

Core packing of well-defined x-ray and NMR structures is the same

Alex T. Grigas^{1,2}, Zhuoyi Liu^{2,3}, Lynne Regan⁴, and Corey S. O’Hern^{1,2,3,5,6}

¹*Graduate Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, 06520, USA*

²*Integrated Graduate Program in Physical and Engineering Biology, Yale University, New Haven, Connecticut, 06520, USA*

³*Department of Mechanical Engineering and Materials Science, Yale University, New Haven, Connecticut 06520, USA*

⁴*Institute of Quantitative Biology, Biochemistry and Biotechnology, Centre for Synthetic and Systems Biology, School of Biological Sciences, University of Edinburgh*

⁵*Department of Physics, Yale University, New Haven, Connecticut 06520, USA*

⁶*Department of Applied Physics, Yale University, New Haven, Connecticut 06520, USA*

Manuscript Pages: 12

Total Manuscript Figures: 6

Total Manuscript Tables: 0

Supporting Information Pages: 12

Total Supporting Information Figures/Tables: 7/1

Abstract: Numerous studies have investigated the differences and similarities between protein structures determined by solution NMR spectroscopy and those determined by x-ray crystallography. A fundamental question is whether any observed differences are due to differing methodologies, or to differences in the behavior of proteins in solution versus in the crystalline state. Here, we compare the properties of the hydrophobic cores of high-resolution protein crystal structures and those in NMR structures, determined using increasing numbers and types of restraints. Prior studies have reported that many NMR structures have denser cores compared to those of high-resolution x-ray crystal structures. Our current work investigates this result in more detail, and finds that these NMR structures tend to violate basic features of protein stereochemistry, such as small non-bonded atomic overlaps and few Ramachandran and side chain dihedral angle outliers. We find that NMR structures solved with more restraints, and which do not significantly violate stereochemistry, have hydrophobic cores that have a similar size and packing fraction as their counterparts determined by x-ray crystallography at high-resolution. These results lead us to conclude that, at least regarding the core packing properties, high-quality structures determined by NMR and x-ray crystallography are the same, and the differences reported earlier are most likely a consequence of methodology, rather than fundamental differences between the protein in the two different environments.

Significance: Dense packing of hydrophobic residues is key to protein structure and stability. Previously, it has been noted that structures solved by NMR spectroscopy have denser cores than x-ray crystal structures. Here, we calculate the core size and packing fraction of NMR structures with experimental restraints in the PDB. Their cores are typically smaller, but denser. However, NMR structures with accurate stereochemistry possess core packing properties that are nearly identical to high-resolution x-ray crystal structures.

Availability: https://github.com/agrigas115/NMR_code

Keywords: hydrophobic core | protein structure | protein design

1 Introduction

We seek to further clarify the key physical properties of proteins that determine their structure and function. This physics-based approach is complementary to that of the remarkably successful recent machine learning methods that predict protein structure from sequence^{1,2}. We focus our analyses on protein hydrophobic cores, because the core region essentially defines the folded structure and thermodynamic stability of proteins^{3,4}.

Over the past three decades, protein structures have been determined primarily by x-ray crystallography and solution NMR spectroscopy. Prior studies have reported differences between structures determined by NMR and those determined by x-ray crystallography⁵⁻¹². A fundamental question is whether the observed differences are due to differing methodologies, or to differences in the behavior of proteins in solution versus in the crystalline state.

For structures determined by x-ray crystallography, it is universally accepted that a good structure has a low R-factor and that high-resolution structures represent a more accurate picture of the structure than low-resolution structures. In previous work, we analyzed a non-redundant set of high-resolution x-ray crystal structures (with resolution $< 1.8 \text{ \AA}$) and found that the cores (residues with zero solvent accessible surface area) of these proteins represent about 8% of the total number of residues in the protein. In addition, the packing fraction, ϕ , in protein cores is $\sim 0.55 \pm 0.01$ ^{13,14}. In earlier work, using a subset of NMR structures with a large number of distance restraints per residue, we concluded that the cores of NMR-determined protein structures had a higher packing fraction than those of high-resolution protein structures determined by x-ray crystallography¹². Denser core packing in NMR structures has also been previously reported in other studies^{5,15}. However, for structures determined by NMR, there is no equivalent universal definition of resolution or metric of quality¹⁶. Here, we filter the NMR dataset by a wide array of structural metrics and track how the physical properties of the protein core vary.

Currently, there are approximately 12,000 protein structures solved via solution-NMR spectroscopy available in the Protein Data Bank (PDB). In this article, we focus on those structures that have restraints accessible via the PDB, which is roughly half. We first analyzed the core packing of all individual structures in these NMR bundles. We found that on average there is a lower fraction of residues in the core, f_c , than in high-resolution x-ray crystal structures. However, the packing fraction, ϕ , of the cores of protein structures determined by NMR is higher than that of the cores determined by x-ray crystallography. Moreover, these results also hold when comparing the structures of the same protein, determined by both NMR and x-ray crystallography.

To investigate the possible origins of the differences seen in NMR core packing compared to high-resolution x-ray crystal structures, we filtered the dataset by a variety of structural metrics. We find that NMR bundles solved with over 10 distance restraints per residue have a fraction of residues in the core that is similar to that of high-resolution x-ray crystal structures, but the average packing fraction of those cores is higher than that observed in the cores of protein structures determined by x-ray crystallography. We investigated this phenomenon further by selecting NMR structures that satisfy common stereochemical metrics, such as low Clashscores and few Ramachandran and sidechain dihedral angle outliers. Applying these filters results in only 72 bundles of the original 6,449. The f_c and packing fraction of the remaining 72 bundles is the same as that observed in high-resolution protein x-ray crystal structures. These results suggest that protein structures are fundamentally the same in solution as in crystals, and that previously noted differences in the core are a consequence of the structure-determination methodology.

2 Results

First, we assembled datasets of x-ray crystal and NMR protein structures from the PDB. For the x-ray crystal structures, we used a dataset of 5,261 high-resolution x-ray crystal structures culled from the PDB using PISCES^{17,18} with resolution $< 1.8 \text{ \AA}$, sequence identity cutoff of $< 20\%$ and R-factor cutoff of < 0.25 . For the NMR structures, we retrieved all NMR bundles from the PDB that have available restraint data (i.e. a subset of distance restraints derived from the nuclear Overhauser effect (NOE), dihedral angle restraints from J-couplings, and bond-vector orientational restraints from residual dipolar couplings (RDCs)) using NMR2GMX¹⁹, resulting in 6,449 NMR bundles with a total of 127,959 individual structures. We refer to these 127,959 NMR structures as the entire NMR dataset. We subsequently filter this dataset according to the number of NOE distance restraints per residue and by several stereochemical validation metrics. Our goal is to understand how the core properties depend on these parameters.

