What AlphaFold teaches us about deep learning using prior knowledge and ontologies

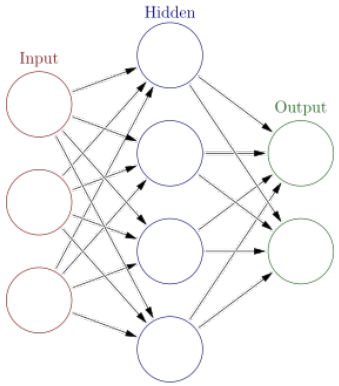
Jobst Landgrebe
September 27th
ICBO 2022 Ann Arbor

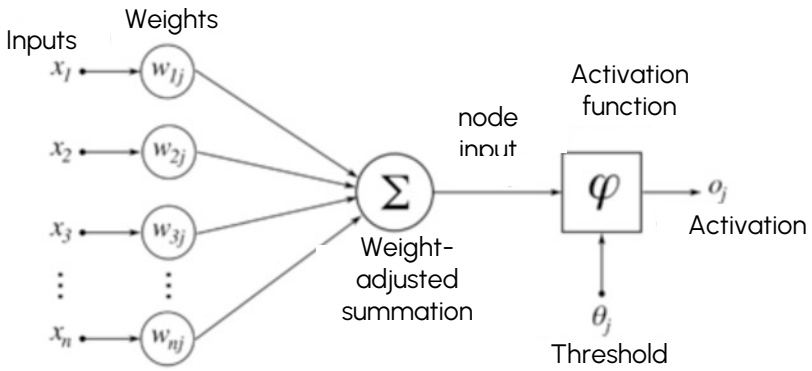
Contact:
jobstlan@buffalo.edu jobst.landgrebe@cognotekt.com

cognotekt

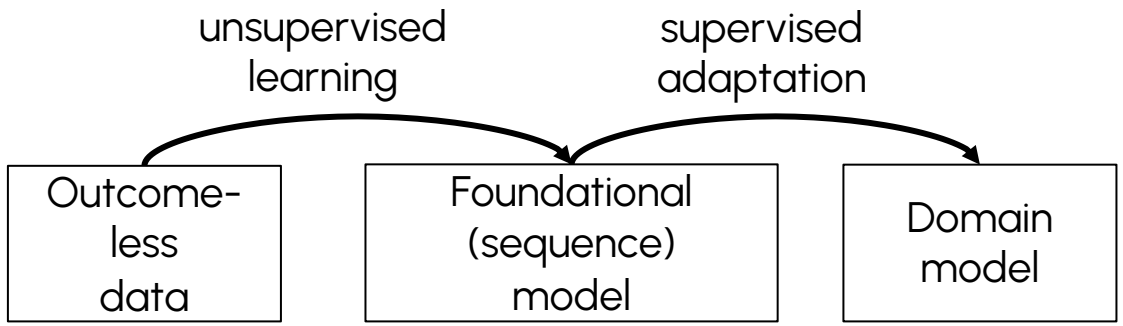
Today's leading stochastic prediction models are parameter-rich supervised dNN or unsupervised foundational models – both are major engineering achievements

Supervsied dNN





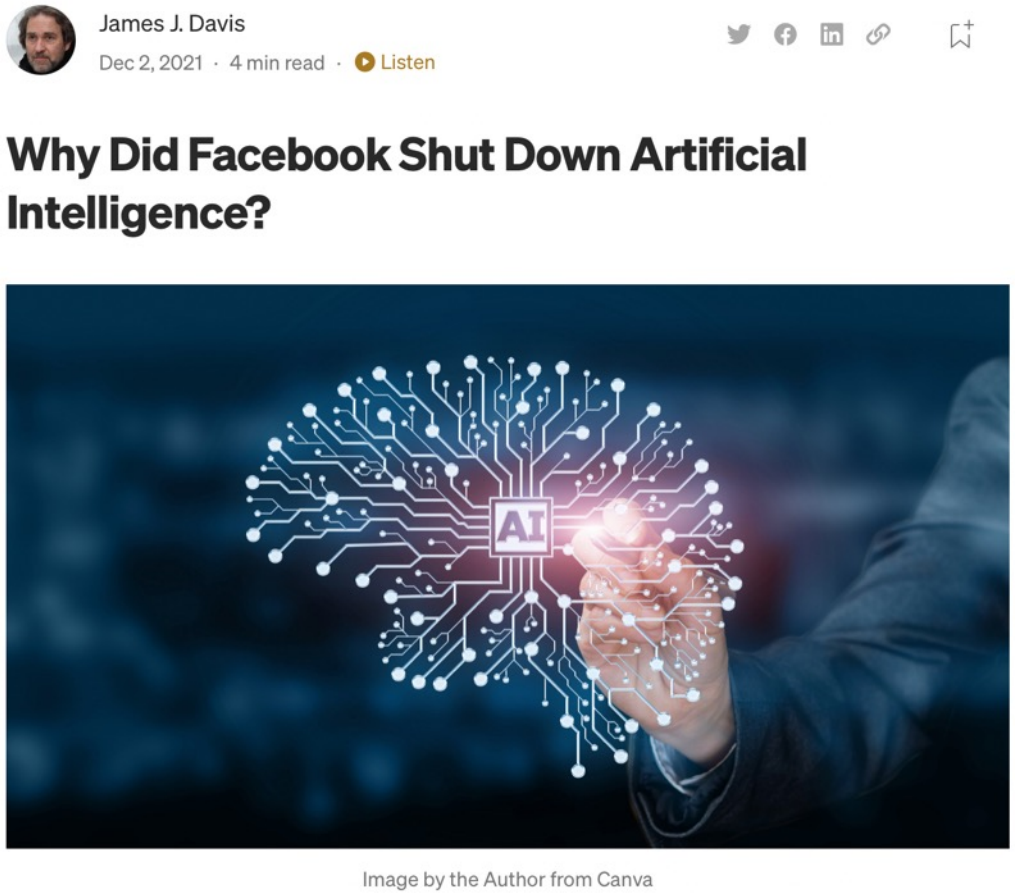
Unsupervised foundational model



Type	Supervised dNN	Unsupervised dNN
Training data	$\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$ input **x** and output **y** (outcome) **x** is *p*-dimensional	$N \times p$ data matrix X *without* outcome
Procedure	Identification of local loss minima using gradient-descent on model loss functions	
Model structure	implicit, non-linear dNN	
Parameters	millions to billions	billions
Popular examples	Microsoft chatbot Xiaoice , translate.google.com	GPT-3, BERT, CLIP

But the models have major weaknesses which become apparent immediately when they get used out of context or for the wrong purpose

dNN limitations



1. Each model is tied to a specific context.
2. Each model is tied to specific training data, and the training data are often inadequate, or available only in insufficient quantities.
3. Each model has user-specified training hyper-parameters and optimisation algorithms.
4. Cases where adequate – which means representative – training data are available in sufficient quantities are few.
5. We cannot obtain adequate models of the behaviour of animate and inanimate complex systems in open environments.

