A coarse-grained approach to NMR-data-assisted modeling of protein structures

Emilia A. Lubecka[*], Adam Liwo[†]

March 17, 2023

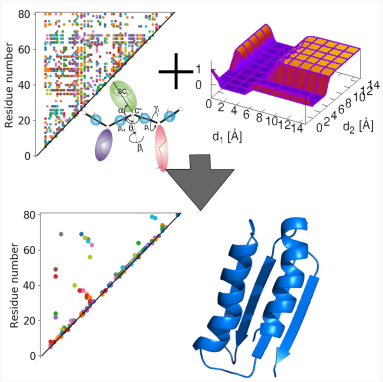
Abstract

The ESCASA algorithm for analytical estimation of proton positions from coarse-grained geometry developed in our recent work has been implemented in modeling protein structures with the highly coarse-grained UNRES model of polypeptide chains (2 sites per residue) and nuclear magnetic resonance (NMR) data. A penalty function with the shape of intersecting gorges was applied to treat ambiguous distance restraints, which automatically selects consistent restraints. Hamiltonian replica exchange molecular dynamics was used to carry out the conformational search. The method was tested with both unambiguous and ambiguous restraints producing good-quality models with GDT_TS from 7.4 units higher to 14.4 units lower than those obtained with the CYANA or MELD software for protein-structure determination from NMR data at the all-atom resolution. The method can thus be applied in modeling the structures of flexible proteins, for which extensive conformational search enabled by coarse graining is more important than high modeling accuracy.

Keywords: NMR-assisted protein-structure modeling, ambiguous restraints, coarse-grained models, UNRES

[*]Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, G. Narutowicza 11/12, 80-233 Gdańsk, Poland

[†]Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-308 Gdańsk, Poland



NMR-data assisted modeling of protein structures with the UNRES coarse-grained force field. The positions of the protons are estimated from coarse-grained geometry by the ESCASA algorithm. Ambiguous restraints are handled through the introduction of a penalty function which has the form of intersecting gorges and takes nearly the same value regardless of whether one only or more components of an ambiguous restraint are satisfied.

INTRODUCTION

Depending on the size of a system, time scale, and the required accuracy, different modeling approaches are used to simulate macromolecules: from quantum mechanics (QM)[1,2] through all-atom approaches[3,4] to coarse-graining (CG) methods.[5–10] Due to the fact that most of the biologically important molecules are too large to handle for the atomistically-detailed simulation methods, coarse-grained models and force fields have become useful in these studies.[9–12] In the coarse-grained models, a single interaction site encompasses several atoms, whole ligand molecule or a polymer unit or even large structural parts of a macromolecule.[13–15] Therefore, the numbers of interaction sites and degrees of freedom are substantially reduced, this resulting in a significant reduction of simulation time. Consequently, the coarse-grained approaches enable us to extend the time- and size-scale of simulations by several orders of magnitude[16]. Moreover, the energy landscapes of the coarse-grained models are smoother that those of the all-atom models, thereby helping the conformational search to avoid local energy minima (kinetic traps)[9]. On the other hand, coarse graining inevitably results in some loss of accuracy, which can be compensated by the introduction of database or experimental information from such technique as the Nuclear Magnetic Resonance (NMR) spectroscopy[17,18].

NMR spectroscopy has been used for protein-structure determination since 1980[19]. It is one of the principal techniques used to determine 3D structures of biomolecules at the atomic precision and to analyze their dynamics in solution under nearly-physiological conditions. The NMR spectroscopy usually provides the information of proton-proton distances, which largely define the spatial structure of a protein, as well as chemical shifts and coupling constants that can be used to determine the local structure[20]. The distance and local-structure information is converted into the distance- and dihedral-angle restraints, which are added to the energy function as penalty terms[20]. It should be noted that, due to errors in peak assignment, some of the NMR distance restraints are wrong, which results in restraining the distances between the atom pairs that are not at contact. Moreover, ambiguous peak assignment often happens[21–23], which results in ambiguous restraints.

Some coarse-grained force fields (e.g., Rosetta[24] and AWSEM[25]) keep part of the atomic

details of the polypeptide chain. However, generally, the atoms (usually hydrogens) between which the distances are to be restrained based on NMR data are not present in coarse-grained representations; this is the case of the most commonly used MARTINI force field[7,26,27] and of the UNited RESidue (UNRES) force field developed in our laboratory[8]. One way to introduce atom-based distance restraints is the approach developed by Latek and Kolinski[28] for the CABS coarse-grained model of polypeptide chains[29], in which all-atom structures are rebuilt from the CG representation and the atom-based restraints are subsequently evaluated. However, while this approach was straightforward to implement with CABS, which uses Monte Carlo dynamics as a conformational-search method, it would be difficult to use with the CG models designed for molecular-dynamics simulations (e.g., MARTINI and UNRES), which require the forces due to the restraints and not only the penalty-function value. Therefore, recently, we developed an analytical approach for the estimation of proton coordinates in proteins from the CG geometry (ESCASA)[30], with which the analytical derivatives of the estimated proton positions in the coarse-grained coordinates and, consequently, the analytical forces due to the atom-based restraints are available. The present version of the method enables us to estimate the coordinates of the backbone and the H^{β} protons.

In this paper, we report the implementation of ESCASA with the UNRES model[8] and tests of the resulting method of NMR-data-assisted modeling of protein structures with 13 proteins for which both X-ray and NMR structures, along with the restraint sets, are available[31] with different secondary structure (α, β, and $\alpha + \beta$), chain length (from 76 to 202 amino-acid residues) and different oligomerization state. We have also included the 80-residue *de novo* designed $\alpha+\beta$ protein Foldit3 (PDB code: 6msp)[32], which was an NMR-data-assisted target in the 13th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP13)[33]. For this protein, we used both the refined NMR restraints deposited in the Protein Data Bank (PDB)[34] and highly ambiguous (up to over 100 alternative assignments per peak) and contradictory "raw" restraints that were released during the CASP13 experiment. Two NMR-assisted targets, N1077 and N1088, with highly-ambiguous restraints, were also released in CASP14[23], of which the models of N1088 were evaluated, while the structure of N1077 was not available by the conclusion of CASP14. In CASP14 we used a preliminary version of the methodology described in this

work, obtaining top-ranked models[23,35] However, because the structure of neither of these two targets has been released yet we did not include these proteins in the present study.

To handle contradictory restraints and ambiguous assignments, we used a penalty function developed in our earlier work[36,37], which produces only a small gradient when a restraints is grossly violated and attains minimum value when only one restraint of the ambiguous set is satisfied. We used the Hamiltonian Replica Exchange Molecular Dynamics (HREMD)[38,39] implemented with UNRES in our earlier work[37,40] for conformational search, with restraint weights varying between replicas. For comparison, we processed the same data with CYANA[41], a standard software for the determination of protein structure from NMR data. We demonstrate that our approach is able to produce correct structures even from highly ambiguous restraint sets.

METHODOLOGY

UNRES model of polypeptide chains

In UNRES[8,42–45], a highly reduced representation of the polypeptide chain is assumed with only kinds of two interaction sites: united side chains (SC) and united peptide groups (p). The α-carbon (C^α) atoms serve to define backbone geometry, but are not interaction sites. United side chains are attached to the α-carbons by virtual bonds and the united peptide groups are located halfway between the adjacent C^αs. The backbone geometry of the polypeptide chain is defined by the $C^\alpha \cdots C^\alpha \cdots C^\alpha$ virtual-bond angles θ and the $C^\alpha \cdots C^\alpha \cdots C^\alpha \cdots C^\alpha$ virtual-bond-dihedral angles γ, whereas the local geometry of the ith side-chain center is defined by the zenith angle α_i and the azimuth angle β_i (Figure 1). The UNRES energy function is based on the physics of interactions and originates from the potential of mean force (PMF) of polypeptide chains in water[44]. It consists of pairwise site-site terms, local terms, and correlation terms corresponding to the coupling between the backbone-local and backbone-electrostatic interactions, which are essential to reproduce regular secondary structures[8,44,45]. In this study we used the latest variant of UNRES, with energy terms derived by using our recently developed scale-consistent formalism[44] and pa-

rameterized by means of the maximum-likelihood approach with 9 small training proteins with different structures[45]. The UNRES energy function depends on temperature, which is a direct consequence of its origin from the PMF; consequently, it contains not only the potential-energy but also the entropy component corresponding to the averaging over the degrees of freedom that are lost when passing from the all-atom to the coarse-grained representation[43].

Restraints and penalty function

We used the interproton distances available from NMR data as restraints in data-assisted modeling. As mentioned, the proton positions are not directly available from the UNRES geometry. To estimate the positions of the backbone-α (H^{α}), backbone-amide (HN) and sidechain-β (H^{β}) protons, we use our recently developed ESCASA algorithm[30]. This algorithm calculates proton positions in the local-coordinate system of the respective $C^{\alpha}\cdots C^{\alpha}\cdots C^{\alpha}$ frame, using approximate analytical formulas. The coordinates of a given H^{α} or H^{β} proton depend on the respective backbone-virtual-bond angle θ, while those of a given HN proton depend on the backbone-virtual-bond-dihedral-angle γ whose axis is the $C^{\alpha}\cdots C^{\alpha}$ axis of the peptide group that contains that proton and the two adjacent virtual-bond angles θ (see Figure 1 for the definition of these angles). The method gives an average error in the proton-proton distance of 0.25 Å, compared to 0.21 Å obtained when proton positions are reconstructed by using the PULCHRA algorithm for the conversion of the C^{α} trace into all-atom backbone[46]. Consistent with averaging the signals of equivalent protons in NMR measurements, the position of the "average" Q^{α} protons are calculated for the glycine residues and the position of the "average" Q^{β} proton is calculated for those residues, which have 2 or 3 H^{β} protons.

We estimated the positions of the H^{γ} and higher-index side-chain protons by assuming that the average positions of these protons are on the $C^{\alpha}\cdots SC$ virtual-bond axes at the distance from the C^{α} atom is proportional to the distance of the projection of this proton on the $C^{\alpha}\cdots SC$ virtual-bond axis given extended side-chain conformation, as given by eq. (1).

$$\mathbf{r}_{H^X} = \mathbf{R}_{C^\alpha} + x_{H^X}\mathbf{R}_{C^\alpha\cdots SC} \tag{1}$$

$$x_{H^X} = \frac{\mathbf{r}_{C^\alpha\cdots H^X_{ext}} \circ \mathbf{R}_{C^\alpha\cdots SC_{ext}}}{d^2_{C^\alpha\cdots SC_{ext}}} \tag{2}$$

where $\mathbf{r}_{C^\alpha\cdots H^X}$ is the estimated position of side-chain proton X, $\mathbf{R}_{C^\alpha}$ is the position of C^α, $\mathbf{R}_{C^\alpha\cdots SC}$ is the vector pointing from C^α to side-chain center, $\mathbf{r}_{C^\alpha\cdots H^X_{ext}}$ is the vector pointing from C^α to the proton in the extended side chain, $\mathbf{R}_{C^\alpha\cdots SC_{ext}}$ is the vector pointing from C^α to the side-chain center in the extended conformation, and $d_{C^\alpha\cdots SC_{ext}}$ is the virtual-bond length of the extended side chain.

