Research Article Proteins: Structure, Function and Bioinformatics
DOI 10.1002/prot.24427

BCL::Fold – Protein topology determination from limited NMR restraints

Short title: Protein topology from limited NMR restraints

[1]Brian E. Weiner, [1]Nathan Alexander, [1]Louesa R. Akin, [1]Nils Woetzel, [1]Mert Karakas, and [1]*Jens Meiler

[1]Department of Chemistry, Center for Structural Biology, Vanderbilt University, Nashville TN, 37232, USA

*Corresponding author:

Jens Meiler

Vanderbilt University

465 21st Ave. S.

Nashville, TN 37232

USA

jens.meiler@vanderbilt.edu; Ph: 615-936-5662; Fax: 615-936-2211

Key words: protein structure prediction, topology prediction, sparse data, NOE, RDC

Received: Jul 01, 2013; Revised: Sep 03, 2013; Accepted: Sep 10, 2013

Abstract

When experimental protein NMR data is too sparse to apply traditional structure determination techniques, *de novo* protein structure prediction methods can be leveraged. Here we describe the incorporation of NMR restraints into the protein structure prediction algorithm BCL::Fold. The method assembles discreet secondary structure elements using a Monte Carlo sampling algorithm with a consensus knowledge-based energy function. New components were introduced into the energy function to accommodate chemical shift, nuclear Overhauser effect, and residual dipolar coupling data. In particular, since side chains are not explicitly modeled during the minimization process, a knowledge based potential was created to relate experimental side chain proton-proton distances to C_β-C_β distances. In a benchmark test of 67 proteins of known structure with the incorporation of sparse NMR restraints, the correct topology was sampled in 65 cases, with an average best model RMSD100 of 3.4 ± 1.3 Å versus 6.0 ± 2.0 Å produced with the *de novo* method. Additionally, the correct topology is present in the best scoring 1% of models in 61 cases. The benchmark set includes both soluble and membrane proteins with up to 565 residues, indicating the method is robust and applicable to large and membrane proteins that are less likely to produce rich NMR datasets.

Introduction

Traditional structure determination via NMR spectroscopy requires a rich dataset with a preference for distance restraints between amino acids that are far apart in sequence which serve to define the protein topology. In cases of sparse or primarily local restraints, identification of the correct topology becomes more difficult as several incorrect topologies may also satisfy the restraints. Additionally, knowledge of the topology is often required to assign otherwise ambiguous nuclear Overhauser effect (NOE) cross peaks that can then be used as additional distance restraints to further refine the structure. Recently, spectroscopists have begun taking advantage of advances in protein NMR

such as perdeuteration, selective labeling, and TROSY to study large proteins that were previously considered outside the realm of protein NMR. Nonetheless, the data collected on these large proteins are often sparse and of reduced quality, making structure determination challenging. Thus computational tools designed to predict protein topology from sparse data could facilitate the structure determination process.

Incorporating sparse NMR data into computational protein structure prediction algorithms has been shown to be extremely successful [1-4]. Rosetta, for example, was able to correctly fold proteins up to 25 kDa using backbone-only NMR data [5]. For larger proteins, the algorithm was unable to sample native-like topologies, which indicates that conformational sampling is still the computational bottleneck, even with the inclusion of experimental restraints. Incorporation of sparse side chain distance restraints from deuterated samples increased the feasible upper limit to 40 kDa [6].

Like many protein structure-prediction methods, Rosetta uses a simplified side chain approximation during the model building stages, so handling of any available side chain-side chain NOE restraints is not directly modeled. In these cases, an arbitrary amount is typically added to the distance restraint in order to represent the restraint as a backbone-backbone distance. This approach however reduces the information content of each restraint. The problem of relating experimentally determined distances to distances measurable during the minimization process is not unique to NMR data. In site-directed spin labeling electron paramagnetic resonance (SDSL-EPR) experiments, distances are reported between two spin labels covalently attached at specific sites on the protein model. A knowledge-based potential has been developed and successfully used to evaluate the probability of observing the C_β-C_β distance given the spin label-spin label distance [7]. We take a similar approach with side chain-side chain proton distances from NOE data to evaluate C_β-C_β distances in the model with the hypothesis that this method will produce more native-like models.

A protein structure prediction method, BCL::Fold, was recently introduced with the goal of efficiently sampling larger and more complex topologies than those accessible to other *de novo* protein structure prediction algorithms [8]. Like most algorithms, BCL::Fold begins with protein secondary structure prediction. The predicted secondary structure elements (SSEs) are then collected into a pool, with loops and side chains being discarded. A Monte Carlo algorithm assembles the SSE building blocks into a viable topology, guided by a consensus knowledge-based energy function. The final model is generated via subsequent loop building and side-chain replacement. Both the assembly and scoring stages are flexible, making the incorporation of experimental restraints possible. This has already been successfully demonstrated with cryo-electron microscopy data [9].

Here we describe the incorporation of three types of NMR restraints – chemical shifts (CSs), NOEs, and residual dipolar couplings (RDCs) – into the BCL::Fold algorithm. A novel NOE knowledge-based potential was developed in order to evaluate C_β - C_β distances observed in the model based on experimental side chain-side chain restraints. The method was benchmarked using 23 structures with experimental restraints and an additional 44 proteins with simulated restraints. The incorporation of restraints enhanced native-like sampling and facilitated the selection of low RMSD models. BCL::Fold is therefore a viable method for rapid identification of protein topology from sparse NMR restraints.

Materials and Methods

INPUT FILES. Chemical shift data is read in indirectly as a TALOS+ [10] secondary structure prediction file (*SS.tab). Both RDC and NOE data are read in directly using the NMR-STAR 3.1 format [11] as supported by the BMRB. RDC data can be normalized to N-H values or have signs adjusted to

account for the negative gyromagnetic ratio of nitrogen via command-line flags, but was not necessary for the selected benchmark proteins.

SELECTION OF BENCHMARK PROTEINS. 67 total proteins were selected from three groups: 1) 6 large proteins from the BCL::Fold benchmark, 2) 38 membrane proteins from the BCL::MP-Fold benchmark, and 3) 23 small, soluble proteins containing experimental NMR data. The experimental benchmark set contains proteins that have both NOE and RDC data available on the BMRB [12], aside from 1CFE [13], 1ULO [14], and 2EE4, which have no RDC data. The benchmark proteins with experimental data contain no ligands, have less than 30% sequence similarity, range in length from 58 to 224 residues, and are soluble, single chains. Additionally, the proteins were selected to have a diverse set of alpha, beta, and alpha/beta topologies with > 50% SSE content.

MODIFICATION TO THE ALGORITHM. The NMR restraint scores are added to the BCL::Fold method as part of the restraint protocol. Refer to the supplementary information for required command line flags and modifications to the stage and score weight set files. Iterative folding rounds were also introduced to better leverage experimental restraint information. After generating 1000 models, the top 10 models were selected by restraint score and used as start models to generate a new set of 1000 models. For the six large, soluble proteins, this process was repeated once more. In the subsequent analysis, only the models produced by the last iteration are considered.

BENCHMARK. 1000 models were generated with and without the incorporation of NMR restraints for each protein in the benchmark set. All CS and RDC data for residues in SSEs were used when available. When CS data was not available for SSE pool generation, it was simulated using SPARTA+ [15]. In order to simulate sparse NOE data, random subsets of the experimental restraints were selected where both atoms were in SSEs and at least five residues apart. Here we exclude short and

medium range distance restraints in order to focus on the long range distance restraints that serve to constrain the topology. Experimental selective labeling strategies also enrich for long range distance restraints since there is an increased chance neighboring atoms are not labeled; instead there is a predominance of side chain methyl groups that engage in long range van der Waals contacts in the protein core. For each protein, ten random subsets were selected, and the subset size was equal to the number of residues in SSEs. These datasets were further reduced (down to 0.1 restraints/residue) and expanded (up to 2.0 restraints/residue) in order to evaluate the effect of restraint density on topology prediction accuracy. To generate the complete 1000 models, 100 models were constructed for each NOE restraint subset. Example command lines for running BCL::Fold can be found in the Supporting Information.

AVAILABILITY. BCL::Fold is implemented as part of the BioChemical Library, a suite of software currently under development in the Meiler laboratory (www.meilerlab.org). BCL software, including BCL::Fold, is freely available for academic use.

Results and Discussion

RESTRAINT SCORE FUNCTIONS. Three scoring functions were introduced into BCL::Fold in order to accommodate evaluation of NMR restraints. RDCs are evaluated using the traditional Q-value measure [16]. To evaluate NOE distance restraints, a knowledge-based score, NOE-KB, and an atom distance penalty score, NOE-pen, are used in conjunction. CS's are evaluated indirectly using the previously described secondary structure prediction agreement score [17] via the program TALOS+ [10].

To evaluate RDC restraints, the optimal tensor is determined using the Saupe order matrix approach [18-20] after each minimization step. This gives a calculated theoretical RDC value for each

supplied experimental value. The Q-value is then calculated, $Q = \sqrt{\sum_{ij}(D^{ij}_{\exp} - D^{ij}_{theor})^2 / \sum_{ij}(D^{ij}_{\exp})^2}$, where D^{ij} is the dipolar coupling between nuclei i and j[16]. The unweighted score is given by, $RDC = Q - 1$, so that a perfect agreement gives a score of -1.

Since BCL::Fold assembles SSEs lacking side chain atoms, a method was needed to relate distance restraints between side chain protons to useable backbone-to-backbone distances. The PISCES databank [21] was used to cull a list of 4379 proteins with less than 25% sequence identity and better than 2.0 Å resolution. Proton atoms were added using the program Reduce [22]. Statistics were then collected in order to relate each H-H distance to the corresponding C_β-C_β distance. A separate histogram was created for the total number of bonds the protons were away from the C_β. For example, a $H_{\beta 3}$-$H_{\delta 2}$ pair totals four bonds away from C_β's. Separate histograms were generated for restraints to H_α or amide H since the coordinates of these atoms can be determined directly from BCL::Fold models. The C_β-C_β distance minus the H-H distance was computed and placed in a corresponding 0.5 Å bin. This process was repeated for each H-H pair at least 5 residues apart in sequence but no more than 6.0 Å apart in space for each of the proteins in the dataset. Each histogram was then converted to a cubic spline such that distances in the most common bin receive a score near -1 and distances not observed receive a score of zero (Figure 1A-C). The unweighted NOE-KB score is set as the mean individual restraint scores.

The NOE-pen score is simply a trigonometric transition between the maximal score, zero, and the ideal score, -1. The width of the transition is set to 25 Å. The curve is generated such that it reaches a value of -1 at a distance of 2 Å greater than the smallest observed distance for the given atom types (Figure 1D). This score was introduced to evaluate moderately to severely violated distance restraints; the NOE-KB score has a rather narrow minimum, and thus cannot adequately discriminate these violations.

The standard BCL::Fold KB energy potentials scale linearly with respect to protein size. For consistency, each restraint score is therefore multiplied by the number of residues in the protein model to achieve the same property. An additional consideration for restraint scores is how to handle scaling of the score with the number of restraints. We chose to have the score scale logarithmically with the number of restraints. This allows for the score to change with additional restraints, but not overwhelmingly so. Finally, each score was given a relative weight of 5.0. With this scaling the experimental data contribute approximately 50% to the total score of the model while the KB potentials contribute the remainder of the score. The final restraint energy is given by the following equation:

$$E_{rest} = N(w_{RDC}(Q-1)\log(M_{RDC}+1) + (w_{KB}\bar{s}_{KB} + w_{Pen}\bar{s}_{Pen})\log(M_{NOE}+1)),$$

where M is the number of restraints, N is the number of amino acids in the target, w is the weight (the default case being 5.0), and $\bar{s}$ is the average NOE score.

SELECTION OF A DIVERSE BENCHMARK SET. A benchmark set of proteins of known structure was collected to test for the ability of the NMR scores to enhance native-like sampling during BCL::Fold minimizations. The set contains 67 total proteins, broken into three groups. 23 proteins are small, soluble proteins, with structures determined by NMR and with CS, NOE, and/or RDC data available on the BMRB. An additional six are large (> 220 residues) proteins from the original BCL::Fold method benchmark test [8]. The final 38 proteins are membrane proteins from the BCL::MP-Fold benchmark test [23]. Membrane proteins are on the frontier of protein NMR, and are therefore more likely to produce sparse, rather than complete, datasets.

The small soluble proteins have complete datasets, so random subsets of NOE restraints were selected for a total of one long-range restraint per residue in SSEs to create sparse data. NMR restraints were simulated for the large soluble proteins and the membrane proteins. Again one restraint per residue

was selected as the initial restraint density. For the membrane proteins, side chain NOE restraints (1 restraint/residue) were limited to isoleucine, leucine, and valine residues to mimic the increasingly popular strategy of specific isotopic labeling of methyl groups [24].

