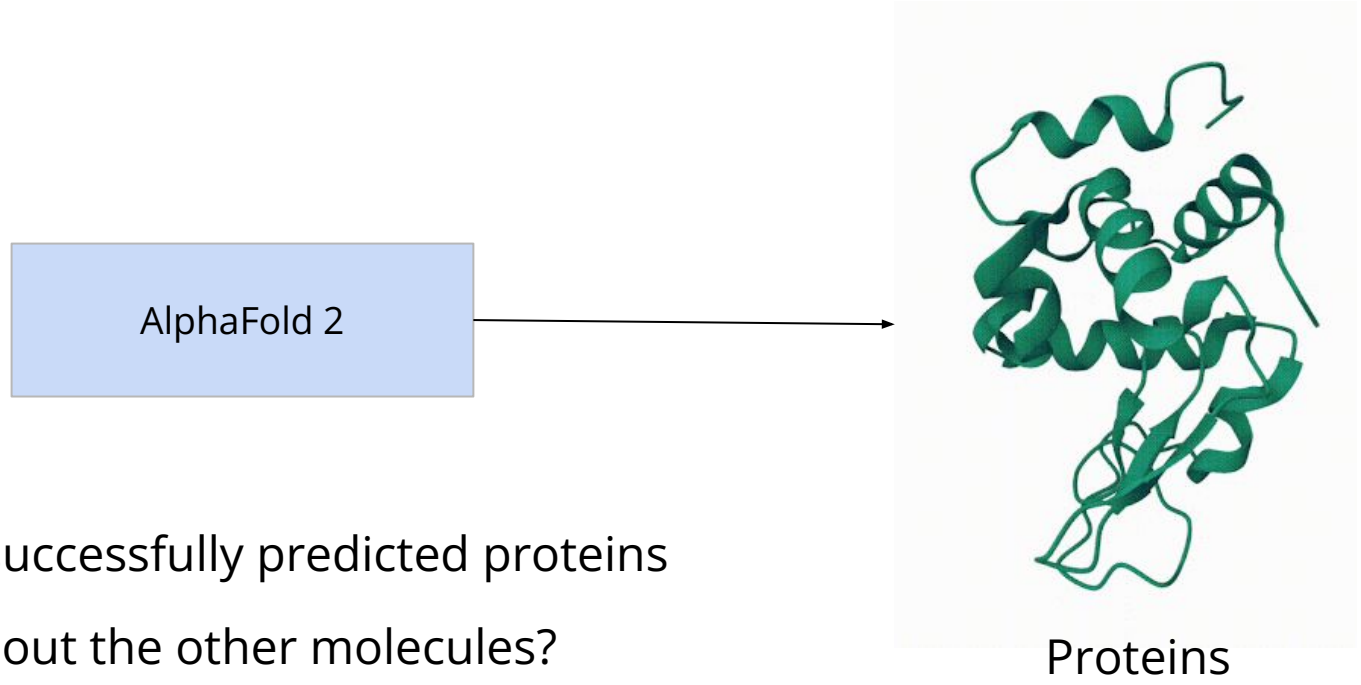
IBISC Deep Learning reading group

AlphaFold 3

Clément Bernard

Abramson, J., Adler, J., Dunger, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500 (2024). https://doi.org/10.1038/s41586-024-07487-w

AlphaFold 2

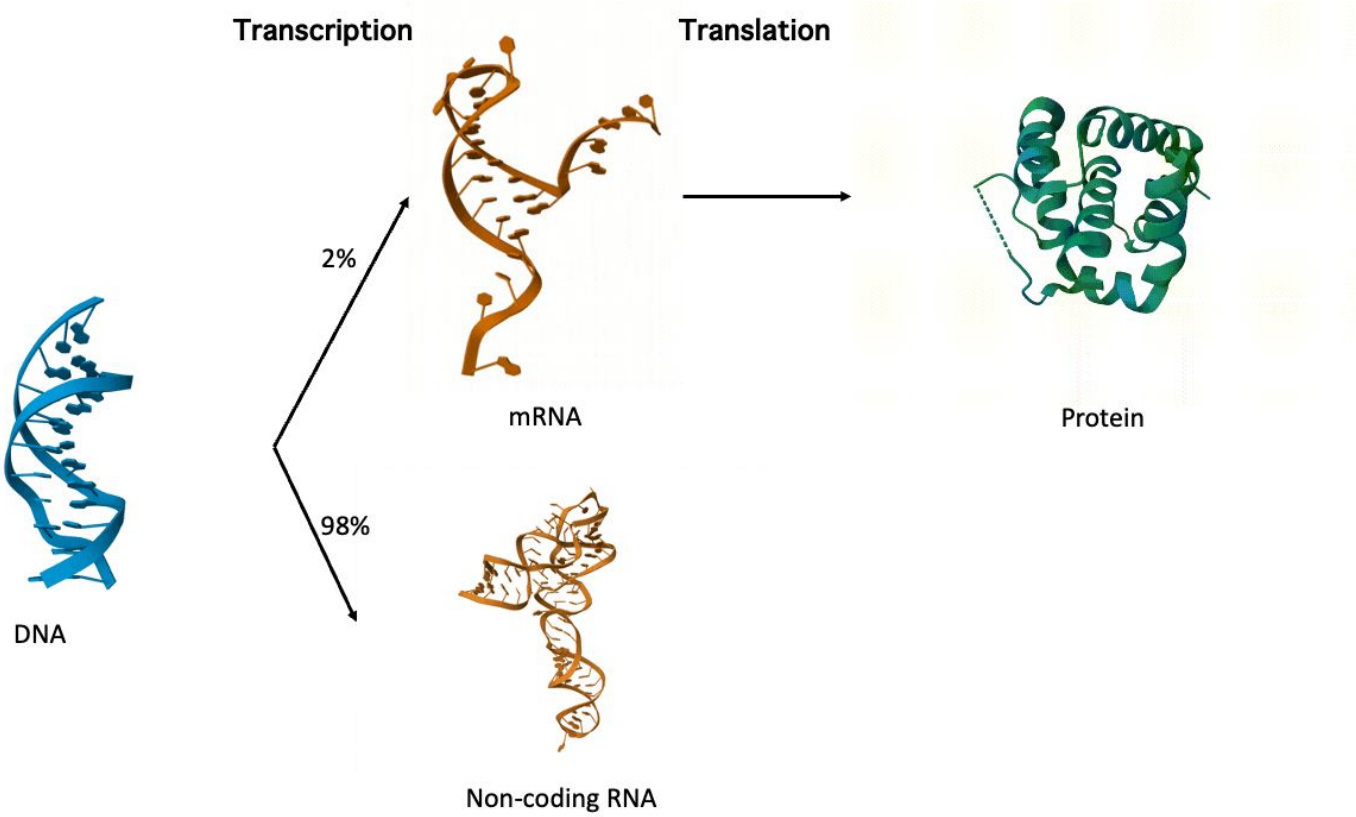


Proteins

AlphaFold 2 successfully predicted proteins

What about the other molecules?

