Structure

Article

# Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction

## Graphical abstract


## Authors

Konstantin Weissenow, Michael Heinzinger, Burkhard Rost

## Correspondence

k.weissenow@tum.de



## In brief

Weissenow et al. leverage protein language models (pLMs) to predict protein structures without using alignments central to state-of-the-art solutions. Speeding up computation more than 10-fold, this method caters to protein design questions, e.g., enabling high-throughput *in silico* point-mutation experiments and predictions for large datasets on almost-laptop-like consumer-grade hardware.


## Highlights

- High-speed protein structure prediction not using alignments
- Predictions for entire human proteome within a week on single machine
- Structure predictions for point mutants correlate with deep mutational scans
- Method, EMBER2, freely available for protein design and stability analysis

Weissenow et al., 2022, Structure *30*, 1169–1177
August 4, 2022 
https://doi.org/10.1016/j.str.2022.05.001

Structure

CellPress

Article

# Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction

Konstantin Weissenow,[1,2,4,*] Michael Heinzinger,[1,2] and Burkhard Rost[1,3]

[1]TUM (Technical University of Munich), Department of Informatics, Bioinformatics and Computational Biology - i12, Boltzmannstr. 3, 85748 Garching/Munich, Germany

[2]TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), Boltzmannstr. 11, 85748 Garching, Germany

[3]Institute for Advanced Study (TUM-IAS), Lichtenbergstr. 2a, 85748 Garching/Munich, Germany & TUM School of Life Sciences Weihenstephan (WZW), Alte Akademie 8, Freising, Germany

[4]Lead contact

*Correspondence: k.weissenow@tum.de

https://doi.org/10.1016/j.str.2022.05.001

## SUMMARY

Advanced protein structure prediction requires evolutionary information from multiple sequence alignments (MSAs) from evolutionary couplings that are not always available. Artificial intelligence (AI)-based predictions inputting only single sequences are faster but so inaccurate as to render speed irrelevant. Here, we described a competitive prediction of inter-residue distances (2D structure) exclusively inputting embeddings from pre-trained protein language models (pLMs), namely ProtT5, from single sequences into a convolutional neural network (CNN) with relatively few layers. The major advance used the ProtT5 attention heads. Our new method, *EMBER2*, which never requires any MSAs, performed similarly to other methods that fully rely on co-evolution. Although clearly not reaching *AlphaFold2*, our leaner solution came somehow close at substantially lower costs. By generating protein-specific rather than family-averaged predictions, *EMBER2* might better capture some features of particular protein structures. Results from using protein engineering and deep mutational scanning (DMS) experiments provided at least a proof of principle for such a speculation.

## INTRODUCTION

### Protein-structure-prediction problem solved

The Critical Assessment of protein Structure Prediction (CASP) has provided the gold standard to evaluate protein structure prediction for almost three decades (Moult et al., 1995). At its first meeting (CASP1, Dec. 1994), the combination of machine learning (ML) and evolutionary information derived from multiple sequence alignments (MSAs) reported a major breakthrough in secondary-structure prediction (Rost and Sander, 1995). This solution expanded into deep-learning inter-residue distances (Jones et al., 2015; Li et al., 2021; Wang et al., 2017), of which the deep dilated residual network(s) of *AlphaFold1* advanced to serve as constraints for subsequent folding pipelines (Kryshtafovych et al., 2019; Senior et al., 2020). 2021's method of the year (Marx, 2022), *AlphaFold2* (Jumper et al., 2021), has combined more advanced artificial intelligence (AI) with more advanced evolutionary information from larger MSAs to essentially solve the protein-structure-prediction problem. *AlphaFold2* predictions directly advance structure determination (Flower and Hurley, 2021). Even this pinnacle of 50 years of research has a shortcoming: predictions are family averaged, not protein specific. On top of that, lack of computing resources may limit proteome-wide predictions; however, predictions will be available for the entire UniProt (Tunyasuvunakool et al., 2021).

All top structure-prediction methods, including *AlphaFold2*, rely on correlated mutations (Marks et al., 2011). Direct coupling analysis (DCA) sharpens this signal (Anishchenko et al., 2017) through pseudolikelihood maximization (Balakrishnan et al., 2011; Seemayer et al., 2014) or through sparse inverse covariance estimation (Jones et al., 2011). Both are challenged by families with low diversity (too little signal) and those with high diversity (too much noise). One solution is to generate multiple MSAs by altering parameters, alignment tools, and databases (Jain et al., 2021; Zhang et al., 2020); these increase runtime (Table 3). Instead of using correlation matrices or Potts parameters derived from MSAs, several recent methods such as *CopulaNet* (Ju et al., 2021) and *rawMSA* (Mirabello and Wallner, 2019) directly process alignments.

### Protein language models (pLMs) decode aspects of the language of life

In analogy to the recent leaps in natural-language processing (NLP), pLMs learn to "predict" masked amino acids given their context using no other annotation than the amino acids for $10^7$–$10^9$ proteins (Alley et al., 2019; Asgari and Mofrad, 2015; Bepler

**Table 1. Performance saturation reached for subset of attention heads (AHs)**

| | MCC (all)[a] | MCC (long range)[a] |
|---|---|---|
| All 768 AHs | 0.30 ± 0.04 | 0.25 ± 0.04 |
| Top 50 AHs | 0.26 ± 0.04 | 0.24 ± 0.04 |
| Top 100 AHs | 0.29 ± 0.04 | 0.24 ± 0.04 |
| Top 120 AHs | 0.29 ± 0.04 | 0.25 ± 0.04 |

[a]Logistic regression (LR) results based on AHs from *ProtT5* for 200 randomly selected training samples for *SetValCASP12*. Methods (rows): first row: results for all 768 AHs from ProtT5; bottom three rows: results for the top 50, top 100, and top 120 most informative AHs, respectively. Performance measures (columns): the ± values indicate ±1.96 standard errors, i.e., 95% confidence interval (CI95; Equation 7) The top 100 AHs reached baseline performance (within the SE).

and Berger, 2019, 2021; Elnaggar et al., 2021; Heinzinger et al., 2019; Madani et al., 2020; Ofer et al., 2021; Rao et al., 2019; Rives et al., 2021; Wu et al., 2021). NLP words/tokens correspond to amino acids in pLMs and sentences to entire proteins. Embeddings extract the information learned by the pLMs. Where NLP embeddings reflect grammar, pLM embeddings decode aspects of the language of life as written in protein sequences (Heinzinger et al., 2019; Ofer et al., 2021). This suffices as exclusive input to many methods predicting aspects of protein structure and function without further pLM optimization through a second step of supervised training (Alley et al., 2019; Asgari and Mofrad, 2015; Elnaggar et al., 2021; Heinzinger et al., 2019; Madani et al., 2020; Rao et al., 2019; Rives et al., 2021) or by refining the pLM through another supervised task (Bepler and Berger, 2019, 2021; Littmann et al., 2021b). Embeddings can outperform homology-based inference based on the traditional sequence comparisons optimized over five decades (Littmann et al., 2021a, 2021b). With little optimization, methods using only embeddings even outperform advanced MSA-based methods (Elnaggar et al., 2021; Stärk et al., 2021). Simple embeddings mirror the last "hidden" states/values of pLMs. Slightly more advanced are weights learned by so-called transformers; in NLP jargon, these are referred to as "attention heads"(Vaswani et al., 2017). These directly capture complex information about protein structure (Rao et al., 2020), e.g., allowing the transformer-based pLM ESM-1b to predict structure without supervision (Rives et al., 2021).

Here, we introduced a novel approach using attention heads (AHs) from pre-trained pLMs to predict inter-residue distances without MSAs at levels of performance similar to methods relying on large MSAs and evolutionary couplings/DCA. Thereby, this approach enables accurate predictions of protein 3D structure substantially faster and at lower computing costs.

