Published online 19 May 2020 *Nucleic Acids Research, 2020, Vol. 48, No. 11* ***5839–5848***
doi: 10.1093/nar/gkaa399

Integrating NMR and simulations reveals motions in the UUCG tetraloop

Sandro Bottaro[1,2,*], Parker J. Nichols[3], Beat Vögeli[3], Michele Parrinello[1] and Kresten Lindorff-Larsen[2,*]

[1]Atomistic Simulations Laboratory, Istituto Italiano di Tecnologia, Genova, Italy, [2]Structural Biology and NMR Laboratory, Department of Biology, University of Copenhagen, Copenhagen, Denmark and [3]Department of Biochemistry and Molecular Genetics, University of Colorado Denver, Anschutz Medical Campus, Aurora, CO, USA

Received January 04, 2020; Revised April 03, 2020; Editorial Decision May 01, 2020; Accepted May 17, 2020

ABSTRACT

We provide an atomic-level description of the structure and dynamics of the UUCG RNA stem–loop by combining molecular dynamics simulations with experimental data. The integration of simulations with exact nuclear Overhauser enhancements data allowed us to characterize two distinct states of this molecule. The most stable conformation corresponds to the consensus three-dimensional structure. The second state is characterized by the absence of the peculiar non-Watson–Crick interactions in the loop region. By using machine learning techniques we identify a set of experimental measurements that are most sensitive to the presence of non-native states. We find that although our MD ensemble, as well as the consensus UUCG tetraloop structures, are in good agreement with experiments, there are remaining discrepancies. Together, our results show that (i) the MD simulation overstabilize a non-native loop conformation, (ii) eNOE data support its presence with a population of ≈10% and (iii) the structural interpretation of experimental data for dynamic RNAs is highly complex, even for a simple model system such as the UUCG tetraloop.

INTRODUCTION

RNA loops are structural elements that cap A-form double helices, and as such are fundamental structural units in RNA molecules. The great majority of known RNA loops contain four nucleotides (1), and these so-called tetraloops are one of the most common and well-studied three-dimensional RNA motifs. The great majority of known RNA tetraloops have the sequence GNRA or UNCG, where N is any nucleotide and R is guanine or adenine (2). Their small size, together with their biological relevance, has made these systems primary targets for nuclear magnetic resonance (NMR) spectroscopy, X-ray-crystallography, and atomistic molecular dynamics (MD) simulation studies (3–5).

The UUCG tetraloop has been long known to be highly stable, and both crystallographic and NMR studies suggest that this tetraloop adopts a well-defined three dimensional structure including a characteristic trans-Sugar-Watson (tSW) interaction between U6 and G9 (6,7) (Figure 1). Experimentally, the UUCG tetraloop is used to stabilize the secondary structure of larger RNA molecules without interacting with other RNAs or proteins (8).

Despite its stability, the UUCG tetraloop is not rigid, and the sequence does not systematically determine a unique structure (2,10). Three recent studies by independent groups indicate the presence of alternative loop conformations (11–13), and earlier NMR studies (7,14) also suggested the presence of loop dynamics, without providing a detailed structural interpretation of the data. More generally, the atomic-detailed characterization of RNA structure and dynamics requires specialized techniques and substantial experimental effort, including NMR measurements of nuclear Overhauser effects (NOE), scalar couplings, chemical shifts, residual dipolar couplings, cross-correlated relaxation rates as well as a wide range of relaxation–dispersion type NMR experiments (15,16).

While NOEs are typically used to determine RNA and protein structures, they also contain dynamic information. Because ensemble-averaged NOEs are highly sensitive to the underlying distance fluctuations, they may contain contributions even from minor populations. Normally, such information is difficult to extract because standard NOE measurements are relatively inaccurate. It has, however, been demonstrated that a substantial part of the information content inherent to these probes can be obtained from exact NOE measurements (eNOEs) (13,17). As opposed to conventional NOEs, eNOEs can be converted into tight upper and lower distance limit restraints (18–20).

[*]To whom correspondence should be addressed. Tel: +45 3532 2027; Email: sandro.bottaro@bio.ku.dk
Correspondence may also be addressed to Kresten Lindorff-Larsen. Email: lindorff@bio.ku.dk

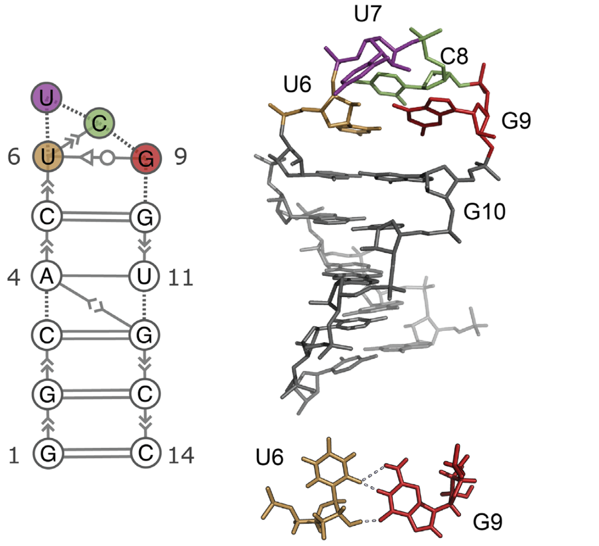

Figure 1. Consensus secondary structure (left) and three dimensional structure (right) of the UUCG tetraloop (7). The stem is formed by 5 consecutive Watson–Crick base-pairs capped by the loop U6–U7–C8–G9. One of the most distinctive feature of this structure is the *trans*-sugar–Watson interaction between U6 and G9 (bottom). Extended secondary structure annotation follows the Leontis–Westhof nomenclature (9).

Previous computational studies of the UUCG tetraloop focused either on the dynamics around the near-native state (21) or on the difficulty in separating force-field inaccuracies from insufficient sampling (22,23). In a previous study, we reported converged free-energy landscape for RNA 8-mer and 6-mer loops, and we have shown that native-like states are not the global free-energy minimum using the current AMBER RNA force-field (24). This problem has been addressed in a new parameterization of the AMBER force-field, that improves the description of the UUCG 14-mer and other RNA systems (25). Nevertheless, it remains difficult to assess the accuracy of these simulations, because experiments alone do not provide an atomic-detailed description of structure and dynamics that serve as a benchmark.

Here, we use extensive atomistic MD simulations to map the conformational landscape of the UUCG tetraloop using enhanced sampling techniques and a recent force-field parameterization. To improve the description of this system further, we perform an a posteriori refinement of the MD simulation using experimental data via a Bayesian/maximum entropy procedure (26,27). We validate the eNOE-refined ensemble against independent NMR measurements and find an agreement that is on average comparable with NMR structures of the UUCG tetraloop deposited in the Protein Data Bank (PDB).

Our experimentally-refined ensemble reveals the presence of two conformational states. The dominant, major state (here called state A) is the consensus UUCG structure shown in Figure 1. The second, previously unreported lowly-populated state (state B) is characterized by the absence of the signature U6–G9 non-Watson–Crick base pair, with the C8 and G9 bases exposed into solution. We employ a random forest classifier to identify the structural properties that distinguish state A from state B. Furthermore, we use the same method in the space of experimental data to identify specific measurements that are most sensitive to the presence of state B.

