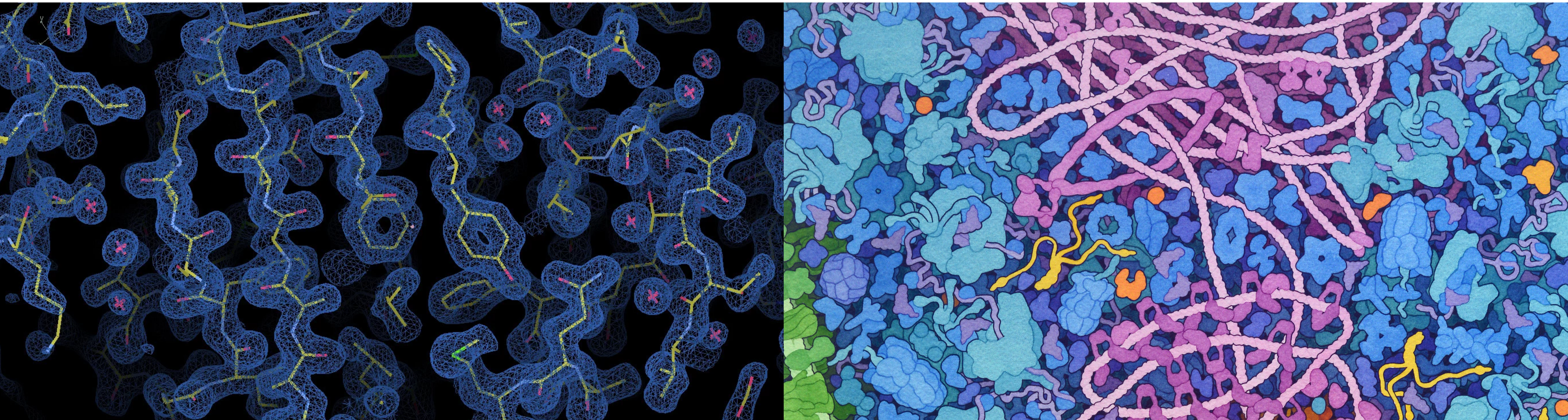


Breakthroughs in protein structure prediction & AlphaFold



Ellen Zhong
Department of Computer Science
Princeton University

Fall 2026
COS 597Y Week 2



PRINCETON
UNIVERSITY

Course Logistics

- Make sure you are added to Slack and have filled out the Google form!
- From the poll: **Weekly classes on Wed @ 2:55-4:55p**
 - TBD Thanksgiving + last week of class scheduling
 - Contact me if you have a conflict
- Presentation sign up + schedule will be shared by Friday this week
 - Likely 2-3 students per week
 - Some weeks will have assigned papers, and some weeks will be topic-only for now
 - For topic-only weeks, presenters will propose papers for approval
 - Final paper selections should be approved and shared with the class ≥ 1 week in advance
- Course materials, readings, and pre-lecture questions will be posted on the course website

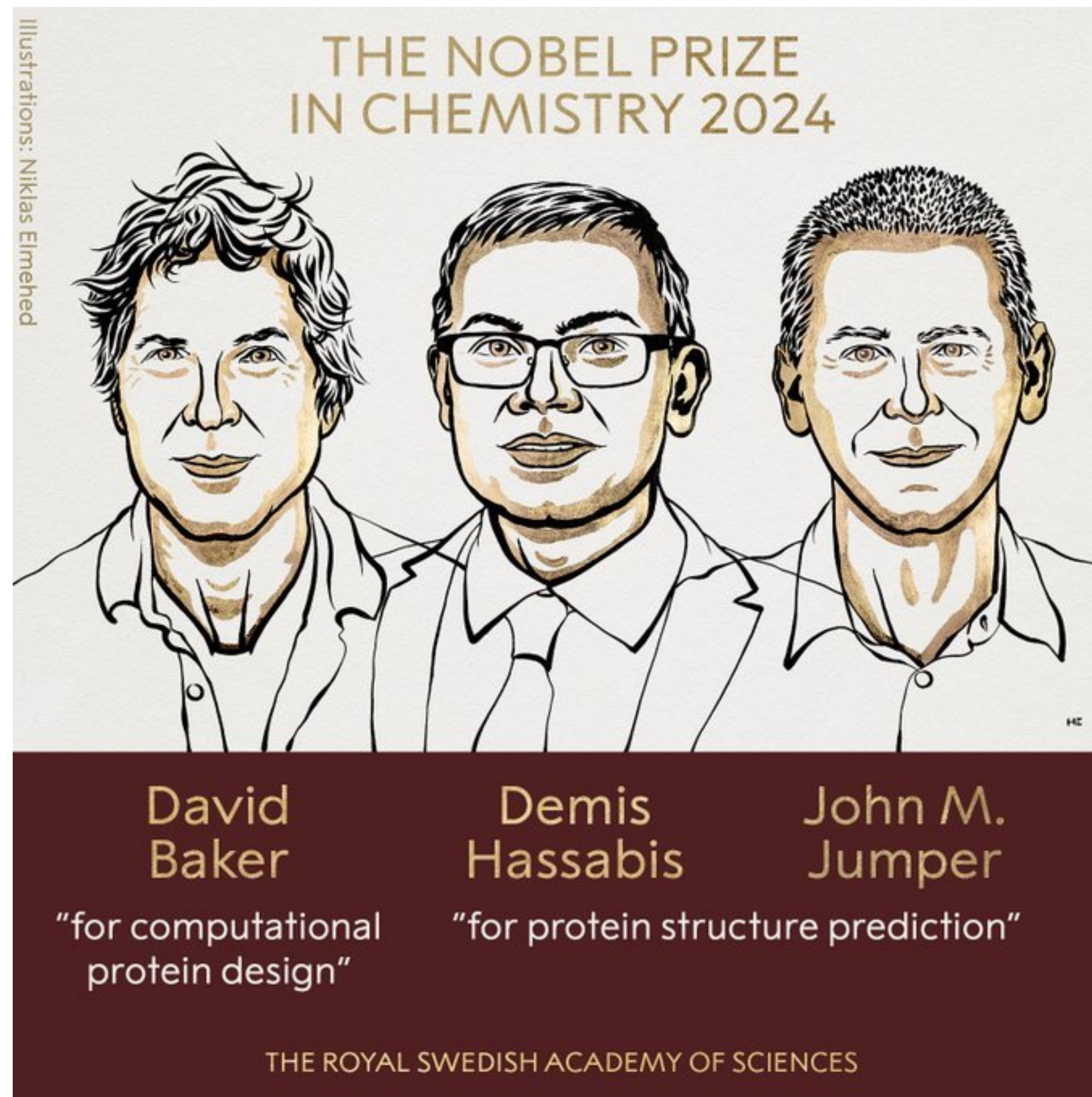
Next Week

- Generative modeling 101
- Required reading:
 - VAE paper
 - CryoDRGN ICLR paper
- Looking for presenters

Outline for this lecture

- Bio 101 and the protein folding problem
- The CASP Benchmark
- AF1
- AF2
- AF3 and beyond
- Discussion

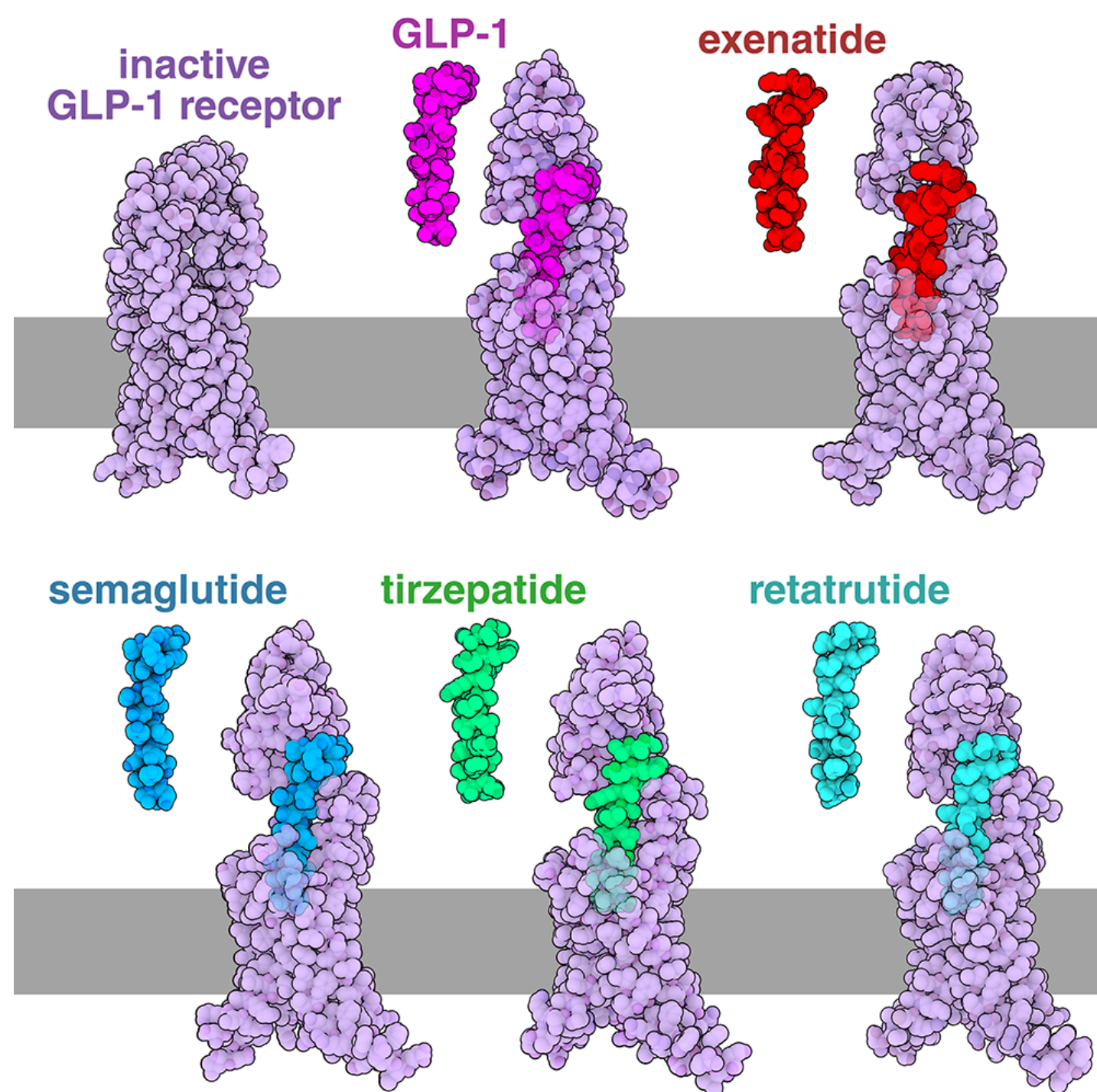
Spring 2024 ...



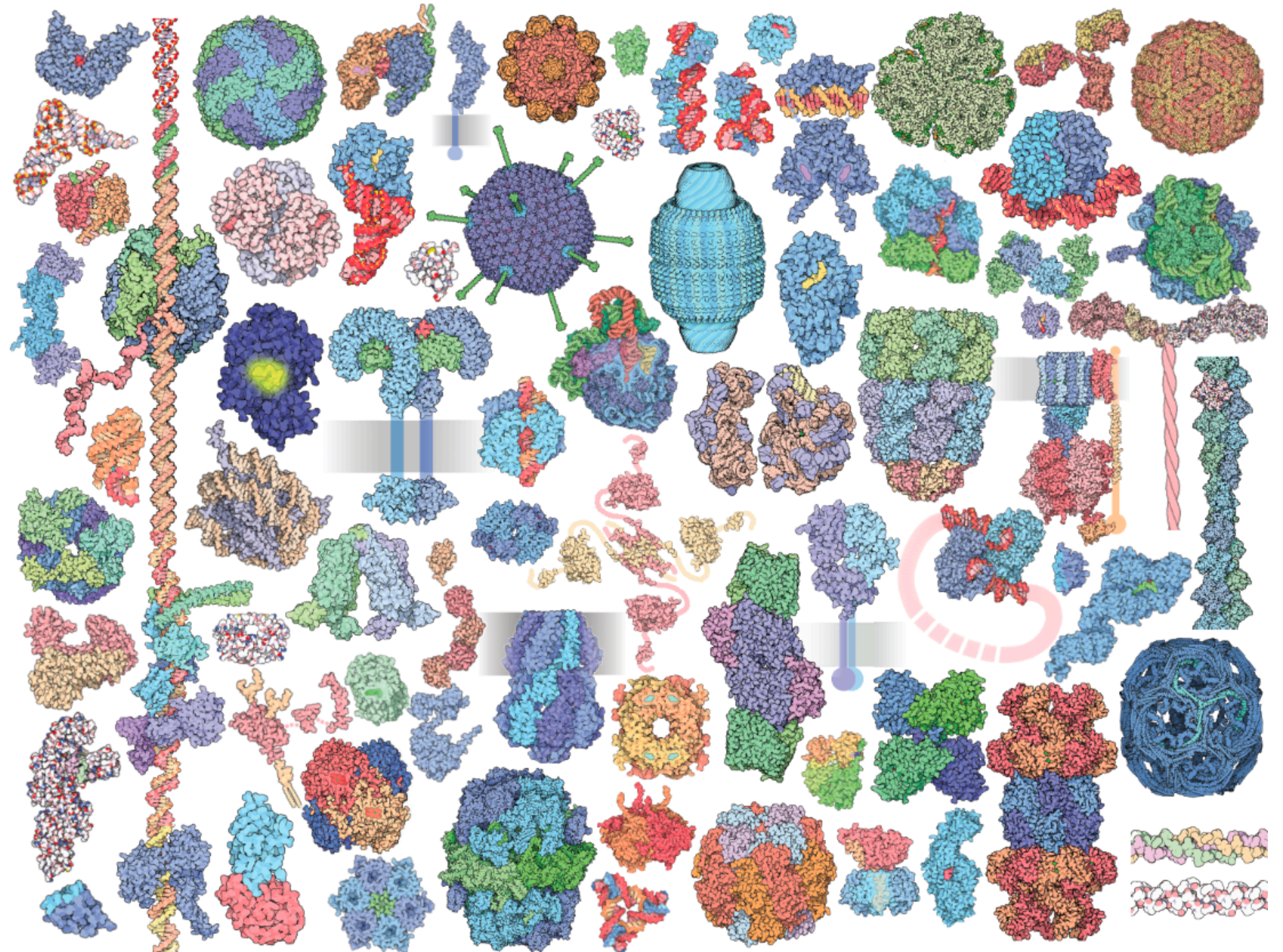
What are proteins?

All essential biological processes are carried out by proteins and protein complexes

- Fundamental molecules of life
- Medicine and health
- Nanotech and biotech



PDB-101 Molecule of the Month



Goodsell et al. PLoS Biology 2015.

The central dogma of molecular biology

DNA sequence

ATGCACTTGAGCAGGGAAGAAA...



RNA sequence

AUGCACUUGAGCAGGGAAGAAA...



Protein sequence

MSTAGKVIKCKAAVLWELKKPF...

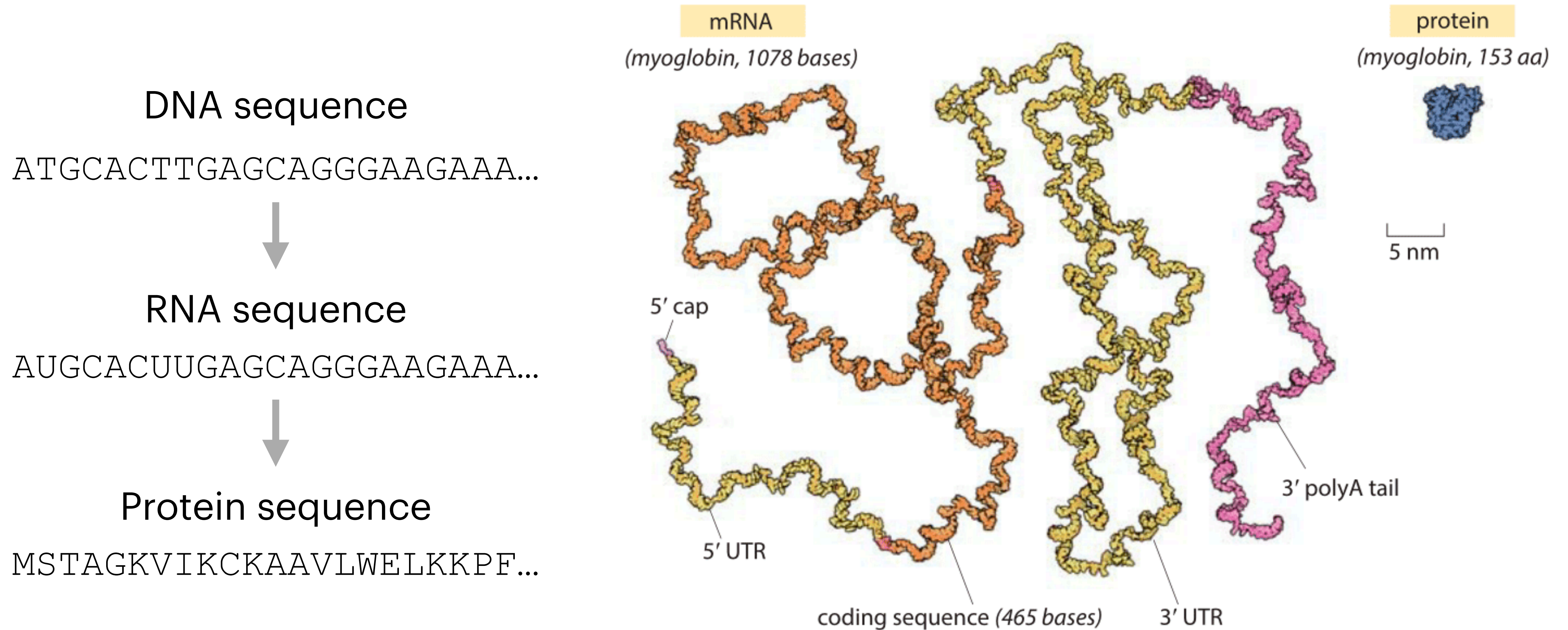
Human genome:

- * Contains around 3 billion base pairs
- * Encodes ~20k genes

Proteins are the final product of the genetic information flow

Modern molecular biology research: how is life implemented by our genetic code?

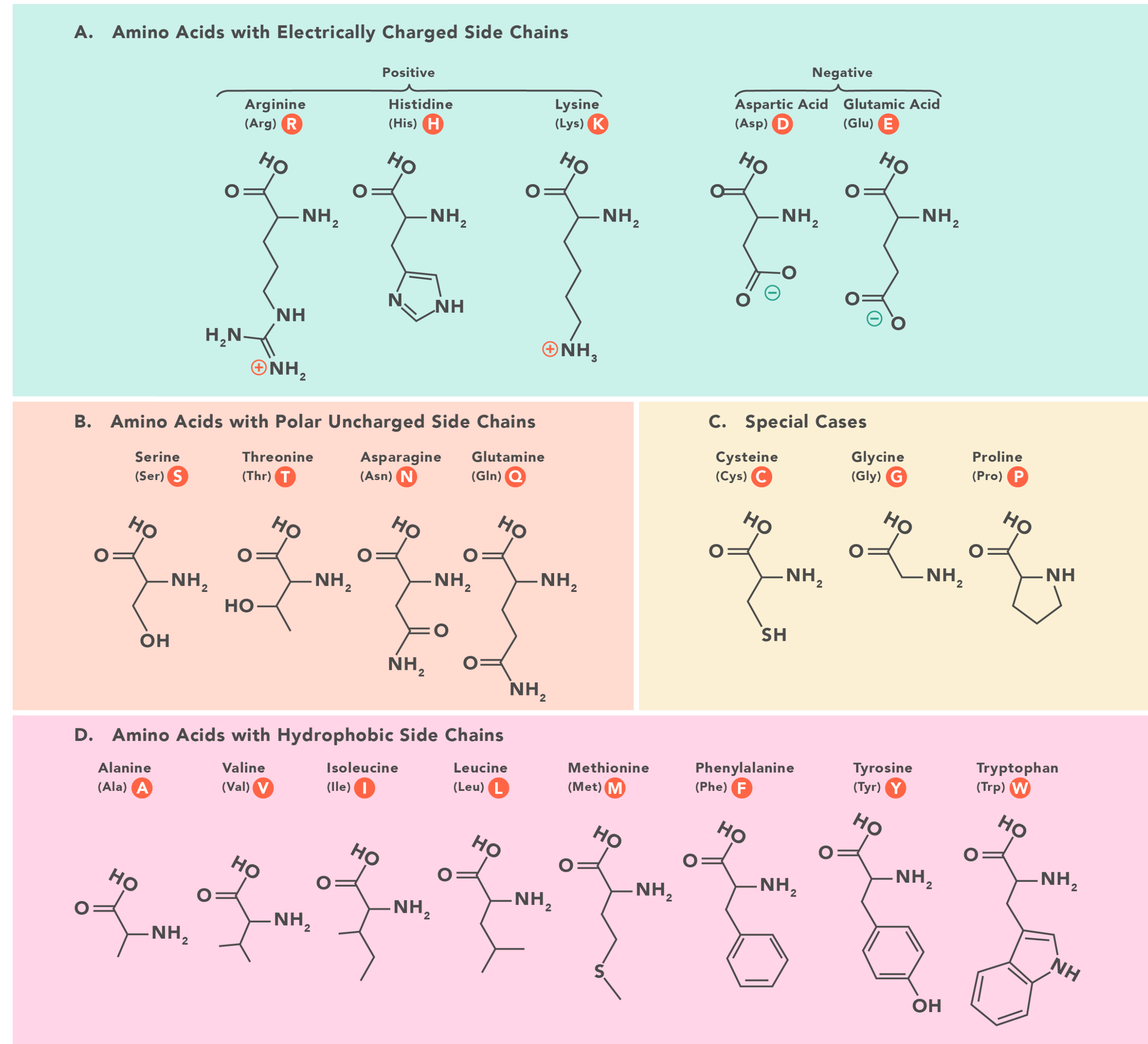
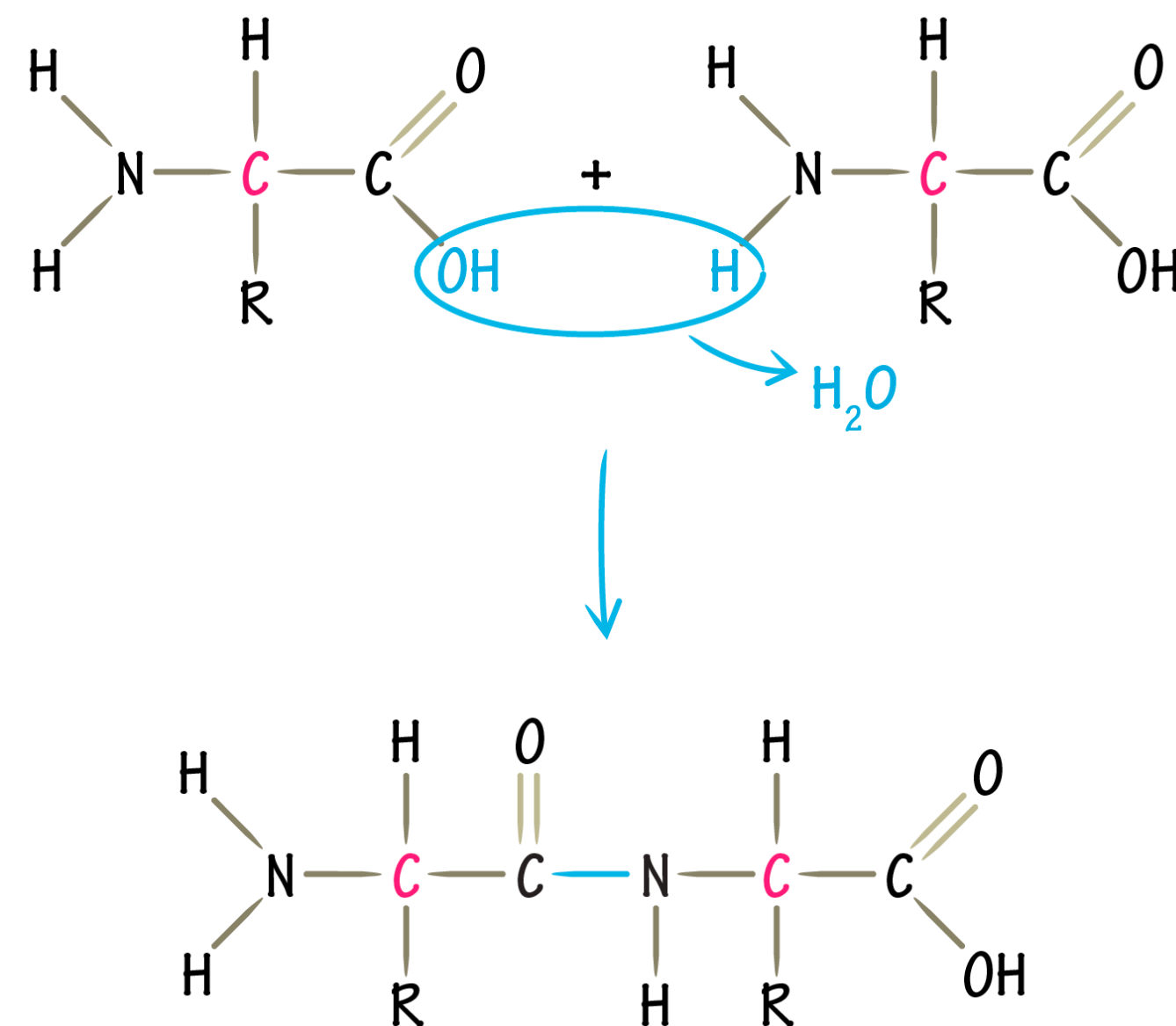
Structural biology: The study of proteins and other biomolecules through their 3D structure



What are proteins?

- A linear sequence of amino acids polymerized in a chain
- An alphabet of twenty possible amino acids
 - Common *backbone* but different *side chains*
- Various non-covalent interactions and other forces drive folding of the chain into a globular 3D structure*

Peptide Bond Formation

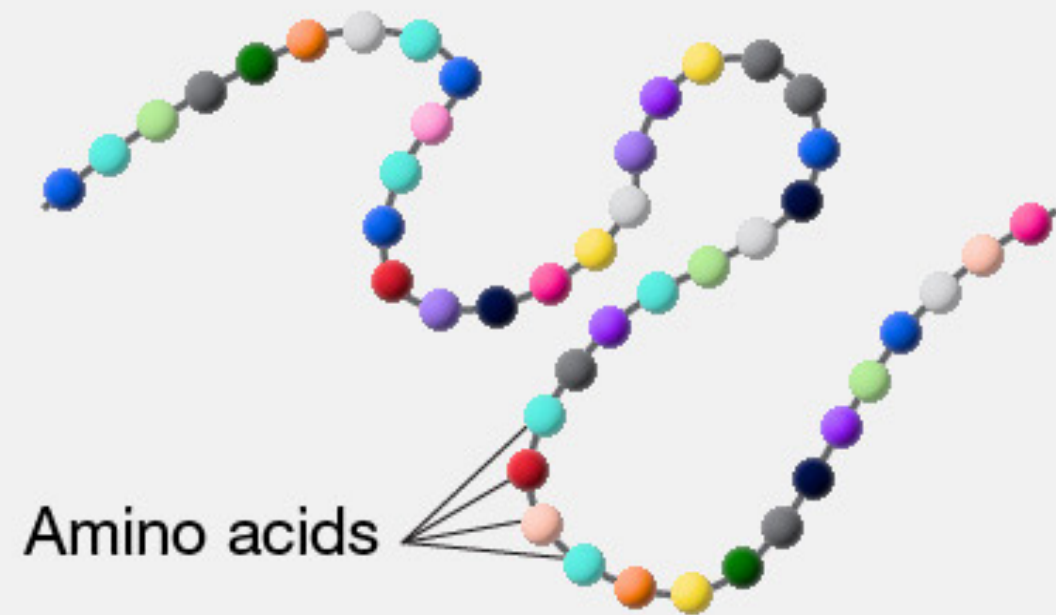


* There have been many different paradigms for thinking about protein folding. See Afinsen's hypothesis, Dill et al. 2008, [The Protein Folding Problem](#)

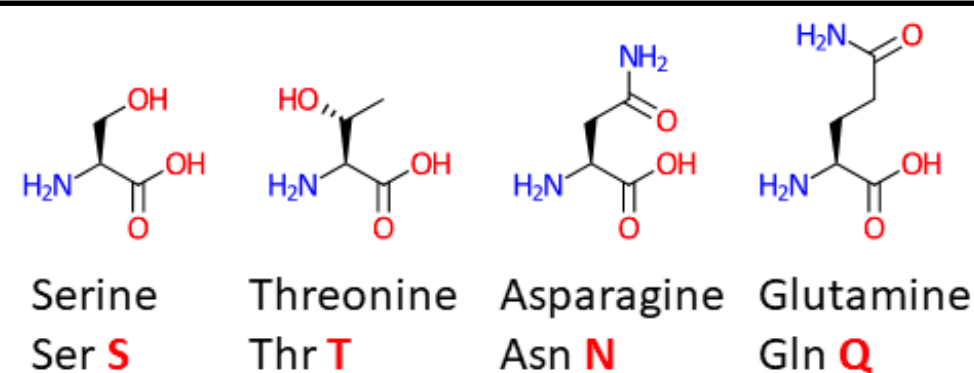
Protein Organization

Levels of protein organization

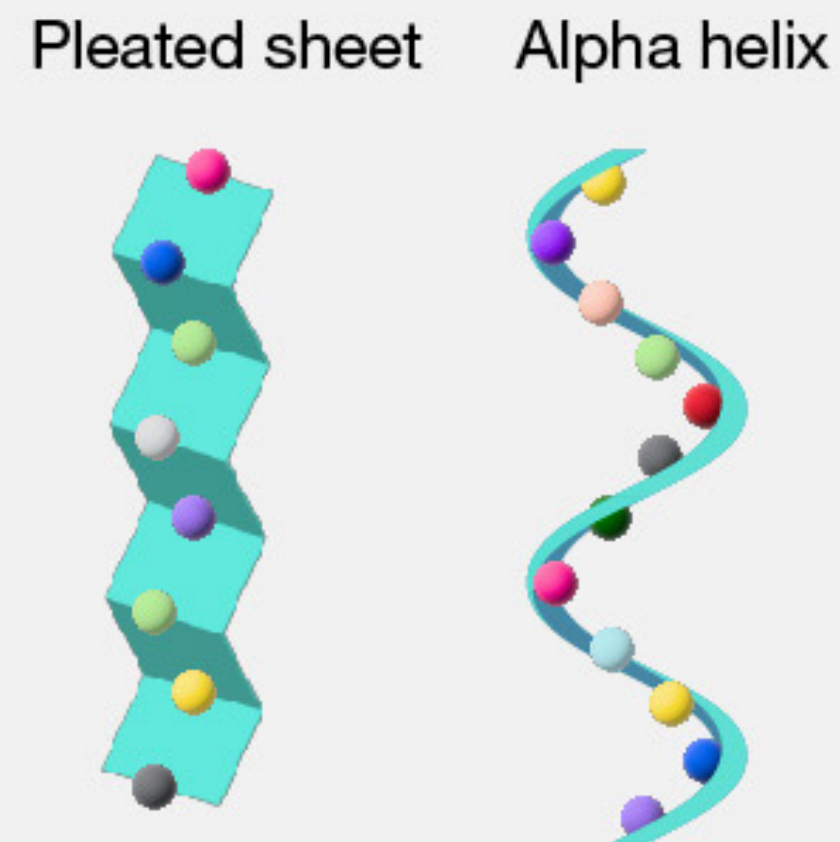
Primary protein structure
is the sequence of a chain of amino acids.



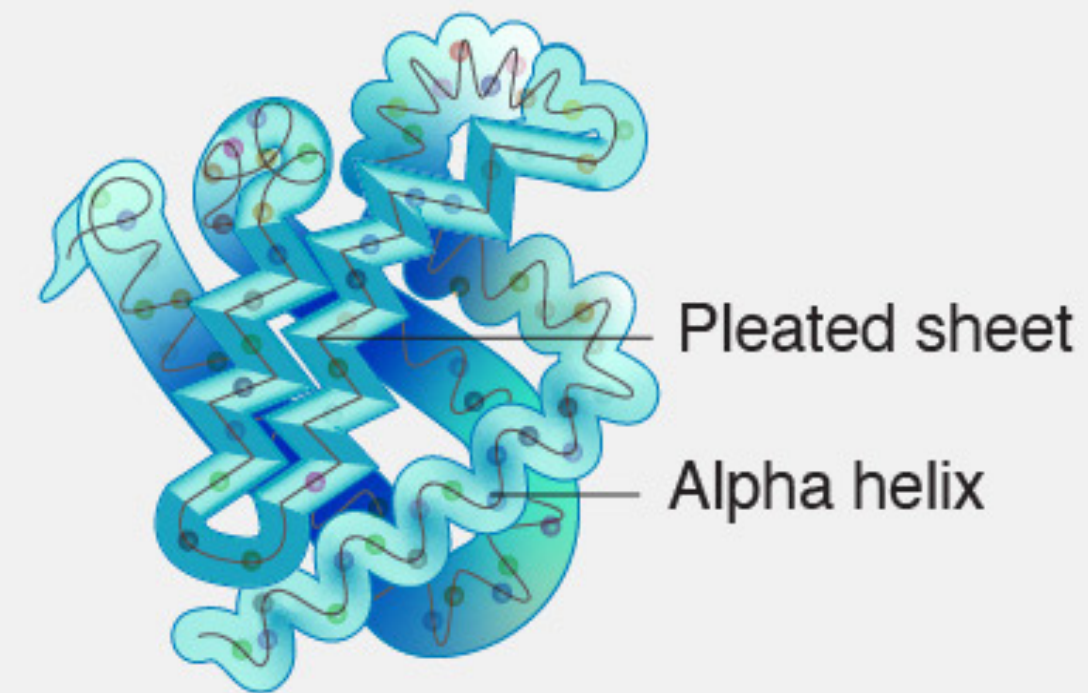
MSTAGKVIKCKAA...



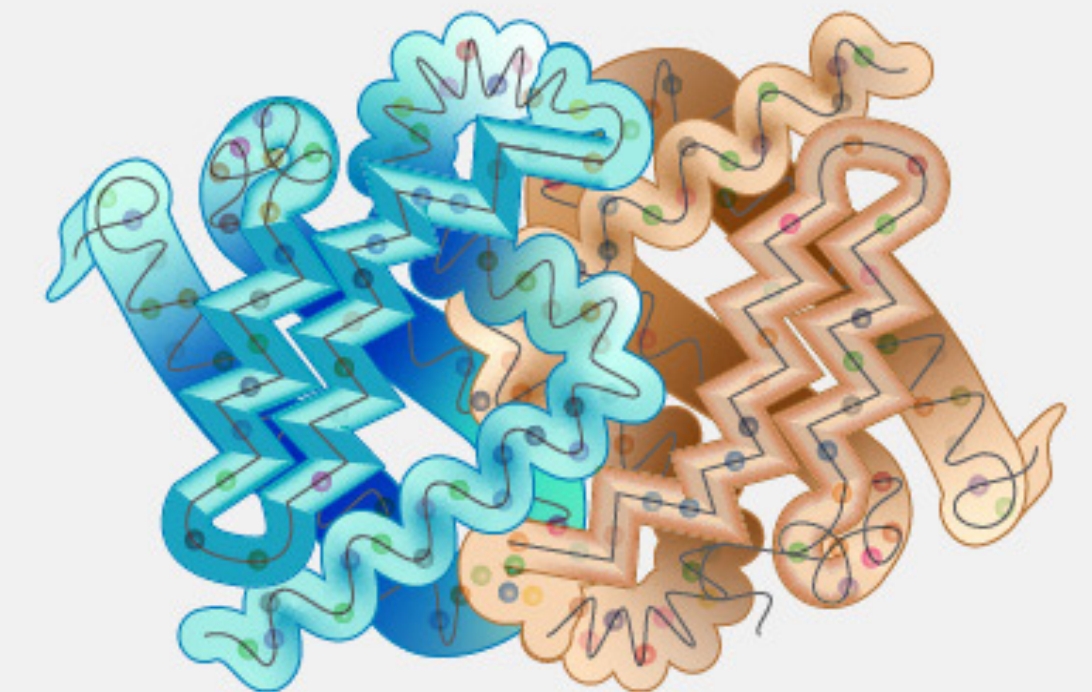
Secondary protein structure
occurs when the sequence of amino acids folds into a three-dimensional shape.



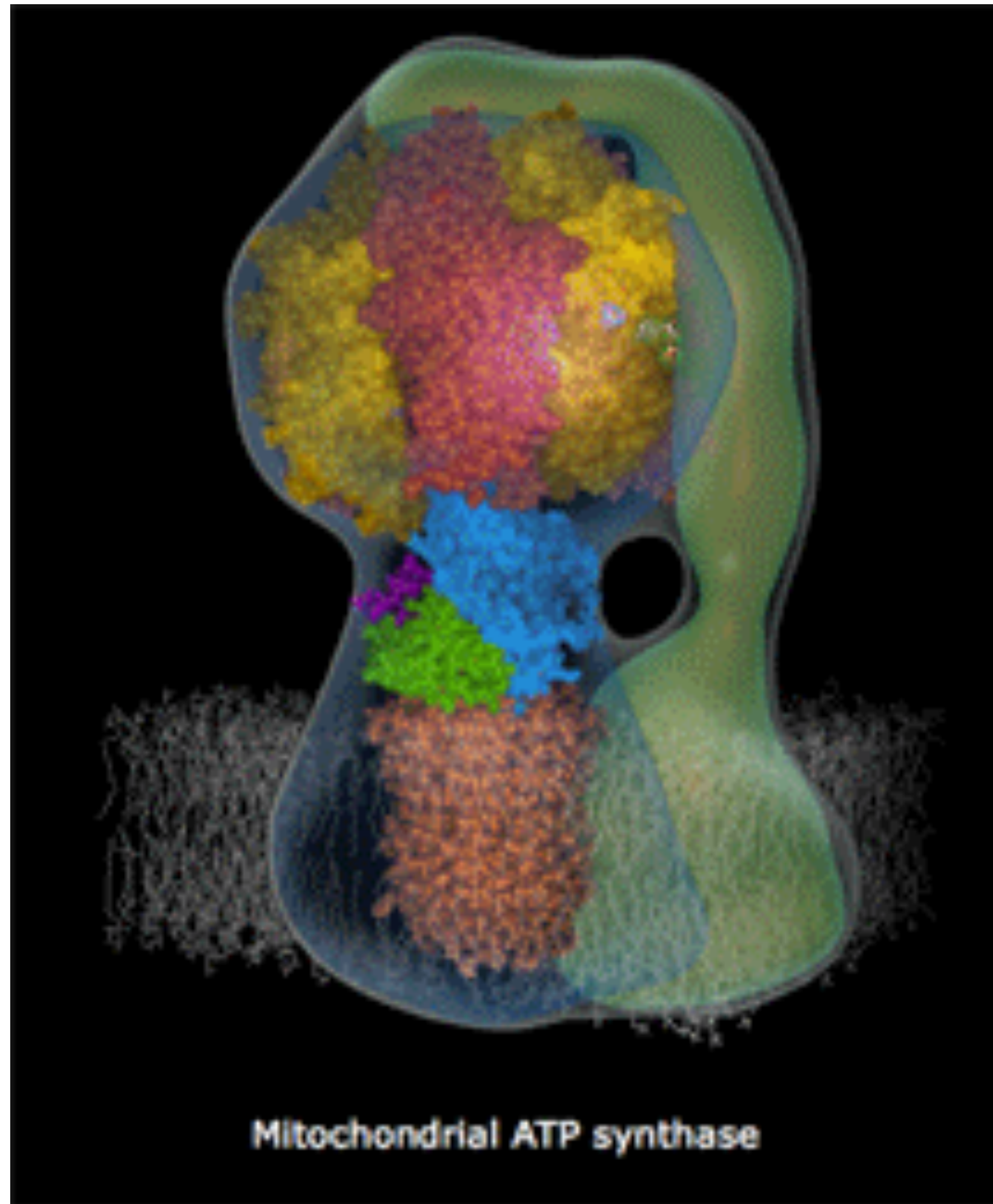
Tertiary protein structure
occurs when a mature protein
folds upon itself.



