Using physical features of protein core packing to distinguish real proteins from decoys

Alex T. Grigas[1,2], **Zhe Mei**[2,3], **John D. Treado**[2,4], **Zachary A. Levine**[5,6], **Lynne Regan**[7], **and Corey S. O'Hern**[1,2,4,8,9]

[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 Chemistry, Yale University, New Haven, Connecticut 06520, USA; [4]Department of Mechanical Engineering and Materials Science, Yale University, New Haven, Connecticut 06520, USA; [5]Department of Pathology, Yale University, New Haven, Connecticut 06520, USA; [6]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, 06520; [7]Institute of Quantitative Biology, Biochemistry and Biotechnology, Centre for Synthetic and Systems Biology, School of Biological Sciences, University of Edinburgh; [8]Department of Physics, Yale University, New Haven, Connecticut 06520, USA; [9]Department of Applied Physics, Yale University, New Haven, Connecticut 06520, USA

This manuscript was compiled on January 3, 2020

The ability to consistently distinguish real protein structures from computationally generated model decoys is not yet a solved problem. One route to distinguish real protein structures from decoys is to delineate the important physical features that specify a real protein. For example, it has long been appreciated that the hydrophobic cores of proteins contribute significantly to their stability. As a dataset of decoys to compare with real protein structures, we studied submissions to the bi-annual CASP competition (specifically CASP11, 12, and 13), in which researchers attempt to predict the structure of a protein only knowing its amino acid sequence. Our analysis reveals that many of the submissions possess cores that do not recapitulate the features that define real proteins. In particular, the model structures appear more densely packed (because of energetically unfavorable atomic overlaps), contain too few residues in the core, and have improper distributions of hydrophobic residues throughout the structure. Based on these observations, we developed a deep learning method, which incorporates key physical features of protein cores, to predict how well a computational model recapitulates the real protein structure without knowledge of the structure of the target sequence. By identifying the important features of protein structure, our method is able to rank decoys from the CASP competitions equally well, if not better than, state-of-the-art methods that incorporate many additional features.

protein decoys | hydrophobic core | protein structure prediction | protein design

It remains a grand challenge of biology to design proteins that adopt user-specified structures and perform user-specified functions. Although there have been significant successes (1–11), the field is still not at the point where we can robustly achieve this goal for any application (12). An inherent problem in protein structure prediction and design is that it is extremely difficult to distinguish between computational models that are apparently low energy (13), but which are different from the real, experimentally determined structures (14–16). This problem is known as "Decoy Detection". For example, in recent Critical Assessment of protein Structure Prediction (CASP) competitions, in which researchers attempt to predict the three-dimensional (3D) structure of a protein, based on its amino acid sequence, many groups produced impressively accurate predictions for certain targets (Fig. 1 (A)). However, for most targets there is a wide spread of prediction accuracy across the submissions from different groups. (Note that the fluctuations in prediction accuracy across groups is comparable to fluctuations within a single group. See Supplementary Information (SI).)

In recognition of this issue, there is a subcategory in CASP, Estimation of Model Accuracy (EMA), in which researchers aim to rank order the submitted models according to their similarity to the backbone of the target structure. The challenge is that researchers must develop such a scoring function for determining model accuracy, yet they do not have access to the target structure (17–23). Although EMA methods are improving (24–34), they are still unable to consistently rank models submitted to CASP in terms of their similarity to the target structure (23).

The protein core has long been known to determine protein stability and provide the driving force for folding (35–43). Additionally, in our previous work, we have found that several features of core packing are universal among well-folded experimental structures, such as the repacking predictability of core residue side chain placement, core packing fraction, and distribution of core void space (44–49). This work suggests that analysis of core residue placement and packing in proteins more generally should be a powerful tool for determining the accuracy of protein decoys. Indeed, the RosettaHoles software uses defects in interior void space to differentiate between high-resolution x-ray crystal structures and protein decoys (50). Nevertheless, a minimal set of features that can determine protein decoy accuracy has not yet been identified.

We demonstrate, that for recent CASP competition predictions, we can determine protein decoy accuracy solely by

Significance Statement

A common problem in both the prediction of a protein's three-dimensional (3D) structure from its amino acid sequence, and also in the design of sequences that will adopt a desired 3D structure, is that one can create low-energy computational models that are wrong. Either the predicted structure does not match the experimentally determined structure, or the designed sequence does not adopt the desired fold. Here, we identify features that differentiate real, experimentally determined protein structures from low-energy, but incorrect, model structures. We subsequently use these features, which focus on packing constraints, to develop a deep learning model, which is able to distinguish real, experimentally determined protein structures from computationally generated structures that are not correct.

