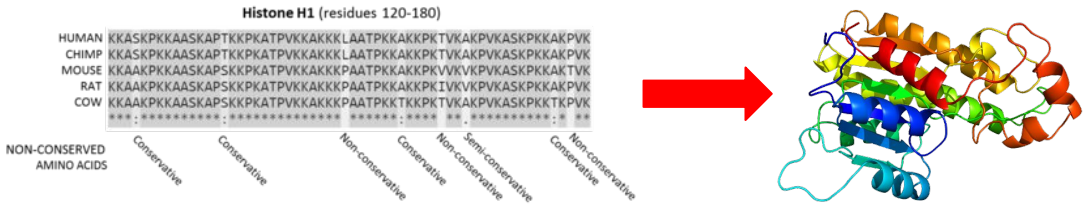
AlphaFold2

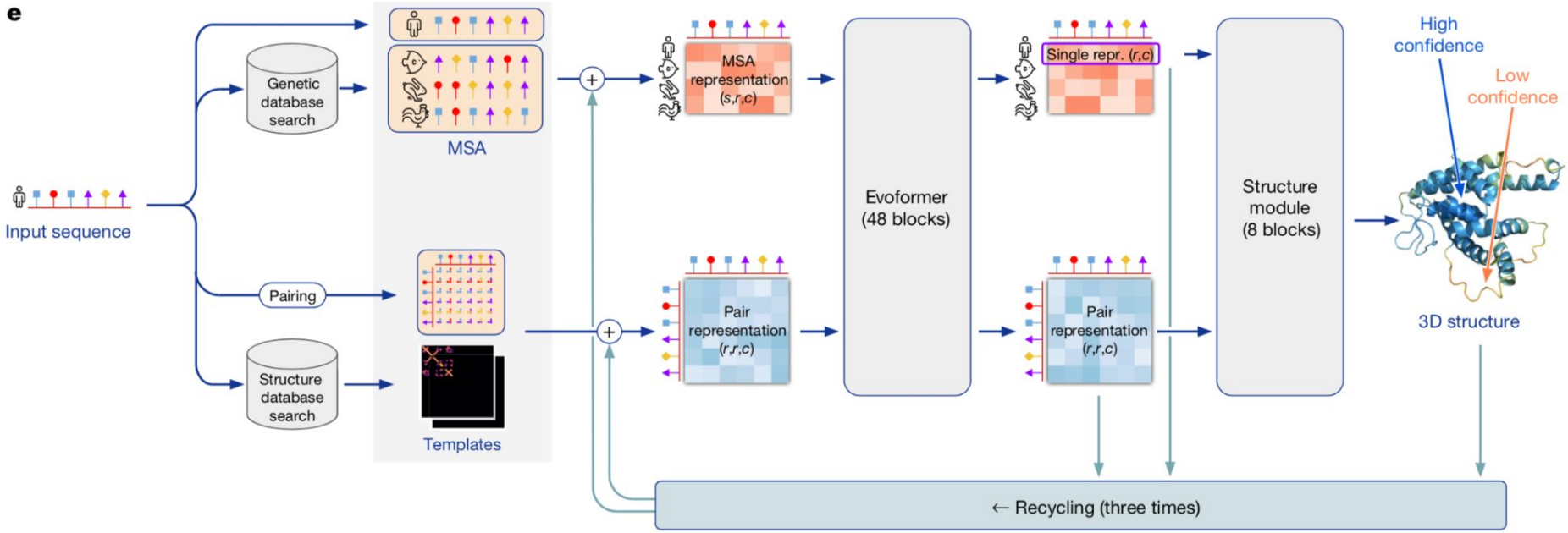
CS182/282, Spring 2022

Protein Structure Prediction

- **Protein structure prediction** is a long-standing goal of computational biology
  - Namely: given a protein sequence, infer its 3D structure
- Two classes of approaches to structure prediction:
  - *Ab initio*- or *de novo*- methods operate solely based on physical principles and the protein sequence
  - *Comparative* methods use previously solved structures to inform the structure of a related sequence
- CASP is a biennial international “competition” for structure prediction algorithms
- AlphaFold2 uses a deep learning (comparative) approach, and greatly outperformed all other models in several categories at CASP14 (2020)
  - Advances in protein structure prediction by AlphaFold2 and RoseTTAFold was [Science’s 2021 Breakthrough of the Year]

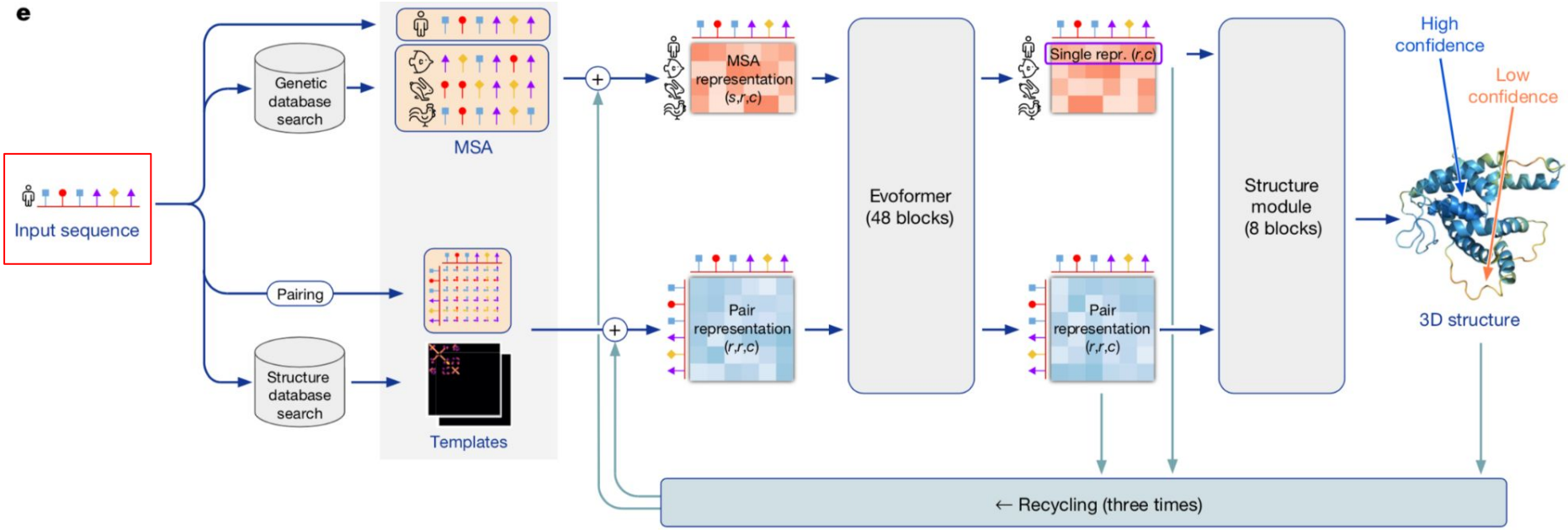


High-level Overview of Architecture



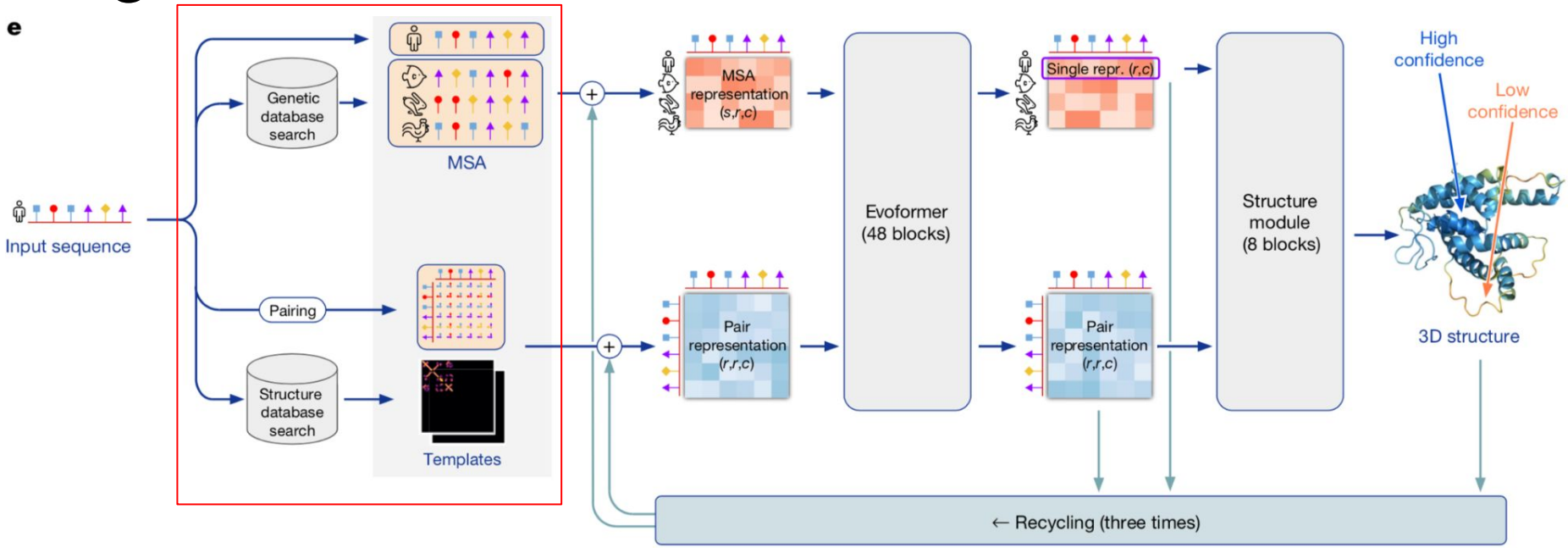
Deep learning uses sequential modules (layers) to progressively extract information (learn) from the input data.