Biological molecules

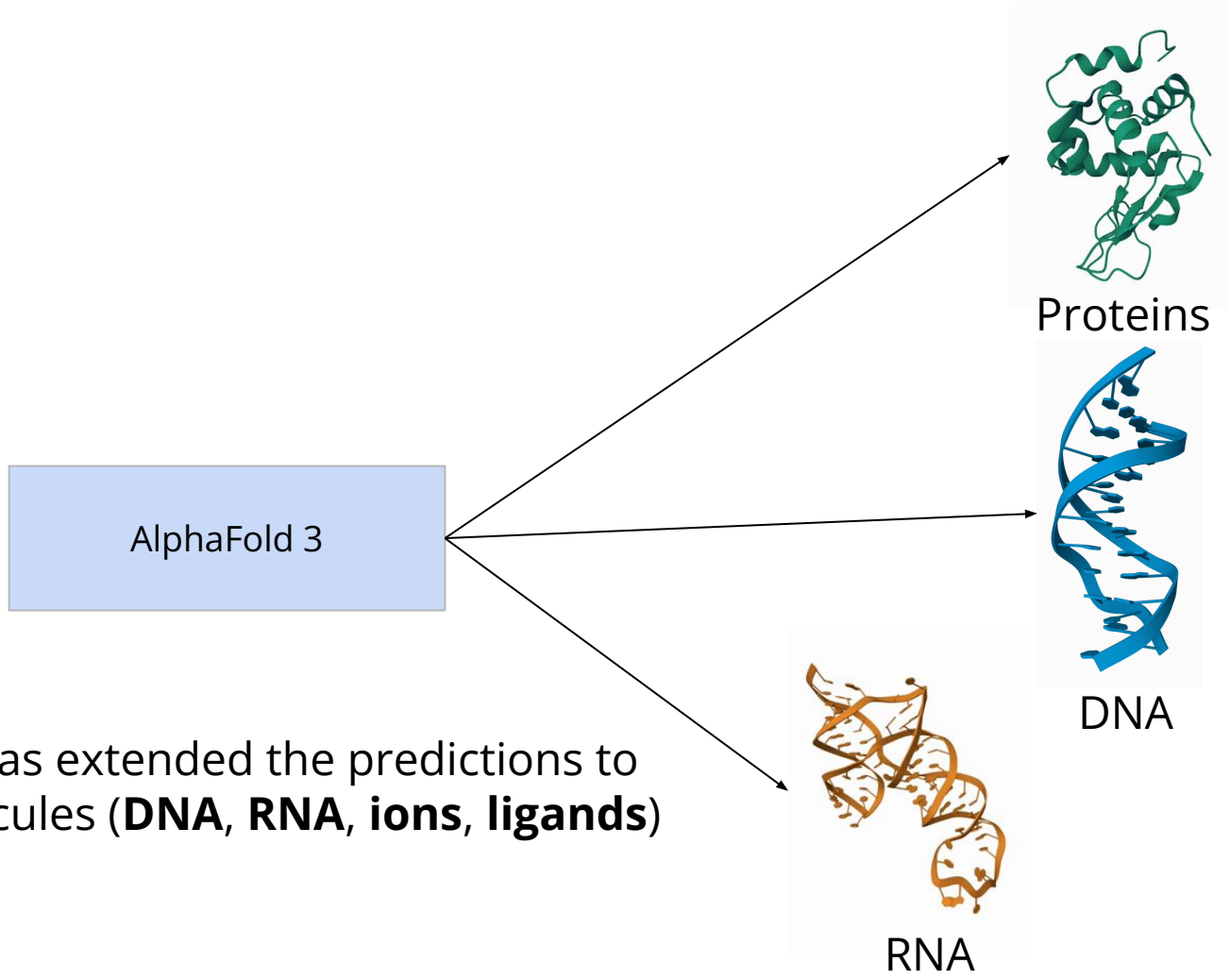


Biological molecules

- **Biological function** of molecules is directly linked to the **3D structure**
- Experimental methods are expensive in both **time** and **money**
- Interest to have **computational methods**: compute 3D structures of molecules from the raw sequence
- Could then be used for gene therapy for instance, to understand diseases, etc.

DNA RNA Proteins

AlphaFold 3

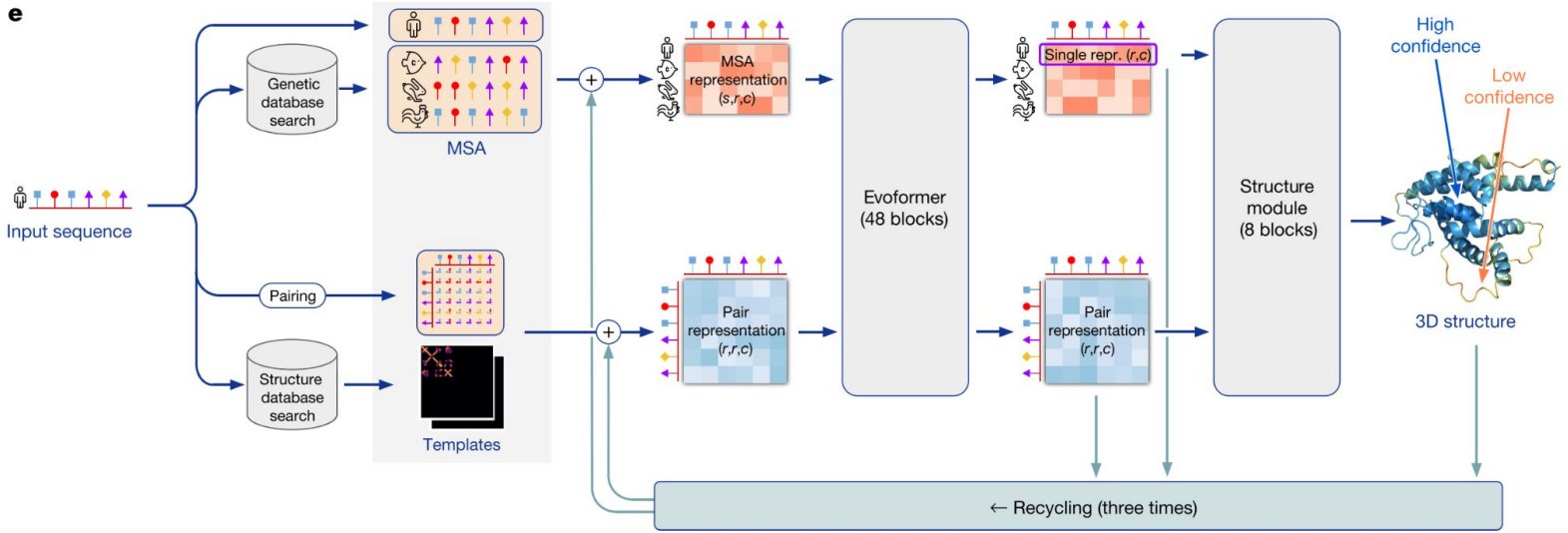


AlphaFold 3 has extended the predictions to different molecules (**DNA**, **RNA**, **ions**, **ligands**)

Table of contents

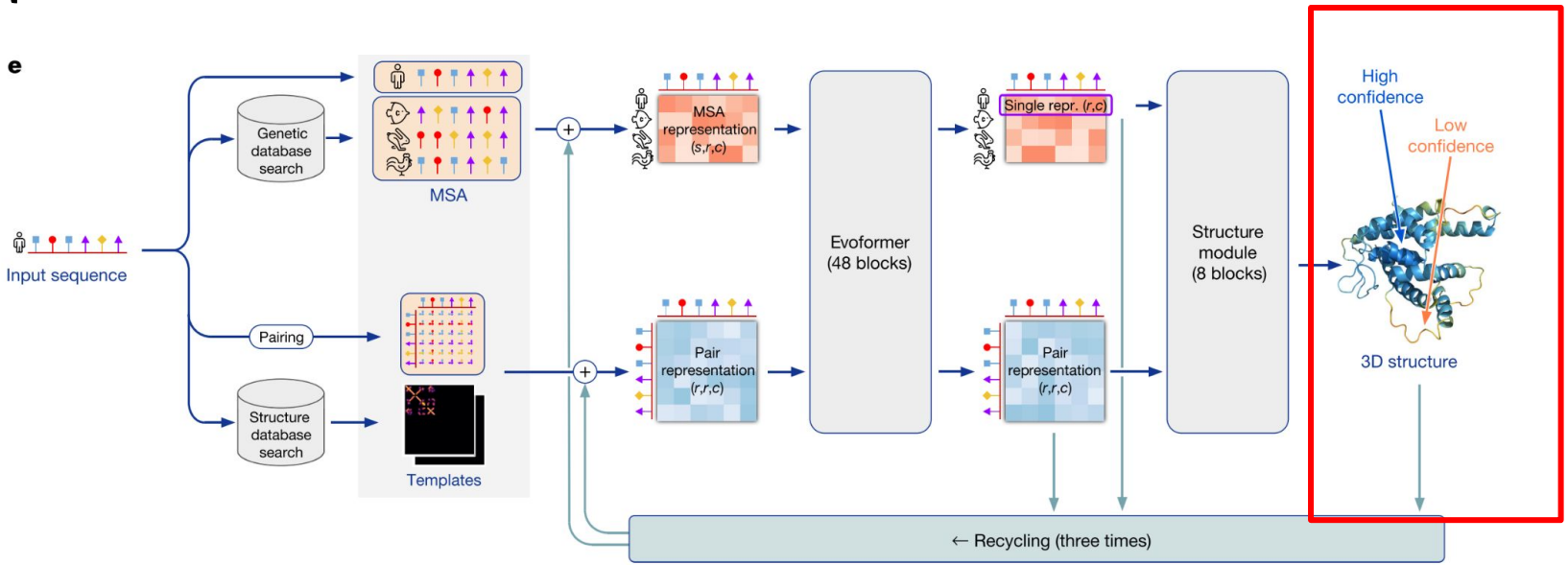
I. AlphaFold 2

AlphaFold 2



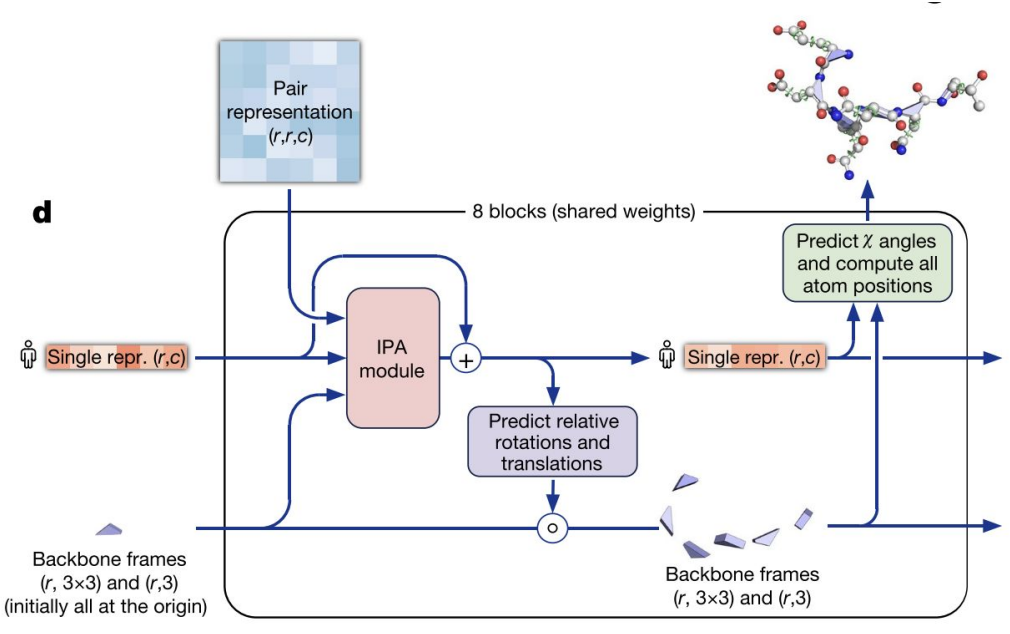
- Let’s have a quick overview of AlphaFold 2 to understand the changes of AlphaFold 3

