Atomic-scale structural insights into the functional roles of the full low complexity domain of the hnRNPA1 protein

Irawati Roy,[1, a)] Sangeetha Balasubramanian,[1, a)] Rajeswari Appadurai,[1, a)] and Anand Srivastava[1, b)]

Molecular Biophysics Unit, Indian Institute of Science, Bangalore, Karnataka 560012, India

[a)]Equal contributions: Irawati Roy, Sangeetha Balasubramanian and Rajeswari Appadurai

[b)]Electronic mail: rajeswari@iisc.ac.in (Co-corresponding author), anand@iisc.ac.in (Corresponding author)

Abstract: hnRNPA1, a protein from the heterogeneous-nuclear ribonucleoprotein family, mediates cellular processes such as RNA metabolism and DNA telomere maintenance. Besides the folded RNA recognition motifs, hnRNPA1 has a $\sim$ 135 amino-acids long low-complexity domain (LCD) consisting of RGG-repeats region and a prion-like domain (PrLD). Biochemical data suggest that RGG-rich region modulates recognition of telomeric repeats, while PrLD is often implicated in formation of biomolecular condensates. Here, we utilize recent experimental data from techniques like NMR, DEER and SAXS on hnRNPA1-LCD and integrated that with $\sim$ 100 μ-seconds molecular simulations trajectory data to elucidate the Boltzmann-weighted atomic-resolution conformational ensemble of hnRNP1-LCD and study its interaction with telomeric G-quadruplexes (GQs). Single chain statistics and abundance of molecular motifs as well as consistency with low-resolution experimentally reported structural data suggest faithful recapitulation of local interactions. We observe multi-domain structural architecture for the LCD suggesting that functional modularity may be encoded in very long IDPs. Our binding simulations studies reveal that some conformations from RGG-rich region destabilize telomeric GQ, which may be an important first step in unfolding activities at the telomeric end. Spontaneous appearances of kinked beta sheet motifs within the hnRNPA1-LCD hint at possible nucleation centres of reversible biomolecular condensate formation.

Keywords: hnRNPA1-LCD, RNA and DNA G-quadruplex, telomere maintenance, integrative modeling, multidomain IDPs

INTRODUCTION

Heterogeneous nuclear ribonucleoprotein A1 (hnRNPA1) is a member of a complex and diverse group of proteins called heterogeneous ribonucleoproteins (hnRNPs) that collectively play an important role in processing heterogeneous nuclear RNAs into mature mRNAs. This abundant nuclear protein modulates a number of RNA metabolism processes, including transcription[1–3], splicing[3–6], RNA stability, export[7–9] and translation[10–12]. However, the role of hnRNAP1 is not just limited to mRNA biogenesis but extends to other processes such as microRNA processing and telomere maintenance[13]. Structurally (Fig. 1(a)), the N-terminus of the protein contains two well-folded RNA recognition motifs (RRMs) connected by a very small linker region. Jointly these two motifs constitute Unwinding protein 1(UP1)[14], which mediate interactions with target RNAs. UP1 domain is shown to bind to telomeric ssDNA and also unfold telomeric quadruplexes[13,15–18]. In addition, a flexible Glycine rich C-terminal region consisting of Arg-Gly-Gly tripeptide repeats (RGG), interspersed with aromatic (Phe, Tyr) residues to form RGG boxes, provide both protein and RNA binding capabilities to hnRNPA1[14]. Downstream from the RGG boxes, the C-terminal domain harbours a prion-like domain (PrLD) and a nuclear shuttling sequence called the M9 sequence[19]. PrLD has been shown to induce protein-protein interaction leading to stress granule assembly and pathological protein aggregation[19]. The RGG and PrLD domain together constitutes the disordered low complexity domain (LCD) of hnRNPA1, which is the focus of our study.

Several line of evidence have demonstrated that both UP1 and RGG-box domain are involved in DNA G-quadruplex (GQ) unfolding activity at the telomeric end of DNA[20,21]. It is believed that the RGG-box interacts with the loop nucleotides of the GQ structure and helps in the initial specific recognition and binding of the GQ DNA[20]. As such, RGG-box acts as an adaptor for the UP1 by bringing it closer to the GQ, which leads to an efficient binding of UP1 followed by enhanced destabilisation of the DNA GQ structure[20]. Finally, UP1 forms a high-affinity complex with the single-stranded DNA and stabilises the extended form. Besides DNA, very recent studies have also reported structure-dependent recognition of RNA GQ by the RGG-box of hnRNPA1, where the long non-coding RNA called the telomeric repeats containing RNA (also known as telomere RNA or TERRA) forms GQ structures and has regulatory functions over telomere replication. It was shown

that the RGG-box of hnRNPA1 binds to the loops containing TERRA GQ but not with the single-stranded RNA, which in turn provides the specificity to the UP1 domain to bind and enhances unfolding of the intramolecular TERRA GQ[22]. Despite tremendous strides made in the mechanistic understanding of hnRNPA1 interactions with GQ using biochemical studies, high-resolution structural and dynamics insights into telomere maintenance is still missing. In our work, we utilize the integrative molecular modeling approach that combines atomic-resolution molecular simulations with available low-resolution ensemble averaged data from various experiments such SAXS, DEER, and NMR to faithfully reconstruct the atomic-resolution ensemble of 3D structures for hnRNPA1-LCD and use the structural data to investigate functional role of LCD in telomere maintenance.

Under stress and pathological conditions, the prion-like hnRNPA1-LCD also has a propensity to phase-separate into amyloid-like fibrils comprising cross-β structure[23]. This property has been proposed to mediate stress granule assembly, which are membrane-less cytosolic bodies containing mRNAs and proteins[24]. This irreversible stress granule formation is central to various pathogenesis such as multisystem proteinopathy (MSP) and amyotrophic lateral sclerosis (ALS)[19]. Elucidation of atomic-resolution structural ensemble of hnRNPA1-LCD could shed important insights into the molecular (and often multivalent and weak) driving forces leading to these assemblies. In the last few years, tremendous progress has been made in the structural and biophysical characterization of hnRNPA1-LCD. The recent availability of experimental data, though often incomplete and ensemble averaged in nature, from multiple techniques such as NMR, SAXS and DEER[22,25–27] provides us with a tremendous opportunity to apply an integrative approach and faithfully reconstruct the ensemble of 3D structures. Along with the experimental techniques, ensemble conformation data is also available from molecular simulation studies on hnRNPA1 system[25,26,28,29] that can further be integrated to elucidate the high-resolution conformational ensemble of hnRNPA1-LCD consistent with the available ensemble averaged low-resolution experimental data.

In this work we have focused on and explored the structural heterogeneity and associated biological activities of hnRNPA1 with the help of our advanced-sampling[30] and machine-learning techniques[31] and followed that up with large scale biomolecular simulations and analyses of hnRNPA1-LCD and its RNA/DNA complexes. In the Results and Discussion section below, we report experimentally-consistent high-resolution conformational ensemble

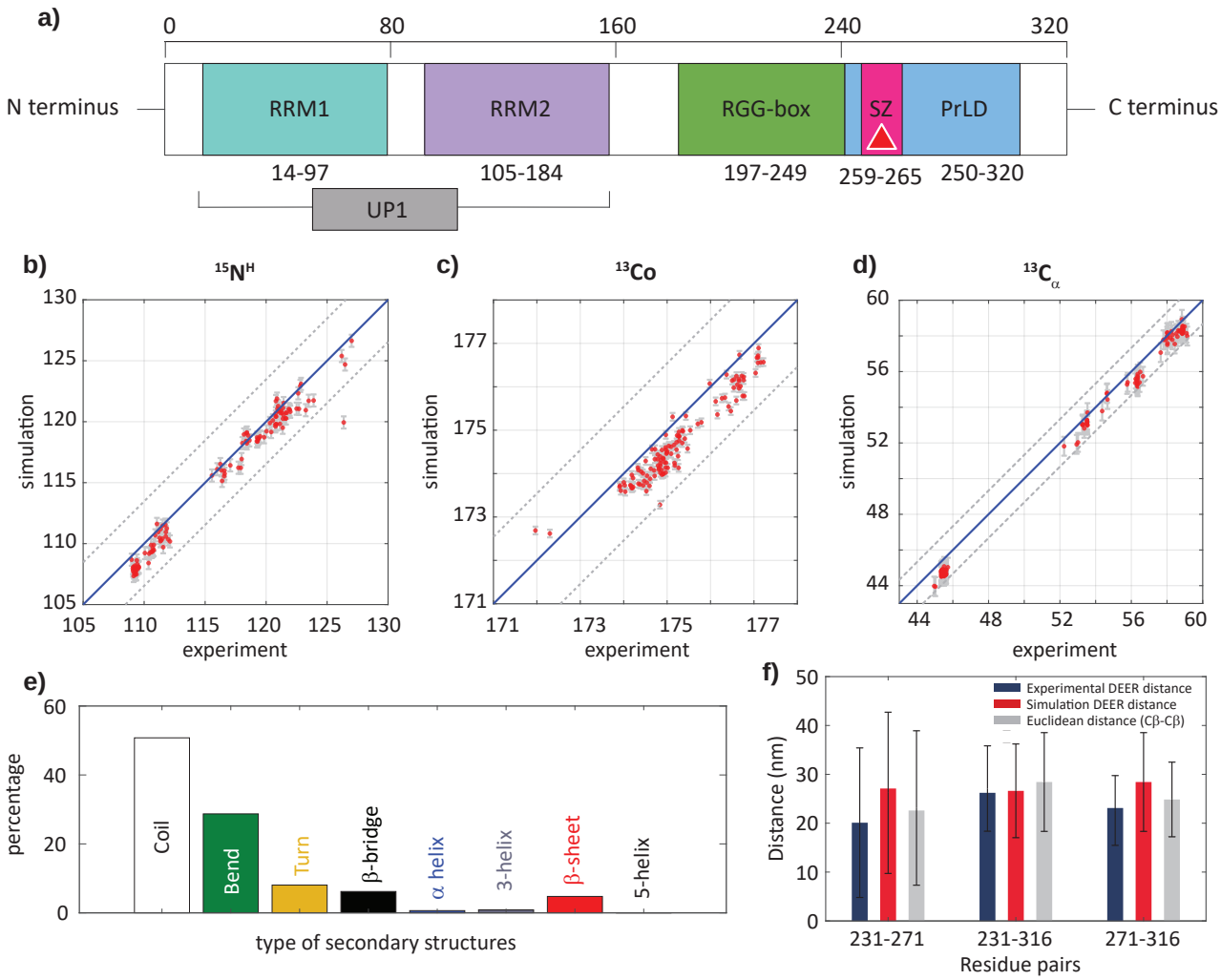

FIG. 1: a) Domain architecture of hnRNPA1. The Teal and Purple coloured domains are the N terminal ordered RNA recognition motifs. Together these two domains are called UP1. The steric zipper motif (259-265) is colored in Magenta. The Green and the Blue domains are intrinsically disordered RGG-box domain and PrLD respectively. The steric zipper region is deleted in our construct. b), c), d) are the $^{15}N^{H}$, $^{1}3CO$ and $^{13}C^{\alpha}$ chemical shift plots for Δhex-LCD respectively. The dashed lines are indicative of the acceptance region observed by SPARTA+. e) Percentages of secondary structures observed in Δhex-LCD trajectory. f) Comparison of experimentally observed DEER distances (Blue), computationally calculated DEER distances (Red) and Euclidean distances between Cβ-Cβ atoms (Gray) of pair 231-271, 231-316 and 271-316.

of hnRNPA1-LCD. Our analyses of the hnRNPA1-LCD ensemble data reveal an extensive residue-wise interaction network with high propensity to form short-lived secondary structures. Another interesting observation that we discuss below is the spontaneous but transient appearance of kinked beta sheets of different types at diverse locations within hnRNPA1-

LCD, reminiscent of the now well-described low-complexity aromatic-rich kinked segments (LARKS)[32] and reversible amyloid core (RACS)[33] motifs. We also discover that the disordered hnRNPA-1 LCD express structural modularity, a fascinating phenomenon in long IDPs, allowing versatile functional boundaries within a single long polypeptide IDP chain. Given the modularity in LCD and available biochemical data about the role of RGG-region in modulating telomere GQ recognition, we carried out extensive simulations of RGG-region interactions with RNA and DNA GQs. Our simulations provide molecular-level insights into the reported role of RGG-region in telomere maintenance. We discuss in detail how the RGG-region binds differently to telomeric DNA and telomeric RNA GQs and we also demonstrate how a very small population of hnRNPA1-LCD conformers are capable of destabilizing the GQs. After the Results section, we briefly summarize our observations and finding in the Conclusions section, which is followed by the Materials and Methods where we provides details of all the methods that we have used. Additional information is included in the Supporting Information (SI) file. All our data including all trajectories and input files for our simulations, all analyses related data and codes are publicly available online (GitHub link: codesrivastavalab/hnRNPA1-LCD).

