Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

The role of local versus nonlocal physicochemical restraints in determining protein native structure

Jeffrey Skolnick and Mu Gao

The tertiary structure of a native protein is dictated by the interplay of local secondary structure propensities, hydrogen bonding, and tertiary interactions. It is argued that the space of known protein topologies covers all single domain folds and results from the compactness of the native structure and excluded volume. Protein compactness combined with the chirality of the protein's side chains also yields native-like Ramachandran plots. It is the many-body, tertiary interactions among residues that collectively select for the global structure that a particular protein sequence adopts. This explains why the recent advances in deep-learning approaches that predict protein side-chain contacts, the distance matrix between residues, and sequence alignments are successful. They succeed because they implicitly learned the many-body interactions among protein residues.

Address
Center for the Study of Systems Biology, School of Biological Sciences, Georgia Institute of Technology, 950 Atlantic Drive, NW, Atlanta, GA 30332, United States

Corresponding authors:
Skolnick, Jeffrey (skolnick@gatech.edu), Gao, Mu (mu.gao@gatech.edu)

Current Opinion in Structural Biology 2021, **68**:1–8

This review comes from a themed issue on **Sequences and topology**

Edited by **Nir Ben-Tal** and **Andrei Lupas**

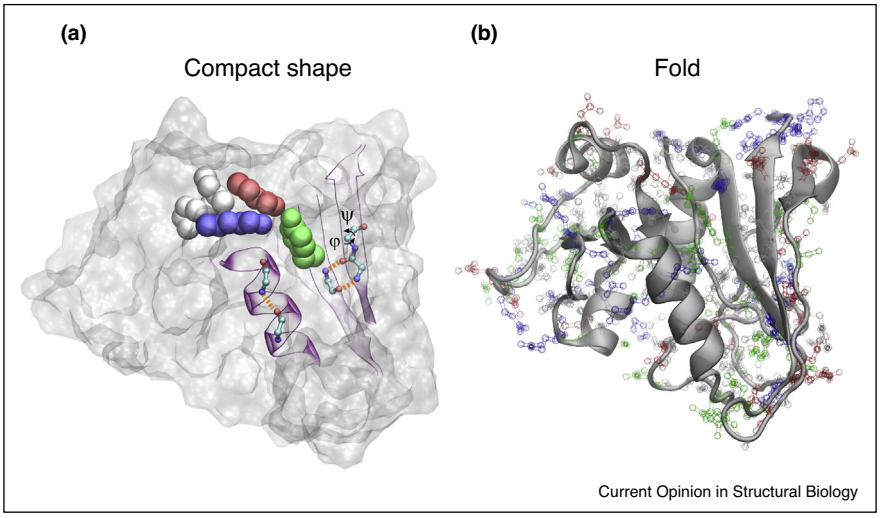
https://doi.org/10.1016/j.sbi.2020.10.008

0959-440X/© 2020 Elsevier Ltd. All rights reserved.

Determination of the native fold of a protein from its amino acid sequence is a fundamental question in biophysics. Given the validity of the Anfinsen hypothesis that the native conformation of the protein is at a global free energy minimum, the native structure is the best compromise of local conformational preferences and tertiary interactions [1]. There are three factors that determine the global fold of a protein: First and foremost is the peptide bond that provides hydrogen bonding; this is the key directional term driving protein folding [2]. One might also expect the peptide bond to provide local chain stiffness. Then, there are the amino acid's side chains which provide the sequence specificity of the protein. Amino acids are chiral, with the majority of native residues comprised of L-amino acids which have a preference for right handed α-helices and a right handed twist of β-sheets [2]. Third, native proteins structures are well packed, compact structures whose average internal packing density is comparable to crystalline solids [3]. Thus, the interplay of conformational preferences as provided by the local in sequence interactions among the side chains and backbone atoms, hydrogen bonding and chain compactness driven by tertiary interactions dictate the particular native structure a protein adopts. But to what extent do the structural properties of proteins depend solely on the intrinsic residue propensities with complete neglect of their interactions with other residues in the protein [4,5]? If this term were dominant, then local geometric and sequence specific effects would be most important. Conversely, what general features of protein structures arise from chain compactness and excluded volume with complete neglect of sequence specificity [6,7]? If compactness were to play a major role in dictating the general geometric and topological features of proteins, then perhaps, many body interactions between protein residues provide the specificity responsible for which structure a protein adopts, with local residue interactions providing the conformational bias to its secondary structure. This contribution explores these issues. Despite the intuitive view that local peptide sequence interactions dominate, as illustrated in Figure 1, we argue that it is chain compactness which explains many general protein features, most especially the space of protein folds, with local sequence specific effects providing a major, but not deterministic, driving force for secondary structure. Another key component is many-body, nonlocal in sequence interactions. While the importance of many body interactions was recognized for decades [8], the advent of deep-learning approaches enables the consideration of a wide spectrum of many-body effects that ultimately select the structure a given sequence adopts [9–12,13•,15••,16••].

