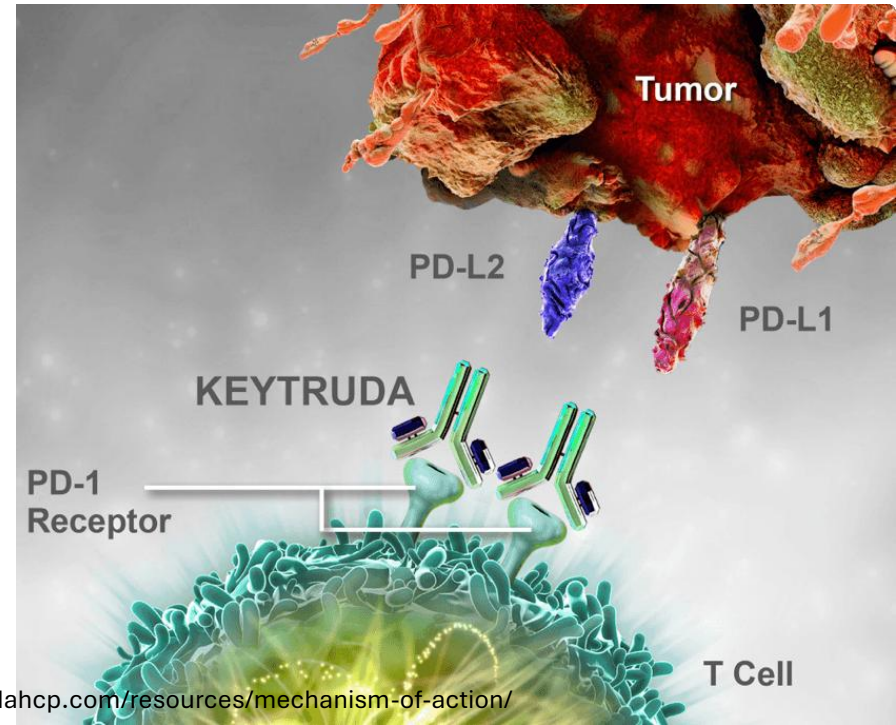
Accurate structure prediction of biomolecular interactions with AlphaFold 3

Abramson et al. *Nature,* April 29 2024

Presented by Maxwell Soh

Motivation

- AF2 gives us the structure of a single protein chain
- To go from structure to function we need to model how proteins interact with other biomolecules
- Almost all drugs bind to a protein using a:
  - Small molecule/ligand – e.g. aspirin, penicillin
  - Protein
    - Antibodies – e.g. Keytruda
    - Shorter peptides – e.g. GLP-1 receptor agonists



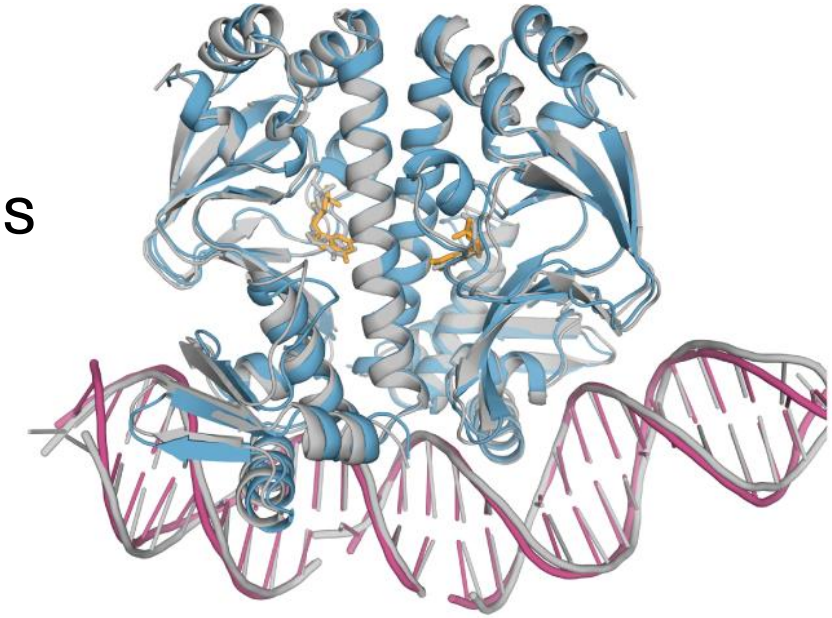
https://www.keytrudahcp.com/resources/mechanism-of-action/

Prior methods

- Specialist methods for one type of biomolecular interaction
  - AutoDock Vina (protein-ligand)
  - RoseTTAFold2NA (protein-nucleic, RNA structures)
  - AlphaFold-Multimer (protein interactions)
- RoseTTAFold All-Atom – can generalize to all types of interactions but worse performance than specialist methods

Challenges

- Prior methods are not accurate enough for many applications
- Specialist methods don’t work if there are multiple types of interactions in a complex
- Less public data for biomolecular complexes
  - A lot of private protein-ligand data
  - But AF3 was only trained on public data



AF3 Inputs and Output

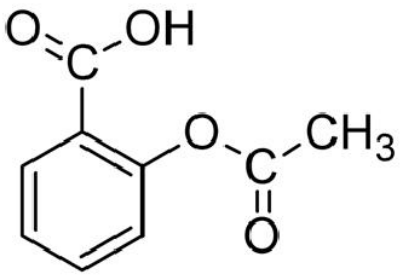
- Amino acid sequence (protein)
- Nucleotide sequence (DNA, RNA)
- SMILES (ligands)
  - String denoting a ligand's atoms and bonds
  - Stands for: simplified molecular-input line-entry system
- Output: 3D position of each heavy atom

Aspirin

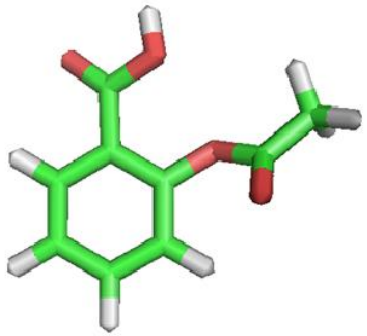
A

CC(=O)Oc1ccccc1C(=O)O

B



C



Polski et al., *Pharma Ed* 2015

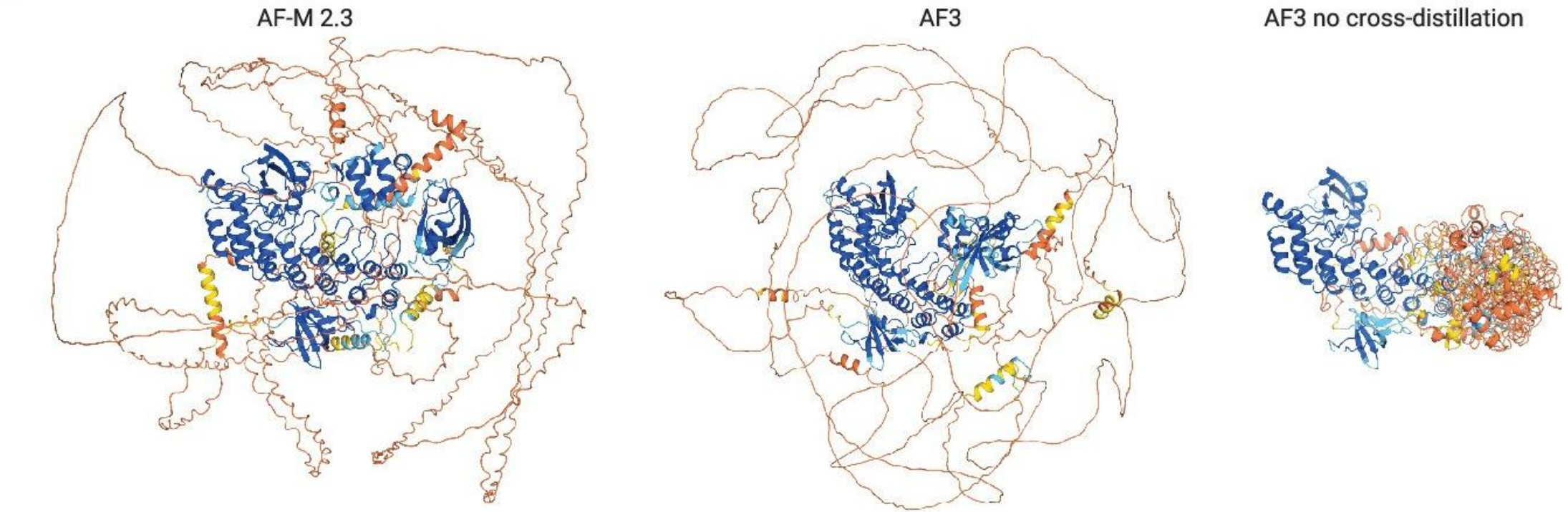
Training data

- All GT data was from the Protein Data Bank
- PDB data is weighted to sample more uncommon proteins and types of complexes

Name	Description	Sampl. strategy	Weight
Weighted PDB	Ground truth PDB structures	weighted	0.5
Disordered protein PDB distillation	Proteins with unresolved residues	weighted	0.02
Protein monomer distillation	Protein monomer predictions from MGnify	uniform	0.495
Short protein monomer distillation	Protein short monomer predictions from MGnify	uniform	0.005
RNA distillation	RNA monomer predictions from Rfam	uniform	0.05
Transcription factor negatives	MGnify protein + random DNA	uniform	0.01[1]
Transcription factor positives	DNA+protein predictions from JASPAR	uniform	0.02[1]

Disordered protein distillation of AF-Multimer

- 25K predicted PDB proteins with >40 unresolved residues in GT
- Reduces structure hallucination in intrinsically disordered regions



Protein distillation from AF2

- AF3 training samples from the AF2 generated structures of 41 million protein sequences from the MGnify database
- Short proteins are 4-200 residues

Name	Description	Sampl. strategy	Weight
Weighted PDB	Ground truth PDB structures	weighted	0.5
Disordered protein PDB distillation	Proteins with unresolved residues	weighted	0.02
Protein monomer distillation	Protein monomer predictions from MGnify	uniform	0.495
Short protein monomer distillation	Protein short monomer predictions from MGnify	uniform	0.005
RNA distillation	RNA monomer predictions from Rfam	uniform	0.05
Transcription factor negatives	MGnify protein + random DNA	uniform	0.01[1]
Transcription factor positives	DNA+protein predictions from JASPAR	uniform	0.02[1]

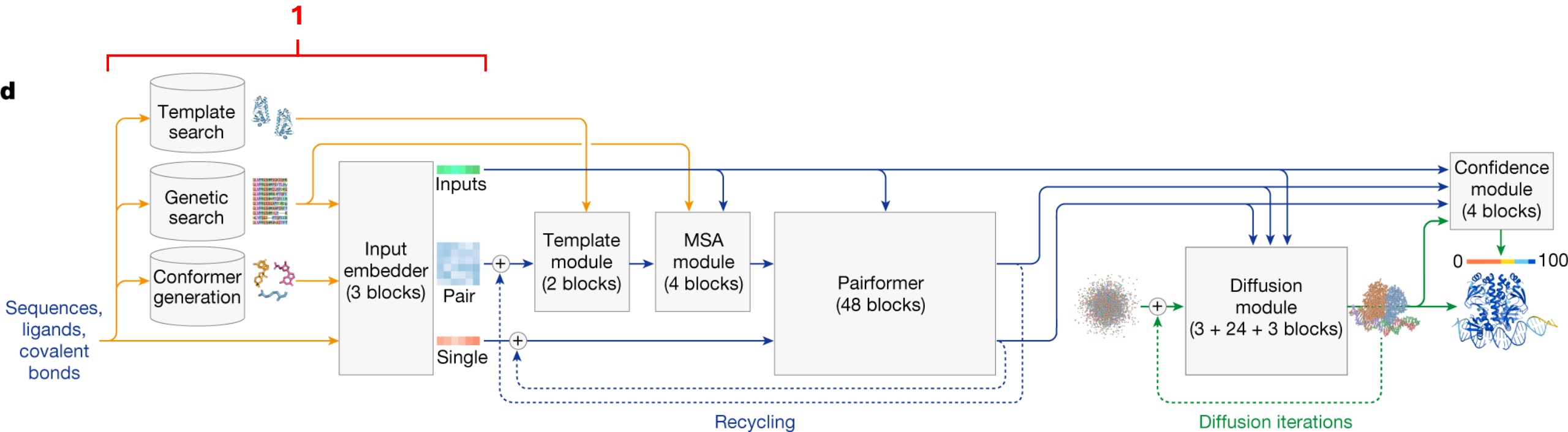
AF3 self-distillation

- Like with AF2 self-distillation, only retain the best structures
- MGnify protein is sampled from the AF2 distilled set and placed away from the AF3 generated DNA structure