A.T.G. compiled the datasets and carried out the computations. A.T.G, Z.M., J.D.T., Z.A.L., L.R., and C.S.O. designed the research and contributed to the analysis of the results. A.T.G., J.D.T., C.S.O. and L.R. wrote the article.

The authors declare no competing interests.

[1] To whom correspondence should be addressed. E-mail: corey.ohern@yale.edu

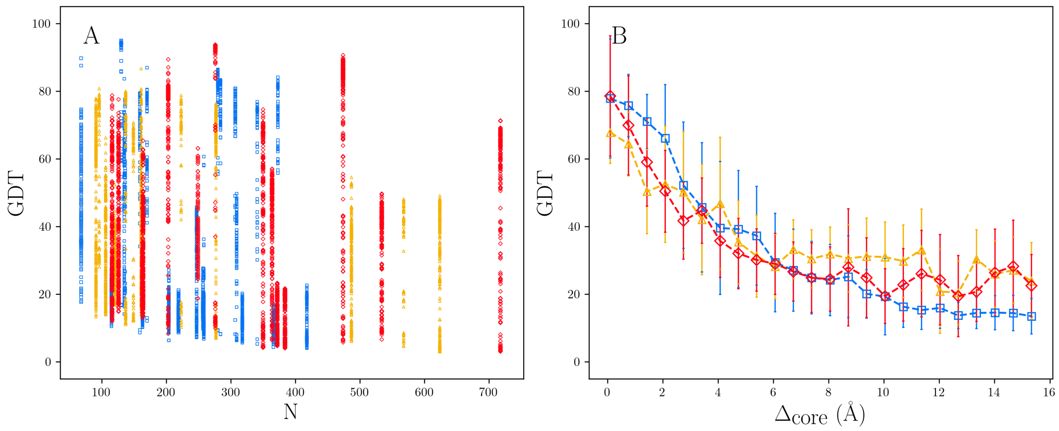

Fig. 1. (A) Scatter plot of the Global Distance Test (GDT) score, which gives the average percentage of C_α atoms that is within a given cutoff distance to the target (averaged over four cutoff distances), versus the number of residues N in the target structure for free modeling submissions to CASP11 (blue squares), CASP12 (orange triangles), and CASP13 (red diamonds). (B) GDT plotted versus the root-mean-square deviations (RMSD) among C$_\alpha$ atoms of core residues defined in the target ($\Delta_{\rm core}$). The symbols represent the average in each $\Delta_{\rm core}$ bin and the error bars represent one standard deviation.

identifying the structures that place the correct residues in the protein core. We also show that only predicted structures that place core residues accurately, measured using the root-mean-squared deviation of the C$_\alpha$ atoms of solvent inaccessible residues (i.e. $\Delta_{\rm core} < 1$Å), can achieve high Global Distance Test (GDT) scores (GDT $\gtrsim 70$) (Fig. 1 (B)), where GDT ranges from 0 to 100 and 100 is a perfect match to the target structure (51). Motivated by these observations, we then analyzed several important attributes of the *cores* of both experimentally-observed and predicted protein structures. Using these results, we developed a decoy detection method based on only five principal features of protein packing that are independent of the target structure. Our method is more effective than many of the methods in the CASP13 EMA. Moreover, all of the methods used in CASP13 EMA employ a far greater number of features than we do (52). For example, in contrast to our approach, the top performing method in the CASP13 EMA, ModFOLD7 (23, 52), uses a neural network to combine 21 scoring metrics, each based on numerous starting features, to reach a "consensus" GDT. The effectiveness of the small number of features in our approach highlights the importance of core residues, which take up $\lesssim 10\%$ of globular proteins on average, and packing constraints in determining the global structure of proteins.

1. Results

First, we identify several key features that distinguish high-resolution x-ray crystal structures and computationally-generated decoys, such as the average core packing fraction, core overlap energy, fraction of residues positioned in the core, and the distribution of the packing fraction of hydrophobic residues throughout the protein. We then show how these features can be used to predict the GDT of CASP submissions, independent of knowing the target structure.

