Conformational Ensembles of Non-Coding Elements in the SARS-CoV-2 Genome from Molecular Dynamics Simulations

Sandro Bottaro[1,*], **Giovanni Bussi**[2], **Kresten Lindorff-Larsen**[1,*]

***For correspondence:**
sandro.bottaro@bio.ku.dk (SB);
lindorff@bio.ku.dk (KLL)

[1]Structural Biology and NMR Laboratory & Linderstrøm-Lang Centre for Protein Science. Department of Biology, University of Copenhagen. Ole Maaløes Vej 5, DK-2200 Copenhagen N, Denmark; [2]Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136, Trieste, Italy

Abstract The 5′ untranslated region (UTR) of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genome is a conserved, functional and structured genomic region consisting of several RNA stem-loop elements. While the secondary structure of such elements has been determined experimentally, their three-dimensional structures are not known yet. Here, we predict structure and dynamics of five RNA stem loops in the 5′-UTR of SARS-CoV-2 by extensive atomistic molecular dynamics simulations, more than 0.5ms of aggregate simulation time, in combination with enhanced sampling techniques. We compare simulations with available experimental data, describe the resulting conformational ensembles, and identify the presence of specific structural rearrangements in apical and internal loops that may be functionally relevant. Our atomic-detailed structural predictions reveal a rich dynamics in these RNA molecules, could help the experimental characterisation of these systems, and provide putative three-dimensional models for structure-based drug design studies.

Introduction

The genome of the SARS-CoV-2 virus is a single, positive-stranded RNA consisting of approximately 30 thousand nucleotides. Coding regions of the genome are translated into several non-structural proteins, as well as in the widely studied spike protein, envelope protein, membrane protein, and others [1]. The 5′ and 3′ untranslated regions (UTR) contain cis-acting sequences required for viral transcription and replication [2, 3]. In particular, the 5′-UTR harbours conserved, functional elements that enhance viral transcription [4, 5, 6], and are involved in discontinuous transcription leading to leader-body fusion [7]. Recent studies suggest that the 5′-UTR plays a key role in liquid-liquid phase separation phenomena with the SARS-CoV-2 nucleocapsid protein [8, 9]. Proof-of-principle studies using locked nucleic acids targeting conserved structural motifs in the SARS-CoV-2 genome showed the potential for inhibiting growth via RNA-interacting molecules [10].

Chemical probing experiments, NMR chemical shift measurements, and computational predictions have shown that the 5′-UTR is highly structured [11, 12, 13, 14, 15, 10], and consists of several stem-loop elements (Fig. 1). The secondary structures of the isolated elements are in close agreement with the full-length construct [11], suggesting that they can fold independently. The three-dimensional structure of stem loop 2 (SL2) from SARS-CoV-1, which is completely conserved

in SARS-CoV-2, has been determined by NMR spectroscopy [16], while no experimental structures exist for the remaining elements in the 5′-UTR. In a recent computational study [17], Rangan et al predicted the three-dimensional fold of all stem-loop elements in the 5′-UTR, 3′-UTR, and of the frameshifting stimulatory element using Rosetta's FARFAR2 algorithm [18, 19]. The predicted model of SL2 in the 5′-UTR was later refined using MD simulations, [20] leading to an improved agreement with available NMR data [16].

While there has been a plethora of structural studies of SARS-CoV-2 proteins, much less is known about the structure of the RNA genome at atomic-resolution. RNA molecules often display a rich, complex dynamical behaviour that is difficult to capture using experiments alone and which can play crucial roles in cellular functions [21, 22]. In this work we have thus used atomistic molecular dynamics simulations to predict the structure and dynamics of five 5′-UTR stem-loops. These elements were chosen because their high degree of conservation in betacoronaviruses, together with their vital role in viral replication, make them attractive, potential targets for small molecules. Experimental studies on the same constructs are currently being pursued within the COVID-19 NMR initiative (https://covid19-nmr.de/). The results presented in this work therefore represent a step in using computation and experiments in a synergistic manner. At the same time, the elements that we have chosen to model represent the largest structural elements for which it is currently feasible to obtain relatively converged simulations even with multi-microsecond enhanced sampling simulations.

With this spirit, we have generated three-dimensional configurations compatible with the secondary structure recently determined by chemical probing and NMR experiments [11]. While several fragment-based tools exist to perform such task [23, 24], we here adapt an MD-based procedure [25]. Similar secondary structure restraining approaches have been employed in all-atom RNA folding simulations [26, 27] and to study internal loop dynamics [28]. These initial structures are subject to extensive MD sampling, amounting to more than 0.5ms of aggregate simulation time. More precisely, we enhance the sampling of apical and internal loop regions using a temperature replica-exchange scheme in which only specific portions of the RNA solute are affected. [29, 30, 31]. For such regions we currently lack detailed structural information, and available experimental data suggest that conformational rearrangements are more likely to occur. Our simulations are thus a Boltzmann-weighted ensemble of three-dimensional conformations constituting our prediction of structure and conformational heterogeneity of these elements.

First, we apply the computational protocol to SL2, and validate the accuracy of our prediction against the three-dimensional structure of SL2 from SARS-CoV-1 and the NMR data used to generate this structure [16]. We then proceed by describing our predictions on four elements of increasing complexity (Fig. 1), and discuss the agreement with available NMR measurements [32, 33]. Stem loop 5a (SL5a) is part of the four-way junction SL5 containing the AUG starting codon, whose function is linked to RNA packaging [34]. Stem loop 1 (SL1) has been suggested to play a role in escaping the viral mechanism inhibiting mRNA translation [35, 6]. The function of stem loop 4 (SL4) is still not completely understood, and may act as a spacing element [36]. Finally, the SL2+SL3 construct consists of SL2 and stem loop 3 (SL3). The latter element contains the transcription regulatory sequence necessary for discontinuous RNA synthesis [37].

Our predictions could help the atomic-detailed experimental characterisation of such systems, as simulations can be easily integrated a posteriori with NMR or SAXS data [38, 39, 40]. The predicted models provide a starting point for structure-based drug design studies [41, 42], and could serve as additional benchmark systems to evaluate the accuracy of atomistic RNA force-field [43]. To maximize reproducibility and enable others to build on our work [44], we have made trajectories, input files, and analysis scripts available via github at https://github.com/KULL-Centre/papers/tree/master/2020/COVID_5UTR_MD.

