

```
results[seq_id] = { 'A': counts.get('A', 0), 'C': counts.get('C', 0), 'G': counts.get('G', 0), 'T': counts.get('T', 0) }
    return results def gc_content(self):
        """Berechnet den GC-Gehalt (%) der Sequenzen"""
        total_count = len(sequence)
        gc_count = sequence.count('G') + sequence.count('C')
        gc_percentage = (gc_count / total_count) * 100 if total_count > 0 else 0
        gc_results[seq_id] = gc_percentage
        return gc_results def find_open_reading_frames(self, min_length=1):
        """Identifiziert offene Leserahmen (ORFs) in den Sequenzen"""
        start_codon = "ATG" stop_codons = ["TAA", "TGA", "TTC"]
        for seq_id, sequence in self.sequences.items():
            orfs[seq_id] = []
            for frame in range(3):
                pos = 0
                while pos < len(sequence) - 2:
                    codon = sequence[pos:pos+3]
                    if codon == start_codon:
                        for stop_codon in stop_codons:
                            stop_pos = pos + sequence.find(stop_codon, pos)
                            if stop_pos != -1:
                                stop_pos += 3
                                orf_length = stop_pos - pos
                                orfs[seq_id].append((pos, stop_pos, orf_length))
                                break
                    pos += 3
            pos += 3
        return orfs def plot_gc_content(self):
        """Visualisiert den GC-Gehalt mit Matplotlib"""
        gc_data = self.gc_content()
        seq_ids = list(gc_data.keys())
        gc_values = list(gc_data.values())
        plt.figure(figsize=(10, 5))
        plt.bar(seq_ids, gc_values)
        plt.xlabel("Sequenz ID")
        plt.ylabel("GC-Gehalt (%)")
        plt.title("GC-Gehalt der DNA-Sequenzen")
        plt.tight_layout()
        plt.show() def plot_nucleotide_distribution(self):
        """Erstellt Balkendiagramme der Nukleotidverteilung"""
        distributions = self.nucleotide_distribution()
        seq_ids = list(distributions.keys())
        distributions = list(distributions.values())
        fig, ax = plt.subplots(figsize=(10, 6))
        width = 0.2 # Balkenbreite für die Gruppen
        for i, seq_id in enumerate(seq_ids):
            x_pos = i * 1.5 + width / 2
            ax.bar([x_pos], distributions[i])
        ax.set_xticklabels(seq_ids, rotation=45)
        ax.set_ylabel("Anzahl der Nukleotide")
        ax.set_title("Nukleotidverteilung")
        plt.tight_layout()
        plt.show() def main():
    parser = argparse.ArgumentParser(description="Bioinformatics tool for DNA sequences")
    parser.add_argument("fasta_file", help="Pfad zur FASTA-Datei")
    parser.add_argument("--gc", action="store_true", help="GC-Gehalt berechnen")
    parser.add_argument("--orf", action="store_true", help="ORFs identifizieren")
    parser.add_argument("--plot_gc", action="store_true", help="GC-Gehaltverteilung visualisieren")
    args = parser.parse_args()
    if args.gc:
        print(f"GC-Gehalt der Sequenzen: {gc_content(args.fasta_file)}%")
    if args.orf:
        print(f"Offene Leserahmen (ORFs) in der Sequenz: {find_open_reading_frames(args.fasta_file)}")
    if args.plot_gc:
        plot_gc_content(args.fasta_file)
```

Protein modeling

Dr. Amanda Souza Câmara

Introduction to Bioinformatics

Spring school for early stage plant scientists
IPK Gatersleben · May 13-16, 2025



Schedule

- 13:30 – 15:00 Part 1 – Theory on protein structure and the challenge of solving it
- 15:00 – 15:30 Coffee break
- 15:30 – 17:00 Part 2 – Hands-on modelling, visualizing and analyzing with AlphaFold and PyMOL

Everything, including this presentation, is on GitHub:

<https://github.com/InsilicoGenebankProteomics/ProteinModeling-Plant2020SpringSchool>

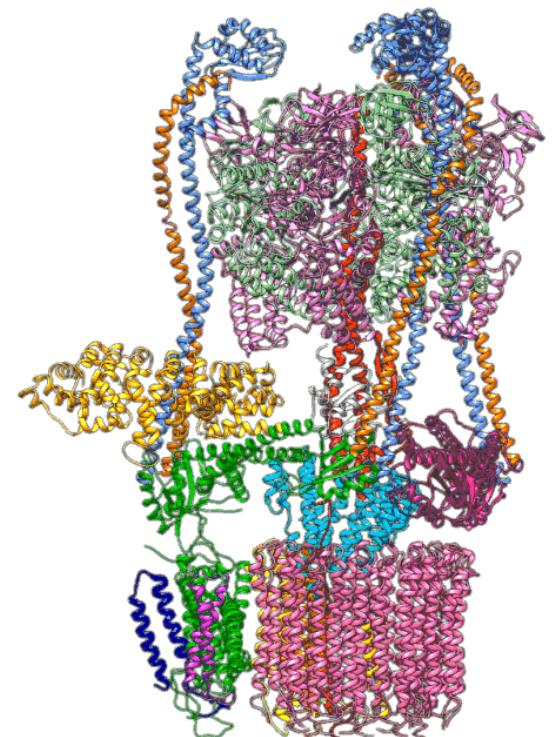
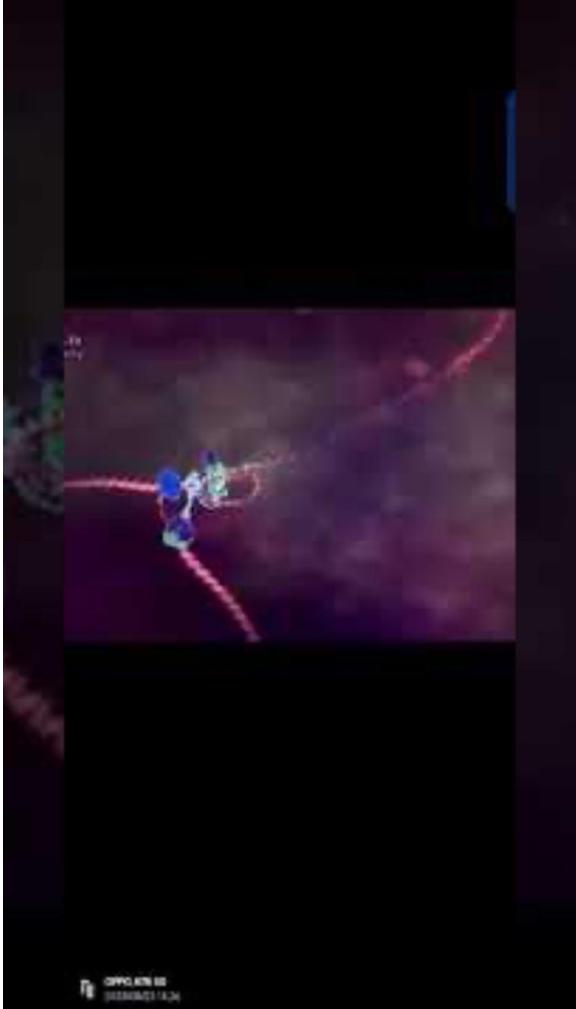
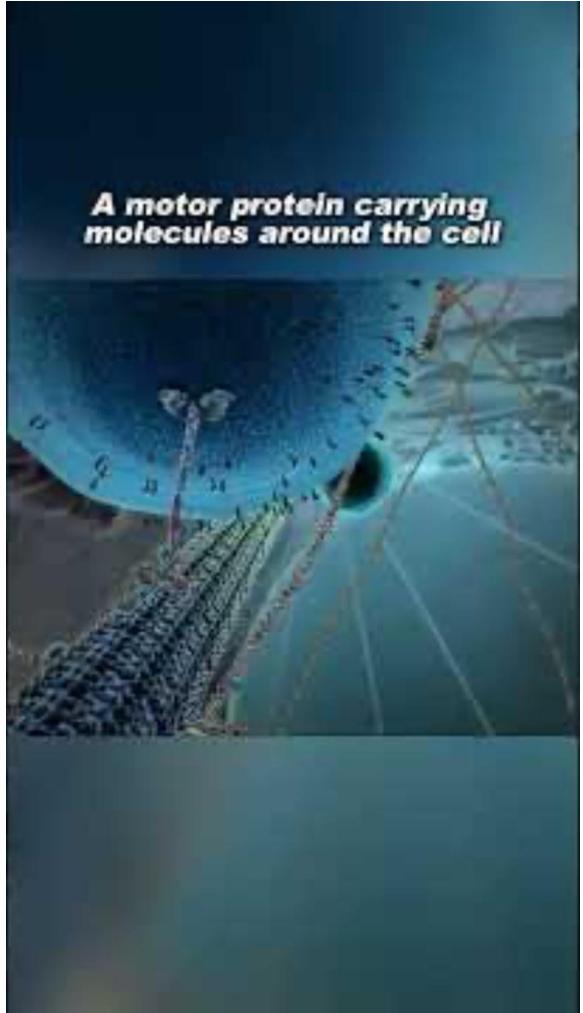
Bioinformatics from 1D

... to 3D!

Goals for part 1

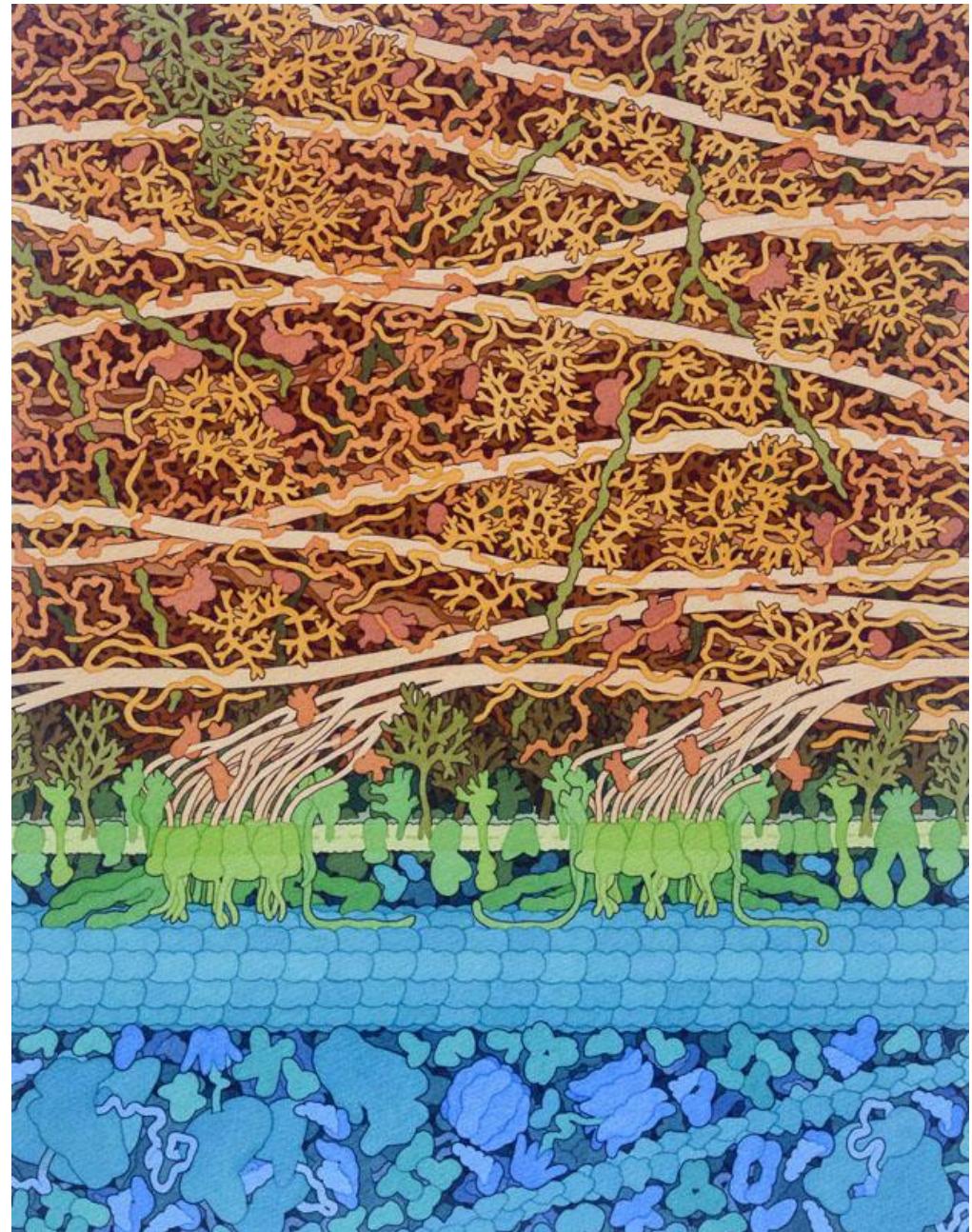
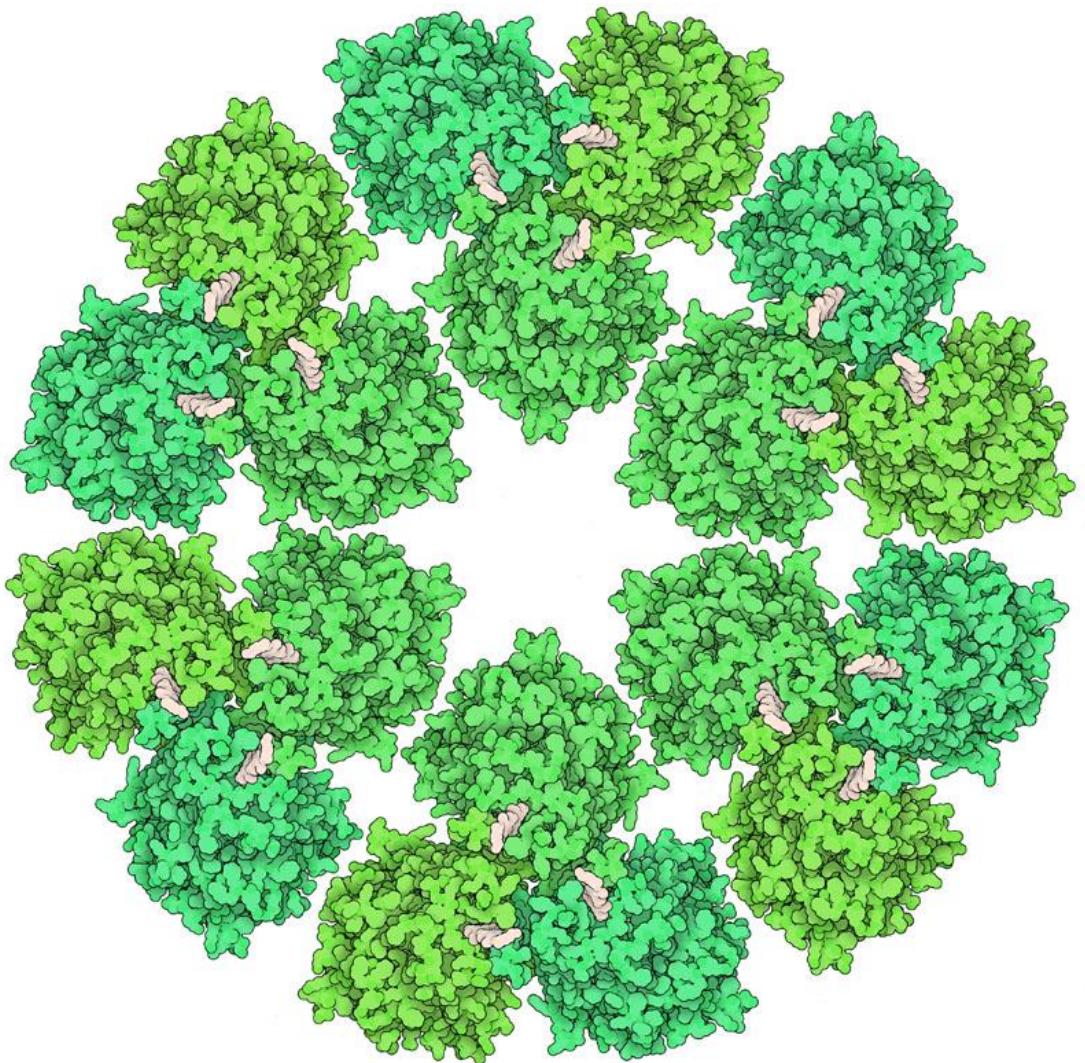
- Diversity of proteins shapes and functions
- Relationship between structure and function, and between sequence and structure
- Main characteristics of protein folding and structure
- The Protein Data Bank and the coordinates file
- AlphaFold, understand and interpret

Proteins variety of shapes and functions

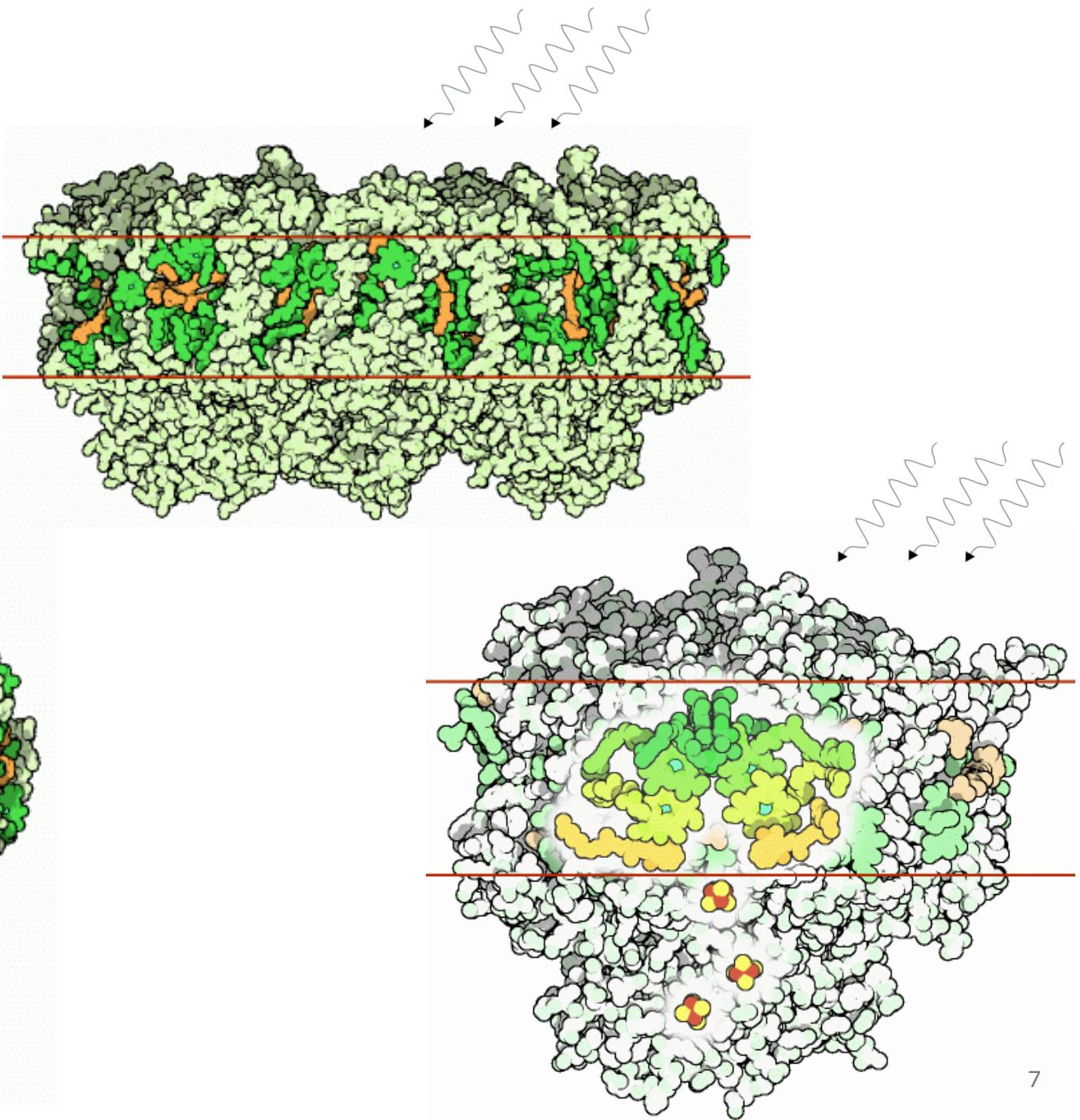
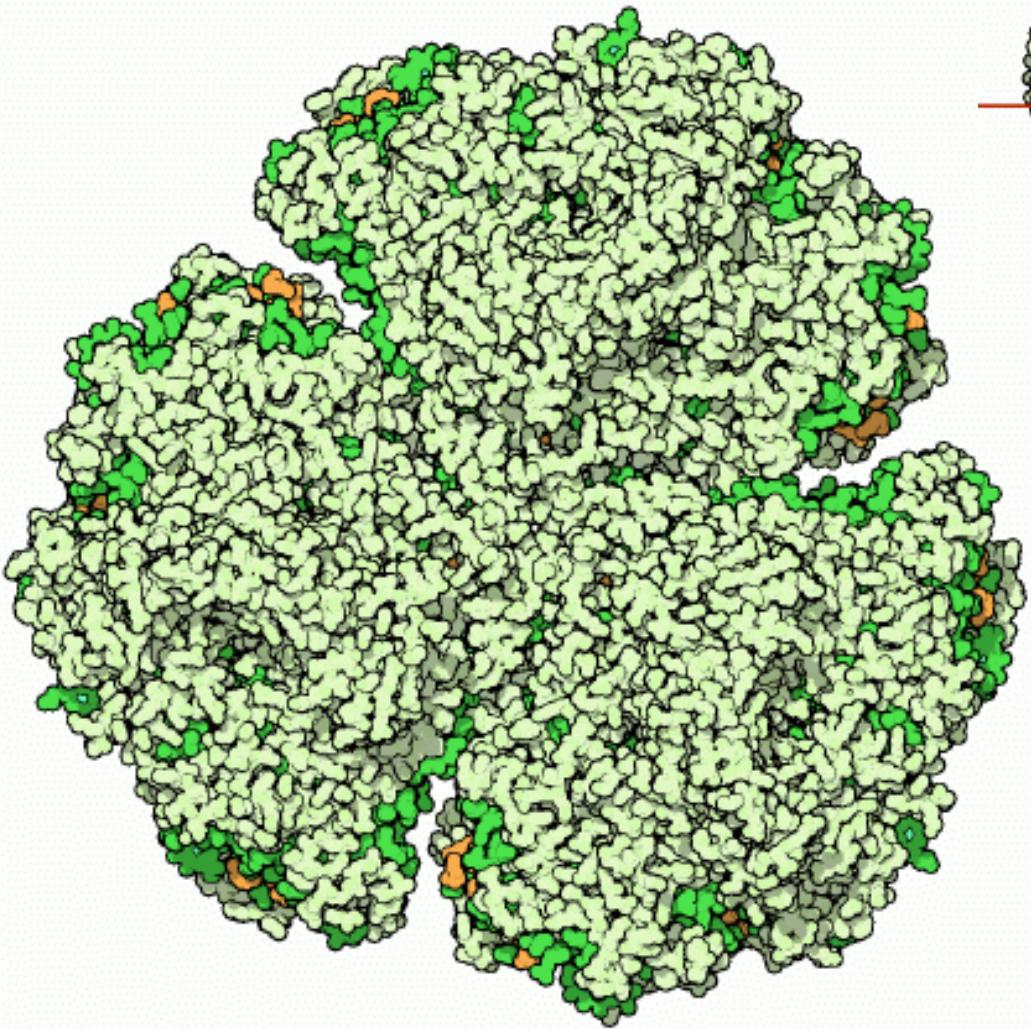


ATP synthase or
Ions pump 5

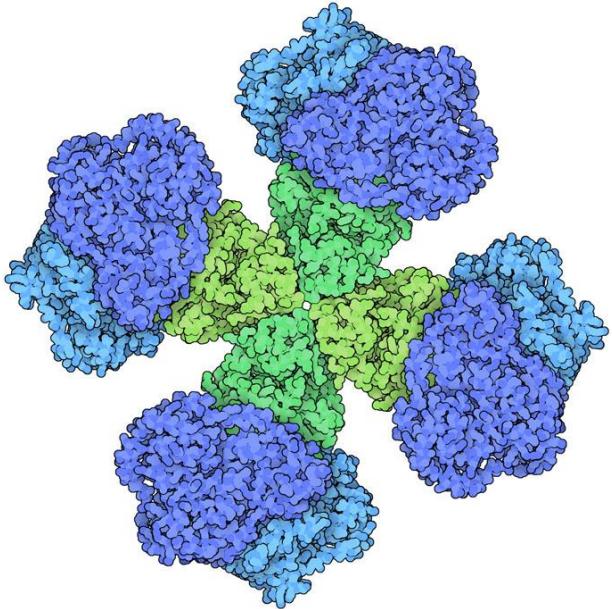
Cellulose synthase



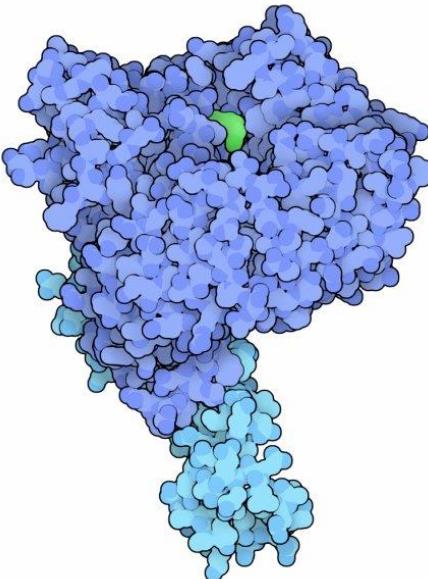
Photosystem I



Other plant specific proteins

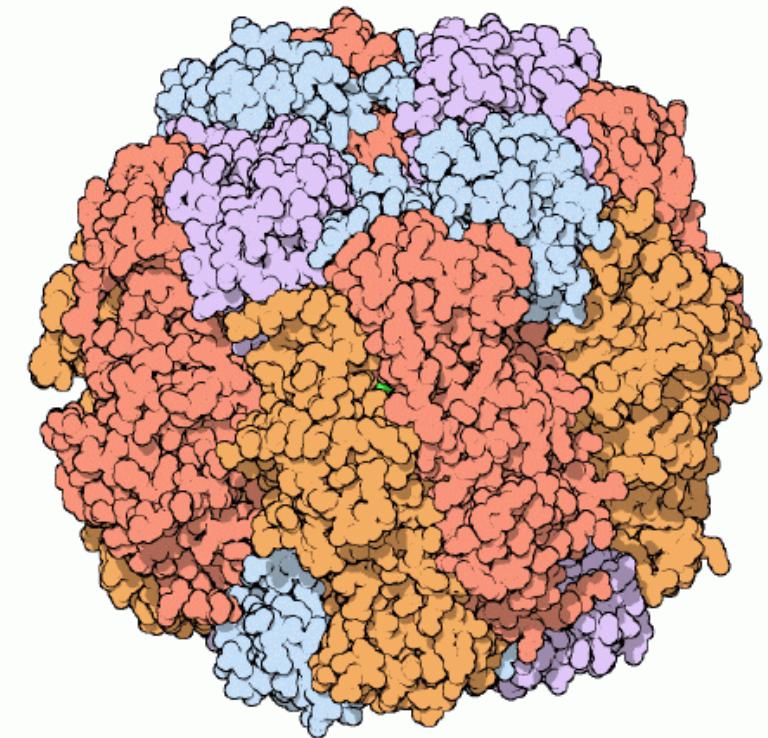
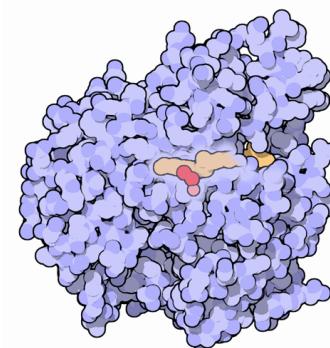


Acetohydroxyacid Synthase
Synthesis of three essential amino acids



TIR ubiquitin ligase with auxin (green)

*Carotenoid oxygenase,
with a carotenoid molecule.*



*Rubisco fixes atmospheric
carbon dioxide into
bioavailable sugar
molecules*

Proteins diversity

Molecular Machinery: A Tour of the Protein Data Bank

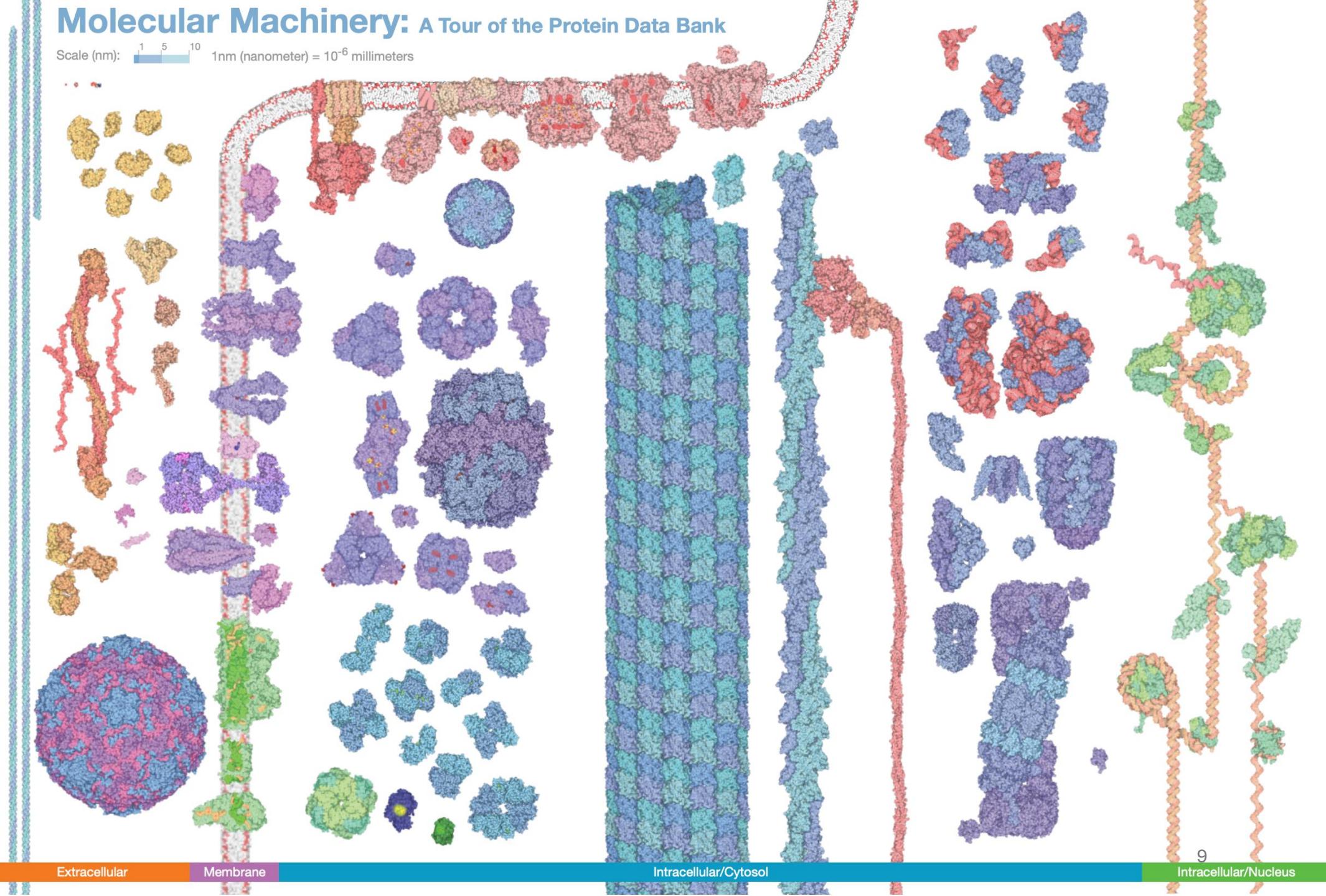
Scale (nm):