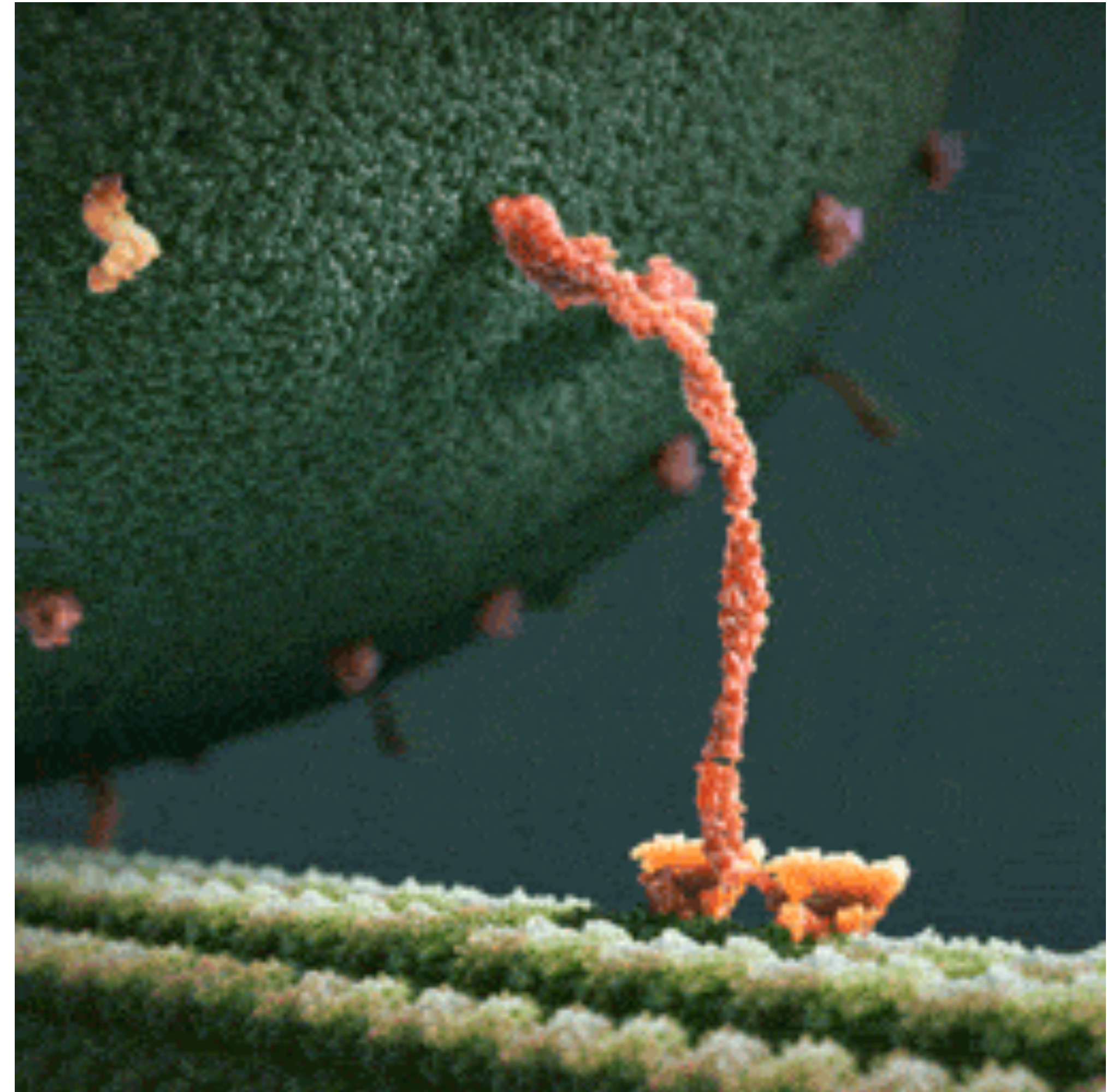
Quaternary protein structure
is a protein consisting of more than one polypeptide chain.



Structure Informs Function



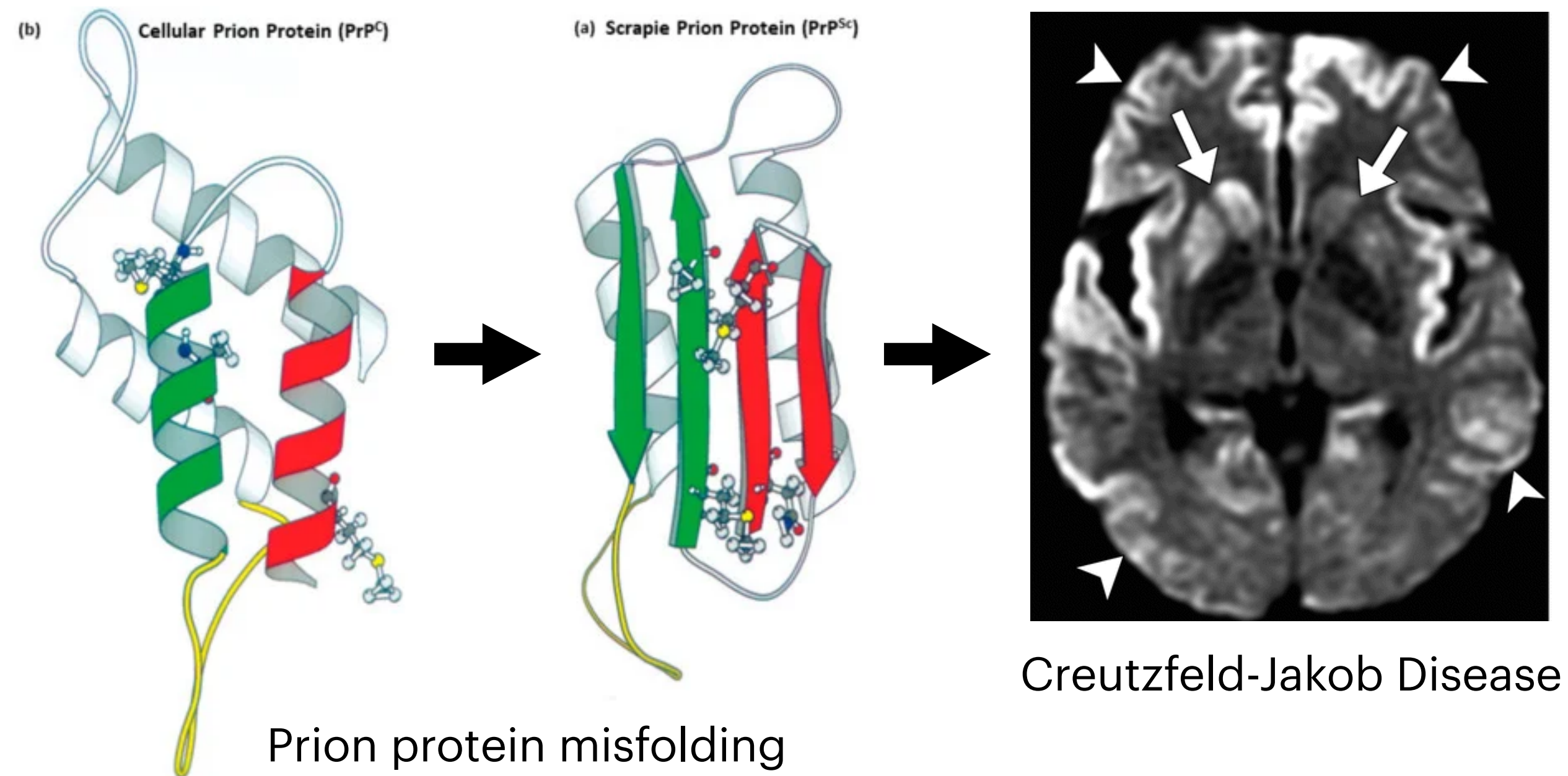
ATP Synthase



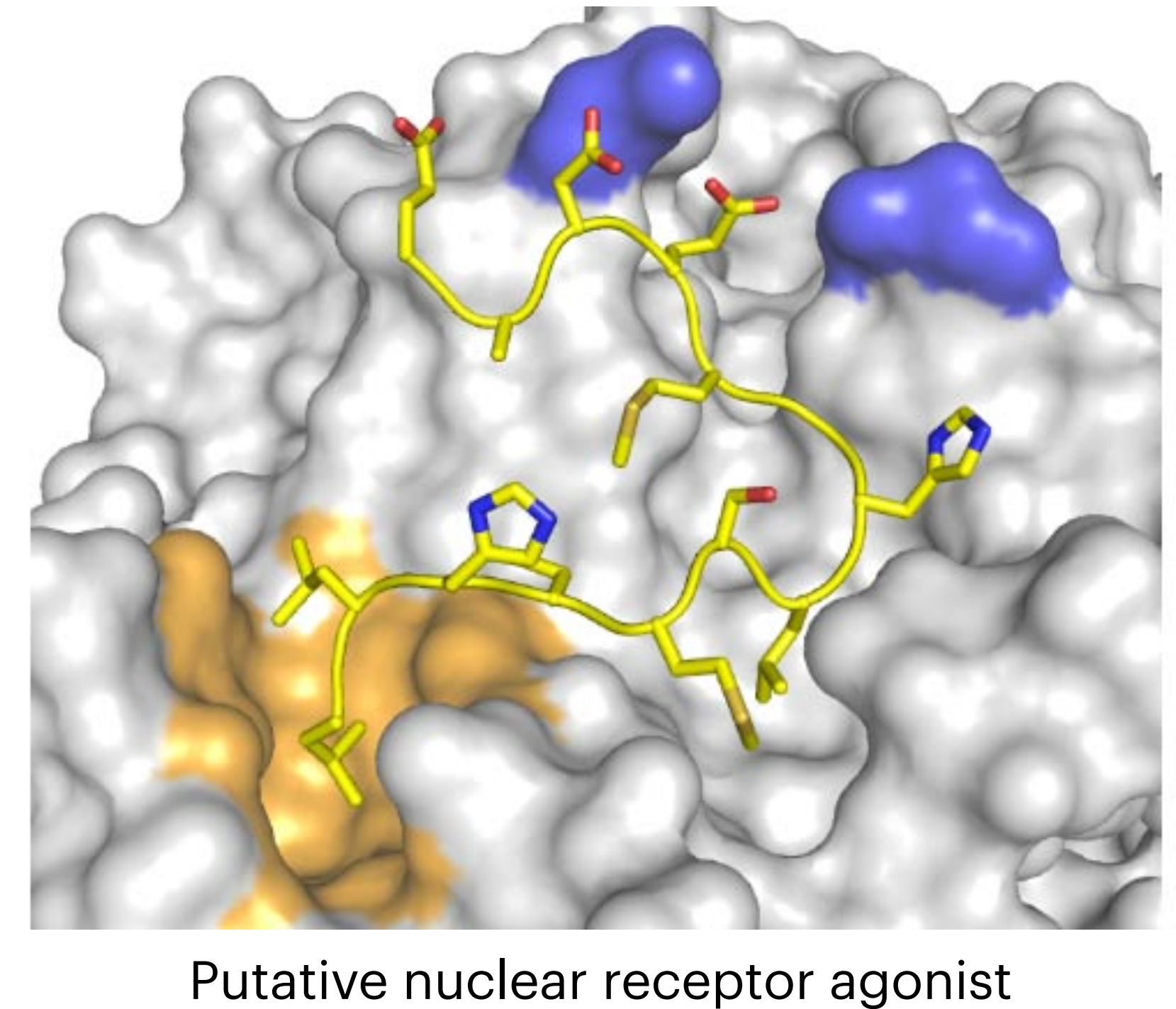
Kinesin

Structure Informs (Dys)Function

Mechanistic Disease Understanding

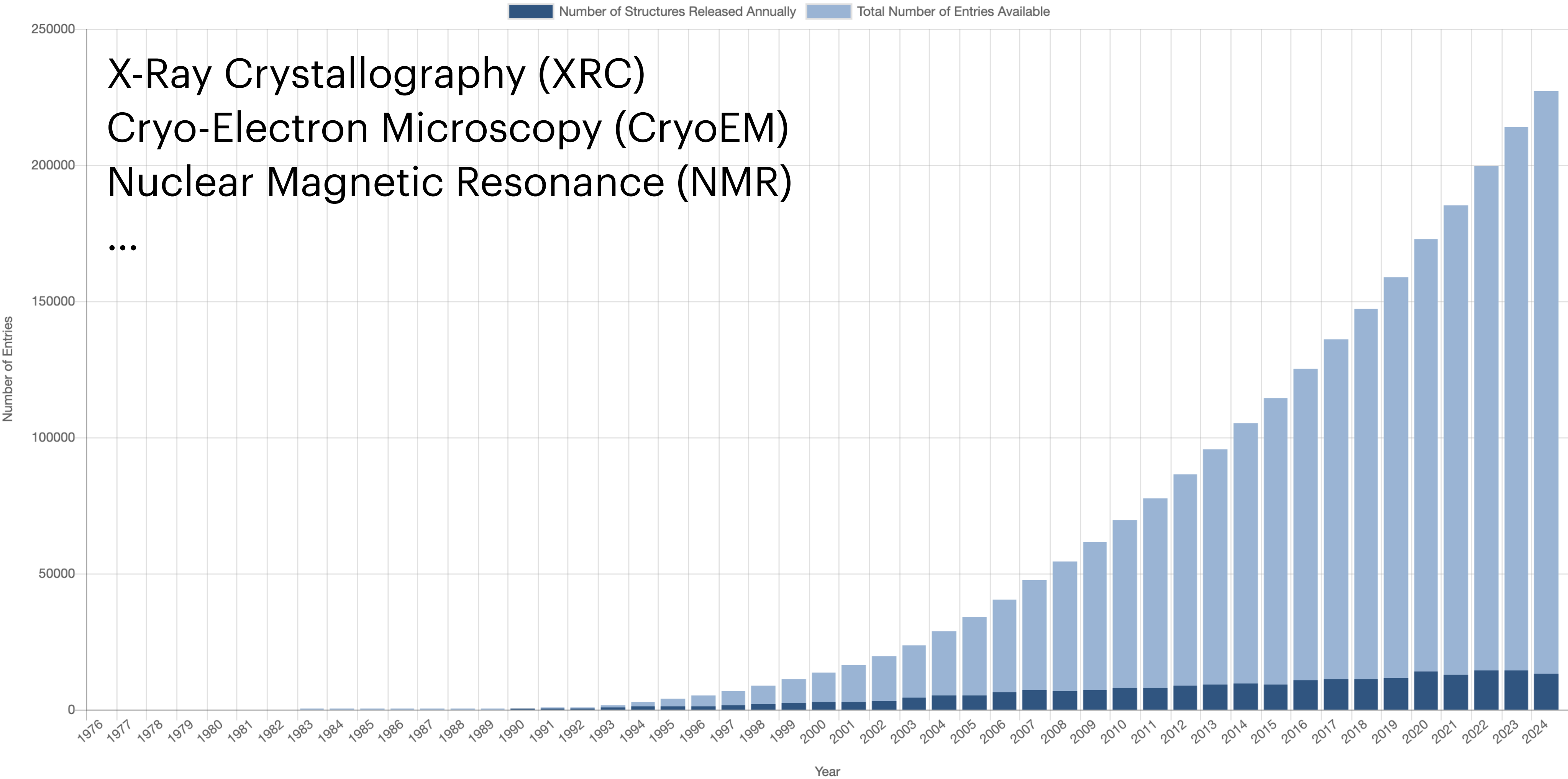


Structure-Based Drug Design



The Protein Data Bank

227,344 experimentally determined structures



“The Protein-Folding Problem, 50 Years on”

Dill & MacCallum, Science 2012

Three broad questions:

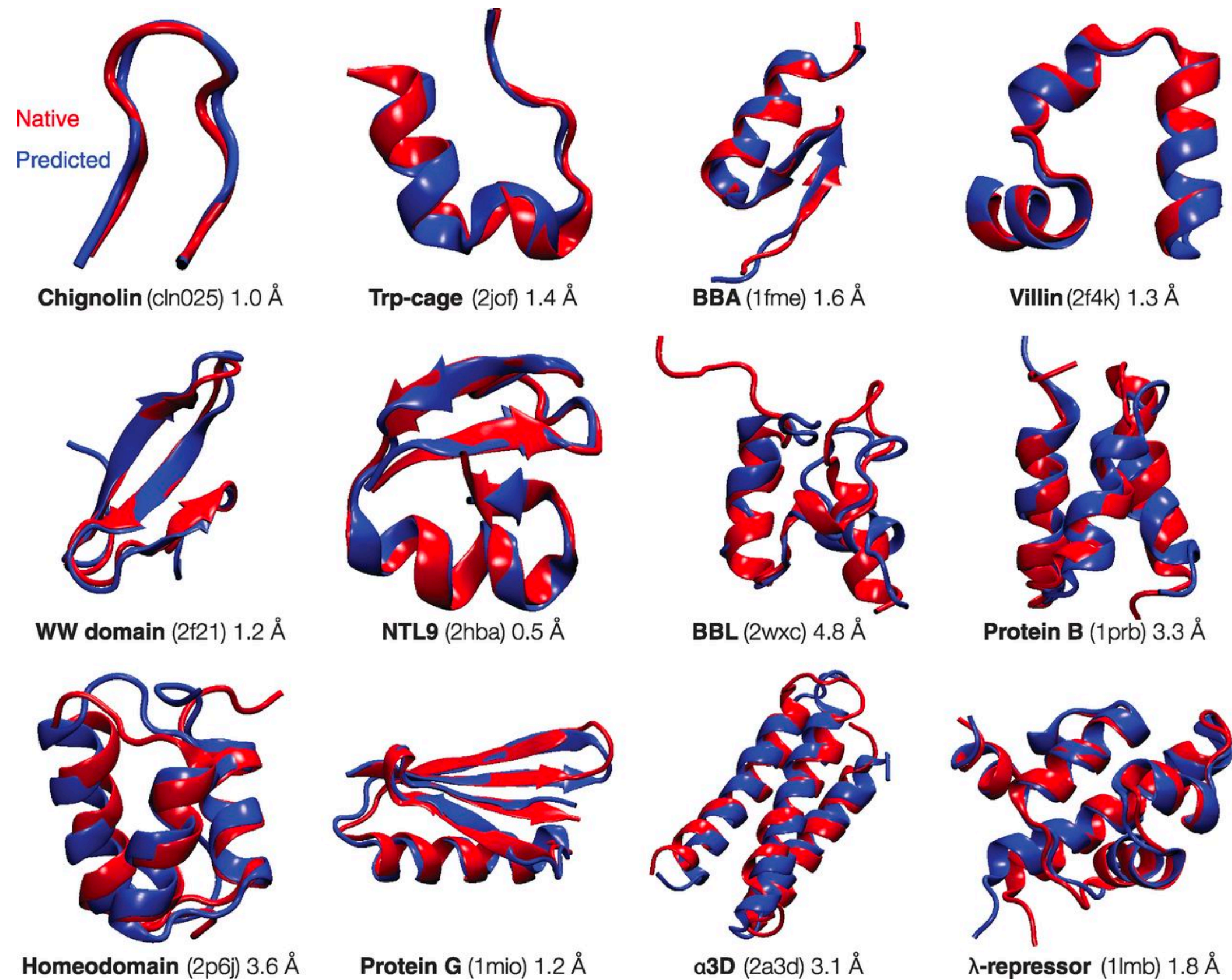
1. What is the physical code by which an amino acid sequence dictates a protein's native structure? (Thermodynamics)
2. How can proteins fold so fast? (Kinetics)
3. Can we devise a computer algorithm to predict protein structures from their sequences?

“Perhaps the most remarkable features of the molecule are its complexity and its lack of symmetry. The arrangement seems to be almost totally lacking in the kind of regularities which one instinctively anticipates, and it is more complicated than has been predicated by any theory of protein structure.”

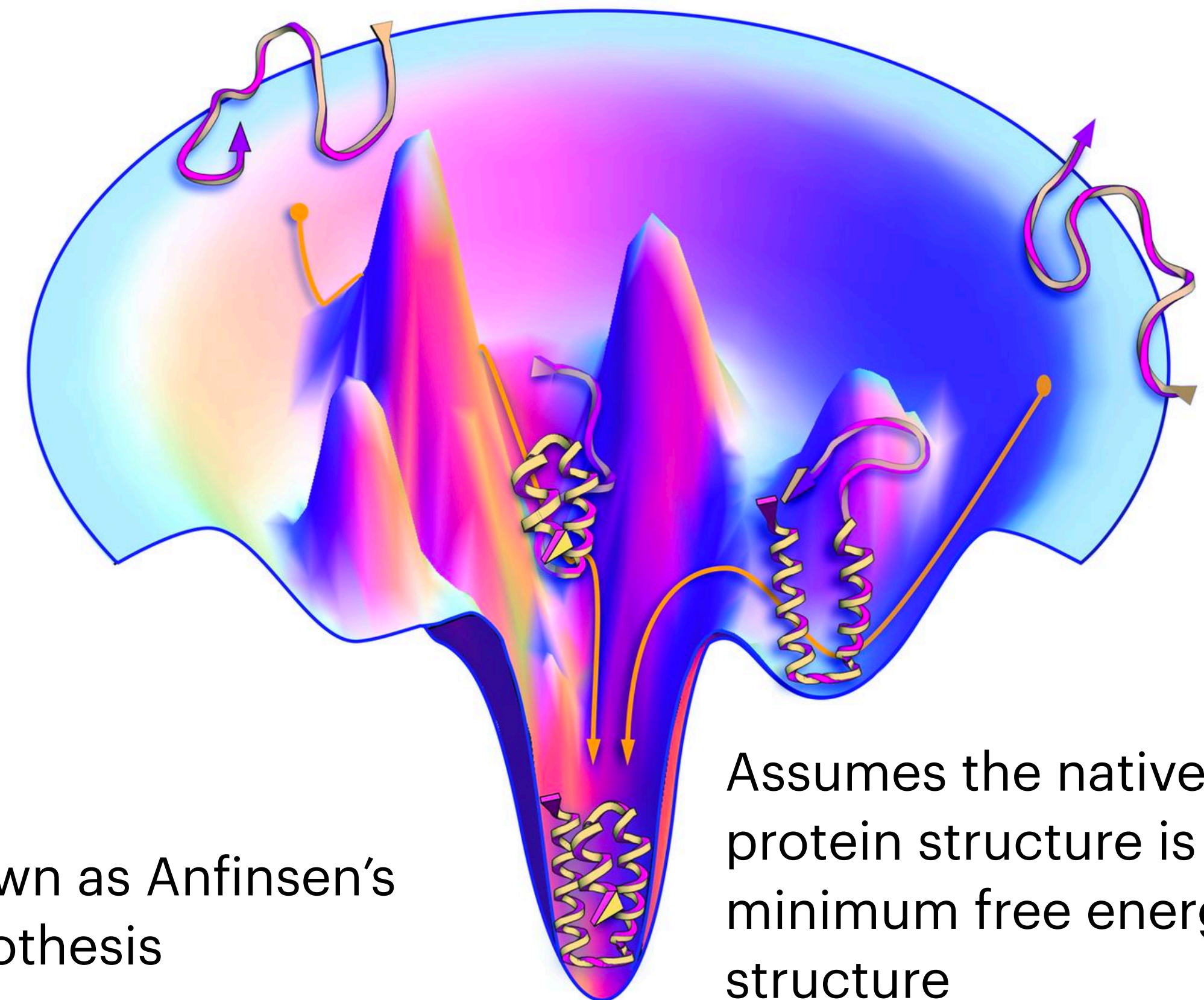


Fig 1: In 1958, Kendrew and coworkers published the first structure of a globular protein, myoglobin at 6Å resolution

Protein folding is driven by physics



Energy landscape theory of protein folding



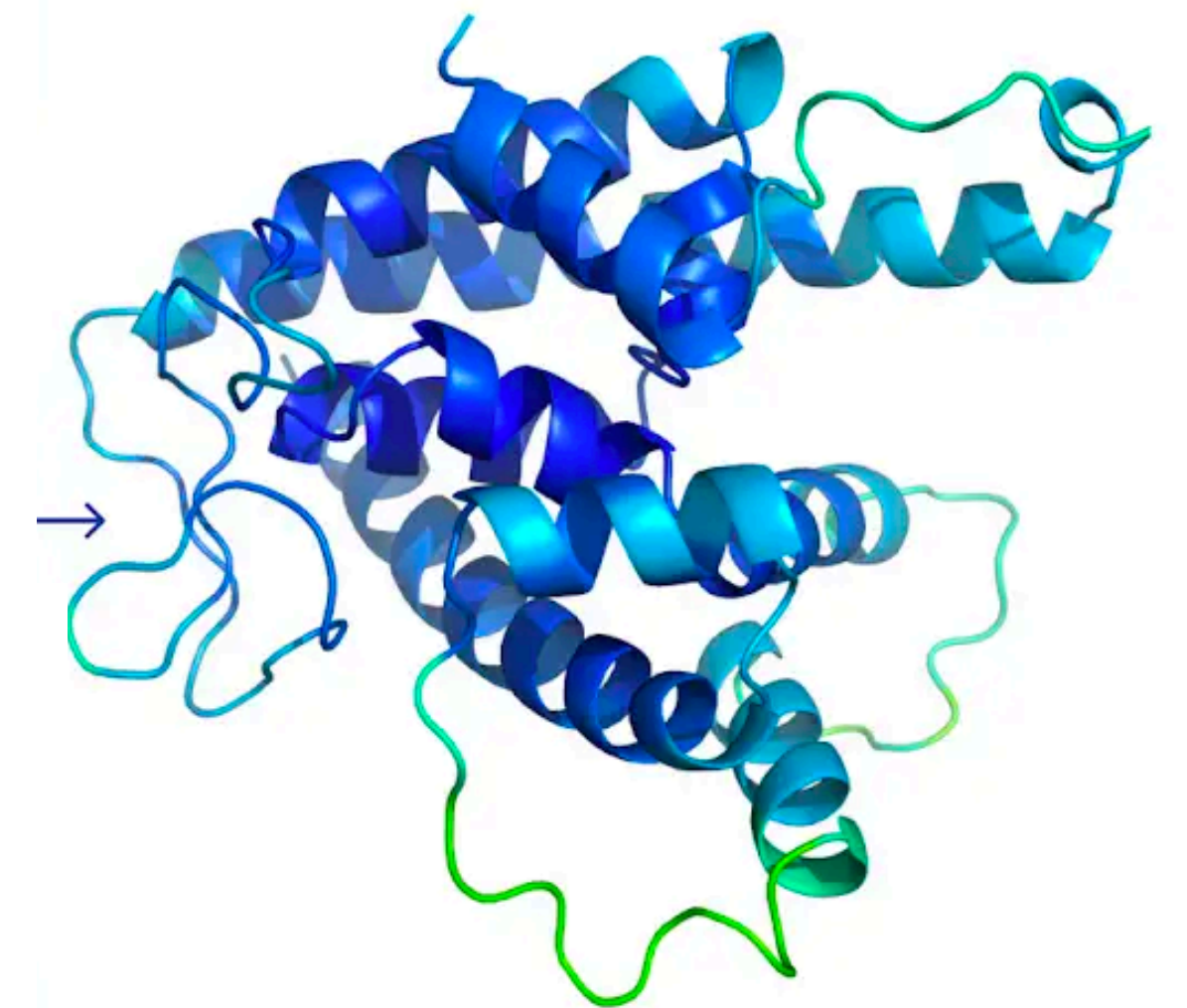
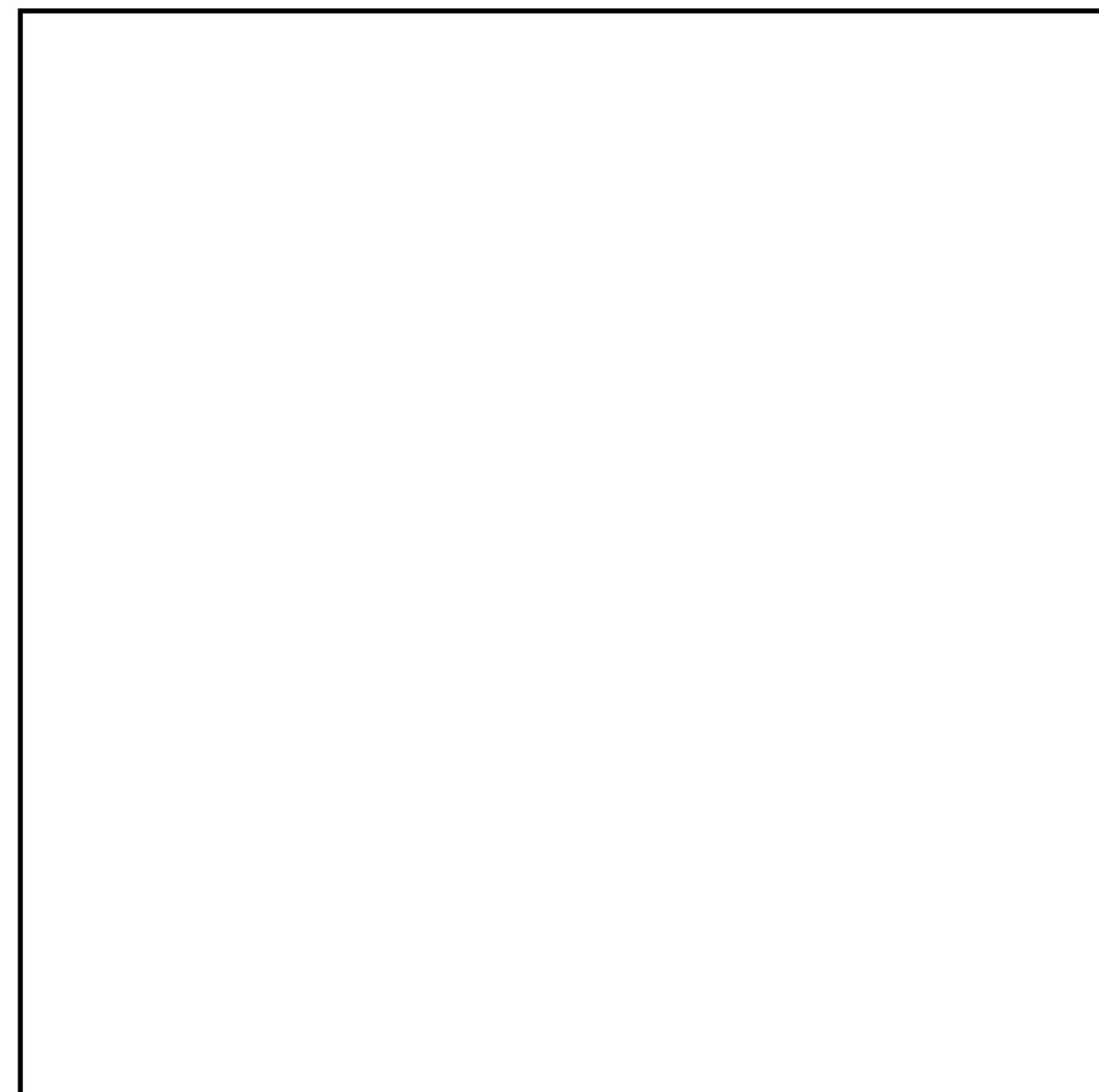
Known as Anfinsen's hypothesis

Protein structure prediction

Task: Predict the 3D coordinates of all atoms of a given input protein sequence

- Experimental structure determination can be challenging and time-consuming
- Large gap between known sequences and structures
 - As of 2021, only 35% of proteins and 17% of residues in the human proteome had experimental structural information

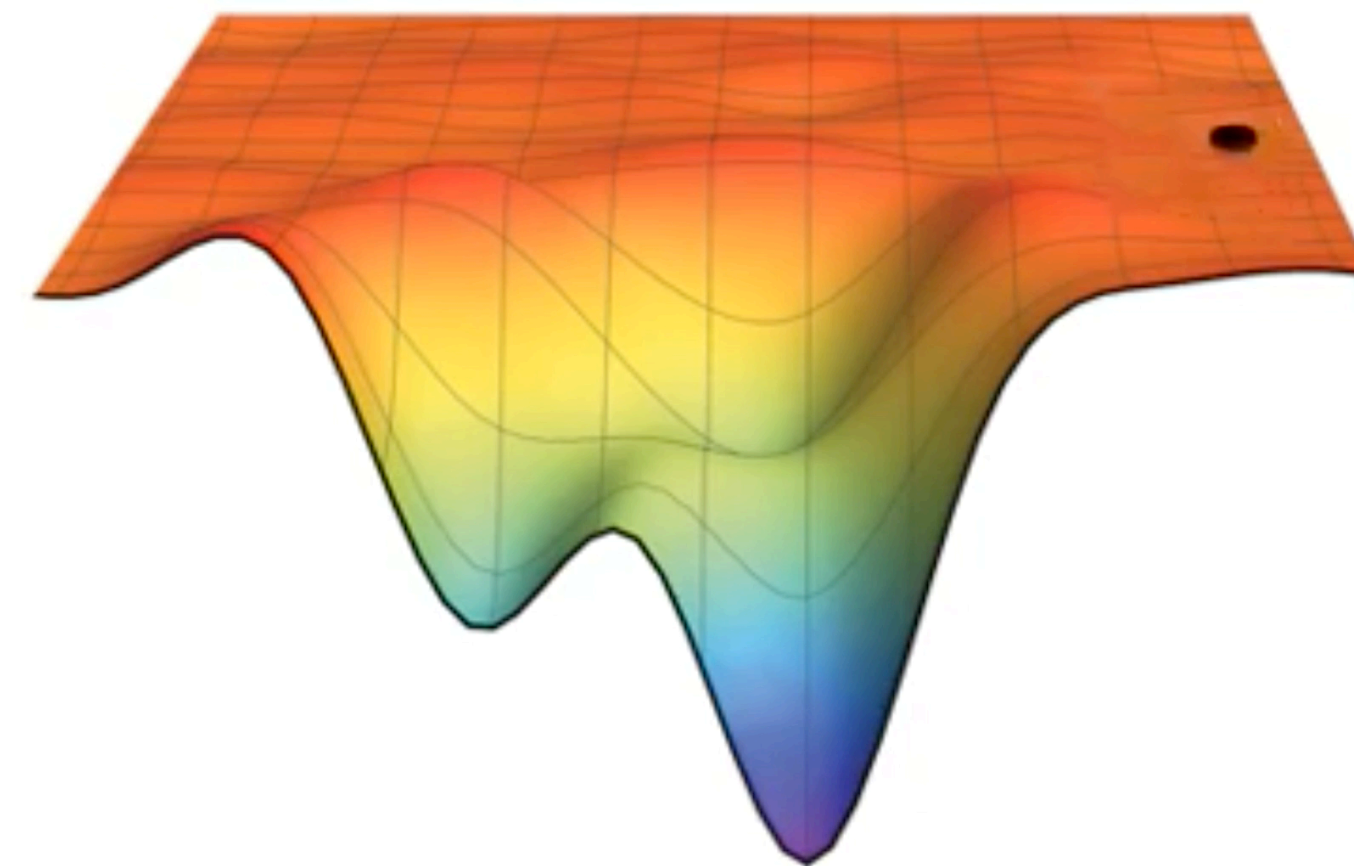
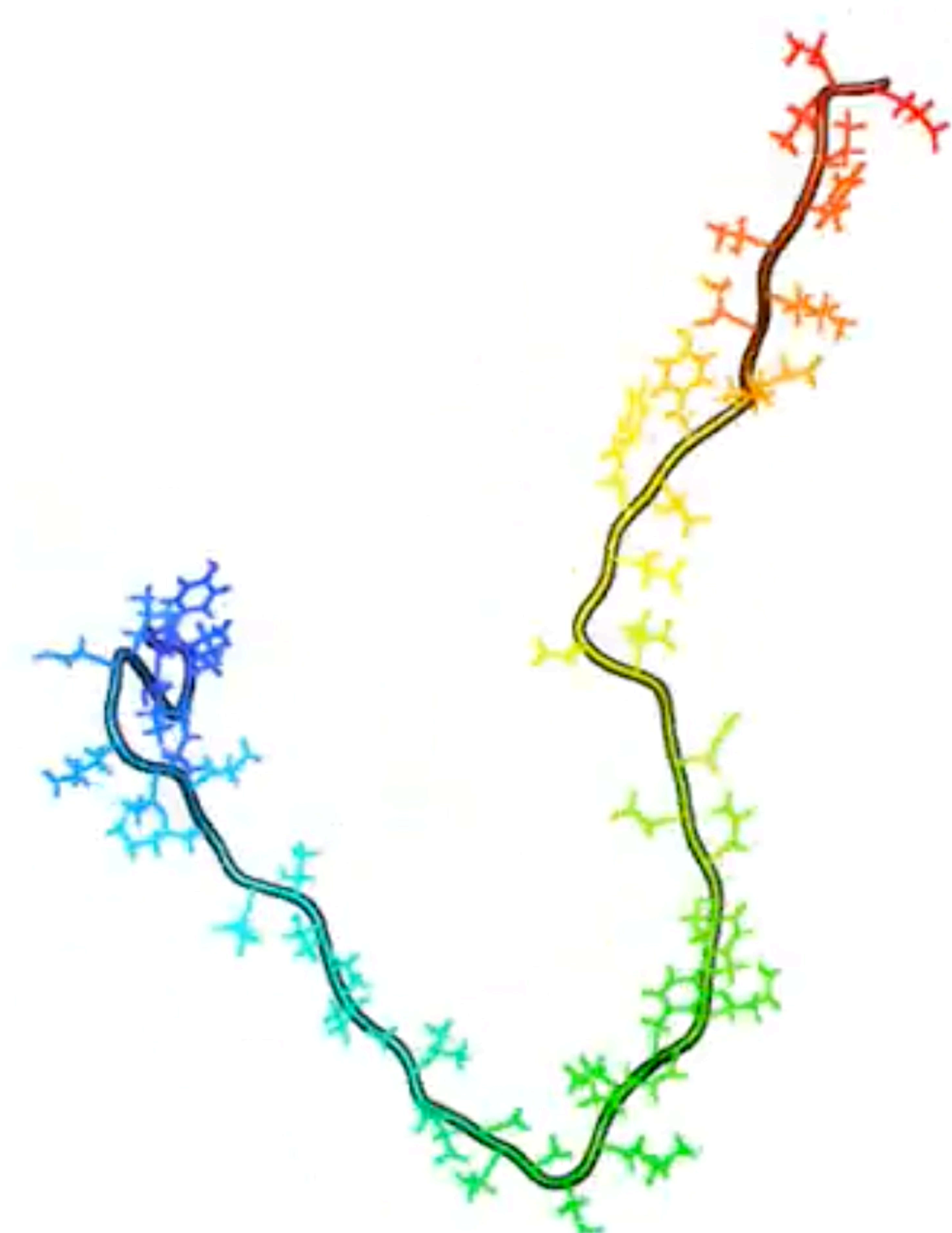
Input sequence
MRKPRTPTFTT...



