THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Article

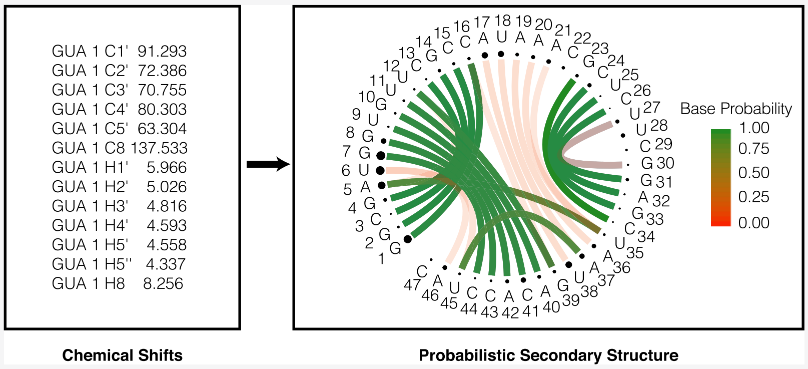
Probabilistic Modeling of RNA Ensembles Using NMR Chemical Shifts

Kexin Zhang and Aaron T. Frank*

Cite This: *J. Phys. Chem. B* 2021, 125, 9970–9978

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information



GUA	1	C1'	91.293
GUA	1	C2'	72.386
GUA	1	C3'	70.755
GUA	1	C4'	80.303
GUA	1	C5'	63.304
GUA	1	C8	137.533
GUA	1	H1'	5.966
GUA	1	H2'	5.026
GUA	1	H3'	4.816
GUA	1	H4'	4.593
GUA	1	H5'	4.558
GUA	1	H5''	4.337
GUA	1	H8	8.256

Chemical Shifts

Probabilistic Secondary Structure

ABSTRACT: NMR-derived chemical shifts are structural fingerprints that are sensitive to the underlying conformational distributions of molecules. Thus, chemical shift data are now routinely used to infer the dynamical or conformational ensembles of peptides and proteins. However, for RNAs, techniques for inferring their conformational ensembles from chemical shift data have received less attention. Here, we used chemical shift data and the Bayesian/maximum entropy (BME) approach to model the secondary structure ensembles of several single-stranded RNAs. Inspection of the resulting ensembles indicates that the secondary structure of the highest weighted (most probable) conformer in the ensemble typically resembled the known NMR structure. Furthermore, using apo chemical shifts measured for the HIV-1 TAR RNA, we found that our framework reproduces the expected structure yet predicts the existence of a previously unobserved base pair, which we speculate may be sampled transiently. We expect that the chemical shift-based BME (CS-BME) framework we describe here should find utility as a general strategy for modeling RNA ensembles using chemical shift data.

■ INTRODUCTION

Determining the conformational states accessible to RNA is a critical first step in establishing links between their sequence, structure, dynamics, and biological function(s).[1] Modeling the RNA conformational ensemble is, however, challenging because the number of unknowns exceeds that which can be measured experimentally. Also, many conformers exist in low abundance and have short lifetimes and therefore fall outside experimental detection. Although recent advances in NMR such as relaxation dispersion and saturation transfer have made it possible to study these previously "invisible" yet functionally important conformational states,[2,3] they still remain difficult to characterize. NMR does, however, provide access to chemical shift signatures of RNA that one could, in principle, use to infer their conformational ensembles, which may include conformers that resemble these transient states. Although there has been some initial exploration of the utility of chemical shifts in inferring RNA ensembles,[4] compared to proteins,[5−21] the use of chemical shifts to infer conformational ensembles of RNA has remained relatively underexplored.

There are two frameworks that one could use to infer conformational ensembles starting from a set of ensemble-averaged experimental data like chemical shifts.[22] First, experimental data could be used as restraints during folding simulations to guide algorithms to regions of conformational space that maximize the agreement between measured and simulated data.[23−27] Second, experimental data could be used to reweight an initial ensemble of structures such that ensemble-averaged, back-calculated data agree with the observed experimental data.[28−30] Compared with restraining, reweighting methods are attractive because they are fast and flexible, and a single initial ensemble can be reweighted using multiple data sets, allowing easy comparison of the resulting

Received: June 27, 2021
Revised: August 9, 2021
Published: August 27, 2021

https://doi.org/10.1021/acs.jpcb.1c05651

ensembles, which are important in cases where each data set corresponds to the RNA under distinct physicochemical conditions.

Here, we applied the Bayesian/maximum entropy (BME)[31,32] reweighting technique to model the ensembles of RNA using chemical shift data. The goal of BME is to find a new distribution of conformations within a given library, such that the data back-calculated from the reweighted library (or the conformational/dynamical ensemble) exhibit good agreement with available experimental data. Using chemical shift-based BME (CS-BME), we generated secondary structure ensembles of 15 single-stranded RNAs. To achieve this, we first trained a set of models that predicted chemical shifts from secondary structure models. Then, for each RNA, we generated a conformational library containing secondary structures that the RNAs are likely to adopt. Using our chemical shift predictors, we next computed chemical shifts associated with each conformer within the conformational library. For each RNA, the chemical shifts back-calculated from the conformers and their associated experimental chemical shifts were then used to reweight each conformational library and produce a conformational ensemble. In general, we found that the most dominant structure in the ensemble closely resembled the known structure. In addition, for the HIV-1 TAR, we analyzed its dynamical ensemble and predict that it may sample a state harboring a previously unobserved base pair.

■ METHODS

RNA Chemical Shifts from Secondary Structures. Our objective in this study was to implement a framework for using chemical shift data to infer RNA secondary structure ensembles. Central to this technique is the comparison between measured chemical shifts and chemical shifts computed from secondary structure models. As such, we trained secondary structure to chemical shifts (**SS2CS**) models, which take the RNA secondary structure as input and output the predicted chemical shifts for different nucleus types. To train the **SS2CS** models, we compiled a data set composed of the secondary structure and chemical shifts for 108 RNAs. The secondary structures were retrieved from model 1 of each NMR bundle using the program DSSR from the 3DNA suite.[33] The NMR chemical shifts were downloaded from the Biological Magnetic Resonance Data Bank (BMRB: http://www.bmrb.wisc.edu/). As we have done in previous work,[34] we corrected ^{13}C data that were predicted to contain systematic referencing errors.[35]

To predict the non-exchangeable chemical shifts of, namely, H1′, H2′, H3′, H4′, H2, H5, H5′, H5″, H6, H8, C1′, C2′, C3′, C4′, C5′, C2, C5, C6, and C8, we first constructed a data set for individual nucleus types. Briefly, for each nucleus type, the chemical shifts associated with this nucleus type from all the RNAs, along with the secondary structure features of each residue, were combined into a large data set. The secondary structural features we encoded from the input structure file include (for residue *i*) (1) the length of the RNA; (2) the residue types of residues *i*, *i* − 1, and *i* + 1; (3) the residue type of residue *i*'s pairing partner *j*, if exists; (4) the residue types of the pairing partner of residues *i* − 1 and *i* + 1, if exist; (5) the residue types of residues *j* − 1 and *j* + 1, if exist; and (6) the residue types of the pairing partner of residues *j* − 1 and *j* + 1, if exist (Figure 1). The features we used to predict the chemical shifts for a central residue *i* consist of the secondary structure features of residues *i* − 3, *i* − 2, *i* − 1, *i*, *i* + 1, *i* + 2, and *i* + 3. Thus, even for single-stranded regions, the features carry information about the unique environment of around individual nucleotides by capturing the extended nearest-neighbor sequence context and local structure associated with each residue and its neighbors. These features used contain similar information to the subset of the "attributes" used previously to train similar SS2CS predictors, namely, RNAShifts[36] and RNAShifts2.[37] These attributes, however, contain additional higher structural information that includes the position in a tetraloop, multiplets, stacking, and pseudoknots.[37] Our secondary structure features do not contain analogous information.