Name	Description	Sampl. strategy	Weight
Weighted PDB	Ground truth PDB structures	weighted	0.5
Disordered protein PDB distillation	Proteins with unresolved residues	weighted	0.02
Protein monomer distillation	Protein monomer predictions from MGnify	uniform	0.495
Short protein monomer distillation	Protein short monomer predictions from MGnify	uniform	0.005
RNA distillation	RNA monomer predictions from Rfam	uniform	0.05
Transcription factor negatives	MGnify protein + random DNA	uniform	0.01[1]
Transcription factor positives	DNA+protein predictions from JASPAR	uniform	0.02[1]

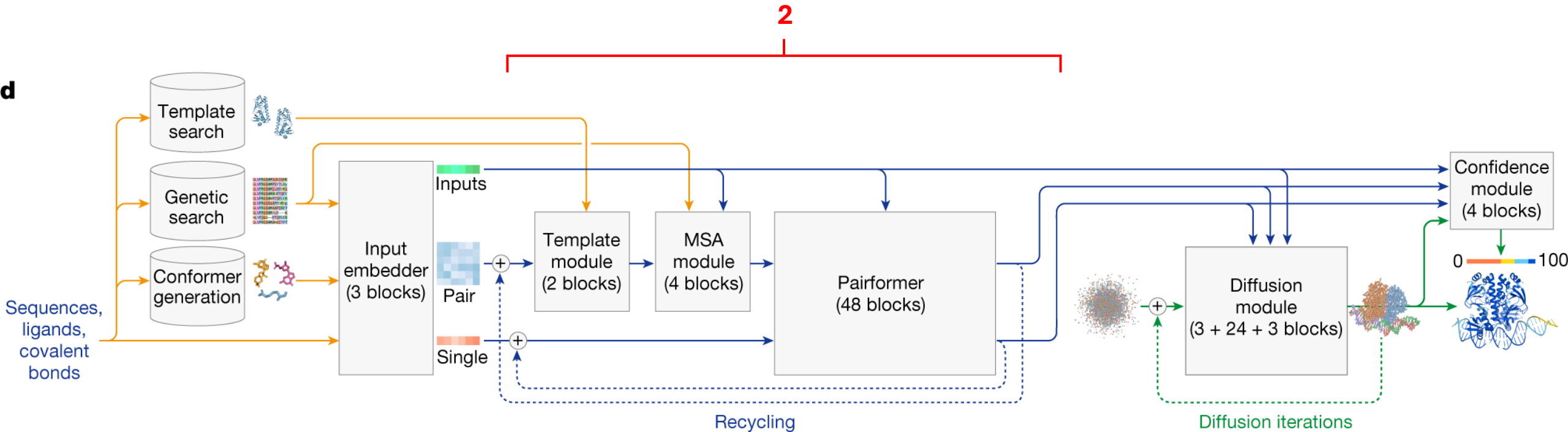
AF3 Method Overview

1. Embed inputs



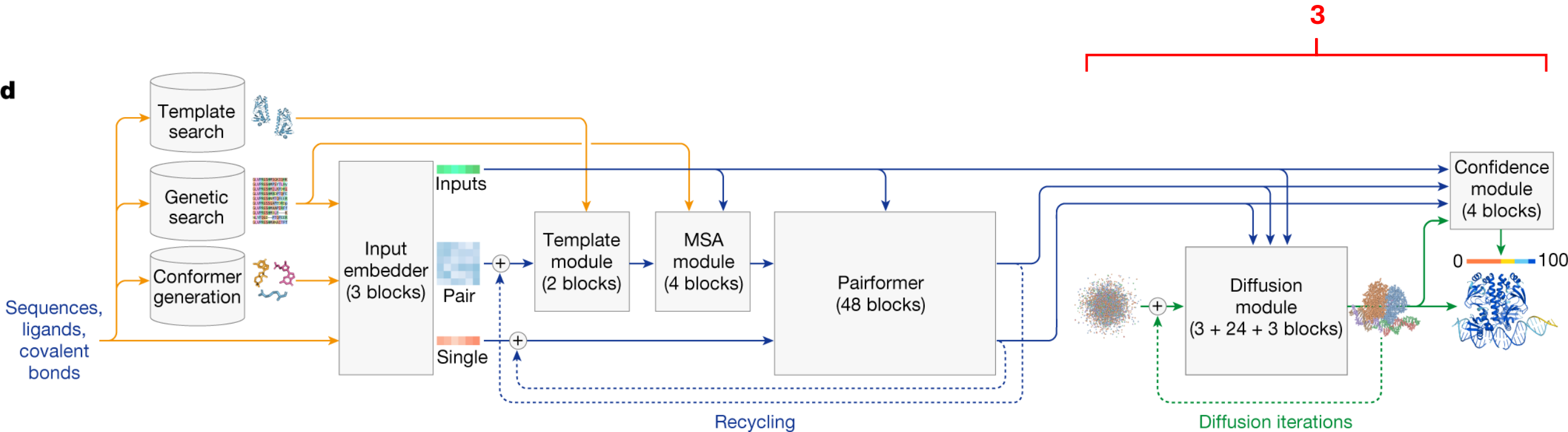
AF3 Method Overview

1. Embed inputs
2. Process embeddings



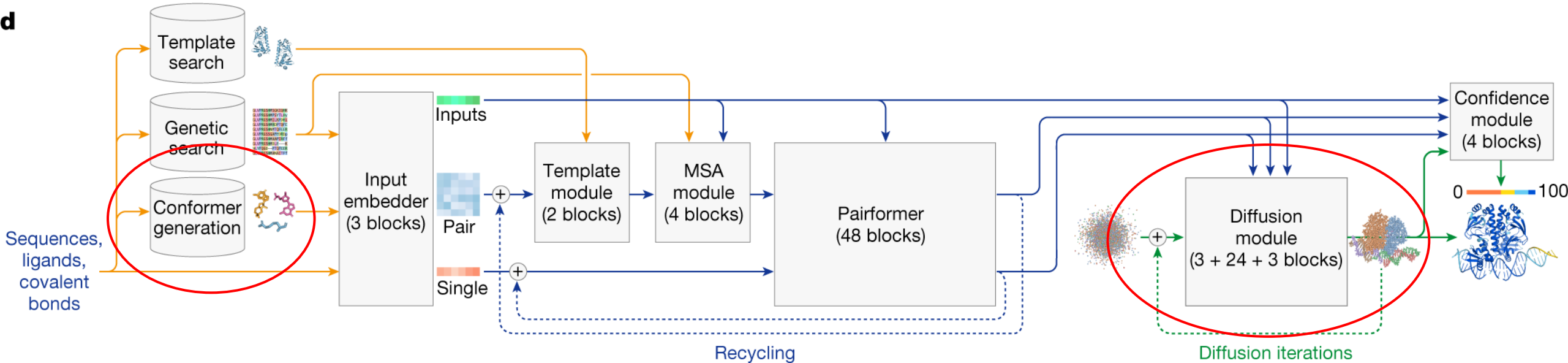
AF3 Method Overview

1. Embed inputs
2. Process embeddings
3. Generate atom positions based on embeddings



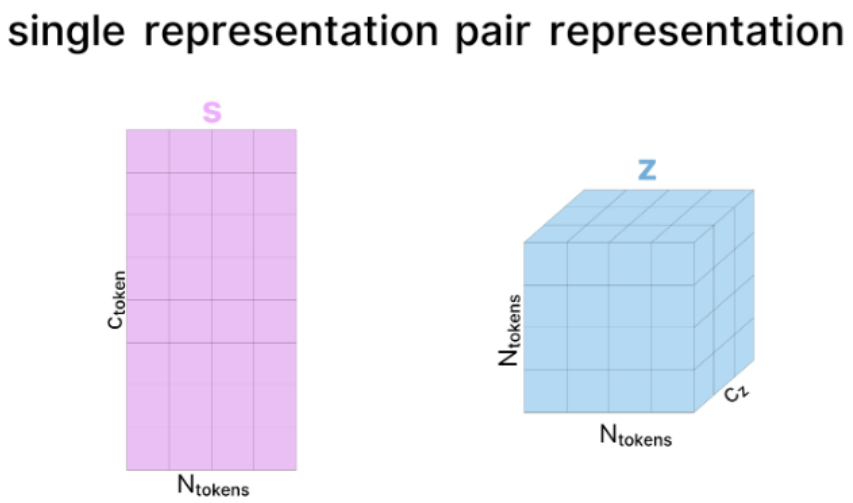
AF3 key differences from AF2

- More use of general chemical information than MSA
  - MSA doesn't exist for ligands and doesn't inform DNA's 3D structure
- Diffusion generates atom positions rather than structure module
  - Reducing unneeded complexity made it easier to generalize the method

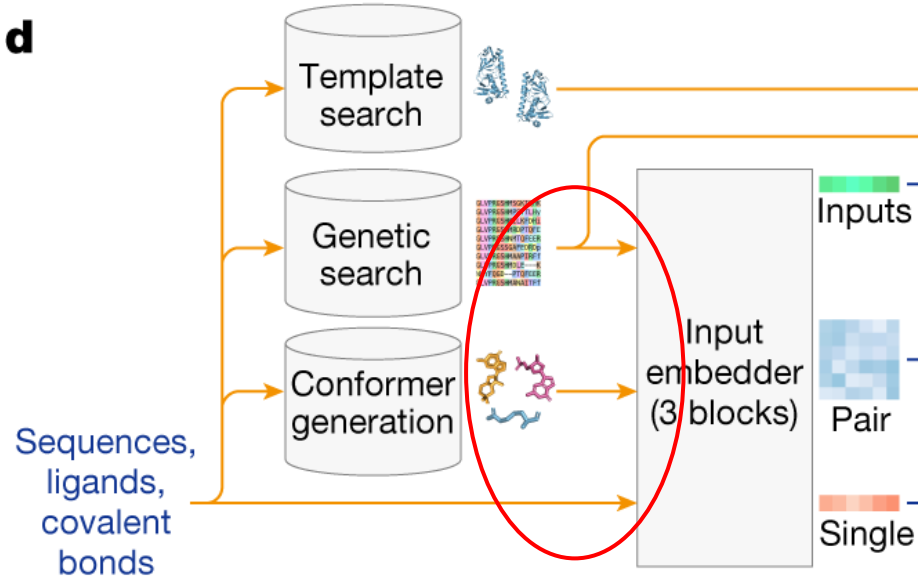


AF3 tokens

- Per-residue – standard AAs and nucleotides
- Per-atom – ligands and modified residues
- Residue type
  - Which AA, nucleotide, or unknown (ligands)
- Reference conformer
  - Each atom's element, initial position, charge
- MSA summary features
  - Each position's residue type distribution