We define core residues as those whose relative solvent accessible surface area, rSASA, is $\leq 10^{-3}$. We calculate the packing fraction, ϕ , of core residues, which is defined as the ratio of the volume of the residue to the volume of the Voronoi polyhedron that encloses it²⁰. We have shown in previous studies of high-resolution x-ray crystal structures that ϕ for core residues is $\sim 0.55 \pm 0.01$. (We also found that ϕ increases with decreasing rSASA²¹, until it reaches a plateau at $\phi \sim 0.55$ for rSASA $< 10^{-3}$.) (See Methods for more details.)

In Fig. 1 (A), we compare the distributions of the average packing fraction $\langle \phi \rangle$ for core residues in the high-resolution x-ray crystal structure dataset and the entire NMR dataset. For the x-ray crystal structures, we average over the core residues in each protein structure, resulting in a single value of $\langle \phi \rangle$ per protein. For the NMR structures, $\langle \phi \rangle$ is calculated by averaging over all core residues across all structures in each bundle, resulting in a single $\langle \phi \rangle$ per protein. We find that for the entire NMR dataset the average value of $\langle \phi \rangle$ is 0.57, (two standard deviations above the value of 0.55 for x-ray crystal structures), and it is clear in Fig. 1 (A) that the distribution of $\langle \phi \rangle$ is very broad, and distinctly skewed towards more densely packed structures.

The average packing fraction in the core, $\langle \phi \rangle$, is essentially independent of the total number of residues in the protein, N . In contrast, the fraction of the protein that is core, f_c , increases with increasing N . (See Supporting Information.) Because f_c depends on N , it is important to compare datasets with the same length distribution. The proteins in the NMR data set tend to be smaller than those in the dataset of x-ray crystal structures, therefore to calculate $P(f_c)$, we resampled the x-ray crystal dataset so that it possesses a $P(N)$ that is similar to that for the NMR dataset for $N < 300$. The resampling was obtained by averaging over 10^3 independent trials. $P(f_c)$ for the NMR dataset and the resampled x-ray crystal structure dataset with $N < 300$ is plotted in Fig. 1 (B). It is clear that $P(f_c)$ has a strong peak for $f_c \sim 0$ (i.e. no core residues) for the NMR dataset. In contrast, $P(f_c)$ shows that most x-ray crystal structures of these lengths have a core of $f_c \sim 0.05$ on average. In summary, when we analyze all structures in the PDB determined by NMR, for which restraint data is available, they on average possess smaller, more densely packed cores compared to high-resolution x-ray crystal structures. Interestingly, smaller, but more densely packed cores have also been found in many computationally generated low-scoring structures with incorrect backbone placement submitted to the Critical Assessment of protein Structure Prediction (CASP) competitions¹⁴.

To investigate the discrepancies in the core packing properties in greater detail, we compared protein structures that had been determined by both x-ray crystallography and NMR. We assembled a dataset of 702 pairs with both a high-resolution x-ray crystal structure ($< 2.0 \text{ \AA}$) and an NMR structure with available restraints, where the pairs have $> 90\%$ sequence similarity. (See Methods.) In Fig. 2, we show a scatter plot of the differences in the average core packing fraction and fraction

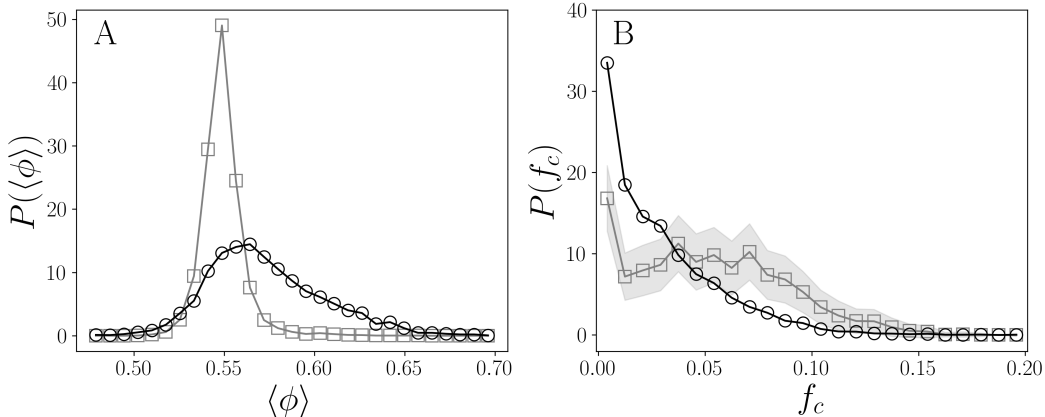


Figure 1: (A) The distribution of the packing fraction of core residues $P(\langle\phi\rangle)$ for high-resolution x-ray crystal structures (grey squares) and NMR structures (black circles). For the x-ray crystal structures, $\langle\phi\rangle$ is averaged over all core residues for each protein and for the NMR structures, $\langle\phi\rangle$ is averaged over all core residues for all structures in the bundle for each protein. (B) The distribution $P(f_c)$ for the resampled x-ray crystal structures (grey squares with shading based on the standard deviation) averaged over 10^3 resampling trials and for NMR structures (black circles), where the protein size is limited to $50 < N < 300$.

of core residues between the NMR and x-ray crystal structures, where $\Delta\langle\phi\rangle = \langle\phi\rangle_{\text{NMR}} - \langle\phi\rangle_{\text{x-ray}}$ and $\Delta\langle f_c\rangle = \langle f_c\rangle_{\text{NMR}} - \langle f_c\rangle_{\text{x-ray}}$. Consistent with Fig. 1, the majority of NMR structures have smaller, more densely packed cores compared to their x-ray crystal structure counterparts, i.e. most data points occur in the upper left corner. However, there is also data in the three other quadrants.

These results raise the interesting question of whether the differences in the size and packing fraction of the core represent true differences between protein structures when they are in crystalline states versus in solution? Or are the differences a consequence of the methods used to determine the atomic structure? Since NMR bundles are determined by a set of restraints (i.e. distance restraints derived from NOEs, dihedral angle restraints from J-couplings, and bond-vector orientational restraints from RDCs), it is reasonable to assume that the more restraints used, the higher the quality of the NMR structure. Note that while requiring a large number of restraints is necessary, previous studies have shown that it is not sufficient for identifying high-quality NMR structures^{22–24}. Nevertheless, it is of interest to track how the core properties vary with the number of NOE measurements per residue. In a typical NMR structure there are many more NOE distance restraints than any other type of restraint, so we chose to analyse their effect on core properties.

