SUPPORTING INFORMATION FOR "NMR-ASSISTED PROTEIN STRUCTURE PREDICTION WITH MELDXMD"

Post-CASP simulations

Our typical MELD protocol has 30 replicas. Every 50 ps we attempt exchanges between replicas and we run for about 1,000,000 ps (1μs). We typically set our replica exchange success at about 30-40 per cent, and expect that each walker will likely sample from a wide set of conditions before the end of the simulation. This ensures that the replica exchange protocol is converged and we are getting the full advantage of this advanced sampling technique. If some walkers remain at high replica indexes without ever sampling low temperatures, the benefits of sampling of these replicas does not make it into the lower ones.

During our participation in the last part of CASP, we were unaware of a bug that got introduced into our sampling protocol affecting our refinement and NMR predictions. We identified the bug after the CASP competition ended. The bug affected replica exchange probabilities, and hence, sampling efficiencies in REMD. In order to increase exchange frequencies during CASP, this necessitated simulations with many more replicas (100) than we typically use. The bug has been recently corrected, and our standard protocol with 30 replicas running for 1μs was tried out after the fix for comparison purposes. Additionally, and for consistency with CASP simulations, runs with 100 replicas were also carried out on selected systems to identify other potential

issues. In both post-CASP simulations, we did a single runs with no re-seeding.

We compare our ability to sample during and after CASP in Fig. S1. Our results show that using 100 or 30 replicas produced structures of similar accuracy – a factor of 3 reduction in computational cost. The second thing it revealed is that the resulting structures were of similar accuracy to the ones produced during CASP. Nevertheless, an advantage in our post-CASP simulations was the automated pipeline in that we no longer needed to stop replicas, analyze the ensembles and seed new replicas with the best structures. The main issue remains the poor REMD exchanges, as detailed below.

| Poor H,T-REMD exchanges

Based on our experience in CASP11 [21], we knew that NMR data could pose problems for efficient replica exchanges. We monitored this closely during CASP, seeing a very inefficient replica exchange rate (see Fig. S7, left panel). As can be seen from the figure, the exchange probability for walkers in neighboring replicas was high enough, but the probability of those walkers progressing to far away replicas was not so good. This was the case even after increasing the number of walkers to 100.

After CASP, we identified the bug that affected the exchanges when using NMR data. We resolved the bug and re-simulated the targets with 30 and 100 replicas and without the need to re-seed. The results from these simulations did not show dramatic improvements compared to CASP results in terms of predicted structures, even though the exchanges were now much better (Fig. S7, right panel). What still hindered overall

sampling in general is the fact that even though exchanges improved and walkers now sampled more diverse set of replica conditions, there were still few round trips between the lowest and highest replicas. In both CASP and post-CASP strategies, the ability to sample correct restraints and obtain near-native ensembles remained similar, however, running the simulations in a more automated framework with 30 replicas poses a significant reduction in computational resources required for structure prediction using MELDxMD.

A major current limitation in MD structure prediction accuracy is backtracking[30]. Some contacts form early on in the simulation, and in order for other NMR contacts to be satisfied and fold the rest of the protein, the chain would need to pass through the formed contacts; something that is not physically possible. We expected that the replica exchange protocol would allow to fold/unfold the protein, or at least parts of it in order to mitigate the backtracking problem. Instead, we see that the abundance of NMR contacts, their cooperativity and redundancy makes it very difficult to unfold regions of the protein that already satisfy many of the restraints, with an overall consequence of walkers not making whole round trips in the replica ladder. Strategies that deal with this issue and favor round trips in such an enhanced sampling protocol are key for future efforts.

Post-CASP simulations with and without evolutionary contact

To assess the quality of the incorporated evolutionary contacts (ECs) that we calculated using jackhmmer and GREMLIN, we performed simulations using no evolutionary contacts and using evolutionary contacts provided by CASP for the NMR targets. For those post-CASP simulations the MELDxMD protocol used the following input:

1. MELD with no evolutionary data: We used secondary structure predictions from PSIPRED, and NMR data (TALOS angles and NOESY peaks);
2. MELD with provided evolutionary data: We used secondary structure predictions from PSIPRED, NMR data (TALOS angles and NOESY peaks) and provided evolutionary information. These provided ECs were filtered to keep the long and medium ranged ones (residue pairs separated by more than 12 amino acids in sequence), and the top N/5 ECs (N=sequence length) were used during MELD simulation.

Figures S9 and S10 show the quality of different ECs and the corresponding MELD prediction. The conclusions are 1) Compared with more accurate provided ECs, the predicted ones hindered MELD predictions (Orange bars of N0968s2-D1 and N0957s1-D2 in Fig. S10), and the results improved when more accurate, provided ECs were used. 2) ECs added little value to MELDxMD when system specific NMR data were used. The advances in generating precise ECs has greatly benefited non-MD

modelers, and we anticipate that they could also help, but we did not see that in these cases.

Case study of sub-domain performances

To understand the variation in MELDxMD prediction quality among sub-domains of a single protein target, we chose N0981 with its five sub-domains. We conducted MD stability tests using the same physics model as in our MELDxMD simulation. Specifically, we performed a MD simulation (100ns long, 300K) of each domain starting from the native structures. The results are shown in Figure S12. It is clear that the native structures of the sub-domains that MELDxMD successfully predicts in CASP13 (N0981-D1,N0981-D4,N0981-D5) are stable in the force field and solvent model used, while the ones for which predictions with MELDxMD was unsuccessful (N0981-D2,N0981-D3) are unstable. This indicates that force field bias is a potential reason for performance variation in independent predictions of sub-domain structures.

In an orthogonal analysis, we checked the density of true contacts from the list of data processed and submitted to MELDxMD. Figure S13 shows the relative sizes of each sub-domain of N0981 and the distribution of the contacts in each sub-domain. The data list for most domains contains both local and non-local contacts, except for domain N0981-D2 that shows mostly local contacts, adding to the challenge of folding it in MELDxMD. Domain N0981-D3 was the largest with 203 amino acids, but with good non-local density of contacts in the input data set. However, as discussed above,

native structure of this domain is unstable in the force field. In addition, the D2 and D3 domains of N0981 had worse templates from server predictions (GDT_TS values below 60%) amongst the five sub-domains. A combination of some or all of force field bias, starting templates and locality of correct experimental contacts contributed to weak performance of MELDxMD on certain sub-domains of the same target protein.

| Could MELDxMD work with extended chains?