To handle ambiguous restraints, we used the penalty function developed in our earlier work[37], which is defined by eq. (3) and has the form of intersecting gorges (Figure 2).

$$V^{dist}_{NMR}(\{d\}; d_l, d_u, A) = -\frac{1}{\alpha}\ln\left\{\sum_{i=1}^{n_{amb}}\exp\left[-\alpha V_{cont}(d_i; d_l, d_u, A)\right]\right\} \tag{3}$$

where $\{d\}$ is the set of distances potentially corresponding to a given ambiguous restraint, α is an arbitrary parameter, and $V_{cont}(d_i; d_l, d_u, A)$ (the contact-distance-restraint function) is a flat-bottom penalty function modified from that introduced in our earlier work[36,47], which is defined by eq. (4). With α large enough, $V_{NMR}(\{d\}; d_l, d_u, A)$ takes a value of nearly 0 independent of whether only one or all restraints of the ambiguous set are satisfied, thus naturally eliminating the restraints of an ambiguous set, which are incompatible with the structure. In this work we set $\alpha = 20$.

$$V_{cont}(d, d_l, d_u, A) = \begin{cases} A\frac{(d-d_l)^4}{\sigma^4+(d-d_l)^4}[1+\kappa\ln\cosh(d-d_l)] & \text{for } d < d_l \\ 0 & \text{for } d_l \le d \le d_u \\ A\frac{(d-d_u)^4}{\sigma^4+(d-d_u)^4}[1+\kappa\ln\cosh(d-d_u)] & \text{for } d > d_u \end{cases} \tag{4}$$

where d is a proton-proton distance estimated from an UNRES structure, d_l and d_u are the lower and upper distance boundaries, respectively, which are taken from NMR data, σ is the thickness of the transition region between zero and maximum restraint height, A is the height of the restraint well, and κ is the slope of the restraint at large distances. In this work we assumed $\sigma = 0.5$ Å and $A = 1.0$ kcal/mol. The original penalty function

from our earlier work[36,47] corresponds to $\kappa = 0$ and quickly approaches the asymptote A, contributing virtually no force when $d \gg d_u$. Thus, the penalty terms do not force incompatible restraints (which usually correspond to wrongly predicted contacts), preventing a simulation from producing non-protein-like structures. With a small $\kappa > 0$, the right asymptote is $A + \kappa(d - d_u)$, which provides a small gradient at large distances, thus mildly guiding the search towards satisfying the restraint but not forcing it if incompatible with other restraints. In this work we used $\kappa = 0.01$ kcal/mol.

Plots illustrating V_{cont} [eq. (4)] for a doubly-degenerated distance restraint and V_{NMR} [eq. (3)] are shown in Figure 2.

The backbone-angular restraints were also used. As in our earlier earlier work[37], we converted the lower and upper boundaries of the backbone ϕ and ψ angles, which were provided in the NMR restraint sets into those in the backbone-virtual-bond and backbone-virtual-bond-dihedral angles (θ and γ, respectively), by using the formulas derived by Nishikawa et al.[48]. The respective restraint function is given by eq. (5).

$$V_{NMR}^{\theta\gamma} = w_\theta g(\theta, \theta_l, \theta_u) + w_\gamma g(\gamma, \gamma_l, \gamma_u) \tag{5}$$

with

$$g(x, x_l, x_u) = \begin{cases} \frac{1}{4}\delta^4 & \text{for } \delta < \frac{x_l - x_u}{2} \\ 0 & \text{for } \frac{x_l - x_u}{2} < \delta < \frac{x_u - x_x}{2} \\ \frac{1}{4}\delta^4 & \text{for } \delta > \frac{x_u - x_l}{2} \end{cases} \tag{6}$$

$$\delta = \left(x - \frac{x_l + x_u}{2} \right) \mod 2\pi \tag{7}$$

where θ_l, θ_u, γ_l, and γ_u are the lower and upper boundaries on the virtual-bond angles θ and virtual-bond-dihedral angles γ, respectively (which are calculated from the boundaries on the ϕ and ψ backbone dihedral angles). In this study we set $w_\theta = 1$ kcal/mol and $w_\gamma = 5$ kcal/mol, respectively.

Hamiltonian Replica Exchange Molecular Dynamics with NMR-data-assisted UNRES

The Replica-Exchange (RE) method[49] and its multiplexed variant (MRE)[50] are enhanced-sampling techniques, which enable a system to overcome kinetic traps by running several canonical Monte Carlo or molecular dynamics simulations at different temperatures and exchanging the temperatures between replicas. When modified energy functions are used for different replicas (e.g., with different restraint strength or different van der Waals repulsive term), the method becomes Hamiltonian Replica Exchange (HRE)[38,39]. When canonical MD is used as a search techniques, the three methods are abbreviated as REMD, MREMD, and HREMD, respectively. All these methods have been implemented with UNRES in our earlier work[40,42,51]. In this work we adapted HREMD to work with NMR-based restraints.

In the HREMD variant implemented in this work, M canonical MD trajectories are run simultaneously at different temperatures and with different weights of the NMR-restraint terms: $M = M_T \times M_H$, where M_T is the number of temperatures and M_H is the number of different weights of the restraint terms. For a given replica (with bi-index ij for temperature and restraint weight, respectively), the pseudo-energy function, V_{ij}, is thus expressed by eq. (8).

$$V_{ij}(\mathbf{X}_{ij}; T_i; w_j) = U_{UNRES}(\mathbf{X}_{ij}; T_i) + V_{NMR}^{\theta\gamma} + w_j V_{NMR}^{dist}(\mathbf{X}_{ij}) \tag{8}$$

where $\mathbf{X}_{ij}$ is the vector of the coarse-grained coordinates of the conformation of replica ij, T_i and w_j are the absolute temperature and weight of the NMR restraints for that replica, U_{UNRES} is the UNRES energy function (that depends on temperature[43]), $V_{NMR}^{\theta\gamma}$ is the penalty function corresponding to angular restraints [eq. (5)] and V_{NMR}^{dist} is the NMR distance-restraint function [eq. (3)]. It should be noted that the angular penalty function has the same weight in all replicas, because these restraints restrict the local chain structure to the desired regions. The replicas constitute a two-dimensional (T_i, V_j) grid. The replicas evolve independently for a given number of steps (10,000 in this work), after which an exchange of temperatures or restraint weights is attempted with its neighboring replica in one or two dimensions (up, down or diagonal on the grid, the direction being selected at

random). The exchange is accepted based on the probability ω expressed by eq. (9):

$$\omega\left(\mathbf{X}_{ij} \rightarrow \mathbf{X}_{kl}\right) = \min[1, \exp(-\Delta)], \quad (k,l) \in \{(i+1,j),(i,j+1),(i+1,j+1)\}$$
$$\Delta = \left[\frac{V_{kl}(\mathbf{X}_{kl}; T_k; w_l)}{RT_k} - \frac{V_{ij}(\mathbf{X}_{kl}; T_i; w_j)}{RT_i}\right] - \left[\frac{V_{kl}(\mathbf{X}_{ij}; T_k; w_l)}{RT_k} - \frac{V_{ij}(\mathbf{X}_{ij}; T_i; w_j)}{RT_i}\right] \quad (9)$$

where R is the universal gas constant. Exchanging the restraint weights between replicas enables the structures that satisfy only part of the restraints but sit in deep energy minima to relax when a high restraint weight is replaced with a smaller one and, consequently, to transform to structures that satisfy more restraints.

To determine the effect of including NMR-based restraints on the quality of the resulting models of the structures of the proteins under study, we also carried out unrestrained MREMD simulations, in which the replicas pertain only to different temperatures.

Test proteins and NMR restraints

We used 13 proteins from the benchmark set created by the G.T. Montelione group[31], for which both X-ray and NMR structures were determined. This set will be referred to as the Montelione/NEF Benchmark Data Set. The information about these proteins, which includes their PDB IDs, numbers of residues, structure types, oligomerization status, and the numbers of distance and angle restraints is collected in Table 1. It can be seen from the Table that these proteins represent all basic structural classes (α, β, $\alpha+\beta$), have a wide span of chain length (from 76 to 202 residues), and four of them are dimers. The respective experimental NMR restraints were taken from the repository available from `https://montelionelab.chem.rpi.edu/databases/nmrdata`. These restraints are refined and only those of 6nbn are weakly ambiguous (Table 1).

To test the ability of our method to handle ambiguous restraints, we used the *de novo* designed protein Foldit3[32] (PDB: 6msp), which corresponds to two of the data-assisted CASP13 targets N1008 and n1008, respectively. For this protein, we used four NMR restraint sets: two variants of refined restraints deposited in the PDB at the 6msp entry[32], denoted as 6msp-v1 and 6msp-v2, respectively, and two raw sets which were provided to the CASP13 participants at `https://predictioncenter.org/download_area/CASP13/extra_experiments` by G.T.

Montelione, under the N1008 and n1008 target entries. The N1008 restraints pertain only to backbone-amide and side-chain protons further to H^{β}, while the restraints for n1008 pertain to all backbone and most of the side-chain protons[33]. The refined 6msp-v1 and 6msp-v2 data sets contain only a few ambiguous peak assignments (1.6 and 1.95 per peak on average, respectively), while those for N1008 and n1008 are highly ambiguous, containing 4.32 and 7.95 assignments per peak on average, respectively, the number of alternative assignments often exceeding 100 (Table 2).

The positions of the NOE peaks for the Foldit3 protein corresponding to all alternative assignments are shown, together with the proton-proton contact-distance maps calculated from the experimental 6msp structure, in Figure 3A-D. It can be seen that the refined NMR distance restraints trace the contact pattern very well (panels A and B of the Figure), while the multiply-assigned peaks cover almost the whole domain for N1008 and n1008 (panels C and D of the Figure). For n1008, the number of possible assignments per peak exceeds 100 for some of the peaks, as illustrated in Figure 4. Moreover, it can be seen from Table 2 that about 48 % of restraints for n1008 and about 57 % of restraints for N1008 are violated by the 6msp structure. We consider a restraint satisfied if, for any of the alternative proton-pair assignments corresponding to this restraint, the proton-proton distance calculated from the structure is not greater than the upper distance boundary. The violations of the upper distance boundaries are quite severe, namely 3.57 Å and 4.21 Å on average for N1008 and n1008, respectively. These violations are averaged over the violated restraints and each violation is the minimum over all alternative proton pair assignments. For the refined 6msp-v1 and 6msp-v2 restraints the number of non-satisfied restraints is much lower (about 14 % and about 17 %, respectively) and the violations of the upper boundaries are below 1 Å(Table 1).

