Are there any differences between x-ray crystal structures and NMR structures of proteins?

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,
U. Edinburgh

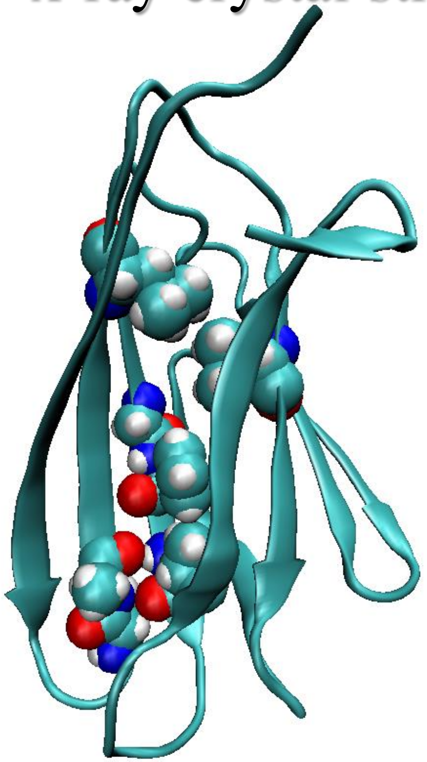
Alex Grigas,
CBB

Jack Treado
MEMS

Grace Meng
Chemistry

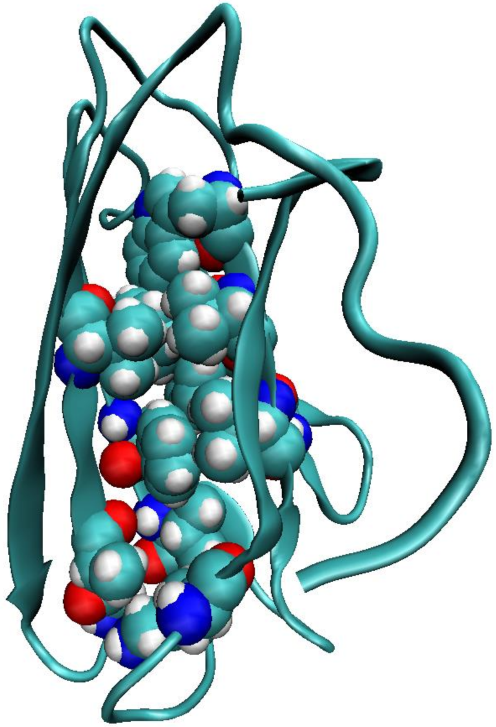
Zhuoyi Liu
MEMS

Program in Physics, Engineering, and Biology (PEB) x-ray crystal structure



PDBID: 2CWR
Resolution: 1.7Å

NMR structure



PDBID: 2CZN
25 NOE restraints/residue

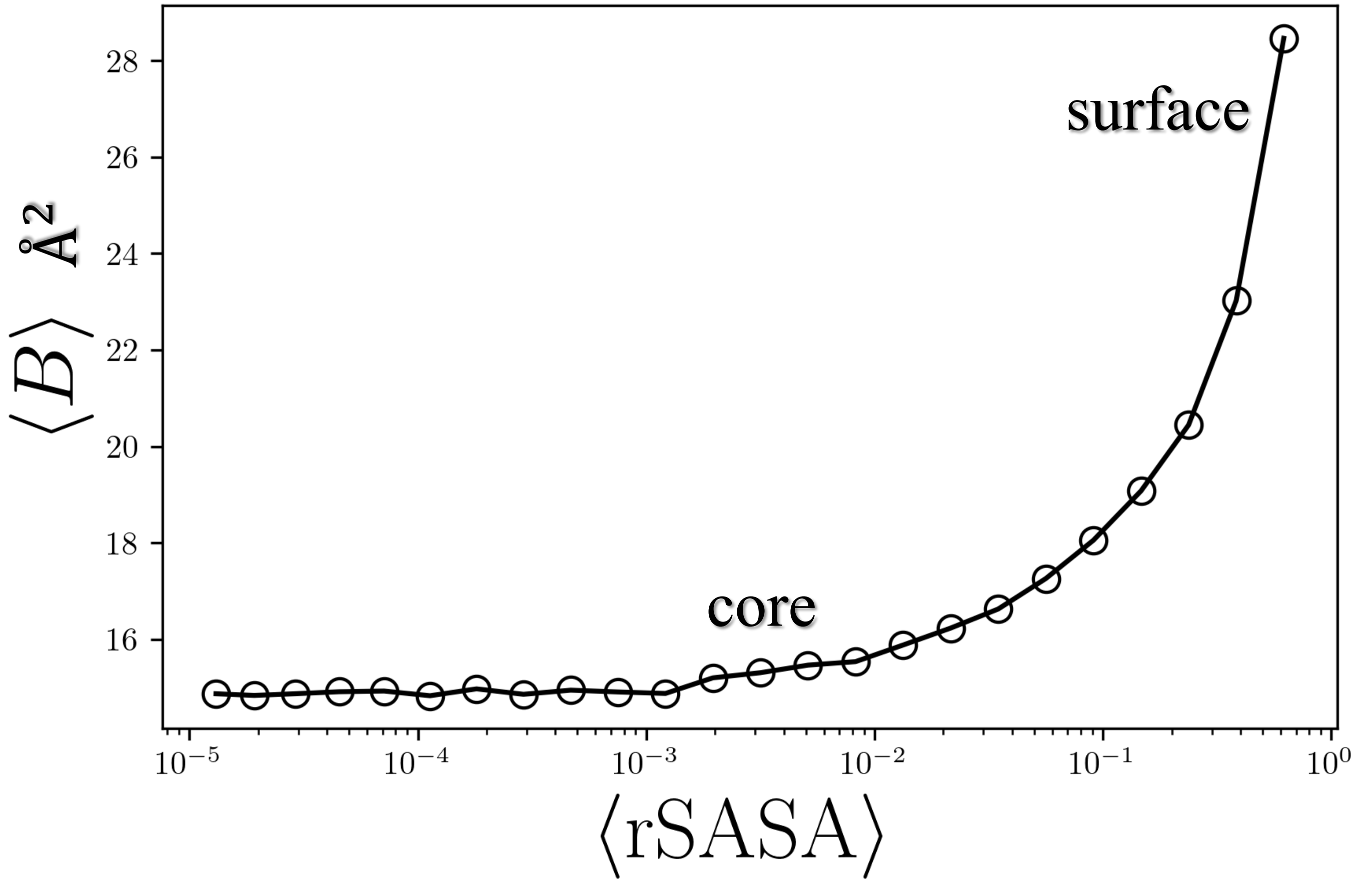
Chitin binding domain
of chitinase from
Pyrococcus furiosus; 103 AA

A. T. Grigas, Z. Liu, L. Regan, and C. S. O'Hern, "Core packing of well-defined X-ray and NMR protein structures is the same," to appear in Protein Science (2022)

To quantify differences between x-ray crystal structures and NMR structures, we will focus on “core” residues.

(a) Compare *average* properties of high-resolution x-ray crystal structures (5621) and high-quality NMR structures (6449).
(b) Compare NMR and x-ray crystal structures of the same protein (702 pairs).

