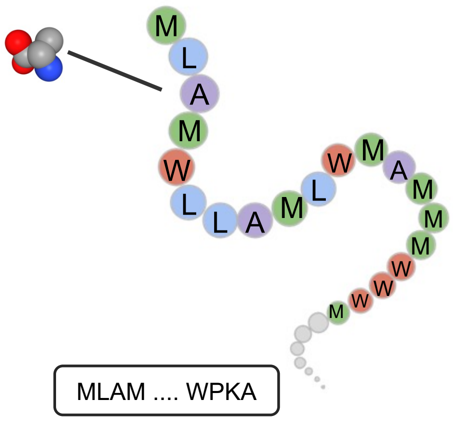
hhu
Heinrich Heine
Universität Düsseldorf

Determining protein 3D structures

Proteins

- Proteins consist of amino acid sequences that fold into 3D structures

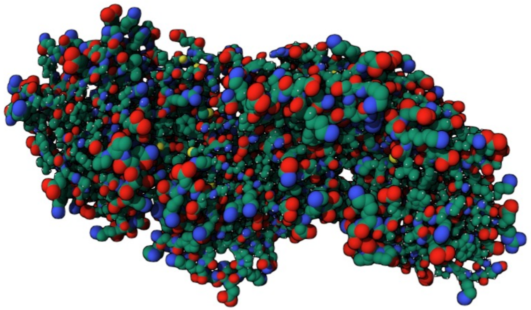
20 different amino acids; each amino acid is a small molecule



MLAM .... WPKA

Folding into

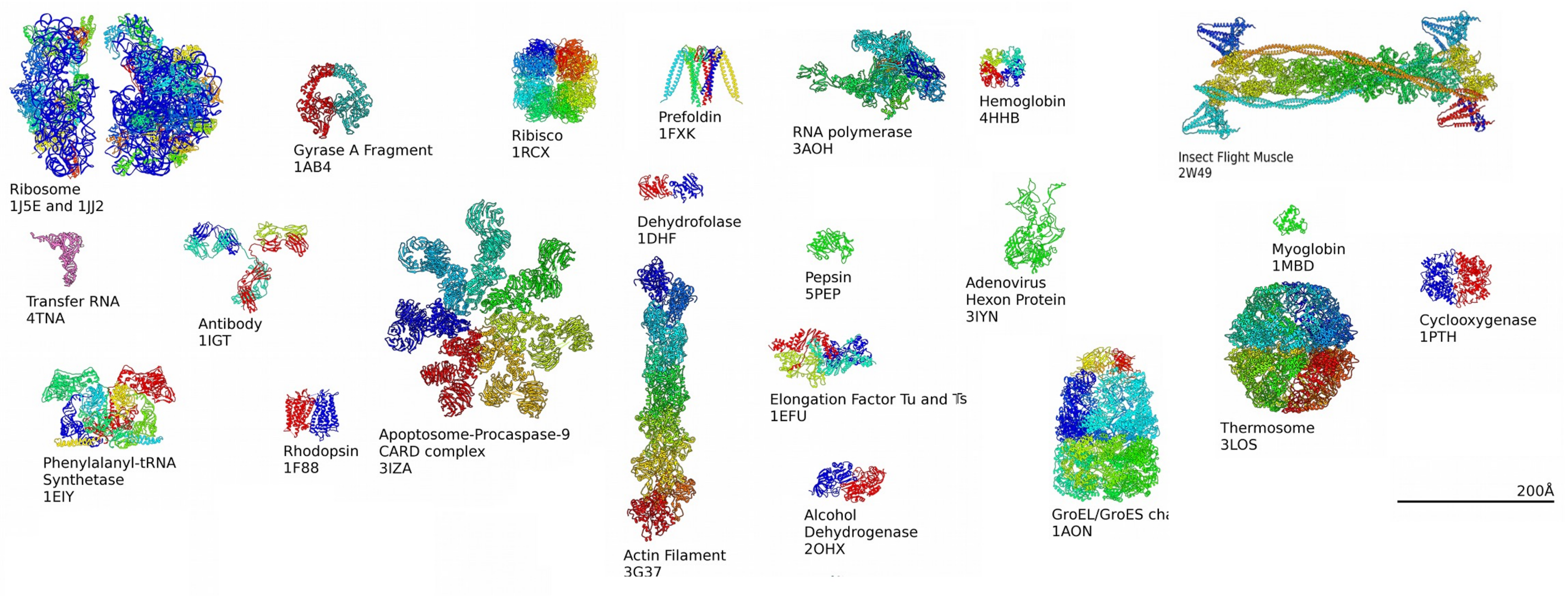
3D structure



Small molecule image: pubchem.ncbi.nlm.nih.gov; Protein image: https://www.rcsb.org/3d-view