I. RESULTS AND DISCUSSION

A. Atomic resolution structural insights into the hnRNPA1-LCD conformation ensemble

1. Conformations consistent with scattering and spectroscopy data reveal extensive interaction network between residues of hnRNPA1-LCD

To gain conformational insights into the functionally important intrinsically disordered region of hnRNPA1, we simulated full low complexity domain employing a Hamiltonion and temperature scaled hybrid replica exchange technique called Replica Exchange with Hybrid Tempering (REHT) to generate the full ensemble of heterogeneous structures[30]. We found inhomogeneous diffusion of replicas, most likely due to water aided strong intermolecular interactions by signature steric zipper motif and insufficient highest effective temperature (Fig. S1(a) in SI). Therefore, we created a construct of hnRNPA1 LCD domain with ag-

gregation promoting steric zipper motif (259-264) deleted (hereafter referred as Δhex-LCD) and performed replica exchange with hybrid tempering on it with proper diffusion of replicas across all temperature in the given range (Fig. S1(b) in SI). Please refer to Material and Methods sections and SI for more details. The construct is exactly the same as used by Tanja Mittag and co-workers[25] in their recent experimental work and we have used the SAXS and NMR Chemical Shifts data from that work to validate our simulation data. A comparison of $^{15}N^{H}$, $^{13}C^{\alpha}$ and ^{13}CO chemical shift data between our simulation and experiment showed a considerable match[25] as shown in Fig. 1(b-d). In Fig. S2, we show the residue-wise superposition of experimentally-available and calculated chemical shift data from conformation ensemble generated from our work. Upon performing a basic secondary structure analysis across the time evolution, we observed that the IDR has a significant propensity to form transient secondary structures, preferentially beta sheet and extended beta structures as shown in Fig.1(e). We show the secondary structure information for each residue as a function of the 1 microsecond simulation time in Fig. S3. These observations indicate an overall inclination of the hnRNPA1-LCD to form local contacts. Concomitantly, when we performed small angle X-ray scattering (SAXS) analysis of Δhex-LCD ensemble (Fig. S4(a)), it showed an excellent match with experimental data in the high scattering vector (q) region indicating that we are able to capture the higher local molecular interaction with a high-degree of reliability.

To understand the chain behaviour of this IDR more, we carried out single chain statistical analyses. We took 16 pairs of residues of varying segment length from the Δhex-LCD construct with one end fixed at residue number 197 and calculated their distances to observe the relationship between the distance distributions and primary sequence separation(ΔN). Unlike that of a random polymer chain, the average distance ($< r >$) of the distributions followed a non-monotonic pattern with respect to ΔN (Fig. S4(b)). Even for a large ΔN, the $< r >$ including the standard deviation were similar enough to indicate a compact behaviour of the protein chain. This is likely a result of extensive interactions among sequence-wise apart residues. Our observation corroborates the observations that were established from spin-label distance distrubtion data on hnRNPA1-LCD from Gunnar Jeschke and co-workers[27,34]. Specifically, in their DEER spectroscopy experiments, Jeschke and co-workers chose to spin-label three pairs of residues (231-271, 271-316 and 231-316) and

reported the ensemble-averaged distance between these residues. We compared our data to these experiments by computationally spin labelling the same sites with the help of DEER-PREdict package[35] and the results are shown in Fig. 1(f). It's evident from the data that the distance distributions statistically belong in the same standard deviation range with that of experimental DEER data. These series of comparisons leads us to believe that we are able to capture the local intramolecular interactions of the Δhex-LCD domain successfully.

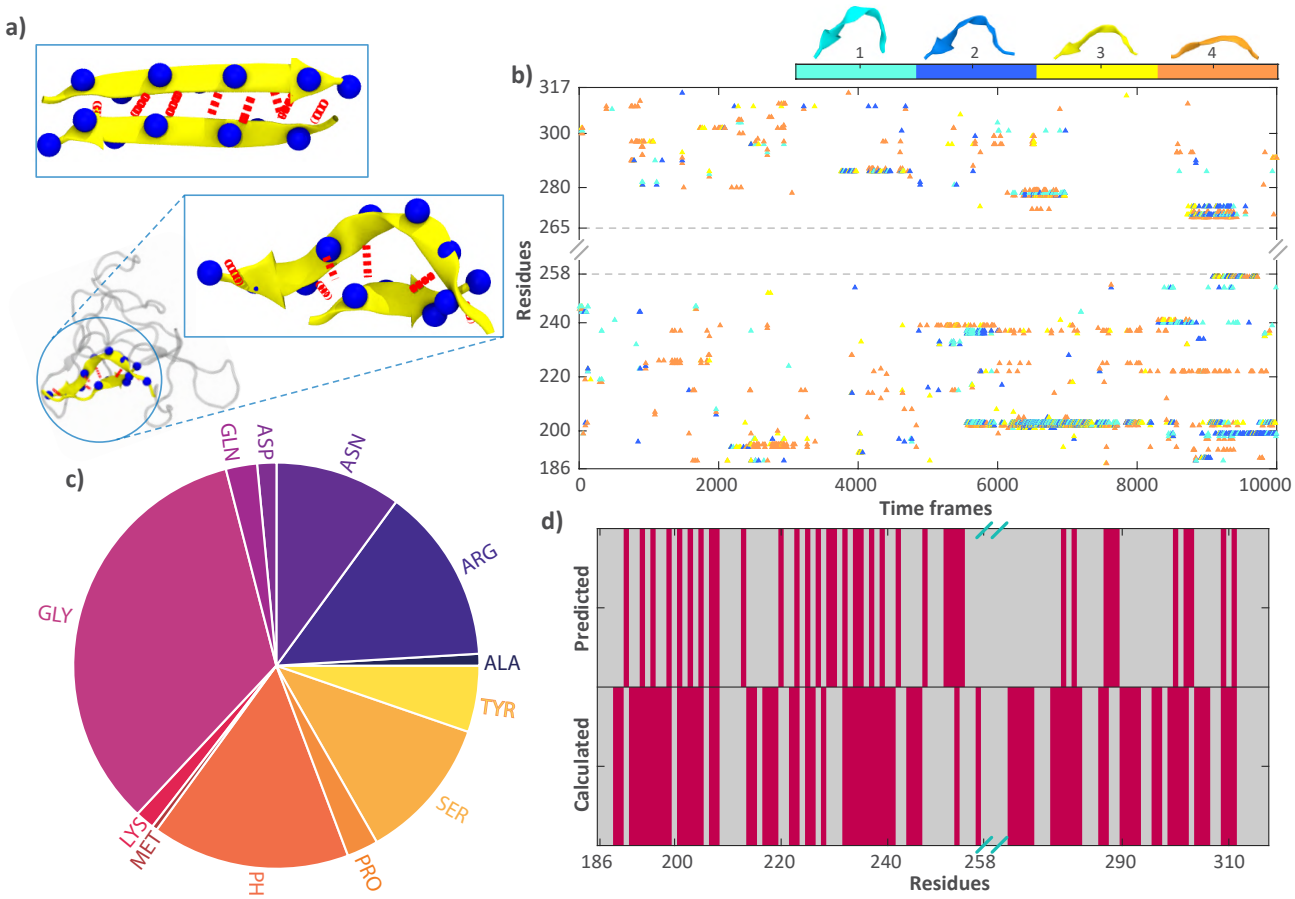

FIG. 2: a) Schematic presentation of kinked-beta sheet and regular beta sheet and corresponding example conformation. The Blue spheres are the consecutive $c\alpha$ dihedral angles and the red dotted lines are the hydrogen bonds. In the kinked beta-sheet one designated hydrogen bond is missing. b) LARKS appearance profile. The $c\alpha$ dihedral angles of Δhex-LCD are plotted over the course of 1 microsecond. The captured LARKS formed by residues are color coded differently based on the acuteness of the $c\alpha$ dihedral angles and corresponding schematics of the angle ranges are given . Cyan: -20 to 20; Blue: -40 to -20 and 20 to 40; Yellow: -60 to -40 and 40 to 60; Orange: -90 to -60 and 60 to 90. c) Pie chart of residue variety and content in the kinked beta sheet. d) Comparison between the LARKS calculated from our simulation of Δhex-LCD region (with a 0.002 percent cut-off) and the LARKS obtained from LARKS database. Red tiles indicate presence of LARKS-like motif. Grey tiles indicate the absence of the same.

2. hnRNPA1-LCD is very rich in low complexity aromatic-rich kinked segments (LARKS) motifs

The intermolecular interactions in protein conformations are responsible for formation of different structural motifs ranging from regular secondary structures to special signature motifs that have important functional roles. Low complexity domain of various intrinsically disordered proteins including FUS[36–38] and TDP43[39–43] are known to take part in stress granule assembly. Studies on small peptide segments of LCDs often reported reversible amyloid formation with the help of specific motifs rich in Glycine, Serine and aromatic residues. David Eisenberg and co-workers group named these key motifs as low complexity aromatic-rich kinked segments (LARKS) as they observed kinks at specific residues of this motifs[32]. Cong Liu's group also observed similar kinks in short peptide based reversible amyloid core (RACs) studies of hnRNPA1[33]. In a separate study, Espinosa and co-workers have shown increased condensate viscosity and moderately high droplet density upon re-organisation of disordered peptide to ordered inter-protein LARKS-sheet accumulation by transforming weak and transient contacts into comparatively permanent contacts[44].

A microscopic view into our Δhex-LCD ensemble showed us the existence of multiple kinks in beta sheets throughout the entire length of Δhex-LCD sequence. These kinked beta-sheets appeared in a transient manner for the full stretch of simulation. The residues involved are mostly aromatic residues, Glycine, Serine, Arginine and Asparagine. Eisenberg's group has sequentially predicted this kinked beta-sheets alias LARKS using Rosetta energy score. We wanted to structurally quantify these unique secondary structures. In order to do this, we used consecutive Cα atoms dihedral angle that are forming beta sheet as our characterizing marker and calculated the angles for the backbone of full sequence to capture all the acute beta-kink angles (Fig. 2(a,b)). The accepted angle range of (0°-90°) is captured and color coded in 4 different categories for ease of understanding. The sporadic reappearance and disappearance of the beta-kinks (hereafter we will refer these as LARKS) emphasizes the transient nature of motif in our ensemble. Fig. 2(c) shows the relative distribution of residue in the kinked beta sheet. The sequence-wise profile obtained from LARKSdb predicted the probability for residue to form LARKS-like motifs. For each of the predictions, we found respective kinked-beta sheet structures in our Δhex-LCD,

RGG and PrLD simulations as shown in Fig. S5 in SI. Our kinked beta sheet analysis on the ensemble-derived structures showed more number of residues forming LARKs-like motif compared to the LARKdb prediction. In Fig. 2(d), we compare our calculations of beta-kinks to the LARKS profile obtained from LARKSdb[45]. The figure indicated that we found more number of LARKs-like motif in our simulation compared to the prediction by LARKSdb. These regions can be thought of as future focuses in the research of nucleation points or fibril formation promoting regions in the disordered region of hnRNPA1.

