

Artificial Intelligence Learns Protein Prediction

Michael Heinzinger¹ and Burkhard Rost^{1,2,3,4}

¹Technical University of Munich (TUM) School of School of Computation, Information and Technology (CIT),
Bioinformatics and Computational Biology - i12, 85748 Garching/Munich, Germany

²Institute for Advanced Study (TUM-IAS), 85748 Garching/Munich, Germany

³TUM School of Life Sciences Weihenstephan (WZW), 85354 Freising, Germany

⁴Department of Biochemistry and Molecular Biophysics, Columbia University, New York,
New York 10032, USA

Correspondence: mheinzing@rostlab.org

From *AlphaGO* over *StableDiffusion* to *ChatGPT*, the recent decade of exponential advances in artificial intelligence (AI) has been altering life. In parallel, advances in computational biology are beginning to decode the language of life: *AlphaFold2* leaped forward in protein structure prediction, and protein language models (pLMs) replaced expertise and evolutionary information from multiple sequence alignments with information learned from reoccurring patterns in databases of billions of proteins without experimental annotations other than the amino acid sequences. None of those tools could have been developed 10 years ago; all will increase the wealth of experimental data and speed up the cycle from idea to proof. AI is affecting molecular and medical biology at giant steps, and the most important might be the leap toward more powerful protein design.

SCIENCE FICTION OR FUTURE SCIENCE?

Walking her dog, Dr. Elena decides to engineer a bacterium efficiently gobbling up all those painkillers she had to swallow after her recent tooth extraction. She hopes to immerse those bugs into a wastewater facility. On her phone, she begins collecting a few dozen enzymes known to catalyze reactions similar to those needed to digest the environmentally toxic ingredients of that medicine. She downloads the bug's sequences and predicts the three-dimensional (3D) structures for all proteins (Box 1). Before reaching home, she has already created the

most likely functional 3D scaffolds relevant for the proteins to bind the toxins, has applied a protein language model (pLM) to generate millions of new sequences that might have similar 3D structures, has selected a few tens of top candidates for experimental testing. She has sent those sequences to her laboratory robot. When she reaches the laboratory, 2 hours later the robot has already gone through the first round of optimization with results being fed back directly to the pLM for further refinement of the top candidates. All is ready for more detailed experimental analysis thanks to the advances in artificial intelligence (AI).

BOX 1. ABBREVIATIONS USED

(1D) One-dimensional (string such as secondary structure), (2D) two-dimensional (interresidue distances or contacts), (3D) three-dimensional (coordinates), (AI) artificial intelligence, *AlphaFold2*: AI-based method reliably predicting protein 3D structure from MSAs (Jumper et al. 2021), (AI) artificial intelligence, (ANN) artificial feedforward neural network, (BFD) Big Fantastic Database (Steinegger et al. 2019), (CASP) critical assessment of protein structure prediction (biannual meeting), (CATH) hierarchical classification of protein 3D structures in the Class, Architecture, Topology and Homologous superfamily, (CNN) convolutional neural network, (EAT) embedding-based annotation transfer, (EI) evolutionary information, (embeddings) fixed-size vectors derived from pretrained pLMs, (IDPs) intrinsically disordered proteins (Dunker et al. 2013), (IDR) intrinsically disordered regions (Dunker et al. 2013), (ML) machine learning (here we drop the distinction between ML and AI considering existing delineations more or less arbitrary for our ends), (MSA) multiple sequence alignment, (pLM) protein language model, (PPI) physical protein–protein interaction, (SOTA) state-of-the-art.

EVOLUTIONARY INFORMATION POWER CHARGES PROTEIN PREDICTION**Secondary Structure Prediction Jumped by Combining AI and Alignments**

The application of advanced machine learning (here for simplicity coined AI) to protein prediction began 35 years ago, with simple artificial feedforward neural networks (ANNs) predicting protein secondary structure (Bohr et al. 1988; Qian and Sejnowski 1988). Although these, along with subsequent publications, provided the proof-of-principle for a powerful new technique, its breakthrough came by combining AI and evolutionary information (EI) as derived from multiple sequence alignments (MSAs) (Rost and Sander 1992, 1993). Secondary structure had been predicted by both AI (Bohr et al. 1988; Qian and Sejnowski 1988) and EI (Zvelebil et al. 1987); the successful novelty was the combination of both. This succeeded because ANNs captured long-range information (sequence separation between residues i and j such that, e.g., $|i - j| > 15$) much better than other statistical analyses of MSAs. The formula AI + EI was so successful that performance rose above what had been published in many textbooks as the theoretical limit, namely, a three-state per-residue accuracy of $Q3 = 65\%$ (Fig. 1; purple dashed horizontal line). The first method, dubbed PHD, surprisingly reached above 72% (Rost 1993, 1996), which pushed the advance more than the three decades of improvements and data col-

lection before (Fig. 1). The trick was to put more complex information (replace single sequences by protein families described by EI) into advanced learning methods capable to mine this complexity. Fine-grained control of the tool (balanced training and stacking of simple ANNs into constructs that resembled some aspects of deep learning a decade before its introduction) allowed to also fix other aspects of the problem not reflected in the simple three-state per-residue accuracy (Rost 1993).

AI + EI Recipe Boosts Other Aspects of Structure Prediction

The successful combination of AI + EI was expanded to other features of protein 1D structure, including the prediction of solvent accessibility and membrane regions (Rost 1996). The initial objective had been to predict interresidue distance maps (2D structure; Rost 1993), but the complexity of this objective required another decade (Punta and Rost 2005a). Although successful in many ways (Punta and Rost 2005b; Schlessinger et al. 2007), such methods did not suffice to generalize from 2D to 3D structure predictions. Overall, AI + EI broke through many ceilings, including predicting molecular function (Rost et al. 2003), but appeared to fail accurately predicting 2D or 3D structure.

In fact, the major advance toward 2D predictions sufficient for the generation of 3D structure

