Structural Bioinformatics

Elodie Laine

Master BIM-BMC Semestre 3, 2022-2023

Laboratoire de Biologie Computationnelle et Quantitative (LCQB)
e-documents: http://www.lcqb.upmc.fr/laine/STRUCT
e-mail: *elodie.laine@sorbonne-universite.fr*
@LaineElodie

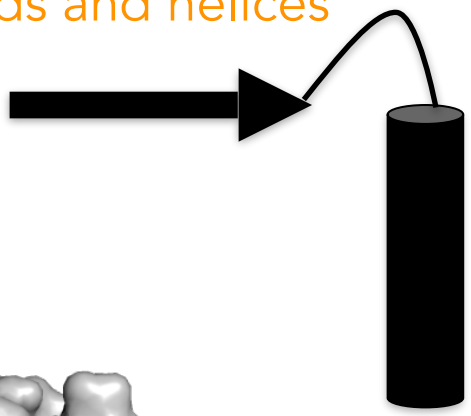
Introduction

What is a protein?

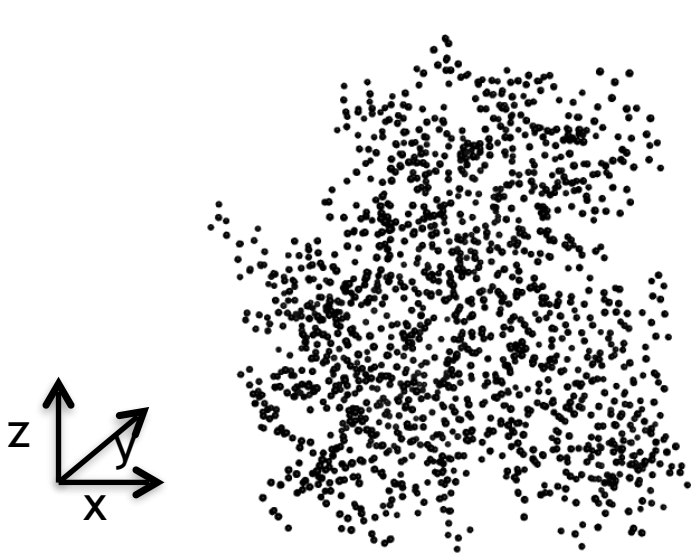
1-Dimensional text

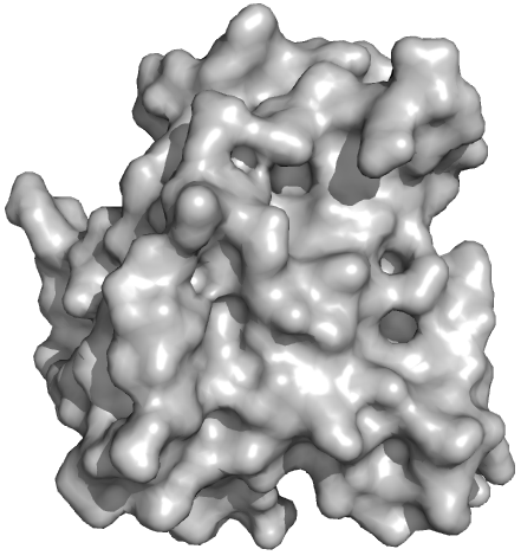
WPLSSSVPSQKTYQGSYGFRLGFLH

2-Dimensional series of strands and helices



3-Dimensional set of points/shape



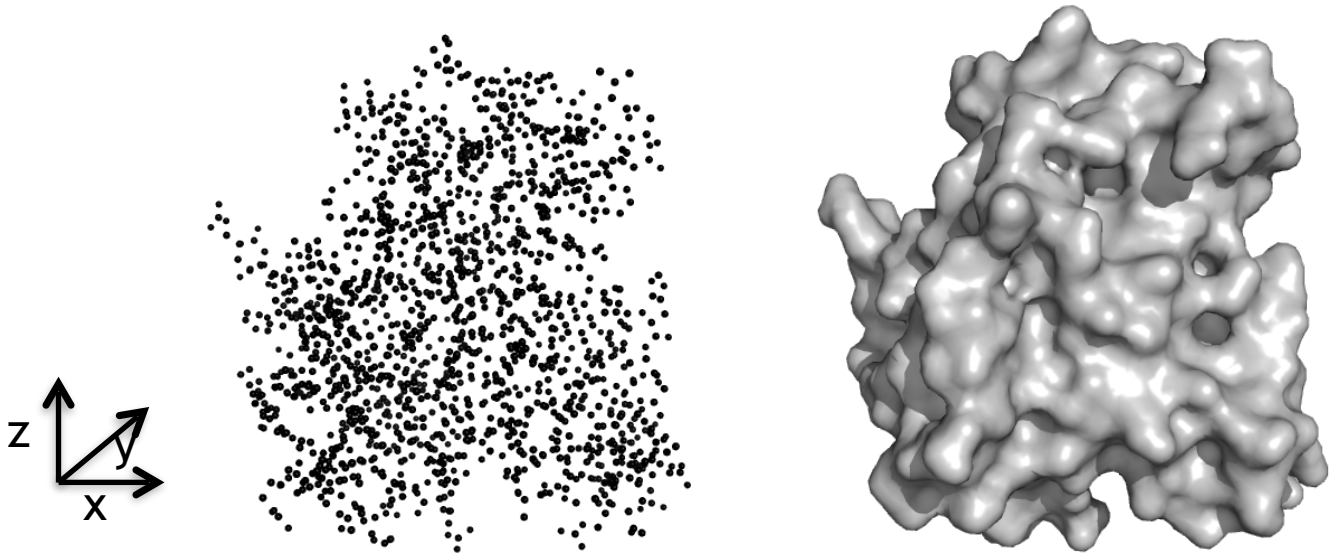


What is a protein?