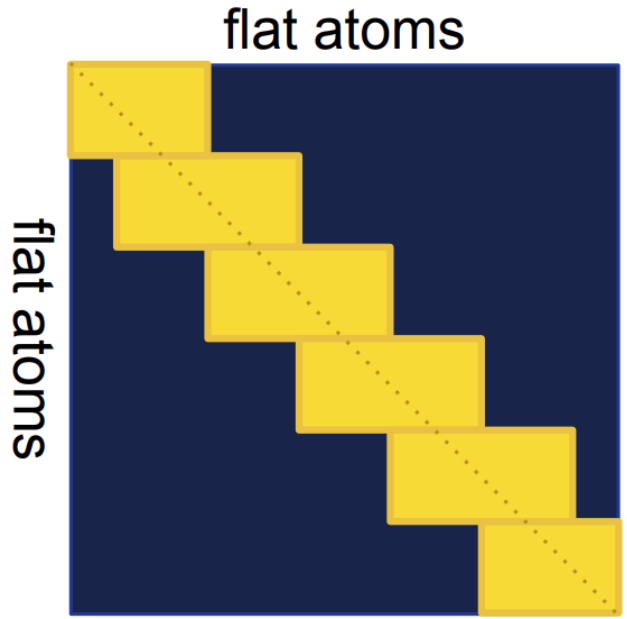


The Illustrated AlphaFold (Simon & Silberg, 2024)



Sequence-local atom attention

- Represent structure as a flat list of atoms
- Each atom attends to nearby 128 atoms
  - Model learns rules about local atom arrangements
  - Bounds compute and memory



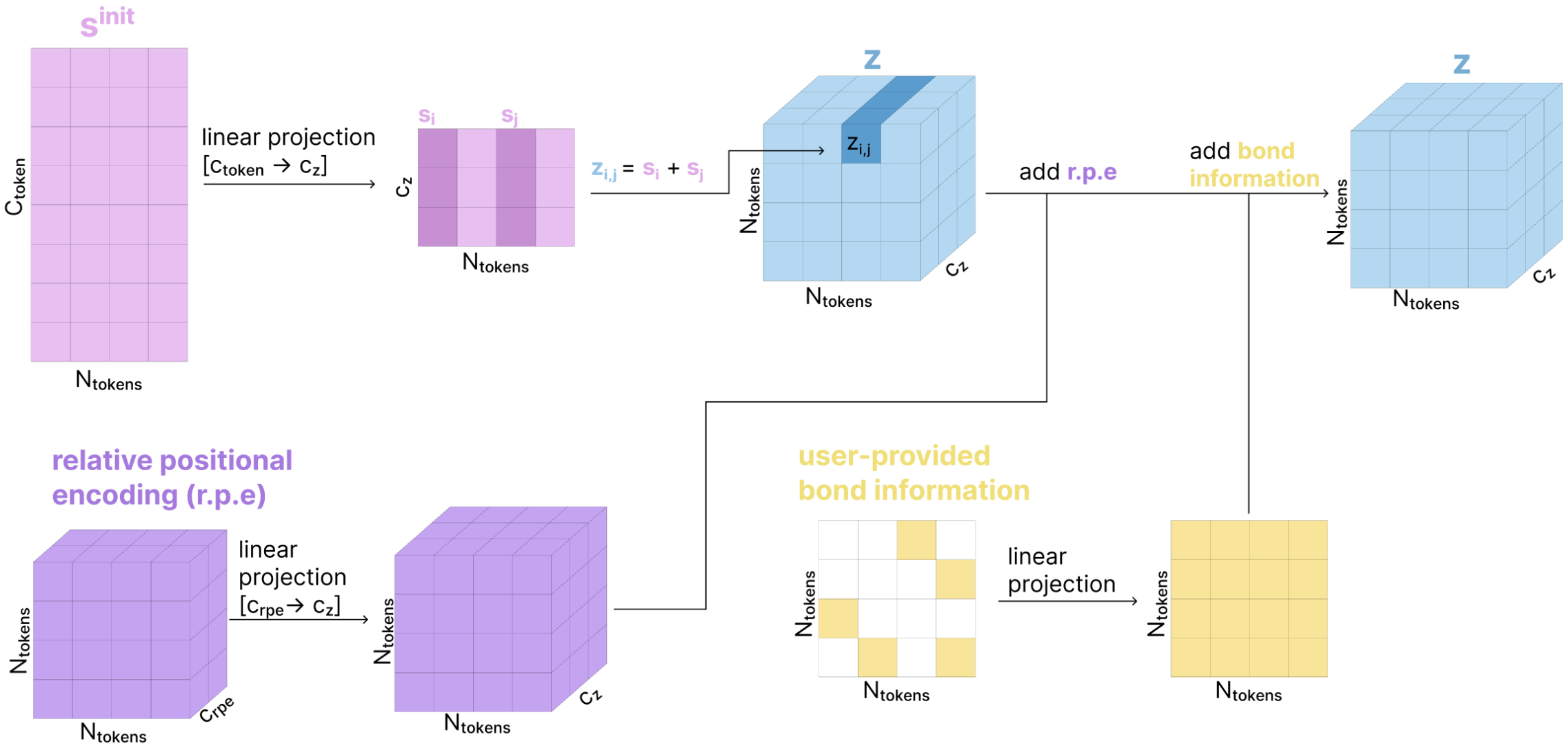
 # Cross attention transformer.

15: $\{\mathbf{q}_l\} = \text{AtomTransformer}(\{\mathbf{q}_l\}, \{\mathbf{c}_l\}, \{\mathbf{p}_{lm}\}, N_{\text{block}} = 3, N_{\text{head}} = 4)$

 # Aggregate per-atom representation to per-token representation

16: $\mathbf{a}_i = \underset{\substack{l \in \{1,\ldots,N_{\text{atoms}}\} \\ \text{tok\_idx}(l)=i}}{\text{mean}} (\text{relu}(\text{LinearNoBias}(\mathbf{q}_l)))$

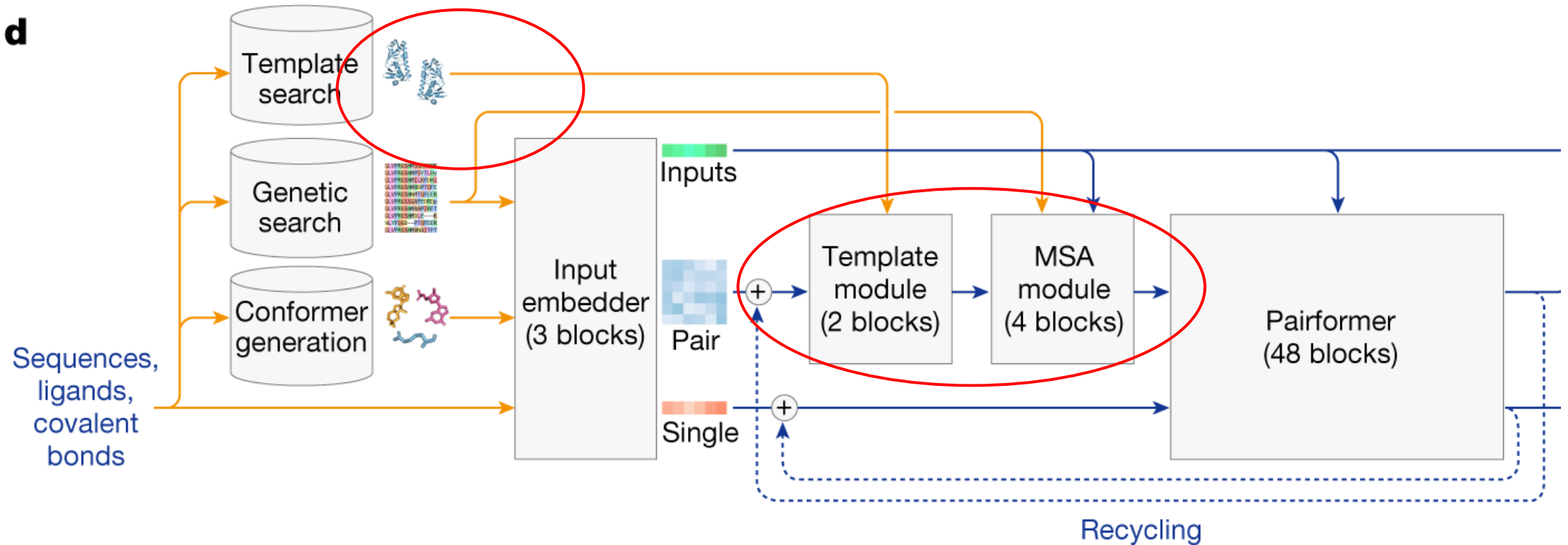
Pair representation



The Illustrated AlphaFold (Simon & Silberg, 2024)

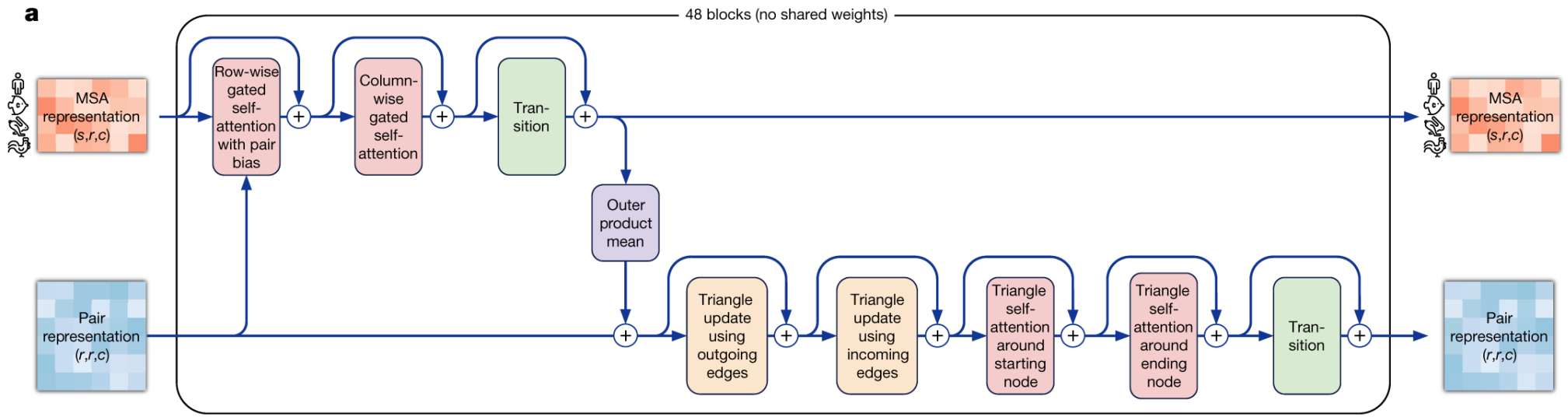
Evolutionary features are used later

- MSA – up to 16K related protein or RNA sequences
- Templates – atom positions for up to 4 related proteins

