Title: Collective Repacking of Protein Cores

Authors: J.C. Gaines[1,2], A. Virrueta[3, 2], S.J. Fleishman[5], C.S. O'Hern[3, 1, 2,4,6] and L. Regan[1,2,7,8]

1. Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA
2. Integrated Graduate Program in Physical and Engineering Biology (IGPPEB), Yale University, New Haven, Connecticut 06520, USA
3. Department of Mechanical Engineering & Materials Science, Yale University, New Haven, Connecticut 06520, USA
4. Department of Physics, Yale University, New Haven, Connecticut 06520, USA
5. Department of Biomolecular Sciences, Weizmann Institute of Science, Rehovot 76100, Israel
6. Department of Applied Physics, Yale University, New Haven, Connecticut 06520, USA
7. Department of Molecular Biophysics & Biochemistry, Yale University, New Haven, Connecticut 06520, USA
8. Department of Chemistry, Yale University, New Haven, Connecticut 06520, USA

Abstract

Protein core repacking provides a meaningful test of computational protein design software. A study of different protein design software showed that they are much more successful at predicting side chain conformations of core compared to surface residues. Motivated by this observation, we investigated to what extent an energy function that includes only stereochemical constraints and repulsive hard-sphere interactions can correctly repack protein cores. Specifically, we tested the ability of the hard-sphere model to predict the side chain conformations of core residues in ~200 proteins. For both single residue and collective repacking, the hard-sphere model accurately recapitulates the observed side chain conformations for Ile, Leu, Phe, Thr, Trp Tyr and Val. This result is important because it shows that there are no alternative, sterically allowed side chain conformations of core residues. Further, we analyzed the same set of protein cores using the protein design software, Rosetta. Both the hard-sphere model and Rosetta performed equally well on Ile, Leu, Phe, Thr, and Val. However, the hard-sphere model performed better on Trp and Tyr, while Rosetta performed better on Ser. This study emphasizes that for many residues steric interactions alone determine side chain conformations in protein cores.

Introduction

A grand challenge in biology is to design new protein-protein interactions for many potential applications including point of care diagnostics (Rusling *et al*. 2010), sensors for proteinaceous biological warfare agents (Sapsford *et al*., 2008), and more effective vaccines (Correia *et al*., 2014). Computational protein design offers a way to test a large number of amino acid sequences efficiently and rapidly. Moreover, computational protein design provides an additional route to gain fundamental insights into protein structure. It is important to benchmark the predictions made by computational design software against known protein crystal structures. A frequently used test for computational design software is side chain conformation recovery, where the side chains are removed from a protein crystal structure and the software attempts to recover the observed side chain conformations of all residues (Peterson *et al*., 2014). In protein core repacking, the side chains of core residues are removed simultaneously, and the design software samples all side chain dihedral angle combinations, predicts the optimal combination, and compares it to the observed structure. (See Fig. 1.) Protein core repacking is a particularly meaningful test for computational design software that is used to assess mutations to protein cores (Borgo *et al*., 2012) and design new protein-protein interactions (Fleishman *et al*., 2011).

In recent work, Peterson and coworkers (Peterson *et al*. 2014) performed side chain recovery for ~200 proteins using six different protein design software suites (SCWRL (Krivov *et al*., 2009), OSCAR (Liang *et al*., 2011), RASP (Miao *et al*., 2011), Rosetta (Kuhlman *et al*., 2000), Sccomp (Eyal *et al*., 2004), and FoldX (Guerois *et al*., 2002)). The key component of computational protein design software is the energy function, which can include many terms: stereochemistry (potentials that enforce equilibrium bond lengths and angles derived from small molecule crystal structures) plus up to eight additional terms---statistical potentials derived from backbone-dependent side chain rotamer libraries (Dunbrack and Cohen 1997, Shapovalov and Dunbrack 2011); repulsive and attractive van der Waals atomic interactions; hydrogen bonding; electrostatics; desolvation energies; disulfide bond energy (RASP-specific), and an *ad hoc* pairwise residue potential (Rosetta-specific). The energy functions differ in the relative weights assigned to each of these terms.

Overall, protein design software performs well for protein side chain recovery. Specifically, Peterson, *et al*. found that all six software packages obtain higher accuracy for their predictions for the side chain dihedral angle conformations for core residues compared to surface residues. In addition, all of the software packages achieve higher accuracy when predicting χ_1 alone (90-95% within 40°) compared to predictions of side chain dihedral angle combinations, *e.g*. χ_1 and χ_2 (82-87% within 40° degrees for each). Because the rotamer recovery prediction accuracy for all of the protein design software tested is higher for core residues, here we investigate to what extent an energy function that only includes stereochemistry and repulsive hard-sphere atomic interactions can repack protein cores. To enable a residue-by-residue comparison with a well-established protein design software package, we performed collective core repacking calculations using both the stereochemistry plus hard-sphere model and Rosetta.

For our core repacking studies, we employed the Dunbrack 1.0Å and HiQ54 databases as our benchmark sets of proteins. Our calculations involve several steps. We first identify the core residues in each protein, where a core residue is defined as a residue with no atom that is solvent accessible. We next identify clusters of interacting core residues. (See Fig. 2.) A residue is defined as a member of an interacting cluster if any side chain dihedral angle conformation brings that residue into contact with any other residue in the cluster, but not residues of another cluster. We will first describe studies of single residue rotations, where we sample all side chain dihedral angle combinations of a single core residue, keeping the side chain conformations of all other residues fixed to their crystal structure values. We evaluate the energy of each side chain dihedral angle combination and compare the lowest energy side chain dihedral angle combination for each core residue (Leu, Ile, Met, Phe, Ser, Thr, Trp, Tyr, Val) to the observed values. We find that the hard-sphere model achieves a prediction accuracy of greater than 90% (within 30°) for all residues except Met (84%) and Ser (38%). (See Fig. 4.) We compare the results of single residue rotations to the results of collective residue rotations, which provides insight into the number of possible ways to pack interacting core residues. In doing so, we address the question: Are the side chain dihedral angle combinations observed in protein crystal structures the only way that core residues can be arranged without steric clashes?