We addressed this topic in the main CASP competition – in the absence of NMR data. We signed up three different predictor groups (Laufer-100, Laufer-ab initio and Laufer) to test the trade-offs between sampling, force fields and data in folding proteins. In all three cases we used the same force field and solvent combination as described in methods. For Laufer-100 and Laufer-ab initio we started from extended chains as produced by AMBER's leap module[31], while Laufer (same protocol as in CASP-NMR) started from 15 template predictions (see methods). Laufer-100 used general heuristics[5] (hydrophobic residues pack at the core and β-strands pair up); this method has only been successful in targets up to 100 residues in length. The other two protocols included co-evolutionary data (see methods). Since they all use the same force field, they should all converge on structures favored by it. If the force field favors the native state, then the three methods should converge on it. Figure S14 shows results for four targets we attempted under 100 residues and were successful with at least one protocol. In all cases, starting from server predictions (cyan line) was

the most successful approach. For one target, T0955-D1 (bottom right), all protocols converged on the native structure as identified by the top cluster – showing that when sampling is long enough, the structures converge on what the force field prefers. Target 974 (top left) shows that both Laufer-ab initio and Laufer converge on the same structure: the added data (co-evolutionary in this case) helped to converge faster than regular heuristics and the correct structure was identified regardless of starting from an extended or server prediction. Finally, T1019s2-D1 (top right) and T0958 (bottom left) only identify the native state when starting from server predictions – showing that longer sampling times were needed in the other protocols. Figure S15 shows that when considering the five structures submitted to CASP, some of the Laufer-100 and Laufer-ab initio predictions are shifted towards higher GDT scores (especially for the ab initio protocol in T1019s2-D1). This highlights the need of longer sampling times to improve convergence. In conclusion, starting from structures that have some native like characteristics improves convergence and reduces the amount of sampling needed – but, using the same force field and sources of data, and with enough sampling time, clusters will likely converge onto the same structure.

| NMR data processing scripts

The scripts used in processing the NMR data for MELDxMD simulation setup are freely available at `https://github.com/laufercenter/CASP-NMR`

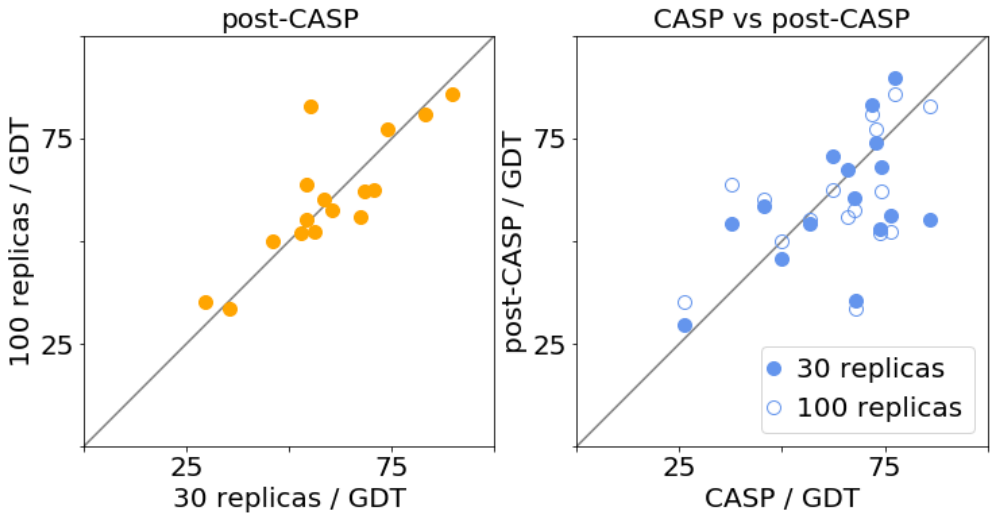

FIGURE S1 **MELD ability to predict structures improves when NMR data is available.** On the left we present the comparison between our two post-CASP protocols. The 30 and 100 replicas protocols have similar performances in predicting the quality of the structure. In the right panel we show how our two post-CASP protocols (full and empty circle identify the 30 and 100 replicas protocols respectively) compare to our CASP submission. While some scattering is present our CASP submission is representative of the quality of our predictions.

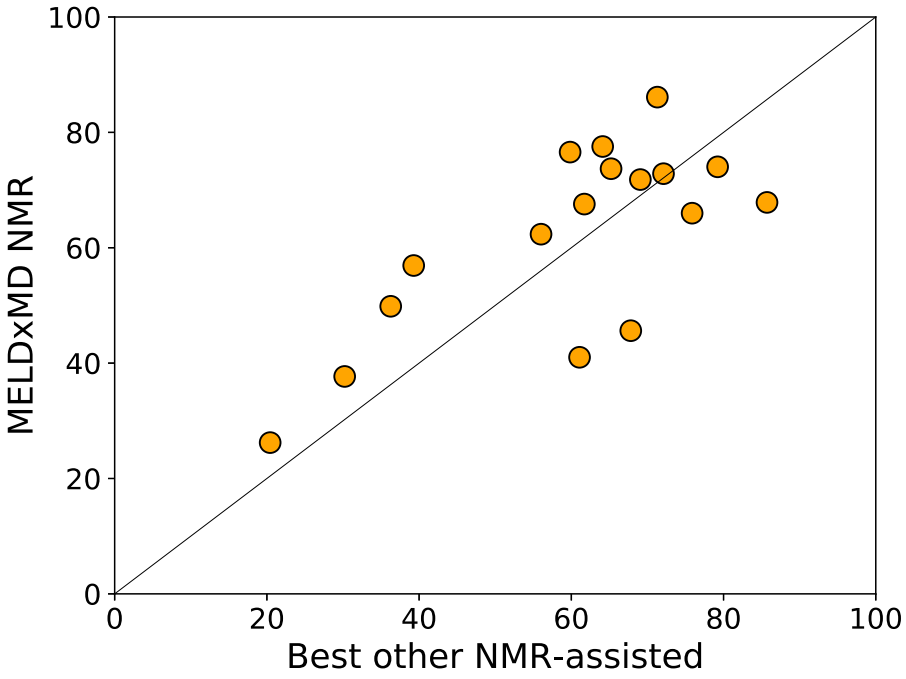

FIGURE S2 **MELDxMD GDT_TS outscores the best models of other predictors in the NMR assisted category of CASP13**. The plot shows a comparison between the best GDT_TS model from the MELDxMD NMR predictions and the best GDT_TS model from any other submission in the NMR category for each of the 17 target domains. MELDxMD performs best in 12 out of the 17 targets (points above the diagonal line).