3. Disordered region of hnRNPA1 shows structural modularity reminiscent of multi-domain folded proteins

In our pursuit to investigate the local contacts, we performed an all-to-all residue wise distance analysis among each amino acid of our Δhex-LCD. The distance analyses revealed an increased proximity among residues in an intra-domain manner and a distinct separation within the two domains (dumbbell shape indicating two separate domains). This observation reflects the modular nature of the conformation ensemble. In order to observe the entirety of the ensemble in more tangible way, we classified our conformational landscape using a recently developed IDP-clustering tool that uses dimensionality reduction technique called t-distributed stochastic neighbour embedding technique (t-SNE)[31]. We mapped and superposed the relevant conformations to each of the 16 clusters as shown in Fig. 3. Each superposition indicates the homogeneous nature of every individual clusters. To show that the modularity is consistent as a function of time, we also calculated the modularity score (see Methods section) and plotted that as a function of time for the duration of our simulation time. The system is consistently modular over time as shown in Fig. S6.

These evidence of structural modularity of RGG-box domain and PrLD gave us the motivation to do separate simulations of these two domains. Besides the full LCD system, we also set up replica exchange systems for the RGG-box and PrLD system (see Table I below for system details). Our REHT simulations on RGG-box as well as on the PrLD showed exhaustive sampling as revealed by the replica diffusion map for the two systems (Fig. S7(a)

for RGG-box and Fig. S7(b) for the PrLD system). After carrying out individual REHT simulation on these two domains, we again carried out the chemical shift calculations using SPARTA+. The results of both RGG-box domain and PrLD showed excellent matches with the experimental data collected from the recently published work of Mittag and co-worker[25]. An important point to be noted here is that the for the simulations of complete Δhex-LCD domain (where RGG-box domain and PRLD are together) and the simulation of these two

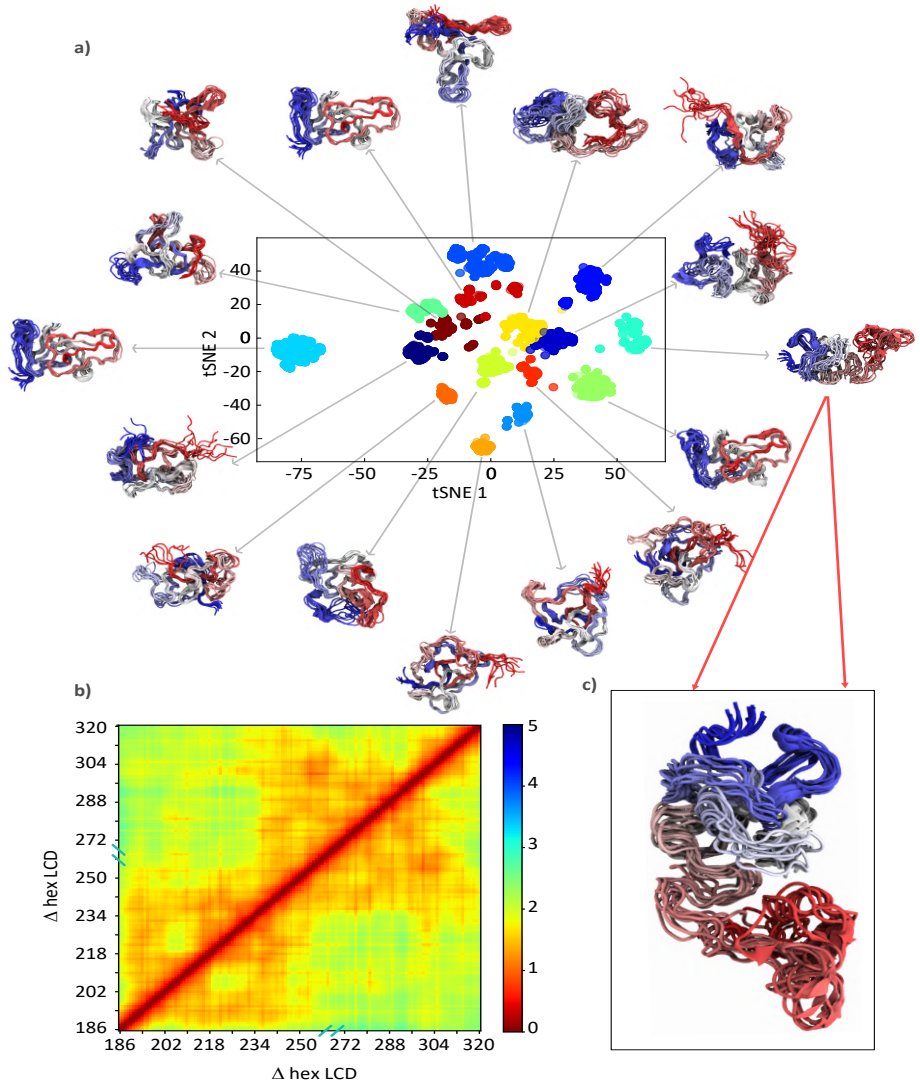

FIG. 3: a) Clustering of Δhex-LCD ensemble. 16 clusters are differently colour coded. 10 conformations collected from the cluster-centers are mapped and superposed to every clusters. b) All-to-all distance analysis plot showing inter-residue distance map of Δhex-LCD. Red to Blue are distances that are color coded from large distances to smaller distances, respectively. c) Zoomed-in modular Δhex-LCD structure.

domains individually, all three of these simulation data showed considerable match with the experimental data (Fig. S8 and Fig. S9), again proving the point that these two domains are modular. This is very interesting because long LCDs are generally not assumed to have distinct domains. Our data suggests that some long IDPs may have evolved with design features that reflect multi-domain architecture to have functional modularity across the LCD.

B. Functional insights from hnRNPA1-LCD conformation ensemble: Role of RGG box in telomere maintenance

The structural and functional modularity motivated us to dig deeper into the individual domain functionality in greater details. Our next objective was to look into the functional relevance of RGG box domain in telomere maintenance. hnRNPA1 expresses a concentration dependent binding with telomeric RNA or DNA, particularly it binds to the GQ motif. Studies have hypothesised that the RGG-box binds to GQ, recruits RRM1 and RRM2 which in turn destabilises the GQ[20,22]. We explore this hypothesis to study RGG-GQ interactions and the first step is to choose the reference structure owing to the intrinsic disorderedness of IDP. The machine learning based dimensionality reduction technique of t-distributed stochastic neighbour embedding technique (t-SNE) - a recently developed tool in our group - helped us to cluster the huge heterogeneous ensemble into 50 groups[31] (see Fig. S10 in SI). The conformations of RGG obtained from the clustering of REHT trajectories represents the different states attained by RGG in solution. Being an IDP, the RGG might interact with other partners via several mechanisms like "folding upon binding" or fuzzy interactions[46]. Though these are dynamic reorientations, docking studies might provide an insight into the formation of "encounter complexes" that might ultimately lead to such re-orientations in solution. In order to identify these encounter complexes and the probable basis for GQ recognition by RGG, we performed docking studies of RGG with telomeric DNA as well as RNA quadruplexes. The docked structures were then subjected to further molecular simulations. A flowchart depicting the protocol adopted to generate RGG-GQ complexes is shown in Fig. 4.

1. RGG box prefers "groove binding" geometry with telomeric GQ DNA and "end stacking" bnding mode with telomeric GQ RNA

We docked each of the 50 conformations of RGGs with 7 different unimolecular, bimolecular and tetramolecular GQs of telomeric RNA (TERRA6, TERRA12 and TERRA24) and

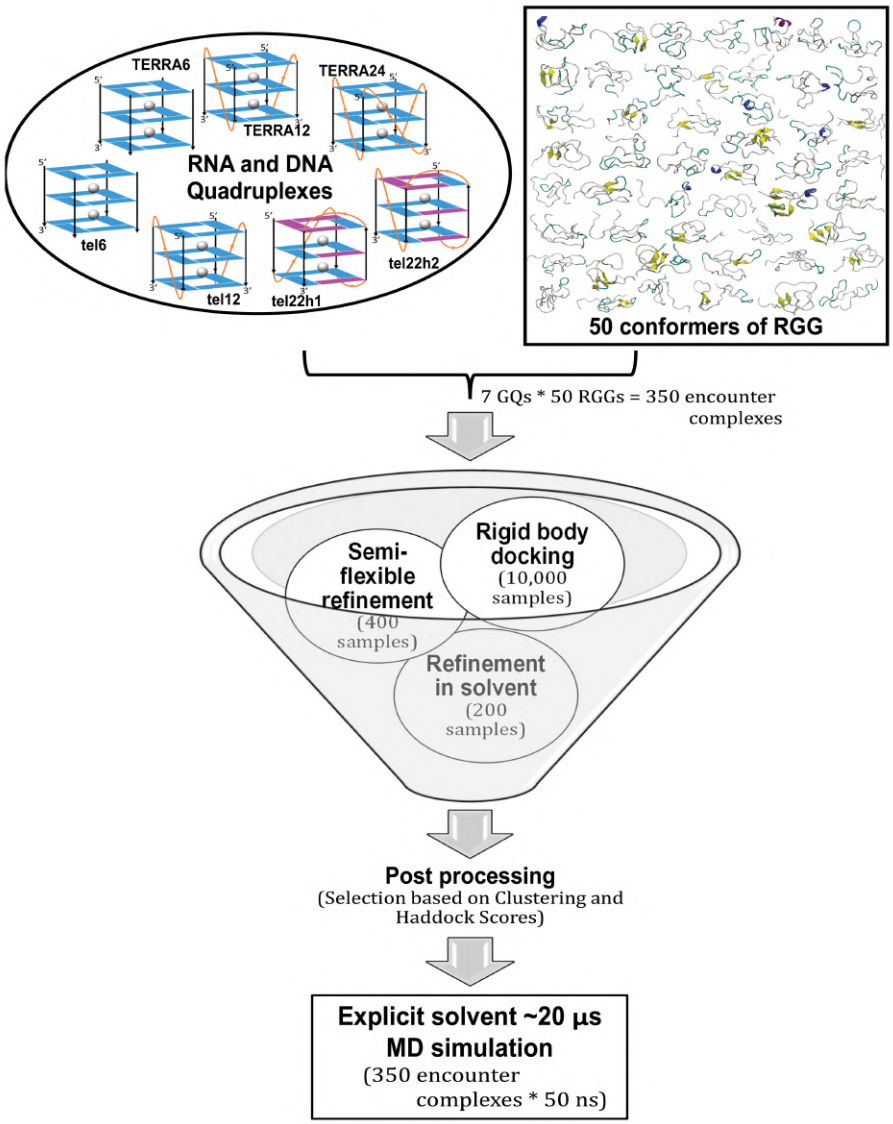

FIG. 4: Flowchart of the protocol adopted to generate RGG-GQ complexes. Seven different telomeric DNA and RNA GQ structures chosen from literature were docked with each of the 50 conformers of RGG obtained by t-sne clustering. HADDOCK protocol with center of mass distance criteria was used to reduce the sample space from 10,000 initial random geometries for each complex. The 350 (7 GQs with 50 RGGs) encounter complexes were further simulated for 50 ns each resulting in 20microseconds ensemble to explore the mode of binding, stability and conformational dynamics of RGG-GQ binding process.

DNA (tel6, tel12, tel22h1 and tel22h2). The systems under consideration are listed in Table II. The sites most preferred by the 50 RGG conformations to bind with each of the 7 GQs are depicted in the occupancy map plotted on a three-dimensional isosurface mimicking the first solvation shell (3.5 Å) shown in Fig. 5. We observed that the mode of interaction of RGG with RNA and DNA GQ varies widely. A majority of the docked RGGs are interacting over the external G-quartets of TERRA in the "end stacking" mode, whereas in case of telomeric DNA, the RGG is dominant in the "groove binding" mode interacting with the grooves and loops of GQs.