1

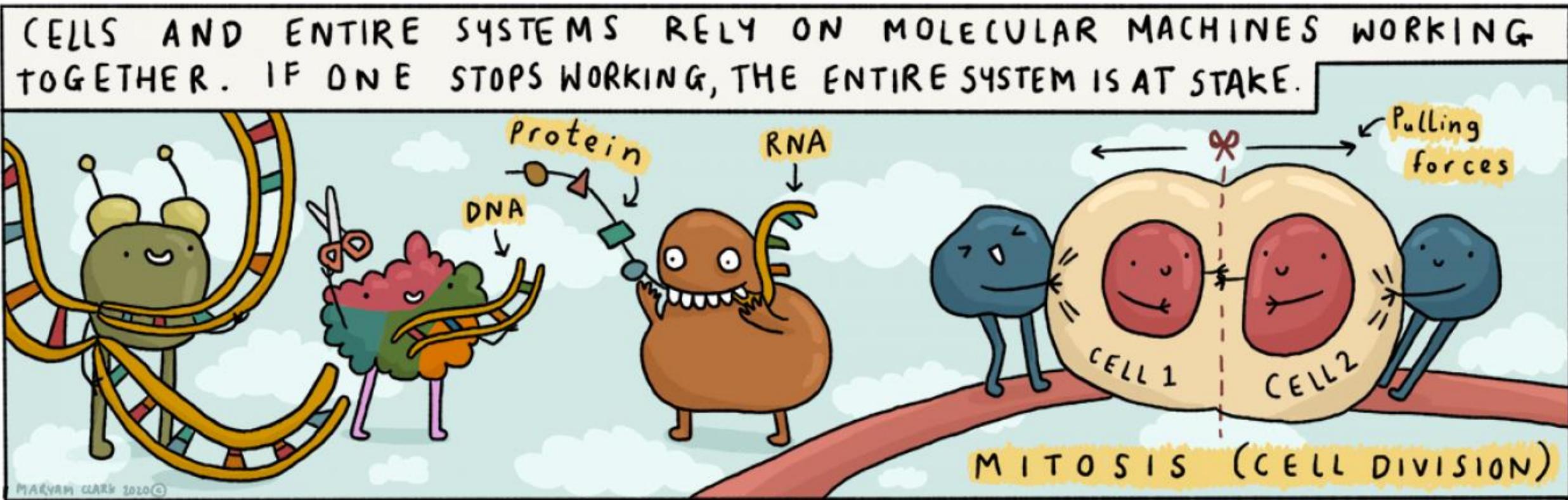
5

10

1nm (nanometer) = 10^{-6} millimeters



Lifeless molecular machines that perform vital processes



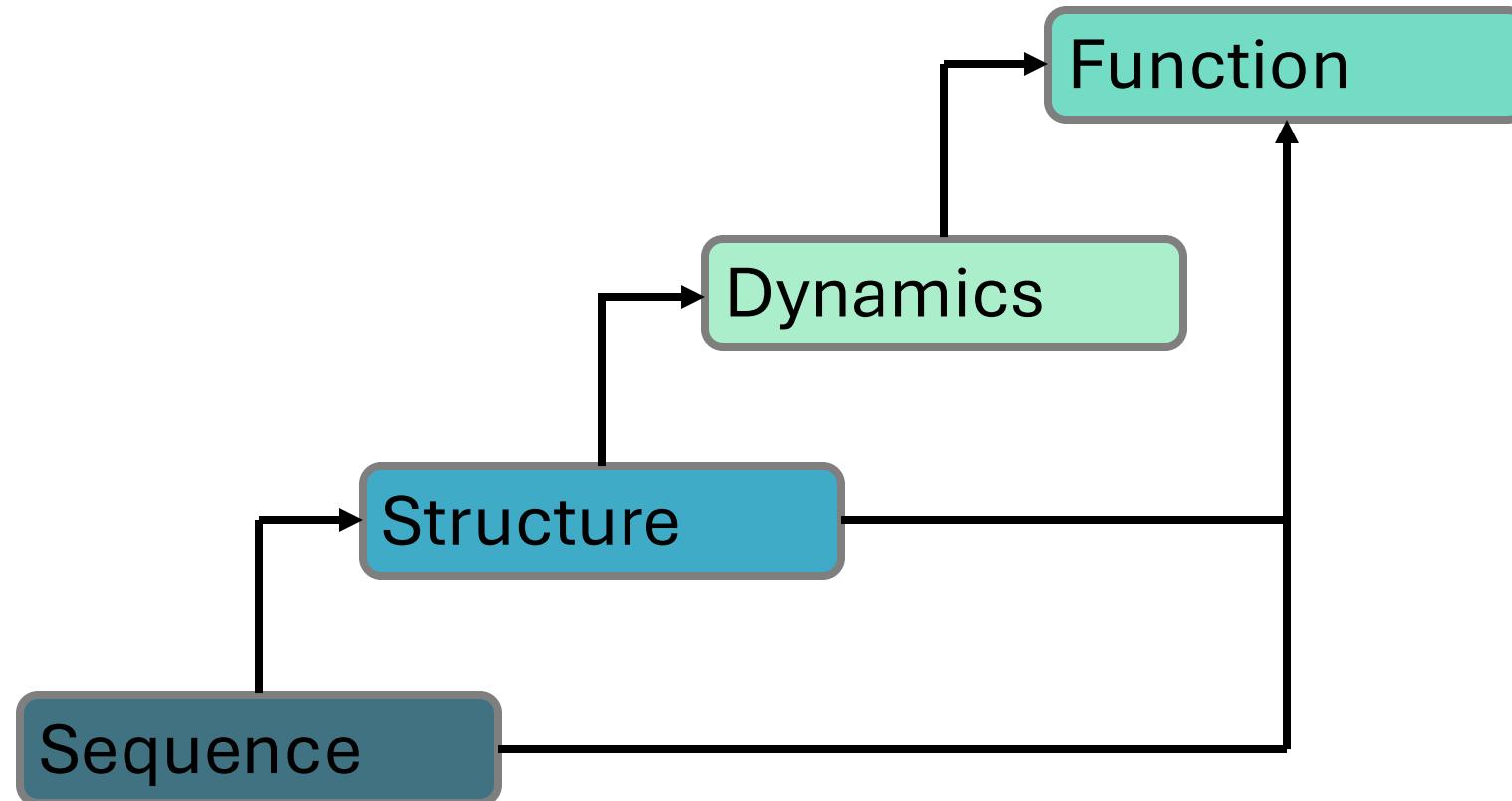
<https://www.ucl.ac.uk/>

They move around and move themselves according to their physical and chemical properties.

*“Everything that living things do
can be understood in terms of the
jigglings and wigglings of atoms.”*

Richard Feynman

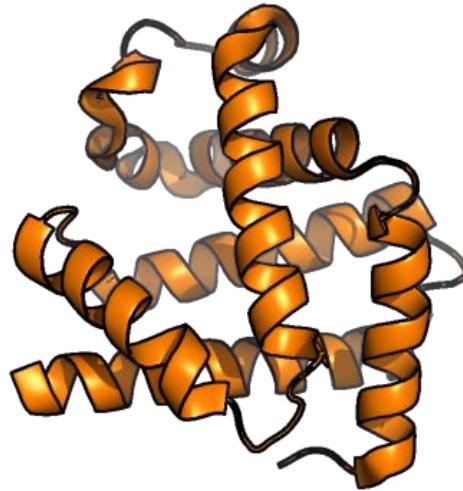
Pillars of function



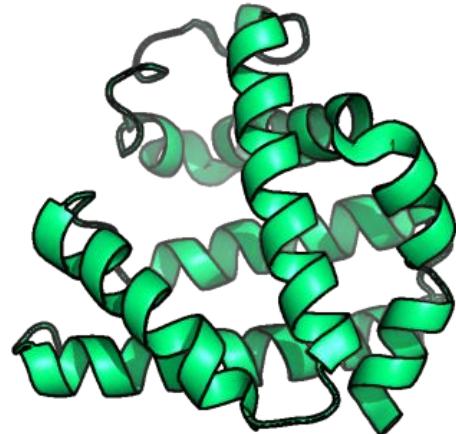
But nothing is straight forward.

Low sequence similarity but high structural homology

Human myoglobin



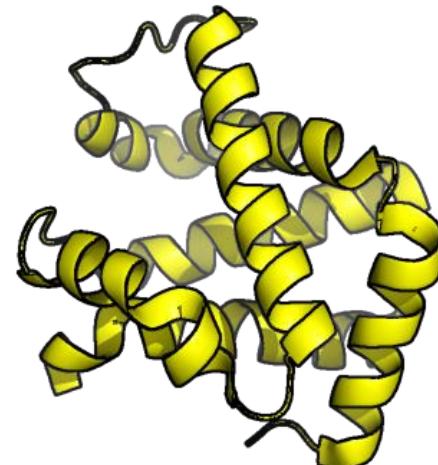
Pigeon myoglobin
25% sequence identity



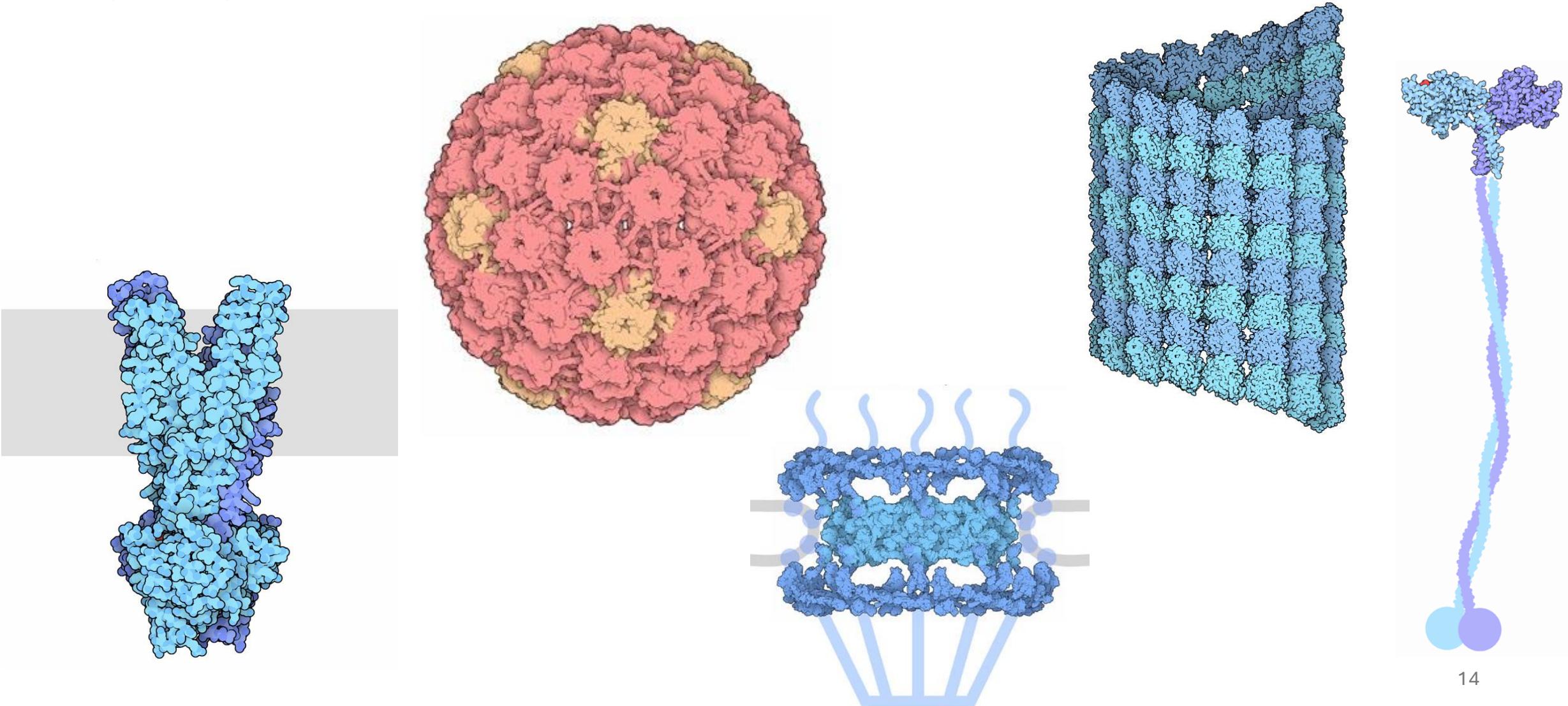
African elephant myoglobin
80% sequence identity



Black-fin tuna myoglobin
45% sequence identity

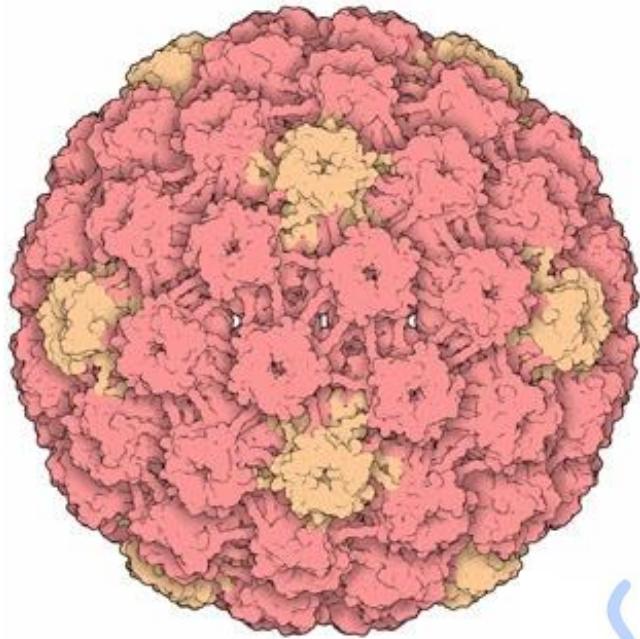
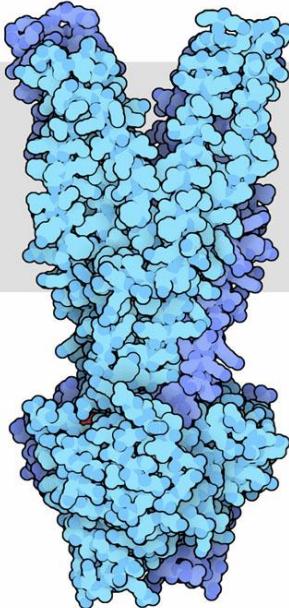


Some structures are of obvious function,
or not...

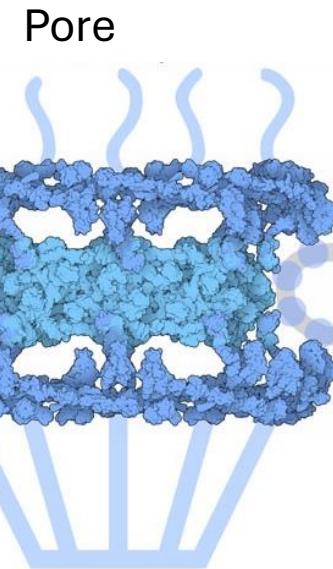


Some structures are of obvious function, or not...

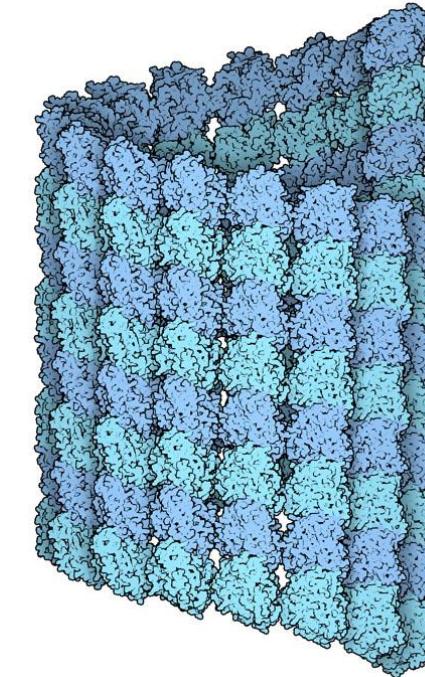
Transporter



Virus capsid

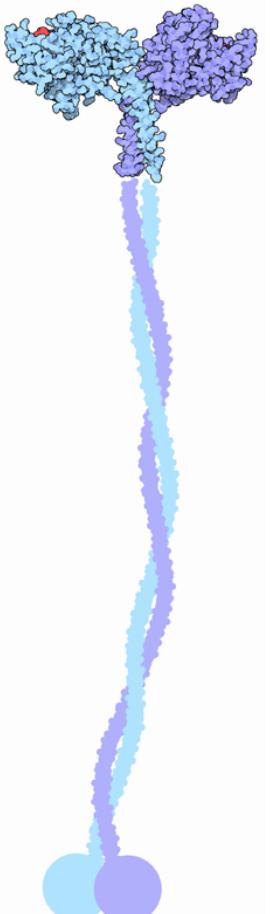


Pore

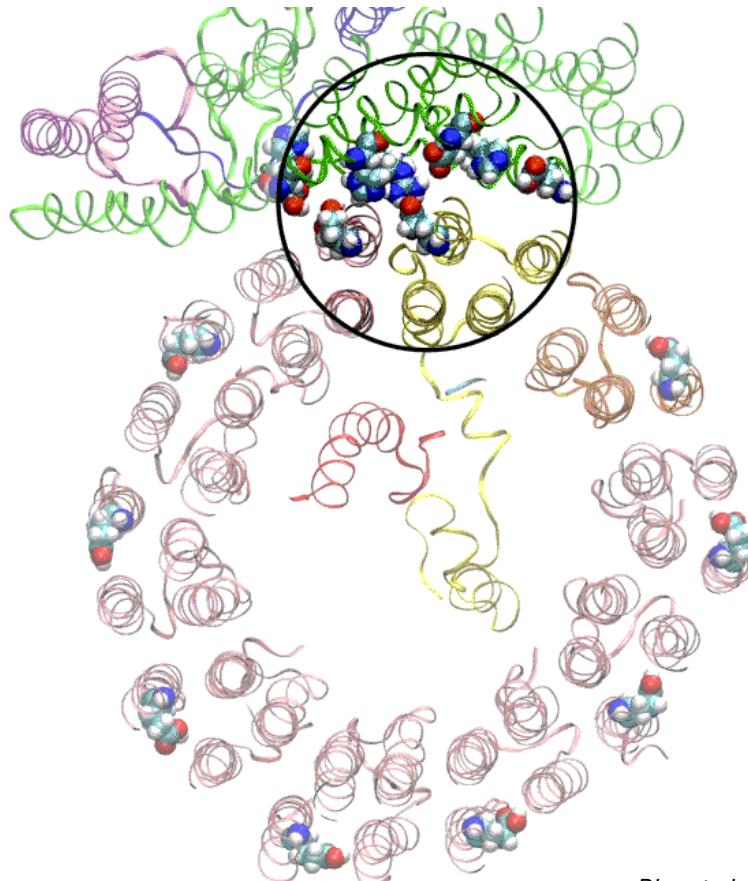


Microtubule

Cargo
transporter

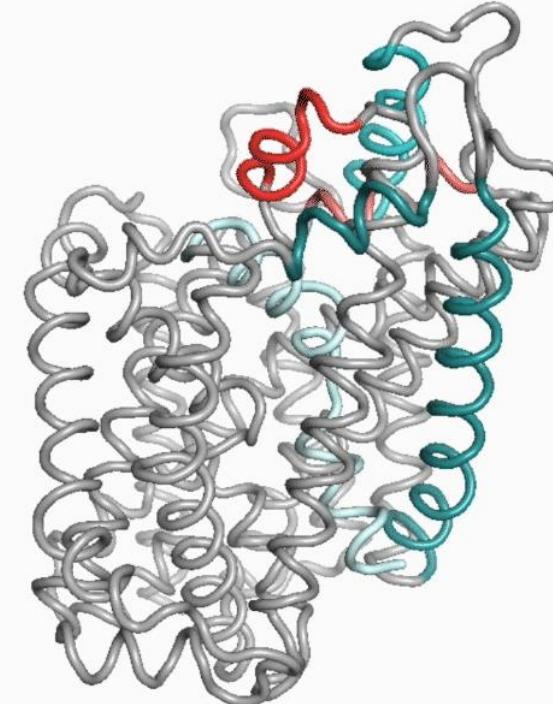


Dynamics can be calculated to see them in action



Proton pump

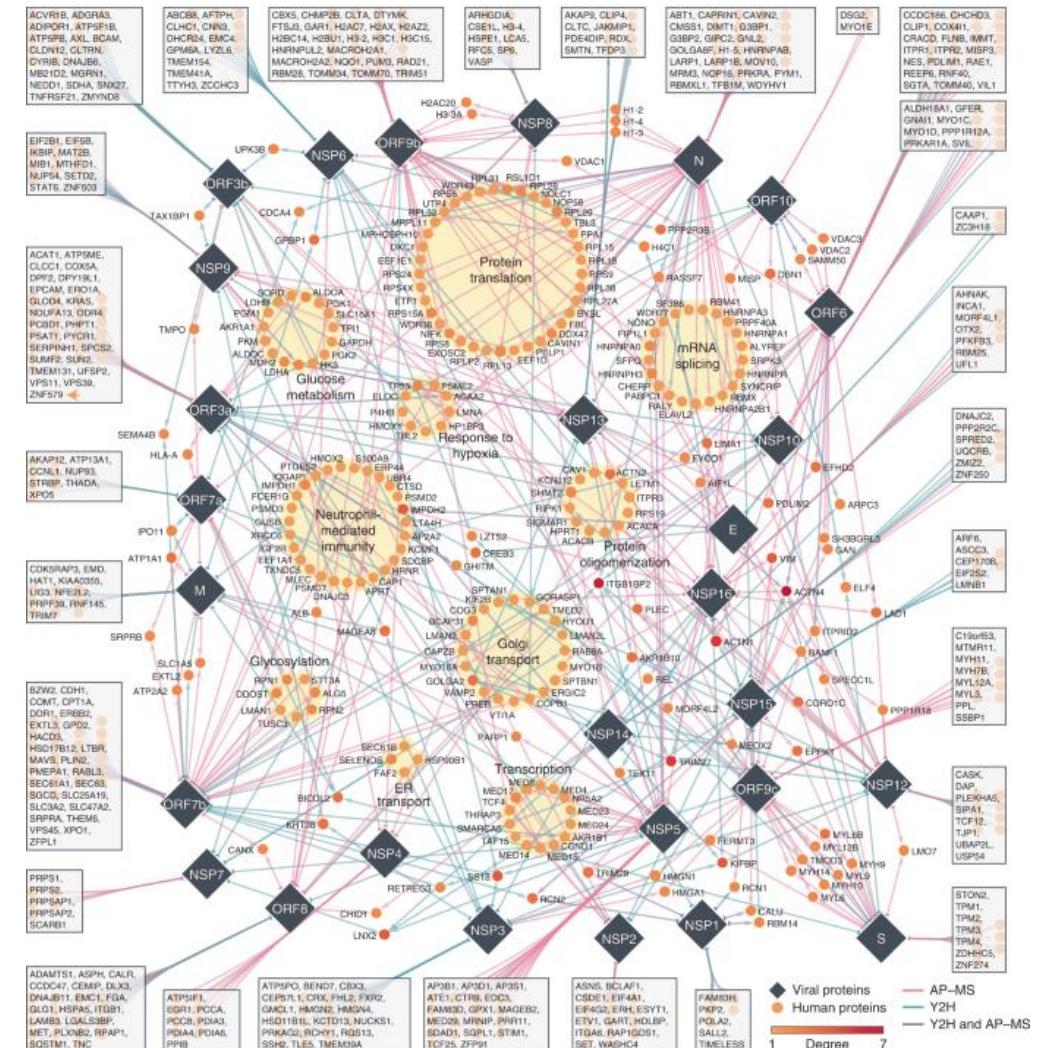
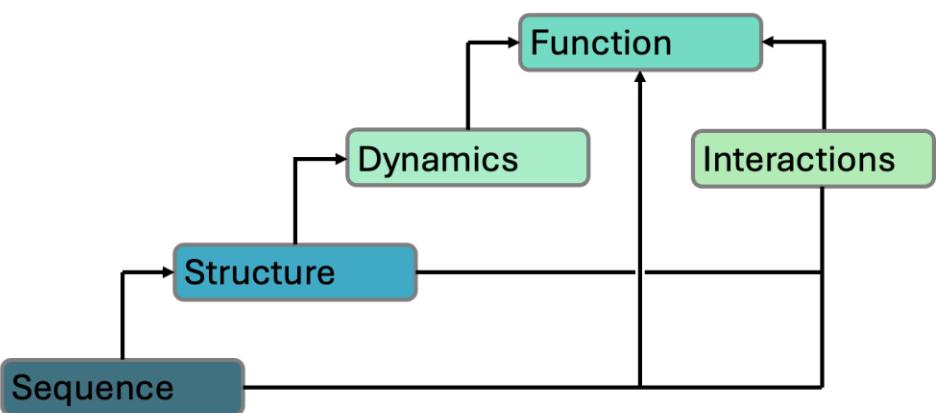
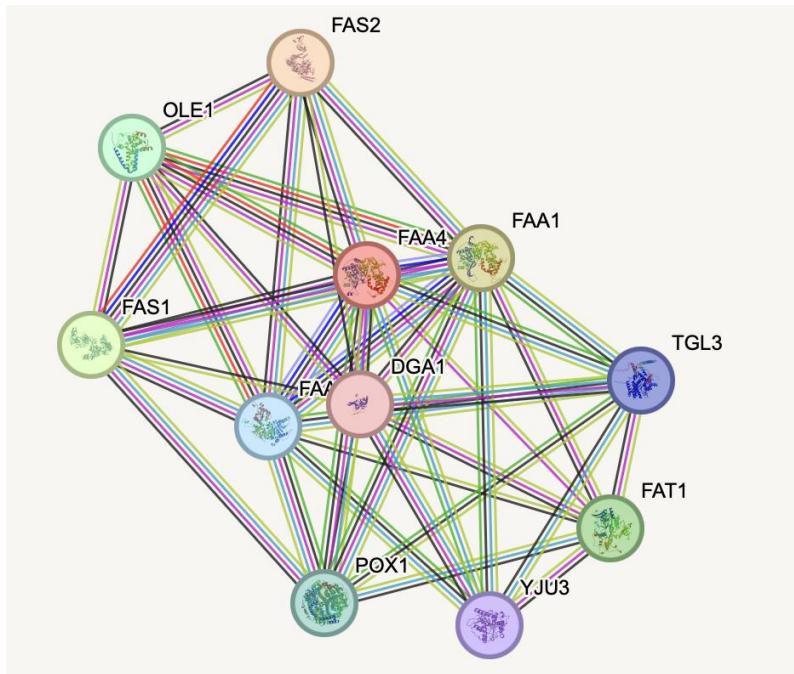
But some previous knowledge is required, and the computational costs are big.



<https://phys.org/news/2020-03-neuroscientists-important-protein-brain.html>

Neurotransmitter transporter

Interactome



Zhou et al., 2023

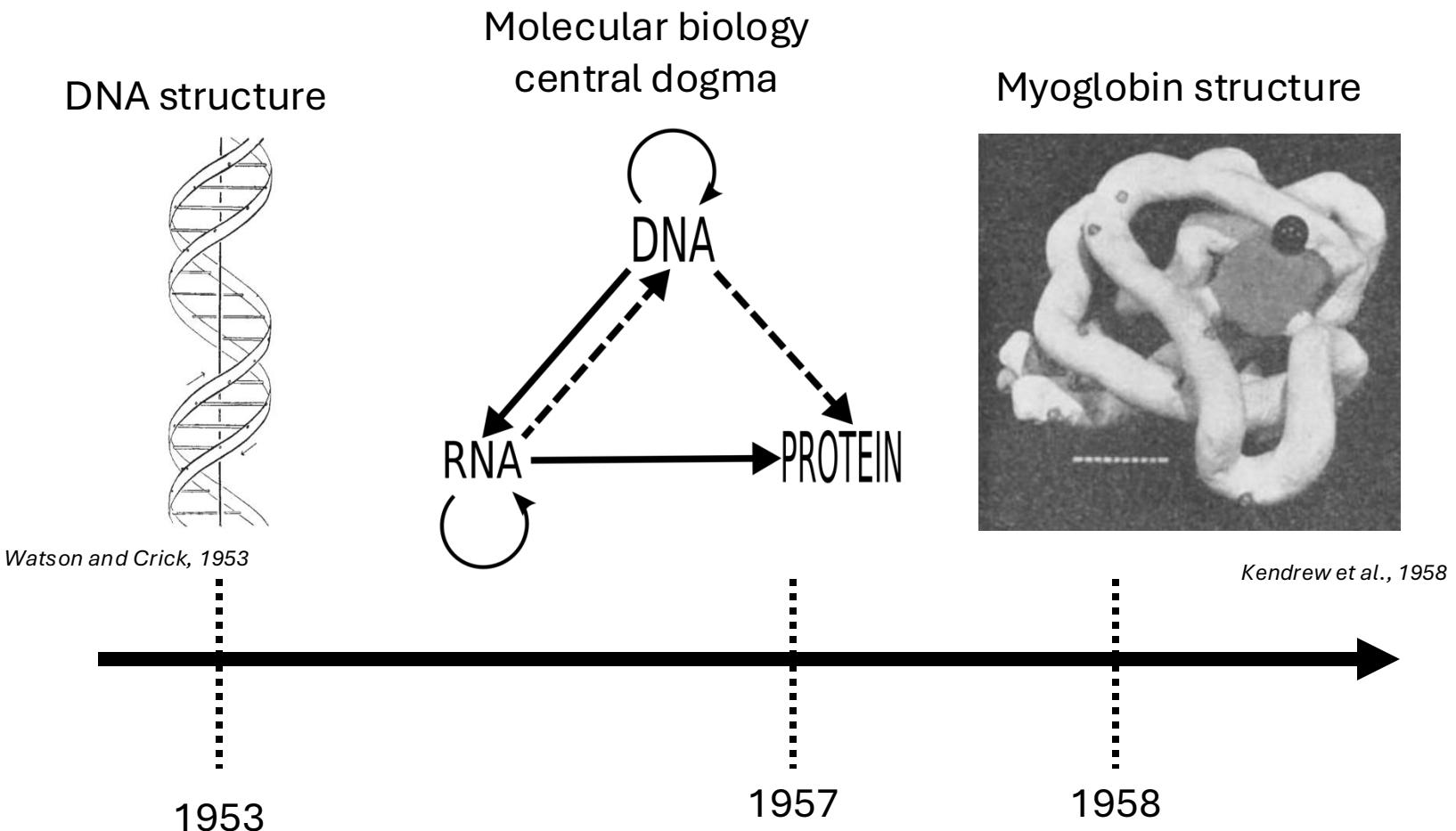
SARS-CoV-2 proteins
 Human proteins

Solve the structure, solve the function

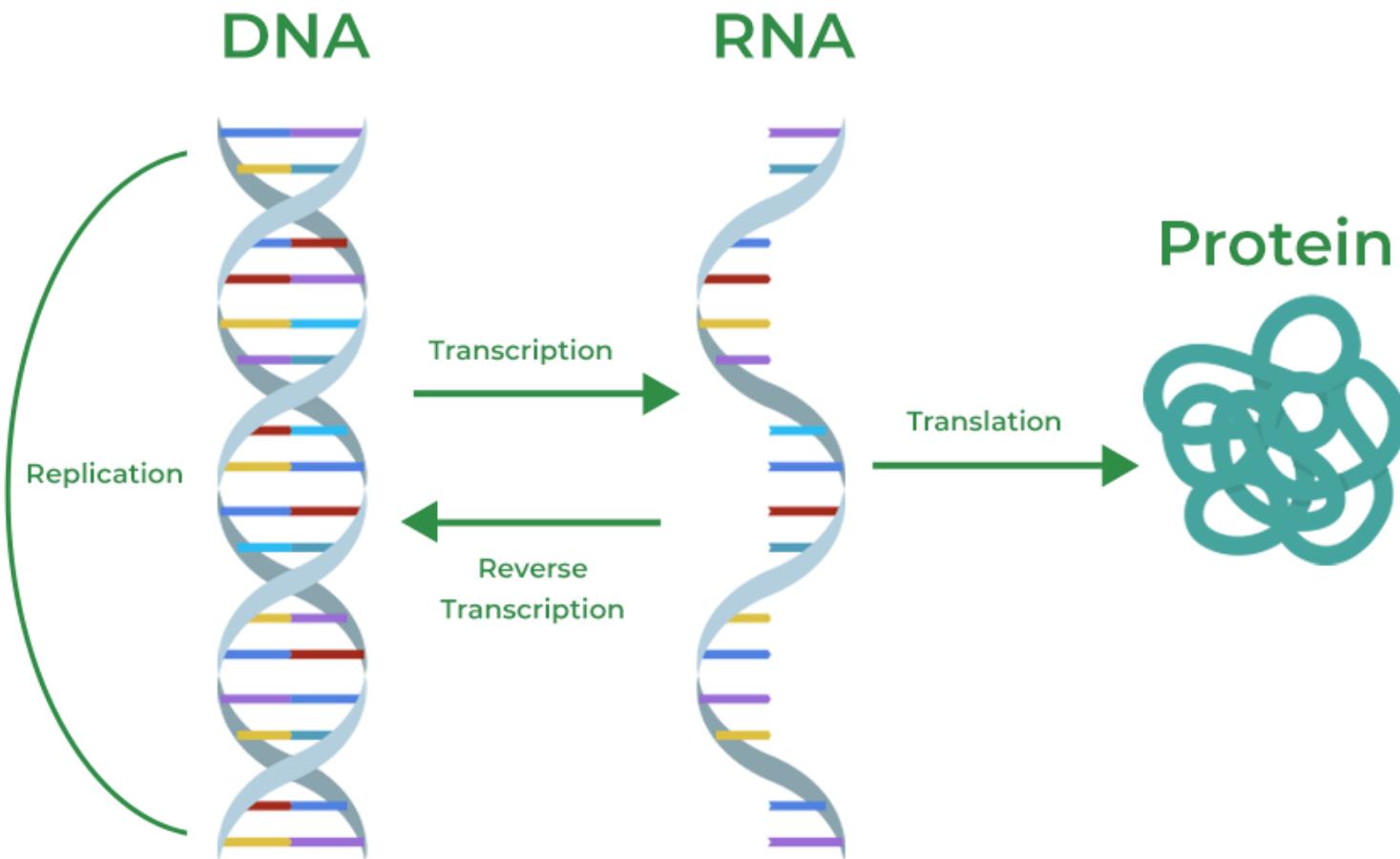


The relationship between structure and function is interdisciplinary.
It helps us build functional objects and understand biological objects.