NOE KNOWLEDGE-BASED FUNCTION ENRICHES FOR NATIVE-LIKE MODELS. Each small, soluble native protein in the benchmark set was scored with the NOE-KB score and the NOE-pen score for agreement with all available long range experimental NOEs. With an ideal score of -1.00, the mean NOE-KB score was -0.84 ± 0.07 BCL energy units (BCLEUs), and the mean NOE-pen score was -1.00 ± 0.00 BCLEUs. The NOE-KB score is not exactly -1.00 BCLEUs due to experimental error and the fact that the score represents a rather wide distribution of observed distances, with only the most commonly occurring receiving scores near -1.00 BCLEUs.

In order to test the ability of NOE scores to select for native-like models, we created a set of decoy models. For each protein, 10,000 decoys were generated by de novo protein structure prediction without restraints using BCL::Fold. These decoys were then also scored with the two NOE scores. We define any model with less than 8.0 Å RMSD100 [25] to the native as "native-like" or a "good" model. RMSD100 is the C_α RMSD normalized to a protein length of 100 residues. This measure is useful when evaluating proteins of varying sizes, such as those used in this benchmark. Using the 8.0 Å cutoff, the enrichment was calculated for those proteins which produced at least 0.1% "good" models [17]. Ranking the models by the sum of the NOE scores produces an average enrichment of 5.5 ± 1.6 out of a maximal 10.0. In contrast, using a quadratic energy function analogous to the bounded energy potential in Rosetta [1] produces an average enrichment of 4.9 ± 1.4 ($p = 0.02$). This demonstrates that the NOE-KB and NOE-pen scoring functions improve the identification of native-like models when compared to the traditional score.

NATIVE-LIKE SAMPLING IS ENHANCED WITH NMR RESTRAINT SCORES. For each protein in the benchmark set, 1000 models were generated using the de novo BCL::Fold method. An additional 1000 models were also constructed using the available NMR restraints in combination with the implemented scoring functions. Over all proteins, the average C_α RMSD100 of the best model to the native structure was 3.4 ± 1.3 Å with restraints and 6.0 ± 2.0 Å without (Table I, Figures 2,3). When a structure with an RMSD100 of less than 8.0 Å is considered to be the correct topology, the inclusion of restraints allows for sampling of the correct topology in 65 of 67 cases (97%) compared to 54 of 67 cases (81%) when no restraints are incorporated. With a cutoff of 6.0 Å, the correct topology is sampled in 64 cases (96%) with restraints and in 41 cases (61%) without. With a cutoff of 4.0 Å, the correct topology is sampled in 54 cases (81%) with restraints and in 9 cases (13%) without. When looking at the top 5% of models produced from the first round, the best dataset contributes 18% of the top models on average (vs 10% expected with a random distribution), with the worst contributing 3% (Table S1). We conclude that while there is a dataset bias, even the 'worst' dataset is capable of producing highly accurate models – possible additional sampling is needed.

Of the small, soluble proteins, 2KYY showed the largest improvement upon the incorporation of restraints, with a best model RMSD100 decrease of 5.8 Å. The protein is a mixed α/β fold with 153 residues. The de novo method assembles a sheet, but the strand order is incorrect and the helices are not properly placed on either side of the major sheet. In contrast, the NMR method is able to build the sheet with the proper ordering and the helices are appropriately placed. Of the proteins with simulated NMR data, 1VIN [26] showed the largest improvement upon the incorporation of restraints, with a best model RMSD100 decrease of 7.5 Å. This protein contains thirteen helices and 252 residues, placing it on the upper edge of de novo BCL::Fold's predictive capabilities; the native topology is sampled however, even without restraints [8]. Here restraints serve to improve accuracy by promoting sampling of those

models with the correct topology. After the first round of iterative folding, the best model produced has an RMSD100 of 4.7 Å. The subsequent iterations then are typically starting their minimizations with the correct topology, making production of an accurate model much more likely.

BCL::FOLD COMPARES FAVORABLY WITH THE ROSETTA METHOD. Rosetta is a well established protein structure prediction method with a proven track record of producing quality models with limited experimental data. The structures of the soluble proteins in the benchmark were also predicted using the same sparse datasets using the AbinitioRelax application in Rosetta. Chemical shift data were used to generate fragments, and both NOE and RDC data were used during the minimizations. Side chain NOE restraints were converted to C_β restraints by adding 1.0 Å to the restraint distance per bond from the side chain proton to the C_β. 1000 models were generated per target, and the top 5% of models selected by RMSD100 to the native were retained for comparison with BCL models. The mean RMSD100 of the top Rosetta models was 4.9 ± 1.8 Å compared to 3.9 ± 1.4 Å for BCL::Fold (Table S2, $p = 0.003$). While BCL::Fold appears to sample topologies slightly better than Rosetta in our experiment, it should be noted that Rosetta is still the method of choice for loop building and side chain replacement once the topology has been constructed.

FEW NOE RESTRAINTS ARE REQUIRED FOR THE SAMPLING IMPROVEMENT. The previously described benchmark test used one NOE restraint per residue in SSEs. As a next step, additional restraint densities (0.1, 0.2, 0.5, and 2.0 restraints/residue) were tested for those proteins containing experimental data (Figure 4). After iterative folding, the top 5% of models by RMSD100 were analyzed from each group. The model quality improves up to 0.5 restraints/residue, but further increasing the number of restraints to 1.0 restraints/residue shows no effective additional improvement (the mean RMSD100 decrease is 0.2 ± 0.9 Å, $p = 0.31$). Analyzed separately, however, sampling for the larger proteins (> 125 residues) does improve overall from 0.5 to 1.0 restraints/residue. For proteins less

than 125 residues, the average improvement in the top 5% of models selected by RMSD100 sampled is 0.0 ± 0.6 Å. For proteins with more than 125 residues, the improvement is 0.6 ± 1.3 Å.

RESTRAINT SCORES FACILITATE MODEL SELECTION. The selection of the best model(s) out of the thousands generated is a difficult problem, especially when using low-resolution energy functions, as is the case with BCL::Fold. Table I highlights this problem by listing the RMSD100 of the lowest energy model. When no restraints are considered, the average RMSD100 is 10.6 ± 2.3 Å. However when NMR restraints are used, the average RMSD100 of the model with the lowest score is 5.4 ± 2.6 Å. Perhaps more strikingly, when the top 1% of models are selected by score, the native topology is contained within this subset in 27 out of 67 cases (40%) without restraints versus 61 out of 67 cases (91%) when using sparse NMR data.

BUILDING FULL ATOM MODELS. In order to explore the feasibility of constructing full atom models from BCL::Fold-generated topologies, we used the protein 1VIN as a test case. For this 252 residue helical protein, BCL::Fold produced models with an RMSD100 down to 1.8 Å compared to the native when sparse restraints were considered. The 50 lowest scoring models of the 1000 generated during the BCL::Fold benchmark test were retained for loop building using the Rosetta CCD loop building protocol. Side chains were then added using the Rosetta FastRelax protocol to generate 1000 complete, full atom models. Of the 20 best scoring final models, the mean backbone C_α RMSD100 was 2.4 ± 0.2 Å RMSD100 to the native SSE residues and 4.5 ± 0.4 Å over all residues.

POTENTIAL APPLICATIONS. One potential use of sparse restraints with BCL::Fold is to assist in the identification of ambiguous NOE assignments. For proteins that are suitable for traditional NMR structure determination methods, this would speed up the process by allowing for more confident NOE assignments during the structure determination process. Additionally, the BCL::Score program

can be used to identify any violated restraints in the given model, which can lead to subsequent NOE re-assignments or model refinement.

Perhaps the most exciting application for BCL::Fold lies with membrane proteins. Membrane proteins constitute roughly 50% of all known drug targets, yet only 2% of the deposited PDB structures [27]. BCL::Fold can sample the native topology in all but 2 of the 38 membrane proteins in the benchmark when combined with sparse NMR data. This includes predicted models of less than 4.0 Å RMSD100 to the native for five proteins larger than 400 residues (with up to 15 transmembrane helices).

Conclusions

The *de novo* protein structure prediction method, BCL::Fold, has been updated to incorporate sparse experimental NMR data. Scoring functions were introduced to evaluate CS, NOE, and RDC data. In particular, a NOE knowledge-based potential was developed to relate experimental side chain proton-proton distance restraints to C_β-C_β distances that are measurable during the BCL::Fold minimization.

The benchmark test using a robust dataset demonstrated that sparse NMR data can be combined with BCL::Fold to produce native topologies in 97% of the cases. Using 1.0 NOE distance restraint per residue produces a mean improvement of 2.6 Å RMSD100 versus the *de novo* method. Reducing the number of restraints to 0.1 per residue still produces a mean improvement of 1.1 Å RMSD100 versus the de novo method. BCL::Fold, therefore, has the potential to provide experimentalists with feasible models that satisfy available NMR data to be used to generate further structure-based hypotheses.

Acknowledgments

The authors thank the Vanderbilt University Center for Structural Biology computational support team for hardware and software maintenance. We also thank the Vanderbilt University Advanced Computing Center for Research and Education for computer cluster access and support. Work in the Meiler laboratory is supported through NIH (R01 GM080403, R01 MH090192, R01 GM099842) and NSF (Career 0742762).

References

1. Rohl CA. Protein structure estimation from minimal restraints using Rosetta. Methods Enzymol 2005;394:244-260.
2. Li W, Zhang Y, Kihara D, Huang YJ, Zheng D, Montelione GT, Kolinski A, Skolnick J. TOUCHSTONEX: protein structure prediction with sparse NMR data. Proteins 2003;53(2):290-306.
3. Latek D, Kolinski A. CABS-NMR--De novo tool for rapid global fold determination from chemical shifts, residual dipolar couplings and sparse methyl-methyl NOEs. J Comput Chem 2011;32(3):536-544.
4. Zheng D, Huang YJ, Moseley HN, Xiao R, Aramini J, Swapna GV, Montelione GT. Automated protein fold determination using a minimal NMR constraint strategy. Protein Sci 2003;12(6):1232-1246.
5. Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu G, Ramelot TA, Eletsky A, Szyperski T, Kennedy MA, Prestegard J, Montelione GT, Baker D. NMR structure determination for larger proteins using backbone-only data. Science 2010;327(5968):1014-1018.
6. Lange OF, Rossi P, Sgourakis NG, Song Y, Lee HW, Aramini JM, Ertekin A, Xiao R, Acton TB, Montelione GT, Baker D. Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. Proc Natl Acad Sci U S A 2012;109(27):10873-10878.
7. Hirst SJ, Alexander N, McHaourab HS, Meiler J. RosettaEPR: an integrated tool for protein structure determination from sparse EPR data. J Struct Biol 2011;173(3):506-514.
8. Karakas M, Woetzel N, Staritzbichler R, Alexander N, Weiner BE, Meiler J. BCL::Fold--de novo prediction of complex and large protein topologies by assembly of secondary structure elements. PLoS One 2012;7(11):e49240.
9. Lindert S, Staritzbichler R, Wotzel N, Karakas M, Stewart PL, Meiler J. EM-fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps. Structure 2009;17(7):990-1003.
10. Shen Y, Delaglio F, Cornilescu G, Bax A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 2009;44(4):213-223.
11. Hall SR, Cook APF. Star Dictionary Definition Language - Initial Specification. Journal of Chemical Information and Computer Sciences 1995;35(5):819-825.
12. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL. BioMagResBank. Nucleic Acids Res 2008;36(Database issue):D402-408.
13. Fernandez C, Szyperski T, Bruyere T, Ramage P, Mosinger E, Wuthrich K. NMR solution structure of the pathogenesis-related protein P14a. J Mol Biol 1997;266(3):576-593.
14. Johnson PE, Joshi MD, Tomme P, Kilburn DG, McIntosh LP. Structure of the N-terminal cellulose-binding domain of Cellulomonas fimi CenC determined by nuclear magnetic resonance spectroscopy. Biochemistry 1996;35(45):14381-14394.

15. Shen Y, Bax A. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. J Biomol NMR 2007;38(4):289-302.
16. Cornilescu G, Marquardt JL, Ottiger M, Bax A. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. Journal of the American Chemical Society 1998;120(27):6836-6837.
17. Woetzel N, Karakas M, Staritzbichler R, Muller R, Weiner BE, Meiler J. BCL::Score--knowledge based energy potentials for ranking protein models represented by idealized secondary structure elements. PLoS One 2012;7(11):e49242.
18. Losonczi JA, Andrec M, Fischer MW, Prestegard JH. Order matrix analysis of residual dipolar couplings using singular value decomposition. J Magn Reson 1999;138(2):334-342.
19. Saupe A. Recent Results in Field of Liquid Crystals. Angewandte Chemie-International Edition 1968;7(2):97-&.
20. Meiler J, Peti W, Griesinger C. DipoCoup: A versatile program for 3D-structure homology comparison based on residual dipolar couplings and pseudocontact shifts. J Biomol NMR 2000;17(4):283-294.
21. Wang G, Dunbrack RL, Jr. PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res 2005;33(Web Server issue):W94-98.
22. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 1999;285(4):1735-1747.
23. Weiner BE, Woetzel N, Karakas M, Alexander N, Meiler J. BCL::MP-Fold: Folding Membrane Proteins through Assembly of Transmembrane Helices. Structure 2013.
24. Rosen MK, Gardner KH, Willis RC, Parris WE, Pawson T, Kay LE. Selective methyl group protonation of perdeuterated proteins. J Mol Biol 1996;263(5):627-636.
25. Carugo O, Pongor S. A normalized root-mean-square distance for comparing protein three-dimensional structures. Protein Sci 2001;10(7):1470-1473.
26. Brown NR, Noble ME, Endicott JA, Garman EF, Wakatsuki S, Mitchell E, Rasmussen B, Hunt T, Johnson LN. The crystal structure of cyclin A. Structure 1995;3(11):1235-1247.
27. Bakheet TM, Doig AJ. Properties and identification of human protein drug targets. Bioinformatics 2009;25(4):451-457.