High-level Overview of Architecture



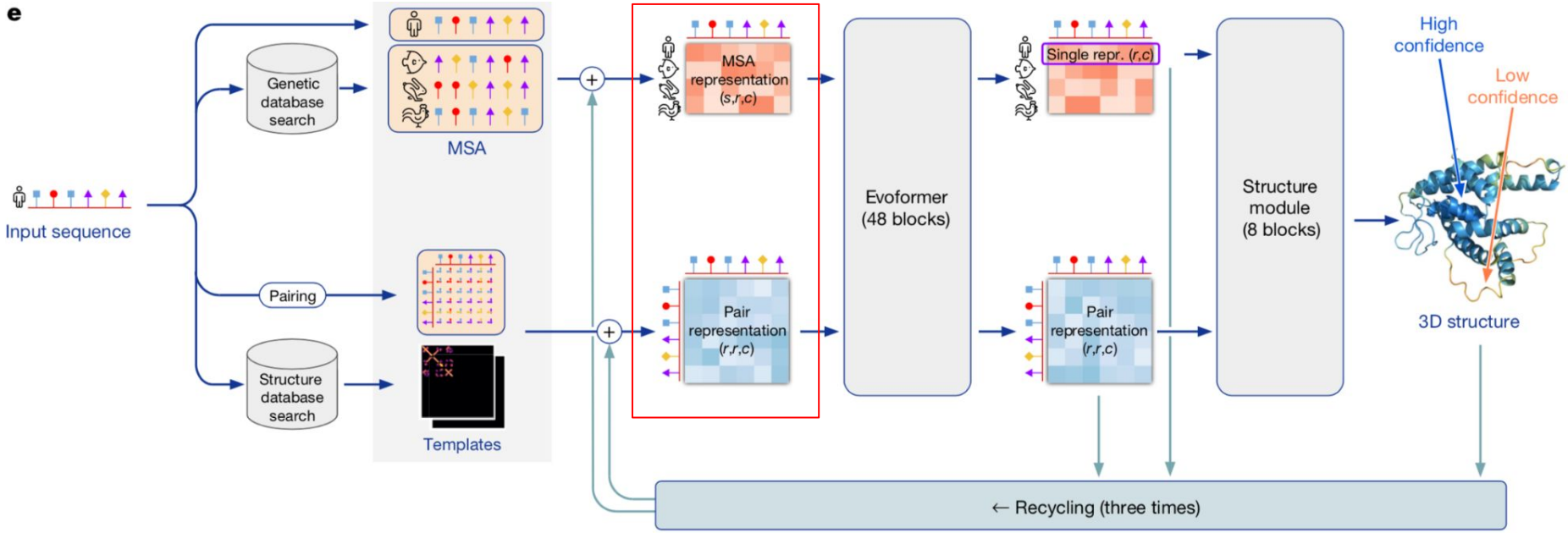
AlphaFold2 takes a protein's primary sequence of amino acids as input.

High-level Overview of Architecture



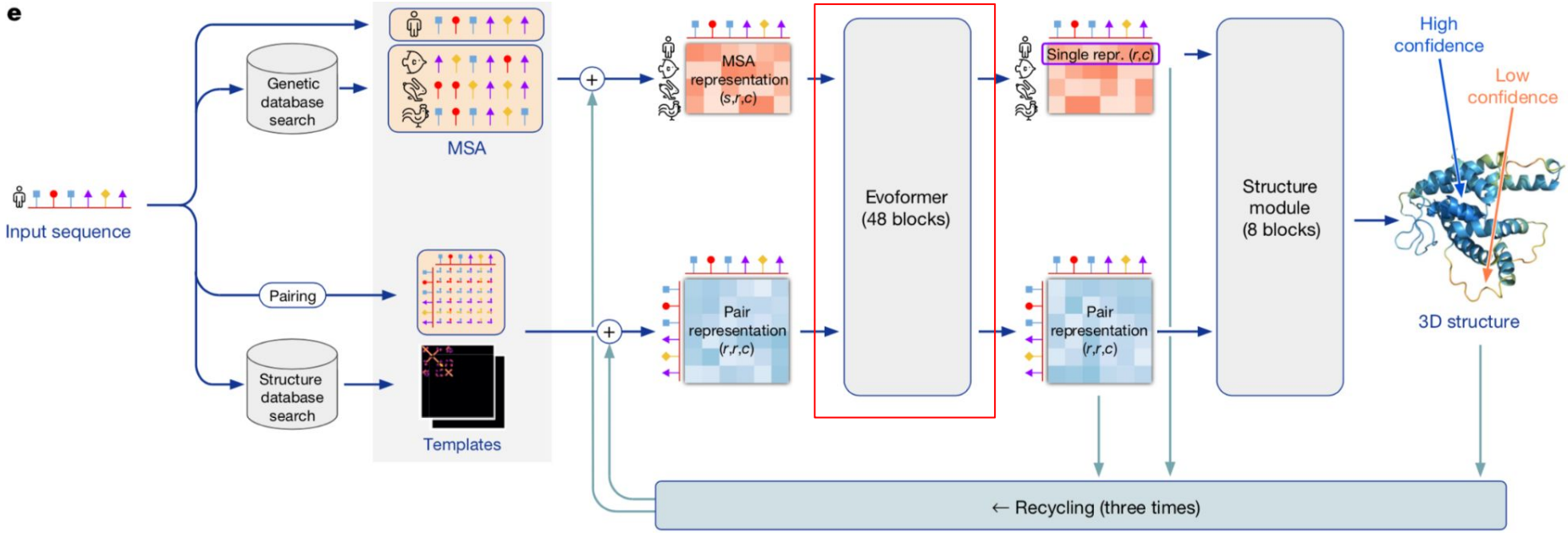
The preprocessing module searches the sequence against protein databases of sequences and known structures.

High-level Overview of Architecture



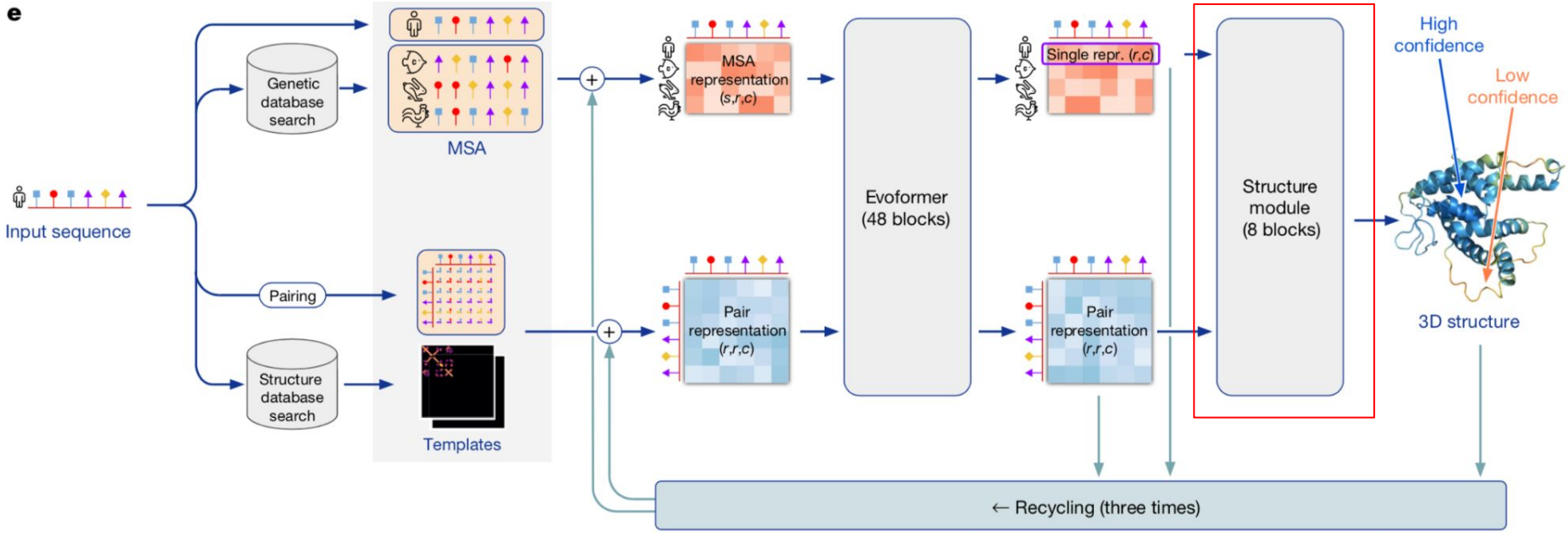
This produces two new types of data: a **multiple sequence alignment** with homologous sequences and a **matrix representation of pairwise interactions** between residues.

