Identifying native-like RNA Structures Using Unassigned Chemical Shift Data: Yet Another Novel Application of the Hungarian Algorithm

William Fuh and Aaron T. Frank*

Departments of Biophysics and Chemistry, University of Michigan, 930 North University Avenue, Ann Arbor, Michigan 48109, USA

E-mail: afrankz@umich.edu

Phone: (734) 615-2053

Abstract

In this work, we demonstrate that *unassigned* NMR chemical shifts, that is, an "anonymized" list of observed chemical shift peak values, can be used to identify native-like conformations of RNAs. We achieve this by casting the problem of "assigning" unassigned chemical shift peaks to specific sites in an RNA as a linear assignment problem, and then solving it using the fast Kuhn–Munkres bipartite matching algorithm – also commonly referred to as the Hungarian algorithm. Using our assignment method, which we refer to as SCAHA (Structure-Based Chemical-Shift Assignment via the Hungarian Algorithm), we found that given an accurate structural model of an RNA, and chemical shift data free of referencing errors, unassigned chemical shift peaks can be assigned to specific sites in the RNA with an accuracy of about ∼0.2 and ∼1.0 ppm for non-exchangeable ^{1}H and ^{1}C nuclei, respectively. By comparing SCAHA-assigned chemical shifts to computed chemical shifts, we demonstrate that we can identify, with high sensitivity, native-like structures in conformational pools

that contain both native-like and non-native structures. Our results suggest that hybrid methods that combine state-of-the-art structure prediction methods with accurate structured-based assignment methods, like SCAHA, will soon enable RNA structure to be rapidly elucidated from *unassigned* NMR spectra.

Introduction

The renewed appreciation of the essential and diverse roles played by ribonucleic acids (RNAs) in the cell,[1] has led to keen interest in determining their atomistic structure. Such information is crucial to uncovering structure-function relationships that govern the regulatory properties of RNAs, and is important in the design and discovery of therapeutics that target RNAs associated with human diseases.[2]

De-novo structure prediction algorithms can, in principle, be used to elucidate the structure of RNAs directly from their sequence. In practice, though, additional experimental data is often needed to guide prediction efforts in what is now referred to as "hybrid" or "integrative" modeling.[3,4] NMR chemical shifts have recently emerged as a viable source of structural information that can be used to guide prediction of RNA structure within this hybrid modeling framework. Along these lines, Wijmenga and coworkers recently showed that the structure of RNA helices could be determined using a *de novo* structure prediction method that utilized assigned non-exchangeable ^{1}H chemical shifts as the only experimental restraints.[5] In addition, Das and coworkers recently demonstrated that assigned non-exchangeable ^{1}H chemical shift data can, in several cases, be used to predict the structure of non-canonical motifs of RNA with near atomic accuracy.[6] However, acquiring assigned chemical shift data, especially for medium-sized to large RNAs, can be both time consuming and expensive. As such, there is immense interest in developing methods that enable multi-resolution structural information to be extracted from "unassigned" chemical shift data.

In the case of proteins, several methods have been developed that utilize unassigned NMR data to extract structural information. Meiler and Baker, for example, developed a

method in which unassigned chemical shifts, intensities of NOESY cross-peaks, and residual dipolar couplings were used to guide protein fold prediction.[7] In related work, Bermejo and Llinás described a method, referred to as SC-CLOUDS, that allowed the global fold of proteins to be determined using only unassigned NOE data.[8] Most recently, Rienstra and coworkers developed the Comparative Objective Measurement of Protein Architectures by Scoring Shifts (COMPASS) approach, in which, ^{13}C-^{13}C correlation spectra are simulated from structural models and then compared to a single actual unassigned ^{13}C-^{13}C correlation spectra.[9] At the core, each of these methods rely on some a structure-based approach to optimally "assign" the unassigned NMR data. In so doing, structural models could then be identified that exhibit the best agreement between computed and optimally assigned NMR data. Promisingly, for each of these methods, the folds in structural models that exhibited the best agreement with unassigned NMR data were typically in excellent agreement with the folds in known structures, and in several cases, near atomic accuracy was achieved.[7,9]

In the case of RNAs, few methods have been developed that enable structural information to be extracted from unassigned NMR data. One notable exception is the NMR-assisted prediction of secondary structure algorithm, referred to as NAPSS, that was developed by Turner and coworkers. NAPSS enables secondary structure information to be extracted from unassigned NMR data.[10] More recently, a chemical shift-based version of NAPSS, NAPSS-CS, was developed that utilizes directional chemical shift constraints to guide secondary structure prediction, and most excitingly, enables pseudoknot RNAs to be accurately identified.[11]

Currently lacking, however, are methods that enable unassigned NMR data to be incorporated into the atomistic modeling of RNAs. As has been done for proteins, state-of-the-art structure prediction algorithms could first be used to generate putative three-dimensional (3D) models of an RNA from its sequence. And then, for each generated model, unassigned chemical shift peaks could be optimally assigned to specific sites in the RNA by minimizing the sum of the differences between the chemical shifts assigned to a given site and the

chemical shift computed for that site. The structural model or set of models that exhibit the best agreement between "assigned" and computed chemical shifts could then be identified as the predicted structure(s) of the RNA.[7–9] Clearly, such an approach rely on access to methods the can accurately predict chemical shifts from 3D model. Moreover, the success of this approach hinges on the assumption that given the availability of accurate prediction methods chemical shifts can be assign with sufficient accuracy that enables native-like and non-native conformations of an RNA to be discriminated. Inherent in this assumption is another assumption, namely, that the native-like conformations of RNA exhibit lower assignment errors than non-native conformations of the same RNA. In this study, we set out to test the hypotheses that (1) native-like conformations of an RNA exhibit the lower assignment errors than non-native conformations and (2) native-like conformations of an RNA exhibit the lower errors between the optimally "assigned" chemical shifts and computed chemical shifts than non-native conformation, the latter being a more direct test of the feasibility of utilizing unassigned chemical shift to disambiguate structural differences between native-like and non-native structures of RNAs.

To test these hypotheses, we make use of a set of 52 RNAs for which NMR structures and assigned chemical shift data are available, and for which we constructed conformational pools containing both native-like and non-native decoys. To mimic unassigned chemical data, we "anonymized" the assigned chemical shift data and then for each RNA, applied a method we refer to as SCAHA, Structure-Based Chemical-Shift Assignment via the Hungarian Algorithm, to optimally "assign" our synthetic *anonymized* unassigned chemical shift peaks to each structure in the conformational pools. Confirming our hypotheses, we found that in general, the conformations that exhibited the lowest assignment errors tended to be native-like, as were the conformations that exhibited the lowest errors between the optimally "assigned" chemical shifts and computed chemical shifts. In what follows, we first describe the theoretical underpinnings of SCAHA, assess its inherent accuracy, and finally describe in detail how we used SCAHA to test, and eventually, confirm our central

hypotheses.

Theoretical Methods

Assigning chemical shift peaks by solving the linear assignment problem

The task of "assigning" a set of n unassigned chemical shift peaks, $\Delta^{\text{actual}} \Rightarrow \{\delta_i^{\text{actual}}\}$, to specific sites in an RNA, $S \Rightarrow \{s_j\}$, can be cast as a one-to-one assignment problem in which: (a) a given peak, δ_i^{actual}, can only be assigned to one site, s_j, and (b) multiple peaks δ_i^{actual} cannot be assigned to the same s_j. Formally, this assignment problem corresponds to the linear assignment problem, the optimal solution of which is the one that minimizes the objective function, χ, which is defined as:

$$\chi = \sum_{i \in \Delta} \sum_{j \in S} C_{i,j} x_{i,j} \tag{1}$$

where $C_{i,j}$ is the "cost" associated with assigning peak δ_i^{actual} to site s_j and $x_{i,j}$ is 1 if and only if δ_i^{actual} is assigned to s_j and 0 otherwise. Clearly, to enforce a one-to-one mapping, the the optimal solution to Eq. 1 must satisfy the constraints that

$$\sum_{i \in \Delta} x_{i,j} = 1 \text{ for } i \in \Delta \tag{2}$$

and

$$\sum_{j \in S} x_{i,j} = 1 \text{ for } j \in S. \tag{3}$$

Many techniques have been developed for solving the linear assignment problem. Of note is the so-called Kuhn–Munkres algorithm (also referred to as the Hungarian algorithm), a bipartite matching algorithm which is able to solve the linear assignment problem in polynomial time.[12,13] The Hungarian algorithm takes as input the cost matrix $\mathbf{C}$, whose

elements are the C_{ij}s defined in Eq. 1, and outputs the optimal one-to-one assignments.

SCAHA: Structure-Based Chemical-Shift Assignment via the Hungarian Algorithm

As has been done for proteins,[7,9,10] unassigned chemical shift data could be incorporated into structural modeling of RNAs by optimally assigning chemical shift peaks based on a assumed structural model such that the differences between the chemical shift (δ_i^{actual}) assigned to a given site (s_j) and the computed (or predicted) chemical shift associated with that s_j ($\delta_j^{\text{computed}}$) are minimized. Within such a structure-based assignment framework, the $C_{i,j}$ in Eq. 1, which are used to construct the cost matrix ($\mathbf{C}$), can be expressed as:

$$C_{i,j} = \left| \delta_j^{\text{computed}} - \delta_i^{\text{actual}} \right| \tag{4}$$

where $\delta_j^{\text{computed}}$ is the chemical computed from the assumed structural model of the RNA. The Hungarian algorithm can then be used to solve the assignment problem described by Eqs. 1-4. Here onward, we will refer to this approach as SCAHA, Structure-based Chemical-Shift Assignment via the Hungarian Algorithm.