Protein structure prediction

How would you do this in 2012?

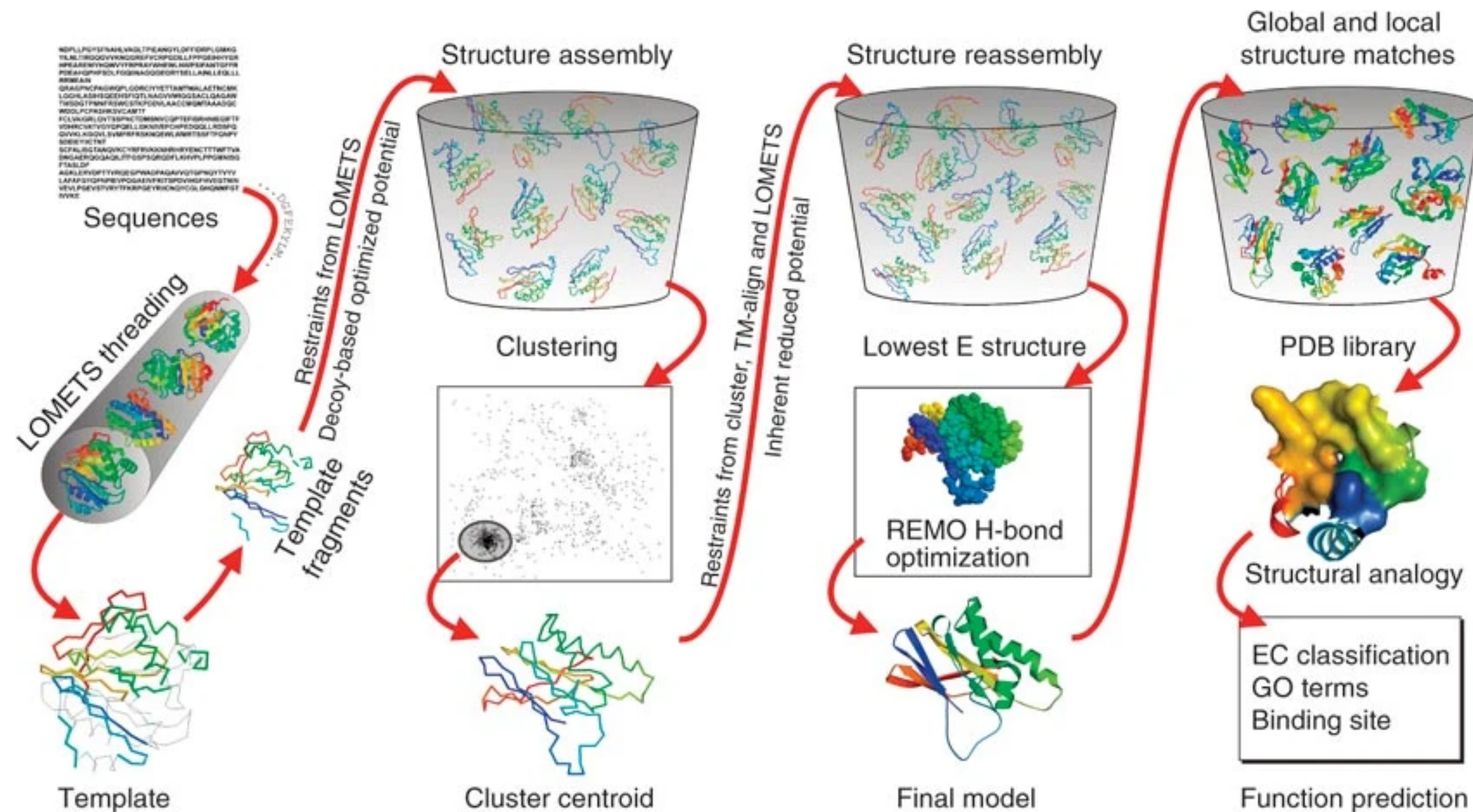
- Molecular dynamics (MD) simulation



Protein structure prediction

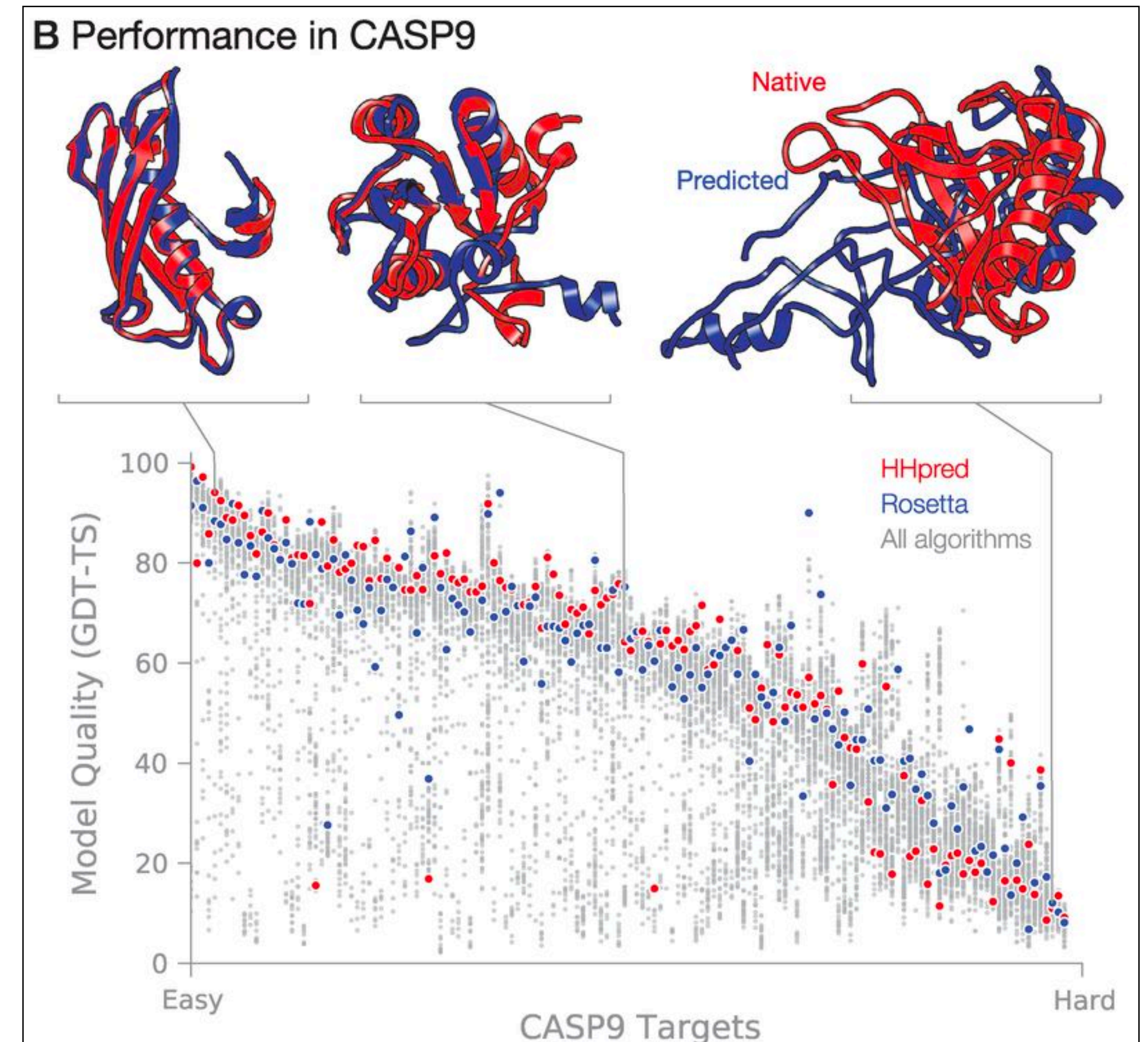
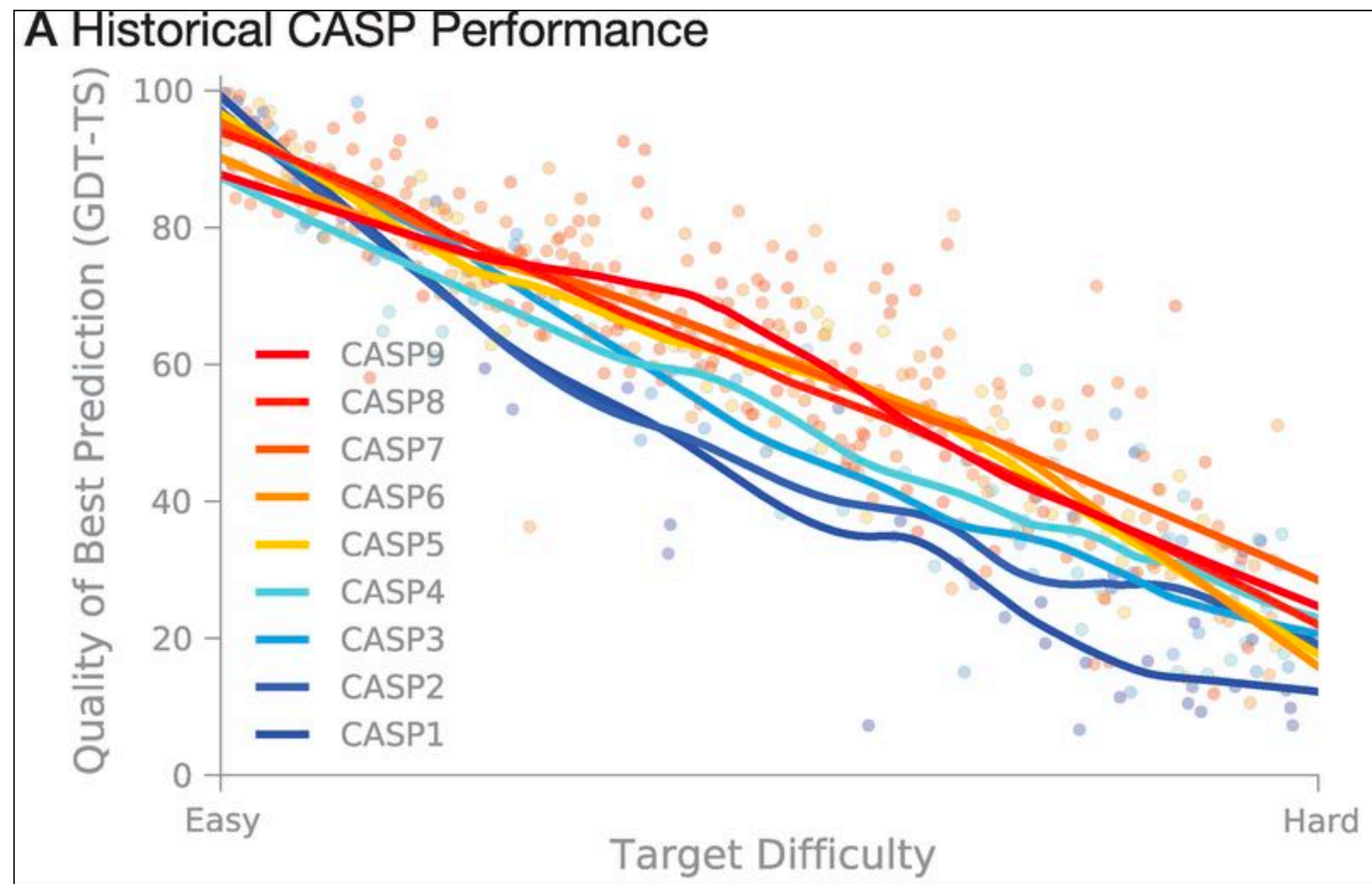
How would you do this in 2012?

- Template-based fragment assembly



CASP: Critical Assessment of Protein Structure Prediction

- Biennial, community-wide blind competition initiated by John Moult and colleagues in 1994
- ~100 target sequences, ~150 research groups
- Release ~1 protein per day, May-August. 3 weeks to return 5 predictions
- What makes a target sequence “easy” vs “hard”?



Evaluation Metrics

- Root Mean Squared Distance (RMSD) ↓
 - Average distance between corresponding residues in experimental and predicted structures
 - Sensitive to large local errors
- Template Modeling Score (TM) ↑
 - Reweights distances to be more sensitive to global structural similarity
- Global Distance Test (GDT) ↑
 - Fraction of residues in predicted structure within a defined distance cutoff of positions in experimental structure
- Local Distance Difference Test (LDDT) ↑
 - Fraction of distances in experimental structure that are preserved (within a cutoff) in the predicted structure
 - Does not require alignment of experimental and predicted structures (superposition-free)

$$\sqrt{\sum_{i=1}^L d_i^2 / L}$$

RMSD

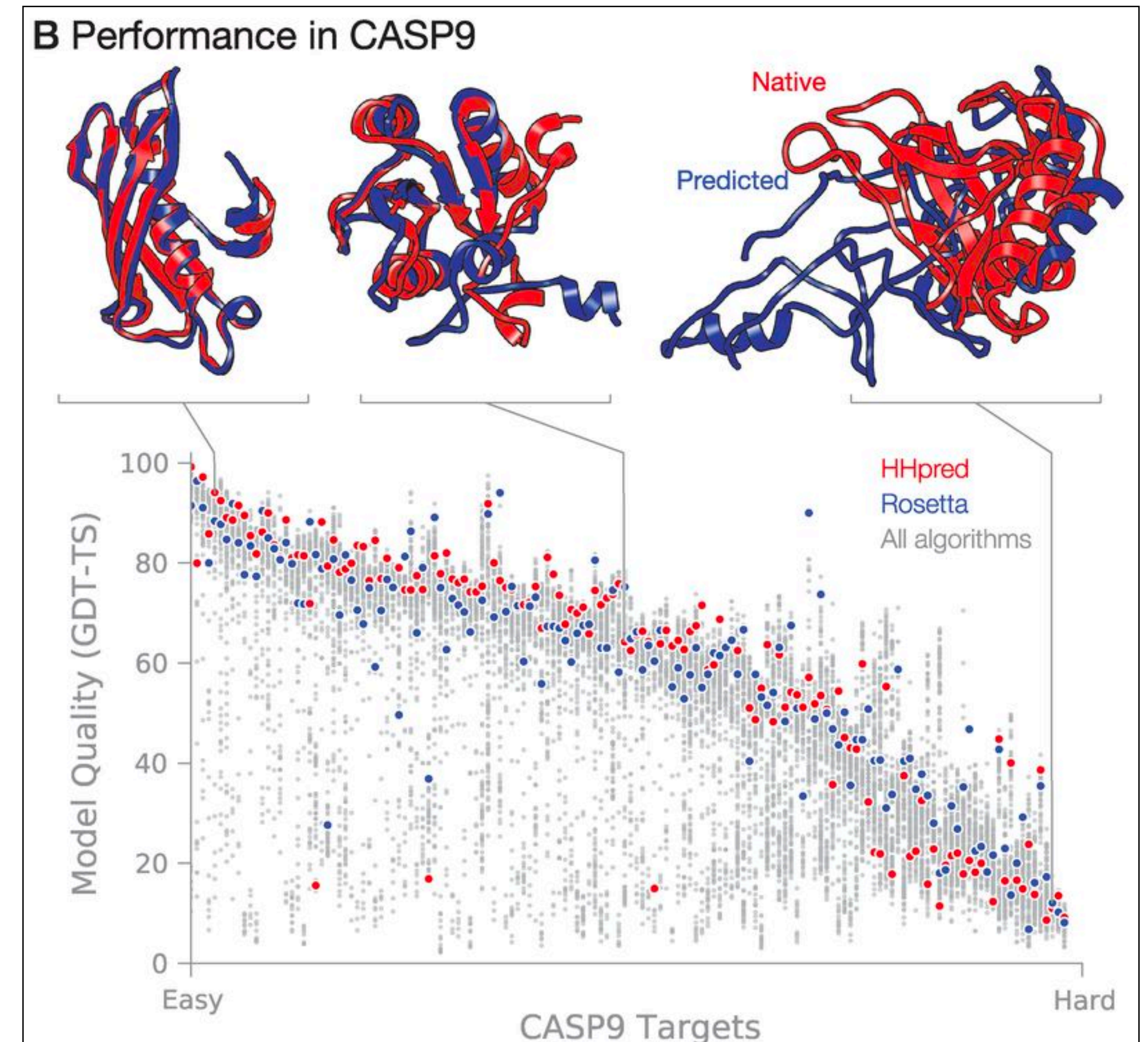
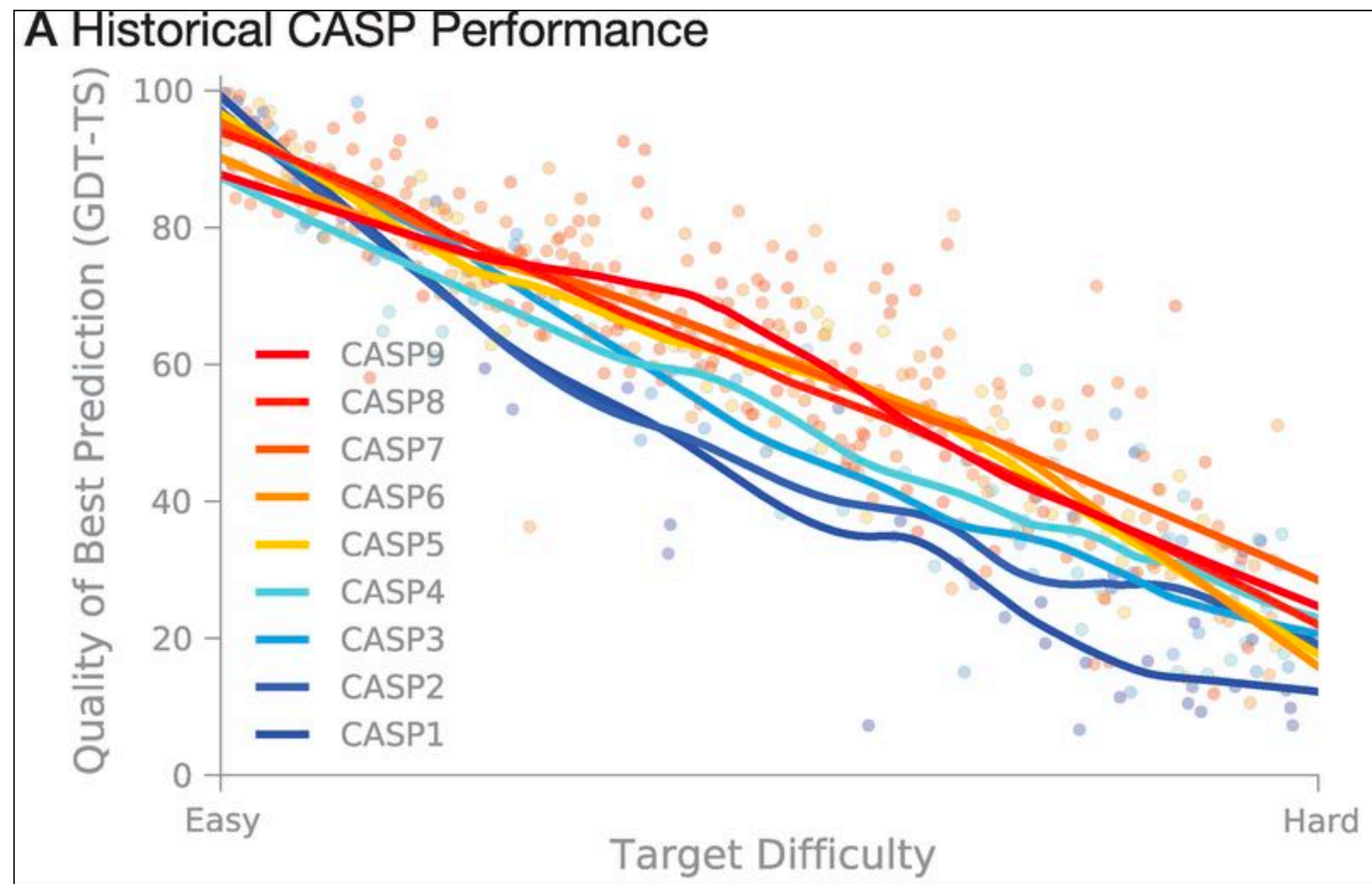
$$\max \left[\frac{1}{L_{\text{target}}} \sum_i^{L_{\text{common}}} \frac{1}{1 + \left(\frac{d_i}{d_0(L_{\text{target}})} \right)^2} \right]$$

TM-Score



CASP: Critical Assessment of Protein Structure Prediction

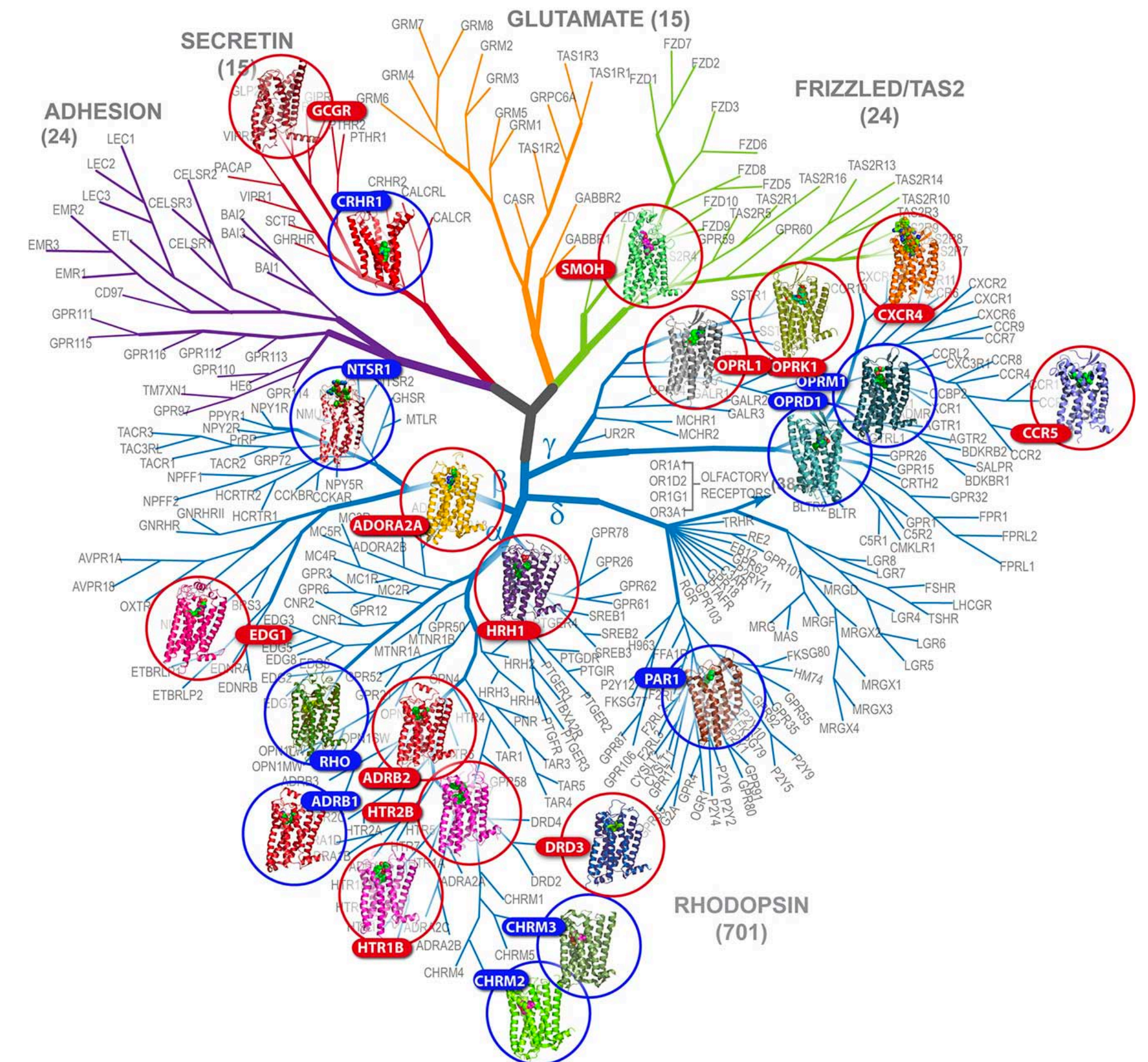
- Biennial, community-wide blind competition initiated by John Moult and colleagues in 1994
- ~100 target sequences, ~150 research groups
- Release ~1 protein per day, May-August. 3 weeks to return 5 predictions
- What makes a target sequence “easy” vs “hard”?



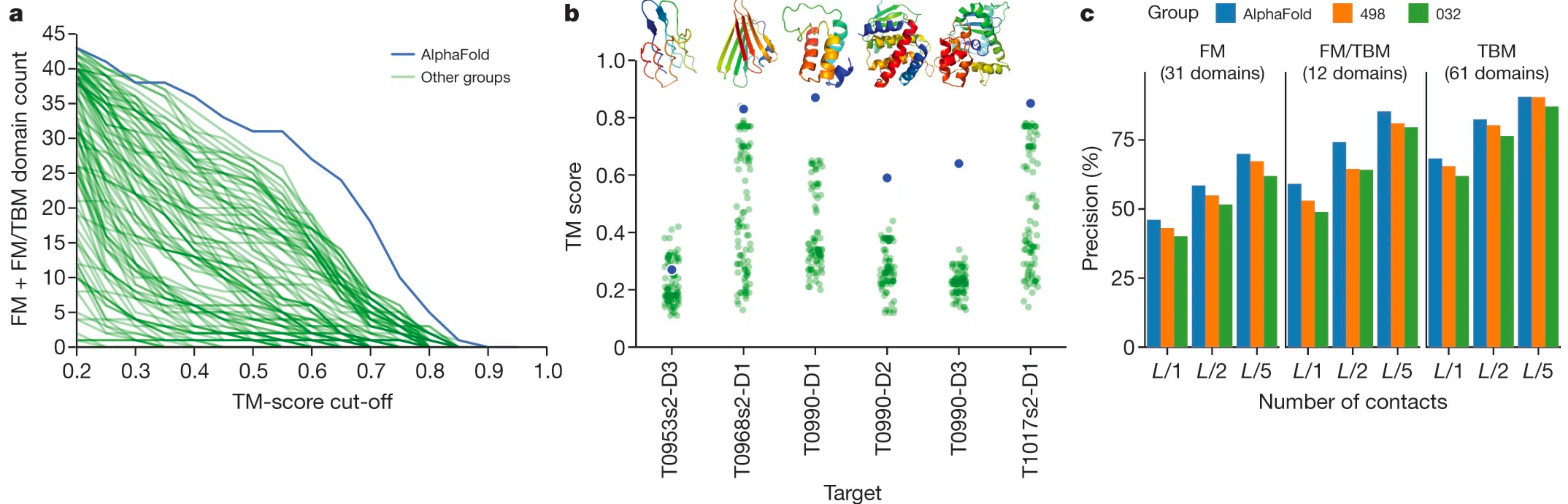
Template-based modeling (TBM) vs. Free modeling (FM)

- Template-based modeling (TBM)
 - Availability of a *homologous* structure (i.e. the template)
- Free modeling (FM)
 - No similar sequence with a known structure
 - Methods typically rely on fragment assembly approaches and/or predict residue-residue correlations from sequence alone
- What data or information do methods use?
 - Assume that similar sequences lead to similar structures.
 - 2.4B protein sequences
 - In 2012: 80,000 structures in the PDB (4000 structural families, 1200 folds)

A *homologous* gene (or homolog) is a gene inherited in two species from a common ancestor.

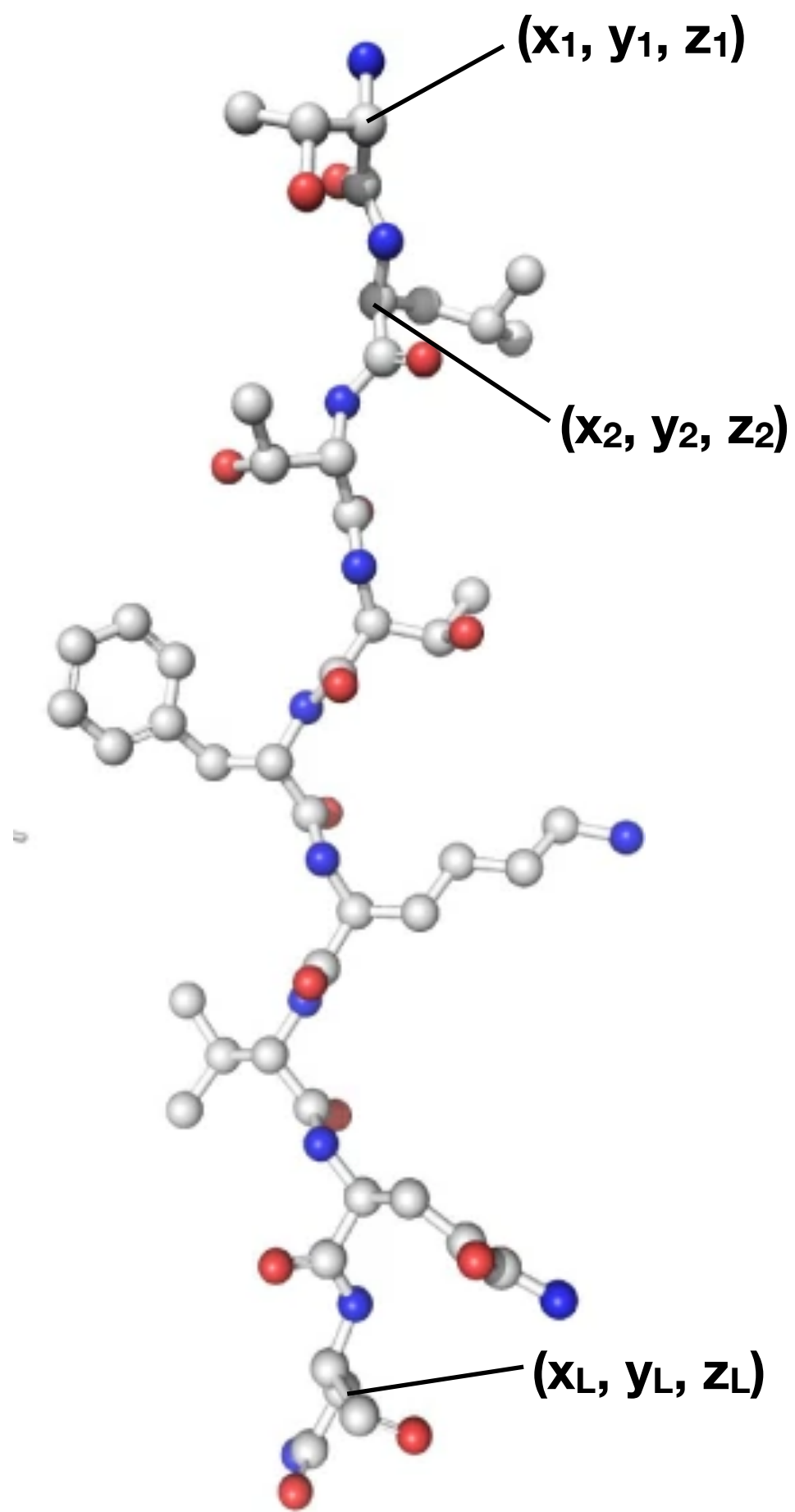


AlphaFold1 at CASP13

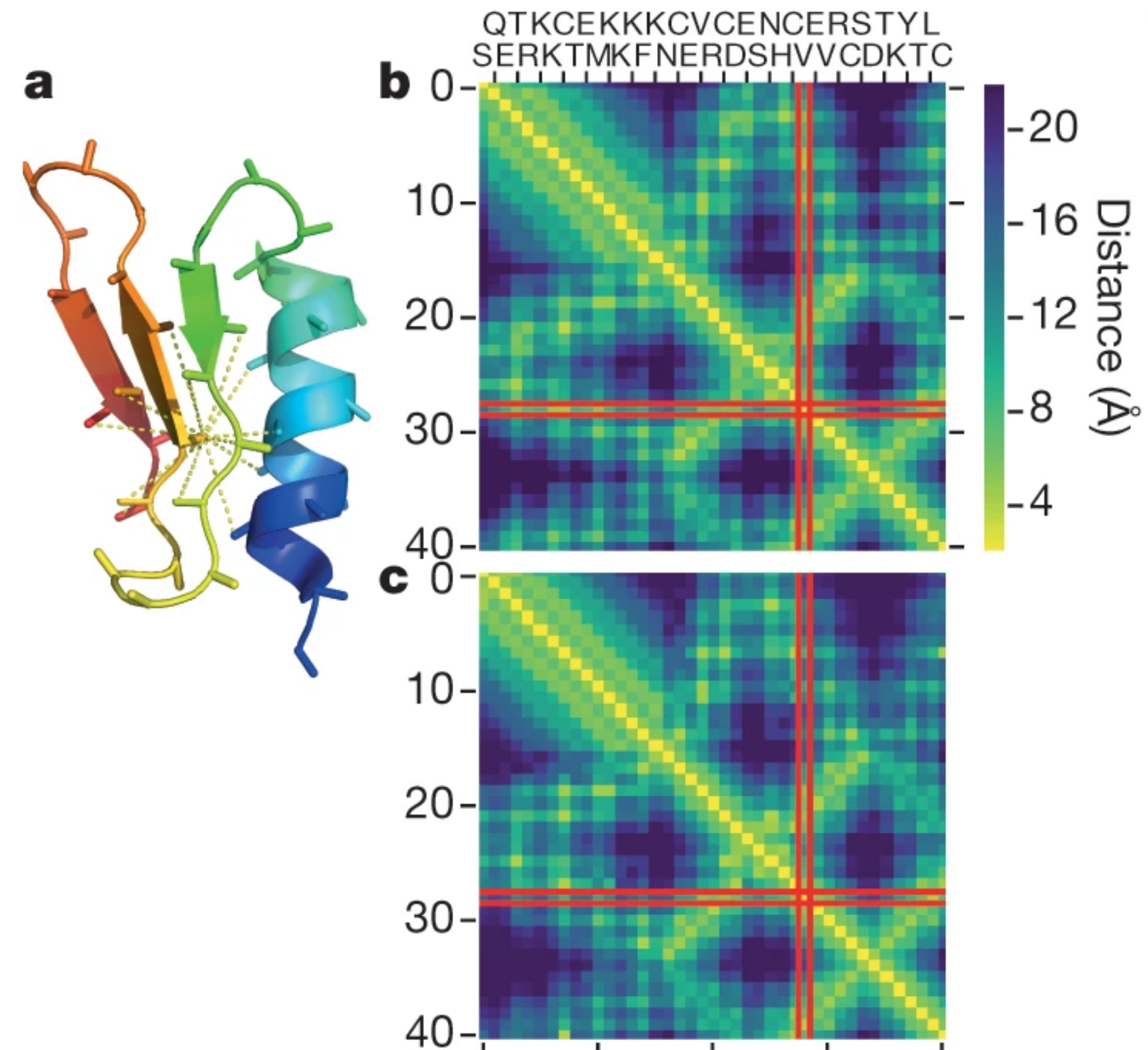


a, Number of FM (FM + FM/TBM) domains predicted for a given TM-score threshold for AlphaFold and the other 97 groups. **b**, For the six new folds identified by the CASP13 assessors, the TM score of AlphaFold was compared with the other groups, together with the native structures. The structure of T1017s2-D1 is not available for publication. **c**, Precisions for long-range contact prediction in CASP13 for the most probable L , $L/2$ or $L/5$ contacts, where L is the length of the domain. The distance distributions used by AlphaFold in CASP13, thresholded to contact predictions, are compared with the submissions by the two best-ranked contact prediction methods in CASP13: 498 (RaptorX-Contact²⁶) and 032 (TripletRes³²) on 'all groups' targets, with updated domain definitions for T0953s2.