~~1-Dimensional~~ text

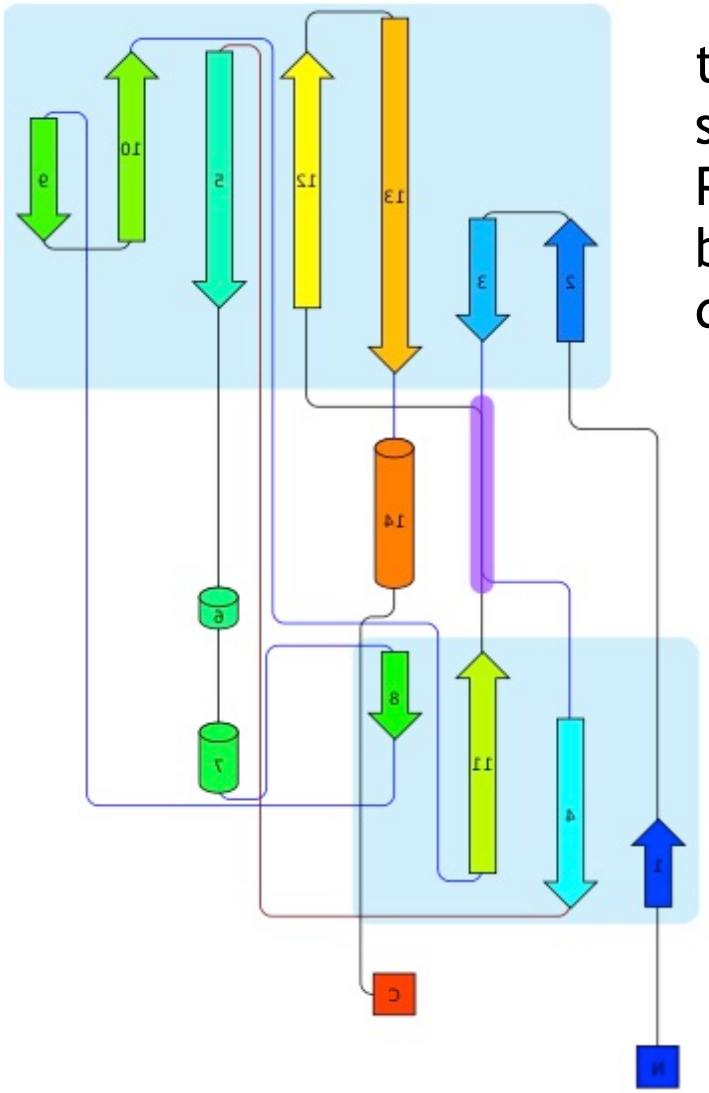
WPISSSVPSQKTYQGSYGFRLGFLH

3-Dimensional set of points/shape

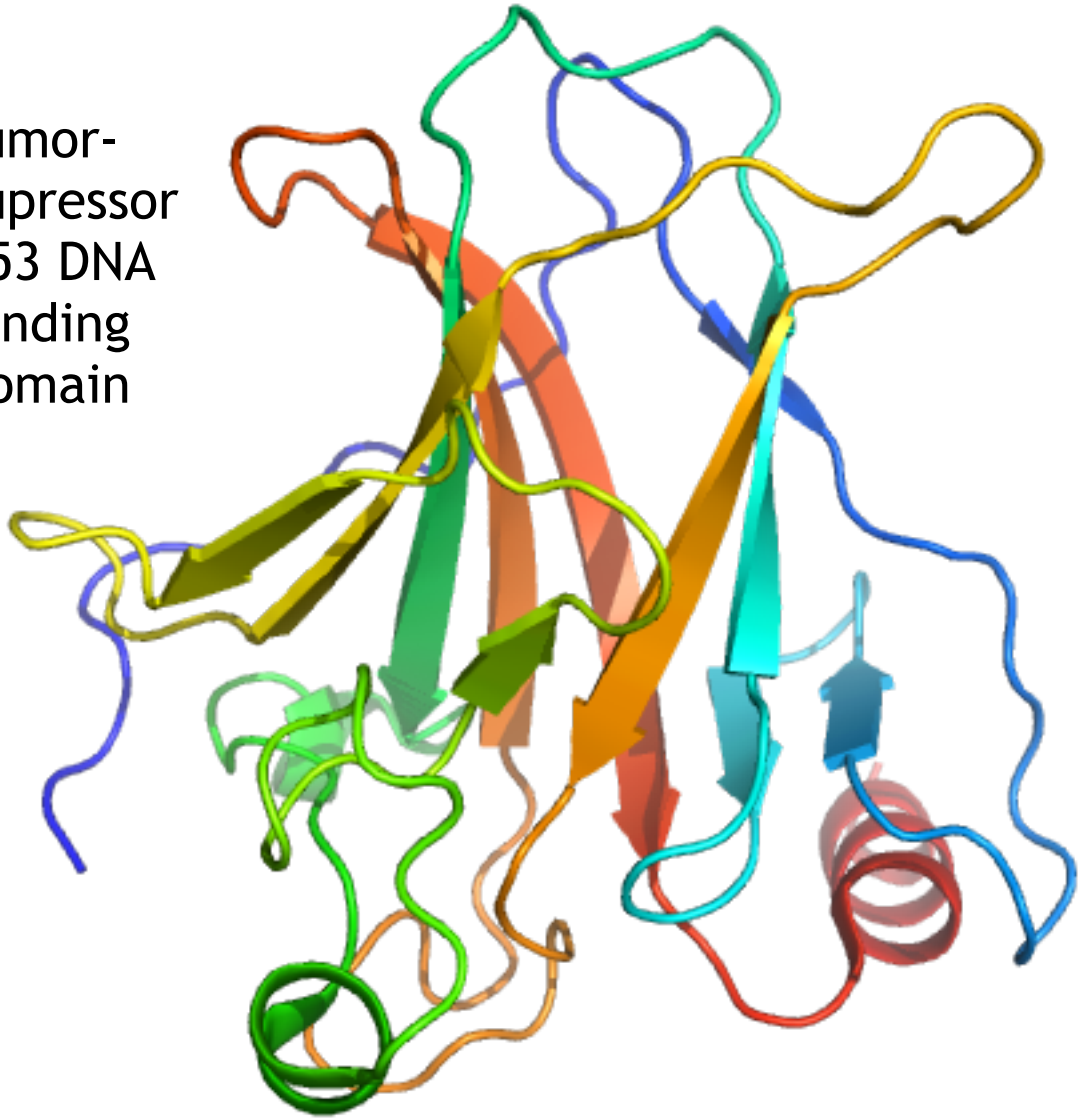


Proteins are objects that exist in the physical world.

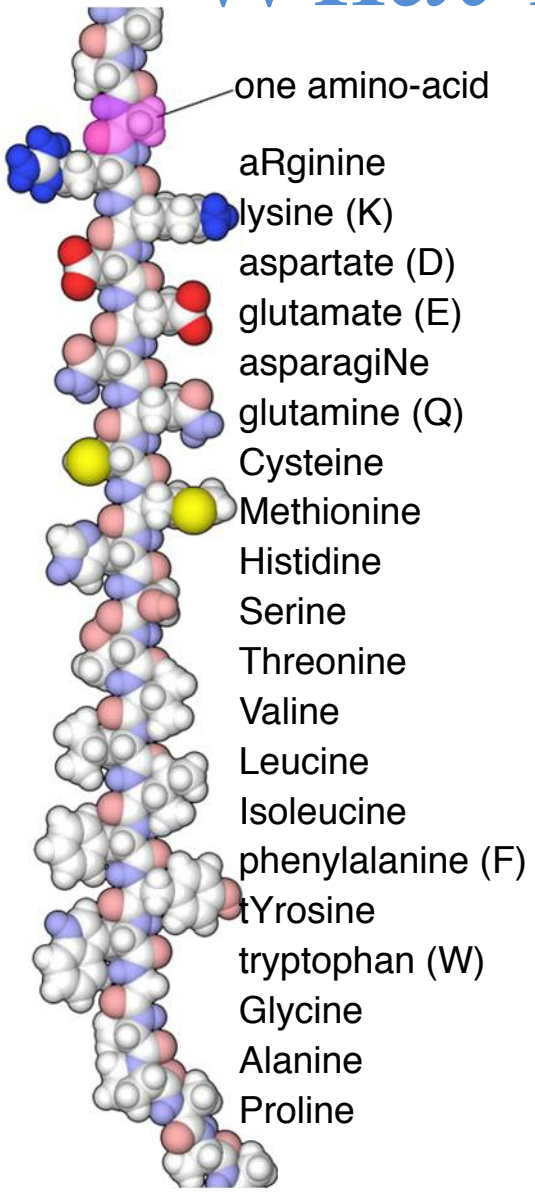
What is a protein?



tumor-supressor P53 DNA binding domain



What is a protein made of?

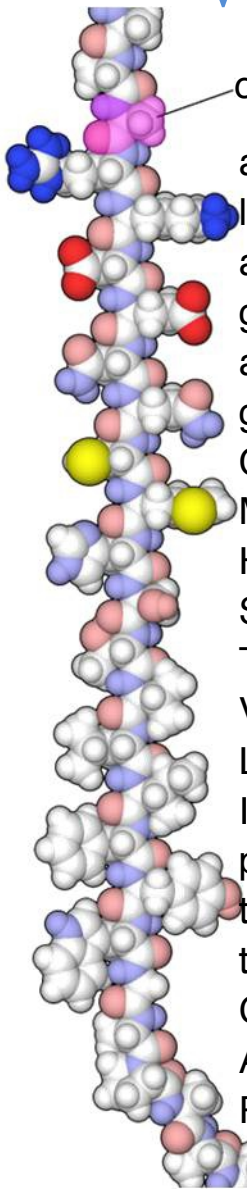


one amino-acid

aRginine
lysine (K)
aspartate (D)
glutamate (E)
asparagiNe
glutamine (Q)
Cysteine
Methionine
Histidine
Serine
Threonine
Valine
Leucine
Isoleucine
phenylalanine (F)
tYrosine
tryptophan (W)
Glycine
Alanine
Proline

20 amino acids

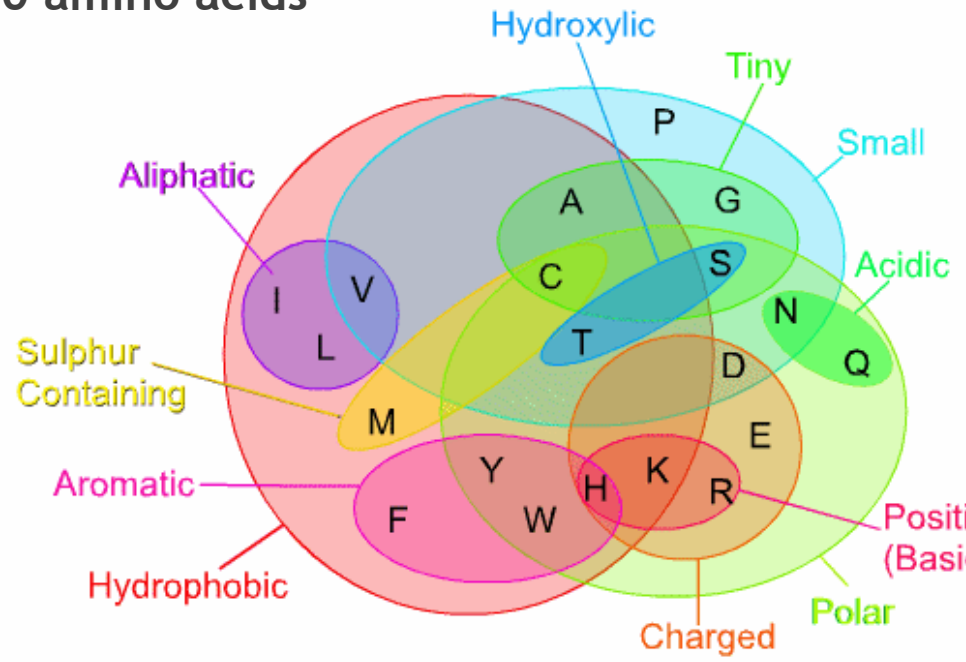
What is a protein made of?



