UNIVERSITÄT
HEIDELBERG

Protein Structure Prediction

L5, Structural Bioinformatics

WiSe 2023/24, Heidelberg University

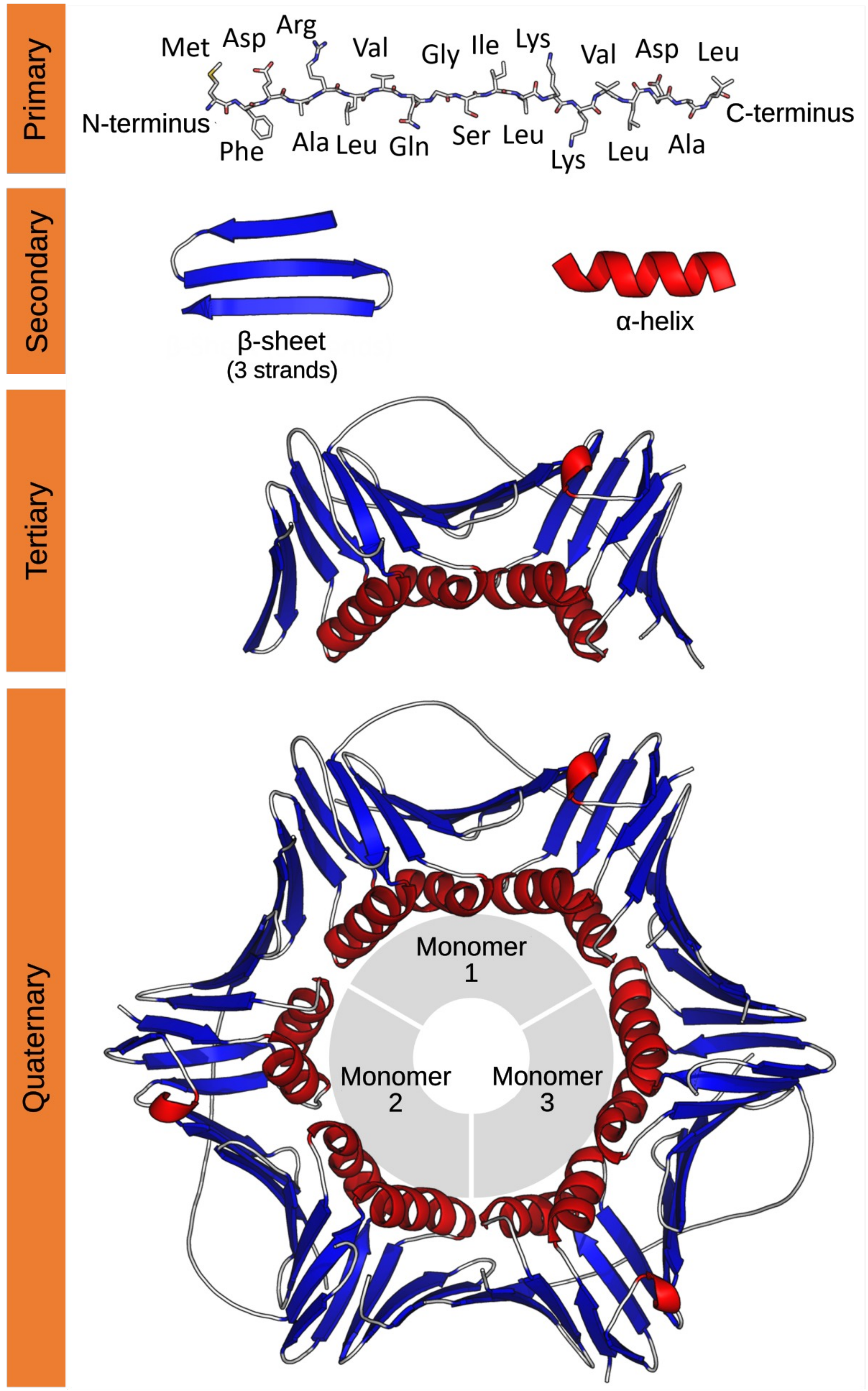
Overview

1. The Problem and its History
2. Pre-AlphaFold2 World
3. AF2: The main ideas
4. AF2: The Evoformer
5. AF2: The Structure Module
6. AF2: Losses and other Details
7. Impact and Outlook

1. The Problem and its History

Protein Structure is important

As the old dogma goes: structure determines function

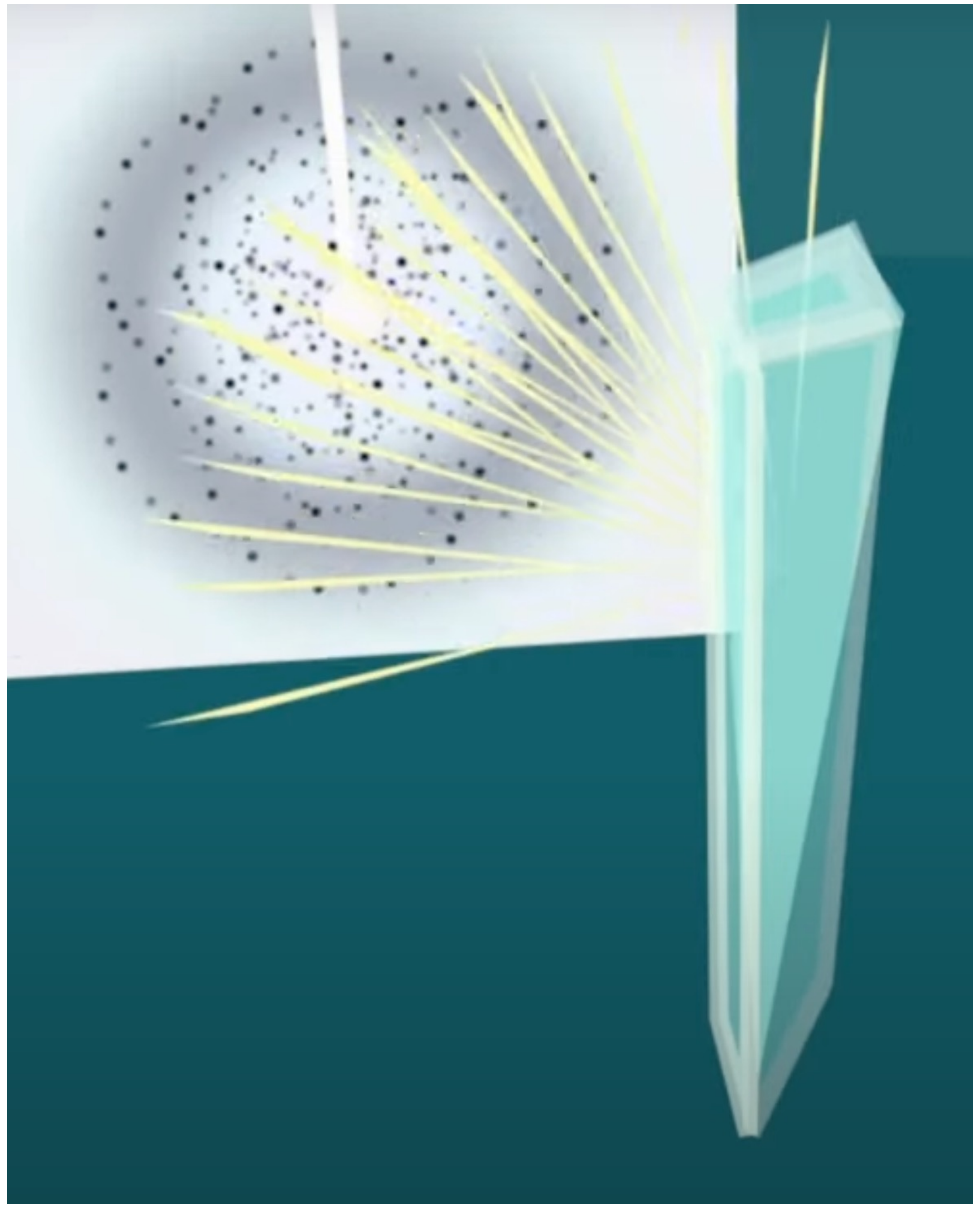


Wikipedia.org

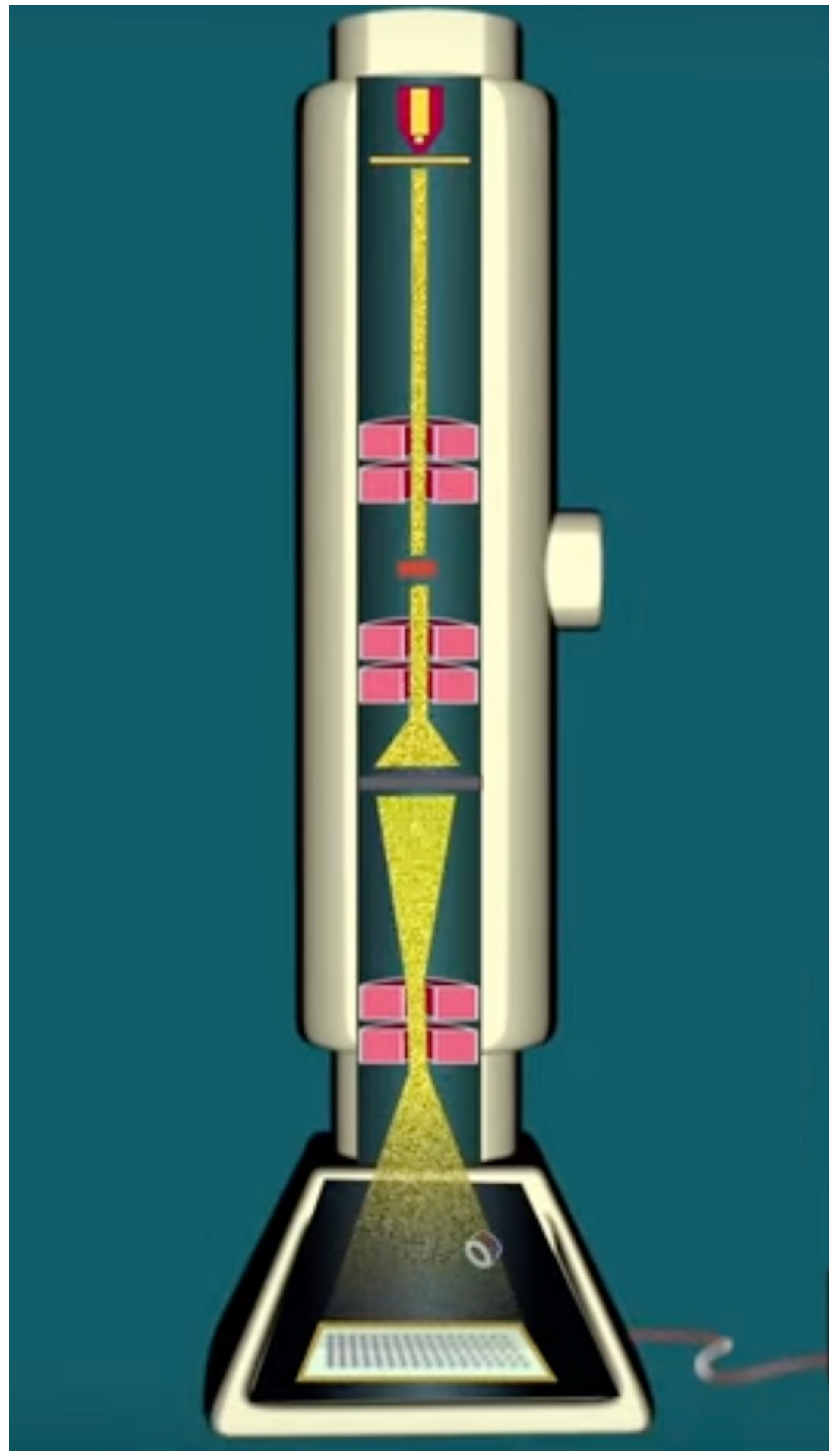
Experimental structure determination

3 main methods, all of them a lot of work

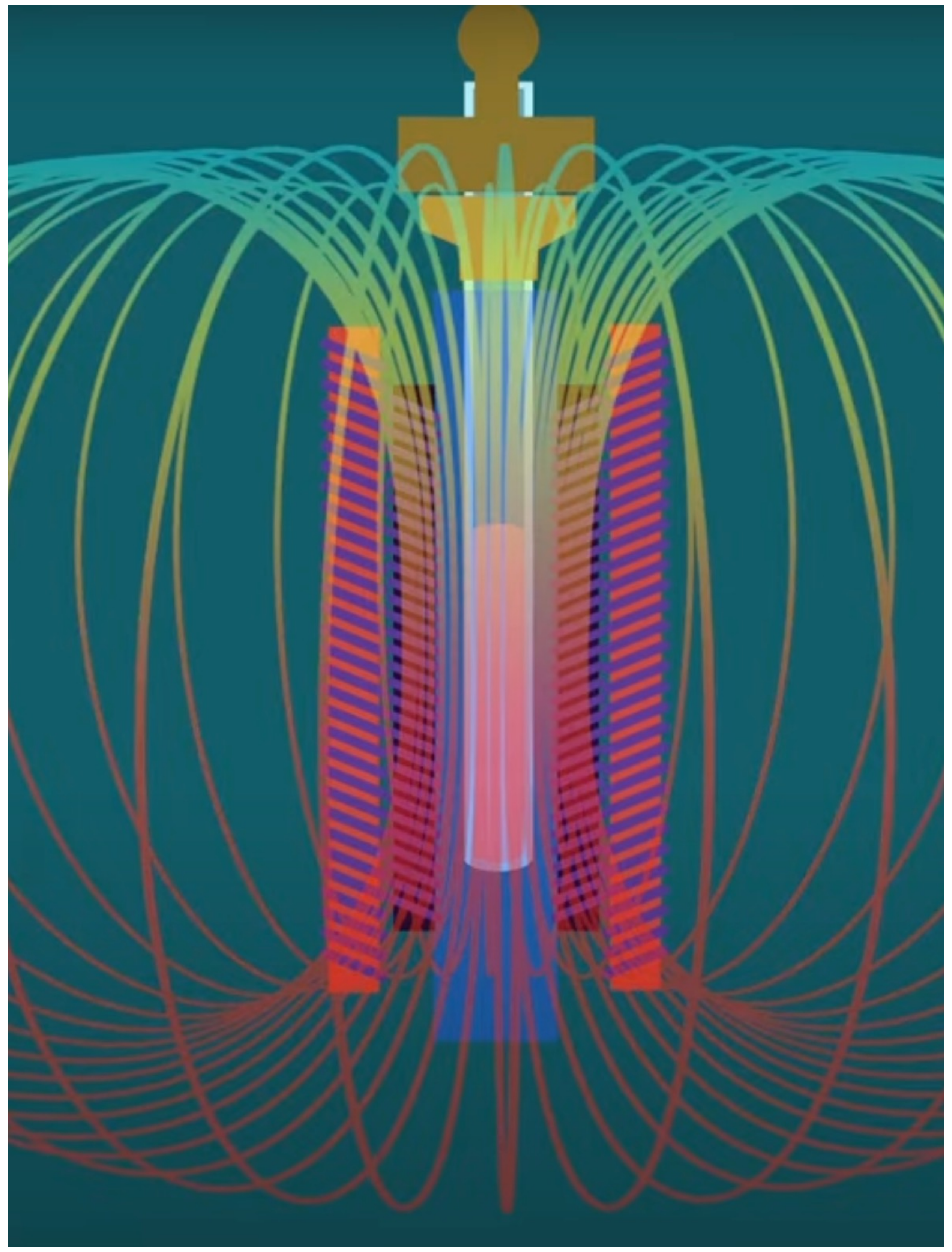
X-Ray Crystallography



Cryo-EM



NMR

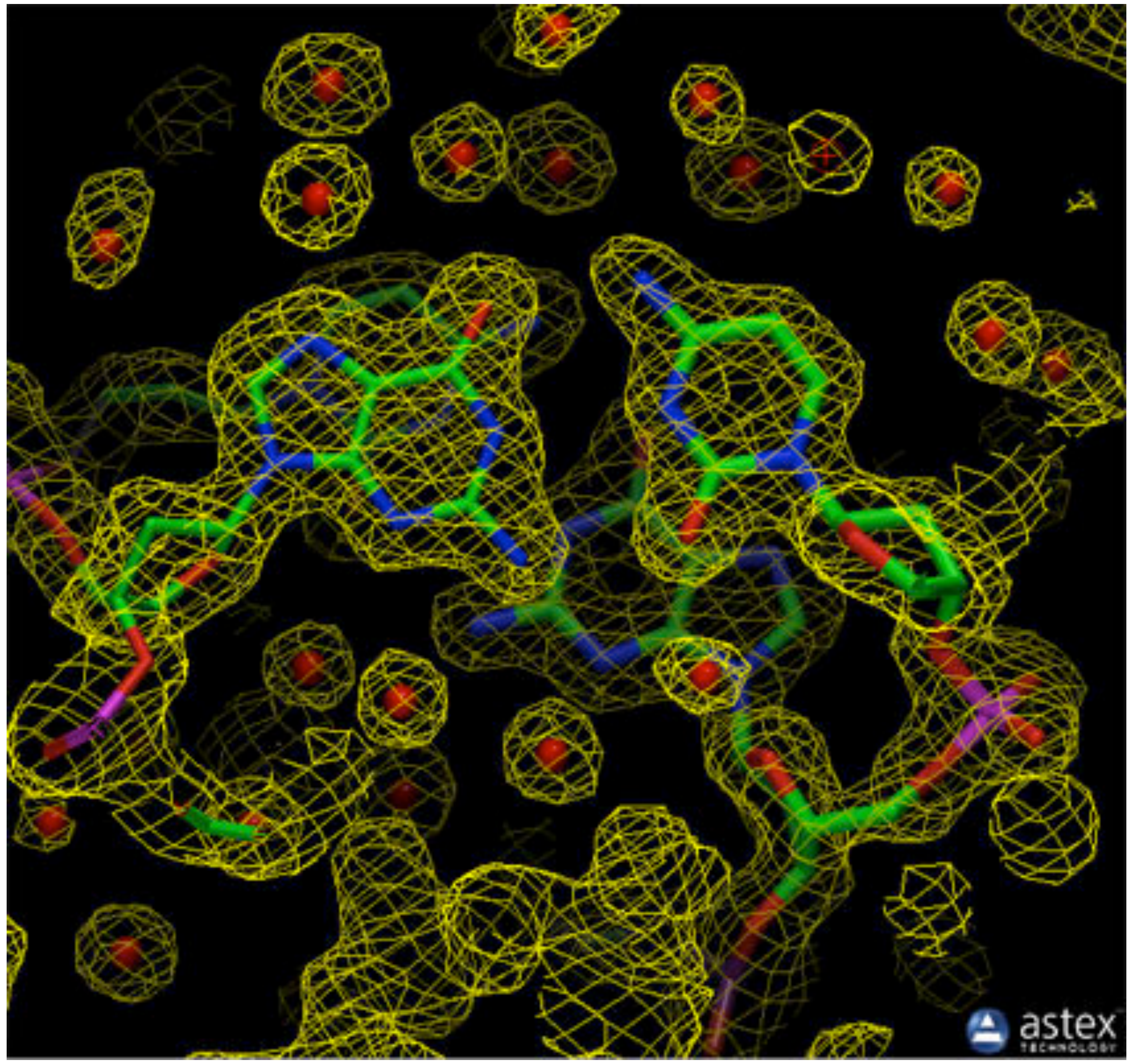


PDB 101 Article

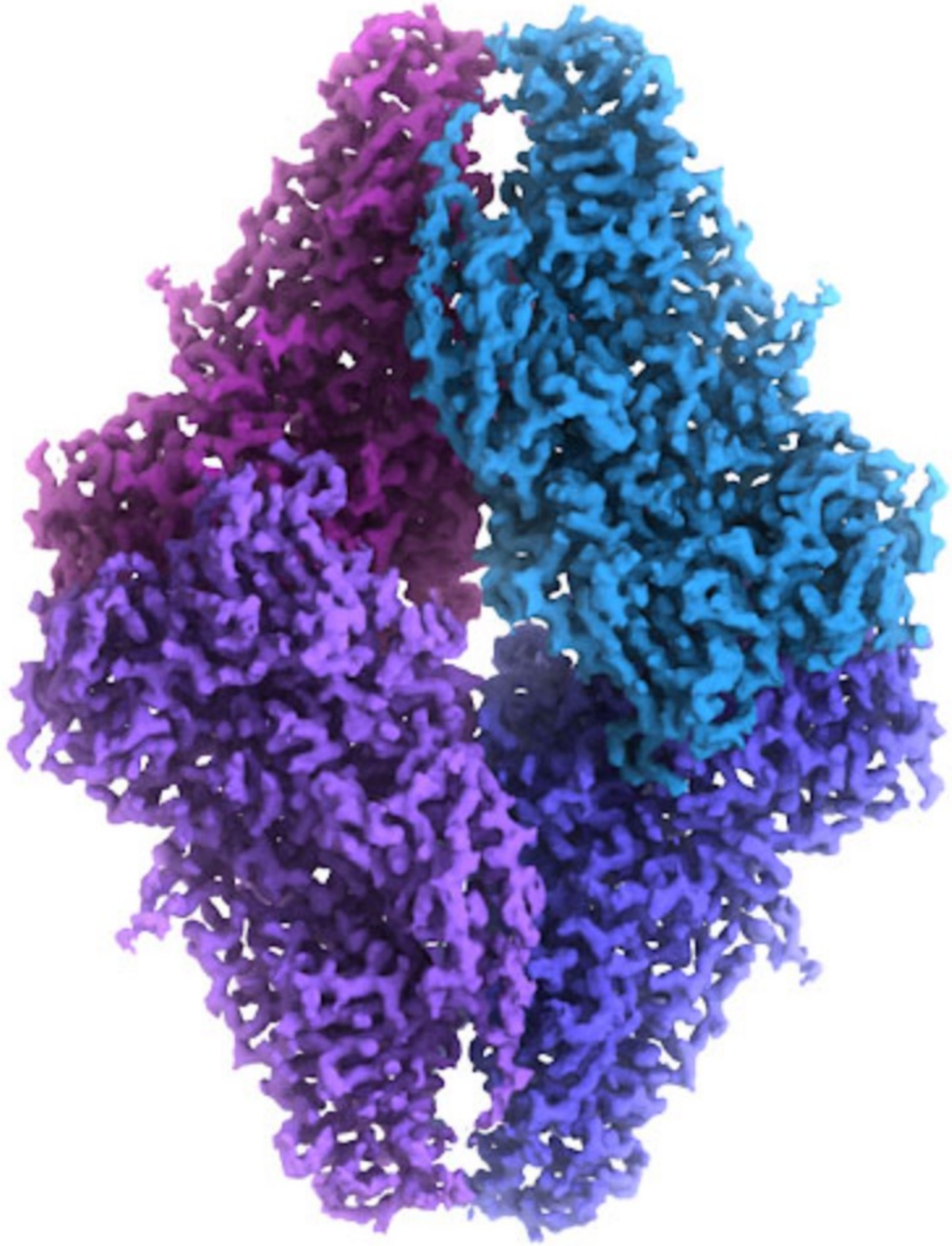
Experimental structure determination

3 main methods, all of them a lot of work

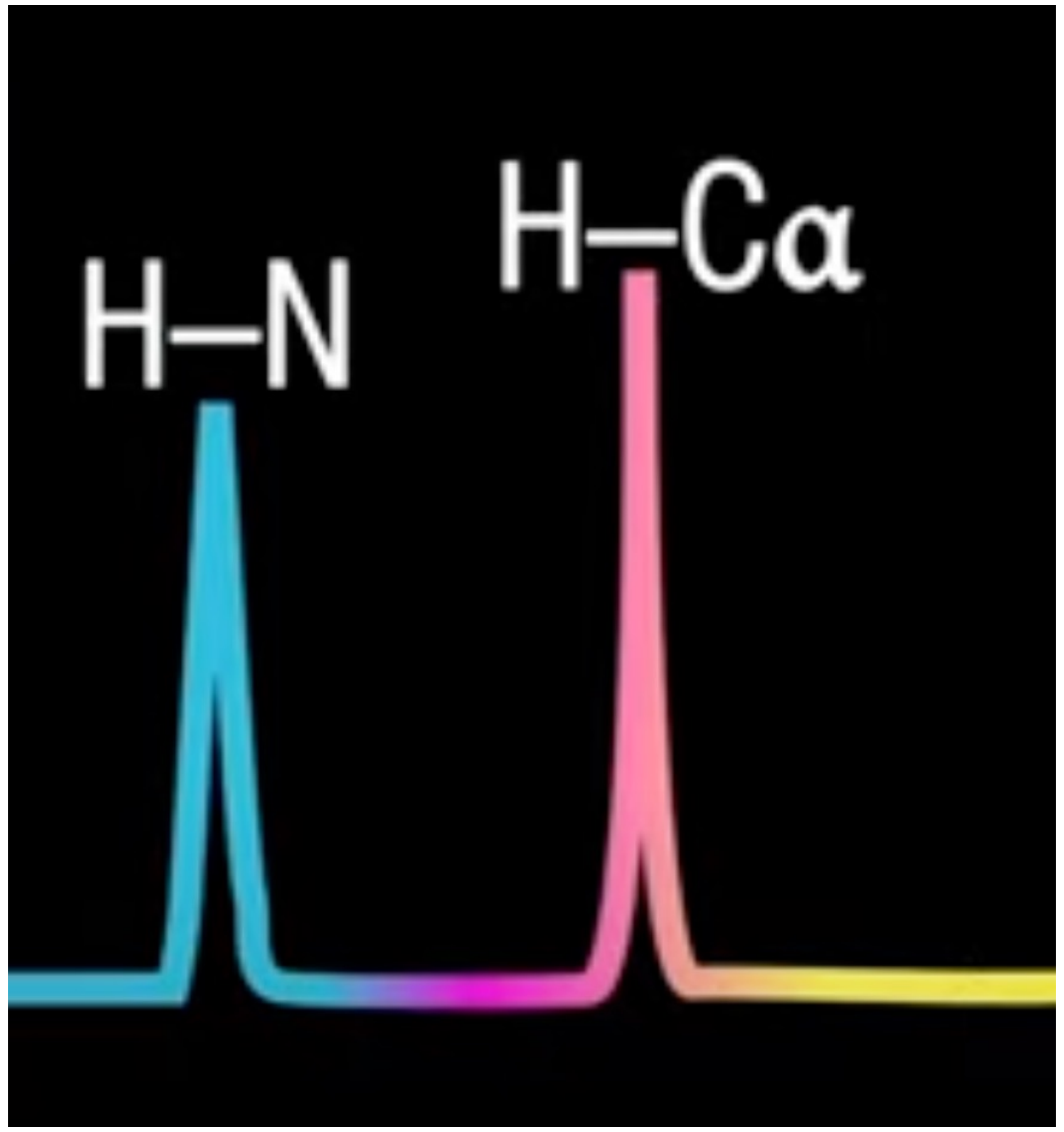
X-Ray Crystallography



Cryo-EM



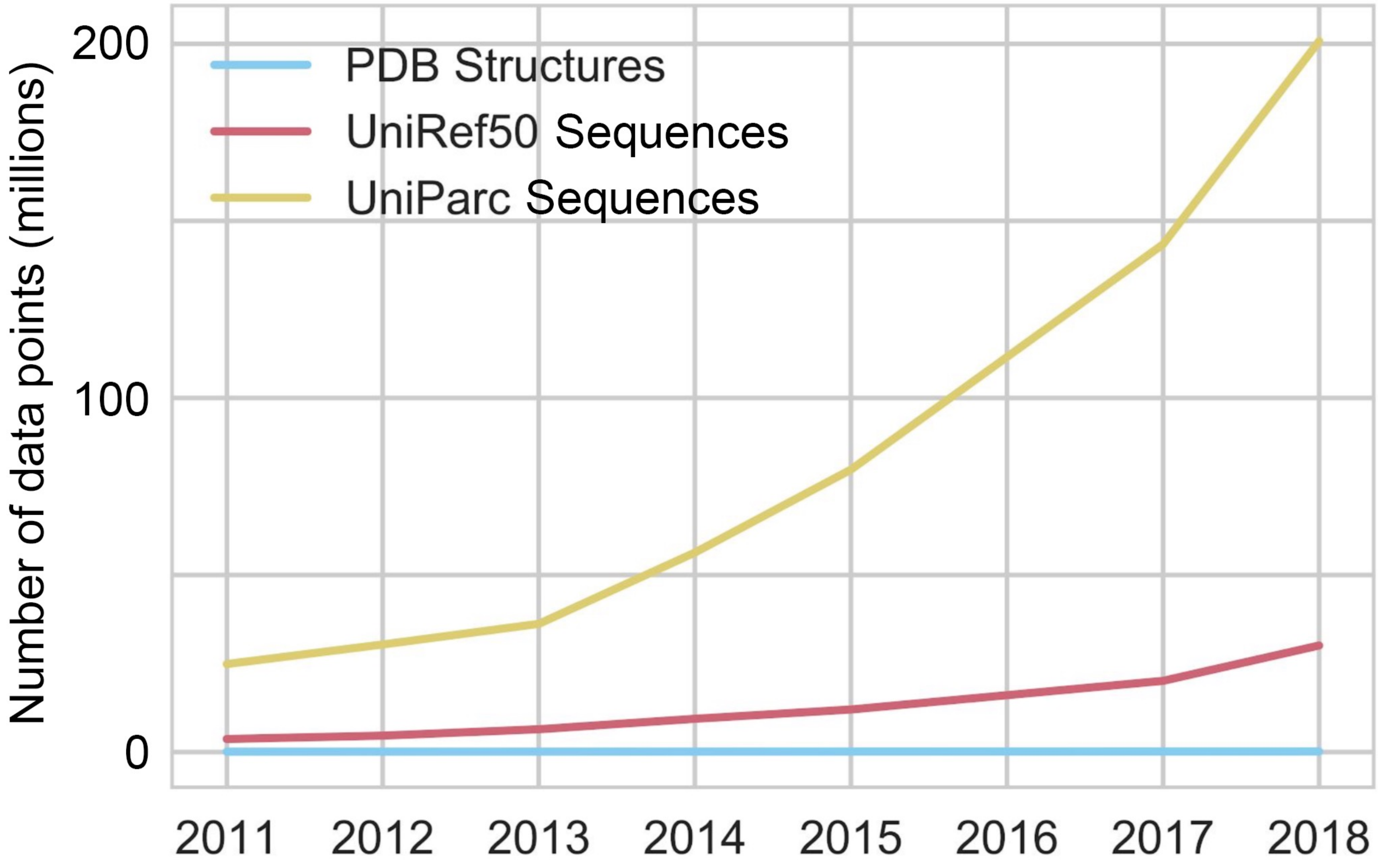
NMR



PDB 101 Article

The sequence-structure gap

Cheaper sequencing widens it every year



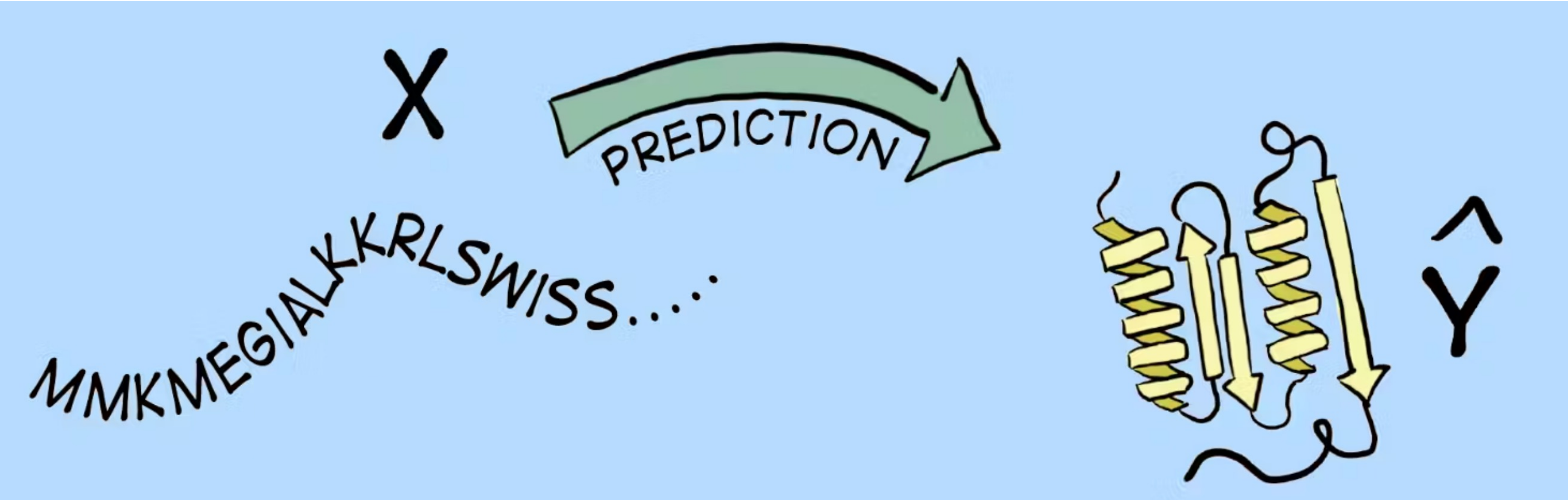
200M

200K

Mohammed AlQuraihi, Blog

Protein Structure Prediction

The “cheap” alternative



My PTNG blog post

Protein Structure Prediction its hard

Called a “grand challenge in biology” for a reason

WHAT DO YOU DO?
I MAKE SOFTWARE THAT PREDICTS HOW PROTEINS WILL FOLD.
IS THAT A HARD PROBLEM?
SOMEONE MAY SOMEDAY FIND A HARDER ONE.
WHY IS IT SO HARD?
HAVE YOU EVER MADE A FOLDED PAPER CRANE?
YEAH.
IMAGINE FIGURING OUT THE FOLDS TO MAKE AN ACTUAL *LIVING* CRANE.
... *JUST* FOLDS? CAN I MAKE CUTS?
IF YOU CAN FOLD A PROTEASE ENZYME.

xkcd

Where do we come from?

The balance between *ab initio* prediction and data-driven methods

Images: [1] PyMolWiki, [2] Aponte-Santamaria, [3] Wikipedia, [4] Wikipedia, [5] Pixabay

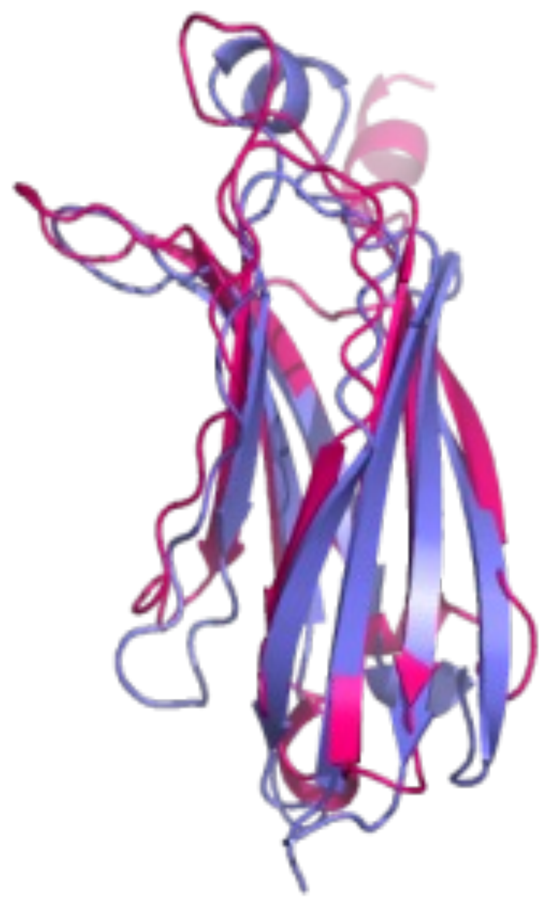
Where do we come from?

The balance between *ab initio* prediction and data-driven methods

1970s

Template-based Modelling (TBM)

Utilise sequence alignments to ”copy” similar residues



1980s

1990s

2000s

2010s

Images: [1] PyMolWiki, [2] Aponte-Santamaria, [3] Wikipedia, [4] Wikipedia, [5] Pixabay

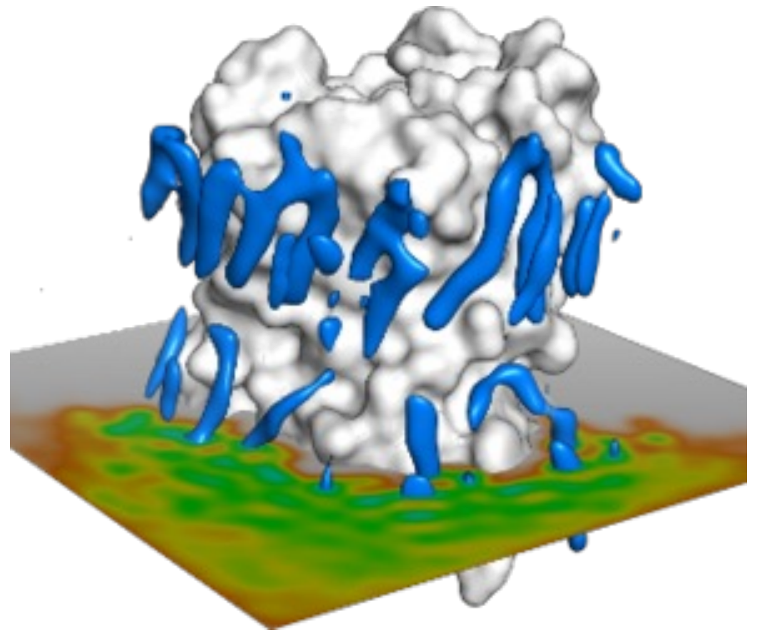
Where do we come from?

The balance between *ab initio* prediction and data-driven methods

1970s

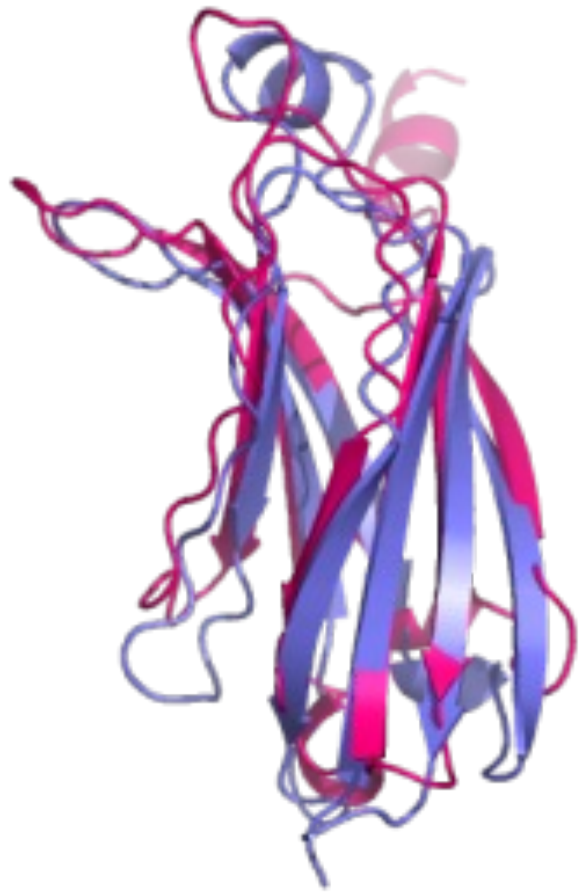
Template-based Modelling (TBM)

Utilise sequence alignments to ”copy” similar residues



1990s

2010s



1980s

Molecular Dynamics

AMBER (‘81), CHARMM (‘83)

2000s

Images: [1] PyMolWiki, [2] Aponte-Santamaria, [3] Wikipedia, [4] Wikipedia, [5] Pixabay

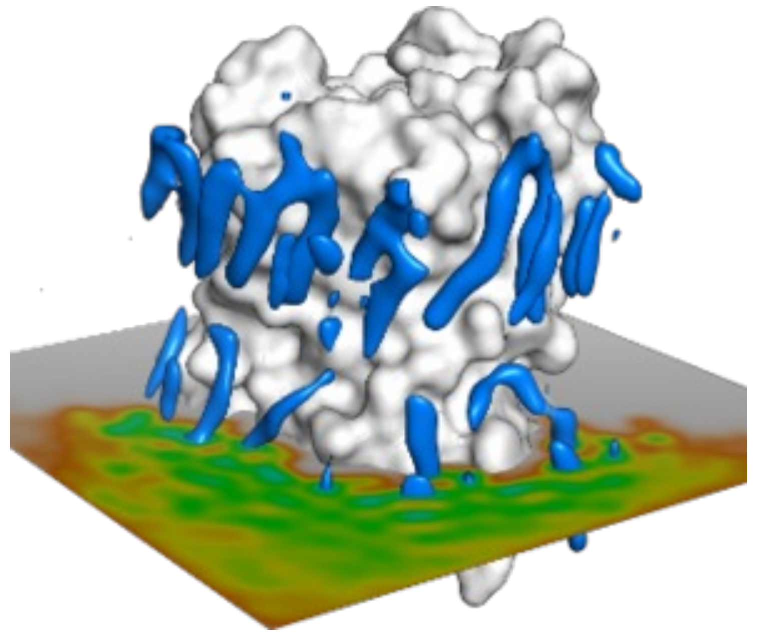
Where do we come from?