The paper is organized as follows: in the Results section we first compare the predictions obtained from MD simulation against different experimental datasets. We then discuss the effect of the refinement procedure, showing how it improves the agreement with experiments and how it affects the population of different conformations. We proceed by identifying the relevant degrees of freedom and contacts that characterize the two states. Finally, we identify experiments sensitive to the presence of state B. We accompany this paper with the code to reproduce step-by-step the complete analysis, including all figures and supplementary results presented in the manuscript.

MATERIALS AND METHODS

MD simulations

We simulate the RNA 14-mer with sequence GGCACUUCG-GUGCC starting from a completely extended conformation. Studying the folding free-energy landscape of this system is computationally expensive: for this reason previous attempts required μs-long simulations in combination with tempering protocols (25,28,29).

Here, we combine two enhanced sampling techniques: solute tempering in the REST2 formulation (30) and well-tempered metadynamics (31). We used a nucleic-acid specific metric, called eRMSD, (32) as a collective variable for enhanced sampling. The MD simulation setup and convergence analysis are presented in supporting information 1 (SI1).

Calculating experimental quantities from structure (forward models)

We calculated NOEs from the ensembles as $\mathrm{NOE}_{\mathrm{CALC}} = [\sum_j^n w_j r_j^{-1/6}]^{-6}$. The index j runs over the n frames/models with associated weight w_j. This is an approximation because the NOE depends also on the timescales of motion and may contain contributions from angular fluctuations (18). In our analyses we circumvent these problems by extracting effective average distances from the NOE buildup curves and calculate these as weighted distance averages, and note that more advanced analyses would require reweighting of long-timescale correlation functions. We calculated ^{3}J scalar couplings using Karplus equations as defined in the software package baRNAba (33). Cross-correlated relaxation rates are predicted with an in-house script available on Github (see Data Availability). Pales (34) is used to compute residual dipolar couplings (RDC), while solvent paramagnetic resonance enhancements (sPRE) are calculated as described in (35). More details and practical examples are available in SI2.

Integrating simulations and experiments

We combine the MD simulation with experimental data using a maximum entropy/Bayesian procedure (26,36,37). In our previous work, we have described this reweighting procedure as Bayesian/MaxEnt (BME) (27,38). In BME we use the experimental data to modify a posteriori the simulation so that the new conformational ensemble has the following properties: (i) the calculated averages are close to the experimental values taking uncertainty into account and (ii) it maximizes the relative Shannon entropy with respect to the original simulation ensemble. The modification comes in the form of a new set of weights w_j^*, one for each simulation frame.

It can be shown that this problem can be cast as a minimization problem, in which one seeks the minimum of the function Γ with respect to the set of Lagrange multipliers $\bar{\lambda} = \lambda_1 \cdots \lambda_m$, with m being the number of experimental constraints.

$$\Gamma(\bar{\lambda}) = \log(Z(\bar{\lambda})) + \sum_i^m \lambda_i F_i^{\text{exp}} + \frac{\theta}{2}\sum_i^m \lambda_i^2 \sigma_i^2 \qquad (1)$$

Here, σ_i are the uncertainties on the experimental measurements F_i^{exp} and include experimental errors and inaccuracies introduced by the calculation of the experimental quantity from the atomic positions ($F(\mathbf{x})$). The partition function Z is defined as

$$Z(\bar{\lambda}) = \sum_{j=1}^n w_j^0 \exp\left[-\sum_i^m \lambda_i F_i(\mathbf{x}_j)\right] \qquad (2)$$

The sum over the index j runs over the n frames in the simulation, and w_j^0 are the original weights. $w^0 = 1/n$ when using plain MD simulations or enhanced sampling techniques that sample directly from the target distribution (e.g. parallel tempering). In this paper we use WT-METAD, and the original weights w^0 are estimated using the final bias potential (39). The minimization of equation (1) yields a set of Lagrange multipliers $\bar{\lambda}^*$ that are used to calculate the optimal weights

$$w_j^* = \frac{1}{Z(\bar{\lambda}^*)} w_j^0 \exp\left[-\sum_i^m \lambda_i^* F_i(\mathbf{x}_j)\right] \qquad (3)$$

We choose the free hyper-parameter of the algorithm (θ) by performing a 5-fold cross-validation procedure as described in SI3.

Random forest classifier

The random forest analysis is set up according to the following procedure:

- Bootstrap $n = 50\,000$ samples from the MD simulation trajectories. The weight of each sample, w_j^*, is calculated as described above.
- Samples are assigned to state A if the eRMSD from the first model of the NMR structure 2KOC is <0.7, and state B otherwise.
- Calculate a set of m features (torsion angles, distances between ring centres, experimental data) for each sample (33).
- Construct a random forest classifier using n samples, m features and 2 classes (state A and state B). In our work, we use the implementation of the random forest algorithm (40) available in sklearn 0.22 using a maximum tree depth of 2. 80% of the samples are used for training, while the remaining 20% are used to evaluate the accuracy of the classifier (>97% in all cases).
- Rank the m features by their importance. In the analysis of the experimental data in Figure 5, only features with importance greater than 0.2 are shown.

The implementation of the procedure is described in SI 4.

RESULTS

First, we compare the computational prediction with available NMR spectroscopy data. More precisely, we consider the following experimental datasets:

- **Dataset A**. Exact eNOEs (13), consisting in 62 bidirectional exact NOE, 177 unidirectional eNOE and 77 generic normalized eNOE (gn-eNOE). This dataset alone was used to determine the structure of the UUCG tetraloop with PDB accession codes 6BY4 and 6BY5. In addition to the original dataset, we added one new eNOE and six new gn-eNOEs.
- **Dataset B**. 96 ^{3}J scalar couplings, 32 RDCs, 251 NOE distances, and 84 cross-correlated relaxation rates. These data were used to calculate the consensus UUCG tetraloop structure (PDB 2KOC (7)).
- **Dataset C**. 39 (RDC1) plus 14 (RDC2) residual dipolar couplings. These RDCs have been used in conjunction with MD simulations to obtain a dynamic ensemble of the UUCG tetraloop. (11).
- **Dataset D**. 91 solvent paramagnetic resonance enhancement (sPRE) measurements (12).

The exact set of experimental data we used is described in SI2 and available in tabular format in SI 5. The orange bars in the four panels of Figure 2 show the agreement between simulation and the different experimental datasets. The agreement between experiment and simulations is expressed using the reduced χ^2 statistics, defined as the average square difference between the experimental measurement (F^{exp}) and the back-calculated ensemble average $\langle F(\mathbf{x})\rangle$normalized by the experimental error σ:

$$\chi^2 = \frac{1}{m}\sum_i^m \frac{(\langle F(\mathbf{x})\rangle_i - F_i^{EXP})^2}{\sigma_i^2} \qquad (4)$$

Hence, the lower the χ^2, the better the agreement. As a rule of thumb, $\chi^2 < 1$ can be considered small, as the average difference between experiment and prediction is within experimental error.