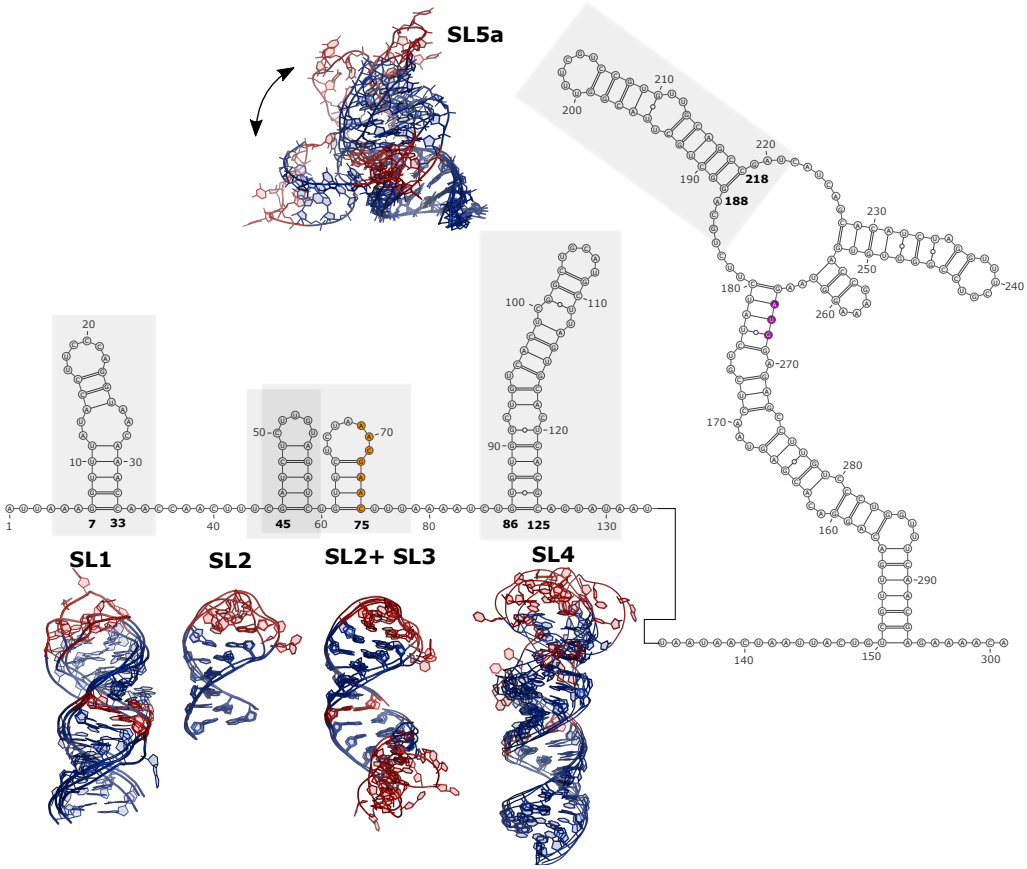

Figure 1. Reference sequence and secondary structure of the 5'-UTR in SARS-CoV-2 genome. The numbering follows the convention in [11]. The five structured elements considered in this work are labeled and highlighted in grey: SL1, SL2, SL2+SL3, SL4, and SL5a. The transcription regulatory sequence and the AUG start codon are highlighted in orange and purple, respectively. Selected centroids of the most populated clusters from MD simulations corresponding to each element are shown.

Results and Discussion

In this section we describe the conformational ensembles obtained from MD simulations. We first show the results obtained on SL2, the simplest system considered in this study. Crucially, in this case it is possible to validate the procedure by comparing with available structural data and NMR measurements.

Stem Loop 2

We run unbiased MD simulations of a single-stranded, extended RNA with sequence `GGAUCUCUUGUAGAUCU`. After 10ns, we apply a biasing force (in a process we term 'pulling') to promote the formation of an ideal A-form RNA structure in the region where the secondary structure is known (blue nucleotides in Fig. 2a). By construction the structural dissimilarity from the template, here evaluated using a nucleic-acids specific distance called ellipsoidal root mean square distance (eRMSD)[45], decreases during pulling (Fig. 2b, top panel). After 40 ns the external bias is removed, and the system fluctuates around a stable structure in which all five Watson-Crick (WC) base-pairs are formed. (Fig. 2b, bottom panel).

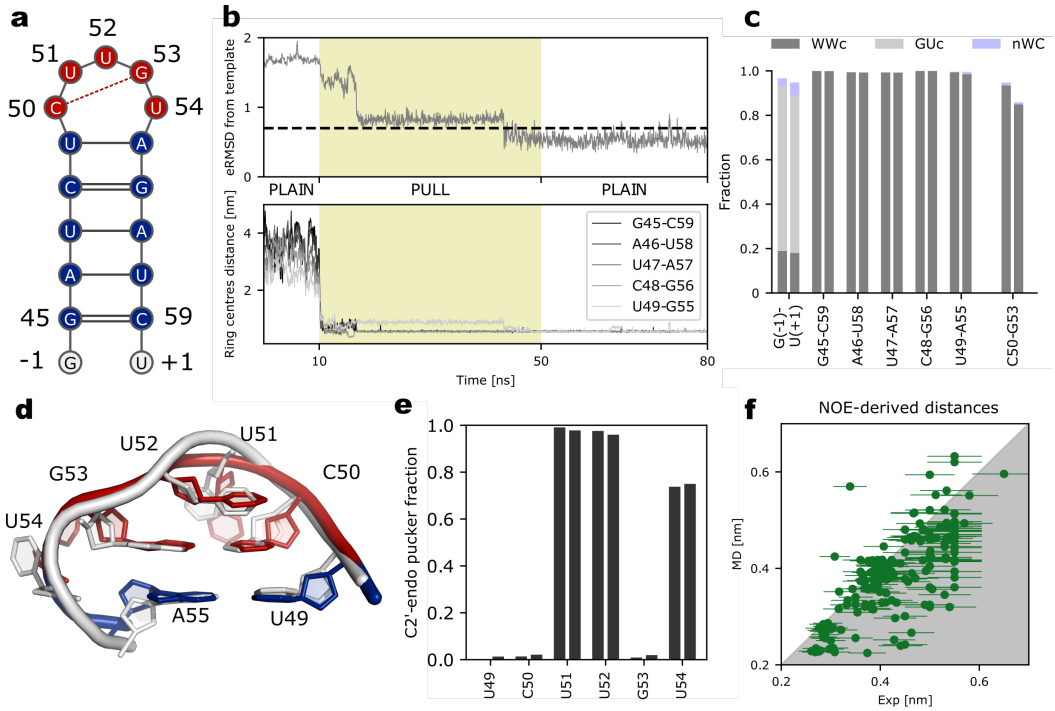

Figure 2. Description of the computational procedure and validation on SL2. **a** Sequence and secondary structure of SL2. The genomic numbering follows the convention in [11]. The region used for constructing the template A-form three-dimensional structure is shown in blue. The sampling in the region shown in red is enhanced by partial tempering. The Watson-Crick base-pair between C50 and G53 predicted in partial tempering simulations is shown as a dashed red line. **b** Behaviour of a successful pulling simulation. eRMSD (top panel) and distance between ring centres of the five WC base-pairs (bottom panel) during the three stages of the procedure: initial plain MD, pulling stage, and final plain MD stage, as labelled. The eRMSD threshold of 0.75 is shown as a dashed horizontal line. **c** Population of base-pairs from 20 replicas $\times$ 2.6 μs partial tempering simulations. Following the Leontis-Westhof classification [46], base-pairs are annotated as cis Watson-Crick/Watson-Crick (WWc), GU wobble cis (GUc) or non-Watson-Crick (nWC). For each interaction, the two bars show the statistics from two independent simulations. **d** Centroid of the most-populated cluster from MD simulations (blue/red) superposed onto the SL2 loop from SARS-CoV-1 structure (white) [16]. **e** Sugar pucker population in the loop region. **f** Calculated and experimental NOE-derived upper bound distances. The 10 of the 179 calculated NOEs that are larger than the experimental values are located in the white upper-triangular region.

While pulling simulations make it possible quickly to generate conformations with a desired

stem structure, these runs are not sufficiently long to allow rearrangements in the loop region. Previous simulation studies have shown these motions to occur on long timescales (μs to ms) that require extensive simulations [47]. We thus enhance the sampling in the loop region (red in Fig. 2a) by running extensive partial tempering simulations (20 replicas, each 2.6 μs-long) initialised from a successful pulling run (see Methods).

Analyses of the enhanced simulations show that all five base-pairs in the stem are stable, and a comparison of two independent simulations shows that our simulations are well converged (Fig. 2c). In addition, we observe the formation of a stable terminal GU wobble base-pair, as well as of a C50-G53 Watson-Crick interaction in the loop, resulting in a structure resembling a CUUG tetraloop [48, 49] with a bulged U at position 54. The interaction between C50 and G53 is also compatible with the decreased sensitive to chemical probing of these two bases relative to the remaining three in the loop [10]. A three-dimensional cluster analysis [50] reveals that the largest cluster ($\approx 90\%$) is remarkably similar to the available NMR loop structure from SARS-CoV-1 [16], whose sequence differs from our construct in the two terminal base-pairs. For example, the centroid of the most populated cluster from MD simulations superimposes extremely well to the first NMR model in PDB 2L6I (Fig. 2d). In the loop, C50 forms a canonical base-pair with G53, U52 stacks on top of C50, and both U51/U54 are solvent exposed. The median distance between the experimental structure and the MD simulation is 1.6 Å heavy-atom root mean square deviation and 0.5 eRMSD, as shown in Supporting Information 1 (SI1).

