Random close packing in protein cores

Corey S. O'Hern
Department of Mechanical Engineering & Materials Science
Department of Applied Physics
Department of Physics
Graduate Program in Computational Biology & Bioinformatics
Yale University, New Haven, CT USA

Lynne Regan,
MB&B,
Chemistry

Jennifer Gaines,
CBB

Diego Caballero
Physics

Wendell Smith
Physics

Alex Virrueta
MEMS

The Raymond and Beverly Sackler Institute
for Biological, Physical and Engineering Sciences
Yale University

NIH NSF

Recent Publications

1. A. T. Grigas, Z. Liu, L. Regan, and C. S. O'Hern, "Core packing of well-defined x-ray and NMR structures is the same," submitted to Protein Science (2022).
2. Z. Mei, J. D. Treado, A. Grigas, Z. A. Levine, L. Regan, and C. S. O'Hern, "Analyses of protein cores reveal fundamental differences between solution and crystal structures," *Proteins: Structure, Function, and Bioinformatics 88* (2020) 1154.
3. A. Grigas, Z. Mei, J. D. Treado, Z. A. Levine, L. Regan, and C. S. O'Hern, "Using physical features of protein core packing to distinguish real proteins from decoys," *Protein Science 29* (2020) 1931.
4. J. D. Treado, Z. Mei, L. Regan, and C. S. O'Hern, "Void distributions reveal structural link between jammed packings and protein cores," *Phys. Rev. E* 99 (2019) 022416.
5. J. C. Gaines S. Acebes, A. Virrueta, M. Butler, L. Regan, and C. S. O'Hern, "Comparing side chain repacking in soluble proteins, protein-protein interfaces, and transmembrane proteins," *Proteins: Structure, Function, and Bioinformatics* 86 (2018) 581.
6. J. C. Gaines, A. Virrueta, D. A. Buch, S. J. Fleishman, C. S. O'Hern, and L. Regan,"Collective repacking reveals that the structures of protein cores are uniquely specified by steric repulsive interactions," *Protein Eng. Des. Sel.* 30 (2017) 387.
7. J. C. Gaines, A. H. Clark, L. Regan, and C. S. O'Hern, "Packing of protein cores," *J. Phys.: Condens. Matter* 29 (2017) 293001.
8. D. Caballero, A. Virrueta, C. S. O'Hern, and L. Regan, "Steric interactions alone determine side chain conformations in protein cores," *Protein Eng. Des. Sel.* 29 (2016) 367.
9. J. C. Gaines, W. W. Smith, L. Regan, and C. S. O'Hern, "Random close packing in protein cores," *Phys. Rev. E* 93 (2016) 032415.
10. A. Virrueta, C. S. O'Hern, and L. Regan, "Understanding the physical basis for the side chain conformations of Met," *Proteins: Structure, Function, and Bioinformatics* 84 (2016) 900.
11. D. Caballero, W. W. Smith, C. S. O'Hern, and L. Regan, "Equilibrium transitions between side chain conformations in leucine and isoleucine," *Proteins, Structure, Function, and Bioinformatics* 83 (2015) 1488.

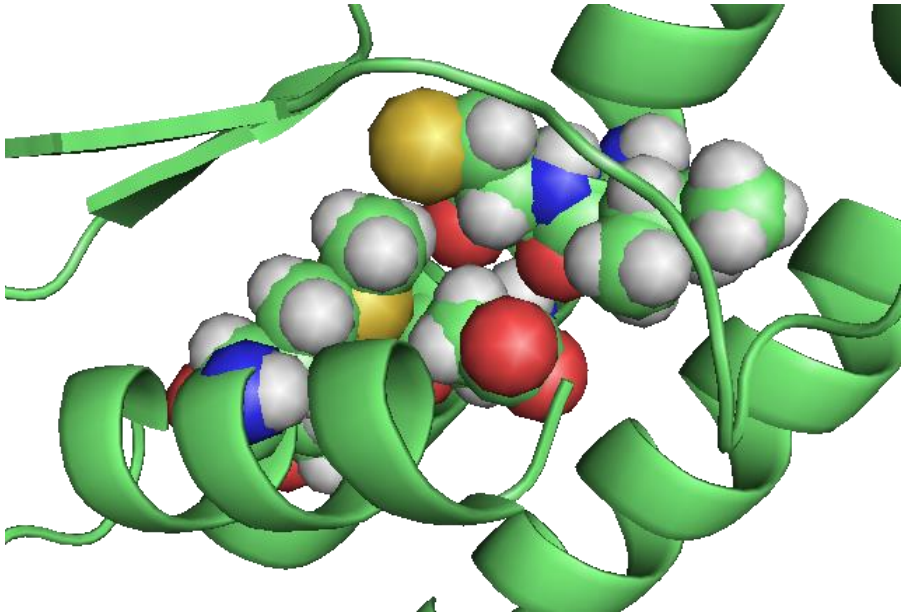
Important questions for protein packing

1. Can the structural properties of protein cores be quantitatively modeled using hard-spheres?

2. What is the packing fraction in protein cores?

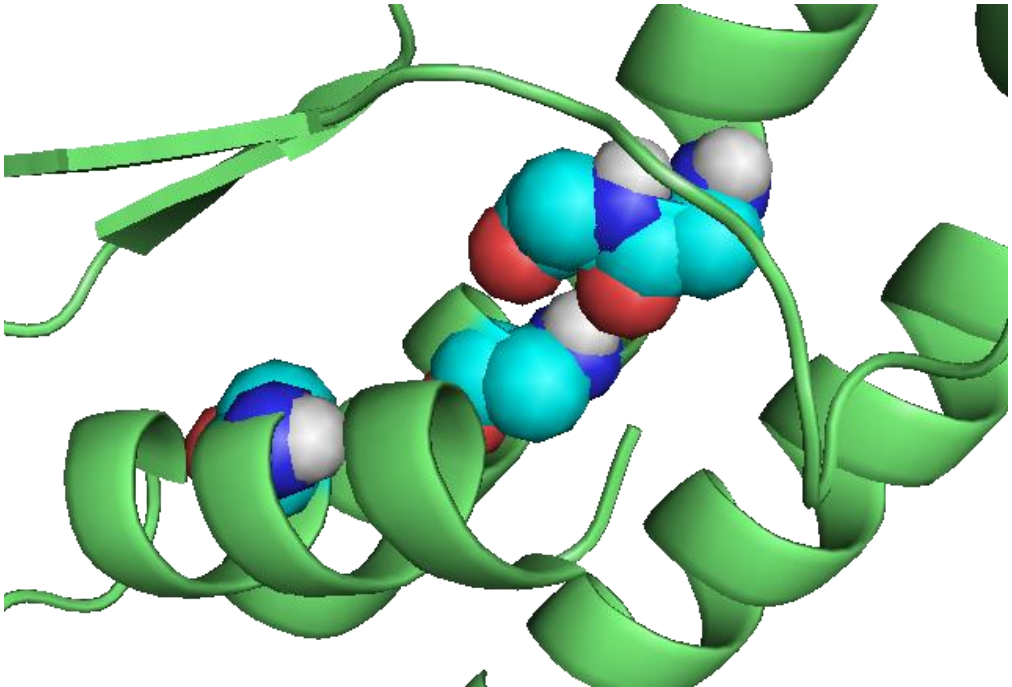
3. Can simple hard-sphere model improve computational design of protein-protein interactions?

Protein Re-packing



1G4I; with sidechains

Met, Tyr, Asp, Ile, Cys, Asp, Ala



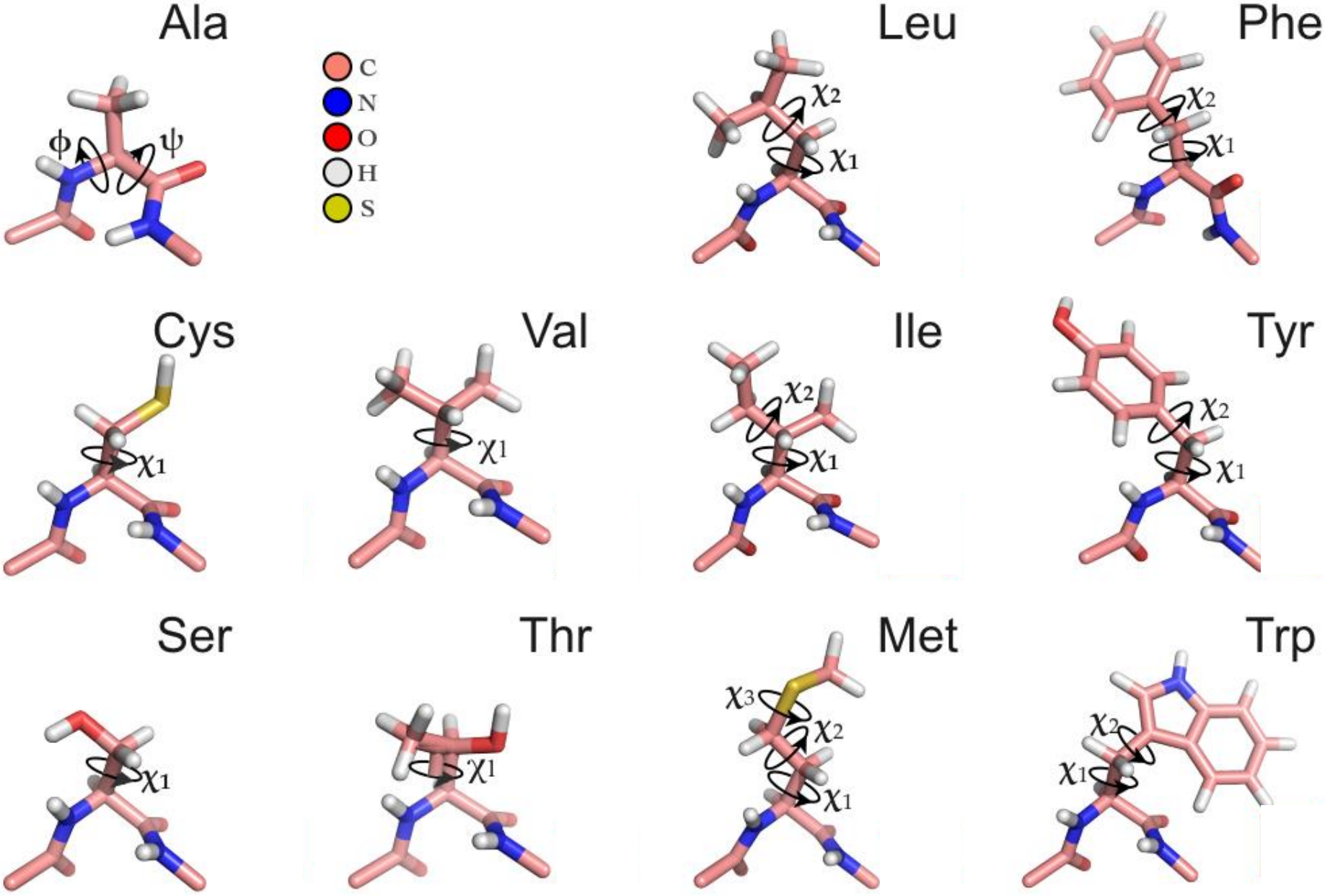
1G4I; without sidechains

What is a good model for the packing inside a protein? Is a protein packed more like a liquid or a solid? Based on the observations of high packing densities (Richards, 1977) and low compressibilities (Gavish et al., 1983), protein cores are often considered to be more like solids than liquids (Chothia, 1984; Murphy and Gill, 1991). Packing in proteins was first analyzed quantitatively by Richards (Richards, 1974) and Finney (Finney, 1975). They used a Voronoi analysis for proteins in a space-filling model, where each atom is taken to be a sphere with a fixed radius, given by the van der Waals radius. These and other classic papers showed that the average packing density in a protein is as high as that inside crystalline solids (Chothia, 1975; Harpaz et al., 1994; Gerstein and Chothia, 1996).

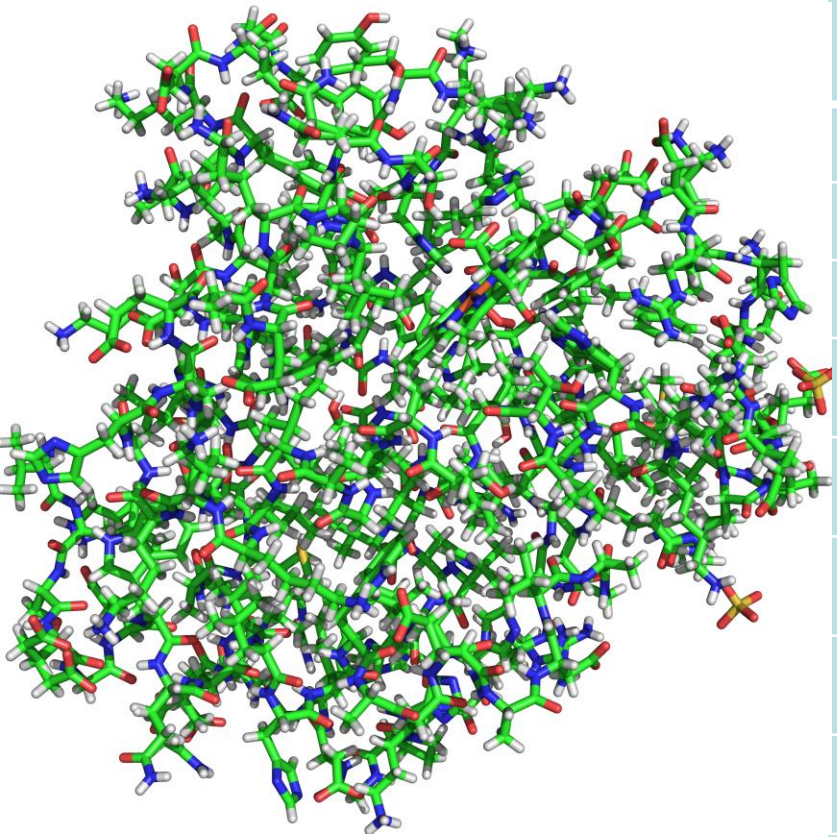
J. Liang and K. A. Dill, "Are proteins well-packed," *Biophys. J.* 81 (2001) 751.

What is a hard-sphere model for protein structure? How does one choose the atom sizes? Does one need explicit hydrogens?

Why does the packing fraction of protein cores have to satisfy $\phi<0.74$?


