Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Biology

Machine learning in protein structure prediction

Mohammed AlQuraishi[1,2]

Abstract

Prediction of protein structure from sequence has been intensely studied for many decades, owing to the problem's importance and its uniquely well-defined physical and computational bases. While progress has historically ebbed and flowed, the past two years saw dramatic advances driven by the increasing "neuralization" of structure prediction pipelines, whereby computations previously based on energy models and sampling procedures are replaced by neural networks. The extraction of physical contacts from the evolutionary record; the distillation of sequence–structure patterns from known structures; the incorporation of templates from homologs in the Protein Databank; and the refinement of coarsely predicted structures into finely resolved ones have all been reformulated using neural networks. Cumulatively, this transformation has resulted in algorithms that can now predict single protein domains with a median accuracy of 2.1 Å, setting the stage for a foundational reconfiguration of the role of biomolecular modeling within the life sciences.

Addresses

[1]Program for Mathematical Genomics, Columbia University, New York, NY, USA

[2]Department of Systems Biology, Columbia University, New York, NY, USA

Corresponding author: AlQuraishi, Mohammed (m.alquraishi@columbia.edu)

Current Opinion in Chemical Biology 2021, **65**:1–8

This review comes from a themed issue on **Machine Learning in Chemical Biology**

Edited by **Connor W. Coley** and **Xiao Wang**

For a complete overview see the Issue and the Editorial

Available online 18 May 2021

https://doi.org/10.1016/j.cbpa.2021.04.005

Keywords

Protein structure prediction, Machine learning, Deep learning, Alphafold, Protein folding, Biophysics, Protein modeling, Protein design, Protein structure.

Introduction

Protein structure prediction (PSP) has long been a central problem in biochemistry, driven by the dogma that sequence determines structure and structure determines function. For much of the 2000s progress seemed stalled but that changed in the early 2010s with the proliferation of so-called coevolutionary methods [1–13]. What started as a trickle of progress accelerated over the subsequent decade and, last year, reached a crescendo with DeepMind's AlphaFold2 [14], a system that arguably solves single apo domain PSP (Figure 1).

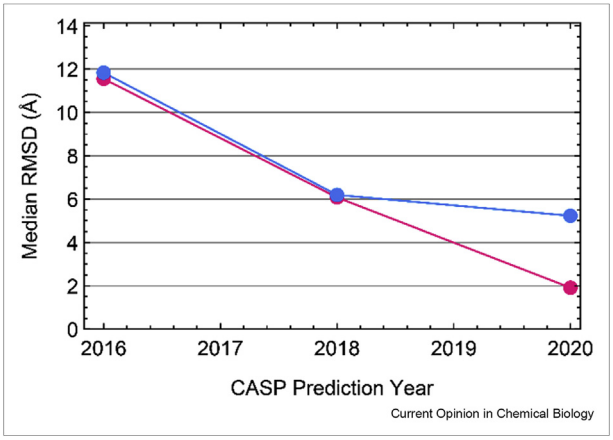
Modern PSP systems generally comprise four components: (i) an input module (Section Inputs) that takes a single protein sequence to generate additional input features, almost always including a multiple sequence alignment (MSA) of homologous proteins, (ii) a 'trunk' (Section Trunks), typically a neural network capable of sophisticated pattern recognition, which transforms features from the input module to spatial information that partially encode the 3D structure, (iii) an output module (Section Outputs) that converts this spatial information into an initial 3D structure, sometimes without explicit side-chain atoms, and (iv) a refinement module (Section Refinement) that improves the initial structure and produces all atomic coordinates.

Traditionally, these modules relied on a mixture of physics-based energy functions, knowledge-based statistical reasoning, and heuristic algorithms. The last few years however have witnessed an infusion of machine learning, particularly neural networks, into every aspect of PSP. In technical parlance, PSP systems have been neuralized, starting out with individual components and now spanning essentially the whole of the prediction process. In this review, we will examine each of these components and how they have been transformed by neuralization both individually and as an integrated whole (Figure 2). While the prediction of single protein domains is now very mature, open problems do remain, including predicting *de novo* designed proteins and capturing the impact of mutations on protein structure and stability, not to mention the broader problems of multidomain proteins and quaternary complexes. These challenges present opportunities for further innovation which we highlight in Section Future Directions.

Inputs

The basic input to a PSP system is the primary protein sequence, which in principle is all that is needed for structure prediction, at least for independently folding proteins. In practice, modern PSP systems use additional information that substantially ease the problem.

