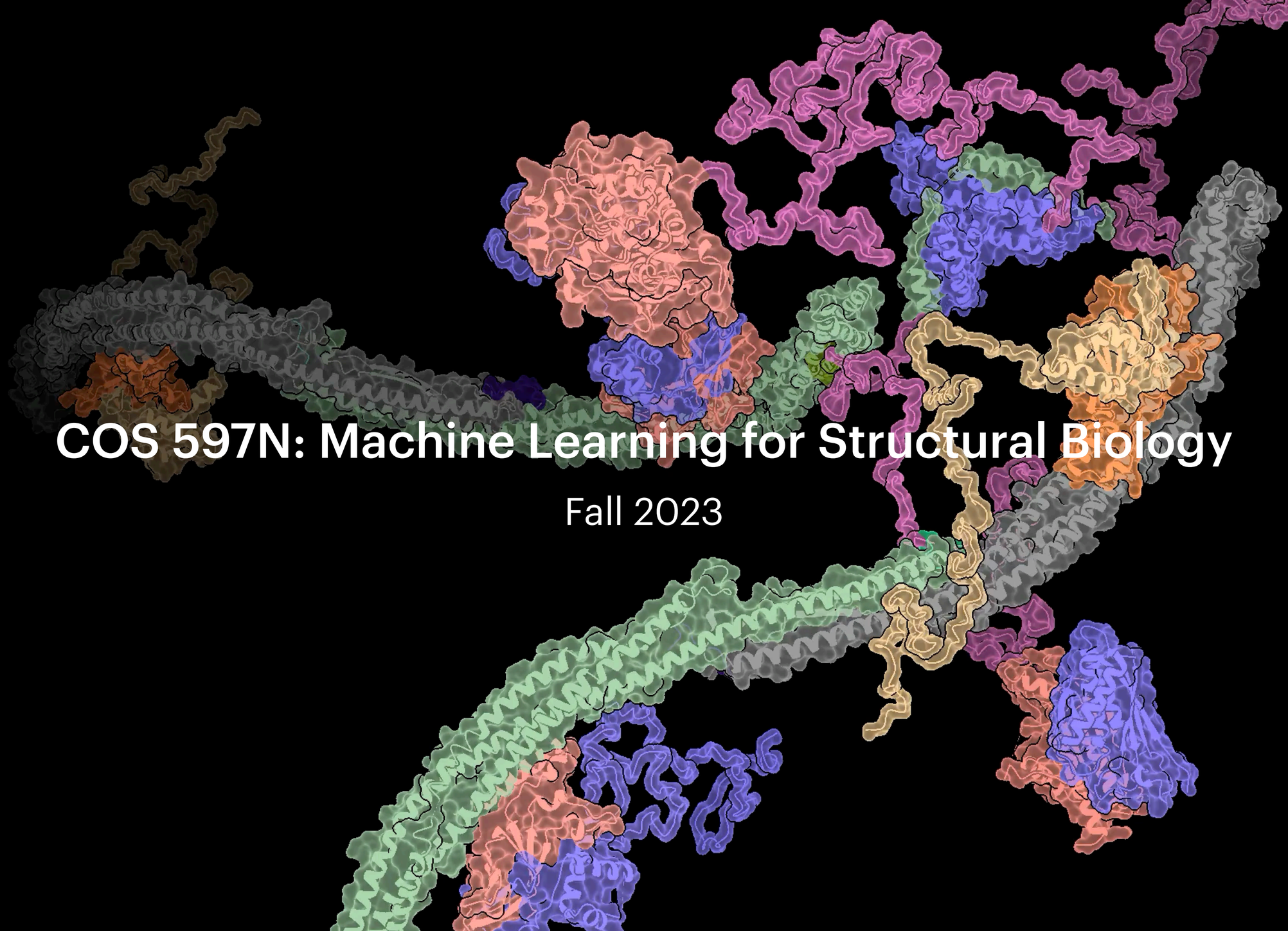


COS 597N: Machine Learning for Structural Biology

Fall 2023

Welcome!

- Introductions:
  - What is your year / background / research interests?
  - What do you want to learn from this class?
  - Fun fact?

- About me:
  - Second time teaching this class — as always, feedback is welcome!
  - Interested in machine learning for structural biology, cryo-EM methods, and 3D computer vision:
    - We will explore this new research area, its various subfields and connections to other topics in computer science throughout this semester

Course logistics

- Website: https://www.cs.princeton.edu/courses/archive/fall23/cos597N/
- Instructor: Ellen Zhong (she/her)
- Office hours: **Wednesdays 4:00-5:00p**, CS 314, via Calendly or by appointment

- Class meetings: **Thursdays 3:00-5:00p**
  - Aside from today, all classes will involve a group discussion of assigned papers
  - Attendance is required — contact me in advance if there are extenuating circumstances

- We will mostly use **slack** for communication and to share helpful resources or related papers, or #random! I will create a slack channel after this meeting and add everyone.

Course Design

- Goals of this course:
  - Learn about machine learning methods applied to problems in structural biology
  - Learn how to critically read and evaluate papers
  - Learn how to pose research problems and practice oral and written scientific communication skills
  - *Bonus:* Exposure to relevant basic and applied ML research in industry from guest speakers

- There are two components of this class:
  - Weekly in-class presentations and discussions on assigned papers
  - Final project

- Syllabus for more details

Prerequisites

- This is an advanced, interdisciplinary paper reading class.
- You should have exposure/working knowledge of machine learning concepts and deep learning architectures.
  - I will provide supplementary reading/primers. Add any helpful resources you find to the #resources channel in slack!
- No prior knowledge of biology is required, however, students should expect to develop a sufficient understanding of each application area to evaluate new developments.

- **Key prerequisite:** Interest in achieving a deep understanding of **both** ML algorithms and structural biology problems
  - Interested in “AI for science”? A key ability is to be able to read and understand papers from both communities

Grading

- Participation (30%)
- Presentation (30%)
- Final project (40%)

- Participation
  - One of the primary ways of learning and engaging with the material
  - Examples: Posing or answering questions, explaining background material, sharing reflections

- Grades are not the goal of this graduate-level seminar
  - The goal is to learn about this research area and engage with your peers!
  - A highly interdisciplinary area — everyone brings a unique perspective.

Additional logistics

- Any important course announcements will be communicated through email
- I will be updating the course website, not Canvas

- We will leave a week open in the latter half of the semester for any late-breaking work!
- There will be a few guest lecturers during the semester, either guiding the discussions (if I am out of town) or giving a presentation.
- Any other questions?

- Full syllabus here
- Website here: https://www.cs.princeton.edu/courses/archive/fall23/cos597N/

Rest of this class

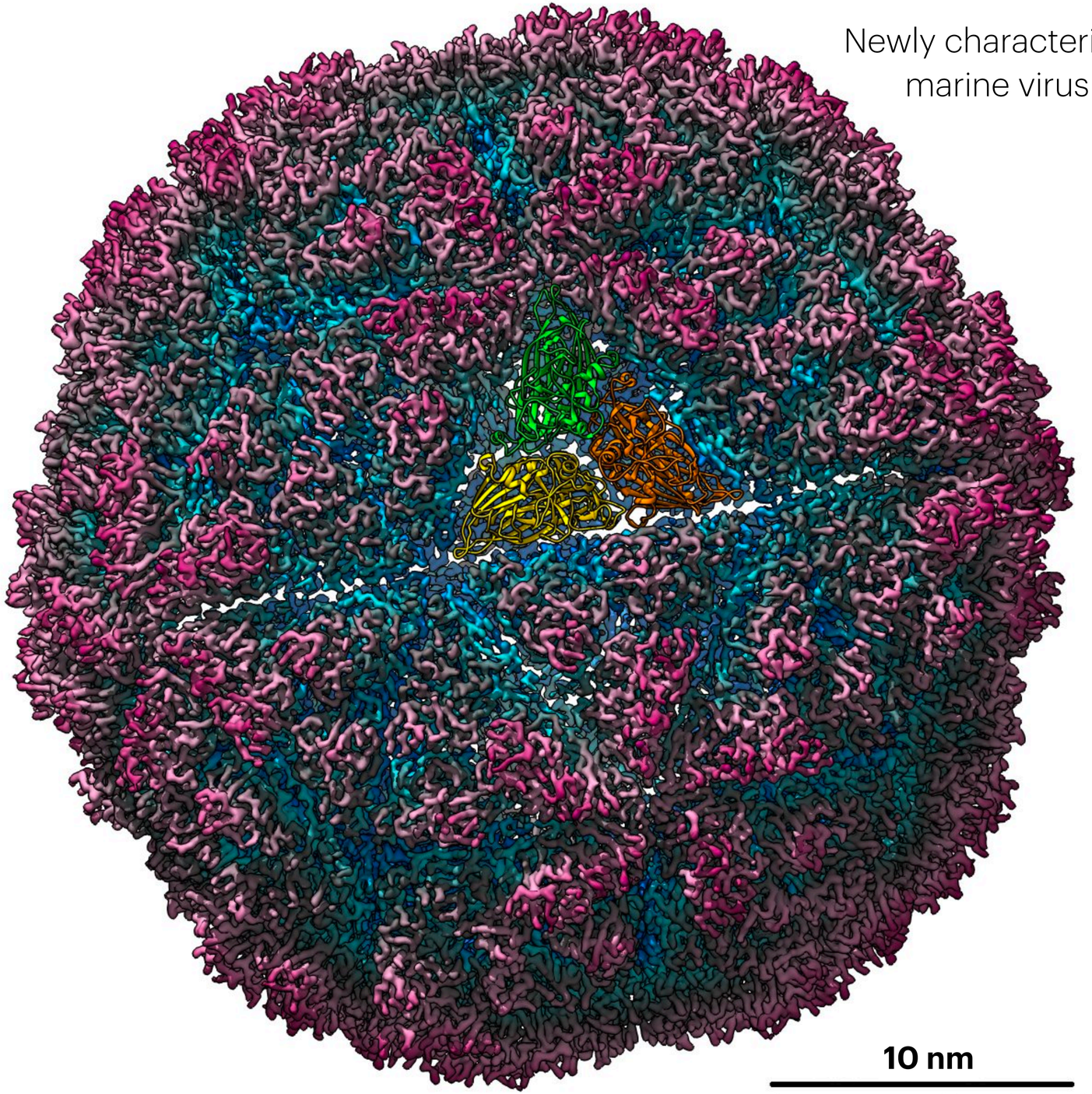
- An introduction to structural biology through the lens of biology, chemistry, physics, and computer science
- Recent breakthroughs in structural biology from machine learning (AlphaFold2)
- An overview of topics in this course
- Discussion on paper reading strategies

An introduction to structural biology

1. Motivation: What is structural biology? What are proteins? Why should you care?

2. Background: History of structural biology and protein structure 101

3. Current moment in machine learning for structural biology



(Image courtesy of Jason Kaelbler)

The central dogma of molecular biology

DNA sequence

`ATGCACTTGAGCAGGGAAGAAA...`

RNA sequence

`AUGCACUUGAGCAGGGAAGAAA...`

Protein sequence

`MSTAGKVIKCKAAVLWELKKPF...`

Human genome:

* Contains around 3 billion base pairs
* Encodes ~20k genes

Proteins are the final product of the genetic information flow

Modern molecular biology research: how is life implemented by our genetic code?

Structural biology: The study of proteins and other biomolecules through their 3D structure

DNA sequence

ATGCACTTGAGCAGGGAAGAAA...

↓

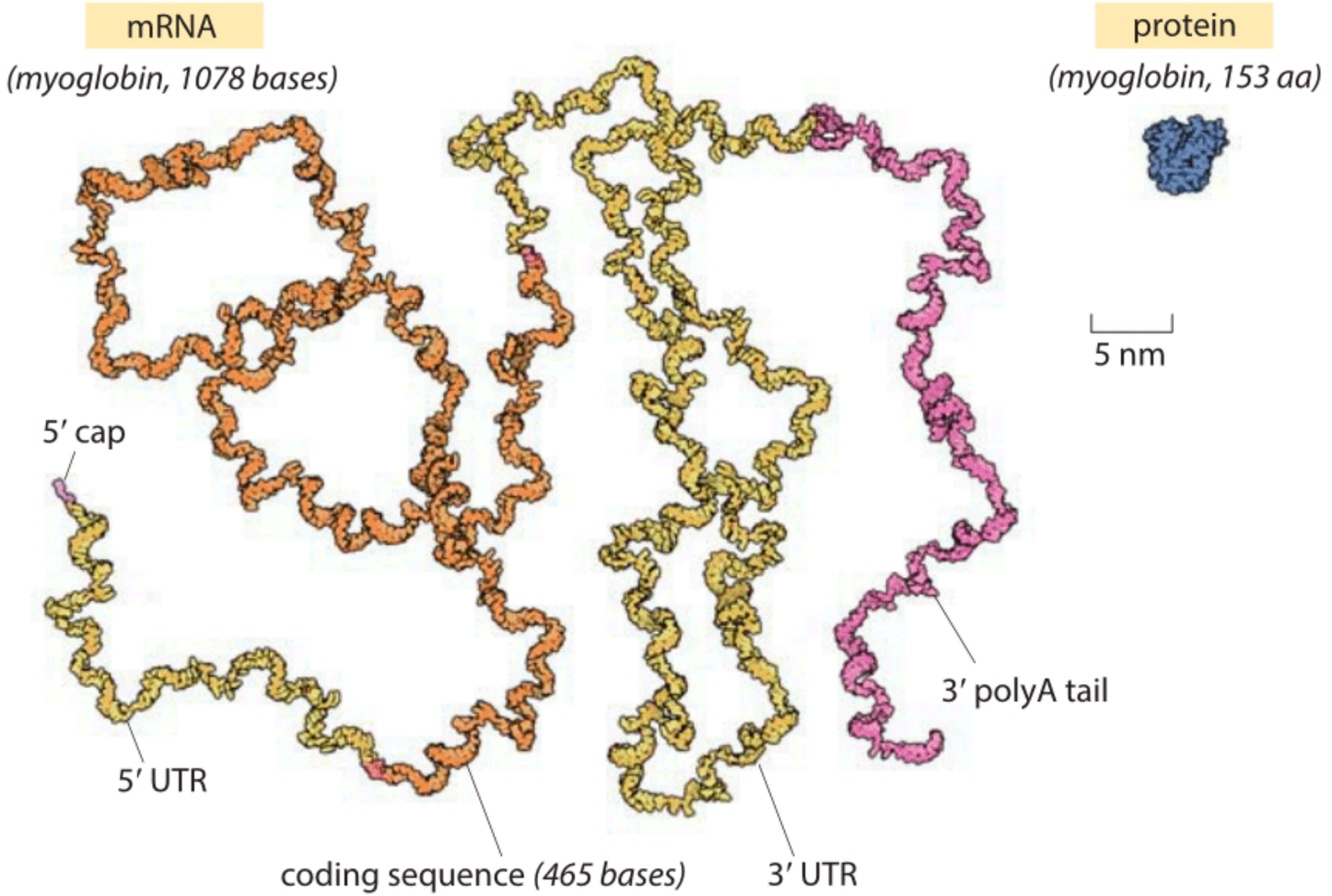
RNA sequence

AUGCACUUGAGCAGGGAAGAAA...