For collective residue rotations, we simultaneously rotate the side chains of all residues in a given interacting cluster. We perform these calculations for all clusters in all proteins. We observe the same high prediction accuracy for collective residue rotations as we did for single residue rotations for the hard-sphere model: greater than 90% accuracy (within 30°) for all core residues except for Met (77%) and Ser (36%). (See Figs. 5 and 6.) For combined rotations, Rosetta and the hard-sphere model give the same high prediction accuracy (≥ 90% within 30°) for Ile, Leu, Thr, Phe, and Val (Fig. 7). The hard-sphere model performs slightly better on aromatic residues than Rosetta, whereas Rosetta achieves much higher accuracy for Ser. We discuss the reasons for these differences in the Results section. The success of the hard-sphere model in repacking protein cores emphasizes that steric interactions play a dominant role in determining structure of protein cores. The cases for which the hard-sphere model does not achieve high prediction accuracy allow us to identify when additional interactions are necessary to predict side chain conformations.

Materials and Methods

Datasets of protein crystal structures and core residues

We use the Dunbrack 1.0Å database (Wang and Dunbrack, 2003, 2005) of high-resolution protein crystal structures as the basis for our protein core repacking studies. The Dunbrack 1.0Å database contains 221 proteins with resolution ≤ 1.0Å, side chain B-factors per residue ≤ 30Å^2, R-factor ≤ 0.2, and sequence identity < 50%. As a way to model the system at non-zero temperature and improve the statistics, variations in bond lengths and angles are implemented by replacing each side chain with different instances of the side chain taken from the Dunbrack 1.7Å database, each with an independent set of side chain bond lengths and angles (Zhou *et al.,* 2014). The Dunbrack 1.7Å database contains ~800 proteins with resolution ≤ 1.7Å (Dunbrack *et al.,* 1997). Additional studies were performed on a second database, the `HiQ54' database (Leaver-Fay *et al.,* 2013), which contains 54 non-redundant, single-chain monomeric proteins with resolution and MolProbity score < 1.4Å.

We have limited our analysis of side chain conformations to residues in protein cores. We have identified all core residues in the Dunbrack 1.0Å database using a method described previously

(Gaines *et al*., 2016; Caballero *et al*., 2016). In brief, non-core atoms are identified that are on the surface of the protein or near an interior void with a radius of ≥ 1.4 Å. Core residues are then defined as any residue containing only core atoms (including hydrogen atoms). The numbers of each amino acid that occur as core residues in the Dunbrack 1.0Å database are given in Table 1.

Hard-sphere model

As described in previous work (Zhou *et al*., 2014; Gaines *et al*., 2016), the 'hard-sphere' model treats each atom *i* as a sphere that interacts pairwise with all other non-bonded atoms *j* via the purely repulsive Lennard-Jones potential:

$$U_{RLJ}(r_{ij}) = \frac{\epsilon}{72}\left[1 - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^6\right]^2 \Theta\left(\sigma_{ij} - r_{ij}\right),$$

where r_{ij} is the center-to-center separation between atoms *i* and *j*, $\Theta\left(\sigma_{ij} - r_{ij}\right)$ is the Heaviside step function, ϵ is the energy scale of the repulsive interactions, $\sigma_{ij} = (\sigma_i + \sigma_j)/2$, and $\sigma_i/2$ is the radius of atom *i*. The values for the atomic radii (C_{sp3}, $C_{aromatic}$: 1.5 Å; C_O: 1.3Å ; O: 1.4Å; N: 1.3Å; H_C: 1.10Å; $H_{O,N}$: 1.00Å and S: 1.75Å) were obtained in prior work (Zhou *et al*., 2014) by minimizing the difference between the side chain dihedral angle distributions predicted by the hard-sphere dipeptide mimetic model and those observed in protein crystal structures for a subset of amino acid types. Hydrogen atoms were added using the REDUCE software program (Word *et al*., 1999), which sets the bond lengths for C-H, N-H, and S-H to 1.1, 1.0 and 1.3Å respectively, and the bond angles to 109.5° and 120° for angles involving C_{sp3} and C_{sp2} atoms. Additional dihedral angle degrees of freedom involving hydrogen atoms are chosen to minimize steric clashes (Word *et al*., 1999).

Predictions of the side chain conformations of single amino acids are obtained by rotating each of the side chain dihedral angles $\chi_1, \chi_2, \ldots, \chi_n$ (with a fixed backbone conformation (Liu *et al*., 2016)) and finding the lowest energy conformations of the residue, where the energy includes both intra- and inter-residue steric interactions (Figure 1 (C)-(E)). If the lowest energy conformation of the residue is degenerate (*i.e*. multiple dihedral angle configurations result in the same minimum energy), all lowest energy configurations are recorded. We then calculate the Boltzmann weight of the lowest energy side chain conformation of the residue, $P_i(\chi_1, \ldots, \chi_n) \propto$

$e^{-U(\chi_1,\ldots,\chi_n)/k_BT}$, where the temperature T/ε=10^{-2} approximates hard-sphere-like interactions. To sample bond length and angle fluctuations, each residue is replaced with random bond length and angle combinations taken from the Dunbrack 1.7Å database and the new lowest energy conformation is found. We select 50 bond length and angle variants, and for each find the lowest energy dihedral angle conformation and corresponding $P_i(\chi_1, \ldots, \chi_n)$ values. We average P_i over the variants to obtain $P_m(\chi_1, \ldots, \chi_n)$. We then compare the particular dihedral angle combination $\{\chi_1^{HS}, \ldots, \chi_n^{HS}\}$ associated with the highest value of P_m to the side chain of the crystal structure $\{\chi_1^{xtal}, \ldots, \chi_n^{xtal}\}$. To assess the accuracy of the hard-sphere model in predicting the side chain dihedral angles of residues in protein cores, we calculated

$$\Delta\chi = \sqrt{\left(\chi_1^{xtal} - \chi_1^{HS}\right)^2 + \cdots + \left(\chi_n^{xtal} - \chi_n^{HS}\right)^2}$$

If multiple side chain configurations were reported in the Protein Databank for a given protein, $\Delta\chi$ was calculated for all reported conformations with an occupancy ≥ 40%.

In addition to single residue rotations, we performed core repacking using combined rotations of interacting core residues in each protein. For the combined rotation method, all residues in an interacting cluster are rotated simultaneously (with fixed backbone conformations), and the global minimum energy conformation is identified (Figure 1 (B)). A cluster of interacting residues is defined such that side chain atoms of each residue in the cluster only interact with other residues in the cluster without interactions with the side chains of other core residues in the protein (Figure 2). Specifically, if an atomic overlap is possible between two residues without an interaction with the protein backbone also occurring, those two residues are considered to be interacting. Examples of interaction networks between core residues in interacting clusters are given in Fig. 3. (C). Ala, Gly, and Pro were excluded from this analysis since these amino acids do not possess side chain dihedral angle degrees of freedom. In addition, we did not include Cys residues because they can form disulfide bonds. The Dunbrack 1.0Å database includes 352 distinct clusters (with greater than 1 residue) with sizes given in Fig. 3. A few clusters contained 10 or more residues, but these were not included in the analyses. We also removed clusters containing the charged residues Arg, Asp, Glu, and Lys and the polar residues Asn, Gln, and His, which are rare in protein cores (less than 10% of core residues). This resulted in a total of 250

clusters and 852 residues from the Dunbrack 1.0Å database. The frequency of each amino acid in these clusters is given in Table 2. The HiQ54 database contains 50 core clusters with 2 to 15 residues per cluster. (See Figure 3 (B).)