The distribution of packing fractions ϕ of core residues in proteins whose structures are determined by x-ray crystallography occur over a relatively narrow range, with a mean of 0.55 and a standard deviation of 0.1 (44, 46, 49). We define core residues as those with small values of the relative solvent accessible surface area, rSASA $< 10^{-3}$. (See the Materials and Methods section for a description of the database of high-resolution protein x-ray crystal structures and definition of rSASA.) In contrast, we find that many of the CASP submissions possess core residues with packing fractions that are much higher than those in experimentally determined proteins structures. One way to achieve such an un-physically high packing fraction would be to allow atomic overlaps. We therefore analyzed the side-chain overlap energy for core residues, using the purely repulsive Lennard-Jones inter-atomic potential,

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

where the sum is taken over all side-chain atoms i and all other atoms not part of the same residue j, ϵ defines the energy scale, $\sigma_{ij} = (\sigma_i + \sigma_j)/2$, σ_i is the diameter of atom i, r_{ij} is the distance between atoms i and j, and $\Theta(x)$ is the Heaviside step function, which is 1 when $x > 0$ and is 0 when $x \leq 0$. For high-resolution x-ray crystal structures, half of core residues have an overlap energy of zero; the remaining half of the residues have very small overlap energies with an average value of $U_{\rm RLJ}/\epsilon \approx 10^{-4}$ (Figs. 2 (A) and (B)). In contrast, the models in the CASP datasets include some extremely high energy residues, with $U_{\rm RLJ}/\epsilon \sim 10^{16}$. The absence of data points in the lower right-hand corner of Fig. 2 (A) clearly highlights that artificially high packing fractions are only found when the overlap energy is high. In Fig. 2 (B), we show the frequency distribution of packing fractions for core residues with $U_{\rm RLJ} = 0$. The differences in peak heights reflect how much more likely it is for core residues from x-ray crystal structures of proteins to have zero overlap energy compared to those in the CASP submissions.

These results demonstrate that individual core residues in the computational models submitted to CASP are typically overpacked. We then asked whether core overpacking is related to the number of residues in the core relative to the number of residues in the protein. In Fig. 2 (C), we plot the probability that a structure, either computationally-generated or experimentally-determined, has a given fraction of its total