At one extreme, one might imagine that every protein family adopts a unique protein fold which bears no statistically similar structural relationship to evolutionary unrelated proteins [17]. That is, the structure adopted by a given protein family is truly unique. Alternatively, due to inherent stereochemical restraints, perhaps there are a limited number of structurally distinct folds that single protein domains adopt [18,19•,20]. This is consistent with the observation of convergent folds and the fact that very diverse sequences adopt similar protein structures [21].

Figure 1



Illustrating the interactions that dictate the **(a)** compact shape and **(b)** overall structural fold of a protein sequence. In (a), hydrogen bonds (dashed orange lines) yields α-helices or β-strands (purple cartoons). Together with various protein sidechain interactions (van der Waals representation), hydrophobic (white), positive (blue), negative (red), and polar (green), lead to the compactness of a protein. The protein backbone (ball-and-stick representation, carbon (cyan), nitrogen (blue), and oxygen(red) atoms) is characterized by the φ,ψ torsion angles. In (b), many-body, long-range interactions among residues (ball-and-stick representations) give rise to the specific protein fold (grey cartoon), as demonstrated for *E. coli* dihydrofolate reductase [57]. Snapshot images were created with VMD [58].

Fold convergence could occur either because the space of single domain protein structures is inherently limited, or it could just reflect the evolutionary process by which all contemporary proteins emerged from a small set of ancestral peptides that were subsequently replicated, modified and shuffled to assemble all native folds [22–25]. Perhaps, the latter occurred but was subject to the inherent global stereochemical constraints of proteins. This would result in population of the same types of global folds independent of how they are generated. However, their relative population might be dictated by the initial primordial sets of primordial peptides. By studying quasi-spherical, random, flexible chains comprised of random sequences protein-like average composition of L-amino acids compacted into the same volume as native proteins, those protein properties that result from compaction and excluded volume were explored [26]. Remarkably, the topology of every single domain native protein (e.g. when the local chain details are ignored, e.g. a helix is replaced by a smooth curve down its principal axis), is found in the structural library of compact, quasi-spherical proteins. This also implies that the structural space of single domain proteins is likely complete. What is more striking is that these compact, quasi-spherical random proteins exhibit protein like distributions of virtual bond angles; this result was *not* built into the model. While their backbone Ramachandran (φ,ψ) plots are more diffuse than native proteins, they occupy very similar regions (Box 1). Thus, in compact structures, backbone geometry and protein side chain chirality generate the approximate native (φ,ψ) distribution without consideration of additional factors. These results suggest that the space of protein global protein folds and local geometric structures arise from protein compactness and local side chain geometry and nothing more. When artificial proteins containing native like secondary structures are folded, as expected, they also cover the space of single domain proteins [26].

Recent work on homopolymeric proteins containing equal numbers of D-residues and L-residues, demi-chiral proteins, shows that they when compacted they too contain all observed native folds, that is, the space of native folds is complete [27••]. But due to defects in chain packing caused by the formation of local secondary structures, they contain all native-like ligand-binding

Box 1 Ramachandran plot: The rotation of a protein backbone chain can be characterized by two dihedral angles called (φ,ψ), see Figure 1, left panel. The two dimensional plot of this pair of angles is known as the Ramachandran plot, which shows energetically preferred angle regions, largely dictated by steric constraints.

Free-modeling: This is the most challenging category of protein targets in the CASP. A free-modeling target does not have a homolog whose atomic structure has been experimentally determined and identified in the Protein Data Bank. Hence, it is a hard target for homology-based structure prediction approaches such as 'threading' and sequence profile methods. Predicted structures in this category are generally of lower quality than when related template structures can be identified.

pockets, including native active site geometries with L-amino acids. At what point following the origin of life was protein fold space fully populated is uncertain. Local chain stiffness is induced by intrinsic local conformational preferences and hydrogen bonding between residues. Because they have fewer internal hydrogen bonds due to their less regular secondary structure, such demi-chiral proteins are thermodynamically less stable. Selection for stability could drive selection toward the amino acid homochirality found in contemporary proteins. Thus, the space of protein folds is due to global tertiary interactions arising from the hydrophobic effect that drives compactness (Box 1).