The sugar puckers of U51, U52, and U54 are preferably in C2'-endo conformation, in full agreement with NMR measurements [16] (Fig. 2e). The great majority (169 out of 179) of upper-bound NOE-derived distances are satisfied in MD simulations (Fig. 2f). The most significant discrepancies consist of short proton-proton distances between G53-U54 that are not sampled in our MD runs. The full list of violated NOEs is shown in Fig. SI2.

Stem loop 5a

SL5a consists of two helices (blue nucleotides in Fig. 3a) connected by an internal loop and capped by an hexaloop with sequence UUUCGU. We use the same procedure as described above for SL2 to generate initial configurations and to sample the ensemble of SL5a by enhancing regions where secondary structure is not known (red regions in Fig. 3a). Helix 1 is highly stable throughout the 20 replicas $\times$ 3.6 μs partial tempering simulations (Fig. 3b).

In the internal loop in SL5a we observe the formation of a U194-U212 base-pair with a population of $\approx 60\%$ (Fig. 3b). The wobble GU base-pair between U195 and G210 in helix 2 is in equilibrium with an alternative WC/WC base-pair between U195 and U211. U212 is highly dynamic (see below), while other nucleotides in the internal loop, including G210, remain in the stack throughout MD simulations. Our predictions are overall compatible with NMR data, suggesting U194 to be base-paired, and U195-G210 at least partially stable [32].

The most frequent conformation of the apical loop is characterised by a wobble interaction between U200 and G204, with U205 completely solvent exposed. Sugar pucker conformation for nucleotides 201 to 204 are predominantly in C2'-endo conformation, at variance with NMR data that indicate C2'-endo conformations only at positions U202 and C203 [32]. In one of the two simulations we observe the transient formation of a U200-U205 base-pair capped by a canonical UUCG tetraloop fold [51]. The low population of this state is not surprising, as the canonical UUCG tetraloop fold has been shown not to be stabilised to a sufficient degree within this force-field [52]. While the detailed architecture of the hexaloop has not yet been solved experimentally [11, 32], our computational prediction and available NMR data cannot be easily reconciled in this case. Instead, we suggest that the simulations might be integrated with structural data collected in the future through a reweighting procedure [40].

In addition to the analysis of individual base-pairs, we identify in the simulations frequent transitions from straight to kinked conformations (Figs. 3c-d). The latter structure is reminiscent of a kink-turn motif, but lacks its signature interactions and sequence propensities [53]. This large-scale

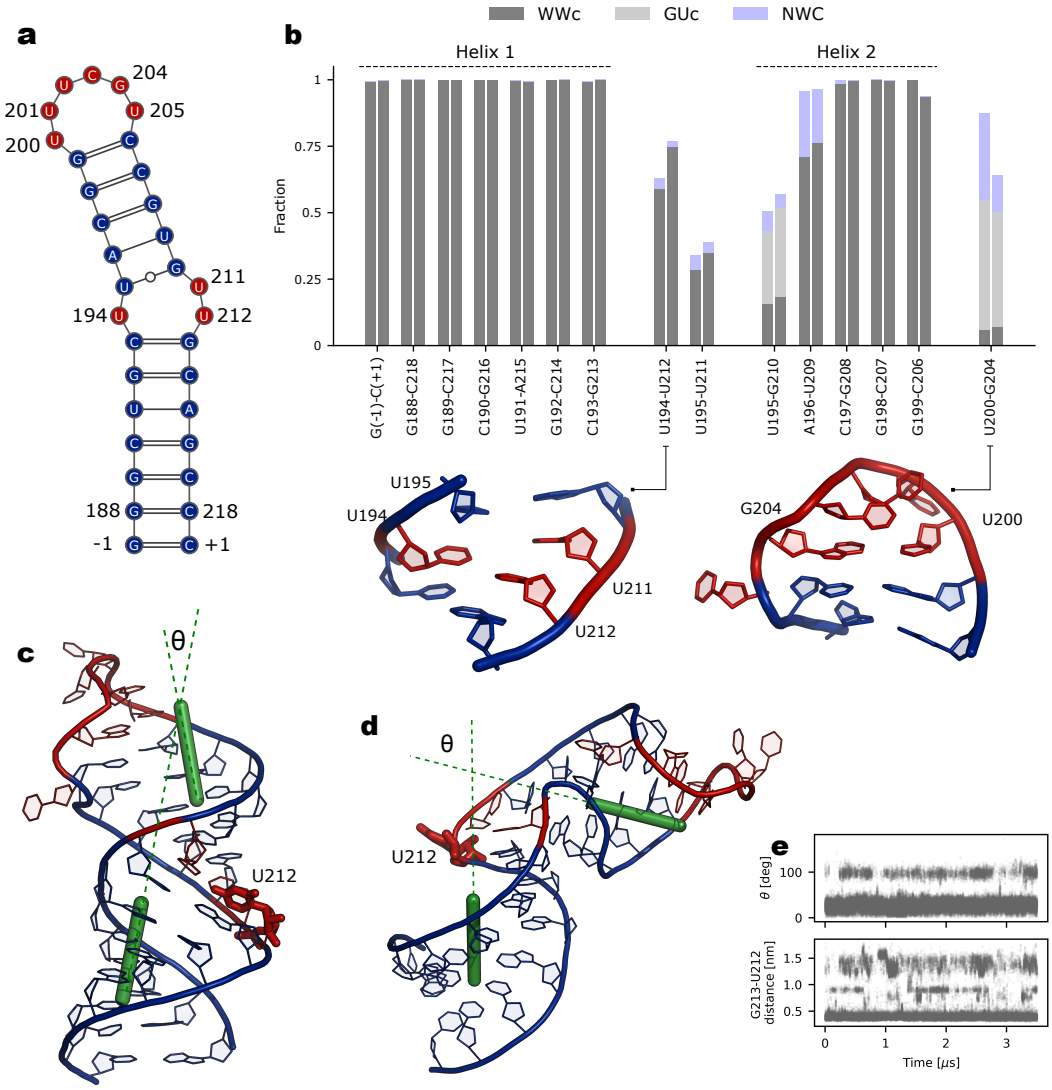

Figure 3. SL5a results. **a** Sequence and secondary structure of SL5a. The genomic numbering follows the convention in [11]. The region used for constructing the A-form templates is shown in blue. The sampling in the region shown in red is enhanced by partial tempering. **b** Population of base-pairs from 20 replicas × 3.6 μs partial tempering simulations. Centroids of the most populated internal and apical loop structures discussed in the text are shown. **c** Representative straight conformation. The axes defined by helix 1 and 2, together with the angle θ between them are shown in green. The helical axes are calculated as the vector normal to the base-pair plane, averaged over all base-pairs within the helix. **d** Representative kinked conformation. **e** Time series of the angle between helices, θ, (top panel) and of the distance between U212 and G213 (bottom panel) in the neutral replica. We note that due to the replica-exchange setup used in our simulations, these simulations do not reflect the timescales needed to switch between the two conformations.

motion does not affect the structures of the individual stems, but only their relative orientation. We also note that kinked conformations are observed only when U212 is solvent exposed, and therefore not engaged in base-pairing interactions with U194 or U195 (Fig. 3e), and suggest that the relative populations could be modulated by binding of proteins or small molecules to this motif.

Stem loop 1

Similarly to SL5a, SL1 consists of two helices connected by an asymmetrical internal loop (Fig. 4a). We use the pulling procedure to initialise our runs from the previously determined secondary structure [11]. While U13-A26/U17-A22 base pairs could not be confirmed by NMR, we have included these two interactions for generating the initial structures as they were present in previous consensus secondary structure assignments [14]. Note that secondary structure restraints are absent during partial tempering simulations, and we indeed observe a rich dynamics involving the loop regions as well as the neighboring residues.