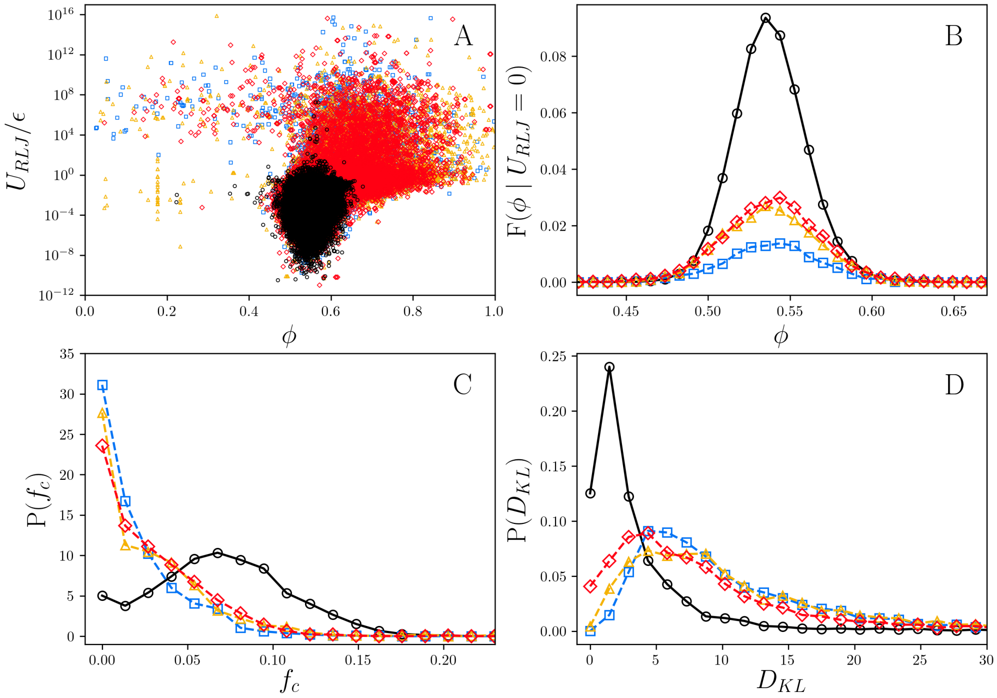

Fig. 2. Packing features of high-resolution x-ray crystal structures (black circles) and submissions to CASP11 (blue squares), CASP12 (orange triangles), and CASP13 (red diamonds). (A) Purely repulsive Lennard-Jones potential energy $U_{\rm RLJ}$ that measures the overlap of core residue sidechain atoms versus packing fraction ϕ. (B) Frequency distribution of the packing fraction $F(\phi|U_{\rm RLJ}=0)$ for core residues with zero overlap energy. (C) Probability distribution $P(f_c)$ of the fraction of core residues f_c. (D) Probability distribution $P(D_{KL})$ of the Kullback-Leibler divergence D_{KL} from the distribution of the packing fractions of all hydrophobic residues in high-resolution x-ray crystal structures.

number of residues in the core. It is clear from this plot that computationally-generated models often have too few residues in the core. Thus, the computationally-generated models not only possess cores with un-physically high packing fraction and overlap energy, but they also, typically, have a smaller fraction of residues in the core compared to x-ray crystal structures of proteins.

Many CASP models have too few residues in the core; how does this affect the distribution of hydrophobic residues outside of the core? We examined the degree to which the packing fractions of all hydrophobic residues in a given protein deviate from the expected distribution from high-resolution x-ray crystal structures (53, 54). (See Fig. 2 (D).) Specifically, we measured the Kullback-Leibler (KL) divergence (D_{KL}) between the overall distribution of packing fractions of hydrophobic residues from a database of high-resolution x-ray crystal structures, and each individual structure's packing fraction distribution for all its hydrophobic residues in that database (55). (See SI for more details.) Additionally, we measured the D_{KL} for all CASP models against the distribution from the database of high-resolution x-ray crystal structures. We find that the distribution of packing fractions of hydrophobic residues for each individual experimentally-observed protein structure is similar to the full distribution, whereas the distributions for the computationally-generated structures differ significantly from the experimentally observed distribution.

Before developing a predictive model for decoy detection, we investigated the correlation between the accuracy of backbone placement and correct identification of core residues. In Fig. 3, we plot the average GDT versus the fraction $f_{\rm core}$ of the predicted core residues that are core residues in the target structure. This plot shows that there is a strong correlation between the accuracy of backbone placement and correct identification of the core residues. In particular, when $f_{\rm core} \to 1$, the average GDT $\gtrsim 80$. However, one does not know the correct set of core residues at the time of the prediction. Yet, the core residues should share the features shown in Fig. 2. Therefore, we should be able to predict the GDT of a model based upon how well the core properties and the distribution of the hydrophobic residues match those of high-resolution x-ray crystal structures of proteins.

While we have shown that many predicted structures submitted to CASP do not recapitulate the packing properties of high-resolution protein x-ray crystal structures, we have not yet made a quantitative link between differences in these properties and the overall backbone accuracy (i.e. GDT). Therefore, we developed a neural network based on the four packing-related features in Fig. 2, plus the number, N, of residues in the protein, to construct the GDT function. (We included N to account for larger fluctuations in packing properties that occur for small N.) We built a simple feed-forward neural network with five hidden layers and a combination of common non-linear activation functions. (For more details, see SI.) The mean-squared error in GDT was used as the loss function. Submissions from CASP11, CASP12, and a large database of high-resolution x-ray crystal structures (53, 54) were used as training data. The model was then tested on CASP13 submissions. The results for the predicted versus actual GDT are plotted in Fig. 4. Our model achieves a Pearson correlation of 0.72, a Spearman correlation of 0.71, a Kendall Tau of 0.51, and an average absolute error of 13 GDT. For comparison, in the most recent assessment of decoy detection (EMA 13), one of the top ranked single-ended methods, ProQ3, reported a correlation between CASP13 actual GDT and predicted GDT of 0.67 (23). Another recent study reported a maximum Pearson correlation of 0.66 for predicted versus actual GDT for several methods that tested on CASP12 structures (27). The best absolute GDT loss reported in the CASP13 EMA competition was 7 GDT and the average GDT loss across all methods was 15 (52).