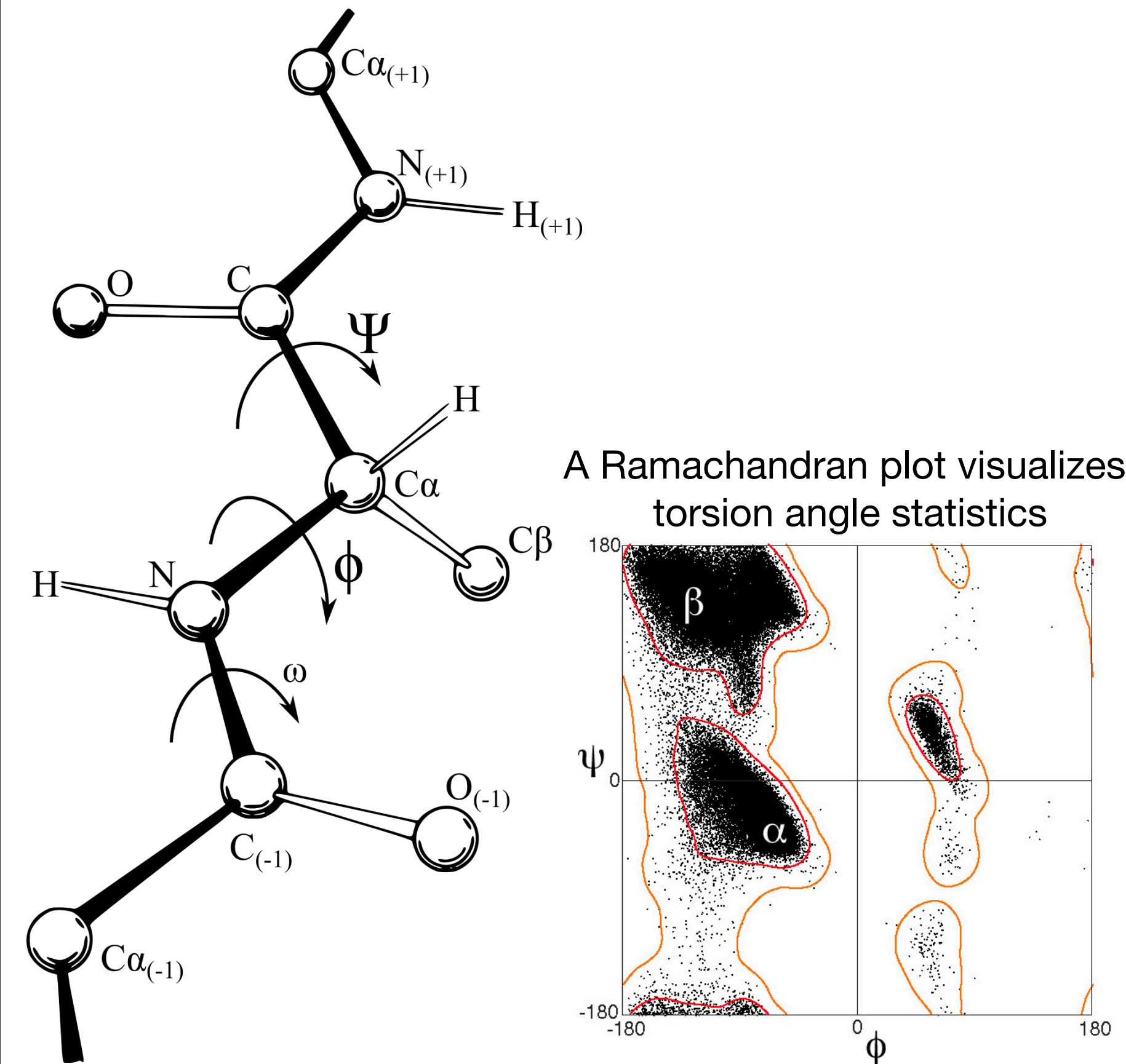
Geometric descriptors for protein backbone



3D coordinates



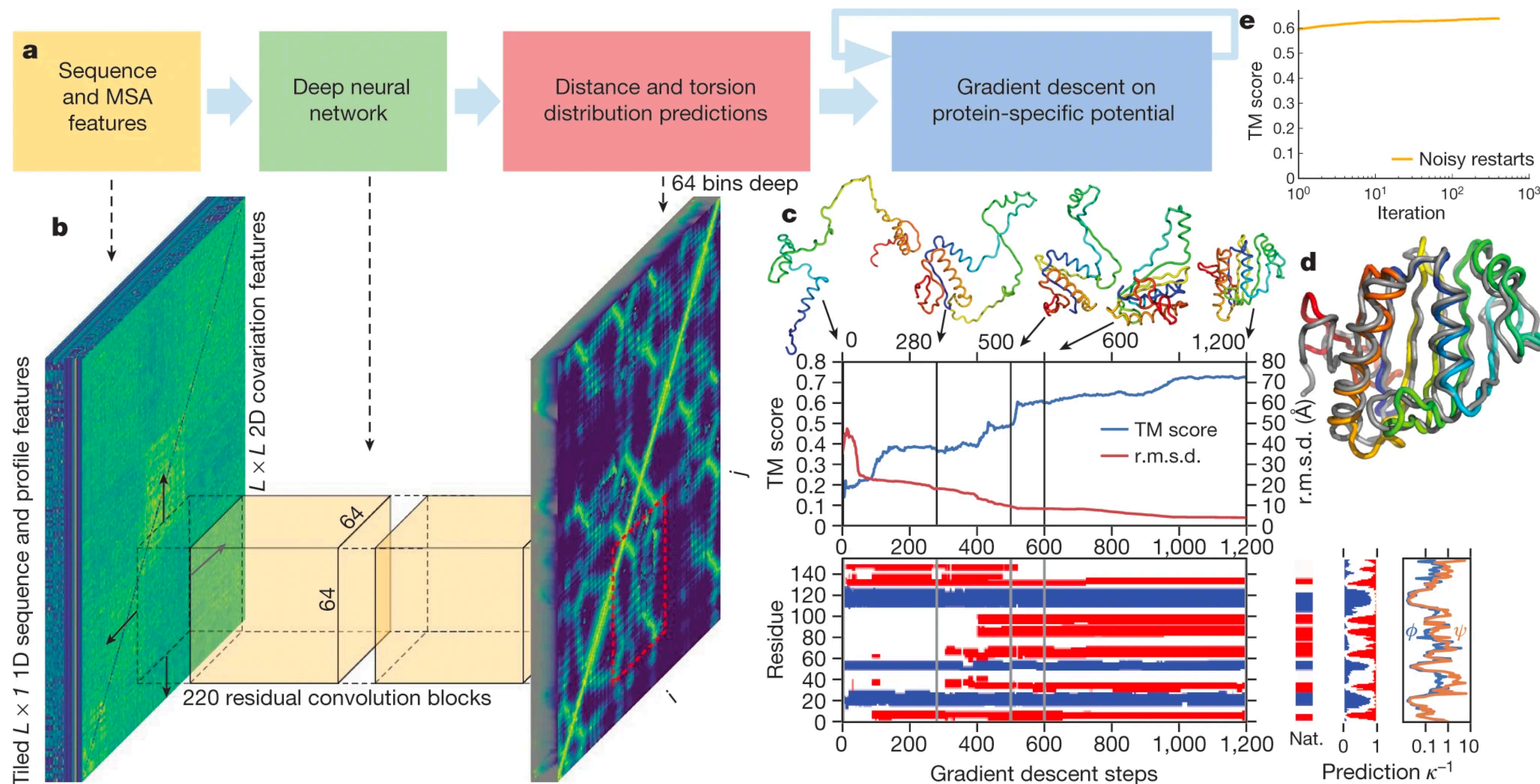
Pairwise residue distances



Backbone torsion angles

How does AlphaFold1 work?

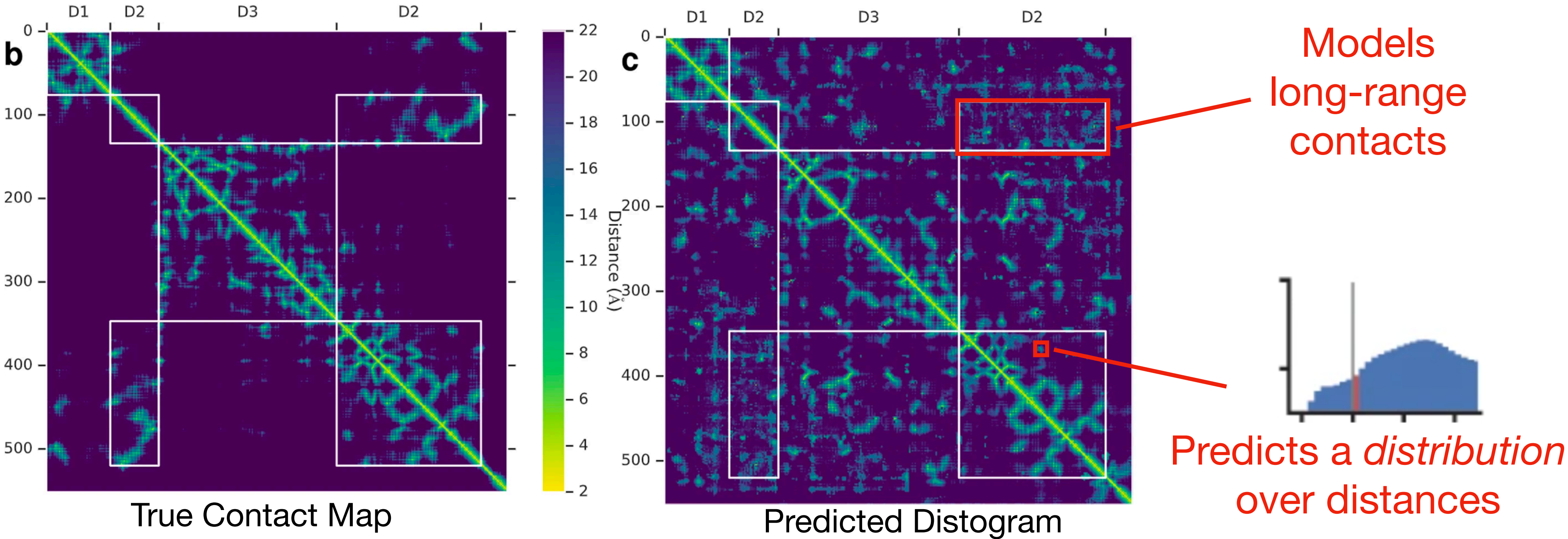
Fig. 2: The folding process illustrated for CASP13 target T0986s2.



What is a distogram?

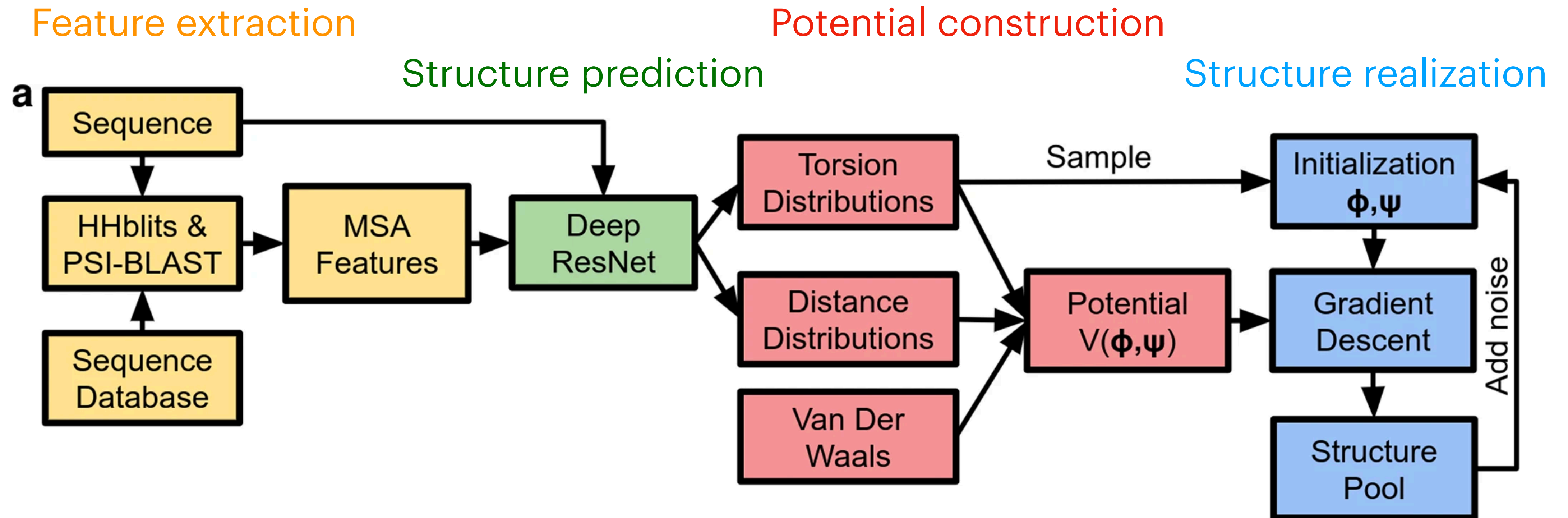
a

Contact precisions		L long			L/2 long			L/5 long		
Set	<i>N</i>	AF	498	032	AF	498	032	AF	498	032
FM	31	46.1	43.1	40.1	58.5	54.9	51.6	69.9	67.3	61.9
FM/TBM	12	59.1	53.0	48.9	74.2	64.5	64.2	85.3	81.0	79.6
TBM	61	68.3	65.5	61.9	82.4	80.3	76.4	90.6	90.5	87.1



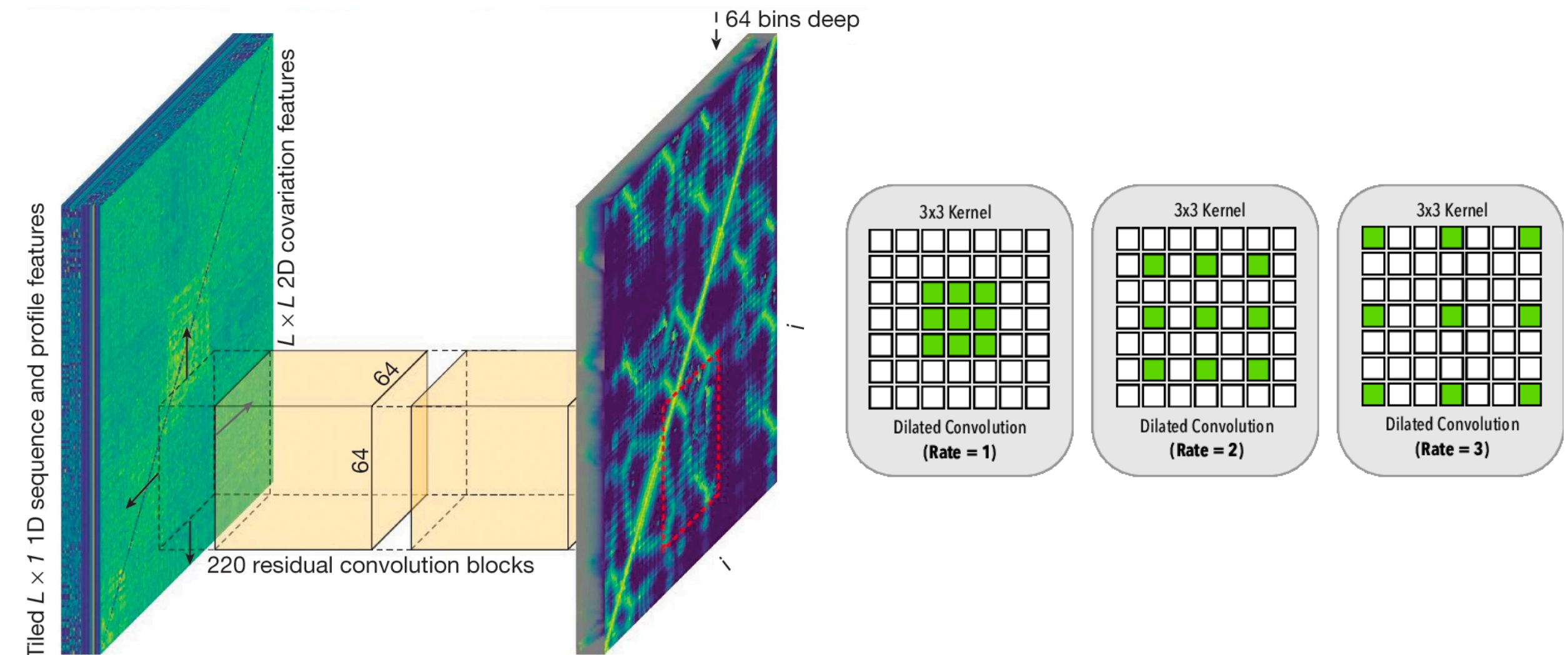
a, Precisions (as shown in Fig. 1c) for long-range contact prediction in CASP13 for the most probable L , $L/2$ or $L/5$ contacts, where L is the length of the domain. The distance distributions used by AlphaFold (AF) in CASP13, thresholded to contact predictions, are compared with submissions by the two best-ranked contact prediction methods in CASP13: 498 (RaptorX-Contact²⁶) and 032 (TripletRes³²), on ‘all groups’ targets, with updated domain definitions for T0953s2. **b**, **c**, True distances (**b**) and modes of the predicted distogram (**c**) for CASP13 target T0990. CASP divides this chain into three domains as shown (D3 is inserted in D2) for which there are 39, 36 and 42 HHblits alignments, respectively (from the CASP website).

The overall folding system



Structure prediction: residual 2D convolutional networks

- A 220 layer residual 2D convolutional network, 21M parameters
- Dilated convolutions for efficient long-range interaction
- Predicts pairwise *distance* distribution instead of contacts
- Mean and max pooling of the 2D activations in the second to last layer to get an 1-D output head:
 - 8-class secondary structure labels (computed by DSSP)
 - Relative accessible surface area
 - Backbone torsions as discretized marginal Ramachandran distributions for each residue
- At test time:
 - Tile crops to cover the whole histogram
 - Average across 4 models



More architecture and training details...

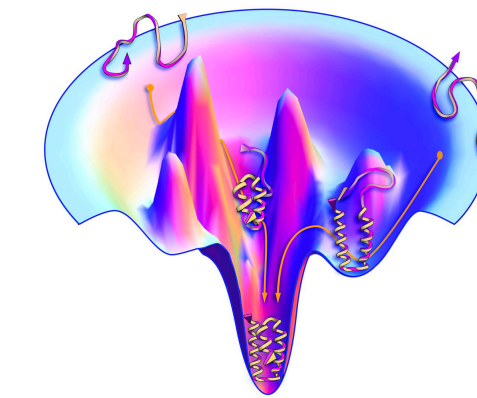
- 7 groups of 4 blocks with 256 channels, cycling through dilations 1, 2, 4, 8.
- 48 groups of 4 blocks with 128 channels, cycling through dilations 1, 2, 4, 8.
- Optimization: synchronized stochastic gradient descent
- Batch size: batch of 4 crops on each of 8 GPU workers.
- 0.85 dropout keep probability.
- Nonlinearity: ELU.
- Learning rate: 0.06.
- Auxiliary loss weights: secondary structure: 0.005; accessible surface area: 0.001. These auxiliary losses were cut by a factor 10 after 100 000 steps.
- Learning rate decayed by 50% at 150,000, 200,000, 250,000 and 350,000 steps.
- Training time: about 5 days for 600,000 steps.
- Data augmentation: MSA subsampling, coordinate noise, distogram cropping

How did they design the system?

Building a protein-specific potential

- To construct a differentiable potential, the predicted distance distribution is interpolated with a cubic spline
- A distance potential is created from the negative log likelihood of the distances, summed over all pairs of residues

$$V_{\text{distance}}(\mathbf{x}) = - \sum_{i,j, i \neq j} \log P(d_{ij} | \mathcal{S}, \text{MSA}(\mathcal{S}))$$

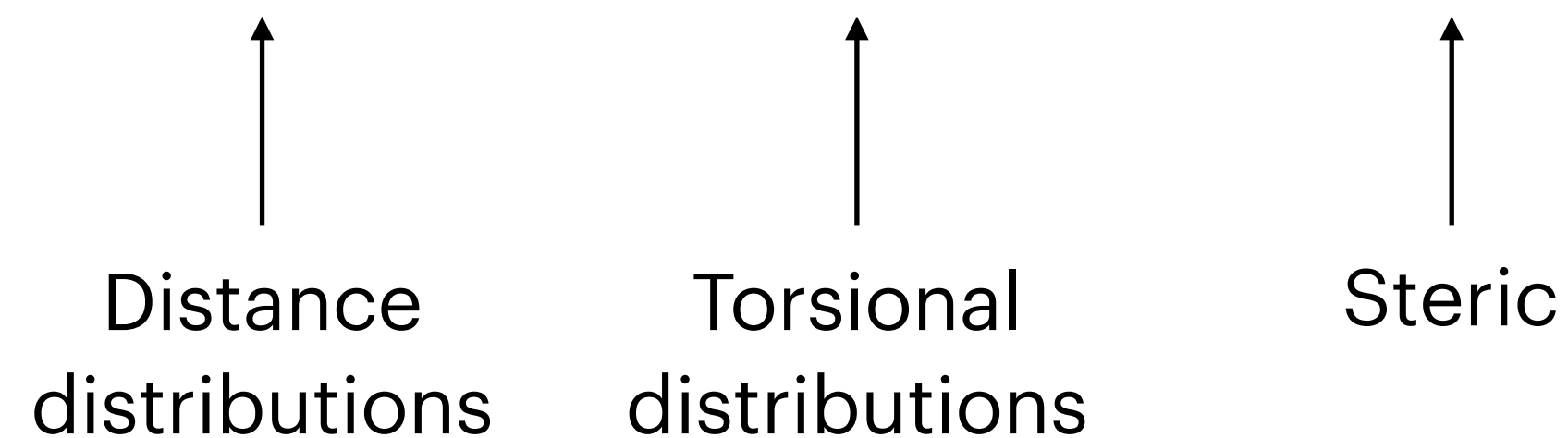


- The distance potential is adjusted based on a predicted reference distribution

$$V_{\text{distance}}(\mathbf{x}) = - \sum_{i,j, i \neq j} \log P(d_{ij} | \mathcal{S}, \text{MSA}(\mathcal{S})) - \log P(d_{ij} | \text{length}, \delta_{\alpha\beta})$$

- The final potential consists of three separate terms

$$V_{\text{total}}(\phi, \psi) = V_{\text{distance}}(G(\phi, \psi)) + V_{\text{torsion}}(\phi, \psi) + V_{\text{score2_smooth}}(G(\phi, \psi))$$



Building a protein-specific potential

Key aspects

- Use a very large number of distributional predictions from a neural network
 - $P_{ij}(\text{distance})$ for all pairs
 - $P_i(\phi, \psi)$ for each residue
- Individual predictions are detailed, calibrated, and smooth
- Averaging the agreement scores over large numbers of distributions predictions (e.g. all distances) gives an accurate and smooth scoring function

Realizing a protein structure with gradient descent

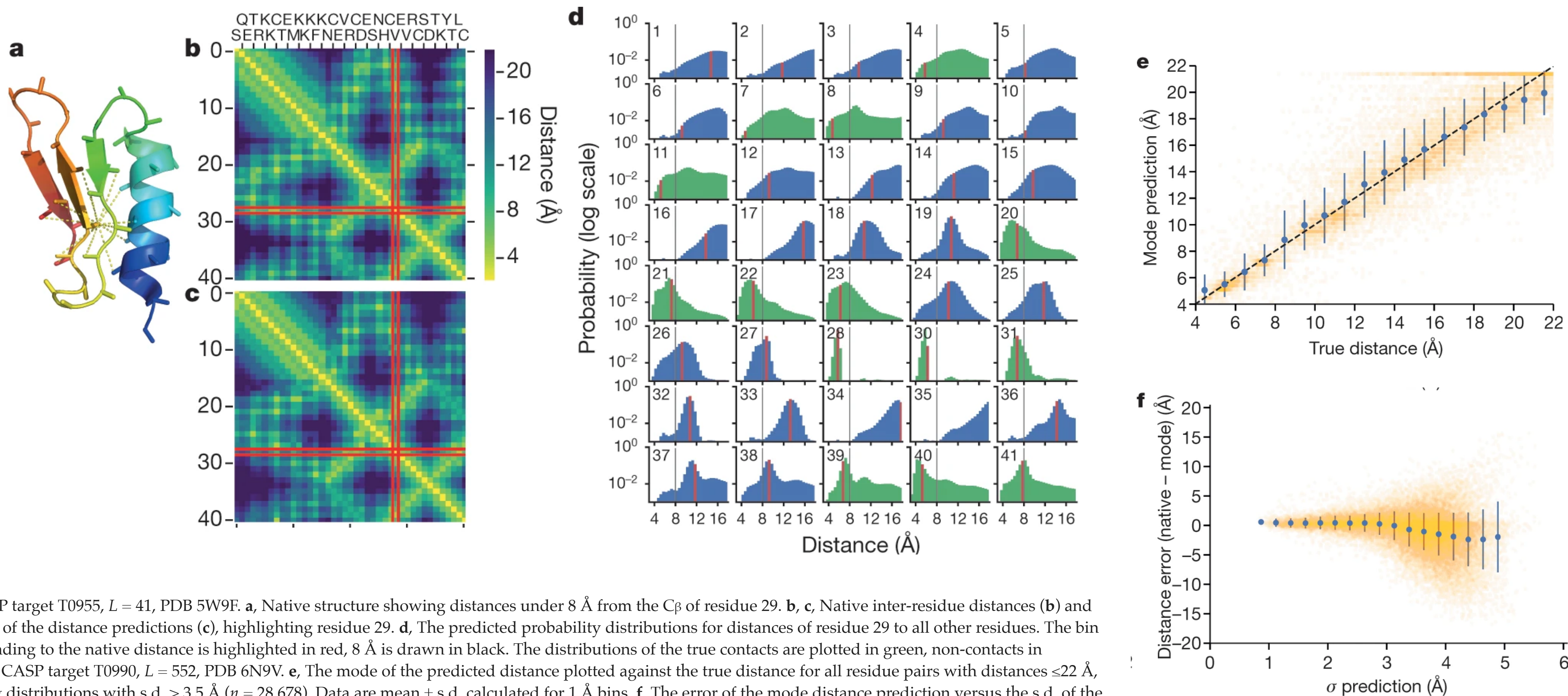
- The potential is differentiable with respect to backbone torsions angles
- Given an initial set of torsions, V_{total} is minimized by optimization with L-BFGS
- Repeat with multiple initializations and add noisy restarts (addition of 30 deg noise to the backbone torsions)
- Folding is performed on full chains, which avoids error-prone domain segmentation
- How is this related to the folding energy funnel?



Data: Input features, training and test sets

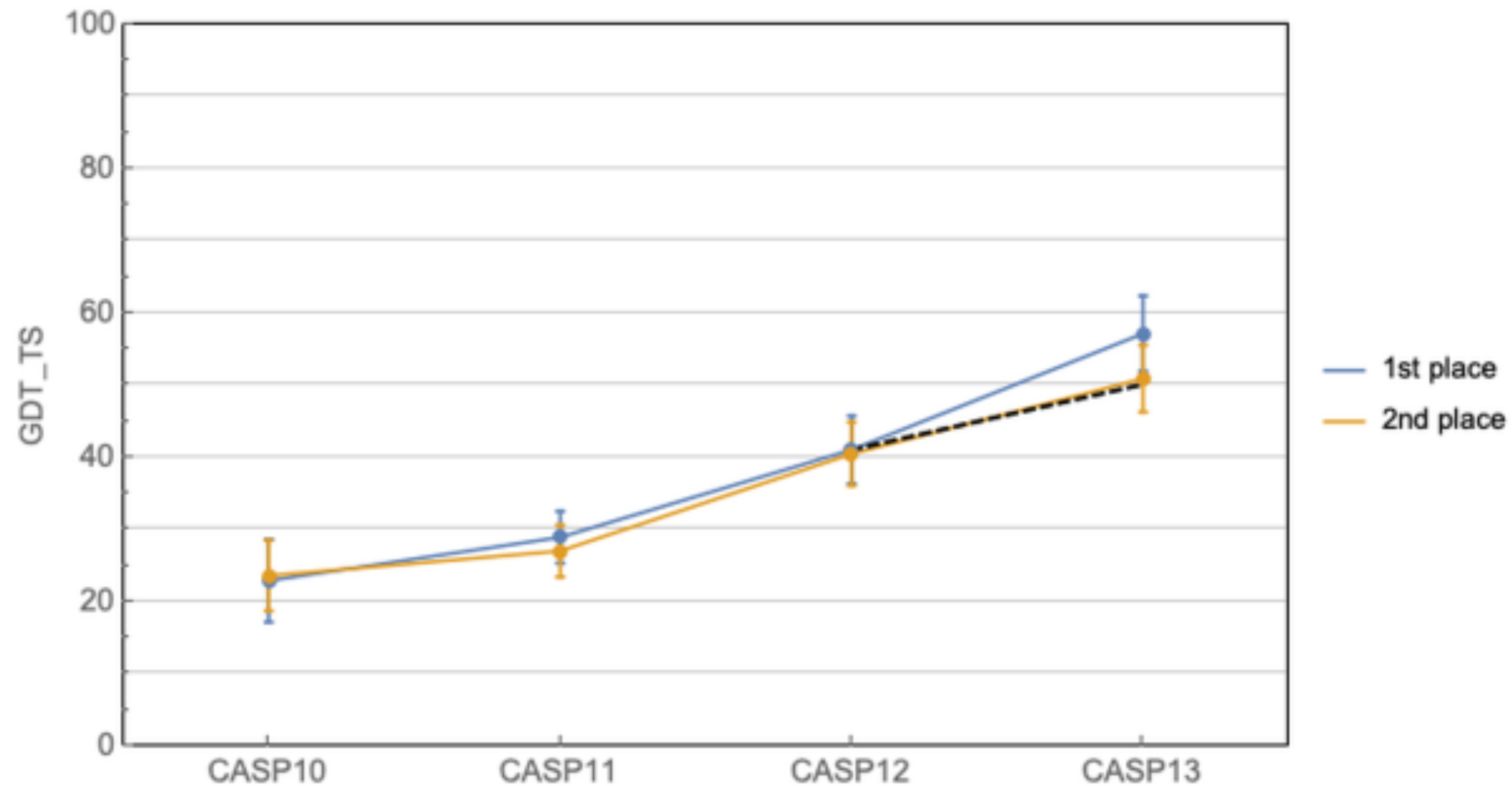
- PDB structures as of March 2018:
 - Extract non-redundant domains by using the CATH 35% sequence similarity cluster representatives
 - 31,247 domains, split into 29,427 train and 1,820 test
 - Keeping all domains from the same homologous superfamilies (H-level) in the same partition
 - CATH superfamilies of FM in CASP11 and CASP12 were excluded from training
- Sequences from Uniclust30 as of Oct 2017:
 - HHblits to generate the MSA and other covariation features
 - Additional features from PSI-BLAST

Predicted distance distributions compared with true distances



a–d, CASP target T0955, $L = 41$, PDB 5W9F. **a**, Native structure showing distances under 8 Å from the C β of residue 29. **b**, **c**, Native inter-residue distances (**b**) and the mode of the distance predictions (**c**), highlighting residue 29. **d**, The predicted probability distributions for distances of residue 29 to all other residues. The bin corresponding to the native distance is highlighted in red, 8 Å is drawn in black. The distributions of the true contacts are plotted in green, non-contacts in blue. **e**, **f**, CASP target T0990, $L = 552$, PDB 6N9V. **e**, The mode of the predicted distance plotted against the true distance for all residue pairs with distances ≤ 22 Å, excluding distributions with s.d. > 3.5 Å ($n = 28,678$). Data are mean $\pm$ s.d. calculated for 1 Å bins. **f**, The error of the mode distance prediction versus the s.d. of the distance distributions, excluding pairs with native distances > 22 Å ($n = 61,872$). Data are mean $\pm$ s.d. are shown for 0.25 Å bins. The true distance matrix and distogram for T0990 are shown in Extended Data Fig. [2b, c](#).

AlphaFold1 @ CASP13: “What just happened?”



- Substantial progress since CASP12
- Substantial distance between 1st and 2nd place

See blog post by Mohammed AlQuraishi for more

<https://moalquraishi.wordpress.com/2018/12/09/alphafold-casp13-what-just-happened/>

Two years later...

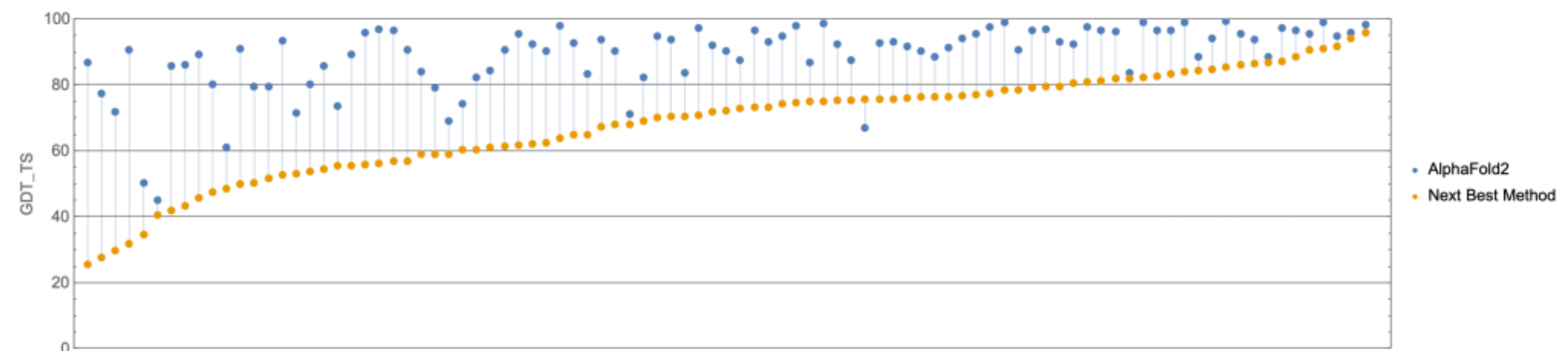
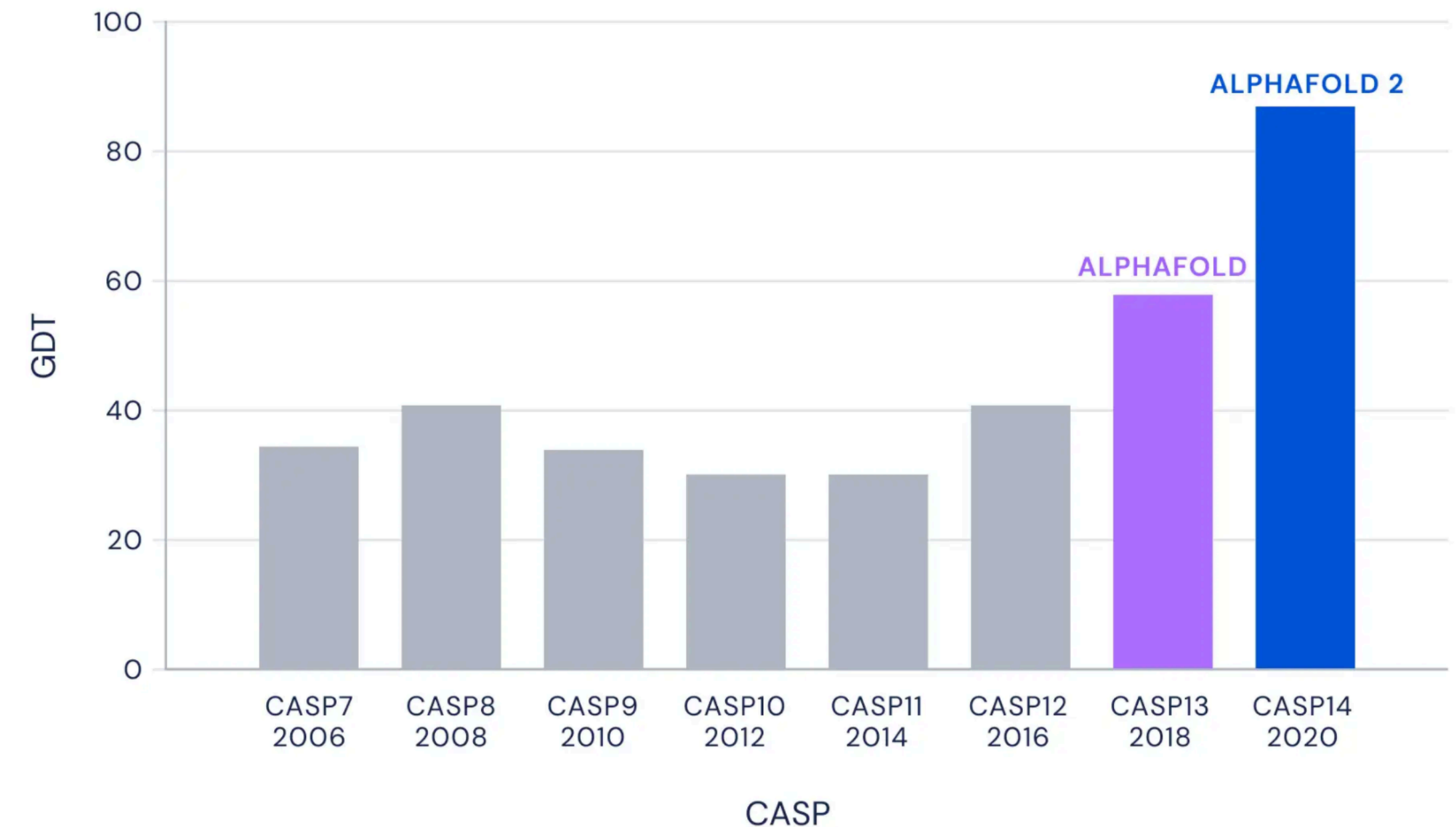
92.4 GDT

AlphaFold performance at CASP14

AlphaFold2 at CASP14

- CASP¹: Biannual community-wide blinded competition on ~100 newly solved proteins
- CASP14 press release: “Artificial intelligence solution to a 50-year-old science challenge could ‘revolutionise’ medical research”
- 92.4 median GDT
 - (global distance test, 0-100)
 - 0.96 Å backbone RMSD, 1.6 Å all-atom RMSD
- **Above >90 GDT considered within experimental error**

Median Free-Modelling Accuracy

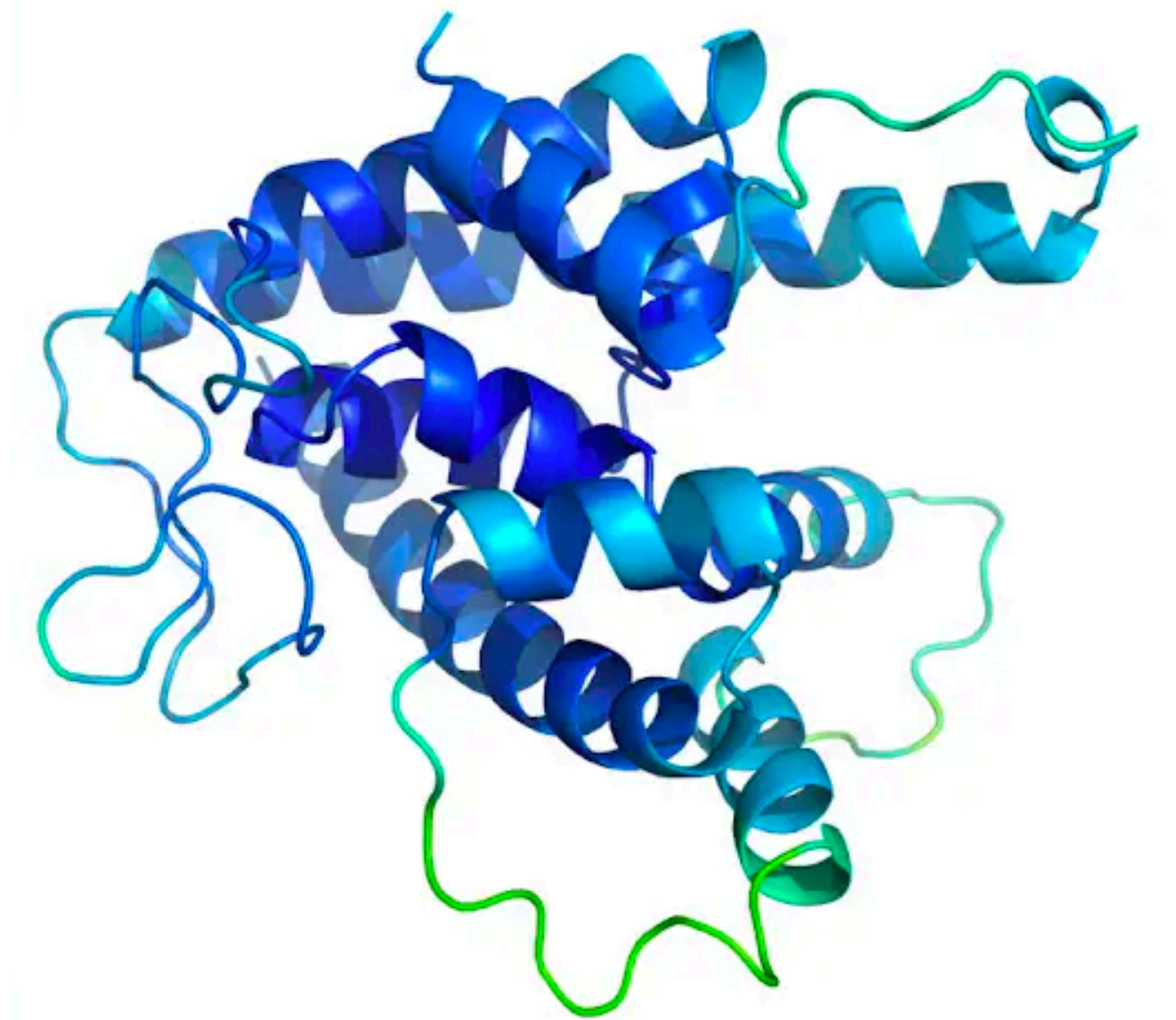


Inside the AlphaFold system

Input sequence
MRKPRTPTFTT...



