Probing the interactions that drive RNA binding and self-association of hnRNPA1 implicated in neurodegeneration

by

Syeda Sakina Fatima

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Science

in

Pharmacy

Waterloo, Ontario, Canada, 2023

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I understand that my thesis may be made electronically available to the public.

Abstract

The heteronuclear ribonucleoprotein A1 (hnRNPA1 or A1) is associated with the pathology of different diseases, including neurological disorders and cancers. In particular, the aggregation and dysfunction of A1 has been identified as a critical driver for neurodegeneration in Multiple Sclerosis (MS). Structurally, A1 includes a low-complexity domain (LCD) and two RNA-recognition motifs (RRMs), and their interdomain coordination may play a crucial role in A1 aggregation. Previous studies propose that RNA-inhibitors or nucleoside analogs that bind to RRMs can potentially prevent A1 self-association. Therefore molecular-level understanding on the RNA recognition by A1 RRMs remains of scientific interest. Although several crystal structures of RNA-bound RRM complexes have been reported in the literature, there are still open questions about which RRM RNA prefers to bind and why only specific RNA sequences tend to bind A1. This thesis aims at probing the structures, dynamics and nucleotide interactions with A1's RRMs using a combination of advanced computational methods. Our research to-date has revealed that adenine and guanine in RNAs (or DNAs), and the key residues from the interdomain linker connecting the two RRM domains contribute significantly for RNA binding to A1 RRMs. Further research will seek to address the impact of RNA length on its binding and how RNA specificities vary between the RRMs. Critical residues for RNA-binding have been identified and their molecular-level insights on their nucleotide preferences have been evaluated. As a final addition, the full-length A1 protein for which a crystal structure in the PDB does not exist, is modeled, to analyze the interactions that occur between the RRMs and the LCD domain that could promote A1's aggregation. Both of A1's known isoforms, isoform A (320aa) and isoform B (372aa) have been modeled and studied, with and without RNA bound to them. Our data suggests that interplay between the LCD and the RRMs may block exposure of critical RNA-binding residues to the

environment when RNA is not already bound to the protein. Taken together, this thesis elaborates on full protein dynamics and nucleotide-protein interactions that may be helpful in designing therapeutics. Nucleotide-based therapies or nucleoside analogs in particular, can be designed based on specific interactions outlined in this thesis.

Acknowledgements

First and foremost, I would like to thank Dr. Aravindhan Ganesan, for his dedicated supervision, much-needed support and for mentoring my graduate studies. His constant reassessment of my methods, valuable suggestions and encouragement, have made this research possible, so he deserves my deepest gratitude. A special thank you to my graduate committee members as well- Dr. William Wong and Dr. Praveen Nekkar, who have ensured I receive their best guidance and feedback, to stay successful during the span of my Masters' program.

I would also like to express my deepest appreciation for previous and current members of ArGan's team, particularly Dr. Mohamed Aboelnga, Maya Petgrave and Shiqi Su, for their constant support and feedback. The friendship extended to me by Maya and Shiqi has been one of the most wonderful outcomes of being in Dr. Ganesan's team.

Research is often impressive when experts with valuable experience come together- I am indebted to my collaborators who entrusted me to conduct valuable experiments. I would like to extend my appreciation to Dr. Michael Levin, Dr. Joseph-Patrick Clarke and Dr. Patricia Thibault, at the University of Saskatchewan, College of Medicine, and Dr. Subha Kalyaanamoorthy at the University of Waterloo, Department of Chemistry.

A very special thank you to my mother, who ensured I would be able to complete my Masters' program and for all her prayers that have made this possible. My father, who always inspired me to accomplish my goals in life and setting an example for what real success looks like. My lovely friends- Maryam, Shafa, Fairooj and Faleeha, for always helping me stay motivated and supported when curveballs felt impossible to handle. Lastly, my husband, Kazim, for being the surprising, but much-needed addition to this journey.

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

List of Tables

Chapter 1

Chapter 3

Chapter 4

List of Abbreviations

CNS- Central Nervous System
DNA- deoxyribonucleic acid
dsRNA – double-stranded RNA
dsDNA – double stranded DNA
hnRNPA1- heteronuclear ribonucleoprotein A1
LCD- Low complexity domain
LLPS- liquid-liquid phase separation
MD- Molecular dynamics
MS- Multiple Sclerosis
ns- nanosecond
ps- picosecond
RBP- RNA-binding protein
RNA- Ribonucleic acid
RNAO – RNA oligonucleotide
RNP- ribonucleoprotein
RRM- RNA recognition motif
SDS- Sodium dodecyl sulfate
ssRNA – single stranded RNA
ssDNA -single stranded DNA
UP1- unwinding protein 1

All nucleotides have been abbreviated as: A (Adenine), G (Guanine), C (Cytosine) and U (Uracil)

All amino acids have been abbreviated using their three-letter form. For example, Arginine 92 has been abbreviated as Arg92.

Chapter 1: Introduction

1.1 The essential roles of RNA-binding proteins and their pathophysiological significance

The widely accepted Central Dogma of Molecular Biology, which explains the flow of information in a living cell, recognizes the significance of ribonucleic acids (RNAs) for cell growth and homeostasis, where disruption to RNA metabolism would be detrimental.[1] However, RNA is a relatively unstable molecule in the cells, and often relying on proteins, called the RNA binding proteins (RBPs) to form ribonucleoprotein complexes.[1,2] RBPs are involved in various cellular roles including, but not limited to: DNA replication, RNA metabolism, regulation of transcriptional and post-translational gene expression, and immune response moderation.[3,4] Given their diverse physiological importance, dysfunctions of RBPs are implicated in an array of diseases including neurodegenerative disorders, diabetes, and cardiovascular diseases.[2,5,6]

Structurally, RBPs have evolutionarily conserved RNA-binding domains such as RNA recognition motifs (RRMs), the K-homology (KH) domain, and zinc-finger (ZF) domains that support specific RNA recognition.[44] In addition to well-folded RNA-binding domains, most RBPs also have an intrinsically disordered region known as a low-complexity domain (LCD) or a prion-like domain, which lacks a defined secondary structure.[7,10,11] The LCDs from RBPs are known to self-aggregate and form fibrils following the liquid-liquid phase separation (LLPS).[8,9] Recently, there is growing evidence on the roles of LLPS in the biogenesis of membrane-less organelles such as cajal bodies, nucleoli and stress granules.[10] During LLPS, biomolecules (e.g., proteins and RNAs) tend to interact with each other and form gel-like condensates that involve in various important physiological mechanisms such as cellular stress responses.[11] Nevertheless, post-LLPS, RBPs with prion-like domains undergo self-association and form amyloid-like fibrils[11,12] that are often linked with neurodegenerative conditions such as amyotrophic lateral sclerosis, frontotemporal dementia, Alzheimer's disease, Parkinson's disease, and Multiple Sclerosis (MS)

to name a few.[13,18,19,20,21] Interestingly, recent research has revealed that RNA-binding RRMs interact with the LCD segments during LLPS,[10] and that binding of RNA to RRMs attenuate LCD aggregation.[14] However, the molecular link between the two domains of RBPs still remain unclear. Until now, ~70 RBPs with an LCD have been identified in humans, which include FUS, TDP-43 and hnRNP A/B family of proteins.[15,16]

1.2 HnRNPA1- A key player in cellular metabolism

This research will focus on gaining molecular-level insights into hnRNP A1 (or A1) protein from the hnRNP A/B family.[17,18] The hnRNP A/B family represents a group of highly conserved RBPs, specifically hnRNP A1, A2/B1, A3 and A0, which are linked with diverse cellular functions and human diseases.[6,13] Of these proteins, the A1 protein is more abundant in the 40S ribonucleoprotein complex, and its role in the moderation of RNA homeostasis and messenger-RNA (mRNA) metabolism have been well characterized.[19] For example, hnRNP A1 is known to associate with promoter sequences and is involved in transcriptional initiation and regulation.[19] When A1 binds to promoters for genes coding for thymidine kinase (TK)[20], γ-fibrinogen[21] and the vitamin D receptor, transcription is blocked.[22] However, it is known to be an activator for certain promoters of genes including ApoE.[23] While the RNA-binding capacity of A1 has been well-studied, the DNA-binding capabilities of A1 is also of interest. G-quadruplex structures that exist in DNA as a result of repeats of Gs, are known to be destabilized by A1 to allow for transcription initiation.[24] G-quadruplex structures in DNA which A1 is known to bind, such as the KRAS and c-myc promoters, to allow their transcription initiation.[25,26]

Telomeric repeats that consist of TTAGGG sequences in vertebrates also exist as G-quadruplex structures.[27] These repeats exist to protect the ends of chromosomes from degradation and genetic loss.[27] When telomeres are shortened excessively, the capacity for cell division may

be reduced.[28] On the contrary, failure to maintain telomeres mediates chromosome instability and malignant transformation in cells. [28] Cell development often times uses telomerase, an enzyme that elongates telomeric repeats, to grow the ends of chromosomes.[27] Abnormal cells such as cancer cells, upregulate this process, allowing for abnormal chromosomal growth with telomerase.[27]

The role of A1 in mediating telomere maintenance is critical, as proven with *in vitro* assays wherein reduced A1 protein in human cells reduces telomerase activity.[29] Similarly, A1-defeicient mouse models display smaller lengths of telomeres, which get elongated with the addition of A1.[30] A1 is known to play a role in telomeric maintenance and elongation by binding to the G-quadruplex structures of telomeres, the RNA component of telomerase and TERRA RNA from the telomere complex.[31] Recent evidence also suggests that A1 mediates end-capping of the telomeres when it phosphorylated by the DNA-PK kinase.[32] Phosphorylated A1 contributes to the removal of TERRA from telomeric ends, which in turn, allows for efficient replication and extension of the S-phase of the cell cycle.[32]

A1 has an M9 nuclear localization sequence that allows for its shuttling properties such that it can aid in nuclear export of mRNA.[33] It can associate with nucleolar and cytosolic poly(A)+ RNA[33] and is a critical member of the hnRNP complex that helps translocate mature mRNA transcripts across nuclear pores.[34] As an example, thorough *in vitro* studies utilizing electron microscopy studies and light sheet microscopy showed that A1 potentially binds to the giant Balbiani ring mRNA in Chironomus tentans to translocate the mRNA to the cytosol.[35,36] Indeed, actinomycin-mediated inhibition of RNA polymerase II in mouse embryos and in HeLa cells indicate that nuclear import of A1 is triggered when mRNA is synthesized in the nucelus.[37,38]

In addition to regulating the expression of other proteins by binding mRNA, A1 autoregulates its own expression.[39] Suzuki and Matsuoka, in their study from 2017[39] showed that A1 has the

capacity to inhibit the splicing of intron10 in the A1 pre-mRNA.[39] They proposed that since unspliced mRNA is degraded by nucleases such as the Xrn2 exonuclease, A1 results in downregulation of its own expression.[39] However, autoregulation of A1 does not occur in the 3' or 5' UTRs, unless a potential RNA-binding protein Quaking, is involved.[40] Quaking has previously been shown to bind the 5' and 3' UTRs of A1 mRNA and encourage its stabilization.[40] The possibility of Quaking being involved in A1 autoregulation is yet to be further explored.[39]

Implications of A1 autoregulation could be linked to an array of diseases.[39] In fact, overexpression of other RBPs such as TDP-43 and FUS, that also possess their own autoregulation mechanisms, results in cell death.[41,42] Certain neurodegenerative diseases have also been linked to errors in autoregulation of TDP-43 and FUS.[43,44] Indeed, it may be reasonable to propose that based on protein structure and function similarity, errors of A1 autoregulation may contribute to disease just like TD-43 and FUS.[39] For example, A1 has been shown to be upregulated in many cancers and gliomas.[45] This may be because cancer cells require A1 to accelerate RNA metabolisms and hence protein expression to enhance cell proliferation.[45] Therefore, it can be hypothesized that in normal cells, however, enhanced A1 expression may be cytotoxic.[39] Whereas reduced levels of A1 are often observed in patients with Alzhiemer's disease and ALS patients who have TDP-43 aggregation.[46,47] This suggests that A1 may be downregulated in certain neurodegenerative conditions. Therefore, A1's autoregulation and maintenance of adequate A1 expression levels is essential for healthy cell metabolism.[39]

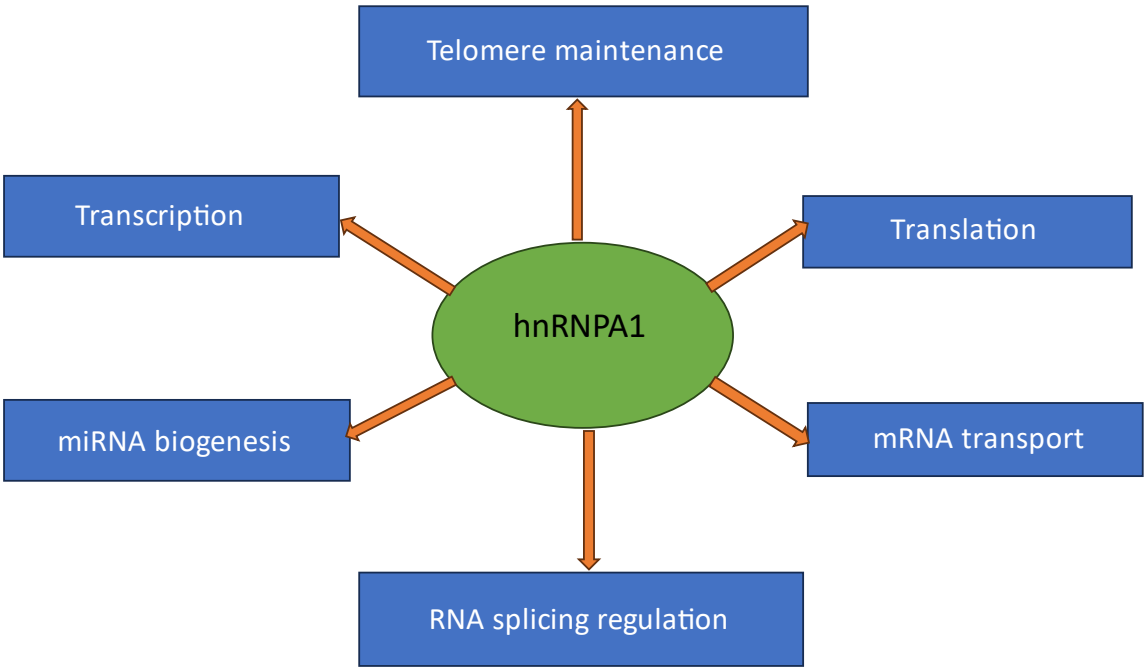


Figure 1.1: A1's major roles in the cell involving nucleic acids.[48] A1 has essential contributions to RNA metabolism, where A1 dysfunction could result in an array of concerns for cell metabolism.

1.3 hnRNPA1's roles in viral replication

A1's roles in gene expression also extend to mechanisms of viral nucleic acid replication. Viral replication depends on a number of host factors to propagate the infection effectively in host cells.[49] While the interactions between the host factors and viral genetic material is complex, there are some host proteins well-characterized that often make-up the hijacked host machinery.[49] One such example is the A1 protein, known to be involved in the propagation and sometimes, the prevention of replication, for an array of pathogenic human viruses.[49] Viruses such as the human rhinovirus,[50] Enterovirus,[51] Sindbis virus,[52] hepatitis C virus,[53] human papilloma virus[54] and even the HIV-1 virus.[55] Indeed, A1's RNA-binding preferences have been studied using the HIV-1 viral transcripts and their interactions with A1.[56] Interestingly, A1 also enhances IRES-mediated

translation initiation of mRNA from viruses[57] and may also support the export of viral RNA from the nucleus.[58]

A1 may play a pro-viral or an anti-viral role depending on the type of virus.[49] For example, during an enterovirus 71 (EV-71) infection, A1 localizes in the cytoplasm to act as an internal trans activating factor (ITAF) by interacting with the 5'UTR in the stem loop of the internal ribosome entry site (IRES), allowing IRES-mediated translation of viral genetic information.[48] Studies have also proven that the stem loop structure adopts a modified conformation to encourage the formation of an A1-RNA complex. Viral replication is impaired if mutations or deletions occur in the stem loop domain.[48] In fact, knockdown of A1 and A2B1, a second member of the hnRNP family, attenuates viral replication completely.[59,60] In contrast, A1 plays a more antiviral role in the replication of the Hepatitis C virus (HCV).[49] A1 has capacity to bind to the HCV RNA-dependent RNA polymerase (NS5B).[61] Increasing A1 expression in Huh-7 cells reduced HCV RNA synthesis, which was rescued after A1 silencing.[62]

1.4 hnRNPA1 and neurodegeneration

A1 is mainly expressed in the central nervous system and aggregation of A1 has been shown as an important basis for neurodegeneration in MS[63] that affects over 2 million people worldwide.[64] MS is a neurodegenerative and autoimmune disease resulting in demyelination of the central nervous system (CNS) and formation of plaques comprising of T cells, macrophages, accompanied by pro-inflammatory cytokines.[63,64] Inflammation eventually damages oligodendrocytes and causes demyelination, disrupting neuronal message transmission and conduction.[65] As damage accumulates, it also becomes irreversible, slowly progressing over the years, with patients losing motor and sensory control.[65] However, it was commonly accepted that autoimmunity causing an attack on the myelin of neuronal axons in the CNS resulted in demyelination.[66] Whereas recent

evidence indicates that neurodegeneration occurs first, leading to an immune response, which results in autoantibodies and further damage to the CNS.[63,66] This is verified with proteins not related to myelin often having autoantibodies produced against them in MS.[63] Unsurprisingly, autoantibodies to A1 have been identified in MS patients, which can be a result of its mis localization from the nucleus to the cytoplasm and aggregation.[63,64,66] To further exacerbate the problem, hnRNPA1 being a key player in cellular stress responses, is dysfunctional if aggregated, so when cellular stress occurs in neurodegeneration, hnRNPA1 is not available to help combat the damage.[64] Henceforth, inflammation continues to increase during the course of MS, with most treatment options being immunomodulatory drugs that do not address the underlying issue of protein aggregation.[66,67]

Apart from aggregation, A1 also has additional pathophysiological implications in neurons. Disruptions to RNA metabolism due to A1's dysfunction has been characterized in an array of neurodegenerative diseases such as Alzheimer's disease (AD) where reduced A1 protein expression has been noted.[68] Although the mechanism is not fully understood, it is well-established that A1 is able to perform alternative splicing on the APP gene which codes for Amyloid beta.[69] To date, most research has coincided with Amyloid beta protein aggregated in AD, resulting in plaque.[69] Therefore, A1 may indirectly influence the pathogenesis of AD by affecting the expression of the APP gene coding for Amyloid beta.[69] Interestingly, in mouse models, inducing a loss of A1 also results in reduced cognitive function.[70]

Spinal muscular atrophy (SMA), amyotrophic lateral sclerosis (ALS), fronto-temporal lobar degeneration (FTLD), HTLV-I associated myelopathy/tropical spastic paraparesis (HAM/TSP) and hereditary spastic paraparesis (HSP) are example of other neurodegenerative diseases that A1 is known to have pathological implications for.[71] In SMA, loss of motor neurons

in the spinal cord results in muscle atrophy in the body.[72] In a healthy body, the Survival Motor Neuron 1 (SMN) protein is coded by the SMN1 gene.[72] An SMN2 paralogous gene, produces low amounts of the SMN protein, which is because A1 acts as a repressor to the protein.[72] Inhibition of A1, has thus been suggested as a treatment for SMA by allowing SMN2 gene expression to allow for sufficient levels of SMA in a patient.[72] In fact, the use of antisense oligonucleotides that masked the splicing regulatory sequence on SMN2 that is recognized by A1, allowed the production of functional SMN proteins.[73]

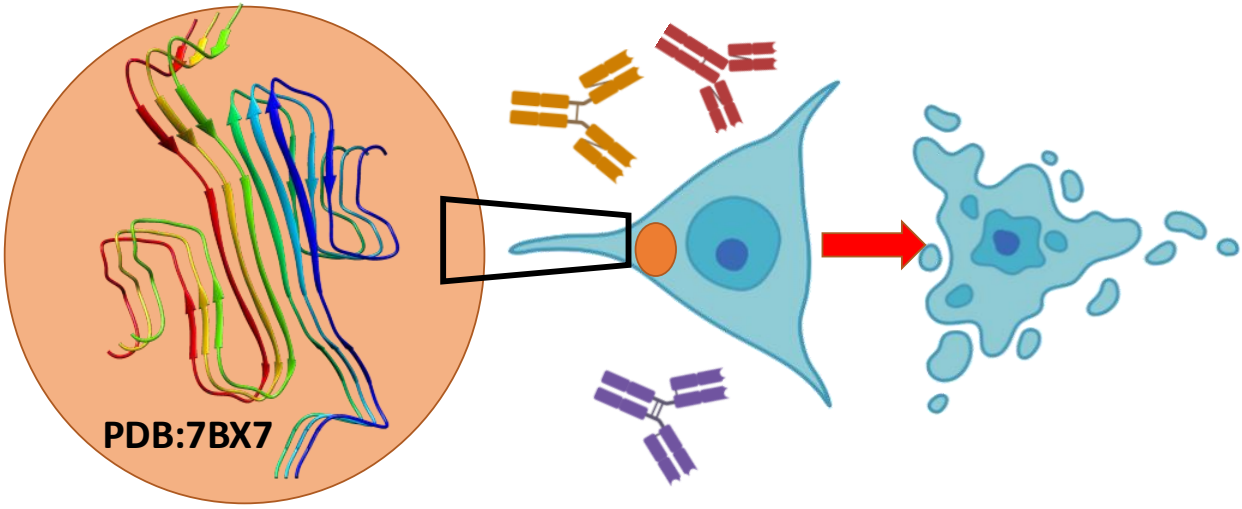


Figure 1.2: A1 aggregation in the cell results in autoimmunity and neuronal loss.[63] An amyloid structure formed by the LCD domain of A1 derived via cryo-EM (PDB:7BX7[74]) demonstrates the capability of A1 to aggregate. This aggregation mechanism and mis localization of A1 to the cytoplasm is thought to be responsible for triggering an immune response.[63] Antibodies produced by the immune system in order to resolve A1 aggregates often results in inflammation and neuronal loss.[63]

1.5 HnRNPA1 and cancer

Aside from hnRNPA1's aggregation propensity, it has also been heavily implicated in an array of cancers such as endometrial cancer[75], bladder cancer[76], gastric cancer[77], liver cancer[78] and prostate cancer[79,80]. This is usually accomplished by A1 having the capacity to either upregulate tumour-

promoting proteins and/or RNAs or downregulating tumour-suppressing proteins and/or RNAs.[77,81,82] Recently, it has also been shown that A1 increases aerobic glycolysis in multiple myeloma cells, thus increasing cancer proliferation by upregulating pyruvate kinase M2.[81]

Unsurprisingly, direct inhibition of A1 has been proposed as a therapeutic option for cancers and even neurodegenerative diseases.[19,79] Indeed, downregulation of A1 via small molecule drugs has been proven useful for castration resistant prostate cancer *in vitro*.[79] Natural tumour-suppressor RNAs such as the miR-490, which suppresses gastric cancer, has been shown to be downregulated by A1, thus promoting tumour proliferation and metastasis.[77] Therefore, the use of RNA-based therapies and small molecule drugs that can inhibit A1 pose desirable outcomes for treating cancer and metastasis. The splicing activity of A1 has also been targeted for treating

1.6 Structural insights on hnRNPA1

It is essential to understand the structure, dynamics and oligonucleotide interactions of A1 to be able to prevent its aggregation and inhibit it in cancers where deemed necessary.[19] The A1 gene is localized on chromosome 12q13.13 and has two major isoforms in the cell namely A1-A (320 amino acids (aa), 34 kDa) and A1-B (372 aa, 38 kDa) due to the differences in mRNA splicing.[6] An alignment of the structures of the two isoforms are provided in Figure 1A for comparison. Previous reports suggest that the A1-A is more abundant than the other isoform.[48] A1 consists of two N-terminal RRM domains (RRM1 and RRM2) and a C-terminal intrinsically disordered LCD with a nuclear localization sequence (Figure 1B).[6] The RRMs are primarily for RNA-binding, collectively known as unwinding protein 1 (UP1), whereas the LCD is associated with self-interactions (or aggregation), as described in Figure 1B.[6] Each RRM has two sub-motifs (RNP1 and RNP2) specific for binding RNA through aromatic stacking and electrostatic interactions with nucleotides (Figure 1C).[48] In RRM1, the RNP1 is comprised of residues 55-

RGFGF-61 and RNP2 is comprised of residues 15-KLFIG-20.[48] Similarly, in RRM2, RNP1 spans residues 146-RGFAF-152 while RNP2 spans residues 106-KIFVG-111 (Figure 1C).[6] A1's RRMs have a 35% amino acid identity and ~60% similarity.[6] Despite the high sequence similarity, the two RRMs are known to be functionally unique with respect to RNA alternative splicing.[83] Until now, several experimental structures of unbound and RNA/DNA-bound A1 RRM complexes have been resolved through X-ray crystallography or solution NMR (see Table 1). In all the reported structures of A1 RRMs in the protein data bank (PDB), the binding modes and the nature of interactions between the residues of A1 and the oligonucleotides were highly conserved, which describe the high specificity in RNA recognition by A1. It was noted that, in almost all the complexes, an adenine (A) and a guanine (G) stack against specific aromatic residues of A1 (Phe17, Phe59, and His101), and hydrogen-bonds (H-bonds) with Arg92 (Figure 2A). However, despite these conserved interactions, the dissociation constants still vary amongst ligands due to factors such as length of the RNA and its base-pairing.[84] For example, Kooshapur et. al[33] demonstrated that a flexible 7-mer RNAO had a lower dissociation constant than a longer 12-mer RNAO (K_D = 3.4 μM and 15.5 nM, respectively).[33] Interestingly, both were derived from an 18-mer miRNA that is A1's biological ligand and both contained the key AG motifs that A1 is known to recognize and bind.[33] Whereas the original 18-nt miRNA ligand has a K_D that falls in between-147 nM.[33] This indicates that even though all three RNAOs contain the same sequence in the binding site, but due to varying lengths, they have different affinities which do not have an obvious pattern. Therefore, exploring and analyzing A1's binding to various types of nucleotides could reveal key insights that *in vitro* data found in literature has not been able to explain thus far.[84] Further, it is interesting to note that, in all the reported experimental structures of bound-A1 RRMs, the oligonucleotides always bound within the binding site of RRM1 domain and not with the

RRM2 (Figure 2B). Efforts to downregulate A1's roles in cancer cells by targeting the RRM1 domain have also demonstrated efficiency.[79] This bias hinted at the RRM1 domain being of primary significance with respect to drug-design, thus this research focuses on RRM1 for RNA-A1 binding (Chapter 3).

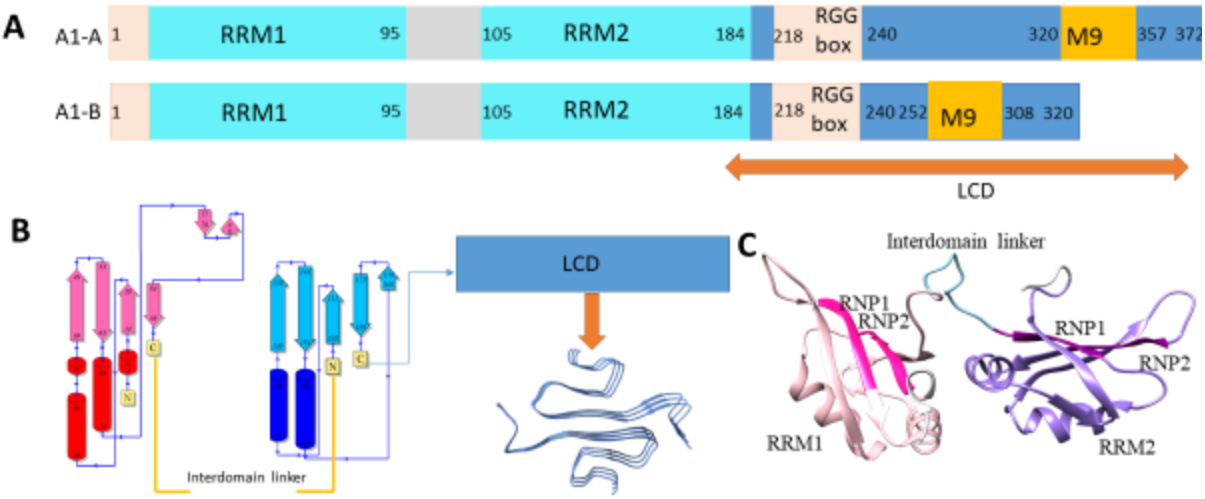


Figure 1.3: The structure of the A1 protein. (A) The two isoforms of A1 vary in length by a few amino acids in the LCD, resulting in isoform B (A1-B) being longer than isoform A (A1-A). Both isoforms have two RNA-binding domains- RRM1 and RRM2, separated by a linker (grey box). The LCD domain contains the RGG-box to bind nucleic acids and the M9 nuclear localization sequence. (B) Structurally, the RRMs fold into beta-sheets while the linker region connecting the two remains a loop. The entire LCD is disordered, which means a defined secondary structure for it is not known. The LCD has capacity to arrange into ordered fibrils (PDB: 7BX7). (C) A crystal structure of A1 with just the RRMs (PDB: 4YOE), demonstrates visual insights into the RRMs. Each RRM has RNP motifs to create a binding pocket for RNA.

With regards to the C-terminal LCD of A1, it is composed of ~173 residues that is rich in RGG sequences, termed RGG boxes.[19] A1 LCD sequence is mostly polar in nature with a few hydrophobic aromatic, and charged residues.[19] A recent study confirmed the main driving forces during LLPS of A1 are aromatic-aromatic and aromatic-arginine interactions. Martin et. al[85] also reported that phenylalanine or tyrosine residues in the LCD that actively play a role in phase separation via deletion constructs.[85,86] Recently, an amyloid-like fibril structure of a segment of A1 LCD (residues 251-295) was resolved using the cryo-EM technique (PDB: 7BX7).[13] This

structure describes that 6 monomers A1 LCD chains are packed in a 3X3 manner (as chains ABC facing chains DEF, see in Figure 3) through β-sheet stacking –a well-known phenomenon of protein aggregation– that is stabilized by electrostatic interactions (of Tyr266, Asp262, Arg284) and hydrophobic contacts (Phe254, Phe263, Phe273, Phe275). This structure provides some insights into A1 LCD fibril architecture. Nevertheless, until now, the complete model of A1 is not available, which limits our understanding on the interplay between RRMs and LCD in A1. For example, binding of ATP to RRMs improves the overall thermal stability of A1, as demonstrated in a previous study using a thermal shift assay.[16] Further, binding of RNAOs on RRMs have also shown to reduce LCD aggregation in A1[14], which suggests an allosteric link between the two domains. Therefore, availability of a full model of A1 can provide a useful tool to understand the interdomain contacts in A1 and the implication of RNA-binding on their interactions. These insights may be relevant for understanding the key drivers of A1's pathological aggregation mechanisms. Chapter 4 of this thesis focuses on the aspect of self-association using the full model of the A1 protein.

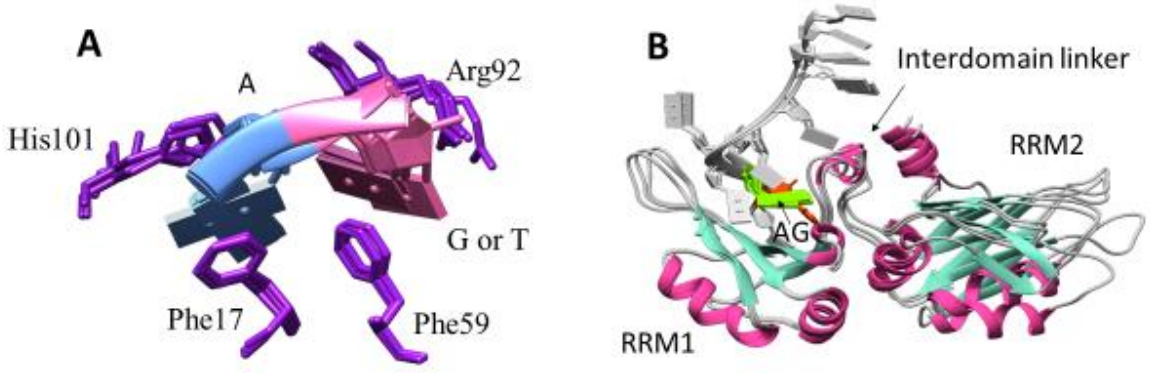


Figure 1.4: The redundancy in RNA-A1 binding found in crystal structures. (A) Adenine tends to stack with Phe17 and His101 while a Guanine or a Thymine tends to stack with Phe59 and Arg92. (B) In the monomeric form of UP1, any DNA or RNA ligand tends to bind to RRM1 instead of RRM2 as seen as this alignment of crystal structures and NMR structures listed in Table 1.

Table 1: Type of ligands bound to the various UP1 structures attained by X-ray crystallography or NMR spectroscopy found in the PDB. A list of structures where UP1 has nucleotide ligand bound has ben provided and surprisingly, all have nucleotide ligands bound to RRM1.

PDB ID	Method	Residues	Ligand
1HA1	X-ray	1-184	ssDNA
4YOE	X-ray	1-196	ssRNA
1PGZ	X-ray	2-196	ssDNA
1PO6	X-ray	8-190	ssDNA
1U1K	X-ray	1-196	ssDNA
1U1L	X-ray	1-196	ssDNA
1U1M	X-ray	1-196	ssDNA
1U1N	X-ray	1-196	ssDNA
1U1O	X-ray	1-196	ssDNA
1U1P	X-ray	1-196	ssDNA
1U1Q	X-ray	1-196	ssDNA
1U1R	X-ray	1-196	ssDNA
1UP1	X-ray	3-184	ssDNA
2LYV	NMR	2-196	ssDNA
2UP1	X-ray	8-190	ssDNA

1.7 Liquid-liquid separation and the aggregation propensity of A1

Understanding the aggregation propensity of A1 requires the understanding of A1's structure and cellular processes that may trigger its dysfunction.[86] A common mechanism frequently mentioned in literature in regards to proteins with low-complexity domains is liquid-liquid phase separation (LLPS).[10] LLPS is a cellular process naturally occurring to form gel-like condensates in the

cytosol for the purpose for compartmentalization.[10] Membrane-less organelles such as nucleoli, P-bodies and stress granules are known to form via LPS. Liquid–liquid phase separation (LLPS) mediates the extensive compartmentalization of cells and leads to the formation of membrane less organelles including nucleoli, stress granules and P bodies amongst many others.[87] Other condensates in the cells with a biological role but not necessarily considered organelles, include heterochromatin.[88] Phase separation can occur in the cytosol involving various macromolecules and may sometimes become irreversible.[10] Multiple multivalent interactions involving biological macromolecules may make some condensates difficult to resolve.[10] Such types of phase separation often involve proteins with an LCD.[86]

The LCD region in proteins often have a unique composition wherein aromatic residues are interspersed between polar residues.[85] Martin et. al[85] recently designed a model called the stickers and spacers framework, based on associative polymers.[89] In their model, they describe stickers as individual residues or even motifs that are mostly responsible for the clustering behaviour of LCD domains.[85] Specifically for A1, they point out aromatic residues present across the LCD that contribute to A1's LCD self-aggregation. Spacer motifs in this model are residues in between the stickers. While the stickers alone are solely involved in noncovalent intra and intermolecular interactions that encourage phase separation of the protein.[90] If the threshold concentration for phase separation is met and the system is saturated enough, phase separation may occur.[90]

While the stickers and spacers in A1 have been well characterized[85], experimental insights into A1's aggregation mechanism has revealed two key observations: A1 can phase separate into condensates *in vitro* and solubilizing the folded RRM domain can increase its phase separation.[86] At high salt concentrations, the full A1 protein can form gel-like droplets because the RRM domains are solubilized by the salt, allowing the LCD to cluster in itself.[86] On the contrary, low

salt concentrations allow for the stickers in the LCD, which are mostly aromatic residues, to mediate compaction of the LCD.[86]Therefore, although the exact mechanism of A1's aggregation is not clear, it is linked to the capacity of the LCD to become compact in a certain environment and preventing its compaction may allow the full protein to stay functional.

1.8 Hypothesis

Since A1 has implications in an array of neurodegenerative diseases[64,66], cancers[45], viral replication[49] and has general aggregation propensity[74], it poses as an attractive target for therapeutic purposes. Drug-design targeting A1 requires greater understanding of how the protein interacts with its native ligands and any interdomain contacts that would affect these native interactions.[91] Detailed insights on how A1 recognizes different types of nucleic acids and its interdomain contacts are not outlined in literature. We hypothesize that deeper insights into RNA recognition by A1 RRMs and the RRM-LCD interdomain interactions in A1 can be gained by probing their structures at molecular level.

It has been proven that computational methods such as molecular modelling and molecular dynamics (MD) methods have been proven as efficient tools for predicting the 3D structures and the dynamic properties of biomolecules and macromolecular complexes to a high-level of accuracy.[92,93] Therefore, we hypothesize that our computational approaches will help in building comprehensive models of RNAO-RRM complexes (objective 1) and the full-length A1 (objective 2), and answering the outstanding questions on the dynamic interactions of RNAOs with A1 RRMs, RNAO-specificity to RRM1 and the structure-dynamics of full-length A1 protein.

Specific objectives for the thesis:

1. **Identifying the key nucleotide-amino acid interactions driving RNA-binding to A1 RRMs.** As illustrated in Figure 1.4 and Table 1, A1 has redundant native contacts found in crystal structures reported to date. However, RNAOs that possess the key AG motif alone do not always display the highest K_d values *in vitro*.[84] This hints at the notion that the AG motif, although necessary, is not the only enhancer for binding affinity between RNA and A1. Additional contact sites in the protein, that may increase A1's binding to

RNAOs, would aid in the design of RNA-based therapies or nucleoside analogs that would specifically bind and target A1. A comprehensive list of RNAOs, with varying sequences and structures, have been probed A1 to address this critical aspect of A1-RNA recognition.

2. **Modelling the complete A1 structure and assess the dynamic RRM-LCD interactions.** A1's aggregation propensity has been well-documented but little is known about its aggregation mechanism. Interdomain contacts between the LCD and RRMs have been proposed as a possible precursor to the large-scale LLPS and aggregation that occurs in cells. To answer this question, the full model of A1 has been subjected to extensive MD simulations to analyze interdomain contacts, their possible repercussions on RNA binding and how these differ between different isoforms of A1.