Computational Details

Challenge sets that contained a mixture of native-like and non-native structural models were constructed for 52 RNAs, for which both NMR structures and chemical shifts were available in the PDB[14] and either from literature or the BMRB,[15] respectively. Each of these RNAs contained non-canonical structural features (e.g., bulges, internal loops, apical loops, and three-way junctions). As such, these RNAs served as an excellent test cases for our current study. The conformational pools were generated as follows: First, for each RNA, the primary sequence was extracted from its PDB file. Next, the 9 most energetically fa-

vorable secondary structures, as predicted using MC-fold,[16] were used to generate base-pair restraints. These, along with base-pair identified in the first model of NMR bundle for a given RNA, were then used to generate structural models using the fragment assembly method, FARNA,[17] implemented in the Rosetta modeling package. For each FARNA run, 50 models were assembled and the number of cycles was set to 10000. In addition, molecular dynamics simulations were carried out starting for the first model from each NMR ensemble. Approximately, 2000 MD-derived conformers were added to the composite pool. The CHARMM simulation package[18] was used for all MD simulations, which were carried out using the GBMV implicit solvent.[19,20] For the GBMV model, we used parameters that have been shown to produce relatively stable trajectories of nucleic acids, in particular, DNA.[21] The ten set of models returned by FARNA, together with individual models of the NMR bundle for a given RNA, and the MD generated structures were combined, and then between 25-30 conformers that evenly spanned the RMSD range between 0 and 7 Å were selected. These 25-30 conformers were then used for the analysis described and discussed below. The set of conformers (referred to as the RNA-NMR decoy set) are made available via https://github.com/atfrank/RNA-NMR-Decoys.

For each of the 52 RNAs included in our challenge set, *synthetic* unassigned chemical shift data was generated from assign chemical shift data (Table 1). To "anonymized" the list of assigned chemical shifts, information about residue numbers, residues names, and nucleus types were all removed for the assigned chemical shift data files, thus resulting in a list of observed chemical shift peak values for each RNA (void of any information that would identify the sites to which they were assigned). Then for each conformer in the conformational pool of a given RNA, we then applied our SCHAHA methodology to optimally reassign the chemical shifts to specific site in the RNA. This was achieved using the Hungarian algorithm implemented in the function `solve_LSAP` in the R package clue.[22] The input to the function was a cost matrix containing elements C_{ij} (Eqs. 1 and 4), in which the δ_{actual} and δ_{computed} correspond to the "anonymized" chemical shifts and chemical

shifts computed from the conformations in the conformational pool using either the chemical shift predictors LARMOR^{D} [23] or RAMSEY,[24] respectively. As such, SCAHA was separately applied using either chemical shifts computed with LARMOR^{D} or RAMSEY.

Table 1: Challenge set RNAs used in this study.

PDB	BMRB	confomers	# $\delta^{\text{actual/computed}}$	PDB	BMRB	confomers	# $\delta^{\text{actual/computed}}$
1Z2J	6543	26	425/689	1KKA	5256	30	293/261
2KOC	5705	30	333/216	2N6S	25780	30	308/566
1YSV	6485	30	409/423	2M24	18894	30	343/451
1OW9	5852	30	436/349	1SCL	n/a	25	327/441
1NC0	5655	30	360/368	2N4L	25671	27	172/835
2LDL	17671	30	341/419	2MFD	19545	30	339/295
1LDZ	4226	30	585/456	2M22	18892	30	369/355
1LC6	5371	30	276/370	2QH4	7405	30	289/278
1PJY	5834	30	254/334	2N2P	25604	20	353/363
1UUU	n/a	30	273/295	2MEQ	18975	30	288/293
2LU0	18503	26	636/757	2M21	18891	30	341/327
2LHP	17860	30	282/573	2LQZ	18336	30	287/419
2LDT	17682	27	204/473	2K66	15859	30	252/340
2LPA	18240	30	114/233	1XHP	6320	30	425/494
4A4U	18035	30	88 /157	2QH2	7403	30	291/370
2LV0	18549	30	408/372	2N2O	25603	27	387/363
2LP9	18239	30	120/248	2M12	18838	30	290/355
2LBL	17565	30	183/263	2JWV	15538	30	480/453
2FDT	10018	30	178/560	2NCI	26024	30	273/430
4A4T	18034	30	89 /158	2MXL	25416	30	166/599
2LK3	17972	30	247/370	2M5U	19081	30	179/342
2LBJ	17563	30	198/261	2N6T	25781	27	314/662
1L1W	5321	30	165/439	2MNC	19887	30	100/248
4A4S	18036	30	89 /157	2M4W	19024	29	257/261
2LUB	18515	25	277/571	2LUN	18532	29	486/432
2LI4	17877	27	232/494	1ZC5	6633	28	649/629

[1] Listed for each RNA in the challenge set are the PDB accession code, the BMRB number, number of diverse conformers in its conformation pool, the number of actual experimental/computed chemical shift peaks.

[2] For 1UUU, chemical shift data was obtained from literature.[25]

[3] For 1SCL, chemical shift data was obtained from literature.[26]

For each RNA, after assigning the "unassigned" chemical shift data to each conformer in the associated conformational pool using SCAHA, three sets of analyses were carried out: First, to assess the accuracy of SCAHA, the assigned chemical shifts obtained by applying

SCAHA to the conformer corresponding to the representative solution structure (taken as model 1 in the NMR bundle obtained from the PDB) for a given nucleus was compared to actual experimental assignment in the reference data obtained from the BMRB or directly from literature sources. For each RNA, the mean-absolute-error (MAE) over all the non-exchangeable ^{1}H and ^{13}C nucleus was then calculated and used as a measure of the assignment accuracy. Second, to determine whether native-like conformations had the lowest assignment errors, for each RNA in our challenge set, the MAE (see above) was calculated for each conformer in its conformational pool, and then the conformer with the lowest assignment error was identified and compared with the representative solution structure by computing the all-atom RMSD between the selected ("best") structure and the representative NMR structure. And third, to determine whether native-like conformations exhibited the smallest

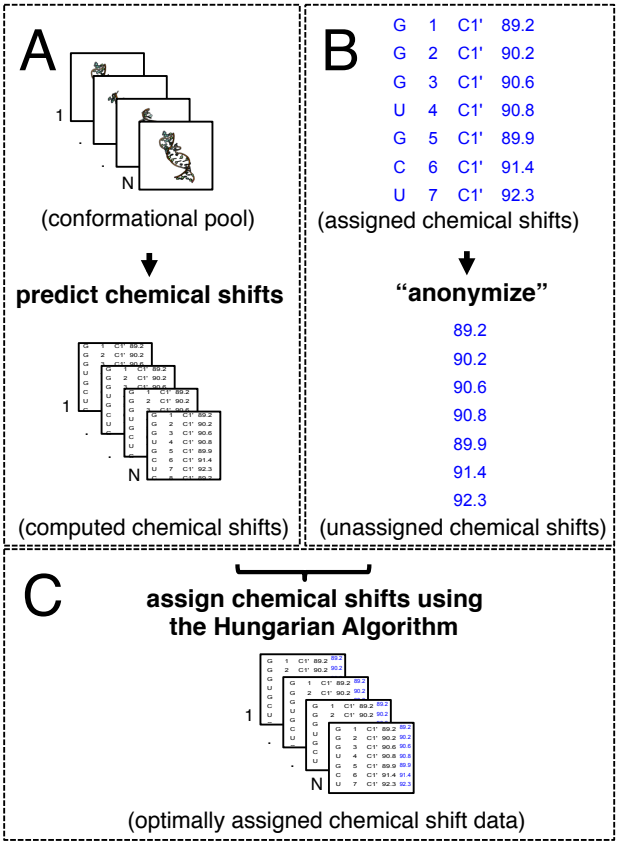


Figure 1: Application of SCAHA to RNAs in our challenge set Shown is structure-based chemical shifts assignment via the Hungarian algorithm (SCAHA) workflow that is applied to each RNA in our challenge set. For a given RNA, (A) chemical shifts are predicted from all the structures in its conformational pool and (B) a set of "synthetic " unassigned chemical shifts is generated by the anonymizing the assigned chemical shifts obtained from the BMRB and literature sources. (C) Next, the unassigned chemical shifts is optimally assigned to each set of computed chemical shifts using the Hungarian algorithm, which minimizes the sum of differences between "assigned" and computed chemical shifts.

difference between SCAHA-assigned chemical shifts and computed chemical shifts, for each RNA in our challenge set, the weighted mean absolute error (wMAE) was calculated for each conformer in its conformational pool, and then the conformer with the smallest wMAE was identified and compared with the representative solution structure by computing the all-atom structural RMSD between the selected ("best") structure and the representative NMR structure. Here the wMAE is given by:

$$w\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} w_i \left| \delta_i^{\text{assigned}} - \delta_i^{\text{computed}} \right| \tag{5}$$

where $\delta_i^{\text{assigned}}$ and $\delta_i^{\text{computed}}$ are the SCAHA-assigned and computed chemical shift associated with site i (s_i) in the RNA, respectively; N is the total number of sites corresponding to non-exchangeable ^{1}H or ^{13}C nuclei in the RNA for which "unassigned" chemical shifts were assigned to using SCAHA; and w_i is a weight factor that is equal to $1/\sigma_i$, where σ_i is the expected accuracy with which chemical shifts are computed for the nucleus type associated with s_i.[27] For comparison, the conformer that exhibited the small wMAE between experimentally assigned and computed chemical shifts was determine for each RNA and then compared to the results obtained using the SCAHA-assigned chemical shifts. *This corresponds to the limiting case in which assignments are perfect.* In all cases, chemical shifts were computed for non-exchangeable ^{1}H and ^{13}C nuclei using the empirical chemical shifts predictors LARMOR$^{\text{D}}$ and RAMSEY, and then the results obtained with each were then compared.

