# Conformational heterogeneity of UCAAUC RNA oligonucleotide from molecular dynamics simulations, SAXS, and NMR experiments

CHRISTINA BERGONZO,[1] ALEXANDER GRISHAEV,[1] and SANDRO BOTTARO[2,3]

[1]National Institute of Standards and Technology and Institute for Bioscience and Biotechnology Research, Rockville, Maryland 20850, USA
[2]Structural Biology and NMR Laboratory, Department of Biology, University of Copenhagen, DK-2200 Copenhagen N, Denmark
[3]Department of Biomedical Sciences, Humanitas University, 20090 Pieve Emanuele, Italy

## ABSTRACT

**We describe the conformational ensemble of the single-stranded r(UCAAUC) oligonucleotide obtained using extensive molecular dynamics (MD) simulations and Rosetta's FARFAR2 algorithm. The conformations observed in MD consist of A-form-like structures and variations thereof. These structures are not present in the pool generated using FARFAR2. By comparing with available nuclear magnetic resonance (NMR) measurements, we show that the presence of both A-form-like and other extended conformations is necessary to quantitatively explain experimental data. To further validate our results, we measure solution X-ray scattering (SAXS) data on the RNA hexamer and find that simulations result in more compact structures than observed from these experiments. The integration of simulations with NMR via a maximum entropy approach shows that small modifications to the MD ensemble lead to an improved description of the conformational ensemble. Nevertheless, we identify persisting discrepancies in matching experimental SAXS data.**

**Keywords: integrative structural biology; molecular dynamics; NMR; SAXS**

## INTRODUCTION

An important step forward in understanding RNA structure is the consideration of ensembles as the best descriptors of flexible systems (Plumridge et al. 2019; Liu et al. 2021). Folding behavior and intermolecular interactions of single-strand RNA molecules have implications in biopharmaceutical drug development, including rational design and predicting delivery properties (Costales et al. 2020), as single stranded antisense oligonucleotides (ASOs) have been approved as drug products starting in 1998 (de Smet et al. 1999). To predict pharmacokinetic properties of native versus modified single stranded RNAs, advanced potential functions and molecular dynamics can be used in conjunction with a variety of biophysical characterization methods.

To this end, solution state nuclear magnetic resonance (NMR) spectroscopy provides an important tool to monitor the dynamics experienced by single-stranded RNAs. A key step in the interpretation of the NMR data is the site-specific resonance assignment, forming the basis of structural analysis. However, the determination of the single-stranded RNA structure, as well as its NMR spectral assignment utilizing nuclear Overhauser effect (NOE) connectivities, is complicated by the fact that oligonucleotides can adopt a variety of conformations which can be difficult to decouple into independent structure contributions to an ensemble (Tubbs et al. 2013; Condon et al. 2015).

Molecular dynamics has been used to aid the interpretation of experimental NMR data into structures and ensembles. Given an adequate amount of experimental data, or a more structured helical system, structure prediction yields good agreement with experimental observables (Bergonzo and Grishaev 2019). Though known to have limited accuracy in predicting complete Boltzmann weighted ensembles of structures (Sponer et al. 2018), MD simulations can be used to effectively explore conformational space and generate accurate structural predictions for well-described free energy minima (Roe et al. 2014).

The accuracy of MD ensembles can be improved by including experimental data in simulations, either a

**Corresponding authors: christina.bergonzo@nist.gov, sandro.bottaro@hunimed.eu**

Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.078888.121. Freely available online through the *RNA* Open Access option.

posteriori, as in reweighting or selection schemes, or during simulations (Ward et al. 2013; Hummer and Köfinger 2015; Bonomi et al. 2017). For flexible systems, it is crucial to consider experimental measurements as averages over multiple conformational states, and this should hold true in simulations as well.

Previous studies have shown how to obtain conformational ensembles of dynamic RNA by integrating solution NMR data with extensive MD simulations using maximum entropy approaches (Borkar et al. 2017; Reißer et al. 2020). Furthermore, through the use of this reweighting technique, it has been possible to identify inaccuracies both in the force field and in the experimental data (Bottaro et al. 2018). Force field improvements can be achieved by investigating systems that are large enough to display a high degree of structural complexity, yet simple enough to be amenable to extensive, converged simulations (Bottaro and Lindorff-Larsen 2018). To expand the repertoire of such systems, here we provide an atomic-level detailed structural description of the UCAAUC RNA hexamer oligonucleotide using extensive MD simulations in combination with available NMR (Zhao et al. 2020) and newly measured solution X-ray scattering (SAXS) experimental data. First, we find that two recent MD force fields, namely AMBER FFLJbb (Bergonzo and Cheatham 2015) with OPC water (Izadi et al. 2014) and ROC-RNA (Aytenfisu et al. 2017) with TIP3P water, both produce ensembles containing single-stranded, A-form-like structures but with different conformational preferences. Both ensembles share almost no overlap with the structures generated using Rosetta's FARFAR2 algorithm (Watkins et al. 2020). Second, we compare the computational results with solution experiments, including chemical shifts, $^{3}J$ scalar couplings, NOE, and SAXS data. We examine the complementary roles played by these experimental data types and show that MD force fields provide a more accurate description of the structure ensemble compared to FARFAR2. Finally, we integrate NMR data into simulations a posteriori using Bayesian maximum entropy reweighting, and describe a conformational ensemble that agrees with all available experimental data.

## RESULTS

### FFLJbb, ROC-RNA, and FARFAR2 produce different conformational ensembles

We compare three conformational ensembles of the UCAAUC RNA hexanucleotide obtained by extensive MD simulations using the FFLJbb (Bergonzo and Cheatham 2015) and ROC-RNA (Aytenfisu et al. 2017) force fields (see Materials and Methods), and using Rosetta's FARFAR2 (Watkins et al. 2020) algorithm. MD simulations of r(UCAAUC) with the widely used FFOL3 force field (Zgarbová et al. 2011) and TIP3P water have been shown to be less accurate compared to ROC-RNA (Zhao et al. 2020) and are thus not included in this study. MD ensembles were generated using multidimensional replica exchange MD, where temperature and biased dihedral force constant scaling were combined (Bergonzo et al. 2014). Two simulations were initiated from either A-form or extended linear structures, and throughout the manuscript we show results from each independent simulation. Convergence was analyzed by comparing the top five principal components as well as the results of cluster analysis and indicate that the independent simulations are sampling highly similar conformational populations (Supplemental Information 1 and 2).