## RESULTS AND DISCUSSION

### Top 100 AHs almost as good as all 768 but faster

A logistic regression (LR) system trained on 200 randomly selected samples from the training set *SetTrnProtNet12* and evaluated on the validation set *SetValCASP12* suggested that about one-seventh of the AHs already sufficed to reach the performance of all 768 AHs (Table 1). This reduced storage requirements of pre-computed inputs (from 3.1 TB to 406 GB) and improved training speed when working with the full training set. The fact that a simple LR sufficed highlighted the remarkably strong structural signal readily available from ProtT5 (Elnaggar et al., 2021) AHs. Although trained on only 200 proteins (100-fold smaller than training *SetTrnProtNet12*), that model outperformed convolutional neural networks (CNNs) completely trained on less complex embeddings (Seqvec [Heinzinger et al., 2019] and ProtAlbert [Elnaggar et al., 2021]; Figure 1B).

### AHs clearly improved contact predictions

When trained on the full training set (*SetTrnProtNet12*), even CNNs with few layers performed well when enriching the embeddings through ProtT5 AHs (Figure 1A). Smaller CNNs with 80 ResNet blocks (Figure S1) even reached numerically higher Matthews correlation coefficients (MCCs) than 50% larger CNNs with 120 ResNet blocks (Figure 1A; difference not statistically significant). Nevertheless, all following results were obtained for the less accurate version with 120 ResNet because we tested smaller CNNs after those results had been collected and decided to reduce energy consumption.

Comparing embeddings from different pLMs, Seqvec (based on ELMo [Peters et al., 2018]) and ProtAlbert (based on Albert [Lan et al., 2020], a leaner version of BERT [Devlin et al., 2019]) performed significantly worse than other transformers (Figure 1B). Top were CNNs inputting ProtT5 AHs (based on T5 [Raffel et al., 2020]; Figure 1B). Although never using MSAs, it numerically outperformed our in-house CNN dependent on evolutionary couplings (DCA; Figure 1B). From here on, we refer to the model using ProtT5 AHs and 120 ResNet blocks as our final model *EMBER2*.

Additional input features (STAR Methods) slightly improved for earlier pLMs but not statistically significantly for the final *EMBER2*. As transformers explicitly encode position, adding positional information might become redundant.

Given that the embedding-based, MSA-free *EMBER2* performed similar as our in-house CNN (DCAdst) relying on evolutionary couplings from MSAs, we expected embeddings to perform better for proteins from families with low diversity (weak evolutionary coupling) and worse for those with large diversity (strong evolutionary coupling). Although some evidence supported this expectation (embeddings outperformed evolutionary couplings for very small families), the embeddings also performed better for some very large families with high diversity. While we could not explain this finding, we might speculate that very large families contain so much structural divergence that embedding-based protein-specific predictions outperform family-averaged predictions. If so, at least the most structurally diverged members might be predicted better without MSAs. As methods using evolutionary couplings benefit from immense diversity (Marks et al., 2012), simply constraining "too large families" might not remedy such a shortcoming of MSA-based solutions. If this speculation were partially correct, we would still have no data as to whether this would only affect the performance of some proteins (outliers) or of most (although most define the average, almost all might deviate substantially from the average).

### EMBER2 reached Raptor-X without MSA

For *SetTst29*, we collected C-alpha contact predictions from the publicly available *Raptor-X*, which performed well at CASP

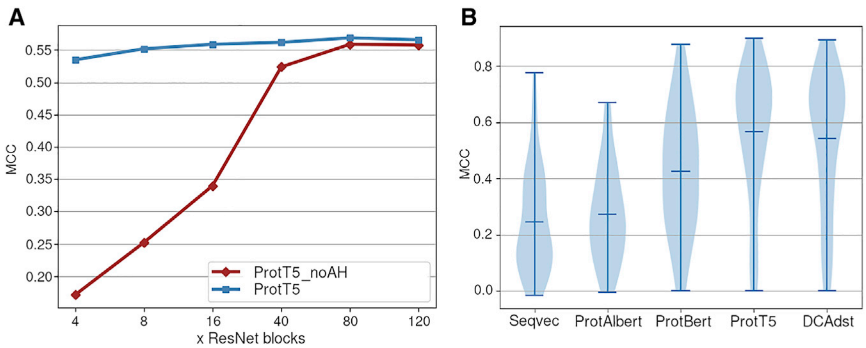


**Figure 1. *ProtT5* attention heads (AHs) best**

y axes: Matthews correlation coefficient (MCC; Equation 4; for medium- and long-range contacts).

(A) *SetValCASP12*: the x axis gives the number of ResNet blocks; values on the left of the x axis describe CNNs with few layers, i.e., those with fewer parameters (ranging from 235,306 for 4 blocks to 6,501,162 for 120 blocks). Both lines used the *ProtT5* pLM: the upper blue line marks the method introduced here with AHs, and the lower red line represents embeddings without AHs. AHs performed well with shallow architectures, while raw embeddings needed ≥40 ResNet blocks to reach MCC-levels >0.5.

(B) *SetValCASP12*: the five violin plots for embeddings from four different pLMs (Elnaggar et al., 2021; Heinzinger et al., 2019) along with our in-house method using evolutionary couplings (DCAdst). The markers indicate highest, lowest, and average MCC, while the width—light blue background cloud—shows the overall distribution (see Figure S2 for more details).

(Wang et al., 2017). With the larger MSAs from today (May 2021) than at CASP12/13, *Raptor-X* likely performed slightly better. Although numerically, the supervised method *EMBER2* using AHs outperformed the version not using AHs (Table S2), this difference was not statistically significant within the 95% confidence interval. *EMBER2* numerically outperformed *Raptor-X* for medium-range contacts (12≤|i-j|≤23); the opposite was the case for long-range contacts (|i-j|>23). None of those differences were statistically significant (Table S2). Overall, the model trained on the top 100 most informative AHs of ESM-1b performed worse than *EMBER2* (Table S2, *EMBER2* versus ESM-1b). Since the supervised ESM-1b-based distance predictions were not made available, we also compared their published performance (Rives et al., 2021) on the CAMEO test set by *trRosetta* (Yang et al., 2020); ProtT5 and ESM-1b performed alike (Table S3).

Comparing the embedding-based approach using AHs and no MSA (*EMBER2*) with the state-of-the-art *Raptor-X* using MSAs and post-processing for evolutionary couplings in detail revealed that MSA-free predictions did perform better for very small families (Figure 2A, darkest points usually above diagonal). For some proteins (e.g., T0960-D2 and T0963D2; Figure 2A, top left), *EMBER2* correctly predicted distances while *Raptor-X* failed; for others (e.g., T1049; Figure 2A, bottom right), the opposite was the case.

## Good 3D structure predictions

The *trRosetta* (Yang et al., 2020) pipeline with pyRosetta (Chaudhury et al., 2010) turned our predicted distance distributions (distograms) into 3D predictions. For the CASP13/14 free-modeling and template-based modeling (TBM)-hard targets (SetTst29), the similarity in the 2D distance prediction performance of *EMBER2* and *Raptor-X* remained essentially similar for the resulting 3D predictions (Figures 2A versus 2B; Table 2). We showcased two proteins with the largest advantage for *EMBER2* in template modeling score (TM-score, (Zhang and Skolnick, 2005, Figure 2C): while both methods predict overall structure with good quality, *Raptor-X* misplaces and swaps helices, resulting in a significant drop in TM-score. As expected, *AlphaFold2* and *RoseTTAFold* outperformed both our approach and *Raptor-X* significantly on 3D predictions (Table 2). Using structural domains, instead of the sequences used in CASP, raised TM-scores about 6% for all methods (Tables S5 and S6).

