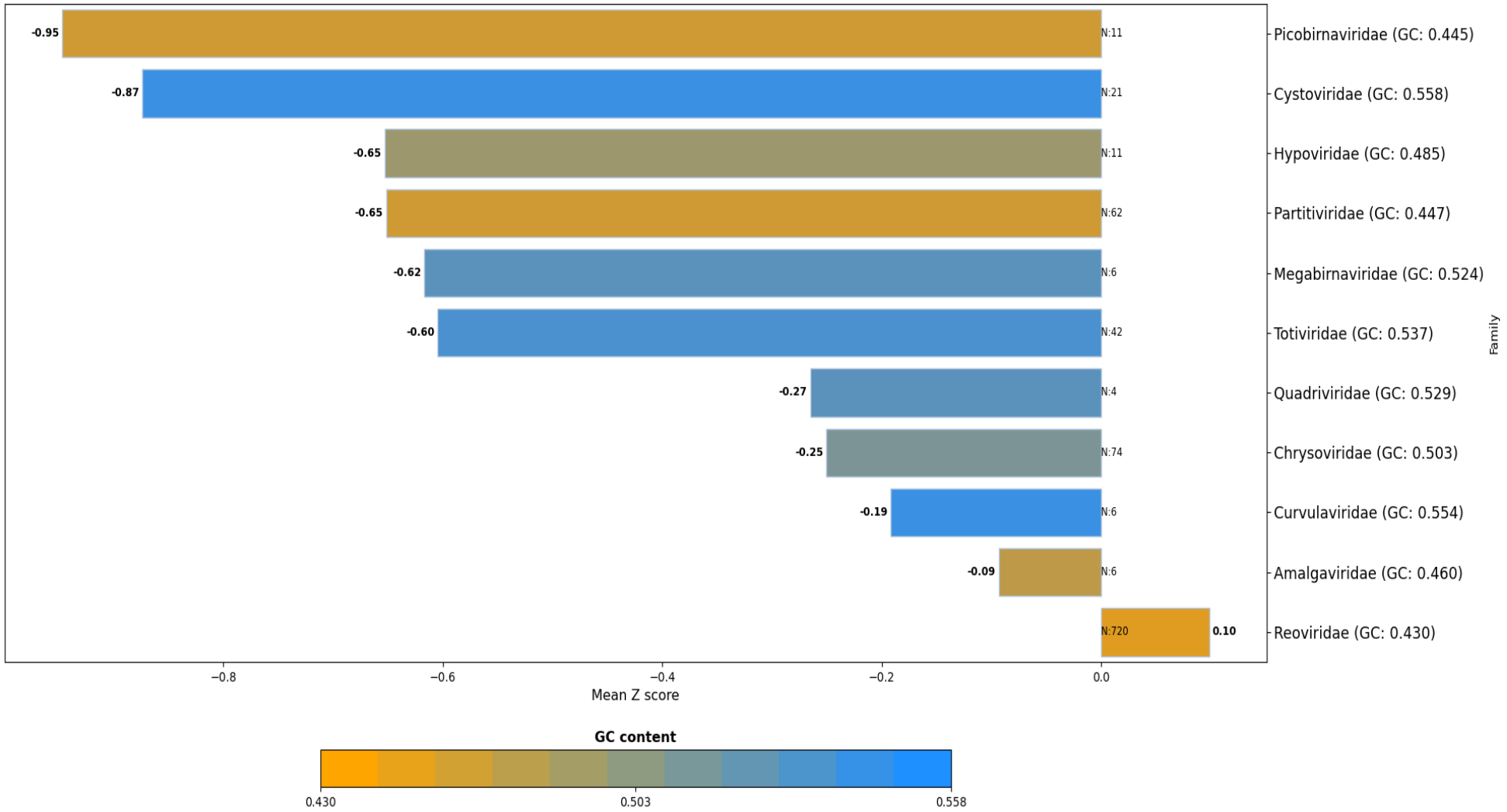

Figure 11: Mean Z-scores and GC content per family in dsRNA, 5'-3' strand. The mean value of the GC content per family is plotted on each bar, and the total range is shown below the plot. The number of species in each family is displayed on the right side of the bars. The mean Z-score values are shown in descending order.

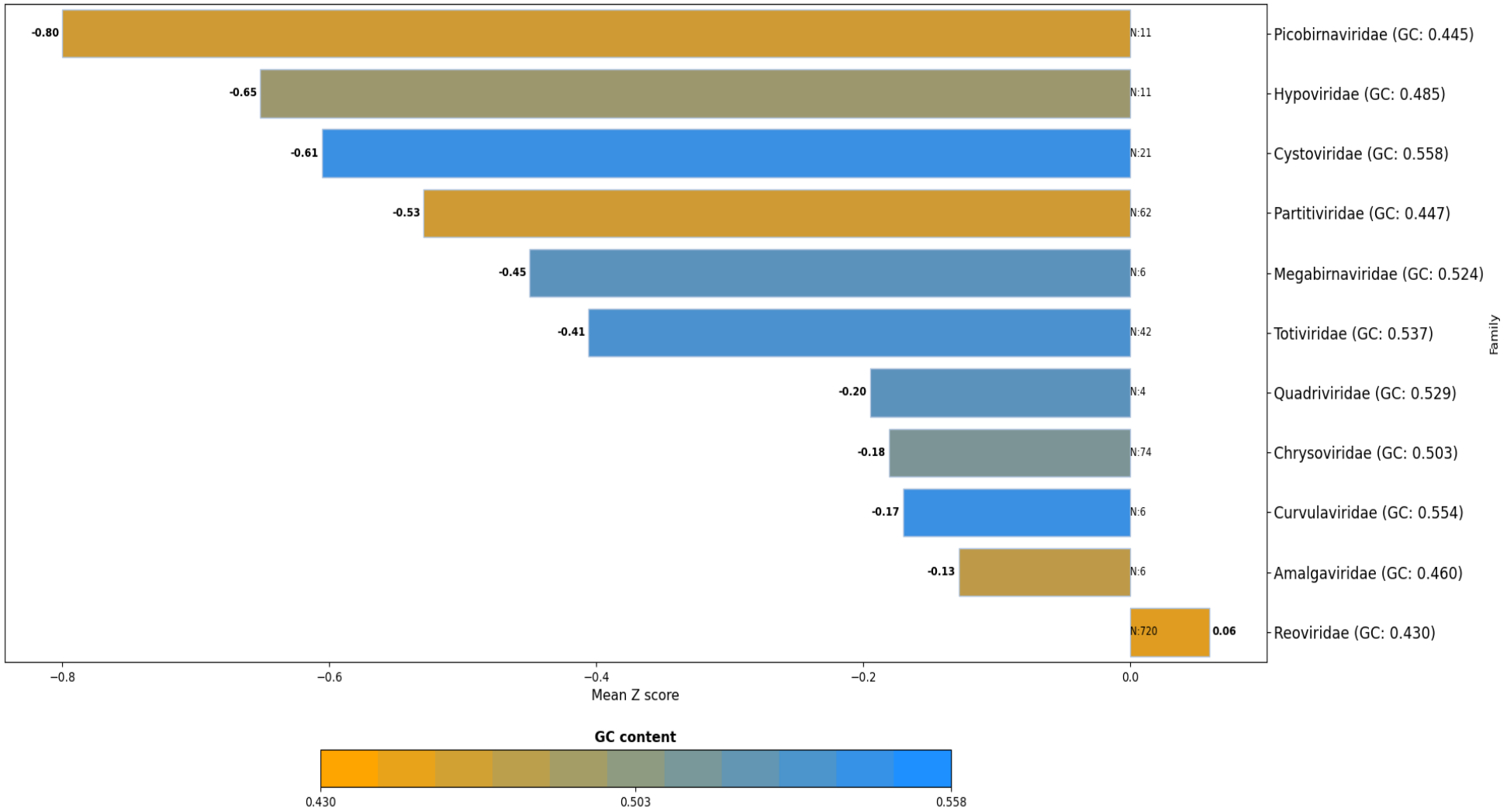

Figure 12: Mean Z-scores and GC content per family in dsRNA, 3'-5' strand. The mean value of the GC content per family is plotted on each bar, and the total range is shown below the plot. The number of species in each family is displayed on the right side of the bars. The mean Z-score values are shown in descending order.

5.4 Mean Z-scores and opening energy values

In this result section, the focus lies on the correlations between the opening energy values and mean Z-score values, in specific regions in the 5'-3' strand of the genome, as well as across the whole genome. The opening energies are a feature of the RNAplfold program in the ViennaRNA package [71] and are described in more detail in the Materials and Methods part. The opening energy option, used together with the -u option parameter, defines the average opening energy needed to unfold a specific region of length u in the genome [14]. Firstly, the mean opening energies for the whole genome have been calculated and plotted together with the mean Z-score values, again classified by family in each Baltimore group. A second analysis included the calculation of mean opening energies and mean Z-scores for the sequence of nucleotides around the position of the first CDS. This was done by parsing the first CDS from the Genbank file of each virus (for which this information was available) and then selecting the sequence of nucleotides so that the position of the first CDS is the median of the sequence. A correlation plot was created to determine the relationship between mean Z-scores and mean opening energies around the first CDS.

5.4.1 Mean Z-scores and mean opening energy values in ssRNA(+)

Figure 13 shows the mean Z-scores (orange) depicted together with the mean opening energies (blue) for the entire genome, for all families in the ssRNA(+) group. The highest value for the opening energy lies at 13.79 kcal/mol for the *Matonaviridae* family, whose mean Z-score has the value -0.22. The *Steitzviridae* family, which has the lowest mean Z-score among all ssRNA(+), has an overall opening energy value of 11.22 kcal/mol, the second highest in this group. High opening energies are an indicator that the nucleotide sequence is particularly structured, and in order to break the nucleotide bonds a lot of energy needs to be used. This energy correlates with the probability of a given sequence to be rather unpaired, meaning that the higher the unpaired probability, the lower the energy.
To further assess the genome's thermodynamic properties regarding both the structure and the free energy needed to unfold the structure, a more in depth analysis was undertaken to investigate how the opening energy behaves around the region of the first start codon.
Table 2 shows, for each family, the mean Z-score of the sequence of nucleotides in which the first coding region is located, together with the opening energy values determined using the RNAplfold program. The highest energy value (16.45 kcal/mol) was predicted for the *Hepeviridae* family, which also exhibits a considerably negative mean Z-score value for the sequence containing the first coding region in their genomes. The opening energy in this region is more than 6 kcal/mol higher than the average mean opening energy for the whole genome in this family. This difference may occur as a result of a much more structured area surrounding the AUG start-codon than the overall structuredness across the whole genome. There are 5 viruses in the *Hepeviridae* family and their average genome length is approximately 7000 nucleotides long.
In contrast, the *Dicistroviridae* family has a mean Z-score value of 0.05 in the region surrounding the first CDS, which is a rather lpositive value, suggesting that the sequence is somewhat unstructured. Its mean opening energy value is one of the

smallest in the ssRNA(+) group, with a value of 4.57 kcal/mol. This would suggest that there is less free energy needed to disturb the structures formed in that particular region, because the region itself is not thermodynamically stable, as reflected in the mean Z-score value.

Family	**Mean Z-score spanning first CDS**	**Opening Energy spanning first CDS (in kcal/mol)**
Calciviridae	-3.562	11.971
Carmotetraviridae	-2.735	10.653
Duinviridae	-2.434	8.771
Gammaflexiviridae	-2.250	16.447
Hepeviridae	-2.223	11.414
Alvernaviridae	-2.169	9.696
Matonaviridae	-2.031	8.344
Endornaviridae	-2.009	5.981
Leviviridae	-1.898	6.094
Narnaviridae	-1.819	10.746
Tornidovirineae	-1.669	7.986
Flaviviridae	-1.666	9.159
Fiersviridae	-1.566	8.769
Ronidovirineae	-1.514	11.173
Deltaflexiviridae	-1.505	6.699
Astroviridae	-1.391	10.188
Cornidovirineae	-1.319	8.978
Alphatetraviridae	-1.296	9.202
Secoviridae	-1.283	8.231
Mayoviridae	-1.075	8.060
Monidovirineae	-1.074	10.020
Togaviridae	-1.052	11.907
Mesnidovirineae	-1.005	6.394
Arnidovirineae	-0.988	9.676
Mitoviridae	-0.983	5.255
Kitaviridae	-0.978	10.038
Solinviviridae	-0.961	8.687
Betaflexiviridae	-0.937	7.445
Nodaviridae	-0.932	6.832
Marnaviridae	-0.838	6.335
Steitzviridae	-0.808	12.647
Barnaviridae	-0.807	4.522
Closteroviridae	-0.712	8.093
Tombusviridae	-0.657	8.624
Virgaviridae	-0.632	5.327
Picornaviridae	-0.6	7.253
Solemoviridae	-0.549	9.144
Bromoviridae	-0.531	6.851
Alphaflexiviridae	-0.529	8.270
Botourmiaviridae	-0.513	7.627
Tymoviridae	-0.437	7.577
Potyviridae	-0.172	6.734
Benyviridae	-0.079	7.525
Sinhaliviridae	-0.068	8.831
Polycipiviridae	0.031	4.703
Dicistroviridae	0.045	4.568
Nanidovirineae	0.101	6.790
Iflaviridae	0.163	6.113

Table 2: Overview of ssRNA(+) families, the mean Z-score and opening energy values spanning the first CDS

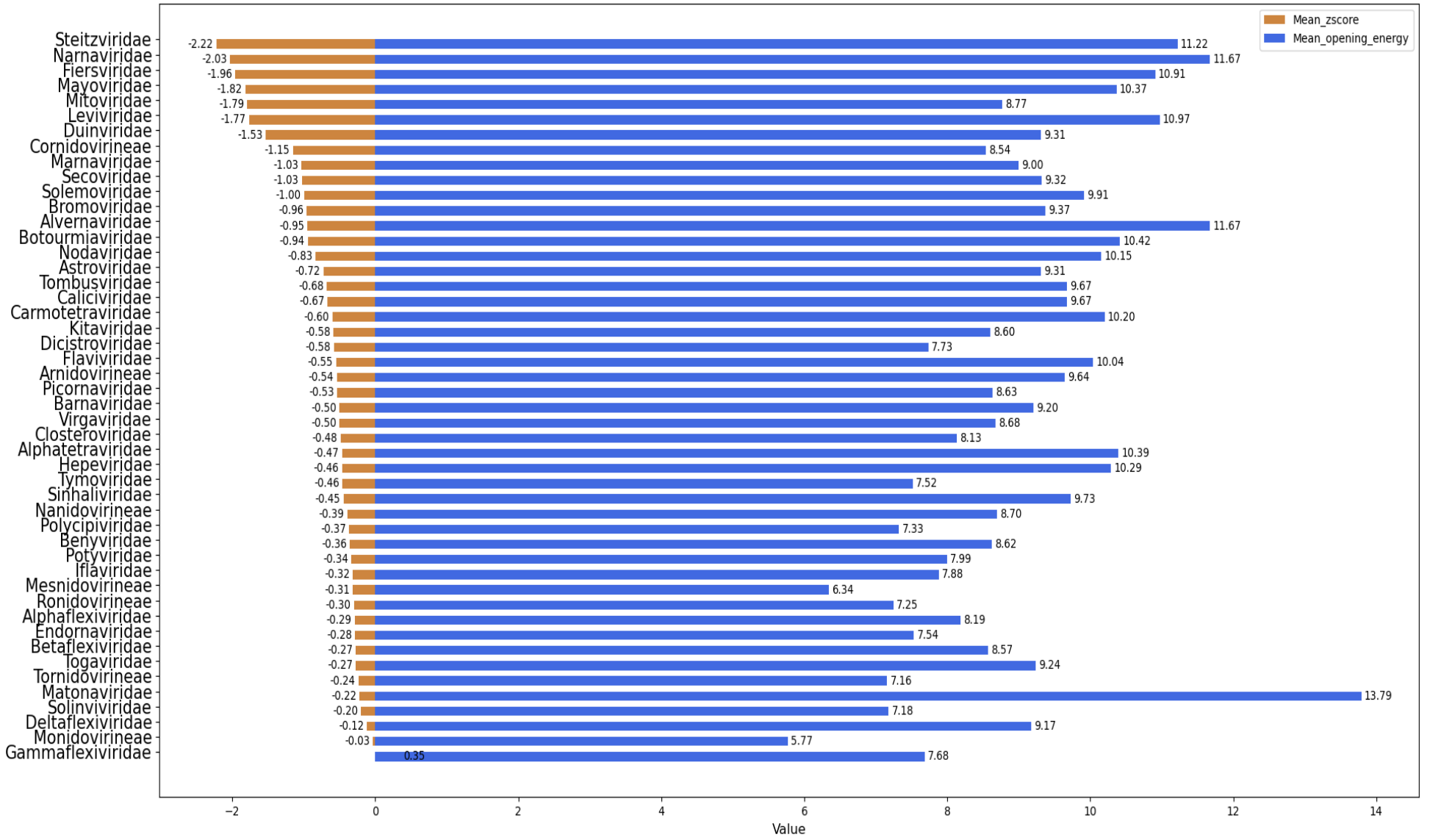

Figure 13: Mean Z-scores and mean opening energy values for the whole genome in ssRNA(+). The mean value of the Z-scores is shown in blue, while the mean value of the opening energies is shown in orange, grouped by taxonomic family

Finally, to determine if there is indeed a correlation between the mean Z-score and the mean opening energies around the first coding region, a correlation plot was created for the families in the ssRNA(+) group, as shown in Figure 14. The Pearson correlation coefficient (Person's r) has a value of -0.366, indicating a moderate trend that the nucleotide sequence surrounding the first CDS has a lower mean Z-score, the higher its mean opening energy value is. This makes sense, because the free energy required to unfold a stable structure needs to be much higher than for the destabilization of a structure whose MFE has a more positive value.

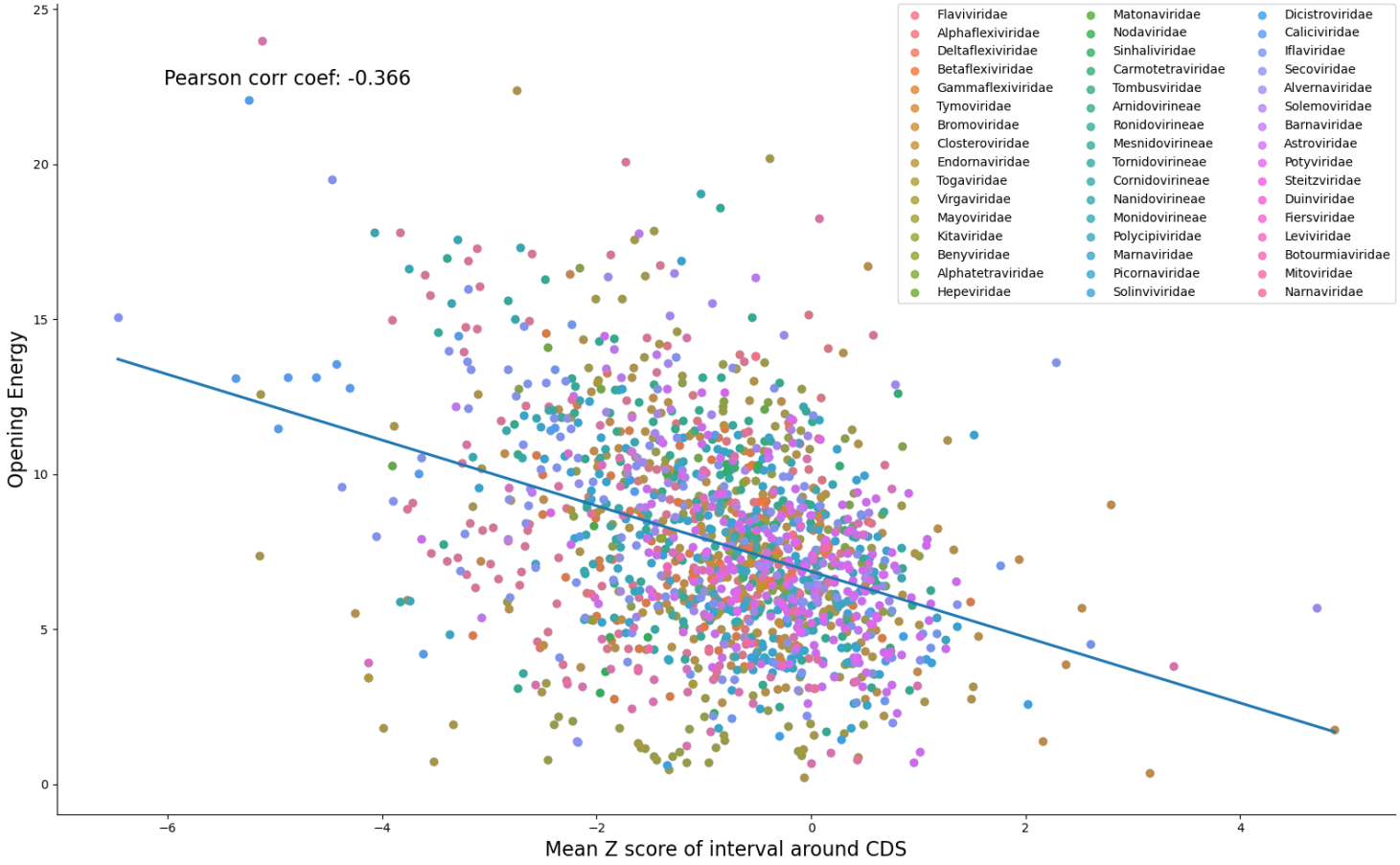

Figure 14: Correlation of the mean Z-scores and mean opening energy values in the region of the first CDS in ssRNA(+). Each data point depicts a species, grouped by taxonomic family (see legend)

5.4.2 Mean Z-scores and mean opening energy values in ssRNA(-)

Similarly to the analysis in the ssRNA(+) viruses, the 5'-3' genome strand of the ssRNA(-) viruses was also investigated to determine the interplay between mean Z-score and mean opening energy, first across the whole genome and then only in that part of the genome where the first AUG codon is located. Figure 15 shows the mean Z-scores (blue) together with the mean opening energies (orange) across the whole nucleotide sequence, grouped again by taxonomic family.
The highest mean opening energy in the ssRNA(-) group is 9.52 kcal/mol. This value is attributed to the *Aliusviridae* family, whose mean Z-score value for the entire genomes lies at -0.29. The *Myriaviriade* family, with the overall smallest mean Z-score value (-0.1) in ssRNA(-), has the second highest mean opening energy value (9.44 kcal/mol). The family with the most structured genomes *Arenaviridae*, has a mean opening energy value of only 7.84 kcal/mol. By looking at the results in this group, it would seem that there is no specific trend for the viral families to follow: there are structured genomes whose mean opening energies have low values and vice-versa.
To further assess these two properties, Table 3 shows the mean Z-scores and mean opening energy values in the 30 nucleotide window spanning the first start codon, on the 5'-3' strand. There are some families (*Tospoviridae, Wupedeviridae, Nyamiviridae, Yueviridae, Lispiviridae, Bornaviridae*) whose mean Z-scores in the first CDS region are positive (0.08, 0.18, 0.3, 0.4, 1.03, 1.47 respectively), as opposed to their mean Z-scores across the whole genome, which would imply that the region around the first annotated gene is much more unstructured that the overall structuredness in the entire genome.
The *Sunviridae* family, comprised of only one virus (Reptile sunshine virus) has the smallest mean opening energy value around the first CDS in the ssRNA(-) group (1.2). The mean Z-score value of -0.73 is a marker that the investigated CDS region has a certain structuredness, while the small mean opening energy denotes that little free energy is needed to make unfold the region around the first CDS. The highest mean opening energy value in the first CDS region (8.36 kcal/mol) is found in the *Hantaviridae* family, whose mean Z-score in the said region lies at -0.5. The *Arenaviridae* family, who has the highest overall mean Z-score across the whole genome (as shown in Figure 15), has a higher mean Z-score (-1.04 vs -0.58) in the 30 nucleotide window surrounding the first AUG codon and a mean opening energy of 6.45 kcal/mol.

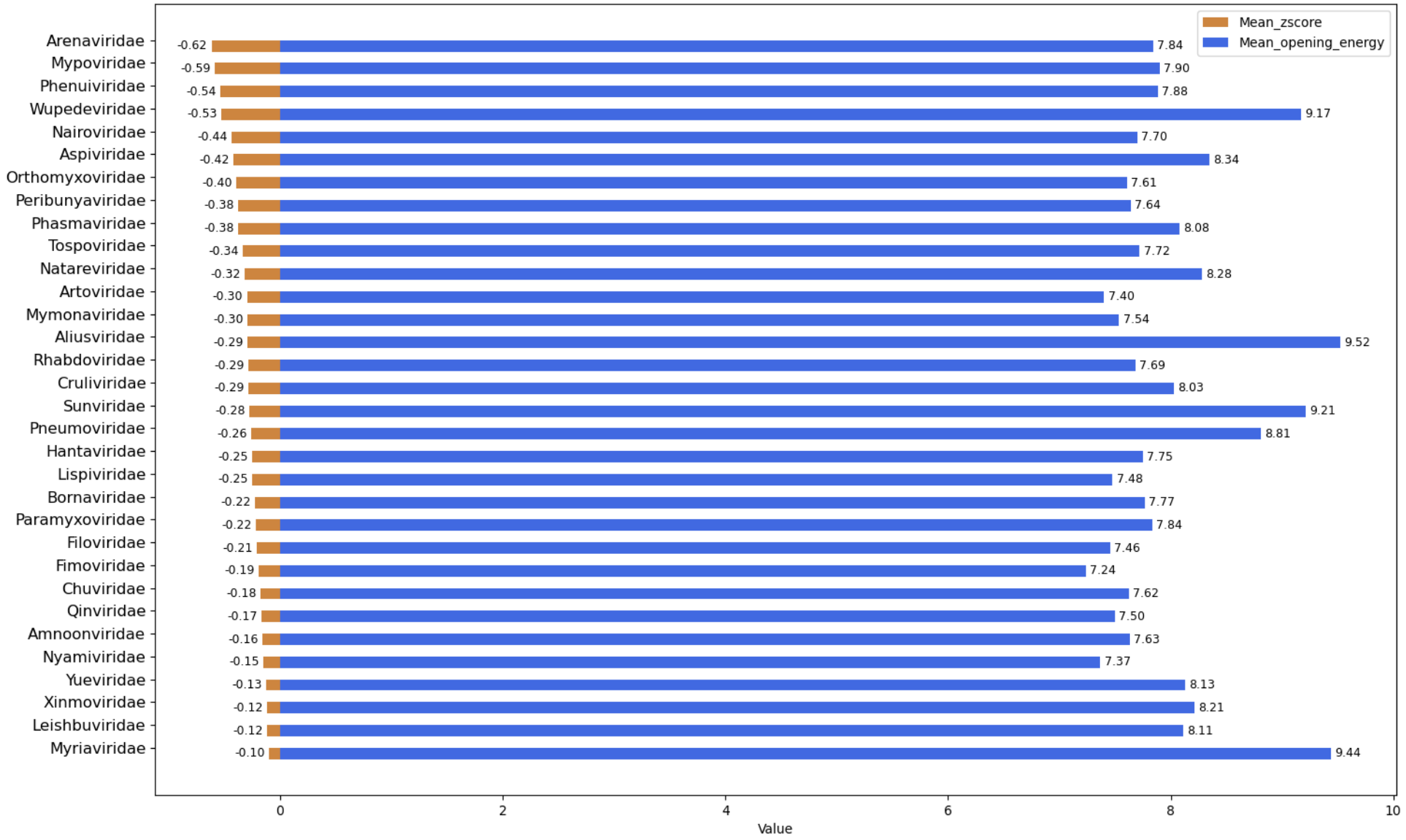

Figure 15: Mean Z-scores and mean opening energy values for the whole genome in ssRNA(-). The mean value of the Z-scores is shown in blue, while the mean value of the opening energies is shown in orange, grouped by taxonomic family

Family	Mean Z-score spanning first CDS	Opening Energy spanning first CDS (in kcal/mol)
Mymonaviridae	-1.661	5.674
Aliusviridae	-1.391	5.179
Arenaviridae	-1.037	6.452
Myriaviridae	-0.911	5.290
Nairoviridae	-0.876	5.846
Phasmaviridae	-0.783	5.809
Peribunyaviridae	-0.782	6.571
Paramyxoviridae	-0.755	7.304
Natareviridae	-0.749	2.675
Sunviridae	-0.727	1.198
Orthomyxoviridae	-0.722	6.593
Filoviridae	-0.648	7.707
Phenuiviridae	-.548	6.434
Rhabdoviridae	-0.504	6.632
Hantaviridae	-0.499	8.357
Leishbuviridae	-0.442	3.948
Chuviridae	-0.427	6.134
Pneumoviridae	-0.417	8.033
Qinviridae	-0.411	4.848
Artoviridae	-0.401	7.091
Cruliviridae	-0.238	7.909
Mypoviridae	-0.028	6.894
Fimoviridae	-0.021	5.484
Xinmoviridae	0.001	5055
Tospoviridae	0.076	6.970
Wupedeviridae	0.175	4.270
Nyamiviridae	0.298	6.119
Yueviridae	0.398	4.834
Lispiviridae	1.032	6.516
Bornaviridae	1.467	7.220

Table 3: Overview of ssRNA(-) families, the mean Z-score and opening energy values spanning the first CDS

Figure 16 shows the correlation plot between the mean Z-score and mean opening energy around the first CDS on the 5'-3' strand for the viruses with single stranded negative genomes. The Pearson's r has the value 0.016, which suggests that there is no relationship between these two properties in this virus group. By looking at each family individually and assessing the relationship between the mean Z-scores and energy values in the entire genome (Supplementary Figure 53), it becomes clear that there is no correlation between these two parameters. Apart from the *Bornaviridae* family, in which there could be a negative trend between mean Z-scores and free opening energy values, in no other family can this observation be made.