Ala
ϕ
ψ
C
N
O
H
S
Leu
χ2
χ1
Phe
χ2
χ1
Cys
χ1
Val
χ1
Ile
χ2
χ1
Tyr
χ2
χ1
Ser
χ1
Thr
χ1
Met
χ3
χ2
χ1
Trp
χ2
χ1

"Dunbrack" database of high-resolution protein crystal structures



PDB:1A6M Oxy-Myoglobin

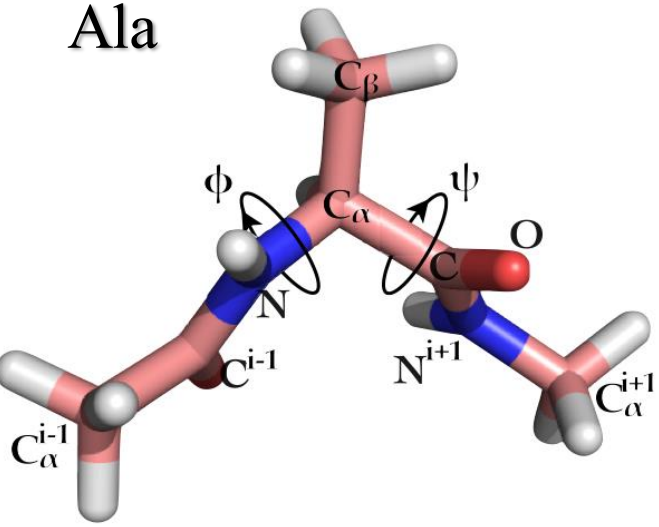
Residue Type	Number
Ala	3471
Gly	3457
Leu	2958
Val	2744
Ser	2624
Thr	2542
Asp	2427
Lys	2190
Glu	2158

Residue Type	Number
Ile	2068
Asn	1884
Pro	1776
Arg	1653
Phe	1451
Gln	1436
Tyr	1409
His	853
Met	695
Cys	623
Trp	597

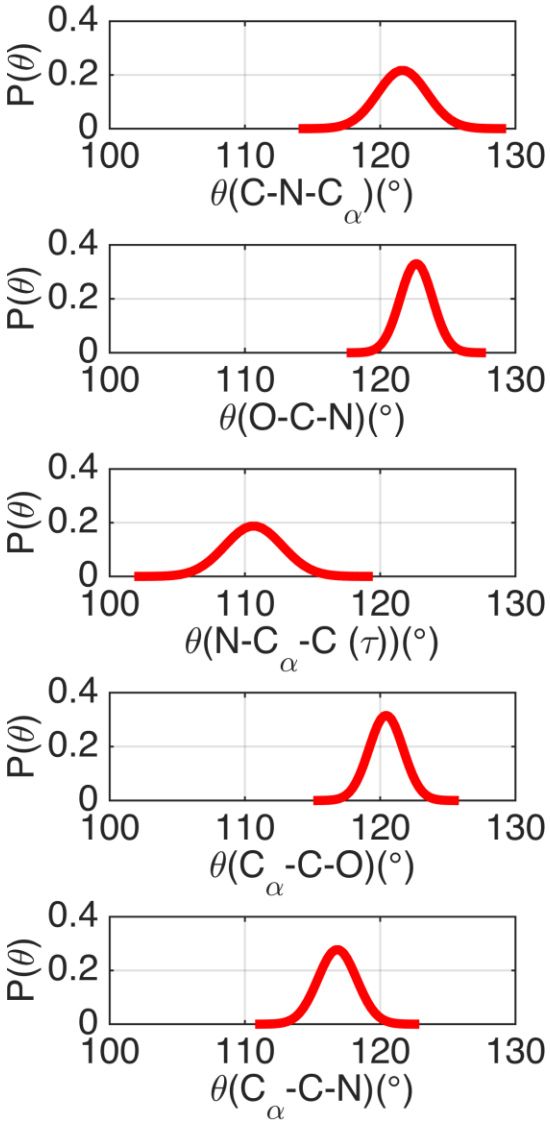
792 structures; resolution ≤ 1.0Å; sequence identity $< 50\%$; B-factor per residue ≤ 30Å^2

G. Wang and R. L. Dunbrack, Bioinformatics 19 (2003) 1589.

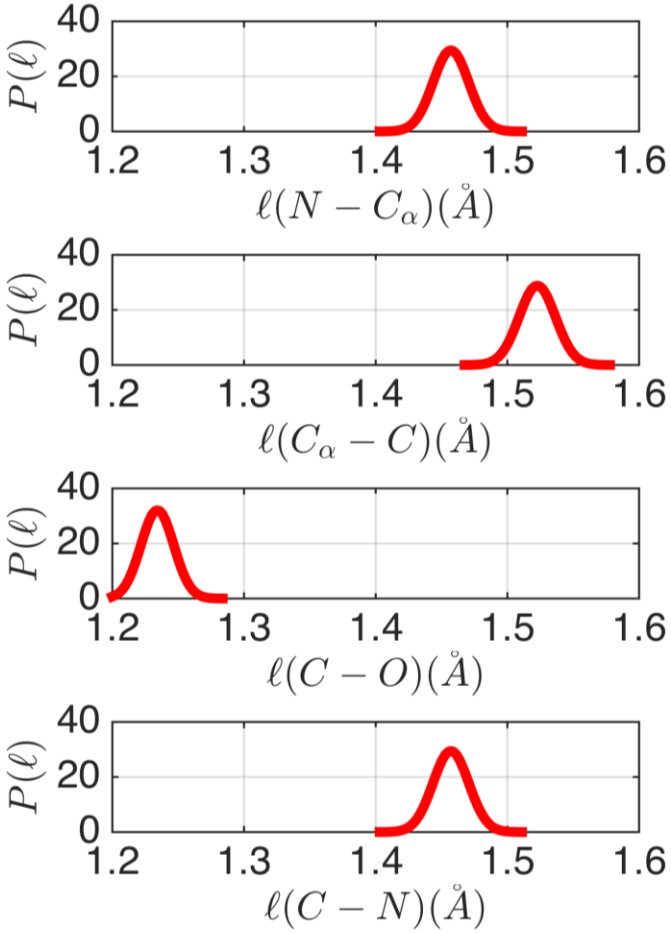
Ala



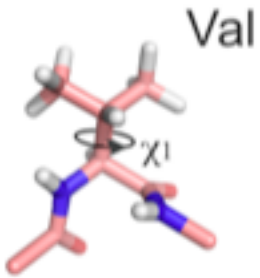
Bond angles, θ

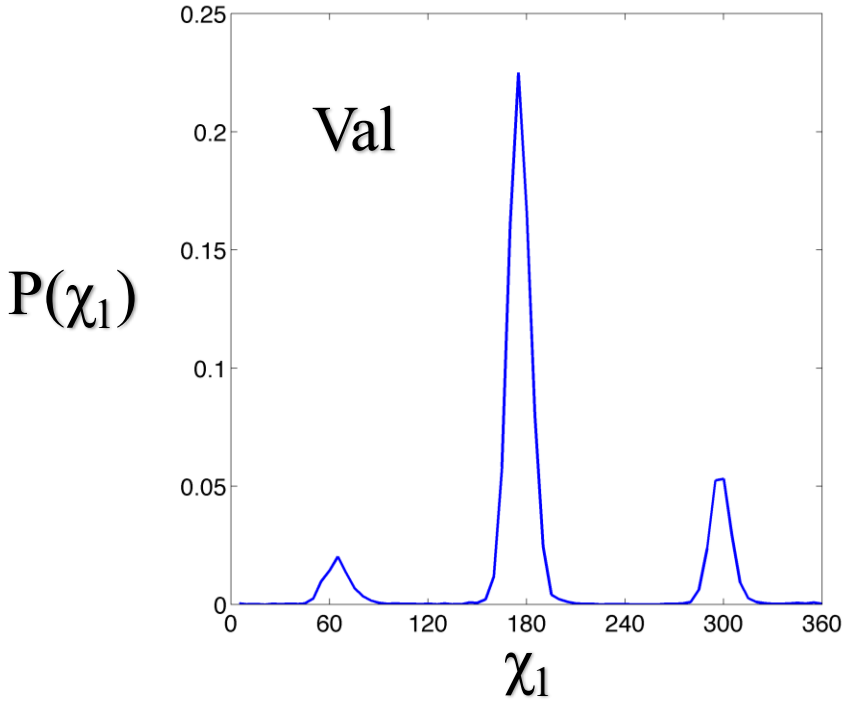


Bond lengths, *l*

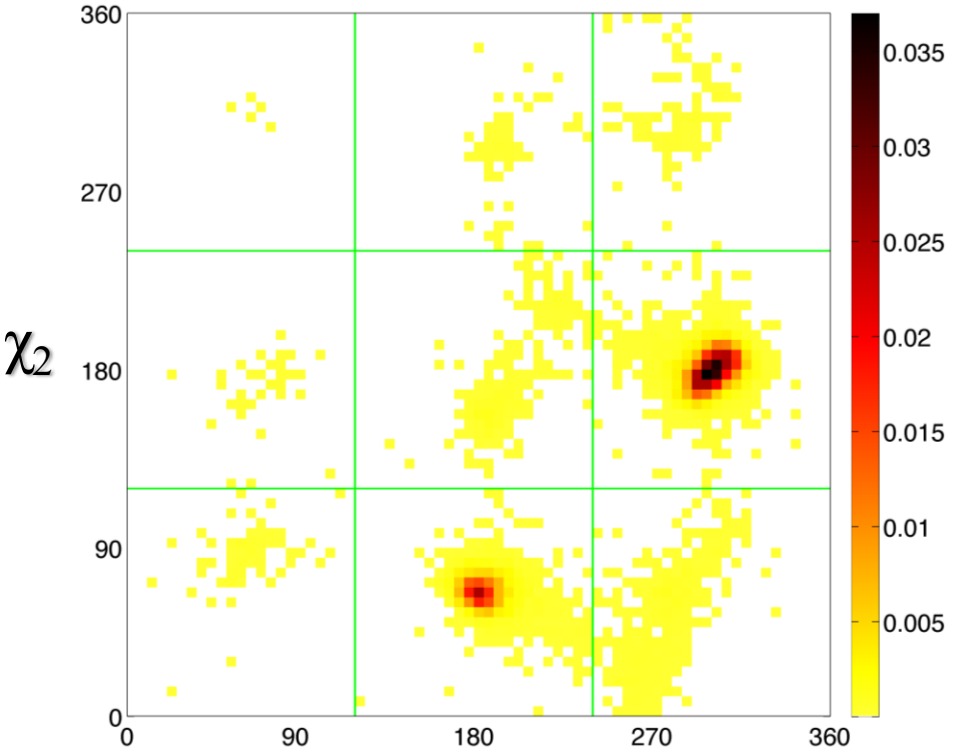


Observed Side-chain Dihedral Angle Distributions

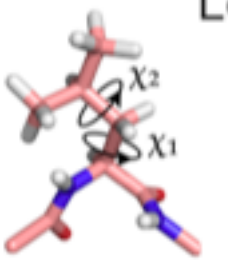




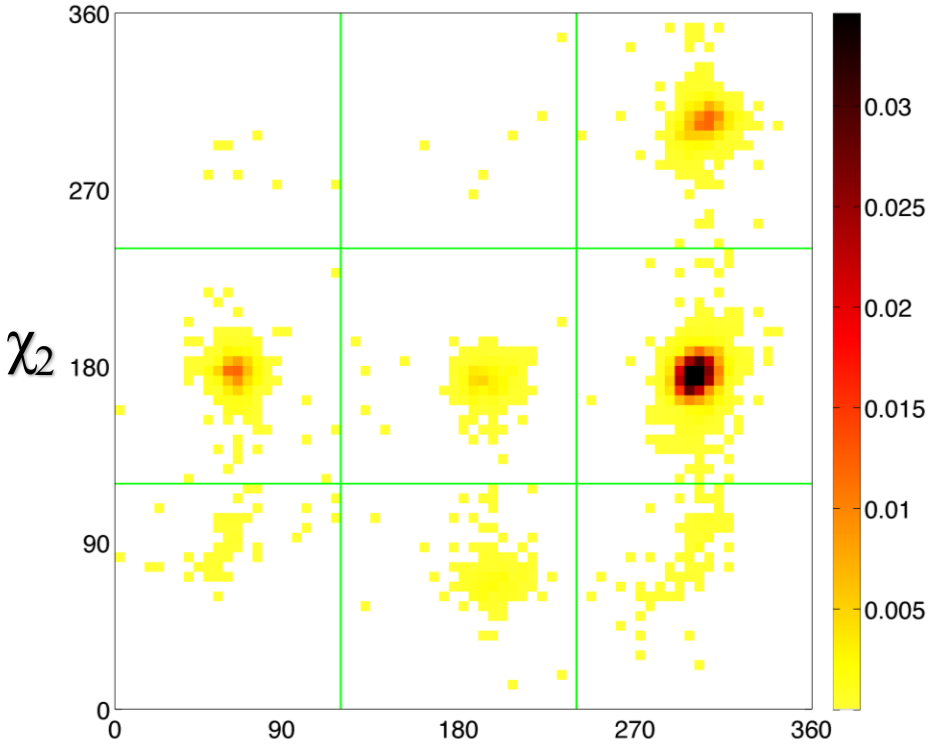
Leu



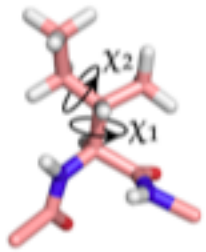
Leu



Ile



Ile



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.

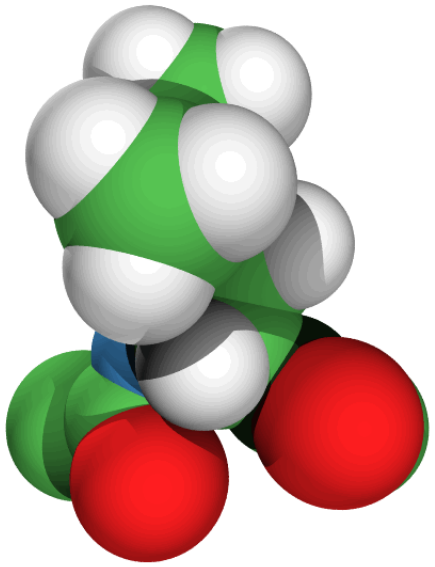
Calculate Distribution of Side-chain Dihedral Angles

$$V_{RLJ}\left(r_{ij}\right)=\begin{cases}\left[\left(\dfrac{s_{ij}}{r_{ij}}\right)^{6}-1\right]^{2} & r_{ij}\le s_{ij}\\ 0 & r_{ij}>s_{ij}\end{cases}$$

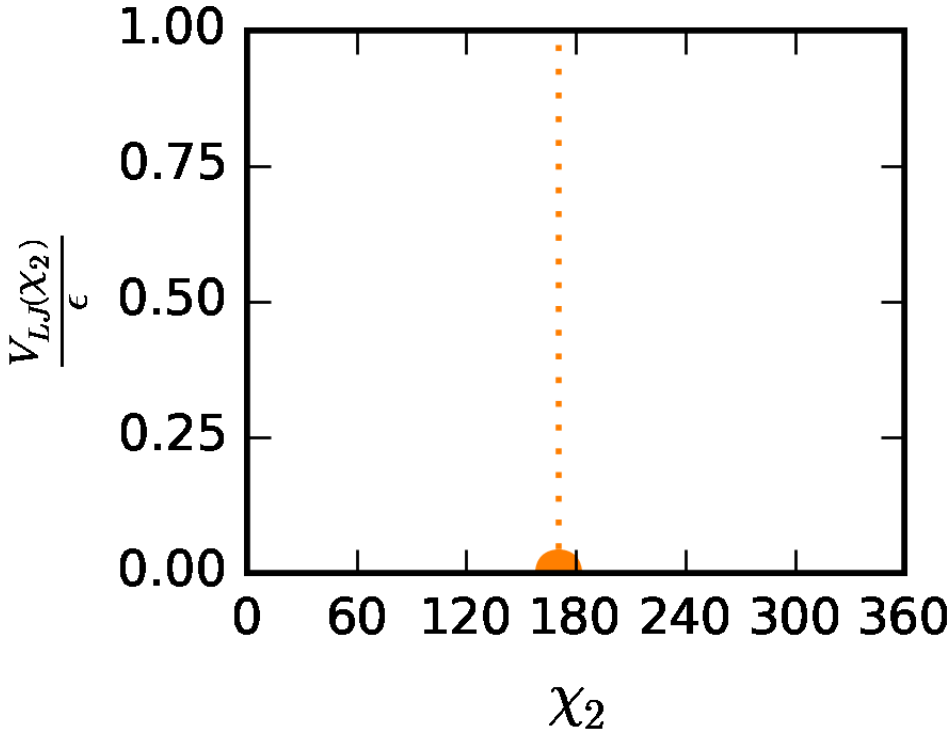
$$\sigma_{ij}=(\sigma_i+\sigma_j)/2$$

$$V_{RLJ}^{Tot}=\frac{1}{N}\sum_{i>j}V_{RLJ}\left(r_{ij}\right)$$

$$P(c_1)\propto\left\langle\exp\left[-\frac{V_{RLJ}^{Tot}(c_1)}{k_BT}\right]\right\rangle \qquad \frac{k_bT}{e}\to 0$$



