Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Machine learning methods for predicting protein structure from single sequences

Shaun M. Kandathil, Andy M. Lau and David T. Jones

Abstract

Recent breakthroughs in protein structure prediction have increasingly relied on the use of deep neural networks. These recent methods are notable in that they produce 3-D atomic coordinates as a direct output of the networks, a feature which presents many advantages. Although most techniques of this type make use of multiple sequence alignments as their primary input, a new wave of methods have attempted to use just single sequences as the input. We discuss the make-up and operating principles of these models, and highlight new developments in these areas, as well as areas for future development.

Addresses

Department of Computer Science, University College London, Gower Street, London, WC1E 6BT, United Kingdom

Corresponding author: Jones, David T (d.t.jones@ucl.ac.uk)

Current Opinion in Structural Biology 2023, **81**:102627

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

Edited by **Madan Babu** and **Rita Casadio**

For a complete overview see the Issue and the Editorial

Available online xxx

https://doi.org/10.1016/j.sbi.2023.102627

Introduction

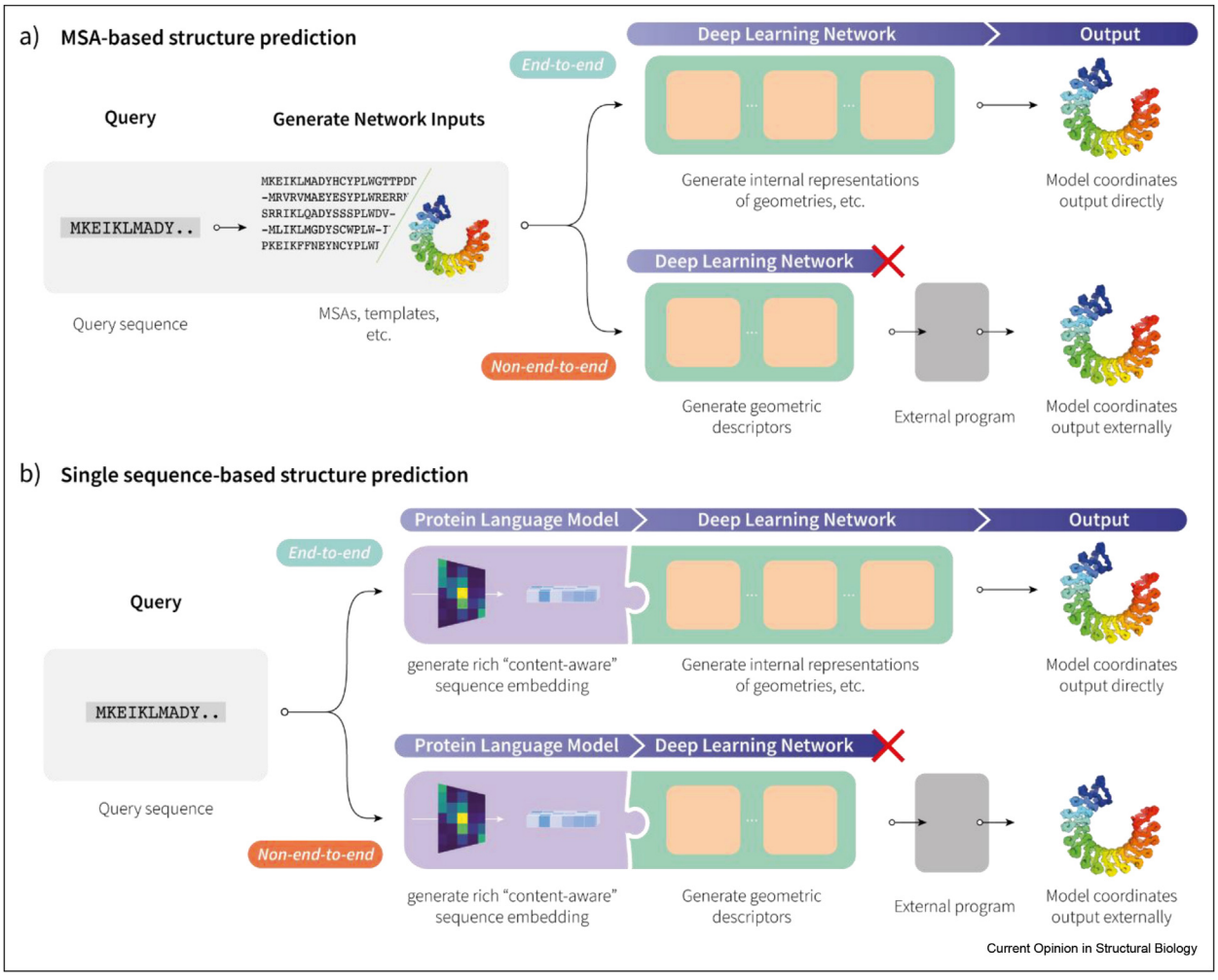
The classic problem of predicting protein structures from amino acid sequence remains unsolved, despite significant advances in recent years. These advances have mainly come about from applications of modern AI techniques, particularly deep neural networks to the problem. The most successful of these methods directly produce atomic coordinates as their outputs, known as "end-to-end" methods, which are distinct from prior methods of accurate structure prediction that first predicted inter-residue distances and then used them to 'solve' the 3D structure in a subsequent step, similar to NMR structures. Prominent examples of this two-step approach are the first version of the AlphaFold system [1,2], trRosetta [3], RaptorX [4], CONFOLD [5], and DMPfold [6,7].