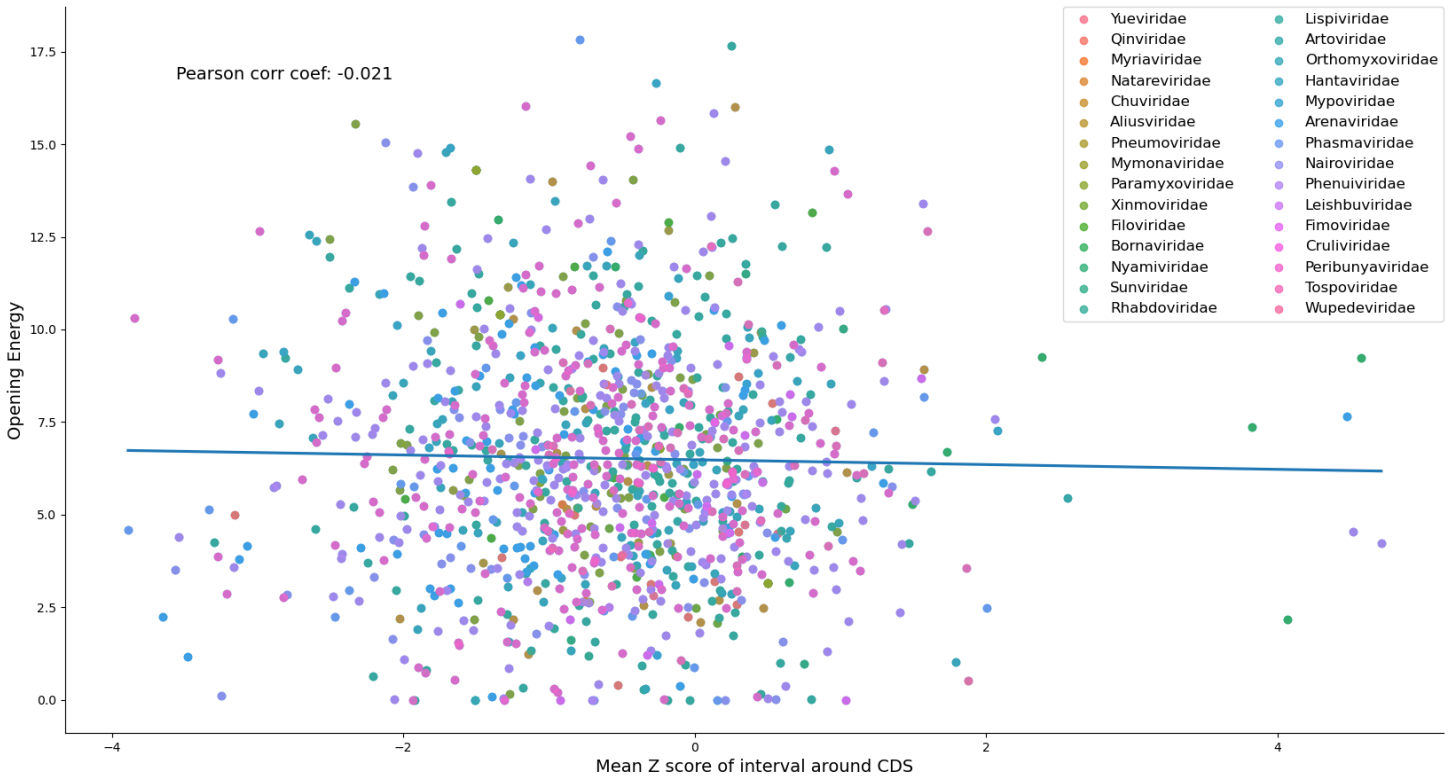

Figure 16: Correlation of the mean Z-scores and mean opening energy values in the region of the first CDS in ssRNA(-). Each data point depicts a species, grouped by taxonomic family (see legend)

5.4.3 Mean Z-scores and mean opening energy values in dsRNA

Further, the 5'-3' genome strand of the dsRNA viruses was analysed using the same approach as described previously in this section. Figure 17 shows the mean opening energy and the mean Z-score for the whole genome of the dsRNA viruses, categorized by family. The highest mean opening energy (11.52 kcal/mol) was determined for the *Cystoviridae* family, whose mean Z-score is also one of the most negative in this group (-0.87). The *Reoviridae* family has the lowest mean opening energy and also the highest mean Z-score overall. Next, to better understand how the mean Z-score and opening energy behave around the first CDS region, Figure 18 shows the result of the two computed parameters for that specific area.
Interestingly, three of the dsRNA families which have negative mean Z-scores overall the genome, have positive mean Z-scores in the 30 nucleotide window spanning the first CDS. The *Quadriviridae, Curvulaviridae* and *Amalgaviridae* familiy, shown in Figure 18, have a mean Z-score of 0.96, 0.79 and 0.15 in the region of the first CDS, respectively. The mean opening energies for these families in the same region are the lowest in the group, with values ranging from 4 to 7 kcal/mol. The *Reoviridae* family, whose mean Z-score across the 5'-3' strand has a positive value of 0.10, has now an even higher positive mean Z-score around the first CDS (0.25). This high mean Z-score value is met with a surprisingly high mean opening energy value, considering the overall values in this group (7.2kcal/mol). The highest mean opening energy value is attributed to the *Picornaviridae* family (8.36 kcal/mol), whose mean Z-score is also the most negative in this case (-1.53).

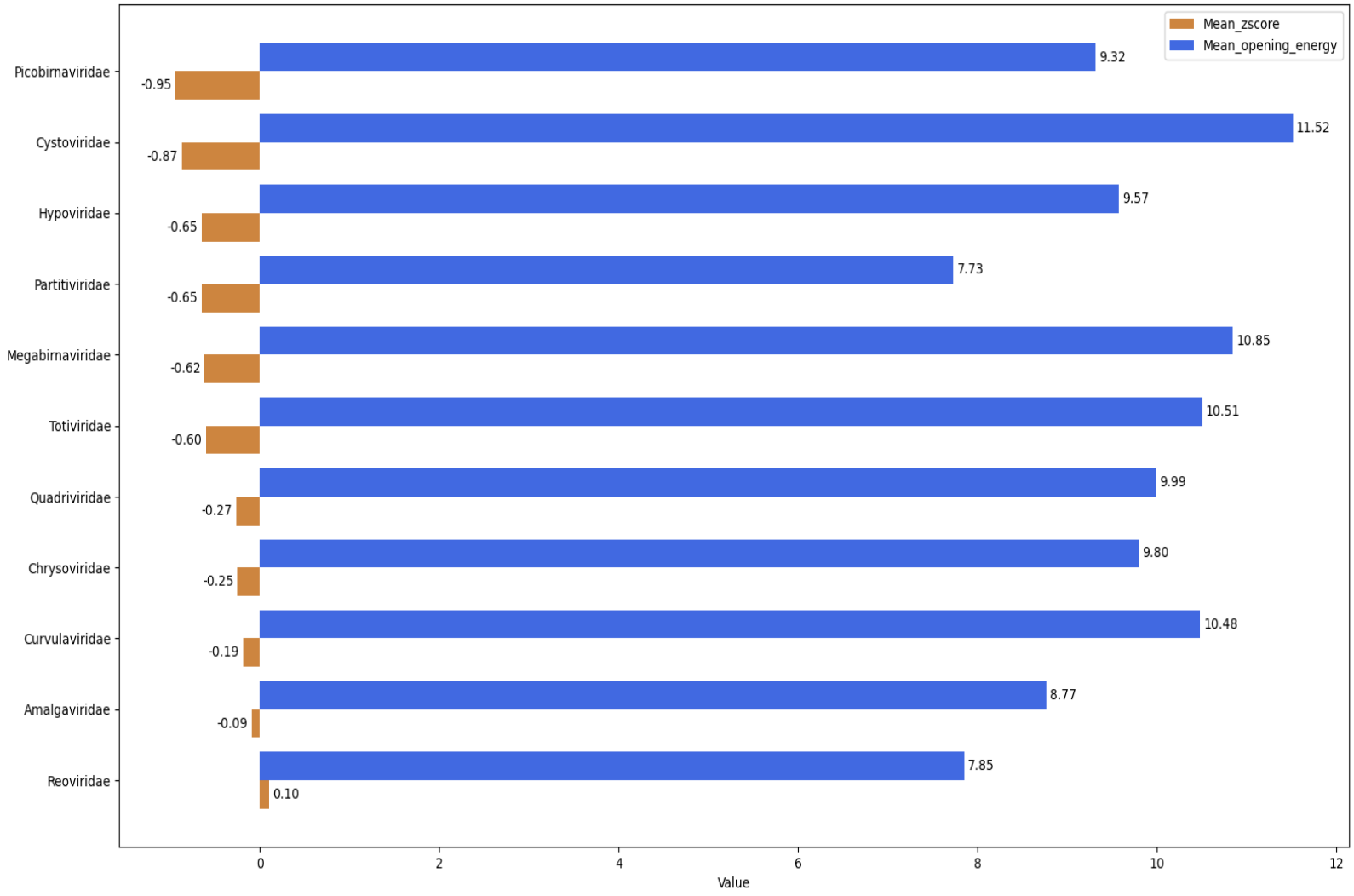

Figure 17: Mean Z-scores and mean opening energy values for the whole genome in dsRNA, 5'-3' strand. The mean value of the Z-scores is shown in blue, while the mean value of the opening energies is shown in orange, grouped by taxonomic family

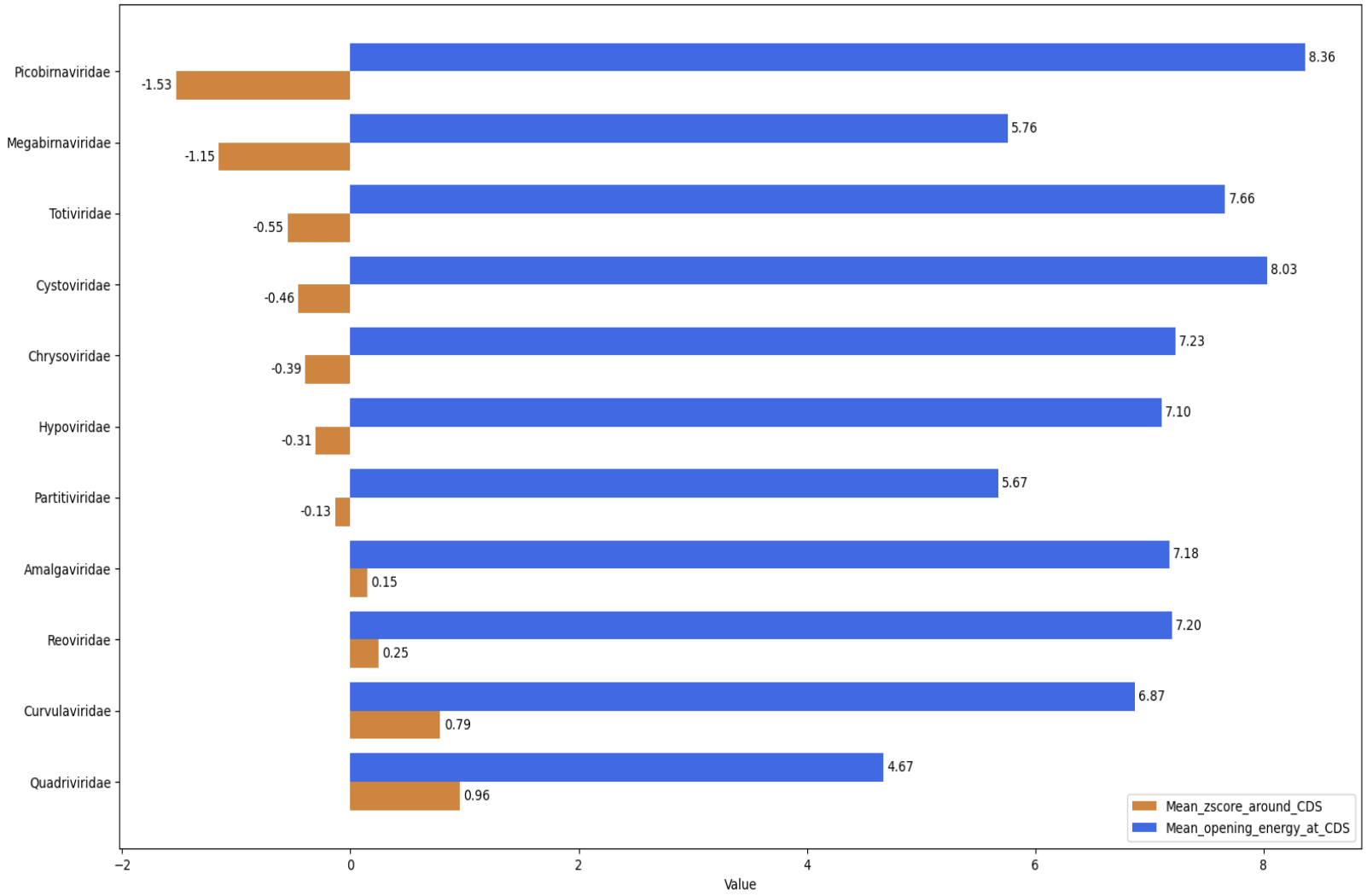

Figure 18: Mean Z-scores and mean opening energy values for the region around the first CDS in dsRNA, 5'-3' strand. The mean value of the Z-scores is shown in blue, while the mean value of the opening energies is shown in orange, grouped by taxonomic family

To determine if there is a correlation between the mean opening energy and the mean Z-score in the region of the first AUG start codon in the dsRNA group, Figure 19 shows the correlation plot between these two paramteres for all the analyzed genomes in this group. Indeed, it appears that the opening energy and Z-scores tend to correlate in the dsRNA viruses. The correlation between the opening energy and the mean Z-score in the region of the first CDS for the dsRNA viruses has a Pearson's r value of -0.348. This hints to a global trend, as in the case of ssRNA(+) group, that the binding region for the ribosome is thermodynamically accessible on the mRNA. This facilitates the gene expression at least in the first coding region and might have important implications for the viral replication and survival in the host.
By looking at each family individually (Supplementary Figure 54), it can be observed that in *Curvulaviridae* and *Quadriviridae*, containing in total three viruses each, the opening energies in all segments have higher values when the mean Z-scores are positive. This would mean that the are high amounts of free energy needed to unfold even an unstructured sequence in their genomes, which is paradoxical. The three small segmented dsRNA viruses, Fusarium graminearum mycovirus 4 and Trichoderma harzianum bipartite mycovirus 1 from the Orthocurvulavirus genus, as well as Rosellinia necatrix quadrivirus 1 from the Quadrivirus genus, are segmented viruses that infect fungi.

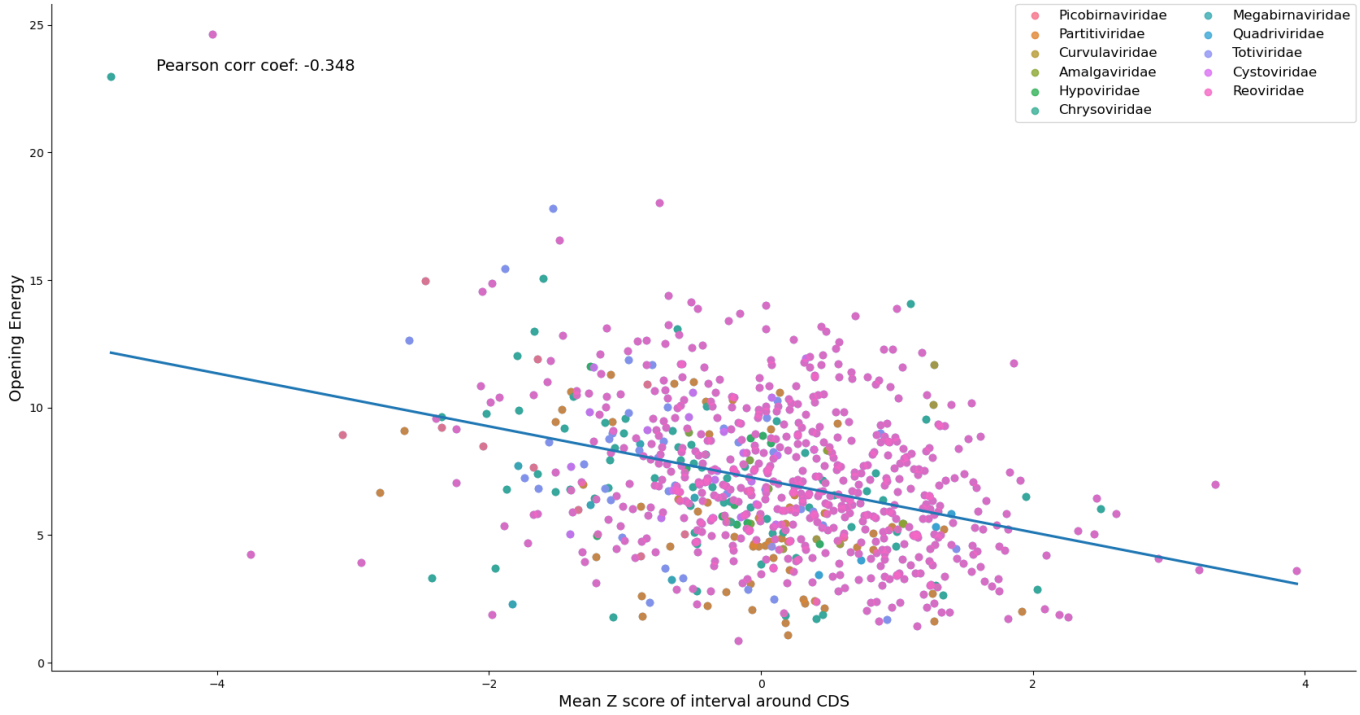

Figure 19: Correlation plot between mean Z-scores and mean opening energy values for the region around the first CDS in dsRNA, 5'-3' strand. Each data point depicts a species, grouped by taxonomic family (see legend)

5.4.4 Correlation of the mean Z-scores and mean opening energy values in the window following the first AUG codon

In order to determine if there is a correlation between the mean Z-score and mean opening energy in other regions in the genomes, the next 30 nucleotide window following the first CDS region was selected. The correlation plots are seen in Figure 20 for the ssRNA(+), Figure 21 for the ssRNA(-) and Figure 22 for the dsRNA virus groups. The correlation coefficient improves for the ssRNA(+) and dsRNA(+), while for the ssRNA(-) the Pearson'r is almost 0 in this case, which means that there is no correlation at all between the two parameters. By comparison with the ssRNA(-) viruses, this means that the gene expression in ssRNA(+) is qualitatively better, because the ribosome access to a binding site on the mRNA is well facilitated, at least in the context of the free energy of the RNA-RNA interaction. This may be explained by the fact that, for the majority of the viruses, these further 30 nucleotides are still part of a coding region, as their CDS are usually longer than 45 nucleotides. The fact that the correlation gets better inside the first CDS, would mean that there is a high translational drive of the first gene in the single stranded positive sense viruses. However, the correlation between mean Z-scores and opening energies for the whole genome is different when looking at each family individually, even for the ssRNA(+) viruses (Supplementary Figure 52). For example, in the *Virgaviridae* and *Alphaflexiviridae* families, there is seems to be no correlation between the values of the Z-scores and the opening energies. The viruses in both families infect plants and/or fungi. On the other hand, the *Coronaviridae* family shows a negative relationship between the mean Z-scores and opening energies in their genomes. This implies that the thermodynamic properties of their genomes not only allow the formation of many stable secondary structures (as previously described in literature), but also facilitate a good translation of the encoded proteins, denoted by the tight relationship between the opening energy values and the mean Z-scores.

For the ssRNA(-) viruses, the analysis of the mean energy value in the 30 nucleotide window spanning the first CDS showed no correlation with the mean Z-score of the same region. This comes as a surprise, as for both the other two Baltimore groups analyzed, there seemed to be at least a global negative trend between the two parameters.

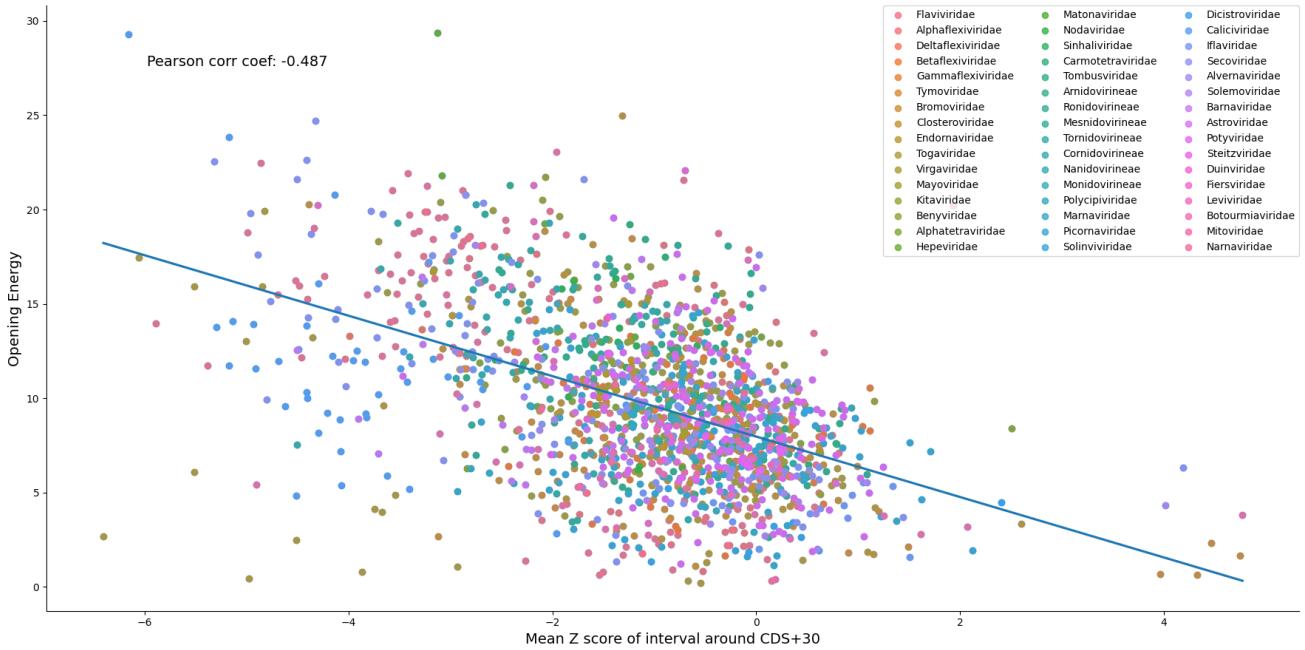

Figure 20: Correlation plot between mean Z-scores and mean opening energy values in the region following the first CDS in ssRNA(+). Each data point depicts a virus genome, grouped by taxonomic family (see legend)

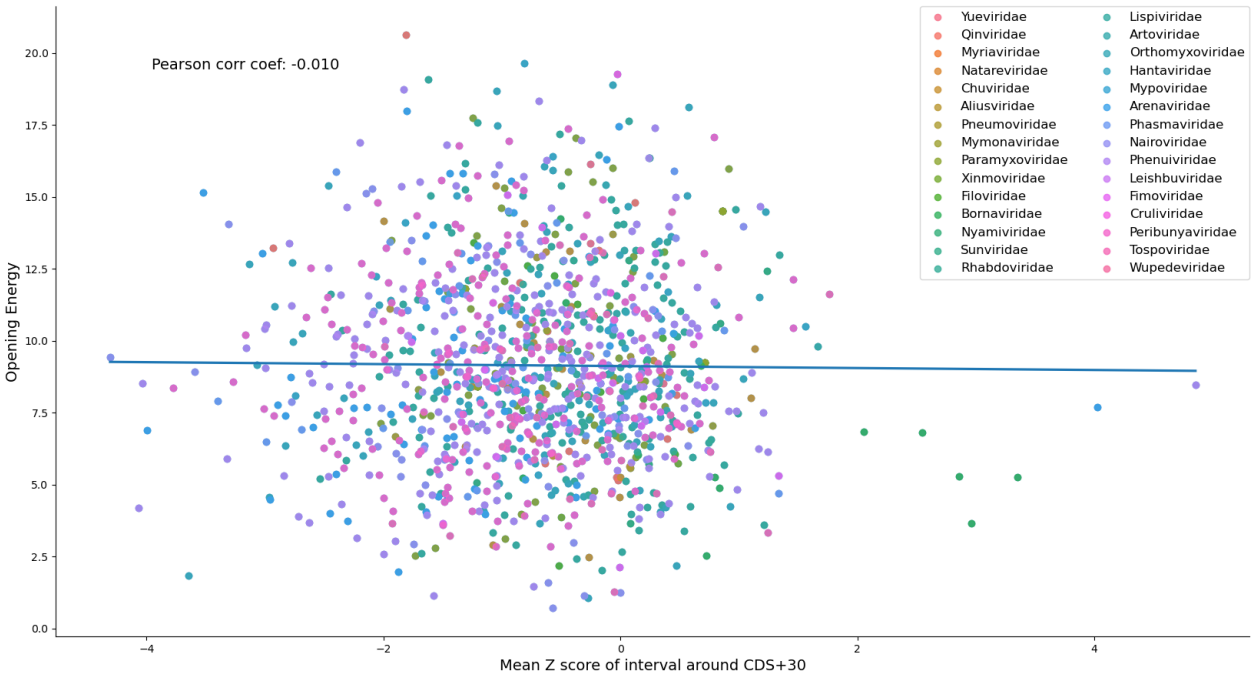

Figure 21: Correlation plot between mean Z-scores and mean opening energy values in the region following the first CDS in ssRNA(-). Each data point depicts a virus genome, grouped by taxonomic family (see legend)

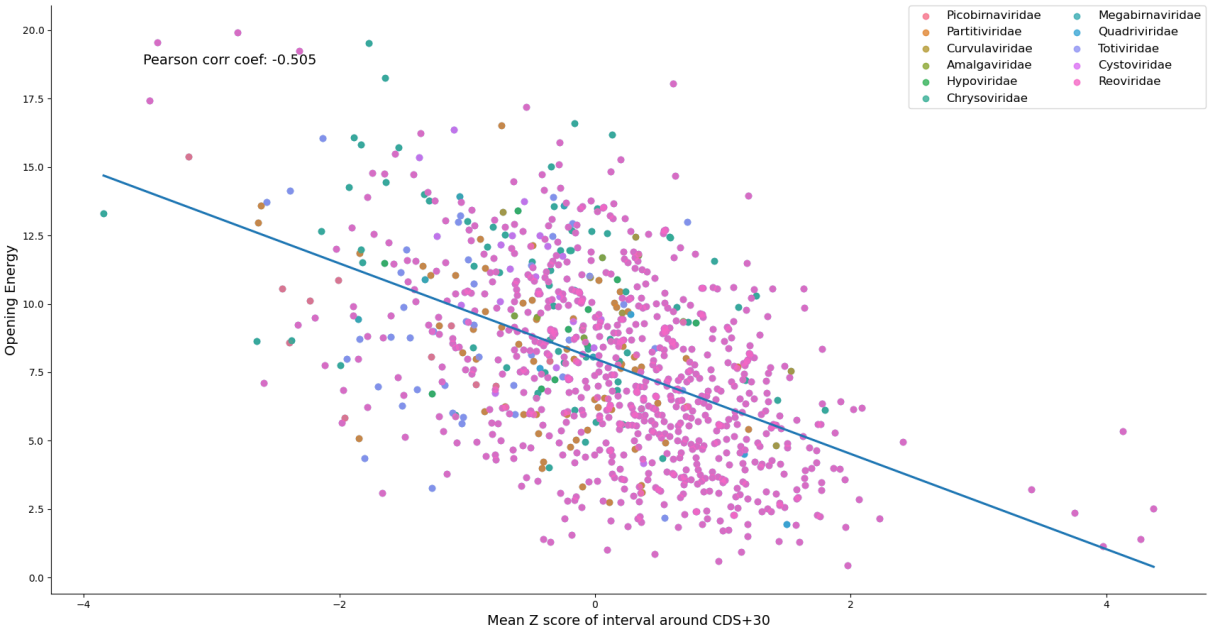

Figure 22: Correlation plot between mean Z-scores and mean opening energy values in the region following the first CDS in dsRNA. Each data point depicts a virus genome, grouped by taxonomic family (see legend)

5.5 Minimum Free Energy Difference

The minimum free energy difference (MFED) values were calculated from the RNAfold module as described in the Material and Methods part. Briefly, the MFED are the result of subtracting the mean free energy of the background model from the native RNA. Since the Z-score approach is the core method used in this work, it was interesting to see if there is a correlation between the Z-scores and MFED values in all Baltimore groups. The difference between the Z-score and the MFED values is solely the fact that the MFED is not divided by the standard deviation from the model, hence it is expected that the correlation between the two parameters is strong. Figures 23, 25 and 24 indicate the Pearson correlation coefficient for the ssRNA(+), as well as dsRNA and ssRNA(-) on both the 5'-3' and 3'-5' strand.
The Pearson's r in the dsRNA group is 0.959 for the forward and 0.938 for the backward strand, denoting a strong linear relationship between the Z-score and the MFED values. For the ssRNA(-) viruses, the correlation coefficient has the value 0.864 on the 5'-3' strand and 0.832 on the 3'-5' strand, which further strengthens the relationship between the two parameters. In the case of the ssRNA(+) viruses, the Pearson correlation coefficient has the value 0.926. These results confirm that indeed there is an indubitably strong linear relationship between the Z-score and the MFED values, regardless of genome structure and type of virus.

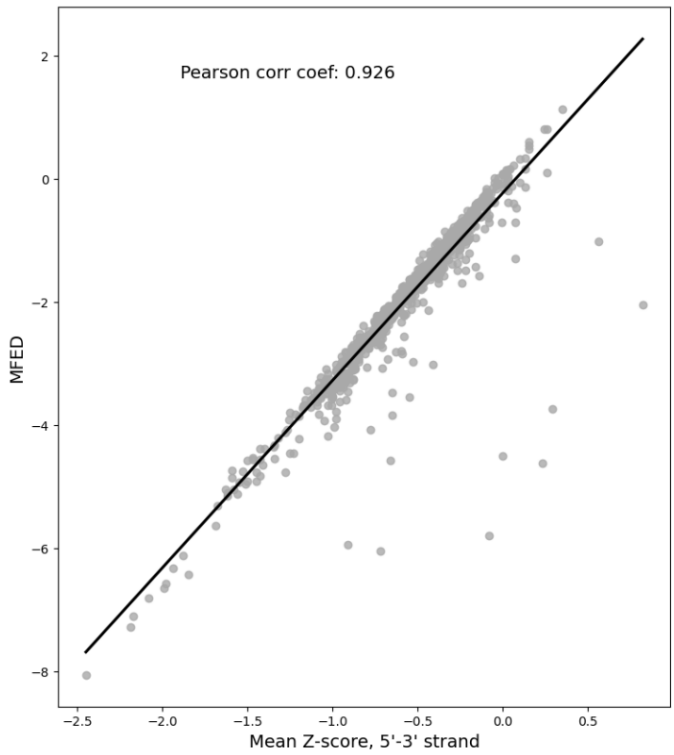

Figure 23: Correlation plot between Z-scores and MFED values in ssRNA(+). Each data point depicts a virus.