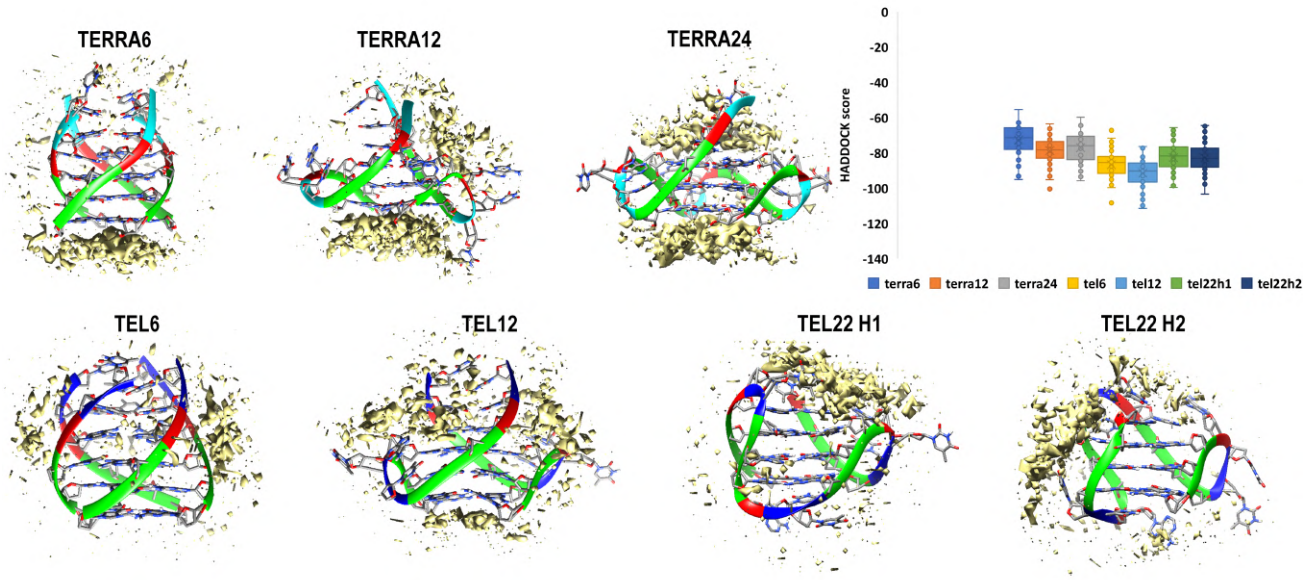

FIG. 5: Occupancy map depicting the extent of RGG-GQ interaction and the high density of interactions at specific sites on GQ are shown as an isosurface mimicking the first solvation shell (Yellow surface). The range of HADDOCK scores obtained for the docking of RGG with the seven GQ structures. Boxplots indicate minimum, median, maximum, and upper and lower quartiles. The GQ backbone is shown as ribbons while the bases are displayed as sticks. The bases are colored as Green: Guanine, Red: Adenine, Blue: Thymine and Cyan: Uracil.

2. Both DNA and RNA GQ with loops and overhangs have stronger binding consistent with recent biochemical observations

The variation in the binding mode is also reflected in the range of HADDOCK scores sampled by the different quadruplexes (Fig. 5). The complexes in end stacking modes namely

TERRA6, TERRA12 and TERRA24 show higher HADDOCK scores in the range of -65 to -85. The tel6 complexes have majority of groove binding interaction and their HADDOCK scores are slightly better (-80 to -90) than TERRA complexes. The tel12 complexes show a mixed mode of interaction that includes both end stacking and groove binding, and interestingly, these complexes possess the best HADDOCK scores (-85 to -100). Another noteworthy observation is that the two hybrid forms of tel22 with their groove binding modes, have a very similar range of HADDOCK scores (-75 to -90). It is previously known that RGG primarily interacts with GQ loops[20,22], and accordingly we observe that the binding with tetramolecular structures without loops is weaker than bi- or unimolecular structures with propeller loops.

3. Encounter complexes identify key interacting residues of RGG with DNA and RNA GQs

The HADDOCK scores of each of the docked complex with all 7 GQs are shown in Fig. S11 in SI. Among these, we analyzed the best three complexes with good HADDOCK scores for their inter-molecular interactions like their hydrogen bonds, electrostatic as well as *pi*-stacking interactions (see Fig. 6). In all these complexes, at the very least two of the four Arg residues from the four RGG motifs are involved in either Hydrogen bonding or electrostatic interactions. Apart from these, several of the Gly residues show hydrogen bonding interactions, while the Phe residues form *pi*-stacking interactions. Interestingly, in the telomeric DNA complexes, hydrogen bonding interactions with Ser residues are higher than RNA complexes. The residue Tyr244 is identified as an important interacting residue by NMR experiments[20,22,25], and accordingly, the complexes of TERRA6 and tel22h2 express *pi*-stacking interactions with Tyr244. In addition, the complexes of tel12 also express *pi*-stacking and hydrogen bonding interactions with Tyr237. The residues in RGG like Ser197, Ser231, Arg232, Gly233, Tyr244, N245 and Gly246 were identified previously by NMR studies to show differences in their chemical shifts upon binding with GQ, however, unambiguous assignment of several of the Arg and Gly chemical shifts was difficult due to their large number in the sequence. In our docking study, we identify several of these residues to

interact with the RNA and DNA GQs.

The docked complexes are encounter complexes that can modify their interactions under physiological conditions. Hence, we simulated all docked complexes, irrespective of their docking scores, for a period of 50 ns each (a total of 17.5 μ-seconds of all-atom simulations) to monitor their structural stability. We also simulated the GQ structures in their unbound state in order to ensure that the structures are stable in their native conformation. All seven GQ structures of telomeric RNA and DNA were stable for a period of 300ns as seen by the time evolution of their G-quartet hydrogen bonds (Fig. S12 in SI). Hence, any changes in stability of the RGG bound structures must be accounted as the effect of RGG binding. During the simulation of RGG bound complexes, we observed that the RGG structure remains dynamic and expresses conformational changes within RGG as well as its binding with GQ. Considering the stability of apo GQ structures and the dynamic behavior of RGG, if a 50 ns simulation time induces any local disruptions in the GQ structure, it can be interpreted as a direct effect of RGG binding. The average number of hydrogen bonds between RGG and GQ as well as within the G-quartets were calculated in all the simulated complexes and a scatter plot correlating these is shown in Fig. 7. A majority of the GQs possess 22 $\pm$ 2 hydrogen bonds and hence, all GQs are stable with very few exceptions, which we discuss separately in the next section.

4. A minor fraction of RGG ensemble can destabilize both RNA and DNA GQs albeit with different molecular mechanism:

Loss of even 2 hydrogen bonds in each of the three G-quartets would elicit severe destabilization in the GQ structure. Therefore, we used an arbitrary cut-off of 15 hydrogen bonds or less (of the total 24) within the G-quartets to consider it as unstable. Fig. 7 shows that at least one among each of the TERRA complexes are unstable. Similarly, in case of the telomeric DNA (Fig. 7), three out of the 50 tel6 and one of the tel22h2 complexes were unstable. Snapshots of these complex structures after the 50 ns molecular dynamics (MD) refinement is also shown. In both the tetramolecular complexes (TERRA6 and tel6), the RGG is a groove binder, while in TERRA12 and TERRA24, the RGG is an end stacker.

As observed previously by NMR experiments, the primary binding sites of RGG are the

sugar and phosphate backbones of the bases in GQ loops. The interaction pattern of each amino acid in RGG with each of the nucleobase in GQ was extracted to reveal the different binding modes in the RGG-GQ complexes (Fig. 6). When the interactions with the external Guanine bases are higher, the RGG is considered to express end stacking mode. Similarly, the loop binding mode of RGG is described by the involvement of TTA loops. Simultaneous contacts of RGG with TTA loops and/or the backbone of Guanines can be considered as the groove binding mode. Among these three modes of binding, end stacking mode can be clearly distinguished while the other two modes are overlapping. In our studied systems, the end stacking mode of TERRA12, TERRA24 and tel12 is clearly shown in Fig. 6 by the higher number of interactions with the external Guanine bases. Similarly, the groove or loop binding mode in TERRA6, tel22h1 and tel22h2 is highlighted by the interactions with the Guanine backbone as well as the loop bases (Fig. 6). Deeper analysis of the all-atom trajectories

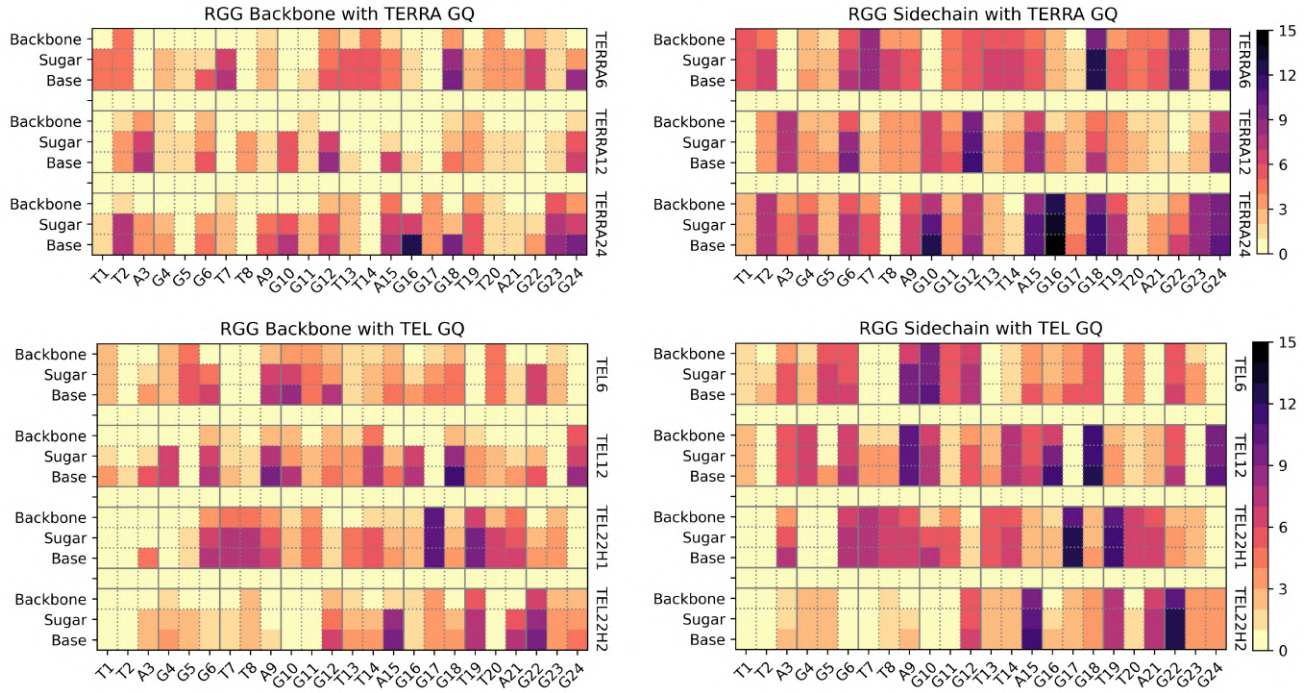

FIG. 6: RGG-GQ interaction patterns and binding modes. The number of interactions between each amino acid in RGG with every base of GQ averaged for the 50 encounter complexes of each GQ. The interactions for the RNA and DNA GQs are plotted separately. The interactions were classified into different pairs, backbone of RGG with backbone/sugar/base of GQ and sidechain of RGG with backbone/sugar/base of GQ, plotted in the rows. x-axis denotes the GQ sequence. The Guanines and TTA loops are differentiated by grid lines loops. The number of interactions are color coded on a yellow to black scale where darker colors indicate a larger number of interactions.