As a reference, we report in Figure 2 the agreement calculated on the PDB ensembles 6BY5 (13) and 2KOC (7). For set A, the agreement of the MD with experiment is considerably poorer than the one calculated on 6BY5. We recall that this latter ensemble was determined by fitting dataset A, we thus expect χ^2 to be small in this case. On datasets B, D the MD better agrees with experiments compared to 2KOC, 6BY5, while differences on set C are smaller. The same conclusions apply when performing the comparison for each type of data and when considering other statistics (SI 6). Note that χ^2 for sPRE, RDC, and scalar couplings

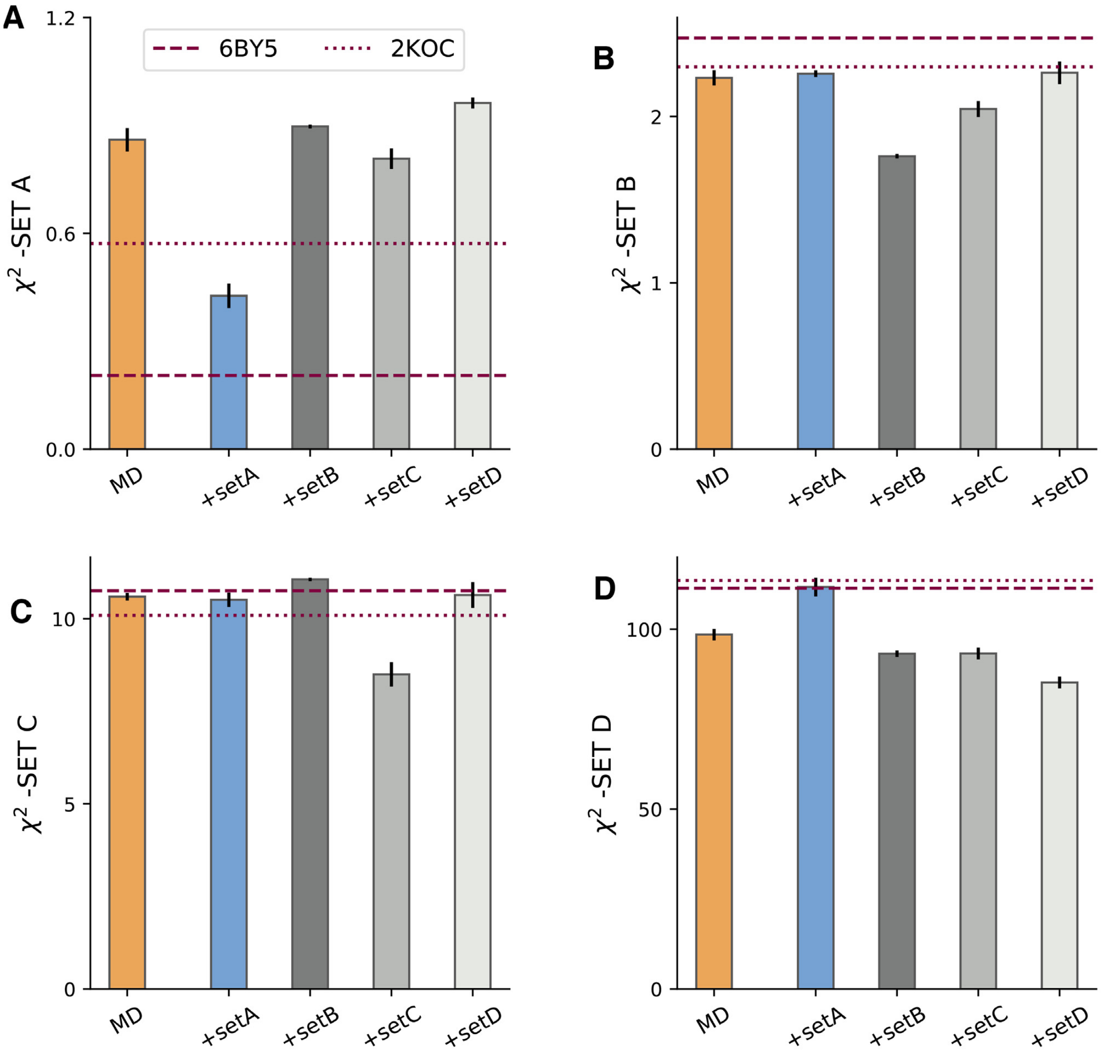

Figure 2. Comparison between experiment and simulations. (**A**) χ^2 between experimental dataset A against the MD ensemble (MD) and against the refined ensembles (MD+set A, MD+set B, MD+set C, MD+set D). As a reference, values calculated from all NMR models from PDB structures 2KOC and 6BY5 are shown as dashed lines. The agreement between the same ensembles and datasets B, C, D, are shown in panels (**B**), (**C**) and (**D**), respectively. Error bars show the standard error estimated using four blocks.

is very large. This discrepancy may arise both from the imperfect ensembles, from the underestimation of the experimental error, as well as from the limitation of the function used to calculate the experimental quantity from the atomic positions (i.e. the forward model). As an example, the parameters in the Karplus equation for HCOP couplings critically depend on a single experimental data point measured in 1969 (41).

Bayesian/Maximum entropy refinement of the MD ensemble

As described above, our MD simulation provides a conformational ensemble consisting of a rich and diverse set of conformations, that, however, as a whole does not match all experimental data perfectly.

In order to improve the description provided by the MD simulation, we calculate a refined conformational ensemble by *a posteriori* including experimental information into simulations. In brief, the refinement is obtained by assigning a new weight to each MD snapshot, in such a way that the averages calculated with these new weights match a set of input (or 'training') experimental data within a given error. Among all the possible solutions to this underdetermined problem, we use the one that maximizes the Shannon cross-entropy (36,42,43).

Since we have four independent experimental datasets, we can refine the simulation by using one of them, using the other three as test set. By construction, the refinement procedure improves the agreement on the training data, but it does not guarantee improved agreement on the test sets. Figure 2 shows the effect of the refinement for different combinations of training/test set. For example, the blue bar in panel A shows the agreement between experiment and the MD simulations refined against set A (MD+set A) evaluated on set A itself (training). The blue bars in panels B, C, D show the χ^2 of MD+set A on datasets B, C, D (test). The MD+set B is trained on set B and tested on datasets A, C, D and so on.

