CS 364 COMPUTATIONAL BIOLOGY

Sara Mathieson
Haverford College

Outline

- RNA secondary structure prediction
- Protein structure prediction (AlphaFold)
- Further application of computational biology
- Concluding thoughts

Notes:

--Project meetings in lab Thursday
--Presentations Thursday!
--Presentation instructions and final deliverables posted
--Writeup and repo due **Friday Dec 20 at noon**

Project Presentation Notes

- Date: in-class Thursday, Dec 11
- Each group will have 10-12 minutes to present (+ time for questions and transition)
- Email me your slides by 12pm on Dec 11! (PDF only)
- I will have a laser pointer / slide advancer clicker

Project Presentation Notes

Your presentation should include

- Motivation and Scientific Question
- Data and Methods
- Results and Interpretation
- Conclusions and Future Work

Presentation Tips

- Speak loudly (to the back of the class)
- Avoid text-heavy slides, use images/diagrams
- Include citations for any figures you did not make
- Ask at least one question to another group

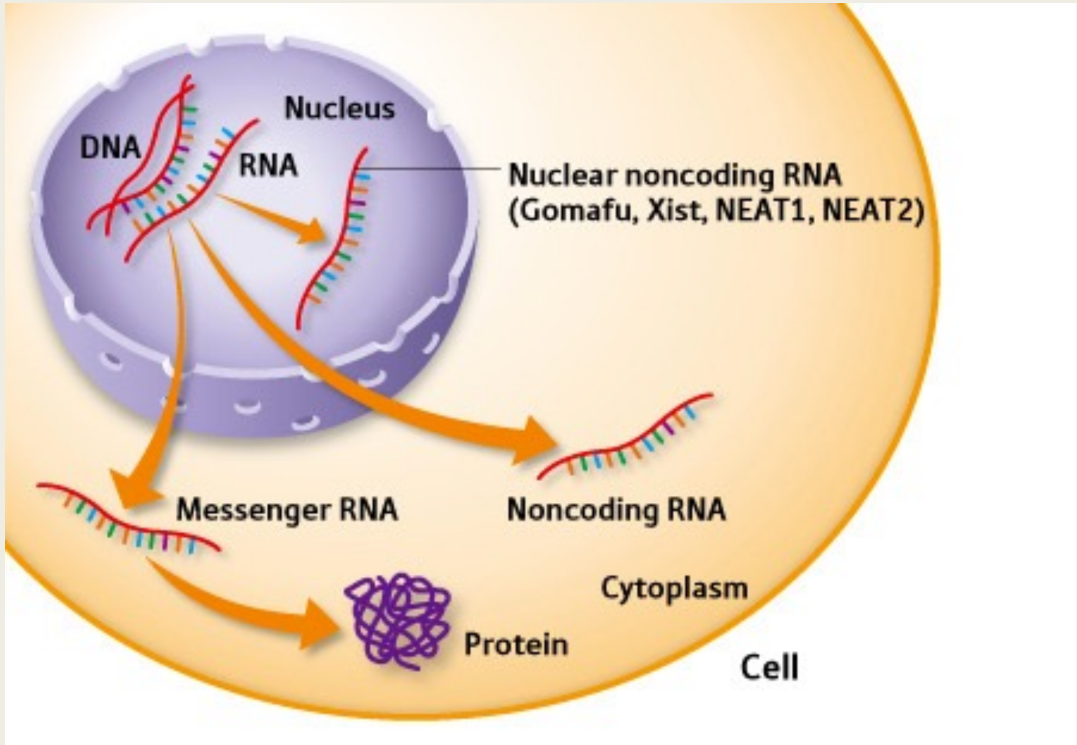
Submit by 12pm on Dec 20 (github)

- Writeup
- README
- All project code

Think about reproducibility!

Beyond a linear sequence…

RNA folding



- RNA does not stay as a linear sequence
- It folds into a secondary structure that minimizes energy

https://www.youtube.com/watch?v=KBI69y2ziXw

Image: genius.com

RNA secondary structure: larger example

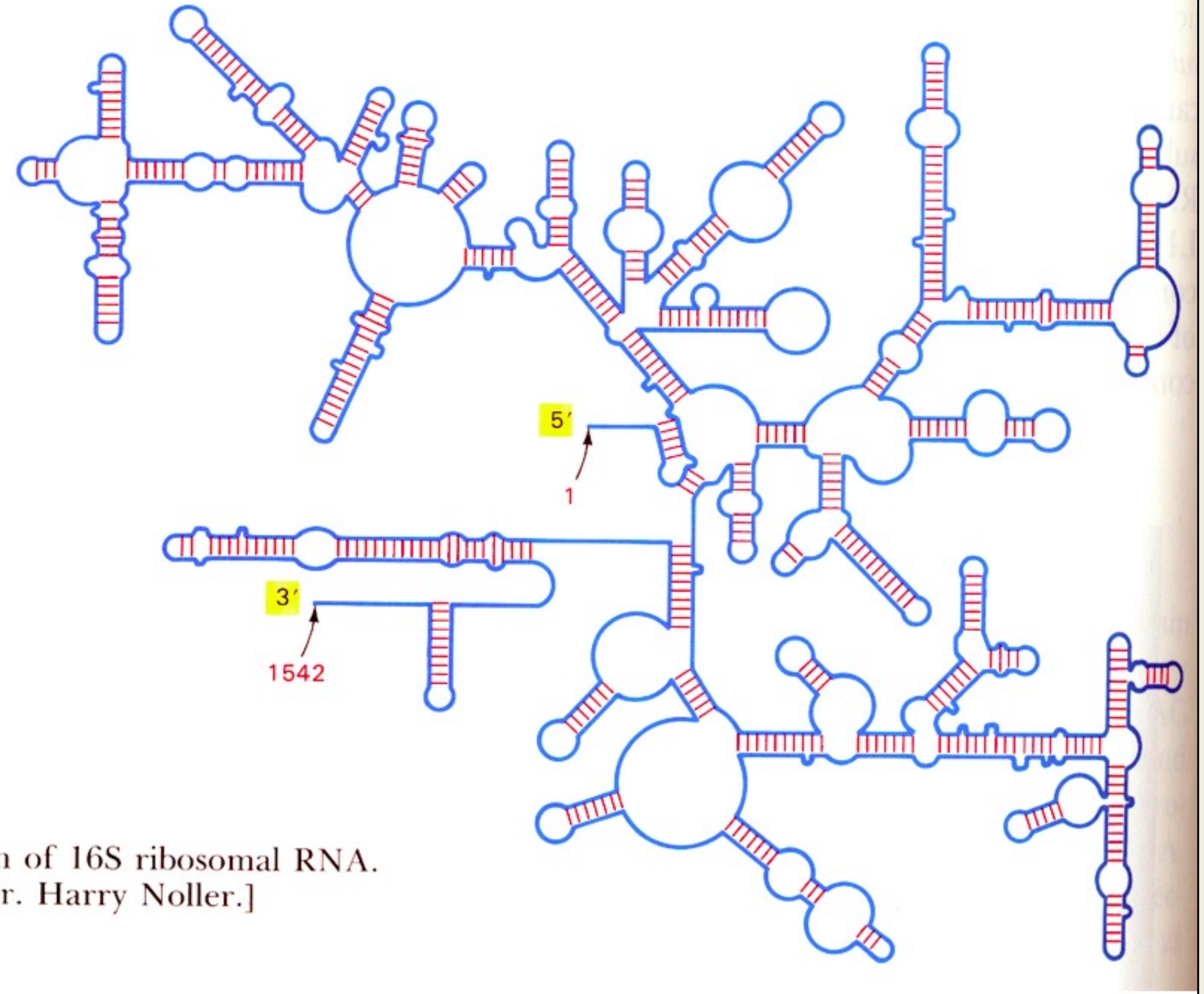
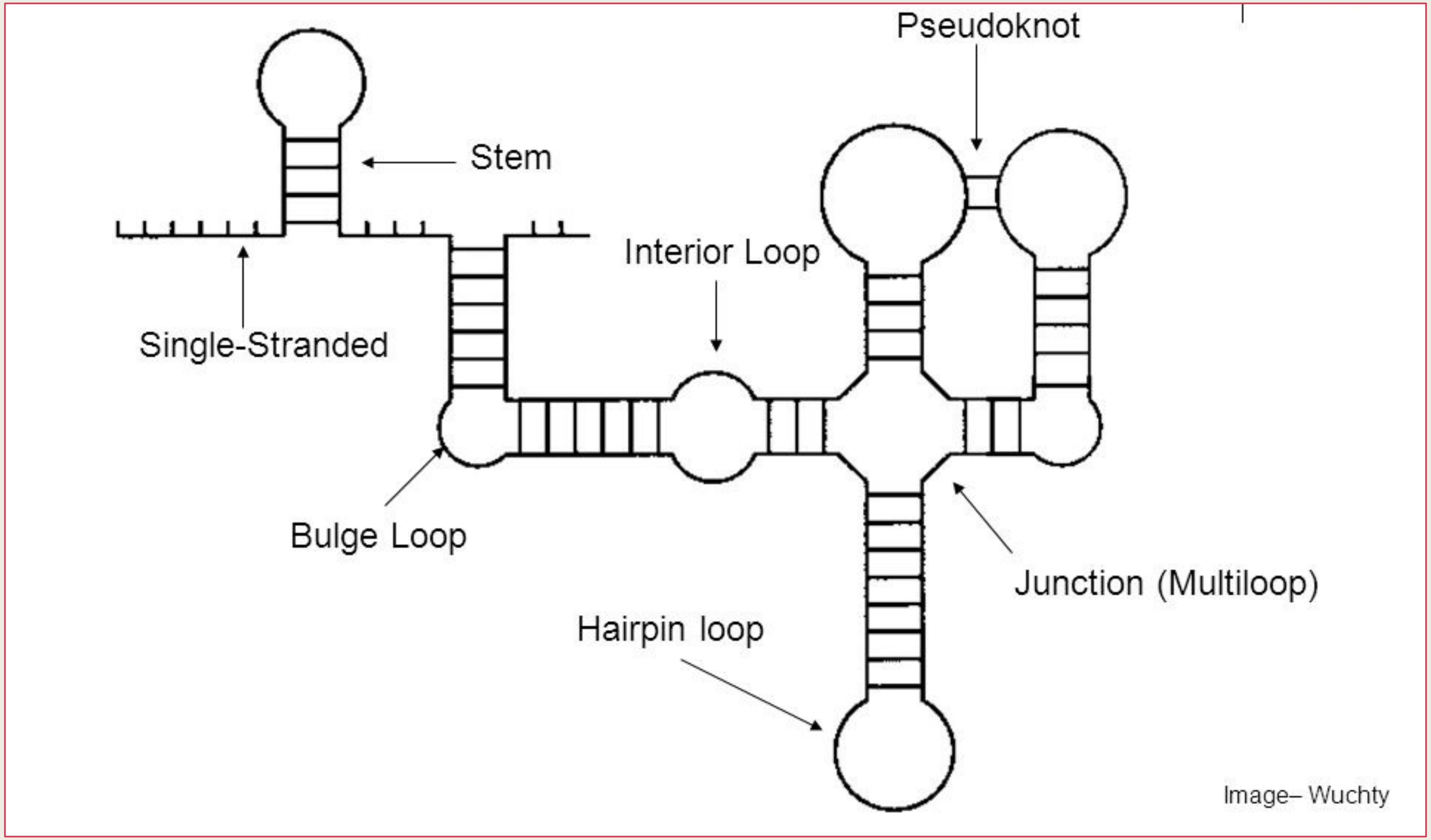
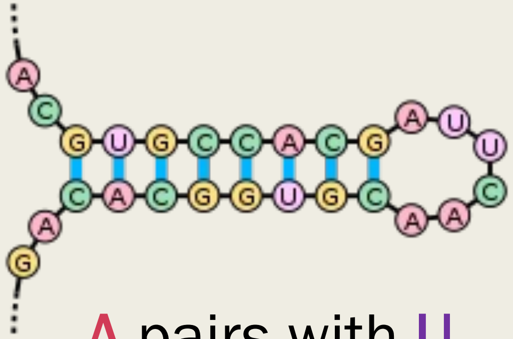


Figure 30-18
Folding pattern of 16S ribosomal RNA.
[Courtesy of Dr. Harry Noller.]

Features of RNA secondary structure



Image– Wuchty



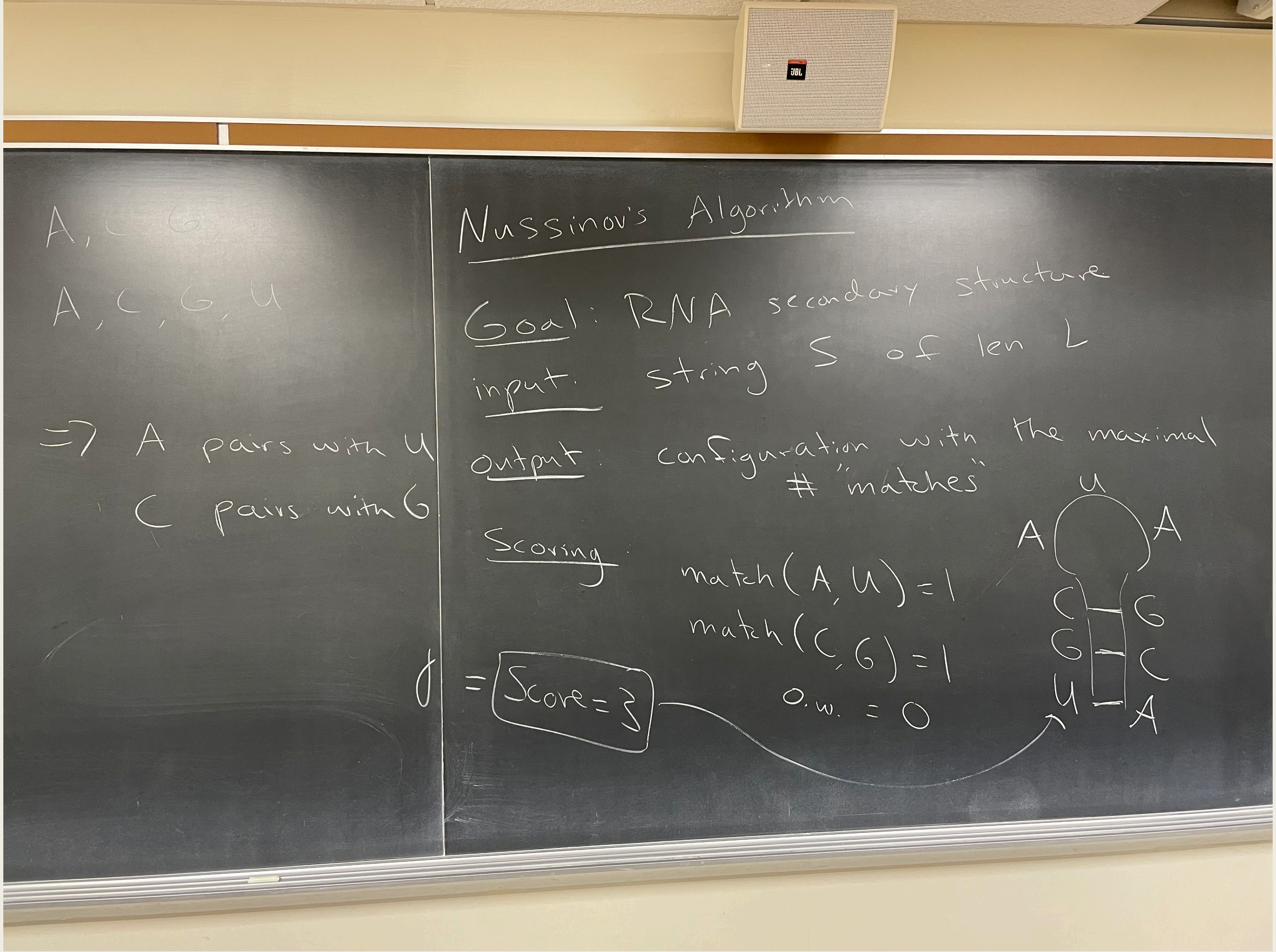
A pairs with U

C pairs with G

Image: wikipedia

Enter: computational biology

- Goal: how could we predict RNA secondary structure?
- Inspiration: sequence alignment
- Answer: dynamic programming (Nussinov’s algorithm)



A, C, G,

A, C, G, U

⇒ A pairs with U

C pairs with G

Nussinov's Algorithm

Goal: RNA secondary structure

input: string S of len L

output: configuration with the maximal # "matches"