Leu

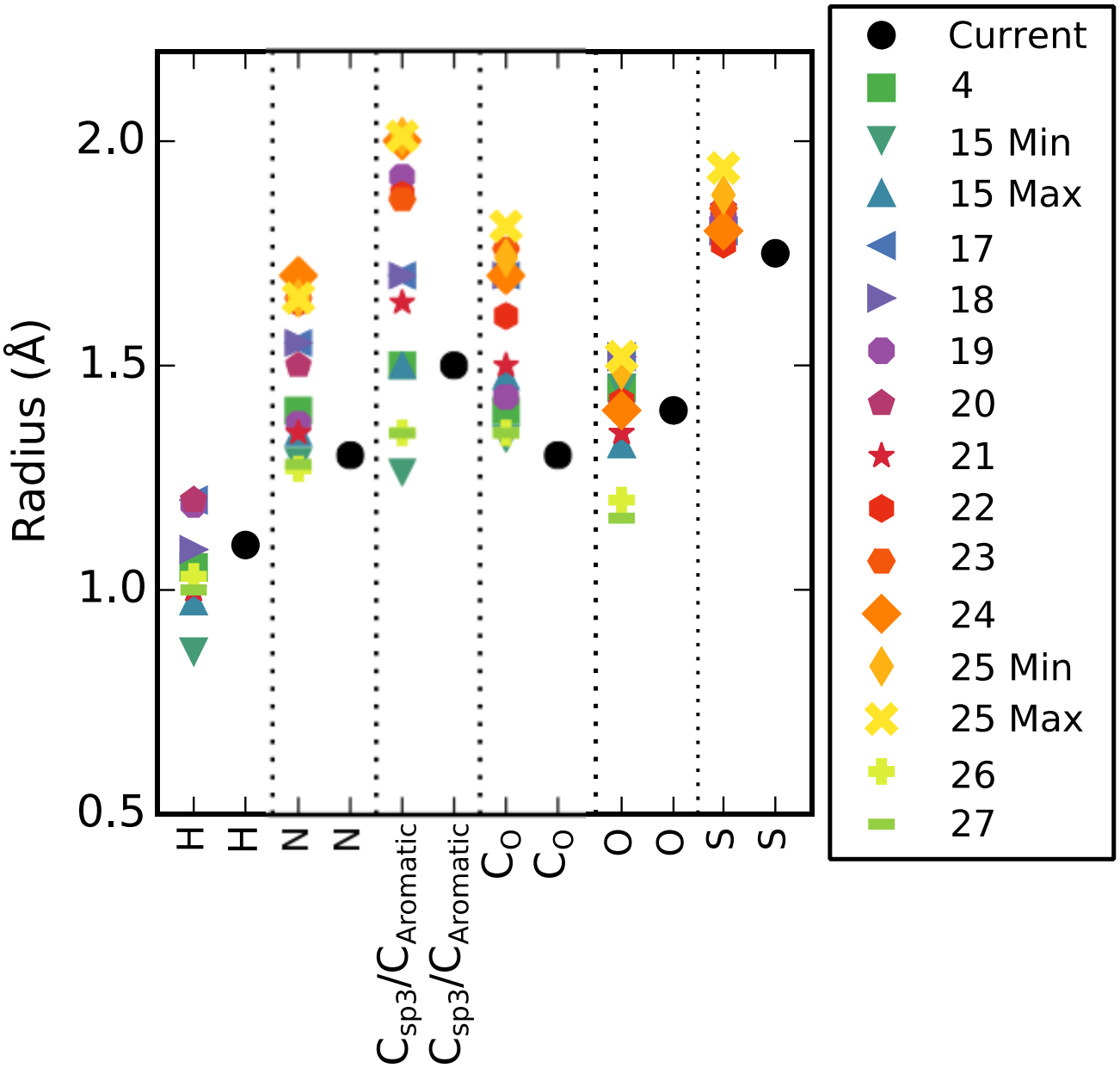

1.00
0.75
0.50
0.25
0.00
$\frac{V_{LJ}(\chi_2)}{\epsilon}$
0
60
120
180
240
300
360
χ_2

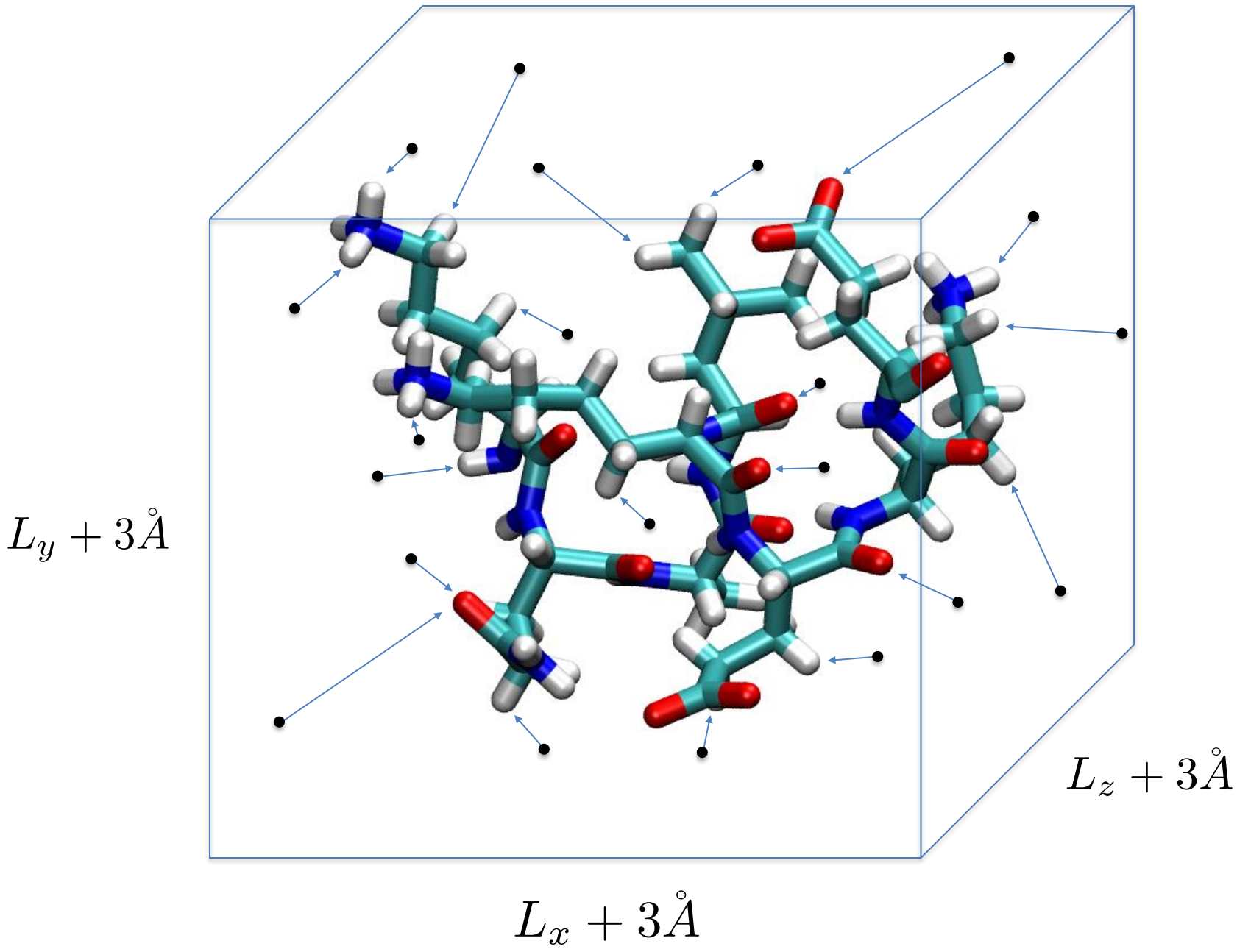
O

C

H

N


Radius (Å)
2.0
1.5
1.0
0.5
H
H
N
N
$C_{sp3}/C_{Aromatic}$
$C_{sp3}/C_{Aromatic}$
C_O
C_O
O
O
S
S
Current
4
15 Min
15 Max
17
18
19
20
21
22
23
24
25 Min
25 Max
26
27


$L_y + 3\mathring{A}$
$L_z + 3\mathring{A}$
$L_x + 3\mathring{A}$

Core Residues

Residue Type	% in Core	% on Surface
Leu	17.5	7.5
Ala	13.2	7.2
Val	12.7	6.2
Ile	12.2	4.6
Gly	9.9	7.9
Phe	8.0	3.7
Cys	5.1	1.5
Thr	3.9	6.4
Ser	3.7	6.4

Residue Type	% in Core	% on Surface
Met	3.2	1.1
Tyr	2.2	4.0
Pro	1.9	4.5
Trp	1.4	1.7
Asn	1.4	5.3
Asp	1.0	6.7
Gln	0.7	4.0
His	0.7	2.1
Glu	0.5	6.7
Lys	0.5	7.2
Arg	0.3	5.2

Intra-residue contacts

Inter-residue contacts

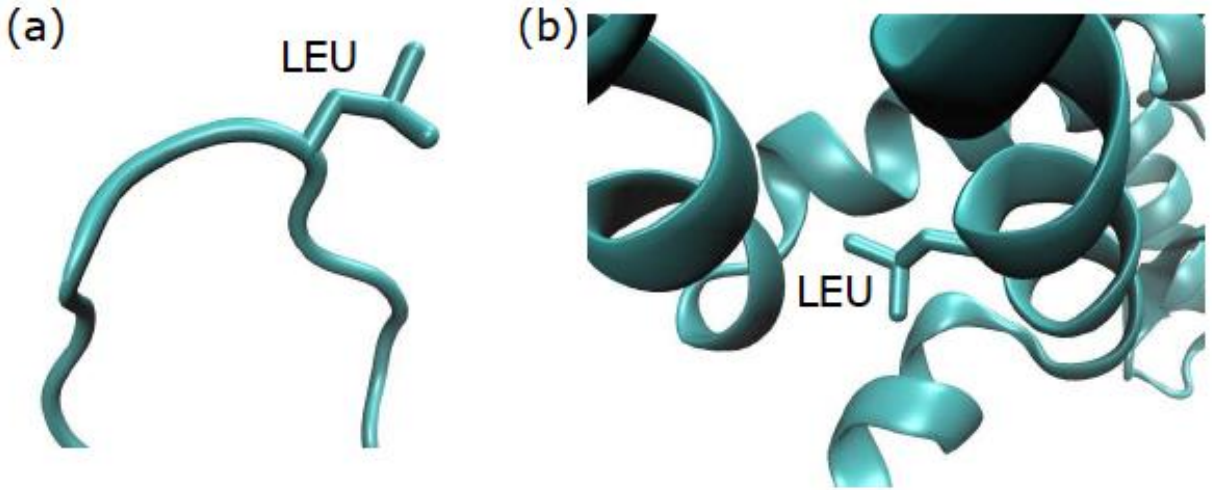
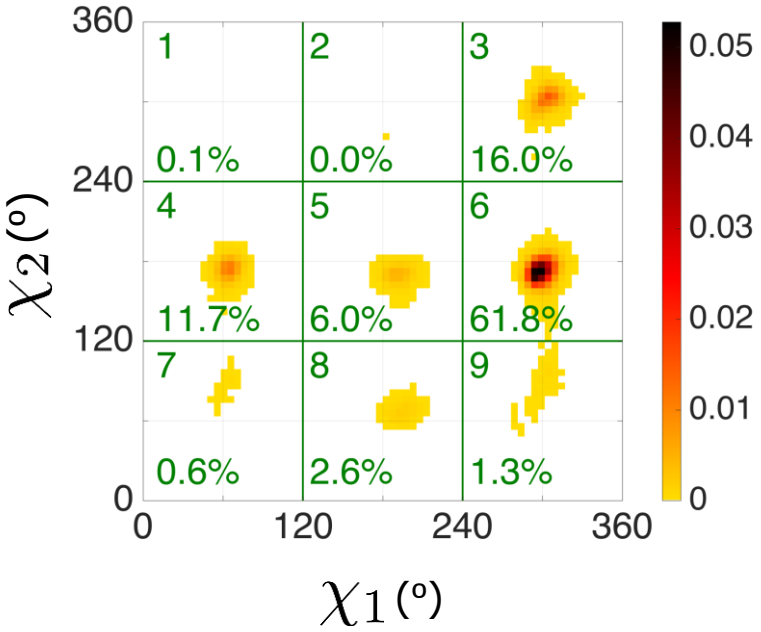
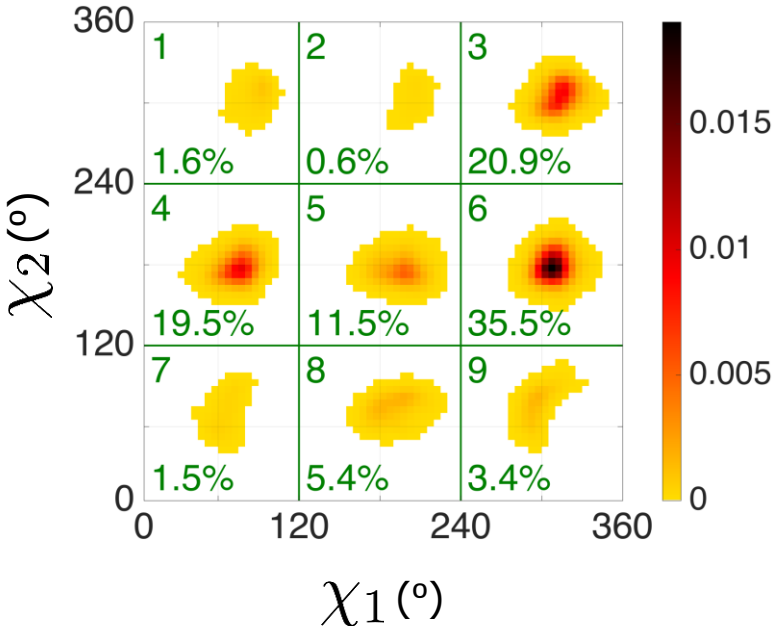


Figure 1: (a) X-ray crystal structure in ribbon representation of a coil region of the protein T4 Lysozyme (PDB ID: 1L63) with Leu 15 highlighted. (b) Close-up of Leu 99, which occurs on an α-helical segment in the core of T4 Lysozyme.