High-level Overview of Architecture



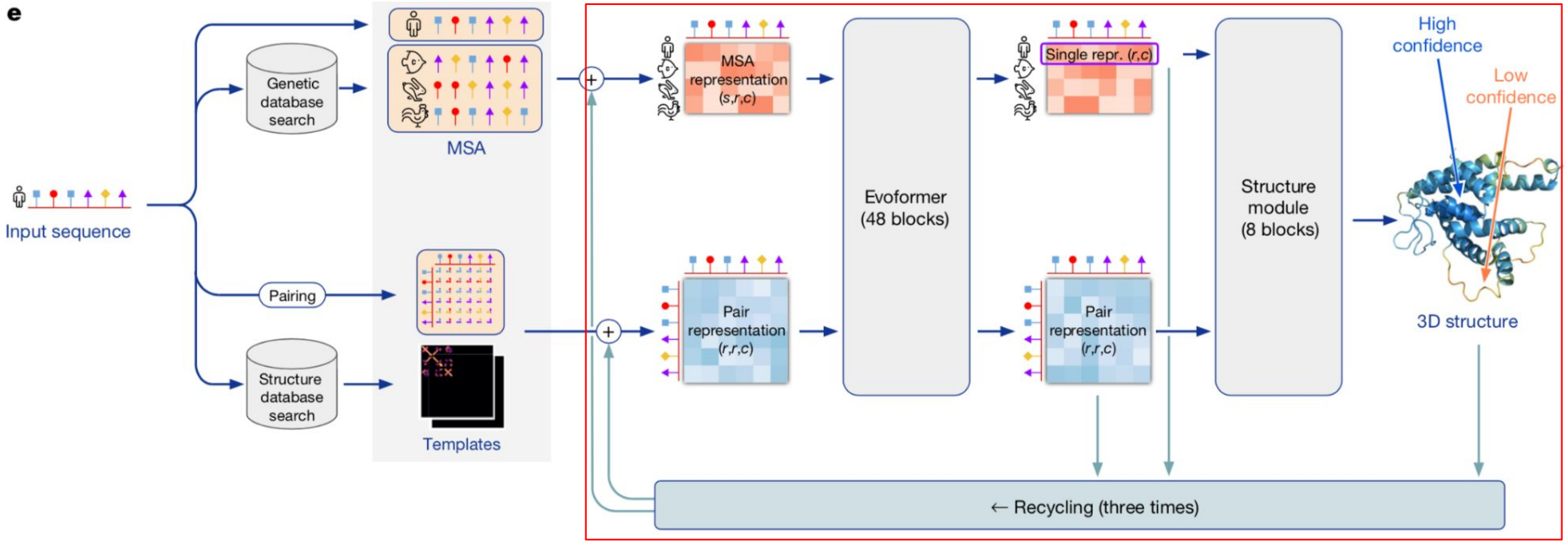
Both the MSA and pair representation are passed into a specialized module for exchanging information between both forms of data and refining them based on physical principles and any relationships it learns.

High-level Overview of Architecture



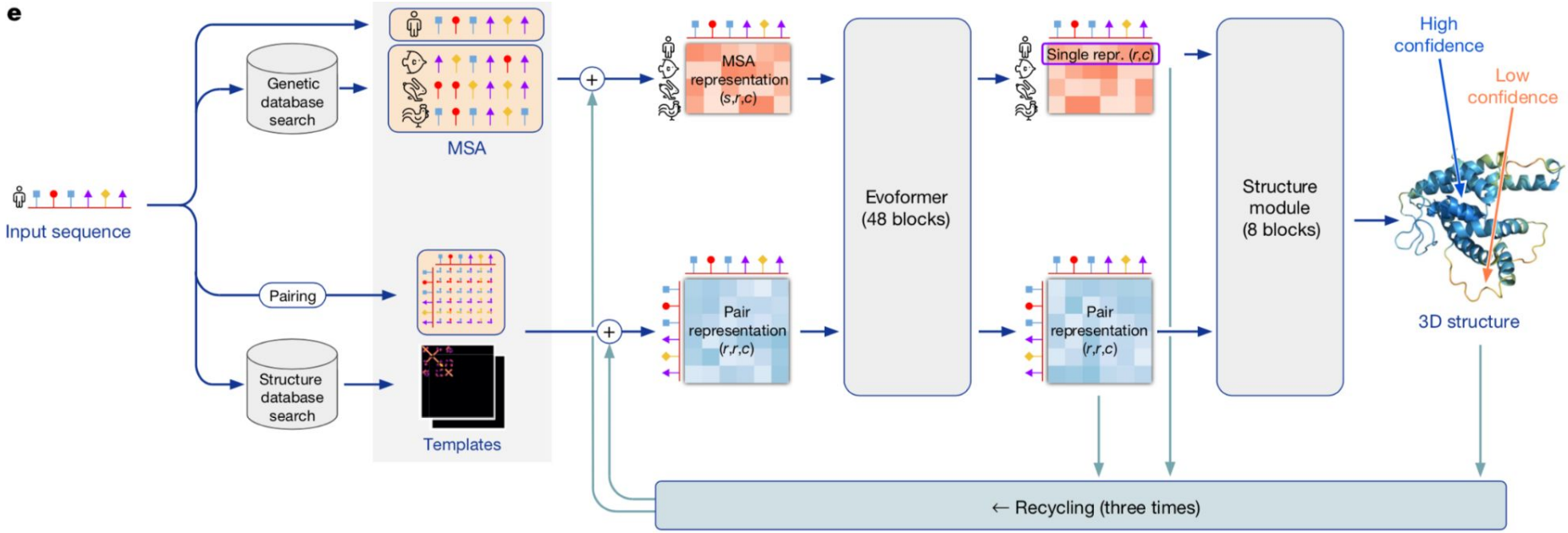
The data is then passed into a different module which converts those representations into a 3D structure (defined by translations and rotations).

High-level Overview of Architecture



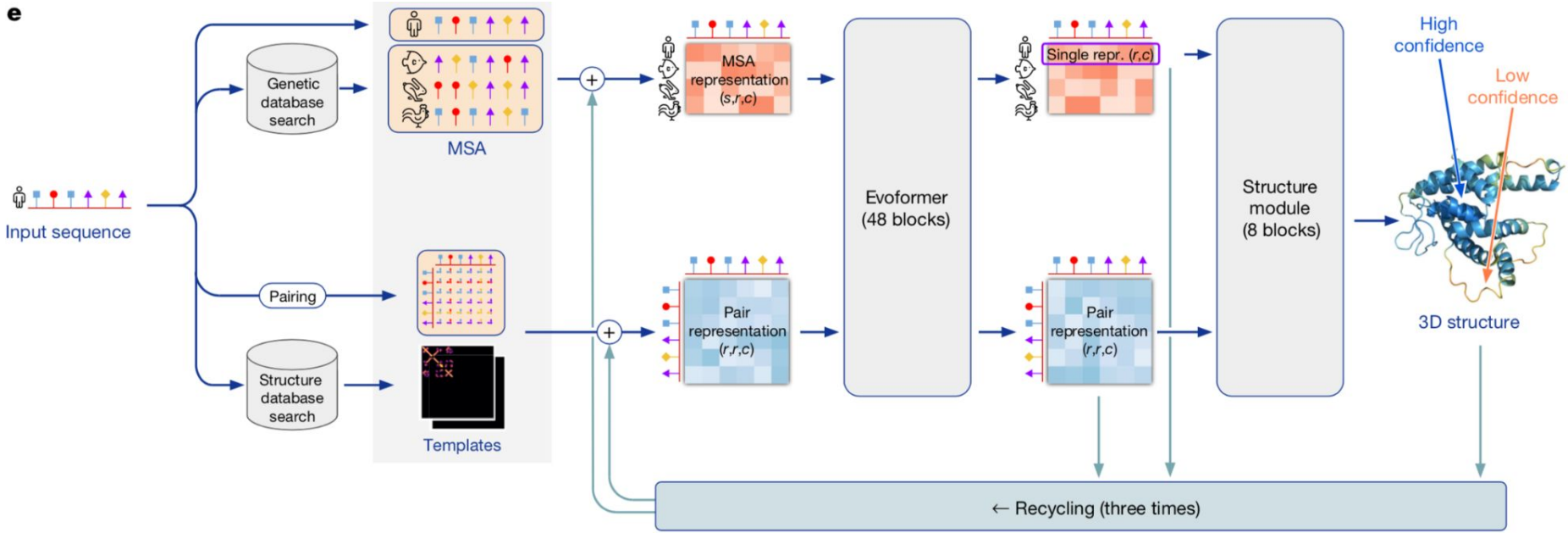
The outputs are passed through the same Evoformer and Structure modules to iteratively improve the structure.

High-level Overview of Architecture



How do we know how to update each module's weights to "improve" accuracy?

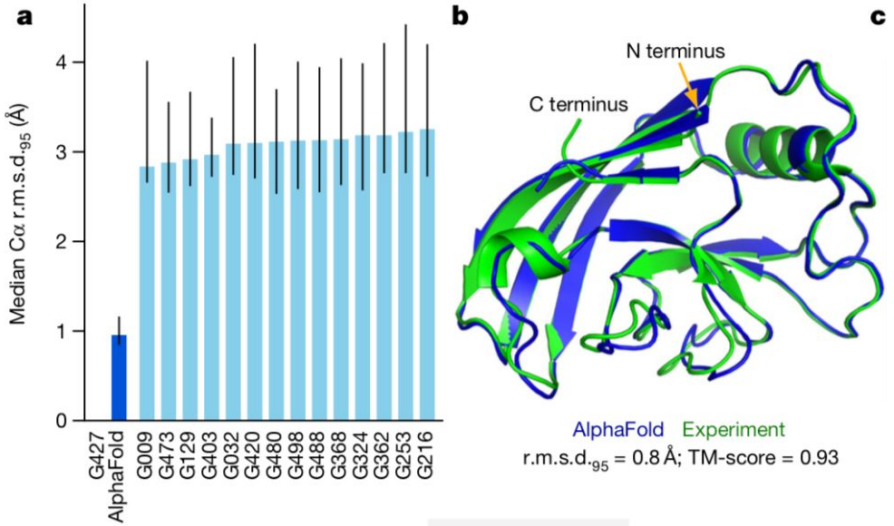
High-level Overview of Architecture



How do we know how to update each module's weights to "improve" accuracy?