Because several of RNAs in our challenge set are known to contain systematic referencing errors,[28] we developed an objective structure-based approach for identify such cases. To identify these "referencing errors" for a given RNA in our challenge set, chemical shifts were computed from the representative solution of that RNA and then secondary chemical shifts, defined as the difference between computed and observed chemical shifts was fitted to linear model using Bayesian regression approach implemented in the MCMCpack package in R.[29]

For a given RNA, the linear model describing the secondary chemical shift of a given nucleus i of type k is given by:

$$\delta_{i,k}^{\text{secondary}} = \delta_{i,k}^{\text{computed}} - \delta_{i,k}^{\text{actual}} = \beta_{i,k} \tag{6}$$

where $\delta_{i,k}^{\text{computed}}$, $\delta_{i,k}^{\text{actual}}$, $\delta_{i,k}^{\text{secondary}}$, and $\beta_{i,k}$ are the chemical shift computed from the representative solution structure of that RNA, the actual chemical shift, the secondary chemical shift, and the fitting parameter (the intercept) associated with the i^{th} nucleus of type k in the RNA. Here k is one of the non-exchangeable ^{13}C nuclei in RNA, namely, C1', C2', C3', C4', C5', C2, C5, C6, C8 non-exchangeable nuclei. Within this scheme, systematic errors manifest themselves as large $\beta_k \equiv \langle \beta_{i,k} \rangle$, where $\langle \cdot \rangle$ denote an average over all instances of nucleus type k in the dataset. For a given RNA in our challenge set, a ^{13}C nucleus type is identified as having a referencing error if and only if $\beta_k > 2.00$ ppm and $\beta_k / \sigma_k > 5$ (where σ_k is the standard deviation of $\beta_{i,k}$). For all RNAs in our challenge set, ^{13}C "referencing errors" were identified using this approach and then used to correct the original chemical shift dataset before converting the data into synthetic unassigned data. Here we report results using uncorrected (raw) unassigned chemical shift and corrected unassigned chemical shift data. Separate sets of referencing errors were determined using chemical shift computed with LARMOR$^{\text{D}}$ and RAMSEY and they are reported in Table 3. The data and R code used to identify these referencing errors are available at: https://github.com/atfrank/SCAHA.

Results and discussion

We were first interested in assessing the accuracy with which a set unassigned chemical shift peaks ($\Delta^{\text{actual}} \Rightarrow \{\delta_i^{\text{actual}}\}$) could be assigned to specific sites ($S \Rightarrow \{s_j\}$) in each of the 52 RNAs we examined in the current study (which we refer to as our challenge set). For each RNA in our challenge set, SCAHA was applied using chemical shifts computed ($\Delta^{\text{computed}} \Rightarrow \{\delta_i^{\text{computed}}\}$) from a representative NMR structure (model 1 in the NMR bundle obtained from the PDB) using the chemical shift predictors LARMOR$^{\text{D}}$ and RAMSEY (two

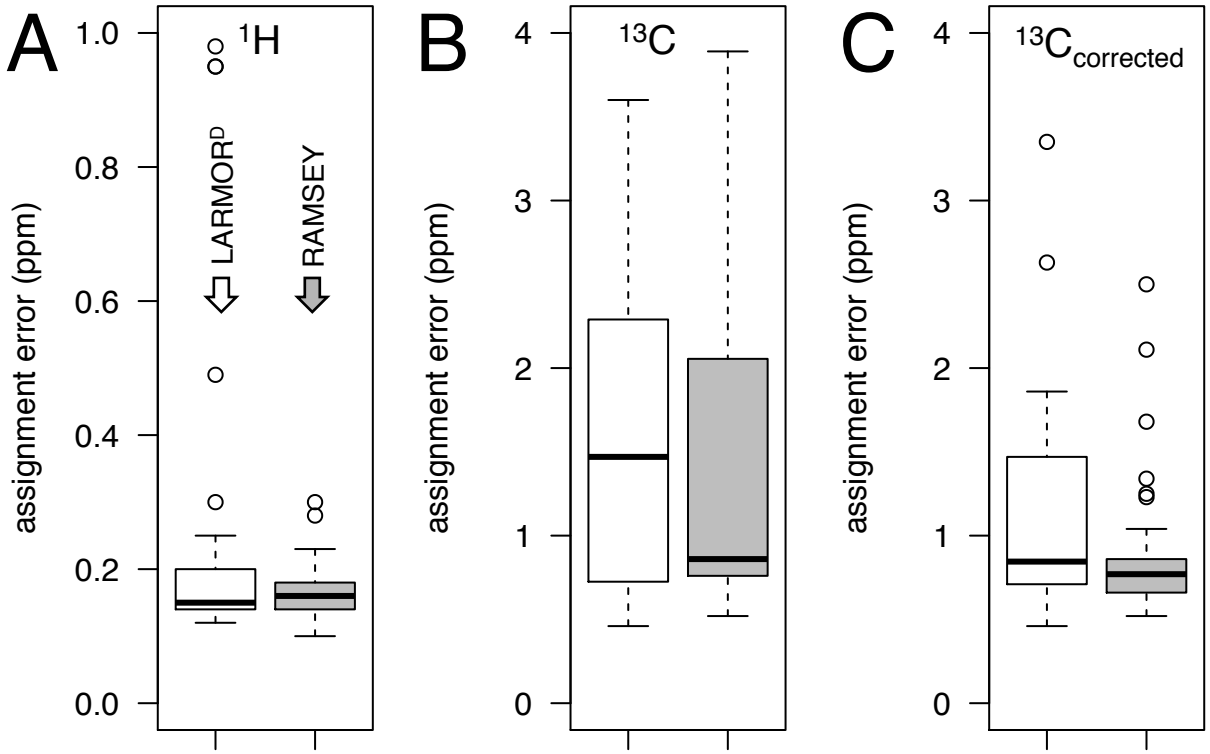


Figure 2: Assignment Errors. Shown are boxplots that summarize the distribution of assignment errors obtained for (A) ^{1}H and (B and C) ^{13}C. For ^{13}C nuclei, results are shown when (B) SCAHA was applied using "unassigned" chemical shift peaks obtained directly from the BMRB and literature source and (C) SCAHA was applied using "unassigned" chemical shift peaks that were corrected directly from the BMRB and literature source.

fast empirical chemical shift prediction methods that are capable of predicting chemical shifts for both ^{1}H and ^{13}C non-exchangeable nuclei in RNA). The mean assignment errors for non-exchangeable ^{1}H and ^{13}C nuclei were then determined.

For ^{1}H nuclei, assignment errors are on-par with the inherent errors in computed ^{1}H chemical shifts.

Figure 2A show the distributions of mean assignment errors for ^{1}H non-exchangeable nuclei in our challenge set. When using LARMOR^{D} computed chemical shifts to carry out SCAHA (LARMOR^{D}-based SCAHA), the median assignment error was 0.15 ppm and the interquartile range was 0.06 ppm. Similar results were obtained when RAMSEY computed chemical shifts were used to carry out SCAHA (RAMSEY-based SCAHA). In this case, the median assignment error and interquartile range were 0.16 and 0.04 ppm, respectively.

Because some of the RNAs included in our challenge set were also used to train LARMOR^{D} and RAMSEY, we were concerned that these initial estimates of the assignment errors (see above) may contain significant bias. Of the 52 RNAs contained in our challenge set, 17 of them were included in the training database used to develop LARMOR^{D} and 10 of them were included in the training database used to develop RAMSEY. To examine the extent to which assignment errors are affected by whether or not the RNAs in our challenge set were also used to train the predictors, we did following: for LARMOR^{D}-based SCAHA, we analyzed separately the assignment errors for the RNAs that were in the LARMOR^{D} training set and those that were not included in the LARMOR^{D} training set. Similarly, for RAMSEY-based SCAHA, we analyzed separately the assignment errors for the RNAs that were in the RAMSEY training set and those that were not included in the RAMSEY training set.

In Table 2, we compared the median assignment errors and interquartile ranges we observed for those RNAs that were included and those that were not included in the LARMOR^{D} and RAMSEY training sets, respectively. For both LARMOR^{D} and RAMSEY, we observed marginal differences between the two sets. For example, for LARMOR^{D}-based SCAHA, the median assignment error and interquartile range, were 0.14 and 0.03 ppm and 0.17 and 0.08 ppm for RNAs included and not included, in the LARMOR^{D} training set, respectively. Similar results were about using RAMSEY (Table 2). On the basis of this analysis, it appears as if the estimates of the ^{1}H assignment errors for LARMOR^{D}- and RAMSEY-based SCAHA exhibit little unbiased with regard to whether or not the RNAs in used in the accuracy analysis were included in the LARMOR^{D} and RAMSEY training database, respectively.

Interestingly, the estimated assignments errors of about $\sim$0.15 and $\sim$0.16 ppm that are associated with LARMOR^{D}- and RAMSEY-based SCAHA, compare favorably with the expected errors of $\sim$0.15[23] and $\sim$0.14 ppm[24] in chemical shifts computed with LARMOR^{D} and RAMSEY, respectively. Therefore, at least for ^{1}H nuclei, the accuracy with which SCAHA could assign the set of unassigned chemical shift peaks to specific sites in the RNAs

in our challenge set is on-par with the estimated accuracy with which ^{1}H chemical shifts are computed from 3D models of RNAs.