AlphaFold 2



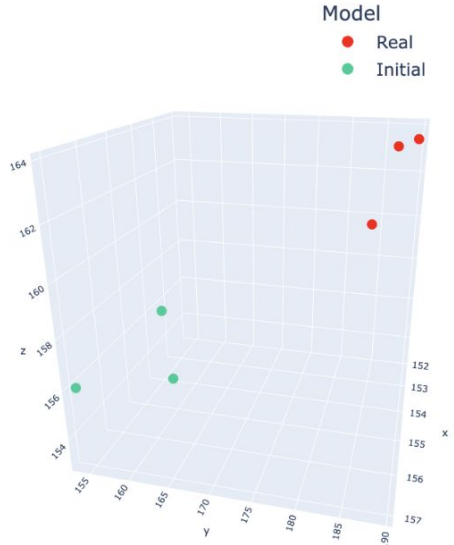
- Let’s understand what is the output of the prediction

AlphaFold 2



- AF2 outputs backbone frames (3x3) and (3)
- Prediction of angles to compute all atom positions

AlphaFold 2

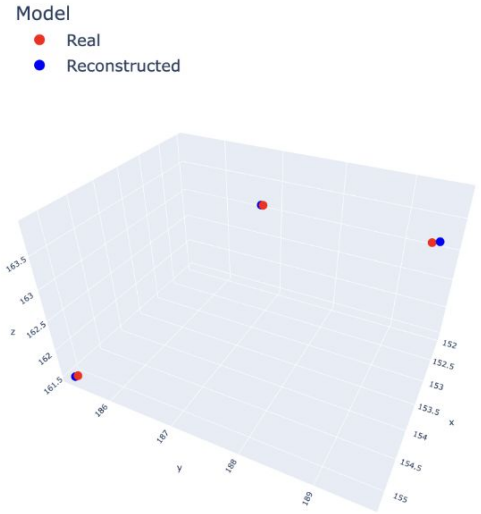


Rotation

0.51	0.85	0.04
0.24	-0.09	-0.96
-0.81	0.51	-0.26

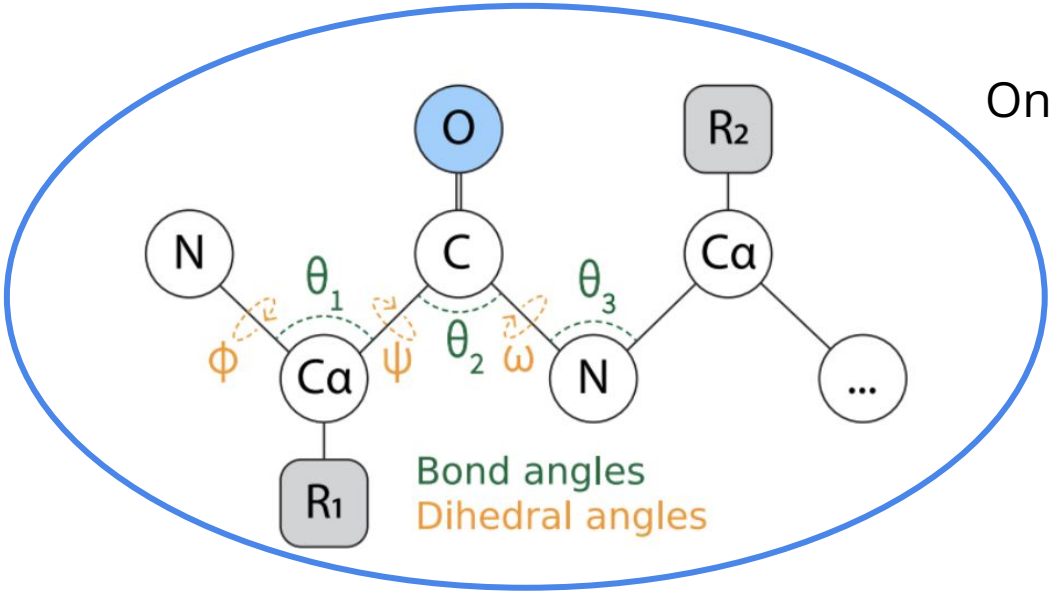
Translation

161
0.24
-0.8



- Given two sets of points, we can find the **rotation** (3x3) and **translation** (3x1) matrices that convert one set of points to the other
- Instead of predicting all the atom positions, **AF2 outputs the rotation/translation** matrices that converts one **base frame** into **global conformation**

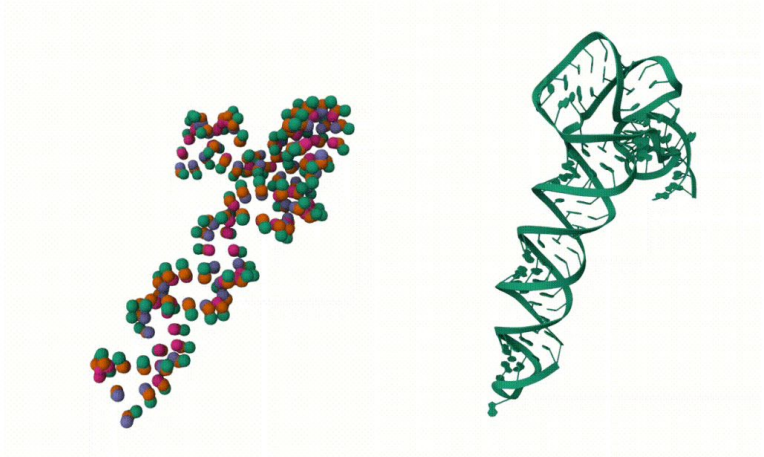
AlphaFold 2



One amino acid

- Amino acid main atoms can be defined as follows
- AF2 wants to only output main atoms per amino acid

AlphaFold 2

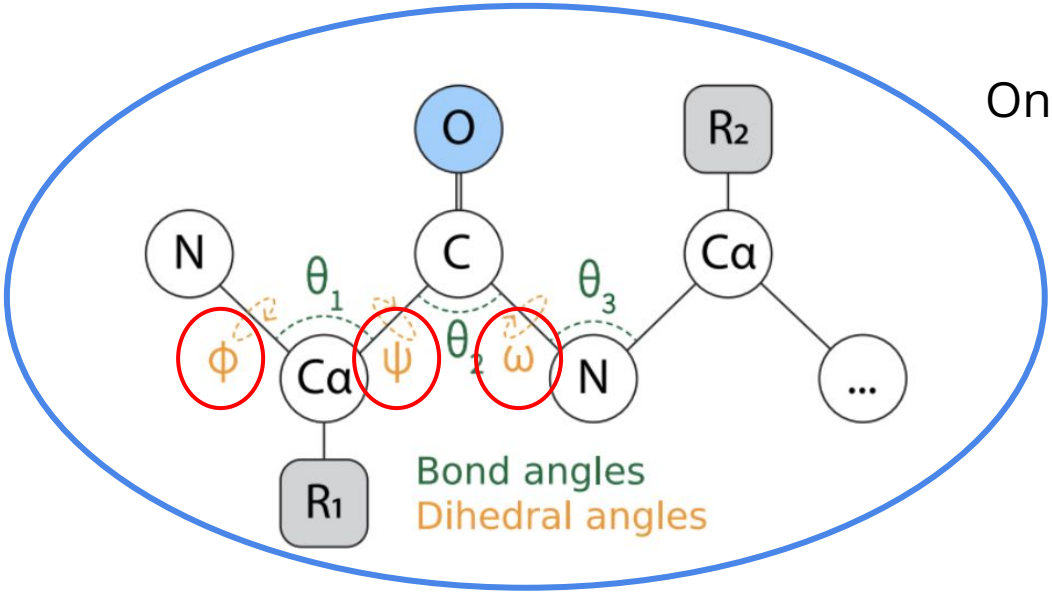


Coarse-grained structure

Full-atom structure

- Example of reconstruction with an RNA (AF2 did it ONLY for proteins)
- It gives only the skeleton, not the full structure