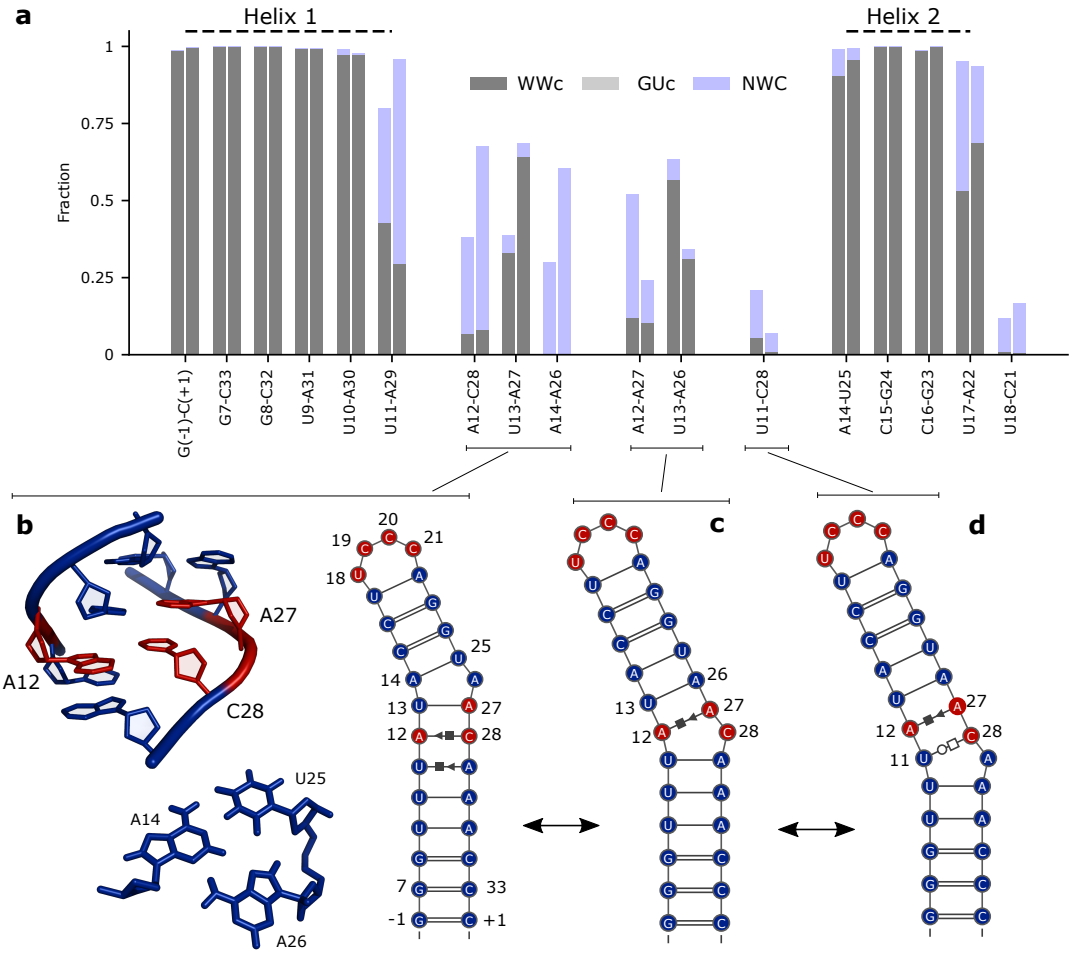

Figure 4. SL1 results. **a** Population of base-pairs from partial tempering simulations. **b** Three-dimensional structure of the most populated internal loop configuration showing the U11-A29 Hogsteen-Sugar, A12-C28 Sugar-Hoogsteen base-pairs, as well as the base A14-U25-U26 base triple. The corresponding secondary structure is shown on the right. The genomic numbering follows the convention in [11]. The base-paired regions selected for creating initial configurations is shown in blue, while partial tempering acts in regions shown in red. **c-d** Secondary structures of two alternative internal loop arrangements.

While helix 1 is stable, the terminal base-pair U11-A29 either interacts through the Watson-Crick/Watson-Crick or Hoogsteen-sugar edges (Fig. 4b). The presence of this non-canonical interaction allows the formation of a A12-C28 sugar-Hoogsteen interaction as well as of a U13-A27

base-pair. A26 is not completely solvent-exposed, but forms a base triple together with A14 and U25. In our simulations we also observe two alternative, lowly populated secondary structures (Fig. 4c-d) where A12-A27/U13-A26 are base-paired, and either C28 or A29 are partially excluded from the stack. Qualitatively, the nucleobase arrangements shown in Fig. 4b-d all preserve the coaxial stacking of the two helices.

The apical loop is highly dynamic and lacks a major, stable conformation with specific base-base interactions. U17-A22 base-pair interconverts from Watson-Crick/Watson-Crick to sugar/Hoogsteen, and U18 can engage a sugar-Hogsteen interaction with C21, albeit with a small population (Fig SI3).

Stem loop 4

The SL4 element has a more complex organisation, and consists of four A-form helices of different lengths linked by short pyrimidine internal loops and capped by a `UGCAU` pentaloop (Fig. 5a). The experimentally-derived secondary structure is generally preserved in the parallel tempering simulations (Fig. 5b). The A-form geometry between helix 1 and helix 2 is stabilised by a WC/WC C92-C119 base-pair (Fig. 5b). U95 is solvent-exposed and transiently engages in non-canonical interactions with G94 (Fig. 5b).

While the interactions observed in the first and second loops are relatively consistent between the two independent simulations, this is not the case for the the third internal loop. Here, the two simulations behave differently suggesting slow motions that are not fully sampled in our simulations. In the first run C100 preferentially base-pairs with U112, while in the second run additional configurations are sampled, where either C100 or U112 are not in the stack. The partial overlap between the two runs, that we ascribe to insufficient sampling, is further confirmed by a principal component analysis of this internal loop region (SI4), as well as from the low number of round trips in temperature space in run 1 (SI5). Experimental NMR measurements on SL4 agree with A-form helix compatible arrangement of the opposing residues C100 and U112 [11], thus supporting the presence of pairing between the two bases. The apical loop structure resembles a tetraloop with U108 solvent-exposed and U104:A107 forming either a Watson-Crick/Watson-Crick or a sugar/Hogsteen non-canonical interaction, as shown in Fig. 5b.

SL2+SL3

We here report the results on the SL2+SL3 construct, consisting of two consecutive stem-loop structures: SL2 (described above) and SL3, which contains the important transcription-regulating sequence required for subgenomic viral RNA synthesis [37]. Although we did not enforce this geometry, we find structures where the two helices are stacked coaxially tail-to-tail, thereby forming a unique A-form-like stem (Fig. 6).

The SL2 loop conformations are similar to those observed for SL2 alone, with a stable C50-G53 base-pair (see above). The `UCUAAAC` heptaloop in SL3 is highly flexible, and adopts several conformations that are partially structured in terms of base-pairing (Fig. 6). A cluster analysis of SL3 conformations reveals the presence of a loop structure where U65 forms a stacking interaction with C66, followed by the strand inversion and by two consecutive stackings between A68-A69-A70. This specific loop architecture can be considered the most RNA-like sampled in simulations, as we found it to be remarkably similar to Helix 58 loop in the archeal large ribosomal subunit [54] as well as to the anticodon loop in a nonproteinogenic tRNA [55] (Fig. SI6).

Helix 2, containing the transcription regulatory sequence `ACGAAC` is fully stable in our simulations. NMR measurements show that helix 2 is less stable than helix 1 in the absence of magnesium [11]. Such decreased stability of helix 2 is in line with a model for the body-to-leader fusion mechanism, in which the leader transcription regulatory sequence, 'hidden' inside SL3, is required to unfold to be active [3]. The tail-to-tail structure observed in simulations, which may not form in the context of the full 5'-UTR construct, may overstabilise helix 2 in this construct. Note that quantifying duplex stability is a challenging problem that is not considered here, as it usually requires dedicated simulation protocols [47].