## Case study: Beta-barrel gene duplication

All known transmembrane beta-barrel proteins, found in the outer membrane of Gram-negative bacteria, feature an even number of between 8 and 36 beta strands (Lauber et al., 2018). For instance OmpX from *Escherichia coli* (outer membrane protein protein X; Swiss-Prot: ompx_ecoli [Boutet et al., 2016]) has an 8-stranded beta barrel. Gene *in vitro* duplication and selective removal of beta hairpins produced new stable beta-barrel proteins, which folded *in vitro* with strand numbers between 8 and 16 (Arnold et al., 2007). We retrained *EMBER2*, excluding proteins in the training set with pairwise sequence identity (PIDE) >25% to OmpX in order to validate our model using the experimental structure (PDB: 1Q9F [Fernández et al., 2004]). *EMBER2* distance predictions refined through trRosetta (Yang et al., 2020) predicted the native OmpX structure accurately reaching a TM score of 0.73 (Figure 3, left). For three of the five engineered variants shown to fold *in vitro* (OmpX64c, OmpX66, and OmpX84), our predicted structures suggested a single larger barrel with 10 and 12 beta strands (Figure 3, three rightmost panels). This was confirmed experimentally (Arnold et al., 2007). As proof of principle, these results suggested that our approach could reliably predict structures of transmembrane proteins that are inherently difficult to predict by comparative modeling and other methods due to their under-representation (Kloppmann et al., 2012; Pieper et al., 2013). The under-representation of membrane proteins in the PDB did not affect the pLMs underlying our predictions because the generation of the pLMs only required sequence information (Elnaggar et al., 2021; Heinzinger et al., 2019) and membrane proteins are likely not under-represented in UniProt (Consortium, 2016).

## EMBER2 mutant predictions weakly correlated with DMS

We expected our protein-specific predictions to be more sensitive to point mutations (dubbed SAV for single amino-acid variants in the following) than family-averaging, MSA-based methods such as *AlphaFold2*. To investigate, we validated predictions on deep mutational scanning (DMS) experiments (Fowler and Fields, 2014) (*DMS5*, STAR Methods). We predicted structures for each possible SAV with *EMBER2* and *AlphaFold2* (via *ColabFold*). For such a large set, *AlphaFold2* predictions require so much computing that we had to choose the five shortest proteins, resulting in 24,639 sequences. For each, we

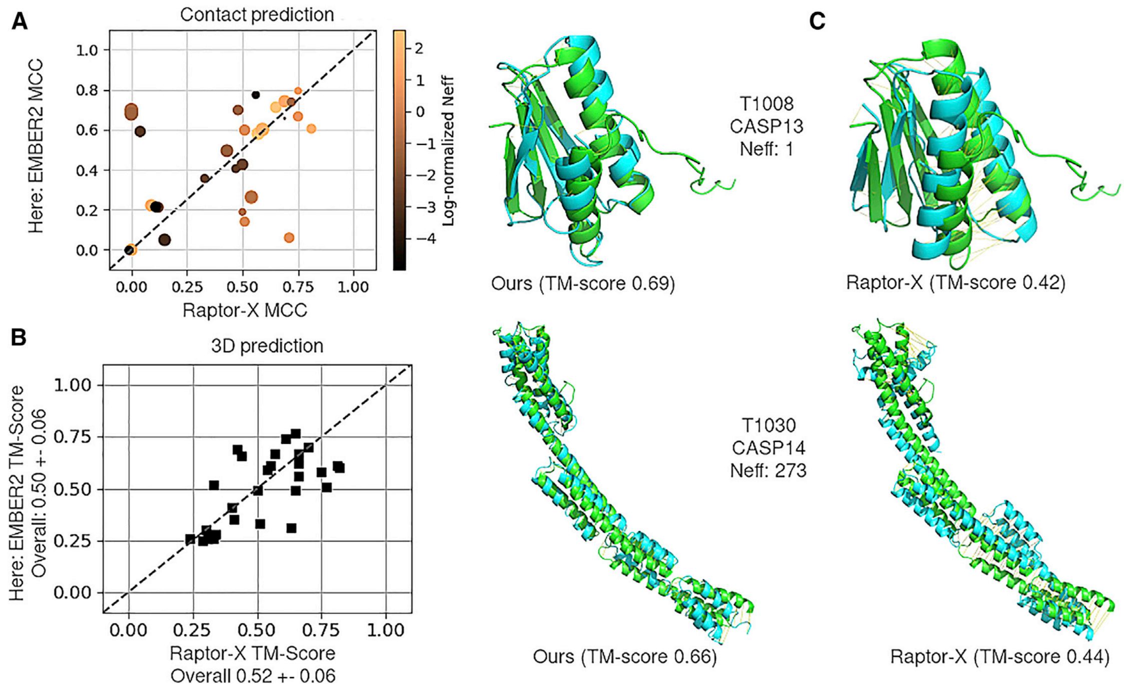


**Figure 2. *EMBER2* beats *Raptor-X* for small families**

Dataset: SetTst29. Methods: *EMBER2* (as introduced here: using AHs without MSAs) and *Raptor-X* (Wang et al., 2017).

(A) Per-protein comparison of MCCs (Equation 5; correlation predicted-observed: higher is better) for medium- and long-range contact predictions.

(B) 3D structure prediction performance: TM-align (Zhang and Skolnick, 2005) computed TM-scores for all predictions (higher numbers indicate better predictions; values <0.17 approach random, values >0.5 indicate that the overall fold is correctly predicted). Overall, the performance was similar for both methods.

(C) Detailed comparison of 3D predictions versus experiment for two proteins (T1008 and T1030; experiment: green, prediction: cyan). One protein gives an example for a small (T1008: PDB: 6MSP: *de-novo*-designed protein FoldIt3 [Koepnick et al., 2019]) and the other for a large (T1030: PDB: 6POO: N-terminal helical domain of BibA [Manne et al., 2020]) family, and for both, *EMBER2* outperformed *Raptor-X*, although overall the performance of these two was similar. For T1008 (CASP13), both predictions captured the overall fold correctly, but *Raptor-X* incorrectly swapped the two helices, reducing the TM score from 0.69 (*EMBER2*) to 0.42. Similarly, for the longer protein T1030 (CASP14), *Raptor-X* misplaced several helices. For both proteins, *EMBER2* predicted above and *Raptor-X* below average, as demonstrated by (B). The ± values indicate ±1.96 standard errors, i.e., 95% confidence interval (CI95; Eq. 7). Images are from PyMol (Schrödinger and DeLano, 2021).

computed the differences in the structures (Equation 8) predicted for mutant and wild type and compared those DMS measures for the effect of SAVs upon protein function (proxied by very different experimental setups for the five experiments).

A small structural change in the binding site can impact binding. Since our coarse-grained perspective of "structural change anywhere" could not capture such changes, the predicted structural impact could, at best, roughly proxy functional impact. Nevertheless, we observed some correlation between predicted structural change (wild type versus point mutant) and DMS scores for *EMBER2* and *AlphaFold2* for most proteins (Figure 4). The protein with the highest correlation for *AlphaFold2* (translation initiation factor IF1 [TIF_IF1] [Kelsic et al., 2016]) had also a relatively high correlation for *EMBER2*. However, the worst *AlphaFold2* set (small ubiquitin-related modifier 1 [SUMO1] [Weile et al., 2017]) was one of the best for *EMBER2*. Overall, *EMBER2* correlated significantly more with DMS than *AlphaFold2* (Figure 4).

Next, we considered the correlation between structure change and DMS exclusively for internal residues (chosen as 50% with highest contact densities based on experimental structures—UBC9: PDB: 2GRR [Yunus and Lima, 2006]; TIF_IF1: PDB: 2N8N [Kim et al., 2017]; SUMO1: PDB: 1A5R [Bayer et al., 1998]; and Hras: PDB: 1AA9 [Ito et al., 1997]). We excluded yeast ubiquitin because of the low coverage of its structures. On this subset of residues of *DMS5*, the correlation with DMS scores increased significantly for *EMBER2* on all but one protein (Table S4).

### Hemagglutinin (HA) predictions might suggest hinge motion

Influenza HA is a viral membrane protein involved in the infection of target cells. The HA2 domain was observed to change conformation to a spike motif at low pH (Caffrey and Lavie, 2021). Hoping to capture aspects of the dynamics of HA, we predicted structures of the wild type of the HA2 domain and all possible SAVs with *EMBER2*. Instead of an "in between" conformation, we obtained a structure closer to the low pH state, which might be due to low quality of the prediction (Figure S5A). *AlphaFold2* did not predict either of the observed conformations (Figure S5D).

Investigating the predicted SAV effects (Figure S5B), the residues around position 110 were predicted to strongly impact structure. The 3D prediction for the strongest-effect point mutant (T111I) suggested a possible hinge motion involved in the conformational flip between states (Figure S5C). It remains unclear to what extent this example captured a generic trend.