Scoring:

match(A, U) = 1

match(C, G) = 1

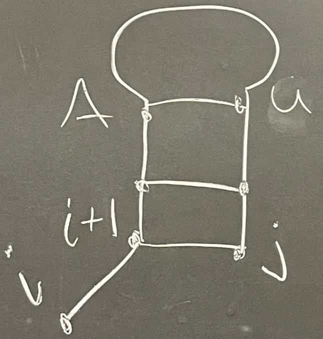
o.w. = 0

= $\boxed{\text{Score} = 3}$

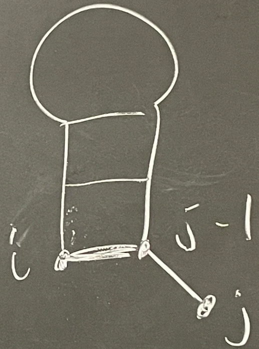
$\gamma(i,j)$ = score for $S[i:j]$ (inclusive)

4 options

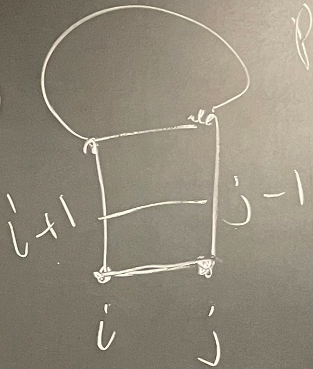
① $i+1, j$ paired
i unpaired



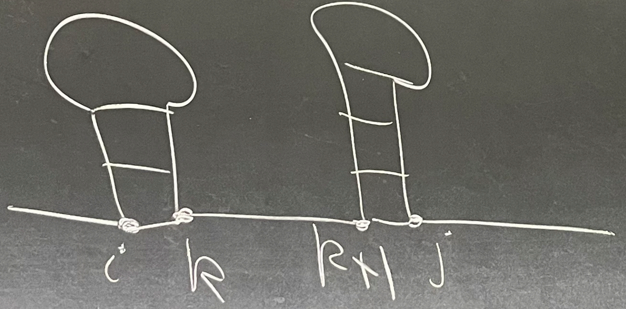
② $i, j-1$ paired
j unpaired

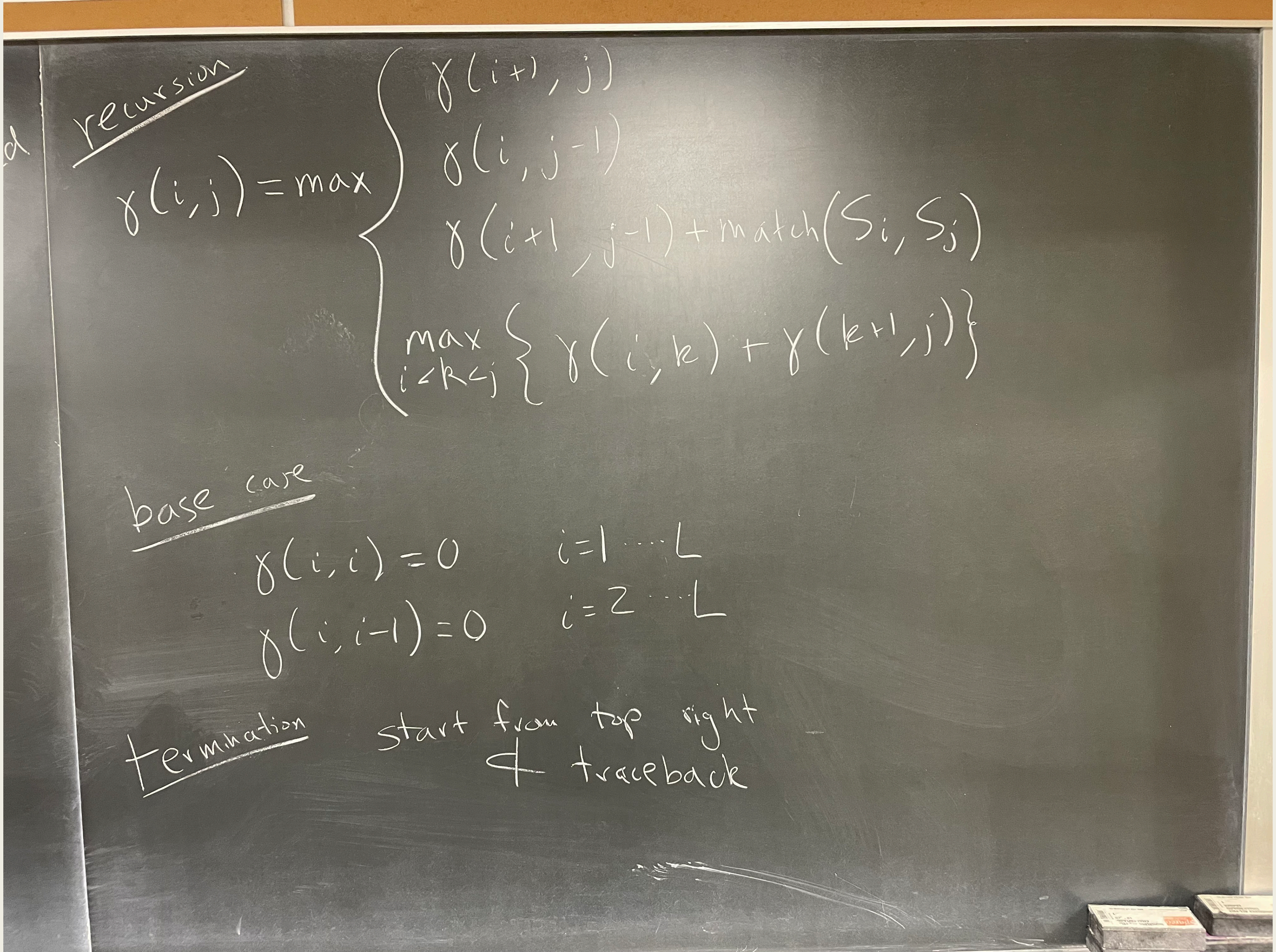


③ i & j paired



④ for some k s.t. $i<k<j$





d

recursion

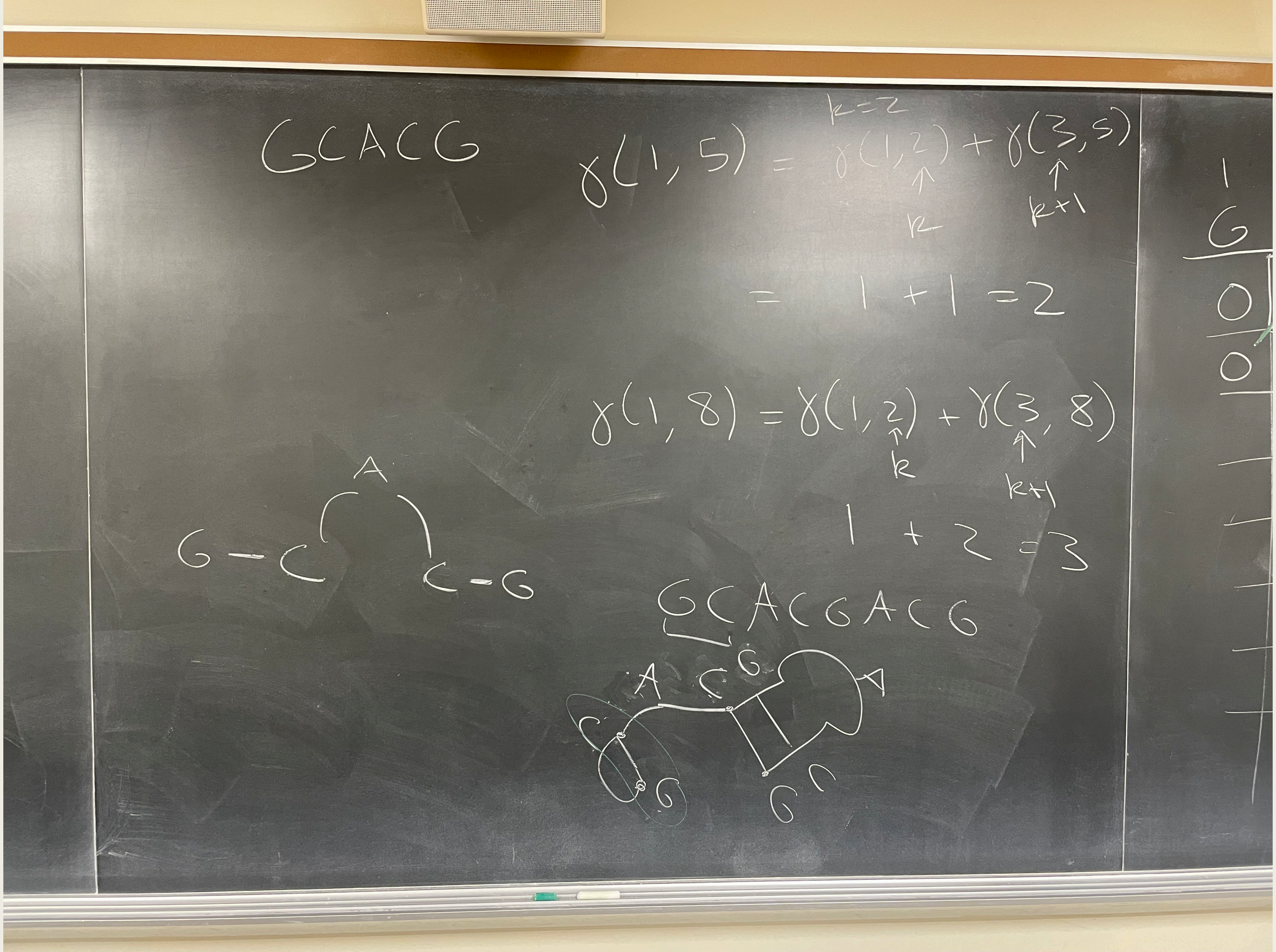
$$\gamma(i,j) = \max \begin{cases} \gamma(i+1, j) \\ \gamma(i, j-1) \\ \gamma(i+1, j-1) + \text{match}(S_i, S_j) \\ \max\limits_{i<k<j} \{ \gamma(i,k) + \gamma(k+1, j) \} \end{cases}$$

base case

$\gamma(i,i) = 0$ $\quad i = 1 \cdots L$

$\gamma(i,i-1) = 0$ $\quad i = 2 \cdots L$

termination start from top right & traceback



GCACG

$k=2$

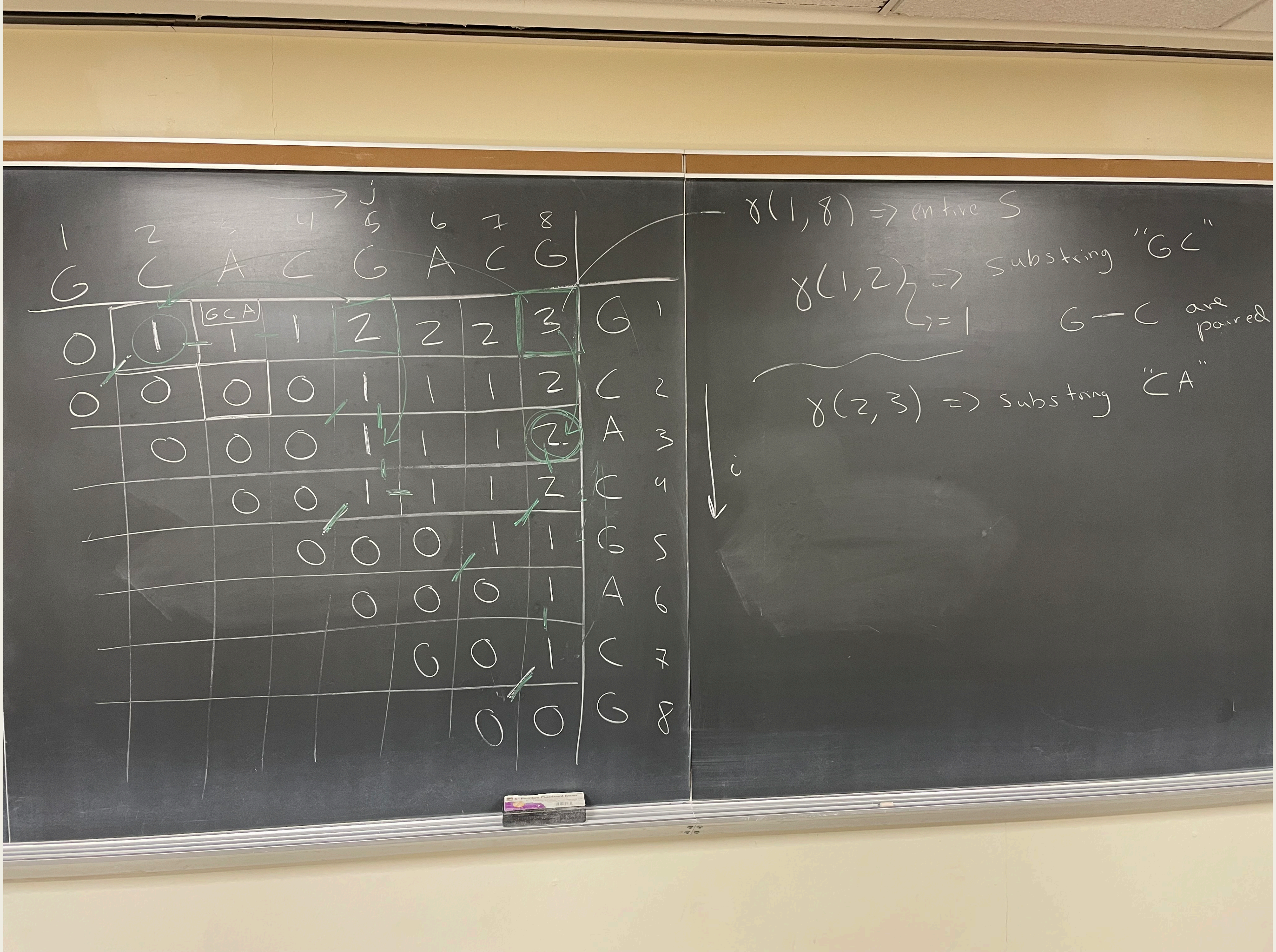
$$\gamma(1,5) = \gamma(1,\underset{\substack{\uparrow\\k}}{2}) + \gamma(\underset{\substack{\uparrow\\k+1}}{3},5)$$

$$= 1 + 1 = 2$$

$$\gamma(1,8) = \gamma(1,\underset{\substack{\uparrow\\k}}{2}) + \gamma(\underset{\substack{\uparrow\\k+1}}{3},8)$$

$$1 + 2 = 3$$

GCACGACG



		1	2	3	4	5	6	7	8
		G	C	A	C	G	A	C	G
G	1	0	1	1	1	2	2	2	3
C	2	0	0	0	0	1	1	1	2
A	3		0	0	0	1	1	1	2
C	4			0	0	1	1	1	2
G	5				0	0	0	1	1
A	6					0	0	0	1
C	7						0	0	1
G	8							0	0