Ile in proteins

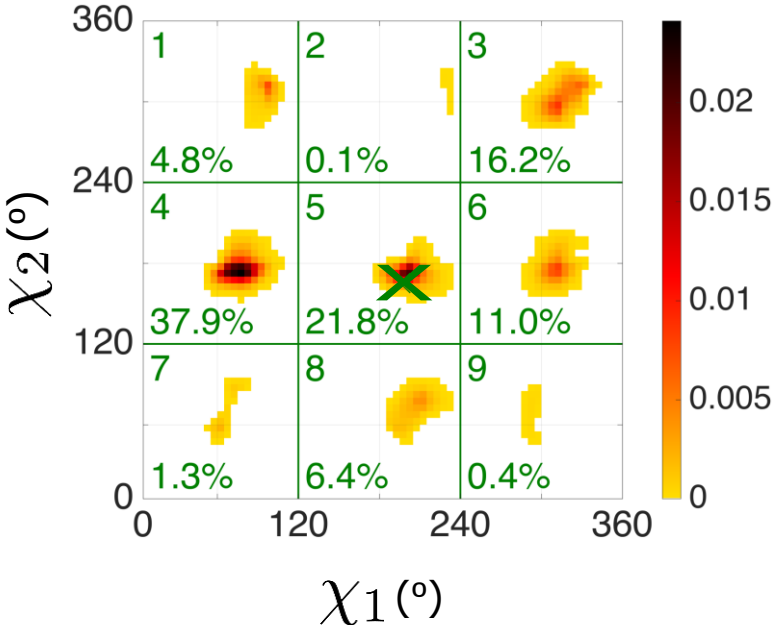


Ile hard-sphere dipeptide prediction

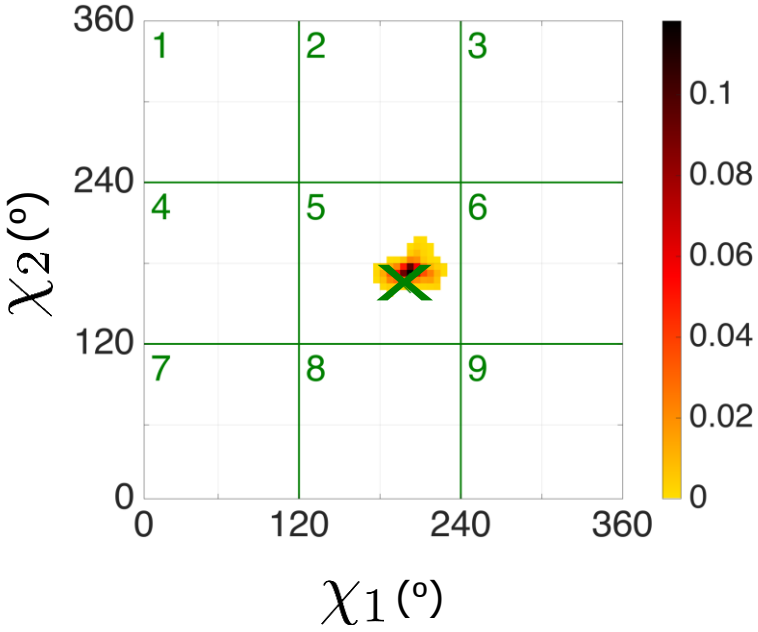


Ile56 in PDB 2NWD

Hard-sphere dipeptide



Hard-sphere protein environment



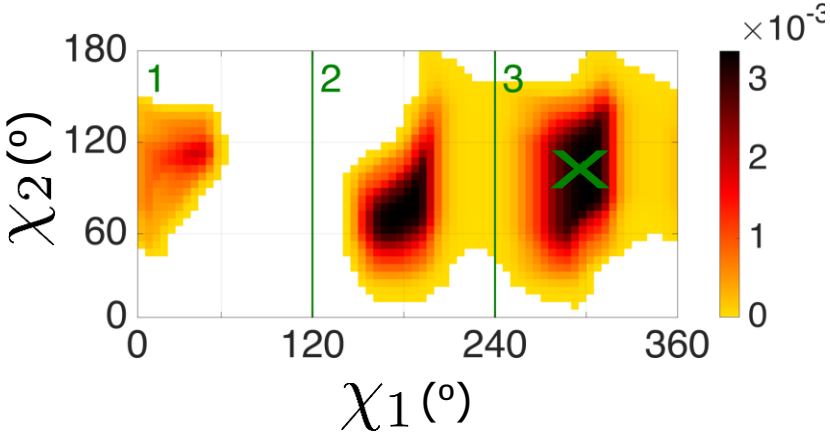
Hard-sphere dipeptide

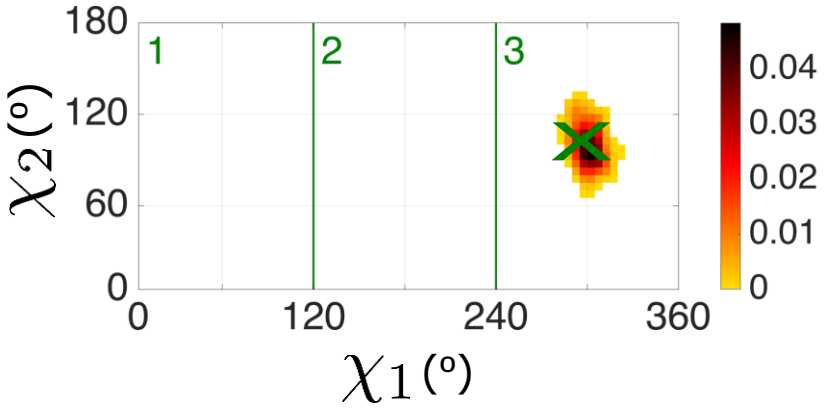
Hard-sphere protein environment

(a)

(b)

Phe

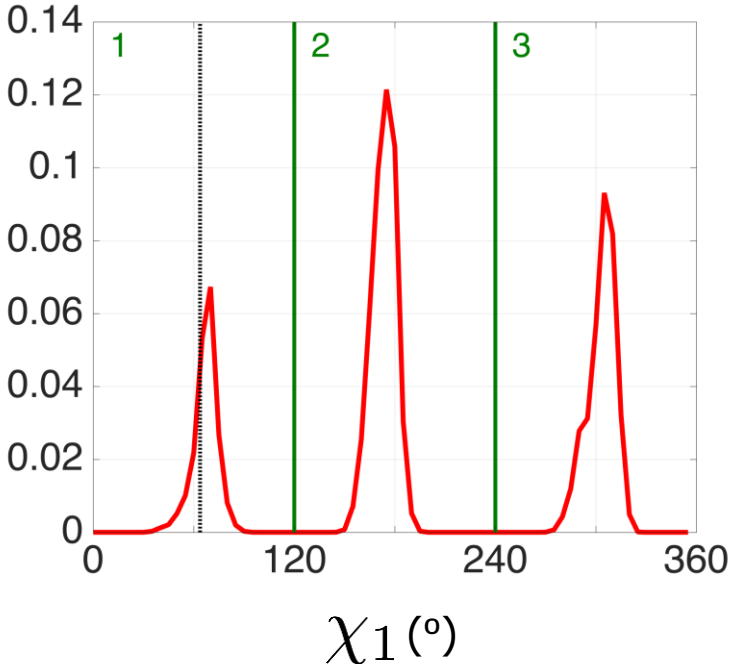


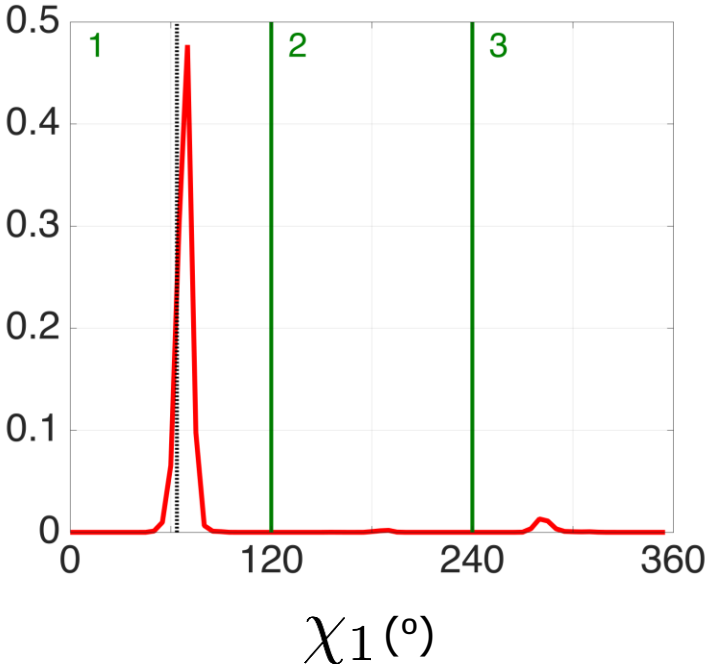


(c)

(d)

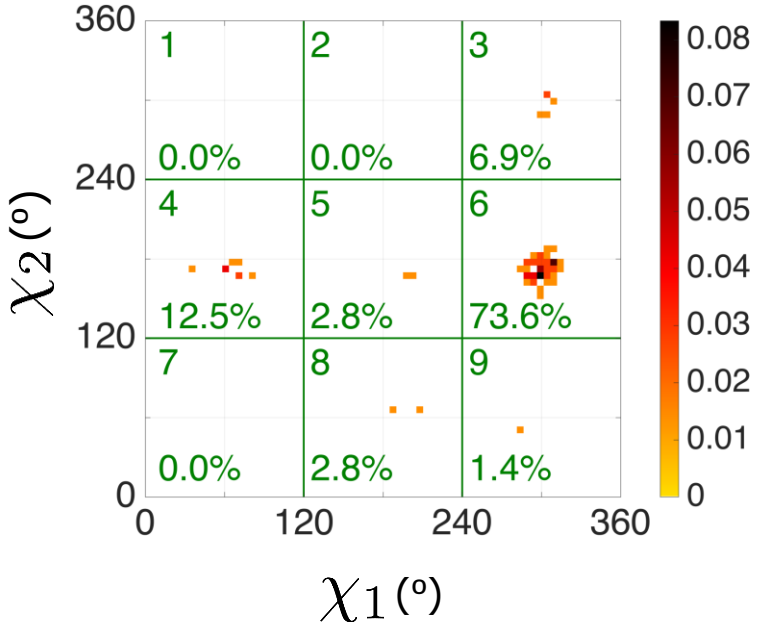
Val



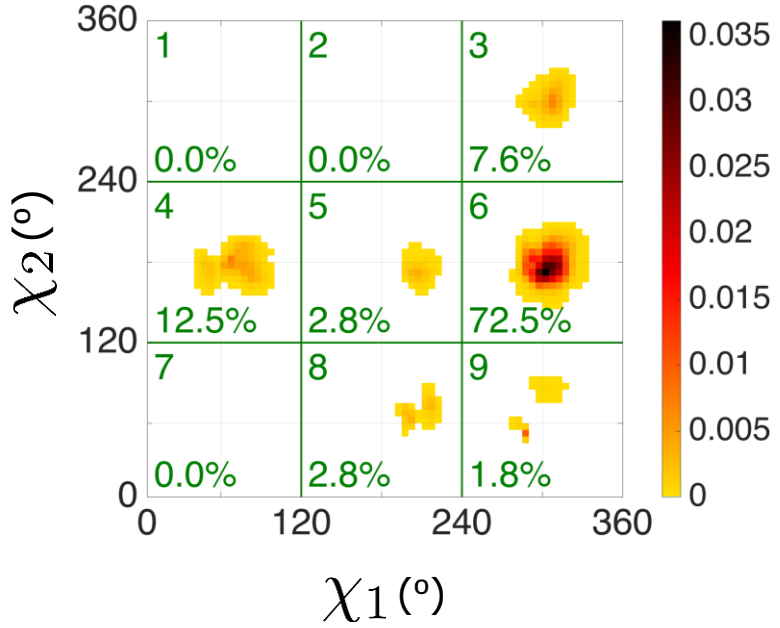


Ile in Protein Environment

(a) PDB-HiQ54



(b) Hard-sphere protein environment



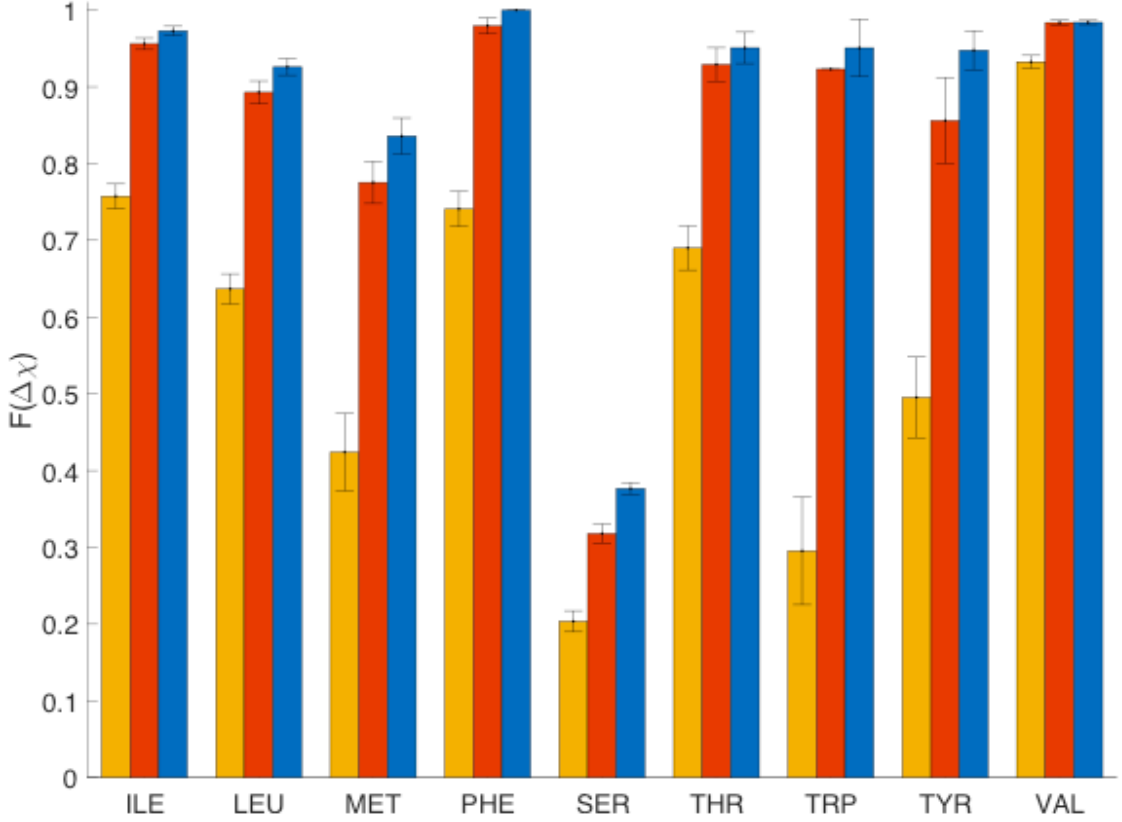


Figure 4: Single residue rotations in the context of the protein core: The fraction (F(Δχ)) of each residue type for which the hard-sphere model prediction of the side chain conformation is Δχ < 10° (yellow), 20° (red), or 30° (blue) from the crystal structure for core residues in the Dunbrack 1.0Å database.

$$Dc = \sqrt{\left(c_1^{xtal} - c_1^{HS}\right)^2 + \ldots + \left(c_n^{xtal} - c_n^{HS}\right)^2}$$