Chapter 2: Methodology

2.1 Protein Modeling

Structural insights on macromolecules are critical for the understanding their functions, implications in disease and mechanisms for reactions they may be involved in.[94] Structural biology is a complex area of determining the 3D representations of major macromolecules in the cell such as proteins.[94] It employs an array of complex methods such as X-ray crystallography, nuclear magnetic resonance spectroscopy, cryo-electron microscopy (cryo-EM), small angle X-ray scattering and more recently, computational modelling.[94] Of these techniques that are *in vitro*, some challenges make them tedious- each step is manual, time-intensive and requires isolation and purification of desired protein from their source.[94]

An alternative to *in vitro* structural modeling could be *in silico* modeling.[95] Given that the native conformation of a protein is its most stable form, computational modeling can be applied to predict such a state.[95] With the amino acid sequence, programs can predict how the protein could fold in a given environment.[95] A template-based or a template-free approach can be used in such cases.[95] Template-based approaches are most useful for proteins with evolutionary ancestors (homology modeling) or some sequence similarity with other proteins that would result in some common folds (threading).[95,96] However, when no such similarities exist, a template-free approach is performed, which is completely ab initio.[97] Ab initio methods can be based on physics, where force fields are used to model atomic interactions, or knowledge-based that used structures form the PDB as a starting point.[97] The modelling of isoform A of A1 for this research was conducted using threading via the I-TASSER method which uses a combination of template-based and pure ab initio approaches (Chapter 4).[98-100] First, I-TASSER tries to model the protein based on available templates.[98-100] When that is not applicable, a pure ab initio method is applied.[98-100]

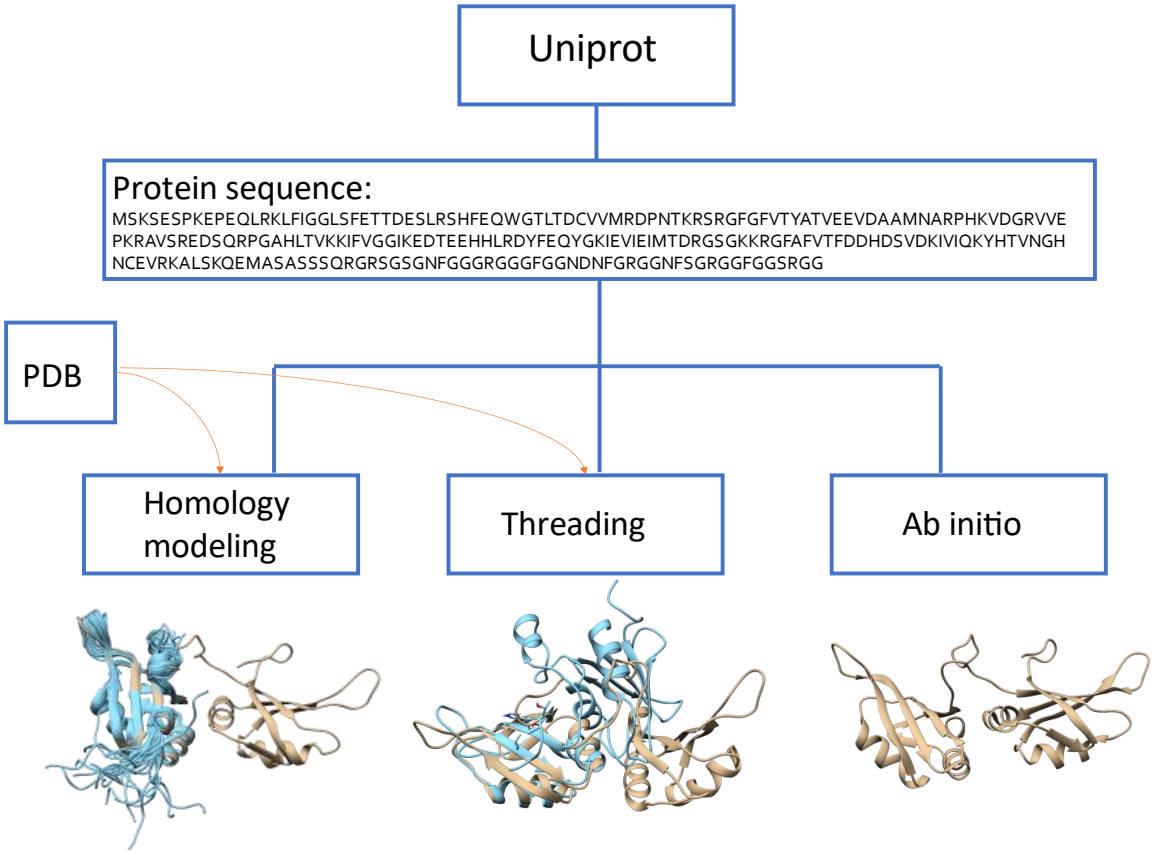


Figure 2.1: Summary of common protein modeling methods.[95]

The rise in computational modeling is gaining increasing acceptance to accomplish goals that *in vitro* studies would find challenging or time-intensive.[96] Artificial intelligence (AI) further enhances the capability of *in silico* research and as such, the Alphafold tool was introduced to make protein structures available to researchers.[101,102,103] Alphafold utilizes the efficiency of deep neural networks in its algorithm that combines templates and multiple sequence alignments to predict protein structures. Only isoform B of A1 was available on Alphafold. Consequently, A1-B was obtained from Alphafold while A1-A was modeled via a mixed homology and ab initio approach. Both models have been used for studying the dynamics of the full protein in Chapter 4. Although A1-B is more abundant in neurons and more likely to aggregate, A1-A is more

ubiquitous in the body but much shorter in size.[104] The structural variations between the two deemed it necessary to study each of them and compare their dynamics to achieve insights on their differing aggregation propensity.

2.2 RNA Modeling

Once only recognized as a passive messenger molecule, RNA has now acclaimed its significance in biological research.[1] The role of RNA in cellular metabolism has been well-characterized in recent years.[1] For example, its ability to catalyze reactions as a ribozyme, sense homeostatic changes as riboswitches and regulate epigenetics as long non-coding RNAs (lcRNAs) has encouraged its structure determination.[105] Understanding RNA folding is a critical step to RNA modeling, which can then be used for RNA structure and function.[105] An example where RNA modeling has proven fruitful includes the understanding of ribosome and the role of ribosomal rRNAs that conduct protein translation.[105]

The rise in studying RNA for the purposes of drug discovery requires an understanding of its structure. It can be thought of as a polymer chain comprised of four basic nitrogen-containing bases, namely Adenine (A), Guanine (G), Cytosine (C) and Uracil (U).[106] The four nucleic acids can make up the primary structure of RNA which would be a linear strand.[106] However, RNA, if long enough, has the potential for canonical or Watson-Crick base pairing, resulting in a folded secondary duplex structure.[106] non-canonical base pairing is also possible, allowing RNA to exist as a duplex structure whenever possible.[106]

As with protein modeling, RNA-modeling also constitutes of general template-based, template-free or machine-learning based methods, or any combination of the three.[105] General machine learning techniques and neural networks are becoming increasingly popular and accurate for RNA- secondary structure prediction.[107] Covariation-based methods and integrative methods

employ the use of previously known RNA structures and multiple sequence alignments available.[105] Energy-based methods for RNA secondary structure prediction remains the most widely-used method to-date, where RNA structures are predicted based on most stable conformations possible using thermodynamic parameters when a template is not available.[105] RNA-modeling has been used in Chapter 3 to predict the structures of RNAOs that have not been reported previously.

2.3 Molecular Docking

Molecular docking is a method used to predict optimal conformation of a given ligand in a given site within a protein.[108] Possible list of conformations of ligand-protein complexes are generated and ranked based on scoring functions that vary within docking programs.[109] Most docking programs do not take protein flexibility into consideration, also known as rigid docking.[108] A more preferable approach to docking that brings protein flexibility into calculations include soft-docking and rotamer libraries.[110] Soft docking works by allowing a small amount of overlap between the receptor and ligand, thereby "softening" the van der Waals terms in the scoring functions.[110] However, soft docking only allows local movements of the residues involved with ligand interactions.[108] To capture conformational changes in the full protein by ligand placement, an ensemble docking algorithm can be used.[110] This type of docking averages the grids of an ensemble of protein structures provided.[108] The ensemble of protein structures will often comprise of unique conformations of the same protein.[108]

RNA-protein docking is an essential step in understanding and/or designing therapeutics for disease since defects in RNA-protein interactions are implicated in neurological diseases and cancers.[111,112] However, experimentally-derived macromolecular structures have remained difficult to obtain, which paved way for computational modeling of protein and protein-ligand

complexes.[113] For RNA-protein docking, the protein is defined as the receptor and RNA as the ligand.[114] software Docking occurs in two steps: conformational sampling and scoring. The conformational sampling explores possible stable orientations that can exist for the complex and scoring determines mathematically how accurate the complex could be.[113]

A major challenge in RNA-protein docking is managing to achieve specific protein-RNA interactions.[115,116] The phosphate backbone of RNA tends to bind with any polar or charged residue in a protein, although it is well-established that aromatic pi-pi stacking are drivers of specific contact in RBPs with RRM domains.[117] In fact, A1's signature contacts as illustrated in Figure 1.4 emphasise on a specific binding mode. For the purposes of this research and to minimize nonspecific contacts given by most docking programs that are not specialized for RNA-protein docking, the MDockPP[118] server has been employed in Chapter 3. Although it uses rigid docking, given the very limited docking servers available that are specialized for RNA-protein docking, MDockPP posed as the best choice as it is a well-established tool for modeling RNA-protein binding.[118] Molecular dynamics were utilized to allow for flexibility in the complexes that may not have been achieved with rigid docking by MDockPP.[118]

2.4 Alignment and Extraction of PDBs

In certain cases, docking may not be necessary if experimentally determined structures are available in the PDB. For example, the RRM1 of A1 has a 7-nt ssRNA bound to it (PDB:5MPG) [17]. The 5MPG structure alone was not sufficient for research as it was only RRM1 and A1 is considered more stable with both RRMs tethered together.[119] Using a software such as Chimera, [120] aligning the 5MPG structure (only RRM1) with a structure containing both RRMs (PDB:4YOE)[121], would allow the 7nt ssRNA from 5MPG to get "extracted" onto the 4YOE structure. Removing the protein component of the 5MPG structure leaves the 4YOE[122] bound to

the 7nt ssRNA. Mutations can be done to the RNA in Chimera to create an array of RNA-protein complexes. This method is heavily relied for most of the experiments in this thesis. Both Chapters 3 and 4 have RNA-protein complexes derived from this method.

Another way extraction has been utilized in this thesis is in Chapter 4 to create a dimer of A1-A. The only crystal structure of A1 in dimeric form was the 6DCL[84] structure, but only contained the RRMs and not the LCD. By aligning two monomers of the full A1-A onto the 6DCL structure and then deleting the 6DCL structure, a dimer of A1-A with their LCD domains was created.

2.5 Molecular Dynamics

The most redundantly used technique in this thesis is Molecular Dynamics (MD). MD simulations use mathematical calculations that are based on Newtonian equations of motions to predict each atom's behaviour in a given environment.[123] This allows for obtaining extensive insights into biomolecules and cellular processes such a protein folding, ligand binding and possible conformations of a molecule.[92] Time-dependent variations in conformations of a system are also useful insights that make MD a valuable tool.[108] Experimentally expensive perturbations such as mutations and post-translational modifications are also much easier to do with MD.[92] This is acquired by treating atoms in the system as solid spheres and the bonds connecting the atoms as springs.[108] Movements and vibrations of atoms thus become possible to capture during an MD simulation.[108]

$$m_i = \frac{\delta^r r_i}{\delta t^2} = F_i \quad \text{(Equation 2.1)}$$

Equation 2.1 explains how Newtonian physics is used to run an MD simulation.[124] F_i is the net force acting on an atom with a mass, m_i, while r_i is the position of the atom at time t.[124]

$$F_i = \frac{-\delta U(r_1, r_2, \ldots, r_n)}{\delta r_i} \qquad \text{(Equation 2.2)}$$

The force can be elaborated with equation 2.2, where, $U(r_i, r_2, \ldots, r_n)$ is the potential energy of a specific conformation and can be explained as a forcefield.[124] A force field is essentially a mathematical function that takes possible bonded and nonbonded interactions to describe the potential energy of the atoms in a system.[124] The parameters differ between different types of forcefields.[123] Common forcefields that exist are AMBER[125], CHARMM[126] and Gromacs[127]. AMBER is the forcefield used for all the experiments in this research to study isolated A1 as well as A1-RNA bound complexes (Chapters 3 and 4).

The functional form of AMBER MD equations is simplified as:[125]

$$E_{total} = \sum_{bonds} k_b(r - r_0)^2 \qquad \textit{Stretch term}$$

$$+ \sum_{angles} k_\theta(\theta - \theta_0)^2 \qquad \textit{Bend term}$$

$$+ \sum_{dihedrals} V_n[1 + \cos(n\phi - \gamma)] \qquad \textit{Torsional term}$$

$$+ \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{12}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right] \qquad \textit{Non-bonded interactions}$$

(Equation 2.3)

In equation 2.3, E_{total} is the total energy of the atoms. It takes into account the stretch terms (bonds), bend terms (bond angles), torsional terms (dihedral angles), and non-bonded interactions. [128] The

terms k_b, r_0, k_θ ,θ_0,V_n, γ, $A_{i\,j}$, $B_{i\,j}$ are parameters that are specified based on the type of molecule modeled [128] For example, AMBER offers forcefields for simulation of proteins, DNA, RNA, carbohydrates, lipids, water, ions and organic molecules.[128] The parameters are adjusted each time to optimal values for each type of molecule. Partial charges (q_i ,q_j) are predetermined values that are automatically assigned by AMBER prior to MD.[128]

Prior to MD, the system of interest must have missing residues added, solvated with an appropriate amount of water and charge-neutralized.[129] Physiological salt concentration can also be achieved by adding additional ions.[129] The temperature and pressure are examples of parameters that can be modified throughout the simulation.[123] Periodic boundary conditions were applied for this thesis when using MD, which essentially allows an object passing through one side of a unit cell to re-appear on the opposite side.[130] Setting periodic boundary conditions in MD helps to allow the extrapolation of the behaviour of a small number of atoms to a larger set of atoms. [130]

2.6 Binding-free energy calculation

Following molecular docking, a fair understanding of the protein-ligand complex and a binding energy to estimate the rank of each docked pose is achieved.[131] However, gaining an even more thorough quantification may be helpful to obtain a deeper understanding of protein-ligand contacts.[108] Binding free energy ΔG_{bind}; is a value frequently determined to quantify the free energy difference between the ligand-bound state (complex) and the corresponding unbound states of proteins and ligands) is used to quantify the affinity of a ligand to its target.[131] The ΔG_{bind} values can help decipher which ligands have best affinity to the target and thus can be useful in drug design. Several computational methods exist to determine ΔG_{bind} that may integrate vigorous thermodynamic calculations (alchemical methods) or simply calculate using an end-point

approach (end-point methods).[131] The thermodynamic integration (TI) and free energy perturbation (FEP) methods are alchemical methods, whereas liner interaction energy (LIE), MM-generalised Born surface area (MM-GBSA), and MM-Poisson–Boltzmann surface area (MM-PBSA) are end-point methods.[131]

Alchemical methods employ a thermodynamic cycle which annihilates the interactions of the ligand molecule with its environment which allows the calculation of what is known as decoupling free energy.[131] The calculation of the decoupling free energy is often considered accurate, however, very computationally expansive and time-intensive.[108] Endpoint methods on the other hand are faster but less accurate compared to alchemical methods.[131] These only consider the endpoints of binding processes by considering the isolated ligand and protein and a ligand-protein complex.[108] Endpoint methods: LIE (linear interaction energy), MM-GBSA (molecular mechanics Generalized Born surface area) and MM-PBSA (molecular mechanics Poisson–Boltzmann surface area).[131] The LIE method analyzes the interactions between the ligand with the environment which results in solvated endpoint states to calculate interaction energies.[131] Essentially, it is similar to the docking scoring method but with the addition of solvent effects.[131] LIE has been utilized in Chapter 4 to calculate non bonded interactions between different residues of the same protein to get a quick trend in binding free energies across the simulations.

MM-PBSA and MM-GBSA are more elaborate than LIE. Both use implicit solvent effects via using the dielectric constant of the solvent to evaluate the binding-free energy of a protein-ligand complex (ΔGbind,aq). ΔGbind,aq is calculated with the following[132,133]:

$$\Delta G_{bind,aq} = \Delta H - T\Delta S \approx \Delta E_{MM} + \Delta G_{bind,solv} - T\Delta S \quad \text{(Equation 2.4)}$$

Where ΔE_{MM}, $\Delta G_{bind,\ solv}$, and $-T\Delta S$ are the change in gas-phase molecular mechanical energy, the change in the solvation free energy, and the change in entropy, respectively.[133]

$\Delta E_{MM} = \Delta E_{covalent} + \Delta E_{electrostatic} + \Delta E_{vdW}$ (Equation 2.5)

ΔE_{MM} is calculated with molecular mechanics (MM) using the change in covalent energy ($\Delta E_{covalent}$), the change in electrostatic energy ($\Delta E_{electrostatic}$), and the change in van der Waals energy (ΔE_{vdW}).

$\Delta E_{covalent} = \Delta E_{bond} + \Delta E_{angle} + \Delta E_{torsion}$ (Equation 2.6)

$\Delta E_{covalent}$ takes into account the changes in bond terms (ΔE_{bond}), the changes in angle terms (ΔE_{angle}), and the changes in torsion terms ($\Delta E_{torsion}$).[131,133]

$\Delta G_{bind,solv} = \Delta G_{polar} + \Delta G_{non-polar}$ (Equation 2.6)

The solvation free energy change ($\Delta G_{bind,\ solv}$) consists of polar and non-polar changes (ΔG_{polar} and $\Delta G_{non\text{-}polar}$).

The entropy term ($T\Delta S$) is the term omitted in MM-GBSA, which is the key difference in the two methods.[131] This is due to the entropy change being computationally more expensive and difficult to calculate.[131] This renders the MM-GBSA method with very little use without a control system to compare it to, as with docking scores since the binding energy calculated with MM-GBSA is never absolute, but only useful for comparison.[131] For the purposes of this research, the MM-GBSA method seemed fit as it is used in the context of comparing various RNA-protein complexes with each other and knowing the absolute energy of the complex is not required.[131] We employed optimal dielectric constants for protein-RNA systems as reported earlier.[134]

2.7 Principal Component and Cross-Correlation Analyses

Principal component analysis (PCA) is a statistical method used to reduce the dimensionality of a complex system, while extracting the variations in the datasets.[135] Using PCA, the dataset is

reduced to a few components that represent sample variations instead of visualizing thousands of variables.[135] With respect to MD data, PCA helps describe the variance in the dynamics and conformation of the systems.[136] The variance of the atomic positional fluctuations captured in each dimension are characterized by their corresponding eigenvalue.[136] Most of the cases have 3-5 dimensions that capture over 70% of the total variance within a given MD trajectory.[137] For MD, PCA is sufficiently calculated via measurements of dihedral angles or atomic coordinates for α-carbon atoms.[136] The x, y and z Cartesian coordinates of the C-α atoms were then used to map a cross-correlation visualization for the trajectory.[136] All PCA and cross-correlation plots were conducted using the Bio3d[137] package in R.

Workflow

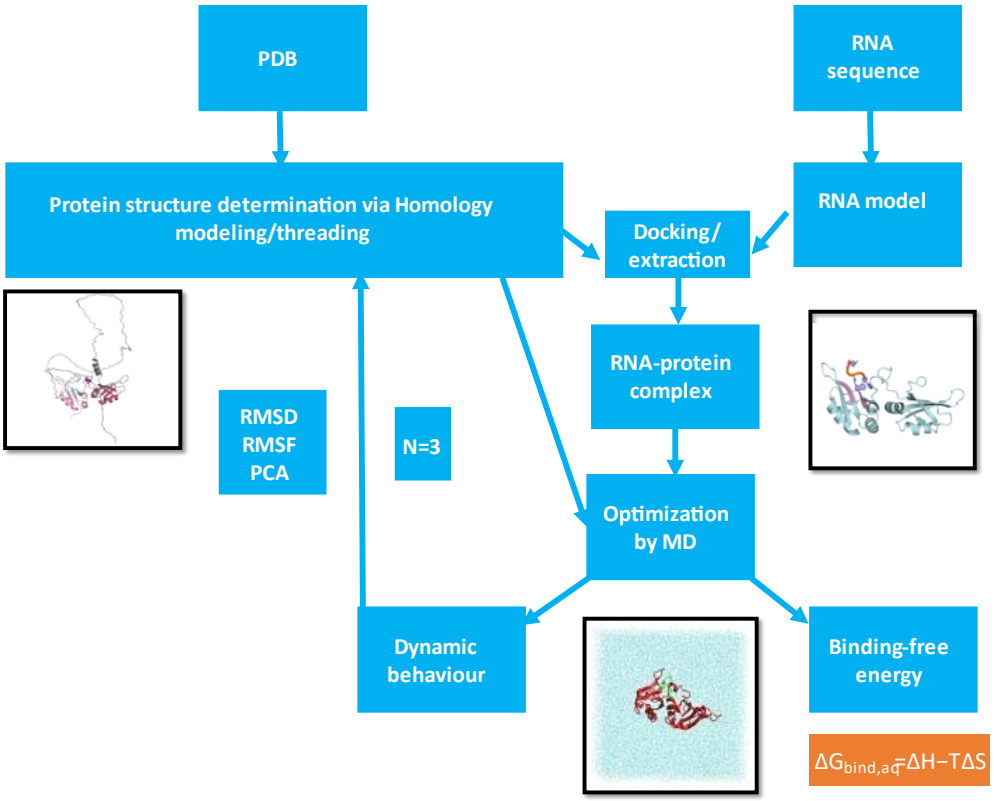


Figure 2.2: The general workflow for this thesis involved modeling the full structure of RNA and the full protein when it was not available. Once the RNA and protein models were obtained, the RNA (ligand) was docked onto the protein (receptor) to create the protein-ligand complex. Docking was replaced by extraction when the RNA positioning on the protein was known and a PDB of a similar complex already existed. The complex was then subjected to MD simulations. For a free system not involving RNA, docking was skipped and MD was directly performed on the full protein. Following MD, the behaviour of the system was studied via RMSD, RMSF and/or PCA, depending on the system. Bind-free energy analysis was performed via MM-GBSA for any RNA-protein complex.

Chapter 3: Molecular interactions driving RNA binding to A1

3.1 Introduction

RNP motifs within the RRM domains have been reported as crucial for RNA adherence to the A1 protein.[1] In reported crystal structures, the RNP motif is observed to interact with a 5′-AG-3′ motif (Table 1 and Figure 3.1). These interactions are mostly based on aromatic stacking of an Adenine between Phe17 and His101, in addition to a Guanine or Thymine stacking with Phe59 and Arg92 (Figure 3.1). The RNA without the signature 5′-AG-3′ motif showed the weakest affinity to A1.[56] Therefore, this motif (5′-AG-3′) has become a well-established tool for probing RNA-A1 complex and their effects on A1 functions.[83] Given the ubiquitous nature of A1 and its ability to bind diverse RNA molecules, it would be useful to understand the effects of sequence variations to 5′-AG-3′ motif on the RNA-A1 RRM complex at the molecular level. This is particularly useful as there are currently no experimental three-dimensional (3D) structures describing RNAs without the specific 5′-AG-3′ motif.

Therefore, this work aims to bridge this gap by performing a systematic molecular-level analyses to understand how different nucleotides (AUGC) in the 5′-AG-3′ motif would alter the structure, dynamics, and binding affinity of RNA-A1 RRM complexes. A combination of *in silico* mutations, molecular dynamics (MD) simulation, binding free energy calculations, and per-residue decomposition analyses are employed. The crystal structure of 5′-AG-3′ motif-bound human A1 RRM complex (PDB: 4YOE) was used as a reference and making single point substitutions/mutations at the positions of the AG motif with different nucleotides. The reference crystal structure (along with others in Table 1) demonstrates that only A and G in the motif are directly involved in binding with RNPs of A1 RRMs (Figure 3.1), and U is does not contribute to binding as it faces away (in the 4YOE structure). Therefore, no substitutions were explored in the position of uracil in this work, but only on nucleotide substitutions at the 5′-AG-3′ positions. The

findings cross-verified by making favourable and unfavourable substitutions at the analogous AG positions within a 7-nucleotide (7-nt) long RNA sequence of 5′-UUAGGUC-3′.

Finally, the significance of specific native interactions mediated by the presence of the AG motif were extended to longer literature-inspired RNAOs in collaboration with the University of Saskatchewan. RNAOs designed by our collaborators were modeled and simulated to assess for RNA-binding. Results from *in silico* and *in vitro* analysis further confirmed insights obtained from the single nucleotide substitutions. Thus, our work extends quantitative and qualitative insights into the roles of nucleotide specificity in RNA recognition by the A1 RRM at molecular level, which should be useful for developing RNA-based therapies to modulate A1 functionalities.

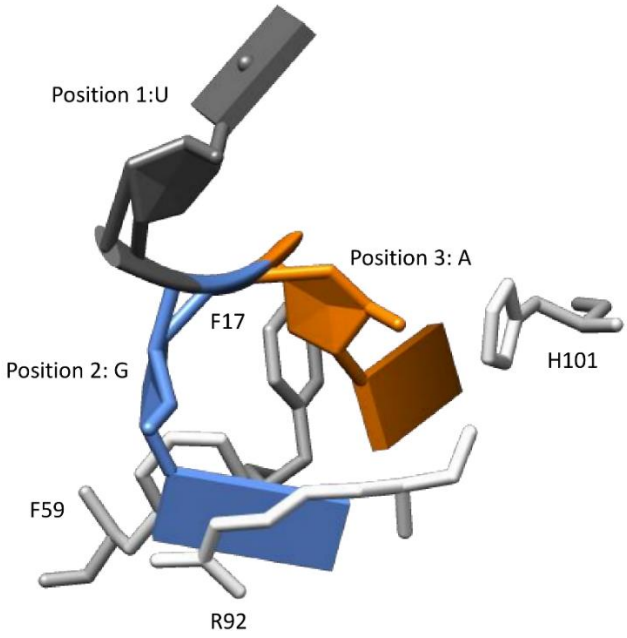


Figure 3.1: The labelling of positions from the crystal structure 4YOE[122] for the experimental setup. Each placement of nucleotide at the RNA-protein interface has been given a name for ease in understanding the methodology for section 3.2.

3.2 Methodology

3.2.1 Modelling the WT and mutant RNA-bound complexes

The 5′-AGU-3′ RNA-bound human A1 RRM complex (PDB: 4YOE) was used as the reference structure, and named as WT^{3NT} in this text. Out of all the experimental structures of RNA-bound A1 RRM complexes, we chose this structure as it describes the binding of the 5′-AGU-3′ motif to the complete RRM domain of A1, which includes both RRM1 and RRM2 tethered (known as UP1). A previous study found that the ATP molecule exhibited a higher binding affinity towards UP1 rather than an isolated RRM1 or RRM2.[119] Further, it is proposed that the binding of RNA supports the relative orientations of RRM1 and RRM2 in A1. Therefore, the 4YOE crystal structure was downloaded and prepared by removing the water molecules, and ions, and adding hydrogen atoms This prepared WT^{3NT} complex was used to perform systematic mutation of the 5′-AGU-3′ RNA to model the mutant structures using UCSF Chimera (version 1.16).[142]As described in the introduction, only the first two positions positions related to AG were modified with other nucleotides, one at a time. The Uracil was left unmodified, as it did not make any significant contact with the protein. (Figure 3.1). When adenine at position 1 was substituted to another nucleotide, these complexes was named as M1*N*, with M referring to mutants, 1 indicating the position, and N denoting the substituted nucleotide for A. For example, if A is substituted with G, then this complex is dubbed as M1G. Similarly, when G at position 2 of WT was substituted, then ligand was named M2*N*, where 2 refers to the position of mutation in the WT RNA (i.e.,G). For example, if G in WT is substituted with C, then this modified RNA is dubbed as M2C. Note that all the mutations were directly performed on the RNA bound in WT^{3NT} complex directly, to probe how the substitution would affect the original RNA-RRM binding pose. This strategy resulted in a total of six mutant RNA-A1 complexes. Next, we mutated all the three positions of the 5′-AGU-

3′ motif in the WT^{3NT} system, in order to probe the interactions and binding affinity of the same nucleotide type against A1 RRM. These systems were named as MA3, MU3, MG3 and MC3, for the substation of adenine, uracil, guanine, and cytosine in all the three positions. In addition, we expanded our work towards probing the effects of mutating the AG positions in a 7-nt long RNA (5′-UU**AG**GUC-3′) that is bound to the A1 RRM (PDB: 5MPG[173]). This structure represents an ensemble of 20 RNA-bound A1 RRM1 complex resolved through NMR technique. Therefore, the NMR structures were clustered using Chimera[142] to find the dominant structure. Since the RRM2 was absent in the 5MPG structure, the dominant model was superposed over the 4YOE model, and the 7-nt long RNA was extracted into the complete UP1 in the latter structure (i.e., 4YOE). This WT structure with bound 7-nt RNA is dubbed as WT^{7NT}. Finally, we mutated AG at the positions 3 and 4 of WT^{7NT} with same nucleotide types to model the mutant structures that were dubbed as MN34, where N refers to the nucleotide replacing both 3 and 4 positions. This resulted in 4 mutant models of WT7NT. This was done to confirm the reproducibility of trends observed for the shorter nucleotide models, WT^{3NT} and its mutant structures. Therefore, a total of 16 A1 RRM complexes (2 WT and 14 mutants) were modelled for analyzing their dynamic stability and binding affinities. Refer to Supplementary Table 1 for a list of all the structures modelled and probed in this work.

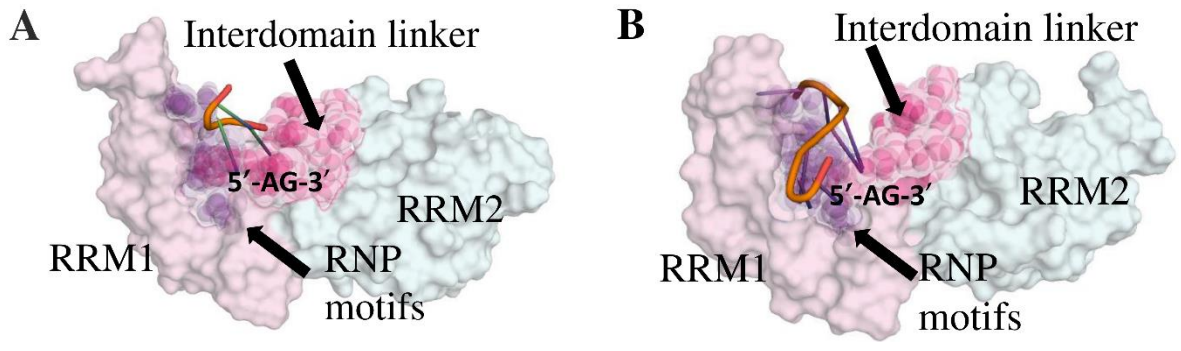


Figure 3.2: The starting conformations of the WT^{3NT} and WT^{7NT}. The WT^{3NT} (A) an WT^{7NT} (B) are shown as a surface representation, with the RRM1 coloured as cyan and the RRM2 in pink. The RNP motifs and the linker that make up the binding pocket for the RNA ligand (ribbon) are shown as beads in purple and grey, respectively.

3.2.2 MD simulation of the RNA-RRM complexes

Each of the WT^{3NT} and its mutant complexes were energy minimized, equilibrated, and followed by a total of 400 ns long production under physiological conditions. All MD simulations were performed with the AMBER20 package[125] and pmemd.cuda engine[139]. A combination of FF14SB[33] and the recent RNA Amber force field developed by a Rochester team (i.e., RNA-ROC) was used for describing the structural parameters of protein and RNA, respectively, for MD simulation. Each complex was solvated in a cubic box of explicit TIP3P water molecules with a distance of 10 Å between the solute and the edge of the box. The solvated systems were charge neutralized and brought to 150 mM concentration of NaCl, which is accepted to be the physiological salt concentration in the cell.[140] All system preparation was performed using the tleap program available within the AMBER package[125].

The prepared complexes were initially energy-minimized in 6 stages with each stage involving 1000 steps of steepest descent minimization and 10000 steps of conjugate gradient minimization with a pre-defined harmonic restraint. In the initial stage, a 100 kcal/mol Å−2 restraint was applied on the solute atoms which was gradually decreased to 70>50>40>30>0 kcal/mol Å−2 in the subsequent rounds of minimization. The energy minimized systems were gradually heated to 310 K (with a 15 kcal/mol Å−2 on the solute atoms) over a duration of 100 ps and, subsequently, subjected to 5 x 0.4 ns equilibration cycles that were performed under isothermal-isobaric (NPT) conditions with periodic boundary conditions. Again, the equilibration was performed with an implied restraints on the solute atoms that gradually reduced as 15>10>5>3>2 kcal/mol Å−2 in each phase. The equilibrated complexes underwent a 10 ns long MD simulation with a low restraint The stability of the protein and RNAs in the complexes during the course of simulation was assessed by computing the evolution of root mean square deviation (RMSD). MD trajectory

analyses were performed using the CPPTRAJ module[35] in Amber and the visual analyses were carried out using VMD (version 1.9.3)[141], UCSF Chimera (version 1.16)[142] and PyMol programs (Version 2.0).[38]

3.2.3 Binding free energy calculations and analysis

The binding free energy calculations of the RNA-A1 RRM complexes in this work were carried out using the molecular mechanics with generalised Born and surface area solvation (MM-GBSA) method with the implicit solvent model of Onufriev and Case (igb=2).[143] The snapshots were sampled at a constant interval of 100 ps from the last 10 ns of the MD trajectory for these calculations. The pairwise decomposition analyses (idecomp=2) were performed to identify the key nucleotides and amino acids that contribute to the binding free energies of the complexes. All computations were performed using MMPBSA.py.MPI script[143] included in the AmberTools 20 [125]. A combination of analyses, including RMSD, RMSF and Binding free energies, were used to assess the effects of single-point nucleotide substitution on the RNA-A1 interactions. Taken together, these analyses should be able to reveal thorough insights into nucleotide-type and key amino acids that play a crucial role in RNA recognition by A1 RRMs.

3.3 Results

3.3.1 Molecular contacts driving the binding affinity of native RNA ligands with the RRM1 of A1

The AGU-RRM1 complex (WT^{3NT}) and the 5MPG-A1 complex (WT^{7NT}) structures were subjected to a 400 ns-long MD simulation. Studying and assessing the stability and key interactions of ligands found bound to A1 in the PDB was a crucial step to understating native contacts. An array of analyses were performed to analyze these dynamics. The overall stability of the protein backbone and full RNA was assessed using the RMSD evolution for 400 ns (Figure 3.3A).

Fluctuations in the RMSD usually indicate movements in the structure, so the simulation required stability before analysis[144]. RMSD fluctuations occurred in the first ~100 ns, after which full stability was achieved by 200 ns. RMSD values stabilized earlier (~180 ns) for the WT^{7NT} complex than the WT^{3NT} complex (~200 ns), perhaps due to a longer strand on RNA enhancing the stability of the RRMs. Stability was further assessed with RMSF analysis that indicates fluctuations of each residue during the course of the MD. Higher RMSF peaks indicate higher fluctuations compared to lower ones. The RMSF analysis in Figure 3.3B describes the RMSF for the protein where the RNP motifs and residues in the linker loop having stable atomic fluctuation rather than sharp peaks as seen for the N-terminal and C-terminal ends of the protein that are disordered and flexible. Further, RMSF analysis was also conducted for the RNA residues as illustrated in Figure 3.3 C. Lower RMSF fluctuations for the AG motif compared to the U provide additional conformation of the AG motif forming stable interactions in Figure $3.3C^{I}$ for the WT^{3NT}. This is explained by the interaction visual provided in Figure $3.3C^{I}$ where the signature AG-RNP contact was maintained for the WT^{3NT}, spanning most of the trajectory and confirming its significance in A1-RNA recognition. This trend is replicated in Figure $3.3C^{II}$. The pi-pi stacking interactions with Phe17 and Phe59 with A3 and G4, respectively, are forming essential contacts. G4 is also forming electrostatic contacts with Arg92. In both the complexes, the 'AG' motif in direct contact with RRM1 had lower RMSF values compared to the non-interacting residues.

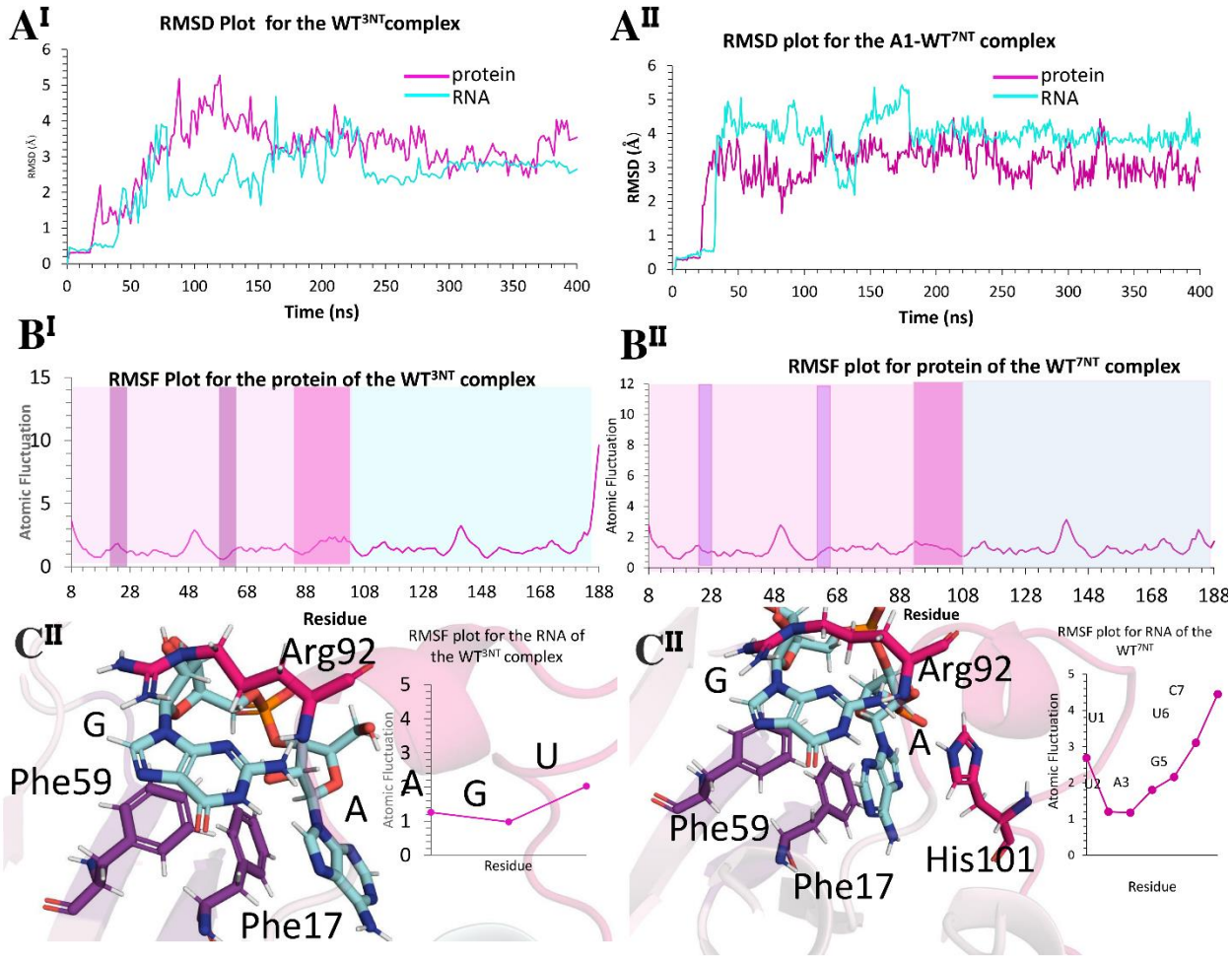


Figure 3.3: Analyzing stability of the WT^{3NT} (left panel) and WT^{7NT} (right panel). (A) RMSD analysis demonstrated stability around 200 ns (A^{I}) and 180 ns (A^{II}). **(B^{I}) (B^{II})** The RMSF plots corresponding to the residues by colour indicate relative stability for the RNP regions, with slightly higher fluctuations for the linker regions for both complexes. **(C^{I}) (C^{II})** Phe17, Phe59 and Arg92, form most of the interactions with the AG motif. The RNA was overall stable in the binding pocket as indicated by the RMSF for RNA with fluctuations by noninteracting residues**.**

Intriguingly, nucleotide sidechains are facing downwards into the RNP motifs, with the AG motif being predominantly responsible for stacking interactions with the RNPs (Figure 3.3 C). Phe17 and Phe59 in particular, contribute to the pi-pi stacking with an upward-facing conformation.[117] Aromatic residues in RBPs adopt an upwards conformation to interact with the nucleotides side chains[117], which is evident in Fig 3.3C, thus providing a qualitative conformation of the AG motif stacking with Phe17 and Phe59., which is quantitatively proven by the MM-GBSA (Figure 3.4A,

3.5A). The MM-GBSA decomposition analysis elaborates on the significant contribution to the binding affinity from the RNPs and the linker-loop which is due to the proximity with the AG motif. The RNA residues contributing to binding are mostly the AG motif and some electrostatic affinity provided by other residues.

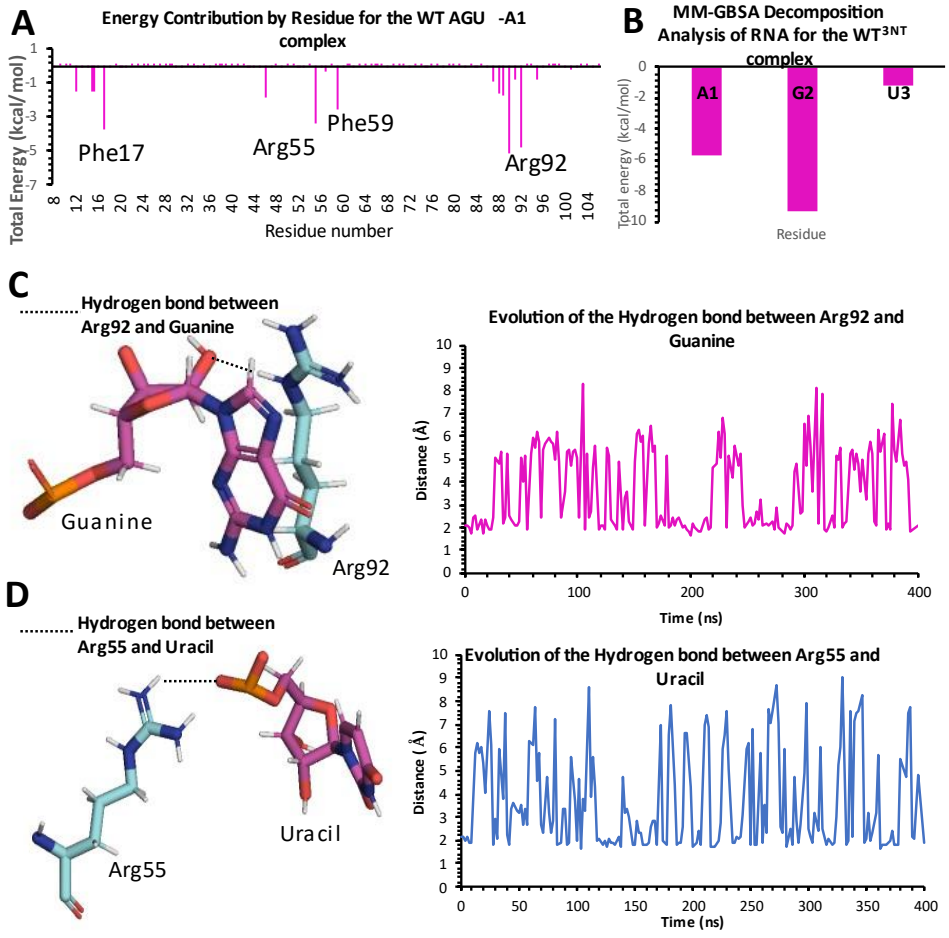


Figure 3.4: Binding affinity analysis and prominent hydrogen bonds in the WT^{3NT} complex. **(A)** Unsurprisingly, the residues contributing the most to the binding affinity as shown in the MM-GBSA analysis plot included the RNP motifs. Other than stacking interactions, some residues from the interdomain linker or RNP motifs provided stability to the complex via hydrogen bonds such as Arg92 and Arg55. **(B)** the RNA MM-GBSA decomposition analysis for the WT^{3NT} complex demonstrates a significant contribution to the RNA binding by the Adenine in position 1, Guanine in position2 but very little contribution by Uracil in position 3. **(C)** A hydrogen bond interaction between the Guanine in position 2 and Arg92 is demonstrated as an example. The selected amino acids and nucleotides are shown as stick representation as the other segments of the binding pose are shown as cartoon representation in the background. Evolution of distances between the side-chain carbonyl group of guanidine group of Arg92 and purine ring of Guanine is represented as a

plot. **(D)** Similarly, the hydrogen bond distance evolution between the guanidine group of ARG55 and phosphate group of Uracil (NH-OP1 bond) is illustrated to provide explanation for the large energy contribution by Arg55 in the MM-GBSA decomposition analysis.

