Supplementary Material for

Protein Structure Prediction Assisted with Sparse NMR Data in CASP13

Davide Sala#, Yuanpeng Janet Huang#, Casey A. Cole, David Snyder, Gaohua Liu, Yojiro Ishida, G.V.T. Swapna, Kelly P. Brock, Chris Sander, Krzysztof Fidelis, Andriy Kryshtafovych, Masayori Inouye, Roberto Tejero, Homayoun Valafar, Antonio Rosato*, Gaetano T. Montelione*
co-first author
* co-submitting authors

Production and Solution NMR Structure Determination of Target 1008 (foldit3).

The synthetic gene for foldit3 [26] without ACA sequences [27, 28] was obtained from Genscript already incorporated into plasmid pET15TEV_NESG, which includes a N-terminal 6xHis purification tag, followed by a TEV protease cleavage site (sequence 'MGHHHHHHGWSENLYFQGS'). For these NMR studies, this affinity purification tag was not removed. Sample preparation followed standard protocols, as outlined in the previous publication on foldit3 [26]. *E. coli* BL21(DE3) cells harboring plasmid pET15TEV_NESG-foldit3 were grown in 1 L MJ9 minimal media [78], supplemented with 100 μg/ml ampicillin at 37 ºC. In order to produce uniformly ^{15}N and ^{13}C enriched protein samples, 1g / L $^{15}NH_4$-salts and 2g / L U-^{13}C glucose were added as sole a nitrogen and a carbon sources, respectively. When $O.D._{600}$ reached around 0.5 units, the culture was transferred to 18 ºC, and the protein production was induced by addition of 1 mM IPTG. After overnight incubation, the cells were collected and resuspended in 20 ml binding buffer, containing 20 mM Tris-HCl pH 8.0, 500 mM NaCl and 20 mM imidazole. After passing the cells through a 16,000-17,000 psi French press twice, cell debris were removed by 10,000 rpm for 30 min. The supernatant was further spun down at 40,000 rpm for 1 hr. The obtained supernatant (soluble fraction) was mixed with 1 mL of Ni-resin and incubated at 4 ºC for 1 hr. The non-specific binding proteins were removed by 20 mL binding buffer and washing buffer, containing 20 mM Tris-HCl pH 8.0, 500 mM NaCl and 50 mM imidazole, and the target protein was eluted by 5 mL elution buffer (20 mM Tris-HCl pH 8.0, 500 mM NaCl and 300 mM imidazole). The protein was dialyzed against gel filtration buffer, containing 20 mM Tris-HCl pH 8.0, 100 mM NaCl), overnight, and gel filtration was carried out using AKTA Express purification system with high-load 26/600 Superdex 200 pg column. Homogeneity (> 97%) was validated by SDS polyacrylamide gel electrophoresis. The purified protein was dialyzed against 20 mM potassium phosphate (pH 6.5), and the protein concentration was adjusted to between 0.3-0.4 mM for NMR studies.

Supplementary Table S1. Statistics for Real and Simulated NMR Data for NMR-Assisted CASP13 Targets