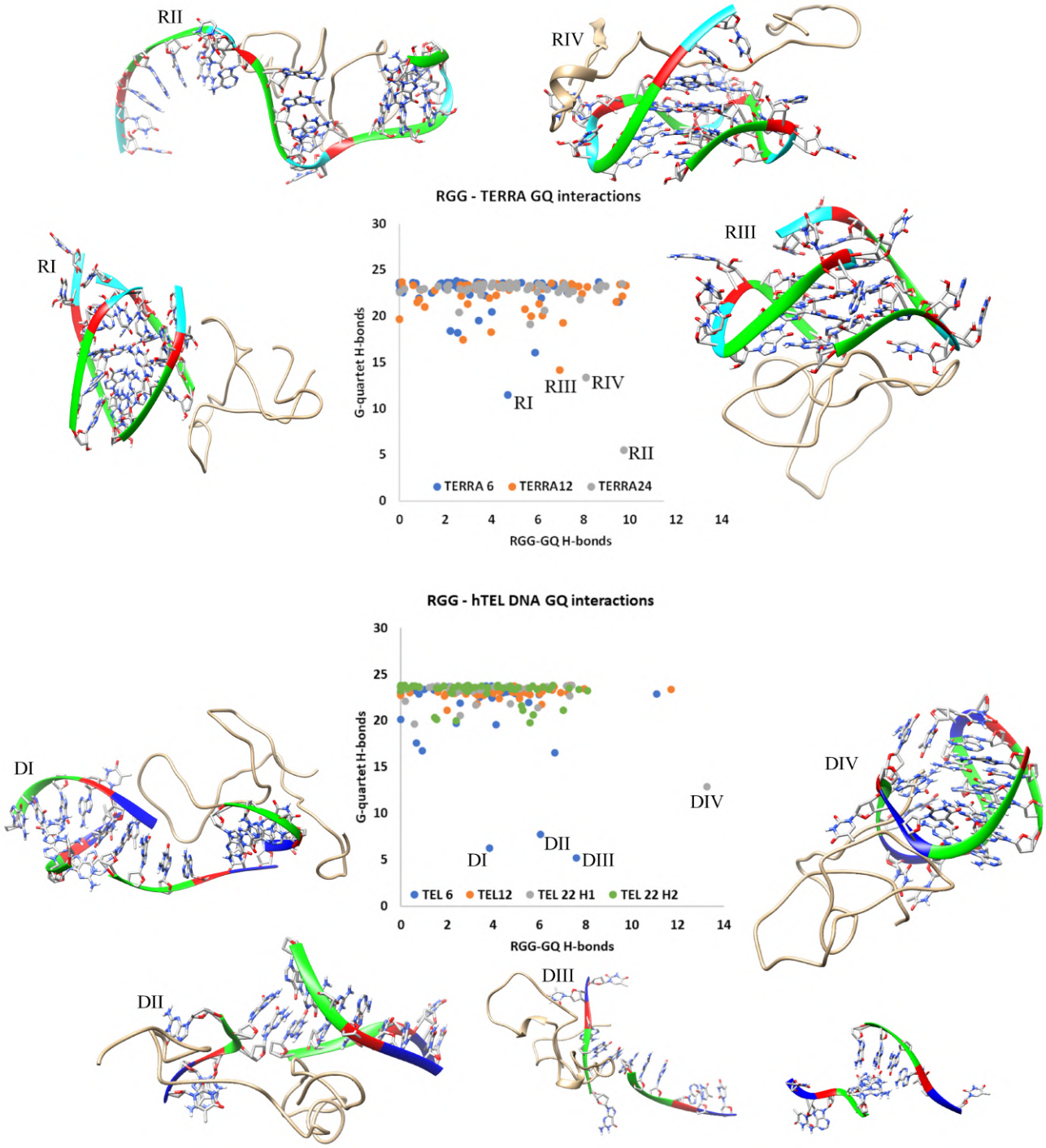

FIG. 7: Scatter plot between the number of intermolecular RGG-GQ (RNA and DNA) hydrogen bonds and intramolecular G-quartet hydrogen bonds highlighting the complex stability. The unfolding complexes are labelled and the structures after 50ns MD are shown.

revealed that in the groove binding mode, particularly in the tetramolecular structures, the RGG interacts weakly with the backbone of GQs, thereby straining the GQ to unfold. On the contrary, in the end stacking mode, the relative positioning of the electronegative functional groups of amino acids over the GQ core hinders the K+ ion coordination with the G-quartets, thereby destabilizing the quadruplex. As an example, we chose one unfolding complex representing each of the two binding modes and we further explored the mode of GQ destabilization by RGG. For the case of destabilization via end stacking, the TERRA12 complex with 7 RGG-GQ H-bonds disrupts the GQ to result in only 15 G-quartet H-bonds. By visualizing this trajectory, we clearly observed the expulsion of K+ ion by RGG (Fig. 8. Closer inspection of the RGG-GQ interaction revealed a *pi*-stacking interaction by Phe26 stabilizing the G12, while the Gly23 is present within the coordination shell of K+ ion (Fig. 8). The Gly23 is positioned to intrude into the core of GQ leading to repositioning and expulsion of K+ ions in the core. However, in case of groove binders like the tel6 complex, the RGG interacts directly with the nucleobases causing the GQs to unfold. For instance, the sidechain atoms of Asn53 in the tel6 complex (labelled DIII in Fig. 7), interacts directly with an oxygen atom in the deoxyribose sugar of G5 and N3 atom in the Watson-crick face of G4 (Fig. 8. These interactions are due to the protrusion of Asn sidechain into the groove between two parallel strands of DNA, thereby disintegrating the tetramolecular GQ into four single strands. Altogether, our observations indicate the protrusion of RGG residues into the GQ structure inducing instability in the GQ topology, strengthening the possibility of an RGG induced mechanism for destabilizing GQs.

It is generally known that RGG rich regions bind to nucleic acids, and the RGGs of several proteins like hnRNP, FMRP and FUS express either unfolding or stabilizing effects on GQ. Previous studies on FMRP RGG-box with an RNA duplex-quadruplex junction revealed a well-formed -turn in RGG is responsible for its binding to the RNA[47]. A similar involvement of β-spiral structure in FUS is reported as necessary for binding a GQ[48]. Our docking study has shown that such β-turns are accommodated well in the grooves of the GQ. The RGG of hnRNP A1 is reported to enhance the unfolding of intramolecular human telomeric GQ by UP1. In our study, we observed only a marginal preference for DNA GQ over RNA GQ and unimolecular TERRA over the tetramolecular TERRA. The individual RGGs were also able to destabilize GQs although to a lesser extent (8 unfolding events out of the total 350

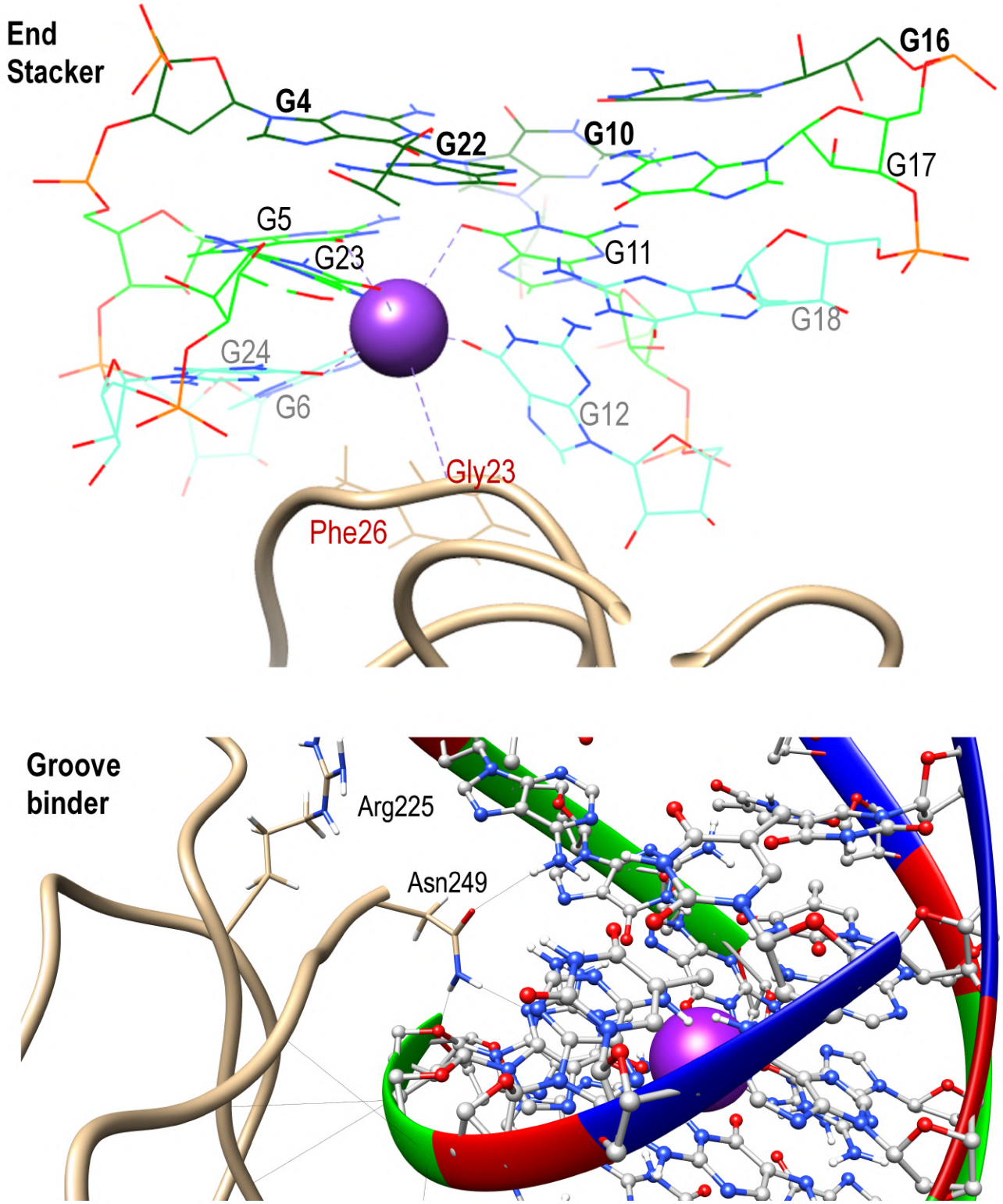

FIG. 8: RGG induced unfolding mechanism in (a) the end stacking mode of TERRA12 and (b) groove binding mode of tel6. The RGG backbone is shows as ribbons while the sidechains are depicted in wire representation. The K+ ion is shown as a purple sphere and its coordination shell is identified with dotted lines. Phosphate: Yellow, Carbon: gray (RGG)/brown (GQ), Nitrogen: Blue, Oxygen: red, Hydrogen: white and K+ ion: purple.

complexes generated). The cellular concentration of TERRA and hnRNP plays a vital role in regulating telomere maintainance. Moreover, the role of hnRNP UP1 on unfolding of GQ motifs is well-known. However, the unfolding mechanism of UP1+RGG is not yet clear, and it is hypothesized that the preference of RGG towards unimolecular TERRA enables UP1 binding. From our study, we propose an addendum to this hypothesis that the initial binding of RGG induces an instability in the GQ which might facilitate UP1 binding and be responsible for the enhanced unfolding by UP1.