AlphaFold 2



One amino acid

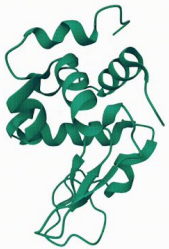
- Dihedral angles were also predicted
- Given torsional angles, we can reconstruct missing atoms

AlphaFold 2

- Torsional angles could only be used on **proteins**, and not on other molecules
- Vocabulary used (in the MSA for instance) only considered the amino acids

RNA

vs

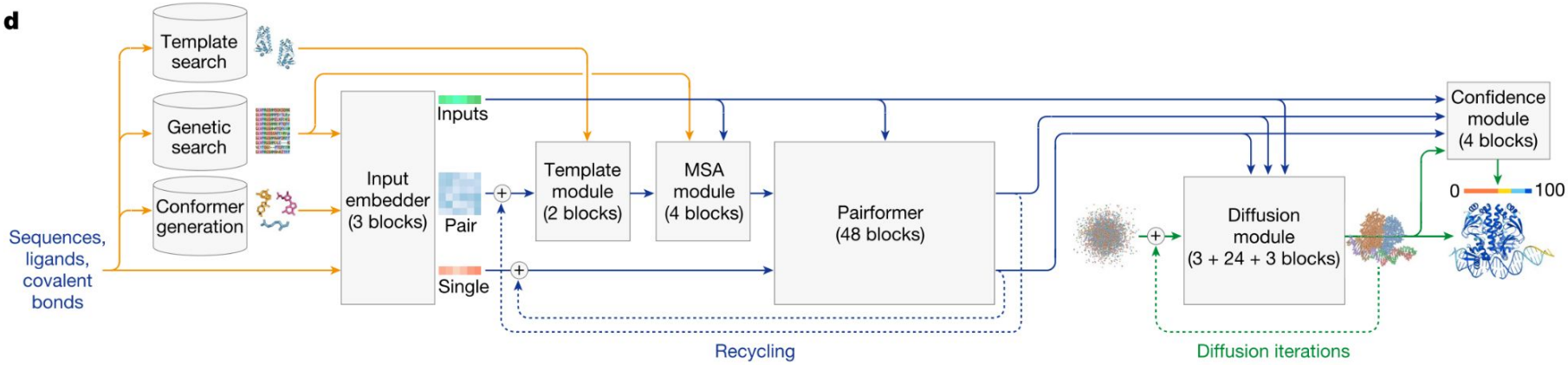


Proteins

-> AF2 was not directly adaptable to DNA, RNA, etc.

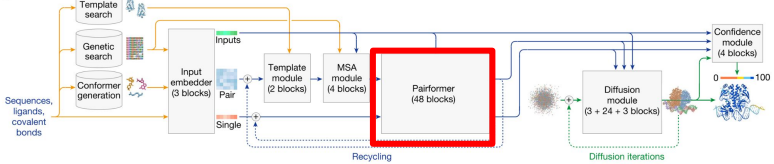
II. AlphaFold 3

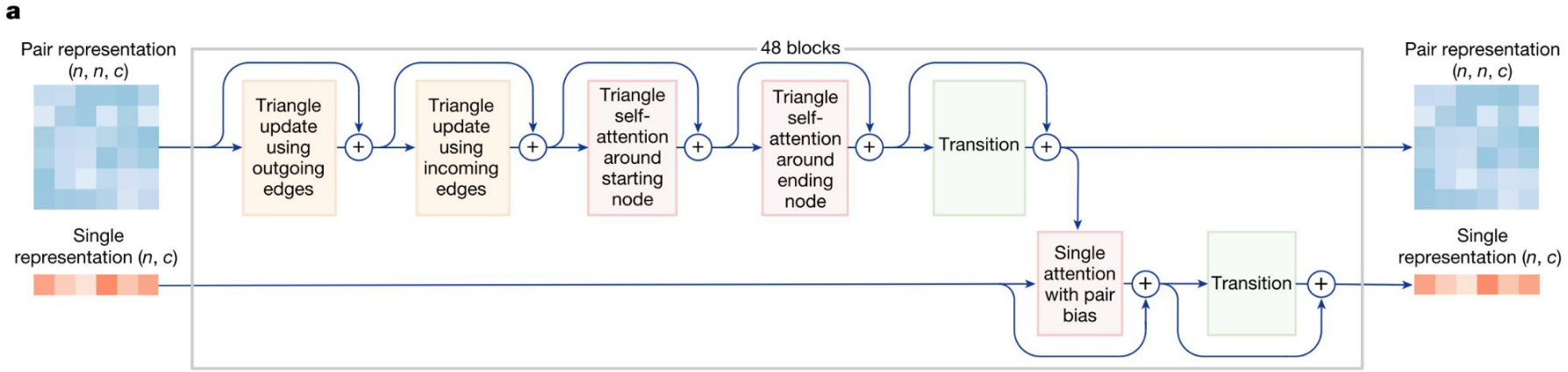
AlphaFold 3



- AF3 architecture:
  - Modify the MSA integration
  - Evoformer changed to Pairformer
  - Do not predict rotation/translation matrices but diffusion module

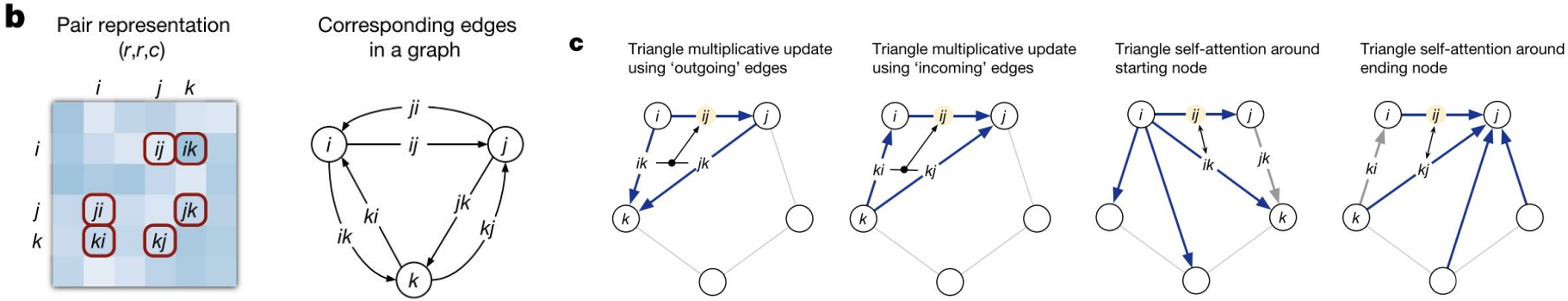
AlphaFold 3





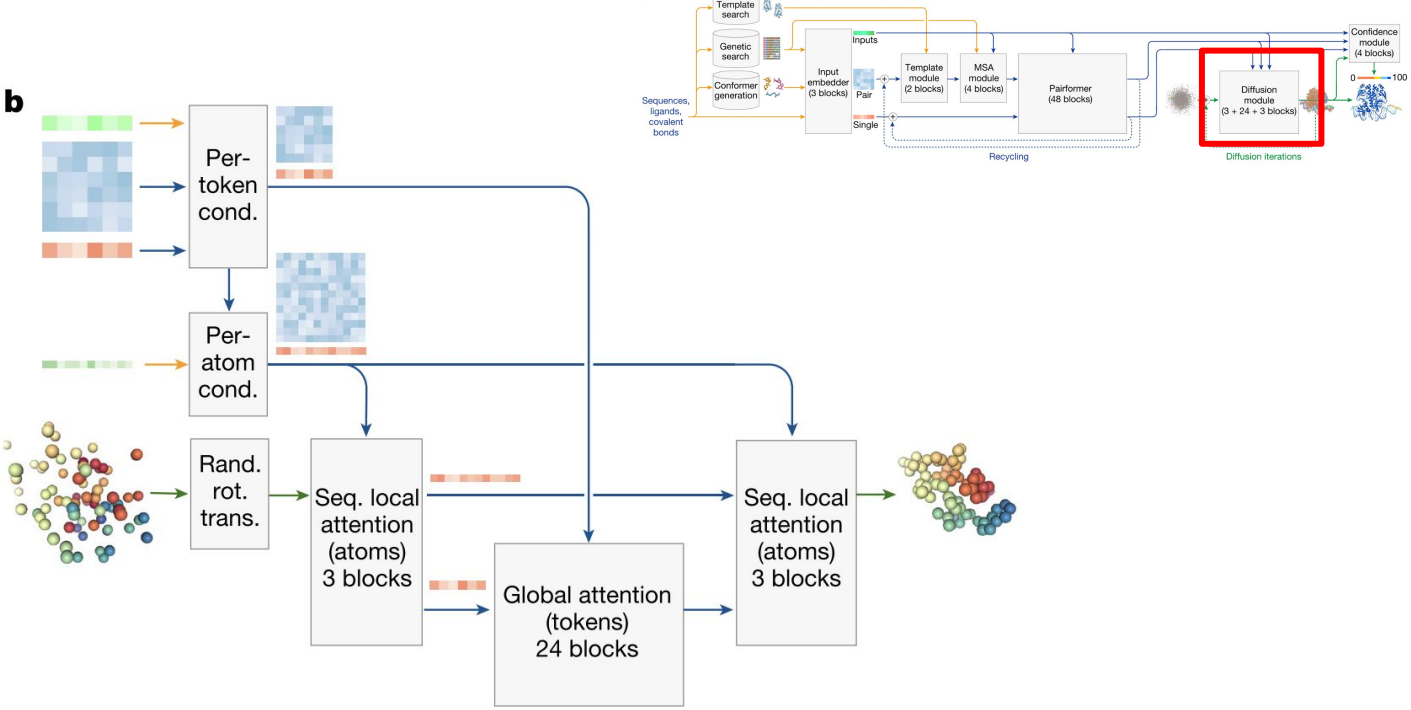
- Pairformer:
  - Triangle update using corresponding graph (same as AF 2)
  - No more MSA representation at this step but pair representation