The balance between *ab initio* prediction and data-driven methods

1970s

Template-based Modelling (TBM)

Utilise sequence alignments to ”copy” similar residues

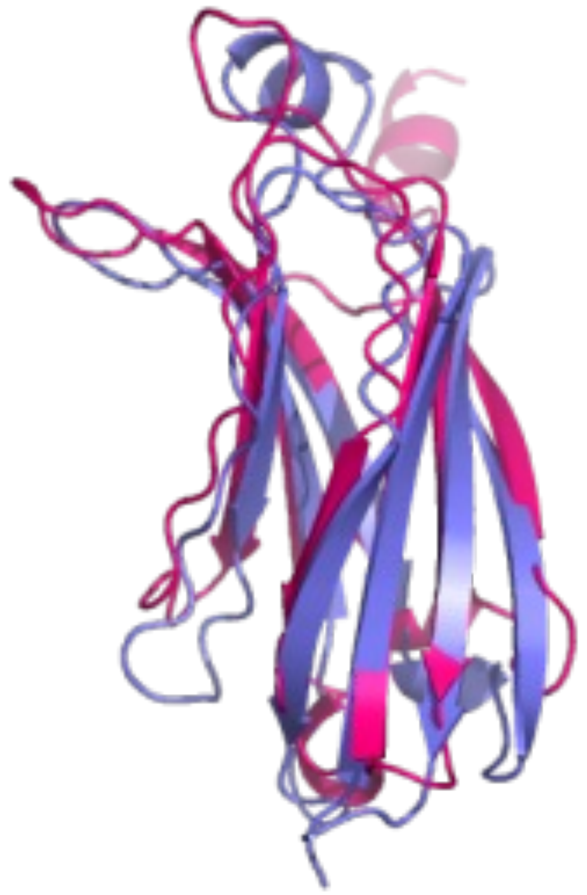


1990s

Fragment Assembly

Rosetta (‘97), 1st CASP (‘94), Threading (‘91), BLAST (‘90)

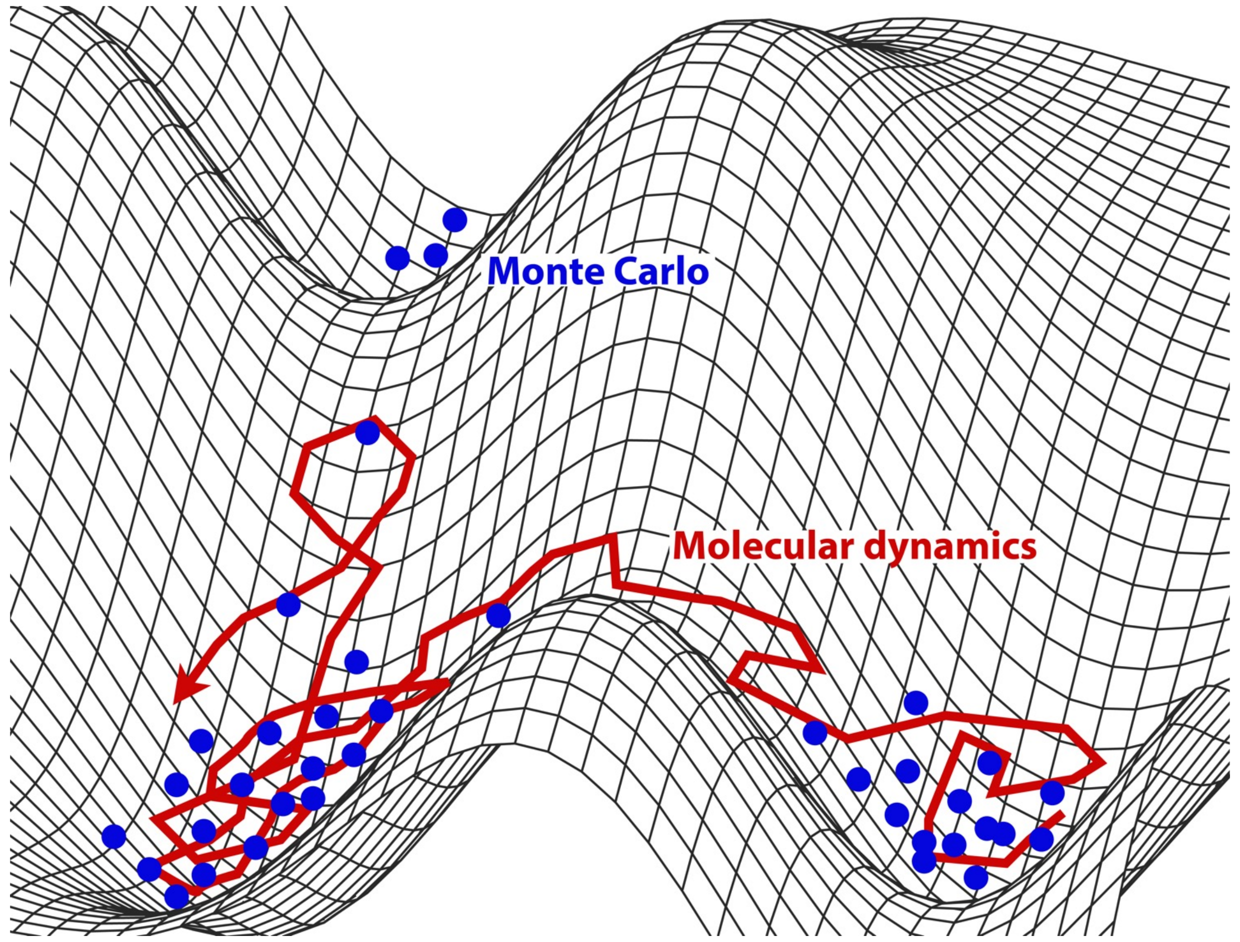
2010s



1980s

Molecular Dynamics

AMBER (‘81), CHARMM (‘83)



2000s

Images: [1] PyMolWiki, [2] Aponte-Santamaria, [3] Wikipedia, [4] Wikipedia, [5] Pixabay

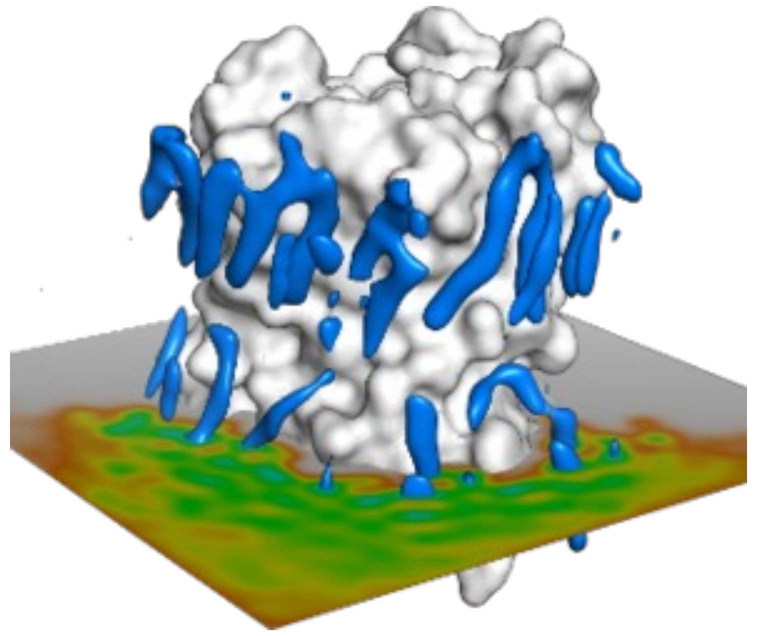
Where do we come from?

The balance between *ab initio* prediction and data-driven methods

1970s

Template-based Modelling (TBM)

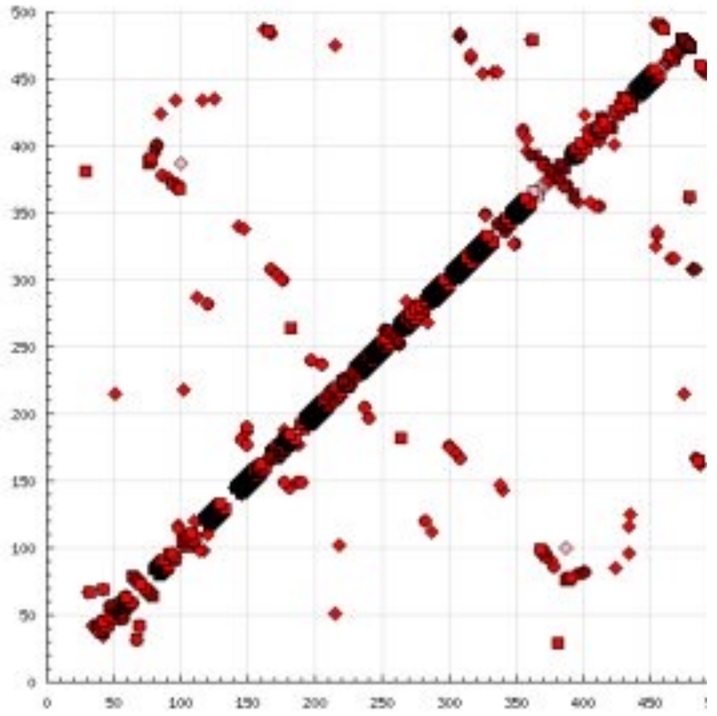
Utilise sequence alignments to ”copy” similar residues



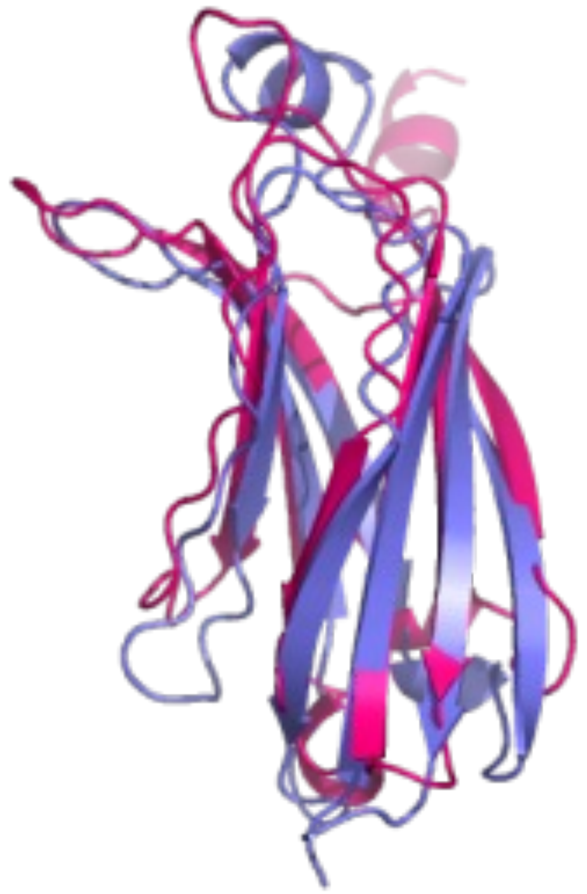
1990s

Fragment Assembly

Rosetta (‘97), 1st CASP (‘94), Threading (‘91), BLAST (‘90)



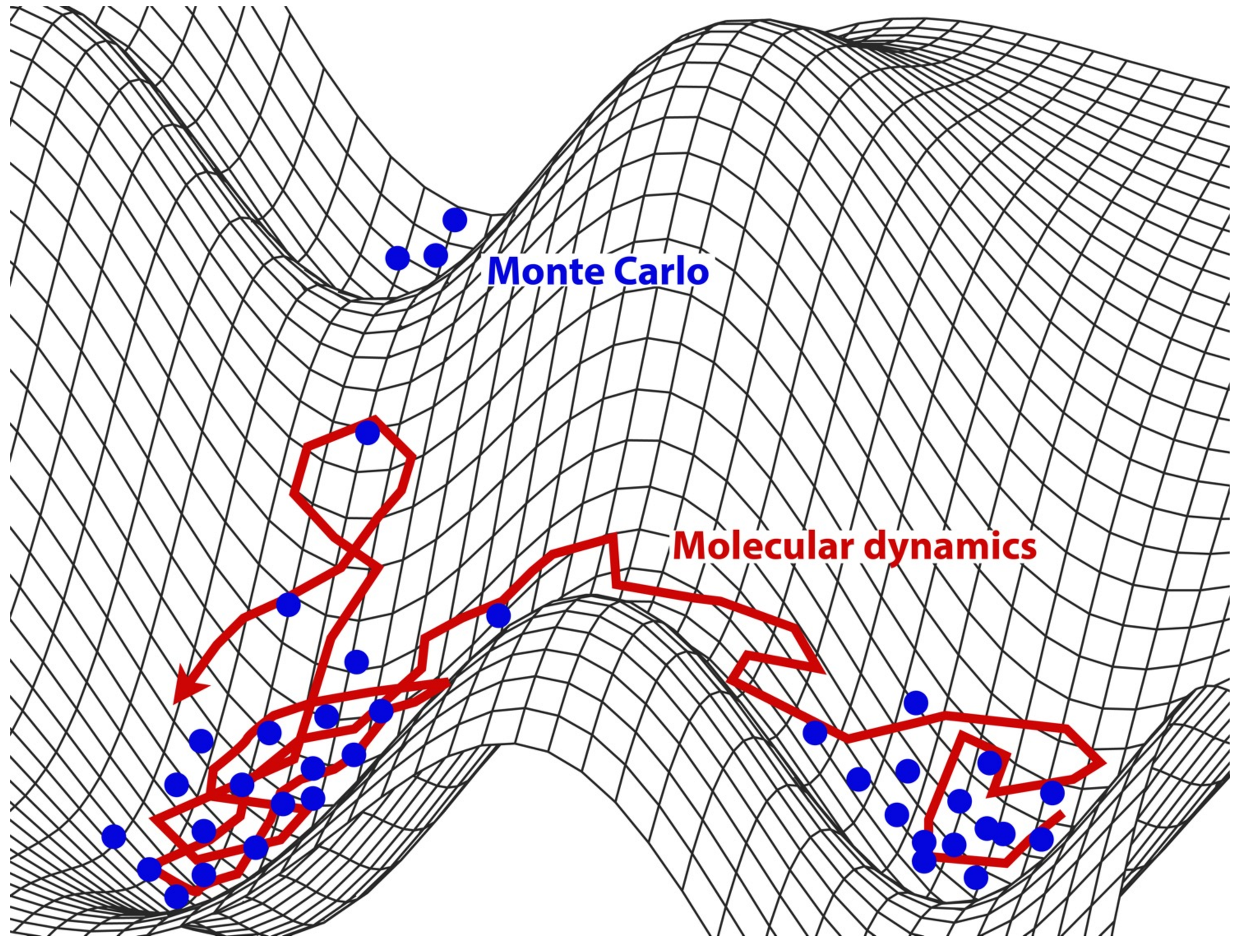
2010s



1980s

Molecular Dynamics

AMBER (‘81), CHARMM (‘83)



2000s

Contact/Distance Map Prediction

Images: [1] PyMolWiki, [2] Aponte-Santamaria, [3] Wikipedia, [4] Wikipedia, [5] Pixabay

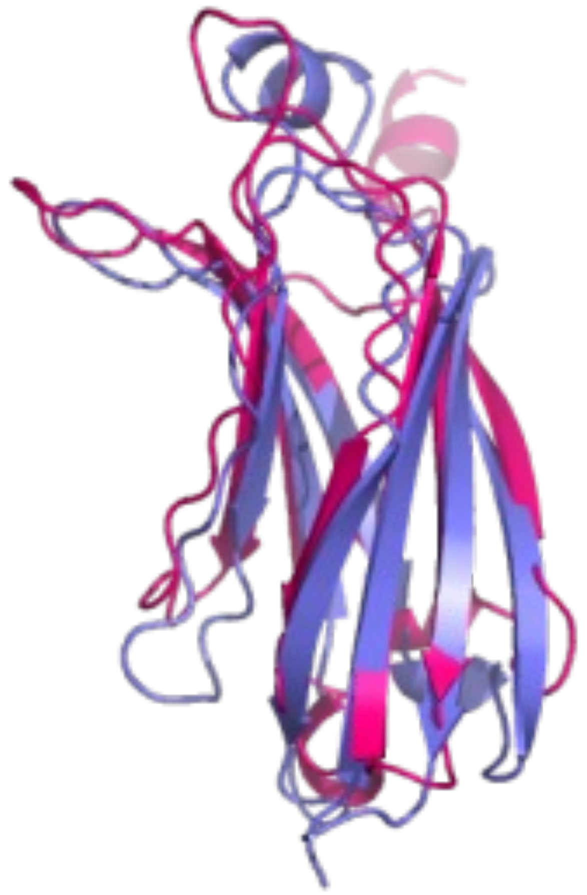
Where do we come from?

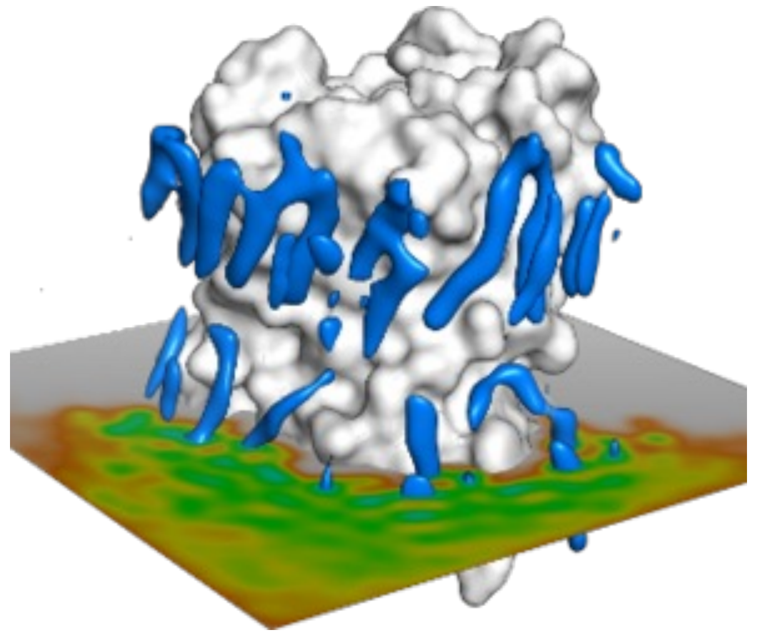
The balance between *ab initio* prediction and data-driven methods

1970s

Template-based Modelling (TBM)

Utilise sequence alignments to ”copy” similar residues

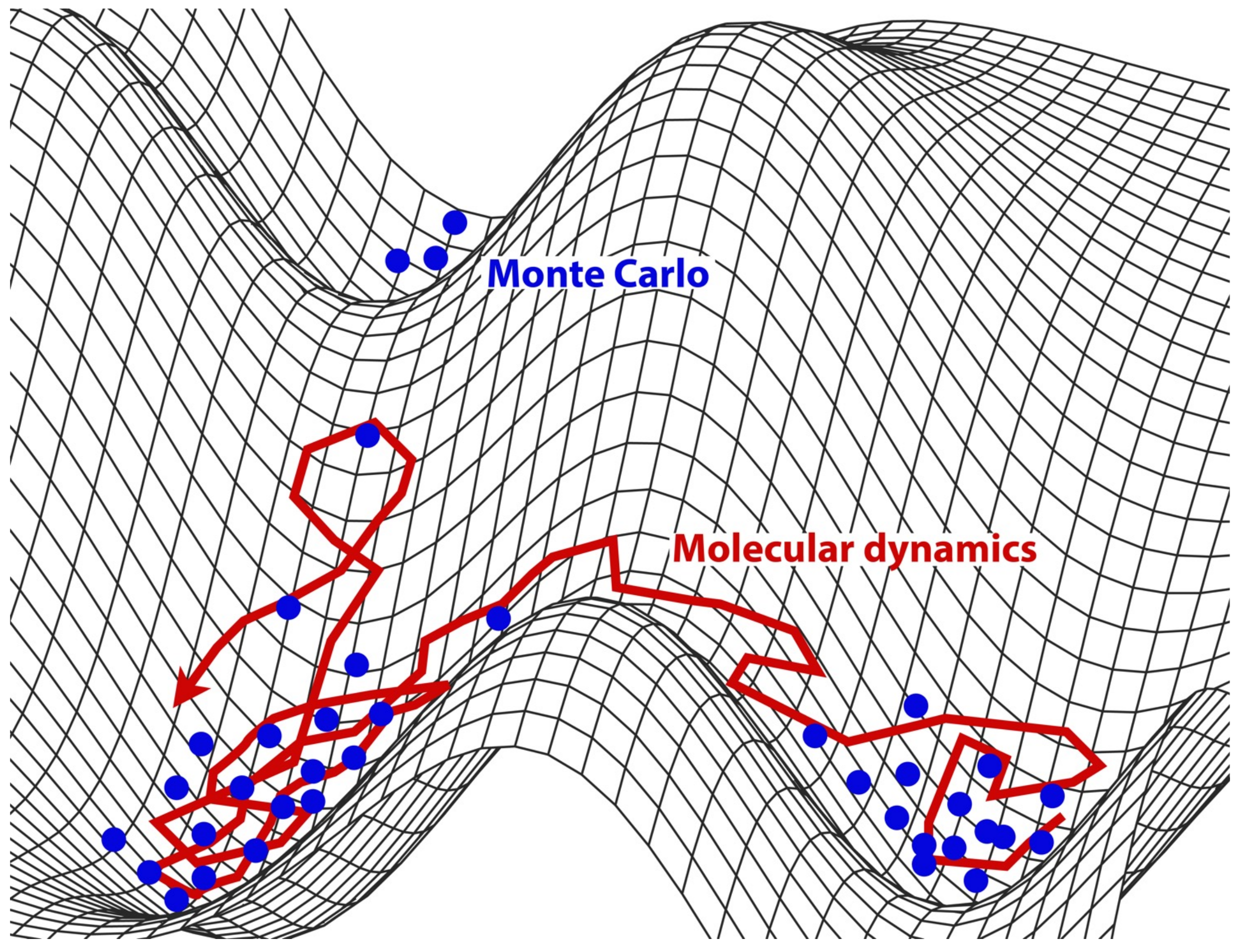




1980s

Molecular Dynamics

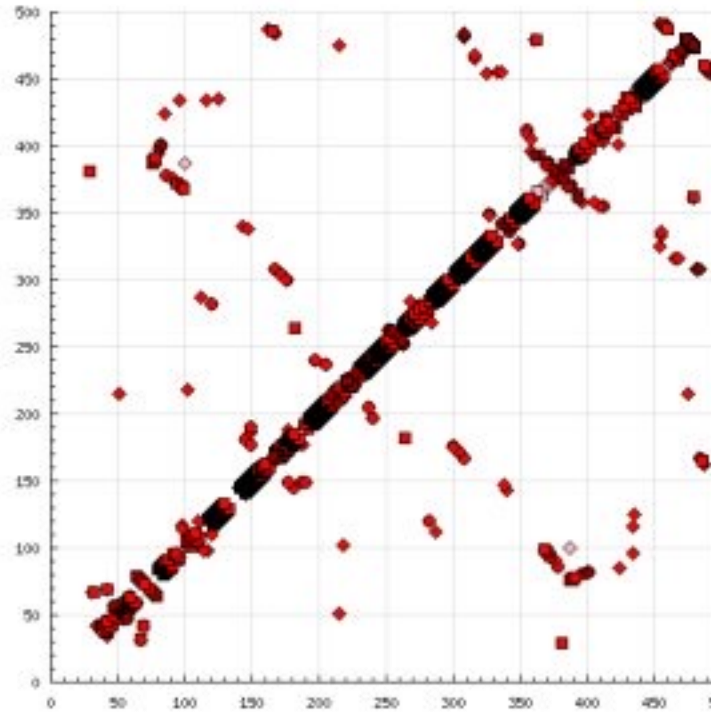
AMBER (‘81), CHARMM (‘83)



1990s

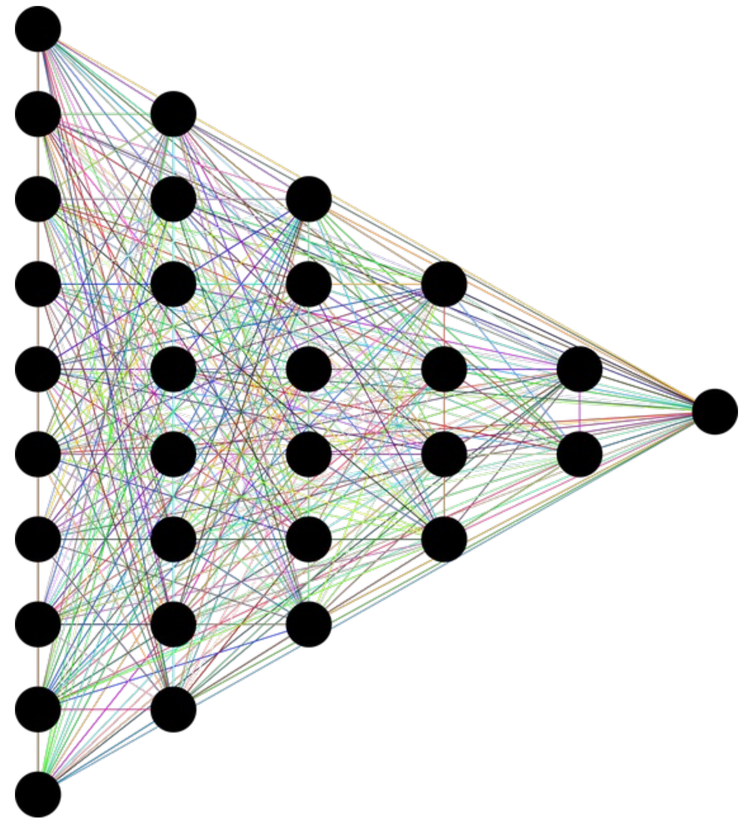
Fragment Assembly

Rosetta (‘97), 1st CASP (‘94), Threading (‘91), BLAST (‘90)



2000s

Contact/Distance Map Prediction



2010s

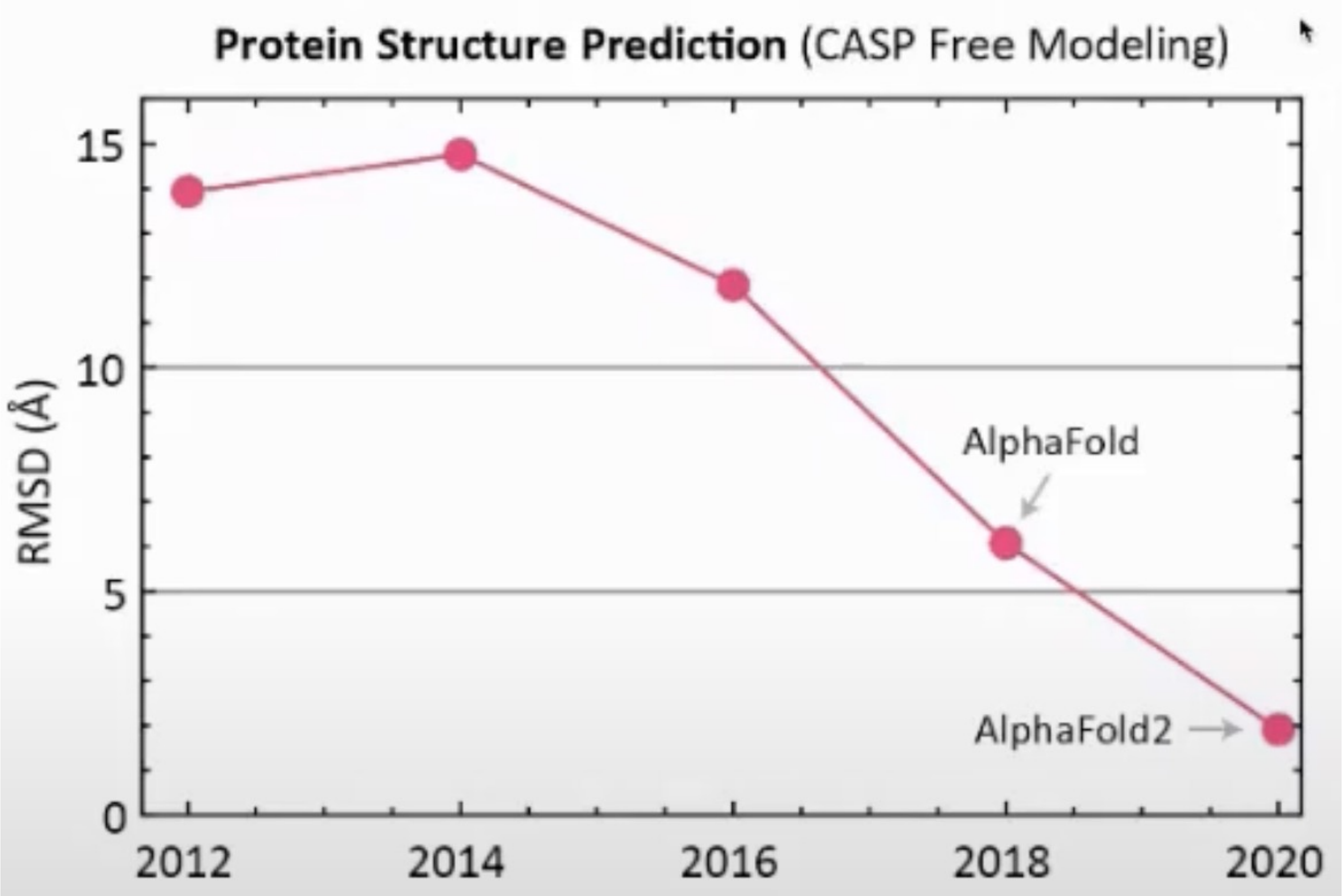
DL, first for maps, then end-to-end

RaptorX (‘17), AF (‘18), RGN (‘19), AF2 (‘20)

Images: [1] PyMolWiki, [2] Aponte-Santamaria, [3] Wikipedia, [4] Wikipedia, [5] Pixabay

Rapid progress in the last years

Deep Learning pushed the latest methods into the usable regime

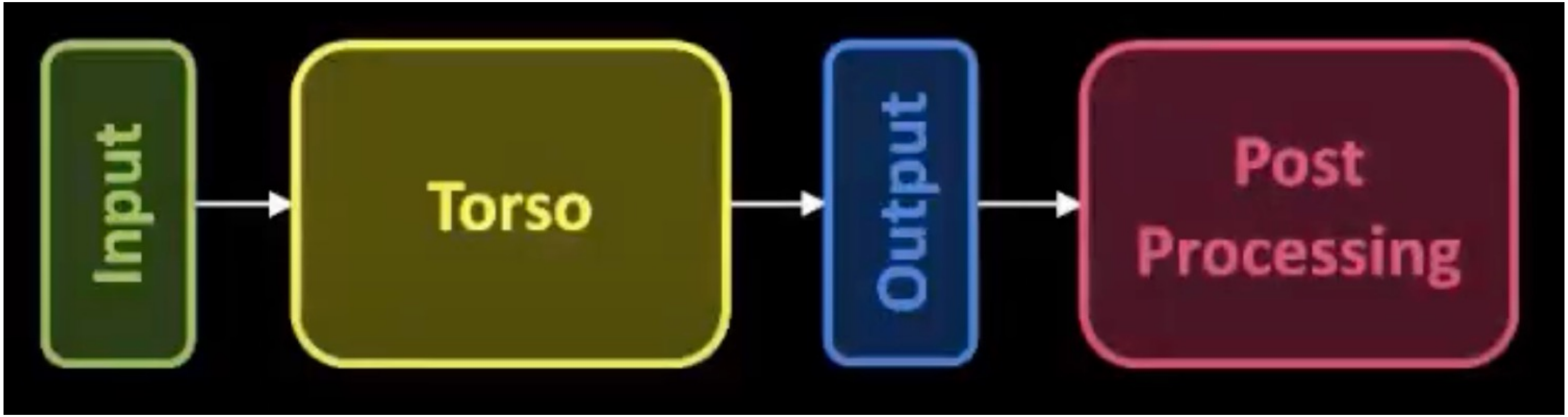


Mohammed AlQuraishi

2. Pre-AlphaFold2 World

How does a folding algorithm look like?

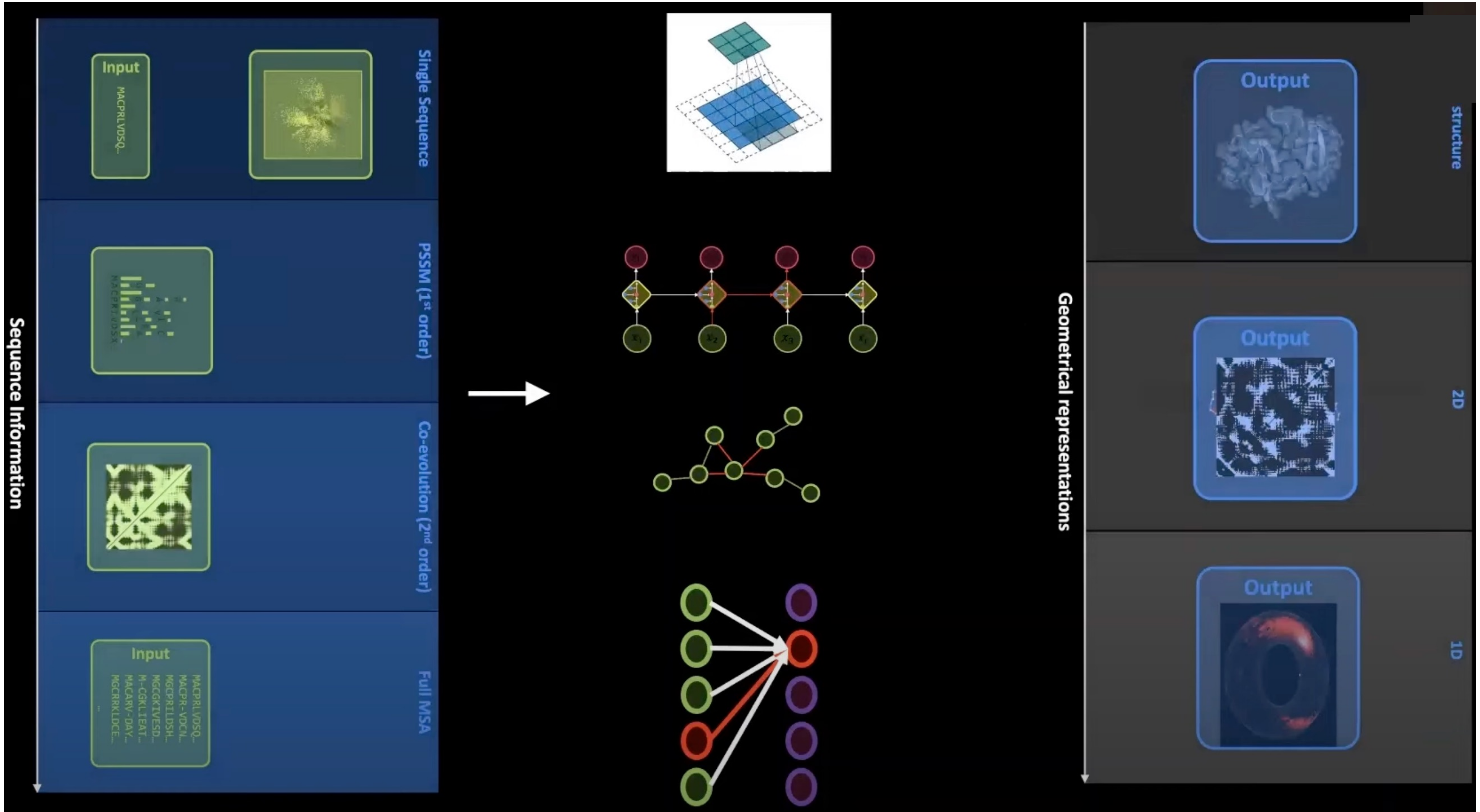
Input and output can vary considerably



Mohammed AlQuraishi/
Nazim Bouatta

How does a folding algorithm look like?

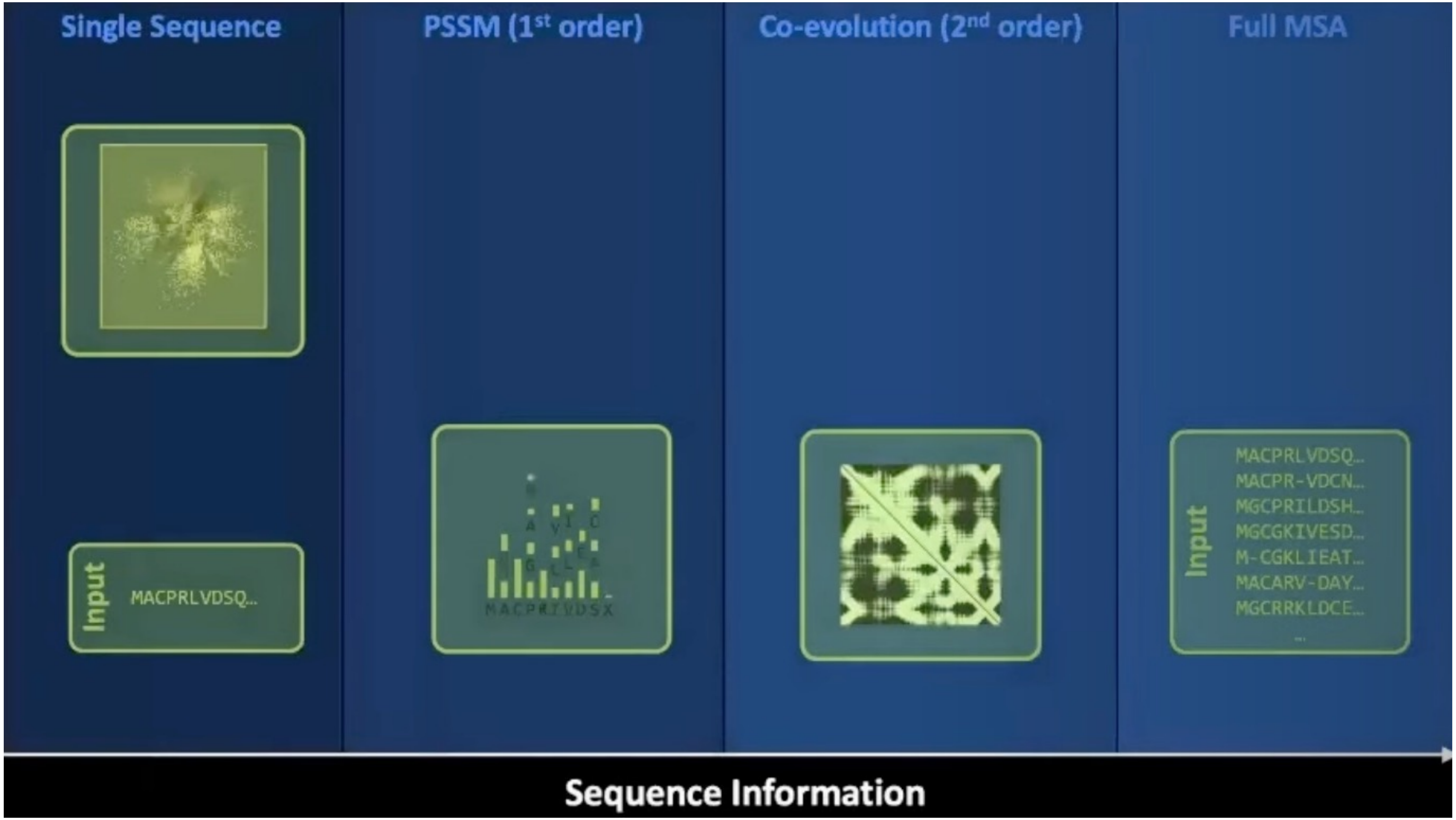
Input and output can vary considerably



Mohammed AlQuraishi/
Nazim Bouatta

What do give our model as input?

