ALPHA-FOLD: HIGHLY ACCURATE PROTEIN STRUCTURE PREDICTION

Presented by:

Siddhartha Dheer

MS in Data Science, University at Buffalo

Course:

CSE 676-B: Deep Learning — Summer 2025

Instructor:

Prof. Alina Vereshchaka

University at Buffalo The State University of New York

Why Protein Structure Prediction Matters

Importance of Protein Structures

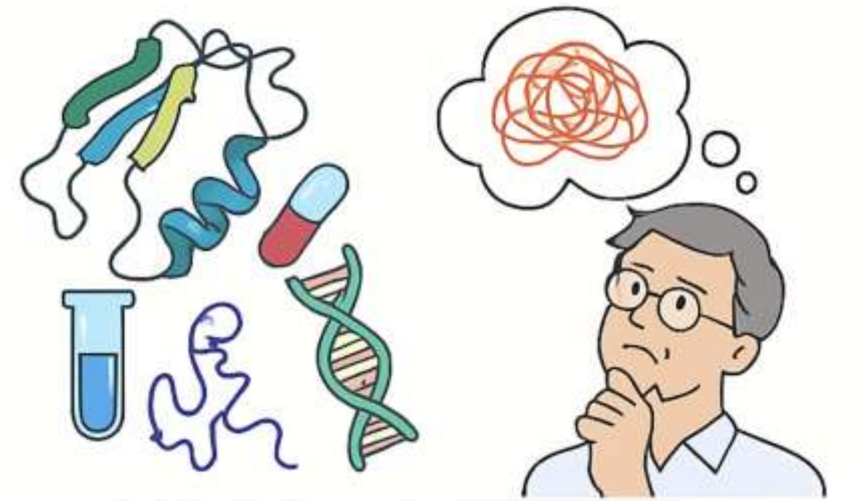
- Understanding protein structures is crucial for predicting biological functions, facilitating drug design, and deciphering disease mechanisms.
- Traditional experimental methods (X-ray crystallography, Cryo-EM, NMR) are slow, costly, and limited, covering less than 0.02% of known protein sequences.

Historical Challenges

- The “Protein Folding Problem”—accurately predicting the 3D structure solely from amino acid sequences—has remained unsolved for over 50 years.
- Physical simulations (like Molecular Dynamics) are computationally infeasible for most real-world proteins due to complexity (Levinthal’s Paradox).

Breakthrough Achieved by AlphaFold

- AlphaFold was trained and validated on challenging CASP13 and CASP14 datasets without relying on structural homology.



- Successfully predicts novel structures at atomic-level resolution through a sophisticated neural network architecture.
- Utilizes a combination of biological priors, advanced deep learning techniques, and geometry-aware attention mechanisms.

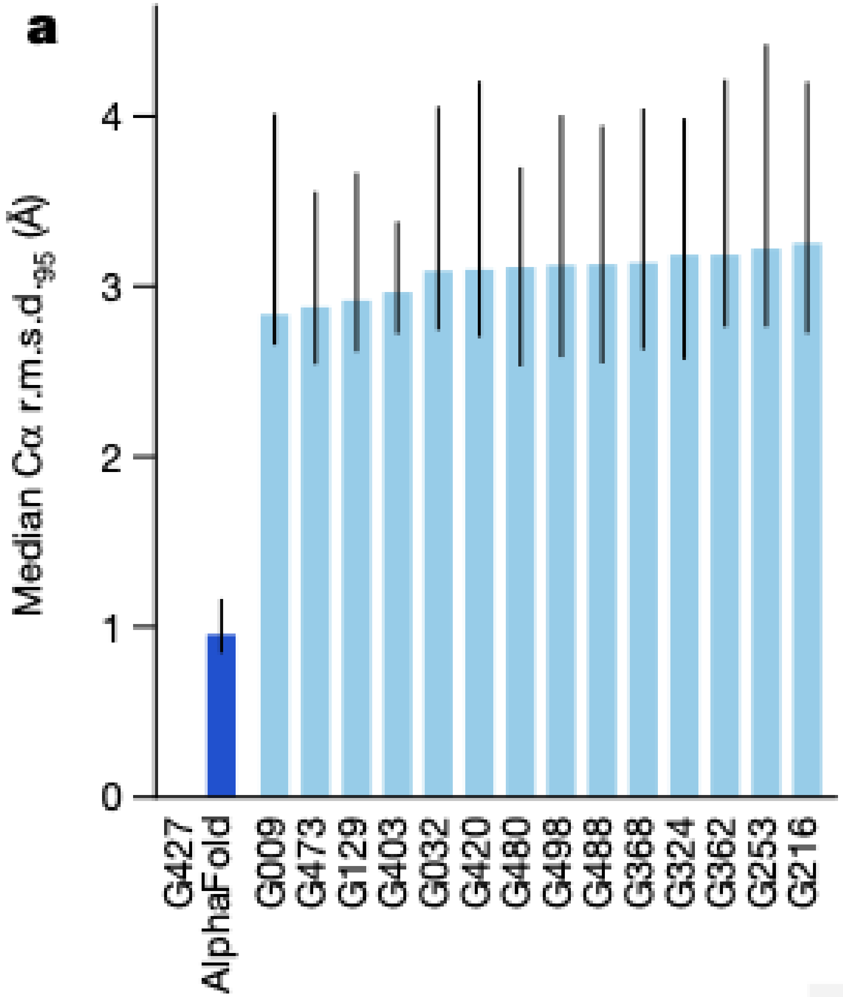
Limitations of Early Computational Methods

- Early machine learning methods lacked generalization capabilities for previously unseen protein folds.
- Evolutionary coupling methods required close homologous structures as templates, severely restricting their predictive scope.

AlphaFold CASP14 Performance Overview

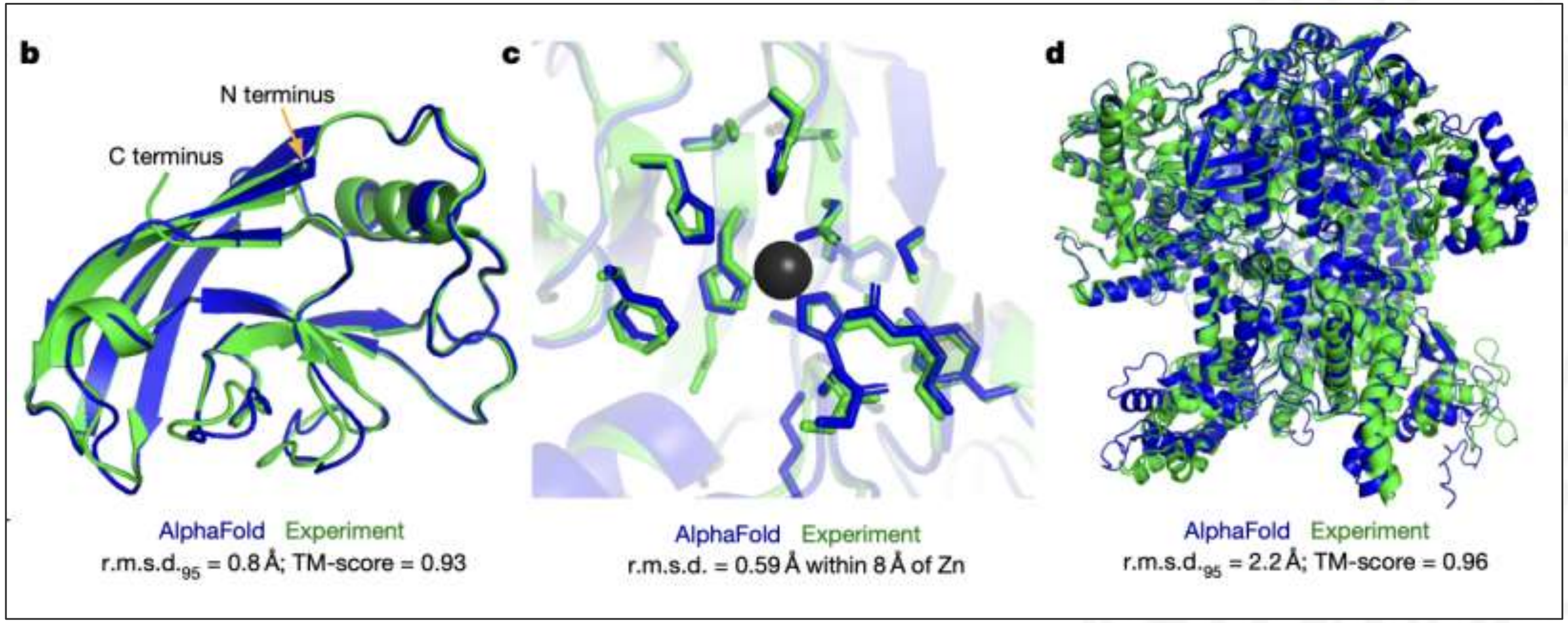
- AlphaFold assessed at CASP14, benchmarking against 87 protein domains.
- AlphaFold took the median Cα r.m.s.d.$_{95}$ of 0.6 Å.
- Compared to next-best methods: AlphaFold outperformed the next-best methods (the next-best median RMSD ~2.8 Å).
- AlphaFold achieved unprecedented accuracy on atomic-level.