Predictions from combined rotations for the side chain dihedral angle combinations of core residues in a given cluster are obtained by rotating each of the side chain dihedral angles $\chi_1, \chi_2, \ldots, \chi_n$ of all residues in that cluster and identifying the lowest energy side chain dihedral angle combination, where the total energy includes the repulsive Lennard-Jones interactions between atoms on a single residue as well as atoms on different residues both in the given cluster and other residues in the protein. We represented the side chain dihedral angle combinations as a tree, where each level represents an amino acid and the nodes at each level represent the allowed side chain dihedral angle conformations for the corresponding residue. We then implement a depth-first search to find the global energy minimum and the corresponding side chain dihedral angle conformation. Bond lengths and angles were varied by sampling 30 bond length and angle variants from the Dunbrack 1.7Å database. The Boltzmann weight P_i for each variant was found and averaged over the variants to obtain $P_m(\chi_1, \ldots, \chi_n)$, and $\Delta\chi$ was calculated as described above.

Rosetta Predictions

The prediction accuracy for collective core repacking using the hard-sphere model was compared to that from Rosetta (Leaver-Fay *et al*., 2011) on all core clusters. We first generated relaxed structures for each protein studied, which were obtained by running Rosetta's fast relax protocol with backbone constraints that maintain the positions of the backbone heavy atoms near the crystal structure locations (Tyka *et al*., 2011, Liu *et al*., 2016). 50 relaxed structures were produced and the five lowest energy structures were chosen for core repacking. Rotamer sampling on all side chain dihedral angles was set to the maximum value (*i.e*. the original rotamer value ±0.25 standard deviations). For each of the 5 relaxed structures, we performed repacking and selected the output conformation with the lowest Rosetta energy. $\Delta\chi$ was calculated for each residue as described above, resulting in five $\Delta\chi$ values for each residue, which were used to obtain the average fraction F($\Delta\chi$) of residues with $\Delta\chi$ less than 10°, 20°, and

30°. A sample Rosetta script and a description of the calculations of the error bars for $F(\Delta\chi)$ can be found in the Supplemental Material.

Results

In previous studies, we have shown that the hard-sphere dipeptide model can recapitulate the observed side chain dihedral angle distributions of nonpolar, aromatic and polar amino acids (Cys, Ile, Leu, Phe, Ser, Thr, Trp, Tyr and Val) (Zhou, *et al.,* 2014). In more recent work (Caballero *et al.,* 2016), we showed that the hard-sphere model including both intra- and inter-residue interactions could predict the side chain dihedral angle conformations of single residues in protein cores. The prediction accuracy (within 20° of the observed structure) was greater than 90% for Ile, Leu, Phe, Thr, Trp, Tyr and Val. This prior work focused on rotations of the side chains of individual residues in protein cores. Here, we expand this work to examine the predictions obtained by the hard-sphere model from simultaneous rotations of multiple residues in protein cores, as well as to a larger database of protein crystal structures. To enable a detailed comparison with a well-established protein design software package, we compare the predictions of the hard-sphere model to those from Rosetta.

In Figure 4, we investigate the accuracy of the hard-sphere model in predicting the side chain dihedral angles of individual residues in protein cores. For each amino acid (Ile, Leu, Met, Phe, Ser, Thr, Trp, Tyr and Val), we find the percentage of residues for which the predicted side chain dihedral angle conformation is within 10°, 20° and 30° of the crystal structure value. Consistent with our prior results, the hard-sphere model accurately predicts (≥ 90% within 30°) the side chain dihedral angle combinations of single residues in the context of the protein for Ile, Leu, Phe, Thr, Trp, Tyr, and Val. This result emphasizes that the purely repulsive hard-sphere model can accurately predict the side chain dihedral angle combinations for nonpolar and uncharged amino acids. The quantitative values of our results differ slightly from those found in Caballero *et al*. (2016) because in the current study we use the much larger Dunbrack 1.0Å database of protein crystal structures.

We find that the hard-sphere model is unable to predict two residues that we studied with high accuracy: Ser and Met. Our results for Met are consistent with those found in Virruetta *et al*. (2016). In this prior work, we found that local steric interactions were insufficient to predict the shape of the $P(\chi_3)$ distribution for Met. It was necessary to add attractive atomic interactions to the hard-sphere model to reproduce the observed $P(\chi_3)$. Here, using only repulsive interactions, we are only able to predict ~80% of Met residues within 30°. Our results for Ser (only 38% within 30°) are also consistent with our prior work in Caballero *et al*. (2016). We speculate that because the side chain of Ser is small, hydrogen-bonding interactions must be included to correctly place its side chain. In contrast, we suggest that the more bulky Thr and Tyr side chains causes steric interactions to determine the positioning of their side chains, even though they are able to form hydrogen bonds (Zhou *et al*. 2012).

We obtain similar results when we perform combined rotations of core residues using the hard-sphere model (Figure 5 and 6). Single and combined rotations have the same prediction accuracy, which shows that there are very few arrangements of the residues in a protein core that are sterically allowed and that the side chain conformations of most core residues are dominated by packing constraints. Slightly lower prediction accuracy is found for a few residues using combined rotations, which is due to the fact that finding the conformation corresponding to the global energy minimum may improve the accuracy for one residue, while the lowering the accuracy for another residue in the same cluster. We also performed single and collective repacking on the HiQ54 dataset and found similar accuracies for both single and combined rotations for both datasets. (These results are shown in the Supplementary Material.)

We now compare the results of core repacking (with combined rotations) using the hard-sphere model to that using Rosetta (Fig. 7). For the residues Ile, Leu, Phe, Thr, and Val, the hard-sphere model achieves a similar prediction accuracy to that obtained by Rosetta. The largest differences occur for Ser: Rosetta gives 85% (within 30°), while the hard-sphere model gives 36% (within 30°). We previously speculated that because the side chain of Ser is small, hydrogen-bonding interactions are more important for Ser than for Thr. Rosetta includes hydrogen-bonding interactions, which is likely the reason for its higher prediction accuracy.