UNRES-based simulation protocol

We used the four-stage protocol developed in our earlier work for protein-structure modeling[35,52]. In the first stage, HREMD simulations with the UNRES force field with geometry restraints from NMR (see sections "Restraints and penalty function" and "Test proteins and NMR restraints"), as well as control MREMD simulations, were carried

out. In the second stage, the results of the simulations were processed with the binless Weighted-Histogram Analysis Method (WHAM)[43,53] to compute the conformational weights to calculate the statistical weight (probability) of each structure of the last section of HREMD simulation. The probabilities were calculated at three temperatures: 260, 280 and 300 K. In the third stage, the conformational ensemble was dissected into 5 clusters for each temperature (15 clusters in summary) by means of Ward's minimum variance method and the clusters were ranked based on the collective probability of all conformations of a given cluster. For NMR-restrained simulations, the conformation of the cluster with the lowest value of the NMR-penalty function (i.e., satisfying the NMR restraints best) was selected as a cluster representative[43], while for the unrestrained simulations the cluster representative was selected as the conformation closest to the cluster-averaged conformations (cf. Ref. 43). Finally, in the fourth stage, the obtained conformations were converted to all-atom structures using the PULCHRA[46] and SCWRL[54] algorithms and subsequently refined using the Assisted Model Building with Energy Refinement, version 2014 (AMBER14) package[3] with the ff14SB force field and Generalized Born Surface Area (GBSA) implicit-solvent model. The refinement was carried out with 500 steps of energy minimization, followed by a short (0.3 ps) canonical MD simulation, and finished with an additional 500-step minimization. Such refined all-atom structures were taken for further analysis.

The replicas of the UNRES/HREMD simulations of the first stage consisted constituted a 2-dimensional grid of 12 temperatures (262, 267, 274, 279, 285, 290, 295, 301, 308, 333, 355, and 370 K, respectively) and 8 distance-restraint weights [eq. (8)] (w_{NMR} = 0, 0.0634678, 0.18963, 0.295432, 0.444649, 0.653766, 0.824766, and 1, respectively), this giving a total of 72 replicas. For 2kzn and 2kcu, for which the conformational search was more demanding, we used 24 temperature replicas, from 260 to 500 K, this giving a total of 144 replicas in HREMD simulations. The final statistical weights of the conformations were calculated for $w_{NMR} = 1$. The temperatures and the weights were selected to maximize the number of walks in the weight space, by using a variant of the Hansmann algorithm[55], adapted to the weight space[35]. The unrestrained UNRES/MREMD simulations were carried out with the same set of 12 temperatures as above, 4 trajectories run at each temperature (48 replicas total). For both the HREMD simulations and the control REMD simulations, each trajectory consisted

of 20,000,000 steps with a 4.89 fs step length and the replicas were exchanged every 10,000 steps. The trajectories were run in the Langevin scheme implemented in UNRES in our earlier work[16], with the viscosity of water scaled by a factor of 0.01 as in our earlier work[16]. The Velocity-Verlet method with the Adaptive Multiple Time Step (A-MTS) algorithm[56] was used to integrate the equations of motion. All runs were started from randomly-generated conformations.

To compare the obtained models with the respective experimental structures, we used the α-carbon Root-Mean-Square Deviation (C^{α}-RMSD or RMSD) and the Global Distance Test Total Score (GDT_TS)[57,58], which is an average of the percentage of C^{α} atoms in the model, which are at a distance not exceeding 1, 2, 4, and 8 Å, respectively, from the corresponding C^{α} atoms in the reference (experimental) structure at the optimal superposition. The GDT_TS is the primary measure used in CASP to compare the structures of the models of proteins or their domains with the respective experimental structures. For the 13 proteins from the Montelione/NEF Benchmark Data Set[31], we used the X-ray structures as reference structures, while for the Foldit3 protein[32] we used the average 6msp structure.

Simulations with CYANA

To compare our results with one of the currently leading method in protein structure determination based on NMR data, we have used CYANA 2.1 software[41]. CYANA calculates structures using simulated annealing[41] with a highly efficient torsion angle space molecular dynamics algorithm[59]. For each system we ran 4,000 MD steps with CYANA, collecting a total of 50 snapshots. The 15 conformers that satisfied the NMR restraints best were selected for further analysis.

RESULTS AND DISCUSSION

Unambiguous and weakly ambiguous restraints

The results of the determination of the structures of the 13 proteins of the Montelione/NEF Benchmark Data Set (Table 1) are summarized, in form of GDT_TS and C^{α}-RMSD bar

plots, in Figure 5A and 5B, respectively, for the best and the first models, respectively. The respective numerical values are collected in Table S1 of the Supplementary Material. For each protein, the *first model* corresponds to the cluster with the greatest summary probability of its constituent conformations, while *best model* is the one with the highest GDT_TS. It should be noted, though, that because the explicit comparison with the experimental structure is not involved in the selection of this model, it is not the highest GDT_TS model obtained in the simulations of a given protein. It can be seen that the experimental structures of these proteins satisfy 47 % – 100 % of restraints (Table 1); however, even for sets with the least percentage of satisfied restraints the mean restraint violations are lower than 1 Å.

As shown in Figure 5, the GDT_TS values of the best CYANA models are generally greater than 80; however, for 3 larger proteins: 2kzn (147 residues), 2kcu (166 residues) and 2kw5 (202 residues) the GDT_TS values of the best models are only 59.7, 64.1 and 66.0, respectively, even if flexible ends and loops are excluded. Inspection of the respective PDB structures shows that the N-terminal α-helix of the 2kcu structure has two alternative positions. The discrepancy between the NMR and the X-ray structures of the two other proteins is likely to be caused by crystal packing. This conclusion is supported by the fact that the experimental X-ray structures corresponding to 2kzn and 2kw5 (3e0o and 3mer, respectively) satisfy less than 50 % of NMR restraints (Table 1).

As can be seen from Figure 5 and Table S1, the quality of the CYANA models is usually, but not always higher than that of the UNRES models. The GDT_TS values of the best and the first models obtained from restrained simulations with UNRES are by 6.34 and 6.45 units, respectively, smaller on average than those of the corresponding CYANA models. However, the first UNRES model of 2ko1 has GDT_TS by 7.37 units higher than its CYANA counterpart. In view of the high degree of coarse graining of UNRES (only 2 centers per residue), the results are very good and suggest that our approach can be used in the modeling tasks in which extensive conformational search enabled by coarse graining is more important than high accuracy [e.g., modeling the conformational ensembles of the intrinsically disordered proteins (IDPs) and proteins with intrinsically disordered regions (IDRs)]. We also note that the GDT_TS values of the UNRES models correlate well with those of the corresponding CYANA models (Figure 6), the correlation coefficients being 0.8154 and

0.8024 for the best and for the first models, respectively.

The results of unrestrained control simulations are collected in Table S3 of the Supplementary Material. As can be seen from this Table, the models obtained in unrestrained simulations are very far from the corresponding experimental structure, except for the best model of 2juw, which has GDT_TS of 45.90. It should be noted that even in this case, addition of NMR restraints dramatically improves the model quality (Table S1). Thus, for these proteins, including NMR restraints is necessary to obtain good-quality models.

The percentages of satisfied distance restraints for the first and the best models obtained in UNRES and CYANA calculations are shown, as bar plots, in Figure 7. As shown, the structures obtained with CYANA satisfy the experimental restraints better, consistent with the fact that CYANA uses all-atom representation of polypeptide chains, in which proton positions are explicit in the model, while the positions of protons are only estimated from the UNRES structures.

Highly ambiguous restraints

To test the ability of our method to handle ambiguous restraints in comparison with other approaches, we used the *de novo* designed protein Foldit3[32]. With the two refined sets of NMR data available for this protein at the 6msp entry (6msp-v1 and 6msp-v2), CYANA has produced good-quality structures with C^{α}-RMSD from 0.98 to 1.63 Å and GDT_TS from 87.81 to 93.75 (Figure 8 and Table S2). The quality of the CYANA models obtained with the 6msp-v2 set is slightly lower. Due to its highly reduced representation of polypeptide chains, UNRES produced lower-quality but still good models with C^{α}-RMSD of from 1.31 to 1.81 and GDT_TS from 78.75 to 87.19, respectively.

The results turn to the favor of our UNRES-based approach when un-refined restraints with high ambiguity and a significant amount of wrong information, corresponding to the CASP13 data-assisted targets n1008 and N1008 are used in modeling. The respective bar plots are shown in Figure 8 and the GDT_TS and RMSD values are collected in Table S2. The GDT_TS of the CYANA and UNRES first models are 23.75 and 80.00 for the n1008 restraints, and 27.19 and 72.81 for the N1008 restraints, respectively, while those of the best models are 26.25 and 80.00 for the n1008 and 33.75 and 75.94 for the n1008 and N1008

restraints, respectively. This is the result of using the V_{NMR}^{dist} penalty function [eq. (3)] to handle ambiguous restraints, the component of which is the modified flat-bottom Lorenz-like penalty function [eq. (4)], which results in a small gradient when some of the restraints are incompatible with the other ones and can, thus, be ignored in building the candidate models. The results are significantly improved with respect to the quality of our UNRES models of the two targets obtained in CASP13,[37], in which we used a "naïve" estimation of proton positions, with the upper-bounds of the proton-proton distances approximated by the distances between the pertinent UNRES centers (C^{α} for the backbone and SC center for the side-chain protons) plus 2 Å, obtaining the GDT_TS values of the best n1008 and N1008 models of only 52.27 and 44.16, respectively.

The best models (with the highest GDT_TS values) of the *de novo* designed protein Foldit3 obtained with UNRES and CYANA, using NMR restraints of different quality, are shown, together with the experimental structure, in Figure 9. It can be seen that the CYANA models obtained with the n1008 (Fig. 9A) and N1008 (Fig. 9B) restraints are different from the experimental structure and consist mainly of unstructured sections, whereas those obtained using the 6msp-v1 (Fig. 9C) or 6msp-v2 (Fig. 9D) restraints are of high quality. In contrast to this, all UNRES models, regardless of the quality of the NMR restraints for the Foldit3 protein, exhibit the same fold as that of the experimental structure (Fig. 9F-I). This observation suggest that the UNRES-based approach is very robust, even though it does not result in high-resolution structures. As for the proteins of the Montelione/NEF set, including NMR restraints is necessary to obtain good-quality model. The best model of the *de novo* designed protein Foldit3 obtained in unrestrained UNRES simulations has GDT_TS of 58.75 and the first model is not native-like (Table S3).