one amino-acid

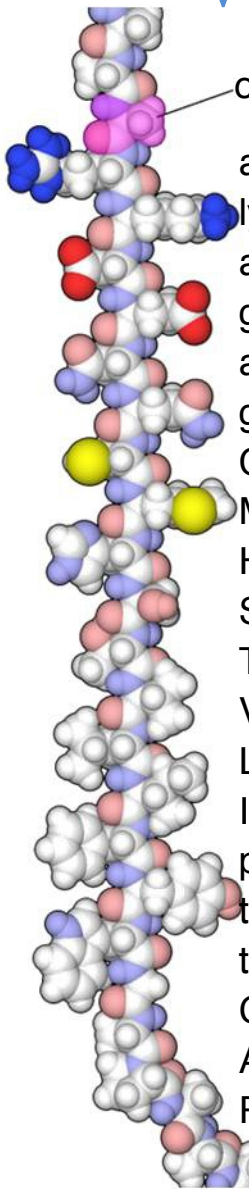
aRginine
lysine (K)
aspartate (D)
glutamate (E)
asparagiNe
glutamine (Q)
Cysteine
Methionine
Histidine
Serine
Threonine
Valine
Leucine
Isoleucine
phenylalanine (F)
tYrosine
tryptophan (W)
Glycine
Alanine
Proline

20 amino acids



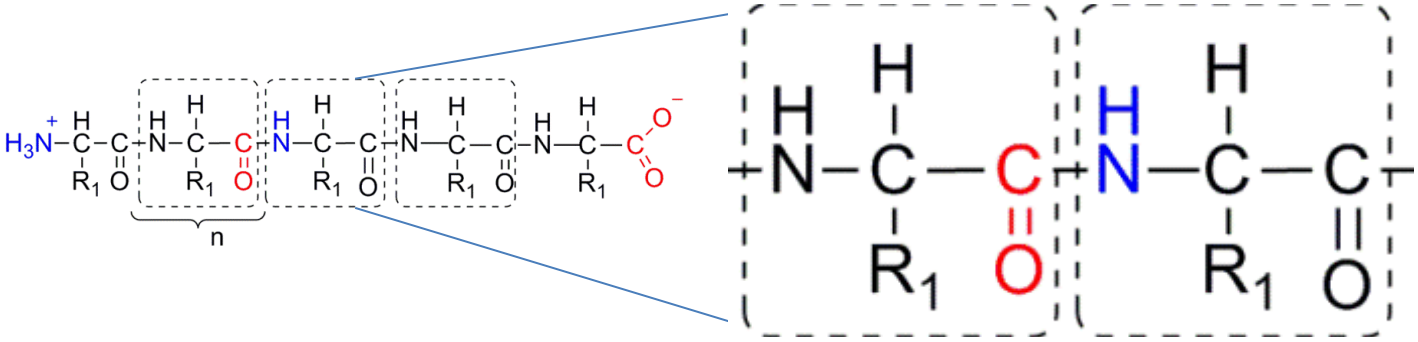
This is 1 possible classification, among others!

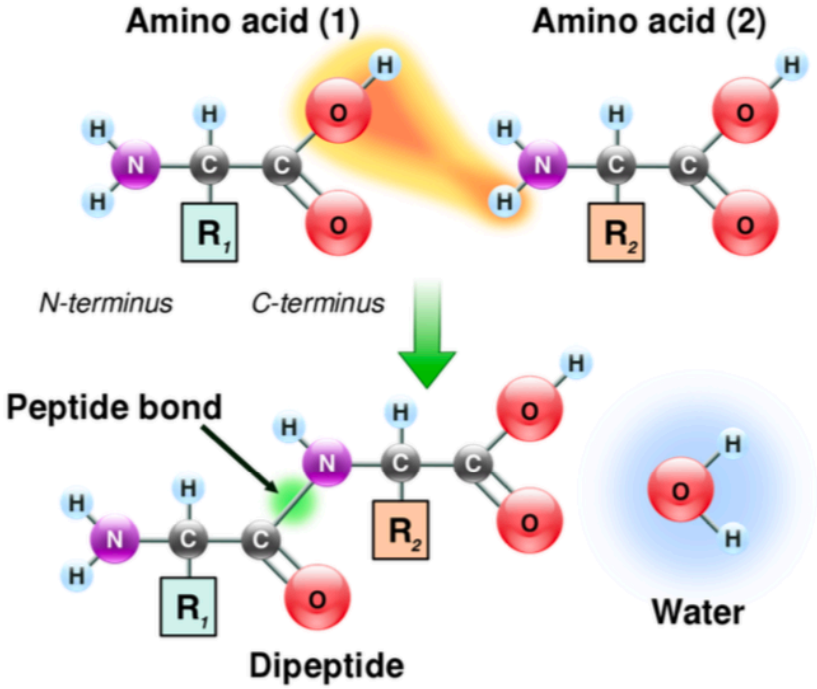
What is a protein made of?



one amino-acid

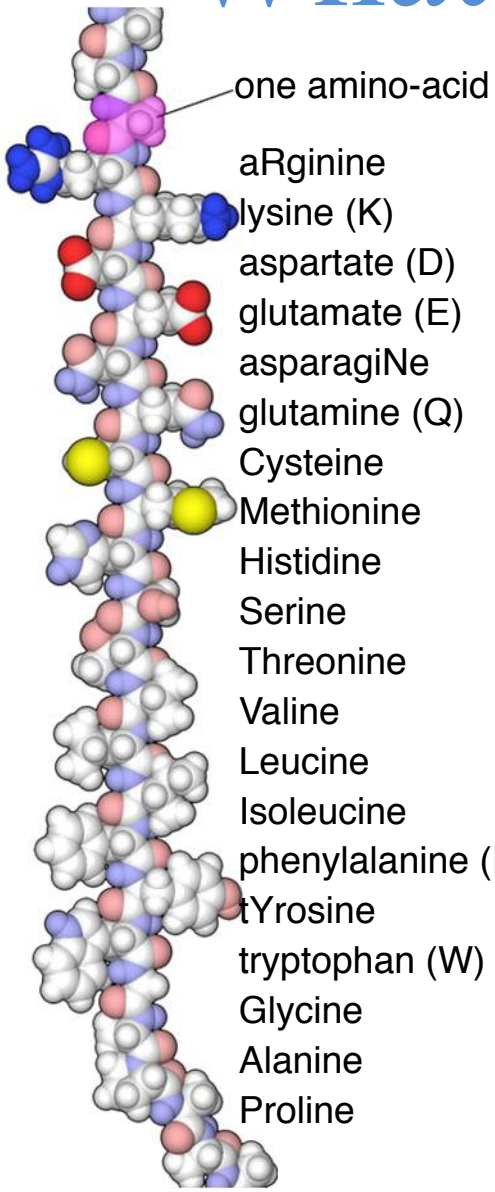
aRginine
lysine (K)
aspartate (D)
glutamate (E)
asparagiNe
glutamine (Q)
Cysteine
Methionine
Histidine
Serine
Threonine
Valine
Leucine
Isoleucine
phenylalanine (F)
tYrosine
tryptophan (W)
Glycine
Alanine
Proline





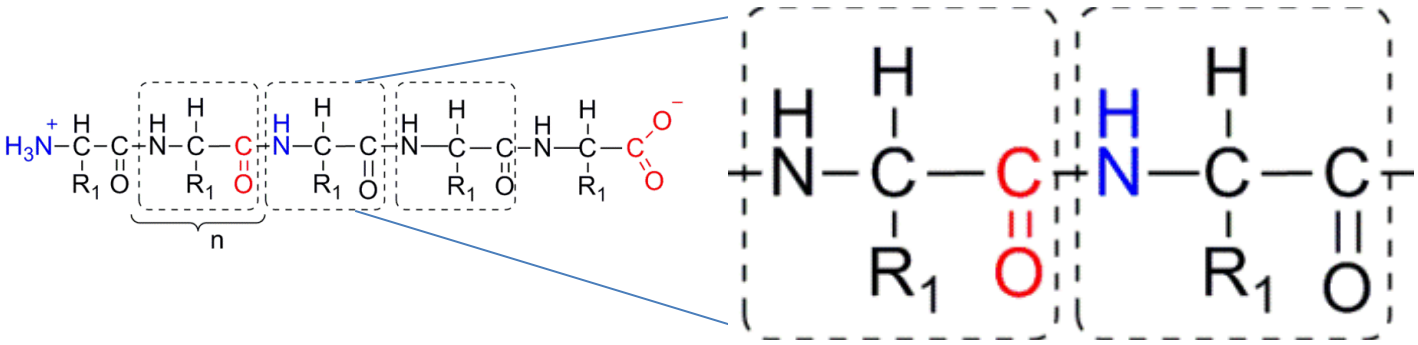
Peptidic bond
covalent
strong

What is a protein made of?