↓

Protein sequence

MSTAGKVIKCKAAVLWELKKPF...



Cell Biology By The Numbers. Illustration by David Goodsell.

Structural biology: The study of proteins and other biomolecules through their 3D structure

DNA sequence

ATGCACTTGAGCAGGGAAGAAA...

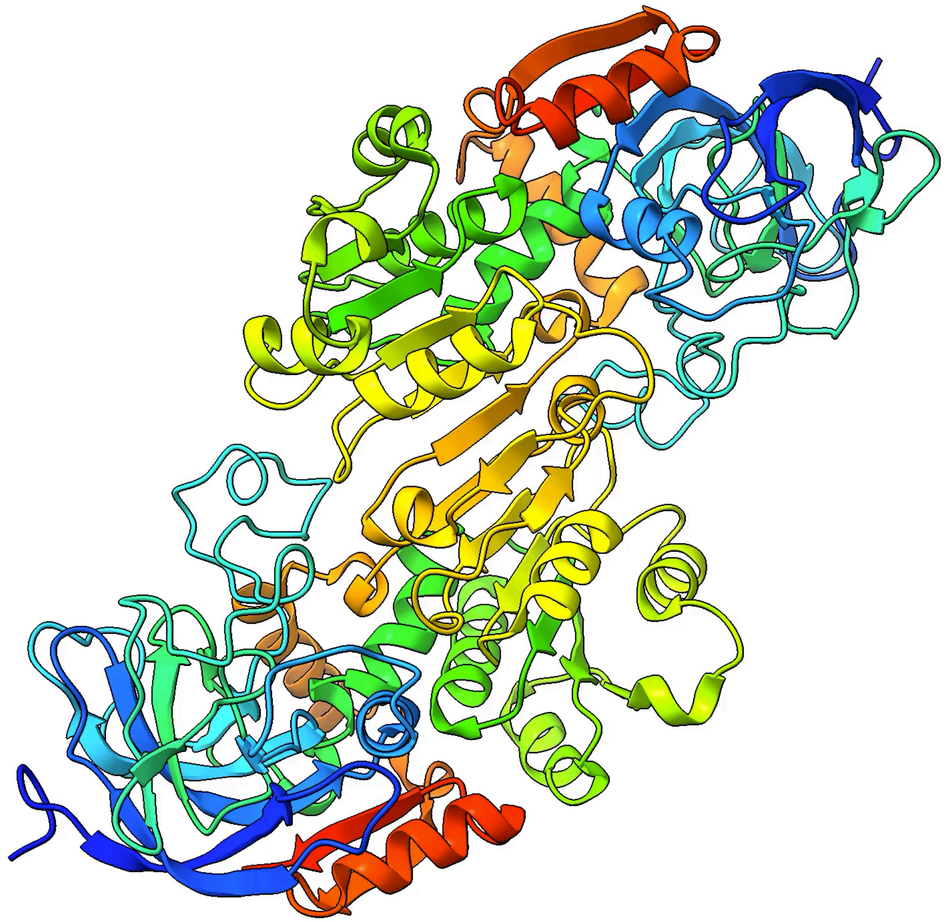
RNA sequence

AUGCACUUGAGCAGGGAAGAAA...

Protein sequence

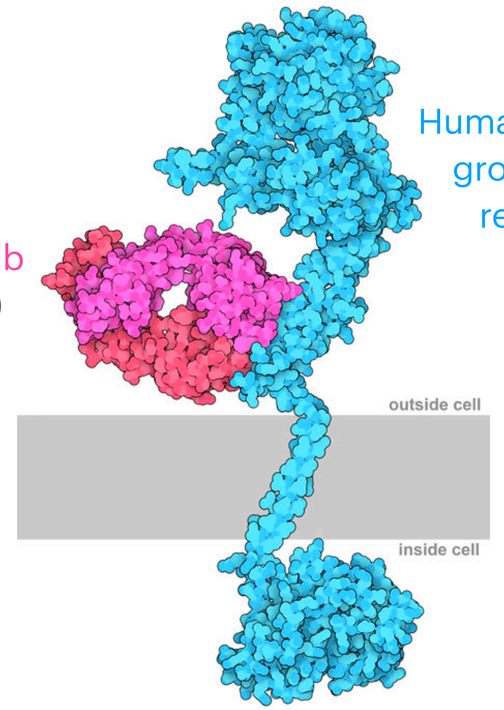
MSTAGKVIKCKAAVLWELKKPF...

Protein folding

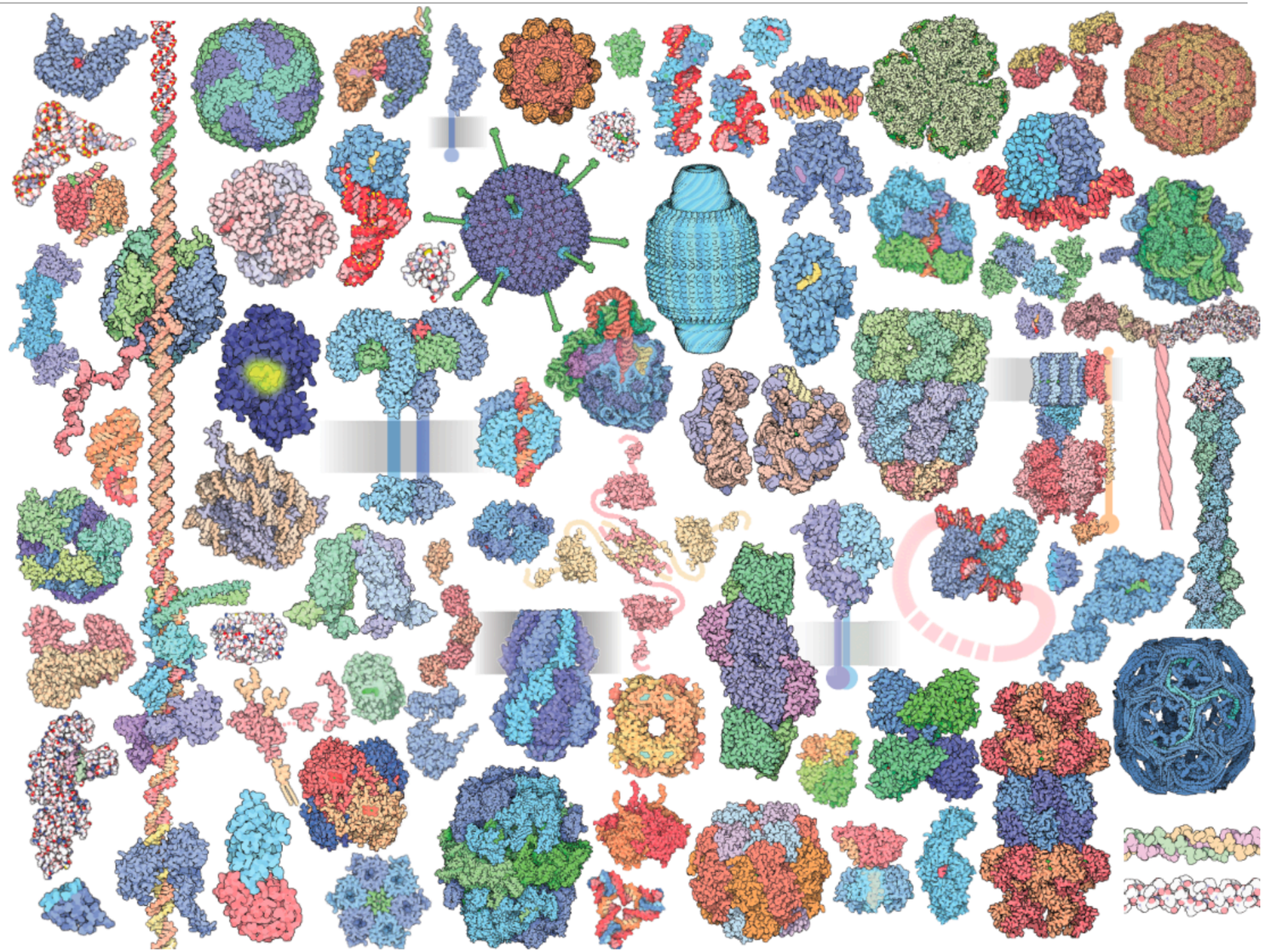


All essential biological processes are carried out by proteins and protein complexes

- Fundamental molecules of life
- Medicine and health
- Nanotech and biotech



PDB-101 Molecule of the Month

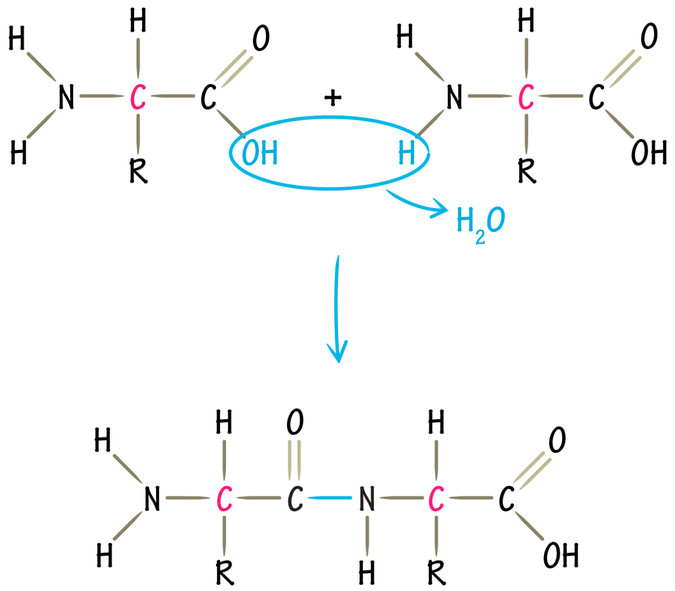


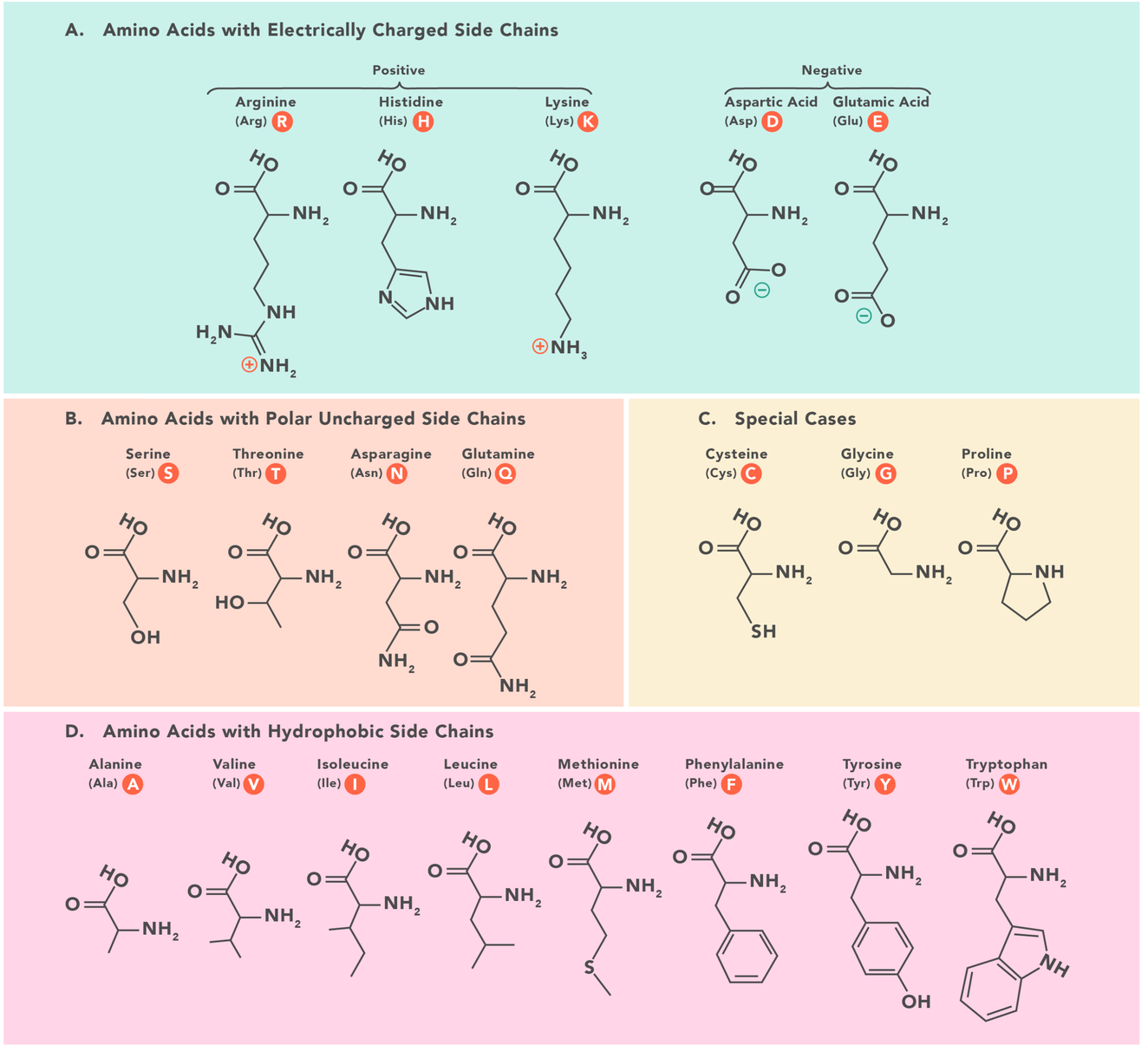
Goodsell et al. PLoS Biology 2015.

