# Highly accurate protein structure prediction with AlphaFold

* Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics.
* Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence - the structure prediction component of the ‘protein folding problem’ – has been an important open research problem for more than 50 years.
* Even recent development in AI the existing methods fall far short of atomic accuracy, especially when no homologous structure is available.
* In this paper the researchers provide the first computational method that can regularly predict the protein structures with atomic accuracy, especially when no homologous structure is available.
* The researchers provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known.
* They validated an entirely redesigned version of their neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction, demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods.
* The latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm.
* The physical interaction program heavily integrates our understanding of molecular driving forces into either thermodynamic or kinetic simulation of protein physics or statistical approximations thereof.
* It may seem appealing but this approach has proven high challenging for even moderate-sized proteins due to the computational intractability of molecular simulation, the context dependence of protein stability of molecular simulation, the context dependence of protein stability and the difficulty of producing sufficiently accurate model of protein physics.
* Despite many technical advancements has been done in recent time contemporary physical and evolutionary-history-based approaches produce predictions that are far short of experimental accuracy in the majority of cases in which a close homologue has not been solved experimentally and this has limited their utility for many biological applications.
* The neural network AlphaFold that researcher developed was entered into the CASP14 assessment.
* AlphaFold structures had a median backbone accuracy of 0.96 Å r.m.s.d. (C root-mean-square deviation at 95% residue coverage) (95% confidence interval=0.85-1.16 Å) whereas the next best performing method had a median backbone accuracy of 2.8 Å r.m.d. (95% confidence interval = 2.7-4.0 Å).
* AlphaFold is able to produce highly accurate side chains when the backbone is highly accurate and considerably improves over template-based methods even when strong templates are available.
* The researcher’s method is scalable to very long proteins with accurate domains and domain-packing. Also, the model is able to provide precise, per-residue estimates of its reliability that should enable the confident use to these predictions.
* AlphaFold demonstrated in CASP14 extends to a large sample of recently released PDB structures; in this dataset, all structures were deposited in the PDB after the training data cut-off and are analyzed as full chains.
* The researchers observe that high side-chain accuracy when the backbone prediction is accurate and they show that their confidence measure, the predicted local-distance test (pLDDT) reliably predicts the C local-distance difference test (IDDT-C) accuracy of the corresponding prediction.
* They also find that the global superposition metric template modelling score (TM-score) can be accurately estimated.
* These analyses validated that the high accuracy and reliability of AlphaFold on CASP14 proteins also transfers to an uncurated collection of recent PDB submissions, as would be expected for confirmation that this high accuracy extends to new folds.
* The AlphaFold network: - AlphaFold greatly improves the accuracy of structure prediction by incorporating novel neural network architectures and training procedures based on the evolutionary, physical and geometric constraints of protein structures.
* The researchers demonstrate new architecture to jointly embed multiple sequence alignments (MSAs) and pairwise features, a new output representation and associated loss that enable accurate end-to-end structure prediction, a new equivariant attention architecture, use of intermediate losses to achieve iterative refinement of predictions, masked MSA loss to jointly train with the structure, learning from unlabelled protein sequences using self-distillation and self-estimates of accuracy.
* The AlphaFold network directly predicts the 3D coordinates of all heavy atoms for a given protein using the primary amino acid sequence and aligned sequence of homologues as input.
* The network comprises two main stages: - first, the trunk of the network processes the inputs through repeated layers of a novel neural network block which the researcher’s term as Evoformer to produce an *N*seq x *N*res array that represents a processed MSA and an *N*res x *N*res array that represents residue pairs. The MSA representations is initialized with the raw MSA.
* The Evoformer blocks contain a number of attention-based and non-attention-based components.
* The key innovations in the Evoformer blocks are new mechanism to exchange information within the MSA and pair representations that enables direct reasoning about the spatial and evolutionary relationships.
* The trunk of the network is followed by the structure module that introduces an explicit 3D structure in the form of a rotation and translation for each residue of the protein.
* The key innovations in this section of the network include breaking the chain structure, a novel equivariant transformer to allow the network to implicitly reason about the unrepresented side-chain atoms and a loss term that places substantial weight on the orientational correctness of the residues.
* The researcher reinforces the notion of iterative refinement be repeatedly applying the final loss to outputs and then feeding the outputs recursively into the same modules.
