

# Supplementary Information

## NMR Data Processing, Visualization, Analysis and Structure Calculation with NMRFX

Ellen Koag<sup>1</sup>, Simon G. Hulse<sup>1</sup>, Gregory L. Helms<sup>1</sup>, Kevin M. Call<sup>1</sup>, Michael F. Summers<sup>2,3</sup>, Jan Marchant<sup>2</sup>, and Bruce A. Johnson<sup>1,\*</sup>

<sup>1</sup>Structural Biology Initiative, Advanced Science Research Center at the CUNY Graduate Center, 85 St. Nicholas Terrace, New York, NY, 10031, USA

<sup>2</sup>Department of Chemistry and Biochemistry, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD, 21250, USA

<sup>3</sup>Howard Hughes Medical Institute, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD, 21250, USA

\*bjohnson@gc.cuny.edu

November 11, 2025

## Contents

<b>S1 Experimental Details</b>	<b>S5</b>
S1.1 Taccalonolide E . . . . .	S5
S1.2 Algae metabolomics . . . . .	S6
<b>S2 Figures</b>	<b>S7</b>
<b>S3 Tables</b>	<b>S11</b>
<b>S4 Code Listings</b>	<b>S22</b>
<b>References</b>	<b>S27</b>

## List of Figures

S1	Protein chemical shift performance of NMRFx compared with that of SHIFTX+, a widely-used structure-based shift prediction algorithm. The RMS deviation between the predicted values and experimentally determined values are shown for backbone atoms. The test set used is the 61 proteins originally used for SHIFTX+. The RMSD values for SHIFTX+ are taken from the original publication [4]. Also shown for NMRFx are the values for 10-fold cross validation of the training set. . . . .	S7
S2	Experimental chemical shifts compared to predicted shifts generated by NMRFx using the 61-protein test set from the ShiftX+ paper. . . . .	S8
S3	Violin plots depicting the distributions of the difference between experimental and predicted chemical shifts from Figure S2. The white bar is the mean, and the box spans the interquartile range between Q1 and Q3. Whiskers include the entire range of values. . . . .	S9
S4	The result of performing a 3D structure calculation on the 36 nt RNA construct, based on distance, angle and RDC constraints determined by NMR assignment. <b>(a)</b> The structure within the computed 50-member ensemble with the lowest RMSD between the experimental and calculated RDC values. <b>(b)</b> A plot of normalized experimental versus estimated RDC values for the structure. Error bars are standard deviations. The RMSD between the points and the black line, which defines equality between the experimental and calculated RDCs, is 0.04. . . . .	S10

## List of Tables

S1	Evaluation of the LARS-based protein chemical shift predictions on the test dataset provided by the SHIFTX+ paper. . . . .	S11
S2	A comparison of updated RNA chemical shift predictions using a newly-adopted FMs approach to a previously reported SVR model. . . . .	S12
S3	Evaluation of structure recalculations by NMRFx and CYANA using deposited restraints in the 100-Protein NMR Spectra dataset. . . . .	S13

S4	Minimal and maximal distance violations across the 10 lowest energy structures generated by NMRFX structure recalculation for each protein in the 100-protein dataset. . . . .	S16
S5	Violations associated with the structure calculation for the 36 nt RNA construct. . . . .	S18
S6	Protein structure attributes and coefficients used by the LARS model to predict chemical shifts for select atoms in glutamine residues. . . . .	S19

## List of Code Listings

S1	Jython script used to process the NUS HNCACB dataset presented in Figure 1 of the main manuscript. . . . .	S22
S2	An example YAML file used to define the layout of spectrum charts in NMRFX.S24	
S3	The YAML file used for the RNA structure calculation presented in Section 3.4 of the main manuscript. . . . .	S26

## Acronyms

**1D** one-dimensional

**2D** two-dimensional

**BMRB** Biological Magnetic Resonance Data Bank

**CV** cross validation

**CYANA** combined assignment and dynamics algorithm for NMR applications

**FID** free induction decay

**FM** factorization machine

**GARP** globally optimized alternating phase rectangular pulse

**GUI** graphical user interface

**HMBC** heteronuclear multiple bond correlation

**HRMAS** high-resolution magic angle spinning

**HSQC** heteronuclear single quantum coherence

**LARS** Least Angle Regression

**MAE** mean absolute error

**NMR** nuclear magnetic resonance spectroscopy

**NUS** non-uniform sampling

**PDB** Protein Data Bank

**RMSD** root-mean-squared deviation

**RNA** ribonucleic acid

**ROESY** rotating frame Overhauser effect spectroscopy

**SVR** support vector regression

**TMSP** trimethylsilylpropanoic acid

**TOCSY** total correlation spectroscopy

## S1 Experimental Details

### S1.1 Taccalonolide E

#### S1.1.1 Sample Preparation

A total of 7 mg of taccalonolide E was dissolved into 0.6 mL of CDCl<sub>3</sub> (Cambridge Isotope Labs, Woburn, MA) and placed into a Wilmad 535pp 5 mm NMR tube (SP-Wilmad Labglass, Vineland, NJ).

#### S1.1.2 NMR Experiments

Data were collected on either a Varian Inova 500 (Varian Associates, Palo Alto, CA) operating at 499.9 MHz for <sup>1</sup>H and 125.7 MHz for <sup>13</sup>C, or a Varian VNMRs 600 operating at 599.7 MHz and 150 MHz for <sup>1</sup>H and <sup>13</sup>C respectively.

A 1D <sup>1</sup>H FID (600 MHz) was acquired with a 5020 Hz spectral width, a 45° pulsewidth, 5 s relaxation delay and 2.5 s acquisition time. 16 transients were accumulated.

A 1D <sup>13</sup>C FID (150 MHz) were acquired with a 36 765 Hz spectral width, a 45° pulsewidth, 0.5 s relaxation delay and a 1.3 s acquisition time using gated broadband <sup>1</sup>H decoupling. 320 transients were accumulated.

<sup>1</sup>H-<sup>13</sup>C HSQC dataset was acquired (600 MHz) using the Varian GHSQCAD pulse program and echo-antiecho for phase discrimination in F1. Spectral widths were 5020 Hz and 23 371 Hz, with acquisition times of 0.23 s and 0.011 s in the <sup>1</sup>H and <sup>13</sup>C dimensions respectively. A total of 256 uniformly-sampled increments were acquired, with 16 transients acquired per increment.

A TOCSY dataset was acquired (600 MHz) using the Varian userlib Ztocsy\_zq pulse program [1] and States-TPPI for phase discrimination in F1. Spectral widths were 5020 Hz in both dimensions and acquisition times were 0.204 s and 0.025 s in the direct and indirect dimensions respectively. A total of 256 increments using 4 transients per increment were acquired using uniform-sampling. The spin lock for TOCSY transfer was 70 ms in duration using a DIPSI-2 mixing sequence.

A ROESY dataset was acquired (500 MHz) using the Varian ROESY pulse program and States-TPPI for phase discrimination in F1. Spectral widths were 5006 Hz in both dimensions and the acquisition times were 0.204 s and 0.02 s in the direct and indirect dimensions respectively. Two hundred increments (uniform-sampling) were collected using 16 transients per increment. A 200 ms spinlock using the transverse ROESY mixing scheme [2] was used to effect cross relaxation.

A <sup>1</sup>H-<sup>13</sup>C HMBC dataset was acquired with the Varian gHMBC pulse program with spectral widths of 5102 Hz and 39 912 Hz with acquisition times of 0.128 s and 0.0065 s in the <sup>1</sup>H and <sup>13</sup>C dimensions respectively. A total of 400 increments using 64 transients per increment were acquired using uniform-sampling. A long range correlation delay (0.0625 s) corresponding to an 8 Hz long-range J value was used and a two-step, 1-bond J suppression sequence was used to lessen the 1-bond <sup>13</sup>C-<sup>1</sup>H correlations.