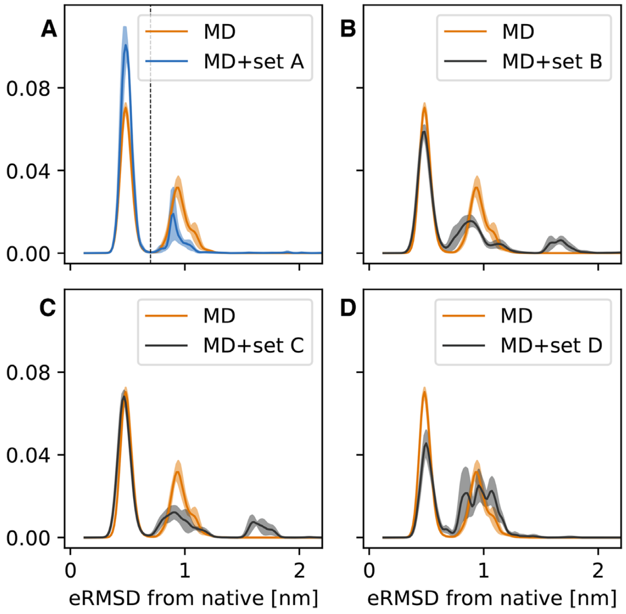

Figure 3. Histograms of the eRMSD from native. The original MD simulation (orange) is compared with the four refined ensembles: MD+set A in panel (**A**), MD+set B in panel (**B**), MD+set C in panel (**C**) and MD+set D in (**D**). Shades show the standard error estimated using four blocks. The vertical dashed line in panel (A) shows the separation between state A (eRMSD < 0.7) and state B (eRMSD ≥ 0.7).

We observe that including into simulations dataset A has a detrimental effect on the agreement with set D, leaving the agreement with dataset B and C unchanged. The MD+set B ensemble shows improved agreement with dataset D, but performs worse than the original MD on set A and C. We observe a similar behavior for MD+set D. Among the four refined ensembles the MD+set C is the one that behaves better, as it shows a smaller or equal χ^2 (relative to MD) on the three tests sets A,B,D. Taken together, our results show that the MD and the refined ensembles fit available experimental data to a degree that is comparable to the one calculated from PDB structures 2KOC and 6BY5 (Figure 2). However, there exists substantial discrepancies with experimental data, and none of the considered ensembles clearly outperforms the others.

Conformational ensemble of the UUCG Tetraloop

In this section we analyse in detail the refined MD ensemble, and discuss the differences with respect to the original simulation and previously determined structures. We consider the histogram of the distance (structural dissimilarity) from the consensus structure (PDB: 2KOC). Distances are measured using the eRMSD, a nucleic-acid specific metric that takes into account both position and orientations between nucleobases (32), although a qualitatively similar picture is obtained using the standard RMSD metric.

The distribution obtained from the original MD ensemble is shown in orange in Figure 3A, and repeated in the four panels. The two peaks correspond to structures where the stem is fully formed, but with different loop conformations. Structures in the left peak (eRMSD < 0.7, state A) display the signature interactions present in the consensus structure, while in the second peak (state B) the loop is disordered. The population of state A in our MD is 60% ± 4, larger than the one calculated from the 180μs simulated tempering simulation by Tan *et al.* (25) (≈ 40%).

The inclusion of experimental data in simulations affects the histogram in different ways. When including eNOE data from dataset A (Figure 3A, blue), we observe an increase in the population of state A to 83% ± 16. The refinement obtained with datasets B and C (Figure 3 panels b,c) has a small effect on the population of state A, i.e. 60% ± 7 and 65% ± 5, respectively. In the MD+set D ensemble, the population of state A is decreased (44% ± 10). Interestingly, a new peak appears at eRMSD >1.5 in the MD+set B and MD+set C ensembles. This peak corresponds to fully extended conformations, and it is present when including RDC measurements. If we trust our simulation, the forward model including the complex issue of estimating the alignment tensor, and the refinement procedure, we can conclude that RDC are sensitive to a small population of either dimers or unfolded structures. Unfortunately, we do not have the possibility to test this hypothesis, which is left for future investigations.

Here, instead, we focus on the results obtained from MD+set A. First, because this ensemble has the highest population of state A, and it is thus less 'surprising'. Second, because we better understand and control the experimental data, and lastly because we can take advantage of our previous experience in using NOE for RNA ensemble refinement (13,27).

Structural differences between state A and B

Having discovered this new B-state, we proceed to analyse its structural features. While state A is known and structurally well-defined (Figures 1 and 4A), it is not trivial from a simple visual inspection to identify which are the main features of state B (Figure 4B). Here, we address this question by using a random forest classifier. In practice, we first extract samples from the MD+set A ensemble. Second, we label each conformation depending on the distance from native (state A if eRMSD<0.7, and state B otherwise). For each conformation we calculate structural properties (e.g. torsion angles, distances) that are used to train a random forest classifier. In this context we are not interested in the decision tree *per se*, but rather in its ability to rank the importance of the input features in the classification problem. Thus, we can find the most relevant degrees of freedom that discriminate the two states. The result of such analysis on all dihedral angles α, β, γ, δ, ϵ, ζ, χ in the 14-mer reveals that the two highest-ranked angles are ζ in C8 and G9. Figure 4D,E show the histograms for these two angles: the angle in C8 is a good classifier, as all samples from state A (empty curve) are in *gauche* + conformation, while all samples from state B (filled curve) are in *gauche* −. Information on state A/B differences are also contained in G9 ζ, although the separation of the two states is not as striking as in C8.

The importance of C8 and G9 is further confirmed when using the distance between the center of the six-membered