Figure 1



Accuracy of protein structure prediction. The median accuracy of the top two performing methods at CASP (Critical Assessment of protein Structure Prediction) is shown over the last four years. In CASP14 (late 2020), the AlphaFold2 system from DeepMind achieved ~2 Å accuracy over all heavy atoms for single domain apo proteins.

These include other predicted quantities such as secondary structure elements, solvent accessibility, propensity for disorder, and most importantly, one or more MSAs of homologous proteins. MSAs provide a snapshot of the evolutionary history of a protein, which in turn contains crucial information about its structure.

Markov Random Fields

The simplest use for MSAs is in constructing position-specific scoring matrices (PSSMs) that describe the observed frequencies of different amino acids at every residue position in the alignment. PSSMs thus encode which residue positions are most highly conserved and their physicochemical preferences.

PSSMs correspond to 'first-order' interactions in an MSA; *i.e.*, how each residue position behaves independently. While useful, they do not encode spatial information. It was observed as early as the 1960s [15] that coevolving residues, *i.e.*, residues with the propensity to mutate in a compensatory response to mutations in other residues (hence 'second-order' interactions), tend to be physically proximal in the protein structure—thus coevolutionary information encodes spatial information. This idea was refined over the subsequent decades and began to be seriously applied to PSP in the early 2010s [7–9,12], coinciding with the growth of genomic data. Effective second-order methods convert MSAs into Markov Random Fields (MRFs) fit to a so-called Potts Model which encodes the statistical coupling of every amino acid combination between every two residue positions. Crucially, these MRFs correct for transitive effects whereby a spurious coupling between two residues is caused by physical interactions each make with a third residue. This is key to making MRFs sufficiently accurate for PSP [16]. Initially, MRFs were directly converted into pairwise residue–residue potentials and used as geometric constraints for knowledge- or physics-based folding engines to reduce the conformational space that must be explored [13,17]. Later PSP systems used MRFs as inputs to neural networks (Section Trunks), either as summarized pairwise potentials [18,19] or as raw objects [20–22], to improve their accuracy before they are used as constraints. Combining multiple coevolutionary features, including ones derived using different sequence search parameters, also improves performance [23,24].

Raw MSAs

While MRFs capture second-order interactions, they ignore higher-order effects (*e.g.*, three-way interactions in catalytic sites) and global aspects of MSAs such as phylogeny. MSA construction is also far from flawless, often including irrelevant, evolutionarily unrelated sequences. This has spurred development of methods that operate directly on raw MSAs. Initial attempts yielded mixed results [25] but AlphaFold2 used raw MSAs to gain a qualitative leap in prediction accuracy [14]. Moreover, training neural networks to reconstruct partially masked MSAs induces them to learn aspects of protein structure despite never having been tasked to do so [26].