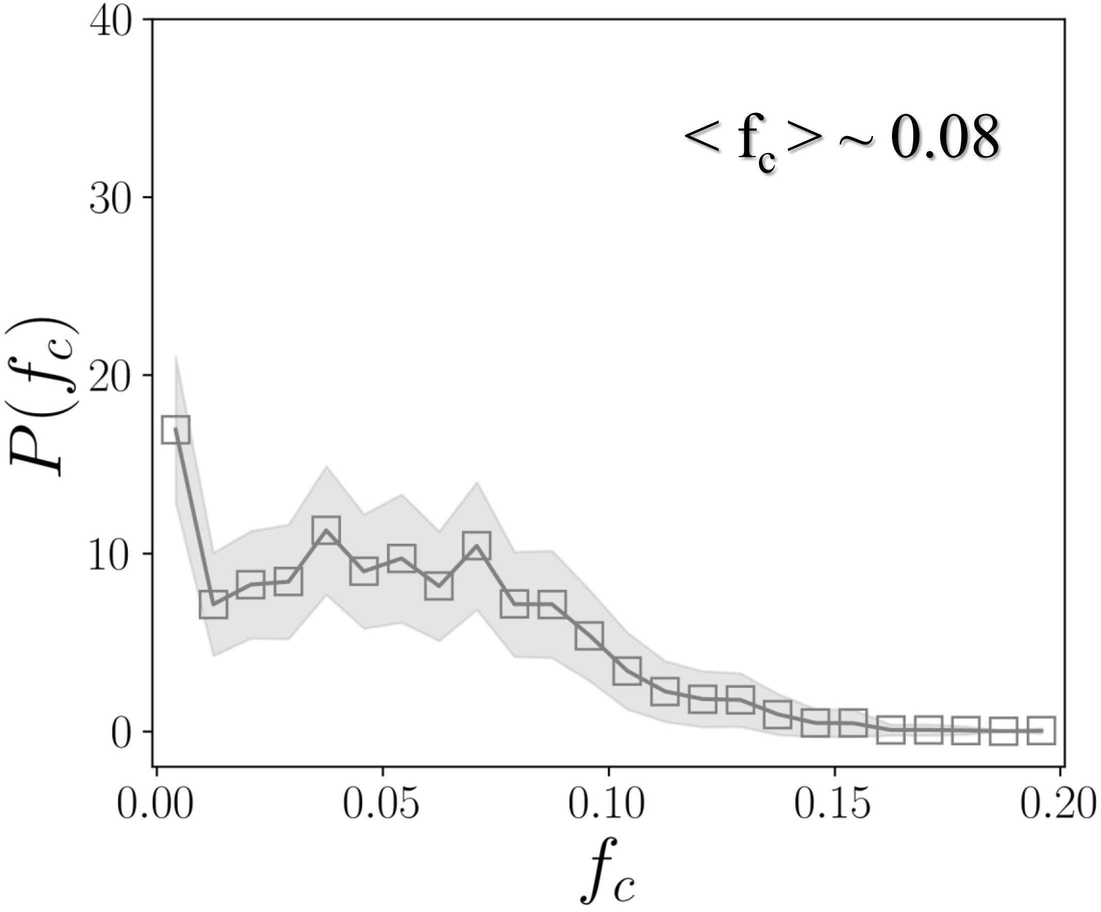
Fluctuations in x-ray crystal structures



$$[0,1] = rSASA = \frac{SASA_{protein}}{SASA_{dipeptide}}$$

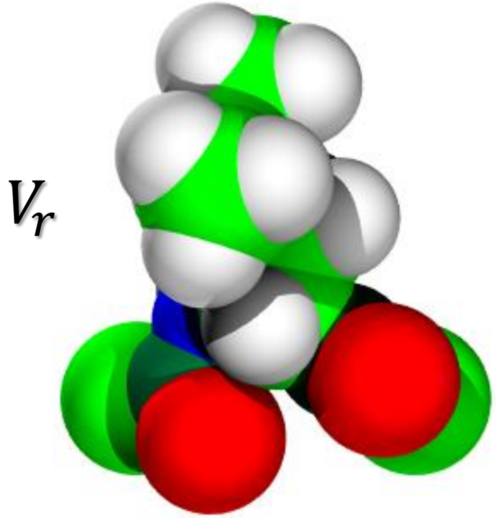
Features of protein cores: 1. fraction of core residues,
2. packing fraction, …

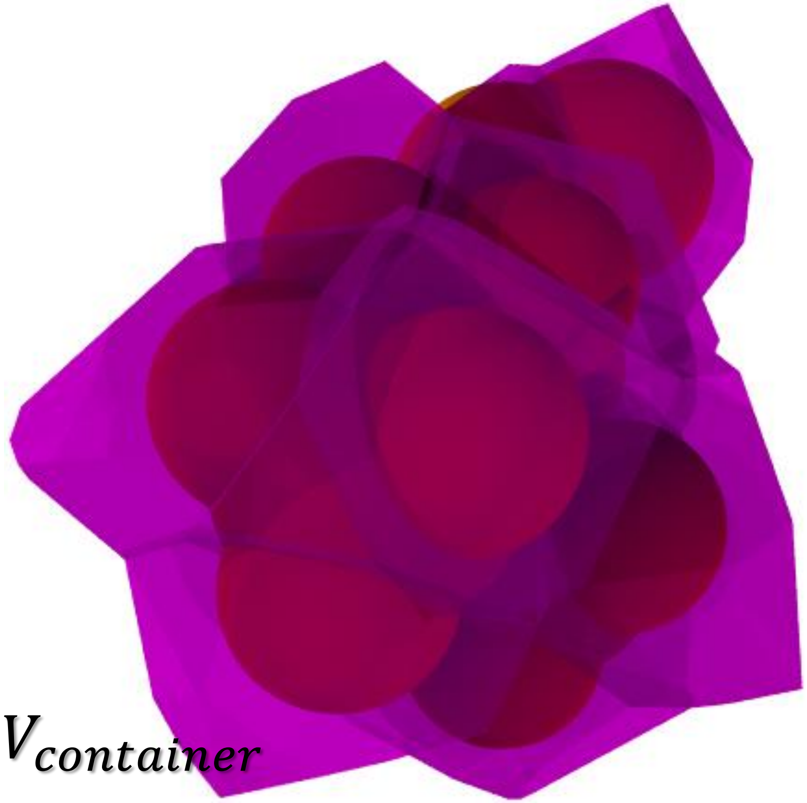
Fraction of core residues in x-ray crystal structures