What are proteins?

- A linear sequence of amino acids polymerized in a chain
- An alphabet of twenty possible amino acids
  - Common *backbone* but different *side chains*
- Various non-covalent interactions and other forces drive folding of the chain into a globular 3D structure*

Peptide Bond Formation

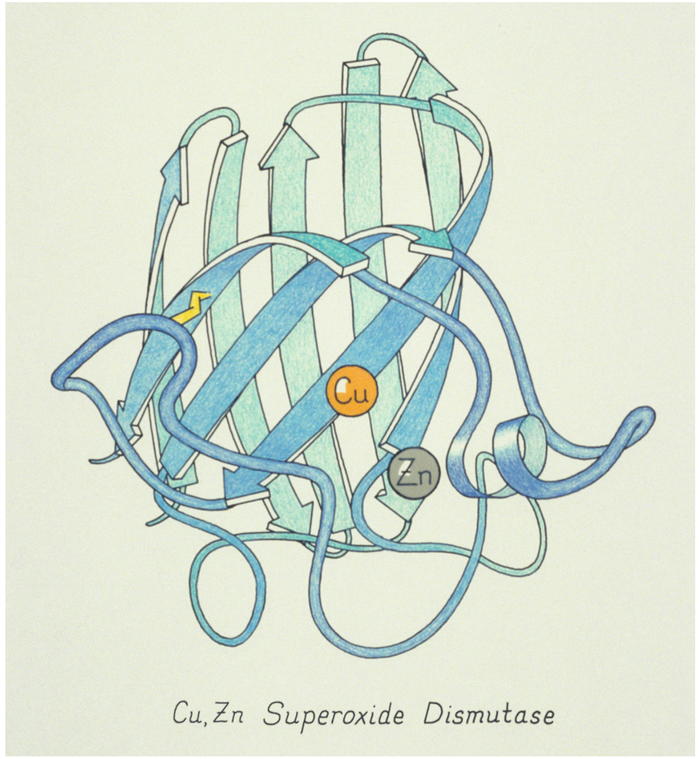


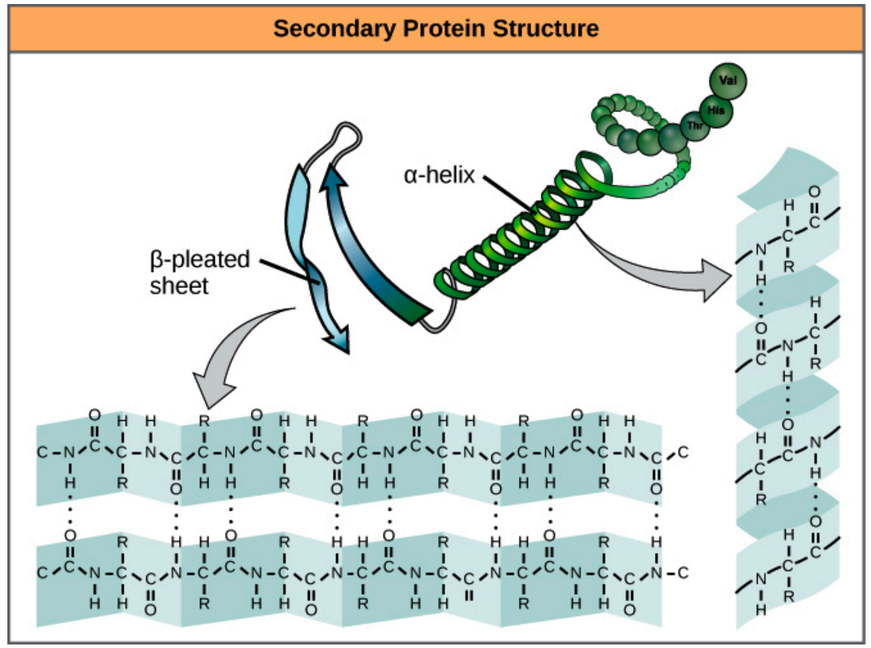


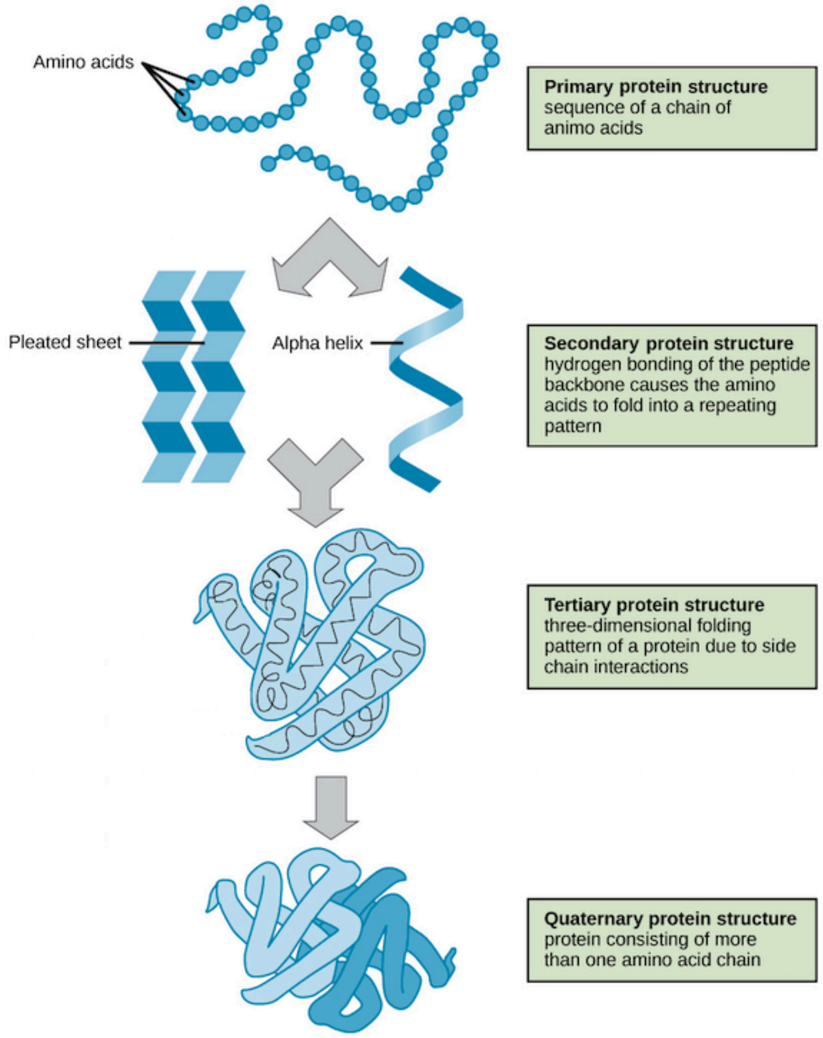
* There have been many different paradigms for thinking about protein folding. See Afinsen's hypothesis, Dill et al. 2008, The Protein Folding Problem

Primary, secondary, and tertiary structure

- Ribbon diagram for visual interpretation of structure, developed by Jane Richardson in the late 1970s - early 1980s
- See her keynote at MLSB 2021 @ NeurIPS for a historical overview





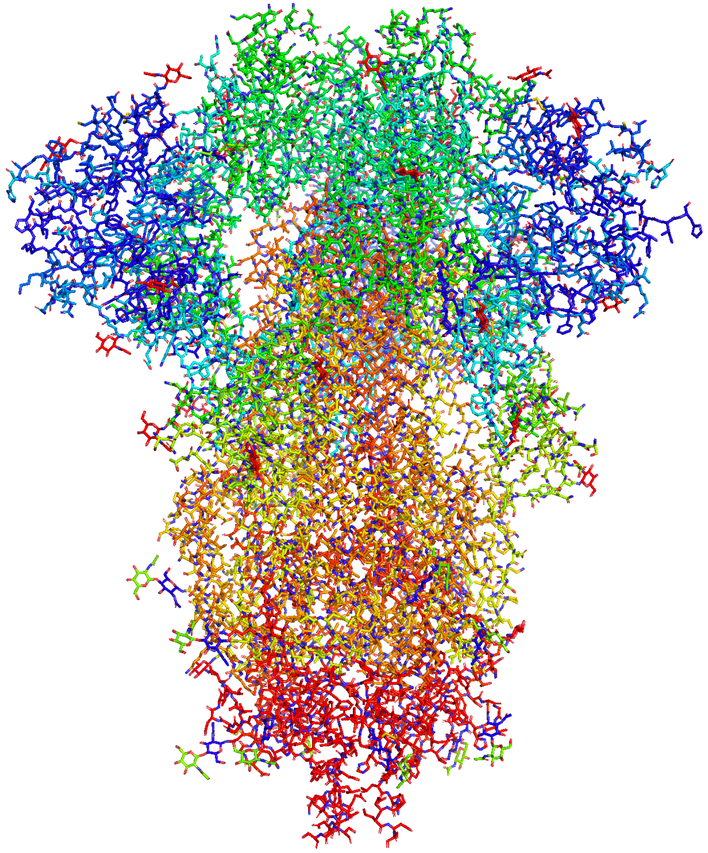


https://stories.duke.edu/sciences-mother-of-ribbon-diagrams-celebrates-50-years-at-duke

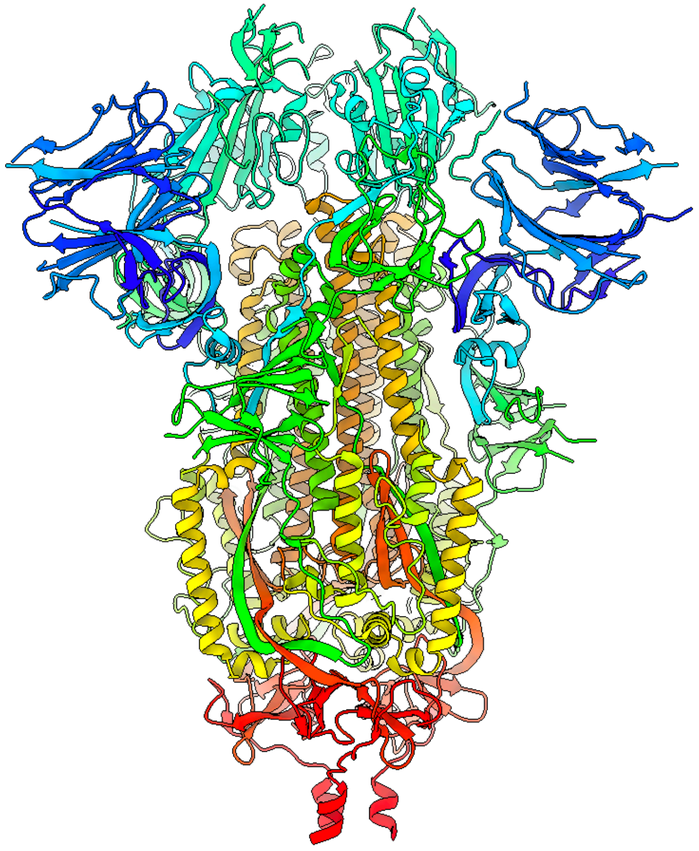
Different choices in visualizing and representing 3D structure