Statistical
machine



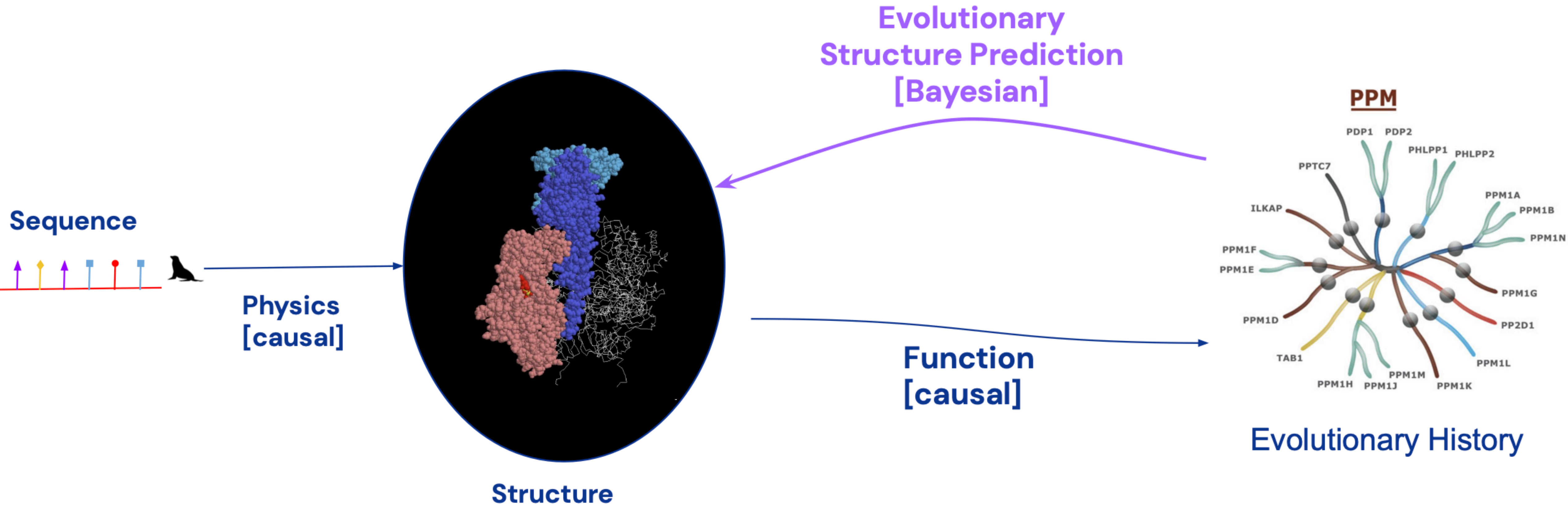
3D structure

DeepMind

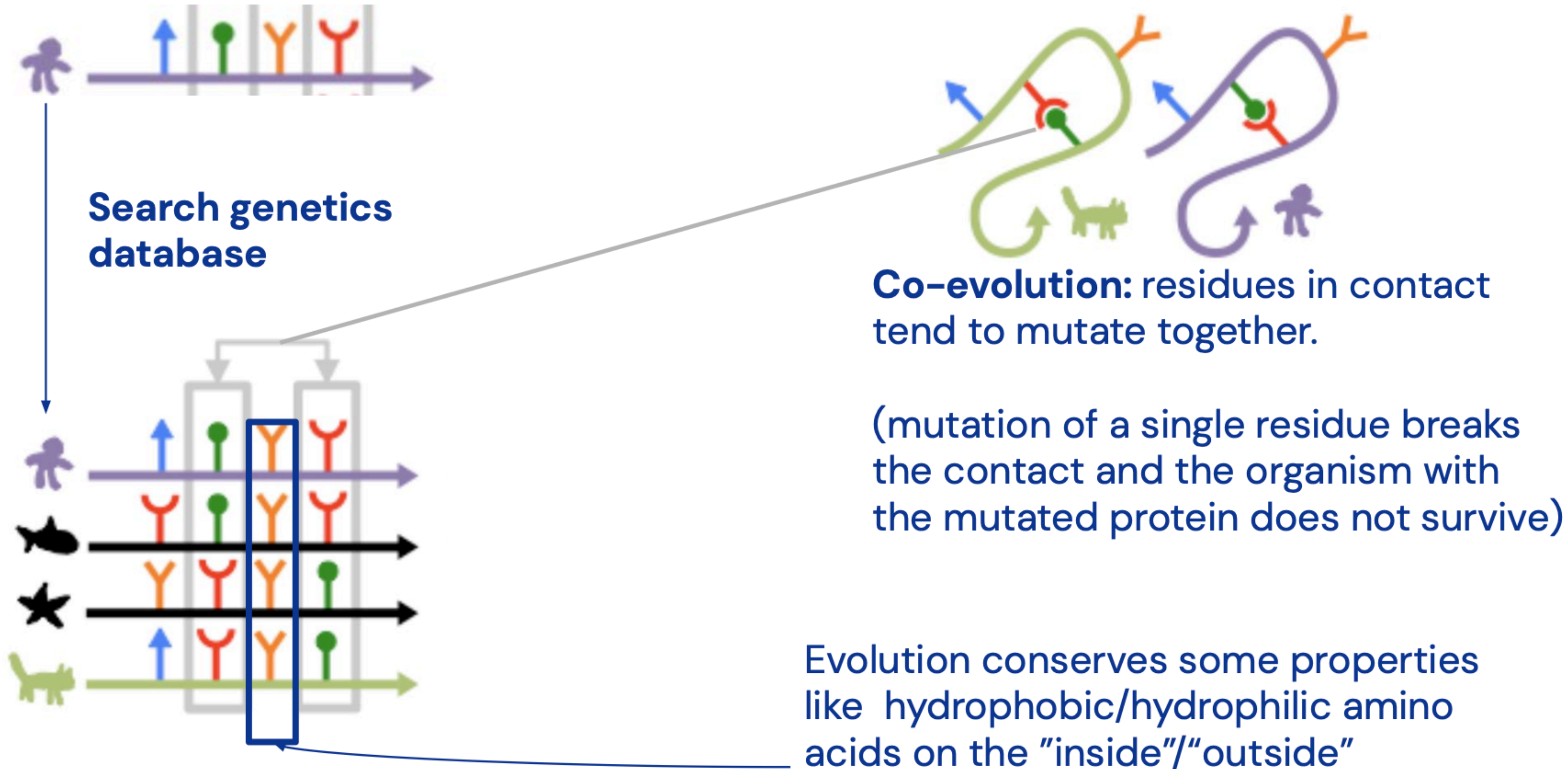
How AlphaFold works



Determining Structure from Evolution - Intuition



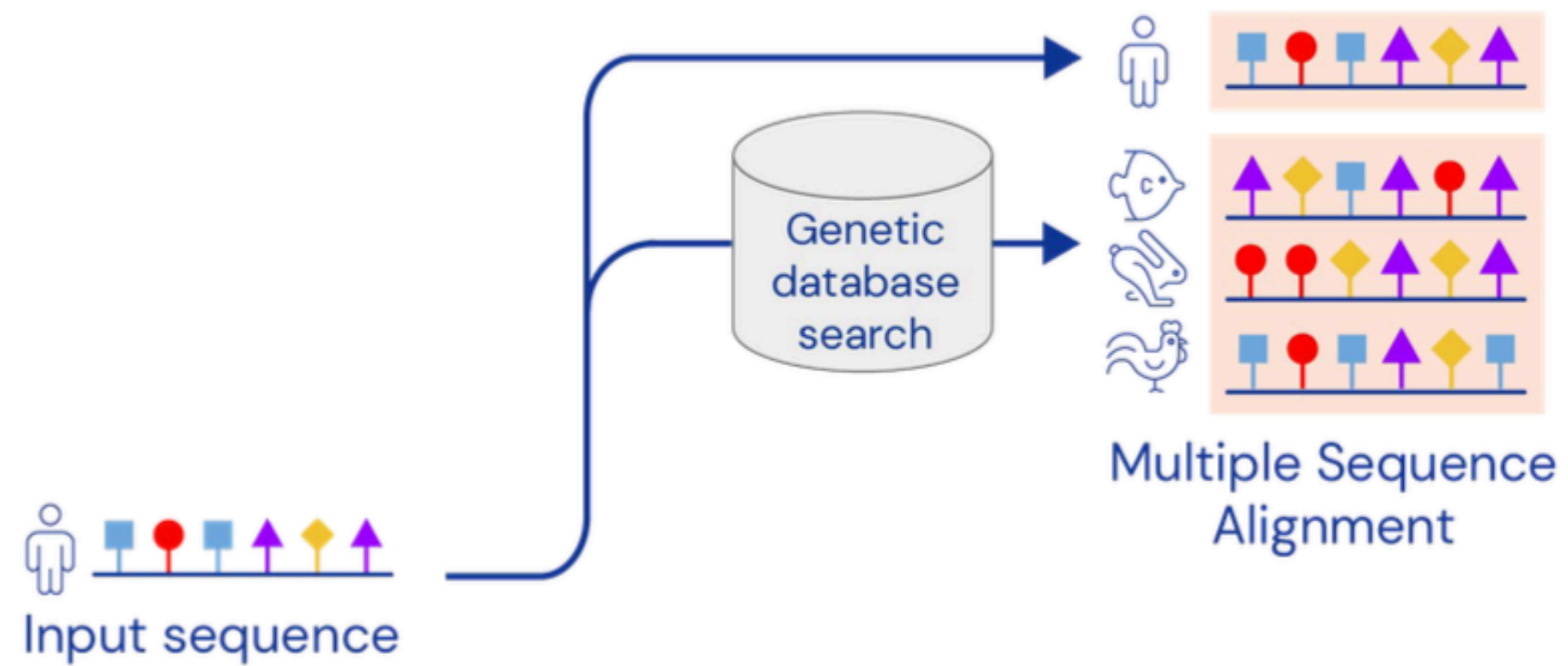
Determining Structure from Evolution - Intuition



*Coevolution cartoon by Sergey Ovchinnikov
(<https://jgi.doe.gov/seeking-structure-metagenome-sequences/cartoon-coevolution-sergey-o/>)*



Inputs

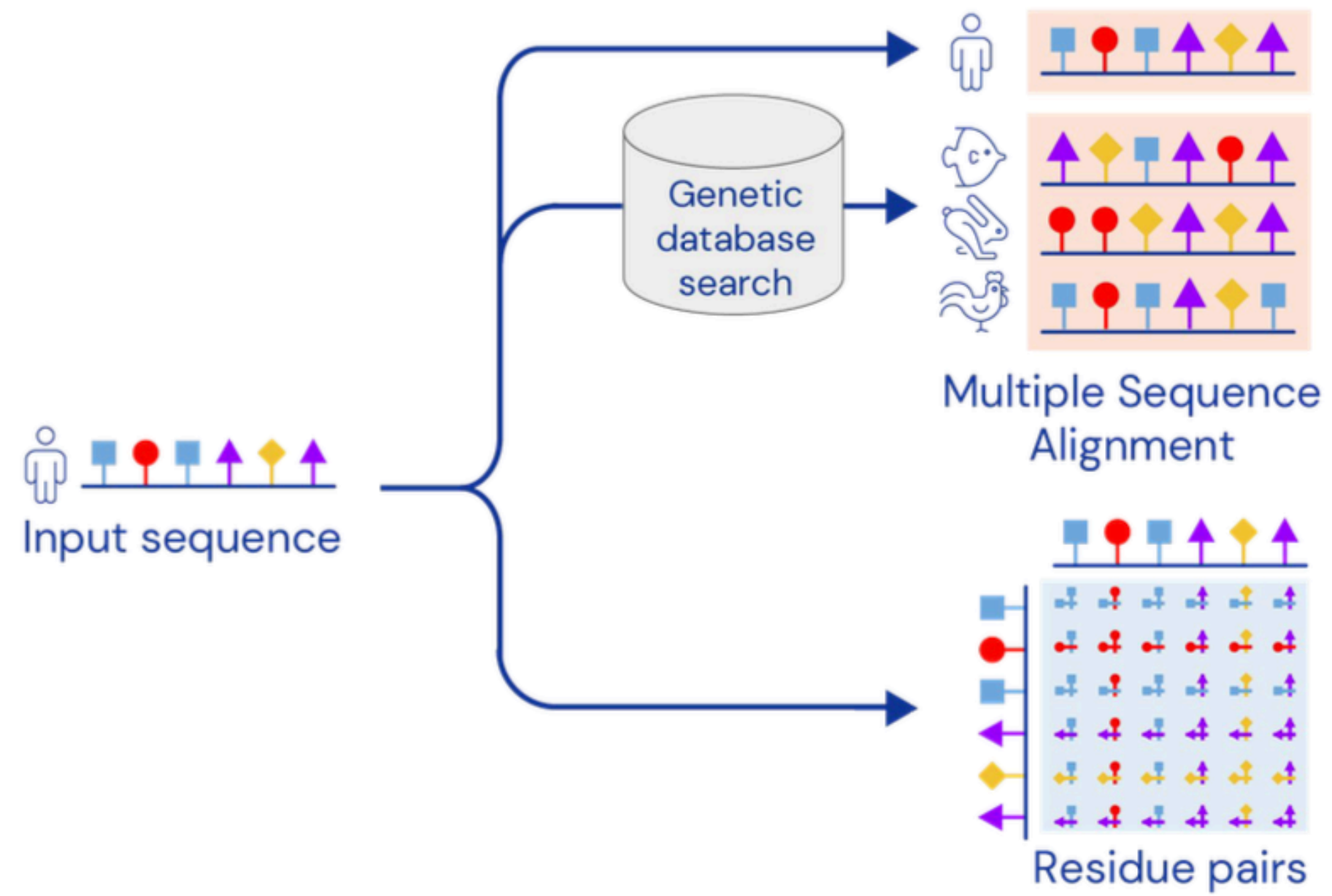


Uses raw MSA, not a pre-computed statistical model (e.g. Potts)

A key AlphaFold input is the MSA, containing sequences evolutionarily related to the target. Related sequences are found using standard tools and public databases.



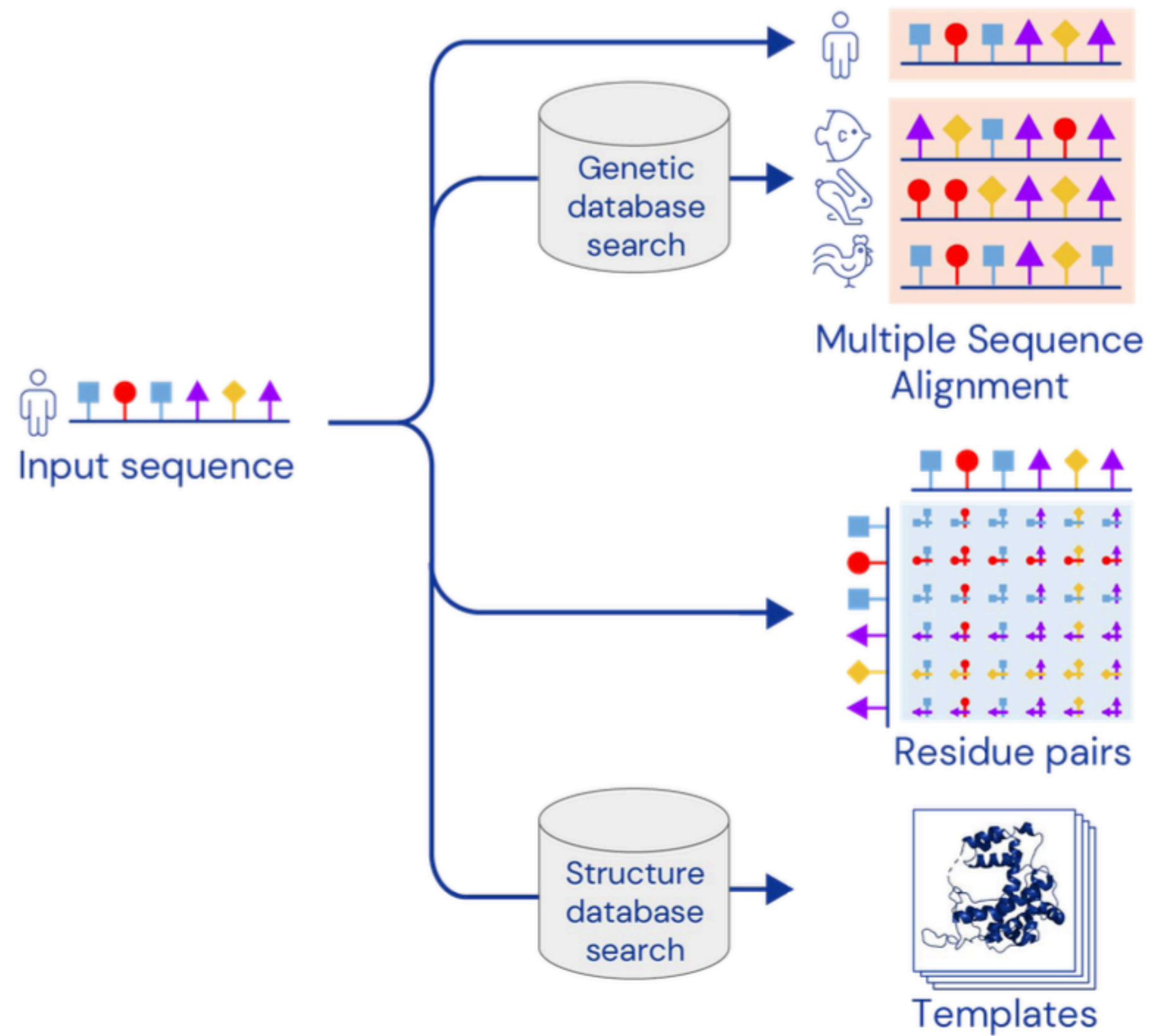
Inputs



The input sequence is used to create an array of representations representing all residue pairs.



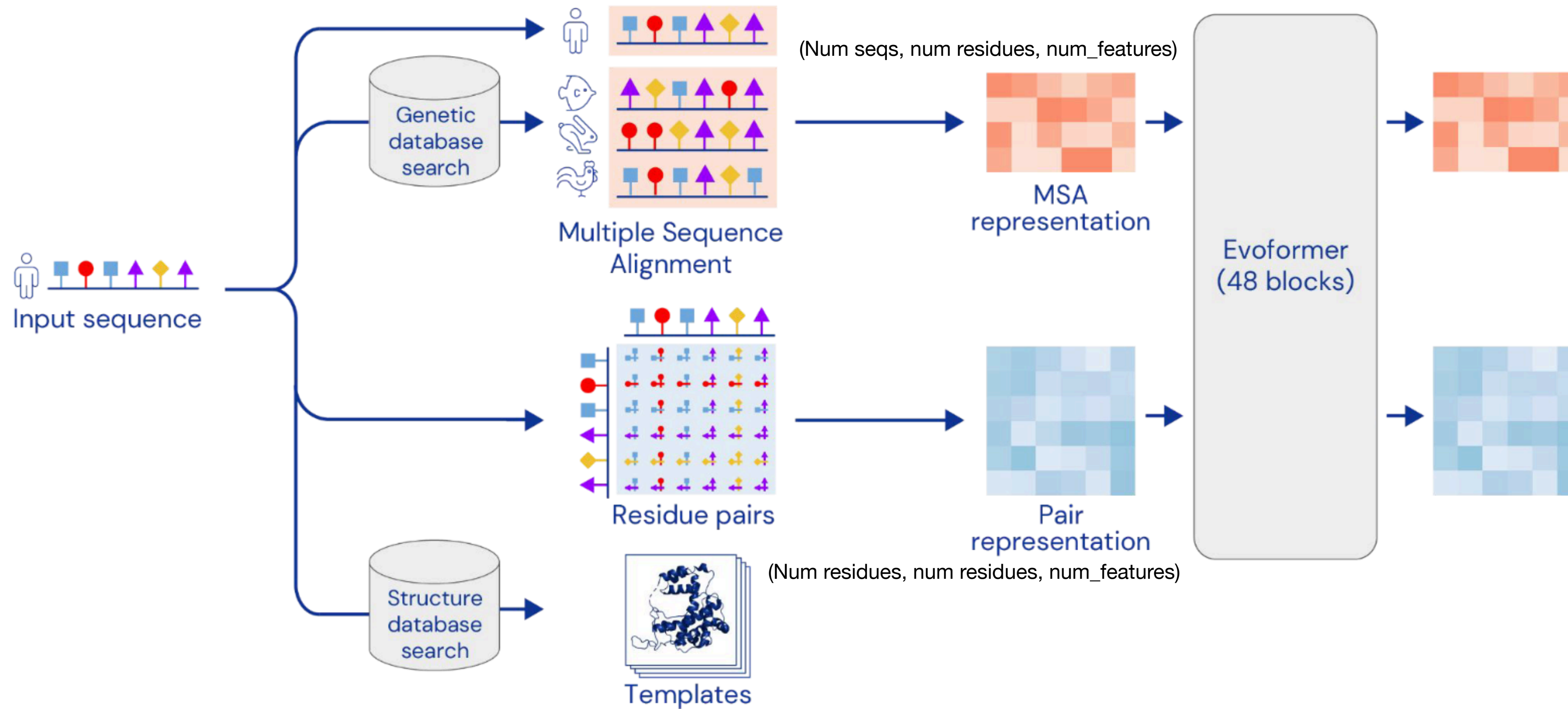
Inputs



AlphaFold can also use template structures from the PDB, found using standard tools. However, it often produces accurate predictions without a template.



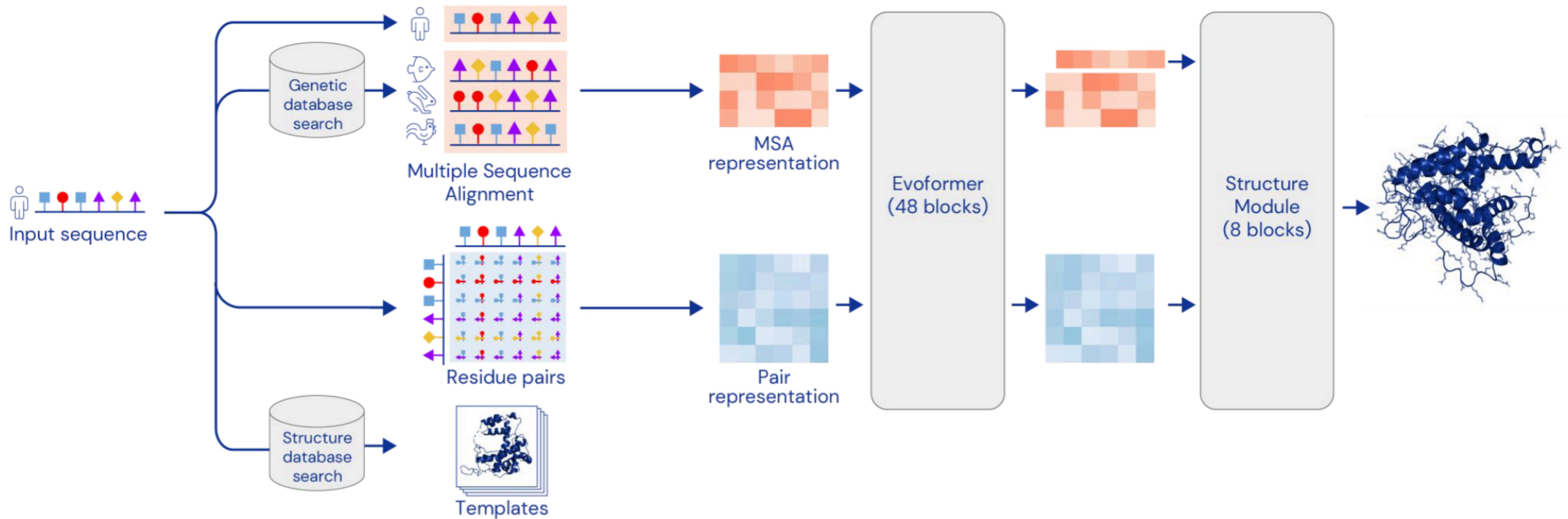
Network



The Evoformer blocks extract information about the relationship between residues. The MSA representation can update the pair representation and vice versa.



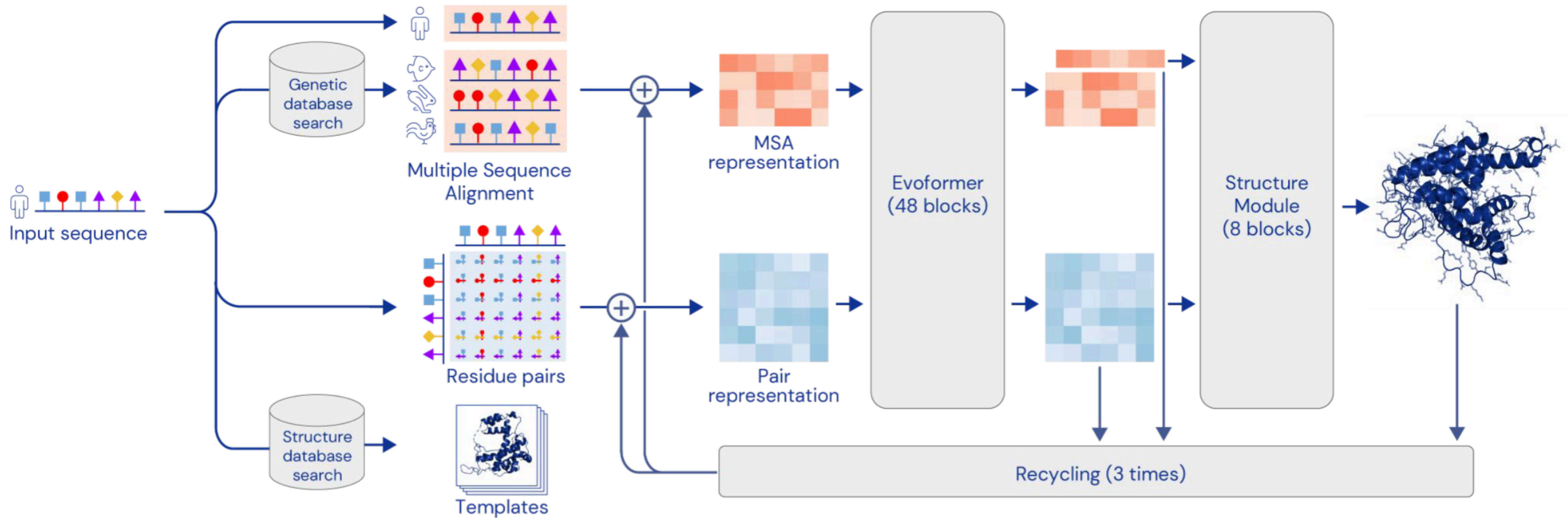
Network



The Structure Module predicts a rotation + translation to place each residue.
A small network predicts side chain chi angles.



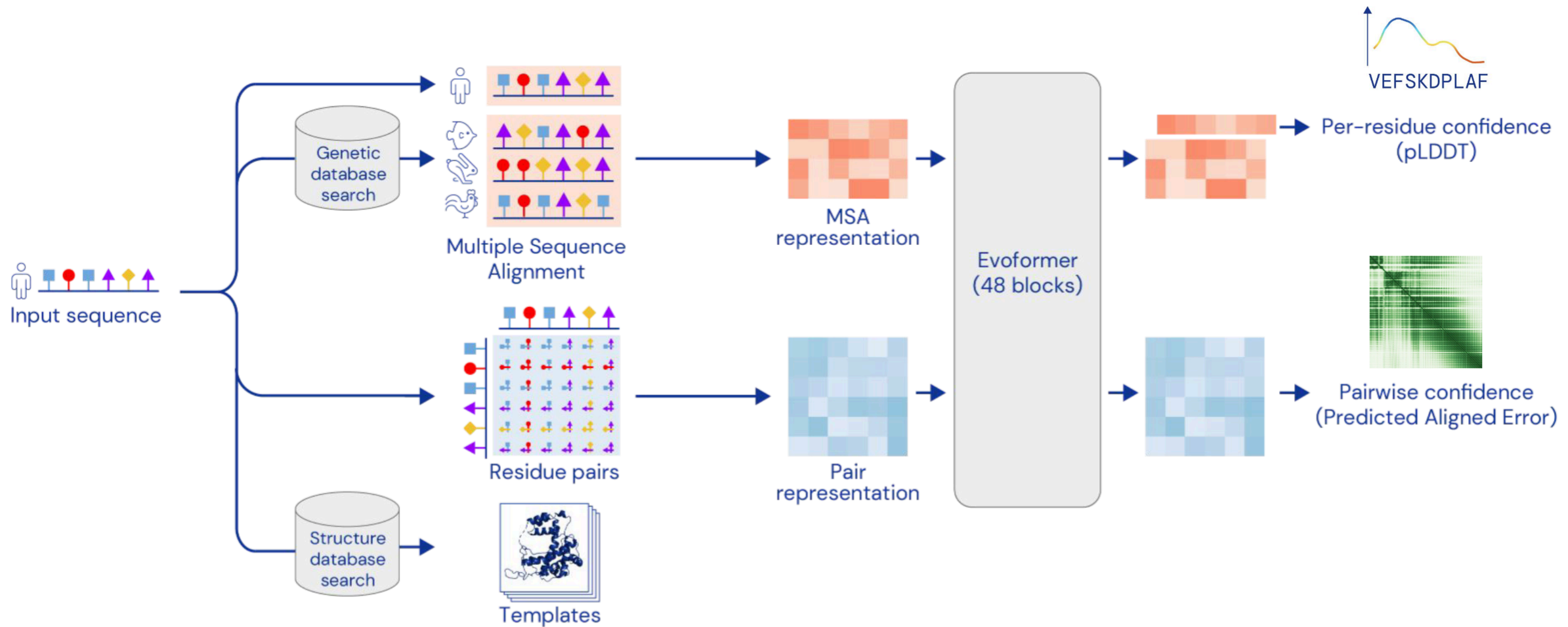
Network



Feeding certain outputs back through the network again improves performance



Confidence estimates

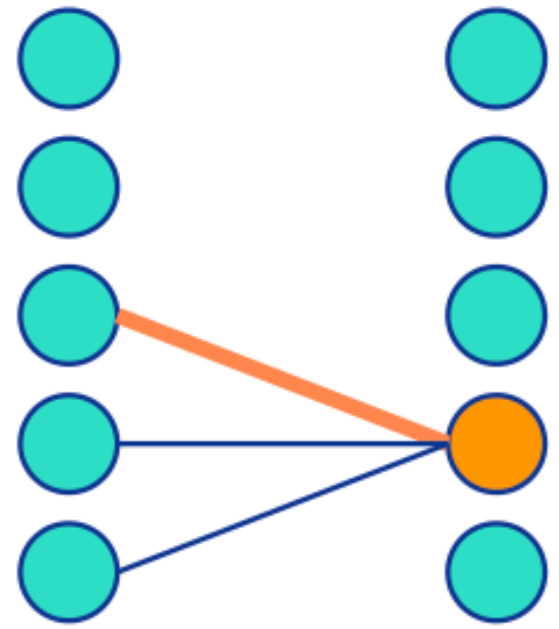


AlphaFold is trained to predict its confidence via simple supervision on the training set

We use these confidences for ranking of predictions

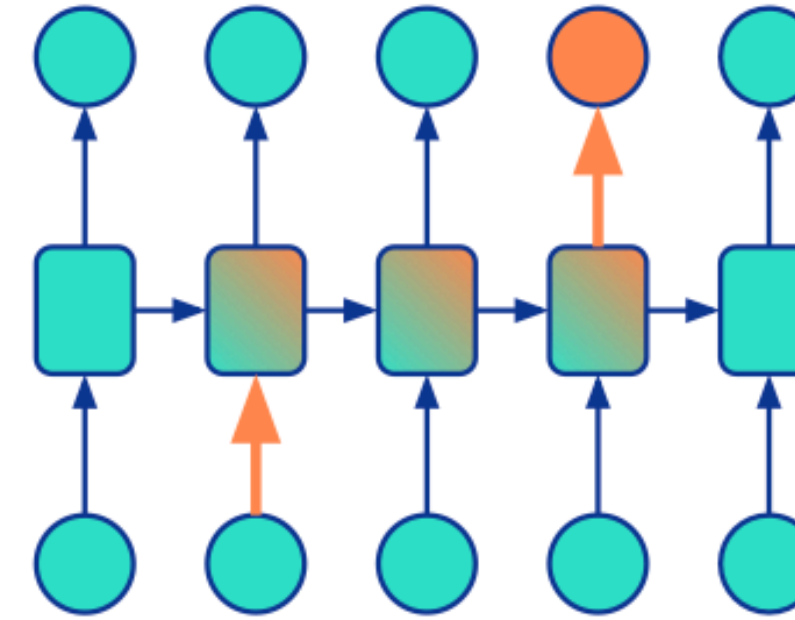


Inductive Bias for Deep Learning Models



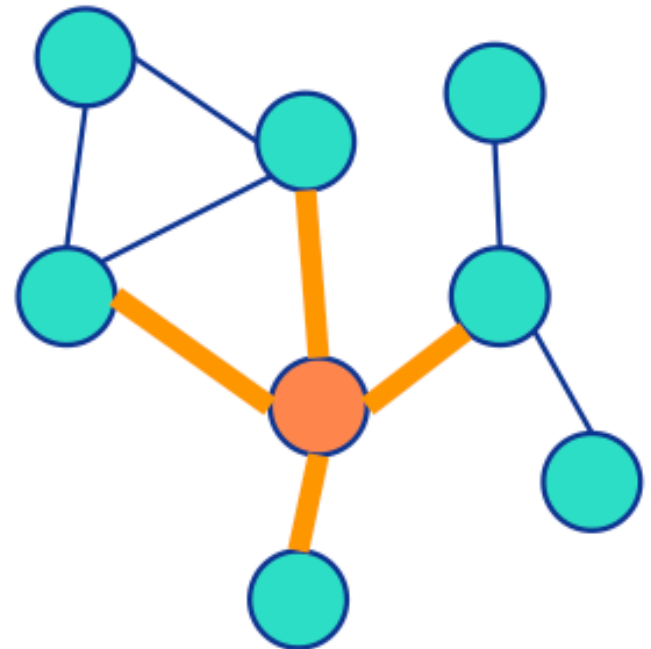
Convolutional Networks (e.g. computer vision)

- data in regular grid
- information flow to local neighbours
- AlphaFold 1



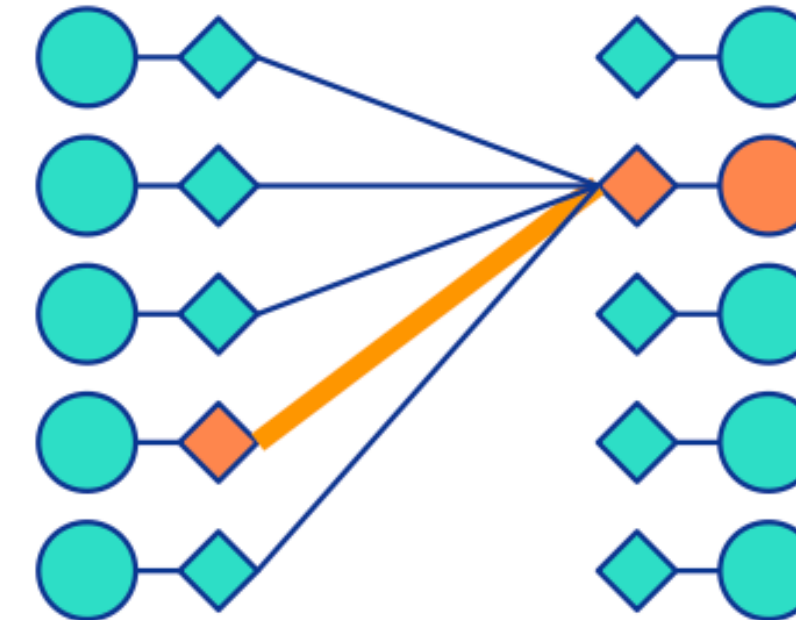
Recurrent Networks (e.g. language)

- data in ordered sequence
- information flow sequentially



Graph Networks (e.g. recommender systems or molecules)