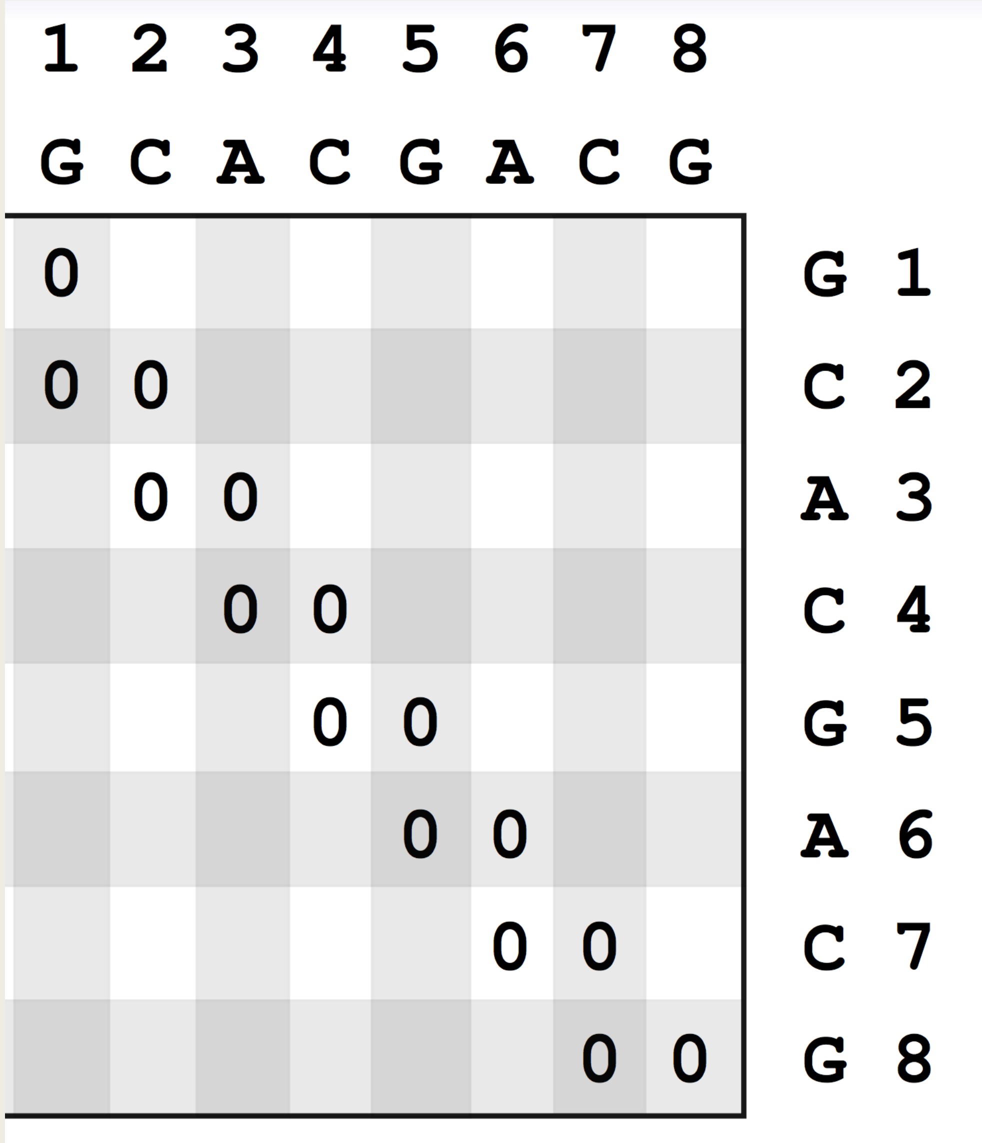
$\gamma(1,8) \Rightarrow$ entire S

$\gamma(1,2) \Rightarrow$ substring "GC"

$\Rightarrow 1$ G—C are paired

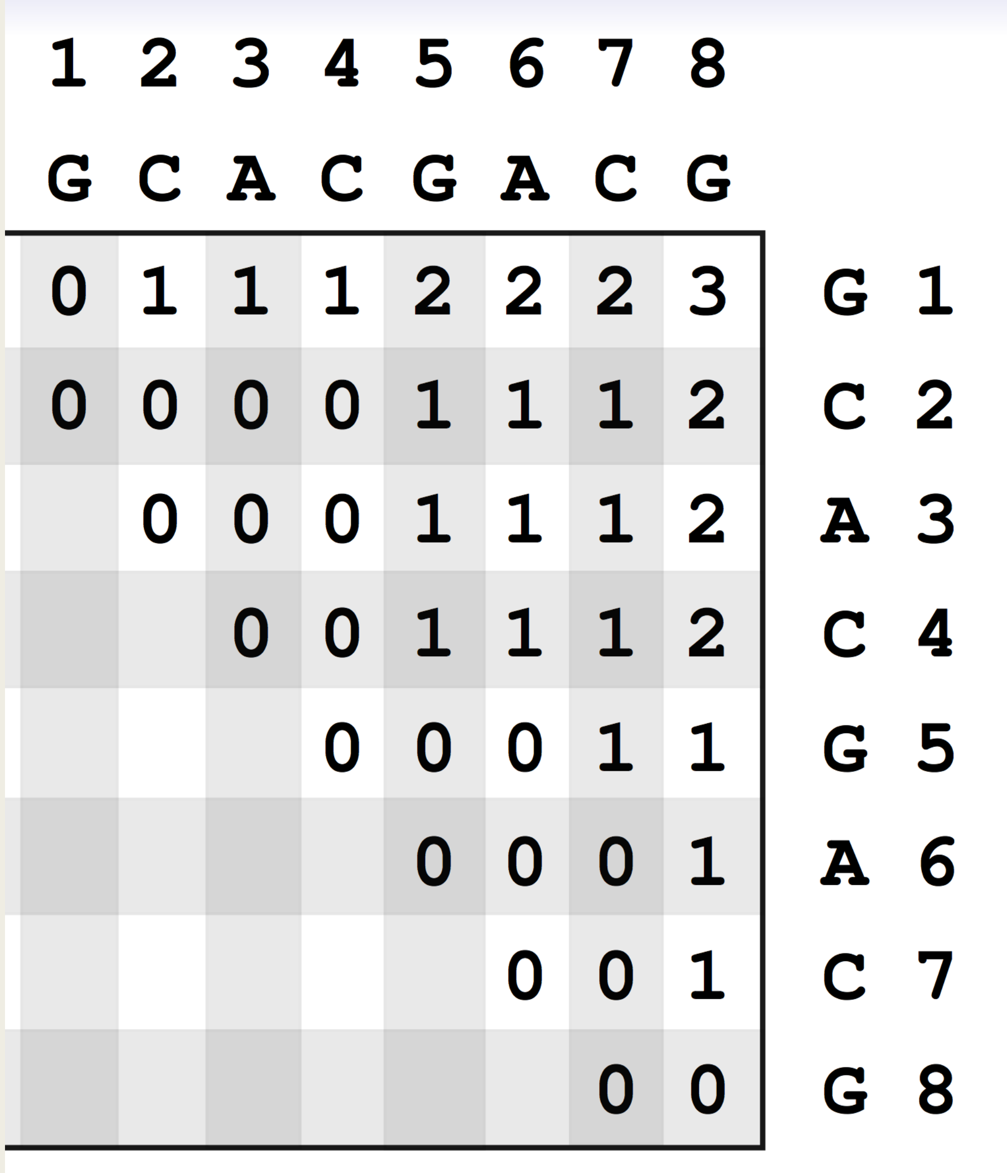
$\gamma(2,3) \Rightarrow$ substring "CA"

Example



1	2	3	4	5	6	7	8		
G	C	A	C	G	A	C	G		
0								G	1
0	0							C	2
	0	0						A	3
		0	0					C	4
			0	0				G	5
				0	0			A	6
					0	0		C	7
						0	0	G	8

Example solution. Exercise: back-tracing

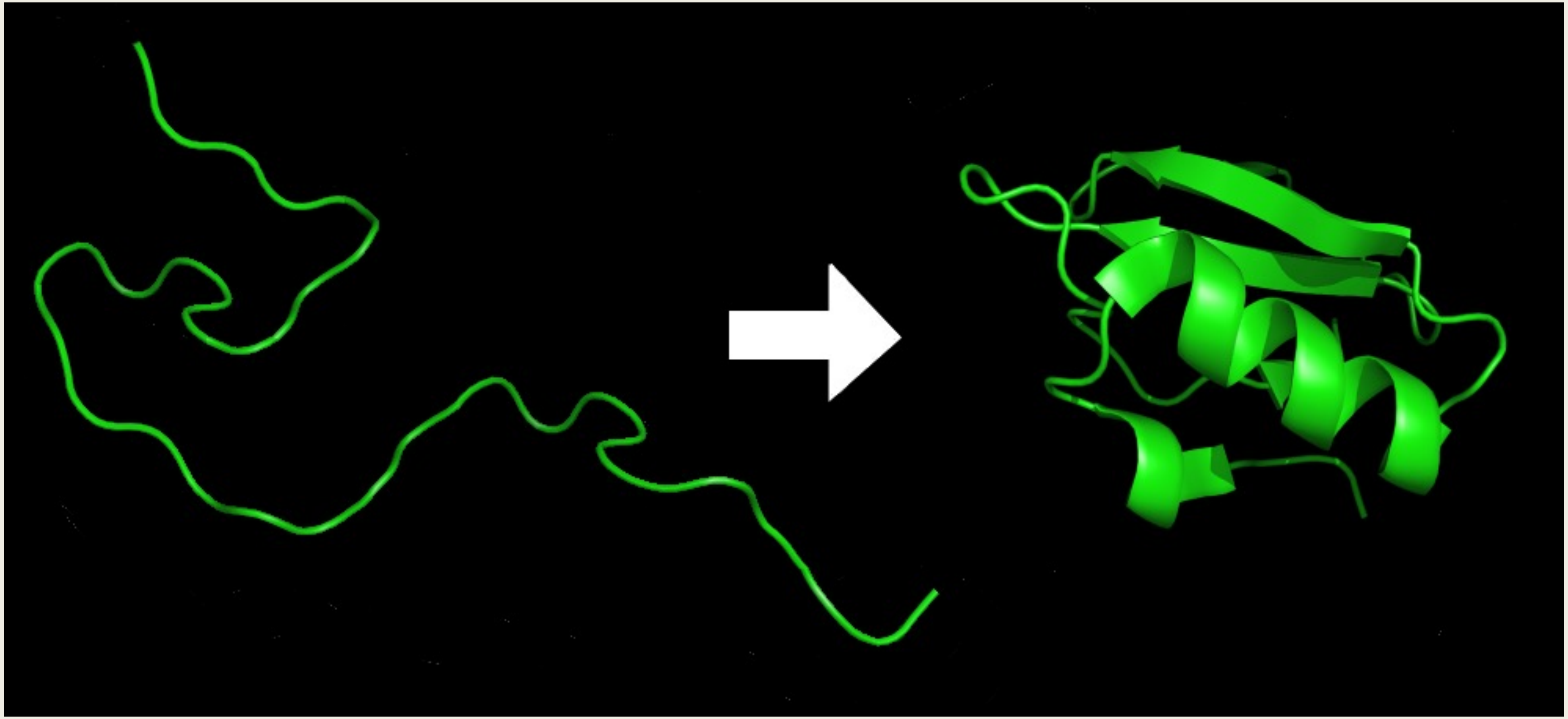


1	2	3	4	5	6	7	8		
G	C	A	C	G	A	C	G		
0	1	1	1	2	2	2	3	G	1
0	0	0	0	1	1	1	2	C	2
	0	0	0	1	1	1	2	A	3
		0	0	1	1	1	2	C	4
			0	0	0	1	1	G	5
				0	0	0	1	A	6
					0	0	1	C	7
						0	0	G	8

S.Will, MIT 18.417, Fall 2011

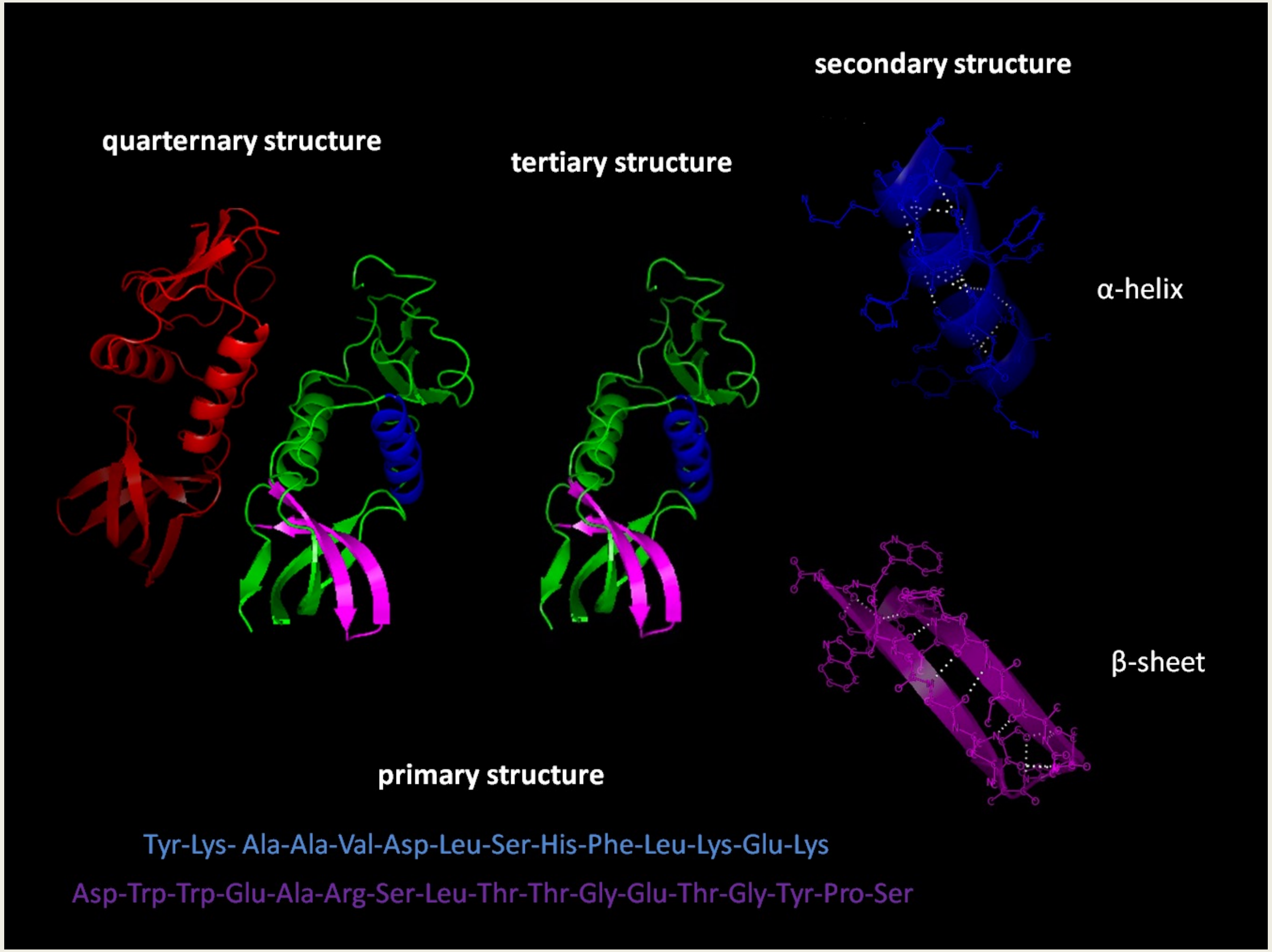
Protein Folding

Protein folding: from sequence to structure



By: DrKjaergaard, Wikipedia

Protein structure beyond the sequence

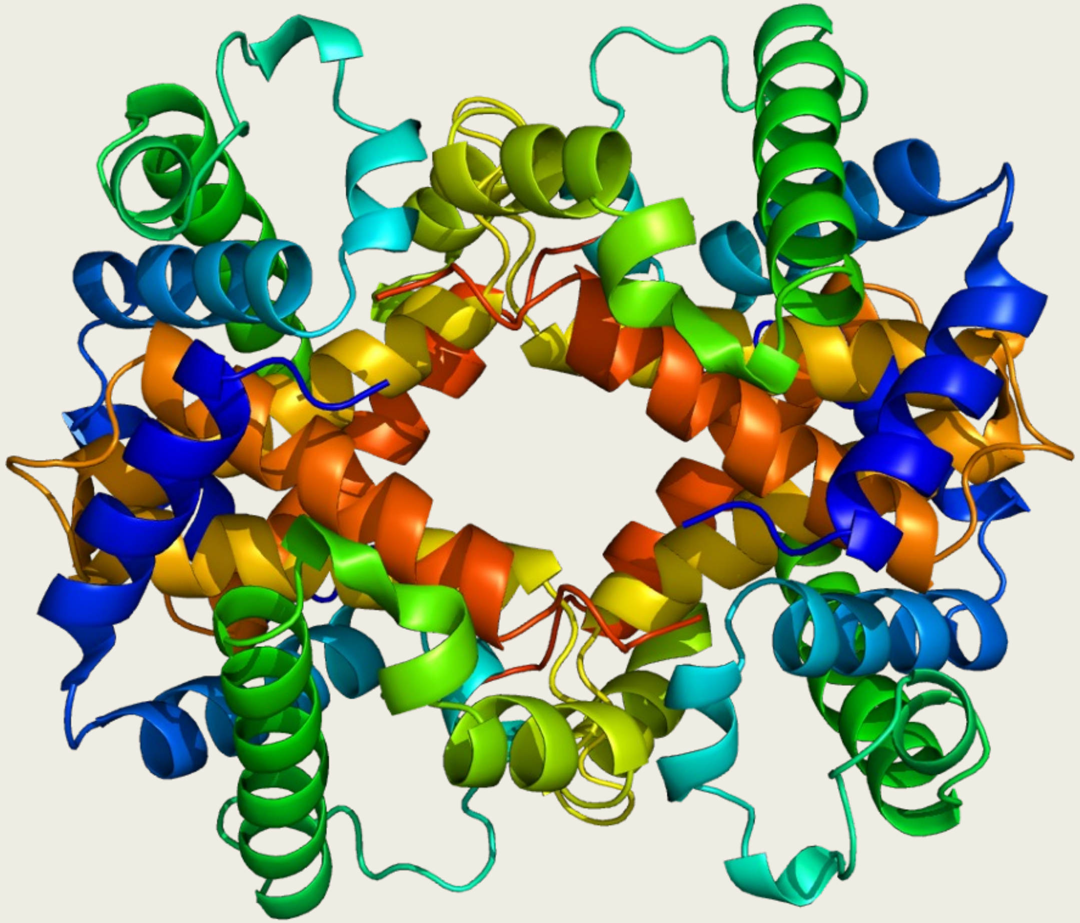

secondary structure
quarternary structure
tertiary structure
α-helix
β-sheet
primary structure
Tyr-Lys- Ala-Ala-Val-Asp-Leu-Ser-His-Phe-Leu-Lys-Glu-Lys
Asp-Trp-Trp-Glu-Ala-Arg-Ser-Leu-Thr-Thr-Gly-Glu-Thr-Gly-Tyr-Pro-Ser