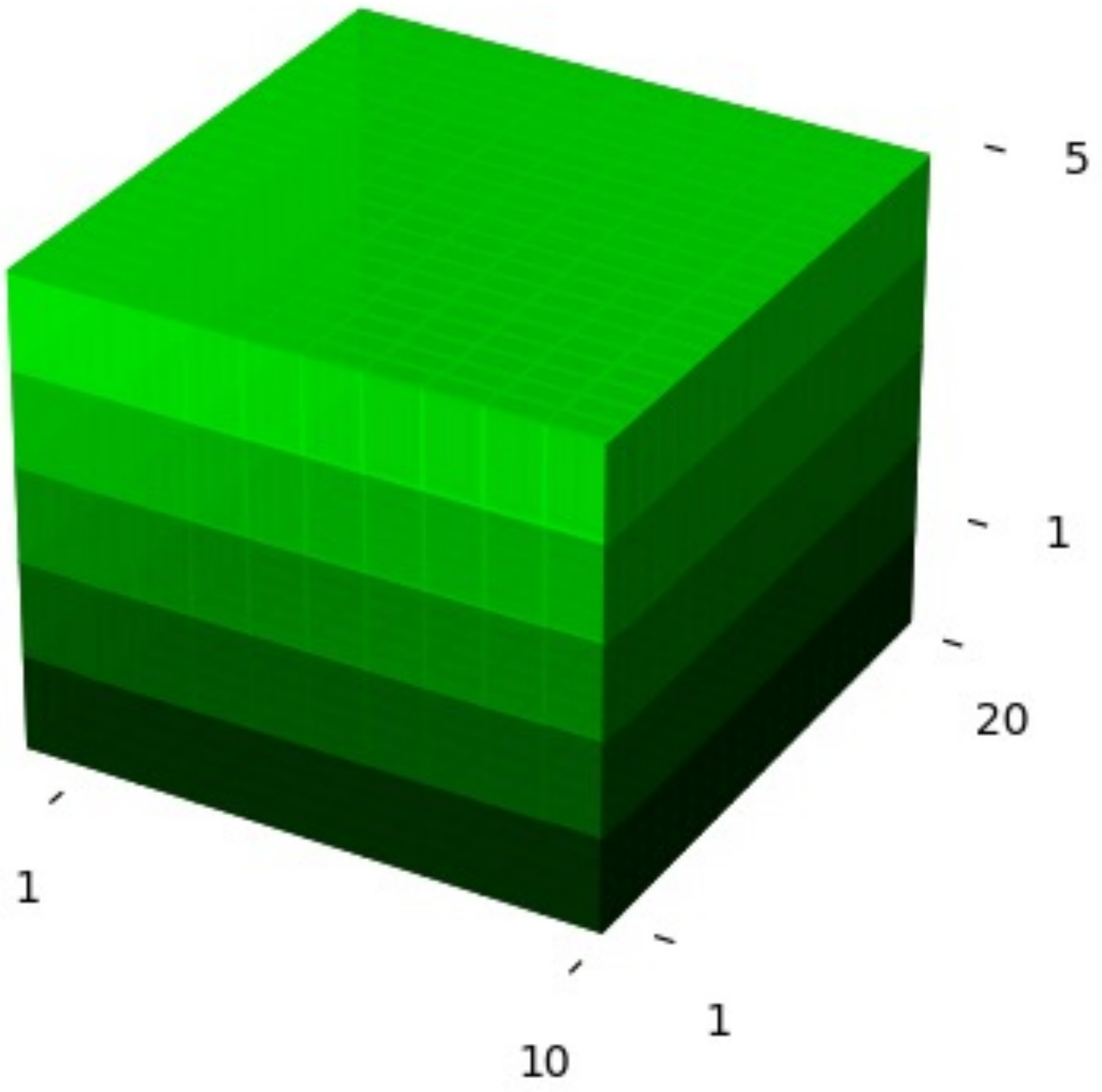
Use evolutionary information to different degrees

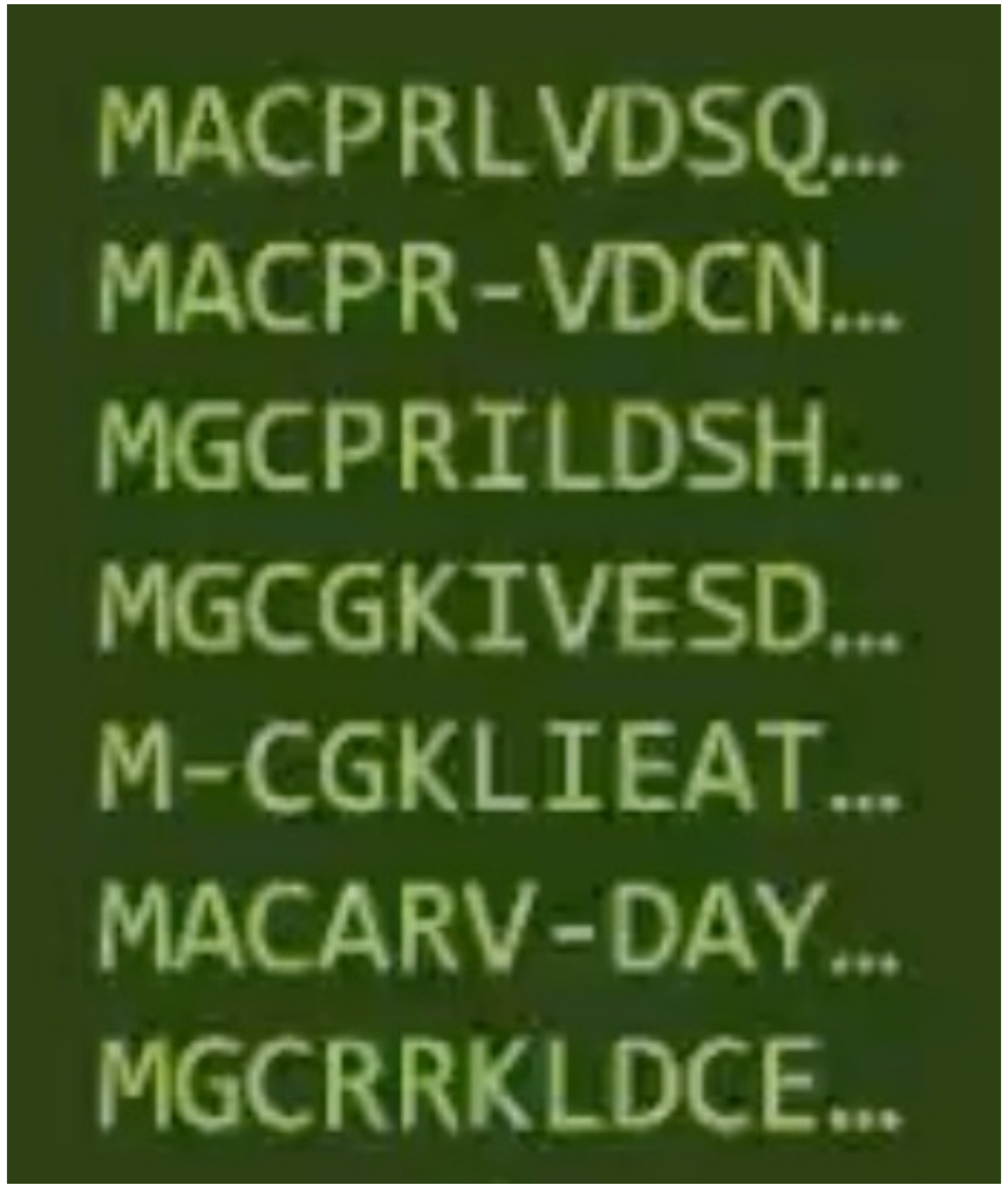


Mohammed AlQuraishi/
Nazim Bouatta

MSA = (#Sequences, Length, 20)

Multiple Sequence Alignment contains all raw information



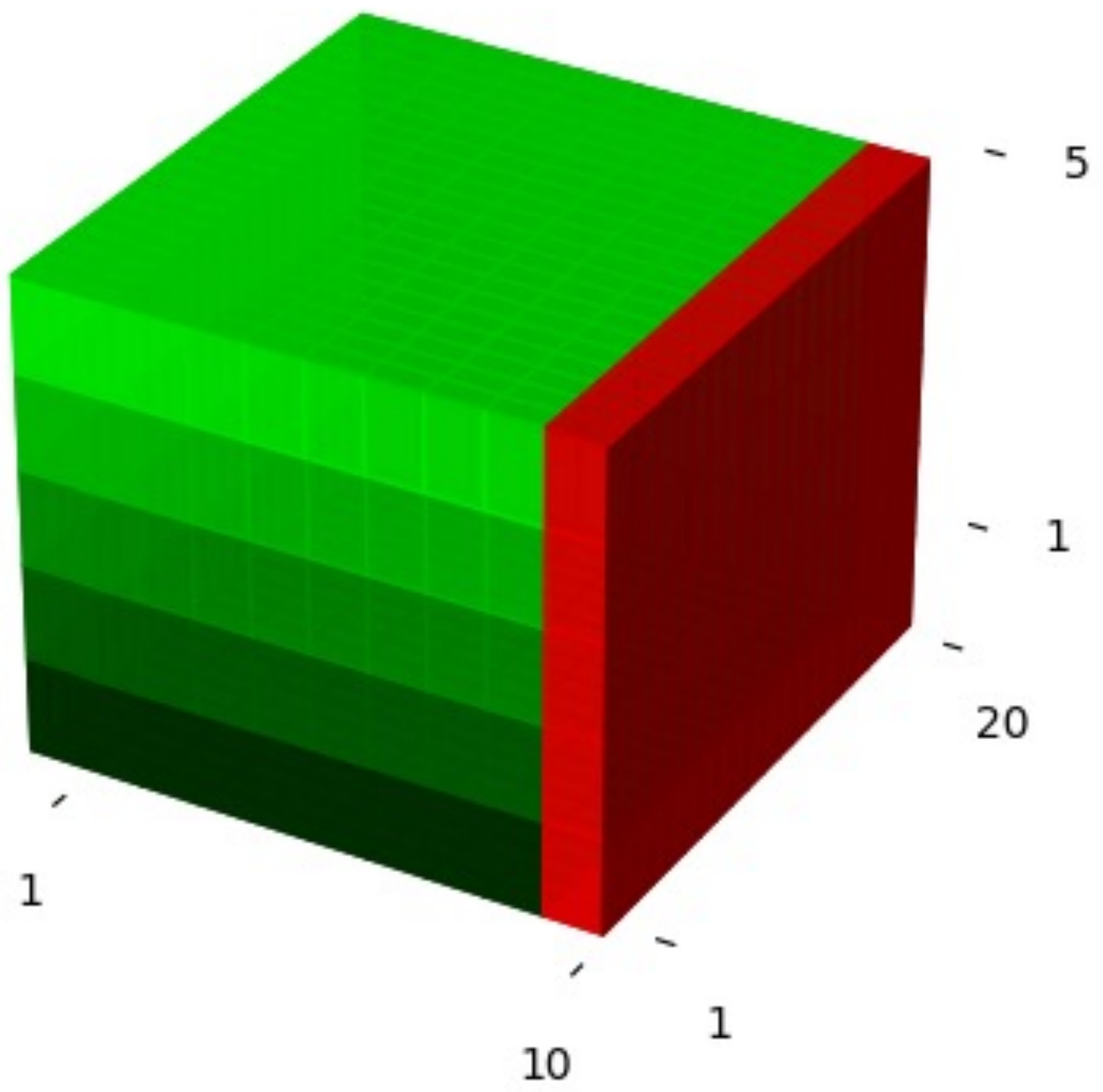


MACPRLVDSQ...
MACPR-VDCN...
MGCPRILDSH...
MGCGKIVESD...
M-CGKLIEAT...
MACARV-DAY...
MGCRRKLDCE...

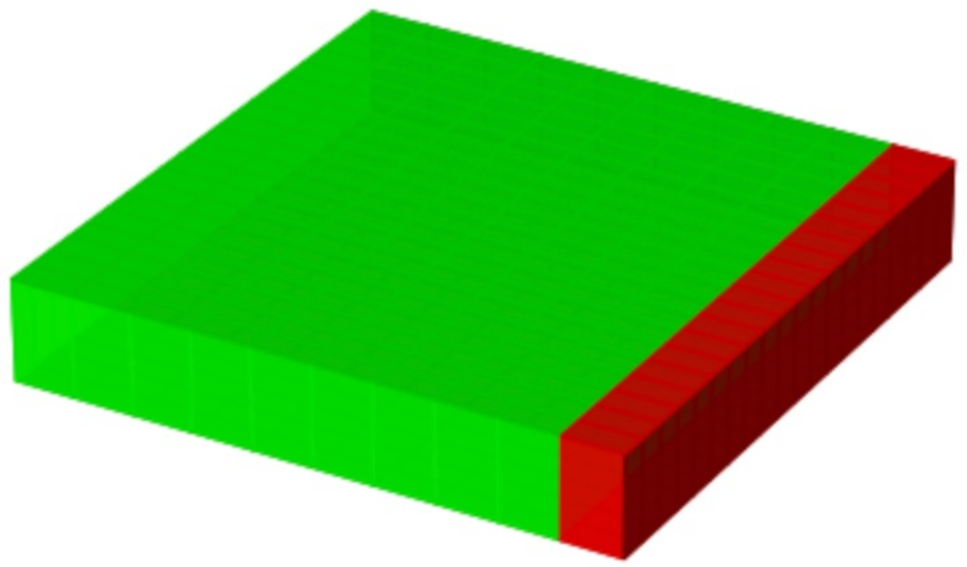
Mohammed AlQuraishi/
Nazim Bouatta

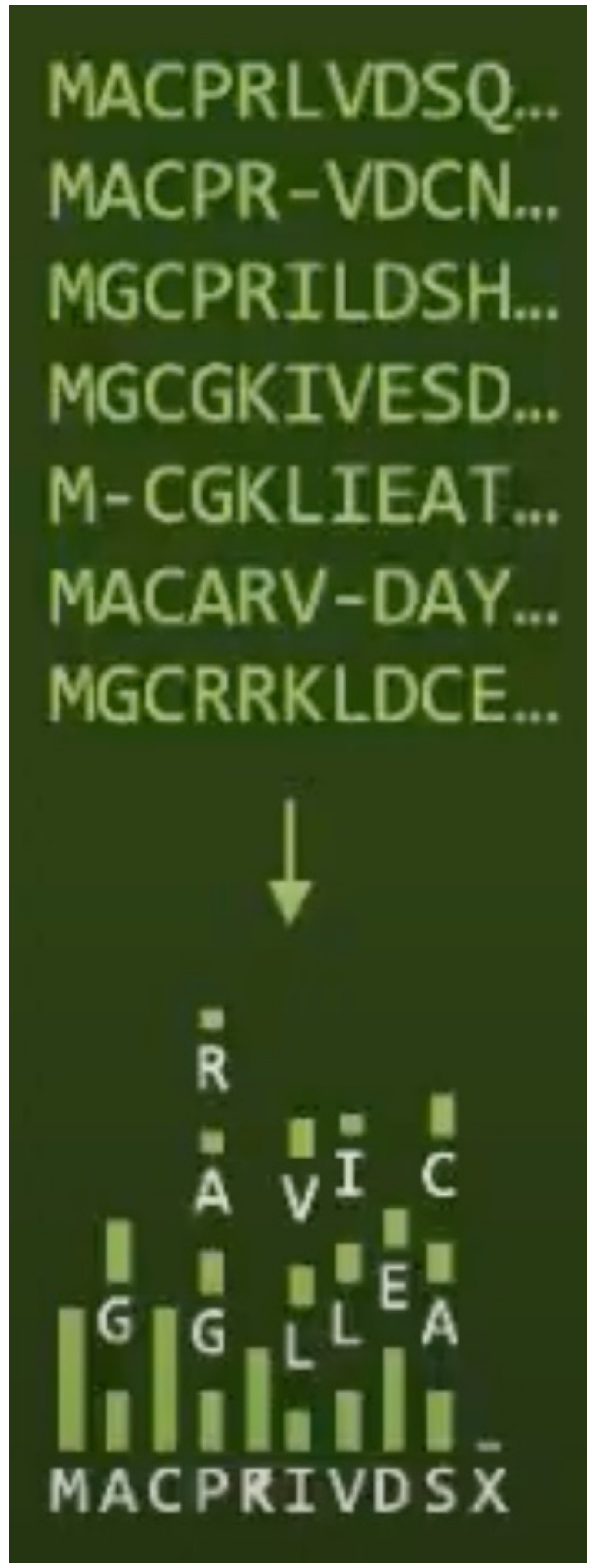
Covariance = (Length, Length)

Covariance conserves 2nd order information



Average across sequences



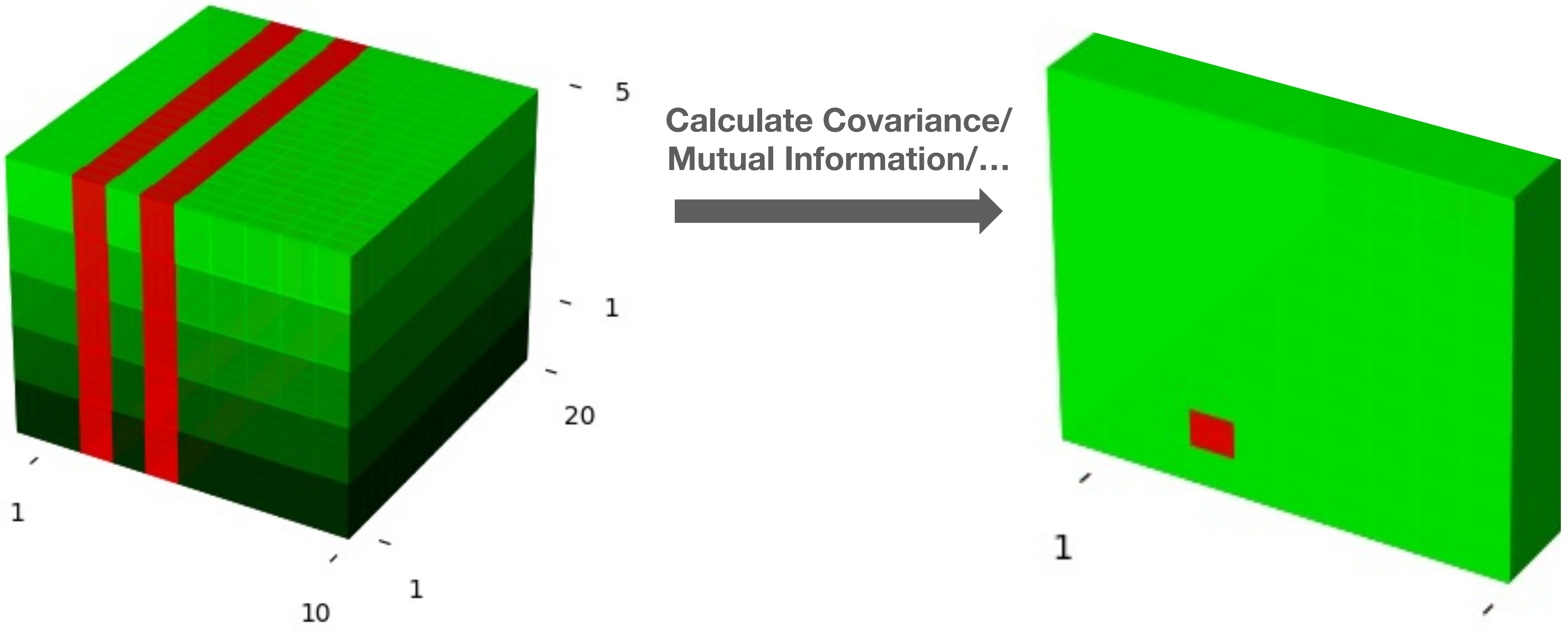


MACPRLVDSQ...
MACPR-VDCN...
MGCPRILDSH...
MGCGKIVESD...
M-CGKLIEAT...
MACARV-DAY...
MGCRRKLDCE...

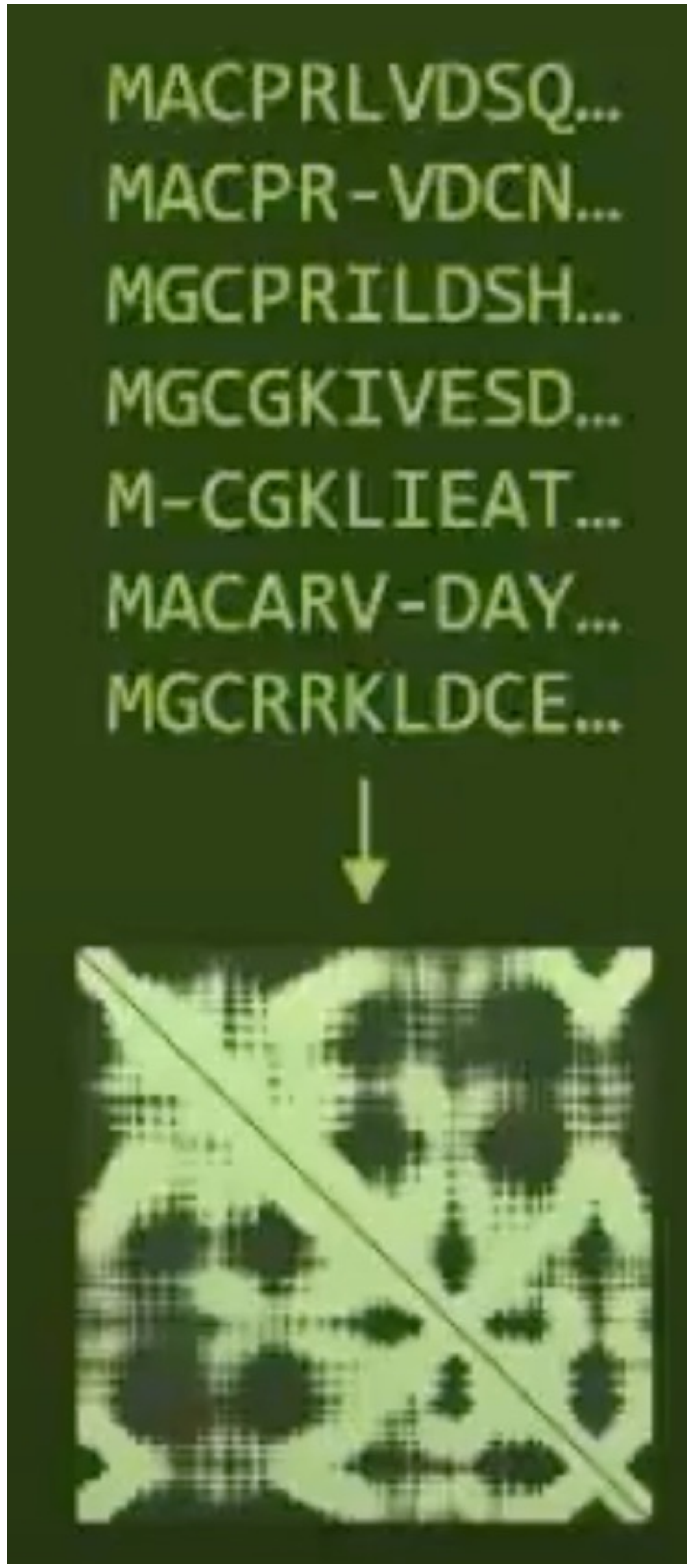
Mohammed AlQuraishi/
Nazim Bouatta

Coevolution = (Length, Length)

Coevolution conserves 2nd order information



Calculate Covariance/
Mutual Information/…

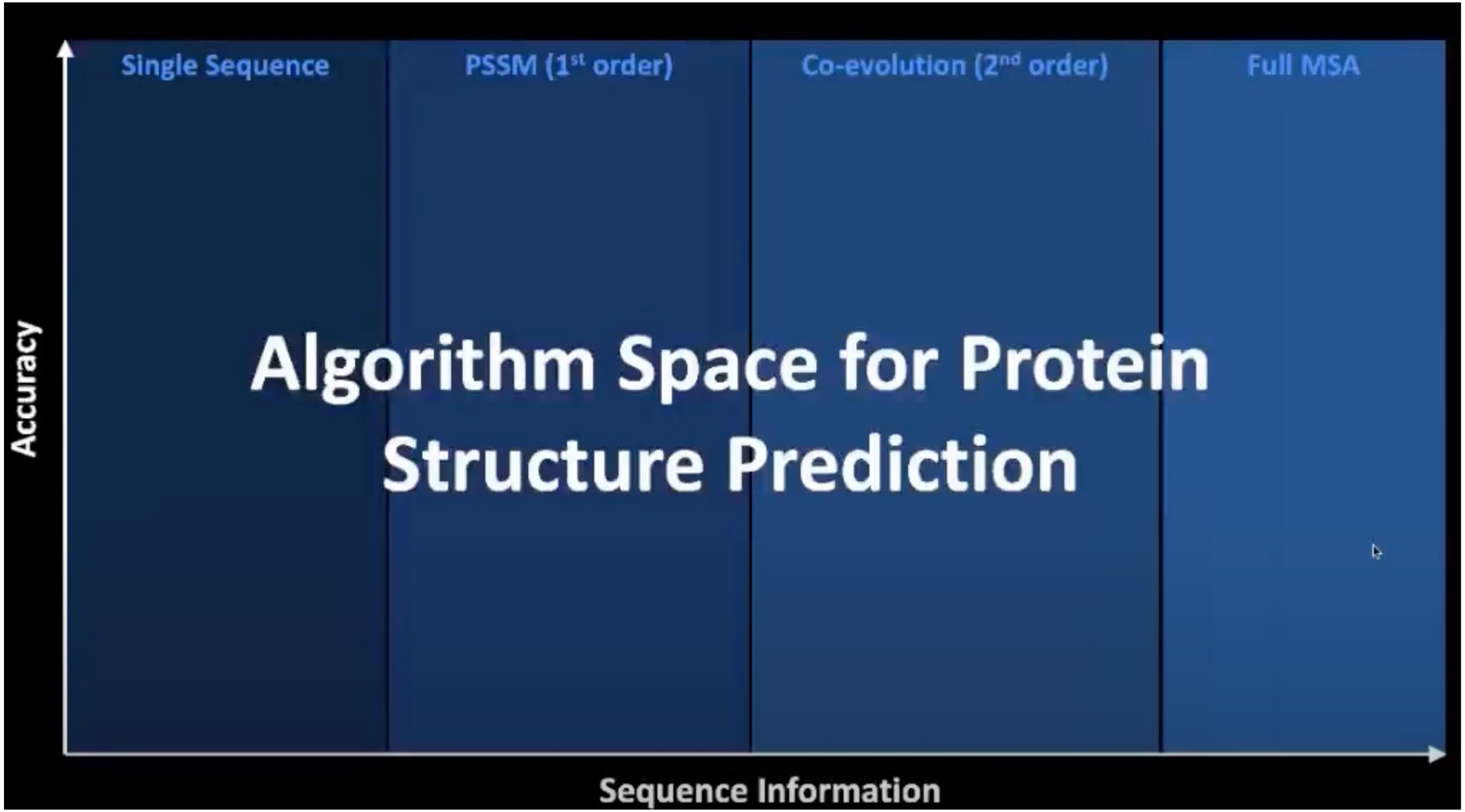


MACPRLVDSQ…
MACPR-VDCN…
MGCPRILDSH…
MGCGKIVESD…
M-CGKLIEAT…
MACARV-DAY…
MGCRRKLDCE…

Mohammed AlQuraishi/
Nazim Bouatta

How do people tackle the problem?

Classifying by what information you feed the model

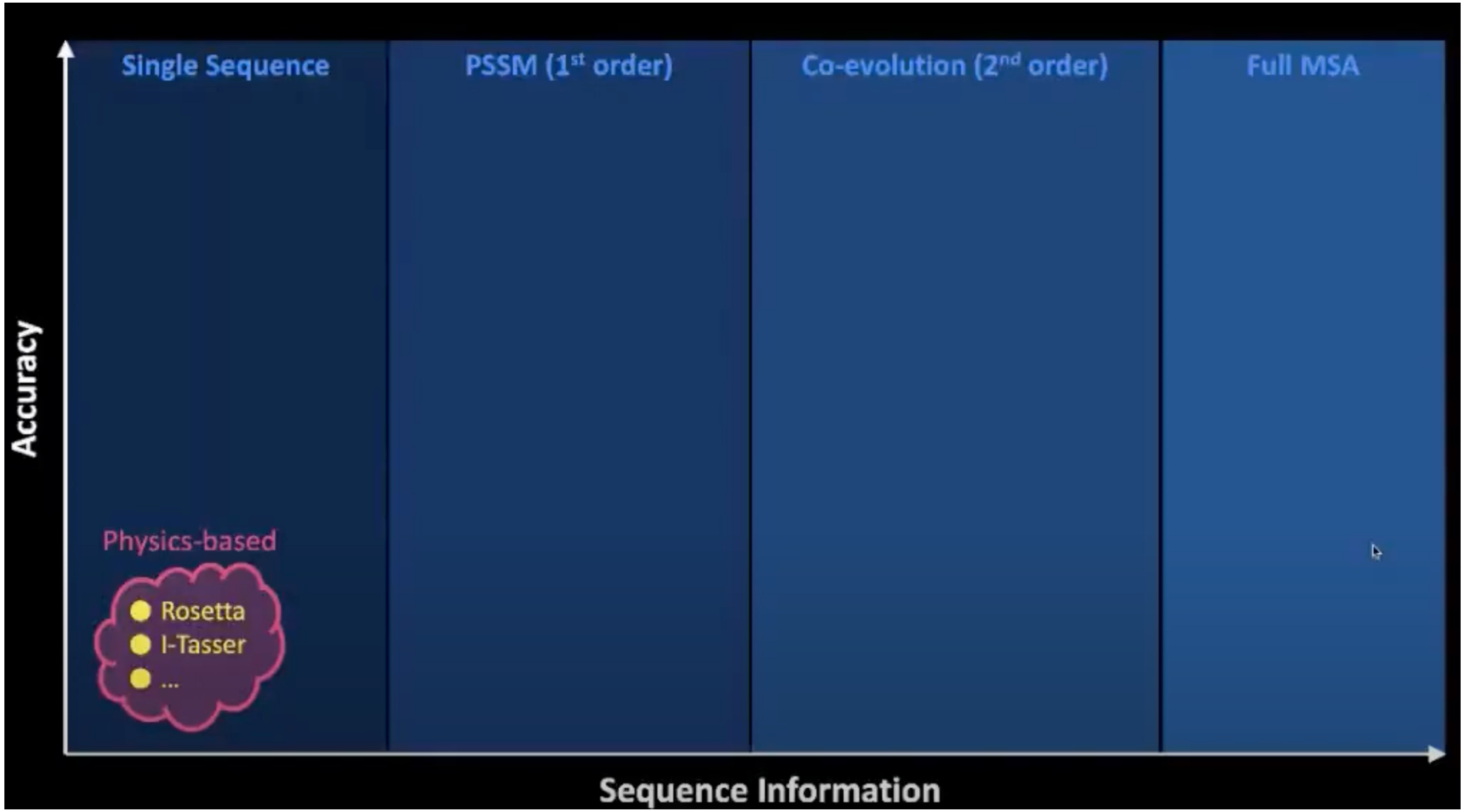


Algorithm Space for Protein Structure Prediction

Mohammed AlQuraishi/
Nazim Bouatta

Physics-based approaches

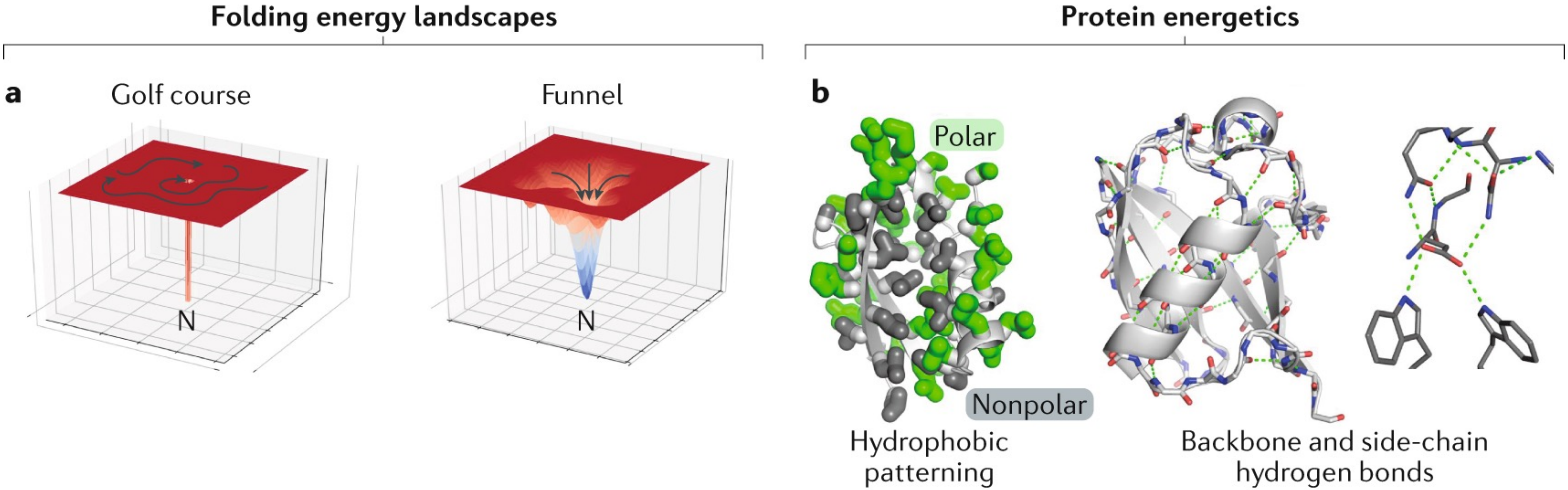
Following Anfinsen to predict structure



Mohammed AlQuraishi/
Nazim Bouatta

Consider energetics to navigate folding

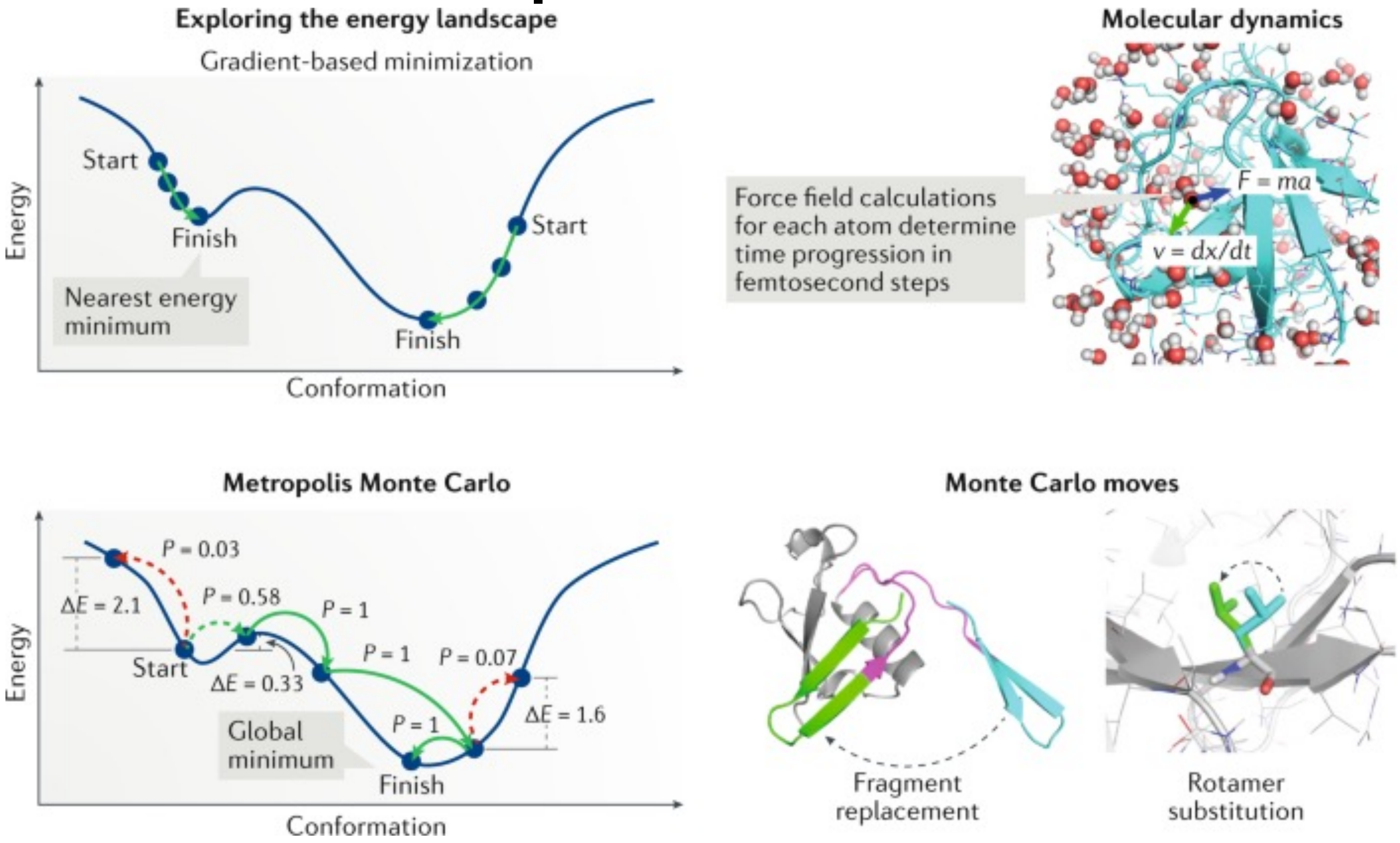
Consistent trends across protein families



Kuhlman, B., Bradley, P. Advances in protein structure prediction and design. *Nat Rev Mol Cell Biol*

Consider energetics to navigate folding

Monte Carlos Methods proved to be most efficient here



Exploring the energy landscape

Molecular dynamics

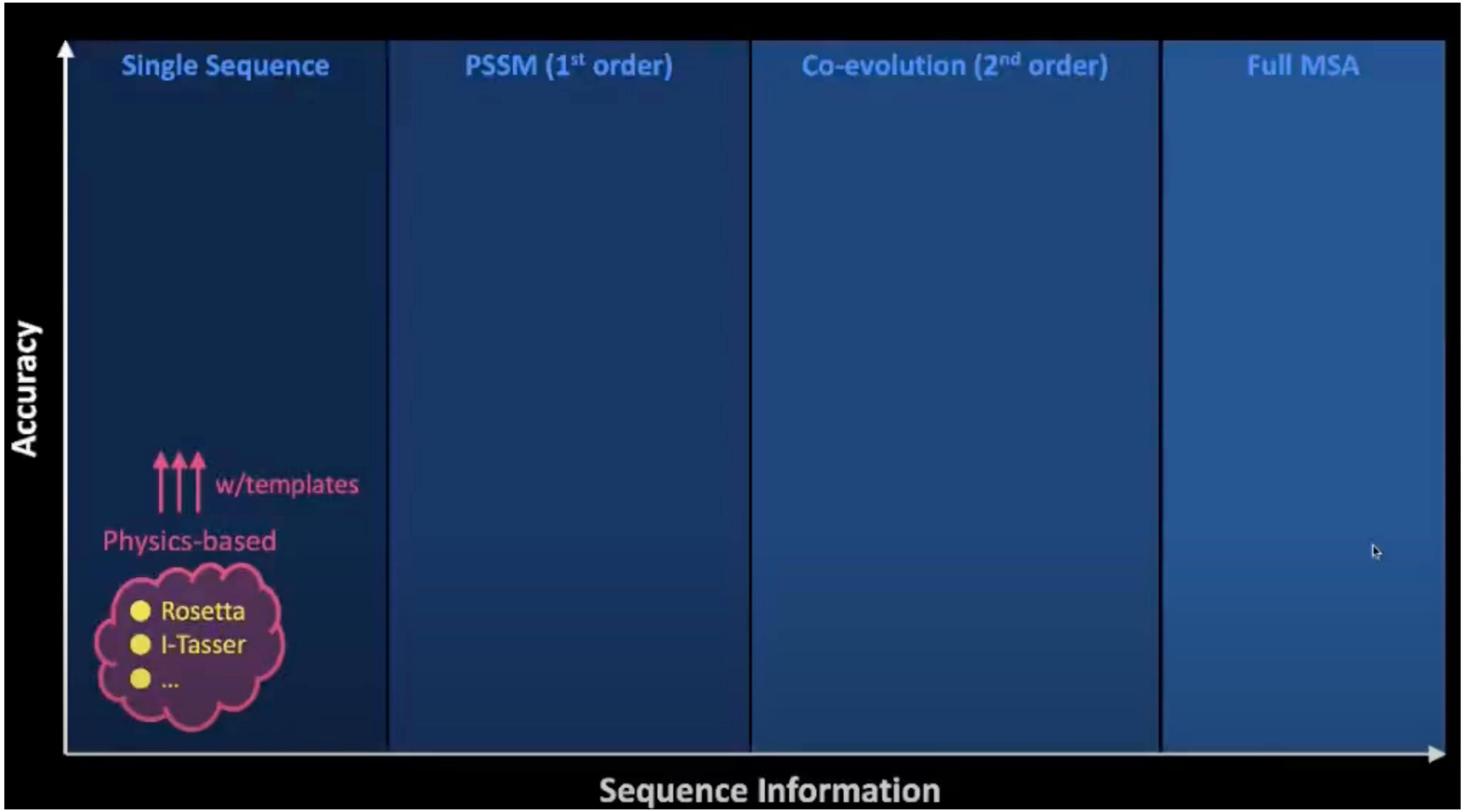
Metropolis Monte Carlo

Monte Carlo moves

Kuhlman, B., Bradley, P. Advances in protein structure prediction and design. *Nat Rev Mol Cell Biol*