We apply updates which minimize the loss function (AlphaFold2 uses several different loss functions to refine specific parts of their model! But the final loss computes the difference between corresponding atoms in the predicted and actual structures)

Performance

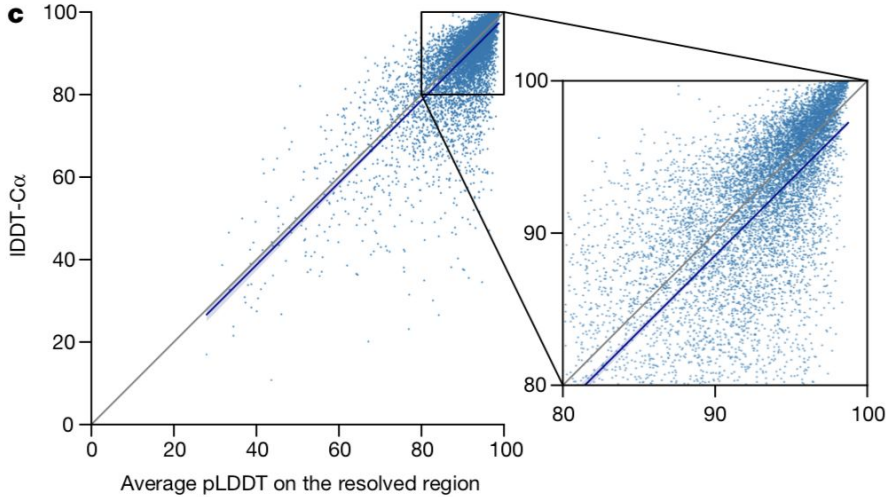


AlphaFold2 achieves atomic-level accuracy (on-par with experimental data), with median distance/atom of less than one angstrom

On the right is T1049, a protein with experimentally verified structure that was not used in AlphaFold2 training

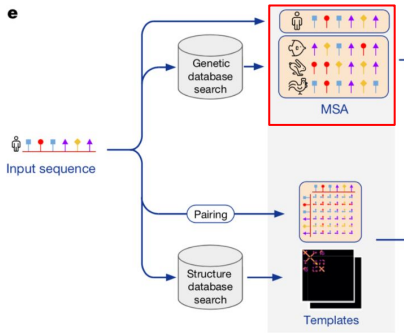
AlphaFold2 also includes a metric of uncertainty in predicted structure: the predicted local-distance difference test (plDDT)

This uncertainty predicts an actual protein structure scoring metric



Preprocessing Modules - Multiple Sequence Alignment

- Multiple Sequence Alignment (MSA) is another seminal problem in computational biology.
  - Alignment of two sequences can be performed in O(nm) time by the Needleman-Wunsch (global) or Smith-Waterman (local) algorithms.
  - But applying an analogous procedure for the optimal alignment of more than two sequences requires $O(n^k)$ time, exponential in the number of sequences, which is computationally intractable.
- Existing approaches to multiple sequence alignment apply heuristics to approximate the optimal solution in a feasible amount of time.
  - Profile HMMs for alignment rather than dynamic programming
  - Guide trees defining sequences of merges for pairwise alignments
  - Most recently, transformers (a specific deep learning architecture)



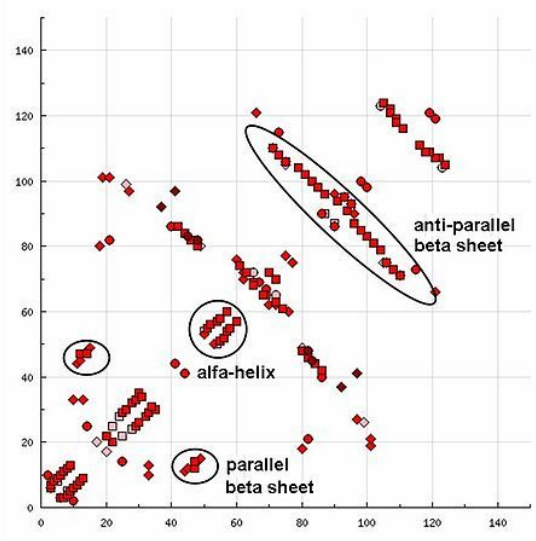
Preprocessing Modules - Multiple Sequence Alignment

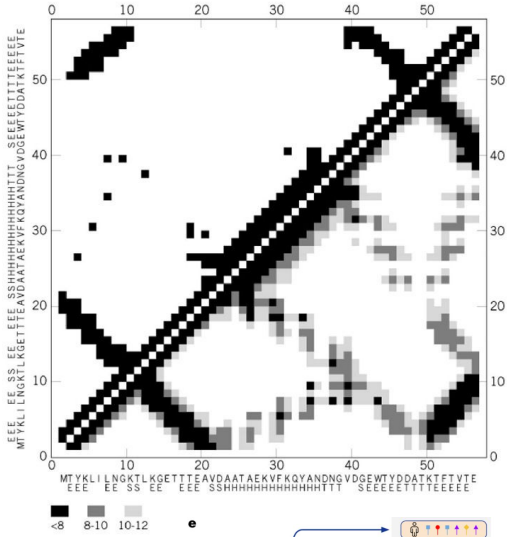
AlphaFold2 uses FAMSA for MSA:

- Compute pairwise similarities
  - Similarity = longest common subsequence / indel distance
- Determine a guide tree
  - SLINK = single-linkage hierarchical clustering
  - Can be performed quickly and incrementally with partial similarity matrix
- Merge profiles progressively according to the guide tree
  - Parallelizable and memory-efficient due to sequences in gapped representations
  - Gap corrections and scaling enables accurate scoring for complex relations between multiple sequences and gap types (terminal/extension/opening penalties)
- Iterative refinement of the final profile
  - For small protein families, can correct gap misalignments using column-wise scans

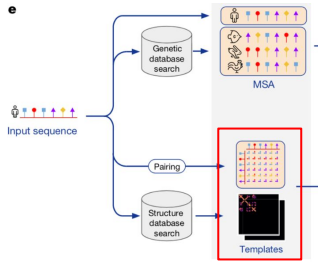
Preprocessing Modules - Contact Maps

- Contact maps encode distances between pairs of residues
- Entries can be computed using distance thresholds or as continuous distances





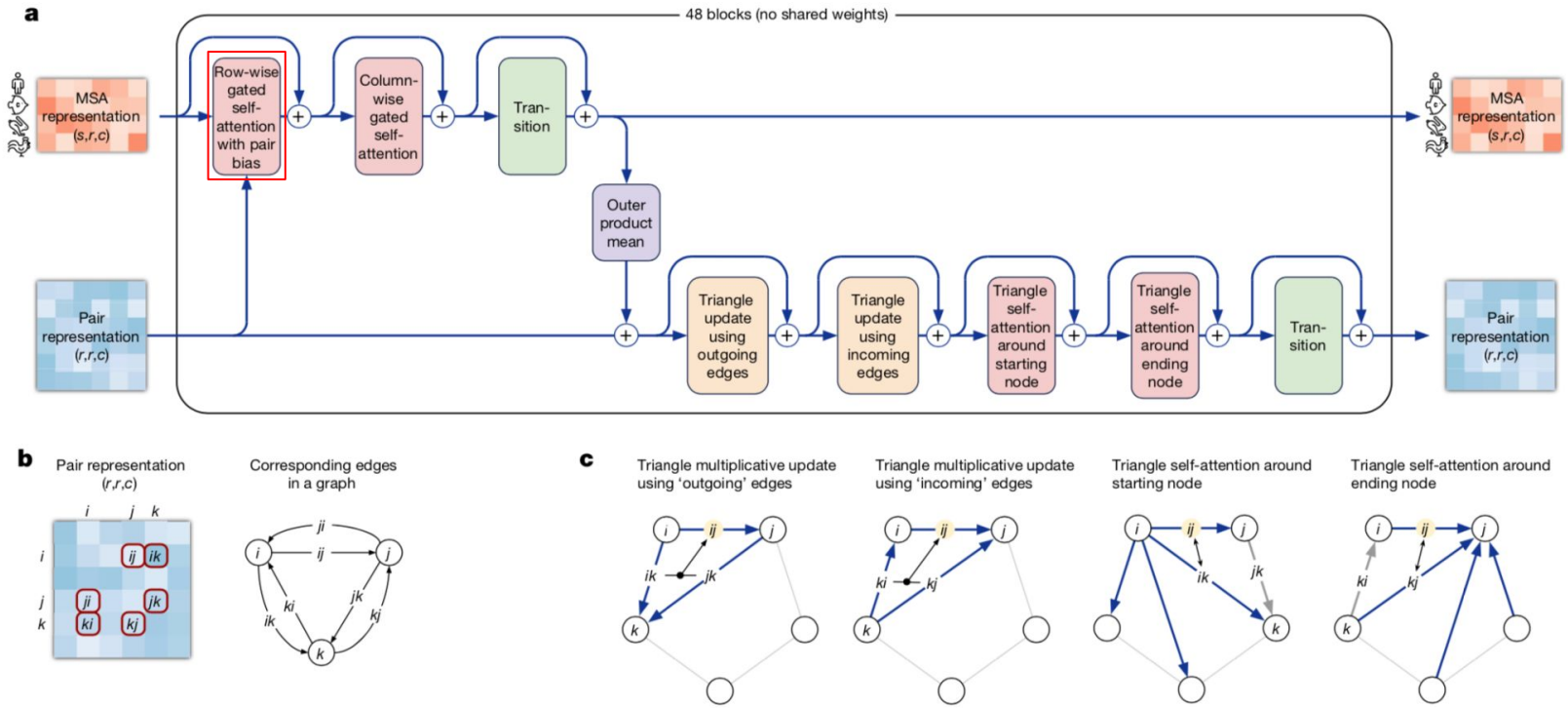
- AlphaFold2 searches for contact maps (distograms), along with other pairwise features to incorporate as inputs into their main Evoformer module