* Evoformer: - the key principle of the building block of the network – named Evoformer - is to view the prediction of protein structures as a graph inference problem in 3D space in which the edges of the graph are defined by residues in proximity.
* The elements of the pair representation encode the information about the relation between the residues.
* The columns of the MSA representation encode the individual residues of the input sequence while the rows represent the sequences in which those residues appear.
* The MSA representation updates the pair representation through an element-wise outer product that is summed over the MSA sequence dimension.
* Within the pair representation, there are two different update patterns. Both are inspired by the necessity of consistency of the pair representation – for a pairwise description of amino acids to be satisfied including the triangle inequality on distances.
* Based on this intuition, the researchers arranged the update operations on the pair representation in terms of triangles of edges involving three different nodes.
* The researchers add an extra logit bias to axial attention to include the ‘missing edge’ of the triangle and the researchers define a non-attention update operation ‘triangle multiplicative update’ that uses two edges to update the missing third edge.
* The researchers also use a variant of axial attention within the MSA representation.
* End-to-end structure prediction: - the structure module operates on a concrete 3D backbone structure using the pair representation and the original sequence row of the MSA representation from the trunk.
* The 3D backbone structure is represented as *Nres* independent rotations and translations, each with respect to the global frame.
* These rotations and translations – representing the geometry of the N-C-C atoms – prioritize the orientation of the protein backbone so that the location of the side chain of each residue is highly constructed within that frame.
* The peptide bond geometry is completely unconstrained and the network is observed to frequently violate the chain constraint during the application of the structure module as breaking this constraint during the application of the structure module as breaking these constraints enables the local refinement of all parts if the chain without solving complex loop closure problems.
* Satisfaction of the peptide bond geometry is encouraged during fine-tuning by a violation loss term.
* Exact enforcement of peptide bond geometry is only achieved in the post-prediction relaxation of the structure by gradient descent in the Amber force field.
* The residue gas representation is updated iteratively in two stages. First, a geometry aware attention operation that the researcher’s term ‘invariant point attention’ (IPA) is used to update an *N*res set a neural activation without changing the 3D positions, then an equivariant update operation is performed on the residue gas using the updated activations.
* The 3D queries and keys also impose a strong spatial/locality bias on the attention, which is well-suited to the iterative refinement of the protein structure.
* After each attention operations and element-wise transition block, the module computers an update to the rotation and transition of each backbone frame.
* Predictions of side-chain angles as well as the final, per-residue accuracy of the structure (pLDDT) are computed with small per-residue networks on the final activations at the end of the network.
* The estimate of the TM-score (pTM) is obtained from a pairwise error prediction that is computed as a linear projection from the final pair representation that is computed as a linear projection from the final pair representation.
* The final loss compares the predicted atom positions to the true positions under many different alignments.
* For each alignment, defined by aligning the predicted frame (Rk, **t***k*) to the corresponding true frame, the researchers compute the distance of all predicted atom positions **x**i from the true atom positions.
* The resulting *N*frames x *N*atoms distances are penalized with a clamped *L*1 loss.
* Training with labelled and unlabeled data: - The AlphaFold architecture is able to train accuracy using only supervised learning on PDB data, but are able to enhance accuracy using an approach similar to noisy student self-distillation.
* For this the researchers used around 350,000 diverse sequences from Uniclust30 and then make them a new dataset which will help to predict the structures filtered to a high-confidence subset.
* After that the researchers train the same architecture again from the scratch using a mixture of PDB data and this new dataset of predicted data augmentations such as cropping and MSA subsampling make it challenging for the network to recapitulate the previously predicted structures.
* The researchers also randomly mask out or mutate in individual residues within the MSA and have a Bidireactional Encoder Representations from Transformer (BERT)- style objective to predict the masked elements of the MSA sequences.
* This objective encourages the network to learn to interpret phylogenetic and covariation relationships without hardcoding a particular correlation statistic into the features.
* Interpreting the neural network: - for understanding the working of AlphaFold predicts protein structure, the researchers trained a separate structure module for each of the 48 Evoformer blocks in the network while keeping all parameters of the main network frozen.
* In the recycling stages, it provides a trajectory of 192 intermediates structures – one per full Evoformer block – in which each intermediate represents the belief of the network of the most likely structure at that block.