$$\boxed{\mathrm{RMSD}\left(=\frac{1}{N}x_i - x_i^{r}\right)^2}$$



Structure Predictions and Side-chain Accuracy (Illustrations)

- AlphaFold predicts not only backbone structures but also side-chain orientations accurately, which is important for functional predictions.
- Example from CASP target T1044 shows accurate prediction of a zinc-binding site without explicit modeling of metal ions.
- Visual Examples:



AlphaFold Experiment
r.m.s.d.$_{95}$ = 0.8 Å; TM-score = 0.93

AlphaFold Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn

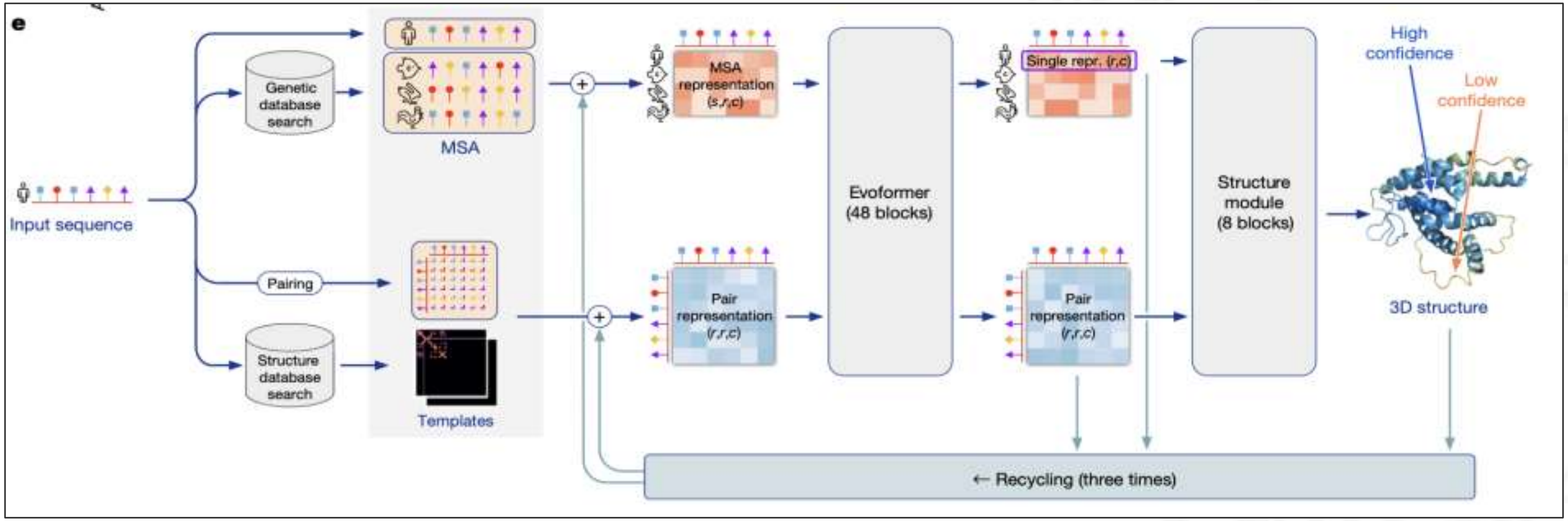
AlphaFold Experiment
r.m.s.d.$_{95}$ = 2.2 Å; TM-score = 0.96

Side-chain∑RMSD accuracy: ≈1.5 Å

AlphaFold Network Architecture Overview

Content:

- Two primary modules:
  - ❖ Evoformer (48 blocks).
  - ❖ Structure Module (8 blocks)
- Inputs: Multiple sequence alignments (MSAs), pairwise residue interactions, structural templates.
- Output: 3D atomic protein structure predictions, refined through iterative recycling (3 cycles).


e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

Recent PDB Structures Validation (Generalization Capability)

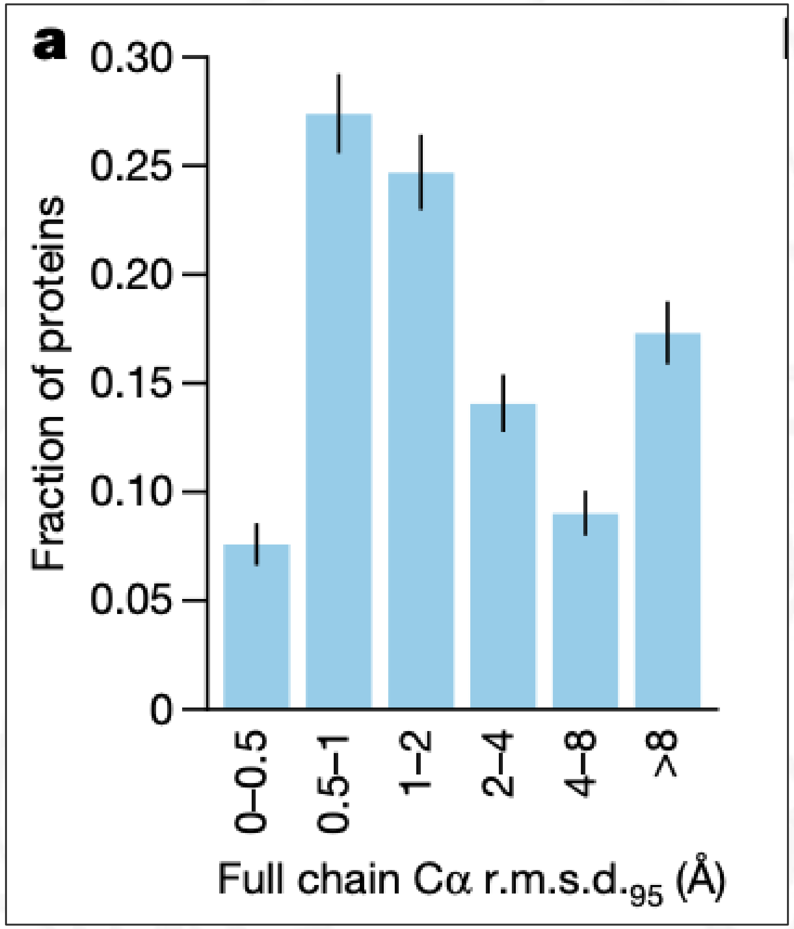
- The accuracy of AlphaFold is excellent on previously unseen structures and shows good generalization.
- Accuracy of AlphaFold assessed on structures deposited in the Protein Data Bank (PDB) after the CASP14 training cutoff.

➢ **Validation Overview:**

- **Dataset**: Protein structures added to the **Protein Data Bank (PDB)** after AlphaFold's training period.
- **Metric Used**: Cα RMSD<sub>95(Root Mean Square Deviation of Alpha Carbon atoms after excluding 5% outliers) across entire protein chains.
- **Evaluation**: Compared predicted structures to experimental structures.

➢ **Graph Interpretation:**

- The majority of proteins have **low RMSD values (0–2 Å)**, indicating **high structural accuracy**.
- A smaller fraction shows deviations >4Å, often corresponding to **intrinsically disordered or flexible proteins**.