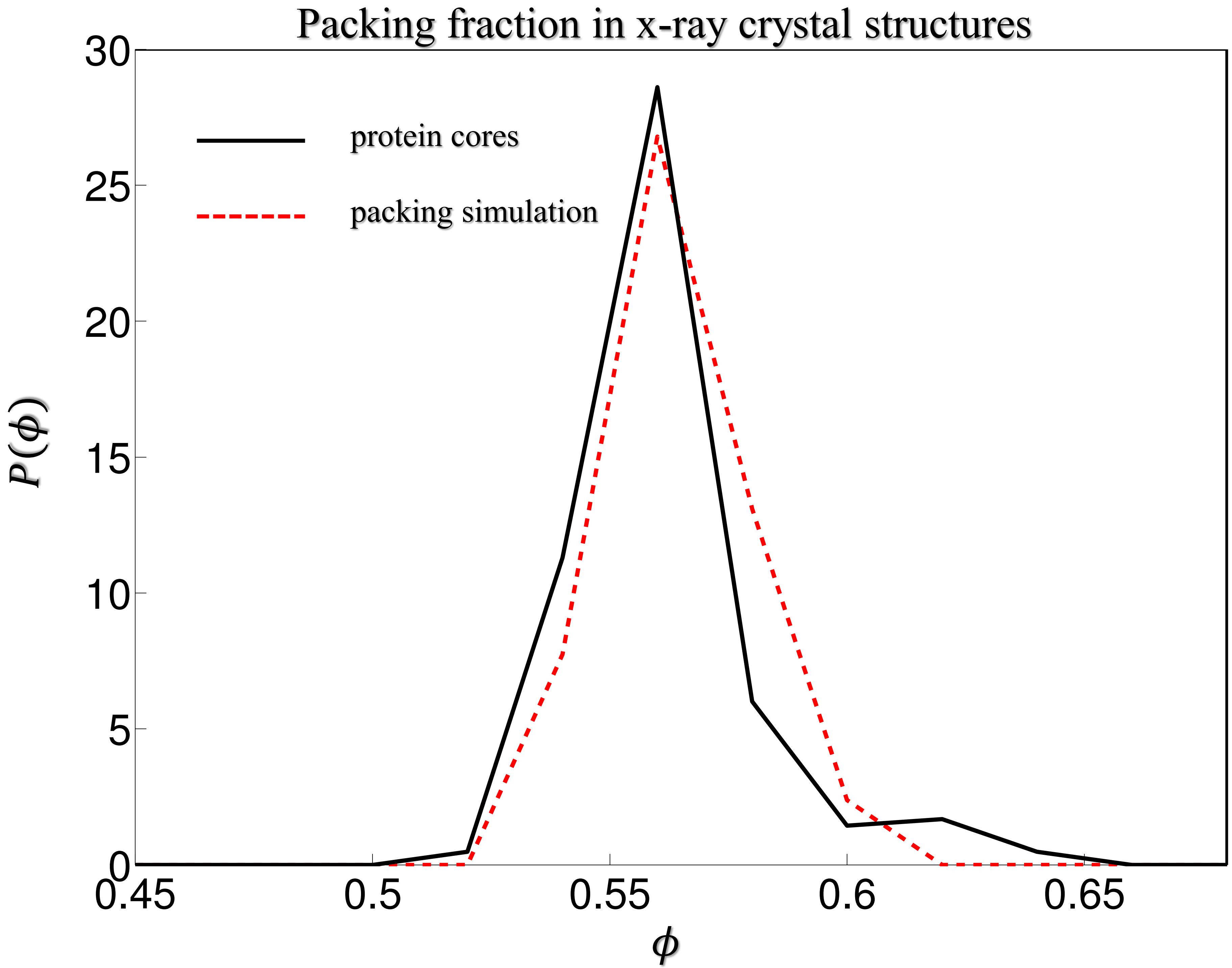


$< f_c > \sim 0.08$

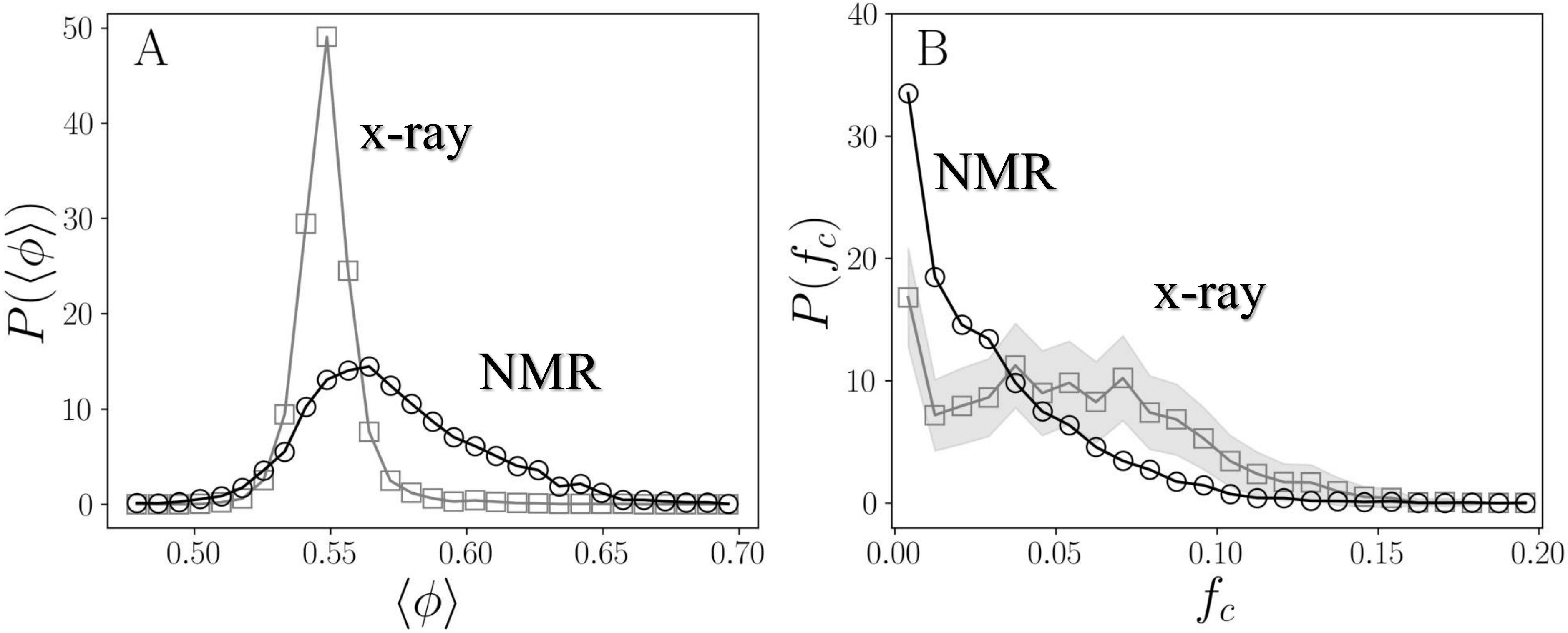
$$\phi = \frac{V_r}{V_{container}}$$





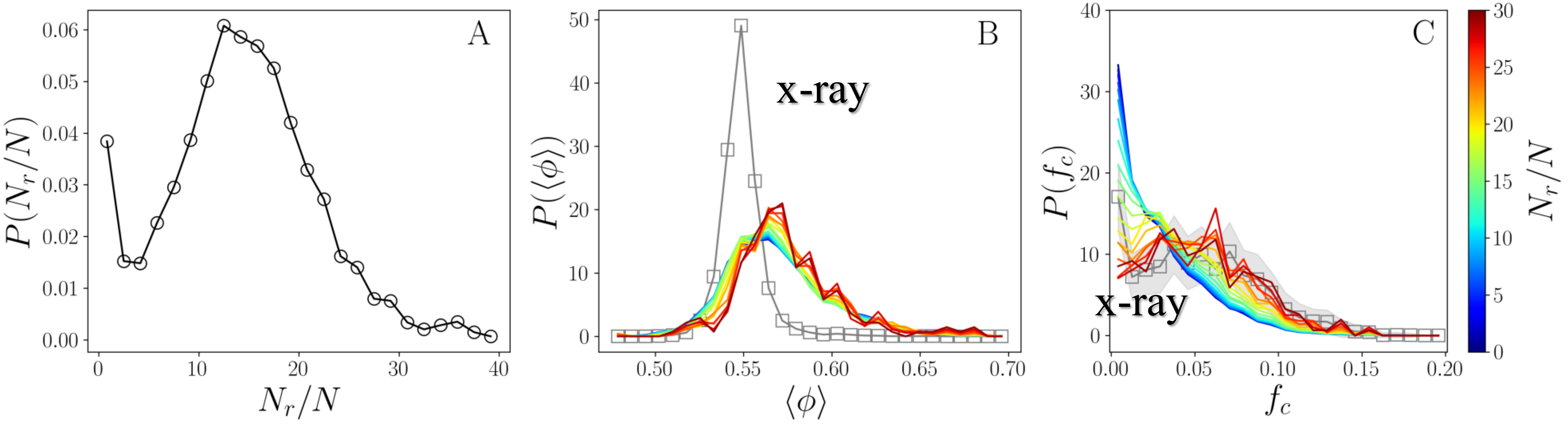

Packing fraction in x-ray crystal structures
protein cores
packing simulation
$P(\phi)$
ϕ
0
5
10
15
20