The above results justify the conceptual approach adopted in threading where one considers a library of protein folds and then finds the structure that the given target sequence prefers [28–30]. But what dictates the preference of a given sequence for a given global structure? How much of the protein's free energy a property of the individual amino acids without consideration of other residues? Clues to the nonlocal nature of secondary structure propensities go back to the pentane effect in alkanes [31] and is reflected in the fact that the exact location of the (φ,ψ) minima of amino acids depends on the neighboring amino acids [32]. The history of secondary structure prediction provides additional insights. Despite the use of deep-learning to predict protein secondary structure, the best secondary structure methods are about 85% accurate, having improved only marginally over the past 15 years from much simpler machine learning approaches which had an accuracy of about 80% [33]. Thus, as secondary structure schemes advanced from consideration of individual residue propensities [5] to neural network models that considered local windows of residues [34] to deep-learning approaches, the accuracy of secondary structure prediction has saturated [33]. One logical explanation is that even more distant, subtle nonlocal interactions perhaps involving the entire protein sequence or a significant fraction of it dictate what secondary structure is adopted for some parts of the protein chain. Indeed, chameleon sequences (identical sequences) adopt different secondary structures depending on context [35]. The importance of nonlocal interactions has been recognized for a long time [36], but until recently the tools did not exist to fully address this issue.

Over the past several years, exciting breakthroughs have been made that include the long-range residue-residue interactions [37]. Using a deep residual convolutional neural network, Xu *et al.* were the first to demonstrate significant success in predicting (C_β–C_β) contacts between individual residues of a protein sequence in CASP12 [9,10], a blind biannual protein structure prediction competition. Such C_β–C_β contacts encode *both the local secondary structure and the global fold, and it is their synergy that could improve model quality*. Despite that, their utility in structure prediction was not shown in CASP12 [38]. Subsequently, several groups demonstrated that better contact predictions yielded significantly better structure predictions, especially for challenging targets [11,12,13•,14,16••]. In CASP13, all four top-ranked groups in the most challenging, free-modeling category (Box 1) used residue-residue contacts or distance matrices predicted via deep-learning [39]. Among them, AlphaFold achieved the best performance using high-quality residue-residue distance matrices to derive statistical folding potentials [16••].

Prior to the arrival of deep-learning algorithms, the most successful approaches for structure prediction were template-based [40] and employed HMM-based sequence profiles [41] or relied on threading [29,42]. Threading combines various evolution based sequence profiles and statistical pair potentials to generate the best alignment with experimentally known protein structures (as templates), using dynamic programming algorithms. Both approaches exploit the completeness of fold space and their scoring functions are either local in sequence as in HMM methods, consider individual pairs of interacting residues, did not consider higher order residue correlations, and in the case of threading because pair potentials are used, the best scores represent, a local and not necessarily, global optimum. Sequence profile and threading based methods fail when the best structural template is evolutionarily too remote to be detected [43].

To overcome these issues, direct co-evolutionary signals detected from the multiple sequence alignment (MSA) to infer inter-residue contacts were employed [44]. Although proposed decades ago [45], success was only possible when more sequences were available for building deep MSAs [46], and better algorithms were developed to disentangle indirect correlations from direct residue-residue contacts [47,48]. However, co-evolutionary approaches (also known as direct-coupling analysis) often yielded sparse contact predictions with many false positives [49]. These false positives occur because they do not consider the higher order correlations between many covarying residues implicit in the protein's structure.

As indicated above, it was recognized long ago that the long-range interactions in a sequence are important to folding [50]. Addressing this shortcoming is the key reason for the success of deep-learning algorithms [9–11] and enables the prediction of native-like clusters of contacts. One example is found in Figure 2 which shows the DESTINI protocol for deep learning based contact/distance matrix prediction. As a representative case depicted in Figure 2, in a systematic clustering analysis of the contacts generated by a deep-learning versus co-evolutionary analysis, most contacting clusters of residues were identified by both methods [11]. In 70% of the analyzed proteins, the difference between the two