Some of the AI limitations can be overcome by using prior knowledge as input for the algorithms

Enhancing machine learning with prior knowledge

$$p(\theta, y) = p(\theta)p(y|\theta)$$

Prior knowledge [$p(\theta)$] can be used to restrict a stochastic space in such a way that models learned from training samples drawn from this space have a higher accuracy than models that do not do so.

Prior knowledge can be formulated using explicit stochastic networks.

But it can also be provided using applied ontology, often described as "knowledge graph" in the computer science literature.

Ontologies can be used in multiple ways to enhance statistical learning – in biology, AlfaFold is an example of sophisticated class feature usage

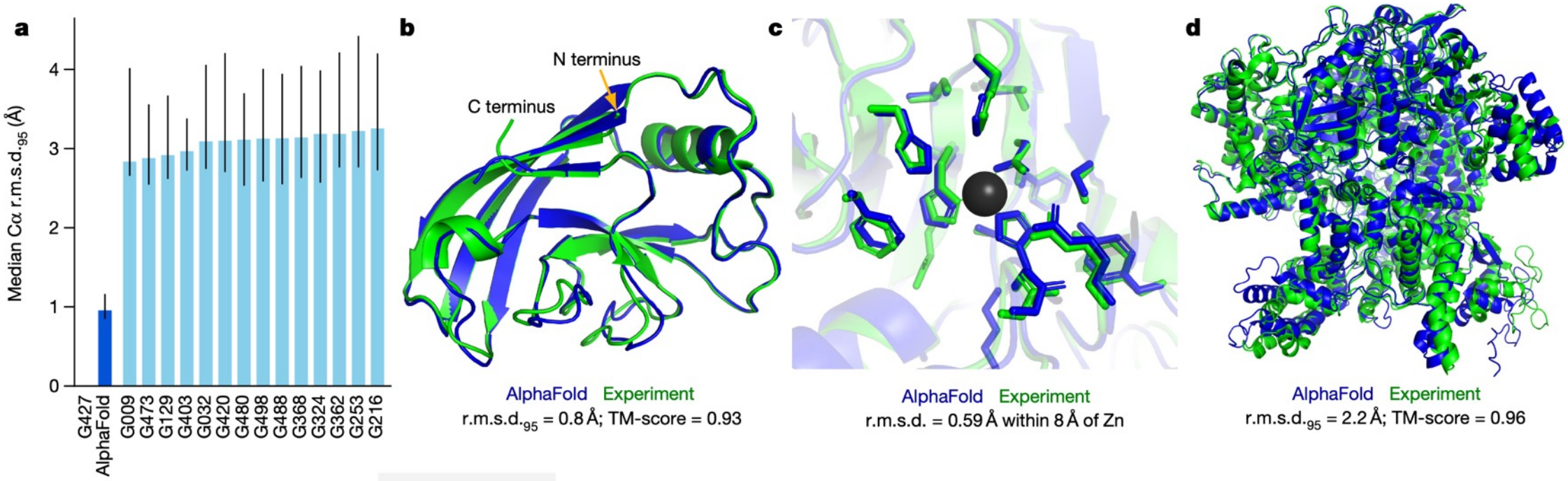
Usage of ontologies for statistical learning

Paradigms	Summary	Categories
Mapping-based	These methods project the input and/or the class into a common vector space where a sample is close to its class w.r.t. some distance metric, and prediction can be implemented by searching the nearest class.	Input Mapping
		Class Mapping
		Joint Mapping
Data Augmentation	These methods generate samples or sample features for the unseen classes, utilizing KG auxiliary information.	Rule-based
		Generation Model-based
Propagation-based	These methods propagate model parameters or a sample's class beliefs from the seen classes to the unseen classes via a KG.	Model Parameter Propagation
		Class Belief Propagation
Class Feature	These methods encode the input and the class into features often with their KG contexts considered, fuse these features and feed them directly into a prediction model.	Text Feature Fusion
		Multi-modal Feature Fusion

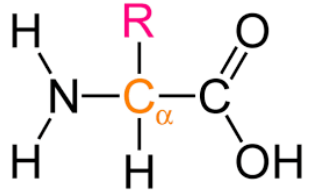
Source: J. Chen et al. (2022) arXiv preprint arXiv:2112.10006v5

AlphaFold, an encoder-decoder sequential dNN, outperformed other approaches to protein 3D structure prediction in the CASP14 competition

AI in biomedical research: Protein folding example



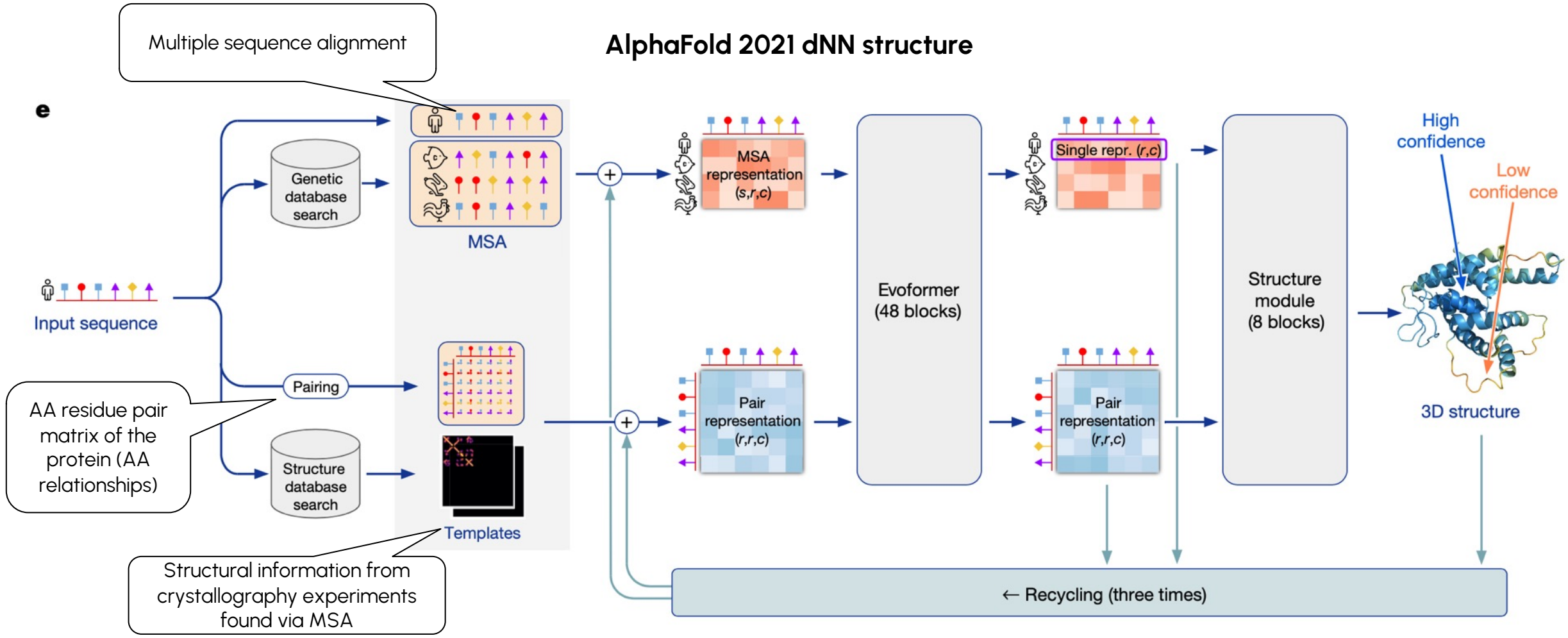
a. The performance of AlphaFold on the CASP14 dataset (n = 87 protein domains) relative to the top-15 entries (out of 146 entries).
b. –d . Examples for predictions experiments with AlphaFold
C_α root-mean-square deviation$_{95}$: measure of the average distance between the C_α atoms of superimposed proteins at 95% coverage of the AA-sequence of the protein. 1 Å = 0,1 nm= 10^{-10} m