25
30
0.45
0.5
0.55
0.6
0.65

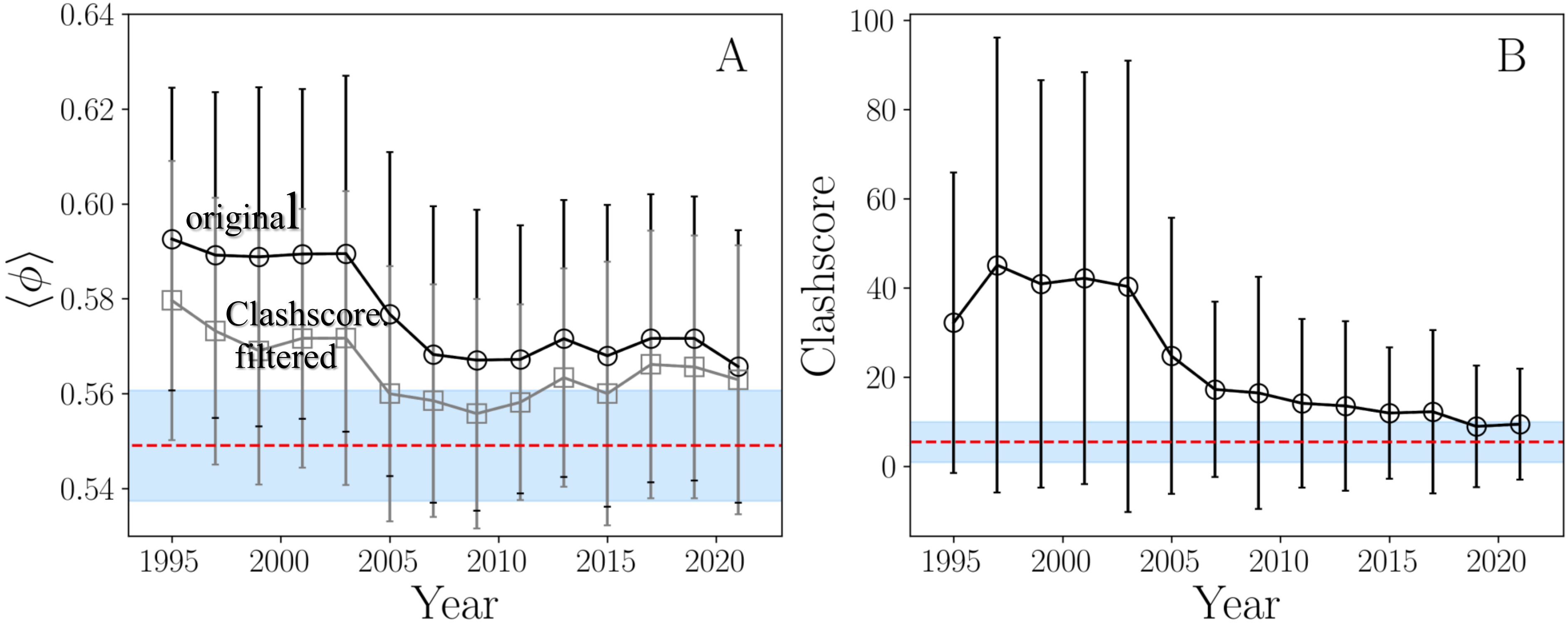
Differences in packing features between NMR and x-ray crystal structures



Are differences in packing features caused by methodological issues or by differences in protein structure in crystalline vs. solution conditions?