Therefore, aromatic residues from the RNP motifs and interdomain linker provide specific RNA-recognition interactions. While MM-GBSA is a good way to know which residues contribute to binding, further analysis can be conducted using hydrogen bond analysis. This is conducted where the distance between atoms forming a hydrogen bond in a complex is measure over the course of the simulation. In the WT^{3NT} complex a detailed hydrogen bond distance analysis for Arg55 and Arg92 elaborate on the stability of the electrostatic contacts provided by the RNPs and the linker together (Figure 3.4C, D). The Arg92 guanidine group interacts with the Guanine, which is also stacking with Phe59, hence keeping it fixed in position 2. This illustrates the role of the interdomain linker that may be essential for more than keeping the two RRM domains attached. Abundance of polar residues in the linker, such as Arg92, may aid in RNA-binding especially to the RRM1 domain.

In the WT^{7NT} complex, additional binding affinity is provided by the additional G in the AGG motif (Figure 3.5 B). The role of this additional guanine is elaborated by analyzing the hydrogen bond analysis in Figure 3.5 C , where it is seen interacting with an Asp42, which did not occur in the WT^{3NT} due to the absence of this guanine. G3 replicates the electrostatic contact with Arg92 as seen in the WT^{3NT}. Additional electrostatic contacts due to a longer length of RNA in the WT^{7NT} also resulted in a higher binding affinity value (-83 kcal/mol) (Table 3.3) as compared to the WT^{3NT} structure (-48 kcal/mol) (Table 3.1) , despite higher fluctuations in the RNA.

All in all, both complexes revealed the conservation of interactions where applicable. Although both had a redundant binding mode, they were both necessary. This due to the fact that shorter ssRNAs tend to be flexible and allow for the bases to be more exposed and available for aromatic

stacking.[145] Henceforth, testing the AGG motif as part of longer RNA strands (7-nt) using the same methodology were tested to validate the insights gained form the WT^{3NT} system. These data altogether elaborate on the essential role of the AG motif in A1-RNA binding with both the WT^{3NT} and WT^{7NT}. While Adenine and Guanine redundantly placed in their respective positions for efficient binding of RNA, this raised the question if either Adenine or Guanine were replaceable.

Finding 1: Our MD and binding-free energy analysis for the WT^{3NT} and WT^{7NT} indicates that the AG motif binds effectively with RRM1 of A1, suggesting that Adenine and Guanine and key preferred nucleotides to bind RRM1

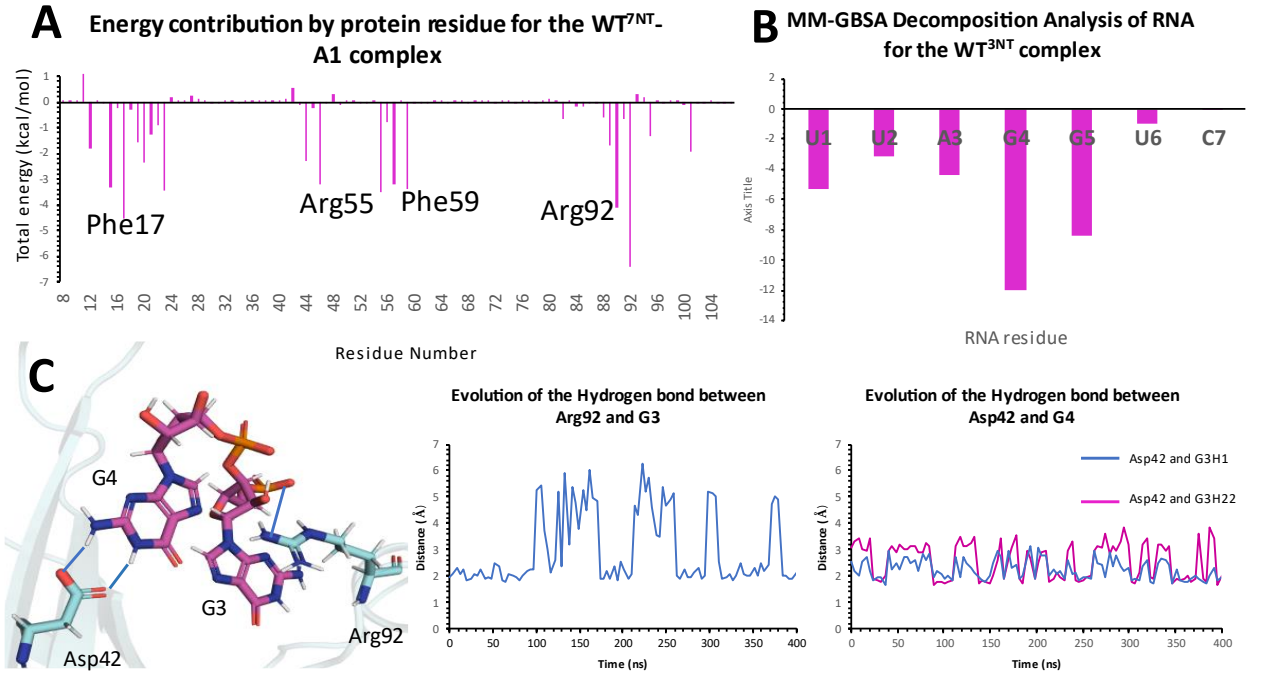


Figure 3.5: Binding affinity and bond analysis for the WT^{7NT}. (A) The highest peaks in the MM-GBSA analysis plot are from prominent RNP residues that are likely stacking, while some residues provided stability to the complex via hydrogen bonds such as Arg92 and Arg55**. (B)** MM-GBSA decoposition analysis confirms the intense additive effect of the position 5 G, even more so than the A in position 3. **(C)** Prominent hydrogen bonds seen in the WT^{7NT} complex. Arg92 and Asp42 are able to form extensive hydrogen bonds with the Guanines in the WT ligand. While the side chain of Asp42 interacts with the polar base amino groups, Arg92 interacts using its guanidine group with the oxygen atom of the phosphate backbone.

3.3.2 Binding-free energy analysis supports the significance of Guanines present in the RNA for binding

The experimental set-up aimed to replace either Adenine or Guanine one at a time with any of the other standard nucleic acids found in RNA. The data for the single-nucleotide mutations on position 1 whereby an Adenine is being replaced (M1G, M1C and M1U) demonstrated lower RNA RMSD values compared to the M2C and M2A (Figure 3.6 B). M2U defies the trend and keeps the RNA stable when in position 2 although M2C and M2A are considerably more unstable. However, the protein of the M2U has higher RMSF fluctuations, in the RNP and linker-loop regions, which is also true for M2A and M2C (Figure 3.6A). M2A, in particular, has a sharp peak for RMSF in the linker loop pertaining to the protein residue instability due to the RNA, followed by M2A and M2C. Surprisingly, M1U and M1G indicate less fluctuations in the linker regions compared to the WT^{3NT} (Figure 3.2A). Indeed, M1U and M1G are the closest in RMSF fluctuations in RNA to the WT^{3NT}. What is interesting to note is that even though position 3 has not been modified in any of the complexes, it indicates a higher fluctuation in RMSF nevertheless.

Upon visual inspection in Figure 3.2 (panels C and D), it becomes clear that the native contacts seen in the WT^{3NT} and crystal structures are present in all the M1*N* complexes, but not maintained in any of the M2*N* complexes. This notion is validated with the MM-GBSA binding energy (Table 2). Position 1 with A was mutated to a G, C or U, has binding affinity that remained comparable to that of the original AGU in WT^{3NT} (-48.19 kcal/mol at 400 ns) (Table 3). M1G had a binding affinity value of -52.71 kcal/mol, whereas M1C and M1U had values of -50.51 and -54.79 kcal/mol, respectively, at 400 ns in the MD.

Table 3.1: Binding affinities (kcal/mol) for the mutations performed at positions 1 (M1*N*) and 2 (M2*N*) with their respective standard deviations. This set of data indicates that position 2 is more sensitive to changes in nucleotide occupancy.

NAME	POSITION 1	POSITION 2	POSITION 3	BINDING AFFINITY (KCAL/MOL)	STANDARD DEVIATION
WT^{3NT} (PDB: 4YOE)	A	G	U	-48.19	6.34
M1G	G	G	U	-52.71	6.33
M1C	C	G	U	-50.51	10.36
M1U	U	G	U	-54.79	5.03
M2A	A	A	U	-25.68	6.71
M2C	A	C	U	-13.43	8.62
M2U	A	U	U	-38.54	3.69

On the contrary, any mutation to the G in position 2 reduced the binding affinity by almost a half or even more (Table 2). The numbers differ drastically from that of the control WT^{3NT} complex, since substituting G in position 2 to an A (M2A), C (M2C) and U (M2U) gave values of -25.68 kcal/mol, -13.43 kcal/mol and -38.54 kcal/mol, at 400 ns, respectively. Collectively, these data demonstrate more sensitivity for the binding affinity for position 2 than position 1. The position 2 Guanine illustrates more vitality in its respective native position over the position 1 Adenine.

Understanding of this trend requires comprehension of chemical bonds that maintain RNA-protein contacts, which includes π-π stacking, hydrogen bonds and salt bridges[146, 147, 148] While all nucleic acids are capable of aromatic π-π stacking, each one has a specific dipole moment, given their unique chemical structures.[146] Guanine having the highest dipole moment, can form stronger hydrogen bonds than the other nucleic acids, which gives it an advantage when interacting with Arg92 in position 2 [146, 149] This observation is congruent with previous data reporting arginine and lysine prefer guanine for hydrogen bonding over other nucleotides.[52,149] Further, since Arginine is

commonly found interacting extensively with nucleic acids in biological systems, it has been suggested that Arginine can contribute to both electrostatic interactions and cation-π interactions.[53,150] Given the highly polar nature of Guanine, having Arg92 interacting with it in position 2 may be highly favourable as it allows for cation-π stacking in addition to simultaneous hydrogen-bonding, to provide the RNA with elevated probability of adherence.[148,150]

Although this set of experiments concluded Guanine as being more necessary than Adenine when binding to the RNP motifs, a further set of experiments were conducted to see if Guanine is also sufficient in holding RNA to the RRM1. These data are illustrated in the next section using additional sets of nucleic acid mutants, named MA3, MG3, MU3 and MU3 (Table 3). These M*N*3 ligands were designed such that only a single type of nucleotide could sit in the RRM1 binding pocket and assessed for binding.

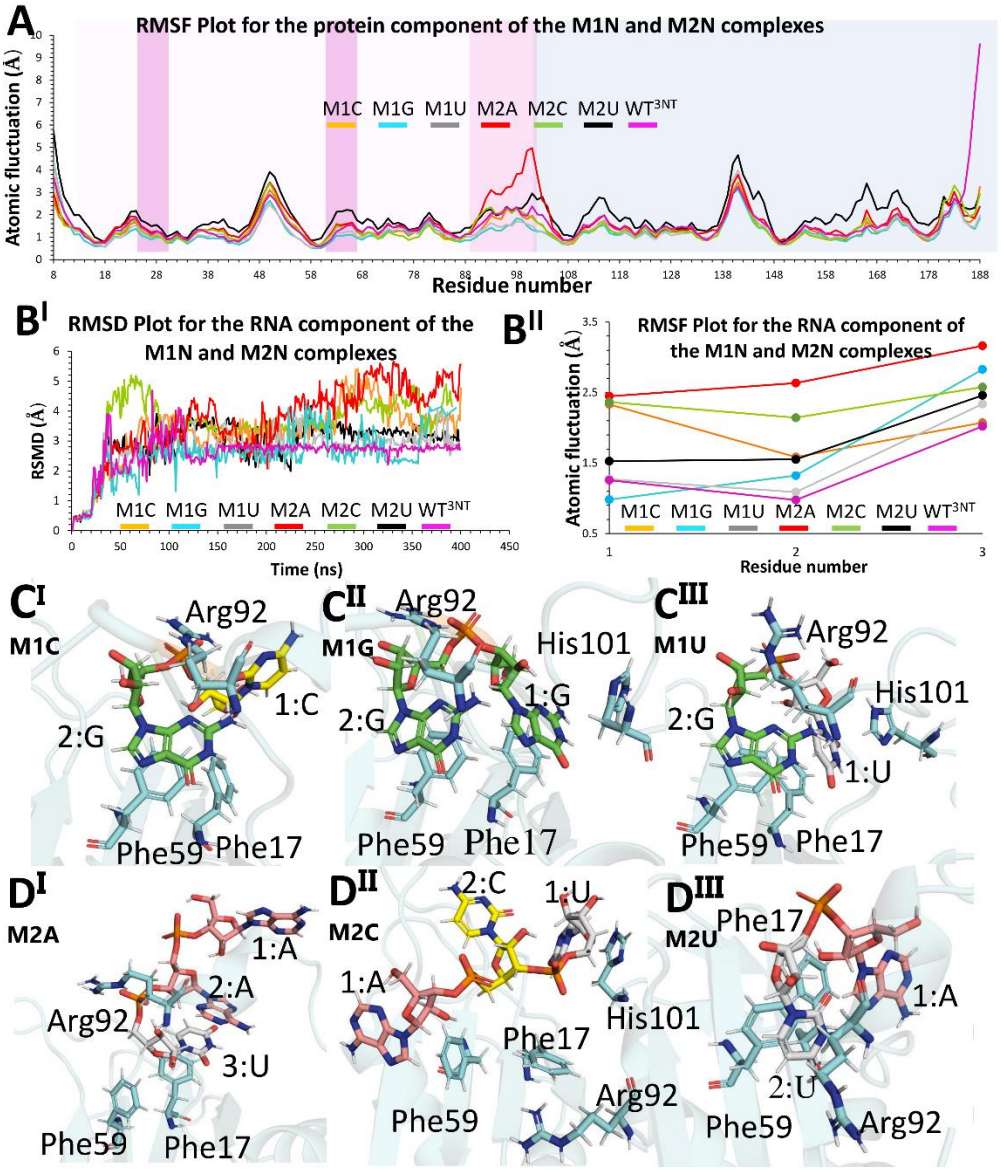


Figure 3.6: Interactions formed by the M1N and M2N complexes at 400 ns to investigate the effects of nucleotide modifications at positions 1 and 2. Nucleotides have been labeled with their positions and coloured green (G), pink (A), yellow (C) or grey (U), while all residues from the protein are in cyan. (A) The protein of the M2U has higher RMSF fluctuations, in the RNP and linker-loop regions, which is also true for M2A and M2C (Figure 4B′). M2A, in particular, has a sharp peak for RMSF in the linker loop pertaining to the protein residue instability due to the RNA, followed by M2A and M2C. Surprisingly, M1U and M1G indicate less fluctuations in the linker regions compared to the WT^{3NT}. (B)Indeed, M1U and M1G are the closest in RMSF fluctuations in RNA to the WT^{3NT}. Panel C: Replacement of A in position 1 mostly keeps the complex intact with most interactions conserved, with the exception of Cytosine in the M1C complex (C^{I}). Panel D: Contrary to Panel A, the modifications at position2 for all complexes has disturbed the native contacts to some extent. While M2A(D^{I}) and M2C(D^{II}) appear to have lost most of their

interactions, M2U(D^{III}) is still somewhat involved in the binding, especially with Phe59 while Arg92 has moved away.

3.3.3 Guanines can replace Adenines in their respective positions, but not vice versa

The M*N*3 complexes were assessed for binding in the same way as their M1*N* and M2*N* counterparts. Although MA3 and MG3 complexes have fairly stable RMSD values, MU3 and MC3 demonstrated similar fluctuations in RMSD (Figure 3.7, Appendix Figure A3). However, a more definitive trend is demonstrated by the RMSF analysis for both the protein and RNA (Figure 3.7B). MC3 overall has the highest fluctuations within RRM1, particularly in the RNPs and the linker loop, followed by MA3. Surprisingly, MG3 and MU3 demonstrate stability that is higher than the WT^{3NT}. RMSF fluctuations in the RNA also reinforce the trend of MA3 and MC3 being significantly more unstable than MU3 and MG3. Nevertheless, the binding free energy for the MG3 complex is overall the highest of ~ -65 kcal/mol, compared to -48.19 kcal/mol of the WT^{3NT} (Table 3).

Table 3.2: Binding affinities (kcal/mol) for the mutations performed at all positions with their respective standard deviations for the M*N*3 complexes. This set of data indicates that Guanine alone in the RNP-RNA interface can maximize the binding.

NAME	POSITION 1	POSITION 2	POSITION 3	BINDING AFFINITY (KCAL/MOL)	STANDARD DEVIATION
WT (PDB: 4YOE)	U	G	A	-48.19	6.33
MG3	G	G	G	-64.87	4.82
MC3	C	C	C	-31.76	6.45
MU3	U	U	U	-36.50	5.59
MA3	A	A	A	-36.33	5.24

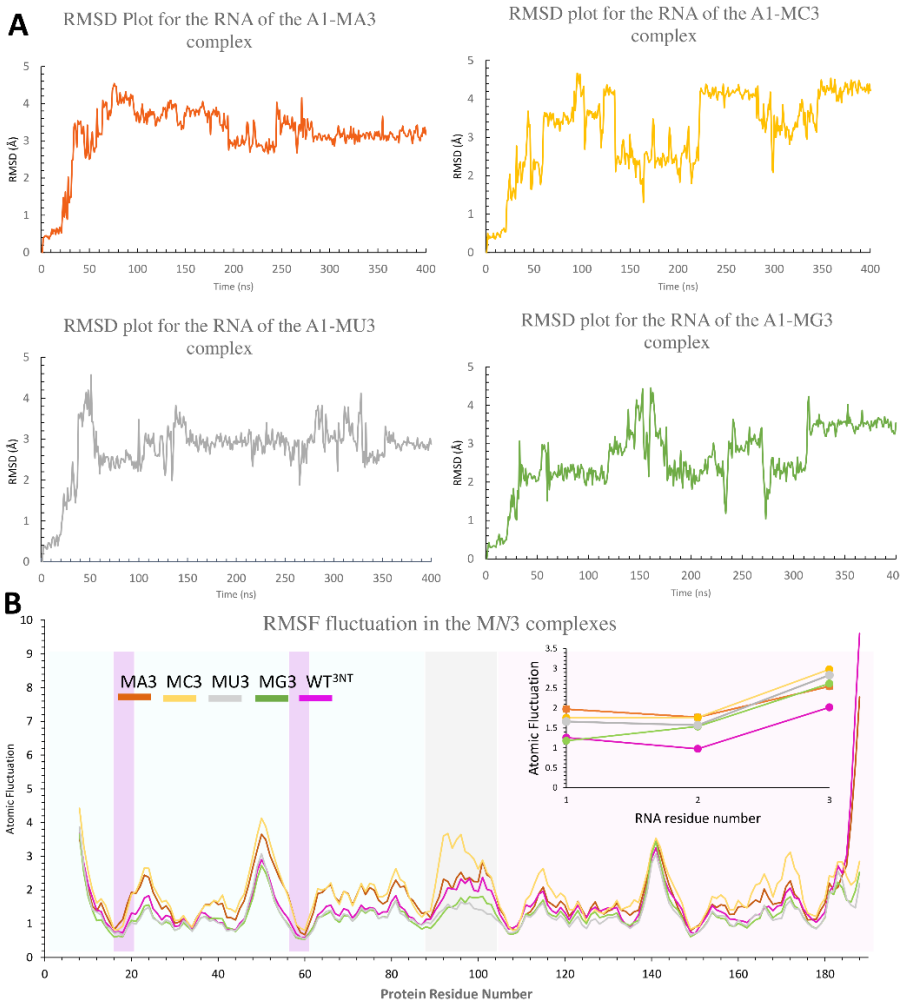


Figure 3.7: Evolution of the RMSD for the RNA component of the M*N*3-A1 complexes and the RMSF fluctuations of the complexes. (A) The stability of the complexes during the MD simulation was assessed by plotting the RMSD for all the atoms of RNA over the course of the simulation. RNA underwent conformational changes overall, adapting to the A1 RRMs and constantly shifting positions. (B) The RMSF fluctuations are low in the RNP regions with a better separation in the linker-loop region. Nevertheless, the trend demonstrates MG3 and MU3 more than the WT^{3NT} and MA3 having higher fluctuations followed by the MC3 having the most fluctuations. However, MG3 and MU3 RNA have distinguishable higher fluctuations relative to the WT^{3NT}. MA3 and MC3 stand out as having the highest fluctuations in RNA as they had for the protein.

Upon visual inspection, the Guanines in position 1 and 2 form stable contacts with Phe17, Phe59, Arg92 and His101, which resembles the signature native contacts seen in crystal structures but not the rest of the M*N*3 (Figure 3.8). While the other nucleotides do not form all of the native contacts, they always stack with Phe17 (Figure 3.8, panel B). However, stacking with Phe59 in position 2 is only seen with G. As mentioned previously and confirmed with Appendix Figure A3, Arg92 significantly supports G in position 2, giving it an advantage over other nucleotides.

This may be responsible for enhanced stacking with Phe59, since G in position 2 can simultaneously interact with both Arg92 and Phe59, keeping it in place. Figure A5 confirms the considerably high-capacity Guanine has for hydrogen bonds, being able to hold its interactions

with additional residues in RRM1. While A, C or U can stack with Phe17, one or more Guanines are needed to stably form interactions with Arg92 which seemingly plays a crucial role in enhancing stacking with Phe59 of the RNP1.

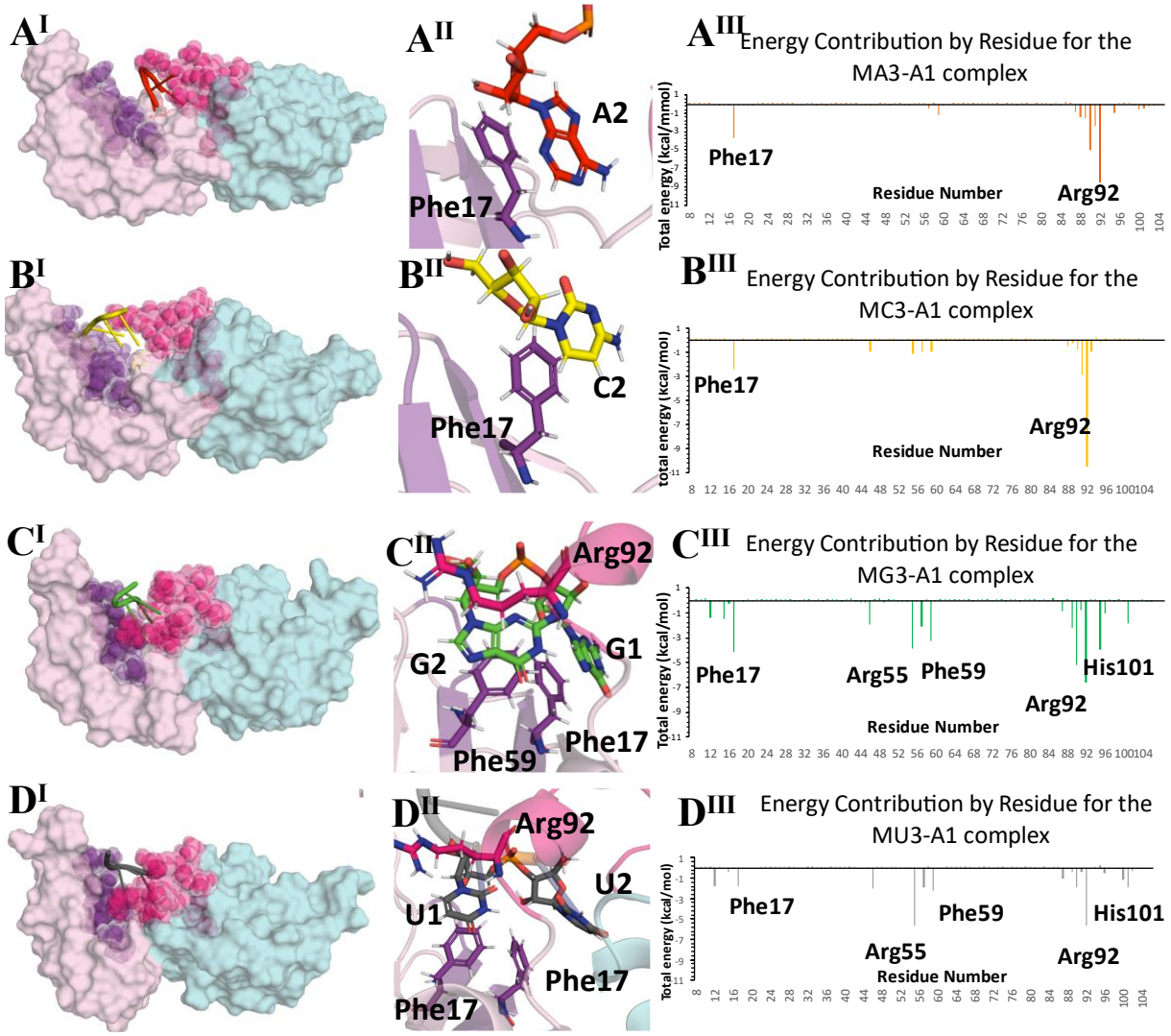


Figure 3.8: The conformation of the M*N*3-A1 complexes at 400 ns with focus on key molecular interactions and residues contributing the most to the binding. **(A)** The MA3-A1 complex comprising of just adenines in the binding pocket adopted a stable conformation although much different from that seen for the WT^{3NT} complex (A^{I}). This can be attributed to the loss of contacts with Arg92 and Phe59 (A^{II}), as also seen in the decomposition analysis (A^{III}). **(B)** A similar trend is seen with the MC3-A1 complex which comprises of cytosines in the binding pocket, which sits in the binding pocket like MA3-A1 (B^{I}) and has also lost contacts with Arg92 and Phe59 (B^{II} and B^{III}). **(C)** The MG3-A1 complex comprising of guanines displayed the highest binding affinity and this trend is visually confirmed with the presence of all notable contacts with Phe17, Phe59 and Arg92(C^{II} and C^{III}) as seen for the WT AGU-A1 complex (A). The GGG nucleotide also adopts a very similar conformation in the binding pocket to that of WT^{3NT} (D^{I}).

(D) Similarly, the MU3-A1 complex has a similar binding mode to WT^{3NT} and MG3-A1 but less affinity for Phe17 which reduced its overall affinity for A1 (D^{II} and D^{III}).

Finding 2: The analyses suggest that guanine is more preferable for binding with the A1 RRM binding site rather than any other types of nucleotides.

3.3.4 Longer ssRNAs reinforce the significance of Guanines

As with the mutants derived from the WT^{3NT}, the WT^{7NT} was used to create mutated versions of the 5PMG RNA ligand. Four mutant RNAs were created, having a GG, AA, CC or UU sequence occupying the positions where the AG motif would have been. The rest of the WT^{7NT} sequence was left unmodified.

As seen for the 3nt M1*N*, M2*N* and M*N*3 systems, the RMSD values for M*N*34 systems do not demonstrate an obvious trend in stability (Figure 3.9A, B). However, the RMSD values are slightly higher than the shorter 3nt systems, which could be explained if the longer length of RNA is taken into consideration (Figure 3.9 B). Longer RNA may allow the nucleotides to adapt to different conformations over the course of the simulation, especially since the 5′ and 3′ ends of the RNA are directly in contact with the solvent, with no protein residues stabilizing them.

The binding affinity values calculated (Table 4) along with visual inspection, (Figure 3.10) verify the trend seen earlier with the 3nt systems. Having additional Guanines in the binding pocket allow for the RNA ligand to adopt an accessible conformation whereby the nucleotide sidechains face downward into the pocket (Figures 3.10 A^{I}-D^{I}). Aromatic sidechains of the protein stack better with Guanines in the RNP site than having Adenines, Cytosines or Uracils (Figures 3.10 A^{II}-D^{II}). Guanine in position 4 of the 7-nt RNA maintains its contacts better with Arg92 and Phe59 comparatively (Figure 3.10 C^{II}). However, with the MMGBSA analysis, the MC34-A1 complex demonstrates higher energy contribution from Arg92 (Figure 3.10 B^{III}) than Guanine does in

position 4 (Figure 3.10 C^{III}). Visual inspection explains that Arg92 uses its guanidine group to interact with the negatively charged phosphate backbone of the RNA in MC34, and not with the nucleotide side chain. However, MG34 demonstrates the highest binding affinity value overall of -88 kcal/mol, even higher than that of the WT 5MPG ligand of -83.7 kcal/mol (Table 4). The trend of a GG motif having a slightly higher binding affinity than the AG motif was seen earlier with the MG3 and the WT^{3NT} complex.

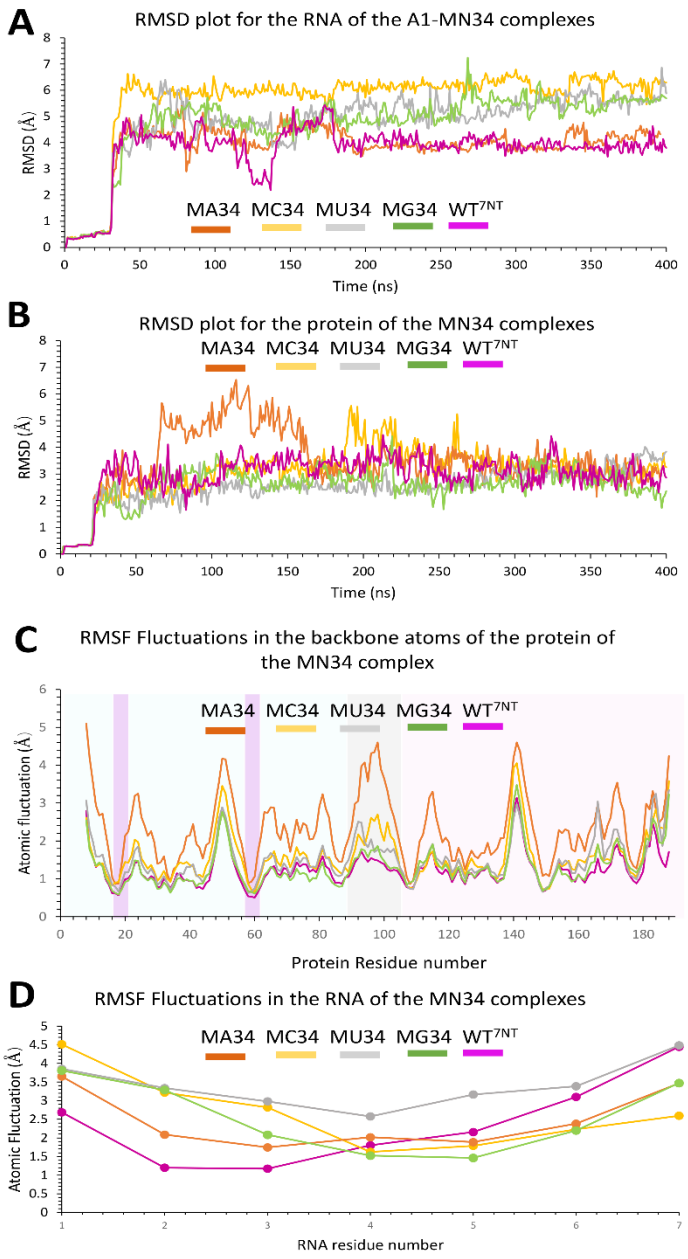


Figure 3.9: The RMSD and RMSF plots constructed for the ligand and protein components for the MN34-A1 complexes. (A) , the RNA component of the complexes overall has higher RMSD values as well as slightly more fluctuations. (B) As expected, the protein component for the RNA-A1 complexes illustrates stability for the entire simulation. Nevertheless, both components demonstrate stability for the complexes, indicating stable binding. (C) Similar to the trend for the MN3 complexes, MA34 has the highest protein RMSF fluctuation, followed by MC34, MU34 and MG34. Although all complexes share similar values for the RNP regions, the dispersion in values becomes mor explicit in the linker loop region. (D) This trend differs significantly for the RNA RMSF values. All the residues share higher fluctuations for positions 1, 5, 6 and 7 due to the lack of protein binding to those residues. Whereas, positions 3 and 4 have similar values for all complexes due to the deep binding into the RRM1 binding pocket.

This MG34 complex the strongest by residue contribution from the RNP motifs (Panel 3.10 C). A detailed hydrogen bond analysis for the prominent bonds mediated by Guanine in position 4 for the MG34 is illustrated in the Appendix Figure A5. In addition to the notable Arg92 contact, G4 also interacts with Asp42, which although is not part of the RNP motifs, but still adds to the binding energy. Further elaborating on the role of Guanines makes the additional G5 noteworthy as mentioned previously with Figure 3.5C and D. The presence of G5 results, at least partially, in the MN34 systems having a higher binding affinity than their 3nt counterparts. Some of the enhanced affinity can also be attributed to the longer length of RNA which may add to electrostatic contacts. All in all, the recurring trend in increased binding affinity by having one or more guanines in the binding pocket may have a physiological significance[151].

Table 3.3: Binding affinities (kcal/mol) for the mutations performed at all positions with their respective standard deviations for the M34*N* complexes. This dataset complements the previous trends identified in the shorter 3nt RNA systems with reinforcing the idea that G is sufficient and necessary for binding.

NAME	POSITION 3	POSITION 4	POSITION 5	BINDING AFFINITY (KCAL/MOL)	STANDARD DEVIATION
WT^{7NT}	A	G	G	-83.7521	4.9535
M34G	G	G	G	-88.04	6.64
M34C	C	C	G	-51.31	5.56
M34U	U	U	G	-48.04	5.76
M34A	A	A	G	-57.69	6.51

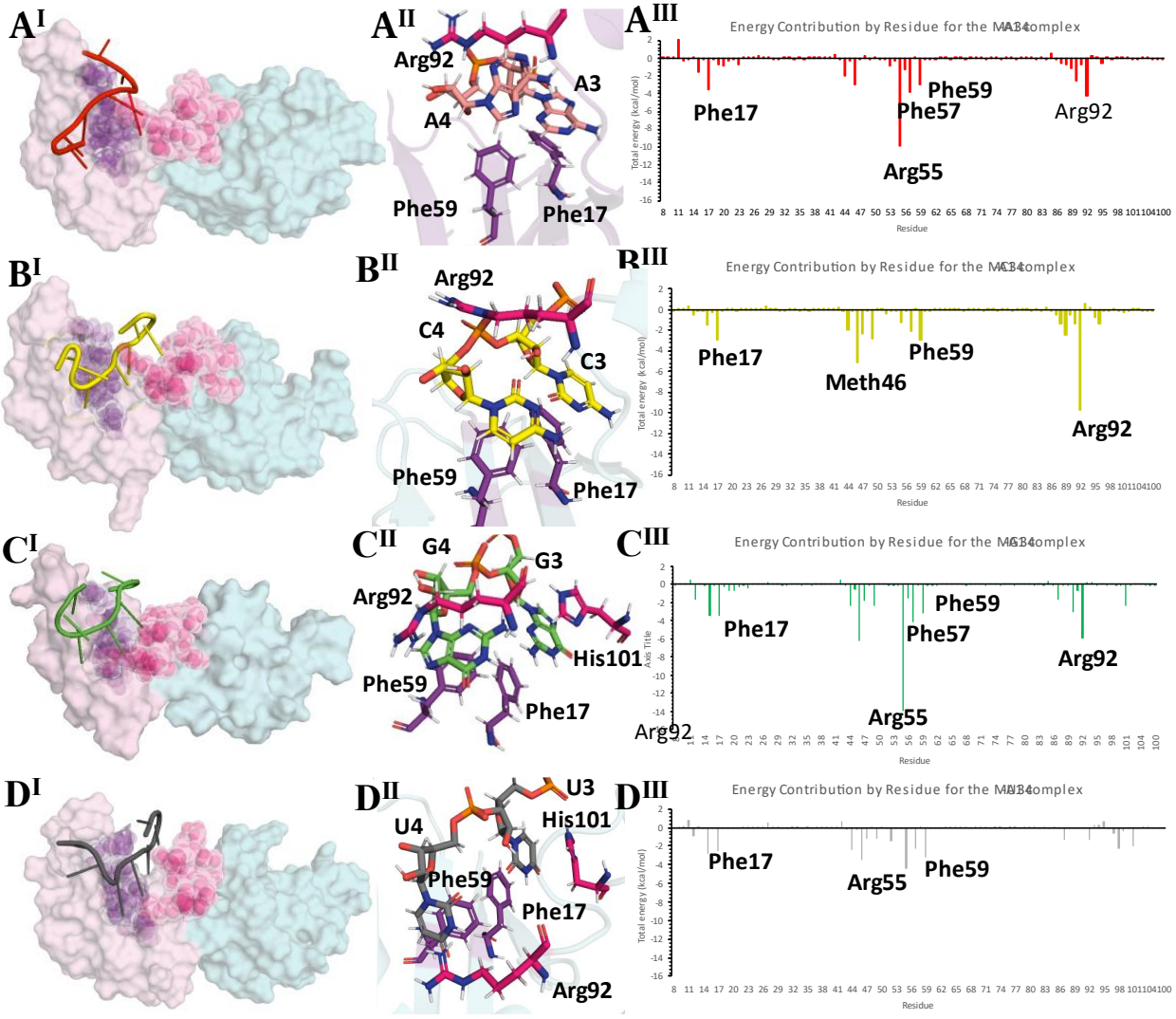


Figure 3.10: The conformation of 7nt RNA ligands and the residues contributing the most to their binding. **(A)** The M34A-A1 complex comprising of adenines in the binding pocket adopted a significantly less stable conformation, with much of the side chains of nucleic acids facing away from the RNP motifs and the 5' end seems to be sticking out and not interacting with the protein surface (A^{I}). This is unsurprisingly confirmed with the close-up visual inspection in B^{II}, with Phe59 unable to stack with A, while Phe17 still can. As per the MM-GBSA decomposition analysis, most of the binding energy is coming from Arg55 which is likely forming electrostatic contacts (A^{III}). **(B)** Similarly, M34C adopts a conformation that is much different and less stable compared to the WT^{7NT} (B^{I}). Zooming in onto the native interactions confirms this disruption. While Phe17 is able to maintain some stacking, Phe59 tries to stack but it is not a proper face-to-face stacking ($B^{I)}$. Arg92 however, is able to interact with the phosphate backbone and this energy contribution is confirmed with the MM-GBSA decomposition analysis (B^{III}). **(C)** M34G can form extensive contacts with the RNPs considering how deep the nucleic sidechains extend into the binding pocket (C^{I}). The M34G-A1 complex displayed the highest binding affinity and this trend is visually confirmed with the presence of all notable contacts with Phe17, Phe59 and Arg92(C^{II}

and C^{III}). A significant contribution also arises from the Arg55 residue of the RNP1. **(D)** Similarly, the M34U-A1 complex adopts a conformation such that its nucleic sidechains face the RNPs to be able to bind and form pi-pi stacking interactions. A closer visual inspection reveals that it is able to stack, however, Arg92 has shifted such that it can no longer form strong electrostatic interactions, thus resulting in reduced affinity (D^{II}). The MM-GBSA decomposition analysis conforms this data (D^{III}).