In contrast to the two-step approach, end-to-end methods do not require additional modelling tools to build a model, and are thus faster and easier to run. More importantly, though, they perform better because in training the neural networks, the whole prediction process can be optimised to produce better models. Another advantage is the ability to use the trained model in ways that were not possible using the two-step approach (e.g. for protein design), where an end-to-end method can effectively be just run in reverse. Most end-to-end methods take multiple sequence alignments (MSAs) as inputs, built from the target protein sequence and homologous sequences retrieved from protein sequence data banks. Accurate structure prediction is dependent on the number and diversity of homologous sequences in the input MSA. Alternatively, a new group of methods are designed to operate only on single sequences, which will be explored in this review. Here, we take the term "single-sequence" to mean that only a single sequence is provided to the method for structure prediction, that is to say, in an entirely MSA-free manner. Figure 1 illustrates the operation of methods based on either single-sequence or MSA inputs, in both the two-step and end-to-end configurations.

Single-sequence methods based on language models

Although deep learning methods that use MSAs have clearly been successful in predicting protein structures, the need for high-quality MSAs can be problematic. Although far fewer than there were, there still exist so-called *orphan* sequences or "lineage-specific genes" with no or very few homologs in current sequence databases [8–11]. Such genes have been theorised to have originated from several sources, including evolution from non-coding sequences in specific species, or through gene neofunctionalisation [12], whereby gene duplication events give rise to novel functions. MSA-based methods typically do not produce good predictions for these sequences. More broadly, there is also the argument that one should not need to resort to using a family of related sequences in order to infer the structure of one member of that family; after all, a protein chain folding *in vivo* has no knowledge of its evolutionary history. From a computational perspective, single-sequence methods also offer advantages. MSA-based methods rely on time-consuming database searching,

Figure 1



Overview of deep learning protein structure prediction methods. Both **a**) MSA-based and **b**) single sequence-based structure prediction methods can be divided into those that produce coordinates in an end-to-end fashion, and those that do not. Single sequence-based methods replace MSA generation with pLMs which generate inputs for a deep learning network. Non-end-to-end methods rely on external programs for model generation.

and this can be a major bottleneck when trying to predict the structures of a large number of proteins.

Consequently, there is now a growing body of work starting to look at prediction methods which can work from just single-sequence inputs, and these methods have started to show promise in a number of scenarios where MSA-based prediction is either impossible or problematic (Table 1). One such scenario would be in predicting the effect of mutations; single sequence-based methods, although far from perfect, have been shown to be at least more sensitive to mutations than using MSA-based AlphaFold2 [27,29]. Improving predictions in this area may require specialised methods that are better equipped to decode variant effects from single sequences.

Recent developments in natural language processing have been rapidly exploited in computational biology, and have led to an emerging group of methods that employ language models (LMs) to learn useful protein representations. In this context, a protein language model (pLM) is a deep learning model that encodes protein sequences into rich, high-dimensional, and "content-aware" representations, which can be decoded to tackle a range of prediction tasks, including fold classification, function prediction, and more recently, in protein structure prediction and design.

Like their natural language cousins, pLMs employ an encoder-decoder architecture, trained in a self- or unsupervised manner on sequence regeneration tasks. The input sequence with some number of positions masked out or randomly mutated is first provided to the encoder, with the task of the decoder being to regenerate the original sequence, solely from the output of the encoder. This encourages the decoder to learn what amino acids to expect at a masked position from the

Table 1

Some applications where effective single-sequence prediction methods would be useful.

Application	Advantages	Reference
Fast protein structure prediction	Rapid execution enables large-scale analyses; reduces barriers to e.g. sequence similarity searching	[26,27]
Protein design	Rapid generation and evaluation of many designed sequences	[28]
Variant effect prediction	Prediction of single and multiple mutations and splice variant effects, which need to be specific to just the modified target sequence.	[27,29]
Orphan protein structure prediction	Hopefully more accurate predictions compared to MSA-based methods	[30,31]
Understanding protein biophysics	May be more representative of real-world protein energy potentials	[30]
Homology detection	Rapid assignment of novel proteins to protein families; detection of novel folds	[20,27]
Antibody modelling	More accurate antibody loop modelling (evolutionary information is not helpful here)	[21,32]

context available in the unmasked positions, while also simultaneously encouraging the encoder to produce embeddings to assist in this endeavour. Training can be done either autoregressively, that is, starting from the first position, all previous positions (none at all for the first position) are used to help predict the next, masked position, or bi-directionally, with a scheme known as masked-language modelling [13,14].