Templates improve structure prediction

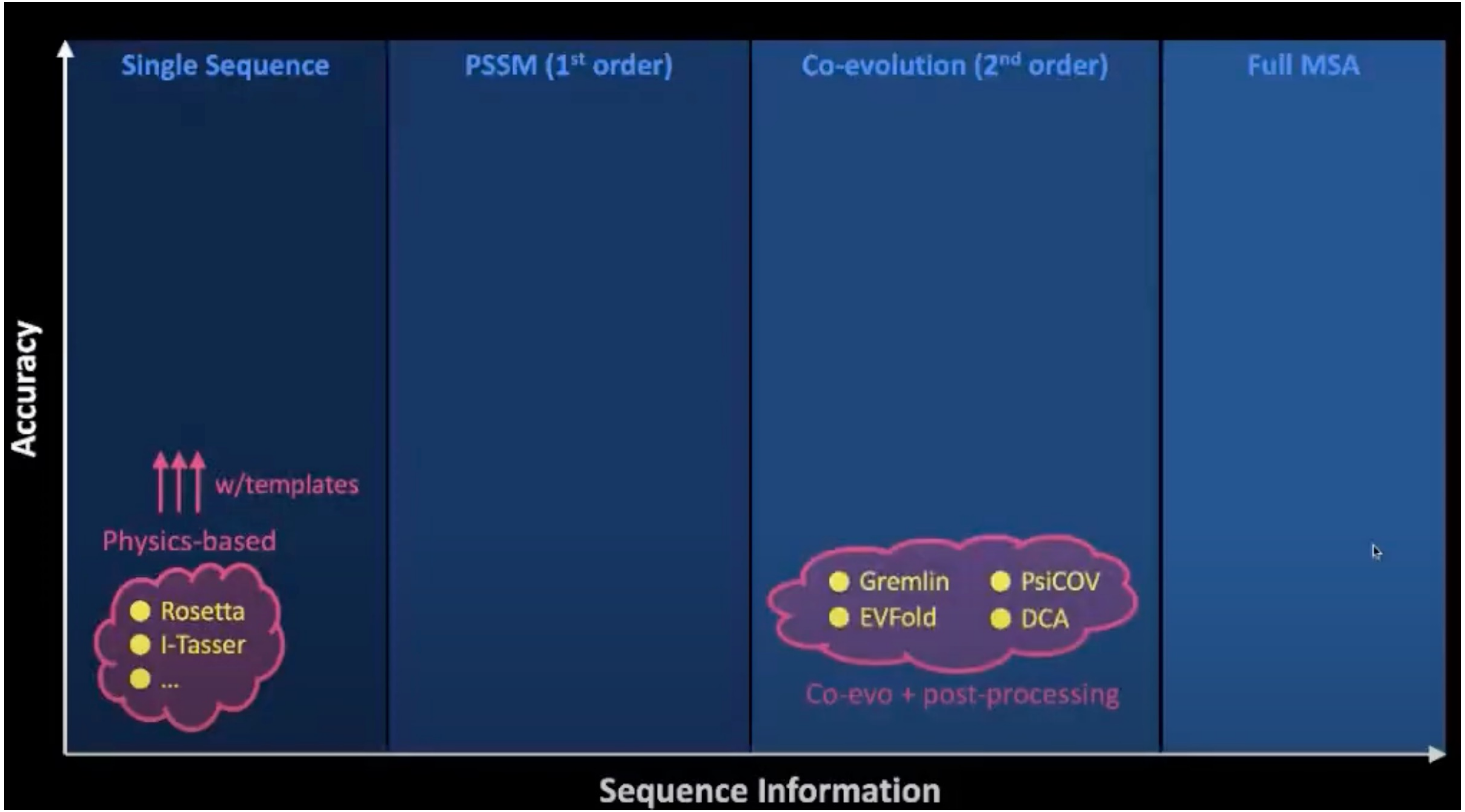
Templates can be found with sequence alignments



Mohammed AlQuraishi/
Nazim Bouatta

Coevolution offered a different approach

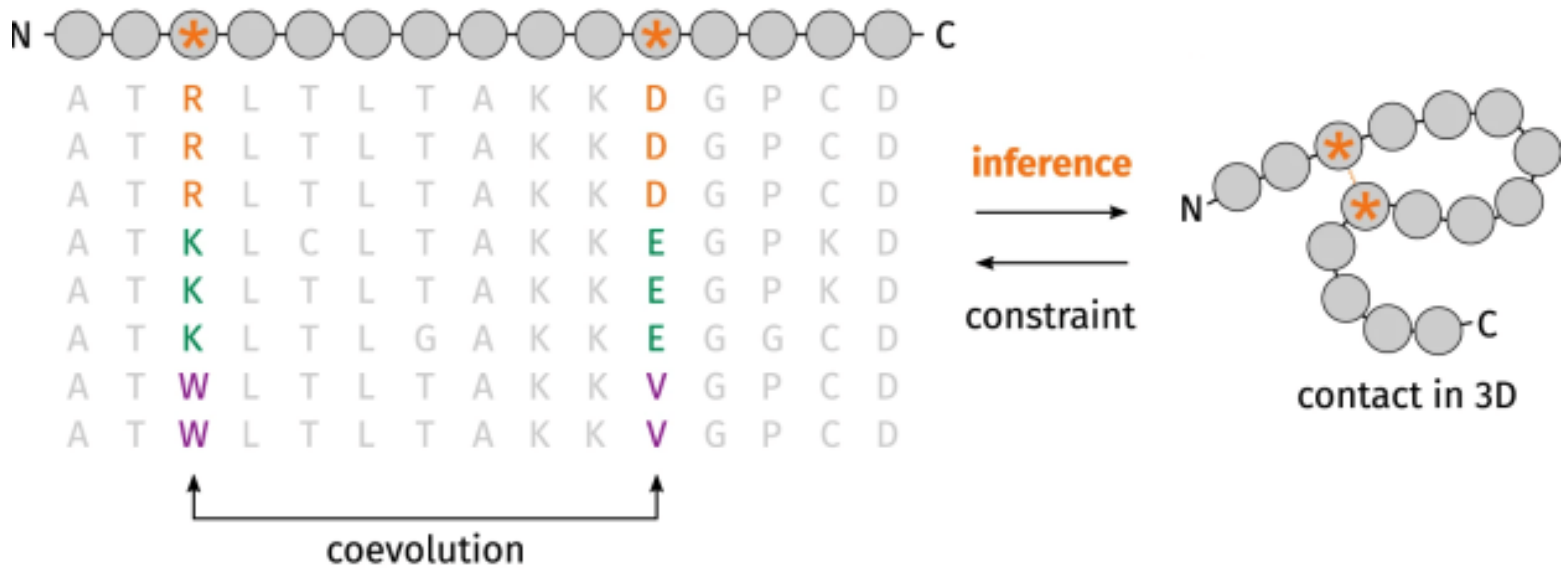
Use evolutionary information to infer geometric constraints



Mohammed AlQuraishi/
Nazim Bouatta

Coevolution: The Idea

Residues that correlate are probably close in space

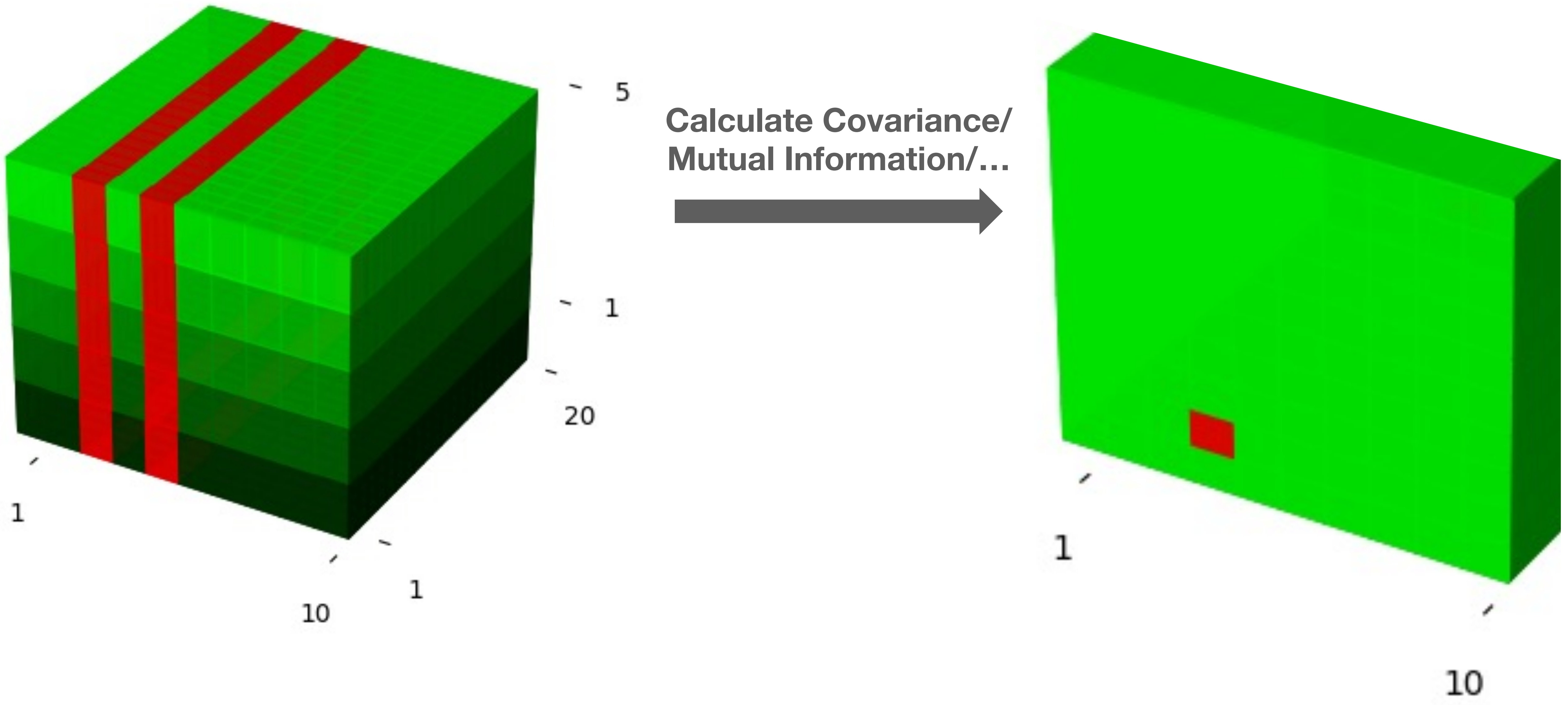


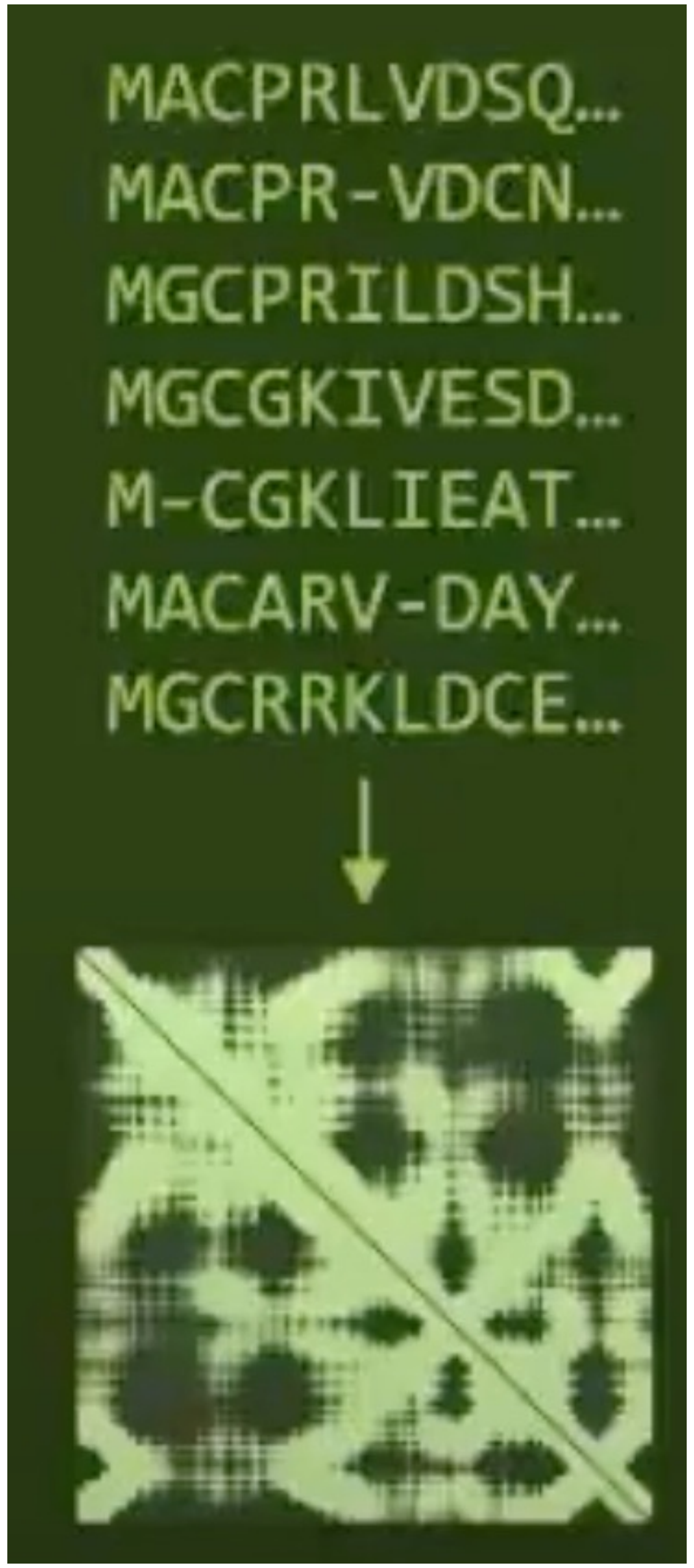
```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
```

Carlos Outeiral Rubiera

Coevolution = (Length, Length)

Coevolution conserves 2nd order information

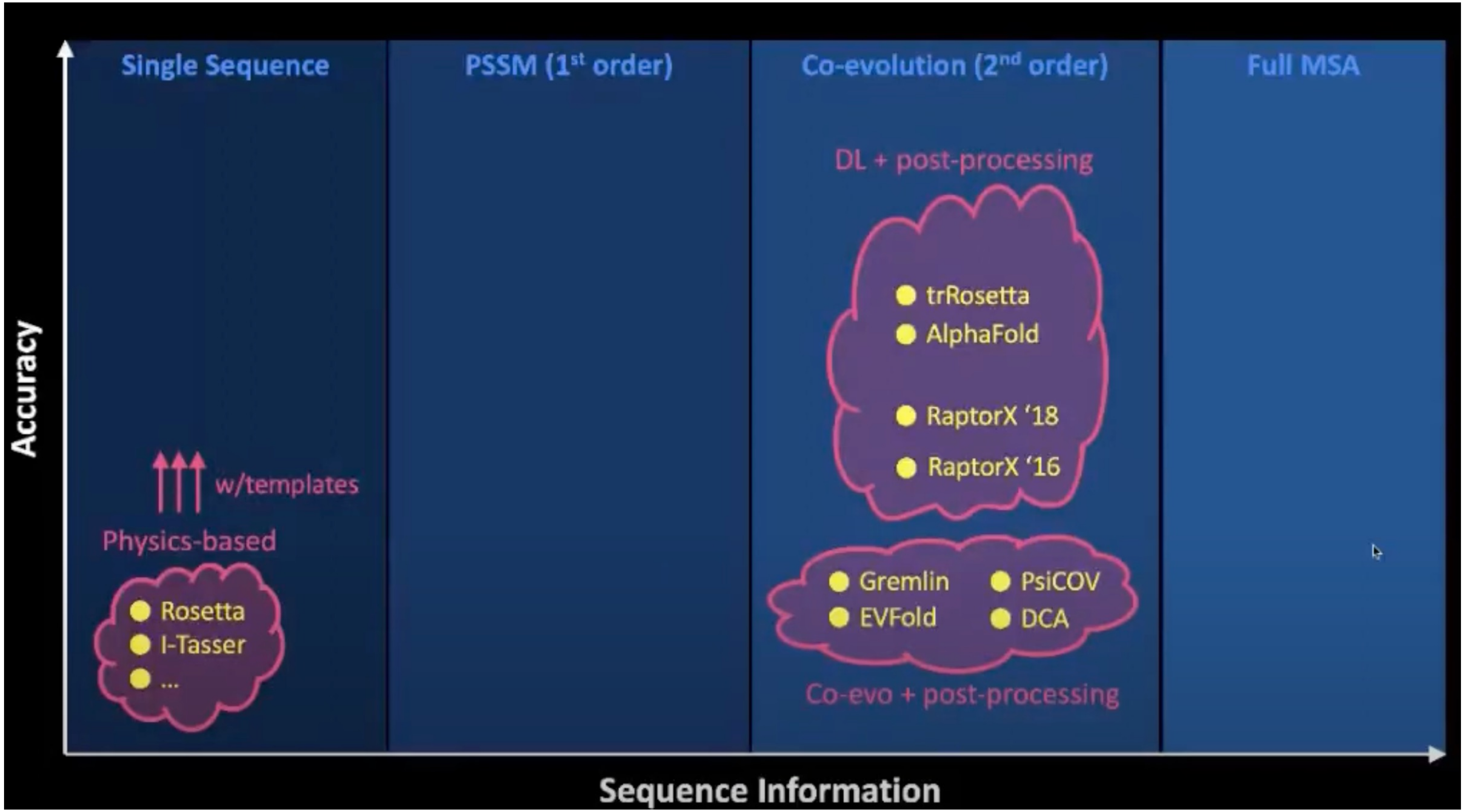




Mohammed AlQuraishi/
Nazim Bouatta

Deep Learning pushed co-evolution methods

Advances from Computer Vision translated to Proteins



Mohammed AlQuraishi/
Nazim Bouatta

Reminder: Image-to-Image

CNNs detect localised patterns

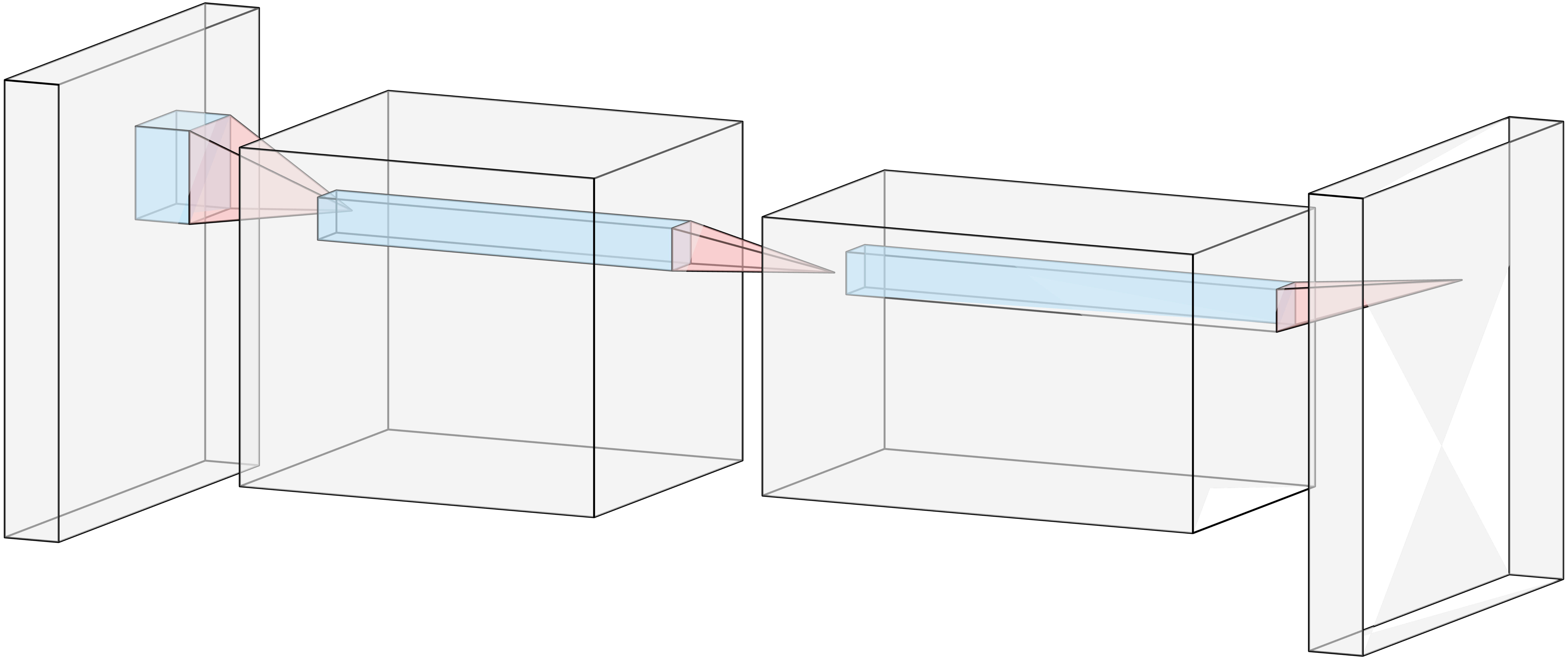
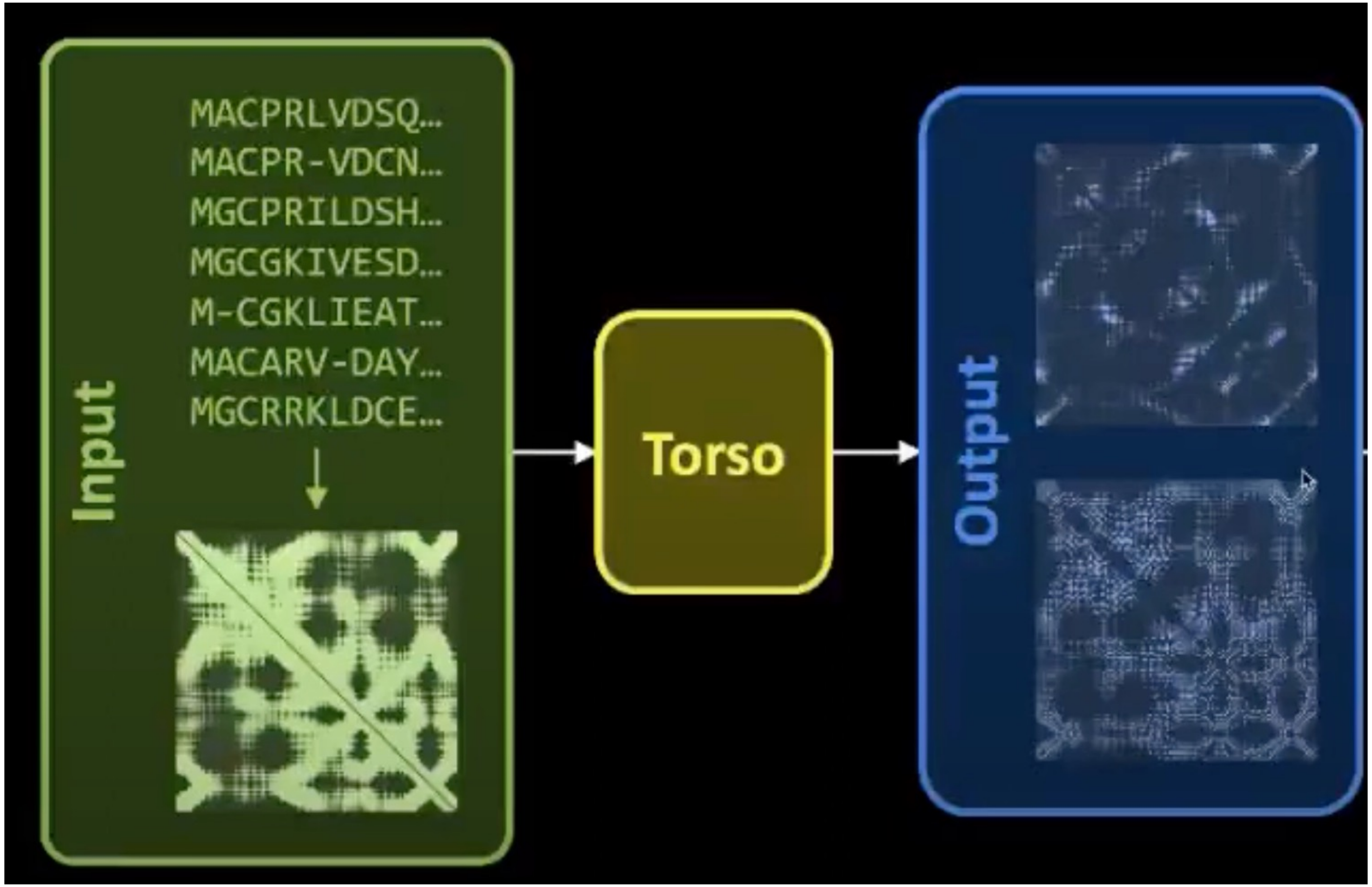


Image-to-Image

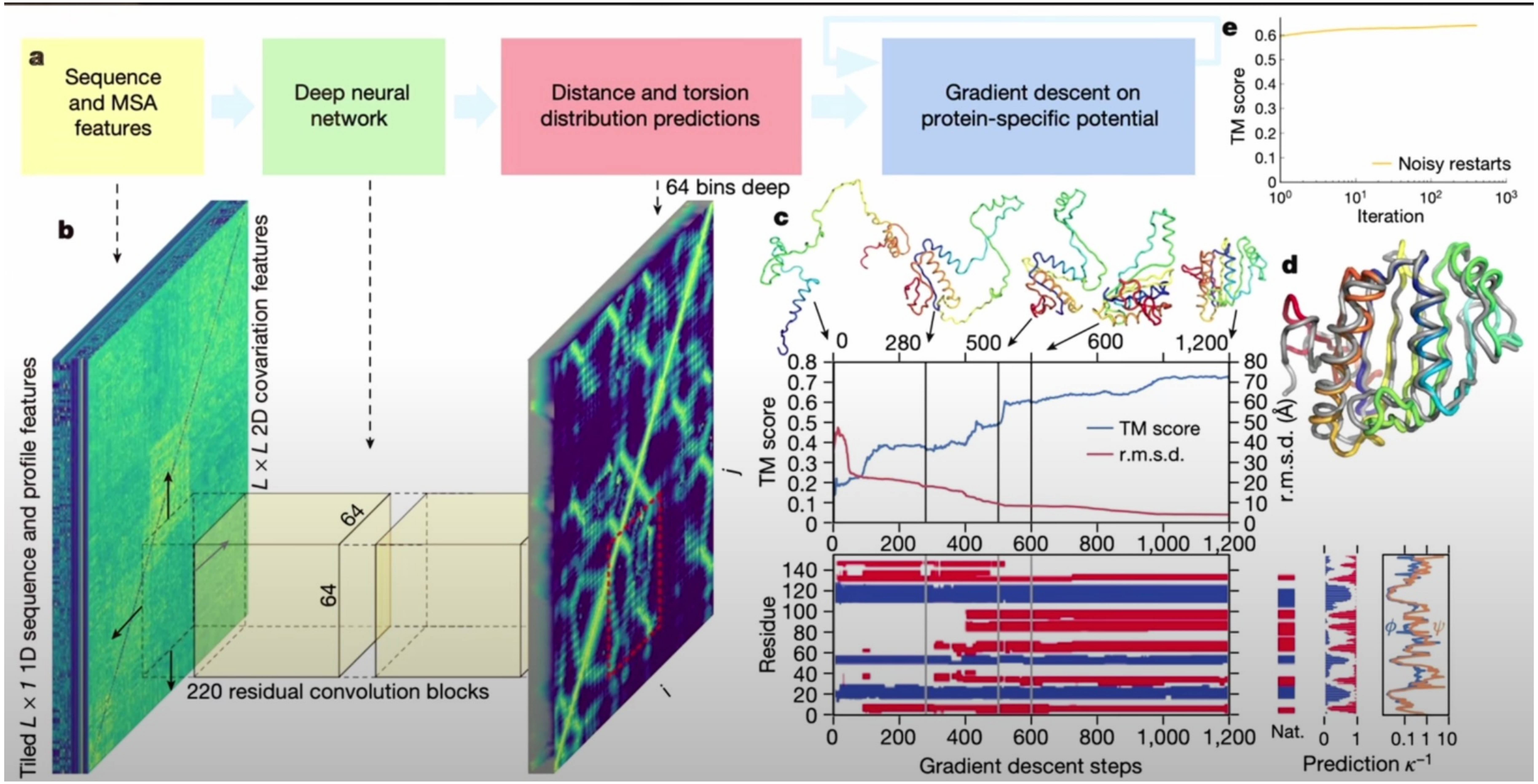
Coevolution data used to predict contact/distance maps



Mohammed AlQuraishi/
Nazim Bouatta

AF1: An Image-to-Image Model

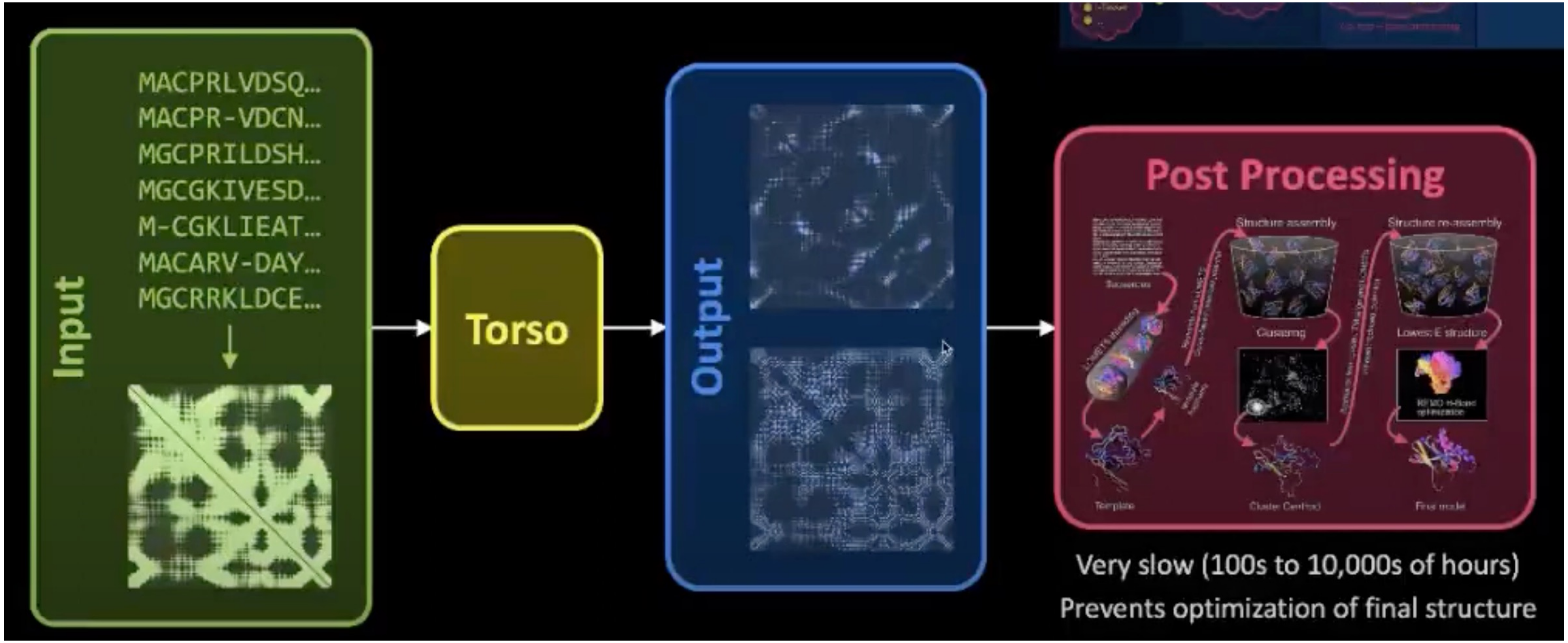
Residual CNN used to predict distances and torsion angles



Senior, A.W., Evans, R., Jumper, J. *et al.* Improved protein structure prediction using potentials from deep learning. Nature

Image-to-Image

Problem: Slow and inconsistent processing into final structure



Input

MACPRLVDSQ...
MACPR-VDCN...
MGCPRILDSH...
MGCGKIVESD...
M-CGKLIEAT...
MACARV-DAY...
MGCRRKLDCE...

Torso

Output

Post Processing

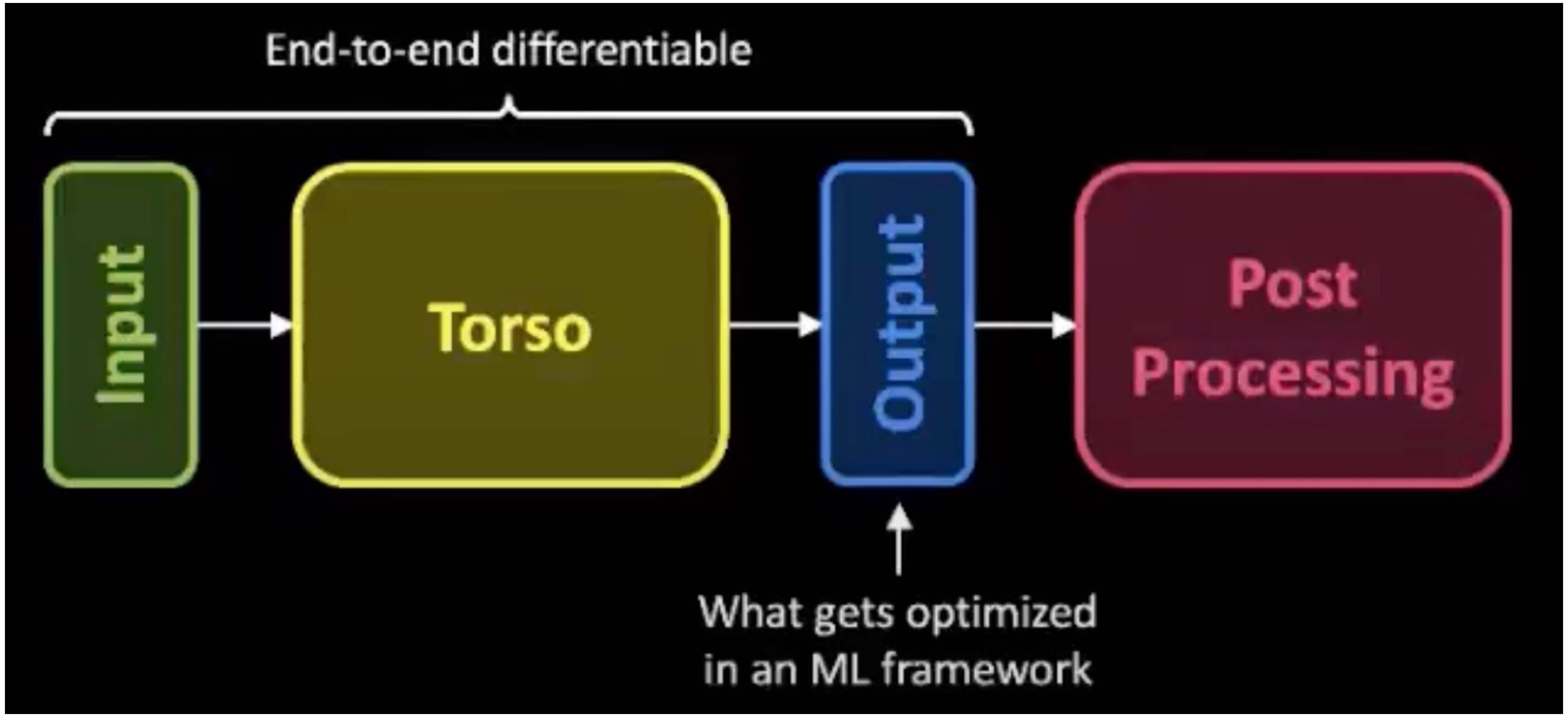
Very slow (100s to 10,000s of hours)

Prevents optimization of final structure

Mohammed AlQuraishi/
Nazim Bouatta

End-to-End Differentiability

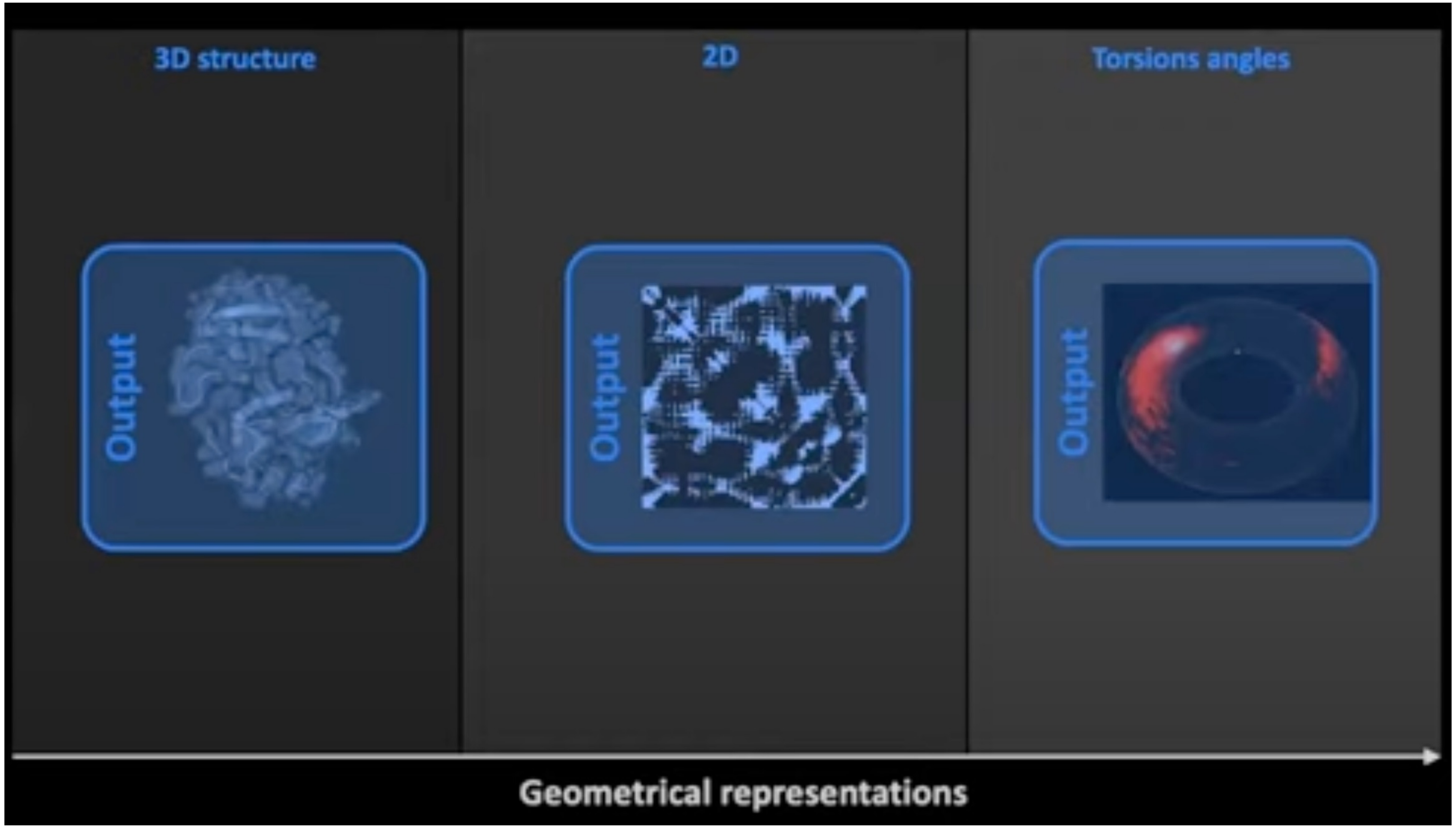
Optimising the output we want to optimise


End-to-end differentiable
Input
Torso
Output
Post Processing
What gets optimized in an ML framework

Mohammed AlQuraishi/
Nazim Bouatta

End-to-End Differentiability

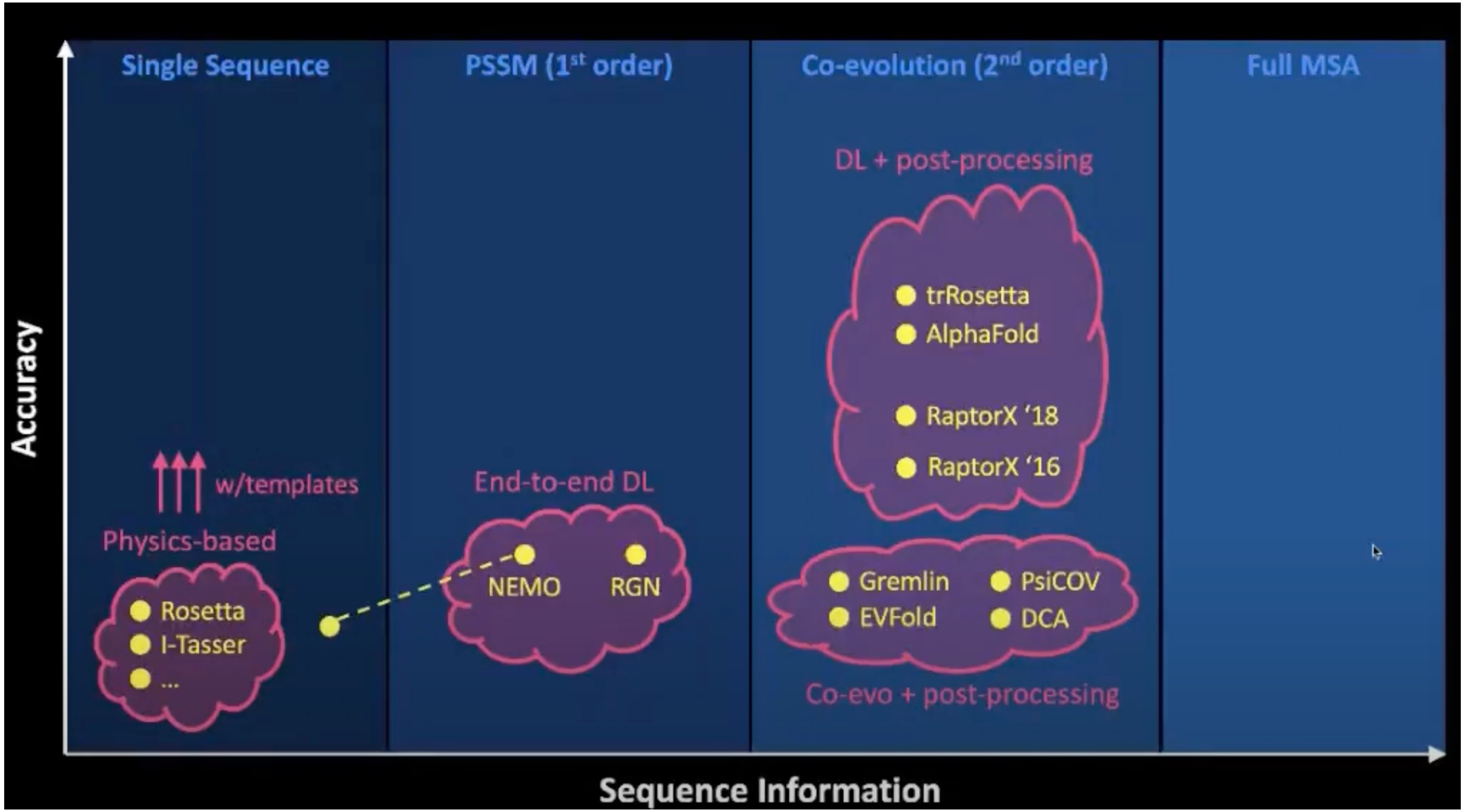
Different geometrical representations of output possible