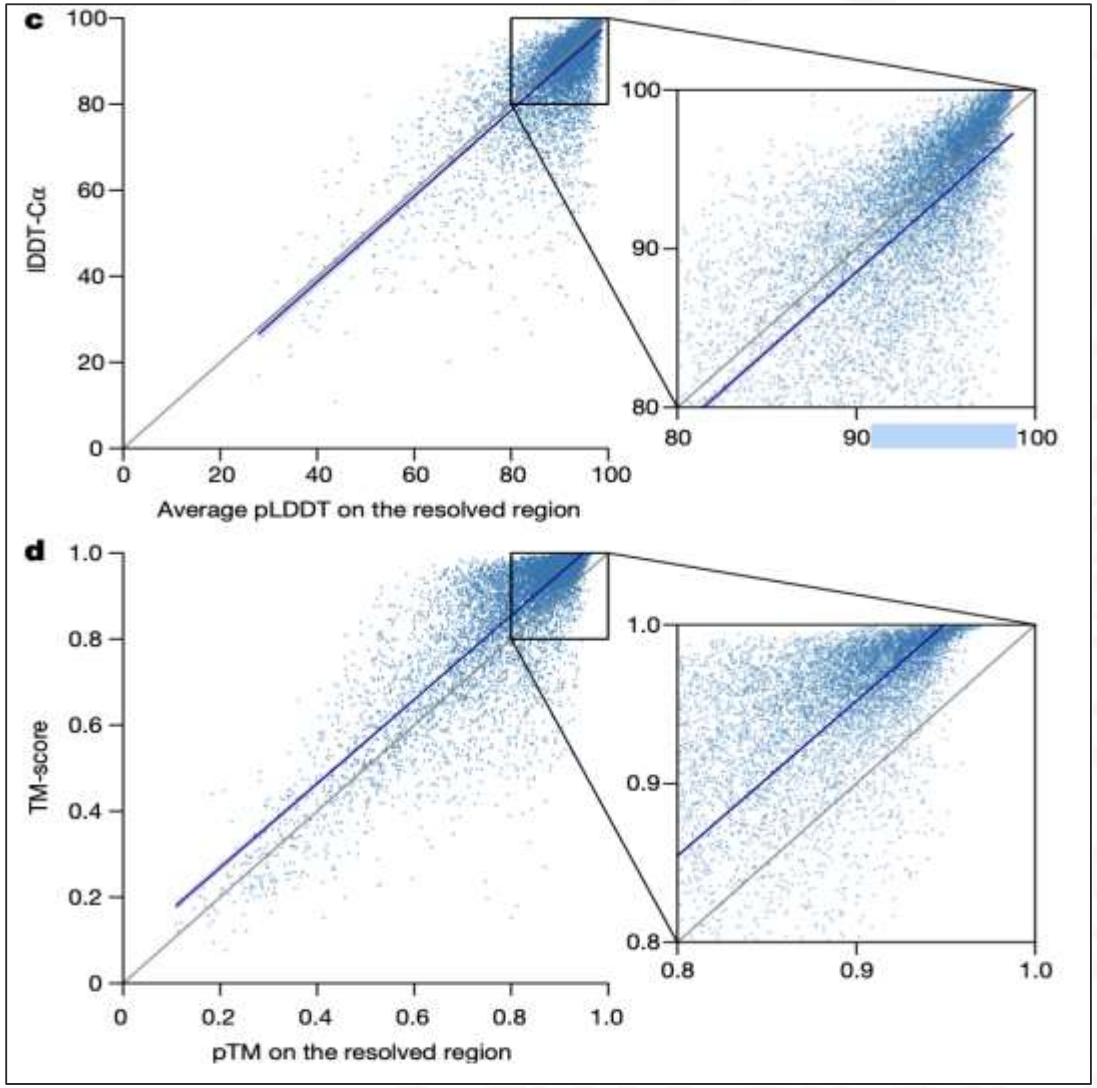


Confidence Estimation with pLDDT

- pLDDT: predicted Local Distance Difference Test, an AlphaFold-generated confidence score for per-residue predictions.

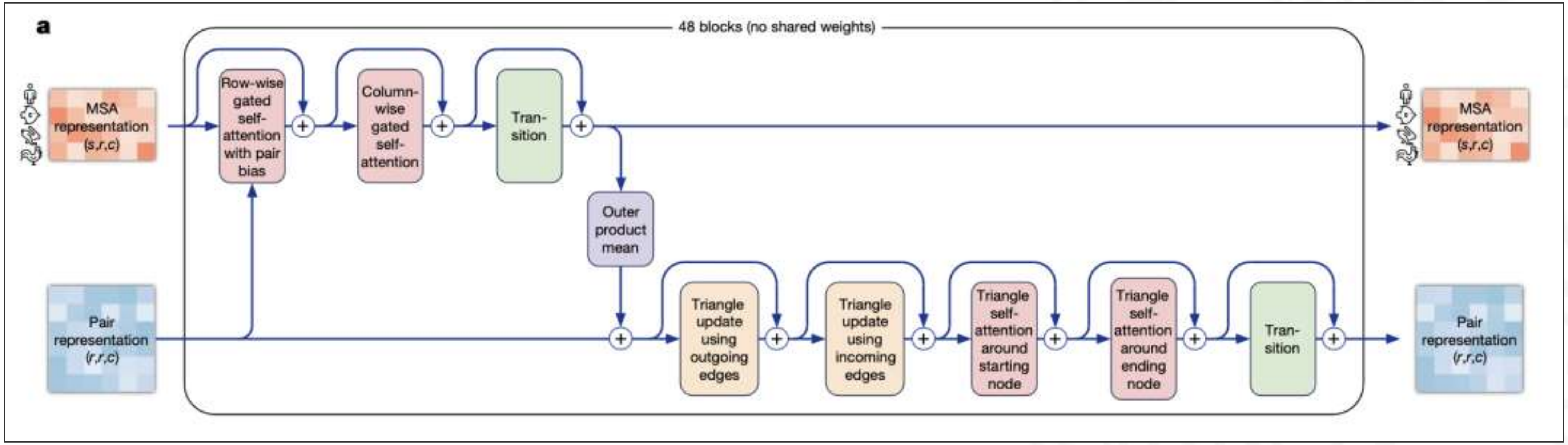
- Strong correlation between predicted pLDDT and actual accuracy (Pearson correlation r ~0.76–0.85).

$$\text{TM-score} = 0.99 \times pLDDT - 1.17$$



Evoformer – Network Core Principle

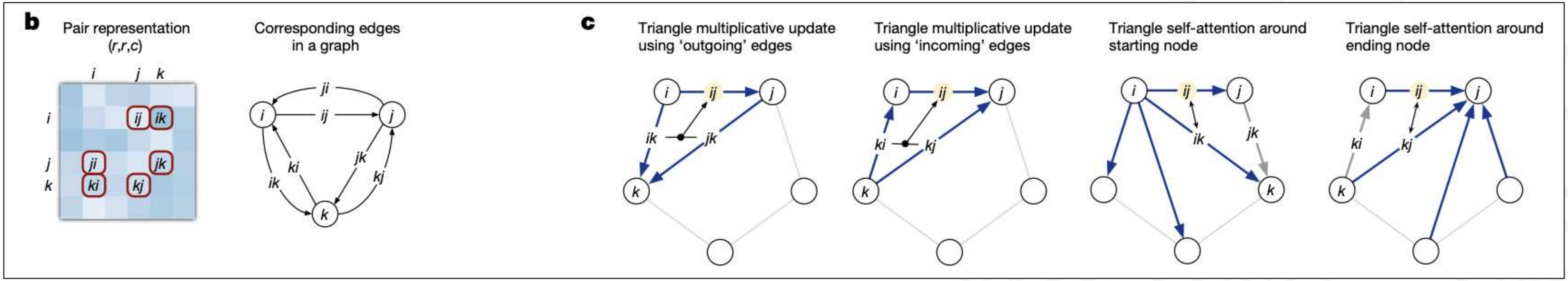
- Evoformer processes MSA and pairwise residue interactions simultaneously.
- Axial gated self-attention applied on MSA representation.
- Triangle updates and attention mechanisms enforce geometric consistency within pair representations.


a
48 blocks (no shared weights)
MSA representation (s,r,c)
Row-wise gated self-attention with pair bias
Column-wise gated self-attention
Tran-sition
Outer product mean
Pair representation (r,r,c)
Triangle update using outgoing edges
Triangle update using incoming edges
Triangle self-attention around starting node
Triangle self-attention around ending node
Tran-sition
MSA representation (s,r,c)
Pair representation (r,r,c)

Triangle multiplicative update:

$$z_{jnew} = f(z_{ij}, z_{ik}, z_{kj})$$

Pair representation (r,r,c)


b
Pair representation (r,r,c)
i j k
i
j
k
ij ik
ji jk
ki kj
Corresponding edges in a graph
i j k
ji ij ik ki jk kj
c
Triangle multiplicative update using 'outgoing' edges
i ij j ik jk k
Triangle multiplicative update using 'incoming' edges
i ij j ki kj k
Triangle self-attention around starting node
i ij j ik jk k
Triangle self-attention around ending node
i ij j ki kj k

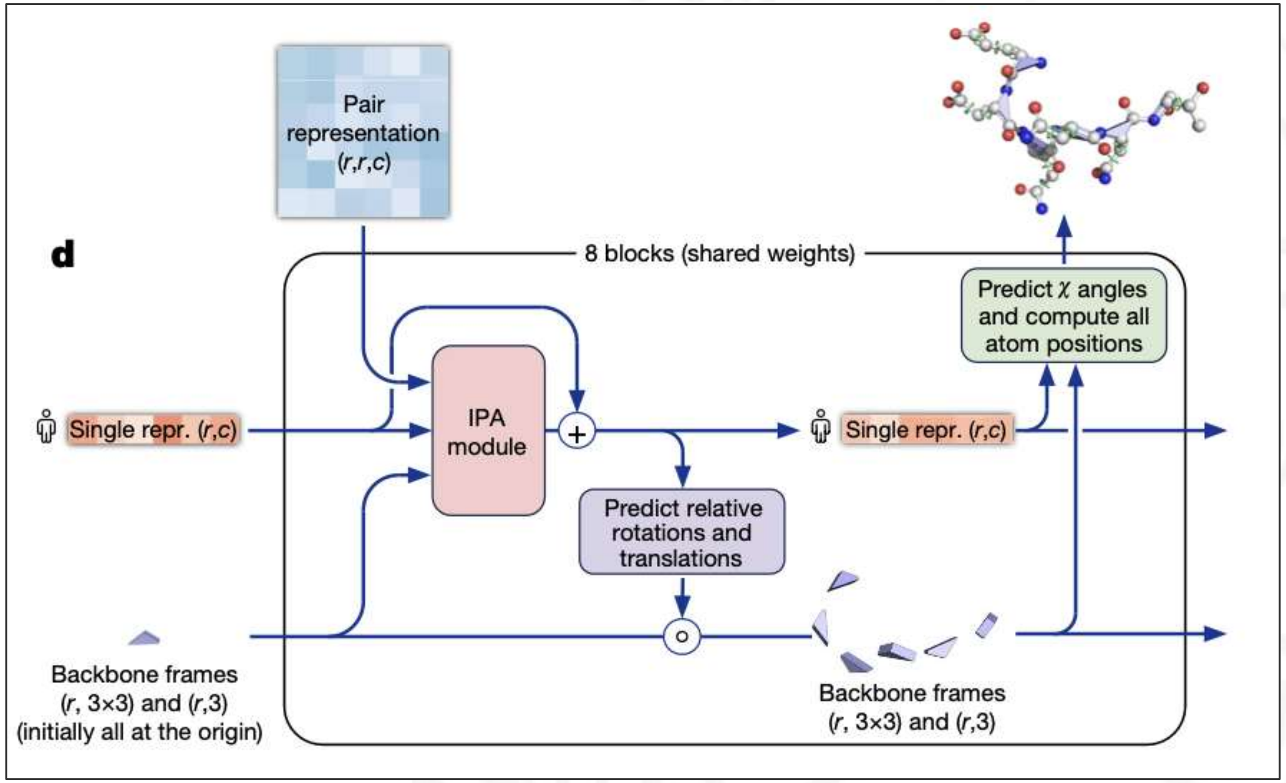
b. The pair representation interpreted as directed edges in a graph.