* It gets smooth after few blocks, which shows that the AlphaFold makes constant increment improvements to the structure until it can no longer improve.
* It also illustrates the role of network depth.
* For very challenging proteins like ORF8 or SARS-CoV-2, the network searchers and rearrange secondary structure elements for many layers before setting on a good structure.
* For other different proteins like LmrP (T1024), the network finds the final structure within the first few layers.
* MSA depth and cross-chain contacts: - the model uses MSAs and the accuracy decreases substantially when the median alignment depth is less than around 30 sequences.
* The researchers also observed that when the threshold effect where improvements in MSA depth over around 100 sequences lead to small gains.
* The researchers hypothesize that the MSA information is needed to coarsely find the correct structure within the early stages of the network, but refinement of that prediction into a high-accuracy model does not depend crucially on the MSA information.
* The researchers also find that the AlphaFold is much weaker for proteins that have few intra-chain or homotypic contacts compared to the number to the number of heterotypic contacts.
* AlphaFold is often able to give high-accuracy predictions for homomers, even when the chains are substantially interviewed.
* Related Work: - even with the long history of applying neural networks to structure prediction we have only recently come to improve structure predictions.
* These approaches effectively leverage the rapid improvement in computer vision systems by treating the problem of protein structure predictions as converting an ‘image’ of evolutionary coupling to an ‘image’ of the protein distance matrix and then integrating the distance predictions into a heuristic system that produces the final 3D coordinate prediction.
* Discussion: - the researcher used combination of the bioinformatics and physical approaches to build the AlphaFold.
* They use a physical and geometric inductive bias to build components that learn from PDB data with minimal imposition of handcrafted features.
* AlphaFold is able to handle missing the physical context and produce accurate models in challenging cases like intertwined homomer or proteins that only fold in the presence of an unknown haem group.
* In common, AlphaFold is trained to produce the protein structure most likely to appear as part of a PDB structure.
* The AlphaFold has already show case its utility to the experimental community, both for molecular replacement and for interpreting cryogenic electron microscopy maps.
* Because the AlphaFold outputs protein coordinates directly, AlphaFold produces predictions in graphics processing unit (GPU) minutes to GPU hours depending on the length of the protein sequence.
* The explosion in available genomic sequencing techniques and data has revolutionized bioinformatics but the intrinsic challenge of experimental structure determination has prevented a similar expansion in our structural knowledge.
* Methods: - IPA – the IPA module combines the pair representation, the single representation and the geometric representation to update the single representation.
* In this each representations contributes affinities to the shared attention weights and then uses these weights to map its values to the output.
* Each residue produces query points, keys points and value points in its local frame.
* These points are then projected into the global frame using the backbone frame of the residue in which they interact with each other.
* The resulting points are then projected back to the local frame.
* The affinity computation transformations ensure the invariance of this module with respect to the global frame.
* The IPA, standard dot product attention is computed on the abstract single representation and a special attention on the pair representation.
* Inputs and data sources: - inputs to the networks are the primary sequence. For both the MSA and templates, the search processes are tuned for high recall; spurious matches will probably appear in the raw MSA but this matches the training condition of the network.
* One of the sequence databases used, Big Fantastic Database (BFD), was custom-made and released publicly and was used by several CASP teams.
* BFD was built in three steps. First, 2,423,213,294 protein sequences were collected from UniProt, and clustered to 30% sequence identity, while enforcing a 90% alignment coverage of the shorter sequences using MMseqs2/Linclust.
* For computational efficiency the researchers remove the clusters with less than three members, which results 61,083,719 clusters.
* Second, they added 166,510,624 representative protein sequences from Metaclust NR by aligning them against the cluster representatives using MMseqs2.
* The remaining 25,347,429 sequences that could not be assigned were clustered separately and added as new clusters, resulting the final clustering.
* Third, for each of the clusters, the researchers computed an MSA using FAMSA and computed the HMMs following the Uniclust HH-suite database protocol.
* For MSA search on BFD + Uniclust30, and templates search against PDB70, the researchers used HHBlits and HHSearch from hh-suite v.3.0-beta.3.
* For MSA search on Uniref90 and clustered MGnify, the researchers used jackhammer from HMMER3.
* For constrained relaxation of structures, they used OpenMM v.7.3.1 with the Amber99sb force field.