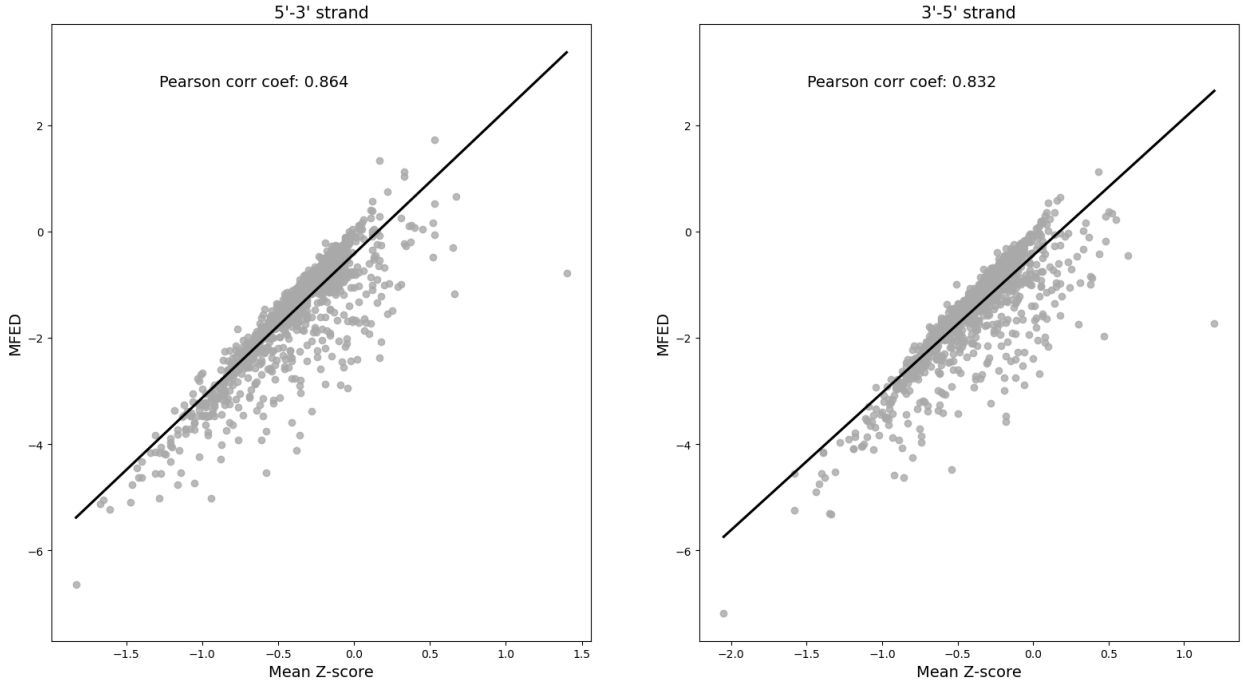

Figure 24: Correlation plot between Z-scores and MFED values in ssRNA(-) for both genome strands. Each data point depicts a virus or segment of a virus.

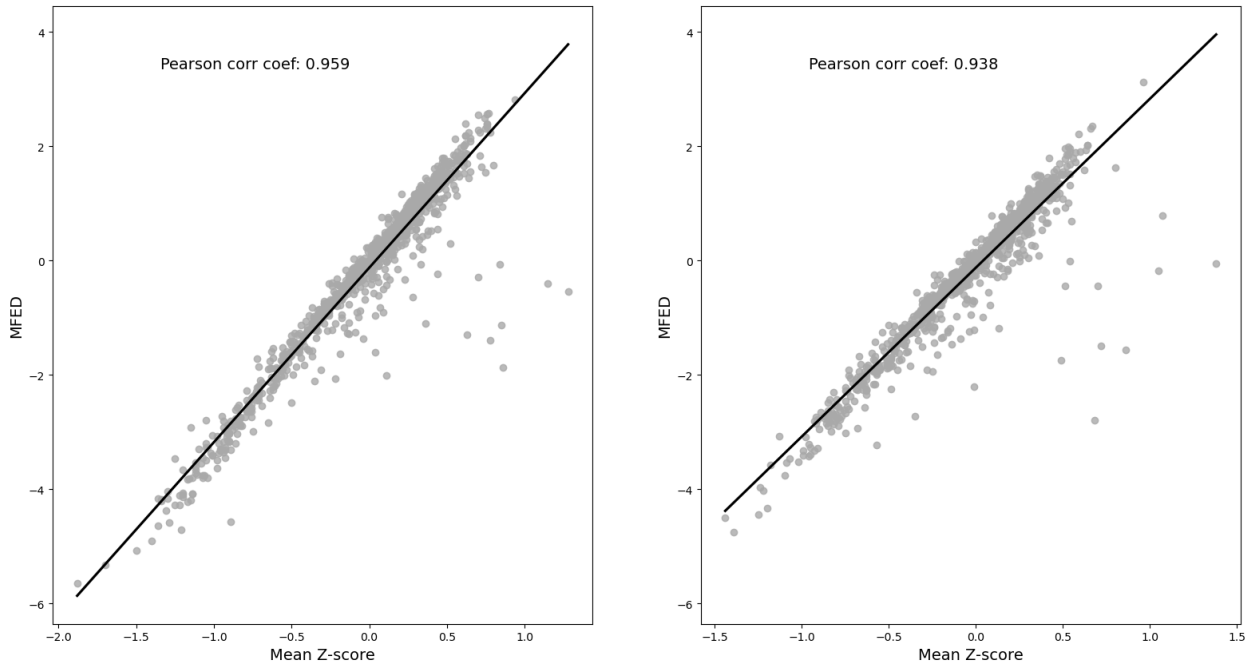

Figure 25: Correlation plot between Z-scores and MFED values in dsRNA. Each data point depicts a virus or segment of a virus.

5.6 Genus and species level analysis

To further assess the structuredness of the different viruses in the Baltimore groups, a complex analysis pipeline was built to pre- and post-process, visualize and annotate all the genomes belonging to every distinct viruses in each group. The analysis pipeline involved statistical tests, genome annotation, differentiation between coding and non-coding regions as well as thermodynamic predictions results from different ViennaRNA modules, integrated as .bash scripts in the Python pipeline. One of the modules used was RNALfold, which predicts the locally stable secondary structures in a given nucleotide sequence. Here, the output from RNALfold was processed to achieve an intuitive visualization of the location, length and confidence level of the predictions, by filtering for only those structures whose Z-score $\leq$ -2. The details of the implementation are described more thoroughly in the Material and Methods section. The majority of the analysis has been done on virus level, but since it was important to determine how the different organisms behave from a thermodynamic point of view inside of a taxonomic genus, the analysis was extended to include different features in a given genus, such as density plots and box plots for the Z-scores for all the genomes included in that particular genus. Also, the distribution of the opening energies around relative to the position of the first start codon in the CDS of each virus in a genus was visualized.
In this section, for each of the ssRNA(+), ssRNA(-) and dsRNA virus groups, the results for one genus and one virus belonging to that genus are described, serving as example for how the pipeline works and how the genomic data is processed and visualized. All results are shown for a sliding-window of length 100.

5.6.1 Genus and level species in ssRNA(+)

The Hepacivirus genus was chosen as an example for the analysis of a ssRNA(+) genus. Hepaciviruses belong to the *Flaviviridae* family in the Amarillovirales taxonomic order. There are 27 viruses in the Hepacivirus genus which are known to cause chronic and acute hepatitis, carcinoma and liver function failure not only in humans, but also in rodents, cows, bats, horses and possums, to name a few [75]. Their genomes are approximately 9500 nucleotides long and encode for a polyprotein, which is co- and posttranslationally cleaved into individual viral proteins [43]. The Hepatitis C virus (HCV), maybe the best known virus in the Hepacivirus genus, is a human pathogen which infects roughly 3% of the global population and causes chronic hepatitis in children (most frequently via vertical transmission) and adults [17]. In order to characterize this genus from a thermodynamic point of view, the density function and empirical cumulative density function (ECDF) of the Z-scores in all the hepaciviruses were determined, as well as the distribution of Z-scores per position (Figure 26). The density function shows that in almost all organisms, the Z-scores seem o follow a normal distribution, denoted by the bell-shaped lines in the upper Figure 26. The distribution of the Z-scores per position denotes multiple structured regions in the genomes and, in only a few hepaciviruses, short regions where the mean Z-scores have positive values. These results align nicely with the previous results denoting a high level of structuredness in the genomes of the ssRNA(+) viruses.

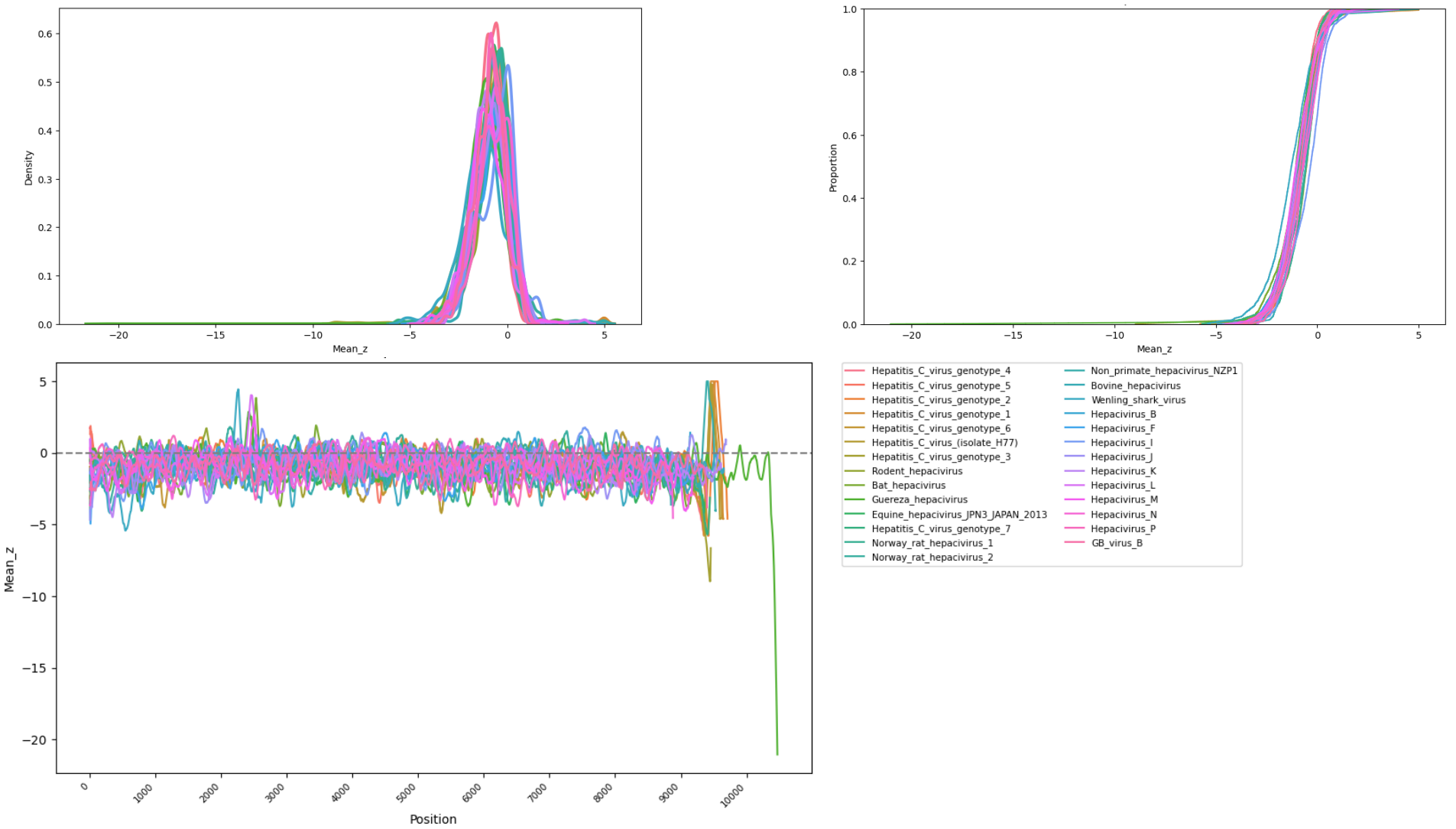

Figure 26: Genus level analysis in Hepacivirus, 5'-3' strand. Upper left: density function of all Z-scores per position per species; upper right: ECDF of all Z-scores per position species; lower: Z-scores vs genomic position. Each line denotes a different species in the Hepacivirus genus

Next, the HCV genotype 3 is shown in Figure 27 as an example for the virus level analysis in ssRNA(+). The HCV is classified into six different genotypes, based on its genetic material. Here, genotype 3 was chosen, because it is known to have a worldwide distribution and it is predominant in Europe, Southeast Asia, India, Australia and the United States of America [8]. The genome annotation, together with the locally stable predicted structures from RNALfold and the mean Z-scores per position are depicted in Figure 27. Throughout the genome, the mean Z-scores maintain a negative value, denoting the genome's global ability to form stable secondary structures. This trend is further supported by the abundance of locally stable structures predicted by the RNALfold module. Especially the 5'-UTR and 3'-UTR regions seem to contain at least one stable structure each, implying the presence of secondary structures flanking the CDS. Indeed, one can observe that in the 3' end of the genome is the location of the most negative Z-score peak, with a value of around -8.

Since both the CDS and non-CDS seem to contain Z-scores that are predominantly negative, it was interesting to determine if the distribution of the mean Z-scores in both regions are alike and whether there is any statistical difference between them. Figure 28 shows the distribution of the Z-scores in CDS and non-CDS for the HCV genotype 3 virus. It is possible to observe that the distribution of the mean Z-scores in non-CDS follows a bimodal distribution, where the predominant peak is around -1 and the second peak is around -8. On the other hand, the tail of the distribution in CDS (in blue) seems to follow a normal distribution, with a peak at approximately -1. The Shapiro-Wilkinson test was used to determine if the data in the underlying distribution is drawn from a normal distribution. The test determined that neither of them follows a normal distribution (CDS: Shapiro-Wilk test statistic=0.97, p-value $<< 0.05$; non-CDS: Shapiro-Wilk test statistic=0.89, p-value $<< 0.05$). Both distributions are slightly left-skewed, denoting a global trend that the vast majority of the genomic positions in the HCV genotype 3 genome have negative mean Z-score values. This observation aligns with the structuredness predictions from RNALfold seen in Figure 27.

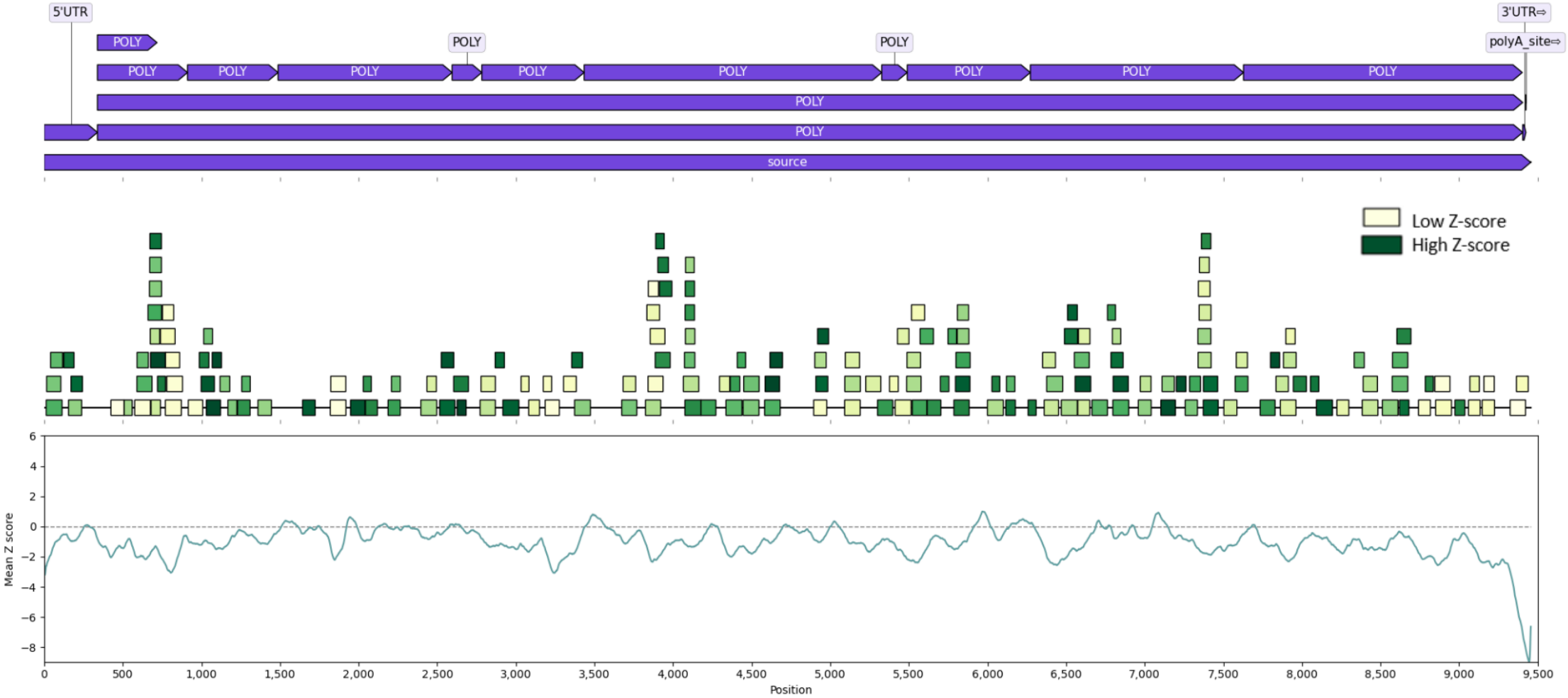

Figure 27: Species level analysis for Hepatitis virus C genotype 3, 5'-3' strand. The rectangles are the predicted RNALfold stable structures, their width is proportional to the length of the structure. All the predictions have a Z-score $\leq$ -2. NCBI accession: `NC_009824`

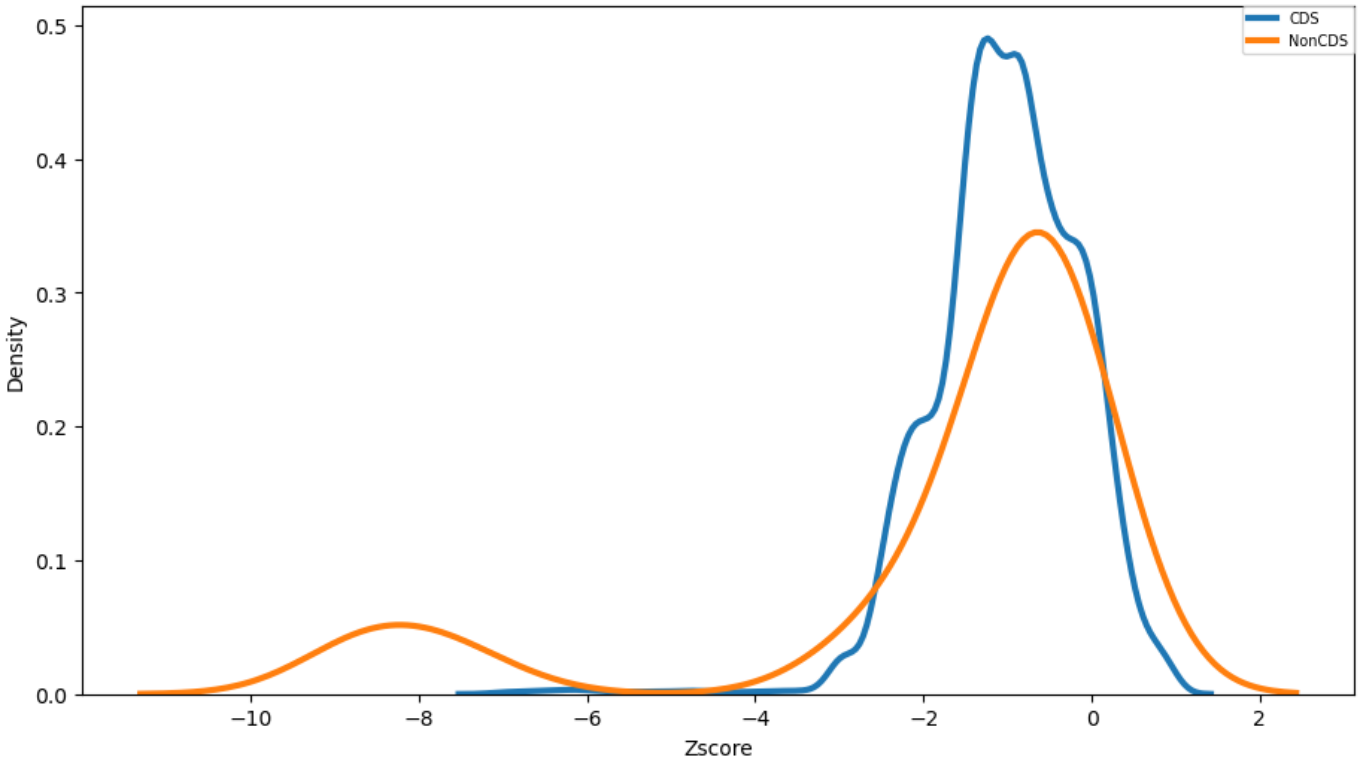

Figure 28: Distribution of mean Z-scores in CDS and non-CDS region of the HCV genotype 3 virus. Blue: CDS Z-scores, orange: non-CDS Z-scores

To facilitate the analysis of the whole Hepacivirus genus, Figure 30 shows the distribution of the mean Z-scores as boxplots, differentiated by coding or non-coding region for all viruses belonging to this genus. There is no global trend for either CDS or non-CDS to be more structured. The Mann-Whitney test (MWU) was used to determine if the underlying distribution of the mean Z-scores in the two regions are significantly different. This test was used because not all species follow a normal distribution of the Z-scores in the CDS or non-CDS regions (data not shown). For ten of the viruses, such as Hepacivirus B or Hepatitis virus C genotype 7, the distribution of the Z-scores are not significantly different between the CDS and non-CDS, while for thirteen others the distribution of the mean Z-scores in CDS are significantly different than the mean Z-scores per position in the non-coding regions.

In order to have an overview of the distribution of the opening energy around the position of the first AUG start codon in each first CDS, firstly, for each virus in the Hepacigenus the location of the first start codon was extracted. The corresponding value of the opening energy was computed using RNAplfold 30 positions before and after the first AUG; thus, position 0 (in red) is denoting the start codon in the first CDS, and the x-axis labels are the relative positions to the first AUG. Figure 29 shows, in form of boxplots, the distribution of the opening energy values relative to the start codon. For a better overview of how the opening energy behaves in the nucleotide regions before and after the first AUG, a span of 60 nucleotides was taken into account, as to show the distribution of the energy values for a window of 60 nucleotides for which the start codon is located in the middle. In the positions before the start codon, the median of the energy values lies between 10 and 15 kcal/mol, however the value drops once the CDS is reached, starting from position 0 to 29. Taking into account the global structuredness of the hepaciviruses (Figure 30) and

the negative correlation between the mean Z-scores and the opening energy across the whole genome for the *Flaviviridae* taxonomic family (Supplementary Figure 52), where the hepaciviruses belong, one possible interpretation is that the unfolding of the CDS takes place successfully before the location of the CDS. This suggests that it is likely that the ribosome binding is thermodynamic feasible for the first coding region. However, since it is expect that the CDS of the hepaciviruses is also well folded, it is curious that the median value of the opening energy lies between 0 and 5 kcal/mol once the CDS begins. This would hint towards a qualitatively low RNA-RNA interaction of the genome with the ribosome for facilitating gene expression, at least in the context of the first CDS in the genomes of the organisms in the Hepacivirus genus.

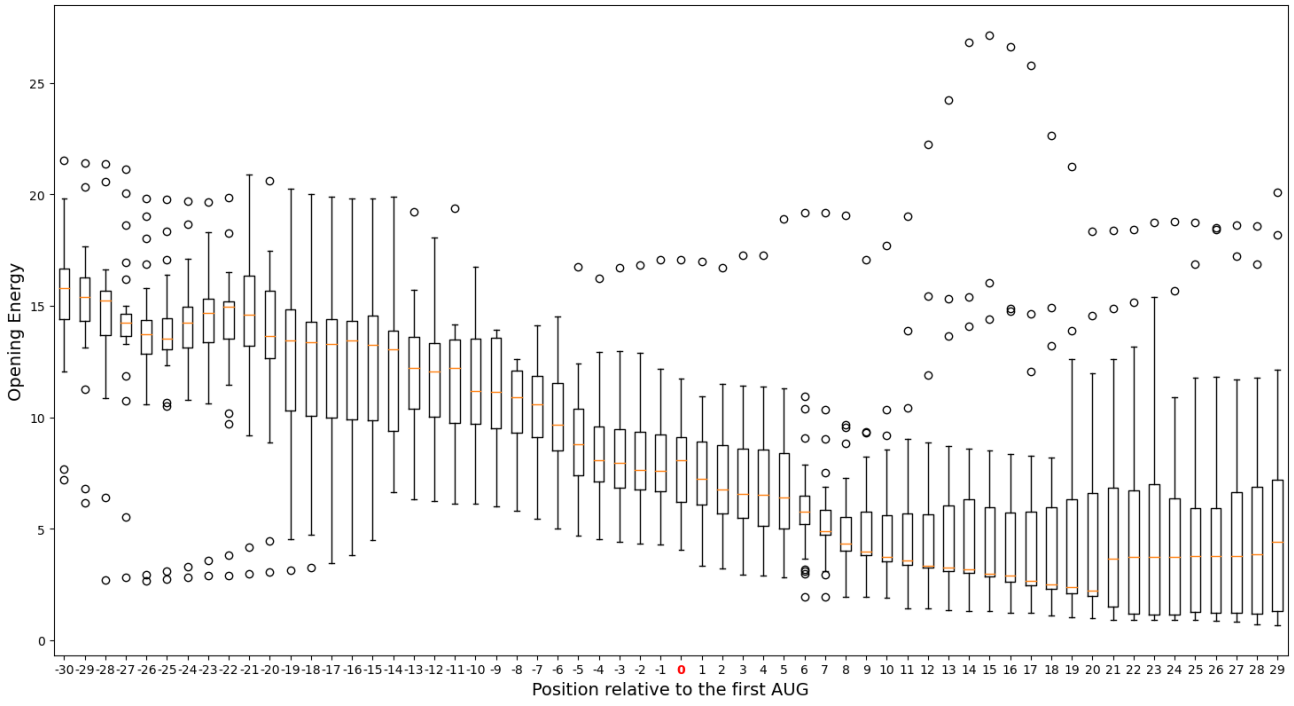

Figure 29: Distribution of opening energy values for the 60 nucleotide window spanning the first start codon in the first CDS of hepaciviruses

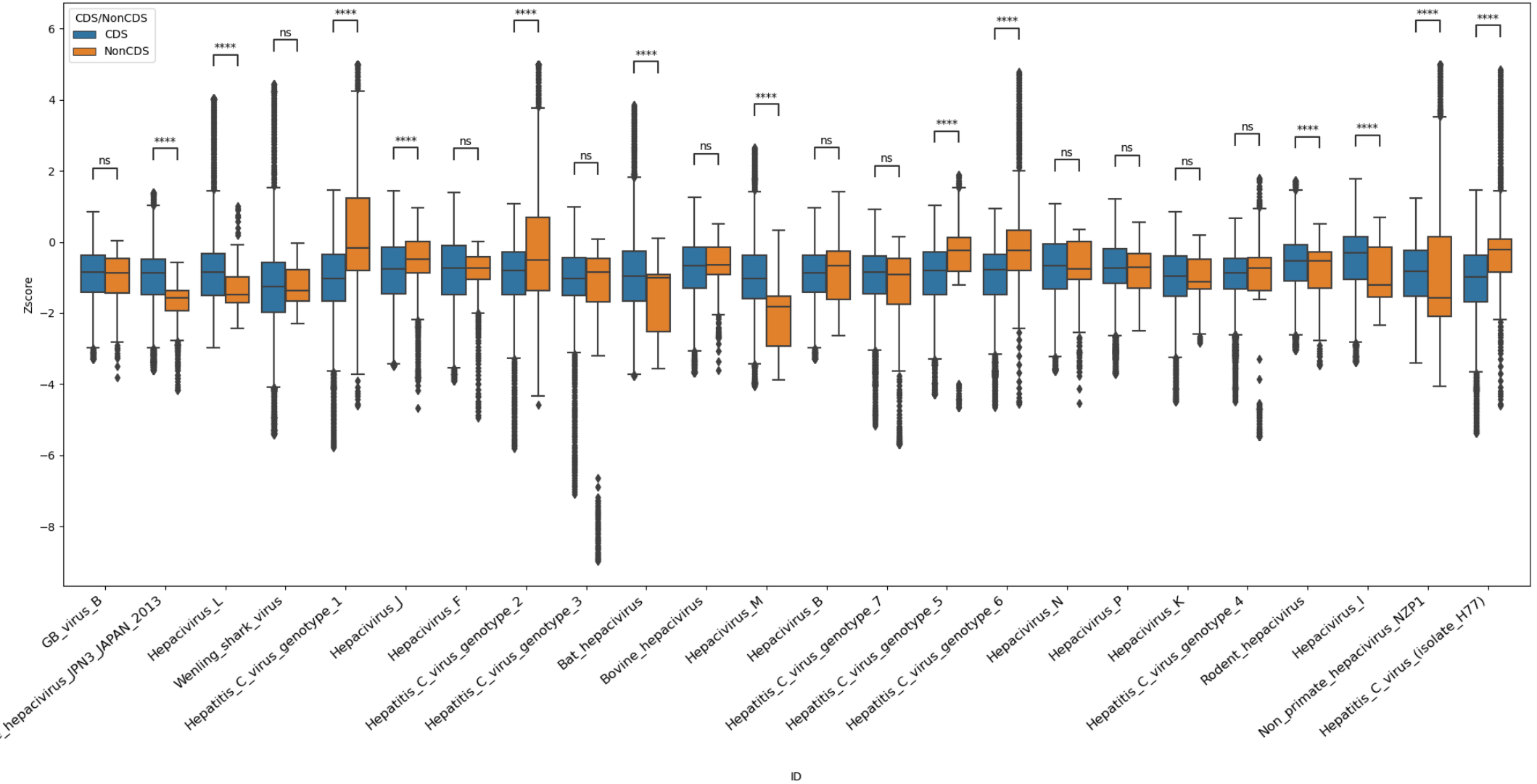

Figure 30: Boxplots of Z-score values in CDS and non-CDS regions for all Hepacivirus. Blue: CDS Z-scores, orange: non-CDS Z-scores. NCBI accessions (preserving the order in the figure): NC_001655.1, NC_024889.1, NC_031916.1, NC_028377.1, NC_004102.1, NC_038429.1, NC_038427.1, NC_009823.1, NC_009824.1, NC_031947.1, NC_026797.1, NC_038430.1, NC_038426.1, NC_030791.1, NC_009826.1, NC_009827.1, NC_038432.1, NC_040815.1, NC_038430.1, NC_009825.1, NC_021153.1, NC_038428.1, NC_038425.1, NC_038882.1; p-value annotation legend: ns: 5.00e-02 $\leq$ 1.00e+00; *: 1.00e-02 $\leq$ p $\leq$ 5.00e-02; **: 1.00e-03 $\leq$ p $\leq$ 1.00e-02; ***: 1.00e-04 $\leq$ p $\leq$ 1.00e-03; ****: p $\leq$ 1.00e-04

5.6.2 Genus and species level analysis in ssRNA(-)

The Ebolavirus genus was chosen as example for single stranded negative sense virus genus. The ebolaviruses belong to the *Filoviridae* family, in the Mononegavirales taxonomic order. There are six viruses in the Ebolavirus genus: Sudan ebolavirus, Tai Forest ebolavirus, Reston ebolavirus, Zaire ebolavirus, Bundibugyo ebolavirus and Bombali ebolavirus. Their genomes encode for seven genes, in the order: 3'-NP-VP35-VP40-GP-VP30-VP24-L-5'. Their genome lengths are around 19000 nucleotides and they are unsegmented viruses, whose lethality upon infection is between 50 and 90% in both humans and non-humans hosts [64]. The first analysis was to determine the density function and ECDF of the Z-scores in these viruses, as well as the visualization of the Z-scores per position for all the genomes, on both strands (Figures 31 and 32).