C_α in an amino acid

Source: Jumper et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature. https://doi.org/10.1038/s41586-021-03819-2

The core idea of AlphaFold is to view the prediction of protein structures as a graph inference problem in 3D space



Edges of the graph are defined by residues in proximity

Source: Jumper et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature. https://doi.org/10.1038/s41586-021-03819-2

AlphaFold directly predicts the 3D coordinates of the heavy atoms for a given protein using the sequence, sequences homology and crystallography results as inputs

AlphaFold 2021 Input features

Protein amino acid pair matrix

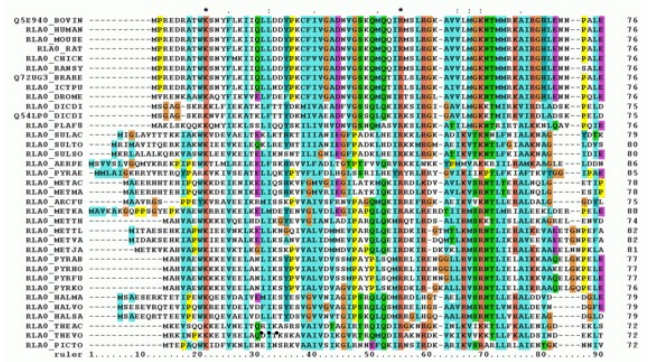
Protein aminoacid

Protein aminoacid

$$\begin{array}{c} \\ 1 \\ 2 \\ \\ \\ n \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & & & & n \end{array} \\ \begin{bmatrix} 0 & d_{12}^2 & d_{13}^2 & \dots & d_{1n}^2 \\ d_{21}^2 & 0 & d_{23}^2 & \dots & d_{2n}^2 \\ d_{31}^2 & d_{32}^2 & 0 & \dots & d_{3n}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n1}^2 & d_{n2}^2 & d_{n3}^2 & \dots & 0 \end{bmatrix} \end{array}$$

32 AA clipped pair distances

Multiple homology-based sequence alignment

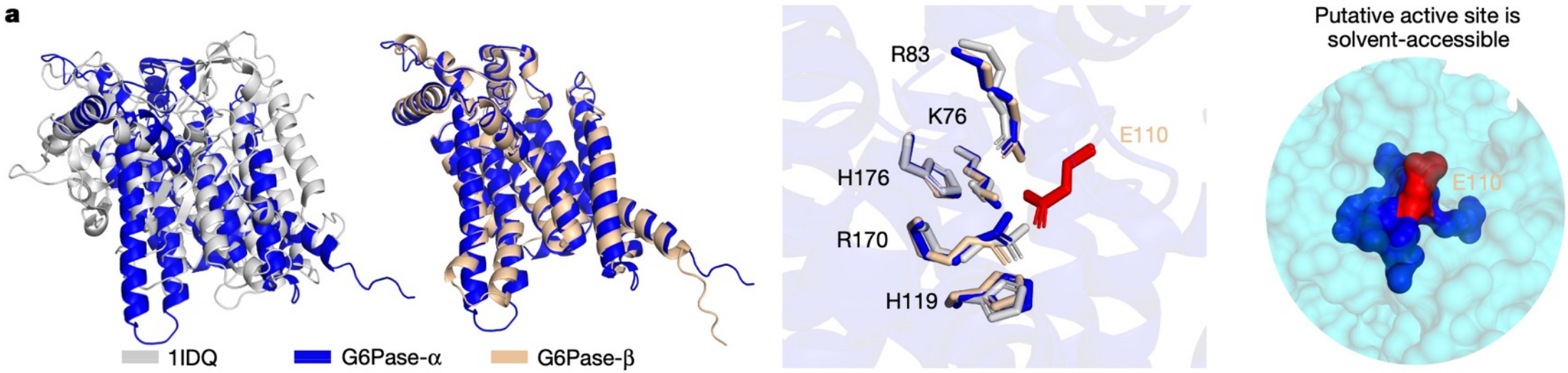


Crystallography Information for homology cluster

Each mmCIF file can be seen as an ontology of the protein structure it describes, and the many mmCIF files from an ontology of the Protein Data Bank (PDB). The collection of mmCIF files is structured into protein homology families, and as UniProt and PDB develop, more and more of the hierarchical ontology structure incorporated into PRO will become explicit in these resources, too.