* Removing BFD reduced the mean accuracy by 0.4 GDT, removing Mgnify reduced the mean accuracy by 0.7 GDT, and removing both reduced the mean accuracy by 6.1GDT.
* They observe that most targets had very small changes in accuracy but a few outliers had very large differences.
* The MSA is relatively unimportant until it approaches a threshold value of around 30sequences when the MSA size effects become quite large.
* They also observe that mostly overlapping effects between inclusion of BFD and Mgnify, but having at least one of these metagenomics databases is very important for target classes that are poorly represented in UniRef, and having both was necessary to achieve full CASP accuracy.
* Training regimen: - chains are sampled in inverse proportion to cluster size of a 40% sequence identity clustering. After that the researchers randomly crop them to 256 residues and assemble into batches of size 128.
* They train the model on Tensor Processing Unit (TPU)v3 with a batch size of 1 per TPU core.
* It is trained until convergence and further fine-tuned using longer crops of 384 residues, larger MSA stack and reduced learning rate.
* The network is supervised by the FAPF loss and a number of auxiliary losses.
* First, the final pair representation is linearly projected to a binned distribution prediction, scored with a cross-entropy loss.
* Second, they use random masking on the input MSAs and require the network to reconstruct the masked regions form the output MSA representation using BERT-like loss.
* Third, the output single representations of the structure module are used to predict binned per-residue IDDT-C values.
* At last, they use an auxiliary side-chain loss during training, and an auxiliary structure violation loss during fine-tuning.
* An initial model trained with the above objectives was used which make the structured predictions for a Uniclust dataset of 355,993 sequences with the full MSAs.
* These predictions are used to train the final model which have same hyperparameters, except for sampling examples 75% of the time from the Uniclust prediction set, with sub-sampled MSAs, and 25% of the time from the clustered PDB set.
* The researcher train five different model using different random seeds, some with templates and some without it, to encourage diversity in the predictions.
* They fine-tuned the model after CASP14 to add a pTM prediction objective and use the obtained models.
* Inference regimen: - the researchers inference five trained models and use the predicted confidence score to select the best model per target.
* Using CASP14 of the researchers for AlphaFold, the trunk of the network is run multiple times with different random choices for the MSA cluster centers.
* The researchers found that the accuracy of the network without ensembling is very close or equal to the accuracy with ensembling.
* Without ensembling the network is also 8x faster and the representative timings for a single model are 0.6 min with 256 residues, 1.1 min with 384 residues and 2.1h with 2,500 residues.
* Large proteins can easily exceed the memory of a single GPU. The memory usage is a approximately quadratic in the number of residues, so a 2,500-residue protein involves using unified memory so that the researchers can greatly exceed the memory of a single V100.
* Metrics: - the predicted structure is compared to the true structure from the PDB in terms of IDDT metric.
* The distances are either computed between all heavy atoms (IDDT) or only the C atoms to measure the backbone accuracy (IDDT-C).
* Domain accuracies in CASP are reported as GDT and the TM-score is used as a full chain global superposition metric.
* The researchers perform five iterations of (1) a least-squares alignment of the predicted structure and the PDB structure on the currently chosen C atoms; (2) selecting the 95% of C atoms with the lowest alignment error;
* The r.m.s.d. metric is more robust to apparent errors that can originate from crystal structure artefacts, although in some cases the removed 5% of residues will contain genuine modelling errors.
* Test set of recent PDB sequences: - the researchers downloaded the PDB dataset for evaluation and remove all the unnecessary data (like, duplicate data, single amino acid data etc.).
* Because the PDB contains many near-duplicate sequences, the chain with the highest resolution was selected from each cluster in the PDB 40% sequence clustering of the data.
* After that they removed all sequences for which fewer than 80 amino acids had the alpha carbon resolved and removed chains with more than 1,400 residues.
* The final dataset contains 10,795 protein sequences.
* Each residue position in a query sequence was assigned the maximum identity of any template hit covering that position. Filtering then proceeded as described in the individual figure legends, based on a combination of maximum identity and sequence coverage.
* The MSA depth analysis was based on computing the normalized number of effective sequences (*N*eff) for each position of a query sequence. Per-residue *N*eff values were obtained by counting the number of non-gap residues in the MSA for this position and weighting the sequences using the *N*eff scheme with a threshold of 80% sequence identity measured on the region that is non-gap in either sequence.