II. CONCLUSION

This study is an attempt to faithfully capture the conformational landscape explored by the low complexity domain of hnRNPA1 through all-atom simulations and further explore its functional significance. Our hybrid replica exchange method has been able to successfully generate the conformational states attained by Δhex-LCD in solution that also matches well with the chemical shifts of an NMR ensemble. By employing the tSNE method, we were able to understand the heterogeneity of the conformational landscape and explore their functional significance. Though the global shape of the generated ensemble shows a compact nature, the local molecular interactions are captured quite well since the SAXS and DEER predictions were an excellent match. By generating an all-atom ensemble, we are able to provide a high-resolution picture of the local interactions in LCD, and LARKS are of particular interest due to their role in fibril formation. We were able to identify even the smallest kinks in the structure and their lifetimes. The transient appearance of LARKS at several sites of LCD and their high frequency of occurrence near the steric zipper motif identifies the most probable nucleation points for fibril formation. We also get a glimpse of how, on a single molecule level, the D262V mutation leads to increased local interaction and thereby more compaction. We speculate that this, in turn, can lead to more inter-fibril interaction across the monomers of amyloid fibrils and a recipe for irreversibility. Our simulations also show that the functionally independent regions of LCD, namely PrLD and RGG also behave like structurally independent domains expressing inter-domain interactions and LARKS. A disordered region expressing functional and structurally modular behavior is a very significant observation. It is known that the RGG binds to GQ structures of human

telomeric DNA and RNA in a groove binding or end stacking mode primarily interacting with the GQ loops. Previous studies have hypothesized that RGGs are required for recruiting UP1, however, our study has clearly showed another purpose for the RGGs. Based on our extensive docking and simulations, we add to the existing hypothesis that the RGG of hnRNP A1 has a destabilizing effect on the GQ structures. The RGG-GQ interactions are mainly dominated by Gly and Arg residues, while Gly and Asn are mainly responsible for GQ destabilization. Hence, we propose that apart from recruiting RRMs, the RGG also plays a crucial role in destabilizing the GQs and this is further enhanced when UP1 is also recruited. Overall in this paper, we were able to capture the modularity of an IDR and the interdomain interactions that create structurally relevant and functionally influential molecular motifs.

The availability of high-resolution conformation ensemble of the hnNPA1-LCD as well as its components allows high-fidelity simulation-assisted molecular recognition and binding geometries studies. We showed the usage of this for RGG interactions with various GQs. In Fig. S13, we report the sub-states of the PrLD system as well and Fig. S14 shows certain sub-states with high secondary structures features for RGG-box and PrLD. These conformations can be used to either explain experimental observable from a high-resolution structural point of view and can also be used to test certain hypothesis conveniently. As an example, we created the known aggregation-prone pathological mutation D262V in one of our PrLD conformation to explore the possible structural changes that could be leading to fibril formation. Preliminary analysis (with 1 μ of simulation time) of the all-atom ensemble of D262V mutant revealed changes in PrLD resulting in a compact structure (Fig. S15 in SI). Fig. S15(a) shows the reduction in radius of gyration when compared to the WT data. Accordingly, the local contact profile indicates higher number of residues within 0.5 nm of V262 (Fig. S15 (c)) than the D262 (Fig S15 b). Increase in local interactions also leads to reduction in the structural dynamics of LCD. All these results highlight that the mutation induces rigidness in LCD that is required for fibril formation, and the availability of high-resolution conformational ensemble data allowed us to test this possibility.

III. METHODS

A. Simulation

We have used an advanced sampling technique called replica exchange with hybrid tempering (REHT) developed in our lab[30]. In this hybrid approach, both the temperature as well as the Hamiltonian across the replicas are changed. The initial unfolded structures of full LCD, Δhex-LCD, RGG domain and PrLD have been obtained from Iterative Threading ASSEmbly Refinement (I-TASSER) which is a widely used server for automated protein structure prediction and structure-based functional annotation[49–51]. Using the amino acid sequence as a starting point, I-TASSER constructs 3D structural models. All four generated structures have collapsed coiled-coil formation. Table I contains detailed information on the full LCD, Δhex-LCD, RGG domain and PrLD systems analysed in this work. The usage of I-TASSER is limited to generate initial coordinates of the protein and to set up the molecular dynamics (MD) simulations from there. All the proteins were solvated in a cubic box with a minimum distance of 1.2 nm from the surface of the protein. 3-site rigid TIP3P water model was used commonly for four systems. The systems were also neutralized to maintain a physiological concentration of NaCl (0.15 M). The forcefield used for all four proteins systems is Charmm36m[52]. The simulations were carried out using Gromacs-2016.5 patched with Plumed-2.4.1[53–55]. Initially, the modelled proteins solutions were energy minimized using steepest descent algorithm for 50,000 steps to avoid any poor contacts. The energy minimized structure was then thermalized and equilibrated sequentially in NVT and NPT ensembles for 1 ns.

Next, the protein and the solvent were coupled separately to the target temperatures using the Nose-Hoover thermostat and the final production simulation was performed in the NVT ensemble. A cut-off of 1 nm was used for calculating the electrostatic and vdw interactions and Particle Mesh Ewald was used for long-range electrostatics. To integrate the equations of motions, the Leap-frog integrator with the time step of 2 fs was used. LINCS algorithm was used to constrain all the hydrogen atoms. Exchanges were attempted at every 1 ps interval.

In this approach, the Hamiltonian of the solute particle in the different replica is scaled-down up to 0.6-0.7 while also raising the temperature of the systems simultaneously up

Systems studied	Number of atoms	Number of replicas	Overall effective temperature range	Cumulative time per replica (in ns)	Average exchange acceptance probability
full LCD	75370	30	300 K-500 K	1000	0.21
Δhex-LCD	101640	25	300 K-430 K	1000	0.12
RGG-box	22948	20	300 K-500 K	1000	0.26
PrLD	23538	24	300 K-500 K	1000	0.3

TABLE I: Advanced sampling system information of simulation for RGG-box domain, PrLD, Δhex-LCD

to a maximum of 430-500 K. This way, as a function of both lambda scaling and explicit thermostat conditions, a very high temperature is ensured to be realized on the protein solute to sufficiently overcome the energy barriers. The solvent is heated up mildly in order for the rapid reorientation of the hydration shell upon conformational change of the protein. The number of replicas used for all the systems are indicated in Table I and the corresponding temperature ranges and average acceptance probabilities are also indicated. Post processing analyses of all the trajectories were performed with the Gromacs analysis tools and the trajectories were visualized using VMD.

For interaction studies of RGG-GQ complexes (listed in Table II), we performed molecular dynamics simulation using Gromacs-2020.4[56]. We used Charmm36m forcefield for the mutation studies and simulated both wild-type and mutated system for a duration of 500 ns. For the RGG-GQ complexes, we used a99SBdisp forcefield of Paul Robestelli for RGG, OL15 force field for DNA quadruplexes and OL3 force field for RNA quadruplexes. Both of the PrLD system and RGG-GQ complex systems (obtained from HADDOCK) were solvated with TIP3P water in a periodic box extending up to 1 nm and 1.2 nm in all directions respectively. The systems were neutralized and additional ions were added to mimic a salt concentration of 0.15 M. The short-range interactions were truncated with a cut-off distance of 1 nm. Electrostatic interactions were treated by particle-mesh Ewald with a real space cut-off value of 1 nm. Bonds containing hydrogens were constrained using the LINCS algorithm. The solvated and neutralized systems were energy minimized using the Steepest Descent algorithm followed by an equilibration of 5 ns and subsequently production runs. The temperature and pressure of the systems were maintained at 310 K and 1 atm using the nose-hoover thermostat and Parrinello-Rahman barostat in an NPT ensemble. Post processing analyses were performed with the Gromacs analysis tools, CPPTRAJ module of AmberTools20 and the trajectories were visualized using Visual Molecular Dynamics[57], UCSF

Systems	PDB ID	Simulated systems		Comments
		Apo (ns)	With the 50 conformers of RGG (ns)	
TERRA 6	based on 6GE1	300	50	tetramolecular RNA
TERRA 12	2KBP	300	50	bimolecular RNA
TERRA 24	based on 1KF1	300	50	unimolecular RNA
tel6	1NB9	300	50	tetramolecular DNA
tel12	1K8P	300	50	bimolecular DNA
tel22h1	2HY9	300	50	unimolecular DNA hybrid 1
tell22h2	2JPZ	300	50	unimolecular DNA hybrid 2

TABLE II: Telomeric RNA and DNA structures modeled in this study

Chimera v1.13. RGG-GQ interaction matrices were generated by calculating all heavy-atom contacts within 0.4 nm that are present for atleast 1% of the total simulation period for each GQ (50 complexes * 50 ns = 2.5 μs), and the number of RGG residues within 0.4 nm of each nucleobase is plotted.

B. Chemical shift calculation

We have used SPARTA+(version 2.90) in conjunction with a software package called Mdtraj[58,59]. SPARTA+ employs a well-trained neural network algorithm to construct quantitative relations between chemical shifts and protein structures, including backbone and side-chain conformation, H-bonding, electric fields and ring-current effects. It's single-level feed-forward multilayer artificial neural network (ANN) is capable of identifying the dependence of $^{15}N^{H}$, $^{1}3CO$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{1}H^{\alpha}$, ^{1}HN chemical shifts on the local structural and dynamic information as well as amino acid type, and those of its immediate neighbors. Mdtraj is a python software package which is designed to analyse trajectory data generated from MD simulations. We have compared our predicted data with the experimentally calculated chemical shift submitted in Biological Magnetic Resonance Data Bank (BMRB entry 50017)[25,60]. In the deposited data, the chemical shift for the residue Y244 and for a hexapeptide region in the range of 259-264 are missing.

C. SAXS calculation

For determination of protein structures, small angle X-ray scattering (SAXS) is a well-established technique which has been regularly used for years now. We have used a cutting-edge program called CRYSOL (version 3.0.3) to calculate the theoretical SAXS profile of Δhex-LCD[61]. This method evaluates solution scattering from a full-atom resolution structure using a multipole expansion for quick computation of spherical averaged scattering pattern. The program factors in the hydration shell as well. It can either predict the solution scattering curve or it can fit the experimental scattering curve by varying three parameters: (i) average displaced solvent volume per atomic group (ii) contrast of the hydration shell (iii) relative background. We have used experimental SAXS data provided by Dr. Tanja Mittag's lab for fitting purpose during calculation. Among the available arguments to compute theoretical SAXS profile, we have used '-eh' which accounts for explicit hydrogens and '-cst' which accounts for possible systematic errors due to mismatched buffers in the experimental data.

D. t-SNE clustering

We have chosen a nonlinear dimensionality reduction method called t-distributed Stochastic Neighbor Embedding (t-SNE) in order to separate high dimensional heterogenous IDP ensemble into meaningful clusters. An important aspect of t-SNE algorithm is the perplexity value, a tuneable parameter in t-SNE that balances the information between the local and global features of the dataset under investigation. Diligent choice of perplexity value for a given data is important to be able to most discretely divide the data into unambiguous clusters. In our work, we have noticed that for very low perplexity values, the clustering is extremely diffusive in nature and for a high value of perplexity parameter, the entire dataset seems to be treated as a single cluster. The performance of t-SNE is fairly robust to changes in the perplexity. In this work, we have explored several perplexity value options ranging from 50-2000. We have chosen the values which most discretely divides the datapoints into separate clusters.

To aid the result of t-SNE clustering in further partitioning, we have employed a widely used centroid based technique called K-means clustering which is one of the favoured and

standard unsupervised machine learning algorithms. We have used SciKit tool for these analyses which is an open source library for machine learning in Python[62].

E. DEER analysis

For DEER calculation, we used DEER-PREdict package developed by Larsen.et.al[35]. We spin labelled three pairs 231-271, 231-316 and 271-316. The calculation was done using the default parameters.

F. Modularity analysis

We have calculated a pairwise distance matrix of the full LCD and Δhex-LCD trajectory using mindist function of Gromacs. This operation calculates minimum distance between one group and multiple other groups as a result of time evolution and computes an average distance from it. We have plotted the output using Matplotlib[63] where the entire Δhex-LCD sequence is on both the axes.