We show the distribution $P(N_r/N)$ of the number of NOE distance restraints per residue in the NMR dataset in Fig. 3 (A). There are many NMR structures with $N_r/N \sim 1$, but the vast majority have $N_r/N \sim 10$ -25, and a few have $N_r/N \gtrsim 30$. In Fig. 3 (B), we plot the distribution of the average packing fraction of core residues, $P(\langle\phi\rangle)$, in NMR structures with different values of N_r/N . We find that $P(\langle\phi\rangle)$ does not depend strongly on N_r/N and differs from $P(\langle\phi\rangle)$ for high-resolution x-ray crystal structures for all values of N_r/N . In contrast, the distribution of the average fraction of core residues, $P(f_c)$, in NMR structures becomes increasingly similar to that of the high-resolution x-ray crystal structure dataset for $N_r/N \gtrsim 20$, as shown in Fig. 3 (C). There is no correlation between N and N_r/N and therefore we used the same resampling of the x-ray crystal structure database in Fig. 1. The number of NOE distance restraints per residue has been reported to be only an approximate proxy of the quality of the protein structure, since the NOE

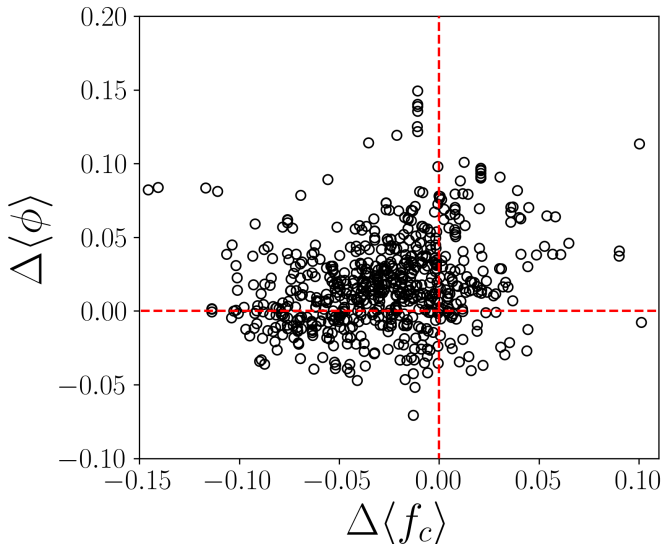


Figure 2: Scatter plot of the differences in the packing fraction $\Delta\langle\phi\rangle$ and fraction of core residues Δf_c between the NMR and x-ray crystal structures in the paired dataset.

distance restraints can be redundant²⁵. However, when we perform the same analyses for the more informative inter-residue distance restraints involving side chain atoms, the results are the same. (See the Supporting Information.)

We were interested in investigating if the packing fraction of core residues in NMR structures changed over time as the methodologies for NMR structure determination have improved^{26–28}. In Fig. 4 (A), we show the average packing fraction of core residues $\langle\phi\rangle$ versus the year that the NMR structures were deposited in the PDB. We find that $\langle\phi\rangle$ was largest before 2004 ($\langle\phi\rangle \sim 0.59$), decreased to $\langle\phi\rangle \sim 0.57$ after 2004, and has remained nearly constant over the past decade. One possible cause of the elevated values of $\langle\phi\rangle$ in the NMR structures, relative to the value of 0.55 observed in high-resolution x-ray crystal structures, is overlap of non-bonded atoms, which can increase the apparent packing fraction. It is therefore of interest to note that the Clashscore, defined by the MolProbity software²⁸ as the number of non-bonded atomic overlaps that are greater than 0.4 Å per 1,000 atoms, has decreased over time for NMR structures, with a rapid decrease around 2004 (the year MolProbity was first published) and then a slow, but steady decrease towards the average value for high-resolution x-ray crystal structures over the last decade. (See Fig. 4 (B).) The wwPDB NMR Validation Task Force has emphasized that validation metrics, such as the Clashscore and Ramachandran and side chain dihedral angle outliers, used to determine the quality of x-ray crystal structures are appropriate for assessing NMR structures¹⁶. In particular, these metrics should be similar for all high-quality protein structures, i.e. structures that are obtained from x-ray crystallography and solution NMR spectroscopy^{26,29}. When we remove NMR structures that have a Clashscore larger than the average value plus one standard deviation of the Clashscore for high-resolution x-ray crystal structures (i.e. Clashscore > 10.5), we find that $\langle\phi\rangle$ decreases significantly, but $\langle\phi\rangle$ for these NMR structures with minimal or no overlap of non-bonded atoms, remains above the value for high-resolution x-ray crystal structures.

These results indicate that while removing NMR structures that contain unphysical atomic overlaps accounts for some of the overpacking found in NMR structures, $\langle\phi\rangle$ for NMR structures is still above the value for x-ray crystal structures. Thus, there is another reason, in addition to atomic

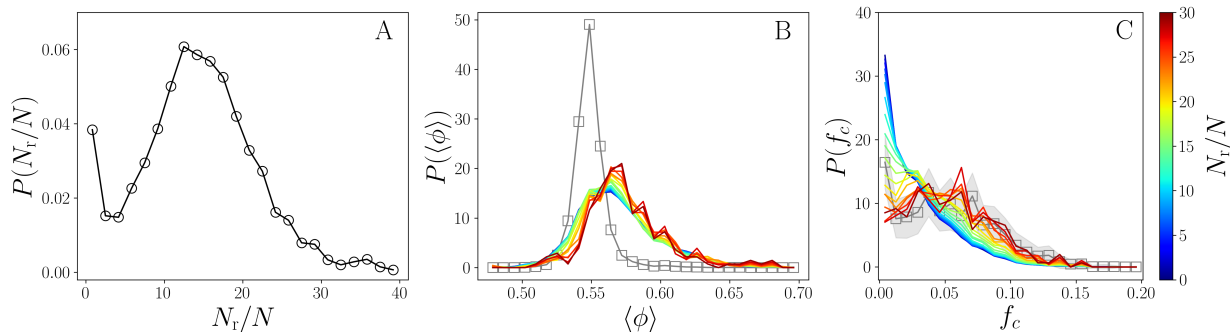


Figure 3: (A) The distribution $P(N_r/N)$ of the number of distance restraints per residue N_r/N for the NMR dataset. (B) The distribution $P(\langle\phi\rangle)$ of the average core packing fraction and (C) distribution $P(f_c)$ of the fraction of core residues for NMR structures with N_r/N greater than a given value (indicated by the color from 0 in violet to 30 in dark red). In (B) and (C), we also show $P(\langle\phi\rangle)$ and $P(f_c)$ for the resampled high-resolution x-ray crystal dataset (grey squares with shading that indicates one standard deviation).

overlaps, that is responsible for the elevated packing fraction observed in some NMR structures. Although many of the metrics are inter-related, we considered several additional quality metrics, and analysed how the properties of the cores of the protein structures depend on them. To this end, we explicitly consider the fraction of Ramachandran and sidechain dihedral angle outliers, bond length and bond angle outliers, fraction of backbone chemical shifts assigned, and strength of the magnet used. If we filter the NMR dataset using each of these metrics one at a time, $\langle\phi\rangle$ for NMR structures decreases toward the value for high-resolution x-ray crystal structures and the $P(f_c)$ distribution for NMR structures more closely resembles that of the x-ray crystal structure distribution. However, whether or not the structure was solved with RDCs makes little difference for the core packing properties. (See Supporting Information.)

