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

Alex T. Grigas1,2, Zhe Mei2,3, John D. Treado2,4, Zachary A. Levine5,6, Lynne Regan7, and Corey S. O’Hern1,2,4,8,9

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, USA; 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.

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
identifying the structures that place the correct residues in
the protein core. We also show that only predicted struc-
tures that place core residues accurately, measured using the
root-mean-squared deviation of the Cα atoms of solvent inac-
sessible residues (i.e. Δcore < 1Å), can achieve high Global
Distance Test (GDT) scores (GDT ≥ 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 decay 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 ≤ 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 crystallog-
raphy 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 sub-
missions 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 there-
fore analyzed the side-chain overlap energy for core residues,
using the purely repulsive Lennard-Jones inter-atomic poten-
tial,

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

where the sum is taken over all side-chain atoms i and all
other atoms not part of the same residue j, \( \epsilon \) defines the
energy scale, \( \sigma_{ij} = (\sigma_i + \sigma_j)/2 \), \( \sigma_i \) 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 \( U_{RLJ}/\epsilon \approx 10^{-4} \) (Figs. 2 (A) and (B)).
In contrast, the models in the CASP datasets include some
extremely high energy residues, with \( U_{RLJ}/\epsilon \sim 10^{16} \). 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 \( U_{RLJ} = 0 \). 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 re-
lated 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. 1. (A) Scatter plot of the Global Distance Test (GDT) score, which gives the average percentage of Cα 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 Cα atoms of core residues defined in the target (Δcore). The symbols represent the average in each Δcore bin and the error bars represent one standard deviation.
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 \(f_{\text{core}}\) 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 \(f_{\text{core}} \rightarrow 1\), 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,

---

**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 \(U_{RLJ}\) that measures the overlap of core residue sidechain atoms versus packing fraction \(\phi\). (B) Frequency distribution of the packing fraction \(F(\phi|U_{RLJ} = 0)\) 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.
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 $\approx 0$. 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 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 $r$SASA 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.

3. 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 Information.
tion. For the decay datasets, we examined CASP11 (2014) (58), CASP12 (2016) (59) and CASP13 (2018) (14) downloaded from the predictioncenter.org data archive. Each target in the competition has a corresponding experimental structure. We selected targets with an x-ray crystal structure under a resolution cutoff. A cutoff of \( \leq 2.0 \text{Å} \) was used in the cases of CASP11 and CASP12, however; a cutoff of \( \leq 2.7 \text{Å} \) was used for CASP13, as very few protein targets fell under \( \leq 2.0 \text{Å} \). 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 \( \leq 1.8 \text{Å} \), a sequence identity cutoff of 20%, and an R-factor cutoff of 0.25.

\[ r_{\text{SASA}} = \frac{\text{SASA}_{\text{context}}}{\text{SASA}_{\text{dipeptide}}}. \]  

Core residues are classified as those that have \( r_{\text{SASA}} \leq 0.3 \). In Fig. 3, “near core” residues are those with \( r_{\text{SASA}} \leq 0.1 \).

**Packaging Fraction.** A characteristic measure of the packing efficiency of a system is the packaging fraction. The packing fraction of residue \( \mu \) is

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

where \( \nu_\mu \) is the non-overlapping volume and \( V_\mu \) is the volume of the Voronoi cell surrounding residue \( \mu \). 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 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 \( \sim 400 \) points per atom, or \( \sim 6000 \) 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 T32GM088283 (A.T.G.) and T23EB019941 (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.**


Grigas et al. PNAS | January 3, 2020 | vol. XXX | no. XX | 5
60. SJ Hubbard, JM Thornton, Naccess (1993).