Rosetta obtains prediction accuracies of 85% and 78% (within 30°) for Trp and Tyr, respectively, while the hard-sphere model obtains 95% and 94% (within 30°) for Trp and Tyr, respectively (Fig. 7). To further investigate this difference, we calculated $\Delta\chi$ for χ_1 and χ_2 separately for both residues (Fig. 8). For Trp, the hard-sphere model performs slightly better than Rosetta at predicting χ_1 and χ_2. For Tyr, Rosetta and the hard-sphere model perform equally well for χ_1, but the hard-sphere model performs better for χ_2.

For Met, both the hard-sphere model and Rosetta obtain prediction accuracies below 80% for $\Delta\chi < 30°$. From Fig. 8, we find that the hard-sphere model and Rosetta accurately predict χ_1 and χ_2 (above 90% within 30°), but have prediction accuracies for χ_3 below 80% within 30°. In previous work, we showed that χ_1 and χ_2 of Met are well predicted using the hard-sphere model, whereas χ_3 is not (Virrueta *et al*., 2016). This result holds true for both the dipeptide model as well as in the context of the protein core. In this previous study, we found that the addition of attractive atomic interactions improves the prediction of χ_3 for Met. The current results for single and collective core repacking showing that the hard-sphere model yields low χ_3 prediction accuracy for Met are consistent with the previous results. The low prediction accuracy for χ_3 for Met using Rosetta is surprising since Rosetta includes a statistical potential in the energy function derived from backbone-dependent side chain dihedral angle rotamer libraries.

Discussion

In this article, we showed several key results. First, single and collective core repacking using the hard-sphere model give the same prediction accuracies for the side chain conformations of six of the most common core residues. This result implies that there are no alternate sterically allowed conformations of core residues other than those in the crystal structure. If alternative sterically allowed conformations existed, we would have found them using the collective repacking method and thus the prediction accuracy would have dramatically decreased relative to the value for single residue rotations. It does not. Second, the hard-sphere model obtains prediction accuracies that are as high or higher than Rosetta for Ile, Leu, Phe, Thr, Val, Trp, and Tyr. Thus, hard-sphere interactions are dominant in determining side chain conformations for

these residues. The hard-sphere model and Rosetta both give < 80% prediction accuracy for Met, which is caused by poor positioning of the side chain dihedral angle χ_3. This result is surprising for Rosetta since its energy function includes statistical potentials that are based on backbone-dependent side chain dihedral angle rotamer libraries. Rosetta performs better on Ser, presumably because Rosetta includes hydrogen-bonding interactions. Interestingly, Thr and Tyr, can both hydrogen bond, but can be accurately predicted using the hard-sphere model alone presumably because they both have bulkier side chains than Ser. Third, we have shown that an energy function that only includes stereochemistry and repulsive hard-sphere atomic interactions can repack protein cores with high accuracy.

Why do the hard-sphere model and six computational protein design software packages studied in Peterson *et al*. obtain similar high prediction accuracies for many core residues? We hypothesize that the weights of the repulsive atomic interactions and statistical potentials derived from backbone dependent side chain dihedral angle rotamer libraries are large in comparison to other terms in the energy functions of the six software packages. Alternatively, but less likely, there could be a special combination of weights of the terms in the energy functions of the six software packages that give rise to the same result.

Acknowledgements

The authors thank R. L. Dunbrack, Jr. and J. S. and D. C. Richardson for providing the Dunbrack 1.0Å and HiQ54 databases of protein crystal structures, respectively, and for their interest in this work and helpful discussions. The authors also thank D. Caballero for his initial development of codes for repacking protein cores using the hard-sphere model.

Funding

The authors acknowledge support from the National Library of Medicine Training Grant No. T15LM00705628 (J.C.G.); National Science Foundation (Grant NSF-PHY-1522467 to L.R. and C.S.O.; Grant NSF-DMR-1307712 to L.R.); the Ford Foundation Pre-Doctoral Fellowship program (to A.V.); the National Science Foundation Graduate Research Fellowships program (to A.V.); and the Raymond and Beverly Sackler Institute for Biological, Physical and Engineering Sciences (to J.C.G., A.V., C.S.O., and L.R.). The authors benefited from the Facilities and staff of the Yale University Faculty of Arts and Sciences High Performance Computing Center and acknowledge the National Science Foundation (Grant No. CNS 08-21132) that in part funded acquisition of the computational facilities.