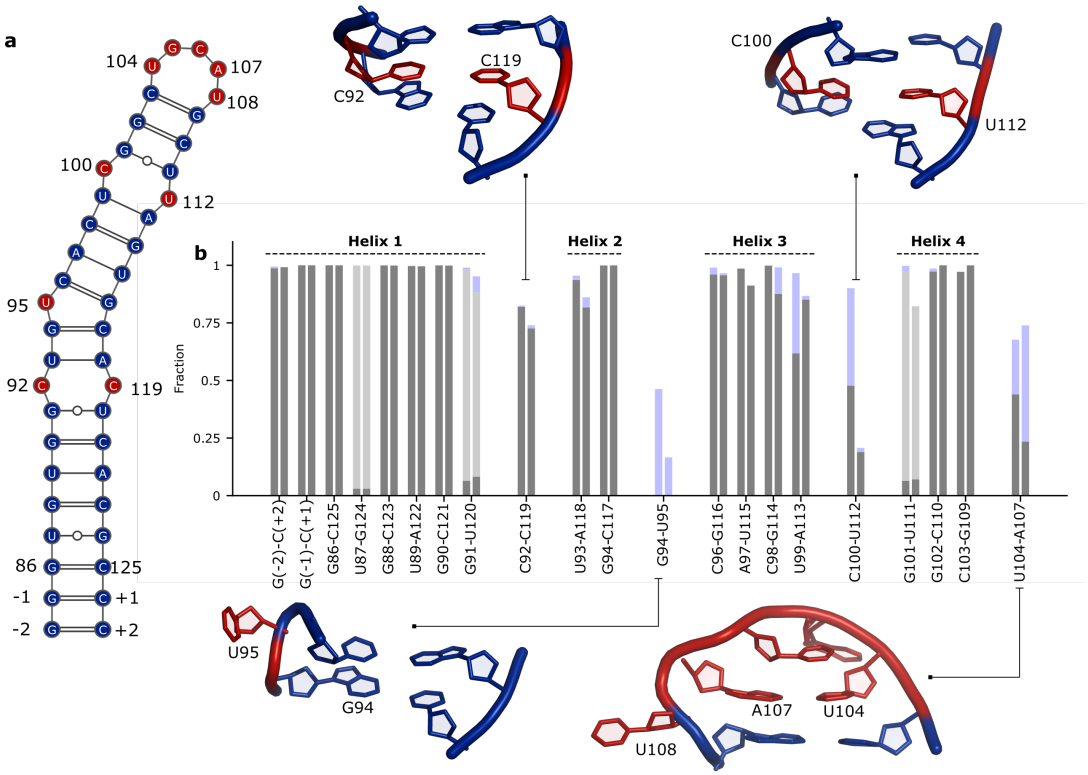

Figure 5. SL4 results. **a** Sequence and secondary structure of SL4. The genomic numbering follows the convention in [11]. The region used for constructing the A-form templates is shown in blue. The sampling in the region shown in red is enhanced by partial tempering. **b** Population of base-pairs from 20 replicas × 1.6 μs partial tempering simulations. Centroid of the most populated internal and apical loops discussed in the text are shown.

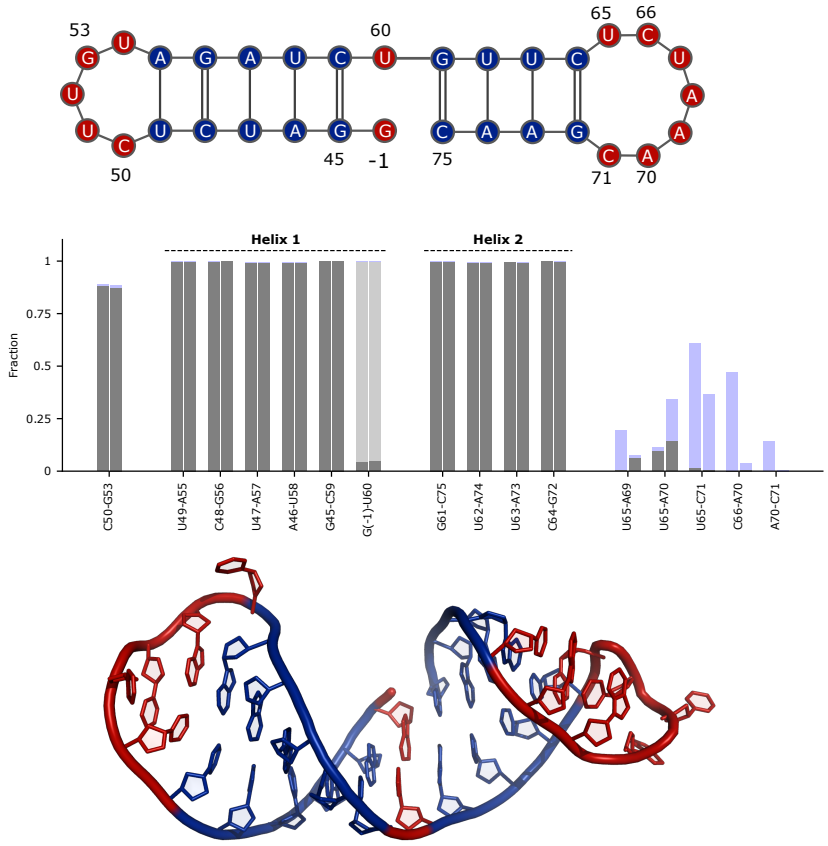

Figure 6. SL2+SL3 results. Top: Sequence and secondary structure of SL2+SL3. The genomic numbering follows the convention in [11]. The region used for constructing the A-form templates is shown in blue. The sampling in the region shown in red is enhanced by partial tempering. Middle panel: Population of base-pairs from partial tempering simulations. Bottom panel: Centroid of the most populated cluster.

Conclusions

We report the prediction of the structure and conformational heterogeneity of five stem-loop elements, ranging in size between 17 and 44 nucleotides, located at the 5'-UTR in SARS-CoV-2 genome. As a starting point for our study we use the secondary structure determined by NMR experiments [11].

Structures of several of these elements have also been predicted using FARFAR2 [17]. A principal component analysis of the structures generated for SL1 and SL2 shows small overlap between FARFAR2 and MD ensembles, suggesting the two methods to be complementary (SI7). In the case of SL2, the MD predictions mostly agree with the available NMR data, albeit small discrepancies remain. For this structure FARFAR2 does not predict the experimental loop architecture with the C50-G53 base-pair (SI1), and it thus appears that the ensemble generated by MD provides a more accurate description in this case. Given the scarcity of available experimental data on the 5'-UTR elements it is not possible at this stage to compare the accuracy of different computational methods. Indeed, FARFAR2 has been shown to better predict experimental data compared to MD on an RNA internal loop [28].

As in most MD studies, the results presented in this work depend on the amount of sampling. We took care in identifying and discussing states that are visited multiple times during one simulation and in both replicates, indicating these conformations to correspond to *bona fide*, highly populated metastable states. Moreover, the overall high agreement between the two independent simulations suggests that the conformational ensembles are relatively well converged. Our simulations of SL4, consisting of 44 nucleotides, represents an important exception with greater deviations between the two simulations, in particular of the third internal loop (SI4, SI5), indicating that our approach would require larger computational resources for molecules of this size.

The modelling protocol used in the present study implicitly considers helix formation and internal loop dynamics as partially decoupled events. While this is a relatively common assumption in computational studies of RNA [18, 31, 27], we can envisage situations in which the rearrangement of a loop requires one or more base pairs to break transiently. In our partial tempering MD simulations the secondary structure is not fixed: these rare events can thus be observed, albeit with a low frequency due to the absence of auxiliary replicas enhancing the breakage and formation of stem regions.

Despite recent improvements of RNA force-fields [56, 47, 43], the accuracy of the simulation results still depend on the accuracy of the physical models used, and we have validated our simulations against available experimental data. As more data becomes available, we note that it should be straightforward to integrate our extensive conformational sampling with data from future solution measurements (e.g. NMR or SAXS) [38, 39, 40] to refine the atomic-level description of SARS-CoV-2 RNA structure and dynamics.