Primary Modules - Introduction to Attention

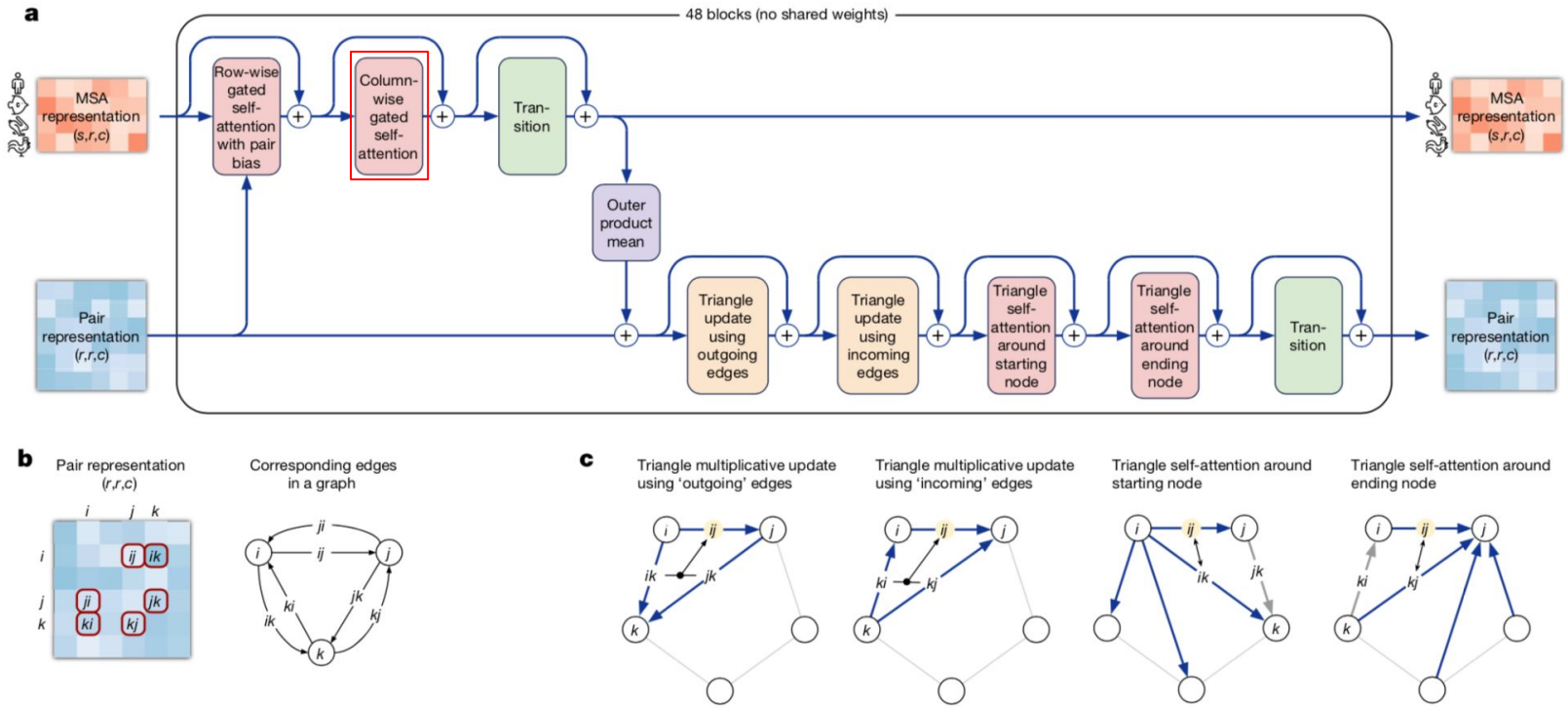
- General principle: want to extract information from a large input with complex structure, for example, an image
- Humans do this by ‘paying attention’ to particular regions of the image
    - Perhaps we’re searching an image for a friend: we would focus our attention on faces in the image
- Computers accomplish the same task by converting the target information (our friend) into “query” vectors and converting the input (an image) into “key” vectors
- Note: we can apply this same strategy to the input itself to encode a more “distributed” representation

Primary Modules - Evoformer



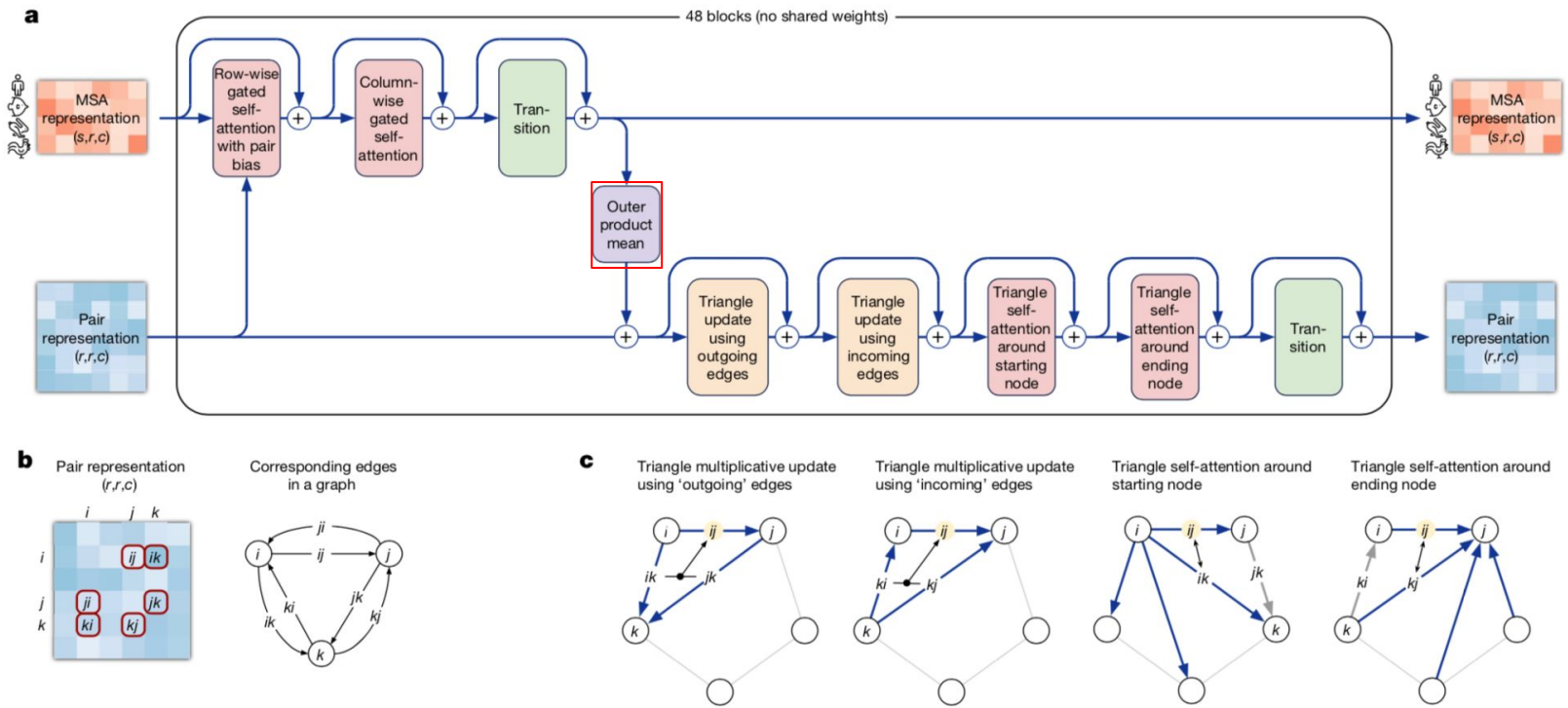
The row-wise self-attention submodule identifies pairs of 'interacting' or correlated amino acids in the aligned proteins, and focuses 'attention' using information from the pair representation