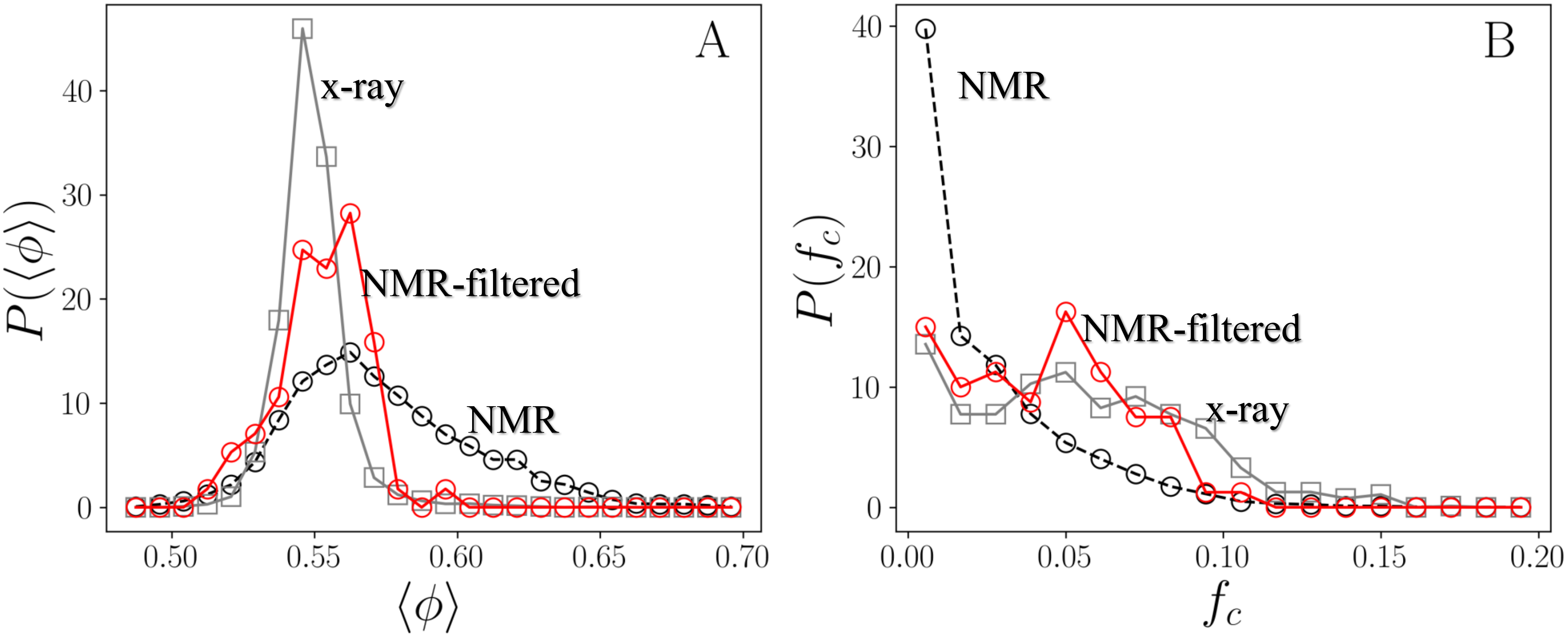
Number of NOE distance restraints



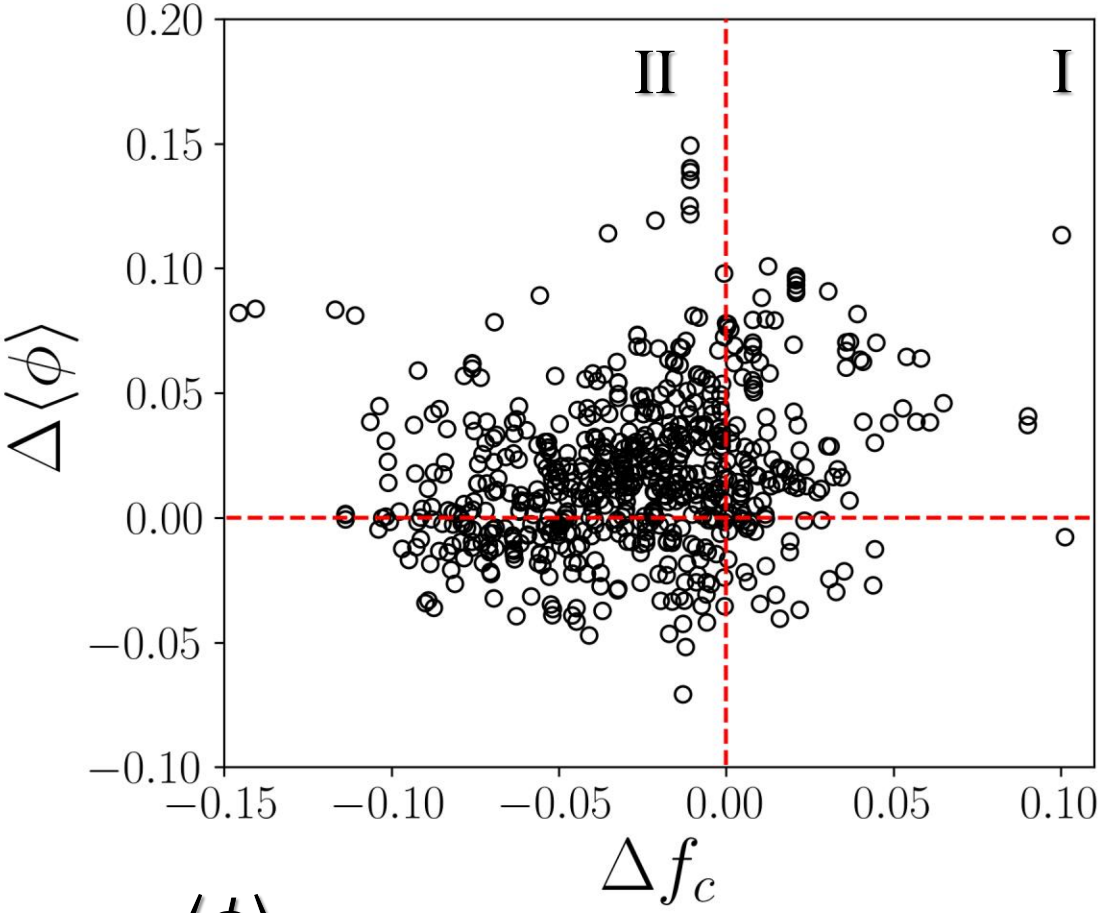


A
B
original
Clashscore-
filtered
⟨φ⟩
0.64
0.62
0.60
0.58
0.56
0.54
Clashscore
100
80
60
40
20
0
1995
2000
2005
2010
2015
2020
Year

Filter NMR structures by clashscore, backbone and sidechain dihedral angle outliers, number of NOE restraints,…