This redundant binding mode can be accredited to A1's role in telomere maintenance, where it binds to the TTAGGG repeats that are abundantly present in telomeres.[151] A1's role in transcription by interacting with G-quadruplex structures using its UP1 domain and its LCD domain has been well-established.[152,153,154,155] It is also known that the RGG box in the LCD is responsible for contributing to G-quadruplex binding and can not bind linear ssRNA.[156] However, A1's co-crystallized ligands in the PDB do not have consecutive Guanines to adopt a G-quadruplex conformation, but have short and linear RNA or DNA ligands that have Guanines interspersed instead of being long stretches. This may account for the redundancy of AG motifs in A1's linear nucleotide ligands to ensure only the UP1 domain interacts with these ligands. Therefore, while stretches of Guanines, by our analysis, are sufficient to drive RNA-A1 cohesion, they may result in RNA adopting a G-quadruplex structure, which may activate LCD binding to it. The interaction contributed by the LCD may not be necessary for all types of physiological RNA ligands for the A1 protein, which may instead contain the AG motifs. Having AG motifs instead of stretches of Gs may ensure the linearity of the RNA ligands as seen with the co-crystalized ligands in the PDB. It would be interesting to see A1 interacting G-quadruplexes in future studies and the effects of nucleotide substitutions on binding within the quadruplex structures.[152]

Finding 3: Collectively, our data reveal promising insights regarding Guanine's enhanced capacity to bind A1's RRM1. Nucleotides rich in G nucleotides have exhibited higher binding affinity to RRM1.

3.3.5 Known RNA oligonucleotides with different guanine content exhibit variable affinity to A1 RRMs[14]

A previous study conducted by Rollins et. al[56] determined 7-nucleotide long RNAOs that bind A1's RRM1 and their dissociation constants (K_D values). Inspired by these RNAOs, our collaborators expanded the 5′ and 3′ ends of the RNAOs to make 27-nucleotide long RNAOs, (Figure 3.11) for which *in vitro* data was supplemented with *in silico* analysis reported below[157].

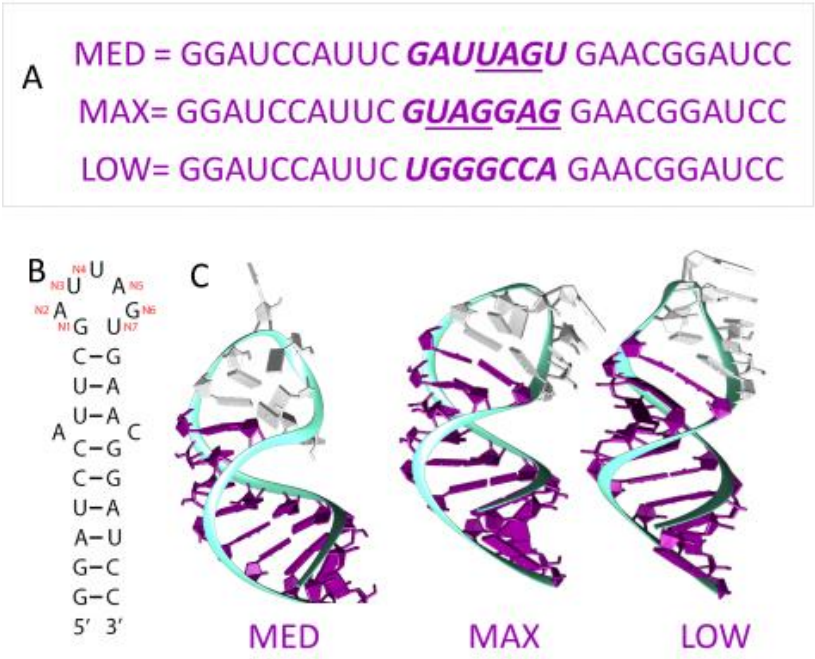


Figure 3.11: The sequences and structures of the three RNAOs that bind A1 *in vitro*.[14] (A) The full sequences of the 27-nt RNAOs are displayed with the bolded nucleotides being what constitutes the apical loop, with the expected nucleotides active in binding underlined. **(B)** An illustration of the full 27-nt RNAO with the apical loop labeled at the top. **(C)** The i*n silico* analysis conducted for MAX, MED and LOW was conducted after modelling them, taking note of the amount of self-stacking dsRNA normally has, with some free bases towards the apical loop.

3.3.5.I Modeling the RNAO-A1 complexes and optimizing them with MD

The secondary structural information of the RNA was predicted using the RNAFold program available within the Vienna RNA package.[95] Both the minimum free energy (MFE) and partition function (PF) algorithms for predicting the optimal secondary structures of RNA oligos with

minimum free energies that are calculated using dynamic programming were employed.96 A recent study[97] compared the performance of different RNA prediction program and noted that RNAfold was able to calculate accurately predict the secondary structures of RNA and that its folding scores were in good agreement with observed free energies.[97] The secondary structural information of RNA oligos (MAX, MED, and LOW RNAs) were used to model their 3D structures using RNAComposer program[98] that functions based on a fragment assembly approach. In RNAComposer, the input secondary structural information is broken into multiple fragments with overlapping native base pairs, which are then matched with the known 3D structural fragments available within RNA FRABASE database.[99] The identified fragments are then assembled together using the overlapping pairs to build the complete RNA 3D models. The predicted 3D structural models of RNA oligos were used to model their interactions against the A1 protein.

Following RNA-modeling, RNA-protein docking calculations were performed to model the structural complexes of RNA oligos with the A1 RRMs. While the structures of RNAOs in this study were modelled computationally, a previously existing high resolution X-ray crystal structure of a monomeric hnRNPA1 RRM2 (PBD: 4YOE) was used for docking. This structure was chosen as it represented the complete RRM domain (including both RRM1 and RRM2) of hnRNPA1 and was also co-crystallized with a 3-nucleotide long RNA. The docking calculations were carried out using the MDockPP program[63], which employs a 2-tier screening approach: In the first step, it involves an altered Fast Fourier Transform (FFT) algorithm to identify putative binding poses based on the shape complementarity within the target molecules; and the initial binding poses are reassessed using an ensemble docking algorithm accounting for molecular flexibility and a knowledge-based scoring function (ITScorePP). The RNA oligos in this work are 25-nucleotide long sequences with only variations seen in the 7 nucleotides (from 11-17 positions), which

suggests that the binding affinity differences amongst the RNAOs are plausibly driven by this short stretch of sequence. Therefore, we selected these 7 nucleotides as active residues for docking against A1. Whereas, for the protein residues, we defined Phe17, Phe59 and His101 as the active residues for docking as they were found to interact with RNAs and DNAs in the previously reported structures in the PDB23 (e.g., PDB: 4YOE). We further specified the interactions of adenines and guanines from the apical loop (i.e., 11-17 positions) with that of the selected protein residues as optional interface residues. The binding poses from molecular docking were visually inspected to choose only the complexes that resembled the native RNAOs-A1 RRM interactions in the PDB, in which an adenine and/or a guanine engaged with the active site residues in A1 RRM1. This resulted in filtering of the best RNA-A1 RRM complexes for further optimization.

MD and binding-free energy calculations were performed in the same manner as done in section 3.2, except for the restraints applied during MD, which differed slightly attributing to the longer length of RNA. To be precise, the restraints applied following equilibration were 0.5 kcal/mol Å−2 on the solute atoms followed by a 10 ns long MD simulation with a lowered restraint of 0.2 kcal/mol Å−2 only on the RNA atoms and a Phe17 residue that was reported to be a key player in RNA recognition in the earlier studies.[65] Subsequently, another 10 ns long simulation with restraints applied only on the apical loops of the RNA oligos (11-17 nucleotide positions). This multi-stage MD simulation protocol was employed so as to allow the A1 and RNA oligos to adapt their interactions (‘induced-fit’ effects) by minimizing non-specific electrostatic interactions that are commonly seen in RNA-protein complexes.[67]

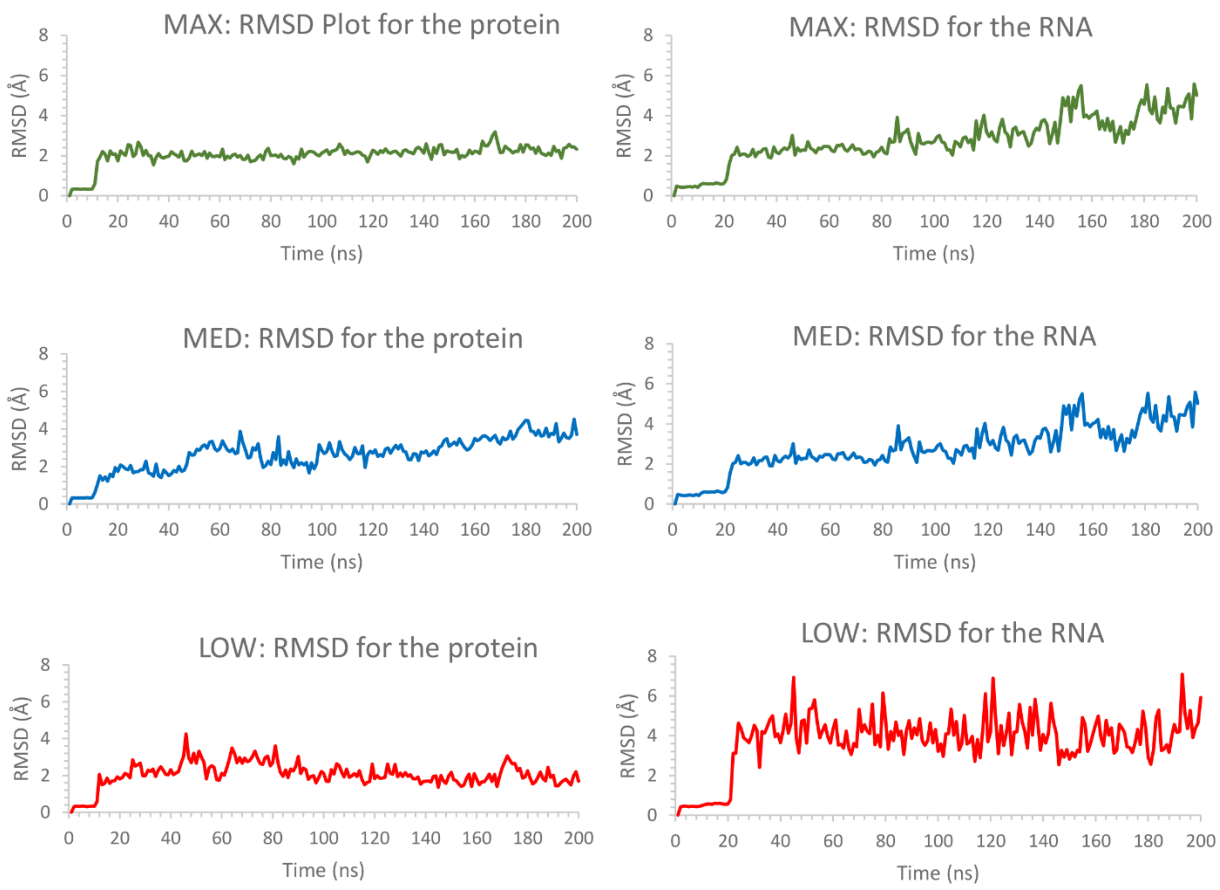


Figure 3.12: Evolution of RMSD of the RNA-A1 complexes. The stability of the complexes during MD simulation was assessed by plotting the evolution of backbone RMSD of A1 RRMs (on the left-side panel) and all the atoms of RNAOs (on the right-side panel). As seen in the plots on the left, the protein has much lower RMSD values than those of RNAOs (right) for all three complexes. RNA underwent more conformations changes to adapt to its binding with A1 RRMs that was relatively stable during simulation.

3.3.5.II The AG motif and G nucleotides overall provide enhanced RNA-binding preference in RNAOs

To gain insights into the binding interactions between RNAOs and A1 RRM at molecular-level, the RNAO-bound complexes were modeled as described in the Methods section. The structural models of the complexes were initially predicted through RNA-protein docking and were optimized through 200 ns long MD simulation so as to allow induced-fit effects and overall conformational dynamics from RNA binding to A1 RRM. Assessment of RMSD evolution during

the course of MD confirmed the overall stability of the complexes (Figure 3.12). While the protein backbone and the RNAO structures in the MAX and LOW complexes reached a plateau > 80 ns, the MED complex exhibited slightly higher fluctuations indicating the changes in the RNA-protein interactions. Binding free energy values in Table 3.4, predicted from the last 10 ns of the equilibrated MD trajectories of the complexes suggested that the full-length MAX RNAO has the strongest affinity to A1 RRM with a value of -53.9 kcal/mol, as compared to that of the MED RNAO (-49.6 kcal/mol) and LOW RNAO (-23.5 kcal/mol). This relative ranking of the predicted binding affinity is in good agreement with the previously reported Kd values from the literature[94] , A1 clustering response (Appendix Figure A8) and the thermal shift data from experiments from our collaborators (Appendix Figure A9), which confirmed the validity of the models. Binding mode analyses of the complexes (Figure 3.13) revealed that the RNAOs bound to the RRM1 domain of A1 and their RNA-protein interactions were driven through the 7-nucleotide apical loop (11-17 sequence positions) from RNAOs and the RNP motifs (residues 15-20 and 55-60) in RRM1 and the RRM1/2 linker loop in A1.

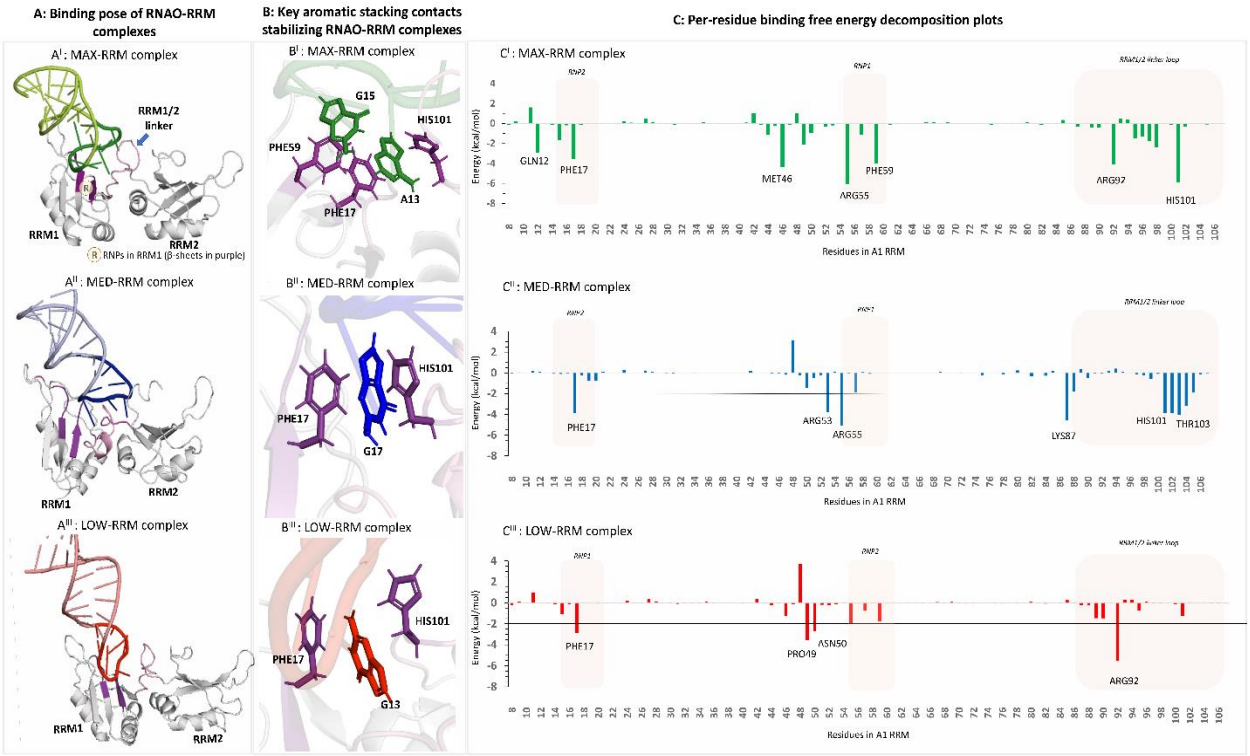


Figure 3.13: 3D structural models of the RNAO-RRM complexes (A-B) and the key residues contributing to their binding free energies (C). A. The structures of the complexes (A^{I}-A^{III}) describe that the binding of RNAOs with A1 RRMs were mediated through the interactions of the apical loops of the RNAOs with that of the RNPs from RRM1 and the RRM1/2 linker loop (shown in purple and marked in A^{I}). B. The close-up views of the binding sites of the complexes reveal key aromatic stacking interactions between RNAOs and RRM. MAX RNAO made two aromatic interactions such as PHE17-A13-HIS101 and PHE59-G15, which are consistent with the interactions reported in the known oligos-A1 RRM complexes in PDB (B^{I}). The MED (B^{II}) and LOW (B^{III}) RNAOs exhibit only a single aromatic stacking rendered by a guanine and PHE17. C. Per-residue decomposition analyses identified other key residues in RRM1 that contributed to the binding free energies of the RNAO-RRM complexes. A number of residues from RNPs and the RRM1/2 linker loop contribute to the binding free energy of the MAX complex (C^{I}). The binding free energy of MED-RRM complex is driven mostly by residues from RRM1/2 linker loop and fewer residues from RNPs (C^{II}). Apart from the stacking contact with PHE17, the binding free energy of the LOW-RRM complex is dominated by non-specific electrostatic interactions of residues not part of RNPs (C^{III}).

Whereas the segment other than the apical loop in RNOs did not make significant contact with the A1 protein surface. The 13AGG15 fragment in MAX RNAO played a central role in its recognition by the A1 RRM: especially, the A13 and G15 nucleotides from MAX formed aromatic stacking interactions with upward facing Phe17 and Phe59, respectively, (Figure 3.13 AII). These interactions can be considered as native signature contacts for nucleotide binding to A1, as

confirmed by the binding poses of different DNA and RNA molecules against A1 RRMs reported in the PDB.[2,74,75,76] Further, it was previously reported that specific binding of RNA to RRMs is supported by the aromatic sidechains in RNPs that adapt upward conformation.100 Per-residue energy decomposition analyses revealed additional key residues that contributed to (< -2 kcal/mol) the MAX RNAO-A1 RRM binding free energy (Figure 3.13 AIII). These include T-shaped aromatic stacking of His101 with A13 and salt-bridge interactions of Arg55 and Arg92 with those of the phosphate groups from G14 and G15 nucleotides (Figure 3.13, 3.14). These findings together explain the superior binding and activity rendered by MAX RNAO against A1. In MED and LOW RNAOs, since the adenine side chains are located inwards, a guanine molecule (G17) engaged in the aromatic stacking contacts with Phe17 and His101 (Figure 3.13 BII-III), which is usually rendered by an Adenine as seen in our MAX model and the previous PDB RNA/DNA-bound A1 RRM structures.

Table 3.4: **Comparison of the predicted binding free energies of RNAO-A1 RRM complexes against the previously reported K_d values and the corresponding Tm values from thermal shift assay experiments from this work.**

Complex	Binding Affinity (kcal/mol)	K_d (nM)[#]	Tm (in °)[$]
MAX	-53.9 +/- 5.3	19.4	61.83
MED	-49.6 +/- 5.8	27.8	61.38
LOW	-23.5 +/- 4.4	598	60.77

[#]The K_d values are from Rollins et al.[159] These values for the binding of only the 7-nucleotide loop from the RNAOs against A1 RRM.
[$]Melting temperature of A1 RRMs when treated with full-length RNAOs at 75 μM in this work.

Nevertheless, the apical loop of MED RNA engaged mostly with the RRM1/2 linker loop formed by residues His101-Val104, rather than the RNP motifs (refer to decomposition plot in Figure 3.13 CII). Whereas nucleotides outside of the apical loop in MED such as A7 and U8 formed

electrostatic interactions with Arg53 and Arg55 residues in A1 RRM1 (Figure 3.15). As a result, the MED RNAO exhibited weaker affinity to A1 RRM when compared to MAX. With respect to LOW, apart from the G13-Phe17 aromatic contact, it did not make any native contacts with A1 RRMs but only involved in non-specific electrostatic interactions with a small set of residues (see in decomposition plot in Figure 3.13 CIII), which led to its weakest affinity to RRMs. Therefore, structural insights from our models helped to explain the differential binding affinities and variable thermal shift response (Appendix Figure A10) of the RNAOs against A1 in this study.

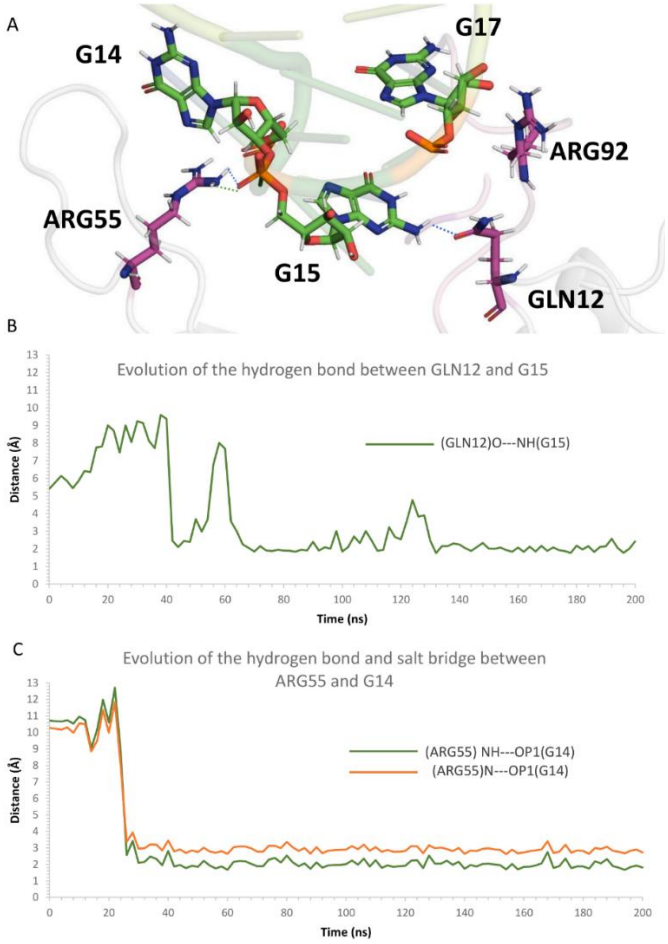


Figure 3.14: Illustration of the prominent electrostatic interactions in the MAX-A1 complex. (A) A 3D representation of key salt-bridge and hydrogen bond interactions between nucleotides such as G14, G15 and G17 against ARG55, GLN12 and ARG92, respectively. The selected amino acids and nucleotides are shown as stick representation as the other segments of the binding pose are shown as cartoon representation in the background. Evolution of distances between the side-chain carbonyl group of GLN12 and purine ring of G15 **(B)** confirmed that these interactions formed after ~60 ns and remained mostly stable until the end of simulation. In a similar nature, the distance evolution between the side-chain amino groups of ARG55 and phosphate group of G15 **(C)** described that this pair established a salt-bridge (N-OP1 shown in orange) and a hydrogen bond (NH-OP1 shown in green) after 20 ns and maintained them throughout the course of MD simulation. These confirm the importance of these interactions in stabilizing the MAX-A1 complex.

Therefore, these results are consistent with our findings (based on WT^{3NT} and WT^{7NT}) that guanine richness in RNAs can help enhance their binding with A1 RRMs and impact its pathophysiological processes.

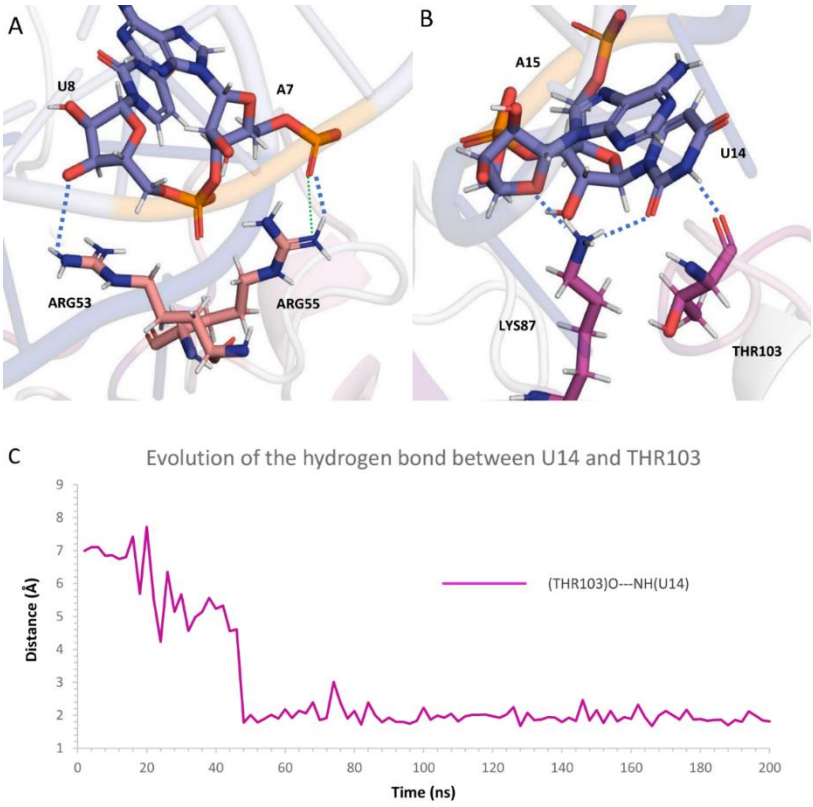


Figure 3.15: Illustration of the prominent electrostatic interactions in the MED-A1 complex. 3D representations of key salt-bridge and hydrogen bond interactions between key nucleotides from MED RNAO and amino acids from RRMs are shown in A-B. ARG53 and ARG55 formed dynamic salt-bridge contacts with A7 and U8 nucleotides from MED RNAO (A); while the side-chain amino group of LYS87 and backbone of THR103 made hydrogen bond contacts with A15 and U14, respectively, (B). The evolution of distance between the backbone carbonyl oxygen atom in THR103 and pyrimidine ring of U14 confirmed the stability of their hydrogen bond during the course of MD simulation.

Finding 4: Consistent with data obtained from experiments on the WT^{3NT} and WT^{7NT} complexes and their mutants, higher Guanine content in RNA can help enhance RNA binding, which may have important physiological or therapeutic consequences.

3.3.6 Mutating critical A1 residues involved in MAX RNAO binding

The previous sections highlighted the importance of guanine composition in RNA recognition by A1's RRM1. Some key residues were highlighted as critical with MD trajectory and binding free energy decomposition analyses of the WT and mutant models of different lengths of RNAs (3NT, 7NT, and 27NT) highlighted a few residues from RRM1 that consistently played critical roles in its binding with RNAs. These residues include Phe17, Arg55, Phe57, Phe59, Arg92, and His101 that are dispersed across the RNPs of RRM1and the interdomain linker of A1. In the previous section, MAX RNAO rescued A1's clustering *in vitro* (Appendix Figure A8) which is explained by its ability to form specific and native interactions with the A1 protein.

The next step was to report novel protein residues in the A1 which may have been overlooked in previous research as RNP residues are given the most significance. Using MAX RNAO-A1 as a positive control, two mutant complexes were created by mutating A1 residues (not mutating MAX). Two systems, His101A-MAX and Arg92-A max were prepared by mutating His101 to Ala and Arg92 to Ala, respectively. This was done in Chimera by directly modifying the the original MAX-A1 complex prior to MD.

RMSD analysis for the protein components of the complexes demonstrates the His101A mutant having relatively less stability than the A1-MAX complex (Figure 3.16A). Since His101 and Arg92 are residues from the linker loop, that increases the probability of inducing displacement in the interdomain region due to the flexibility, thus increasing RMSD values. This notion is confirmed with the RMSF plots for the protein (Figure 3.16C) where the MAX-A1 complex has the lowest fluctuations by residue for the protein. The trend is clearer in the linker loop region (residues 85 to 102) while some residues in the RRM2 region (residues 138 to 148) do not follow the trend. While the rest of the residues in RRM2 also have comparable values between the three

systems, it could be hypothesized that MAX binding to RRM1 does not impact the RRM2 as much due to the distance separating the two RRMs.

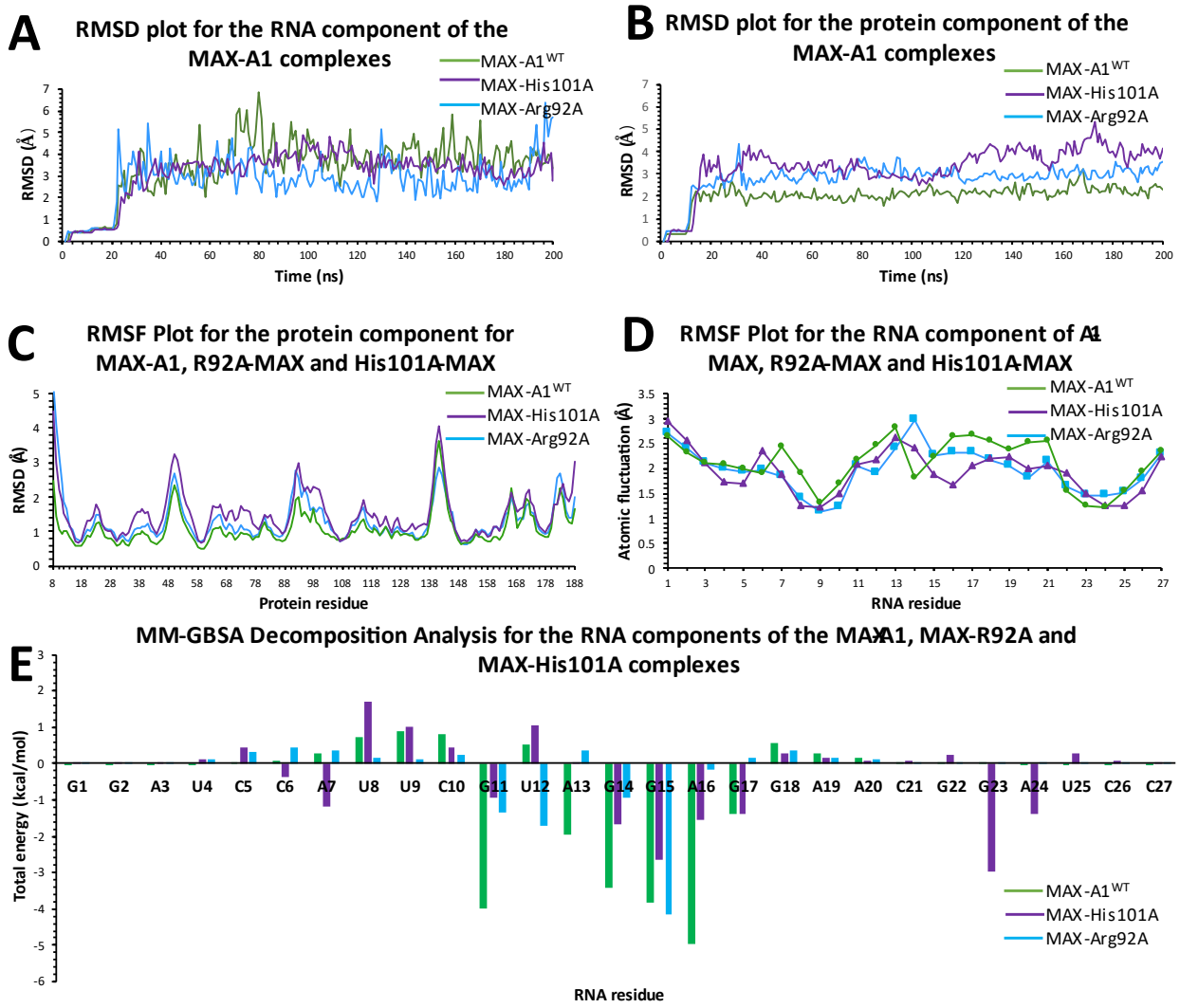


Figure 3.16: RMSD and RMSF analysis of the His101A-MAX and R92A-MAX in comparison to the A1-MAX. (A) The three systems displayed comparable values for the RNA components of the complexes. (B) The protein component of the MAX-A1 was more stable compared to either mutant A1, where the His101A mutant had the highest amount of instability. (C) Similarly, the RMSF fluctuations by residue for MAX were most stable with the exception in some of the RRM2 residues (138-148). (D) RNA residues demonstrate no obvious trends overall with residues varying in stability based on position. The 3′ and 5′ terminal ends with less overall contact with the protein have similar RMSF values. However, residues 11-17 that make up the apical loop hence directly interact with the RNP binding pocket have MAX-A1 G14 more stable followed by MAX-His101A and then MAX-R92A, while A13 has similar fluctuations. The His101A-MAX G15 has less fluctuations than the G15 in MAX-A1 or MAX-R92A, which have similar values. (E) The MM-GBSA contributions by RNA explains interactions by RNA residues

contributing to binding. A13 in MAX-His101A has almost no contribution to binding affinity while Arg92A-MAX has a repulsive effect and A1-MAX has a decent contribution to binding by A13. The subsequent RNA residues in the apical loop, residues 14-17 demonstrate a clear trend where MAX-A1 has the highest binding affinity values, followed by the A1 mutants.

However, the trend in RNA RMSD fluctuations demonstrate no obvious trend (Figure 3.16B). While the MAX-A1 complex has stable RNA residues in the beginning of the simulation, all three complexes have similar values after ~130 ns with a mild separation. The rend in RNA stability has MAX-R92A as the most stable, followed by His101A-MAX and then MAX-A1. RMSF fluctuation by residue indicates that the terminal ends of MAX regardless of the complex have very little differences in stability, most likely due to less surface-area contact with the protein. The trend shows separation in the apical loop (residues 11-17), where only G14 from the MAX-A1 complex has the lowest fluctuation, while the rest of the MAX-A1 RNA residues have higher fluctuations compared to the mutants. The MM-GBSA plot (Figure 3.16 E) offers no explanation as to why the trend in RNA RMSF is unique, since most of the residues in the apical loop in MAX-A1 contribute to much higher binding affinity values compared to the mutants. In fact, A13 almost contributes to no binding affinity in MAX-His101A and is repulsive in its interactions in MAX-R92A.

Both mutant proteins have a decline in binding affinity compared to the MAX-A1 protein (Table 6). This is confirmed with residues 1-9 in the RNA having more repulsions in the mutant complexes than the MAX-A1 while the apical loop overall (residues 11-17) demonstrate more binding affinity in the MAX-A1 complex (Figure 3.16 E). Interestingly, there is almost a 16 kcal/mol difference in binding affinity values between the MAX-His101A and MAX-Arg92A. This difference in binding affinity where the drop is more significant when Arg92 is mutated versus His101 is surprising (Table 3.5). Considering the fact that the MAX-A1 complex reported earlier had increased binding affinity contribution by His101 than Arg92, the drop in binding

affinity was expected to be more significant when mutating His101. Two observations help explain the decline in binding affinity in Table 3.5. The compensatory interactions by residues G23 and A24 in the His101A mutant (Figure 3.16 E) add to the binding affinity in MAX-His101A. Arg55 from the protein in His101A1 also adds to the binding affinity values where it has a contribution even higher than in the MAX-A1 complex, while the Arg92A mutant is unable to do so (Figure 3.17 A^{II}, B^{III}). These two compensatory mechanisms by both the protein and the RNA in the MAX-His101A complex explain the differences in binding affinity values. Interestingly, the MAX-His101A complex, even with compensatory interactions, has a lower binding affinity than the WT RNAO-A1 complex (Table 3.4), which confirms the overall lack of affinity that the His101A mutant protein has.

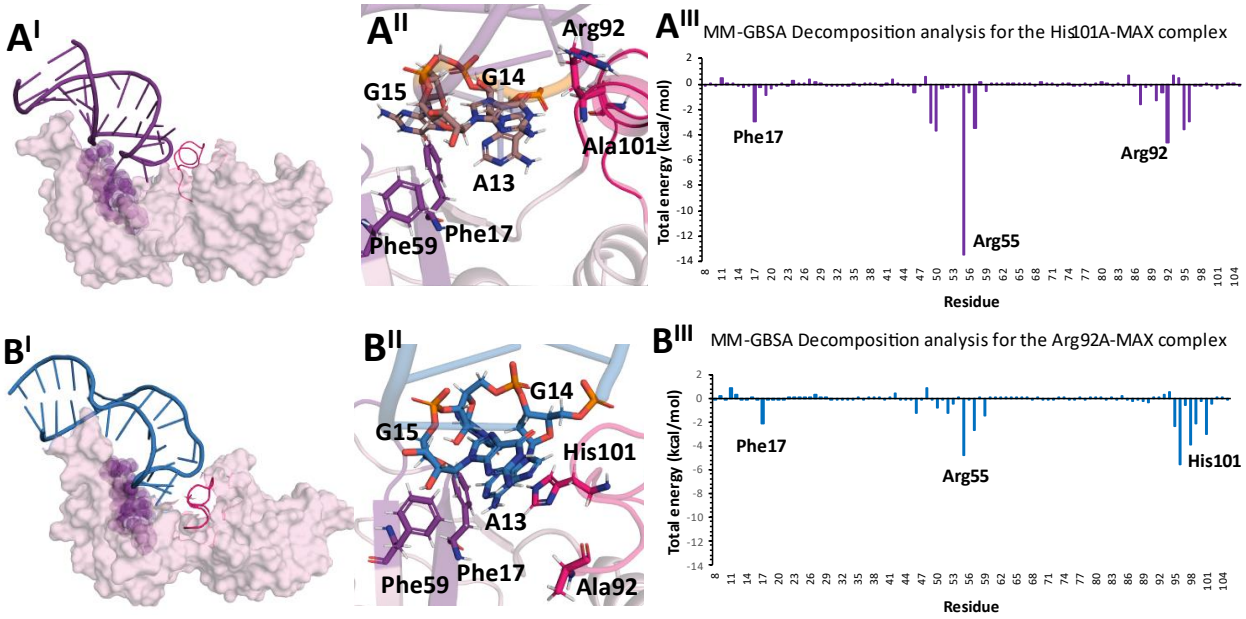


Figure 3.17: The MAX RNAO is unable to bind the A1 protein when His101 and Arg92 are mutated to alanine. (A) At 200 ns, The His101A-MAX complex demonstrates loss of native contacts. (A^{I}) It has nucleotide sidechains mostly facing away from the RNP pocket (purple). (A^{II})This is elaborated on by the fact that the linker loop (magenta) has been displaced, resulting in Arg92 loosing contact with G15 which simultaneously breaks contact with Phe59. A13 is the only residue interacting actively with Phe17 in a native conformation, via a T-shaped stacking

interaction. MM-GBSA binding affinity analysis confirms the reduction overall in contacts with the RNPs with the exception of Arg55, which forms electrostatic contacts with the phosphate backbone of the RNAO. **(B)** Similarly, the Arg92A mutation has the MAX-RNA reduce contacts with the RNPs. Very few nucleotide side chains face the RNP binding pocket and this is explained by the G15 moving completely away from the Phe59 and Ala92 (B^{I}) and (B^{II}). Phe17 stacks with A13, and His101 but with reduced affinity. This is confirmed by the reduction in binding affinity for both Phe17 and His101 and even Arg55. Non-specific interactions from the linker-loop keep the complex intact (B^{III}).

Table 3.5: **Binding affinity values for the MAX-A1, MAX-His101A and MAX-Arg92A complexes calculated at 200 ns.**

Complex	Binding Affinity (kcal/mol)
MAX-A1	-53.9 +/- 5.3
MAX-His101A	-48.5+/- 6.3
MAX-Arg92A	-32.6+/- 6.6

However, having compensatory interactions in the MAX-His101A does not result in stable interactions with the RNP motifs (Figure 3.17 panel A). While visual inspection demonstrates a T-shaped interaction between A13 and Phe17 (Figure 3.17 A^{II}), the MM-GBSA decomposition for A13 in Figure 3.16 E illustrates no contribution from A13 to the binding affinity. Therefore, a native pi-pi stacking interaction can not be confirmed between A13 and Phe17. This is also true for Phe59 which has lost the stacking interaction with G15 found originally in the MAX-A1 complex. However, Arg92 is able to form electrostatic contacts with the phosphate backbone of the MAX RNAO, resulting in significant contribution (Figure 3.17 A^{III}). Similarly, the massive increase in the binding affinity value for Arg55 can be explained as a compensatory interaction by electrostatic contacts with the phosphate backbone (Appendix Figure A7).

Indeed, both the His101A mutation and the Arg92A mutation impede MAX binding to the protein *in vitro*.[38] This is proven by clustering analysis of the A1 protein that can be induced by Optogenetics. MAX binding to A1 prevents its clustering[38], but if A1 has an His101A mutation (Appendix Figure A10) or an Arg92A mutation, (Appendix Figure A11), the clustering of A1 can

not be prevented by MAX RNAO. Since there is no statistical significance in A1 clustering with or without MAX RNA binding to the protein, it effectively confirms the essential role both residues play in binding RNA.

.

Chapter 4: The influence of Interdomain contacts on the dynamics of the full A1 protein

4.1 Introduction

The hnRNPA1 protein, as mentioned earlier, has N-terminal RNA binding domains and a disordered C-terminal domain or low-complexity domain (LCD).[6,83] While the LCD is often considered to be the main contributor for the protein's LLPS and pathophysiological implications [6], it does have crucial roles such as mediating nuclear localization[74], binding certain RNAs [160] and protein-protein contacts[83].