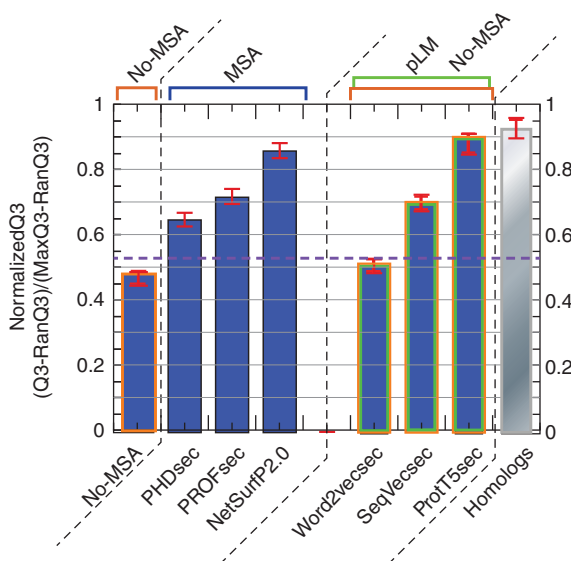


Figure 1. Rise in secondary structure prediction. Protein secondary structure prediction might be the simplest and best-understood aspect of structure prediction. Therefore, we use it as a proxy to compare different methods. The y-axis shows the performance in terms of normalized Q3 (Rost et al. 1994). This is defined as follows: $\text{NormalizedQ3} = (\text{Q3} - \text{RandomQ3}) / (\text{MaxQ3} - \text{RandomQ3})$; Q3: three-state per-residue accuracy (helix, strand, other); MaxQ3 = 92% approximates the secondary structure string agreement between alternate experimental structures for the same protein (Andersen et al. 2002); and minimal performance, namely, random, $\text{RandomQ3} = 35\%$ (Rost et al. 1994). By definition, NormalizedQ3 ranges from 0 (random) to 1 for predictions reaching the experimental resolution. Methods: *No-MSA* refers to simple statistical models or simple artificial feedforward neural networks (ANNs) not using evolutionary information (EI) from multiple sequence alignments (MSAs), *MSA-based*: *PHD_{sec}* marks the first stacked system of ANNs combining artificial intelligence (AI) and EI (Rost and Sander 1992), *PROF_{sec}* (Rost 2001) uses richer MSAs (from PSI-BLAST rather than BLAST; Altschul et al. 1997), *NetSurfP2* (Klausen et al. 2019) marks the top recent MSA-based predictions; *pLM-based* (protein language model-based methods): *Word2vec_{sec}* (Heinzinger et al. 2019) is the context-independent first generation of language models (LMs) (bag-of-words; Mikolov et al. 2013), *SeqVec_{sec}* (Heinzinger et al. 2019) uses the LM called ELMo (Peters et al. 2018), and *ProtT5_{sec}* (Elnaggar et al. 2021) is based on transformers; *Homologs* marks the ~88% agreement of secondary structure between proteins with similar sequences (Rost et al. 1994). The red-arrowed error bars roughly approximate a comparison between methods developed over the course of three decades and assessed on a diversity of data sets. The horizontal dashed purple line (marking $\text{Q3} \sim 65\%$) has been considered the top reachable for many years. In fact, this was true before AI + EI. The noncontext-aware *Word2vec* value confirms this level. On a side note, the rise for MSA-based solutions from *PHD_{sec}* (1992) to *NetSurfP2.0* (2019; *middle* set of blue bars) required about three orders of magnitude larger databases combined with much advanced AI (from long short-term memories [LSTMs] to convolutional neural networks [CNNs]) and took almost three decades. The comparable advance for pLM-based solutions from *Word2vec* to *ProtT5* took a little more than 3 years.

originated from a combination of advanced processing of EI in the form of evolutionary couplings (Marks et al. 2011). The successful signal-to-noise filtering that turned those couplings into a breakthrough solution required statistical models (Lapedes et al. 1999; Weigt et al. 2009; Balakrishnan et al. 2011; Marks et al. 2011; Jones et al. 2012; Seemayer et al. 2014). This advance was orthogonal to another set of tools

largely dominating CASP (critical assessment of protein structure prediction) (Moult et al. 1995, 1999, 2007; Kryshchuk et al. 2007), namely, programs using more or less directly comparative modeling (Baker and Sali 2001; Bonneau et al. 2001; Pieper et al. 2011; Biasini et al. 2014). Although comparative modeling and evolutionary coupling-based advances remained AI-free, the next step, once again, com-

bined the simpler statistical models with AI-based models (Wang et al. 2017; Yang et al. 2020). This approach peaked in *AlphaFold1* (Senior et al. 2020) (officially *AlphaFold*), the last method still on the way toward accurate 3D prediction.

ALPHAFOLDODOLOGY IS ALL THERE IS FROM NOW ON?

Leap in 3D Prediction by AI Just in Time

In December 2020, *AlphaFold2* (Jumper et al. 2021) broke through in protein structure prediction at CASP14 (Kryshtafovych et al. 2021). Like all top structure predictions since *PHD_{sec}* (Fig. 1), *AlphaFold2* succeeded through using EI from MSAs. Why did it take 28 years to get to this point? Simply put, the breakthrough could not have happened any earlier because it was rooted in three crucial advances coinciding at that point in time: (1) software (i.e., the advanced layered architectures) (deep neural networks—especially, the flexibility offered by *Transformers* paired with geometry-aware attention operations rendering the input 3D structure invariant to global rotations and translations), (2) hardware (advanced GPUs and TPUs), and (3) database sizes (*BFD*; Steinegger et al. 2019) is 10-times *UniProt* (The UniProt Consortium 2021). That the tool of the year (Marx 2022) came at all in 2020 required many successful novel solutions cleverly engineered by a large team of advanced experts fueled with sufficient resources (Jumper et al. 2021). Among many other inventions, the novelties included learning explicitly to predict (1) reliability (in the form of the so-called *pLDDT* score); (2) particular shapes (no method had ever implemented this explicitly); (3) to iterate over and thereby refine its own predictions; and (4) to increase data set size and diversity by training on its own predictions (which in turn needed a method as successful as *AlphaFold2*).

All these advances required ingenious AI engineering coding physical and geometrical constraints (inductive biases) directly into the components of the models, or more precisely: into its architecture. Joining these components with learning features in a way known as end-to-

end (i.e., by backpropagating the gradient from the predicted 3D structure to the MSA) allowed the model to directly learn features from the wealth of experimental 3D structures deposited in the PDB (Burley et al. 2023). This solution was in stark contrast to previous approaches that either relied on expertly crafted features or required external tools for computing input (evolutionary couplings) or targets (3D structures).

AlphaFold2 also leveraged the concept of distograms (Rost 1993) introduced unsuccessfully 30 years earlier. The DeepMind engineers made it work. One way to showcase the jump: While *AlphaFold1* outperformed any single method at CASP13 in 2018, it was not the best for any protein. In contrast, *AlphaFold2* at CASP14 in 2020 clearly outperformed each method for every protein, and immediately helped experimental structure determination (Millán et al. 2021). At CASP15 in 2022, *AlphaFold2* became the method to judge others by. Although many methods have reached performance levels that would have stunned the world 30 months ago, none has consistently reached the top, yet. So far, this seems as true for methods attempting to fully reengineer *AlphaFold2* as for those seeking “smarter” ways to improve.

AlphaFold2 Changes Experimental and Structural Biology

Molecular biologists have reacted by using the new tool (e.g., to advance experimental high-resolution determination of protein 3D structure) (Millán et al. 2021; Akdel et al. 2022; Bryant et al. 2022; Laurents 2022; Thorn 2022) to optimize docking and complement molecular dynamics (Guo et al. 2022; Laurents 2022; Tsaban et al. 2022). Many colleagues use the method successfully even for tasks for which it was not designed, such as predicting regions of intrinsically disordered proteins (Bryant et al. 2022; Guo et al. 2022; Ilzhöfer et al. 2022), or of permanent protein-protein interactions (PPIs) (Evans et al. 2021; Bryant et al. 2022; Johansson-Åkhe and Wallner 2022).

Although apparently successful for modeling the interaction between permanently bound

constituents of proteins, neither *AlphaFold2* nor the interaction specialist *AlphaFold-Multimer* (Evans et al. 2021) appear to rise up to the challenges of predicting binding of transient, physical PPIs for experimentally uncharacterized protein pairs (Burke et al. 2023; L Kaindl and B Rost, unpubl.). In fact, most existing methods either infer or predict such PPIs through the simple annotation transfer referred to as comparative modeling or homology-based modeling/inference applying the simple logic: if the sequence similarity between two proteins Q and A exceeds an empirically established threshold T, copy annotation of A to Q. PPI prediction is an even tougher nut to crack than 3D structure prediction, and the appropriate assessment of methods is easier to get wrong than right (Park and Marcotte 2012; Hamp and Rost 2015). Another aspect that appears not fully covered by *AlphaFold2* is the prediction of the effect of sequence variation (Weissenow et al. 2022a,b). The reason is that *AlphaFold2* reaches its peak performance by generating a family average rather than a protein-specific prediction. We might expect that this limitation could be bypassed because *AlphaFold2* has internal components that allow it to weigh the query sequence against the MSA, thereby moving between protein-specific and family-average. However, there is way too little experimental data on the mutational effect on 3D structure to leverage this effect. Thirty-one years ago, when the successful combination of AI and EI won the day (Rost and Sander 1992), the distinction between the two never became relevant due to lower performance (Fig. 1: no bar reaches the level of Homologs to the very right; although we have not assessed this, we assume that *AlphaFold2* would approach the experimental error, i.e., approach NormalizedQ3~1). Despite undisputed success, there are limits—even for systems such as *AlphaFold2*.