#### S1.1.3 Data Processing

The 1D <sup>1</sup>H FID was apodized with 0.3 Hz of exponential line broadening, zero filled and Fourier transformed. The spectrum was subsequently integrated, peak picked and the multiplets defined using the Analyze function within NMRfx.

The 1D  $^{13}\text{C}$  FID was apodized with 0.5 Hz of exponential line broadening, zero filled and Fourier transformed.

The HSQC, TOCSY, ROESY and HMBC datasets were apodized (unshifted sine bell for HMBC,  $90^\circ$  shifted sine bell squared for the others) in both dimensions, zero filled, and the indirect dimension was doubled in length using the NESTA algorithm prior to Fourier transform.

## S1.2 Algae metabolomics

### S1.2.1 Sample Preparation

*Chlorella vulgaris* UTEX 395 was cultured using the methods described in [3] where  $\text{NaH}^{13}\text{CO}_3$  (Cambridge Isotope Laboratories, Woburn, MA) was substituted for the unlabeled bicarbonate. The culture was grown for 38 h, and 1.5 mL aliquots were taken at various time points. The aliquots were centrifuged for 1 min at  $5000 \times g$ , and the pellets were suspended in 34  $\mu\text{L}$  of  $\text{D}_2\text{O}$  containing  $0.1 \text{ mg mL}^{-1}$  of TMSP as an internal standard. The suspended cells were transferred to a 4 mm glass rotor for analysis by HRMAS  $^1\text{H}$  NMR spectroscopy.

### S1.2.2 NMR Experiments

The HRMAS spectra were acquired on a 600 MHz Varian VNMRS spectrometer ( $^1\text{H}$  and  $^{13}\text{C}$  carrier frequencies of 599.7 MHz and 150.8 MHz, respectively), spinning at a rate of 2.7 kHz, and oriented at the magic angle. The spectral width was 8000 Hz with an acquisition time of 0.4 s. The relaxation delay was 3.5 s during which time the residual  $\text{H}_2\text{O}$  was saturated with a weak RF field of 50 Hz.  $^{13}\text{C}$  decoupling was accomplished using a GARP-1 modulated sequence with the decoupler offset placed at 60 ppm.  $^1\text{H}$  spectra were acquired in an interleaved fashion to minimize heating effects from the heteronuclear decoupling with 16 scans without  $^{13}\text{C}$  decoupling followed by 16 scans with  $^{13}\text{C}$  decoupling until 128 scans were acquired for both modes. The coupled and decoupled FIDs for each time point were separated into two files denoted by the string **nnn** for coupled and **nnny** for decoupled.

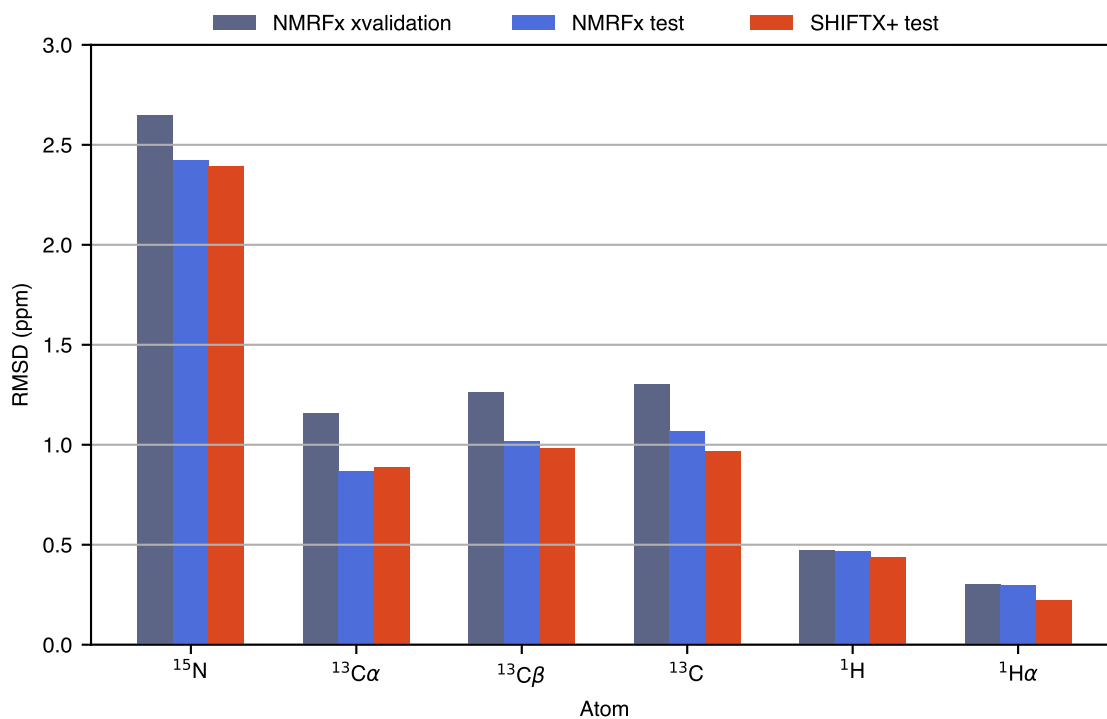


Figure S1: Protein chemical shift performance of NMRFx compared with that of SHIFTX+, a widely-used structure-based shift prediction algorithm. The RMS deviation between the predicted values and experimentally determined values are shown for backbone atoms. The test set used is the 61 proteins originally used for SHIFTX+. The RMSD values for SHIFTX+ are taken from the original publication [4]. Also shown for NMRFx are the values for 10-fold cross validation of the training set.