Figure 2

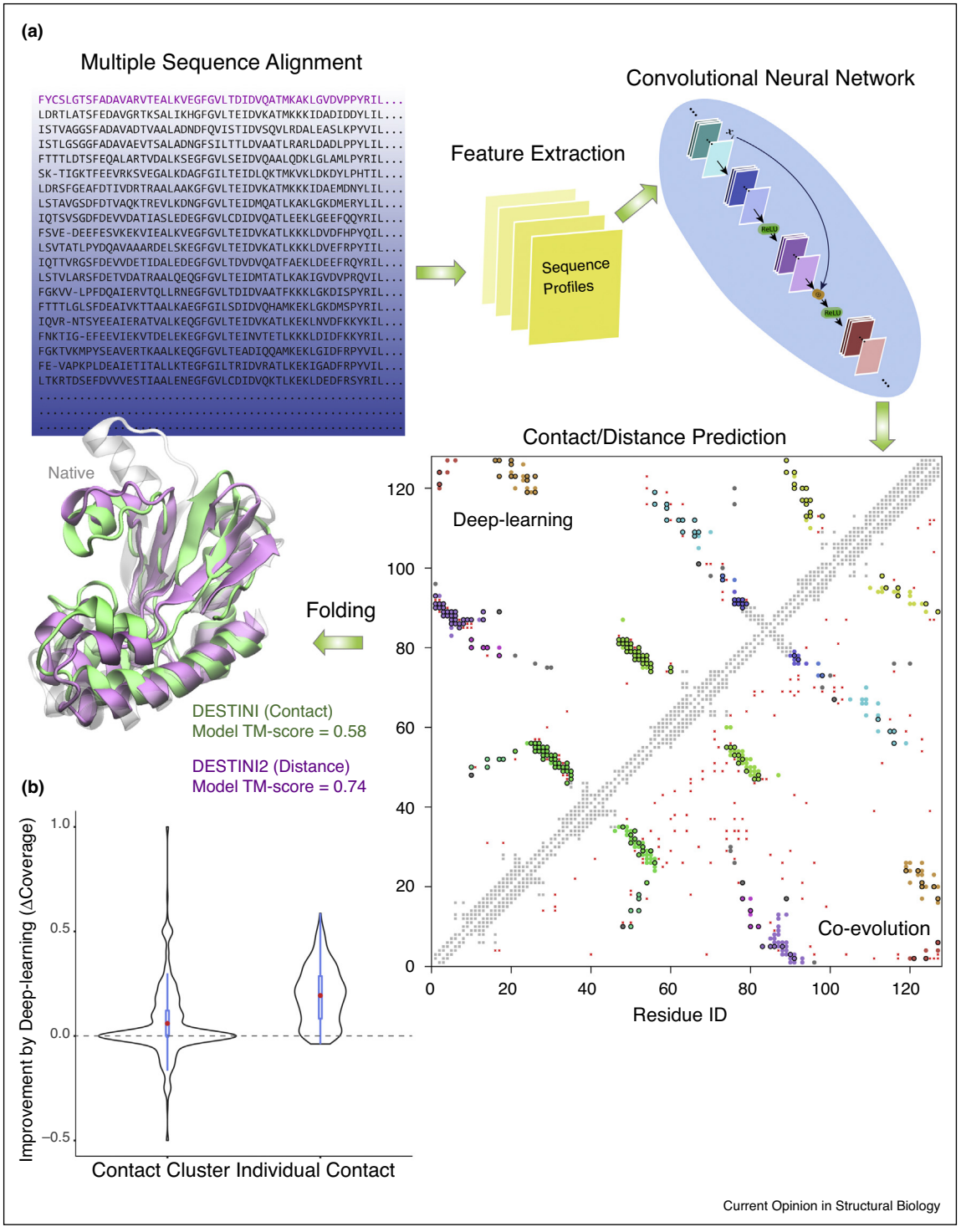


Illustration of the deep-learning algorithm DESTINI for protein structural prediction [11]. **(a)** Flowchart of DESTINI using a representative example (TT1751 of *Thermus thermophilus* HB8, PDB code 1J3M). A multiple sequence alignment is constructed for the query sequence (purple) from scanning homolog hits in a large sequence library containing many millions of sequences. Sequence profiles, both 1D and 2D (e.g. direct-coupling scores), are extracted as features fed into a deep convolutional neural network, which outputs a residue-residue contact or distance matrix. The contact map is shown for the query sequence from DESTINI [9] (upper triangle) in comparison with predictions by co-evolutionary analysis by CCMPred [59] (lower triangle). Medium and long-range native contacts are represented by circles filled in different colors to differentiate contacting clusters of side chains, and isolated sparse contacts are represented by grey circles. Correctly predicted contacts by either method are indicated by black borders surrounding the circles. False positives are represented by red dots. Other contacts are local or short range shown

approaches is less than or equal to one cluster. The success of deep-learning is mainly due to the ability to generate coherent, native-like contact patterns and eliminate spurious isolated contacts [11]. Deep-learning discovered how protein secondary structures pack with the corresponding pattern of native, nonlocal side chain interactions.