AlphaFold 3



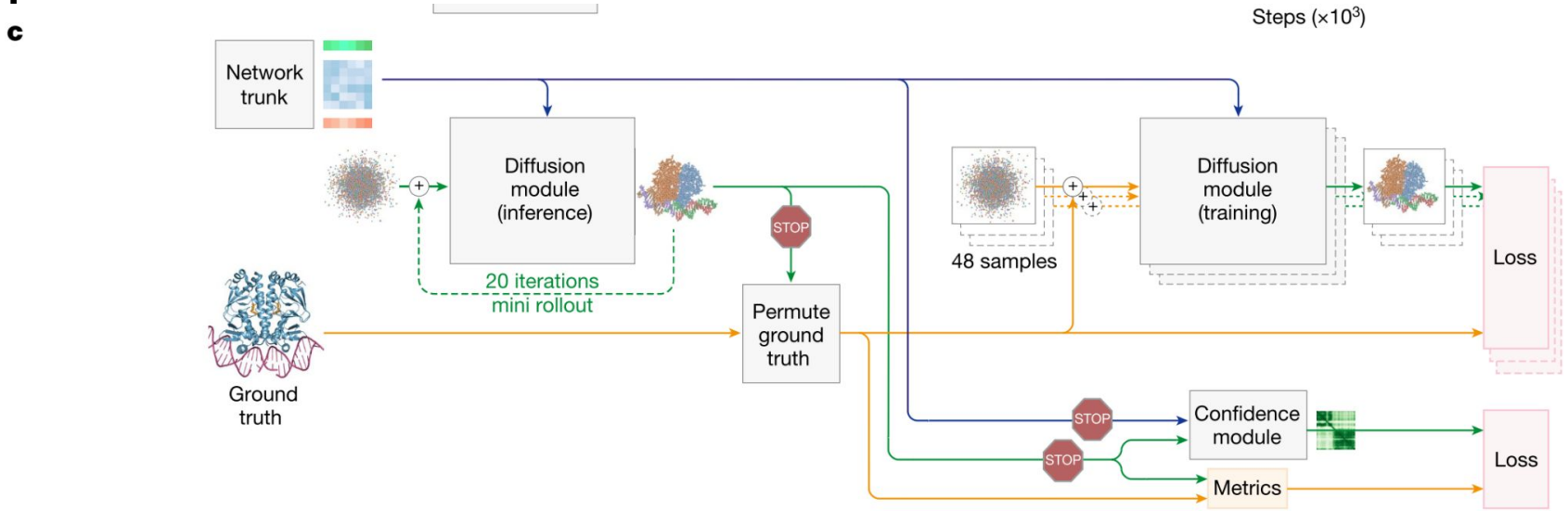
- Triangle multiplicative updates
- Convert pair representation to edges of graph
- Updates only on given type of edges

AlphaFold 3



- Diffusion module
- Conditional attention used on the different modalities
- Go from atom representation to tokens to atom again (full-atom representation)

AlphaFold 3



- Global training process
- Loss for the diffusion and loss for the global model
- Training is done in three steps:
  - Initial training with 384 tokens
  - Fine tuning with 640 tokens
  - Final tuning with 768 tokens

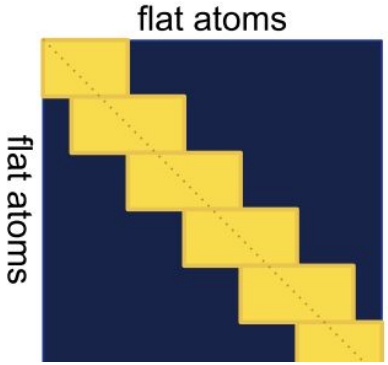
AlphaFold 3

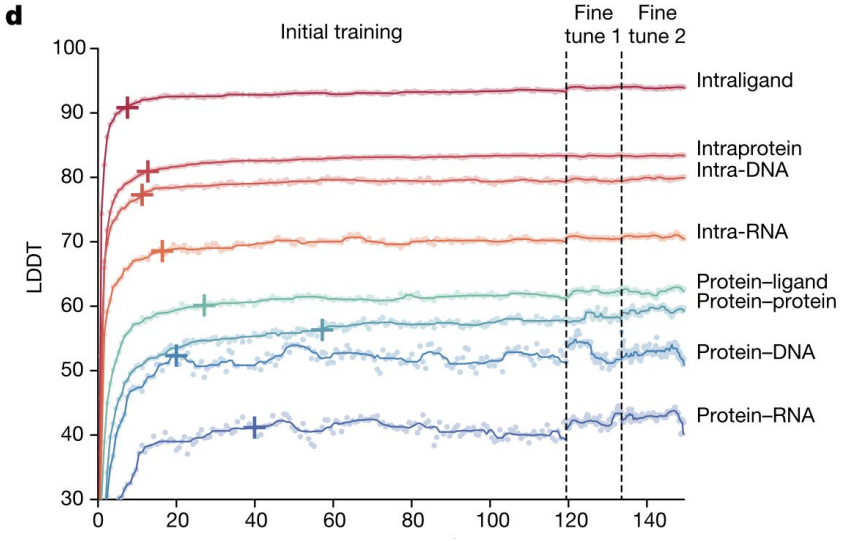
- AF3 uses **5 datasets** with the training process:
  1. **Sample a dataset** according to the weights
  2. **Draw an example** from the dataset
  3. **Sample a structural crop** from the example (with cropping strategy: contiguous, spatial or spatial interface)
- Distillation sets obtained from self prediction or AlphaFold 2 predictions

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]

AlphaFold 3

- Sequence-local atom attention: reduce cost of training
- Diffusion module:
  - Train a denoiser to remove Gaussian noise from the positions of all heavy atoms
  - No geometrical biases involved
  - Create 48 random versions of the input by applying random translation/rotation
- Combined loss (frame alignment, distogram, confidence head, etc)

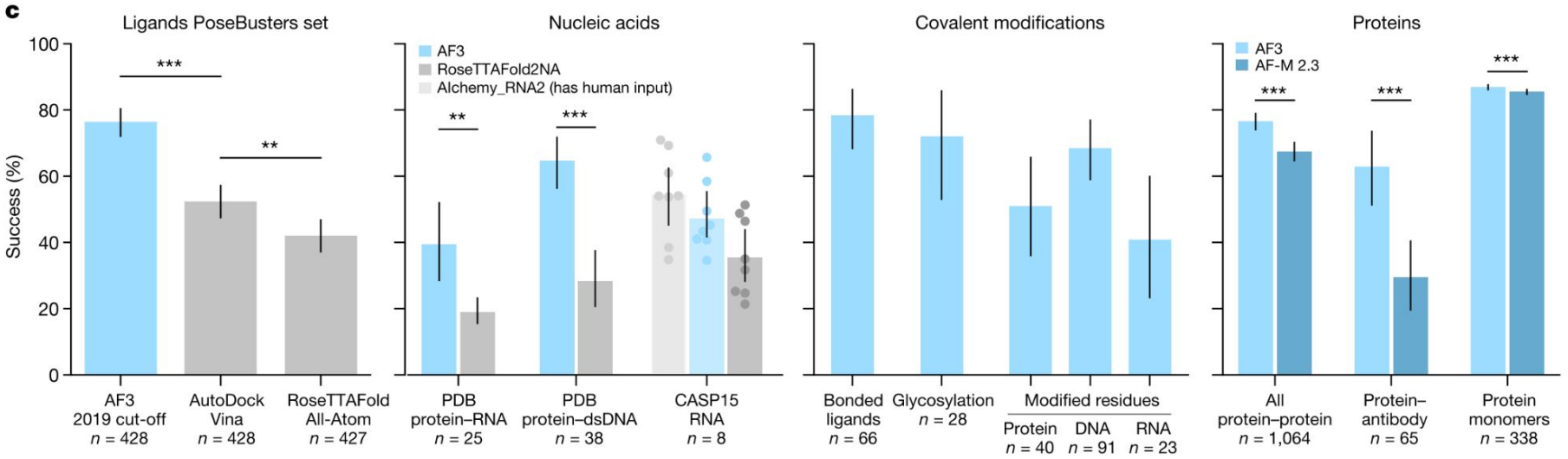




$$\mathcal{L}_{\text{loss}} = \alpha_{\text{confidence}} \cdot (\mathcal{L}_{\text{plddt}} + \mathcal{L}_{\text{pde}} + \mathcal{L}_{\text{resolved}} + \alpha_{\text{pae}} \cdot \mathcal{L}_{\text{pae}}) + \alpha_{\text{diffusion}} \cdot \mathcal{L}_{\text{diffusion}} + \alpha_{\text{distogram}} \cdot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