Table 2. TM-scores on *SetTst29*

| Method | TM score[a] |
|---|---|
| *EMBER2* (method introduced) | 0.50 ± 0.06 |
| *Raptor-X* | 0.52 ± 0.06 |
| *RoseTTAFold* | 0.81 ± 0.05 |
| *ColabFold* (*AlphaFold2* weights) | 0.79 ± 0.07 |

[a]Data: *SetTst29* combined CASP13 + 14. Methods: *RoseTTAfold/Robetta* (Baek et al., 2021) and *ColabFold/AlphaFold2* (Mirdita et al., 2021). Performance measure (TM score): the ± values indicate ±1.96 standard errors, i.e., 95% confidence interval (CI95; Equation 7). TM scores reflect the performance for the entire sequences as submitted to CASP (not to structural domains, for which TM scores are higher for all methods [Tables S5 and S6]).

## Reducing computation saves resources

The experimental determination of high-resolution protein structures is so costly that good predictions are valuable even when consuming substantial resources. The first compute-intensive tasks of state-of-the-art (SOTA) structure prediction methods is the generation of MSAs along with the processing of evolutionary couplings (Balakrishnan et al., 2011; Marks et al., 2011; Seemayer et al., 2014). Depending on hardware, alignment method, and sequence database, the average time needed to create MSAs varies substantially. For 29 CASP proteins (*SetTst29*), *EMBER2* was almost 100-fold faster than the in-house MSA-based *DCAdst* (Table 3). However, these numbers compared 2D predictions for *EMBER2* and 3D predictions for *ColabFold/AlphaFold2*. Turning 2D into 3D (as used for the comparisons in Figure 2 and Table 2) took extra.

For the ∼25,000 predictions of point mutants, the speed up from *EMBER2* to *ColabFold* was about 35-fold (note: *ColabFold* is more than 10 times faster than the original *AlphaFold2* that it optimizes [Mirdita et al., 2021]). These numbers included all that was needed to correlate predicted structure change with DMS (Figure 4) since the analysis was based on 2D distance maps (which, for *ColabFold*, were computed from their 3D predictions).

The runtime measures included loading pre-trained models (embeddings), amounting to a one-time cost of ∼25 s regardless of the number of proteins predicted. We computed predictions for almost the entire human proteome (proteins with <3,000 residues due to graphics processing unit [GPU] memory limits) in about 8 days using the same hardware. The numbers excluded pLM pre-training (ProtT5) because that method had been made available before we started and has not been changed to predict protein structure.

The time required for development varied substantially with input type and network depth. The final model (*EMBER2*) with 120 ResNet blocks converged after 20 epochs and 35 h (1.75 h/epoch). The smallest architecture sampled with ProtT5 AHs and 4 ResNet blocks converged after 46 epochs and 11 h.

## Distances predicted for 99% of human proteins

Within about a week on a single machine (as specified in Table 3), we predicted inter-residue distances for all human proteins below 3,000 residues (constituting 99% of the human reference proteome); 3D predictions from *AlphaFold2* for all human proteins have been made available before (Tunyasuvunakool et al., 2021). For the subset of the human proteome for which we had predictions from *AlphaFold2* and our method *EMBER2*, our method predicted, on average, lower contact densities, suggesting a tendency in our method not to have fully "folded the protein" (Figure S6). The lower average density was expected given that *EMBER2* is less accurate than *AlphaFold2* (Table 2). Nevertheless, at this point, our protein-specific 2D predictions are the only alternative to *AlphaFold2*, and, as demonstrated by the increase in correlation with DMS for the fraction of the residues with highest contact density (Figure 4 versus Table S4), contact densities might help directly for certain analyses. Similarly, downstream methods inputting structural information might benefit from the simplicity of 2D over 3D. Clearly, our data are readily usable for methods operating directly on contacts/distances (Punta and Rost, 2005).

It remains unclear to what extent more protein-specific versus more family-averaged predictions will matter. The example of the hinge of HA might, or might not, evidence how predicted contacts could become helpful. Overall, embedding-based predictions might be better for protein design (Wu et al., 2021) than those based on MSAs.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

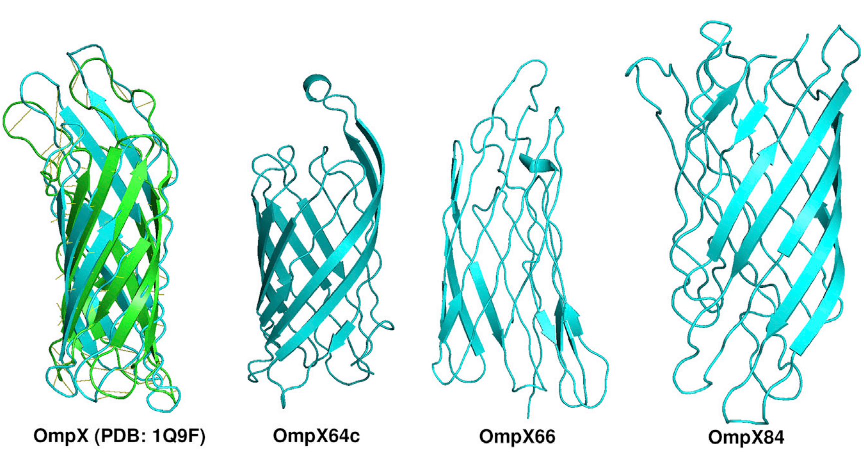


Figure 3. 3D predictions for OmpX and 3 variants (OmpX64x, OmpX66, and OmpX84)
The experimental structure is shown in green (PDB: 1Q9F [Fernández et al., 2004]) and predictions in cyan (images generated using PyMol [Schrödinger and DeLano, 2021]). Prediction and experiment matched with a TM score of 0.73 for the native protein of known structure. While the predictions for the protein-engineered sequence variants (OmpX64x, OmpX66, and OmpX84) suggested less compact structures, our predictions confirmed the experimental findings of larger single beta barrels (Arnold et al., 2007).