c. Triangle multiplicative update and triangle self-attention. The circles represent residues. Entries in the pair representation are illustrated as directed edges and in each diagram, the edge being updated is ij.

Invariant Point Attention (IPA)

- IPA is AlphaFold’s mechanism to perform attention in 3D space, crucial for precise structural geometry predictions.
- IPA uses query/key/value representations of residues as rigid transformations in 3D.
- Key Mathematical Insight: IPA attention scores defined as invariant to global translations/rotations:

$$\mathrm{IPA}_{ij} = \exp\left(-|R_i x_i - R_j x_j||^2\right)$$

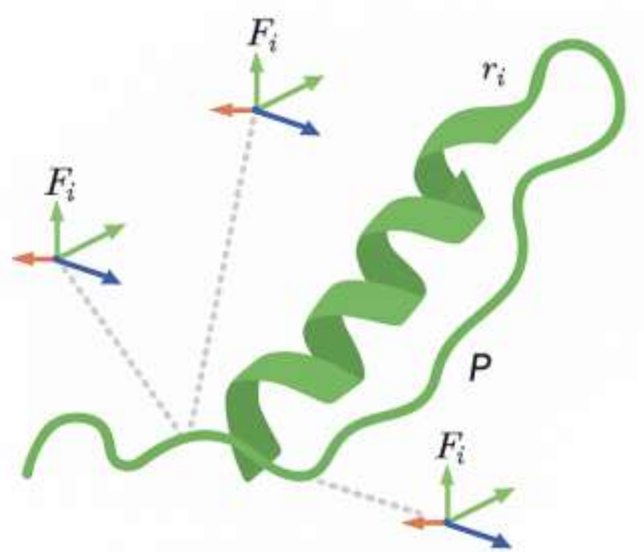


d
Pair representation (r,r,c)
8 blocks (shared weights)
Single repr. (r,c)
IPA module
+
Single repr. (r,c)
Predict χ angles and compute all atom positions
Predict relative rotations and translations
∘
Backbone frames (r, 3×3) and (r,3) (initially all at the origin)
Backbone frames (r, 3×3) and (r,3)

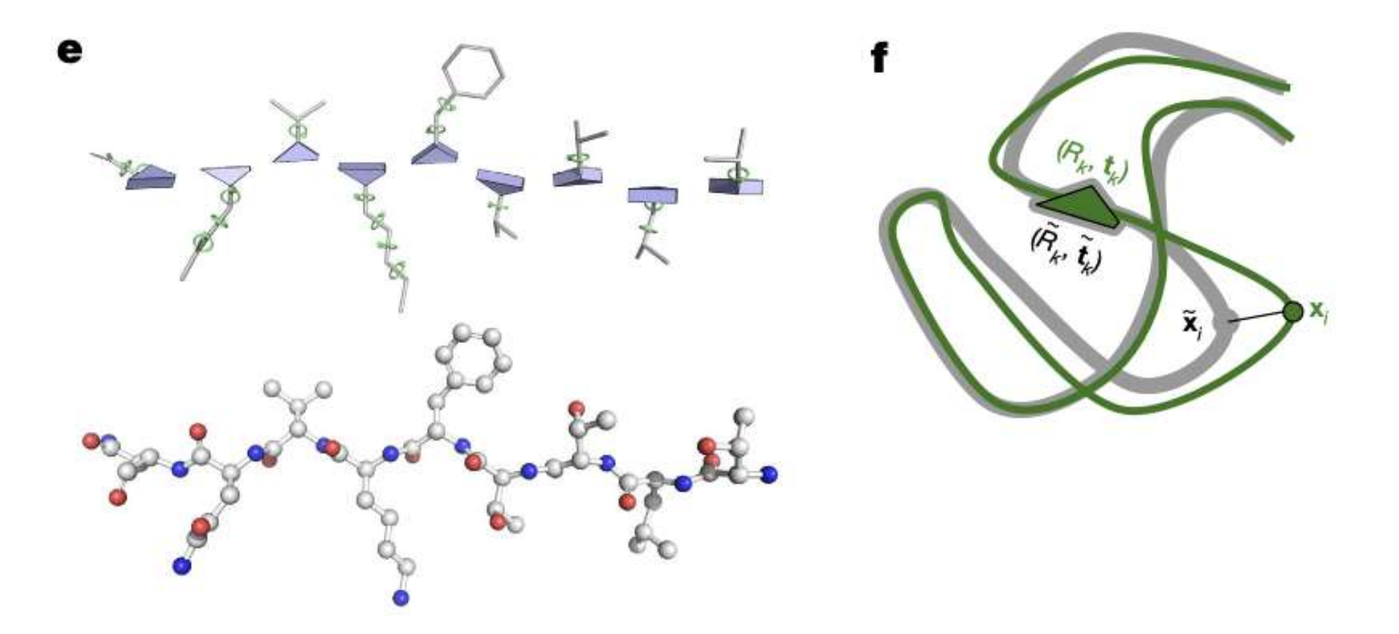
End-to-end Structure Prediction & FAPE Loss

- Structure module converts intermediate representations into precise 3D atomic structures.
- Loss function: Frame-Aligned Point Error (FAPE), measures structural deviations precisely.

$$FAPE = \frac{1}{N}\sum_{i}|R_i + t_i - (R_i^{\iota} g_i^{\alpha} + t_{\iota}^{i} \,||_1$$