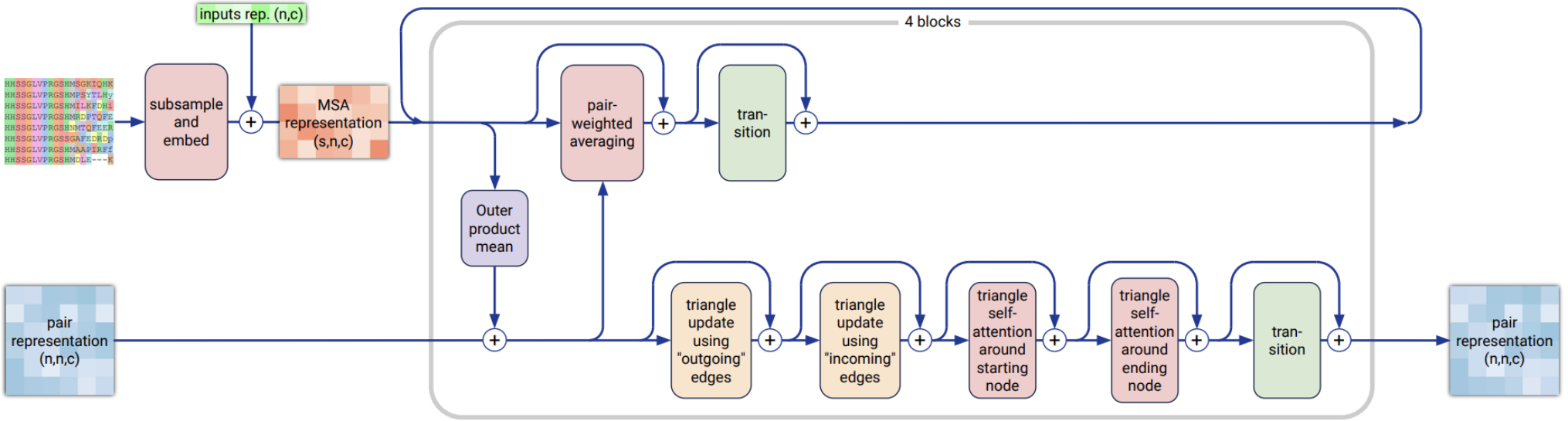


MSA in AlphaFold 2 vs 3

AF2 Evoformer



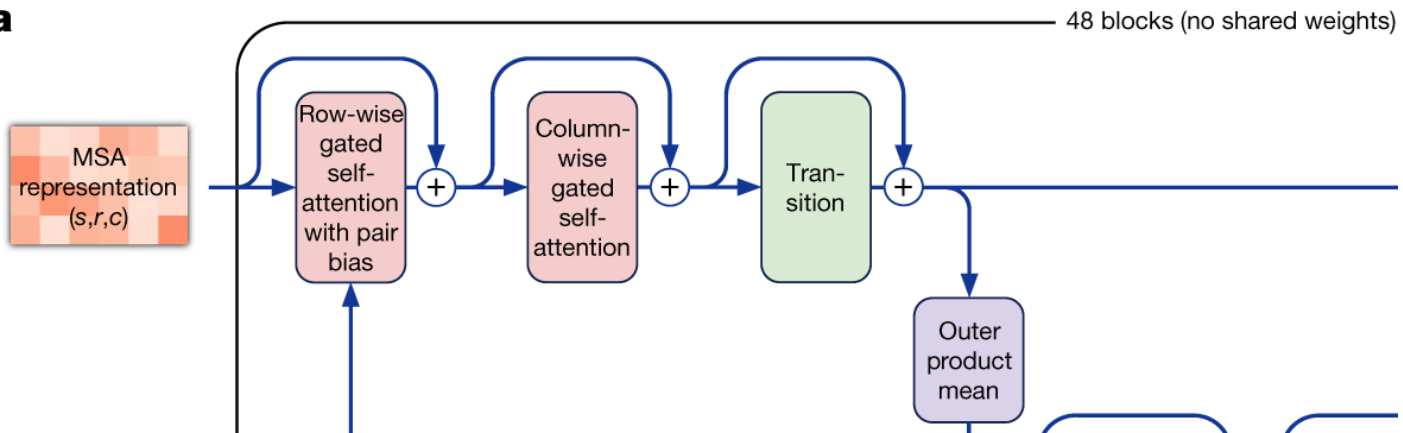
AF3 MSA Module



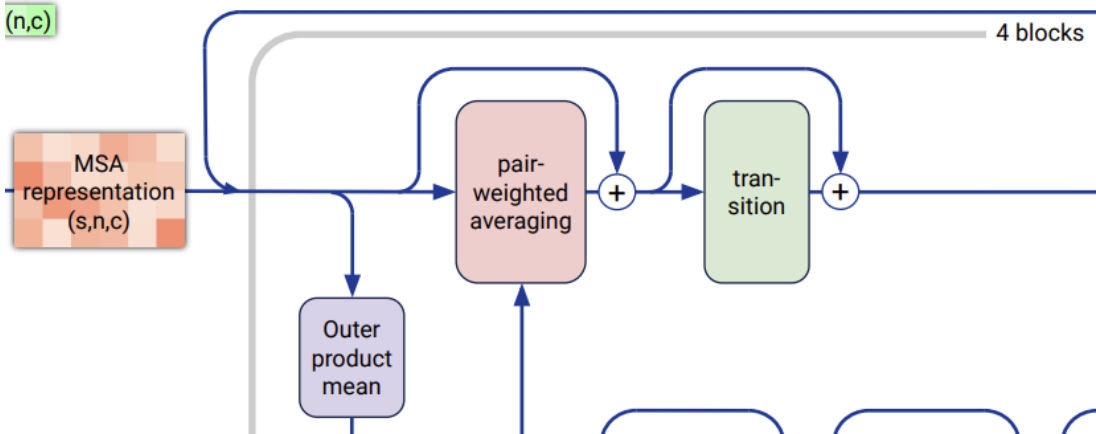
Deprioritizing MSA in AF3

- Only 4 blocks vs AF2's 48
- Pair-weighted average – row attn but attn scores are from pair rep
- No column attention – less info exchanged between sequences
- MSA is not used for downstream steps, only pair rep

AF2



AF3



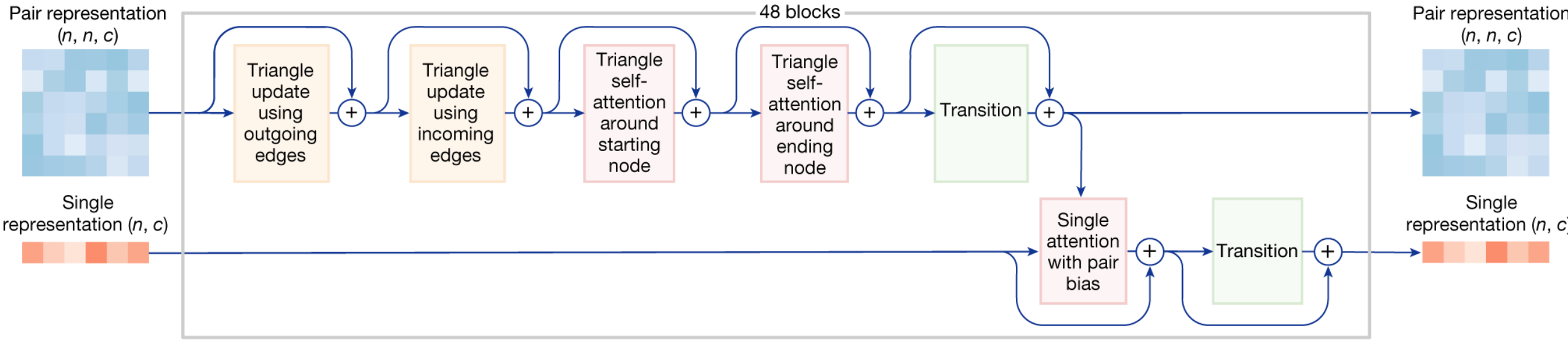
Pairformer

- Triangular updates to pair rep
- Uses pair rep to bias single rep attention scores

$$b_{ij}^h \leftarrow \text{LinearNoBias}(\text{LayerNorm}(\mathbf{z}_{ij}))$$

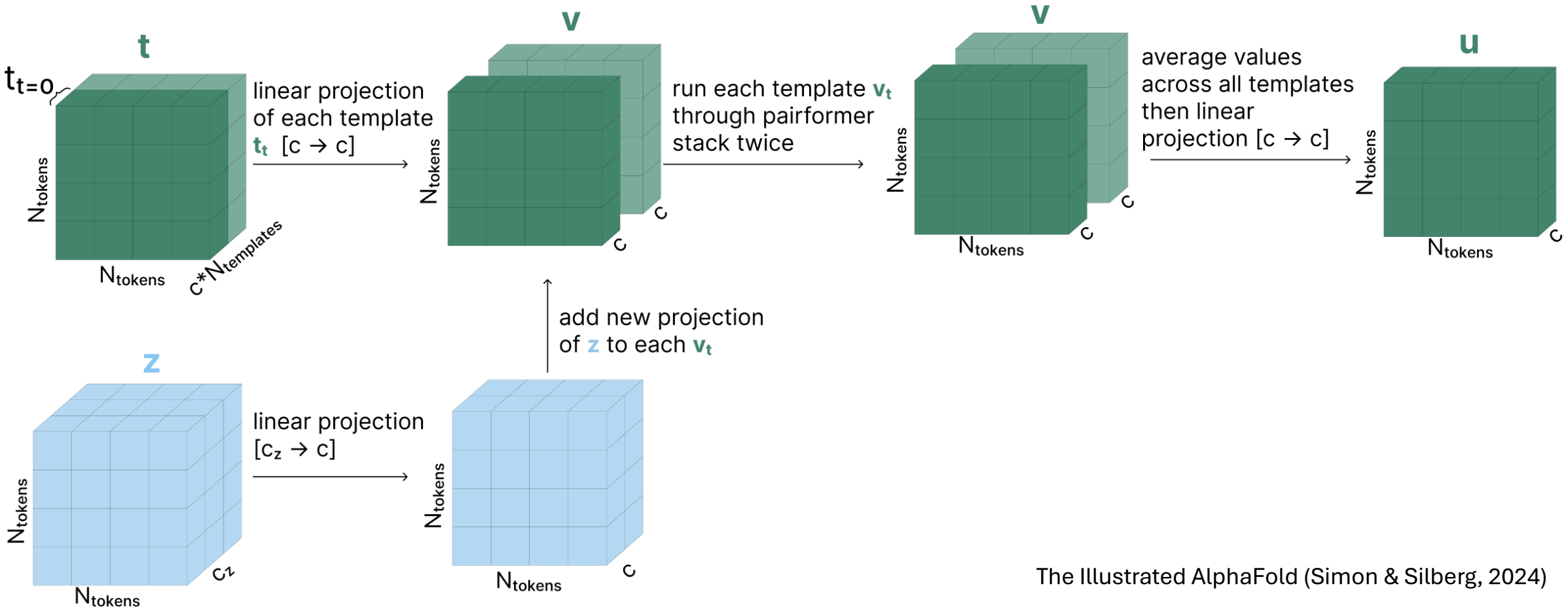
 # Attention

$$A_{ij}^h \leftarrow \text{softmax}_j\left(\frac{1}{\sqrt{c}}\,\mathbf{q}_i^{h\top}\mathbf{k}_j^h + b_{ij}^h\right)$$