Figure Legends

Figure 1. NOE knowledge based potentials. The energy potential for each cumulative bond distance is plotted versus the measured C_β-C_β distance subtracted from the expeimental H-H distance. The bond distance is the number of bonds between the measured proton and the C_β atom of the same residue. For example, an NOE between $H_{\beta 3}$ and $H_{\delta 2}$ would have a cumulative bond distance of four. (A) Potentials for side chain-side chain NOEs. (B) Potentials H_α-side chain NOEs. (C) Potentials for backbone amide H-side chain NOEs. (D) The NOE-KB and NOE-pen potentials are plotted for a cumulative bond distance of 5.

Figure 2. NMR restraints improve native-like sampling. (A) The mean RMSD100 values of the best 10 models sampled with and without restraints are plotted. Soluble proteins are represented by circles and membrane proteins by squares. Proteins are colored according to size: < 150 residues (green), ≥ 150 and < 250 residues (yellow), ≥ 250 and < 400 residues (orange), and ≥ 400 residues (red). The dashed line at 8.0 Å indicates the cutoff for the correct topology, and the dashed line at 4.0 Å indicates a feasible target for continuing with full atom refinement. The error bars are ± 1 S.D. (B) Of the top 10 models by score, the RMSD100 value of the best model is plotted for folding with and without restraints. Marker shapes and colors are the same as in panel A.

Figure 3. Gallery of select benchmark results. Left column – Distribution of RMSD100 to native SSE values for models produced by the de novo method (red) and the restraint-based method (green). Right column – Superimposition of the best model produced by the restraint method (rainbow) with the native protein (gray). Refer to the supporting information for the complete gallery of benchmark results.

Figure 4. Sampling efficiency depends upon restraint density. The size of the random subset of NOEs selected for folding was adjusted relative to the total number of residues in native SSEs. Each of the 23 proteins with experimental data was folded at varying restraint densities (0.0, 0.1, 0.2, 0.5, 1.0, and 2.0 restraints/residue). The distribution of the mean RMSD100 for the top 5% (selected by RMSD100) of models for each benchmark protein are shown. The boxes contain values within one standard deviation of the mean (of mean RMSD100 values) and the lines represent the minimum and maximum values observed from the 23 proteins for that restraint density. *Improvement over previous restraint density ($p < 0.01$).

Figure 5. Core side chain conformations can be accurately predicted. Native protein model 1VIN is shown in gray, with side chain atoms displayed for His63, Leu64, Tyr68, and Phe97. The corresponding side chains from the best scoring Rosetta model after full-atom refinement are shown in black.

Figure S1. Gallery of benchmark results with experimental data. Left column – Distribution of RMSD100 to native SSE values for models produced by the de novo method (red) and the restraint-based method (green). Right column – Superimposition of the best model produced by the restraint method (rainbow) with the native protein (gray). Refer to the supplementary information for the complete gallery of benchmark results.

Figure S2. Gallery of soluble protein benchmark results with simulated data. Left column – Distribution of RMSD100 to native SSE values for models produced by the de novo method (red) and the restraint-based method (green). Right column – Superimposition of the best model produced by the restraint method (rainbow) with the native protein (gray). Refer to the supplementary information for the complete gallery of benchmark results.

Figure S3. Gallery of membrane protein benchmark results with simulated data. Left column – Distribution of RMSD100 to native SSE values for models produced by the de novo method (red) and the restraint-based method (green). Right column – Superimposition of the best model produced by the restraint method (rainbow) with the native protein (gray). Refer to the supplementary information for the complete gallery of benchmark results.

Table I. Benchmark statistics and results.

PDB	Statistics					Restraints	RMSD100 (Å)				Score (Å)			
	AA	Type	Helices	Strands	rco		Best		Top 5%		Best		Top 5%	
							NMR	*dn*	NMR	*dn*	NMR	*dn*	NMR	*dn*
1Q2N	58	A	3	0	0.41	exp	4.0	2.8	4.7	4.0	5.2	8.4	5.5	10.4
2KIQ	62	A	4	0	0.37	exp	3.0	4.9	4.3	6.7	5.1	13.8	5.3	12.5
2L9R	69	A	3	0	0.27	exp	3.0	4.1	3.5	5.2	9.5	13.3	5.2	10.8
1WCL	76	A	5	0	0.27	exp	2.9	5.5	3.2	7.1	3.9	10.7	4.4	11.0
2L7K	76	A	4	0	0.45	exp	3.3	5.9	3.9	7.3	7.4	10.4	7.2	11.0
1OP1	82	A	3	0	0.31	exp	2.8	3.4	3.2	4.2	5.1	12.9	6.9	10.7
2AMW	83	A	3	0	0.38	exp	4.1	4.2	4.5	6.2	6.7	7.1	7.0	10.2
2KYW	87	B	0	7	0.37	exp	4.8	7.0	5.3	8.5	5.4	10.7	6.7	11.7
2BG9	91	A (MP)	3	0	0.41	sim	2.3	2.8	2.5	3.4	2.8	9.9	2.8	6.7
1W09	92	A	3	0	0.44	exp	1.9	3.5	2.0	4.5	4.4	10.1	2.8	11.3
1NKZ	93	A (MP)	3	0		sim	5.8	4.3	6.8	4.6	16.2	11.2	12.0	8.3
2KCT	94	B	0	6	0.27	exp	4.0	8.6	4.6	9.3	10.0	12.0	9.1	12.3
2H45	95	B	0	6	0.32	exp	4.1	4.1	6.1	5.7	10.2	13.3	8.2	9.7
2L35	95	A (MP)	3	0		sim	2.6	3.1	2.8	3.7	3.5	17.2	3.5	9.7
2KLC	101	A/B	1	5	0.28	exp	3.6	4.6	4.4	7.2	7.2	11.8	5.4	11.6
2KSF	107	A (MP)	4	0	0.34	sim	2.9	3.9	3.1	4.5	3.6	5.1	3.3	5.6
2JV3	110	A	6	0	0.28	exp	2.5	5.1	3.2	7.1	5.7	8.9	4.9	10.0
2A7O	112	A	3	0	0.34	exp	1.7	2.3	2.1	4.2	4.1	11.6	3.7	11.0
2KCK	112	A	6	0	0.18	exp	3.0	5.7	3.8	7.8	6.3	12.6	5.3	10.0
1J4N	116	A (MP)	4	0	0.40	sim	2.6	4.9	3.2	5.9	4.6	9.6	4.9	9.0
2KD1	118	A	5	0	0.25	exp	2.6	4.5	2.8	5.5	5.0	9.8	4.6	9.2
3SYO	122	A (MP)	4	0	0.33	sim	4.9	5.2	5.4	6.3	7.6	9.7	8.6	10.0
1PY7	123	A (MP)	4	0	0.28	sim	2.4	3.9	2.7	4.7	3.1	5.4	3.2	6.4
2PNO	130	A (MP)	4	0	0.29	sim	1.8	5.0	2.3	6.7	2.8	5.4	3.1	8.6
1CFE	135	A/B	4	4	0.35	exp	2.8	5.7	3.2	8.3	3.9	12.2	4.3	10.8
2L3W	143	A	7	0	0.32	exp	2.8	6.2	3.3	8.1	3.4	9.6	5.3	10.3
2BL2	145	A (MP)	6	0	0.37	sim	2.2	2.9	2.5	3.8	3.2	6.7	3.6	7.3
1CMZ	152	A	9	0	0.26	exp	4.4	7.7	5.0	9.6	5.7	12.2	5.8	12.6
1ULO	152	B	0	10	0.34	exp	4.1	6.9	4.6	8.7	5.4	12.4	6.1	11.3
2KYY	153	A/B	3	6	0.31	exp	3.2	9.0	3.6	9.8	4.8	11.5	4.3	12.0
2K73	164	A (MP)	6	2	0.33	sim	3.3	4.7	4.1	5.9	9.0	10.1	6.8	9.1
1RHZ	166	A (MP)	6	0	0.33	sim	3.8	6.7	4.3	8.0	5.7	9.9	5.4	10.4
1IWG	168	A (MP)	7	0	0.31	sim	2.4	4.3	2.9	5.6	3.2	8.5	3.6	8.3
3P5N	179	A (MP)	8	0	0.24	sim	2.6	5.8	3.3	7.4	4.4	8.3	4.5	9.8
2IC8	182	A (MP)	8	0	0.25	sim	2.9	6.0	3.8	7.2	4.3	9.5	5.2	9.3
2YVX	188	A (MP)	5	0	0.34	sim	3.3	5.1	4.1	6.9	5.5	9.2	5.5	9.4
1PV6	189	A (MP)	11	0	0.42	sim	2.6	5.7	2.8	6.8	3.4	10.6	4.1	9.4

1OCC	191	A (MP)	5	0	0.33	sim	2.2	4.6	2.5	5.9	3.2	8.5	3.7	8.0
2NR9	192	A (MP)	8	0	0.24	sim	3.5	5.7	4.1	7.2	4.7	8.7	5.0	9.5
4A2N	192	A (MP)	6	2	0.31	sim	3.7	4.3	4.0	6.2	4.0	8.1	4.7	8.8
1RW5	199	A	5	0	0.38	exp	1.6	4.7	1.8	7.9	2.3	11.5	3.0	11.1
1KPL	203	A (MP)	8	0	0.31	sim	3.0	8.7	3.4	10.5	6.6	14.4	4.9	12.5
2EE4	209	A	12	0	0.23	exp	2.8	7.5	3.5	9.4	3.6	12.8	4.6	11.4
2ZW3	216	A (MP)	8	3	0.35	sim	2.6	4.0	3.2	5.1	5.3	9.2	5.8	8.1
2BS2	217	A (MP)	8	0	0.27	sim	3.4	5.4	3.9	6.9	5.1	11.0	4.8	9.2
1L0V	221	A (MP)	9	0		sim	3.3	5.2	3.9	7.2	8.2	9.0	7.5	9.4
1UAI	223	B	0	16	0.25	sim	5.8	7.9	6.7	9.1	8.2	11.0	8.2	10.8
2KSY	223	A (MP)	9	2	0.26	sim	2.1	5.1	2.6	6.3	3.4	9.3	3.2	8.6
1PY6	227	A (MP)	7	2	0.27	sim	2.1	4.8	2.5	5.9	2.4	6.1	3.3	8.4
1VIN	252	A	13	0	0.12	sim	1.8	9.3	2.3	10.1	2.9	12.3	2.7	11.9
3KCU	252	A (MP)	14	0	0.29	sim	3.5	7.3	4.0	8.5	3.8	11.2	4.8	10.5
1XQO	253	A	14	0	0.23	sim	6.6	8.8	7.6	10.1	9.7	12.6	9.3	12.2
1FX8	254	A (MP)	12	0	0.28	sim	4.0	6.4	4.7	7.6	5.5	9.3	5.7	9.8
2OF3	266	A	15	0	0.13	sim	3.4	9.6	3.9	11.2	4.7	13.5	4.8	13.6
1U19	278	A (MP)	10	2	0.24	sim	3.0	5.3	3.9	6.6	3.8	8.9	4.2	8.8
2ZCO	284	A	15	0	0.17	sim	2.3	8.9	2.7	10.2	2.7	13.0	3.1	12.3
2R0S	285	A	14	0	0.20	sim	3.1	9.1	3.4	10.0	4.8	11.2	4.0	11.9
1OKC	292	A (MP)	11	0	0.25	sim	4.4	7.1	4.9	8.2	5.6	9.9	8.1	10.3
3KJ6	311	A (MP)	15	0	0.28	sim	3.5	5.9	4.8	7.4	3.5	10.5	5.5	10.0
3B60	319	A (MP)	11	0	0.27	sim	4.7	9.5	5.6	10.8	7.3	12.4	7.4	13.2
3HD6	403	A (MP)	15	2	0.23	sim	3.5	7.2	4.1	8.2	4.5	11.0	4.6	10.3
3GIA	433	A (MP)	18	0	0.34	sim	3.0	9.6	3.6	10.7	6.6	13.4	7.3	12.6
3O0R	449	A (MP)	18	0	0.15	sim	2.9	6.9	3.6	8.2	2.9	10.2	4.1	10.3
2XUT	488	A (MP)	24	0	0.22	sim	8.8	7.7	9.6	9.0	12.1	10.2	11.6	11.4
3HFX	493	A (MP)	18	0	0.36	sim	3.2	8.9	3.7	9.7	4.1	13.1	4.6	11.4
1YEW	528	A (MP)	20	3		sim	8.2	9.7	9.6	11.5	10.4	14.1	11.8	13.3
2XQ2	565	A (MP)	28	0	0.29	sim	3.5	8.2	4.0	10.1	5.4	12.2	5.7	12.1
Mean	199		8	1	0.30		3.4	6.0	4.0	7.3	5.4	10.6	5.5	10.3
SD	119		6	3	0.07		1.3	2.0	1.5	2.1	2.6	2.3	2.1	1.7

Protein types are "A" for alpha-helical and "B" for beta-strands. "MP" denotes a membrane protein. The NMR restraints used were from published experimental data ("exp") or simulated computationally ("sim"). The best models were selected by either RMSD100 ("RMSD100" columns) or score ("Score" columns). RMSD100 values are displayed for both the best model and the mean of top 5% of models.