Overall, very typical for AI applications, the performance of *AlphaFold2* may even exceed the hype it created at CASP14 (Kryshtafovych et al. 2021). The *AlphaFoldDB* database now (June 2023) holds over 217 million 3D predictions (Tunyasuvunakool et al. 2021), and if you want to run the method on your set of sequences without requiring too much computing resources, *Colab-*

Fold (Mirdita et al. 2022) offers easy and fast access to the tool (essentially gaining speed by flexibly adjusting the number of iterations and by more efficient MSA generation).

Better Resolution of Structure Space

How much do the millions of accurate 3D structure predictions change our perception of structure space (i.e., the CATH [Sillitoe et al. 2021] or SCOP [Andreeva et al. 2020] classification of proteins by their 3D structure)? A recent analysis of 370,000 confident *AlphaFold2* 3D predictions could assign more than 90% of these to one of the known CATH domain superfamilies (Bordin et al. 2023). Expert analysis of the nonmatching human proteins revealed 25 novel superfamilies. Overall, the 130,000 predicted structures increased the number of “known folds” (i.e., compact 3D structural scaffolds typically shared between many distantly related proteins by more than 30%). An impressive tool for viewing protein structure space, which incidentally relies on *FoldSeek* (van Kempen et al. 2024) for comparing structures (below), has recently been made available by the developers of *ESMFold* (Lin et al. 2023). The confluence of two tools on the opposite side of the spectrum of data analysis, namely, on the side of great detail, the CATH-based analysis having experts visually compare the similarities and differences in detail for about a thousand proteins and on the side of large numbers the *ESMFold*-based threshold-driven automatic comparisons of 700 million proteins. The resulting message is so strong because it originates from both ends (detailed *AlphaFold2*/CATH and coarse-grained *ESMFold* perspective): The most important gain from 3D structure predictions is in refining known superfamilies, in linking previously isolated proteins to known groups, and providing some evidence for where to hunt for new “folds.”

New Solutions Boost the Power of *AlphaFold2*

Methods in computational and experimental biology that benefit from the results of *AlphaFold2* are mushrooming. In fact, just 2 years after the publication of *AlphaFold2*, their number and di-

versity are already beyond the scope of this perspective. Nevertheless, we want to highlight what we consider possibly the most important impact from accurate 3D structure predictions of the day, namely, *Foldseek* (van Kempen et al. 2024). For three decades, the most important criterion for the development of methods comparing protein 3D structures was accuracy (Taylor and Orengo 1989; Kolodny et al. 2005): The number of known 3D structures added at any time point were so small that computing resources were only of minor concern. With *AlphaFoldDB* (Tunyasuvunakool et al. 2021) and *ESMFold* (Lin et al. 2022) making hundreds of millions 3D structure predictions available, the field needed another revolution: Reliable 3D structure comparisons are three orders of magnitude faster than existing tools to make the gigantic amount of new data searchable and thereby useable. *FoldSeek* found a genius solution toward this end by mapping 3D structure through a so-called vector quantised-variational autoencoder (VQ-VAE) (van den Oord et al. 2017) onto an alphabet of 20 different letters. The resulting 3Di “states” can be imagined as configurations in the backbone angles most informative for known protein structures (possibly reminding more experienced scholars of the informative 3D motifs used early on to predict structure; Byströff and Baker 1998). The 3Di states are described by 20 letters that can be used to tap into the amazing accelerations realized for sequence comparisons through the blazingly fast *MMseqs2* (Mirdita et al. 2019). This engineering marvel enables *Foldseek* to reliably compare 3D structures (predictions or observations) at the speed of the faster ever reliable sequence comparisons. Just like *AlphaFold2*, *Foldseek* is a solution that would have been impossible 7 years ago, and it could never have been as useful as after the publication of *AlphaFoldDB* (Tunyasuvunakool et al. 2021). No matter how big this story in itself may be, it seeds an even bigger rift between past and future: Ultimately, *Foldseek* is the beginning of the end for sequence comparisons as they have been perfected for more than 42 years (Smith and Waterman 1981). *Foldseek* substitutes 1D sequence comparisons of proteins by 3D structure comparisons (although technically projected onto 1D sequences). Comparisons using 3D predic-

tions already outperform sequence-based alignment methods in many ways (Heinzinger et al. 2022; Schütze et al. 2022; van Kempen et al. 2024), even for relatively inaccurate 3D prediction methods (Weissenow et al. 2022b).

The term *AlphaFoldology* has been used half-jokingly by colleagues to capture some of the results of the amount of change brought about by the breakthrough *AlphaFold2*: From here on, all that is needed in structure prediction is to understand what we can and cannot do with *AlphaFold2*, to actually spread and dive into the word, reason, or discourse signifying the Greek word λόγος. Clearly, this has described many activities over the last 2 years. Another flurry of activity has been dedicated toward inventing tools that provide simpler access to *AlphaFold2*-level predictions, such as *ColabFold*, or that render the results from *AlphaFold2* even more useful, such as *Foldseek*. Undoubtedly, we will witness many more improvements.

The idea that chaining the tools *AlphaFold2/ColabFold-Foldseek* will replace traditional ways to generate MSAs leads to an interesting circle: *AlphaFold2* is so successful because it uses large MSAs. How can we maintain the success when removing the foundation? Today, we may or may not quite get there yet. However, thanks to pLMs, tomorrow we might easily accomplish the seemingly impossible.

PROTEIN LANGUAGE MODELS (pLMs) TRIGGER PARADIGM SHIFTS

Protein Language Models (pLMs): Learning the Language of Life

Advances in natural language processing (NLP) spawned pLMs (Alley et al. 2019; Bepler and Berger 2019, 2021; Heinzinger et al. 2019; Rives et al. 2021; Elnaggar et al. 2022). pLMs leverage protein sequence databases that have outgrown computers for 28 years; they require no annotation (*label* in AI-jargon) except for the amino acid sequence. Through transfer learning, pLMs bridge the sequence-annotation gap (Ofra et al. 2005) after *AlphaFold2* has substantially quenched the sequence-structure gap (Porta-Pardo et al. 2022) (i.e., the difference between proteins of known se-

quence and known structure) (Rost and Sander 1996). Explicitly, pLMs capture sequential patterns in the input by either predicting the next residue (amino acid; earlier pLMs) or by recovering masked-out residues from those surrounding it (later pLMs, typically 15%). Thereby, language models (LMs) from NLP implicitly learn the grammar of written languages. Similarly, pLMs implicitly learn aspects of the language of life as written in proteins (Fig. 2; Step 1). The information learned by such pLMs (e.g., by inputting a protein sequence into the network and construct-

ing vectors from the values representing the last hidden layers of the pLM) yields a representation of protein sequences referred to as embeddings (Fig. 2). This allows to transfer features learned by the pLM to any downstream (prediction) task requiring numerical protein representations (i.e., *transfer learning*) (Fig. 2; Step 2). In fact, the so-called *EvoFormer* in *AlphaFold2* that extracts information from an MSA is a modified version of the pLM-based solution learning directly from MSAs, namely, the *MSA-Transformer* (Rao et al. 2021). However, instead of pretraining on large,