G. Kinked-beta structure analysis

In this work, we wanted to predict and quantify the beta-kinks observed in of Δhex-LCD, RGG and PrLD for the entire trajectory. For this purpose, first we obtained the beta sheet information from DSSP and STRIDE. Then we came up with the collective variable: Cα dihedral angles of consecutive amino acids in the peptide backbone. The Cα atoms of regular beta sheets exist in an alternate side manner. On the other hand, in case of kinks, the i and i+3 Cα atoms stay on onside of the beta strand and the i+1, i+2 are situated on the other side of the strand. Hence the dihedral angles formed by the four consecutive Cα atoms fall in the acute angle category. We roughly divided the angle range 0-90 into four parts. Followed by this, we have taken a window of 4 amino acids and scanning the entire sequence for a specific angle range among the residues containing beta structures. Upon finding the angles in the allotted range, we have assigned them as beta-kinks. For each 4-residue window, we have also calculated the beta-kink percentage.

H. Structures of RGG and GQ

The 50 conformations of RGG obtained by clustering the REHT trajectories were docked with the human telomeric DNA and RNA GQ (also known as GQ). Unimolecular, bimolecular and tetramolecular forms of both DNA and RNA quadruplexes were modeled in this study (Table II). The tetramolecular TERRA (or TERRA6) was modeled using the solution structure with PDB ID: 6GE1, while the unimolecular TERRA24 was modeled based on the telomeric DNA quadruplex with PDB ID: 1KF1. The NMR structure of bimolecular telomeric RNA (or TERRA12) is available in the Protein Databank with the ID 2KBP. The structures of all forms of telomeric DNA quadruplexes were taken directly from the Protein Databank with IDs 1NB9 (tetramolecular or tel6), 1K8P (bimolecular tel12), 2HY9 (unimolecular form 1 tel22h1) and 2JPZ (unimolecular form 2 tel22h2). The unimolecular telomeric DNA in K+ ion solution adopts two different hybrid conformations that are equally populated and both the structures (differentiated as tel22h1 and tel22h2) were taken for our docking studies. The numbers behind TERRA* or tel* indicates the length of each RNA or DNA strand forming the GQ. Additional bases were added to the 5‘ or 3‘ ends of the GQ structures to match the sequences used in ITC studies by Ghosh et al[20,22]. Two K+ ions were placed manually between the three G-quartets of all 7 GQs. The modeled structures were simulated for a period of 300 ns and the structures extracted at 10 ns simulation were used as starting structures for docking studies.

I. HADDOCK Docking protocol

HADDOCK utilizes a data-driven approach for molecular docking. The stand-alone version of HADDOCK 2.4 was used to drive the docking of RGG with telomeric RNA and DNA GQs. Since the RGG conformations are randomly coiled structures of 53 amino acids, all residues are highly solvent exposed. In order to perform an unbiased docking study, the distance between the center of mass of the protein and GQ alone were used as a restraint criterion. The docking process is performed in three steps: rigid-body docking, semi-flexible refinement and water refinement. During the initial stage, 10,000 structures were generated starting from random orientations and the center of mass distance restraint. Also, the generated structures were allowed to undergo 180° rotations during the rigid body energy

minimization to eliminate any false positives. The best 4% of these structures were subsequently used in the second stage of semi-flexible refinement. During this stage, simulated annealing process with default settings was performed by allowing flexibility to the residues at the complex interface. The last stage of water refinement was performed for 200 complexes over 100 MD heating steps at 100, 200, and 300 K followed by 1250 sampling steps at 300 K and 500 MD cooling steps at 300, 200, and 100 K all with 2 fs time steps. ["Alternate concise version: About 10,000 complexes were generated during the first stage rigid-body docking by using the standard HADDOCK protocol [40]. Among these, 400 lowest-energy structures were selected for subsequent semi-flexible simulated annealing and explicit solvent (water) refinement, to optimize the interface."] HADDOCK scoring was performed according to the weighted sum (HADDOCK score) of different energy terms, including the van der Waals energy, electrostatic energy, distance restraints energy, inter-vector projection angle restraints energy, diffusion anisotropy energy, dihedral angle restraints energy, symmetry restraints energy, binding energy, desolvation energy and buried surface area. The final structures were clustered using the fraction of common contacts (FCC) with a cutoff of 0.6 and a minimum cluster size of 4.

IV. DATA AVAILABILITY

Input files needed to initiate molecular simulations and full trajectory data of all simulations for all systems considered in this work are available on our server for download. The server data can be publicly accessed via our laboratory GitHub link: codesrivastavalab/hnRNPA1. The files can also be accessed directly from our SharePoint location here.

AUTHOR CONTRIBUTIONS

AS conceived the idea. AS and RA designed and formulated the project with help of IR and SB. IR constructed and set up the hnRNPA1-LCD systems and SB constructed and set up models for hnRNPA1-GQ complex systems. IR and SB carried out all simulations and generated all the data, figures and plots for the manuscripts. IR, SB, RA and AS analyzed the data. IR and SB created the first draft of the manuscript, RA and AS refined the draft.

ACKNOWLEDGEMENTS

IR would like to acknowledge financial support from Ministry of Education, Government of India. R.A. would like to thank the early career fellowship from the DBT-Wellcome Trust India Alliance (Grant number: IA/E/18/1/504308). S.B. would like to thank the National Supercomputing Mission grant for financial support. A.S. also thanks the DST for the National Supercomputing Mission grant. Financial support from the Indian Institute of Science-Bangalore through I.o.E and the high-performance computing facility "Beagle" setup from grants by a partnership between the Department of Biotechnology of India and the Indian Institute of Science (IISc-DBT partnership programme) are greatly acknowledged.

REFERENCES

[1]Hsin-Pai Li, Xuming Zhang, Roger Duncan, Lucio Comai, and Michael MC Lai. Heterogeneous nuclear ribonucleoprotein a1 binds to the transcription-regulatory region of mouse hepatitis virus rna. Proceedings of the National Academy of Sciences, 94(18):9544–9549, 1997.

[2]Julie S. Lau, Peter Baumeister, Edmund Kim, Binayak Roy, Tsai-Yuan Hsieh, Michael Lai, and Amy S. Lee. Heterogeneous nuclear ribonucleoproteins as regulators of gene expression through interactions with the human thymidine kinase promoter. Journal of Cellular Biochemistry, 79(3):395–406, 2000.

[3]Jianguo Feng, Jianlong Zhou, Yunxiao Lin, and Wenhua Huang. hnrnp a1 in rna metabolism regulation and as a potential therapeutic target. Frontiers in Pharmacology, 13:986409, 2022.

[4]Akila Mayeda and Adrian R. Krainer. Regulation of alternative pre-mrna splicing by hnrnp a1 and splicing factor sf2. Cell, 68(2):365–375, 1992.

[5]Maurice S Swanson and Gideon Dreyfuss. Rna binding specificity of hnrnp proteins: a subset bind to the 3 end of introns. The EMBO journal, 7(11):3519–3529, 1988.

[6]Gideon Dreyfuss, Michael J Matunis, Serafin Pinol-Roma, and Christopher G Burd. hnrnp proteins and the biogenesis of mrna. Annual review of biochemistry, 62(1):289–321, 1993.

[7]Sara Nakielny and Gideon Dreyfuss. Transport of proteins and rnas in and out of the nucleus. Cell, 99(7):677–690, 1999.

[8]Stavroula Mili, Hong Jun Shu, Yingming Zhao, and Serafin Pinol-Roma. Distinct rnp complexes of shuttling hnrnp proteins with pre-mrna and mrna: candidate intermediates in formation and export of mrna. Molecular and cellular biology, 21(21):7307–7319, 2001.

[9]Jacques Jean-Philippe, Sean Paz, and Massimo Caputi. hnrnp a1: The swiss army knife of gene expression. International Journal of Molecular Sciences, 14(9):18999–19024, 2013.

[10]Sophie Bonnal, Frédéric Pileur, Cécile Orsini, Fabienne Parker, Françoise Pujol, Anne-Catherine Prats, and Stéphan Vagner. Heterogeneous nuclear ribonucleoprotein a1 is a novel internal ribosome entry site trans-acting factor that modulates alternative initiation of translation of the fibroblast growth factor 2 mrna. Journal of Biological Chemistry, 280(6):4144–4153, 2005.

[11]Oak D Jo, Jheralyn Martin, Andrew Bernath, Janine Masri, Alan Lichtenstein, and Joseph Gera. Heterogeneous nuclear ribonucleoprotein a1 regulates cyclin d1 and c-myc internal ribosome entry site function through akt signaling. Journal of Biological Chemistry, 283(34):23274–23287, 2008.

[12]Joseph P. Clarke, Patricia A. Thibault, Hannah E. Salapa, and Michael C. Levin. A comprehensive analysis of the role of hnrnp a1 function and dysfunction in the pathogenesis of neurodegenerative disease. Frontiers in Molecular Biosciences, 8, 2021.

[13]Qing-Shuo Zhang, Lisa Manche, Rui-Ming Xu, and Adrian R. Krainer. hnrnp a1 associates with telomere ends and stimulates telomerase activity. RNA, 12(6):1116–1128, 2006.

[14]Uriya Bekenstein and Hermona Soreq. Heterogeneous nuclear ribonucleoprotein a1 in health and neurodegenerative disease: From structural insights to post-transcriptional regulatory roles. Molecular and Cellular Neuroscience, 56:436–446, 2013. RNA and splicing regulation in neurodegeneration.

[15]Hélène LaBranche, Sophie Dupuis, Yaacov Ben-David, Maria-Rosa Bani, Raymund J Wellinger, and Benoit Chabot. Telomere elongation by hnrnp a1 and a derivative that interacts with telomeric repeats and telomerase. Nature genetics, 19(2):199–202, 1998.

[16]Frederic Dallaire, Sophie Dupuis, Stephan Fiset, and Benoit Chabot. Heterogeneous nuclear ribonucleoprotein a1 and up1 protect mammalian telomeric repeats and modulate telomere replication in vitro. Journal of Biological Chemistry, 275(19):14509–14516, 2000.

[17]Hirokazu Fukuda, Masato Katahira, Naoto Tsuchiya, Yoshiaki Enokizono, Takashi Sugimura, Minako Nagao, and Hitoshi Nakagama. Unfolding of quadruplex structure in the

g-rich strand of the minisatellite repeat by the binding protein up1. Proceedings of the National Academy of Sciences, 99(20):12685–12690, 2002.

[18]Patricia A Thibault, Aravindhan Ganesan, Subha Kalyaanamoorthy, Joseph-Patrick WE Clarke, Hannah E Salapa, and Michael C Levin. hnrnp a/b proteins: an encyclopedic assessment of their roles in homeostasis and disease. Biology, 10(8):712, 2021.

[19]Wang YD Scarborough EA Moore J Diaz Z MacLea KS Freibaum B Li S Molliex A Kanagaraj AP Carter R Boylan KB Wojtas AM Rademakers R Pinkus JL Greenberg SA Trojanowski JQ Traynor BJ Smith BN Topp S Gkazi AS Miller J Shaw CE Kottlors M Kirschner J Pestronk A Li YR Ford AF Gitler AD Benatar M King OD Kimonis VE Ross ED Weihl CC Shorter J Taylor JP Kim HJ, Kim NC. Mutations in prion-like domains in hnrnpa2b1 and hnrnpa1 cause multisystem proteinopathy and als. Cell, 163(1):123–133, 2015.

[20]Meenakshi Ghosh and Mahavir Singh. RGG-box in hnRNPA1 specifically recognizes the telomere G-quadruplex DNA and enhances the G-quadruplex unfolding ability of UP1 domain. Nucleic Acids Research, 46(19):10246–10261, 09 2018.

[21]Xiao Liu, Takumi Ishizuka, Hong-Liang Bao, Kei Wada, Yuma Takeda, Keisuke Iida, Kazuo Nagasawa, Danzhou Yang, and Yan Xu. Structure-dependent binding of hnrnpa1 to telomere rna. Journal of the American Chemical Society, 139(22):7533–7539, 2017. PMID: 28510424.