Atomic model

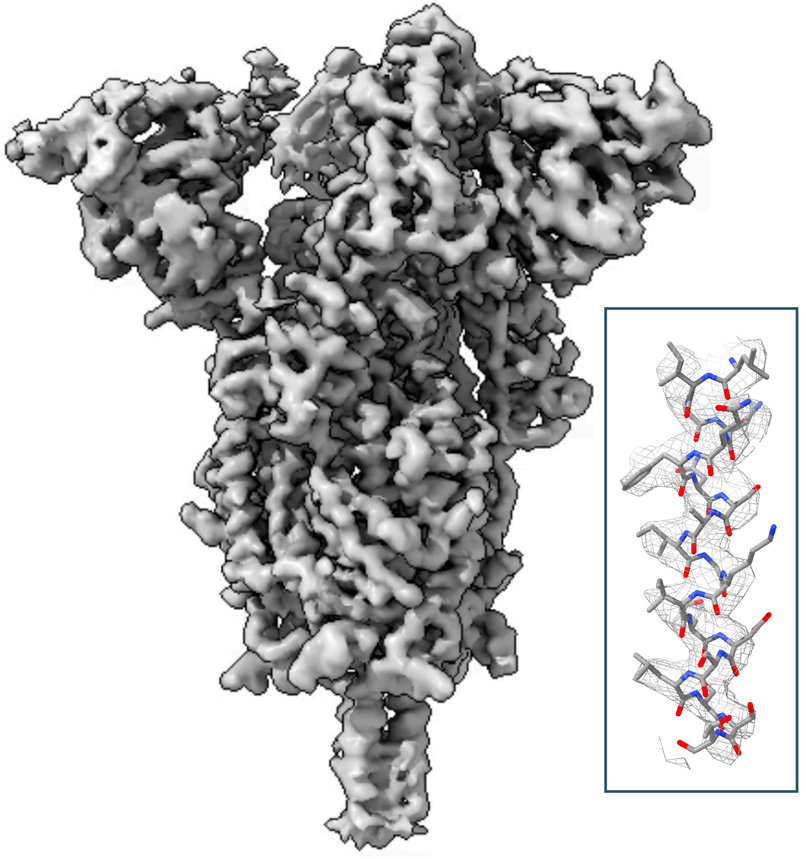
Density volume



(full atom representation)



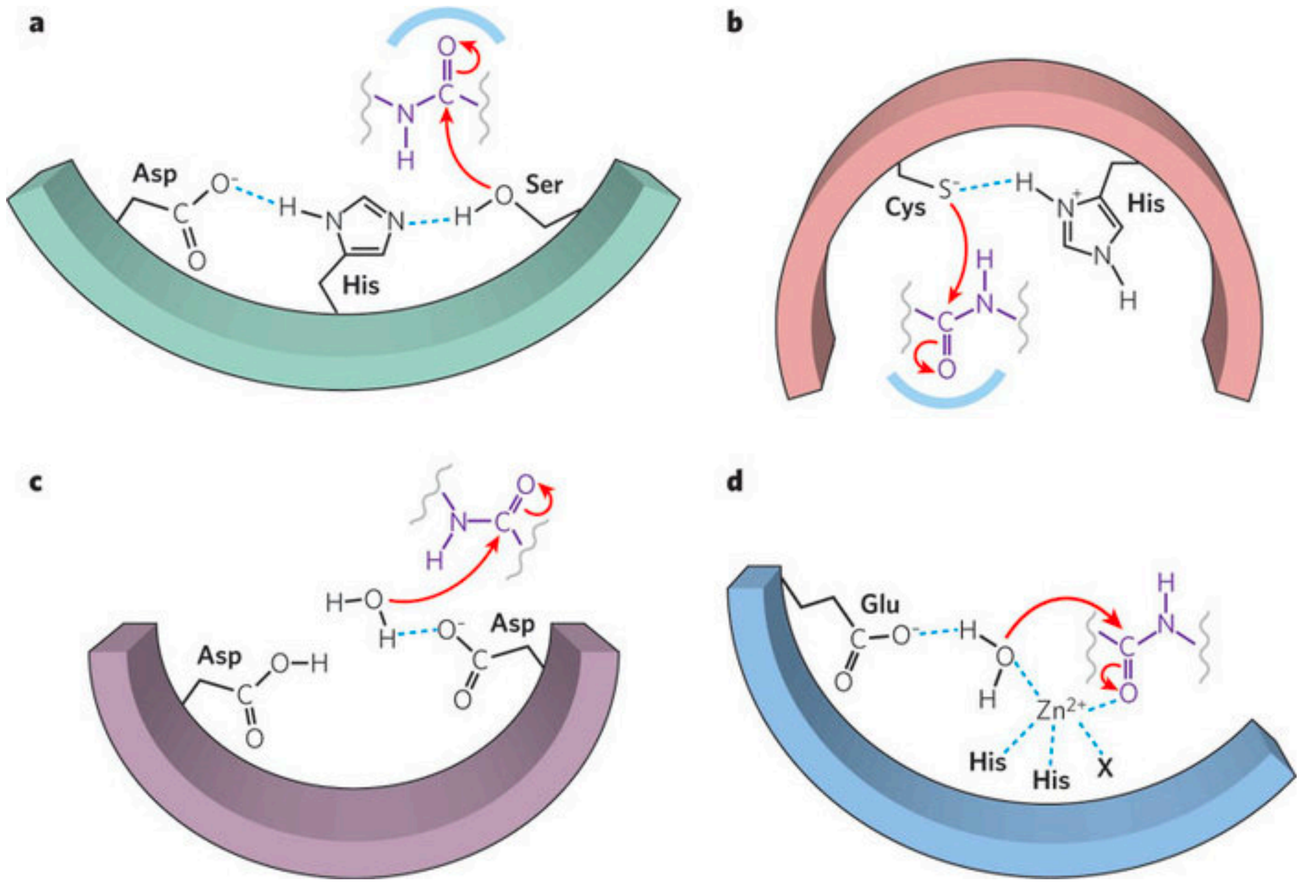
(backbone representation)



(isosurface contour)

Many proteins are enzymes that catalyze chemical reactions

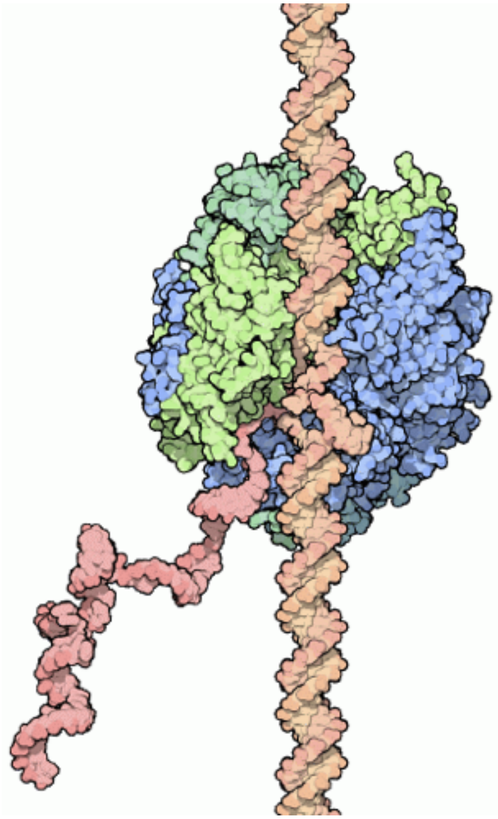
- The precise structural arrangement of amino acid residues creates the opportunity to bind and catalyze chemical reactions
- Catalysis is carried out at an active site or binding site
- What are some examples of enzymes?



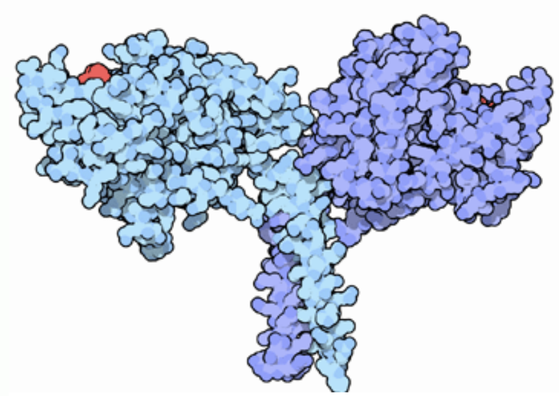
Polypeptides can be cleaved either chemically or enzymatically. Enzymes that catalyse the hydrolytic cleavage of peptide bonds are called proteases.

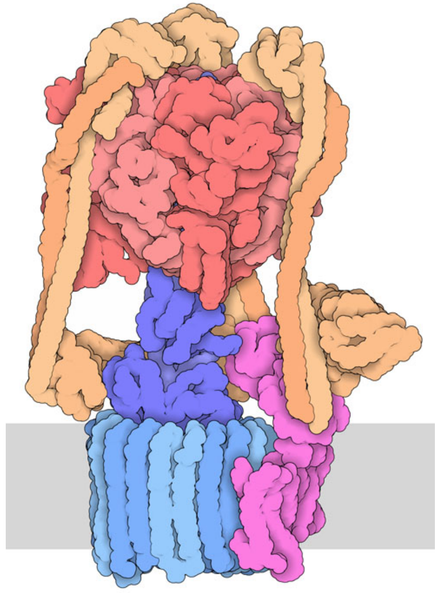
https://www.nature.com/scitable/content/protease-mechanisms-14462487