Primary Modules - Evoformer



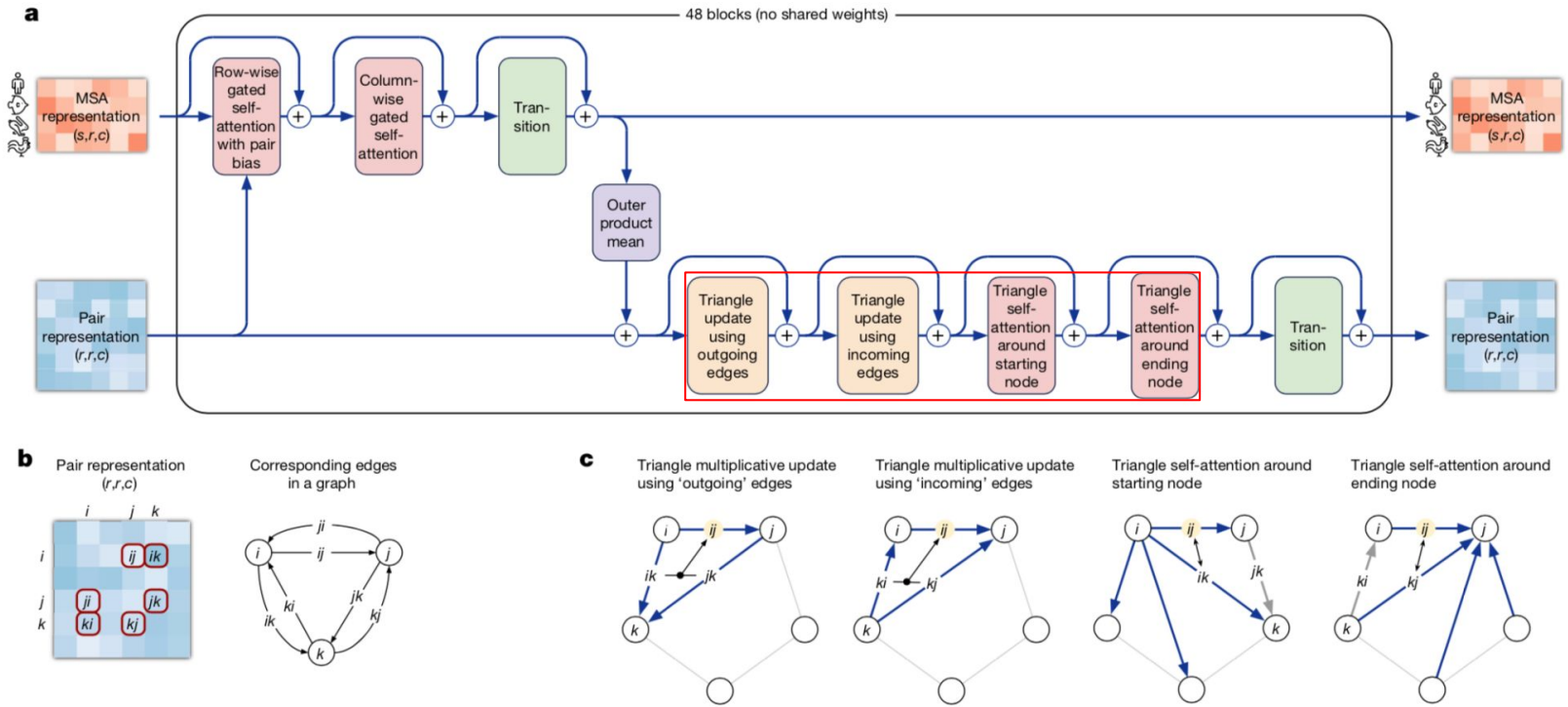
The column-wise self-attention submodule identifies pairs of strongly correlated amino acid sequences; heuristically, the model 'learns' which aligned sequences are 'informative' for the target structure

Primary Modules - Evoformer



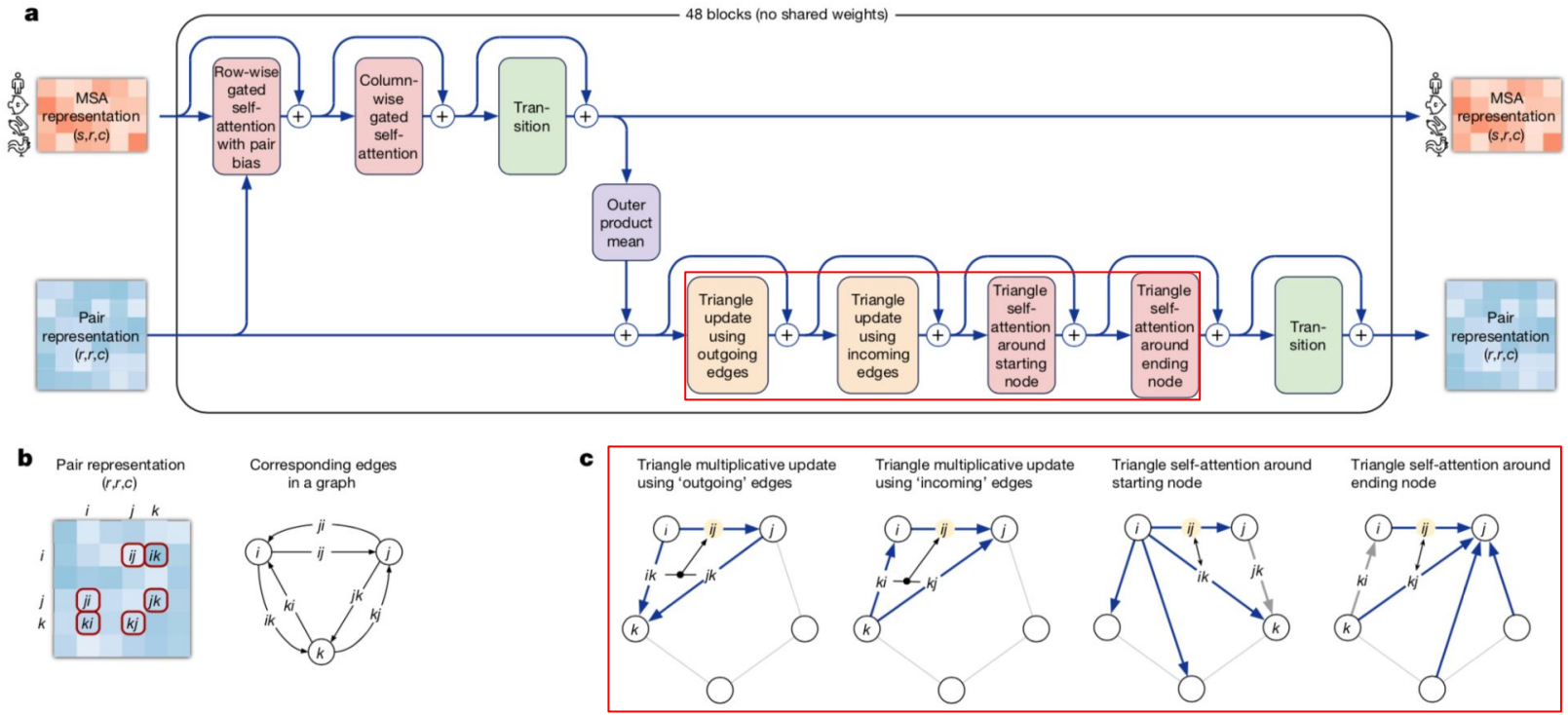
The outer product mean 'returns information' from the MSA to the pair representation by computing an outer product between columns *i* and *j*, which is used to update the *i*,*j*th entry of the pair representation

Primary Modules - Evoformer



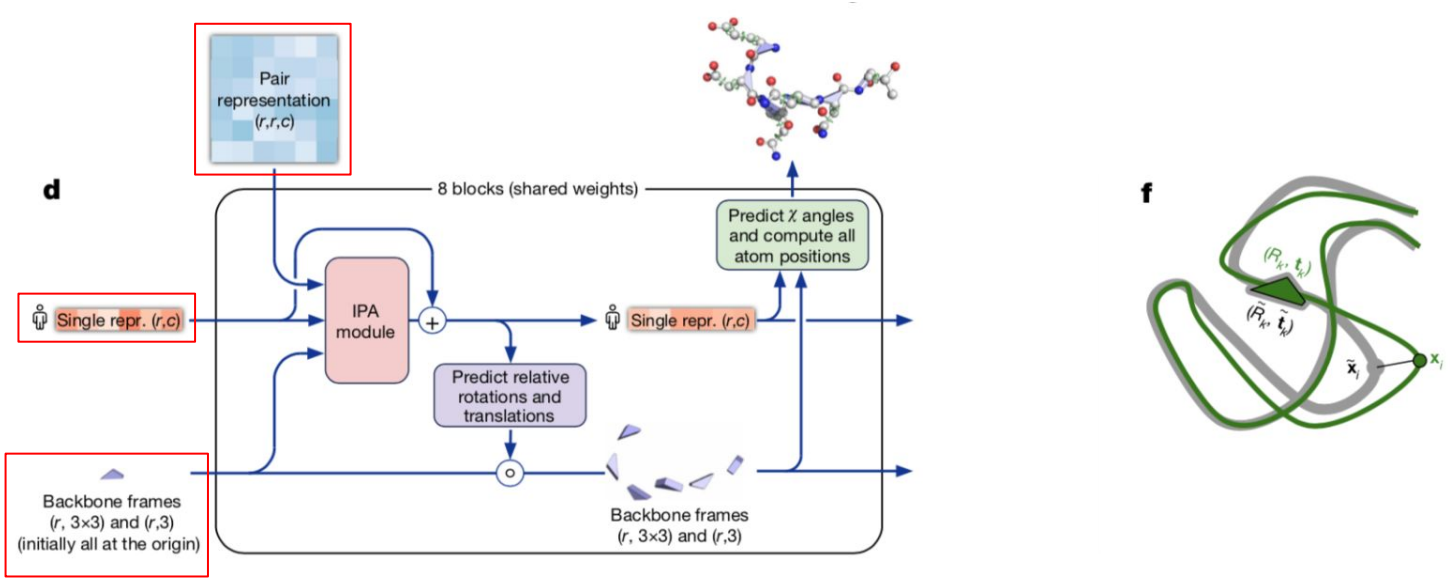
The pair representation is intended to represent a set of distances between residues of the protein, but distances in Euclidean space respect the triangle inequality, so we need to enforce this rule