Table 2: Assignment Accuracy

	^{1}H		^{13}C		^{13}C$_{\text{corrected}}$	
dataset	$\Delta\delta$ (ppm)	IQR (ppm)	$\Delta\delta$ (ppm)	IQR (ppm)	$\Delta\delta$ (ppm)	IQR (ppm)
all	0.15/0.16	0.06/0.04	1.47/0.86	1.56/1.29	0.84/0.77	0.76/0.20
training	0.14/0.16	0.03/0.01	0.67/0.77	0.16/0.66	0.71/0.74	0.29/0.66
not training	0.17/0.16	0.08/0.06	2.02/0.99	1.76/1.53	1.13/0.78	0.73/0.13

[1] $\Delta\delta$ and IQR is the median assignment error and interquartile range.
[2] Separate analyses carried out using all RNAs in challenge set (all), challenge set RNAs *included* in the LARMOR^{D} and the RAMSEY training sets (training), and challenge set RNAs *not included* in the LARMOR^{D} and the RAMSEY training sets (not training).
[3] Separated by '/' are results associated with LARMOR^{D}- and RAMSEY-based SCAHA assignments.

After Accounting for Referencing Errors, ^{13}C Assignment Errors are also On-par with the inherent errors in computed ^{13}C chemical shifts

For ^{13}C nuclei, the median ^{13}C assignment error was 1.47 and 0.86 ppm when SCAHA was carried out utilizing LARMOR^{D} and RAMSEY computed chemical shifts, respectively, and the interquartile range was 1.56 and 1.29 ppm, respectively. These results indicate that, RAMSEY was able to more accurately and robustly assign chemical shift peaks for ^{13}C nuclei. These results are not too surprising given the comparatively simple distance-based model the LARMOR^{D} uses to estimate chemical shift from structure. RAMSEY, in contrast, predicts chemical shifts using more sophisticated machine-learning based algorithm (random forest) and a rich feature set that explicitly including information that includes, but not limited to, ring current effects, electrostatic bond polarization, magnetic anisotropy, hydrogen bonding and stacking.

Compared with the expected errors in computed chemical shifts, 0.81 and 0.84 ppm, for LARMOR^{D} and for RAMSEY, respectively, the assignment errors we observed, which range

between ∼1.0 to ∼1.5 ppm, are large. However, several of the ^{13}C chemical shift datasets we used in this study are known to contain referencing errors. One would expect, therefore, that the existence of such error would significantly compromise any of our attempts to assign these chemical shift peaks using SCAHA and result in larger than expected assignment errors. As such, we repeated the SCAHA assignments using "unassigned" chemical shifts that were corrected prior to being "annoymized" (Table 3).

For LARMOR^{D}-based SCAHA, correcting the chemical shifts significantly decreased the median assignment error from 1.47 to 0.89 and the interquartile range was from 1.56 and 0.90 ppm, respectively. For RAMSEY, the decrease in the median assignment error was less dramatic, it decreased marginal from 0.86 to 0.83 ppm. However, the decrease in interquartile range was very pronounced; it decreased from 1.29 to 0.31 ppm. Collectively, these results indicated that after accounting for references errors in the ^{13}C chemical shift data, the SCAHA assignment errors (of 0.8 to 0.9 ppm) is on-par with the estimated of the inherent errors (of 0.8 to 0.9 ppm) in the chemical shift predictors used in this study.

As we did for ^{1}H nuclei, we also investigated the impact on our assessment of the assignment errors. In contrast to ^{1}H, however, we did observe a significant difference in the assignment errors for those RNAs included in the training database of LARMOR^{D} and RAMSEY, and those that were not included. For example, for LARMOR^{D}, even after accounting for referencing errors, the median assignment errors for RNAs included it's training database was 0.71 ppm, compared with a value of 1.13 ppm for those RNAs in our testing set that were not included in the LARMOR^{D} training database. Additionally, the interquartile ranges were dramatically different, for RNA included in the training database the IQR value was only 0.29 ppm, compared to 0.90 ppm for RNAs not included in the training database. For RAMSEY, similar results were obtained.

Table 3: Estimated Referencing Errors

Method	PDB ID	C1'	C2'	C3'	C4'	C5'	C2	C5	C6	C8
LARMOR^{D}	1SCL	2.70	2.71	2.27	2.93	n/a	2.36	2.38	2.77	2.94
LARMOR^{D}	1UUU	n/a	n/a	n/a	n/a	n/a	2.31	n/a	n/a	n/a
LARMOR^{D}	1XHP	2.49	2.08	2.14	2.10	2.25	n/a	n/a	n/a	2.19
LARMOR^{D}	1ZC5	2.46	2.53	n/a	2.09	2.49	n/a	n/a	n/a	n/a
LARMOR^{D}	2JWV	n/a	n/a	n/a	n/a	n/a	n/a	2.01	n/a	n/a
LARMOR^{D}	2K66	2.45	n/a	n/a	n/a	n/a	2.52	2.09	2.77	2.40
LARMOR^{D}	2M21	2.57	2.83	2.53	2.98	2.53	2.90	3.83	2.74	2.73
LARMOR^{D}	2M24	6.05	5.90	n/a	n/a	n/a	5.41	5.23	4.83	5.15
LARMOR^{D}	2M4W	2.56	2.75	3.10	2.90	3.33	3.21	2.36	2.65	2.86
LARMOR^{D}	2MEQ	2.99	2.58	n/a	2.54	2.95	3.03	3.04	3.11	3.10
LARMOR^{D}	2N6S	2.34	n/a	n/a	n/a	n/a	3.72	3.00	3.01	3.02
LARMOR^{D}	2N6T	2.40	n/a	n/a	n/a	n/a	3.08	3.22	3.13	2.99
LARMOR^{D}	2QH2	2.73	2.72	n/a	2.42	2.14	n/a	2.14	2.49	2.37
LARMOR^{D}	2QH4	2.40	2.90	n/a	2.50	2.04	n/a	n/a	n/a	2.26
RAMSEY	1PJY	n/a	n/a	n/a	n/a	n/a	2.06	n/a	n/a	n/a
RAMSEY	1SCL	2.61	2.83	2.99	3.11	n/a	2.27	2.91	3.05	3.19
RAMSEY	1XHP	2.21	2.15	2.31	2.23	2.23	2.46	3.07	2.09	2.29
RAMSEY	1ZC5	2.34	2.46	2.07	n/a	2.31	n/a	n/a	n/a	n/a
RAMSEY	2JWV	n/a	n/a	n/a	n/a	n/a	n/a	2.02	n/a	n/a
RAMSEY	2K66	2.58	n/a	n/a	n/a	n/a	2.53	2.52	3.31	2.62
RAMSEY	2M21	2.77	2.91	3.07	2.90	2.85	3.16	3.00	2.92	2.84
RAMSEY	2M24	5.88	6.32	n/a	n/a	n/a	5.27	5.29	5.23	5.20
RAMSEY	2M4W	2.81	2.98	3.34	2.96	2.90	3.14	2.98	2.65	3.19
RAMSEY	2MEQ	2.80	2.78	2.81	2.73	3.06	2.41	2.63	3.32	3.24
RAMSEY	2N6S	2.52	n/a	n/a	n/a	n/a	3.94	2.72	3.06	3.13
RAMSEY	2N6T	2.22	n/a	n/a	n/a	n/a	3.33	2.85	3.04	3.23
RAMSEY	2QH2	2.44	2.97	2.38	2.76	2.34	2.43	2.63	2.75	2.85
RAMSEY	2QH4	2.51	3.34	2.02	2.81	2.56	2.47	2.21	3.25	2.91

[1] Listed are RNAs in our challenge set whose chemical shifts, obtained from either the BMRB or literature sources, are predicted to contain referencing errors (Methods).

SCAHA Assignments are Rarely "Perfect"

As a further gauge of the overall performance of SCAHA, for each RNA in our challenge set, we determined the fraction of the assignments that were perfect (f_{perfect}), that is, the fraction SCAHA assignments that matched the actual experimental assignments. In general, we found that the f_{perfect} values were very small. For example, for LARMOR^{D} and RAMSEY-