Figures

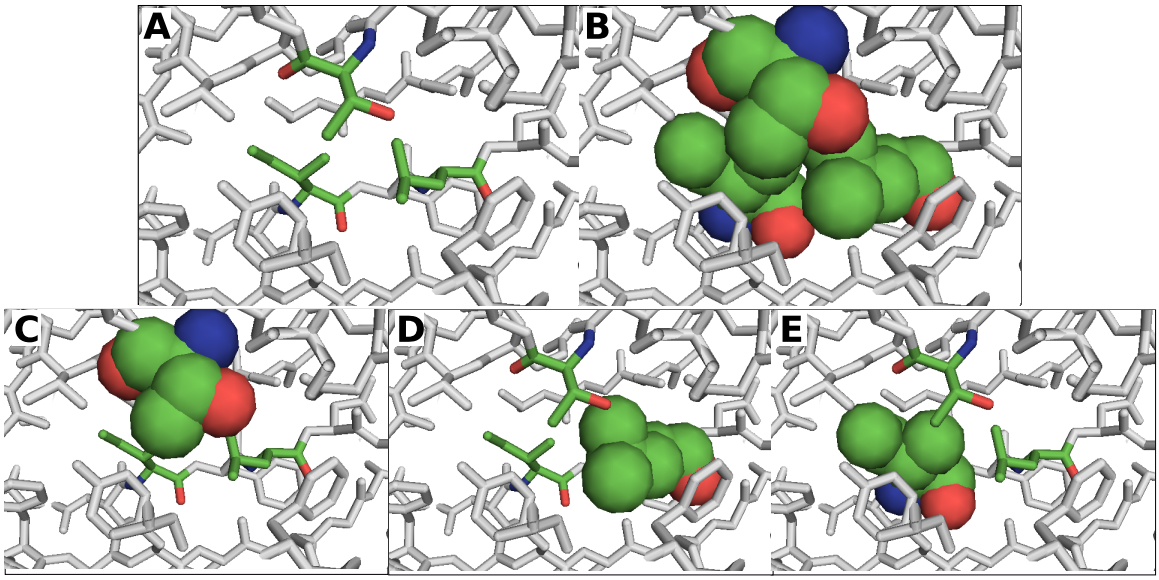

Figure 1: Illustration of single and combined rotations for protein core repacking studies using PDB: 1C7K. (A) We show a cluster of 3 interacting core residues (Thr, Leu, Val) shaded in green using stick representation with the rest of the protein shaded in grey. (B) For combined rotations, all three core residues, with atoms represented as spheres (C: green, N: blue, O: oxygen), are rotated simultaneously and the repulsive steric interactions are calculated between atoms in the three moving residues as well as between atoms in the residues with fixed side chains. (C-E) For single rotations, only one core residue ((C) Thr, (D) Leu, or (E) Val) in the cluster is rotated at a time, while the others remain fixed. Steric interactions are calculated between atoms in the moving residue and atoms of all other residues in the protein. In all cases, each atom in the protein is represented as a sphere, but stationary atoms are shown here as sticks to highlight the residues that are not rotated.

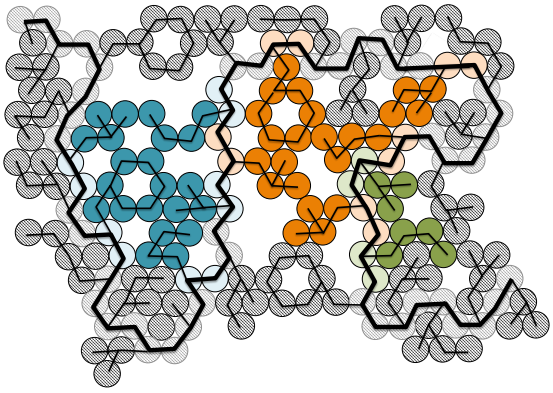

Figure 2: Schematic in two dimensions of a protein that contains three core clusters. Each amino acid is represented by disk-shaped atoms that are connected by lines. The protein backbone is indicated by a thick

black line, and the thinner lines form the side chains. Each residue contains two side chain atoms and between one and seven side chain atoms. "Surface" residues are shaded grey. Any residue that is completely surrounded by other atoms is designated as a core residue. Each core cluster contains residues that interact with each other but do not interact with the side chains of residues in another cluster. For example, the cluster in blue has atoms that touch the backbone of the cluster in orange, but these atoms do not interact with the side chains of residues in the orange cluster without clashing with the backbone first. The three core clusters shown here contain five (blue), five (orange), and two (green) residues.

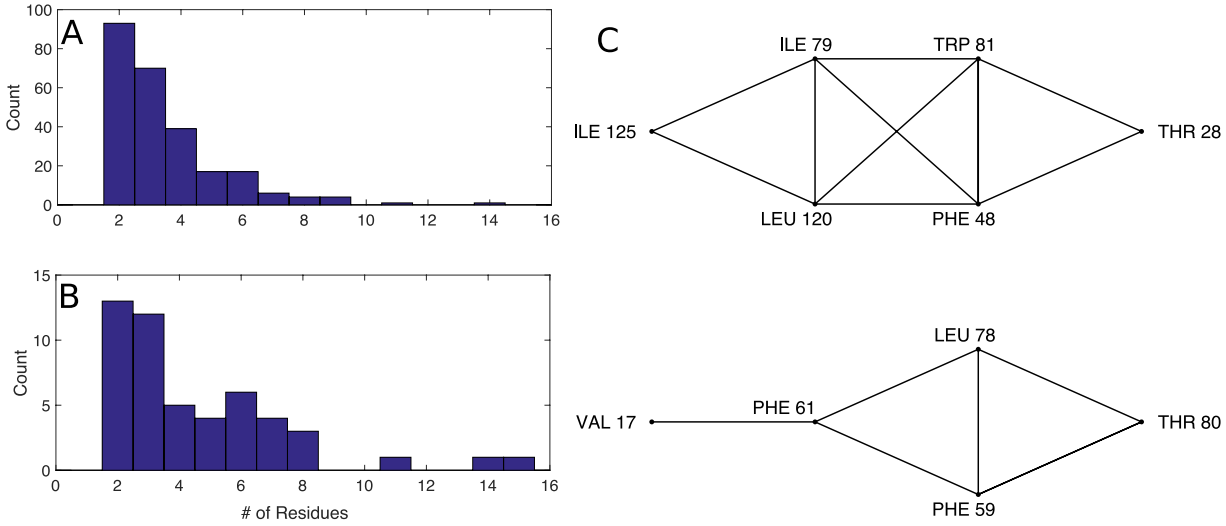

Figure 3: The distribution of cluster sizes in the (A) Dunbrack 1.0Å and (B) HiQ54 databases. Each cluster is defined as a set of residues in a protein core that interact with each other, but not with any other core residues. (C) Examples of interaction networks based on two clusters of core residues from protein PDB:1T3Y. The clusters contain eight (top) and five (bottom) residues respectively. Each line in the network indicates interactions between two residues. For example, in the top cluster Ile 125 interacts with Ile 79 and Leu 120, but does not interact with Trp 81 or Val 17.

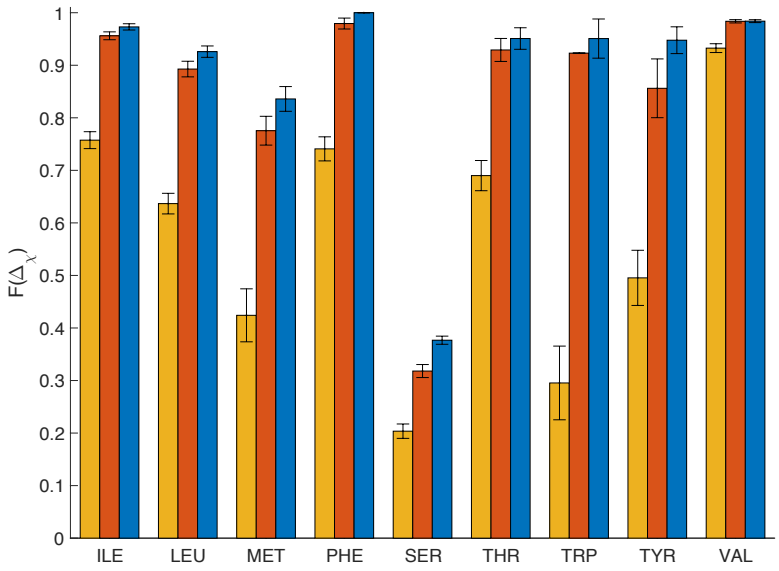

Figure 4: Single residue rotations in the context of the protein core: The fraction (F(Δχ)) of each residue type for which the hard-sphere model prediction of the side chain conformation is Δχ < 10° (yellow), 20° (red), or 30° (blue) from the crystal structure for core residues in the Dunbrack 1.0Å database.