As demonstrated by DESTINI [11], having accurate residue-residue contact predictions, the chance of folding into native-like protein structure greatly improves. Then, as demonstrated in DESTINI2, generalization to residue-residue distance matrix prediction dramatically boosts model accuracy. In a challenging test set [43], as shown in Figure 3, using DESTINI2, 69% of targets have a TM-score > 0.4 versus 41% using the contact matrix of DESTINI and just 9% by the classic threading-based approach TASSER [51]. The mean model TM-score is 0.52 by DESTINI2 versus 0.39 by DESTINI. Even when there is no improvement in medium/long range (whose residue separation ≥ 12) contact predictions, an average TM-score improvement of 0.10 is observed in models built by distance predictions. Remarkably, in cases where the precision of contact prediction decreases slightly (<10%), probably due to a comprise for better overall distance prediction, an improvement of mean TM-score at 0.08 is achieved. This improvement largely comes from better assembly of secondary structure segments assisted by accurate long-range distance restraints. For over 95% of targets, 80% of secondary segments (3 states: α, β, coil) have at least one long distance restraint (>10 Å) correctly predicted. Overall, using the distance matrix for protein structure prediction is a small yet significant step forward. Although other deep-learning predictions such as torsion angles were also utilized, they seem to add very minor improvement at best [13•,16••,52], as propagation of very small (∼5°) errors in backbone angle predictions from canonical values can significantly impact the overall fold [53].

The spectacular success of deep-learning in protein structure prediction benefits from three major factors: First, because each residue-residue pair becomes a training data point, there are now enough training data points for the millions of parameters typically found in a deep neural network model. Second, one only needs a fraction of native long-range residue contacts to correctly, or approximately, fold a protein, though the exact criteria are fold-dependent. For a globular fold, it is estimated that about $L/4$ to $L/5$ accurate long-range contacts are sufficient for folding a sequence of N amino acids [11,54]. The distance matrix then acts to fine-tune this approximate structure. Lastly, since protein fold space is already complete in the current protein data bank [55], fold prediction is much easier when the answer is more or less known and trained to recognize, than in a scenario where most proteins folds are novel.

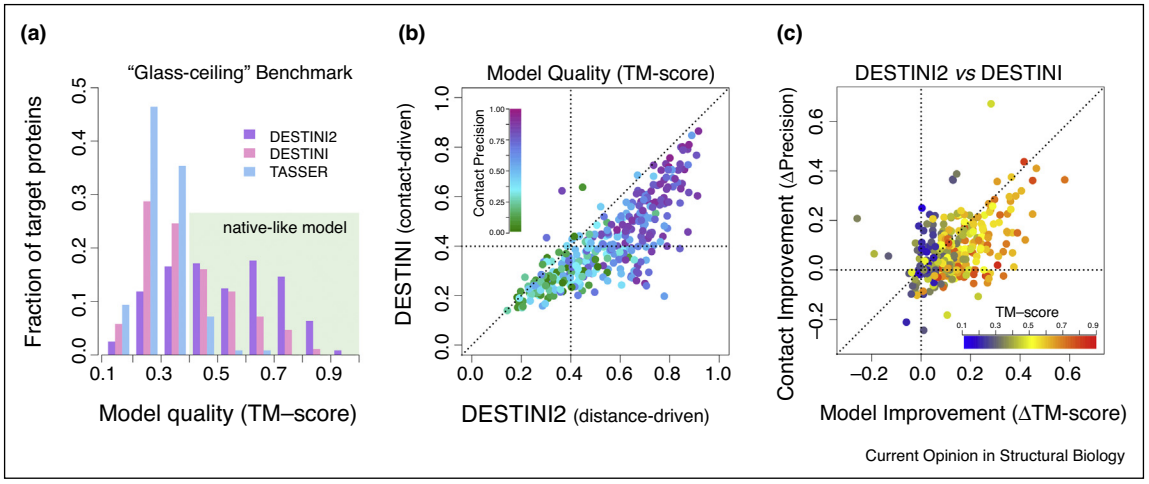
Problem solved? Not quite. As with other machine-learning algorithms, distance prediction via deep-learning is knowledge-based and is expected to perform poorly on novel sequences. To drive the inference process, the quality of the MSA is essential. When the MSA is 'shallow', for example, for a novel sequence with no known homologs (beyond the sensitivity of existing sequence alignment algorithms), the success rate dramatically decreases. For example, when the effective number of sequences in an MSA is less than 50, DESTINI2 only has a 23% chance of predicting an acceptable fold with a TM-score > 0.4 for hard targets.