FFLJbb populates conformations very close to the canonical A-form, as well as distant ones as judged by the histogram in Figure 1A. The structural dissimilarity is here evaluated using a nucleic acid-specific distance called ellipsoidal root-mean-square distance (eRMSD) that only considers the relative arrangements between nucleobases in a molecule (Bottaro et al. 2014). For ROC-RNA, we observe one large peak around eRMSD = 0.75, indicating this ensemble to be less heterogeneous compared to FFLJbb. In FARFAR2, instead, all structures are distant from A-form. The distribution of the radii of gyration (Fig. 1B) shows that FFLJbb is more extended compared to ROC-RNA and FARFAR2, and all three ensembles are on average more compact with respect to the ideal A-form (dashed line in Fig. 1B). To evaluate the degree of overlap between the ensembles, we show in Figure 1C a uMAP projection (McInnes et al. 2018) obtained by aggregating all samples. While FFLJbb and ROC-RNA partially overlap, FARFAR2 samples are located in a different region of this plane. By cluster analysis and visual inspection, we identified in the MD samples A-form-like structures (Aform), C6-inverted (C6-I) (Zhao et al. 2020), C6-bulged (C6-B), U5-bulged (U5-B), C2-bulged (C2-B), and U1-bulged (U1-B) conformations. These structures represent altogether ≈50% of the FFLJbb ensemble, with the remainder being intermediate or other lowly populated conformations. C6 inverted conformations are highly populated (75%) in ROC-RNA ensembles (Fig. 1D); A-form and C6 bulged structures are observed as well, but with lower populations. FARFAR2 ensemble, instead, is composed by a diverse set of compact structures with several intra-molecular interactions, none of them being similar to A-form-like structures (Supplemental Information 3).

### MD simulations better agree with NMR and SAXS measurements compared to FARFAR2

The conformational ensemble from MD simulations is in qualitative agreement with NMR data that suggest a dominant A-form state in equilibrium with C6-inverted, C6 bulged, and C2-bulged conformations (Zhao et al. 2020).

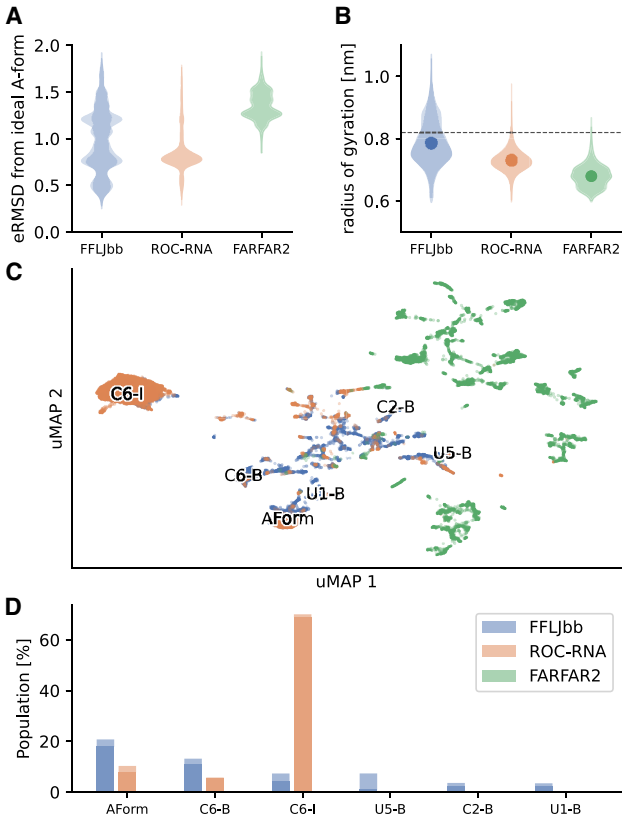


**FIGURE 1.** Structural analysis of r(UCAAUC) simulations. (*A*) eRMSD from ideal A-form histograms. For each ensemble, the distributions calculated for two replicates are shown in shade. (*B*) Radii of gyration distributions of the three ensembles, as labeled. Averages are shown as dots, and the *horizontal* dashed line indicates the radius of gyration of an ideal A-form structure. (*C*) MD simulations and FARFAR2 samples projected onto a uMAP 1/uMAP2 plane. The color scheme follows panels *A* and *B*: FFLJbb in blue, ROC-RNA in orange, and FARFAR2 in green. The structures discussed in the main text are labeled on the plane: A-form-like structures (Aform), C6-inverted (C6-I), C6-bulged (C6-B), U5-bulged (U5-B), C2-bulged (C2-B), and U1-bulged (U1-B) structures. (*D*) Population of different clusters, as labeled. The populations in FARFAR2 ensembles are zero for all clusters.

These structures are observed in the FFLJbb ensemble (Fig. 1C,D), and to a lesser extent in ROC-RNA as well. To quantify this agreement, we compare experimental data with ensemble averages calculated for different data sets: chemical shifts (CS), NOE, $^3$J scalar couplings, unobserved NOE (uNOE), ambiguous NOE (ambNOE) (Zhao et al. 2020), and SAXS. For each data type, we report the $\mathcal{X}^2$ calculated for two independent simulations in Figure 2 (see Materials and Methods section). Since $\mathcal{X}^2$ is defined as the average squared difference between calculated and experimental measurements normalized by the experimental errors (see Equation 3 in Materials and Methods), values below one are typically considered acceptable (Gull and Daniell 1978). Scatter plots showing experimental versus calculated averages for NMR data are shown in Supplemental Information 4–8. We find that FFLJbb is in better agreement with NOEs, scalar couplings, and SAXS data compared to ROC-RNA and FARFAR2, and that ROC-RNA better agrees with ambiguous and unobserved NOEs. Note also that the sensitivity, positive predictive values, and accuracy calculated from NMR data also suggest FFLJbb to be more accurate compared to ROC-RNA and FARFAR2, as shown in Table 1 (Zhao et al. 2020).

Note that a single, ideal A-form structure is in poor agreement with experimental data, except for unobserved NOEs. These results show that (i) a single A-form structure could not explain the NMR data, as previously described (Zhao et al. 2020) and (ii) at least a fraction of conformations in the computational ensembles do violate unobserved NOEs, indicating the presence of structures not compatible with experimental evidence.