Solving the structure of biomacromolecules

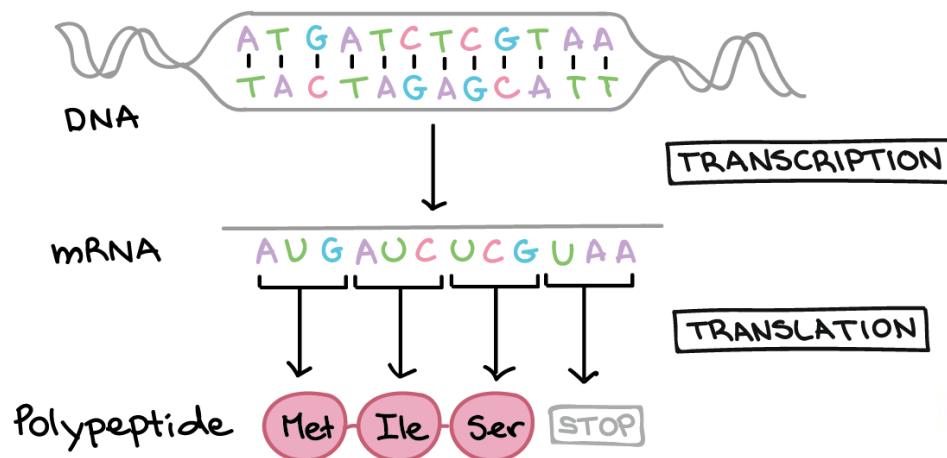


Molecular biology central dogma

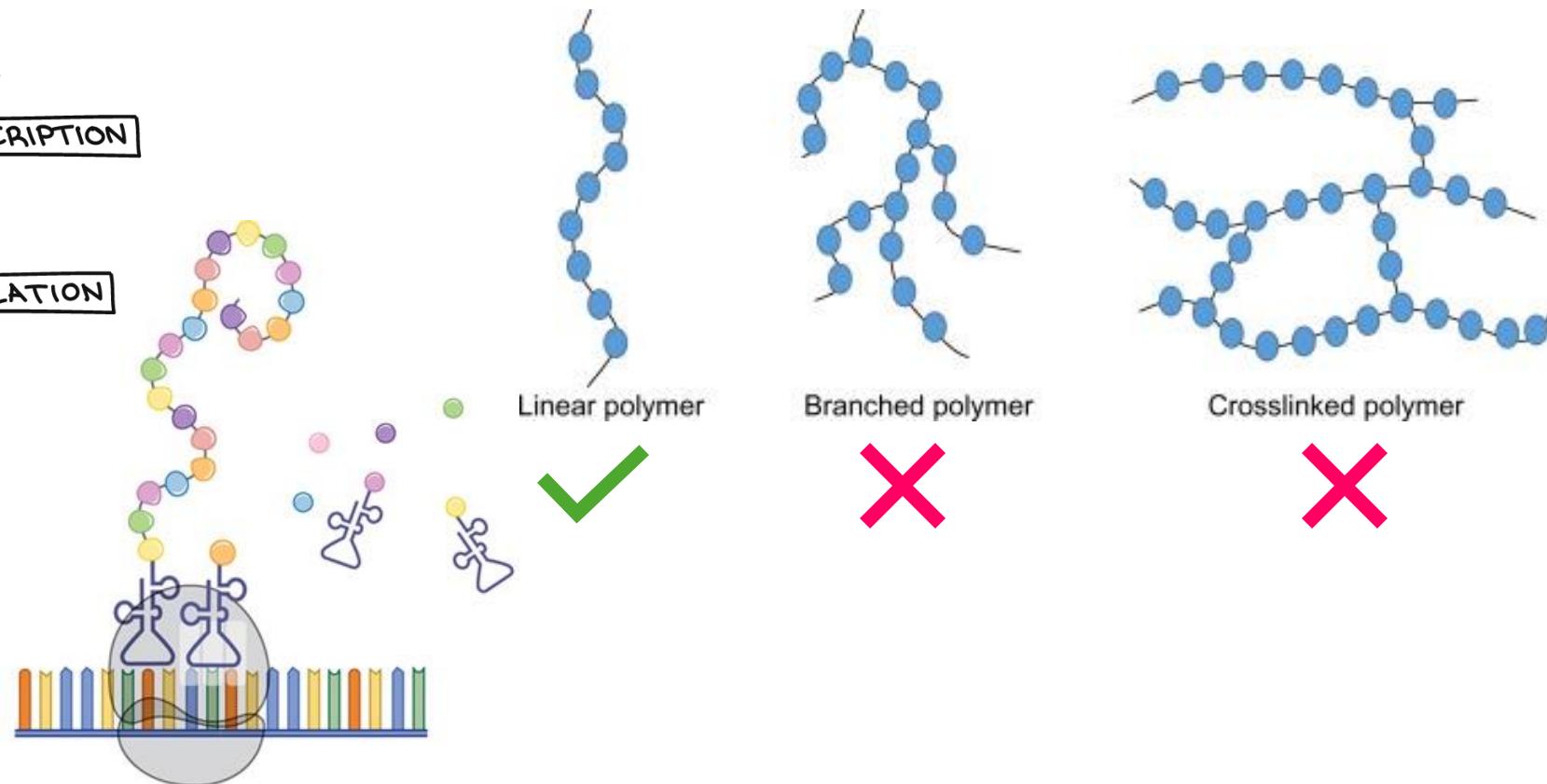


Proteins are linear heteropolymers

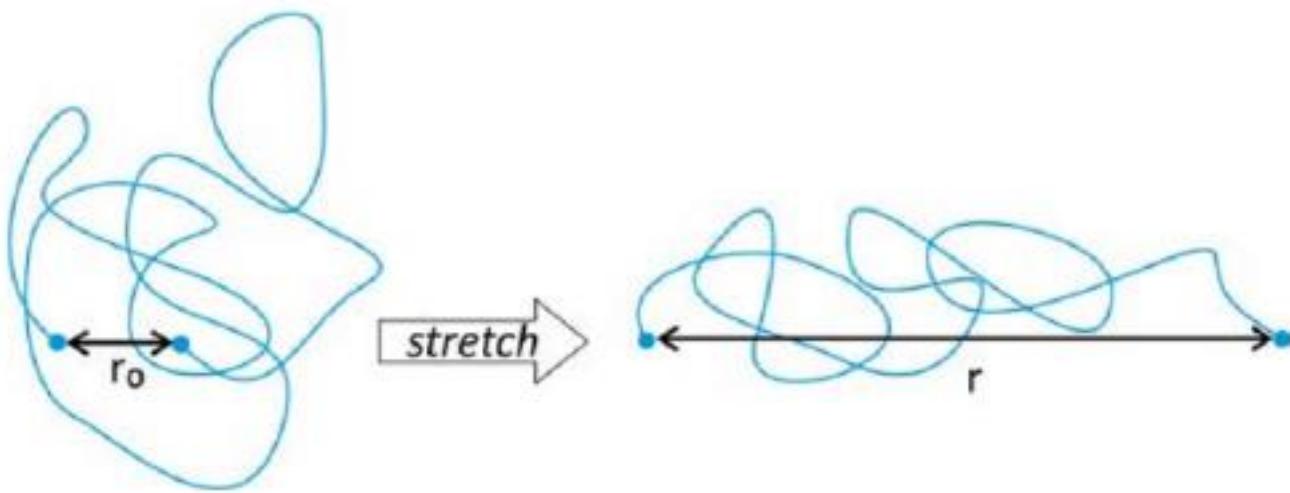
THE CENTRAL DOGMA



<https://www.khanacademy.org>

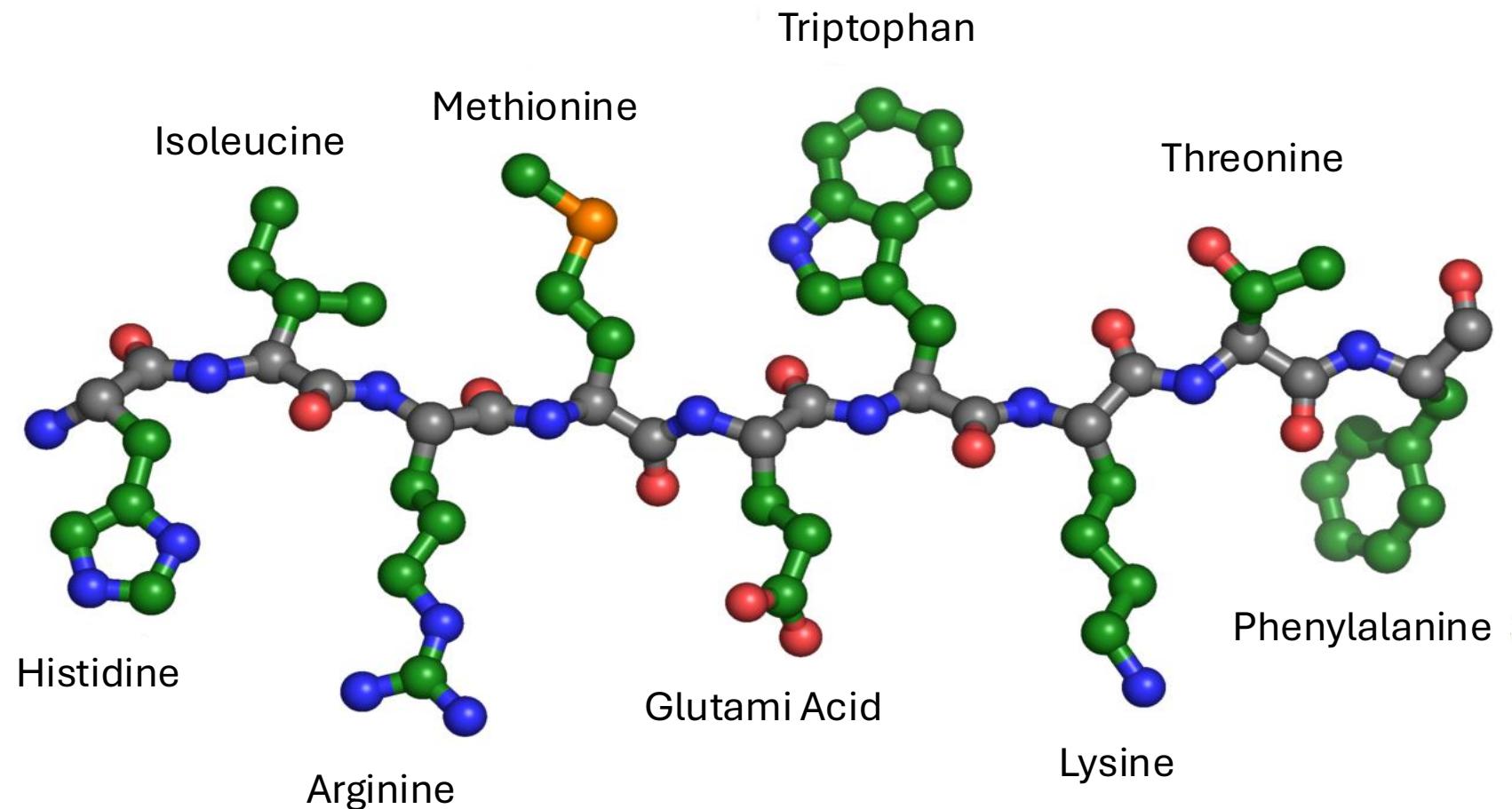


Conformations of a linear polymer



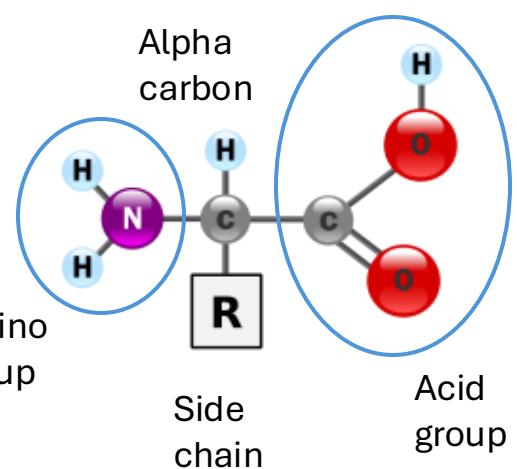
The completely stretched conformation is just one among countless possibilities.
But these conformations have **restrictions!**

Amino acids are the building blocks

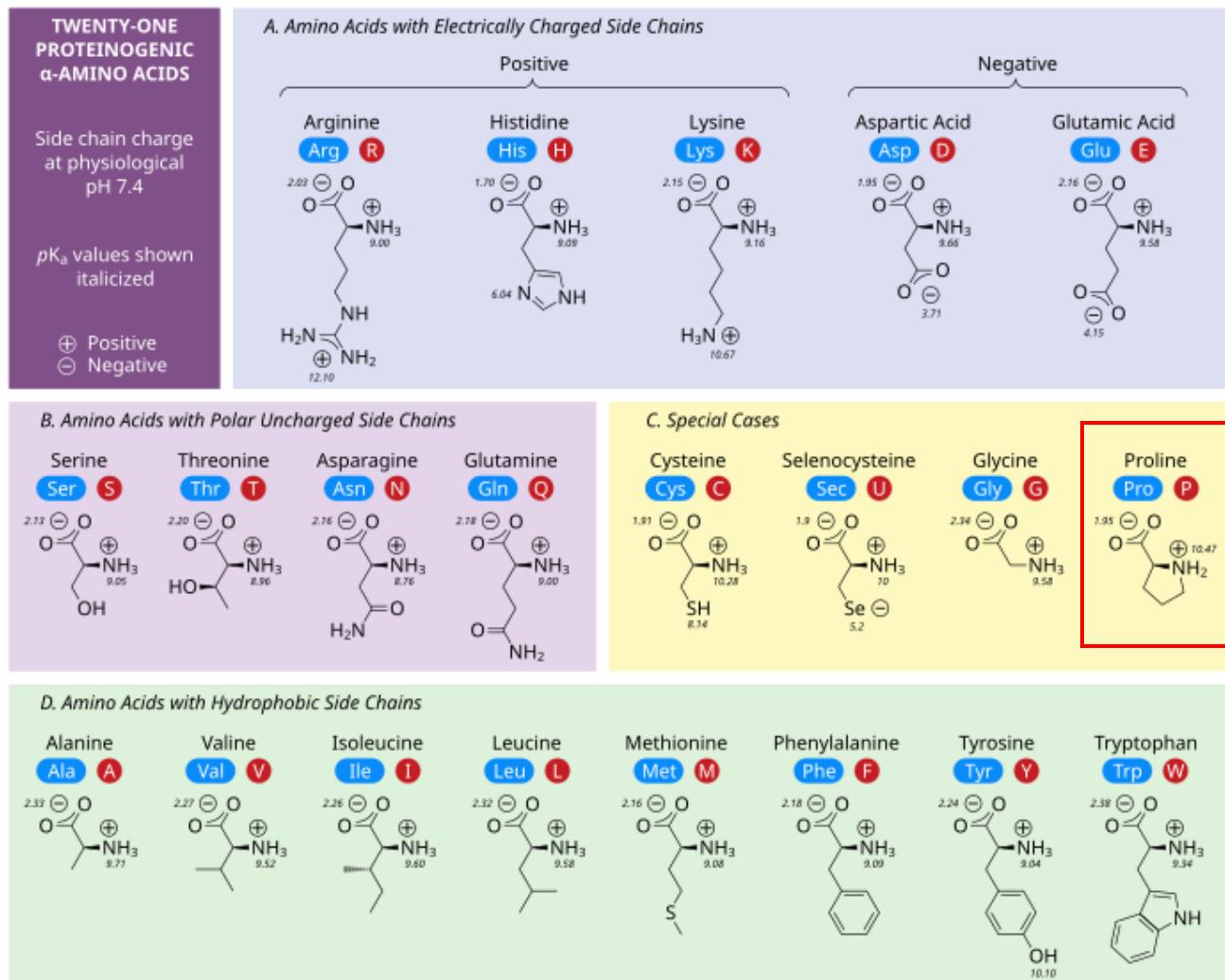


The twenty amino acids

Basic structure



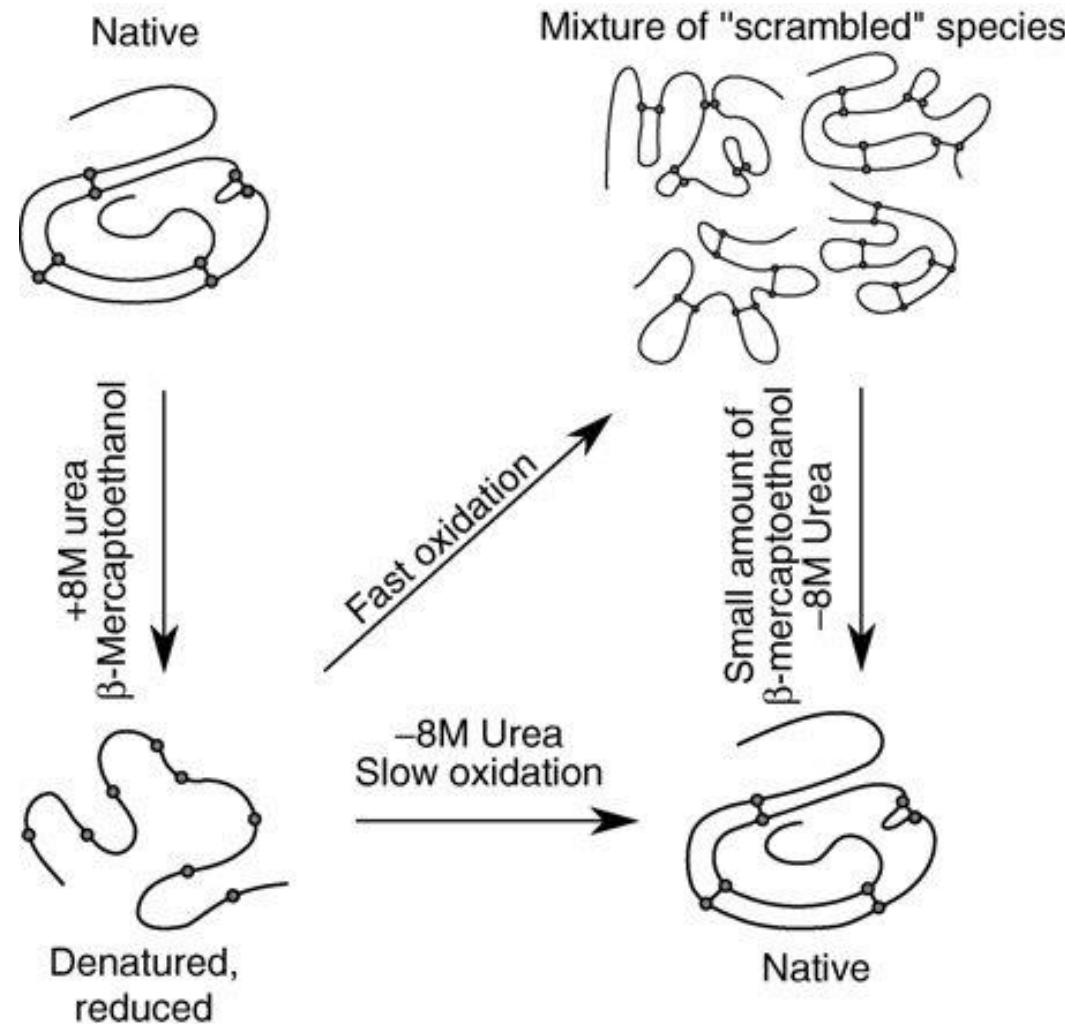
+
different side chains



Anfinsen's experiment (1961)

Urea
disrupts non-covalent
interactions like hydrogen
bonds

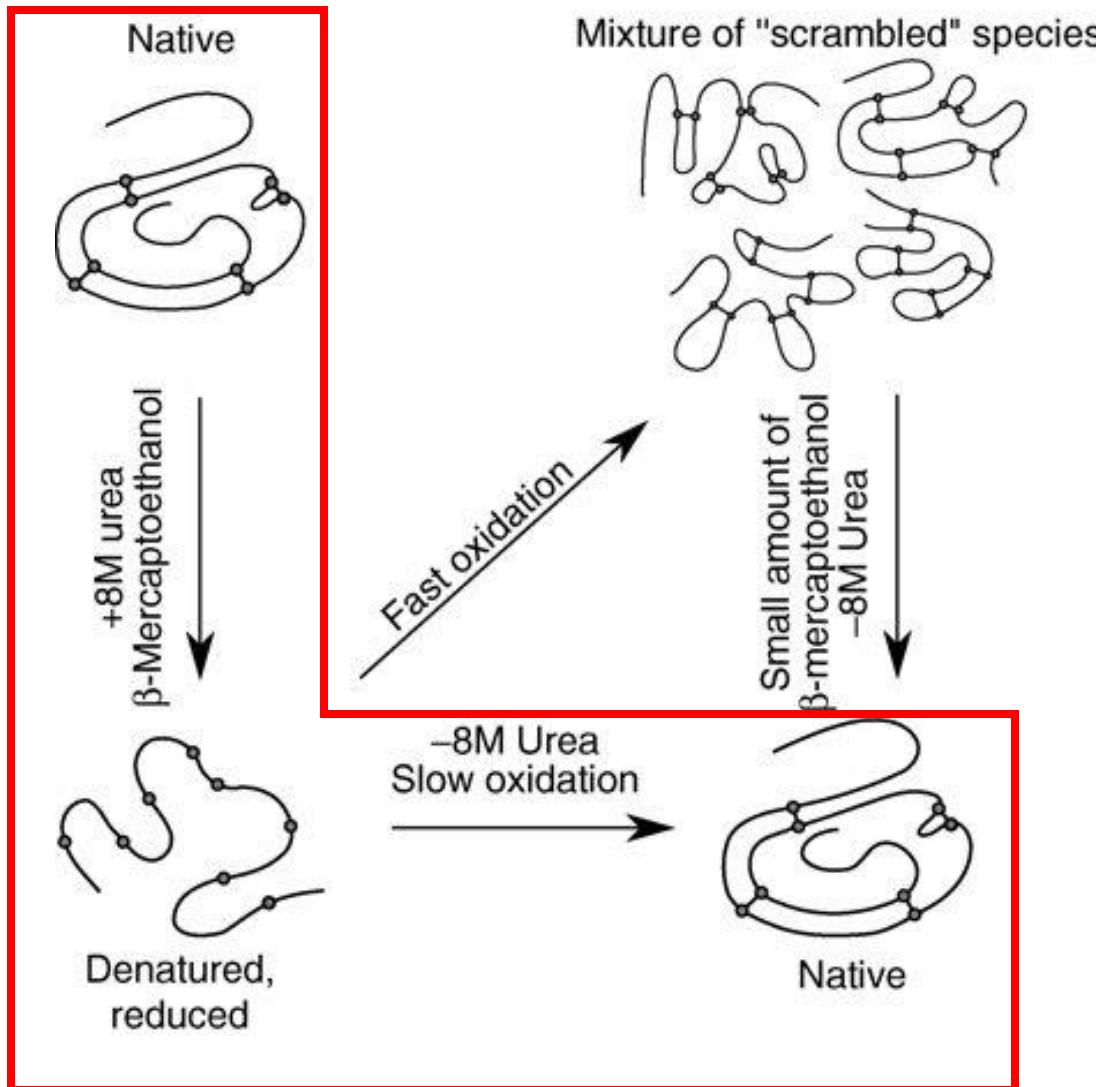
β -mercaptoethanol
reduces disulfide bonds



Christian B. Anfinsen

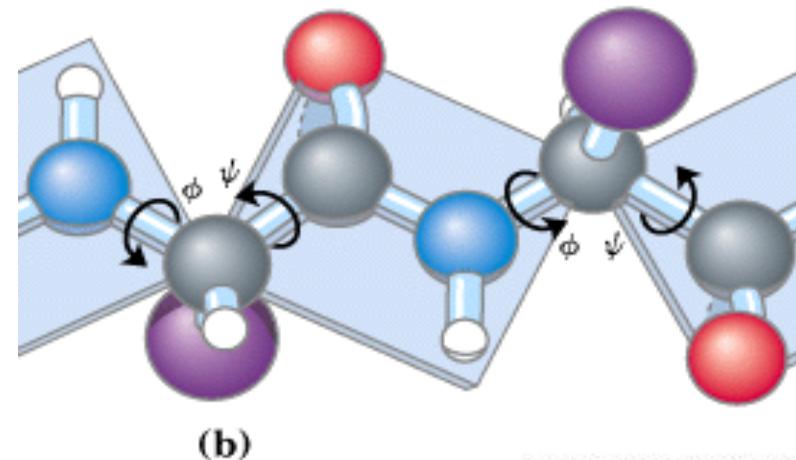
Anfinsen's hypothesis

The information for correct protein folding is entirely contained in its amino acid sequence.



Levinthal's paradox (1969)

- 3 stable conformations per angle (alpha helix, beta sheet and loop)
- For a protein with 101 amino acids,
 $3^{100} = 5 \times 10^{47}$ conformations
- Bond rotation rate of 10^{13} conformations per second
- It will take **10^{27} years to try them all!**

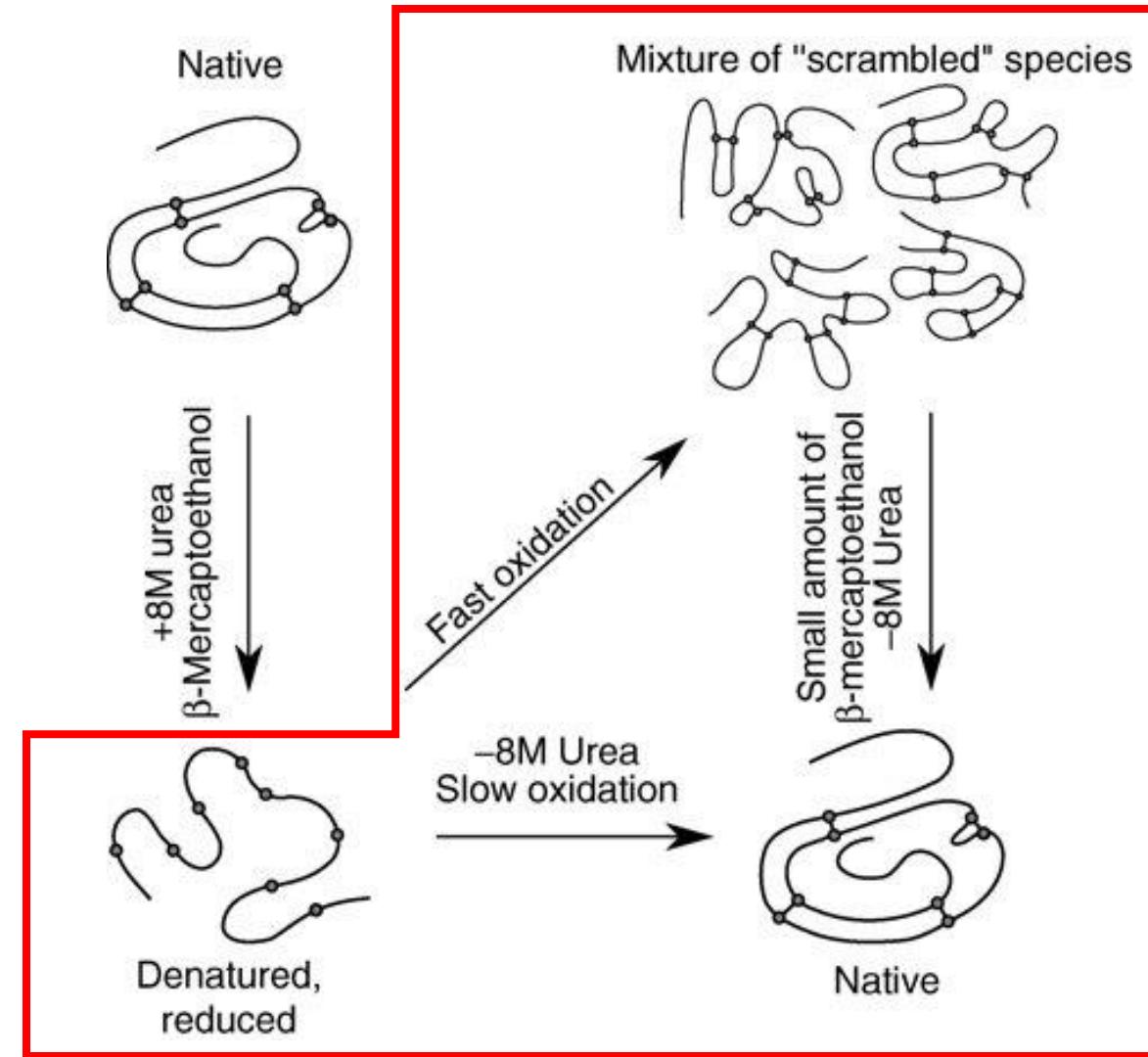


©1999 GARLAND PUBLISHING INC.
A member of the Taylor & Francis Group

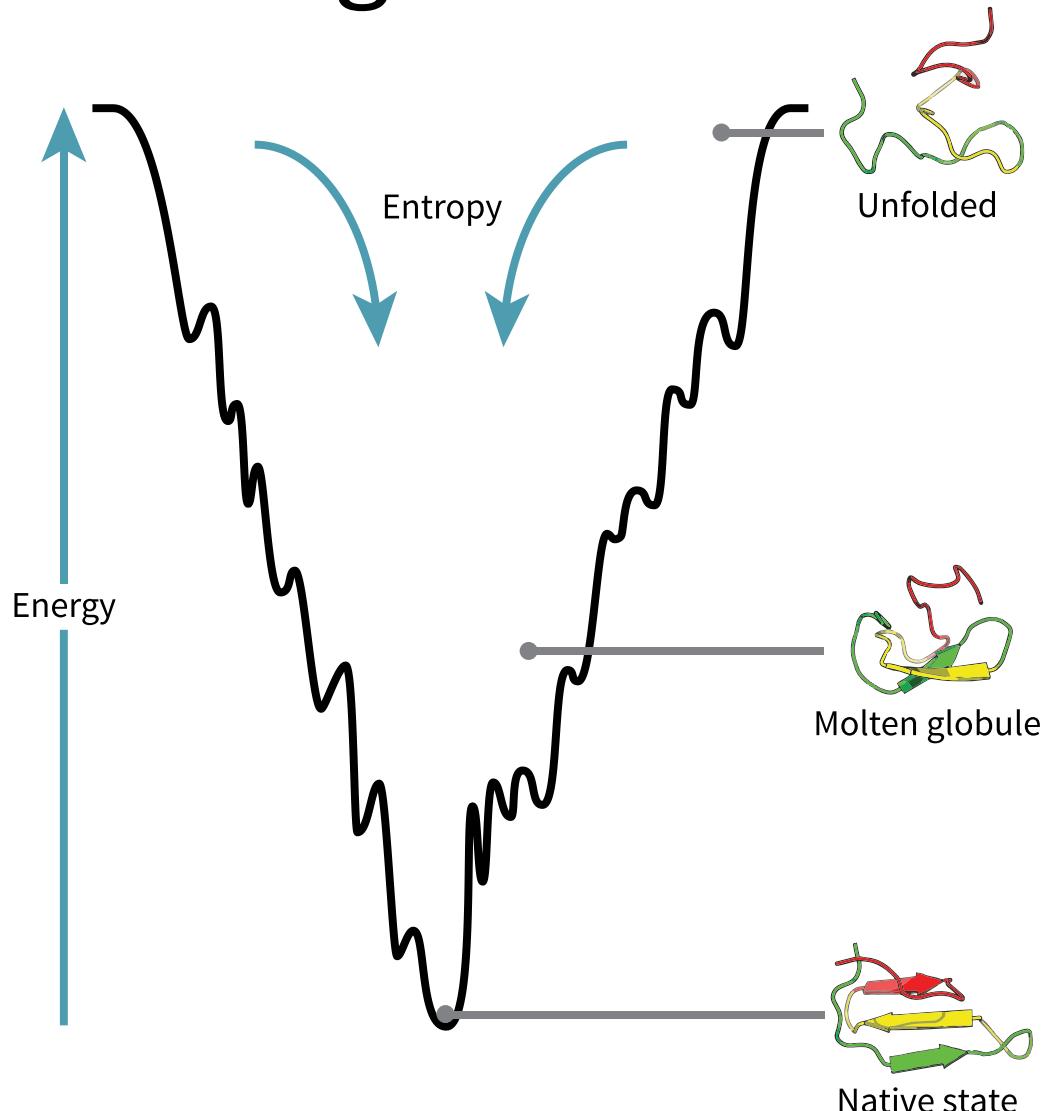
Order of hydrogen bonds and disulfide bonds

The order and timing of bond formation (non-covalent before covalent) are crucial.

Hydrogen bonds help *find* the native structure, while disulfide bonds help *stabilize* it.