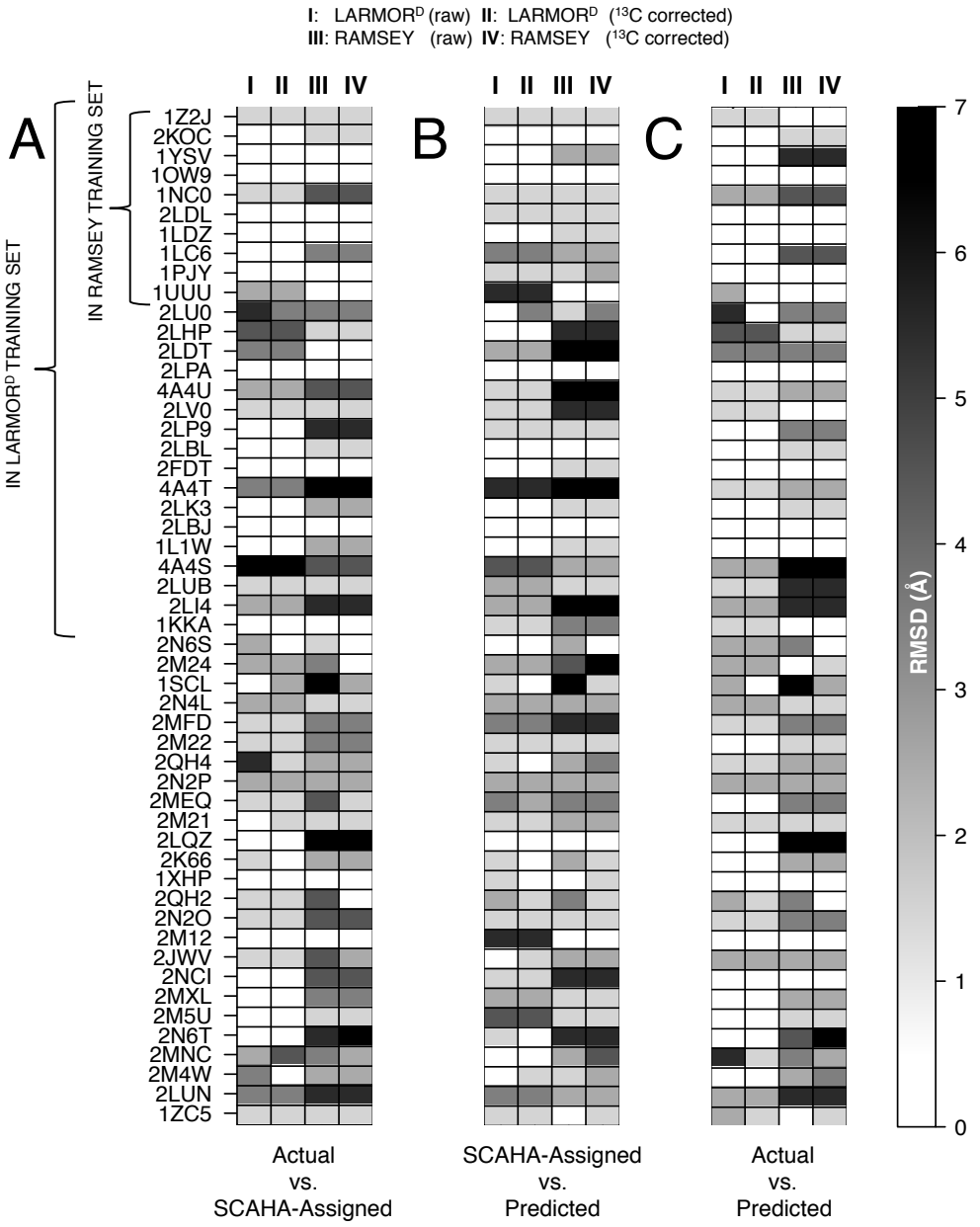


Figure 3: Comparison between "best" (i.e, low-error) structures and representative solution NMR structures of RNAs in our challenge set. Shown are levelplots of the RMSD between the representative solution NMR structure of each RNA in our challenge set and the conformer in the conformational exhibiting (A) the lowest mean absolute error (MAE) between actal experimentially assigned chemical shifts and the SCAHA-assigned chemical shifts, (B) the lowest weighted mean absolute error (wMAE; Eq. 6) between SCAHA-assigned chemical shifts and computed chemical shifts, and (C) the lowest$w$MAE between actual experimentally-assigned chemical shifts and computed chemical shifts. For each levelplot, results are shown when using LARMOR$^{\mathrm{D}}$ computed chemical shifts and (column **I**) raw uncorrected chemical shifts data and (column **II**) corrected chemical shifts data (Methods), and RAMSEY computed chemical shifts and (column **III**) raw uncorrected chemical shifts data and (column **IV**) corrected chemical shifts data, respectively. Indicated (upper left) are the RNAs in our challenge set that were included in the datasets used to train LARMOR$^{\mathrm{D}}$ and RAMSEY, respectively. Note that only ^{13}C chemical shift re-referenced (where appropriate; see Methods)

based SCAHA, the median f_{perfect} (calculated over all the RNAs in our challenge set) were only 0.066 (6.6 %) and 0.057 (5.7 %), respectively (Table 4). Similar results were also

Table 4: Mean RMSD between the "best" (i.e, low-error) structure and the representative solution NMR structure for RNAs in our challenge set.

dataset	f_{perfect}	$f_{<\text{error}}$
all	0.066/0.057	0.596/0.528
training	0.078/0.075	0.639/0.536
not training	0.059/0.057	0.548/0.521

[1] f_{perfect}: median fraction of perfect assignments
[2] $f_{<\text{error}}$: median fraction of assignments less than the associated prediction error
[3] Separated by '/' are results associated with LARMOR$^{\text{D}}$- and RAMSEY-based SCAHA assignments.

obtained for the RNAs the were included and those not included in the LARMOR$^{\text{D}}$ and RAMSEY training sets, respectively (Table 4). Similarly, we also computed the fractions of assignments that were less than the expected prediction errors ($f_{<\text{error}}$). For LARMOR$^{\text{D}}$ and RAMSEY-based SCAHA, the median $f_{<\text{error}}$ (calculated over all the RNAs in our challenge set) were 0.596 (59.6 %) and 0.528 (52.8 %), respectively (Table 4). For the RNAs the were included and those not included in the LARMOR$^{\text{D}}$ and RAMSEY training sets, respectively, the corresponding values were 0.639 (63.9 %) and 0.536 (53.6 %) and 0.548 (54.8 %) and 0.521 (52.1 %), respectively (Table 4).

Though SCAHA Assignments are Rarely "Perfect", native-like Structures Appear to Exhibit the Lowest Assignment Errors

Though the fraction of perfect assignment made by SCAHA are typically small, the fact that ∼50 % of the assignments were within the expected errors in the computed chemical shifts suggest that it might still be possible to use structure-based assigned chemical shifts to guide structural modeling of RNA. Indeed, previous work on proteins have demonstrated that even when there are assignment discrepancies between the structure-based assignments and the actual, experimental assignments,[7–9] unassigned NMR data could still be used acquire structural information about proteins by identifying structure or set of structures that

exhibited the smallest difference between optimally assigned and computed chemical shifts.

Accordingly, given the current accuracy with which SCAHA could be used to assign unassigned chemical shift, we then tested the first of the hypotheses described in the introduction, namely, that native-like conformations of an RNA exhibit the lowest assignment errors. To test this hypothesis, for each RNA in our challenge set we optimally assigned the unassigned chemical shifts using SCAHA to each of the conformations associated with that RNA. For each nucleus in a given RNA for which chemical shift data was available, we compared the SCAHA-assigned chemical shift to the actual, experimentally assigned chemical shift, computed the mean absolute assignment error for each conformation, and then identified the "best" conformation as the one exhibiting the lowest assignment error. If our hypothesis is correct, then we would expect that for the majority of RNAs in our challenge set, the "best" conformation is native-like. Indeed, this is what we observed (Figure 3; Table 6). For LARMOR$^{\mathrm{D}}$-based SCAHA, the mean RMSD for the "best" conformation was 1.53 and 1.53 Å prior to and after correcting for ^{13}C referencing errors, respectively. For RAMSEY-based SCAHA, the corresponding values were 2.66 and 2.34 Å, respectively.

Given that some of the RNAs in our challenge set were included in the training sets used to develop LARMOR$^{\mathrm{D}}$ and RAMSEY, we carried out separate analyses for RNAs in our challenge set that were included in the respective training sets and those that were not included in those training sets. For RAMSEY-based SCAHA, when comparing the results for RNAs that were included in the RAMSEY training set and those that were not included, we observed an increase in the mean RMSD of the "best" conformations: using uncorrected chemical shift data, the mean RMSD of the "best" conformations was 1.41 and 2.96 Å, respectively.

Interestingly, for LARMOR$^{\mathrm{D}}$-based SCAHA, we did not observed such a large discrepancy when comparing the results for RNAs that were included in the training the LARMOR$^{\mathrm{D}}$ and those that were not included: using uncorrected chemical shift data, the mean RMSD of the "best" conformations were both 1.53 and 1.53 Å, respectively. Taken together, these results

indicate that in general, and especially for LARMOR^{D}-based SCAHA, the conformation of an RNA that exhibit the lowest assignment error do tend to be native-like. Here we note that the training set used to train RAMSEY predictors only contained 19 RNAs, as compared to 36 RNAs for LARMOR^{D}. As such, this difference probably explains why LARMOR^{D}-based SCAHA appeared to be more robust than RAMSEY-based SCAHA performance of LARMOR^{D} and RAMSEY: LARMOR^{D} predictors were trained on a more large and diverse set of RNAs and so is more sensitive to structural differences in these RNAs. (Counter-intuitively, it was RAMSEY-based SCAHA that was able to more accurately assigned the "unassigned" chemical shift data (Figure 2).)

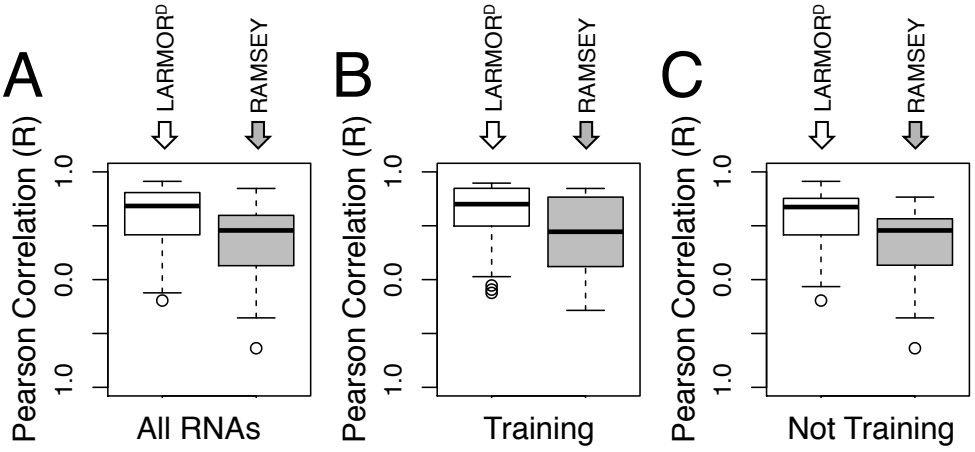


Figure 4: Assessing the ability to identifying native-like conformations XXX

The use of structured-based assignment approaches to first assign unassigned chemical shift peaks and then identify native-like conformations of RNA hinge on the assumption that the non-nativeness (e.g., as measured by the RMSD) is positively correlated with assignment error. In other words, it hinges on the assumption that the more non-native the conformation the larger the assignment error. To investigate more closely this relationship between assignment error and structure, for each RNA in our challenge set, we determined the extent to which assignment error was correlated to structural (dis)similarity relative to the representative solution structure.