Recently, Mondal and Pérez[60] used the MELD protocol coupled with the all-atom AMBER force field to model, among others, the n1008 and N1008 CASP13 targets. They obtained GDT_TS and RMSD values of 82.14/1.60 Å, 77.27/1.77 Å, for n1008 and N1008, respectively, where the first number in a pair is C^{α} RMSD and the second number is GDT_TS, compared to 80.00/1.66 Å and 75.94/2.12 Å obtained with UNRES. The higher GDT_TS and lower RMSD values obtained with MELD presumably result not only from the low resolution of the UNRES model but also from extensive all-atom refinement of the MELD

models[60]. It should also be noted that MELD performs iterative elimination of violated restraints, while the penalty function given by eq. (3) performs this task implicitly, without having to interrupt the simulations.

After UNRES simulations with contradictory and ambiguous restraints, we can also prune the original set of ambiguous restraints by keeping only those which correspond to the interproton contacts determined from the UNRES model after its conversion to all-atom geometry. We set a distance cut-off of the upper interproton-distance boundary plus 2 Å to select the sets of consistent restraints out of the raw n1008 and N1008 restraints and used the obtained restraints in CYANA runs. With these restraints, we obtained the GDT_TS values of the best CYANA models of 85.31 and 71.56, for n1008 and N1008, respectively. The GDT_TS of the best CYANA models obtained with the refined 6msp-v1 and 6msp-v2 restraint sets are 93.75 and 91.25, respectively (Table S2).

Although CYANA produces poor structures in simulations with both n1008 and N1008 restraints (Fig. 9), while our UNRES-based approach with the penalty function designed to handle ambiguous/inaccurate restraints produced good-quality structures, the UNRES structures satisfy a smaller percentage of NMR restraints than the CYANA structures do for both n1008 but not for N1008 (Figure 10). A plausible explanation of this fact is that about a half and more than a half of the n1008 and N1008 restraints, respectively, are violated by the experimental 6msp structure (Table 2), this suggesting that many of the restraints are inconsistent. The inconsistent restraints are effectively ignored by the Lorenz-like V_{cont} penalty function but they contribute to the penalty function used in CYANA, hence the algorithm used there tries to satisfy also those restraints, this resulting in the deterioration of the quality of the models.

CONCLUSIONS

We have developed a multiscale approach to NMR-data-assisted modeling of protein structures, in which the main part of the conformational search is carried out at the coarse-grained level, with explicit NMR-based restraints imposed at simulation time. We have implemented with UNRES the ESCASA algorithm for calculating the approximate positions of the back-

bone and H^{β} protons from the coarse-grained geometry[30], extended in this work to other side-chain protons. To handle inaccurate and ambiguous restraints, we have implemented the "intersecting-gorge-like" function [eq. (3) and Figure 2][37], which is based on the flat-bottom Lorentz-like penalty function introduced in our earlier work[36,47] modified in this work to provide a mild slope at large distances [eq. (4)].

We have tested our approach with both unambiguous restraints (the Montelione/NEF Benchmark Data Set[31]; Table 1) and the less accurate "raw" restraints, some of which corresponded to over 100 possible assignments (the 2 sets of restraints resulting from the NMR experiments with the the *de novo* designed Foldit3 protein[32], Table 2). The GDT_TS values (computed with respect to the corresponding crystal structures) of the data-assisted first-choice UNRES models of the 13 proteins of the first set range from 44.37 to 87.69, while those obtained with the standard CYANA software[41] range from 57.84 to 91.21, the GDT_-TS values of the CYANA models being higher by 6.45 on average. For the best (highest GDT_TS) models the ranges of GDT_TS are from 48.06 to 92.91 for UNRES and from 59.68 to 92.77 for CYANA, respectively, the GDT_TS values of the CYANA models being higher by 6.34 on average (Figure 5 and Table S1). Remarkably, for some of the benchmark proteins the GDT_TS of the models obtained with the UNRES-based approach are higher than those from CYANA (Figure 5 and Table S1). The obtained results are very good in view of the fact that UNRES is a highly reduced model, with only 2 interaction sites per residue.

For the Foldit3 protein and ambiguous data sets, CYANA failed to produce reasonable models, while our UNRES-based approach produced models that were similar to the models of this protein obtained with refined restraints, with GDT_TS from the average NMR structure deposited in the PDB from 72.81 to 80.00 (Figures 8 and 9 and Table S2). The MELD algorithm[60], which was used by Mondal and Pérez to treat these data sets produced higher GDT_TS values. However, the difference is only about 2.5 GDT_TS units. Moreover, the final structures obtained with MELD were subjected to extensive refinement at the all-atom level, while only cursory refinement with the AMBER force field and implicit solvent was carried out for UNRES model.

The above results suggest that NMR data can be used as restraints in data-assisted modeling with UNRES and other heavily coarse-grained protein models, which can be of

importance if the structures of large proteins are to be determined. Clearly, to make our approach reach the quality of all-atom approaches for NMR-data-assisted modeling, an algorithm for the refinement of UNRES models at the all-atom level is necessary. This work is currently underway in our laboratory. On the other hand, our UNRES-based approach in its present form seems to be appropriate when extensive conformational search (enabled by the highly simplified UNRES model) is more important than model accuracy, e.g., in NMR-data-assisted determination of the conformational ensembles of the IDPs or the IDRs. Another advantage of our approach is its ability to handle ambiguous restraints without having to iteratively filter the experimental data.

DATA AVAILABILITY

The source code of the version of UNRES with the NMR-assisted-simulation feature, which uses the algorithm for handling ambiguous and contradictory restrains and the ESCASA algorithm for analytical estimation of proton positions from C^{α}-trace geometry, and ESCASA parameters are available from the Downloads section of the UNRES package page (https://unres.pl/downloads, files unres-src-HCD-5D_nmr-May-5-2021.tar.gz and PARAM-May-5-2021.tar.gz, respectively).

ACKNOWLEDGEMENTS

We thank Prof. Gaetano T. Montelione, Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute for providing the benchmark set of proteins with both X-ray and NMR structures (Ref. 31).

This work was supported by grant No. UMO-2021/40/Q/ST4/00035 from the National Science Center of Poland (Narodowe Centrum Nauki). For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission. Computational resources were provided by (a) the Interdisciplinary Center of Mathematical and Computer Modeling (ICM) the University of Warsaw under grant No. GA71-23, (b) the Centre of Informatics - Tricity Academic Supercomputer & Network (CI TASK) in Gdańsk, (c) the Academic Computer Centre Cyfronet

AGH in Krakow under grants: unres2021, and (d) our 796-processor Beowulf cluster at the Faculty of Chemistry, University of Gdańsk.

References

[1] P. Atkins and R. Friedman, *Molecular Quantum Mechanics* (Oxford University Press, 2010).

[2] R. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, 1989).

[3] D. A. Case, V. Babin, J. T. Berryman, R. M. Betz, Q. Cai, D. S. Cerutti, T. E. Cheatham, T. A. Darden, R. E. Duke, H. Gohlke, et al., *Amber 14* (2014), University of California: San Francisco.

[4] H. Berendsen, D. van der Spoel, and R. van Drunen, Comput. Phys. Commun. **91**, 43 (1995).

[5] A. Kolinski and J. Skolnick, Polymer **45**, 511 (2004).

[6] A. Kolinski and J. Skolnick, Proteins **32**, 475494 (1998).

[7] S. J. Marrink and D. P. Tieleman, Chem. Soc. Rev. **42**, 6801 (2013).

[8] A. Liwo, M. Baranowski, C. Czaplewski, E. Gołaś, Y. He, D. Jagieła, P. Krupa, M. Maciejczyk, M. Makowski, M. A. Mozolewska, et al., J. Mol. Model. **20**, 2306 (2014).

[9] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, and A. Kolinski, Chem. Rev. **116**, 7898 (2016).

[10] N. Singh and W. Li, Int. J. Mol. Sci. **20**, 3774 (2019).

[11] S. Takada, R. Kanada, C. Tan, T. Terakawa, W. Li, and H. Kenzaki, Acc. Chem. Res. **48** (2015).

[12] G. Voth, *Coarse-Graining of Condensed Phase and Biomolecular Systems* (CRC Press, Taylor & Francis Group, 2008), 1st ed.

[13] J. Dama, A. Sinitskiy, M. McCullagh, J. Weare, B. Roux, A. Dinner, and G. Voth, J. Chem. Theory Comput. **9**, 2466 (2013).

[14] A. Davtyan, J. Dama, A. Sinitskiy, and G. Voth, J. Chem. Theory Comput. **10**, 5265 (2014).

[15] M. Nanias, M. Chinchio, J. Pillardy, D. Ripoll, and H. Scheraga, Proc. Nat. Acad. Sci. U. S. A. **100**, 1706 (2003).

[16] M. Khalili, A. Liwo, A. Jagielska, and H. A. Scheraga, J. Phys. Chem. B **109**, 13798 (2005).

[17] D. A. Case, H. J. Dyson, and P. E. Wright, in *NMR in Proteins*, edited by G. Clore and A. Gronenborn (MacMillan, New York, 1993), pp. 53–91.

[18] K. Joo, I. Joung, J. Lee, J. Lee, W. Lee, B. Brooks, S. J. Lee, and J. Lee, Proteins: Struct., Funct., Bioinf. **83**, 2251 (2015).

[19] M. P. Williamson, T. F. Havel, and K. Wüthrich, J. Mol. Biol. **182**, 295 (1985).

[20] J. Cavanagh, W. J. Fairbrother, A. G. Palmer, M. Rance, and N. J. Skelton, eds., *Protein NMR Spectroscopy* (Academic Press, Burlington, 2007), second edition ed.

[21] M. Nilges, J. Mol. Biol. **245**, 645 (1995).

[22] Y. J. Huang, K. P. Brock, Y. Ishida, G. V. T. Swapna, M. Inouye, D. S. Marks, C. Sander, and G. T. Montelione, Biol. NMR A. Methods Enzymol. **614**, 363 (2019).

[23] Y. J. Huang, N. Zhang, B. Bersch, K. Fidelis, M. Inouye, Y. Ishida, A. Kryshtafovych, N. Kobayashi, Y. Kuroda, G. Liu, et al., Proteins **89**, 1959 (2021).

[24] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker, in *Numerical Computer Methods, Part D* (Academic Press, 2004), vol. 383 of *Methods in Enzymology*, pp. 66 – 93.

[25] A. Davtyan, N. P. Schafer, W. Zheng, C. Clementi, P. G. Wolynes, and G. A. Papoian, J. Phys. Chem. B **116**, 8494 (2012).

[26] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, J. Phys. Chem. B **111**, 7812 (2007).