- data in fixed graph structure
- information flow along fixed edges

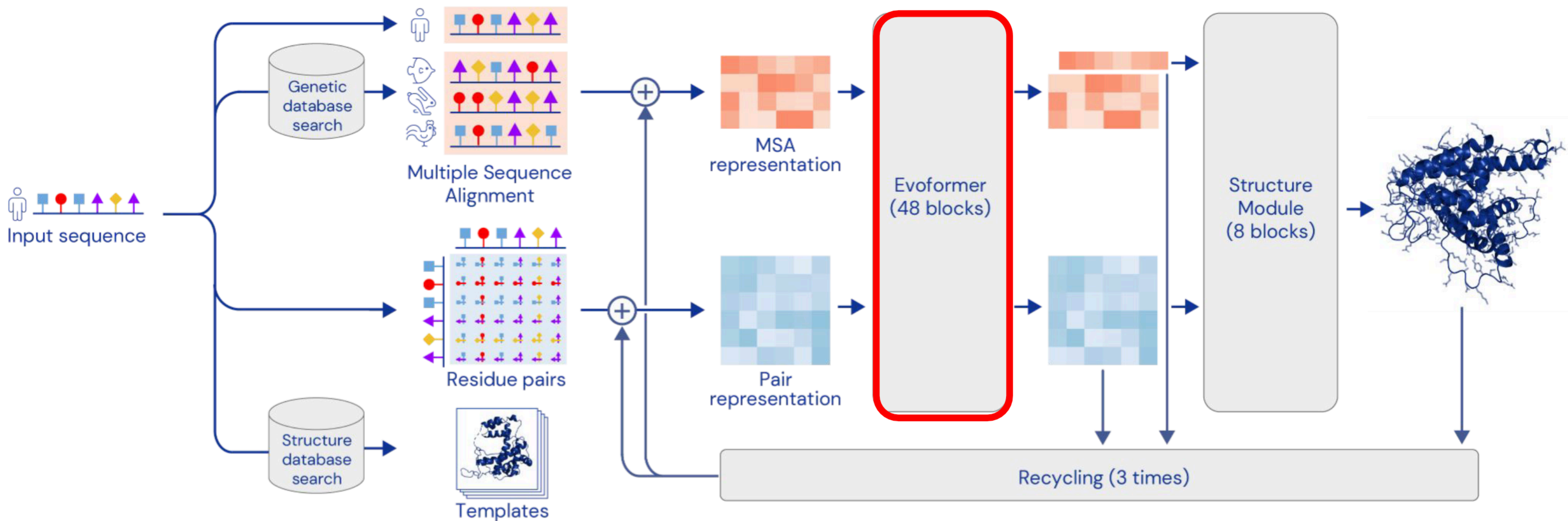


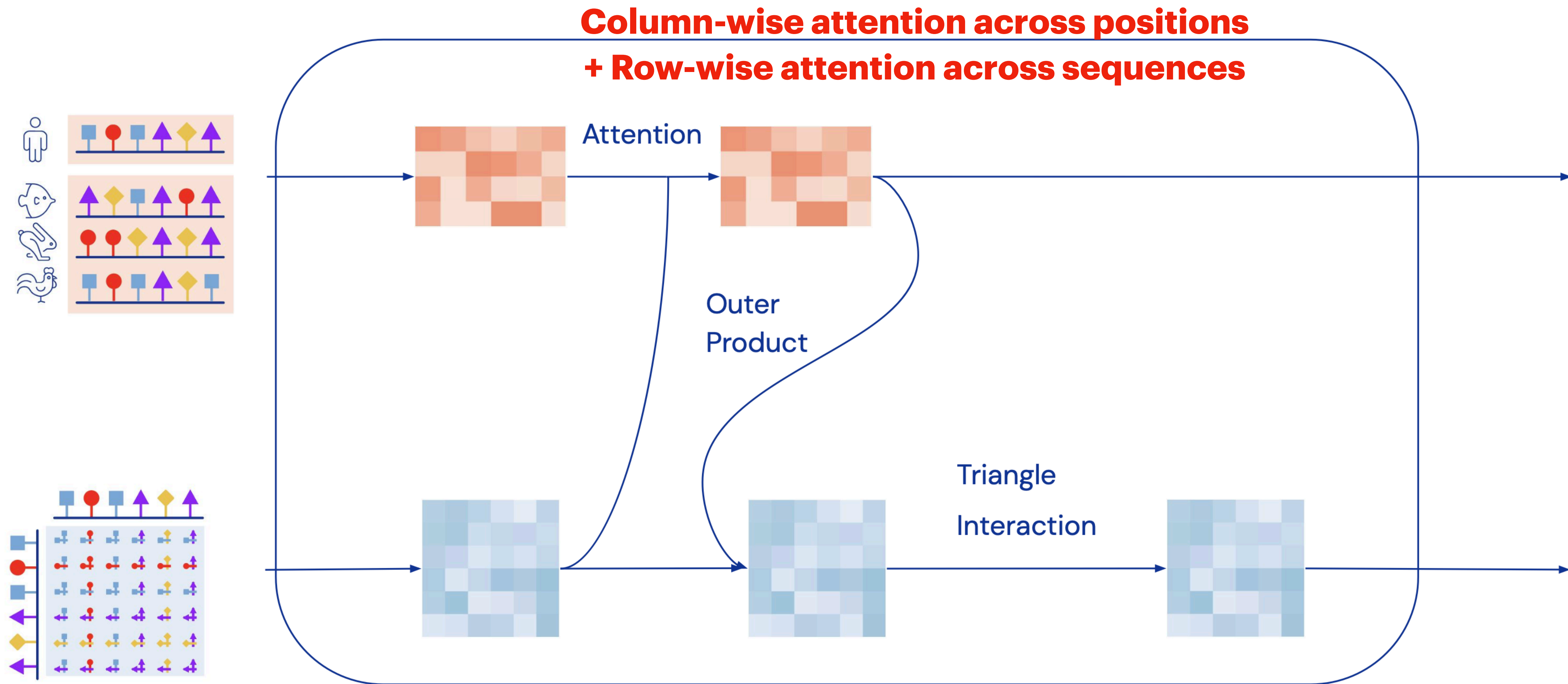
Attention Module (e.g. language)

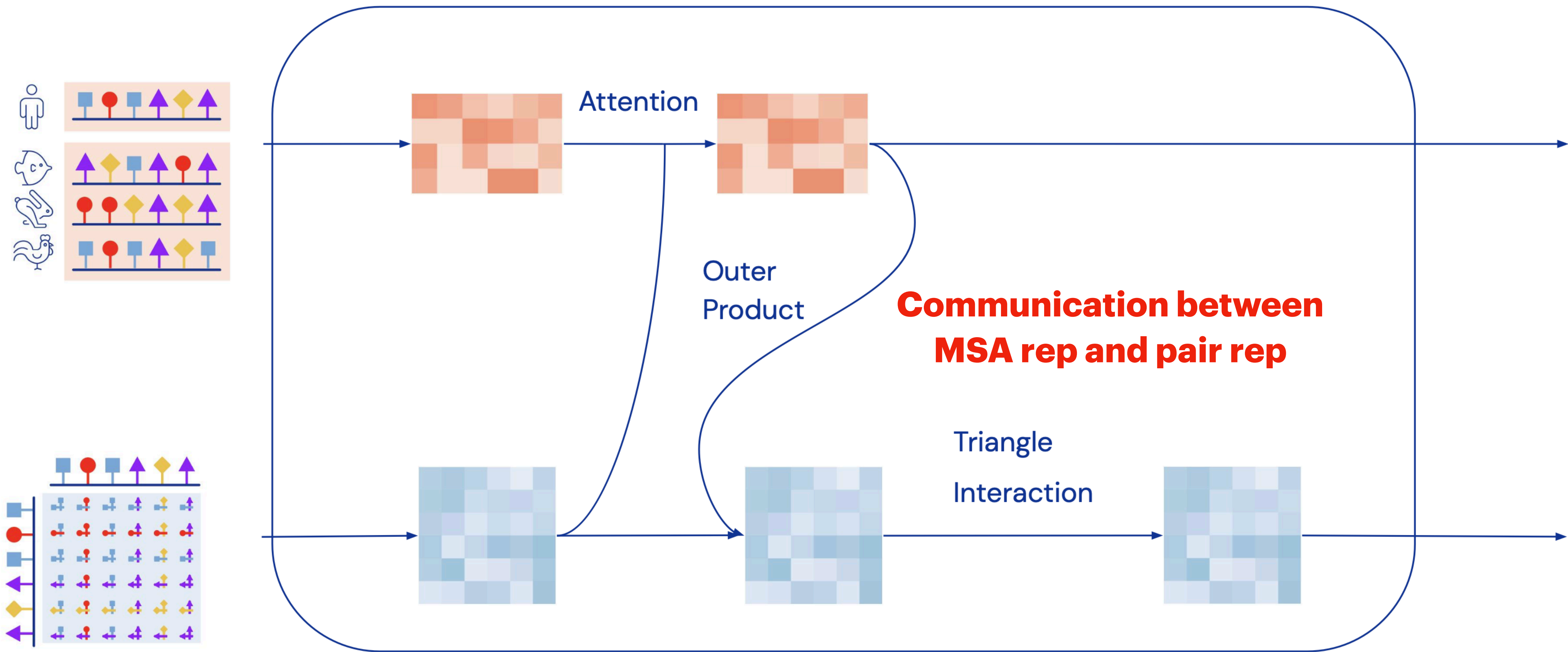
- data in unordered set
- information flow dynamically controlled by the network (via keys and queries)



Network

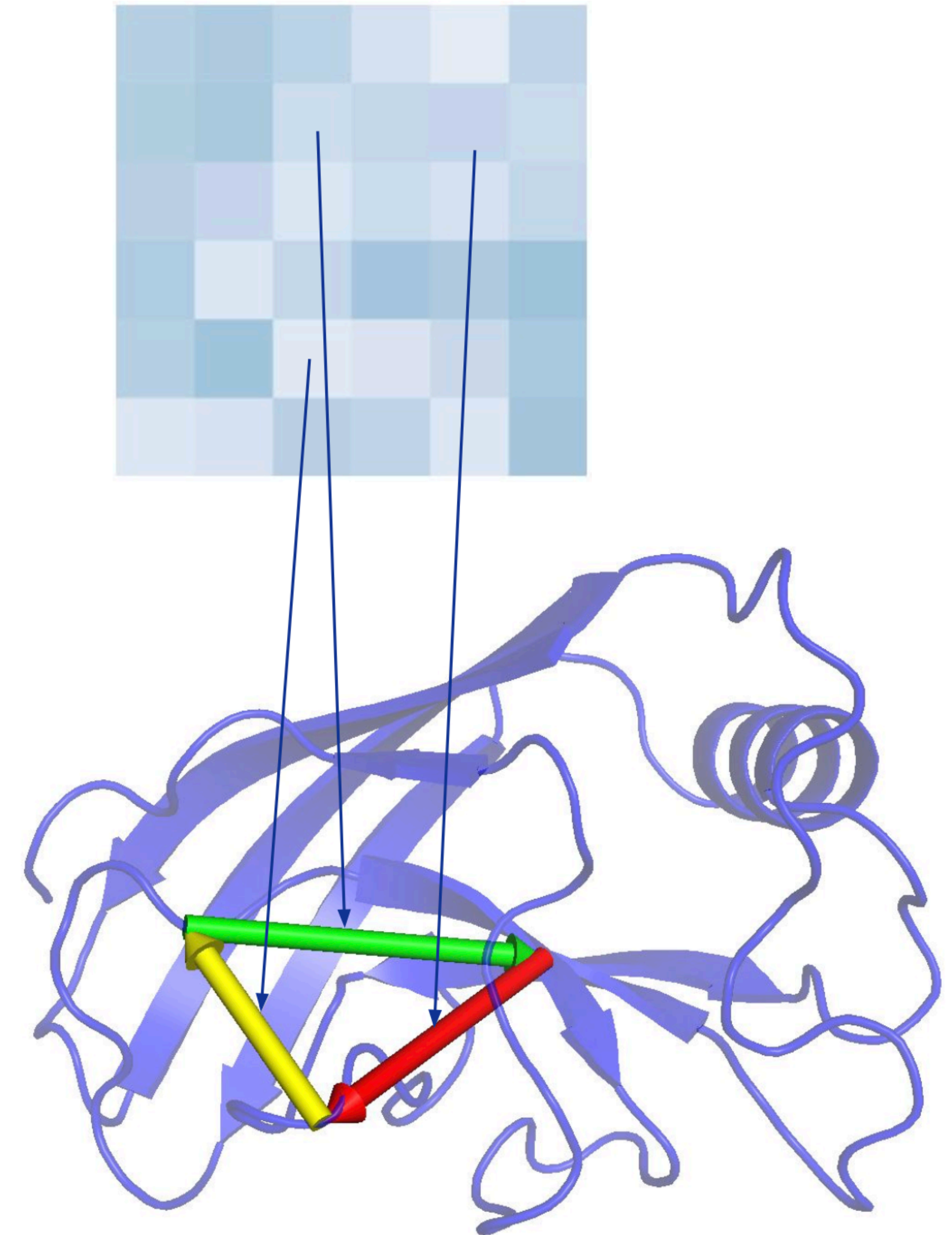




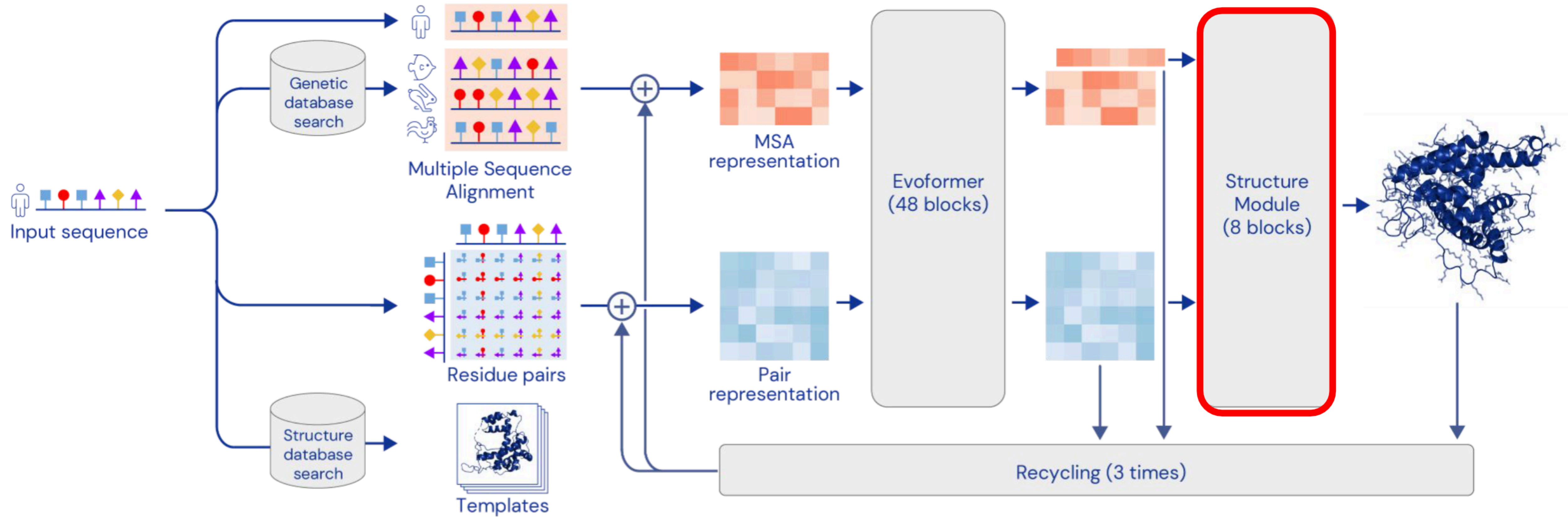


Triangle Interaction

- **Take 3 points A, B, C**
 - ◆ If Distance AB and distance BC known strong constraint on AC (triangle inequality)
 - ◆ Evolution & Sequence gives information about relations between residues
- **Pair Embedding encodes relations**
 - ◆ Update for pair AC should depend on BC, AB



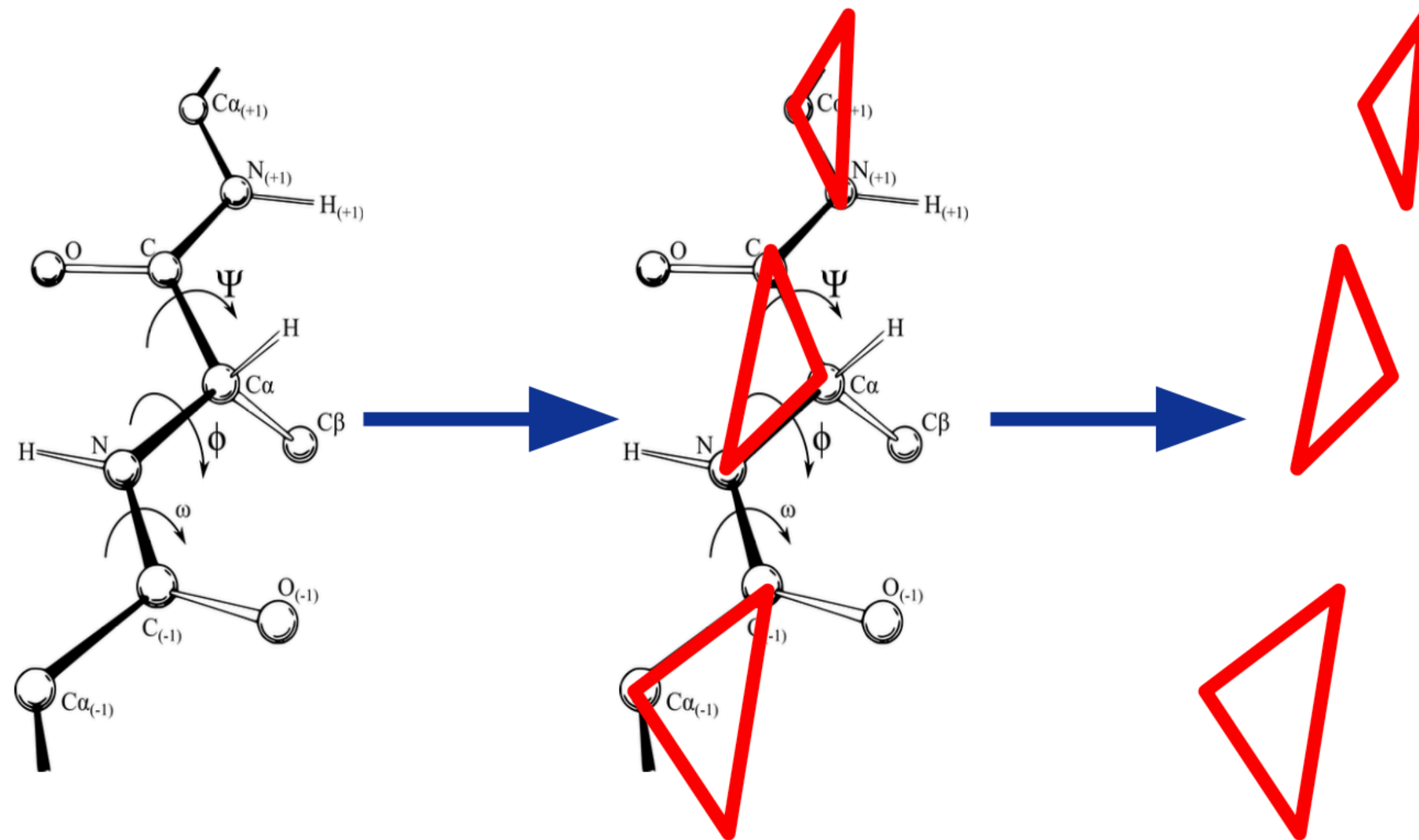
Network



Structure module

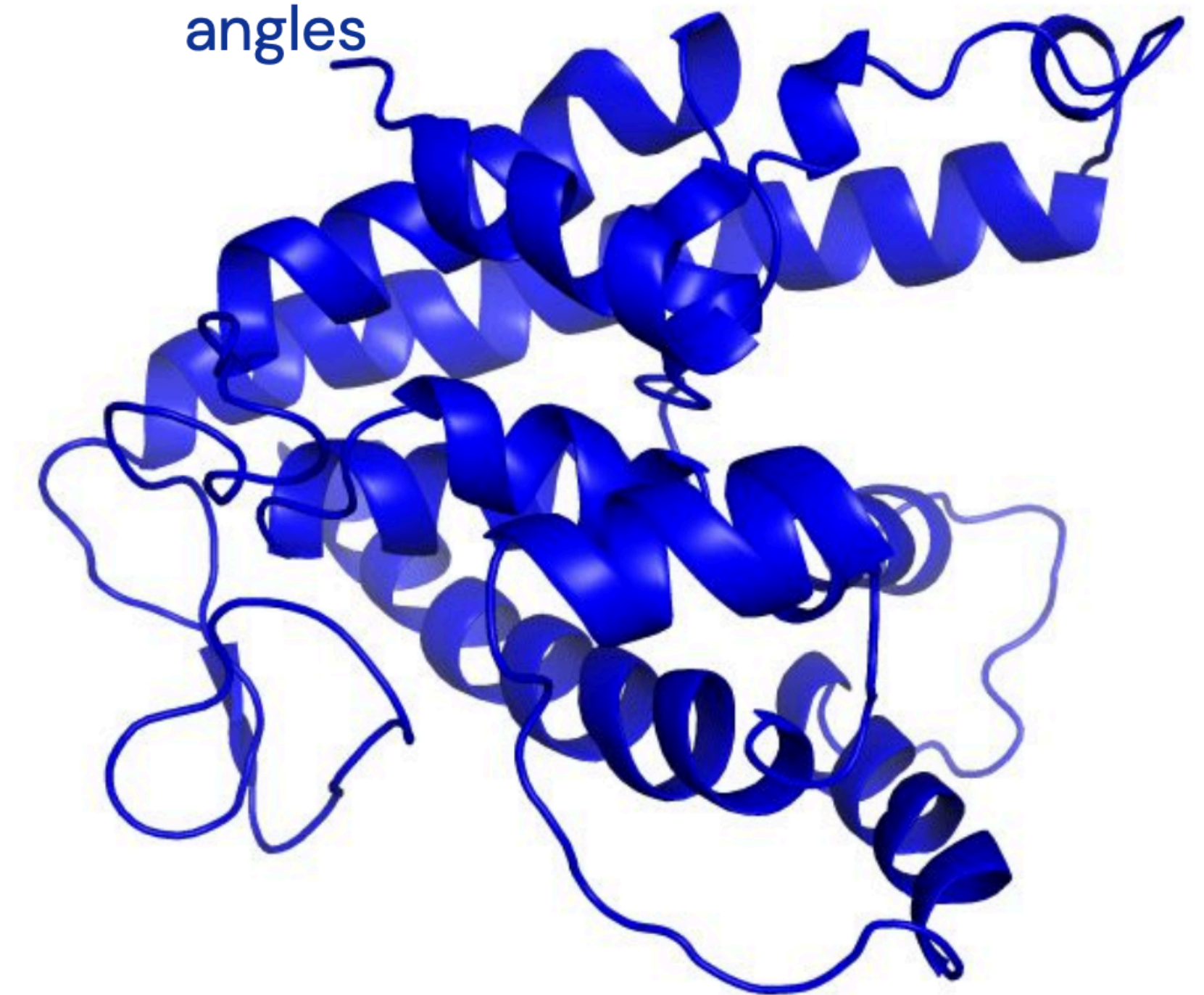
→ **End-to-end folding** instead of gradient descent

→ Protein backbone = gas of 3-D rigid bodies
(chain is learned!)



→ **3-D equivariant transformer architecture**
updates the rigid bodies / backbone
(Invariant Point Attention)

- Also builds the side chains from torsion angles



Iteration 1

Target: T1041

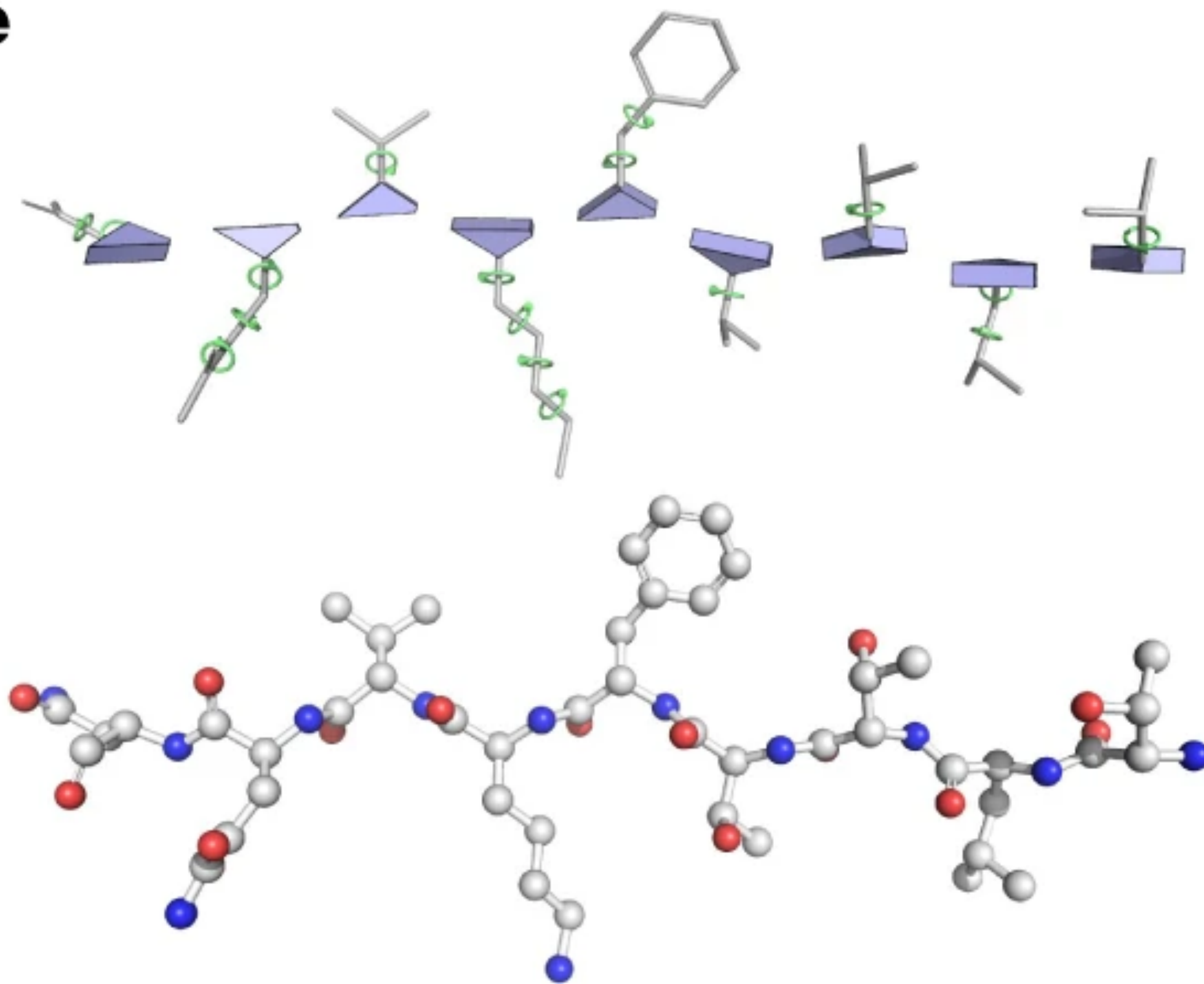
Image: Dcrjsr, vectorised Adam Rędzikowski (CC BY 3.0, Wikipedia)



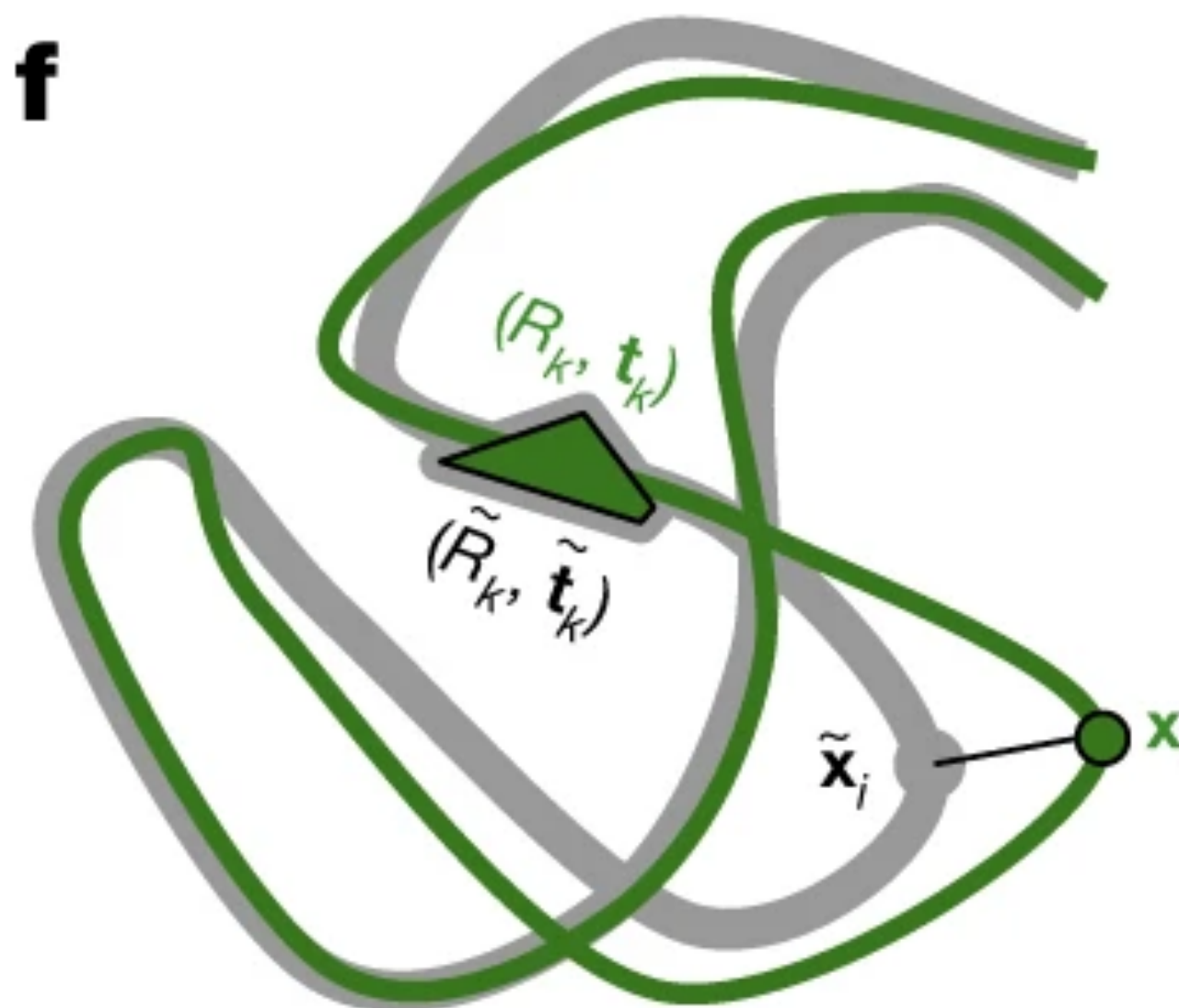
Loss: Frame-Aligned Point Error

Compute distance between true and predicted positions in residue pairs, across many different alignments

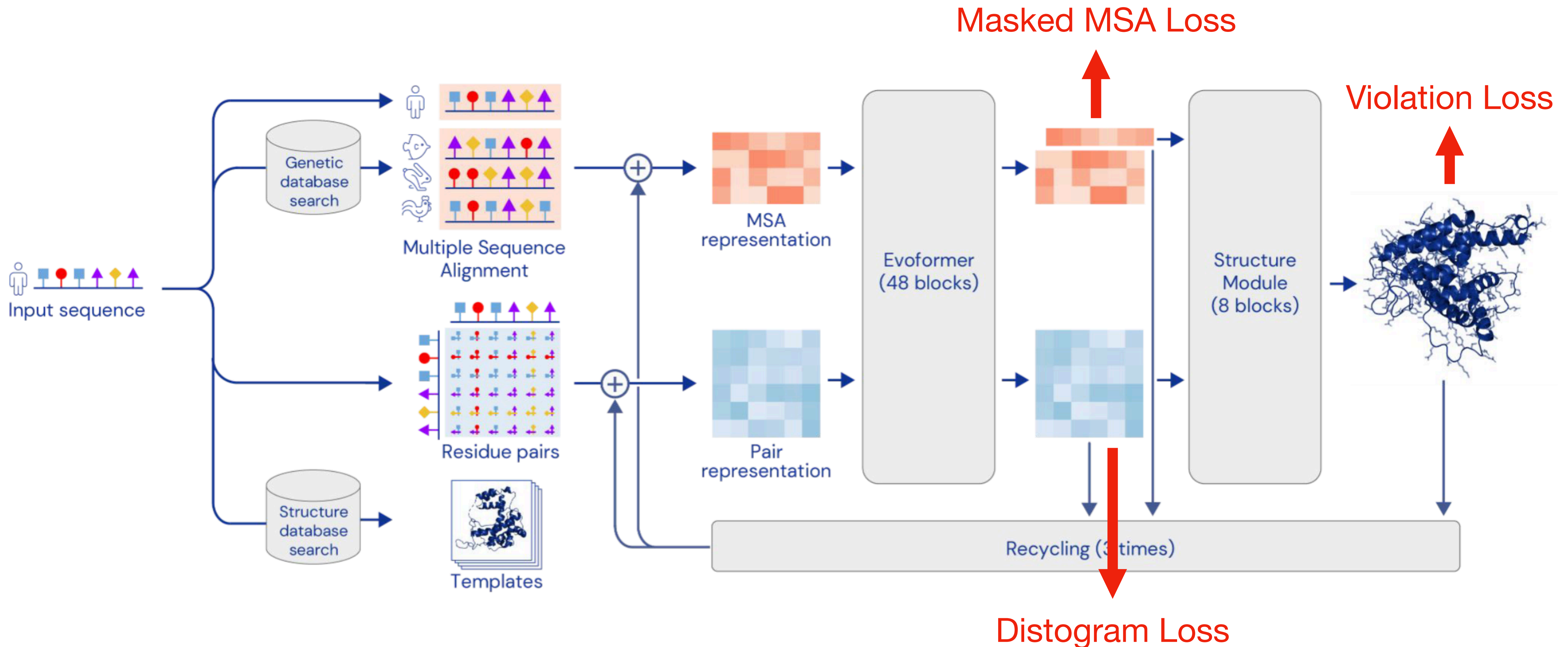
e



f



Auxiliary Losses



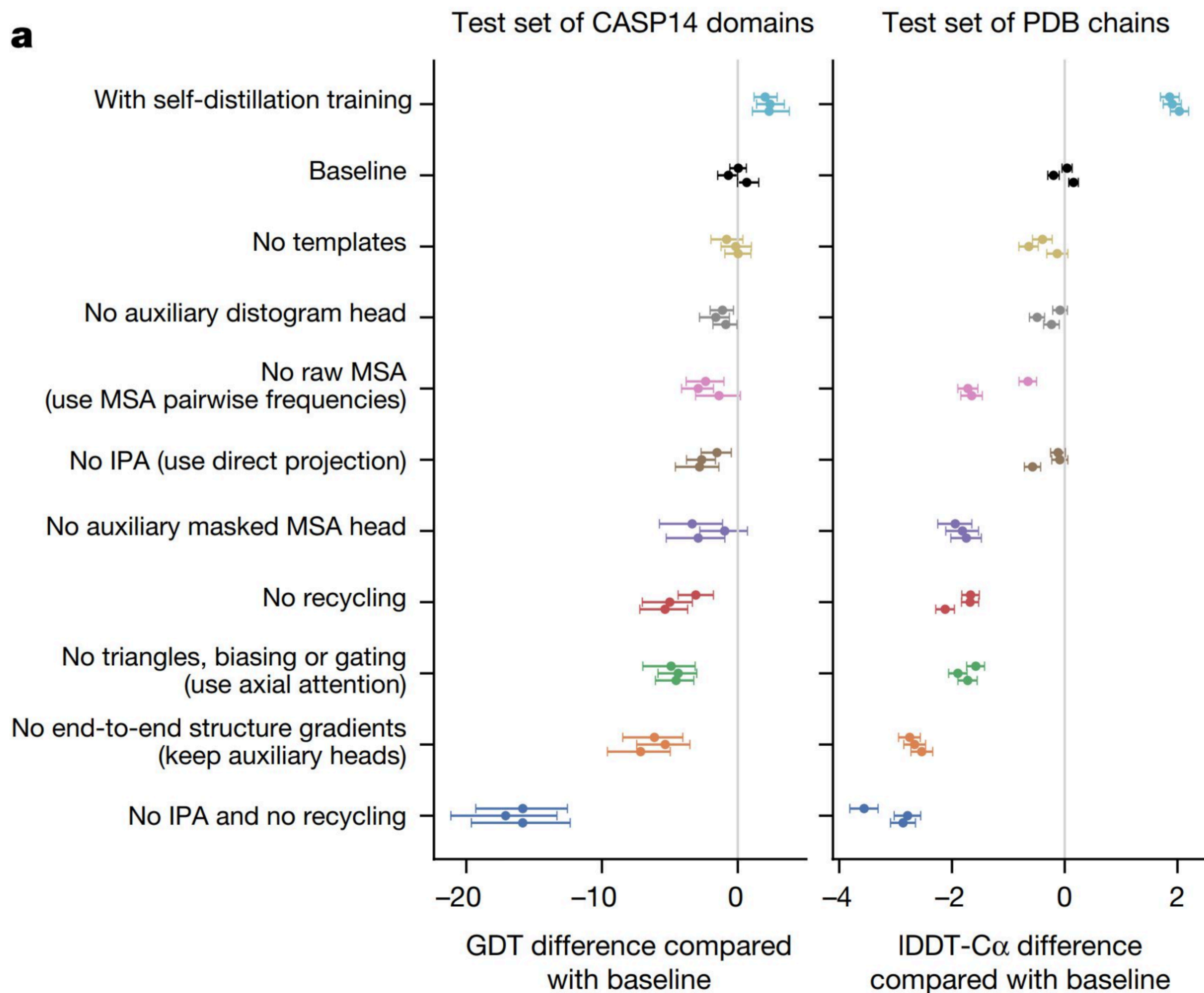
And more ...