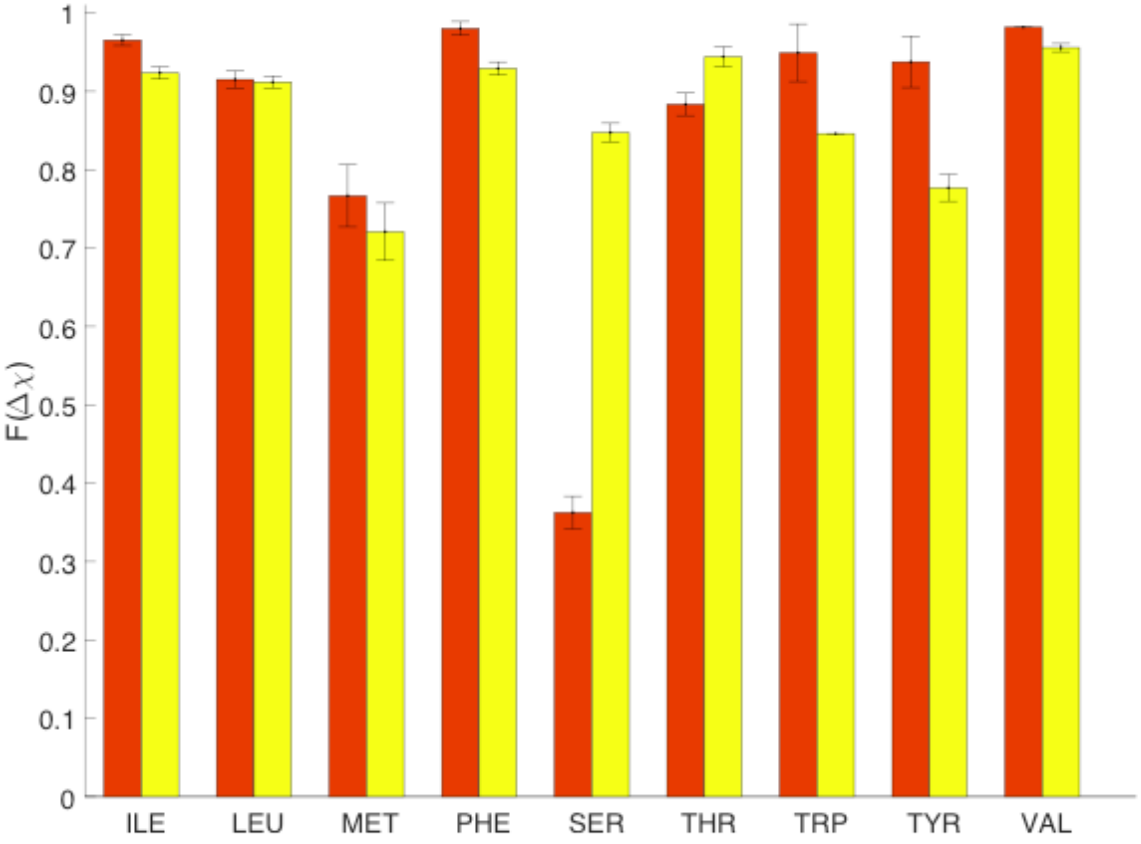


Figure 7: Comparison of the accuracy of combined rotations for core residues in the Dunbrack 1.0Å database using the hard-sphere model (red) and Rosetta (yellow). Each bar shows the fraction F(Δχ) of residues for which the model prediction was $\Delta\chi < 30°$.

Studies of Packing in Protein Cores

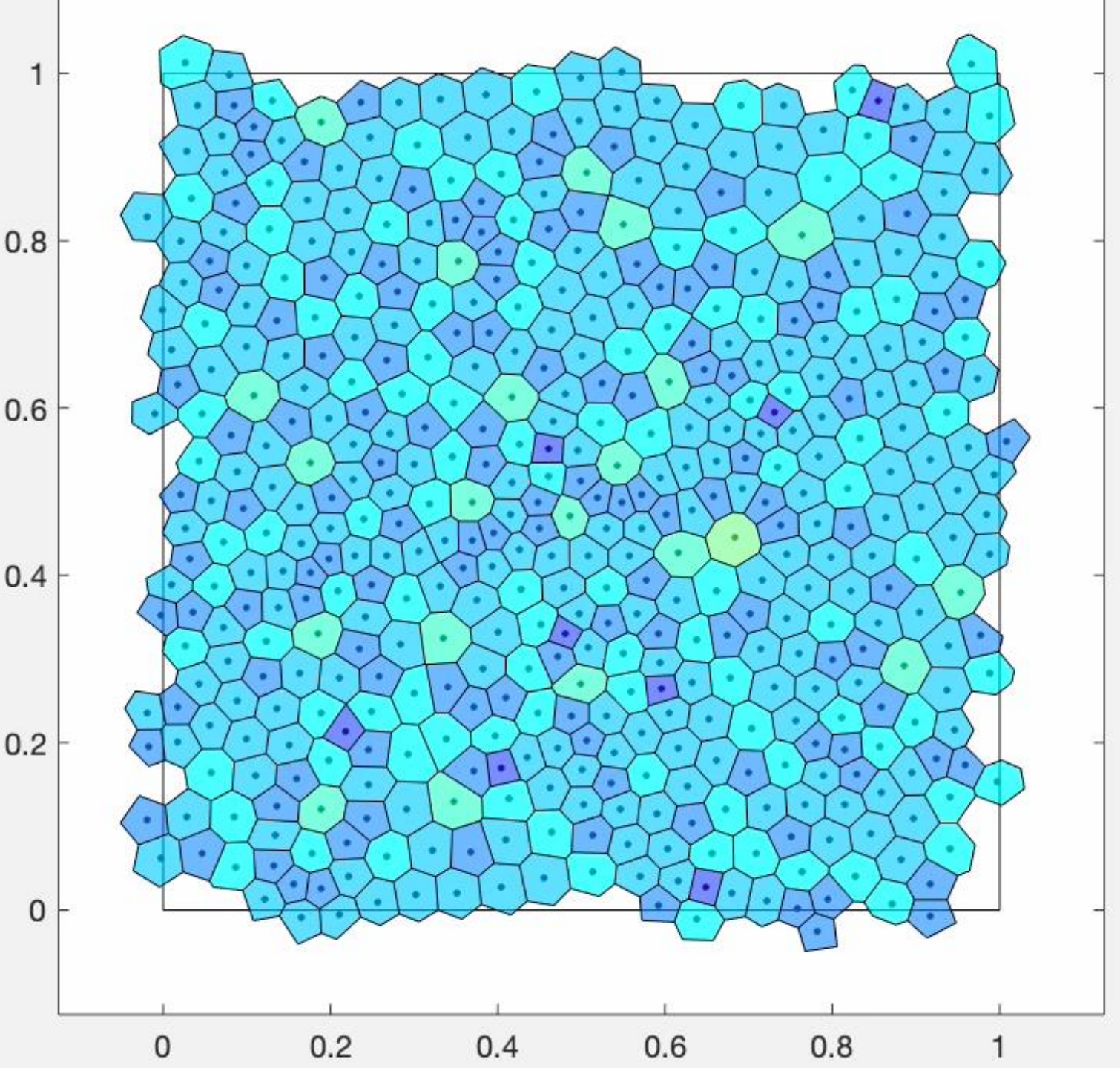
F. M. Richards, "The interpretation of protein structures: Total volume, group volume distributions, and packing fraction," *J. Mol. Biol.* 82 (1974) 1.

J. Liang and K. A. Dill, "Are proteins well-packed," *Biophys. J.* 81 (2001) 751.

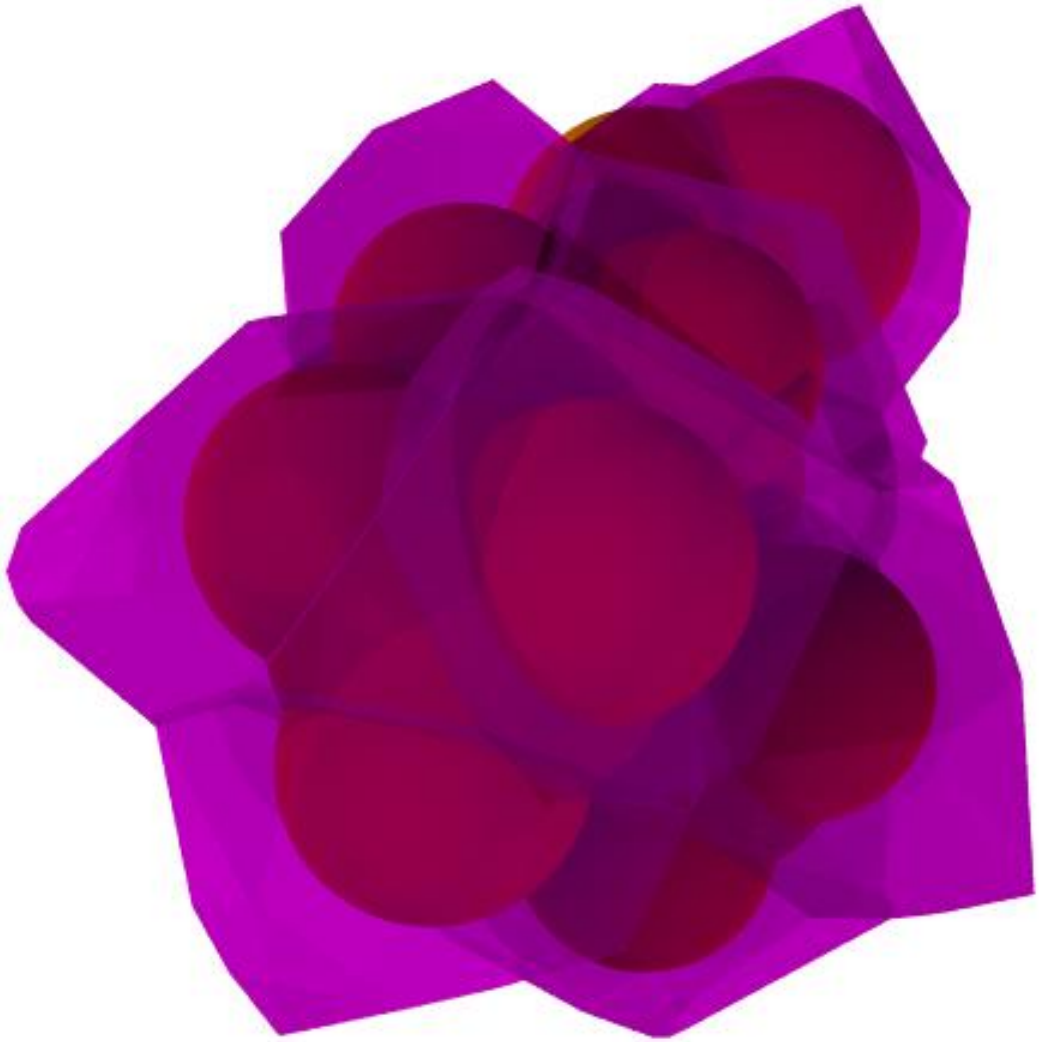
$$f_r = \frac{V_r}{V_{container}} \qquad f_r = \frac{V_r}{\sum_{i=1}^{N_a} v_{ri}}$$