Mohammed AlQuraishi/
Nazim Bouatta

End-to-End Differentiability

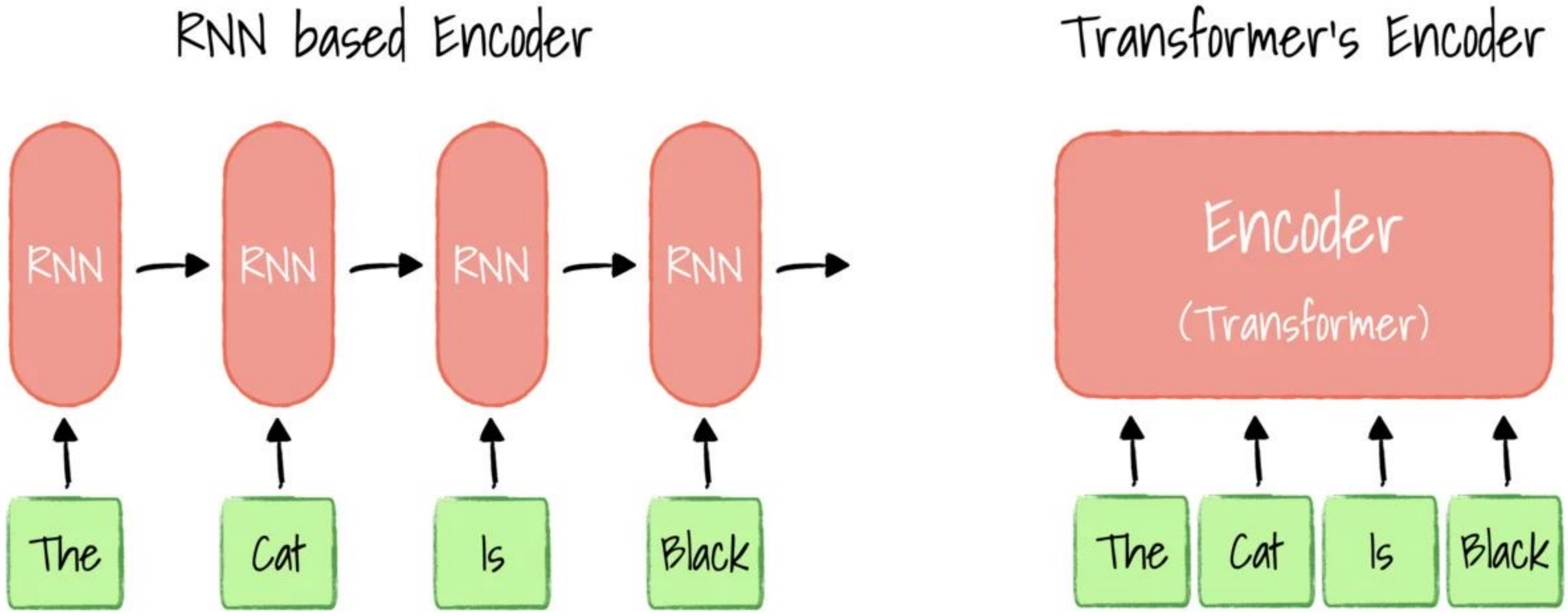
First of these models predicted torsion angles



Mohammed AlQuraishi/
Nazim Bouatta

Reminder: Sequence-to-Sequence

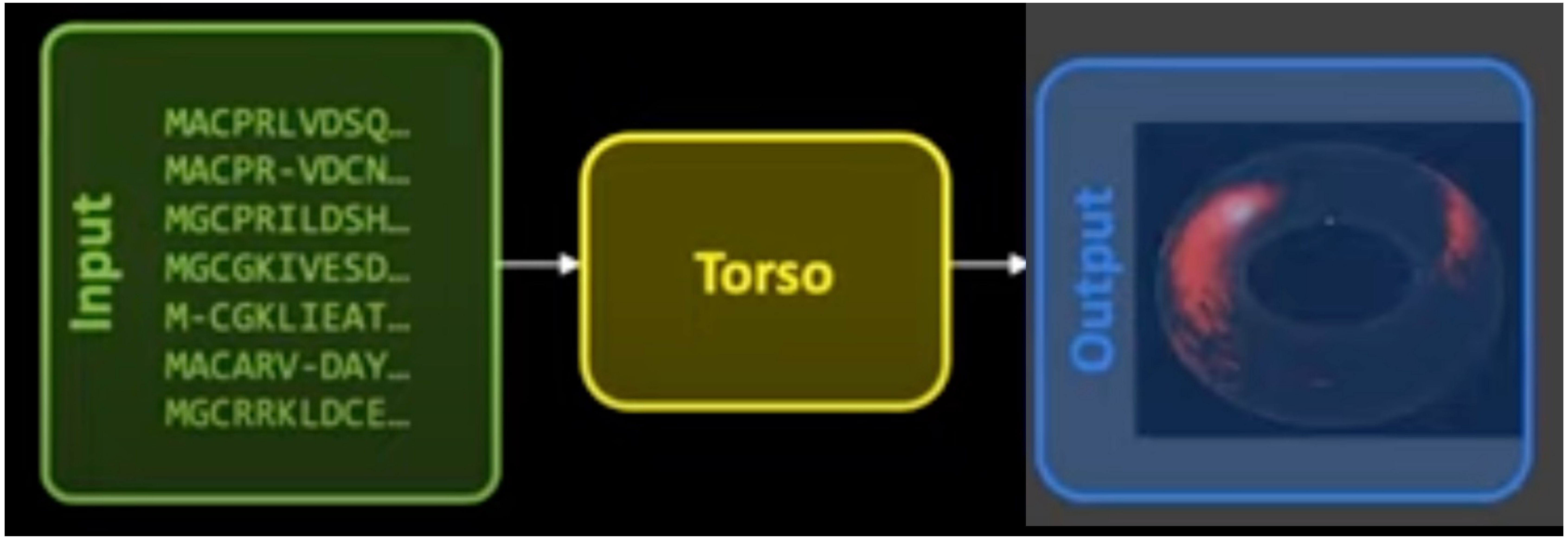
RNNs update a hidden state, transformers process in parallel



jinglescode.github.io

Sequence-to-Sequence

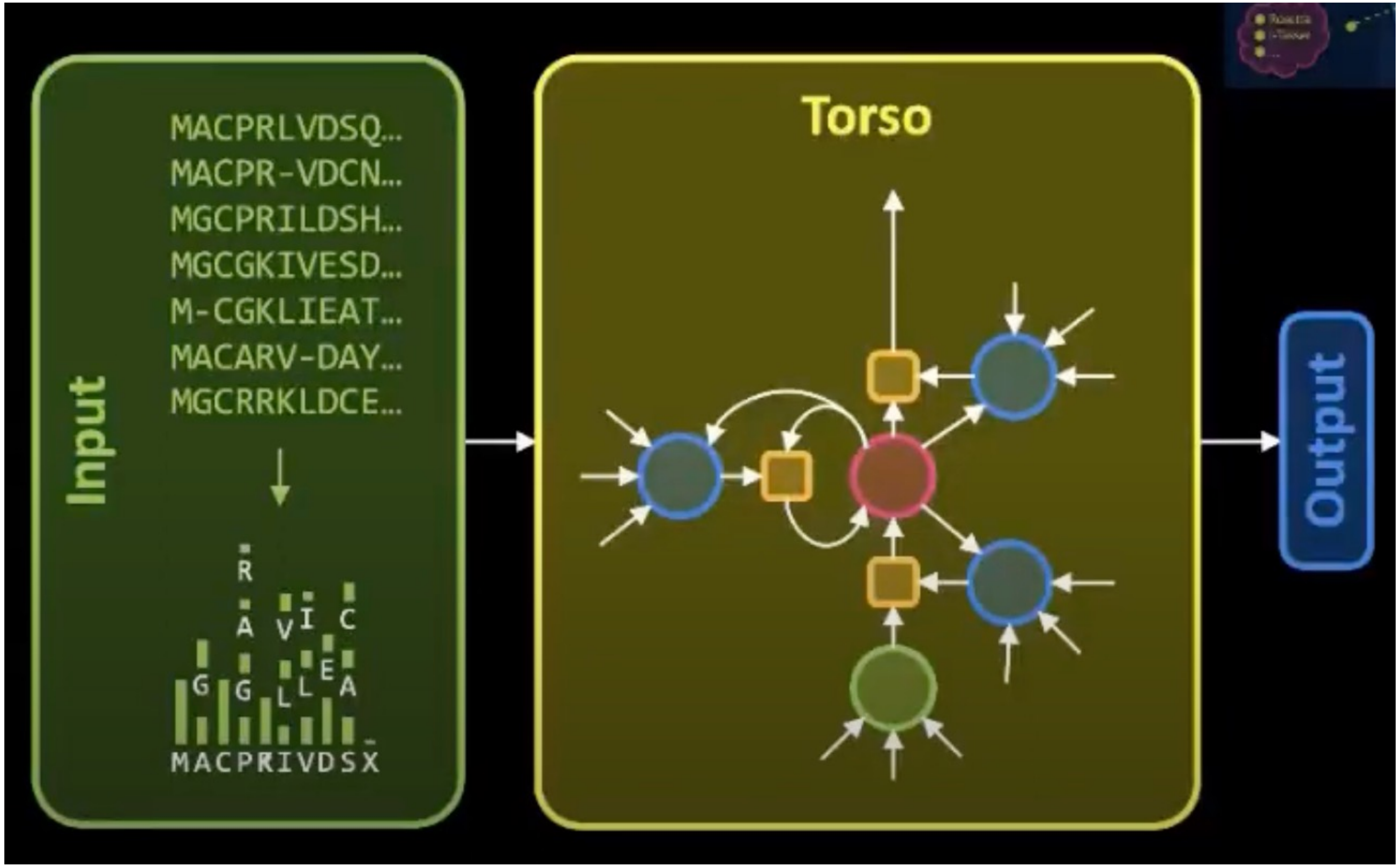
Use MSAs/PSSMs/… to predict a torsion angle sequence



Mohammed AlQuraishi/
Nazim Bouatta

RGN: End-to-end, but still an RNN

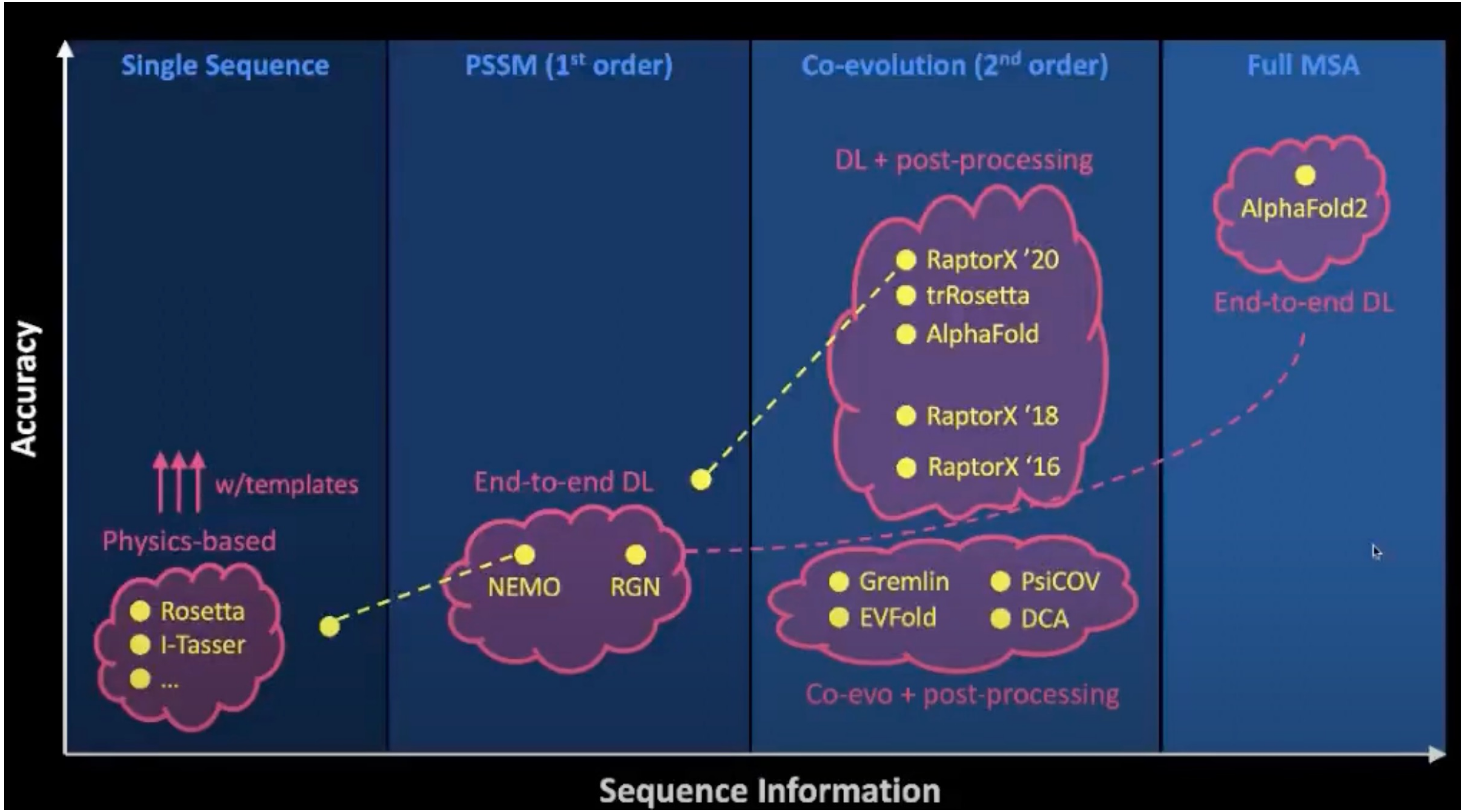
RNNs struggle with long-range interactions, important in proteins



Mohammed AlQuraishi/
Nazim Bouatta

AF2: End-to-end DL with full MSA

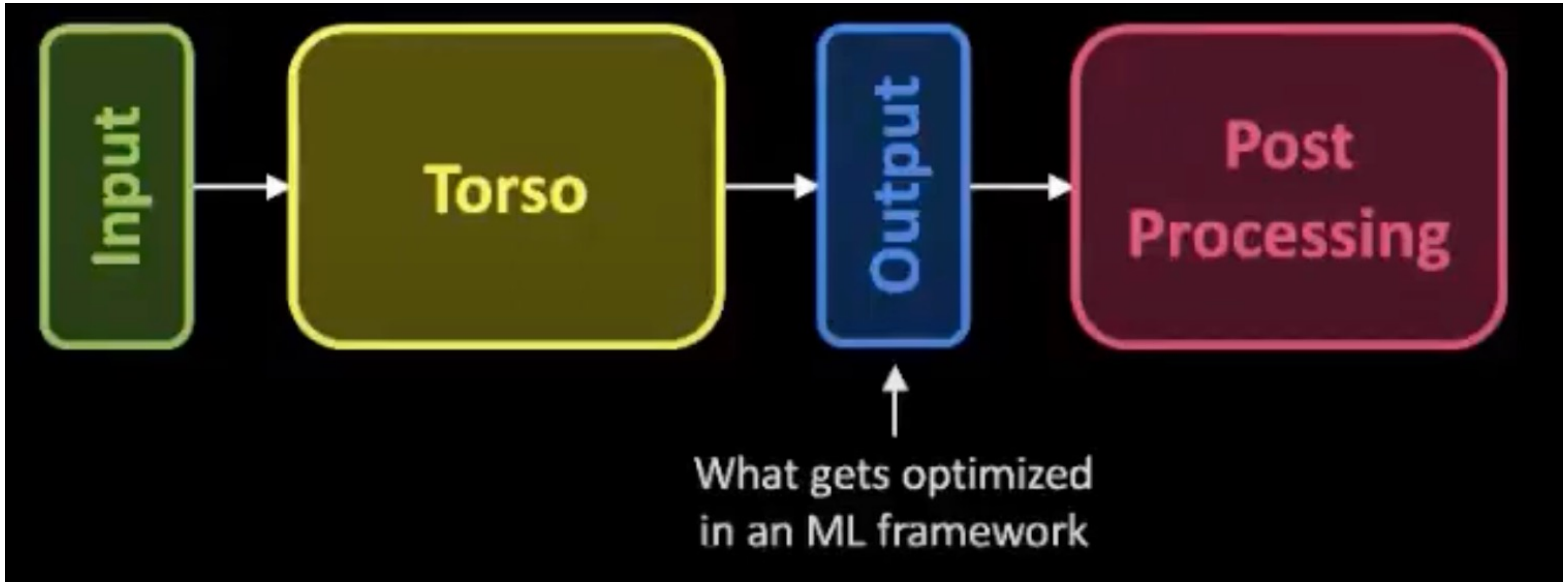
The DL Mantra: Use your model as feature extractor



Mohammed AlQuraishi/
Nazim Bouatta

End-to-End Differentiability

We want to optimise the output we are interested in: 3D Structures!



Mohammed AlQuraishi/
Nazim Bouatta

Sequence-to-Sequence

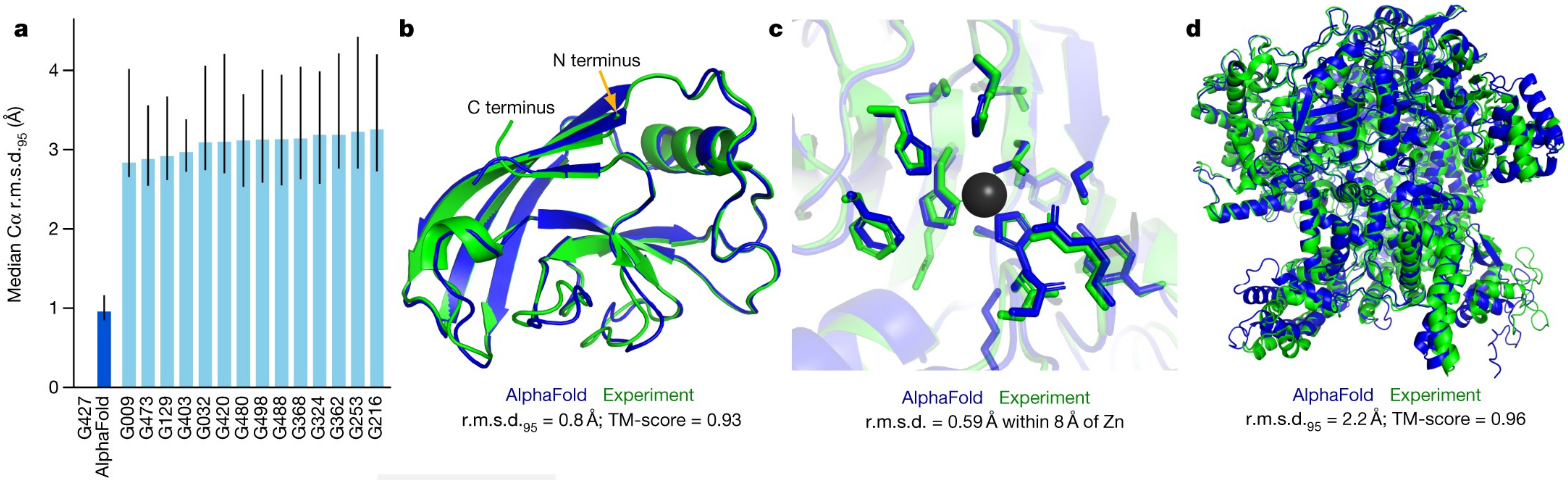
Use MSA to predict 3D structure directly

Input
MACPRLVDSQ...
MACPR-VDCN...
MGCPRILDSH...
MGCGKIVESD...
M-CGKLIEAT...
MACARV-DAY...
MGCRRKLDCE...
Torso
Output

Mohammed AlQuraishi/
Nazim Bouatta

AF2: solving the structure prediction problem?

New records in terms of prediction accuracy



Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*

3. AF2: The main ideas

The road to understanding AF2

Ranking based on difficulty, not quality (all these are great!)

OPIG Blog Post + YT Video 1&2

AF2 Paper +This lecture

Nazim Bouatta's lecture series

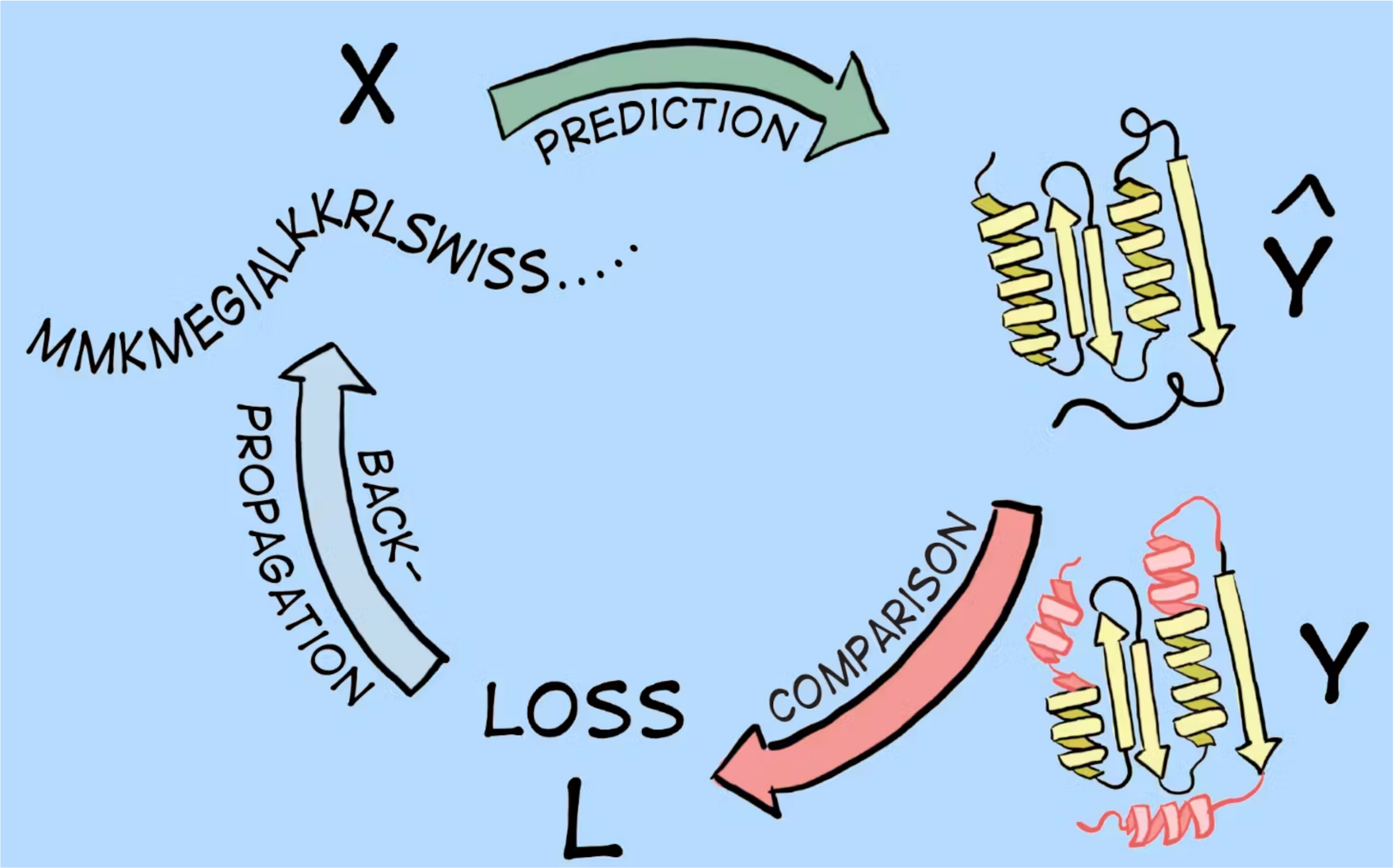
Castorina/Burkov post + OpenFold

AlQuraishi Blogpost and AF2 SI

My PTNG blog post

End-to-End Differentiability

Directly supervise on the output we care about



My PTNG blog post

End-to-End Differentiability

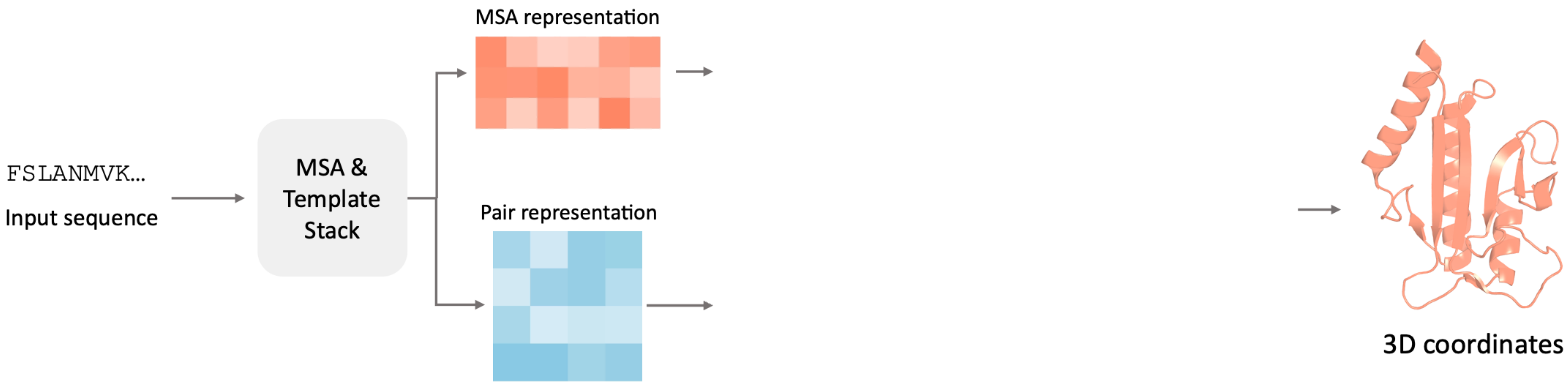
Directly supervise on the output we care about

FSLANMVK...
Input sequence
3D coordinates

Lukas Jarosch

Use both coevolution and geometric constraints

Both MSA and templates leveraged



FSLANMVK...

Input sequence

MSA & Template Stack

MSA representation

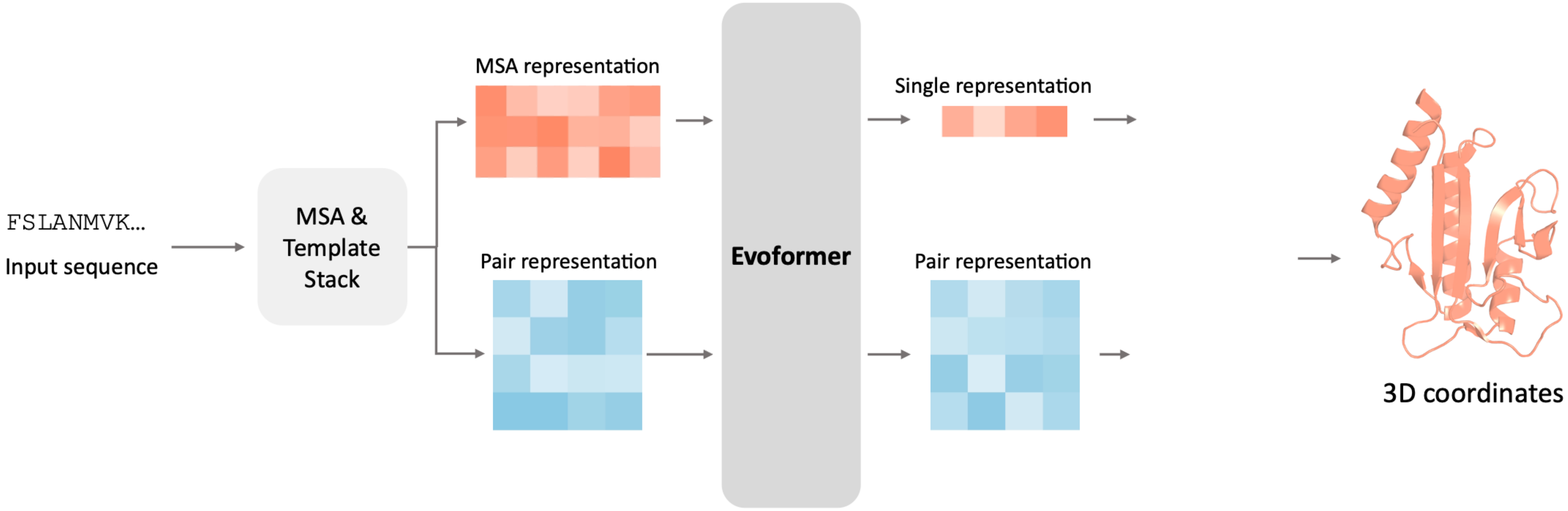
Pair representation

3D coordinates

Lukas Jarosch

Inductive Biases reflect protein biophysics

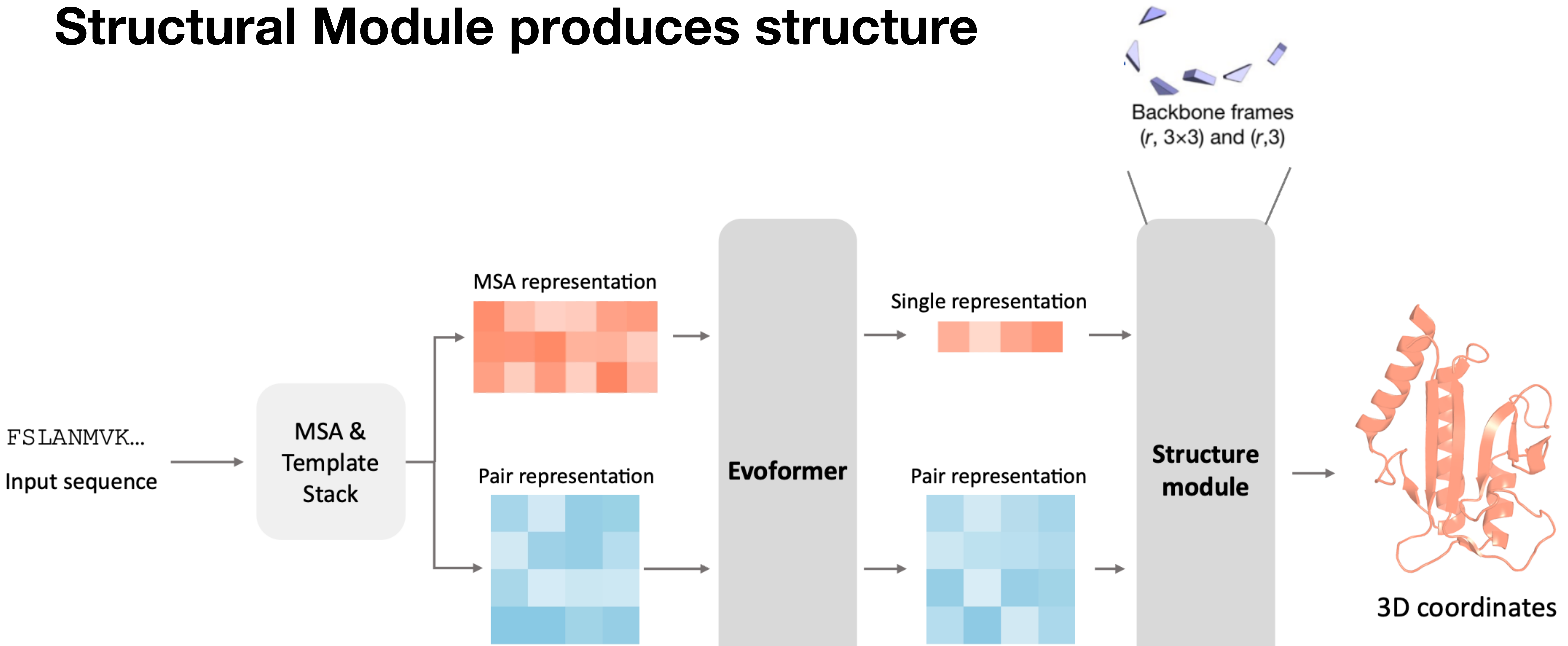
Communication encouraged between residues close in space



Lukas Jarosch

(Some) Physical constraints built-in

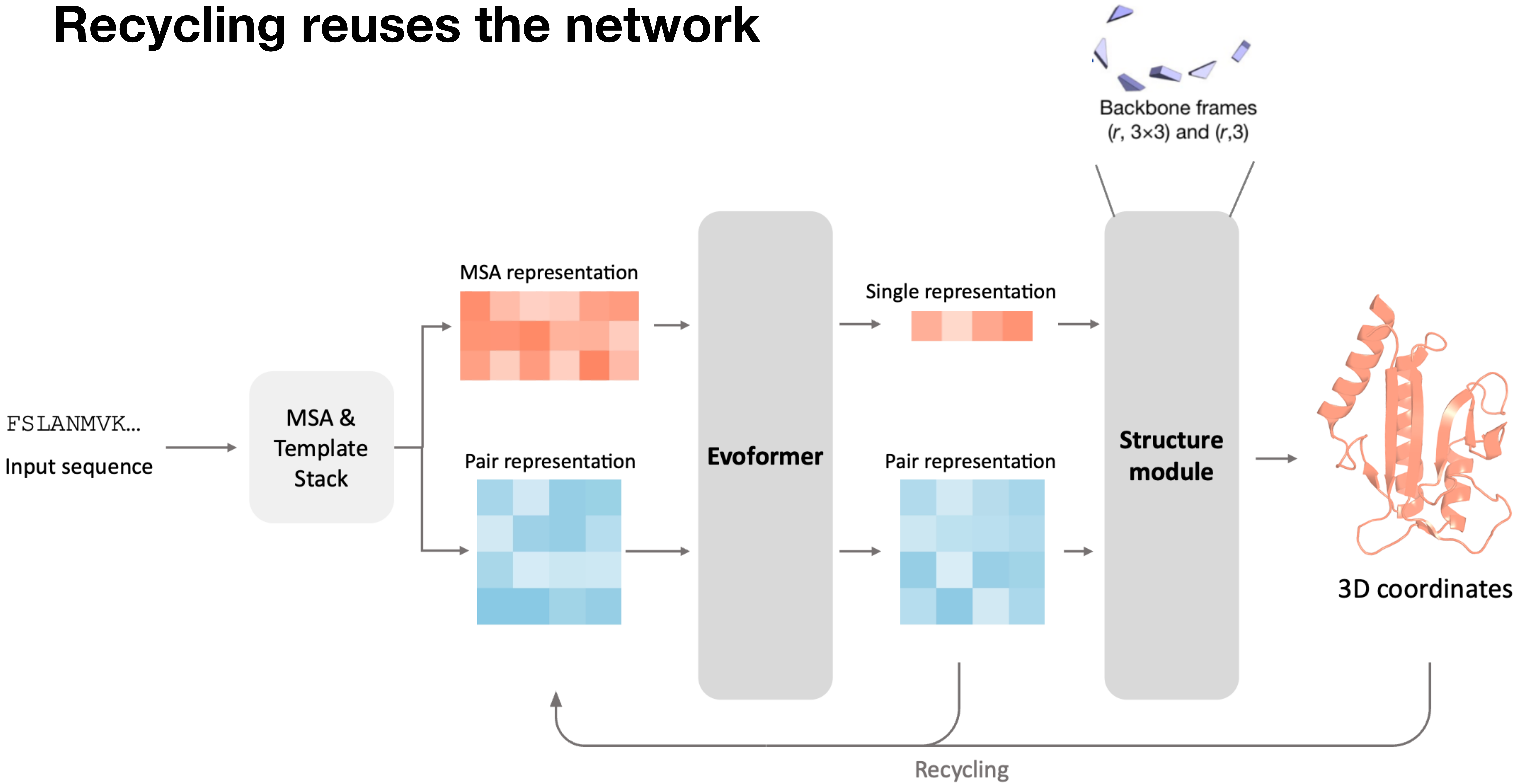
Structural Module produces structure



Lukas Jarosch

Iterative Refinement of Results

Recycling reuses the network



Lukas Jarosch

AF2 Architecture Overview

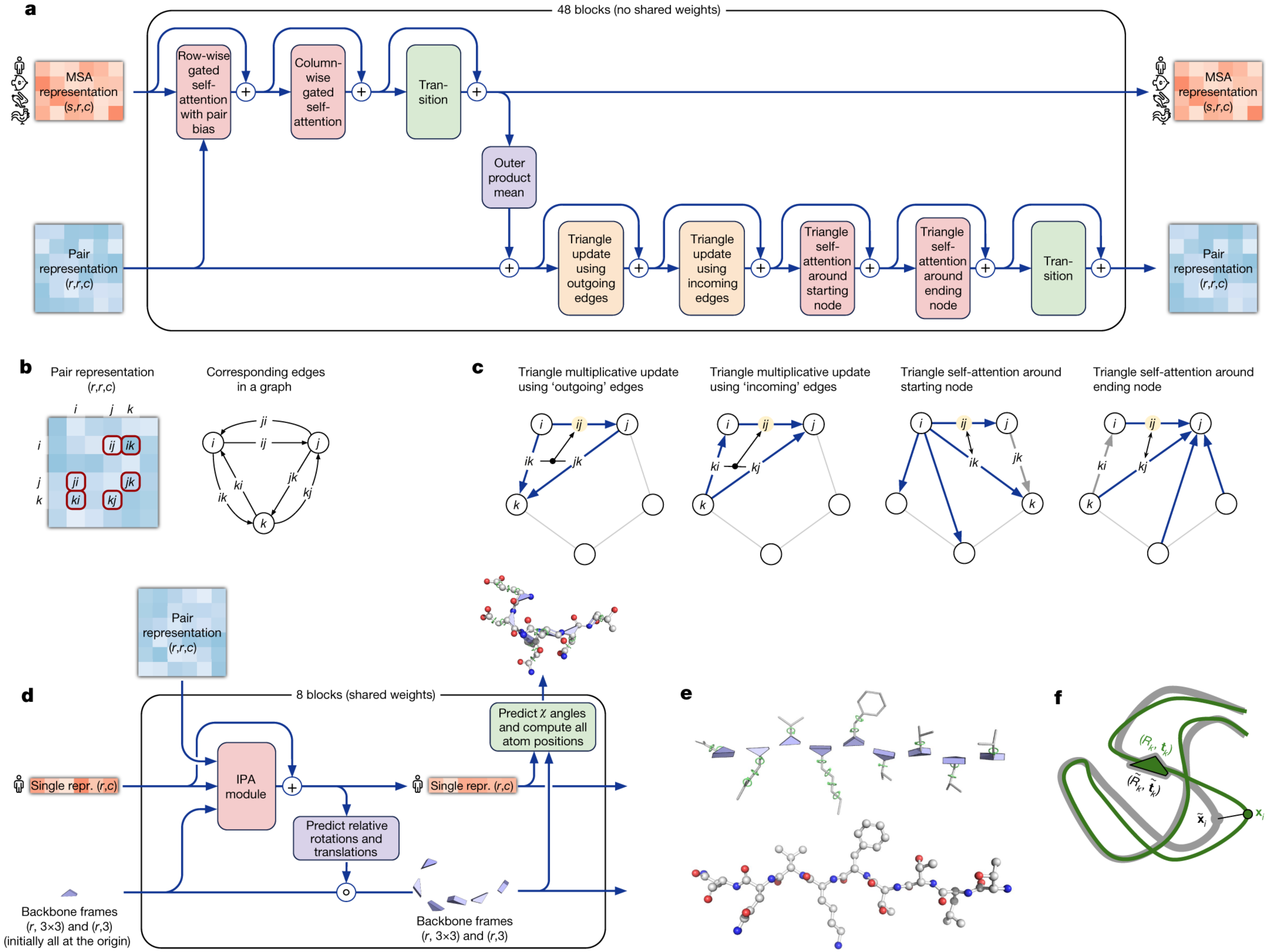
Reflects the main ideas discussed

Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*

The devil is in the detail…

A lot of superb engineering determined the final architecture



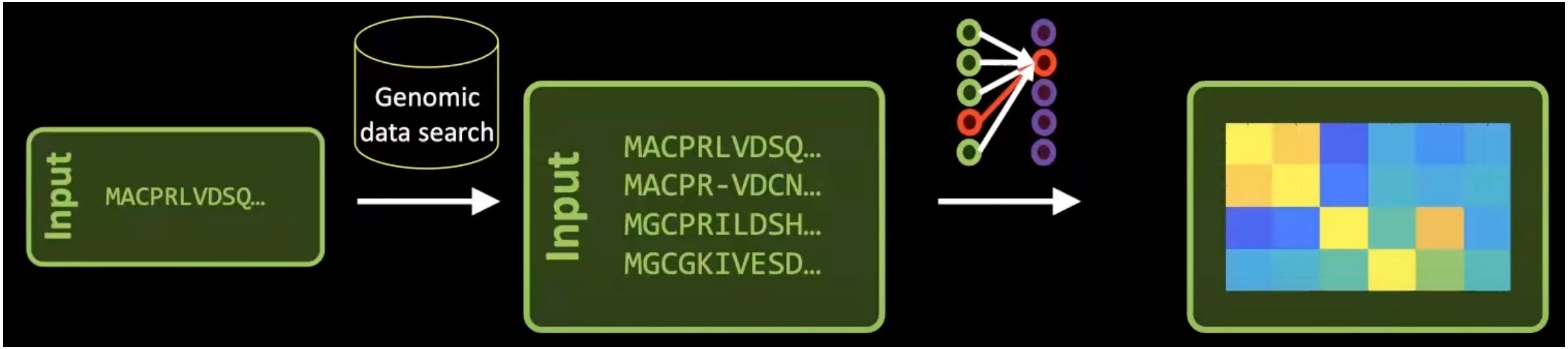
a
48 blocks (no shared weights)
MSA representation (s,r,c)
Row-wise gated self-attention with pair bias
Column-wise gated self-attention
Tran-sition
Outer product mean
Pair representation (r,r,c)
Triangle update using outgoing edges
Triangle update using incoming edges
Triangle self-attention around starting node
Triangle self-attention around ending node
Tran-sition
MSA representation (s,r,c)
Pair representation (r,r,c)
b
Pair representation (r,r,c)
i j k
i
j
k
ij ik
ji jk
ki kj
Corresponding edges in a graph
ji
ij
ki
jk
ik
kj
i
j
k
c
Triangle multiplicative update using 'outgoing' edges
Triangle multiplicative update using 'incoming' edges
Triangle self-attention around starting node
Triangle self-attention around ending node
d
Pair representation (r,r,c)
8 blocks (shared weights)
Single repr. (r,c)
IPA module
Predict relative rotations and translations
Single repr. (r,c)
Predict χ angles and compute all atom positions
Backbone frames (r, 3×3) and (r,3) (initially all at the origin)
Backbone frames (r, 3×3) and (r,3)
e
f

Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*

4. AF2: The Evoformer

The Evoformer

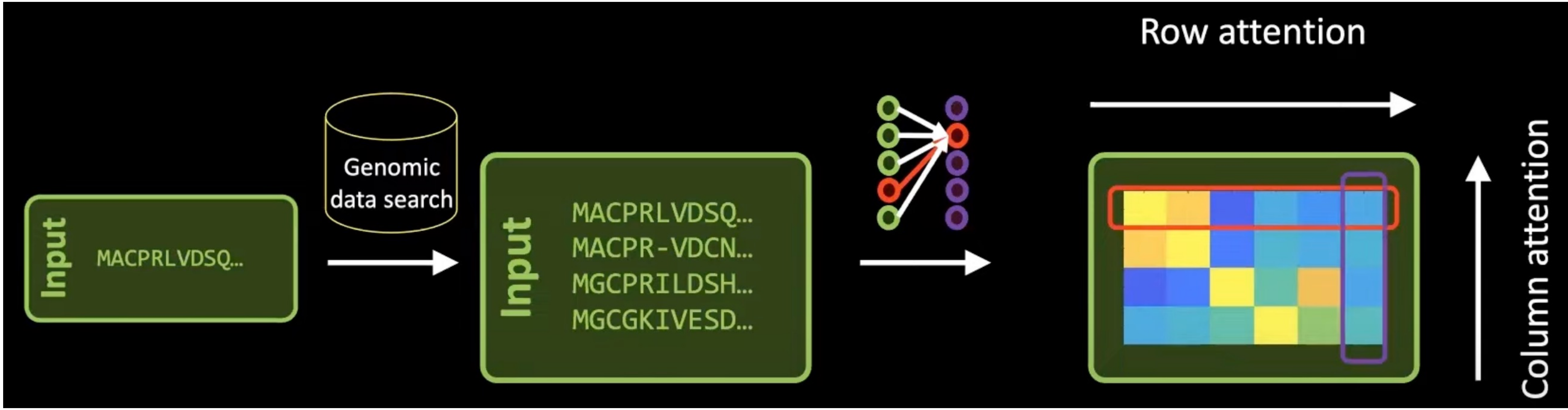
Building an MSA and processing it via a transformer



Mohammed AlQuraishi/
Nazim Bouatta

Communication in the MSA Stack

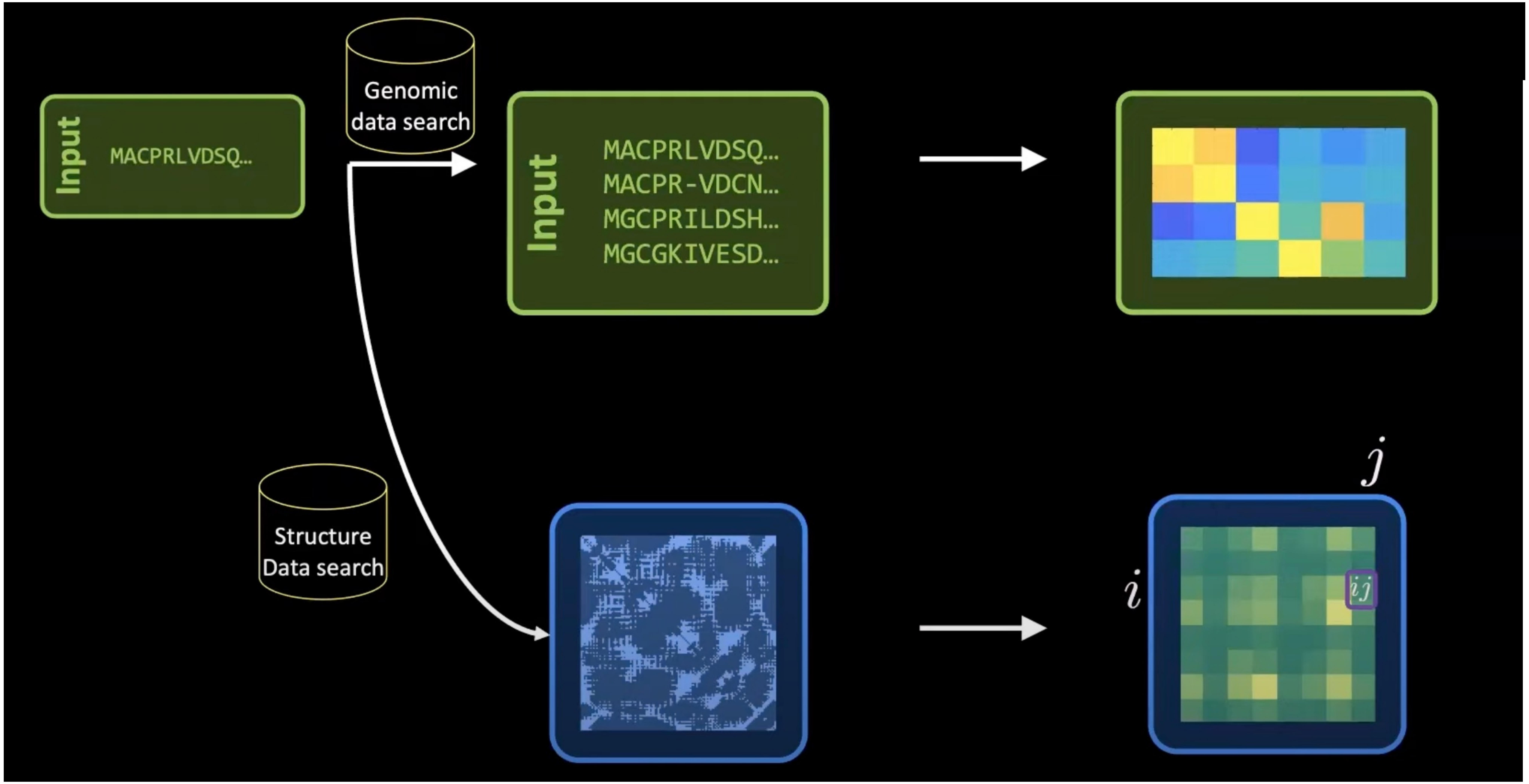
Row attention in a sequence; column attention between sequences



Mohammed AlQuraishi/
Nazim Bouatta

2-track architecture: Pair Representation

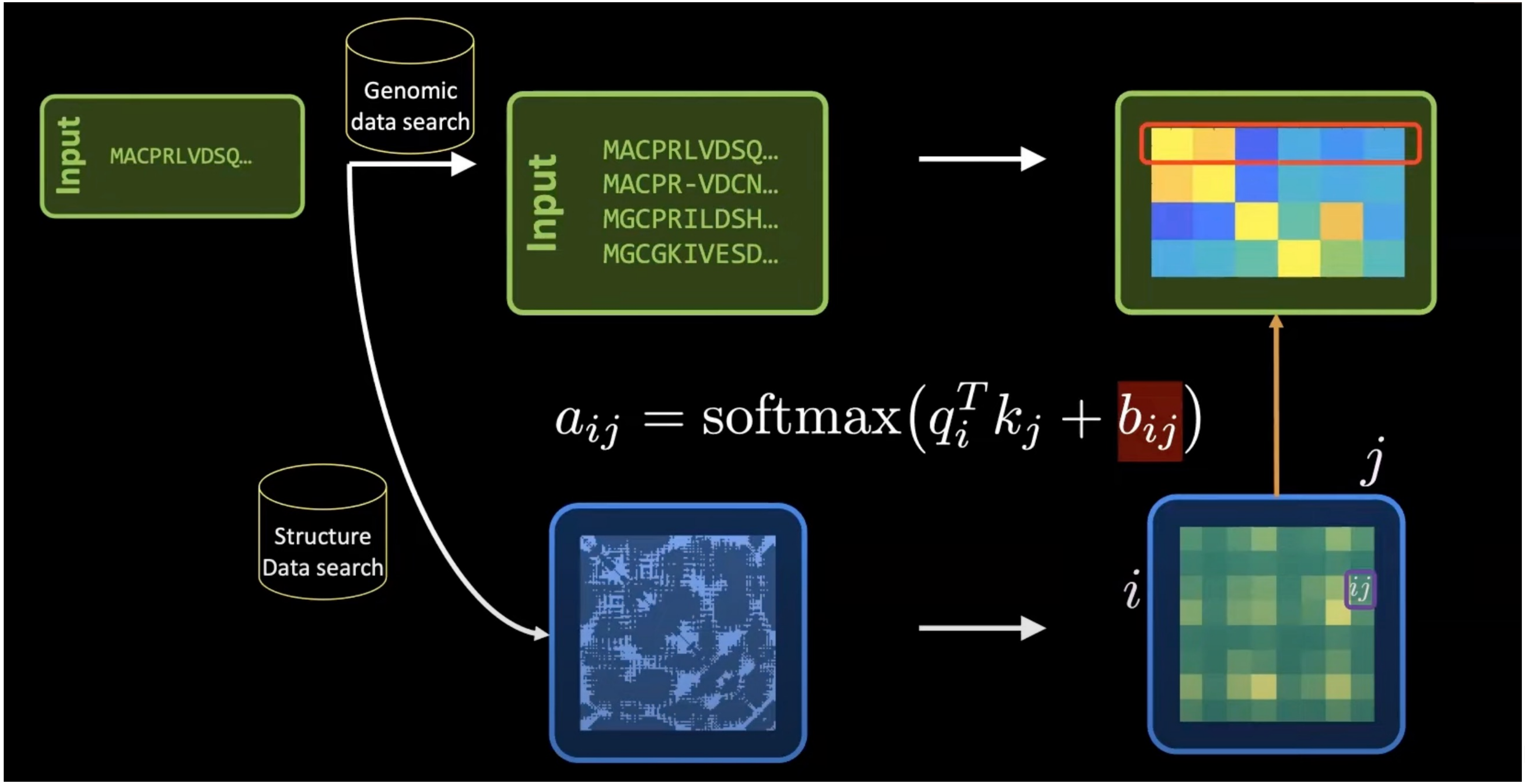
Reason not only over evolution but also over geometry



Mohammed AlQuraishi/
Nazim Bouatta

Row-wise attention with pair bias

Tell the MSA representation which residues to pay attention to



$$a_{ij} = \mathrm{softmax}(q_i^T k_j + b_{ij})$$

Mohammed AlQuraishi/
Nazim Bouatta

Pair representation updates MSA stack

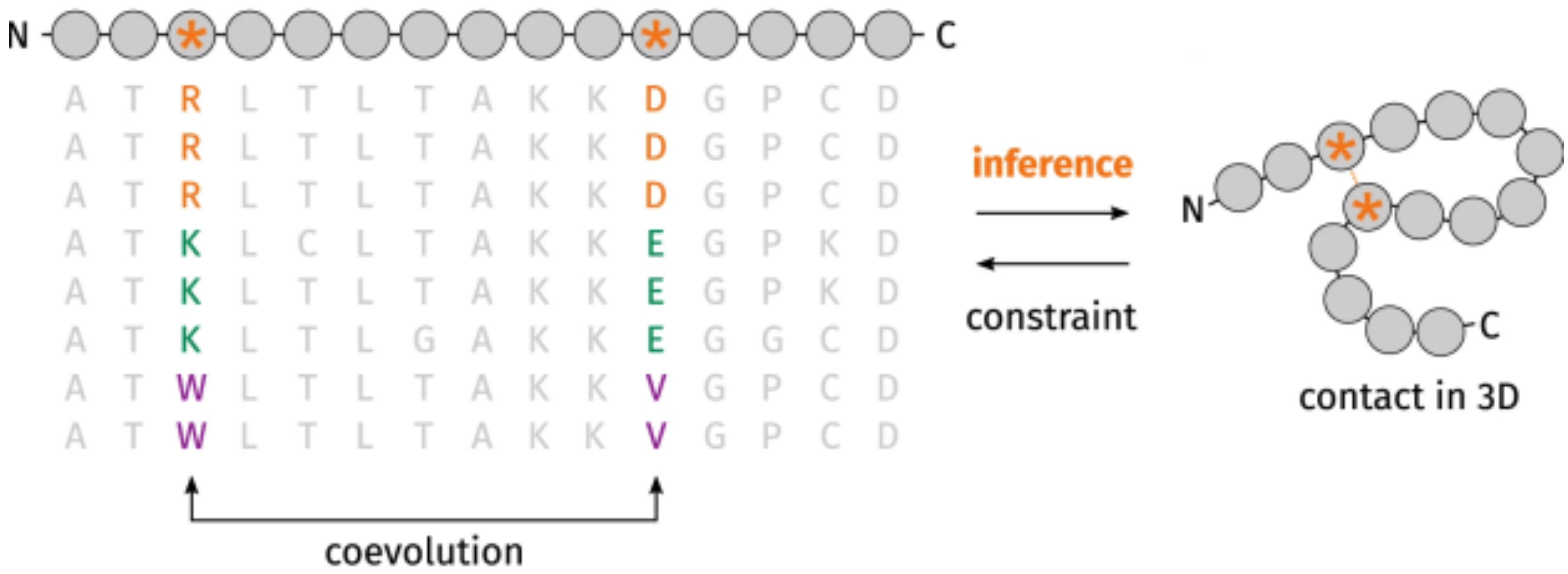
Geometrical constraints inform coevolutionary search

Input
MACPRLVDSQ...
Genomic data search
Input
MACPRLVDSQ...
MACPR-VDCN...
MGCPRILDSH...
MGCGKIVESD...
Structure Data search

Mohammed AlQuraishi/
Nazim Bouatta

Coevolution: The Idea

Residues that correlate are probably close in space



```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
```

Carlos Outeiral Rubiera

MSA stack updates pair representation

Coevolution infers geometrical constraints (outer product mean)

Input
MACPRLVDSQ...
Genomic data search
Input
MACPRLVDSQ...
MACPR-VDCN...
MGCPRILDSH...
MGCGKIVESD...
Structure Data search

Mohammed AlQuraishi/
Nazim Bouatta

Our old nemesis: Self-inconsistency

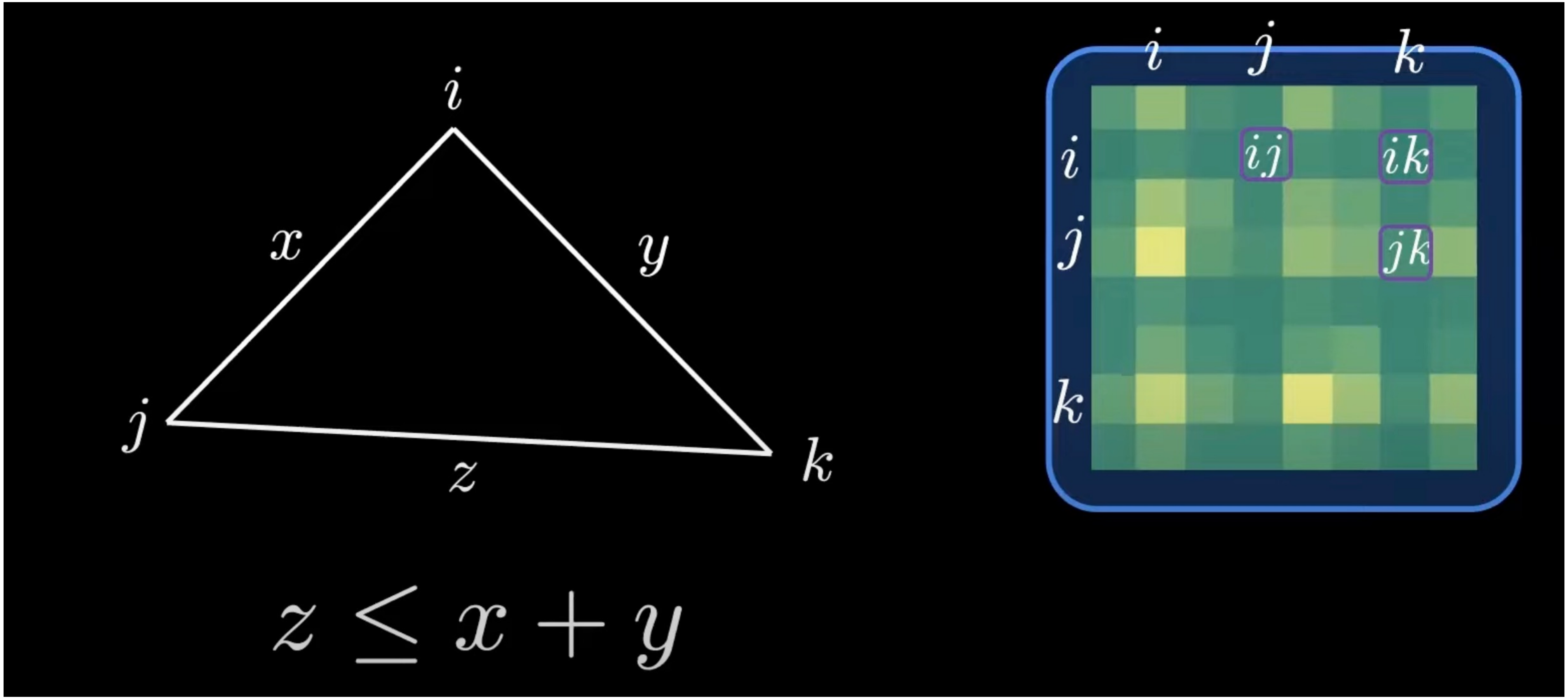
As in AF1, “image” representations can contradict themselves

Input
MACPRLVDSQ...
Genomic data search
Input
MACPRLVDSQ...
MACPR-VDCN...
MGCPRILDSH...
MGCGKIVESD...
Structure Data search
Inconsistent

Mohammed AlQuraishi/
Nazim Bouatta

The Triangular Inequality

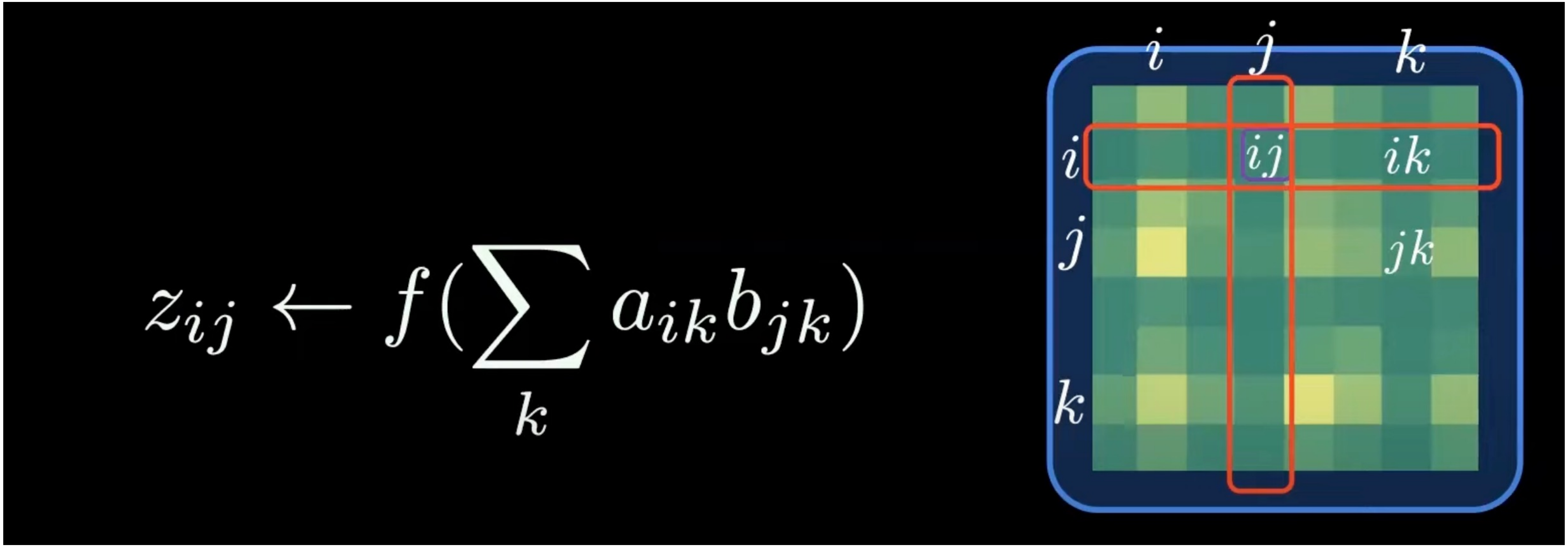
How to enforce this geometric constraint?



Mohammed AlQuraishi/
Nazim Bouatta

Triangular Updates

Update pair representation in consistent manner



$$z_{ij} \leftarrow f(\sum_k a_{ik} b_{jk})$$

Mohammed AlQuraishi/
Nazim Bouatta

Communication is key

How to go now from Evoformer output to structure?

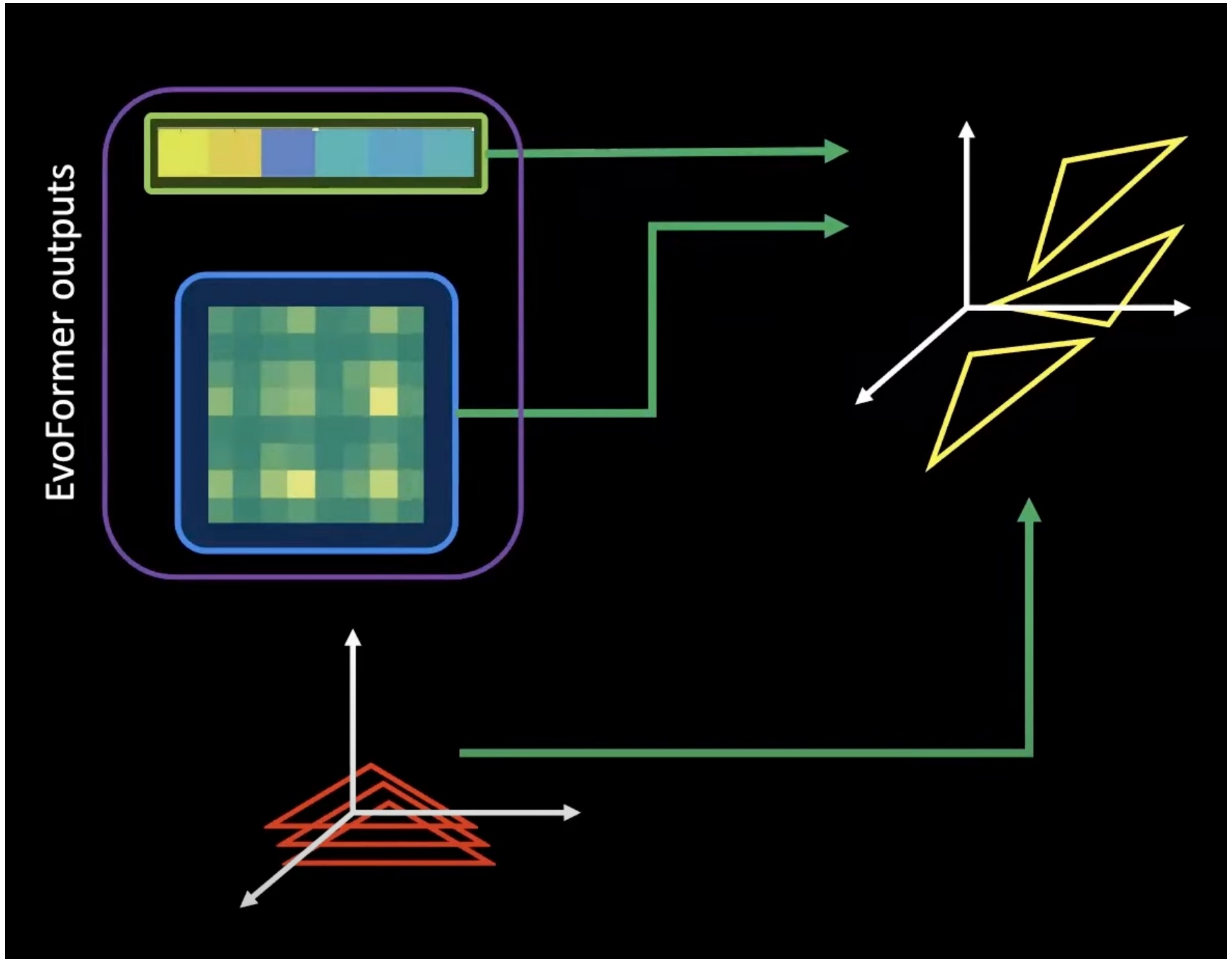
Input
MACPRLVDSQ...
Genomic data search
Input
MACPRLVDSQ...
MACPR-VDCN...
MGCPRILDSH...
MGCGKIVESD...
Structure Data search
?

Mohammed AlQuraishi/
Nazim Bouatta

5. AF2: The Structure Module

How to get from Evoformer to structure?

Clever part: No post-processing, everything end-to-end



Mohammed AlQuraishi/
Nazim Bouatta

Protein as a triangle gas

Break up the chain to allow structural exploration

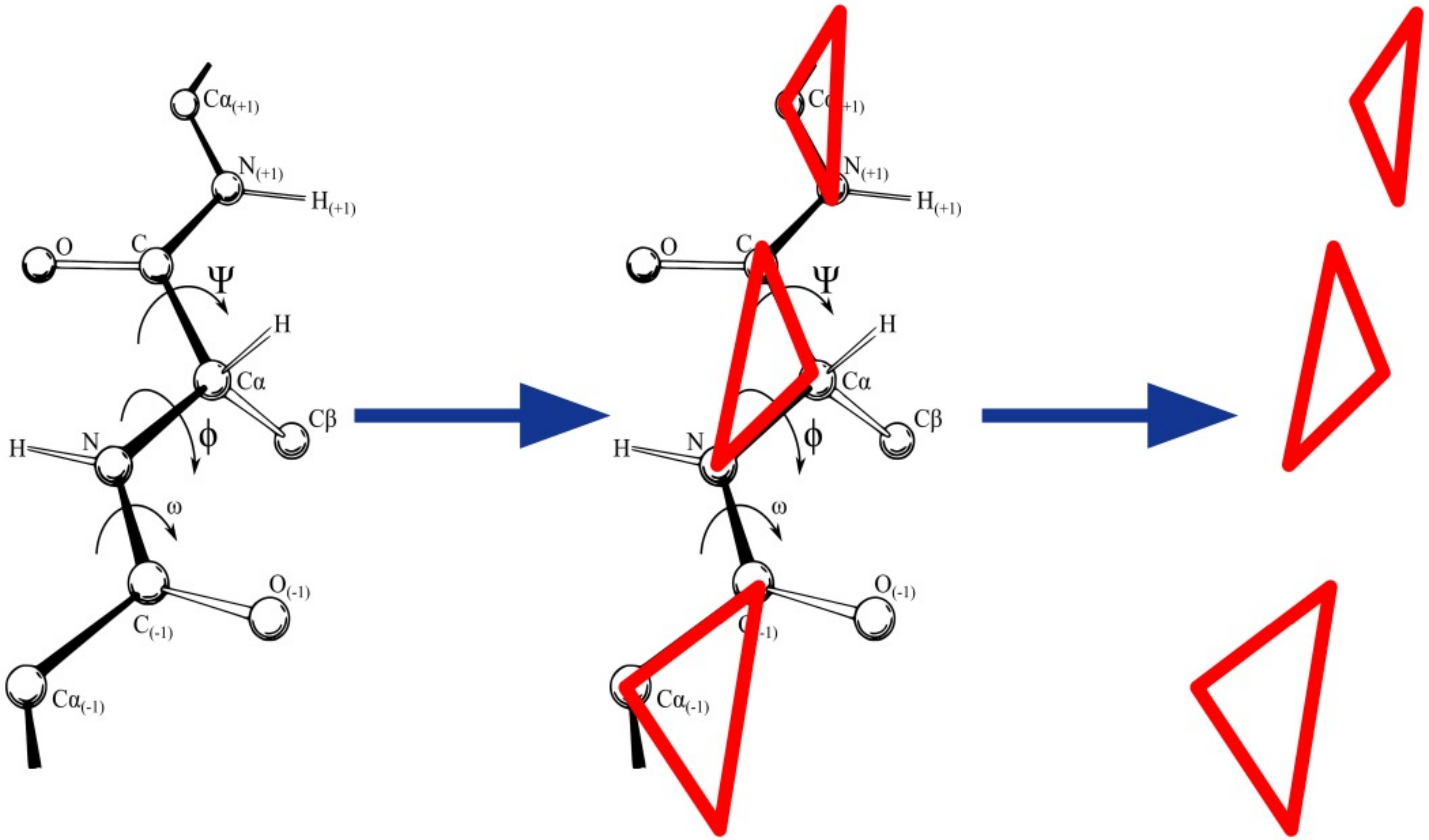
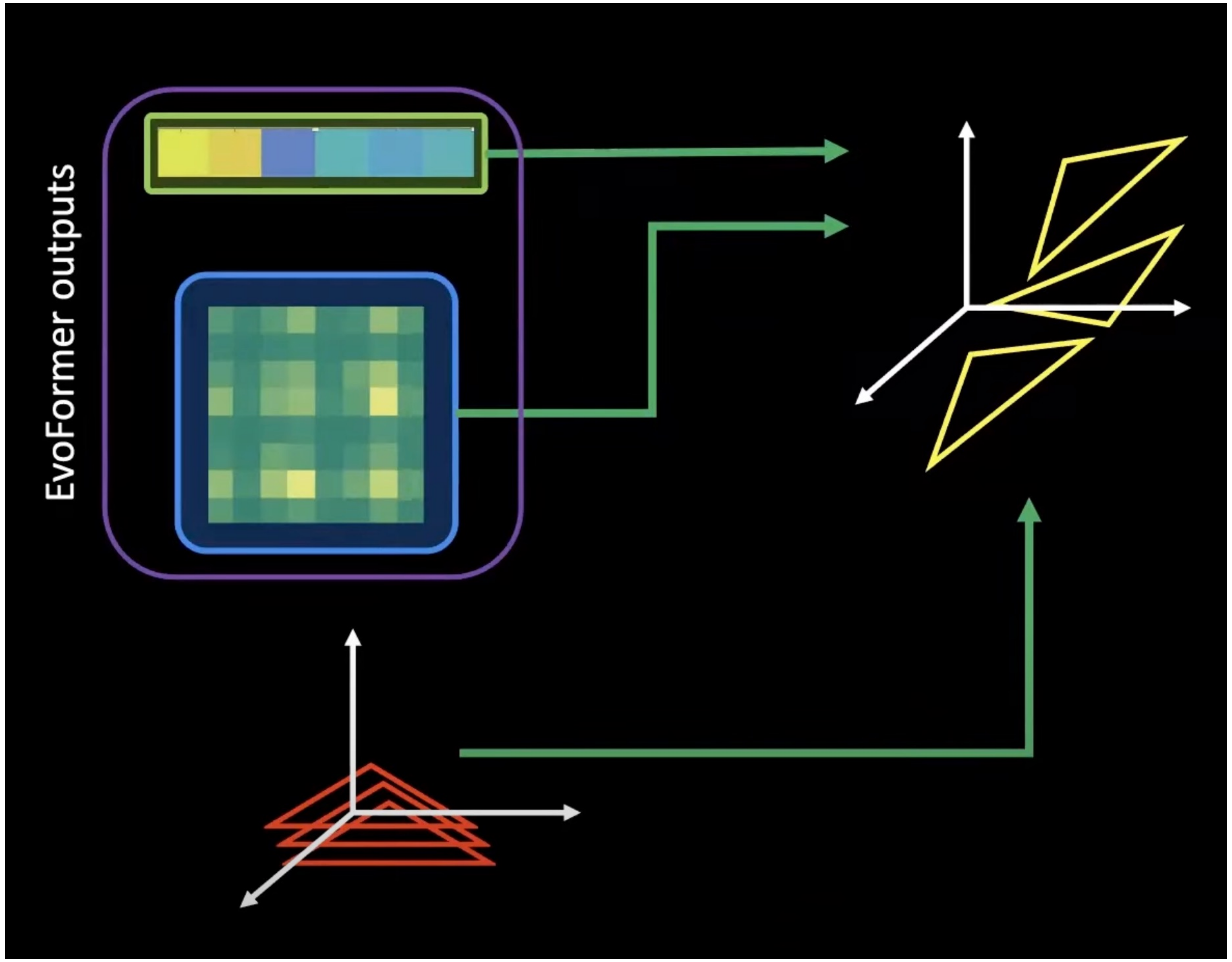


Image: Dcrjsr, vectorised Adam Rędzikowski (CC BY 3.0, Wikipedia)

AF2 Presentation, John Jumper

Black Hole Initialisation

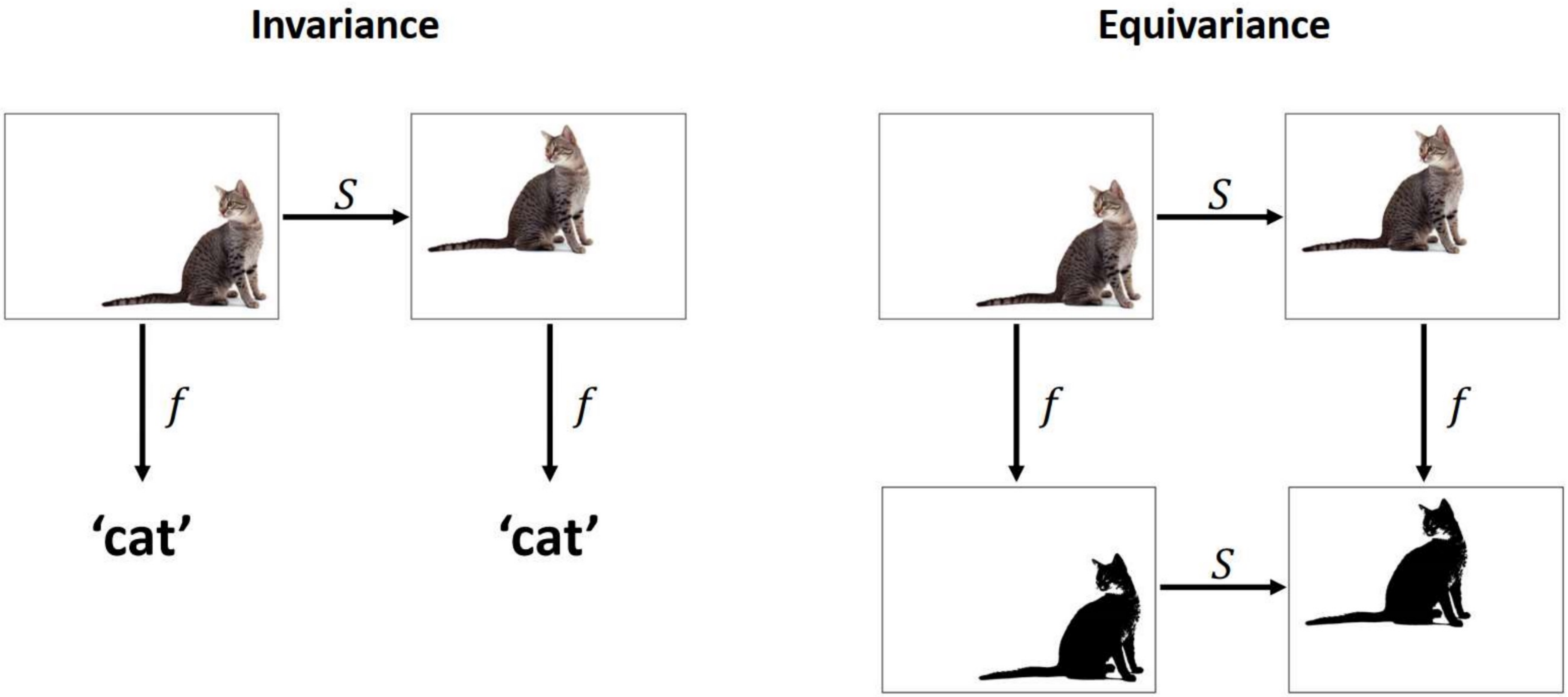
Place all triangles at the origin intially