[27] L. Monticelli, S. Kandasamy, X. Periole, R. Larson, D. Tieleman, and S. Marrink, J. Chem. Theory Computat. **4** (2008).

[28] D. Latek and A. Koliński, J. Comput. Chem. **32**, 536 (2011).

[29] A. Kolinski, Acta Biochim. Polym. **51**, 349 (2004).

[30] E. Lubecka and A. Liwo, J. Comput. Chem. **42**, 1579 (2021).

[31] J. K. Everett, R. Tejero, S. B. K. Murthy, T. B. Acton, J. M. Aramini, M. C. Baran, J. Benach, J. R. Cort, A. Eletsky, F. Forouhar, et al., Prot. Sci. **25**, 30 (2016).

[32] B. Koepnick, J. Flatten, T. Husain, A. Ford, D.-A. Silva, M. J. Bick, A. Bauer, G. Liu, Y. Ishida, A. Boykov, et al., Nature **570**, 390 (2019).

[33] D. Sala, Y. J. Huang, C. A. Cole, D. A. Snyder, G. Liu, Y. Ishida, G. V. T. Swapna, K. P. Brock, C. Sander, K. Fidelis, et al., Proteins **87**, 1315 (2019).

[34] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucl. Acid Res. **28**, 235 (2000).

[35] A. Antoniak, I. Biskupek, K. Bojarski, C. Czaplewski, A. Giełdoń, M. Kogut, M. Kogut, P. Krupa, A. Lipska, A. Liwo, et al., J. Mol. Graph. Model. **108**, 108008 (2021).

[36] E. A. Lubecka and A. Liwo, J. Comput. Chem. **40**, 2164 (2019).

[37] E. Lubecka, A. Karczyńska, A. Lipska, A. Sieradzan, K. Zieba, C. Sikorska, U. Uciechowska-Kaczmarzyk, S. Samsonov, P. Krupa, M. Mozolewska, et al., J. Mol. Graph. Model. **92** (2019).

[38] H. Fukunishi, O. Watanabe, and S. Takada, J. Chem. Phys. **116**, 9058 (2002).

[39] F. Zeller and M. Zacharias, J. Chem. Theory Comput. **10**, 703 (2013).

[40] A. S. Karczyńska, C. Czaplewski, P. Krupa, M. A. Mozolewska, K. Joo, J. Lee, and A. Liwo, J. Comput. Chem. **38**, 2730 (2017).

[41] P. Güntert, C. Mumenthaler, and K. Wüthrich, J. Mol. Biol. **273**, 283 (1997), ISSN 0022-2836.

[42] A. Liwo, S. Ołdziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, J. Comput. Chem. **18**, 849 (1997).

[43] A. Liwo, M. Khalili, C. Czaplewski, S. Kalinowski, S. Ołdziej, K. Wachucik, and H. A. Scheraga, J. Phys. Chem. B **111**, 260 (2007).

[44] A. K. Sieradzan, M. Makowski, A. Augustynowicz, and A. Liwo, J. Chem. Phys. **146**, 124106 (2017).

[45] A. Liwo, A. K. Sieradzan, A. G. Lipska, C. Czaplewski, I. Joung, W. Żmudzińska, A. Hałabis, and S. Ołdziej, J. Chem. Phys. **150**, 155104 (2019).

[46] P. Rotkiewicz and J. Skolnick, J. Comput. Chem. **29**, 1460 (2008).

[47] A. K. Sieradzan and R. Jakubowski, J. Comput. Chem. **38**, 553 (2017).

[48] K. Nishikawa, F. A. Momany, and H. A. Scheraga, Macromolecules **7**, 797 (1974).

[49] U. H. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).

[50] Y. M. Rhee and V. S. Pande, Biophys. J. **84**, 775 (2003).

[51] C. Czaplewski, S. Kalinowski, A. Liwo, and H. A. Scheraga, J. Chem. Theor. Comput. **5**, 627 (2009).

[52] P. Krupa, M. Mozolewska, M. Wiśniewska, Y. Yin, Y. He, A. Sieradzan, R. Ganzynkowicz, A. Lipska, A. Karczyńska, M. Ślusarz, et al., Bioinformatics **32**, 3270 (2016).

[53] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, J. Comput. Chem. **13**, 1011 (1992).

[54] Q. Wang, A. A. Canutescu, and R. L. Dunbrack, Nat. Protoc. **3**, 1832 (2008).

[55] S. Trebst, M. Troyer, and U. H. E. Hansmann, J. Chem. Phys. **124**, 174903 (2006).

[56] F. Rakowski, P. Grochowski, B. Lesyng, A. Liwo, and H. A. Scheraga, J. Chem. Phys. **125**, 204107 (2006).

[57] A. Zemla, Nucleic Acids Res. **31**, 3370 (2003).

[58] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, Proteins **82 (Suppl. 2)**, 1 (2013).

[59] A. Jain, N. Vaidehi, and G. Rodriguez, **106** (1993), ISSN 0021-9991.

[60] A. Mondal and A. Perez, Front. Mol. Biosci. **8**, 774394 (2021).

Figure 1: UNRES model of polypeptide chains. The interaction sites are united peptide groups located between the consecutive α-carbon atoms (light-blue spheres) and united side chains attached to the α-carbon atoms (spheroids with different colors and dimensions). The backbone geometry of the simplified polypeptide chain is defined by the $C^{\alpha}\cdots C^{\alpha}\cdots C^{\alpha}$ virtual-bond angles θ (θ_i has the vertex at C_i^{α}) and the $C^{\alpha}\cdots C^{\alpha}\cdots C^{\alpha}\cdots C^{\alpha}$ virtual-bond-dihedral angles γ (γ_i has the axis passing trough C_i^{α} and C_{i+1}^{α}). The local geometry of the ith side-chain center is defined by the polar angle α_i (the angle between the bisector of the respective angle θ_i and the $C_i^{\alpha}\cdots SC_i$ vector) and the azimuth angle β_i (the angle of counter-clockwise rotation of the $C_i^{\alpha}\cdots SC_i$ vector about the bisector from the $C_{i-1}^{\alpha}\cdots C_i^{\alpha}\cdots C_{i+1}^{\alpha}$ plane, starting from C_{i-1}^{α}). For illustration, the bonds of the all-atom chains, except for those to the hydrogen atoms connected with the carbon atoms, are superposed on the coarse-grained picture. Reproduced with permission from Zaborowski et al., J. Chem. Inf. Model., 55, 2050 (2015). Copyright 2015 American Chemical Society.

Figure 2: (A) Illustration of the V_{NMR}^{dist} penalty function to handle ambiguous restraints defined by eq. (3) for a doubly-degenerate distance restraint, which has the shape of intersecting gorges. As shown in the plot, satisfying only one of the restraints of the degenerate set results in nearly as low a value of the penalty function as satisfying both restraints. (B) Plot of the flat-bottom Lorenz-like V_{cont} distance-restraint function with a small slope at large distances.

Figure 3: Illustration of the four NMR distance restraints sets for the *de novo* designed Foldit3 protein used in this work: 6msp-v1 (A), 6msp-v2 (B), N1008 (C) and n1008 (D), of which those shown in panels A and B are unambiguous and were taken from the 6msp PDB entry, while those shown in panels C and D are ambiguous and correspond to the N1008 and n1008 CASP13 targets[33]. The multi-colored crosses in the lower diagonal represent the restraints, all components of an ambiguous restraint having the same color, while the green circles in the lower diagonal represent the proton-proton contacts in the 6msp experimental structure[32].

Figure 4: Bar plot of the number of assignments per peak for the CASP14 n1008 target (*de novo* designed Foldit3 protein). The peaks have been sorted by the number of assignments in the descending order. The red horizontal line corresponds to average number of assignments per peak.

Figure 5: Bar plots of (A) the GDT_TS and (B) C^{α}-RMSD for the experimental NMR structures (gray), CYANA best models (green), CYANA first models (light green), UNRES best models (blue), and UNRES first models (light blue) of the 13 proteins selected from the Montelione/NEF Benchmark Data Set[31]. The X-ray structures were the reference structures in the computation of GDT_TS and C^{α}-RMSD. Flexible ends and flexible loops were excluded from comparison.

Figure 6: Correlation of the GDT_TS values of the best (purple) and first (green) CYANA and UNRES models. The slopes and intercepts in the equation $\mathrm{GDT\_TS_{UNRES}} = a\mathrm{GDT\_TS_{CYANA}} + b$ are $a = 1.12(0.16)$, $b = -16.(13.)$ for the best-model GDT_TS and $a = 1.10(0.17)$, $b = -16.(12.)$ for average GDT_TS, where the numbers in parentheses are the standard deviations of the parameters. The correlation coefficients are 0.9030 and 0.8958 for the best- and first-model GDT_TS, respectively.

Figure 7: Bar plots of the percentages of satisfied distance restraints of the NMR experimental structures (gray), CYANA best models (green), CYANA first models (light green), UNRES best models (blue) and UNRES first models (light blue) of the 13 proteins selected from the set of 41 proteins for which both X-ray and NMR structures are available[31].

Figure 8: Bar plots of (A) the GDT_TS and (B) C^{α}-RMSD for the best models (darker colors) and first models (lighter colors) of the CYANA (green) and UNRES (blue) models of the *de novo* designed Foldit3 protein[32] obtained with the 6msp-v1, 6msp-v2 (unambiguous), n1008 and N1008 (ambiguous) restraint sets. The mean 6msp structure was the reference structure in the computation of GDT_TS and RMSD.

Figure 9: Cartoon diagrams of the best models of *de novo* designed Foldit3 protein, obtained with different NMR distance restraints sets: n1008 (A and F), N1008 (B and G), 6msp-v2 (C and H) and 6msp-v1 (D and I), obtained with CYANA (A-D) and with the UNRES-based approach developed in this work (F-I). The experimental 6msp structure is shown in the middle panel (E), the chain colored from blue to red from the N- to the C-terminus, respectively.

Figure 10: Bar plots of the percentage of satisfied 6msp-v1, 6msp-v2 (unambiguous) and n1008 and N1008 (ambiguous) distance restraints of the *de novo* designed Foldit3 protein by the experimental 6msp structure (gray), the CYANA best models (green), the CYANA first models (light green), the UNRES best models (blue), and the UNRES first models (light blue) obtained with these restraints.