Folding funnel



Extended conformations have low probability and high energy

Globular conformations have a hydrophobic core

Lacking water, polar groups of the main chain make hydrogen bonds with each other, forming Secondary Structures (helices and sheets)

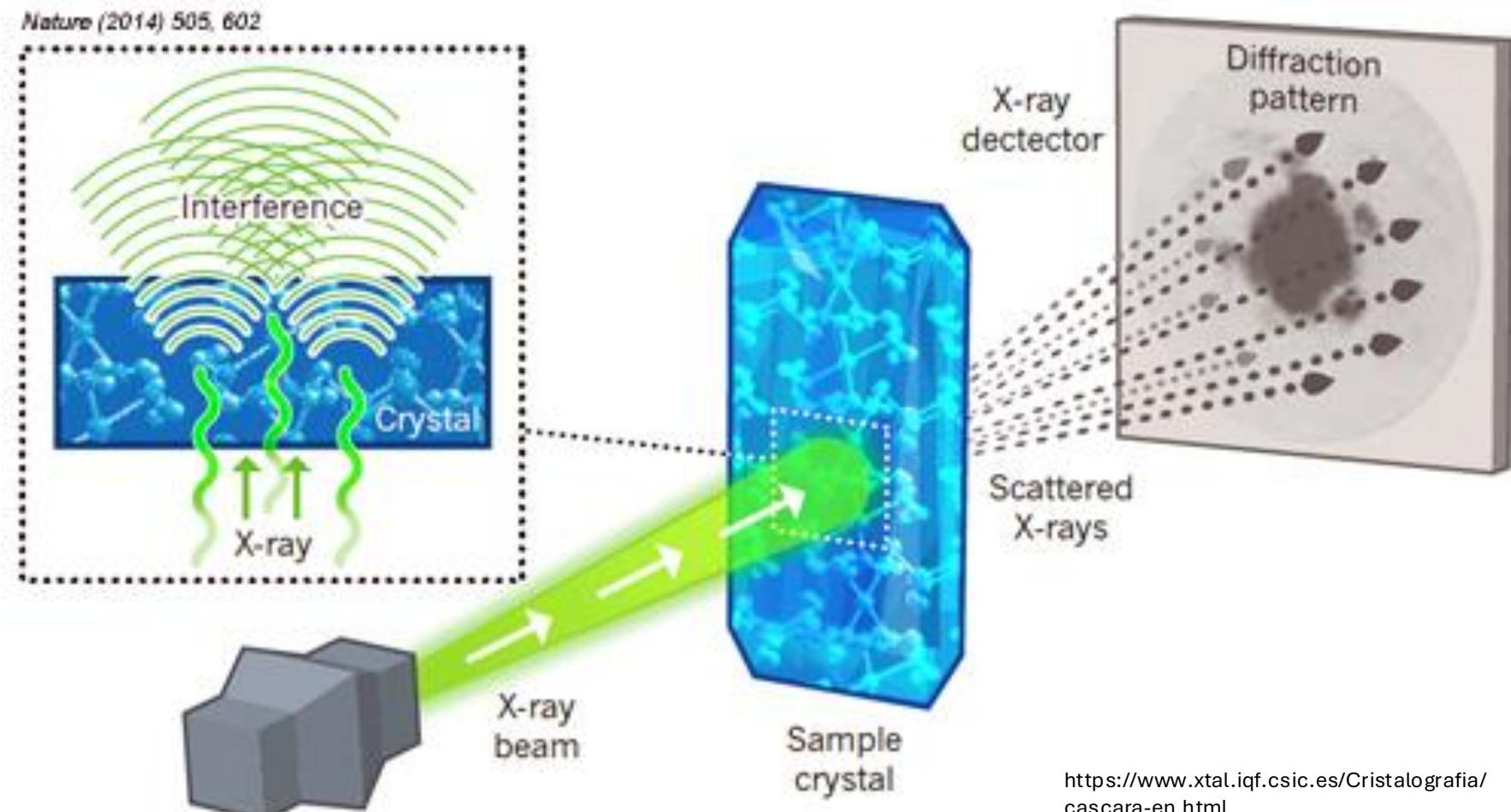
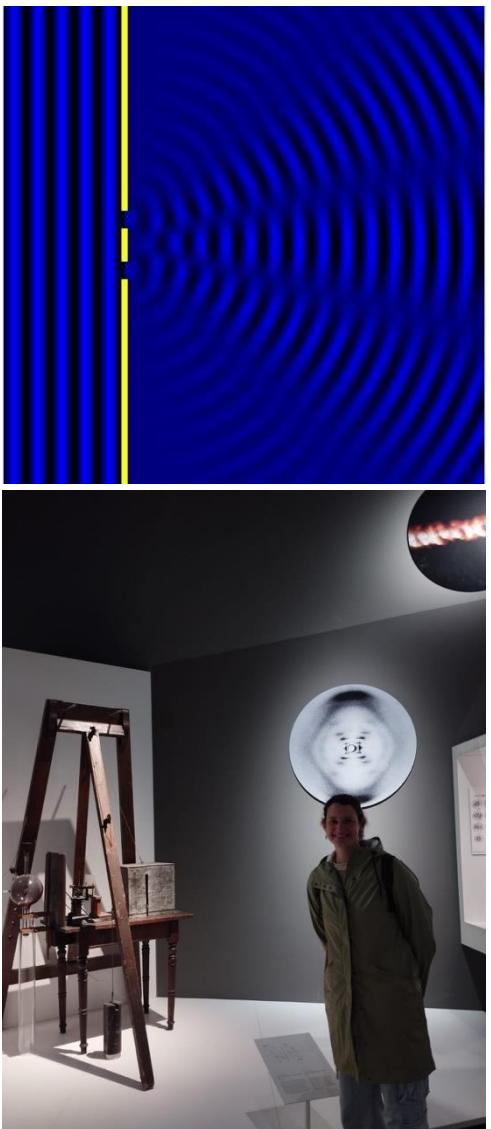
Secondary Structures arrange into motives and domains, which arrange into Tertiary Structures

Other protein chains arrange to form Quaternary Structures

The quest to solve protein structure

- Experiments – Crystallography, Nuclear Magnetic Resonance and Cryo-Electron Microscopy
- Modeling – Simulate folding and by homology with known structures

Crystallography



Max von Laue's
x-ray machine, 1912

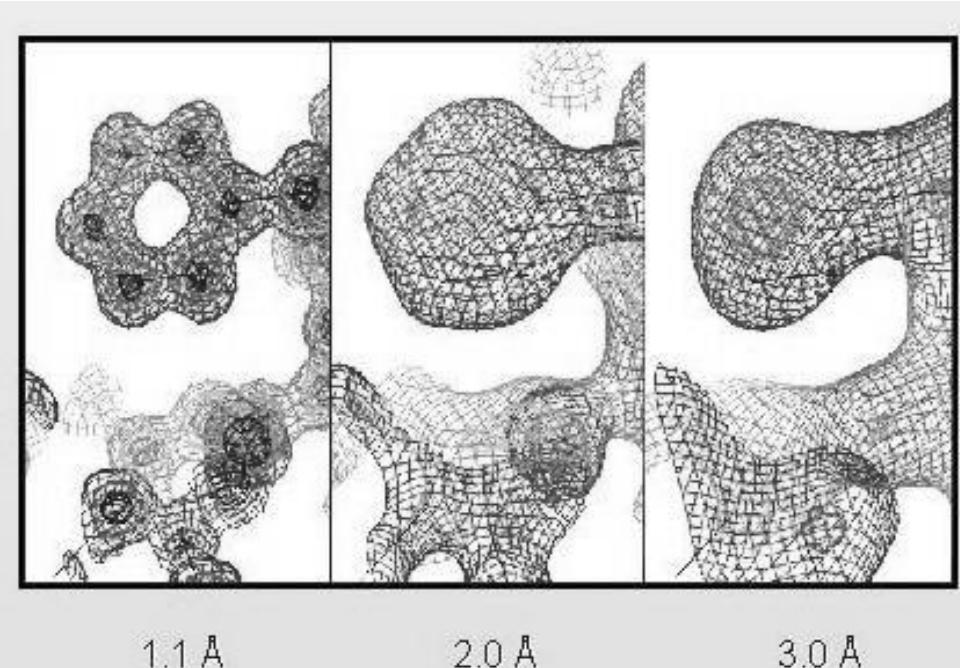
[https://www.xtal.iqf.csic.es/Cristalografia/
cascara-en.html](https://www.xtal.iqf.csic.es/Cristalografia/cascara-en.html)

The phase problem

Diffraction data gives the **intensities** of reflected X-rays, which are proportional to the **amplitudes** of the structure factors. However, **phases** are lost in the experiment, and you need both **amplitudes and phases** to reconstruct the electron density map (via Fourier transform).

Molecular replacement solves the phase problem by:

- 1.Using a previously solved **homologous structure** as a **model**.
- 2.Placing and orienting the model within the new crystal's unit cell.
- 3.Calculating initial phase estimates from this positioned model.
- 4.Refining the model against the observed data.



<http://www.bmsc.washington.edu/people/verlinde/experiment.html>

A model that fits the experimental data.

Identification of common structural features

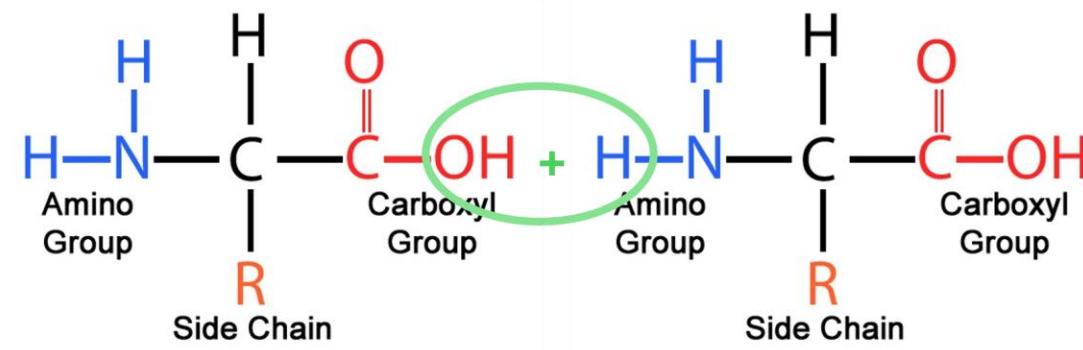
- Not all conformations are allowed because some cause clashes between atoms
- Foldings are hierarchically organized into secondary, tertiary and quaternary structures

Prohibited conformations

Calculated prohibited conformations coming from the restrictions of the peptide bond.



G.N. Ramachandran

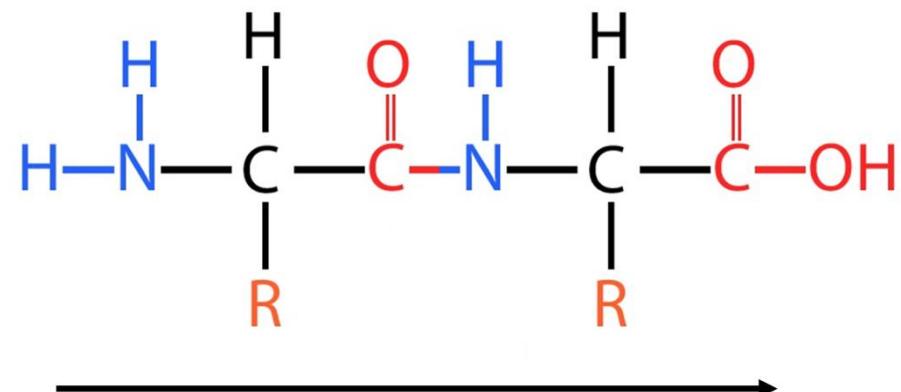


Prohibited conformations

Calculated prohibited conformations coming from the restrictions of the peptide bond.



G.N. Ramachandran



From the amino-terminal end to the carboxyl-terminal end

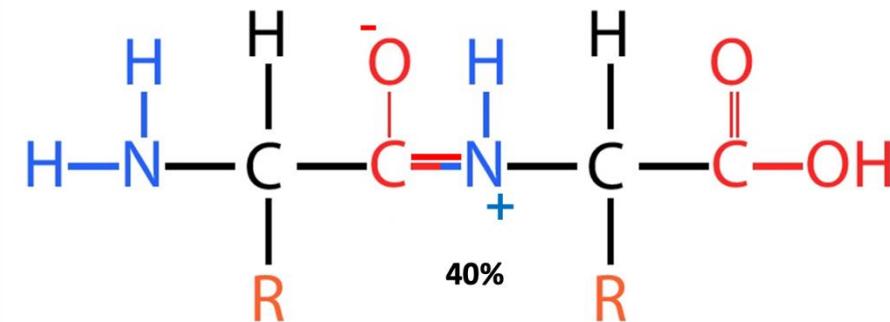
2 amino acid residues

Prohibited conformations

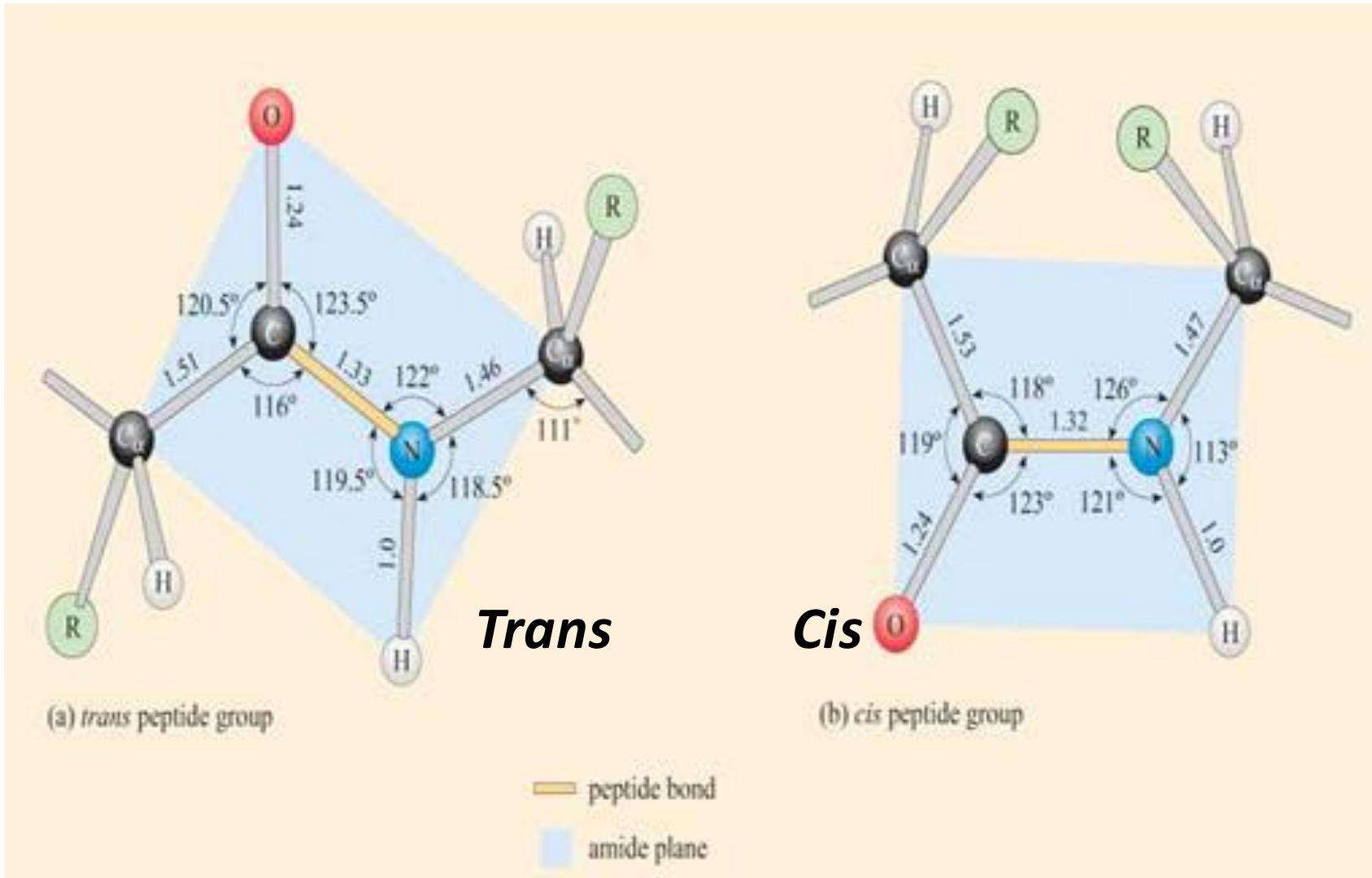
Calculated prohibited conformations coming from the restrictions of the peptide bond.



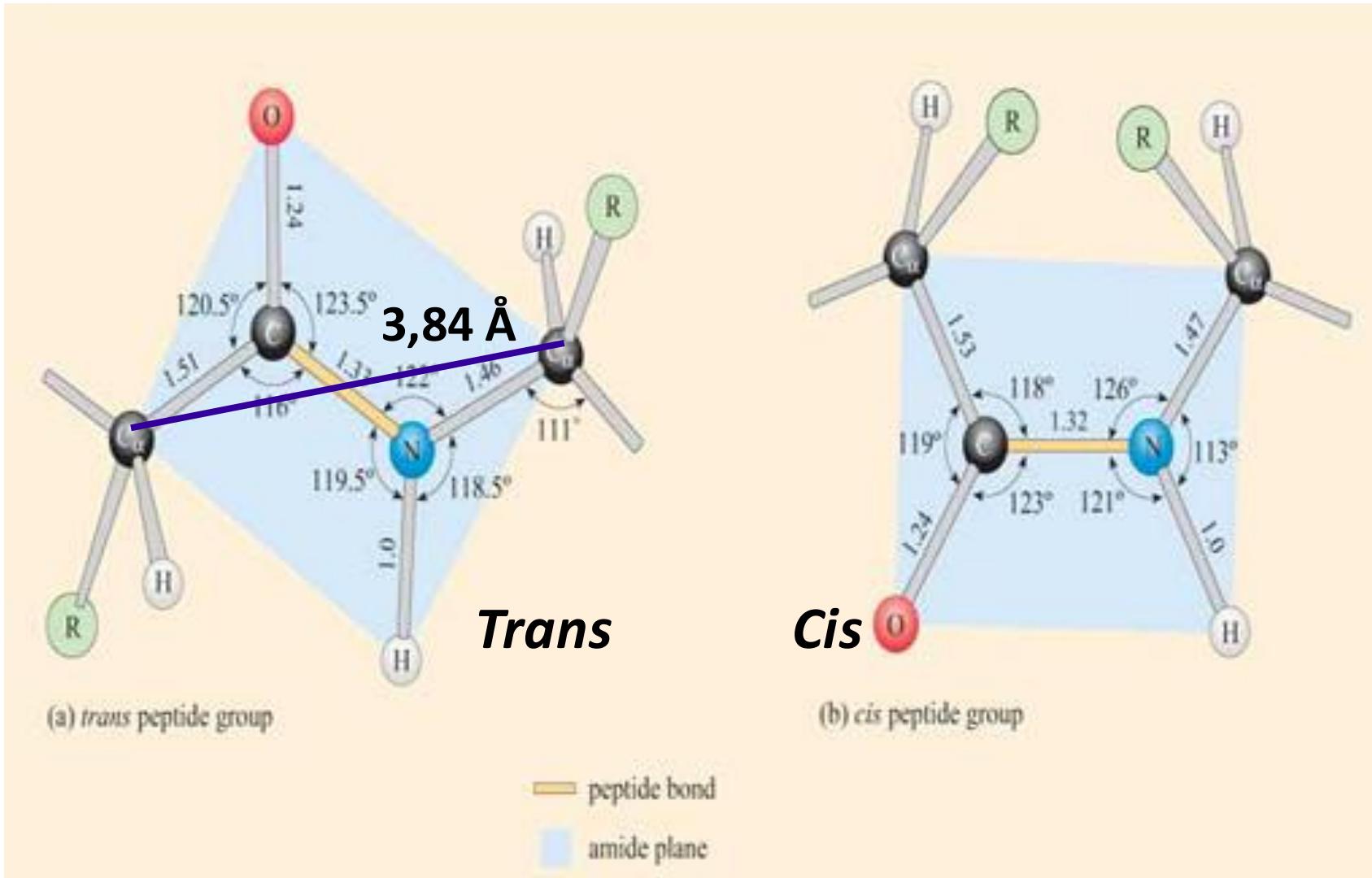
G.N. Ramachandran



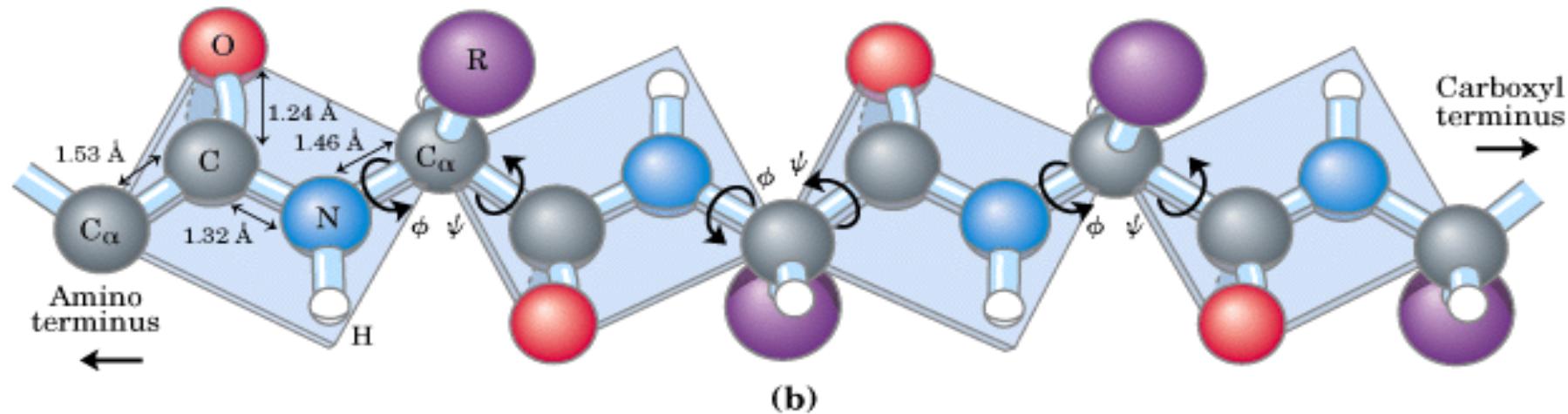
The peptide bond, cis or trans conformations



The peptide bond, cis or trans conformations



Between peptide planes are two torsion angles

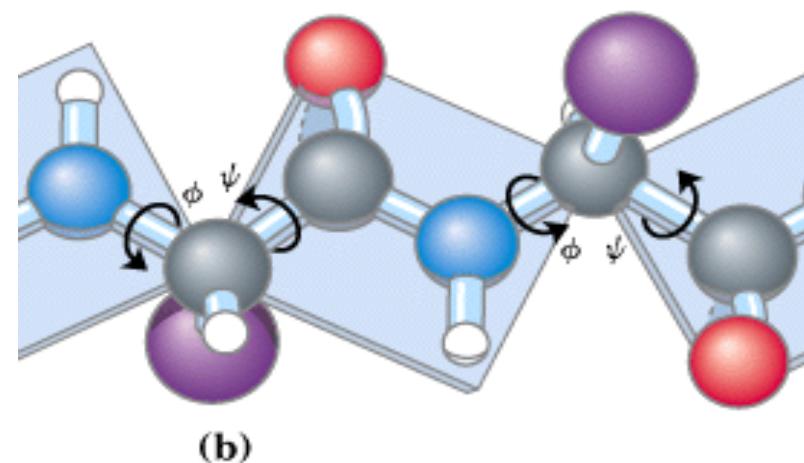
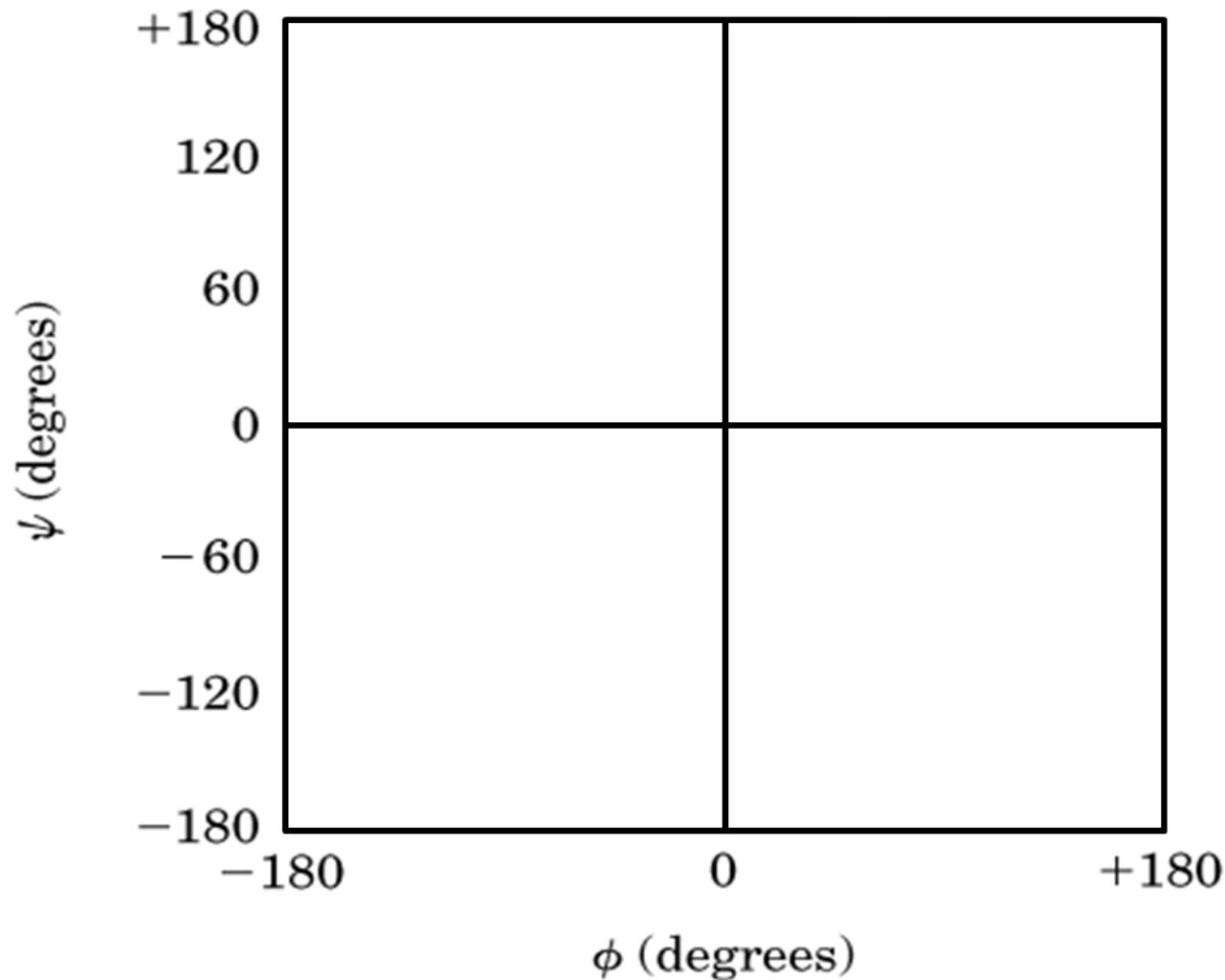


©1999 GARLAND PUBLISHING INC.
A member of the Taylor & Francis Group

Only the ϕ and ψ angles are free to rotate

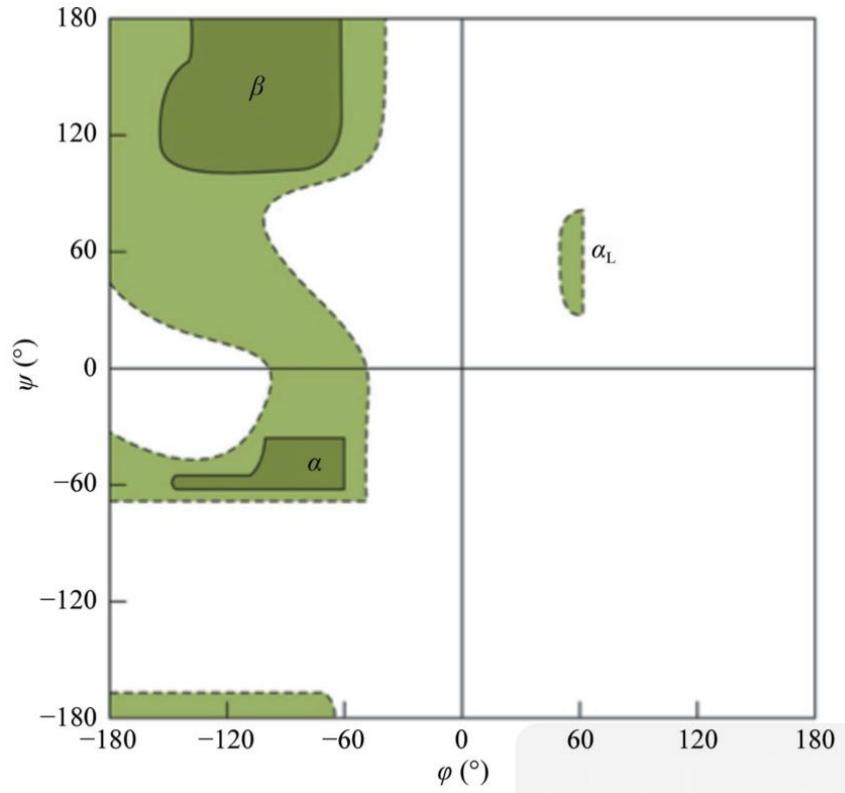
And the positions of the main-chain atoms can be entirely defined by them.