Shown in Figure 4 is are distributions of the Pearson correlation coefficient (R) between

the assignment error and the structural RMSD. Overall, we observed a positive correlation between the assignment errors associated with both LARMOR$^{\mathrm{D}}$- and RAMSEY-based SCAHA exhibited a positive correlation with the "non-nativeness", and these results did not depend on whether or not the RNAs were included in the LARMOR$^{\mathrm{D}}$ and RAMSEY training sets: over all the RNAs in our challenge set, the median R between the RMSD and LARMOR$^{\mathrm{D}}$- and RAMSEY assignment errors were 0.68 and 0.46, respectively; over the RNAs in the included in the LARMOR$^{\mathrm{D}}$ and RAMSEY training sets, respectively, these values 0.70 and 0.44; and over the RNAs in the not included in the LARMOR$^{\mathrm{D}}$ and RAMSEY training sets, these values 0.67 and 0.46. Though both the assignment errors associated with both LARMOR$^{\mathrm{D}}$- and RAMSEY-based SCAHA exhibited a positive correlation with the "non-nativeness", the correlations is strongest for LARMOR$^{\mathrm{D}}$, which is consistent with our observations that the assignment errors associated LARMOR$^{\mathrm{D}}$-based SCAHA were better able to identify native-like conformations of RNAs.

Taken together, the results presented above show that though the fraction of perfect assignments that achieved with SCAHA are low, the conformations of each RNA in our challenge set that exhibit the lowest assignment errors do indeed tend to be native-like.

native-like Structures Also Exhibit the Lowest Chemical Errors Between SCAHA-Assigned and Computed Chemical Shifts.

Encouraged by the fact that native-like structures exhibit the lowest assignment errors, and that there was a positive correlation between assignment error and "non-nativeness", we next addressed the second hypothesis laid out in the introduction, namely, that native-like conformations of an RNA exhibit the small differences (errors) between "assigned" and computed chemical shifts. Within a structured-based assignment framework, in which unassigned data is assign to specific sites in an RNA by assuming a structural model of an RNA and then optimally assigning the unassigned data based on chemical shifts computed from that structure, the more direct test of the feasibility of utilizing unassigned chemical shift to identify

native like structures is whether the conformations of an RNA that exhibit the smallest difference between the optimally "assigned" and computed chemical shifts are native-like. If, in general, the conformation(s) that exhibits the small difference between "assigned" and computed chemical shifts are indeed native-like, then it would suggest that unassigned chemical shift data could be used to together with RNA structure prediction approaches to identify native-like conformations, thus enabling useful structural information about an RNA to be acquired from unassigned chemical shift data.

Table 5: Mean RMSD between the "best" (i.e, low-error) structure and the representative solution NMR structure for RNAs in our challenge set.

	Actual vs. Assigned		Assigned vs. Predicted		Actual vs. Predicted	
dataset	LARMOR^{D}	RAMSEY	LARMOR^{D}	RAMSEY	LARMOR^{D}	RAMSEY
all	1.53/1.43	2.66/2.34	1.73/1.67	2.54/2.60	1.38/1.07	2.35/2.22
training	1.53/1.47	1.41/1.41	1.38/1.53	1.47/1.52	1.30/1.09	1.69/1.69
not training	1.53/1.40	2.96/2.57	2.06/1.79	2.80/2.86	1.46/1.06	2.51/2.35

[1] RMSD are reported in Å.

[2] Separated by '/' are results associated with raw uncorrected chemical shift data and corrected chemical shift. Note that only ^{13}C chemical shift re-referenced (where appropriate; see Methods); ^{1}H chemical shift used as reported in the BMRB or literature sources.

We therefore repeated the analysis described above, but instead of calculating the differences between SCAHA-assigned chemical shifts and the experimentally assigned chemical shifts, we calculated the differences between SCAHA-assigned chemical shifts and LARMOR^{D} and RAMSEY computed chemical shifts. The results we obtained mirrored closely the results that was obtained when comparing SCAHA-assigned chemical shifts to experimentally assigned chemical shifts. For example, over the entire set of RNAs in our challenge set, for LARMOR^{D}- and RAMSEY-based SCAHA, the mean RMSD of the "best" conformations (i.e., those that exhibit the smallest differences) were 1.73 and and 2.54 Å, respectively when using the uncorrected ^{13}C data and 1.67 and 2.60 Å respectively when using corrected ^{13}C data. For RNAs included in the LARMOR^{D} and RAMSEY training, these values were 1.38 and and 1.47 Å, respectively when using the uncorrected chemical shift data and 1.53 and 1.52 Å respectively when using the corrected chemical shift data. For RNAs not included in

the training sets, the corresponding values were 2.06 and and 2.80 Å, respectively and 1.79 and 2.86 Å respectively.

Comparing computed chemical shifts to the actual, experimental assigned chemical shifts, enabled us to assess the ability to identify native-like in the limit of "perfect" assignment (or when fully assigned chemical shift is available). For RNAs not included training sets, the mean RMSD of the "best" conformations based on LARMOR^{D}- and RAMSEY-based SCAHA were 1.46 and and 2.51 Å, respectively when using the uncorrected chemical data, and 1.06 and 2.35 Å, respectively when using the corrected chemical data, compared with the values of 2.06 and and 2.80 Å, respectively and 1.79 and 2.86 Å respectively that obtained using SCAHA-based assignments.

Focusing on the results obtained for RNAs not included in the training sets and utilizing data free of referencing, the results we obtain indicate that: (i) Starting from completely unassigned chemical shift data, SCAHA-based assignments could be used to identify native-like structures of RNA in our challenge set to within ∼2.0 and 3.0 Å when using LARMOR^{D} and RAMSEY computed chemical shifts, respectively; and (ii) In the limit of perfect assignment (or when fully assigned chemical shift data is available) native-like structures could be identified to within ∼1.0 and 2.5 Å when using LARMOR^{D} and RAMSEY computed chemical shifts, respectively.

Table 6: Mean RMSD between five "best" (i.e, five low-error) structures and representative solution NMR structures of RNAs in our challenge set.

	Actual vs. Assigned		Assigned vs. Predicted		Actual vs. Predicted	
dataset	LARMOR^{D}	RAMSEY	LARMOR^{D}	RAMSEY	LARMOR^{D}	RAMSEY
all	1.76/1.70	2.77/2.56	1.99/1.99	2.93/2.95	1.51/1.52	2.57/2.43
training	1.59/1.57	2.25/2.25	1.76/1.80	2.76/2.77	1.50/1.50	1.93/1.93
not training	1.92/1.82	2.89/2.64	2.19/2.17	2.97/2.99	1.52/1.53	2.73/2.55

[1] RMSD are reported in Å.

[2] Separated by '/' are results associated with raw uncorrected chemical shift data and corrected chemical shift. Note that only ^{13}C chemical shift re-referenced (where appropriate; see Methods); ^{1}H chemical shift used as reported in the BMRB or literature sources.

Collectively, the results we presented above suggest that unassigned chemical shift data

may find utility in guiding structural modeling of RNAs within a framework in which structure prediction methods are first used to sample or generate putative conformations of an RNA, then a structure-based assignment method is used to optimally assign the unassigned chemical shift data to each conformation, and finally the conformation(s) that exhibit the best agreement between the optimally "assigned" chemical shifts and computed chemical shifts is/are identified. Similar to how assigned chemical shift data as been used to guide modeling,[6] within this framework, the unassigned chemical shift data is used to disambiguate structural differences between the native-like vs non-native conformations of RNAs. The results presented thus far seem to suggest that the errors between optimally "assigned" chemical shifts and computed chemical shifts can – for the majority of RNAs in our challenge set – accurately resolve these structural differences. To more directly probe this, we computed the normalized sum of logarithmic ranks (NSLR),[30] a popular performance metric which quantities the ability of some "measure" (here, the error between SCAHA-assigned chemical shifts and computed chemical shifts) to separate or resolve groups within a dataset (here the native-like and non-native conformers in the conformational pools of the RNAs). The NSLR ranges between 0 and 1, with 1 corresponding to complete separation (or, perfect "resolvability") of the two groups and is well suited for quantifying the resolving power of unassigned chemical shift data over the RNAs in our challenge set.[27] Figure 5 summarizes the results of this analysis.

In contrast to the theoretical NSLR that one would expect if the conformers in each conformational pool was randomly prior to its calculation (NSLR $\sim$ 0.50), we found that the NSLR obtained using the errors between computed chemical shifts and the chemical shifts assigned using LARMOR$^{\mathrm{D}}$ were large. For example, over the entire set of RNAs in our challenge set, the median NSLR associated with LARMOR$^{\mathrm{D}}$-based SCAHA assignments were 0.8. Similar results were obtained regardless of whether or not the RNAs that were analyzed were included the respective LARMOR$^{\mathrm{D}}$ (Figure 5B and 5C). These results indicating that for most of the RNAs, the errors between the computed chemical shifts and chemical shifts

optimally assigned using LARMOR^{D}-based SCAHA were effective at resolving structural differences between the native and non-native conformation of the RNAs in our challenge set. In contrast, the median NSLR associated with RAMSEY-based SCAHA was consistently near $\sim$0.5 (Figure 5A, 5B and 5C), indicating that only about half the RNAs median NSLR in excess of what one would expect for NSLR associated random sorting of the conformation. Similarly, the NSLR obtained using the conformational energies obtained using the Rosetta-RNA all-atom energy function[31] tended to be $\sim$0.5.