The lack of defined secondary structural characteristics of the LCD has posed a few hindrances for the study of proteins with disordered domain.[96] The most significant setback is the lack of structural knowledge for A1 due to no crystal or NMR structures available for the full protein, or even the LCD alone.[83] An alternative to *in vitro* structural modeling could be *in silico* modeling. The rise in computational modeling is gaining increasing acceptance to accomplish goals that *in vitro* studies would find challenging or time-intensive.[96] Artificial intelligence (AI) further enhances the capability of *in silico* research and as such, the Alphafold tool was introduced to make protein structures available to researchers.[101,102,103] Although AI-generated models may not be as popular as crystal or NMR structures available in the Protein Data Band (PDB), some situations deem it necessary for use when the classical aforementioned techniques can not be employed.

The research question for this project is an example where the Alphafold model for hnRNPA1 was the only full structure available for use. The alternative to Alphafold was to model the full protein from scratch (*de novo*), which is a complex area of research, requiring expertise and intensive training.[96] Our team had earlier modeled isoform A of A1 which is shorter but also less abundant in neurons. [104]9 This resulted in a need for an isoform B model, for which Alphafold was utilized. Time limitations for this project rendered *de novo* modeling for the full structure of isoform B too

ambitious to be attainable. Nevertheless, both models have been investigated to study their unique properties that may render isoform B more prone to aggregation in neurons. Figure 16 illustrates all the systems relevant to this chapter.

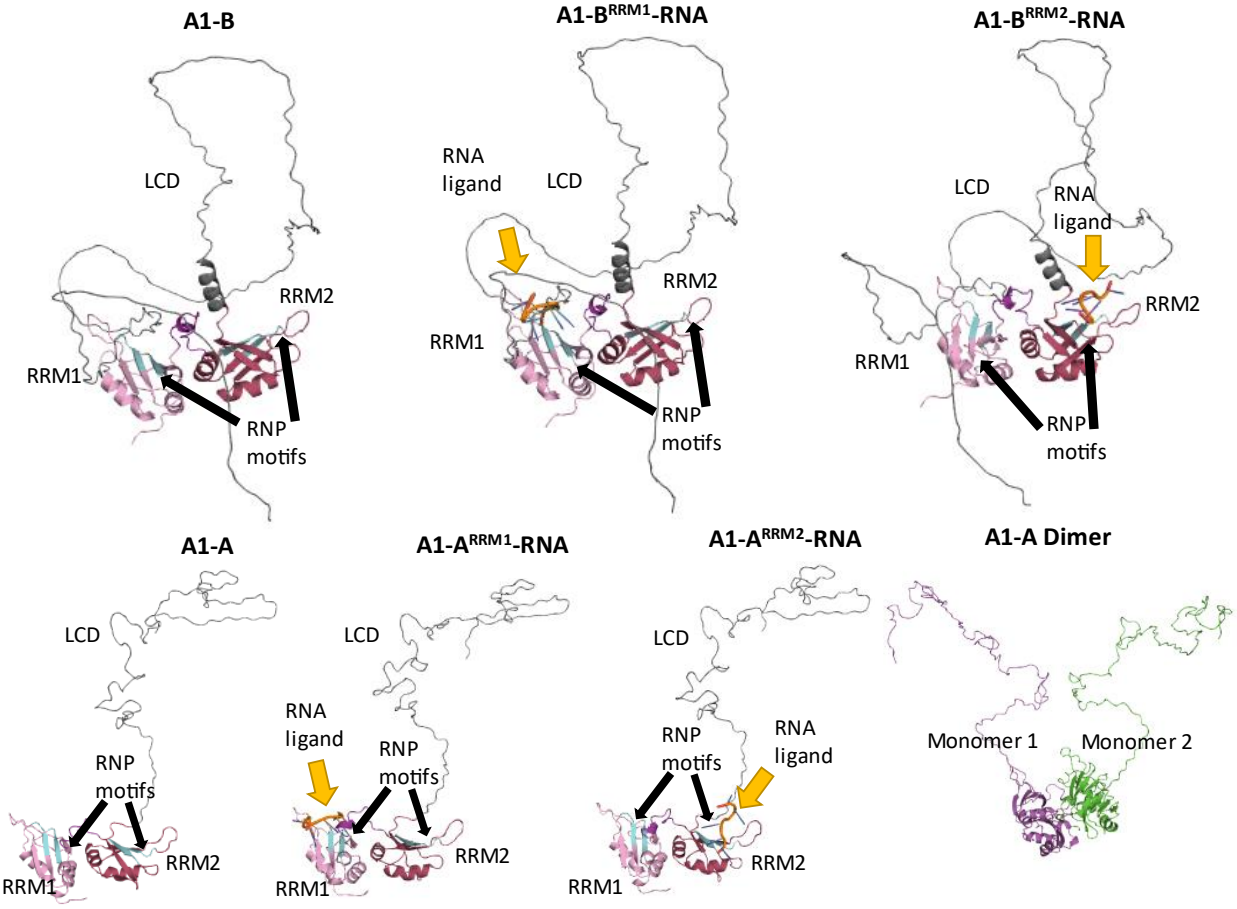


Figure 4.1: The structure and composition of the systems studied to analyze the full A1 protein. Isoform B (upper panel) was obtained from Alphafold. Two different types of RNA, a 7-nt one, was extracted onto RRM1, and a 6-nt one, was extracted onto RRM2. Isoform A (lower panel) was constructed by our team using I-TASSER [98,99,100] .Along with the free system, RNA molecules were extracted onto each RRM, similar to isoform B. This resulted in 2 free systems and 4 RNA-bound systems in total for the monomers. The last unique system was a dimer constructed by aligning two monomers of Isoform A together.

4.2 Methodology

IV.2.1 Obtaining monomer PDB structures for the full model and construction of the dimer

Isoform A was modleled using the I-TASSER server.[98,99,100] The amino acid sequence of the 320 AA monomer obtained fron Uniprot was used as the input and the top predicted sructure was downloaded as a PDB file.

Isoform B was obtained from the Alphafold server.[102,103]

The RNA ligand from the 5MPG NMR structure was extracted onto the RRM1 of both isoforms. Similarly, the RNA ligand from the 5MPL NMR structure was extracted onto RRM2 of both isoforms. Both RNA sequences differ in length and sequence, although both have the signature AG motif facing the RNPs. The use of different RNA sequences was necessary to avoid the process of docking and use simple extraction of native RNA ligands found in the PDB for hnRNPA1.

See Figure 4.1 for an illustration of all systems.

The 6DCL crystal structure was used to construct the dimer system from 2 monomers of A1-A. The 6DCL model only consists of 2 monomers with their RRMs (no LCD) and two ssRNAs bound. Two monomers of A1-A were extracted onto the 6DCL structure to get them aligned, following which the original 6DCL was removed. The ssRNA was removed as well for the purposes of this study as it was too long to allow the full A1-A LCD to align without clashing with the ssRNA.

4.2.2 Molecular Dynamics and binding-free energy calculations

The RNA-A1 complexes (no mutations, in the WT 4YOE control) were relaxed under physiological conditions using 1000 ns long MD simulation each performed with the AMBER 20 package[125]10 and pmemd.cuda engine[139]. A combination of FF14SB[161] and the recent RNA Amber force field developed by a Rochester team (i.e., RNA-ROC) was used for describing the structural parameters of protein and RNA, respectively, for MD simulation. Each complex was solvated in a cubic box of explicit TIP3P water molecules with a distance of 7 Å between the solute and the

edge of the box. The solvated systems were charge neutralized. All system preparation was performed using the tleap program available within the AMBER package[125].

The prepared complexes were initially energy-minimized in 6 stages with each stage involving 1000 steps of steepest descent minimization and 10000 steps of conjugate gradient minimization with a pre-defined harmonic restraint. In the initial stage, a 100 kcal/mol Å−2 restraint was applied on the solute atoms which was gradually decreased to 70 > 50>40>30>0 kcal/mol Å−2 in the subsequent rounds of minimization. The energy minimized systems were gradually heated to 310 K (with a 15 kcal/mol Å−2 on the solute atoms) over a duration of 100 ps and, subsequently, subjected to 5 x 0.4 ns equilibration cycles that were performed under isothermal-isobaric (NPT) conditions with periodic boundary conditions. Again, the equilibration was performed with an implied restraints on the solute atoms that gradually reduced as 15>10>5>3>2 kcal/mol Å−2 in each phase. The equilibrated complexes underwent a 10 ns long MD simulation with a low restraint of 0.5 kcal/mol Å−2 on the solute atoms followed by another 10 ns of 0.01 kcal/mol Å−2 on just the RNA ligand. Finally, an unrestrained production MD simulation of the three systems were carried out for 1000 ns time scale to probe their molecular interactions. The stability of the protein and RNAOs during the course of simulation was assessed by computing the evolution of root mean square deviation (RMSD), after which MD was stopped after reaching a stable trajectory. MD trajectory analyses were performed using the CPPTRAJ module[162] in Amber and the VMD[141], UCSF Chimera[120] and PyMol programs[163].

Following the MD simulation, the last 500 ns of the 1000-ns long MD trajectories of the RNA-A1 complexes were used to compute their (relative) binding free energies that were computed using the MM-GBSA method with the implicit solvent model of Onufriev and Case (igb=2)[157].The snapshots were sampled at a constant interval of 100 ps from the last 10 ns of the MD trajectory

for these calculations. The pairwise decomposition analyses (idecomp=2) were performed to identify the key nucleotides and amino acids that contribute to the binding free energies of the complexes. All computations were performed using MMPBSA.py.MPI script[143] included in the AmberTools 20[125].

Analysis was done in Cpptraj[162].Total electrostatic energy was calculated lie. Native contacts were calculated by specifying protein chains forming contacts within 3.5 Å. RMSD and RMSF calculations were performed using cpptraj[162]. All plots were made in Microsoft excel.

4.2.3 Principal Component and Cross-Correlation Analyses

Principal component analysis (PCA) is a statistical method used to reduce the dimensionality of a complex system, while extracting the variations in the datasets.[135] Using PCA, the dataset is reduced to a few components that represent sample variations instead of visualizing thousands of variables.[135] With respect to MD data, PCA helps describe the variance in the dynamics and conformation of the systems.[136] The variance of the atomic positional fluctuations captured in each dimension are characterized by their corresponding eigenvalue.[136] Most of the cases have 3-5 dimensions that capture over 70% of the total variance within a given MD trajectory.[137] For MD, PCA is sufficiently calculated via measurements of dihedral angles or atomic coordinates for α-carbon atoms.[136] The x, y and z Cartesian coordinates of the C-α atoms were then used to map a cross-correlation visualization for the trajectory.[136] All PCA and cross-correlation plots were conducted using the Bio3d[137] package in R.

4.3 Results

4.3.1 The full unbound A1-B model frequently displays blocking of critical RNP residues

The A1-B isoform had an initial conformation that distributed the LCD such that it was equidistant from the RRMs (Figure 4.1, top panel). Subjecting the model to 1000 ns of simulation time in

triplicate displayed structural agreement and stability (Figure 4.2). The LCD domain had a preference of interdomain contacts with the RRMs which resulted in clustering of RRM-LCD contacts. RMSD analysis indicated high fluctuations in the beginning of the simulation followed by rapid steadiness (Figure 4.2 A). The LCD coiled into a globular structure compared to the extended starting conformation, as RMSD stabilized, which is indicative of more preference and stability for a globular, clustered conformation (Figure 4.2 B). However, the overall fluctuations for the entire LCD domain from residues 220 to 372 indicated higher fluctuations than the RRMs. This was also true for the N-terminal disordered region from residues 1 to 8 (Figure 4.2 C). The three trials were congruent with the trend in RMSF fluctuations.

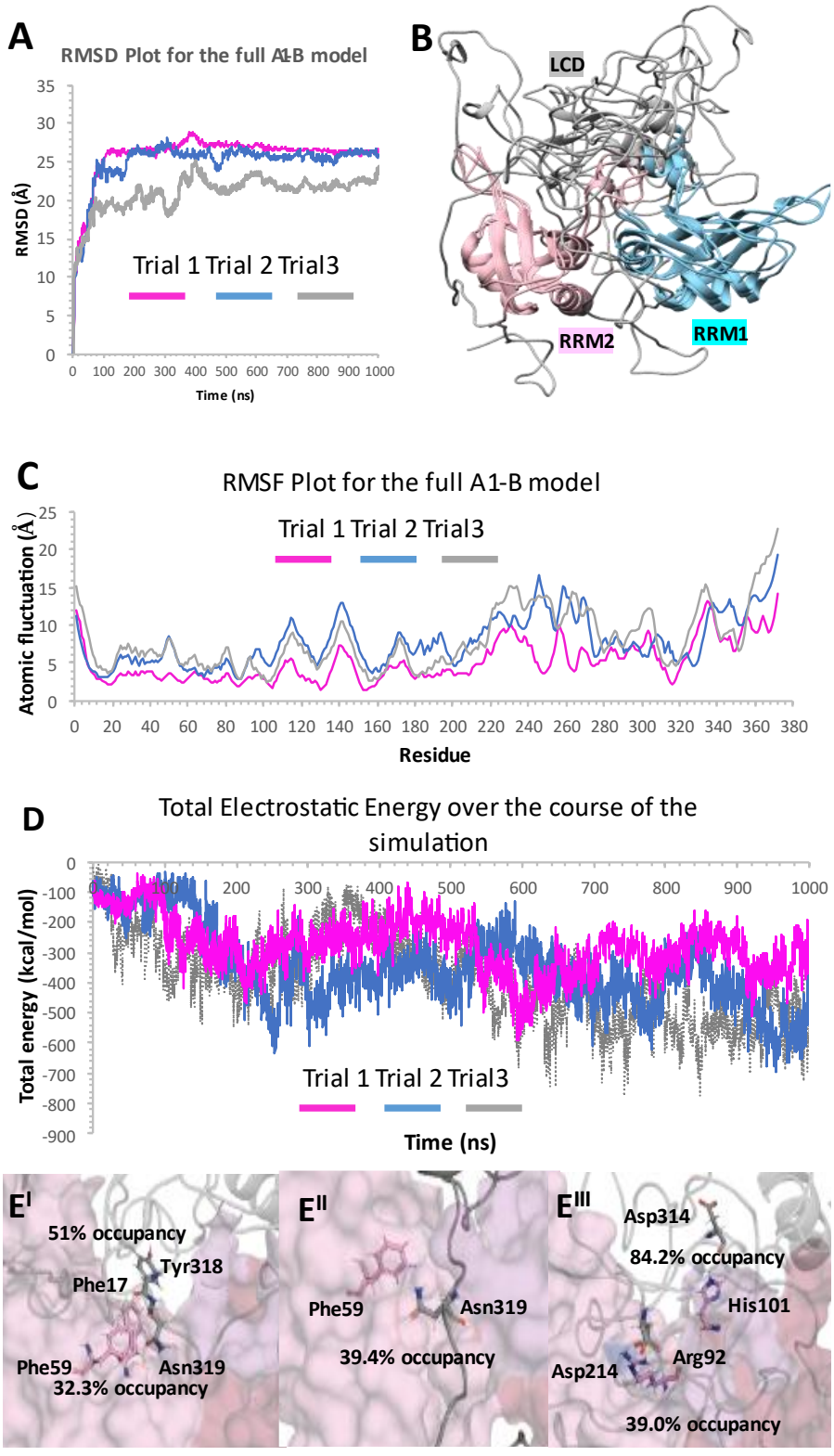


Figure 4.2: Structural dynamics and Interdomain contacts of the A1-B model taken from Alphafold. (A) Stability in the RMSD values are achieved for all three trials after ~400 ns. (B) Alignment of the three trials for the A1-B simulation demonstrates perfect alignment of the folded RRM domains, while the LCD domains for each trial occupies various conformations. (C) Unsurprisingly, the LCD is the only region with high RMSF fluctuations due to its disordered state, along with a few N-terminal residues that do not have a secondary structure. (D) Total electrostatic energy for the three trials demonstrate a gradual increase after 400 ns, which is due to the LCD adopting stable conformations that allow maximum interdomain contacts. (E) The Interdomain contacts resulting in blocking of critical residues involved in RNA binding. Calculations were performed by scanning for native contacts from 500 ns to 100 ns of the simulation. (E^{I}) In Trial 1, Phe17 and Phe59 are both forming contacts with the LCD domain. (E^{II}) However, in Trial 2, of the major RNP residues, only Phe59 is forming contacts with the LCD. (E^{III})Whereas in Trial3, His101 and Arg92 of the interdomain linker are in contact with the LCD.

The overall increase in electrostatic energy between the RRMs and the LCD may be attributed to increased interdomain contact that occurred for all three trials, gradually increasing and fluctuating

throughout the simulation (Figure 4.2 D). While the electrostatic energy does not specify which residues contribute to the binding of the LCD to the RRMs, an analysis of native contacts provides detailed insights into the potential pathophysiology of the LCD-RRM interactions (Figure 4.2 E). Prominent contacts directly involving RNA-binding residues were noted for each trial, which included Phe17 (Trial 1), Phe59 (Trial 1 and 2), His101 (Trial 3) and Arg92 (Trial 3). The aforementioned residues were proven earlier to be critical for RNA-A1 complex stability. Hence, the LCD forming interactions with these residues may prevent RNA from entering the RNP binding pocket of RRM1.

Another way the dynamics were captured for the A1-B systems was using PCA. The eigenvalues give by PCA from 500 ns to 1000 ns elaborated on the diversity of conformations adopted by the protein in all three trials (Figure 4.3). In fact, it takes about 7 eigenvalues to reach 78.3% in trial 1, 76.5% in trial 2 and 72.3% for trial 3. While neither PC1, PC2 or PC3 in any of the trials sufficiently capture the dominant state of the protein based on eigen values, the protein backbone RMSD illustrations provide more insight on where conformational changes occur. Figure 4.3 A demonstrates that Trial 1 has significant variations occurring in the dynamics of the LCD in all three PCs. This is expected because a large time frame 500 ns to 1000 ns was chosen for PC analysis and while all three systems had stable RMSD values (Figure 4.2 A), that stability does not translate into the LCD conformations. In fact, most of the backbone stability comes form the RRMs, which is expected due to their stable folded structures.

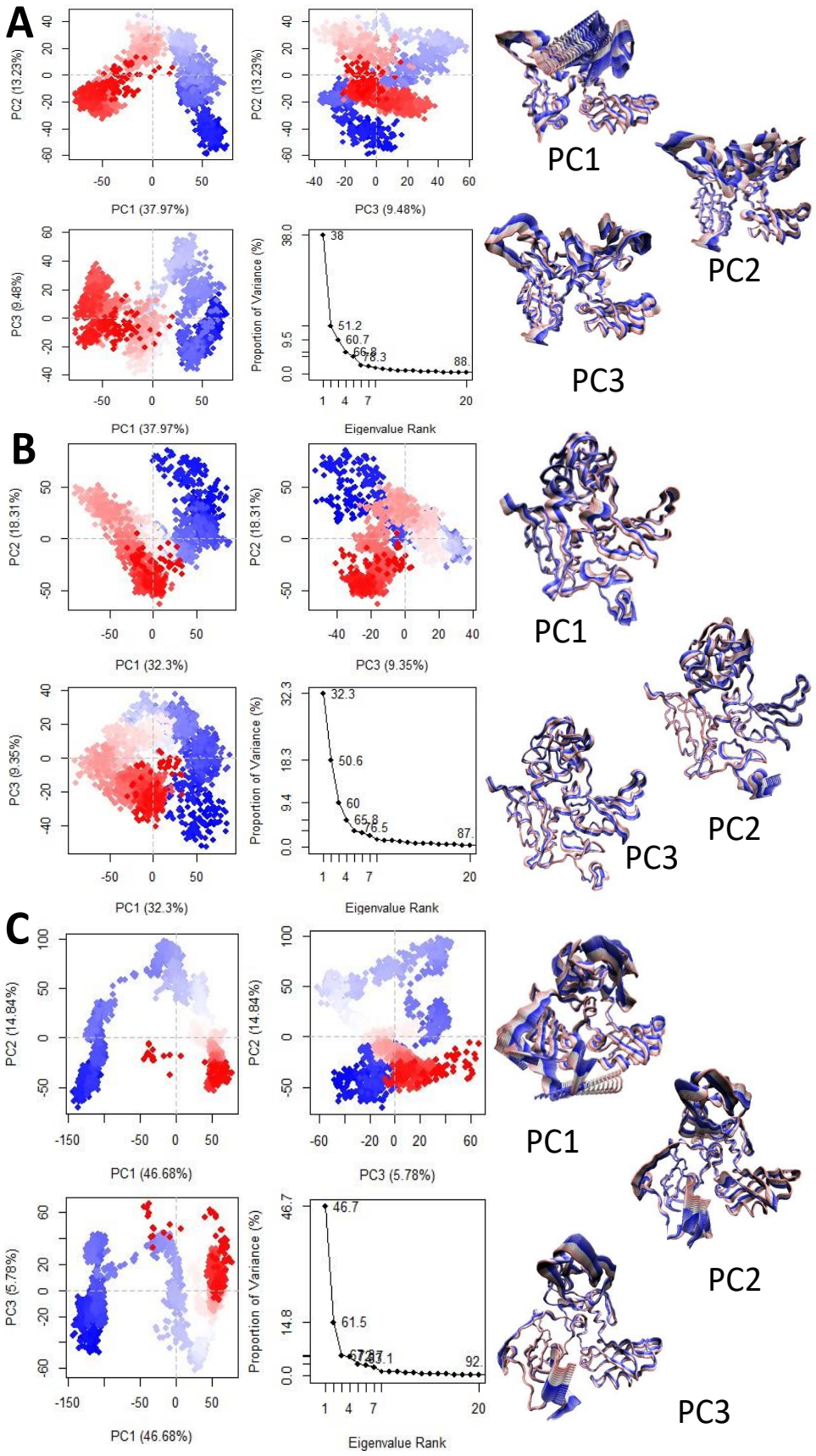


Figure 4.3: PCA analysis and the fluctuations in the backbones for trials 1, 2 and 3 of the A1-B free system. The fluctuations in backbones are only illustrated for PC1, PC2 and PC3. (A) Trial 1 eigenvalue 3 captures 60.7% of the dominant conformations and reaching 78.3% at eigenvalue 7. Most of the conformational changes occur in the LCD as illustrated by PC1, PC2 and PC3. (B) Trial 2 eigenvalue 3 captures 60% of the dominant conformations and reaching 76.5% at eigenvalue 7. Conformational changes predominantly occur in the core LCD as well as the C-terminal end of the LCD. (C) Trial 3 eigenvalue 3 captures 67.2% of the dominant conformations and reaching 72.3% at eigenvalue 7. Most of the conformational changes occur in the LCD above the RRMs and some in the C-terminal end.

Recent evidence has indicated that the mislocalization of A1 is observed in Multiple Sclerosis from the nucleus to the cytoplasm.[63,158] While the nucleus has a high concentration of RNA, the cytoplasm does not.[164] Henceforth, the lack of RNA concentration in the cytoplasm has been proposed as a mechanism of aggregation of RNA-binding proteins in the cytoplasm.[165]

While A1 is designed to optimize RNA adherence, the LCD blocking the binding pocket due to interdomain contacts may prevent RNA from being able to bind. This may explain the cascade of events that allow aggregation of A1 to begin in the cytoplasm, wherein low concentrations of RNA leave the RRMs with room for LCD contact, simultaneously preventing any new RNA molecules from binding. To address how RNA-bound proteins would form interdomain contacts, two systems of A1-B with RNA bound to either RRM1 or RRM2 were simulated.

4.3.2 RNA bound to either RRM results in LCD interacting with the free RRM

The RNA-RRM bound systems were run for single trials, nevertheless delivering insightful data. Both systems, as with the free A1 model, stabilized within the first 100 ns of the simulation (Figure 4.4 A) (Figure 4.5 A). For the A1-B^{RRM1}-RNA system, the LCD had a high preference for forming interactions with RRM2 with less surface-area contact with RRM1 (Figure 4.4B). This allowed the RNA to maintain stable binding affinity for the RRM1 binding pocket, as confirmed with the RMSF data (Figure 4.4C). RMSF values for the terminal ends of the RNA ligand have high fluctuations due to less contact with the protein, but the central residues remained anchored. The RMSF values for the protein, however, followed the general trend as with the RNA-free systems where the LCD and N-terminal end of the protein have highly unsteady fluctuations compared to the RRMs. Similarly, this trend is replicated for the A1-B^{RRM2}-RNA system.

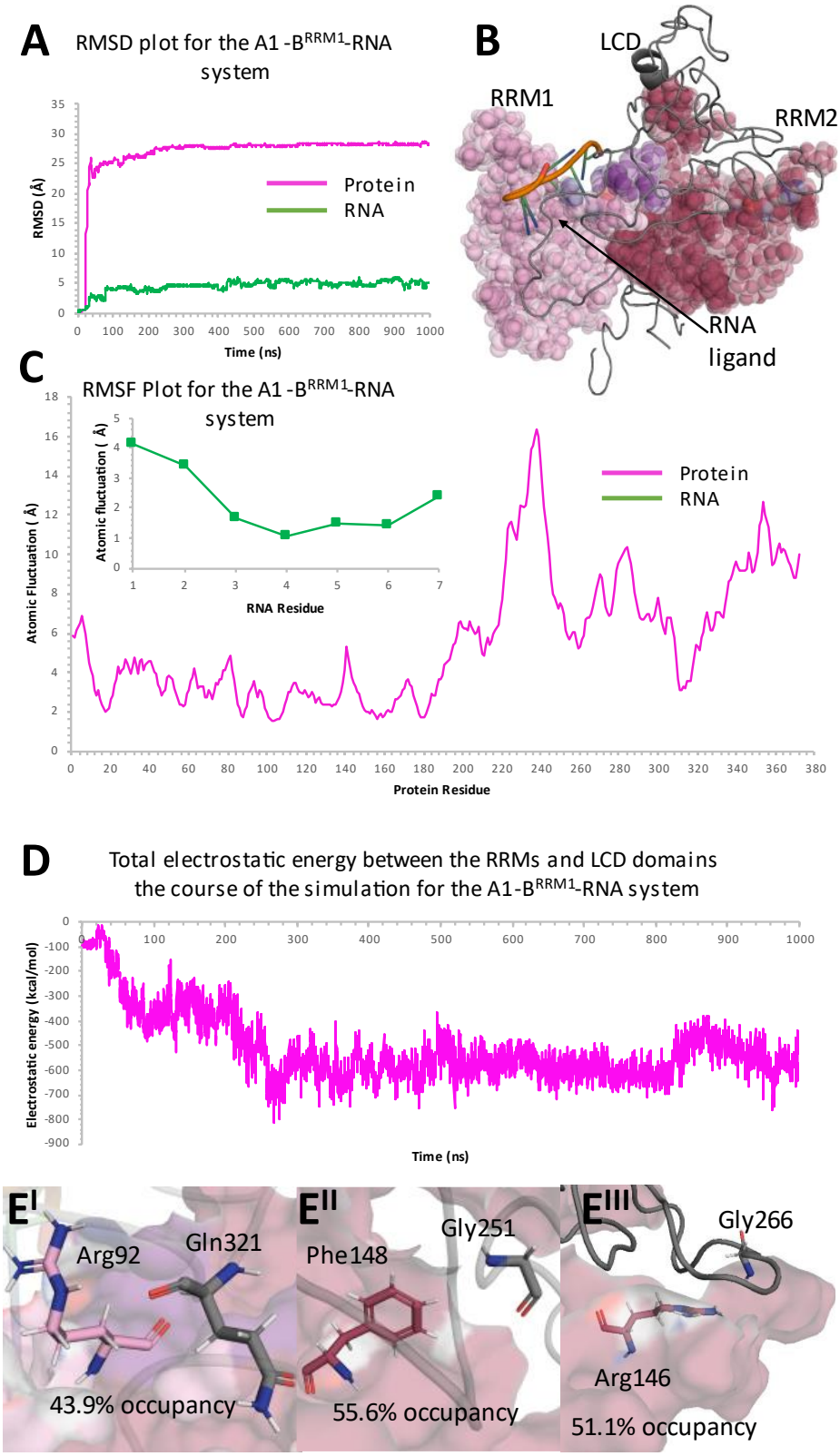


Figure 4.4: Structural dynamics and Interdomain contacts of the A1-B model with an RNA ligand bound to RRM1. (A) RMSD fluctuations for the protein and RNA demonstrate significantly higher values for the protein. This may be due to the protein having a significantly larger proportion of disordered regions. However, both components stabilized early on in the simulation as seen with the plateau. (B) The overall structure of the complex at 1000 ns demonstrating significant Interdomain contacts. (C) Unsurprisingly, the LCD is the only region with high RMSF fluctuations due to its disordered state, along with a few N-terminal residues that do not have a secondary structure. The RNA residues with high fluctuations were also the terminal residues with less surface contact with the RRMs. (D) Total electrostatic contact energy calculated between the RRMs and all LCD residues gives a general increase in values, confirming the interdomain bonds. (E) Interdomain contacts resulted in blocking of critical residues involved in RNA binding. Calculations were performed by scanning for native contacts from 500 ns to 1000 ns of the simulation. (E^{I}) Arg92 made extensive contacts with Gln 321 while RNPs of the RRM2-Phe148 and R146 were in contact with Gly251 and Gly266, respectively (E^{II} ,E^{III}).

The total electrostatic energy for interdomain contacts in both systems demonstrated consistent and overlapping increases from 0 kcal/mol at 0 ns to ~ -500kcal/mol at 1000 ns (Figure 4.4D) (Figure 4.5 D). The increases in interdomain contacts, however, differ structurally at the atomic level depending on which RRM the RNA ligand is bound to. When RNA is bound to RRM1 (Figure 4.4 Panel E), the LCD interacts with some of the interdomain linker residues such as the backbone of Arg92, such that the guanidine group is still available for RNA-binding (Figure 4.4 EI). More importantly, the LCD interacts with the RNP residues of RRM2, such as Phe148 and Arg146 (Figure 4.4 EII, EIII) which was not observed with the RNA-free systems. Hence, when the RRM1 binding pocket is preoccupied, the LCD forms contacts with RRM2 due to more surface area available in that region. The opposite is true for the RNA bound to RRM2 (Figure 4.5 Panel E). RNPs of RRM1, such as Arg55 and Gly20 form interactions with the LCD when the RNA is bound to RRM2 (Figure 4.5 EI, EII). The essential linker residues with RNA-adherence roles such as His101 are also a mediator of interdomain contacts (Figure 4.4 EII).

The dynamic behavior of both systems was finally analyzed with PCA (Figure 4.6). When RNA is bound to RRM1, the first three eigenvalues represent 52% of the conformations. As expected, most of the fluctuations in the protein backbone are a direct result of the disordered LCD. PC1, PC2 an PC3 illustrate unique dominant conformations for the protein while there is some overlap between PC2 and PC3. Interestingly, even though RNA was bound to RRM1, there are no significant fluctuations observed for the RRM1 or the interdomain linker which was in contact with RNA. This indeed confirms the stability of RNA in the RRM1 binding pocket whereas the conformational changes in the system can mostly be attributed to the LCD, whether in its clustered region or the C-terminal end. It takes eigenvalue 7 to capture 69.4% of the dominant

structures for this system, which is similar to the values seen for the unbound A1-B protein (Figure 4.3).

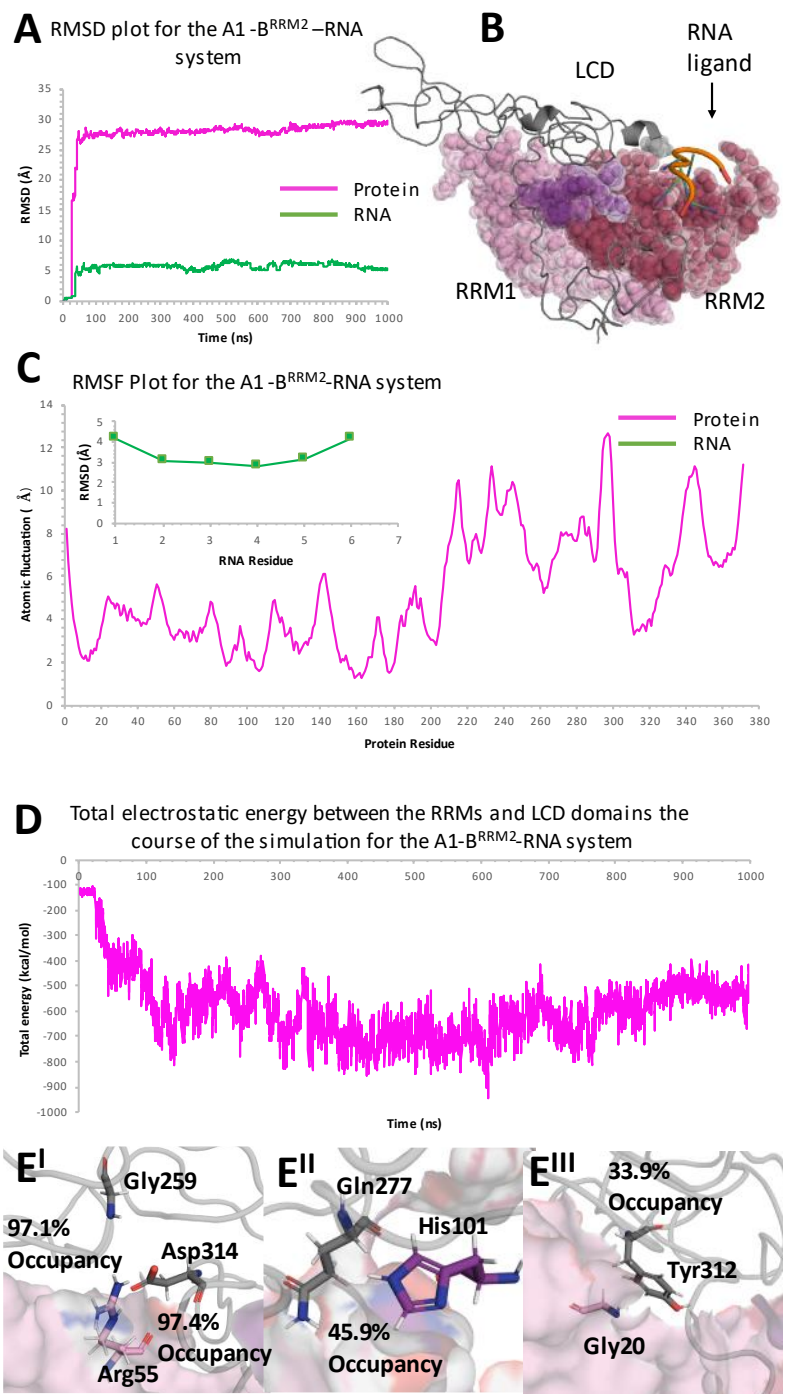


Figure 4.5: Structural dynamics and Interdomain contacts of the A1-B model with an RNA ligand bound to RRM2. (A) RMSD fluctuations for the protein and RNA demonstrate significantly higher values for the protein. This may be due to the protein having a significantly larger proportion of disordered regions. However, both components stabilized early on in the simulation as seen with the plateau. (B) The overall structure of the complex at 1000 ns demonstrating significant Interdomain contacts. (C) Unsurprisingly, the LCD is the only region with high RMSF fluctuations due to its disordered state, along with a few N-terminal residues that do not have a secondary structure. The RNA residues with high fluctuations were also the terminal residues with less surface contact with the RRMs. (D) Total electrostatic contact energy calculated between the RRMs and all LCD residues gives a general increase in values, confirming the interdomain bonds. (E) Interdomain contacts resulted in blocking of critical residues involved in RNA binding. Calculations were performed by scanning for native contacts from 500 ns to 1000 ns of the simulation. The shifting of the LCD towards RRM1 resulted in the blocking of RNPs: (E^{I}) Arg55 made extensive contacts with Asp314 and Gly259 while Gly20 made contacts with Tyr312 (E^{II}). (E^{III})Additionally, His101 from the linker loop was in contact with Gln277 from the LCD.

The PCA results for the RNA^{RRM2}-A1 have first three eigenvalues representing 74% of the conformations, with the majority of 61.6% coming from PC1 alone (Figure 4.7). There is not a

significant overlap in conformations between PC1 and PC2 nor PC1 and PC3, but PC2 and PC3 demonstrate some overlap. PC1 displays a significant amount of fluctuations in the protein backbone of disordered LCD while there is also some instability in the RRMs in contrast to the systems seen earlier.

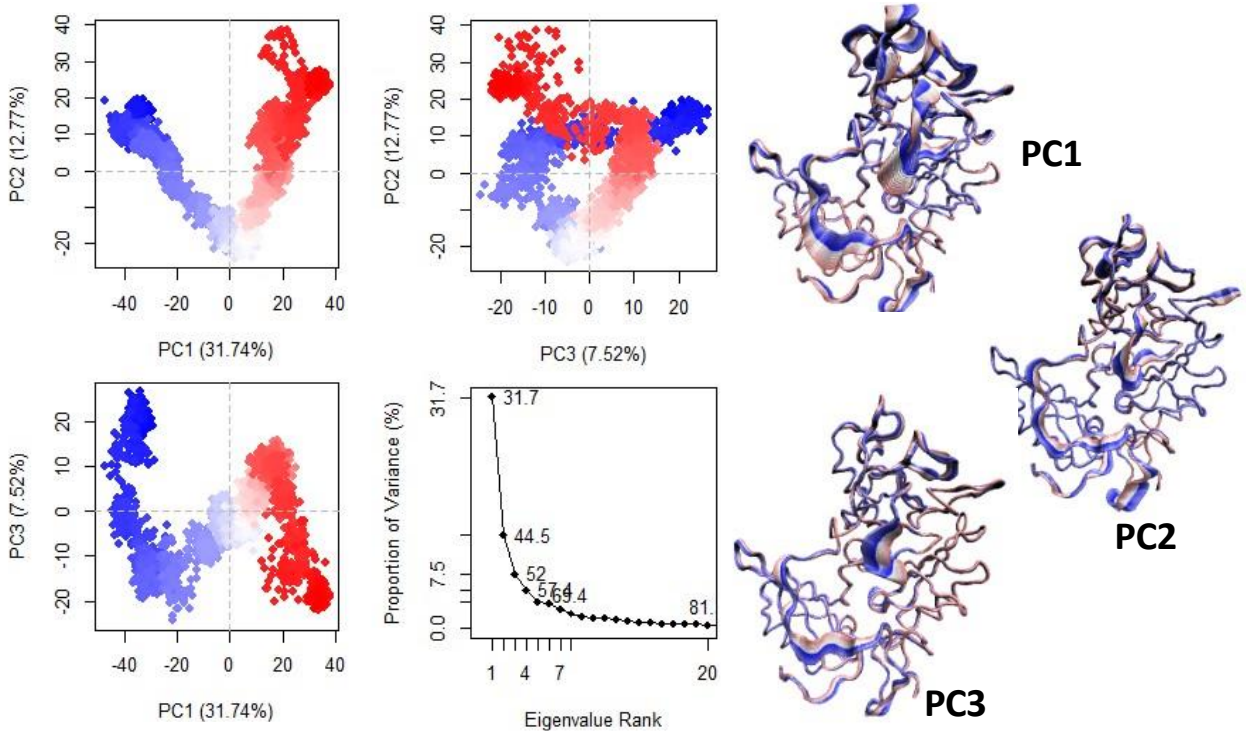


Figure 4.6 PCA analysis and the fluctuations in the backbones for the A1B^{RRM1}-RNA free system. The fluctuations in backbones are only illustrated for PC1, PC2 and PC3. PC1-PC2 and PC1-PC3 do not have a significant overlap. However, PC2 and PC3 demonstrate some level of overlap. The first three eigenvalues together represent 52% of the conformations with the first 7 eigenvalues representing 69.4%.