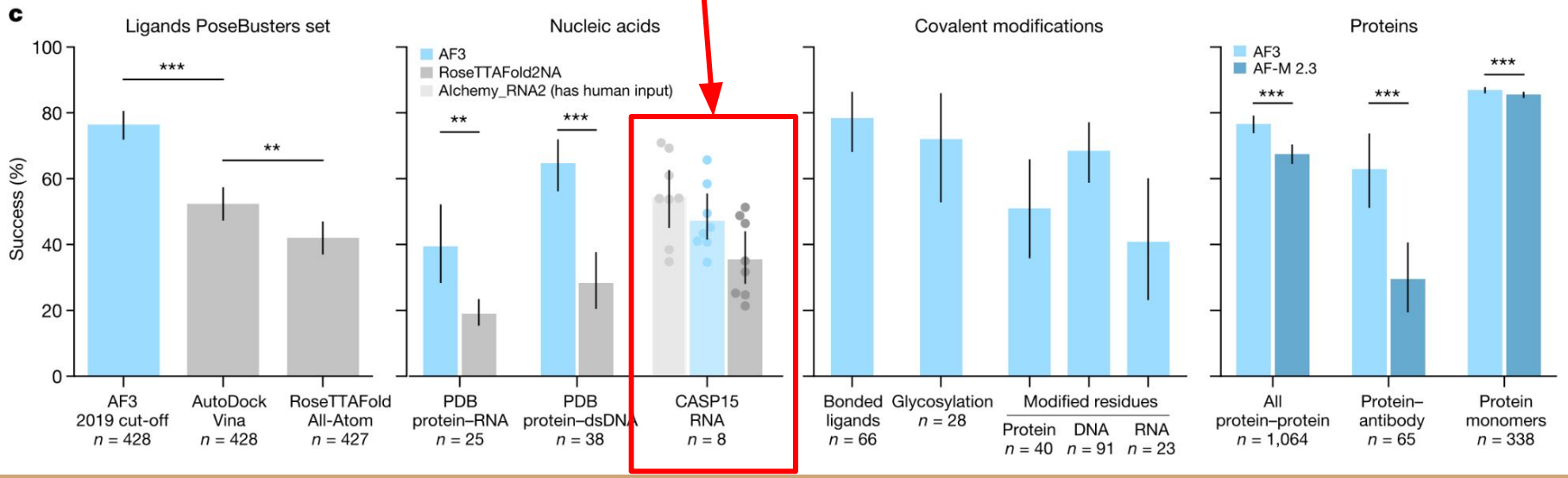
AlphaFold 3

- AF3 has competitive results on different tasks: ligands, RNA, docking, proteins, etc.
- Can also input structures up to 5000 tokens (1 token = 1 residue: nucleic acid or amino acid)



AlphaFold 3

What about the results for RNAs?



III. AlphaFold 3 for RNAs?

AlphaFold 3 for RNAs

We did a work on benchmarking AF3 to state-of-the-art methods for RNAs

New Results

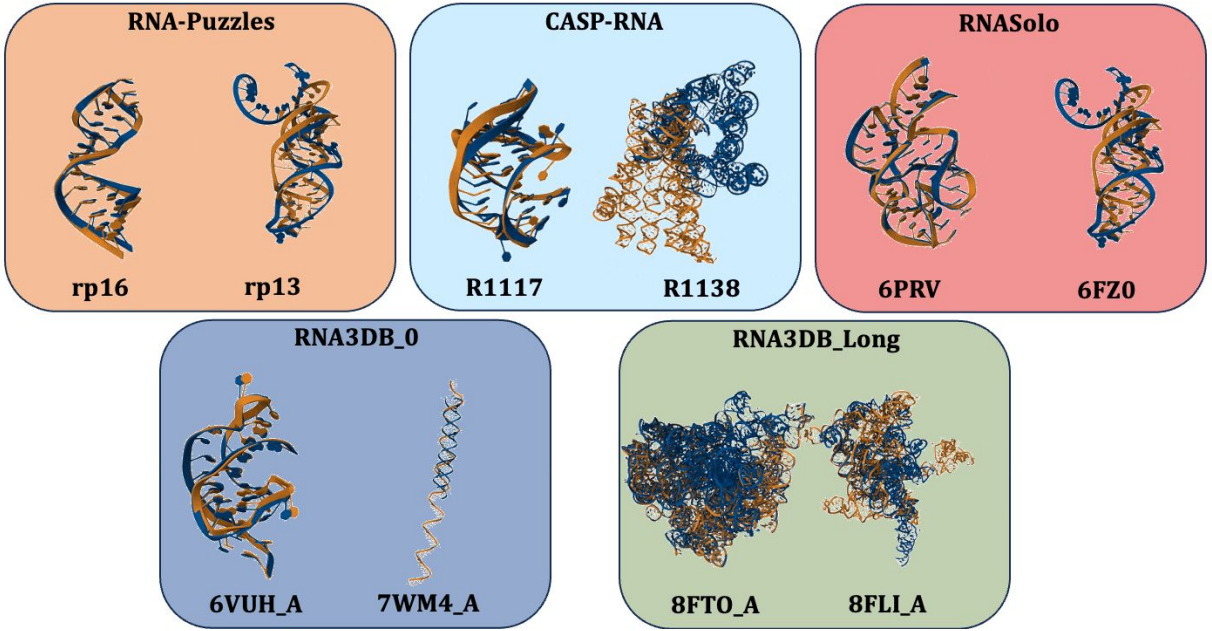
Has AlphaFold 3 reached its success for RNAs?

Clément Bernard, Guillaume Postic, Sahar Ghannay, Fariza Tahi

doi: https://doi.org/10.1101/2024.06.13.598780

AlphaFold 3 for RNAs

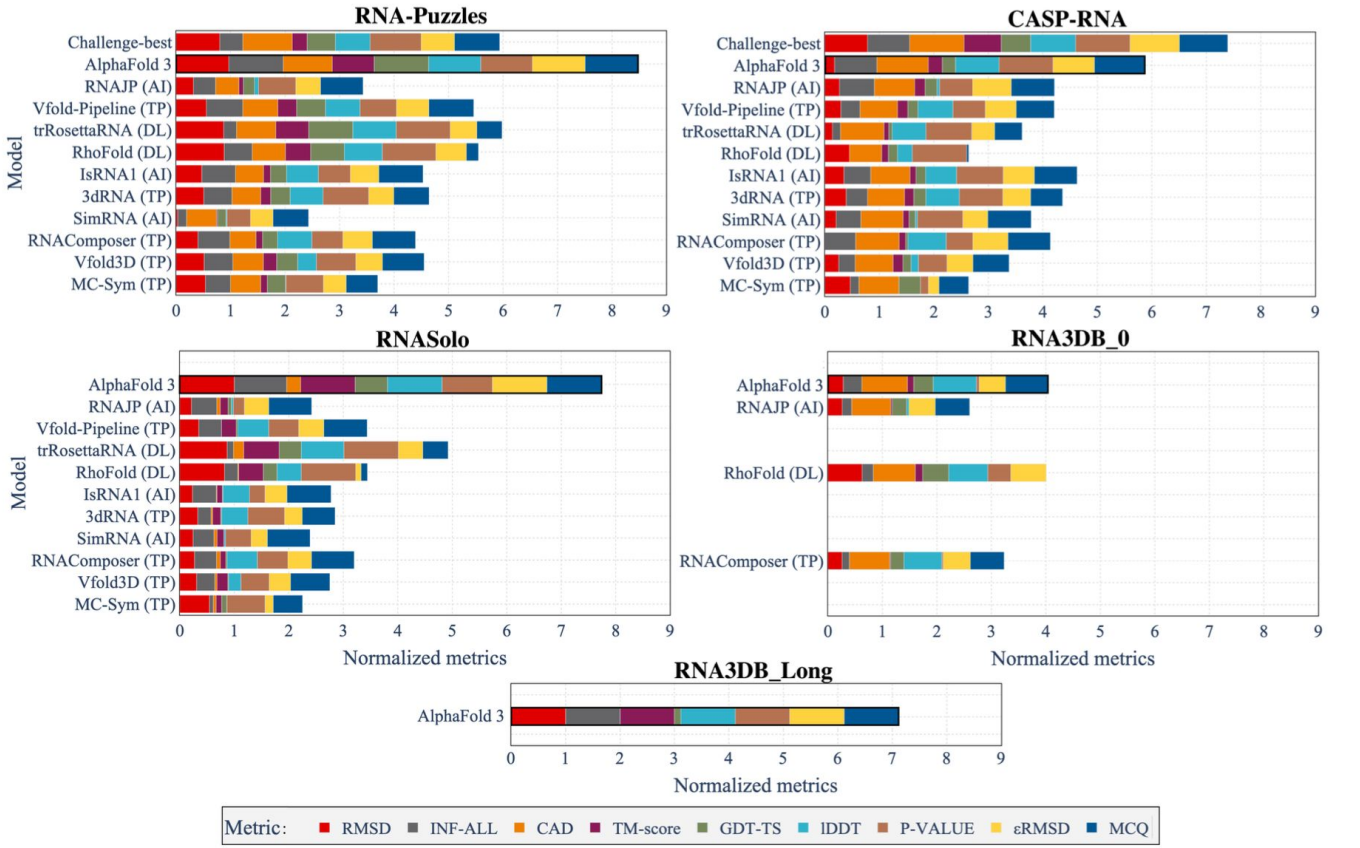
- Benchmark of **ten existing predictive methods** for RNA 3D structure prediction
- Use of **five different datasets**
- More than **300 predictions of AlphaFold 3** made (online …)