By: Holger87, Wikipedia

Protein structure prediction

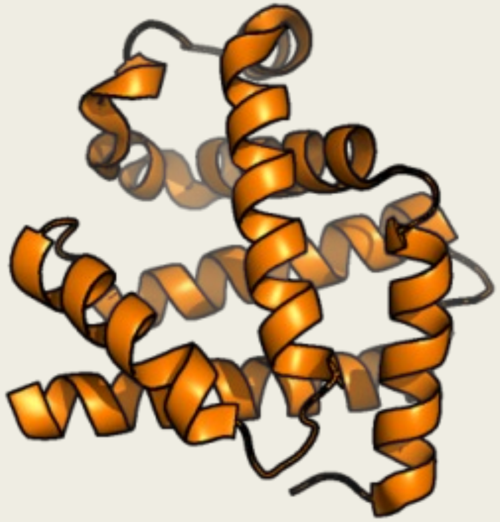
```
MVHLTPEEKSAVTALWGKVNVDEVGGEALG
RLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLS
ELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
```

→ ??? →

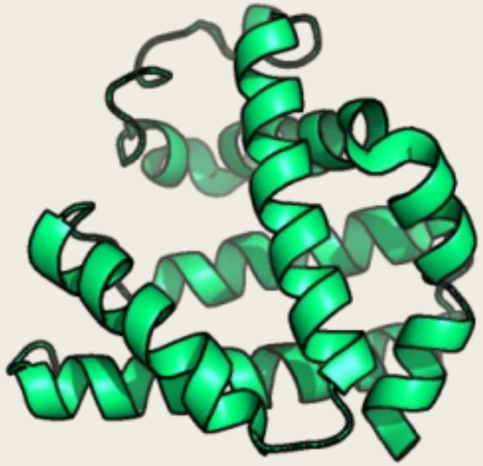


Why do we want to do this?

Human Myoglobin



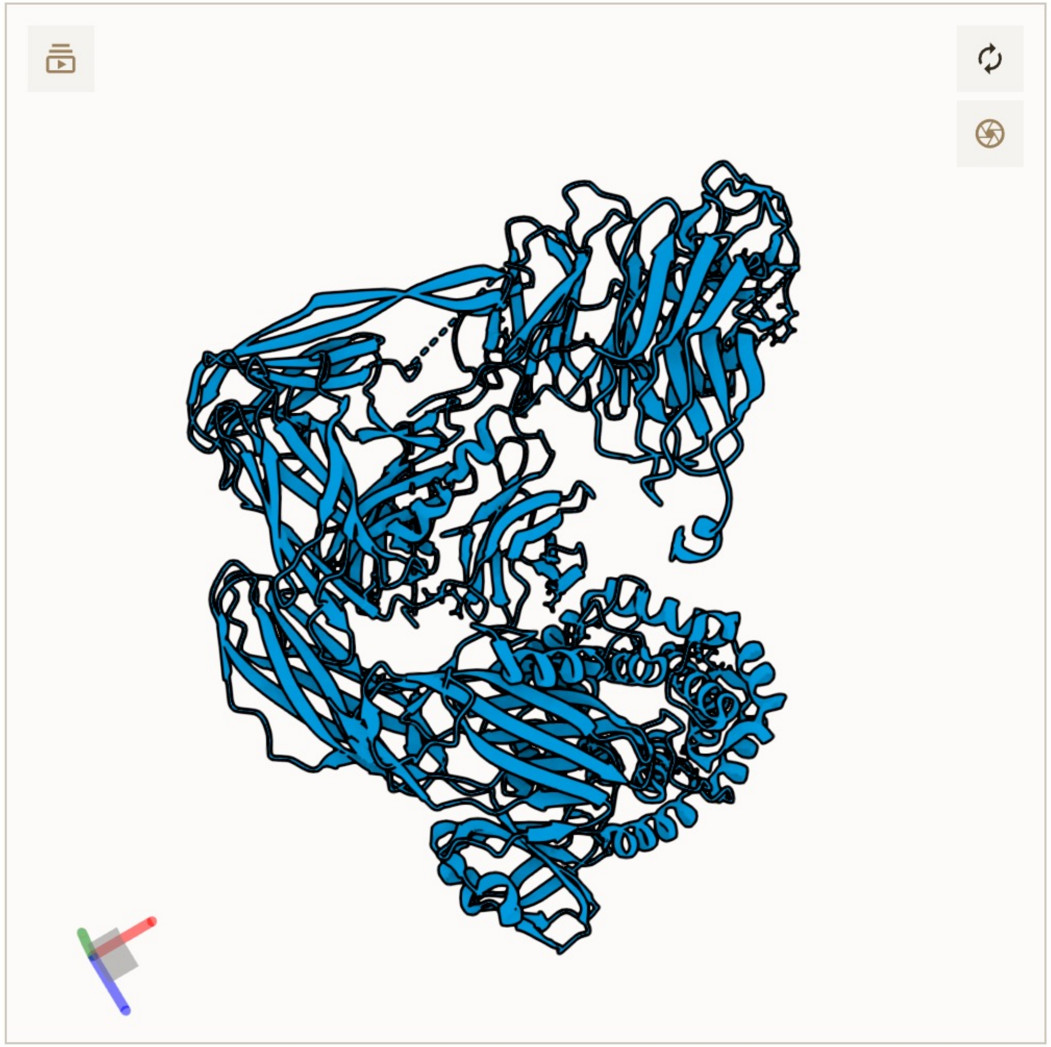
Elephant (80% identity)



Tuna (45% identity)

Pigeon (25% identity)

https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/



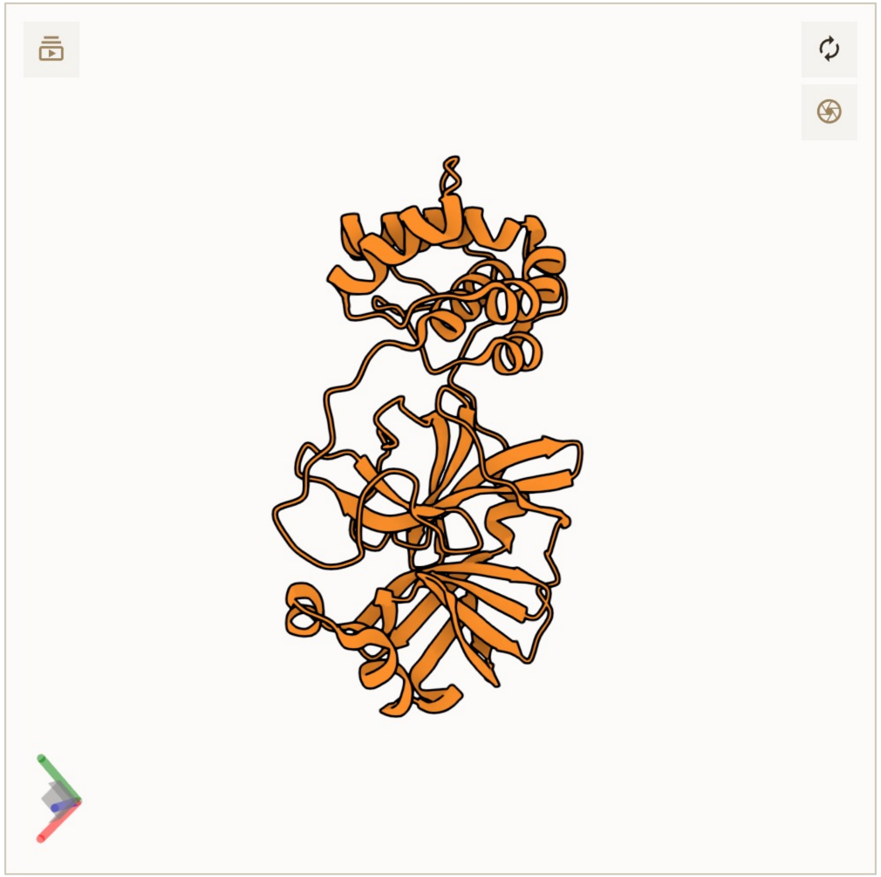
Salmonella alpha-2-macroglobulin (PDB ID: 4U48)

Provides protection against proteases secreted from hosts

HIV-1 Reverse Transcriptase (PDB ID: 8U6H)

One of the key proteins responsible for the replication cycle of HIV-1 in the host.

https://www.ebi.ac.uk/training/online/courses/alphafold/



SARS-CoV main protease (PDB ID: 6Y7M)

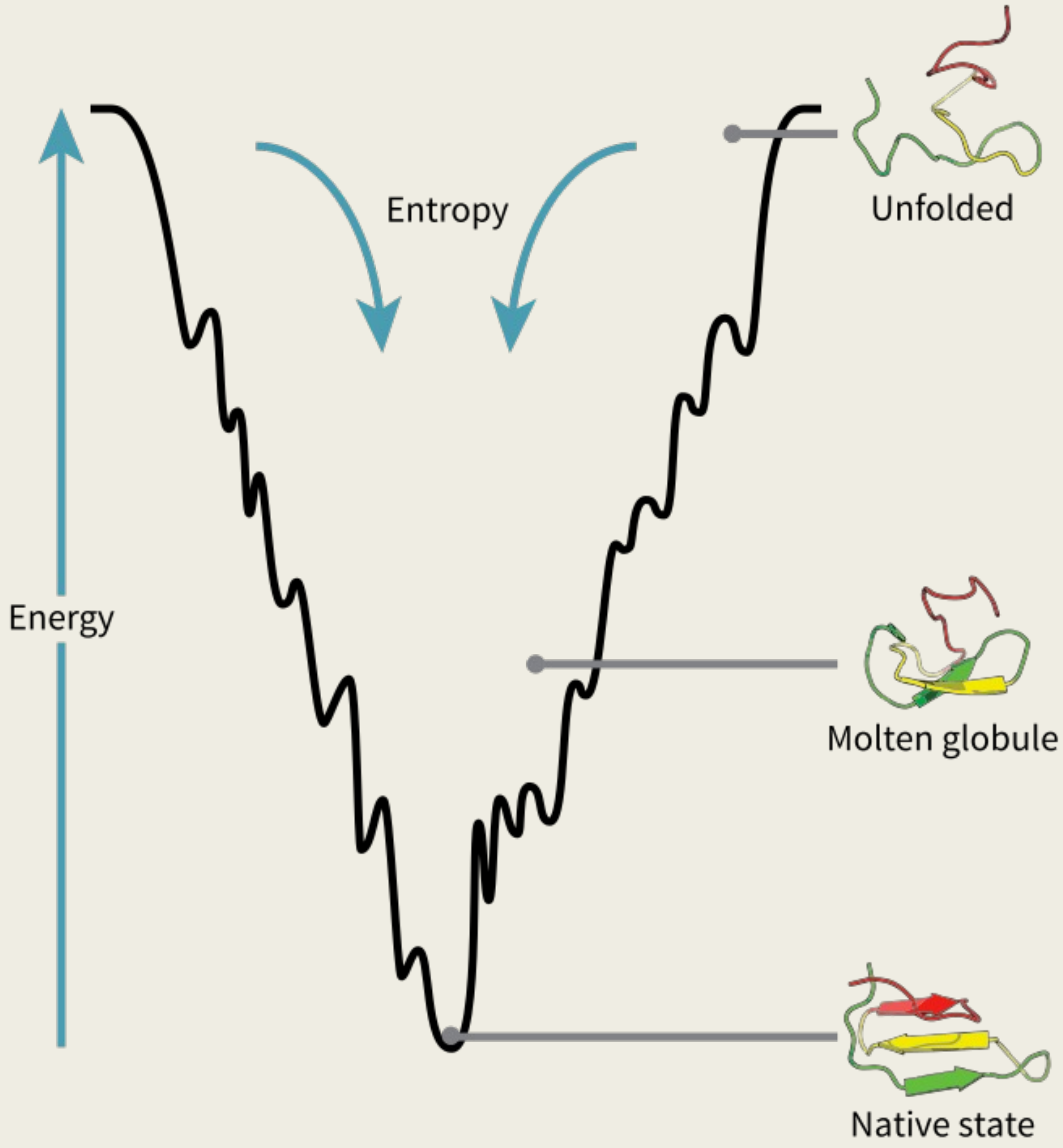
This essential coronavirus protease processes the polyproteins translated from the viral RNA

Circadian Clock Protein KaiA (PDB ID: 1R8J)

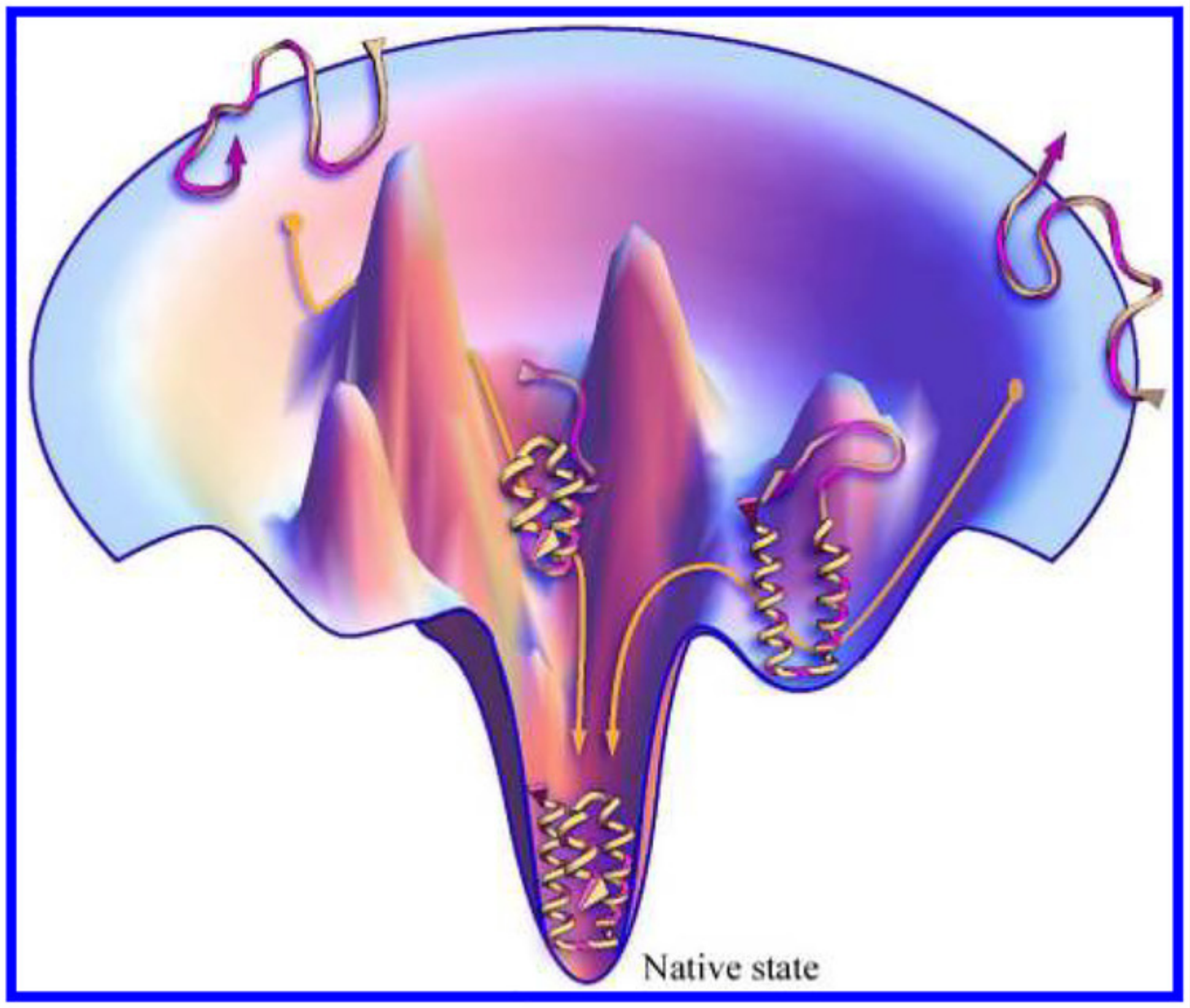
A tiny clockmaker that orchestrates the daily rhythms of cyanobacteria. KaiA regulates other proteins in a 24-hour cycle, like a metronome for time. To keep the rhythm on track, it senses light and other cues