The deep learning approaches described above focus on an individual protein and then predict pairs of residue contacts or the distance matrix between residues. Threading and sequence profile based methods take a complementary approach to protein structure prediction and use the structure of an existing template and then predict the best alignment of the target sequence of interest to that of the template. Can deep learning be applied to generate the best alignment between two different proteins? If so, this would not only extend the ability to recognize evolutionary distant sequences, but if successful it could further help learn the folding code. Towards this goal, we have developed SAdLSA, a sequence alignment algorithm that uses a deep convolutional neural network to learn from hundreds of thousands of structural alignments [15••]. It can detect remote structural relationships without protein structures, yet still benefits from the power of deep-learning the protein folding code. Thus, the successes of deep-learning strongly argues that while the local side chain geometry provides a bias toward the appropriate regions within the space of torsion angles, and compaction provides the approximate space of protein folds, it is the many-body, nonlocal interactions between residues that ultimately select the fold for a given protein sequence.

Given a deeper understanding of the factors that dictate protein structure from sequence, could this information be used to reverse engineer how protein structures

in light grey squares along the diagonal. Finally, structural models are folded using the contact matrix (DESTINI, green) and distance matrix (DESTINI2, purple), respectively. Both are superimposed onto the native structure (grey). **(b)** The advantage of deep-learning (DESTINI) over a co-evolutionary approach (CCMPred) is revealed by clustering analysis of the predicted contacts. Violin plots of the improvement in coverage by deep learning for contact clusters and individual clustered contacts, respectively, for 362 targets, curated from the 'glass-ceiling' set after removing NMR structural targets from the original set [11]. In each violin plot, the black contour is proportional to the probability density; the blue box inside indicates the interquartile range from 25% to 75%; the median is represented by a black bar within the box; and the whisker extends up to 1.5 times the interquartile range. The red circle is positioned at the mean value.

Figure 3



Performance analysis of the deep-learning approaches (contact-driven DESTINI and distance-driven DESTINI2), and classic threading-based TASSER on the 'glass-ceiling' set. **(a)** Histograms of TM-scores for each protein target. The shaded area represents good, native-like models. **(b)** Model quality improvement from contact-driven to distance matrix-driven. Each circle represents a target protein, color-coded according to the predicted precision of the top *L* medium/long range contact predictions. **(c)** Correlation between model quality improvement and the precision of the top *L* medium/long range contact predictions. Target points are color-coded according to the TM-score of DESTINI2 models.

evolve? On the most straightforward level, one could use existing deep learning algorithms to drive the design of novel sequences. However, this does not directly address the question as to how new proteins folds arise in nature. To address this issue, one could imagine applying deep learning tools to existing structures whose substructures have been identified as putative primordial fragments [24,56] to explore if they have stronger signals than later evolving fragments. If so, perhaps, deep learning could be used to identify strongly predicted, and possibly oldest, protein fragments. It could then be used to construct a structure based evolutionary analysis as to how existing folds emerged and perhaps suggest possible ways of morphing existing sequences to adopt different structures.

Conflict of interest statement

Nothing declared.

Acknowledgements

This work was supported in part by the Division of General Medical Sciences of the National Institute Health (NIH grant R35GM118039).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Anfinsen CB: **Principles that govern the folding of protein chains**. *Science* 1973, **181**:223-230.

2. Brändén C-I, Tooze J: *Introduction to Protein Structure*. edn 2. New York: Garland Pub.; 1999.

3. Liang J, Dill KA: **Are proteins well-packed?** *Biophys J* 2001, **81**:751-766.

4. Chou PY, Fasman GD: **Prediction of protein conformation**. *Biochemistry* 1974, **13**:222-245.

5. Chou PY, Fasman GD: **Prediction of the secondary structure of proteins from their amino acid sequence**. *Adv Enzymol Relat Areas Mol Biol* 1978, **47**:45-148.

6. Honig B, Cohen FE: **Adding backbone to protein folding: why proteins are polypeptides**. *Fold Des* 1996, **1**:R17-R20.

7. Rose GD, Fleming PJ, Banavar JR, Maritan A: **A backbone-based theory of protein folding**. *Proc Natl Acad Sci U S A* 2006, **103**:16623-16633.

8. Kolinski A, Galazka W, Skolnick J: **On the origin of the cooperativity of protein folding: implications from model simulations**. *Proteins* 1996, **26**:271-287.

9. Wang S, Sun S, Li Z, Zhang R, Xu J: **Accurate de novo prediction of protein contact map by ultra-deep learning model**. *PLOS Comput Biol* 2017, **13**:e1005324.