AlphaFold 3 for RNAs

- Each metric is normalised to be better when close to 1
- Consider the sum of cumulative metrics

-> the higher, the better

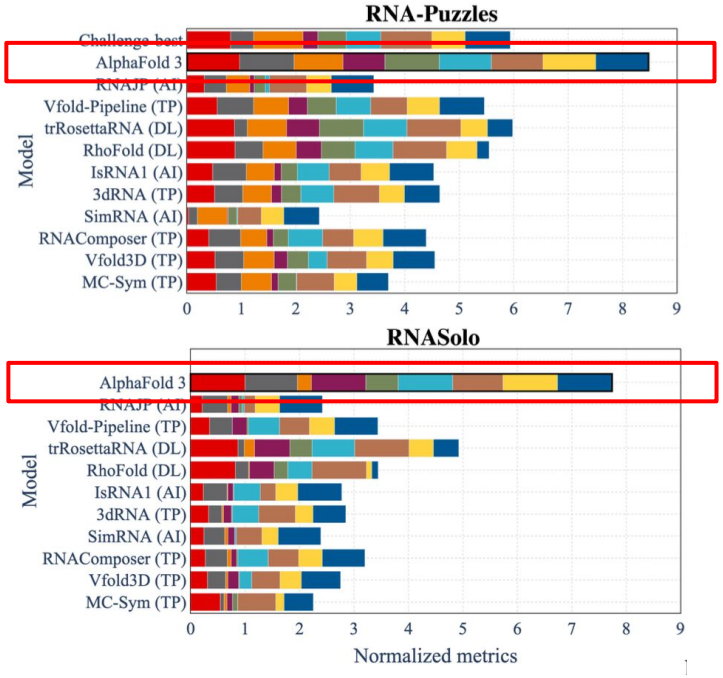


Sum of normalised metrics to assess RNA 3D structural quality

AlphaFold 3 for RNAs

Main results

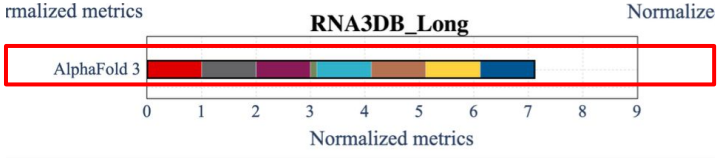
- AF3 has **competitive results**, outperforming state-of-the-art methods on two datasets



AlphaFold 3 for RNAs

Main results

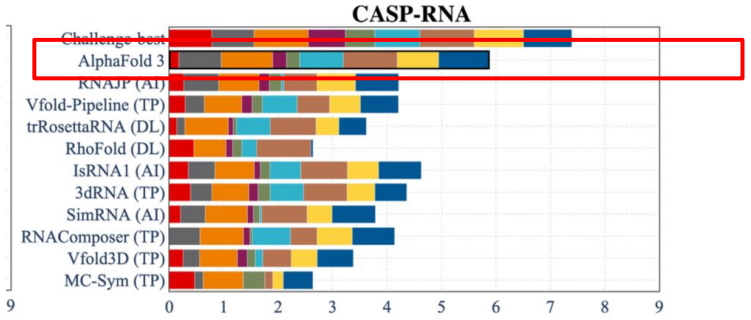
- AF3 has **competitive results**, outperforming state-of-the-art methods on two datasets
- Achieve very good results for **long RNAs** (higher than 1000 nt)



AlphaFold 3 for RNAs

Main results

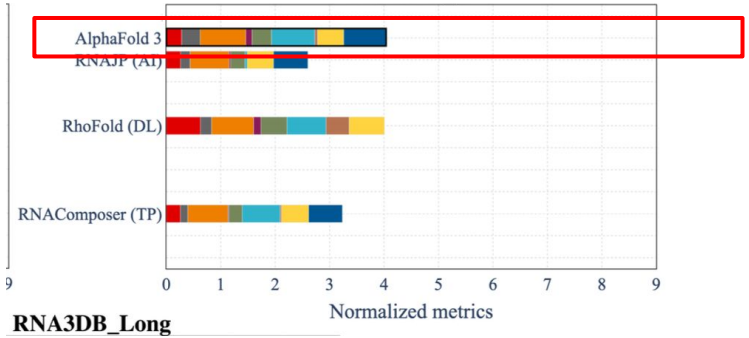
- AF3 has **competitive results**, outperforming state-of-the-art methods on two datasets
- Achieve very good results for **long RNAs** (higher than 1000 nt)
- Is outperformed by **human-guided solutions**



AlphaFold 3 for RNAs

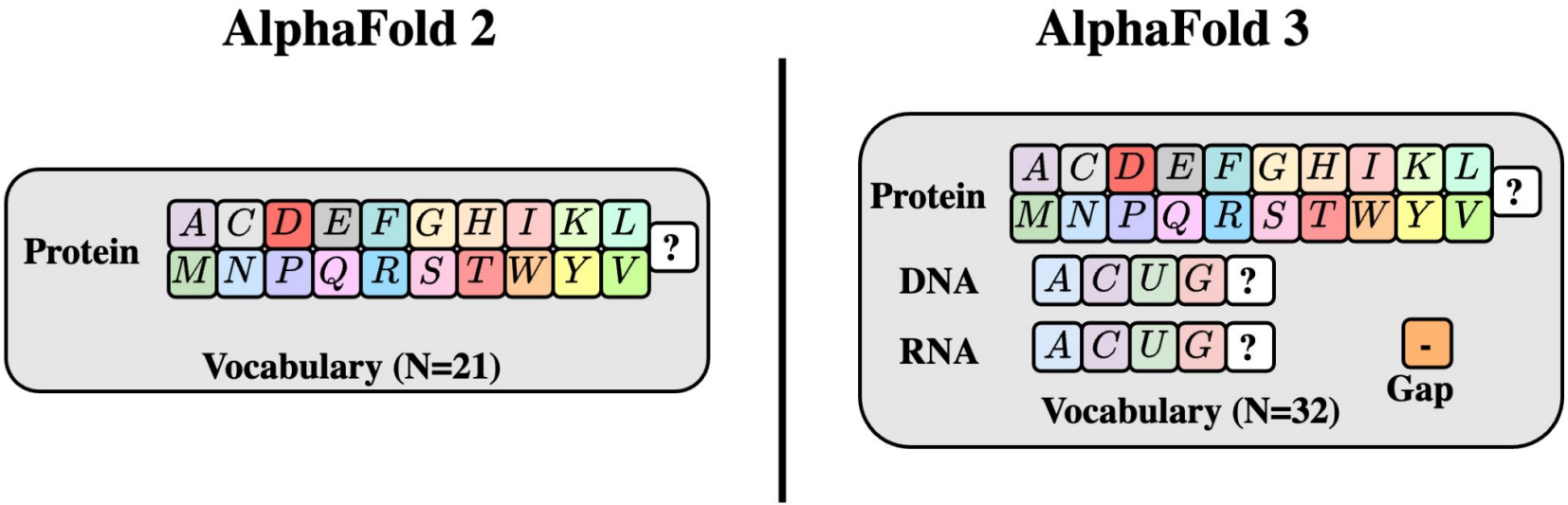
Main results

- AF3 has **competitive results**, outperforming state-of-the-art methods on two datasets
- Achieve very good results for **long RNAs** (higher than 1000 nt)
- Is outperformed by **human-guided solutions**
- Bad results for **orphan structures** (structure without any known RNA families)