Structural templates

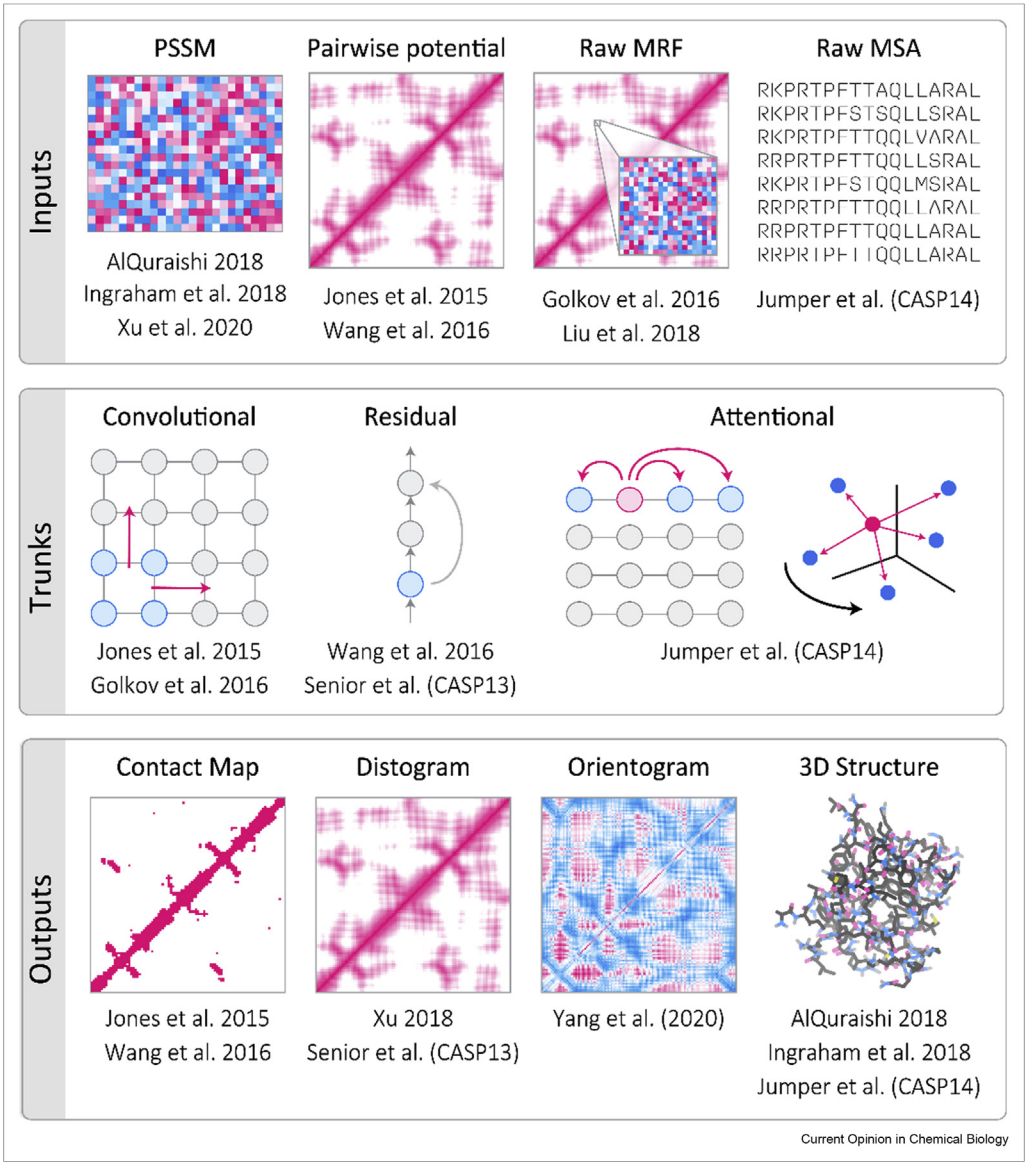
Beyond homologous sequences, homologous structures can serve directly as templates for PSP; template-based prediction was in fact the dominant approach before the advent of coevolutionary methods. Such systems traditionally use one or more templates as starting structures that are refined to the desired target sequence using energy-based conformational sampling. CASP14 saw a reemergence of template-based modeling with both AlphaFold2 and trRosetta/trRefine [27] using structural templates as raw inputs to neuralized systems. Unlike pre-neuralized systems, which have separate pipelines for template-based and *de novo* prediction—relying on human-engineered heuristics to determine which to use—the latest neuralized systems operate simultaneously on homologous sequences and structures and learn from data how best to combine the two.

Trunks

Shallow convolutional networks

Prior to the advent of neuralized PSP, the bulk of the prediction problem fell on folding engines that relied on physical models of protein folding and empirical knowledge of sequence-structure propensities ('fragment-based' methods). Rosetta [28] and I-Tasser [29] exemplified this approach. This began to change when neural networks were applied to either the summarized statistical potentials derived from MRFs [18] or directly to raw MRFs [20] to predict contact maps—binary

Figure 2



Evolution of protein structure prediction (PSP). PSP pipelines are comprised of multiple modules, including inputs, computational trunks, and outputs. Each of these modules have undergone substantial evolution in the past few years, with some of the major innovations highlighted above. See text for more details.

matrices encoding which pairs of residues are likely to be in contact (typically defined as having a Cβ-Cβ distance of less than 8 Å.) Contact maps are not fundamentally different in structure or expressive power from the MRF-derived statistical potentials used to predict them; the neural networks merely acted as denoising filters. Furthermore, initial systems used shallow (three to four layers) convolutional neural networks (CNNs) [30,31] and as a result showed modest performance gains.