For determining the validation metric cutoffs, we sought a compromise between strict cutoffs near the high-resolution x-ray crystal structure values and loose cutoffs that admit more data. When we filter the NMR dataset so that the NMR structures 1) possess values for the Clashscore, Ramachandran and sidechain dihedral angle outliers that are within one standard deviation of the average for the high-resolution x-ray crystal structures (i.e. Clashscore < 10.5 , Ramachandran dihedral angle outliers $< 0.5\%$ and side chain dihedral angle outliers $< 3\%$), 2) have at least 10 NOE distance restraints per residue, 3) have 60% or more backbone chemical shifts assigned, and 4) were solved on a 650 MHz magnet or stronger, $\langle\phi\rangle$ and $P(f_c)$ for the NMR structures are nearly identical to those for the high-resolution x-ray crystal structures as shown in Fig. 5.

These filters are quite stringent, as the filtered dataset contains only 72 NMR structures of the 6,449 from the entire NMR dataset. The average packing fraction of core residues in this filtered NMR dataset is $\langle\phi\rangle = 0.55 \pm 0.02$ compared to the value of $\langle\phi\rangle = 0.55 \pm 0.01$ for the high-resolution x-ray crystal structure dataset. The distribution of the fraction of core residues is also similar for the filtered NMR structures and high-resolution x-ray crystal structures. (See Figure 5.) Similar results can be achieved when only the stereochemical criteria are considered. (See Supporting Information.) None of the filtered NMR bundles is included the x-ray/NMR paired dataset.

To investigate the origin of the deviation in the packing properties between NMR and x-ray crystal structures, we used GROMACS to carry out all-atom MD simulations of a globular protein, the chitin binding domain of chitinase from *Pyrococcus furiosus*, which has an x-ray crystal structure with a resolution of 1.7 Å (PDBID: 2CWR) and an NMR bundle with a large number of distance

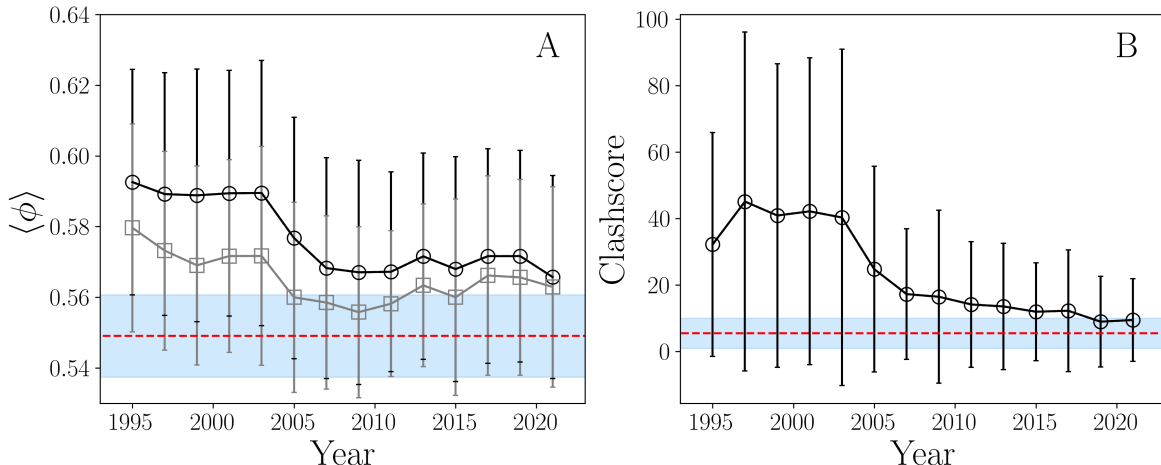


Figure 4: (A) The average packing fraction $\langle\phi\rangle$ of core residues and (B) Clashscore in NMR structures versus the year that they were deposited in the PDB for the entire NMR dataset (black circles) and those with a Clashscore that is lower than the average value for high-resolution x-ray crystal structures plus one standard deviation (grey squares). The average values and standard deviations of $\langle\phi\rangle$ and Clashscore for high-resolution x-ray crystal structures are indicated by red dashed lines and blue shading, respectively.

restraints (PDBID: 2CZLN). There are 38 individual models in the NMR bundle, solved with an average of 25 distance restraints per residue. Nevertheless, the bundle has a high Clashscore of 124. We used MD simulations to investigate if we can remove the atomic overlaps in the NMR bundle and match the NMR distance restraints to the same extent as the deposited bundle. When retrieving NOE distance restraints from the PDB, the restraints are reported as a lower bound, a first upper bound, and a second upper bound on the pairwise atomic distances. Here, we define a distance restraint as being satisfied if the distance is less than the first upper bound. In the case of ambiguous NOE restraints, where the NOE restraint is attributed to more than a single pair of atoms (e.g. methyl hydrogens), we count the restraint as satisfied if at least one of the set of pairs has a distance less than the first upper bound. To assess the resulting structures, we then calculated $\langle\phi\rangle$ and $\langle f_c\rangle$ over the MD simulation trajectory.

In Fig. 6 (A), we plot the average fraction of core residues $\langle f_c\rangle$ versus the average packing fraction of core residues $\langle\phi\rangle$ from the MD simulations and for the x-ray crystal and NMR structures. For this particular protein, the NMR bundle has both a higher packing fraction and a larger fraction of core residues than the x-ray crystal structure, but it has a large Clashscore and many Ramachandran and sidechain dihedral angle outliers. The color scale in Fig. 6 (A) indicates the average fraction F_r of the distance restraints that are satisfied (using the first upper bound).