The models generated with NMR restraints ("NMR") and without ("*dn*").

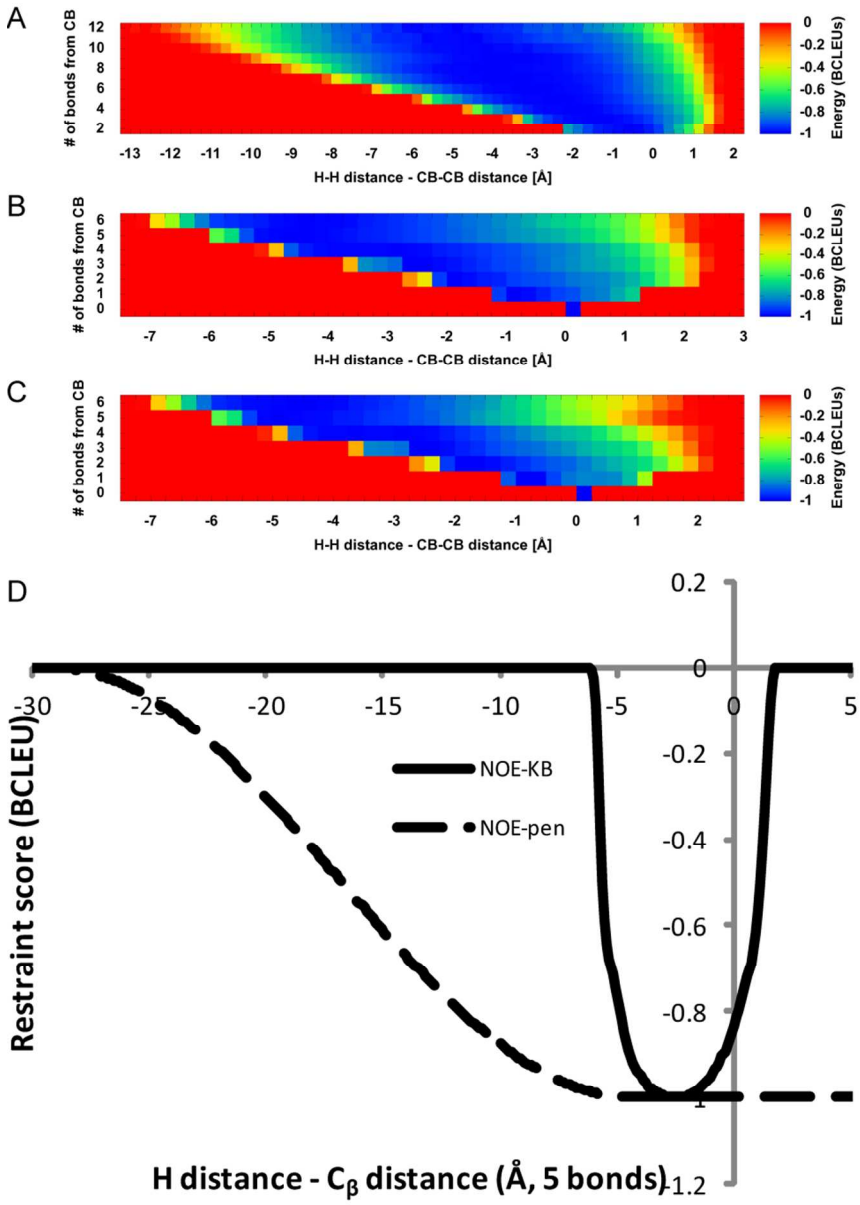


Figure 1. NOE knowledge based potentials. The energy potential for each cumulative bond distance is plotted versus the measured Cβ-Cβ distance subtracted from the expeimental H-H distance. The bond distance is the number of bonds between the measured proton and the Cβ atom of the same residue. For example, an NOE between Hβ3 and Hδ2 would have a cumulative bond distance of four. (A) Potentials for side chain-side chain NOEs. (B) Potentials Hα-side chain NOEs. (C) Potentials for backbone amide H-side chain NOEs. (D) The NOE-KB and NOE-pen potentials are plotted for a cumulative bond distance of 5.
82x115mm (300 x 300 DPI)

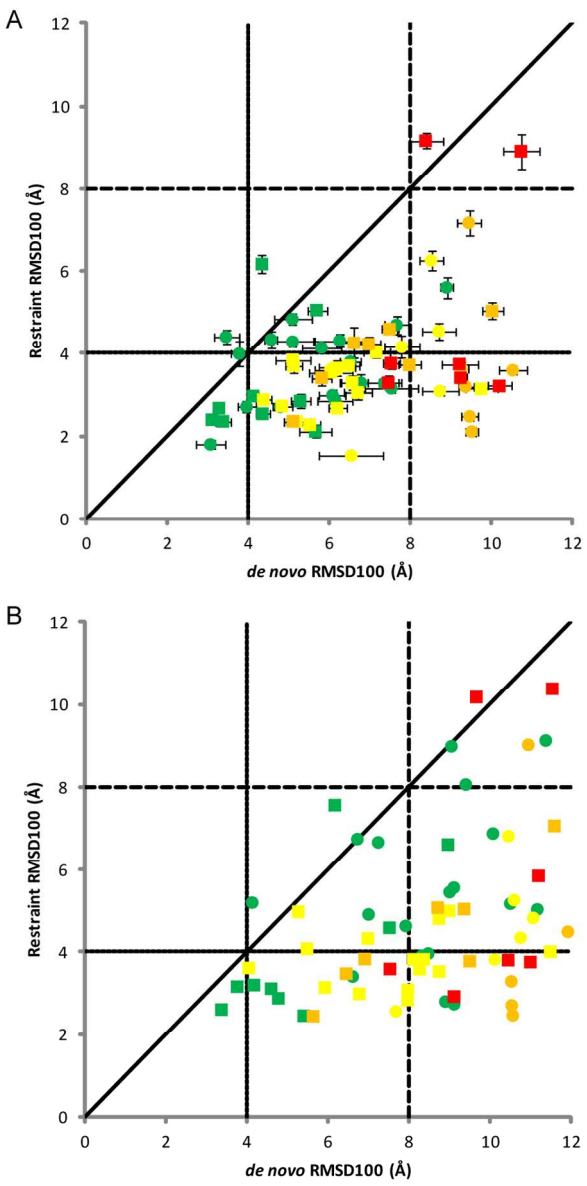


Figure 2. NMR restraints improve native-like sampling. (A) The mean RMSD100 values of the best 10 models sampled with and without restraints are plotted. Soluble proteins are represented by circles and membrane proteins by squares. Proteins are colored according to size: < 150 residues (green), ≥ 150 and < 250 residues (yellow), ≥ 250 and < 400 residues (orange), and ≥ 400 residues (red). The dashed line at 8.0 Å indicates the cutoff for the correct topology, and the dashed line at 4.0 Å indicates a feasible target for continuing with full atom refinement. The error bars are ± 1 S.D. (B) Of the top 10 models by score, the RMSD100 value of the best model is plotted for folding with and without restraints. Marker shapes and colors are the same as in panel A.
84x168mm (300 x 300 DPI)

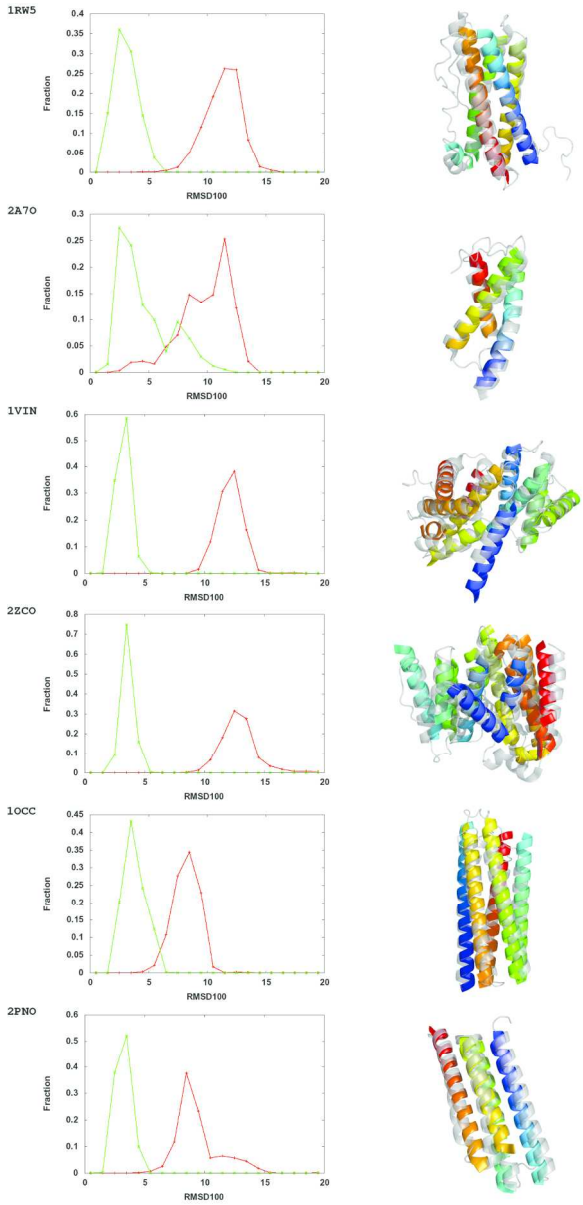


Figure 3. Gallery of select benchmark results. Left column – Distribution of RMSD100 to native SSE values for models produced by the de novo method (red) and the restraint-based method (green). Right column – Superimposition of the best model produced by the restraint method (rainbow) with the native protein (gray). Refer to the supporting information for the complete gallery of benchmark results.
127x240mm (300 x 300 DPI)

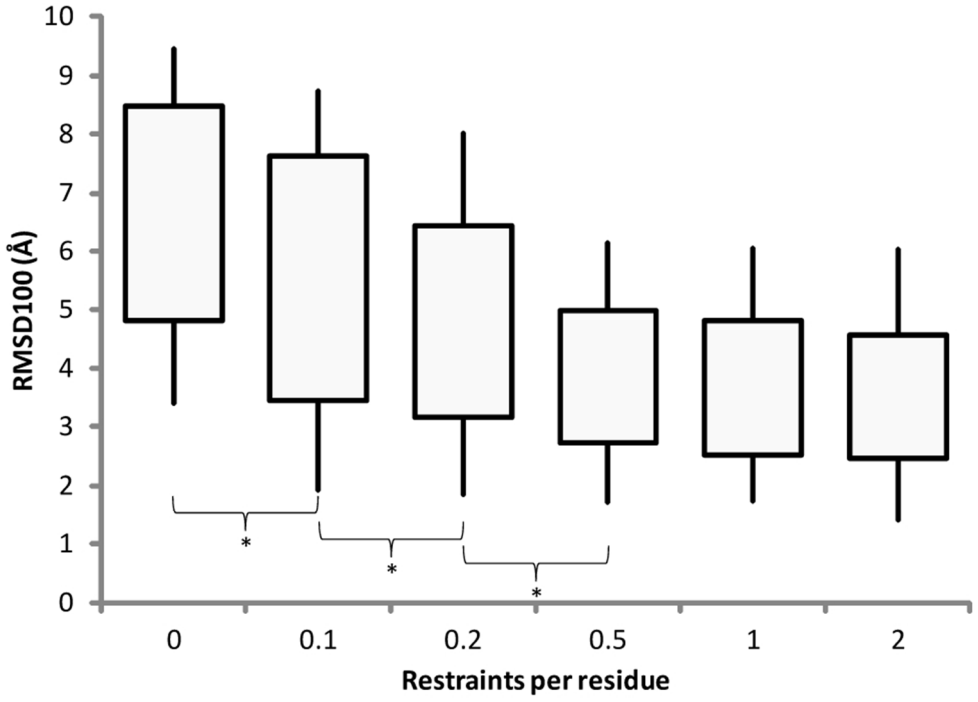


Figure 4. Sampling efficiency depends upon restraint density. The size of the random subset of NOEs selected for folding was adjusted relative to the total number of residues in native SSEs. Each of the 23 proteins with experimental data was folded at varying restraint densities (0.0, 0.1, 0.2, 0.5, 1.0, and 2.0 restraints/residue). The distribution of the mean RMSD100 for the top 5% (selected by RMSD100) of models for each benchmark protein are shown. The boxes contain values within one standard deviation of the mean (of mean RMSD100 values) and the lines represent the minimum and maximum values observed from the 23 proteins for that restraint density. *Improvement over previous restraint density ($p < 0.01$).
84x61mm (300 x 300 DPI)