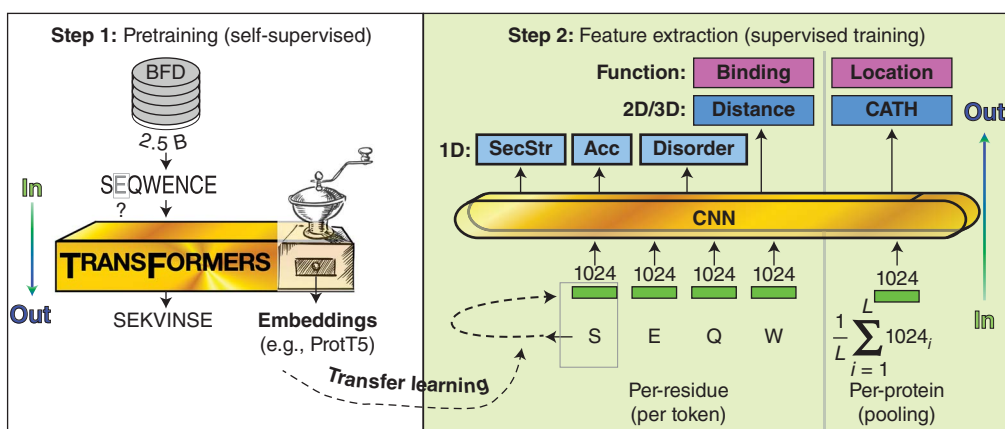


Figure 2. Embeddings from protein language model (pLM) for transfer learning. The sketch illustrates generic pLMs. *Step 1 (left)*: pictures the self-supervised learning of the context of protein sequences from training on very large sequence databases (e.g., *ProtT5* [Elnaggar et al. 2021] required *BFD* [Steinegger et al. 2019] with 2.5 billion sequences). Large transformer models (*ProtT5* has 3 billion connections, *Ankh* 1.7b, and *ESM2* 15b) inputting and outputting the same sequences, learn the *grammar of life* (i.e., implicitly extract the rules underlying the generation of known protein sequences). This extracted information is contained in the hidden layers of the transformers, which can be extracted in the form of vector representations, so-called embeddings. In *Step 2 (right)*: the embeddings are used as exclusive input for subsequent artificial intelligence (AI) trained in supervised manner on different tasks (e.g., per-residue prediction of secondary structure, accessibility, disorder, binding residues, or per-protein prediction of subcellular location or *CATH* numbers). *Step 2* is considered *transfer learning* because the *grammar* implicitly extracted from sequences in *Step 1* is transferred to increase the efficiency and success of *Step 2*. The number of units of the last hidden layers differs between pLMs (e.g., *ProtT5*) describes each residue (amino acid position) by a vector of 1024 dimensions. Therefore, the subsequent prediction method (here assumed to be a convolutional neural network, CNN) will use 1024 input units for each residue (i.e., $L \cdot 1024$ for a protein with L residues). The simplest way to condense this information for problems that require per-protein rather than per-residue predictions (e.g., per-residue: three-dimensional [3D] structure coordinate, per-protein: structural class such as *CATH* [Sillitoe et al. 2021; Nallapareddy et al. 2023]) is to average over each dimension). Although this resembles the concept of amino acid composition, which is well known to be informative of protein function, it is a priori not evident that such a crude average carries any meaningful description of protein features. In addition to per-residue (raw embeddings learned by pLM) and per-protein (pooling/averaging) embeddings, transformers also learn so-called *Attention Heads* (e.g., 732 for *ProtT5*) reflecting the importance (attention) from the entire protein onto a particular residue i . The latter is crucial when using embeddings to predict 2D and 3D structures (Weissenow et al. 2022a,b).

unlabeled data and transferring the resulting knowledge, the *EvoFormer* is only trained on the orders of magnitude smaller but labeled structure-prediction data set. Yet, the task of reconstructing masked amino acids from unmasked context, which is usually used to pretrain pLMs, was also used as part of *AlphaFold2* training.

Generic pLMs benefit from the transformer technology (Vaswani et al. 2017), which uses large neural networks to recover masked-out amino acids from the context of the entire input sequence. Several aspects seem crucial after 3 years of experience with the tool. Firstly, we need to begin with a very large sequence database (Elnaggar et al. 2022). UniProt (The UniProt Consortium 2021), with more than 220 million sequences, clearly outperforms Swiss-Prot (curated fraction of UniProt) and for some aspects the 10-fold larger BFD (Steinegger et al. 2019) might be needed. This implies that directly building generic pLMs from minute subsets such as *Homo sapiens* or eukaryotes, at this point, constitute a rather bad idea unless such specializations are used to refine existing models. Secondly, solving the task of optimally recovering the sequence from the masked version appears another ill-advised objective (Heinzinger et al. 2019; Elnaggar et al. 2022). After all, the number of pre-training tokens outweighs the number of model parameters usually by orders of magnitude, which prevents the model from learning a perfect reconstruction but instead forces the model to learn some compression of the input. This, in fact, implies that training pLMs is a task relatively free of the typical challenge for AI, namely, over-training or overfitting. Instead, pLMs are optimized to detect, combine, and compress reoccurring patterns in the input, which is an optimal starting point for a variety of transfer learning tasks (Fig. 2B), and this objective cannot easily be summarized by any reasonable single number (Heinzinger et al. 2019; Elnaggar et al. 2022, 2023). The emphasis here is on *reasonable*: while there are many nonsense averages, a solution successfully processing different prediction tasks remains wanted. This strategy optimizes the generic usefulness of embeddings from pLMs, thereby also considering the aspect of energy consumption: Training pLMs is extremely resource-inten-

sive (Heinzinger et al. 2019; Elnaggar et al. 2022, 2023; Lin et al. 2023), with the latest versions even needing the new Google TPUs (Elnaggar et al. 2023). This investment appears justified if and only if the resulting pLMs are sufficiently generic to serve for a large diversity of transfer learning solutions.

Successful transfer solutions exist for diverse aspects of protein prediction (Fig. 2) ranging from 3D structure (Rao et al. 2020; Bhattacharya et al. 2021; Chowdhury et al. 2022; Wang et al. 2022; Weissenow et al. 2022a,b; Wu et al. 2022; Lin et al. 2023), transmembrane regions (Bernhofer and Rost 2022; Hallgren et al. 2022), and intrinsically disordered regions (IDR/IDP) (Ilzhöfer et al. 2022) to various aspects of function (Littmann et al. 2021a; Stärk et al. 2021; Villegas-Morcillo et al. 2021; Bileschi et al. 2022; Heinzinger et al. 2022; Nallapareddy et al. 2023). Distance in embedding space correlates more with protein function than with sequence similarity (Littmann et al. 2021a) and can help with clustering proteins into families (Littmann et al. 2021a; Bileschi et al. 2022; Heinzinger et al. 2022).

Embedding-Based Outperform MSA-Based Predictions

Using embeddings from pLMs instead of EI from MSAs simplifies and speeds up protein prediction. For several tasks, including the prediction of secondary structure (Elnaggar et al. 2022, 2023), transmembrane helices and strands (Bernhofer and Rost 2022; Hallgren et al. 2022), signal peptides (Teufel et al. 2022), subcellular location (Stärk et al. 2021), or binding residues (Littmann et al. 2021b). The latter might be the most impressive example for the concept of transfer learning: the reliable experimental data about binding continues to be so sparse that machine learning cannot easily manage the complexity of the problem. The result is that solutions require models with minimal input information (e.g., neural networks with fewer than 20 input units). Using embeddings from pLMs allows for scaling up two orders of magnitude, simply because the embeddings are extremely informative (Littmann et al. 2021b).

Despite substantial effort, 2 years after the first introduction of *AlphaFold2*, no method outperforms this remarkable solution despite ample reengineering attempts. What about embedding-based methods? In fact, the first pLMs have not even been able to reach the level of performance of the previous state-of-the-art (SOTA) reached by, for example, *AlphaFold1* (Senior et al. 2020) or *Raptor-X* (Wang et al. 2016); instead it required the more advanced transformers to reach parity (Weissenow et al. 2022a). However, even the better pLM-based methods, such as *ESMFold* (Lin et al. 2022), still remain substantially below *AlphaFold2* (and below other solutions reengineering that solution such as *RoseTTAFold*; Baek et al. 2021). For the time being, it remains unclear whether or not future pLMs might leap above the *AlphaFold2* mark.