Nonetheless, these early (2015) methods introduced an important idea to the field: two-dimensional MRFs can be treated as images to leverage the rapidly evolving machinery of computer vision, translating successes in deep learning to PSP. CNNs in particular relied on inductive priors that would prove useful in the short-term (but limiting in the long-term.) First, CNNs assume locality by detecting patterns in spatially localized receptive fields (in images this translates to adjacent pixels; in MRFs, to pairs of contiguous stretches of residues). This is somewhat relaxed by deep CNNs that induce wider receptive fields. Second, CNNs assume translational invariance, *i.e.*, they are indifferent to where patterns occur in the overall input. This makes sense for images; a cat is a cat irrespective of where it

appears. It sometimes makes sense for proteins as well, *e.g.*, β sheet hydrogen bonding patterns, but sometimes does not, *e.g.*, patterns occurring along the diagonal of the MRF matrix (pertaining to adjacent residues) versus patterns occurring in the off-diagonal regions (pertaining to distant protein segments.)

Deep residual networks

CNNs began to dramatically improve contact map quality with the use of residual networks (ResNets) [32], first introduced to PSP by RaptorX [19]. ResNets facilitate training of very deep CNNs with dozens to hundreds of layers, enabling wider receptive fields and greater expressive power. The resulting contact maps were sufficiently accurate to materially simplify the folding problem for existing engines such as Rosetta. In essence, RaptorX kicked off PSP neuralization by focusing on the core problem of converting coevolutionary inputs into useful geometric constraints. ResNets proved incredibly powerful and were the architecture of choice for all systems prior to 2020, including the first AlphaFold [22] and the AlphaFold-inspired trRosetta [24]. These later systems improved upon the RaptorX trunk in minor ways (bigger changes were made to the outputs—see Section Outputs) but preserved the ResNet-based core.

Attention networks

The next major breakthrough in trunk design occurred at CASP14 (2020) with the debut of AlphaFold2. As previously noted, CNNs (including ResNets) induce priors that only partly reflect protein structure, a legacy of their roots in computer vision. AlphaFold2's trunk makes liberal use of Transformers [33], a type of neural network that makes minimal inductive assumptions. Transformers rely exclusively on 'attention', a computational operation in which different parts of an input (*e.g.*, words in a sentence or residues in a protein) learn to intentionally focus on other parts of the input most relevant to the task at hand (*e.g.*, answering a question or predicting a structure). Instead of assuming that local interactions are the most important, Transformers place *a priori* equal weight on all possible interactions, then learn the relevant ones during training. This generality has enabled Transformers to dramatically improve the capabilities of natural language processing systems. Hence it is no surprise to see their impact on PSP given the importance of both short- and long-range interactions in protein folding.

Outputs

Trunks perform the heavy-lifting of detecting and homogenizing sequence-structure patterns, but these implicit patterns must ultimately be converted into explicit 3D structures. This is the purview of the output module which, within a machine learning framework, additionally plays the crucial role of defining the optimization objective used to fit all model parameters. Output modules therefore determine not only what gets predicted but how a model is trained, and both aspects have evolved significantly over the past few years.

Binary contact maps

After the coevolution revolution fully permeated the field (2016—2018), virtually all neuralized PSP systems standardized on binary contact maps as their output modality (there were key exceptions—see Section Three-Dimensional Structures and End-to-End Differentiability). This meant that all the employed neural networks, irrespective of their sophistication, were comparable in their output expressiveness to the input MRFs. Consequently, even the most advanced realizations of this approach still relied on non-neuralized physics- or knowledge-based folding engines to turn binary contact maps into folded proteins, by feeding them as geometric constraints to *e.g.*, Rosetta.

Distograms and orientograms

A new RaptorX system [34] began to change this by predicting discretized inter-residue distances (in increments of 0.5 Å bins) instead of binary contacts. These quantities were much closer to 3D structure and optimized RaptorX's neural network to hew more closely to the final desired object. RaptorX was again pathbreaking, but not sufficiently refined to yield marked improvements in PSP accuracy. A similar approach, contemporaneously developed by DeepMind and appearing under the moniker 'A7D' at CASP13 (2018), was. A7D, or AlphaFold, incorporated additional enhancements including a more powerful trunk and, importantly, converted its predicted distributions over distances ('distograms') into smooth energy potentials that could be minimized using a simple gradient descent algorithm to obtain a folded 3D structure without using a traditional folding engine. As a result, and despite being simpler than leading methods such as Rosetta, AlphaFold was the best performing system at CASP13 [35].

Elaborations of AlphaFold were rapidly developed including trRosetta [24], which augments distograms with 'orientograms' that discretize angular relationships between residues. This feature improved performance and was adopted by multiple methods at CASP14, including a new RaptorX [36].