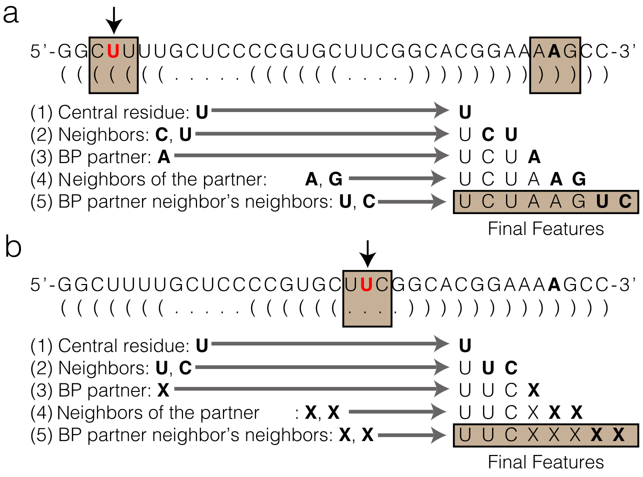


Figure 1. Illustration of the feature extraction technique we applied to describe individual residues in an RNA based on their secondary structure. Shown are examples of the secondary structure features associated with a residue in the (a) stem and (b) single-stranded regions. Here, we used "X" to denote a null residue. To estimate chemical shifts for a central residue *i*, our predictors take as input long-range sequence-structure information that consists of the secondary structure features of residues *i* − 3, *i* − 2, *i* − 1, *i*, *i* + 1, *i* + 2, and *i* + 3.

We applied this featurization to all residues in each RNA and mapped these features to the corresponding chemical shifts to construct our final data set (Table S3). To train **SS2CS** predictors, we used the random forest approach.[38] When training a random forest model, a collection (or ensemble) of independent decision trees are trained such that the aggregate predictions of the ensemble agree with the target, which in this study are measured chemical shifts. We trained 19 random forest models, one for each nucleus type (H1′, H2′, H3′, H4′, H2, H5, H5′, H5″, H6, H8, C1′, C2′, C3′, C4′, C5′, C2, C5, C6, and C8). To predict chemical shifts for a given residue *i*, the secondary structure features for that residue *i* and *i* − 3, *i* − 2, *i* − 1, *i* + 1, *i* + 2, and *i* + 3 are combined and fed into the **SS2CS** predictors for each nucleus. The resulting estimates for each nucleus type are then compiled to provide estimates for the chemical shifts associated with that residue. Because the features used to predict chemical shifts contain long-range sequence and structure information that captures the local context within which a given nucleotide in the RNA resides, chemical shift prediction is dependent on this context.

Reweighting RNA Secondary Structures Using Chemical Shifts. To model the secondary structure ensembles of RNAs using chemical shift data, we employed a probabilistic approach in which we assigned weights to a collection (or library) of RNA structures conditioned on available and experimentally measured chemical shift data. To do this, we utilized the BME. BME has emerged as a robust framework for integrating experimental measurements with simulation data like computed chemical shifts. The central idea is to optimize the weights assigned to each member or conformer in a conformational library of (a priori) structures to maximize the agreement between the reweighted ensemble-averaged properties and the experimental observations, while minimally changing the initial weights (or priors). Techniques developed based on maximum entropy reweighting have been successfully applied to protein and RNA for structure determination and force field refinement.[39−41]

Here, we applied a BME[31] approach where the error or uncertainty of experimental data is also taken into account. From previous studies,[42−44] it can be shown that the optimal weights will minimize the following loss function:

$$\mathcal{L}(w_1 ... w_n) = \frac{m}{2}\chi^2(w_1 ... w_n) - \theta S_{\mathrm{rel}}(w_1 ... w_n) \tag{1}$$

in which

$$\chi^2(w_1 ... w_n) = \frac{1}{m}\sum_i^m \left(\frac{\left[\left(\sum_j^n w_j F(x_j)\right) - F_i^{\mathrm{exp}}\right]^2}{\sigma_i^2} \right) \tag{2}$$

$$S_{\mathrm{rel}} = -\sum_i^n w_j \ln\left(\frac{w_j}{w_j^0}\right) \tag{3}$$

Here, w_j and w_j^0 are the new and original weights of the jth member in the ensemble, respectively, n is the population of the ensemble, σ_i is the uncertainty of the measurement F_i^{exp}, and $F(x_j)$ is the back-calculated property from the jth member in the ensemble. In this study, σ_i (the uncertainty) corresponds to the error in chemical shifts computed using our **SS2CS** predictors and does not necessarily correspond to the specific error in a system. Additional sources of error in the experimental data may similarly be present, but here, we assume that they are small relative to the errors in the computed chemical shifts. In the loss function $\mathcal{L}$ (eq 1), the first term (χ^2) describes the agreement between the experimental data and the back-calculated properties from structure models and the second term (S_{rel}) describes the deviation of the new weights from the original weights. The original weights, in our case, are $1/n$ if there are n members in the ensemble. Others have shown that the optimal weights can also be calculated through Maximum A Posteriori (MAP) estimation[14] by minimizing the negative log-likelihood of the posterior distribution. This method is called Bayesian ensemble refinement, and it is mathematically equivalent to the maximum entropy with error or the BME approach described above.

Testing Chemical Shift-Based Reweighting of RNA Secondary Structures. To test the ability of chemical shift-based reweighting to resolve RNA secondary structures, we applied BME to reweight the conformational libraries of 15 RNAs based on their sets of measured and computed chemical shifts. We first created an ensemble of low-energy secondary structures for a given RNA sequence using the tool MC-Fold from the MC-Sym suite[45] (as it allows the formation of pseudoknotted structures). However, for large RNAs, it may take a long time to generate decoys using MC-Fold. For this reason, we used AllSub from the RNAstructure modeling suite instead for the largest RNA in our data set, the HIV-1 RNA (PDB ID: 2N1Q), which has 155 nts. Using MC-Fold, we generated 10 different structures whose folding free energies are within 30% of the lowest energy structure. The exception was the fluoride riboswitch for which we combined decoys generated from AllSub and MC-Fold to ensure that the pool of structures was diverse and contained both pseudoknotted and non-pseudoknotted structures. The structures were then combined with the native structure to form the final conformational library. If the native structure contained non-canonical base pairs, then native structures with and without the non-canonical base pairs are included in the final conformational library.