IV. Summary AF2 vs AF3

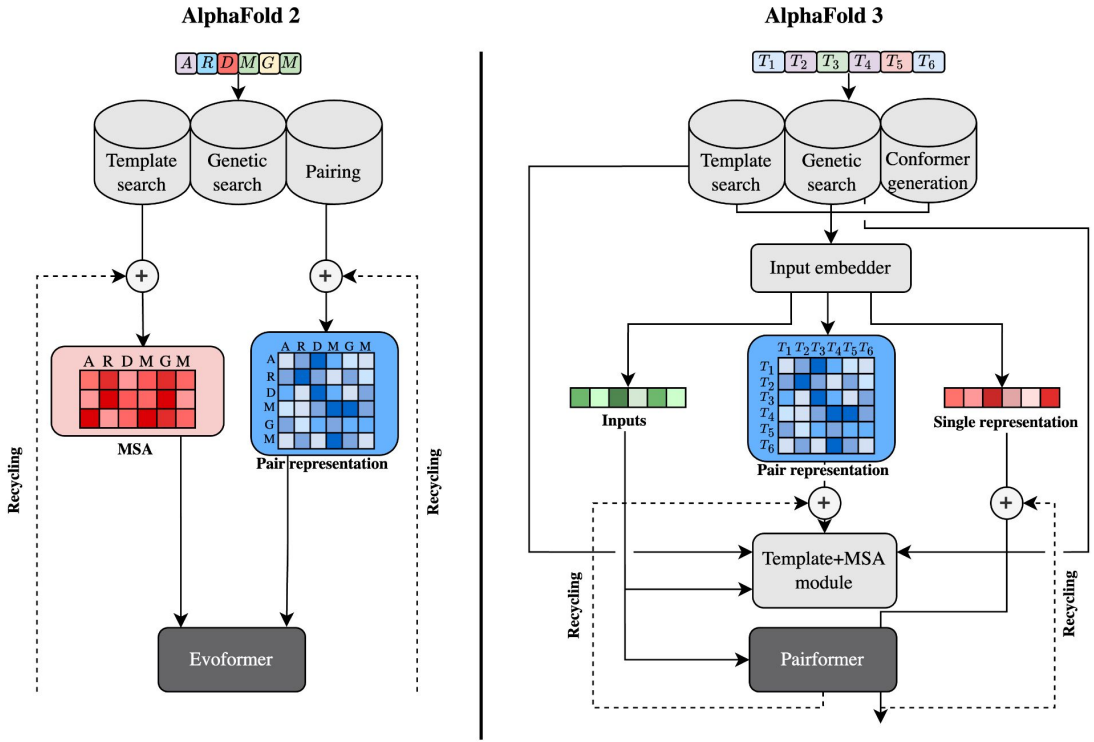
AlphaFold 2 vs AlphaFold 3



Difference of vocabulary: add tokens for DNA and RNA

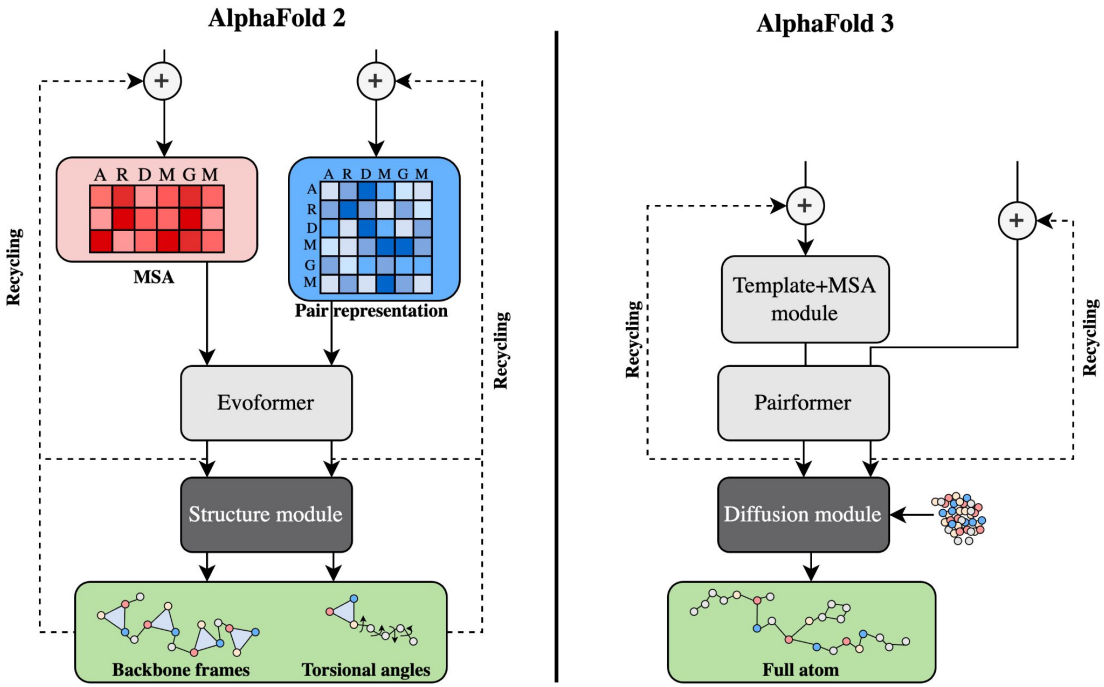
AlphaFold 2 vs AlphaFold 3

- **MSA integration** that has **less impact** compared to AF 2
- **Evoformer** changed to **Pairformer**



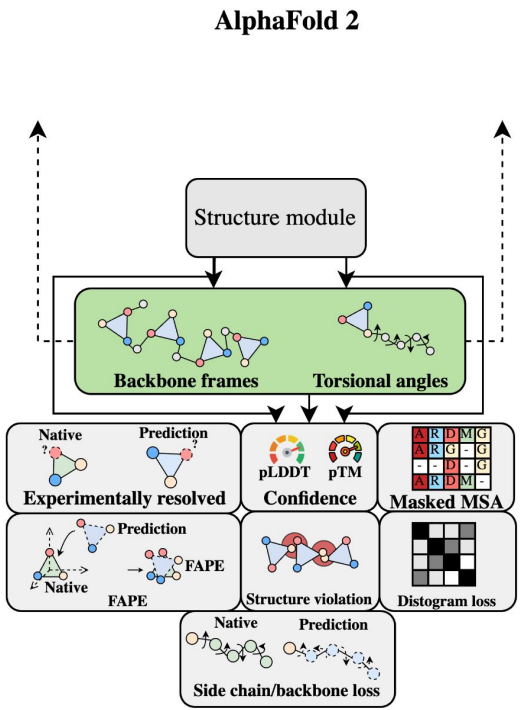
AlphaFold 2 vs AlphaFold 3

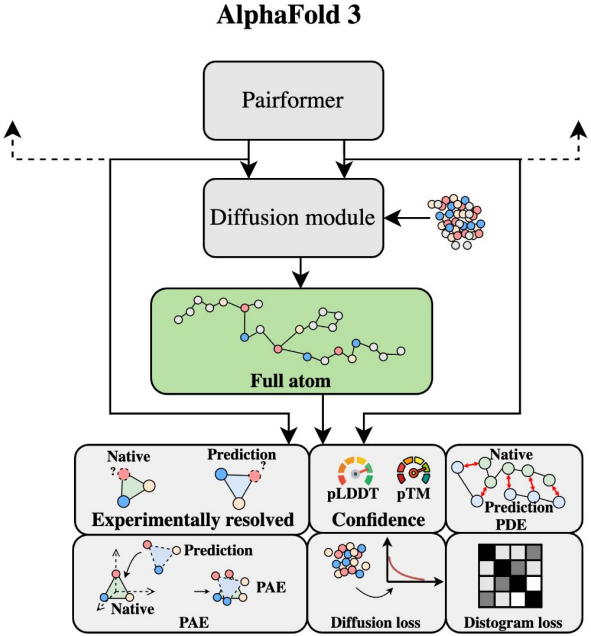
- **Structure module removed** to have a generative module with the use of **diffusion**
- Output directly **full atom positions** instead of backbone + dihedral angles



AlphaFold 2 vs AlphaFold 3

- Confidence head differs:
  - Use **diffusion loss**
  - Don’t have **masked MSA**
  - No **explicit structure violation score**
- Still predict:
  - **Experimentally resolved** score
  - **pLDDT, pTM**
  - **Distogram loss**

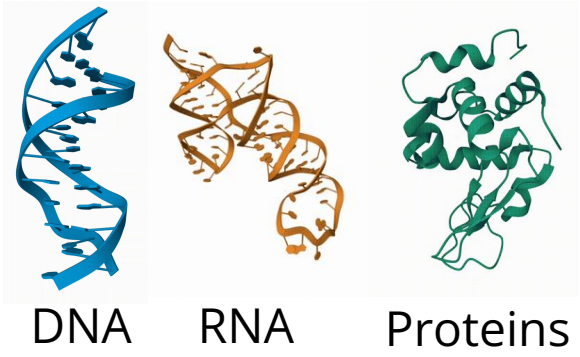




V. Conclusion

Advantages

- **Predict multiple molecules** at the same time
- Produce overall **high quality structures**
- Inputs can have until **5000 tokens**
- **Open-source** since yesterday



Limits

- It is mentioned in the paper the following limits:
  - **Stereochemistry**: non respect to chirality and produce overlapping clashing atoms
  - **Hallucinations** in disordered region
  - Some targets remain **challenging** to predict
  - Do not predict the **dynamic of the folding** process
- What we experienced:
  - Struggle to **unseen RNA families** (orphan)
  - Performances do not **exceed best human-guided** methods

Interested in this research area?

- I have done ~**10 Medium posts** where I discuss around the **RNA 3D structure subject**
- **Tutorials using Python** on how to implement PDB manipulations, visualisation, ...
- Step by step guide: **available for everyone.**
- Link