The Ramachandran plot

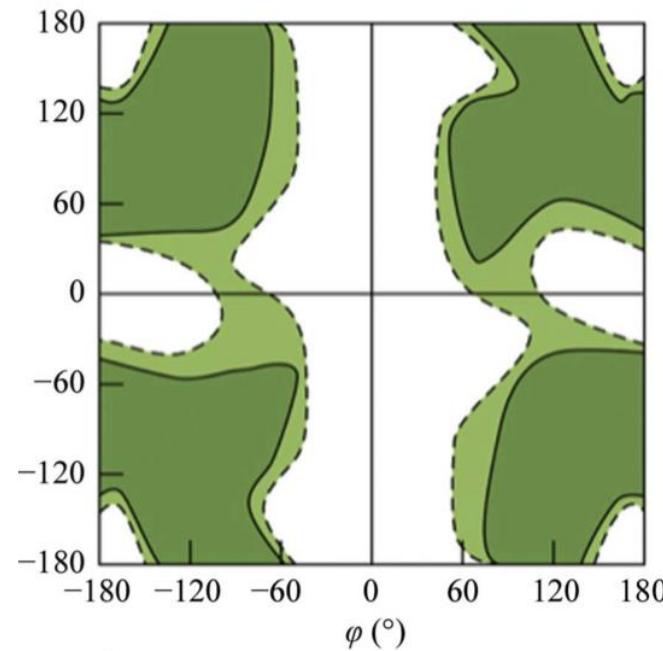
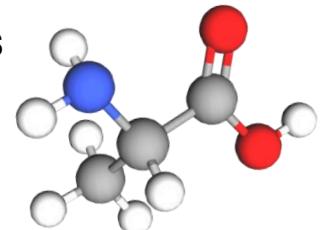


©1999 GARLAND PUBLISHING INC.
A member of the Taylor & Francis Group

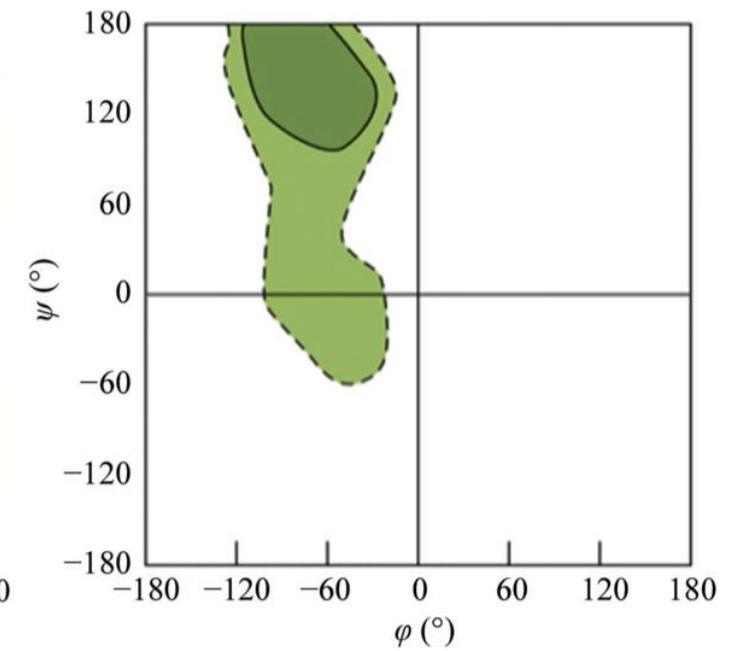
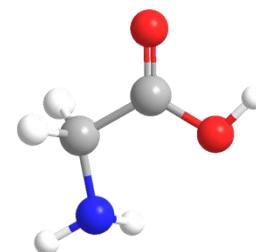
Allowed angles depending on side chain



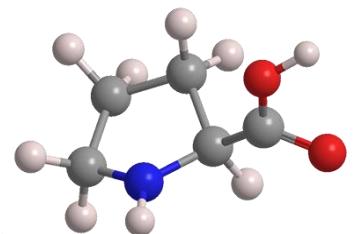
Most amino acids

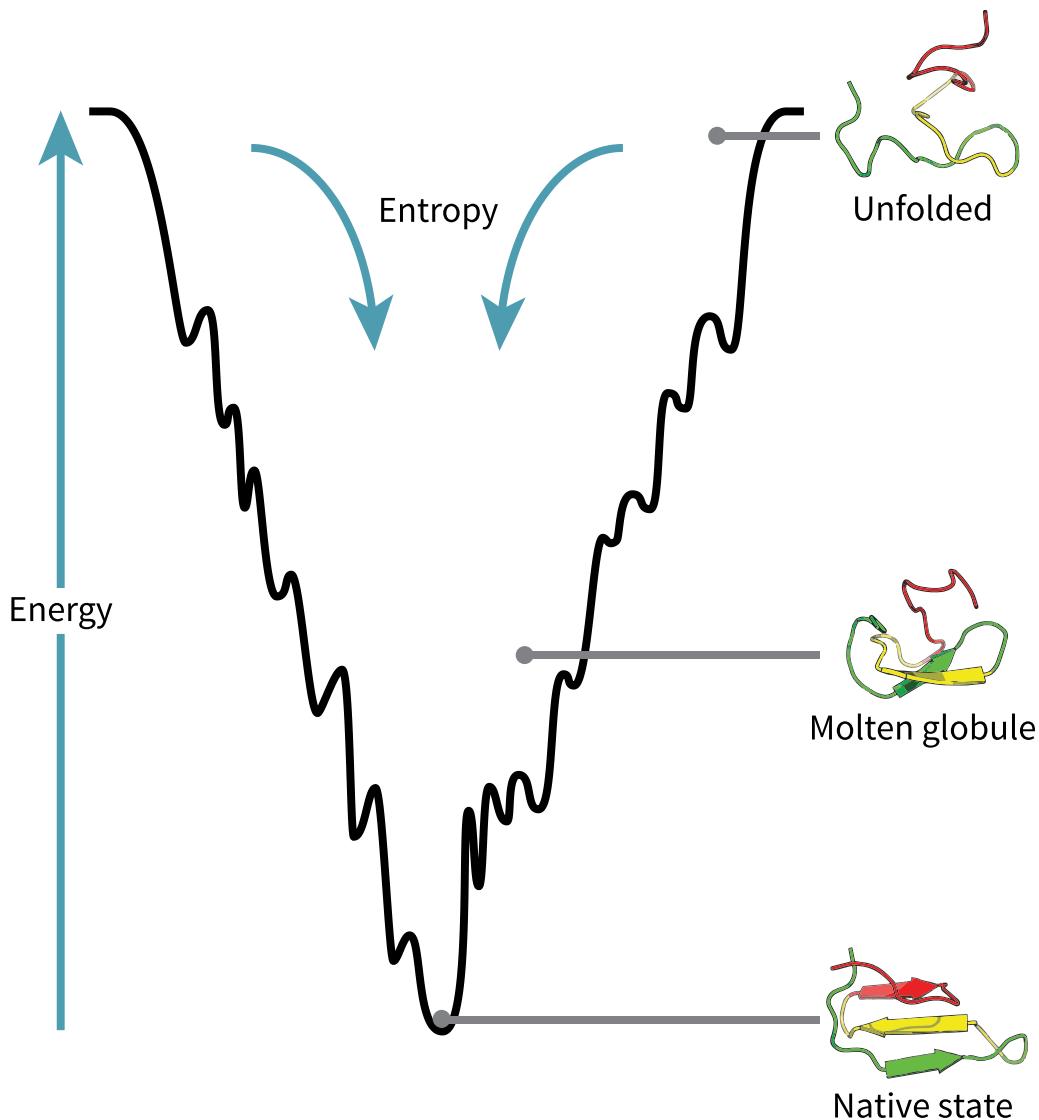


Glycine



Proline





Extended conformations have low probability and high energy

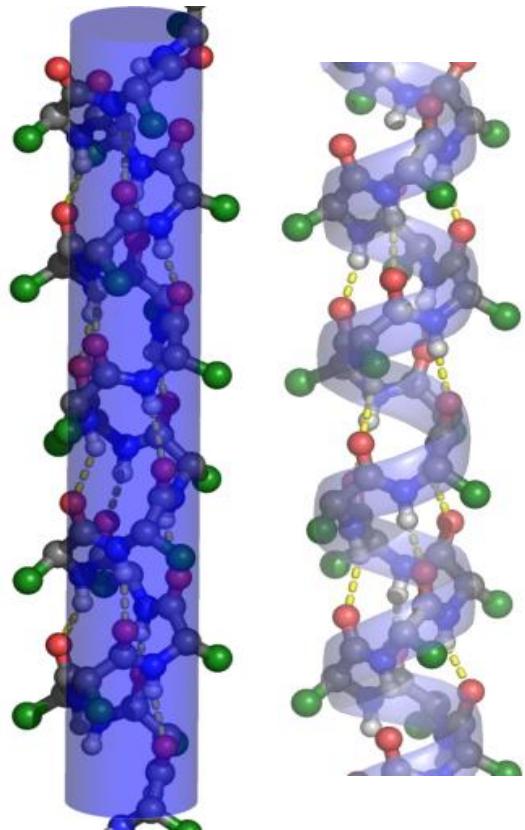
Globular conformations have a hydrophobic core

Lacking water, polar groups of the main chain make hydrogen bonds with each other, forming Secondary Structures (helices and sheets)

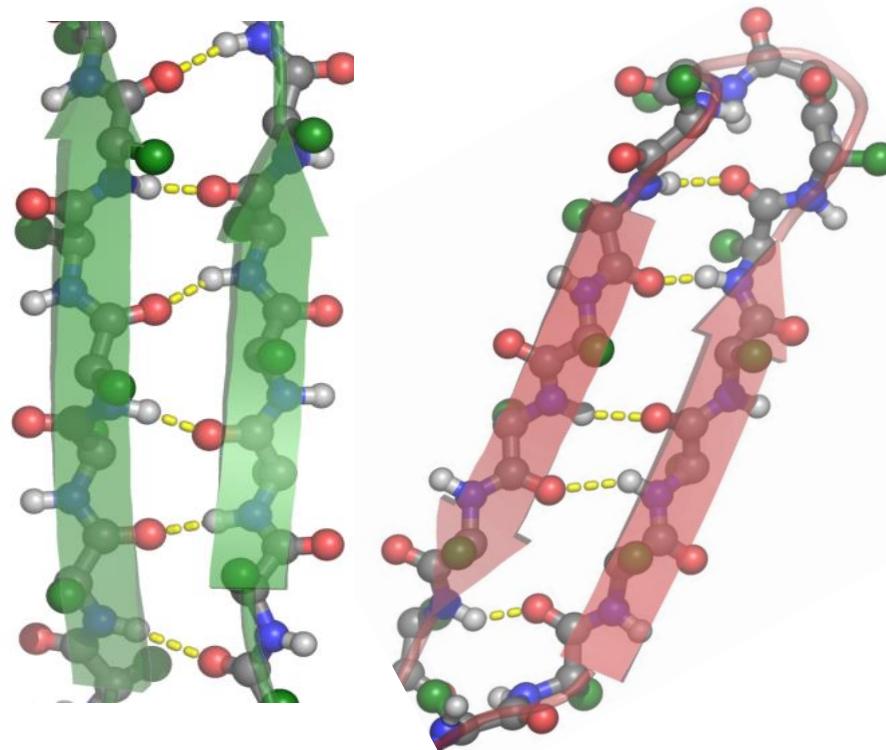
Secondary Structures arrange into motives and domains, which arrange into Tertiary Structures

Other protein chains arrange to form Quaternary Structures

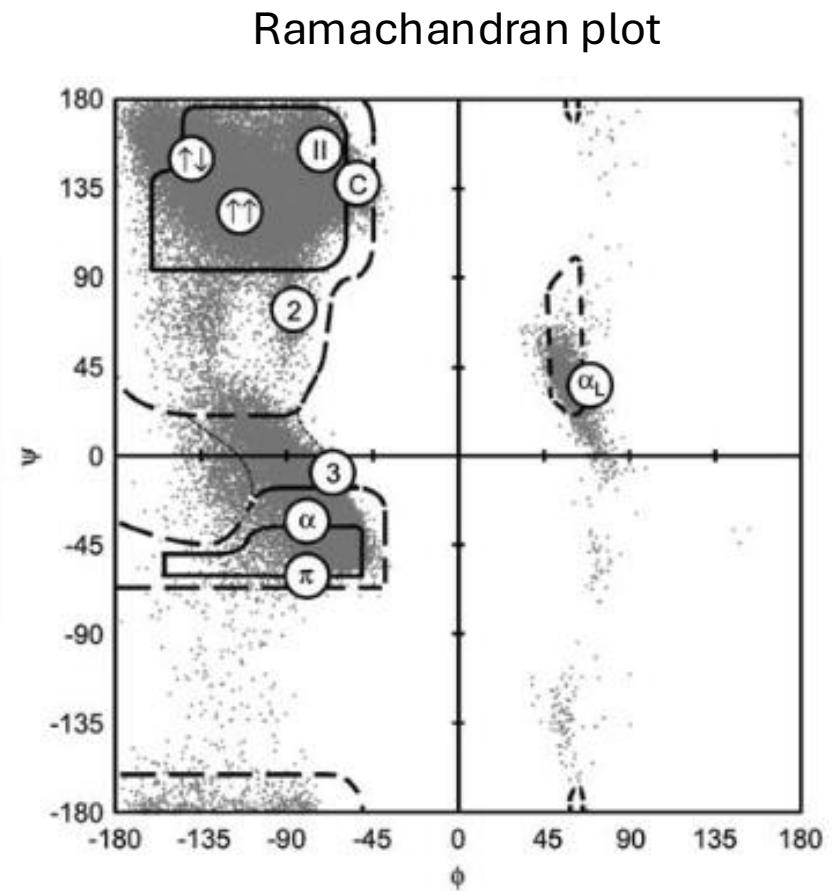
Secondary structure



α -helices



β -sheets



α -helices

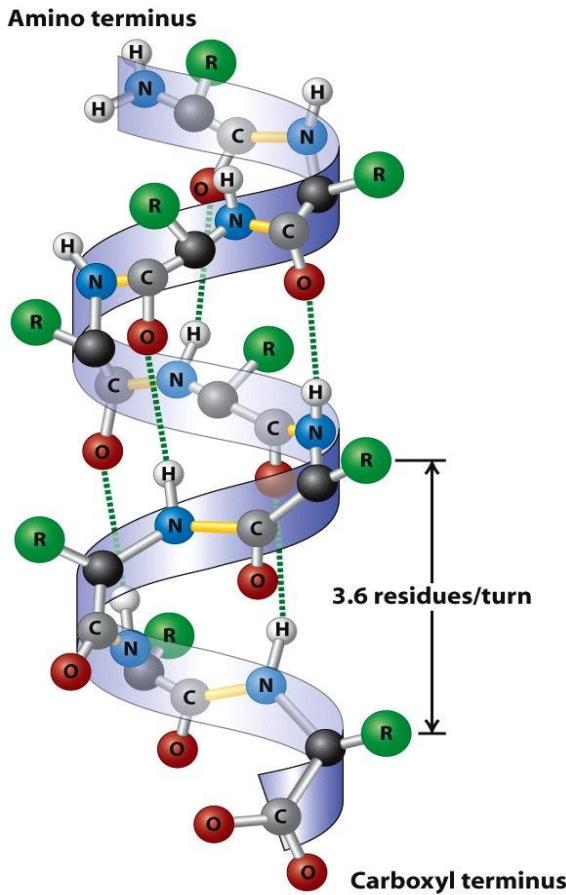
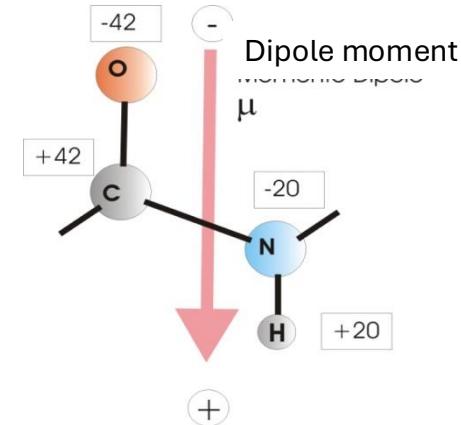
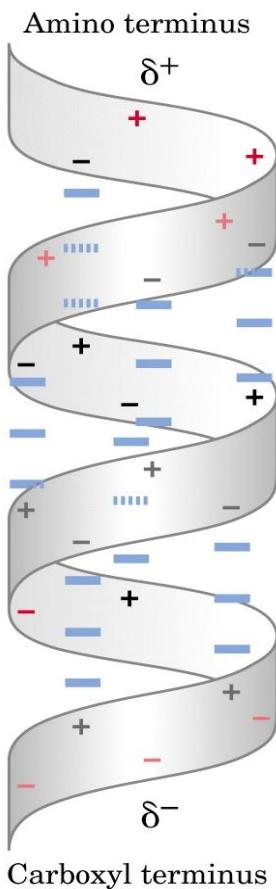


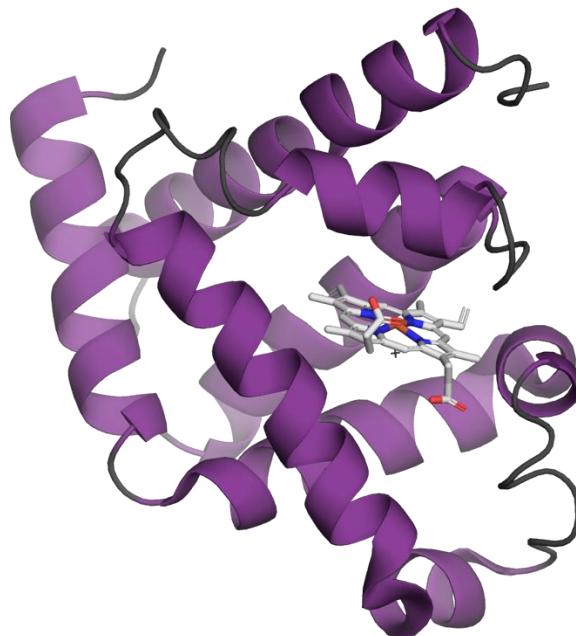
Figure 3-4
Molecular Cell Biology, Sixth Edition
© 2008 W.H. Freeman and Company



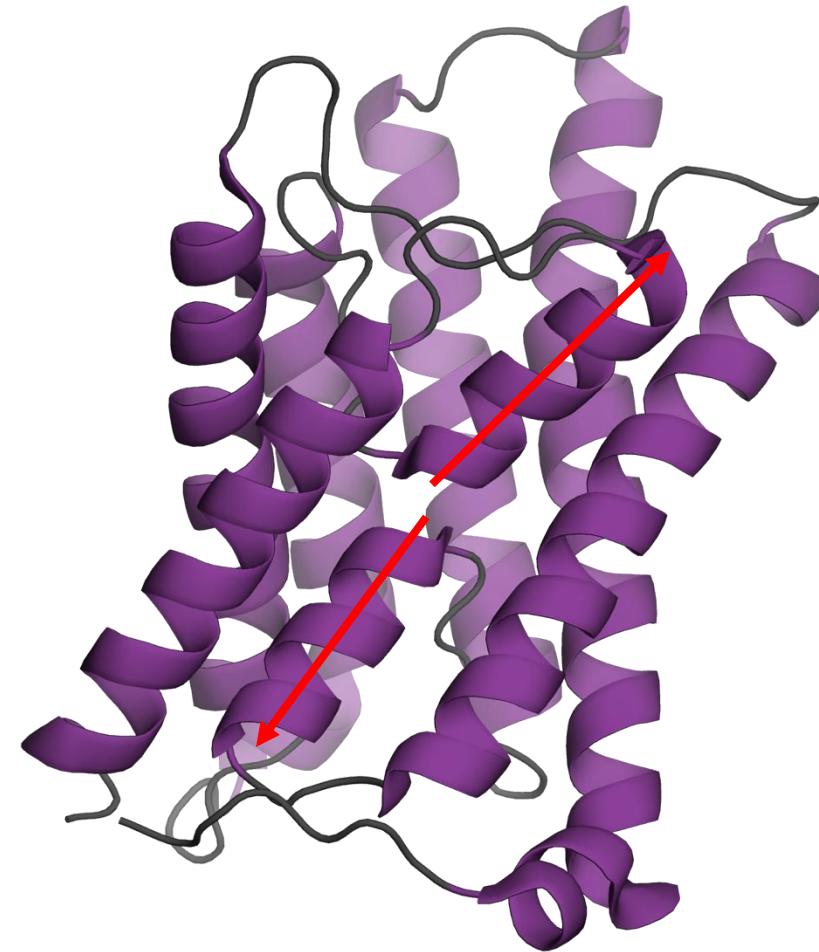
Biological Function

- Help in binding charged cofactors
- Long range attraction (K^+ and Cl^- channels)
- Change the nucleophilic properties of neighbouring residues for catalysis

All alpha protein example



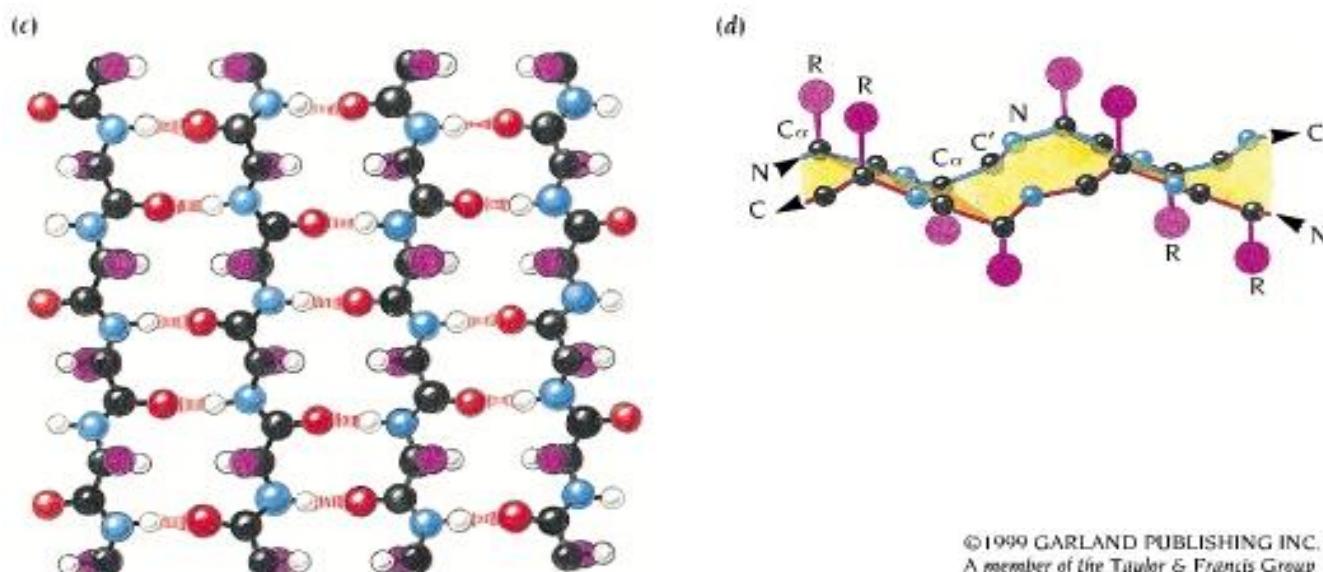
Myoglobin, 1976



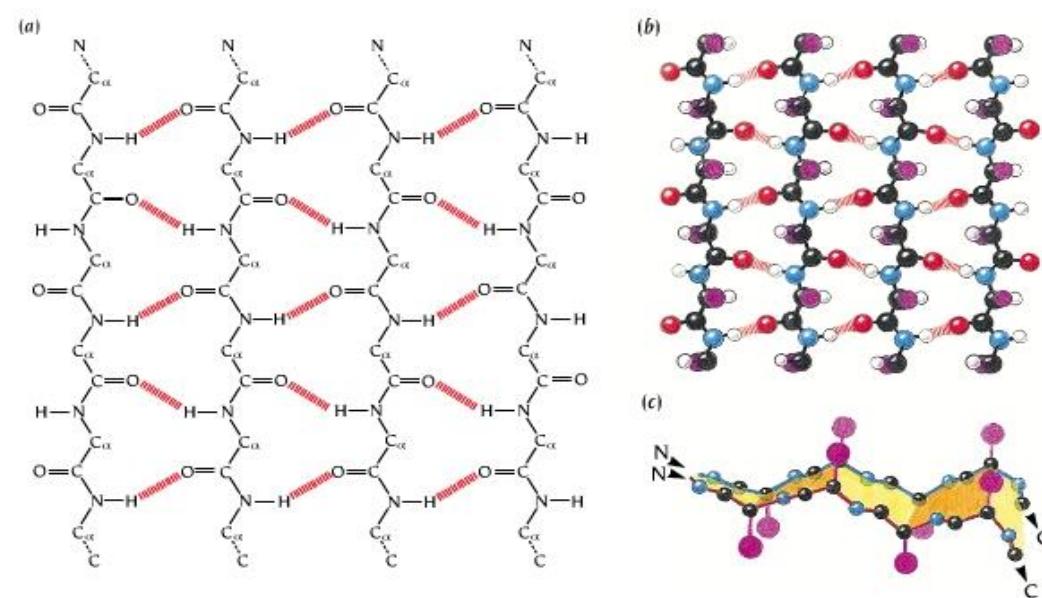
Aquaporin

β -sheets

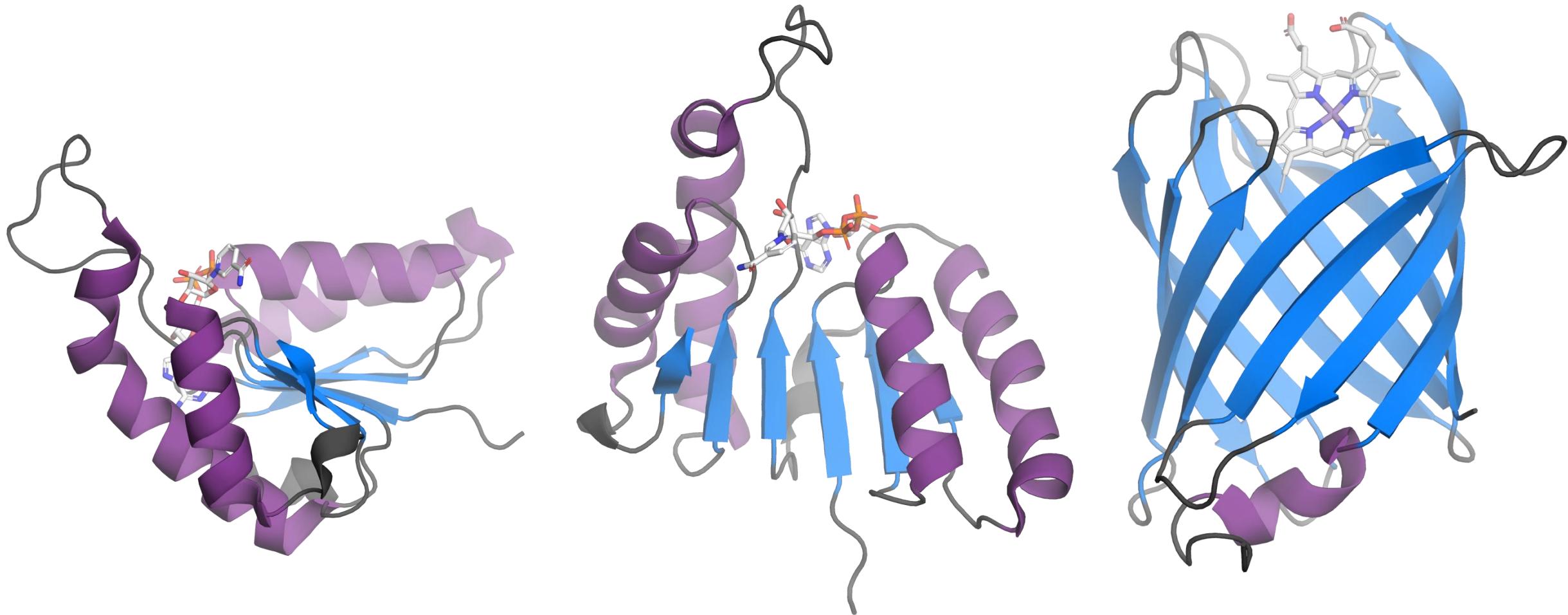
Anti-parallel β -sheet



Parallel β -sheet

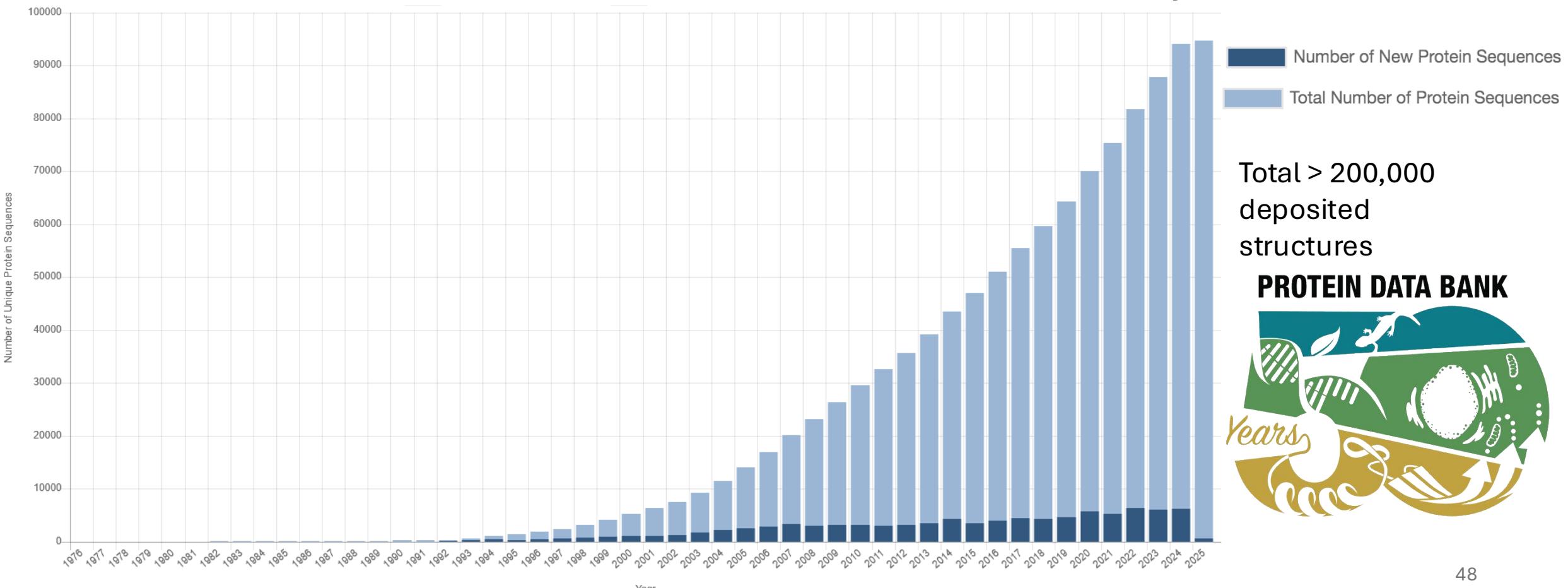


Example of beta sheet with twisted beta strands



Protein structure as a digital resource

- 1971 was established the Protein Data Bank with only 7 structures



Protein structure as a digital resource

- 2003 was announced the worldwide Protein Data Bank - wwPDB



The PDB website

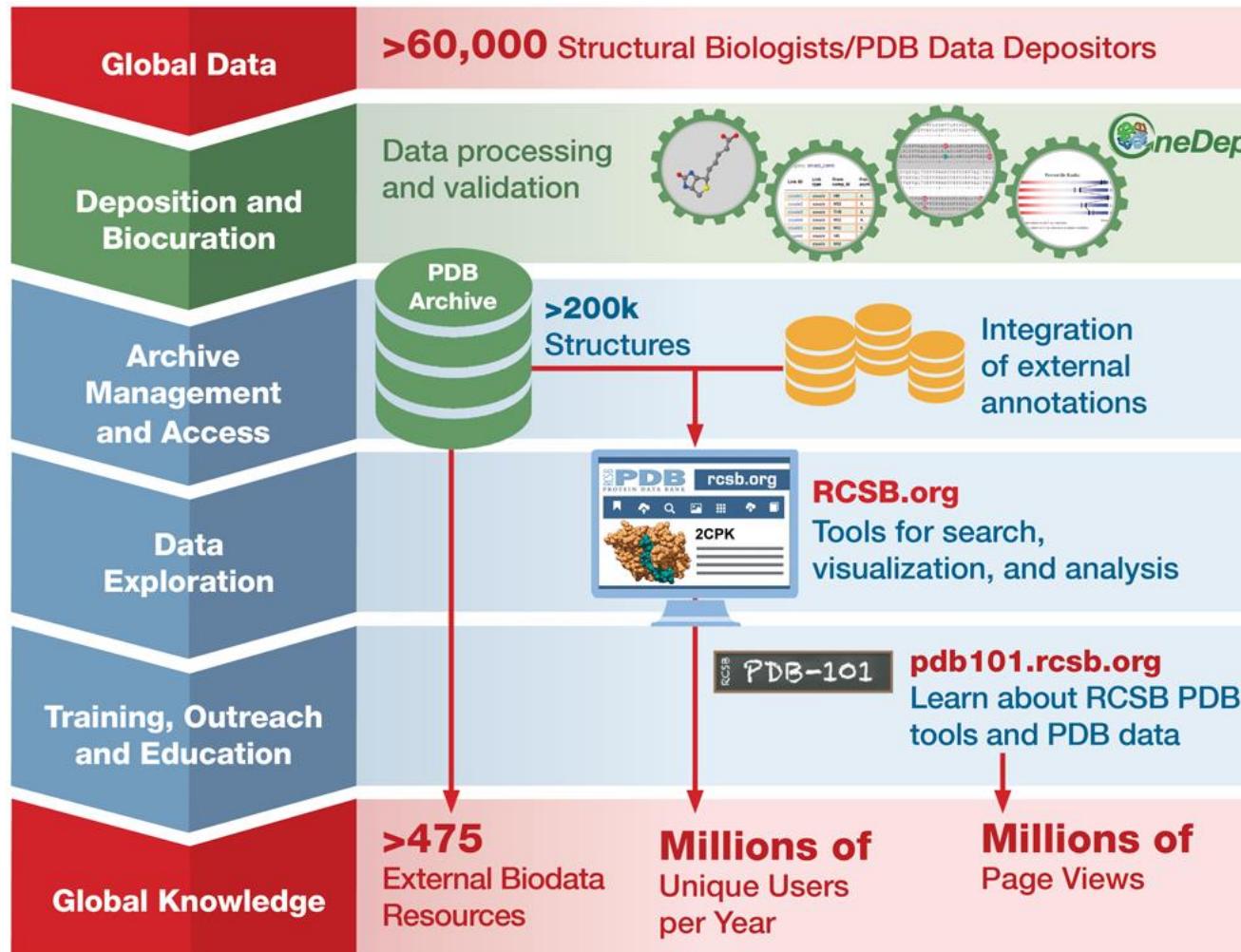
The screenshot shows the main homepage of the RCSB PDB website. At the top, there is a navigation bar with links to Deposit, Search, Visualize, Analyze, Download, Learn, About, Documentation, Careers, COVID-19, MyPDB, and Contact us. Below the navigation bar, the RCSB PDB logo is displayed, along with statistics: 208,347 Structures from the PDB and 1,068,577 Computed Structure Models (CSM). A search bar allows users to enter search terms, ID(s), or sequences, with options to include CSM and perform a search. Below the search bar, there are links to Advanced Search and Browse Annotations, and a Help button.

In the center, a banner announces "New: More Computed Structure Models (CSM) available" with a "Learn more" link. To the left, a sidebar menu includes Welcome, Deposit, Search, Visualize, Analyze, Download, and Learn. The main content area features a section about RCSB Protein Data Bank, highlighting experimentally-determined 3D structures and Computed Structure Models (CSM) from AlphaFold DB and ModelArchive. It also mentions the August Molecule of the Month, which is ATM and ATR Kinases, shown as a purple and pink molecular structure. Below this, there are sections for Latest Entries (listing entry 8G4E, a Green Fluorescence Protein imaged on a cryo-EM imaging scaffold), Features & Highlights (mentioning updated annotation and standardization of peptide residues, PDB NextGen Archive, and DNS name changes), News (mentioning Bragg Your Pattern at IUCr, a new poster available for download, and the summer newsletter), and Publications (mentioning the Bragg Your Pattern at IUCr event).

At the bottom, there are links to PDB at a Glance, CSM at a Glance, and various statistics: 63,919 Structures of Human Sequences, 16,507 Nucleic Acid Containing Structures, 999,251 AlphaFoldDB, and 69,326 ModelArchive. There is also a "More Statistics" link.