If you wanted to apply pLMs to your prediction task, how to know whether it will work? Although we have no comprehensive answer, a few empirical observations might help. (1) Assume that you try to predict a feature with too little reliable data for the complexity of the task. The odds are good that embeddings might help (e.g., prediction of binding residues) (Littmann et al. 2021b), unless existing methods are extremely complex and optimized for this particular task. For instance, impressively complex solutions predict intrinsically disordered regions (Del Conte et al. 2023). Similarly, transmembrane helices, signal peptides, and subcellular location are predicted by well-tuned methods. In all of these cases, however, there are pLM-based solutions that reach or outperform the SOTA after some additional adjustments (Stärk et al. 2021; Bernhofer and Rost 2022; Hallgren et al. 2022; Ilzhöfer et al. 2022; Teufel et al. 2022). Predicting the effect of sequence variation is an example for lack of data for which advanced MSA-based methods still tend to have the upper hand. (2) Conversely, for tasks for which ample data exist (e.g., the prediction of protein inter-residue distances [2D structure] or actual 3D structure), solutions using very informative MSAs clearly dominate (although when inputting single sequence rather than MSAs *AlphaFold2* is outperformed by pLM-based solutions; Lin et al. 2022; Weissenow et al. 2022b; Wu et al.

2022). However, the same is not true for the much simpler secondary structure prediction for which rich, informative MSAs also yield much better results, but for which pLM-based predictions appear to approach the ultimate possible. The reason why 1D secondary structure and 2D interresidue distances or 3D coordinates behave so differently remains unknown. In lieu of knowledge, we might speculate two possible answers: (1) saturation effect: secondary structure prediction might be closer to the top level of performance possible, namely, the “experimental error” than 3D prediction, or (2) simplicity: possibly pLMs condense less information than MSAs and advanced modules in *AlphaFold2* might turn any additional information into better predictions.

Protein-Specific Rather than Family-Averaged Predictions

One technical aspect of pLM-based prediction methods is that they do not need MSAs. This has two advantages. Firstly, with growing databases MSA generation requires resources. Although *MMseqs2* (Mirdita et al. 2019) as used for *ColabFold* (Mirdita et al. 2022) is now so fast that this hardly matters in everyday applications (Bernhofer et al. 2021). Secondly, for some proteins, such as intrinsically disordered proteins (Dunker et al. 2013) or the *dark proteome* (Perdigão et al. 2015) MSA generation becomes problematic. Possibly more substantial, however, is another advantage: pLM-based methods enable protein-specific rather than family-averaged predictions. For instance, this permits to predict the effects of sequence variation upon function (Meier et al. 2021; Marquet et al. 2022; Dunham et al. 2023). Unfortunately, solutions predicting the effects upon structure appear to be largely confined to signals captured by the transformers rather than by the differential structure prediction (Weissenow et al. 2022a, b). Even a feature as intricately linked to EI as the “conservation” within a family can be predicted surprisingly well by rather simplistic convolution neural networks (Marquet et al. 2022). Does this imply that pLMs have captured evolutionary information? Although we have

collected ample indirect evidence pointing to an affirmative answer (K Erckert and B Rost, unpubl.), we continue to doubt this to be the case. One reason for this is that pLMs are apparently extremely sensitive to differential data sets: A difference of less than an order of magnitude between the least (cysteine) and most frequent amino acid (leucine) leads to a substantially poorer resolution of cysteine. In contrast, the “dilution” of family relations is at least five orders of magnitude smaller: large families have $<10^4$ members, BFD holds $>10^9$ proteins (i.e., the difference between the number of related (same family) and unrelated (different family) sequences any protein will pick up exceeds 10^5). Although not impossible, this is unlikely to be picked up by today’s pLMs. If so, why do they capture information relevant to predict family conservation? Most likely because the grammar of the language of life as written in proteins is coined by the same constraints written into the conservation profiles reflected by MSAs. In other words, both pLMs and MSAs capture aspects of the grammar, one due to biophysical constraints imprinted into sequences, the other due to constraints regulating what is observed—in MSAs—and what is not.

Beginning of the End for Alignment Methods?

Leveraging the sparse experimental annotations available 70 years ago, methods comparing protein sequences (Schwartz and Dayhoff 1978; Smith and Waterman 1981) have arguably been at the center of development for the field of computational biology since its existence. Although protein sequence databases have been growing faster than the speed at which computers can cope with this wealth of data since the mid-1990s, blazingly fast solutions such as *MMseqs2* (Mirdita et al. 2019) have succeeded in staying on top of the seemingly lost challenge through finding shortcuts with acceptable performance loss for most users. In fact, *MMseqs2* is even able to cope with challenges amounting to *Gargantuan* tasks such as the more than 3.1 quintillion ($2.5 \times 1.25 \times 10^{18}$) pair comparisons required for an all-on-all of BFD (Steinegger et al. 2019). However, except for requiring resources, sequence com-

parisons have been limited by the simple assumption that the alignment at positions i and j are statistically independent of each other. It continues to be stunning how successful an entire field can become even when built upon such a blatantly incorrect assumption. The only approach surmounting this problem, the Genetic Algorithm-based T-Coffee (Notredame et al. 2000), was already too slow for comparing to entire databases 20 years ago. The advent of pLMs offers another stab at this problem: if we could replace sequence comparisons by the generalized sequences generated by embeddings, we might be able to capture correlations between residues i and j . In its simplest implementation of per-protein comparisons, this approach indeed works very well to predict GeneOntology (GO) numbers (Littmann et al. 2021a), Pfam families (Bileschi et al. 2022), and CATH numbers (Heinzinger et al. 2022; Nallapareddy et al. 2023), ultimately, because similarity in embedding space is more informative for inferring similarity in function than the similarity in sequence space (Littmann et al. 2021a). Existing embedding-based protein comparisons also benefit from immense speed: by describing the protein as one average number (e.g., with 1024 dimensions for ProtT5), they project the task of comparison to a simple vector product. To picture the power of embeddings, just imagine you wanted to compare two proteins based on their 20D vectors of amino acid composition: clearly not sufficient to distinguish between two proteins with similar GO numbers from among hundreds of thousands! The devil in the details is that most proteins have more than one domain, and as many have three or more than that (Liu and Rost 2002). As long as we use a per-protein average embedding, we will not succeed in capturing domain similarities unless we base the comparison on domains such as those described by Pfam and CATH (Schütze et al. 2022). When comparing multidomain full-length proteins, we need to refine by actually comparing k -mers of per-residue embeddings between pairs of proteins (Schütze et al. 2022). Although this is both possible and successful, it comes with an overhead in runtime, and for the time being the costs seem not to justify the gain (Schütze et al. 2022). This result brings

up another issue: so far, due to limitations in resources, generic pLMs have largely been built upon the software tools developed for NLP (i.e., the LMs). Ultimately, LMs will become available for handheld telephones. At that point, pLMs might retire traditional protein alignment methods.