Target	Data available	No. of residues	Assessment Units	No. of peaks in the final list	No. of peaks not assignable	No. of peaks per residue	No. of peaks removed	No. of residues for which resonance assignments were deleted before NOESY simulation	No. of residues for which resonance assignments were deleted after NOESY simulation
N1008	Only backbone resonance assignments, dihedrals	80	N1008	665	163 (19.7%)	8.4	N/A	0	0
n1008	Essentially complete resonance assignments. dihedrals	80	n1008-D1	3422	6 (0.2%)	43.3	N/A	0	0
N1005	simNOE, dihedrals, 2x RDC's	326	N1005	4367	342 (7.3%)	12.6	245 (5.2%)	83	46
N0980s1	simNOE, dihedrals, 2x RDC's	105	N0980s1	623	41 (6.2%)	5.1	89 (13.4%)	24	15
N0989	simNOE, dihedrals, 2x RDC's	246 132 134	N0989-D1.D2 N0989-D1 N0989-D2	1407	119 (7.8%)	9.3	157 (10.3%)	56	35
N0981-D1	simNOE, dihedrals, 2x RDC's	86	N0989-D1	349	31 (8.2%)	3.5	48 (12.6%)	28	14
N0981-D2	simNOE, dihedrals, 2x RDC's	80	N0989-D2	359	36 (9.1%)	3.6	70 (17.7%)	28	14
N0981-D3	simNOE, dihedrals, 2x RDC's	203	N0981-D3	1186	106 (8.2%)	5.1	155 (12.0%)	50	36
N0981-D4	simNOE, dihedrals, 2x RDC's	111	N0981-D4	553	41 (6.9%)	4.4	68 (11.4%)	29	18
N0981-D5	simNOE, dihedrals, 2x RDC's	127	N0981-D5	698	59 (7.8%)	4.7	97 (12.8%)	30	19
N0968s2	simNOE, dihedrals, 2x RDC's	116	N0968s2	592	41 (6.5%)	4.5	67 (10.6%)	30	18
N0968s1	simNOE, dihedrals, 2x RDC's	123	N0968s1	751	52 (6.5%)	5.4	83 (10.3%)	32	20
N0957s1	simNOE, dihedrals, 2x RDC's	163 108 54	N0957-D1.D2 N0957-D1 N0957-D2	1123	105 (8.6%)	5.9	165 (13.4%)	40	20

Data (real or simulated) provided for each target are listed are listed in the second column. NOESY peaks which cannot be accounted by combined analysis of the chemical shift list and the coordinates of the reference structure are not assignable. The number of residues deleted either before simulating the NOESY peak list, or after simulating the NOESY peak list, are reported in the last two columns, respectively.

Supplementary Table S2. Correlation Coefficients Between Various CASP13 Metrics.

	GDT_HA	GDT_SC	RPF	SphGrdr	CAD_AA	MolPrbty
GDT_HA		0.959	0.923	0.907	0.929	0.518
GDT_SC	0.952		0.902	0.891	0.937	0.521
RPF	0.918	0.902		0.952	0.969	0.557
SphGrdr	0.901	0.895	0.947		0.927	0.555
CAD_AA	0.915	0.932	0.966	0.920		0.588
MolPrbty	0.546	0.554	0.573	0.562	0.610	

Friedman's Test indicates, aside from the MolProbity packing metric, different scoring techniques do not give significantly different rankings. Upper right – Pearson coefficient.
Lower left – Spearman coefficient.

Supplementary Table S3. Principal Component Analysis of Key Structure Assessment Metrics

Component	GDT_HA	GDT_SC	RPF	SphGrdr	CAD_AA	MolPrbty	% Variance Explained
1	0.442	0.449	0.425	0.428	0.433	0.227	86.702
2	-0.146	-0.188	-0.067	-0.056	-0.040	0.966	8.351
3	-0.388	-0.562	0.389	0.608	0.050	-0.104	2.511
4	-0.371	-0.034	0.373	-0.567	0.632	-0.044	1.331
5	0.655	-0.548	0.380	-0.319	-0.156	-0.007	0.800
6	0.256	-0.383	-0.616	0.145	0.621	-0.044	0.306

Supplemental Table S4: NMR Data and Refinement Statistics for Foldit3 [26]

Summary of conformationally-restricting experimental restraints [a]

NOE-based distance restraints:	
Total	1725
intra-residue [i = j]	448
sequential [| i - j | = 1]	441
medium range [1 < | i - j | < 5]	344
long range [| i - j | ≥ 5]	492
NOE restraints per restrained residue [b]	21.3
Hydrogen bond restraints:	
Total	66
long range [| i - j | ≥ 5]	22
Dihedral-angle restraints:	118
Total number of restricting restraints [b]	1909
Total number of restricting restraints per restrained residue [b]	23.6
Restricting long-range restraints per restrained residue [b]	6.3
Total structures computed	100
Number of structures used	20