We also investigated the importance of each feature in the neural network model. To do this, we randomly permuted the values of a given feature after training. This procedure decorrelates each structure with its feature value to effectively remove that feature from the model. In Fig. 5, we display the Pearson correlation between the predicted and actual GDT following feature permutations, averaged over 200 different random permutations. All of the features are important, although eliminating the sequence length, N, as a feature still yields a Pearson correlation of 0.65, indicating it is the least important. The two largest single feature changes come from permuting either the fraction of core residues or the KL divergence from the hydrophobic residue packing fraction distribution, leading to Pearson correlations of 0.42 and 0.39, respectively. Also, permuting both of these features together leads to the largest pair-wise drop in the Pearson correlation to ≈ 0. These results indicate that the most important pair of features to include in protein decoy detection are the fraction of core residues and packing fraction distribution of hydrophobic residues. The packing fraction and overlap energy of core residues are slightly less important features. We believe this is because including the wrong residue in the core will give rise to a low GDT (Fig. 3), even if the packing fraction and overlap energy of the misplaced residues are typical of those for core residues in high-resolution protein x-ray crystal structures.

2. Discussion

We have identified several important features characterizing protein packing that allow us to distinguish protein decoys from experimentally realizable structures. We developed a machine learning model, using deep learning on a small number of packing features, that is able to predict the GDT of CASP13 structures with high accuracy and without knowledge of the target structures. In addition to developing a highly predictive model, this work also demonstrates the importance of the core and packing constraints for protein structure prediction and points out potential improvements to current prediction methods by properly modeling protein cores. Importantly, the machine learning model we developed can be used to identify protein decoys beyond those generated by CASP. For example, molecular dynamics (MD) simulations are often used to analyze thermal fluctuations in folded proteins. To what extent do the protein conformations sampled in such MD simulations recapitulate the packing properties of experimentally observed protein structures (56)? The model developed here can be used in concert with MD simulations to filter out un-physical conformations, which will have low values of GDT, without using knowledge of the experimentally observed protein structure. Thus, such an approach can be used to improve protein structure prediction. Additionally, our model can be used to assist protein design methods by selecting designs that are more likely to be experimentally attainable.

We expect future improvements to our basic model will increase its accuracy. For example, we have shown that the identification of core residues is one of the most important aspects for determining a predicted structure's accuracy. Thus, we will also implement recurrent neural networks to predict the rSASA values for each residue (57). This model can then be concatenated with the model developed here. In addition, we will incorporate predictions of GDT into MD folding simulations to improve the accuracy of computationally-generated protein structures. In addition to appreciating the overall success of our approach, it will also be informative to study in greater depth cases where there are large deviations in GDT. For example, investigating examples of high predicted GDT, but low actual GDT (or *vice versa*) has the potential to provide key insights into native protein structures.

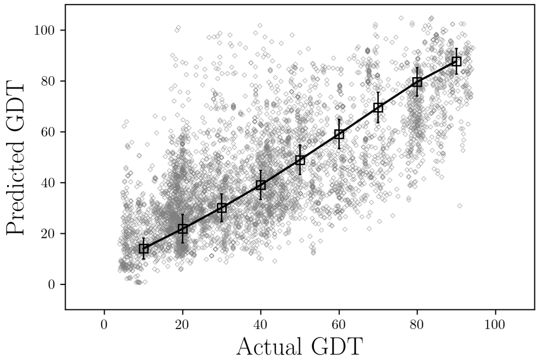

Fig. 4. Predicted versus actual GDT of CASP13 structures (gray diamonds) from a model that was developed from the four features in Fig. 2 plus N input into a neural network. The open squares represent the average value of the predicted GDT in each GDT bin and the error bars represent one standard devation.

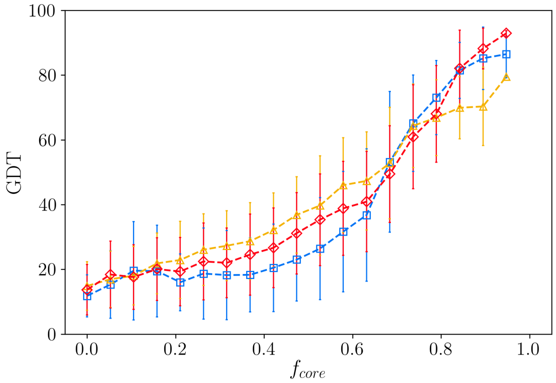

Fig. 3. The average GDT of CASP predictions that correctly identify each given fraction of near core residues with rSASA $\leq 10^{-1}$, $f_{\rm core}$, for CASP11 (blue squares), CASP12 (orange triangles), and CASP13 (red diamonds) structures. Error bars represent one standard deviation.