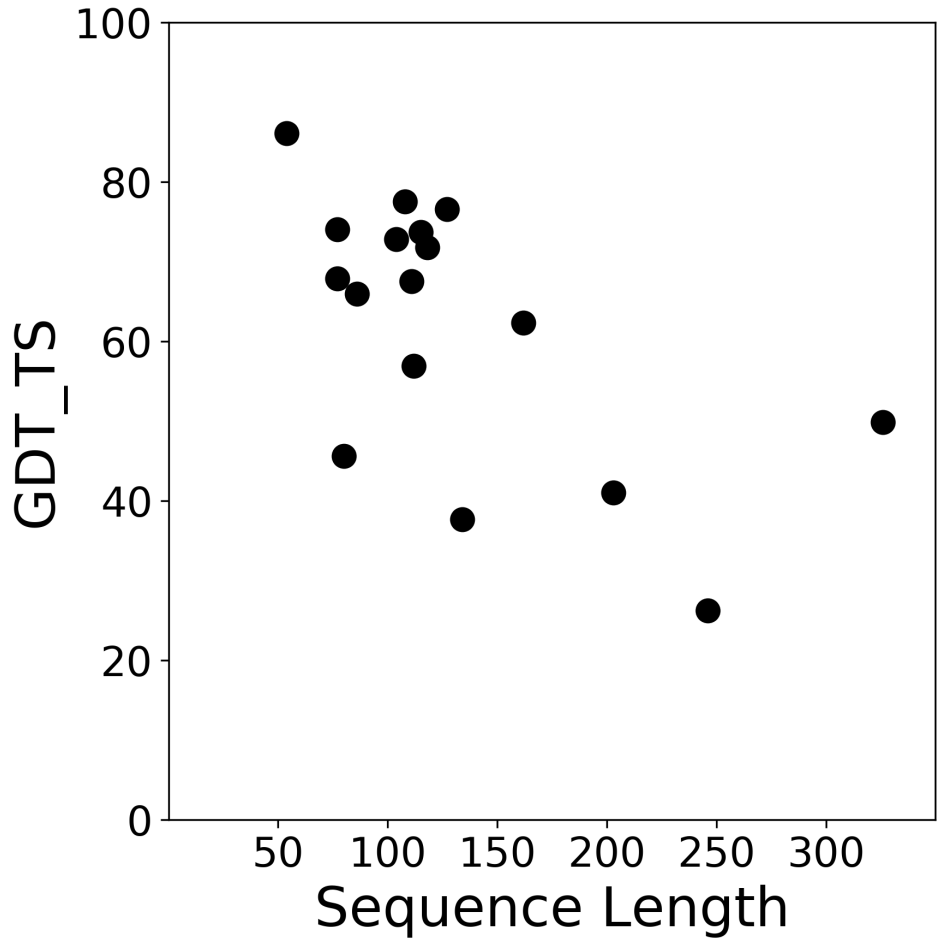

FIGURE S3 **GDT_TS scores were independent of protein size.** The GDT_TS are shown for 17 assessment units modeled by MELDxMD with NMR data. Systems range from 54 to 326 residues and there was little relationship between sequence length and prediction accuracy.

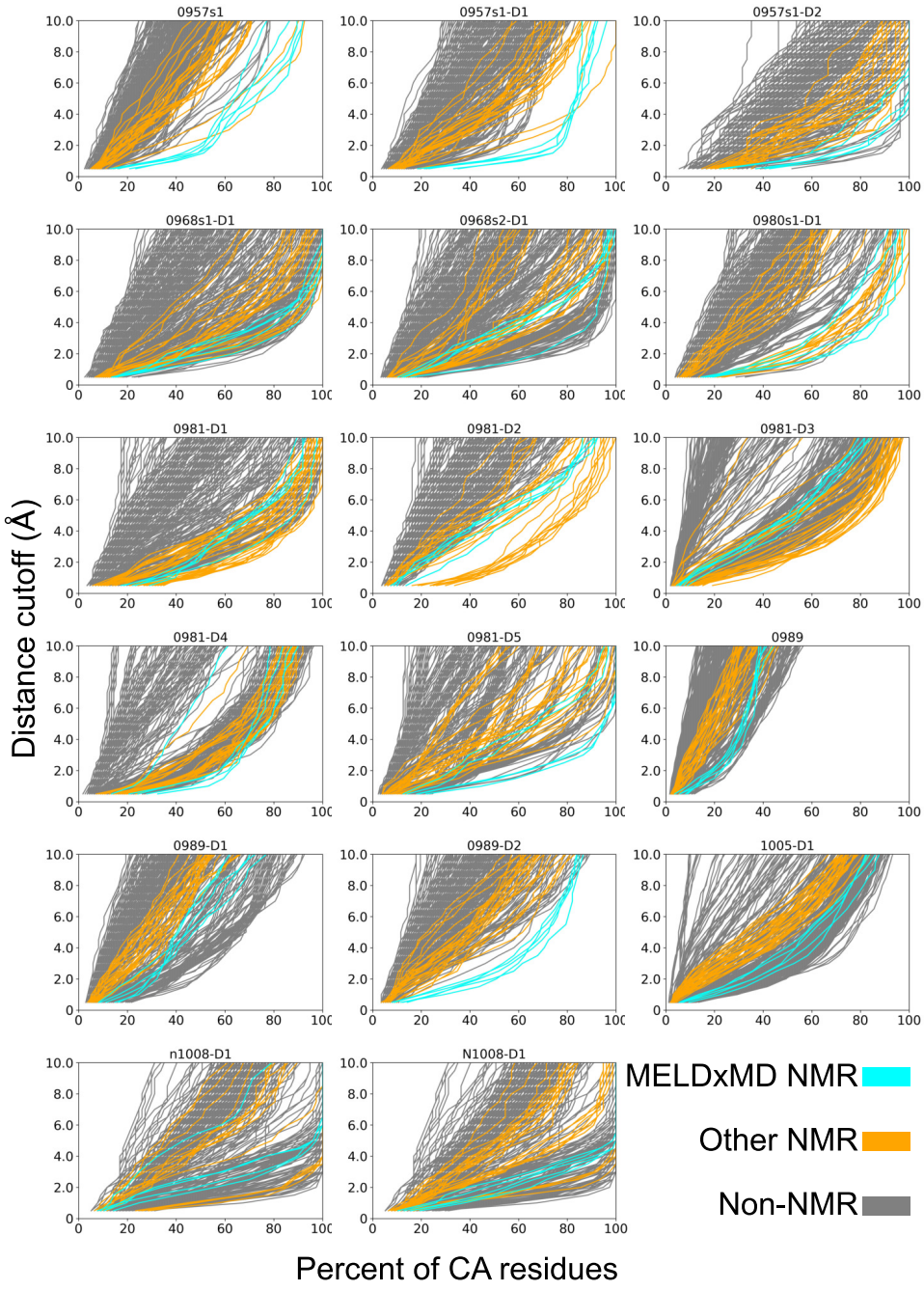

FIGURE S4 **GDT plots show the performance of MELDxMD (Laufer group 431) on CASP13 targets.** The GDT plots of 17 evaluation units from the NMR data-assisted category are shown. The MELDxMD models are highlighted in cyan, the other NMR predictors are shown in orange, and the non-data-assisted human and server predictions from the main CASP13 category are shown in grey. All 5 models are shown for all predictors.