Primary Modules - Evoformer



Roughly, all of these submodules operate by updating an edge's value using information from the other edges of the triangle, whether by direct linear transformations or using attention

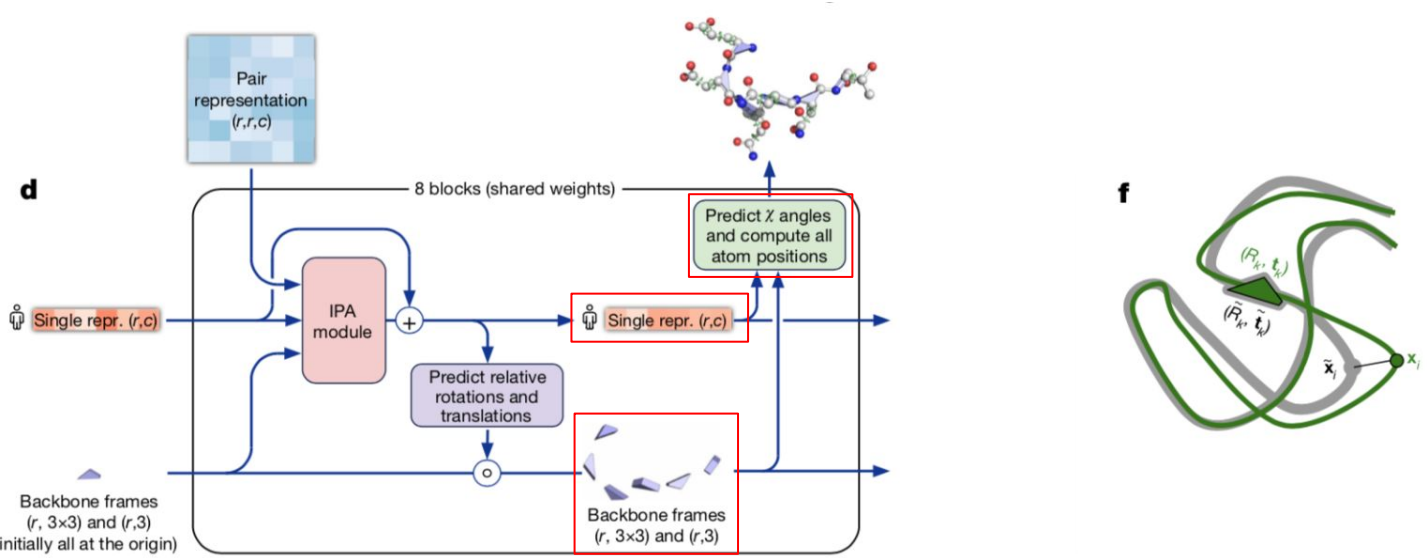
Primary Modules - Structure Module



The input to the structure module roughly consists of three items:

1. The pair representation outputted by the sequence of Evoformers - pairwise interaction information between residues
2. A single representation of the target protein - local frame information about each residue
3. Backbone frames - a set of rotation matrices and translation vectors for each residue

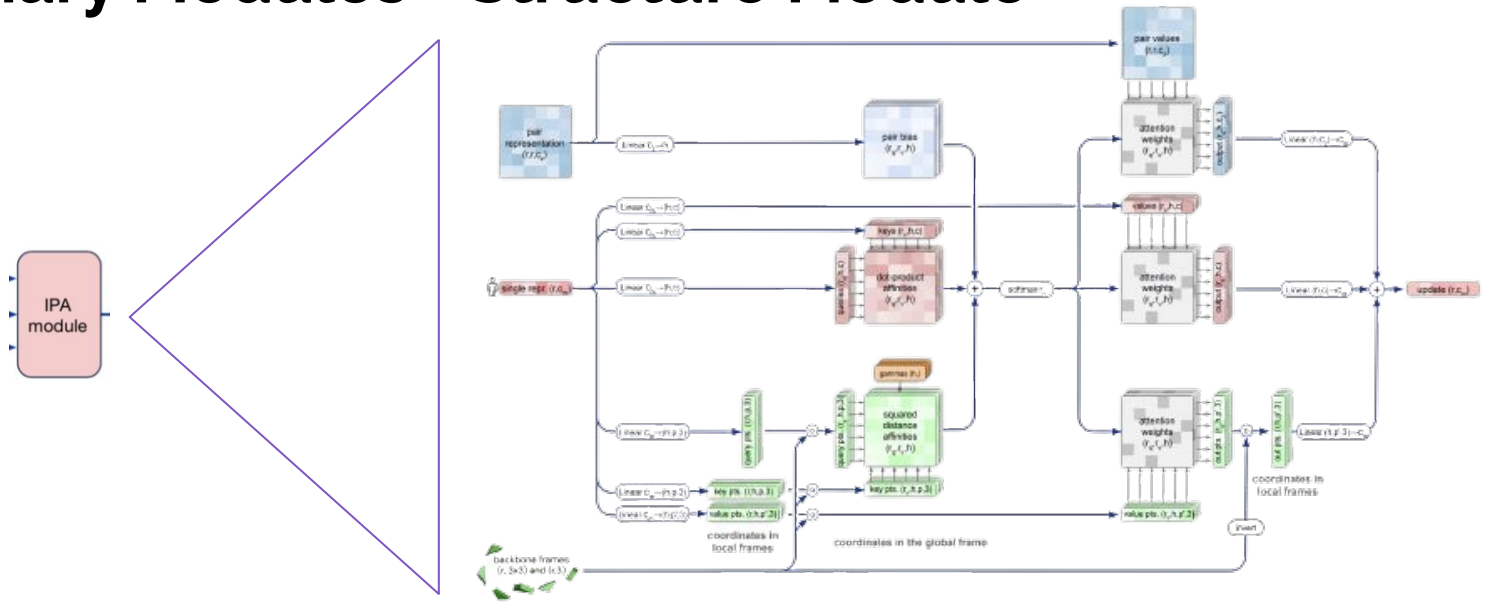
Primary Modules - Structure Module



The final output is a set of points in 3D space for each atom of the target protein, so the refinement happens on inputs 2 and 3: the abstract local frames for each residue, and the backbone frames which embed these residues in 3D space.

We already know the fixed bond lengths in amino acids, so it remains only to compute torsion angles!

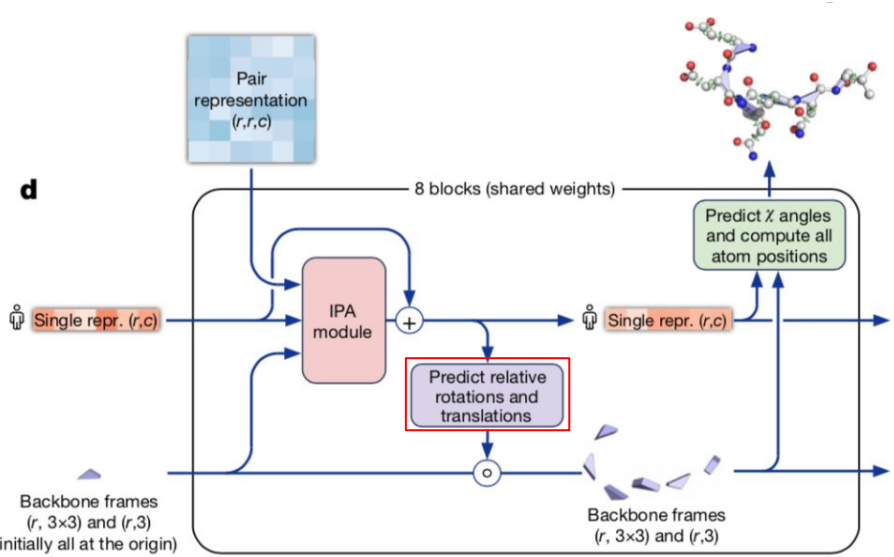
Primary Modules - Structure Module

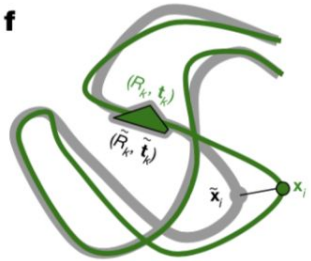


The IPA module updates the abstract single representation using multiple “channels” of attention:

1. Compute the usual self-attention for the single representation
2. Convert the pair representation into a simple element-wise “pair bias”
3. Convert the backbone frames into “distance affinities”, which can be thought of simply as representing distances
4. Add all of these to obtain a final set of attention weights, which are then combined with vector representations of each input type and summed for the final single representation update

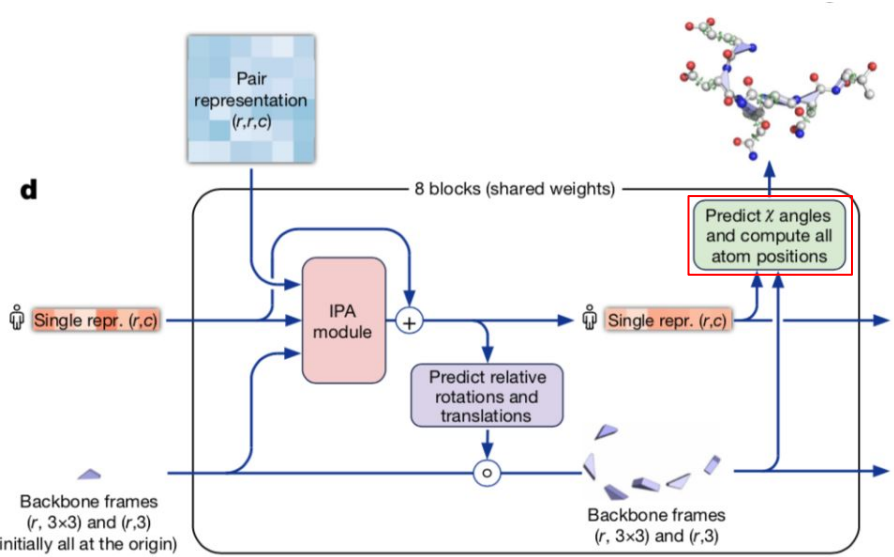
Primary Modules - Structure Module

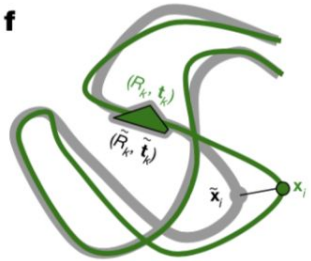




From the single representation, we can easily derive the backbone frames by simply mapping to a quaternion (with fixed first coordinate) and a translation vector