Template module

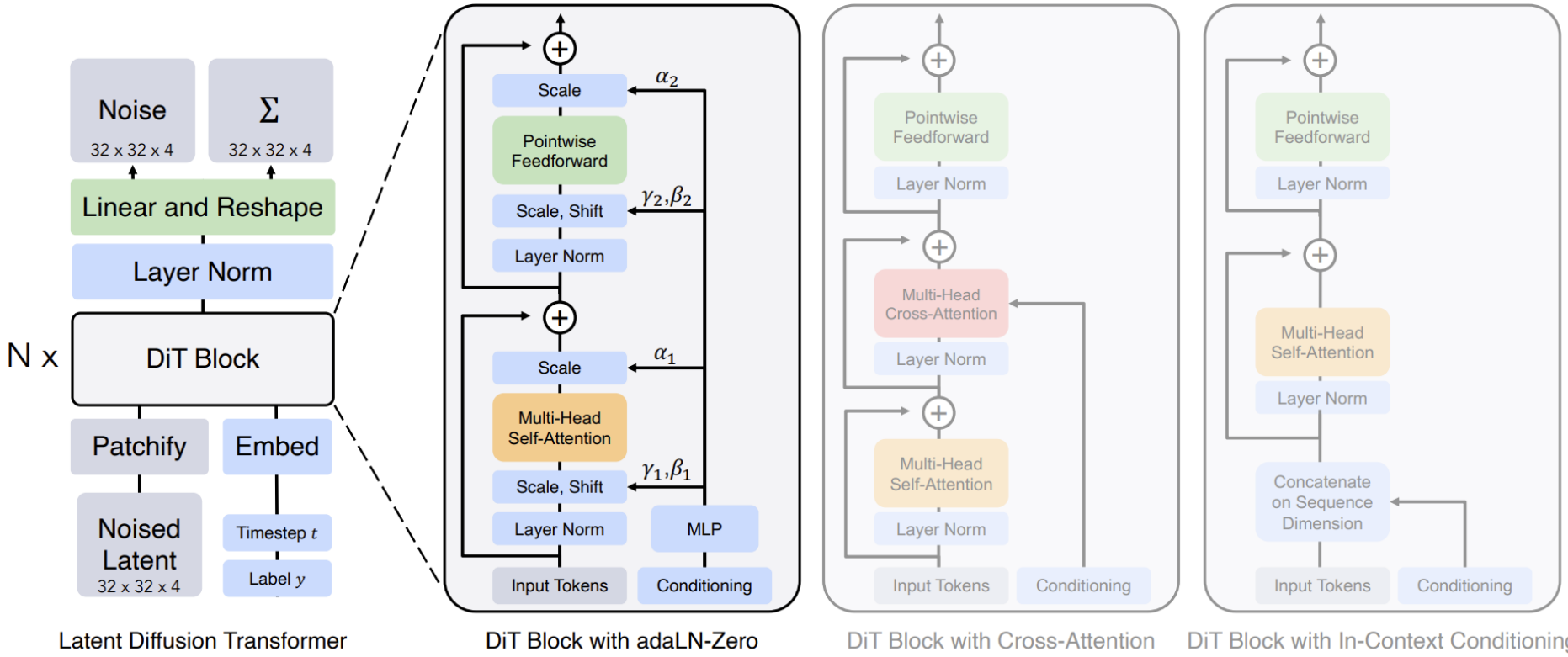
- z is the pair rep; afterwards u is added to z



The Illustrated AlphaFold (Simon & Silberg, 2024)

Diffusion transformer

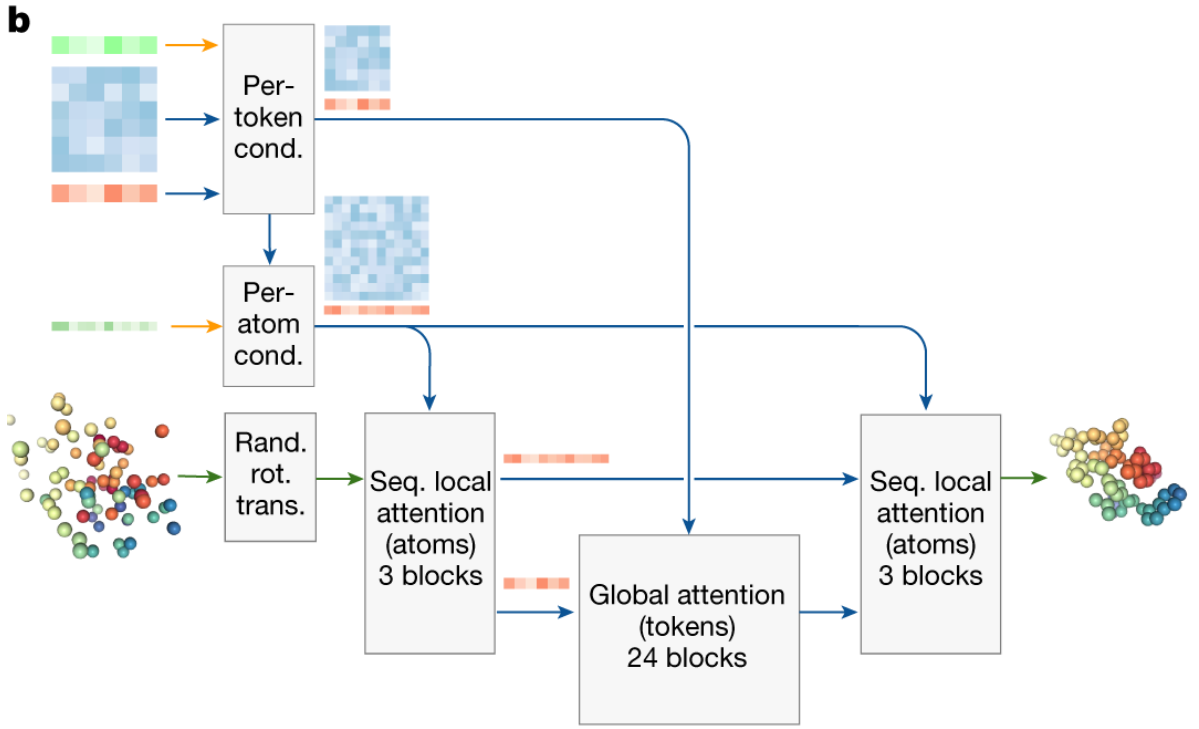
- Gaussian noise is added to the input atom positions
- Model iteratively denoises input using conditioning embedding



Peebles & Xie, *CVPR* 2023

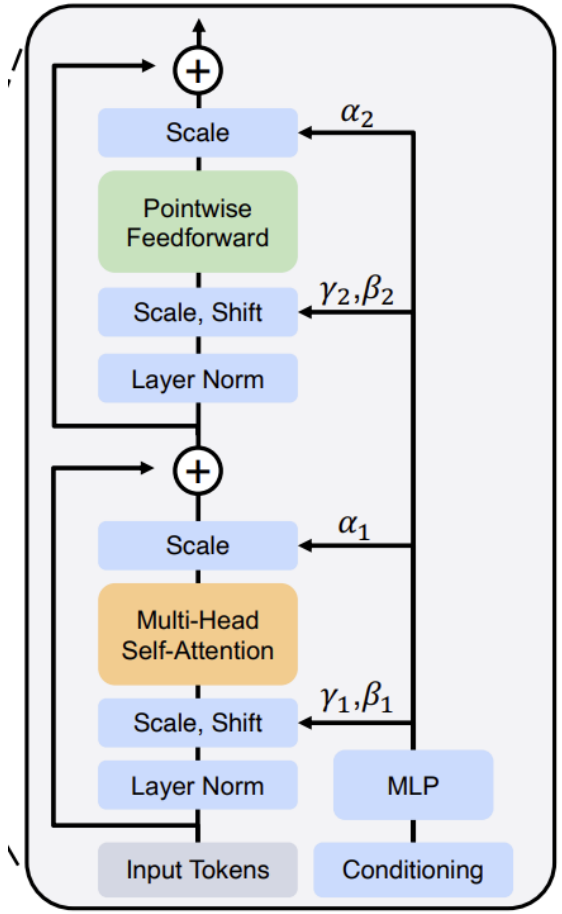
Diffusion module

- Converts random positions into predicted structure with input features, pair rep, and single rep
- No geometric bias like SE(3) invariance
  - Randomly rotate and translate positions to encourage invariance



Conditioning diffusion

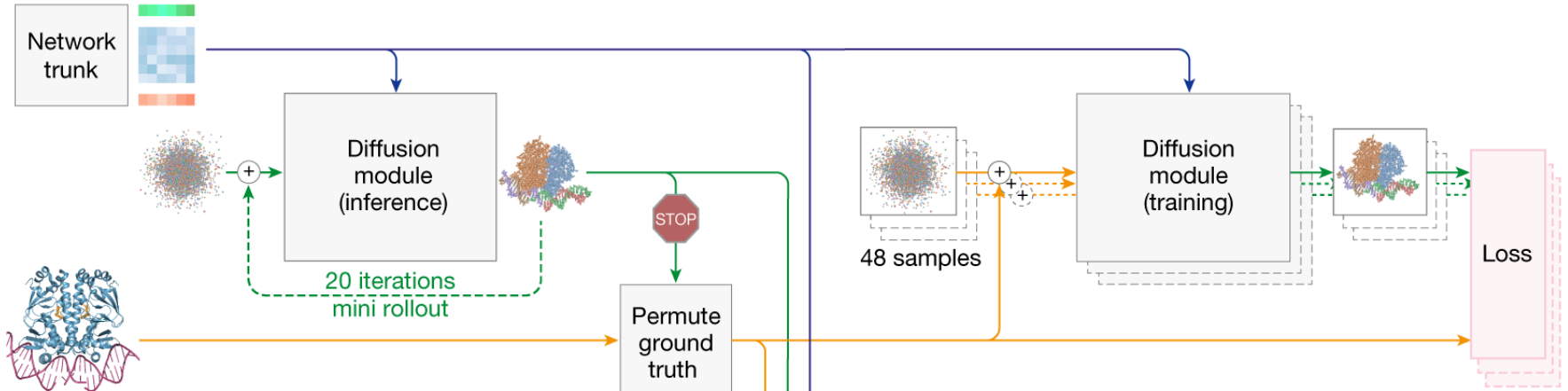
- AF3 is unusual as most of its compute is for creating the conditioning embeddings, not the diffusion itself
- Accordingly, there are several layers of conditioning
- Input features – added to the initial atom positions
- Pair rep – bias the attention scores
- Single rep – adaptive layernorm scale and shift



DiT Block with adaLN-Zero

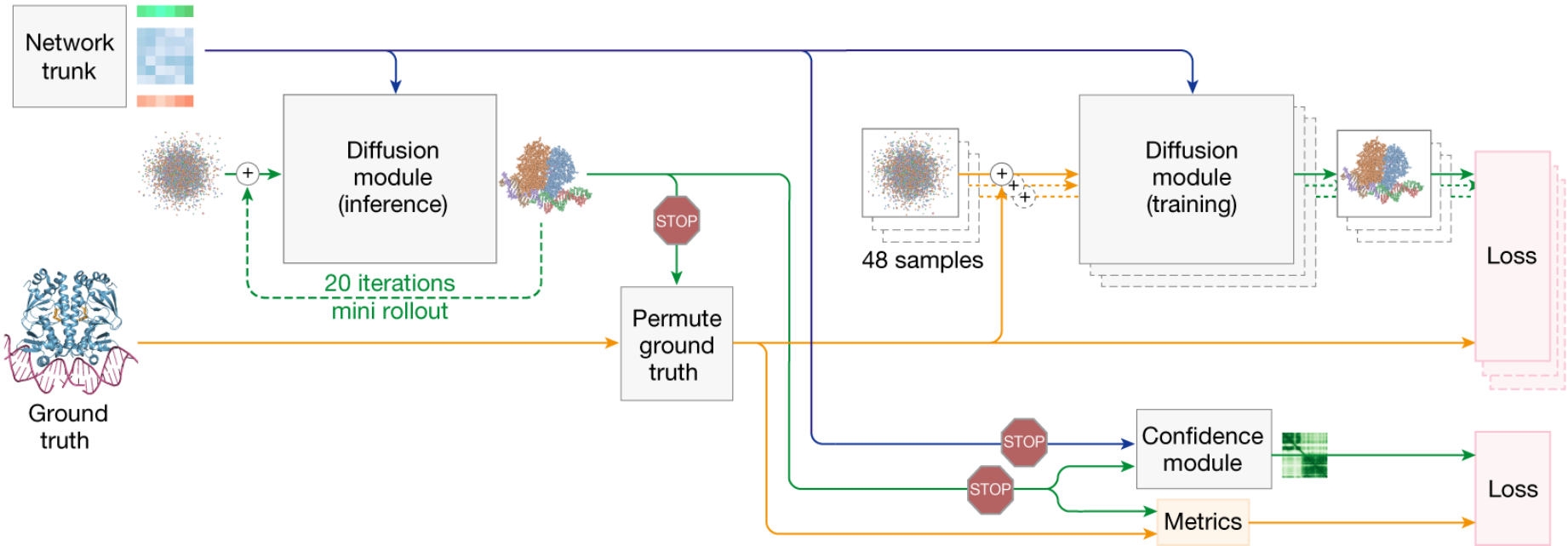
Diffusion training

- Each time trunk is run, diffusion module is trained on 48 different versions of the input structure in parallel
  - Each version has different random rotations, translations, noise
  - Improves training efficiency since diffusion module is much cheaper
- Primarily uses MSE loss, upweighting nucleotide and ligand atoms
- Auxiliary losses for bond length and smooth LDDT