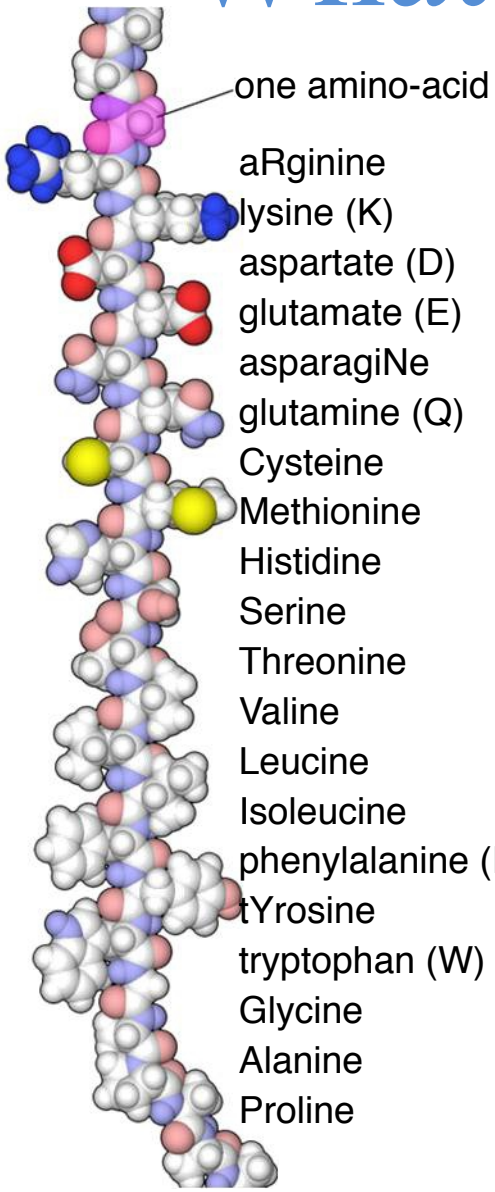
one amino-acid

aRginine
lysine (K)
aspartate (D)
glutamate (E)
asparagiNe
glutamine (Q)
Cysteine
Methionine
Histidine
Serine
Threonine
Valine
Leucine
Isoleucine
phenylalanine (F)
tYrosine
tryptophan (W)
Glycine
Alanine
Proline



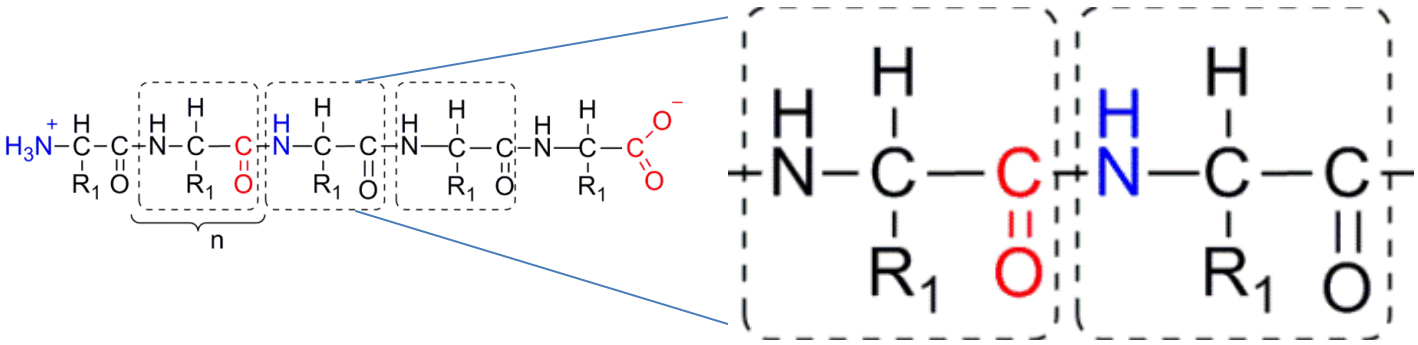
The amino acids are <u>chiral</u> molecules

What is a protein made of?

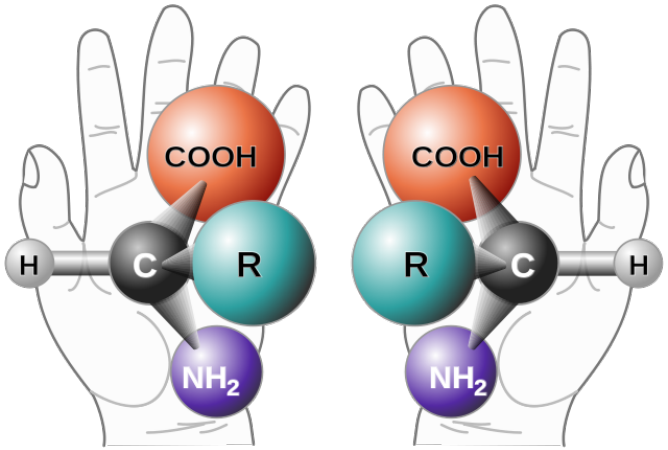


one amino-acid

aRginine
lysine (K)
aspartate (D)
glutamate (E)
asparagiNe
glutamine (Q)
Cysteine
Methionine
Histidine
Serine
Threonine
Valine
Leucine
Isoleucine
phenylalanine (F)
tYrosine
tryptophan (W)
Glycine
Alanine
Proline



The amino acids are <u>chiral</u> molecules



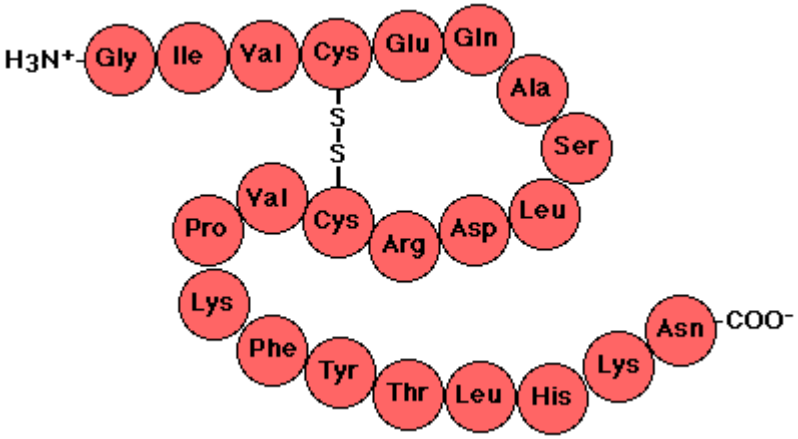
The translation machinery of protein has evolved to utilize only one of the chiral forms of amino acids : **the L-form**

How are proteins organised?

1st level of organisation : primary structure

…QNCQLRPSGWQCRPTRGDCDLPEFCPGDSSQCPDVSLGDG…

~10's to ~1000's of amino acid residues

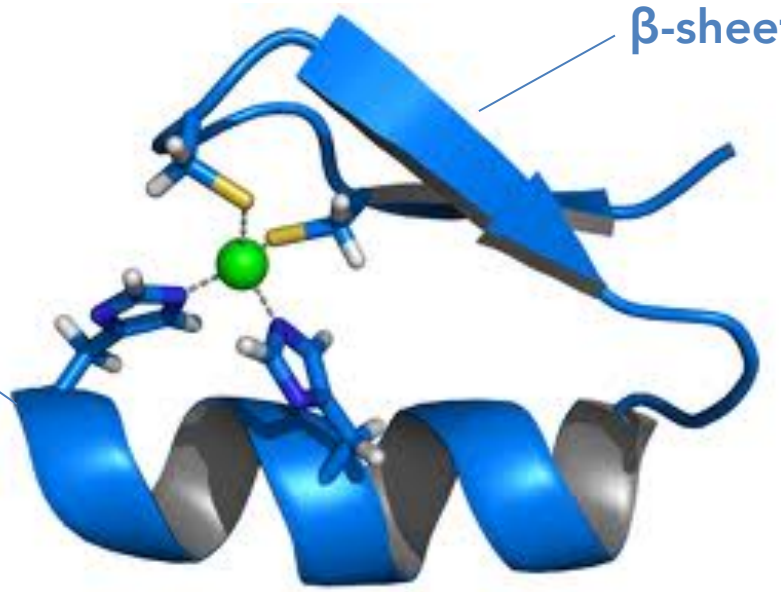


Covalent bonds

1 protein = 1 polypeptidic chain

How are proteins organised?