Protein structure space

Heinrich Heine Universität Düsseldorf



Modified image from: https://commons.wikimedia.org/wiki/File%3AProtein_structure_examples.png

Experimentally determining 3D structures

- Nuclear Magnetic Resonance (NMR) Spectroscopy
- Cryogenic Electron Microscopy (Cryo-EM)
- Electron Crystallography
- X-ray Crystallography
  - 1. Protein Purification
  - 2. Protein Crystallization
  - 3. X-ray Diffraction
  - 4. Electron Density Maps
  - 5. Atom Model
  - 6. Refinements
- For novel proteins, this process typically takes several months or even up to over a year
- Protein Data Bank (PDB): stores ~180,000 experimentally determined protein 3D structures (~600M proteins are sequenced)

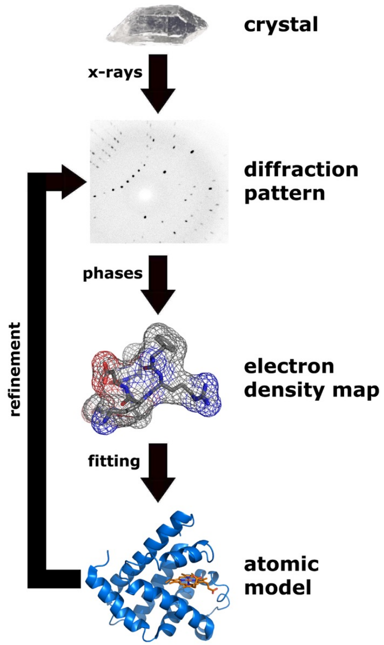
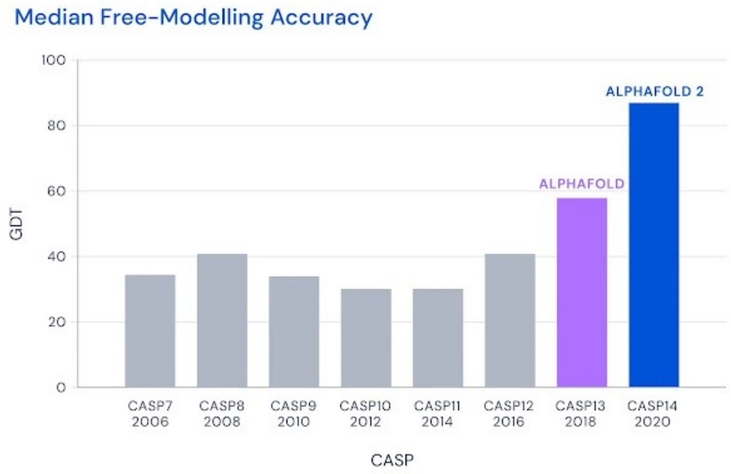


Image copied from: https://en.wikipedia.org/wiki/X-ray_crystallography#/media/File:X_ray_diffraction.png