Proteins form large macromolecular machines

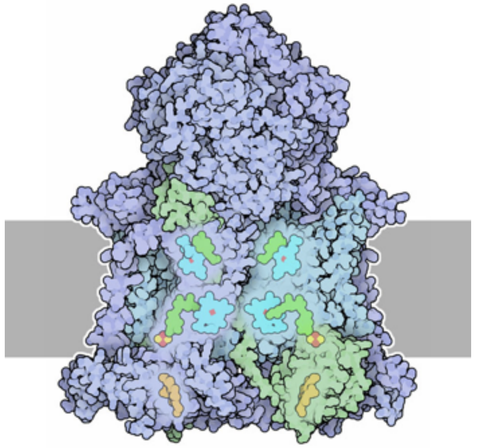


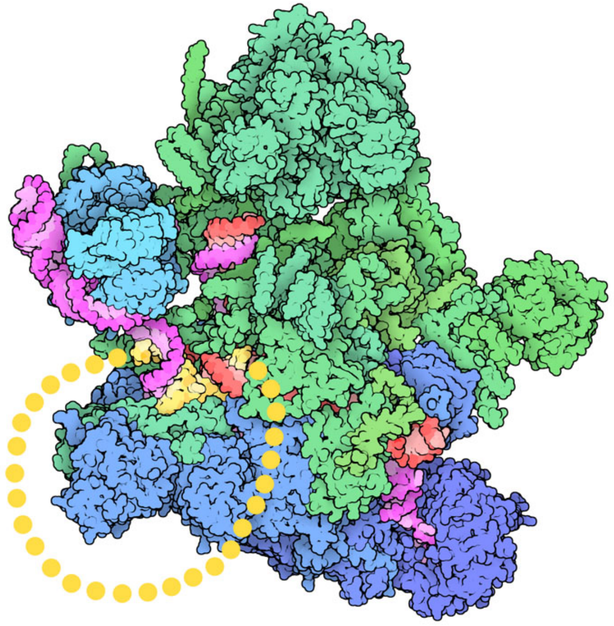
RNA polymerase

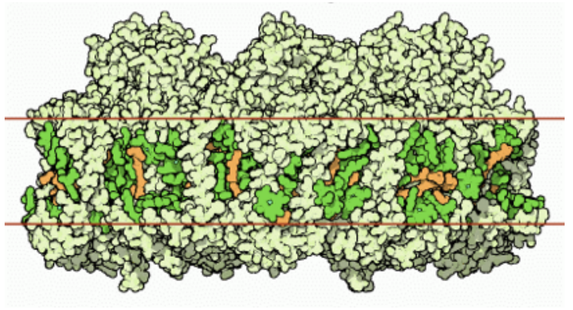


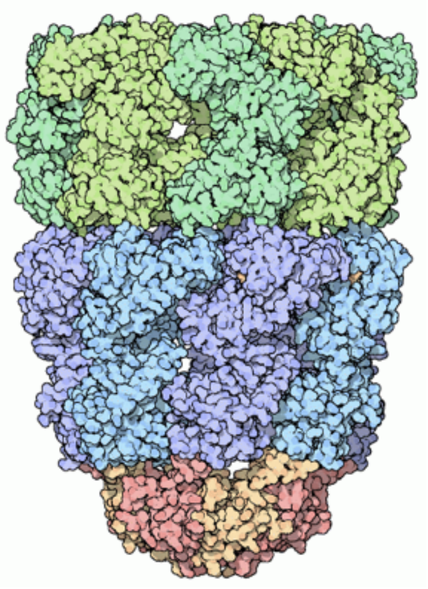


ATP synthase





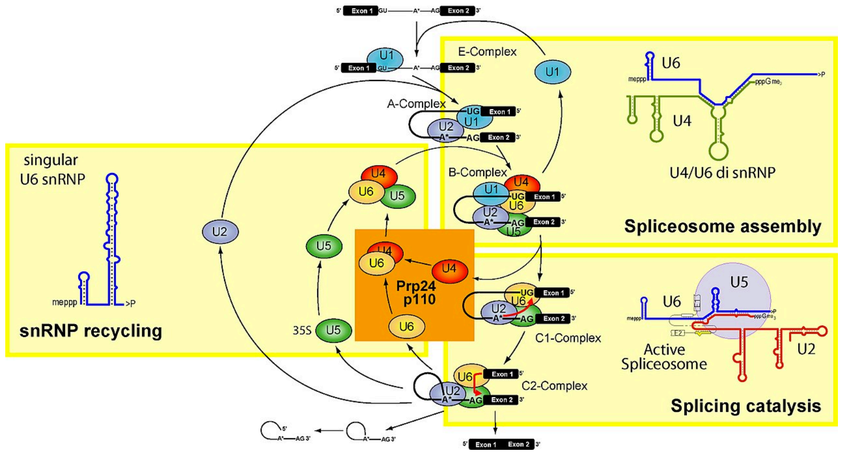




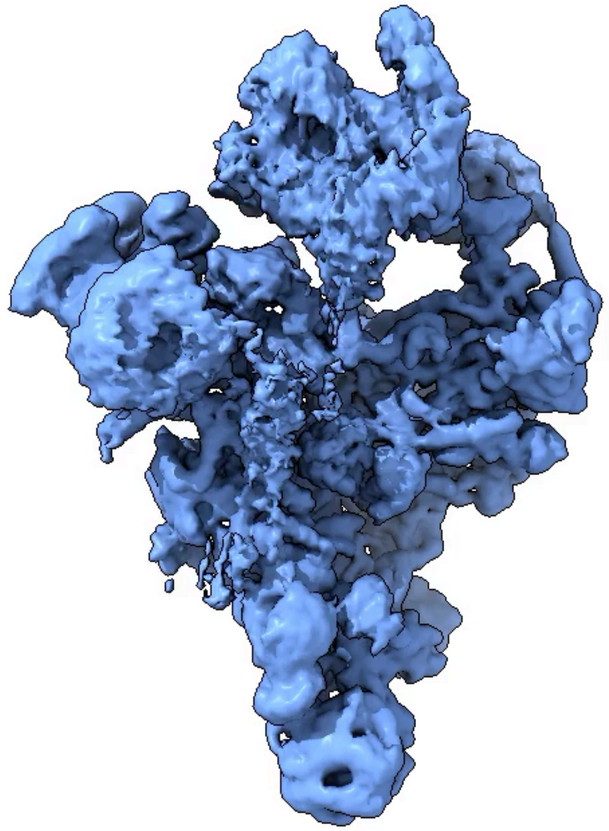
https://pdb101.rcsb.org/

Proteins form large, dynamic macromolecular machines

Spliceosome splicing cycle

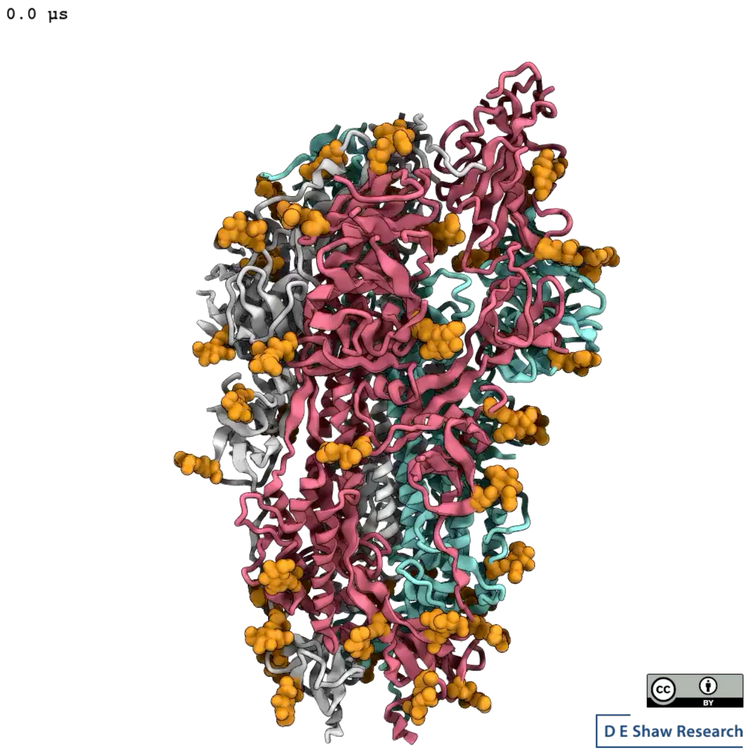


https://en.wikipedia.org/wiki/Spliceosome



Zhong et al, Nature Methods 2021

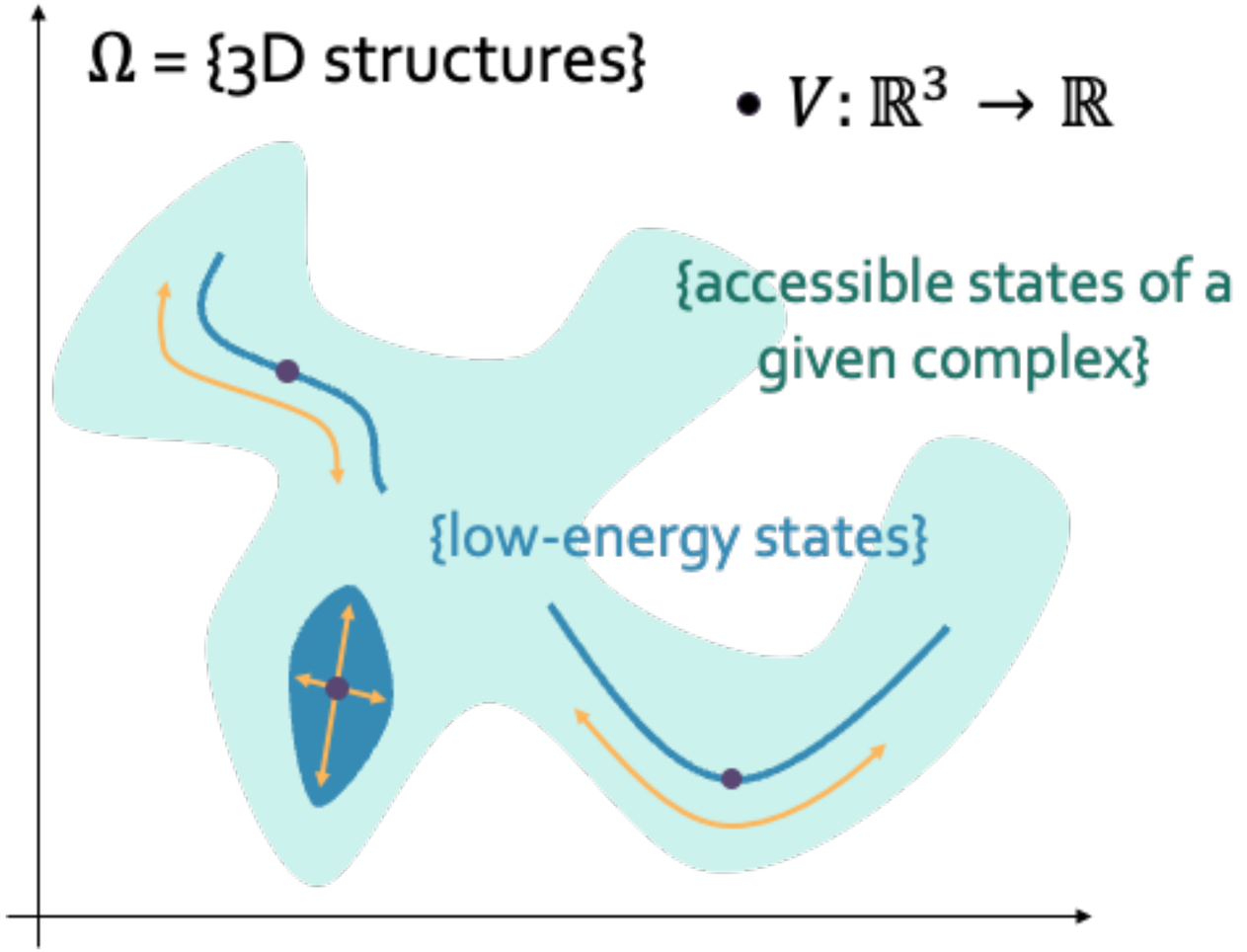
cryoDRGN trajectory of the pre-catalytic spliceosome



MD simulation of SARS CoV-2 Spike

How do we model structural variability?

Molecules can move, bind and unbind



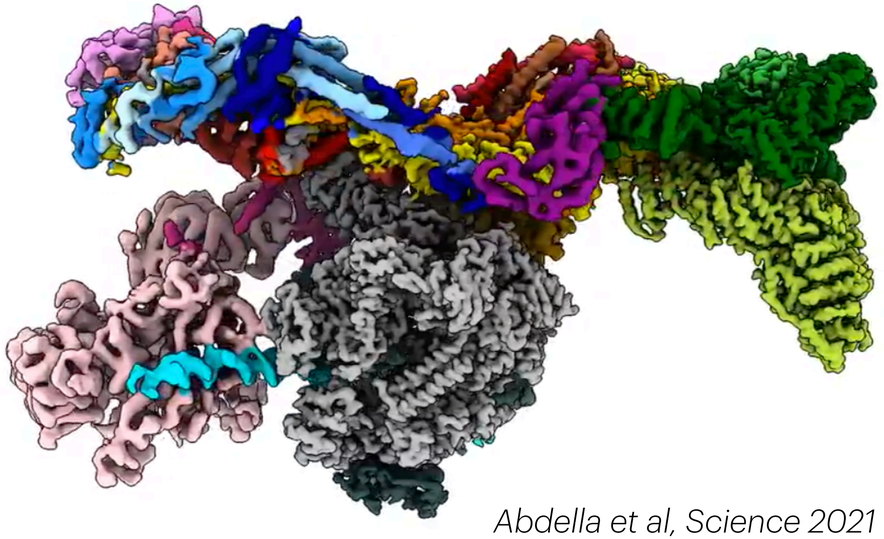
Low-energy states
=
"Conformational space"
=
Discrete union of **continuous** subspaces

Compositional heterogeneity

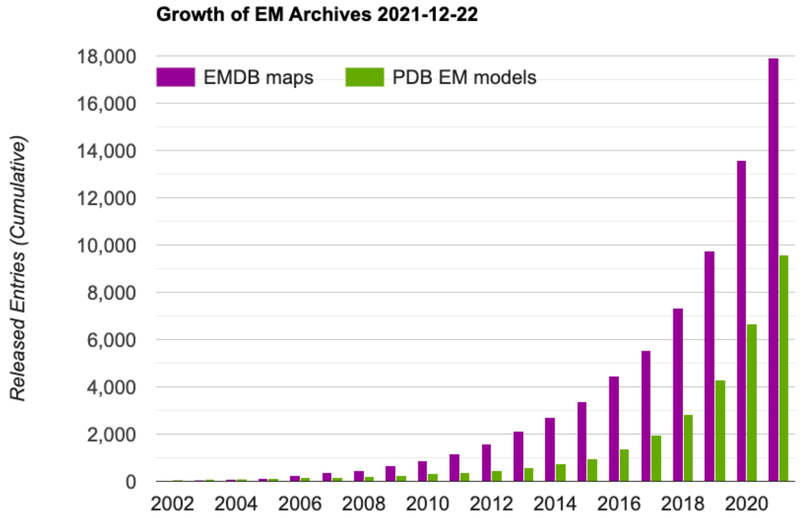
Conformational heterogeneity

Experimental approaches for protein structure determination

- The first protein structure by Linus Pauling, Robert Corey, and Herman Branson in 1951
- NMR spectroscopy
- X-ray crystallography
- Cryo-electron microscopy (cryo-EM)
  - 2017 Nobel prize in Chemistry
  - Opened up new areas of structural biology through recent technological advances
  - New computational challenges and opportunities

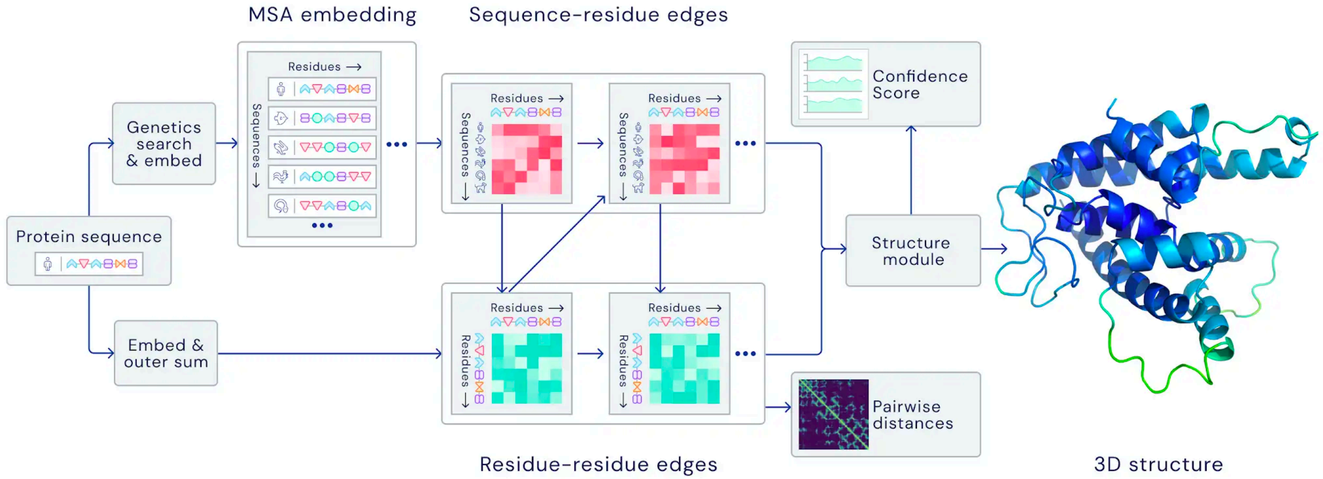


Abdella et al, Science 2021

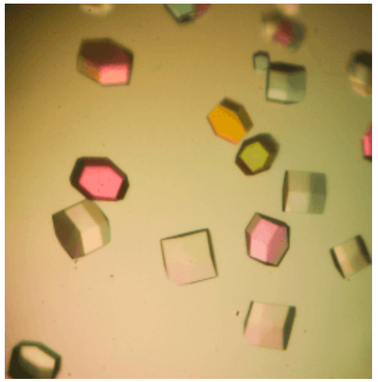


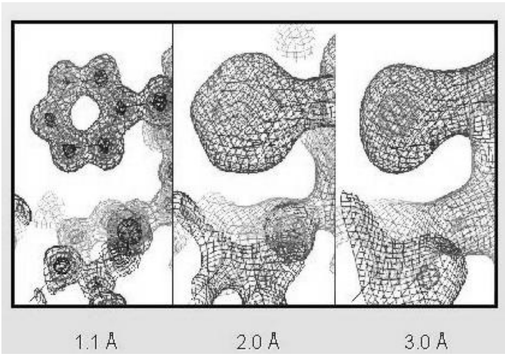
Different levels of “accuracy”: from *in silico* to *in vitro* to *in situ*