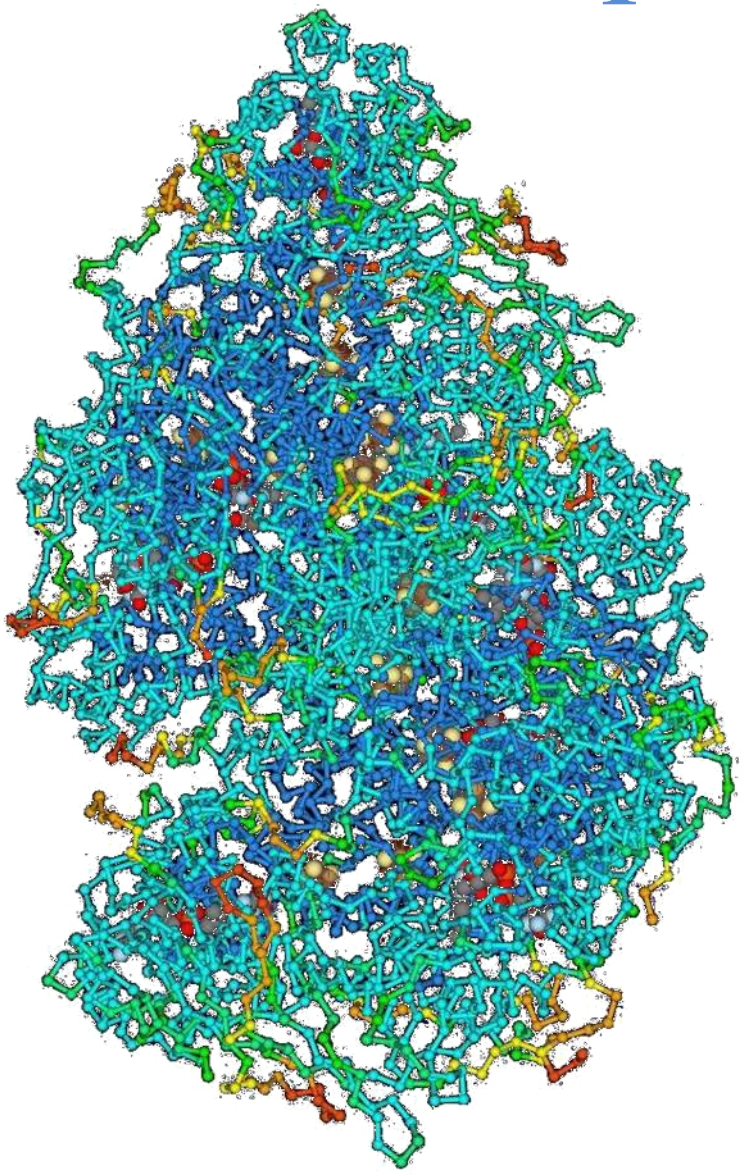
2nd level of organisation : secondary structure



Backbone-backbone weak chemical bonds

Other elements: 3_{10} helix > turns > loops > random coil

How are proteins organised?

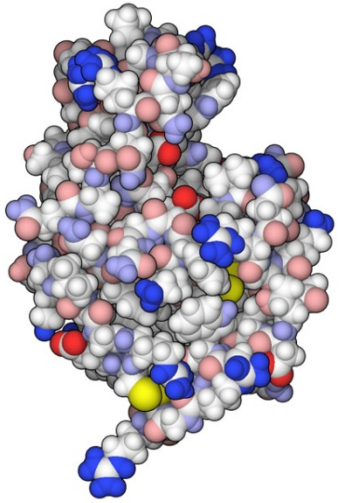


3rd level of organisation : tertiary structure

A protein sequence adopts a particular fold in solution, which corresponds to a free energy minimum

Types of non-covalent interactions:

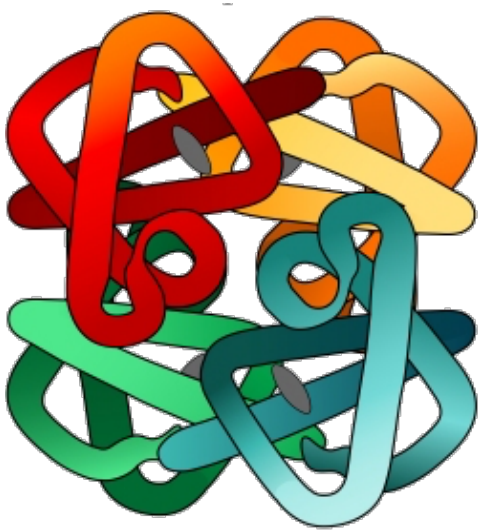
- *salt bridges*
- *hydrogen bonds*
- *hydrophobic contact*
- *pi-pi stacking…*



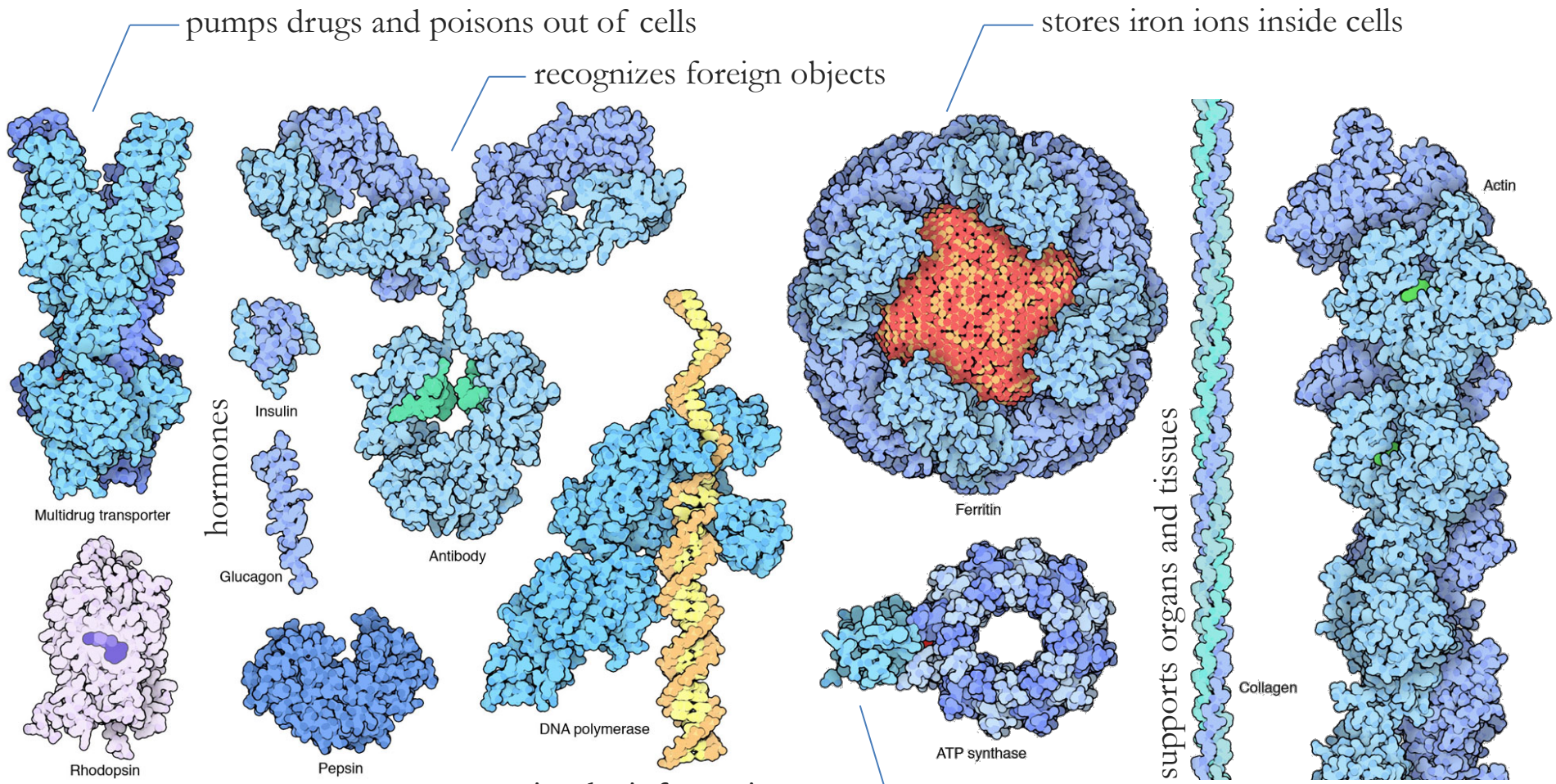
How are proteins organised?