Confidence module

- Mini rollout – generate a final structure to compare with GT
- Predict the min error between generated structure and any permutation of the GT chains and ligands
  - Ordering of identical chains or ligands in a complex is arbitrary



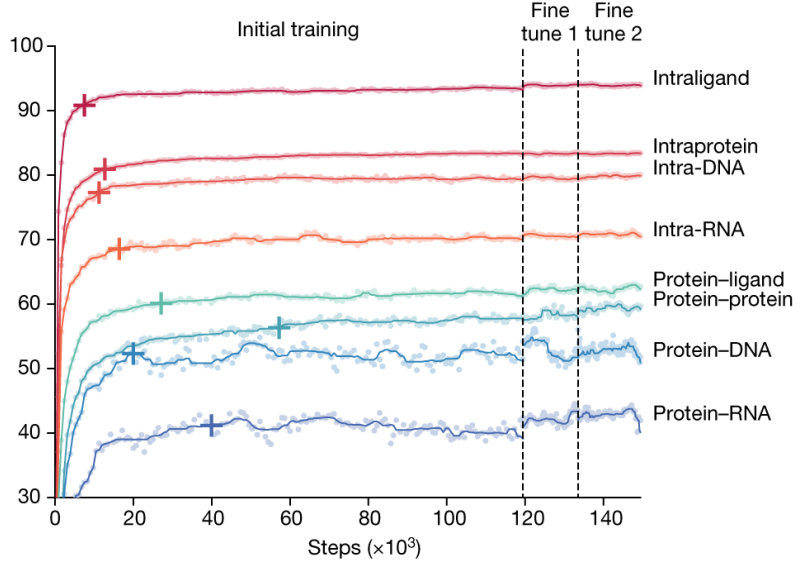
Overall loss

- Predicted local distance difference test (pLDDT)
  - How accurate an atom's distances are to nearby atoms
- Predicted distance error (PDE)
  - Expected error in distance between a pair of tokens
- Predicted aligned error (PAE)
  - Expected error in token *i*'s position assuming token *j* is at its GT position
- Resolved – predict whether each atom is resolved in GT
- Distogram – binned distances between tokens, same as AF2

$$\mathcal{L}_{\text{loss}} = \alpha_{\text{confidence}} \cdot (\mathcal{L}_{\text{plddt}} + \mathcal{L}_{\text{pde}} + \mathcal{L}_{\text{resolved}} + \alpha_{\text{pae}} \cdot \mathcal{L}_{\text{pae}}) + \alpha_{\text{diffusion}} \cdot \mathcal{L}_{\text{diffusion}} + \alpha_{\text{distogram}} \cdot \mathcal{L}_{\text{distogram}} \quad (15)$$

Where $\alpha_{\text{confidence}} = 10^{-4}$, $\alpha_{\text{diffusion}} = 4$, $\alpha_{\text{distogram}} = 3 \cdot 10^{-2}$ and $\alpha_{\text{pae}} = 0$ for all except for the final training stage, where it is set to 1.

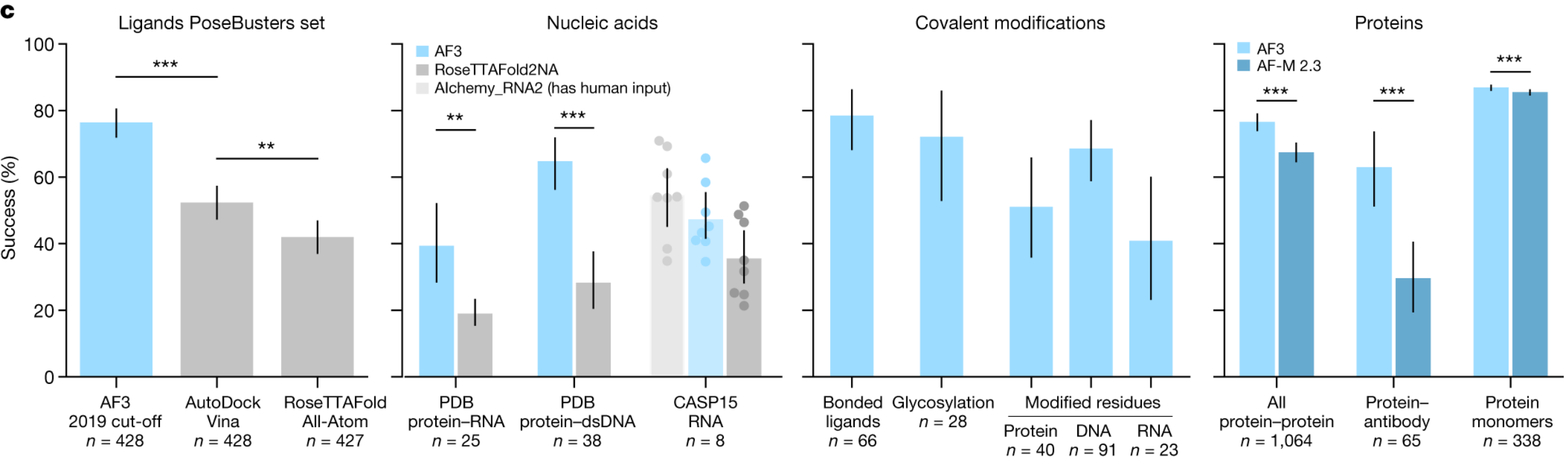
Training stages



	Initial training	Fine tuning 1	Fine tuning 2	Fine tuning 3
Sequence crop size N_{token}	384	640	768	768
Parameters initialized from	Random	Initial training	Fine tuning 1	Fine tuning 2
Sampling weight for disorder PDB distillation	0.02	0.01	0.02	0.02
Train on transcription factor distillation sets	False	False	True	True
Masked diffusion loss for non-protein in disorder PDB distillation	True	False	False	False
Train structure and distogram	True	True	True	False
Train PAE head	False	False	False	True
Diffusion batch size	48	32	32	32
Training samples ($\cdot 10^6$)	≈20	≈1.5	≈1.5	≈1.8
Training times (days on 256 A100s)	≈10	≈3	≈5	≈2
Polymer-ligand bond loss weight	0	1	1	1
Max number of chains	20	20	20	50

Results

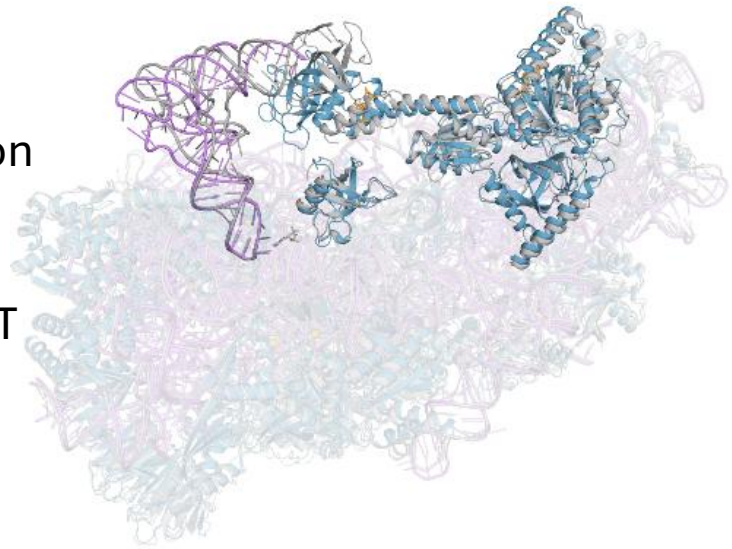
- Success – atoms in pocket have <2 Å rmsd