FARFAR2 is a widely used RNA structure prediction algorithm that has been recently used to perform ensemble refinement of flexible regions with improved results compared to MD (Shi et al. 2020). While the agreement between FARFAR2 and experiments is poorer compared to FFLJbb and ROC-RNA, $\mathcal{X}^2$ do not significantly exceed unity for all NMR measurements. This is to some extent surprising, as A-form structures are completely absent from the FARFAR2 ensemble (Fig. 1A,C,D; Supplemental Information 3). We observe even larger deviations for

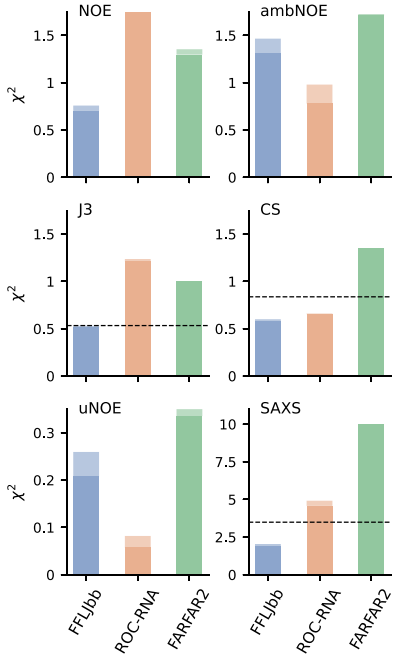


**FIGURE 2.** Agreement between NOE, $^3$J scalar couplings, unobserved NOE (uNOE), chemical shifts (CS), ambiguous NOE (ambNOE), and SAXS data, as labeled. As a reference, the $\mathcal{X}^2$ relative to the ideal A-form structure are 6.13 (NOE), 0.02 (uNOE), and 4.82 (ambNOE), while the values for CS, J3, and SAXS are shown as dashed lines. The two bars show the statistics calculated on two independent runs.

**TABLE 1.** Sensitivity, positive predictive value (PPV), and accuracy for different ensembles as defined in Zhao et al. (2020)

| | Sensitivity | PPV | Accuracy |
|---|---|---|---|
| A-FORM | 0.61 | 0.61 | 0.61 |
| FFLJbb | 0.77/0.76 | 0.73/0.70 | 0.75/0.73 |
| ROC-RNA | 0.67/0.67 | 0.66/0.66 | 0.66/0.66 |
| FARFAR2 | 0.64/0.63 | 0.61/0.60 | 0.63/0.62 |

Values from two independent simulations are reported.

SAXS data in the FARFAR2 ensemble, in agreement with the high degree of compaction of FARFAR2 structures (Fig. 1B). Note that here we are using FARFAR2 outside its original scope, as it has been designed to predict the structure of large molecules, and not to sample the equilibrium distribution of a flexible single-stranded RNA. Nonetheless, common applications of FARFAR2 include regions of possible conformational heterogeneity such as loops and bulges, which strongly affect the long-range RNA structure. Our observation of overcompactness, consistent with exaggerated formation of intra-RNA interactions, suggests the need for better validation of conformations generated via this approach.

Figure 3 shows the experimental and calculated SAXS intensities for the three ensembles. We find persistent discrepancies between the ensemble-predicted and measured data, particularly at the lowest scattering angles, suggesting the structures sampled in simulations to be overly compact. This result is consistent with previous computational studies (Tan et al. 2018).

Ensemble-averaged SAXS profiles calculated via FoXS (Schneidman-Duhovny et al. 2010), one of the most popular data modeling methods, exhibit differences with the measured data throughout the entire experimental resolution range, but particularly pronounced at lowest $q$ values. Indeed, the calculated radius of gyration (0.79 nm) is smaller than the experimental one (0.88 nm). SAXS-reported radii of gyration are affected by both the RNA coordinates and the structure of the surface solvent layer. With FoXS placing surface solvent directly on top of the surface-facing atoms rather than offset from them, its use could lead to underestimation of the predicted radii of gyration. To investigate this issue, we repeated SAXS data prediction via two additional techniques, both of which position surface solvent layer outside the volume occupied by the RNA—Crysol (Svergun et al. 1995) and AXES (Grishaev et al. 2010). Crysol uses a layer of uniform electron density to represent the surface solvent implicitly, while AXES uses all-atom molecular representation via frames from an MD simulation of a water-filled volume, with solvent represented explicitly. We find the agreement between the measured and ensemble-predicted data improving from FoXS to Crysol to AXES, with $\chi^2$ values decreasing by factors of 2.2, and 1.9, respectively. Since AXES leads to the best agreement between the measured and model-predicted data, we use it throughout the manuscript. The impact of the effects such as the distribution of the surface counter ions on the low-angle scattering data is minor as the differences between the SAXS-extracted radii of gyration for the buffers containing 150 mmol/L and 75 mmol/L NaCl do not exceed 0.01 nm.

## Conformational ensemble of r(UCAAUC) by integrating MD and NMR

The agreement with experimental data can be further improved by using reweighting techniques. In such approaches, the weight of each sampled set of coordinates is adjusted so that the resulting ensemble averages more closely match experimental measurements compared to the original ensemble. In previous studies, we have used the BME approach (Bottaro et al. 2020a) to provide an accurate description of structure and conformational heterogeneity of RNA tetranucleotides and tetraloops (Bottaro et al. 2018, 2020b). We use a similar approach here and integrate experimental data with simulations. In BME, as in many optimization procedures, the user needs to set the value of a regularization parameter ($\theta$). Small values of $\theta$ correspond to a better fit (small $\chi^2$), whereas in the limit of large $\theta$ one approaches the $\chi^2$ obtained using the original ensemble. At the same time, it is possible to monitor the $\chi^2$ relative to data that were not used for reweighting

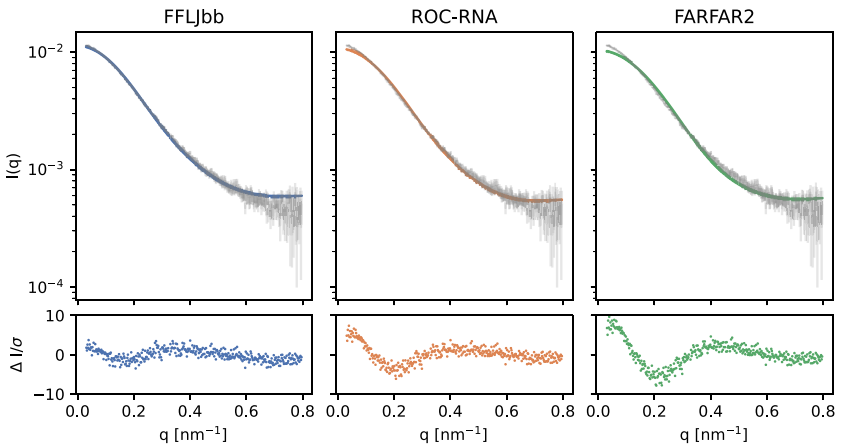


**FIGURE 3.** (*Top* panels) Experimental SAXS curve (gray) and average intensities calculated from the ensembles, as labeled. The normalized difference between the two is shown in the *bottom* panels.