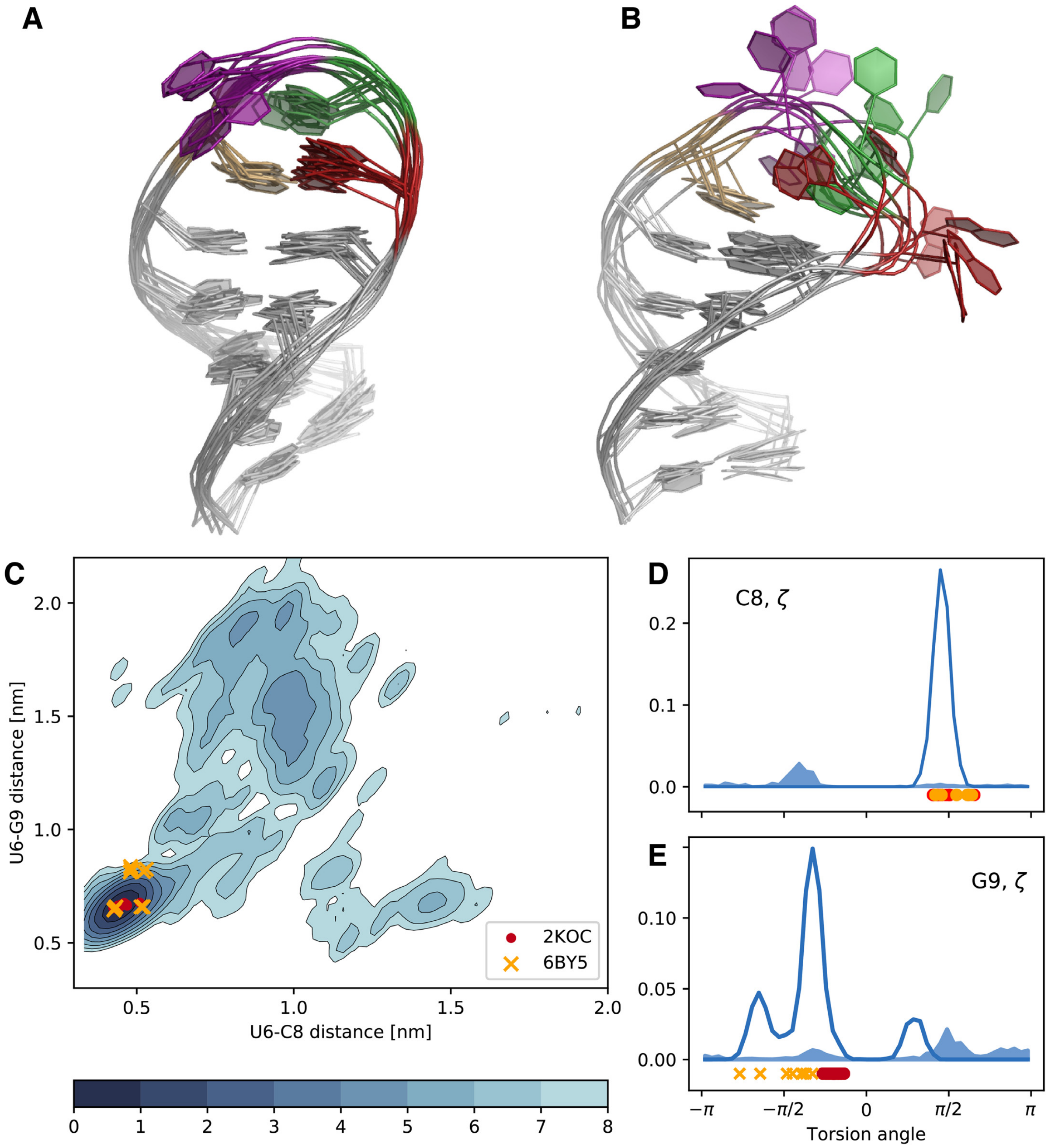

Figure 4. (**A**) Representative (random) conformations sampled from state A. (**B**) Representative (random) conformations sampled from state B. The color code is identical to Figure 1: U6 in ochre, U7 in purple, C8 in green and G9 in red. (**C**) Free energy surface projected onto the the U6-C8/U6-G9 distance between ring centres. The units of the colorbar are in k_BT. (**D**) Histogram of ζ dihedral angle in C8. The open filled area indicates conformations belonging to state A, and the filled area indicates conformations belonging to state B. (**E**) Histogram of ζ dihedral angle in G9.

rings in the nucleobases as input features. In this case, the distances between U6–C8 and U6–G9 are the two most important degrees of freedom that distinguish state A from state B. The two-dimensional log-histogram (i.e. the free energy surface) of these distances is shown in Figure 4C. In the consensus structure U6 and G9 interact through a trans sugar-Watson base-pair and U6 and C8 are stacked. On top of the free-energy surface, in Figure 4C we plot the two tetraloop structures 2KOC and 6BY5. As expected, both structures lie in the left-bottom region. Note also that the original experimental study described the presence of two sub-states in 6BY5, that are separated along the y projection in Figure 4C. State B is characterized instead by the absence of the stacking interaction (large U6–C8 distance), and of the non-Watson-Crick base-pair (large U6–G9 distance). PDB structures of both states A and B are available in supplementary material as well as in the Github repository (see data availability).

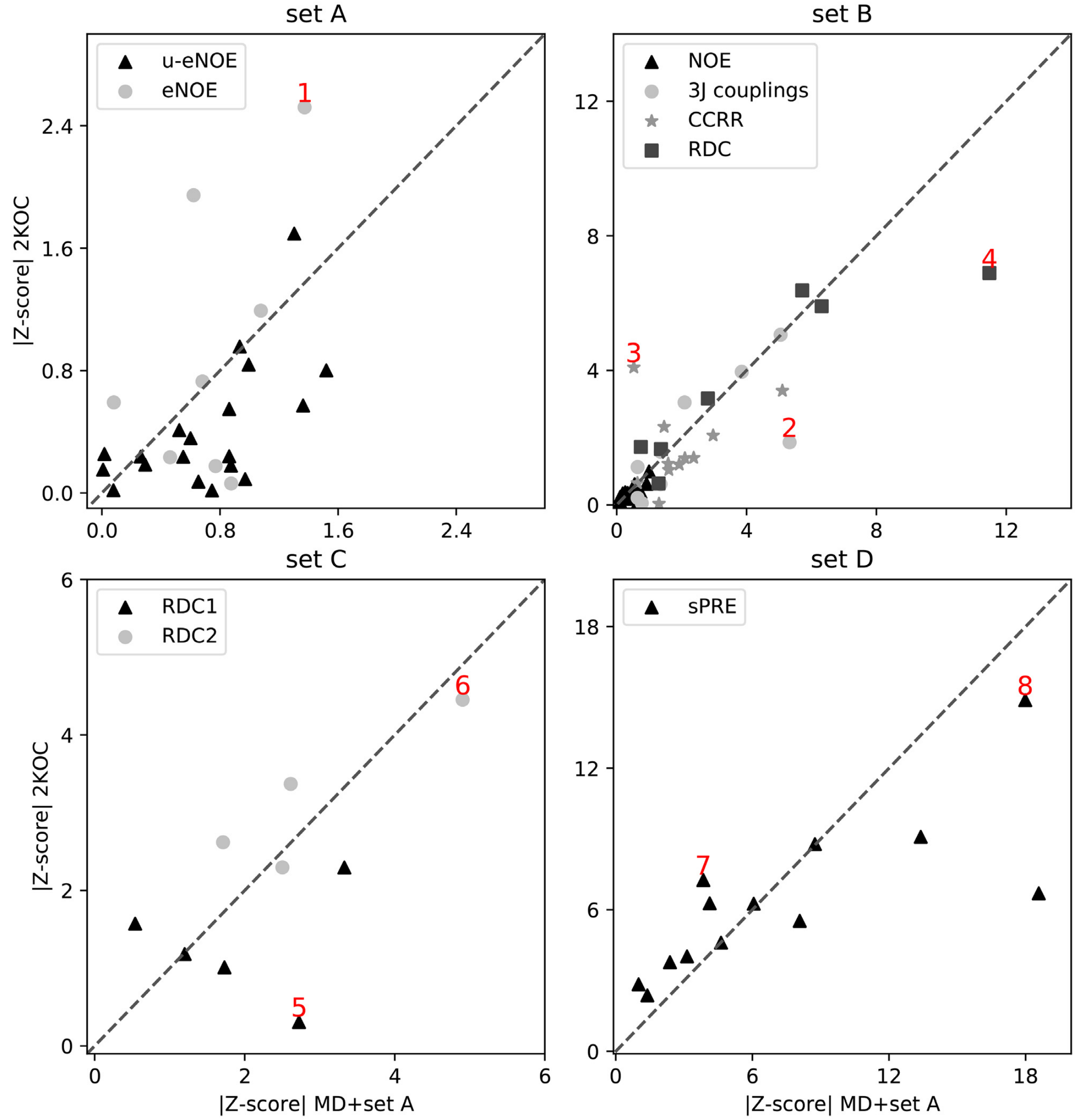

Figure 5. Comparison between 2KOC and MD+set A ensemble on experimental data that are most sensitive to the presence of state B. Scatter plots show the Z-score calculated on the MD+set A ensemble (x-axis) versus the same quantity calculated on the PDB ensemble 2KOC (y-axis). The four panels show data belonging to the four datasets: set A in panel (**A**), set B in panel (**B**), set C in panel (**C**) and set D in panel (**D**). Points discussed in the main text and shown in Figure 6 are labeled in red.