Residual constraint violations [a,c]

Distance violations / structure	
0.1 - 0.2 Å	10.05
0.2 - 0.5 Å	2.35
> 0.5 Å	0
RMS of distance violation / restraint	0.01 Å
Maximum distance violation [d]	0.42 Å
Dihedral angle violations / structure	
1 - 10 °	17
> 10 °	0
RMS of dihedral angle violation / restraint	1.12 °
Maximum dihedral angle violation [d]	8.40 °

RPF scores

Recall	Precision	F-measure	DP-score
0.945	0.956	0.95	0.842

RMSD Values

	all	ordered[e]	Selected[f]
All backbone atoms	6.9 Å	0.6 Å	0.6 Å
All heavy atoms	7.7 Å	1.1 Å	1.1 Å

Structure Quality Factors

	Mean score	SD	Z-score [g]

Procheck G-factor [e] (phi / psi only)	-0.18	N/A	-0.39
Procheck G-factor [e] (all dihedral angles)	-0.21	N/A	-1.24
Verify3D	0.25	0.0282	-3.37
ProsaII (-ve)	0.87	0.0739	0.91
MolProbity clashscore	4.97	2.6461	0.67

General linear model RMSD prediction 1.14 Å

Ramachandran Plot Summary from Procheck [f]

Most favored regions	94.4%
Additionally allowed regions	5.5%
Generously allowed regions	0.1%
Disallowed regions	0.0%

Ramachandran Plot Statistics from Richardson's lab

Most favored regions	97.3%
Allowed regions	2.5%
Disallowed regions	0.1%

---

[a] Analyzed for residues 1 to 97, Including N-terminal purification tag.
[b] There are 81 residues with conformationally-restricting restraints.
[c] Calculated for all restraints for the given residues, using sum over r^{-6}
[d] Largest restraint violation among all the reported structures.
[e] Residues with sum of phi and psi order parameters > 1.8.

Ordered residue ranges: 21A-45A,48A-54A,57A-78A,80A-87A,90A-96A
[f] Residues selected based on: dihedral angle order parameter, with $S(\phi)+S(\psi) >= 1.8$