On the 5'-3' genome strand, the distribution of the Z-scores per genome position seem to be normally distributed, as seen in upper part of Figure 31. As depicted in the lower part of the figure, the Z-scores per position are fluctuating around 0, which is also denoted in the ECDF plot, as about a half of all the Z-scores are below 0. There are some regions where the value of the Z-scores are positive in all viruses, for example around position 7500. The 5'-UTR region seems to be structured in these viruses, as all the Z-scores are negative around the first positions of the genome. Overall, the 5'-3' strand does not seem to be particularly structured, as the Z-scores are not consistently negative across the whole genome. There are certain regions in all viruses in which the Z-scores values get as low as -3, but as shown in Figure 31, this is not consistent across the whole genome.

The same analysis was conducted for the 3'-5' strand and the results can be seen below in Figure 32. The distribution of the Z-scores on the 3'-5' seems to be similar to the 5'-3' strand. Additionally, the ECDF plots of the ebolaviruses show no notable differences between the two strands. The lower plot in Figure 32 shows the Z-score per genome position. Here, there are also some regions where the Z-scores have positive values, for example around positions 11000 and 16000. There is a consistently structured region in the 5'-UTR, as all the Z-scores take values around -2.

By visually comparing the Z-scores per position on both strands, is would seem that the Z-score values do not vary much, which is consistent with the results presented in Figure 5, where it can be seen that across the ssRNA(-) genome strands, the values of the mean Z-scores are very close.

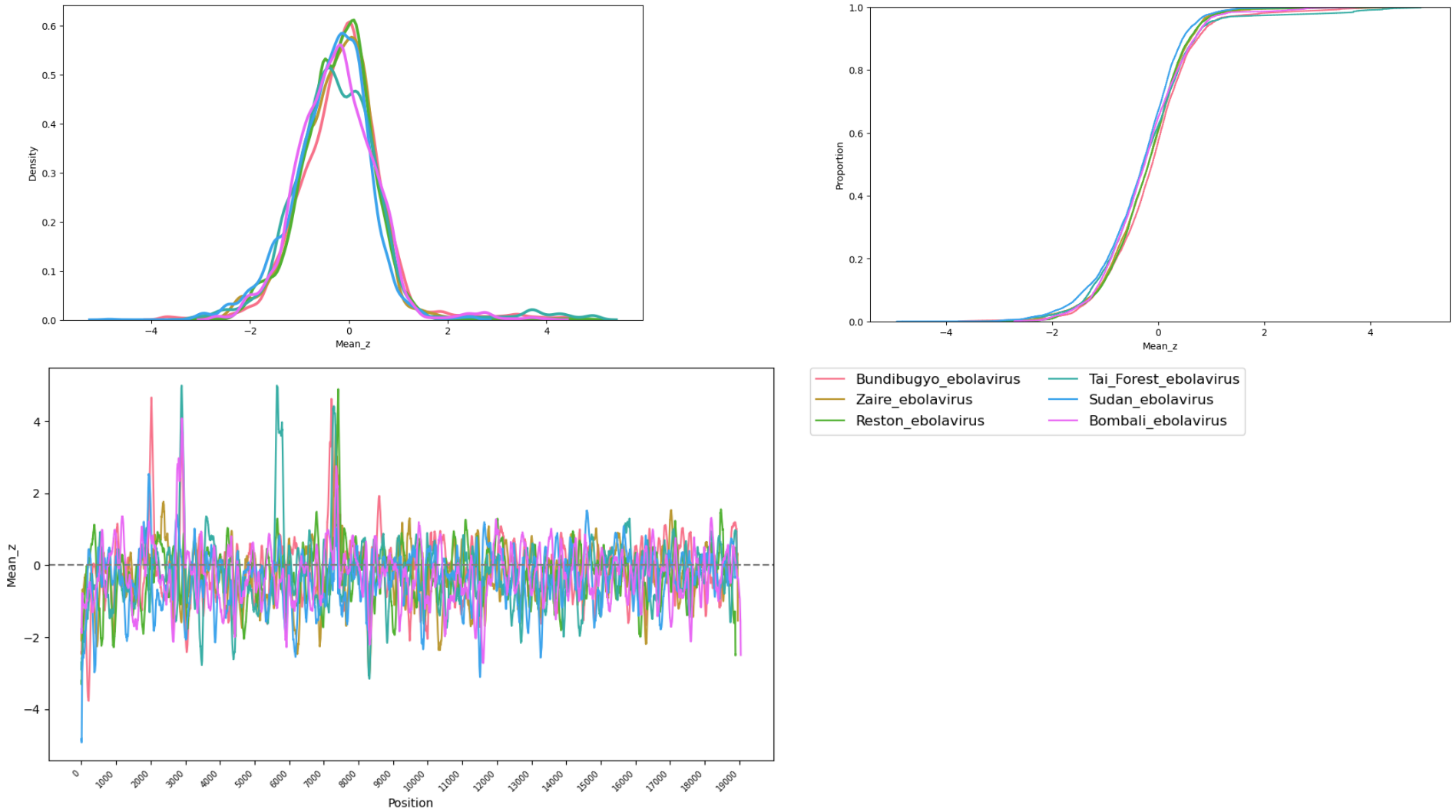

Figure 31: Genus level analysis in Ebolavirus, 5'-3' strand. Upper left: density function of all Z-scores per position per species; upper right: ECDF of all Z-scores per position species; lower: Z-scores vs genomic position. Each line denotes a different species in the Ebolavirus genus

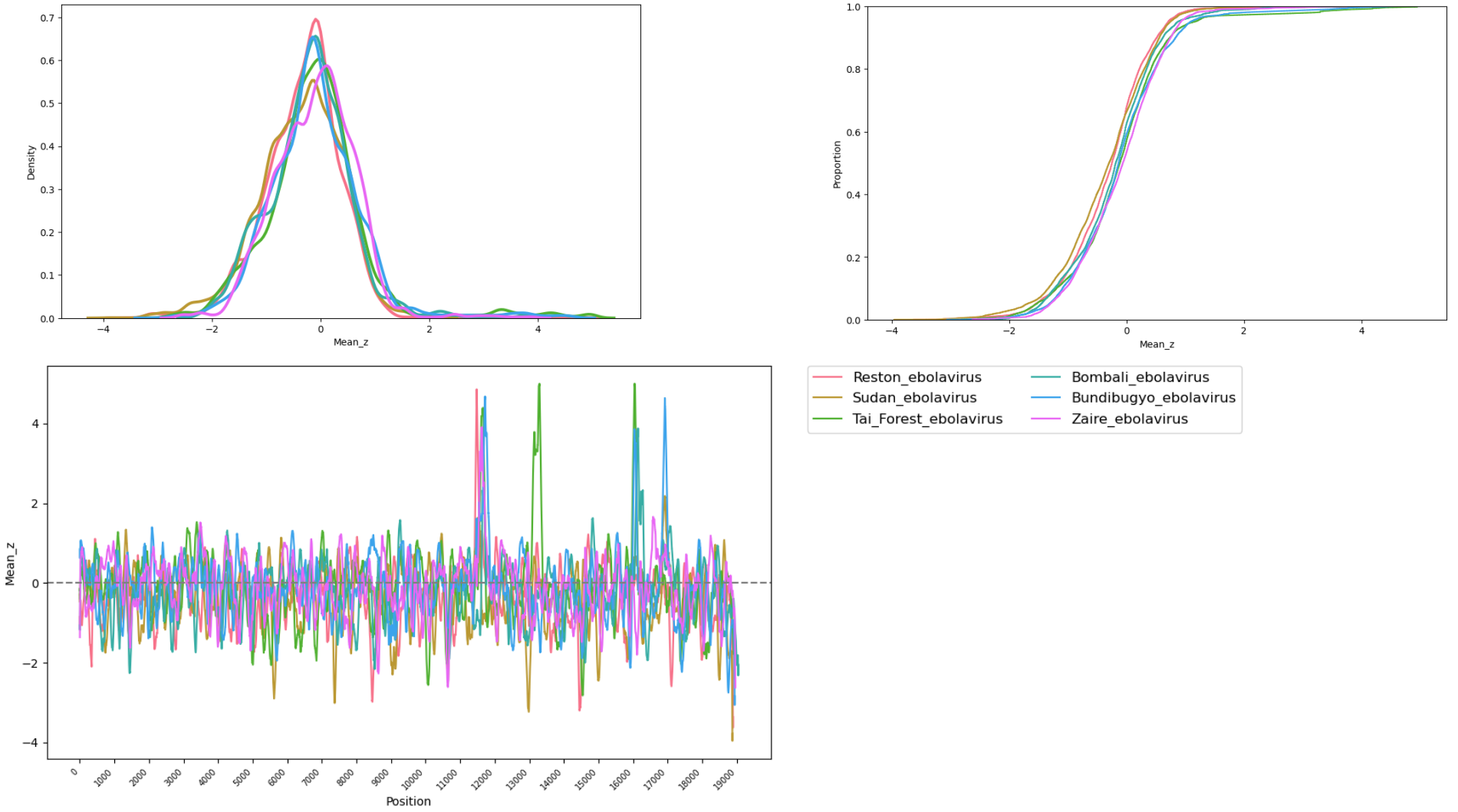

Figure 32: Genus level analysis in Ebolavirus, 3'-5' strand. Upper left: density function of all Z-scores per position per species; upper right: ECDF of all Z-scores per position species; lower: Z-scores vs genomic position. Each line denotes a different species in the Ebolavirus genus

Next, for each virus in the Ebolavirus genus, the genome annotation together with the mean Z-score per position was visualized, as shown in Figure 33. The virus chosen as example is the Zaire ebolavirus, which was first recognized in 1976 and is still one of the deadliest human viruses [41]. The mean Z-score per position on the 3'-5' genome, together with the calculated minimum free energy (MFE) values in each sliding window are depicted in red, while the mean Z-scores per position and the MFE values in the 5'-3' antigenome strand are shown in blue. Visualizing the distribution of the Z-scores on a one to one scale with the genome annotation makes it possible to determine where in the genome there are more structured regions as indicated by the negative values of the mean Z-scores.
In the Zaire ebolavirus genome, there is one predominant positive peak between the positions 7000 and 8000, where the value of the Z-score reaches, on the antigenome strand, the value 5 and on the genome strand the value 4. This particular region is located in the GP coding region. The GP gene is responsible for glycoprotein synthesis and is the only protein that is expressed on the virion's surface [64].
Overall, the Zaire ebolavirus is not particularly structured, as the mean Z-scores per position fluctuate around 0, which is a feature of all ebolaviruses, as presented in Figures 31 and 32. Figure 34 shows the overlap of the Z-scores per position in the two strands, to better asses if indeed both the genome and antigenome have the expected similar behaviour with respect to the property of being structured. By visually inspecting the figure below, it is clear that the overlap of the two lines is almost identical, and the regions that are more positive or negative are at the same location on both strands.
Further on, using the RNALfold module from the ViennaRNA package, it was determined whether there are locally stable structures in the genomes. From the RNALfold output, the length and location of the structured regions was processed and filtered to get those structures whose Z-score $\leq$ -2, as to locate only the most reliable locally stable structures. The visualization of the RNALfold output, together with the mean Z-score per position and genome annotation can be seen in Figures 35 and 36.

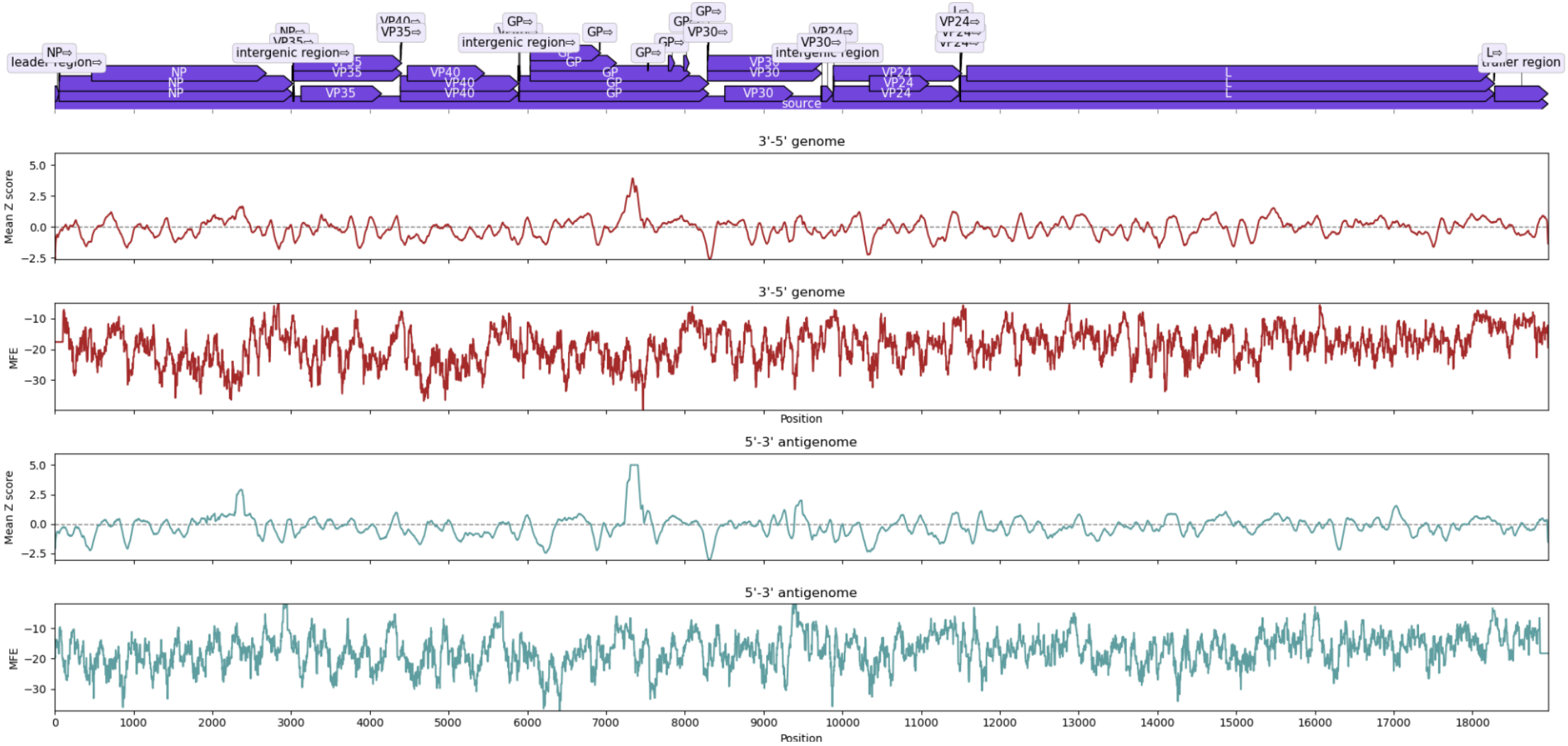

Figure 33: Species level analysis for the Zaire ebolavirus. The annotation of the genome is visualized together with the mean Z-score per position and the MFE value of each window of length 100. Red: 3'-5' genome values; Blue: 5'-3' genome values. NCBI accession: `NC_002549.1`

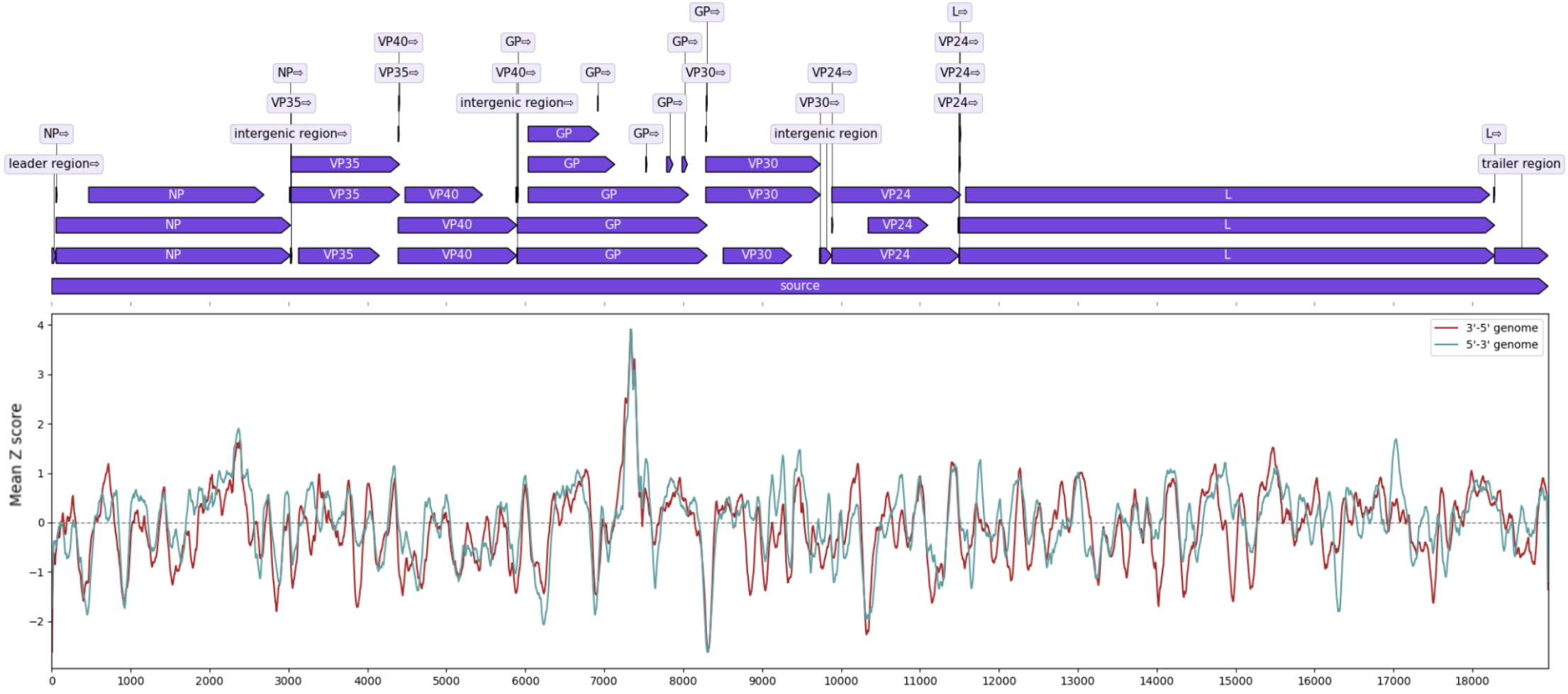

Figure 34: Species level analysis for the Zaire ebolavirus. The annotation of the genome is visualized together with the mean Z-score per position on both strands. Red: 3'-5' genome values; Blue: 5'-3' genome values. NCBI accession: `NC_002549.1`

There are some structured regions on the 5'-3' antigenome strand, for example around position 6000 and 10000, denoted by both the RNALfold output as well as negative mean Z-scores near those positions. From all intergenic regions, the one between VP40 and GP seems to be particularly structured. In the VP24 gene location, around position 10300, the mean Z-scores reach a value of -2, which denotes the possibility of a highly structured region and RNALfold predicts as well at least a few structures of different lengths in that region, whose Z-scores are definitely negative. The leader and trailer regions have negative Z-scores, however there is no locally stable prediction from RNALfold, at least not for structures with a Z-score lower than the selected threshold. The positive peak, around position 7000, is located in the GP gene region and, as expected, there is no RNALfold structure prediction around that position.
Figure 36 shows the RNALfold output, with the genome annotation and mean Z-score per position for the 3'-5' genome strand. A high density of RNALfold predicted structures can be seen in the VP30 and GP intergenic region, as well as in the beginning of the GP gene. The positive peak in also located between positions 7000 and 8000, similarly to the 5'-3' strand. There are some positions in the trailer region, whose mean Z-scores are positive and for which RNALfold predicts one locally stable structure, which is a distinct feature compared to the 5'-3' strand.
The final step of the analysis was to differentiate between CDS and non-CDS regions in the genomes, on the genome level as well as genus level. For all the six ebolaviruses, there is no gene annotation available for the 3'-5' genome, as all of the coding regions are found of the antigenome strand 5'-3'. Figure 37 shows the distribution of the Z-scores in the CDS or non-CDS positions in the Zaire ebolavirus genome. The Shapiro-Wilk test showed that the data in the CDS and non-CDS does not come from a normal distribution (CDS: Shapiro-Wilk test statistic = 0.89, p-value $<<$ 0.05; non-CDS: Shapiro-Wilk test statistic = 0.98, p-value $<<$ 0.05). The non-CDS Z-score curve is slightly skewed to the left, indicating that there are more negative Z-scores in the non-CDS region that in the CDS region.

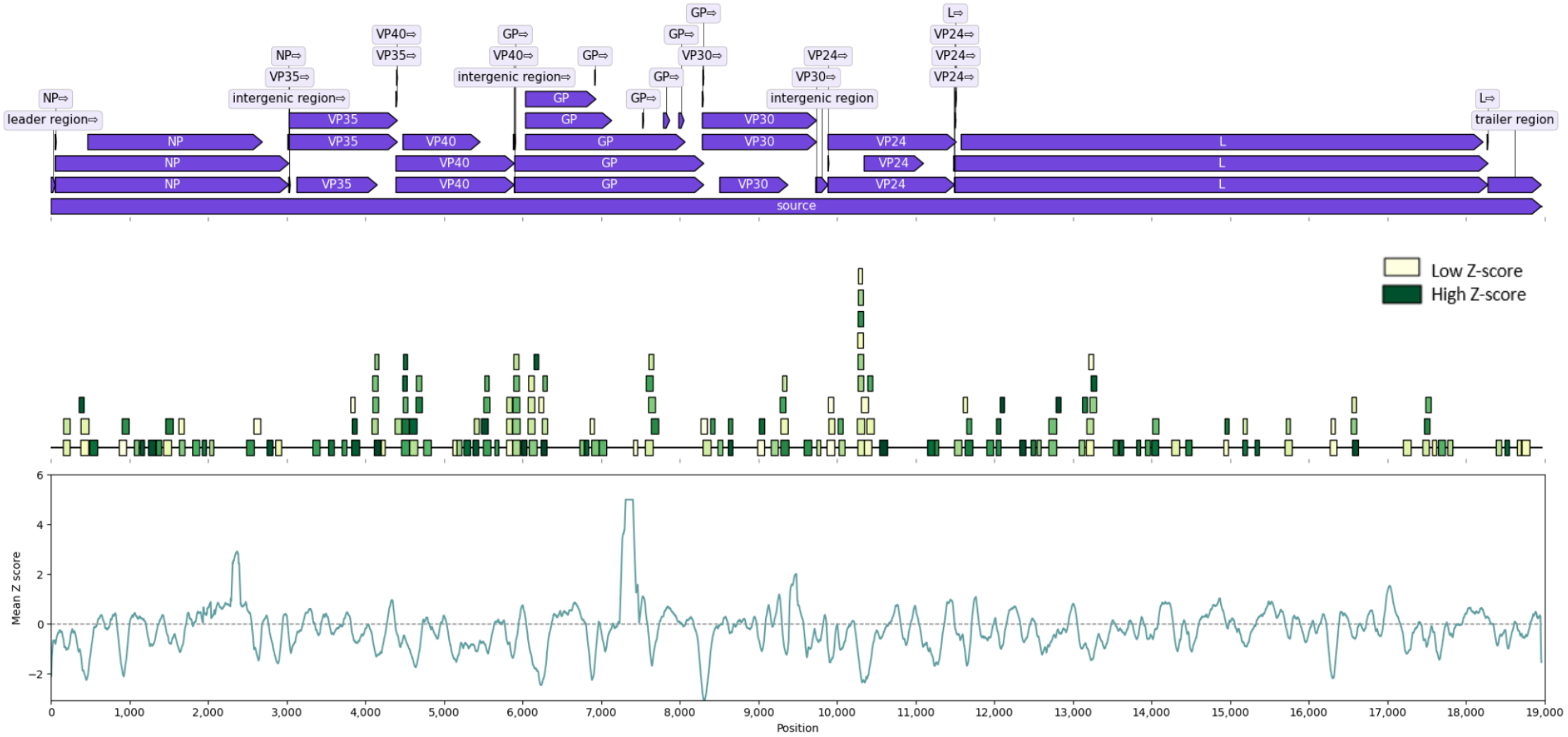

Figure 35: Genus level analysis in Ebolavirus, 5'-3' strand. The rectangles are the predicted RNALfold stable structures, their width is proportional to the length of the structure. All the predictions have a Z-score $\leq$ -2. NCBI accession: `NC_002549.1`

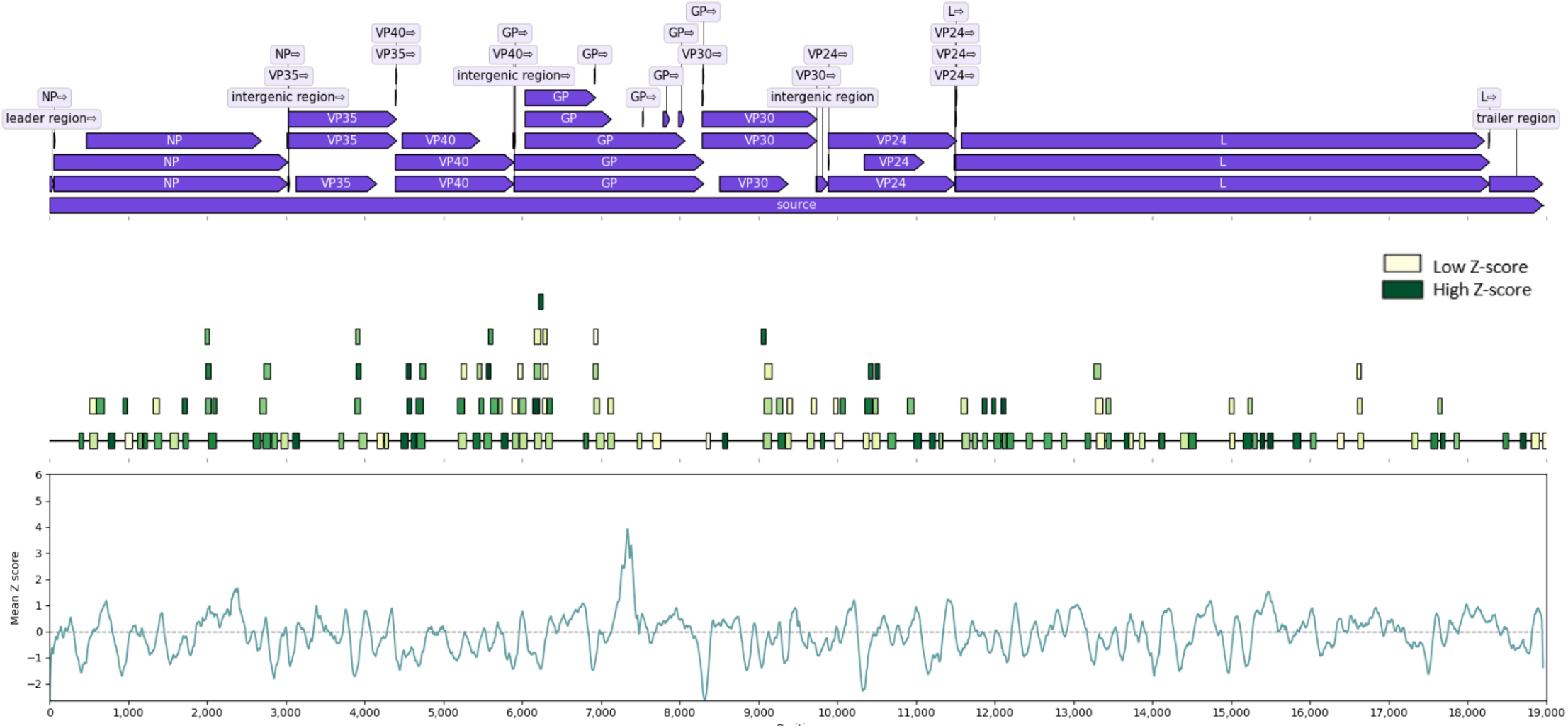

Figure 36: Species level analysis in Ebolavirus, 3'-5' strand. The rectangles are the predictied RNALfold stable structures, their width is proportional to the length of the structure. RNALfold output is filtered for structures whose Z-score $\leq$ -2. NCBI accession: `NC_002549.1`