CONCLUSIONS

In the first approximation, *AlphaFold2* solved the protein 3D structure prediction through a combination of advanced AI with advanced EI from MSAs generated from ever-growing protein sequence databases. Combining AI and EI as it peaked in *AlphaFold2* has been the winning card for almost three decades (Fig. 1). The *AlphaFold2* leap that could not have been realized more than 5 years ago has already spawned a flurry of important new developments. One of those, *Foldseek*, enables extremely reliable and fast comparison of millions of predicted or observed protein 3D structures. This tool immensely increases the value of databases with hundreds of millions of accurate 3D predictions from *AlphaFold2* (*AlphaFoldDB*) or *ESMFold*, and it might do more than any other solution to complement (or even replace at some point) the tools for protein sequence comparisons (alignment methods) upon which *Foldseek* is based. Another equally explosive and orthogonal development has been the introduction and rapid improvement of pLMs (Fig. 2). Through successful transfer learning (Fig. 2) that leverages information from unannotated protein sequences to help in learning from (even very little) experimental data, these tools begin to replace MSAs and open the age of making protein-specific trump family-averaged predictions. The combination of those two revolutions from successful 3D prediction and pLMs begins to spawn successful protein design beyond what was possible before and will help fictive Dr. Elena to solve real-world problems that we are already facing today.

ACKNOWLEDGMENTS

Thanks primarily to Chris Dallago (NVIDIA), Martin Steinegger (Seoul National University),

and Christine Orengo (UCL) for invaluable collaborations that have strongly supported our work, and to Ahmed Elnaggar (TUM) for his immense panache in helping to push the development of protein language models in our group at TUM. M.H. and B.R. were supported by the Bavarian Ministry of Education through funding to the TUM, by a grant from the Alexander von Humboldt Foundation through the German Ministry for Research and Education (BMBF: Bundesministerium für Bildung und Forschung), and by a grant from Deutsche Forschungsgemeinschaft (DFG-GZ: RO1320/4-1). Last but not least, thanks to all those who maintain public sequence databases, in particular Alex Bateman (UniProt, EBI Hinxton), Johannes Söding (MPI Göttingen), Martin Steinegger (Seoul National University) and their crews, and to all experimentalists who enabled this analysis by making their data publicly available.

REFERENCES

- Akdel M, Pires DEV, Pardo EP, Jänes J, Zalevsky AO, Mézáros B, Bryant P, Good LL, Laskowski RA, Pozzati G, et al. 2022. A structural biology community assessment of AlphaFold2 applications. *Nat Struct Mol Biol* **29**: 1056–1067. doi:10.1038/s41594-022-00849-w
- Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. 2019. Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods* **16**: 1315–1322. doi:10.1038/s41592-019-0598-1
- Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped blast and PSI-blast: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402. doi:10.1093/nar/25.17.3389
- Andersen CAF, Palmer AG, Brunak S, Rost B. 2002. Continuum secondary structure captures protein flexibility. *Structure* **10**: 175–184. doi:10.1016/s0969-2126(02)00700-1
- Andreeva A, Kulesha E, Gough J, Murzin AG. 2020. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res* **48**: D376–D382. doi:10.1093/nar/gkz1064
- Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, et al. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**: 871–876. doi:10.1126/science.abj8754
- Baker D, Sali A. 2001. Protein structure prediction and structural genomics. *Science* **294**: 93–96. doi:10.1126/science.1065659
- Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ. 2011. Learning generative models for protein fold families. *Proteins* **79**: 1061–1078. doi:10.1002/prot.22934

M. Heinzinger and B. Rost

- Bepler T, Berger B. 2019. Learning protein sequence embeddings using information from structure. *arXiv* doi:10.48550/arXiv.1902.08661
- Bepler T, Berger B. 2021. Learning the protein language: evolution, structure, and function. *Cell Syst* **12**: 654–669.e3. doi:10.1016/j.cels.2021.05.017
- Bernhofer M, Rost B. 2022. TMbed: transmembrane proteins predicted through Language Model embeddings. *BMC Bioinformatics* **23**: 326. doi:10.1186/s12859-022-04873-x
- Bernhofer M, Dallago C, Karl T, Satagopam V, Heinzinger M, Littmann M, Olenyi T, Qiu J, Schoetze K, Yachdav G, et al. 2021. PredictProtein—predicting protein structure and function for 29 years. *Nucleic Acids Res* **49**: W535–W540. doi:10.1093/nar/gkab354
- Bhattacharya N, Thomas N, Rao R, Dauparas J, Koo PK, Baker D, Song YS, Ovchinnikov S. 2021. Interpreting pots and transformer protein models through the lens of simplified attention. *Bioinformatics* **2022**: 34–45. doi:10.1142/9789811250477_0004
- Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Gallo Cassarino T, Bertoni M, Bordoli L, et al. 2014. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res* **42**: W252–W258. doi:10.1093/nar/gku340
- Bileschi ML, Belanger D, Bryant DH, Sanderson T, Carter B, Sculley D, Bateman A, DePristo MA, Colwell LJ. 2022. Using deep learning to annotate the protein universe. *Nat Biotechnol* **40**: 932–937. doi:10.1038/s41587-021-01179-w
- Bohr H, Bohr J, Brunak S, Cotterill RMJ, Lautrup B, Nørskov L, Olsen OH, Petersen SB. 1988. Protein secondary structure and homology by neural networks. The α -helices in rhodopsin. *FEBS Lett* **241**: 223–228. doi:10.1016/0014-5793(88)81066-4
- Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CE, Baker D. 2001. Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins* **45**: 119–126. doi:10.1002/prot.1170
- Bordin N, Sillitoe I, Nallapareddy V, Rauer C, Lam SD, Waman VP, Sen N, Heinzinger M, Littmann M, Kim S, et al. 2023. AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Commun Biol* **6**: 160. doi:10.1038/s42003-023-04488-9
- Bryant P, Pozzati G, Zhu W, Shenoy A, Kundrotas P, Elofsson A. 2022. Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. *Nat Commun* **13**: 6028. doi:10.1038/s41467-022-33729-4
- Burke DF, Bryant P, Barrio-Hernandez I, Memon D, Pozzati G, Shenoy A, Zhu W, Dunham AS, Albanese P, Keller A, et al. 2023. Towards a structurally resolved human protein interaction network. *Nat Struct Mol Biol* **30**: 216–225. doi:10.1038/s41594-022-00910-8
- Burley SK, Bhikadiya C, Bi C, Bittrich S, Chao H, Chen L, Craig PA, Crichlow GV, Dalenberg K, Duarte JM, et al. 2023. RCSB protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res* **51**: D488–D508. doi:10.1093/nar/gkac1077
- Bystroff C, Baker D. 1998. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol* **281**: 565–577. doi:10.1006/jmbi.1998.1943
- Chowdhury R, Bouatta N, Biswas S, Floristean C, Kharkar A, Roy K, Rochereau C, Ahdritz G, Zhang J, Church GM, et al. 2022. Single-sequence protein structure prediction using a language model and deep learning. *Nat Biotechnol* **40**: 1617–1623. doi:10.1038/s41587-022-01432-w
- Del Conte A, Bouhraoua A, Mehdiabadi M, Clementel D, Monzon AM, CAID predictors, Tosatto SCE, Piovesan D. 2023. CAID prediction portal: a comprehensive service for predicting intrinsic disorder and binding regions in proteins. *Nucleic Acids Res* **51**: W62–W69. doi:10.1093/nar/gkad430
- Dunham AS, Beltrao P, AlQuraishi M. 2023. High-throughput deep learning variant effect prediction with sequence UNET. *Genome Biol* **24**: 110. doi:10.1101/2022.05.23.493038
- Dunker AK, Babu MM, Barbash E, Blackledge M, Bondos SE, Dosztányi Z, Dyson HJ, Forman-Kay JD, Fuxreiter M, Gsponer J, et al. 2013. What's in a name? Why these proteins are intrinsically disordered. *Intrinsically Disordered Proteins* **1**: e24157. doi:10.4161/idp.24157
- Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Yu W, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, et al. 2022. Prottrans: toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* **44**: 7112–7127. doi:10.1109/TPAMI.2021.3095381
- Elnaggar A, Essam H, Salah-Eldin W, Mousafa W, Elkerdawy M, Rochereau C, Rost B. 2023. Ankh: optimized protein language model unlocks general-purpose modelling. *bioRxiv* doi:10.48550/arXiv.2301.06568
- Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, Židek A, Bates R, Blackwell S, Yim J, et al. 2021. Protein complex prediction with AlphaFold-Multimer. *bioRxiv* doi:10.1101/2021.10.04.463034
- Guo HB, Perminov A, Bekele S, Kedziora G, Farajollahi S, Varaljay V, Hinkle K, Molinero V, Meister K, Hung C, et al. 2022. Alphafold2 models indicate that protein sequence determines both structure and dynamics. *Sci Rep* **12**: 10696. doi:10.1038/s41598-022-14382-9
- Hallgren J, Tsigirigou KD, Pedersen MD, Almagro Armenteros JJ, Marcatili P, Nielsen H, Krogh A, Winther O. 2022. DeepTMHMM predicts α and β transmembrane proteins using deep neural networks. *bioRxiv* doi:10.1101/2022.04.08.487609
- Hamp T, Rost B. 2015. More challenges for machine-learning protein interactions. *Bioinformatics* **31**: 1521–1525. doi:10.1093/bioinformatics/btu857
- Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, Rost B. 2019. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics* **20**: 723. doi:10.1186/s12859-019-3220-8
- Heinzinger M, Littmann M, Sillitoe I, Bordin N, Orengo C, Rost B. 2022. Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genom Bioinform* **4**: lqac043. doi:10.1093/nargab/lqac043
- Ilzhöfer D, Heinzinger M, Rost B. 2022. SETH predicts nuances of residue disorder from protein embeddings. *Front Bioinform* **2**: 1019597. doi:10.3389/fbinf.2022.1019597