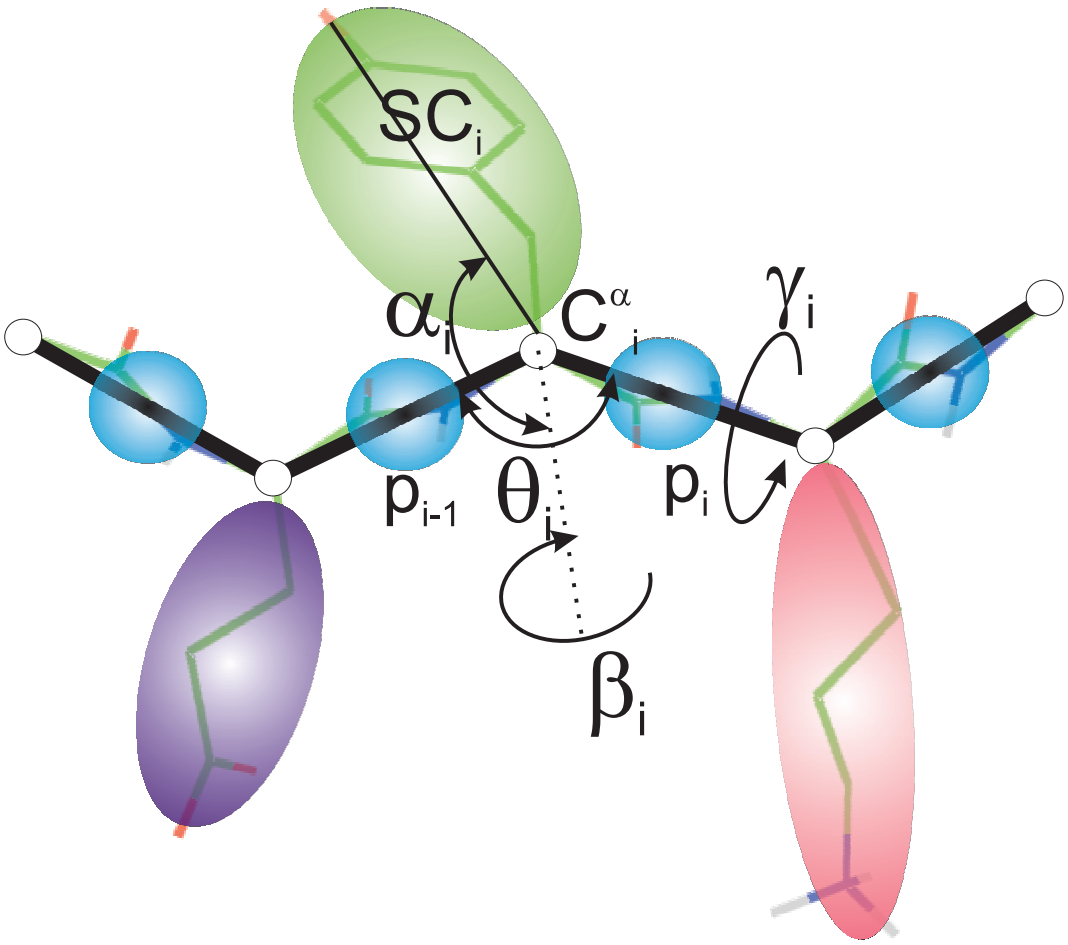
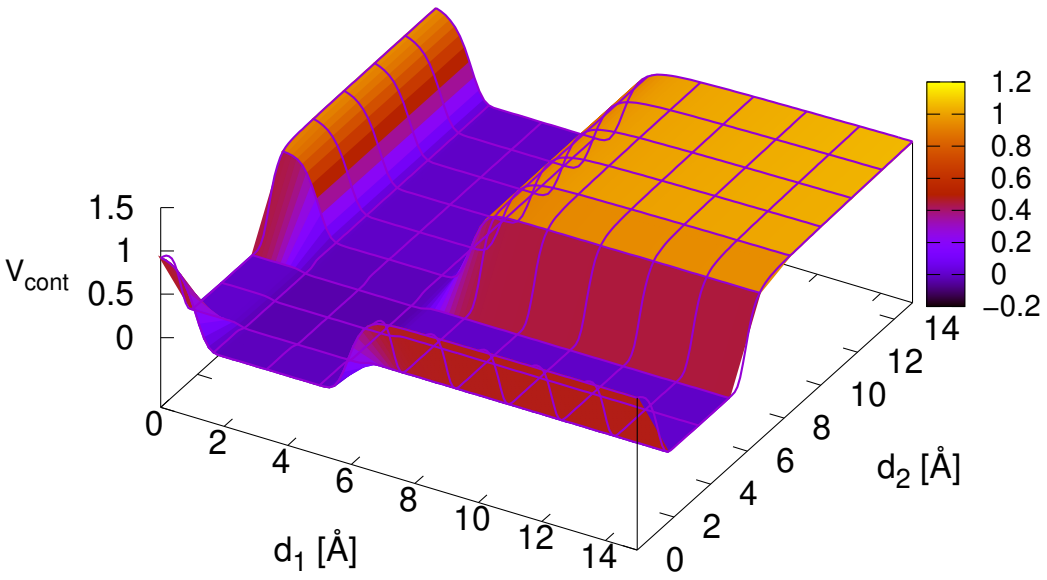
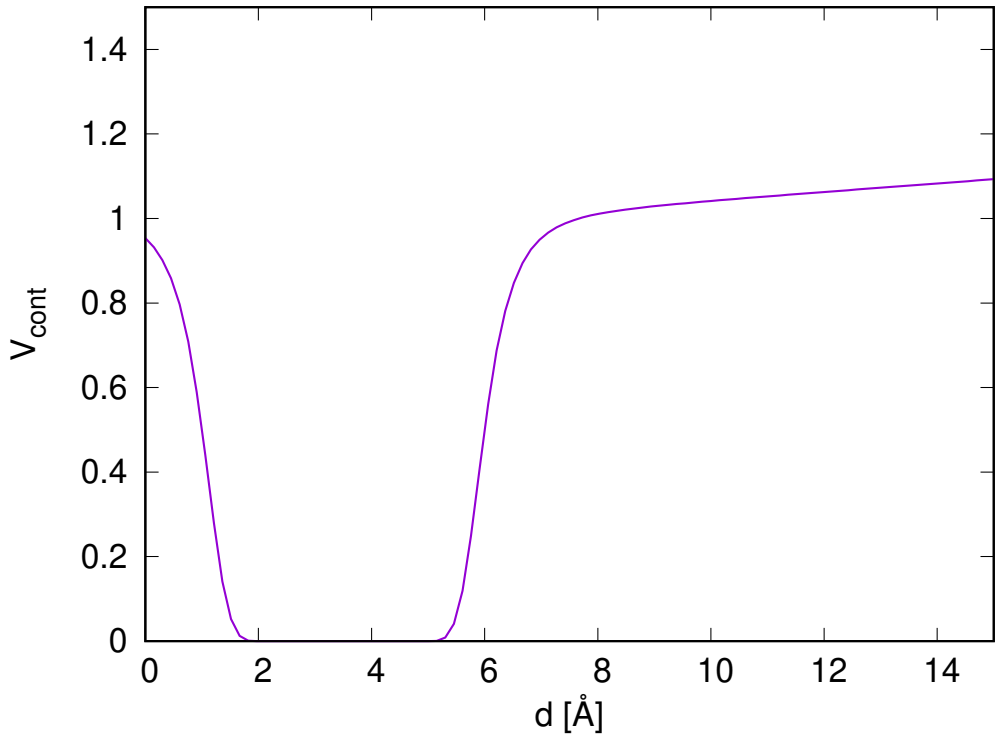


Figure 1
E.A. Lubecka, A. Liwo
J. Comput. Chem.



A



B

Figure 2
E.A. Lubecka, A. Liwo
J. Comput. Chem.

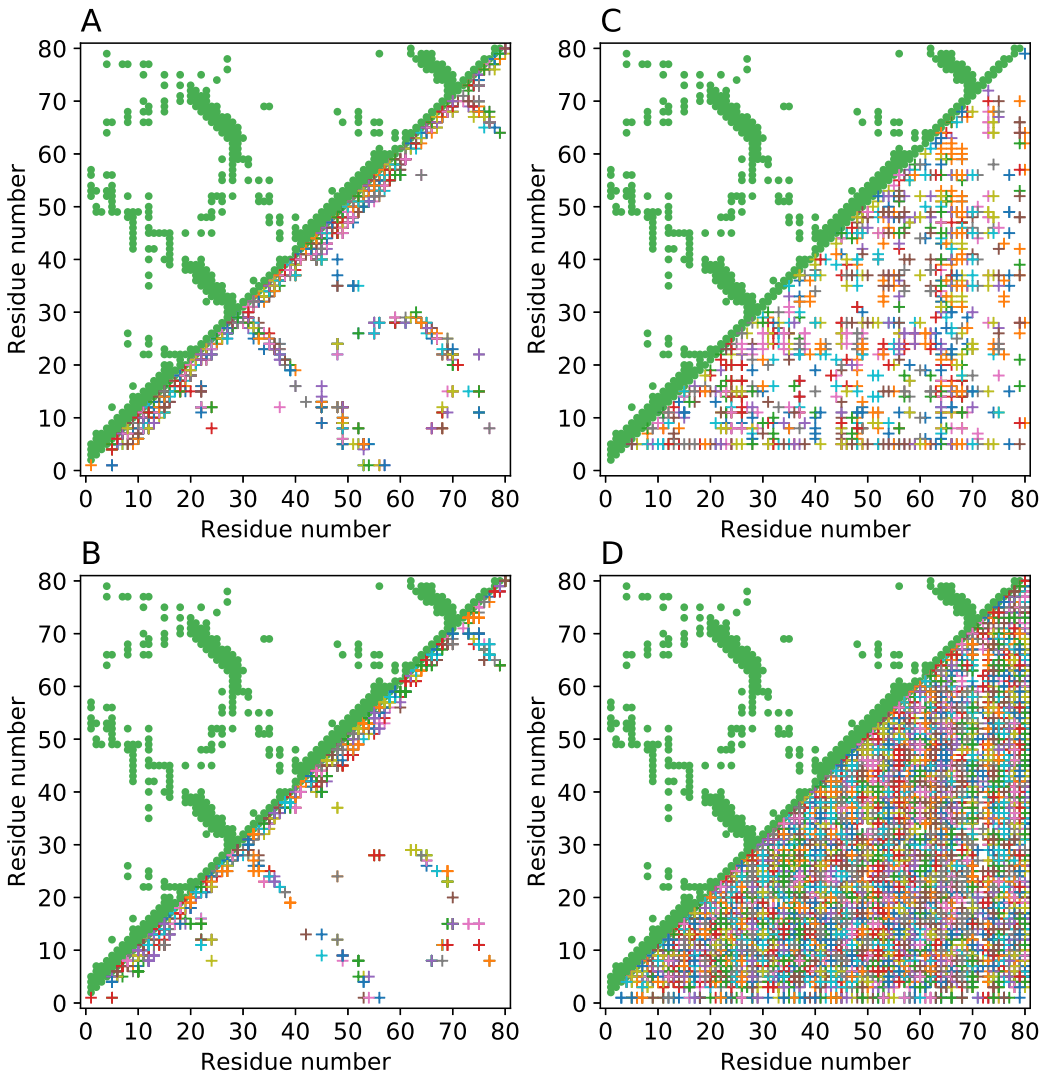


Figure 3
E.A. Lubecka, A. Liwo
J. Comput. Chem.