## S2 Figures

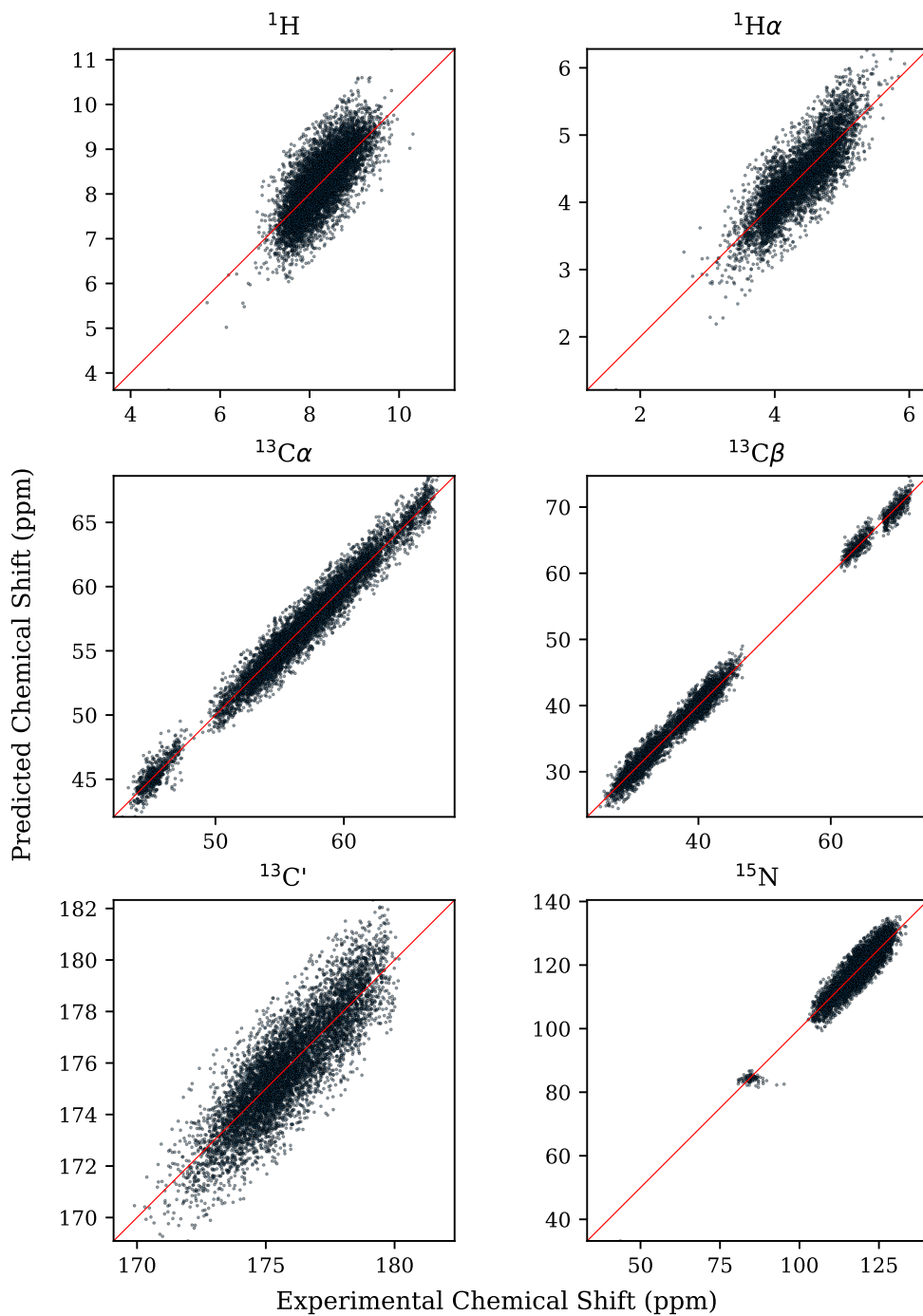


Figure S2: Experimental chemical shifts compared to predicted shifts generated by NMRFx using the 61-protein test set from the ShiftX+ paper [4].

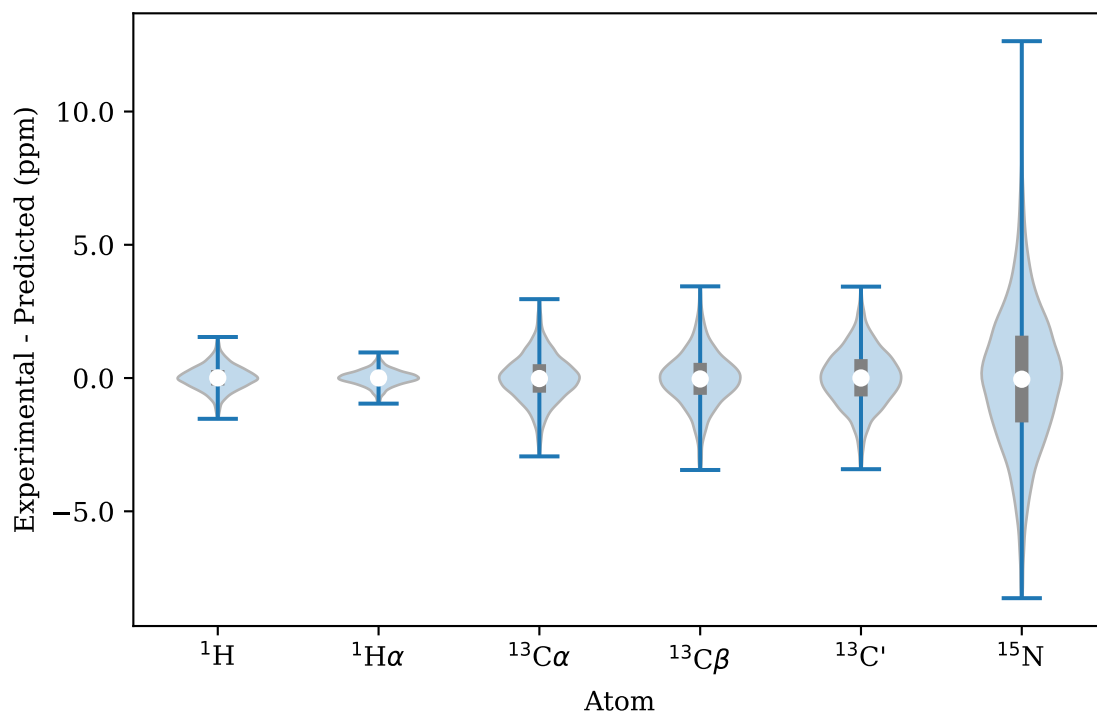


Figure S3: Violin plots depicting the distributions of the difference between experimental and predicted chemical shifts from Figure S2. The white bar is the mean, and the box spans the interquartile range between Q1 and Q3. Whiskers include the entire range of values.

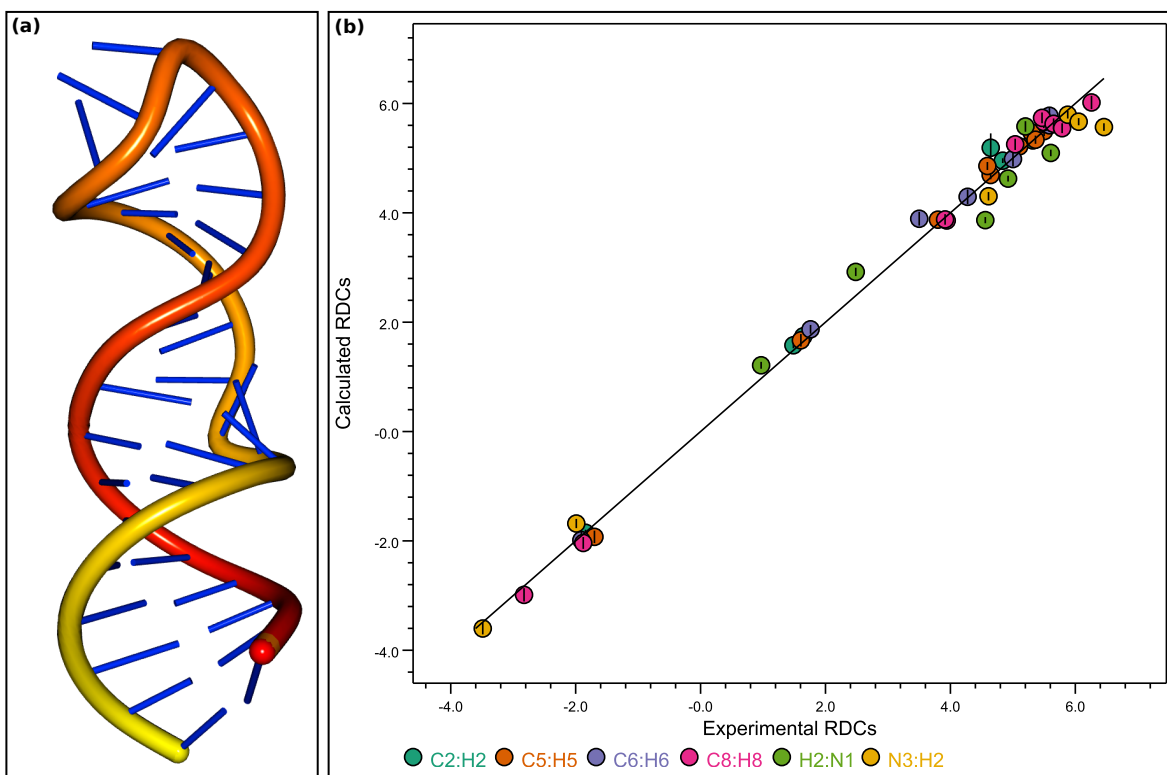


Figure S4: The result of performing a 3D structure calculation on the 36 nt RNA construct, based on distance, angle and RDC constraints determined by NMR assignment. **(a)** The structure within the computed 50-member ensemble with the lowest RMSD between the experimental and calculated RDC values. **(b)** A plot of normalized experimental versus estimated RDC values for the structure. Error bars are standard deviations. The RMSD between the points and the black line, which defines equality between the experimental and calculated RDCs, is 0.04.