https://www.ebi.ac.uk/training/online/courses/alphafold/

Proteins seek a low-energy configuration



By: Thomas Splettstoesser, Wikipedia

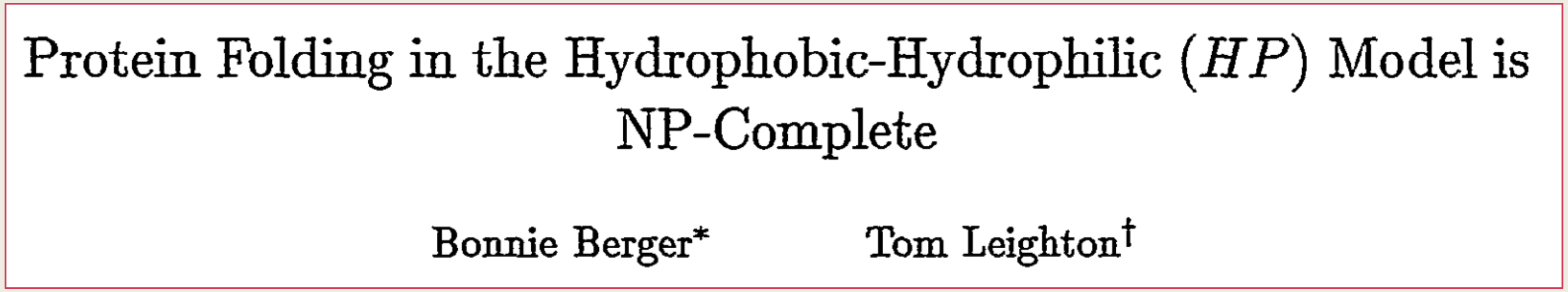


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6407873/

Breakthrough in protein folding

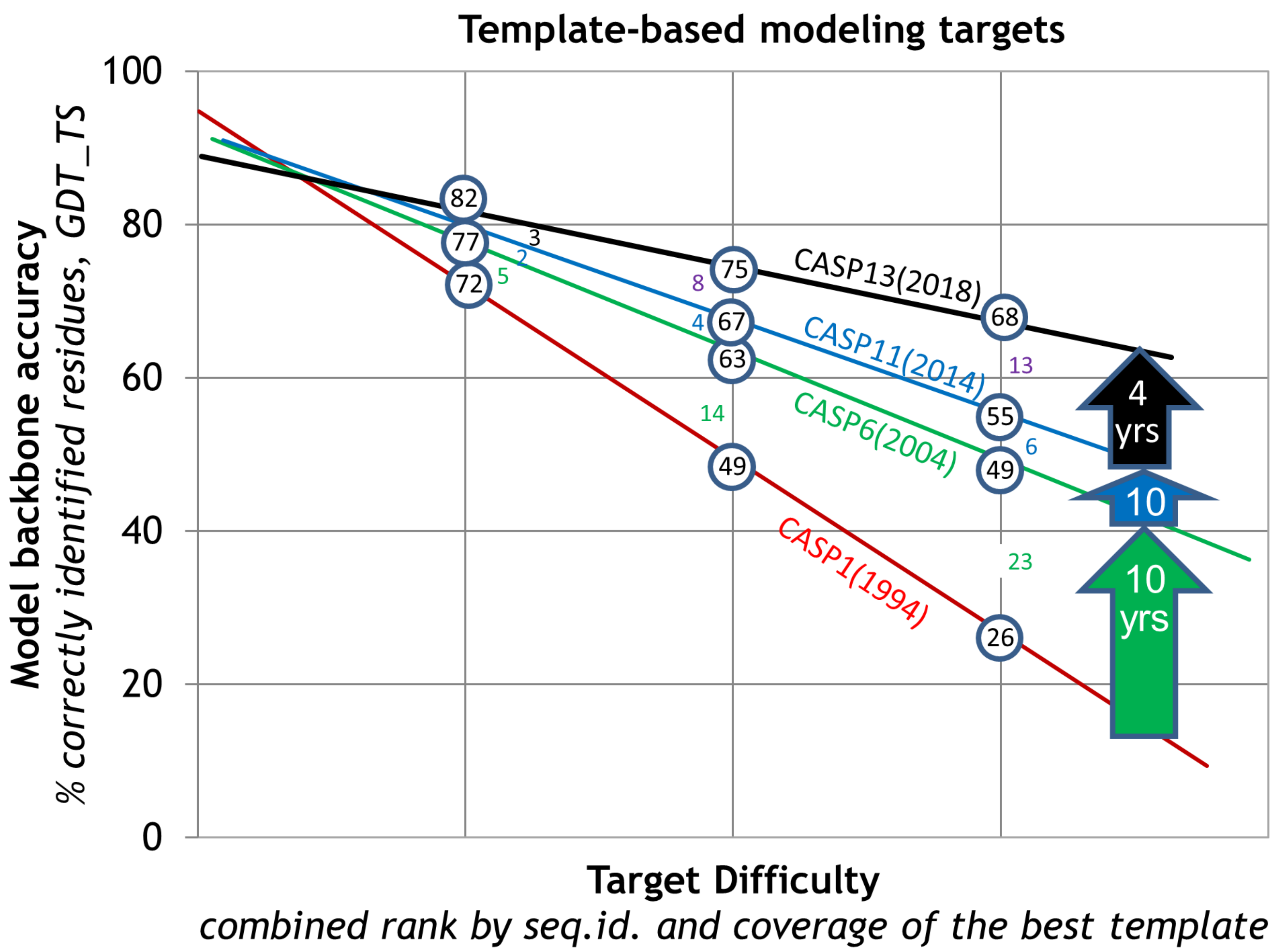
- Bonnie Berger and Tom Leighton prove protein folding is NP-Complete (1998)

- Helped pave the way for approximation algorithms



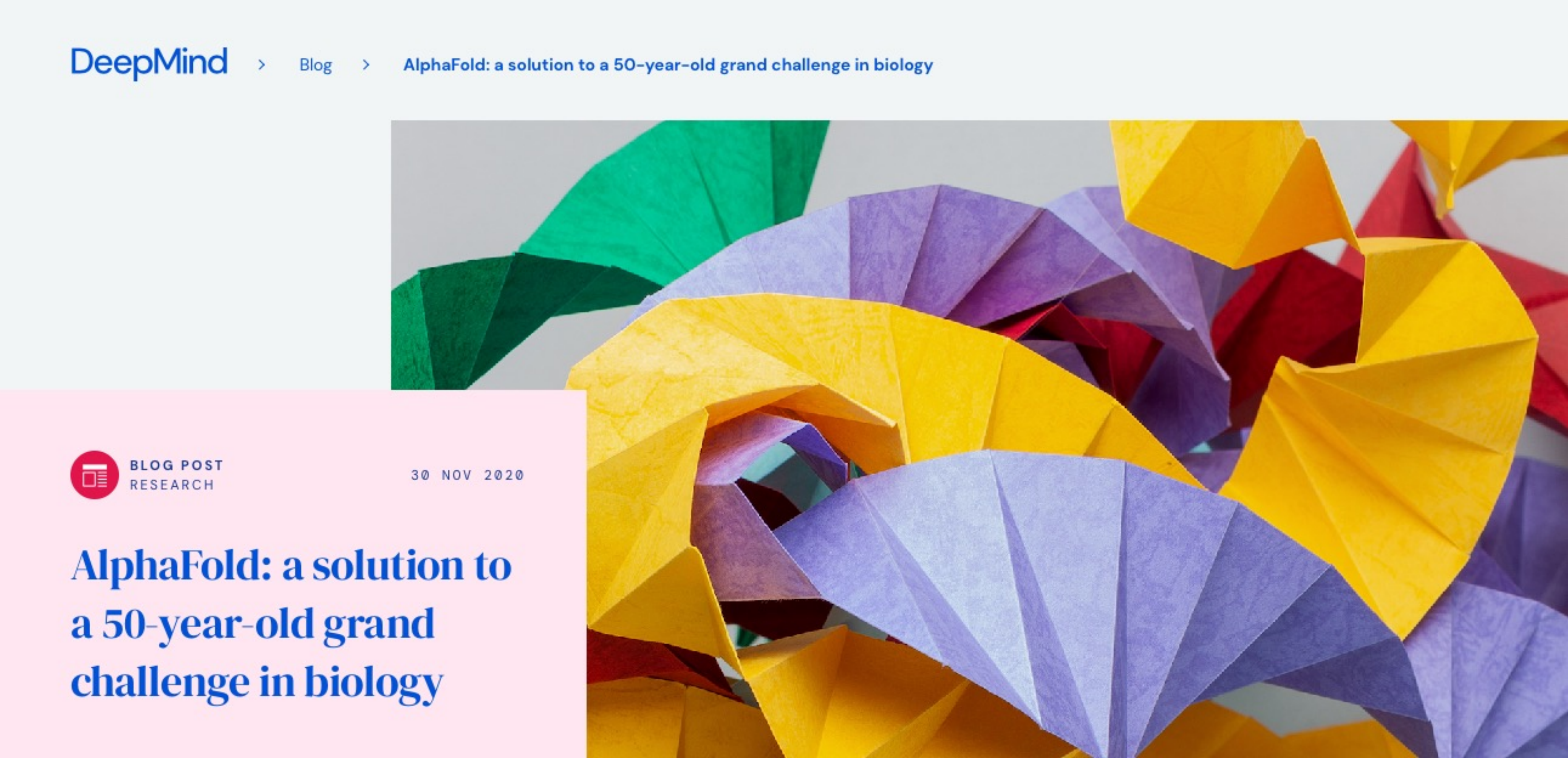
Protein Folding in the Hydrophobic-Hydrophilic (HP) Model is NP-Complete

Bonnie Berger[*] Tom Leighton[†]



Template-based modeling targets

Alphafold2 and protein structure prediction



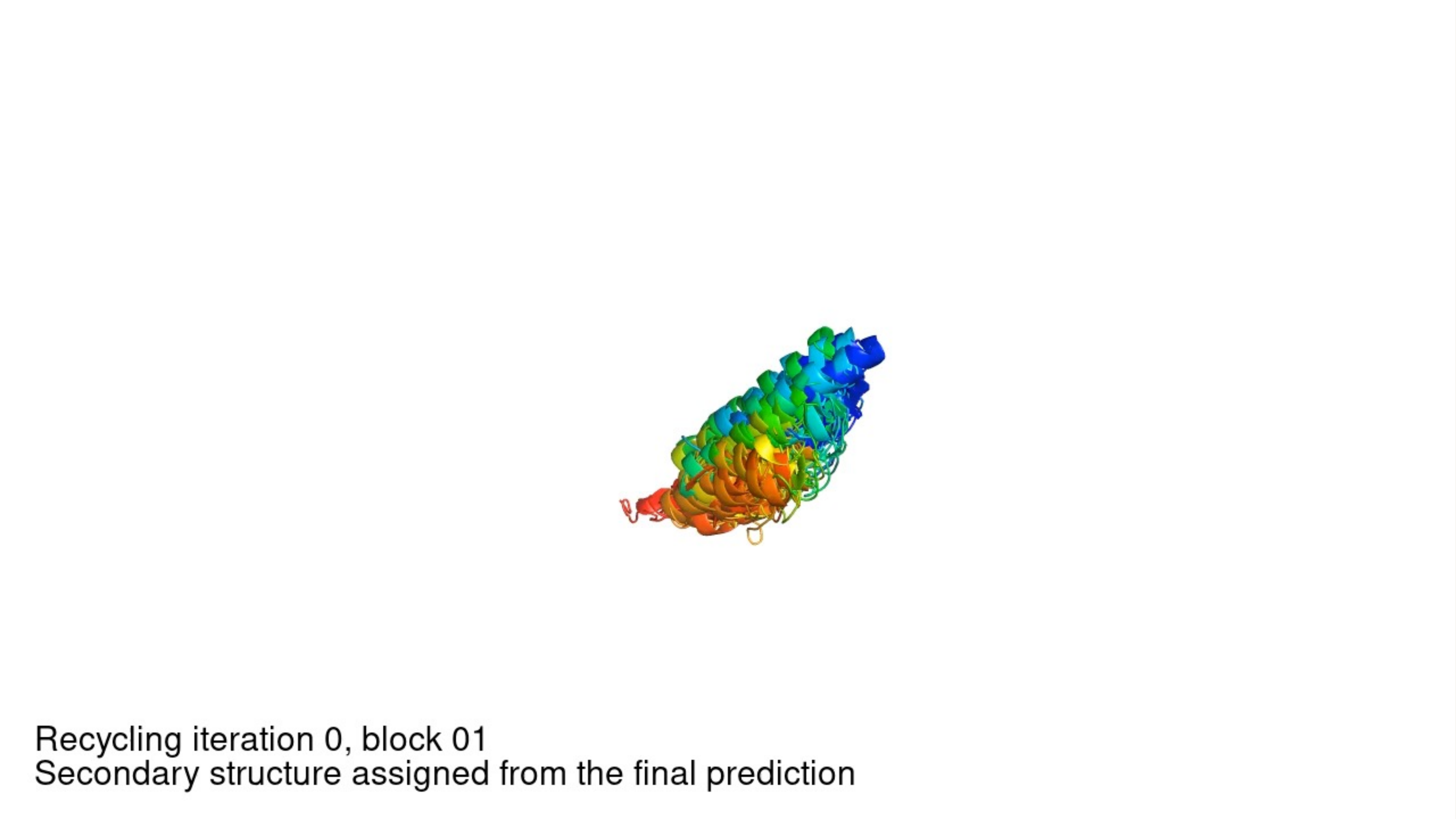
DeepMind › Blog › AlphaFold: a solution to a 50-year-old grand challenge in biology

BLOG POST
RESEARCH

30 NOV 2020

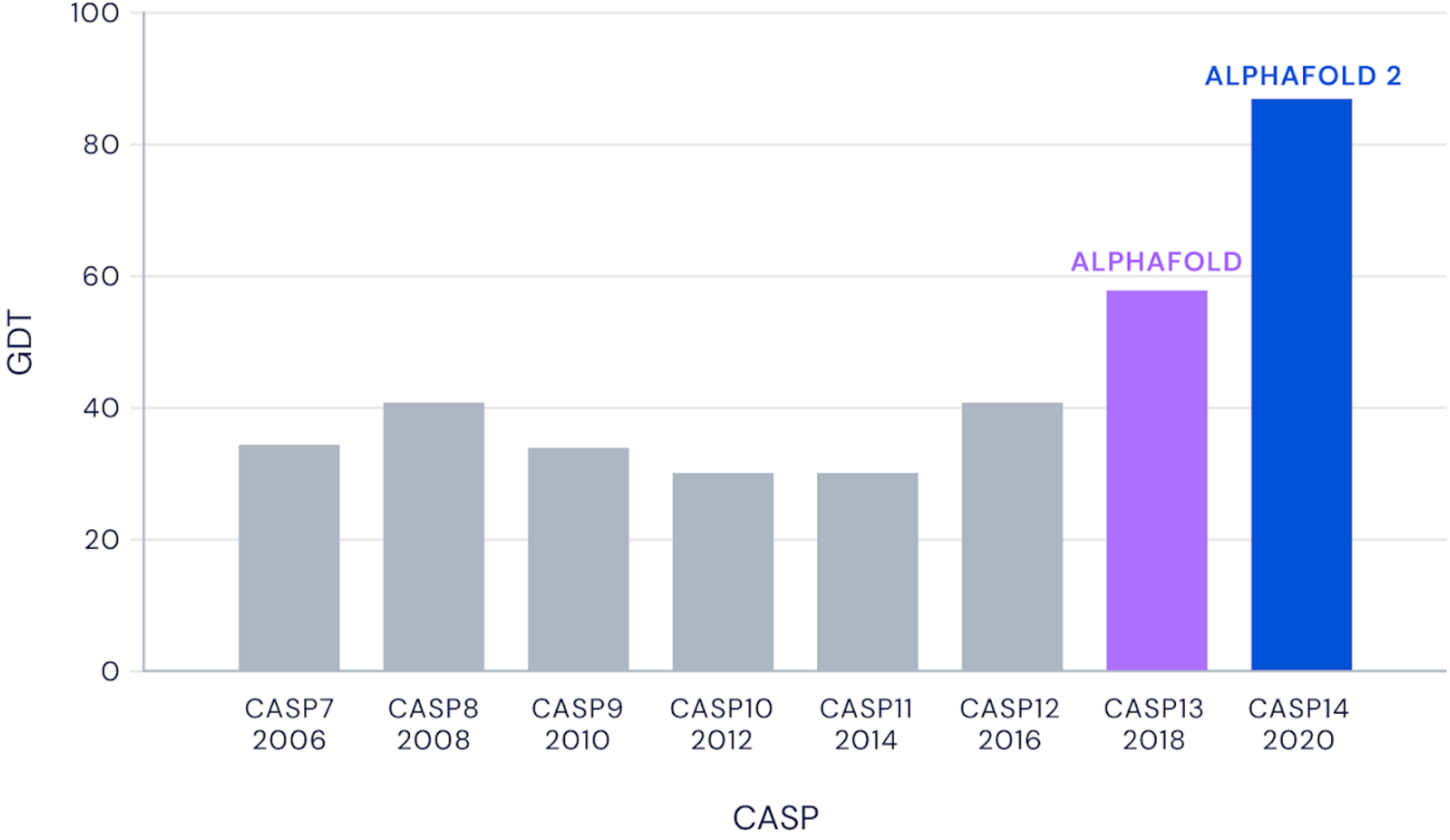
AlphaFold: a solution to a 50-year-old grand challenge in biology