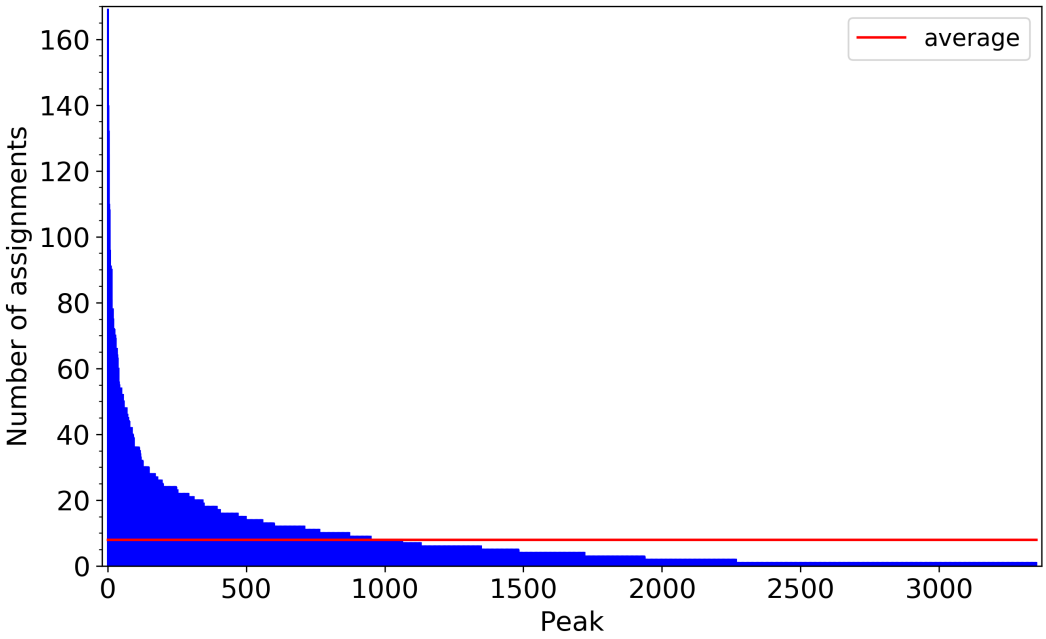


Figure 4
E.A. Lubecka, A. Liwo
J. Comput. Chem.

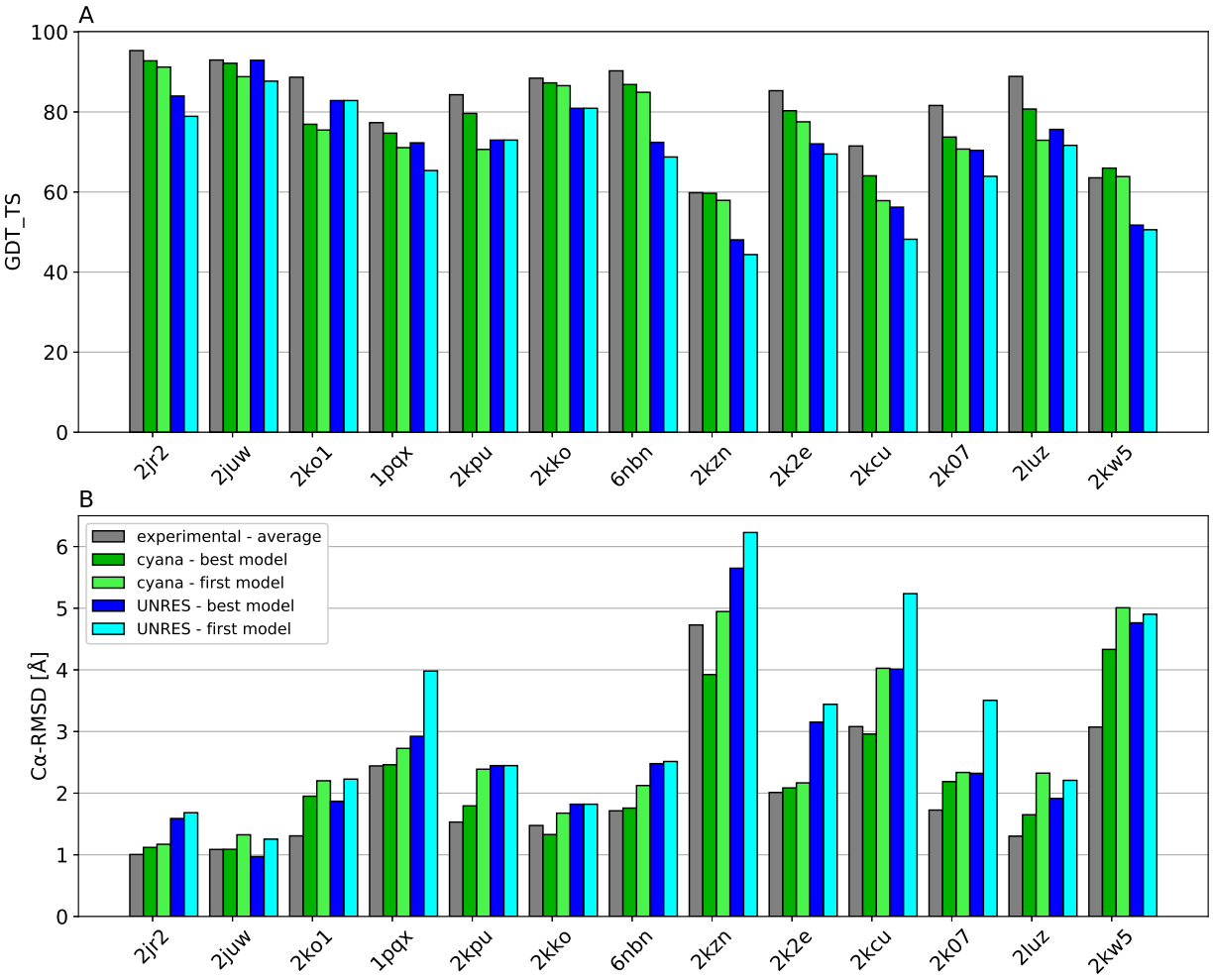


Figure 5
E.A. Lubecka, A. Liwo
J. Comput. Chem.

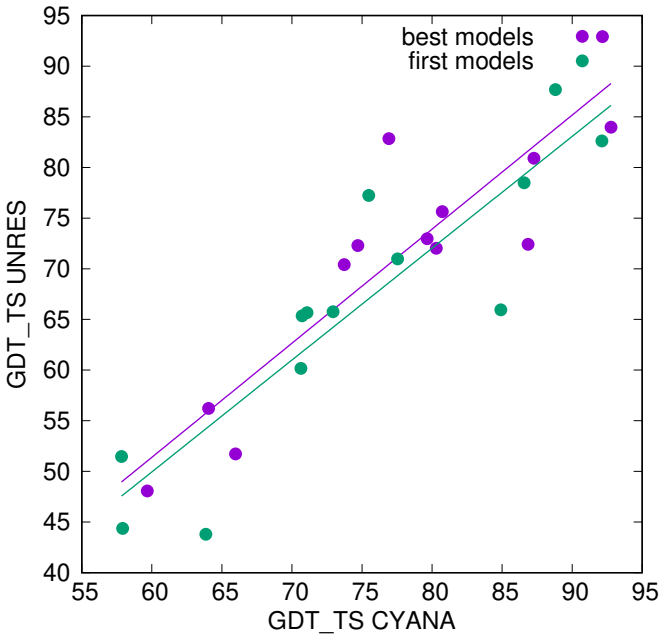


Figure 6
E.A. Lubecka, A. Liwo
J. Comput. Chem.

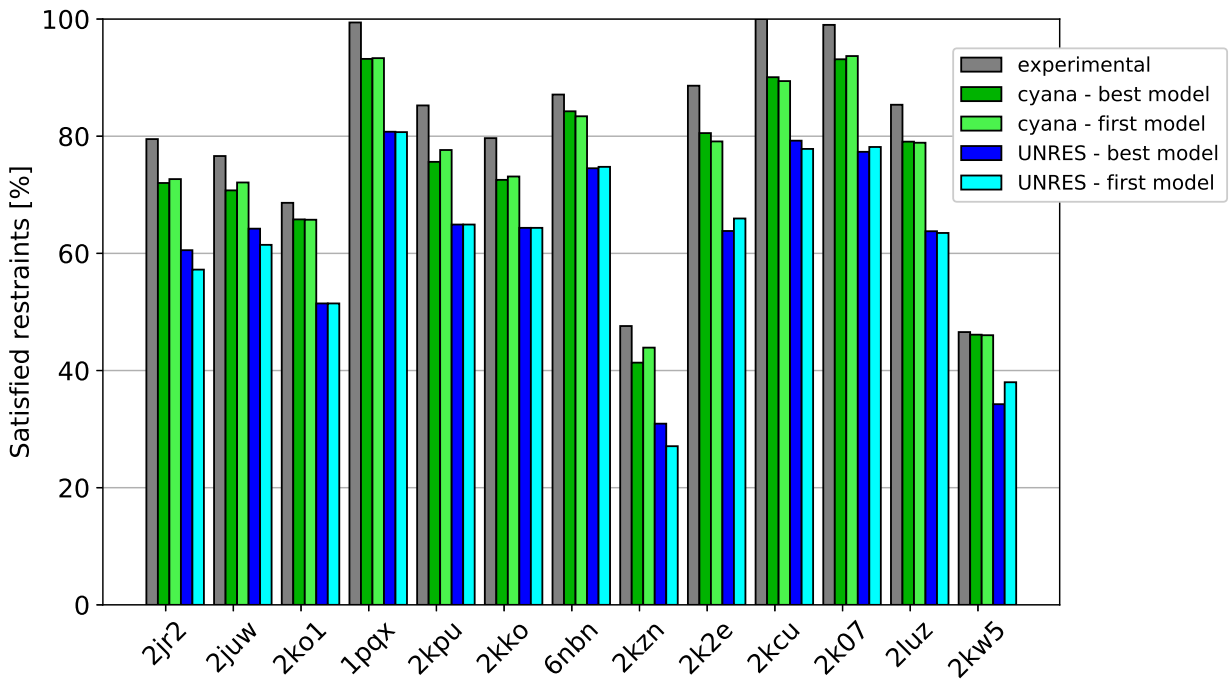


Figure 7
E.A. Lubecka, A. Liwo
J. Comput. Chem.