4th level of organisation : quaternary structure

Arrangements of domains within a protein or of proteins within a macro-molecular assembly



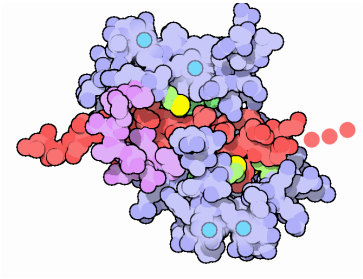
What do proteins do?



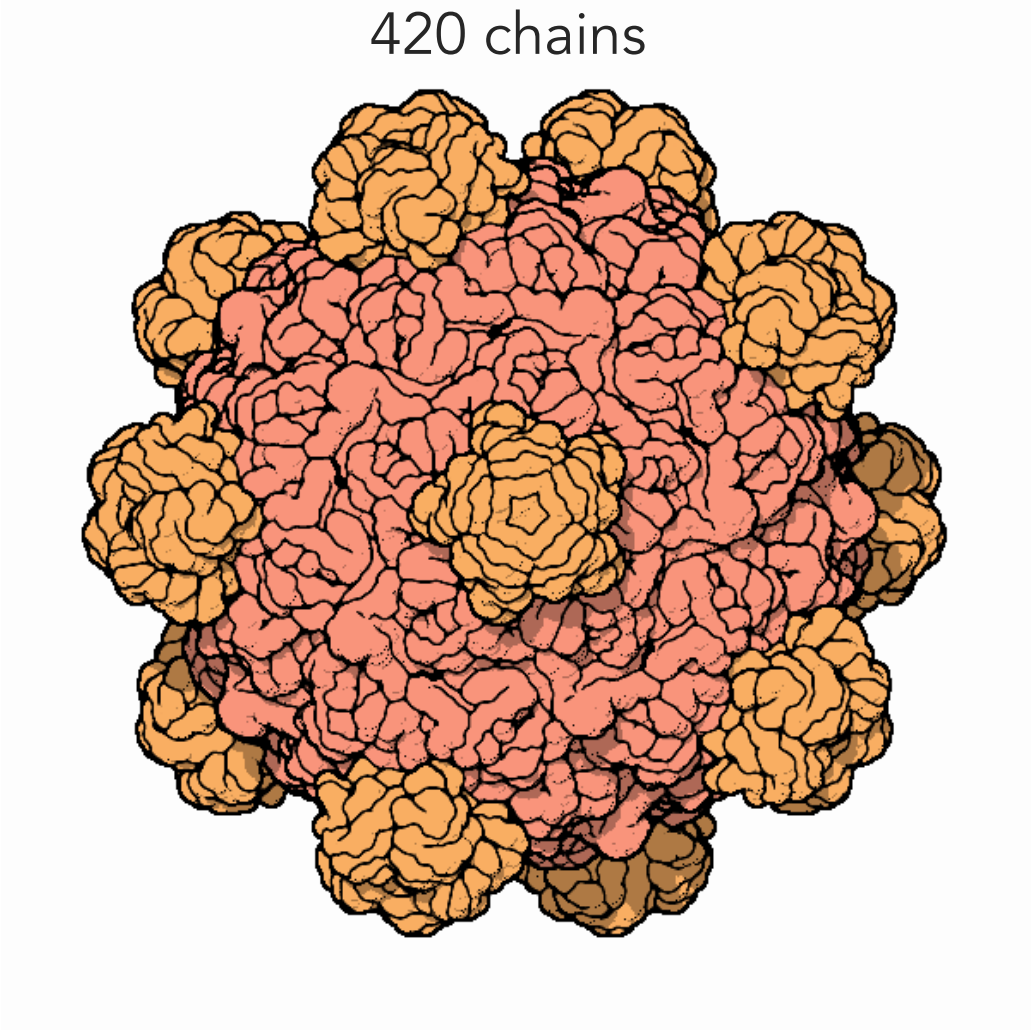
D. Goodsell

Proteins do not act alone

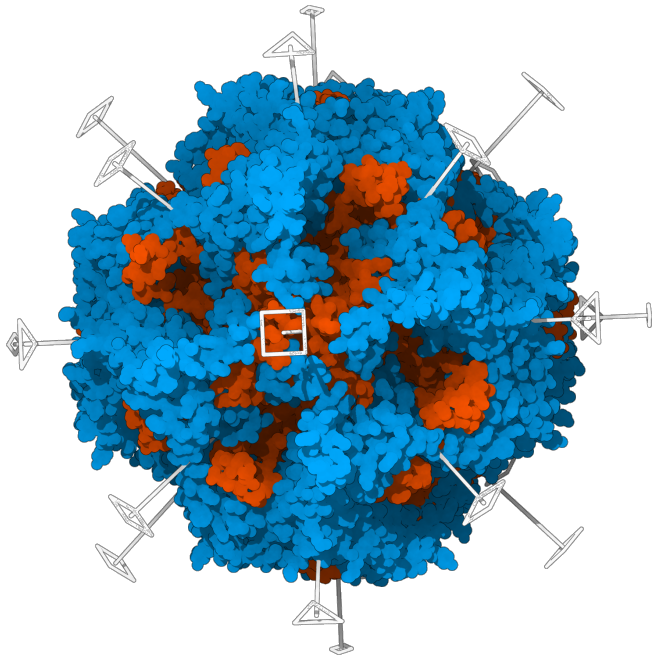
Binary complex



Bacteriophage
420 chains

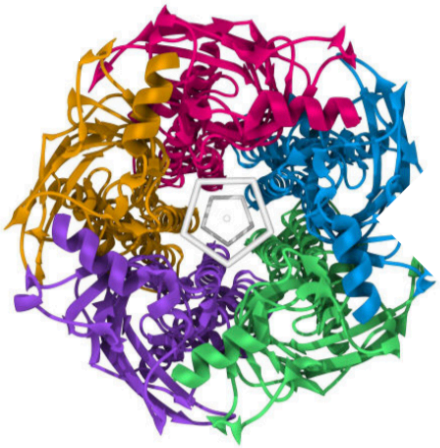


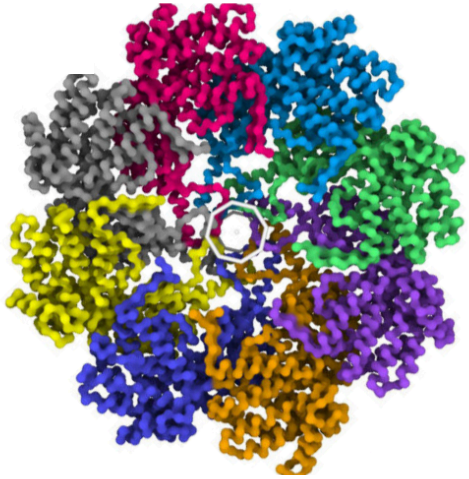
Proteins do not act alone



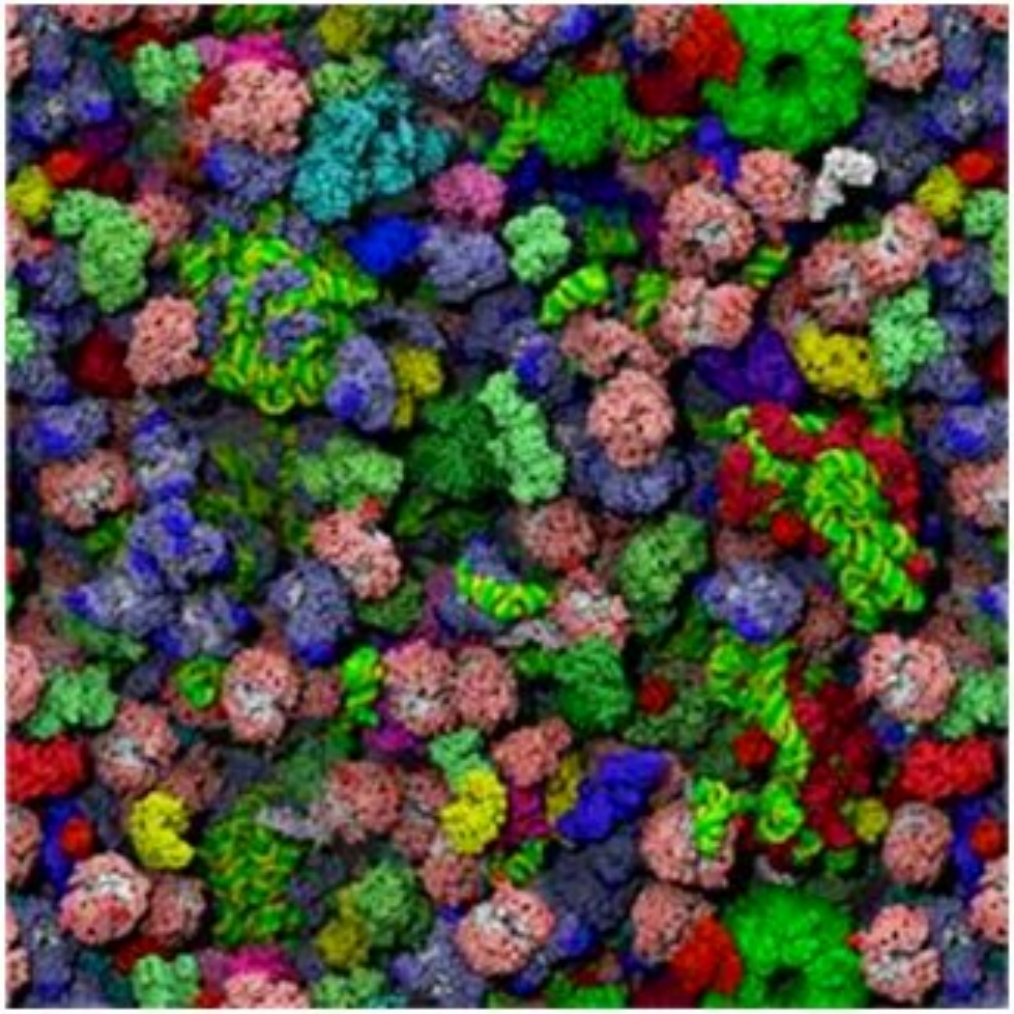
Symmetrical protein complexes are very common in nature. Molecular symmetries are important for evolution, folding, stability and function.

Ananas: automated tool to determine symmetry order and axes
https://team.inria.fr/nano-d/software/ananas/





Proteins do not act alone

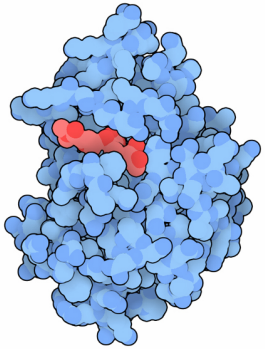