AlphaFold is impressive for examples with high homology to known structures

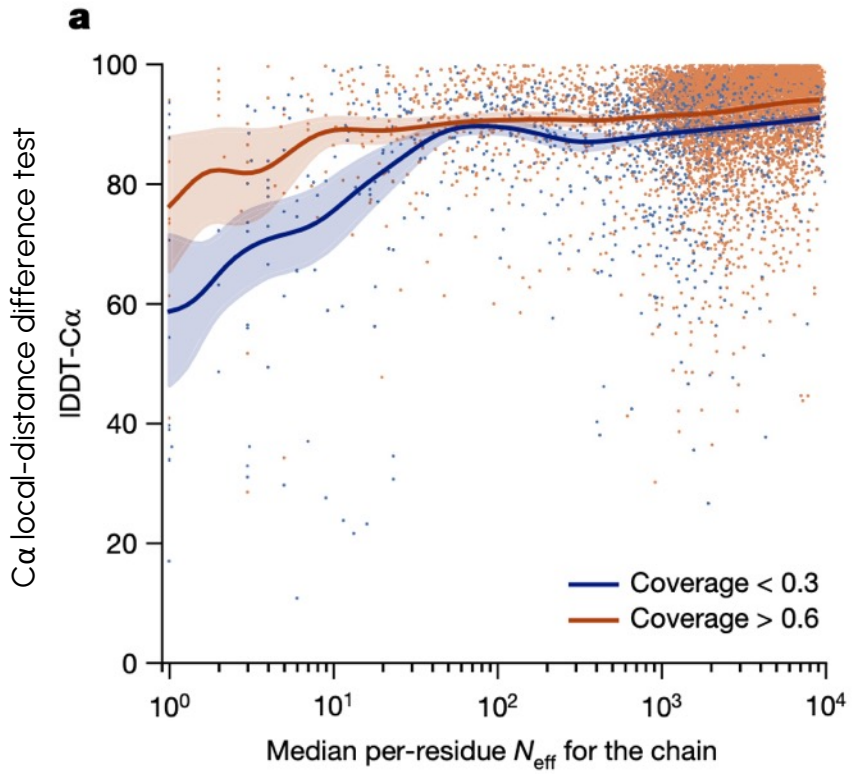
AlphaFold 2021: Achievement example



"To our knowledge, no experimental structure [of Glucose-6-phosphatase] exists, but previous studies have attempted to characterize the transmembrane topology and active site. [...] In the G6Pase-α binding pocket face, opposite the residues shared with the chloroperoxidase, we predict a conserved glutamate (Glu110) that is also present in our G6Pase-β prediction (Glu105) but not in the chloroperoxidase (Fig. 3a). The glutamate could stabilize the binding pocket in a closed conformation, forming salt bridges with positively charged residues there."

Tunyasuvunakool et al (2021) Highly accurate protein structure prediction for the human proteome, Nature.
https://doi.org/10.1038/s41586-021-03828-1

AlphaFold is mostly limited to predicting proteins which are homologous to known structures and for which rich homology groups are available



"The model uses MSAs and the accuracy decreases substantially when the median alignment depth is less than around 30 sequences."

Fig. 5 | Effect of MSA depth and cross-chain contacts. a, Backbone accuracy (lDDT-Cα) for the redundancy-reduced set of the PDB after our training data cut-off, restricting to proteins in which at most 25% of the long-range contacts are between different heteromer chains. We further consider two groups of proteins based on template coverage at 30% sequence identity: covering more than 60% of the chain ($n = 6,743$ protein chains) and covering less than 30% of the chain ($n = 1,596$ protein chains). MSA depth is computed by counting the

Biased analysis because long-range contacts cannot be modelled

Source: Jumper et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature. https://doi.org/10.1038/s41586-021-03819-2

The success of AlphaFold, which models known structure families, will enhance the usage of ontologies in statistical learning

AlphaFold 2021: design style, assumptions and results

- dNN design is like alchemy – input data and architecture are chosen in a heuristic manner using different patterns until a design emerges that can tackle the problem and yield good predictions that can also cope with independent test data
- AlphaFold is also made like this, but it contains some design decisions which show the nature of the cognitive style of this scientific community:
  - The core assumptions about the spatial relationships between amino acids totally abstract the biological reality of the protein folding, which is a highly contextual process depending on multiple adjuvant proteins in the endoplasmatic reticulum
  - Instead, simple mathematical (triangular) relationships between the AA residues are used to design the AA-matrix transformations
  - The operations performed on the MSA representation are even less motivated by biology

- Nevertheless, the performance on proteins homologous to known structures is excellent, illustrating once more the excellent ability of dNN to identify regular, recurring patterns

- **What the dNN models is the conversation of protein folding mechanisms in evolution as far as we have learned about it from crystallography experiments**
- **AlphaFold will enhance the usage of ontologies in statistical learning**

Thanks for your interest

jobstlan@buffalo.edu

The input features characterise the AA sequence, the MSA clustering results and the crystallography properties

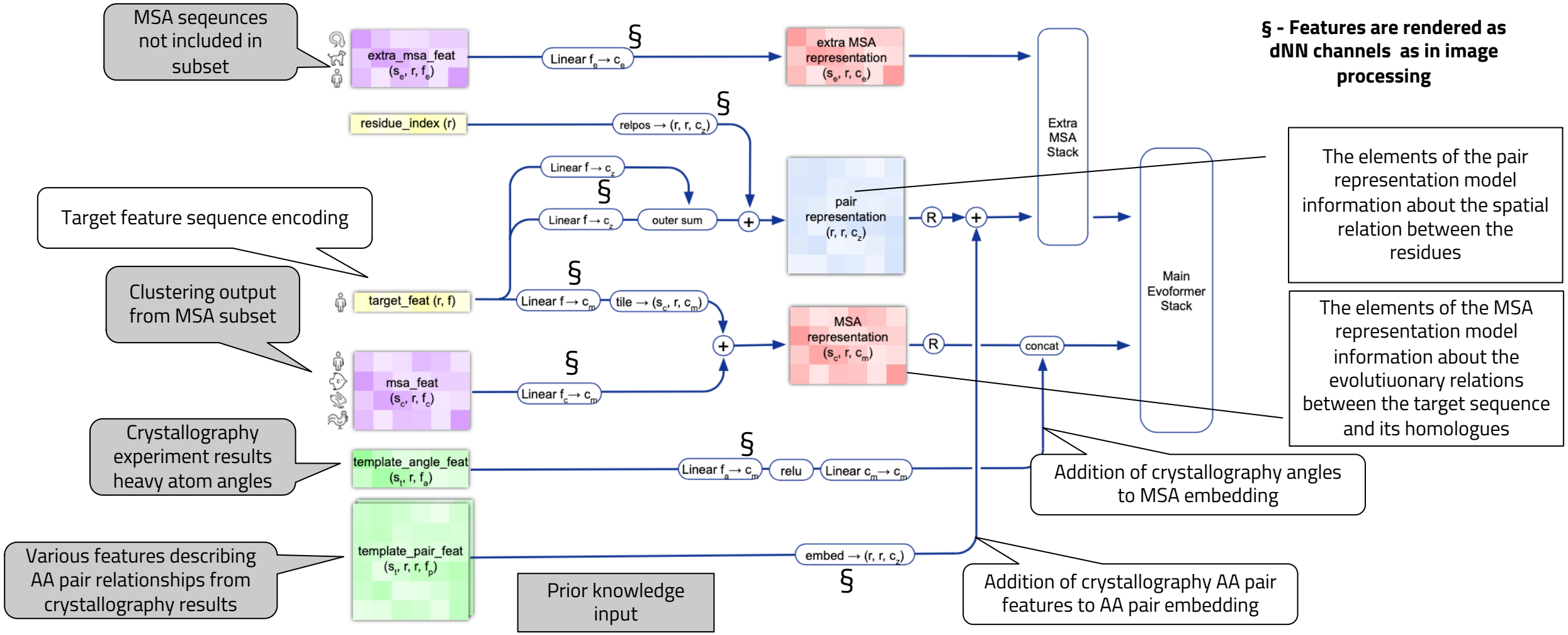
Table 1 | Input features to the model. Feature dimensions: N_{res} is the number of residues, N_{clust} is the number of MSA clusters, $N_{\text{extra\_seq}}$ is the number of additional unclustered MSA sequences, and N_{templ} is the number of templates.