Predicting protein 3D structures

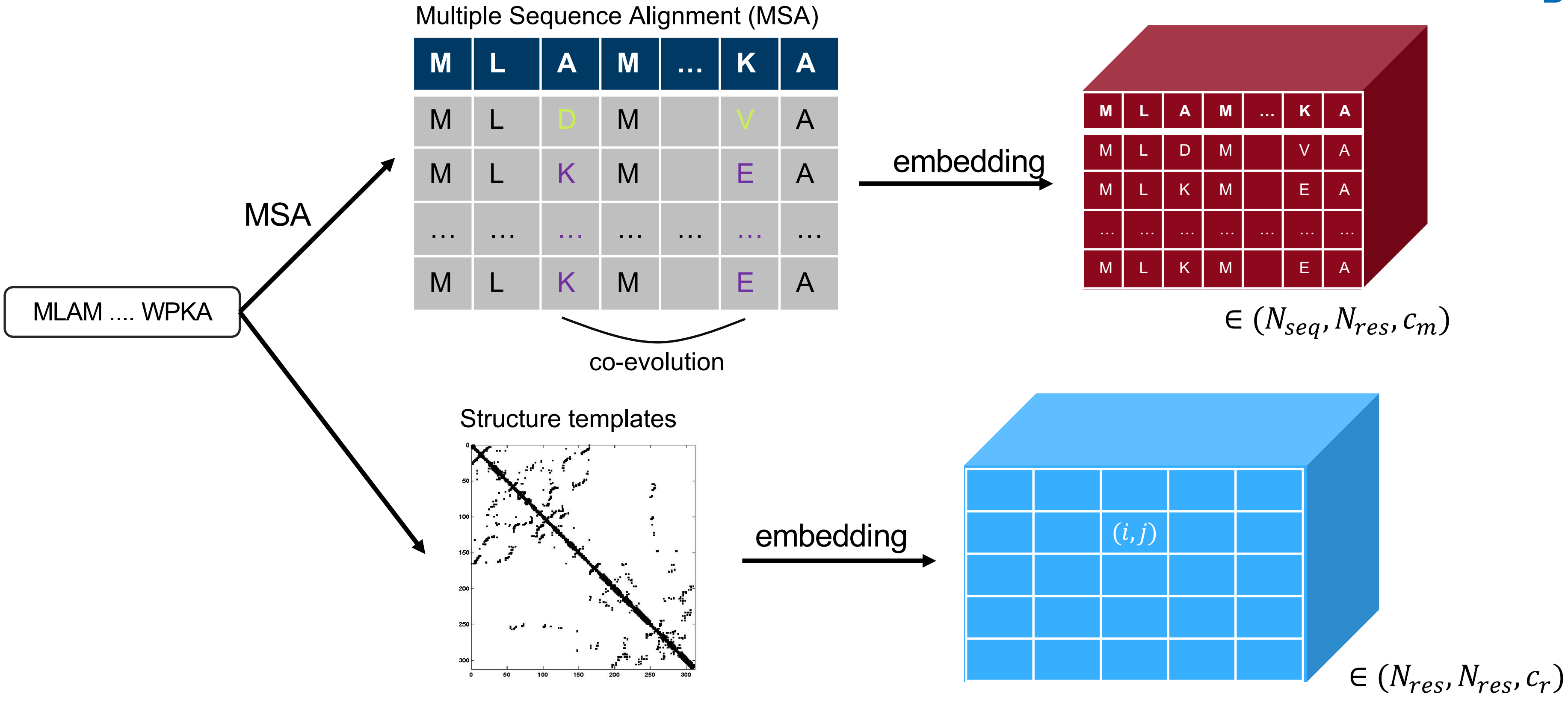
- Predicting protein 3D structure from protein sequence is desirable
  - Was not possible (with high accuracy) until 2020:



- Deep Learning Models that can predict protein 3D structures
  - AlphaFold 2
  - ESMFold
  - RoseTTAFold

Image copied from: https://deepmind.google/discover/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology/

AlphaFold2 - Input



MLAM .... WPKA

MSA

Multiple Sequence Alignment (MSA)