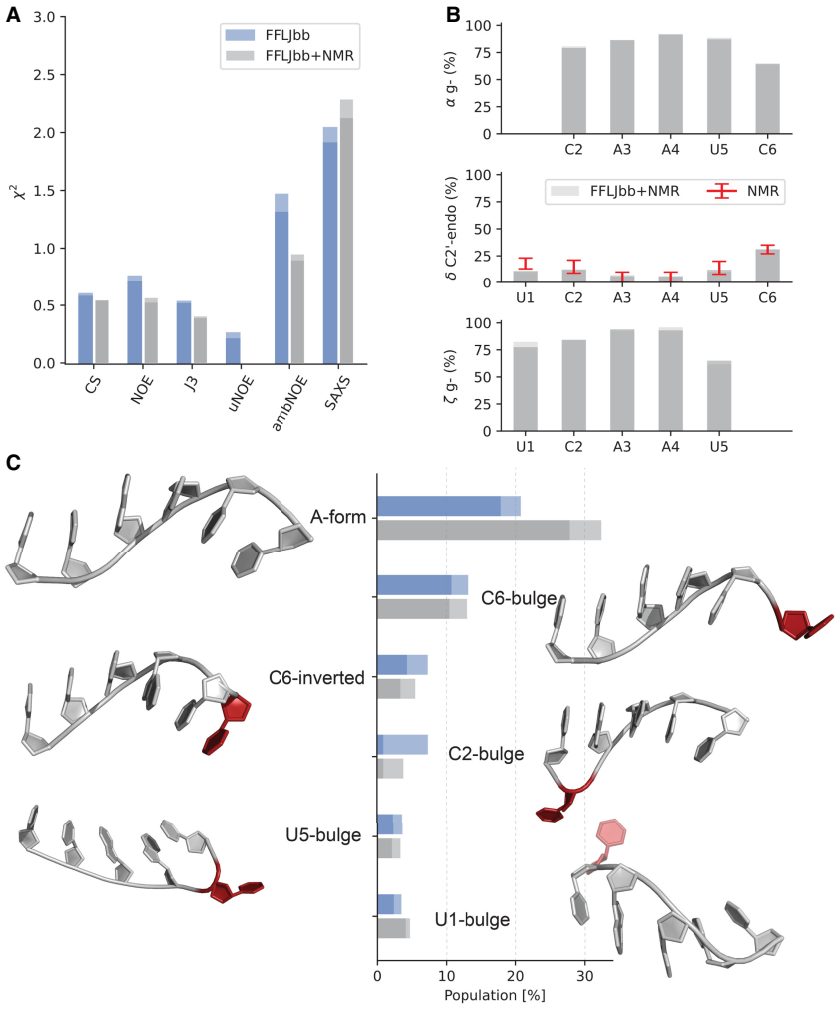


**FIGURE 4.** FFLJbb MD ensemble reweighted using NMR data. (*A*) Agreement with experimental data measured as $\chi^2$ before (blue bars) and after reweighting (gray bars). (*B*) Population of rotameric states for each residue: $\alpha$ gauche$^-$, $\delta$ in C2′-endo, and $\zeta$ in gauche$^-$ conformation, as indicated. (*C*) Cluster populations before and after NMR reweighting. The two bars in all panels indicate averages calculated over two independent simulations. Representative three-dimensional structures are shown next to each bar. Nucleotides that deviate from the canonical A-form conformations are shown in red.

(i.e., a validation set): by doing so it is usually possible to identify a range of $\theta$ values that provide a trade-off between overfitting and underfitting.

What we observe here is that when we reweight using NMR data, the $\chi^2$ relative to SAXS (not used for reweighting) is marginally affected (Supplemental Information 9). This observation holds across a wide range of $\theta$ values and for all ensembles, indicating that SAXS data are sensitive to the detailed structural information contained in NMR measurements to a small extent.

Conversely, fitting to SAXS data results in a sharp increase of $\chi^2$ relative to NMR data for small values of $\theta$. By inspecting the structural properties of SAXS-reweighted MD ensembles, we observe that large weights are associated with extended non-A-form structures. Taken together, these results indicate that at least a fraction of the samples in MD simulations are correct on a local scale, but they all possess a degree of compaction that is not fully compatible with SAXS measurements.

This inherent limitation of all computational ensembles makes it difficult to combine both NMR and SAXS data at the same time. Therefore, we proceed by describing the conformational ensemble obtained by reweighting the ensemble that best agrees with all available data (FFLJbb) using only NMR data (Fig. 4).

By construction, the agreement with chemical shifts, NOE, and scalar couplings improve in the reweighted ensemble. The most significant improvement is observed for uNOE, and the $\chi^2$ goes to zero, meaning that no uNOE violations are present in the reweighted ensemble. Ideally, we would expect the $\chi^2$ for SAXS data (not used for reweighting) to become smaller. Instead, it increases by a small amount, likely consistent with the previously noted complementarity between the global long-range SAXS data and local short-range NMR restraints such as NOEs, scalar couplings and chemical shifts.

The weights obtained from integrating FFLJbb and NMR can be used to calculate any observable of choice. As an example, in Figure 4B we report the population of the most variable torsion angles. Residues U1 to U5 are compatible with A-form-like conformations, where pucker angles are predominantly in C3′-endo conformation and $\alpha$/$\zeta$ in gauche$^-$. C6 shows a higher degree of flexibility, both around the phosphate group as well as in the sugar. From the FFLJbb + NMR ensemble, we estimate a population of C2′-endo in C6 of 32%. Note that this percentage is compatible with the estimate obtained from $^3$J scalar couplings alone (31% ± 5%) (Zhao et al. 2020). We note the presence of structures with $\chi$ angles in high-anti-conformations (>270°) that have been described in previous simulation studies as possible force field artifacts (Mlynsky et al. 2010; Chen and García 2013; Bergonzo et al. 2015). The presence of high-anti-rotameric states in FFLJbb is below 5% for pyrimidines and in the range 10%–15% for A3 and A4. Since their population is marginally smaller in the reweighted ensembles, we conclude

that these high-anti-states might be force-field artifacts, but their presence could not be ruled out completely given the available experimental and simulation data.

We report in Figure 4C NMR-adjusted cluster populations (see also Fig. 1D). The inclusion of NMR data brings the population of A-form structures between 27.8% and 32.4% (≈+11% with respect to the original FFLJbb ensemble). The population of C6 inverted conformations is 3.4%–6%, in agreement with the NOE-based estimate of 6% (Zhao et al. 2020), while the populations of other clusters are below 5% and are essentially unchanged upon NMR reweighting.