Feature & Shape	Description
aatype $[N_{\text{res}}, 21]$	One-hot representation of the input amino acid sequence (20 amino acids + unknown).
cluster_msa $[N_{\text{clust}}, N_{\text{res}}, 23]$	One-hot representation of the msa cluster centre sequences (20 amino acids + unknown + gap + masked_msa_token).
cluster_has_deletion $[N_{\text{clust}}, N_{\text{res}}, 1]$	A binary feature indicating if there is a deletion to the left of the residue in the MSA cluster centres.
cluster_deletion_value $[N_{\text{clust}}, N_{\text{res}}, 1]$	The raw deletion counts (the number of deletions to the left of every position in the MSA cluster centres) are transformed to the range $[0, 1]$ using $\frac{2}{\pi} \arctan \frac{d}{3}$ where d are the raw counts.
cluster_deletion_mean $[N_{\text{clust}}, N_{\text{res}}, 1]$	The mean deletions for every residue in every cluster are computed as $\frac{1}{n} \sum_{i=1}^{n} d_{ij}$ where n is the number of sequences in the cluster and d_{ij} is the number of deletions to the left of the ith sequence and jth residue. These are then transformed into the range $[0, 1]$ in the same way as for the cluster_deletion_value feature above.
cluster_profile $[N_{\text{clust}}, N_{\text{res}}, 23]$	The distribution across amino acid types for each residue in each MSA cluster (20 amino acids + unknown + gap + masked_msa_token).

Feature & Shape	Description
extra_msa $[N_{\text{extra\_seq}}, N_{\text{res}}, 23]$	One-hot representation of all MSA sequences not selected as cluster centres (20 amino acids + unknown + gap + masked_msa_token).
extra_msa_has_deletion $[N_{\text{extra\_seq}}, N_{\text{res}}, 1]$	A binary feature indicating if there is a deletion to the left of the residue in the extra MSA sequences.
extra_msa_deletion_value $[N_{\text{extra\_seq}}, N_{\text{res}}, 1]$	The raw deletion counts to the left of every residue in the extra_msa, converted to the range $[0, 1]$ using the same formula as for cluster_deletion_value.
template_aatype $[N_{\text{templ}}, N_{\text{res}}, 22]$	One-hot representation of the amino acid sequence (20 amino acids + unknown and gap).
template_mask $[N_{\text{templ}}, N_{\text{res}}]$	Mask indicating if a template residue exists and has coordinates.
template_pseudo_beta_mask $[N_{\text{templ}}, N_{\text{res}}]$	Mask indicating if the beta carbon (alpha carbon for glycine) atom has coordinates for the template at this residue.
template_backbone_frame_mask $[N_{\text{templ}}, N_{\text{res}}]$	A mask indicating if the coordinates of all the required atoms to compute the backbone frame (used in the template_unit_vector feature) exist.
template_distogram $[N_{\text{templ}}, N_{\text{res}}, N_{\text{res}}, 39]$	A one-hot pairwise feature indicating the distance between beta carbons (alpha carbon used for glycine) atoms. The pairwise distances are discretized into 38 bins of equal width between 3.25 Å and 50.75 Å; and one more bin contains any larger distances.
template_unit_vector $[N_{\text{templ}}, N_{\text{res}}, N_{\text{res}}, 3]$	The unit vector of the displacement of the alpha carbon atom of all residues within the local frame of each residue. These local frames are computed in the same way as for the target structure, see subsubsection 1.8.1. (Current models were trained with this feature set to zero.)
template_torsion_angles $[N_{\text{templ}}, N_{\text{res}}, 14]$	The 3 backbone torsion angles and up to 4 side-chain torsion angles for each residue represented as sine and cosine encoding.
template_alt_torsion_angles $[N_{\text{templ}}, N_{\text{res}}, 14]$	Alternative torsion angles for side chain parts with 180^{o}-rotation symmetry.
template_torsion_angles_mask $[N_{\text{templ}}, N_{\text{res}}, 14]$	A mask indicating if the torsion angle is present in the template structure.
residue_index $[N_{\text{res}}]$	The index into the original amino acid sequence.

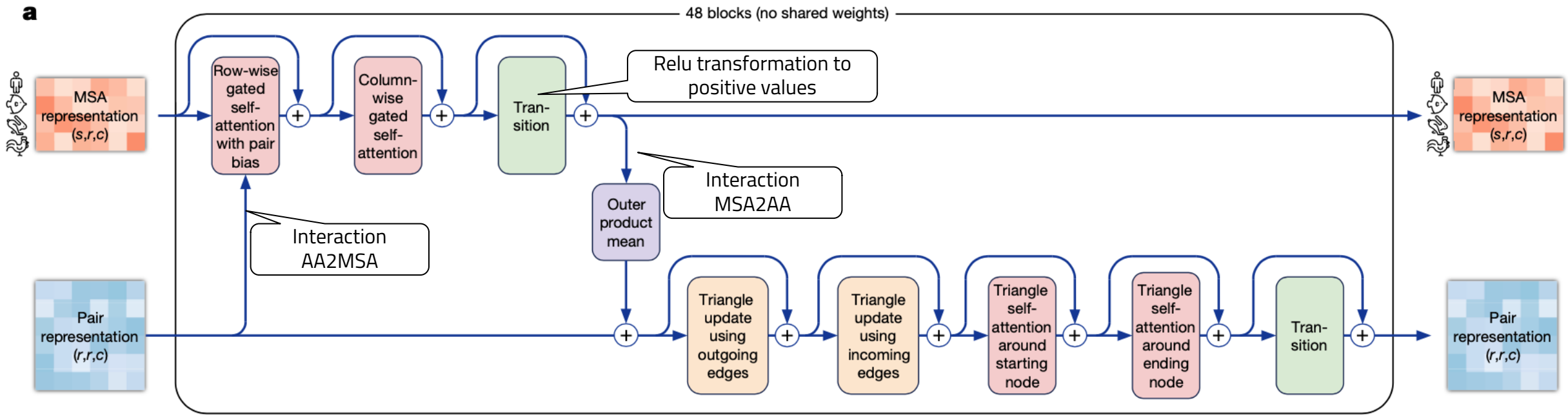
The input features are transformed in the first layer of the neural network