Selected residue ranges: 21A-45A,48A-54A,57A-78A,80A-87A,90A-96A

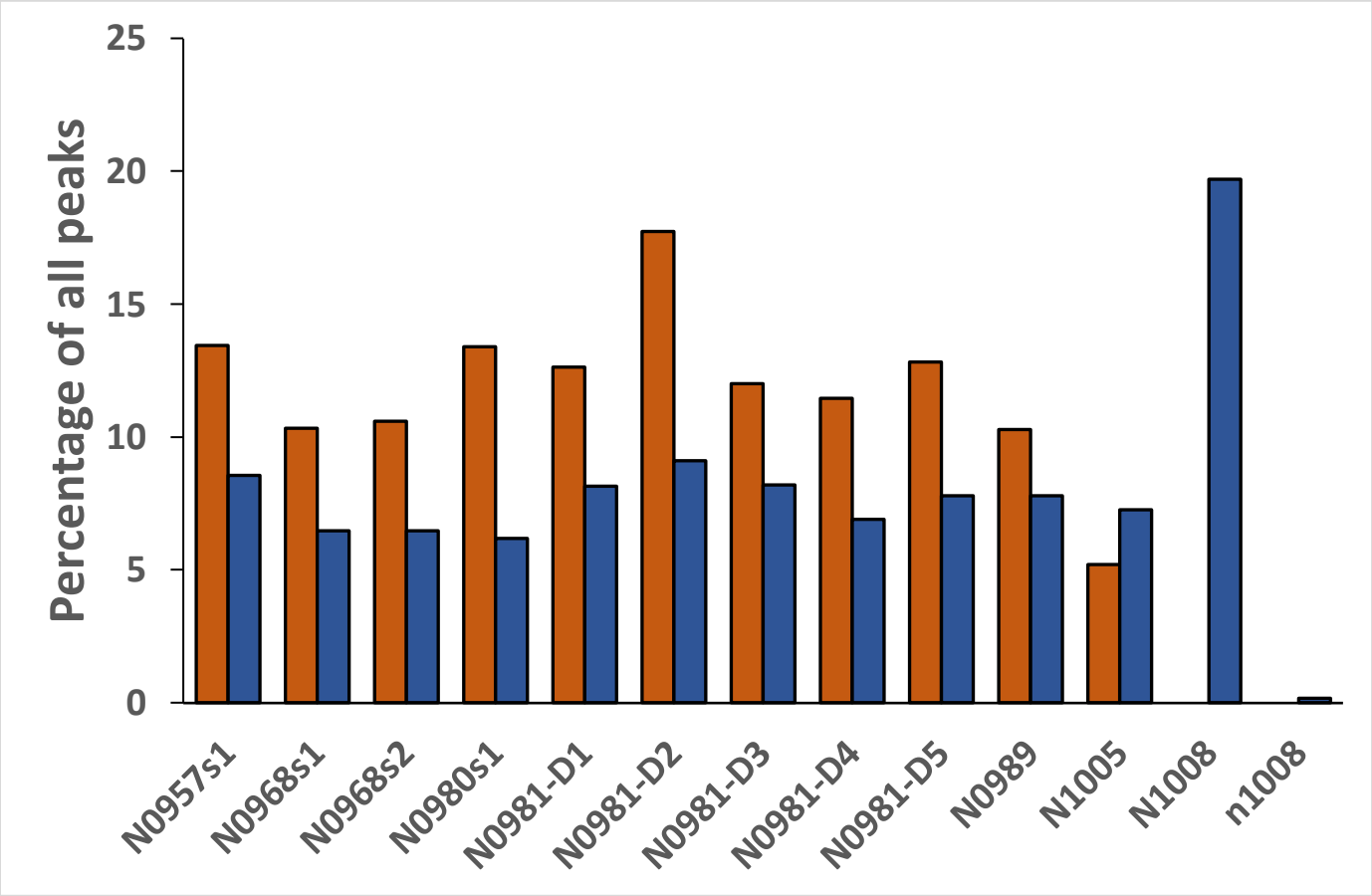
[g] With respect to mean and standard deviation for for a set of 252 X-ray structures < 500 residues, of resolution <= 1.80 Å, R-factor <= 0.25 and R-free <= 0.28; a positive value indicates a 'better' score

Generated using PSVS 1.5

Supplementary Table S5. Assessment of Contact Ambiguity

	No. of Residues	No. of Possible Contacts	Average Ambiguity per Contact	Maximum Ambiguity per Contact	Unique Long-range Contacts Total / Per Residue	Unique Long-range H^N-H^N Contacts Total / Per Residue
Simulated NMR Data						
N0957s1	163	5582	5	50	110 / 0.67	39 / 0.24
N0968s1	123	1506	2	16	138 / 1.12	29 / 0.24
N0968s2	115	2088	4	32	93 / 0.81	51 / 0.44
N0980s1	105	1489	3	18	92 / 0.88	34 / 0.32
N0981-D1	86	538	2	10	126 / 1.47	44 / 0.51
N0981-D2	80	504	2	8	127 / 1.59	60 / 0.75
N0981-D3	203	4701	4	32	193 / 0.95	67 / 0.33
N0981-D4	111	1093	2	10	100 / 0.90	49 / 0.44
N0981-D5	127	1983	3	21	135 / 1.06	74 / 0.58
N0989	246	7095	5	90	200 / 0.81	91 / 0.37
N1005	326	49,887	11	92	263 / 0.81	90 / 0.28
Real NMR Data						
N1008	97[a]		5	54	53 / 0.54	27 / 0.28
n1008	97[a]		9	169	200 / 2.06	19 / 0.20