Three-dimensional structures and end-to-end differentiability

All aforementioned methods generate outputs that are proxies of the actual desired object: the set of atomic protein coordinates. The use of proxies has two major limitations: (i) neural network outputs can only be formulated as constraints for downstream use by folding engines and, more critically, (ii) neural network trunks are optimized by a learning process that does not fully leverage the information contained in 3D structures,

including the coordination of three or more atoms, distant in the protein chain, in a small region of physical space.

Prior to CASP13 and the distance-based RaptorX/AlphaFold we and another group developed systems (RGN [37] and NEMO [38], respectively) that predicted 3D structures directly from PSSMs or primary sequences. These systems were thus 'end-to-end differentiable' as they fully neuralized the PSP pipeline from input to output using differentiable neural primitives (derivation of PSSMs from sequence databases still relied on traditional search, as is the case for raw MSAs used by AlphaFold2). RGN implicitly folded proteins using recurrent neural networks then sequentially placed the backbone atoms. NEMO folded proteins using an explicit 3D simulator that is fully neuralized and integrated within its PSP pipeline.

Neither system utilized coevolutionary information and as a result exhibited inferior performance relative to the best CASP13 methods. Nonetheless, RGN and NEMO anticipated the fully neuralized PSP vision that was ultimately realized in AlphaFold2, which predicts the 3D structure, including all side-chain atoms, in an end-to-end differentiable fashion (as discussed previously, AlphaFold2 also uses full raw MSAs as inputs, utilizing more information than even MRF-based methods.) While it remains unclear which component of AlphaFold2 is most responsible for its dramatic leap in performance, its reformulation as an end-to-end differentiable system likely played a substantial role.

Refinement

PSP systems typically produce structures that require further refinement using physics-based engines, for example to place side-chain atoms (if missing) and optimize their conformations. Refinement has resisted neuralization until recently and remains a nascent research area that will likely see significant future growth, particularly for molecular systems involving more than a single protein domain.

NEMO was first to neuralize protein refinement although echoes of the idea existed before [39]. Starting with predicted backbone coordinates, NEMO iteratively updates their positions using hundreds of steps of Langevin dynamics to minimize a protein-specific energy potential (itself learned). This approach restricts refinement to be quasi-physical which, while elegant, may limit its expressiveness and render it overly computationally demanding.

AlphaFold2 follows NEMO's lead in using neuralized refinement but does so using a Transformer that respects the translational and rotational symmetries of space. Unlike NEMO, AlphaFold2 appears to refine proteins using only a handful of iterations that can move protein coordinates in non-physical ways; details of the method remain sparse however. This is followed by limited physics-based refinement to remove any remaining steric clashes.

Future directions

The development of AlphaFold2 is undoubtedly a watershed moment in the multi-decade history of PSP, achieving near-angstrom accuracy for single apo domain prediction given sufficiently deep MSAs. Nonetheless, multiple important use cases exist in which the latest PSP systems remain uneven and untested. These include (i) MSA-free prediction from individual protein sequences, important for *de novo* designed proteins, rapidly evolving viral proteins, and evolutionarily young mammalian proteins, (ii) ultra-high accuracy prediction ($<$0.5 Å), important for drug discovery and enzymology, and (iii) predictions sensitive to minor sequence changes that lead to major structure changes, important for understanding the molecular bases of genetic diseases. The failure modes of PSP systems remain poorly characterized particularly for the last problem, as error metrics are traditionally reported over whole proteins, averaging away local but biologically important deviations.

Beyond single domains lie multidomain proteins, quaternary complexes, and protein-ligand complexes, all of which are outside the scope of current systems. We expect rapid developments in these areas and neuralization of PSP-adjacent tasks such as docking. We describe these expectations in greater detail next.

Inputs

A central PSP trend of the past decade has been the use of increasing amounts of input data, both in totality and as representations of individual proteins, culminating with raw MSAs in AlphaFold2. Yet, while all PSP components have now been neuralized, the MSA construction process itself has not, with leading methods [40–42] incorporating no machine learning. A natural direction is thus to neuralize the search for and alignment of sequence homologs.

In the opposite direction, reducing the reliance on homologous sequences would benefit many biological applications and return PSP to its single-sequence roots, emphasizing the fundamental sequence-to-structure map that underlies protein folding. Here RGN and NEMO, as well as the most recent RaptorX-based system [36], have been pushing the frontier. One exciting prospect is the use of learned protein representations that generalize across all of sequence space instead of siloing it into individual families [43–45]. Such representations transform amino acid residues within proteins into semantically rich objects, much as they have done for words in natural language sentences,

implicitly capturing many of the features that MSAs encode [46]. Single-sequence PSP based on learned representations is more likely to succeed than previous efforts, in particular because said representations leverage all available sequence data.