We first ran unrestrained MD simulations, starting from a structure in the NMR bundle. This data point is plotted as a triangle in Fig. 6 (A). In the unrestrained MD simulations, 85% of the distance restraints are satisfied compared to $\sim 93\%$ in the x-ray crystal structure (x) and $\sim 100\%$ in the NMR bundle (+). Note that in the unrestrained MD simulations $\langle\phi\rangle$ decreases from $\langle\phi\rangle \sim 0.63$ for the initial NMR structure to the value for the x-ray crystal structure, $\langle\phi\rangle \sim 0.57$. However, the fraction of core residues, $\langle f_c\rangle \sim 0.03$, is much smaller than that for the initial NMR bundle with $\langle f_c\rangle \sim 0.11$ and the x-ray crystal structure with $f_c \sim 0.08$. The unrestrained MD simulation was run for 500 ns, during which time the packing properties rapidly equilibrate, and therefore the restrained

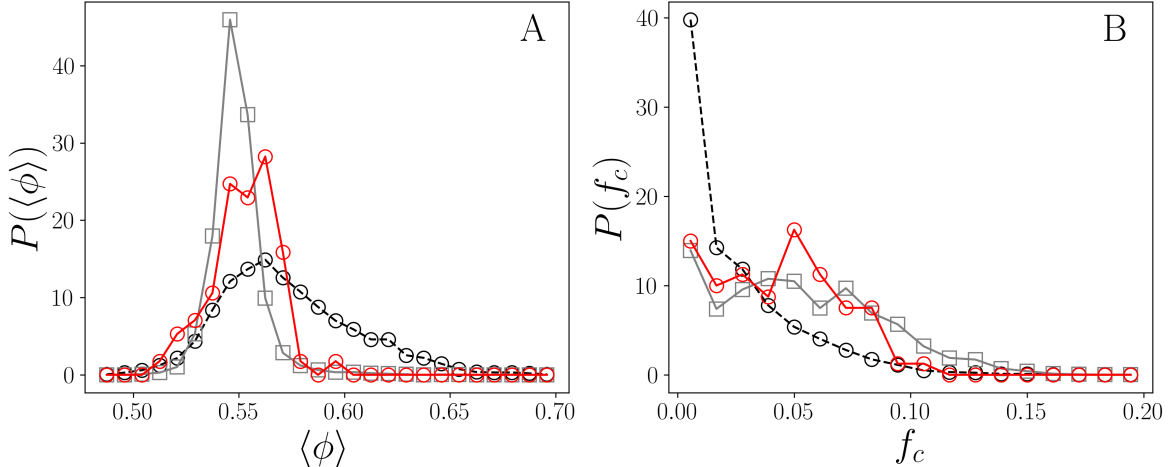


Figure 5: (A) The distribution $P(\langle\phi\rangle)$ of the average packing fraction for core residues and (B) the distribution $P(f_c)$ of the fraction of core residues for high-resolution x-ray crystal structures (grey squares), the filtered NMR structures (red circles), and the entire NMR dataset (black circles dashed).

MD simulations described below were run for shorter times. (See Supporting Information.)

We also ran MD simulations with restraints to recapitulate the NOE distances. GROMACS enforces NMR distance restraints using the following potential energy:

$$V_{dr}(r_{ij}) = \begin{cases} \frac{k}{2}(r_{ij} - r_0)^2, & r_{ij} < r_0, \\ 0, & r_0 \leq r_{ij} < r_1, \\ \frac{k}{2}(r_{ij} - r_1)^2, & r_1 \leq r_{ij} < r_2, \\ \frac{k}{2}(r_2 - r_1)(2r_{ij} - r_2 - r_1), & r_2 \leq r_{ij}, \end{cases} \quad (1)$$

where r_{ij} is the separation between NOE atom pairs i and j , r_0 , r_1 , and r_2 are the lower bound, first upper bound, and second upper bound distances for the NOE atom pairs. Additionally, the average separation $\langle r_{ij} \rangle$ is weighted by a factor of r_{ij}^{-3} multiplied by an exponentially decaying function of time with decay constant τ . (See Methods.) This restraint potential is similar to the potentials used in NMR structure determination software^{30–32}. Each restrained simulation was initialized using a single structure from the NMR bundle and run for 20 ns following energy minimization, and the NOE pair separations and packing properties were obtained every 0.1 ns and averaged. We considered several spring constants $k = 10^3$, 10^4 , and 10^5 kJ/mol/nm² and a decay time $\tau = 0.5$ ns, which increases the average fraction $\langle F_r \rangle$ of distance restraints that are satisfied from 0.86 to 0.90. For $k > 10^5$ kJ/mol/nm², $\langle F_r \rangle$ saturates and cannot reach $\langle F_r \rangle \sim 1$, which is the value for the NMR bundle deposited in the PDB. (See Supporting Information.) Restrained MD simulations using the flat-bottomed potential in Eq. 1 yield cores with similar values of $\langle f_c \rangle$ and $\langle \phi \rangle$ as the x-ray crystal structure. To obtain $\langle F_r \rangle \sim 1$, we implemented a new flat-bottomed distance restraint potential, where r_1 is set half way between the original lower bound and original first upper bound, using spring constants $k = 10^3$, 10^4 , and 10^5 kJ/mol/nm² and a decay time $\tau = 0.5$ ns and a final highly restrained simulation with a spring constant $k = 10^5$ kJ/mol/nm² and $\tau = 0$ ns. (See Methods.) With these modifications to the restraint potential and $\tau = 0$ ns, we are able to obtain structures that have $\langle F_r \rangle \sim 1$ and values of $\langle f_c \rangle$ and $\langle \phi \rangle$ that match those for the NMR bundle. However, in Fig. 6 (B), we show that as $\langle F_r \rangle$ increases toward 1, the Clashscore rises steeply, reaching values

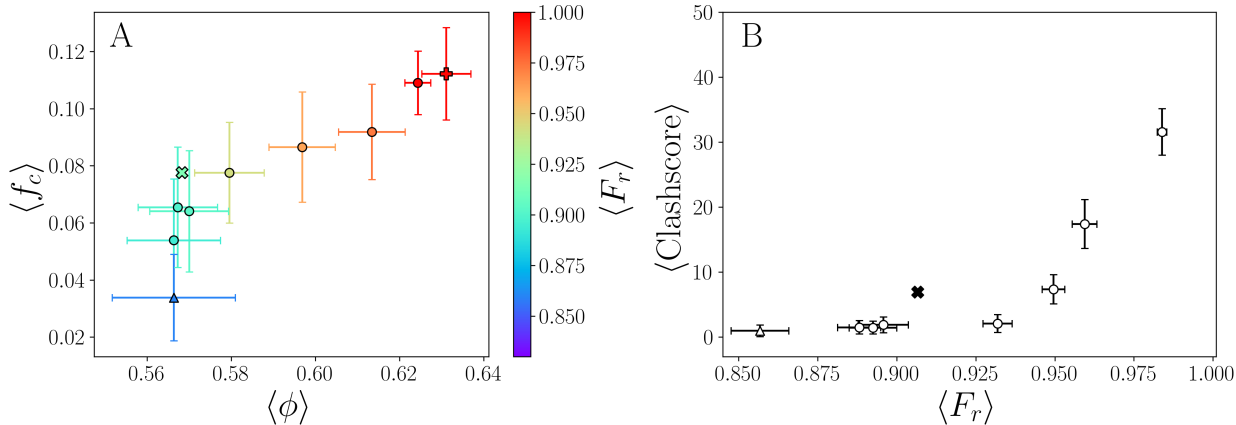


Figure 6: (A) Average packing fraction of core residues $\langle \phi \rangle$ and average fraction of core residues $\langle f_c \rangle$ for the x-ray crystal structure with PDBID: 2CRW (x), the NMR bundle with PDBID: 2CZN (+), and unrestrained (upward triangle) and restrained (circles) MD simulations colored by the average fraction of NOE distance restraints $\langle F_r \rangle$ that are satisfied. (B) The average Clashscore plotted versus the average fraction of NMR distance restraints satisfied $\langle F_r \rangle$ for the systems in (A). The data for the NMR bundle with a high Clashscore of 124 is not shown.