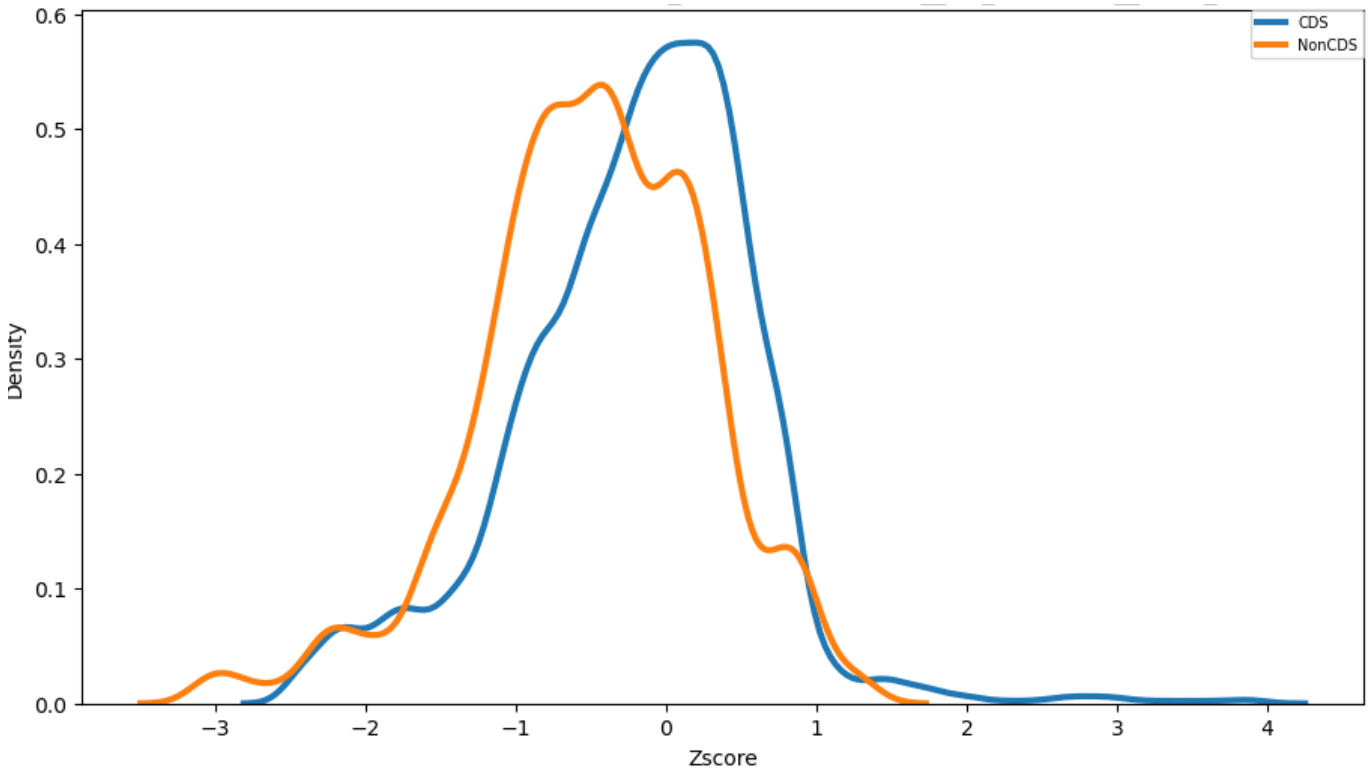

Figure 37: Distribution of mean Z-scores in CDS and non-CDS region of the antigenome in Zaire ebolavirus. Blue: CDS Z-scores, orange: non-CDS Z-scores. NCBI accession: `NC_002549.1`

Figure 39 summarizes the Z-scores in the CDS and non-CDS region for all the organisms in the Ebolavirus genus. The Mann-Whitney test (MWU) was used for determining if the underlying distribution of the mean Z-scores in the two regions are significantly different. This test was used because not all species follow a normal distribution of the Z-scores in the CDS or non-CDS regions (data not shown). For all species, the MWU-test indicates that the distribution of Z-scores in CDS vs non-CDS regions is statistically significant different. As seen in Figure 39, all viruses in the Ebolavirus genus behave similarly, with the median of the Z-scores in both CDS and non-CDS regions slightly below 0. Since there is no gene annotation on the 3'-5' strand in either of the six viruses, only the overview for the 5'-3' strand is presented in this case.
Figure 38 shows the distribution of the opening energy values for the antigenome strand spanning the position of the first start codon in the first coding region (marked as position 0, in red). The median of the opening energies in the region immediately before the first CDS range from 7 to 11 kcal/mol, while the median of the distribution at position 0 has approximately the value 8. However, the values increase at the CDS starts, suggesting the unfolding of the secondary structures and possibly facilitating RNA-RNA interactions. However, the correlation between the mean opening energy and the mean Z-score across the whole genomes in the *Filoviridae* family shows that the values of the opening energies increase as the Z-score are more positive (Supplementary Figure 53), meaning that the unfolding of the quite unstructured genome still requires a high amount of energy. The behaviour of the opening energy spanning the region of the first AUG codon in ebolaviruses denotes an unfolding trend that is otherwise not observable globally, throughout the entire genome in these viruses.

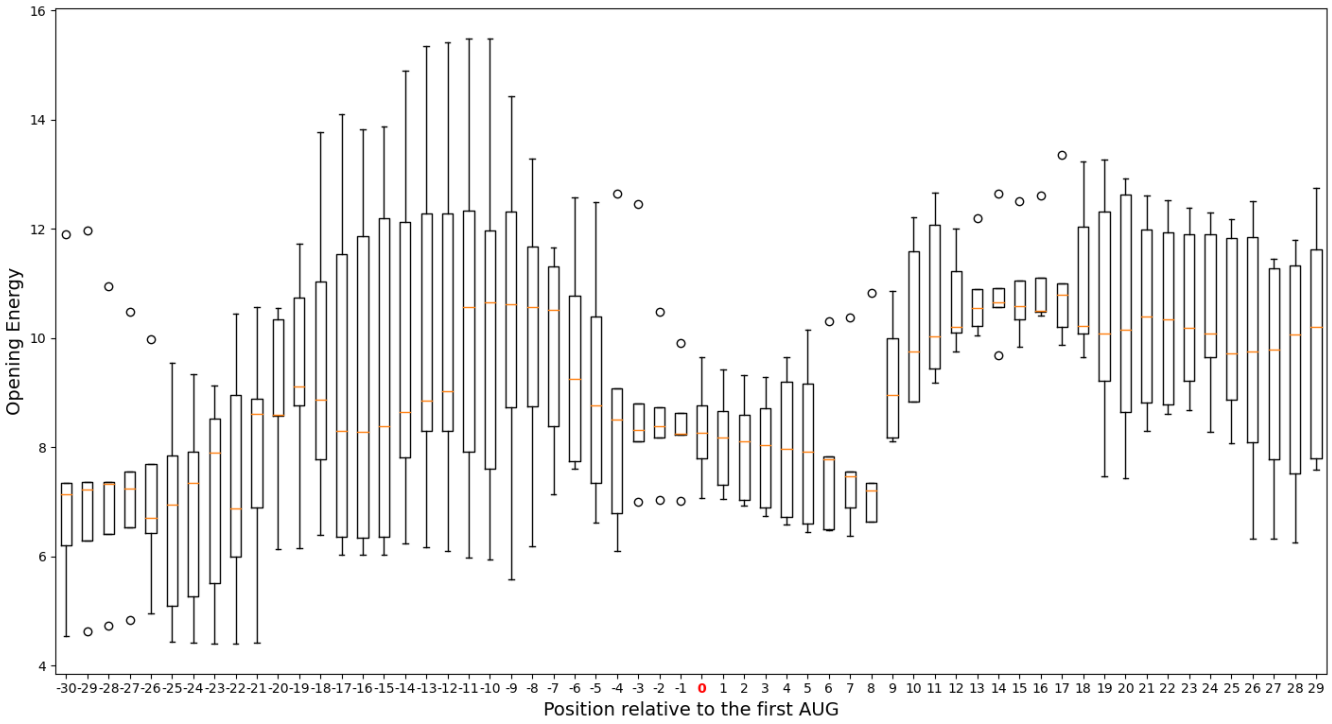

Figure 38: Distribution of opening energy values for the 60 nucleotide window spanning the first start codon in the first CDS of ebolaviruses

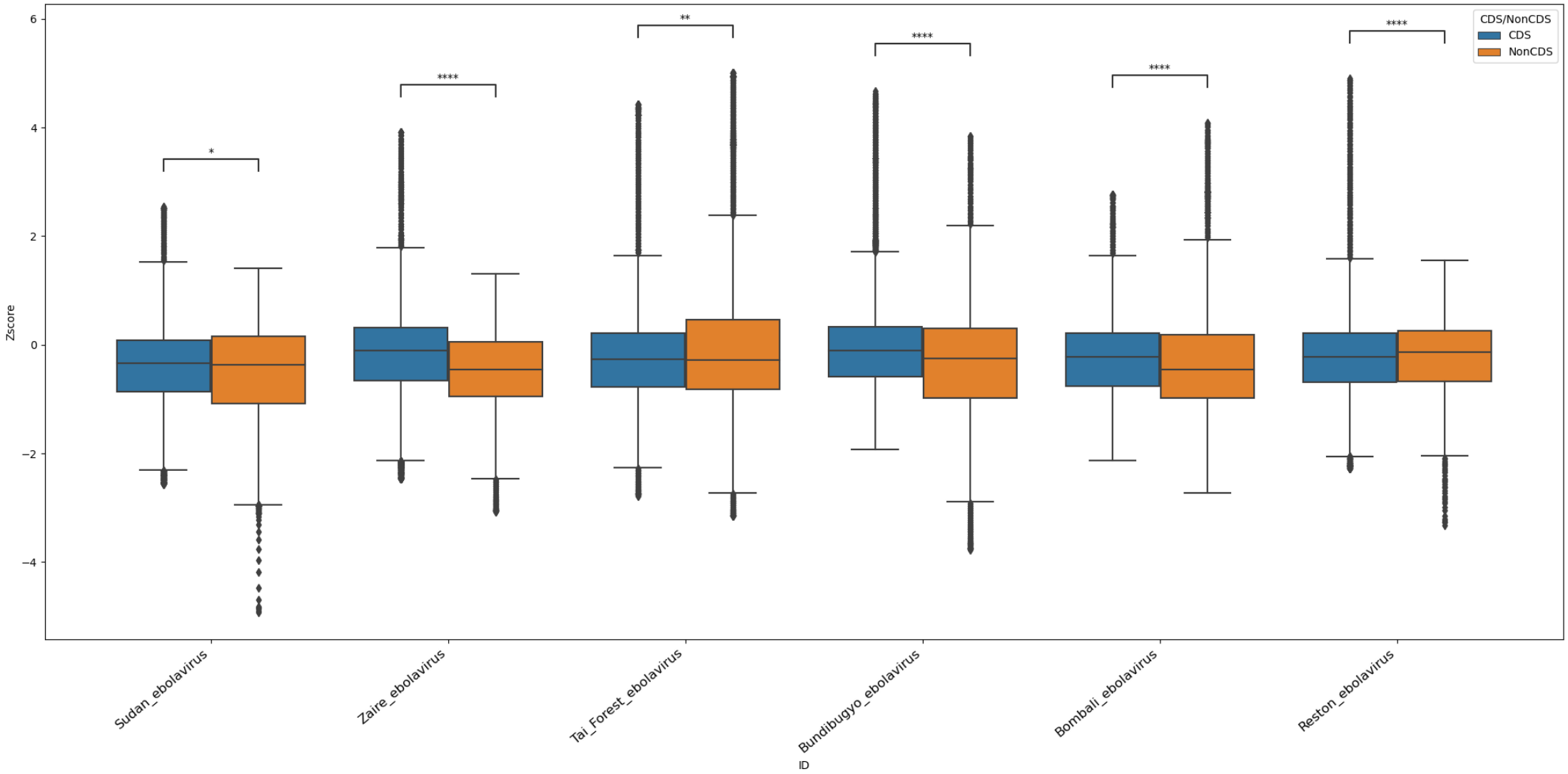

Figure 39: Boxplots of Z-score values in CDS and non-CDS regions for all six Ebolaviruses. Blue: CDS Z-scores, orange: non-CDS Z-scores. NCBI accessions (preserving the order in the figure): `NC_006432.1`, `NC_002549.1`, `NC_014372.1`, `NC_014373.1`, `NC_039345.1`, `NC_004161.1`; p-value annotation legend: ns: 5.00e-02 $\leq$ 1.00e+00; *: 1.00e-02 $\leq$ p $\leq$ 5.00e-02; **: 1.00e-03 $\leq$ p $\leq$ 1.00e-02; ***: 1.00e-04 $\leq$ p $\leq$ 1.00e-03; ****: p $\leq$ 1.00e-04

5.6.3 Segmented species level analysis from dsRNA

The Rotavirus genus, from the dsRNA group, belongs to the *Sedoreoviridae* family and contains six RefSeq approved viruses, with each genome containing 11 segments. The Rotavirus A is one of the best known examples in this family and is responsible for being the most common cause of severe diarrhea infections in young children [94] and is therefore of medical importance in the scientific community. The 11 segments of the Rotavirus A have lengths ranging from 660 to 3300 nucleotides and are enclosed by a tricapsid structure, resembling a wheel-like shape, hence the name rotavirus [15].
Figures 40 and 41 depict the density function and the ECDF plots of the Z-scores in all segments, as well as the distribution of Z-scores per position across all segments for the 5'-3' and 3'-5' strands, respectively. Since there are such big differences in the segment sizes, the x-axis has been normalized to range from 0 to 1 for the entire data set. By looking at the density function plots on both strands, it is visible that the distribution of the Z-scores does not follow a normal distribution in all segments, which can be attributed to the short length of some segments of the Rotavirus A; the ECDF plots also support this observation, as the curves do not resemble the shape of a normal distribution. The mean Z-score per position plot shows regions with structured elements, with most of the negative peaks in the 3'-UTR region of the genome on the 5'-3' strand. The most predominant positive peaks are attributed to segment 5, as shown in Figure 40.
The mean Z-scores per position on the 3'-5' strand fluctuate around 0, which the most negative peak being in the 3'-UTR region, belonging to segment 6. The most positive peak, with value 3, is attributed to segment 5, which also has the positive peaks on the 5'-3' strand. Overall, the genome of the Rotavirus A does not seem to be highly structured, which is also reflected in the ECDF plot, where it can be observed that only about 40% of the Z-scores have values smaller than 0. This results are consistent with the mean Z-score overview provided in Figure 5, as the values of the mean Z-scores per position are similar on both strands, reflected in the Rotavirus A virus very well for all its segments.

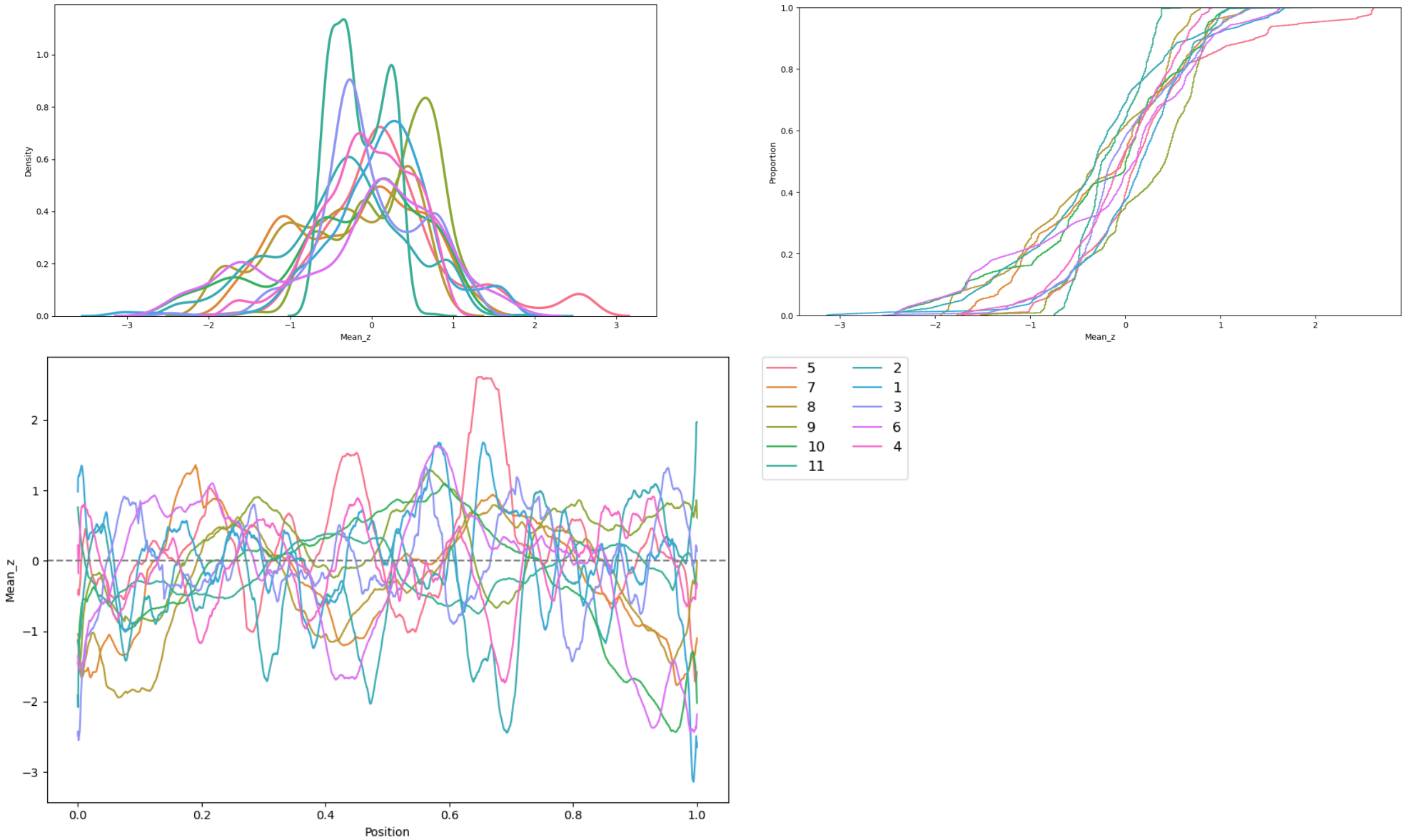

Figure 40: Species level analysis for Rotavirus A, 5'-3' strand. Upper left: density function of all Z-scores per position per segment; upper right: ECDF of all Z-scores per position per segment; lower: Z-scores vs genomic position, normalized on the x-axis. Each line denotes a different segment of the Rotavirus A species (see legend)

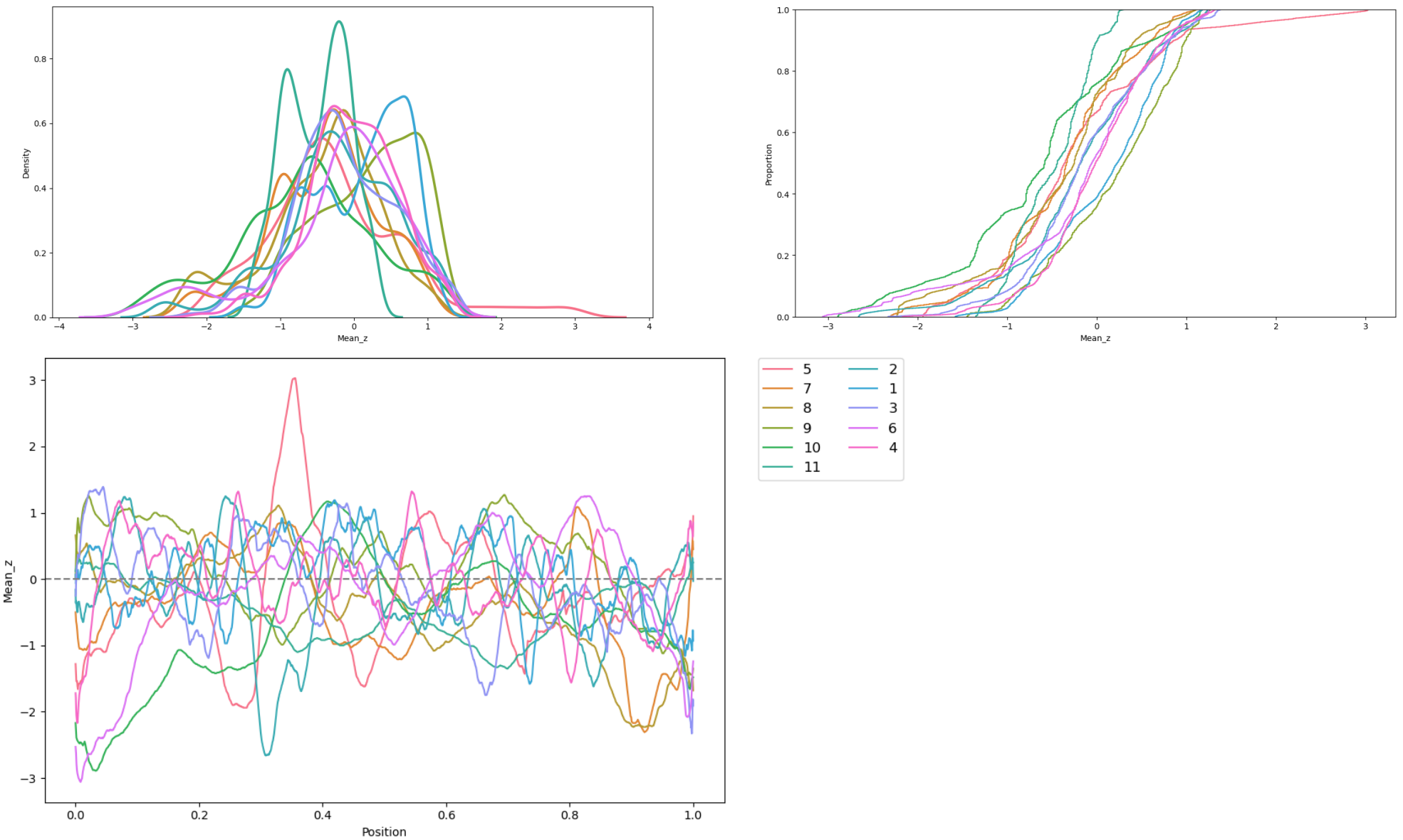

Figure 41: Species level analysis for Rotavirus A, 3'-5' strand. Upper left: density function of all Z-scores per position per segment; upper right: ECDF of all Z-scores per position per segment; lower: Z-scores vs genomic position, normalized on the x-axis. Each line denotes a different segment of the Rotavirus A species (see legend)

To detect possible locally stable structures in the genome, the RNALfold module was run for all the segments of the Rotavirus A. Here, only the results for segment 2 are shown, for both the 5'-3' and 3'-5' strand (Figures 42 and 43). The second segment of the Rotavirus A species has a length of 2693 nucleotides and encodes for the VP2 gene. There are a few stable structures predicted by RNALfold, the majority being between positions 1000 and 1500, where the values of the mean Z-scores per position are also below 0. The most negative peak, near position 2000, is matched by at least one thermodynamically stable structure from RNALfold, whose Z-score is very low, denoting a high confidence for that region to be structured. The 5'-3' strand seems to be more structured, observation based on both the mean Z-score distribution per position and also the RNALfold result. The positions of the most negative peak (approximately around position 2000) are included in the RNALfold prediction of at least one folded region with low Z-score. This peak matches the one on the 5'-3' strand, however there are a low regions on the 3'-5' strand where the mean Z-scores per position have slightly positive values as opposed to the 5'-3' strand, for example around position 1000 in the genome.
Finally, the distribution of the Z-scores in the CDS and non-CDS regions of segment 2 in Rotavirus A genome were analysed and plotted, as seen in Figure 44. The shape of the distributions does not resemble a normal Gaussian distribution. The Z-scores in the non-CDS seem to follow a bimodal distribution, with the main peak at around 1 and the second lower peak at around -1. On the other hand, the distribution of the Z-scores in the CDS region seem to follow a multimodal distribution, with the main peak at around 0 and the two additional peaks at around -1.5 and 1. The MWU test was used to determine if the two sample distributions come from the same distribution; the null hypothesis was rejected with p-value $<< 0.05$.

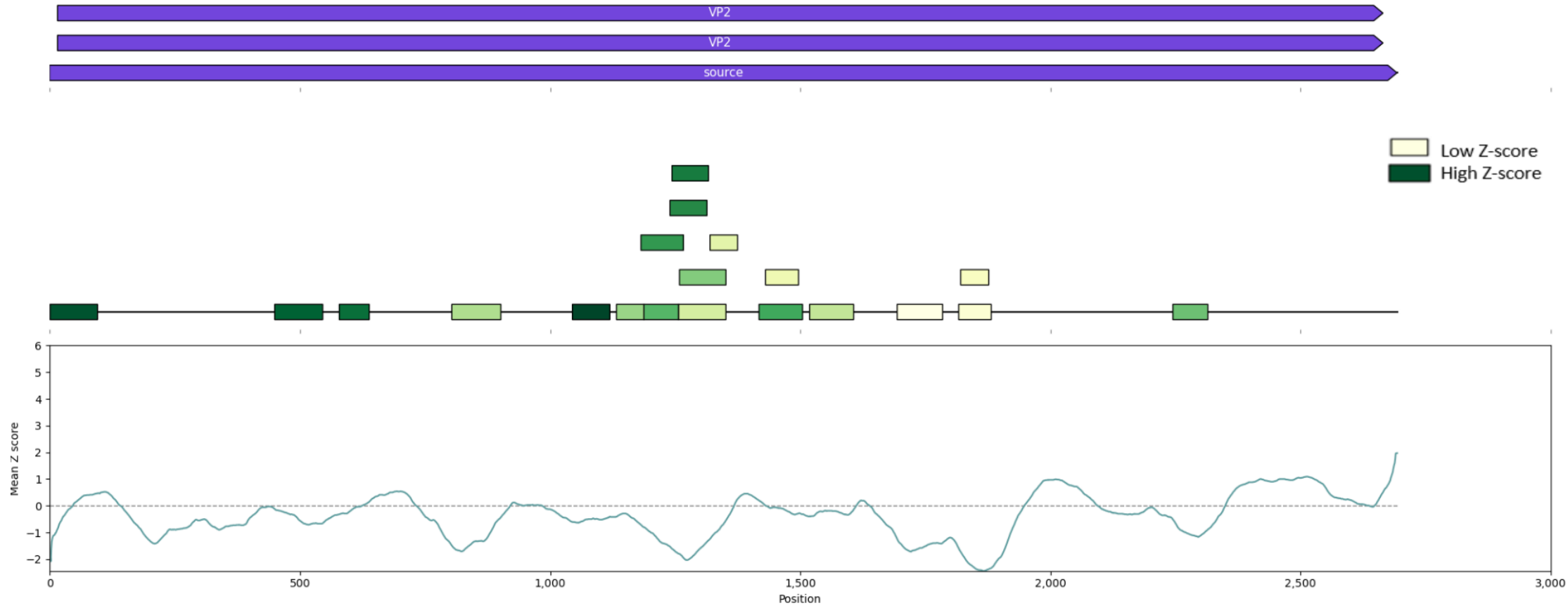

Figure 42: Species level analysis for Rotavirus A, 5'-3' strand. The rectangles are the predicted RNALfold stable structures, their width is proportional to the length of the structure. RNALfold output is filtered for structures whose Z-score $\leq$ -2. NCBI accession: `NC_011506.2`