Protein structure prediction

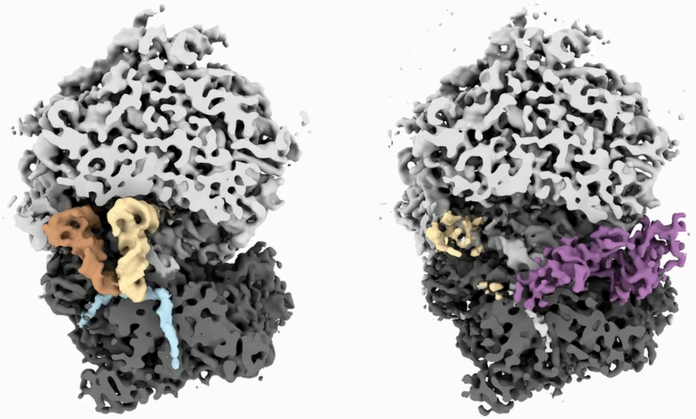


Experimental 3D structure determination From crystals

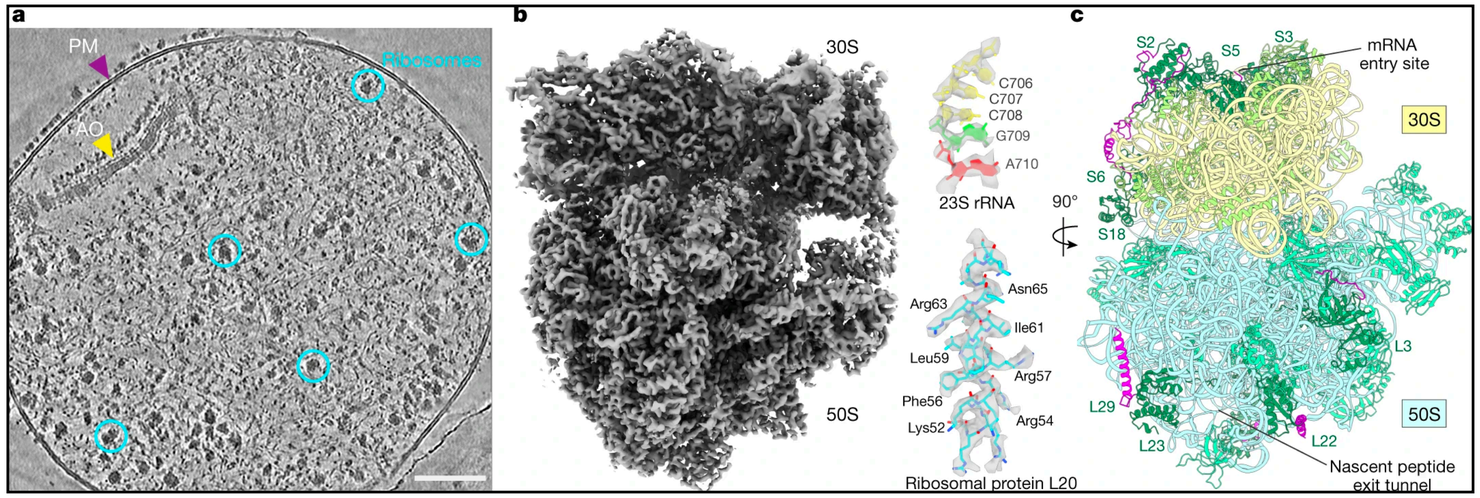




From purified solutions (cryo-EM)



Directly in the cell i.e. the “native” environment



What is the protein folding problem?

- The protein folding problem is the question of how a protein's amino acid sequence dictates its three-dimensional atomic structure.
- Emerged as a “problem” in the 1960s
- The “protein folding problem” consists of three closely related puzzles:
  - (a) What is the folding code? (Thermodynamics)
  - (b) What is the folding process or mechanism? (Kinetics)
  - (c) The computational problem of can we predict the native structure of a protein from its amino acid sequence
- From Dill et al, Annual Review of Biophysics, 2008

Recap: Understanding protein structure through different lenses

- (Biology) — Protein function (and dysfunction), genomic and cellular contexts
- (Chemistry) — Amino acids, pKa, biological catalysts
- (Physics) — Statistical mechanics, the Boltzmann distribution, and free energy landscapes
- In this class, our goal is to explore a computer science perspective on problems in structural biology

Motivations for this course

- Structural biology poses a rich set of algorithmic challenges and scientific opportunities
- A ***new*** and ***rapidly-evolving*** field
  - 1st NeurIPS workshop on MLSB (2020)
    - “...structural biology... has emerged as an area of great promise for machine learning"
  - 2nd NeurIPS workshop on MLSB (2021)
    - “Structural biology ... is a field on the cusp of transformation.... recent machine-learning based modeling approaches have shown that it will become routine to predict and reason about structure at proteome scales with unprecedented atomic resolution.”

3rd NeurIPS workshop on MLSB (2022)

- In only a few years, structural biology... has been transformed by breakthroughs from machine learning algorithms. Machine learning models are now routinely being used by experimentalists:
  - to **predict structures** that can help answer real biological questions (e.g. AlphaFold),
  - accelerate the experimental process of **structure determination** (e.g. computer vision algorithms for cryo-electron microscopy), and
  - have become a new industry standard for **bioengineering new protein therapeutics** (e.g. large language models for protein design).

  - More info: mlsb.io

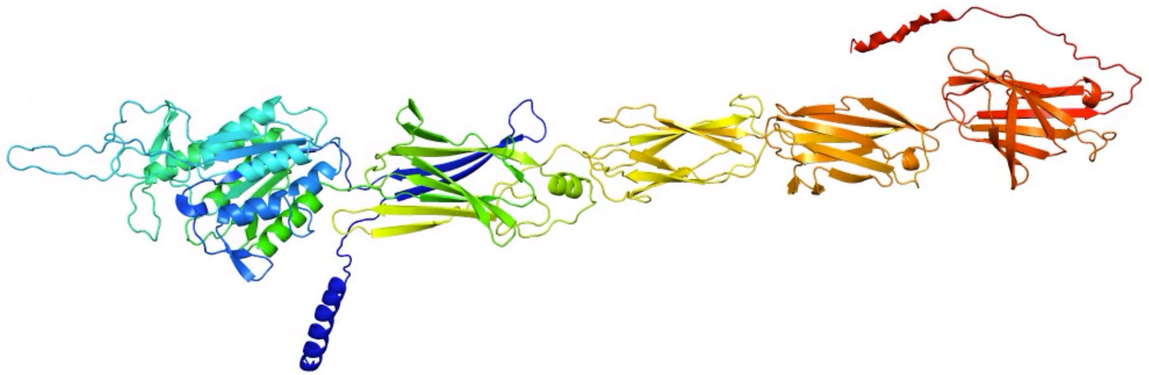
Let’s talk about AlphaFold2

MIT Technology Review

Featured Topics Newsletters Events Podcasts

Artificial intelligence

DeepMind’s protein-folding AI has solved a 50-year-old grand challenge of biology



Recycling iteration 1, block 46
Secondary structure assigned from the final prediction

Vox

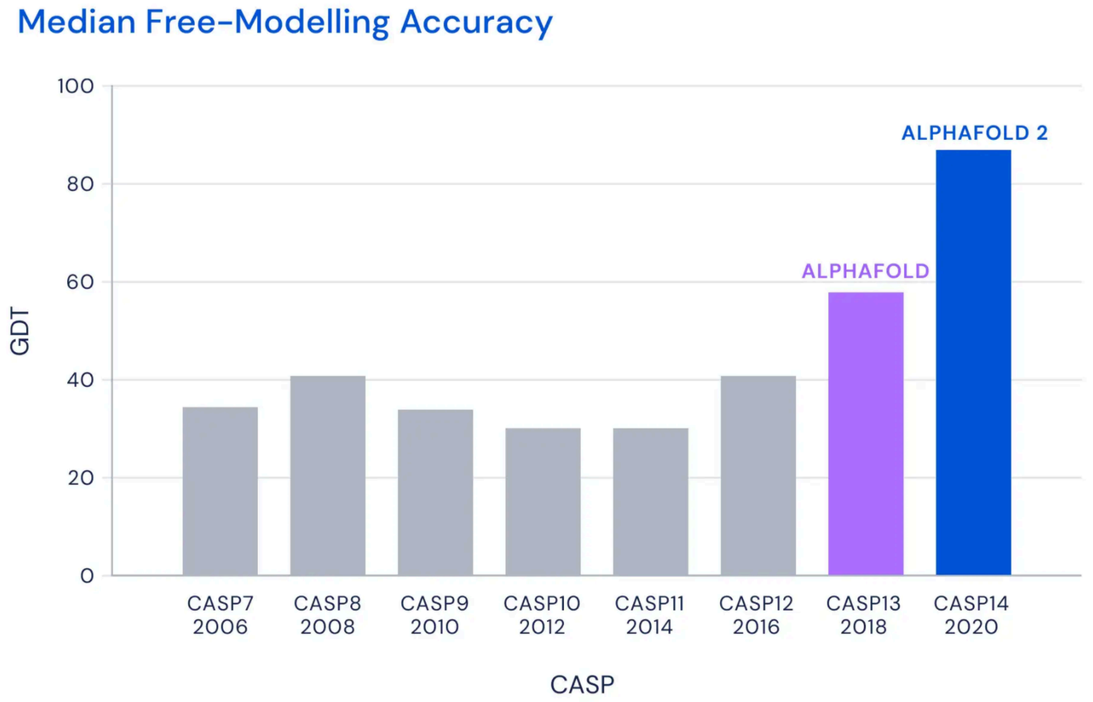
AI has cracked a problem that stumped biologists for 50 years. It’s a huge deal.

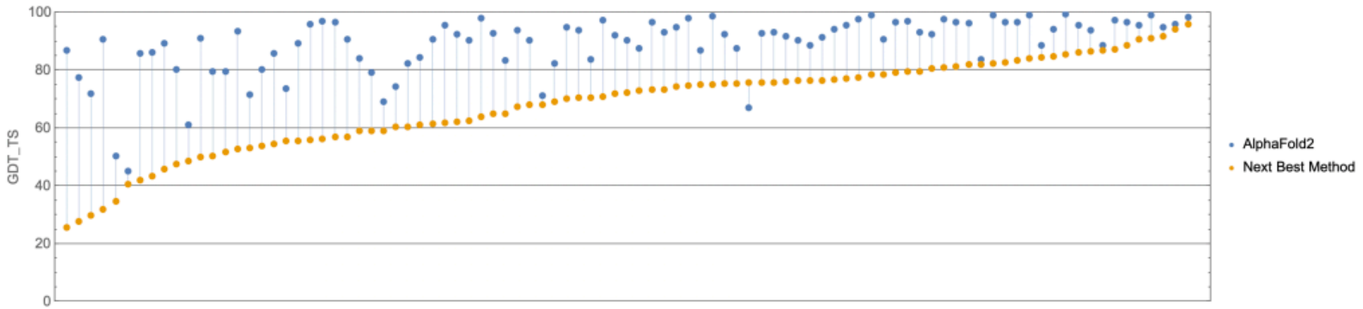
92.4 GDT

AlphaFold performance at CASP14

AlphaFold at CASP14

- CASP[1]: Biannual community-wide blinded competition on ~100 newly solved proteins
- CASP14 press release: “Artificial intelligence solution to a 50-year-old science challenge could ‘revolutionise’ medical research”
- 92.4 median GDT
  - (global distance test, 0-100)
  - 1.6 A RMSD error
  - Above >90 GDT considered within experimental error



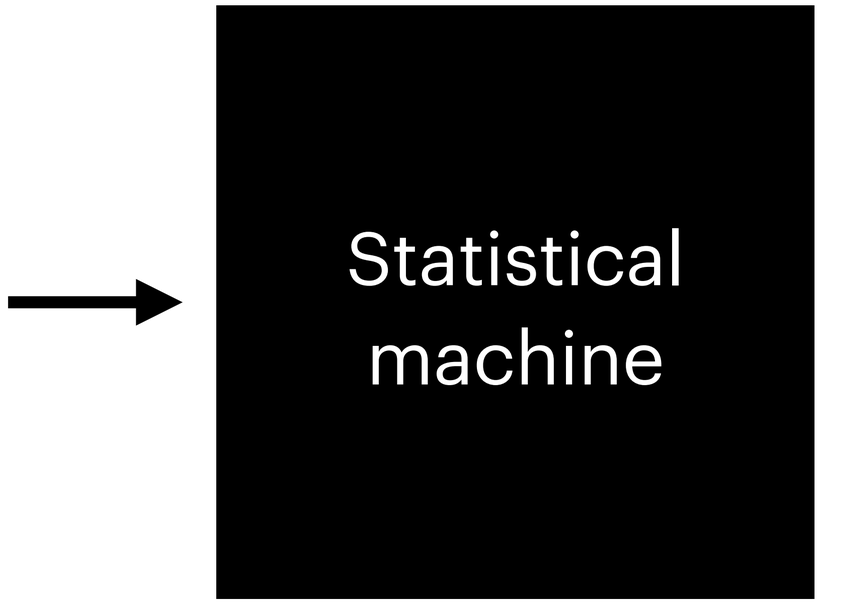


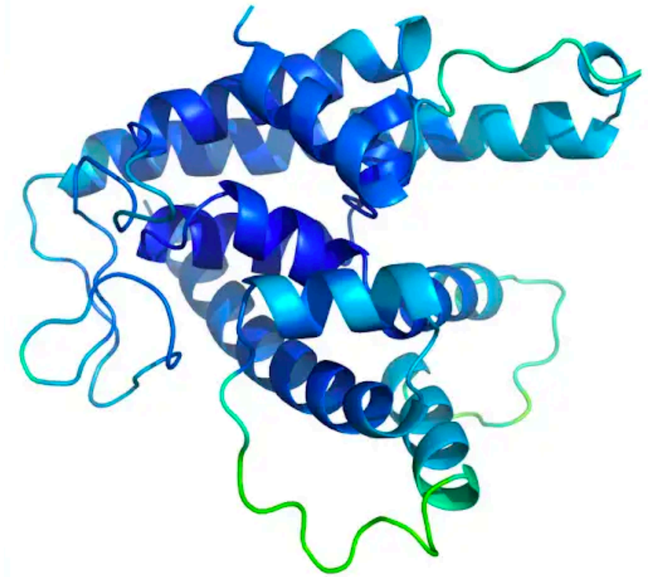
[1] Moult et al 1995

Mohammed AlQuraishi’s blog post, “AlphaFold2 @ CASP14: “It feels like one’s child has left home.”