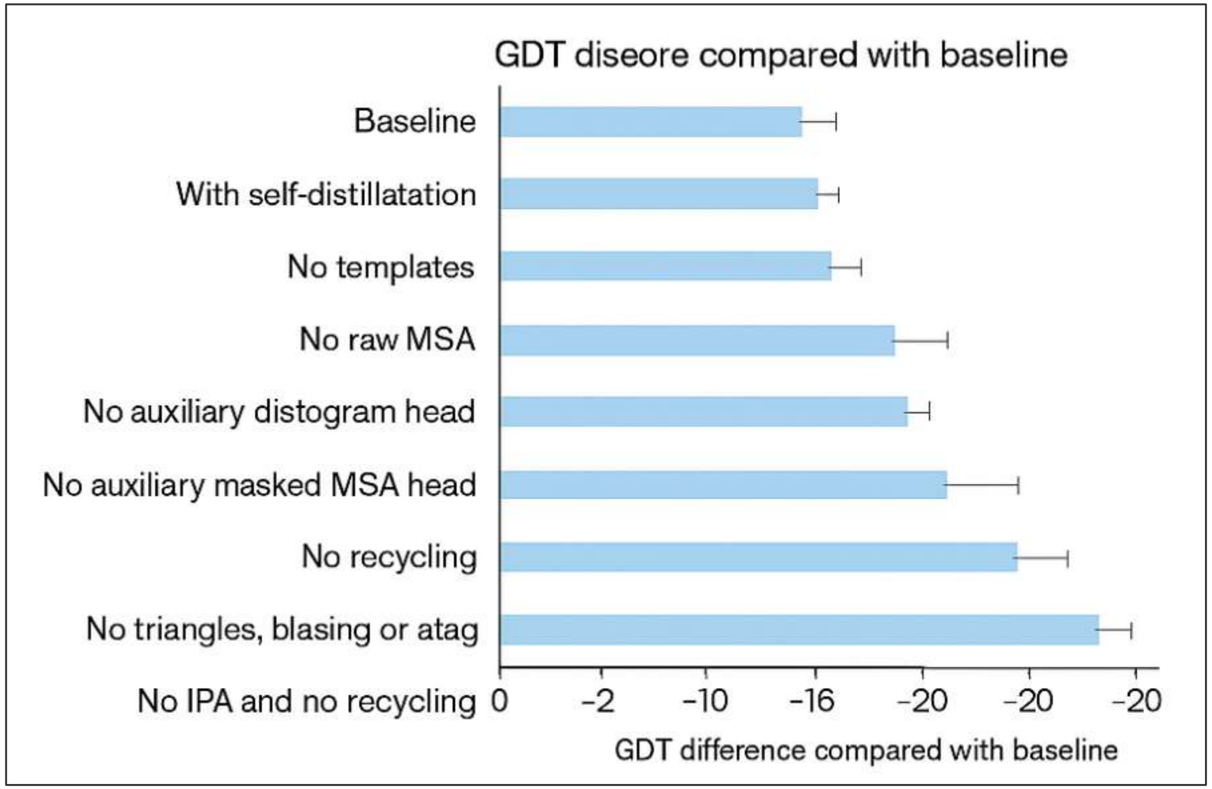


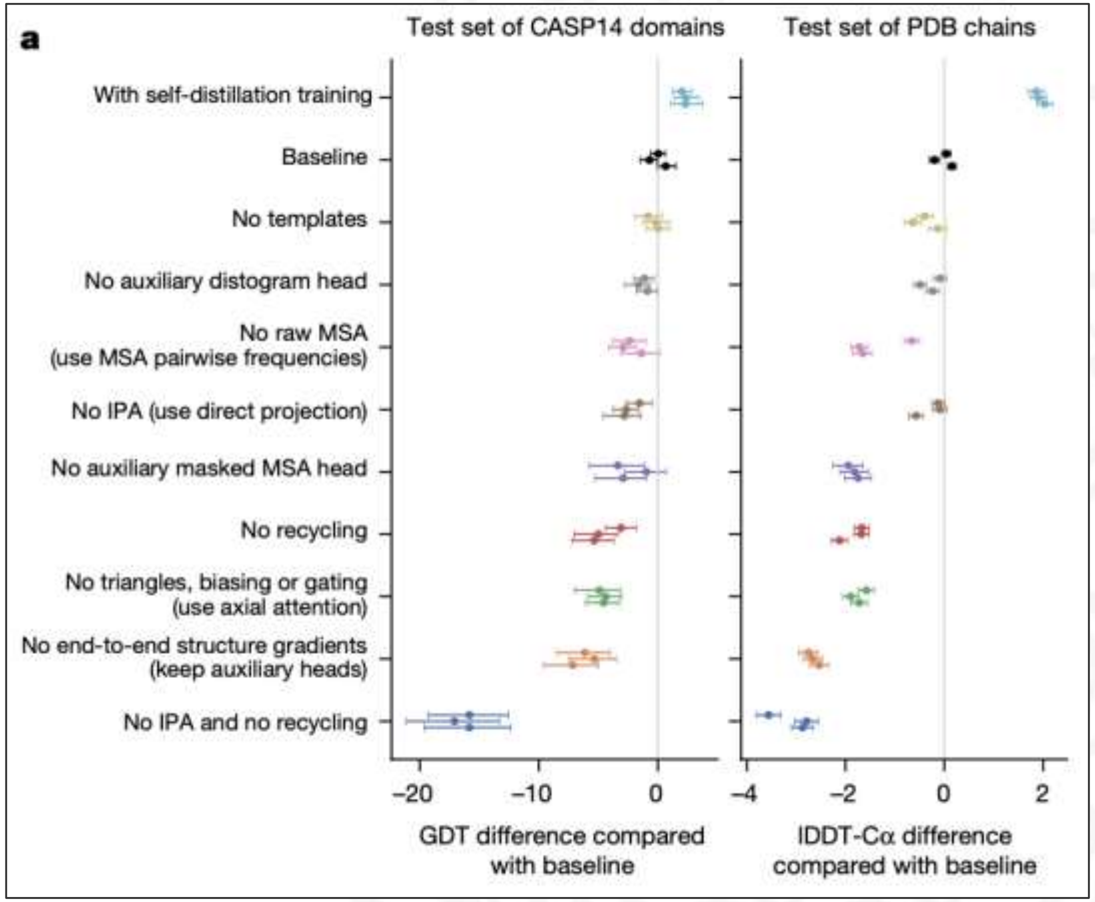
Residual frames and corresponding errors



Ablation Studies (Understanding AlphaFold's Improvements)

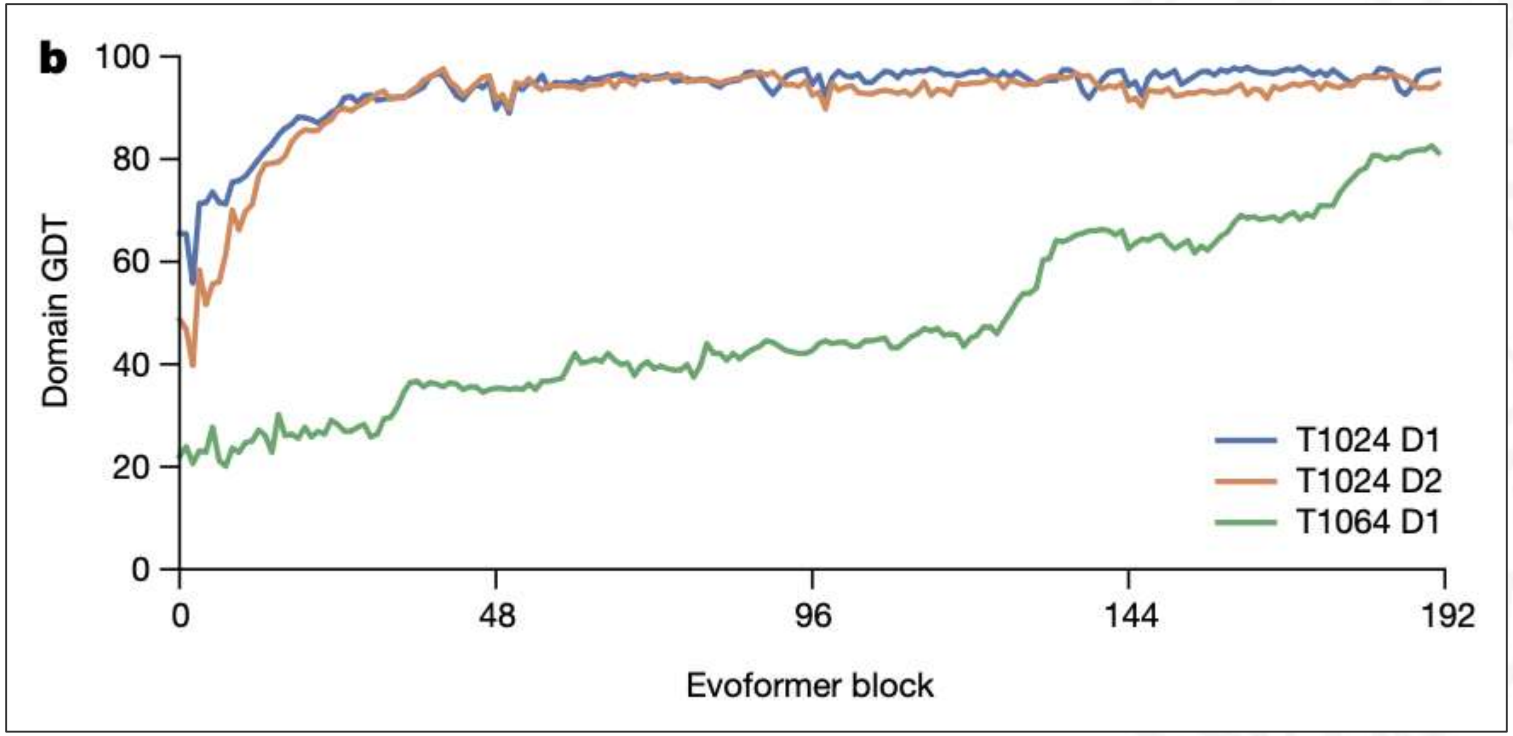
- Removal of key components significantly decreases AlphaFold’s accuracy.
- Recycling, IPA, and Evoformer triangle updates were critical for high-accuracy predictions.





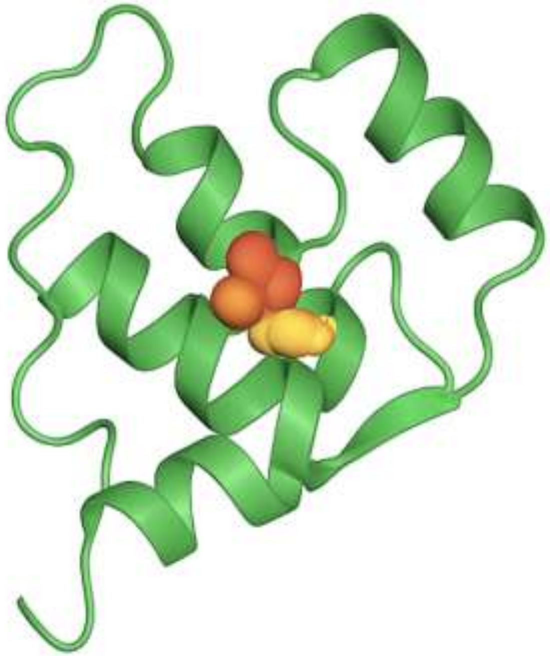
Neural Network Interpretability (Structural Trajectories)

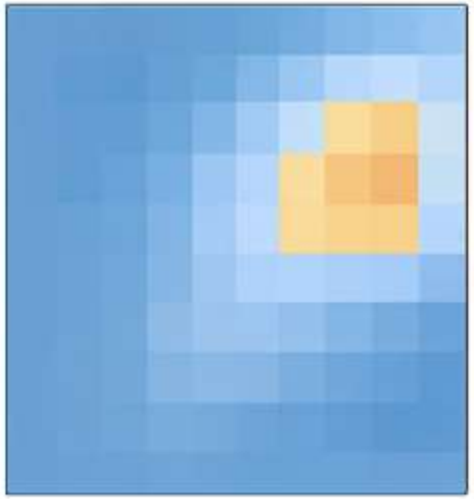
- Intermediate structure predictions demonstrate network refinement over multiple Evoformer blocks.
- AlphaFold progressively refines predictions until convergence is achieved.