AlphaFold 2021 Input feature embedding
-- heavily heuristic and using massive prior knowlede



The "Evoformer" component refines a matrix representation of the processed MSA and the distance matrix of the input sequence's residue pairs

AlphaFold 2021: Evoformer Module input transformation chain

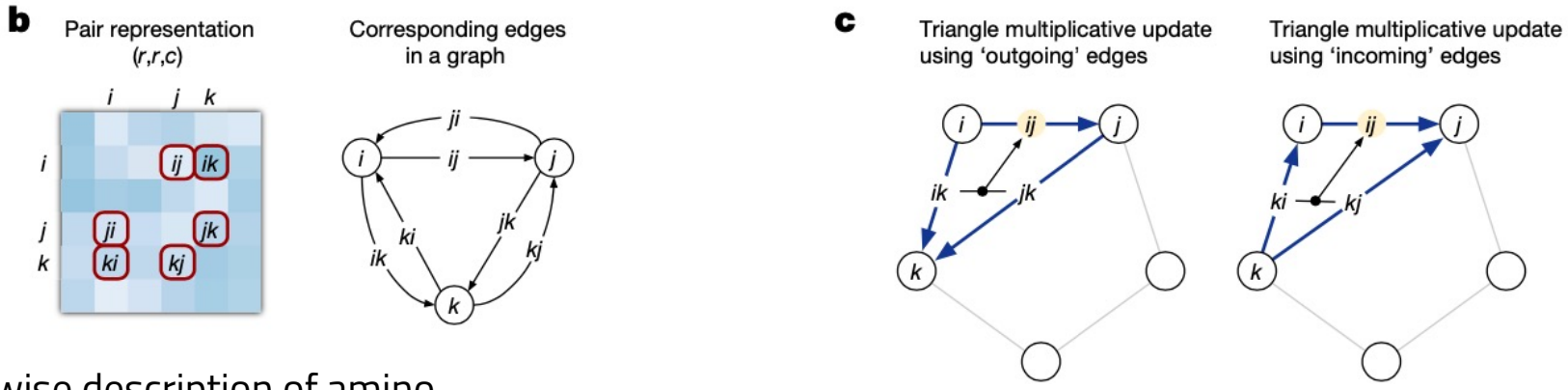


- MSA: evolutionary information
- Pair representation: spatial relationships between the amino-acids within the protein
- (Attention: see next slide)
- The Evoformer block exchanges information within the MSA and pair representations that enables the modelling of interactions between the evolutionary and spatial relationships

Source: Jumper et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature. https://doi.org/10.1038/s41586-021-03819-2

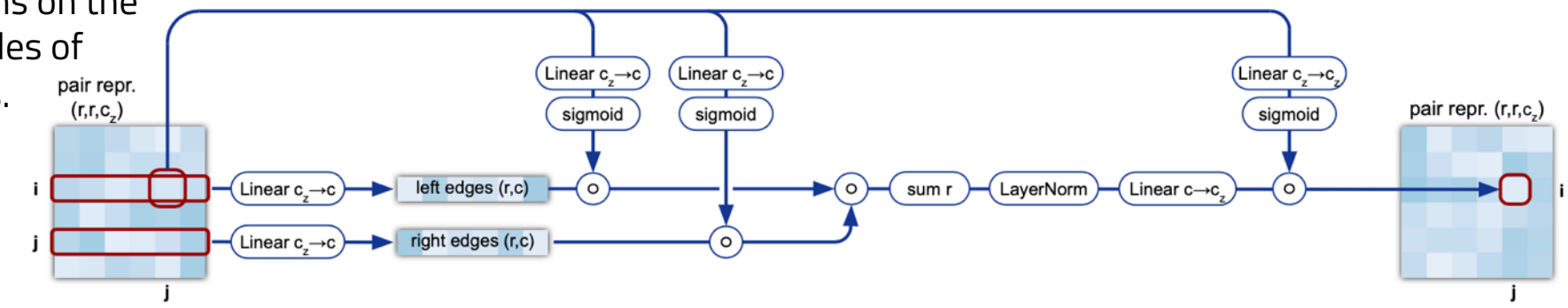
The AA-residue distance matrix is refined using a geometric triangle logic

AlphaFold 2021: AA-residue representation refinement Step 1



Basic idea: for pairwise description of amino acids to be representable as a single 3D structure, many constraints must be satisfied including the triangle inequality on distances. Authors arrange the update operations on the pair representation in terms of triangles of edges involving three different nodes.