Interestingly, if experimentally assigned chemical shifts are used instead of the SCAHA assigned chemical shifts and compared with computed chemical shifts, the NSLR we observed were relatively large (Figure 5D-F), indicative of strong resolving power. For example, for the RNAs not included in the LARMOR^{D} and RAMSEY training sets respectively, the median NSLR values were 0.90 and 0.70, respectively (Figure 5F). This is to be compared with values of 0.75 and 0.50 that are obtained using chemical shifts assigned with LARMOR^{D}- and RAMSEY-based SCAHA, respectively (Figure 5C). Though not the central focus of this study, these results do serve to highlight potential of assigned NMR chemical shifts data in disambiguating native-like from non-native structures. In addition do serves as an indication of the limiting resolving power one might expect as more accurate structured-assignment strategies are developed.

In summary, by comparing computed chemical shifts to SCAHA-assigned chemical shifts, and then identifying the structure that exhibits the lowest difference between them, the native-like conformation could be identified for most of the RNA in our challenge set. Moreover, as evidence by larger NSLR, LARMOR^{D} associated chemical shifts errors exhibited a superior ability to separate (or resolve) the set of all native-like conformations with the decoy pool of each RNA from the set of non-native decoys than RAMSEY-associated errors (Figure 5). More importantly, LARMOR^{D} chemical shifts errors also exhibited superior ability to separate (or resolve) native-like conformations from non-native decoys than the Rosetta RNA all-atom energy function (Figure 5), a clear demonstration of the potential value that

can be added by incorporating unassigned chemical shifts into the structural modeling of RNAs through the use of structured-based assignment approaches like SCAHA.

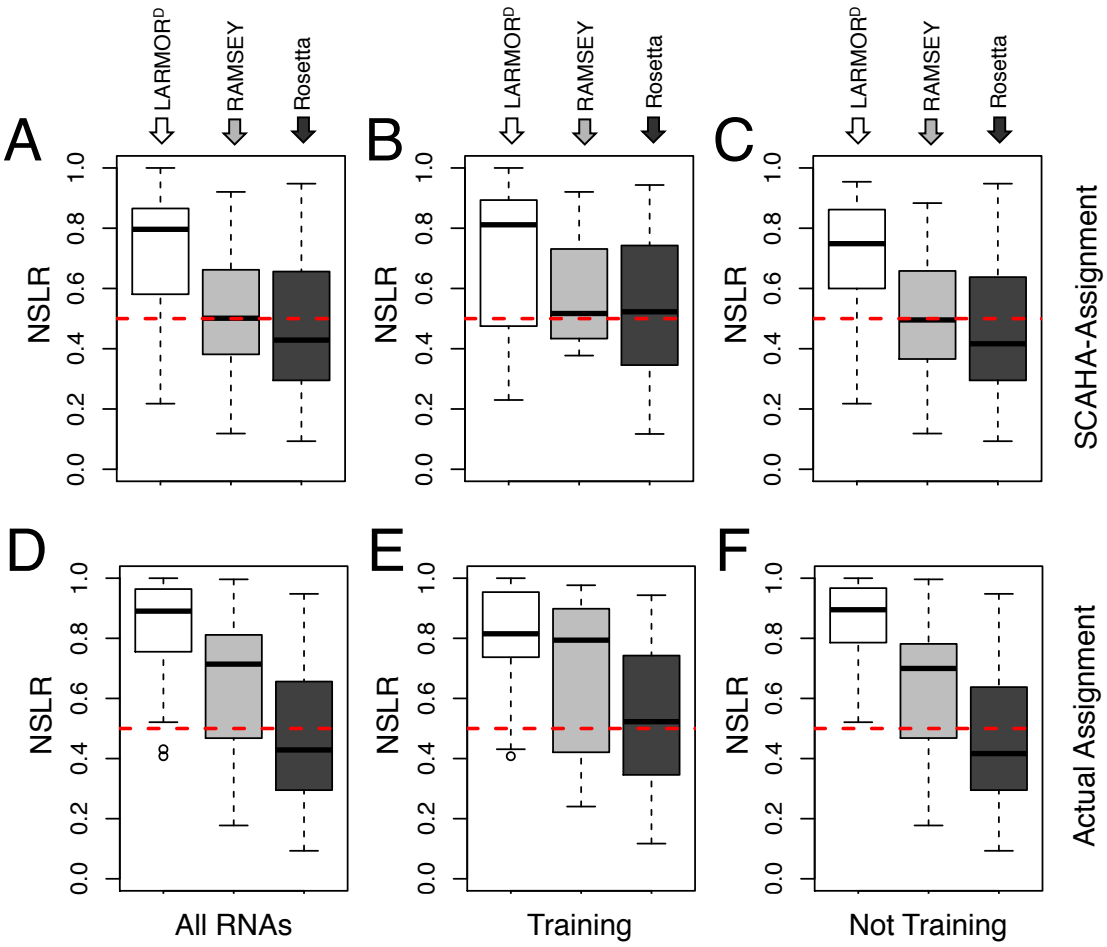


Figure 5: Resolving native-like from non-native using unassigned chemical shift data. Distributions of the normalized-sum-of-logarithmic ranks (NSLR) for (A, D) all the RNAs in our challenge set, (B, E) RNAs in our challenge set that were *included* in the LARMOR^{D} and RAMSEY training set, and (C, F) RNAs in our challenge set that were *not included* in the LARMOR^{D} and RAMSEY training set. Shown in each boxplot are the NSLRs that were obtained based on the weighted mean-absolute-error (wMAE; Eq. 6) between (A-C) computed chemical shifts and chemical shifts assigned using LARMOR^{D}- and RAMSEY-based SCAHA, respectively and (D-F) computed chemical shifts and actual experimentally assigned chemical shifts. Also reported are the NSLR distribution obtained by using the energies obtained by minimizing and scoring conformations using the Rosetta all-atom energy function.[31] For reference, the NSLR that one would expect to observe if conformer are randomly sorted is indicated by the red dashed line.

Discussion

Motivated by our interest in utilizing unassigned chemical shift data to aid in the structural modeling of RNAs, in the current study we focused on addressing two related hypotheses, namely, that (1) "native-like conformations of an RNA exhibit the lowest assignment errors

and (2) native-like conformations of an RNA exhibit the small differences errors between optimally "assigned" chemical shifts and computed chemical shifts. To test these hypotheses, we generated "synthetic" unassigned chemical shift data for 52 RNA by "anonymizing" assigned chemical shift data that we obtained from either the BMRB or literature sources. Our results, which confirm both of these hypotheses, are encouraging and bode well for the potential use of unassigned chemical shift data in modeling the 3D structure of RNAs. In particular, unassigned NMR data could be used to guide the modeling of RNAs by combining state-of-the-art structure prediction methods with SCAHA in a hybrid scheme in which a diverse set of RNA structures are generated using these structure prediction methods, and then SCAHA, taking as input unassigned chemical shift data, could be used to identify the model(s) that exhibit the smallest difference between the optimally assigned chemical shifts and the computed chemical shifts. Our results using synthetic unassigned chemical shift data suggest that, in general, these models tend to be native-like.

Most empirical structure-based chemical shift method can confidently predict chemical shifts for non-exchangeable ^{1}H or ^{13}C nuclei.[23,24,32,33] Though LARMOR$^{\mathrm{D}}$ can also, predict chemical shift imino and amino nuclei, the prediction errors for these nuclei are relatively large ($\sim$0.40 and $\sim$1.32 ppm, respectively)[23] when compared to the expected errors in non-exchangeable ^{1}H and ^{13}C predictions ($\sim$0.15 and $\sim$0.81 ppm, respectively).[23] As such, using empirical predictors, SCAHA that utilize chemical shifts computed with these methods can only confidently be used to assign chemical shifts for non-exchangeable ^{1}H and ^{13}C nuclei, even though, in a practical setting, unassigned peaks list will contain chemical shifts peaks for other ^{1}H, ^{13}C, and ^{15}N nuclei as well, depending the type of NMR experiment carried out. As an alternative, chemical shifts for all NMR-active nuclei in RNAs can be computed using quantum mechanics, but even for small RNAs (<40-nt), such calculations can be computationally demanding. Recently described fragmentation schemes, however, go a long way towards making these quantum mechanical calculation more feasible.[34] The expense of these quantum mechanically calculation notwithstanding, SCAHA can be used in conjunction with

quantum mechanically computed chemical shifts, thus enabling, in principle, all the observed ^{1}H, ^{13}C, and ^{15}N peaks to be assigned.

In the current study, we did not attempt to optimize SCAHA, instead, in the current version of SCAHA we made use the Hungarian algorithm already implemented in the clue package in R.[22] In our hands, the mean SCAHA-runtime (for a single conformer) over our challenge set of 52 RNAs was 1.361 s, and the SCAHA-runtimes ranged between ∼0.133 to ∼25.96 s for RNAs that had cost matrices (Eq. 1) that contained 88 × 157 and 172 × 835 elements, respectively. In the context of modeling large RNAs, for which the use of unassigned data might be most useful, and for which the cost matrices that are inputted into the Hungarian algorithm are larger, the longer runtimes will possibly limit the number of conformations that can processed using SCAHA. Fortunately, a graphical processing unit (GPU)-accelerated implementation of the Hungarian algorithm was recently developed[35] and so future work will center on implementing a fast GPU-accelerated version of SCAHA. To facilitate use and further testing of the current version of SCAHA, we have released the source code under a GNU license and make it freely available to the academic community https://github.com/atfrank/SCAHA.

We note that though our development of SCAHA was motivated by our interest in utilizing unassigned chemical shift data to guide modeling of RNA structure SCAHA is completely general and so can be used to assign chemical shift peaks using chemical shifts computed from structural models of *any molecule*; SCAHA takes as input a simple text file containing a list of observed chemical shift peak value (which are to be assigned to specific sites in a given molecule) and a file containing chemical shifts computed from a structural model or set of models of that molecule. Beyond RNA, SCAHA can therefore also be used to assign chemical shift peaks in other biomolecules, for example, proteins, and so can be used to guiding structural modeling of those molecules.