Inside the AlphaFold system

Input sequence
MRKPRTPFTT...





3D structure

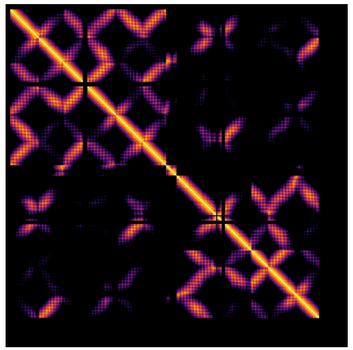
Inside the AlphaFold system

Input sequence
MRKPRTPFTT…

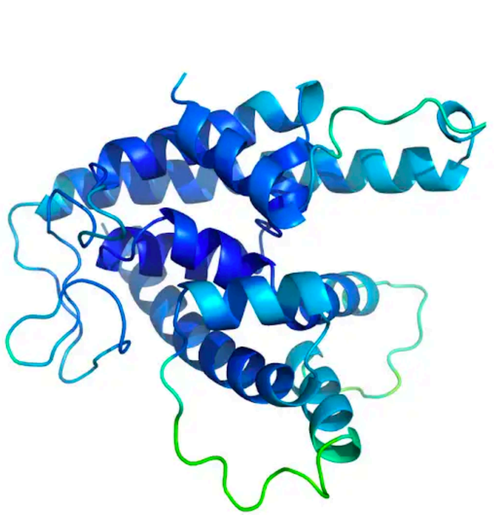
MSA

MRKPRTPFTT…
MRKPRSPFTT…
MRKPATPFTT…
MRKPATPFST…
MRKPRTPFTS…
…

Templates



Statistical machine



3D structure

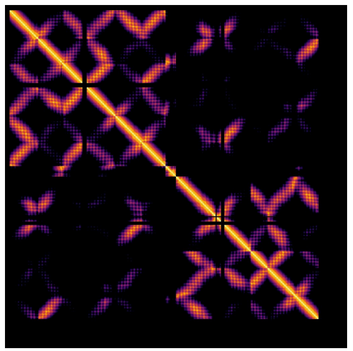
Inside the AlphaFold system

Input sequence
MRKPRTPFTT…

MSA

MRKPRTPFTT…
MRKPRSPFTT…
MRKPATPFTT…
MRKPATPFST…
MRKPRTPFTS…
…

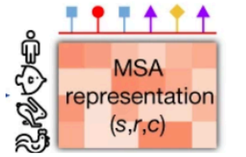
Templates

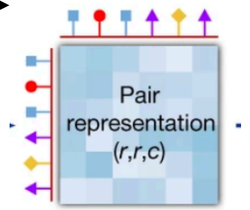


Trunk

“Evoformer”

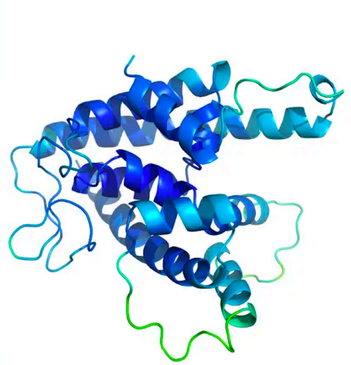
Latent





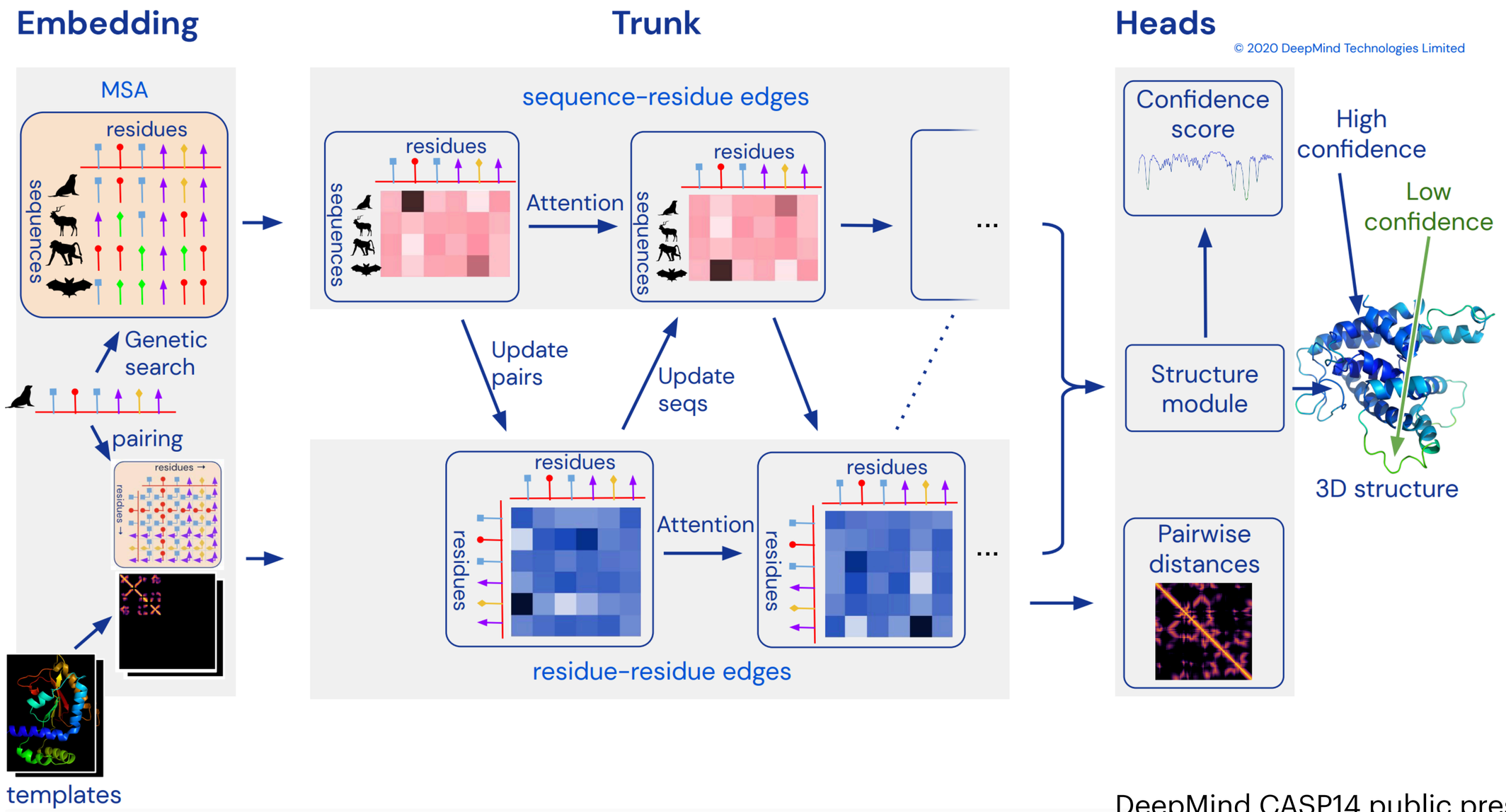
Head

“Structure Module”



3D structure

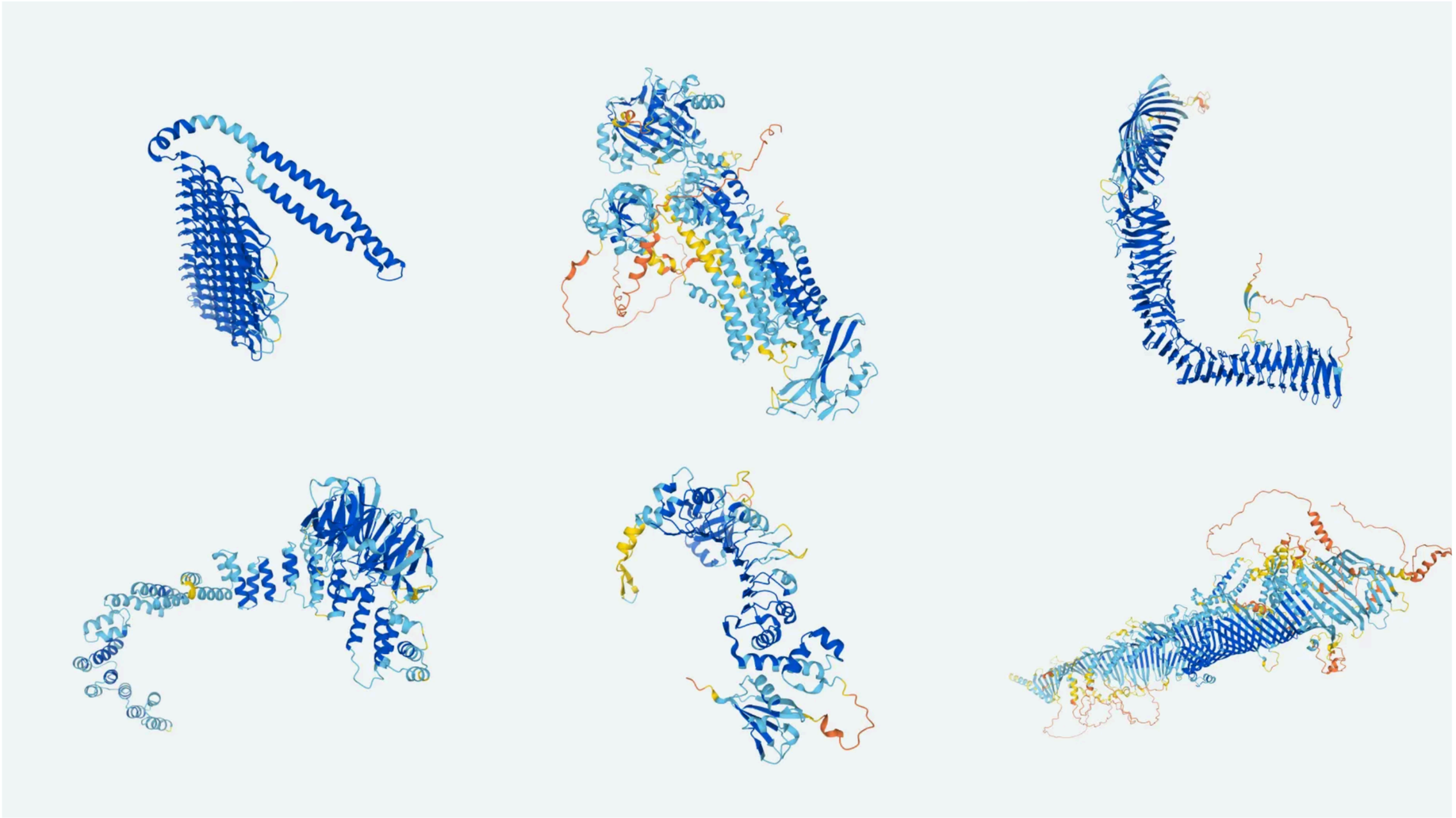
Inside the AlphaFold system



DeepMind CASP14 public presentation

A broad liberation of 3D structure

Before: ~100k unique structures — After: >350k predictions today, all ~100M UniProt sequences

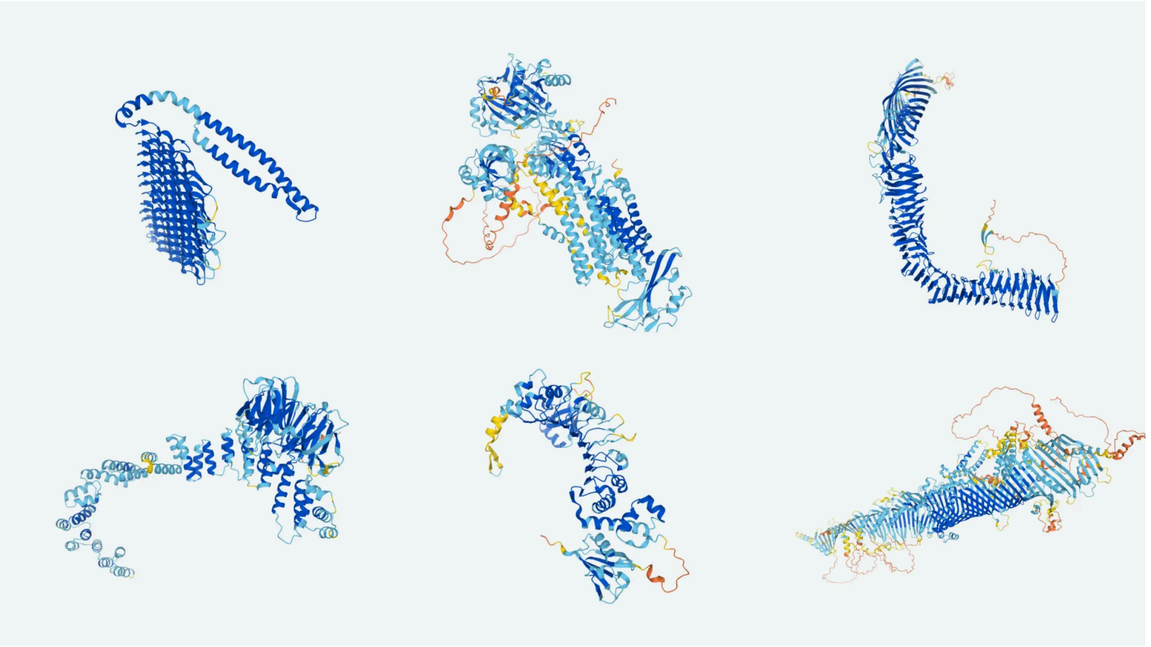


Tunyasuvunakool et al, 2021

A broad liberation of 3D structure

Before: ~100k unique structures — After: >350k predictions today, all ~100M UniProt sequences

- Whole proteome coverage for humans and 20 other model organisms
- Predicted Local Distance Difference Test score (pLDDT) as a well-calibrated measure of confidence
- State-of-the-art predictor of disorder?