Experimental measurements sensitive to the presence of state B

In the previous section we have described the structural differences between state A and state B, and we now seek for additional experimental validation. In particular, we would like to answer the following question: does the MD+set A ensemble provide a better loop description compared to the consensus NMR structure 2KOC?

The four experimental sets contain >1000 datapoints in total, and it is therefore not trivial to identify specific measurements that probe directly the presence of state B. In machine learning this is called a feature importance problem, that we solve using a random forest classifier, as we did in the previous section. Here, however, the features are not structural properties, but back-calculated experimental data. We construct a random forest classifier for each experimental dataset, and we obtain four sub-datasets with the features (measurements) that are most important for classifying (i.e. separating) state A and B. Consistently with Figure 4, we find that these measurements involve nucleotides in the loop region.

For each measurement in the sub-datasets, we calculate the difference between the ensemble average and experiments divided by the error σ (i.e. the Z-score).

$$|\text{Z-score}|_i = \frac{|\langle F(\mathbf{x})\rangle_i - F_i^{EXP}|}{\sigma_i} \quad (5)$$

Figure 5 shows the *Z*-score calculated on the MD+set A ensemble (x-axis) against the same quantity calculated on the 2KOC structures. Points above the diagonal (such as

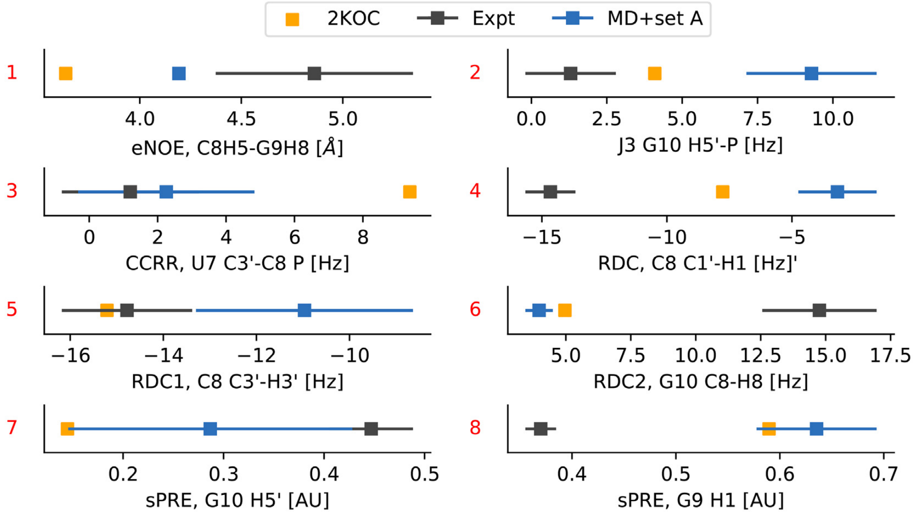

Figure 6. Comparison between calculated and experimental data for selected measurements discussed in the text and labelled as in Figure 5.

1 in panel a, 3 in panel b, and 7 in panel d) indicate measurements for which the MD+set A ensemble better agrees with experimental data ($|Z\text{-score}|_{\mathrm{MD + setA}} < |Z\text{-score}|_{\mathrm{2KOC}}$), while points below the diagonal (e.g. 2, 4, 5, 6, 8) indicate the opposite.

Overall, there are more points below the diagonal (56) than above (40), suggesting that the original 2KOC ensemble provide a better description of the loop region. For example, the eNOE C8 H5’ to G9 H8 in dataset A is significantly closer to the experiment in the MD+set A ensemble (point 1, Figure 6). This is also true for the cross-correlated relaxation rate U7 C3’–C8 P (point 3). In other cases the 2KOC value is within experimental error, but not in our MD+set A ensemble (points 2,5 Figure 6). Note that major discrepancies are present in both ensembles, such as point 4 (RDC in C8 C1’–H1’), point 6 (RDC G10 C8–H8), and point 8 (sPRE G9 H1). Again, we stress that these discrepancies can be ascribed to errors in the ensembles, but also to inaccuracies in the empirical model employed to calculate experimental data from structures, or to errors in the data.

DISCUSSION

Based on our extensive MD simulations and integrating them with exact NOE data, we report the free energy landscape of a prototype stem-loop RNA 14-mer known as the UUCG tetraloop. By combining a recently refined force field for RNA with enhanced sampling MD we were able to fold the tetraloop to its native conformation(s) as judged by the agreement with several sets of experiments. The main finding of the present study is the presence of a low populated, non-native conformation (state B). The low-populated state differs from the consensus structure (state A) only in the loop region, and it is characterized by the absence of the tSW base-pair between U6 and G9, with C8 and G9 partially exposed into solution (Figure 4). This result has been obtained by using atomistic MD simulations and eNOE, without the need of additional data.

The free-energy surfaces and estimated population provided here are based on the available experimental data, on the employed model, and the extent of our sampling. Therefore, they are subject to inaccuracies. However, both simulations and eNOE data are consistent with the presence of the B state as described in this paper. This interpretation is qualitatively consistent with several NMR studies, that also suggested the presence of dynamics in G9 (7,12,14). Conversely, on-resonance 13C R1ρ relaxation dispersion experiments on a UUCG tetraloop with a different stem sequence showed no significant exchange contributions, indicating the absence of motions with substantial chemical shift variation in the μ-ms timescale (44). Note also that G9-exposed structures were reported in previous MD simulations (23,28,45), suggesting our finding to be robust with respect to the choice of the force-field and water model.

In this work, we have used eNOEs to reweight *a posteriori* the ensemble generated via enhanced sampling MD simulations. This refinement procedure is a computationally cheap post-processing (26,27,46) that allows one to try different combinations of training and test sets, as we did in this work (43). Note that refinement is in principle less powerful compared to on-the-fly methods that samples directly from the target probability distribution (47,48).

In our study, we refine the simulation by matching RDC data (set B, C) or solvent PRE (set D) as well. Only when we use set C for training we obtain an improved or equal agreement on the test sets relative to the original MD simulation (Figure 2). Additionally, different data affect the MD conformational ensemble in different ways (Figure 3). Several reasons can contribute to this behavior. First, we do not expect all experimental data to be perfectly compatible one with the other, because measurements were conducted in similar, but not identical conditions. Second, the forward models might not be accurate for arbitrary molecular conformations. For example, if the forward model accurately predicts the RDC given the native structure, but fails on unfolded/misfolded conformations, we obtain artefacts that cannot be easily accounted for in our refinement procedure. Note that this problem is typically less relevant when using experimental RDC, sPRE or chemical shift data for scoring structures (12,44,49).