Interpretability / Attention Map Visualization

- AlphaFold’s attention maps can reveal bioological features
- Example: focus on a binding site



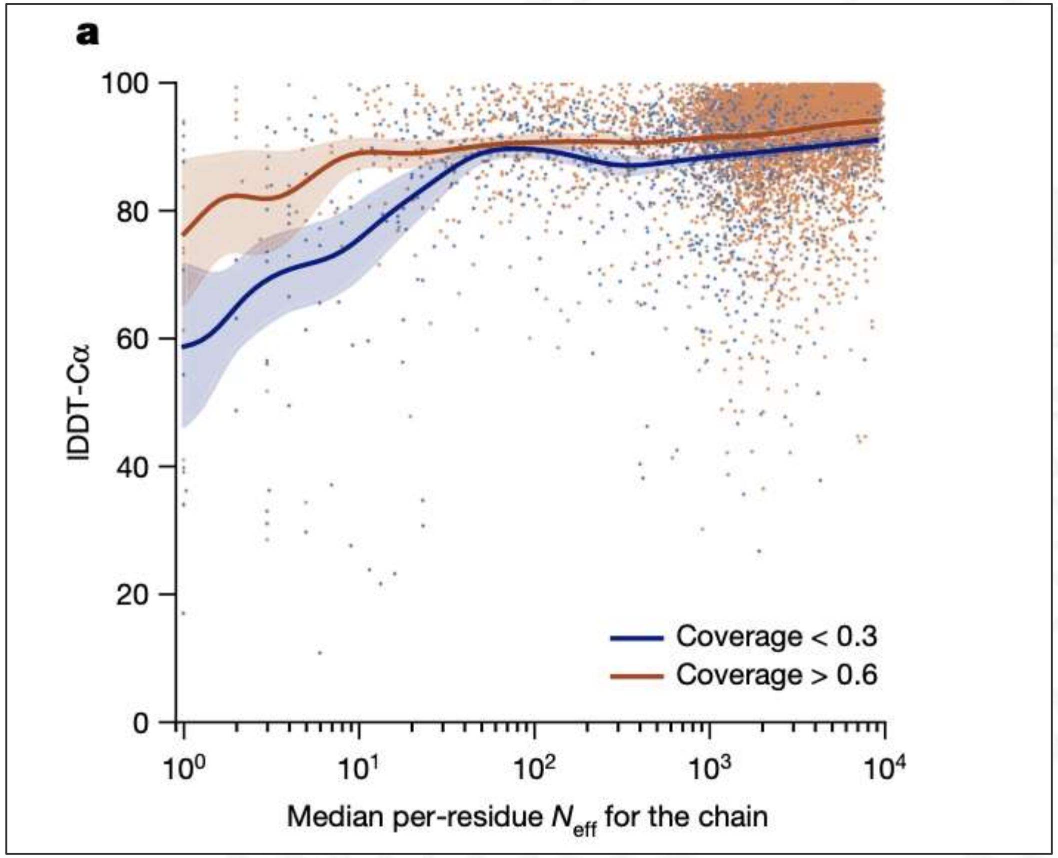


Attention map

MSA Depth and Cross-chain Contacts

- Accuracy declines with shallow MSAs (<30 sequences).
- Optimal accuracy achieved with MSA depth of around 100 sequences, plateaus beyond 500 sequences.
- Cross-chain predictions remain challenging.

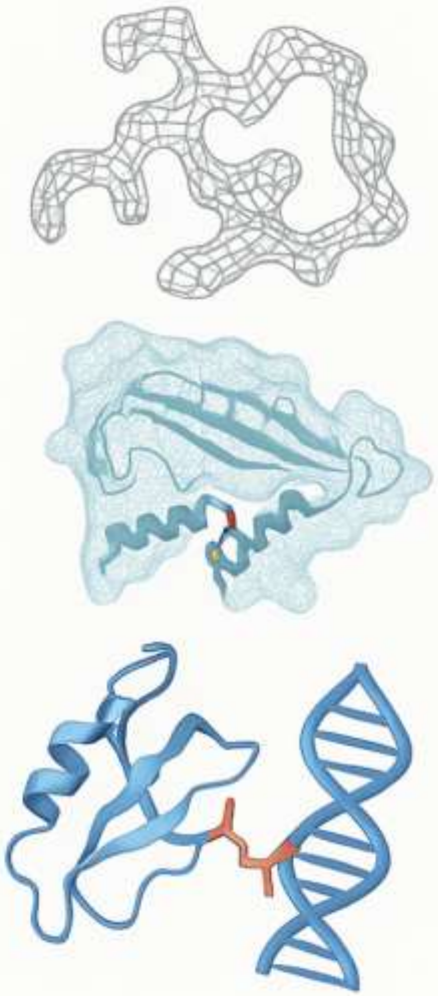
$$\text{IPA}=\left(\frac{-Rx_{q}-R_j}{x^2}\right)$$



Real-World Use Cases

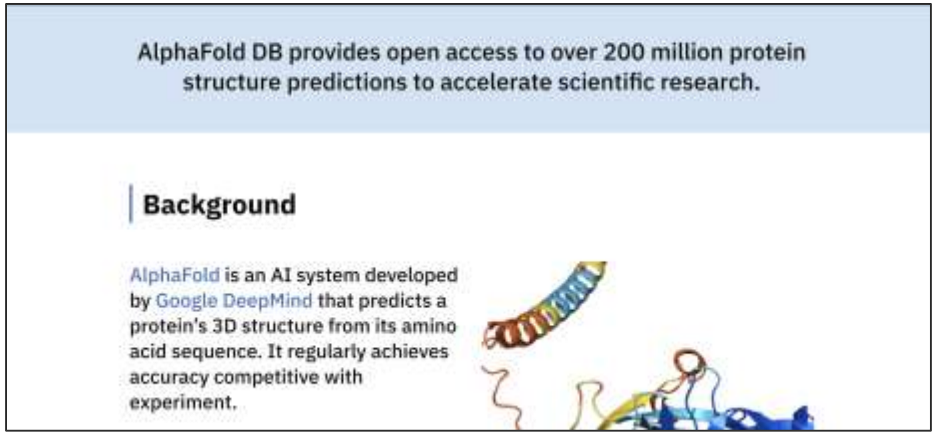
Applications:

- **X-ray Crystallography Support**
  AlphaFold-predicted models enable *molecular replacement* for solving X-ray diffraction data, reducing reliance on experimental phase information.
- **Fitting Cryo-EM Density Maps**
  Predicted protein structures can be accurately *fit into cryo-EM maps*, especially useful for low-resolution experimental data.
- **Drug Discovery & Target Validation**
  Used to predict 3D structures of drug targets, allowing *structure-based drug design* and *ligand docking* in early-phase drug discovery.
- **Mutation Impact Prediction**
  AlphaFold aids in understanding how *disease-associated mutations* (e.g., missense) impact protein structure and function.



AlphaFold Protein Structure Database (EMBL-EBI)