Mohammed AlQuraishi/
Nazim Bouatta

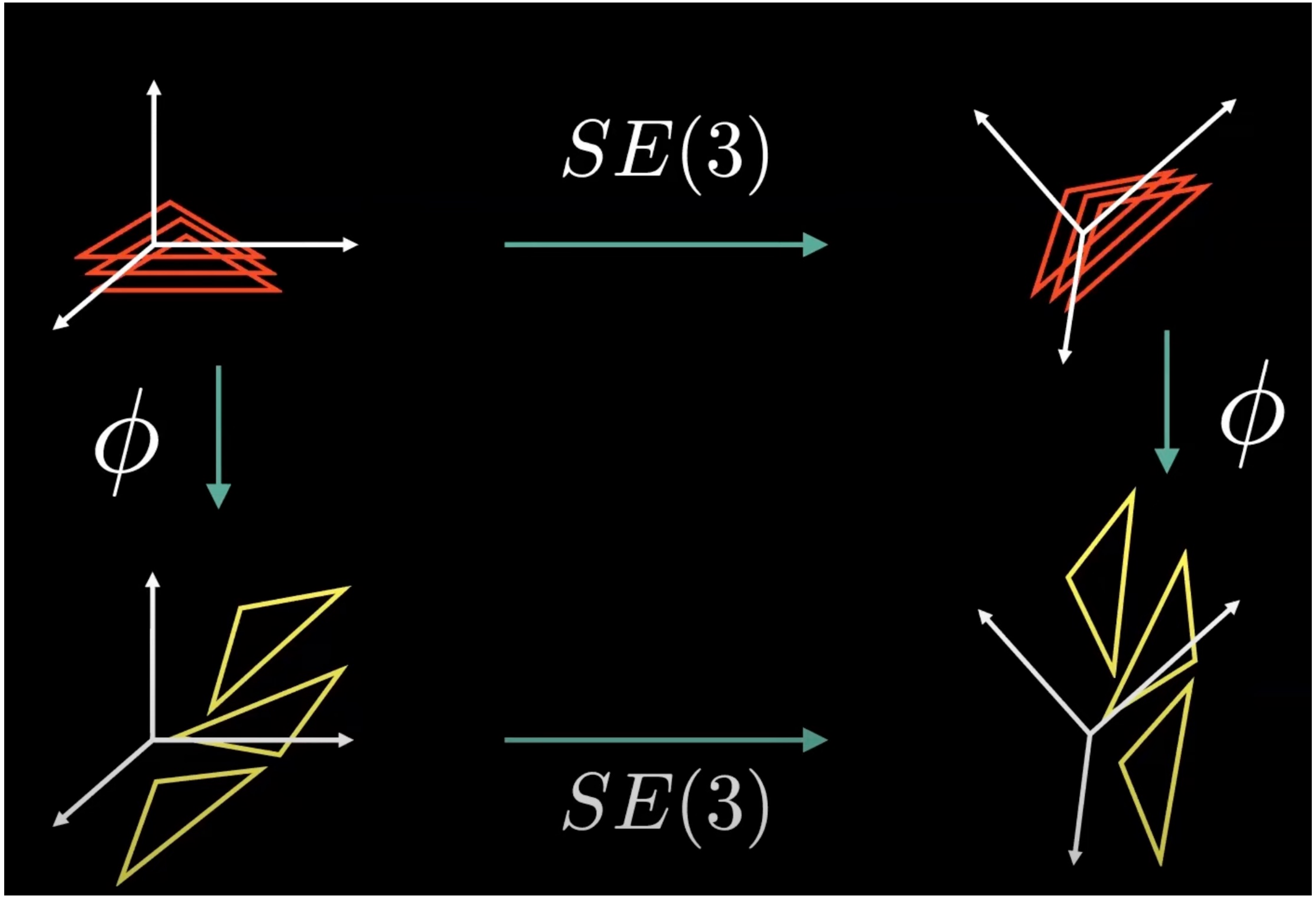
Reminder: Equivariance

Leverage the symmetry of your data



Reminder: Equivariance

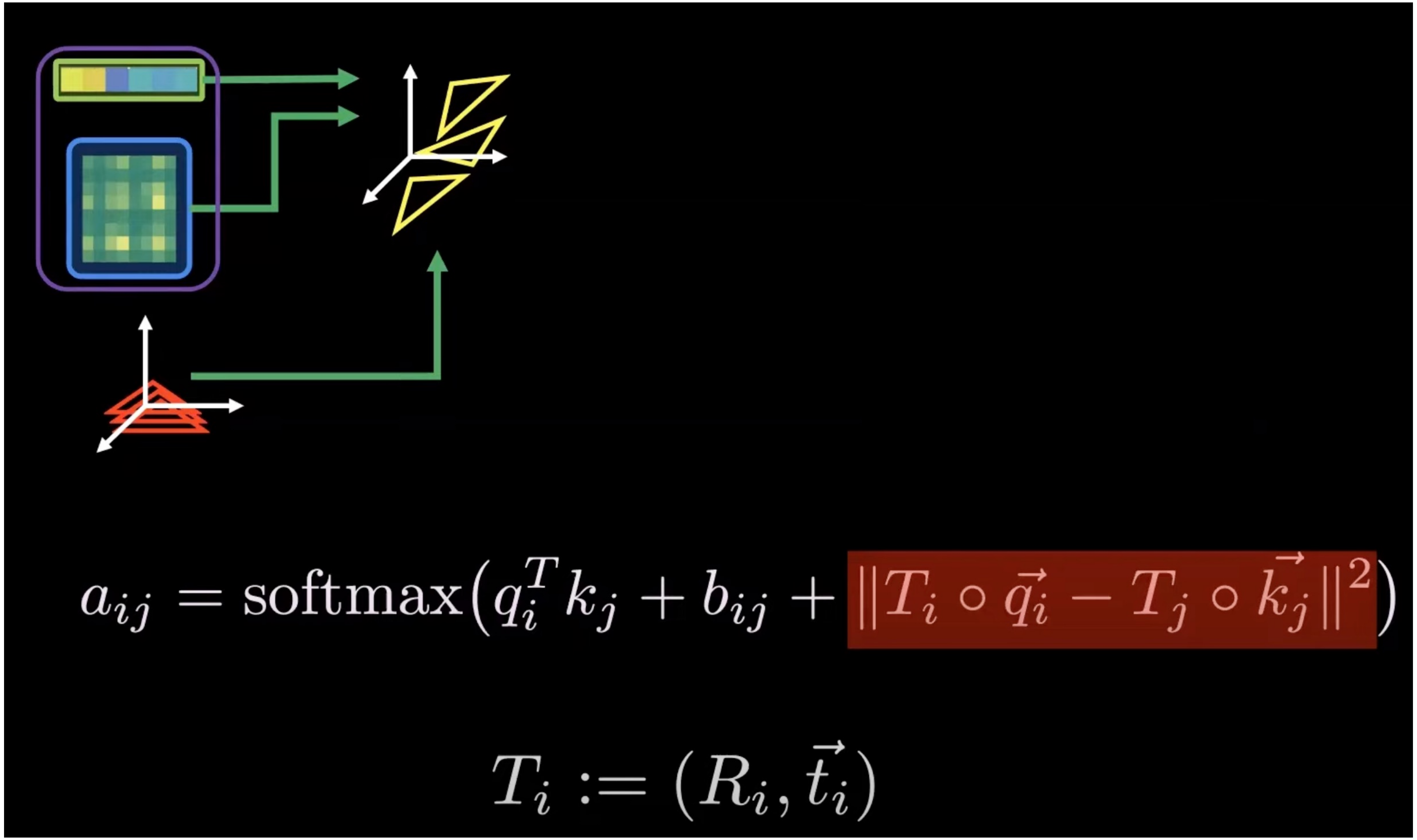
Leverage the symmetry of your data



Mohammed AlQuraishi/
Nazim Bouatta

Geometric keys and queries

Backbone Update via IPA (Invariant Point Attention)



$$a_{ij} = \text{softmax}\left(q_i^T k_j + b_{ij} + \|T_i \circ \vec{q_i} - T_j \circ \vec{k_j}\|^2\right)$$

$$T_i := (R_i, \vec{t_i})$$

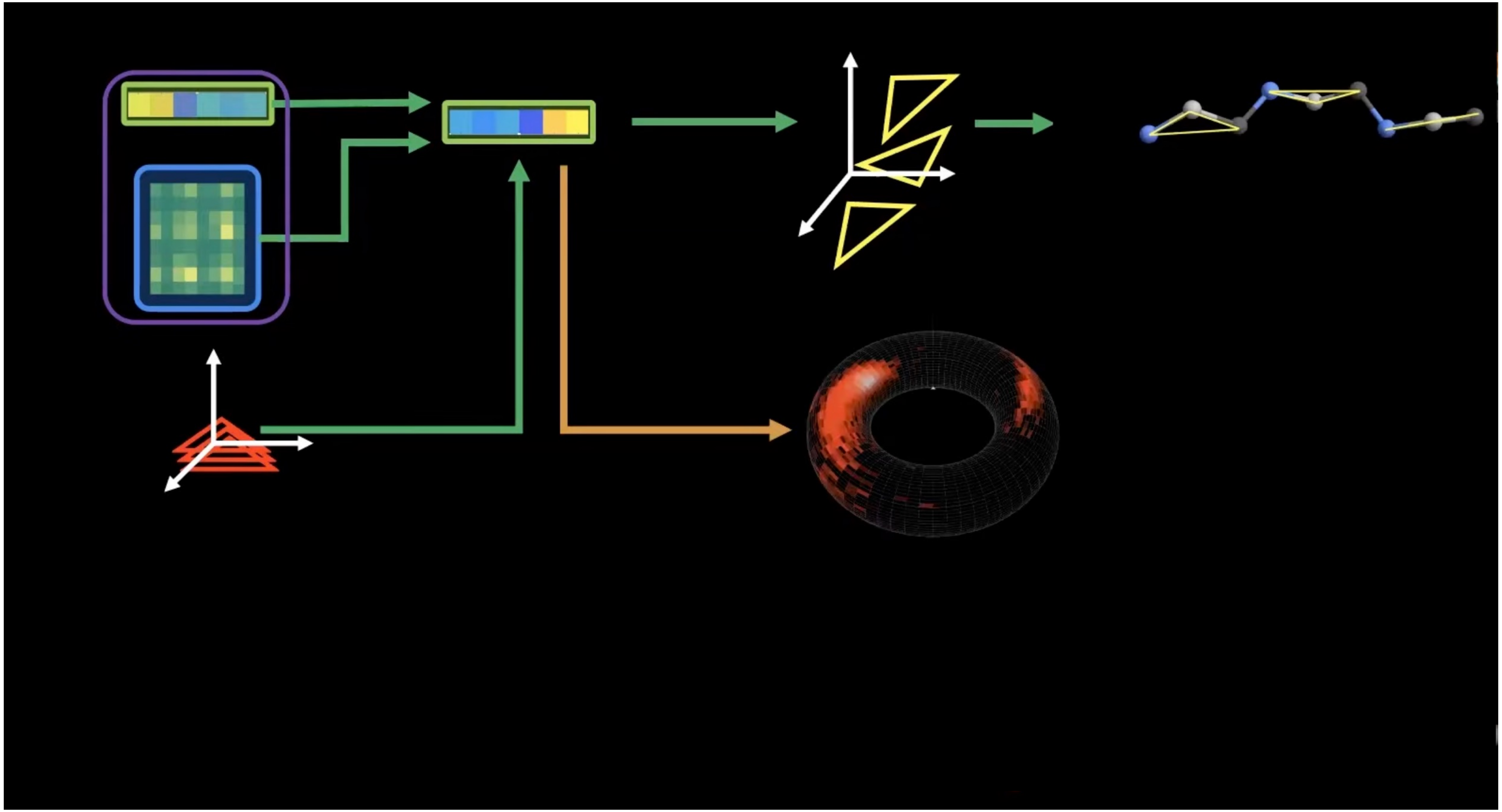
Mohammed AlQuraishi/
Nazim Bouatta

Spraying key and query vectors

IPA: Invariant Point Attention

Predicting the final structure

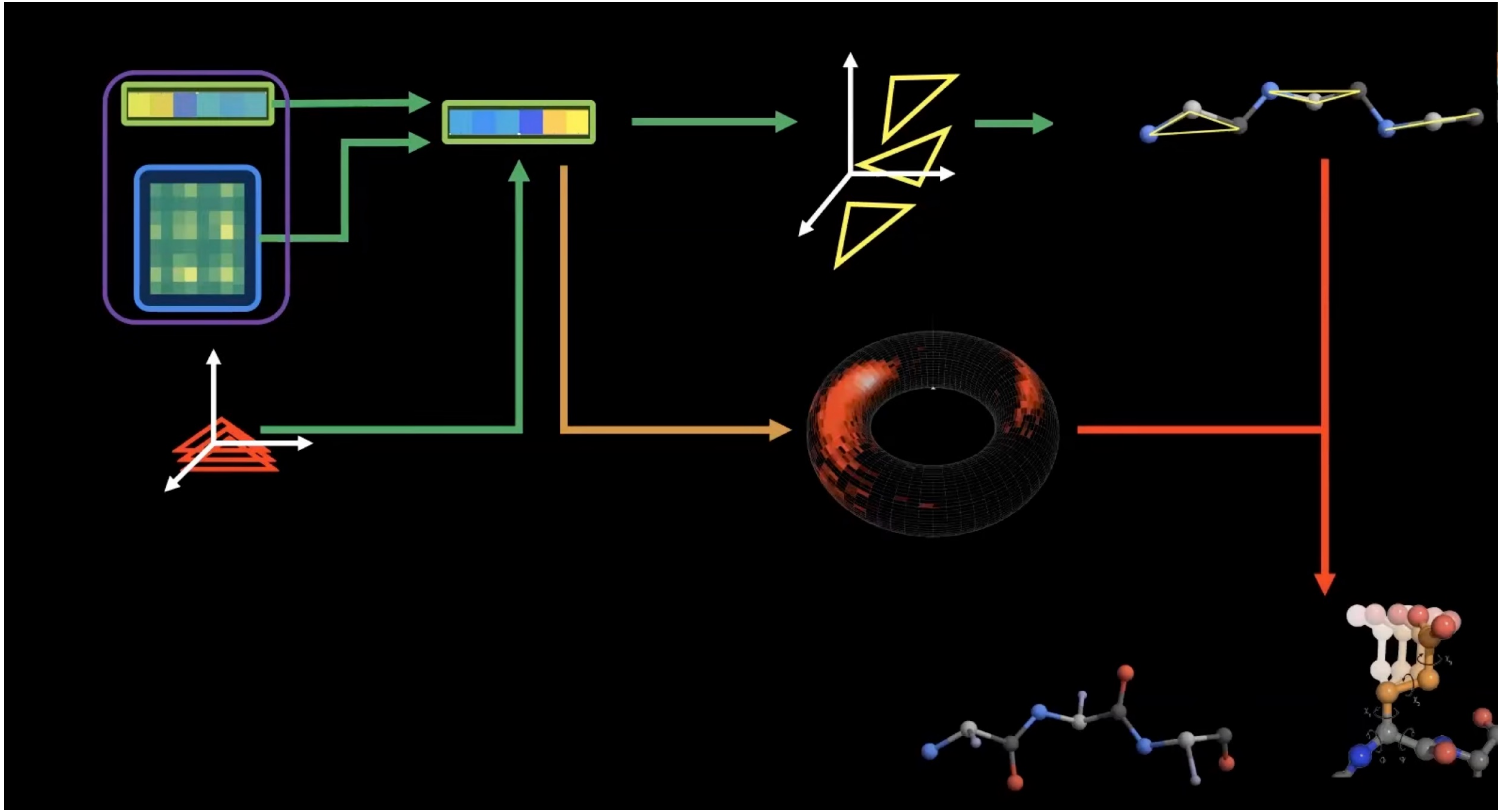
Predict triangle positions+orientations+torsion angles



Mohammed AlQuraishi/
Nazim Bouatta

Predicting the final structure

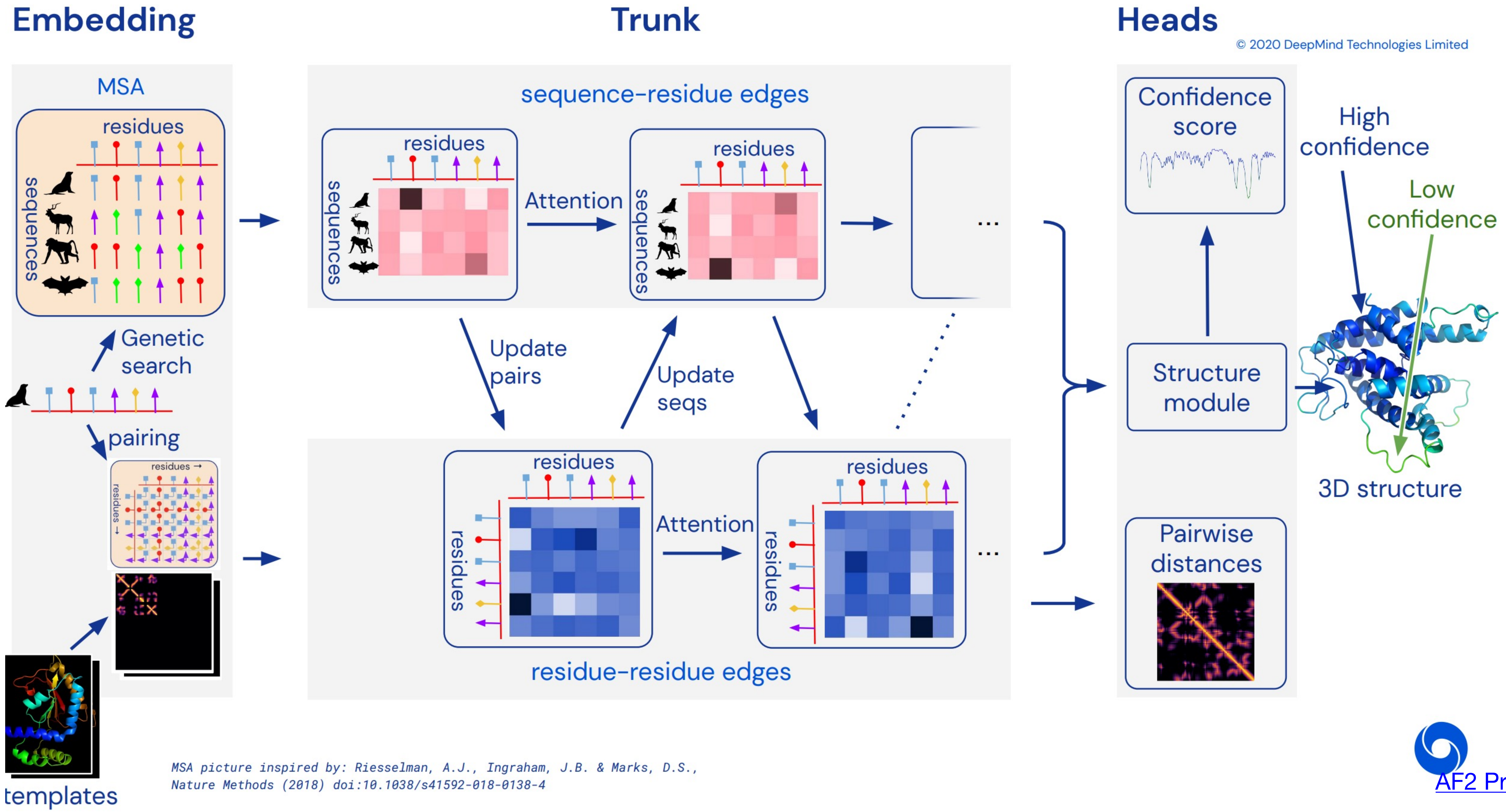
Use torsion angles to reconstruct side-chains



Mohammed AlQuraishi/
Nazim Bouatta

AF2 Overview

Communication in the trunk allow accurate head predictions



MSA picture inspired by: Riesselman, A.J., Ingraham, J.B. & Marks, D.S., Nature Methods (2018) doi:10.1038/s41592-018-0138-4

AF2 Presentation, John Jumper

6. AF2: Losses and other Details

AF2: Loss Functions, one per submodule

Nudging the network to biophysically plausible predictions

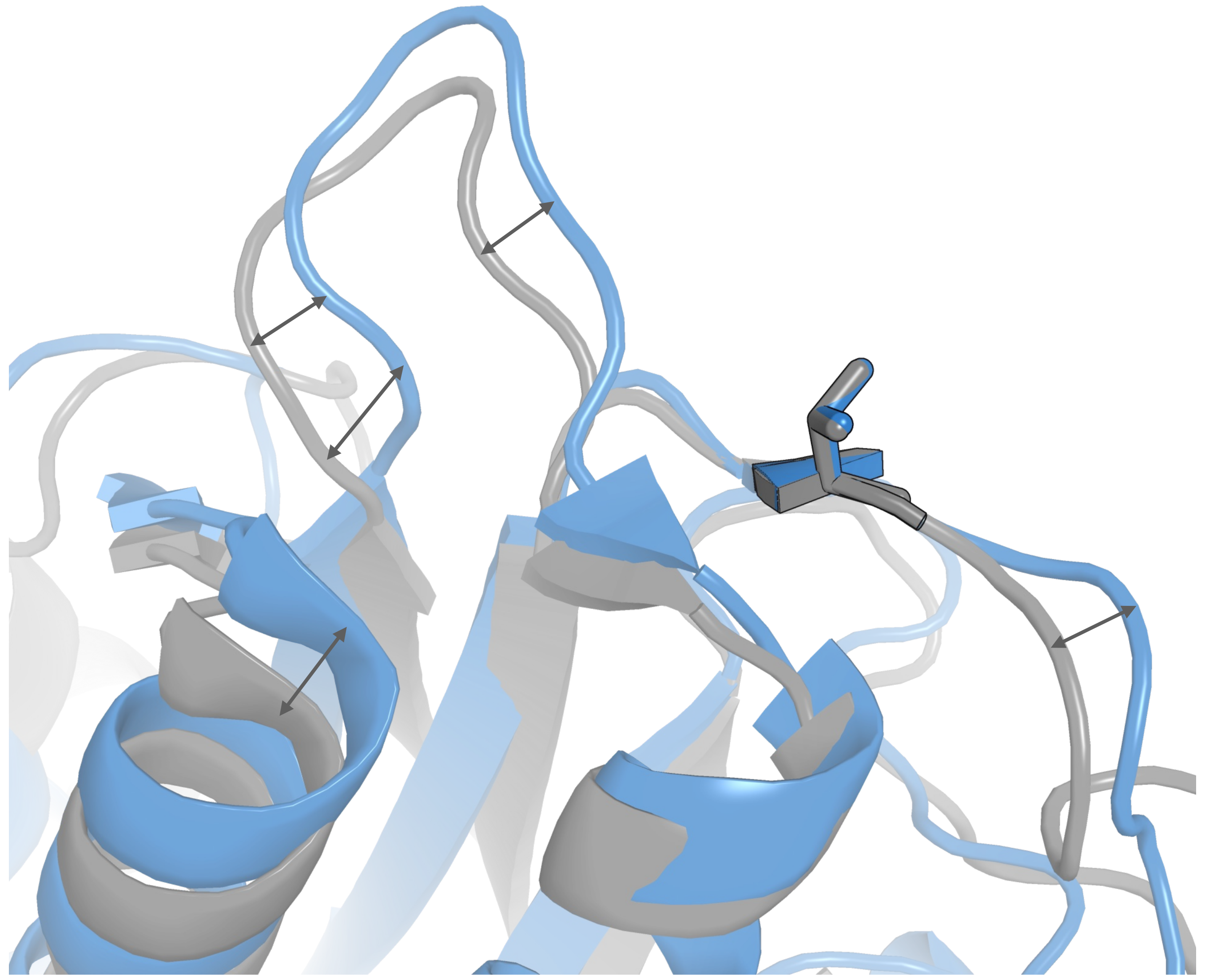
$$\mathcal{L} = \begin{cases} 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} & \text{training} \\ 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} + 0.01\mathcal{L}_{\text{exp resolved}} + 1.0\mathcal{L}_{\text{viol}} & \text{fine-tuning} \end{cases}$$

Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*

FAPE Loss for Structure Module

FAPE loss supervises relative residue positions

$$\mathcal{L} = \begin{cases} 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} & \text{training} \\ 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} + 0.01\mathcal{L}_{\text{exp resolved}} + 1.0\mathcal{L}_{\text{viol}} & \text{fine-tuning} \end{cases}$$



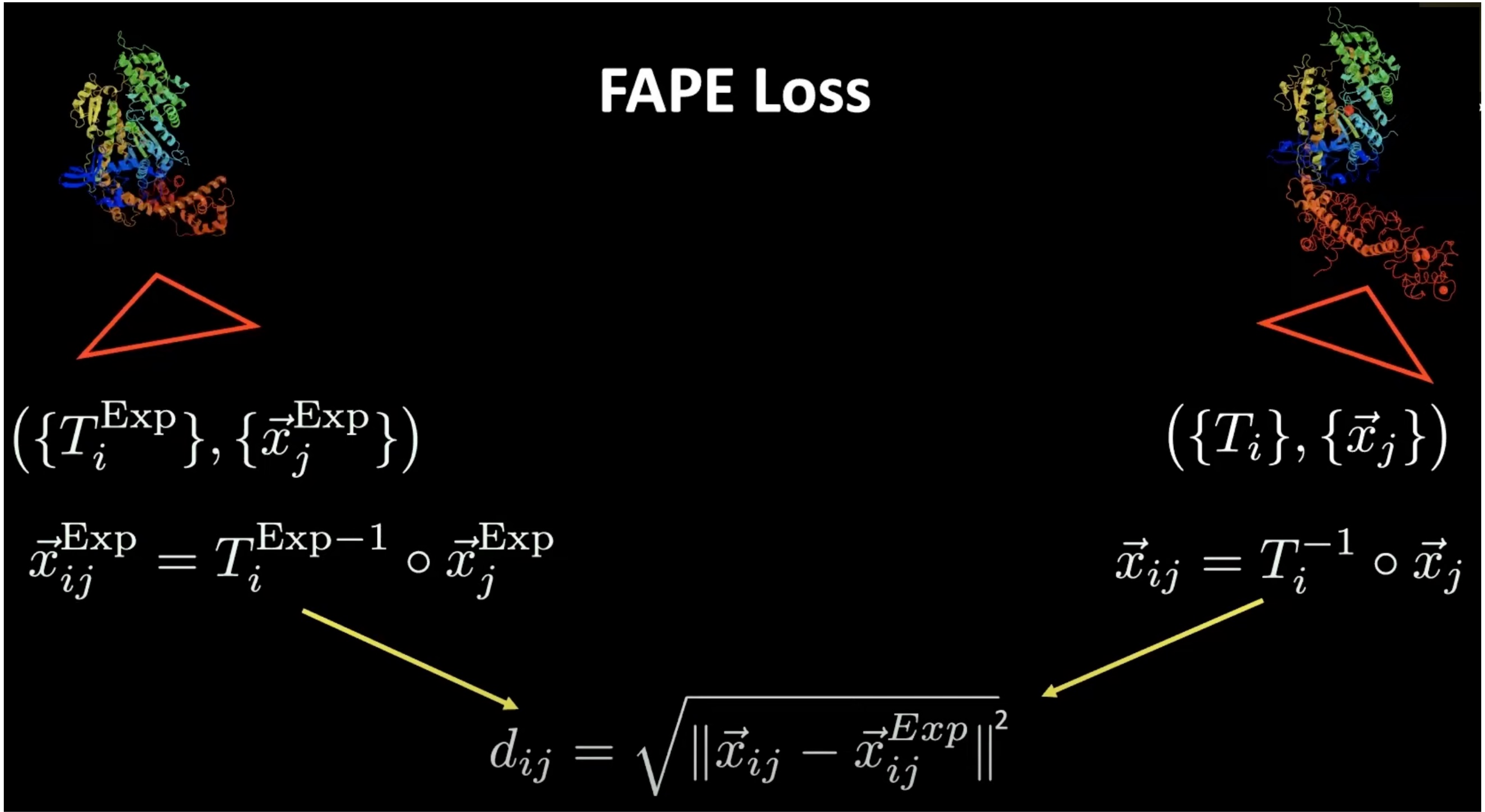
Protein-protein FAPE
(single residue alignment)

Protein-protein FAPE loss

Loss on protein coordinates under local residue frame alignments
(same as in AlphaFold)
→ **relative positioning** of protein residues

FAPE Loss for Structure Module

Again needs to take care of equivariance



FAPE Loss

$$(\{T_i^{\text{Exp}}\}, \{\vec{x}_j^{\text{Exp}}\})$$

$$(\{T_i\}, \{\vec{x}_j\})$$

$$\vec{x}_{ij}^{\text{Exp}} = T_i^{\text{Exp}-1} \circ \vec{x}_j^{\text{Exp}}$$

$$\vec{x}_{ij} = T_i^{-1} \circ \vec{x}_j$$

$$d_{ij} = \sqrt{\left\|\vec{x}_{ij} - \vec{x}_{ij}^{Exp}\right\|^2}$$

Mohammed AlQuraishi/
Nazim Bouatta

Aux Loss for Structure Module

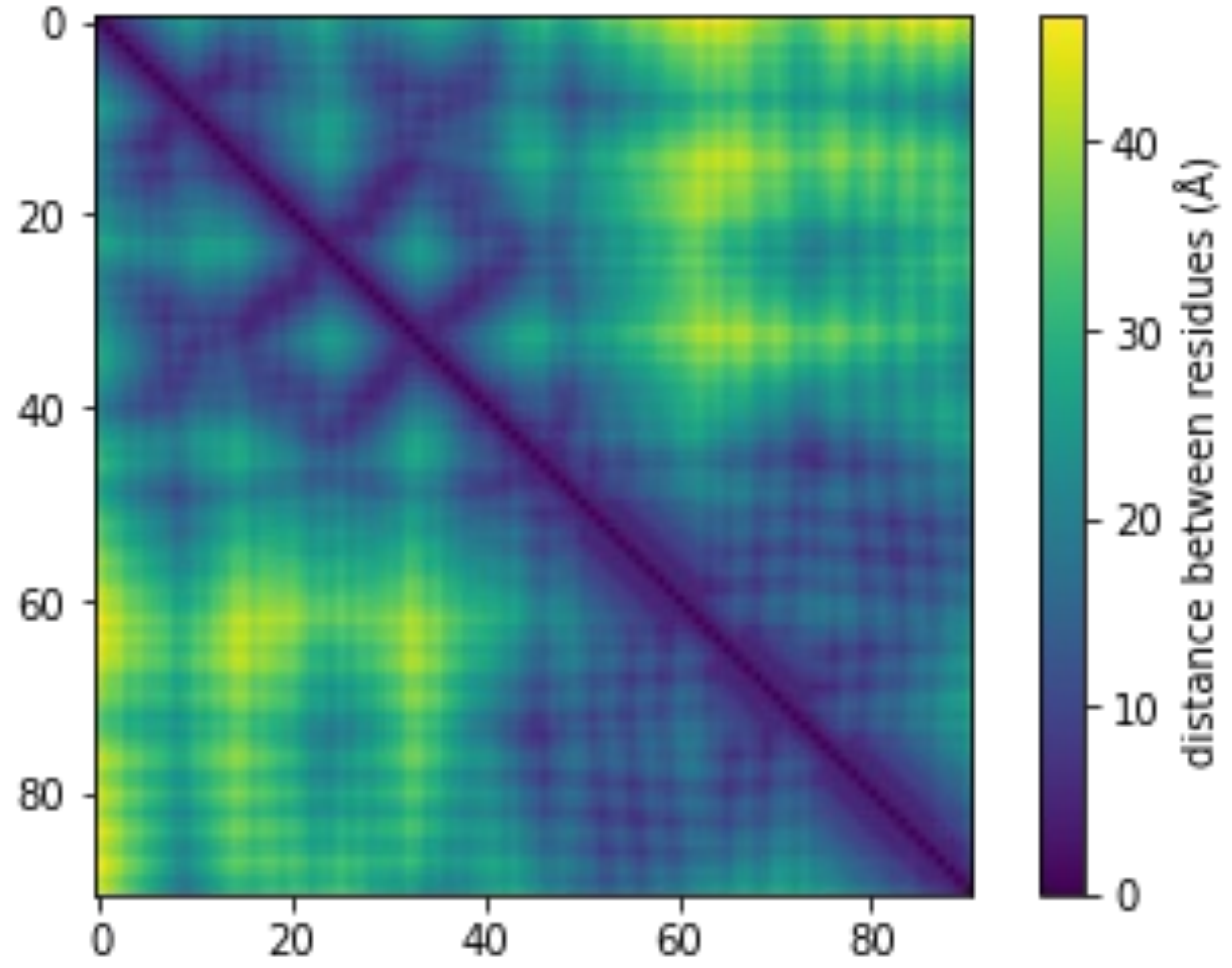
Nudging the network to biophysically plausible predictions

$$\mathcal{L} = \begin{cases} 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} & \text{training} \\ 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} + 0.01\mathcal{L}_{\text{exp resolved}} + 1.0\mathcal{L}_{\text{viol}} & \text{fine-tuning} \end{cases}$$

Distogram loss: For pair representation

Forcing the network to reason about structure

$$\mathcal{L} = \begin{cases} 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + \boxed{0.3\mathcal{L}_{\text{dist}}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} & \text{training} \\ 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + \boxed{0.3\mathcal{L}_{\text{dist}}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} + 0.01\mathcal{L}_{\text{exp resolved}} + 1.0\mathcal{L}_{\text{viol}} & \text{fine-tuning} \end{cases}$$



Distogram loss

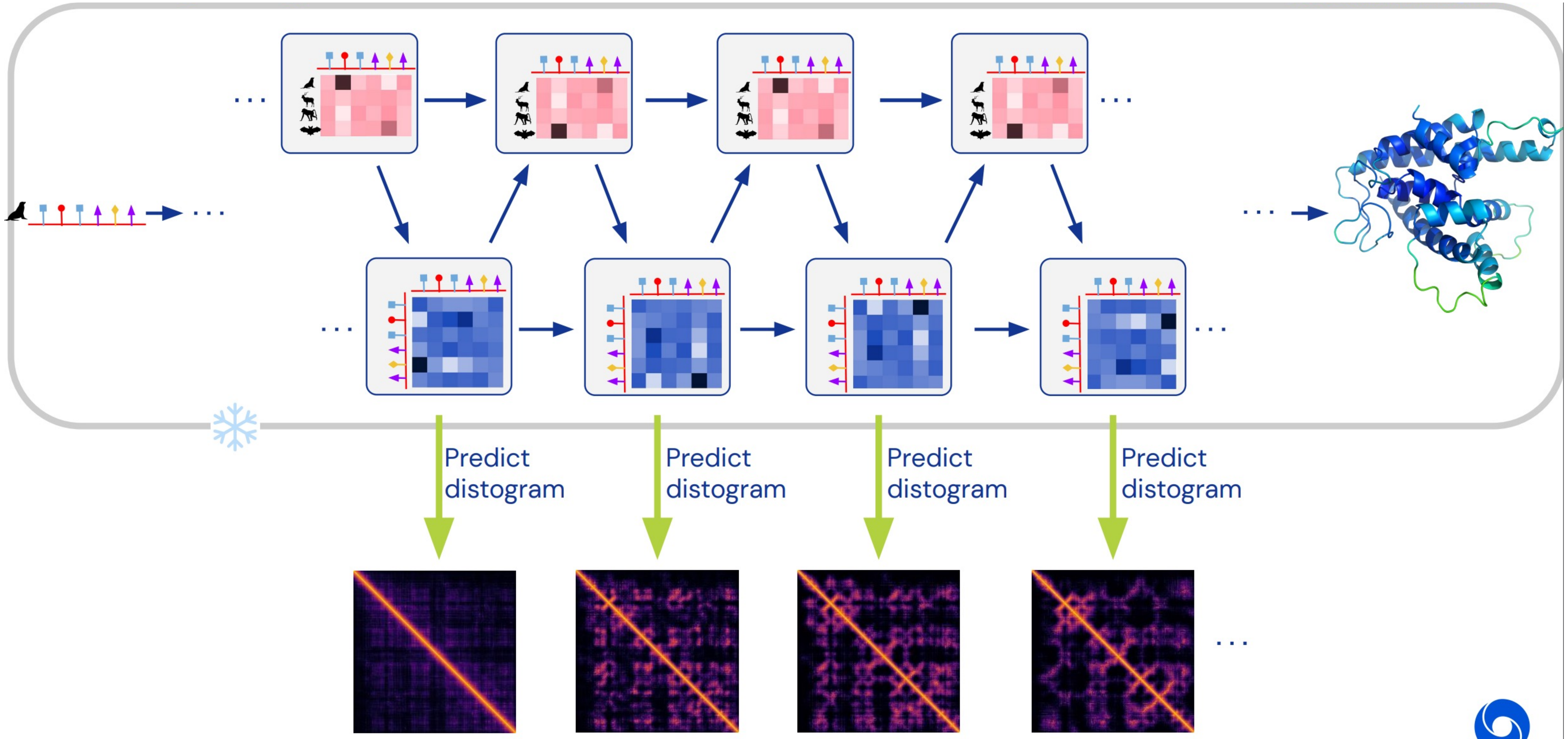
- prediction of Cα distogram

→early **structural hypothesis** in Evoformer

$$\mathcal{L}_{\text{dist}} = -\frac{1}{N_{\text{res}}^2} \sum_{i,j} \sum_{b=1}^{64} y_{ij}^b \log p_{ij}^b \;.$$

Distogram loss: For pair representation

Forcing the network to reason about structure

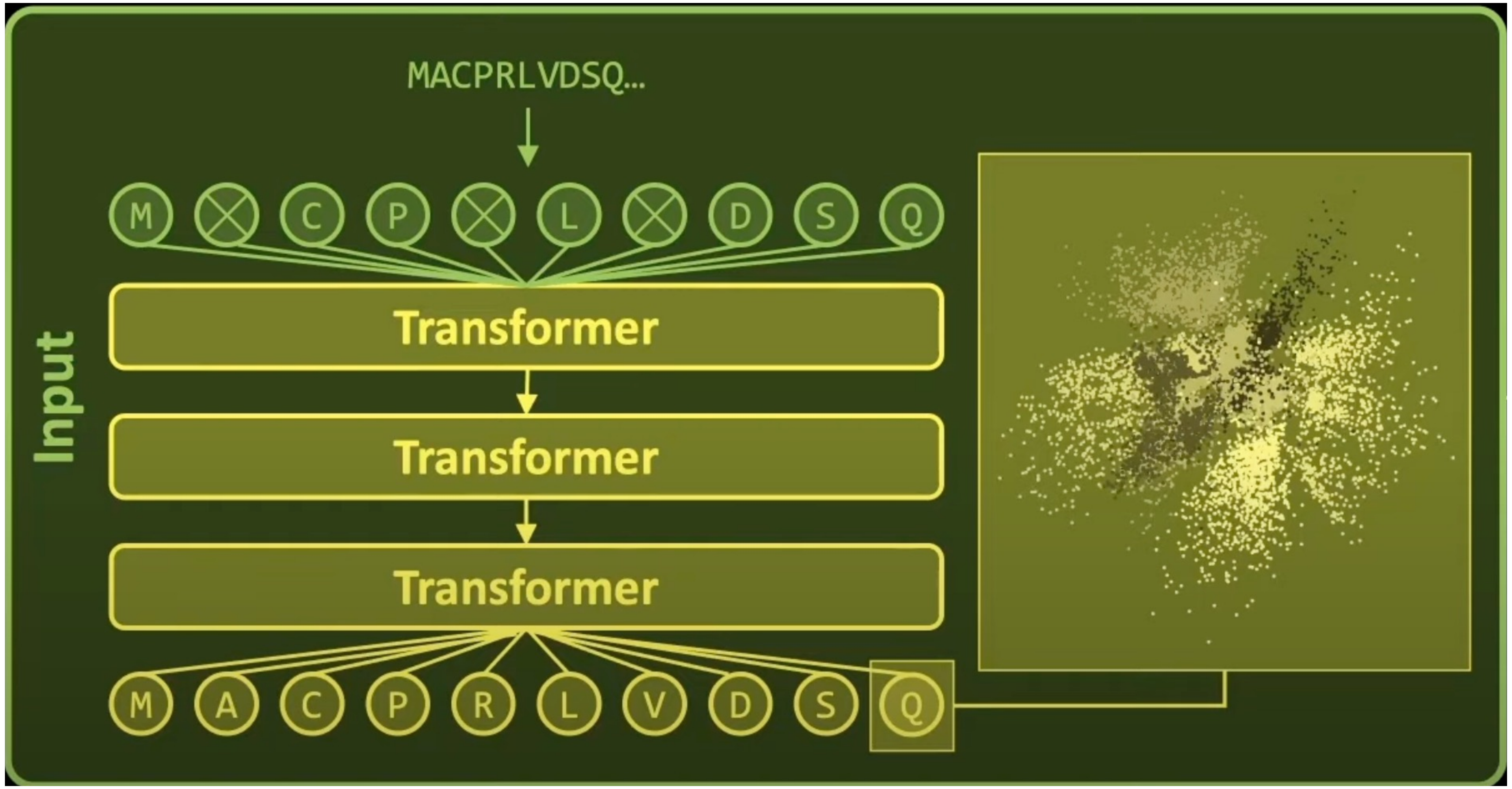


AF2 Presentation, John Jumper

MSA Loss for MSA representation

Force network to infer coevolutionary patterns

$$\mathcal{L} = \begin{cases} 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} & \text{training} \\ 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} + 0.01\mathcal{L}_{\text{exp resolved}} + 1.0\mathcal{L}_{\text{viol}} & \text{fine-tuning} \end{cases}$$



$$\mathcal{L}_{\text{msa}} = -\frac{1}{N_{\text{mask}}} \sum_{s,i \in \text{mask}} \sum_{c=1}^{23} y_{si}^{c} \log p_{si}^{c}$$

Conf Loss allows pLDDT metric

Small to not destroy the prediction accuracy

$$\mathcal{L} = \begin{cases} 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} & \text{training} \\ 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} + 0.01\mathcal{L}_{\text{exp resolved}} + 1.0\mathcal{L}_{\text{viol}} & \text{fine-tuning} \end{cases}$$

Carlos Outeiral Rubiera

AF2: Loss Functions

Nudging the network to biophysically plausible predictions

$$\mathcal{L} = \begin{cases} 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} & \text{training} \\ 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} + \boxed{0.01\mathcal{L}_{\text{exp resolved}}} + 1.0\mathcal{L}_{\text{viol}} & \text{fine-tuning} \end{cases}$$

$$\mathcal{L}_{\text{exp resolved}} = \text{mean}_{(i,a)} \left(-y_i^a \log p_i^{\text{exp resolved},a} - (1 - y_i^a) \log(1 - p_i^{\text{exp resolved},a}) \right)$$

Carlos Outeiral Rubiera

Viol Loss for biophysically plausible structures

Only used during fine-tuning, otherwise accuracy drop

$$\mathcal{L} = \begin{cases} 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} & \text{training} \\ 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} + 0.01\mathcal{L}_{\text{exp resolved}} + \boxed{1.0\mathcal{L}_{\text{viol}}} & \text{fine-tuning} \end{cases}$$

$$\mathcal{L}_{\text{viol}} = \mathcal{L}_{\text{bondlength}} + \mathcal{L}_{\text{bondangle}} + \mathcal{L}_{\text{clash}}$$