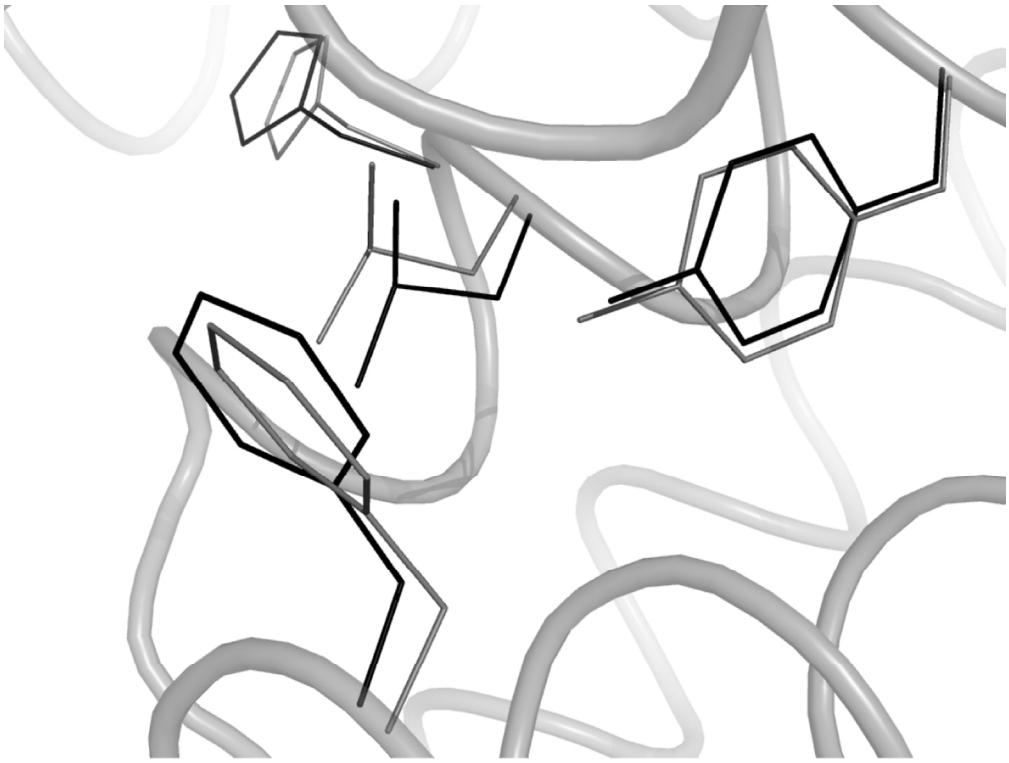


Figure 5. Core side chain conformations can be accurately predicted. Native protein model 1VIN is shown in gray, with side chain atoms displayed for His63, Leu64, Tyr68, and Phe97. The corresponding side chains from the best scoring Rosetta model after full-atom refinement are shown in black.
101x76mm (300 x 300 DPI)

BCL::Fold – Protein topology determination from limited NMR restraints

Brian E. Weiner, Nathan Alexander, Louesa R. Akin, Nils Woetzel, Mert Karakas, and *Jens Meiler

Department of Chemistry, Center for Structural Biology, Vanderbilt University, Nashville TN, 37232, USA

Supporting Materials and Methods

Creating an SSE pool

The BCL application, "CreateSSEPool" is used to create pools from TALOS+ predictions. A sample command line is:

```
./bcl.exe CreateSSEPool -ssmethods TALOS -pool_min_sse_lengths 5 3 -sse_threshold 0.0 0.0 0.0 -chain_id -prefix input/1CMZA -join_separate -factory SSPredMC
```

This will create a pool for protein 1CMZ using the TALOS+ predictions. The "input" folder must contain 1CMZA.fasta and 1CMZASS.tab.

Folding with NMR restraints

Below is a sample command line for using BCL::Fold in combination with sparse NMR data to predict protein structure:

```
./bcl.exe Fold -nmodels 100 –native input/1CMZ.pdb -pool_separate 1 -pool input/1CMZA_TALOS.pool -sspred TALOS JUFO PSIPRED -sspred_path_prefix input 1CMZ -pool_min_sse_lengths 5 3 -mc_temperature_fraction 0.5 0.2 500 10 -quality RMSD GDT_TS -superimpose RMSD -stages_read stages.txt -function_cache -message_level Critical -protein_storage output/ Overwrite -restraint_types NOE RDC -restraint_prefix input/1CMZ -prefix 1CMZA -random_seed 1
```

Input files are placed in the "input" folder. This command will generate 100 models in the "output" folder. The "input" folder should contain:

- 1CMZ.pdb – Native PDB file for quality measurements. Use "-fasta" with a FASTA formatted file if no native model is available.
- 1CMZ_OCT.pool – Pool generated from TALOS+ predictions, which for this example, contains:

```
bcl::assemble::SSEPool

HELIX   1   1 PRO A   14  TRP A   20  1                          7

HELIX   2   2 PRO A   31  THR A   43  1                         13

HELIX   3   3 GLU A   47  LYS A   60  1                         14

HELIX   4   4 GLN A   65  TYR A   79  1                         15

HELIX   5   5 SER A   92  LYS A  101  1                         10

HELIX   6   6 PHE A  110  LEU A  130  1                         21

HELIX   7   7 PRO A  133  LEU A  138  1                          6

END
```

- 1CMZA.jufo – JUFO secondary structure predictions
- 1CMZA.psipred – PSIPRED secondary structure predictions
- 1CMZASS.tab – TALOS+ secondary structure predictions
- stages.txt – Stage file, which contains:

```
NUMBER_CYCLES 1

STAGE Stage_assembly_1

 SCORE_PROTOCOLS Default Restraint

 SCORE_WEIGHTSET_FILE input/assembly_01.scoreweights

 MUTATE_PROTOCOLS Default Assembly

 NUMBER_ITERATIONS 2000 400

STAGE_END

STAGE Stage_assembly_1

 SCORE_PROTOCOLS Default Restraint

 SCORE_WEIGHTSET_FILE input/assembly_02.scoreweights

 MUTATE_PROTOCOLS Default Assembly

 NUMBER_ITERATIONS 2000 400
```

```
STAGE_END

STAGE Stage_assembly_3

 SCORE_PROTOCOLS Default Restraint

 SCORE_WEIGHTSET_FILE input/assembly_03.scoreweights

 MUTATE_PROTOCOLS Default Assembly

 NUMBER_ITERATIONS 2000 400

STAGE_END

STAGE Stage_assembly_4

 SCORE_PROTOCOLS Default Restraint

 SCORE_WEIGHTSET_FILE input/assembly_04.scoreweights

 MUTATE_PROTOCOLS Default Assembly

 NUMBER_ITERATIONS 2000 400

STAGE_END

STAGE Stage_assembly_5

 SCORE_PROTOCOLS Default Restraint

 SCORE_WEIGHTSET_FILE input/assembly_05.scoreweights

 MUTATE_PROTOCOLS Default Assembly

 NUMBER_ITERATIONS 2000 400

STAGE_END

STAGE Stage_refinement

 SCORE_PROTOCOLS Default Restraint

 SCORE_WEIGHTSET_FILE input/refine.scoreweights

 MUTATE_PROTOCOLS Default Refinement

 NUMBER_ITERATIONS 4000 400

STAGE_END
```

- assembly_0[1..5].scoreweights – Table of score weights for each assembly stage. The contents of the first file are shown below; subsequent files differ by increasing aaclash and sseclash by 125 for each new stage. The file contains two lines, headers and weights.

```
bcl::storage::Table<double>  aaclash  aadist  aaneigh  aaneigh_ent  loop  loop_closure_gradient  rgyr  sseclash  ssepack_fr  strand_fr  co_score
ss_PSIPRED  ss_PSIPRED_ent  ss_JUFO  ss_JUFO_ent  noe_restraint  noe_penalty  rdc_restraint  ss_TALOS  ss_TALOS_ent

weights  0  0.35  50  50.0  10.0  50000  5.0  0  8.0  20  0.5  20.0  20.0  5.0  5.0  500  500  500  10  10
```

- refinement.scoreweights – Table of score weights for the refinement stage, which contains:

```
bcl::storage::Table<double>  aaclash  aadist  aaneigh  aaneigh_ent  loop  loop_closure_gradient  rgyr  sseclash  ssepack_fr  strand_fr  co_score
ss_PSIPRED  ss_PSIPRED_ent  ss_JUFO  ss_JUFO_ent  noe_restraint  noe_penalty  rdc_restraint  ss_TALOS  ss_TALOS_ent

weights  500  0.35  50  50.0  10.0  50000  5.0  500  8.0  20  0.5  20.0  20.0  5.0  5.0  5  5  5  10  10
```

- 1CMZ.noe_star – NOE restraints in NMR-STAR 3.1 format.
- 1CMZ.rdc_star – RDC restraints in NMR-STAR 3.1 format.