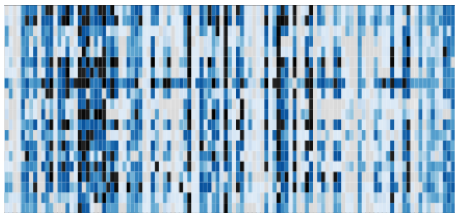
The cell is a very crowded environment!

Simulation of the cytoplasm of *E. coli McGuffee 2010* doi:https://doi.org/10.1371/journal.pcbi.1000694

Sequence → Structure → Function

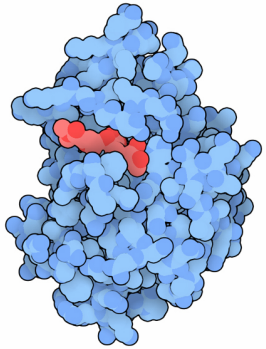
...TLRKLLTGELLTL
ASRQQLIDWMEADK
VGGPLLRSALPAGW
FIADKSGAGERGSR
GI...

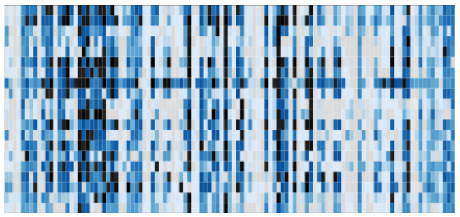




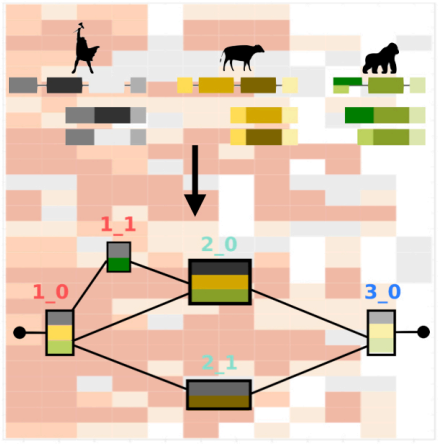
Sequences! → Structures! → Functions!

...TLRKLLTGELLTL
ASRQQLIDWMEADK
VGGPLLRSALPAGW
FIADKSGAGERGSR
GI...

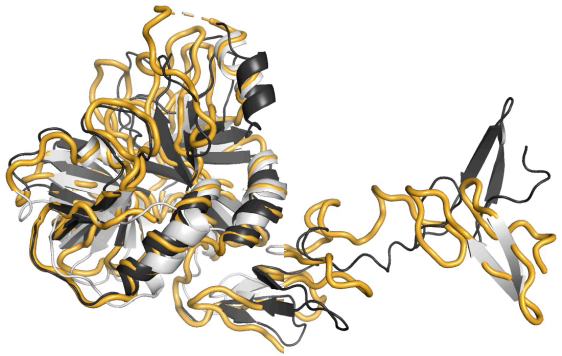




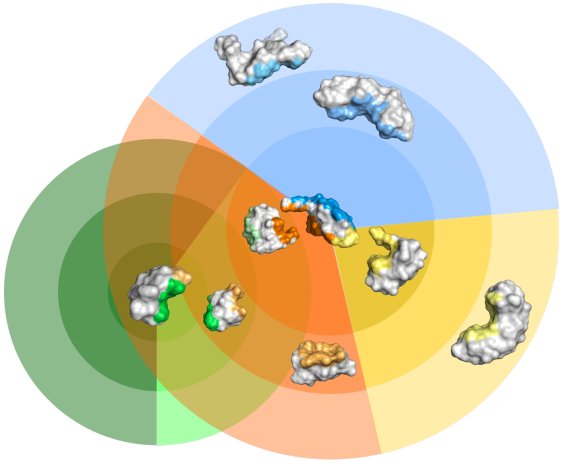
Splice variants



Dynamics

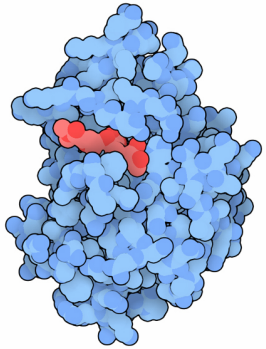


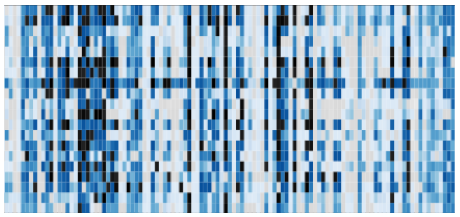
Interactions



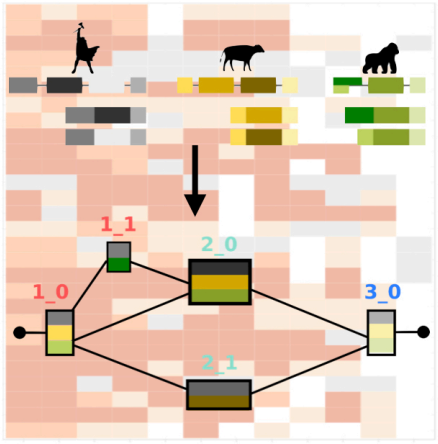
Sequences! → Structures! → Functions!

...TLRKLLTGELLTL
ASRQQLIDWMEADK
VGGPLLRSALPAGW
FIADKSGAGERGSR
GI...