that are indicative of a poor quality protein structure. In addition, as shown in the Supporting Information, the same behavior is found for the fraction of Ramachandran and side chain dihedral angle outliers. The final increase in $\langle F_r \rangle$ toward 1 is caused by the increase in the number of satisfied inter-residue distance restraints involving side chain atoms. (See the Supporting Information.) By enforcing the distance restraints such that $\langle F_r \rangle \sim 1$ in the MD simulations, we obtain the same core packing properties as the deposited NMR bundle, but these structures violate metrics of protein stereochemistry. Thus, matching all of the deposited NOE restraints to the same extent as the deposited NMR bundle is likely the cause of the overpacking observed in this particular NMR bundle.

3 Discussion

An important question in protein science is to what extent are the structures of proteins in the crystalline state different from those of proteins in solution. Here, we have shown that most NMR structures deposited in the PDB possess smaller cores, and that these cores are more densely packed than those observed in high-resolution x-ray crystal structures. More densely packed cores in NMR structures have also been reported previously for smaller datasets^{5,12,15}. However, we find that by only considering NMR bundles that have a large number of restraints and that pass protein stereochemistry validation metrics (i.e. those that have small Clashscores and few Ramachandran and sidechain dihedral angle outliers), both the core size and packing fraction of these NMR structures match the values found for high-resolution x-ray crystal structures. The validation metrics are quite stringent: Only 72 NMR bundles of the original 6,449 bundles in the dataset pass the validation metrics. We interpret these results to show that the packing properties of the cores of proteins in the crystalline state (determined by high-resolution x-ray crystallography) and the cores of proteins in solution (determined by state-of-the-art NMR spectroscopy) are the same. However, since none of the filtered NMR structures have high-resolution x-ray pairs, it remains possible that particular proteins may have non-zero $\Delta \langle \phi \rangle$ or Δf_c .

A recent study reported that many NMR structures have surface loops that are often under-restrained and too floppy³³. This observation is consistent with our finding that the cores of NMR structures solved with few distance restraints are typically too small. In particular, inappropriately positioned surface residues will affect the solvent accessibility of the interior residues.

Over-packing occurs simultaneously with poor protein stereochemistry (large non-bonded atomic clashes and many dihedral angle outliers) in NMR structures in the PDB. However, this result does not rule out the possibility that denser core packing than that found in high-resolution x-ray crystal structures could occur in principle without violating protein stereochemistry. For example, packings of amino acid-shaped particles can become denser with the addition of thermal fluctuations, in contrast to packings that are rapidly compressed and energy-minimized¹². Therefore, we carried out MD simulations with NOE distance restraints to attempt to remove non-bonded atomic overlaps and dihedral angle outliers and then measured the resulting core packing properties. Our MD simulations on a particular globular protein, the chitin binding domain of chitinase from *Pyrococcus furiosus*, showed that forcing the NOE distance restraints to be satisfied to the same extent as the deposited NMR bundle leads to incorrect core packing features, non-bonded atomic clashes, and dihedral angle outliers.

The NOE restraint data clearly contains important structural information, as the NMR and the x-ray crystal structures have similar overall folds (with an average C_α RMSD = 1.6 Å) and loosely enforcing all of the NOE restraints in the MD simulations leads to core packing properties that are similar to the high-resolution x-ray crystal structure. However, enforcing the NOE distance restraints to the same extent of the deposited NMR bundle appears to be inconsistent with protein stereochemistry. The fact that simultaneously satisfying all of the NOE distance restraints leads to violations of protein stereochemistry (for this particular protein) is likely related to the fact that the conversion of NOE measurements to distance bounds is approximate and NOEs report on time-averaged atomic separations, which can be challenging to implement in protein structure determination^{23,34,35}. We emphasize that our MD simulations were performed on a single protein whose structure was determined by both x-ray crystallography and NMR (from quadrant 2 of Fig. 2). In future studies, it will be important to study protein structure pairs that represent all four quadrants from Fig. 2 and to determine what restrained MD protocol would be needed to improve the modeling of the protein core. Additionally, our unrestrained MD simulations show that the core packing fraction is recapitulated using the Amber99SB-ILDN forcefield and TIP3P water model, but the core size is not. Therefore, we will explore whether the same result is observed with different force fields and solvation models, which would indicate how well the hydrophobic effect is currently captured in MD simulations of proteins.

4 Materials and Methods

Datasets

We collected two datasets of protein structures from the Protein Data Bank (PDB). We first compiled a dataset of 5,261 high-resolution x-ray crystal structures from the PDB using PISCES^{17,18} with resolution < 1.8 Å, a sequence identity cutoff of $< 20\%$, an R-factor cutoff of < 0.25 , and lengths greater than 50 residues. We also assembled a large dataset of NMR structures from the PDB. We identified all NMR bundles that have available restraint data in the NMR-STAR format using NMR2GMX¹⁹, resulting in 6,499 NMR bundles, each with at least 10 structures in the bundle. In this dataset, 5,830 bundles have NOE distance restraints, 3,994 have dihedral angle restraints, and 507 have RDC restraints. The NMR dataset was not filtered for similarity in amino acid sequence, and thus it has multiple bundles for some proteins. PYPDB was used to retrieve all of

the validation metrics for each protein from the PDB³⁶. In addition, PYPDB was used to identify a set of NMR and x-ray crystal structure pairs. We selected pairs for which the x-ray crystal structure sequence matched $> 90\%$ via MMSeqs2³⁷, the lengths differ by $< 10\%$, the x-ray crystal structure resolution is $\leq 2.0 \text{ \AA}$, and the NMR bundle has restraint data. This dataset has 702 pairs, made up of 514 x-ray crystal structures and 525 NMR structures, as some NMR structures have multiple x-ray crystal structure matches and vice versa.

rSASA

To identify core residues, we measured each residue’s solvent accessible surface area (SASA). To calculate SASA, we use the NACCESS software package³⁸, which implements an algorithm originally proposed by Lee and Richards³⁹. To normalize the SASA, we take the ratio of the SASA within the context of the protein ($\text{SASA}_{\text{context}}$) and the SASA of the same residue (X) extracted from the protein structure as a dipeptide (Gly-X-Gly) with the same backbone and side-chain dihedral angles:

$$\text{rSASA} = \frac{\text{SASA}_{\text{context}}}{\text{SASA}_{\text{dipeptide}}}. \quad (2)$$

Core residues are classified as those that have $\text{rSASA} \leq 10^{-3}$. The choice of this threshold has been discussed in several previous studies^{20,21}.