NOE-KB histograms

Displayed below is the histogram file used by BCL::Fold for the NOE-KB score. The following histograms first list the BCL atom type corresponding the observed distance. “CB” is a side chain-side chain distance. “H” is a backbone amide proton-side chain distance. “HA” is an H_α-side chain distance. The following line is the sum of bonds from the side chain atoms to the corresponding C_β. Then the bin centers are listed followed by the observed counts for the given atom types and bond distance. The bin refers to the H-H distance - C_β-C_β distance. This raw data is converted to an energy potential using the inverse Boltzmann relation.

```
bcl::biol::AtomTypes::Enum
  "CB"
2
bcl::math::Histogram
          ...<    <..>    <..>    <..>    <..>    <..>    <..>    <..>    <..>    <..>    <..>    <..>
        <..>    <..>    <..>    <..>    <..>    <..>    <..>    >...
  center    -2.500   -2.375   -2.125   -1.875   -1.625   -1.375   -1.125   -0.875   -0.625   -0.375   -0.125    0.125
        0.375    0.625    0.875    1.125    1.375    1.625    1.875    2.000
  counts    0.000    0.000 31146.000 50547.000 85134.000 116487.000          133230.000          140441.000
        156505.000          147030.000          125014.000          90527.000 45714.000 22612.000 8622.000  1821.000   169.000     6.000
           0.000    0.000

bcl::biol::AtomTypes::Enum
  "CB"
3
bcl::math::Histogram
          ...<    <..>    <..>    <..>    <..>    <..>    <..>    <..>    <..>    <..>    <..>    <..>
        <..>    <..>    <..>    <..>    <..>    <..>    <..>    <..>    <..>    <..>    <..>    <..>
        <..>    >...
```

```
  center             -4.000    -3.875    -3.625    -3.375    -3.125    -2.875    -2.625    -2.375    -2.125    -1.875    -1.625    -1.375
           -1.125    -0.875    -0.625    -0.375    -0.125     0.125     0.375     0.625     0.875     1.125     1.375     1.625     1.875
            2.125     2.250
  counts              0.000     0.000    55.000   540.000  29726.000 87533.000 148012.000          197587.000          226916.000
         244783.000          253551.000          258450.000          255853.000          235557.000          209837.000
         181385.000          146896.000          103488.000          72117.000 50464.000 27561.000 9698.000  1543.000    79.000     2.000
           0.000     0.000

bcl::biol::AtomTypes::Enum
  "CB"
4
bcl::math::Histogram
                  ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
           <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
           <..>      <..>      <..>      <..>      <..>      <..>      >...
  center             -5.250    -5.125    -4.875    -4.625    -4.375    -4.125    -3.875    -3.625    -3.375    -3.125    -2.875    -2.625
           -2.375    -2.125    -1.875    -1.625    -1.375    -1.125    -0.875    -0.625    -0.375    -0.125     0.125     0.375     0.625
            0.875     1.125     1.375     1.625     1.875     2.125     2.250
  counts              0.000     0.000     2.000   464.000  8263.000  35122.000 75956.000 133996.000          196068.000
         255676.000          299176.000          325759.000          338233.000          344543.000          348726.000
         345106.000          323100.000          290098.000          251653.000          215081.000          179098.000
         142763.000          107127.000          73637.000 46783.000 25055.000 9501.000  1725.000   109.000     4.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "CB"
5
bcl::math::Histogram
                  ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
           <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
           <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      >...
  center             -6.500    -6.375    -6.125    -5.875    -5.625    -5.375    -5.125    -4.875    -4.625    -4.375    -4.125    -3.875
           -3.625    -3.375    -3.125    -2.875    -2.625    -2.375    -2.125    -1.875    -1.625    -1.375    -1.125    -0.875    -0.625
           -0.375    -0.125     0.125     0.375     0.625     0.875     1.125     1.375     1.625     1.875     2.000
  counts              0.000     0.000     4.000   145.000  4103.000  17312.000 33612.000 56121.000 94303.000 131996.000
         167793.000          199082.000          225422.000          242961.000          257142.000          268348.000
         274669.000          275610.000          272801.000          261601.000          242884.000          222270.000
         198752.000          173014.000          144763.000          114644.000          86050.000 61461.000 42523.000 27344.000 14109.000
         4843.000   780.000    62.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "CB"
6
bcl::math::Histogram
                  ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
           <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
           <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
           <..>      <..>      >...
  center             -7.500    -7.375    -7.125    -6.875    -6.625    -6.375    -6.125    -5.875    -5.625    -5.375    -5.125    -4.875
           -4.625    -4.375    -4.125    -3.875    -3.625    -3.375    -3.125    -2.875    -2.625    -2.375    -2.125    -1.875    -1.625
           -1.375    -1.125    -0.875    -0.625    -0.375    -0.125     0.125     0.375     0.625     0.875     1.125     1.375     1.625
            1.875     2.125     2.250
  counts              0.000     0.000    22.000   788.000  4556.000  10562.000 20444.000 33862.000 47886.000 65755.000 85477.000
         104880.000          121401.000          135117.000          147036.000          156653.000          161817.000
         164682.000          166578.000          167418.000          165436.000          161498.000          154666.000
         147480.000          137977.000          124577.000          109472.000          93488.000 76787.000 61116.000 45997.000 32823.000
         22259.000 14145.000 7149.000  2162.000   308.000    17.000     5.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "CB"
7
bcl::math::Histogram
                  ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
           <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
           <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
           <..>      <..>      <..>      <..>      <..>      <..>      >...
  center             -8.750    -8.625    -8.375    -8.125    -7.875    -7.625    -7.375    -7.125    -6.875    -6.625    -6.375    -6.125
           -5.875    -5.625    -5.375    -5.125    -4.875    -4.625    -4.375    -4.125    -3.875    -3.625    -3.375    -3.125    -2.875
           -2.625    -2.375    -2.125    -1.875    -1.625    -1.375    -1.125    -0.875    -0.625    -0.375    -0.125     0.125     0.375
            0.625     0.875     1.125     1.375     1.625     1.875     2.000
  counts              0.000     0.000     9.000   119.000   915.000  2003.000  4935.000  9854.000  16332.000 23256.000 30363.000 37157.000
         44194.000 51180.000 58012.000 64300.000 68566.000 71321.000 73116.000 74210.000 75133.000 74976.000 74787.000 74035.000 73525.000
         71395.000 70136.000 67122.000 64066.000 61562.000 57073.000 51214.000 44617.000 37091.000 29148.000 22484.000 16358.000 11340.000
         7280.000  3683.000  1242.000   202.000    19.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "CB"
8
bcl::math::Histogram
                  ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
           <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
           <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
           <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      >...
  center             -9.750    -9.625    -9.375    -9.125    -8.875    -8.625    -8.375    -8.125    -7.875    -7.625    -7.375    -7.125
           -6.875    -6.625    -6.375    -6.125    -5.875    -5.625    -5.375    -5.125    -4.875    -4.625    -4.375    -4.125    -3.875
           -3.625    -3.375    -3.125    -2.875    -2.625    -2.375    -2.125    -1.875    -1.625    -1.375    -1.125    -0.875    -0.625
           -0.375    -0.125     0.125     0.375     0.625     0.875     1.125     1.375     1.625     1.875     2.125     2.250
  counts              0.000     0.000    18.000   165.000   469.000  1002.000  1848.000  3348.000  5632.000  8322.000  11034.000 13673.000
         16486.000 19357.000 22559.000 25453.000 27733.000 30133.000 31753.000 33363.000 34376.000 35276.000 35770.000 35847.000 36123.000
         35807.000 35074.000 34668.000 33779.000 33147.000 31847.000 30761.000 29241.000 27652.000 25540.000 22955.000 20008.000 16623.000
         13130.000 9685.000  6940.000  4739.000  2867.000  1462.000   532.000    85.000     6.000     1.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "CB"
9
bcl::math::Histogram
                  ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
           <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
           <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
```

```
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      >...
  center              -11.000   -10.875   -10.625   -10.375   -10.125    -9.875    -9.625    -9.375    -9.125    -8.875    -8.625    -8.375
            -8.125    -7.875    -7.625    -7.375    -7.125    -6.875    -6.625    -6.375    -6.125    -5.875    -5.625    -5.375    -5.125
            -4.875    -4.625    -4.375    -4.125    -3.875    -3.625    -3.375    -3.125    -2.875    -2.625    -2.375    -2.125    -1.875
            -1.625    -1.375    -1.125    -0.875    -0.625    -0.375    -0.125     0.125     0.375     0.625     0.875     1.125     1.375
             1.625     1.875     2.000
  counts                0.000     0.000     1.000    12.000    50.000   150.000   361.000   621.000  1136.000  1525.000  2259.000  3085.000
          4127.000  4878.000  5985.000  6939.000  8015.000  9313.000 10262.000 11384.000 12451.000 13608.000 14516.000 15018.000 15752.000
          16919.000 17303.000 17656.000 18055.000 18230.000 18181.000 18176.000 18133.000 17564.000 17253.000 16572.000 16070.000 15209.000
          13770.000 12727.000 11263.000 9634.000  7916.000  6126.000  4583.000  3405.000  2361.000  1530.000   748.000   289.000    41.000
             1.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "CB"
10
bcl::math::Histogram
                ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      >...
  center              -12.000   -11.875   -11.625   -11.375   -11.125   -10.875   -10.625   -10.375   -10.125    -9.875    -9.625    -9.375
            -9.125    -8.875    -8.625    -8.375    -8.125    -7.875    -7.625    -7.375    -7.125    -6.875    -6.625    -6.375    -6.125
            -5.875    -5.625    -5.375    -5.125    -4.875    -4.625    -4.375    -4.125    -3.875    -3.625    -3.375    -3.125    -2.875
            -2.625    -2.375    -2.125    -1.875    -1.625    -1.375    -1.125    -0.875    -0.625    -0.375    -0.125     0.125     0.375
             0.625     0.875     1.125     1.375     1.625     1.875     2.000
  counts                0.000     0.000     1.000     7.000    23.000    60.000   128.000   265.000   432.000   514.000   782.000  1058.000
          1208.000  1556.000  1864.000  2129.000  2426.000  2729.000  2943.000  3249.000  3475.000  3776.000  4023.000  4257.000  4373.000
          4797.000  5020.000  5103.000  5167.000  5244.000  5377.000  5364.000  5519.000  5295.000  5246.000  5018.000  5075.000  4685.000
          4738.000  4539.000  4320.000  4178.000  3953.000  3557.000  3376.000  3025.000  2653.000  2153.000  1629.000  1289.000   971.000
           598.000   282.000    98.000    25.000     1.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "CB"
11
bcl::math::Histogram
                ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      >...
  center              -12.500   -12.375   -12.125   -11.875   -11.625   -11.375   -11.125   -10.875   -10.625   -10.375   -10.125    -9.875
            -9.625    -9.375    -9.125    -8.875    -8.625    -8.375    -8.125    -7.875    -7.625    -7.375    -7.125    -6.875    -6.625
            -6.375    -6.125    -5.875    -5.625    -5.375    -5.125    -4.875    -4.625    -4.375    -4.125    -3.875    -3.625    -3.375
            -3.125    -2.875    -2.625    -2.375    -2.125    -1.875    -1.625    -1.375    -1.125    -0.875    -0.625    -0.375    -0.125
             0.125     0.375     0.625     0.875     1.125     1.375     1.625     1.750
  counts                0.000     0.000     3.000    11.000    24.000    61.000    85.000   130.000   179.000   226.000   320.000   382.000
           453.000   520.000   614.000   673.000   818.000   839.000   886.000  1042.000  1059.000  1188.000  1130.000  1342.000  1353.000
          1426.000  1404.000  1460.000  1593.000  1530.000  1556.000  1554.000  1592.000  1623.000  1504.000  1464.000  1470.000  1407.000
          1334.000  1287.000  1225.000  1234.000  1162.000  1122.000  1165.000  1104.000  1039.000   926.000   841.000   678.000   472.000
           367.000   326.000   153.000    98.000    37.000     3.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "CB"
12
bcl::math::Histogram
                ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      >...
  center              -13.250   -13.125   -12.875   -12.625   -12.375   -12.125   -11.875   -11.625   -11.375   -11.125   -10.875   -10.625
            -10.375   -10.125    -9.875    -9.625    -9.375    -9.125    -8.875    -8.625    -8.375    -8.125    -7.875    -7.625    -7.375
             -7.125    -6.875    -6.625    -6.375    -6.125    -5.875    -5.625    -5.375    -5.125    -4.875    -4.625    -4.375    -4.125
             -3.875    -3.625    -3.375    -3.125    -2.875    -2.625    -2.375    -2.125    -1.875    -1.625    -1.375    -1.125    -0.875
             -0.625    -0.375    -0.125     0.125     0.375     0.625     0.875     1.125     1.375     1.625     1.750
  counts                0.000     0.000     7.000     2.000     8.000    28.000    39.000    66.000    76.000   117.000   136.000   192.000
           226.000   285.000   360.000   372.000   421.000   523.000   553.000   662.000   671.000   758.000   825.000   857.000   897.000
           949.000  1009.000  1087.000  1047.000  1156.000  1184.000  1139.000  1139.000  1170.000  1227.000  1138.000  1120.000  1106.000
          1094.000  1056.000  1045.000   923.000   925.000   882.000   926.000   855.000   891.000   754.000   740.000   655.000   599.000
           495.000   423.000   326.000   237.000   163.000   108.000    45.000    15.000     5.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "H"
0
bcl::math::Histogram
                ...<      <..>      <..>      <..>      >...
  center              -0.250    -0.125     0.125     0.375     0.500
  counts                0.000     0.000  306240.000             0.000     0.000

bcl::biol::AtomTypes::Enum
  "H"
1
bcl::math::Histogram
                ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      >...
  center              -1.500    -1.375    -1.125    -0.875    -0.625    -0.375    -0.125     0.125     0.375     0.625     0.875     1.125
             1.375     1.500
  counts                0.000     0.000  87595.000 208069.000          205215.000          111596.000          66249.000 55240.000 57924.000
          50599.000 28106.000 3222.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "H"
2
bcl::math::Histogram
                ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      >...
```

```
  center              -3.000    -2.875    -2.625    -2.375    -2.125    -1.875    -1.625    -1.375    -1.125    -0.875    -0.625    -0.375
            -0.125     0.125     0.375     0.625     0.875     1.125     1.375     1.625     1.875     2.125     2.375     2.500
  counts              0.000     0.000   295.000  1329.000 79338.000 132643.000          132407.000          118023.000
         100467.000          86022.000 69498.000 58035.000 52464.000 39955.000 29820.000 23063.000 16589.000 11434.000 5949.000  2088.000
           511.000    75.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "H"
3
bcl::math::Histogram
                    ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      >...
  center              -4.500    -4.375    -4.125    -3.875    -3.625    -3.375    -3.125    -2.875    -2.625    -2.375    -2.125    -1.875
            -1.625    -1.375    -1.125    -0.875    -0.625    -0.375    -0.125     0.125     0.375     0.625     0.875     1.125     1.375
             1.625     1.875     2.125     2.375     2.625     2.750
  counts              0.000     0.000     1.000     1.000   422.000  27472.000 30483.000 35312.000 93681.000 92521.000 81600.000 72879.000
          62965.000 51105.000 42534.000 37637.000 34103.000 28569.000 23535.000 17574.000 13305.000 10881.000 8316.000  5116.000  2068.000
            669.000   180.000    20.000     5.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "H"
4
bcl::math::Histogram
                    ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      >...
  center              -5.500    -5.375    -5.125    -4.875    -4.625    -4.375    -4.125    -3.875    -3.625    -3.375    -3.125    -2.875
            -2.625    -2.375    -2.125    -1.875    -1.625    -1.375    -1.125    -0.875    -0.625    -0.375    -0.125     0.125     0.375
             0.625     0.875     1.125     1.375     1.625     1.875     2.125     2.375     2.625     2.750
  counts              0.000     0.000    28.000   282.000 10606.000 17419.000 18949.000 19380.000 19728.000 19092.000 19347.000 19800.000
          18951.000 17984.000 16584.000 14760.000 13331.000 11242.000 9434.000  8032.000  6799.000  5519.000  4776.000  4252.000  2993.000
          2127.000  1391.000   755.000   366.000   198.000    60.000    18.000     2.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "H"
5
bcl::math::Histogram
                    ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      >...
  center              -6.250    -6.125    -5.875    -5.625    -5.375    -5.125    -4.875    -4.625    -4.375    -4.125    -3.875    -3.625
            -3.375    -3.125    -2.875    -2.625    -2.375    -2.125    -1.875    -1.625    -1.375    -1.125    -0.875    -0.625    -0.375
            -0.125     0.125     0.375     0.625     0.875     1.125     1.375     1.625     1.875     2.125     2.375     2.625     2.750
  counts              0.000     0.000  1064.000  1587.000  5445.000  7613.000  7822.000  7534.000  7330.000  7160.000  6900.000  6551.000
          6090.000  5787.000  5595.000  5041.000  4720.000  4072.000  3657.000  3065.000  2588.000  2273.000  1905.000  1470.000  1111.000
           863.000   737.000   554.000   317.000   172.000    97.000    41.000    15.000     8.000     4.000     1.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "H"
6
bcl::math::Histogram
                    ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      >...
  center              -7.500    -7.375    -7.125    -6.875    -6.625    -6.375    -6.125    -5.875    -5.625    -5.375    -5.125    -4.875
            -4.625    -4.375    -4.125    -3.875    -3.625    -3.375    -3.125    -2.875    -2.625    -2.375    -2.125    -1.875    -1.625
            -1.375    -1.125    -0.875    -0.625    -0.375    -0.125     0.125     0.375     0.625     0.875     1.125     1.375     1.625
             1.875     2.125     2.375     2.625     2.750
  counts              0.000     0.000     7.000   286.000   711.000  2293.000  4416.000  6846.000  7415.000  7814.000  8031.000  8362.000
          8517.000  8590.000  8418.000  8129.000  7558.000  6977.000  6444.000  6028.000  5459.000  4900.000  4545.000  3957.000  3596.000
          3036.000  2332.000  1893.000  1496.000  1076.000   813.000   646.000   517.000   418.000   317.000   183.000    95.000    51.000
             7.000     3.000     1.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "HA"
0
bcl::math::Histogram
                    ...<      <..>      <..>      <..>      >...
  center              -0.250    -0.125     0.125     0.375     0.500
  counts              0.000     0.000  1125600.000            0.000     0.000

bcl::biol::AtomTypes::Enum
  "HA"
1
bcl::math::Histogram
                    ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      >...
  center              -1.500    -1.375    -1.125    -0.875    -0.625    -0.375    -0.125     0.125     0.375     0.625     0.875     1.125
             1.375     1.500
  counts              0.000     0.000   138735.000          252203.000          213171.000          175214.000          134758.000
         115447.000          78461.000 66429.000 39990.000 24668.000    0.000     0.000

bcl::biol::AtomTypes::Enum
  "HA"
2
bcl::math::Histogram
                    ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      >...
  center              -3.000    -2.875    -2.625    -2.375    -2.125    -1.875    -1.625    -1.375    -1.125    -0.875    -0.625    -0.375
            -0.125     0.125     0.375     0.625     0.875     1.125     1.375     1.625     1.875     2.125     2.375     2.625     2.750
  counts              0.000     0.000   320.000  1390.000  109270.000          193935.000          170375.000          152456.000
         136466.000          109328.000          95665.000 92318.000 77404.000 62613.000 55519.000 50589.000 40548.000 28678.000 15987.000
          7575.000  3528.000   532.000     1.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "HA"
3
```

```
bcl::math::Histogram
                ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      >...
  center              -4.250    -4.125    -3.875    -3.625    -3.375    -3.125    -2.875    -2.625    -2.375    -2.125    -1.875    -1.625
              -1.375    -1.125    -0.875    -0.625    -0.375    -0.125     0.125     0.375     0.625     0.875     1.125     1.375     1.625
               1.875     2.125     2.375     2.625     2.750
  counts              0.000     0.000     1.000   486.000  34917.000 42496.000 51350.000 133932.000          132579.000
         122487.000          111280.000          96763.000 84433.000 75031.000 66669.000 58072.000 49619.000 41456.000 34143.000 28769.000
         24444.000 19993.000 13728.000 8337.000  5256.000  2322.000   573.000    56.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "HA"
4
bcl::math::Histogram
                ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      >...
  center              -5.500    -5.375    -5.125    -4.875    -4.625    -4.375    -4.125    -3.875    -3.625    -3.375    -3.125    -2.875
              -2.625    -2.375    -2.125    -1.875    -1.625    -1.375    -1.125    -0.875    -0.625    -0.375    -0.125     0.125     0.375
               0.625     0.875     1.125     1.375     1.625     1.875     2.125     2.375     2.625     2.875     3.000
  counts              0.000     0.000    25.000   364.000  15193.000 23937.000 26183.000 27989.000 27746.000 27410.000 27613.000 28842.000
         27559.000 26106.000 24365.000 22644.000 21176.000 19595.000 17645.000 16031.000 14351.000 12685.000 11583.000 10164.000 8375.000
         6734.000  5074.000  3674.000  2519.000  1633.000   721.000   227.000    21.000     1.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "HA"
5
bcl::math::Histogram
                ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          >...
  center              -6.500    -6.375    -6.125    -5.875    -5.625    -5.375    -5.125    -4.875    -4.625    -4.375    -4.125    -3.875
            -3.625    -3.375    -3.125    -2.875    -2.625    -2.375    -2.125    -1.875    -1.625    -1.375    -1.125    -0.875    -0.625
            -0.375    -0.125     0.125     0.375     0.625     0.875     1.125     1.375     1.625     1.875     2.125     2.375     2.625
             2.750
  counts              0.000     0.000     2.000  1310.000  2208.000  6858.000 10135.000 10737.000 11104.000 11146.000 10851.000 10252.000
         9905.000  9443.000  8978.000  8437.000  8014.000  7460.000  6934.000  6301.000  5732.000  5306.000  4932.000  4359.000  4130.000
         3645.000  3291.000  2878.000  2398.000  1896.000  1501.000  1072.000   733.000   374.000   173.000    31.000     2.000     0.000
            0.000

bcl::biol::AtomTypes::Enum
  "HA"
6
bcl::math::Histogram
                ...<      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>      <..>
          <..>      <..>      <..>      <..>      >...
  center              -7.500    -7.375    -7.125    -6.875    -6.625    -6.375    -6.125    -5.875    -5.625    -5.375    -5.125    -4.875
            -4.625    -4.375    -4.125    -3.875    -3.625    -3.375    -3.125    -2.875    -2.625    -2.375    -2.125    -1.875    -1.625
            -1.375    -1.125    -0.875    -0.625    -0.375    -0.125     0.125     0.375     0.625     0.875     1.125     1.375     1.625
             1.875     2.125     2.375     2.625     2.750
  counts              0.000     0.000     6.000   426.000   874.000  2785.000  5849.000  9015.000  9696.000 10303.000 11002.000 11683.000
         12583.000 12954.000 12562.000 12407.000 11893.000 11233.000 10582.000 9920.000  9155.000  8493.000  7817.000  7097.000  6438.000
         5899.000  5269.000  4732.000  4064.000  3389.000  2940.000  2620.000  2201.000  1862.000  1457.000  1098.000   667.000   344.000
          120.000    33.000     5.000     0.000     0.000

bcl::biol::AtomTypes::Enum
  "Undefined"
```