Which parts mattered? All of it

No single improvement is dominant

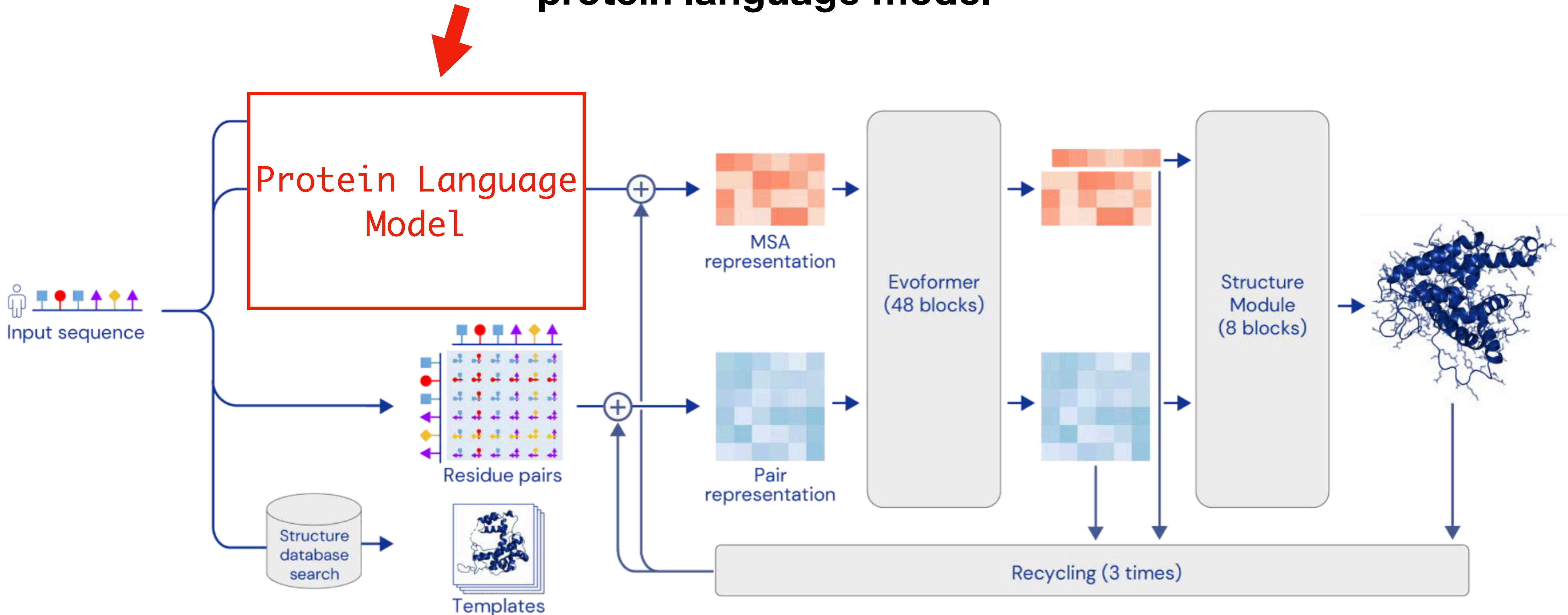
More important was the methodology of building the protein intuition into the model

Multiple ablations suggest strong interactions between many of the components

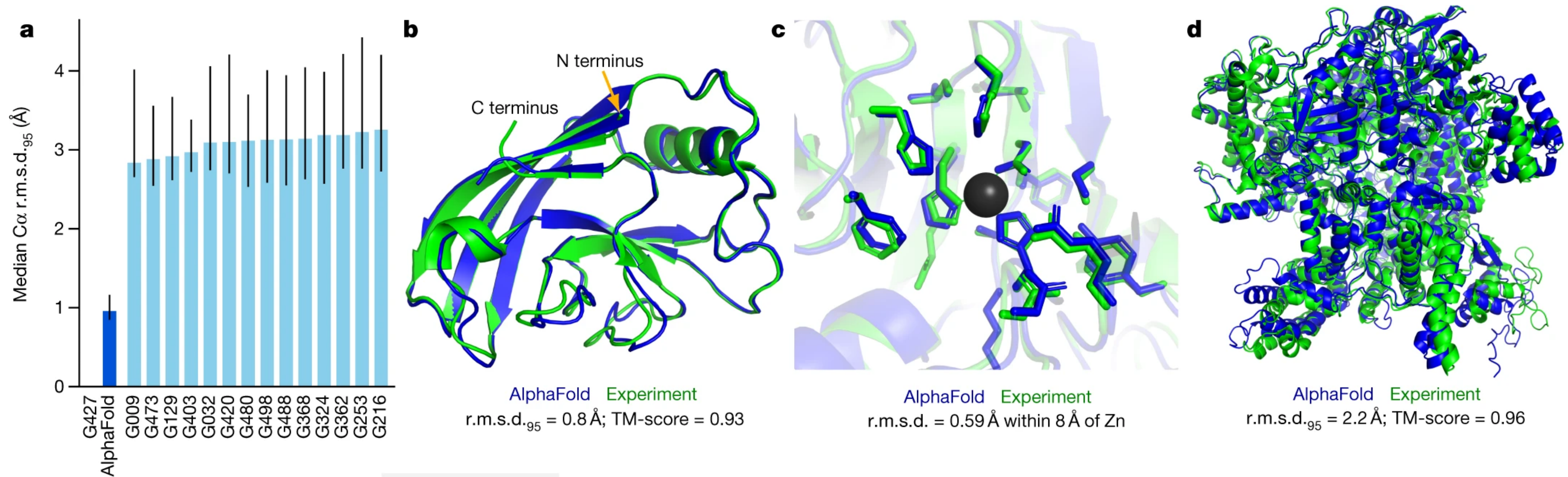


Aside: ESMFold

Same overall design as AlphaFold, but extracts co-evolution info from protein language model



Structure Prediction with Atomic Accuracy



a, The performance of AlphaFold on the CASP14 dataset ($n = 87$ protein domains) relative to the top-15 entries (out of 146 entries), group numbers correspond to the numbers assigned to entrants by CASP. Data are median and the 95% confidence interval of the median, estimated from 10,000 bootstrap samples. **b**, Our prediction of CASP14 target T1049 (PDB 6Y4F, blue) compared with the true (experimental) structure (green). Four residues in the C terminus of the crystal structure are B-factor outliers and are not depicted. **c**, CASP14 target T1056 (PDB 6YJ1). An example of a well-predicted zinc-binding site (AlphaFold has accurate side chains even though it does not explicitly predict the zinc ion). **d**, CASP target T1044 (PDB 6VR4)—a 2,180-residue single chain—was predicted with correct domain packing (the prediction was made after CASP using AlphaFold without intervention).

Structure Trajectories

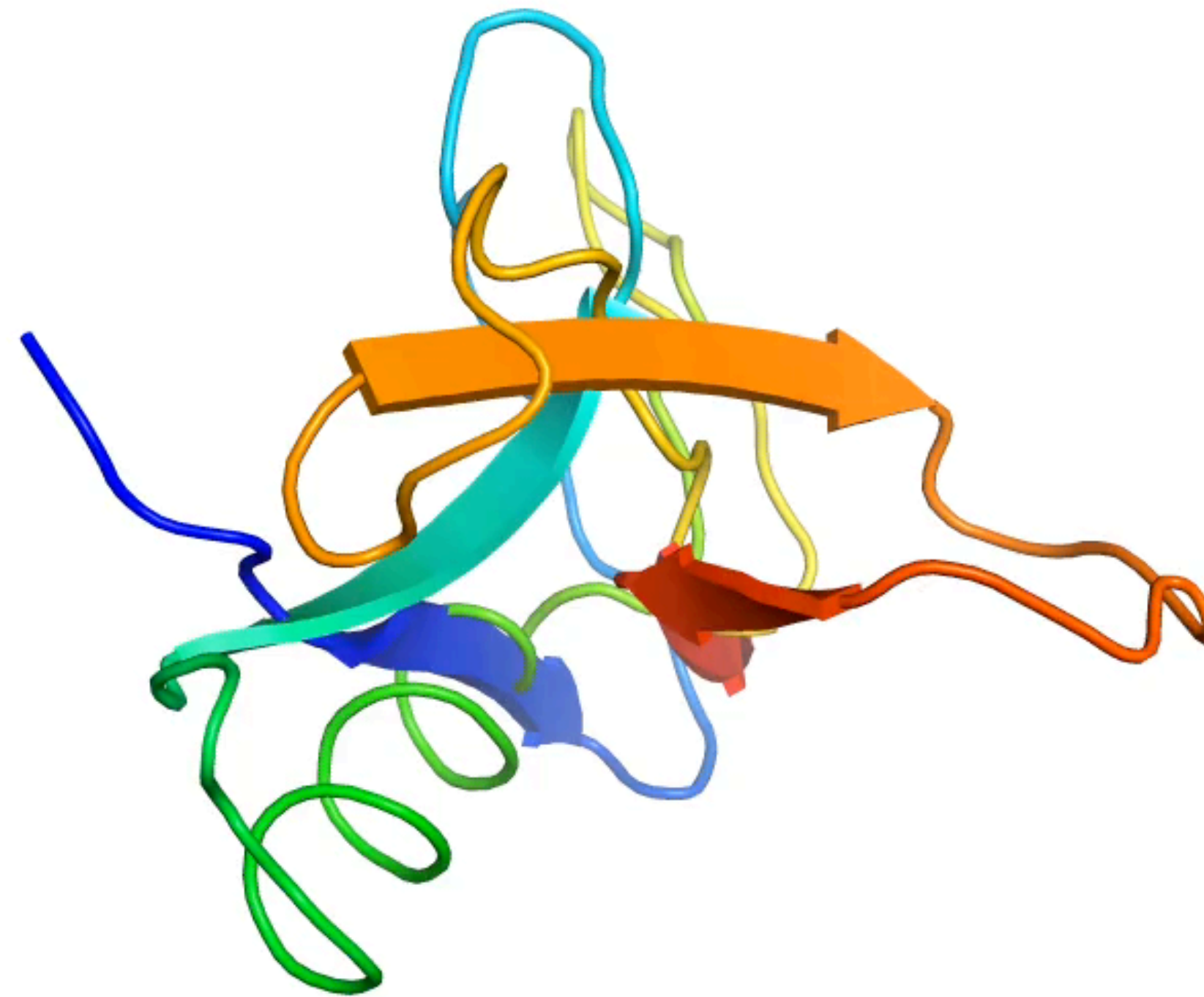
LmrP



Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

Structure Trajectories

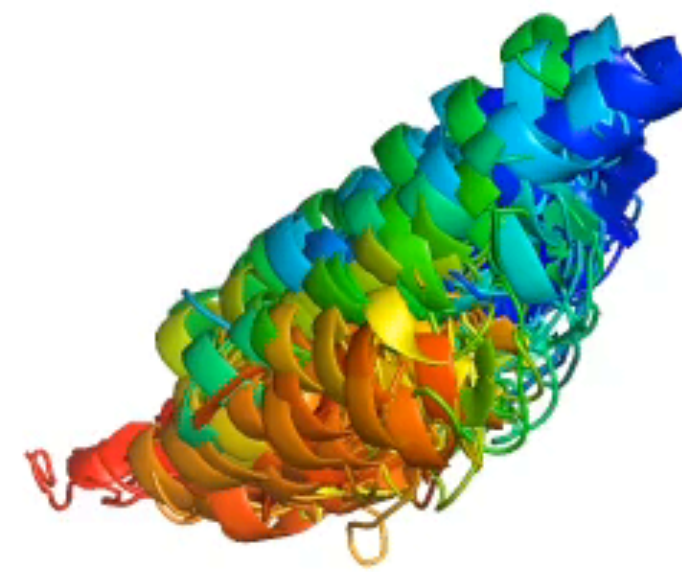
Orf8



Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

Structure Trajectories

RNA Polymerase



Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

Structure Trajectories

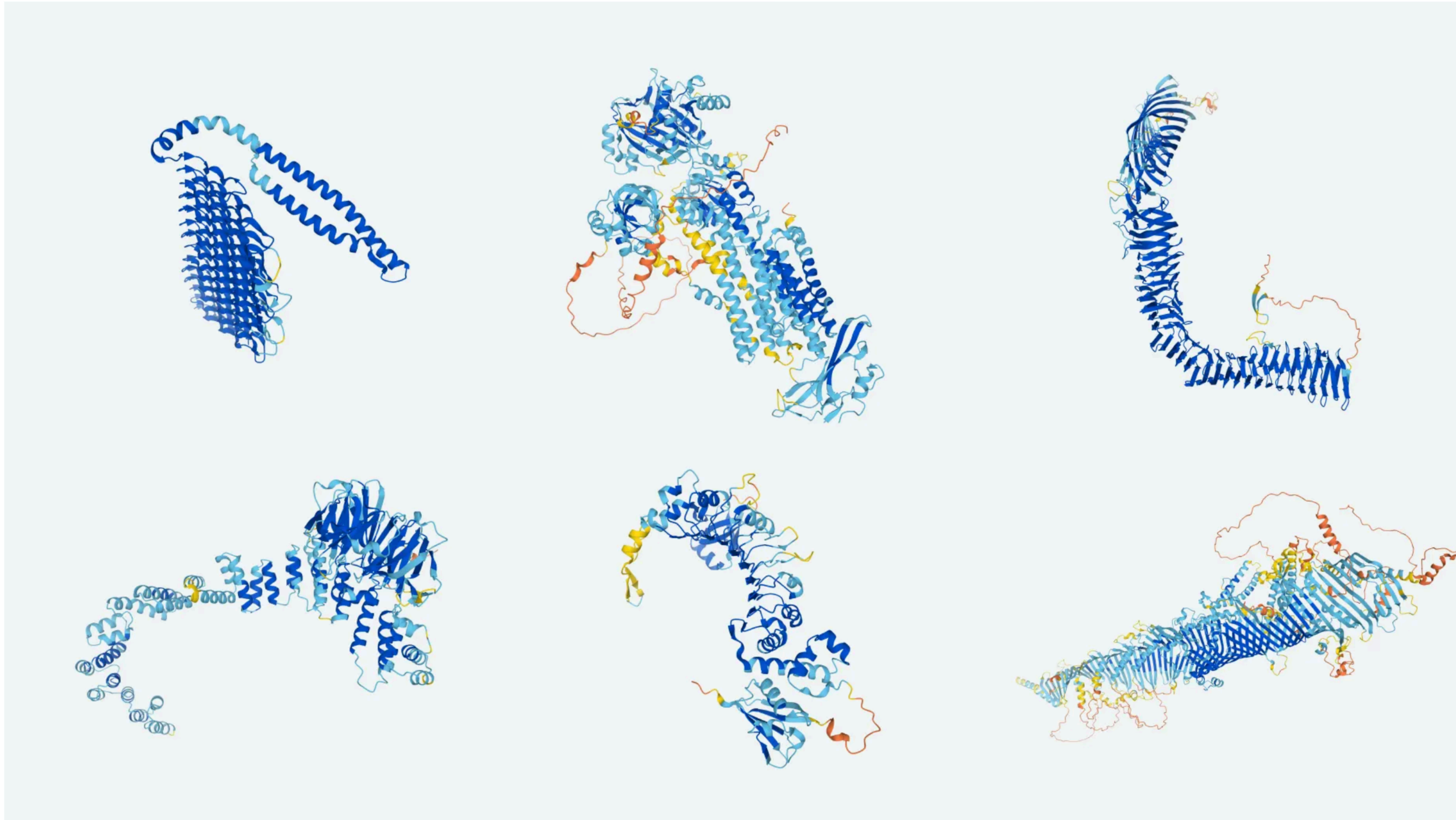
CASP14 Target T1091



Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

A broad liberation of 3D structure

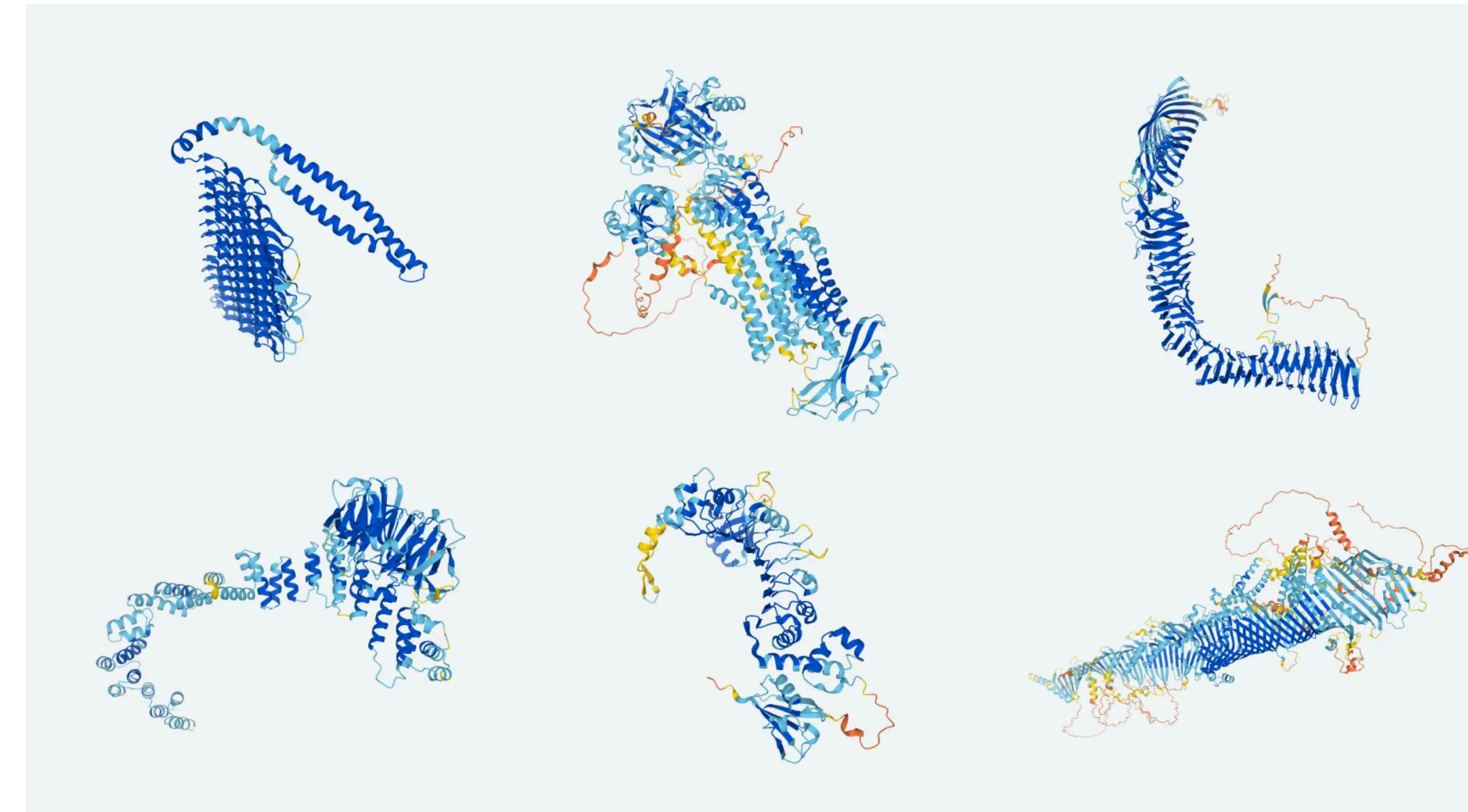
Before: ~100k unique structures — After: >350k predictions today, all ~100M UniProt sequences



A broad liberation of 3D structure

Before: ~100k unique structures — After: >350k predictions today, all ~100M UniProt sequences

- Whole proteome coverage for humans and 20 other model organisms
- Predicted Local Distance Difference Test score (pLDDT) as a well-calibrated measure of confidence
- State of the art predictor of disorder



Tunyasuvunakool et al, 2021



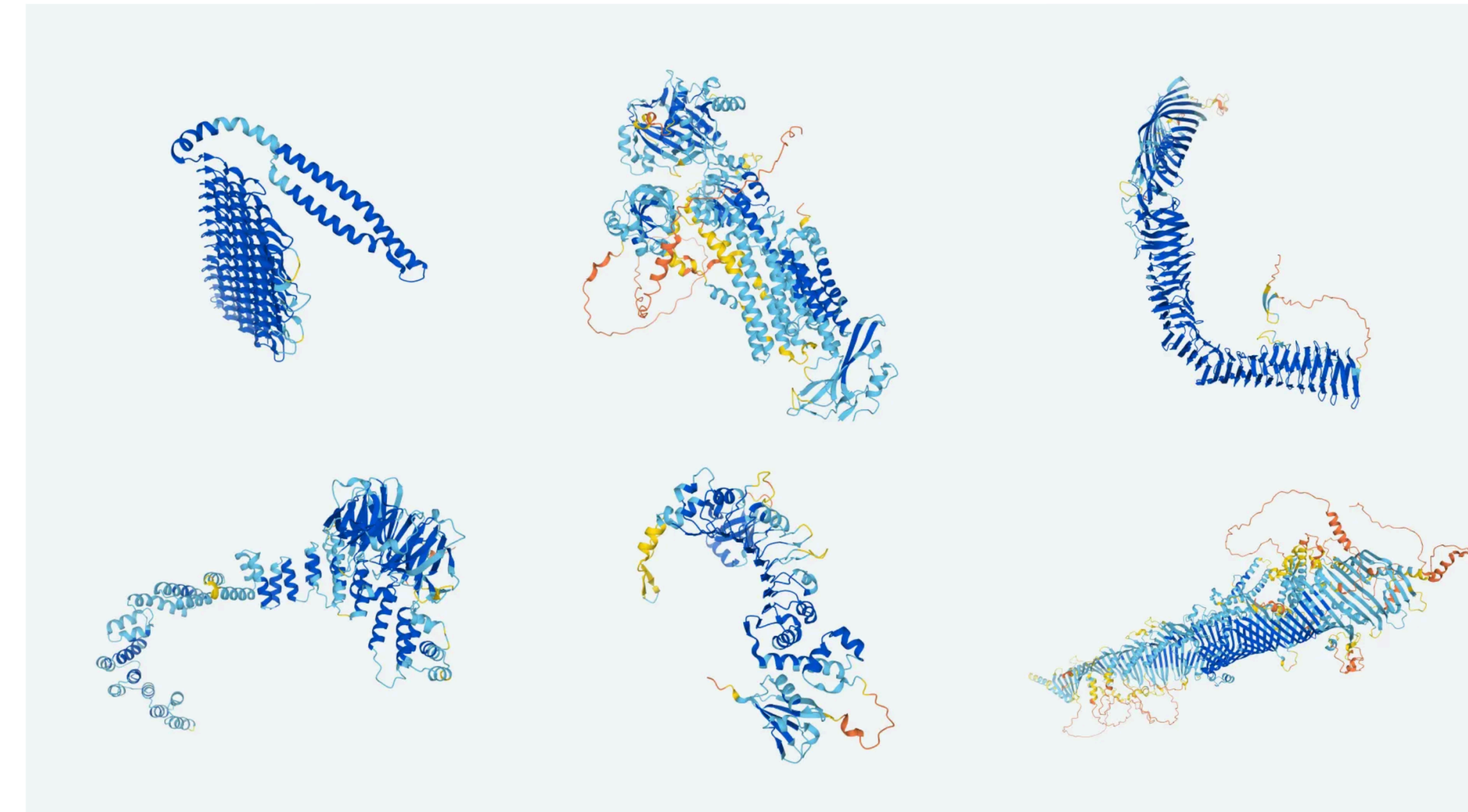
A broad liberation of 3D structure

Before: ~100k unique structures — After: >350k predictions today, all ~100M UniProt sequences

- Whole proteome coverage for humans and 20 other model organisms
- Predicted Local Distance Difference Test score (pLDDT) as a well-calibrated measure of confidence
- State of the art predictor of disorder

Coverage Statistics (pulled from Akdel et al, biorXiv)

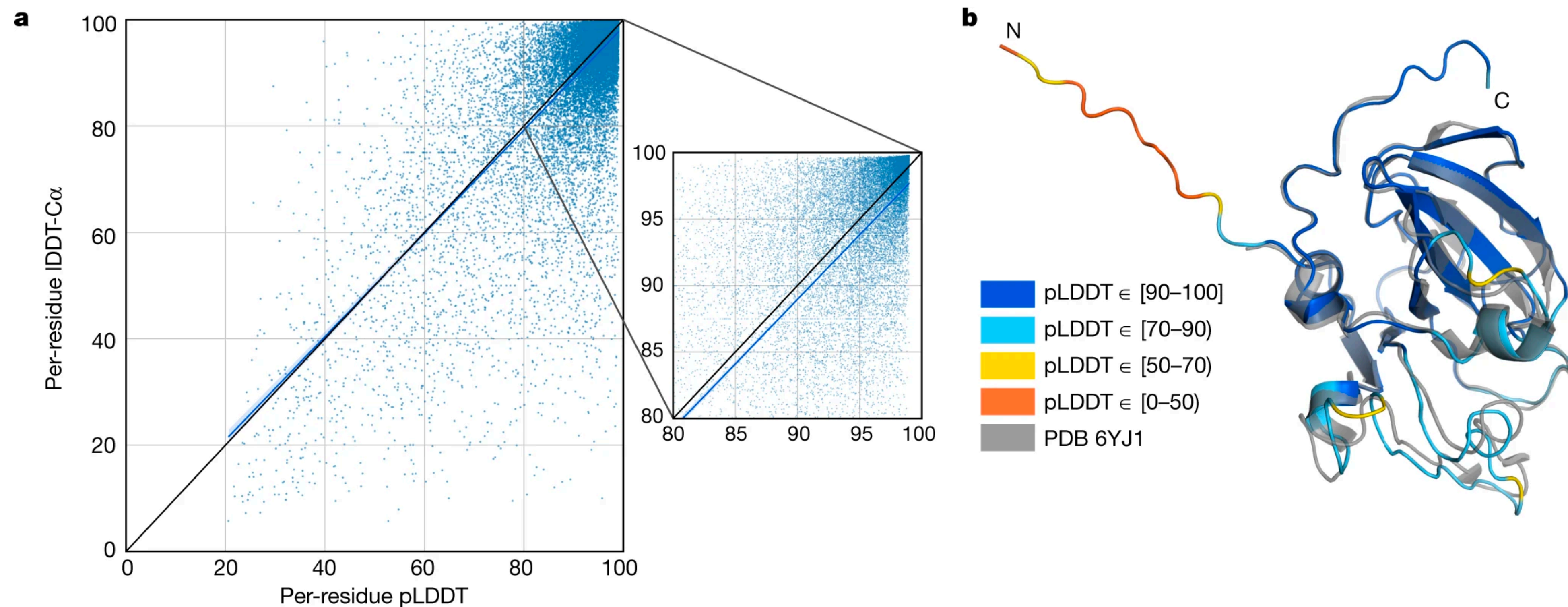
- Confident predictions (pLDDT > 0.7):
 - 27% for *P. falciparum*
 - 77% for *E. coli*
- Highly confident predictions (pLDDT > 0.9) for 25% of all residues
- ~25% of residues of the proteomes covered with novel and confident predictions



Tunyasuvunakool et al, 2021



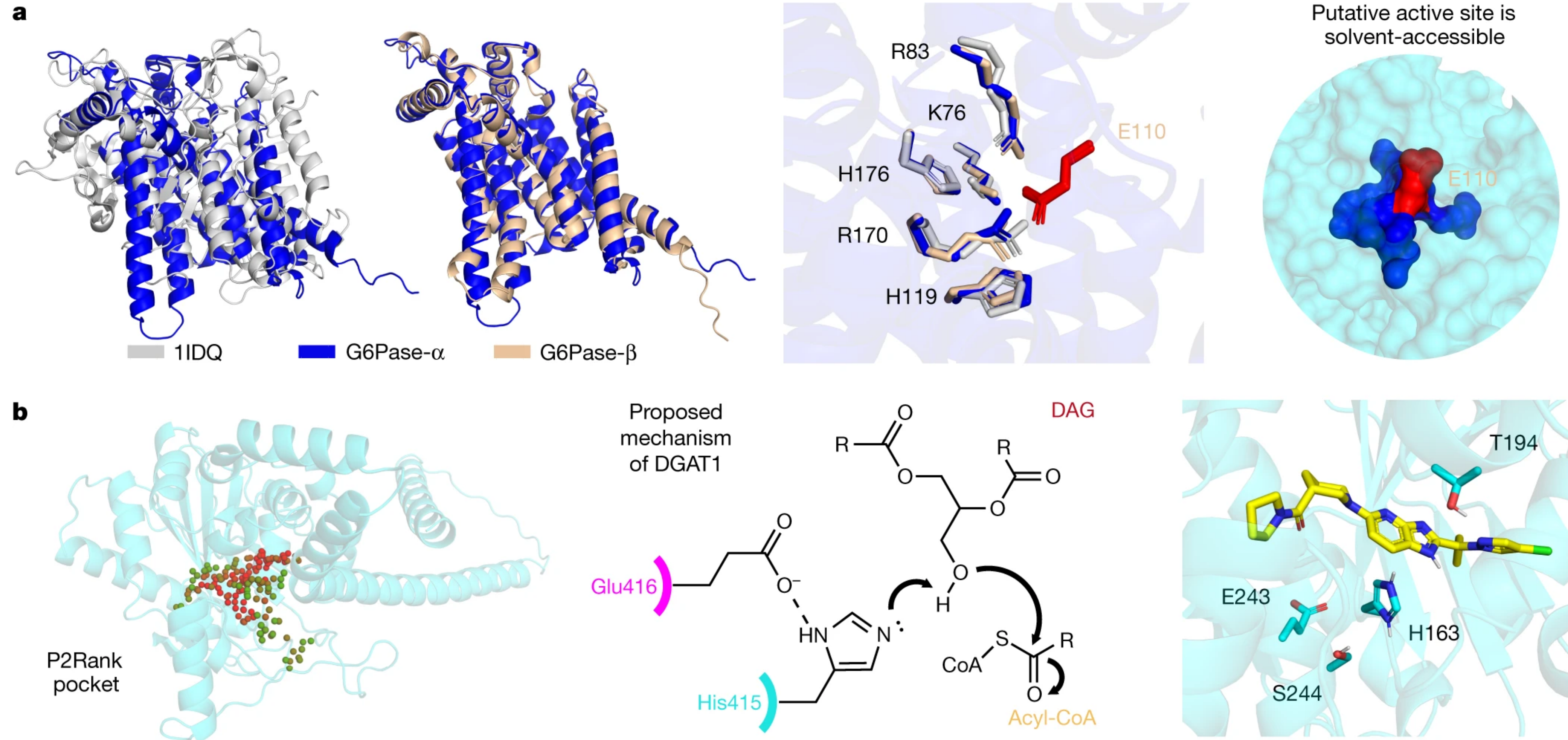
Model Confidence is Well-Calibrated



a, Correlation between per-residue pLDDT and IDDT-C α . Data are based on a held-out set of recent PDB chains (Methods) filtered to those with a reported resolution of <3.5 Å ($n = 10,215$ chains and 2,756,569 residues). The scatterplot shows a subsample (1% of residues), with the blue line showing a least-squares linear fit and the shaded region a 95% confidence interval estimated with 1,000 bootstrap samples. The black line shows $x = y$, for comparison. The smaller plot is a magnified region of the larger one. On the full dataset, the Pearson's $r = 0.73$ and the least-squares linear fit is $y = (0.967 \pm 0.001) \times x + (1.9 \pm 0.1)$. **b**, AlphaFold prediction and experimental structure for a CASP14 target (PDB: 6YJ1)⁶⁴. The prediction is coloured by model confidence band, and the N terminus is an expression tag included in CASP but unresolved in the PDB structure.

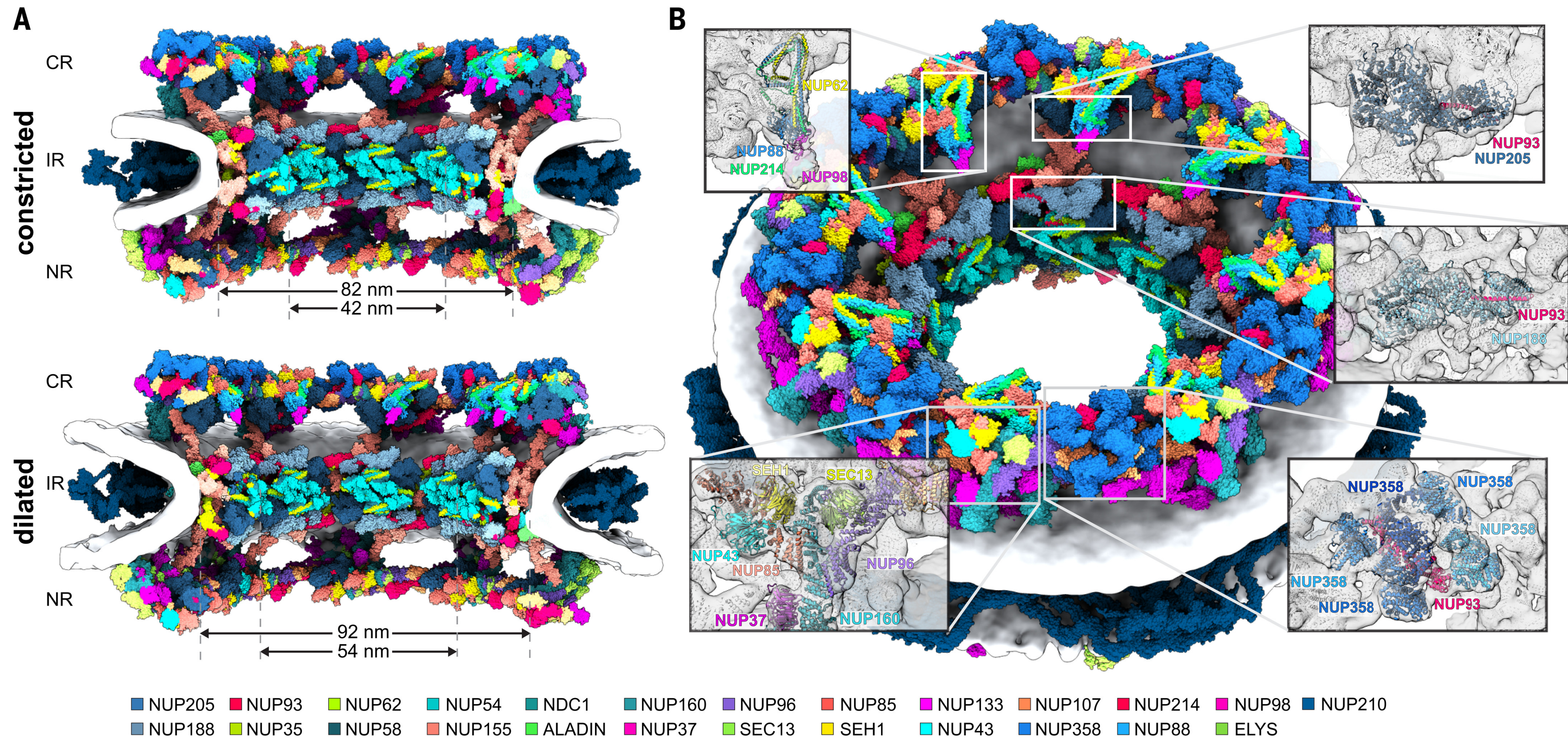
Applications: Novel Mechanistic Hypotheses