Based on the above observations, and considering our previous experience with eNOE data, we here analyse in detail the results obtained using MD refined using set A (Figure 3A). The structural features that are most important to discriminate between state A and state B are identified using a random forest classifier. The problem of concisely interpret differences in biomolecular conformations has been recently pursued using a variety of machine learning methods, including linear discriminant analysis (50), decision trees (51), and others (52). In this work we extend this idea, and use back-calculated experimental data as input features for the random forest classifier. In this way, we identify individual (available) measurements that are most sensitive to the presence of state B. Note that this approach can also be used in a generative fashion to design experiments that probe the existence of specific conformational states.

We closely inspect the selected set of measurements that are sensitive to state B (Figure 5). In the majority of the cases, we find the presence of the additional state to provide a worse agreement with experiment compared to the consensus NMR structure (PDB code 2KOC) (see e.g. Figure 6, points 2, 4, 5). In other cases (Figure 6, points 1, 3, 7, instead, the MD+set A performs better than 2KOC. Several other data significantly deviate from experiments in both ensembles (Figure 6, points 6,8). This suggest the possibility that conformations that are different from state A are indeed present, but do not correspond to the state B as described in Figure 4B.

Finally, we note that the approach taken here is general and it is applicable to other RNA or protein systems (53,54). Previous characterization of slow, larger motions in RNA molecules have mostly relied on relaxation-dispersion, chemical exchange saturation transfer or related NMR experiments that probe chemical shift differences between different conformational states. We hope that the integration of MD simulations and eNOE measurements provides further opportunities for characterizing the free energy landscapes of RNA molecules.

DATA AVAILABILITY

Jupyter notebooks to reproduce the analysis and all figures are included as supporting information. The MD trajectory, together with the experimental data are hosted on github (https://github.com/KULL-Centre/papers/tree/master/2020/UUCG-dynamics-Bottaro-et-al). The Plumed input file to reproduce the simulation is hosted on the Plumed nest (55) under the accession code 19.070.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Giovanni Bussi for helpful comments. We also thank Tobias Madl, Harald Schwalbe and coworkers for useful discussions and for sharing their data.

FUNDING

Velux Foundations (to S.B. and K.L.-L.); Lundbeck Foundation BRAINSTRUC initiative (to K.L.-L.); University of Colorado (to B.V.); NSF [1917254] for Infrastructure Innovation for Biological Research (to B.V.). Funding for open access charge: Internal funding.

Conflict of interest statement. None declared.

REFERENCES

1. Wolters,J. (1992) The nature of preferred hairpin structures in 16S-like rRNA variable regions. *Nucleic Acids Res.*, **20**, 1843–1850.
2. Bottaro,S. and Lindorff-Larsen,K. (2017) Mapping the universe of RNA tetraloop folds. *Biophys. J.*, **113**, 257–267.
3. Cheong,C., Varani,G. and Tinoco,I. Jr (1990) Solution structure of an unusually stable RNA hairpin, 5GGAC (UUCG) GUCC. *Nature*, **346**, 680.
4. Woese,C., Winker,S. and Gutell,R. (1990) Architecture of ribosomal RNA: constraints on the sequence of 'tetra-loops'. *Proc. Natl. Acad. Sci. U.S.A.*, **87**, 8467–8471.
5. Ferner,J., Villa,A., Duchardt,E., Widjajakusuma,E., Wöhnert,J., Stock,G. and Schwalbe,H. (2008) NMR and MD studies of the temperature-dependent dynamics of RNA YNMG-tetraloops. *Nucleic Acids Res.*, **36**, 1928–1940.
6. Ennifar,E., Nikulin,A., Tishchenko,S., Serganov,A., Nevskaya,N., Garber,M., Ehresmann,B., Ehresmann,C., Nikonov,S. and Dumas,P. (2000) The crystal structure of UUCG tetraloop1. *J. Mol. Biol.*, **304**, 35–42.
7. Nozinovic,S., Fürtig,B., Jonker,H.R., Richter,C. and Schwalbe,H. (2010) High-resolution NMR structure of an RNA model system: the 14-mer cUUCGg tetraloop hairpin RNA. *Nucleic Acids Res.*, **38**, 683–694.
8. Hall,K.B. (2015) Mighty tiny. *RNA*, **21**, 630–631.
9. Leontis,N.B. and Westhof,E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499–512.
10. d'Ascenzo,L., Leonarski,F., Vicens,Q. and Auffinger,P. (2017) Revisiting GNRA and UNCG folds: U-turns versus Z-turns in RNA hairpin loops. *RNA*, **23**, 259–269.
11. Borkar,A.N., Vallurupalli,P., Camilloni,C., Kay,L.E. and 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.
12. Hartlmüller,C., Günther,J.C., Wolter,A.C., Wöhnert,J., Sattler,M. and Madl,T. (2017) RNA structure refinement using NMR solvent accessibility data. *Sci. Rep.*, **7**, 5393.
13. Nichols,P.J., Henen,M.A., Born,A., Strotz,D., Güntert,P. and Vögeli,B. (2018) High-resolution small RNA structures from exact nuclear Overhauser enhancement measurements without additional restraints. *Comm. Biol.*, **1**, 61.
14. Duchardt,E. and Schwalbe,H. (2005) Residue specific ribose and nucleobase dynamics of the cUUCGg RNA tetraloop motif by MNMR 13 C relaxation. *J. Biomol. NMR*, **32**, 295–308.
15. Salmon,L., Yang,S. and Al-Hashimi,H.M. (2014) Advances in the determination of nucleic acid conformational ensembles. *Annu. Rev. Phys. Chem.*, **65**, 293–316.
16. Marušič,M., Schlagnitweit,J. and Petzold,K. (2019) RNA dynamics by NMR. *ChemBioChem*, **20**, 2685–2710.
17. Boelens,R., Koning,T., Van der Marel,G., Van Boom,J. and Kaptein,R. (1989) Iterative procedure for structure determination from proton-proton NOEs using a full relaxation matrix approach. Application to a DNA octamer. *J. Magn. Reson.*, **82**, 290–308.
18. Vögeli,B. (2014) The nuclear Overhauser effect from a quantitative perspective. *Prog. Nucl. Mag. Res. Sp.*, **78**, 1–46.
19. Nichols,P., Born,A., Henen,M., Strotz,D., Orts,J., Olsson,S., Güntert,P., Chi,C. and Vögeli,B. (2017) The exact nuclear overhauser enhancement: recent advances. *Molecules*, **22**, 1176.
20. Nichols,P.J., Born,A., Henen,M.A., Strotz,D., Celestine,C.N., Güntert,P. and Vögeli,B. (2018) Extending the Applicability of Exact Nuclear Overhauser Enhancements to Large Proteins and RNA. *ChemBioChem*, **19**, 1695–1701.