- Johansson-Åkhe I, Wallner B. 2022. Improving peptide-protein docking with AlphaFold-multimer using forced sampling. *Front Bioinform* **2**: 959160. doi:10.3389/fbinf.2022.959160
- Jones DT, Buchan DWA, Cozzetto D, Pontil M. 2012. PSI-COV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**: 184–190. doi:10.1093/bioinformatics/btr638
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**: 583–589. doi:10.1038/s41586-021-03819-2
- Klausen MS, Jespersen MC, Nielsen H, Jensen KK, Jurtz VI, Sønderby CK, Sommer MOA, Winther O, Nielsen M, Petersen B, et al. 2019. NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning. *Proteins* **87**: 520–527. doi:10.1002/prot.25674
- Kolodny R, Koehl P, Levitt M. 2005. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J Mol Biol* **346**: 1173–1188. doi:10.1016/j.jmb.2004.12.032
- Kryshchukovych A, Fidelis K, Moulton J. 2007. Progress from CASP6 to CASP7. *Proteins* **69** (Suppl 8): 194–207. doi:10.1002/prot.21769
- Kryshchukovych A, Schwede T, Topf M, Fidelis K, Moulton J. 2021. Critical assessment of methods of protein structure prediction (CASP)—round XIV. *Proteins* **89**: 1607–1617. doi:10.1002/prot.26237
- Lapedes AS, Liu L, Stormo GD. 1999. Correlated mutations in models of protein sequences: phylogenetic and structural effects. In *Proceedings of the IMS/AMS International Conference on Statistics in Molecular Biology and Genetics*, pp. 236–256. Institute for Mathematical Statistics, Hayward, CA.
- Laurents DV. 2022. Alphafold 2 and NMR spectroscopy: partners to understand protein structure, dynamics and function. *Front Mol Biosci* **9**: 906437. doi:10.3389/fmolb.2022.906437
- Lin Z, Akin H, Rao R, Hie BL, Zhu Z, Lu W, dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, et al. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv doi:10.1101/2022.07.20.500902
- Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**: 1123–1130. doi:10.1126/science.ade2574
- Littmann M, Heinzinger M, Dallago C, Olenyi T, Rost B. 2021a. Embeddings from deep learning transfer GO annotations beyond homology. *Sci Rep* **11**: 1160. doi:10.1038/s41598-020-80786-0
- Littmann M, Heinzinger M, Dallago C, Weissenow K, Rost B. 2021b. Protein embeddings and deep learning predict binding residues for various ligand classes. *Sci Rep* **11**: 23916. doi:10.1038/s41598-021-03431-4
- Liu J, Rost B. 2002. Target space for structural genomics revisited. *Bioinformatics* **18**: 922–933. doi:10.1093/bioinformatics/18.7.922
- Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. 2011. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**: e28766. doi:10.1371/journal.pone.0028766
- Marquet C, Heinzinger M, Olenyi T, Dallago C, Erckert K, Bernhofer M, Nechaev D, Rost B. 2022. Embeddings from protein language models predict conservation and variant effects. *Hum Genet* **141**: 1629–1647. doi:10.1007/s00439-021-02411-y
- Marx V. 2022. Method of the year: protein structure prediction. *Nat Methods* **19**: 5–10. doi:10.1038/s41592-021-01359-1
- Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A. 2021. Language models enable zero-shot prediction of the effects of mutations on protein function. bioRxiv doi:10.1101/2021.07.09.450648
- Mikolov T, Chen K, Corrado G, Dean J. 2013. Efficient estimation of word representations in vector space. arXiv doi:10.48550/arXiv.1301.3781
- Millán C, Keegan RM, Pereira J, Sammito MD, Simpkin AJ, McCoy AJ, Lupas AN, Hartmann MD, Rigden DJ, Read RJ. 2021. Assessing the utility of CASP14 models for molecular replacement. *Proteins* **89**: 1752–1769. doi:10.1002/prot.26214
- Mirdita M, Steinegger M, Söding J. 2019. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* **35**: 2856–2858. doi:10.1093/bioinformatics/bty1057
- Mirdita M, Schütze K, Moriwaiki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. Colabfold: making protein folding accessible to all. *Nat Methods* **19**: 679–682. doi:10.1038/s41592-022-01488-1
- Moulton J, Pedersen JT, Judson R, Fidelis K. 1995. A large-scale experiment to assess protein structure prediction methods. *Proteins* **23**: ii–iv. doi:10.1002/prot.340230303
- Moulton J, Hubbard T, Bryant SH, Fidelis K, Pedersen JT. 1999. Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins* **3**: 2–6. doi:10.1002/(SICI)1097-0134(1999)37:3<+2::AID-PROT2>3.0.CO;2-2
- Moulton J, Fidelis K, Kryshchukovych A, Rost B, Hubbard T, Tramontano A. 2007. Critical assessment of methods of protein structure prediction—round VII. *Proteins* **69** (Suppl 8): 9–10. doi:10.1002/prot.21767
- Nallapareddy V, Bordin N, Sillitoe I, Heinzinger M, Littmann M, Waman VP, Sen N, Rost B, Orengo C. 2023. CATH: detection of remote homologues for CATH superfamilies using embeddings from protein language models. *Bioinformatics* **39**: btad029. doi:10.1093/bioinformatics/btad029
- Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**: 205–217. doi:10.1006/jmbi.2000.4042b
- Ofran Y, Punta M, Schneider R, Rost B. 2005. Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. *Drug Discov Today* **10**: 1475–1482. doi:10.1016/S1359-6446(05)03621-4
- Park Y, Marcotte EM. 2012. Flaws in evaluation schemes for pair-input computational predictions. *Nat Methods* **9**: 1134–1136. doi:10.1038/nmeth.2259
- Perdigão N, Heinrich J, Stolte C, Sabir KS, Buckley MJ, Tabor B, Signal B, Gloss BS, Hammang CJ, Rost B, et al. 2015.