Examples of predicted complexes

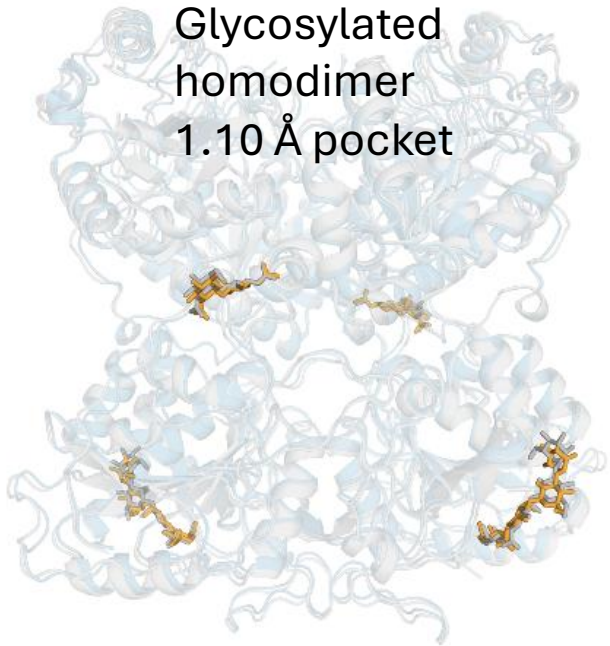
a

Ribosome
rRNA
tRNA
Translation initiation factors
83.0 LDDT
83.1 GDT

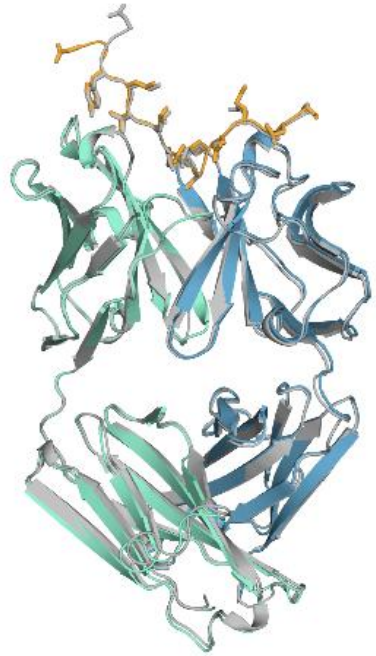


b

Glycosylated homodimer
1.10 Å pocket



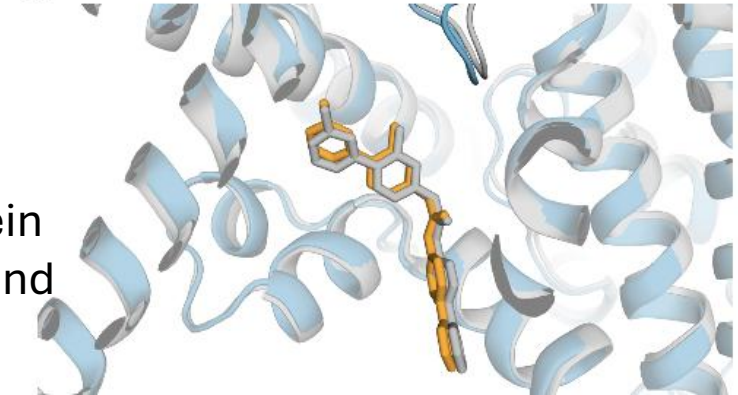
c



Antibody binding to peptide overexpressed in cancer
0.85 DockQ

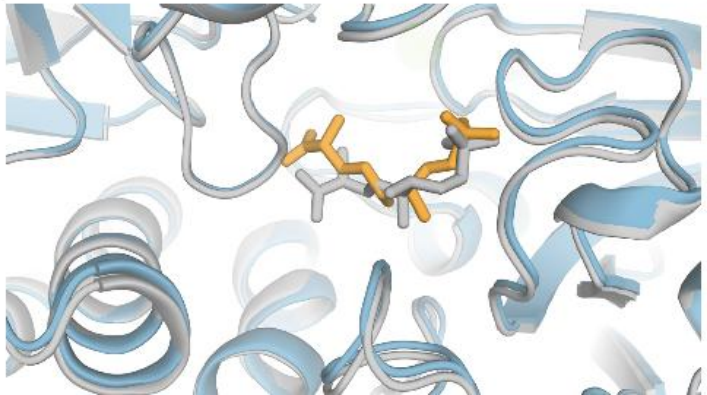
d

Drug candidate inhibiting oncoprotein
1.00 Å ligand

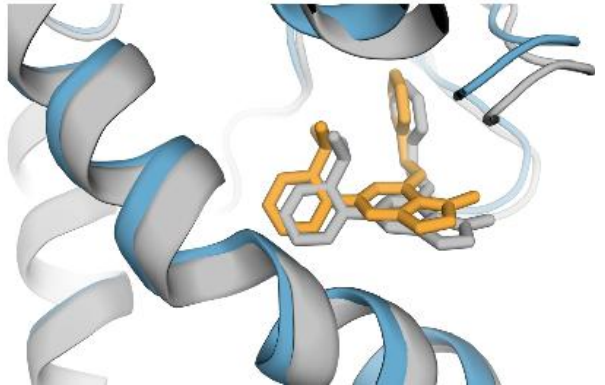


e

Novel protein fold 1.92 Å

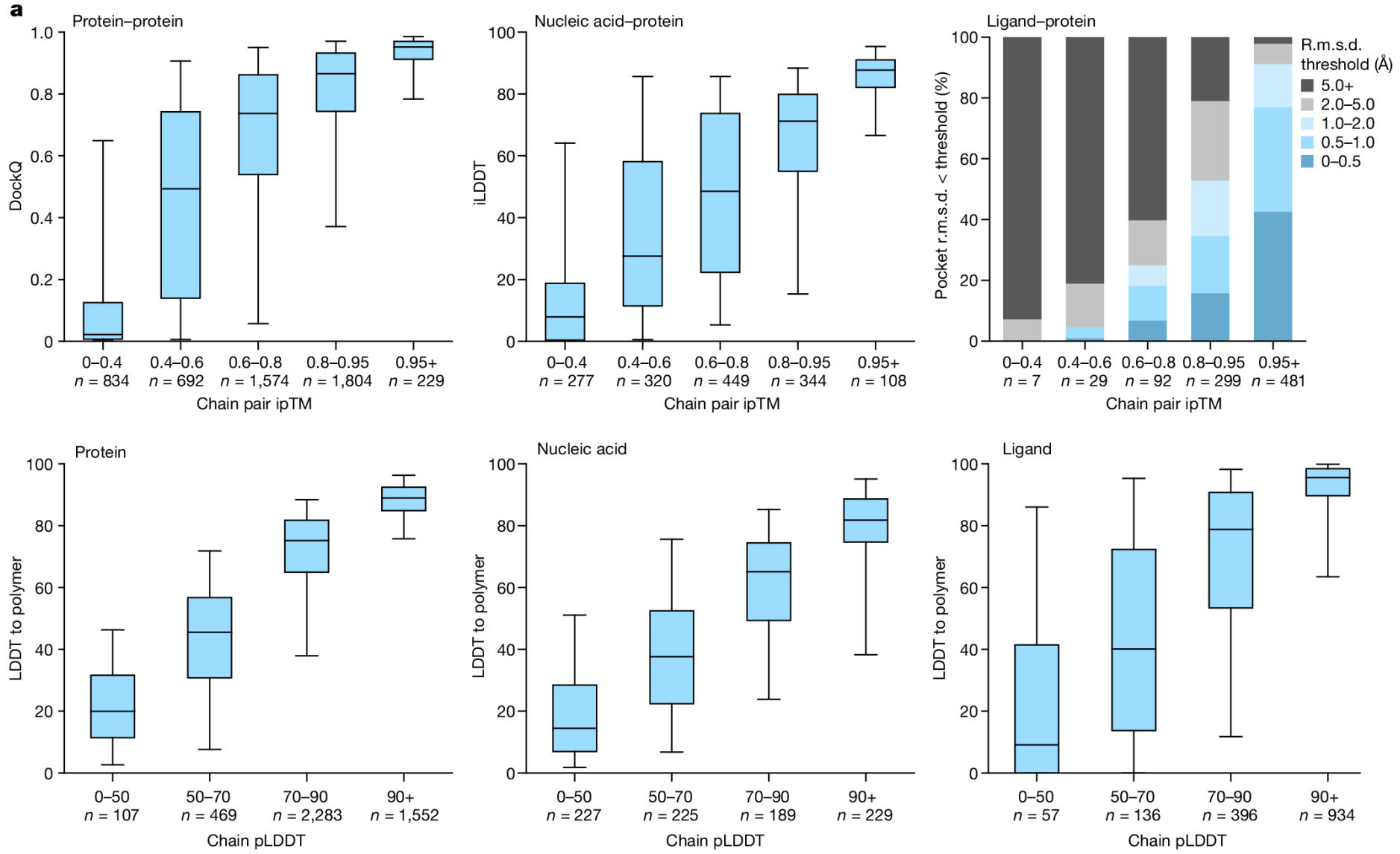


f



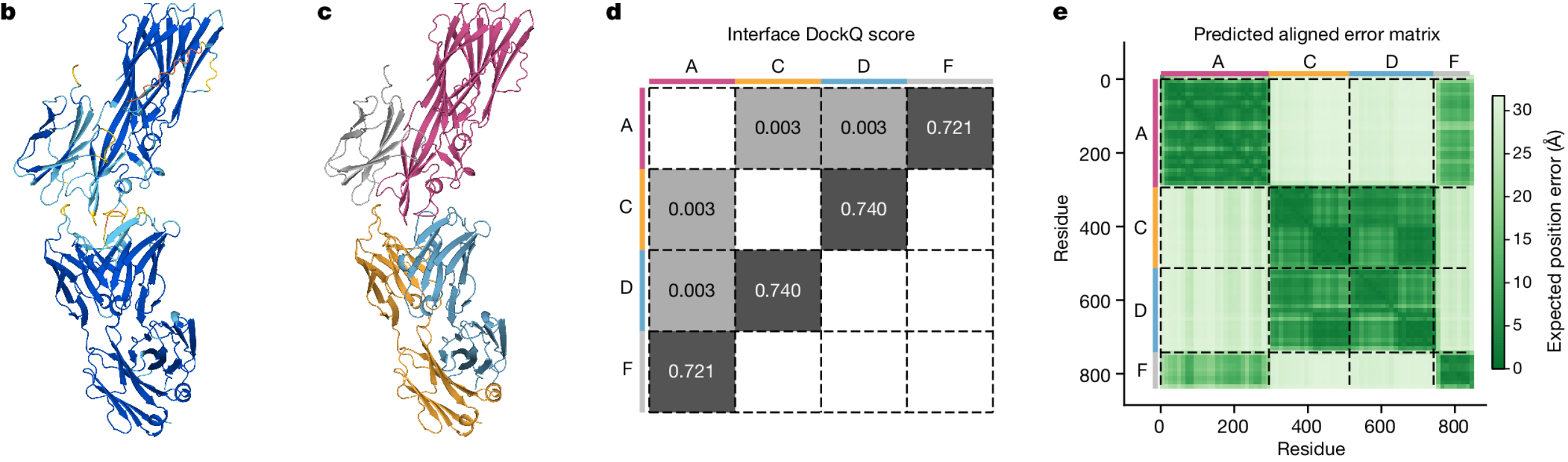
Allosteric binding
0.37 Å

AF3 confidence tracks accuracy measures



PAE is a reliable estimate of inter-chain error

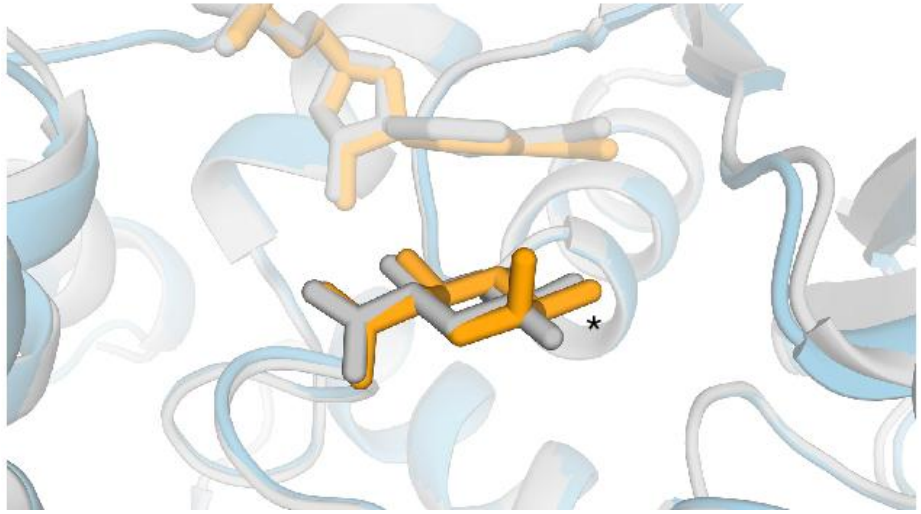
- Interface between A to C, D has lower predicted and GT accuracy



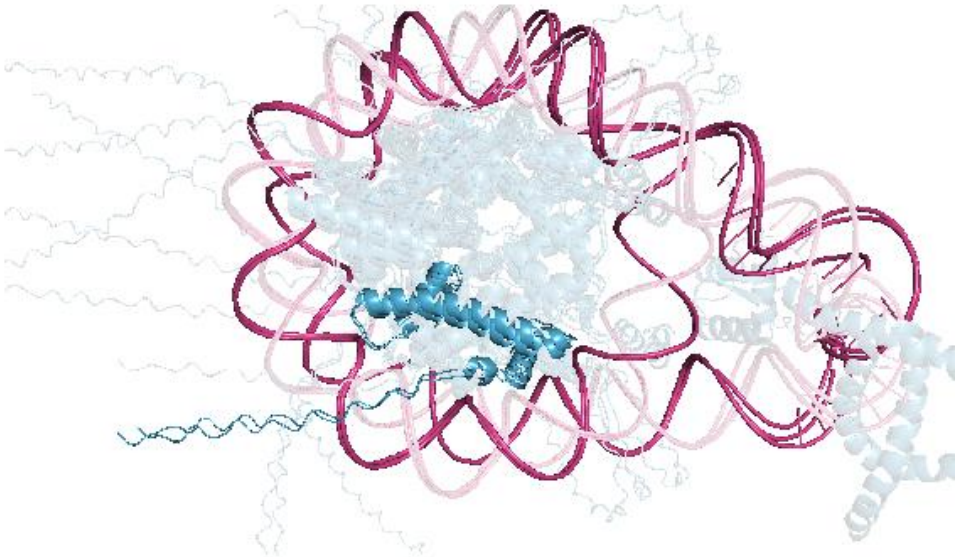
Stereochemistry limitations

- Violations persist despite ranking 1K predictions with 100x penalty
  - e.g. 4.4% of top predictions for PoseBusters had a chirality violation