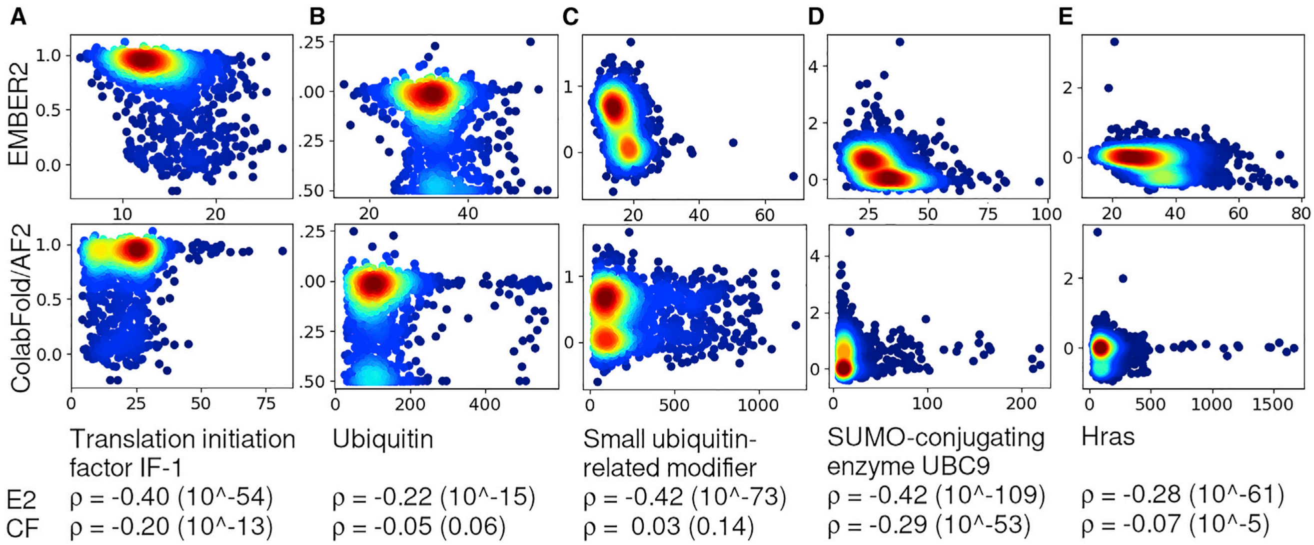


**Figure 4. *EMBER2* predictions correlated better with DMS data than *AlphaFold2***

The x axes give the differences in the predictions of wild type and mutant (Equation 8) for the methods *EMBER2* (introduced here) and *AlphaFold2* (Jumper et al., 2021) as implemented in *ColabFold* (Mirdita et al., 2021); the y axes give the experimentally measured effects for each SAV. The five proteins chosen were the shortest taken from a dataset prepared previously (Bandaru et al., 2017; Kelsic et al., 2016; Mavor et al., 2016; Riesselman et al., 2018; Weile et al., 2017): (A) Translation initiation factor IF-1, (B) ubiquitin, (C) small ubiquitin-related modifier, (D) SUMO-conjugating enzyme UBC9, and (E) Hras. Each point corresponds to a structure prediction with a single amino-acid variant (SAV; i.e., point mutant); color indicates point density. Neither method reached anywhere near expert methods trained on those data (Riesselman et al., 2018), but the protein-specific *EMBER2* consistently outperformed the family-averaged *AlphaFold2*. The Spearman rank correlations and associated p values (in brackets) are given in the table under the plots along with the protein names for each DMS experiment (using the acronym E2 for *EMBER2* and CF for *ColabFold/AlphaFold2* at the left).

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Data sets
  - Protein language models (pLMs)
  - Model architecture
  - Training
  - Input
  - 3D predictions
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Performance measures
  - Error estimates
  - Measure structural difference

**Table 3. Computing resources**

| Method data | Data[a] | Runtime (1,000 s)[b] |
|---|---|---|
| *EMBER2* 2D | SetTst29 | 0.12 |
| DCAdst 2D | SetTst29 | 12.70 |
| *ColabFold* (*AlphaFold2*) 3D | SetTst29 | 6.90 |
| *EMBER2* | DMS5 | 15.66 |
| *ColabFold* (*AlphaFold2*) | DMS5 | 547.20 |

[a]Data: *SetTst29* from CASP13 + 14; *DMS5*: 5 shortest proteins from DMS experiments using 24,639 predictions total. Methods: *EMBER2*: embedding-based method introduced here; *DCAdst*: in-house MSA-based method creating MSAs through *HHblits* (Steinegger et al., 2019) on UniClust30 (2018_8) (Mirdita et al., 2016) to obtain MSAs and *CCMpred* (Seemayer et al., 2014) to generate couplings.

[b]Runtime: measured in multiples of 1,000 s; machines: Intel Xeon Gold 6248 (100 GB RAM) and a single Nvidia Quadro RTX (46 GB VRAM) with all data on a local SSD. Times shown are for *EMBER2* predicting 2D and *ColabFold/AlphaFold2* 3D. However, for the DMS correlations, we used 2D distance maps, which for *ColabFold/AlphaFold2* were computed from their 3D predictions.

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.str.2022.05.001.

## ACKNOWLEDGMENTS

We thank Tim Karl (TUM) for invaluable help with hard- and software and Inga Weise (TUM) for supporting many aspects of this work. Thanks to Jinbo Xu and the *Raptor-X* co-developers (U Chicago) for making their method available; thanks to Jianyi Yang (Nankai U) and his co-developers for publishing the *trRosetta* source code; thanks to Martin Steinegger and Milo Mirdita (Seoul) for making *AlphaFold2* available through *ColabFold*; and particular thanks to the anonymous reviewers who helped considerably to improve this work. This work was supported by the Alexander von Humboldt Foundation (BMBF) and by the German Research Foundation (DFG–GZ: RO1320/4–1). We gratefully acknowledge the support of NVIDIA with the donation of a Titan GPU used for development. Furthermore, the B.R. lab gladly acknowledges support from Google Cloud and the Google Cloud Research Credits program to fund the earlier stages of this project under the COVID19 HPC Consortium grant. Last not least, thanks to all who make their experimental data publicly available and all those who maintain such databases, in particular to Steve Burley and his team at the PDB.

## AUTHOR CONTRIBUTIONS

Conceptualization, K.W. and B.R.; methodology, K.W., M.H., and B.R.; software: K.W. and M.H.; investigation, K.W.; writing – original draft, K.W. and M.H.; writing – review & editing, M.H. and B.R.; supervision, B.R.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 9, 2021
Revised: February 25, 2022
Accepted: April 29, 2022
Published: May 23, 2022

## REFERENCES

Alley, E.C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G.M. (2019). Unified rational protein engineering with sequence-based deep representation learning. Nat. Methods *16*, 1315–1322. https://doi.org/10.1038/s41592-019-0598-1.

AlQuraishi, M. (2019). ProteinNet: a standardized data set for machine learning of protein structure. BMC Bioinformatics *20*, 311. https://doi.org/10.1186/s12859-019-2932-0.

Anishchenko, I., Ovchinnikov, S., Kamisetty, H., and Baker, D. (2017). Origins of coevolution between residues distant in protein 3D structures. Proc. Natl. Acad. Sci. *114*, 9122–9127. https://doi.org/10.1073/pnas.1702664114.

Arnold, T., Poynor, M., Nussberger, S., Lupas, A.N., and Linke, D. (2007). Gene duplication of the eight-stranded beta-barrel OmpX produces a functional pore: a scenario for the evolution of transmembrane beta-barrels. J. Mol. Biol. *366*, 1174–1184. https://doi.org/10.1016/j.jmb.2006.12.029.

Asgari, E., and Mofrad, M.R.K. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS One *10*, e0141287. https://doi.org/10.1371/journal.pone.0141287.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science *373*, 871–876. https://doi.org/10.1126/science.abj8754.

Balakrishnan, S., Kamisetty, H., Carbonell, J.G., Lee, S.I., and Langmead, C.J. (2011). Learning generative models for protein fold families. Proteins *79*, 1061–1078. https://doi.org/10.1002/prot.22934.

Bandaru, P., Shah, N.H., Bhattacharyya, M., Barton, J.P., Kondo, Y., Cofsky, J.C., Gee, C.L., Chakraborty, A.K., Kortemme, T., Ranganathan, R., and Kuriyan, J. (2017). Deconstruction of the Ras switching cycle through saturation mutagenesis. Elife *6*, e27810. https://doi.org/10.7554/elife.27810.

Bayer, P., Arndt, A., Metzger, S., Mahajan, R., Melchior, F., Jaenicke, R., and Becker, J. (1998). Structure determination of the small ubiquitin-related modifier SUMO-1. J. Mol. Biol. *280*, 275–286. https://doi.org/10.1006/jmbi.1998.1839.

Bepler, T., and Berger, B. (2019). Learning protein sequence embeddings using information from structure. Preprint at arXiv. arXiv:1902.08661. https://doi.org/10.48550/arXiv.1902.08661.

Bepler, T., and Berger, B. (2021). Learning the protein language: evolution, structure, and function. Cell Syst. *12*, 654–669.e3. https://doi.org/10.1016/j.cels.2021.05.017.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A.J., Poux, S., Bougueleret, L., and Xenarios, I. (2016). UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. In Plant Bioinformatics: Methods and Protocols, D. Edwards, ed. (Springer), pp. 23–54. https://doi.org/10.1007/978-1-4939-3167-5_2.

Burley, S.K., Berman, H.M., Kleywegt, G.J., Markley, J.L., Nakamura, H., and Velankar, S. (2017). Protein data bank (PDB): the single global macromolecular structure archive. Methods Mol. Biol. *1607*, 627–641. https://doi.org/10.1007/978-1-4939-7000-1_26.

Caffrey, M., and Lavie, A. (2021). pH-dependent mechanisms of influenza infection mediated by hemagglutinin. Front. Mol. Biosci. *8*, 777095. https://doi.org/10.3389/fmolb.2021.777095.

Chaudhury, S., Lyskov, S., and Gray, J.J. (2010). PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics *26*, 689–691. https://doi.org/10.1093/bioinformatics/btq007.

Consortium, T.U. (2016). UniProt: the universal protein knowledgebase. NAR *45*, D158–D169. https://doi.org/10.1093/nar/gkw1099.