Materials and Methods

Datasets. In the main text, we show results for the free modeling CASP submissions, and the corresponding results for template-based modeling data are provided in the Supplementary Informa-

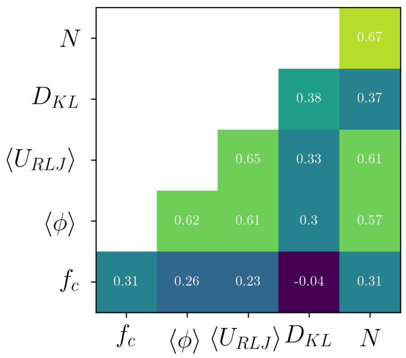

Fig. 5. Pearson correlation coefficients between the predicted and actual GDT of CASP13 structures following permutations of single features (along the diagonal) and pairs of features (for the off-diagonal components). The color ranges from purple (0) to yellow (1) corresponding to the Pearson correlation coefficient.

tion. For the decoy datasets, we examined CASP11 (2014) (58), CASP12 (2016) (59) and CASP13 (2018) (14) downloaded from the `predictioncenter.org` data archive. Each target in the competitions has a corresponding experimental structure. We selected targets with an x-ray crystal structure under a resolution cutoff. A cutoff of ≤ 2.0 Å was used in the cases of CASP11 and CASP12, however; a cutoff of ≤ 2.7 Å was used for CASP13, as very few protein targets fell under ≤ 2.0 Å . These cutoffs resulted in a dataset of 16,905 predictions based on 49 target structures. For the x-ray crystal structure dataset, we compiled a dataset of 5547 x-ray crystal structures culled from the PDB using PISCES (53, 54) with resolution ≤ 1.8 Å, a sequence identity cutoff of 20%, and an R-factor cutoff of 0.25.

rSASA. To identify core residues, we measured each residue's solvent accessible surface area (SASA). To calculate SASA, we use the NACCESS software package (60), which implements an algorithm originally proposed by Lee and Richards (61). 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 extracted from the protein structure as a dipeptide (Gly-X-Gly) with the same backbone and side-chain dihedral angles:

$$\text{rSASA} = \frac{\text{SASA}_{\text{context}}}{\text{SASA}_{\text{dipeptide}}}. \qquad [2]$$

Core residues are classified as those that have $\text{rSASA} \leq 10^{-3}$. In Fig. 3, "near core" residues are those with $\text{rSASA} \leq 10^{-1}$.

Packing Fraction. A characteristic measure of the packing efficiency of a system is the packing fraction. The packing fraction of residue μ is

$$\phi_\mu = \frac{\nu_\mu}{V_\mu}, \qquad [3]$$

where ν_μ is the non-overlapping volume and V_μ is the volume of the Voronoi cell surrounding residue μ. The Voronoi cell represents the local free space around the residue. To calculate the Voronoi tessellation for a protein structure, we use the 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 package (62). 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.

ACKNOWLEDGMENTS. The authors acknowledge support from NIH training grants T32GM008283 (A.T.G.) and T32EB019941 (J.D.T.), the Integrated Graduate Program in Physical and Engineering Biology (Z.M.), and NSF Grant No. PHY-1522467 (C.S.O.). This work also benefited from the facilities and staff of the Yale University Faculty of Arts and Sciences High Performance Computing Center.

References.