Figure 2. Unlocking protein structures. Three experimental methods used for determining protein structure: X-ray, NMR and cryo-EM. AlphaFold2, a powerful AI-driven method, has revolutionised the field by predicting protein structures with remarkable accuracy. Source: What Is AlphaFold? | NEJM , “A Holy Grail — The Prediction of Protein Structure” (Altman, 2023)

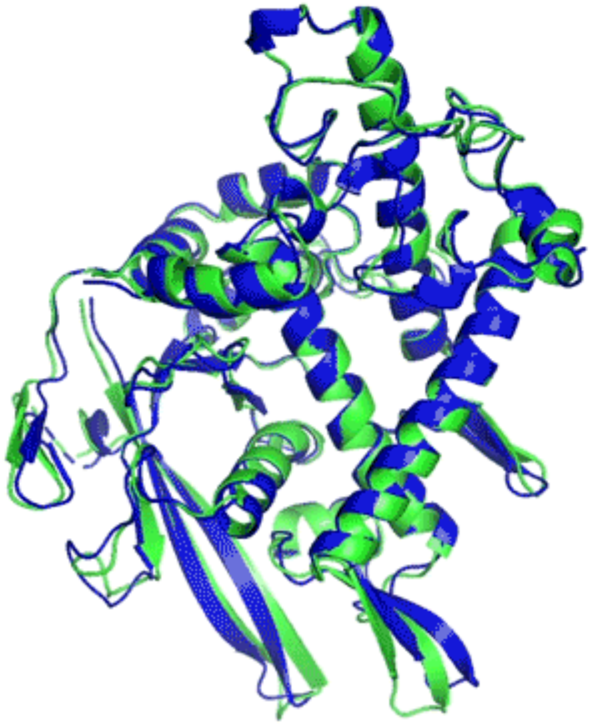

Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

This video presents the intermediate structure trajectory of the CASP14 target T1044, a large (2180 residues) and multi-domain RNA polymerase, predicted by AlphaFold2. Observe the differential folding rates of individual domains, with some folding quickly and others requiring more time. Watch the AlphaFold's prediction process, as it recycles its predictions to refine the final structure (Jumper et al., 2021).

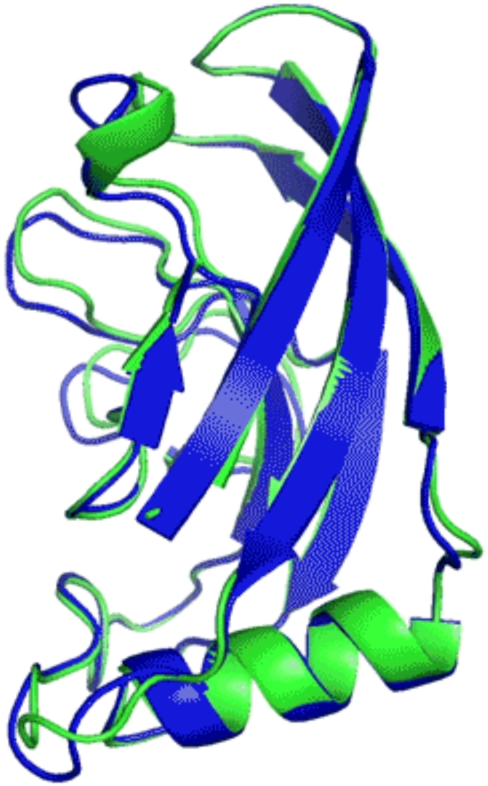
Median Free-Modelling Accuracy



DeepMind



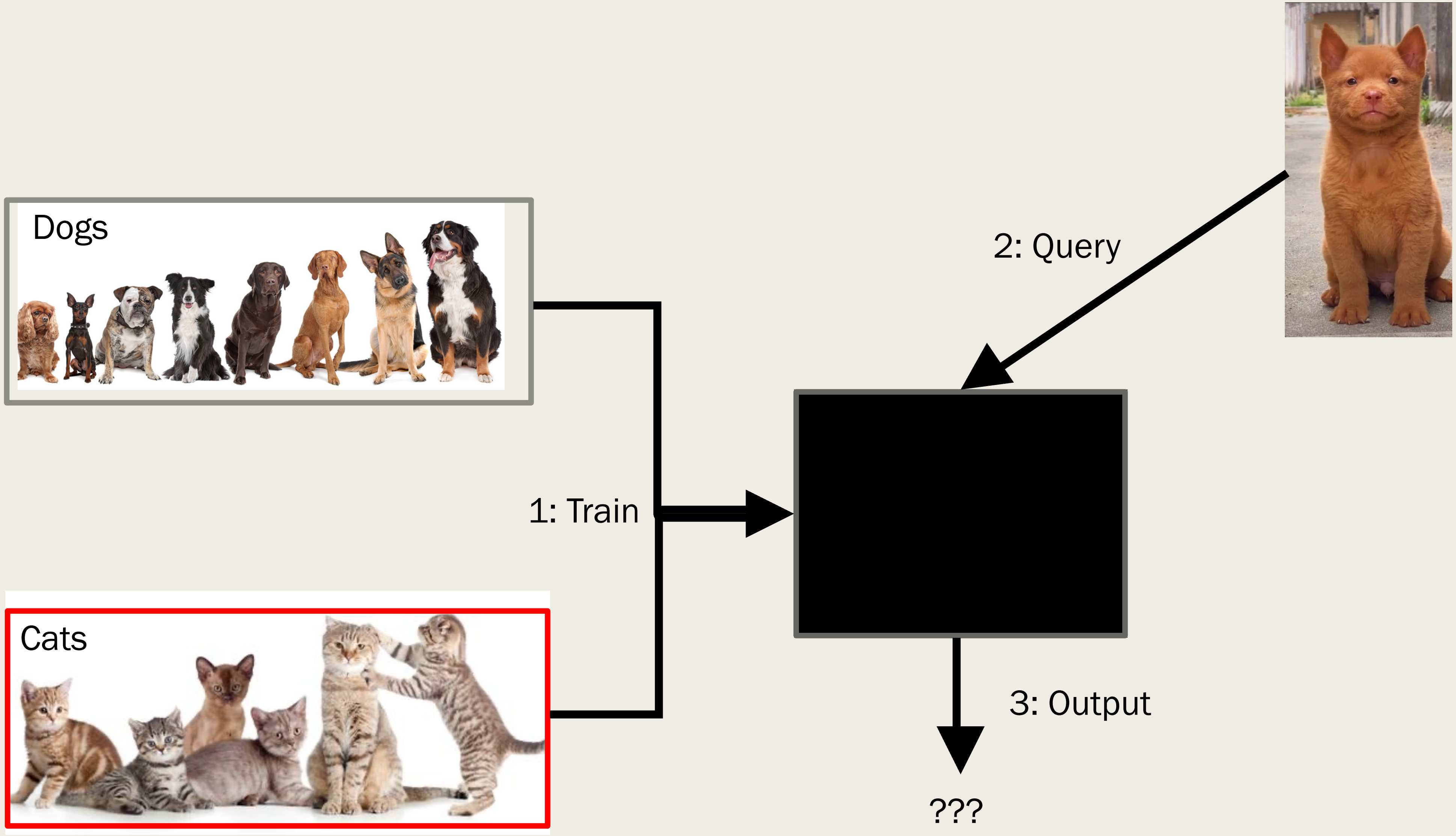
T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



T1049 / 6y4f
93.3 GDT
(adhesin tip)

🟢 Experimental result

🔵 Computational prediction



Dogs

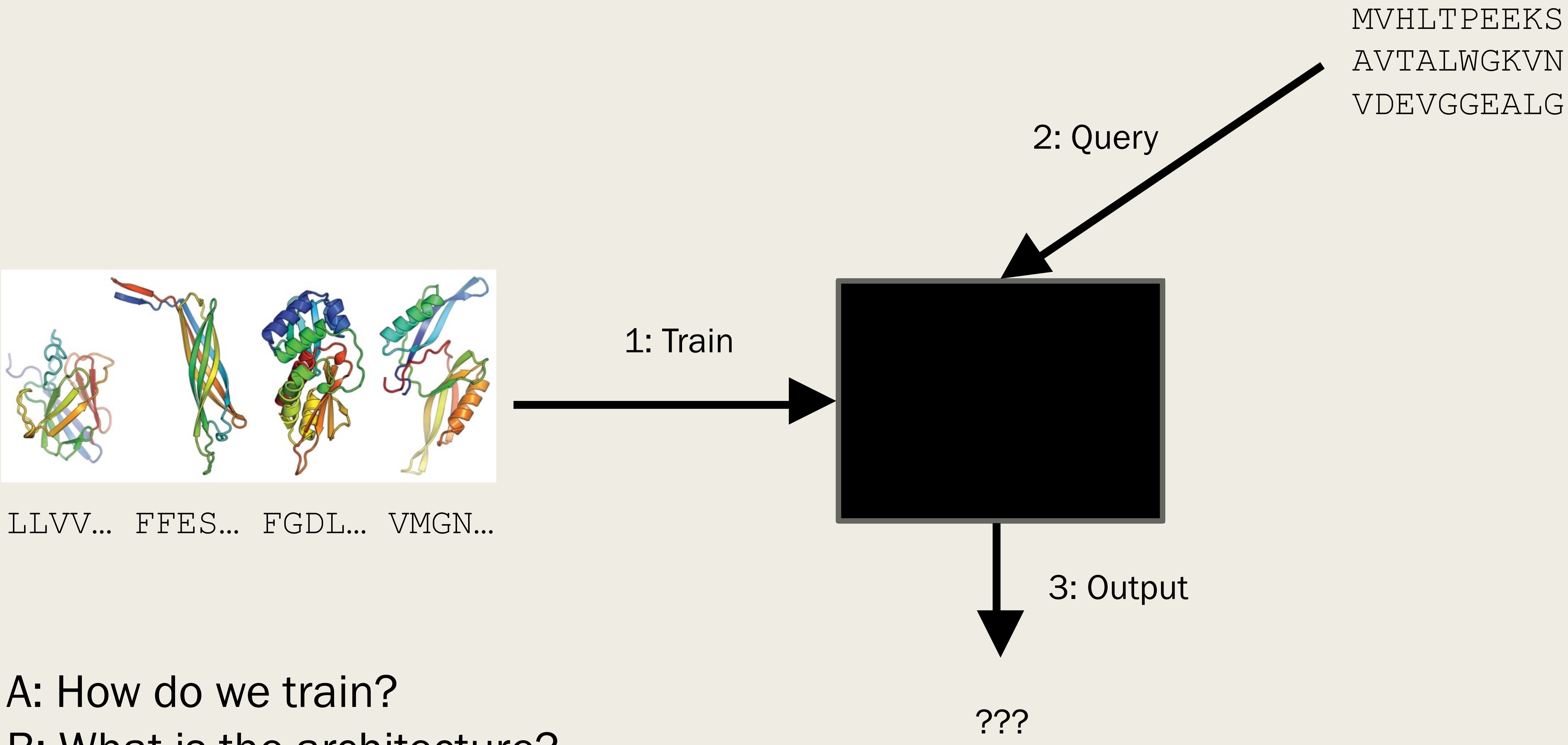
Cats

1: Train

2: Query

3: Output

???



MVHLTPEEKS
AVTALWGKVN
VDEVGGEALG

2: Query

LLVV... FFES... FGDL... VMGN...

1: Train

3: Output

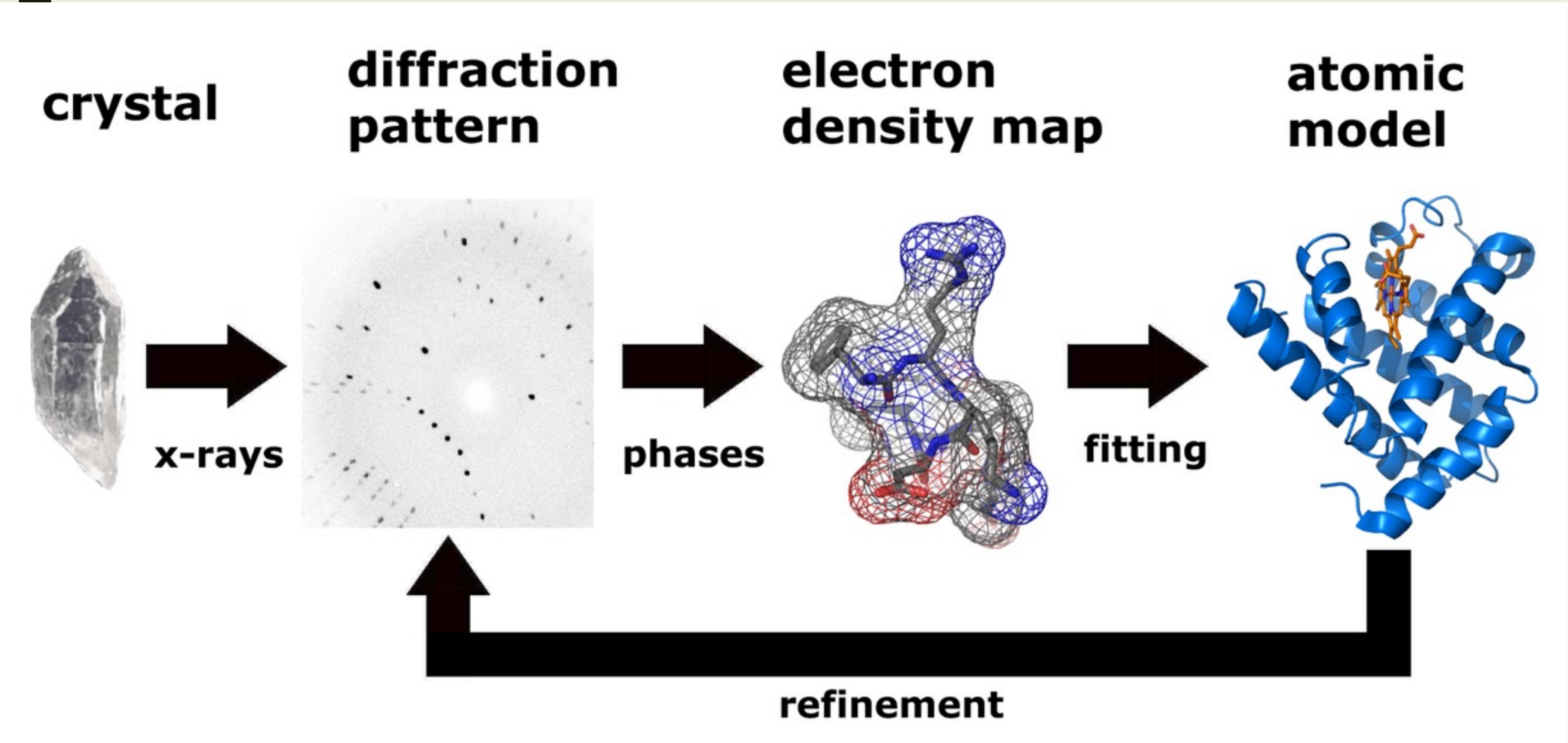
???

A: How do we train?