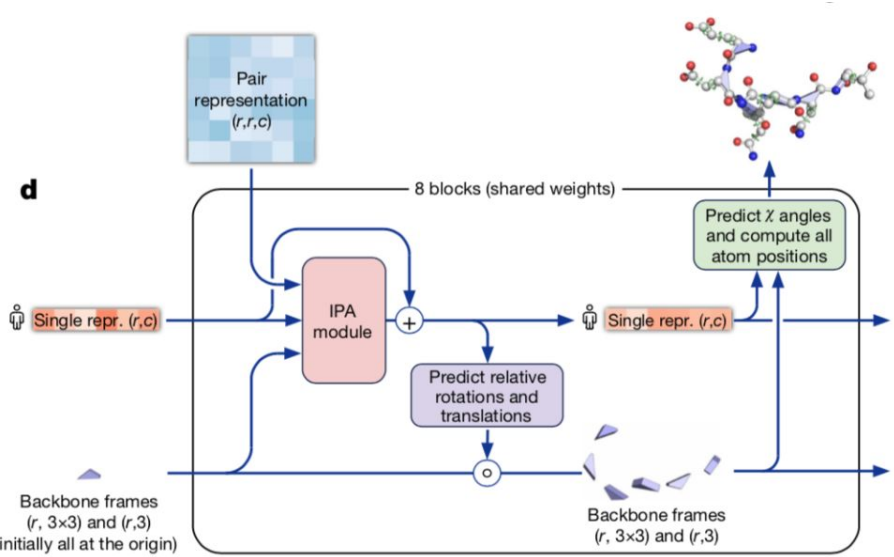
Primary Modules - Structure Module

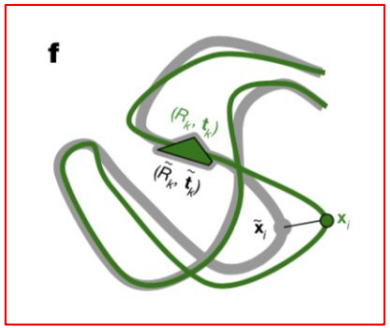




Finally, using the single representation and the backbone frames to embed in 3D space, we can predict optimal torsion angles and use them to compute each atom's position

Primary Modules - Structure Module

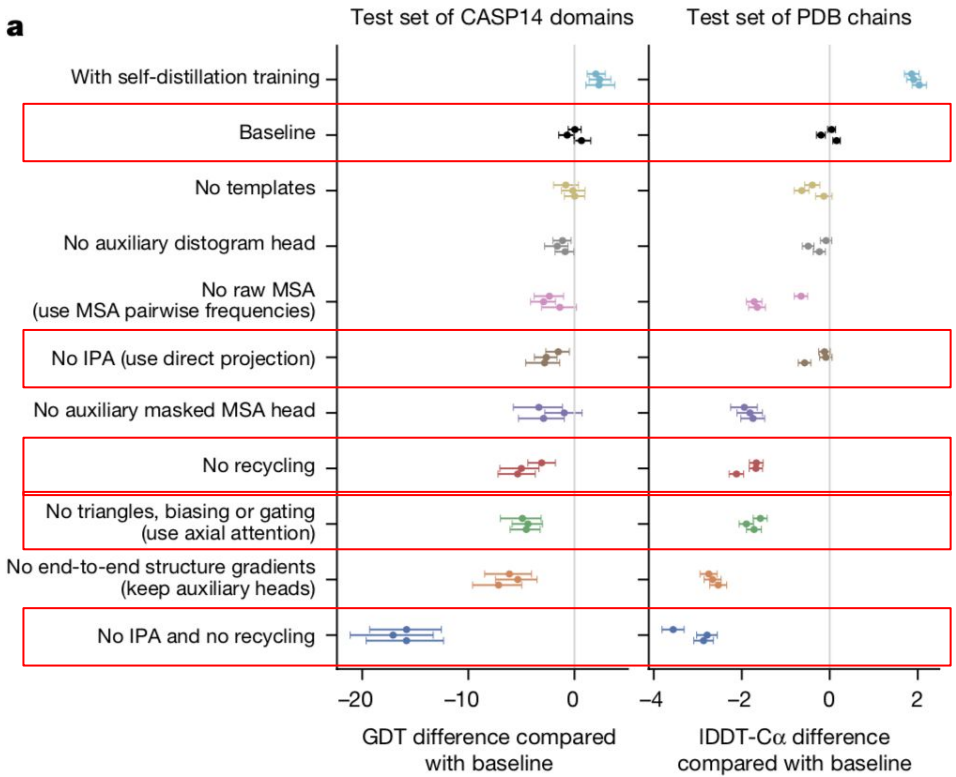




How do we optimize each of these steps?

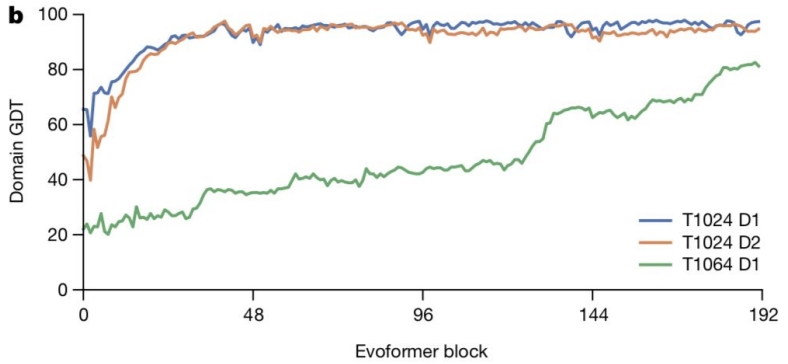
AlphaFold2 incorporates a special structure loss called the frame aligned point error (FAPE), which is roughly computed by viewing every atom under a number of different frames for both the predicted and true protein structures, then taking the average of a simple L2 norm of their differences.

Importance of Each Module



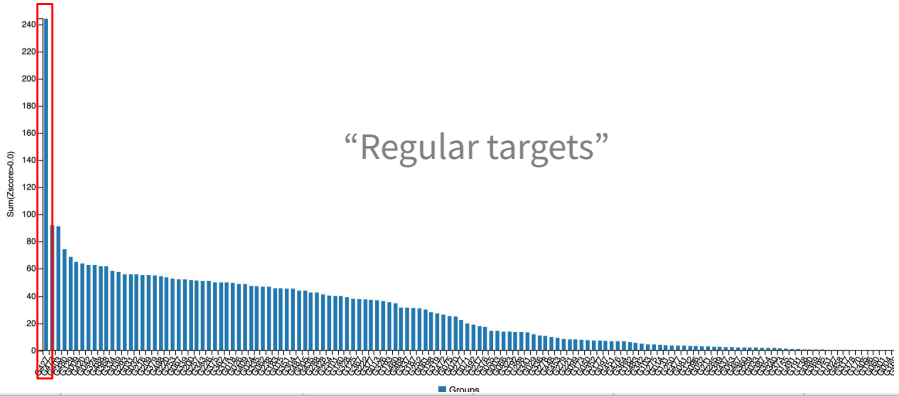
Each of the submodules/processes that we have described contributes significantly to AlphaFold2's performance

In particular, additional iterations of the Evoformer block drastically improve accuracy, although not uniformly for all proteins

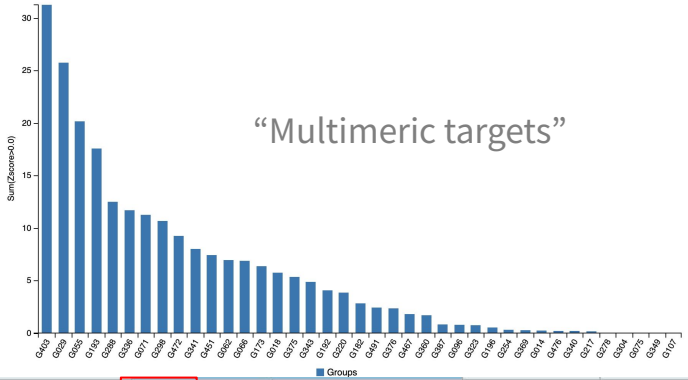


Conclusions

- The comparative approach is “completely” successful given sufficient databases and computing power
  - Alphafold2 incorporates no physical priors
  - Primary improvements in performance can be traced to the Evoformer block
  - Probably a result of representative experimental data



“Regular targets”



“Multimeric targets”

1	427	AlphaFold2	92	244.0217	1	2.6524	1	244.0217	1	2.6524	1

CASP14 results