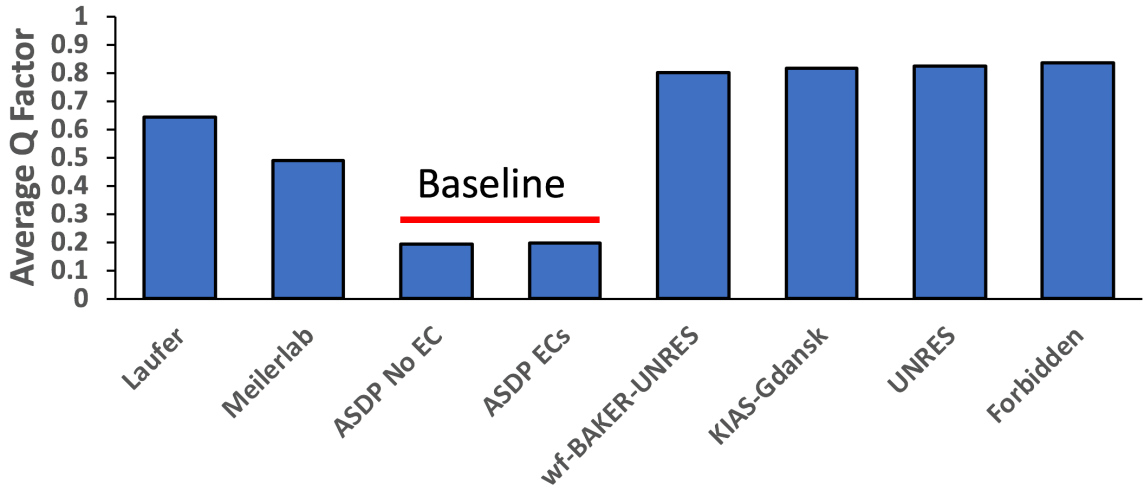
[a]Statistics for target 1008 include the N-terminal 17-residue polypeptide tail.



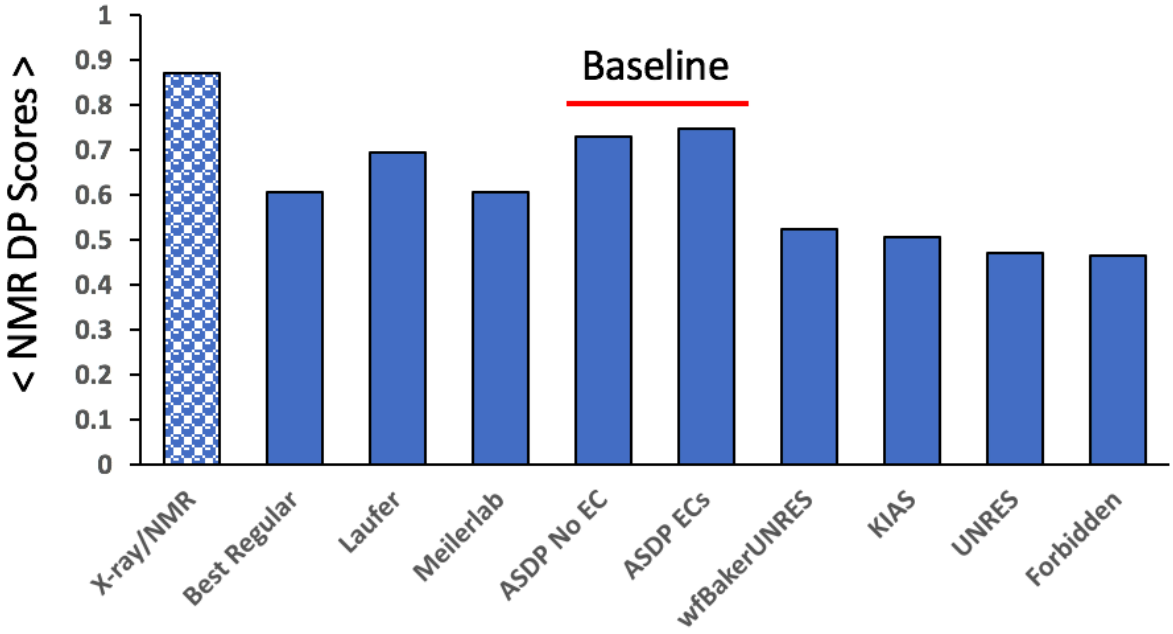
Supplementary Fig. S1. Analysis of NOESY peak lists against the reference atomic coordinates. Orange bars – percentage of all possible NOESY peaks that are removed by simulated deletions of "exchange broadened" resonances from the resonance assignment list. Blue bars – Percentage of all NOESY peaks in the real or simulated spectra that cannot be correctly assigned based on the information provided in the Ambiguous Contact Lists. "Unassignable peaks" arise either from noise peaks, which do not correspond to a true NOE interaction, or for real NOESY peaks when the true resonance that gives rise the cross peak is not assigned in the chemical shift list, leading to erroneous assignments of the NOESY cross peak. This problem is particularly severe for data set N1008 in which many sidechain-backbone NOEs are present in the NOESY peak list, but no sidechain assignments are available in the chemical shift list.