To date, exact residues contributing towards aggregation of A1 have not been evaluated. While this work does not delve into the complex revelation of residues mediating contacts between multiple macromolecules, it has revealed potentially useful insights for a single monomer. The preference of the LCD to bind to RNP residues where the RNA needs to reside, whether in RRM1 or RRM2, may have disease-specific implications. While the RRM1 is required for binding an

array of RNA molecules in the nucleus, the RRM2 is involved in cooperative binding of RNA and also for telomere maintenance.[31] The LCD has the RGG domain and the M9 sequence with crucial roles, but random residues of the LCD for which roles are not known tend to form the interdomain contacts to the RRMs. Residues 218-240 constitute the RGG-box used for RNA binding and telomeric maintenance, while residues 320-357 make up the NLS to interact with nucleoporin complexes. In the simulations evaluated thus far, most LCD residues blocking RNP residues do not have known roles for the protein to date. However, it does confirm the lack of capability the RGG box and the M9 sequence have to bind ssRNA, as mentioned in literature.[156]

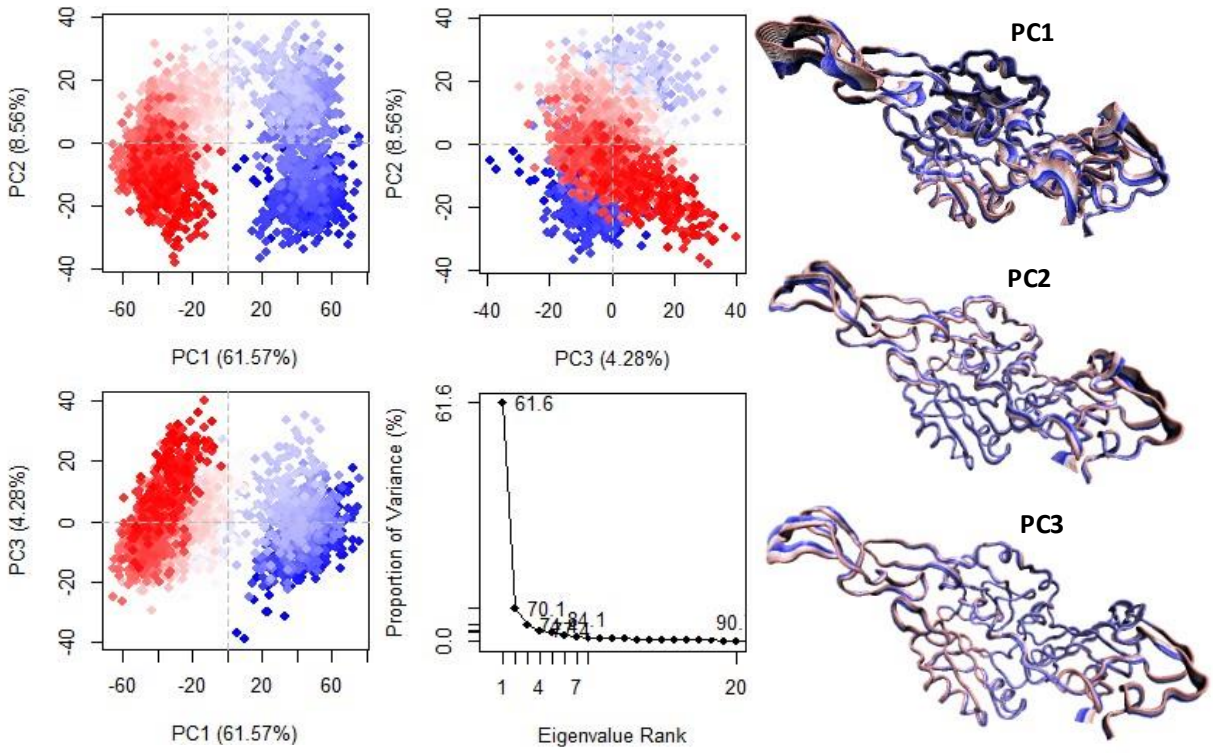


Figure 4.7 PCA analysis and the fluctuations in the backbones for the A1B^{RRM1}-RNA free system. The fluctuations in backbones are only illustrated for PC1, PC2 and PC3. PC1-PC2 and PC1-PC3 do not have a significant overlap. However, PC2 and PC3 demonstrate some level of overlap. The first three eigenvalues together represent 74% of the conformations.

4.3.3 The LCD for Isoform A has an overall preference for RRM2

The results for the A1-A systems differ significantly than those for isoform B. Both isoforms share the exact sequence, including the RRMs, the RGG box and the M9 localization sequence. However, isoform B has an insertion from residues 252 to 307 which does not add to the RGG box or the M9 sequence, but is a disordered region with an unknown role. Therefore, it can be suggested that the structural differences in the isoforms as a result of differences in length are the direct cause for unique conformations adopted by each of them.

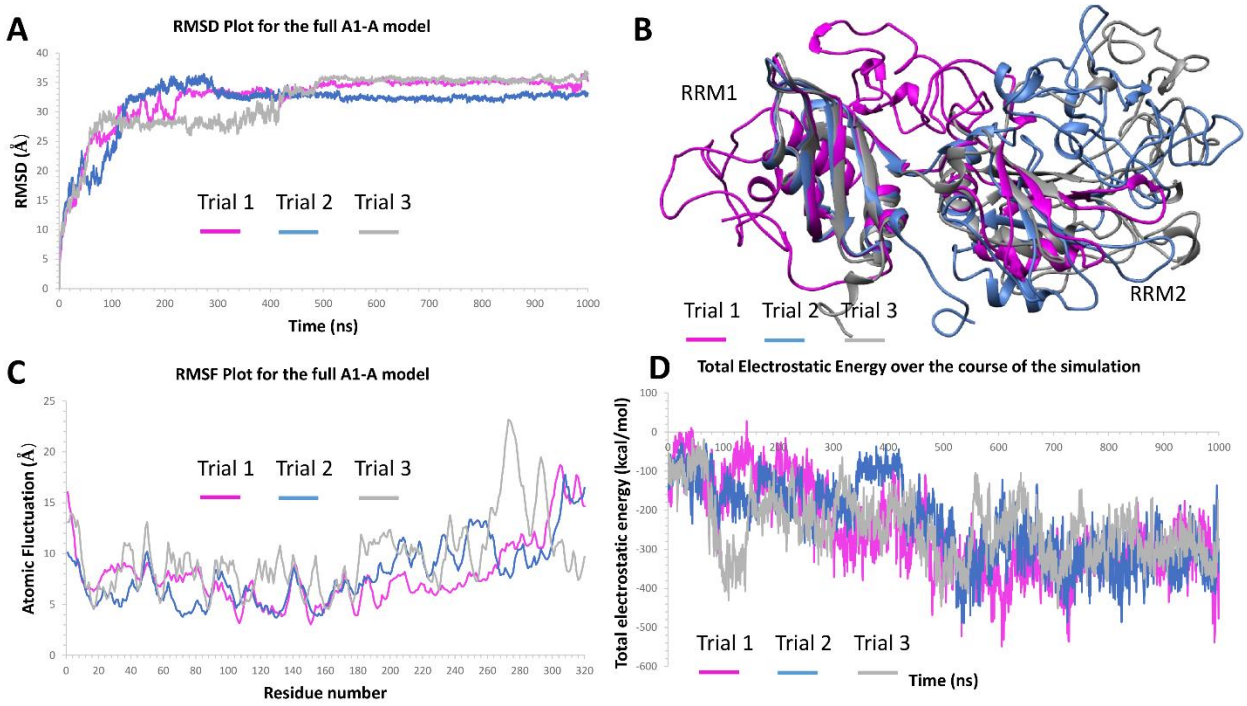


Figure 4.8: Structural insights to the free A1-A monomer system subjected to 1000 ns of simulation. (A) RMSD values remain unchanging after the first hundred nanoseconds of the simulation for all three trials. (B) Evaluation of the structure of the aligned monomers sampled at the end of each trial confirms the LCD-RRM contact preference. The LCD clusters around RRM1 in Trial 1 and around RRM2 for trials 2 and 3. (C) RMSF fluctuations by residue overlap between each trial where the disordered N-terminal residues and the entire LCD is relatively less stable than the folded RRM domains. The exception to this trend is significantly more instability in residues 265-285 in Trial 1. This is due to that specific section of the LCD adopting multiple conformations throughout the simulation. (D) An overall increase in electrostatic energy was expected due to the interdomain interactions, although the fluctuations stabilized after 500 ns.

All three trials indicated a similar trend in stability (Figure 4.8A). Trials 2 and 3 prefer accumulating the LCD near RRM2 while Trial 1 uniquely displays LCD placement near RRM1 (Figure 4.8 B). Nevertheless, placement of the LCD over the RRMs did not impact protein stability as confirmed with RMSF fluctuations (Figure 4.8 C). The RMSF values across all three trials consistently remain stable for the RRM domains and unsteady for the N-terminal end and the entire LCD domain.

Interestingly, the A1-A isoform, unlike the A1-B isoform, does not directly have LCD residues interacting with critical RNA-binding residues, thus leaving the binding pocket unoccupied. Interaction analysis did not confirm any RNP residue interacting with the LCD, therefore are not reported here. However, there were still significant interactions occurring between the LCD and the RRMs (Figure 4.8D). Electrostatic energy overall increased after 500 ns for each trial and became relatively stable at ~700 ns. This increase in electrostatic contacts, along with visual inspection in Figure 4.8A illustrate significant interdomain interactions that allow the LCD to cluster near the RRMs, without impacting the critical RNA-binding pocket.

PCA provided additional valuable insights into the dynamic behavior of each A1-A unbound system (Figure 4.9). The first three eigenvalues in Trial 1 represented 68.2% of the structures, followed by 70.9% in trial 2 and only 57.1% in trial 3. Trial 1 surprisingly demonstrated significant fluctuations in the RRM2 domain in PC1, which comprised 46.4% of the conformations (Figure 4.9 A). In fact, the placement of the LCD over RRM1 resulted in significant fluctuations in backbone RMSD overall for all three eigenvalues in trial 1. This varies from trial 2 (Figure 4.9 B) and trial 3 (Figure 4.9 C) where the backbone RMSD values odo not describe major conformational changes even in the LCD.

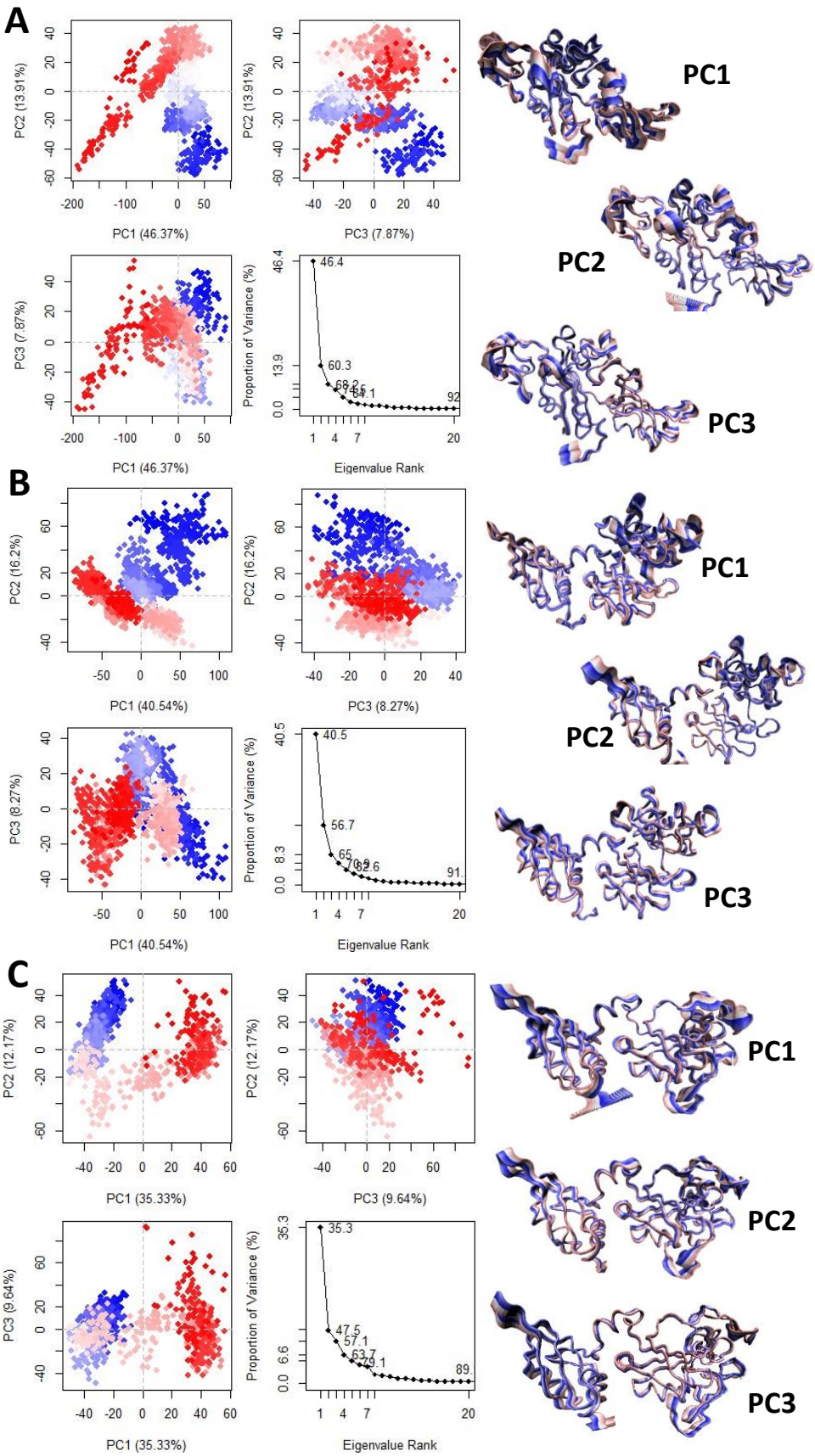


Figure 4.9: PCA and the fluctuations in the backbones for trials 1, 2 and 3 of the A1-B free system. The fluctuations in backbones are only illustrated for PC1, PC2 and PC3. (A) Trial 1 eigenvalue 3 captures 46.4% of the dominant conformations and reaching 84.1% at eigenvalue 7. Most of the conformational changes occur in the LCD and some in the RRM2 as illustrated by PC1, PC2 and PC3. (B) Trial 2 eigenvalue 3 captures 40.5% of the dominant conformations and reaching 82.6% at eigenvalue 7. Conformational changes predominantly occur in the core LCD. (C) Trial 3 eigenvalue 3 captures 35.3% of the dominant conformations and reaching 79.1% at eigenvalue 7. Most of the conformational changes occur in the LCD above the RRMs and some in the C-terminal end.

Each RNA-bound system demonstrated similar trend in interdomain contacts and overall protein dynamics (Figures 4.10, 4.11). RMSD fluctuations for the full protein and RNA bound to either RRM1 or RRM2, achieved a plateau withing the first 100 ns of simulation (Figure 4.10A) (Figure 4.11A), as seen consistently with the full A1-B models (Figure 4.4C) (Figure 4.5C). RMSF analysis demonstrated that the LCD and the N-terminal ends of the protein had higher instability due to their disordered form (Figure 4.10C, 4.11C). Similarly, the central RNA residues that made maximum contact with the RNA binding pocket had lower RMSF values compared to the 5′ and 3′ ends of the RNA (Figure 4.10C, 4.11C).

However, the LCD preferred accumulating near the RRM2 (Figure 4.10B, 4.11B) similar to Trials 2 and 3 of the unbound A1-A models (Figure 4.8A). This accumulation near the RRM2 of the A1-A protein occurs independent of where RNA is bound, providing interesting structural insights to how the differences between length can impact interdomain interactions between isoforms. While A1-B had unbiased accumulation of the LCD across both RRMs and the interdomain linker, the smaller length of A1-A gives its LCD an inclination to fold over RRM2 as it may provide faster stability in the simulation.

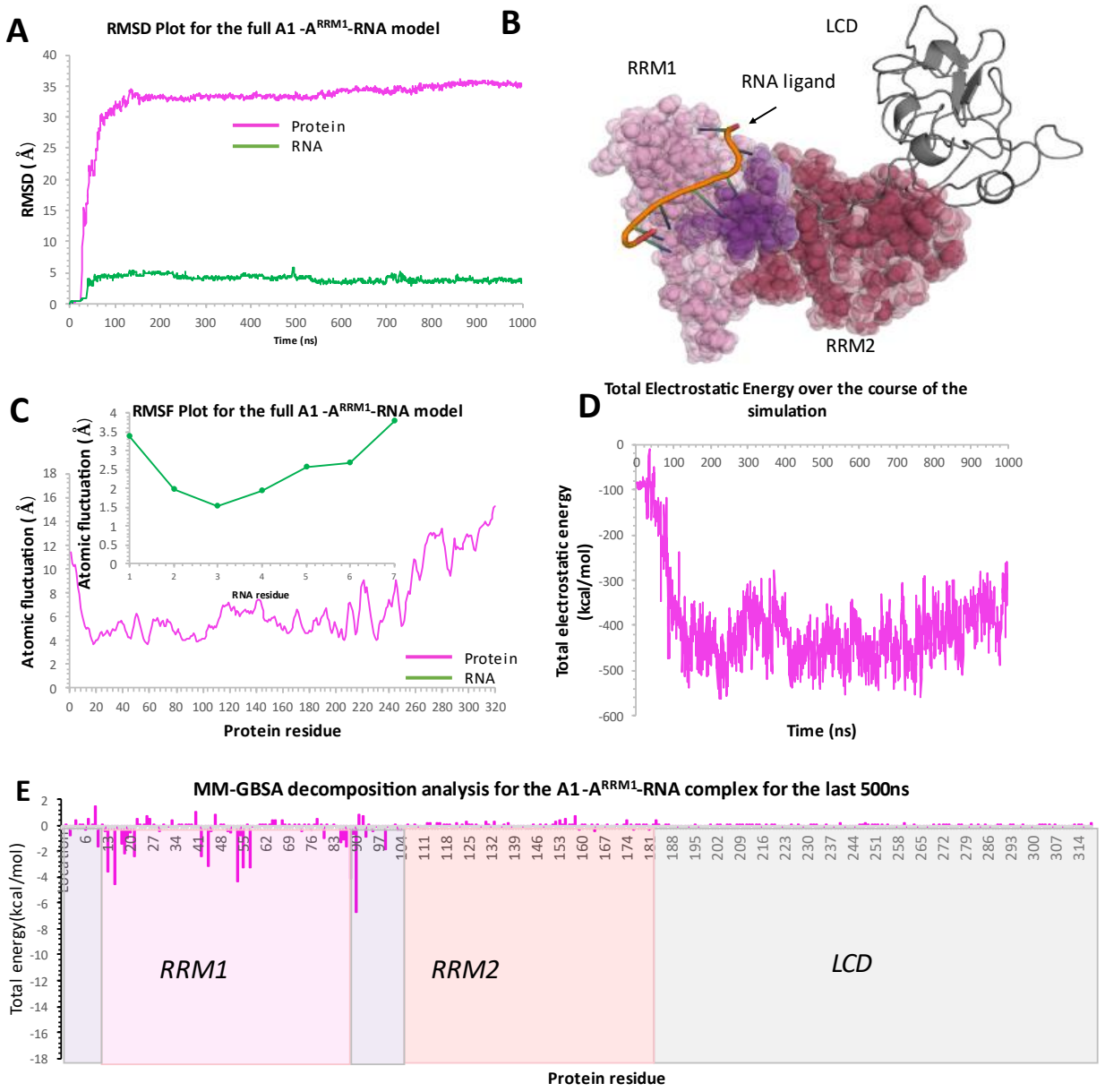


Figure 4.10: Evaluation of the binding of an RNA ligand to RRM1 of the A1-A isoform. (A) The RMSD values for the RNA and protein remained unchanged after the first few nanoseconds of restraints being removed. The RMSD for the protein was significantly higher than that for RNA due to the added changing conformations of the LCD. (B) Structural insights into the A1-ARRM1 –RNA complex at 1000 ns. (C) RMSF fluctuations for both the protein and RNA residues are higher for the terminal ends. This may be due to the terminal ends of the protein being disordered and unstable whereas for RNA, fluctuations are a result of less contact with the protein's binding pocket. (D) An increase in electrostatic energy between the LCD and RRMs is observed after ~200 ns, which is also the time point when the RMSD is stabilized in (A), confirming that the LCD-RRM contacts helped stabilize the protein. (E) MM-GBSA decomposition analysis illustrated the RRM1 residues, mostly the RNPs and the interdomain linker, were responsible for RNA binding.

While electrostatic contacts are ensured when RNA is bound to RRM1 (Figure 4.10C), the binding of RNA to RRM2 distracts the LCD into binding with RNA exclusively such that the contact between the LCD and RRM2 is a net repulsion in electrostatic energy (Figure 4.11D). Thus far, this is the only system where the interdomain contacts between the LCD and the RRMs are unstable and unpreferred. This is confirmed with the MM-GBSA analysis for both systems where RNA bound to RRM1 only has residues from RRM1 and not the LCD contributing to the binding affinity (Figure 4.10D). This is not true for the RNA-RRM2 system: residues from the RRM2 and the LCD both contribute to the binding affinity significantly (Figure 4.11D). As such, the binding affinity differences between the A1-ARRM1-RNA and A1-RRM2-RNA is ~ -70 kcal/mol (Table 4.1).

Table 4.1: Comparison of binding affinity values calculated for the full A1-A protein bound to the 5MPG or 5MPL ligand, from 500 ns to 1000 ns for a single trial. MM-GBSA binding free energy calculations were performed for half of the full simulation as all systems had adequately stabilized early on in the simulation. Overall, RNA bound to RRM1 results in higher binding affinity values.

SYSTEM	BINDING AFFINITY	ST.DEV.
A1-A^{RRM1}-RNA	-85.1689	8.5486
A1-A^{RRM2}-RNA	-155.36	11.4098

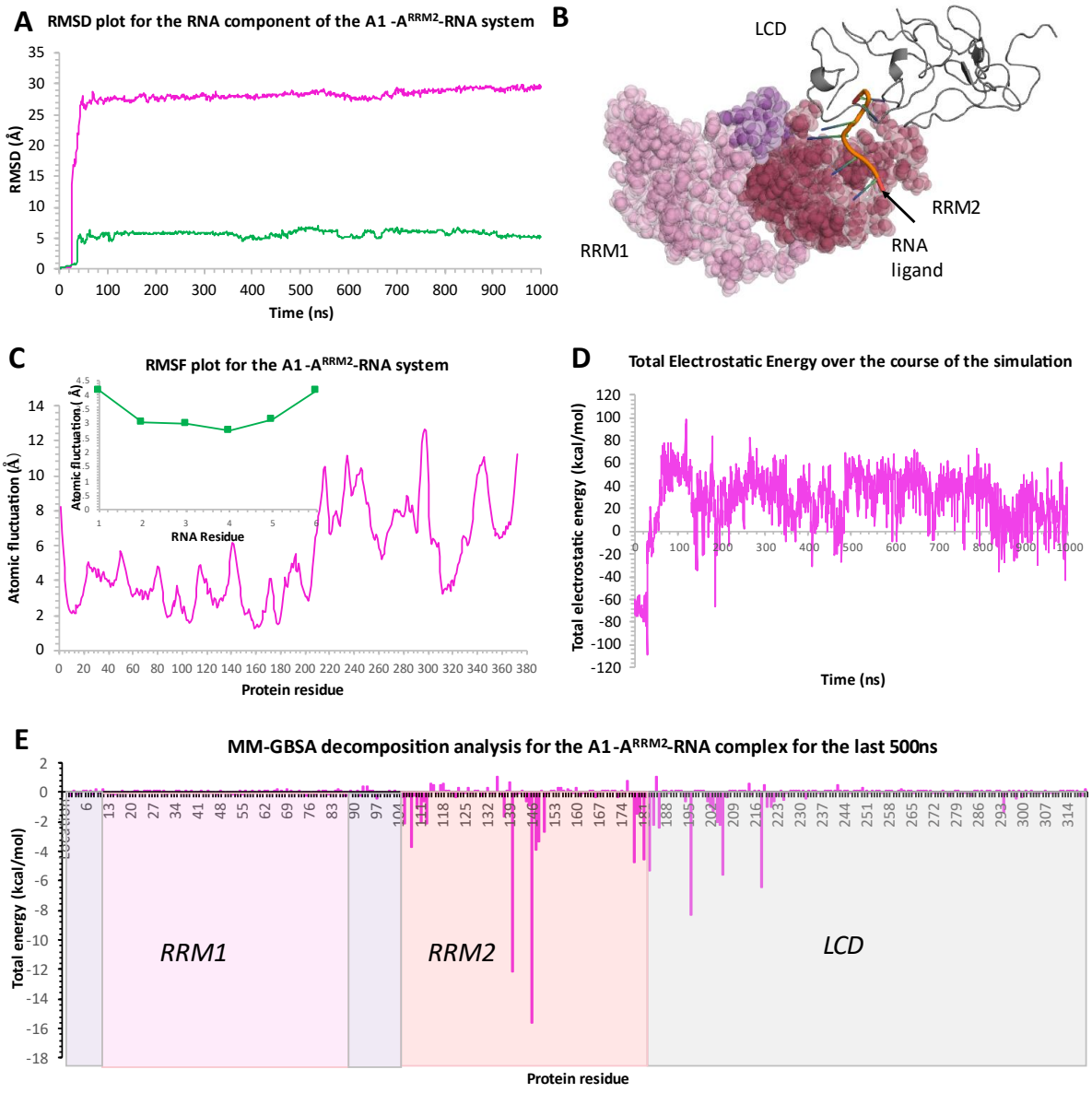


Figure 4.11: Evaluation of the binding of an RNA ligand to RRM2 of the A1-A isoform. (A) The RMSD values for the RNA and protein remained unchanged after the first few nanoseconds of restraints being removed. The RMSD for the protein was significantly higher than that for RNA due to the added changing conformations of the LCD. (B) Structural insights into the A1-ARRM2 –RNA complex at 1000 ns indicates very little interdomain contact. (C) RMSF fluctuations for the protein indicate significant instability for the terminal ends and the entire LCD domain. This may be due to the terminal ends of the protein being disordered, whereas for RNA, fluctuations are a result of less contact with the protein's binding pocket for the 5' and 3' ends. (D) An increase in electrostatic energy followed by significant repulsion between the LCD and RRMs. This may be due to the LCD preferring interactions with the RNA, as proven with the MM-GBSA analysis (E). Both the RRM2 and LCD contribute to RNA binding.

In congruence with the previous A1-A and A1-B systems, the RNA-bound A1-A systems display backbone RMSD stability for the RRMs while the LCD has dynamic changes for the first three eigenvalues (Figures 4.12, 4.13). A1-A^{RRM1}-RNA has the first eigenvalue representing 35.3% of the dominant conformations and 57.1% for the first three eigenvalues (Figure 4.12). It takes 7 eigenvalues to represent 79.1% of the dominant conformations. Nevertheless, the first three eigenvalues overlap significantly with each other and visual inspection of the backbone fluctuations demonstrate differences with the LCD dynamics, which is expected for disordered regions. RNA bound to RRM1 does not significantly impact its fluctuations, confirming that the RNA-binding pocket residues in RRM1 are not constantly adapting conformations and are stable with respect to RNA-binding.

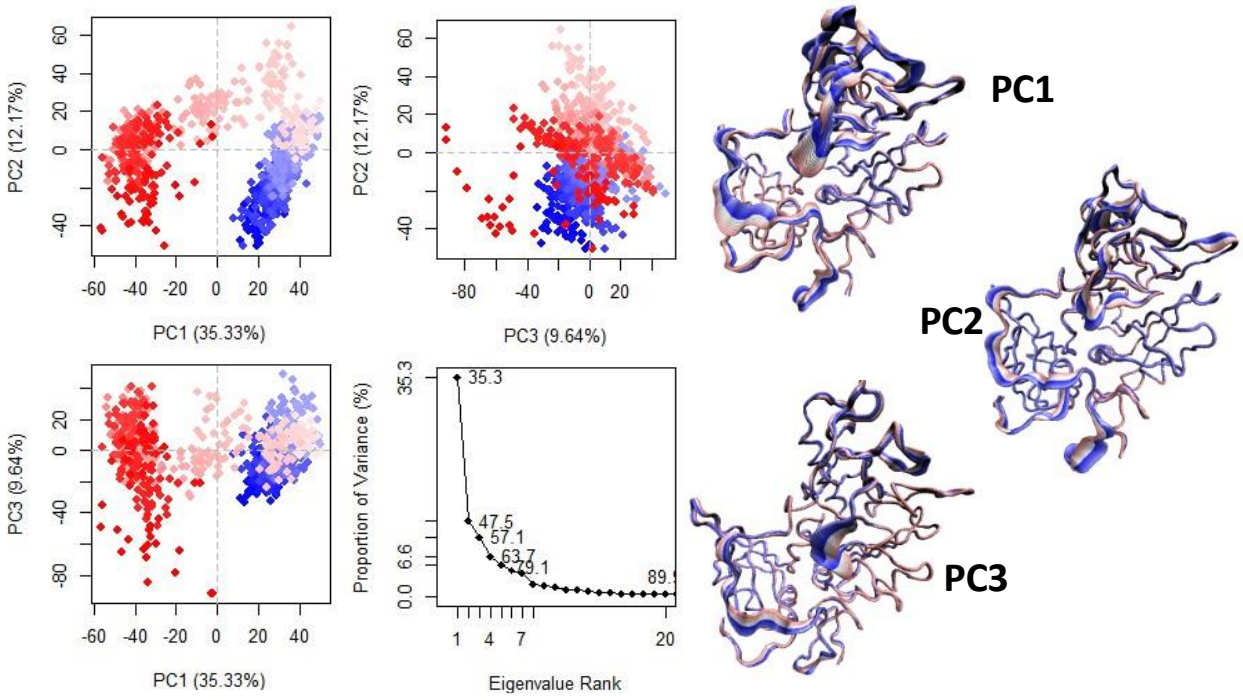


Figure 4.12: **PCA analysis and the fluctuations in the backbones for the A1A^{RRM1}-RNA system.** The fluctuations in backbones are only illustrated for PC1, PC2 and PC3. All principal components have some amount of overlap with each other. The first three eigenvalues together represent 57.3% of the conformations.

Overlap for the first three eigenvalues is more significant for the A1-A^{RRM2}-RNA system (Figure 4.13). The first three eigenvalues represent 64.5% of the conformations for the protein and reaching 82.9% with the seventh eigenvalue. Visual inspection of the backbone RMSD fluctuations surprisingly reveals RRM1 conformational changes along with that of the LCD. This is interesting considering that RNA was bound to RRM2 in this system and not RRM1. This was seen earlier in figure 4.9 where minor fluctuations in the RRM1 of the A1-A unbound system are observed due to RRM1 having no RNA binding nor interdomain contacts, similar to what is observed for Figure 4.13.

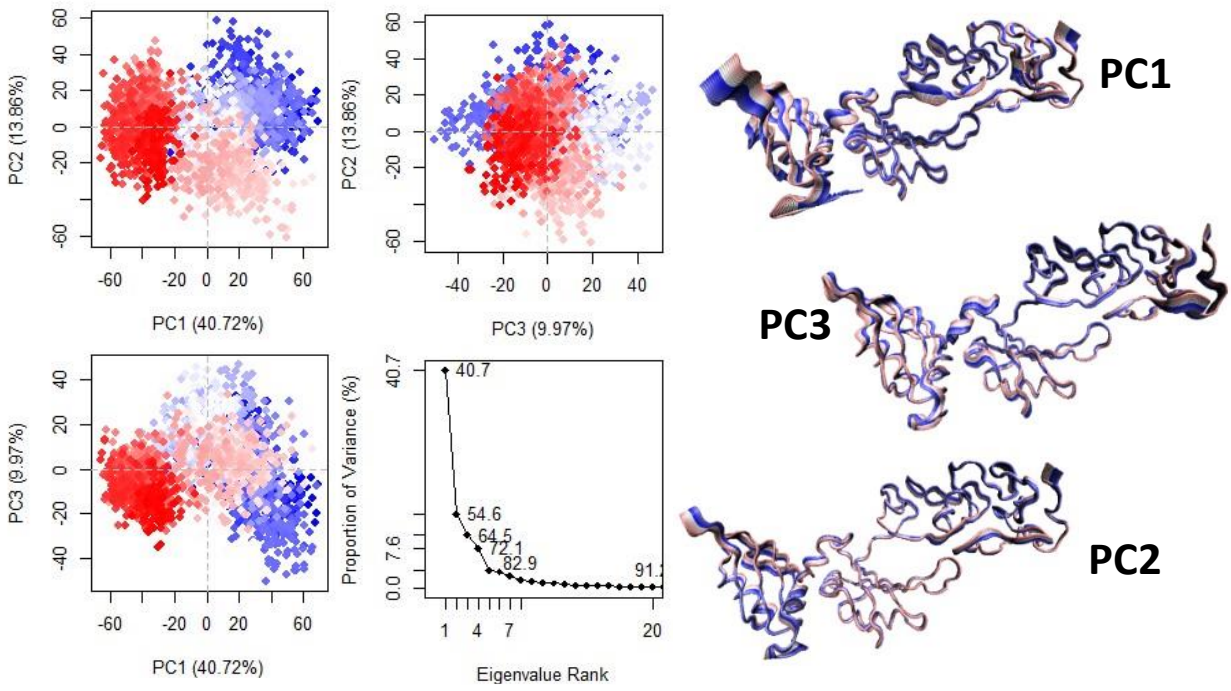


Figure 4.13: **PCA analysis and the fluctuations in the backbones for the A1A^{RRM2}-RNA system.** The fluctuations in backbones are only illustrated for PC1, PC2 and PC3. All principal components have a significant overlap with each other. The first three eigenvalues together represent 64.5% of the conformations.

4.3.4 Preference of LCDs to bind the opposite monomer in a dimer system

The dimer system, for this research, was only constructed for the A1-A isoform. After 1000 ns of simulation time, the LCD domains of each monomer comfortably shifted to the RRMs of the opposite monomer, albeit not blocking the binding pocket (Figure 4.14A). While Monomer 1 displays overall higher RMSD values compared to Monomer 2, the overall trend in RMSF fluctuations are highly similar for both monomers, with slight differences, which are expected in a dynamic system (Figure 4.14 B, C).

The tendency of the A1-A isoform LCD to not block critical RNP residues or even coming into close contact with the RRM binding pocket is a trend seen earlier for the free A1-A system and the RNA-bound systems. The interesting difference however, is that the LCD-RRM interactions no longer exist with the same monomer. The cross interactions of the LCD and RRMs of opposite monomers replaced the general trend seen previously (Figure 4.13 D). Snapshots from the simulation overtime indicate the LCDs of each monomer folding over to reach the opposite monomer (Figure 4.13D). Cumulative electrostatic energy differences confirm this notion with increments only observed over time for cross-monomer interactions but not for domains within the same monomer. However, some residues of the LCD of monomer 1 have a positive cross correlation with the RRMs of monomer 1 and same for monomer 2 (pink box). However, the RRM1 of monomer 1 is also positively correlated with the RRMs of monomer 2 (red box), demonstrating that the RRMs of the two monomers are also capable of interactions.

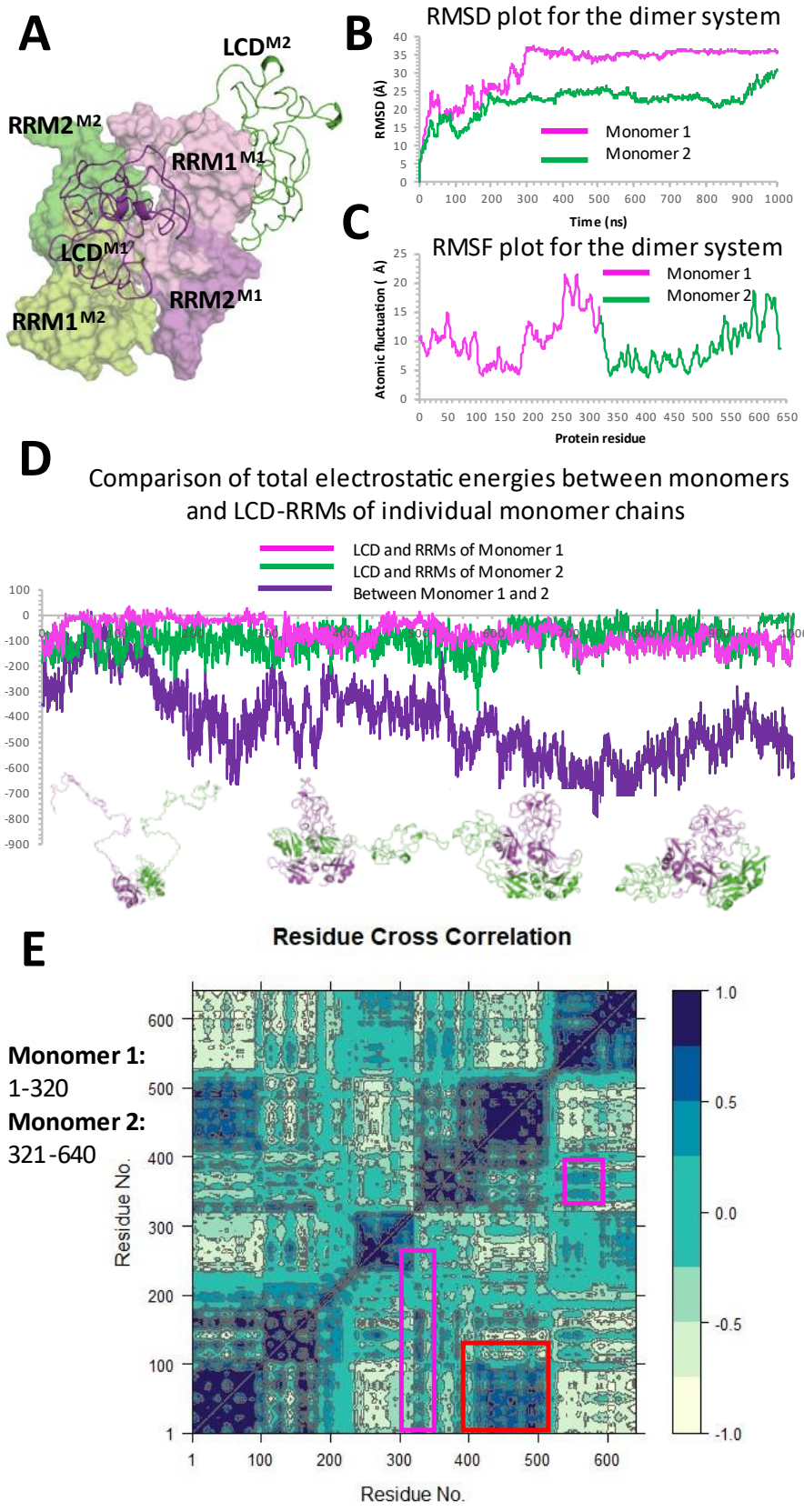


Figure 4.14: Structural dynamics and Interdomain contacts of the A1-A dimer model. (A) The dimer model at 1000 ns displayed a significant amount of overlap of the LCDs with the opposite monomers. The LCD of monomer 1 mainly interacted with the linker loop of monomer 2. The LCD of monomer 2 interacted with RRM1 of monomer 1. However, no significant RNP residues were affected. **(B)** RMSD fluctuations between the monomers display surprising differences in stability. Although both monomers stabilize ~300 ns, there are differences in values of the RMSD. This is due to the fact that the monomer 1 LCD has less surface contact compared to the monomer 2 LCD. (C) Both monomer have high fluctuations for the N-terminal loops while the RRM residues have significantly more stability. The LCD in both monomers have the highest fluctuations, although monomer 1 has higher ones. This is due to the fact that the monomer 1 LCD has less surface contact compared to the monomer 2 LCD.(D) Total electrostatic energy analysis demonstrates an interesting trend. The LCD and RRMs of monomer 1 do not form significant electrostatic contacts, which is same also for the monomer 2 LCD and RRMs. However, there is significant gradual increase in electrostatic contacts between the two individual monomers.(E) some residues of the LCD of monomer 1 have a positive cross correlation with the RRMs of monomer 1 and same for monomer 2 (pink box). The RRM1 of monomer 1 is positively correlated with the LCD of monomer 2 (red box).

4.4 Discussion

A previous study analyzing spatial isoform expression of hnRNPA1 demonstrated the CNS-specific roles of hnRNPA1-B across various species, including humans.[104] In humans, mice and other mammals, hnRNPA1-A is present as an addition, for more ubiquitous roles in the body, but not found in chicken or frogs.[166] Thus, it is suggested that hnRNPA1-A has mammalian-specific roles that arose due to evolutionary differences, hence a reason for other animals to not express it.[104] Nevertheless, hnRNPA1-A is not seen aggregating or mislocalized anywhere in the human body.[19] However, hnRNPA1-B, found exclusively in the CNS, has been reported multiple times to be mislocalized to the cytoplasm, co-localized with stress granules and clustered in the cytosol.[6,63,66,104,167]

Although the dynamics of aggregation in a cellular environment are both difficult to capture *in vitro* and *in silico*, the probability in all A1-A systems for the LCD to not intervene with the RNPs is an interesting observation. This can not be confirmed without supplementary *in vitro* data, but the theory of A1-A not aggregating in the human body could possibly be related to the smaller size of the LCD that structurally does not render the protein dysfunctional spontaneously, unlike with the A1-B system.

Another possibility is A1-B simply having more cytoplasm-specific roles than A1-A that naturally require it to shuttle to the cytoplasm more often.[104,167] This is supported by the fact that a research team at the University of Montreal[104] discovered that A1-B had formed cytosolic granular structures in a human neuroblastoma cell line and mice cortical neurons. Outside of the neurons *in vivo*, A1-B was also detected as granules along the intra-axonal regions of the sciatic nerve although it did not co-localize with myelin.[104] However, A1-A displayed no such presence. Historically, RBPs have been considered crucial for actin cytoskeleton maintenance and neuronal

development. So A1-B may provide with additional roles in neuronal processes that require its cytosolic presence which the A1-A does is exempted from.[104]

Previous literature explained that A1-A and A1-B aid in transport of mRNA to the cytoplasm whereafter transportin-1 (TNPO-1) binds the M9 sequence to allows A1 to return to the nucleus.[74] Since A1 is a target for post-translational modifications such as SOMOylation and phosphorylation, there may be isoform-specific modifications done to A1-B that keep it in the cytoplasm.[168] Collectively, these data suggest that A1-B having extra residues and thus a large LCD than A1-A may have a higher probability of dysfunctional RBPs due to less LCD-RRM interactions. This observation, along with the possibility of A1-B having higher cytosolic presence than A1-A due to additional roles in neurons, may together explain how A1-B has more pathophysiological roles than A1-A with respect to neurodegeneration.