1. B Kuhlman, et al., Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).
2. GL Butterfoss, B Kuhlman, Computer-based design of novel protein structures. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 49–65 (2006).
3. H Yin, et al., Computational design of peptides that target transmembrane helices. *Science* **315**, 1817–1822 (2007).
4. L Jiang, et al., De novo computational design of retro-aldol enzymes. *Science* **319**, 1387–1391 (2008).
5. GJ Rocklin, et al., Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).
6. L Regan, W DeGrado, Characterization of a helical protein designed from first principles. *Science* **241**, 976–978 (1988).
7. JW Bryson, et al., Protein design: A hierarchic approach. *Science* **270**, 935–941 (1995).
8. CJ Lanci, et al., Computational design of a protein crystal. *Proc. Natl. Acad. Sci.* **109**, 7304–7309 (2012).
9. AR Thomson, et al., Computational design of water-soluble -helical barrels. *Science* **346**, 485–488 (2014).
10. WM Dawson, GG Rhys, DN Woolfson, Towards functional de novo designed proteins. *Curr. Opin. Chem. Biol.* **52**, 102 – 111 (2019).
11. ER Main, Y Xiong, MJ Cocco, L D'Andrea, L Regan, Design of stable -helical arrays from an idealized TPR motif. *Structure* **11**, 497 – 508 (2003).
12. D Baker, What has de novo protein design taught us about protein folding and biophysics? *Protein Sci.* **28**, 678–683 (2019).
13. Y Zhang, Protein structure prediction: When is it useful? *Curr. Opin. Struct. Biol.* **19**, 145 – 155 (2009).
14. A Kryshtafovych, T Schwede, M Topf, K Fidelis, J Moult, Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Struct. Funct. Bioinforma.* **87**, 1011–1020 (2019).
15. P Robustelli, S Piana, DE Shaw, Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci.* **115**, E4758–E4766 (2018).
16. K Lindorff-Larsen, S Piana, RO Dror, DE Shaw, How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
17. D Cozzetto, A Kryshtafovych, M Ceriani, A Tramontano, Assessment of predictions in the model quality assessment category. *Proteins: Struct. Funct. Bioinforma.* **69**, 175–183 (2007).
18. D Cozzetto, A Kryshtafovych, A Tramontano, Evaluation of CASP8 model quality predictions. *Proteins: Struct. Funct. Bioinforma.* **77**, 157–166 (2009).
19. A Kryshtafovych, K Fidelis, A Tramontano, Evaluation of model quality predictions in CASP9. *Proteins: Struct. Funct. Bioinforma.* **79**, 91–106 (2011).
20. A Kryshtafovych, et al., Assessment of the assessment: Evaluation of the model quality estimates in CASP10. *Proteins: Struct. Funct. Bioinforma.* **82**, 112–126 (2014).
21. A Kryshtafovych, et al., Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. *Proteins: Struct. Funct. Bioinforma.* **84**, 349–369 (2016).
22. A Kryshtafovych, B Monastyrskyy, K Fidelis, T Schwede, A Tramontano, Assessment of model accuracy estimations in CASP12. *Proteins: Struct. Funct. Bioinforma.* **86**, 345–360 (2018).
23. J Cheng, et al., Estimation of model accuracy in CASP13. *Proteins: Struct. Funct. Bioinforma.* **87**, 1361–1377 (2019).
24. My Shen, A Sali, Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507–2524 (2006).
25. J Zhang, Y Zhang, A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLOS ONE* **5**, 1–13 (2010).
26. M Lu, AD Dousis, J Ma, OPUS-PSP: An orientation-dependent statistical all-atom potential derived from side-chain packing. *J. Mol. Biol.* **376**, 288 – 301 (2008).
27. M Karasikov, G Pagès, S Grudinin, Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. *Bioinformatics* **35**, 2801–2808 (2018).
28. A Ray, E Lindahl, B Wallner, Improved model quality assessment using ProQ2. *BMC Bioinforma.* **13**, 224 (2012).
29. K Uziela, N Shu, B Wallner, A Elofsson, ProQ3: Improved model quality assessments using rosetta energy terms. *Sci. Reports* **6**, 33509 (2016).
30. P Benkert, M Biasini, T Schwede, Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* **27**, 343–350 (2010).
31. A Waterhouse, et al., SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
32. H Zhou, J Skolnick, GOAP: A generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys. J.* **101**, 2043 – 2052 (2011).
33. H Zhou, Y Zhou, Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **11**, 2714–2726 (2009).
34. K Olechnovič, Č Venclovas, Voromqa: Assessment of protein structure quality using interatomic contact areas. *Proteins: Struct. Funct. Bioinforma.* **85**, 1131–1145 (2017).
35. KA Dill, Dominant forces in protein folding. *Biochemistry* **29**, 7133–7155 (1990).
36. FM Richards, WA Lim, An analysis of packing in the protein folding problem. *Q. Rev. Biophys.* **26**, 423–498 (1993).