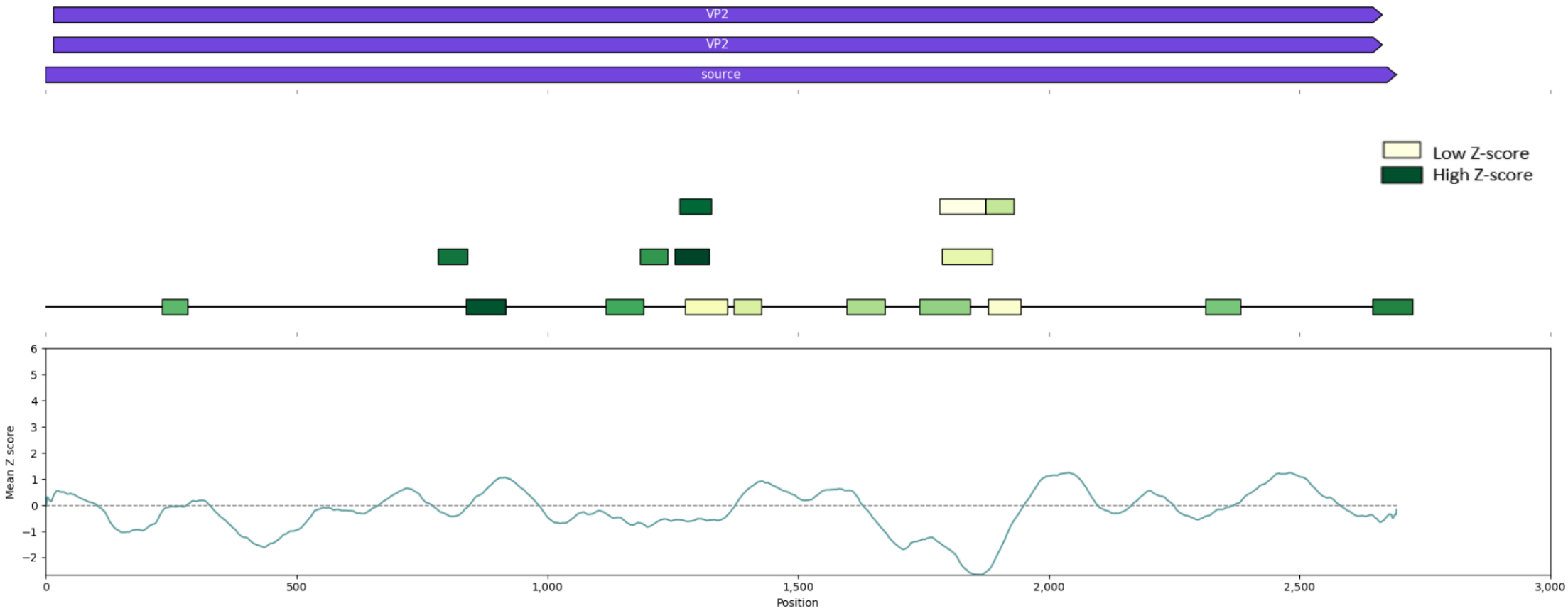

Figure 43: Species level analysis for Rotavirus A, 3'-5' strand. The rectangles are the predicted RNALfold stable structures, their width is proportional to the length of the structure. RNALfold output is filtered for structures whose Z-score $\leq$ -2. NCBI accession: `NC_011506.2`

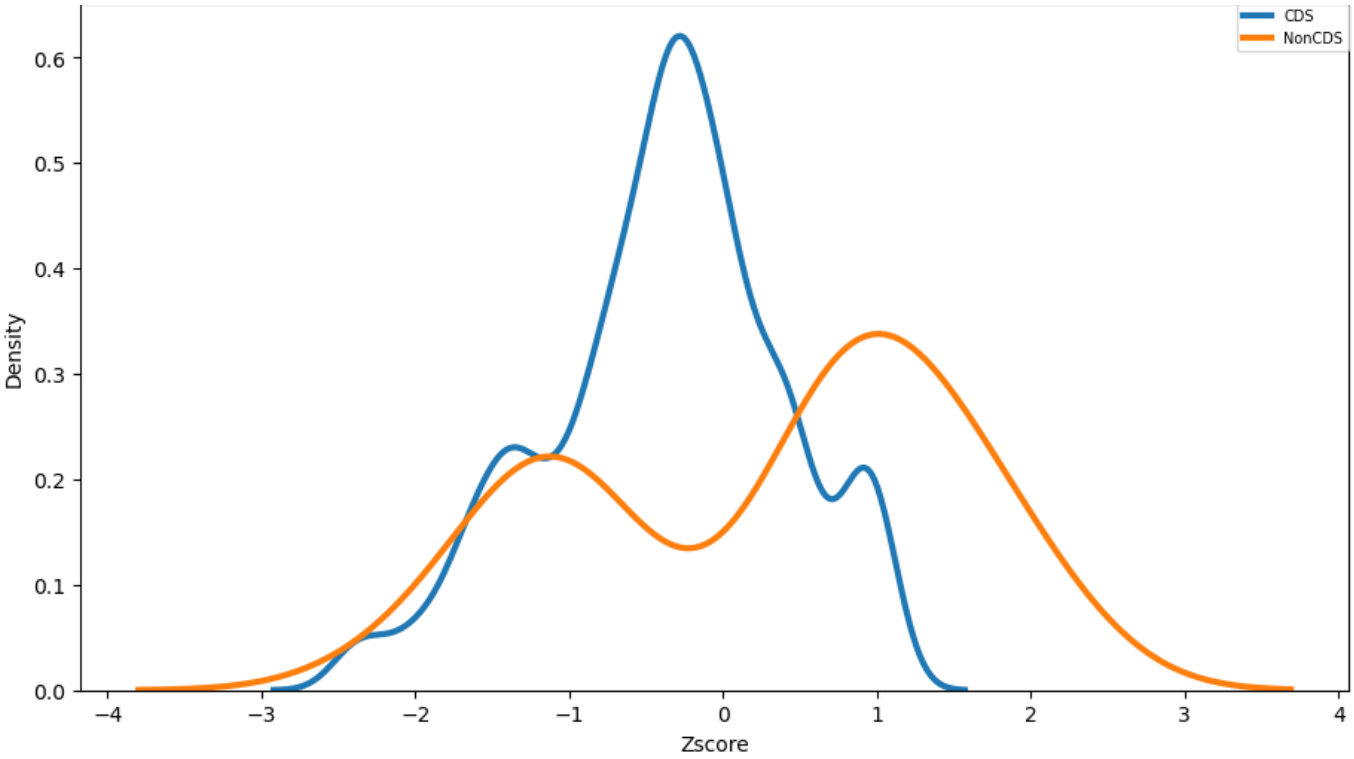

Figure 44: Distribution of mean Z-scores in CDS and non-CDS region of the antigenome in segment 2 of Rotavirus A. Blue: CDS Z-scores, orange: non-CDS Z-scores. NCBI accession: `NC_011506.2`

Figure 46 shows an overview of all segments of the Rotavirus A genome, where the mean Z-scores per position have been differentiated by their location in either a CDS or non-CDS region. The Mann-Whitney test (MWU) was used for determining if the underlying distribution of the mean Z-scores in the two regions are significantly different. Only segments 1, 9 and 11 seem to have non-significant differences in their mean Z-scores in CDS and non-CDS regions. It can be immediately observed that the median values of the Z-scores differ a lot between segments and there is no trend for either region to be have more negative Z-scores than the other.
Figure 45 shows the distribution of the opening energy values in the region spanning the first AUG start codon in the first CDS of each Rotavirus virus segment. It can be seen that the median value of the distribution for the majority of the positions is ranging between 7.5 and 11 kcal/mol, with the median for position 0 (location of the start codon) being approximately 10 kcal/mol. The fact that the values do not fluctuate very much between the non-CDS and CDS location is not surprising, considering the fact that the mean Z-scores in the CDS and non-CDS regions in the dsRNA group are similar. This suggests that there is a thermodynamic trend to facilitate RNA-RNA interactions throughout the segmented genomes of the Rotavirus genus, supported also by the global negative correlation between the mean Z-scores and opening energy values in the *Reoviridae* family, as shown in Supplementary Figure 54.

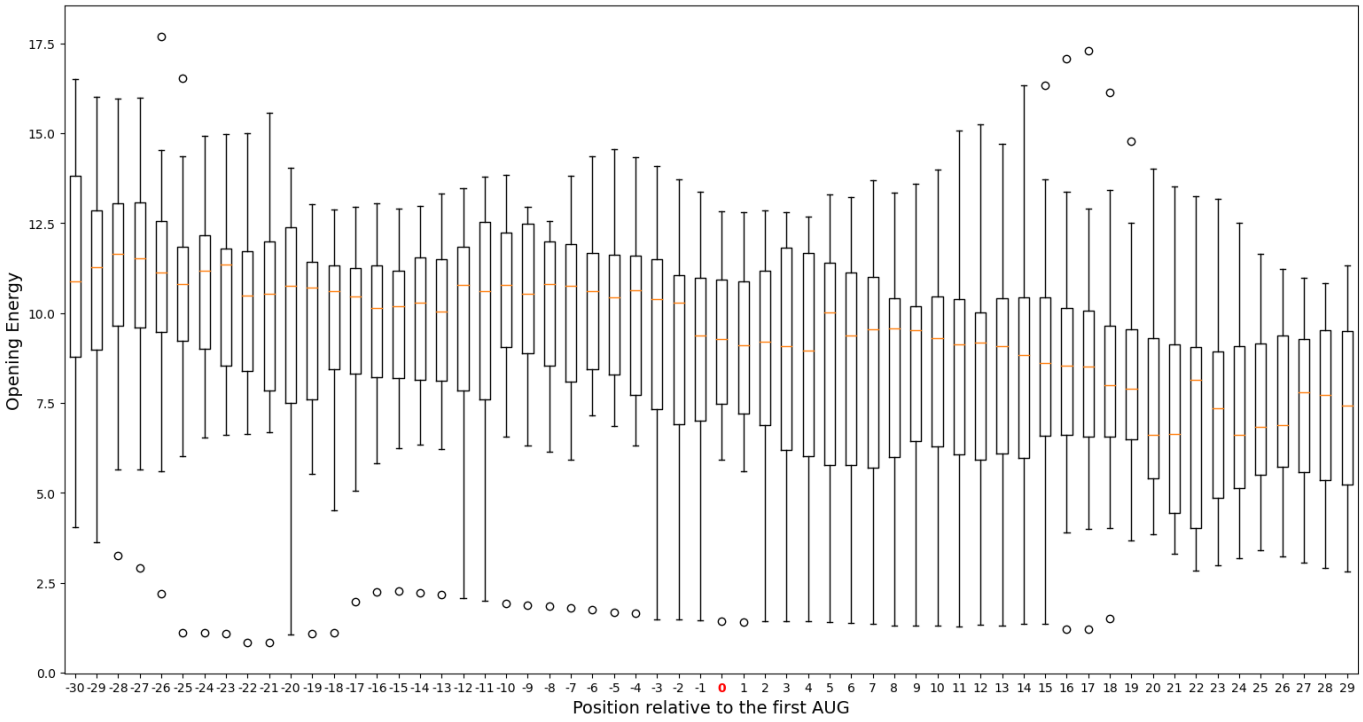

Figure 45: Distribution of opening energy values for the 60 nucleotide window spanning the first start codon in the first CDS of rotaviruses

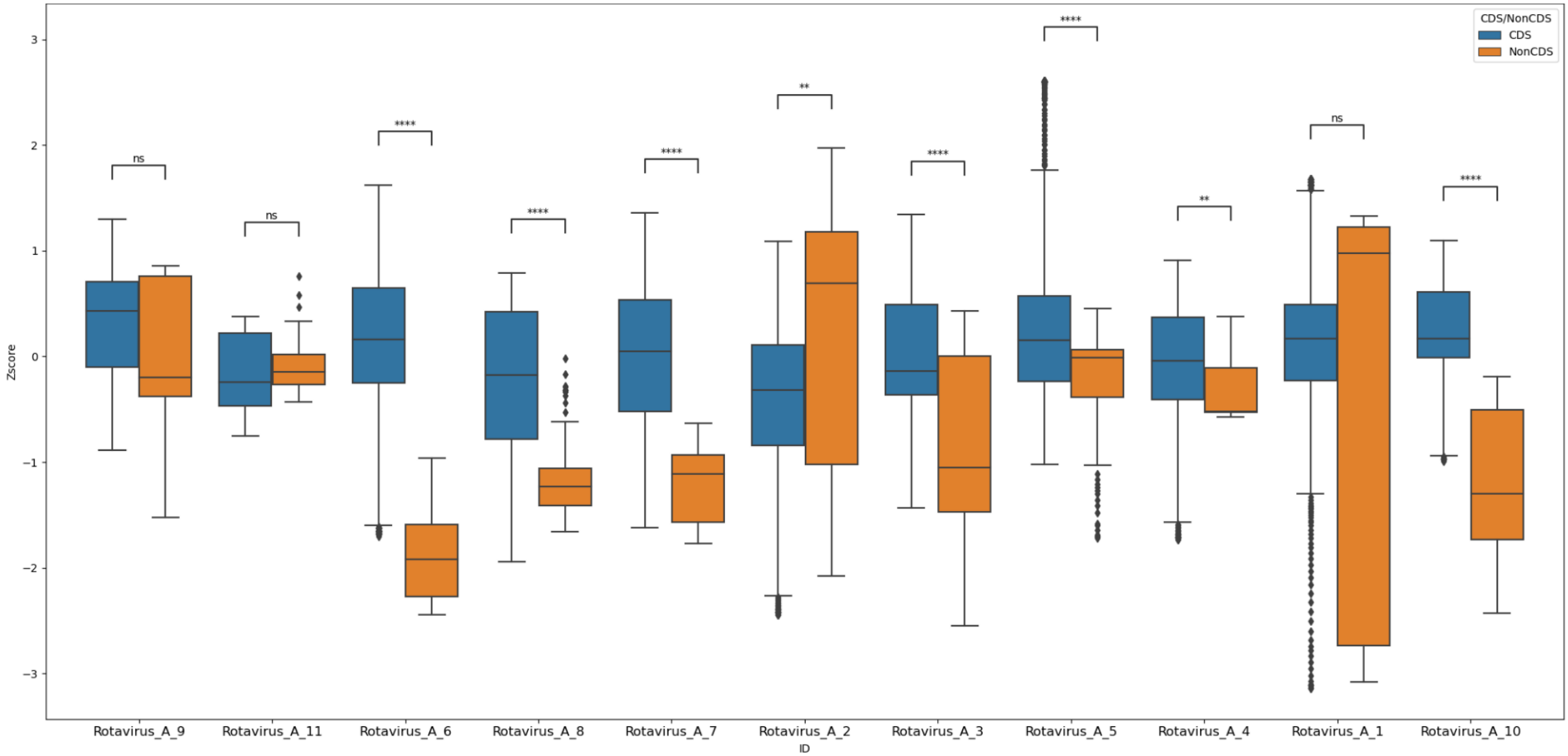

Figure 46: Boxplots of Z-score values in CDS and non-CDS regions for all eleven segments in Rotavirus A. Blue: CDS Z-scores, orange: non-CDS Z-scores. NCBI accessions (preserving the order in the figure): NC_011503, NC_011505, NC_011509, NC_011502, NC_011501, NC_011506, NC_011508, NC_011500, NC_011510, NC_011507, NC_011504. p-value annotation legend: ns: 5.00e-02 $\leq$ 1.00e+00; *: 1.00e-02 $\leq$ p $\leq$ 5.00e-02; **: 1.00e-03 $\leq$ p $\leq$ 1.00e-02; ***: 1.00e-04 $\leq$ p $\leq$ 1.00e-03; ****: p $\leq$ 1.00e-04

5.7 SARS-CoV-2 CoVariants analysis

Because the pandemic from December 2019 raised awareness of the violent virulence of SARS-CoV-2 virus from the *Coronaviridae* family, the results in these section are dedicated to the analysis conducted on the different SARS-CoV-2 strains, which were responsible for infections across the whole globe. Using the background information found at https://covariants.org/ in May 2022, six different CoVariants (SARS-COV-2 virus strains) were taken into account for an in depth analysis with focus on structuredness: 20I (Alpha), 20H (Beta), 20J (Gamma), 21A (Delta), 21K (Omicron), 21L (Omicron). Sequencing data made available in the GisAid database [56] was used for the analysis, as described in the Material and Methods part. For each CoVariant, 20 fully sequenced genomes from different continents were retrieved and structuredness in the means of Z-score computation was measured for each sample individually.

The data presented here summarizes the results obtained for the 20J (Gamma) strain. 20J emerged in Brazil in November 2020 and contains 17 mutations, three of which in the spike protein. This very aggressive strain was associated with high mortality rates and ability to spread very fast [34]. Nucleotide sequencing data from 20 patients infected with 20J, as well the SARS-CoV-2 reference sequence (`EPI_ISL_402124`) used by the research community were retrieved from GisAid website. Figure 47 shows the distribution of mean Z-scores per position for all downloaded sequences isolated from individuals infected with the Gamma SARS-CoV-2 strain.

Except for two, all sequences align almost perfectly with the reference genome. The two sequences with different Z-scores values between positions 3000 and 4000, 6000 and 7000, 11000 and 12000, 22000 and 23000. In these genomic intervals, their Z-scores reach a value of 5, the upper cut-off of the Z-score computation program. After manually inspecting the respective Fasta files with the nucleotide sequence of the two genomes, the conclusion was drawn that the particularly positive values are due to sequencing errors, as exactly in at the positions were the mean Z-scores reach such high values, some nucleotides were missing and were therefore annotated by the sequencing facility with the letter 'N'. The most negative peak lies approximately at position 26000, which is included in the genomic position encoding for the ORF3a. The distribution of the mean Z-scores per position denotes a high level of structuredness throughout the genome, as rarely the mean Z-scores reach positive values. The fact that the genome of the 20J variant appears to be very structured and shows negligible differences from the reference genome (which originates from China), denotes the evolutionary pressure of forming stable secondary structures in order to survive and maintain a successful viral life-cycle, regardless of geographic localization and epigenetic factors.

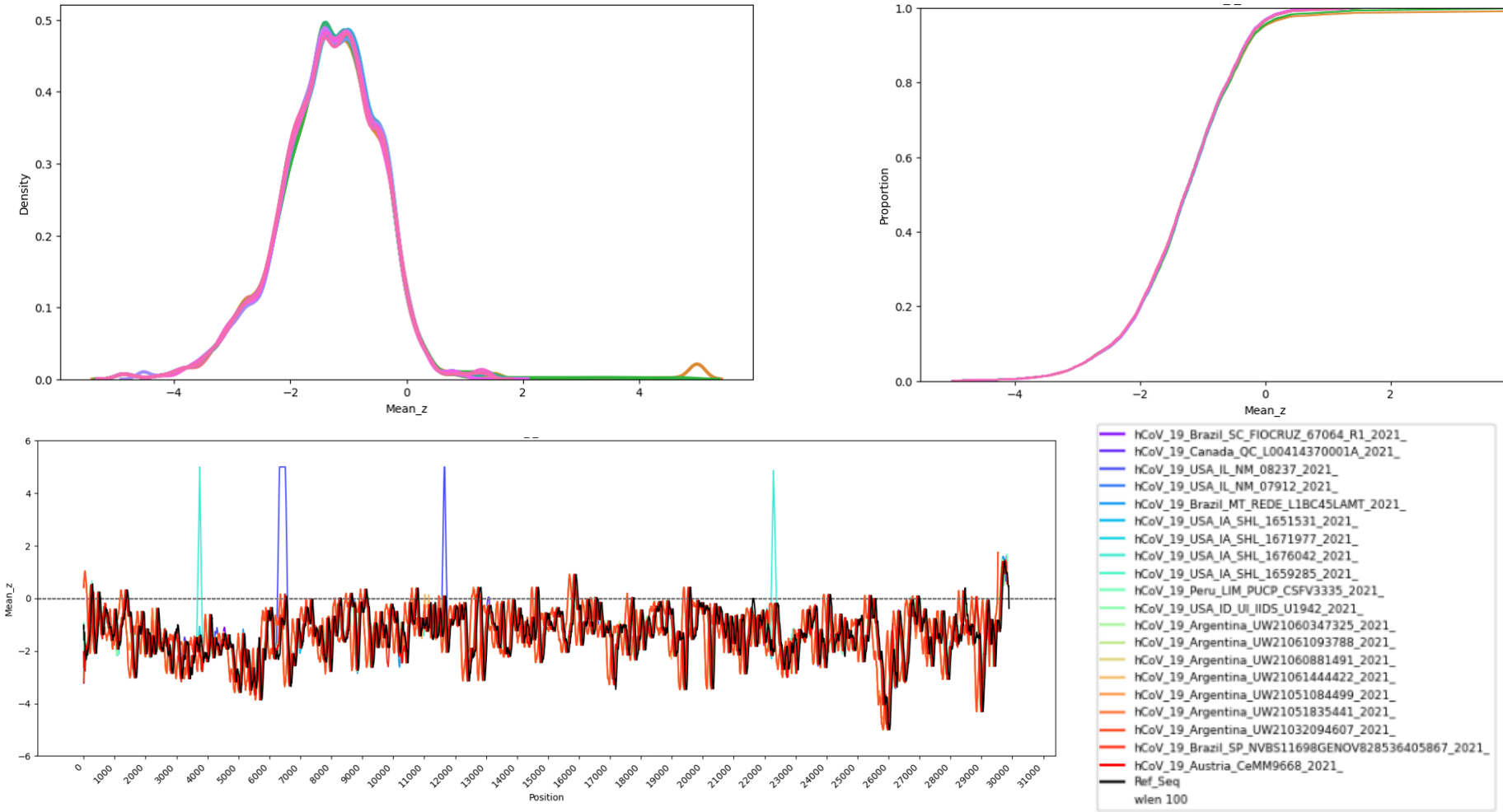

Figure 47: Analysis of SARS-COV-2 Covariant 20J from different hosts. Upper left: density function of all Z-scores per position per genome; upper right: ECDF of all Z-scores per position per genome; lower: distribution of mean Z-scores per genomic position. Each line denotes a different nucleotide sequence isolated from a different individual; the reference sequence is shown in black

For a better visualization and interpretation of the results, Figure 48 shows one 20J sequence, extracted from a patient from Buenos Aires in Argentina, together with the genome annotation, RNALfold output and mean Z-scores per genomic position. As expected, based on the negative mean Z-scores observed in Figure 48, the locally stable secondary structures predicted by RNALfold are abundant in the entire sequence, especially around position 7500 and 24000, which are included in the genomic locations that encode for the ORF1ab and spike protein, respectively. The famous frameshifting element, located approximately between positions 13.405 and 13.417 is well conserved among all CoVariants. The secondary structure prediction programs in the ViennaRNA package are not able to detect pseudoknot-like strctures, due to the high computational implications of their inclusion in the algorithms. Nevertheless, the local predictions by RNALfold show that the region spanning the location of the frameshifting element is particularly structured, suggesting the presence of stable secondary structures and thus aligning nicely with the previous results reported in the literature. The mean Z-scores in those positions reach negative values of around -2 and the RNALfold predictions marks the presence of at least three stable secondary structures in that region. Many bioinformatic analysis have proved the high drive of SARS-CoV-2 genomes to fold into stable secondary structures, more so than the best known example of a structured RNA virus, the HCV [27]. These previously reported observations match the results in this thesis, as the mean Z-score of the SARS-CoV-2 genome (-1.34) is more negative than the mean Z-score of the HCV virus, regardless the genotype (HCV genotype 1: -0.92, HCV genotype 2: -0.8, HCV genotype 3: -1.04, HCV genotype 4: -0.95, HCV genotype 5: -0.87, HCV genotype 6: -0.88, HCV genotype 7: -0.99). In the *Coronavirineae* family, where the coronaviruses are included, the correlation between the opening energy value and the mean Z-score shows a negative relationship between these parameters (Supplementary Figure 52). This denotes a global trend that their highly structured genomes, implicitly the CDS, have the thermodynamic potential to be unfolded in order to facilitate the RNA-RNA interaction between their nucleotide sequence and the ribosome. Since the structuredness pattern in the SARS-CoV-2 reference genome and all the genomes for each of the six CoVariants mentioned before shows no notable differences, it can be assumed effect of the opening energy on their genomes is similar. Both the mean Z-score per position, as well as the RNALfold output delivered very similar results, showing how well conserved the folding pattern is in SARS-Cov-2 and how the different mutations accumulated by the different variants did not change the global thermodynamic stability of the virus. These observations for the different CoVariants are made solely based on the body of data analysed in this work and rely on the sequencing data made available in the GisAid database.

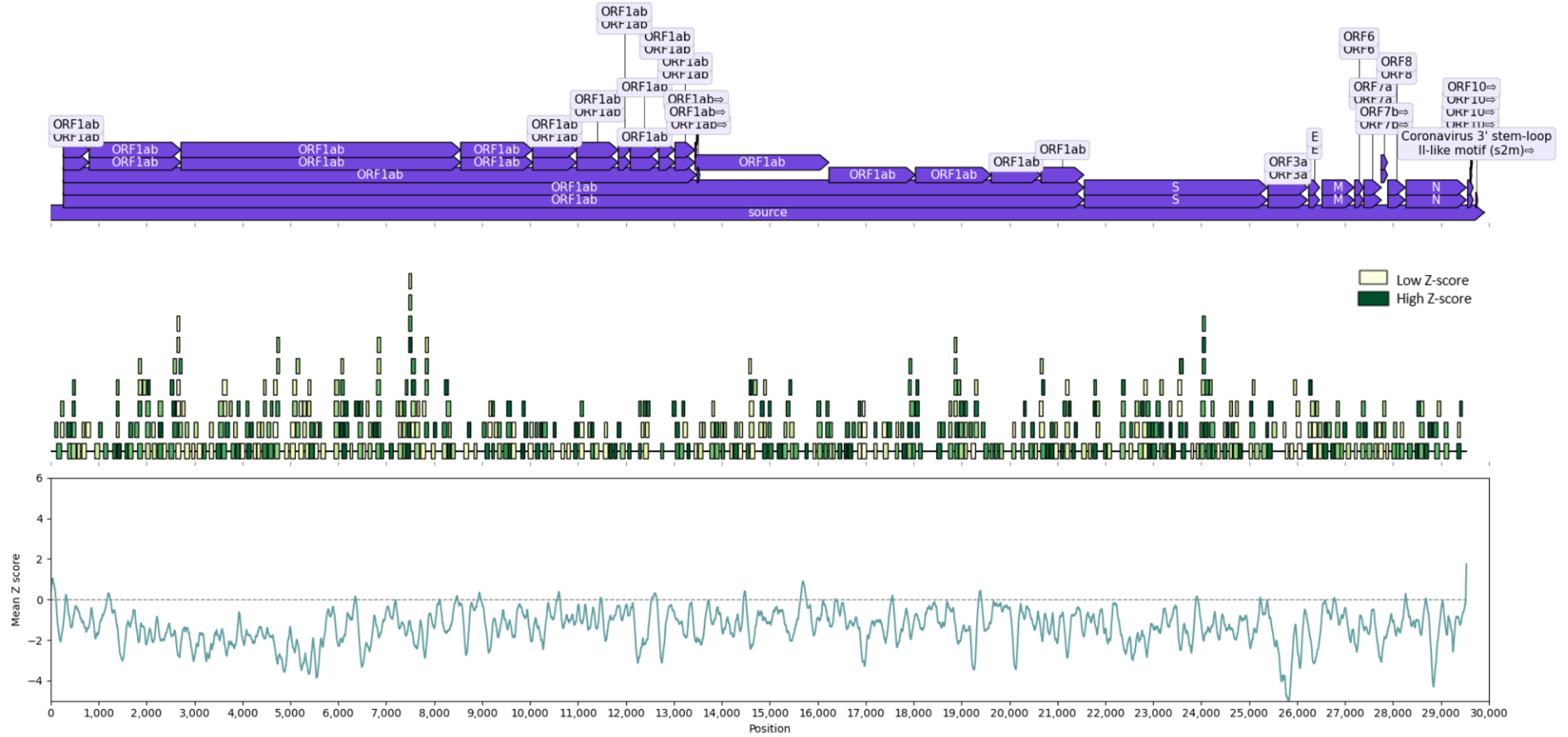

Figure 48: Analysis of one nucleotide sequence belonging to the 20J CoVariant. The rectangles are the predictied RNALfold stable structures, their width is proportional to the length of the structure. RNALfold output is filtered for structures whose Z-score $\leq$ -2. GisAid database accession: `EPI_ISL_10515211`

Since the SARS-CoV-2 virus belongs, from a taxonomic point of view, to the Betacoronavirus genus, it was interesting to determine how the distribution of the opening energy behaves across the whole genus in the region spanning the first AUG start codon, in the first CDS. Figure 49 shows the distribution of the opening energy values for each position as boxplots, with position 0 marking the beginning of the first CDS in the betacoronaviruses. It seems like the values of the opening energy take lower values in the immediate positions following the start codon, suggesting a very possible thermodynamic interaction between their genomes and the ribosome, thus denoting at least a thermodynamic potential for a qualitatively good gene expression. The high opening energy values make sense, as the genome in betacoronaviruses are likely to be very structured (mean Z-score value $\sim$ -1). These observations align nicely with the results shown in Supplementary Figure 52, where the members of the *Coronaviridae* family have a clear negative correlation between their mean Z-scores and opening energy values.