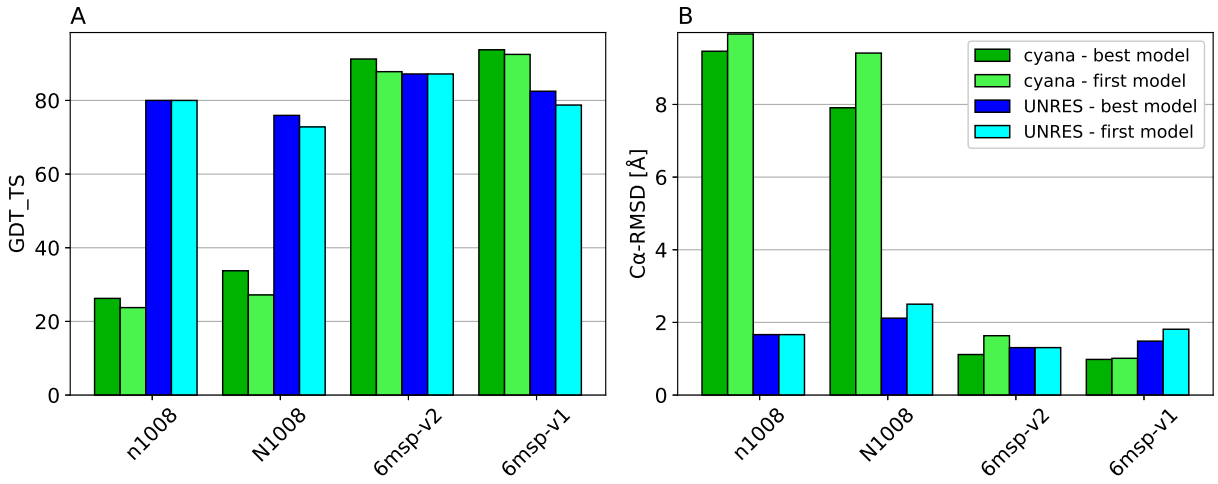


Figure 8
E.A. Lubecka, A. Liwo
J. Comput. Chem.

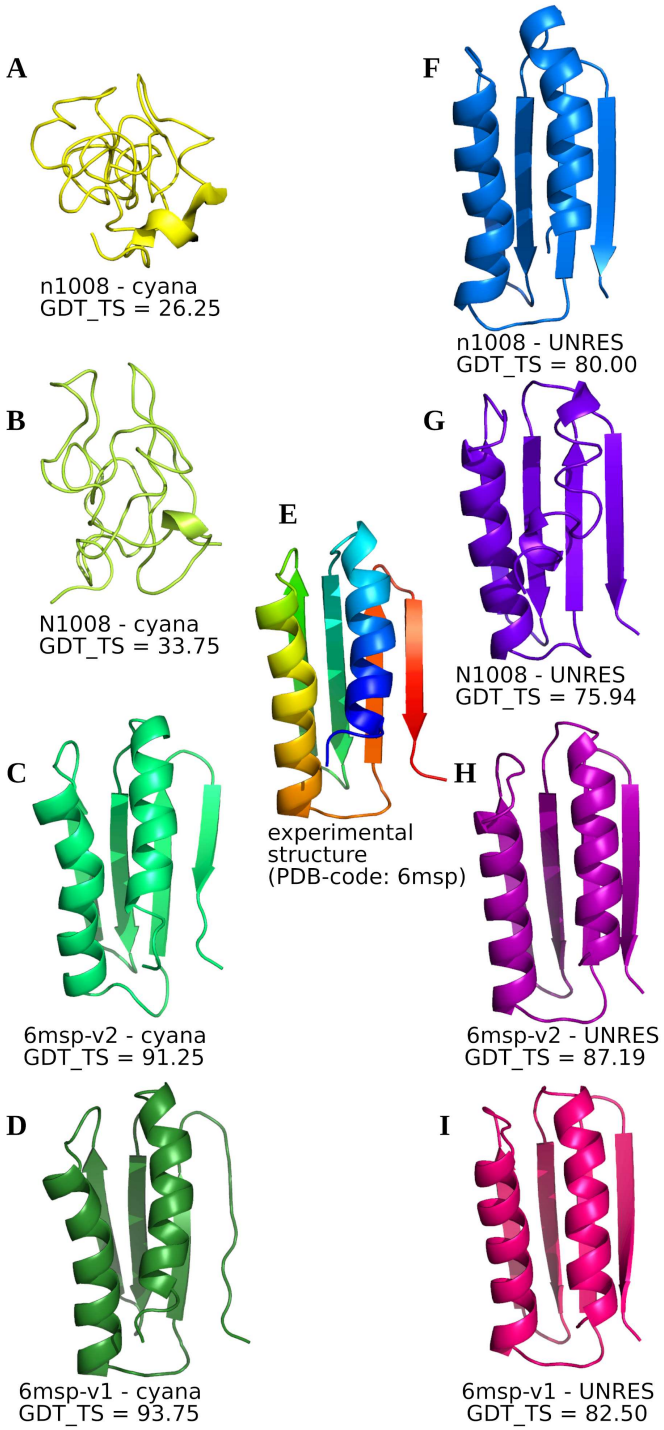


Figure 9
E.A. Lubecka, A. Liwo
J. Comput. Chem.

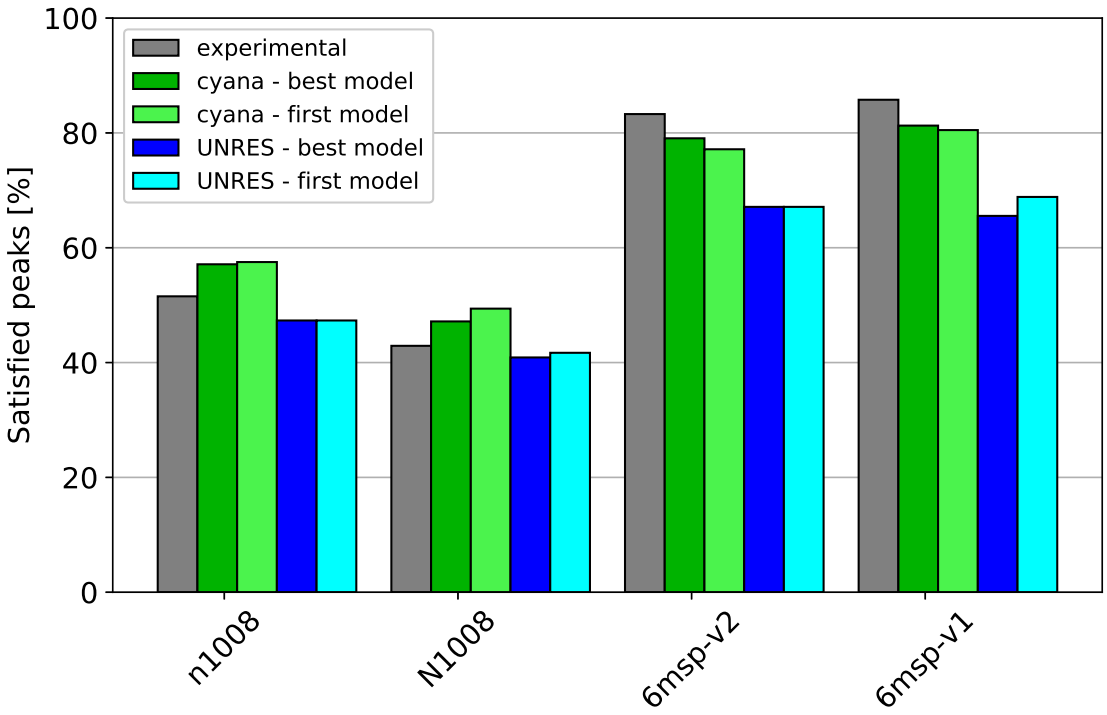


Figure 10
E.A. Lubecka, A. Liwo
J. Comput. Chem.

Table 1: Characteristics of the 13 proteins of the Montelione/NEF Benchmark Data Set[31] and of the corresponding experimental NMR distance restraints. The X-ray structures were used as reference structures in the computation of percentages of satisfied distance restraints.

PDB code		Structure	Number of			Satisfied	Average
NMR	crystal	type	residues	peaks	assignments	peaks[a] [%]	difference[b]
2jr2*	2ota*	α	76	1508	1508	79.51	0.49
2juw*	2qti	α	80	1484	1484	76.62	0.63
2ko1*	3ibw *	$\alpha+\beta$	88	5263	5263	68.63	0.66
1pqx	2ffm	$\alpha+\beta$	91	1544	1544	99.42	0.04
2kpu	3lyw	β	96	841	841	85.26	0.62
2kko*	3gw2	$\alpha+\beta$	108	2612	2612	79.67	0.56
6nbn	6og0	α	123	1645	1915	87.11	0.53
2kzn	3e0o	$\alpha+\beta$	147	624	624	47.60	0.80
2k2e	3cpk	$\alpha+\beta$	158	1125	1125	88.62	0.60
2kcu	3e0h	$\alpha+\beta$	166	1209	1209	100.00	0.00
2k07	3evx	$\alpha+\beta$	175	3627	3627	99.01	0.25
2luz	4fpw	$\alpha+\beta$	182	2714	2714	85.37	0.52
2kw5	3mer	$\alpha+\beta$	202	1121	1121	46.57	0.80

$*$ Dimers;

[a] Restraints for which at least one assignment is valid (the measured interproton distance in the crystal structure is less or equal than the respective upper boundary);

[b] Average difference between the upper-boundary values and the respective interproton distances of the 6msp structures for the violated restraints. For ambiguous restraints, the smallest differences were considered.

Table 2: Characteristics of the 6msp-v1, 6msp-v2 (unambiguous), n1008 and N1008 (ambiguous) NMR distance restraints sets of the *de novo* designed Foldit3 protein[32]. The mean 6msp structure was used as the reference structure in the computation of percentages of satisfied distance restraints.

NMR set	n1008	N1008	6msp-v2	6msp-v1
Number of assignments	26626	2128	2444	1766
Number of peaks	3351	493	1256	1666
Assignments/peak	7.95	4.32	1.95	1.06
Not satisfied peaks[a] [%]	48.46	57.09	16.72	14.23
Satisfied peaks[b] [%]	51.54	42.91	83.28	85.77
Average difference [Å][c]	4.21	3.57	0.92	0.78

[a]Peaks for which none of the assignments is valid (the measured inter-proton distances in the experimental structure are greater than the respective upper boundaries);

[b]Peaks for which at least one assignment is valid (the measured inter-proton distances in the experimental structures are less than or equal to the respective upper boundaries);

[c]Average differences between the upper-boundary values and the respective interproton distances for the violated restraints. For ambiguous restraints, the smallest differences are considered.