Trunks

Current PSP trunks largely borrow from advances in the broader deep learning field, including CNNs, ResNets, and most recently Transformers. While these architectures do reflect aspects of protein structure, including roto-translational invariance, they do not incorporate any of the known biophysics of protein folding (one exception is NEMO, which does employ a minimal physical energy model.) Judging by the advances at CASP14, this has not visibly impeded progress, but as more challenging biomolecular modeling problems are neuralized, particularly ones with more limited data (*e.g.*, quaternary complexes), we expect biophysically informed models to pay dividends.

This prediction stands in contrast with the 'Bitter Lesson', an idea put forth in an essay by the eminent computer scientist Richard Sutton and one that has proven remarkably prescient—namely that scaling up neural networks and adding more data tends to overpower, eventually, any clever architectural designs. Our contention here is that unlike in sensory-driven machine learning, where we simultaneously have limited theoretical understanding of the underlying phenomena and enjoy exponentially growing datasets, scientific problems, and in particular biomolecular problems, do not suffer from a lack of theoretical grounding but do suffer from a chronic data shortage. This makes them uniquely suited for machine learning models that integrate prior formal knowledge.

Outputs and refinement

Neuralizing 3D structure generation proved crucial to improving single domain prediction, a trend that will likely continue in multidomain proteins and protein complexes, as they fundamentally present the same static structure problem. Neuralizing the prediction of conformational ensembles and protein dynamics on the other hand presents a more fundamental challenge to current approaches. On the experimental side, far less data exist on alternate protein conformations, although cryo-electron microscopy is changing this. On the machine learning side, representing protein energy landscapes is difficult, as are the computational requirements of optimizing (*i.e.*, backpropagation) through temporal trajectories. As a result we expect this to be an exciting and rapidly evolving research area. Efforts to neuralize force fields [47–50], conformational ensembles [51], and dynamic trajectories [52–54] are emerging, including software frameworks [55]. As in static PSP, progress may come unexpectedly fast given the ongoing and dramatic developments in machine learning broadly. One thing is certain: the future of protein machine learning never looked brighter.

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..

References

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

* of special interest
** of outstanding interest

1. Thomas J, Ramakrishnan N, Bailey-Kellogg C: **Graphical models of residue coupling in protein families**. *IEEE ACM Trans Comput Biol Bioinf* 2008, **5**:183–197.
2. Dunn SD, Wahl LM, Gloor GB: **Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction**. *Bioinformatics* 2008, **24**:333–340.
3. Bartlett GJ, Taylor WR: **Using scores derived from statistical coupling analysis to distinguish correct and incorrect folds in de-novo protein structure prediction**. *Proteins* 2008, **71**: 950–959.
4. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T: **Identification of direct residue contacts in protein–protein interaction by message passing**. *Proc Natl Acad Sci Unit States Am* 2009, **106**: 67–72.
5. Balakrishnan S, Kamisetty H, Carbonell JG, Lee S-I, Langmead CJ: **Learning generative models for protein fold families**. *Proteins* 2011, **79**:1061–1078.
6. Sadowski MI, Maksimiak K, Taylor WR: **Direct correlation analysis improves fold recognition**. *Comput Biol Chem* 2011, **35**:323–332.
7. 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 USA* 2011, **108**:E1293–E1301.
8. 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.
9. Taylor WR, Jones DT, Sadowski MI: **Protein topology from predicted residue contacts**. *Protein Sci* 2012, **21**:299–305.
10. Jones DT, Buchan DWA, Cozzetto D, Pontil M: **PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments**. *Bioinformatics* 2012, **28**:184–190.
11. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E: **Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models**. *Phys Rev E* 2013, **87**, 012707.
12. Sułkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN: **Genomics-aided structure prediction**. *Proc Natl Acad Sci U S A* 2012, **109**:10340–10345.
13. 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 USA* 2013, https://doi.org/10.1073/pnas.1314045110.
14. ** John Jumper, Evans Richard, Alexander Pritzel, Green Tim, Figurnov Michael, Tunyasuvunakool Kathryn, Ronneberger Olaf, Bates Russ, Augustin Zídek, Bridgland Alex, *et al.*: *High accuracy protein structure prediction using deep learning*. 2020.
AlphaFold2, the CASP14-winning algorithm that brought single domain PSP within striking distance of experimental methods (median RMSD

of 2.1 Å). First highly performant end-to-end differentiable PSP system. First to effectively use raw MSAs and to perform iterative refinement using equivariant 3D neural networks. Currently unpublished.