The goal of pLMs is to model the probability distribution of the training sequences, though where the outputs are conditioned on the single input sequence. This last point is important, because it means that the representation of the input sequence will model variations expected for that sequence, and this can be thought of as analogous to information we might try to extract from an MSA and represent in the form of sequence profiles or Hidden Markov Models. The embeddings produced by pLMs can be considered as alternatives to sequence profiles, but with higher order statistics (pairs, triplets, etc.) taken into account. In theory, such an embedding could model all of the information that could be found in an MSA, as long as there is sufficient data to learn from and the model is sufficiently large to represent it.

Although pLMs generally do not make explicit use of MSAs for training, the embeddings they produce will be strongly biased by the same evolutionary information, and this must be taken into account when evaluating results. The outputs of pLMs are therefore not "homology free" in the sense of true ab initio methods.

The first attempts at using LMs to learn universal representations of protein sequences were in preprint articles published by three different groups in early 2019. Two of these, UniRep [15] and SeqVec [16], were Recurrent Neural Network (RNN) pLMs trained on UniRef sequences. These studies showed that encoded latent representations of protein space had the capacity to predict multiple levels of features, such as amino acid-level physicochemistry, secondary structure, protein disorder, subcellular location, whether a protein was membrane-bound, as well as recognising the relationship between sequences in different model organisms [15,16]. Rather than using RNNs, ESM-1b [17] made use of self-attention transformer models, and this produced embeddings that could be used to produce protein contact maps, which could be offloaded to programs such as RaptorX [4] to output structure coordinates. Although not end-to-end, the methodology importantly demonstrated a way to utilise single-sequence inputs for structure prediction, bypassing the MSA generation bottleneck, as well as being potentially capable of handling orphan and designed sequences which lack homologs.

Most recent single-sequence structure prediction methods are end-to-end in terms of inference, but not in terms of training. Generally, they follow the two-step paradigm of first encoding the query sequence with a pre-trained pLM, followed by coordinate generation using either a neural network-based prediction head, or by external programs. It is interesting to speculate whether training a pLM-based method, fully end-to-end from sequence to structure, may have different properties to those trained using the two-step method, although such a training method would be limited by the number of training examples. So far, coordinate generation by end-to-end methods has typically been based on variations of AlphaFold2's Evoformer stack plus structure module, or RoseTTAFold's three-track architecture. A list of published single-sequence methods can be found in Table 2.

Single-sequence structure prediction accuracy is approaching that of state-of-the-art MSA-based methods

The state of the art in protein structure prediction is evaluated biennially in the Critical Assessment of Structure Prediction (CASP) experiments [18]. In the most recent CASP (CASP15), methods that utilised pLMs

Table 2

Single-sequence protein structure prediction methods.

Method	e2e	pLM	Trunk	Published[a]	Reference
RGN2	Y	AminoBERT	Recurrent geometric network	4 Aug 2021	[30]
trRosettaX-Single	N	s-ESM-1b	pyRosetta	18 Jan 2022	[31]
ESMFold	Y	ESM-2	AlphaFold2-like	21 Jul 2022	[26]
OmegaFold	Y	OmegaPLM	AlphaFold2-like	22 Jul 2022	[33]
HelixFold-Single	Y	No name	AlphaFold2-like	28 Jul 2022	[34]
EquiFold[b]	Y	None	SE (3) equivariant network	8 Oct 2022	[21]
MonoFold	Y	ESM-2	AlphaFold2-like	18 Oct 2022	[35]
tFold-Ab[b]	Y	ProtXLNet	AlphaFold2-like	13 Nov 2022	[32]
EMBER3D	Y	ProtT5-XL	RoseTTAFold-like	18 Nov 2022	[27]

[a] First preprint date.

[b] Designed for specific types of targets.

included EMBER3D (group 140: 'EMBER3D'), ESMFold (group 067: 'ESM-single-sequence'), and a variant of OpenFold using only single-sequence inputs (group 433: 'OpenFold-SingleSeq'). As this excludes most of the methods listed in Table 2, we conducted an in-house benchmark on the performance of each method on CASP15 targets, using the domain sequences from pre-computed alignments (generated by A. Elofsson during the prediction season; Figure 2). As baselines, we also compare against AlphaFold2 using either only single-sequence ('AlphaFold2'), or full-alignment inputs ('AlphaFold2 (MSA)'). Several methods were excluded, including EquiFold and tFold-Ab, which were trained to predict specific types of target, as well as HelixFold-Single and MonoFold, which we were not able to run.