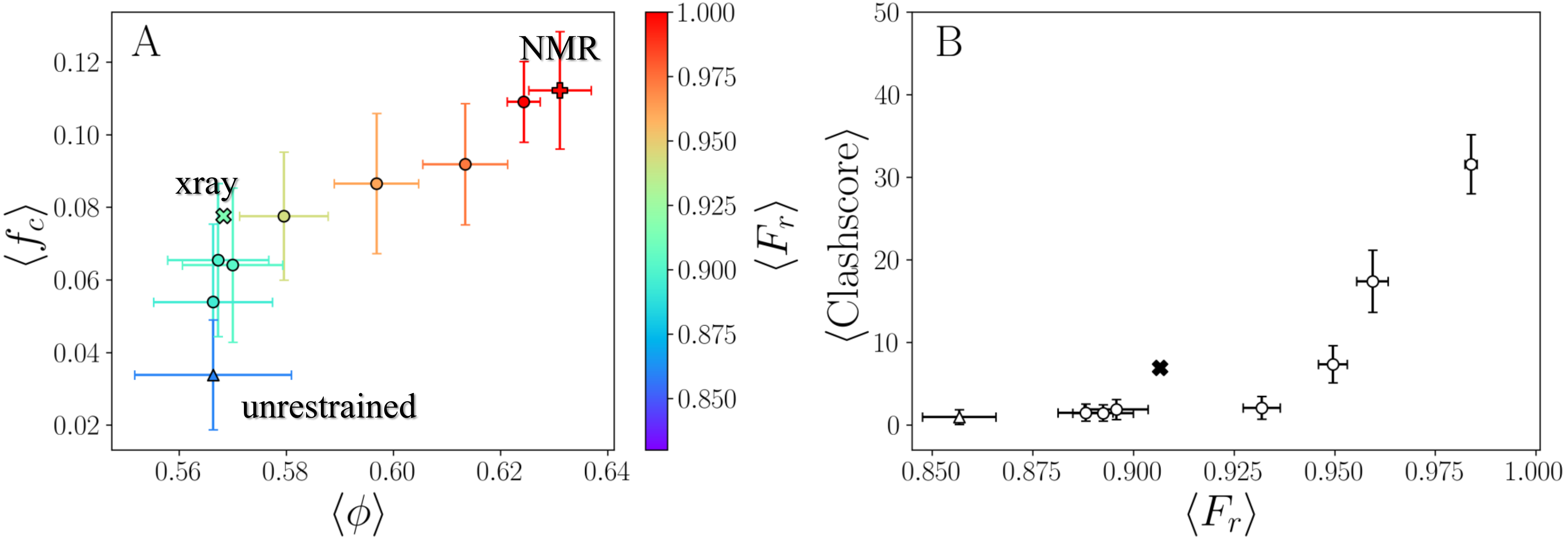
NMR and x-ray crystal structures pairs



$$\Delta\langle\phi\rangle=\langle\phi\rangle_{NMR}-\langle\phi\rangle_{xray}$$

$$\Delta f_c=f_c^{NMR}-f_c^{xray}$$

Unrestrained and NOE distance-restrained all-atom MD simulations



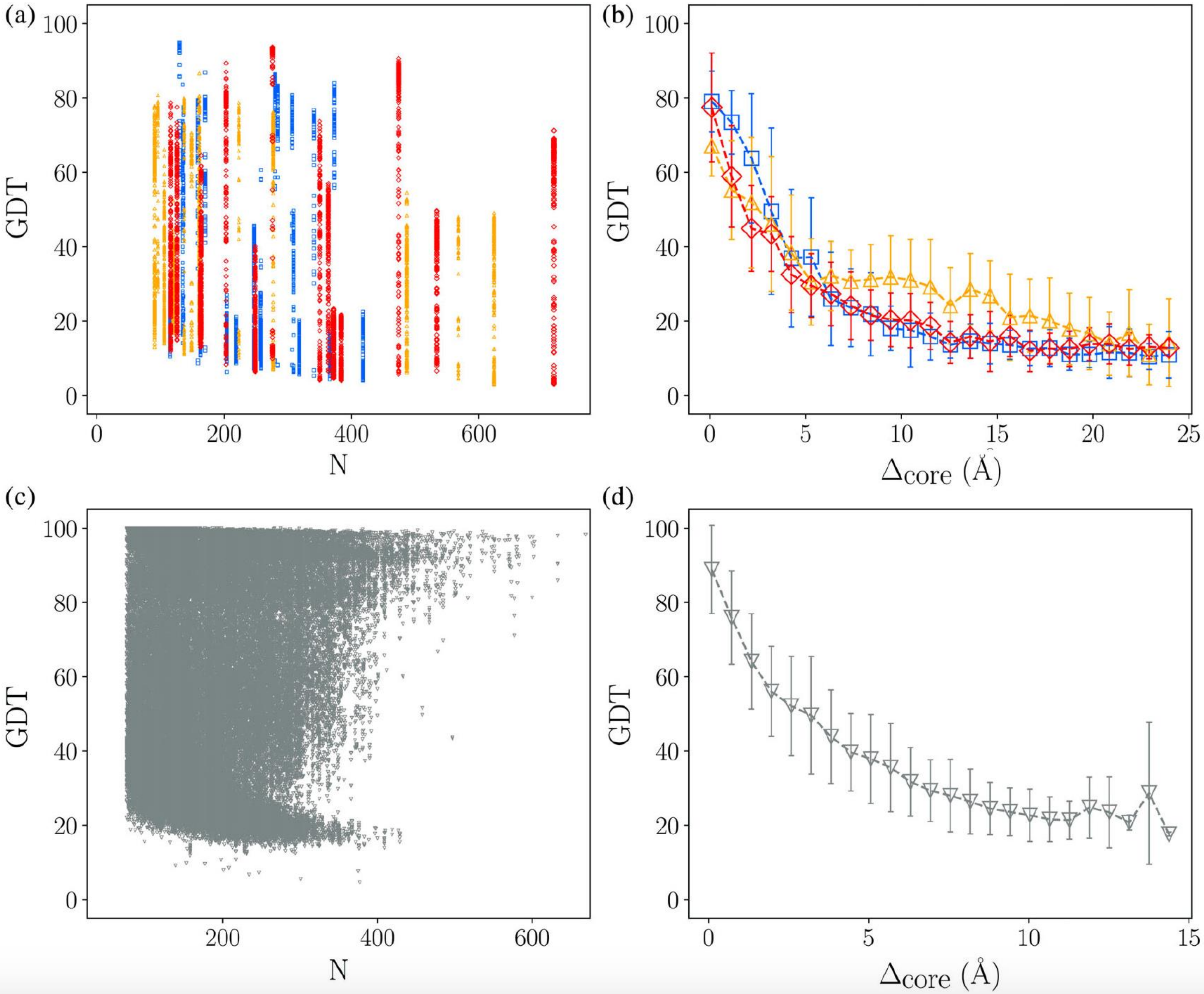
Chitin binding domain
of chitinase from
Pyrococcus furiosus; 103 AA

Conclusions

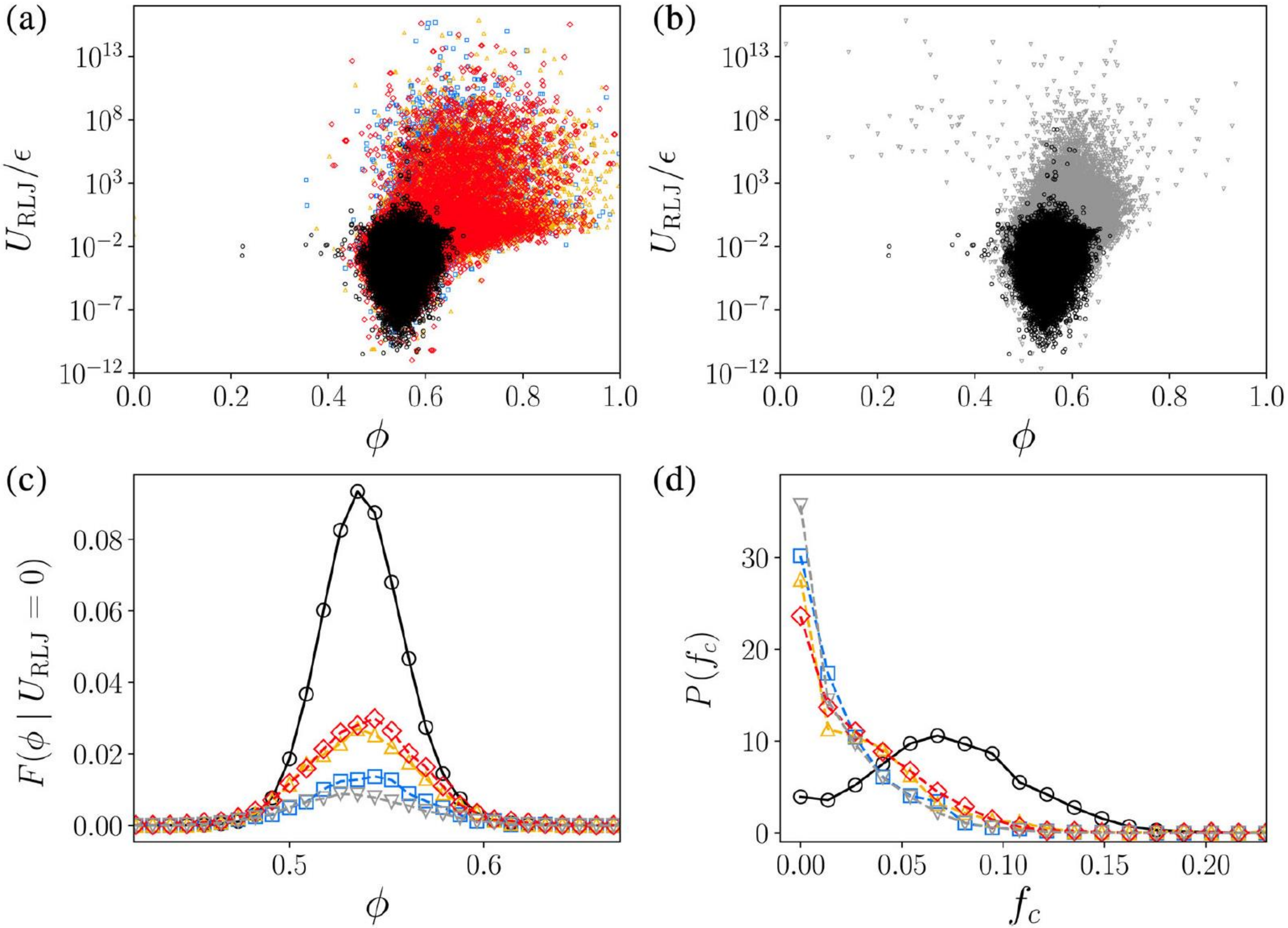
1. There is no agreed upon quality metric for NMR structures
2. When we consider full NMR data set, protein cores are smaller and overpacked compared to those for x-ray crystal structures.
3. When we limit NMR structures to those with large number of restraints, P(f_c) for NMR matches that for x-ray crystal structures. However, NMR structures remain overpacked.
4. When we filter NMR structures according to number of NOE restraints, clashscore, backbone and sidechain dihedral angle outliers…P($\langle \phi \rangle$) is same for NMR and x-ray crystal structures.
5. MD simulations suggest that there are no structures that satisfy protein stereochemistry and *all* NOE restraints
6. Investigate all 702 NMR and x-ray crystal structure pairs