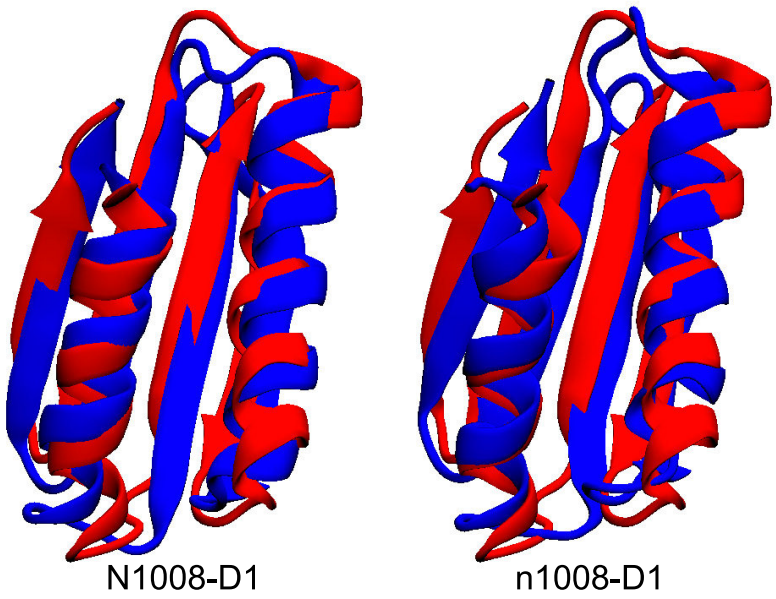

FIGURE S5 **MELDxMD made accurate predictions from real NMR data.** The 1008-D1 target was provided with two sets of NMR data, the N1008-D1 data more noisy than n1008-D1. In both cases, MELDxMD handled the data and predicted structures with GDT_TS scores of 74.03 (N1008-D1) and 67.86 (n1008-D1). Reference structure in red, MELDxMD prediction in blue.

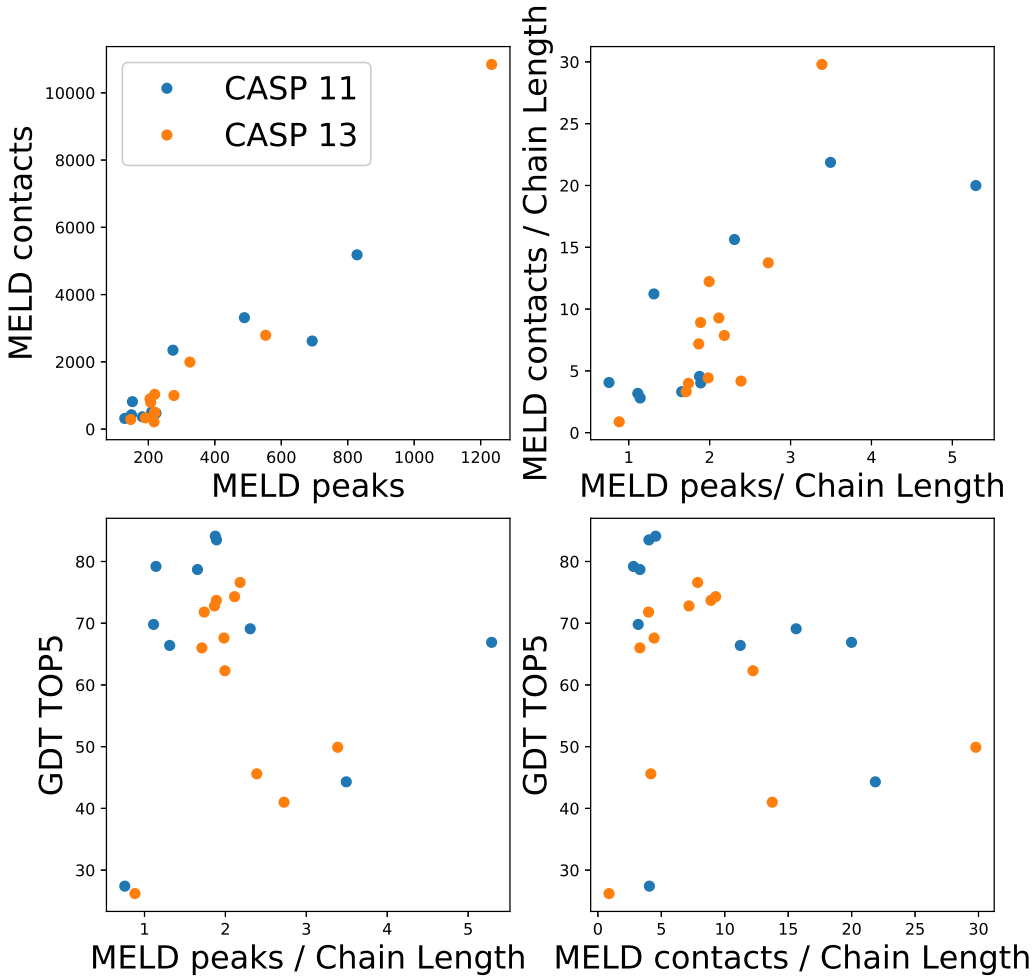

FIGURE S6 **The ambiguity in data we extract from the NMR NOE peaks is about the same in CASP 11 and 13, and MELD performances are comparable between the two CASPS.** The top two panels give a measure of the ambiguity of the data we put into MELD in CASP 11 (blue dots) and 13 (orange dots). The left panel shows the absolute number of peaks and contacts, while the right pabel is normalized over the chain length. The lower two panels show the MELD performance as function of the normalized number of peaks (left) and contacts (right).