[22]Meenakshi Ghosh and Mahavir Singh. Structure specific recognition of telomeric repeats containing RNA by the RGG-box of hnRNPA1. Nucleic Acids Research, 48(8):4492–4506, 03 2020.

[23]Lee J Coughlin M Kanagaraj AP Kim HJ Mittag T Taylor JP Molliex A, Temirov J. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. Cell, 163(1):123–133, 2015.

[24]Veronica H. Ryan, Gregory L. Dignon, Gül H. Zerze, Charlene V. Chabata, Rute Silva, Alexander E. Conicella, Joshua Amaya, Kathleen A. Burke, Jeetain Mittal, and Nicolas L. Fawzi. Mechanistic view of hnrnpa2 low-complexity domain structure, interactions, and phase separation altered by mutation and arginine methylation. Molecular Cell, 69(3):465–479.e7, 2018.

[25]Erik W. Martin, Alex S. Holehouse, Ivan Peran, Mina Farag, J. Jeremias Incicco, Anne Bremer, Christy R. Grace, Andrea Soranno, Rohit V. Pappu, and Tanja Mittag. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. Science, 367(6478):694–699, 2020.

[26]Erik W Martin, F Emil Thomasen, Nicole M Milkovic, Matthew J Cuneo, Christy R Grace, Amanda Nourse, Kresten Lindorff-Larsen, and Tanja Mittag. Interplay of folded domains and the disordered low-complexity domain in mediating hnRNPA1 phase separation. Nucleic Acids Research, 49(5):2931–2945, 02 2021.

[27]Irina Ritsch, Elisabeth Lehmann, Leonidas Emmanouilidis, Maxim Yulikov, Frédéric Allain, and Gunnar Jeschke. Phase separation of heterogeneous nuclear ribonucleoprotein a1 upon specific rna-binding observed by magnetic resonance**. Angewandte Chemie International Edition, 61(40):e202204311, 2022.

[28]Bruna Baumgarten Krebs and Joelma Freire De Mesquita. Amyotrophic lateral sclerosis type 20 - in silico analysis and molecular dynamics simulation of hnrnpa1. PLOS ONE, 11(7):1–18, 07 2016.

[29]D. Janka Bauer, Lukas S. Stelzl, and Arash Nikoubashman. Single-chain and condensed-state behavior of hnrnpa1 from molecular simulations. The Journal of Chemical Physics, 157(15):154903, 2022.

[30]Rajeswari Appadurai, Jayashree Nagesh, and Anand Srivastava. High resolution ensemble description of metamorphic and intrinsically disordered proteins using an efficient hybrid parallel tempering scheme. Nature communications, 12(1):1–11, 2021.

[31]Rajeswari Appadurai, Jaya Krishna Koneru, Massimiliano Bonomi, Paul Robustelli, and Anand Srivastava. Clustering heterogeneous conformational ensembles of intrinsically disordered proteins with t-distributed stochastic neighbor embedding. Journal of Chemical Theory and Computation, 19(14):4711–4727, 2023. PMID: 37338049.

[32]Boyer DR Goldschmidt L Rodriguez JA Cascio D Chong L Gonen T Eisenberg DS Hughes MP, Sawaya MR. Atomic structures of low-complexity protein segments reveal kinked sheets that assemble networks. Science, 359(6376):698–701, 2018.

[33]Xinrui Gui, Feng Luo, Yichen Li, Heng Zhou, Zhenheng Qin, Zhenying Liu, Jinge Gu, Muyun Xie, Kun Zhao, Bin Dai, Woo Shik Shin, Jianhua He, Lin He, Lin Jiang, Minglei Zhao, Bo Sun, Xueming Li, Cong Liu, and Dan Li. Structural basis for reversible amyloids

of hnRNPA1 elucidates their role in stress granule assembly. Nature Communications, 10(1):2006, 2019.

[34]Irina Ritsch, Laura Esteban-Hofer, Elisabeth Lehmann, Leonidas Emmanouilidis, Maxim Yulikov, Frédéric H-T Allain, and Gunnar Jeschke. Characterization of weak protein domain structure by spin-label distance distributions. Frontiers in Molecular Biosciences, 8:636599, 2021.

[35]Giulio Tesei, João M Martins, Micha BA Kunze, Yong Wang, Ramon Crehuet, and Kresten Lindorff-Larsen. Deer-predict: Software for efficient calculation of spin-labeling epr and nmr data from conformational ensembles. PLOS Computational Biology, 17(1):e1008551, 2021.

[36]Dorothee Dormann, Ramona Rodde, Dieter Edbauer, Eva Bentmann, Ingeborg Fischer, Alexander Hruscha, Manuel E Than, Ian RA Mackenzie, Anja Capell, Bettina Schmid, et al. Als-associated fused in sarcoma (fus) mutations disrupt transportin-mediated nuclear import. The EMBO journal, 29(16):2841–2857, 2010.

[37]Dorothee Dormann, Tobias Madl, Chiara F Valori, Eva Bentmann, Sabina Tahirovic, Claudia Abou-Ajram, Elisabeth Kremmer, Olaf Ansorge, Ian RA Mackenzie, Manuela Neumann, et al. Arginine methylation next to the py-nls modulates transportin binding and nuclear import of fus. The EMBO journal, 31(22):4258–4275, 2012.

[38]Jozsef Gal, Jiayu Zhang, David M Kwinter, Jianjun Zhai, Hongge Jia, Jianhang Jia, and Haining Zhu. Nuclear localization sequence of fus and induction of stress granules by als mutants. Neurobiology of aging, 32(12):2323–e27, 2011.

[39]Karli K McDonald, Anaïs Aulas, Laurie Destroismaisons, Sarah Pickles, Evghenia Beleac, William Camu, Guy A Rouleau, and Christine Vande Velde. Tar dna-binding protein 43 (tdp-43) regulates stress granule dynamics via differential regulation of g3bp and tia-1. Human molecular genetics, 20(7):1400–1410, 2011.

[40]Liqun Liu-Yesucevitz, Aylin Bilgutay, Yong-Jie Zhang, Tara Vanderwyde, Allison Citro, Tapan Mehta, Nava Zaarur, Ann McKee, Robert Bowser, Michael Sherman, et al. Tar dna binding protein-43 (tdp-43) associates with stress granules: analysis of cultured cells and pathological brain tissue. PloS one, 5(10):e13250, 2010.

[41]Daryl A Bosco, Nathan Lemay, Hae Kyung Ko, Hongru Zhou, Chris Burke, Thomas J Kwiatkowski Jr, Peter Sapp, Diane McKenna-Yasek, Robert H Brown Jr, and Lawrence J

Hayward. Mutant fus proteins that cause amyotrophic lateral sclerosis incorporate into stress granules. Human molecular genetics, 19(21):4160–4175, 2010.

[42]Colleen M Dewey, Basar Cenik, Chantelle F Sephton, Daniel R Dries, Paul Mayer III, Shannon K Good, Brett A Johnson, Joachim Herz, and Gang Yu. Tdp-43 is directed to stress granules by sorbitol, a novel physiological osmotic and oxidative stressor. Molecular and cellular biology, 31(5):1098–1108, 2011.

[43]Claudia Colombrita, Eleonora Zennaro, Claudia Fallini, Markus Weber, Andreas Sommacal, Emanuele Buratti, Vincenzo Silani, and Antonia Ratti. Tdp-43 is recruited to stress granules in conditions of oxidative insult. Journal of neurochemistry, 111(4):1051–1061, 2009.

[44]Andres R Tejedor, Ignacio Sanchez-Burgos, Maria Estevez-Espinosa, Adiran Garaizar, Rosana Collepardo-Guevara, Jorge Ramirez, and Jorge R Espinosa. Protein structural transitions critically transform the network connectivity and viscoelasticity of rna-binding protein condensates but rna can prevent it. Nature communications, 13(1):1–15, 2022.

[45]Eisenberg DS Hughes MP, Goldschmidt L. Prevalence and species distribution of the low-complexity, amyloid-like, reversible, kinked segment structural motif in amyloid-like fibrils. Journal of Biological Chemistry, 297(4):101194, 2021.

[46]Isaacson RL. Morris OM, Torpey JH. Intrinsically disordered proteins: modes of binding with emphasis on disordered domains. Open Biol., 11(10), 10 2021.

[47]Nikita Vasilyev, Anna Polonskaia, Jennifer C Darnell, Robert B Darnell, Dinshaw J Patel, and Alexander Serganov. Crystal structure reveals specific recognition of a g-quadruplex rna by a β-turn in the rgg motif of fmrp. Proceedings of the National Academy of Sciences, 112(39):E5391–E5400, 2015.

[48]Ryota Yagi, Takatsugu Miyazaki, and Takanori Oyoshi. G-quadruplex binding ability of TLS/FUS depends on the Î²-spiral structure of the RGG domain. Nucleic Acids Research, 46(12):5894–5901, 05 2018.

[49]Jianyi Yang, Renxiang Yan, Ambrish Roy, Dong Xu, Jonathan Poisson, and Yang Zhang. The i-tasser suite: protein structure and function prediction. Nature methods, 12(1):7–8, 2015.

[50]Ambrish Roy, Alper Kucukural, and Yang Zhang. I-tasser: a unified platform for automated protein structure and function prediction. Nature protocols, 5(4):725–738, 2010.

[51]Yang Zhang. I-tasser server for protein 3d structure prediction. BMC bioinformatics, 9(1):1–8, 2008.

[52]Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L De Groot, Helmut Grubmüller, and Alexander D MacKerell. Charmm36m: an improved force field for folded and intrinsically disordered proteins. Nature methods, 14(1):71–73, 2017.

[53]Promoting transparency and reproducibility in enhanced molecular simulations. Nature methods, 16(8):670–673, 2019.

[54]Gareth A Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. Plumed 2: New feathers for an old bird. Computer physics communications, 185(2):604–613, 2014.

[55]Massimiliano Bonomi, Davide Branduardi, Giovanni Bussi, Carlo Camilloni, Davide Provasi, Paolo Raiteri, Davide Donadio, Fabrizio Marinelli, Fabio Pietrucci, Ricardo A Broglia, et al. Plumed: A portable plugin for free-energy calculations with molecular dynamics. Computer Physics Communications, 180(10):1961–1972, 2009.

[56]Lindahl, Abraham, Hess, and van der Spoel. Gromacs 2020.4 manual, October 2020.

[57]William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: visual molecular dynamics. Journal of molecular graphics, 14(1):33–38, 1996.

[58]Bax A Shen Y. Sparta+: a modest improvement in empirical nmr chemical shift prediction by means of an artificial neural network. J Biomol NMR, 48(1):13–22, 2010.

[59]Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee-Ping Wang, Thomas J. Lane, and Vijay S. Pande. Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. Biophysical Journal, 109(8):1528 – 1532, 2015.

[60]Eldon L. Ulrich, Hideo Akutsu, Jurgen F. Doreleijers, Yoko Harano, Yannis E. Ioannidis, Jundong Lin, Miron Livny, Steve Mading, Dimitri Maziuk, Zachary Miller, Eiichi Nakatani, Christopher F. Schulte, David E. Tolmie, R. Kent Wenger, Hongyang Yao, and John L. Markley. BioMagResBank. Nucleic Acids Research, $36(\text{suppl}_1) : D402 - -D408, 112007$.

[61]D. Franke, M. V. Petoukhov, P. V. Konarev, A. Panjkovich, A. Tuukkanen, H. D. T. Mertens, A. G. Kikhney, N. R. Hajizadeh, J. M. Franklin, C. M. Jeffries, and D. I. Svergun. *ATSAS 2.8*: a comprehensive data analysis suite for small-angle scattering from macromolecular

solutions. Journal of Applied Crystallography, 50(4):1212–1225, Aug 2017.

[62]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[63]J. D. Hunter. Matplotlib: A 2d graphics environment. Computing in Science & Engineering, 9(3):90–95, 2007.