Next, for each RNA, the **SS2CS** predictors were used to predict the chemical shifts of all non-exchangeable nuclei in each conformer of the corresponding ensemble. After generating chemical shift predictions using **SS2CS**, we then applied BME to the measured and computed chemical shifts and assigned weights to individual conformers in the structural ensemble. According to eq 1, θ is a global scaling factor that controls the relative contribution of the entropy terms in the overall loss function $\mathcal{L}$. It reflects the trade-off between two terms: (1) χ^2, which is the agreement between experimental data and predicted chemical shifts, and (2) S_{rel}, which is the deviation of the new distribution from the original uniform distribution. Theoretically, the smaller θ is, the more $\mathcal{L}$ is dependent on χ^2 and the better agreement we should be able to achieve between experimental reweighted ensemble-averaged chemical shifts. In reality, we found that for some RNAs, θ and χ^2 may not be positively correlated. To find the best θ, we scanned different values from 1.0 to 200.0 (with a step of 1.0) and calculated χ^2 (using eq 2) at each θ.[31] The optimal θ was chosen as the one that minimized χ^2 versus θ (Figure S3a). Across the 15 RNAs in our CS-BME benchmark set, the optimal θ ranged between 3.0 and 102.0 (with a mean of 28.4) (Figure S3b).

Visualizing RNA Secondary Structures. Throughout this article, we used circular plots to display RNA secondary structures. In these plots, the RNA residues are arranged along a circle, and the base pairs are represented as lines between individual residues. To summarize the CS-BME ensembles, we used circular secondary structure base pair probability plots ($\mathrm{CS^2BP^2}$plots),[46] circular secondary structure plots in which base pair probability information is encoded in the color and thickness of the lines respecting base pairs. We generated these plots using in-house R-scripts.

■ RESULTS AND DISCUSSION

Random forest models predict chemical shifts from secondary structures with similar accuracy to models that estimate them from atomic structures. We began by developing 2D chemical shift estimators. To accomplish this, we used existing proton (^{1}H) and carbon (^{13}C) chemical shifts for 108 RNAs to train a set of machine learning models that estimate the non-exchangeable H1′, H2′, H3′, H4′, H2, H5, H5′, H5″, H6, H8, C1′, C2′, C3′, C4′, C5′, C2, C5, C6, and C8 chemical shifts based on features extracted from the 2D structure of an RNA (see Methods; Tables S1 and S3). We trained machine learning models for each nucleus type using random forest regression, which aggregates the predictions from an ensemble of decision trees. We chose the random forest method because in preliminary testing using cross-

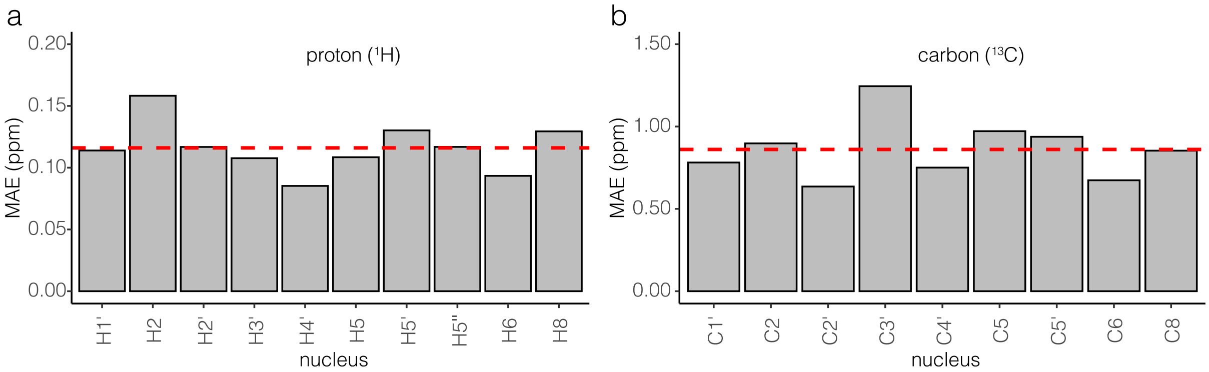


Figure 2. Estimated accuracy of our secondary structure chemical shift predictors. The barplots show the MAE (mean absolute error) between measured chemical shifts and chemical shifts computed from secondary structures in a validation set (Table S1). The red dashed lines indicate the average MAE.

Table 1. Summary of CS-BME Ensembles of RNAs in the Validation Set

PDB ID	description	size	N	$\Delta\chi^2$	weight	TPR	PPV
1HWQ	VS ribozyme substrate stem loop	30	11	−0.299	0.42	0.75	1.00
1YMO	P2b-P3 pseudoknot from human telomerase	47	12	−0.179	0.30	0.88	1.00
2L3E	P2a-J2a/b-P2b of human telomerase RNA	35	11	−0.632	0.57	0.86	0.92
2LU0	κ-ζ region of S.cerevisiae group II intron ai5(γ)	49	12	−1.110	0.80	0.84	1.00
2LUB	helix H1 of the human HAR1	37	11	−0.892	0.86	0.94	1.00
2N1Q	HIV-1 core packaging signal	155	14	−0.119	0.29	0.53	0.73
2N6X	CssA5 of CssA thermometer	43	12	−1.179	0.43	0.94	1.00
2N7X	microRNA 20b pre-element	23	11	−0.380	0.62	1.00	0.90
2N82	microRNA 20b pre-element with Rbfox RRM	23	12	−0.722	0.53	0.83	0.62
2NBY	J domain of EMCV IRES	39	11	−0.190	0.49	0.88	1.00
2NC0	St domain of EMCV IRES	28	11	−0.149	0.27	0.83	1.00
5KH8	apo state fluoride riboswitch	47	18	−0.618	0.61	0.94	1.00
5KMZ	tetrahymena telomerase RNA pseudoknot	31	11	−0.189	0.27	0.33	0.31
5V16	enterovirus IRES domain to stimulate viral translation	41	12	−0.344	0.43	0.82	0.93
6GZK	tetramethylrhodamine (TMR) aptamer 3	48	13	−1.204	0.53	0.76	1.00
mean		45	13	−0.547	0.50	0.81	0.90

validation on the training set, it outperformed other baseline methods (Figure S1). Shown in Figure 2 is the mean absolute error (MAE) between measured chemical shifts and chemical shifts computed from the secondary structure in a test set; the test set consisted of 20% of the data that were excluded from the data set used to train the random forest predictors (Table S1). For protons, the MAE ranged between 0.09 and 0.16 ppm, with a mean of 0.12 ppm (Figure 2a), and for carbons, the MAE ranged between 0.64 and 1.25 ppm, with a mean of 0.86 ppm (Figure 2b). These MAE values show similar trends to those calculated over the independent set of RNAs used to benchmark the CS-BME approach described below (Figure S2). In this case, the MAEs were higher and ranged between 0.11 and 0.25 ppm, with a mean of 0.16 ppm for protons (Figure S2a) and 0.66 and 2.01 ppm, with a mean of 1.17 ppm, for carbons (Figure S2b). Overall, these MAE values are comparable to the accuracy achieved by RNAShifts2 (~0.12 and ~0.80 ppm and protons and carbons, respectively[37]). They are also similar to the accuracy achieved using models that, instead of secondary structures, predict chemical shifts from atomic structures of RNA; the MAEs for such predictors are ~0.15 and 0.81 ppm for protons and carbons, respectively.[47]