Table S1: Evaluation of the LARS-based protein chemical shift predictions with ten-fold CV and on a test dataset comprising of 61 proteins, introduced in the paper outlining SHIFTX+ [4]. All numerical values are in units of ppm.

Atom	CV RMSD	Test RMSD	Test MAE
$^1\text{H}$	0.472	0.470	0.360
$^1\text{H}^\alpha$	0.305	0.300	0.230
$^1\text{H}^\beta$	0.256	0.250	0.190
$^1\text{H}^\gamma$	0.231	0.230	0.170
$^1\text{H}^\delta$	0.259	0.290	0.225
$^1\text{H}^\epsilon$	0.315	0.327	0.250
$^1\text{H}$ Methyl	0.188	0.180	0.130
$^1\text{H}$ Aromatic	0.288	0.290	0.220
$^{13}\text{C}$	1.304	1.070	0.840
$^{13}\text{C}^\alpha$	1.16	0.870	0.670
$^{13}\text{C}^\beta$	1.262	1.020	0.780
$^{13}\text{C}^\gamma$	1.067	0.900	0.645
$^{13}\text{C}^\delta$	2.028	0.620	0.460
$^{13}\text{C}^\epsilon$	0.676	0.390	0.280
$^{13}\text{C}$ Methyl	1.315	1.170	0.910
$^{13}\text{C}$ Aromatic	1.236	1.120	0.820
$^{15}\text{N}$	2.647	2.423	1.767

### S3 Tables

Table S2: A comparison of updated RNA chemical shift predictions using a newly-adopted FMs approach to a previously reported SVR model with analysis using the Automated-Plus approach [5]. The structure of the table is identical to that of Table 1 in [5], except that results for  $^{15}\text{N}$  nuclei, previously unconsidered, are reported for the FM approach. The database of shifts was also expanded with new RNA entries from the BMRB. RMSD values were calculated between the predicted and experimental shift values.  $n$  denotes the number of shifts included in the calculation for the given category. <sup>a</sup>Ten-fold cross validation performed during model training for all atoms. <sup>b</sup>Canonical bases are the central base in a 5 base stretch in which all 5 base pairs have GC or AU base pairing and no other attributes such as being in a triplet, kissing interaction or pseudoknots are present. <sup>c</sup>Non-canonical bases are the same as canonical, but the first and/or fifth bases may be GU wobble base pairs, mismatched, unpaired (e.g. loops) or not-present (e.g. the 5' or 3' termini). <sup>d</sup>Other bases are all bases that are in neither the canonical nor non-canonical categories. <sup>e</sup>All denotes the average across all categories (b-d).

Category	SVR		FM	
	RMSD	$n$	RMSD	$n$
$^1\text{H}$				
Cross-validated <sup>a</sup>	0.13	18,774	0.11	38,099
Canonical <sup>b</sup>	0.06	3,020	0.09	7,422
Non-canonical <sup>c</sup>	0.07	2,903	0.09	6,120
Other <sup>d</sup>	0.11	12,851	0.11	24,557
All <sup>e</sup>	0.10	18,774	0.10	38,099
$^{13}\text{C}$				
Cross-validated	0.83	9,642	0.84	17,956
Canonical	0.46	1,630	0.44	3,367
Non-canonical	0.47	1,526	0.49	2,786
Other	0.85	6,486	0.70	11,803
All	0.75	9,642	0.64	17,956
$^{15}\text{N}$				
Cross-validated	–	–	1.32	3,460
Canonical	–	–	0.36	667
Non-canonical	–	–	0.33	536
Other	–	–	0.36	2,257
All	–	–	0.39	3,460

Table S3: Evaluation of structure recalculations by NMRFX and CYANA [6] using deposited restraints in the 100-Protein NMR Spectra dataset [7]. RMSD values were calculated by superimposing the backbone (i.e.  $^{15}\text{N}$ ,  $^{13}\text{C}'$ ,  $^{13}\text{C}^\alpha$ ) atoms of the mean calculated structure and the mean PDB structure, using NMRFX’s command-line program [super](#). PDB IDs labeled with an asterisk do not have an associated CYANA recalculation.

	PDB ID	Residue ranges	RMSD (Å)	
			CYANA	NMRFX
1	6SVC*	7–29	–	–
2	2JVD	4–37	0.43	0.33
3	2K57	5–52	0.71	0.70
4	6SOW	8–55	0.48	0.49
5	2LX7	5–59	1.02	1.27
6	2MA6	10–57	1.02	1.15
7	2JRM	6–47	0.61	0.67
8	1YEZ	15–25, 29–66	0.76	0.84
9	2L9R	13–56	0.55	0.49
10	2K52	7–70	0.94	0.99
11	2KRS	2–61	0.49	0.62
12	2K53	8–28, 39–66	0.69	0.49
13	2JT1	5–57, 66–69	0.72	0.78
14	2JVO	6–71	0.76	0.71
15	2ERR	2–75	1.02	1.11
16	2L1P	19–78	1.29	1.69
17	2LN3	6–72	0.56	0.77
18	2HEQ	17–20, 34–68	0.65	0.71
19	2KK8	10–82	1.12	1.16
20	2KD0	13–81	1.01	1.21
21	2LML	3–78	1.30	1.15
22	2K3D	2–81	0.84	0.72
23	2LK2	14–65	0.95	1.01
24	MH04*	–	–	–
25	1PQX	13–19, 28–34, 38–65, 70–81	1.14	0.98
26	2L33	19–36, 46–79	0.43	0.44
27	2KZV	9–80	1.17	1.24
28	2KCT	11–37, 44–83	0.98	1.10
29	2MDR	9–89	1.37	1.47
30	2FB7	20–52, 74–87	0.45	0.51
31	2MB0	8–41, 49–85	0.65	0.66