M. Heinzinger and B. Rost

- Unexpected features of the dark proteome. *Proc Natl Acad Sci* **112**: 15898–15903. doi:10.1073/pnas.1508380112
- Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. 2018. Deep contextualized word representations. arXiv doi:10.48550/arXiv.1802.05365
- Pieper U, Webb BM, Barkan DT, Schneidman-Duhovny D, Schlessinger A, Braberg H, Yang Z, Meng EC, Pettersen EF, Huang CC, et al. 2011. Modbase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res* **39**: D465–D474. doi:10.1093/nar/gkq1091
- Porta-Pardo E, Ruiz-Serra V, Valentini S, Valencia A. 2022. The structural coverage of the human proteome before and after AlphaFold. *PLoS Comput Biol* **18**: e1009818. doi:10.1371/journal.pcbi.1009818
- Punta M, Rost B. 2005a. PROFcon: novel prediction of long-range contacts. *Bioinformatics* **21**: 2960–2968. doi:10.1093/bioinformatics/bti454
- Punta M, Rost B. 2005b. Protein folding rates estimated from contact predictions. *J Mol Biol* **348**: 507–512. doi:10.1016/j.jmb.2005.02.068
- Qian N, Sejnowski TJ. 1988. Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol* **202**: 865–884. doi:10.1016/0022-2836(88)90564-5
- Rao R, Meier J, Sercu T, Ovchinnikov S, Rives A. 2020. Transformer protein language models are unsupervised structure learners. bioRxiv doi:10.1101/2020.12.15.422761
- Rao RM, Liu J, Verkuil R, Meier J, Canny J, Abbeel P, Sercu T, Rives A. 2021. MSA transformer. In *Proceedings of the 38th International Conference on Machine Learning* (ed. Marina M, Tong Z), pp. 8844–8856. *PMLR* **139**: 8844–8856.
- Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci* **118**: e2016239118. doi:10.1073/pnas.2016239118
- Rost B. 1993. *Neural networks and evolution—advanced prediction of protein secondary structure*. Departments of Physics and Astronomy, University of Heidelberg, Germany.
- Rost B. 1996. PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol* **266**: 525–539. doi:10.1016/s0076-6879(96)66033-9
- Rost B. 2001. Protein secondary structure prediction continues to rise. *J Struct Biol* **134**: 204–218. doi:10.1006/jjsbi.2001.4336
- Rost B, Sander C. 1992. Jury returns on structure prediction. *Nature* **360**: 540. doi:10.1038/360540b0
- Rost B, Sander C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* **232**: 584–599. doi:10.1006/jmbi.1993.1413
- Rost B, Sander C. 1996. Bridging the protein sequence-structure gap by structure predictions. *Annu Rev Biophys Biomol Struct* **25**: 113–136. doi:10.1146/annurev.bb.25.060196.000553
- Rost B, Sander C, Schneider R. 1994. Redefining the goals of protein secondary structure prediction. *J Mol Biol* **235**: 13–26. doi:10.1016/s0022-2836(05)80007-5
- Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofra Y. 2003. Automatic prediction of protein function. *Cell Mol Life Sci* **60**: 2637–2650. doi:10.1007/s00018-003-3114-8
- Schlessinger A, Punta M, Rost B. 2007. Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics* **23**: 2376–2384. doi:10.1093/bioinformatics/btm349
- Schütze K, Heinzinger M, Steinegger M, Rost B. 2022. Nearest neighbor search on embeddings rapidly identifies distant protein relations. *Front Bioinform* **2**: 1033775. doi:10.3389/fbinf.2022.1033775
- Schwartz RM, Dayhoff MO. 1978. Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. *Science* **199**: 395–403. doi:10.1126/science.202030
- Seemayer S, Gruber M, Söding J. 2014. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics* **30**: 3128–3130. doi:10.1093/bioinformatics/btu500
- Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Židek A, Nelson AWR, Bridgland A, et al. 2020. Improved protein structure prediction using potentials from deep learning. *Nature* **577**: 706–710. doi:10.1038/s41586-019-1923-7
- Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, Pang CSM, Woodridge L, Rauer C, Sen N, et al. 2021. CATH: increased structural coverage of functional space. *Nucleic Acids Res* **49**: D266–D273. doi:10.1093/nar/gkaa1079
- Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J Mol Biol* **147**: 195–197. doi:10.1016/0022-2836(81)90087-5
- Stärk H, Dallago C, Heinzinger M, Rost B. 2021. Light attention predicts protein location from the language of life. *Bioinform Adv* **1**: vbab035. doi:10.1093/bioadv/vbab035
- Steinegger M, Mirdita M, Söding J. 2019. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat Methods* **16**: 603–606. doi:10.1038/s41592-019-0437-4
- Taylor WR, Orengo CA. 1989. Protein structure alignment. *J Mol Biol* **208**: 1–22. doi:10.1016/0022-2836(89)90084-3
- Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G, Nielsen H. 2022. Signalp 6.0 predicts all five types of signal peptides using protein language models. *Nat Biotechnol* **40**: 1023–1025. doi:10.1038/s41587-021-01156-3
- The UniProt Consortium. 2021. Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**: D480–D489. doi:10.1093/nar/gkaa1100
- Thorn A. 2022. Artificial intelligence in the experimental determination and prediction of macromolecular structures. *Curr Opin Struct Biol* **74**: 102368. doi:10.1016/j.sbi.2022.102368
- Tsaban T, Varga JK, Avraham O, Ben-Aharon Z, Khramushin A, Schueler-Furman O. 2022. Harnessing protein folding neural networks for peptide-protein docking. *Nat Commun* **13**: 176. doi:10.1038/s41467-021-27838-9
- Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Židek A, Bridgland A, Cowie A, Meyer C, Laydon A, et al. 2021. Highly accurate protein structure prediction for the human proteome. *Nature* **596**: 590–596. doi:10.1038/s41586-021-03828-1

- van den Oord A, Vinyals O, Kavukcuoglu K. 2017. Neural discrete representation learning. arXiv doi:10.48550/arXiv.1711.00937
- van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. 2024. Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* **42**: 243–246. doi:10.1038/s41587-023-01773-0
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010. Curran, Long Beach, CA.
- Villegas-Morcillo A, Makrodimitris S, van Ham R, Gomez AM, Sanchez V, Reinders MJT. 2021. Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function. *Bioinformatics* **37**: 162–170. doi:10.1093/bioinformatics/btaa701
- Wang S, Li W, Liu S, Xu J. 2016. RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Res* **44**: W430–W435. doi:10.1093/nar/gkw306
- Wang S, Sun S, Li Z, Zhang R, Xu J. 2017. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLOS Comput Biol* **13**: e1005324. doi:10.1371/journal.pcbi.1005324
- Wang G, Fang X, Wu Z, Liu Y, Xue Y, Xiang Y, Yu D, Wang F, Ma Y. 2022. Helixfold: an efficient implementation of AlphaFold2 using PaddlePaddle. arXiv doi:10.48550/arXiv.2207.05477
- Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. 2009. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci* **106**: 67–72. doi:10.1073/pnas.0805923106
- Weissenow K, Heinzinger M, Rost B. 2022a. Protein language model embeddings for fast, accurate, and alignment-free protein structure prediction. *Structure* **30**: 1169–1177.e4. doi:10.1016/j.str.2022.05.001
- Weissenow K, Heinzinger M, Steinegger M, Rost B. 2022b. Ultra-fast protein structure prediction to capture effects of sequence variation in mutation movies. bioRxiv doi:10.1101/2022.11.14.516473
- Wu R, Ding F, Wang R, Shen R, Zhang X, Luo S, Su C, Wu Z, Xie Q, Berger B, et al. 2022. High-resolution de novo structure prediction from primary sequence. bioRxiv doi:10.1101/2022.07.21.500999
- Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. 2020. Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci* **117**: 1496–1503. doi:10.1073/pnas.1914677117
- Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJE. 1987. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J Mol Biol* **195**: 957–961. doi:10.1016/0022-2836(87)90501-8