Methods

MD simulations were conducted using the Amber ff99SB force field with parmbsc0 [57] and OL3 [58] corrections in conjunction with OPC water model [59]. The initial conformations were minimised in vacuum and then solvated in a dodecahedral box with minimum distance from box edge of 1.2nm. Potassium ions were added to neutralise the box [60]. The systems were equilibrated in the NVT and NPT ensembles at 1 bar for 1 ns. Production runs were performed in the canonical ensemble using a stochastic velocity rescaling thermostat [61]. All bonds were constrained using the LINCS [62] algorithm and equations of motion were integrated with a time step of 2 fs. Simulations were performed using GROMACS 2019.4 [63] patched with PLUMED 2.5.3 [64]. Force-field parameters are available at http://github.com/srnas/ff/tree/opc-water-model/amber_na.ff.

Pulling simulations

The aim of the pulling simulations is quickly to generate three-dimensional RNA conformations with a specific input secondary structure. We adapt a procedure originally introduced for reconstructing

atomistic structures from coarse-grained models [25]. Our protocol consists of three stages, as illustrated in Fig. 2a-b for SL2:

1. Plain MD simulation initialised from extended, single stranded, A-form RNA structures.
2. Pulling. We construct a reference structure with an ideal A-form double stranded RNA template for regions with known secondary structure (shown in blue in Fig. 2a) and perform pulling simulations to minimise the eRMSD [45] distance between the RNA molecule and the reference. We here use adiabatic bias MD simulations [65], that introduce a biasing potential damping the fluctuations only when the system moves further away from the template. In this way the pulling procedure follows the thermal fluctuations of the system, and does not require a pre-defined scheduling for the moving restraint [25]. When more than one stem is present, the bias acts on multiple templates simultaneously. The constant of the harmonic damping bias is set to 2 kJ/mol.
3. Plain MD. After 40ns of pulling simulation, the external bias is removed and the system freely fluctuates without external forces.

At the end of the procedure we extract samples from the final plain MD stage satisfying the secondary structure constraints, i.e. with eRMSD from template < 0.75. This criterion guarantees that all base-pairs corresponding to those in the target structure are correctly formed (Fig. 2b, bottom panel). Note that individual pulling simulations may fail: stems are sometimes not formed correctly and unfold during the final plain MD stage (Fig. SI8). The success rate of this procedure ranges from $\approx 80\%$ for SL2 to $\approx 10\%$ for SL4, which consists of 4 stem structures, consistent with previously reported benchmarks [25]. During preliminary tests we have observed that the formation of the correct secondary structure is hampered by early-forming non-native contacts. For this reason we conduct pulling simulations at 340K.

Partial tempering

We resolvate and minimise the RNA structures obtained from pulling as described above. We use partial tempering to enhance sampling in regions with unknown secondary structure [66, 26]. One reference replica exchanges via the Metropolis-Hastings algorithm with parallel simulations that are conducted with a modified Hamiltonian. The modification is such that interactions in the cold region (blue and grey in Fig. 2a, together with ions and water molecules) are kept at the reference temperature T=300K. Interactions in the hot region (shown in red in Fig. 2a) are rescaled by a factor $1/\lambda$, while the interactions between cold and hot regions are at an intermediate effective temperature of $T/\sqrt{\lambda}$, with $1 \leq \lambda < 0$. In this study we use 20 geometrically distributed replicas in the effective temperature range of 300 to 1000 K, and attempting exchanges every 240 steps. Hot regions are shown in red in Figs. 2-6, and include the phosphate group at the 3'-end of the selected region. We have found this small addition to be beneficial in preliminary test runs. System length, simulation time per replica, average acceptance rate, and number of round trips in temperature space are reported in Table 1, and additional statistics are available in Fig. SI5. To help the identification of slow degrees of freedom and assess convergence we conduct two independent runs with different initial conformations obtained from two separate pulling simulations (Fig. SI9).

Analysis and visualisation

Ideal A-form structures were constructed using the nucleic acids builder (NAB) software in Ambertools [67]. Trajectories were analysed using PLUMED [64] and MDtraj [68]. Base-pair annotation and puckering angles were calculated with Barnaba [50].

Supporting Information

1. Root mean squared deviation between PDB:2L6I, MD and FARFAR2 simulations.
2. NOE measurements violated in partial tempering simulations.
3. Three-dimensional centroids of the first five clusters in SL1 apical loop.

System	Lenght (nt)[(a)]	Simulation time [μs][(b)]	Acceptance rate [%] [(c)]	Round trips [(d)]
SL1	29	3.6	15-41	44-191
SL2	17	2.6	26-51	220-435
SL2+SL3	32	1.6	4-23	0-10
SL4	44	1.9	7-30	0-20
SL5a	33	3.6	12-34	6-33

Table 1. Simulated systems and related statistics. (a) number of nucleotides (nt). (b) Per-replica simulation time. Each simulation consists of 20 replicas and two independent runs for each system are carried out. The first 100ns of each run are discarded during analysis. (c) Minimum and maximum acceptance rate across all replicas and runs (d) Minimum and maximum number of round trips in temperature space across all replicas and runs.

4. Partial tempering simulations of SL4 projected onto the first two principal components.
5. Average acceptance rate and number of round trips in temperature space in partial tempering simulations.
6. SL3 structure similar to experimental loop structures in the PDB database.
7. Principal component analysis performed on FARFAR2 and MD ensembles for SL1 and SL2.
8. Example of unsuccessful pulling simulation.
9. Initial configurations in partial tempering simulations.

Acknowledgements

We acknowledge Drs. Harald Schwalbe, Anna Wacker, Julia E. Weigand, Andreas Schlundt, and other members of the Covid19-NMR consortium for discussions and insights. SB and KLL acknowledge funding from the Lundbeck Foundation BRAINSTRUC structural biology initiative (R155-2015-2666). We acknowledge access to computational resources from PRACE for the COVID-RNA project (COVID19-72).

Data Availability

Analysis scripts, MD trajectories, and input files are available via https://github.com/KULL-Centre/papers/tree/master/2020/COVID_5UTR_MD PLUMED input files are available in the PLUMED NEST [69] under the accession code 20.034 https://www.plumed-nest.org/eggs/20/034/.

References

[1] D. Kim, J.-Y. Lee, J.-S. Yang, J. W. Kim, V. N. Kim, and H. Chang, "The architecture of sars-cov-2 transcriptome," *Cell*, vol. 181, no. 4, pp. 914–921, 2020.

[2] S. K. Lal, *Molecular Biology of the SARS-Coronavirus*. Springer Science & Business Media, 2010.

[3] D. Yang and J. L. Leibowitz, "The structure and functions of coronavirus genomic 3 and 5 ends," *Virus Research*, vol. 206, pp. 120–133, 2015.

[4] S.-C. Chen and R. C. Olsthoorn, "Group-specific structural features of the 5'-proximal sequences of coronavirus genomic rnas," *Virology*, vol. 401, no. 1, pp. 29–41, 2010.

[5] R. Madhugiri, N. Karl, D. Petersen, K. Lamkiewicz, M. Fricke, U. Wend, R. Scheuer, M. Marz, and J. Ziebuhr, "Structural and functional conservation of cis-acting rna elements in coronavirus 5'-terminal genome regions," *Virology*, vol. 517, pp. 44–55, 2018.

[6] K. Schubert, E. D. Karousis, A. Jomaa, A. Scaiola, B. Echeverria, L.-A. Gurzeler, M. Leibundgut, V. Thiel, O. Mühlemann, and N. Ban, "Sars-cov-2 nsp1 binds the ribosomal mrna channel to inhibit translation," *Nature structural & molecular biology*, vol. 27, no. 10, pp. 959–966, 2020.

[7] I. Sola, F. Almazan, S. Zuniga, and L. Enjuanes, "Continuous and discontinuous rna synthesis in coronaviruses," *Annual review of virology*, vol. 2, pp. 265–288, 2015.

[8] C. Iserman, C. A. Roden, M. A. Boerneke, R. S. Sealfon, G. A. McLaughlin, I. Jungreis, E. J. Fritch, Y. J. Hou, J. Ekena, C. A. Weidmann, *et al.*, "Genomic rna elements drive phase separation of the sars-cov-2 nucleocapsid," *Molecular Cell*, pp. 1078–1091, 2020.