	PDB ID	Residue ranges	RMSD (Å)	
			CYANA	NMRFX
32	2L05	19–89	0.81	0.98
33	2KJR	15–23, 28–94	0.69	0.77
34	2M5O	17–91	0.75	1.11
35	MDM2*	–	–	–
36	2LNA	16–49, 59–94	0.90	0.89
37	2LA6	15–97	0.79	0.69
38	6FIP	11–93	0.86	0.91
39	2LEA	15–45, 55–88	0.35	0.56
40	2LL8	4–90	0.94	1.04
41	2KPN	12–84	0.83	0.67
42	2K0M	7–70, 76–93	0.75	0.76
43	2K5V	2–29, 38–78, 83–94	1.03	1.04
44	2MQL	15–84	0.53	0.64
45	2K75	3–92	0.91	1.14
46	2LTM	14–99	0.81	0.73
47	2KOB	3–93	0.96	0.63
48	2KHD	31–97	1.15	1.40
49	2RN7	10–55	0.51	0.59
50	2LXU	9–95	0.73	1.41
51	2KIF	3–97	0.72	0.73
52	2KBN	5–29, 34–54, 58–76, 81–94	0.40	0.62
53	2MK2	14–108	1.17	1.42
54	2K50	10–34, 42–85, 92–105	0.93	0.83
55	2KL5	12–53, 58–66, 76–86, 93–99	0.68	0.64
56	2LTA	4–98	1.00	1.21
57	2KIW	2–81	1.63	1.29
58	2LVB	3–102	0.79	0.84
59	2LND	3–48, 52–101	0.69	0.76
60	1WQU	8–106	0.49	0.61
61	2KL6	6–106	0.85	0.73
62	6GT7	7–30, 40–87, 94–113	0.56	0.63
63	2JN8	12–26, 31–109	0.78	0.87
64	2K5D	19–50, 55–84, 98–107	0.70	0.76
65	2KD1	7–89	1.23	0.99
66	2LTL	19–35, 39–41, 46–110	1.12	0.95
67	2KVO	3–23, 28–103	0.72	0.89

	PDB ID	Residue ranges	RMSD (Å)	
			CYANA	NMRFX
68	1T0Y	4-83	0.88	0.74
69	2KCD	3-108	0.99	1.65
70	2KRT	6-114	1.30	1.24
71	2LFI	2-104	1.03	2.16
72	2JQN	3-111	1.18	1.98
73	2L7Q	12-37, 46-101, 105-114	0.76	1.31
74	2KFP	3-115	0.71	0.81
75	1SE9*	17-84, 94-101	-	-
76	2L3G	13-123	0.65	0.74
77	2L3B	14-38, 45-113	0.66	1.18
78	2LRH	3-122	0.91	1.15
79	1VEE	6-123	0.38	0.51
80	2K1G	5-78, 83-122	0.77	0.69
81	2KKZ	5-80, 86-118	1.17	1.14
82	1VDY	9-102, 113-128	0.45	0.52
83	2KKL	33-90, 96-125	0.50	0.89
84	2N4B*	2-26, 40-54, 66-134	-	-
85	2L8V	4-22, 37-65, 73-129	1.02	1.39
86	2LGH	2-109, 113-135	1.37	1.22
87	2K1S	3-140	0.94	1.74
88	2M4F	23-46, 51-57, 63-94, 103-114, 120-129, 136-148	0.25	0.36
89	2JXP	16-144	1.61	1.47
90	2L06	15-38, 45-141	0.77	1.10
91	2LAH	14-25, 33-158	0.83	1.10
92	2LAK	10-37, 68-77, 93-139	0.55	0.87
93	2L82	3-151	0.81	0.74
94	2M47	5-25, 40-56, 66-156	0.93	2.09
95	2K3A	57-102, 108-127, 138-153	0.75	0.86
96	2M7U*	12-151	-	-
97	2B3W	16-162	1.01	1.07
98	KRAS*	-	-	-
99	2G0Q	18-54, 60-126	0.91	1.16
100	2LF2	6-44, 53-69, 76-105, 111-165	1.24	1.67

Table S4: Minimal and maximal distance violations across the 10 lowest energy structures generated by NMRFX structure recalculation for each protein in the 100-protein dataset (see Table S3). Values were not obtained for proteins whose PDB IDs are labeled with an asterisk.

		Distance (Å)				Distance (Å)	
	PDB ID	Min	Max		PDB ID	Min	Max
1	6SVC*	–	–	51	2KIF	0.10	0.40
2	2JVD	0.00	0.00	52	2KBN	0.00	0.21
3	2K57	0.00	0.23	53	2MK2	0.00	0.22
4	6SOW	0.31	0.45	54	2K50	0.17	0.42
5	2LX7	0.00	0.11	55	2KL5	0.11	0.53
6	2MA6	0.57	0.57	56	2LTA	0.00	0.29
7	2JRM	0.30	0.49	57	2KIW	0.00	0.16
8	1YEZ	0.00	0.29	58	2LVB	0.12	0.27
9	2L9R	0.00	0.14	59	2LND	0.30	0.41
10	2K52	0.00	0.43	60	1WQU	0.47	0.80
11	2KRS	0.00	0.15	61	2KL6	0.21	0.28
12	2K53	0.30	0.83	62	6GT7	0.76	0.86
13	2JT1	0.00	0.31	63	2JN8	0.11	0.30
14	2JVO	0.59	1.47	64	2K5D	0.18	0.43
15	2ERR	0.75	1.02	65	2KD1	0.22	0.59
16	2L1P	0.00	0.22	66	2LTL	0.00	0.24
17	2LN3	0.00	0.21	67	2KVO	0.00	0.25
18	2HEQ	0.00	0.41	68	1T0Y	0.17	0.58
19	2KK8	0.00	0.20	69	2KCD	0.00	0.29
20	2KD0	0.00	0.18	70	2KRT	0.00	0.26
21	2LML	0.00	0.33	71	2LFI	0.14	0.49
22	2K3D	0.00	0.36	72	2JQN	0.00	0.36
23	2LK2	0.11	0.24	73	2L7Q	0.00	0.41
24	MH04*	–	–	74	2KFP	0.20	0.41
25	1PQX	0.13	0.56	75	1SE9*	–	–
26	2L33	0.11	0.39	76	2L3G	0.00	0.27
27	2KZV	0.00	0.23	77	2L3B	0.00	0.25
28	2KCT	0.00	0.28	78	2LRH	0.14	0.48
29	2MDR	1.56	1.60	79	1VEE	0.61	1.01
30	2FB7	0.42	0.60	80	2K1G	0.13	0.23
31	2MB0	0.39	0.70	81	2KKZ	0.00	0.21
32	2L05	0.00	0.34	82	1VDY	0.52	0.74
33	2KJR	0.00	0.10	83	2KKL	0.14	0.56

	PDB ID	Distance (Å)			PDB ID	Distance (Å)	
		Min	Max			Min	Max
34	2M5O	0.13	0.43	84	2N4B*	–	–
35	MDM2*	–	–	85	2L8V	0.00	0.16
36	2LNA	0.18	0.37	86	2LGH	0.73	0.76
37	2LA6	0.13	0.38	87	2K1S	0.14	0.38
38	6FIP	1.10	1.42	88	2M4F	0.26	0.58
39	2LEA	1.04	1.43	89	2JXP	0.20	0.38
40	2LL8	0.00	0.37	90	2L06	0.00	0.39
41	2KPN	0.13	0.28	91	2LAH	0.34	0.76
42	2K0M	0.11	0.18	92	2LAK	0.14	0.23
43	2K5V	0.00	0.21	93	2L82	0.12	0.49
44	2MQL	0.14	0.41	94	2M47	0.15	0.42
45	2K75	0.00	0.19	95	2K3A	0.37	0.49
46	2LTM	0.00	0.51	96	2M7U*	–	–
47	2KOB	0.00	0.21	97	2B3W	0.18	0.25
48	2KHD	0.00	0.13	98	KRAS*	–	–
49	2RN7	0.00	0.13	99	2G0Q	0.52	1.18
50	2LXU	0.00	0.26	100	2LF2	0.00	0.16

Table S5: NMRFx generates an output file, `analysis.txt`, after batch mode generation of multiple structures. The file lists the names of the involved atoms, the bound (upper or lower) that is violated, the number of structures with that violation, the mean value of that violation, the violation in the structure with the largest violation, and a text string where + indicates a structure with a violation.