Rosetta Folding

The soluble proteins in the benchmark set were folded using Rosetta with the available NMR data. Fragments were generated using any available CS data with homologs excluded. 1000 models were generated for each target using the AbinitioRelax application. RMSD100 was calculated (using the BCL application, ScoreProtein) to native SSEs to allow for a direct comparison to BCL::Fold-produced topologies. An example command line is:

```
./AbinitioRelax.linuxgccrelease -out:nstruct 100 -out:output -out:overwrite -in:file:fasta input/1CMZA.fasta -in:file:frag3 input/cs_aa1CMZA03_06.200_v1_3 -in:file:frag9
input/cs_aa1CMZA09_06.200_v1_3 -in:file:native input/1CMZA.pdb -stage2_patch input/weights.wts -stage3a_patch input/weights.wts -stage3b_patch input/weights.wts -
stage4_patch input/weights.wts -abinitio:rg_reweight 0.5 -abinitio:rg_reweight 0.5 -abinitio:rsd_wt_helix 0.5 -abinitio:rsd_wt_loop 0.5 -abinitio:fastrelax -
```

```
residues:patch_selectors CENTROID_HA -in:file:rdc input/1CMZA.dpl -score:weights score12_full -score:patch input/weights.wts -in:path:database ./rosetta_database -
constraints:cst_file input/1CMZA_0.cst -out:user_tag cst_1000_0 -out:file:silent output/1CMZA_cst_1000_0.silent -out:sf output/1CMZA_cst_1000_0.score -
run:constant_seed -run:jran 1
```

Input files are placed in the "input" folder. This command will generate 100 models in the "output" folder. The "input" folder should contain:

- 1CMZA.fasta – FASTA file
- cs_aa1CMZA0[3,9]_06.200_v1_3 – Fragement files generated from make_fragments.pl
- weights.wts – Weights file:

```
rdc = 1.0

atom_pair_constraint = 1.0
```

- 1CMZA.dpl –Rosetta formated RDC restraints
- 1CMZA_0.cst – Rosetta formated NOE constraints. Side chain NOE restraints were converted to C_β restraints by adding 1.0 Å to the restraint distance per bond from the side chain proton to the C_β.

Supporting Results

Table S1. Contribution of random data sets to best first round models.

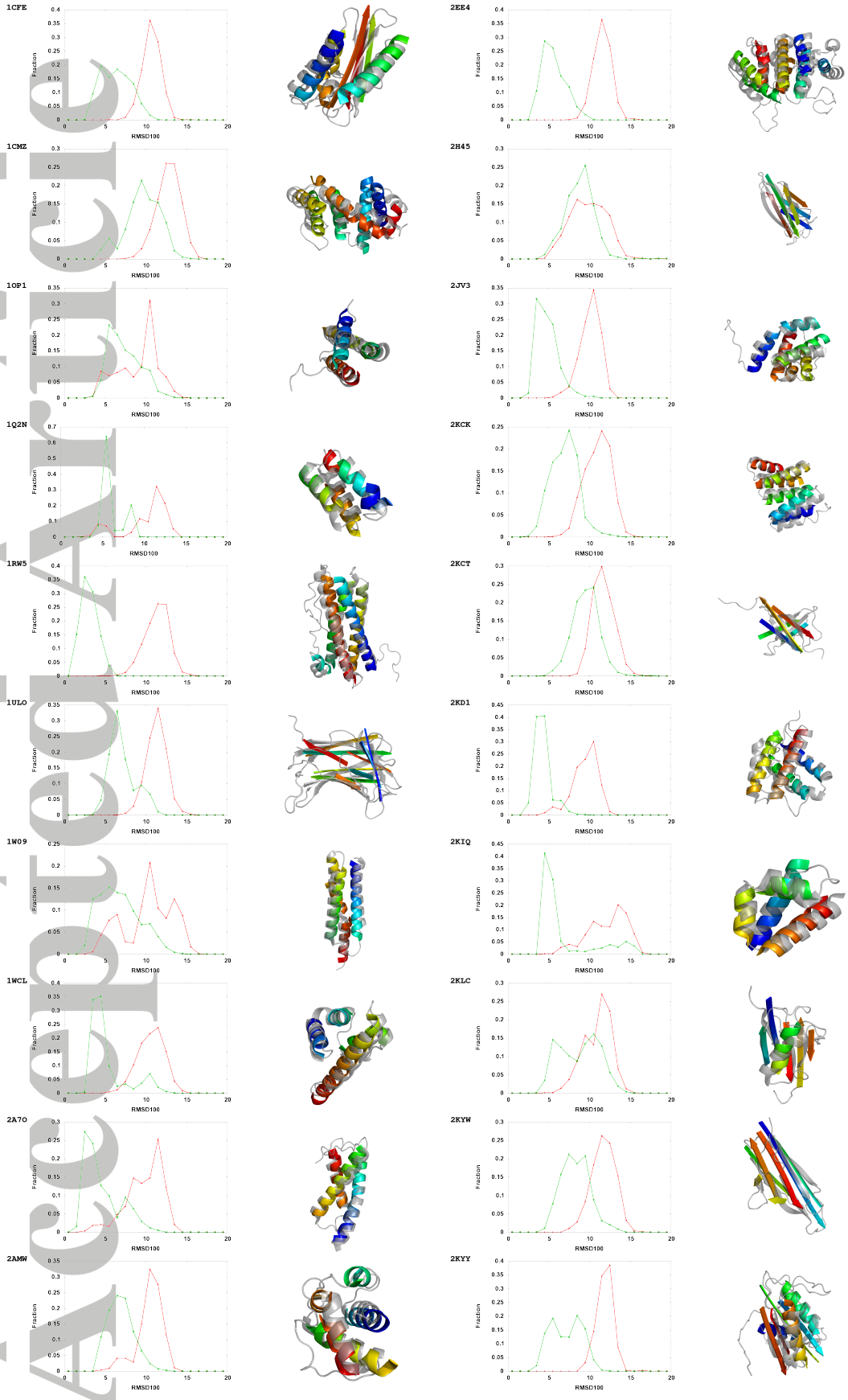
Protein	Best	Worst
1CFE	22%	4%
1CMZ	14%	6%
1OP1	14%	6%
1Q2N	22%	2%
1RW5	18%	4%
1ULO	18%	4%
1W09	22%	2%
1WCL	18%	2%
2A7O	18%	2%
2AMW	16%	6%
2EE4	14%	4%
2H45	16%	4%
2JV3	16%	0%
2KCK	22%	4%
2KCT	20%	2%
2KD1	18%	2%
2KIQ	20%	0%
2KLC	16%	6%
2KYW	26%	4%
2KYY	20%	4%
2L3W	16%	0%
2L7K	16%	2%
2L9R	26%	2%
Mean	19%	3%
SD	3%	2%

Of the top 5% of models (by RMSD100) produced in the first round of folding, the contribution of the data set contributing the most ("Best") and least ("Worst") are shown.