Chirality violation

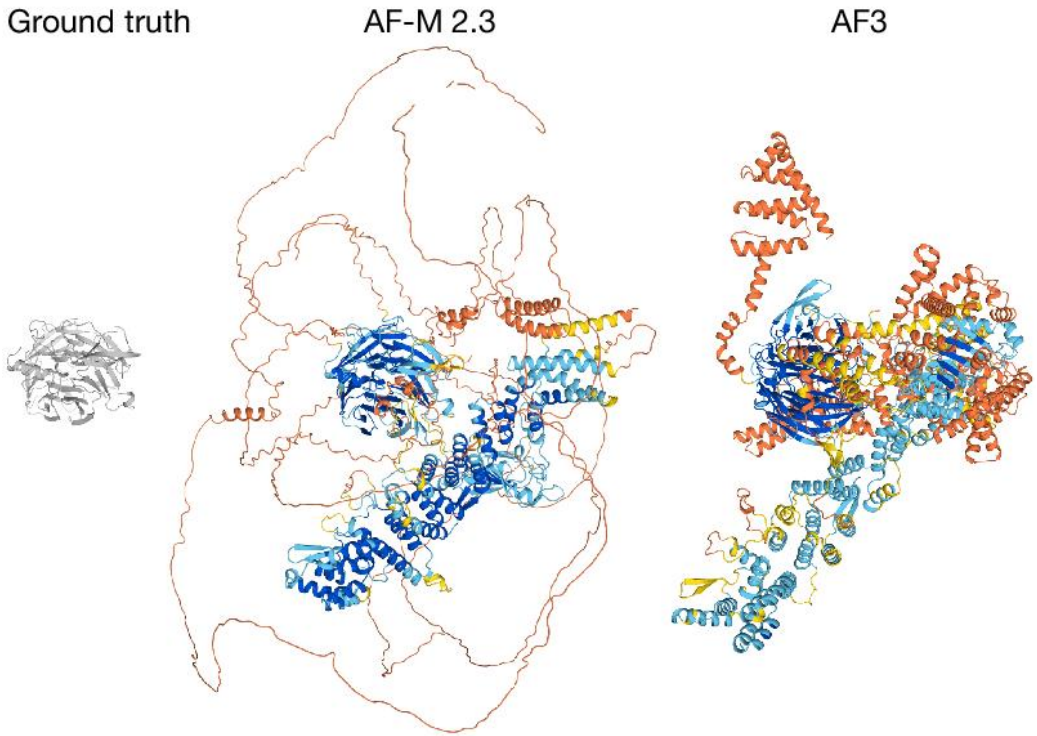


Overlapping atoms



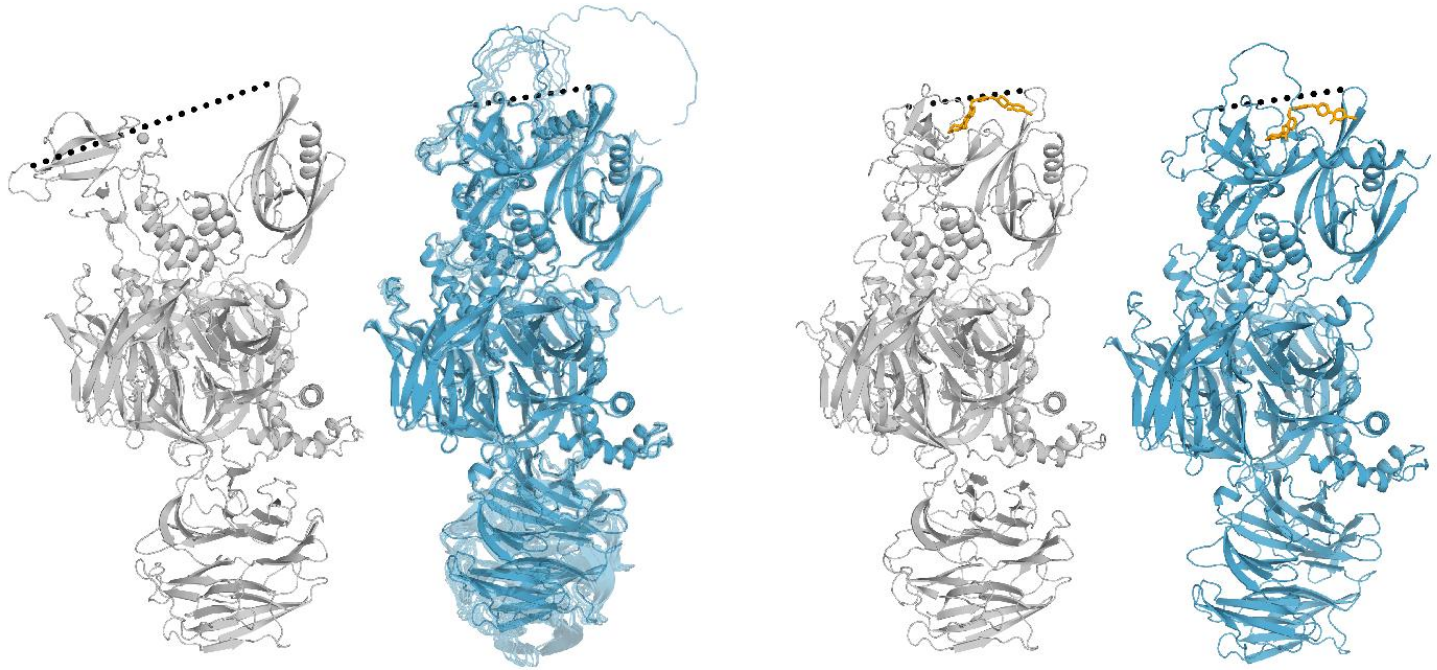
Hallucination and dynamics limitations

AF3 may hallucinate structure



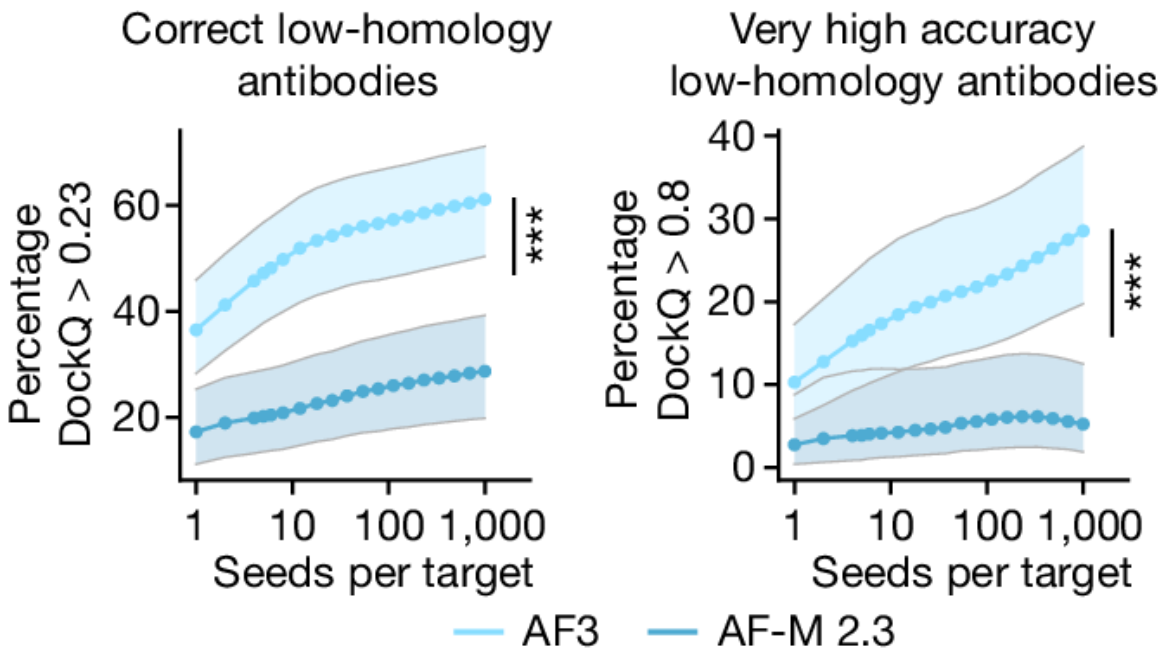
GT has 1,854 unresolved residues

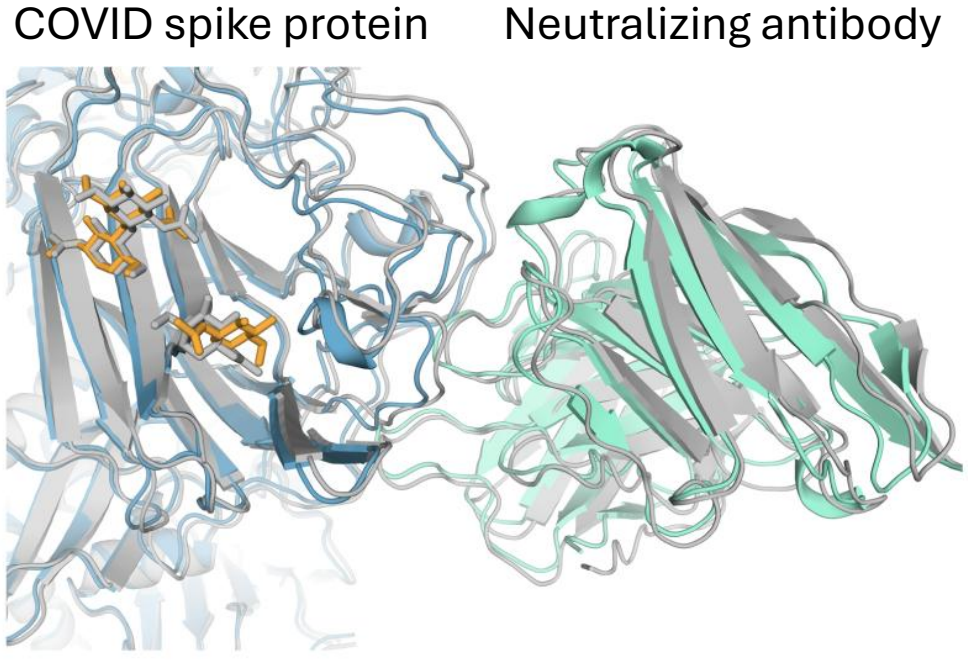
AF3 only predicts some conformations



Many seeds are needed for hard targets

- Antibodies lack reliable MSA because every B cell has a distinctly shuffled and highly mutated sequence
- More seeds didn't improve performance for other molecules





Conclusions

- AF3 achieves SOTA accuracy for predicting many types of biomolecular interactions
- Major improvement in predicting structures with weak MSA
- Bridges protein structure prediction and ligand docking

Future Directions

- Leverage advancements in deep learning to build better models from the same data
- Leverage new sources of training data like cryo-electron microscopy and tomography (cryo-EM, cryo-ET) to model biomolecular dynamics
- <1 Å resolution is often needed for small molecule drug discovery but less important for protein-based drugs

Discussion Questions

- Is it surprising that so much of AF2 was unnecessary?
- How important is distillation?
  - Could it enable a worse architecture to have comparable performance?
- AF3 bucks the trend of using stronger domain specific biases
  - Going forward, what is the role of such biases? e.g. SE(3) invariance
- AF3 does a lot of feature engineering (24 input features!)
  - Will feature complexity increase, or could something simplify this?