<https://www.rcsb.org/>

Global outreach of the Protein Data Bank



Standardization

Format, syntax, data standards

Automation

Software amenable, tools for data management

mmCIF/PDBx

PDB

Extensibility

Method development Basis for other dictionaries

FAIR

Enables Findable, Accessible, Interoperable, and Reusable data

The coordinates files

.pdb

```

ATOM 1 N PRO A 27   18.254 54.186 -22.797 1.00 53.35      N
ATOM 2 CA PRO A 27  17.690 52.844 -22.682 1.00 53.37      C
ATOM 3 C PRO A 27  18.564 51.908 -21.783 1.00 53.24      C
ATOM 4 O PRO A 27  19.115 50.895 -22.257 1.00 50.82      O
ATOM 5 CB PRO A 27 17.659 52.369 -24.147 1.00 51.91      C
ATOM 6 CG PRO A 27 18.673 53.271 -24.894 1.00 52.61      C
ATOM 7 CD PRO A 27 19.231 54.261 -23.898 1.00 53.07      C
ATOM 8 N TYR A 28  18.669 52.246 -20.493 1.00 51.56      N
ATOM 9 CA TYR A 28 19.634 51.581 -19.613 1.00 49.67      C
ATOM 10 C TYR A 28 19.310 50.171 -19.219 1.00 49.67      C
ATOM 11 O TYR A 28 20.210 49.351 -19.134 1.00 49.06      O
ATOM 12 CB TYR A 28 19.819 52.335 -18.325 1.00 48.01      C
ATOM 13 CG TYR A 28 20.340 53.693 -18.509 1.00 48.13      C
ATOM 14 CD1 TYR A 28 21.705 53.916 -18.629 1.00 46.64      C
ATOM 15 CD2 TYR A 28 19.460 54.772 -18.573 1.00 48.85      C
ATOM 16 CE1 TYR A 28 22.204 55.188 -18.792 1.00 48.05      C
ATOM 17 CE2 TYR A 28 19.939 56.058 -18.741 1.00 50.76      C
ATOM 18 CZ TYR A 28 21.319 56.261 -18.850 1.00 50.78      C
ATOM 19 OH TYR A 28 21.781 57.551 -19.020 1.00 52.62      O

```

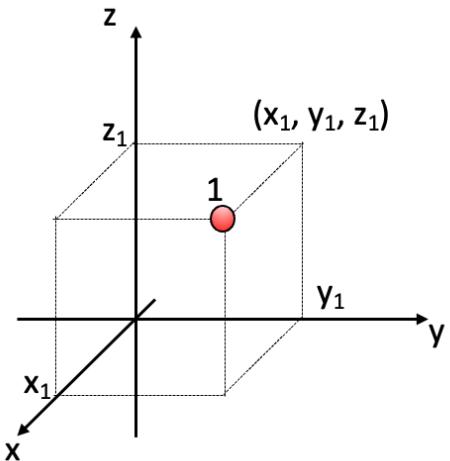
.cif

```

ATOM 1 N N . PRO A 1 27 ? 18.254 54.186 -22.797 1.00 53.35 ? 27 PRO A N 1
ATOM 2 C CA . PRO A 1 27 ? 17.690 52.844 -22.682 1.00 53.37 ? 27 PRO A CA 1
ATOM 3 C C . PRO A 1 27 ? 18.564 51.908 -21.783 1.00 53.24 ? 27 PRO A C 1
ATOM 4 O O . PRO A 1 27 ? 19.115 50.895 -22.257 1.00 50.82 ? 27 PRO A O 1
ATOM 5 C CB . PRO A 1 27 ? 17.659 52.369 -24.147 1.00 51.91 ? 27 PRO A CB 1
ATOM 6 C CG . PRO A 1 27 ? 18.673 53.271 -24.894 1.00 52.61 ? 27 PRO A CG 1
ATOM 7 C CD . PRO A 1 27 ? 19.231 54.261 -23.898 1.00 53.07 ? 27 PRO A CD 1
ATOM 8 N N . TYR A 1 28 ? 18.669 52.246 -20.493 1.00 51.56 ? 28 TYR A N 1
ATOM 9 C CA . TYR A 1 28 ? 19.634 51.581 -19.613 1.00 49.67 ? 28 TYR A CA 1
ATOM 10 C C . TYR A 1 28 ? 19.310 50.171 -19.219 1.00 49.67 ? 28 TYR A C 1
ATOM 11 O O . TYR A 1 28 ? 20.210 49.351 -19.134 1.00 49.06 ? 28 TYR A O 1
ATOM 12 C CB . TYR A 1 28 ? 19.819 52.335 -18.325 1.00 48.01 ? 28 TYR A CB 1
ATOM 13 C CG . TYR A 1 28 ? 20.340 53.693 -18.509 1.00 48.13 ? 28 TYR A CG 1
ATOM 14 C CD1 . TYR A 1 28 ? 21.705 53.916 -18.629 1.00 46.64 ? 28 TYR A CD1 1
ATOM 15 C CD2 . TYR A 1 28 ? 19.460 54.772 -18.573 1.00 48.85 ? 28 TYR A CD2 1
ATOM 16 C CE1 . TYR A 1 28 ? 22.204 55.188 -18.792 1.00 48.05 ? 28 TYR A CE1 1
ATOM 17 C CE2 . TYR A 1 28 ? 19.939 56.058 -18.741 1.00 50.76 ? 28 TYR A CE2 1
ATOM 18 C CZ . TYR A 1 28 ? 21.319 56.261 -18.850 1.00 50.78 ? 28 TYR A CZ 1
ATOM 19 O OH . TYR A 1 28 ? 21.781 57.551 -19.020 1.00 52.62 ? 28 TYR A OH 1

```

Atom indices and type



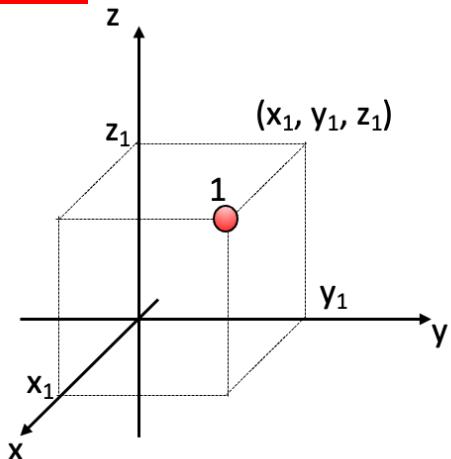
ATOM 1 C CA 1 A x₁ y₁ z₁

The coordinates files

.pdb

x, y and z coordinates

ATOM	1	N	PRO A	27	18.254	54.186	-22.797	1.00	53.35	N
ATOM	2	CA	PRO A	27	17.690	52.844	-22.682	1.00	53.37	C
ATOM	3	C	PRO A	27	18.564	51.908	-21.783	1.00	53.24	C
ATOM	4	O	PRO A	27	19.115	50.895	-22.257	1.00	50.82	O
ATOM	5	CB	PRO A	27	17.659	52.369	-24.147	1.00	51.91	C
ATOM	6	CG	PRO A	27	18.673	53.271	-24.894	1.00	52.61	C
ATOM	7	CD	PRO A	27	19.231	54.261	-23.898	1.00	53.07	C
ATOM	8	N	TYR A	28	18.669	52.246	-20.493	1.00	51.56	N
ATOM	9	CA	TYR A	28	19.634	51.581	-19.613	1.00	49.67	C
ATOM	10	C	TYR A	28	19.310	50.171	-19.219	1.00	49.67	C
ATOM	11	O	TYR A	28	20.210	49.351	-19.134	1.00	49.06	O
ATOM	12	CB	TYR A	28	19.819	52.335	-18.325	1.00	48.01	C
ATOM	13	CG	TYR A	28	20.340	53.693	-18.509	1.00	48.13	C
ATOM	14	CD1	TYR A	28	21.705	53.916	-18.629	1.00	46.64	C
ATOM	15	CD2	TYR A	28	19.460	54.772	-18.573	1.00	48.85	C
ATOM	16	CE1	TYR A	28	22.204	55.188	-18.792	1.00	48.05	C
ATOM	17	CE2	TYR A	28	19.939	56.058	-18.741	1.00	50.76	C
ATOM	18	CZ	TYR A	28	21.319	56.261	-18.850	1.00	50.78	C
ATOM	19	OH	TYR A	28	21.781	57.551	-19.020	1.00	52.62	O



.cif

ATOM	1	N	N	.	PRO A	1	27	?	18.254	54.186	-22.797	1.00	53.35	?	27	PRO A	N	1
ATOM	2	C	CA	.	PRO A	1	27	?	17.690	52.844	-22.682	1.00	53.37	?	27	PRO A	CA	1
ATOM	3	C	C	.	PRO A	1	27	?	18.564	51.908	-21.783	1.00	53.24	?	27	PRO A	C	1
ATOM	4	O	O	.	PRO A	1	27	?	19.115	50.895	-22.257	1.00	50.82	?	27	PRO A	O	1
ATOM	5	C	CB	.	PRO A	1	27	?	17.659	52.369	-24.147	1.00	51.91	?	27	PRO A	CB	1
ATOM	6	C	CG	.	PRO A	1	27	?	18.673	53.271	-24.894	1.00	52.61	?	27	PRO A	CG	1
ATOM	7	C	CD	.	PRO A	1	27	?	19.231	54.261	-23.898	1.00	53.07	?	27	PRO A	CD	1
ATOM	8	N	N	.	TYR A	1	28	?	18.669	52.246	-20.493	1.00	51.56	?	28	TYR A	N	1
ATOM	9	CA	CA	.	TYR A	1	28	?	19.634	51.581	-19.613	1.00	49.67	?	28	TYR A	CA	1
ATOM	10	C	C	.	TYR A	1	28	?	19.310	50.171	-19.219	1.00	49.67	?	28	TYR A	C	1
ATOM	11	O	O	.	TYR A	1	28	?	20.210	49.351	-19.134	1.00	49.06	?	28	TYR A	O	1
ATOM	12	CB	CB	.	TYR A	1	28	?	19.819	52.335	-18.325	1.00	48.01	?	28	TYR A	CB	1
ATOM	13	C	CG	.	TYR A	1	28	?	20.340	53.693	-18.509	1.00	48.13	?	28	TYR A	CG	1
ATOM	14	C	CD1	.	TYR A	1	28	?	21.705	53.916	-18.629	1.00	46.64	?	28	TYR A	CD1	1
ATOM	15	C	CD2	.	TYR A	1	28	?	19.460	54.772	-18.573	1.00	48.85	?	28	TYR A	CD2	1
ATOM	16	C	CE1	.	TYR A	1	28	?	22.204	55.188	-18.792	1.00	48.05	?	28	TYR A	CE1	1
ATOM	17	C	CE2	.	TYR A	1	28	?	19.939	56.058	-18.741	1.00	50.76	?	28	TYR A	CE2	1
ATOM	18	C	CZ	.	TYR A	1	28	?	21.319	56.261	-18.850	1.00	50.78	?	28	TYR A	CZ	1
ATOM	19	O	OH	.	TYR A	1	28	?	21.781	57.551	-19.020	1.00	52.62	?	28	TYR A	OH	1

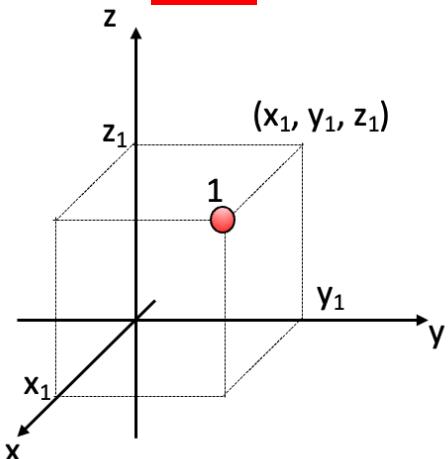
ATOM 1 C CA 1 A x_1 y_1 z_1

The coordinates files

Because proteins are flexible!!!

.pdb

Atomic displacement										
ATOM	1	N	PRO A	27	18.254	54.186	-22.797	1.00	53.35	N
ATOM	2	CA	PRO A	27	17.690	52.844	-22.682	1.00	53.37	C
ATOM	3	C	PRO A	27	18.564	51.908	-21.783	1.00	53.24	C
ATOM	4	O	PRO A	27	19.115	50.895	-22.257	1.00	50.82	O
ATOM	5	CB	PRO A	27	17.659	52.369	-24.147	1.00	51.91	C
ATOM	6	CG	PRO A	27	18.673	53.271	-24.894	1.00	52.61	C
ATOM	7	CD	PRO A	27	19.231	54.261	-23.898	1.00	53.07	C
ATOM	8	N	TYR A	28	18.669	52.246	-20.493	1.00	51.56	N
ATOM	9	CA	TYR A	28	19.634	51.581	-19.613	1.00	49.67	C
ATOM	10	C	TYR A	28	19.310	50.171	-19.219	1.00	49.67	C
ATOM	11	O	TYR A	28	20.210	49.351	-19.134	1.00	49.06	O
ATOM	12	CB	TYR A	28	19.819	52.335	-18.325	1.00	48.01	C
ATOM	13	CG	TYR A	28	20.340	53.693	-18.509	1.00	48.13	C
ATOM	14	CD1	TYR A	28	21.705	53.916	-18.629	1.00	46.64	C
ATOM	15	CD2	TYR A	28	19.460	54.772	-18.573	1.00	48.85	C
ATOM	16	CE1	TYR A	28	22.204	55.188	-18.792	1.00	48.05	C
ATOM	17	CE2	TYR A	28	19.939	56.058	-18.741	1.00	50.76	C
ATOM	18	CZ	TYR A	28	21.319	56.261	-18.850	1.00	50.78	C
ATOM	19	OH	TYR A	28	21.781	57.551	-19.020	1.00	52.62	O

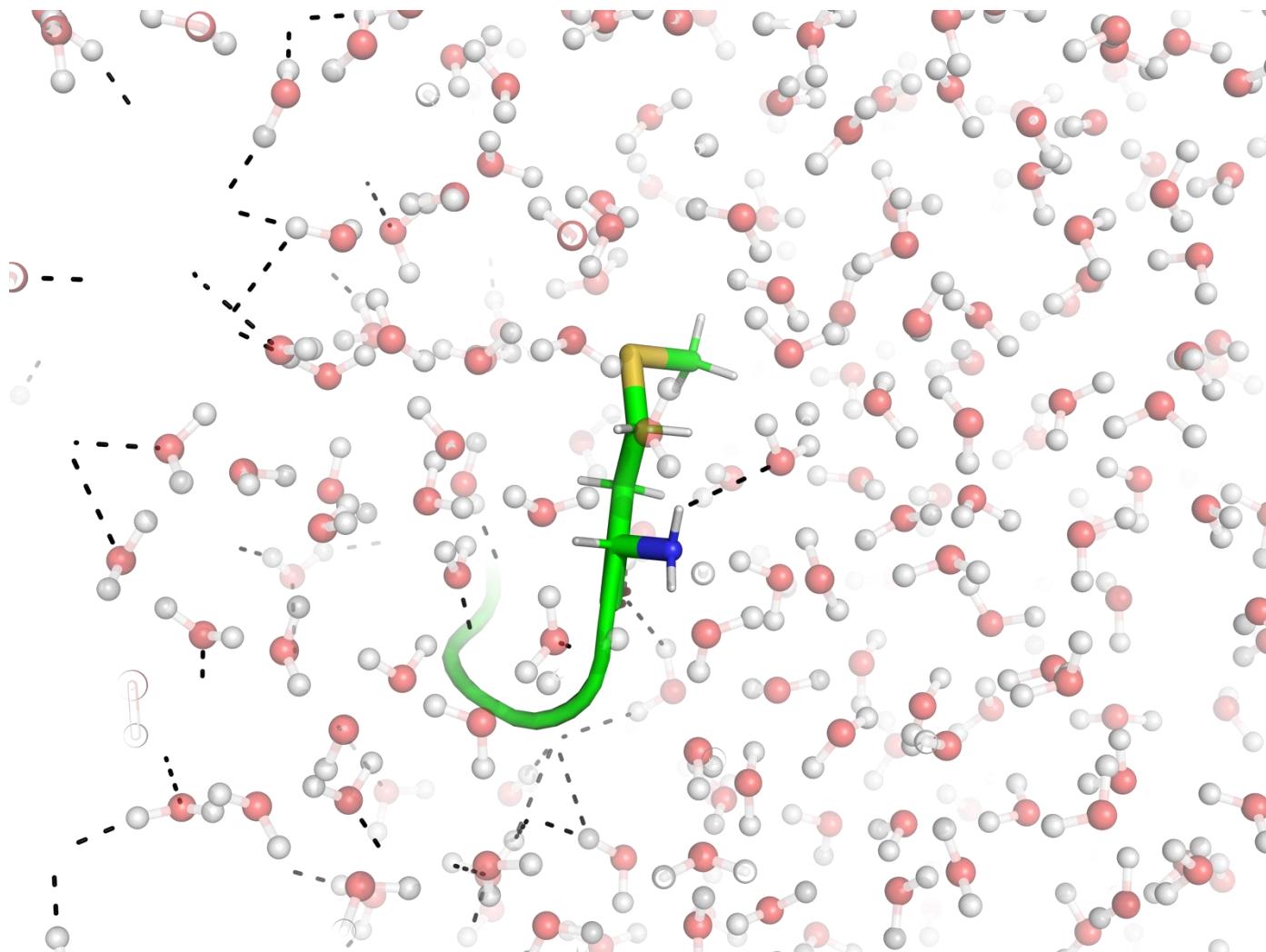


.cif

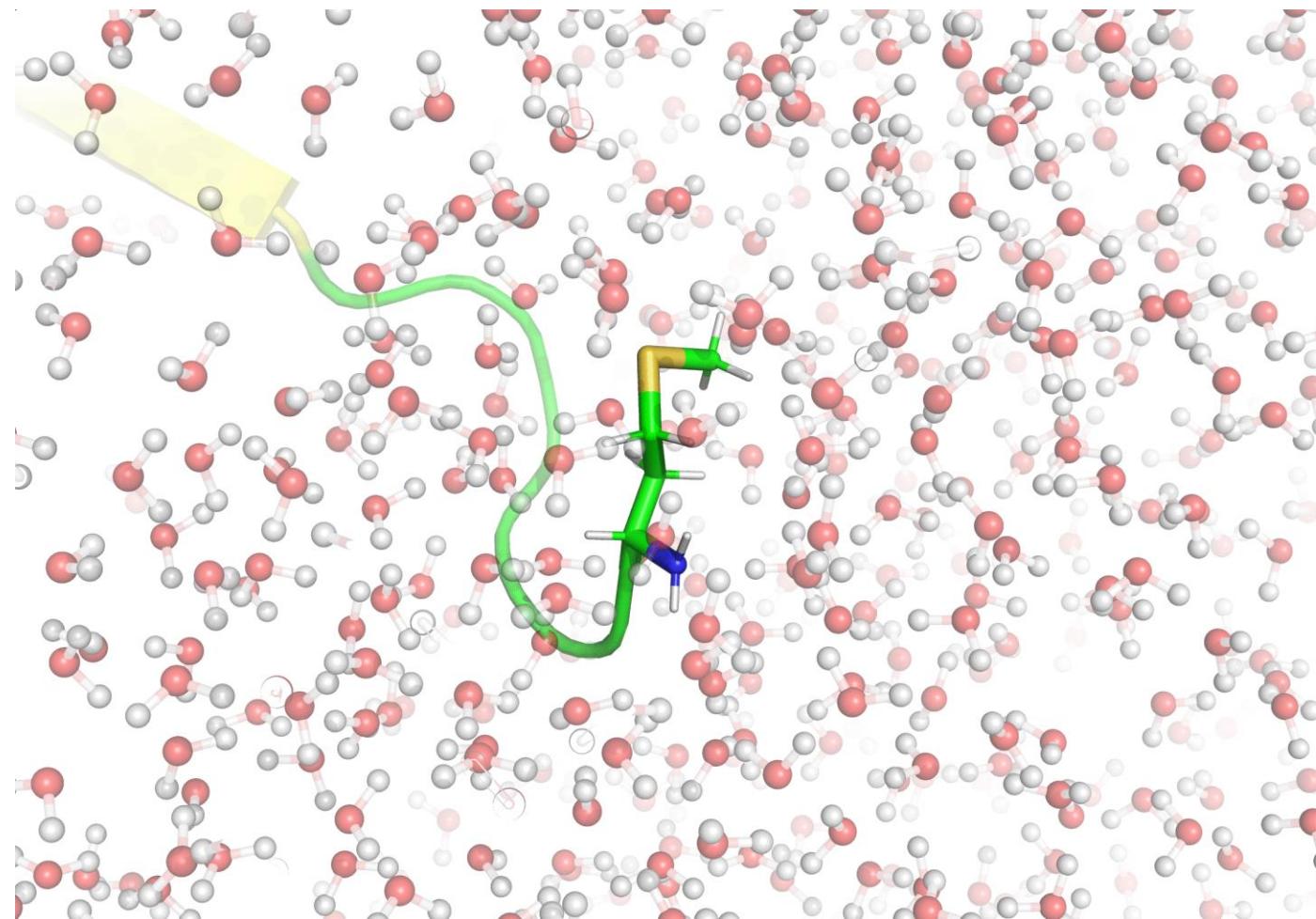
ATOM	1	N	N	.	PRO	A	1	27	?	18.254	54.186	-22.797	1.00	53.35	?	27	PRO	A	N	1
ATOM	2	C	CA	.	PRO	A	1	27	?	17.690	52.844	-22.682	1.00	53.37	?	27	PRO	A	CA	1
ATOM	3	C	C	.	PRO	A	1	27	?	18.564	51.908	-21.783	1.00	53.24	?	27	PRO	A	C	1
ATOM	4	O	O	.	PRO	A	1	27	?	19.115	50.895	-22.257	1.00	50.82	?	27	PRO	A	O	1
ATOM	5	C	CB	.	PRO	A	1	27	?	17.659	52.369	-24.147	1.00	51.91	?	27	PRO	A	CB	1
ATOM	6	C	CG	.	PRO	A	1	27	?	18.673	53.271	-24.894	1.00	52.61	?	27	PRO	A	CG	1
ATOM	7	C	CD	.	PRO	A	1	27	?	19.231	54.261	-23.898	1.00	53.07	?	27	PRO	A	CD	1
ATOM	8	N	N	.	TYR	A	1	28	?	18.669	52.246	-20.493	1.00	51.56	?	28	TYR	A	N	1
ATOM	9	CA	CA	.	TYR	A	1	28	?	19.634	51.581	-19.613	1.00	49.67	?	28	TYR	A	CA	1
ATOM	10	C	C	.	TYR	A	1	28	?	19.310	50.171	-19.219	1.00	49.67	?	28	TYR	A	C	1
ATOM	11	O	O	.	TYR	A	1	28	?	20.210	49.351	-19.134	1.00	49.06	?	28	TYR	A	O	1
ATOM	12	C	CB	.	TYR	A	1	28	?	19.819	52.335	-18.325	1.00	48.01	?	28	TYR	A	CB	1
ATOM	13	C	CG	.	TYR	A	1	28	?	20.340	53.693	-18.509	1.00	48.13	?	28	TYR	A	CG	1
ATOM	14	C	CD1	.	TYR	A	1	28	?	21.705	53.916	-18.629	1.00	46.64	?	28	TYR	A	CD1	1
ATOM	15	C	CD2	.	TYR	A	1	28	?	19.460	54.772	-18.573	1.00	48.85	?	28	TYR	A	CD2	1
ATOM	16	C	CE1	.	TYR	A	1	28	?	22.204	55.188	-18.792	1.00	48.05	?	28	TYR	A	CE1	1
ATOM	17	C	CE2	.	TYR	A	1	28	?	19.939	56.058	-18.741	1.00	50.76	?	28	TYR	A	CE2	1
ATOM	18	C	CZ	.	TYR	A	1	28	?	21.319	56.261	-18.850	1.00	50.78	?	28	TYR	A	CZ	1
ATOM	19	O	OH	.	TYR	A	1	28	?	21.781	57.551	-19.020	1.00	52.62	?	28	TYR	A	OH	1

ATOM 1 C CA 1 A x_1 y_1 z_1

Atomic coordinates in 3D

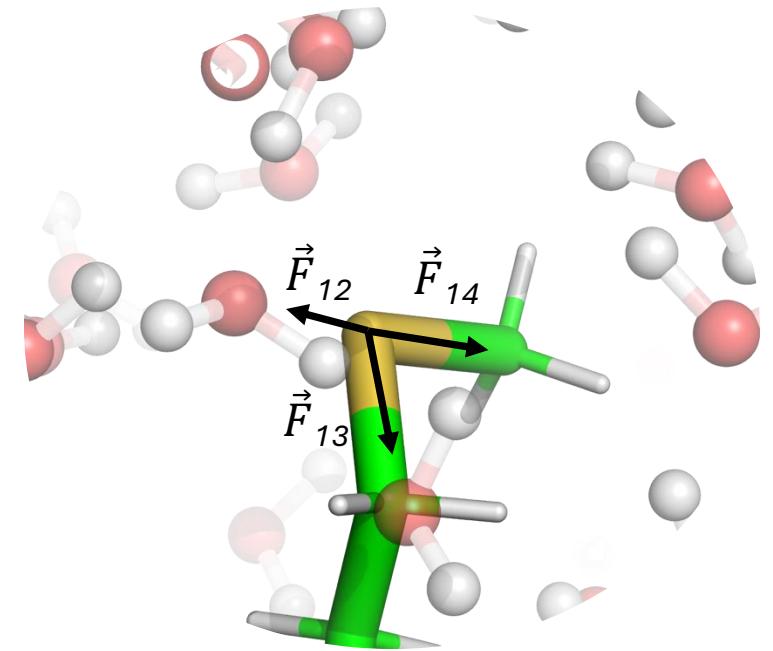
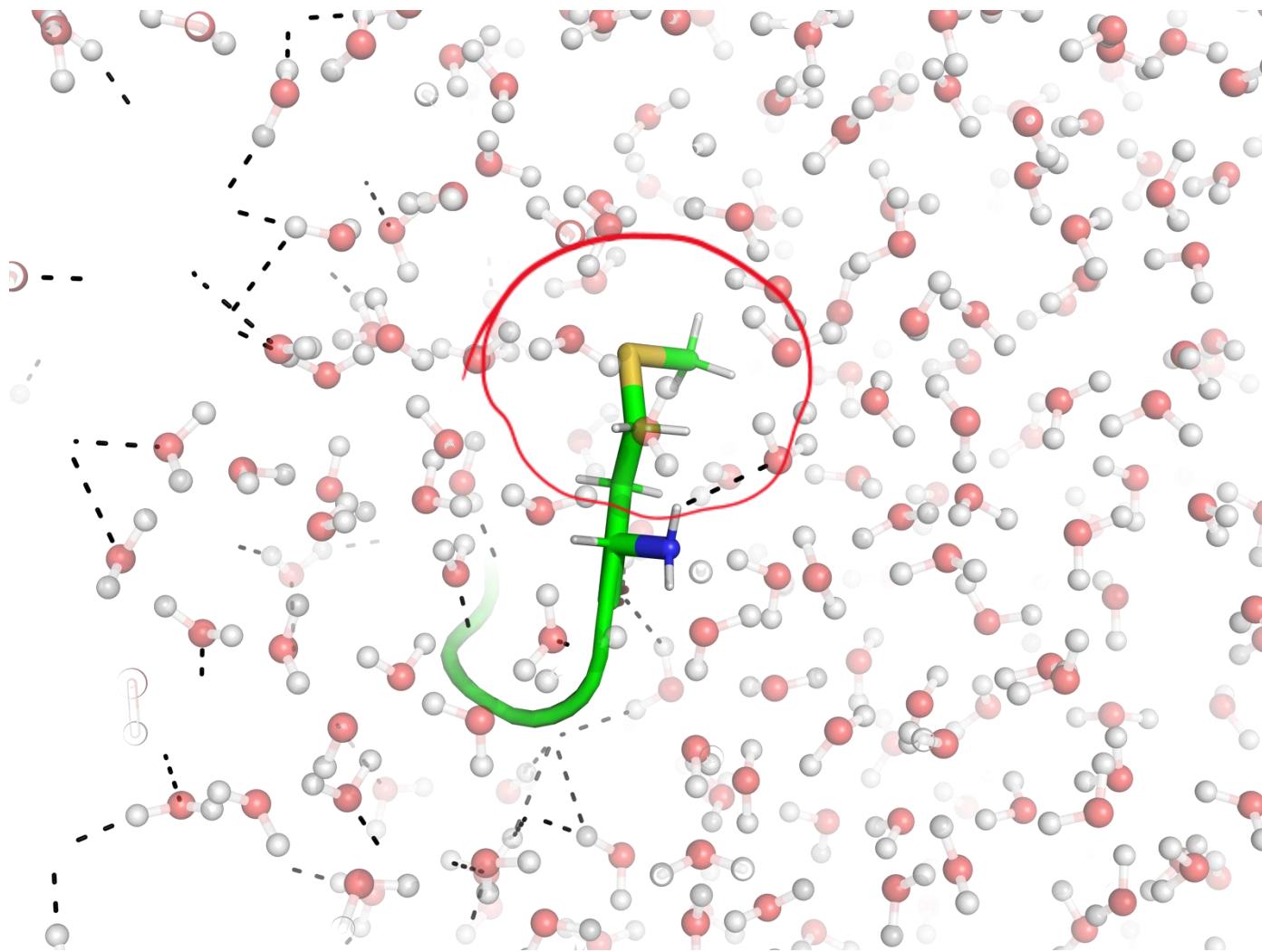


Wigglings and jigglings of atoms



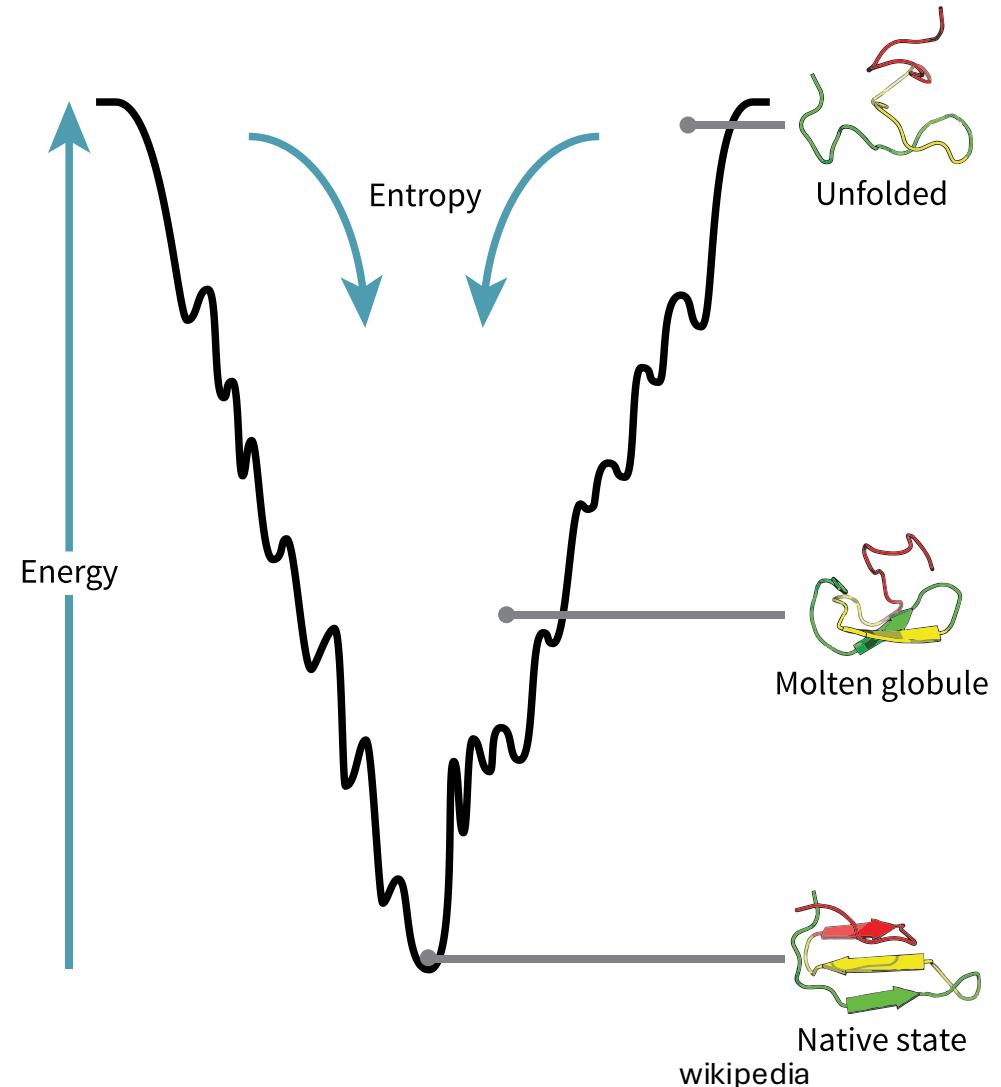
... but 10^{10} times faster!