Carlos Outeiral Rubiera

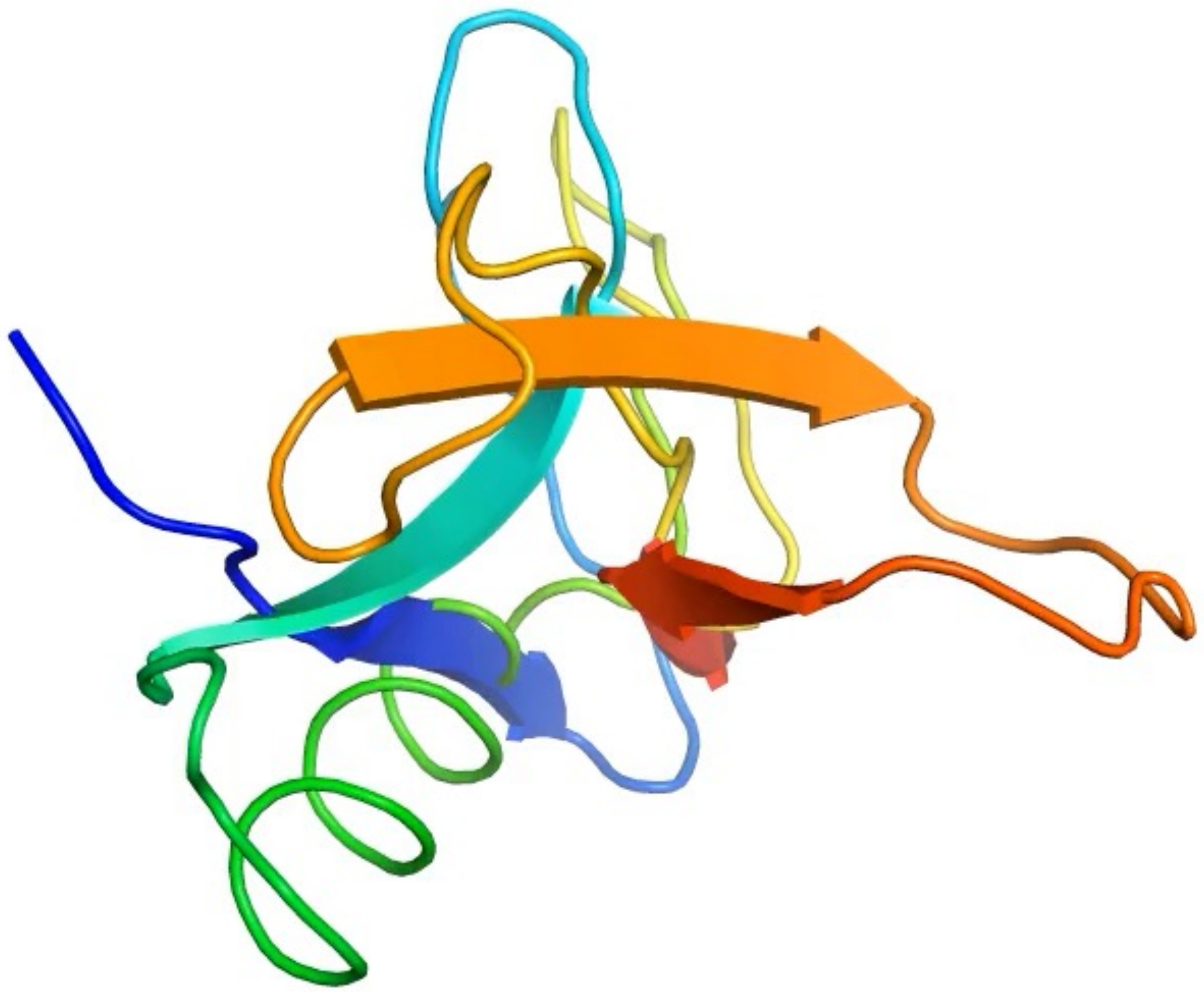
7. Impact and Outlook

Easy targets – early structure hypothesis

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

Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*

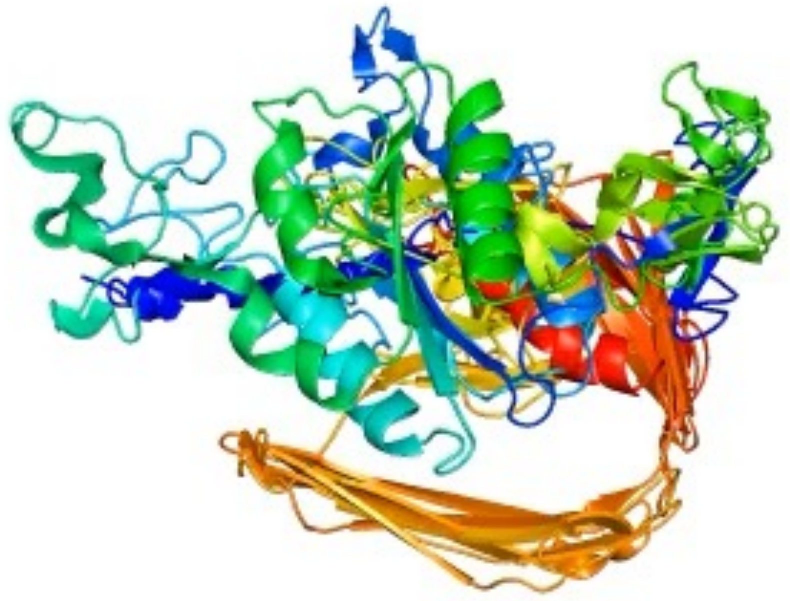
Hard targets – late structure hypothesis



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

Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*

Unphysical structures explored



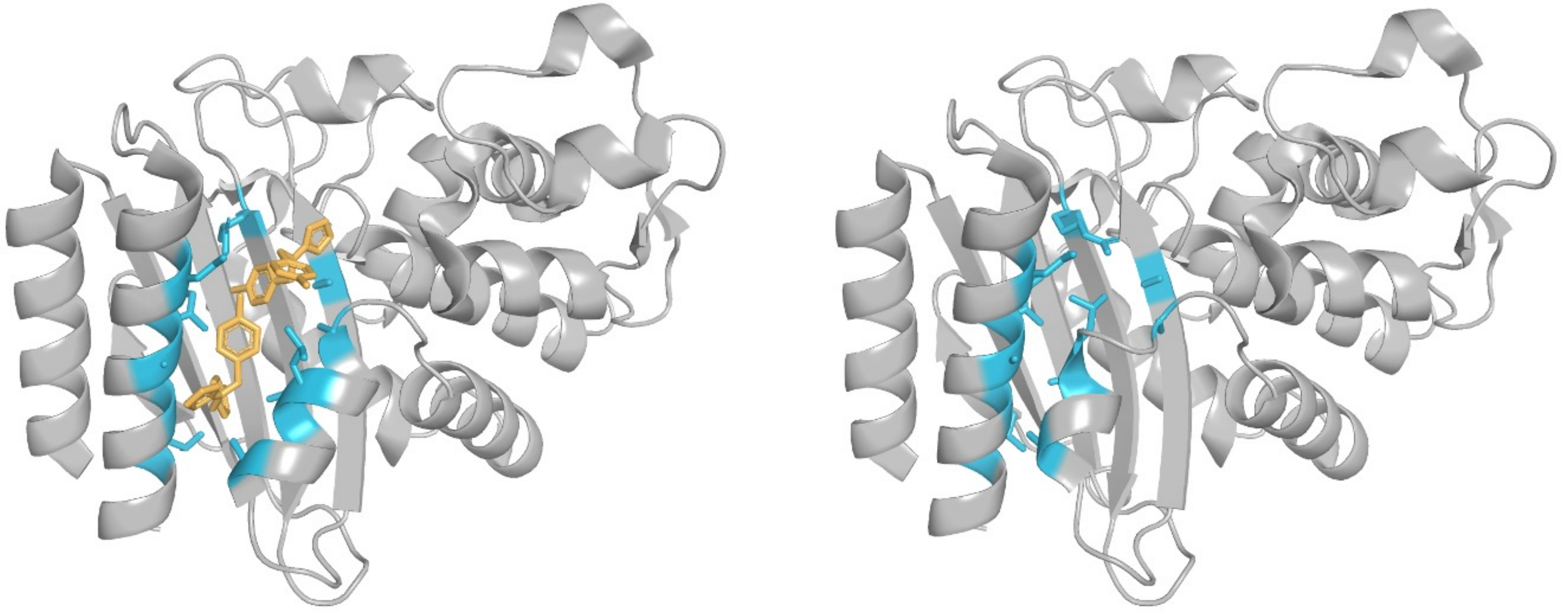
Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*

AF2: Limitations

Unaware of bound/unbound states

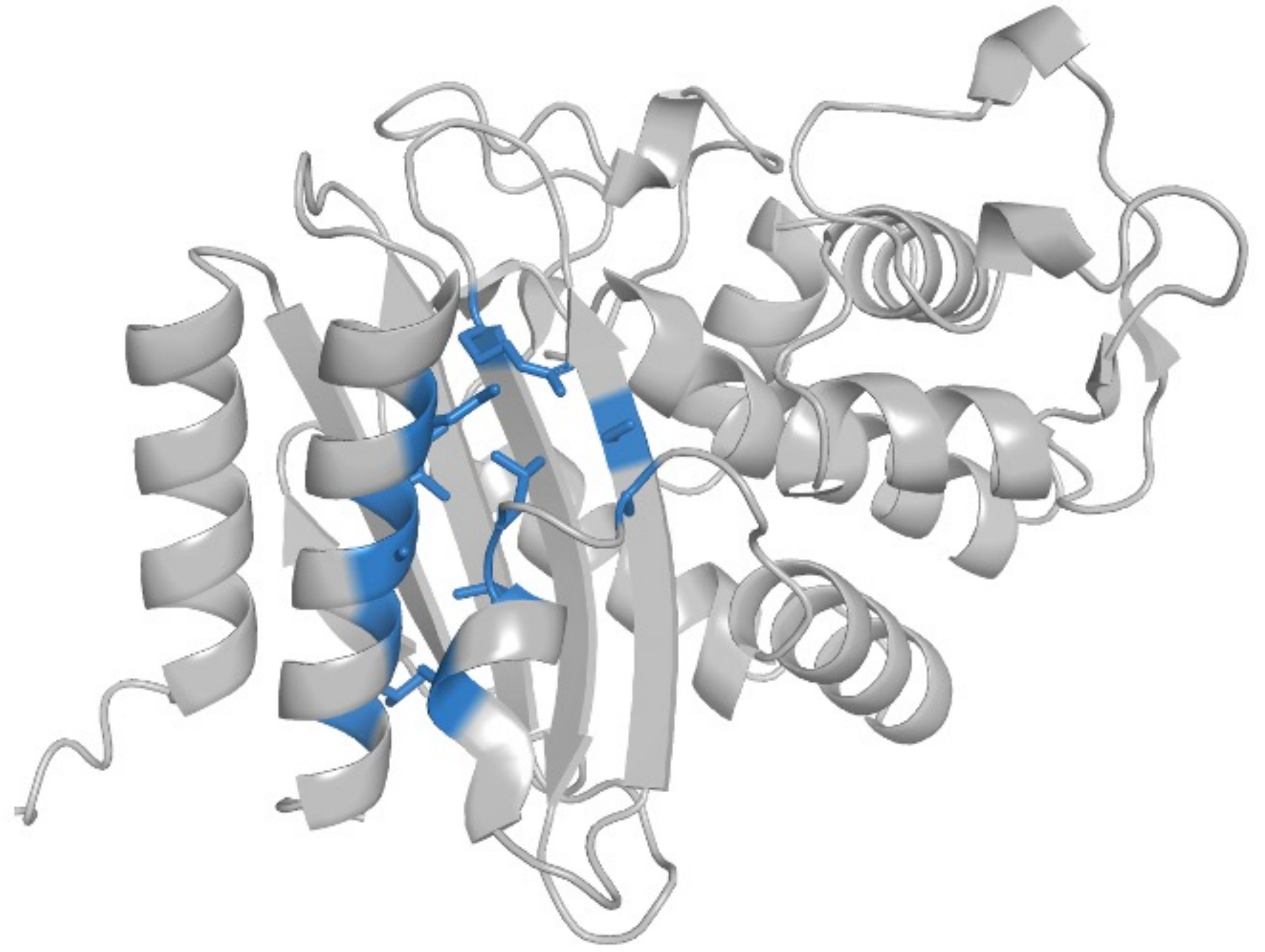
Example: beta-lactamase in complex with inhibitor molecules



holo-state
(PDB: 1PZO)

apo-state
(PDB: 1JWP)

Crystal structures



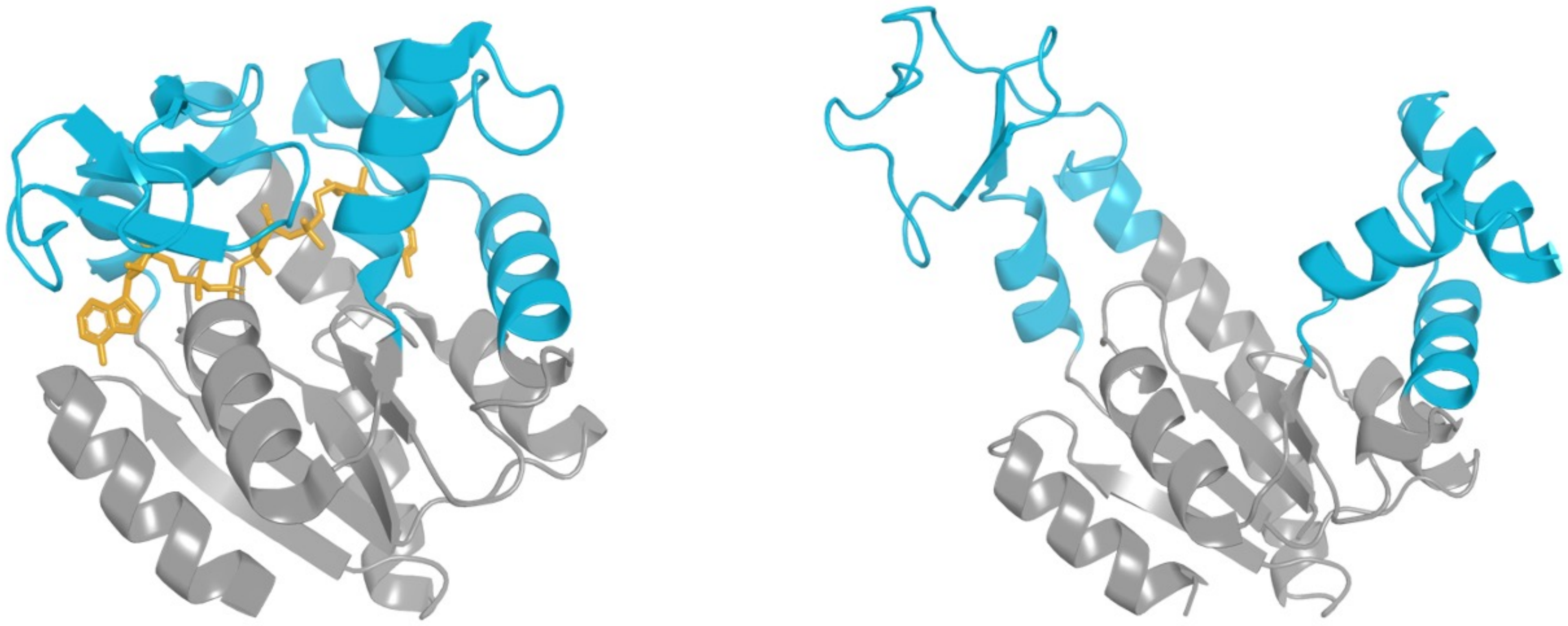
AlphaFold2 prediction

Lukas Jarosch

AF2: Limitations

Unaware of bound/unbound states

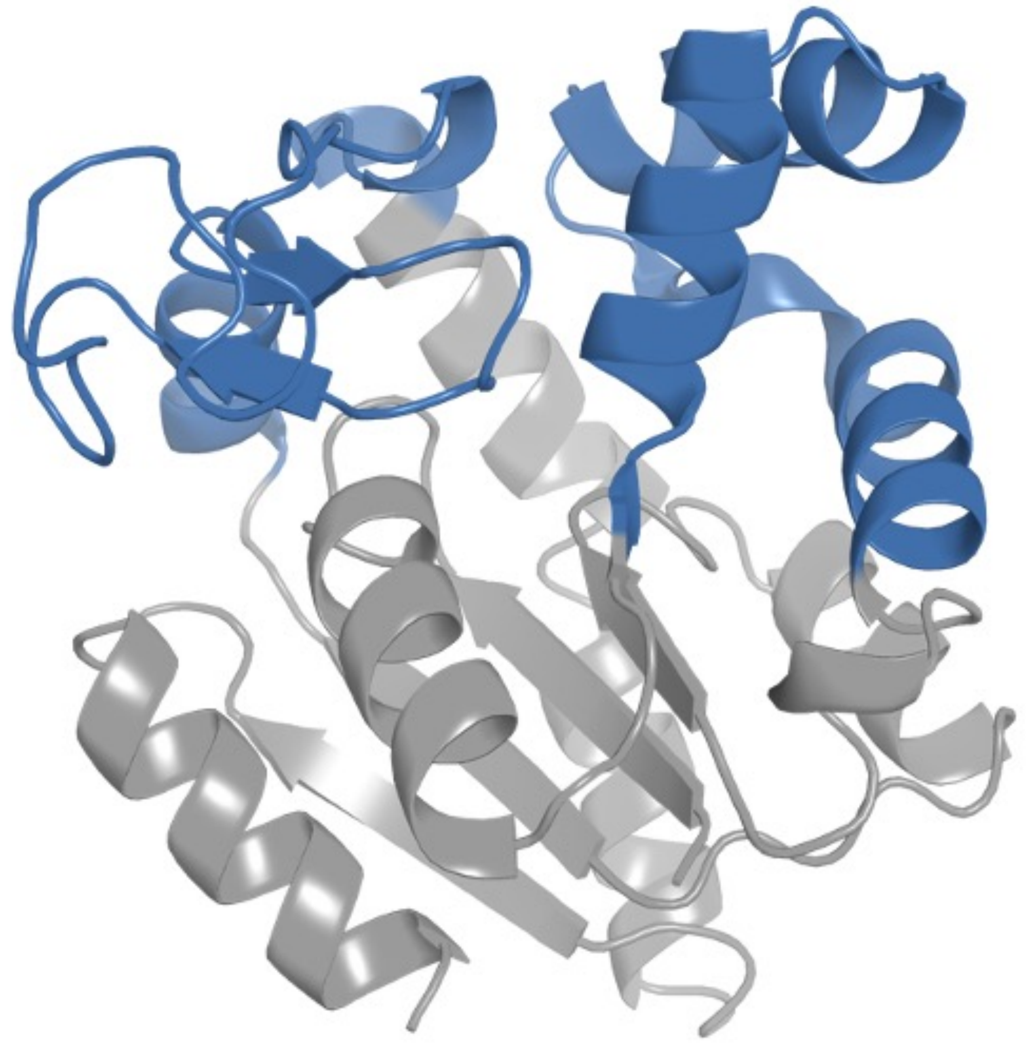
Example: adenylate-kinase binding to substrate



holo-state
(PDB: 1AKE)

apo-state
(PDB: 4AKE)

Crystal structures

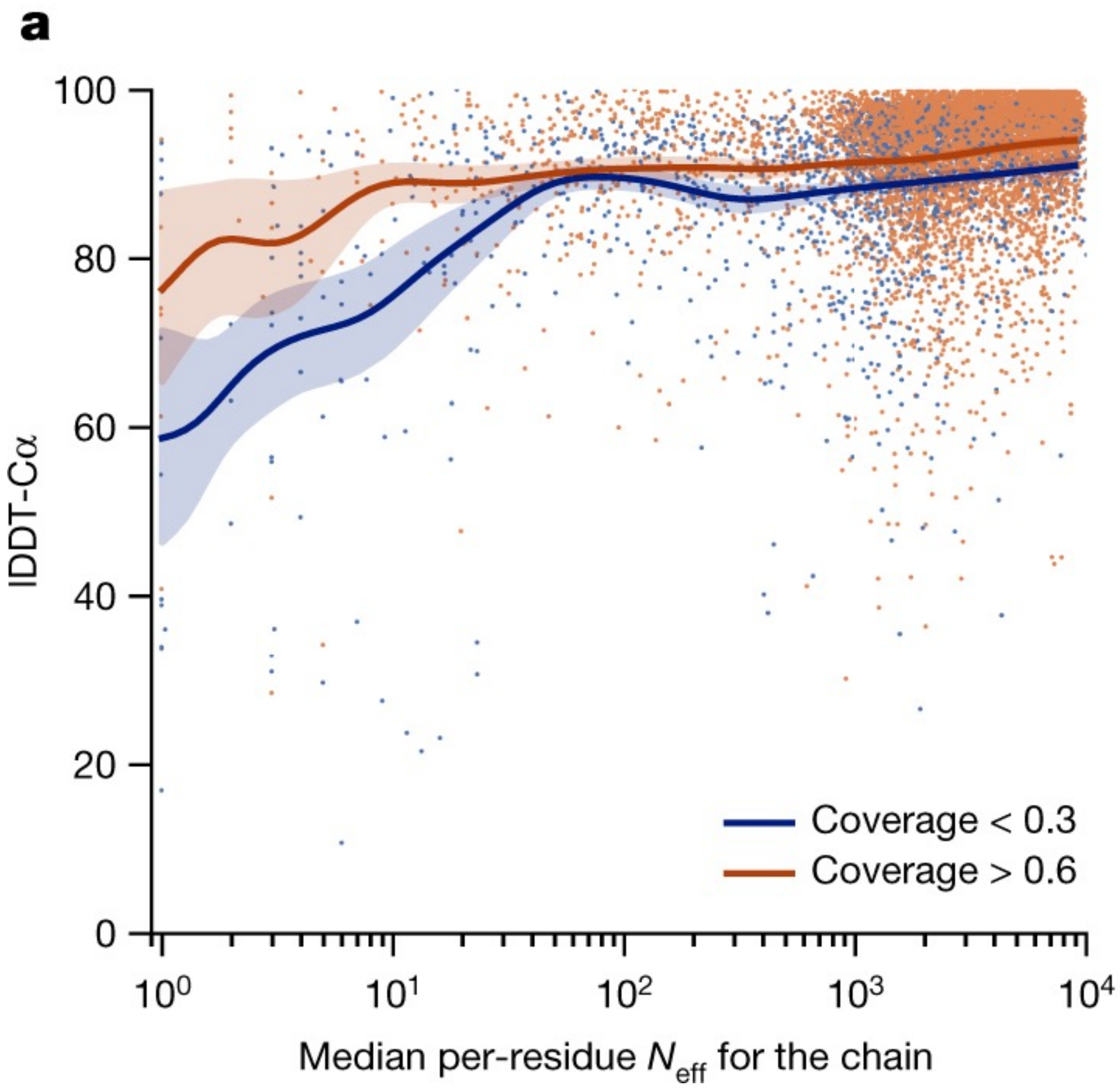


AlphaFold2 prediction

Lukas Jarosch

AF2: Limitations

Susceptible to shallow MSAs



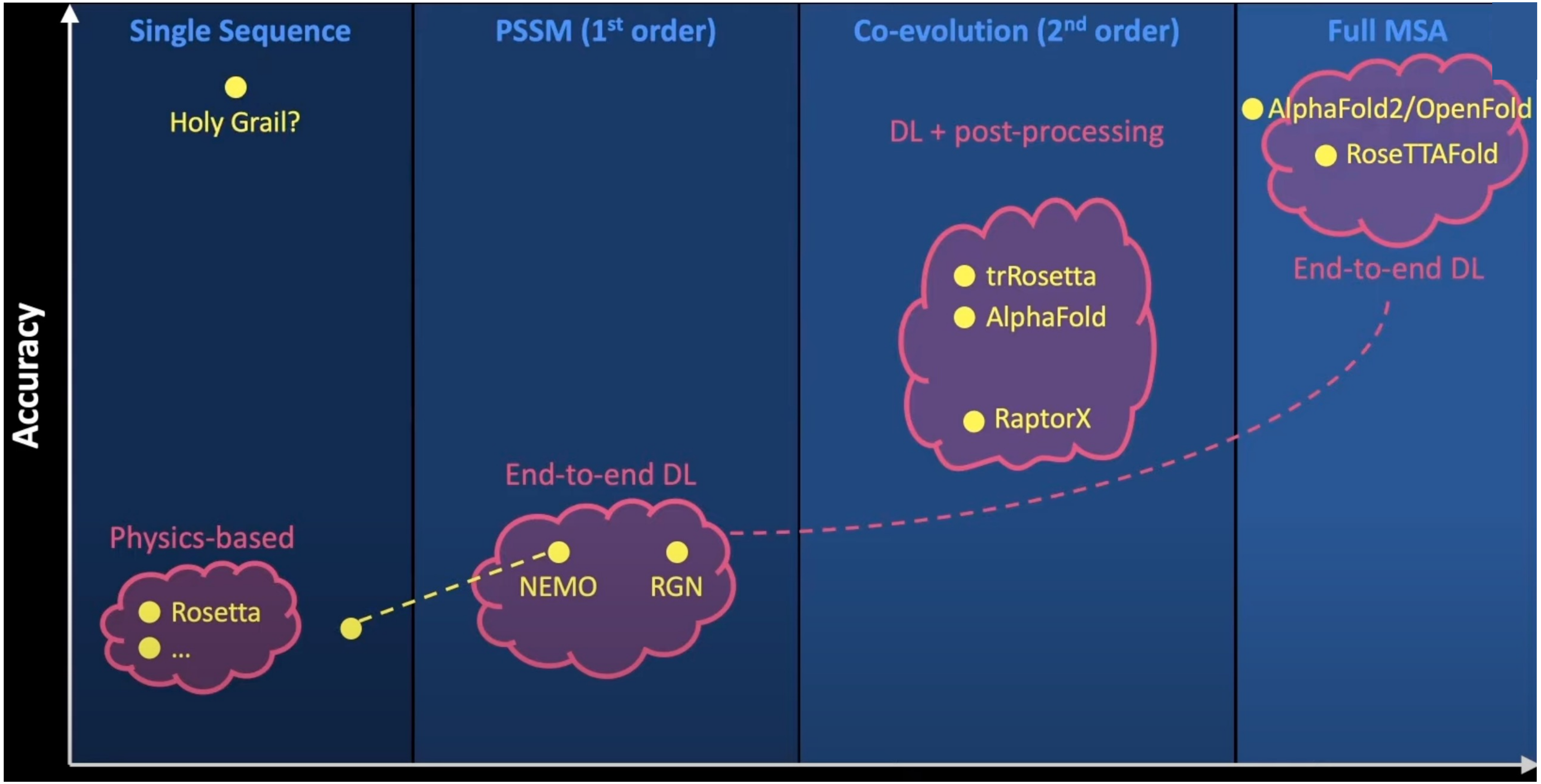
Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*

AF2: Limitations

Problems with less structured/more variable protein families

How to improve protein structure prediction?

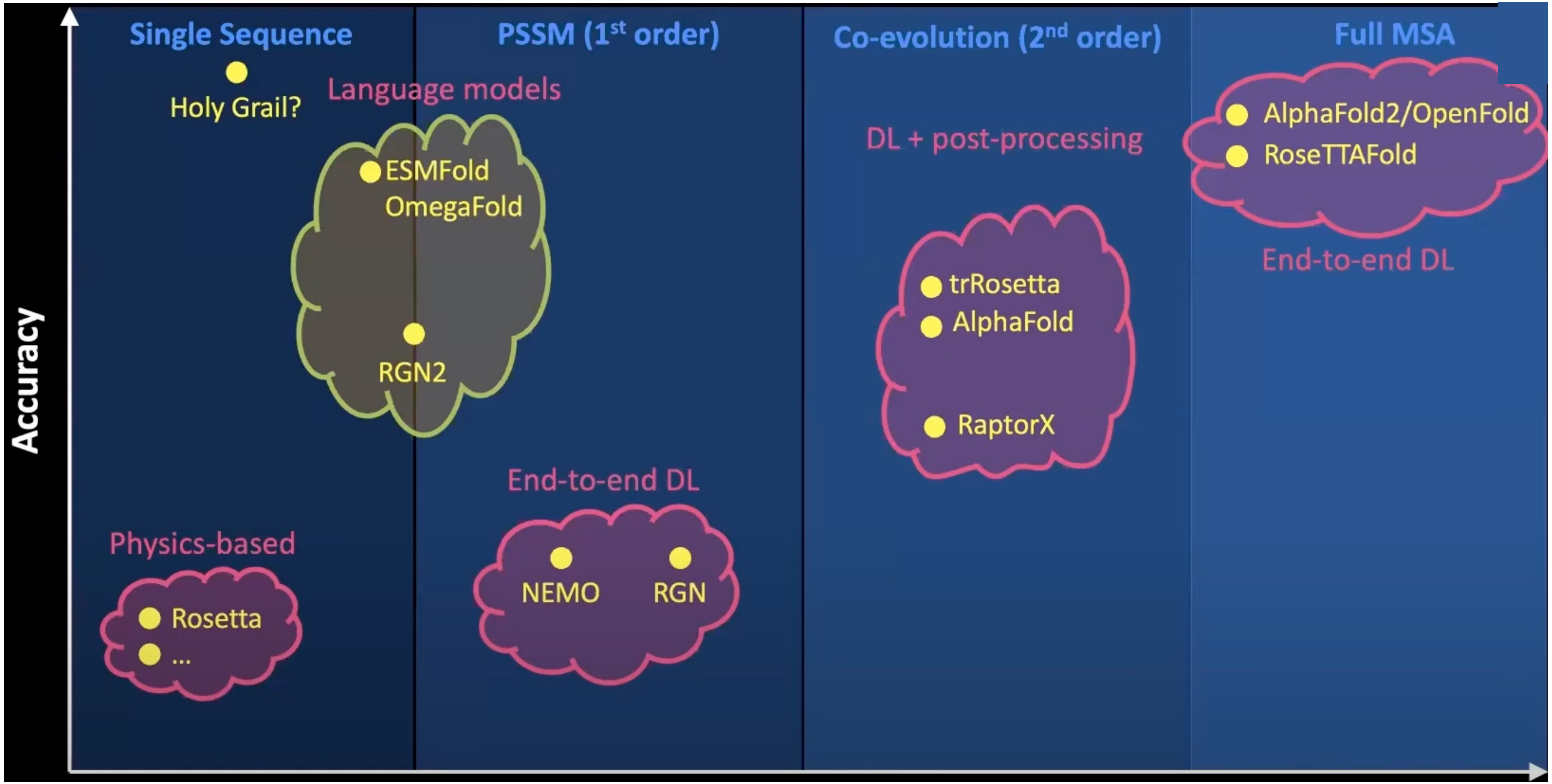
Subheading



Mohammed AlQuraishi/
Nazim Bouatta

How to improve protein structure prediction?

Subheading



Mohammed AlQuraishi/
Nazim Bouatta

Are MSAs the best input we can use?

Subheading

Input

MACPRLVDSQ...
MACPR-VDCN...
MGCPRILDSH...
MGCGKIVESD...
M-CGKLIEAT...
MACARV-DAY...
MGCRRKLDCE...

Torso

Output

Mohammed AlQuraishi/
Nazim Bouatta

Are MSAs the best input we can use?

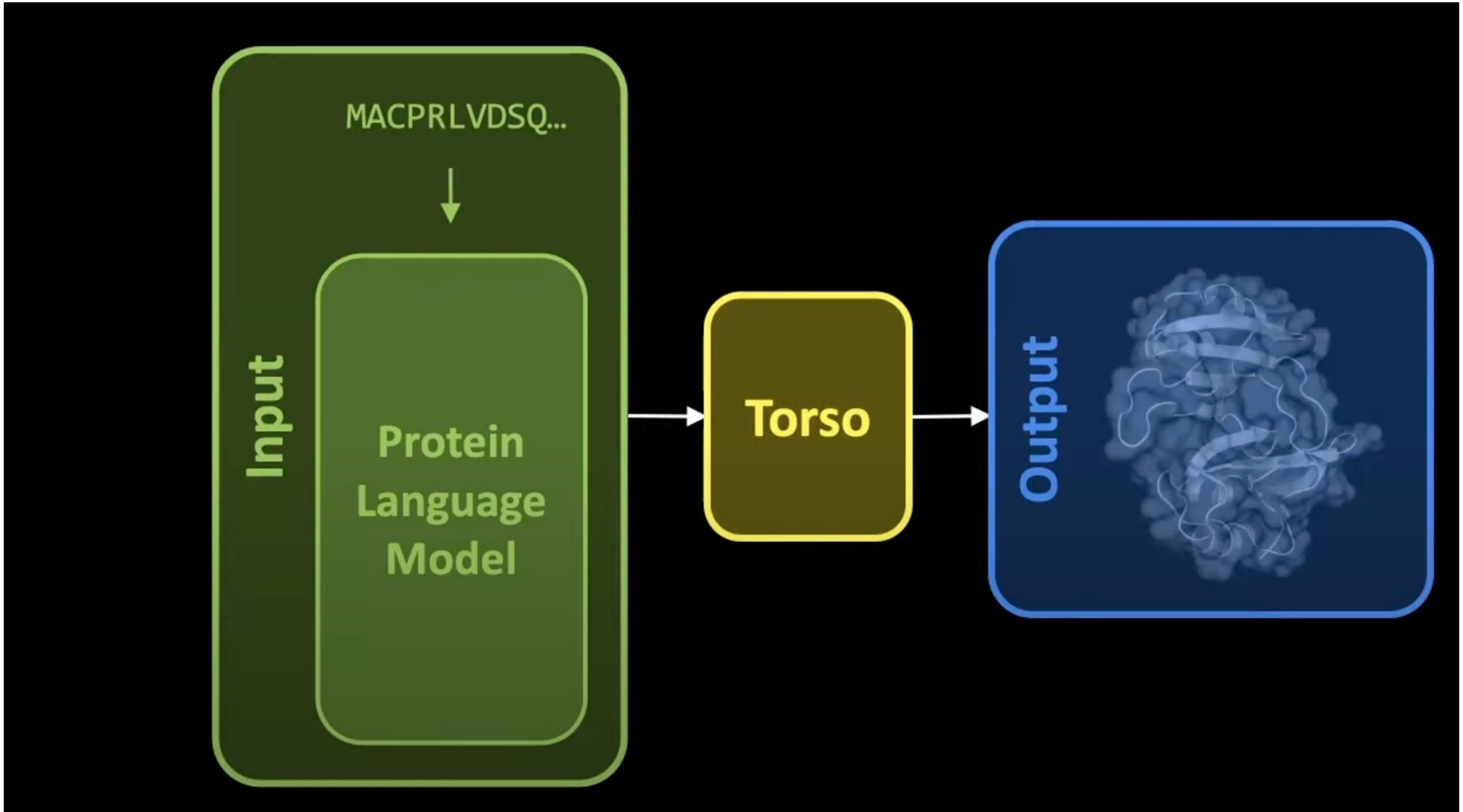
Subheading

Input
MACPRLVDSQ...
Torso
Output

Mohammed AlQuraishi/
Nazim Bouatta

Are MSAs the best input we can use?

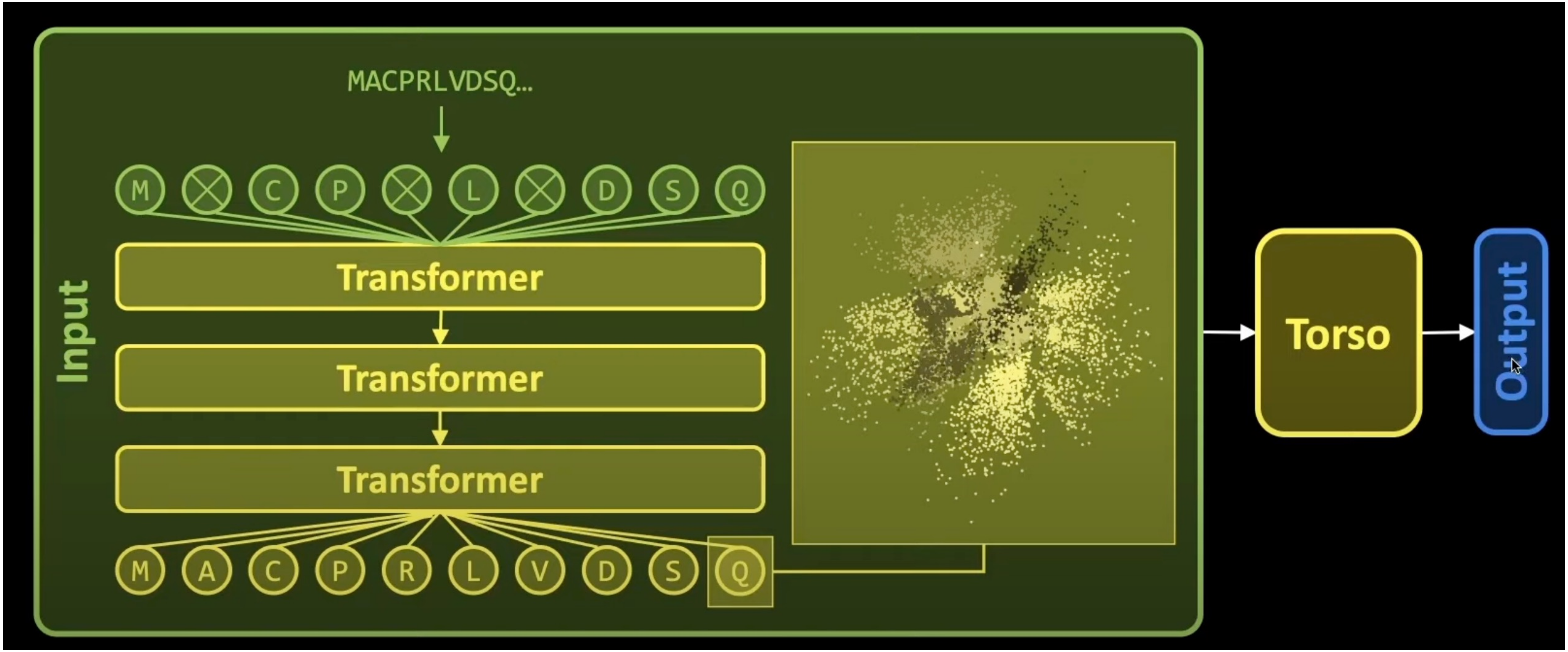
Subheading



Mohammed AlQuraishi/ Nazim Bouatta

Protein Language Models

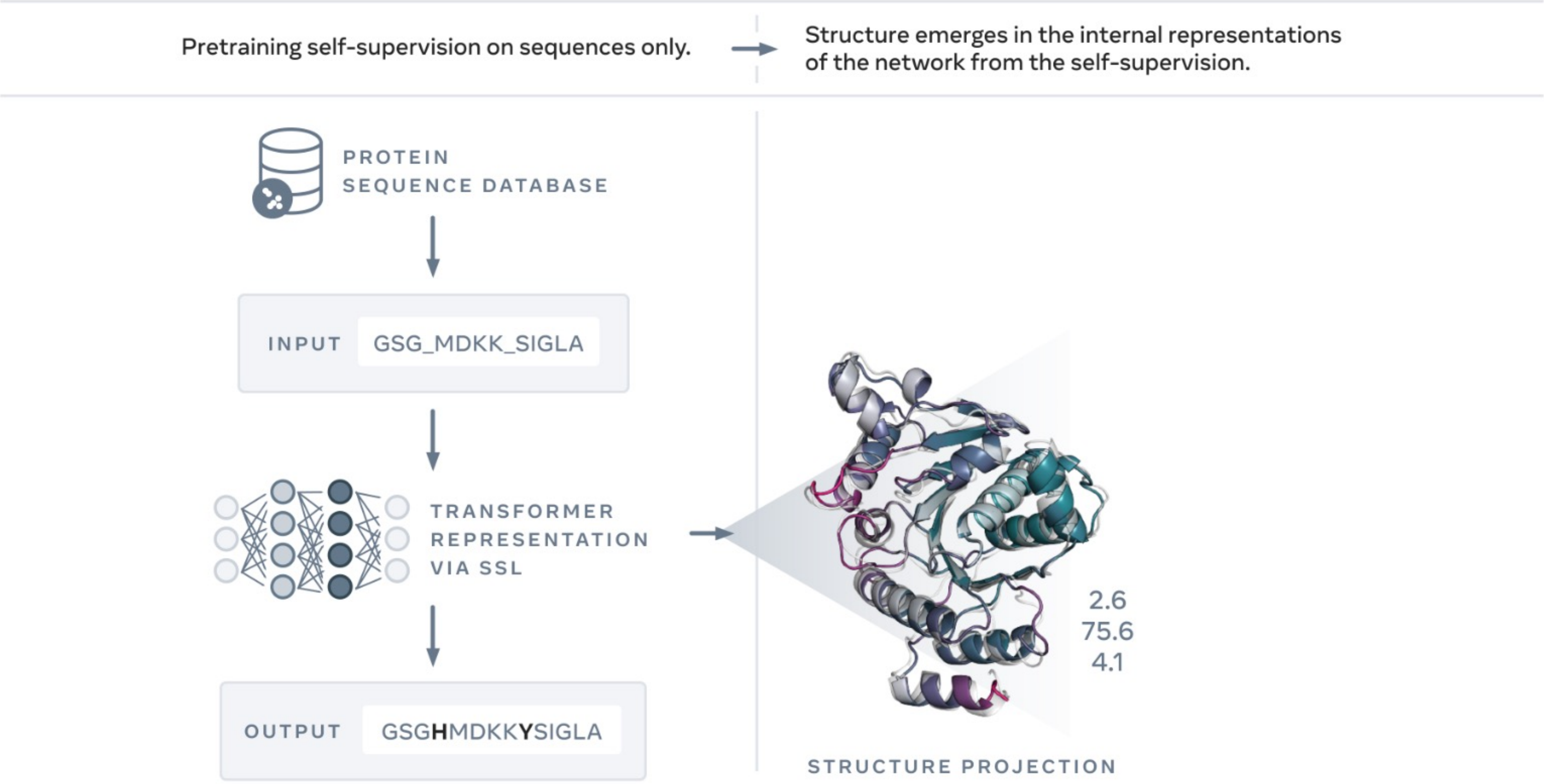
Subheading



Mohammed AlQuraishi/
Nazim Bouatta

Structural Information emerges

Unexpected consequence of self-supervised pretraining



Pretraining self-supervision on sequences only. → Structure emerges in the internal representations of the network from the self-supervision.

PROTEIN SEQUENCE DATABASE

INPUT GSG_MDKK_SIGLA

TRANSFORMER REPRESENTATION VIA SSL

OUTPUT GSG**H**MDKK**Y**SIGLA

STRUCTURE PROJECTION

The ESM-2 language model is trained to predict amino acids that have been masked out of sequences across evolution. We discovered that, as a result of this training, information about the protein's structure emerges in the internal states of the model. This is surprising because the model has been trained only on sequences.

ESMFold Blog

🥡 Takeaway 🥡

Deep Learning revolutionized protein structure prediction, but for applications many important challenges remain.