The secondary structure of the most probable conformer within CS-BME ensembles typically resembles the secondary structure of the reference NMR model. Having trained robust 2D chemical shift predictors, we next attempted to assess the utility of chemical shifts in modeling 2D, RNA dynamical ensembles. Dynamical ensembles are defined by the structure and relative population of states that a molecule is likely to populate. Specifically, we sought to use a probabilistic modeling framework to assign weights to conformers in conformational libraries of RNAs by comparing measured chemical shifts to chemical shifts computed using our estimators. First, we compiled libraries of likely secondary structures that can be adopted by each of the 15 RNAs in our validation/benchmark set. The number of conformers in these libraries ranged between 11 and 18 (Table 1). For each conformer in the libraries, we computed the set of non-exchangeable proton and carbon chemical shifts and then used BME to assign weights to each conformer; using BME, we assigned weights to conformers in a library in a manner that maximized agreement between measured chemical shifts and chemical shifts computed from our **SS2CS** models while also accounting for the inherent uncertainty in the predictions (Methods).

To check if BME reweighting was successful, we computed $\Delta\chi^2$, which is the difference of χ^2 (eq 2) after and before BME reweighting. Negative values of $\Delta\chi^2$ indicate instances where BME reweighting identified conformational weights that improved the agreement between measured and computed, ensemble-averaged chemical shifts. Across the 15 RNAs in our validation set, $\Delta\chi^2$ ranged between −0.149 and −1.204 with a

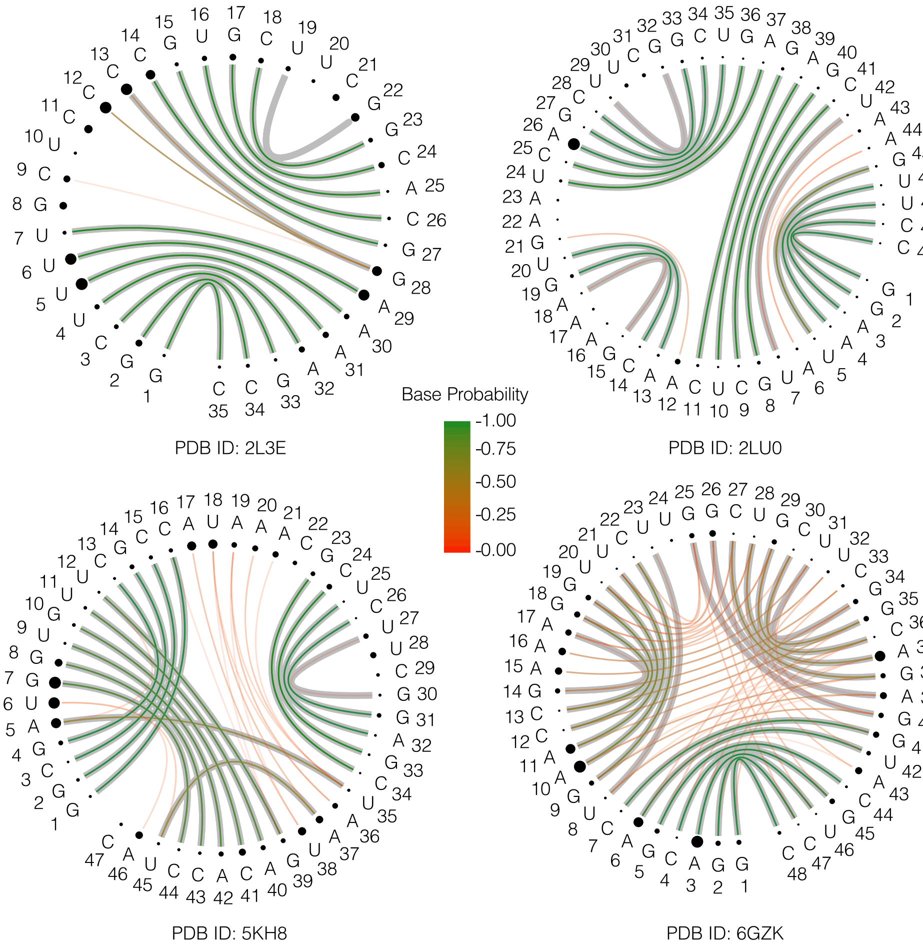
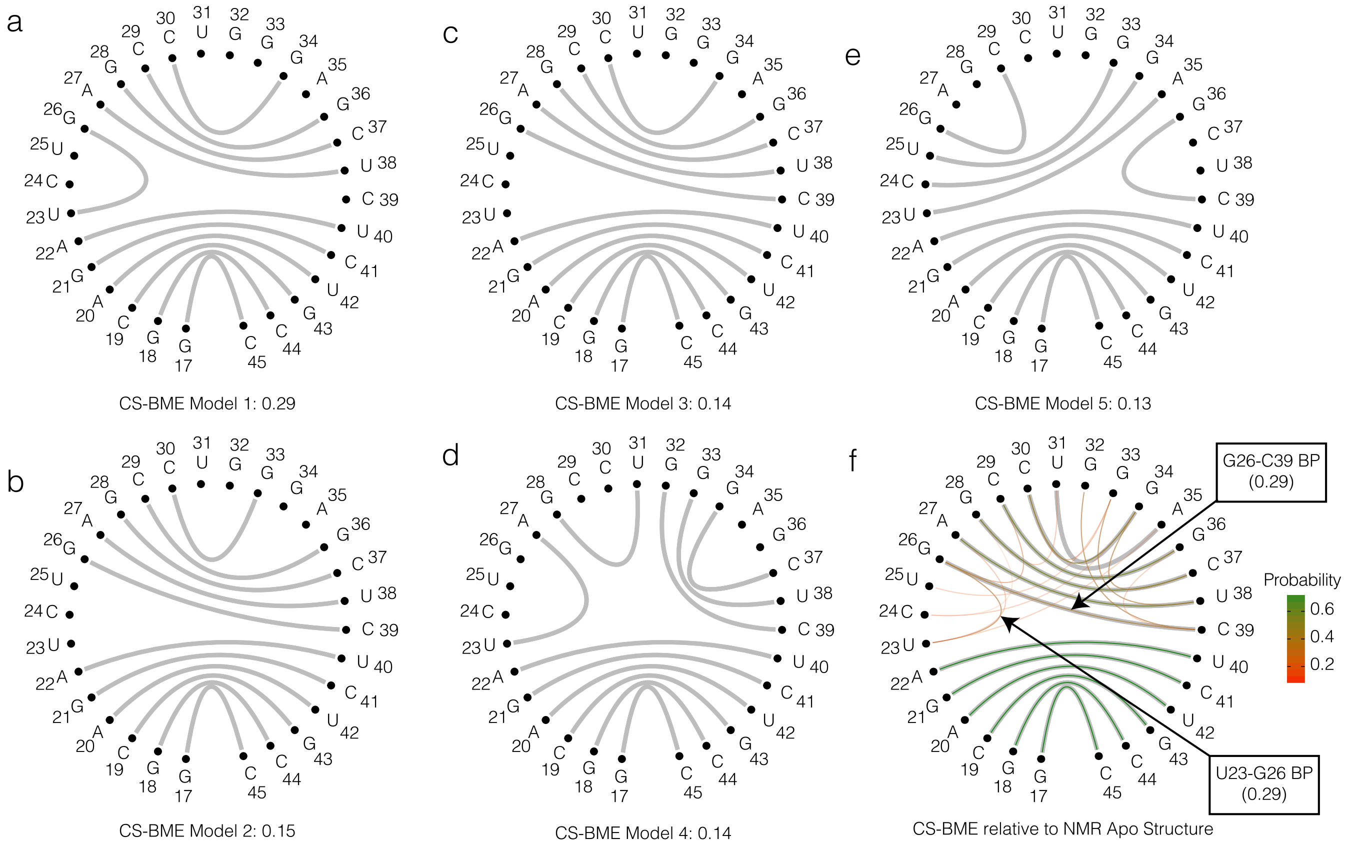