Atom 1	Atom 2	$n_{\text{viol}}$	Bound	Mean	Max	Structures
1:31.H6	1:32.H6	3	5.00	0.13	0.28	..++.....+
1:29.H61	1:8.O4	6	2.00	0.22	0.28	..++++.++
1:10.H5'	1:10.H8	4	3.20	0.16	0.33	...++++...
1:10.P	1:18.P	9	13.00	0.26	0.37	++.+++++++
1:2.H8	1:3.H6	10	5.00	0.46	0.52	+++++++
1:3.H3	1:34.N1	6	1.92	0.20	0.31	+.+.++.++
1:25.H3'	1:25.H8	6	3.50	0.47	0.82	++++. . .++
1:28.N3	1:9.H1	8	1.89	0.24	0.33	+.++++.++
1:22.H6	1:23.H6	5	5.00	0.24	0.46	+.+. . . .++
1:34.H8	1:35.H6	8	5.00	0.25	0.40	++.+. . .++++
1:27.P	1:4.P	4	11.00	0.21	-0.30	+. . . . .++++.
1:10.H8	1:11.H8	8	5.00	0.44	0.62	++.++++.++
1:10.H5''	1:10.H8	7	3.80	0.18	0.28	...+++++++

Table S6: Protein structure attributes and their coefficient values used by the LARS model to predict chemical shifts for select atoms in glutamine residues. Other amino acids use the same attributes, but with different coefficient values. Attributes are sorted in descending order by the sum, over the atom types, of the absolute value of the coefficients, so more important attributes are near the top.

Attribute	$^{13}\text{C}^\alpha$	$^{13}\text{C}^\beta$	$^{13}\text{C}'$	$^{15}\text{N}$	$^1\text{H}^\text{N}$	$^1\text{H}^\alpha$
hydrogen shift1	0.0000	0.0000	0.0000	0.0652	-1.6809	0.5948
hydrogen shift3	0.0000	0.0000	0.0000	-0.0176	1.7780	-0.5381
$\cos(\psi_i)$	0.1952	-0.2550	0.0425	-0.0886	0.3558	-0.1756
$\sin(\psi_i) \sin(\phi_i)$	0.1519	-0.1201	0.0952	-0.2172	-0.0837	-0.2804
$\cos(\phi_i)$	0.1103	-0.1775	0.0377	0.1650	0.1308	-0.1833
$\cos(\psi_{i-1})$	-0.0039	-0.0325	-0.0244	-0.2673	-0.4064	-0.0552
$\sin(2\phi_i)$	-0.1704	0.3059	-0.0625	-0.0539	0.0432	0.1113
ring current shift	0.0336	0.0881	0.0440	0.0021	0.2123	0.2356
$\cos(3\psi_{i-1})$	-0.0200	0.0189	-0.0208	0.3886	0.0657	0.0754
$\sin(\psi_i)$	-0.1177	-0.0967	0.0127	-0.0578	-0.1189	-0.1786
$\cos(\psi_i) \sin(\phi_i)$	0.0654	-0.1504	-0.0298	-0.0521	0.0756	-0.1164
$\sin(2\psi_i)$	-0.1099	-0.1867	-0.0841	-0.0322	-0.0405	-0.0313
$\cos(\phi_{i+1})$	0.0273	-0.0916	0.1112	0.0027	0.0061	-0.2196
$\cos(\chi_i)$	-0.0528	-0.0877	0.0534	-0.1073	0.1096	0.0357
$\sin(\phi_i)$	0.0711	-0.0706	-0.0094	-0.1707	-0.0234	-0.0988
$\sin(\psi_{i-1})$	0.0061	0.0066	-0.0581	0.1980	0.1215	-0.0306
$\cos(\psi_i) \cos(\phi_i)$	-0.0700	0.0698	-0.0201	0.0374	0.0470	0.1616
$\sin(\chi_i)$	-0.0646	-0.0692	-0.1045	0.0402	0.0800	0.0250
$\cos(\psi_{i+1})$	0.0919	-0.0704	0.1332	0.0529	0.0204	-0.0013
$\sin(\phi_i) \cdot \text{size}_{i-1}$	-0.0305	-0.0006	-0.0300	0.2757	0.0267	-0.0046
$\cos(\chi_{2i})$	-0.0972	-0.0501	-0.0189	0.0545	-0.0837	0.0511
$\cos(2\chi_{2i})$	-0.0624	0.0739	-0.0067	-0.0295	-0.1613	0.0107
$\sin(\psi_i) \cdot \text{size}_{i+1}$	-0.0321	0.0517	-0.0306	-0.0279	0.0701	0.1212
$\cos(2\psi_{i-1})$	0.0139	-0.0338	-0.0196	-0.2107	-0.0436	-0.0109
$\cos(\phi_{i-1})$	0.0168	-0.0854	0.0400	-0.0738	0.0211	-0.0863
$\cos(\psi_i) \cdot \text{size}_{i+1}$	0.0070	0.1169	0.0005	-0.0251	-0.0747	0.0821
$\sin(\chi_i) \sin(\chi_{2i})$	-0.0259	0.0513	0.0304	-0.0117	-0.1731	-0.0110
$\cos(3\phi_{i-1})$	0.0292	-0.0594	-0.0171	-0.0834	0.0841	-0.0089
$\cos(2\psi_i)$	-0.0434	0.0485	-0.0036	-0.1289	0.0112	-0.0420
$\sin(\phi_i) \cdot \text{hydrophobicity}_{i-1}$	-0.0348	-0.0225	-0.0255	0.1638	0.0023	0.0287
$\sin(\phi_i) \cos(\chi_i)$	-0.0906	-0.0567	0.0194	-0.0106	0.0435	0.0547
$\cos(\chi_i) \cos(\chi_{2i})$	0.0505	-0.0246	-0.0064	0.0159	0.1313	-0.0416
$\cos(3\psi_i)$	-0.0426	0.0273	-0.0515	-0.0804	-0.0116	0.0401
$\sin(\psi_{i+1})$	-0.0584	-0.0273	-0.0320	-0.0588	-0.0475	0.0289
$\cos(\phi_i) \cdot \text{size}_{i-1}$	0.0552	-0.0664	0.0475	-0.0484	-0.0201	-0.0052
$\sin(\phi_i) \cdot \text{aromaticity}_{i-1}$	-0.0112	0.0108	-0.0211	-0.0868	-0.0860	-0.0219
$\sin(\psi_i) \sin(\chi_i)$	0.0311	0.0412	0.0373	0.0318	0.0566	-0.0389
$\sin(3\psi_i)$	-0.0081	0.0755	0.0218	-0.0636	-0.0009	0.0633
$\cos(2\phi_i)$	-0.0232	0.0564	-0.0247	0.0184	0.0639	-0.0403
$\sin(\psi_i) \sin(\psi_{i-1})$	-0.0563	0.0191	0.0160	-0.0333	0.0405	0.0544