v_{ri}=Voronoi volume of atom i in residue r

$$f_{core} = \frac{1}{N_r} \sum_{i=1}^{N_r} f_{ri}$$

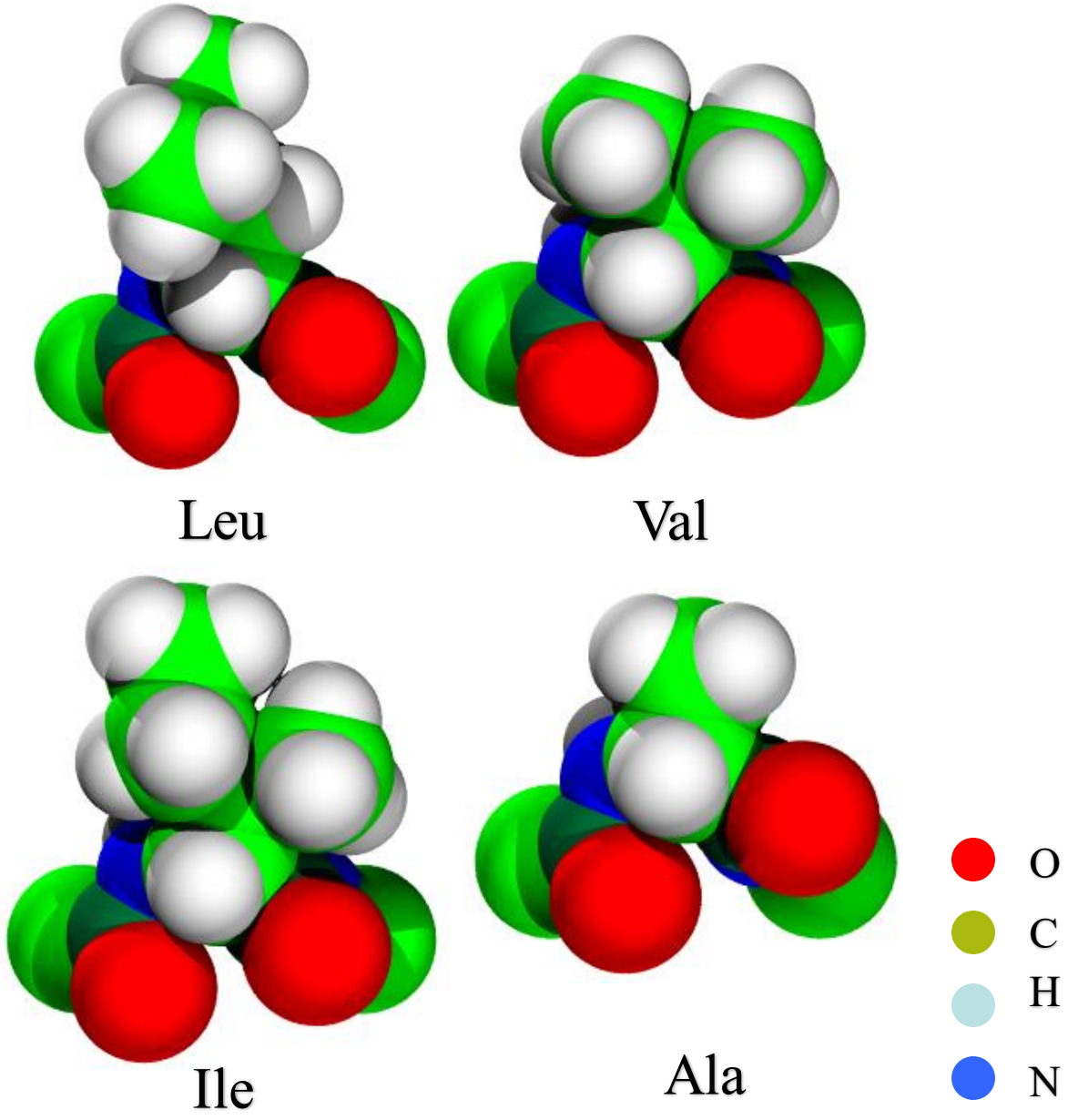

1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

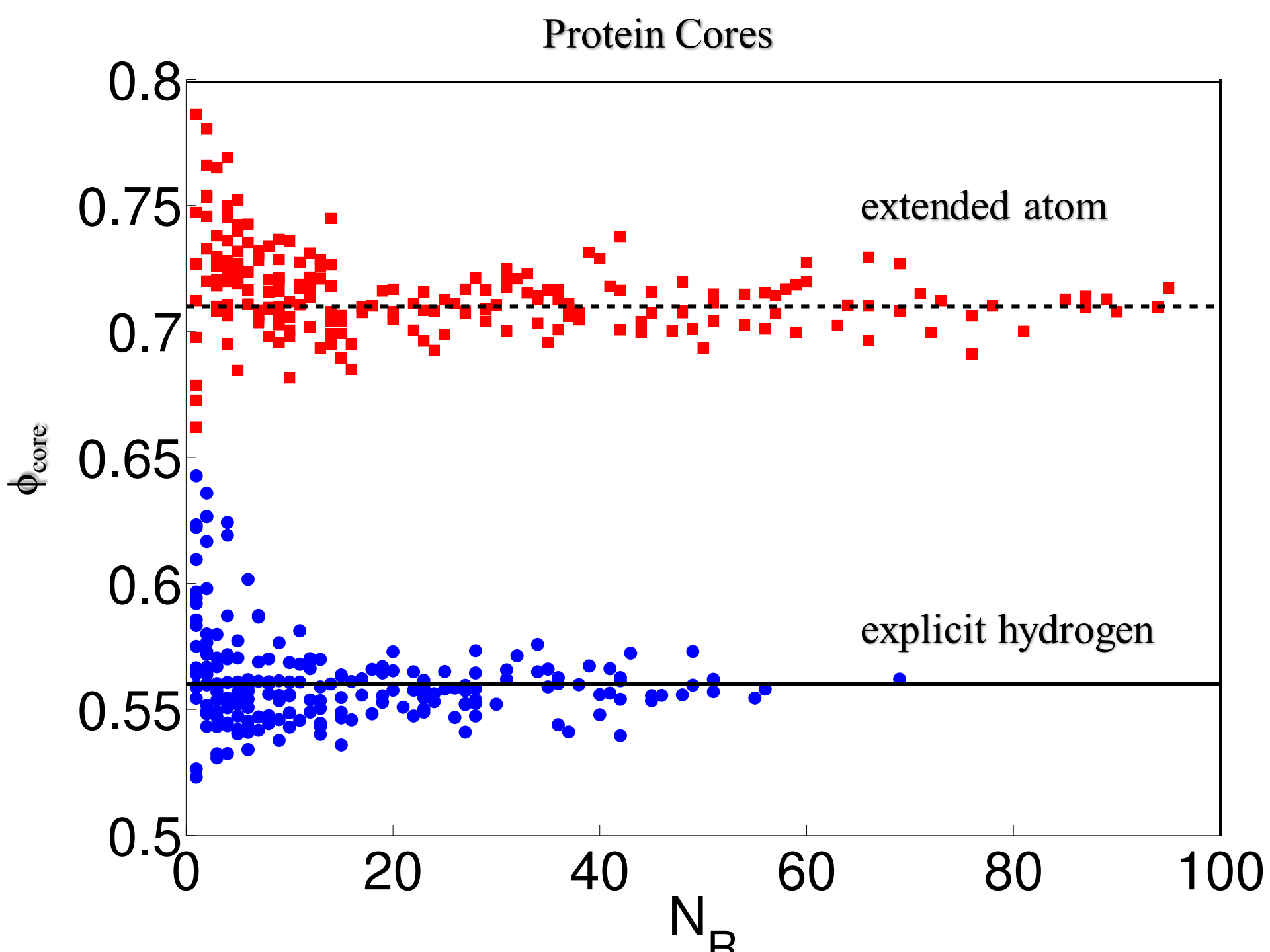
Voronoi Tessellation

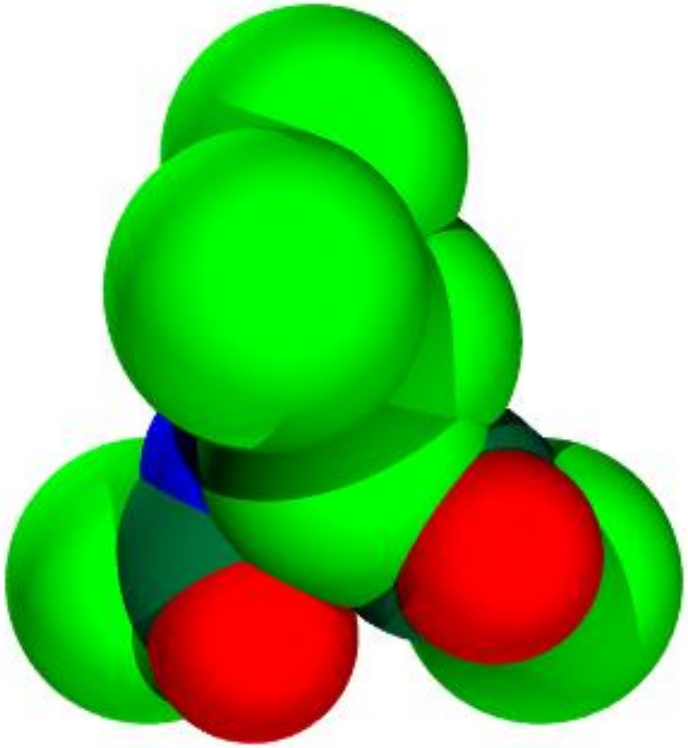


(Non-overlap) Volumes of Core Residues w/ Explicit Hydrogens

Residue Type	Volume V_r (Å^3)
Leu	88.06
Ala	43.62
Val	72.45
Ile	88.13
Gly	30.46
Phe	100.66
Cys	61.20
Met	87.49

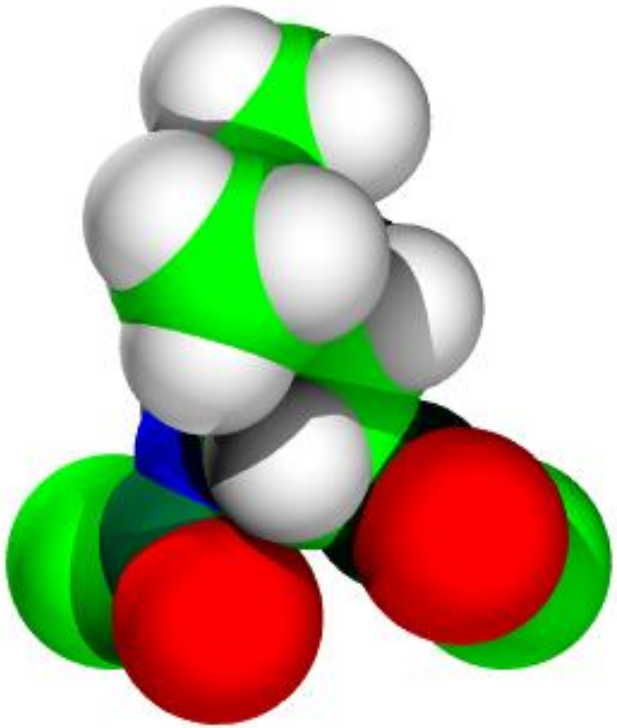



Protein Cores
0.8
0.75
0.7
0.65
0.6
0.55
0.5
ϕ_{core}
extended atom
explicit hydrogen
0
20
40
60
80
100
N_R



Leu
extended atom

V_r=107.2Å^3

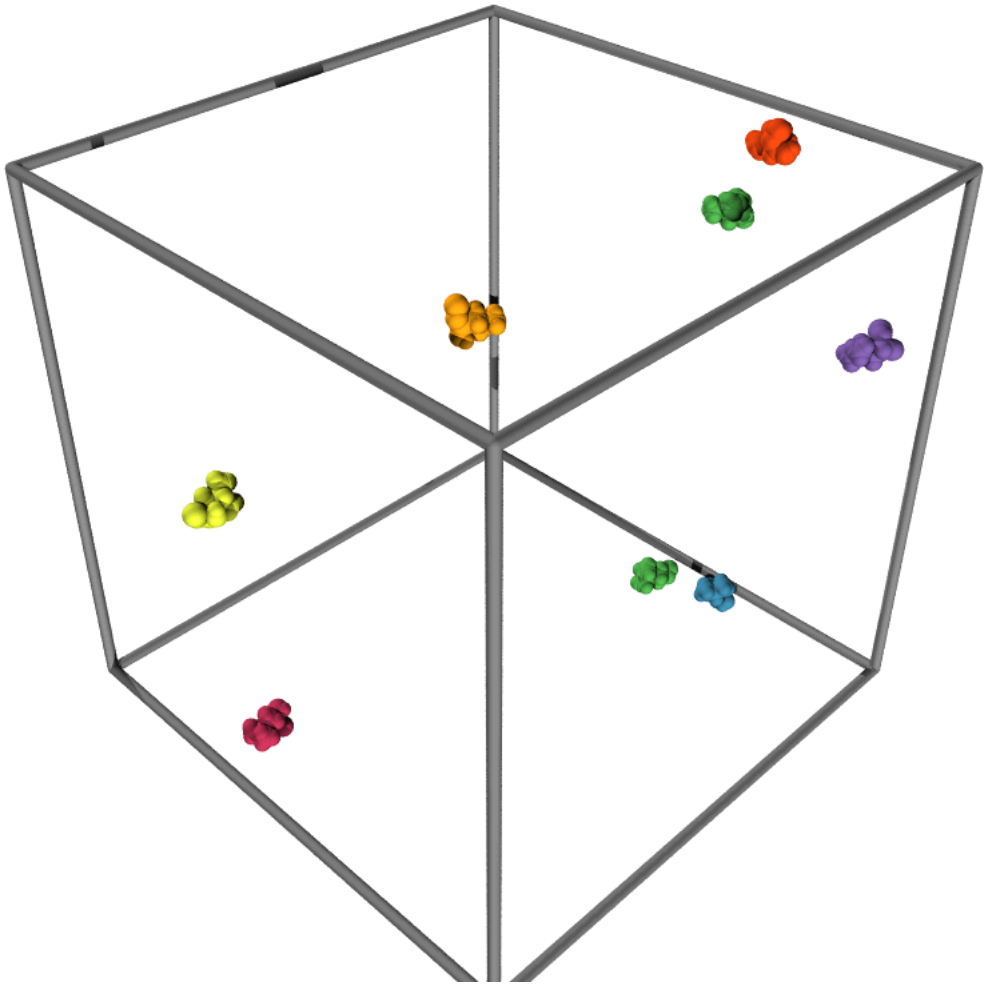


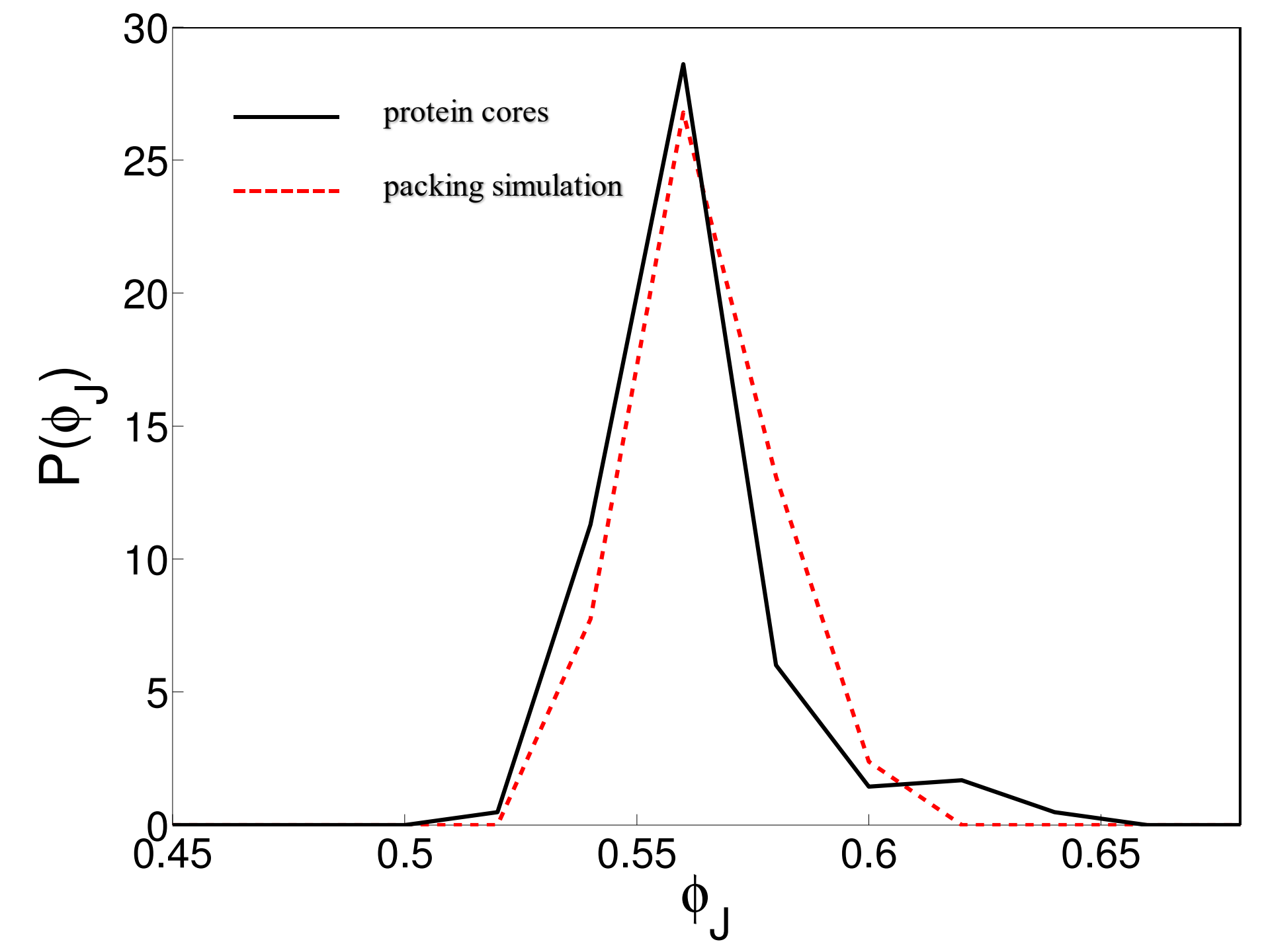
Leu
explicit hydrogens

V_r=88.1Å^3

O
C
H
N

Packing Simulations




protein cores
packing simulation
P(ϕJ)
ϕJ
0
5
10
15
20
25
30
0.45
0.5
0.55
0.6
0.65

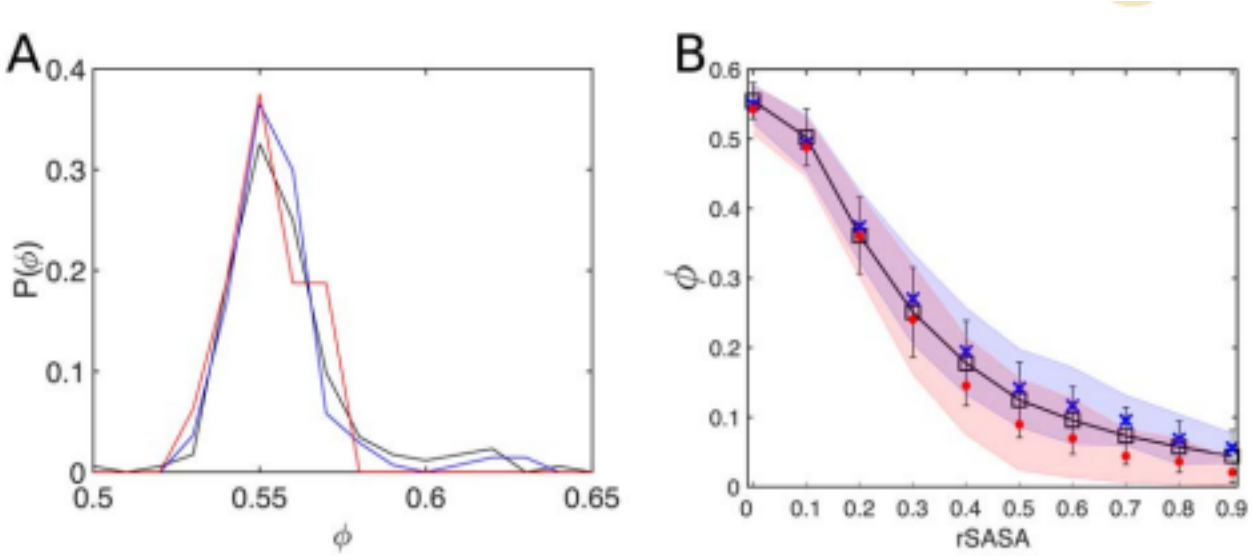
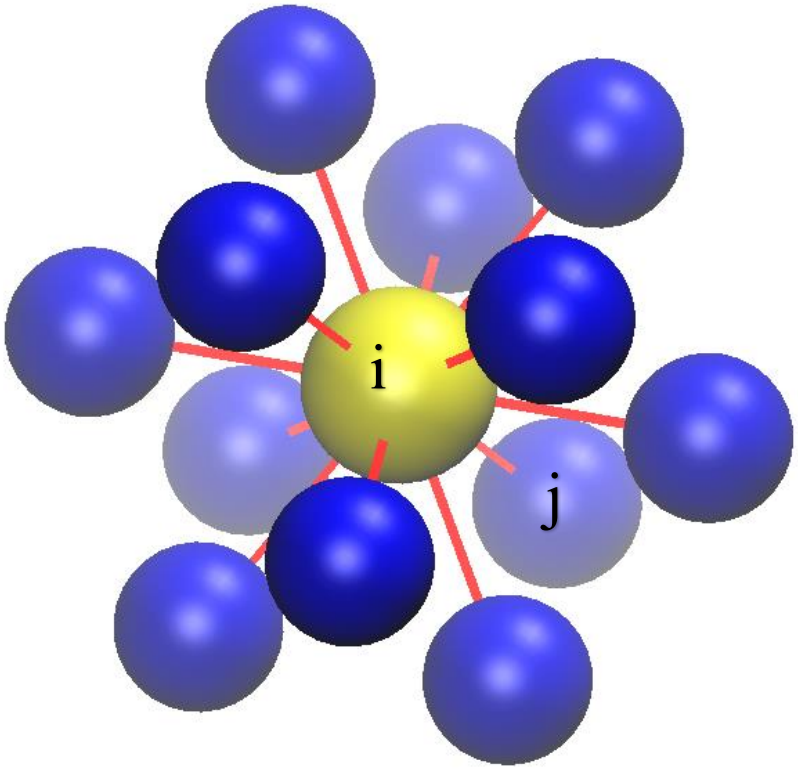