10. Liu Y, Palmedo P, Ye Q, Berger B, Peng J: **Enhancing evolutionary couplings with deep convolutional neural networks**. *Cell Syst* 2018, **6** 65–74.e63.

11. Gao M, Zhou H, Skolnick J: **DESTINI: a deep-learning approach to contact-driven protein structure prediction**. *Sci Rep* 2019, **9**:3514.

12. Hou J, Wu T, Cao R, Cheng J: **Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13**. *Proteins: Struct Funct Bioinformatics* 2019, **87**:1165-1178.

13. • Xu J: **Distance-based protein folding powered by deep learning**. *Proc Natl Acad Sci U S A* 2019, **116**:16856-16865
One of the first publications that introduced a novel deep-learning algorithm to protein structure prediction.

14. Zheng W, Li Y, Zhang C, Pearce R, Mortuza SM, Zhang Y: **Deep-learning contact-map guided protein structure prediction in CASP13**. *Proteins* 2019, **87**:1149-1164.

15. •• Gao M, Skolnick J: **A novel sequence alignment algorithm based on deep learning of the protein folding code**.

Bioinformatics 2020 http://dx.doi.org/10.1093/bioinformatics/btaa810
This is the first deep-learning algorithm for protein sequence alignment. For remote sequences, it may reveal hidden significant relationships previously revealed only by structural comparison.

16. •• Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A *et al.*: **Improved protein structure prediction using potentials from deep learning**. *Nature* 2020, **577**:706-710
Convincingly demonstrated that deep-learning significantly improves protein structure prediction.

17. Xie L, Bourne PE: **Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments**. *Proc Natl Acad Sci U S A* 2008, **105**:5441-5446.

18. Sillitoe I, Dawson N, Thornton J, Orengo C: **The history of the CATH structural classification of protein domains**. *Biochimie* 2015, **119**:209-217.

19. • Andreeva A, Kulesha E, Gough J, Murzin AG: **The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures**. *Nucleic Acids Res* 2020, **48**:D376-D382
This work presents a timely update of the SCOP structural database that has been extremely useful to the structural biology community.

20. Schaeffer RD, Liao Y, Cheng H, Grishin NV: **ECOD: new developments in the evolutionary classification of domains**. *Nucleic Acids Res* 2017, **45**:D296-D302.

21. Lupas AN, Ponting CP, Russell RB: **On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world?** *J Struct Biol* 2001, **134**:191-203.

22. Eck RV, Dayhoff MO: **Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences**. *Science* 1966, **152**:363-366.

23. Romero Romero ML, Rabin A, Tawfik DS: **Functional proteins from short peptides: Dayhoff's hypothesis turns 50**. *Angew Chem Int Ed Engl* 2016, **55**:15966-15971.

24. Nepomnyachiy S, Ben-Tal N, Kolodny R: **Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths**. *Proc Natl Acad Sci U S A* 2017, **114**:11703-11708.

25. Alva V, Lupas AN: **From ancestral peptides to designed proteins**. *Curr Opin Struct Biol* 2018, **48**:103-109.

26. Brylinski M, Gao M, Skolnick J: **Why not consider a spherical protein? Implications of backbone hydrogen bonding for protein structure and function**. *Phys Chem Chem Phys* 2011, **13**:17044-17055.

27. •• Skolnick J, Zhou H, Gao M: **On the possible origin of protein homochirality, structure, and biochemical function**. *Proc Natl Acad Sci U S A* 2019 http://dx.doi.org/10.1073/pnas.1908241116
This paper examined the structure and functional properties of demichiral proteins composed of equimolar mixtures of D and L amino acids. It demonstrated that the fold space of single domain demichiral proteins covers all native protein topologies and that demi-chiral proteins could engage in contemporary biochemistry with the most ancient metabolism functions most frequently generated at random.

28. Bowie JU, Luthy R, Eisenberg D: **A method to identify protein sequences that fold into a known three-dimensional structure**. *Science* 1991, **253**:164-170.

29. Zhou H, Zhou Y: **SPARKS2 and SP^3 servers in CASP6**. *Proteins (Supplement CASP issue)* **(Suppl. 7)**:2005:152-156.

30. Zheng W, Zhang C, Wuyun Q, Pearce R, Li Y, Zhang Y: **LOMETS2: improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins**. *Nucleic Acids Res* 2019, **47**:W429-W436.