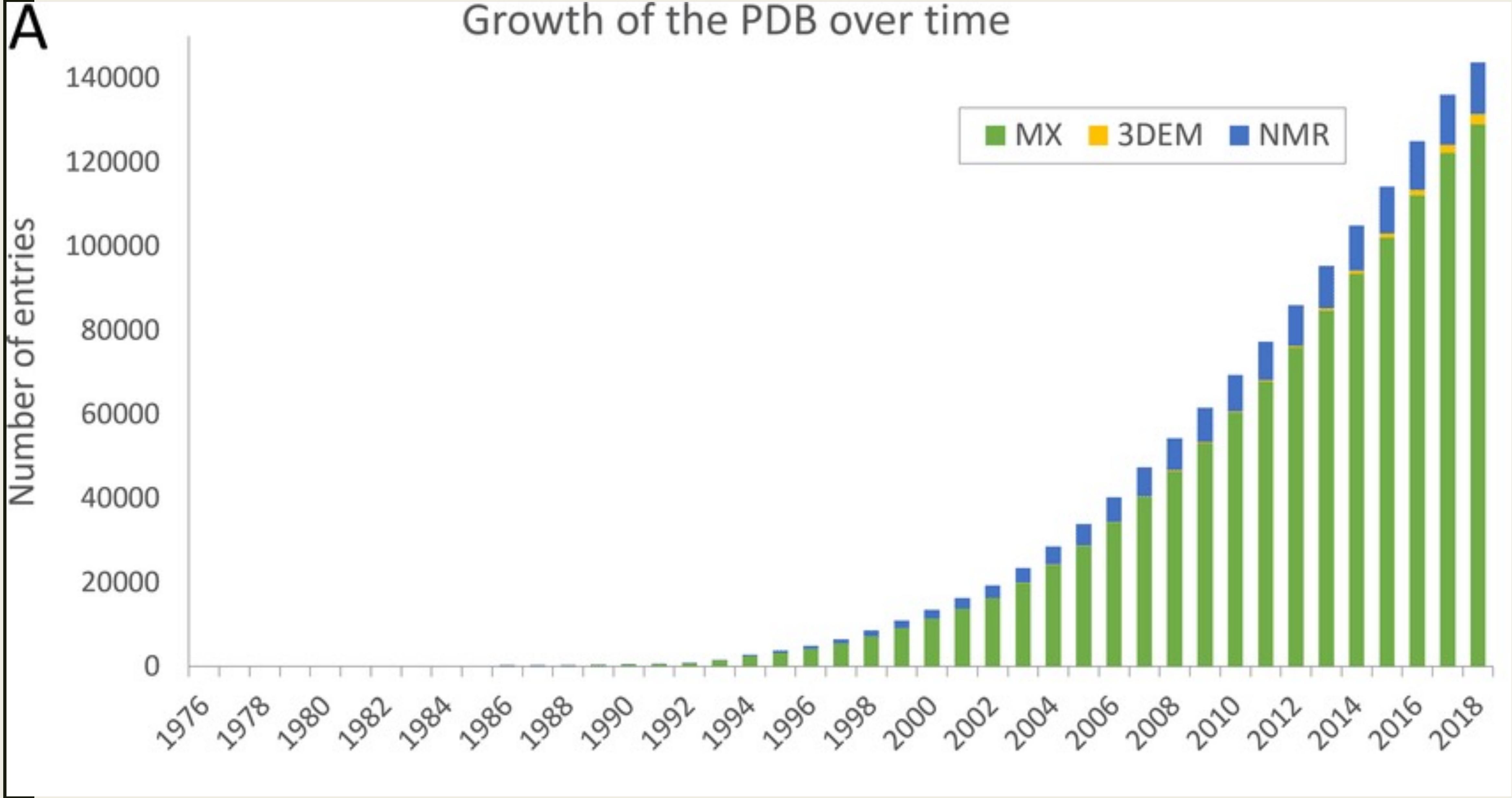
B: What is the architecture?

C: What are the training data?

What are the training data?



https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/structure-factors-and-electron-density

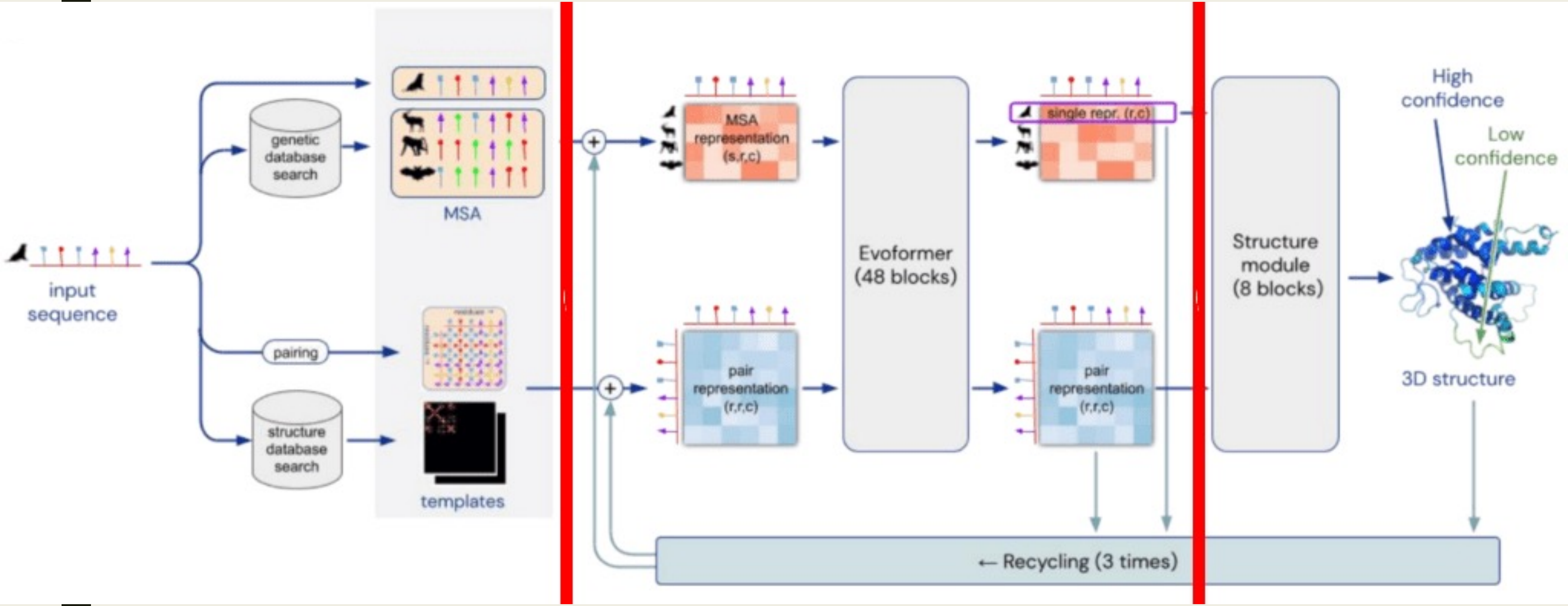


https://www.rcsb.org

Search and alignment

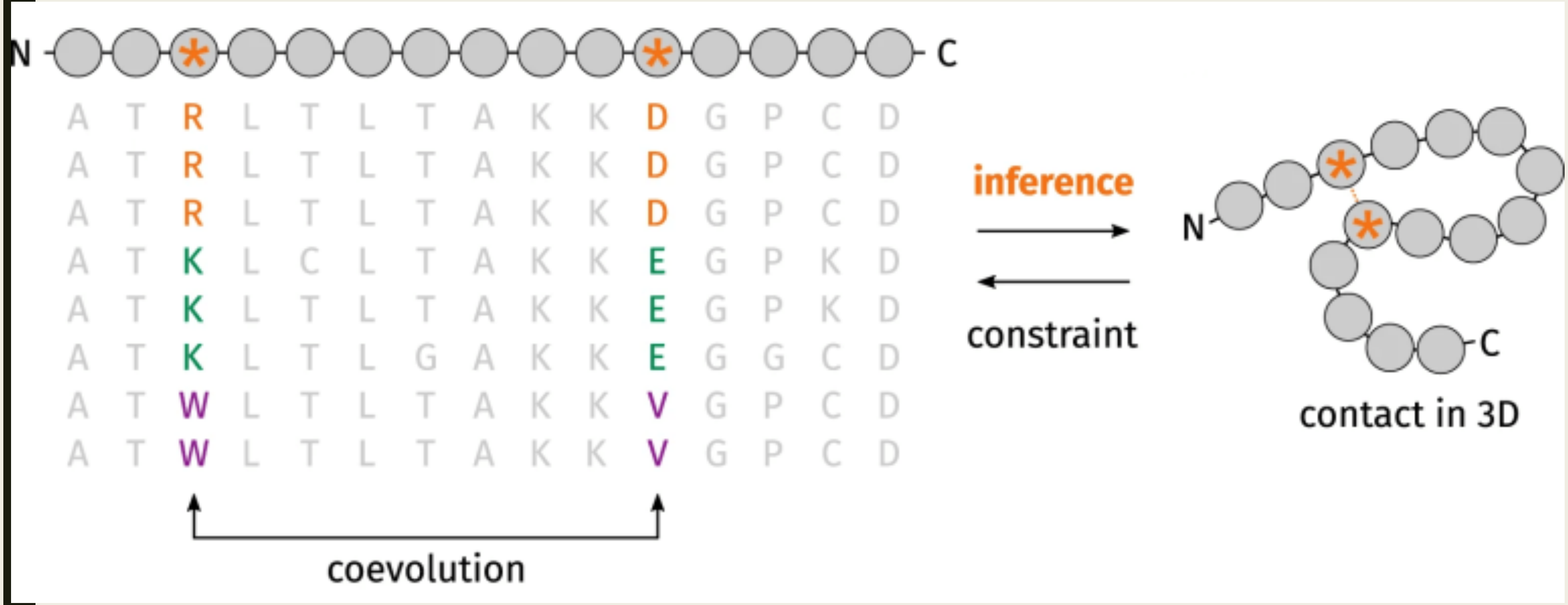
Encoding

Prediction



Jumper et al. modified by Carlos Outeiral Rubiera

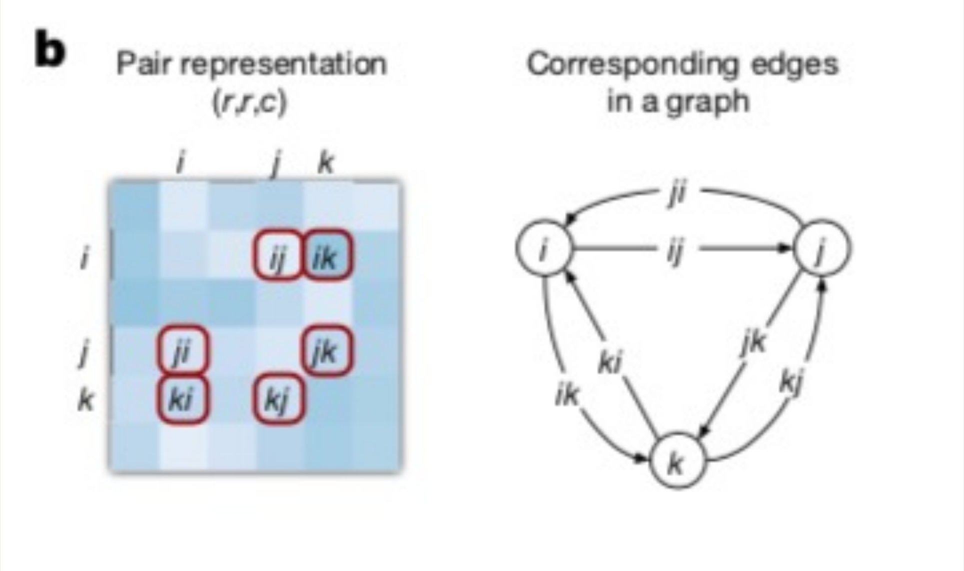
Transformers

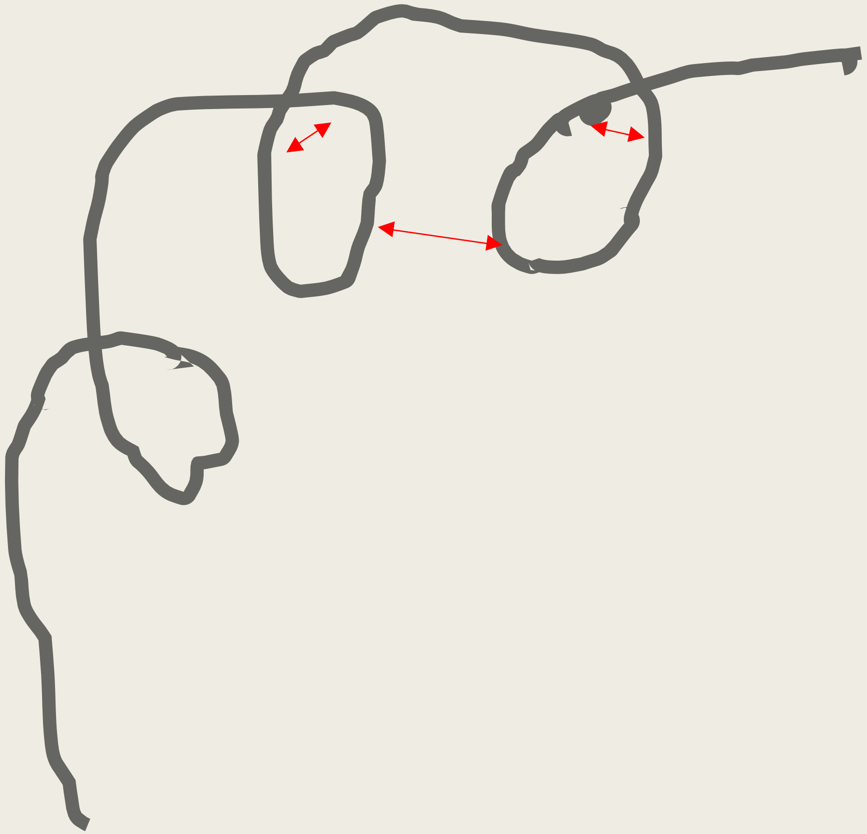


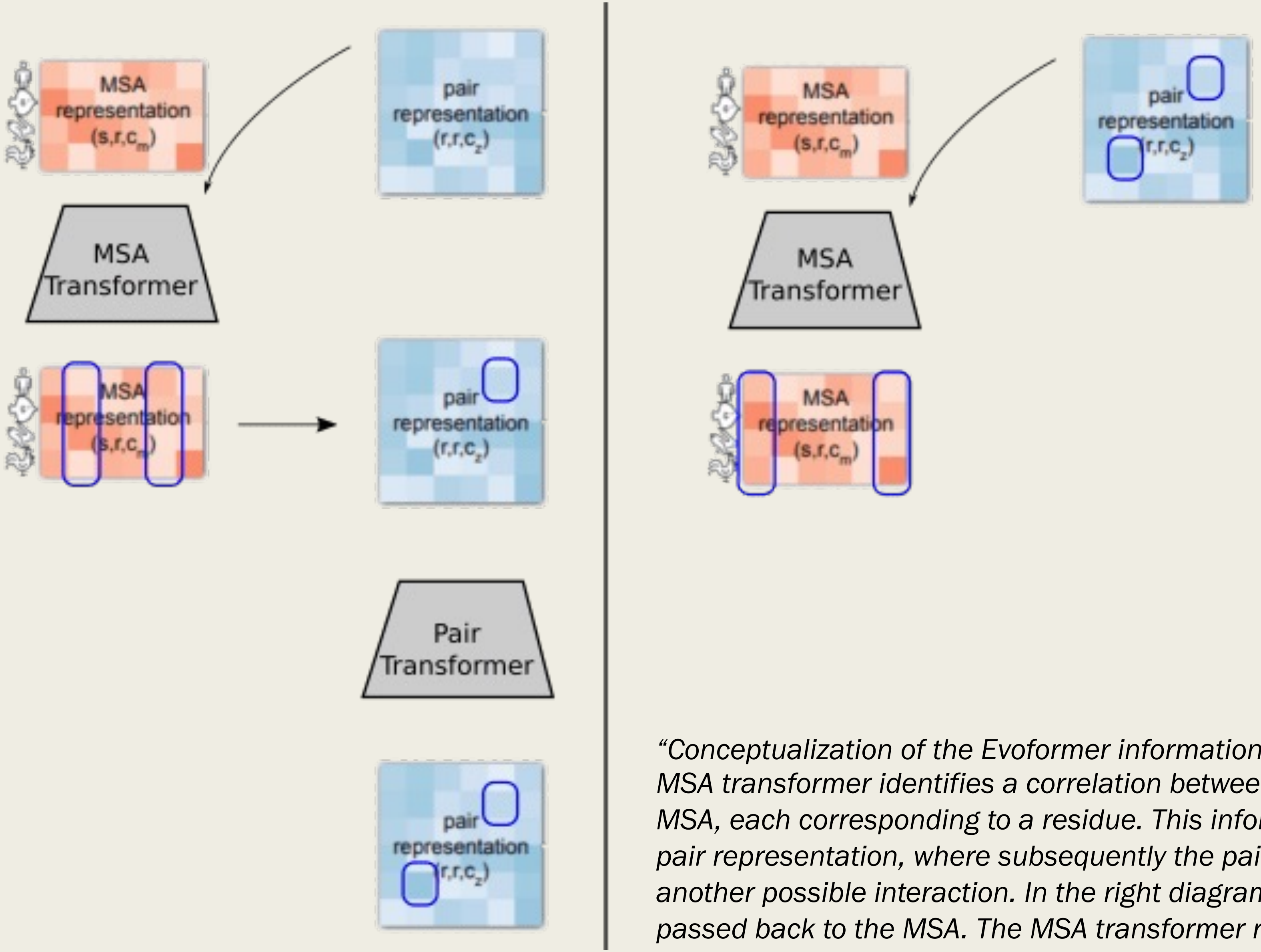
```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
```