The dimer system, although only constructed for the A1-A isoform, demonstrates the extent of protein-protein contact maintained for 1000 ns. The monomers were not bound by cooperative RNA-interactions, and yet, not only stayed intact, but increased contacts over the course of the simulation (Figure 4.14). While multiple macromolecules were not placed in the system to simulate LLPS-like environments,[10] the maintenance of close contacts of two A1 monomers may hint at the potential dangers of local LCD concentrations increasing in the cell, wherein disassembly of the complex interaction networks get difficult.

Chapter 5: Summary and Conclusion

Rollins et. al[56] had reported the significance of the 5'-AG motif for RNA-A1 interactions with the use of biochemical assays, while Beusch et. al[17] delved into deeper insights regarding protein residues driving RNA-RRM specificity. However, for the purposes of drug design, the exact interactions at the molecular level had remained unclear. Drug discovery, especially pharmacophore-based modeling, often requires the knowledge of specific interactions by natural ligands that can be replicated by a candidate drug[169]. Since protein-RNA interfaces are not well-understood, as a result of the variety of structural diversity displayed by RNA molecules, a more specific approach was taken for A1.[145] The use of an array of computational modeling techniques was employed to bridge the knowledge gap of A1's RNA-binding preferences and use those interactions as a basis for drug-design. This research was able to bridge some of the knowledge gap by (1) gaining a deeper understanding of nucleotide preferences in the RRM1 binding pocket and (2) how interdomain contacts could potentially disrupt these interactions.

With this work, the nucleotide binding preferences of each critical residue in the RNP motifs and interdomain linker have been identified, with 5′-AG being a key recognition motif but being replaceable with a 5′-GG motif (Chapter 3 Section 3.1). Optogenetic clustering data and binding-free energy analysis on dsRNA also demonstrated that an overall increase in G nucleotides in the RNA contributes contribute to higher RNA adherence (Chapter 3 Section 3.2). Such insights may aid in the design of RNA-based therapies or nucleoside analogs as it may allow the selection of candidate drugs that mimic key RNA-A1 interactions. Thorough insights from Chapter 3 altogether capture critical aspects of A1-RNA binding that may aid in therapeutic design.

The second key aspect of designing therapeutics for A1 required taking into account the interdomain contact which could potentially impact A1's RNA recognition. To date, no other work has reported on key hotspot residues in the RRMs that the LCD has potential to block. In this

research, extensive 1μs long simulations allowed observation of the LCD's potential of folding over the RRMs in isoform B, blocking RNA-binding residues such as Phe17, Phe59, His101 and Arg92 (Chapter 4). The aforementioned residues were proven earlier to be critical for RNA-A1 complex stability (Chapter 3). Hence, the LCD forming interactions with these residues may prevent RNA from entering the RNP binding pocket of RRM1. The design of therapeutics, as depicted in Figure 4.1, may prevent A1's dysfunction by preventing interdomain contacts that prevent RNA from binding the protein.

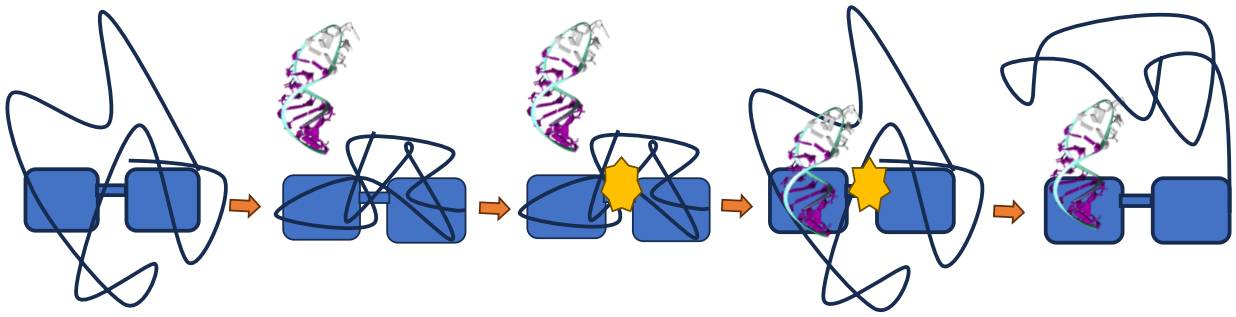


Figure 5.1: Hypothetical representation of the effect of a candidate drug on A1's interdomain contacts. When the LCD folds over the RRMs, it could potentially leave the RNA-binding site occupied and make it difficult to bind RNA. Finding a suitable druggable site that would help keep the LCD free and stabilized may aid in maintaining the functionality of A1.

However, critical RNA sites overall do not seem to be affected in isoform A of A1 in this work. Research has proven earlier that A1-B is expressed higher in neurons and has a higher aggregation-propensity than A1-A, which renders A1-A.[104] One of the limitations of this study is that complex systems with multiple proteins were not set up to thoroughly investigate aggregation. Therefore, correlating a single protein monomer's increased interdomain contacts or lack thereof with its ability to aggregate, may not be suitable since protein aggregation depends heavily on multivalent interactions.[10] A1-A having little or no interdomain contacts compared to A1-B leaves little room

for interpretation other than the longer length of A1-B allowing more LCD-RRM interactions that could be potentially hinder RNA-binding.

A1's mechanism of aggregation, however, remains an interesting question that is yet to be explored. Although LLPS is a well-accepted theory for how proteins with LCDs or prion-like domains aggregate in the cell[7,10], it is unclear how LLPS is triggered. Diseases such as Frontotemporal dementia[7], Alzheimer's disease[170], ALS[171], to name a few, are examples of diseases where LLPS occurs and preventing aggregation and LLPS in these diseases have remained a challenge for the past few years. Understanding the key interactions that induce LLPS and result in irreversible condensates in the cell is an active area of research which holds potential for designing novel therapeutics[10]. However, studying LLPS requires the observation of a complex system with multiple macromolecules in a physiological environment to capture the dynamics at the molecular level[172]. However, this remains a challenge both *in vitro*[7,173] and *in silico*[174] to study since the system is too dynamic for *in vitro*[7,173]characterization and often too complex to be set up *in silico*[174].

While studying A1's aggregation mechanism is important to understand the big picture of its pathophysiology, there are other experiments that can be done to gain some understanding for its dysfunction. Understanding few interactions at a time can aid in piecing together a hypothesized mechanism for aggregation[10]. For instance, A1 is known to shuttle between the nucleus and cytoplasm and is only known to aggregate in the cytoplasm due to mutations or dysfunction[83]. In order for A1 to return to the cytoplasm, it requires interactions with the nuclear pore protein, Kaβ2[74]. Kaβ2 recognizes the nuclear localization sequence (NLS) present in the LCD of A1[74]. As seen in Chapter 4 and hypothesized previously[74], the LCD may cluster over the RRMs, potentially

preventing any access to the NLS, thus preventing its re-entry into the nucleus. In fact, it has been proven that *in vivo,* Kapβ2 prevents the association of RBPs such as FUS and hnRNPA2 into stress granules found in the cytoplasm.[175] Based on this evidence, it may be hypothesized that the presence of KaPβ2, A1's association with stress granules, and potentially, the ability to undergo LLPS, may be reduced[74]. Therefore, *in silico* modeling of Kapβ2-A1 interactions would be an interesting study as a next step. Similar to how experiments were conducted in Chapter 4, the effect of Kapβ2 binding on A1's interdomain contacts may help provide insights which could potentially help in understanding A1's aggregation. For example, mutations that may reduce A1's binding to Kapβ2 may help explain how A1 becomes dysfunctional in the cytoplasm by not be imported into the nucleus due to reduced interactions with Kapβ2.[74]

Other avenues to explore to better understand A1's aggregation include the study of post-translational modifications[168]. Earlier, it had been proven that methylation of Arginines in the RGG box of A1 is often required to associate A1 with stress granules[168]. Stress granules are condensates in the cell that are formed via LLPS to store RNA and RBPs[10]. They become pathological when they become irreversible[10]. A1 associating with stress granules often stays mislocalized in the cytoplasm and unable to perform its RNA-splicing functions[176]. Therefore, considering the link between stress granules, LLPS and RBPs, studying the methylated A1 protein to observe how interdomain contacts are affected may aid in understanding additional aspects of A1's initial aggregation and mis localization mechanism[168].

Overall, the study of A1's aggregation requires further investigation to make progress towards effective therapies in diseases it is implicated in. While this research elaborated on A1's RNA-binding capacity and interdomain interactions, the design of therapeutics requires more thorough

investigation on A1's pathophysiological roles at the molecular level. Progress towards understating and finding treatments for diseases A1 is involved in, or neurodegeneration in general, is an active area of research due to their detrimental implications[66,7,12]. Neurodegenerative diseases remain complex problems that require progress using an array of different approaches to string together a hopeful solution for those affected by the misfortune.

References

1. Pérez-Ortín JE, Tordera V, Chávez S. Homeostasis in the central dogma of molecular biology: The importance of mRNA instability. *RNA Biol*. 2019;16(12):1659-1666. doi: 10.1080/15476286.2019.1655352 [doi].

2. Gebauer F, Schwarzl T, Valcárcel J, Hentze MW. RNA-binding proteins in human genetic disease. *Nature Reviews Genetics*. 2021;22(3):185-198. doi: 10.1038/s41576-020-00302-y.

3. Newman R, McHugh J, Turner M. RNA binding proteins as regulators of immune cell biology. *Clin Exp Immunol*. 2016;183(1):37-49. doi: 10.1111/cei.12684 [doi].

4. Oliveira C, Faoro H, Alves LR, Goldenberg S. RNA-binding proteins and their role in the regulation of gene expression in trypanosoma cruzi and saccharomyces cerevisiae. *Genet Mol Biol*. 2017;40(1):22-30. doi: S1415-47572017000100022 [pii].

5. Kelaini, SophiaAU - Chan, CelineAU - Cornelius, Victoria A.AU - Margariti, AndrianaTI - RNA-Binding Proteins Hold Key Roles in Function, Dysfunction,and Disease. *Biology*. 2021;10(5). doi: 10.3390/biology10050366.

6. Clarke JP, Thibault PA, Salapa HE, Levin MC. A comprehensive analysis of the role of hnRNP A1 function and dysfunction in the pathogenesis of neurodegenerative disease. *Front Mol Biosci*. 2021;8:659610. doi: 10.3389/fmolb.2021.659610 [doi].

7. Molliex A, Temirov J, Lee J, et al. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell*. 2015;163(1):123-133. doi: S0092-8674(15)01176-9 [pii].

8. Tsoi PS, Quan MD, Choi KJ, Dao KM, Ferreon JC, Ferreon ACM. Electrostatic modulation of hnRNPA1 low-complexity domain liquid-liquid phase separation and aggregation. *Protein Sci*. 2021;30(7):1408-1417. doi: 10.1002/pro.4108 [doi].

9. Molliex A, Temirov J, Lee J, et al. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell*. 2015;163(1):123-133. doi: S0092-8674(15)01176-9 [pii].

10. Aledo JC. The role of methionine residues in the regulation of liquid-liquid phase separation. *Biomolecules (Basel, Switzerland)*. 2021;11(8):1248. doi: 10.3390/biom11081248.

11. Liu M, Li H, Luo X, et al. RPS: A comprehensive database of RNAs involved in liquid–liquid phase separation. *Nucleic Acids Res*. 2022;50:D347-D355. doi: 10.1093/nar/gkab986.

12. Babinchak WM, Surewicz WK. Liquid-liquid phase separation and its mechanistic role in pathological protein aggregation. *J Mol Biol*. 2020;432(7):1910-1925. doi: S0022-2836(20)30225-4 [pii].

13. Harrison AF, Shorter J. RNA-binding proteins with prion-like domains in health and disease. *Biochem J*. 2017;474(8):1417-1438. doi: 10.1042/BCJ20160499 [doi].

14. Clarke J, Thibault P, Fatima S, et al. Sequence and structure-specific RNA oligonucleotide biding attenuates protein A1 dysfunction. *Frontiers in molecular biosciences*, *10*, 1178439. https://doi.org/10.3389/fmolb.2023.1178439

15. Onufriev A, Bashford D, Case DA. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins*. 2004;55(2):383-394. doi: 10.1002/prot.20033 [doi].

16. Dang M, Li Y, Song J. Tethering-induced destabilization and ATP-binding for tandem RRM domains of ALS-causing TDP-43 and hnRNPA1. *Sci Rep*. 2021;11(1):1034-6. doi: 10.1038/s41598-020-80524-6 [doi].

17. Beusch I, Barraud P, Moursy A, Cléry A, Allain FH. Tandem hnRNP A1 RNA recognition motifs act in concert to repress the splicing of survival motor neuron exon 7. *Elife*. 2017;6:10.7554/eLife.25736. doi: e25736 [pii].

18. Geuens T, Bouhy D, Timmerman V. The hnRNP family: Insights into their role in health and disease. *Hum Genet*. 2016;135(8):851-867.doi: 10.1007/s00439-016-1683-5.

19. Jean-Philippe J, Paz S, Caputi M. hnRNP A1: The swiss army knife of gene expression. *Int J Mol Sci*. 2013;14(9):18999-19024. doi: 10.3390/ijms140918999 [doi].

20. Lau JS, Baumeister P, Kim E, et al. Heterogeneous nuclear ribonucleoproteins as regulators of gene expression through interactions with the human thymidine kinase promoter. *Journal of cellular biochemistry*. 2000;79(3):395-406. doi: 10.1002/1097-4644(20001201)79:33.0.CO;2-M.

21. Xia H. Regulation of γ-fibrinogen chain expression by heterogeneous nuclear ribonucleoprotein A1. *The Journal of biological chemistry*. 2005;280(13):13171-13178. doi: 10.1074/jbc.M414120200.

22. Chen H, Hewison M, Hu B, Adams JS. Heterogeneous nuclear ribonucleoprotein (hnRNP) binding to hormone response elements: A cause of vitamin D resistance. *Proceedings of the National Academy of Sciences - PNAS*. 2003;100(10):6109-6114. doi: 10.1073/pnas.1031395100.

23. Campillos M, Lamas JR, García MA, Bullido MJ, Valdivieso F, Vázquez J. Specific interaction of heterogeneous nuclear ribonucleoprotein A1 with the –219T allelic form modulates APOE promoter activity. *Nucleic acids research*. 2003;31(12):3063-3070. doi: 10.1093/nar/gkg435.

24. Ferino A, Marquevielle J, Choudhary H, et al. hnRNPA1/UP1 unfolds KRAS G-quadruplexes and feeds a regulatory axis controlling gene expression. *ACS Omega*. 2021;6(49):34092-34106. doi: 10.1021/acsomega.1c05538.

25. Paramasivam M, Membrino A, Cogoi S, Fukuda H, Nakagama H, Xodo LE. Protein hnRNP A1 and its derivative Up1 unfold quadruplex DNA in the human KRAS promoter: Implications for transcription. *Nucleic Acids Research*. 2009;37(9):2841-2853. doi: 10.1093/nar/gkp138.

26. Takimoto M, Tomonaga T, Matunis M, et al. Specific binding of heterogeneous ribonucleoprotein particle protein K to the human c-myc promoter, in vitro. *The Journal of biological chemistry*. 1993;268(24):18249-18258. doi: 10.1016/S0021-9258(17)46837-2.

27. Blackburn EH. Switching and signaling at the telomere. *Cell*. 2001;106(6):661-673.

28. Meeker A, De Marzo A. Recent advances in telomere biology: Implications for human cancer. *Current opinion in oncology*. 2004;16(1):32-38. doi: 10.1097/00001622-200401000-00007.

29. Zhang Q, Manche L, Xu R, Krainer AR. hnRNP A1 associates with telomere ends and stimulates telomerase activity. *RNA (Cambridge)*. 2006;12(6):1116-1128. doi: 10.1261/rna.58806.

30. LaBranche H, Dupuis S, Wellinger RJ, Chabot B, Bani M, Ben-David Y. Telomere elongation by hnRNP A1 and a derivative that interacts with telomeric repeats and telomerase. *Nature genetics*. 1998;19(2):199-202. doi: 10.1038/575.

31. Fiset S, Chabot B. hnRNP A1 may interact simultaneously with telomeric DNA and the human telomerase RNA in vitro. *Nucleic Acids Res*. 2001;29(11):2268-2275. doi: 10.1093/nar/29.11.2268.

32. Le PN, Maranon DG, Altina NH, Battaglia CLR, Bailey SM. TERRA, hnRNP A1, and DNA-PKcs interactions at human telomeres. *Frontiers in oncology*. 2013;3:91. doi: 10.3389/fonc.2013.00091.

33. Mili S, Shu HJ, Zhao Y, Piñol-Roma S. Distinct RNP complexes of shuttling hnRNP proteins with pre-mRNA and mRNA: Candidate intermediates in formation and export of mRNA. *Molecular and Cellular Biology*. 2001;21(21):7307-7319. doi: 10.1128/MCB.21.21.7307-7319.2001.

34. Piñol-Roma S, Dreyfuss G. Shuttling of pre-mRNA binding proteins between nucleus and cytoplasm. *Nature (London)*. 1992;355(6362):730-732. doi: 10.1038/355730a0.

35. Visa N, Alzhanova-Ericsson AT, Sun X, et al. A pre-mRNA-binding protein accompanies the RNA from the gene through the nuclear pores and into polysomes. *Cell*. 1996;84(2):253-264. doi: 10.1016/S0092-8674(00)80980-0.

36. Siebrasse JP, Kaminski T, Kubitscheck U. Nuclear export of single native mRNA molecules observed by light sheet fluorescence microscopy. *Proceedings of the National Academy of Sciences - PNAS*. 2012;109(24):9426-9431. doi: 10.1073/pnas.1201781109.

37. Piñol-Roma S, Dreyfuss G. Shuttling of pre-mRNA binding proteins between nucleus and cytoplasm. *Nature (London)*. 1992;355(6362):730-732. doi: 10.1038/355730a0.

38. Vautier D, Chesne P, Cunha C, Calado A, Renard JP, Carmo-Fonseca M. Transcription-dependent nucleocytoplasmic distribution of hnRNP A1 protein in early mouse embryos. *Journal of Cell Science*. 2001;114(8):1521-1531. doi: 10.1242/jcs.114.8.1521.

39. Suzuki H, Matsuoka M. hnRNPA1 autoregulates its own mRNA expression to remain non-cytotoxic. *Mol Cell Biochem*. 2017;427(1-2):123-131. doi: 10.1007/s11010-016-2904-x.

40. Zearfoss NR, Clingman CC, Farley BM, McCoig LM, Ryder SP. Quaking regulates Hnrnpa1 expression through its 3′ UTR in oligodendrocyte precursor cells. *PLoS Genetics*. 2011;7(1):e1001269. doi: 10.1371/journal.pgen.1001269.

41. Suzuki H, Lee K, Matsuoka M. TDP-43-induced death is associated with altered regulation of BIM and bcl-xL and attenuated by caspase-mediated TDP-43 cleavage. *The Journal of biological chemistry*. 2011;286(15):13171-13183. doi: 10.1074/jbc.M110.197483.

42. Suzuki H, Matsuoka M. Overexpression of nuclear FUS induces neuronal cell death. *Neuroscience*. 2015;287:113-124. doi: 10.1016/j.neuroscience.2014.12.007.

43. Ayala YM, De Conti L, Avendaño-Vázquez SE, et al. TDP-43 regulates its mRNA levels through a negative feedback loop. *The EMBO journal*. 2011;30(2):277-288. doi: 10.1038/emboj.2010.310.

44. Zhou Y, Liu S, Liu G, Oztürk A, Hicks GG. ALS-associated FUS mutations result in compromised FUS alternative splicing and autoregulation. *PLoS Genetics*. 2013;9(10):e1003895. doi: 10.1371/journal.pgen.1003895.

45. Roy R, Huang Y, Seckl MJ, Pardo OE. Emerging roles of hnRNPA1 in modulating malignant transformation. *Wiley Interdiscip Rev RNA*. 2017;8(6):10.1002/wrna.1431. Epub 2017 Aug 8. doi: 10.1002/wrna.1431.

46. Berson A, Barbash S, Shaltiel G, et al. Cholinergic-associated loss of hnRNP-A/B in alzheimer's disease impairs cortical splicing and cognitive function in mice. *EMBO Molecular Medicine*. 2012;4(8):730-742. doi: 10.1002/emmm.201100995.

47. Honda H, Hamasaki H, Wakamiya T, et al. Loss of hnRNPA1 in ALS spinal cord motor neurons with TDP-43-positive inclusions. *Neuropathology*. 2015;35(1):37-43. doi: 10.1111/neup.12153.

48. Levengood JD, Tolbert BS. Idiosyncrasies of hnRNP A1-RNA recognition: Can binding mode influence function. *Semin Cell Dev Biol*. 2019;86:150-161. doi: S1084-9521(17)30470-6 [pii].

49. Kaur R, Lal SK. The multifarious roles of heterogeneous ribonucleoprotein A1 in viral infections. *Reviews in Medical Virology*. 2020;30(2):e2097-n/a. doi: 10.1002/rmv.2097.

50. Cammas A, Pileur F, Bonnal S, et al. Cytoplasmic relocalization of heterogeneous nuclear ribonucleoprotein A1 controls translation initiation of specific mRNAs. *Molecular Biology of the Cell*. 2007;18(12):5048-5059. doi: 10.1091/mbc.E07-06-0603.

51. Shih S, Stollar V, Li M. Host factors in enterovirus 71 replication. *Journal of Virology*. 2011;85(19):9658-9666. doi: 10.1128/JVI.05063-11.

52. Lin J, Shih S, Manjing Pan, et al. hnRNP A1 interacts with the 5′ untranslated regions of enterovirus 71 and sindbis virus RNA and is required for viral replication. *Journal of Virology*. 2009;83(12):6106-6114. doi: 10.1128/JVI.02476-08.

53. Kim CS, Seol SK, Song O, Park JH, Jang SK. An RNA-binding protein, hnRNP A1, and a scaffold protein, septin 6, facilitate hepatitis C virus replication. *Journal of Virology*. 2007;81(8):3852-3865. doi: 10.1128/JVI.01311-06.

54. Zhao X, Schwartz S. Inhibition of HPV-16 L1 expression from L1 cDNAs correlates with the presence of hnRNP A1 binding sites in the L1 coding region. *Virus Genes*. 2008;36(1):45-53. doi: 10.1007/s11262-007-0174-0.

55. Karn J, Stoltzfus CM. Transcriptional and posttranscriptional regulation of HIV-1 gene expression. *Cold Spring Harbor perspectives in medicine*. 2012;2(2):a006916. doi: 10.1101/cshperspect.a006916.

56. Rollins C, Levengood JD, Rife BD, Salemi M, Tolbert BS. Thermodynamic and phylogenetic insights into hnRNP A1 recognition of the HIV-1 exon splicing silencer 3 element. *Biochemistry*. 2014;53(13):2172-2184. doi: 10.1021/bi500180p [doi].

57. Monette A, Ajamian L, López-Lastra M, Mouland AJ. Human immunodeficiency virus type 1 (HIV-1) induces the cytoplasmic retention of heterogeneous nuclear ribonucleoprotein A1 by disrupting nuclear import: Implications for HIV-1 gene expression. *The Journal of biological chemistry*. 2009;284(45):31350-31362. doi: 10.1074/jbc.M109.048736.

58. Najera I, Krieg M, Karn J. Synergistic stimulation of HIV-1 rev-dependent export of unspliced mRNA to the cytoplasm by hnRNP A1. *Journal of molecular biology*. 1999;285(5):1951-1964. doi: 10.1006/jmbi.1998.2473.

59. Lin J, Li M, Brewer G. mRNA decay factor AUF1 binds the internal ribosomal entry site of enterovirus 71 and inhibits virus replication. *PLoS ONE*. 2014;9(7):e103827. doi: 10.1371/journal.pone.0103827.

60. Lin J, Chen T, Weng K, Chang S, Chen L, Shih S. Viral and host proteins involved in picornavirus life cycle. *Journal of Biomedical Science*. 2009;16(1):103. doi: 10.1186/1423-0127-16-103.

61. Kim CS, Seol SK, Song O, Park JH, Jang SK. An RNA-binding protein, hnRNP A1, and a scaffold protein, septin 6, facilitate hepatitis C virus replication. *Journal of Virology*. 2007;81(8):3852-3865. doi: 10.1128/JVI.01311-06.

62. Ríos-Marco P, Romero-López C, Berzal-Herranz A. The cis-acting replication element of the hepatitis C virus genome recruits host factors that influence viral replication and translation. *Scientific Reports*. 2016;6(1):25729. doi: 10.1038/srep25729.

63. Levin MC, Lee S, Gardner LA, Shin Y, Douglas JN, Salapa H. Autoantibodies to heterogeneous nuclear ribonuclear protein A1 (hnRNPA1) cause altered 'ribostasis' and neurodegeneration; the legacy of HAM/TSP as a model of progressive multiple sclerosis. *J Neuroimmunol*. 2017;304:56-62. doi: S0165-5728(16)30155-2 [pii].

64. Salapa HE, Hutchinson C, Popescu BF, Levin MC. Neuronal RNA-binding protein dysfunction in multiple sclerosis cortex. *Ann Clin Transl Neurol*. 2020;7(7):1214-1224. doi: 10.1002/acn3.51103 [doi].

65. Karussis D. The diagnosis of multiple sclerosis and the various related demyelinating syndromes: A critical review. *J Autoimmun*. 2014;48-49:134-142. doi: S0896-8411(14)00025-0 [pii].

66. Salapa HE, Lee S, Shin Y, Levin MC. Contribution of the degeneration of the neuro-axonal unit to the pathogenesis of multiple sclerosis. *Brain Sci*. 2017;7(6):69. doi: 10.3390/brainsci7060069. doi: 10.3390/brainsci7060069 [doi].

67. Díaz C, Zarco LA, Rivera DM. Highly active multiple sclerosis: An update. *Mult Scler Relat Disord*. 2019;30:215-224. doi: S2211-0348(19)30038-0 [pii].

68. Kim HJ, Kim NC, Wang Y, et al. Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. *Nature (London)*. 2013;495(7442):467-473. doi: 10.1038/nature11922.

69. DONEV R, NEWALL A, THOME J, SHEER D. A role for SC35 and hnRNPA1 in the determination of amyloid precursor protein isoforms. *Molecular psychiatry*. 2007;12(7):681-690. doi: 10.1038/sj.mp.4001971.

70. Berson A, Barbash S, Shaltiel G, et al. Cholinergic-associated loss of hnRNP-A/B in alzheimer's disease impairs cortical splicing and cognitive function in mice. *EMBO Molecular Medicine*. 2012;4(8):730-742. doi: 10.1002/emmm.201100995.

71. Bekenstein U, Soreq H. Heterogeneous nuclear ribonucleoprotein A1 in health and neurodegenerative disease: From structural insights to post-transcriptional regulatory roles. *Molecular and cellular neurosciences*. 2013;56:436-446. doi: 10.1016/j.mcn.2012.12.002.

72. Manley JL, Kashima T. A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. *Nature genetics*. 2003;34(4):460-463. doi: 10.1038/ng1207.

73. Hua Y, Sahashi K, Hung G, et al. Antisense correction of SMN2 splicing in the CNS rescues necrosis in a type III SMA mouse model. *Genes & development*. 2010;24(15):1634-1644. doi: 10.1101/gad.1941310.

74. Sun Y, Zhao K, Xia W, et al. The nuclear localization sequence mediates hnRNPA1 amyloid fibril formation revealed by cryoEM structure. *Nat Commun*. 2020;11(1):6349-8. doi: 10.1038/s41467-020-20227-8 [doi].

75. Jiang R, Su G, Chen X, et al. Esculetin inhibits endometrial cancer proliferation and promotes apoptosis via hnRNPA1 to downregulate BCLXL and XIAP. *Cancer letters*. 2021;521:308-321. doi: 10.1016/j.canlet.2021.08.039.

76. Yan Q, Zeng P, Zhou X, et al. RBMX suppresses tumorigenicity and progression of bladder cancer by interacting with the hnRNP A1 protein to regulate PKM alternative splicing. *Oncogene*. 2021;40(15):2635-2650. doi: 10.1038/s41388-021-01666-z.

77. Zhou B, Wang Y, Jiang J, et al. The long noncoding RNA colon cancer-associated transcript-1/miR-490 axis regulates gastric cancer cell migration by targeting hnRNPA1. *IUBMB life*. 2016;68(3):201-210. doi: 10.1002/iub.1474.

78. Zhou Z, Dai Z, Zhou S, et al. Overexpression of HnRNP A1 promotes tumor invasion through regulating CD44v6 and indicates poor prognosis for hepatocellular carcinoma. *Int J Cancer*. 2013;132(5):1080-1089. doi: 10.1002/ijc.27742.

79. Carabet LA, Leblanc E, Lallous N, et al. Computer-aided discovery of small molecules targeting the RNA splicing activity of hnRNP A1 in castration-resistant prostate cancer. *Molecules*. 2019;24(4):763. doi: 10.3390/molecules24040763.

80. Tummala R, Lou W, Gao AC, Nadiminty N. Quercetin targets hnRNPA1 to overcome enzalutamide resistance in prostate cancer cells. *Molecular cancer therapeutics*. 2017;16(12):2770-2779. doi: 10.1158/1535-7163.MCT-17-0030.

81. Gu Z, Xia J, Xu H, Frech I, Tricot G, Zhan F. NEK2 promotes aerobic glycolysis in multiple myeloma through regulating splicing of pyruvate kinase. *Journal of Hematology & Oncology*. 2017;10(1):17. doi: 10.1186/s13045-017-0392-4.

82. Patry C, Bouchard L, Labrecque P, et al. Small interfering RNA-mediated reduction in heterogeneous nuclear ribonucleoparticule A1/A2 proteins induces apoptosis in human cancer cells but not in normal mortal cell lines. *Cancer Res*. 2003;63(22):7679-7688.

83. Thibault PA, Ganesan A, Kalyaanamoorthy S, Clarke JWE, Salapa HE, Levin MC. hnRNP A/B proteins: An encyclopedic assessment of their roles in homeostasis and disease. *Biology (Basel)*. 2021;10(8):712. doi: 10.3390/biology10080712. doi: 10.3390/biology10080712 [doi].

84. Kooshapur H, Choudhury NR, Simon B, et al. Structural basis for terminal loop recognition and stimulation of pri-miRNA-18a processing by hnRNP A1. *Nat Commun*. 2018;9(1):2479-9. doi: 10.1038/s41467-018-04871-9 [doi].

85. Martin EW, Holehouse AS, Peran I, et al. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science*. 2020;367(6478):694-699. doi: 10.1126/science.aaw8653 [doi].

86. Martin EW, Thomasen FE, Milkovic NM, et al. Interplay of folded domains and the disordered low-complexity domain in mediating hnRNPA1 phase separation. *Nucleic Acids Res*. 2021;49(5):2931-2945. doi: 10.1093/nar/gkab063 [doi].

87. Shin Y, Brangwynne CP. Liquid phase condensation in cell physiology and disease. *Science (American Association for the Advancement of Science)*. 2017;357(6357):1253. doi: 10.1126/science.aaf4382.

88. Larson AG, Elnatan D, Keenen MM, et al. Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature*. 2017;547(7662):236-240. doi: 10.1038/nature22822.

89. Rubinstein M, Dobrynin Av. Solutions of associative polymers. *Trends in polymer science. (Regular ed.)*. 1997;5(6):181-186.

90. Wang J, Choi J, Holehouse AS, et al. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell*. 2018;174(3):688-699.e16. doi: 10.1016/j.cell.2018.06.006.

91. Clarke JP, Thibault PA, Fatima S, et al. Sequence- and structure-specific RNA oligonucleotide binding attenuates heterogeneous nuclear ribonucleoprotein A1 dysfunction. *Front Mol Biosci*. 2023;10:1178439. doi: 10.3389/fmolb.2023.1178439.

92. Liu X, Shi D, Zhou S, Liu H, Liu H, Yao X. Molecular dynamics simulations and novel drug discovery. *Expert Opin Drug Discov*. 2018;13(1):23-37. doi: 10.1080/17460441.2018.1403419 [doi].

93. Collier TA, Piggot TJ, Allison JR. Molecular dynamics simulation of proteins. *Methods Mol Biol*. 2020;2073:311-327. doi: 10.1007/978-1-4939-9869-2_17 [doi].

94. Carugo O, Djinović-Carugo K. Structural biology: A golden era. *PLoS Biology*. 2023;21(6):e3002187. doi: 10.1371/journal.pbio.3002187.

95. Jisna VA, Jayaraj PB. Protein structure prediction: Conventional and deep learning perspectives. *Protein J*. 2021;40(4):522-544. doi: 10.1007/s10930-021-10003-y.

96. Peterson LX, Roy A, Christoffer C, Terashi G, Kihara D. Modeling disordered protein interactions from biophysical principles. *PLoS Comput Biol*. 2017;13(4):e1005485. doi: 10.1371/journal.pcbi.1005485.

97. Hardin C, Pogorelov TV, Luthey-Schulten Z. Ab initio protein structure prediction. *Current Opinion in Structural Biology*. 2002;12(2):176-181.

98. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER suite: Protein structure and function prediction. *Nat Methods*. 2015;12(1):7-8. doi: 10.1038/nmeth.3213.

99. Roy A, Kucukural A, Zhang Y. I-TASSER: A unified platform for automated protein structure and function prediction. *Nat Protoc*. 2010;5(4):725-738. doi: 10.1038/nprot.2010.5.

100. Zhang Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*. 2008;9:40-40. doi: 10.1186/1471-2105-9-40.

101. Ruff KM, Pappu RV. AlphaFold and implications for intrinsically disordered proteins. *J Mol Biol*. 2021;433(20):167208. doi: 10.1016/j.jmb.2021.167208.

102. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589. doi: 10.1038/s41586-021-03819-2.

103. Varadi M, Anyango S, Deshpande M, et al. AlphaFold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res*. 2022;50:D439-D444. doi: 10.1093/nar/gkab1061.

104. Gagné M, Deshaies JE, Sidibé H, et al. hnRNP A1B, a splice variant of HNRNPA1, is spatially and temporally regulated. *Front Neurosci*. 2021;15:724307. doi: 10.3389/fnins.2021.724307 [doi].

105. Zhang J, Fei Y, Sun L, Zhang QC. Advances and opportunities in RNA structure experimental determination and computational modeling. *Nature methods*. 2022;19(10):1193-1207. doi: 10.1038/s41592-022-01623-y.

106. Kasprzak WK, Ahmed NA, Shapiro BA. Modeling ligand docking to RNA in the design of RNA-based nanostructures. *Current opinion in biotechnology*. 2020;63:16-25. doi: 10.1016/j.copbio.2019.10.010.

107. Ziv O, Gabryelska MM, Lun ATL, et al. COMRADES determines in vivo RNA structures and interactions. *Nature methods*. 2018;15(10):785-788. doi: 10.1038/s41592-018-0121-0.

108. Ganesan A, Coote ML, Barakat K. Molecular dynamics-driven drug discovery: Leaping forward with confidence. *Drug discovery today*. 2017;22(2):249-269. doi: 10.1016/j.drudis.2016.11.001.

109. Huang S, Grinter SZ, Zou X. Scoring functions and their evaluation methods for protein-ligand docking: Recent advances and future directions. *Physical chemistry chemical physics : PCCP*. 2010;12(4):12899-1298. doi: 10.1039/c0cp00151a.

110. Ferrari AM, Wei BQ, Costantino L, Shoichet BK. Soft docking and multiple receptor conformations in virtual screening. *Journal of medicinal chemistry*. 2004;47(21):5076-5084. doi: 10.1021/jm049756p.

111. Cooper TA, Wan L, Dreyfuss G. RNA and disease. *Cell*. 2009;136(4):777-793. doi: 10.1016/j.cell.2009.02.011.

112. Lukong KE, Chang K, Khandjian EW, Richard S. RNA-binding proteins in human genetic disease. *Trends Genet*. 2008;24(8):416-425. doi: 10.1016/j.tig.2008.05.004.

113. Nithin C, Ghosh P, Bujnicki JM. Bioinformatics tools and benchmarks for computational docking and 3D structure prediction of RNA-protein complexes. *Genes (Basel)*. 2018;9(9):432. doi: 10.3390/genes9090432. doi: 10.3390/genes9090432.

114. Ferreira LG, Dos Santos RN, Oliva G, Andricopulo AD. Molecular docking and structure-based drug design strategies. *Molecules*. 2015;20(7):13384-13421. doi: 10.3390/molecules200713384 [doi].

115. Balcerak A, Trebinska-Stryjewska A, Konopinski R, Wakula M, Grzybowska EA. RNA-protein interactions: Disorder, moonlighting and junk contribute to eukaryotic complexity. *Open Biol*. 2019;9(6):190096. doi: 10.1098/rsob.190096 [doi].

116. Zagrovic B, Bartonek L, Polyansky AA. RNA-protein interactions in an unstructured context. *FEBS Lett*. 2018;592(17):2901-2916. doi: 10.1002/1873-3468.13116 [doi].

117. Diarra Dit Konté N, Krepl M, Damberger FF, et al. Aromatic side-chain conformational switch on the surface of the RNA recognition motif enables RNA discrimination. *Nat Commun*. 2017;8(1):654-3. doi: 10.1038/s41467-017-00631-3 [doi].

118. Huang SY, Zou X. MDockPP: A hierarchical approach for protein-protein docking and its application to CAPRI rounds 15-19. *Proteins*. 2010;78(15):3096-3103. doi: 10.1002/prot.22797 [doi].

119. Dang M, Li Y, Song J. Tethering-induced destabilization and ATP-binding for tandem RRM domains of ALS-causing TDP-43 and hnRNPA1. *Sci Rep*. 2021;11(1):1034-6. doi: 10.1038/s41598-020-80524-6 [doi].

120. Pettersen EF, Goddard TD, Huang CC, et al. UCSF chimera--a visualization system for exploratory research and analysis. *J Comput Chem*. 2004;25(13):1605-1612. doi: 10.1002/jcc.20084 [doi].

121. Morgan CE, Meagher JL, Levengood JD, et al. The first crystal structure of the UP1 domain of hnRNP A1 bound to RNA reveals a new look for an old RNA binding protein. *J Mol Biol*. 2015;427(20):3241-3257. doi: S0022-2836(15)00293-4 [pii].

122. Morgan CE, Meagher JL, Levengood JD, et al. The first crystal structure of the UP1 domain of hnRNP A1 bound to RNA reveals a new look for an old RNA binding protein. *J Mol Biol*. 2015;427(20):3241-3257. doi: S0022-2836(15)00293-4 [pii].

123. Hollingsworth SA, Dror RO. Molecular dynamics simulation for all. *Neuron*. 2018;99(6):1129-1143. doi: 10.1016/j.neuron.2018.08.011.

124. Kalyaanamoorthy S, Chen YP. Modelling and enhanced molecular dynamics to steer structure-based drug discovery. *Progress in biophysics and molecular biology*. 2014;114(3):123-136. doi: 10.1016/j.pbiomolbio.2013.06.004.

125. D.A. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, J.T. Berryman, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, G.A. Cisneros, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, M.C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, A. Shajan, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J Wang, J. Wang, H. Wei, R.M. Wolf, X. Wu, Y. Xiong, Y. Xue, D.M. York, S. Zhao, and P.A. Kollman. Amber 2020. . 2020.

126. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry*. 1983;4(2):187-217. doi: 10.1002/jcc.540040211.

127. Abraham MJ, Murtola T, Schulz R, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*. 2015;1-2(C):19-25. doi: 10.1016/j.softx.2015.06.001.

128. The amber force fields. https://ambermd.org/AmberModels.php. Updated 2023.

129. Basu S, Alagar S, Bahadur RP. Unusual RNA binding of FUS RRM studied by molecular dynamics simulation and enhanced sampling method. *Biophys J*. 2021;120(9):1765-1776. doi: S0006-3495(21)00205-8 [pii].

130. Shakhno DV, Shakhno AV, Paulechka E. Efficient implementation of periodic boundary conditions in monte carlo simulation. *Journal of computational chemistry*. 2019;40(5):734-739. doi: 10.1002/jcc.25757.

131. Reif MM, Zacharias M. Computational tools for accurate binding free-energy prediction. In: *Computational methods for estimating the kinetic parameters of biological systems.* Vol 2385. New York, NY: Springer US; 2022:255-292. 10.1007/978-1-0716-1767-0_12.

132. Wang C, Greene D, Xiao L, Qi R, Luo R. Recent developments and applications of the MMPBSA method. *Frontiers in Molecular Biosciences*. 2018;4:87. doi: 10.3389/fmolb.2017.00087.

133. Kollman PA, Massova I, Reyes C, et al. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Accounts of chemical research*. 2000;33(12):889-897. doi: 10.1021/ar000033j.

134. Chen F, Sun H, Wang J, et al. Assessing the performance of MM/PBSA and MM/GBSA methods. 8. predicting binding free energies and poses of protein-RNA complexes. *RNA*. 2018;24(9):1183-1194. doi: 10.1261/rna.065896.118 [doi].