Figure 3. CS-BME-derived CS^2BP^2Plots for the core domain of the human telomerase RNA (PDB ID: 2L3E), the group II intron ai5γ RNA (PDB ID: 2LU0), the fluoride riboswitch RNA (PDB ID: 5KH8), and the tetramethylrhodamine (TMR) aptamer 3 RNA (PDB ID: 6GZK). In these plots, the size of the spheres encodes information about the relative residue-wise chemical shift error; the larger the sphere, the larger the chemical shift error associated with a given residue. The thick lines that are shaded gray indicate base pairs in the reference NMR model.

mean value of −0.547, indicating that BME reweighting was successful.

Reported in Table 1 are the true positive rate (TPR) and positive predictive values (PPV) between the secondary structure in reference NMR models and the secondary structure of the conformers that were assigned the highest CS-BME weight. TPR and PPV values approaching 1 correspond to conformers that are identical to the reference secondary structure. Across the 15 RNAs, the mean TPR and PPV were 0.81 and 0.90, respectively. For 13 of the 15 RNAs, the conformers with the highest CS-BME weight had TPR or PPV values greater than 0.80. The exceptions were 2N1Q and 5KMZ, the two RNAs for which only proton chemical shifts were available and thus used for reweighting. Together, our results indicate that CS-BME reweighting was generally successful, resulting in conformational weights that improved the agreement between measured and computed chemical shifts. Furthermore, the highest weighted conformer in the CS-BME ensembles resembled the reference NMR structures, indicating that the CS-BME reweighting of the libraries led to ensembles in which the ground-state (GS; most probable) structures were consistent with GS-state structures observed via NMR solution studies.

To highlight the latter point further, we show in Figure 3 circular secondary structure base pair probability plots (CS^2BP^2Plots)[46] of four representative RNAs, namely, the core domain of the human telomerase RNA (PDB ID: 2L3E), the group II intron ai5γ RNA (PDB ID: 2LU0), the fluoride riboswitch RNA (PDB ID: 5KH8), and the TMR aptamer 3 RNA (PDB ID: 6GZK). The complete set of CS^2BP^2Plots can be found in the Supporting Information (Figures S4−S18). These modified CS^2BP^2Plots contain information about the probability of contacts between any two residues across the CS-BME ensemble. In these plots, we encode information about the probabilities of finding two bases paired in terms of both the thickness of the lines connecting pairs of residues and their color, thick and green corresponding to higher probabilities. Also, in these plots, the size of the sphere associated with each residue encodes information about the relative chemical shift error of that residue across the CS-BME ensemble; the larger the sphere, the larger the mean error between measured chemical shifts and ensemble-averaged chemical shifts computed for atoms residing on the associated residue. In these CS^2BP^2Plots, the pair probabilities are derived from the CS-BME weights by summing base pair probabilities across the respective ensembles. Inspection of these plots confirms that, across the ensembles, high-probability base pairs coincide with the base pairs found in the reference NMR models. Interestingly, many of the base pairs that are missed, that is, those base pairs that across the ensemble have low base



CS-BME Model 1: 0.29

CS-BME Model 3: 0.14

CS-BME Model 5: 0.13

CS-BME Model 2: 0.15

CS-BME Model 4: 0.14

CS-BME relative to NMR Apo Structure

Figure 4. CS-BME results for apo-TAR. Shown in (a−e) are circular plots of the secondary structure of the five highest weighted conformers in the CS-BME TAR ensemble. Shown in (f) is the CS^{2}BP^{2}Plot of apo TAR, which was derived from its CS-BME ensemble. In (f), the thick lines that are shaded gray indicate base pairs in the reference NMR model (PDB ID: 7JU1).[48]

pair probabilities yet are present in the reference NMR model, are flanking base pairs in terminal regions of their respective stems or correspond to non-canonical pairs. We initially speculated that these discrepancies were the result of local errors in our predicted chemical shifts. However, we saw no obvious pattern between the location of these missed base pairs and the chemical shift errors (encoded by the sphere size in Figure 3).

The CS-BME ensemble of apo HIV-1 TAR predicts the existence of a previously unobserved U23-G26 base pair. We concluded our study by applying the CS-BME approach to the HIV-1 transactivating response (TAR) element RNA. Because it is a small yet dynamically complex RNA,[49] TAR is an excellent model system for developing techniques to model RNA dynamical ensembles.[48,50,51] Using measured apo-TAR chemical shifts,[48] we modeled the secondary structure ensemble of TAR using CS-BME. To achieve this, we generated a conformational library of TAR using an in-house tool, ss-Sampler (https://github.com/atfrank/SS-Sampler). Given an input sequence, ss-Sampler generates structures that maximize base-pairing and stem-stacking interactions. To achieve this, ss-Sampler enumerates the stems associated with a given sequence and then assembles complete structures using a genetic algorithm guided by a simple hydrogen-bonding counting and base-stacking fitness function.[52] Because the genetic algorithm is stochastic, many cycles can be run to generate a collection of secondary structures and so sample the conformational space of RNAs. Over the course of 200 cycles, we generated a collection of 48 unique structures using ssSampler. We then selected the conformers with energies that were at least 1 standard deviation lower than the median energy of the initial set of 48 conformers, and reweighted the resulting 12-membered library using CS-BME and apo-TAR chemical shifts.