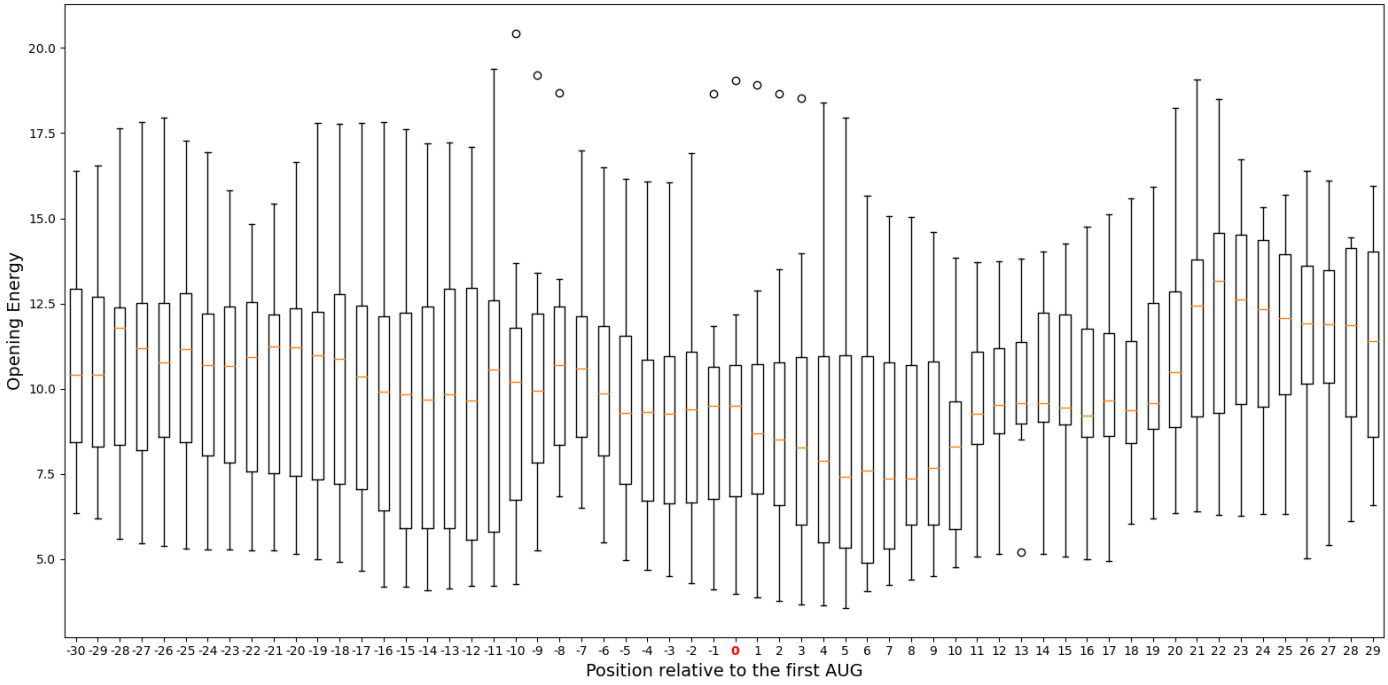

Figure 49: Distribution of opening energy values for the 60 nucleotide window spanning the first start codon in the first CDS of rotaviruses

6 Discussion

6.1 RNA structuredness in the different Baltimore groups

The results presented in this thesis showed that there is no considerable difference between the overall mean Z-score value of the forward and backward strands in the ssRNA(-) and dsRNA viruses. This means that the thermodynamic properties of the positive and negative RNA strands of a virus are similar in the context of secondary structure formation and stability. The genomes in the ssRNA(+) are the most structured ones in all groups. The global trend of the ssRNA(+) to have more structured genomes than the other RNA virus groups could be explained by the underlying biology of these viruses. As their genomes act as mRNAs, whose function is to further provide the proteins essential for viral replication, there needs to be an efficient control of the mRNA translation, not only to ensure a qualitatively good gene expression, but also to avoid the RNA degradation [29]. This translational control has been attributed to the highly conserved DEAD-box RNA helicase Dhh1, which which targets mRNAs with long structured CDS [53].
The overall difference in structuredness across the different Baltimore groups can be ultimately also linked to the length of their RNA genome: it was shown that there is a positive correlation between the length of the mRNA and the stability of RNA structures [35]. This trend can be observed in this work, too. The Pearson correlation coefficient between the genome length and the number of RNALfold predicted structures for in each genome shows a positive correlation in all groups (Pearson's r in dsRNA: 0.62, Pearson's r in ssRNA(-): 0.55; Pearson's r in ssRNA(+): 0.61), as shown in Supplementary Figure 50 and shows that, as expected, the length of the genome correlates with the number locally stable structures predicted. Another interesting finding is that the abundance of the RNALfold hits is not dependent of genome length, as shown in the Supplementary Figure 51: the number of locally stable secondary structures, in the case of ssRNA(+) and ssRNA(-), is not related to the number of nucleotides in the RNA sequence. This shows that the ability of a genome to be structured which offeres protection against degradation can be achieved regardless of its length. In the case of the dsRNA viruses, the abundance of local structures is more high in the genomes where the nucleotide sequence is shorter (approximately 6000 nucleotides). A cut-off for RNALfold predictions of -4 was used in order to filter for the most stable secondary structures. The global structuredness can be also assessed by comparing the number of locally stable structures predicted by RNALfold in the groups.
The average number of structures predicted in ssRNA(+) normalized by the genome length is 0.0016, while in ssRNA(-) it is 0.0012 on both strands. This supports the previous results that the mean Z-score in ssRNA(+) is much negative than in ssRNA(-) and dsRNA and demonstrates that genomes in the ssRNA(+) viral group are more likely to fold in highly stable structures than the genomes of ssRNA(-). For the dsRNA viruses, the average number of RNALfold predicted structures with a Z-score $\leq$ -4 is 0.0008, after normalizing for the length of the genome. By comparison, the dsRNA viruses are the least likely to contain a large number of stable elements, which is also reflected in the mean Z-scores presented in this thesis. In the data set used in this work, the ssRNA(+) have a higher average genome length (8388,7 nucleotides) than the ssRNA(-) (6087.2 nucleotides) and dsRNA (2534.7 nucleotides). The results reported in [35] align nicely with the results of this thesis, as

the mean Z-score in ssRNA(+) is more negative than in ssRNA(-), while the mean Z-score in dsRNA has the highest value across all compared groups.

6.2 RNA structuredness in CDS and non-CDS

There is an impressive body of published research that focused on the CDS and non-CDS in viral genomes and the importance of RNA folding in those regions for the survival and replication of a virus. The presence of secondary structures in ncRNA were shown to be the most important factor in fulfilling functions such as translation and replication, and thus would explain the Z-score value of -0.9 in the non-CDS regions in ssRNA(+) viral group. Nevertheless, the CDS region, with its mean Z-score of -0.64, also appears more structured than expected by chance. Indeed, there are many studies which have shown that the both the CDS and non-CDS regions across different ssRNA(+) families contained structured regions. The HCV, which was shown in this work as an example for the ssRNA(+) species level analysis, demonstrated a high thermodynamic ability for form stable secondary structures not only in the non-CDS but also in the CDS. Moreover, the results reported here determined the presence of at least one highly stable secondary structure in the 3'-UTR, confirmed by both the negative Z-score peak in that region as well as the RNALfold prediction. The results in this thesis are in line with the reported results from the literature, showing that indeed there is a high level of structuredness in the 5' and 3' UTR regions in the HCV virus.
For the ssRNA(-) group, on a global scale, their CDS and non-CDS regions have relatively negative Z-scores, which hints towards an indispensable structuredness on both the forward and backward strands. It should also be mentioned, that the majority of the ssRNA(-) viruses are small, segmented viruses whose annotations may be lacking some CDS regions and therefore this could also contribute to the value of the Z-scores presented in this work, when differentiating between the two regions.
In the case of the dsRNA viruses, the lack of annotation or coding regions on the 3'-5' makes it difficult to interpret the influence of structuredness on this strand. Of the 963 analysed dsRNA genomes, only nine have annotation entries in their Genbank file regarding the location of CDS on the 3'-5' strand. However, the differentiation between CDS and non-CDS on the 5'-3' is based on much more data, and thus should be more reliable. In the dsRNA 5'-3' strand, the mean Z-scores in the non-CDS have a lower mean Z-score than in the CDS, which suggests that the non-coding regions in the genome present more stable secondary structures than the coding regions.

6.3 GC-content, opening energies and structuredness

The ability of a genome to form stable secondary structures is inevitably linked to the GC-content of the nucleotide sequence. Therefore, it was of great interest, related to this work, to determine how the mean GC content and mean Z-score of all the genomes across taxonomic families interplay in the Baltimore groups. Interestingly enough, the results show that a high GC content is not necessarily required to achieve structuredness in either of the analyzed viral data sets. The mean energy values, computed via RNAplfold, determined the amount of free energy needed to unfold a structure, in order to facilitate, for example, the binding of the ribosome for

translation initiation. In the ssRNA(+) viruses, the correlation between the mean opening energy and the mean Z-score for the 30 nucleotide window surrounding the first position of a coding region revealed a Pearson's r of -0.366. The same correlation was performed for the ssRNA(-) and dsRNA group, where the Person's r was 0.010 and -0.348. An even better correlation (Person's r = -0.48) was found between the 30 nucleotide window after the first CDS and the mean Z-score in the ssRNA(+) and dsRNA.

Despite the absence of a corrrelation between Z-score and RNAplfold computed energy values in the ssRNA(-), the genome of the Zaire ebolavirus contains some regions that could potentially present stable secondary structures. In the *Filoviridae* family from the ssRNA(-), all the genes are flanked by conserved gene start and stop signals, that either overlap or are separated by short intergenic regions [51]. The exact structure of their conserved elements is still under research and the results presented in this thesis could shed some light in the structuredness of the filoviruses genomes. For example, in the Zaire ebolavirus, RNALfold predictions on both strands strand show that the intergenic regions between VP30 and VP 24, VP35 and VP40, and VP40 and GP contain at least one stable secondary structure each, whose Z-score is less than -2. These probable structures could play an important role in protein translation and viral replication. Interestingly, as depicted in Supplementary Figure 53, there is no evidence of a correlation between the opening energy and mean Z-scores throughout the whole genomes of the species included in this family. This observation can be extend to the majority of the ssRNA(-) taxonomic families. Thus, it is difficult to come up with a biological explanation or hypothesis as to why the ssRNA(-) behave this way. One potential cause could be the incomplete annotation in some of the segmented viruses and as such their exclusion from the analysis due to missing annotation of the coding region. Some of the viruses from the ssRNA(-) group are the most dangerous pathogens for humans (for example Zaire ebolavirus, Marburgvirus, Lassa virus, Influenzavirus etc) and only dedicated laboratories can perform *in vivo* research on these organisms. As such, the results shown in this work could at least give hints about interesting structured regions that could be the subject of extensive research that may assist in vaccine and/or treatment development upon infection.

For the dsRNA viruses, the GC content values in diverse families have a smaller range than in ssRNA(+) and ssRNA(-). As shown before, the rotaviruses are the most important pathogen in the *Reoviridae* family. Their genomes are not particularly structured, denoted by the low number of locally stable structures predicted by RNALfold throughout their segments. Supplementary Figure 54 also denotes a high affinity of ribosome binding to the mRNA, as the opening energy value become higher the negative the mean Z-scores get. This work may provide the drive to assess in much more detail the genomic architecture of these viruses, based on the local stable predicted structures as well as their very likely structured genomes, which could reveal the function of the secondary structure elements in different parts of their genome.

6.4 MFED and Z-scores as a method to assess global structuredness in other studies

The (minimum free energy difference) MFED as well as Z-score models have been used in other studies to assess the global structuredness of RNA genomes [106, 105]. However, it is still unclear if the computation of MFED in this work and in Simmonds *et al.* is based on the same mathematical formula. In an attempt to reproduce the MFED values in the Simmonds *et al.* paper, the results were inconsistent with the ones reported in the study, which may be due to differences in MFED computation and mathematical formula. The almost perfect correlation between the MFED and Z-score shown here offers the possibility of additionally using MFED as a measure of RNA structuredness, which was expected since the calculation of the Z-score cannot be done without using the MFED (see Materials and Methods). Nevertheless, Simmonds *et al.* also uses the Z-score as a method to determine RNA structures in the genomes of various families in the ssRNA(+) group. The results in Simmonds *et al.* determined that the distribution of the Z-scores is centered around 0 in some genera of the families *Flaviviridae* and *Picornaviridae* such as Enterovirus, Hepatovirus etc., or is slightly left-skewed in the case of other (Hepacivirus, Aphthovirus, etc.). For example, the reported mean Z-scores for Hepacivirus and Flavivirus genera in [106] are -2.5 and 0.39, respectively. The mean Z-scores for the same genera using the approach described in this thesis show values of -0.88 in hepaciviruses and -0.41 in flaviviruses, while the distribution of the mean Z-scores does not follow a Gaussian distribution in all organisms. One important aspect which is very likely to lead to the discrepancy in the reported results is the size of the sliding window length used to scan the genome. In [106], a window length size of 498 is used, which is almost 5 times larger than the window size used in this work. The calculation of MFE and implicitly of MFED depends on the chosen sequence size, so this factor is hypothesized to be the major contributor in the different results. Another reason could be the genomic sequences used for analysis, as it is not mentioned if only RefSeq approved RNA viral sequences were analysed, so the number of viral genomes could may also be a potential factor contributing to the result discrepancy.
Consistently with the results presented in this thesis, the author does acknowledge the global structuredness in the species of the *Coronaviridae* family, which were the subject of analysis in the context of the SARS-CoV-2 pandemic [105]. The size of the sliding window differs in this study too (350 vs 100), but the reported conclusion that the SARS-CoV-2 genome folds into stable secondary structures throughout the genome meets the results described here, as the mean Z-score of the reference SARS-CoV-2 genome (accession number NC045512) is -1.34. These results are also matched with the work presented in [6], where it was reported that the SARS-CoV-2 is highly structured.

6.5 Concluding remarks and outlook

The aim of the research during the master thesis was to assess the structuredness across all available RNA viral genomes. As such, a Python pipeline was developed to obtain, for each genomic position in every taxonomic species, a mean Z-score, which is a robust method for determining the thermodynamic stability of a given sequence. Not only the Z-score, but also MFED, GC content as well as opening

energies were among the methods used to answer the research question. For each taxonomic species, an in depth analysis was performed to assess differences in Z-score distributions between coding and non-coding regions, to annotate the genome, obtain the locally stable secondary structures and compute the free opening energies in the entire genome and in particular coding regions. At a genus level, statistical comparisons between the different distributions in each viruses were performed and at family level, correlations between opening energies and mean Z-scores determined different traits among distinct genera. The results showed, on a global level, that the viruses in the ssRNA(+) group have more structured genomes than ssRNA(-), while the dsRNA viruses are at least structured among all groups. In all categories, the GC content did not necessarily mirror the affinity for secondary structure formation across a genome, as such the analysis showed that despite a low percentage of guanine cytosine base pairs, structuredness can still be achieved. The negative correlation between the opening energies and Z-scores in the 30 nucleotide spanning the beginning of the CDS, as well as in the 30 nucleotide window following it, suggest a thermodynamic feasible RNA-RNA interaction, thus facilitating a qualitative protein translation and gene expression. Intriguingly, the ssRNA(-) group was the only one in which no correlation was found. Apart from the global structuredness computation, for each virus the locally stable secondary structures were obtained, processed, annotated and visualized together with the genome annotation provided in each virus Genbank file. Different traits of structuredness were observed not only at family level, but also inside one taxonomic genus between the different organisms. Differentiating between CDS and non-CDS parts of the genome allowed for a more extensive analysis of how secondary structures are distributed at each genome level. Differences in distributions were observed at genus level, suggesting that there is no common trend inside a taxonomic category for which either region provides better secondary structure folding.

The results obtained in this work match the literature with regard to secondary structures in different RNA viral genomes, which were determined theoretically or experimentally. Thus, corroborating the results with already established patterns of secondary structure formation further validate the work. Not the least, there are still many virus families across all groups which have not yet been under the focus of the research community and the evidence of structuredness in their genomes in still pending; therefore, the analysis conducted in this thesis may provide a starting reference for future research in the field of Virology and RNA bioinformatics.

7 Supplementary material

Figure 50 shows correlation plots between the genome length and the number of predicted RNALfold secondary structures, with a Z-score $\leq$ -4. Figure 51 shows the correlation between the genome length and the abundance of RNALfold hits, normalized for the genome length. Figures 52, 53 and 54 show, for each taxonomic family in each viral group, the relationship between the mean Z-score and opening energy.

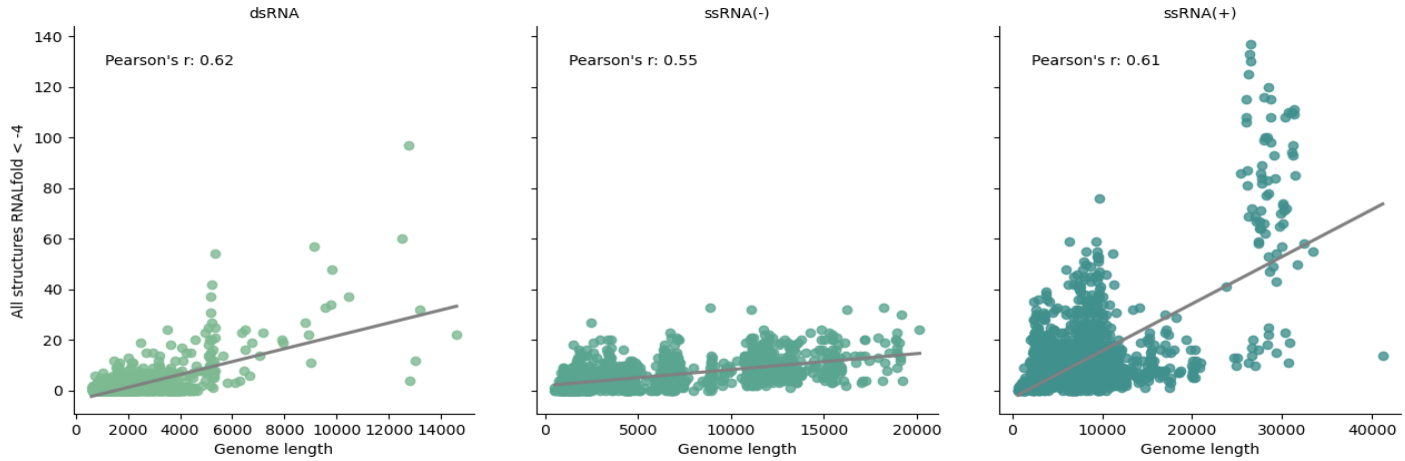

Figure 50: Correlation between genome length and number of RNALfold predictions in all virus groups

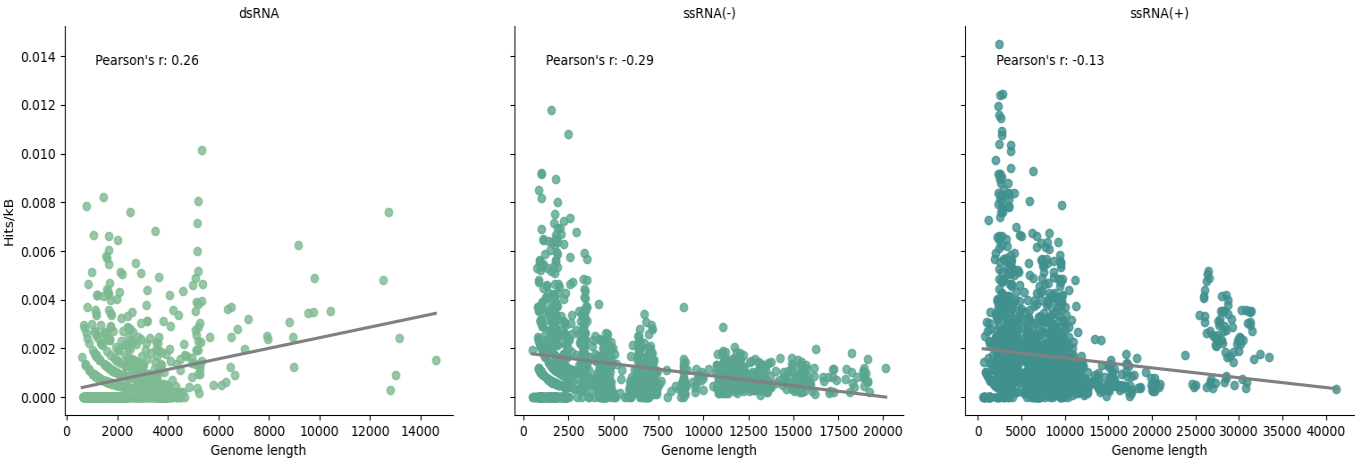

Figure 51: Correlation between genome length and density of RNALfold predictions in all virus groups

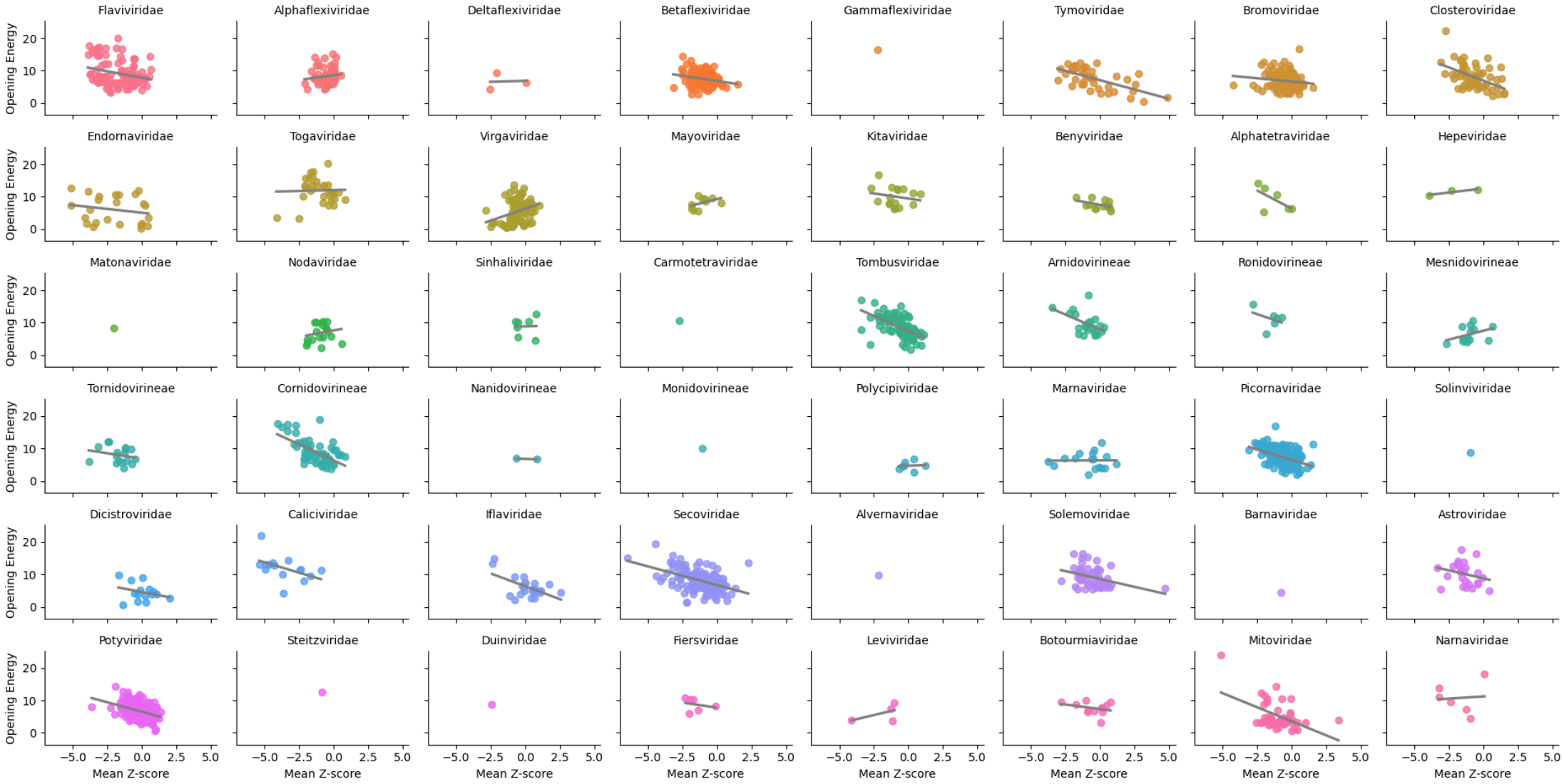

Figure 52: Scatterplot of mean opening energy and mean Z-score per family in ssRNA(+)

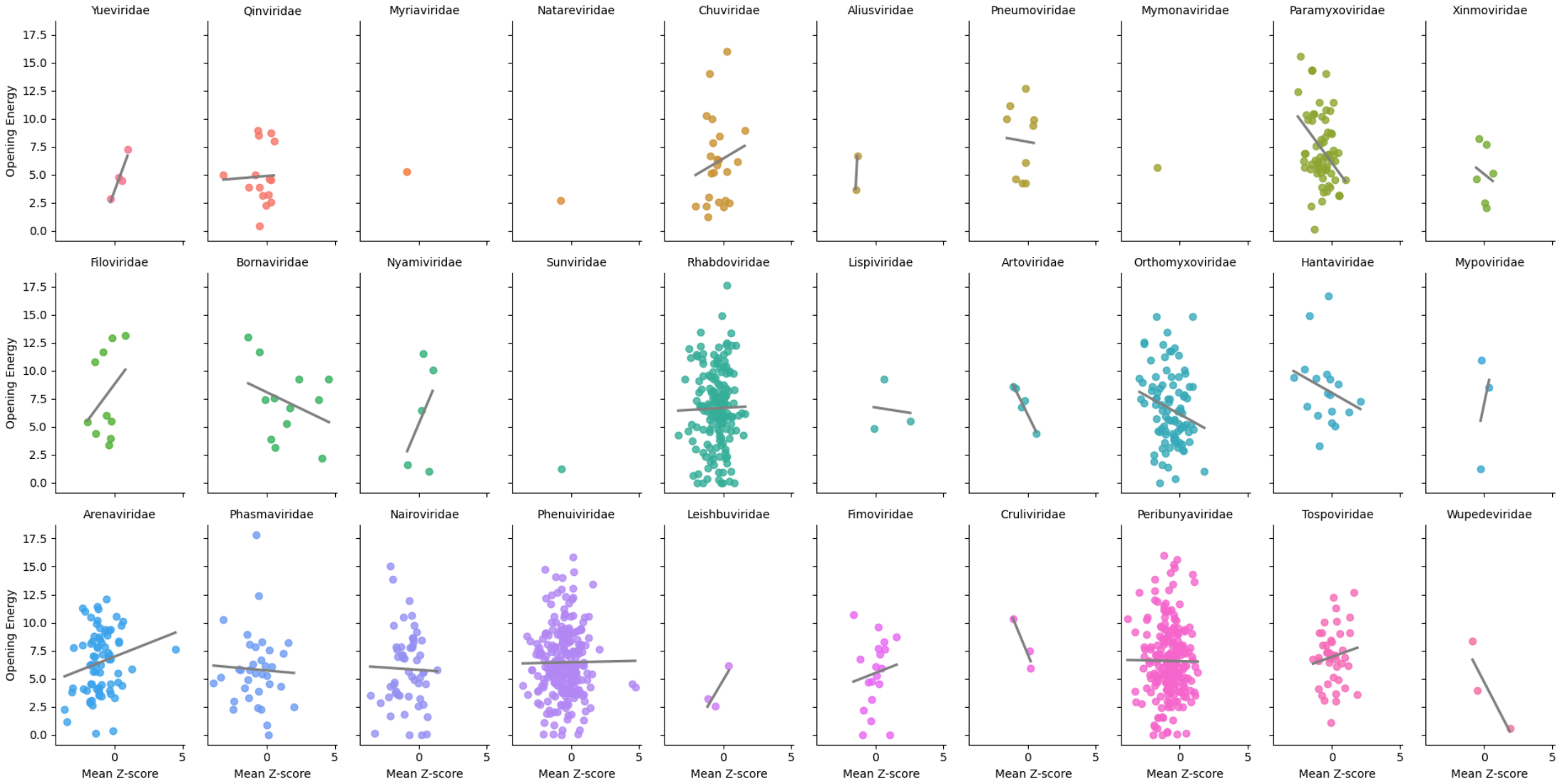

Figure 53: Scatterplot of mean opening energy and mean Z-score per family in ssRNA(-)

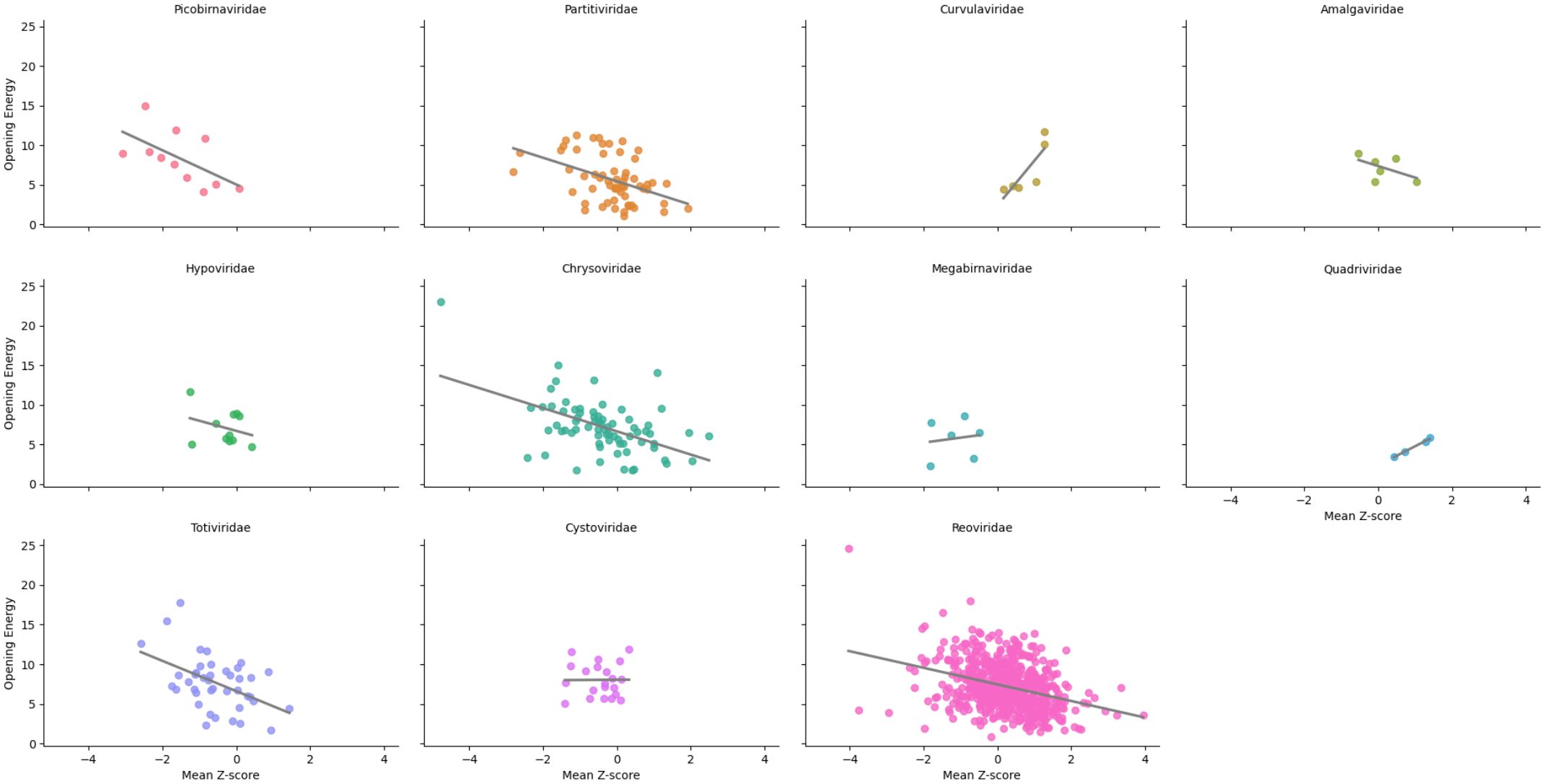

Figure 54: Scatterplot of mean opening energy and mean Z-score per family in dsRNA

References

[1] *Medical Microbiology. 4th Edition.* University of Texas Medical Branch at Galveston., 1996.