15. Yanofsky C, Horn V, Thorpe D: **Protein structure relationships revealed by mutational analysis**. *Science* 1964, **146**: 1593–1594.

16. Marks DS, Hopf TA, Sander C: **Protein structure prediction from sequence variation**. *Nat Biotechnol* 2012, **30**: 1072–1080.

17. Kosciolek T, Jones DT: **De novo structure prediction of globular proteins aided by sequence variation-derived contacts**. *PloS One* 2014, **9**, e92197.

18. * Jones DT, Singh T, Kosciolek T, Tetchner S: **MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins**. *Bioinformatics* 2015, **31**:999–1006.
First application of (shallow) convolutional neural networks to predict contact maps from summarized pairwise potentials derived from co-evolution data.

19. ** 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.
First method to demonstrate dramatic improvements in contact prediction accuracy due to deep learning, specifically residual neural networks.

20. * Golkov V, Skwark MJ, Golkov A, Dosovitskiy A, Brox T, Meiler J, Cremers D: **Protein contact prediction from amino acid Co-evolution using convolutional networks for graph-valued images**. In *Annual conference on neural information processing systems (NIPS)*; 2016.
First application of convolutional neural networks to predict contact maps from raw MRFs derived from co-evolution data.

21. Liu Y, Palmedo P, Ye Q, Berger B, Peng J: **Enhancing evolutionary couplings with deep convolutional neural networks**. *cels* 2018, **6**:65–74.e3.

22. ** 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.
AlphaFold, the CASP13-winning algorithm and DeepMind's debut in the PSP field. Made effective use of distograms and a robustly-engineered neural network to substantially outperform the next best method.

23. Li Y, Zhang C, Bell EW, Yu D-J, Zhang Y: **Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13**. *Proteins: Structure, Function, and Bioinformatics* 2019, **87**:1082–1091.

24. * Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D: **Improved protein structure prediction using predicted interresidue orientations**. *Proc Natl Acad Sci Unit States Am* 2020, **117**:1496–1503.
trRosetta, a system inspired by the first AlphaFold that additionally incorporates inter-residue orientations and refinement using classic Rosetta.

25. ** Mirabello C, Wallner B: **rawMSA: end-to-end deep learning using raw multiple sequence alignments**. *PloS One* 2019, **14**, e0220182.
First method to use raw MSAs for contact prediction. Introduces idea but does not demonstrate performance gains over MRFs.

26. * Rao R, Liu J, Verkuil R, Meier J, Canny JF, Abbeel P, Sercu T, Rives A: **MSA transformer**. *bioRxiv* 2021, https://doi.org/10.1101/2021.02.12.430858.
Method for unsupervised learning of MSAs using Transformers. Architecture likely similar to the first portion of the AlphaFold2 trunk.

27. * Anishchenko Ivan, Baek Minkyung, Park Hahnbeom, Dauparas Justas, Hiranuma Naozumi, Mansoor Sanaa, Humphrey Ian, Baker David: *Protein structure prediction guided by predicted inter-residue geometries*. 2020.
First variant of the trRosetta algorithm that incorporates structural templates in a neuralized manner.

28. Leman JK, Weitzner BD, Lewis SM, Adolf-Bryfogle J, Alam N, Alford RF, Aprahamian M, Baker D, Barlow KA, Barth P, *et al.*: **Macromolecular modeling and design in Rosetta: recent methods and frameworks**. *Nat Methods* 2020, **17**:665–680.

29. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y: **The I-TASSER Suite: protein structure and function prediction**. *Nat Methods* 2015, **12**:7–8.

30. Fukushima K: **Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position**. *Biol Cybern* 1980, **36**:193–202.

31. Goodfellow I, Bengio Y, Courville A: *Deep learning*. The MIT Press; 2016.

32. He K, Zhang X, Ren S, Sun J: **Deep residual learning for image recognition**. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)*; 2016:770–778.

33. * Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I: *Attention is all you need*. 2017. arXiv:170603762 [cs].
Introduced Transformers, neural network primitives that have revolutionized natural language processing by enabling very large models (hundreds of billions of parameters) with minimal inductive biases. Transformers are a key building block of AlphaFold2.