Shown in Figure 4a−e are the top five highest weighted conformers in the CS-BME ensemble of TAR (which accounted for 0.85 of the total conformational weights). The secondary structure of top three conformers resembles the typical stem-bulge-stem-loop structure of TAR. The CS-BME-derived CS^{2}BP^{2}Plot of apo TAR is shown in Figure 4f. The plot captures the ensemble view of TAR base pairs derived from the CS-BME reweighting. In general, the lower stem region (residues 17−22 and 40−45) is predicted to be more stable than the upper stem region (residues 26−29 and 36−39) of TAR. Notably, the CS-BME ensemble predicts G26-C39, which is observed in the reference NMR structure of apo TAR, to have a lower-than-expected base pair probability ($p \sim 0.29$). This is because the ensemble also predicts that G26 forms a novel base pair with U23 ($p \sim 0.29$). Interestingly, NMR studies of wt-TAR have detected a spectroscopic signature of a yet-to-be-characterized transient state that exhibits relaxation−dispersion at G26 C8 and A27 C1′.[53] The electronic environment near these sites is likely to be altered by the formation of the U23-G26 base pair. It is intriguing to speculate that TAR may sample a structure like model 1 (Figure 4a) and that such a structure may in part explain this observed dispersion. Further work will be needed to test whether TAR transiently samples a state that harbors a U23-G26 base pair.

■ DISCUSSION AND CONCLUSIONS

In this study, we explored the use of chemical shift data to model the secondary structure, dynamical/conformational ensemble of RNAs. First, we trained estimators that take the secondary structure of an RNA as input and output estimates of the non-exchangeable proton and carbon chemical shifts. We found that we could estimate these chemical shifts with an accuracy that is on par with models that estimate chemical shifts from atomic coordinates. Next, using our secondary structure-based chemical shift predictors, we explored the utility of chemical shifts to reweight conformational libraries to generate ensemble descriptions of RNAs. Across a set of 15 RNAs, we used measured and computed chemical shifts along with the BME approach (CS-BME) to construct ensembles. The resulting ensembles recover most of the base pairs in the solution NMR models of these RNAs. Finally, we applied CS-BME to generate secondary structure ensembles of apo-TAR. The ensemble predicts the existence of a previously unobserved and, possibly, a transient U23-G26 base pair.

CS-BME is a potentially powerful approach to model the RNA dynamical ensembles at the secondary structure level and potentially at the atomic level. However, CS-BME has several limitations. First, as a reweighting approach, CS-BME is sensitive to sampling. If sampling is incomplete, the resulting ensemble could fail to include conformers that, although not sampled (in silico), may actually be sampled in solution. Second, BME contains a hyperparameter θ that must be tuned.[31] In this study, we scanned θ values and identified the largest θ value that minimized χ^2. The weights associated with this θ value were then used to model the ensemble. Third, CS-BME is limited by the accuracy with which we could compute chemical shifts. Compared to experimental chemical shift errors, our proton and carbon prediction errors, 0.12 and 0.86 ppm, are large. Fortunately, BME uses the estimated uncertainty to guide weight optimization. Despite this, because the large fraction of base pairs in most RNAs correspond to canonical, Watson−Franklin−Crick base pairs, our predictors may produce large chemical shift errors for residues involved in non-canonical pairing (e.g., Figure S4; G6-A25 and A7-G24 base pairs). In some instances, these larger errors may negatively impact the ability of CS-BME ensembles to recover non-canonical base pairs.

More accurate predictors should lead to higher-quality ensembles and an enhanced ability to resolve conformational states of RNA, including those containing non-canonical pairs. A barrier toward realizing more accurate predictors is the scarcity of paired data sets of chemical shifts and structures of RNA. More fundamentally, because measured chemical shifts are conformationally averaged, the mapping of individual structures to measured chemical shifts, as has been done by us when training our predictors, introduces errors that limit the accuracy of the resulting predictors. One path forward would be to rely on data in which one-to-one structure-shift mapping is guaranteed. This could, in principle, be achieved by mapping individual structures to quantum mechanically derived chemical shifts and then training chemical shift predictors using the resulting chemical shift-structure database. Finally, the CS-BME framework we tested relies on assigned chemical shifts. Assigning chemical shifts is currently the bottleneck in NMR spectroscopy. However, histogram-based BME might be a path toward leveraging unassigned 2D NMR chemical shift data for CS-BME.

An additional caveat to our CS-BME approach is that it assumes that the distinct conformational states that an RNA samples are in fast exchange, with the overall chemical shift signature of the RNA being a weighted average over these states. Under slow exchange, certain peaks may have distinct chemical shifts associated with each state an RNA samples. If these peaks can be assigned and resolved, they could be used to compile distinct chemical shift data sets, and CS-BME could then be carried out separately using each of these chemical shift data sets.

Despite these limitations, CS-BME provides a framework to construct a probabilistic representation of RNA secondary structure based on NMR chemical shift data. These can be used to essentially convert NMR chemical shifts into a base pair probability matrix, which one can use to construct probabilistic restraints for RNA structure prediction or selection. Though CS-BME currently requires assigned chemical shifts as an input, it still is a valuable framework for generating hypotheses about the likely conformational states RNA samples based on their chemical shifts, particularly in cases where NMR spectroscopy was used for mechanistic studies and not structure determination. It should also prove helpful when chemical shifts are the only available data (for instance, when probing RNA transient states). Moreover, CS-BME can generate predictions to corroborate or falsify structural predictions generated using other orthogonal techniques.

Finally, we stress that CS-BME is a general technique. Thus, it can be powered using chemical shifts computed using predictors other than the ones we implemented in this study, for example, RNAShifts2.[37] Also, one can use it with chemical shifts computed using 3D-based chemical shift prediction methods, opening up the possibility of using CS-BME to model the atomic ensembles of RNA.[47,54−58]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.1c05651.

> Brief description and access information for the tools (SS2CS, SS-Sampler, and CS-Reweight) associated with this study; PDB IDs for the SS2CS training set; PDB IDs for the CS-BME benchmark set; breakdown of chemical shift data in the training set; SS2CS accuracy breakdown on a random test set and the CS-BME benchmark set; illustrative plot of χ^2 versus θ; CS^2BP^2Plots for 1HWQ; CS^2BP^2Plots for 1YMO; CS^2BP^2Plots for 2L3E; CS^2BP^2Plots for 2LU0; CS^2BP^2Plots for 2LUB; CS^2BP^2Plots for 2N1Q; CS^2BP^2Plots for 2N6X; CS^2BP^2Plots for 2N7X; CS^2BP^2Plots for 2N82; CS^2BP^2Plots for 2NBY; CS^2BP^2Plots for 2NC0; CS^2BP^2Plots for 5KH8; CS^2BP^2Plots for 5KMZ; CS^2BP^2Plots for 5V16; and CS^2BP^2Plots for 6GZK (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Aaron T. Frank − *Biophysics Program, University of Michigan, Ann Arbor, Michigan 48109, United States;* orcid.org/0000-0002-7782-2200; Phone: (734) 615-2053; Email: afrankz@umich.edu

Author

Kexin Zhang − *Chemistry Department, University of Michigan, Ann Arbor, Michigan 48109, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.1c05651

Author Contributions

A.T.F. conceived the project, K.Z. and A.T.F. conducted the computational experiments, and K.Z. and A.T.F. analyzed the results and wrote the manuscript. All authors reviewed the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the members of the Frank lab for many valuable discussions about this work. This research was supported by start-up funds from the University of Michigan and by the National Science Foundation grant MCB-2046005 (to A.T.F.). We acknowledge support for this work through Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562; the allocation identification for this project was TG-MCB190114. We specifically thank the consultants Eroma Abeysinghe, Suresh Marru, and Sudhakar Pamidighantam for their assistance with bringing the SMALTR Science Gateway online and adding SS2CS and CS-Reweight to SMALTR. The latter was made possible through the XSEDE Extended Collaborative Support Service (ECSS) program. SMALTR gateway is hosted with the SciGaP gateway platform, which is supported by the National Science Foundation award number 1339774.