The weighted average radius of gyration is 0.78 nm, unchanged or marginally smaller compared to the original FFLJbb ensemble. Note that this value is smaller than the experimental value of 0.89 nm ± 0.01 nm obtained by SAXS measurements, and that both ROC-RNA and FARFAR2 provide more compact structures (Fig. 1B).

## DISCUSSION

We started this work as an exercise in calculating the structure ensemble for an RNA hexamer that was reported to have a more complicated composition than a single A-form conformation (Zhao et al. 2020). Through calculating a highly converged ensemble using enhanced sampling, we hypothesized that any errors with respect to the experimental NMR data could be used to quantify errors by the force field. Due to the diversity of the available experimental data including NOEs, chemical shifts, $^3$J coupling constants, we expected to correlate deviations from forward predictions based on the complete structure ensemble to specific force field terms, i.e., matching $^3$J coupling constants to limitations in specific torsions, or chemical shifts to balance electrostatic interactions between RNA and water. What we found was that the current FFLJbb force field, used with the OPC water model, reproduced experimental observables from NMR to a high degree—in fact, reweighting only minimally changed the ensemble. By running better converged simulations of the ROC-RNA force field, we were able to confirm the overstabilization of C6 bulged and inverted structures, resulting in a much more homogenous ensemble compared to FFLJbb.

A key result here, that has been recorded elsewhere as well (Zhao et al. 2020), is that a single conformer description of these short and seemingly simple oligonucleotides is insufficient. It is incorrect to describe even these short RNAs as "a structure," when they are an ensemble that contributes to a whole. A single ideal A-form structure deviated from the experimental NOEs including ambiguous as well as chemical shifts, lacking the structure diversity to fit these observables.

We tried to generate other ensembles to test whether or not having an accurate force field made a difference, versus a non-MD-based functional in Rosetta, FARFAR2, which has been used recently to predict nucleotide bulge conformations via a sample-and-select scheme (Shi et al. 2020). In summary, the results tell us that simply using any ensemble is insufficient—ensembles better matching experimental data are generated using FFLJbb. While the FARFAR2 ensemble tends to fit the NOEs relatively well, there are slight deviations from experiment for $^3$J coupling constants and more so for the chemical shifts and SAXS data.

This result suggests that multiple, independent experimental observables are important to assess the accuracy of heterogeneous structural ensembles. Furthermore, it is essential to sample the "correct" structures in order for reweighting to be accurate (Rangan et al. 2018).

We find that residual discrepancies with respect to the SAXS data cannot be accounted for by systematic errors of SAXS curve prediction from the RNA coordinates, which may be reflecting a degree of imbalance between the intra-RNA and RNA-solvent interactions carried by the force fields tested here (Salsbury and Lemkul 2021). More work needs to be done to correct for these effects before SAXS data can be used to reweight conformationally heterogeneous ensembles in combination with the NMR data (Plumridge et al. 2017; Bernetti et al. 2021).

## MATERIALS AND METHODS

### MD simulation setup

An initial RNA structure was built linearly from the sequence (U5 C A A U C3) using tLEaP in Amber18. The A-form RNA was built using the NAB functionality in Amber18 (Case et al. 2018), specifying an A-RNA helix and deleting the complementary strand. The force field used, abbreviated FFLJbb (Bergonzo and Cheatham 2015), combines modified phosphate oxygen van der Waals parameters (Steinbrecher et al. 2012) with previous revisions to FF99 (Wang et al. 2000; Pérez et al. 2007; Zgarbová et al. 2011). The RNA built with FFLJbb was solvated with 3000 OPC water molecules (Izadi et al. 2014). The RNA built with ROC-RNA was solvated with 3000 TIP3P water (Jorgensen et al. 1983). A truncated octahedron box with 1.2 nm buffer from the nucleic acid atoms was used, and five $Na^+$ ions were added to neutralize the RNA hexamer's charge, with the addition of six $Na^+$ ions and $Cl^-$ ions to yield a concentration of 80 mmol/L, where Joung–Cheatham (Joung and Cheatham 2008) monovalent ion parameters were used. A canonical ensemble was implemented with a Langevin thermostat to regulate temperature, with a collision frequency of 5 psec$^{-1}$ (Loncharich et al. 1992). Exchange between neighboring replicas was attempted every 1 psec. Cutoff for calculating direct space electrostatic interactions was 0.9 nm. SHAKE was used to constrain bonds to hydrogens (Ryckaert et al. 1977), and hydrogen mass repartitioning was implemented, increasing the mass of each solute hydrogen to 3.02 Da and decreasing the mass of the heavy atom to which each solute hydrogen atom is bonded by the same amount, allowing a 4 fsec time step (Hopkins et al. 2015). Two 1 microsecond multidimensional replica exchange MD (M-REMD) (Bergonzo et al. 2014) simulations were performed, starting from each of the initial configurations described above. The temperature range used was as follows, in Kelvin: 277, 279.98, 282.98,

286.01, 289.07, 292.15, 295.26, 298.39, 301.55, 304.77, 307.99, 311.23, 314.5, 317.8, 321.13, 324.48, 327.86, 331.28, 334.72, 338.19, 341.69, 345.21, 348.77, 352.36. The range was calculated using an online generator (Patriksson and van der Spoel 2008). The dihedral force constant bias (Bergonzo et al. 2015) was used over the following range: 1.000 0.971 0.942 0.912 0.883 0.854 0.825 0.796 0.766 0.737 0.708 0.679 0.650 0.620 0.591 0.562 0.533 0.504 0.474 0.445 0.416 0.387 0.358 0.328.

## NMR data and forward models

We compare the results of our MD simulations with previously published NMR data (Zhao et al. 2020). Specifically, we consider 64 NOE measurements, 577 unobserved NOEs, 94 chemical shifts and 39 $^3$J scalar couplings. NOEs are calculated from 5000 evenly spaced frames from the last half of each simulated trajectory (total 10,000 frames) as

$$\mathrm{NOE}_{\mathrm{CALC}} = \left[\sum_{j}^{n} w_j r_j^{-6}\right]^{-\frac{1}{6}}. \quad (1)$$

The index $j$ runs over the $n$ frames/models with associated weight $W_j$, and $r$ is the proton–proton distance. The original data set contains 13 ambiguous NOEs that are calculated by summing the contribution from both nuclei pairs (Zhao et al. 2020):

$$\mathrm{ambNOE}_{\mathrm{CALC}}\left[\sum_{j}^{n} w_j (r1_j^{-6} + r2_j^{-6})\right]^{-\frac{1}{6}}. \quad (2)$$

We use Larmor D (Frank et al. 2014) to calculate hydrogen and carbon chemical shifts from simulations, while scalar couplings are computed using Karplus equations as defined in the software package BaRNAba (Davies 1978; Lankhorst et al. 1984; Ippel et al. 1996; Marino et al. 1999; Condon et al. 2015; Bottaro et al. 2019).