135. Ringner M. What is principal component analysis? *Nature biotechnology*. 2008;26(3):303-304. doi: 10.1038/nbt0308-303.

136. Majumder S, Giri K. An insight into the binding mechanism of viprinin and its morpholine and piperidine derivatives with HIV-1 vpr: Molecular dynamics simulation, principal component

analysis and binding free energy calculation study. *Journal of biomolecular structure & dynamics*. 2022;40(21):10918-10930. doi: 10.1080/07391102.2021.1954553.

137. Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD. Bio3d: An R package for the comparative analysis of protein structures. *Bioinformatics*. 2006;22(21):2695-2696. doi: 10.1093/bioinformatics/btl461.

138. Case DA, Belfon K, Ben-Shalom I, et al. Amber 2020. . 2020.

139. Le Grand S, Götz AW, Walker RC. SPFP: Speed without compromise—A mixed precision model for GPU accelerated molecular dynamics simulations. *Comput Phys Commun*. 2013;184(2):374-380. doi: https://doi.org/10.1016/j.cpc.2012.09.022.

140. Terry CA, Fernández MJ, Gude L, Lorente A, Grant KB. Physiologically relevant concentrations of NaCl and KCl increase DNA photocleavage by an N-substituted 9-aminomethylanthracene dye. *Biochemistry*. 2011;50(47):10375-10389. doi: 10.1021/bi200972c [doi].

141. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. *J Mol Graph*. 1996;14(1):33-8. doi: 0263785596000185 [pii].

142. Pettersen EF, Goddard TD, Huang CC, et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci*. 2021;30(1):70-82. doi: 10.1002/pro.3943 [doi].

143. Miller BR,3rd, McGee TD,Jr, Swails JM, Homeyer N, Gohlke H, Roitberg AE. MMPBSA.py: An efficient program for end-state free energy calculations. *J Chem Theory Comput*. 2012;8(9):3314-3321. doi: 10.1021/ct300418h [doi].

144. Coutsias EA, Wester MJ. RMSD and symmetry. *J Comput Chem*. 2019;40(15):1496-1508. doi: 10.1002/jcc.25802.

145. Hu W, Qin L, Li M, Pu X, Guo Y. A structural dissection of protein-RNA interactions based on different RNA base areas of interfaces. *RSC Adv*. 2018;8(19):10582-10592. doi: 10.1039/c8ra00598b [doi].

146. Sponer J, Leszczynski J, Hobza P. Electronic properties, hydrogen bonding, stacking, and cation binding of DNA and RNA bases. *Biopolymers*. 2001;61(1):3-31. doi: 10.1002/1097-0282(2001)61:13.0.CO;2-4 [pii].

147. Wilson KA, Kung RW, D'souza S, Wetmore SD. Anatomy of noncovalent interactions between the nucleobases or ribose and π-containing amino acids in RNA-protein complexes. *Nucleic Acids Res*. 2021;49(4):2213-2225. doi: 10.1093/nar/gkab008.

148. Singh NJ, Min SK, Kim DY, Kim KS. Comprehensive energy analysis for various types of π-interaction. *J Chem Theory Comput*. 2009;5(3):515-529. doi: 10.1021/ct800471b.

149. Luscombe NM, Laskowski RA, Thornton JM. Amino acid–base interactions: A three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic Acids Res*. 2001;29(13):2860-2874. doi: 10.1093/nar/29.13.2860.

150. Wells RA, Kellie JL, Wetmore SD. Significant strength of charged DNA-protein π-π interactions: A preliminary study of cytosine. *J Phys Chem B*. 2013;117(36):10462-10474. doi: 10.1021/jp406829d.

151. Shishkin SS, Kovalev LI, Pashintseva NV, Kovaleva MA, Lisitskaya K. Heterogeneous nuclear ribonucleoproteins involved in the functioning of telomeres in malignant cells. *Int J Mol Sci*. 2019;20(3):745. doi: 10.3390/ijms20030745. doi: 10.3390/ijms20030745 [doi].

152. Fukuda H, Katahira M, Tsuchiya N, et al. Unfolding of quadruplex structure in the G-rich strand of the minisatellite repeat by the binding protein UP1. *Proc Natl Acad Sci U S A*. 2002;99(20):12685-12690. doi: 152456899 [pii].

153. Paramasivam M, Membrino A, Cogoi S, Fukuda H, Nakagama H, Xodo LE. Protein hnRNP A1 and its derivative Up1 unfold quadruplex DNA in the human KRAS promoter: Implications for transcription. *Nucleic Acids Res*. 2009;37(9):2841-2853. doi: 10.1093/nar/gkp138.

154. Krüger AC, Raarup MK, Nielsen MM, et al. Interaction of hnRNP A1 with telomere DNA G-quadruplex structures studied at the single molecule level. *Eur Biophys J*. 2010;39(9):1343-1350. doi: 10.1007/s00249-010-0587-x.

155. Ferino A, Marquevielle J, Choudhary H, et al. hnRNPA1/UP1 unfolds KRAS G-quadruplexes and feeds a regulatory axis controlling gene expression. *ACS Omega*. 2021;6(49):34092-34106. doi: 10.1021/acsomega.1c05538 [doi].

156. Ghosh M, Singh M. Structure specific recognition of telomeric repeats containing RNA by the RGG-box of hnRNPA1. *Nucleic Acids Res*. 2020;48(8):4492-4506. doi: 10.1093/nar/gkaa134.

157. Onufriev A, Bashford D, Case DA. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins*. 2004;55(2):383-394. doi: 10.1002/prot.20033 [doi].

158. Clarke JWE, Thibault PA, Salapa HE, Kim DE, Hutchinson C, Levin MC. Multiple sclerosis-associated hnRNPA1 mutations alter hnRNPA1 dynamics and influence stress granule formation. *International journal of molecular sciences*. 2021;22(6):2909. doi: 10.3390/ijms22062909.

159. Rollins C, Levengood JD, Rife BD, Salemi M, Tolbert BS. Thermodynamic and phylogenetic insights into hnRNP A1 recognition of the HIV-1 exon splicing silencer 3 element. *Biochemistry*. 2014;53(13):2172-2184. doi: 10.1021/bi500180p [doi].

160. Ghosh M, Singh M. RGG-box in hnRNPA1 specifically recognizes the telomere G-quadruplex DNA and enhances the G-quadruplex unfolding ability of UP1 domain. *Nucleic Acids Res*. 2018;46(19):10246-10261. Accessed 8/29/2023. doi: 10.1093/nar/gky854.

161. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput*. 2015;11(8):3696-3713. doi: 10.1021/acs.jctc.5b00255 [doi].

162. Roe DR, Cheatham TE,3rd. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput*. 2013;9(7):3084-3095. doi: 10.1021/ct400341p [doi].

163. Schrödinger L. The PyMOL molecular graphics system. . ;2.0.

164. Ottoz DSM, Berchowitz LE. The role of disorder in RNA binding affinity and specificity. *Open Biol*. 2020;10(12):200328. doi: 10.1098/rsob.200328 [doi].

165. Chakrabarti P, Chakravarty D. Intrinsically disordered proteins/regions and insight into their biomolecular interactions. *Biophys Chem*. 2022;283:106769. doi: 10.1016/j.bpc.2022.106769.

166. Gueroussov S, Weatheritt RJ, O'Hanlon D, et al. Regulatory expansion in mammals of multivalent hnRNP assemblies that globally control alternative splicing. *Cell*. 2017;170(2):324-339.e23. doi: 10.1016/j.cell.2017.06.037.

167. Jahanbazi Jahan-Abad A, Salapa HE, Libner CD, Thibault PA, Levin MC. hnRNP A1 dysfunction in oligodendrocytes contributes to the pathogenesis of multiple sclerosis. *Glia*. 2022. doi: 10.1002/glia.24300.

168. Wall ML, Lewis SM. Methylarginines within the RGG-motif region of hnRNP A1 affect its IRES trans-acting factor activity and are required for hnRNP A1 stress granule localization and formation. *J Mol Biol*. 2017;429(2):295-307. doi: 10.1016/j.jmb.2016.12.011.

169. Giordano D, Biancaniello C, Argenio MA, Facchiano A. Drug design by pharmacophore and virtual screening approach. *Pharmaceuticals (Basel, Switzerland)*. 2022;15(5):646. doi: 10.3390/ph15050646.

170. Carlomagno Y, Manne S, DeTure M, et al. The AD tau core spontaneously self-assembles and recruits full-length tau to filaments. *Cell Rep*. 2021;34(11):108843. doi: S2211-1247(21)00157-1 [pii].

171. Lu Y, Lim L, Song J. RRM domain of ALS/FTD-causing FUS characteristic of irreversible unfolding spontaneously self-assembles into amyloid fibrils. *Scientific Reports*. 2017;7(1):1043-14. doi: 10.1038/s41598-017-01281-7.

172. Paloni M, Bussi G, Barducci A. Arginine multivalency stabilizes protein/RNA condensates. *Protein Sci*. 2021;30(7):1418-1426. doi: 10.1002/pro.4109 [doi].

173. Alberti S, Gladfelter A, Mittag T. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell*. 2019;176(3):419-434. doi: 10.1016/j.cell.2018.12.035.

174. Szała-Mendyk B, Phan TM, Mohanty P, Mittal J. Challenges in studying the liquid-to-solid phase transitions of proteins using computer simulations. *Current opinion in chemical biology*. 2023;75:102333. doi: 10.1016/j.cbpa.2023.102333.

175. Guo L, Kim HJ, Wang H, et al. Nuclear-import receptors reverse aberrant phase transitions of RNA-binding proteins with prion-like domains. *Cell*. 2018;173(3):677-692.e20. doi: 10.1016/j.cell.2018.03.002.

176. Salapa HE, Johnson C, Hutchinson C, Popescu BF, Levin MC. Dysfunctional RNA binding proteins and stress granules in multiple sclerosis. *Journal of neuroimmunology*. 2018;324:149-156. doi: 10.1016/j.jneuroim.2018.08.015.

Appendix

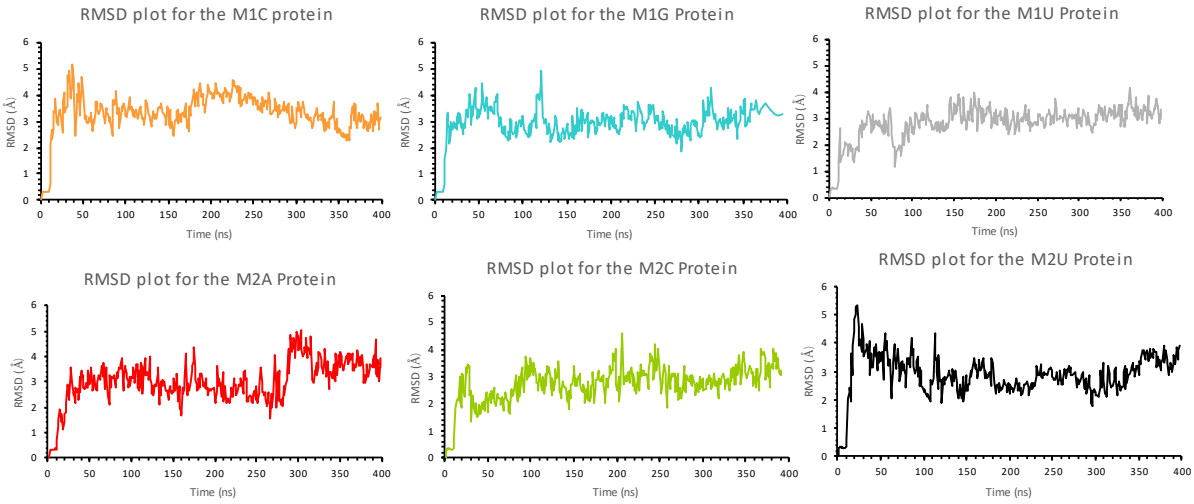


Figure A1: The RMSD Plots for the protein component of the M1N and M2N demonstrate predominant stability in all complexes following the first ~50 ns. This is explained by the RRM1 of A1 is a folded structure, consisting of beta sheets, which are not likely to change conformation extensively.

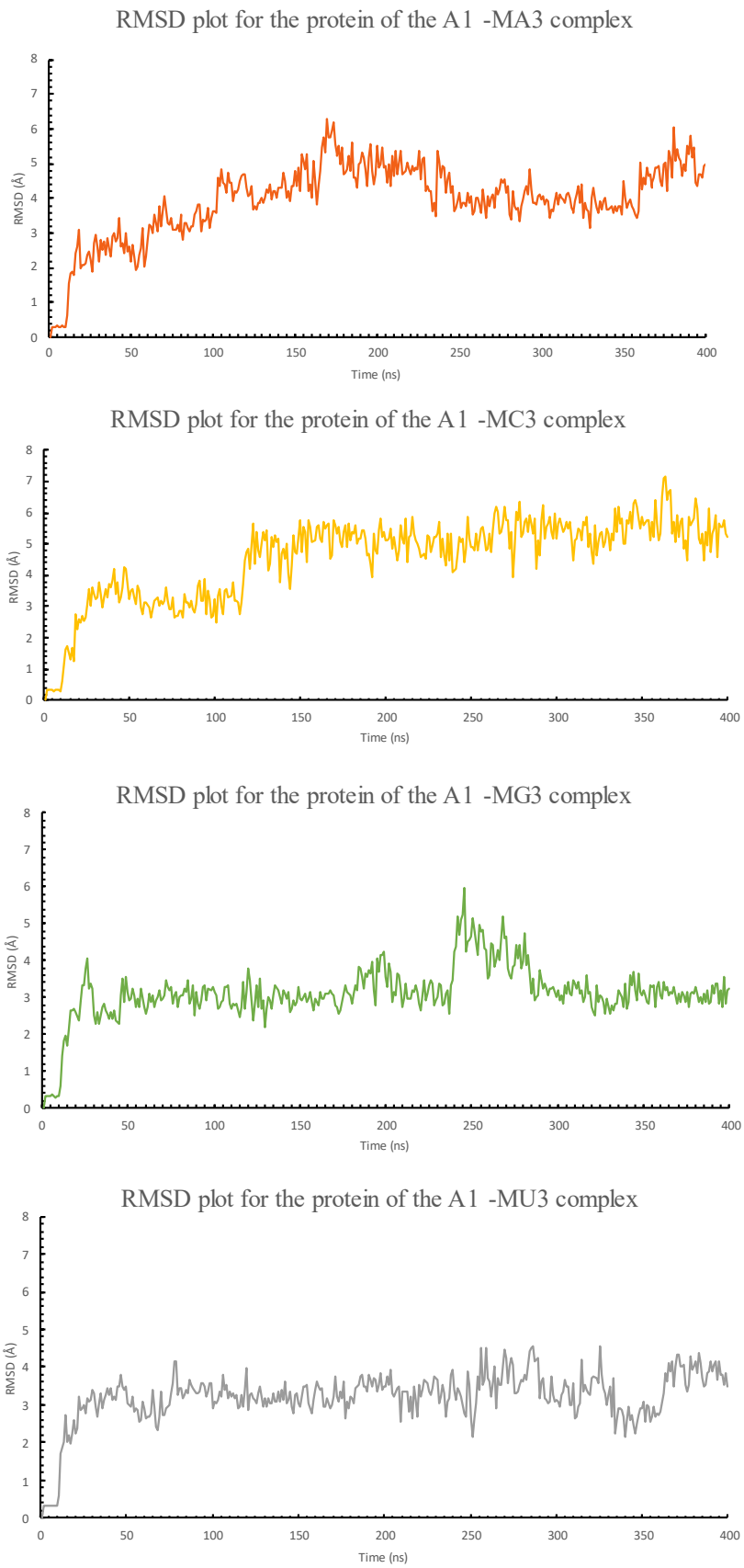


Figure A2: Evolution of the RMSD of the protein component of the RNA-A1 complexes. The stability of the complexes during MD simulation was assessed by plotting the evolution of backbone RMSD of A1 RRMs. As seen in the plots on the left, the protein has low, comparatively unfluctuating RMSD values, indicating the stability of the RRMs during simulation.

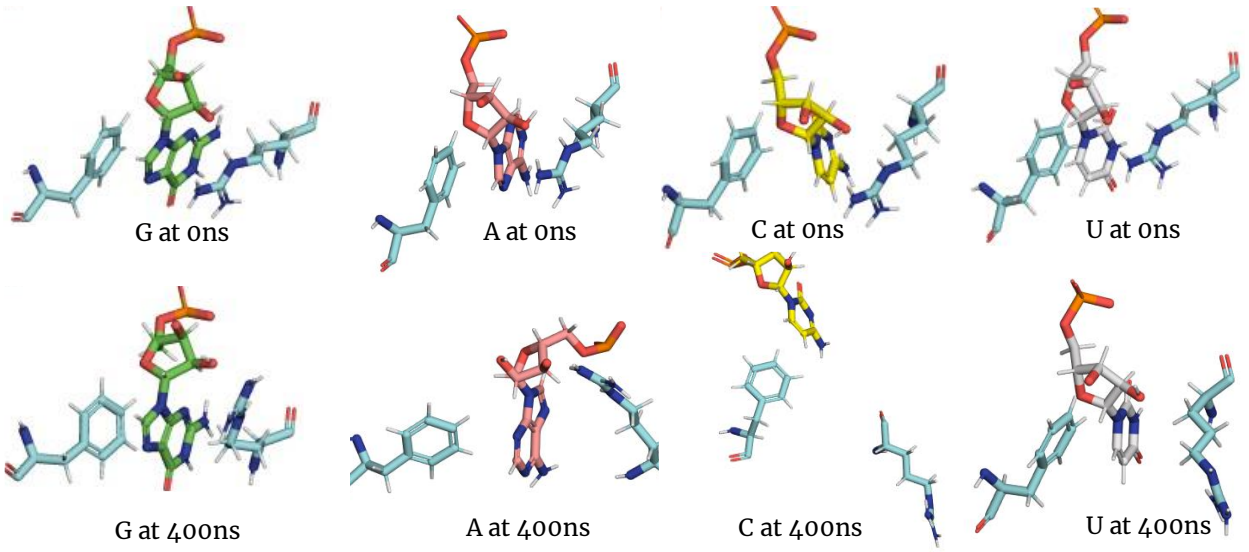


Figure A3: The snapshots of each M*N*3-A1 complex for effective comparison of the starting position and position at 400 ns relative to Arg92 . All nucleotides in position 3 are displayed. This set of data determines Guanine as being most stable in position 2 compared to the other nucleotides that have significantly changes their orientation. Therefore, Guanine seems to provide most stability to the RNA-A1 complex when in position 2. While A and U can still form electrostatic contacts with arg92 using their phosphate backbone atoms, C has drifted significantly.

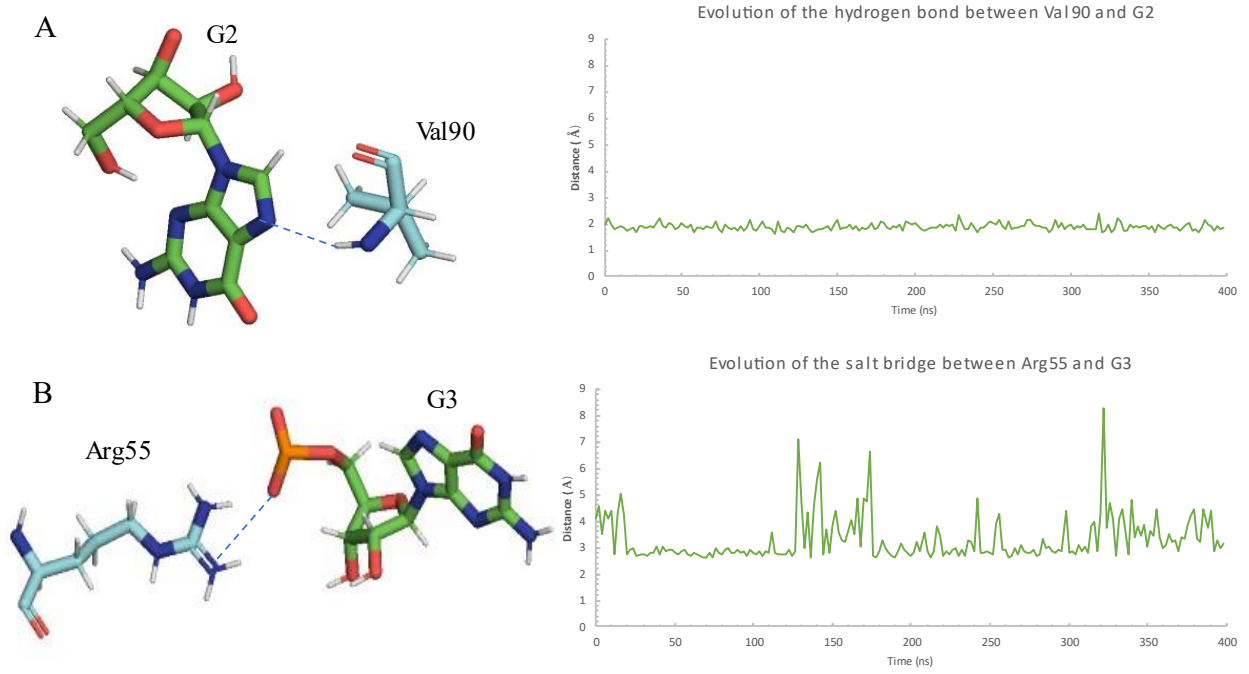


Figure A4: Illustration of the prominent electrostatic interactions in the MG3-A1 complex. (A) A 3D representation of key salt-bridge and hydrogen bond interactions between Val90 and G2 has been illustrated with a plot of the distance (right) and the dotted line (left) indicating the bond forming between the purine ring of Guanine and the backbone amino group of Valine. **(B)** Similarly, the distance evolution between the guanidine group of ARG55 and phosphate group of G3 described that this pair established a salt-bridge, with the distances shown in a plot on the right.

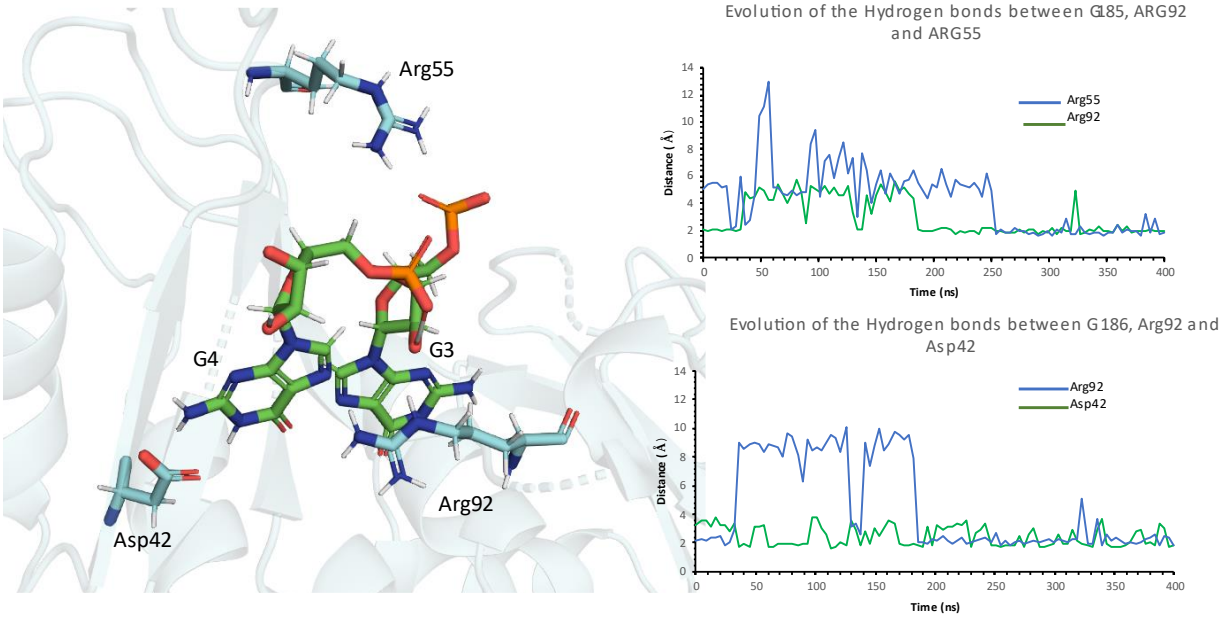


Figure A5: Hydrogen bonds adding to the binding energy for the M34G complexes. Arg55 is able to form a hydrogen bond with the phosphate backbone oxygen atom using it guanidine group. Asp42 is also able to form hydrogen bonds with its negatively charged side chain and the amino group from G4's base. However, Arg92 is able to form two hydrogen bonds: one with the phosphate backbone oxygen atom of G3 and one with G4's base oxygen atom.

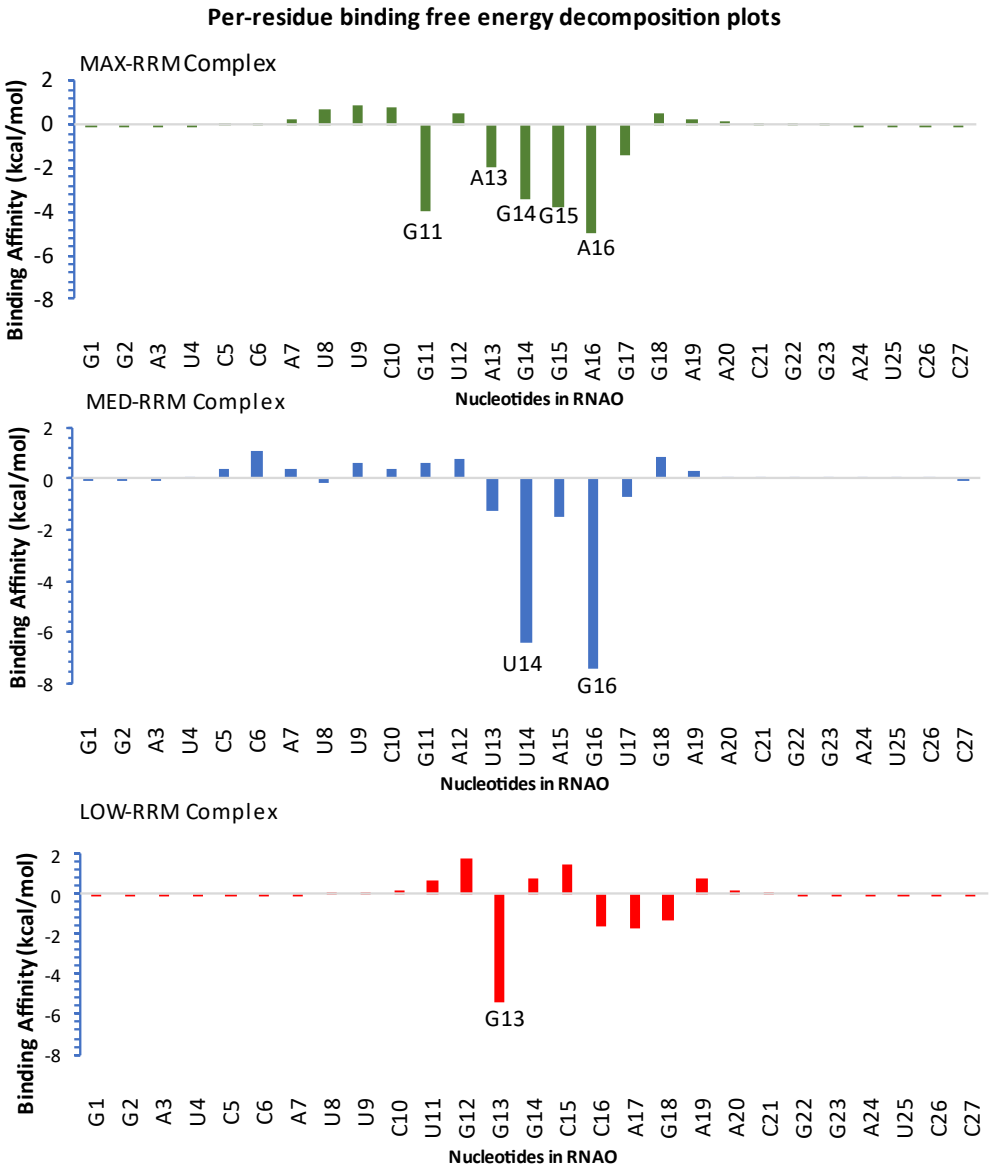


Figure A6: MM-GBSA decomposition analysis for the RNA component of the MAX, MED and LOW complexes. The apical loop (residues 11-17) is mostly responsible for binding affinity as it faces the RNP binding pocket. Of the three complexes, MAX has the highest affinity contribution given the higher number of As and Gs.

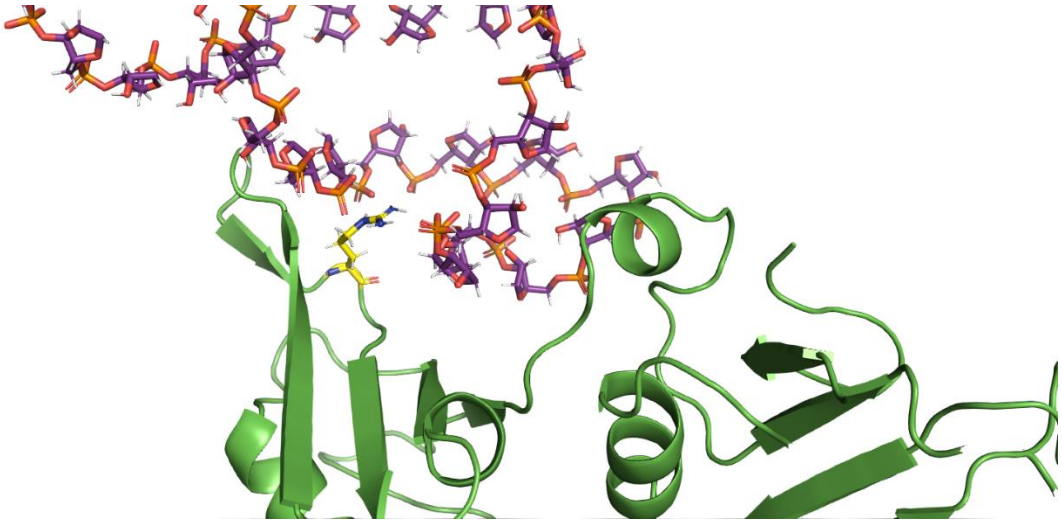
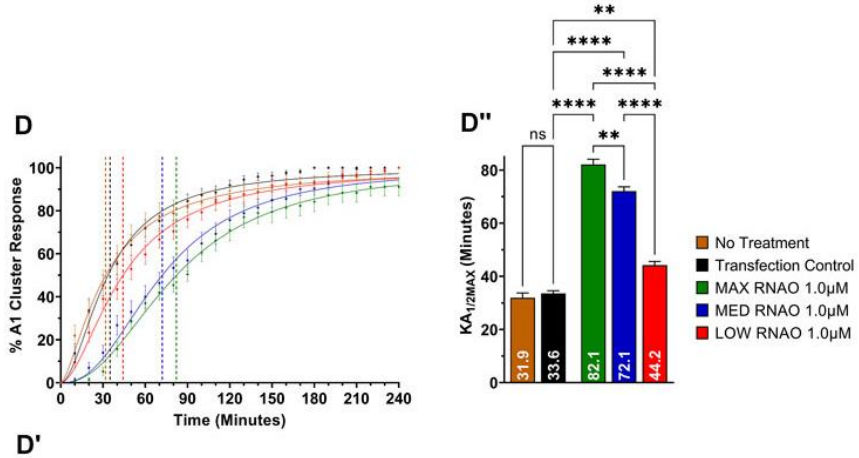


Figure A7: The nonspecific electrostatic contribution by Arg55 in the His101A-MAX complex. Arg55 (stick representation, yellow) interacts with the negatively charged phosphate backbone of the MAX RNAO (bases not shown), thus significantly increasing its binding affinity contribution to the complex. This aids in compensating for the loss of contacts provided by His101.

RNAOs that bind A1 can rescue A1 clustering *in vitro*

(Note: We would like to thank our collaborators, Dr. Michael Levin's team at the University of Saskatchewan, particularly his research fellow, Dr. Joseph Patrick-Clarke for the assays in Figures A8-11.)



D'

Time (Min)	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240
MAX RNAO 1.0μM	ns	ns	****	****	****	****	****	****	****	****	****	***	**	**	**	**	**	**	**	*	*	*	ns	ns	ns
MED RNAO 1.0μM	ns	ns	****	****	****	****	****	****	****	****	***	***	**	**	**	**	**	*	*	*	ns	ns	ns	ns	ns
LOW RNAO 1.0μM	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
P-Values																									

FIGURE A8. OptoA1 clustering is attenuated with the addition of sequence- and structure-specific RNAOs. Representative images of OptoA1 blue light (BL) stimulated cells treated with either **(A)** 1.0 µM MAX RNAO, **(B)** 1.0 µM MED RNAO or **(C)** 1.0 µM LOW RNAO. **(D)** Quantification of A1 cluster formation during a 240-min BL stimulation protocol with the addition of either MAX RNAO (Green), MED RNAO (Blue) or LOW RNAO (Red). Results are plotted as a percent maximum to the highest cluster response at 240 min for each RNA treatment, resulting in a kinetics curve for association dynamics. No Treatment = no treatment with RNA; Transfection Control = cells only transfected with RNAiMAX. Dashed lines indicate $KA_{1/2Max}$. **(D^{I})** Tabular results of a two-way ANOVA, with a Bonferroni post-hoc test from the curves illustrated in **(D)**. **(D^{II})** Bar graphs and one-way ANOVA, with a Tukey post-hoc test analysis of $KA_{1/2Max}$ from the curves illustrated in **(D)**. Data shown are mean ± S.E.M. for three biological replicates. Arrows indicate the formation of OptoA1 clusters. Scale bars = 10 µm *p < 0.05; **p < 0.01; ***p < 0.001; ****p < 0.0001; 95% Confidence Interval.

The optogenetics assay reported by our collaborators in a previous study[158], was performed for this study to use Blue light stimulation to induce clustering of the A1 protein. In this research, A1 cluster response by the A1 protein and the transfection control displayed fast cluster responses. Adding MAX, MED or LOW RNA, reduced A1 clustering with statistical significance between the three where MAX reduced the most clustering followed by MED and then LOW (Figure 3.16). The trend in decreasing clustering by MAX, MED and LOW may be attributed to each type of RNA having different affinity for the protein. To further confirm this, a complementary thermal shift assay was performed by our collaborators.

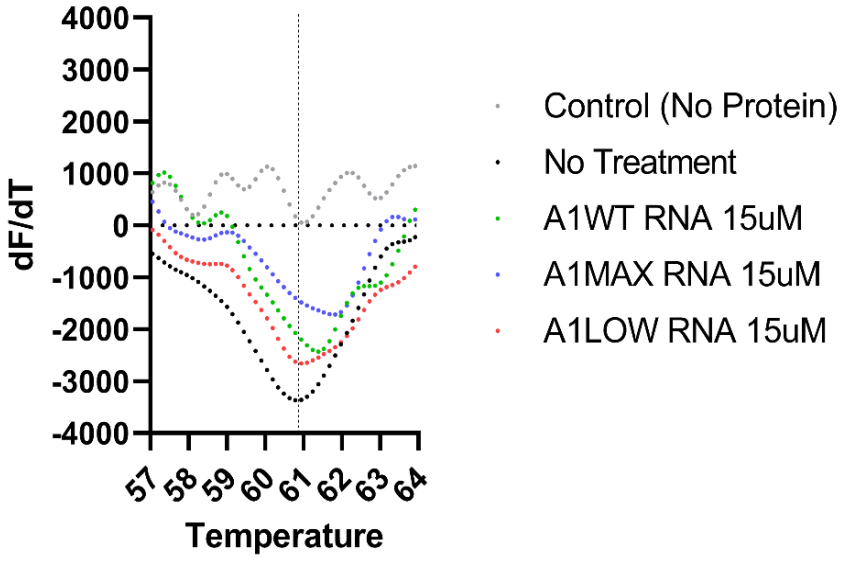


Figure A9: The thermal shift assay for MAX, MED (WT) and LOW. Without any treatment, the A1 has the lowest thermal shift which gradually increases when the 27-nt RNAOs are bound to them. The highest thermal shift can be seen for MAX, followed by MED (WT) and the LOW, at the same concentration for each.

A thermal shift assay measures changes in the thermal denaturation temperature, and hence stability of a protein, under varying conditions such as variations in drug concentration (e.g., RNA type and concentration), buffer pH or ionic strength, redox potential, or sequence mutation.[101] Therefore, binding of MAX, MED and LOW increased the thermal shift and hence the stability as seen in Figure 3.16. While MAX enhanced protein stability the most, MED was next close in value followed by LOW. In fact, since the apical loop regions of MAX, MED and LOW were inspired by literature, the Kd values for RNAOs congaing the apical loop residues for each type of RNA also follows the trend (Table 5). This replication of the trend seen *in silico* with the MD and BFE analysis validates the findings that the multiple AG motifs present in RNA enhances A1's affinity for it and has the potential to rescue protein aggregation.

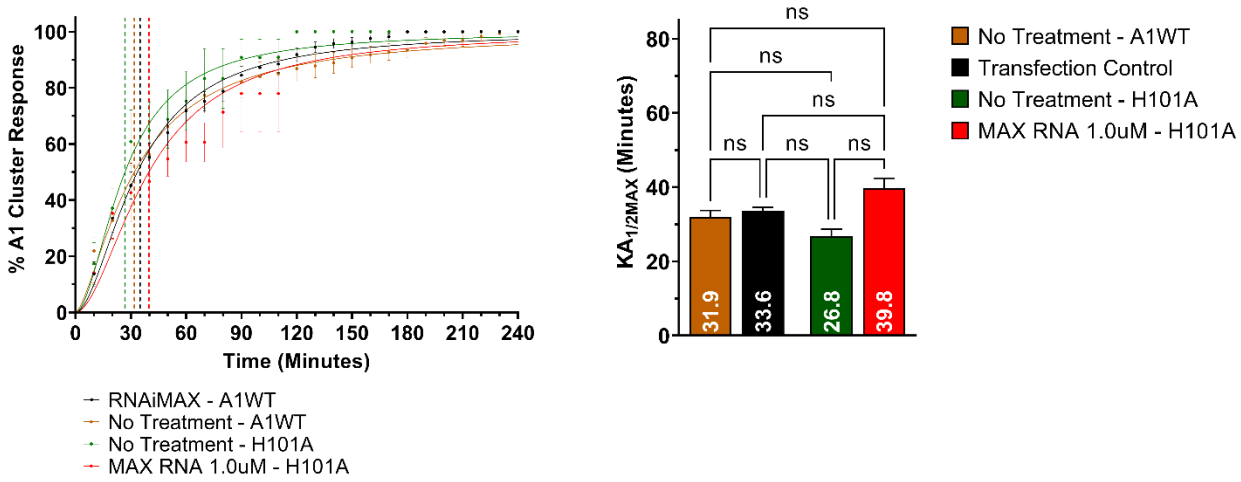


Figure A10: His101A mutation renders MAX-binding ineffective against clustering. While MAX can rescue A1 clustering, it can not do so for the H101A mutant. There is no statistical significance in the clustering response of the untreated A1 and the His101A-MAX complex.

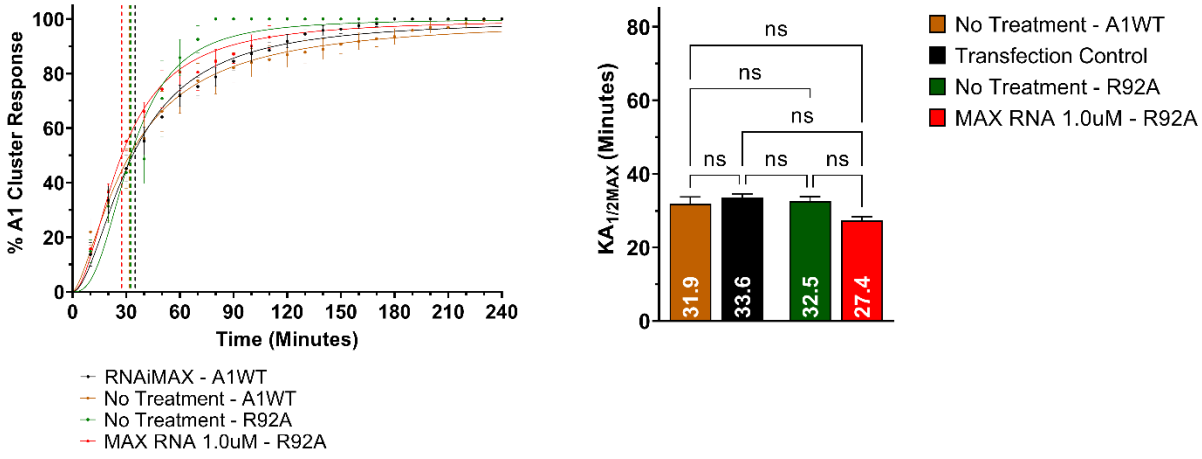


Figure A11: R92AA mutation renders MAX-binding ineffective against clustering. While MAX can rescue A1 clustering, it can not do so for the R92A mutant. There is no statistical significance in the clustering response of the untreated A1 and the R92A-MAX complex.