Packing Fraction

A characteristic measure of the packing in an atomic system is the packing fraction. The packing fraction of residue μ is defined by

$$\phi_{\mu} = \frac{\nu_{\mu}}{V_{\mu}}, \quad (3)$$

where ν_{μ} is the non-overlapping volume of residue μ and V_{μ} is the volume of the Voronoi cell surrounding residue μ . The volume of the residue is determined by the atom sizes and their locations. The particular atom sizes that we used have been previously selected from the range of literature values to reproduce the backbone and side chain dihedral angle distributions found in high-resolution x-ray crystal structures^{40–42}. The Voronoi cell represents the local free space around the residue. We calculate the non-overlapping residue volume with a grid-based volume estimation. To calculate the Voronoi polyhedra for a protein structure, we use surface Voronoi tessellation, which defines a Voronoi cell as the region of space in a given system that is closer to the bounding surface of the residue than to the bounding surface of any other residue in the system. We calculate the surface Voronoi tessellations using the POMELO software package⁴³. This software approximates the bounding surfaces of each residue by triangulating points on the residue surfaces. We find that using ~ 400 points per atom, or ~ 6400 surface points per residue, gives an accurate representation of the Voronoi cells and the results do not change if more surface points are included.

Molecular Dynamics Simulations

The unrestrained and restrained all-atom molecular dynamics (MD) simulations of the globular protein, the chitin binding domain of chitinase from *Pyrococcus furiosus*, were performed using the GROMACS software package⁴⁴. The MD simulations used a cubic box with edge lengths 63 \AA , which ensures that the protein is $> 10 \text{ \AA}$ away from the box edges, and the box is filled with water molecules modeled using TIP3P⁴⁵ at neutral pH and 0.15M NaCl. Periodic boundary

conditions are applied in the x -, y -, and z -directions. Short-range van der Waals and screened Coulomb interactions were truncated at 1.2 nm, while longer-ranged electrostatics were tabulated using the Particle Mesh Ewald summation method. We performed two short energy minimization runs to relax the protein first and then the water molecules and protein together using the steepest decent method. We then simulated the system for 20 ns in the NPT ensemble at temperature $T = 300\text{K}$ and pressure $P = 1$ bar using the weakly coupled Berendsen thermostat and barostat and calculated the $\langle\phi\rangle$ and f_c every 0.1 ns. The time constant of the Berendsen thermostat was set to 2 ps and the isothermal compressibility for the Berendsen barostat was set to $4.5 \times 10^{-5}\text{bar}^{-1}$. (See the Supporting Information for more details about the equilibration procedure.) The equations of motion for the atomic coordinates and velocities were integrated using a leapfrog algorithm with a 1 fs time step. For both the unrestrained and restrained MD simulations, we used the AMBER99SB-ILDN force field^{46,47}. For the restrained MD simulations, we implemented the flat-bottomed spring potential $V_{dr}(r_{ij})$ given in Eq. 1 with spring constants $k = 10^3, 10^4$, and 10^5 kJ/mol/nm² to enforce the NOE distance restraints between atom pairs i and j . We calculate the time average NOE pairwise atomic separations as:

$$\langle r_{ij} \rangle = \left\langle \left(r_{ij}^{-3}(t - \Delta t)e^{(-\Delta t/\tau)} + r^{-3}(t) \left(1 - e^{(-\Delta t/\tau)} \right) \right)^{-1/3} \right\rangle_t, \quad (4)$$

where Δt is the time step and τ is the decay time. $\tau = 0.5$ ns for all restraint simulations, except for the most highly restrained simulation. This time-averaging method has been suggested to more accurately represent NOE distance restraints and allows for larger fluctuations in the atomic separations if NOE distance restraints are incompatible. To further increase the fraction of NOE restraints that are satisfied during the MD simulations to $\langle F_r \rangle \sim 1$, we also implemented a narrower flat-bottomed spring potential, where the first upper bound r_1 is set to the separation between the NOE atom pairs of the initial NMR structure, and $\tau = 0$ when calculating $\langle r_{ij} \rangle$.

References

- [1] Pereira J, Simpkin AJ, Hartmann MD, Rigden DJ, Keegan RM, Lupas AN (2021) High-accuracy protein structure prediction in CASP14. *Proteins Struct Funct Bioinf* 89:1687–1699.
- [2] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–589.
- [3] Dill KA (1990) Dominant forces in protein folding. *Biochemistry* 29:7133–7155.
- [4] Pace CN, Fu H, Fryar KL, Landua J, Trevino SR, Shirley BA, Hendricks MM, Iimura S, Gajiwala K, Scholtz JM, Grimsley GR (2011) Contribution of hydrophobic interactions to protein stability. *J Mol Biol* 408(3):514–528.
- [5] Garbuzynskiy SO, Melnik BS, Lobanov MY, Finkelstein AV, Galzitskaya OV (2005) Comparison of X-ray and NMR structures: Is there a systematic difference in residue contacts between X-ray and NMR-resolved protein structures? *Proteins Struct Funct Bioinf* 60:139–147.
- [6] Andrec M, Snyder DA, Zhou Z, Young J, Montelione GT, Levy RM (2007) A large data set comparison of protein structures determined by crystallography and NMR: Statistical test for structural differences and the effect of crystal packing. *Proteins Struct Funct Bioinf* 69:449–465.
- [7] Schneider M, Fu X, Keating AE (2009) X-ray vs. NMR structures as templates for computational protein design. *Proteins Struct Funct Bioinf* 77: 97–110.
- [8] Sikic K, Tomic S, Carugo O (2010) Systematic comparison of crystal and NMR protein structures deposited in the Protein Data Bank. *Open Biochem J* 4:83–95.
- [9] Mao B, Tejero R, Baker D, Montelione GT (2014) Protein NMR structures refined with Rosetta have higher accuracy relative to corresponding X-ray crystal structures. *J Am Chem Soc* 136:1893–1906.
- [10] Everett JK, Tejero R, Murthy SBK, Acton TB, Aramini JM, Baran MC, Benach J, Cort JR, Eletsky A, Forouhar F, Guan R, Kuzin AP, Lee HW, Liu G, Mani R, Mao B, Mills JL, Montelione AF, Pederson K, Powers R, Ramelot T, Rossi P, Seetharaman J, Snyder D, Swapna GVT, Vorobiev SM, Wu Y, Xiao R, Yang Y, Arrowsmith CH, Hunt JF, Kennedy MA, Prestegard JH, Szyperski T, Tong L, Montelione GT (2016) A community resource of experimental data for NMR / X-ray crystal structure pairs. *Protein Sci*, 25:30–45.
- [11] Leman JK, D’Avino AR, Bhatnagar Y, Gray JJ (2018) Comparison of NMR and crystal structures of membrane proteins and computational refinement to improve model quality. *Proteins Struct Funct Bioinf* 86:57–74.
- [12] Mei Z, Treado JD, Grigas AT, Levine ZA, Regan L, O’Hern CS (2020) Analyses of protein cores reveal fundamental differences between solution and crystal structures. *Proteins Struct Funct Bioinf* 88:1154–1161.
- [13] Gaines JC, Smith WW, Regan L, O’Hern CS (2016) Random close packing in protein cores. *Phys Rev E* 93:032415.