Table S2. Top 5% of models by RMSD100 produced by BCL::Fold and Rosetta with restraints.

Protein	BCL::Fold (Å)	Rosetta (Å)
1CFE	3.2	5.8
1CMZ	5.0	4.6
1OP1	3.2	3.5
1Q2N	4.7	4.0
1RW5	1.8	5.3
1ULO	4.6	5.9
1W09	2.0	2.2
1WCL	3.2	1.9
2A7O	2.1	3.4
2AMW	4.5	4.6
2EE4	3.5	4.6
2H45	6.1	7.6
2JV3	3.2	4.6
2KCK	3.8	3.7
2KCT	4.6	8.4
2KD1	2.8	4.7
2KIQ	4.3	2.6
2KLC	4.4	3.9
2KYW	5.3	6.8
2KYY	3.6	8.2
2L3W	3.3	4.9
2L7K	3.9	3.9
2L9R	3.5	4.7
1UAI	6.7	8.8
1VIN	2.3	4.8
1XQO	7.6	5.6
2OF3	3.9	3.5
2R0S	3.4	6.9
2ZCO	2.7	3.4
Mean	3.9	4.9
SD	1.4	1.8

The mean RSMD100 of the top 5% of models (selected by RMSD100) are shown for BCL::Fold and Rosetta with sparse NMR restraints.



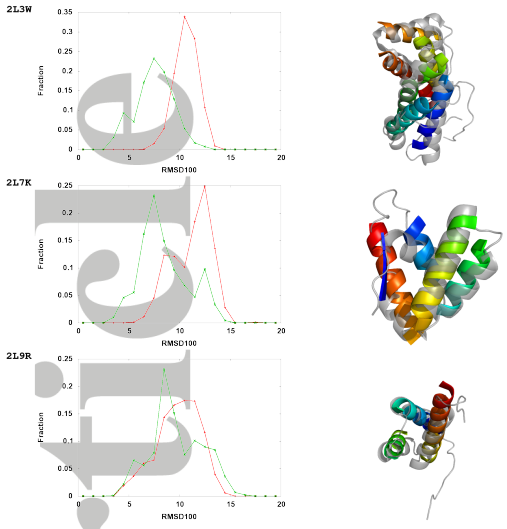


Figure S1. Gallery of benchmark results with experimental data. Left column – Distribution of RMSD100 to native SSE values for models produced by the de novo method (red) and the restraint-based method (green). Right column – Superimposition of the best model produced by the restraint method (rainbow) with the native protein (gray). Refer to the supplementary information for the complete gallery of benchmark results.

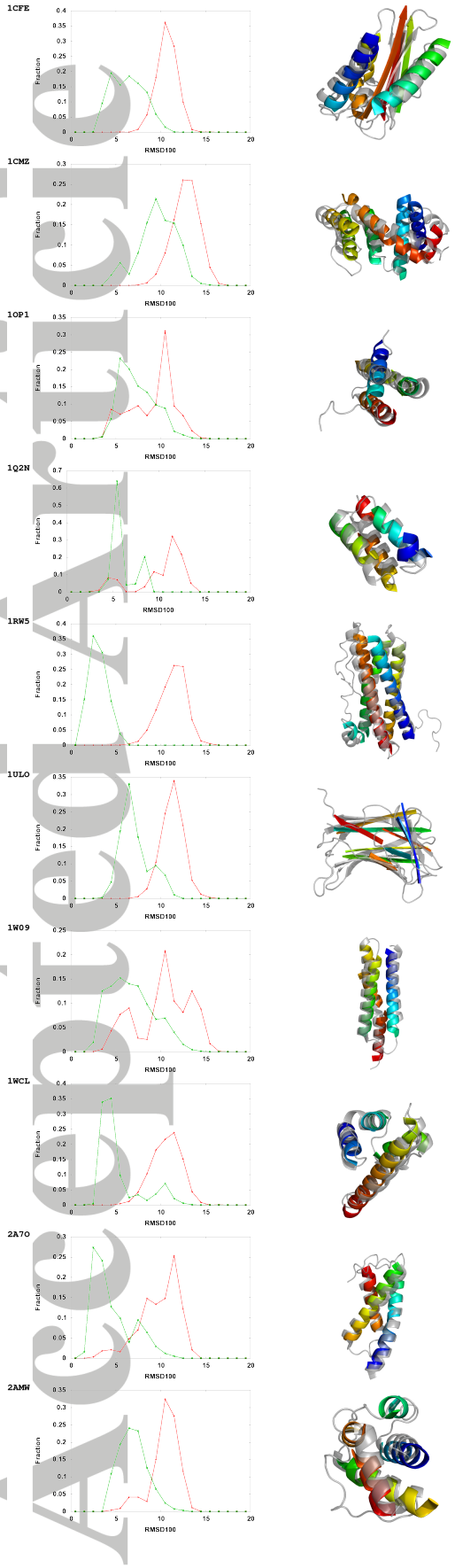
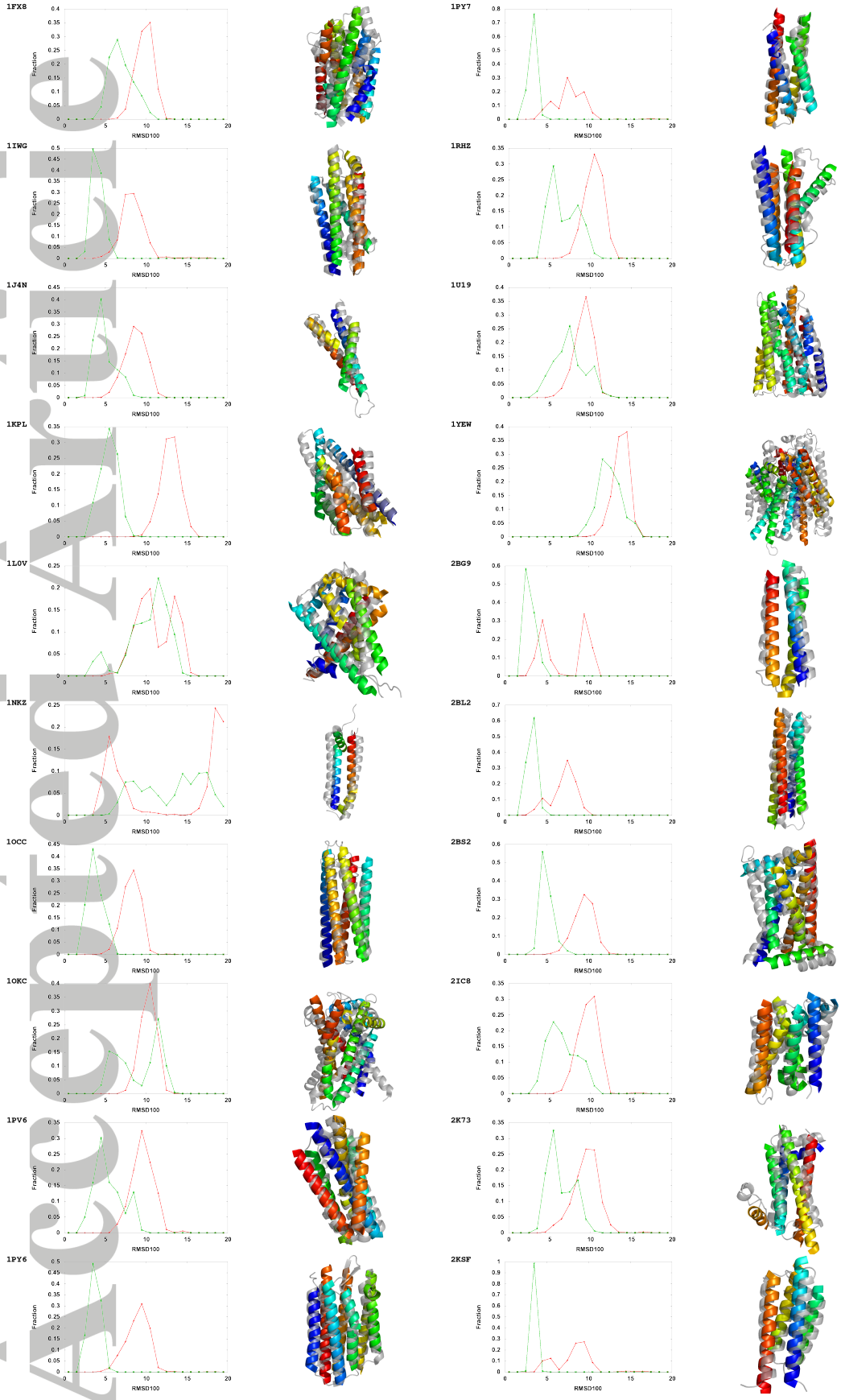


Figure S2. Gallery of soluble protein benchmark results with simulated data. Left column – Distribution of RMSD100 to native SSE values for models produced by the de novo method (red) and the restraint-based method (green). Right column – Superimposition of the best model produced by the restraint method (rainbow) with the native protein (gray). Refer to the supplementary information for the complete gallery of benchmark results.



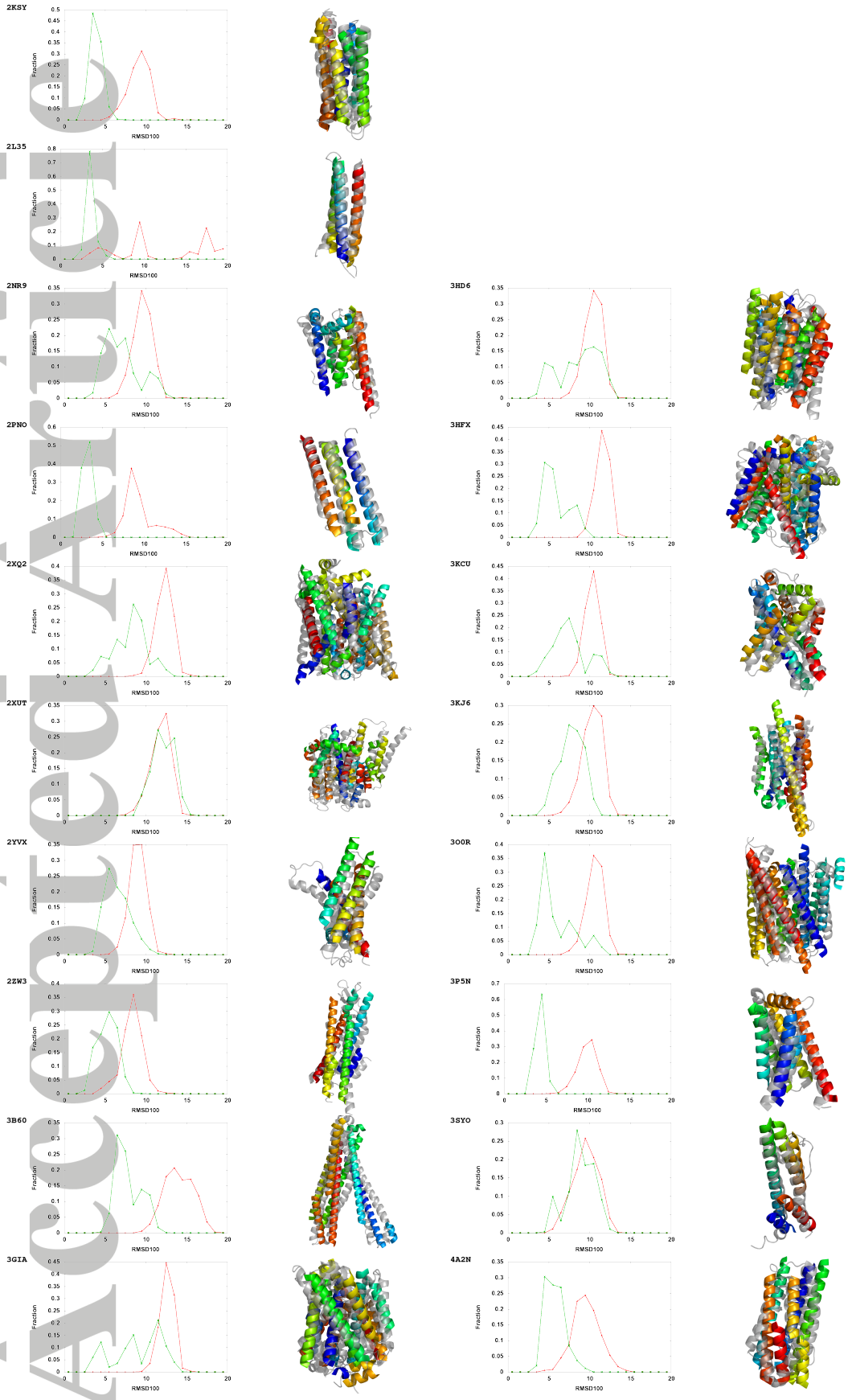


Figure S3. Gallery of membrane protein benchmark results with simulated data. Left column – Distribution of RMSD100 to native SSE values for models produced by the de novo method (red) and the restraint-based method (green). Right column – Superimposition of the best model produced by the restraint method (rainbow) with the native protein (gray). Refer to the supplementary information for the complete gallery of benchmark results.