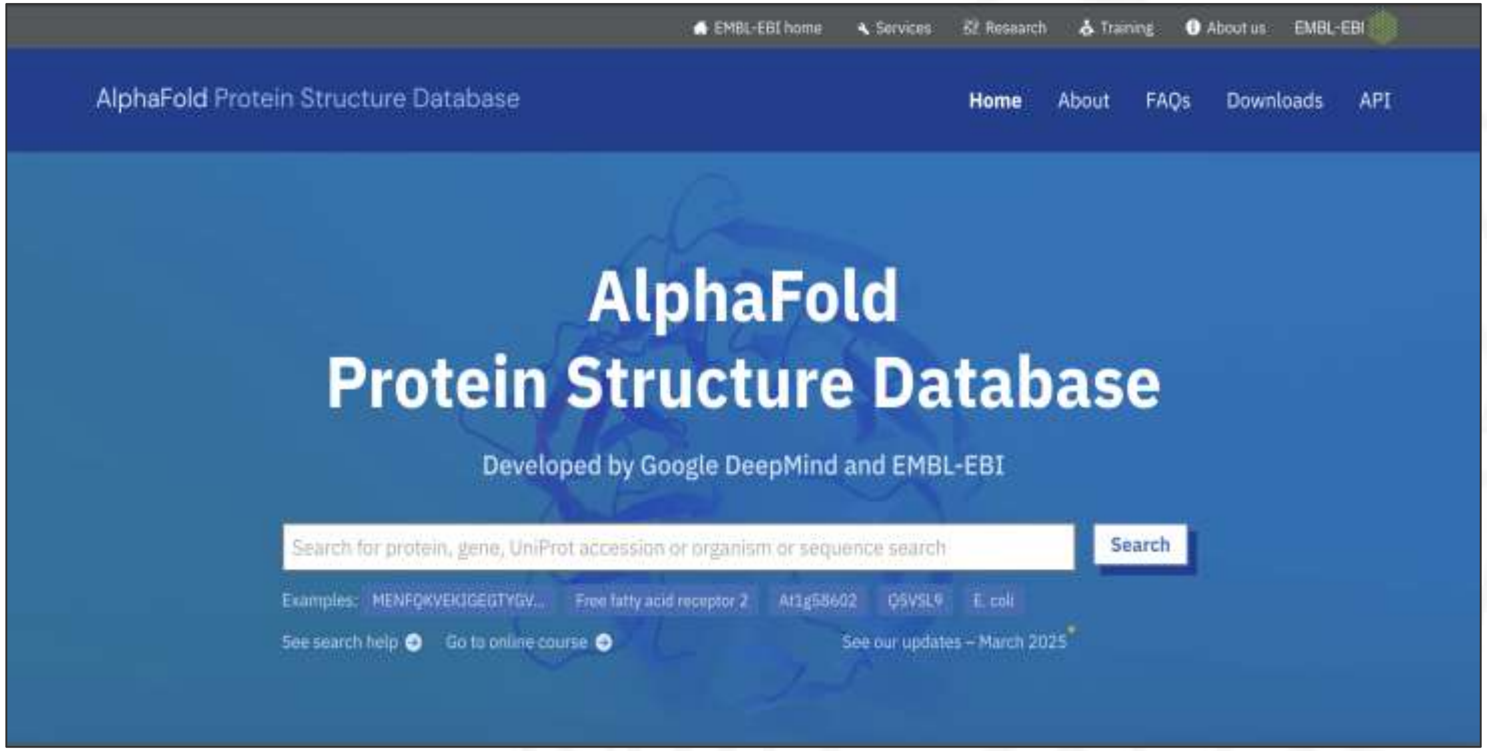
- ~200 million protein structures.
- Search by UniProt ID or gene.
- Free and open access.



AlphaFold DB provides open access to over 200 million protein structure predictions to accelerate scientific research.

Background

AlphaFold is an AI system developed by Google DeepMind that predicts a protein's 3D structure from its amino acid sequence. It regularly achieves accuracy competitive with experiment.



EMBL-EBI home
Services
Research
Training
About us
EMBL-EBI
AlphaFold Protein Structure Database
Home
About
FAQs
Downloads
API

AlphaFold
Protein Structure Database

Developed by Google DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism or sequence search
Search
Examples: MENFQKVEKIGEGTYGV... Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli
See search help
Go to online course
See our updates – March 2025

Reliability:

•Each prediction includes **pLDDT scores** (Predicted Local Distance Difference Test) for residue-level confidence:

- *>90*: High confidence
- *70–90*: Medium
- *<50*: Low (disordered regions)

Use Case:

Researchers use AlphaFold DB to validate targets for *antimicrobial resistance* and *oncology drug discovery*.

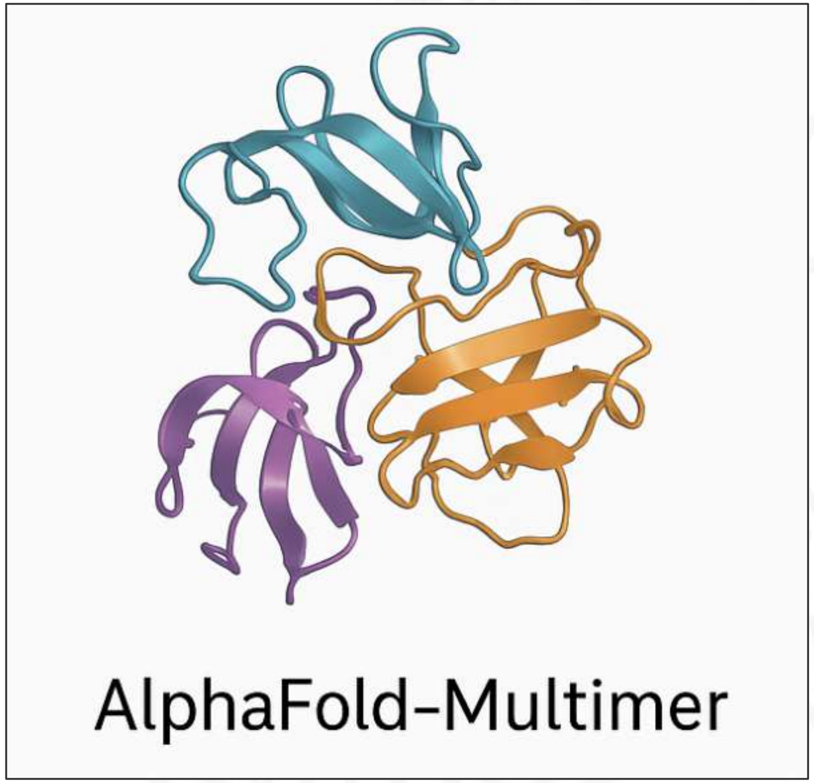
AlphaFold-Multimer

- ***Description:***
  - ➢ Extension for protein–protein complex prediction.
  - ➢ Adds inter-chain attention.
  - ➢ Works well for stable complexes.

✓ **Note:** Lacks explicit modeling of transient or flexible interactions.

Limitation:

- Currently less accurate for **transient**, **disordered**, or **flexible** complexes (e.g., signal transduction complexes).
- **Real Use Case:**
  AlphaFold-Multimer successfully predicted the **interleukin-12 receptor complex**, previously unresolved via experimental means.