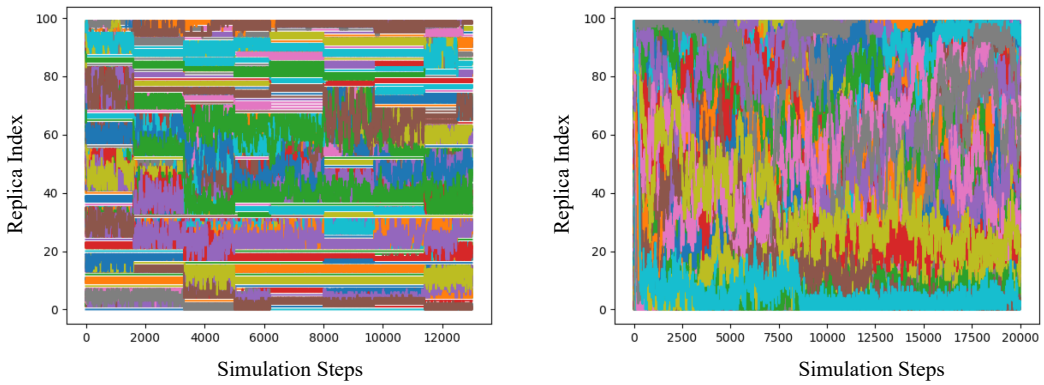

FIGURE S7 **Poor replica exchanges reduced sampling and limited accurate structure prediction.** The replica exchange condition of target N0981-D2 is shown as an example. The left panel shows the poor exchange among replicas due to the software bug. The right one shows the same simulation after the bug-fix. The replica exchange becomes much better: high temperature replicas (bigger replica indices) mixed well with low temperature ones. This indicates better sampling that should lead to improved results, which was the case. The best structure in the left simulation had GDT_TS score of 45.62 while the right one was 63.12.

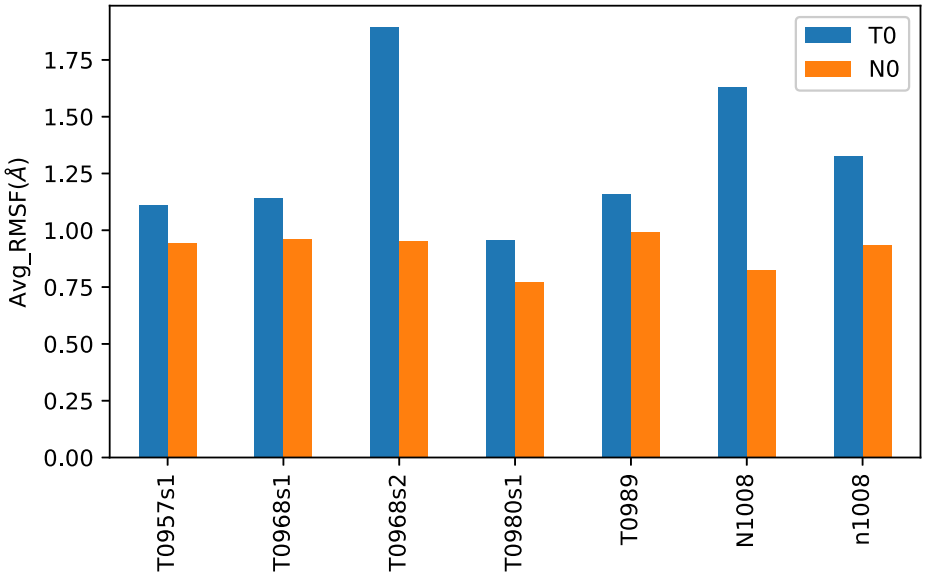

FIGURE S8 **Ensemble fluctuation was lower using only heuristic restraints compared to add NMR restraints**. The MD simulation at 300K with NMR restraints (orange bars) has lower root-mean-square fluctuation (Avg_RMSF) than the the ones with heuristic restraints (blue bars).

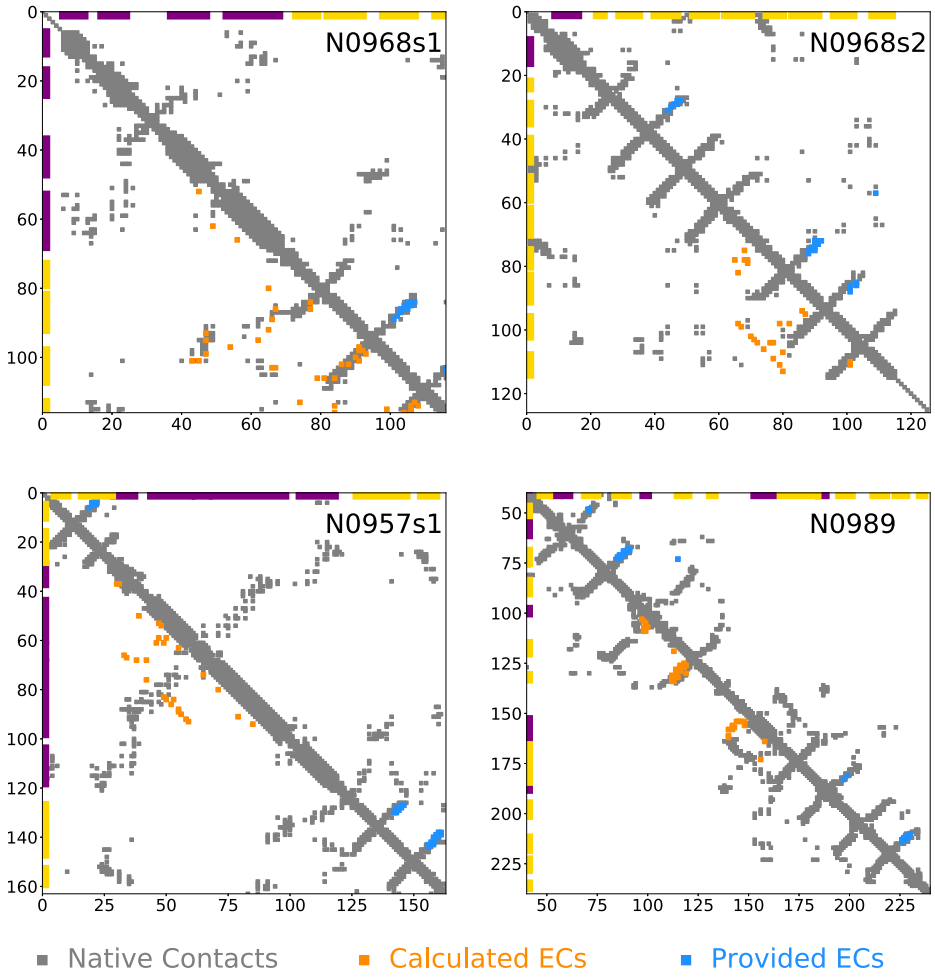

FIGURE S9 **The provided co-evolutionary contacts were more accurate than the ones we used in CASP13.** Contacts below the diagonal in each map are the native contacts (grey) and our predicted contacts using metagenomic sequence libraries and Gremlin (orange). Here, contacts were defined when any two heavy atoms between residues were within 6 Å. Above the diagonal are also native contacts (grey) and the provided evolutionary contacts (blue). In these, contacts were defined if the CB of two residues were within 8 Å. Secondary structure elements are shown along the axes (alpha helices in purple and beta strands in yellow).