M	L	A	M	…	K	A
M	L	D	M		V	A
M	L	K	M		E	A
…	…	…	…	…	…	…
M	L	K	M		E	A

co-evolution

embedding

M	L	A	M	…	K	A
M	L	D	M		V	A
M	L	K	M		E	A
…	…	…	…	…	…	…
M	L	K	M		E	A

$\in (N_{seq}, N_{res}, c_m)$

Structure templates

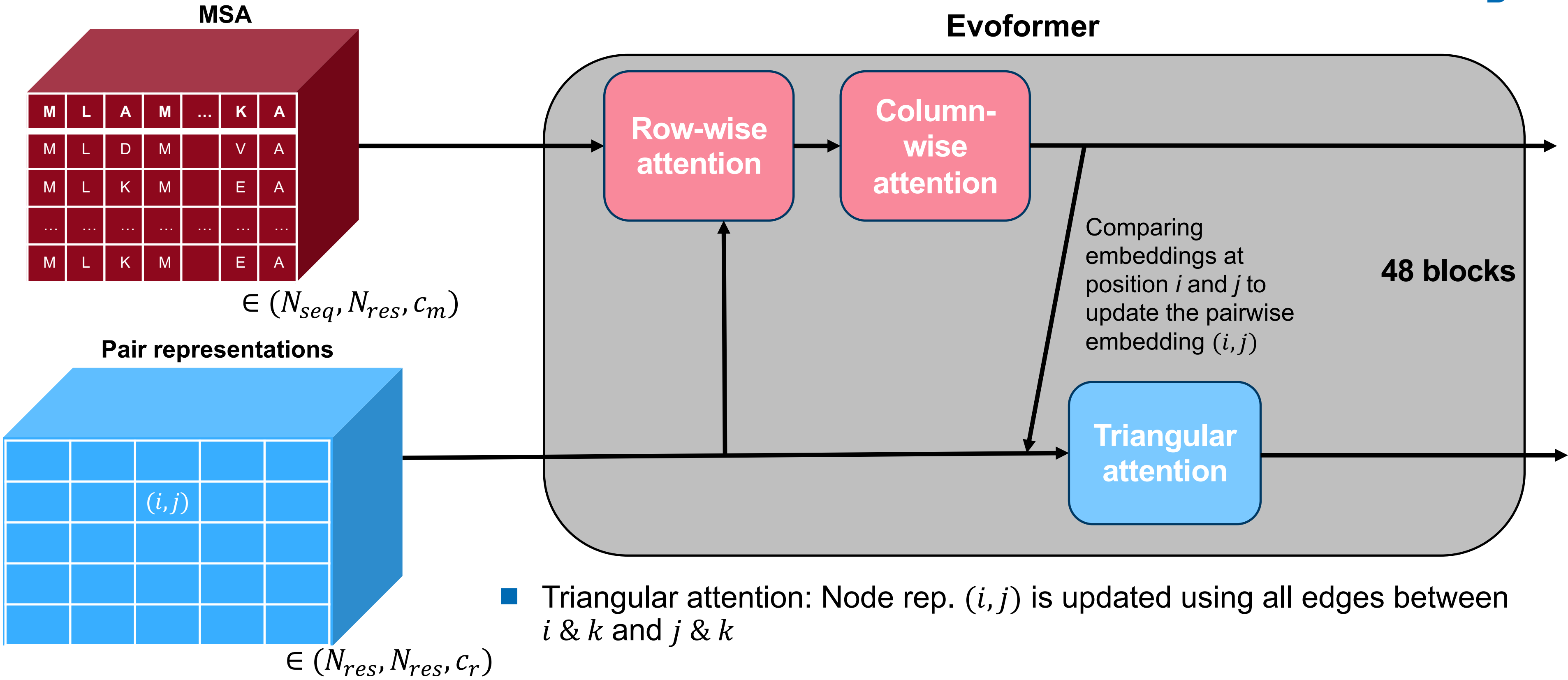
embedding

(i, j)

$\in (N_{res}, N_{res}, c_r)$

Contact map image copied from: Ramanathan, A et al. (2008). *Using tensor analysis to characterize contact-map dynamics of proteins* (PhD thesis)

AlphaFold2 - Evoformer



M	L	A	M	…	K	A
M	L	D	M		V	A
M	L	K	M		E	A
…	…	…	…	…	…	…
M	L	K	M		E	A

$\in (N_{seq}, N_{res}, c_m)$

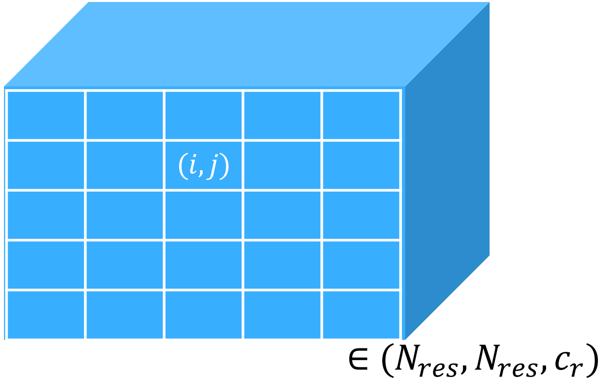
Comparing embeddings at position *i* and *j* to update the pairwise embedding (i,j)

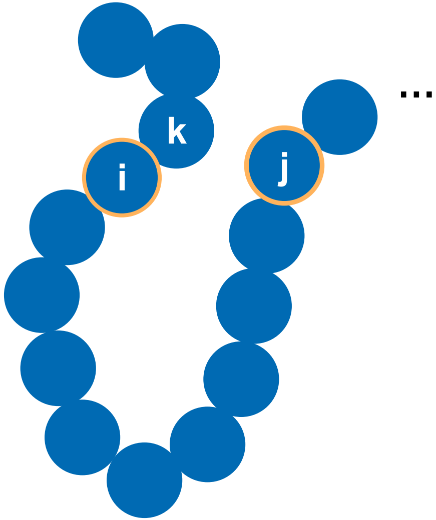
- Triangular attention: Node rep. (i,j) is updated using all edges between i & k and j & k