■ REFERENCES

(1) Ganser, L. R.; Kelly, M. L.; Herschlag, D.; Al-Hashimi, H. M. The roles of structural dynamics in the cellular functions of RNAs. *Nat. Rev. Mol. Cell Biol.* **2019**, *20*, 474−489.

(2) Baldwin, A. J.; Kay, L. E. NMR spectroscopy brings invisible protein states into focus. *Nat. Chem. Biol.* **2009**, *5*, 808.

(3) Marušič, M.; Schlagnitweit, J.; Petzold, K. RNA dynamics by NMR spectroscopy. *ChemBioChem* **2019**, *20*, 2685−2710.

(4) Frank, A. T.; Horowitz, S.; Andricioaei, I.; Al-Hashimi, H. M. Utility of 1H NMR chemical shifts in determining RNA structure and dynamics. *J. Phys. Chem. B* **2013**, *117*, 2045−2052.

(5) Kragelj, J.; Ozenne, V.; Blackledge, M.; Jensen, M. R. Conformational Propensities of Intrinsically Disordered Proteins from NMR Chemical Shifts. *ChemPhysChem* **2013**, *14*, 3034−3045.

(6) Krzeminski, M.; Fuentes, G.; Boelens, R.; Bonvin, A. M. J. J. MINOES: A new approach to select a representative ensemble of structures in NMR studies of (partially) unfolded states. Application to Δ25-PYP. *Proteins: Struct., Funct., Bioinf.* **2009**, *74*, 895−904.

(7) Marsh, J. A.; Forman-Kay, J. D. Ensemble modeling of protein disordered states: Experimental restraint contributions and validation. *Proteins: Struct., Funct., Bioinf.* **2012**, *80*, 556−572.

(8) Kashtanov, S.; Borcherds, W.; Wu, H.; Daughdrill, G. W.; Ytreberg, F. M. Using Chemical Shifts to Assess Transient Secondary Structure and Generate Ensemble Structures of Intrinsically Disordered Proteins. *Methods Mol. Biol.* **2012**, *895*, 139−152.

(9) Baskaran, K.; Brunner, K.; Munte, C. E.; Kalbitzer, H. R. Mapping of protein structural ensembles by chemical shifts. *J. Biomol. NMR* **2010**, *48*, 71−83.

(10) Sahakyan, A. B.; Vranken, W. F.; Cavalli, A.; Vendruscolo, M. Structure-based prediction of methyl chemical shifts in proteins. *J. Biomol. NMR* **2011**, *50*, 331.

(11) Li, D.-W.; Brüschweiler, R. PPM: a side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles. *J. Biomol. NMR* **2012**, *54*, 257−265.

(12) Vammi, V.; Song, G. Ensembles of a small number of conformations with relative populations. *J. Biomol. NMR* **2015**, *63*, 341−351.

(13) Liu, D.; Chen, X.; Long, D. NMR-Derived Conformational Ensemble of State 1 of Activated Ras Reveals Insights into a Druggable Pocket. *J. Phys. Chem. Lett.* **2020**, *11*, 3642−3646.

(14) Camilloni, C.; Robustelli, P.; Simone, A. D.; Cavalli, A.; Vendruscolo, M. Characterization of the Conformational Equilibrium between the Two Major Substates of RNase A Using NMR Chemical Shifts. *J. Am. Chem. Soc.* **2012**, *134*, 3968−3971.

(15) Berlin, K.; Castañeda, C. A.; Schneidman-Duhovny, D.; Sali, A.; Nava-Tudela, A.; Fushman, D. Recovering a Representative Conformational Ensemble from Underdetermined Macromolecular Structural Data. *J. Am. Chem. Soc.* **2013**, *135*, 16595−16609.

(16) Brookes, D. H.; Head-Gordon, T. Experimental Inferential Structure Determination of Ensembles for Intrinsically Disordered Proteins. *J. Am. Chem. Soc.* **2016**, *138*, 4530−4538.

(17) Shrestha, U. R.; Smith, J. C.; Petridis, L. Full structural ensembles of intrinsically disordered proteins from unbiased molecular dynamics simulations. *Commun. Biol.* **2021**, *4*, 243.

(18) Lincoff, J.; Haghighatlari, M.; Krzeminski, M.; Teixeira, J. M. C.; Gomes, G.-N. W.; Gradinaru, C. C.; Forman-Kay, J. D.; Head-Gordon, T. Extended experimental inferential structure determination method in determining the structural ensembles of disordered protein states. *Commun. Chem.* **2020**, *3*, 74.

(19) Gong, H.; Zhang, S.; Wang, J.; Gong, H.; Zeng, J. Constructing Structure Ensembles of Intrinsically Disordered Proteins from Chemical Shift Data. *J. Comput. Biol.* **2016**, *23*, 300−310.

(20) He, Y.; Nagpal, S.; Sadqi, M.; de Alba, E.; Muñoz, V. Glutton: a tool for generating structural ensembles of partly disordered proteins from chemical shifts. *Bioinformatics* **2018**, *35*, 1234−1236.

(21) Ytreberg, F. M.; Borcherds, W.; Wu, H.; Daughdrill, G. W. Using chemical shifts to generate structural ensembles for intrinsically disordered proteins with converged distributions of secondary structure. *Intrinsically Disord. Proteins* **2015**, *3*, No. e984565.

(22) Rangan, R.; Bonomi, M.; Heller, G. T.; Cesari, A.; Bussi, G.; Vendruscolo, M. Determination of structural ensembles of proteins: restraining vs reweighting. *J. Chem. Theor. Comput.* **2018**, *14*, 6632−6641.

(23) Best, R. B.; Vendruscolo, M. Determination of protein structures consistent with NMR order parameters. *J. Am. Chem. Soc.* **2004**, *126*, 8090−8091.

(24) Chen, Y.; Campbell, S. L.; Dokholyan, N. V. Deciphering protein dynamics from NMR data using explicit structure sampling and selection. *Biophys. J.* **2007**, *93*, 2300−2306.