Atomic forces



$$\overline{E_K} = \frac{1}{2} m \bar{v}^2 = \frac{3}{2} k_B T$$

- The higher the temperature the more agitated are the molecules
- Interactions forming secondary structures are lost with higher temperature

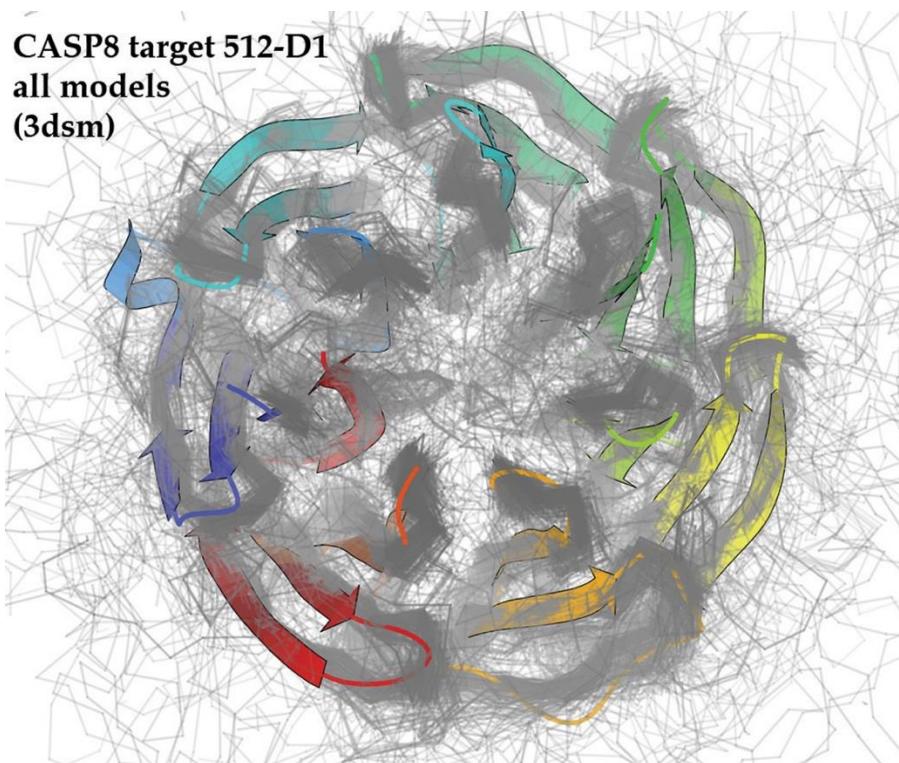


Modeling strategies

- Homology modelling – comparison to homologous proteins
- *De novo* or *ab initio* – prediction from primary structure

CASP

- **Critical Assessment of Structure Prediction**



- Every 2 years since 1994
- An experiment to objectively evaluate the predictions of a community
- Help advance prediction methods
- „world championship“
- Target structures were very recently solved and not yet published

AlphaFold will change everything

nature

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

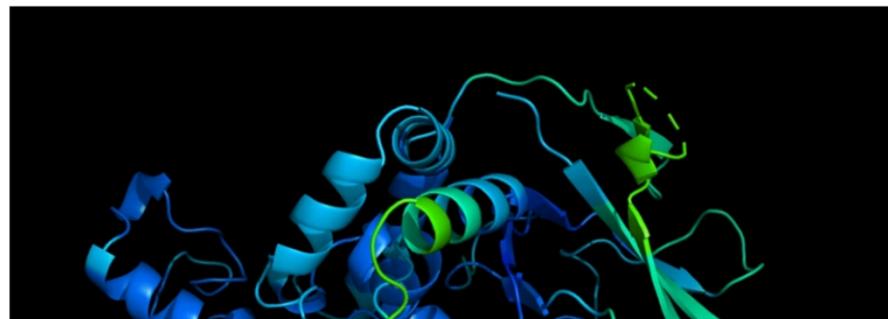
nature > news > article

NEWS | 30 November 2020

'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures

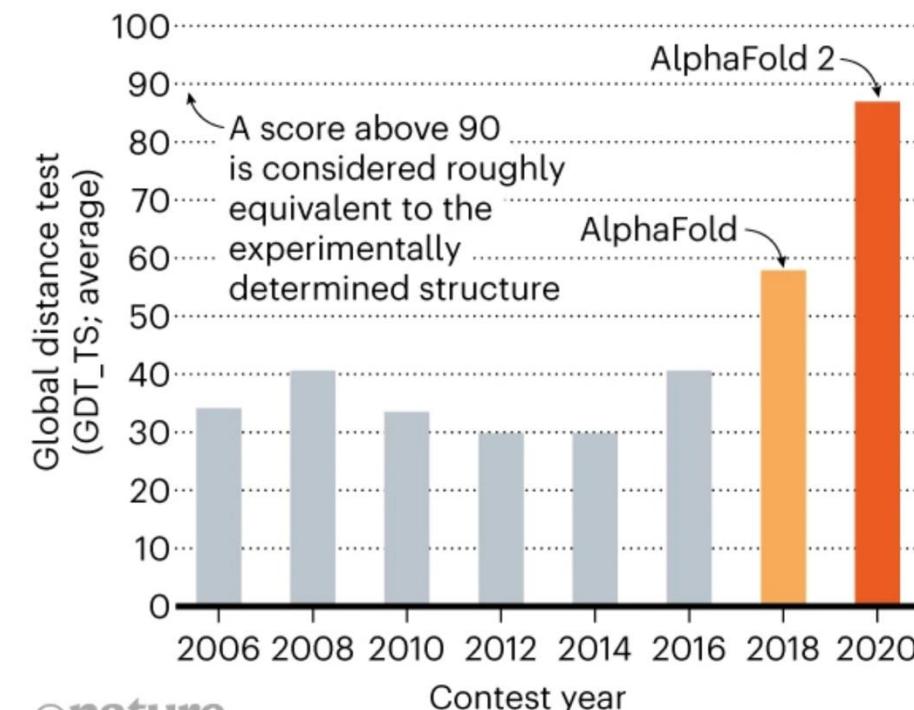
Google's deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.

Ewen Callaway



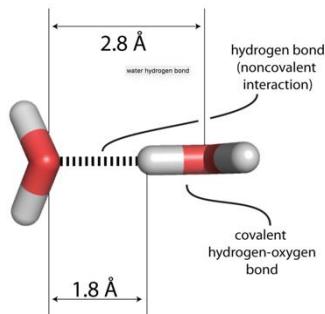
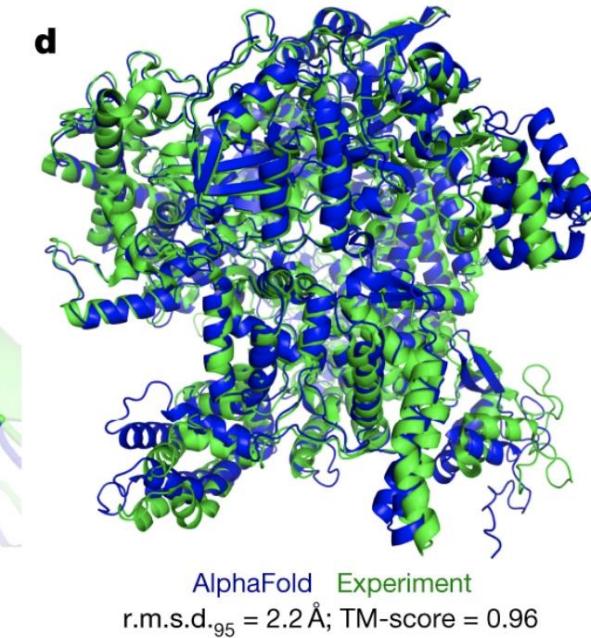
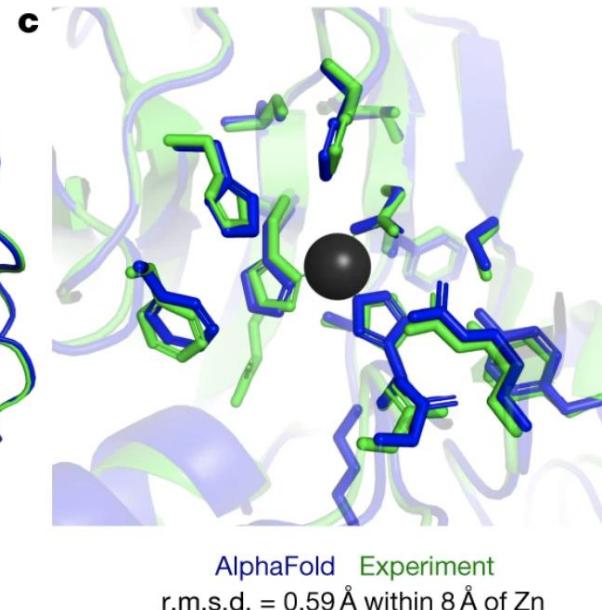
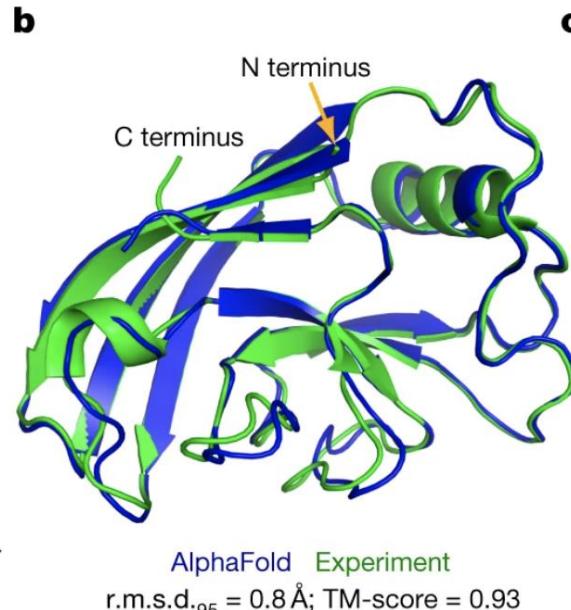
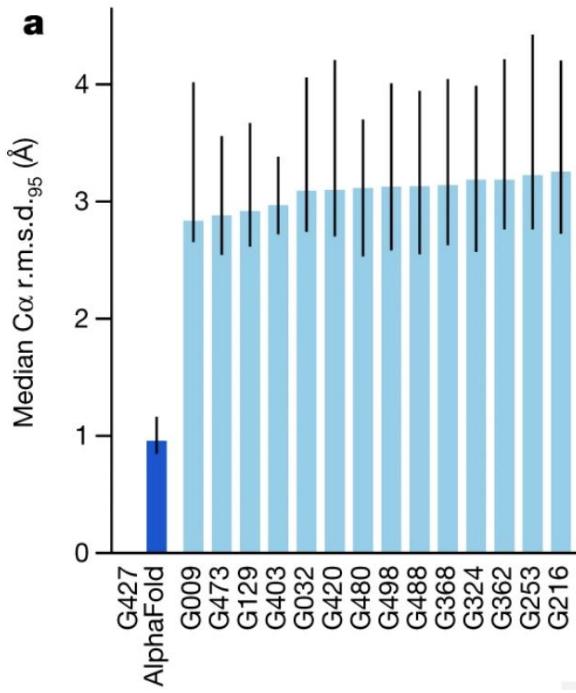
STRUCTURE SOLVER

DeepMind's AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version's performance at the last CASP.



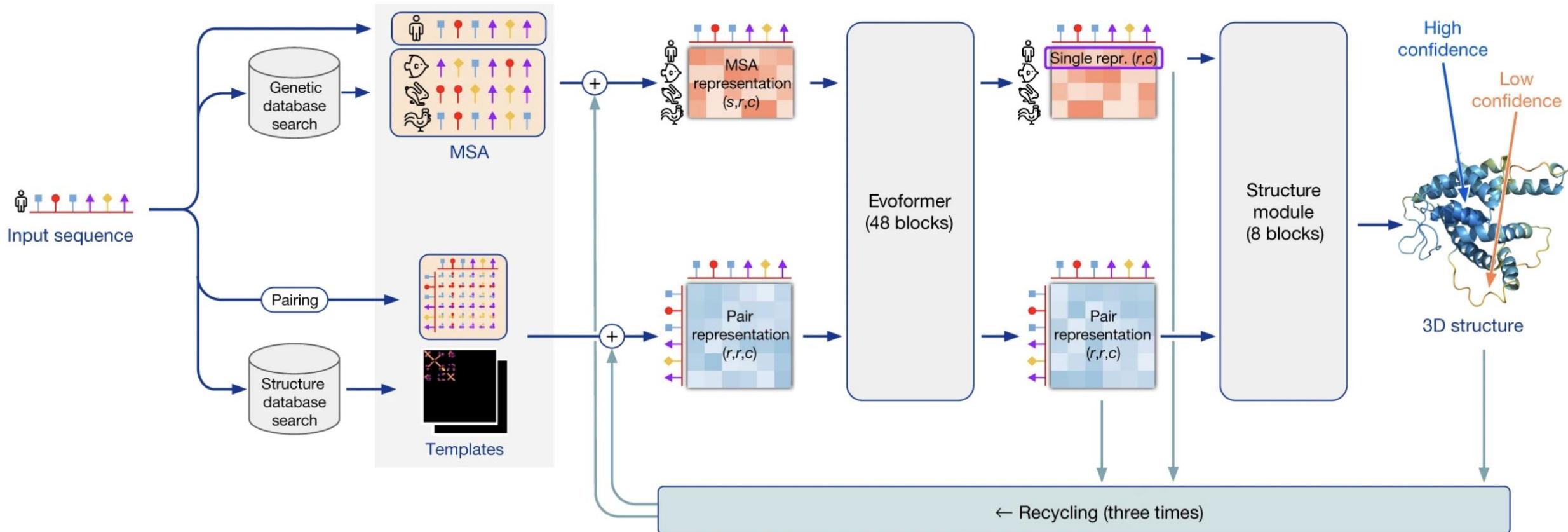
©nature

AlphaFold accuracy



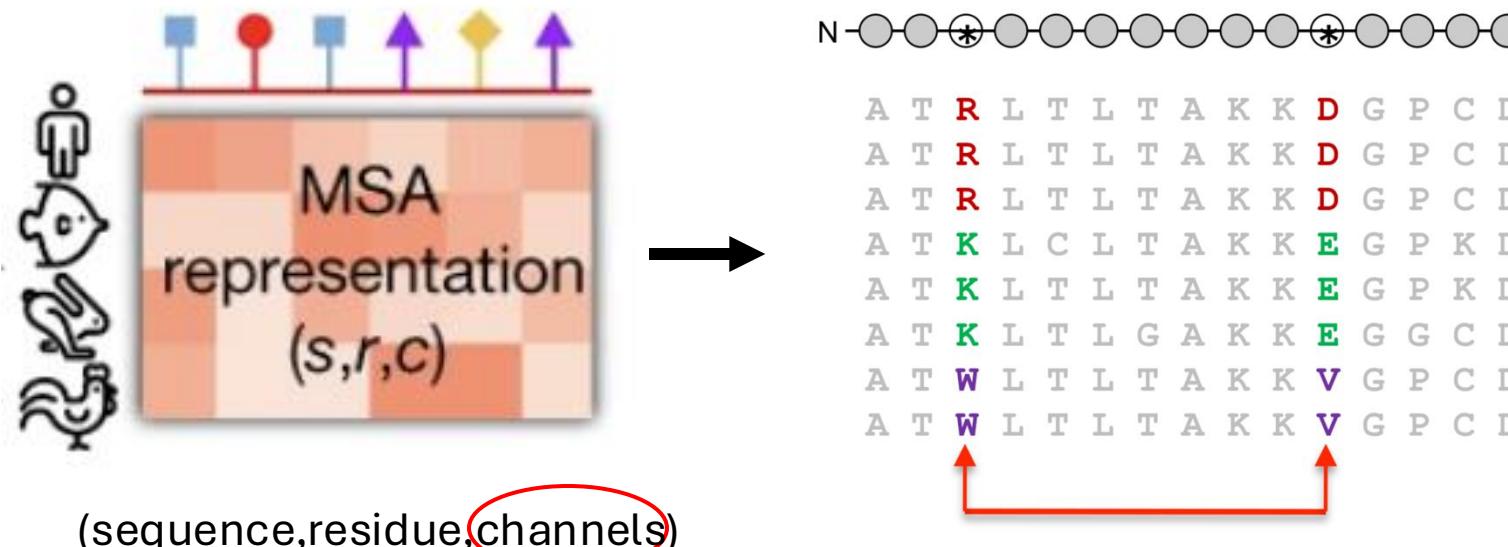
- Compared by RMSD and TM-score

AlphaFold pipeline



Jumper et al., 2021

Multiple Sequence Alignment representation



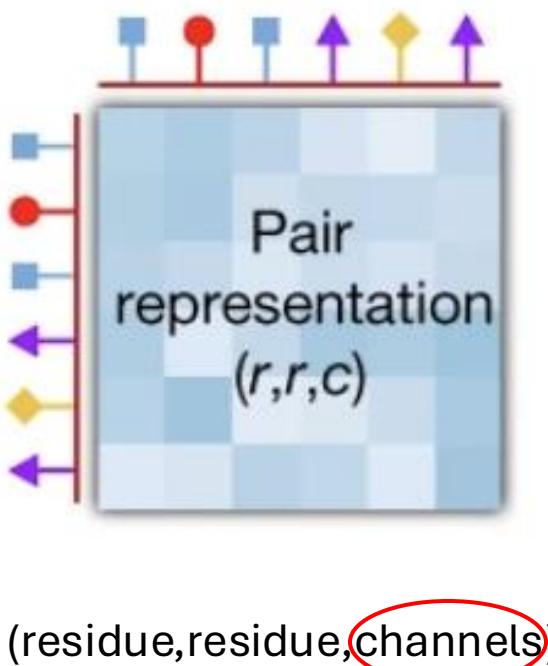
constraint
inference



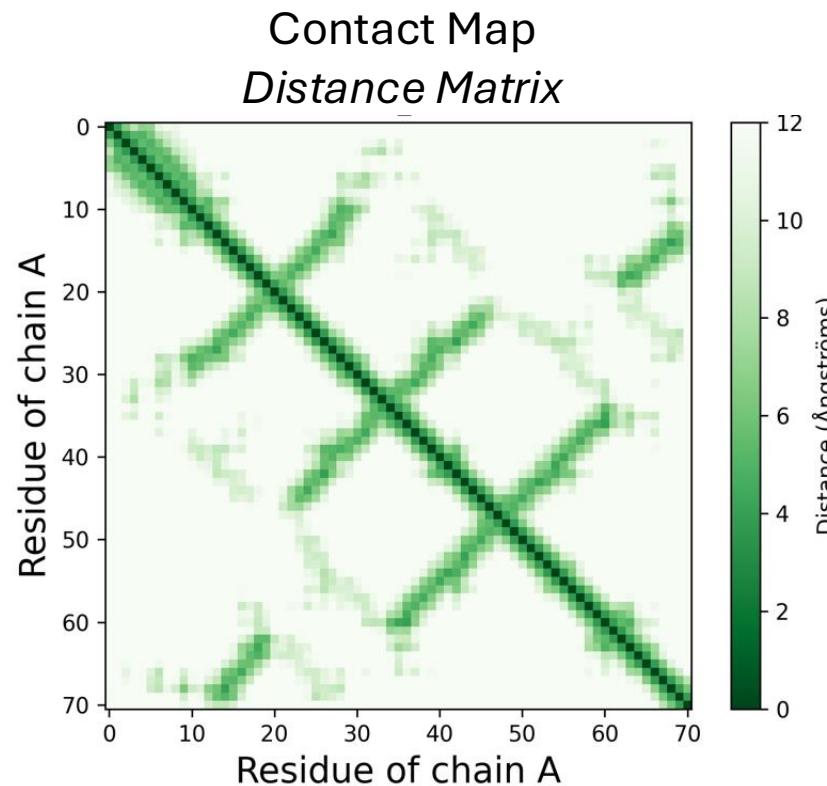
Marks et al., 2011

Many layers with learned features, like more geometric or evolutionary relationships and attention features

Pair representation



Many layers with learned features, like more geometric or evolutionary relationships and attention features



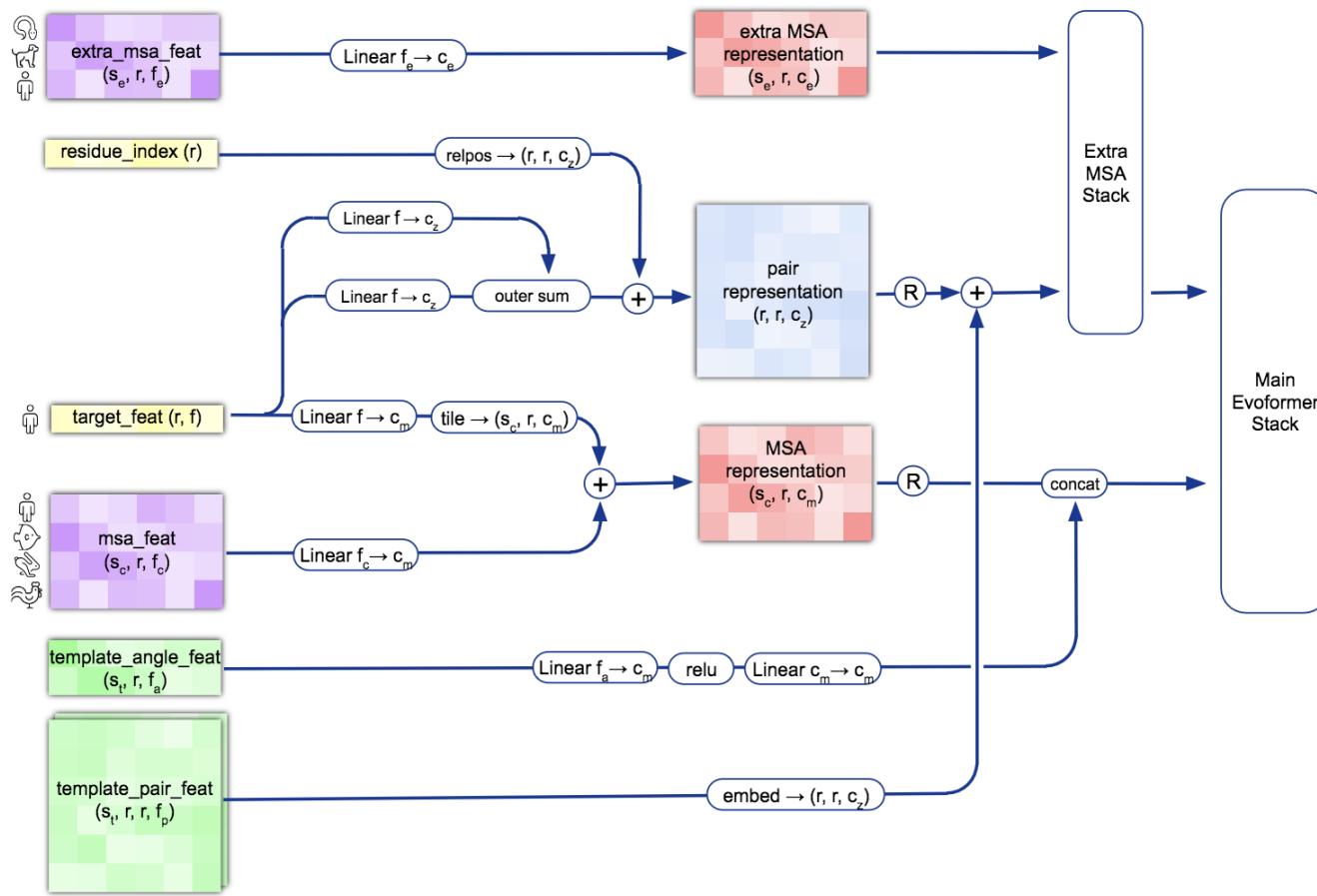
Like a contact map, but with more than 100x more information

Keys to success of AlphaFold2?

- A great database
 - **Big Fantastic Database** - 65,983,866 families , 2,204,359,010 protein sequences from reference databases, metagenomes and metatranscriptomes
 - **PDB**
 - **PDB70**
 - **Uniref90**
 - **Uniclust30**
 - **Uniprot**
 - **MGnify**

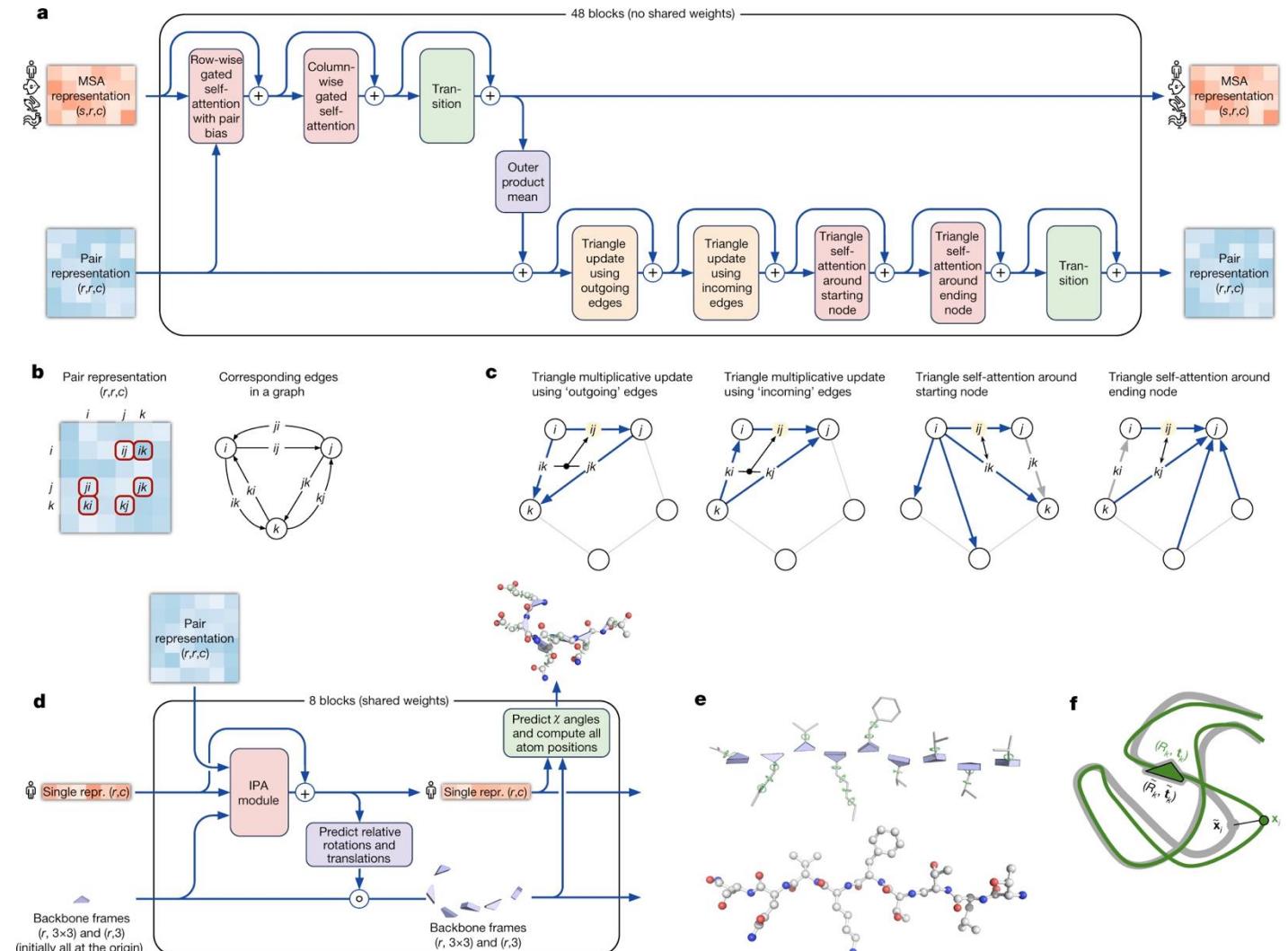
Keys to success of AlphaFold2?

- Ability to handle complex (a great team of experts and a great infrastructure)



Keys to success of AlphaFold2?

- A neural network that alternates between structural and geometrical data and evolutionary data (Evoformer)
- And other clever architectural tricks and assumptions like a „gas of residues“



More on how Alphafold works

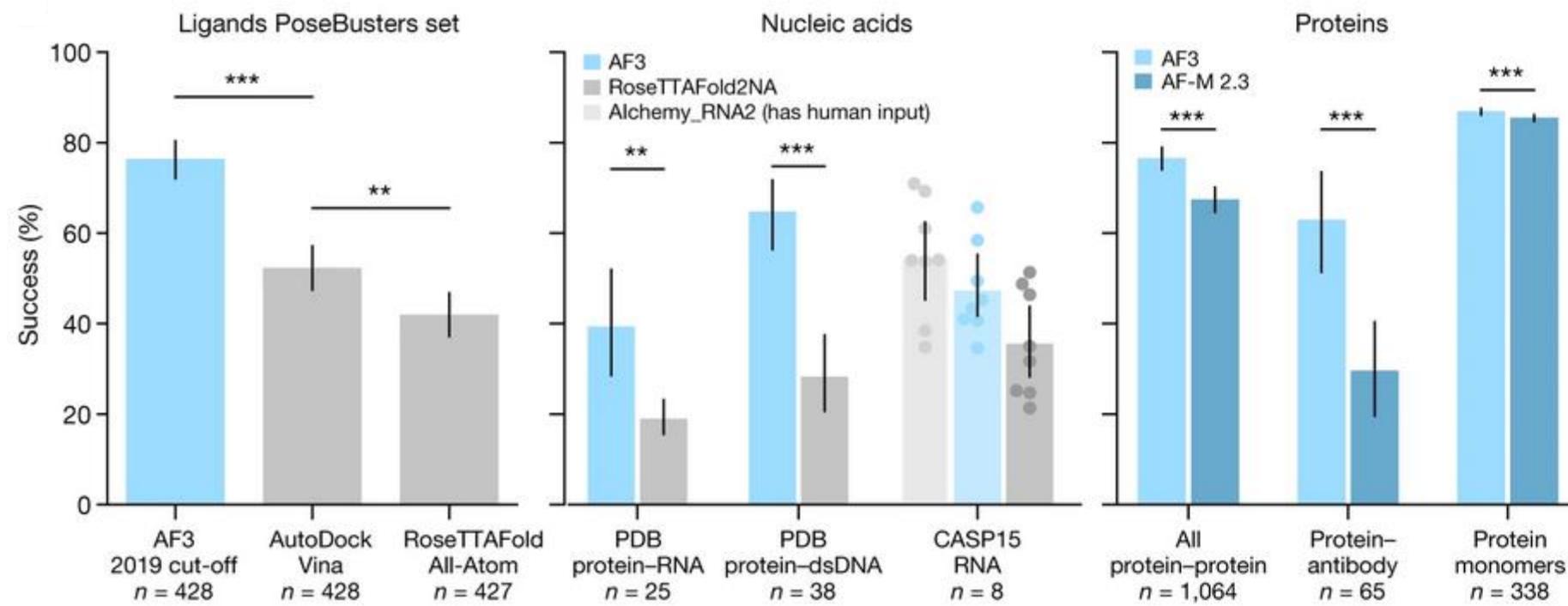
- **Highly Accurate Protein Structure Prediction with AlphaFold | SimonKohl**
 - <https://www.youtube.com/watch?v=tTN0MM2CQLU>
- Oxford Protein Informatics Group - **AlphaFold 2 is here: what's behind the structure prediction miracle**
 - <https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/>

Novelties of AlphaFold3

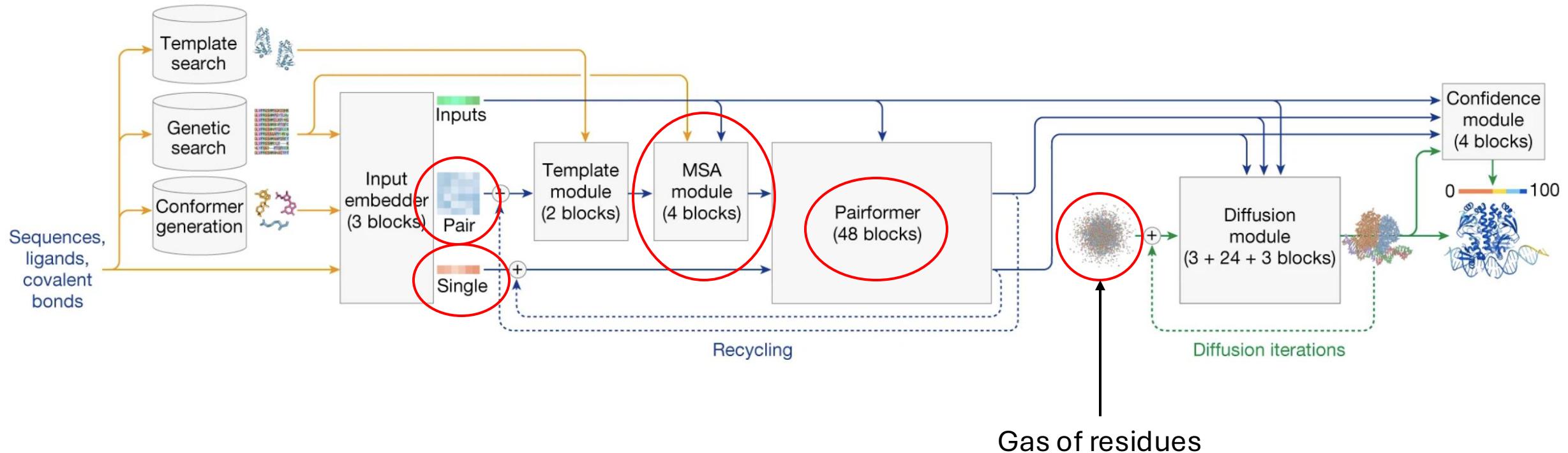


Prediction of proteins with DNA and some small ligands.