Conclusion

Recent studies have demonstrate that guiding modeling using assigned chemical shift data can significantly enhance the accuracy of RNA structure prediction. In this study, we attempted to assess the feasibility of utilizing *unassigned chemical shift data* to guide the atomistic modeling of RNAs. As a first step, we developed a structure-based assignment method, which we refer to as SCAHA, that takes as input a list of unassigned chemical shift peaks and computed chemical shifts for specific sites within an RNA, and then outputs the optimally assigned chemical shifts. Using SCAHA, we then tested two hypotheses that are crucial to assessing the feasibility of utilizing unassigned chemical shift data to guide the structural modeling of RNAs. In particular, we tested the hypotheses that (1) native-like conformations of an RNA exhibit the lower errors between the optimally assigned chemical shifts and actual experimentally assigned chemical shifts than non-native conformations and (2) that native-like conformations of an RNA exhibit the smaller errors between the optimally chemical shifts and computed chemical shifts than non-native conformations. Confirmation of these hypotheses would suggest that unassigned chemical shift data could be used to identified native-like conformations within sets that contained both of native-like and non-native conformations.

By applying SCAHA to a challenge set containing 52 RNAs, for each of which, we constructed conformational pools containing both native-like and non-native decoys, we were able to directly test these two hypotheses. For most of the RNAs in our challenge, we found that the conformations that exhibited the lowest assignment error tended to be native-like, and as are the conformations that exhibited the smallest difference between optimally assigned chemical shifts and computed chemical shifts. Therefore, by utilizing structured-based assignment methods, like SCAHA, native-like conformations of RNAs could be identified from a set of putative structural models using unassigned chemical shift data by identifying the model or set of models that exhibit the lowest error between optimally "assigned" chemical shifts and chemical shifts computed from the models. These results should pave the way

for developing workflows that make use of unassigned chemical shift data in RNA structure prediction and determination.

References

(1) Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489*, 57–74.

(2) Cooper, T. A.; Wan, L.; Dreyfuss, G. RNA and Disease. *Cell* **2009**, *136*, 777–793.

(3) Burke, J. E.; Sashital, D. G.; Zuo, X.; Wang, Y.-X.; Butcher, S. E. Structure of the Yeast U2/U6 SnRNA Complex. *RNA* **2012**, *18*, 673–683.

(4) Lee, W.; Cornilescu, G.; Dashti, H.; Eghbalnia, H. R.; Tonelli, M.; Westler, W. M.; Butcher, S. E.; Henzler-Wildman, K. A.; Markley, J. L. Integrative NMR for Biomolecular Research. *Journal of Biomolecular NMR* **2016**, *64*, 307–332.

(5) van der Werf, R. M.; Tessari, M.; Wijmenga, S. S. Nucleic Acid Helix Structure Determination From NMR Proton Chemical Shifts. *Journal of Biomolecular NMR* **2013**, *56*, 95–112.

(6) Sripakdeevong, P.; Cevec, M.; Chang, A. T.; Erat, M. C.; Ziegeler, M.; Zhao, Q.; Fox, G. E.; Gao, X.; Kennedy, S. D.; Kierzek, R.; Nikonowicz, E. P.; Schwalbe, H.; Sigel, R. K. O.; Turner, D. H.; Das, R. Structure Determination of Noncanonical RNA Motifs Guided by 1H NMR Chemical Shifts. *Nature Methods* **2014**, *11*, 413–416.

(7) Meiler, J.; Baker, D. Rapid Protein Fold Determination Using Unassigned NMR Data. *Proceedings of the National Academy of Sciences* **2003**, *100*, 15404–15409.

(8) Bermejo, G. A.; Llinás, M. Deuterated Protein Folds Obtained Directly from Unassigned NOE Data. *Journal of the American Chemical Society* **2008**, *130*, 3797.

(9) others,, et al. Experimental Protein Structure Verification by Scoring with a Single, Unassigned NMR Spectrum. *Structure* **2015**, *23*, 1958–1966.

(10) Hart, J. M.; Kennedy, S. D.; Mathews, D. H.; Turner, D. H. NMR-Assisted Prediction of RNA Secondary Structure: Identification of a Probable Pseudoknot in the Coding Region of an R2 Retrotransposon. *Journal of the American Chemical Society* **2008**, *130*, 10233–10239.

(11) Chen, J. L.; Bellaousov, S.; Tubbs, J. D.; Kennedy, S. D.; Lopez, M. J.; Mathews, D. H.; Turner, D. H. Nuclear Magnetic Resonance-Assisted Prediction of Secondary Structure for RNA: Incorporation of Direction-Dependent Chemical Shift Constraints. *Biochemistry* **2015**, *54*, 6769.

(12) Kuhn, H. W. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly* **1955**, *2*, 83–97.

(13) Munkres, J. Algorithms for the Assignment and Transportation Problems. *Journal of the Society for Industrial and Applied Mathematics* **1957**, *5*, 32–38.

(14) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank. *European Journal of Biochemistry* **1977**, *80*, 319–324.

(15) Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z. BioMagResBank. *Nucleic Acids Research* **2008**, *36*, D402–D408.

(16) Parisien, M.; Major, F. The MC-Fold and MC-Sym Pipeline Infers RNA Structure from Sequence Data. *Nature* **2008**, *452*, 51–55.

(17) Das, R.; Karanicolas, J.; Baker, D. Atomic Accuracy in Predicting and Designing Non-canonical RNA Structure. *Nature Methods* **2010**, *7*, 291–294.

(18) Brooks, B.; Bruccoleri, R.; Olafson, B.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry* **1983**, *4*, 187–217.

(19) Lee, M. S.; Salsbury Jr, F. R.; Brooks III, C. L. Novel Generalized Born Methods. *The Journal of Chemical Physics* **2002**, *116*, 10606–10614.

(20) Lee, M. S.; Feig, M.; Salsbury, F. R.; Brooks, C. L. New Analytic Approximation to the Standard Molecular Volume Definition and Its Application to Generalized Born Calculations. *Journal of Computational Chemistry* **2003**, *24*, 1348–1356.

(21) Chocholoušová, J.; Feig, M. Implicit Solvent Simulations of DNA and DNA- Protein Complexes: Agreement with Explicit Solvent vs Experiment. *The Journal of Physical Chemistry B* **2006**, *110*, 17240–17251.

(22) Hornik, K. A CLUE for CLUster Ensembles. *Journal of Statistical Software* **2005**, *14*.

(23) Frank, A. T.; Law, S. M.; Brooks III, C. L. A Simple and Fast Approach for Predicting 1H and 13C Chemical Shifts: Toward Chemical Shift-Guided Simulations of RNA. *The Journal of Physical Chemistry B* **2014**, *118*, 12168–12175.

(24) 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.

(25) Sich, C.; Ohlenschläger, O.; Ramachandran, R.; Görlach, M.; Brown, L. R. Structure of an RNA Hairpin Loop with a 5'–CGUUUCG–3' Loop Motif by Heteronuclear NMR Spectroscopy and Distance Geometry. *Biochemistry* **1997**, *36*, 13989–14002.

(26) Szewczak, A.; Moore, P. The Sarcin/Ricin Loop, a Modular RNA. *Journal of Molecular Biology* **1995**, *247*, 81–98.

(27) Frank, A. T. Can Holo NMR Chemical Shifts be Directly Used to Resolve RNA–Ligand Poses? *Journal of Chemical Information and Modeling* **2016**, *56*, 368–376.

(28) Aeschbacher, T.; Schubert, M.; Allain, F. H.-T. A Procedure to Validate and Correct the 13C Chemical Shift Calibration of RNA Datasets. *Journal of Biomolecular NMR* **2012**, *52*, 179–190.

(29) Martin, A. D.; Quinn, K. M.; Park, J. H. Mcmcpack: Markov chain monte carlo in r. **2011**,

(30) Venkatraman, V.; Pérez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Data Set Reveals Limitations of Current 3D Methods. *Journal of Chemical Information and Modeling* **2010**, *50*, 2079–2093.

(31) Alford, R. F. et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **2017**, acs.jctc.7b00125.

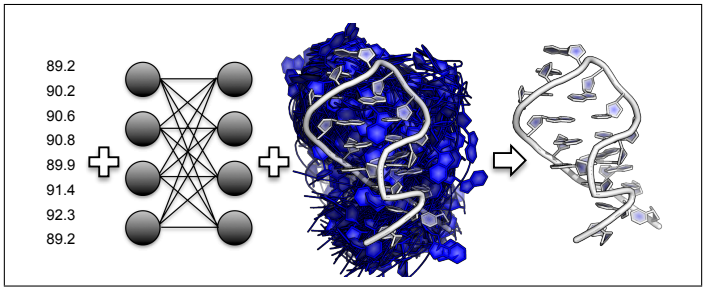
(32) Dejaegere, A.; Bryce, R. A.; Case, D. A. An Empirical Analysis of Proton Chemical Shifts in Nucleic Acids. *ACS Symp. Ser. 732*, 194–206.

(33) Cromsigt, J. A.; Hilbers, C. W.; Wijmenga, S. S. Prediction of Proton Chemical Shifts in RNA–Their Use in Structure Refinement and Validation. *Journal of Biomolecular NMR* **2001**, *21*, 11–29.

(34) Swails, J.; Zhu, T.; He, X.; Case, D. A. AFNMR: Automated Fragmentation Quantum Mechanical Calculation of NMR Chemical Shifts for Biomolecules. *Journal of Biomolecular NMR* **2015**, *63*, 125–139.

(35) Date, K.; Nagi, R. GPU-Accelerated Hungarian Algorithms for the Linear Assignment Problem. *Parallel Computing* **2016**, *57*, 52–72.

Graphical TOC Entry



Illustrating the use of unassigned chemical shift to identify native-like structures of an RNA