31. Flory PJ, Jackson JG: *Statistical Mechanics of Chain Molecules*. Hanser Publishers; 1989.

32. Jagielska A, Skolnick J: **Origin of intrinsic 3(10)-helix versus strand stability in homopolypeptides and its implications for the accuracy of the Amber force field**. *J Comput Chem* 2007, **28**:1648-1657.

33. Torrisi M, Pollastri G, Le Q: **Deep learning methods in protein structure prediction**. *Comput Struct Biotechnol J* 2020, **18**:1301-1310.

34. Rost B: **Review: protein secondary structure prediction continues to rise**. *J Struct Biol* 2001, **134**:204-218.

35. Li W, Kinch LN, Karplus PA, Grishin NV: **ChSeq: a database of chameleon sequences**. *Protein Sci* 2015, **24**:1075-1086.

36. Go N, Taketomi H: **Respective roles of short- and long-range interactions in protein folding**. *Proc Natl Acad Sci U S A* 1978, **75**:559-563.

37. LeCun Y, Bengio Y, Hinton G: **Deep learning**. *Nature* 2015, **521**:436-444.

38. Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin A: **Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age**. *Proteins* 2018, **86(Suppl. 1)**:51-66.

39. Abriata LA, Tamò GE, Dal Peraro M: **A further leap of improvement in tertiary structure prediction in CASP13 prompts new routes for future assessments**. *Proteins: Struct Funct Bioinformatics* 2019, **87**:1100-1112.

40. Zhang Y: **Progress and challenges in protein structure prediction**. *Curr Opin Struct Biol* 2008, **18**:342-348.

41. Alva V, Nam SZ, Soding J, Lupas AN: **The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis**. *Nucleic Acids Res* 2016, **44**: W410-W415.

42. Skolnick J, Kihara D, Zhang Y: **Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm**. *Proteins: Struct Funct Bioinformatics* 2004, **56**:502-518.

43. Skolnick J, Zhou H: **Why is there a glass ceiling for threading based protein structure prediction methods?** *J Phys Chem B* 2017, **121**:3546-3554.

44. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C: **Protein 3D structure computed from evolutionary sequence variation**. *PLoS One* 2011, **6**:e28766.

45. Gobel U, Sander C, Schneider R, Valencia A: **Correlated mutations and residue contacts in proteins**. *Proteins* 1994, **18**:309-317.

46. Kamisetty H, Ovchinnikov S, Baker D: **Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era**. *Proc Natl Acad Sci U S A* 2013, **110**:15674-15679.

47. Jones DT, Buchan DW, Cozzetto D, Pontil M: **PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments**. *Bioinformatics* 2012, **28**:184-190.

48. Ekeberg M, Lovkvist C, Lan Y, Weigt M, Aurell E: **Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models**. *Phys Rev E Stat Nonlin Soft Matter Phys* 2013, **87**:012707.

49. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M: **Direct-coupling analysis of residue coevolution captures native contacts across many protein families**. *Proc Natl Acad Sci U S A* 2011, **108**:E1293-E1301.

50. Go N: **Theoretical studies of protein folding**. *Annu Rev Biophys Bioeng* 1983, **12**:183-210.

51. Zhou HY, Skolnick J: **Template-based protein structure modeling using TASSERVMT**. *Proteins: Struct Funct Bioinformatics* 2012, **80**:352-361.

52. AlQuraishi M: **End-to-end differentiable learning of protein structure**. *Cell Syst* 2019, **8** 292–301. e293.

53. Burgess A, Scheraga H: **Assessment of some problems associated with prediction of the three dimensional structure of a protein from its amino-acid sequence**. *Proc Natl Acad Sci U S A* 1975, **72**:1221-1225.

54. Aszodi A, Gradwell MJ, Taylor WR: **Global fold determination from a small number of distance restraints**. *J Mol Biol* 1995, **251**:308-326.

55. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J: **On the origin and highly likely completeness of single-domain protein structures**. *Proc Natl Acad Sci U S A* 2006, **103**:2605-2610.

56. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim BH, Grishin NV: **ECOD: an evolutionary classification of protein domains**. *PLoS Comput Biol* 2014, **10**:e1003926.

57. Cao H, Gao M, Zhou H, Skolnick J: **The crystal structure of a tetrahydrofolate-bound dihydrofolate reductase reveals the origin of slow product release**. *Commun Biol* 2018, **1**:226.

58. Humphrey W, Dalke A, Schulten K: **VMD: visual molecular dynamics**. *J Mol Graph* 1996, **14**:33-38.

59. Seemayer S, Gruber M, Soding J: **CCMpred–fast and precise prediction of protein residue-residue contacts from correlated mutations**. *Bioinformatics* 2014, **30**:3128-3130.