## Bayesian/maximum entropy (MaxEnt) ensemble refinement

The MD simulation ensemble is refined a posteriori by including experimental information into the simulation's ensemble. The refinement is obtained by assigning a new weight to each MD frame, in such a way that the averages calculated with these new weights match more closely a set of input (or "training") experimental data within a given error. This is achieved by minimizing the following function of the weights $\boldsymbol{W} = W_1 \ldots W_n$ (Hummer and Köfinger 2015):

$$\mathrm{T}(\boldsymbol{w}) = x^2(\boldsymbol{w}) - \theta S_{REL}(\boldsymbol{w})$$

$$\mathrm{T}(\boldsymbol{w}) = \frac{1}{\mathrm{m}} \sum_{\mathrm{i}}^{\mathrm{m}} \frac{(<F(x)>_i - F_i^{EXP})^2}{\sigma_{\mathrm{i}}^2} + \theta \sum_{i=1}^{N} \mathrm{w_i} \log\left(\frac{w_i}{w_i^0}\right). \quad (3)$$

This corresponds to minimizing the deviation from the experimental measurements ($\chi^2$) with an entropic regularization term ($S_{\mathrm{REL}}$). The index $i$ runs over $m$ experimental measurements $F_i^{EXP}$ with associated uncertainty $\sigma_i$, while $<F(x)>_i$ indicates the calculated measurement averaged over the ensemble. The initial weights $W^0{}_i$ are set to $1/N\ \forall\ i$. In the present work, we perform the ensemble refinement using the Bayesian/MaxEnt (BME) code (Bottaro et al. 2020a). The regularization parameter $\theta$ was chosen using a $k$-fold cross-validation procedure.

## SAXS

The r(UCAAUC) hexamer was ordered from Integrated DNA Technologies and purified using high-performance liquid chromatography. It was dialyzed into buffers containing 50 mmol/L tris(hydroxymetyl) aminomethane (TRIS), 0.1 mmol/L ethylenediaminetetraacetic acid (EDTA), and either 75 mmol/L NaCl or 150 mmol/L NaCl at pH 6.2 to a final concentration of 2.4 mg/mL. Data were collected at 283K and 298K for RNA concentrations of 0.6 mg/mL, 1.2 mg/mL, and 2.4 mg/mL using MOLMEX Ganesha instrument at IBBR with the sample to a detector distance of 355 mm. As data collected at two lowest concentrations were completely consistent, buffer-subtracted scattering intensity profiles at 1.2 mg/mL RNA were used for further analysis. Theoretical SAXS profiles were predicted from the coordinates of the MD trajectories using FoXS (Schneidman-Duhovny et al. 2010), Crysol (Svergun et al. 1995), and AXES (Grishaev et al. 2010). Ensemble averages were fitted by linear regression to the experimental profile. The fitting parameters were used to scale and shift the calculated average to the experimental profile before calculating the $\chi^2$ as defined in Equation 3 (Gull and Daniell 1978).

## FARFAR2

The FARFAR2 program in Rosetta was used to generate the "FARFAR2" ensemble. 10,000 structures were generated based on the six-residue sequence, and later refined with the high-resolution Rosetta potential (Watkins et al. 2020).

## Cluster analysis

Cluster analysis was performed based on similarity, using the heavy atom root mean square deviation (RMSD) of all residues. The last half (500 nsec) of each M-REMD simulation was used. K-means clustering was used to generate eleven clusters, after optimization of the Davies–Bouldin index and pseudo F clustering metrics (see Supplemental Information). The initial set of points chosen was randomized.

## uMAP

The low-dimensional uMAP projection (McInnes et al. 2018) was performed using the uMAP Python package version 0.5.1 with *nneighbors* = 60 and using the collection of G-vectors (Bottaro et al. 2014) as input features. Cluster members were defined as all structures with eRMSD < 0.6 from one of the reference structures described in Figures 1 and 4.

## DATA DEPOSITION

MD and FARFAR2 structure ensembles and experimental data in tabular format are hosted on github (https://github.com/sbottaro/UCAAUC).

## SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

## ACKNOWLEDGMENTS

We would like to thank Giovanni Bussi and Kresten Lindorff-Larsen for useful discussions. Certain commercial equipment, instruments, and materials are identified in this paper in order to specify the experimental procedure. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the material or equipment identified is necessarily the best available for the purpose. S.B. acknowledges funding from the Lundbeck Foundation BRAINSTRUC structural biology initiative (R155-2015-2666).

*Received July 2, 2021; accepted March 17, 2022.*

## REFERENCES

Aytenfisu AH, Spasic A, Grossfield A, Stern HA, Mathews DH. 2017. Revised RNA dihedral parameters for the amber force field improve RNA molecular dynamics. *J Chem Theory Comput* **13:** 900–915. doi:10.1021/acs.jctc.6b00870

Bergonzo C, Cheatham TE III. 2015. Improved force field parameters lead to a better description of RNA Structure. *J Chem Theory Comput* **11:** 3969–3972. doi:10.1021/acs.jctc.5b00444

Bergonzo C, Grishaev A. 2019. Maximizing accuracy of RNA structure in refinement against residual dipolar couplings. *J Biomol NMR* **73:** 117–139. doi:10.1007/s10858-019-00236-6

Bergonzo C, Henriksen NM, Roe DR, Swails JM, Roitberg AE, Cheatham TE III. 2014. Multidimensional replica exchange molecular dynamics yields a converged ensemble of an RNA tetranucleotide. *J Chem Theory Comput* **10:** 492–499. doi:10.1021/ct400862k

Bergonzo C, Henriksen NM, Roe DR, Cheatham TE. 2015. Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields. *RNA* **21:** 1578–1590. doi:10.1261/rna.051102.115

Bernetti M, Hall KB, Bussi G. 2021. Reweighting of molecular simulations with explicit-solvent SAXS restraints elucidates ion-dependent RNA ensembles. *Nucleic Acids Res* **49:** e84. doi:10.1093/nar/gkab459

Bonomi M, Heller GT, Camilloni C, Vendruscolo M. 2017. Principles of protein structural ensemble determination. *Curr Opin Struct Biol* **42:** 106–116. doi:10.1016/j.sbi.2016.12.004

Borkar AN, Vallurupalli P, Camilloni C, Kay LE, Vendruscolo M. 2017. Simultaneous NMR characterisation of multiple minima in the free energy landscape of an RNA UUCG tetraloop. *Phys Chem Chem Phys* **19:** 2797–2804. doi:10.1039/C6CP08313G