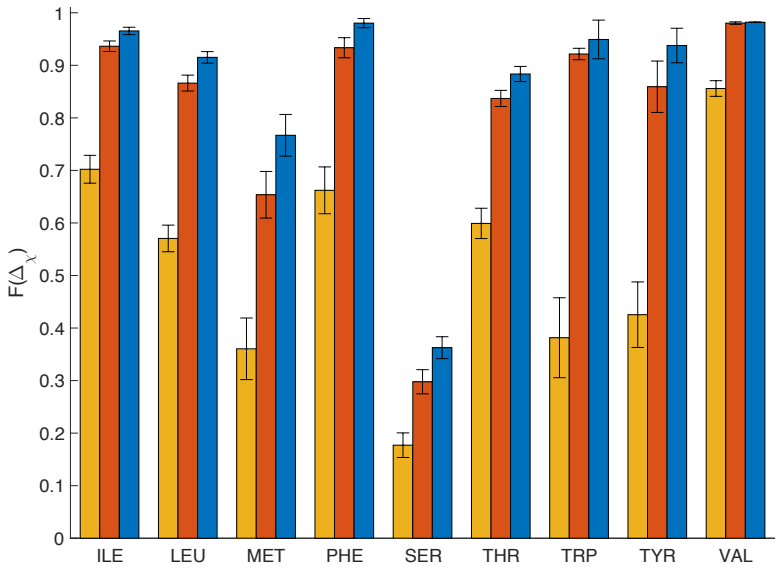

Figure 5: Combined rotations in the context of the protein core: The fraction (F(Δχ)) of each residue type for which the hard-sphere model prediction of the side chain conformation is Δχ < 10° (yellow), 20° (red), or 30° (blue) from the crystal structure for core residues in the Dunbrack 1.0Å database.

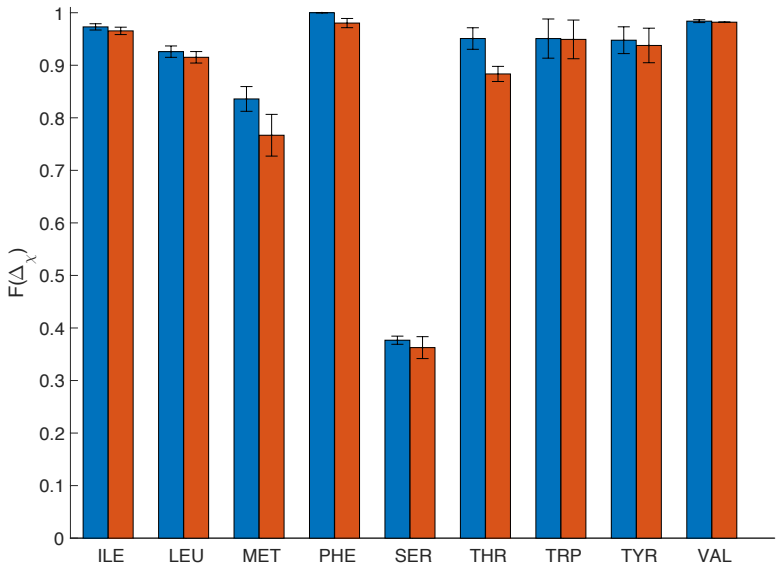

Figure 6: Comparison of the accuracy of single and combined rotations for core residues in the Dunbrack 1.0Å database. Each bar shows the fraction of residues for which the hard-sphere model prediction of the side chain conformation is $\Delta\chi < 30°$ for single (blue) or combined (red) rotations.

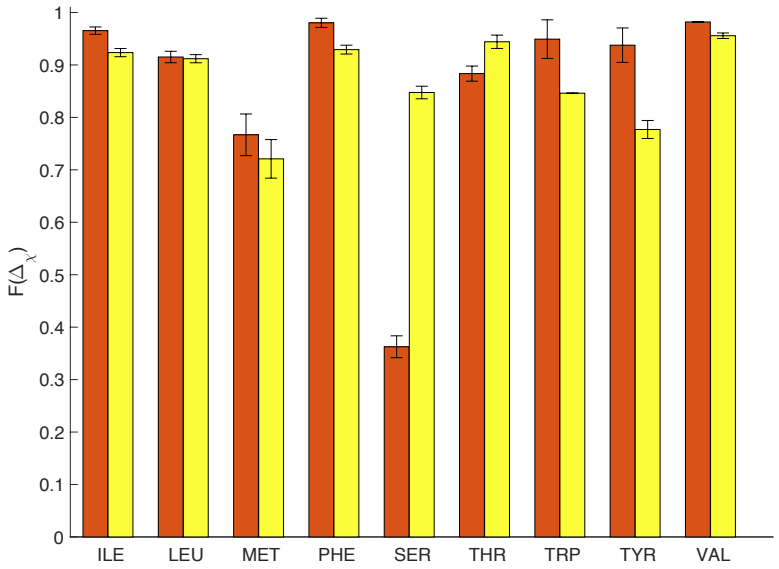

Figure 7: Comparison of the accuracy of combined rotations for core residues in the Dunbrack 1.0Å database using the hard-sphere model (red) and Rosetta (yellow). Each bar shows the fraction F($\Delta\chi$) of residues for which the model prediction was $\Delta\chi < 30°$.

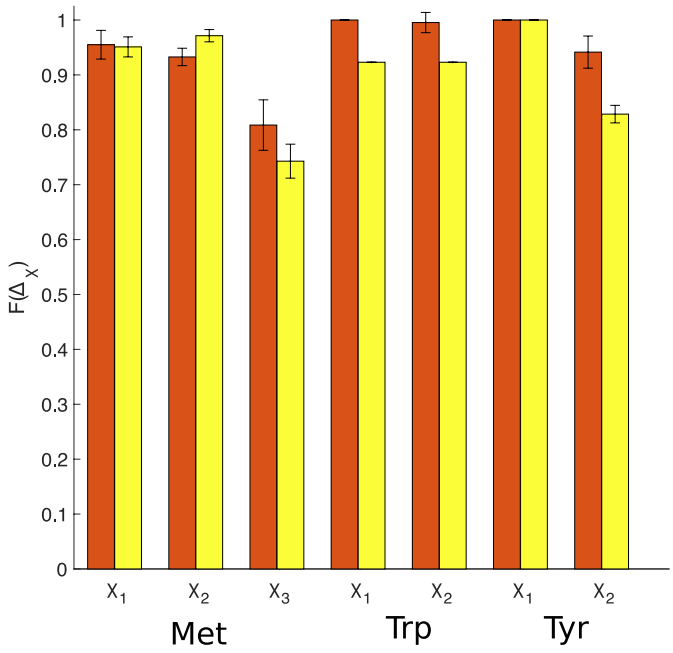

Figure 8: Comparison of the accuracy of combined rotations for core Met, Trp, and Tyr residues in the Dunbrack 1.0Å database using the hard-sphere model (red) and Rosetta (yellow). Each bar shows the fraction F(Δχ) of residues for which the model prediction was Δχ < 30° for each side chain dihedral angle separately.