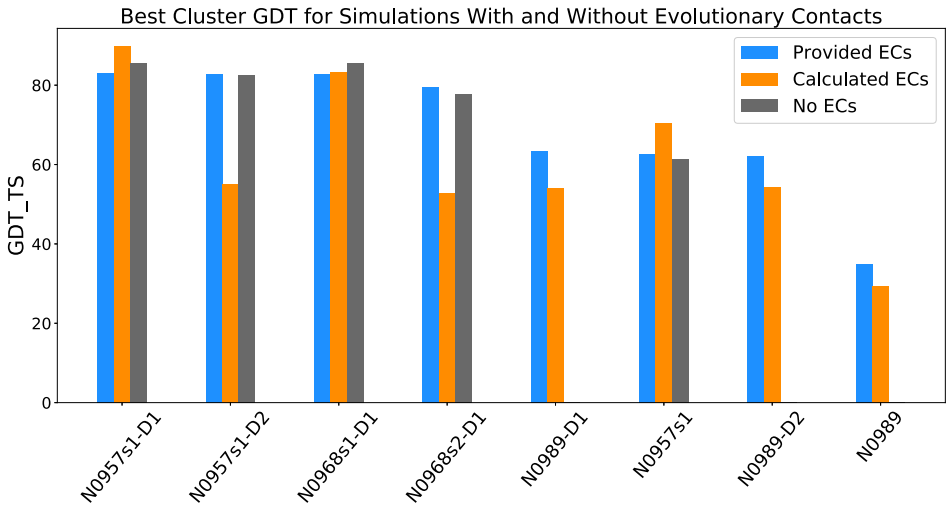

FIGURE S10 **More accurate evolutionary contacts improved MELDxMD structure prediction.** MELDxMD was re-run on 8 systems, post-CASP, to determine the extent that inaccurate evolutionary contacts (ECs) hurt MELDxMD structure prediction. Systems were run with no ECs, with the same ECs used during CASP13, and with ECs provided by CASP that we did not use during the competition. The provided ECs were more accurate and improved GDT_TS for nearly every target. However, the provided ECs did not outperform the absence of ECs, as expected.

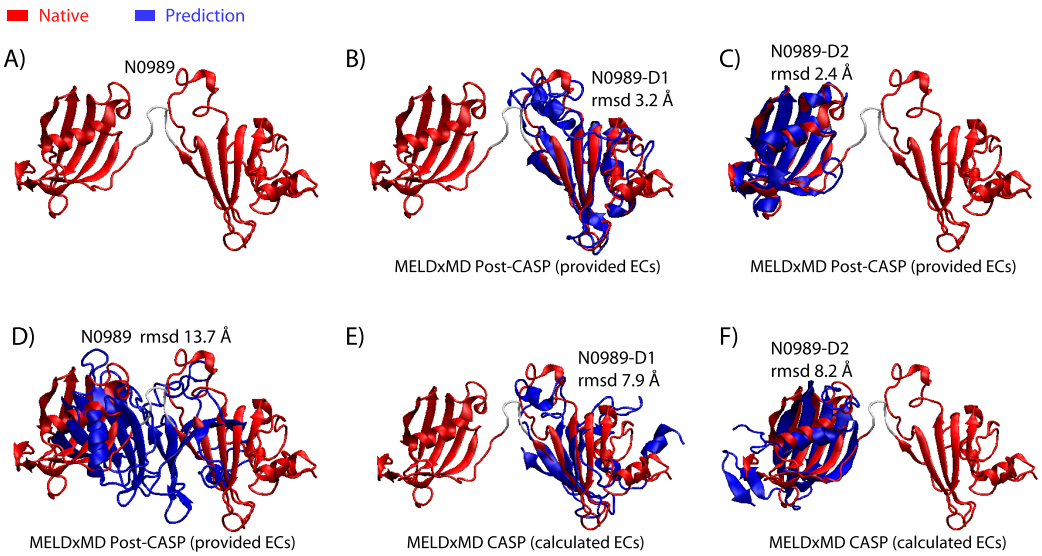

FIGURE S11 **The independent domains of N0989 were modeled well, and improved with better evolutionary contacts, but the full protein was a challenging target.** A) Native structure of N0989, with domain 2 on the left, domain 1 on the right, separated by a flexible linker (white; residues 135-140). B,C,D) Simulations performed post-CASP using NMR data and provided ECs. The lowest RMSD structures from the low temperature trajectories are shown aligned to (B) domain 1 residues 25-134 (the N-terminus segment 1-25 was excluded from the alignment and RMSD calculations as it was flexible during the simulations), (C) domain 2 residues 141 to 246, (D) both domains. When both domains are considered, the protein structure is much more compact than native, likely due to force field and solvation model artifacts. E,F) Structures submitted to CASP that used NMR data and in-house ECs. Alignment and lowest RMSD structures of domain 1 (E) and domain 2 (F).