Residue 1	Residue 2	Peak No.	Upper-bound		Atom 1	Atom 2
R1	**R2**	**P#**	**UPL**	**Confid**	**A1**	**A2**
79	77	17	5.0	0.95	H	H
79	177	20	6.0	0.67	H	HD2
79	135	20	6.0	0.97	H	HD1
79	249	20	6.0	0.96	H	HD1
79	50	20	6.0	0.81	H	HD2
79	217	23	5.0	0.68	H	H
79	230	23	5.0	0.75	H	H
79	232	23	5.0	0.72	H	H
79	106	23	5.0	0.76	H	H
79	166	23	5.0	0.83	H	H
79	100	23	5.0	0.83	H	H
79	82	23	5.0	0.74	H	H
79	246	23	5.0	0.71	H	H
79	216	23	5.0	0.67	H	H
45	37	28	7.5	0.84	HD2	HG1

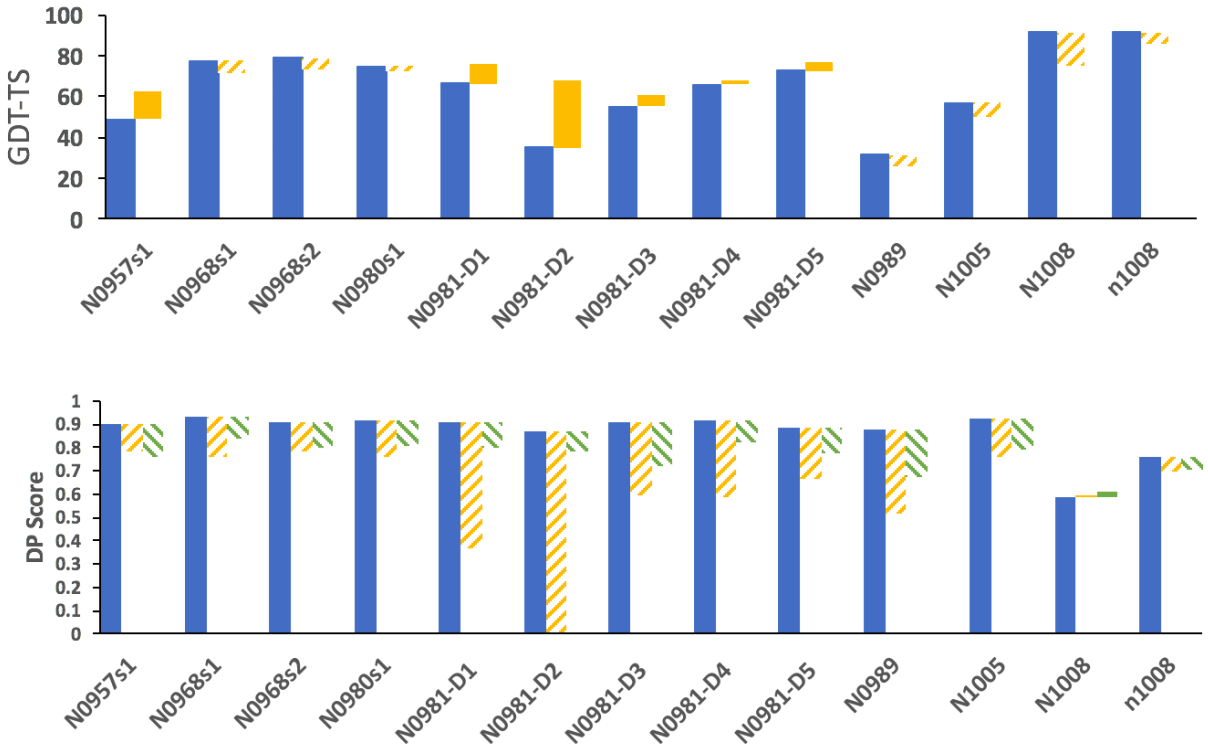
Supplementary Fig. S2. Format of Ambiguous Contact Lists. These data were provided in place of NOESY peak list data to CASP13 predictors. For each peak in the ^{15}N-edited or ^{13}C-edited 3D NOESY peak list (column P#), a set of ambiguous contacts were determined based on the simulated chemical shift list, using the Cycle 0 protocol of the NOESY peak assignment program *ASDP*. Possible contacts are listed between H atom 1 (Residue number R1, and Atom A1), and H atom 2 (Residue number R2, and Atom A2), together with an upper bound distance (UPL) in Å. Early Ambiguous Contact Lists included an assignment confidence score (Confid) ranging from 0 to 1, based on the quality of the match between the chemical shift values of the NOESY peak and the chemical shift values of candidate interacting atoms in the resonance assignment list. Since the Confid score was not used in CASP11, it was phased out of use during CASP13. Atom types include amide H^N protons (H) and