Figure 2a shows that ESMFold and OmegaFold perform the best amongst the pLM-based methods, greatly outperforming AlphaFold2, but only when it's forced to use single-sequence inputs alone. When AlphaFold2 is used with MSA inputs, it outperforms the pLM methods significantly. Furthermore, although pLM-based methods sometimes produced models of comparable quality to AlphaFold (using MSAs), this prediction success correlates with the number of homologs available for the target (Figure 2b). This suggests that prediction quality is still limited by the size of the protein family that the pLM has seen during training. The same conclusions can be made when comparing the CASP15 performance of ESMFold and AlphaFold2 (MSA) in Figure 2c, where both methods perform similarly for targets in the Template-Based Modelling (TBM) category, but ESMFold performs markedly worse on Free-Modelling (FM) targets (which lack known structural homologs in the Protein Data Bank).

Outstanding issues in modelling the language of proteins

Suboptimal positional encoding schemes

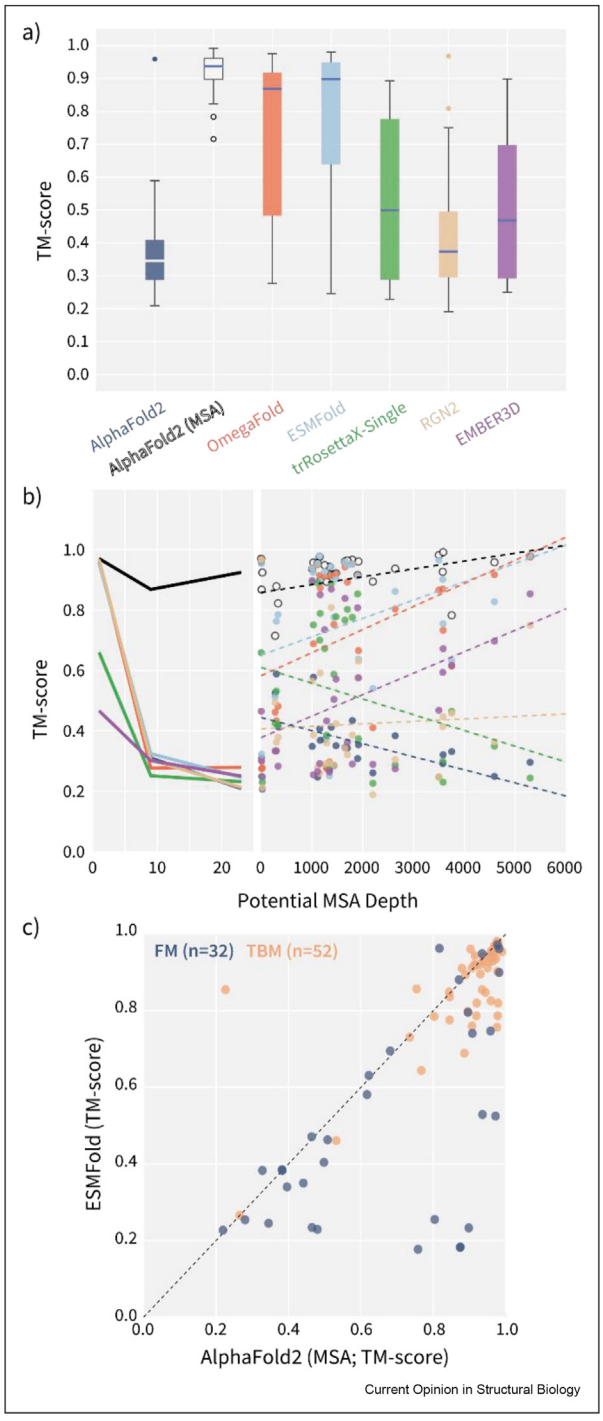
To date, most pLMs have used the same model architectures used for tasks involving human languages, most notably the Transformer [19]. At the core of the Transformer is the attention mechanism, which provides a way of learning dependencies between words or sequence positions that are separated by arbitrary distances in the sequence. A key property of "vanilla" attention is that it is permutation-invariant, that is, it produces the same result for any given sentence and any permutation of that sequence. This is of course problematic, both for natural and biological languages. Therefore, most practical implementations of Transformer models include mechanisms for positional encoding (PE), so that dependencies between tokens can be learned as a function of their absolute or relative positions in the sequence.