(25) Low, J. T.; Weeks, K. M. SHAPE-directed RNA secondary structure prediction. *Methods* **2010**, *52*, 150−158.

(26) Borkar, A. N.; De Simone, A.; Montalvao, R. W.; Vendruscolo, M. A method of determining RNA conformational ensembles using structure-based calculations of residual dipolar couplings. *J. Chem. Phys.* **2013**, *138*, 215103.

(27) Cavalli, A.; Camilloni, C.; Vendruscolo, M. Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *J. Chem. Phys.* **2013**, *138*, 094112.

(28) Jensen, M. R.; Salmon, L.; Nodet, G.; Blackledge, M. Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. *J. Am. Chem. Soc.* **2010**, *132*, 1270−1272.

(29) Krzeminski, M.; Marsh, J. A.; Neale, C.; Choy, W.-Y.; Forman-Kay, J. D. Characterization of disordered proteins with ENSEMBLE. *Bioinformatics* **2013**, *29*, 398−399.

(30) He, W.; Chen, Y.-L.; Pollack, L.; Kirmizialtin, S. The structural plasticity of nucleic acid duplexes revealed by WAXS and MD. *Sci. Adv.* **2021**, *7*, No. eabf6106.

(31) Bottaro, S.; Bengtsen, T.; Lindorff-Larsen, K. *Methods Mol. Biol.* **2020**, *2112*, 219−240.

(32) Bonomi, M.; Camilloni, C.; Cavalli, A.; Vendruscolo, M. Metainference: A Bayesian inference method for heterogeneous systems. *Sci. Adv.* **2016**, *2*, No. e1501177.

(33) Lu, X.-J.; Bussemaker, H. J.; Olson, W. K. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.* **2015**, *43*, No. e142.

(34) Zhang, K.; Frank, A. T. Conditional Prediction of Ribonucleic Acid Secondary Structure Using Chemical Shifts. *J. Phys. Chem. B* **2019**, *124*, 470−478.

(35) Aeschbacher, T.; Schubert, M.; Allain, F. H.-T. A procedure to validate and correct the 13 C chemical shift calibration of RNA datasets. *J. Biomol. NMR* **2012**, *52*, 179−190.

(36) Barton, S.; Heng, X.; Johnson, B. A.; Summers, M. F. Database proton NMR chemical shifts for RNA signal assignment and validation. *J. Biomol. NMR* **2013**, *55*, 33−46.

(37) Brown, J. D.; Summers, M. F.; Johnson, B. A. Prediction of hydrogen and carbon chemical shifts from RNA using database mining and support vector regression. *J. Biomol. NMR* **2015**, *63*, 39−52.

(38) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.

(39) Choy, W.-Y.; Forman-Kay, J. D. Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J. Mol. Biol.* **2001**, *308*, 1011−1032.

(40) Bottaro, S.; Bussi, G.; Kennedy, S. D.; Turner, D. H.; Lindorff-Larsen, K. Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations. *Sci. Adv.* **2018**, *4*, No. eaar8521.

(41) Cesari, A.; Gil-Ley, A.; Bussi, G. Combining simulations and solution experiments as a paradigm for RNA force field refinement. *J. Chem. Theor. Comput.* **2016**, *12*, 6192−6200.

(42) Hummer, G.; Köfinger, J. Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* **2015**, *143*, 243150.

(43) Pitera, J. W.; Chodera, J. D. On the use of experimental observations to bias simulated ensembles. *J. Chem. Theor. Comput.* **2012**, *8*, 3445−3451.

(44) Köfinger, J.; Stelzl, L. S.; Reuter, K.; Allande, C.; Reichel, K.; Hummer, G. Efficient ensemble refinement by reweighting. *J. Chem. Theor. Comput.* **2019**, *15*, 3390−3401.

(45) Parisien, M.; Major, F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **2008**, *452*, 51−55.

(46) Léger, S.; Costa, M. B. W.; Tulpan, D. Pairwise visual comparison of small RNA secondary structures with base pair probabilities. *BMC Bioinf.* **2019**, *20*, 293.

(47) Frank, A. T.; Law, S. M.; Brooks, C. L., III A simple and fast approach for predicting 1H and 13C chemical shifts: toward chemical shift-guided simulations of RNA. *J. Phys. Chem. B* **2014**, *118*, 12168−12175.

(48) Shi, H.; Rangadurai, A.; Abou Assi, H.; Roy, R.; Case, D. A.; Herschlag, D.; Yesselman, J. D.; Al-Hashimi, H. M. Rapid and accurate determination of atomistic RNA dynamic ensemble models using NMR and structure prediction. *Nat. Commun.* **2020**, *11*, 5531.

(49) Dethoff, E. A.; Petzold, K.; Chugh, J.; Casiano-Negroni, A.; Al-Hashimi, H. M. Visualizing transient low-populated structures of RNA. *Nature* **2012**, *491*, 724.

(50) Frank, A. T.; Stelzer, A. C.; Al-Hashimi, H. M.; Andricioaei, I. Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: new insights into RNA dynamics and adaptive ligand recognition. *Nucleic Acids Res.* **2009**, *37*, 3670−3679.

(51) Lu, J.; Kadakkuzha, B. M.; Zhao, L.; Fan, M.; Qi, X.; Xia, T. Dynamic ensemble view of the conformational landscape of HIV-1 TAR RNA and allosteric recognition. *Biochemistry* **2011**, *50*, 5042−5057.

(52) Chou, F.-C.; Kladwang, W.; Kappel, K.; Das, R. Blind tests of RNA nearest-neighbor energy prediction. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 8430−8435.

(53) Merriman, D. K.; Xue, Y.; Yang, S.; Kimsey, I. J.; Shakya, A.; Clay, M.; Al-Hashimi, H. M. Shortening the HIV-1 TAR RNA bulge by a single nucleotide preserves motional modes over a broad range of time scales. *Biochemistry* **2016**, *55*, 4445−4456.

(54) Dejaegere, A.; Bryce, R. A.; Case, D. A. *An empirical analysis of proton chemical shifts in nucleic acids*; ACS Publications, 1999.

(55) Cromsigt, J. A.; Hilbers, C. W.; Wijmenga, S. S. Prediction of proton chemical shifts in RNA−their use in structure refinement and validation. *J. Biomol. NMR* **2001**, *21*, 11−29.

(56) Frank, A. T.; Bae, S.-H.; Stelzer, A. C. Prediction of RNA 1H and 13C chemical shifts: a structure based approach. *J. Phys. Chem. B* **2013**, *117*, 13497−13506.

(57) Swails, J.; Zhu, T.; He, X.; Case, D. A. AFNMR: automated fragmentation quantum mechanical calculation of NMR chemical shifts for biomolecules. *J. Biomol. NMR* **2015**, *63*, 125−139.

(58) Wang, Y.; Han, G.; Jiang, X.; Yuwen, T.; Xue, Y. Chemical shift prediction of RNA imino groups: application toward characterizing RNA excited states. *Nat. Commun.* **2021**, *12*, 1595.