Attribute	$^{13}\text{C}^\alpha$	$^{13}\text{C}^\beta$	$^{13}\text{C}'$	$^{15}\text{N}$	$^1\text{H}^\text{N}$	$^1\text{H}^\alpha$
$\sin(2\phi_{i+1})$	0.0569	0.0144	0.0430	0.0044	0.0845	0.0144
$\cos(3\phi_i)$	-0.0240	-0.0854	0.0153	-0.0037	0.0256	0.0620
$\cos(3\phi_{i+1})$	0.0024	0.0316	-0.0499	-0.0051	-0.0853	-0.0354
$\cos(\psi_i) \cos(\psi_{i-1})$	-0.0156	0.0358	0.0206	-0.0375	0.0570	0.0312
$\cos(\psi_i) \sin(\chi_i)$	0.0306	0.0199	0.0258	0.0130	0.0872	-0.0181
$\sin(3\psi_{i-1})$	0.0150	0.0095	0.0204	-0.0247	-0.0629	-0.0542
$\sin(\psi_i) \cdot \text{proline}_{i+1}$	0.0022	0.0827	0.0169	-0.0475	0.0189	-0.0101
$\sin(\phi_i) \cos(\chi_{2i})$	-0.0292	0.0244	-0.0023	0.0270	-0.0453	0.0494
$\sin(\phi_{i+1})$	0.0271	0.0326	-0.0391	0.0142	-0.0056	-0.0589
$\cos(\phi_i) \cdot \text{hydrophobicity}_{i-1}$	0.0376	-0.0516	0.0183	-0.0358	-0.0131	-0.0197
$\sin(\chi_{2i})$	-0.0181	0.0105	0.0426	-0.0113	-0.0699	0.0225
$\cos(3\psi_{i+1})$	-0.0149	0.0384	-0.0465	0.0286	0.0430	0.0024
$\sin(\phi_i) \sin(\chi_i)$	-0.0213	-0.0230	-0.0577	0.0014	-0.0162	-0.0509
$\sin(\psi_i) \cdot \text{hydrophobicity}_{i+1}$	-0.0077	0.0063	-0.0072	-0.0494	0.0472	0.0469
$\cos(\psi_i) \cdot \text{proline}_{i+1}$	-0.0244	0.0900	0.0071	-0.0088	0.0023	0.0308
$\sin(\phi_{i-1})$	-0.0014	0.0262	-0.0347	0.0077	0.0751	-0.0125
$\cos(2\chi_i)$	0.0102	0.0413	0.0028	0.0101	-0.0805	-0.0097
$\sin(\psi_i) \cos(\phi_i)$	-0.0607	-0.0297	0.0243	-0.0261	0.0101	-0.0019
$\cos(\psi_i) \cdot \text{hydrophobicity}_{i+1}$	-0.0126	0.0349	0.0007	-0.0187	-0.0363	0.0480
$\cos(2\phi_{i-1})$	0.0091	-0.0536	-0.0127	-0.0115	0.0528	0.0105
$\sin(\chi_i) \cos(\chi_{2i})$	-0.0121	-0.0100	0.0059	0.0726	0.0398	0.0059
$\sin(\phi_i) \sin(\chi_{2i})$	-0.0391	-0.0045	0.0440	0.0116	-0.0001	0.0417
$\sin(\psi_i) \cos(\psi_{i-1})$	0.0048	0.0021	-0.0097	-0.0527	0.0307	-0.0389
$\sin(2\psi_{i-1})$	0.0203	-0.0220	0.0195	0.0124	-0.0525	-0.0110
$\cos(\phi_i) \cos(\chi_{2i})$	-0.0062	0.0274	0.0027	-0.0006	0.0563	0.0442
$\sin(\phi_i) \cdot \text{proline}_{i-1}$	-0.0161	0.0192	-0.0023	0.0367	-0.0279	0.0317
$\cos(2\psi_{i+1})$	0.0111	-0.0422	0.0392	0.0081	0.0291	-0.0034
$\sin(3\phi_{i+1})$	-0.0064	0.0236	-0.0265	-0.0260	0.0098	-0.0351
$\sin(3\phi_{i-1})$	0.0139	-0.0039	-0.0374	0.0294	0.0250	-0.0114
$\sin(2\chi_{2i})$	0.0407	0.0023	-0.0005	0.0085	0.0422	-0.0229
$\cos(2\phi_{i+1})$	-0.0291	0.0057	0.0048	0.0149	-0.0124	0.0431
$\sin(2\chi_i)$	0.0182	0.0090	0.0085	-0.0082	-0.0370	-0.0288
$\cos(\phi_i) \sin(\chi_i)$	-0.0110	0.0140	0.0011	0.0112	-0.0260	-0.0448
$\sin(2\psi_{i+1})$	-0.0441	-0.0094	-0.0424	0.0064	-0.0020	-0.0030
$\sin(\psi_i) \cos(\chi_i)$	0.0215	-0.0097	-0.0289	0.0082	0.0165	-0.0168
$\sin(3\phi_i)$	-0.0578	-0.0050	0.0251	0.0051	-0.0052	-0.0009
$\cos(\psi_i) \cos(\chi_{2i})$	0.0102	0.0389	0.0145	-0.0018	0.0159	0.0156
$\cos(\phi_i) \cos(\chi_i)$	0.0160	0.0134	-0.0025	0.0237	-0.0130	-0.0281
$\cos(\psi_i) \cos(\chi_i)$	0.0045	-0.0078	-0.0215	0.0259	0.0206	0.0114
$\sin(\psi_i) \cos(\chi_{2i})$	0.0028	0.0398	-0.0009	0.0145	-0.0027	-0.0306
$\cos(\psi_i) \cdot \text{aromaticity}_{i+1}$	0.0020	-0.0209	-0.0044	0.0078	0.0517	-0.0009
$\sin(3\psi_{i+1})$	0.0043	0.0017	-0.0279	0.0389	-0.0004	0.0110
$\cos(\phi_i) \cdot \text{aromaticity}_{i-1}$	-0.0058	0.0103	-0.0150	0.0249	0.0020	-0.0248
$\cos(\psi_i) \cdot \text{charge}_{i+1}$	0.0045	0.0127	-0.0019	0.0183	-0.0196	0.0176
$\cos(\chi_i) \sin(\chi_{2i})$	0.0272	0.0036	0.0058	0.0249	0.0106	-0.0022
$\cos(\phi_i) \cdot \text{charge}_{i-1}$	0.0047	-0.0222	-0.0104	0.0176	-0.0056	-0.0111

Attribute	$^{13}\text{C}^\alpha$	$^{13}\text{C}^\beta$	$^{13}\text{C}'$	$^{15}\text{N}$	$^1\text{H}^\text{N}$	$^1\text{H}^\alpha$
electrostatic interaction shift	0.0000	0.0000	0.0000	-0.0436	0.0131	0.0122
$\cos(\psi_i) \sin(\psi_{i-1})$	-0.0093	0.0141	-0.0157	-0.0027	0.0048	-0.0204
$\sin(\psi_i) \cdot \text{charge}_{i+1}$	0.0031	0.0054	-0.0037	0.0189	0.0050	0.0254
$\sin(\psi_i) \sin(\chi_{2i})$	0.0008	-0.0168	-0.0068	-0.0216	0.0006	-0.0145
$\cos(\psi_i) \sin(\chi_{2i})$	0.0060	-0.0198	-0.0094	-0.0074	-0.0004	-0.0165
$\cos(\phi_i) \cdot \text{proline}_{i-1}$	-0.0026	0.0015	-0.0016	-0.0060	0.0077	0.0339
$\sin(\phi_i) \cdot \text{charge}_{i-1}$	-0.0057	0.0070	-0.0073	0.0205	0.0087	0.0012
$\sin(2\phi_{i-1})$	-0.0096	0.0028	-0.0029	-0.0118	0.0021	-0.0144
$\sin(\psi_i) \cdot \text{aromaticity}_{i+1}$	0.0020	0.0008	-0.0028	-0.0174	0.0092	0.0010
$\cos(\phi_i) \sin(\chi_{2i})$	-0.0030	0.0018	0.0007	0.0012	0.0067	0.0007
mean shift	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
methyl bond	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
contact sum	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
disulfide bond	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

## S4 Code Listings

---

**Code Listing S1:** Jython script (called `process.py` by default) used to process the NUS HNCACB dataset presented in Figure 1 of the main manuscript. A script of this form (minus the comments which have been added for edification) can be generated indirectly using the Processing Accordion from the NMRFX GUI.