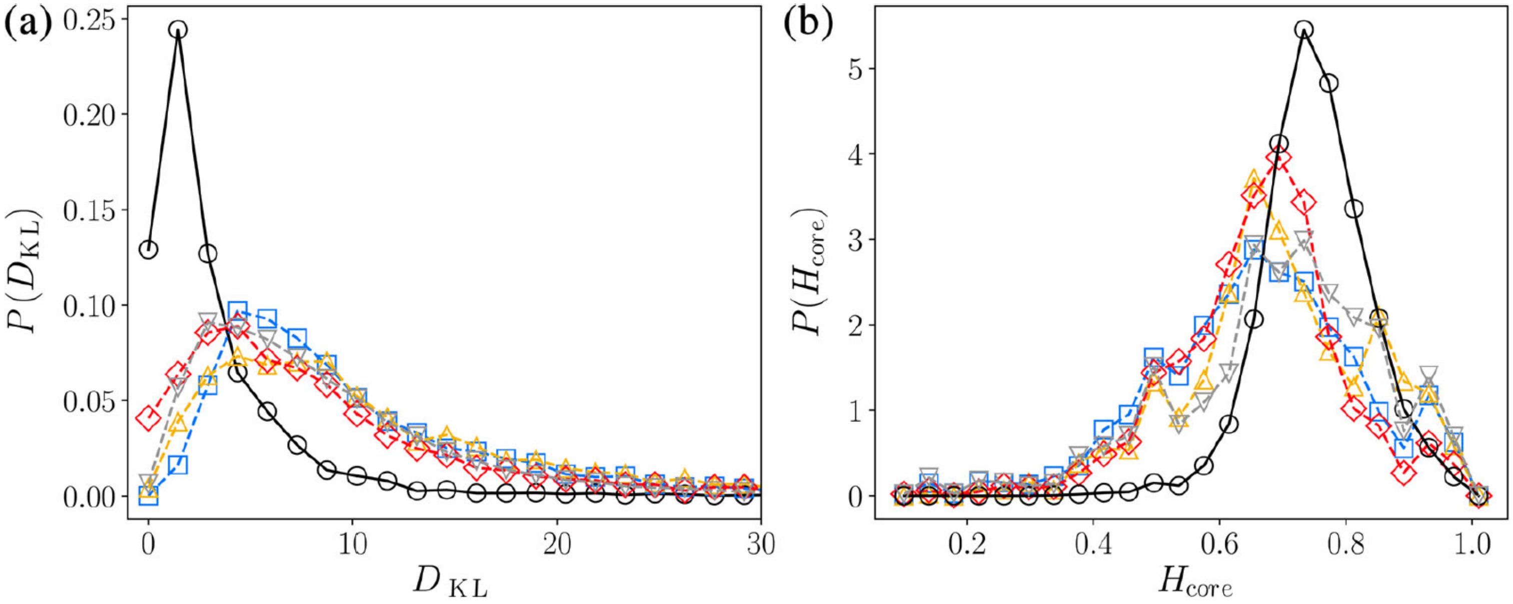
Protein Decoy Detection

A. T. 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.


(a)
GDT
N
0
20
40
60
80
100
0
200
400
600
(b)
GDT
$\Delta_{\rm core}$ (Å)
0
5
10
15
20
25
(c)
GDT
N
200
400
600
(d)
GDT
$\Delta_{\rm core}$ (Å)
0
5
10
15

Distinguishing Features





$$D_{KL} = \int d\phi P(\phi) log\left[\frac{P(\phi)}{Q(\phi)}\right]$$

$$0.5 \leq \langle \phi \rangle \leq 0.62, \quad \langle U_{\mathrm{RLJ}}/\varepsilon \rangle \leq 10^2, \quad f_c \geq 0.02, \quad D_{\mathrm{KL}} \leq 15, \text{ and } H_{\mathrm{core}} \geq 0.5.$$

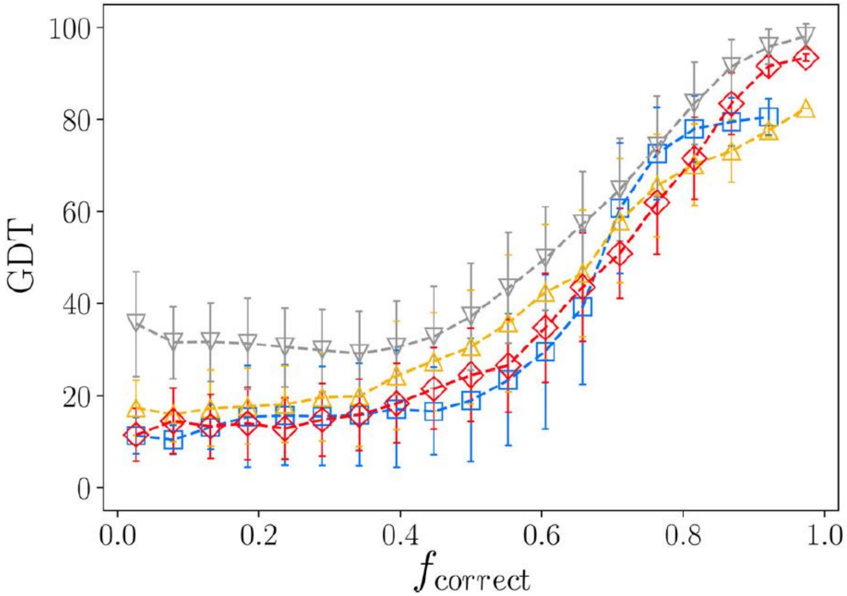


FIGURE 4 The average Global Distance Test (GDT) of Critical Assessment of protein Structure Prediction (CASP) predictions that correctly identify each given fraction of near core residues with r*SASA* ≤ 10^{-1}, f_{correct}, for CASP11 (blue squares), CASP12 (orange triangles), CASP13 (red diamonds), and 3DRobot (gray inverted triangles) structures. Error bars represent one *SD*