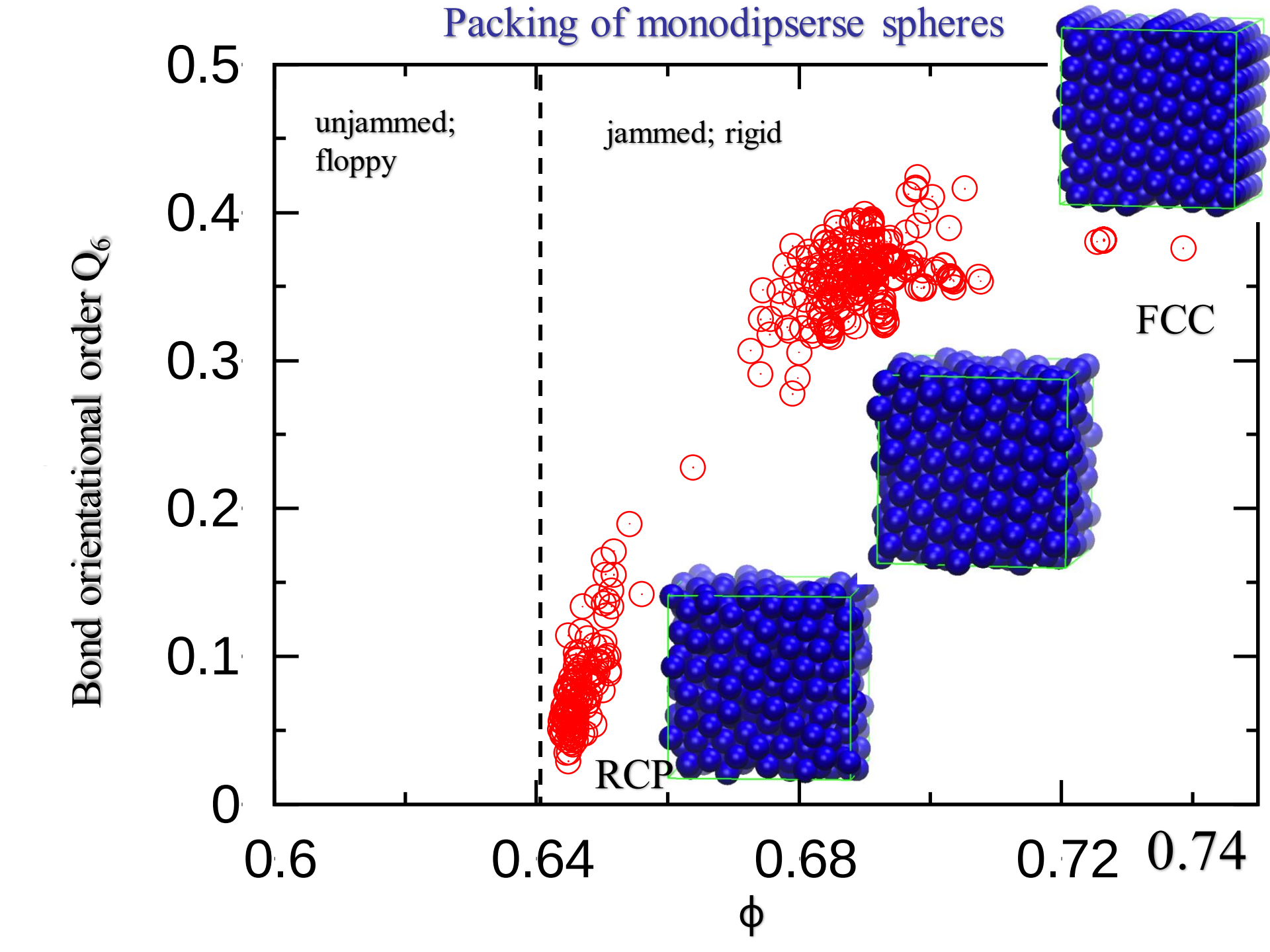
FIGURE 6 A, Distribution of packing fractions $P(\phi)$ of core residues in the Dun1.0 (black), PPI (blue), and TM (red) datasets. ϕ is calculated using Equation 3, where the summation is over all atoms of all core residues in each protein. B, Packing fraction ϕ of residues as a function of the relative solvent accessibility (rSASA) for the Dun1.0 (black line and squares), PPI (blue crosses), and TM (red circles) datasets. The error bars indicate the standard deviation for the Dun1.0 dataset and the blue and red shaded regions indicate the standard deviations for the PPI and TM datasets, respectively [Color figure can be viewed at wileyonlinelibrary.com]

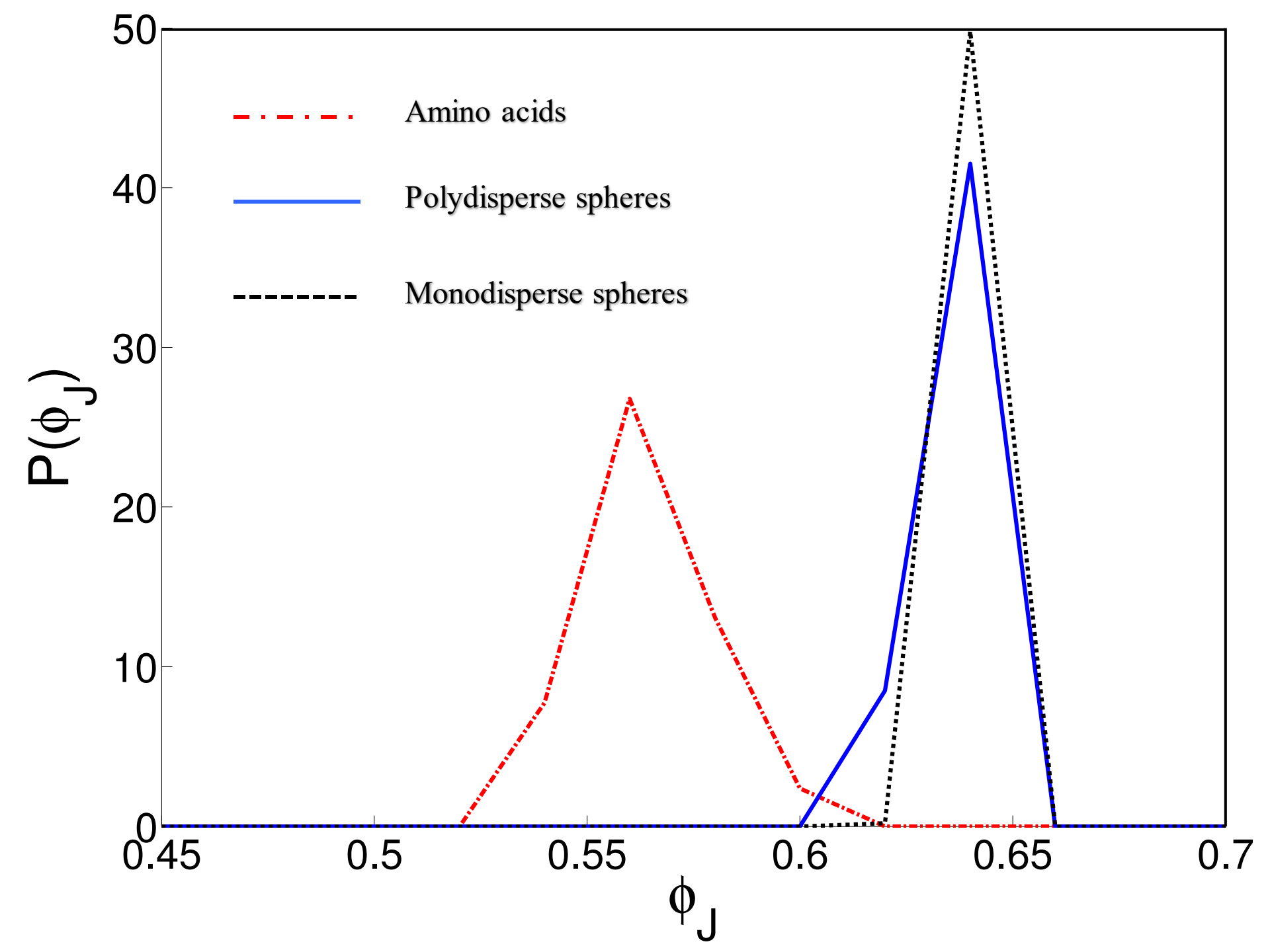
Why is the packing fraction of protein cores $\phi<0.74$?

Bond-orientational order parameter Q_l



$$Q_l^{local} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{4\pi}{2l+1}\sum_{m=-l}^{l}\left|\frac{1}{n_i}\sum_{j=1}^{n_i} Y_l^m\left(q_{ij}, f_{ij}\right)\right|^2\right)^{1/2}$$


Packing of monodipserse spheres
unjammed;
floppy
jammed; rigid
FCC
RCP
Bond orientational order Q_6
0.5
0.4
0.3
0.2
0.1
0
0.6
0.64
0.68
0.72
0.74
ϕ


Amino acids
Polydisperse spheres
Monodisperse spheres
P(ϕJ)
ϕJ
0
10
20
30
40
50
0.45
0.5
0.55
0.6
0.65
0.7

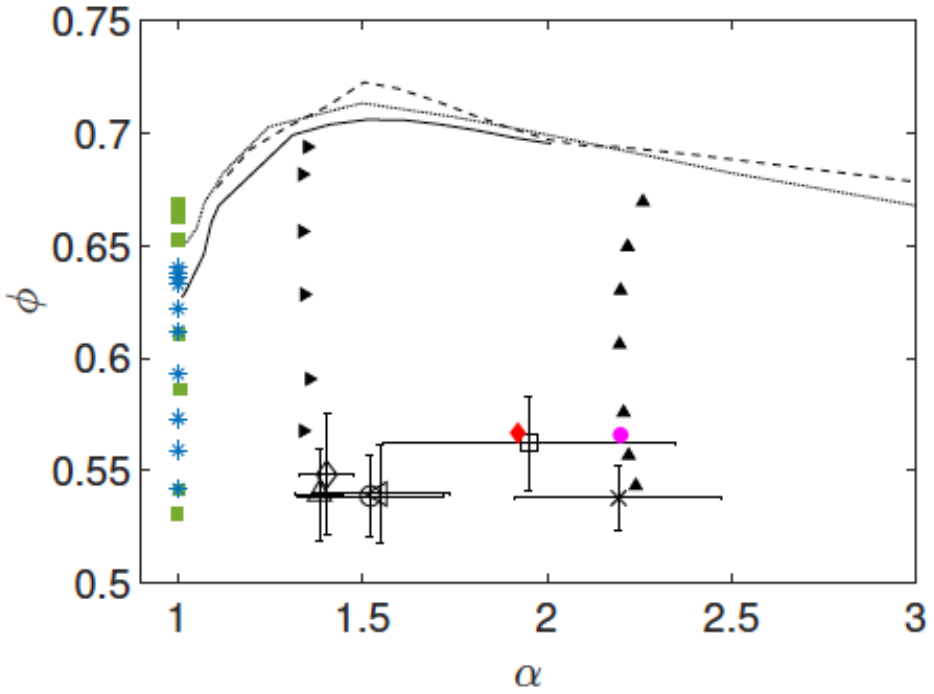


Figure 9: Jammed packing fraction ϕ versus aspect ratio α for frictional spheres (blue asterisks) from Ref. [57], bumpy (green triangles) spheres, smooth, prolate ellipsoids of revolution from Refs. [59] (dotted line) and [60] (solid line) and spherocylinders (dashed line) from Ref. [58]. The static friction coefficient for the frictional spheres varies from $\mu = 10^{-4}$ to 10 from top to bottom. For the bumpy spheres (Fig. 10 (a) and (b)), twelve bumps are placed on the vertices of an icosohedron, and the relative sizes of the bumps are decreased to increase the bumpiness B from $\approx 10^{-2}$ to 0.15 from top to bottom. We also show the packing fraction and aspect ratio for Ala (open diamond), Ile (open leftward triangle), Leu (open circle), Met (open square), Phe (x), and Val (open upward triangle) residues in protein cores. The error bars indicate the root-mean-square fluctuations from averaging over instances of each residue with different backbone and side chain conformations. Results for bumpy ellipsoids are indicated by the filled rightward and upward triangles and results for the non-axisymmetric shapes in Fig. 10 (g) and (h) are indicated by the filled diamond and circle, respectively.