Improvements of AlphaFold3



Novelties of AlphaFold3



Interpretation of predicted structures

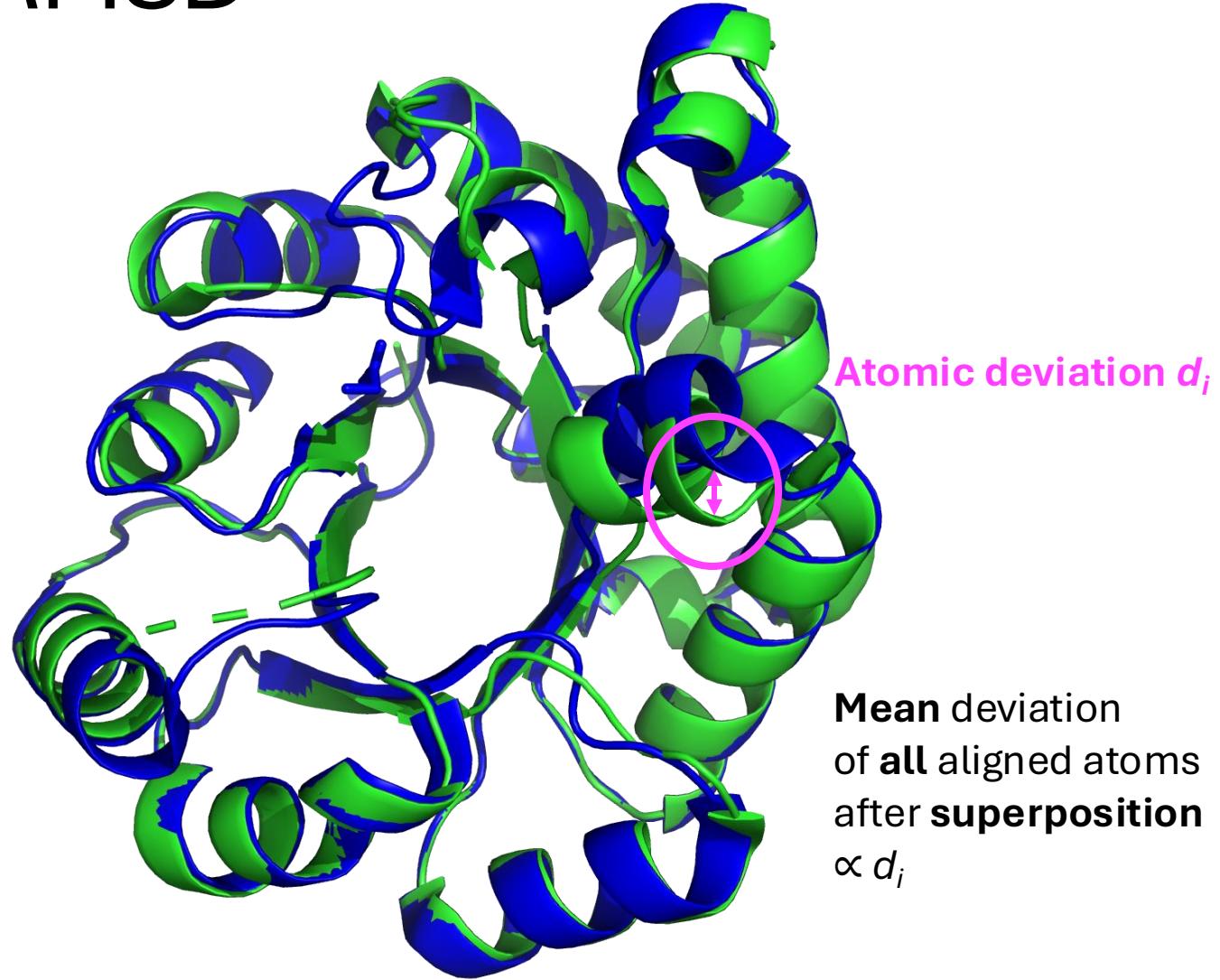
Usual metrics

- RMSD – Root Mean Square Deviation
- TM-score – Template Modelling score
- LDDT – Local Distance Difference Test
- GDT – Global Distance Test (used by CASP)

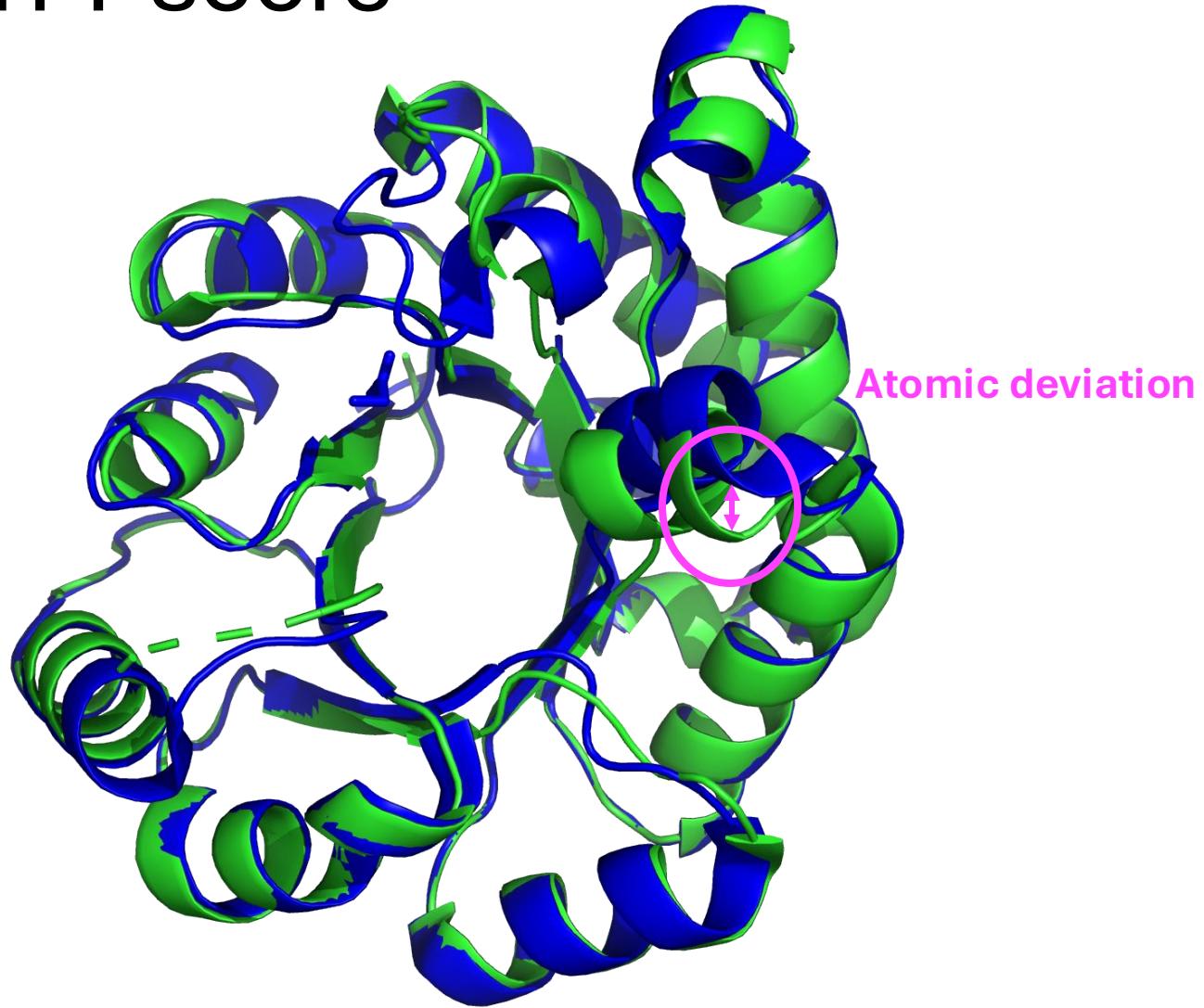
AlphaFold metrics

- pLDDT – predicted LDDT
- PAE – Predicted Aligned Errors

RMSD



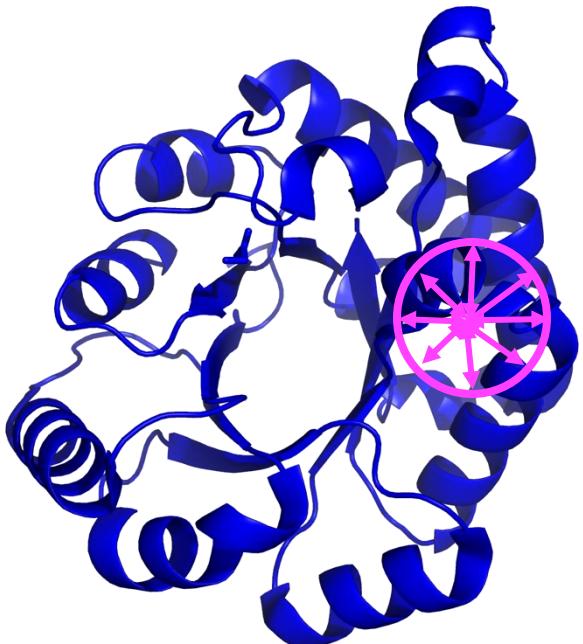
TM-score



A **score** between (0,1]
considering **all** aligned
atoms after **superposition**
 $\propto 1/d_i$

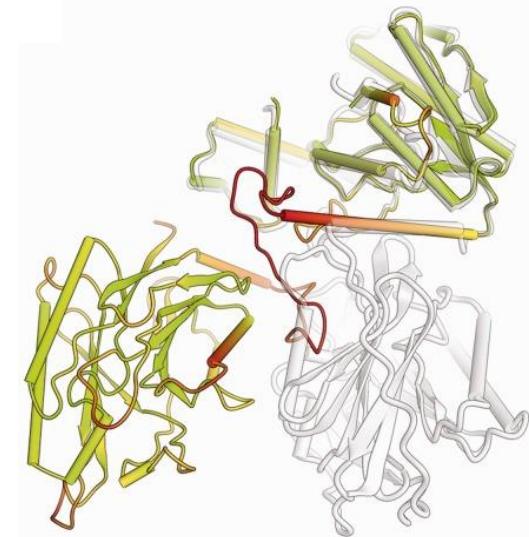
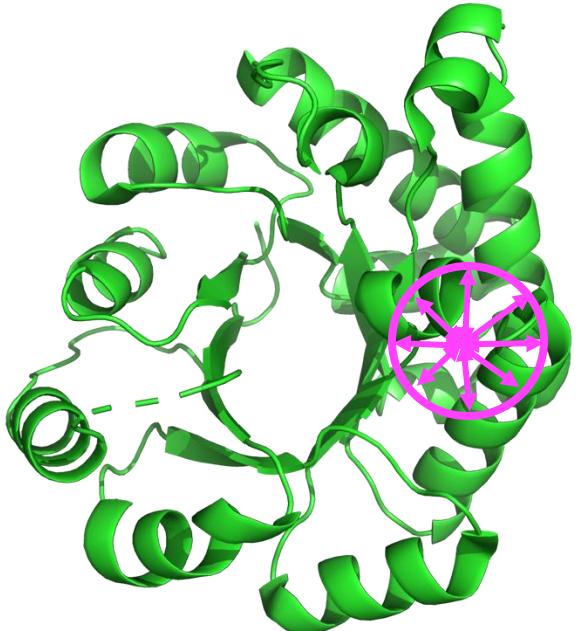
Intended to be more
accurate because
smaller deviations weigh
more than larger
deviations.

LDDT



List of d_i within a threshold

Superposition free,
percentage of conserved
local inter-atomic distances



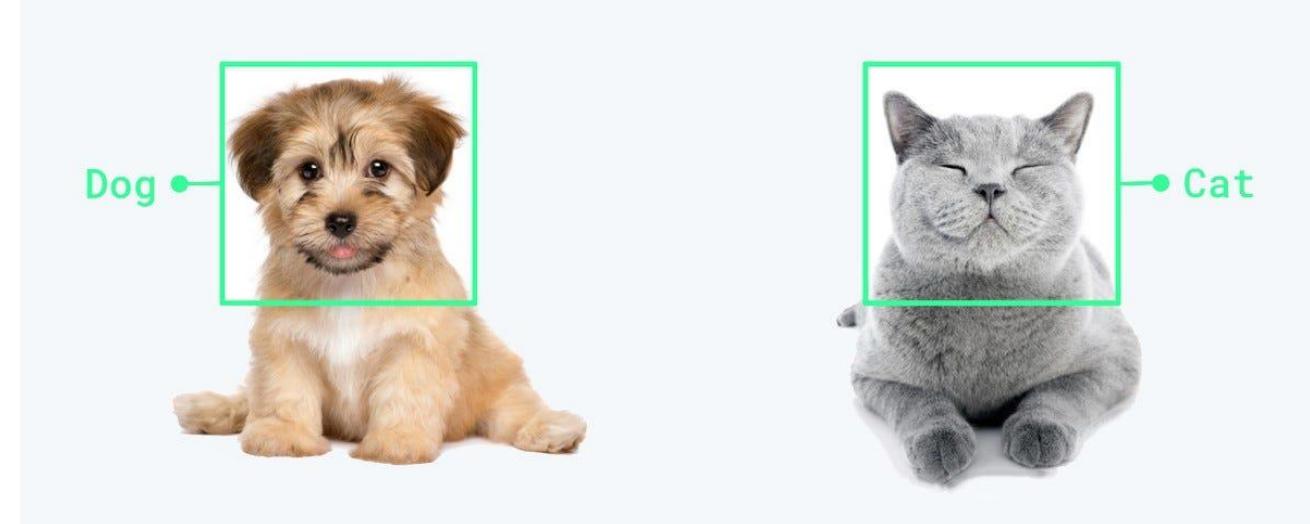
Mariani et al., 2013

Good for comparison of
structures with separate
domains

Correct in relation to others

Ground truth

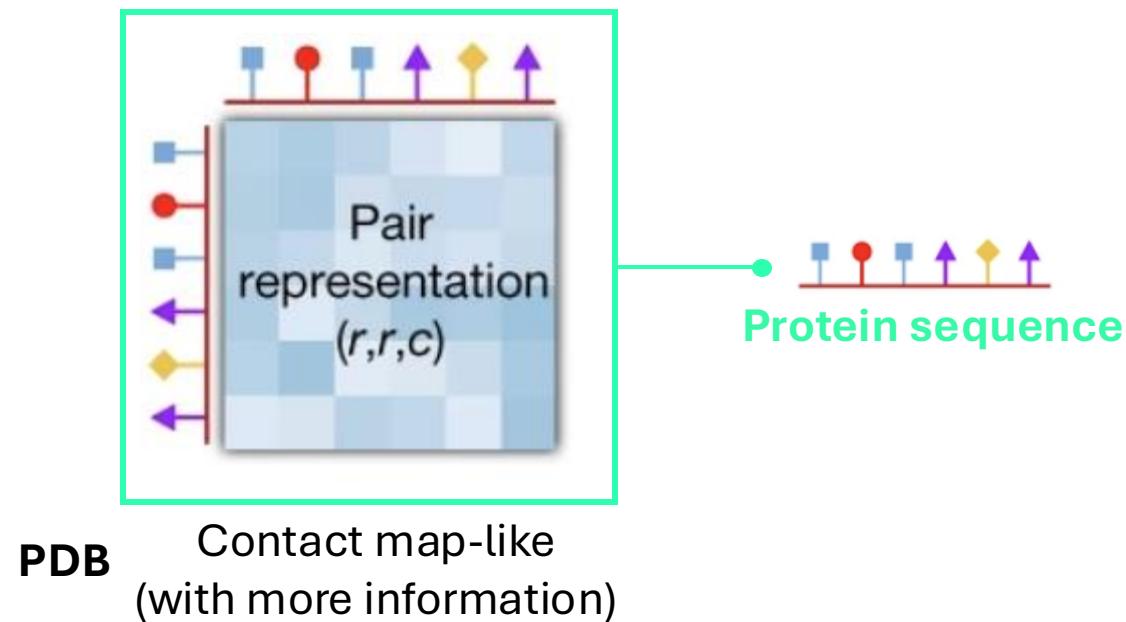
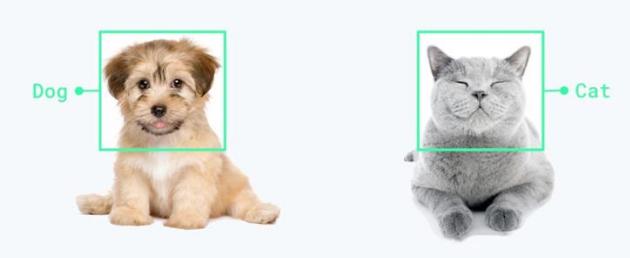
Datasets and labels



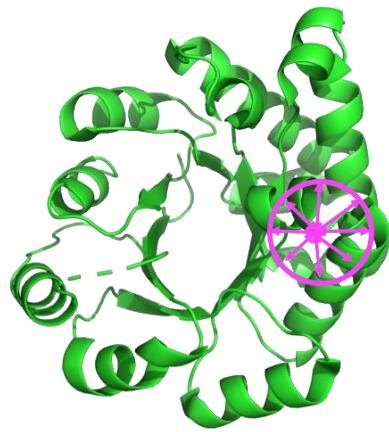
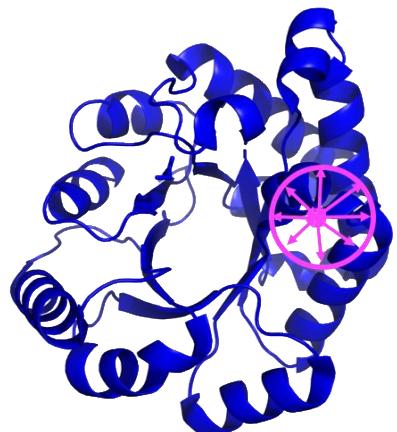
Correct in relation to others

Ground truth

Datasets and labels



pLDDT



Ground truth

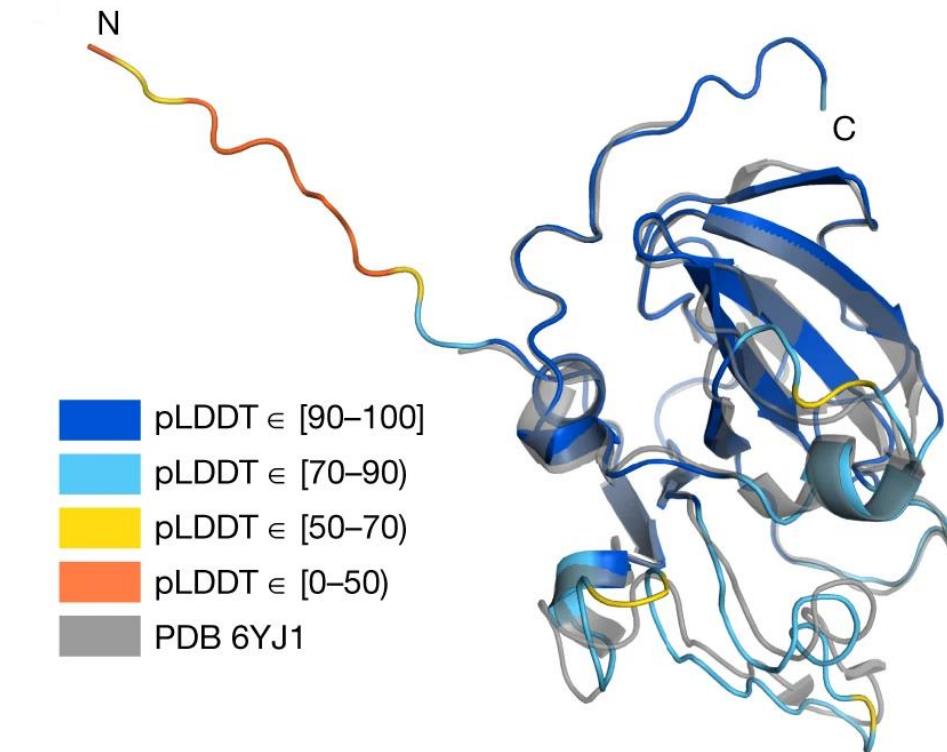
pLDDT > 90 – high confidence

pLDDT > 70 – confident / generally correct backbone

pLDDT < 70 – low confidence

pLDDT < 50 - not confident / disordered

Per residue score

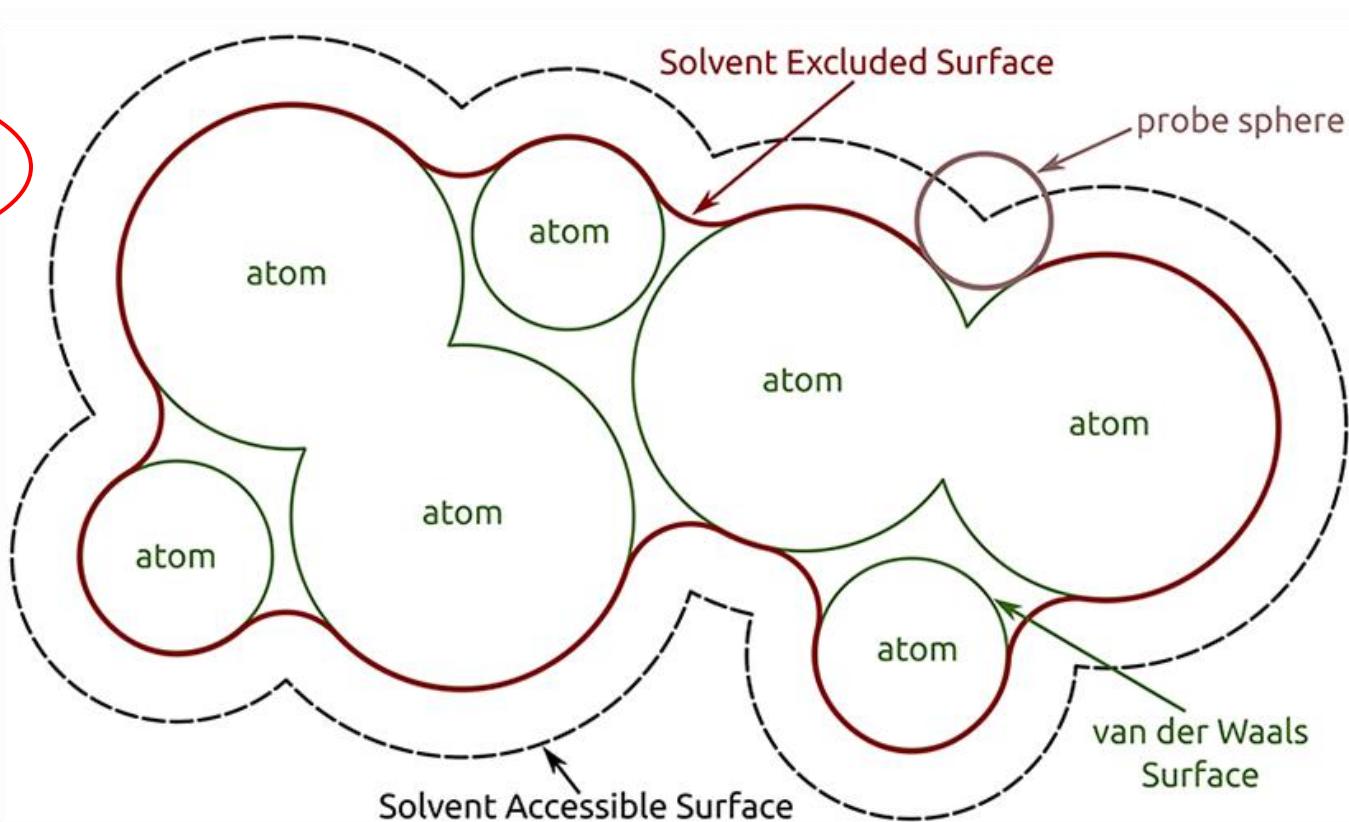


- █ pLDDT ∈ [90–100]
- █ pLDDT ∈ [70–90)
- █ pLDDT ∈ [50–70)
- █ pLDDT ∈ [0–50)
- █ PDB 6YJ1

RASA with AlphaFold3

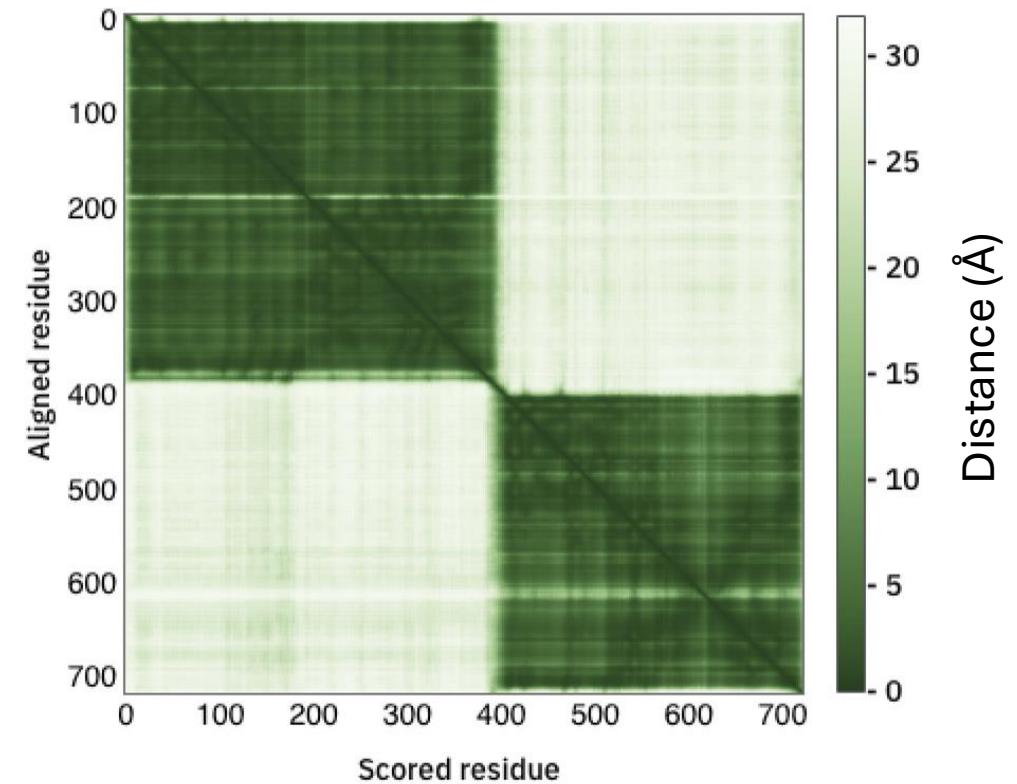
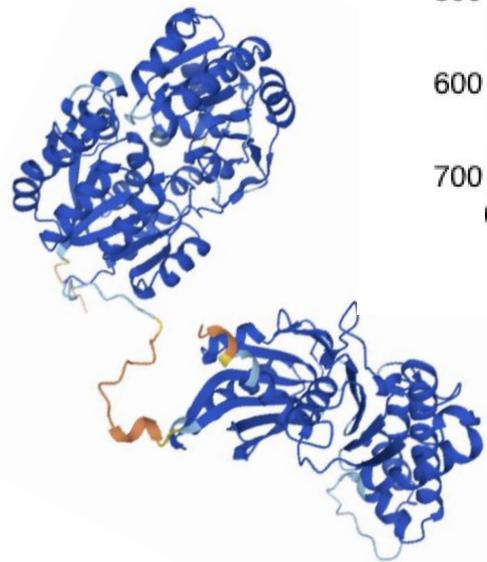
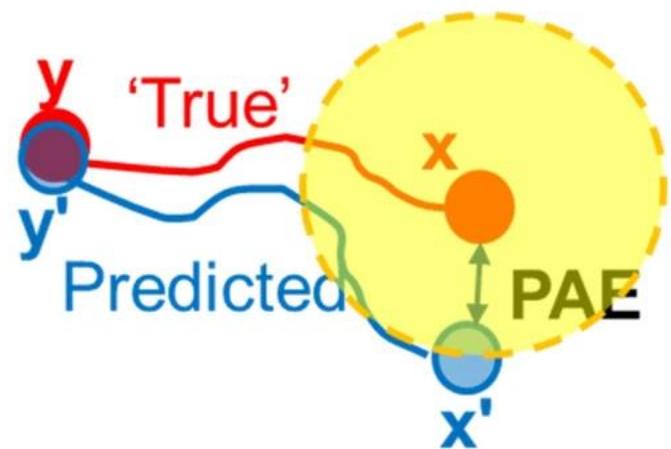
- Relative solvent Accessible Surface Area

Compared to extended conformation



PAE

Predicted aligned error (PAE)



What are the difficulties of AlphaFold?

- Proteins that are experimentally hard to solve are also hard to predict because they lack background data
- The effects of single point mutations are underestimated
- Multimeric predictions will always give a result, even when there is no real interaction

Did AlphaFold solve how proteins fold?

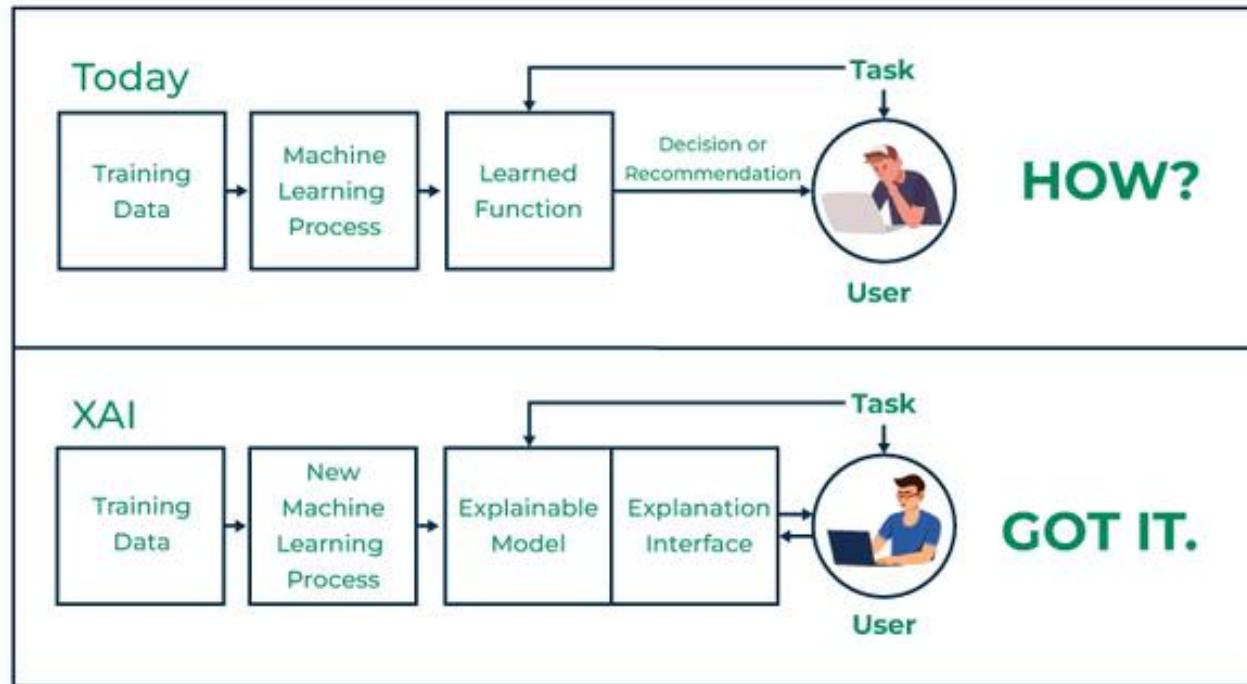
No, but it helps a lot!

The screenshot shows the AlphaFold Protein Structure Database homepage. The title "AlphaFold Protein Structure Database" is prominently displayed in white on a blue background. Below the title, it says "Developed by Google DeepMind and EMBL-EBI". A search bar at the top allows users to "Search for protein, gene, UniProt accession or organism or sequence search". Below the search bar are several examples: MENFQKVEKIGEGTYGV..., Free fatty acid receptor 2, At1g58602, Q5VSL9, and E. coli. At the bottom, there are links for "See search help" and "Go to online course", along with a note about updates: "See our updates – March 2025".

Millions of protein structures
Several entire proteomes

The screenshot shows the AlphaFold Server homepage. The title "AlphaFold Server" is at the top, followed by "Powered by AlphaFold 3". A "Continue with Google" button is visible. The background features a colorful, abstract representation of protein structures. At the bottom, it says "AlphaFold 3 model is a Google DeepMind and Isomorphic Labs collaboration".

Explainable AI



If an AI predicts that a loan should be denied, XAI might reveal:

- 60% of the decision was due to low income,
- 30% due to high debt,
- 10% due to recent credit inquiries.

Transparency, interpretability, justification and trust

Coffee break

***“All models are wrong,
but some are useful”.***

George Box

AlphaFold outputs many models

Model	initial training	first fine-tuning		second fine-tuning				
	1	1.1	1.2	1.1.1	1.1.2	1.2.1	1.2.2	1.2.3
Parameters initialized from	Random	Model 1	...	Model 1.1	...	Model 1.2
Number of templates N_{templ}	4	4	0	4	...	0
Sequence crop size N_{res}	256	384
Number of sequences N_{seq}	128	512
Number of extra sequences $N_{\text{extra_seq}}$	1024	5120	1024	5120	...	1024
Initial learning rate	10^{-3}	$5 \cdot 10^{-4}$
Learning rate linear warm-up samples	128000	0
Structural violation loss weight	0.0	1.0
“Experimentally resolved” loss weight	0.0	0.01
Training samples ($\cdot 10^6$)	9.2	1.1	1.7	0.3	0.6	1.4	1.1	2.4
Training time	6d 6h	1d 10h	2d 3h	20h	1d 13h	4d 1h	3d	5d 12h



5 models

AlphaFold Multimer runs each of the 5 models with 5 different random seeds for MSA sampling, resulting in 25 different models.

equations

- Rmsd

$$\begin{aligned}\text{RMSD}(\mathbf{v}, \mathbf{w}) &= \sqrt{\frac{1}{n} \sum_{i=1}^n \|v_i - w_i\|^2} \\ &= \sqrt{\frac{1}{n} \sum_{i=1}^n ((v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2)}\end{aligned}$$

- Tm score

$$\text{TM-score} = \max \left[\frac{1}{L_{\text{target}}} \sum_i^{L_{\text{common}}} \frac{1}{1 + \left(\frac{d_i}{d_0(L_{\text{target}})} \right)^2} \right]$$

- https://en.wikipedia.org/wiki/Template_modeling_score

- iptm

<https://www.biorxiv.org/content/10.1101/2021.10.04.463034v1.full.pdf>

Pillars of function