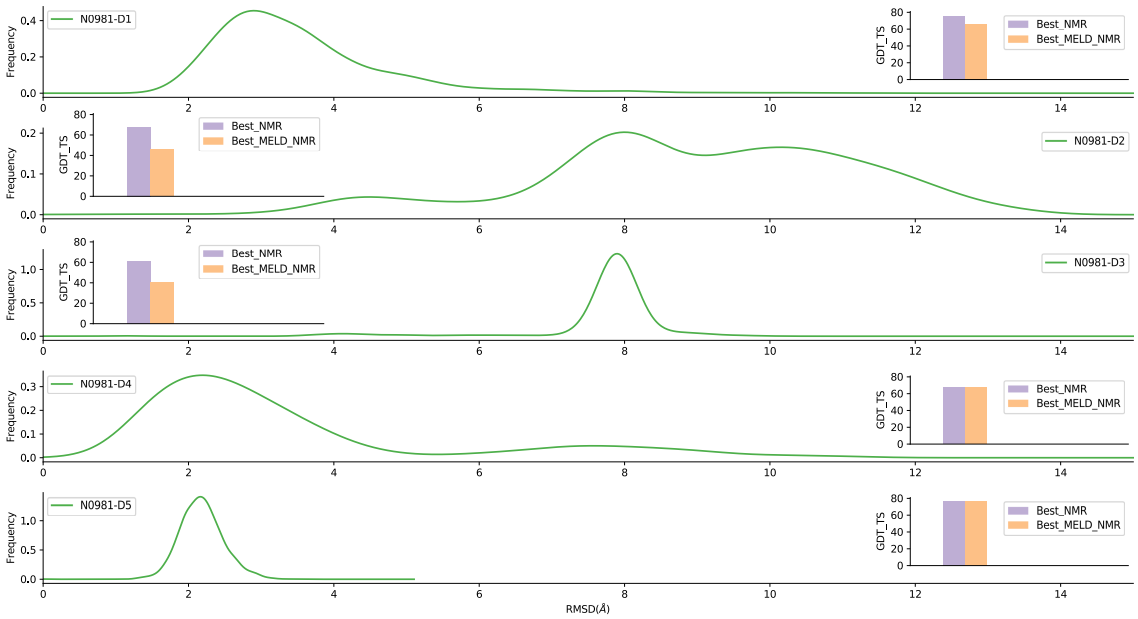

FIGURE S12 **MELDxMD tended to make accurate prediction on domains that were stable in the force field.** The plots show RMSD distributions from MD stability tests, and include CASP GDT_TS scores of MELDxMD compared to the best NMR prediction for each sub-domain of N0981. Stability tests started from native sub-domain structure and ran at 300K for 100ns. The distribution of CA_RMSD is depicted as green lines. Inset bars show the best NMR-assisted and best MELDxMD prediction GDT_TS in purple and orange, respectively. For the 3 domains that were stable (D1, D4, and D5), MELDxMD did well, including making the best predictions on D4 and D5. Domains that did not stay near native in the stability tests (D2 and D3), MELDxMD did a relativley poor job of predicting their structures in CASP13. The implication is that force field errors contributed to poor MELDxMD predictions of D2 and D3.

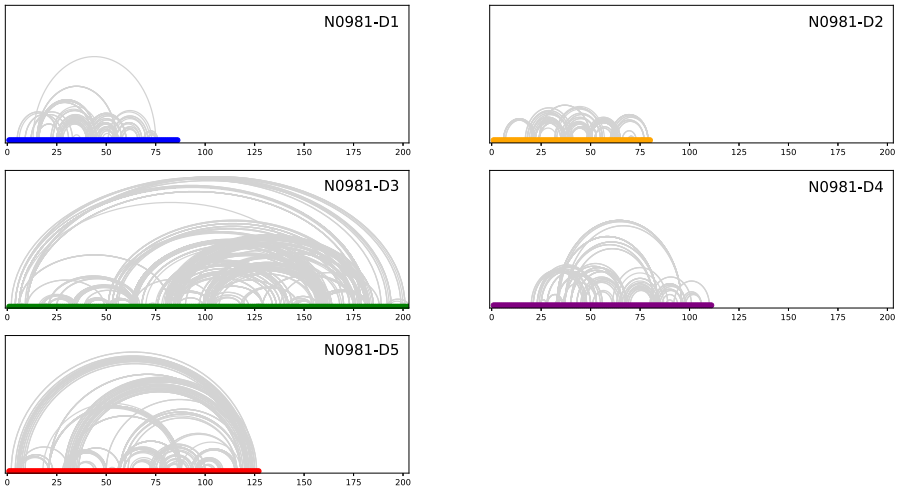

FIGURE S13 **The filtered NOESY-derived contacts used in MELDxMD simulations contained correct local and non-local contacts for all sub-domains, with D2 showing the least amount of non-local contacts**. The pairwise contacts plotted here are the ones present in the NMR input data and in the NMR structure (true positives) as analyzed Post-CASP. The x-axes were scaled to show relative domain size. Amongst all sub-domains, D3 and D5 contained the highest number of non-local contacts in the NMR data used in MELDxMD simulations. However, MELDxMD could not successfully predict the structure of domain D3 as it was heavily influenced by the force field favoring non-native structures (see SI Fig. 11). Domain D2 contained mostly local contacts and was not predicted successfully by MELDxMD. Sub-domains D1, D4 and D5 contained both local and non-local contacts and were also favored by the force field allowing for accurate structure predictions in MELDxMD simulations for those sub-domains.

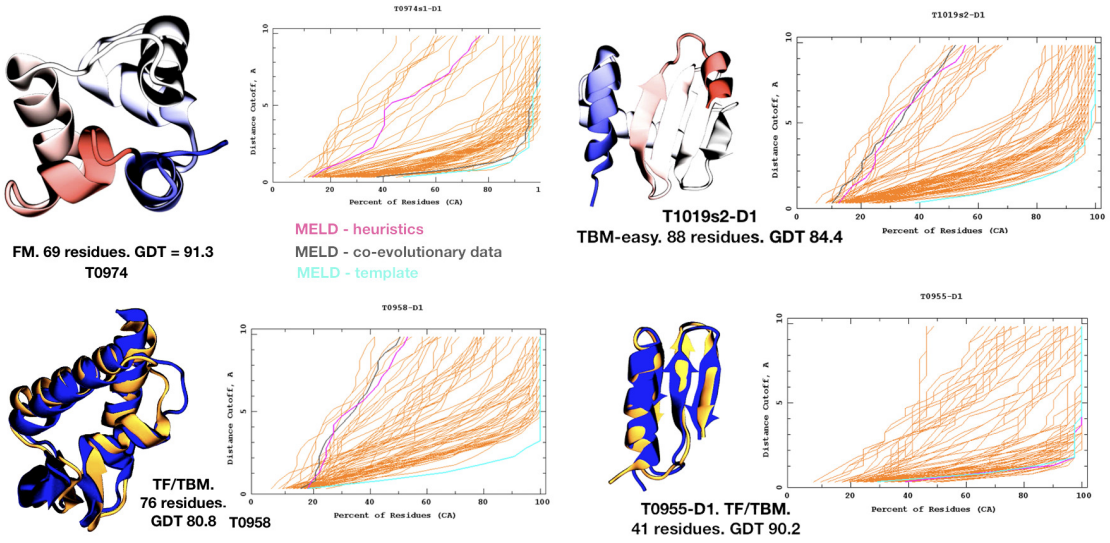

FIGURE S14 **Sampling and force fields influence MELD convergence on the native structure**. In the main portion of CASP we ran some proteins with three different protocols. The figure shows for proteins for which at least one method identified the native state. The GDT plots shows the first model from each group in the T0 category, and highlights three different MELD protocols: using heuristics and starting from extended conformations (pink), adding co-evolutionary data and starting from extended (grey) and using co-evolutionary data and seeding from 15 server predictions (cyan, protocol used for NMR data as well). For two targets the pdb has not been released and our top model is shown blue (N-termini) to red (C-termini). For the other two models (bottom) the blue structure is the experimental one and the orange structure is our model. When sampling is good all protocols converge on the same structure (e.g. T0955). Using co-evolutionary data can help over just using heuristics (T0974). And starting from server predictions accelerates identifying the right answer in all cases.