[9] C. R. Carlson, J. B. Asfaha, C. M. Ghent, C. J. Howard, N. Hartooni, M. Safari, A. D. Frankel, and D. O. Morgan, "Phosphoregulation of phase separation by the sars-cov-2 n protein suggests a biophysical basis for its dual functions," *Molecular Cell*, pp. 1092–110, 2020.

[10] N. C. Huston, H. Wan, M. S. Strine, R. d. C. A. Tavares, C. B. Wilen, and A. M. Pyle, "Comprehensive in vivo secondary structure of the sars-cov-2 genome reveals novel regulatory motifs and mechanisms," *Molecular cell*, vol. 81, no. 3, pp. 584–598, 2021.

[11] A. Wacker, J. E. Weigand, S. R. Akabayov, N. Altincekic, J. K. Bains, E. Banijamali, O. Binas, J. Castillo-Martinez, E. Cetiner, B. Ceylan, *et al.*, "Secondary structure determination of conserved sars-cov-2 rna elements by nmr spectroscopy," *Nucleic Acids Research*, 2020.

[12] R. J. Andrews, J. M. Peterson, H. S. Haniff, J. Chen, C. Williams, M. Grefe, M. D. Disney, and W. N. Moss, "An in silico map of the sars-cov-2 rna structurome," *BioRxiv*, 2020.

[13] I. Manfredonia, C. Nithin, A. Ponce-Salvatierra, P. Ghosh, T. K. Wirecki, T. Marinus, N. S. Ogando, E. J. Snijder, M. J. van Hemert, J. M. Bujnicki, *et al.*, "Genome-wide mapping of sars-cov-2 rna structures identifies therapeutically-relevant elements," *Nucleic Acids Research*, 2020.

[14] R. Rangan, I. N. Zheludev, R. J. Hagey, E. A. Pham, H. K. Wayment-Steele, J. S. Glenn, and R. Das, "Rna genome conservation and secondary structure in sars-cov-2 and sars-related viruses: a first look," *RNA*, vol. 26, no. 8, pp. 937–959, 2020.

[15] Z. Miao, A. Tidu, G. Eriani, and F. Martin, "Secondary structure of the sars-cov-2 5'-utr," *RNA Biology*, pp. 1–10, 2020.

[16] C. W. Lee, L. Li, and D. P. Giedroc, "The solution structure of coronaviral stem-loop 2 (sl2) reveals a canonical cuyg tetraloop fold," *FEBS Letters*, vol. 585, no. 7, pp. 1049–1053, 2011.

[17] R. Rangan, A. M. Watkins, J. Chacon, R. Kretsch, W. Kladwang, I. N. Zheludev, J. Townley, M. Rynge, G. Thain, and R. Das, "De novo 3d models of sars-cov-2 rna elements from consensus experimental secondary structures," *Nucleic Acids Research*, 2021.

[18] A. M. Watkins, R. Rangan, and R. Das, "Farfar2: Improved de novo rosetta prediction of complex global rna folds," *Structure*, pp. 963–976, 2020.

[19] K. Zhang, I. N. Zheludev, R. J. Hagey, M. T.-P. Wu, R. Haslecker, Y. J. Hou, R. Kretsch, G. D. Pintilie, R. Rangan, W. Kladwang, *et al.*, "Cryo-electron microscopy and exploratory antisense targeting of the 28-kda frameshift stimulation element from the sars-cov-2 rna genome," *Biorxiv*, 2020.

[20] C. Bergonzo and A. L. Szakal, "Using all-atom potentials to refine rna structure predictions of sars-cov-2 stem loops," *International Journal of Molecular Sciences*, vol. 21, no. 17, p. 6188, 2020.

[21] L. R. Ganser, M. L. Kelly, D. Herschlag, and H. M. Al-Hashimi, "The roles of structural dynamics in the cellular functions of rnas," *Nature reviews Molecular cell biology*, vol. 20, no. 8, pp. 474–489, 2019.

[22] T. R. Alderson and L. E. Kay, "Nmr spectroscopy captures the essential role of dynamics in regulating biomolecular function," *Cell*, vol. 184, no. 3, pp. 577–595, 2021.

[23] M. Parisien and F. Major, "The mc-fold and mc-sym pipeline infers rna structure from sequence data," *Nature*, vol. 452, no. 7183, pp. 51–55, 2008.

[24] C. Y. Cheng, F.-C. Chou, and R. Das, "Modeling complex rna tertiary folds with rosetta," in *Methods in Enzymology*, vol. 553, pp. 35–64, Elsevier, 2015.

[25] S. Poblete, S. Bottaro, and G. Bussi, "Effects and limitations of a nucleobase-driven backmapping procedure for nucleic acids using steered molecular dynamics," *Biochemical and Biophysical Research Communications*, vol. 498, no. 2, pp. 352–358, 2018.

[26] P. Ebrahimi, S. Kaur, L. Baronti, K. Petzold, and A. A. Chen, "A two-dimensional replica-exchange molecular dynamics method for simulating rna folding using sparse experimental restraints," *Methods*, vol. 162, pp. 96–107, 2019.

[27] M. Y. Angela, P. M. Gasper, L. Cheng, L. B. Lai, S. Kaur, V. Gopalan, A. A. Chen, and J. B. Lucks, "Computationally reconstructing cotranscriptional rna folding from experimental data reveals rearrangement of non-native folding intermediates," *Molecular Cell*, vol. 81, no. 4, pp. 870–883, 2021.

[28] H. Shi, A. Rangadurai, H. Abou Assi, R. Roy, D. A. Case, D. Herschlag, J. D. Yesselman, and H. M. Al-Hashimi, "Rapid and accurate determination of atomistic rna dynamic ensemble models using nmr and structure prediction," *Nature communications*, vol. 11, no. 1, pp. 1–14, 2020.

[29] L. Wang, R. A. Friesner, and B. Berne, "Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (rest2)," *The Journal of Physical Chemistry B*, vol. 115, no. 30, pp. 9431–9438, 2011.

[30] M. Kamiya and Y. Sugita, "Flexible selection of the solute region in replica exchange with solute tempering: Application to protein-folding simulations," *The Journal of chemical physics*, vol. 149, no. 7, p. 072304, 2018.

[31] P. Chyży, M. Kulik, S. Re, Y. Sugita, and J. Trylska, "Mutations of n1 riboswitch affect its dynamics and recognition by neomycin through conformational selection," *Frontiers in molecular biosciences*, vol. 8, p. 12, 2021.

[32] R. Schnieders, S. A. Peter, E. Banijamali, M. Riad, N. Altincekic, J. K. Bains, B. Ceylan, B. Fürtig, J. T. Grün, M. Hengesbach, *et al.*, "1 h, 13 c and 15 n chemical shift assignment of the stem-loop 5a from the 5'-utr of sars-cov-2," *Biomolecular NMR assignments*, pp. 1–9, 2021.

[33] M. Novakovic, E. Kupce, T. Scherf, A. Oxenfarth, R. Schnieders, T. Grün, J. Wirmer-Bartoschek, C. Richter, H. Schwalbe, and L. Frydman, "Magnetization transfer to enhance noe cross-peaks among labile protons: Applications to imino-imino sequential walks in sars-cov-2-derived rnas," *Angewandte Chemie International Edition*, 2021.

[34] D. Escors, A. Izeta, C. Capiscol, and L. Enjuanes, "Transmissible gastroenteritis coronavirus packaging signal is located at the 5' end of the virus genome," *Journal of Virology*, vol. 77, no. 14, pp. 7890–7902, 2003.

[35] A. Tidu, A. Janvier, L. Schaeffer, P. Sosnowski, L. Kuhn, P. Hammann, E. Westhof, G. Eriani, and F. Martin, "The viral protein nsp1 acts as a ribosome gatekeeper for shutting down host translation and fostering sars-cov-2 translation," *RNA*, vol. 27, no. 3, pp. 253–264, 2021.