In pLMs, absolute PE essentially assigns residue numbers to each amino acid (token). This makes sense, but can prevent generalisation to longer or shorter sequences. This is particularly problematic for proteins, where different domain architectures can arise in related longer sequences. Relative PE can be the better choice here, however, long sequence separations are handled poorly, resulting in tandem repeats of domains being indistinguishable from one another (i.e. they all produce the exact same embedding). This might be useful in some cases, but not in many others (e.g. 3-D modelling). Ideally, a hybrid of the two styles of PE would make sense for proteins, but as yet, no well-tailored PE, dedicated to protein sequences, has been proposed. Solving this small but important aspect of pLMs could have a big positive effect on future results.

Inconsistencies in predictive accuracy across different protein families

It's clear that the accuracy of pLM-based methods varies greatly from target to target. Clearly, some learned representations are more informative than others. An obvious factor must be the underlying sizes of the protein families; we would expect proteins from large families to be predicted more accurately than those from smaller

Figure 2



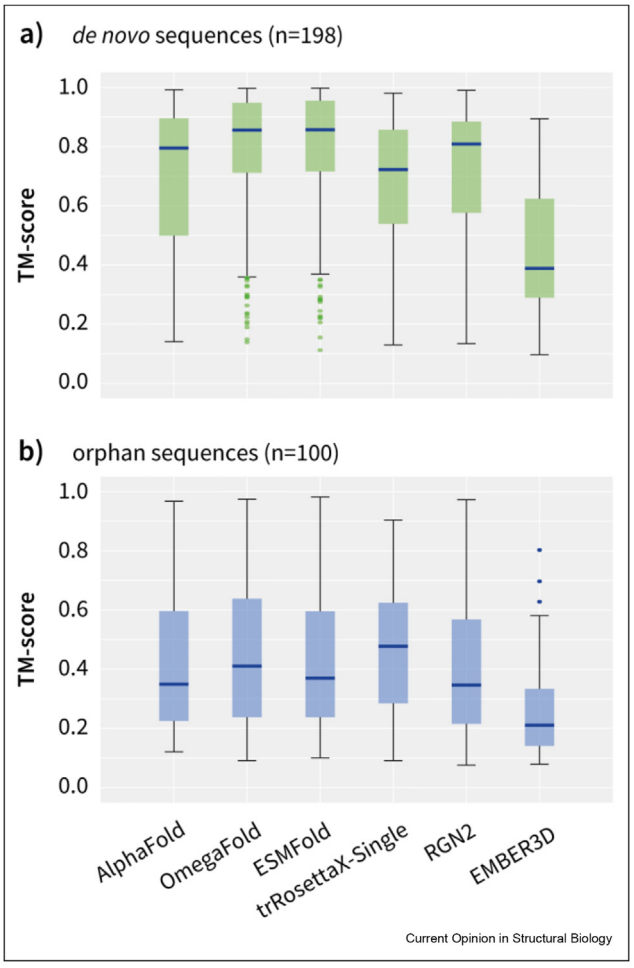
Performance of single-sequence methods and AlphaFold2 on CASP15 targets. a-b) Benchmark on 30 CASP15 targets (9 FM, 2 FM/TBM, 19 TBM), taken as those with ground truth structures released by CASP organisers and with pre-computed alignments. Targets were modelled using the domain sequences from pre-computed alignments. **b**) TM-scores for each target plot against potential MSA depth. Shallow alignments of fewer than 100 sequences have been isolated for clarity. **c**) Comparison of AlphaFold2 (MSA) against ESMFold for 84 CASP15 FM and TBM domains (TM-scores were accessed from CASP15 website). AlphaFold2 scores were taken from entry 'NBIS-AF2-standard' (group 270).

families as larger families should have produced more accurate probability distributions due to greater sampling. However, we have observed numerous deviations from this. DnyF from *Micromonospora chersina* (PDB ID 6UBL; target T1010 in CASP13), which current single-sequence methods fail to predict accurately, belongs to a medium-sized family with >200 homologs in sequence databases. For MSA-based methods, this is sufficient data to produce an excellent 3-D model. We speculate that this might be attributable to the highly scattered distribution of indels/gaps between DynF and its homologs. In our opinion, a likely reason for this failure is that Transformer-based neural networks with only simplistic positional encoding schemes are not able to deal with such variable insertions. If there were many more related sequences to learn from, though, we would expect a case like this to improve. The current position is that both pLMs and MSAs work well for large families, but for smaller families, using an MSA remains the best approach.

Similar inconsistencies in predictive performance can be seen when modelling so-called orphan sequences (sequences with few or no known homologs in protein sequence data banks) and *de novo*-designed proteins. When designing protein sequences *de novo*, it is commonly desired that the designed sequences bear no detectable evolutionary similarity to any sequence in current databases, so that the success of the design is not simply a function of its similarity to naturally occurring sequences. Recent deep-learning methods (both single-sequence and MSA-based) seem to predict structure accurately for most of these designed sequences (Figure 3a), even though the MSA-based methods usually require at least dozens or ideally hundreds of related sequences for accurate prediction. However, neither performs well for the majority of orphan proteins.