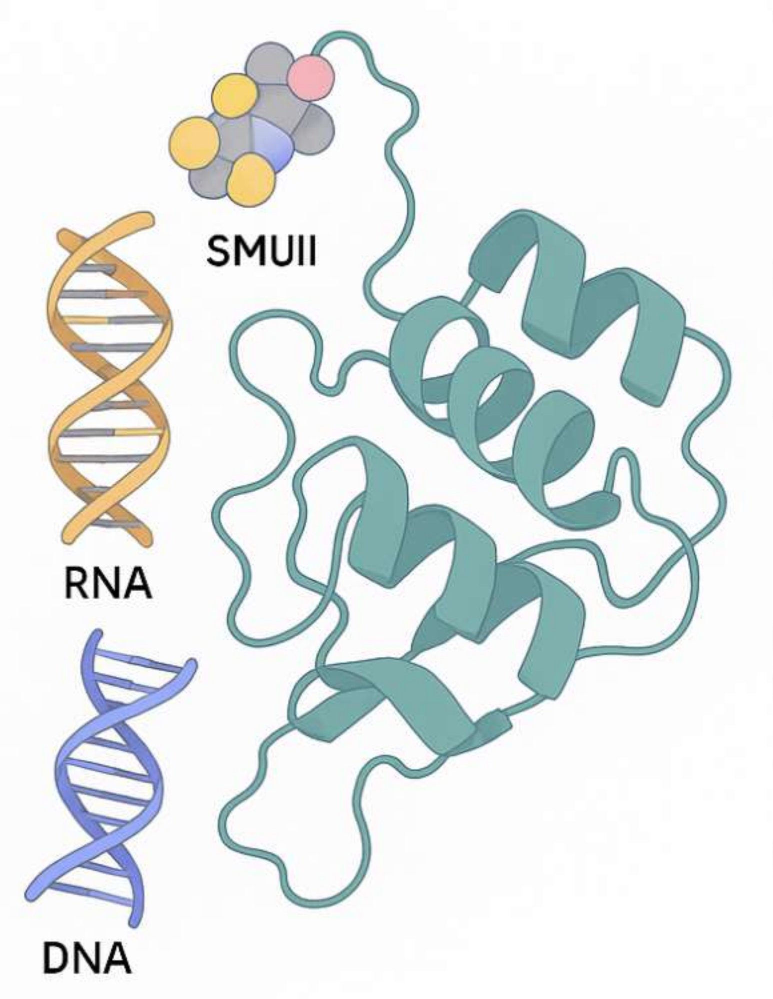


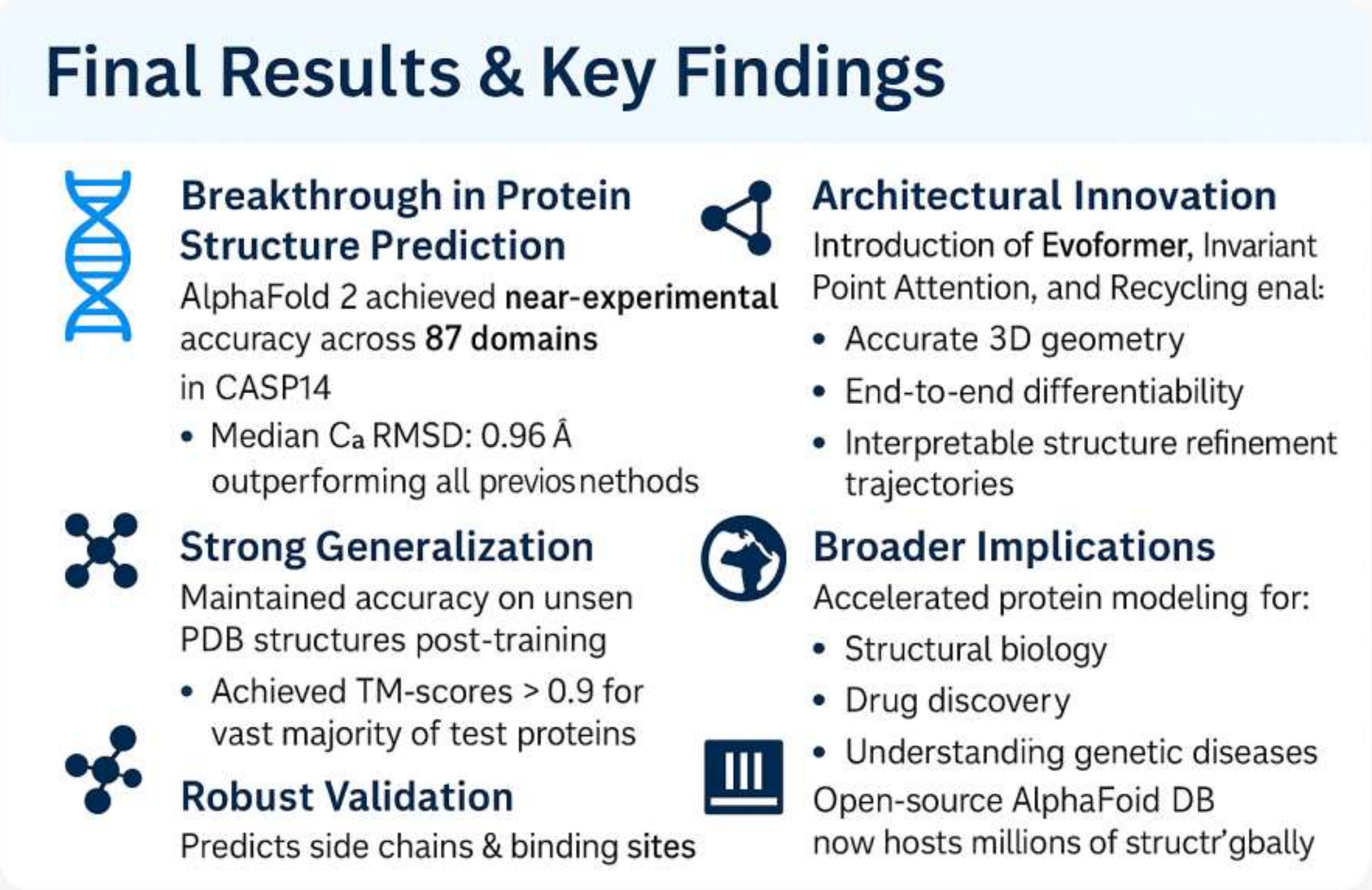
AlphaFold3 (Outlook) Coming soon / Future potential:

- Predicts **interactions with small molecules**, **RNA**, and **DNA** — unlocking new avenues in structural biology.
- Integrates **ligand binding** modeling and **cofactor coordination** (e.g., ATP, metals).
- Built using **diffusion-based architectures** + expanded **language modeling** of protein sequences..

Key Capabilities:

- **Multi-modal modeling**: protein + RNA/DNA + ligand
- **Biological relevance**: Regulatory proteins, ribonucleoprotein complexes, DNA-repair enzymes





Final Results & Key Findings

Breakthrough in Protein Structure Prediction

AlphaFold 2 achieved **near-experimental** accuracy across **87 domains** in CASP14

- Median C_a RMSD: 0.96 Å outperforming all previos nethods

Strong Generalization

Maintained accuracy on unsen PDB structures post-training

- Achieved TM-scores > 0.9 for vast majority of test proteins

Robust Validation

Predicts side chains & binding sites

Architectural Innovation

Introduction of **Evoformer**, Invariant Point Attention, and Recycling enal:

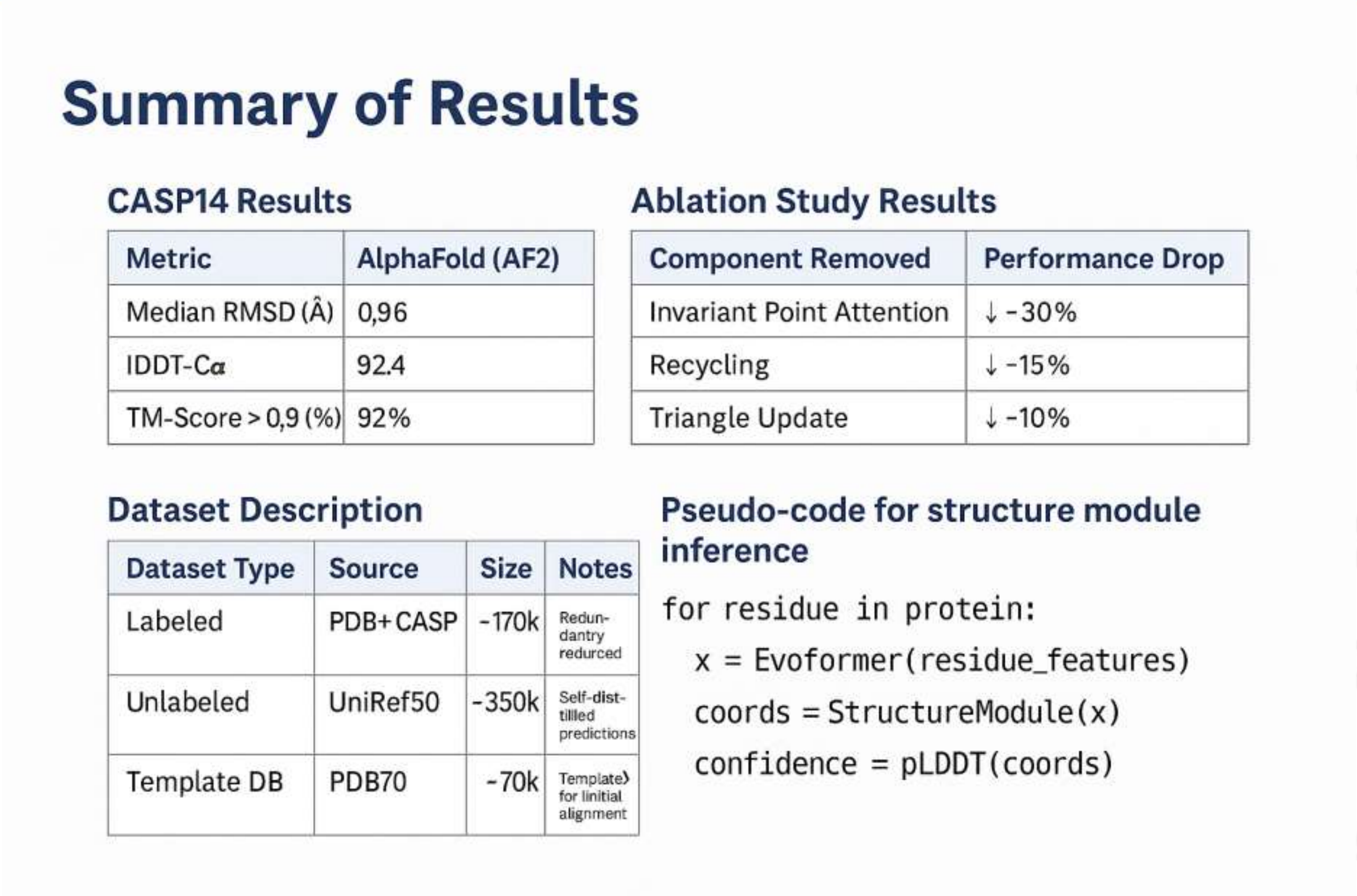
- Accurate 3D geometry
- End-to-end differentiability
- Interpretable structure refinement trajectories

Broader Implications

Accelerated protein modeling for:

- Structural biology
- Drug discovery
- Understanding genetic diseases

Open-source AlphaFoid DB now hosts millions of structr'gbally



Summary of Results

CASP14 Results

Metric	AlphaFold (AF2)
Median RMSD (Â)	0,96
IDDT-Cα	92.4
TM-Score > 0,9 (%)	92%

Ablation Study Results

Component Removed	Performance Drop
Invariant Point Attention	↓ -30%
Recycling	↓ -15%
Triangle Update	↓ -10%

Dataset Description

Dataset Type	Source	Size	Notes
Labeled	PDB+ CASP	-170k	Redundantry reducred
Unlabeled	UniRef50	-350k	Self-disttillled predictions
Template DB	PDB70	-70k	Template) for linitial alignment

Pseudo-code for structure module inference

```
for residue in protein:
    x = Evoformer(residue_features)
    coords = StructureModule(x)
    confidence = pLDDT(coords)
```

THANK YOU