Dallago, C., Schütze, K., Heinzinger, M., Olenyi, T., Littmann, M., Lu, A.X., Yang, K.K., Min, S., Yoon, S., Morton, J.T., and Rost, B. (2021). Learned embeddings from deep learning to visualize and predict protein sets. Curr. Protoc. *1*, e113. https://doi.org/10.1002/cpz1.113.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.04805.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Yu, W., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. (2021). ProtTrans: towards cracking the language of lifes code through self-supervised deep learning and high performance computing. IEEE Trans. Pattern Anal. Mach Intell. https://doi.org/10.1109/TPAMI.2021.3095381.

Fernández, C., Hilty, C., Wider, G., Güntert, P., and Wüthrich, K. (2004). NMR structure of the integral membrane protein OmpX. J. Mol. Biol. *336*, 1211–1221. https://doi.org/10.1016/j.jmb.2003.09.014.

Flower, T.G., and Hurley, J.H. (2021). Crystallographic molecular replacement using an in silico-generated search model of SARS-CoV-2 ORF8. Protein Sci. *30*, 728–734. https://doi.org/10.1002/pro.4050.

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nat. Methods *11*, 801–807. https://doi.org/10.1038/nmeth.3027.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., and Rost, B. (2019). Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics *20*, 723. https://doi.org/10.1186/s12859-019-3220-8.

Hopf, T.A., Green, A.G., Schubert, B., Mersmann, S., Schärfe, C.P.I., Ingraham, J.B., Toth-Petroczy, A., Brock, K., Riesselman, A.J., Palmedo, P., et al. (2019). The EVcouplings Python framework for coevolutionary sequence analysis. Bioinformatics *35*, 1582–1584. https://doi.org/10.1093/bioinformatics/bty862.

Ito, Y., Yamasaki, K., Iwahara, J., Terada, T., Kamiya, A., Shirouzu, M., Muto, Y., Kawai, G., Yokoyama, S., Laue, E.D., et al. (1997). Regional polysterism in the GTP-bound form of the human c-Ha-Ras protein. Biochemistry *36*, 9109–9119. https://doi.org/10.1021/bi970296u.

Jain, A., Terashi, G., Kagaya, Y., Maddhuri Venkata Subramaniya, S.R., Christoffer, C., and Kihara, D. (2021). Analyzing effect of quadruple multiple sequence alignments on deep learning based protein inter-residue distance prediction. Sci. Rep. *11*, 7574. https://doi.org/10.1038/s41598-021-87204-z.

Jones, D.T., Buchan, D.W.A., Cozzetto, D., and Pontil, M. (2011). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics *28*, 184–190. https://doi.org/10.1093/bioinformatics/btr638.

Jones, D.T., and Kandathil, S.M. (2018). High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics *34*, 3308–3315. https://doi.org/10.1093/bioinformatics/bty341.

Jones, D.T., Singh, T., Kosciolek, T., and Tetchner, S. (2015). MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics *31*, 999–1006. https://doi.org/10.1093/bioinformatics/btu791.

Ju, F., Zhu, J., Shao, B., Kong, L., Liu, T.-Y., Zheng, W.-M., and Bu, D. (2021). CopulaNet: learning residue co-evolution directly from multiple sequence alignment for protein structure prediction. Nat. Comm. *12*, 2535. https://doi.org/10.1038/s41467-021-22869-8.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kelsic, E.D., Chung, H., Cohen, N., Park, J., Wang, H.H., and Kishony, R. (2016). RNA structural determinants of optimal codons revealed by MAGE-seq. Cell Syst. *3*, 563–571.e6. https://doi.org/10.1016/j.cels.2016.11.004.

Kim, D.H., Kang, S.J., Lee, K.Y., Jang, S.B., Kang, S.M., and Lee, B.J. (2017). Structure and dynamics study of translation initiation factor 1 from Staphylococcus aureus suggests its RNA binding mode. BBA Proteins Proteom. *1865*, 65–75. https://doi.org/10.1016/j.bbapap.2016.10.009.

Kloppmann, E., Punta, M., and Rost, B. (2012). Structural genomics plucks high-hanging membrane proteins. Cur Opin. Struct. Biol. *22*, 326–332. https://doi.org/10.1016/j.sbi.2012.05.002.

Koepnick, B., Flatten, J., Husain, T., Ford, A., Silva, D.A., Bick, M.J., Bauer, A., Liu, G., Ishida, Y., Boykov, A., et al. (2019). De novo protein design by citizen scientists. Nature *570*, 390–394. https://doi.org/10.1038/s41586-019-1274-4.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)—round XIII. Proteins: Struct. Funct. Bioinformatics *87*, 1011–1020. https://doi.org/10.1002/prot.25823.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). ALBERT: a lite BERT for self-supervised learning of language representations. Preprint at arXiv. arXiv:1909.11942.

Lauber, F., Deme, J.C., Lea, S.M., and Berks, B.C. (2018). Type 9 secretion system structures reveal a new protein transport mechanism. Nature *564*, 77–82. https://doi.org/10.1038/s41586-018-0693-y.

Li, Y., Zhang, C., Bell, E.W., Zheng, W., Zhou, X., Yu, D.-J., and Zhang, Y. (2021). Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. PLOS Comput. Biol. *17*, e1008865. https://doi.org/10.1371/journal.pcbi.1008865.

Littmann, M., Bordin, N., Heinzinger, M., Schütze, K., Dallago, C., Orengo, C., and Rost, B. (2021a). Clustering FunFams using sequence embeddings improves EC purity. Bioinformatics *37*, 3449–3455. https://doi.org/10.1093/bioinformatics/btab371.

Littmann, M., Heinzinger, M., Dallago, C., Olenyi, T., and Rost, B. (2021b). Embeddings from deep learning transfer GO annotations beyond homology. Sci. Rep. *11*, 1160. https://doi.org/10.1038/s41598-020-80786-0.

Madani, A., McCann, B., Naik, N., Shirish Keskar, N., Anand, N., Eguchi, R.R., Huang, P., and Socher, R. (2020). ProGen: language modeling for protein generation. Preprint at arXiv. arXiv:2004.03497.

Manne, K., Chattopadhyay, D., Agarwal, V., Blom, A.M., Khare, B., Chakravarthy, S., Chang, C., Ton-That, H., and Narayana, S.V.L. (2020). Novel structure of the N-terminal helical domain of BibA, a group B streptococcus immunogenic bacterial adhesin. Acta Crystallogr. D Struct. Biol. *76*, 759–770. https://doi.org/10.2210/pdb6poo/pdb.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS One *6*, e28766. https://doi.org/10.1371/journal.pone.0028766.

Marks, D.S., Hopf, T.A., and Sander, C. (2012). Protein structure prediction from sequence variation. Nat. Biotechnol. *30*, 1072–1080. https://doi.org/10.1038/nbt.2419.

Marx, V. (2022). Method of the year: protein structure prediction. Nat. Methods *19*, 5–10. https://doi.org/10.1038/s41592-021-01359-1.

Mavor, D., Barlow, K., Thompson, S., Barad, B.A., Bonny, A.R., Cario, C.L., Gaskins, G., Liu, Z., Deming, L., Axen, S.D., et al. (2016). Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. Elife *5*, e15802. https://doi.org/10.7554/elife.15802.

Mirabello, C., and Wallner, B. (2019). rawMSA: end-to-end deep learning using raw multiple sequence alignments. PLoS One *14*, e0220182. https://doi.org/10.1371/journal.pone.0220182.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2021). ColabFold - making protein folding accessible to all. Preprint at bioRxiv. https://doi.org/10.1101/2021.08.15.456425.

Mirdita, M., von den Driesch, L., Galiez, C., Martin, M.J., Söding, J., and Steinegger, M. (2016). Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. *45*, D170–D176. https://doi.org/10.1093/nar/gkw1081.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2018). Critical assessment of methods of protein structure prediction (CASP)-Round XII. Proteins *86*, 7–15. https://doi.org/10.1002/prot.25415.

Moult, J., Pedersen, J.T., Judson, R., and Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. Proteins *23*. ii–v. https://doi.org/10.1002/prot.340230303.