Bottaro S, Lindorff-Larsen K. 2018. Biophysical experiments and biomolecular simulations: A perfect match? *Science* **361:** 355–360. doi:10.1126/science.aat4010

Bottaro S, Di Palma F, Bussi G. 2014. The role of nucleobase interactions in RNA structure and dynamics. *Nucleic Acids Res* **42:** 13306–13314. doi:10.1093/nar/gku972

Bottaro S, Bussi G, Kennedy SD, Turner DH, Lindorff-Larsen K. 2018. Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations. *Sci Adv* **4:** eaar8521. doi:10.1126/sciadv.aar8521

Bottaro S, Bussi G, Pinamonti G, Reisser S, Boomsma W, Lindorff-Larsen K. 2019. Barnaba: software for analysis of nucleic acid structures and trajectories. *RNA* **25:** 219–231. doi:10.1261/rna.067678.118

Bottaro S, Bengtsen T, Lindorff-Larsen K. 2020a. Integrating molecular simulation and experimental data: a Bayesian/maximum entropy reweighting approach. In *Structural bioinformatics*, pp. 219–240, Springer, NY.

Bottaro S, Nichols PJ, Vögeli B, Parrinello M, Lindorff-Larsen K. 2020b. Integrating NMR and simulations reveals motions in the UUCG tetraloop. *Nucleic Acids Res* **48:** 5839–5848. doi:10.1093/nar/gkaa399

Case DA, Ben-Shalom IY, Brozell SR, Cerutti DS, Cheatham TE III, Cruzeiro VWD, Darden TA, Duke RE, Ghoreishi D, Gilson MK, et al. 2018. Amber 2018, University of California, San Francisco.

Chen AA, García AE. 2013. High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *Proc Natl Acad Sci* **110:** 16820–16825. doi:10.1073/pnas.1309392110

Condon DE, Kennedy SD, Mort BC, Kierzek R, Yildirim I, Turner DH. 2015. Stacking in RNA: NMR of four tetramers benchmark molecular dynamics. *J Chem Theory Comput* **11:** 2729–2742. doi:10.1021/ct501025q

Costales MG, Childs-Disney JL, Haniff HS, Disney MD. 2020. How we think about targeting RNA with small molecules. *J Med Chem* **63:** 8880–8900. doi:10.1021/acs.jmedchem.9b01927

Davies DB. 1978. Conformations of nucleosides and nucleotides. *Prog Nucl Magn Reson Spectrosc* **12:** 135–225. doi:10.1016/0079-6565(78)80006-5

de Smet MD, Meenken C, van den Horn GJ. 1999. Fomivirsen–a phosphorothioate oligonucleotide for the treatment of CMV retinitis. *Ocul Immunol Inflamm* **7:** 189–198. doi:10.1076/ocii.7.3.189.4007

Frank AT, Law SM, Brooks CL III. 2014. A simple and fast approach for predicting 1H and 13C chemical shifts: toward chemical shift-guided simulations of RNA. *J Phys Chem B* **118:** 12168–12175. doi:10.1021/jp508342x

Grishaev A, Guo L, Irving T, Bax A. 2010. Improved fitting of solution X-ray scattering data to macromolecular structures and structural ensembles by explicit water modeling. *J Am Chem Soc* **132:** 15484–15486. doi:10.1021/ja106173n

Gull SF, Daniell GJ. 1978. Image reconstruction from incomplete and noisy data. *Nature* **272:** 686–690. doi:10.1038/272686a0

Hopkins CW, Le Grand S, Walker RC, Roitberg AE. 2015. Long time step molecular dynamics through hydrogen mass repartitioning. *J Chem Theory Comput* **11:** 1864–1874. doi:10.1021/ct5010406

Hummer G, Köfinger J. 2015. Bayesian ensemble refinement by replica simulations and reweighting. *J Chem Phys* **143:** 243150. doi:10.1063/1.4937786

Ippel JH, Wijmenga SS, de Jong R, Heus HA, Hilbers CW, de Vroom E, van der Marel GA, van Boom JH. 1996. Heteronuclear scalar couplings in the bases and sugar rings of nucleic acids: their determination and application in assignment and conformational analysis. *Magn Reson Chem* **34:** S156–S176. doi:10.1002/(SICI)1097-458X(199612)34:13<S156::AID-OMR68>3.0.CO;2-U

Izadi S, Anandakrishnan R, Onufriev A V. 2014. Building water models: a different approach. *J Phys Chem Lett* **5:** 3863–3871. doi:10.1021/jz501780a

Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. 1983. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79:** 926–935. doi:10.1063/1.445869

Joung IS, Cheatham TE III. 2008. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J Phys Chem B* **112:** 9020–9041. doi:10.1021/jp8001614

Lankhorst PP, Haasnoot CAG, Erkelens C, Altona C. 1984. Carbon-13 NMR in conformational analysis of nucleic acid fragments 2. A reparametrization of the Karplus equation for vicinal NMR coupling constants in CCOP and HCOP fragments. *J Biomol Struct Dyn* **1:** 1387–1405. doi:10.1080/07391102.1984.10507527

Liu B, Shi H, Al-Hashimi HM. 2021. Developments in solution-state NMR yield broader and deeper views of the dynamic ensembles of nucleic acids. *Curr Opin Struct Biol* **70:** 16–25. doi:10.1016/j.sbi.2021.02.007

Loncharich RJ, Brooks BR, Pastor RW. 1992. Langevin dynamics of peptides: the frictional dependence of isomerization rates of *N*-acetylalanyl-*N*′-methylamide. *Biopolymers* **32:** 523–535. doi:10.1002/bip.360320508

Marino JP, Schwalbe H, Griesinger C. 1999. *J*-coupling restraints in RNA structure determination. *Acc Chem Res* **32:** 614–623. doi:10.1021/ar9600392

McInnes L, Healy J, Melville J. 2018. Umap: uniform manifold approximation and projection for dimension reduction. *arXiv* doi:10.48550/arXiv.1802.03426

Mlynsky V, Banas P, Hollas D, Réblová K, Walter NG, Sponer J, Otyepka M. 2010. Extensive molecular dynamics simulations showing that canonical G8 and protonated A38H$^+$ forms are most consistent with crystal structures of hairpin ribozyme. *J Phys Chem B* **114:** 6642–6652. doi:10.1021/jp1001258

Patriksson A, van der Spoel D. 2008. A temperature predictor for parallel tempering simulations. *Phys Chem Chem Phys* **10:** 2073–2077. doi:10.1039/b716554d