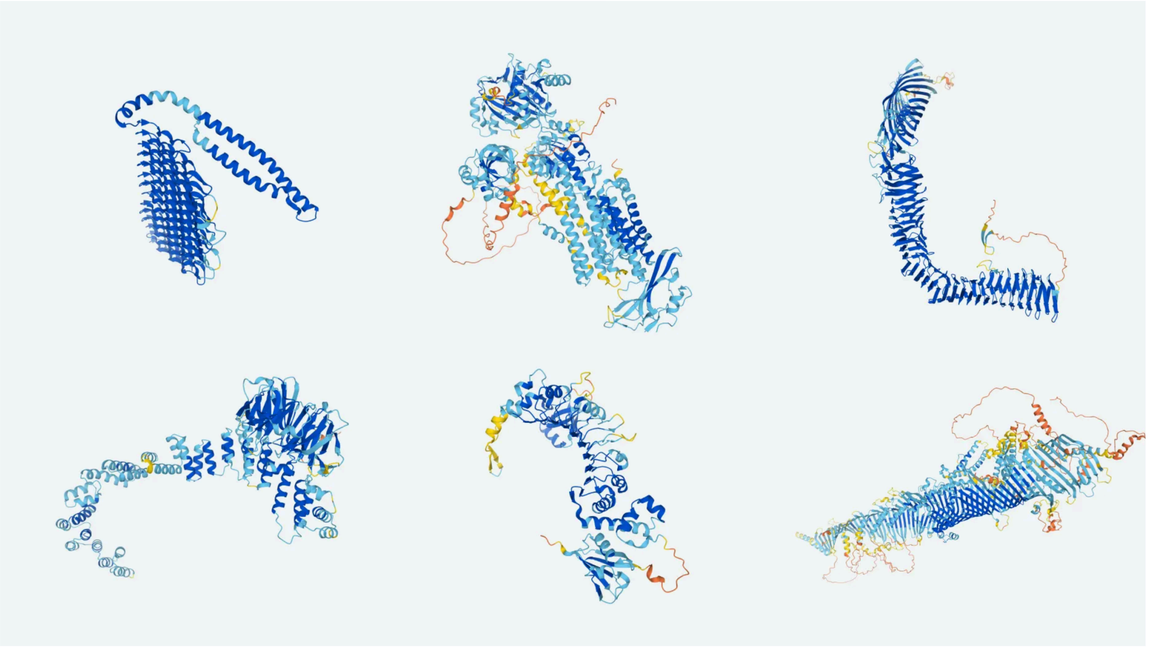


Tunyasuvunakool et al, 2021

A broad liberation of 3D structure

Before: ~100k unique structures — After: >350k predictions today, all ~100M UniProt sequences

- Whole proteome coverage for humans and 20 other model organisms
- Predicted Local Distance Difference Test score (pLDDT) as a well-calibrated measure of confidence
- State-of-the-art predictor of disorder?



Tunyasuvunakool et al, 2021

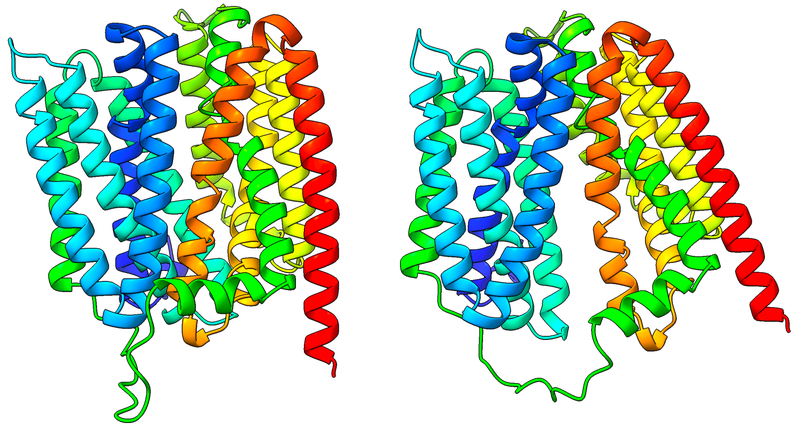
Coverage Statistics (pulled from Akdel et al, biorXiv)

- Confident predictions (pLDDT > 0.7):
  - 27% for *P. falciparum*
  - 77% for *E. coli*
- Highly confident predictions (pLDDT > 0.9) for 25% of all residues
- ~25% of residues of the proteomes covered with novel and confident predictions

Scope and limitations

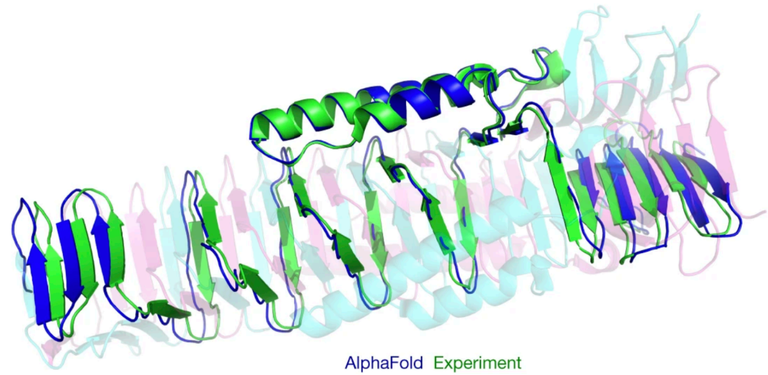
- Machine learning (defn.): Learning patterns from *data*
- The protein sequence to structure prediction problem is *underspecified*

Multiple conformations

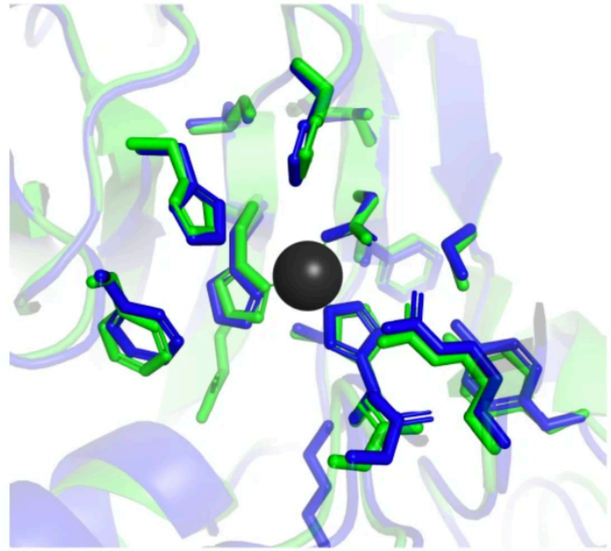


CASP target T1024

A homotrimeric complex



Ligands/ions



Jumper et al, 2021

Outlook for the post-AlphaFold era

- “Discovery” of performant neural network architectures for reasoning over protein sequences and structures

Outlook for the post-AlphaFold era

- “Discovery” of performant neural network architectures for reasoning over protein sequences and structures
- Many interesting problems remain
  - Multiple conformations and dynamics
  - Protein design
  - Interactions with DNA/RNA/small molecules

Outlook for the post-AlphaFold era

- “Discovery” of performant neural network architectures for reasoning over protein sequences and structures
- Many interesting problems remain
  - Multiple conformations and dynamics
  - Protein design
  - Interactions with DNA/RNA/small molecules
- Protein structure prediction vs. protein structure determination
  - Close the loop? Deliberate experimental design?

Overview of selected topics

- Weeks 2-3: Protein structure prediction, AlphaFold2
- Week 4: Cryo-EM reconstruction — cryoDRGN
- Week 5: Atomic modeling — ModelAngelo
- Week 6: Protein design I: Inverse folding
- Week 8: Protein design II: Generative modeling of sequence and structure
- Week 9: Protein language modeling
- Week 10: Physics-based modeling — MD simulation, Boltzmann Generators
- Week 11: Geometric deep learning and drug discovery
- Week 13: Structural bioinformatics
- Week 14: Wildcard

The first two papers

- The protein structure prediction component of the *protein folding problem*
- Looking for volunteers to present next week

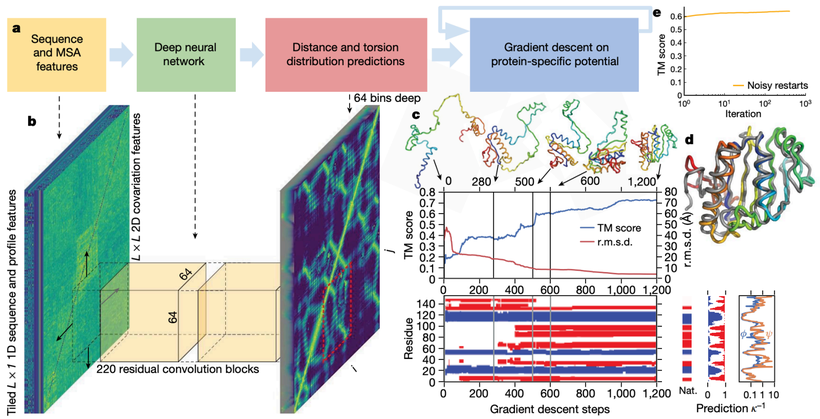
Article

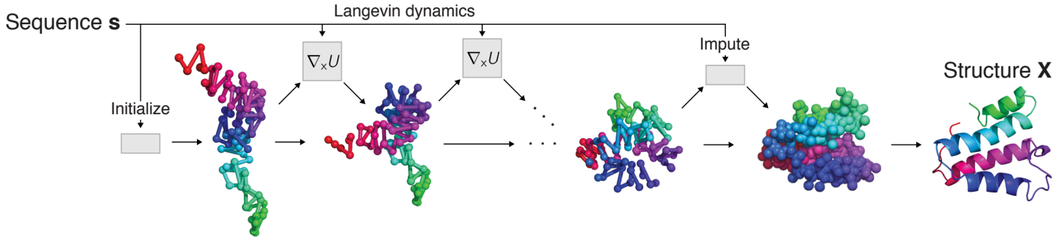
Improved protein structure prediction using potentials from deep learning

https://doi.org/10.1038/s41586-019-1923-7
Received: 2 April 2019
Accepted: 10 December 2019
Published online: 15 January 2020

Andrew W. Senior[1,4]*, Richard Evans[1,4], John Jumper[1,4], James Kirkpatrick[1,4], Laurent Sifre[1,4], Tim Green[1], Chongli Qin[1], Augustin Žídek[1], Alexander W. R. Nelson[1], Alex Bridgland[1], Hugo Penedones[1], Stig Petersen[1], Karen Simonyan[1], Steve Crossan[1], Pushmeet Kohli[1], David T. Jones[2,3], David Silver[1], Koray Kavukcuoglu[1] & Demis Hassabis[1]

Protein structure prediction can be used to determine the three-dimensional shape of a protein from its amino acid sequence[1]. This problem is of fundamental importance as the structure of a protein largely determines its function[2]; however, protein





Published as a conference paper at ICLR 2019

LEARNING PROTEIN STRUCTURE WITH A DIFFERENTIABLE SIMULATOR

John Ingraham[1,*], **Adam Riesselman**[1], **Chris Sander**[1,2,3], **Debora Marks**[1,3]
[1]Harvard Medical School [2]Dana-Farber Cancer Institute
[3]Broad Institute of Harvard and MIT

ABSTRACT

The Boltzmann distribution is a natural model for many systems, from brains to materials and biomolecules, but is often of limited utility for fitting data because Monte Carlo algorithms are unable to simulate it in available time. This gap between the expressive capabilities and sampling practicalities of energy-based models is exemplified by the protein folding problem, since energy landscapes underlie contemporary knowledge of protein biophysics but computer simulations

Paper reading strategies

- Biology journal papers vs. ML conference papers
- What are your current practices?

- Additional consideration when reading papers:
  - What is the historical/social context of this work?
  - (How) did this paper impact the field / other research? Who is citing this work? Why?
  - Think about the differences between carrying out the research and writing the paper. Usually somewhat decoupled.
  - For more ideas on different “roles” in paper reading, see https://colinraffel.com/blog/role-playing-seminar.html