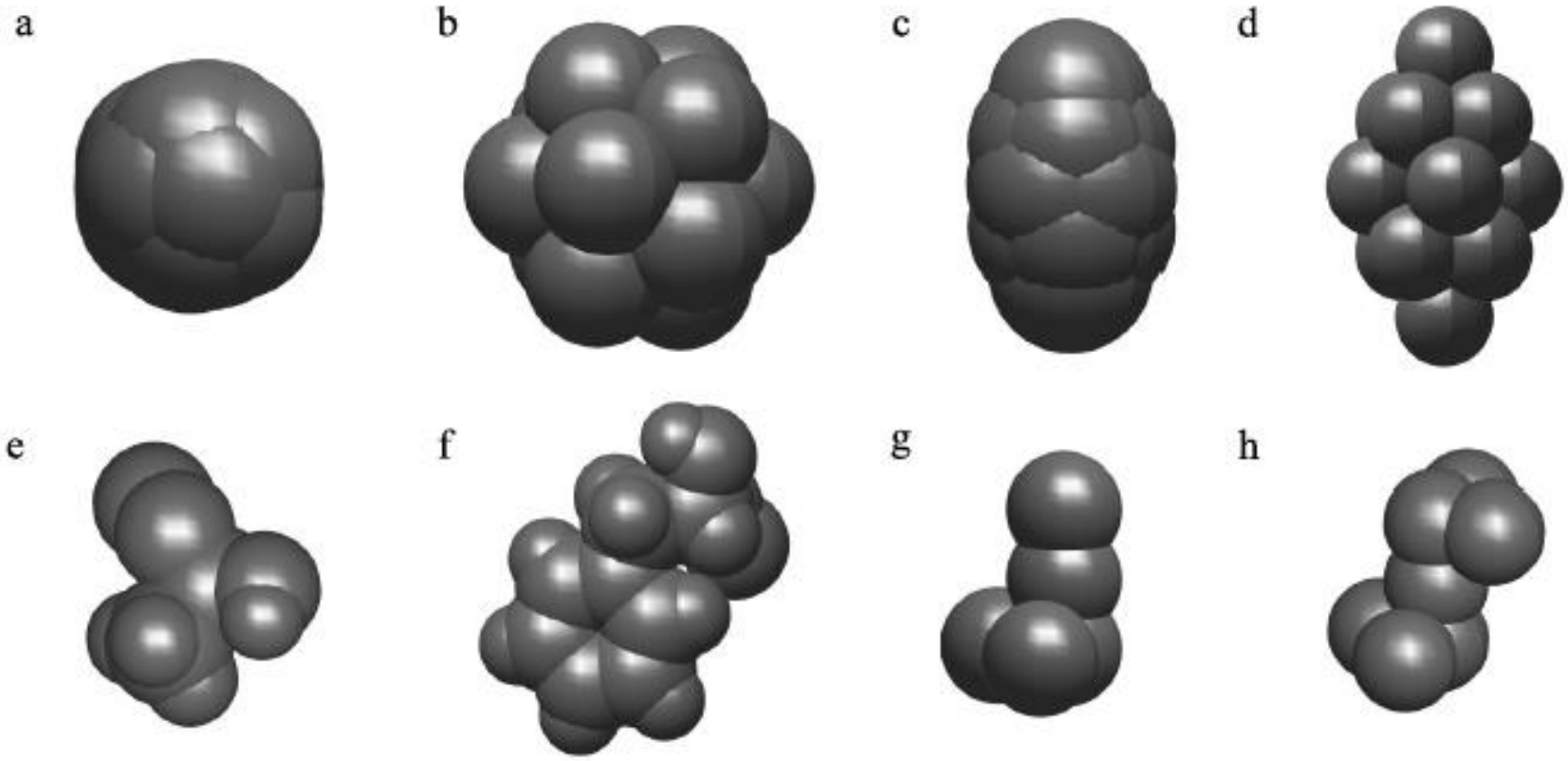


Figure 10: Examples of the composite particle shapes investigated in the packing simulations: bumpy spheres with (a) $B = 0.008$, $\alpha = 1.00$ and (b) $B = 0.113$, $\alpha = 1.00$; bumpy ellipsoids with (c) $B = 0.015$, $\alpha = 1.40$ and (e) $B = 0.162$, $\alpha = 1.40$; (e) Ala and (f) Phe residues; and (g,h) two examples of non-axisymmetric composite particles.

Conclusions

- Previous calculations of ϕ_{core}~0.70-0.74, are likely too high for protein core packing; would imply structural order
- We predict ϕ_{core}~0.56 using a hard-sphere model with explicit hydrogens and atomic radii calibrated to observed side-chain dihedral angle distributions
- Showed that ϕ_{core}~0.56 can be obtained from hard-particle simulations of individual amino acids under isotropic compression
- Apply results to mutations of protein cores and interfaces

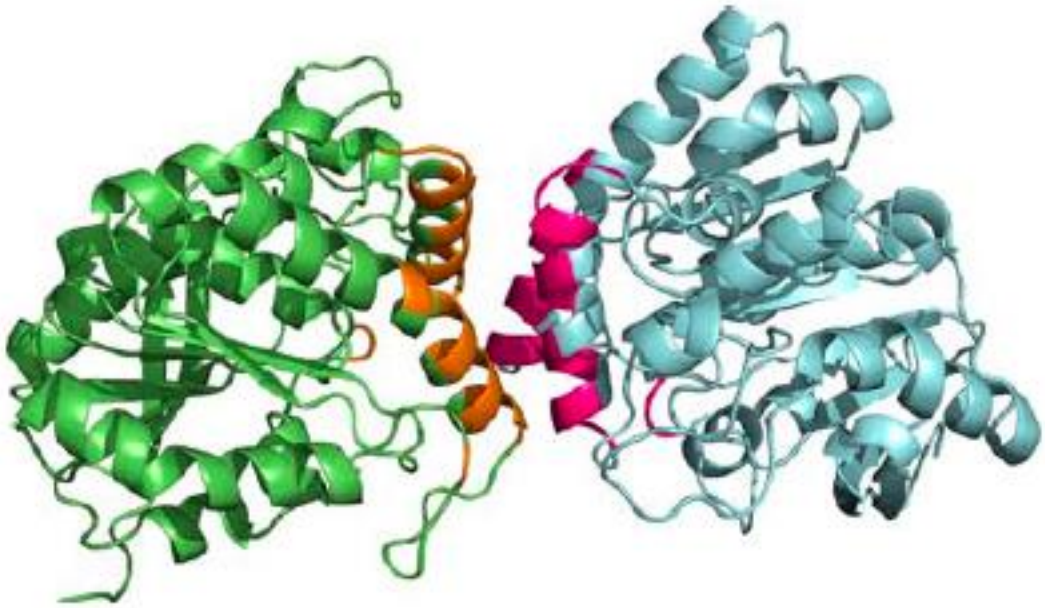


FIGURE 2 Ribbon representation of a protein-protein complex (PDB identifier: 1DQZ). The 2 protein chains are shown in green and blue. The interface residues (displayed in orange and pink) were identified as those residues with a change in SASA, $\Delta SASA_{Res} > 0.1$ Å^2, between the monomer and the complex [Color figure can be viewed at wileyonlinelibrary.com]

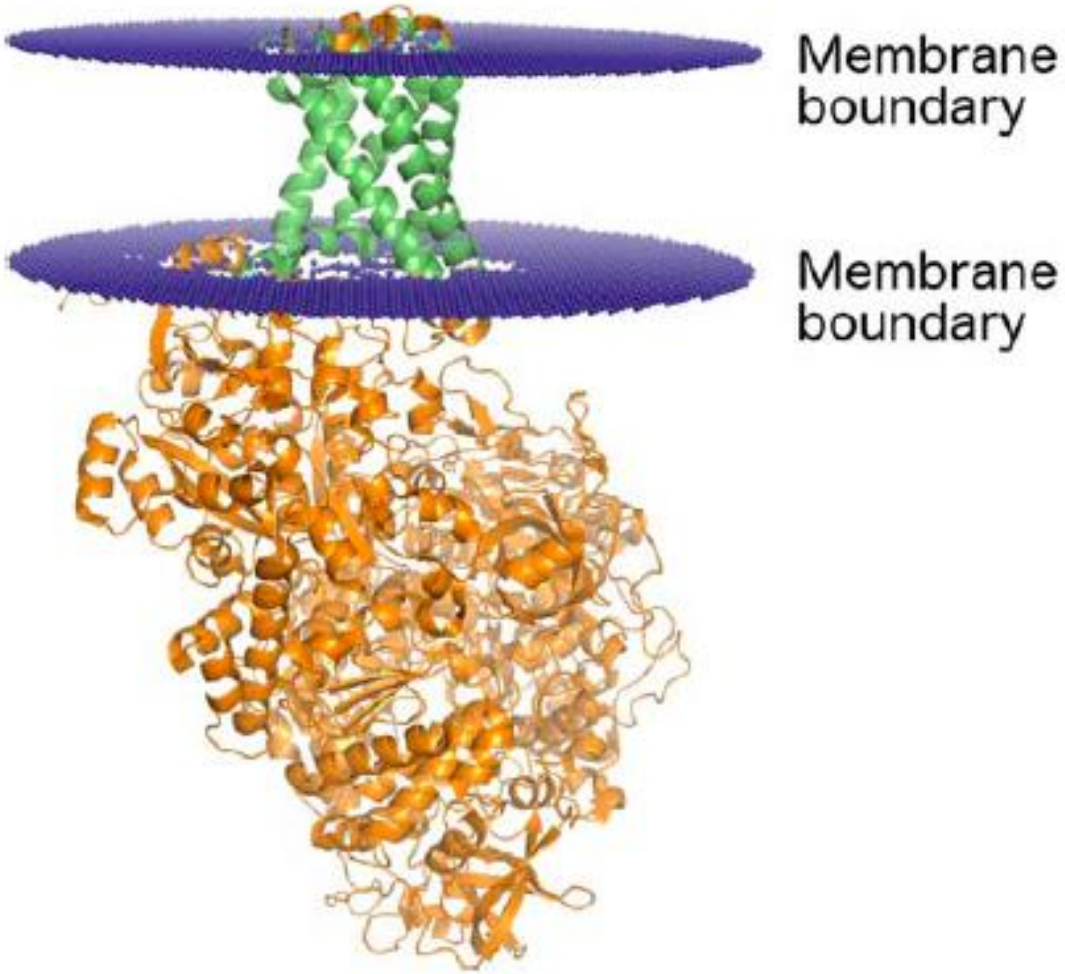


FIGURE 3 Ribbon representation of a transmembrane protein (PDB identifier: 1Q16). The membrane boundary planes (displayed in blue) were obtained from the Positioning of Proteins in Membranes (PPM) server [52]. The region of the protein that spans the membrane is shown in green, and the portion of the protein that extends beyond the membrane is shown in orange [Color figure can be viewed at wileyonlinelibrary.com]

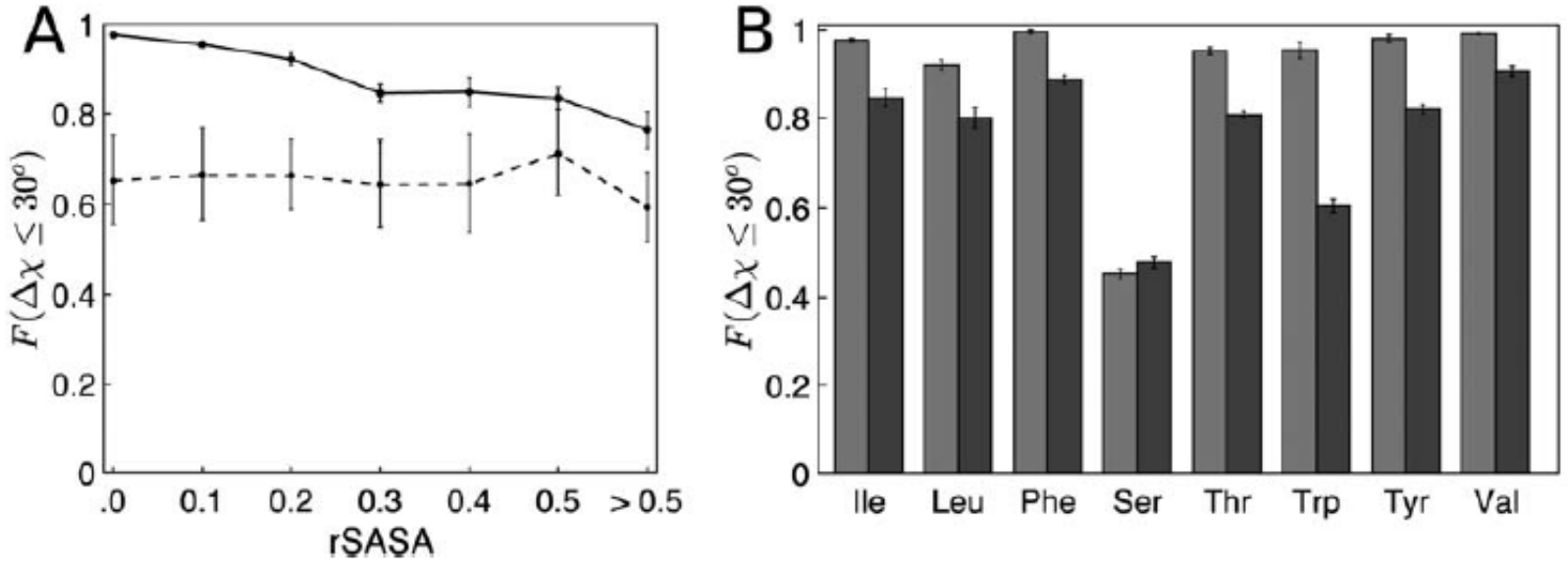


FIGURE 7 A, Fraction of residues predicted within 30° ($F(\Delta\chi \leq 30^\circ)$) for Ile residues in the Dun1.0 database (solid line) and their corresponding dipeptide mimetics (dotted line) as a function of rSASA values. The dotted line provides lower bounds for the prediction accuracy for the residues in each rSASA bin. Due to the low frequency of uncharged residues in the non-core region, we have combined all residues with rSASA > 0.5 into 1 bin. B, $F(\Delta\chi \leq 30^\circ)$ for non-charged amino acids for rSASA $< 10^{-3}$ (light grey) and $0.2 <$ rSASA ≤ 0.3 (dark grey)

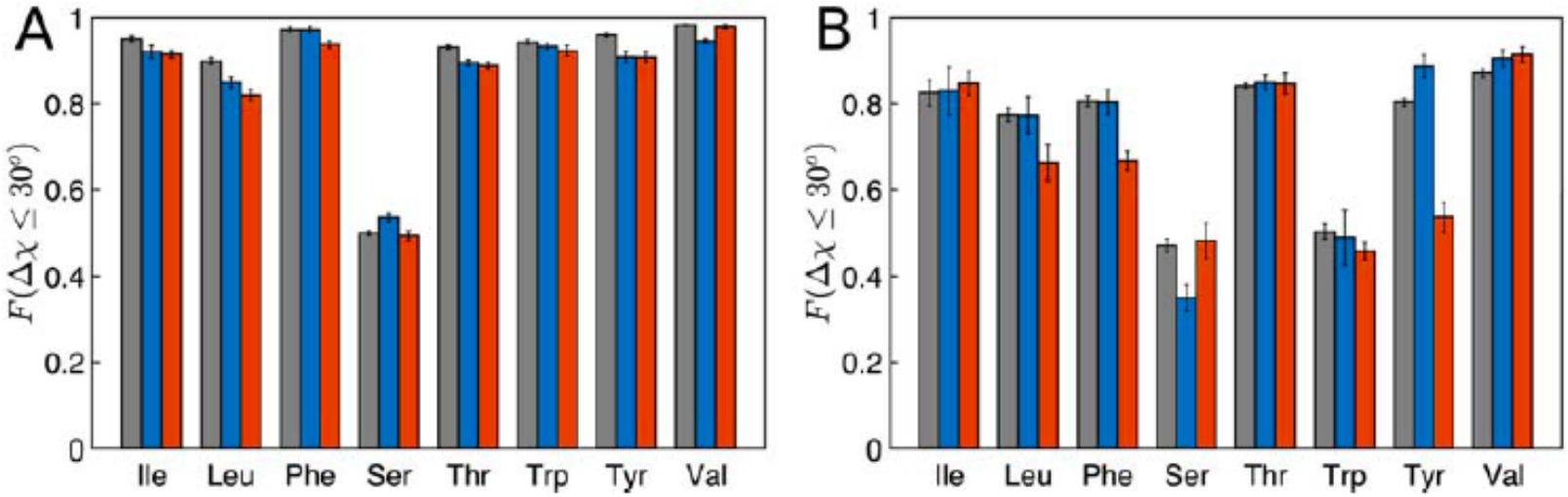


FIGURE 8 $F(\Delta\chi \leq 30^{\circ})$ for non-charged amino acids for A, rSASA < 0.1 and B, $0.2 <$ rSASA ≤ 0.3 for the Dun1.0 (grey), PPI (blue), and TM (red) datasets [Color figure can be viewed at wileyonlinelibrary.com]