Gating Mechanism in G6Pase, Inhibition of DGAT2

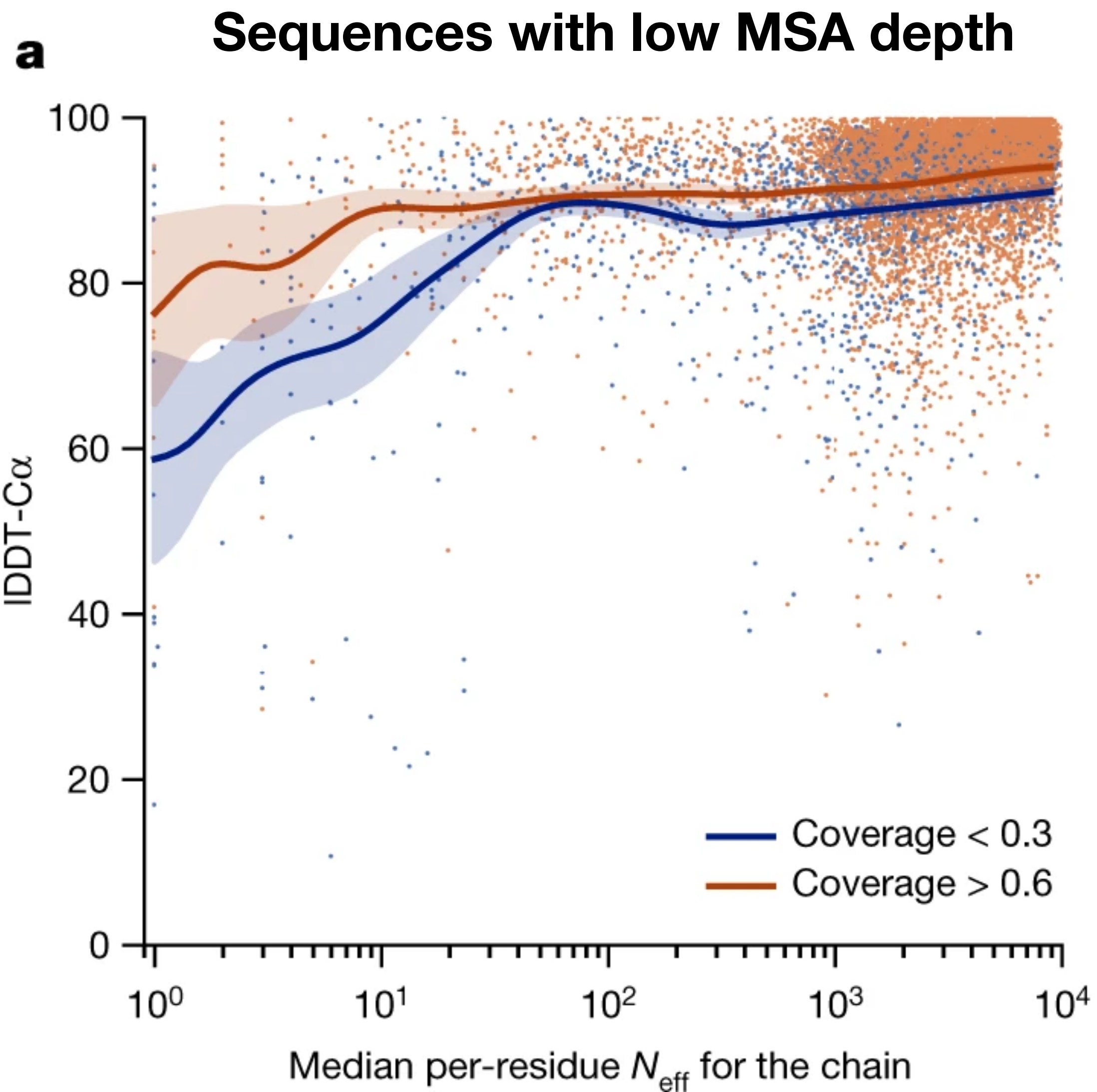


Applications: Integrative Structural Modeling

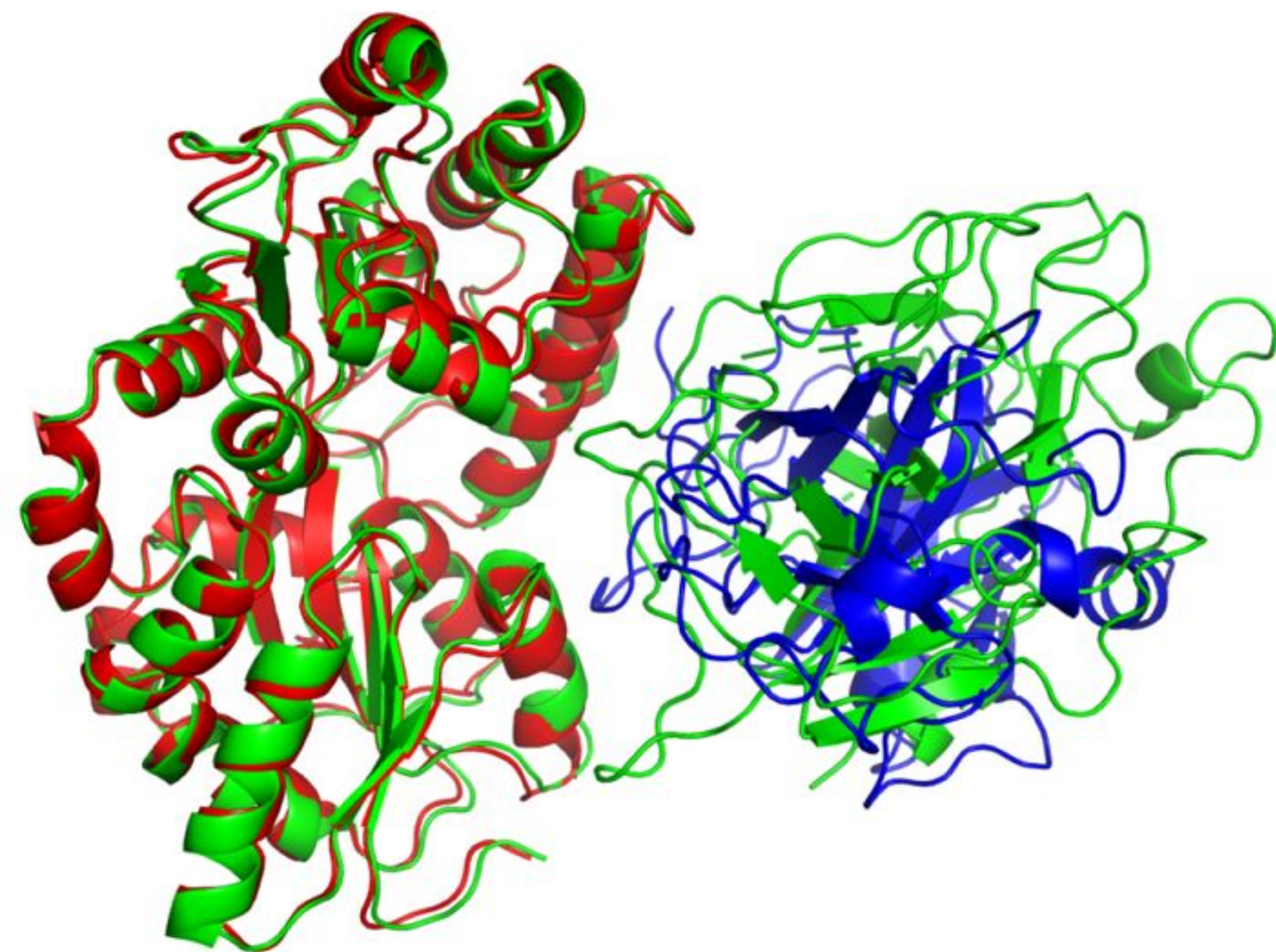
Building ~1000 proteins into the human nuclear pore complex



Limitations



Heteromeric Complexes



TM-score = 66.9, DockQ = 0.035, N_{res} = 628, PDB ID = 6QF7

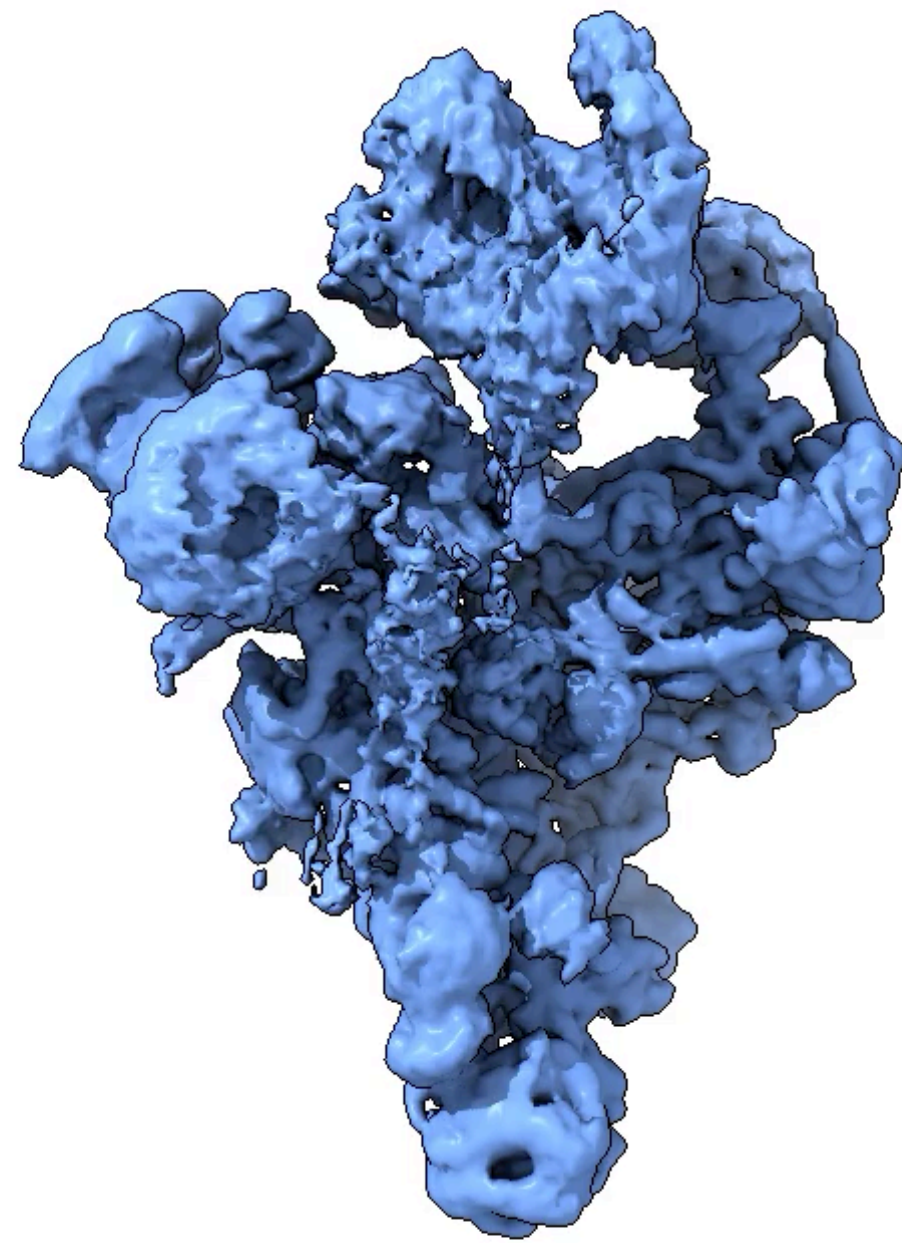
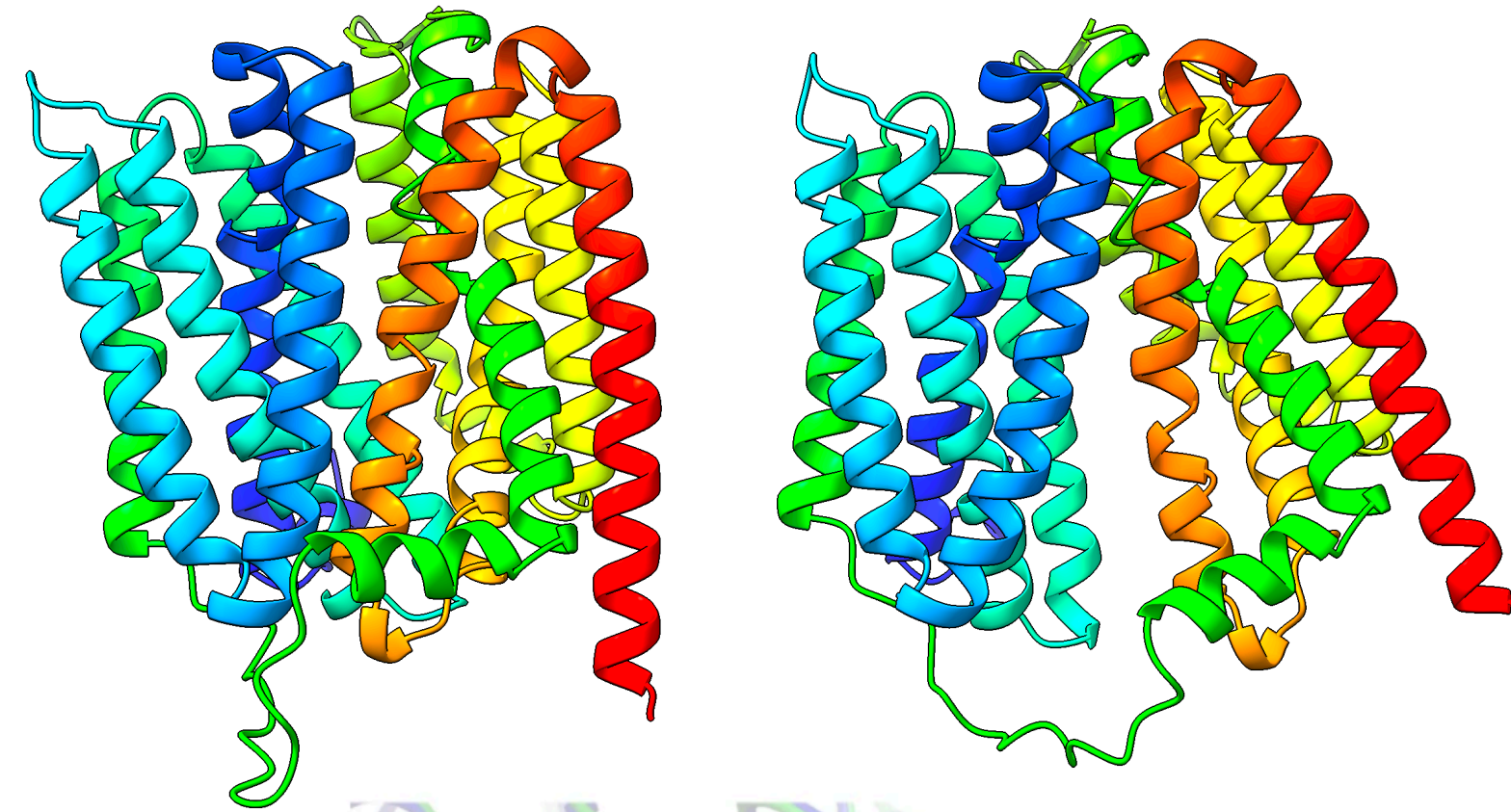
AlphaFold-Multimer, Evans et al. 2021

Summary

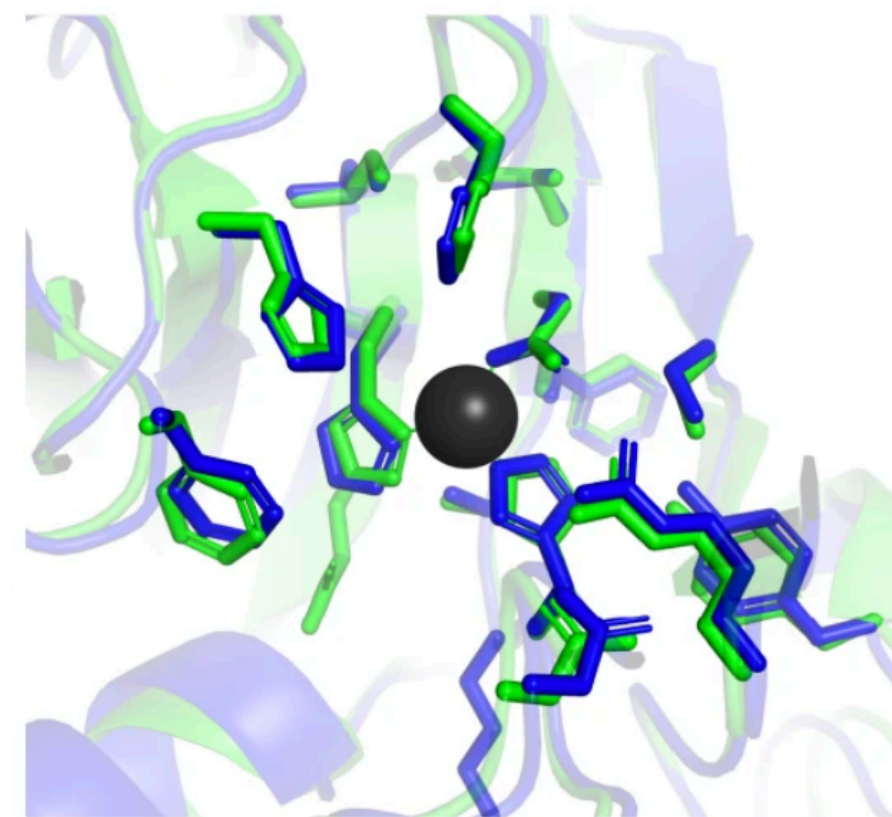
- Impressive performance at CASP14
- New architectures
 - A combination of bioinformatics and physical approaches
- Solves a long-standing problem
 - “A seismic and unprecedented shift so profound it literally turns a field upside down overnight”
- Provides a database of predicted structures (AFDB)

Frontiers for MLSB: Complexes, dynamics, interactions

Input sequence
MRKPRTPTFT...



Zhong et al, Nature Methods 2021

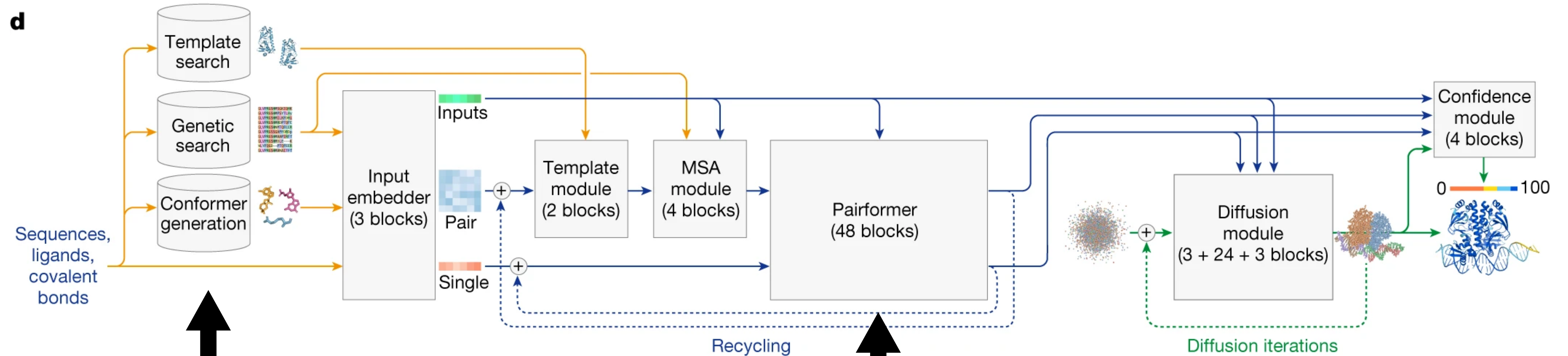


AlphaFold Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn



Illustration of the E. coli bacterium. David S Goodsell.

AlphaFold3: Biomolecular Complexes

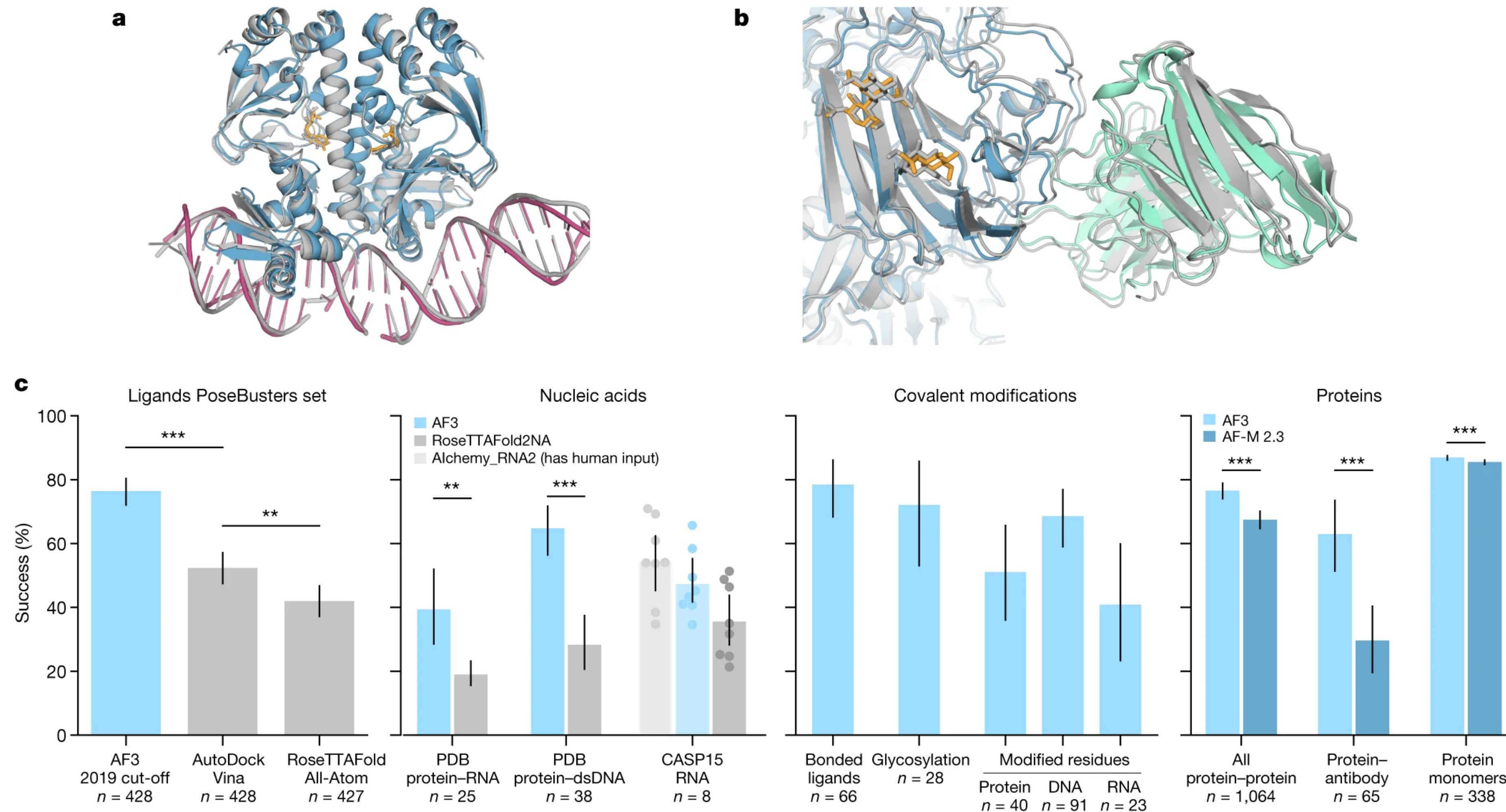


Other bimolecular inputs:
Ligands, nucleic acids,
covalent modifications

Simplified architecture
for MSA and pair
representation updates

Diffusion model
replaces IPA as
structure module

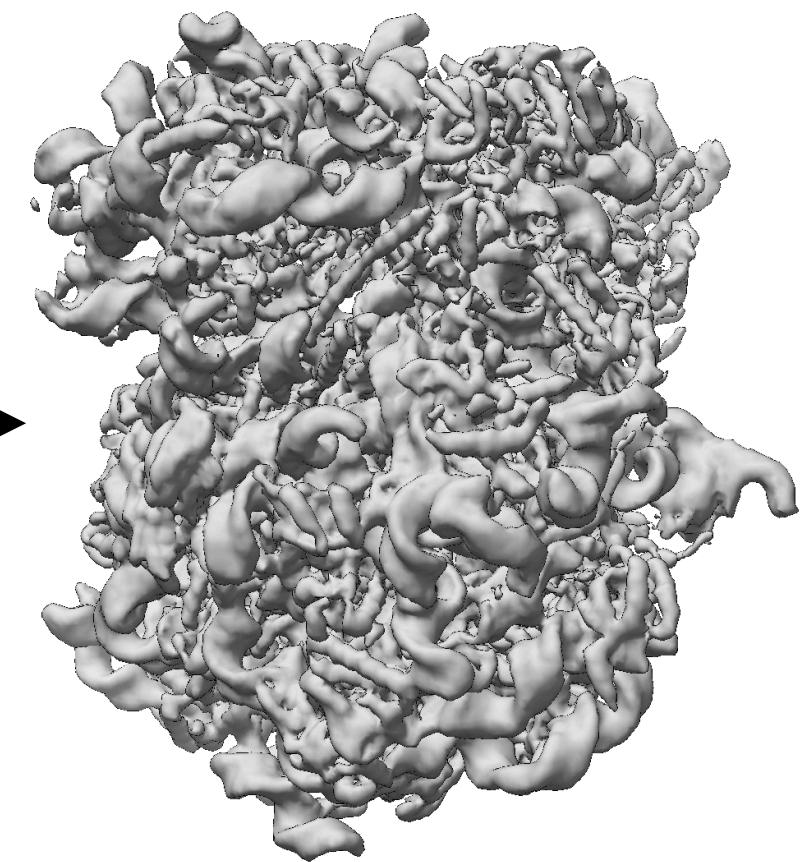
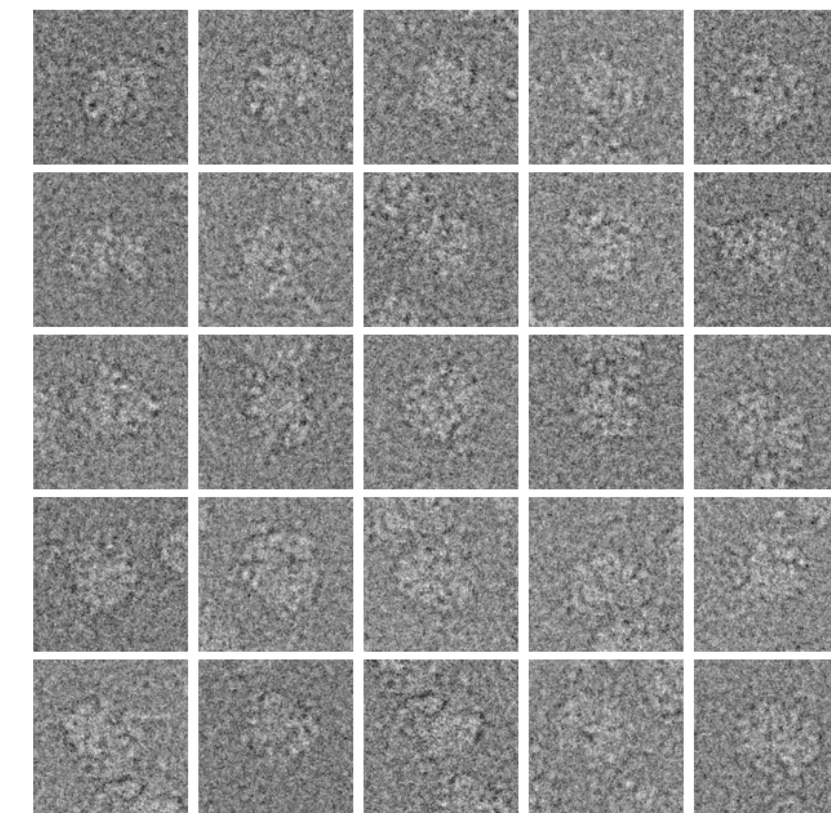
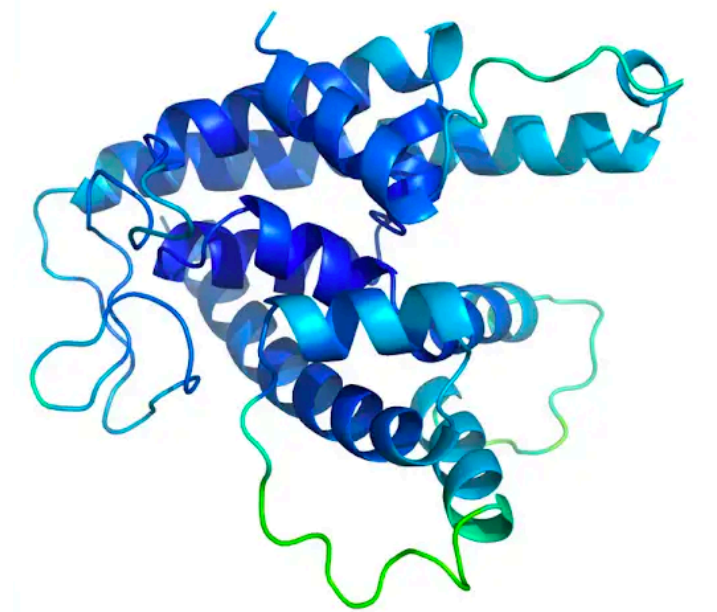
AlphaFold3: Biomolecular Complexes



Outlook for the post-AlphaFold era

- “Discovery” of performant neural network architectures for reasoning over protein sequences and structures

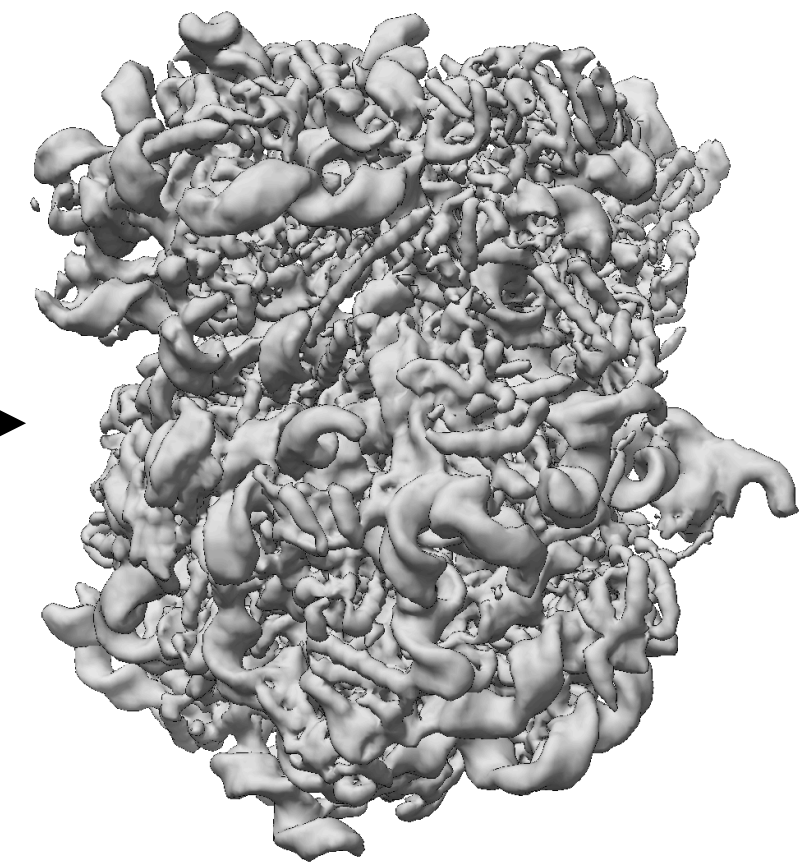
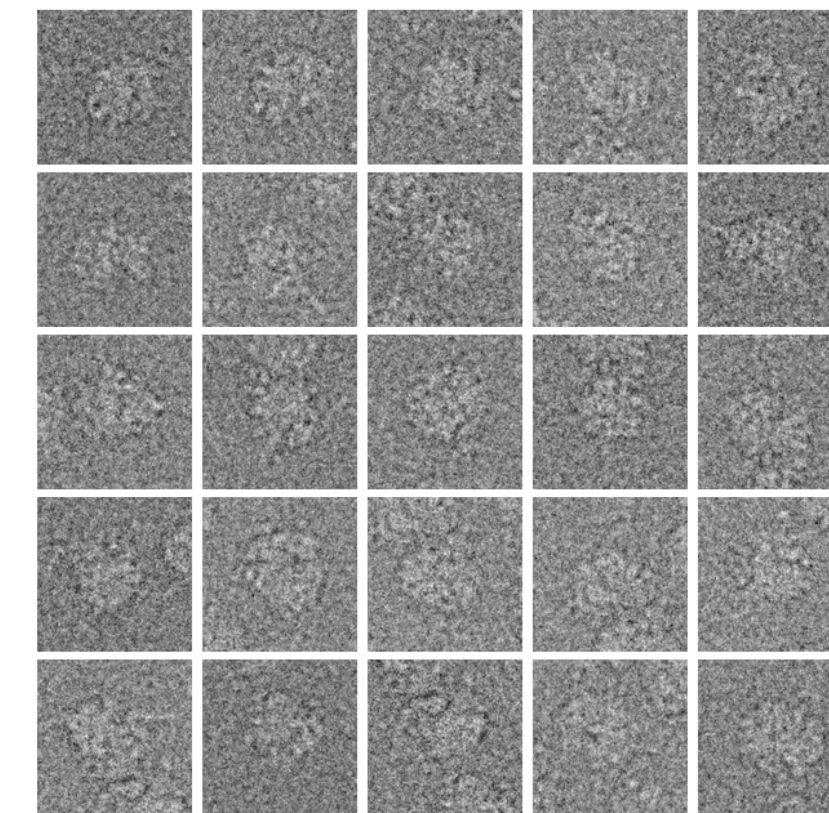
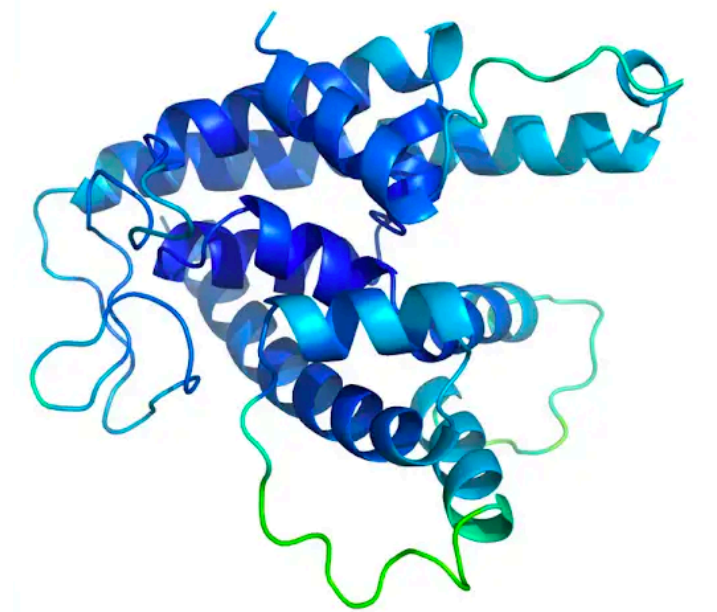
MRKPRTPTFTT...



Outlook for the post-AlphaFold era

- “Discovery” of performant neural network architectures for reasoning over protein sequences and structures
- Do we need additional data sources?
 - Multiple conformations and dynamics
 - Protein design
 - Interactions with DNA/RNA/small molecules

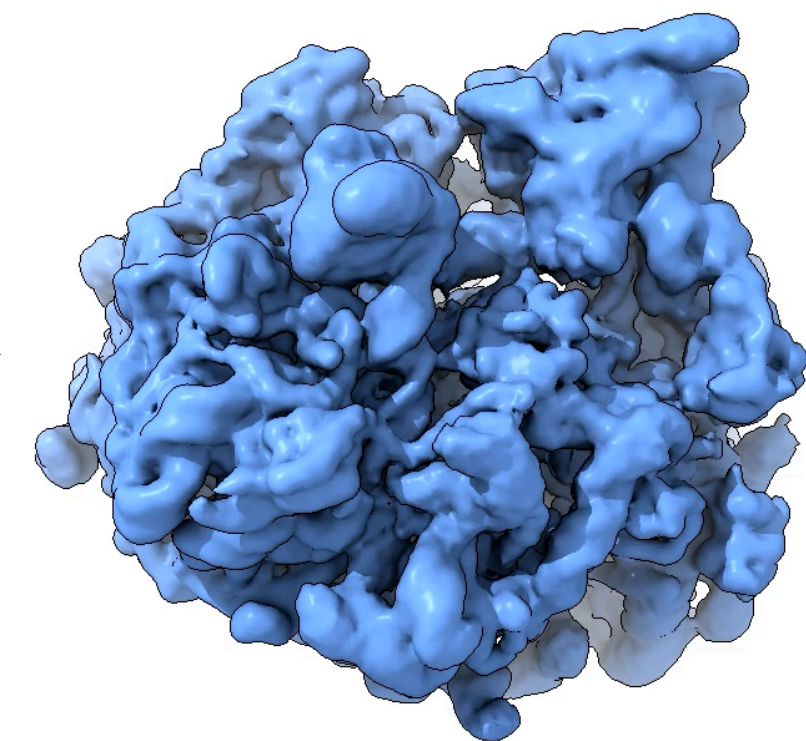
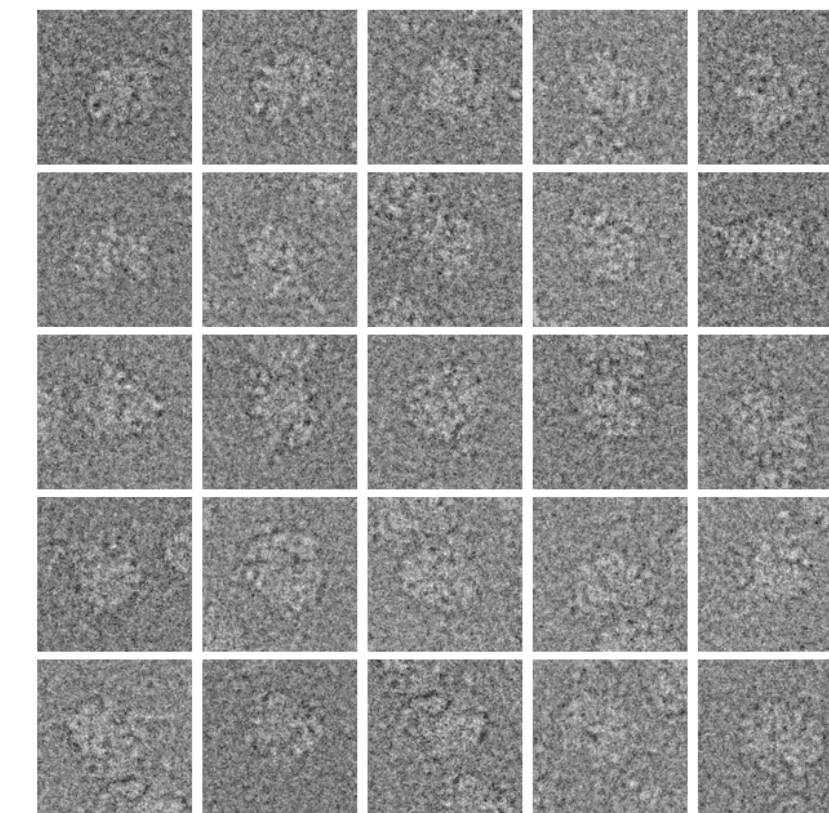
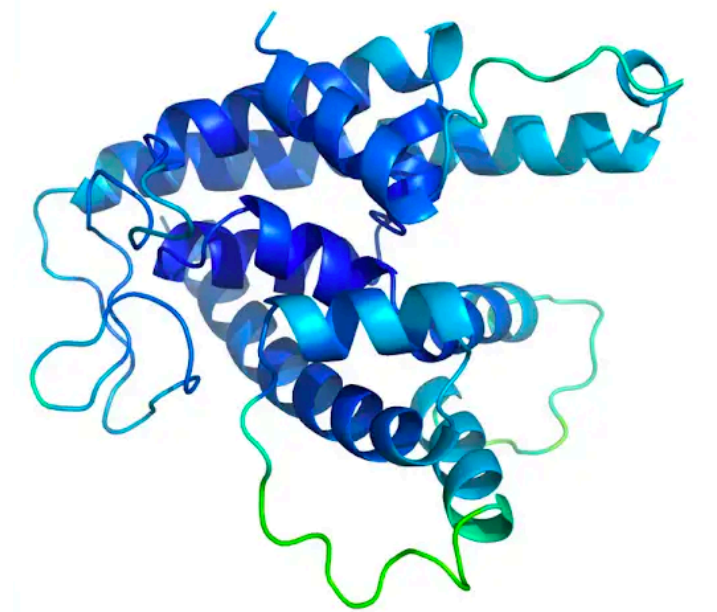
MRKPRTPTFTT...



Outlook for the post-AlphaFold era

- “Discovery” of performant neural network architectures for reasoning over protein sequences and structures
- Do we need additional data sources?
 - Multiple conformations and dynamics
 - Protein design
 - Interactions with DNA/RNA/small molecules
- Protein structure prediction vs. protein structure determination
 - Deliberate experimental design for machine learning systems
 - New paradigms for cryo-EM imaging and processing?

MRKPRTPTTT...



Pre-lecture questions

- Q1: What is pLDDT and why is it a useful metric?
- Q2: Define SE(3) equivariance. Describe one aspect of AlphaFold2 that is SE(3) equivariant.
- Q3: How important is self-distillation in training AlphaFold2? Why do you think distillation works?
- Q4: Describe one difference between the Evoformer architecture and a standard transformer
- Q5: What's next after Alphafold?

Next Week

- Generative modeling 101
- Required reading:
 - VAE paper
 - CryoDRGN ICLR paper
- Looking for presenters