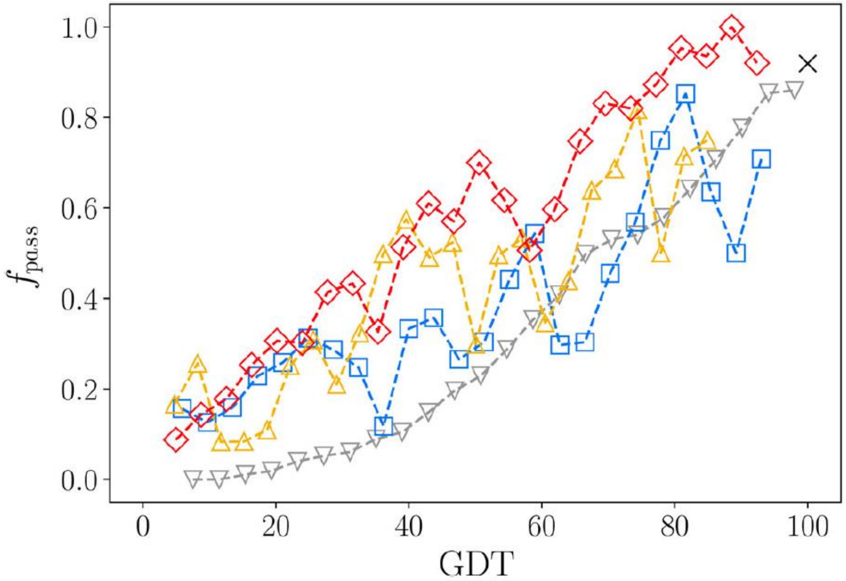


FIGURE 5 Fraction of decoys f_{pass} in a Global Distance Test (GDT) bin that are within the cutoffs for the X-ray crystal structure packing features for submissions to CASP11 (blue squares), CASP12 (orange triangles), and CASP13 (red diamonds) and 3DRobot structures (gray inverted triangles). The fraction of X-ray crystal structures that fall within the packing feature cutoffs is represented as an *x* (at $f_{\mathrm{pass}} = 0.92$)

TABLE 1 Performance of all of the tested methods on the CASP13 dataset. To estimate an average error for VoroMQA, SBROD, ProQ2, and ProQ3, the predicted scores were normalized so that they ranged from 0 to 1. The AUC depends on the cutoff that defines a good versus a bad prediction. Thus, the AUC values were averaged over GDT cutoffs from 40 to 70

Method	Pearson	Spearman	Kendall tau	Avg error	AUC
Cutoffs	—	—	—	—	0.7
Core packing	0.72	0.72	0.53	15.2	0.85
Core/near-core packing	0.75	0.75	0.56	12.9	0.89
VoroMQA	0.76	0.78	0.58	17.2	0.9
SBROD	0.8	0.8	0.58	17.24	0.9
3DCNN	0.82	0.82	0.63	12	0.94
ProQ2	0.8	0.82	0.63	27.2	0.93
ProQ3	0.83	0.84	0.63	17.7	0.95

Abbreviations: AUC, area under the curve; CASP, Critical Assessment of protein Structure Prediction; GDT, Global Distance Test.

TABLE 2 Performance of all of the tested methods on the 3DRobot decoy dataset. VoroMQA, SBROD, ProQ2, and ProQ3 return scores that do not range from 0 to 1. To estimate an average error, the predicted scores were normalized so that they fall within 0 to 1. The AUC values were averaged over GDT cutoffs from 40 to 70

Method	Pearson	Spearman	Kendall tau	Avg error	AUC
Cutoffs	—	—	—	—	0.83
Core/near-core packing	0.8	0.79	0.6	13.7	0.9
VoroMQA	0.87	0.87	0.69	14.3	0.95
SBROD	0.81	0.81	0.61	17.6	0.93
3DCNN	0.93	0.93	0.77	18	0.98
ProQ2	0.76	0.78	0.58	14.8	0.91
ProQ3	0.74	0.75	0.55	15.6	0.9

Abbreviations: AUC, area under the curve; GDT, Global Distance Test.

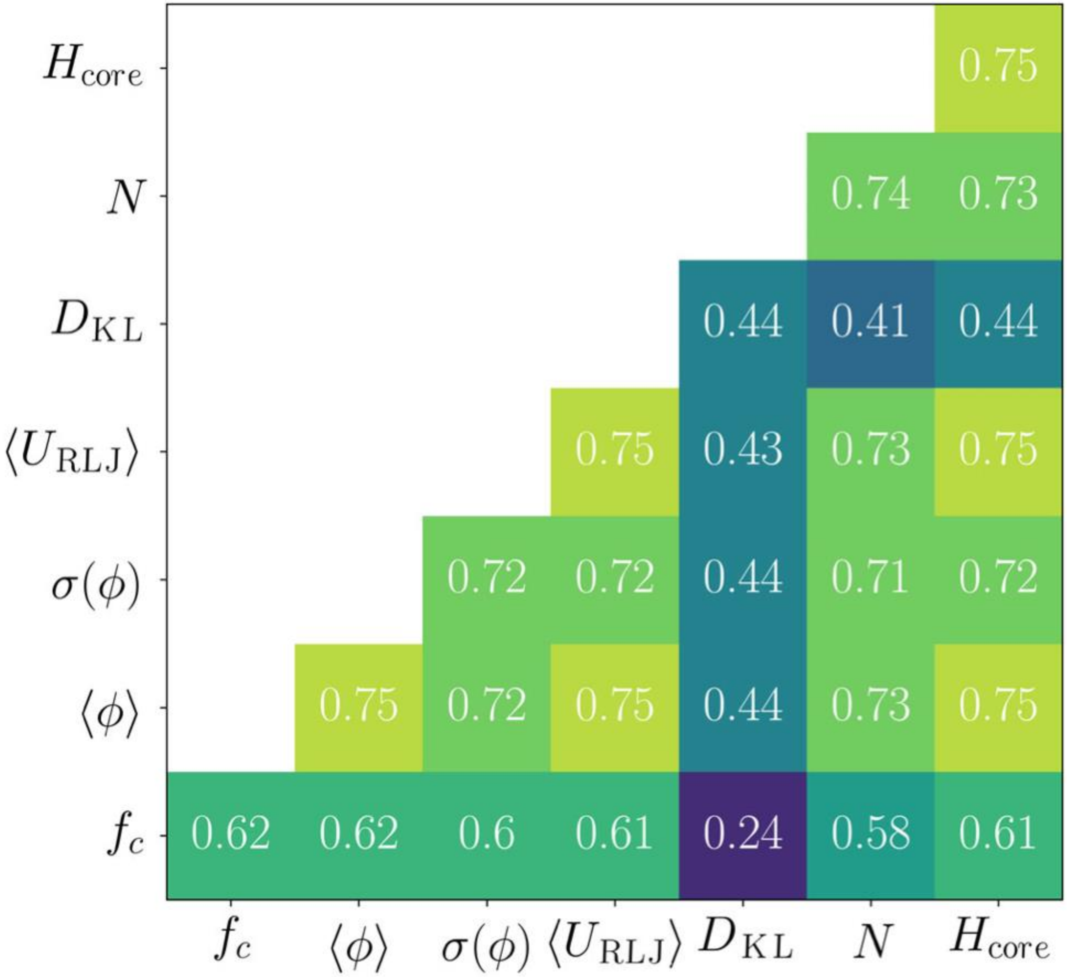


FIGURE 7 Pearson correlation coefficients between the predicted and actual Global Distance Test (GDT) of CASP13 structures following permutations of single features (along the diagonal) and pairs of features (for the off-diagonal components). The color ranges from purple (minimum) to yellow (maximum) corresponding to the values of Pearson correlation coefficient

How do we obtain 100% correlation and zero error bars?