Ofer, D., Brandes, N., and Linial, M. (2021). The language of proteins: NLP, machine learning & protein sequences. Comput. Struct. Biotechnol. J. *19*, 1750–1758. https://doi.org/10.1016/j.csbj.2021.03.022.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. Preprint at arXiv. 1802.05365.

Pieper, U., Schlessinger, A., Kloppmann, E., Chang, G.A., Chou, J.J., Dumont, M.E., Fox, B.G., Fromme, P., Hendrickson, W.A., Malkowski, M.G., et al. (2013). Coordinating the impact of structural genomics on the human $\alpha$-helical transmembrane proteome. Nat. Struc Mol. Biol. *20*, 135–138. https://doi.org/10.1038/nsmb.2508.

Punta, M., and Rost, B. (2005). Protein folding rates estimated from contact predictions. J. Mol. Biol. *348*, 507–512. https://doi.org/10.1016/j.jmb.2005.02.068.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P.J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Preprint at arXiv. https://doi.org/10.48550/arXiv.1910.10683.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., and Song, Y.S. (2019). Evaluating protein transfer learning with TAPE. Preprint at arXiv. https://doi.org/10.48550/arXiv.1906.08230.

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

Riesselman, A.J., Ingraham, J.B., and Marks, D.S. (2018). Deep generative models of genetic variation capture the effects of mutations. Nat. Methods *15*, 816–822. https://doi.org/10.1038/s41592-018-0138-4.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., and Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. PNAS *118*. https://doi.org/10.1073/pnas.2016239118.

Rost, B., and Sander, C. (1995). Progress of 1D protein structure prediction at last. Proteins: Struct. Funct. Genet. *23*, 295–300. https://doi.org/10.1002/prot.340230304.

Schrödinger, L., and DeLano, W. (2021). The pymol molecular graphics system. http://www.pymol.org/pymol.

Seemayer, S., Gruber, M., and Söding, J. (2014). CCMpred–fast and precise prediction of protein residue-residue contacts from correlated mutations. Bioinformatics *30*, 3128–3130. https://doi.org/10.1093/bioinformatics/btu500.

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. Nature *577*, 706–710. https://doi.org/10.1038/s41586-019-1923-7.

Stärk, H., Dallago, C., Heinzinger, M., and Rost, B. (2021). Light attention predicts protein location from the language of life. Preprint at bioRxiv. https://doi.org/10.1101/2021.04.25.441334.

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., and Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics *20*, 473. https://doi.org/10.1186/s12859-019-3019-7.

Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechn *35*, 1026–1028. https://doi.org/10.1038/nbt.3988.

Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., and Wu, C.H.; the UniProt Consortium (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics *31*, 926–932. https://doi.org/10.1093/bioinformatics/btu739.

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, 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. https://doi.org/10.1038/s41586-021-03828-1.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. Preprint at arXiv. https://doi.org/10.48550/arXiv.1706.03762.

Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. (2017). Accurate de novo prediction of protein contact map by ultra-deep learning model. PLOS Comput. Biol. *13*, e1005324. https://doi.org/10.1371/journal.pcbi.1005324.

Weile, J., Sun, S., Cote, A.G., Knapp, J., Verby, M., Mellor, J.C., Wu, Y., Pons, C., Wong, C., Lieshout, N., et al. (2017). A framework for exhaustively mapping functional missense variants. Mol. Syst. Biol. *13*, 957. https://doi.org/10.15252/msb.20177908.

Wu, Z., Johnston, K.E., Arnold, F.H., and Yang, K.K. (2021). Protein sequence design with deep generative models. Curr. Opin. Chem. Biol. *65*, 18–27. https://doi.org/10.1016/j.cbpa.2021.04.004.

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. *117*, 1496–1503. https://doi.org/10.1073/pnas.1914677117.

Yu, F., and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. Preprint at arXiv. https://doi.org/10.48550/arXiv.1511.07122.

Yunus, A.A., and Lima, C.D. (2006). Lysine activation and functional analysis of E2-mediated conjugation in the SUMO pathway. Nat. Struct. Mol. Biol. *13*, 491–499. https://doi.org/10.1038/nsmb1104.

Zhang, C., Zheng, W., Mortuza, S.M., Li, Y., and Zhang, Y. (2020). DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics *36*, 2105–2112. https://doi.org/10.1093/bioinformatics/btz863.

Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. *33*, 2302–2309. https://doi.org/10.1093/nar/gki524.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Deposited data | | |
| ProteinNet12 | (AlQuraishi, 2019) | https://github.com/aqlaboratory/proteinnet |
| UniRef100 | (Suzek et al., 2015) | https://www.uniprot.org/help/uniref |
| Outer membrane protein OmpX (E.coli) structure | (Fernández et al., 2004) | PDB: 1Q9F |
| Deep mutational scan data | (Riesselman et al., 2018) | https://github.com/debbiemarkslab/DeepSequence |
| CASP sets | (Kryshtafovych et al., 2019) | https://predictioncenter.org/ |
| CAMEO evaluation set | (Yang et al., 2020) | https://yanglab.nankai.edu.cn/trRosetta/ |
| Inter-residue contact and distance predictions for the human reference proteome | This paper | https://doi.org/10.5281/zenodo.6461213 |
| Software and algorithms | | |
| EMBER2 protein structure prediction model | This paper | https://doi.org/10.5281/zenodo.6412497 |
| ProtT5 protein language model | (Elnaggar et al., 2021) | https://github.com/agemagician/ProtTrans |
| MMseqs2 | (Steinegger and Söding, 2017) | https://github.com/soedinglab/MMseqs2 |
| CCMpred | (Seemayer et al., 2014) | https://github.com/soedinglab/CCMpred |
| PyMOL 2.5 | (Schrödinger and DeLano, 2021) | https://pymol.org/2/ |
| Bio Embeddings | (Dallago et al., 2021) | https://github.com/sacdallago/bio_embeddings |
| Python v3 | Python Software Foundation | https://www.python.org |

### RESOURCE AVAILABILITY

#### Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Konstantin Weißenow (k.weissenow@tum.de).

#### Materials availability

This study did not generate new unique reagents.

#### Data and code availability

- Data: Predicted inter-residue contacts and distances for the human reference proteome have been deposited at https://rostlab.org/%7Econpred/ProtT5dst/pred_all_human/as well as https://github.com/kWeissenow/EMBER2_human and are freely publicly available as of the date of publication. The DOI is listed in the key resources table.
- Source code/Methods: The original code and trained weights to run our model have been deposited at https://github.com/kWeissenow/ProtT5dst and are freely publicly available as of the date of publication. The DOI is listed in the key resources table.
- Other: Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

### EXPERIMENTAL MODEL AND SUBJECT DETAILS

All data are generated from the dataset provided in the Key resources table.

### METHOD DETAILS

#### Data sets

We obtained 77,864 three-dimensional (3D) structures from *ProteinNet12* (AlQuraishi, 2019) compiled from the PDB (Burley et al., 2017) before the CASP12 submission deadline (Moult et al., 2018) to replicate the CASP12 conditions. To save energy, we trained on redundancy-reduced data letting MMseqs2 (Steinegger and Söding, 2017) select representatives at 20% pairwise sequence identity (PIDE), ultimately using 21,240 of 77,864 proteins (SetTrnProtNet12).

*ProteinNet12* included a validation set with 41 protein chains from the CASP12 model optimization targets (SetValCASP12). For assessment, we used the so called "template-based modeling-hard" (TBM-hard) and "free-modeling" (FM) targets from CASP13/14 (Kryshtafovych et al., 2019) with publicly available experimental structures (CASP13: 13, CASP14: 16; combined into *SetTst29*). To compare to ESM-1b, we added a CAMEO set provided by trRosetta developers (Yang et al., 2020).

To compare, we also trained a model on evolutionary couplings using the *ProteinNet12* MSAs (DCAdst); EVcouplings (Hopf et al., 2019) generated the alignments against UniRef100 (Suzek et al., 2015) with bitscore thresholds 0.1-0.7. CCMpred (Seemayer et al., 2014) optimized Potts model hyperparameters.

DMS5 marked the five shortest proteins with experimental deep mutational scanning (DMS) (Fowler and Fields, 2014) data prepared to develop the SOTA method for predicting sequence variation effects, namely *DeepSequence* (Riesselman et al., 2018).