AlphaFold2 – Triangular Attention

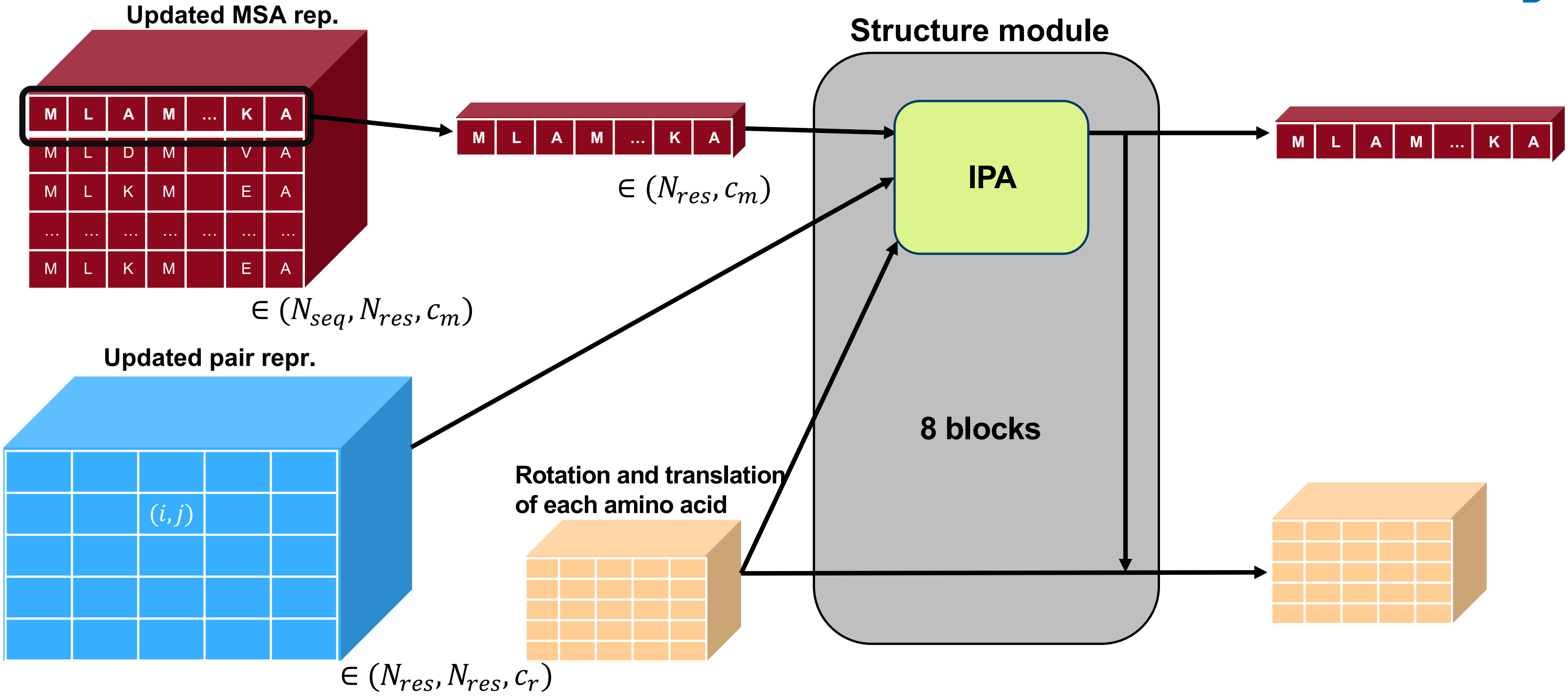
- Triangular attention: Node rep. (i,j) is updated using all edges between i & k and j & k
  - Integrating Euclidean geometry into the network:
    Knowing distance between i & k and between j & k put a strong constraint of distance between i & j