---

```
1  # Processing operations are exposed by NMRFX's `pyproc` library
2  from pyproc import *
3
4  # Use 10 processor cores.
5  # By default, half of the total number of cores are used.
6  procOpts(nprocess=10)
7  FID("/path/to/raw/nus/data") # Load data
8  CREATE("/path/to/raw/nus/data/spec.nv") # Create new NMRFX dataset
9  acqOrder("21")
10 acqarray(0, 0, 0)
11 fixdsp(True)
12 label("HN", "15N", "13C")
13 ref("", "", "")
14 skip(False, False, False)
15 acqmode("complex", "hyper", "hyper")
16
17 # === 1H dimension processing ===
18 DIM(1)
19 SUPPRESS() # Water signal suppression
20 SB() # Sine-bell apodization
21 ZF() # Zero-filling
22 FT() # Fourier Transform
23 PHASE(ph0=-94.4, ph1=0.0) # Phasing
24 EXTRACTP(start=10.75, end=5.9, mode="region") # Region selection
25
26 # === Indirect matrix (15N & 13C) processing ===
27 DIM(2, 3)
28 # Specify NUS schedule
29 SCHEDULE("/path/to/raw/nus/data/nuslist.scd")
30 ZFMAT(zfy=1, zfz=1) # Zero-filling
31 SB()
32 # `PHASE_ID` Fourier transforms the data, performs
33 # phase-correction, and then performs inverse Fourier
34 # transformation, returning the data to the time-domain
35 # for processing with GRINS.
36 # Each argument is a 2-tuple corresponding to the 15N
37 # and 13C dimensions, respectively.
38 # `negateImag` indicates whether the FID points should be
39 # transformed to their complex conjugates.
40 # `negatePairs` indicates whether every second FID point should
```

```

41 # be multiplied by -1.
42 PHASE_ID(
43     ph0=[0.00, 0.00],
44     ph1=[0.00, 0.00],
45     negateImag=[False, False],
46     negatePairs=[True, True],
47 )
48 # Perform NUS reconstruction with the GRINS algorithm
49 GRINS(noiseRatio=5.0)
50
51 # === 15N dimension processing ===
52 DIM(2)
53 SB()
54 ZF()
55 FT()
56
57 # === 13C dimension processing ===
58 DIM(3)
59 SB()
60 ZF()
61 FT()
62
63 # === Baseline correction (Whittaker smoothing) ===
64 DIM(1)
65 BCWHIT()
66 DIM(2)
67 BCWHIT()
68 DIM(3)
69 BCWHIT()
70
71 # === Run commands ===
72 run()

```

---

---

**Code Listing S2:** An example of a YAML file used to define the layout of spectrum charts in NMRfX. This layout is that used to generate the arrangement of NOESY, HMQC and TOCSY experiments in Figure 10.

---

```
1  layouts :
2    - name : RNA-UMBC
3      layout :
4
5        # === Top-left panel ===
6        - type : [noesy,tocsy] # Datasets whose names contain
7                                #   "noesy" or "tocsy" will be in
8                                #   this chart
9          loadpeaks : true      # If the dataset has an associated
10                                 #   peak list, load it
11          row : 0               # Display in first row of the grid
12          column : 0           # Display in first column of the grid
13          rowspan : 2          # Chart spans 2 rows of the grid
14          colspan : 2          # Chart spans 2 columns of the grid
15          x : [6.4, 8.5]       # X-axis spans 6.4 to 8.5 ppm
16          y : [4.8, 6.2]       # Y-axis spans 4.8 to 6.2 ppm
17          xsync : arom         # Synchronize x-axis of this chart
18                                 #   with other axes labelled "arom"
19          ysync : h1py         # Synchronize y-axis of this chart
20                                 #   with other axes labelled "h1py"
21
22        # === Bottom-left panel ===
23        - type : [hmqc]
24          loadpeaks : true
25          row : 2
26          column : 0
27          rowspan : 1
28          colspan : 2
29          x : [6.4, 8.5]
30          y : [133.5, 157]
31          xsync: arom
32
33        # === Top-right panel ===
34        - type : [noesy,tocsy]
35          loadpeaks : true
36          row : 0
37          column : 2
38          rowspan : 2
39          colspan : 1
40          x : [4.8, 6.2]
41          y : [4.8, 6.2]
42          xsync : h1p
43          ysync: h1py
44
```

```
45     # === Bottom-right panel ===
46     - type : [hmqc]
47       loadpeaks : true
48       row : 2
49       column : 2
50       rowspan : 1
51       colspan : 1
52       x : [4.8, 6.2]
53       y : [88, 106.5]
54       xsync : h1p
```

---

---

**Code Listing S3:** The YAML file used for the structure of the 36 nt RNA construct considered in Section 3.4 of the main manuscript. The file specifies the molecular topology, the files used for restraints, and the parameters used in the torsion angle dynamics and optimization.

---

```
1  molecule :
2    entities :
3      - sequence : GGUUGAGUGGAACUGUGAAGUUCGGAACACUCAACC
4        ptype : RNA
5        chain : A
6
7  tree:
8
9  distances :
10   - file : slcm-hb
11     type : cyana
12   - file : slcm-pp
13     type : cyana
14   - file : slcm-noe
15     type : cyana
16
17  angles :
18   - file : all.aco
19     type : cyana
20
21  rdcs :
22   - file : slca_rdcs_for_analyst.dat
23     type : nmrfx
24
25  anneal:
26    dynOptions :
27      steps : 15000
28      highTemp : 5000.0
29      dfreeSteps : 10000
30    force :
31      tors : 0.1
32      dih : 10
33      irp : -0.2
34      rdc : 0.5
35    stage_low:
36      force :
37        repel : 4.0
```

---

## References

- [1] M. J. Thrippleton and J. Keeler. Elimination of zero-quantum interference in two-dimensional NMR spectra. *Angewandte Chemie International Edition*, 42(33):3938–3941, 2003.
- [2] T. L. Hwang and A. J. Shaka. Cross relaxation without TOCSY: transverse rotating-frame Overhauser effect spectroscopy. *Journal of the American Chemical Society*, 114(8):3157–3159, 1992.
- [3] E. J. Lohman, R. D. Gardner, T. Pedersen, B. M. Peyton, K. E. Cooksey, and R. Gerlach. Optimized inorganic carbon regime for enhanced growth and lipid accumulation in *Chlorella vulgaris*. *Biotechnology for Biofuels*, 8(1):82, 2015.
- [4] B. Han, Y. Liu, S. W. Ginzinger, and D. S. Wishart. SHIFTX2: significantly improved protein chemical shift prediction. *Journal of Biomolecular NMR*, 50(1):43–57, 2011.
- [5] J. D. Brown, M. F. Summers, and B. A. Johnson. Prediction of hydrogen and carbon chemical shifts from RNA using database mining and support vector regression. *Journal of Biomolecular NMR*, 63(1):39–52, 2015.
- [6] P. Güntert. Automated NMR structure calculation with CYANA. In A. Kristina Downing, editor, *Protein NMR Techniques*, pages 353–378. Humana Press, 2004.
- [7] P. Klukowski, F. F. Damberger, F. H.-T. Allain, H. Iwai, H. Kadavath, T. A. Ramelot, G. T. Montelione, R. Riek, and P. Güntert. The 100-protein NMR spectra dataset: A resource for biomolecular NMR data analysis. *Scientific data*, 11(1):30–30, 2024.