### Protein language models (pLMs)

As input to predict inter-residue distances, we compared two types of hidden states derived from pre-trained pLMs: (1) The hidden state output by the last layer of the pLM (for SeqVec (Heinzinger et al., 2019) the last LSTM layer; for the transformer-based models, ProtBert, ProtAlbert, and ProtT5 (Elnaggar et al., 2021), the last attention layer), or (2) the scores for each attention head (AH) of transformers (not for SeqVec). The latter benefitted from the comparisons between all residues in proteins generating an LxL representation for a protein of length L. As detailed elsewhere (Elnaggar et al., 2021), we used only the Encoder-part of ProtT5, creating embeddings in half-precision for speed-up.

When training on AHs extracted from ProtT5, the resulting pairwise tensors of dimension LxLx768 (24 attention layers with 32 AHs each yield 768 attention score matrices of LxL) require immense memory and substantially prolong training. To save resources, we trained a logistic regression (LR) on 200 random samples from SetTrnProtNet12 to predict distance probability distributions, evaluated performance on medium- and long-range contact performance for the CASP12 validation set and selected the Top-50, Top-100 and Top-120 AHs based on the absolute value of the LR weights. Following others (Rao et al., 2020), we enforced symmetry in the attention scores and applied average product correction (APC). For each AH of dimension LxL, we computed the APC as follows:

$$F_{ij}^{APC} = F_{ij} - \frac{F_i F_j}{F} \qquad \text{(Equation 1)}$$

where $F_i$ and $F_j$ were the sums over the i-th row and j-th column, and $F$ the sum over the full matrix.

### Model architecture

Irrespective of the input, our deep learning (DL) models consisted of deep dilated residual networks similar to *AlphaFold1* (Senior et al., 2020). Each residual block consisted of three consecutive layers (Figure S1): (1) a convolution with kernel size 1 reduced the number of feature channels from 128 used in the residual connections to 64, (2) a dilated convolution with kernel size 3 (Yu and Koltun, 2015), and (3) a convolution scaling the number of feature channels back up to 128. The dilation factor cycled between 1, 2, 4 and 8 in successive residual blocks. In each layer, we used batch normalization, followed by exponential linear units (ELU) for non-linearity (Figure S1). Expecting the optimal number of residual blocks necessary to vary for different inputs, we tried depths between 4 and 220 blocks.

Inputting co-evolution information, a narrow window/square around a pair of residues sufficed to correctly infer contacts (Jones and Kandathil, 2018). Like *AlphaFold1*, we addressed this through cropping, i.e. by training and evaluating on patches of 64x64 residue pairs extracted from the full distance map.

The two input types, 1D protein embeddings (string of numbers) and 2D AHs (matrix of probabilities), required two different architectures. The model predicting distances from 2D AHs and evolutionary couplings resembled *AlphaFold 1* (Figure S1), that inputting 1D embeddings accounted for the change in input shapes as follows. (A) The architecture for 1D embeddings used residual blocks (Figures S1 and S3), with 1D convolutions for the first half of all residual blocks (for N residual blocks, N/2 were 1D convolutions, the other N/2 blocks were 2D convolutions). (B) Between the 1D and 2D parts, the 1D representations with length L were expanded to pairwise representations of LxL (Figure S4).

Our models inferred a distance probability distribution (distogram) over 42 bins representing distance intervals 0-22 Ångstrøm (0-2.2 nm). The 40 central bins represented distance intervals of 0.5 Å, the first 0-2Å and the last everything else (>22 Å). To convert distances into contacts (binding/not), we summed the predicted probabilities of the first 14 bins representing distances below 8 Å (Wang et al., 2017).

We trained deep learning systems on protein embeddings from a variety of pLMs, and for comparison on co-evolution. When using co-evolution, ProtBert-BFD (subsequently referred to as *ProtBert*) *Seqvec* or *ProtAlbert* as input, 220 residual blocks were needed. Using *ProtT5-U50* (subsequently referred to as *ProtT5*) we could already reach peak performance with 80 blocks (Figure 1).

We trained on non-overlapping crops, including patches up to 32 residues off-edge with zero-padding and masking at the edges. To avoid introducing bias by similar structural motifs in the protein ends, we randomly picked the initial offsets for each training sample between −32 and 0 (Senior et al., 2020).

We used overlapping crops with a stride of 32 for evaluation (cross-training, i.e. hyper-parameter optimization) and 16 for testing, i.e. estimating performance to speed up training without affecting testing. Predictions for residue pairs were averaged across patches to obtain full distance maps. As distances near the crop center were predicted better, we weighted overlapping predictions through a Gaussian kernel, emphasizing central pairs.

### Training

We trained our models on our local cluster using NVidia Quadro RTX GPUs with 48 GB of VRAM. We used the Adamax optimizer with an initial learning rate of 10^-2, cross-entropy loss over the 42 distance bins and a batch size of 75. We stopped early and saved the best model when the MCC on our validation set (CASP12) did not improve over ten iterations.

### Input

The main input were either representations from the pLMs, or the co-evolution signal as Potts model parameters for comparison to a baseline (DCAdst). To both, we added normalized residue positions (relative position in the protein between 0 and 1), normalized protein length and the log-normalized number of effective sequences as additional input channels. We also masked residues not resolved experimentally, both as single amino acid input and as residue pair during the loss computation.

### 3D predictions

We used pyRosetta (Chaudhury et al., 2010) to compute 3D structures by using a modified version of the trRosetta folding protocol (Yang et al., 2020). In contrast to trRosetta, we dropped any constraints on angular information and adapted the script to use C-alpha rather than C-beta distances as constraints. We first generated 150 coarse-grained decoys using short-, mid- and long-range distances from the predicted distograms at varying levels of distance probability thresholds (here: [0.05, 0.5]) as constraints and relaxed the top-50 models through pyRosetta's FastRelax protocol. The final model and decoys were selected based on the lowest total energy reported by Rosetta.

For comparison with MSA-based SOTA methods, we obtained 3D models and distance predictions for the test sets (SetTst29) from Raptor-X (Wang et al., 2017) (accessed June 2021). We submitted the query sequences instead of MSAs to allow Raptor-X internal optimization. For additional comparisons, we computed predictions from our local *ColabFold* installation with *AlphaFold2* weights (Mirdita et al., 2021) and obtained predictions through the Robetta webserver from RoseTTAFold (Baek et al., 2021).

## QUANTIFICATION AND STATISTICAL ANALYSIS

### Performance measures

We assessed performance through the metrics established by CASP, including precision (Equation 2), recall (Equation 3), F1-score (Equation 4), Matthew's correlation coefficient (MCC, Equation 5) and Top-L precision measuring the positive predictive value for the L long-range contacts predicted with the highest probability (L: protein length). Specifically, we reported performance for the top L/1, L/2, L/5 and L/10 residue pairs per protein. We adopted the common thresholds of >4 and >23 residues sequence separation to define medium- and long-range contacts respectively and omitted evaluating short-range contacts ($|i\text{-}j| \leq 4$).

$$P = Precision = 100\frac{TP}{TP+FP} \quad \text{(Equation 2)}$$

$$R = Recall = 100\frac{TP}{TP+FN} \quad \text{(Equation 3)}$$

$$F1 = 2*\frac{precision*recall}{precision+recall} \quad \text{(Equation 4)}$$

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad \text{(Equation 5)}$$

The resulting 3D predictions were assessed through TM-align (Zhang and Skolnick, 2005).

### Error estimates

Standard errors computed as usual:

$$stderr = \frac{stddev}{\sqrt{n}} \quad \text{(Equation 6)}$$

With n as the number of proteins, and *stddev* as the standard deviation obtained by NumPy (Harris et al., 2020). We reported the 95% confidence interval (CI95), i.e. 1.96 standard errors in results:

$$CI95 = 1.96 * stderr \quad \text{(Equation 7)}$$

### Measure structural difference

We computed the difference between structures from their distance maps $d_1$ and $d_2$ as the sum of absolute differences of distances for all residue pairs i,j:

$$\Delta s = \sum_{i,j} |d_1 - d_2| \quad \text{(Equation 8)}$$