[36] D. Yang, P. Liu, D. P. Giedroc, and J. Leibowitz, "Mouse hepatitis virus stem-loop 4 functions as a spacer element required to drive subgenomic rna synthesis," *Journal of virology*, vol. 85, no. 17, pp. 9199–9209, 2011.

[37] D. Dufour, P. A. Mateos-Gomez, L. Enjuanes, J. Gallego, and I. Sola, "Structure and functional relevance of a transcription-regulating sequence involved in coronavirus discontinuous rna synthesis," *Journal of Virology*, vol. 85, no. 10, pp. 4963–4973, 2011.

[38] A. Cesari, S. Reißer, and G. Bussi, "Using the maximum entropy principle to combine simulations and solution experiments," *Computation*, vol. 6, no. 1, p. 15, 2018.

[39] S. Bottaro and K. Lindorff-Larsen, "Biophysical experiments and biomolecular simulations: A perfect match?," *Science*, vol. 361, no. 6400, pp. 355–360, 2018.

[40] S. Bottaro, T. Bengtsen, and K. Lindorff-Larsen, "Integrating molecular simulation and experimental data: a bayesian/maximum entropy reweighting approach," in *Structural Bioinformatics*, pp. 219–240, Springer, 2020.

[41] A. C. Stelzer, A. T. Frank, J. D. Kratz, M. D. Swanson, M. J. Gonzalez-Hernandez, J. Lee, I. Andricioaei, D. M. Markovitz, and H. M. Al-Hashimi, "Discovery of selective bioactive small molecules by targeting an rna dynamic ensemble," *Nature Chemical Biology*, vol. 7, no. 8, p. 553, 2011.

[42] M. G. Costales, J. L. Childs-Disney, H. S. Haniff, and M. D. Disney, "How we think about targeting rna with small molecules," *Journal of medicinal chemistry*, vol. 63, no. 17, pp. 8880–8900, 2020.

[43] P. D. Dans, D. Gallego, A. Balaceanu, L. Darré, H. Gómez, and M. Orozco, "Modeling, simulations, and bioinformatics at the service of rna structure," *Chem*, vol. 5, no. 1, pp. 51–73, 2019.

[44] R. E. Amaro and A. J. Mulholland, "A community letter regarding sharing biomolecular simulation data for covid-19," *Journal of chemical information and modeling*, vol. 60, no. 6, pp. 2653–2656, 2020.

[45] S. Bottaro, F. Di Palma, and G. Bussi, "The role of nucleobase interactions in RNA structure and dynamics," *Nucleic Acids Research*, vol. 42, no. 21, pp. 13306–13314, 2014.

[46] N. B. Leontis and E. Westhof, "Geometric nomenclature and classification of rna base pairs.," *RNA*, vol. 7, no. 4, pp. 499–512, 2001.

[47] J. Šponer, G. Bussi, M. Krepl, P. Banáš, S. Bottaro, R. A. Cunha, A. Gil-Ley, G. Pinamonti, S. Poblete, P. Jurečka, N. G. Walter, and M. Otyepka, "RNA structural dynamics as captured by molecular simulations: a comprehensive overview," *Chemical Reviews*, vol. 118, no. 8, pp. 4177–4338, 2018.

[48] F. M. Jucker and A. Pardi, "Solution structure of the cuug hairpin loop: a novel rna tetraloop motif," *Biochemistry*, vol. 34, no. 44, pp. 14416–14427, 1995.

[49] S. Bottaro and K. Lindorff-Larsen, "Mapping the universe of rna tetraloop folds," *Biophysical Journal*, vol. 113, no. 2, pp. 257–267, 2017.

[50] S. Bottaro, G. Bussi, G. Pinamonti, S. Reißer, W. Boomsma, and K. Lindorff-Larsen, "Barnaba: software for analysis of nucleic acid structures and trajectories," *RNA*, vol. 25, no. 2, pp. 219–231, 2019.

[51] C. Cheong, G. Varani, and I. Tinoco, "Solution structure of an unusually stable rna hairpin, 5ggac (uucg) gucc," *Nature*, vol. 346, no. 6285, pp. 680–682, 1990.

[52] K. Mráziková, V. Mlynsky, P. Kuhrova, P. Pokorná, H. Kruse, M. Krepl, M. Otyepka, P. Banas, and J. Sponer, "Uncg rna tetraloop as a formidable force-field challenge for md simulations," *bioRxiv*, 2020.

[53] D. Klein, T. Schmeing, P. Moore, and T. Steitz, "The kink-turn: a new rna secondary structure motif," *The EMBO journal*, vol. 20, no. 15, pp. 4214–4221, 2001.

[54] N. Ban, P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz, "The complete atomic structure of the large ribosomal subunit at 2.4 å resolution," *Science*, vol. 289, no. 5481, pp. 905–920, 2000.

[55] A. T. Chang and E. P. Nikonowicz, "Solution nuclear magnetic resonance analyses of the anticodon arms of proteinogenic and nonproteinogenic trnagly," *Biochemistry*, vol. 51, no. 17, pp. 3662–3674, 2012.

[56] S. Vangaveti, S. V. Ranganathan, and A. A. Chen, "Advances in rna molecular dynamics: a simulator s guide to rna force fields," *Wiley Interdisciplinary Reviews: RNA*, vol. 8, no. 2, p. e1396, 2017.

[57] A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham III, C. A. Laughton, and M. Orozco, "Refinement of the AMBER force field for nucleic acids: Improving the description of alpha gamma conformers.," *Biophysical Journal*, vol. 92, no. 11, pp. 3817–3829, 2007.

[58] M. Zgarbová, M. Otyepka, J. Sponer, A. Mlaadek, P. Banas, T. E. Cheatham III, and P. Jurecka, "Refinement of the cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles," *Journal of Chemical Theory and Computation*, vol. 7, no. 9, pp. 2886–2902, 2011.

[59] S. Izadi, R. Anandakrishnan, and A. V. Onufriev, "Building water models: a different approach," *The Journal of Physical Chemistry Letters*, vol. 5, no. 21, pp. 3863–3871, 2014.

[60] I. S. Joung and T. E. Cheatham III, "Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations," *Journal of Physical Chemistry B*, vol. 112, no. 30, pp. 9020–9041, 2008.

[61] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling.," *Journal of Chemical Physics*, vol. 126, no. 1, p. 014101, 2007.

[62] B. Hess, H. Bekker, H. J. Berendsen, and J. G. Fraaije, "LINCS: a linear constraint solver for molecular simulations.," *Journal of Computational Chemistry*, vol. 18, no. 12, pp. 1463–1472, 1997.

[63] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindhal, "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit.," *Bioinformatics*, vol. 29, no. 7, pp. 845–854, 2013.

[64] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, "Plumed 2: New feathers for an old bird," *Computer Physics Communications*, vol. 185, no. 2, pp. 604–613, 2014.

[65] M. Marchi and P. Ballone, "Adiabatic bias molecular dynamics: a method to navigate the conformational space of complex molecular systems," *The Journal of Chemical Physics*, vol. 110, no. 8, pp. 3697–3702, 1999.

[66] G. Bussi, "Hamiltonian replica exchange in gromacs: a flexible implementation," *Molecular Physics*, vol. 112, no. 3-4, pp. 379–384, 2014.

[67] T. J. Macke and D. A. Case, "Modeling unusual nucleic acid structures," in *Molecular Modeling of Nucleic Acids*, pp. 379–393, ACS Publications, 1998.

[68] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande, "Mdtraj: a modern open library for the analysis of molecular dynamics trajectories," *Biophysical Journal*, vol. 109, no. 8, pp. 1528–1532, 2015.

[69] M. Bonomi, G. Bussi, C. Camilloni, G. A. Tribello, P. Banáš, A. Barducci, M. Bernetti, P. G. Bolhuis, S. Bottaro, D. Branduardi, *et al.*, "Promoting transparency and reproducibility in enhanced molecular simulations," *Nature Methods*, vol. 16, no. 8, pp. 670–673, 2019.