37. M Munson, L Regan, R O'Brien, JM Sturtevant, Redesigning the hydrophobic core of a four-helix-bundle protein. *Protein Sci.* **3**, 2015–2022 (1994).

38. M Munson, et al., What makes a protein a protein? Hydrophobic core designs that specify stability and structural properties. *Protein Sci.* **5**, 1584–1593 (1996).

39. MA Willis, B Bishop, L Regan, AT Brunger, Dramatic structural and thermodynamic consequences of repacking a protein's hydrophobic core. *Structure* **8**, 1319 – 1328 (2000).

40. S Dalal, S Balasubramanian, L Regan, Transmuting helices and sheets. *Fold. Des.* **2**, R71 – R79 (1997).

41. S Dalal, L Regan, Understanding the sequence determinants of conformational switching using protein design. *Protein Sci.* **9**, 1651–1659 (2000).

42. L Regan, et al., Protein design: Past, present, and future. *Pept. Sci.* **104**, 334–350 (2015).

43. FM Richards, Areas, volumes, packing, and protein structure. *Annu. Rev. Biophys. Bioeng.* **6**, 151–176 (1977).

44. JC Gaines, WW Smith, L Regan, CS O'Hern, Random close packing in protein cores. *Phys. Rev. E* **93**, 032415 (2016).

45. JD Treado, Z Mei, L Regan, CS O'Hern, Void distributions reveal structural link between jammed packings and protein cores. *Phys. Rev. E* **99**, 022416 (2019).

46. JC Gaines, et al., Comparing side chain packing in soluble proteins, protein-protein interfaces and transmembrane proteins. *Proteins: Struct. Funct. Bioinforma.* **86(5)**, 581–591 (2018).

47. D Caballero, A Virrueta, C O'Hern, L Regan, Steric interactions determine side-chain conformations in protein cores. *Protein Eng. Des. Sel.* **29**, 367–376 (2016).

48. J Gaines, et al., Collective repacking reveals that the structures of protein cores are uniquely specified by steric repulsive interactions. *Protein Eng. Des. Sel.* **30**, 387–394 (2017).

49. JC Gaines, AH Clark, L Regan, CS O'Hern, Packing in protein cores. *J. Physics: Condens. Matter* **29**, 293001 (2017).

50. W Sheffler, D Baker, Rosettaholes: Rapid assessment of protein core packing for structure prediction, refinement, design, and validation. *Protein Sci.* **18**, 229–239 (2009).

51. A Zemla, LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31**, 3370–3374 (2003).

52. J Won, M Baek, B Monastyrskyy, A Kryshtafovych, C Seok, Assessment of protein model structure accuracy estimation in CASP13: Challenges in the era of deep learning. *Proteins: Struct. Funct. Bioinforma.* **87**, 1351–1360 (2019).

53. G Wang, J Dunbrack, Roland L., PISCES: A protein sequence culling server. *Bioinformatics* **19**, 1589–1591 (2003).

54. G Wang, J Dunbrack, Roland L., PISCES: Recent improvements to a PDB sequence culling server. *Nucleic Acids Res.* **33**, W94–W98 (2005).

55. S Kullback, RA Leibler, On information and sufficiency. *The Annals Math. Stat.* **22**, 79–86 (1951).

56. Z Mei, et al., Analyses of protein cores reveal fundamental differences between solution and crystal structures. *arXiv:1907.08233* (2019).

57. R Heffernan, et al., Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning. *Sci. Reports* **5**, 11476 (2015).

58. J Moult, K Fidelis, A Kryshtafovych, T Schwede, A Tramontano, Critical assessment of methods of protein structure prediction: Progress and new directions in Round XI. *Proteins: Struct. Funct. Bioinforma.* **84**, 4–14 (2016).

59. J Moult, K Fidelis, A Kryshtafovych, T Schwede, A Tramontano, Critical assessment of methods of protein structure prediction (CASP)—Round XII. *Proteins: Struct. Funct. Bioinforma.* **86**, 7–15 (2018).

60. SJ Hubbard, JM Thornton, Naccess (1993).

61. B Lee, F Richards, The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* **55**, 379 – 400 (1971).

62. S Weis, PWA Schönhöfer, FM Schaller, M Schröter, GE Schröder-Turk, Pomelo, a tool for computing generic set Voronoi diagrams of aspherical particles of arbitrary shape. *EPJ Web Conf.* **140**, 06007 (2017).