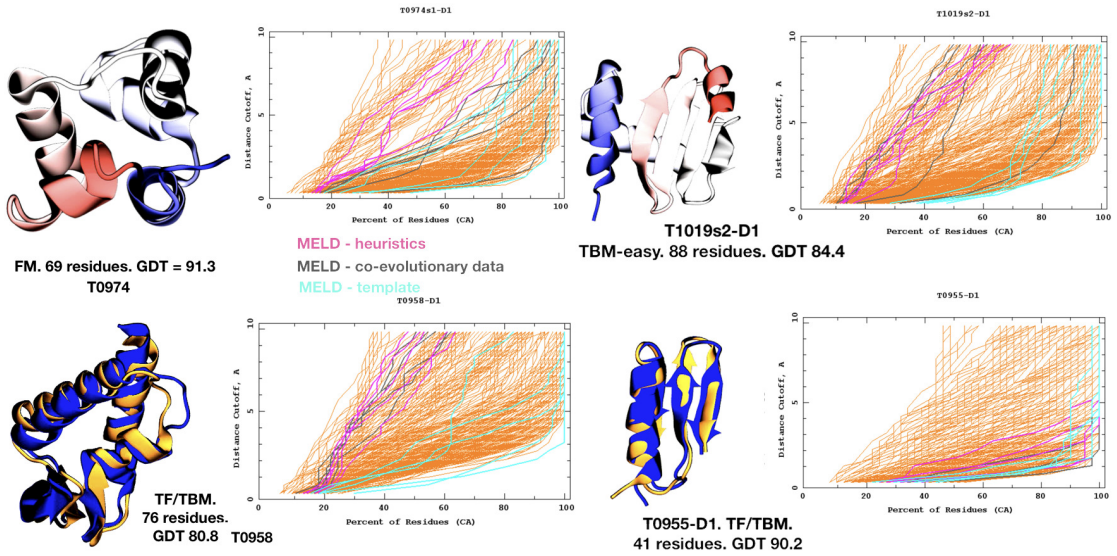

FIGURE S15 **Sampling and force fields influence MELD convergence on the native structure**. In the main portion of CASP we ran some proteins with three different protocols. The figure shows for proteins for which at least one method identified the native state. The GDT plots shows the five models submitted from each group in the T0 category, and highlights three different MELD protocols: using heuristics and starting from extended conformations (pink), adding co-evolutionary data and starting from extended (grey) and using co-evolutionary data and seeding from 15 server predictions (cyan, protocol used for NMR data as well).

TABLE S1 **CASP13 NMR-Assisted Targets Results Summary.** The GDT_TS scores are shown for all 5 models for every assessment unit. The best of the 5 models is in bold and highlighted red.

Target	Seq. Length	1	2	3	4	5
N0957s1	162	52.93	57.87	**62.35**	56.94	60.19
N0957s1-D1	108	74.07	74.07	**77.55**	70.14	76.62
N0957s1-D2	54	76.39	71.76	**86.11**	83.80	84.26
N0968s1-D1	118	59.53	**71.82**	63.35	61.23	62.71
N0968s2-D1	115	**73.70**	52.39	51.74	53.04	48.70
N0980s1-D1	104	67.79	72.11	**72.84**	57.93	59.62
N0981-D1	86	58.43	**65.99**	60.76	57.85	55.52
N0981-D2	80	40.00	38.44	37.81	**45.62**	36.88
N0981-D3	203	**41.01**	34.24	35.59	35.10	37.31
N0981-D4	111	65.77	**67.57**	59.91	60.59	35.81
N0981-D5	127	**76.58**	72.05	71.65	48.23	53.74
N0989	246	23.58	25.61	24.70	**26.22**	24.09
N0989-D1	134	**37.69**	34.89	33.21	36.94	37.13
N0989-D2	112	51.34	56.25	54.46	**56.92**	51.34
N1005-D1	326	**49.85**	43.33	45.32	42.41	48.62
N1008-D1	77	68.18	65.26	62.01	71.10	**74.03**
n1008-D1	77	57.47	56.17	**67.86**	63.64	33.44

TABLE S2 **NMR and MELD restraints.** Reported are the number of peaks and possible contacts provided by the NMR data, and the reduced number of those we input into MELD. GDT of the TOP1 and TOP5 predictions are also reported. The last 10 targets are relative to what we did in the CASP 11 competition.[21]

Target	Length	NMR		MELD		GDT	
		peaks	contacts	peaks	restraints	TOP1	TOP5
N0957s1	163	1124	6299	325	1993	52.9	62.3
N0968s1	126	803	1506	219	503	59.5	71.8
N0968s2	116	650	2088	219	1034	73.7	73.7
N0980s1	111	665	1489	207	797	67.7	72.8
N0981-D1	86	380	538	147	285	58.4	66
N0981-D2	80	395	504	191	334	40.0	45.6
N0981-D3	203	1269	4710	553	2790	41.0	41.0
N0981-D4	111	594	1093	220	492	65.8	67.6
N0981-D5	127	746	1983	277	999	76.6	76.6
N0989	246	1526	7095	217	216	23.5	26.2
N1005	364	4709	49887	1233	10844	49.9	49.9
N1008	97	627	2273	205	900	68.1	74.3
Ts761	237	3867	29210	828	5183	36.1	44.3
Ts763	131	2537	11516	693	2619	66.9	66.9
Ts785	112	1072	4009	210	510	84.1	84.1
Ts800	212	2251	19759	489	3313	68.6	69.1
Ts802	118	900	2014	223	475	81.1	83.5
Ts810	113	1174	3627	129	317	73.2	79.2
Ts818	134	873	2228	149	426	64.6	69.8
Ts824	110	867	1600	182	365	78.7	78.7
Ts826	201	2531	23959	152	816	23.7	27.4
Ts832	209	2146	17630	274	2346	66.4	66.4