21. Giambaşu,G.M., York,D.M. and Case,D.A. (2015) Structural fidelity and NMR relaxation analysis in a prototype RNA hairpin. *RNA*, **21**, 963–974.
22. Banás,P., Hollas,D., Zgarbová,M., Jurecka,P., Orozco,M., Cheatham,T.E. III, Sponer,J. and Otyepka,M. (2010) Performance of molecular mechanics force fields for RNA simulations: stability of UUCG and GNRA hairpins. *J. Chem. Theory Comput.*, **6**, 3836–3849.
23. Bergonzo,C., Henriksen,N.M., Roe,D.R. and Cheatham,T.E. (2015) Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields. *RNA*, **21**, 1578–1590.
24. Bottaro,S., Banas,P., Sponer,J. and Bussi,G. (2016) Free energy landscape of GAGA and UUCG RNA tetraloops. *J. Phys. Chem. Lett.*, **7**, 4032–4038.
25. Tan,D., Piana,S., Dirks,R.M. and Shaw,D.E. (2018) RNA force field with accuracy comparable to state-of-the-art protein force fields. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, E1346–E1355.
26. Hummer,G. and Köfinger,J. (2015) Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.*, **143**, 243150.
27. Bottaro,S., Bengtsen,T. and Lindorff-Larsen,K. (2020) Integrating molecular simulation and experimental data: a Bayesian/maximum entropy reweighting approach. *Struct. Bioinform.*, **2112**, 219–240.
28. Kuhrova,P., Banas,P., Best,R.B., Sponer,J. and Otyepka,M. (2013) Computer folding of RNA tetraloops? Are we there yet?. *J. Chem. Theory Comput.*, **9**, 2115–2125.
29. Chen,A.A. and García,A.E. (2013) High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 16820–16825.
30. Wang,L., Friesner,R.A. and Berne,B. (2011) Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). *J. Phys. Chem. B*, **115**, 9431–9438.
31. Barducci,A., Bussi,G. and Parrinello,M. (2008) Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.*, **100**, 020603.
32. Bottaro,S., Di Palma,F. and Bussi,G. (2014) The role of nucleobase interactions in RNA structure and dynamics. *Nucleic Acids Res.*, **42**, 13306–13314.
33. Bottaro,S., Bussi,G., Pinamonti,G., Reisser,S., Boomsma,W. and Lindorff-Larsen,K. (2019) Barnaba: software for analysis of nucleic acid structures and trajectories. *RNA*, **25**, 219–231.
34. Zweckstetter,M. (2008) NMR: prediction of molecular alignment from structure using the PALES software. *Nat. Protoc.*, **3**, 679.
35. Gong,Z., Schwieters,C.D. and Tang,C. (2018) Theory and practice of using solvent paramagnetic relaxation enhancement to characterize protein conformational dynamics. *Methods*, **148**, 48–56.
36. Boomsma,W., Ferkinghoff-Borg,J. and Lindorff-Larsen,K. (2014) Combining experiments and simulations using the maximum entropy principle. *PLoS Comput. Biol.*, **10**, e1003406.
37. Beauchamp,K.A., Pande,V.S. and Das,R. (2014) Bayesian energy landscape tilting: towards concordant models of molecular ensembles. *Biophys. J.*, **106**, 1381–1390.
38. Bottaro,S., Bussi,G., Kennedy,S.D., Turner,D.H. and Lindorff-Larsen,K. (2018) Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations. *Sci. Adv.*, **4**, eaar8521.
39. Branduardi,D., Bussi,G. and Parrinello,M. (2012) Metadynamics with adaptive Gaussians. *J. Chem. Theory Comput.*, **8**, 2247–2254.
40. Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
41. Bentrude,W.G. and Hargis,J.H. (1969) Conformations of 6-membered-ring phosphorus heterocycles: the 5-t-butyl-2-oxo-1, 3, 2-dioxaphosphorinans. *J. Chem. Soc. D*, **19**, 1113b–1114.
42. Pitera,J.W. and Chodera,J.D. (2012) On the use of experimental observations to bias simulated ensembles. *J. Chem. Theory Comput.*, **8**, 3445–3451.
43. Orioli,S., Larsen,A.H., Bottaro,S. and Lindorff-Larsen,K. (2020) How to learn from inconsistencies: integrating molecular simulations with experimental data. *Comput. Approach. Understand. Dyn. Syst.: Protein Fold. Assembly*, **170**, 123–176.
44. Salmon,L., Giambasu,G.M., Nikolova,E.N., Petzold,K., Bhattacharya,A., Case,D.A. and Al-Hashimi,H.M. (2015) Modulating RNA alignment using directional dynamic kinks: application in determining an atomic-resolution ensemble for a hairpin using NMR residual dipolar couplings. *J. Am. Chem. Soc.*, **137**, 12954–12965.
45. Cesari,A., Bottaro,S., Lindorff-Larsen,K., Banáš,P., Sponer,J. and Bussi,G. (2019) Fitting corrections to an RNA force field using experimental data. *J. Chem. Theory Comput.*, **15**, 3425–3431.
46. Graf,J., Nguyen,P.H., Stock,G. and Schwalbe,H. (2007) Structure and dynamics of the homologous series of alanine peptides: a joint molecular dynamics/NMR study. *J. Am. Chem. Soc.*, **129**, 1179–1189.
47. Bonomi,M., Heller,G.T., Camilloni,C. and Vendruscolo,M. (2017) Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.*, **42**, 106–116.
48. Reißer,S., Zucchelli,S., Gustincich,S. and Bussi,G. (2019) Conformational ensembles of an RNA hairpin using molecular dynamics and sparse NMR data. *Nucleic Acids Res.*, **48**, 1164–1174.
49. Sripakdeevong,P., Cevec,M., Chang,A.T., Erat,M.C., Ziegeler,M., Zhao,Q., Fox,G.E., Gao,X., Kennedy,S.D., Kierzek,R. *et al.* (2014) Structure determination of noncanonical RNA motifs guided by 1 H NMR chemical shifts. *Nat. Met.*, **11**, 413.
50. Piccini,G., Mendels,D. and Parrinello,M. (2018) Metadynamics with discriminants: a tool for understanding chemistry. *J. Chem. Theory Comput.*, **14**, 5040–5044.
51. Brandt,S., Sittel,F., Ernst,M. and Stock,G. (2018) Machine learning of biomolecular reaction coordinates. *J. Phys. Chem. Lett.*, **9**, 2144–2150.
52. Fleetwood,O., Kasimova,M.A., Westerlund,A.M. and Delemotte,L. (2020) Extracting molecular insights from conformational ensembles using machine learning. *Biophys. J.*, **118**, 765–780.
53. Escobedo,A., Topal,B., Kunze,M.B.A., Aranda,J., Chiesa,G., Mungianu,D., Bernardo-Seisedos,G., Eftekharzadeh,B., Gairi,M., Pieratelli,R. *et al.* (2019) Side chain to main chain hydrogen bonds stabilize polyglutamine helices in transcription factors. *Nat. Commun.*, **10**, 2034.
54. Crehuet,R., Buigues,P.J., Salvatella,X. and Lindorff-Larsen,K. (2019) Bayesian-maximum-entropy reweighting of IDP ensembles based on NMR chemical shifts. *Entropy*, **21**, 898.
55. Bonomi,M., Bussi,G., Camilloni,C., Tribello,G., Bonas,P., Barducci,A., Bernetti,M., Bolhuis,P.G., Bottaro,S., Branduardi,D. *et al.* (2019) Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Met.*, **16**, 670–673.