Tables

Amino Acid	No. in Dunbrack 1.0Å database
Ala	529
Asn	50
Asp	78
Arg	6
Cys	142
Gln	17
Glu	31
Gly	453
His	24
Ile	453
Leu	355

Lys	3
Met	90
Phe	141
Pro	63
Ser	193
Thr	136
Trp	28
Tyr	69
Val	438
Total	849

Table 1 : The number of each amino acid designated as core in the Dunbrack 1.0Å database.

Amino Acid	No. in clusters in Dunbrack 1.0Å database
Ile	163
Leu	179
Met	50
Phe	70
Ser	68
Thr	48
Trp	13
Tyr	29
Val	229
Total	849

Table 2: The number of each uncharged amino acid found in interacting clusters (with size greater than 1 residue) in the Dunbrack 1.0Å database

References

1. Borgo, B. and Havranek, J. J. (2012) Proc. Natl. Acad. Sci. USA, **109**, 1494-1499.
2. Caballerao, D., Virrueta, A., O'Hern, C.S. and Regan, L. (2016) *PEDS*, **29,** 367-376.
3. Correia, B. E., Bates, J. T., Loomis, R. J., Baneyx, G., Carrico, C., Jardine, J. G., Rupert, P., Correnti, C., Kalyuzhiniy, O., Vittal, V., Connell, M. J., Stevens, E., Schroeter, A., Chen, M., Macpherson, S., Serra, A. M., Adachi, Y., Holmes, M. A., Li, Y., Klevit, R. E., Graham, B. S., Wyatt, R. T., Baker, D., Strong, R. K., Crowe, J. E., Jr., Johnson, P. R., and Schief, W. R. (2014) Nature **507**, 201-206.
4. Dantas, G., Corrent, C., Reichow, SL., Havranek, J.J., Eletr, Z.M., Isern, N.G., Kuhlman, B., Varani, G., Merritt, E.A., and Baker,D. (2007) *J. Mol. Biol.,* **366**, 1209-1221.
5. Dobson, N., Dantas, G., Baker, D., and Varani, G. (2006) *Structure*, **14,** 847-856.
6. Dunbrack, R.L. and Cohen, F.E. (1997) *Prot. Sci.,* **6,** 1661-1681.
7. Eyal, E., Najmanovich, R., McConkey, B.J., Edelman, M., and Sobelev, V. (2004) *J. Comput. Chem.,* **25,** 712-724.
8. Fleishman, S. J., Khare, S. D., Koga, N. and Baker, D. (2011) *Protein Science*, **20**, 753–757.
9. Gaines, J.C., Smith, W.W., Regan, L. and O'Hern, C.S. (2016) *Phys. Rev. E*, **93,** 032416.
10. Guerois, R., Nielsen, J.E., and Serrano, L. (2002) *J. Mol. Biol.*, **320,** 369-387.
11. Krivov, G.G., Shapovalov, M.M., and Dunbrack, R.L. (2009) *Proteins: Stuct. Funct. Bioinformatics*, **77**, 778-795.
12. Kuhlman, B., and Baker, D. (2000) *Proc. Natl. Acad. Sci. USA,* **97**, 10383-10388.
13. Leaver-Fay, A., O'Meara, M.J., Tyka, M., *et al.* (2013) *Methods Enzymol.,* **523**, 109-143.
14. Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Kristian R.J., Kaufman, W., Renfrew, P.D., Smith, C.A., Sheffler, W., *et al.* (2011) *Methods Enzymol.*, **487,** 545-574.
15. Liang, S., Zheng, D., Zhang, C., and Standley, D.M. (2011) *Bioinformatics,* **27,** 2913-2914.
16. Liu, H. and Chen, Q. (2016) *Curr. Opin. Struct. Biol.*, **39**, 89-95.
17. Miao, Z., Cao, Y., and Jiang, T. (2011) *Bioinformatics,* **27**, 3117-3122.
18. Peterson, L. X., Kang, X. and Kihara, D. (2014) *Proteins*, **82**, 1971–1984.
19. Rusling, J.F., Kumar, C.V., Gutkind, J.S., and Patel, V. (2010) *Analyst*, **135**, 2496-2511.
20. Sapsford, K.E., Bradburne, C., Delehanty, J.B., and Medintz, I.L. (2008) *Materials Today*, **11**, 38-49.
21. Shapovalov, M.M. and Dunbrack R.L., Jr. (2011) *Structure*, **19**, 844-58.

22. Tyka, M.D., Keedy, D.A., Andre, I., Dimaio, F., Richardson, J.S., Baker, D. (2011) *J Mol. Biol.,* **405**, 607–618.
23. Virrueta, A., O'Hern, C.S. and Regan, L. (2016) *PROTEINS,* **84,** 900-911.
24. Wang. G., and Dunbrack, R.L., Jr. (2003) *Bioinformatics*, **19**, 1589-1591.
25. Wang. G., and Dunbrack, R.L., Jr. (2005) *Nucleic Acids Res.,* **33,** W94-W98.
26. Word, J.M., Lovell, S.C., Richardson, J.S. and Richardson, D.C. (1999) *J. Mol. Biol.,* **285,** 1735-1747.
27. Zhou, A. Q., O'Hern, C. S. and Regan, L. (2012) *Biophys. J.,* **102,** 2345-2352.
28. Zhou, A.Q., O'Hern, C.S. and Regan, L. (2014) *PROTEINS,* **82**, 2574-2584.