34. ** Xu J: **Distance-based protein folding powered by deep learning**. *Proc Natl Acad Sci Unit States Am* 2019, **116**: 16856–16865.
First published method to predict distances, as opposed to binary contact maps, from co-evolutionary data.

35. AlQuraishi M: **AlphaFold at CASP13**. *Bioinformatics* 2019, **35**: 4862–4865.

36. * Xu J, Mcpartlon M, Li J: **Improved protein structure prediction by deep learning irrespective of co-evolution information**. *bioRxiv* 2020, https://doi.org/10.1101/2020.10.12.336859.
First application of a ResNet architecture to predict structures from PSSMs without co-evolutionary data, outperforming the PSSM-based RGN.

37. ** AlQuraishi M: **End-to-End differentiable learning of protein structure**. *cels* 2019, **8**:292–301.e3.
RGN, first end-to-end differentiable PSP system. Fully neuralized PSP pipeline predicted backbone atoms from PSSMs without co-evolutionary data.

38. ** Ingraham J, Riesselman A, Sander C, Marks D: **Learning protein structure with a differentiable simulator**. In *ICLR*; 2019.
NEMO, first end-to-end differentiable PSP system to incorporate iterative refinement of 3D structures, anticipating one of AlphaFold2's key innovations. Fully neuralized pipeline predicted all heavy atoms from primary sequence or PSSMs without co-evolutionary data.

39. 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**.

40. Johnson LS, Eddy SR, Portugaly E: **Hidden Markov model speed heuristic and iterative HMM search procedure**. *BMC Bioinf* 2010, **11**:431.

41. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J: **HH-suite3 for fast remote homology detection and deep protein annotation**. *BMC Bioinf* 2019, **20**:473.

42. Zhang C, Zheng W, Mortuza SM, Li Y, Zhang Y: **DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins**. *Bioinformatics* 2020, **36**:2105–2112.

43. 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.

44. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Guo D, Ott M, Zitnick CL, Ma J, Fergus R: **Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences**. *bioRxiv* 2020, https://doi.org/10.1101/622803.

45. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, *et al.*: **ProtTrans: towards cracking the language of life's code**

through self-supervised deep learning and high performance computing. *bioRxiv* 2020, https://doi.org/10.1101/2020.07.12.199554.

46. Rao R, Meier J, Sercu T, Ovchinnikov S, Rives A: **Transformer protein language models are unsupervised structure learners**. *bioRxiv* 2020, https://doi.org/10.1101/2020.12.15.422761.

47. Smith JS, Isayev O, Roitberg AE: **ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost**. *Chem Sci* 2017, **8**:3192–3203.

48. Wang J, Olsson S, Wehmeyer C, Pérez A, Charron NE, de Fabritiis G, Noé F, Clementi C: **Machine learning of coarse-grained molecular dynamics force fields**. *ACS Cent Sci* 2019, https://doi.org/10.1021/acscentsci.8b00913.

49. Wang Y, Fass J, Chodera JD: *End-to-End differentiable molecular mechanics force field construction*. 2020. arXiv:201001196 [physics].

50. Noé F, Tkatchenko A, Müller K-R, Clementi C: **Machine learning for molecular simulation**. *Annu Rev Phys Chem* 2020, **71**:361–390.

51. Ramaswamy VK, Willcocks CG, Degiacomi MT: *Learning protein conformational space by enforcing physics with convolutions and latent interpolations*. 2019. arXiv:191004543 [physics].

52. Chen W, Ferguson AL: **Molecular enhanced sampling with autoencoders: on-the-fly collective variable discovery and accelerated free energy landscape exploration**. *J Comput Chem* 2018, **39**:2079–2102.

53. Sultan MM, Wayment-Steele HK, Pande VS: **Transferable neural networks for enhanced sampling of protein dynamics**. *J Chem Theor Comput* 2018, **14**:1887–1894.

54. Noé F: **Machine learning for molecular dynamics on long timescales**. In *Machine learning meets quantum physics*. Edited by Schütt KT, Chmiela S, von Lilienfeld OA, Tkatchenko A, Tsuda K, Müller K-R, Springer International Publishing; 2020: 331–372.

55. Doerr S, Majewsk M, Pérez A, Krämer A, Clementi C, Noe F, Giorgino T, De Fabritiis G: **TorchMD: a deep learning framework for molecular simulations**. *J Chem Theor Comput* 2021, https://doi.org/10.1021/acs.jctc.0c01343.