various methyl proton groups (HB, HG1, HG2, HD1, HD2, etc). In this example, ASDP has uniquely assigned peak P# 20 to an interaction between the amide H^N of residue 77 and the amide H^N of residue 79, while peak P# 20 has four ambiguous assignments, H^N of residue 79 and methyl resonances of residues 177, 135, 249, and 50



Supplemental Fig. S3. Average ^{15}N-^{1}H RDC Q-Factors, averaged over submitted evaluation units, for each predictor group.



Supplemental Fig. S4. NMR DP scores, averaged over submitted first-ranked models, for each predictor group.



Supplementary Fig. S5. **NMR DP Scores for X-ray Crystal Structures, Experimental NMR Structures Compared with Best Regular or Best NMR-Assisted Models**. Top (same data as main text Fig 7A): GDT-TS scores for the "best" model submitted by any NMR-assisted prediction group (blue bars) compared with the "best" model submitted by any regular prediction group (yellow solid bars show improved accuracy, and hashed yellow bars show average accuracy, due to addition of sparse NMR data). Bottom: DP scores for experimental structures determined by X-ray or NMR (blue bars) compared with the "best" model submitted by any regular prediction group (yellow bars) and the "best" model submitted by any NMR-assisted prediction group (green bars). Hashed yellow or green bars indicate targets are less accurate than the experimental structures, while solid yellow or green bars indicate targets are more accurate.