Considering the median scores in Figure 3, structures for designed sequences can be predicted much more readily than for orphan sequences in general. Why should there be such a difference between natural orphan sequences and designed sequences? The likely explanation is that the vast majority of designed sequences are designed with only one goal — to fold into a specified 3-D structure. Most designed sequences are not functional, and therefore do not need to balance the sequence constraints of folding with those of function. In other words, every residue is selected purely to optimise the sequence—structure relationship. This presumably makes the structures of most designed sequences "easy" to predict, as their sequences are overspecified in terms of what is needed to find a stable 3-D structure (and are likely hyperstable). In contrast, naturally occurring sequences will have some amino acids selected for functional reasons (e.g. a binding site), and some amino acids not selected at all, and simply present due to neutral drift. Without the overspecification of the 3-D structure

Figure 3



Benchmark on 198 *de novo* designed and 100 orphan sequences. Sequences were taken from the RGN2 and trRosettaX-Single publications. All methods including AlphaFold2 used only the target sequence as the sole input.

that most designed sequences demonstrate, natural orphan sequences remain challenging prediction targets.

So what about the 24% of orphan sequences that *are* predicted well (TM-score >0.7) by at least one of the single-sequence methods considered? Barring the possibility of inclusion of the target structures in the training data, the most likely explanation is simply that those orphan sequences are actually distant members of an existing family, but that detecting the relationship is beyond the abilities of available sequence-based homology detection methods. In that case, efficiently comparing the predicted structures to known structures (e.g. with methods such as Foldseek) [20] might help reveal such relationships. Maintaining a benchmark set of "reliable" orphan sequences, that have been run through different deep-homology tests [9,12] to determine whether they are likely to be distant relatives of existing protein families might help stimulate progress in this difficult area. Such a list would need to be regularly updated, however, as we would expect that as the sequence databases grow, fewer and fewer true orphans will remain.

At least for the sequences considered here, though, we have already tried to use the best methods to detect distant sequence relatives, so it is possible that high predictive performance could be down to a genuine ability of the pLMs to identify relationships between sequences that are beyond the ability of current homology detection algorithms. Further work is clearly needed here.

Discussion and future directions

It is evident that structure prediction tools able to use just single target sequence inputs could present a great number of advantages in computational biology, compared to MSA-based methods. Prediction based on single-sequence inputs is still, however, a nascent area of research, certainly in the area of modern AI methods, but one that is likely to very rapidly advance. As these methods become better at predicting monomers, a natural next step would be to extend the single sequence approach to predicting protein complexes.

Although the current family of single-sequence methods do show great promise, they are still dependent on their ability to leverage similarity information between sequences. We are starting to see methods that move beyond this framework and instead truly use individual sequences for prediction. EquiFold [21] produces structures from single sequences without the use of MSA or even pLM inputs. Although limited to very narrow use cases (mini-proteins and antibody structures), the study does suggest that it may be possible to do away with separate large LMs altogether, though this remains to be seen in any meaningfully general or large-scale setting. Furthermore, there are speculations that AlphaFold2 may have inadvertently learned a sufficiently accurate energy function that permits it to map co-evolutionary signals from MSAs into structures, given that it can be used to assess model quality even without an MSA input [22].

One note of caution here, however, is that for true single-sequence prediction using machine learning, we may be heavily data-limited in terms of what can be done. A recent study from our lab [10] suggested that to improve secondary structure prediction for orphan sequences, as many as 160 billion labelled protein sequences might be needed to reach the same levels as reached by MSA-based methods. Obviously, this number is highly speculative and based just on secondary structure prediction, but nonetheless, we should not forget how little experimental structural data we really have once we exclude homologs.

More broadly, current end-to-end structure prediction methods still only produce a single most-likely structure

as an end result from a single run of each method, and no information about folding pathways or thermodynamic ensembles can be gleaned from such predictions. Molecular simulations can provide these, but accurate simulations are often prohibitively expensive, especially for larger systems. Consequently, there is a growing body of work that seeks to accelerate molecular simulation using machine learning. For example, Jumper et al. [23] used a contrastive learning approach to fit the parameters of a coarse-grained force field. Differentiable molecular simulation [24,25] is another approach which works by explicitly modelling the folding process using the same mathematical and computational frameworks used to develop neural networks, which allows the force fields and other parameters to be trained to improve the end results. We speculate that these directions for research will provide more feasible routes to identifying folding intermediates, alternative conformations, and other key dynamic features of proteins, that are currently very difficult or impossible to obtain, but which are likely to be crucial for understanding function.