Pérez A, Marchán I, Svozil D, Sponer J, Cheatham TE III, Laughton CA, Orozco M. 2007. Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophys J* **92:** 3817–3829. doi:10.1529/biophysj.106.097782

Plumridge A, Meisburger SP, Pollack L. 2017. Visualizing single-stranded nucleic acids in solution. *Nucleic Acids Res* **45:** e66. doi:10.1093/nar/gkx140

Plumridge A, Andresen K, Pollack L. 2019. Visualizing disordered single-stranded RNA: connecting sequence, structure, and electrostatics. *J Am Chem Soc* **142:** 109–119. doi:10.1021/jacs.9b04461

Rangan R, Bonomi M, Heller GT, Cesari A, Bussi G, Vendruscolo M. 2018. Determination of structural ensembles of proteins: restraining vs reweighting. *J Chem Theory Comput* **14:** 6632–6641. doi:10.1021/acs.jctc.8b00738

Reißer S, Zucchelli S, Gustincich S, Bussi G. 2020. Conformational ensembles of an RNA hairpin using molecular dynamics and sparse NMR data. *Nucleic Acids Res* **48:** 1164–1174. doi:10.1093/nar/gkz1184

Roe DR, Bergonzo C, Cheatham TE III. 2014. Evaluation of enhanced sampling provided by accelerated molecular dynamics with Hamiltonian replica exchange methods. *J Phys Chem B* **118:** 3543–3552. doi:10.1021/jp4125099

Ryckaert J-P, Ciccotti G, Berendsen HJC. 1977. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J Comput Phys* **23:** 327–341. doi:10.1016/0021-9991(77)90098-5

Salsbury AM, Lemkul JA. 2021. Recent developments in empirical atomistic force fields for nucleic acids and applications to studies of folding and dynamics. *Curr Opin Struct Biol* **67:** 9–17. doi:10.1016/j.sbi.2020.08.003

Schneidman-Duhovny D, Hammel M, Sali A. 2010. FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res* **38:** W540–W544. doi:10.1093/nar/gkq461

Shi H, Rangadurai A, Abou Assi H, Roy R, Case DA, Herschlag D, Yesselman JD, Al-Hashimi HM. 2020. Rapid and accurate determination of atomistic RNA dynamic ensemble models using NMR and structure prediction. *Nat Commun* **11:** 5531. doi:10.1038/s41467-020-19371-y

Sponer J, Bussi G, Krepl M, Banas P, Bottaro S, Cunha RA, Gil-Ley A, Pinamonti G, Poblete S, Jurecka P, et al. 2018. RNA structural dynamics as captured by molecular simulations: a comprehensive overview. *Chem Rev* **118:** 4177–4338. doi:10.1021/acs.chemrev.7b00427

Steinbrecher T, Latzer J, Case DA. 2012. Revised AMBER parameters for bioorganic phosphates. *J Chem Theory Comput* **8:** 4405–4412. doi:10.1021/ct300613v

Svergun D, Barberato C, Koch MHJ. 1995. CRYSOL: a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J Appl Crystallogr* **28:** 768–773. doi:10.1107/S0021889895007047

Tan D, Piana S, Dirks RM, Shaw DE. 2018. RNA force field with accuracy comparable to state-of-the-art protein force fields. *Proc Natl Acad Sci* **115:** E1346–E1355. doi:10.1073/pnas.1713318115

Tubbs JD, Condon DE, Kennedy SD, Hauser M, Bevilacqua PC, Turner DH. 2013. The nuclear magnetic resonance of CCCC RNA reveals a right-handed helix, and revised parameters for AMBER force field torsions improve structural predictions from molecular dynamics. *Biochemistry* **52:** 996–1010. doi:10.1021/bi3010347

Wang J, Cieplak P, Kollman PA. 2000. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J Comput Chem* **21:** 1049–1074. doi:10.1002/1096-987X(200009)21:12<1049::AID-JCC3>3.0.CO;2-F

Ward AB, Sali A, Wilson IA. 2013. Integrative structural biology. *Science* **339:** 913–915. doi:10.1126/science.1228565

Watkins AM, Rangan R, Das R. 2020. FARFAR2: improved de novo Rosetta prediction of complex global RNA folds. *Structure* **28:** 963–976. doi:10.1016/j.str.2020.05.011

Zgarbová M, Otyepka M, Sponer J, Mladek A, Banas P, Cheatham TE III, Jurecka P. 2011. Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J Chem Theory Comput* **7:** 2886–2902. doi:10.1021/ct200162x

Zhao J, Kennedy SD, Berger KD, Turner DH. 2020. Nuclear magnetic resonance of single-stranded RNAs and DNAs of CAAU and UCAAUC as benchmarks for molecular dynamics simulations. *J Chem Theory Comput* **16:** 1968–1984. doi:10.1021/acs.jctc.9b00912

*See the following page for* ***Meet the First Author***

## MEET THE FIRST AUTHOR

Christina Bergonzo

***Meet the First Author(s)* is a new editorial feature within *RNA*, in which the first author(s) of research-based papers in each issue have the opportunity to introduce themselves and their work to readers of *RNA* and the RNA research community. Christina Bergonzo is the first author of this paper, "Conformational heterogeneity of UCAAUC RNA oligonucleotide from molecular dynamics simulations, SAXS, and NMR experiments." Christina is a research chemist in the Biomolecular Measurement Division at the National Institute of Standards and Technology.**

**What are the major results described in your paper and how do they impact this branch of the field?**

Multiple, independent experimental observables are necessary to assess the accuracy of heterogeneous RNA, which is best described as an ensemble instead of a single structure. The impact of these results to the field details how to generate and assess these ensembles, telling us that molecular dynamics (MD) simulations yield more descriptive ensembles versus non-MD-based functionals.

**What led you to study RNA or this aspect of RNA science?**

RNA is more interesting than just what's going on in its helical/ordered regions. The questions of how to represent something inherently flexible, so we can understand it better, has always interested me. MD simulations make a great tool for investigating the dynamics of RNA.

**During the course of these experiments, were there any surprising results or particular difficulties that altered your thinking and subsequent focus?**

The surprising results were from the Rosetta FARFAR2 predicted ensemble. While the ensemble "fits" the NOE data, that is really because the NOE data tells us less about different conformations than it does about intra-residue distances, which are all pretty fixed. It helped reinforce that we should be comparing to multiple experiments, and a robust MD force field will do a good job approximating all of them.

**Are there specific individuals or groups who have influenced your philosophy or approach to science?**

Professor Cheatham, who advised my postdoctoral research, is a big influence on my approach to science—it should be open, shared, and communicated in good faith.