- [14] Grigas AT, Mei Z, Treado JD, Levine ZA, Regan L, O’Hern CS (2020) Using physical features of protein core packing to distinguish real proteins from decoys. *Protein Sci* 29:1931–1944.
- [15] Ratnaparkhi GS, Ramachandran S, Udgaonkar JB, Varadarajan R. (1998) Discrepancies between the NMR and X-ray structures of uncomplexed barstar: Analysis suggests that packing densities of protein structures determined by NMR are unreliable. *Biochemistry*, 37:6958–6966.
- [16] Montelione GT, Nilges M, Bax Ad, Güntert P, Herrmann T, Richardson JS, Schwieters CD, Vranken WR, Vuister GW, Wishart DS, Berman HM, Kleywegt GJ, Markley JL (2013) Recommendations of the wwPDB NMR Validation Task Force. *Structure* 21:1563–1570.
- [17] Wang G, Dunbrack RL (2003) PISCES: A protein sequence culling server. *Bioinformatics* 19:1589–1591.
- [18] Wang G, Dunbrack RL (2005) PISCES: Recent improvements to a PDB sequence culling server. *Nucleic Acids Res*, 33:W94–W98.
- [19] Sinelnikova A, van der Spoel D (2021) NMR refinement and peptide folding using the GRO-MACS software. *J Biomol NMR*, 75:143–149.
- [20] Treado JD, Mei Z, Regan L, O’Hern CS (2019) Void distributions reveal structural link between jammed packings and protein cores. *Phys Rev E*, 99:022416.
- [21] Gaines JC, Acebes C, Virrueta A, Butler M, Regan L, O’Hern CS (2018) Comparing side chain packing in soluble proteins, protein-protein interfaces and transmembrane proteins. *Proteins Struct Funct Bioinf* 86:581–591.
- [22] Spronk CAEM, Nabuurs SB, Bonvin AMJJ, Krieger E, Vuister GW, Vriend G (2003) The precision of NMR structure ensembles revisited. *J Biomol NMR*, 25:225–234.
- [23] Zagrovic B, van Gunsteren WF. Comparing atomistic simulation data with the NMR experiment: How much can NOEs actually tell us? *Proteins Struct Funct Bioinf* 63:210–218.
- [24] Rosato A, Tejero R, Montelione GT (2013) Quality assessment of protein NMR structures. *Curr Opin Struct Biol* 23:715–724.
- [25] Nabuurs SB, Spronk CAEM, Krieger E, Maassen H, Vriend G, Vuister GW (2003) Quantitative evaluation of experimental NMR restraints. *J Am Chem Soc* 125:12026–12034.
- [26] Vuister GW, Fogh RH, Hendrickx PMS, Doreleijers JF, Gutmanas A (2014) An overview of tools for the validation of protein NMR structures. *J Biomol NMR* 58:259–285.
- [27] Buchner L, Güntert P (2015) Increased reliability of nuclear magnetic resonance protein structures by consensus structure bundles. *Structure* 23:425–434.
- [28] Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, Deis LN, Verma V, Keedy DA, Hintze BJ, Chen VB, Jain S, Lewis SM, Arendall WB, Snoeyink J, Adams PD, Lovell SC, Richardson JS, Richardson DC (2018) Molprobity: More and better reference data for improved all-atom structure validation. *Protein Sci* 27:293–315.
- [29] Spronk CAEM, Nabuurs SB, Krieger E, Vriend G, Vuister GW (2004) Validation of protein structures derived by NMR spectroscopy. *Prog Nucl Magn Reson Spectrosc* 45:315–337.

- [30] Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. *J Magn Reson* 160: 65–73.
- [31] Schwieters CD, Kuszewski JJ, Clore GM (2006) Using Xplor-NIH for NMR molecular structure determination. *Prog Nucl Magn Reson Spectrosc* 48:47–62.
- [32] Güntert P, Automated NMR Structure Calculation With CYANA. In: *Protein NMR Techniques* (2004) *Methods in Molecular Biology*, vol 278. Humana Press, pp 353-378
- [33] Fowler NJ, Sljoka A, Williamson MP (2021) The accuracy of NMR protein structures in the Protein Data Bank. *Structure* 29:1430–1439.
- [34] Vögeli B (2014) The Nuclear Overhauser Effect from a quantitative perspective. *Prog Nucl Magn Reson Spectrosc* 78:1–46.
- [35] Chalmers G, Glushka JN, Foley BL, Woods RJ, Prestegard JH (2016) Direct NOE simulation from long MD trajectories. *J Magn Reson* 265:1–9.
- [36] Gilpin W (2015) PyPDB: a Python API for the Protein Data Bank. *Bioinformatics* 32:159–160.
- [37] Steinegger M, Söding J (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 35:1026–1028.
- [38] S. J. Hubbard and J. M. Thornton. *Naccess*, 1993.
- [39] Lee B, Richards FM (1971) The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol* 55:379–400.
- [40] Zhou AQ, O’Hern CS, Regan L (2012) The power of hard-sphere models: Explaining side-chain dihedral angle distributions of Thr and Val. *Biophys J* 102:2345–2352.
- [41] Zhou AQ, O’Hern CS, Regan L (2014) Predicting the side-chain dihedral angle distributions of nonpolar, aromatic, and polar amino acids using hard sphere models. *Proteins Struct Funct Bioinf* 82:2574–2584.
- [42] Caballero D, Jukka Määttä, Zhou AQ, Sammalkorpi M, O’Hern CS, Regan L (2014) Intrinsic α -helical and β -sheet conformational preferences: A computational case study of alanine. *Protein Sci* 23:970–980.
- [43] Weis S, Schonhoferer PWA, Schaller FM, Schroter M, Schroder-Turk GE (2017) Pomelo: a tool for computing Generic Set Voronoi Diagrams of Aspherical Particles of Arbitrary Shape. *EPJ Web of Conferences* 140:06007
- [44] Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1-2:19–25.
- [45] Jorgensen WL (1981) Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water. *J Am Chem Soc* 103:335–340.
- [46] Best RB, Hummer G (2009) Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J Phys Chem B*, 113:9004–9015.
- [47] Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct Funct Bioinf* 78:1950–1958.