https://doi.org/10.5281/zenodo.1405369

Pair representation



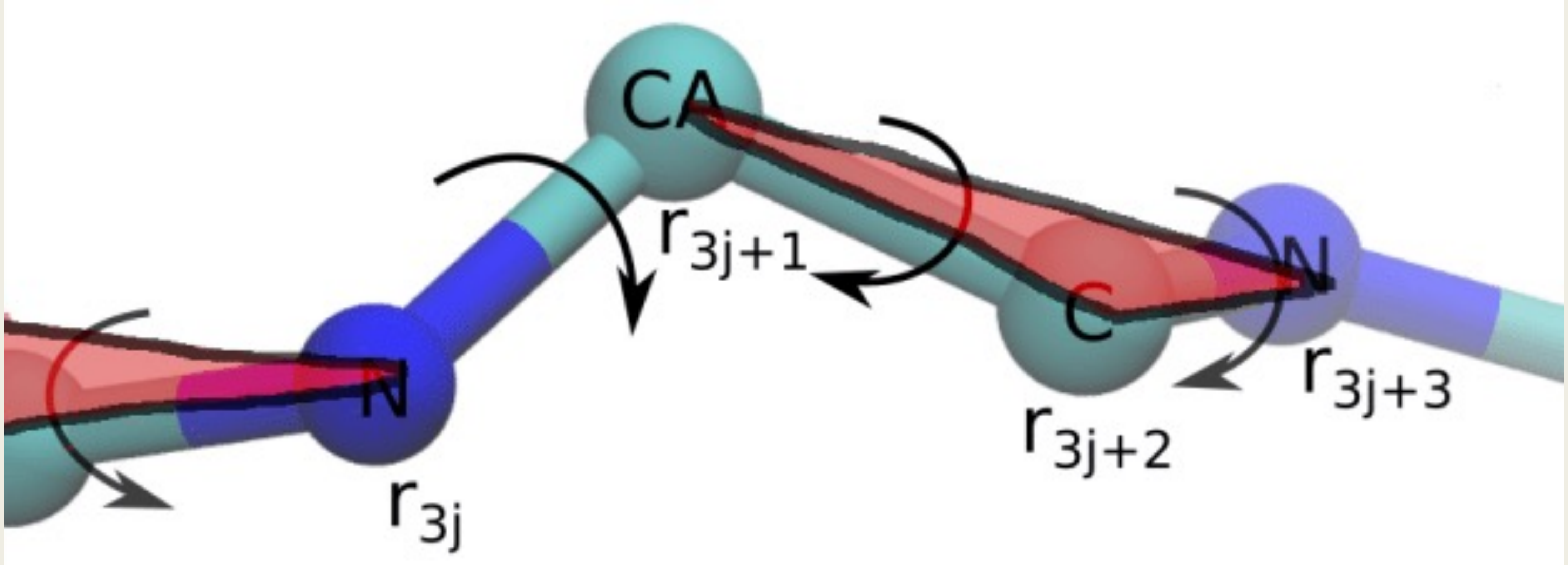



MSA representation (s,r,c_m)
pair representation (r,r,c_z)
MSA Transformer
MSA representation (s,r,c_m)
pair representation (r,r,c_z)
Pair Transformer
pair representation (r,r,c_z)
MSA representation (s,r,c_m)
pair representation (r,r,c_z)
MSA Transformer
MSA representation (s,r,c_m)

“Conceptualization of the Evoformer information. In the left diagram, the MSA transformer identifies a correlation between two columns of the MSA, each corresponding to a residue. This information is passed to the pair representation, where subsequently the pair representation identifies another possible interaction. In the right diagram, the information is passed back to the MSA. The MSA transformer receives an input from the pair representation, and observes that another pair of columns exhibits a significant correlation.”

https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/

Prediction



AlphaFold takeaways

- Key ideas here are
  - *1) Architecture*
  - *2) Training data*
  - *3) computing power*

- In some sense, this approach seems to be completely independent of how the proteins actually fold.

- What might be the limitations of this approach?

- Does this seem satisfactory to you?

Final thoughts

Other areas of Computational Biology

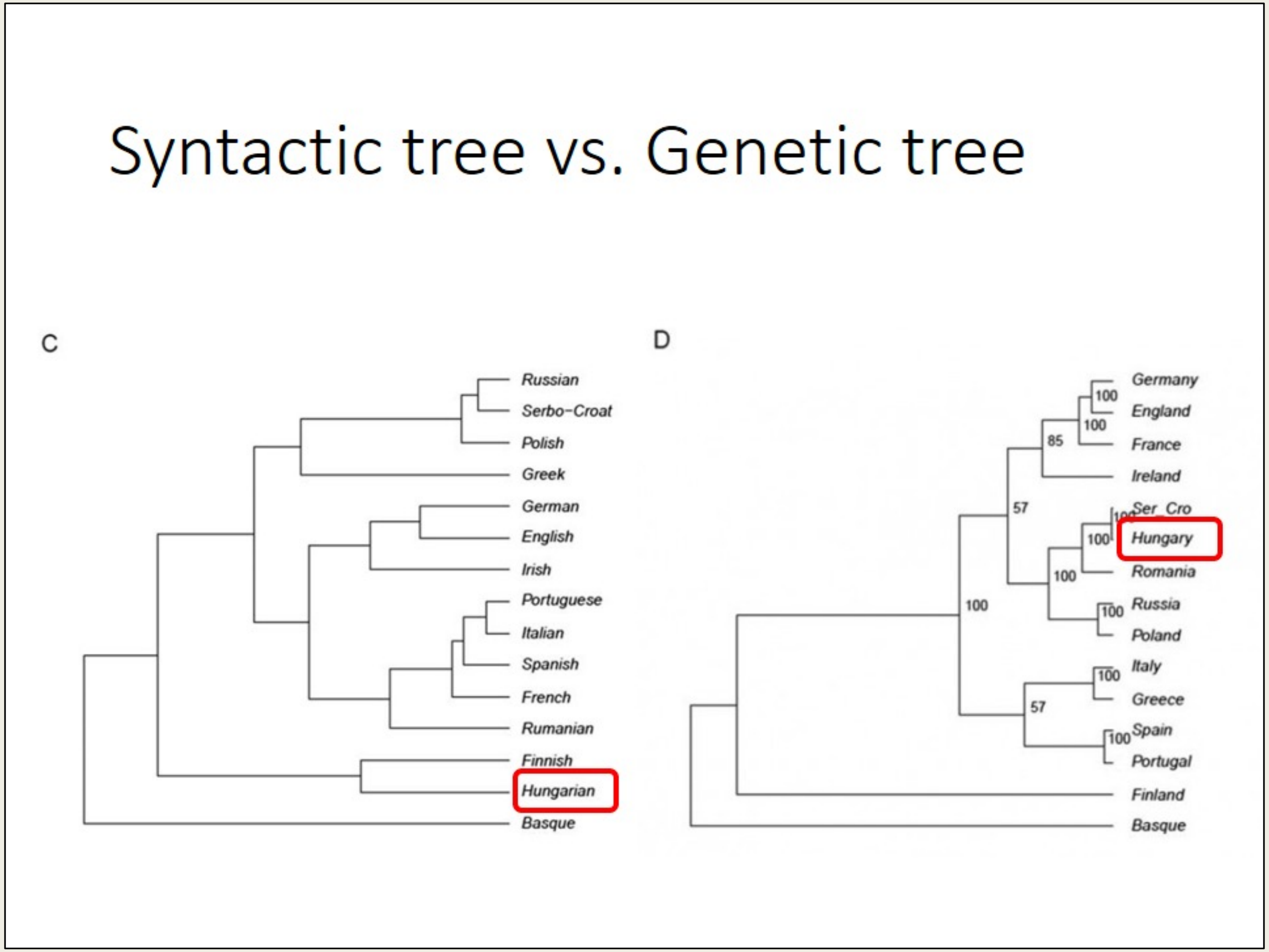
- **Biological Modeling**
  - *Drug entering the body*
  - *Tissue and surgical modeling*
  - *Gene networks*
  - *Intersects with computer vision, computer graphics, and graph theory*

- **Secondary and Tertiary Structure**
  - RNA secondary structure prediction
  - Protein folding

- **Neuroscience**
  - *Modeling the brain*

- **Disease biology**
  - *Pedigree analysis*
  - *Infectious disease models*
  - *Cancer biology*

Combining linguistics and genetics

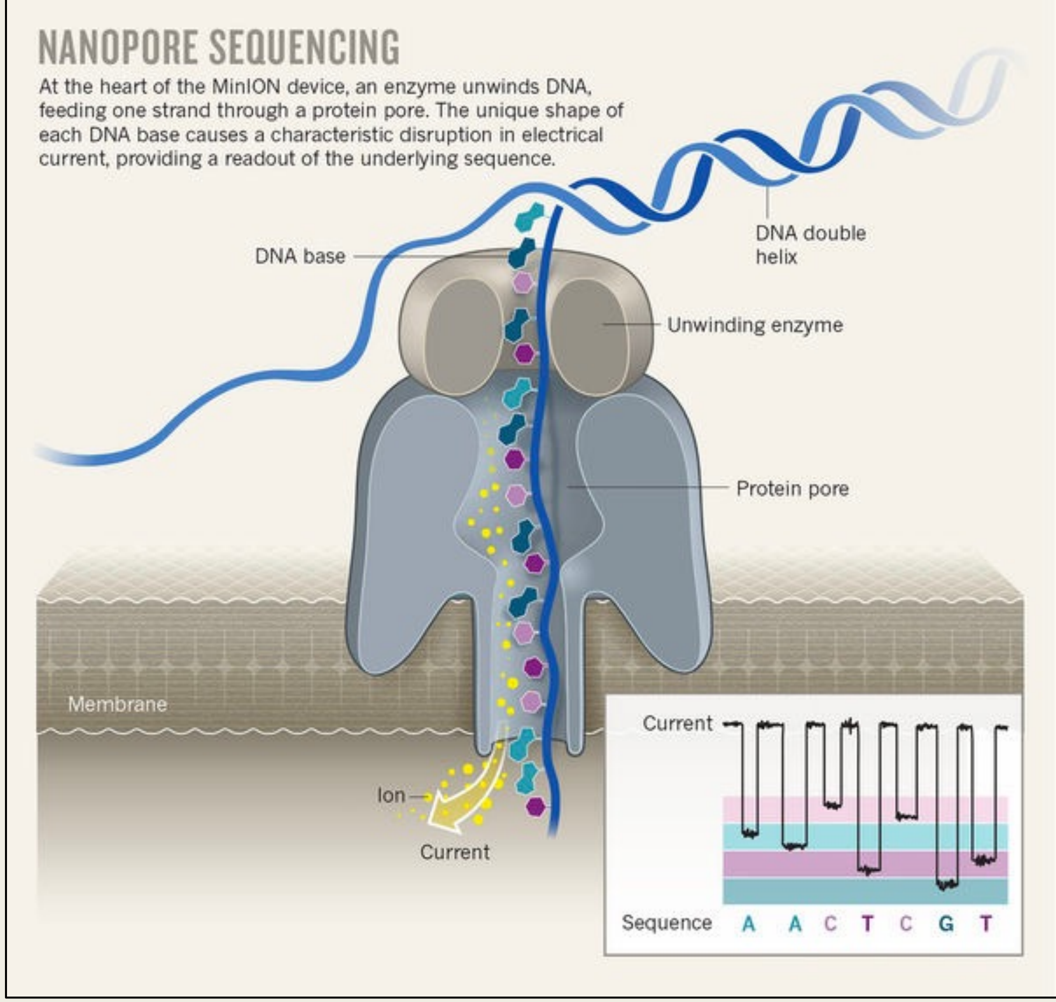


Syntactic tree vs. Genetic tree

Asya Pereltsvaig

Areas of Opportunity

- *Managing and analyzing data quickly and in a more automated way*
- *Intersecting with biochemistry to make sequencing better*
- *Sequencing more species, especially to assist conservation efforts*
- *Microbiome sequencing and understanding*

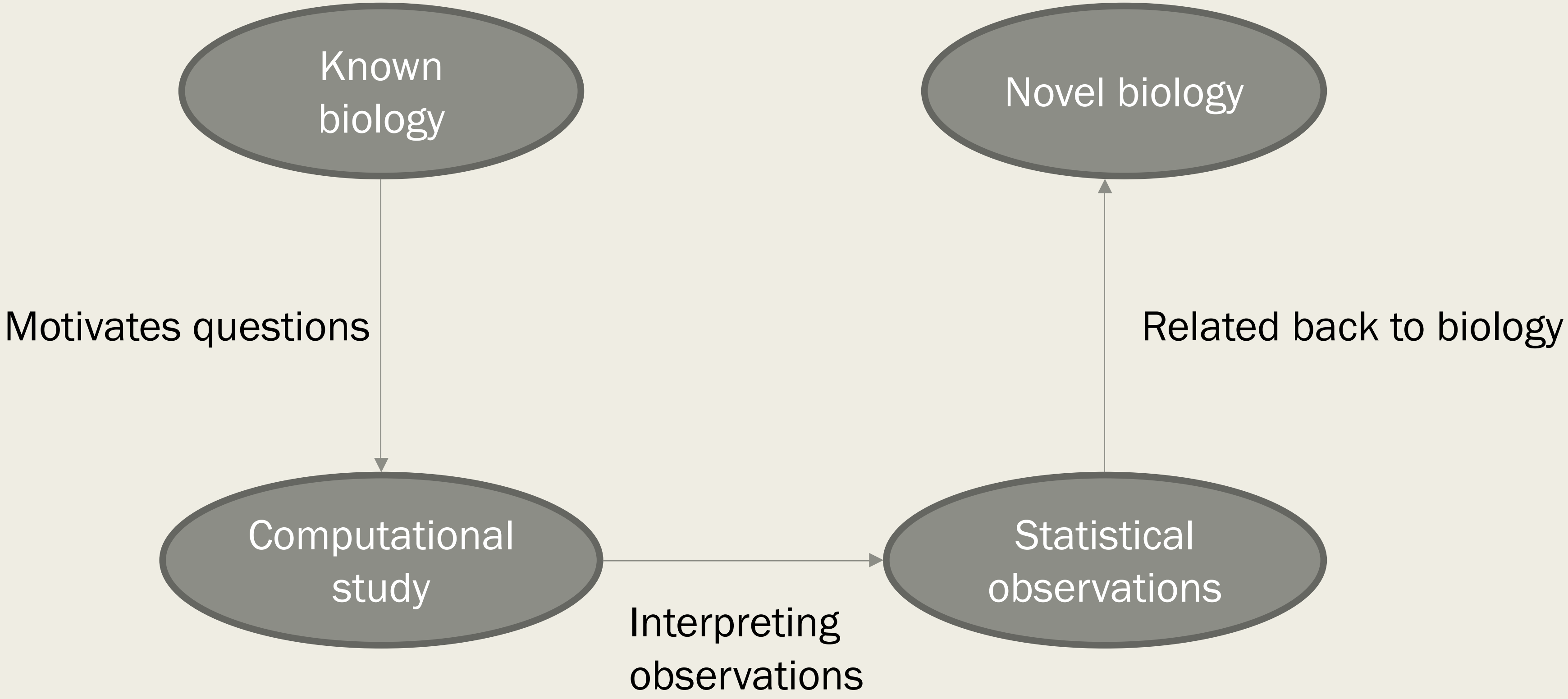


NANOPORE SEQUENCING

At the heart of the MinION device, an enzyme unwinds DNA, feeding one strand through a protein pore. The unique shape of each DNA base causes a characteristic disruption in electrical current, providing a readout of the underlying sequence.

Example: Oxford Nanopore

Image: blogs.nature.com



Known biology

Motivates questions

Computational study

Interpreting observations

Statistical observations

Related back to biology

Novel biology