[2] Family - reoviridae. In Andrew M.Q. King, Michael J. Adams, Eric B. Carstens, and Elliot J. Lefkowitz, editors, *Virus Taxonomy*, pages 541–637. Elsevier, San Diego, 2012.

[3] Paul Ahlquist, Amine O. Noueiry, Wai-Ming Lee, David B. Kushner, and Billy T. Dye. Host factors in positive-strand RNA virus genome replication. *J Virol*, 77(15):8181–8186, Aug 2003.

[4] Sanaâ Alaoui Amine, Marouane Melloul, Moulay Abdelaziz El Alaoui, Hassan Boulahyaoui, Chafiqa Loutfi, Nadia Touil, and Elmostafa El Fahime. Evidence for zoonotic transmission of species a rotavirus from goat and cattle in nomadic herds in morocco, 2012–2014. *Virus Genes*, 56(5):582–593, Jul 2020.

[5] S Amor, L Fernández Blanco, and D Baker. Innate immunity during SARS-CoV-2: evasion strategies and activation trigger hypoxia and vascular damage. *Clin Exp Immunol*, 202(2):193–209, Oct 2020.

[6] Ryan J Andrews, Collin A O'Leary, Van S Tompkins, Jake M Peterson, Hafeez S Haniff, Christopher Williams, Matthew D Disney, and Walter N Moss. A map of the SARS-CoV-2 RNA structurome. *NAR Genom Bioinform*, 3(2), Apr 2021.

[7] Yuzo Arima, , May Chiew, and Tamano Matsui. Epidemiological update on the dengue situation in the western pacific region, 2012. *WPSAR*, 6(2):82–89, Apr 2015.

[8] Ahmed Abdel Aziz. Chapter 1.1 - hepatitis c virus: Virology and genotypes. In Sanaa M. Kamal, editor, *Hepatitis C in Developing Countries*, pages 3–11. Academic Press, 2018.

[9] Simone Bach, Jana-Christin Demper, Nadine Biedenkopf, Stephan Becker, and Roland K. Hartmann. RNA secondary structure at the transcription start site influences EBOV transcription initiation and replication in a length- and stability-dependent manner. *RNA Biol*, 18(4):523–536, Oct 2020.

[10] D Baltimore. Expression of animal virus genomes. *Bacteriol Rev*, 35(3):235–241, Sep 1971.

[11] Sandra Bamberg, Larissa Kolesnikova, Peggy Möller, Hans-Dieter Klenk, and Stephan Becker. VP24 of marburg virus influences formation of infectious particles. *J Virol*, 79(21):13421–13433, Nov 2005.

[12] Nicolás Bejerman, Humberto Debat, and Ralf G. Dietzgen. The plant negative-sense RNA virosphere: Virus discovery through new eyes. *Front Microbiol*, 11, Sep 2020.

[13] Richard Bellman. On the theory of dynamic programming. *P Natl Acad Sci-Biol*, 38(8):716–719, Aug 1952.

[14] S. H. Bernhart, I. L. Hofacker, and P. F. Stadler. Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22(5):614–615, Dec 2005.

[15] Ruth Bishop. Discovery of rotavirus: Implications for child health. *J Gastroen Hepatol*, 24:S81–S85, Oct 2009.

[16] Ákos Boros, Hajnalka Fenyvesi, Péter Pankovics, Hunor Biró, Tung Gia Phan, Eric Delwart, and Gábor Reuter. Secondary structure analysis of swine pasivirus (family picornaviridae) RNA reveals a type-IV IRES and a parechovirus-like 3' UTR organization. *Arch Virol*, 160(5):1363–1366, Feb 2015.

[17] Flavia Bortolotti, Massimo Resti, Raffaella Giacchino, Carlo Crivellaro, Lucia Zancan, Chiara Azzari, Nadia Gussetti, Loredana Tasso, and Stefania Faggion. Changing epidemiologic pattern of chronic hepatitis c virus infection in italian children. *J Pediatr*, 133(3):378–381, Sep 1998.

[18] Linda Bruslind. Introduction to viruses, 2019.

[19] Leah D.B. Carter and Andrew Aronsohn. Overcoming injustice: A roadmap to improve access to hepatitis c virus therapy for our medicaid patients. *Hepatology*, 65(5):1735–1740, Mar 2017.

[20] Chi Yu Chan, C Steven Carmack, Dang D Long, Anil Maliyekkel, Yu Shao, Igor B Roninson, and Ye Ding. A structural interpretation of the effect of GC-content on efficiency of RNA interference. *BMC Bioinformatics*, 10(S1), Jan 2009.

[21] Shi-Jie Chen. Graph, pseudoknot, and SARS-CoV-2 genomic RNA: A biophysical synthesis. *Biophys J*, 120(6):980–982, Mar 2021.

[22] Nilay Chheda and Manish Gupta. Rna as a permutation. 03 2014.

[23] David P. Clark, Nanette J. Pazdernik, and Michelle R. McGehee. Viruses, viroids, and prions. In *Molecular Biology*, pages 749–792. Elsevier, 2019.

[24] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, Mar 2009.

[25] Nkerorema Djodji Damas, Nicolas Fossat, and Troels K. H. Scheel. Functional interplay between RNA viruses and non-coding RNA in mammals. *Non-Coding RNA*, 5(1):7, Jan 2019.

[26] The Interational HapMap Consortium David Altshuler, Donnelly Peter. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, Oct 2005.

[27] Rafael de Cesaris Araujo Tavares, Gandhar Mahadeshwar, Han Wan, Nicholas C. Huston, and Anna Marie Pyle. The global and local distribution of RNA structure throughout the SARS-CoV-2 genome. *J Virol*, 95(5), Feb 2021.

[28] Pallavi Deol, Jobin Kattoor, Shubhankar Sircar, Souvik Ghosh, Krisztián Bányai, Kuldeep Dhama, and Yashpal Malik. Avian group d rotaviruses: Structure, epidemiology, diagnosis, and perspectives on future research challenges. *Pathogens*, 6(4):53, Oct 2017.

[29] Juana Díez and Jennifer Jungfleisch. Translation control: Learning from viruses, again. *RNA Biol*, 14(7):835–837, Jun 2017.

[30] Jonathan D. Dinman. Programmed –1 ribosomal frameshifting in SARS coronavirus. In *Molecular Biology of the SARS-Coronavirus*, pages 63–72. Springer Berlin Heidelberg, Oct 2009.

[31] Kishore J Doshi, Jamie J Cannone, Christian W Cobaugh, and Robin R Gutell. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for rna secondary structure prediction. *BMC Bioinformatics*, 5(1):105, 2004.

[32] Mariola Dutkiewicz and Jerzy Ciesiołka. Form confers function: Case of the 3'x region of the hepatitis c virus genome. *Wolrd J Gastroentero*, 24(30):3374–3383, Aug 2018.

[33] Orgel Leslie E. Prebiotic chemistry and the origin of the RNA world. *Crit Rev Biochem Mol*, 39(2):99–123, Jan 2004.

[34] Nuno R. Faria, Thomas A. Mellan, Charles Whittaker, Ingra M. Claro, Darlan da S. Candido, Swapnil Mishra, Myuki A. E. Crispim, Flavia C. S. Sales, Iwona Hawryluk, John T. McCrone, Ruben J. G. Hulswit, Lucas A. M. Franco, Mariana S. Ramundo, Jaqueline G. de Jesus, Pamela S. Andrade, Thais M. Coletti, Giulia M. Ferreira, Camila A. M. Silva, Erika R. Manuli, Rafael H. M. Pereira, Pedro S. Peixoto, Moritz U. G. Kraemer, Nelson Gaburo, Cecilia da C. Camilo, Henrique Hoeltgebaum, William M. Souza, Esmenia C. Rocha, Leandro M. de Souza, Mariana C. de Pinho, Leonardo J. T. Araujo, Frederico S. V. Malta, Aline B. de Lima, Joice do P. Silva, Danielle A. G. Zauli, Alessandro C. de S. Ferreira, Ricardo P. Schnekenberg, Daniel J. Laydon, Patrick G. T. Walker, Hannah M. Schlüter, Ana L. P. dos Santos, Maria S. Vidal, Valentina S. Del Caro, Rosinaldo M. F. Filho, Helem M. dos Santos, Renato S. Aguiar, José L. Proença-Modena, Bruce Nelson, James A. Hay, Mélodie Monod, Xenia Miscouridou, Helen Coupland, Raphael Sonabend, Michaela Vollmer, Axel Gandy, Carlos A. Prete, Vitor H. Nascimento, Marc A. Suchard, Thomas A. Bowden, Sergei L. K. Pond, Chieh-Hsi Wu, Oliver Ratmann, Neil M. Ferguson, Christopher Dye, Nick J. Loman, Philippe Lemey, Andrew Rambaut, Nelson A. Fraiji, Maria do P. S. S. Carvalho, Oliver G. Pybus, Seth Flaxman, Samir Bhatt, and Ester C. Sabino. Genomics and epidemiology of the p.1 SARS-CoV-2 lineage in manaus, brazil. *Science*, 372(6544):815–821, May 2021.

[35] Guilhem Faure, Aleksey Y. Ogurtsov, Svetlana A. Shabalina, and Eugene V. Koonin. Role of mRNA structure in the control of protein folding. *Nucleic Acids Res*, 44(22):10898–10911, Jul 2016.

[36] Markus Fricke, Nadia Dünnes, Margarita Zayas, Ralf Bartenschlager, Michael Niepmann, and Manja Marz. Conserved RNA secondary structures and long-range interactions in hepatitis c viruses. *RNA*, 21(7):1219–1232, May 2015.

[37] Peter Friebe and Ralf Bartenschlager. Genetic analysis of sequences in the 3′ nontranslated region of hepatitis c virus that are important for RNA replication. *J Virol*, 76(11):5326–5338, Jun 2002.

[38] Alexander E. Gorbalenya, Luis Enjuanes, John Ziebuhr, and Eric J. Snijder. Nidovirales: Evolving the largest RNA virus genome. *Virus Res*, 117(1):17–37, Apr 2006.

[39] Eli Goz and Tamir Tuller. Widespread signatures of local mRNA folding structure selection in four dengue virus serotypes. *BMC Genomics*, 16(S10), Oct 2015.

[40] Andreas R. Gruber, Sven Findeiß, Stefan Washietl, Ivo L. Hofacker, and Peter F. Stadler. RNAZ 2.0:. In *Biocomputing 2010*, pages 69–79. WORLD SCIENTIFIC, Oct 2009.

[41] Richard L. Guerrant, David H. Walker, and Peter F. Weller. *Tropical infectious diseases*. Saunders/Elsevier, 2011.

[42] Kathryn A. Hanley and Scott C. Weaver, editors. *Frontiers In Dengue Virus Research*. Caister Academic Press, 2010.

[43] Alex S. Hartlage, John M. Cullen, and Amit Kapoor. The strange, expanding world of animal hepaciviruses. *Ann Rev Virol*, 3(1):53–75, 2016. PMID: 27741408.

[44] Paul G. Higgs. RNA secondary structure: physical and computational aspects. *Q Rev Biophys*, 33(3):199–253, Aug 2000.

[45] Paul G. Higgs and Teresa K. Attwood. *Bioinformatics and Molecular Evolution*. Blackwell Publishing Ltd., Dec 2004.

[46] Emma B. Hodcroft. Covariants: Sars-cov-2 mutations and variants of interest. 2021.

[47] Stephen R Holbrook. RNA structure: the long and the short of it. *Curr Opin Struc Biol*, 15(3):302–308, Jun 2005.

[48] E.C. Holmes. RNA viruses, evolution of. In *Encyclopedia of Evolutionary Biology*, pages 476–483. Elsevier, 2016.

[49] M Honda, EA Brown, and SM Lemon. Stability of a stem-loop involving the initiator aug controls the efficiency of internal initiation of translation on hepatitis c virus rna. *RNA*, 1996.

[50] Chantal Hulo, Edouard de Castro, Patrick Masson, Lydie Bougueleret, Amos Bairoch, Ioannis Xenarios, and Philippe Le Mercier. ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res*, 39(suppl_1):D576–D582, Oct 2010.

[51] Adam J. Hume and Elke Mühlberger. Distinct genome replication and transcription strategies within the growing filovirus family. *J mol biol*, 431(21):4290–4320, Oct 2019.

[52] Christon J. Hurst. *Biological Role of a Virus*. Springer International Publishing AG, 2021.

[53] Jennifer Jungfleisch, Danny D. Nedialkova, Ivan Dotu, Katherine E. Sloan, Neus Martinez-Bosch, Lukas Brüning, Emanuele Raineri, Pilar Navarro, Markus T. Bohnsack, Sebastian A. Leidel, and Juana Díez. A novel translational control mechanism involving RNA structures within coding sequences. *Genome Res*, 27(1):95–106, Nov 2016.

[54] Andris Kazaks, Tatyana Voronkova, Janis Rumnieks, Andris Dishlers, and Kaspars Tars. Genome structure of caulobacter phage phiCb5. *J Virol*, 85(9):4628–4631, May 2011.

[55] Jamie A. Kelly, Michael T. Woodside, and Jonathan D. Dinman. Programmed -1 ribosomal frameshifting in coronaviruses: A therapeutic target. *Virology*, 554:75–82, Feb 2021.

[56] Shruti Khare, , Céline Gurry, Lucas Freitas, Mark B Schultz, Gunter Bach, Amadou Diallo, Nancy Akite, Joses Ho, Raphael TC Lee, Winston Yeo, GISAID Core Curation Team, and Sebastian Maurer-Stroh. GISAID's role in pandemic response. *CCDC Weekly*, 3(49):1049–1051, 2021.

[57] Chang Wook Kim and Kyong-Mi Chang. Hepatitis c virus: virology and life cycle. *Clin Mol Hepatol*, 19(1):17, 2013.

[58] Johnson KM. Ebola haemorrhagic fever in zaire, 1976. *Bull World Heath Organ*, 1978.

[59] Satoshi Komoto and Koki Taniguchi. Reverse genetics systems of segmented double-stranded RNA viruses including rotavirus. *Future virol*, 1(6):833–846, Nov 2006.

[60] Eugene V. Koonin, Yuri I. Wolf, Keizo Nagasaki, and Valerian V. Dolja. The big bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. *Nat Rev Microbiol*, 6(12):925–939, Nov 2008.

[61] Richard Kormelink, Maria Laura Garcia, Michael Goodin, Takahide Sasaya, and Anne-Lise Haenni. Negative-strand RNA viruses: The plant-infecting counterparts. *Virus Res*, 162(1-2):184–202, Dec 2011.

[62] Mart Krupovic, Valerian V Dolja, and Eugene V Koonin. Plant viruses of the amalgaviridae family evolved via recombination between viruses with double-stranded and negative-strand RNA genomes. *Biol Direct*, 10(1), Mar 2015.

[63] Tammy C. T. Lan, Matty F. Allan, Lauren E. Malsick, Jia Z. Woo, Chi Zhu, Fengrui Zhang, Stuti Khandwala, Sherry S. Y. Nyeo, Yu Sun, Junjie U. Guo, Mark Bathe, Anders Näär, Anthony Griffiths, and Silvi Rouskin. Secondary structural ensembles of the SARS-CoV-2 RNA genome in infected cells. *Nat Commun*, 13(1), Mar 2022.

[64] Jeffrey E Lee and Erica Ollmann Saphire. Ebolavirus glycoprotein structure and mechanism of entry. *Future virol*, 4(6):621–635, Nov 2009.

[65] Ci-Xiu Li, Mang Shi, Jun-Hua Tian, Xian-Dan Lin, Yan-Jun Kang, Liang-Jun Chen, Xin-Cheng Qin, Jianguo Xu, Edward C Holmes, and Yong-Zhen Zhang. Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *ELife*, 4, Jan 2015.

[66] Hui-Chun Li. Production and pathogenicity of hepatitis c virus core gene products. *Wolrd J Gastroentero*, 20(23):7104, 2014.

[67] Hui-Chun Li. Hepatitis c virus: Virology, diagnosis and treatment. *World J. Hepatol.*, 7(10):1377, 2015.

[68] Wilson Li, Emily Manktelow, Johann C. von Kirchbach, Julia R. Gog, Ulrich Desselberger, and Andrew M. Lever. Genomic analysis of codon, sequence and structural conservation with selective biochemical-structure mapping reveals highly conserved and dynamic structures in rotavirus RNAs with potential cis -acting functions. *Nucleic Acids Res*, 38(21):7718–7735, Jul 2010.

[69] Yuanzhi Liu, Yu Zhang, Mingshu Wang, Anchun Cheng, Qiao Yang, Ying Wu, Renyong Jia, Mafeng Liu, Dekang Zhu, Shun Chen, Shaqiu Zhang, XinXin Zhao, Juan Huang, Sai Mao, Xumin Ou, Qun Gao, Yin Wang, Zhiwen Xu, Zhengli Chen, Ling Zhu, Qihui Luo, Yunya Liu, Yanling Yu, Ling Zhang, Bin Tian, Leichang Pan, and Xiaoyue Chen. Structures and functions of the 3′ untranslated regions of positive-sense single-stranded RNA viruses infecting humans and animals. *Front Cell Infect Mi*, 10, Aug 2020.

[70] Maria F. Lodeiro, Claudia V. Filomatori, and Andrea V. Gamarnik. Structural and functional studies of the promoter element for dengue virus RNA replication. *J Virol*, 83(2):993–1008, Jan 2009.

[71] Ronny Lorenz, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA package 2.0. *Algorithms Mol Biol*, 6(1), Nov 2011.

[72] Ronny Lorenz, Michael T. Wolfinger, Andrea Tanzer, and Ivo L. Hofacker. Predicting RNA secondary structures from sequence and probing data. *Methods*, 103:86–98, Jul 2016.

[73] J. Robin Lytle, Lily Wu, and Hugh D. Robertson. The ribosome binding site of hepatitis c virus mRNA. *J Virol*, 75(16):7629–7636, Aug 2001.

[74] Yashpal S. Malik and Souvik Ghosh. Etymologia: Picobirnavirus. *Emerg Infect Dis*, 26(1), Jan 2020.

[75] Tauqeer Hussain Mallhi, Nida Bokharee, and Yusra Habib Khan. Togaviridae and flaviviridae. In Nima Rezaei, editor, *Encyclopedia of Infection and Immunity*, pages 100–112. Elsevier, Oxford, 2022.

[76] David H. Mathews, Matthew D. Disney, Jessica L. Childs, Susan J. Schroeder, Michael Zuker, and Douglas H. Turner. Incorporating chemical modification

constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *P Natl Acad Sci-Biol*, 101(19):7287–7292, May 2004.

[77] David H. Mathews, Jeffrey Sabina, Michael Zuker, and Douglas H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J mol biol*, 288(5):911–940, May 1999.

[78] Jelle Matthijnssens, Peter H. Otto, Max Ciarlet, Ulrich Desselberger, Marc Van Ranst, and Reimar Johne. VP6-sequence-based cutoff values as a criterion for rotavirus species demarcation. *Arch Virol*, 157(6):1177–1182, Mar 2012.

[79] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, May 1990.

[80] Tim R. Mercer, Marcel E. Dinger, and John S. Mattick. Long non-coding RNAs: insights into functions. *Nat Rev Genet*, 10(3):155–159, Mar 2009.

[81] Peter Mertens. The dsRNA viruses. *Virus Res*, 101(1):3–13, Apr 2004.

[82] Susanne Modrow, Dietrich Falke, Uwe Truyen, and Hermann Schätzl. Viruses with single-stranded, positive-sense RNA genomes. In *Molecular Virology*, pages 185–349. Springer Berlin Heidelberg, 2013.

[83] Elke Mühlberger. Filovirus replication and transcription. *Future virol*, 2(2):205–215, Mar 2007.

[84] Elke Mühlberger, Sabine Trommer, Christa Funke, Viktor Volchkov, Hans-Dieter Klenk, and Stephan Becker. Termini of all mRNA species of marburg virus: Sequence and secondary structure. *Virology*, 223(2):376–380, Sep 1996.

[85] Zachary R. Newman, Janet M. Young, Nicholas T. Ingolia, and Gregory M. Barton. Differences in codon bias and GC content contribute to the balanced expression of TLR7 and TLR9. *P Natl Acad Sci-Biol*, 113(10), Feb 2016.

[86] I.P. O'Carroll and A. Rein. Viral nucleic acids. In *Encyclopedia of Cell Biology*, pages 517–524. Elsevier, 2016.

[87] Roman Ochsenreiter, Ivo Hofacker, and Michael Wolfinger. Functional RNA structures in the 3′UTR of tick-borne, insect-specific and no-known-vector flaviviruses. *Viruses*, 11(3):298, Mar 2019.

[88] Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S.

Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, and Kim D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, 44(D1):D733–D745, Nov 2015.

[89] Nicolas Papageorgiou, Maria Spiliopoulou, Thi-Hong Van Nguyen, Afroditi Vaitsopoulou, Elsie Yekwa Laban, Karine Alvarez, Irene Margiolaki, Bruno Canard, and François Ferron. Brothers in arms: Structure, assembly and function of arenaviridae nucleoprotein. *Viruses*, 12(7):772, Jul 2020.

[90] Janusz T. Paweska and Petrus Jansen van Vuren. Rift valley fever virus. In *The Role of Animals in Emerging Viral Diseases*, pages 169–200. Elsevier, 2014.

[91] Susan Payne. Family reoviridae. In *Viruses*, pages 219–226. Elsevier, 2017.

[92] Kayla M. Peck and Adam S. Lauring. Complexities of viral mutation rates. *J Virol*, 92(14), Jul 2018.

[93] Ewan P Plant, Gabriela C Perez-Alvarado, Jonathan L Jacobs, Bani Mukhopadhyay, Mirko Hennig, and Jonathan D Dinman. A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. *PLoS Biology*, 3(6):e172, May 2005.

[94] Mustafizur Rahman, Sukalyani Banik, Abu S. G. Faruque, Koki Taniguchi, David A. Sack, Marc Van Ranst, and Tasnim Azim. Detection and characterization of human group c rotaviruses in bangladesh. *J Clin Microbiol*, 43(9):4460–4465, Sep 2005.

[95] Meghana Rastogi, Neha Pandey, Astha Shukla, and Sunit K. Singh. SARS coronavirus 2: from genome to infectome. *Resp Res*, 21(1), Dec 2020.

[96] Juan Reguera. Negative single-stranded RNA viruses (mononegavirales): A structural view. In *Encyclopedia of Virology*, pages 345–351. Elsevier, 2021.

[97] M. P. Robertson and G. F. Joyce. The origins of the RNA world. *CSH Perspect Biol*, 4(5):a003608–a003608, Apr 2010.

[98] Christina Roman, Anna Lewicka, Deepak Koirala, Nan-Sheng Li, and Joseph A. Piccirilli. The SARS-CoV-2 programmed -1 ribosomal frameshifting element crystal structure solved to 2.09 a using chaperone-assisted RNA crystallography. *ACS Chem. Biol.*, 16(8):1469–1481, Jul 2021.

[99] Sead Sabanadzovic, Rodrigo A. Valverde, Judith K. Brown, Robert R. Martin, and Ioannis E. Tzanetakis. Southern tomato virus: The link between the families totiviridae and partitiviridae. *Virus Res*, 140(1-2):130–137, Mar 2009.

[100] Anthony Sanchez and Michael P. Kiley. Identification and analysis of ebola virus messenger RNA. *Virology*, 157(2):414–420, Apr 1987.

[101] Eric W Sayers, Jeffrey Beck, Evan E Bolton, Devon Bourexis, James R Brister, Kathi Canese, Donald C Comeau, Kathryn Funk, Sunghwan Kim, William Klimke, Aron Marchler-Bauer, Melissa Landrum, Stacy Lathrop, Zhiyong Lu, Thomas L Madden, Nuala O'Leary, Lon Phan, Sanjida H Rangwala, Valerie A Schneider, Yuri Skripchenko, Jiyao Wang, Jian Ye, Barton W Trawick, Kim D Pruitt, and Stephen T Sherry. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 49(D1):D10–D17, Oct 2020.

[102] Claudia S. Sepúlveda, Sandra M. Cordo, Cecilia A. Vázquez, Cybele C. García, and Elsa B. Damonte. Arenaviruses. In *Encyclopedia of Infection and Immunity*, pages 278–291. Elsevier, 2022.

[103] Ethan C Settembre, James Z Chen, Philip R Dormitzer, Nikolaus Grigorieff, and Stephen C Harrison. Atomic model of an infectious rotavirus particle. *EMBO J*, 30(2):408–416, Dec 2010.

[104] Reed S. Shabman, Thomas Hoenen, Allison Groseth, Omar Jabado, Jennifer M. Binning, Gaya K. Amarasinghe, Heinz Feldmann, and Christopher F. Basler. An upstream open reading frame modulates ebola virus polymerase translation and virus replication. *PLoS Pathog*, 9(1):e1003147, Jan 2013.

[105] P. Simmonds. Pervasive RNA secondary structure in the genomes of SARS-CoV-2 and other coronaviruses. *mBio*, 11(6), Dec 2020.

[106] Peter Simmonds, Andrew Tuplin, and David J. Evans. Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence. *RNA*, 10(9):1337–1351, Jul 2004.

[107] P Somogyi, A J Jenner, I Brierley, and S C Inglis. Ribosomal pausing during translation of an RNA pseudoknot. *Mol Cell Biol*, 13(11):6931–6940, Nov 1993.

[108] Nancy Sullivan, Zhi-Yong Yang, and Gary J. Nabel. Ebola virus pathogenesis: Implications for vaccines and therapies. *J Virol*, 77(18):9733–9737, Sep 2003.

[109] Yoshiyuki Suzuki. A possible packaging signal in the rotavirus genome. *Genes Genet Syst*, 89(2):81–86, 2014.

[110] Seng-Lai Tan. *Hepatitis C Viruses*. Taylor & Francis, 2006.

[111] Jacqueline E. Tate, Anthony H. Burton, Cynthia Boschi-Pinto, and Umesh D. Parashar. Global, regional, and national estimates of rotavirus mortality in children &ltundefined years of age, 2000–2013. *Clin Infect Dis*, 62(suppl 2):S96–S105, Apr 2016.

[112] Mir Md Khademul Islam AB Turjya RR, Khan MA. Perversely expressed long noncoding rnas can alter host response and viral proliferation in sars-cov-2 infection. *Future virol*, 2020.

[113] Kazimierz T. Tycowski, Yang Eric Guo, Nara Lee, Walter N. Moss, Tenaya K. Vallery, Mingyi Xie, and Joan A. Steitz. Viral noncoding RNAs: more surprises. *Gene Dev*, 29(6):567–584, Mar 2015.

[114] Tirth Uprety, Dan Wang, and Feng Li. Recent advances in rotavirus reverse genetics and its utilization in basic research and vaccine development. *Arch Virol*, 166(9):2369–2386, Jul 2021.

[115] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat Methods*, 17:261–272, 2020.

[116] V E Volchkov, V A Volchkova, A A Chepurnov, V M Blinov, O Dolnik, S V Netesov, and H Feldmann. Characterization of the l gene and 5' trailer region of ebola virus. *J Gen Virol*, 80(2):355–362, Feb 1999.

[117] Peter J. Walker, Stuart G. Siddell, Elliot J. Lefkowitz, Arcady R. Mushegian, Donald M. Dempsey, Bas E. Dutilh, Balázs Harrach, Robert L. Harrison, R. Curtis Hendrickson, Sandra Junglen, Nick J. Knowles, Andrew M. Kropinski, Mart Krupovic, Jens H. Kuhn, Max Nibert, Luisa Rubino, Sead Sabanadzovic, Peter Simmonds, Arvind Varsani, Francisco Murilo Zerbini, and Andrew J. Davison. Changes to virus taxonomy and the international code of virus classification and nomenclature ratified by the international committee on taxonomy of viruses (2019). *Arch Virol*, 164(9):2417–2429, Jun 2019.

[118] Stefan Washietl, Ivo L. Hofacker, and Peter F. Stadler. Fast and reliable prediction of noncoding RNAs. *P Natl Acad Sci-Biol*, 102(7):2454–2459, Jan 2005.

[119] M.S. Waterman and T.F. Smith. RNA secondary structure: a complete mathematical analysis. *Math Biosci*, 42(3-4):257–266, Dec 1978.

[120] Michael Weik, Jens Modrof, Hans-Dieter Klenk, Stephan Becker, and Elke Mühlberger. Ebola virus VP30-mediated transcription is regulated by RNA secondary structure formation. *J Virol*, 76(17):8532–8539, Sep 2002.

[121] C. Witwer. Conserved RNA secondary structures in picornaviridae genomes. *Nucleic Acids Res*, 29(24):5079–5089, Dec 2001.

[122] Aiping Wu, Yousong Peng, Baoying Huang, Xiao Ding, Xianyue Wang, Peihua Niu, Jing Meng, Zhaozhong Zhu, Zheng Zhang, Jiangyuan Wang, Jie Sheng, Lijun Quan, Zanxian Xia, Wenjie Tan, Genhong Cheng, and Taijiao Jiang. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in china. *Cell Host Microbe*, 27(3):325–328, Mar 2020.

[123] MinKyung Yi and Stanley M. Lemon. 3′ nontranslated RNA signals required for replication of hepatitis c virus RNA. *J Virol*, 77(6):3557–3568, Mar 2003.

[124] Hao Zhang, Chunhe Zhang, Zhi Li, Cong Li, Xu Wei, Borui Zhang, and Yuanning Liu. A new method of RNA secondary structure prediction based on convolutional neural network and dynamic programming. *Front. genet*, 10, May 2019.

[125] Tiejun Zhao. Long noncoding RNA and its role in virus infection and pathogenesis. *Front Biosci*, 24(4):777–789, 2019.

[126] Valentin Zulkower and Susan Rosser. DNA features viewer: a sequence annotation formatting and plotting library for python. *Bioinformatics*, 36(15):4350–4352, Jul 2020.