Funding

This research was funded in whole, or in part, by the UKRI [Grant numbers BB/T019409/1 and BB/W008556/1]. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

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

* of special interest
** of outstanding interest

1. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Zidek A, Nelson AWR, Bridgland A, *et al.*: **Improved protein structure prediction using potentials from deep learning**. *Nature* 2020, **577**:706–710.

2. ** Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, *et al.*: **Highly accurate protein structure prediction with AlphaFold**. *Nature* 2021, **596**:583–589.
Describes the AlphaFold2 method, currently accepted as the state of the art in accurate protein structure prediction. The method uses an end-to-end deep neural network approach, based on both self- and cross-attention modules, starting from multiple sequence alignment inputs. Of particular interest is the "Evo-former" module, used to process multiple sequence alignments along with linked distogram representations. Further novelty lies in the final stages which derive 3-D coordinates from the learned abstract representations.

3. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D: **Improved protein structure prediction using predicted interresidue orientations**. *Proc Natl Acad Sci U S A* 2020, **117**:1496–1503.

4. Xu J: **Distance-based protein folding powered by deep learning**. *Proc Natl Acad Sci U S A* 2019, **116**:16856–16865.

5. Adhikari B, Bhattacharya D, Cao R, Cheng J: **CONFOLD: residue-residue contact-guided ab initio protein folding**. *Proteins* 2015, **83**:1436–1449.

6. Greener JG, Kandathil SM, Jones DT: **Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints**. *Nat Commun* 2019, **10**:3977.

7. Kandathil SM, Greener JG, Lau AM, Jones DT: **Ultrafast end-to-end protein structure prediction enables high-throughput exploration of uncharacterized proteins**. *Proc Natl Acad Sci U S A* 2022:119.

8. Basile W, Salvatore M, Elofsson A: **The classification of orphans is improved by combining searches in both proteomes and genomes**. *bioRxiv* 2019, 185983.

9. Fischer D, Eisenberg D: **Finding families for genomic ORFans**. *Bioinformatics* 1999, **15**:759–762.

10. Moffat L, Jones DT: **Increasing the accuracy of single sequence prediction methods using a deep semi-supervised learning framework**. *Bioinformatics* 2021, **37**:3744–3751.

11. Tautz D, Domazet-Loso T: **The evolutionary origin of orphan genes**. *Nat Rev Genet* 2011, **12**:692–702.

12. * Weisman CM, Murray AW, Eddy SR: **Many, but not all, lineage-specific genes can be explained by homology detection failure**. *PLoS Biol* 2020, **18**, e3000862.
Describes a simple test to identify lineage-specific, or orphan, genes that may have been incorrectly identified as such due to failures in homology detection tools. The test is based on BLASTP similarity scores (bit-scores) and a simple evolutionary model. The analysis is performed on yeast and Drosophila genes and depends on gene length and evolutionary rates. The authors show that many such lineage-specific genes can be identified as being homologous to genes in other species, and that the number of truly novel genes is likely considerably less than widely reported.

13. Devlin J, Chang M-W, Lee K, Toutanova K: *BERT: pre-training of deep bidirectional transformers for language understanding*. 2018, 04805. arXiv:1810.

14. Taylor WL: **"Cloze procedure": a new tool for measuring readability**. *Journal Q* 1953, **30**:415–433.

15. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM: **Unified rational protein engineering with sequence-based deep representation learning**. *Nat Methods* 2019, **16**:1315–1322.

16. Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, Rost B: **Modeling aspects of the language of life through transfer-learning protein sequences**. *BMC Bioinf* 2019, **20**:723.

17. ** Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, *et al.*: **Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences**. *Proc Natl Acad Sci U S A* 2021:118.
Describes the ESM family of protein language models, which are useful for many downstream tasks including secondary structure and interresidue contact prediction. The models use the Transformer architecture. They take single sequences as input and are trained on a large protein sequence database. The ESM model is considerably larger than prior protein language models and appears to be more performant.

18. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J: **Critical assessment of methods of protein structure prediction (CASP)-Round XIV**. *Proteins* 2021, **89**:1607–1617.

19. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I: *Attention is all you need*. 2017, 03762. arXiv:1706.

20. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Gilchrist CLM, Söding J, Steinegger M: **Foldseek: fast and accurate protein structure search**. *bioRxiv* 2022, 479398. 2022.2002.2007.

21. * Lee JH, Yadollahpour P, Watkins A, Frey NC, Leaver-Fay A, Ra S, Cho K, Gligorijević V, Regev A, Bonneau R: **EquiFold: protein structure prediction with a novel coarse-grained structure representation**. *bioRxiv* 2023, 511322. 2022.2010.2007.