Supplementary Material:

Calculation of error bars:

To assess the accuracy of the hard-sphere model in predicting the side chain dihedral angle conformations of residues in protein cores, repacking calculations were performed using N_v=300 bond length and angle variants for each core residue. We then randomly selected N instances of a given residue and M bond length and angle variants for each. For each variant, we identified the optimal side chain dihedral angle combination and calculated $\Delta\chi$, which yields a set of NxM $\Delta\chi$ values for each residue type. (See Fig. S1 (A), where N=M=50). We then calculated the mean fraction of residues F($\Delta\chi$'), which satisfy $\Delta\chi < \Delta\chi$'= 10°, 20°, or 30° (Fig. S1 (B)), and the standard deviation. We used N=50 and M=50 for single residue rotations and N=50 and M=30 for combined rotations.

For the Rosetta studies, one $\Delta\chi$ value was obtained for each of the five relaxed structures we considered for each core residue. The mean fraction F($\Delta\chi$') that satisfied $\Delta\chi < \Delta\chi$'= 10°, 20° or 30° and standard deviation shown in Figs. 7 and 8 were obtained by averaging over the five relaxed structures.

Rosetta Methods

For each protein, relaxation was performed using the command

```
relax.default.linuxgccrelease @relax.options -s this.pdb > relax.out
```

with the following options in the relax.options file:

```
-linmem_ig 100
-nstruct 50
-relax:fast
-relax:constrain_relax_to_start_coords
-scorefile relax.fasc
-score:weights talaris2013
```

From the 50 relaxed structures, the 5 structures with the lowest Rosetta energy were repacked using the command:

```
rosetta_scripts.default.linuxgccrelease @design_multi_relaxed.options -
parser:protocol design_multi_fixed.xml -out:suffix _design_relaxed  -
scorefile design_relaxed.fasc
```

with the following options in the design_multi_relaxed.options file:

```
-extrachi_cutoff 1
-linmem_ig 100 interactions
-nstruct 100
```

```
-s this_best_1.pdb
```

and design_multi_fixed.xml file:

```
<ROSETTASCRIPTS>
        <SCOREFXNS>
        </SCOREFXNS>
        <TASKOPERATIONS>
                Include rotamer options from the command line
                <InitializeFromCommandline name=ifcl />
                Design and repack residues based on resfile
                <ReadResfile name=rrf filename=this.resfile/>
        </TASKOPERATIONS>
        <MOVERS>
                Design the antibody interface
                <PackRotamersMover name=design scorefxn=talaris2013
task_operations=ifcl,rrf />
        </MOVERS>
        <FILTERS>
        </FILTERS>
        <APPLY_TO_POSE>
        </APPLY_TO_POSE>
        <PROTOCOLS>
                Run the design protocol
                <Add mover=design />
        </PROTOCOLS>
        <OUTPUT scorefxn=talaris2013 />
</ROSETTASCRIPTS>
```

this.resfile contains the core residues that are repacked for a given interacting cluster. In resfile, we specified extra rotamer sampling (*e.g.* using the flag EX 1 LEVEL 7).

A

	Δχ(1)	Δχ(2)	Δχ(3)	Δχ(4)	Δχ(5)	Δχ(6)	Δχ(7)	Δχ(8)	…	Δχ(50)
Val_1	8.7	5.7	5.4	5.4	8.7	8.7	5.4	5.4	…	5.4
Val_2	13.4	10.0	10.0	13.4	10.0	13.4	10.0	10.0	…	13.4
Val_3	9.1	12.2	16.0	13.7	9.1	13.7	16.0	9.6	…	118.2
Val_4	6.1	1.1	8.0	8.0	8.0	6.1	6.1	8.0	…	8.0
Val_5	2.7	7.7	2.7	2.7	2.7	2.7	10.9	6.1	…	2.7
Val_6	4.8	4.8	4.8	4.8	4.8	4.8	4.8	15.9	…	4.8
Val_7	5.8	12.2	12.2	8.1	7.2	5.8	12.2	12.2	…	5.8
Val_8	0.5	0.5	0.5	0.5	4.6	0.5	0.5	7.0	…	0.5
Val_9	12.2	22.8	12.2	22.8	23.3	12.2	22.8	23.3	…	12.2
Val_{10}	1.5	26.6	1.5	26.6	9.7	1.5	26.6	20.8	…	1.5

B

F_{10}	80	50	60	60	80	70	40	50	…	70
F_{20}	100	80	100	80	90	100	80	80	…	90
F_{30}	100	100	100	100	100	100	100	100	…	100

C

$F_{10} = 62.2 \pm 13.9$
$F_{20} = 88.9 \pm 9.3$
$F_{30} = 100 \pm 0$

Figure S1: Description of the calculation of error bars for the fraction of residues F(Δχ') with deviation Δχ<Δχ'. (A) Δχ for Val residues with randomly assigned bond length and angle variants. (B) Fraction of residues F(10°) , F(20°), and F(30°)for the data in (A). (C) Average and standard deviations for F(10°) , F(20°), and F(30°) in (B).

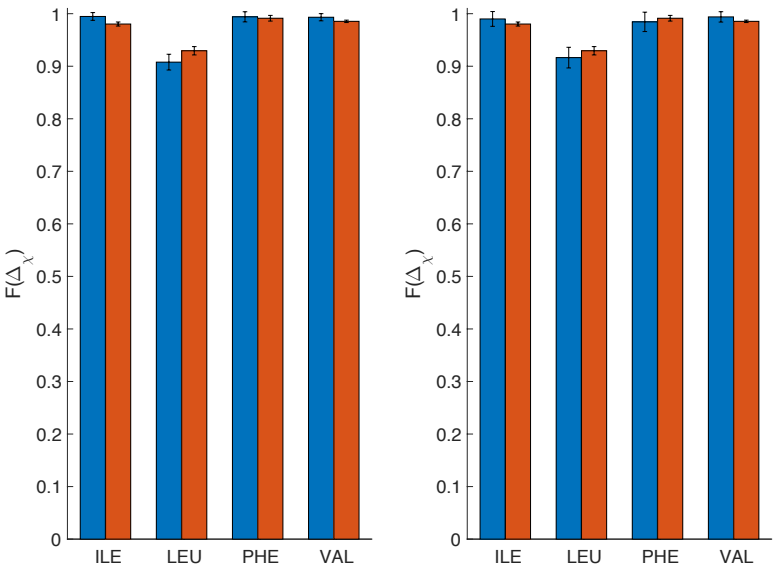

Figure S2: Comparison of the results for single residue (left) and collective (right) repacking of protein cores using the hard-sphere model for the Dunbrack 1.0Å (blue) and HiQ54 (red) protein crystal databases. Each bar shows the fraction of residues with Δχ < 30°.