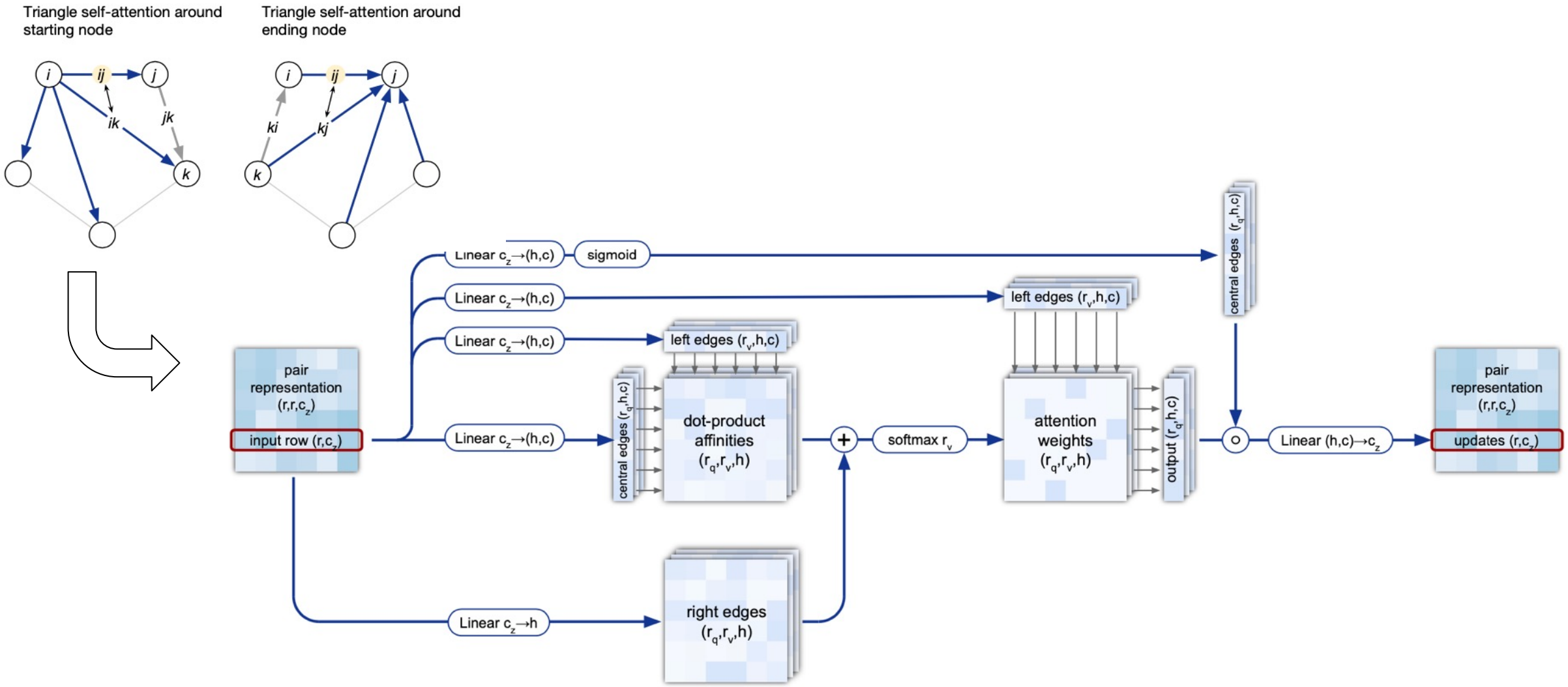
Each edge ij receives an update from the other two edges of all triangles where it is involved to model interactions with the sequence:



Source: Jumper et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature. https://doi.org/10.1038/s41586-021-03819-2

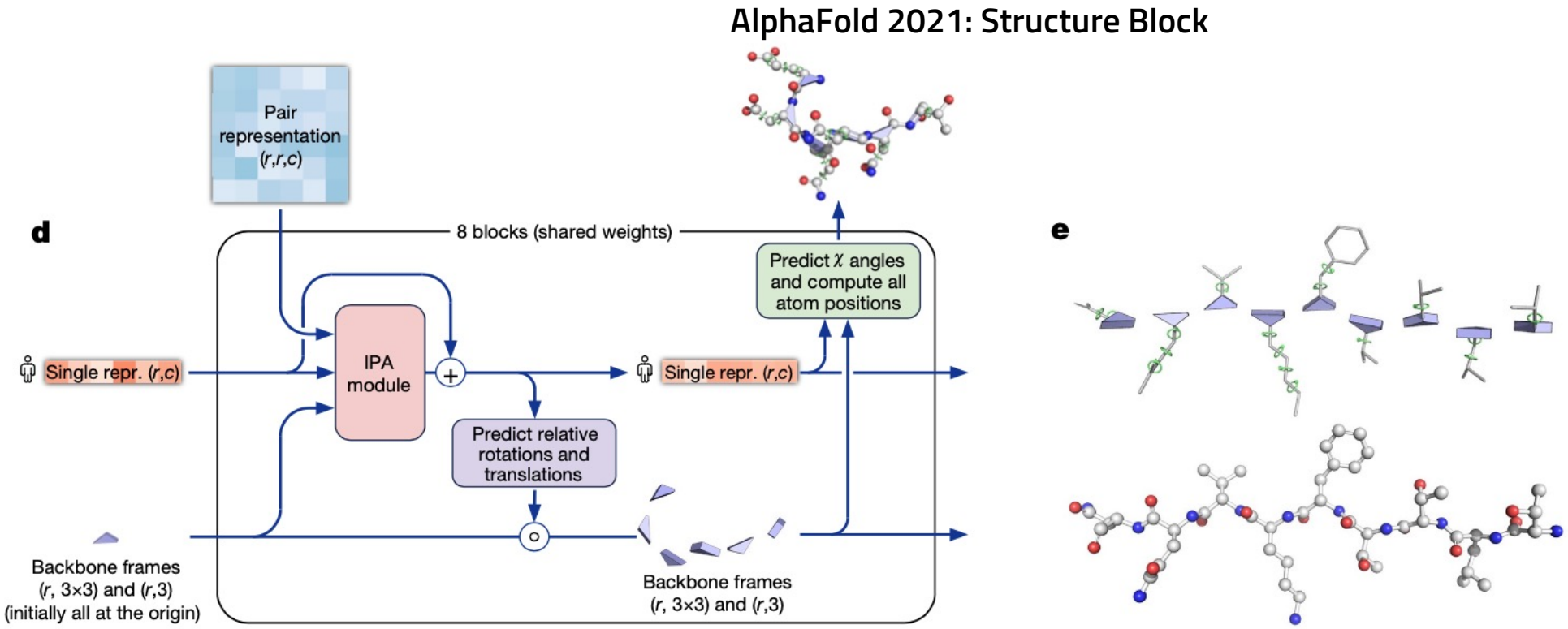
Attention is used to model contextual relationships between the AA mimicing their interactions upon folding

AlphaFold 2021: AA-residue representation refinement Step 2



Source: Jumper et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature. https://doi.org/10.1038/s41586-021-03819-2

The structure decoder block of the sequential dNN assigns a rotation and translation to each residue before computing

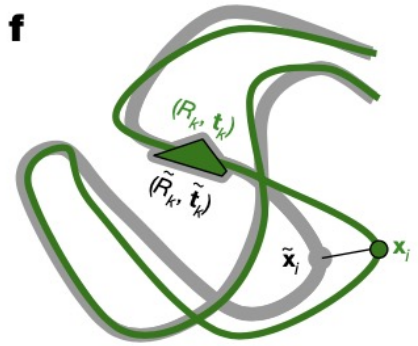


The 3D backbone structure is represented as N_{res} independent rotations and translations, each with respect to the global frame (residue gas) (Fig. 3e). These rotations and translations—representing the geometry of the N-Cα-C atoms—prioritize the orientation of the protein back- bone so that the location of the side chain of each residue is highly constrained within that frame.