Describes a novel coarse-grained representation of protein backbones and sidechains, together with SE(3)-equivariant neural network modules to predict structures. The method is trained using the Frame Aligned Point Error loss from AlphaFold2 and is notable for its representation of sidechains throughout the model, and for being the first structure prediction tool based on geometric deep learning that does not use MSAs or pLM embeddings as inputs, though its applicability is currently limited to designed miniproteins and antibody loops.

22. Roney JP, Ovchinnikov S: **State-of-the-Art estimation of protein model accuracy using AlphaFold**. *Phys Rev Lett* 2022, **129**, 238101.

23. Jumper JM, Faruk NF, Freed KF, Sosnick TR: **Trajectory-based training enables protein simulations with accurate folding and Boltzmann ensembles in cpu-hours**. *PLoS Comput Biol* 2018, **14**, e1006578.

24. Greener JG, Jones DT: **Differentiable molecular simulation can learn all the parameters in a coarse-grained force field for proteins**. *PLoS One* 2021, **16**, e0256990.

25. Ingraham J, Riesselman A, Sander C, Marks D: **Learning protein structure with a differentiable simulator**. In *International conference on learning representations*; 2018.

26. ** Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, *et al.*: **Evolutionary-scale prediction of atomic level protein structure with a language model**. *bioRxiv* 2022, 500902. 2022.2007.2020.

Building on the ESM-1b model, this work describes ESM-2, and its application to protein structure prediction (ESMfold). The speed and accuracy of this single-sequence method makes it applicable to a set of more than 600 million proteins from the MGnify metagenomic sequence database, and the predictions are accessible via a web application. Experiments show that the method is more accurate than AlphaFold2 when using just single sequences (but not comparable to AlphaFold2 using MSAs). Scaling up the ESM-2 model (increasing its number of parameters) produces significant improvements in performance (as has been shown for natural language applications).

27. Weissenow K, Heinzinger M, Steinegger M, Rost B: **Ultra-fast protein structure prediction to capture effects of sequence variation in mutation movies**. *bioRxiv* 2022, 516473. 2022.2011.2014.

28. Ferruz N, Heinzinger M, Akdel M, Goncearenco A, Naef L, Dallago C: **From sequence to function through structure: deep learning for protein design**. *Comput Struct Biotechnol J* 2023, **21**:238–250.

29. Brandes N, Goldman G, Wang CH, Ye CJ, Ntranos V: **Genome-wide prediction of disease variants with a deep protein language model**. *bioRxiv* 2022, 505311. 2022.2008.2025.

30. * Chowdhury R, Bouatta N, Biswas S, Floristean C, Kharkar A, Roy K, Rochereau C, Ahdritz G, Zhang J, Church GM, *et al.*: **Single-sequence protein structure prediction using a language model and deep learning**. *Nat Biotechnol* 2022, **40**: 1617–1623.

Describes the second iteration of the recurrent geometric network model (RGN2) for end-to-end differentiable structure prediction starting from a single sequence. The AminoBERT language model is used to generate representations that are input to a geometry moule, which uses a translationally and rotationally invariant neural network model to derive structure. Though not as performant as MSA-based methods leveraging MSAs, the system is significantly faster and still exhibits good performance.

31. Wang W, Peng Z, Yang J: **Single-sequence protein structure prediction using supervised transformer protein language models**. *Nature Computational Science* 2022, **2**: 804–814.

32. Wu J, Wu F, Jiang B, Liu W, Zhao P: **tFold-ab: fast and accurate antibody structure prediction without squence homologs**. *bioRxiv* 2022, 515918. 2022.2011.2010.

33. * Wu R, Ding F, Wang R, Shen R, Zhang X, Luo S, Su C, Wu Z, Xie Q, Berger B, *et al.*: **High-resolution de novo structure prediction from primary sequence**. *bioRxiv* 2022, 500999. 2022.2007.2021.

Describes OmegaFold, which uses a self-supervised pLM (OmegaPLM) to derive per-residue and pairwaise embeddings that are then fed into a stack of custom neural network modules inspired by features of the AlphaFold2 "trunk." The key idea is to reinforce geometric relationships between the learned embeddings from OmegaPLM, such as the triangle inequality, with a view to improve the self-consistency and quality of the structures output by the model. The model performs comparably to AlphaFold2 in a number of scenarios, and outperforms it on orphan sequences and antibody loops.

34. Fang X, Wang F, Liu L, He J, Lin D, Xiang Y, Zhang X, Wu H, Li H, Song L: *HelixFold-single: MSA-free protein structure prediction by using protein language model as an alternative*. 2022. arXiv: 2207.13921.

35. Barrett TD, Villegas-Morcillo A, Robinson L, Gaujac B, Adméte D, Saquand E, Beguir K, Flajolet A, So ManyFolds, So Little Time: *Efficient protein structure prediction with pLMs and MSAs*. bioRxiv; 2022, 511553. 2022.2010.2015.