Pair representations

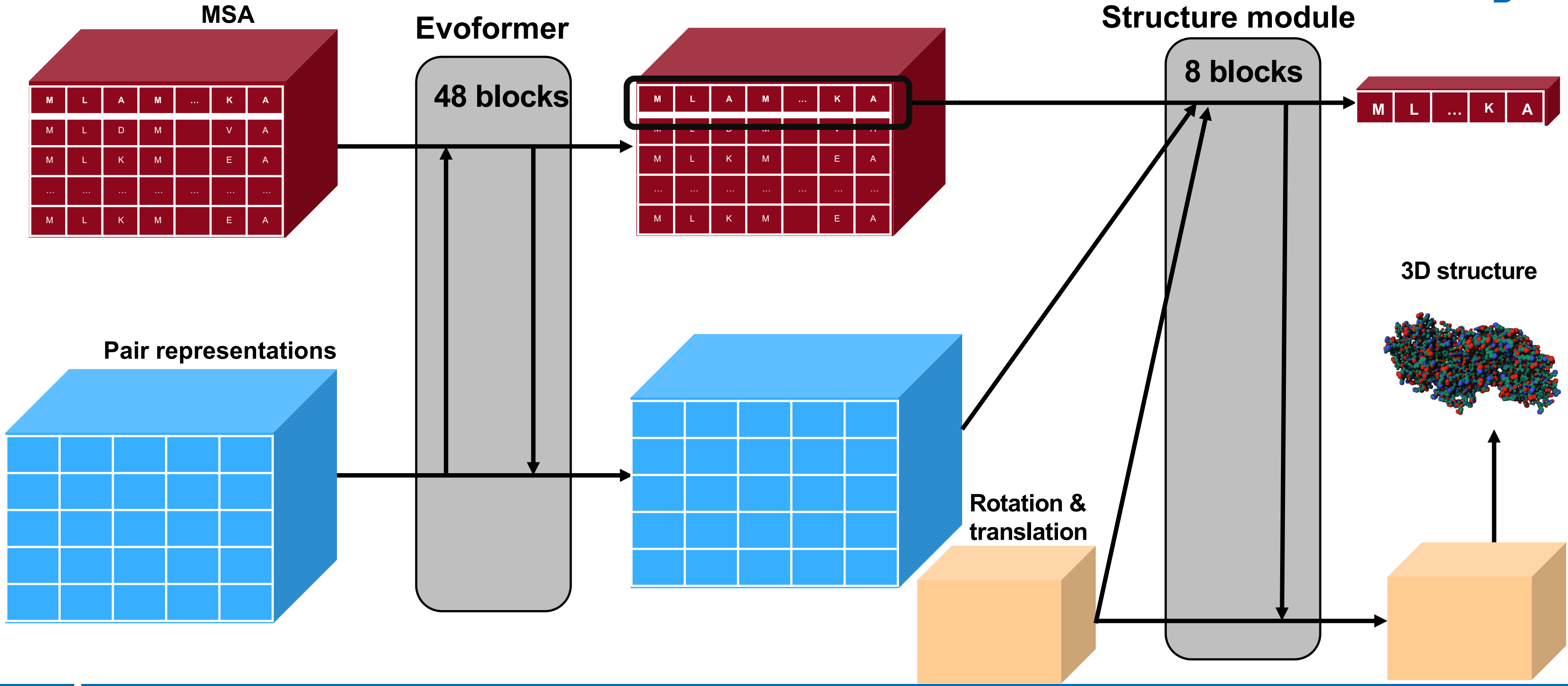




AlphaFold2 – Structure module

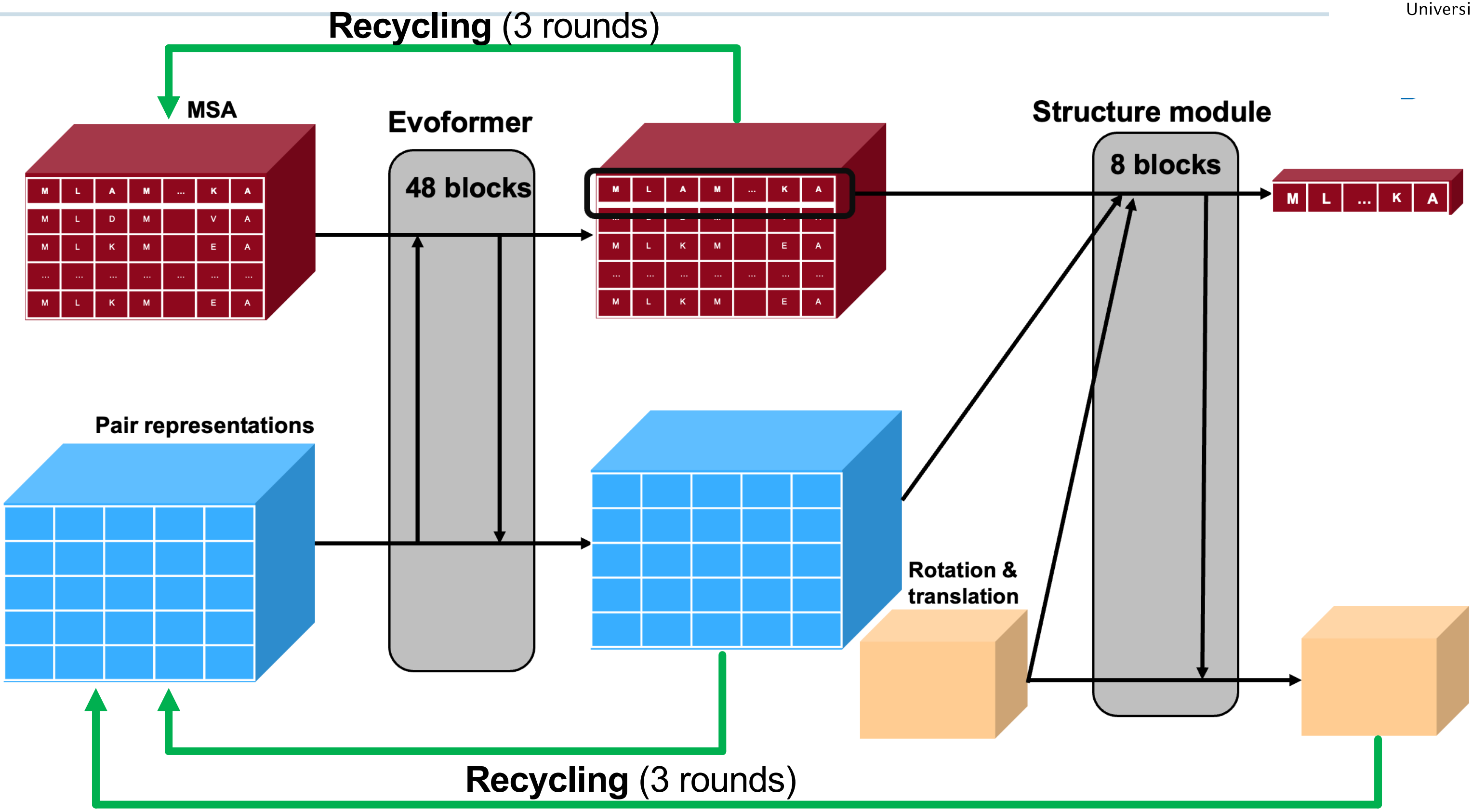


AlphaFold2 – Full Model



M	L	A	M	...	K	A
M	L	D	M		V	A
M	L	K	M		E	A
...	...	...	...	...	...	...
M	L	K	M		E	A

AlphaFold2 – Recycling



M	L	A	M	...	K	A
M	L	D	M		V	A
M	L	K	M		E	A
...	...	...	...	...	...	...
M	L	K	M		E	A

AlphaFold2 – Training

- Loss function:
  - Structural loss named FAPE: Similar to RMSD (root mean squared deviation) of atomic positions
  - Auxiliary losses:
    - Distogram loss: Comparing the pairwise distances between amino acids to ground truth
    - MSA masking: Some tokens of the MSA are masked out and are predicted during training
    - …
- Training data:
  - ~120k experimentally determined protein 3D structures from the PDB
  - After first run of training:
    - Predicting structures of ~350k proteins with yet unknown structures
    - New training on experimental and predicted 3D structures

AlphaFold2 – Capabilities and limitations

- Capabilities
  - Single protein chains
  - Protein multimers
  - Protein-protein complexes
  - AlphaFold2 can often predict protein structure, even if there are no known related protein structures, if a sequence has many related sequences
- Limitations
  - Struggles to predict the structures of protein with few closely related protein sequences
  - Cannot predict effect of point mutations
  - AlphaFold2 does not capture such conformational changes
  - Cannot predict interactions with other molecules, e.g., small molecules

List of limitations and capabilities based on: https://www.ebi.ac.uk/training/online/courses/alphafold/an-introductory-guide-to-its-strengths-and-limitations/strengths-and-limitations-of-alphafold/