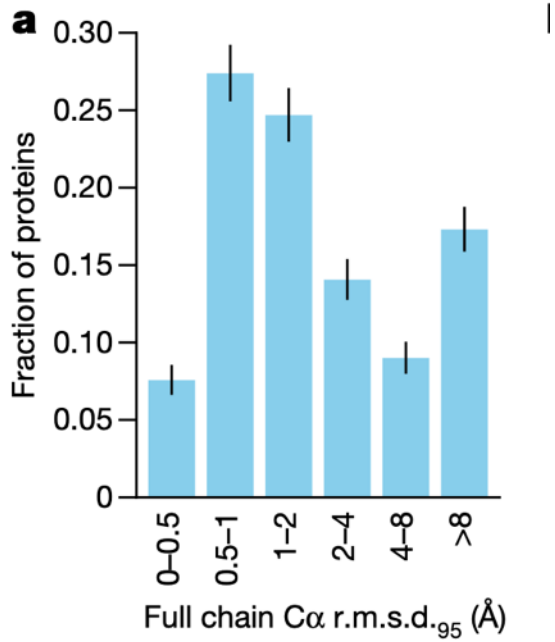
Source: Jumper et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature. https://doi.org/10.1038/s41586-021-03819-2

The final steps of the network compute side chain angles and the pre-residue positions

AlphaFold 2021: Structure Block



Predictions of side-chain χ 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. The finalloss(which we term the frame-aligned point error (FAPE) (Fig.3f)) compares the predicted atom positions to the true positions under many different alignments.

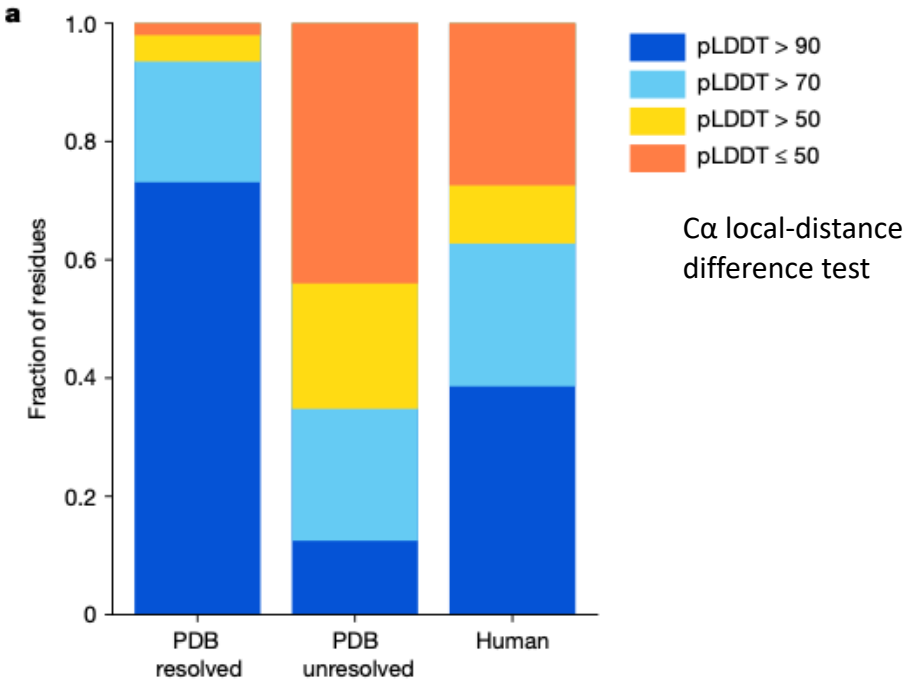


Histogram of backbone r.m.s.d. for full chains (Cα r.m.s.d. at 95% coverage). Error bars are 95% confidence intervals (Poisson). This dataset excludes proteins with a template (identified by hmmsearch) from the training set with more than 40% sequence identity covering more than 1% of the chain (n = 3,144 protein chains).

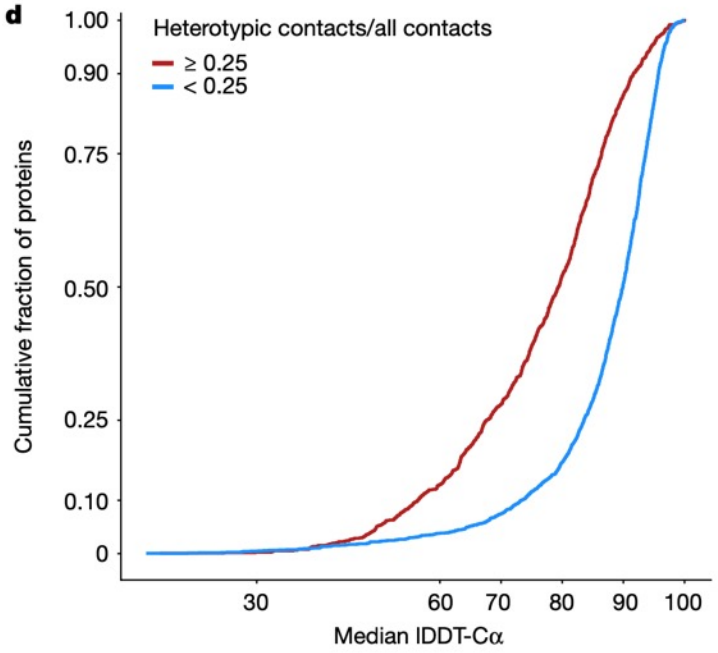
Source: Jumper et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature. https://doi.org/10.1038/s41586-021-03819-2

Intrinsic protein disorder does not explain AlphaFold's insufficient coverage of the human proteome – rather, it cannot cope with unknown and challenging structures

Biased analysis (residue view only)*



Cα local-distance difference test



Cα local-distance difference test

"The other substantial limitation that we have observed is that AlphaFold is much weaker for proteins that have few intra-chain or homotypic contacts compared to the number of heterotypic contacts. " – in other words, the model, like all sequence model, suffers from its context myopy which is able only to take into account a fairly limited context. Intrinsic disorder of proteins in solution (protein domains which do not form a stable 3D conformation) alone cannot explain the prediction gap in AlphaFold.

Tunyasuvunakool et al (2021) Highly accurate protein structure prediction for the human proteome, Nature. https://doi.org/10.1038/s41586-021-03828-1

* At the per-protein level, 43.8% of proteins have a confident prediction on at least three quarters of their sequence, while 1,290 proteins contain a substantial region (more than 200 residues) with pLDDT ≥ 70 and no good template.