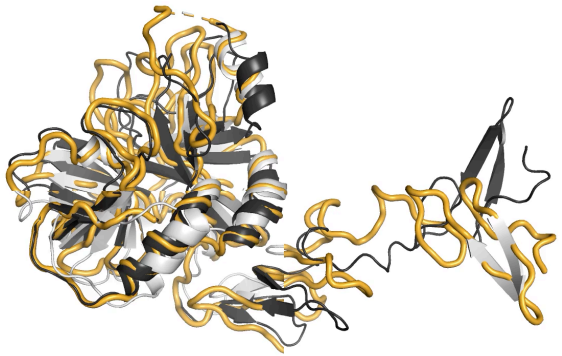




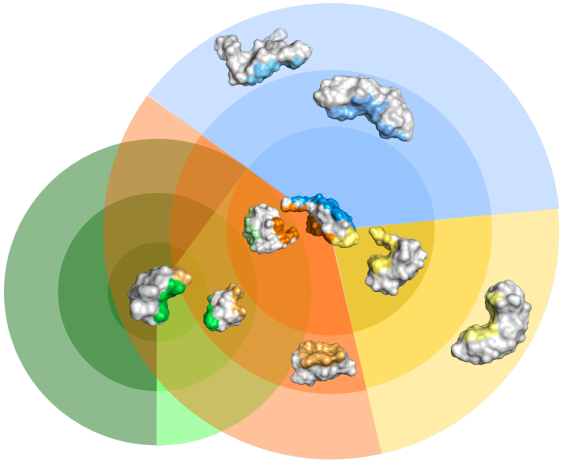
Splice variants



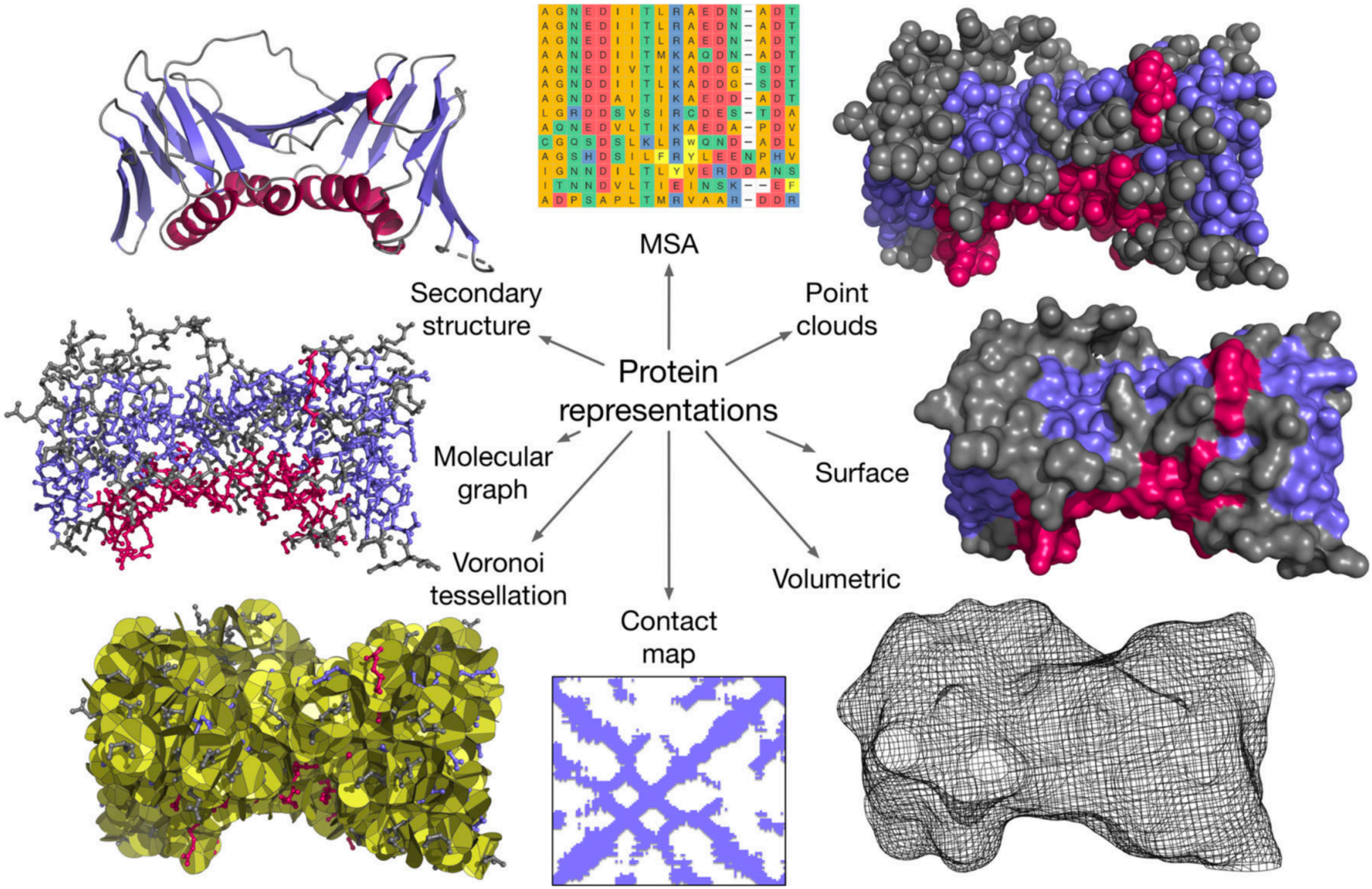
Dynamics



Interactions

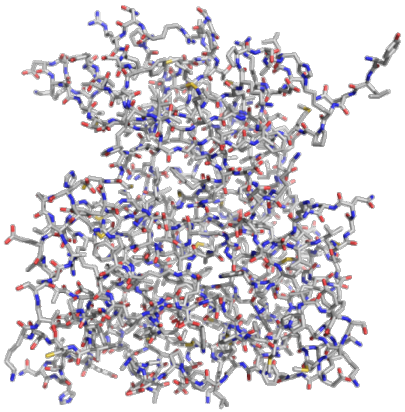


How do we represent proteins?



Laine, Eismann, Elofsson & Grudinin, 2021

How do we represent protein structures?



protein kinase

~ 300 amino acid residues

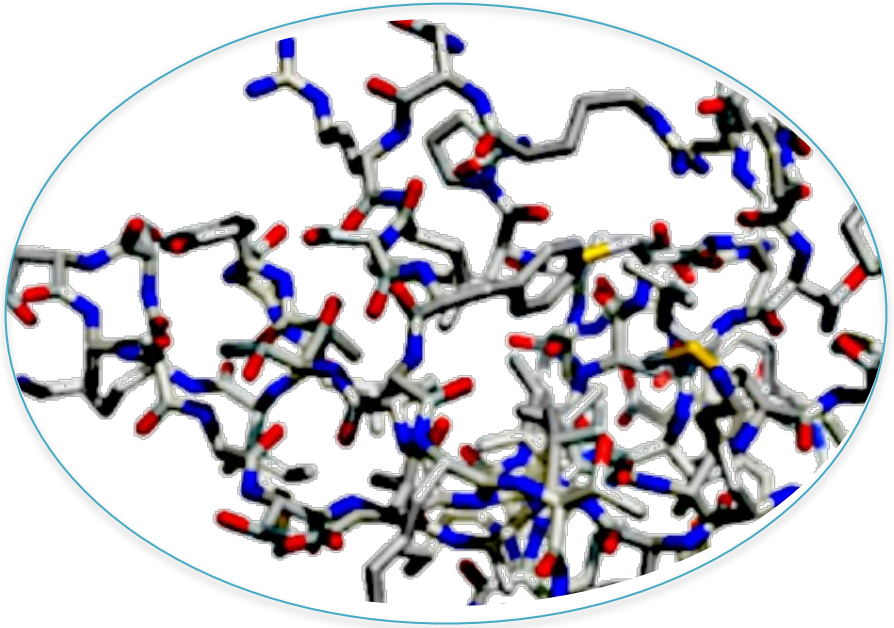
~ 5000 atoms (15 000 dof)

Each atom is colored according to its element.

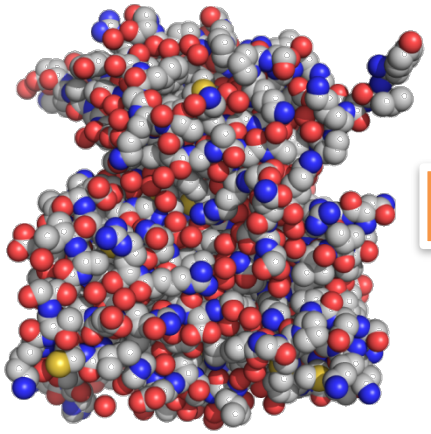
N: blue, O: red, C: grey

The sticks represent the covalent bonds (~1.5 Angstroms) between atoms.

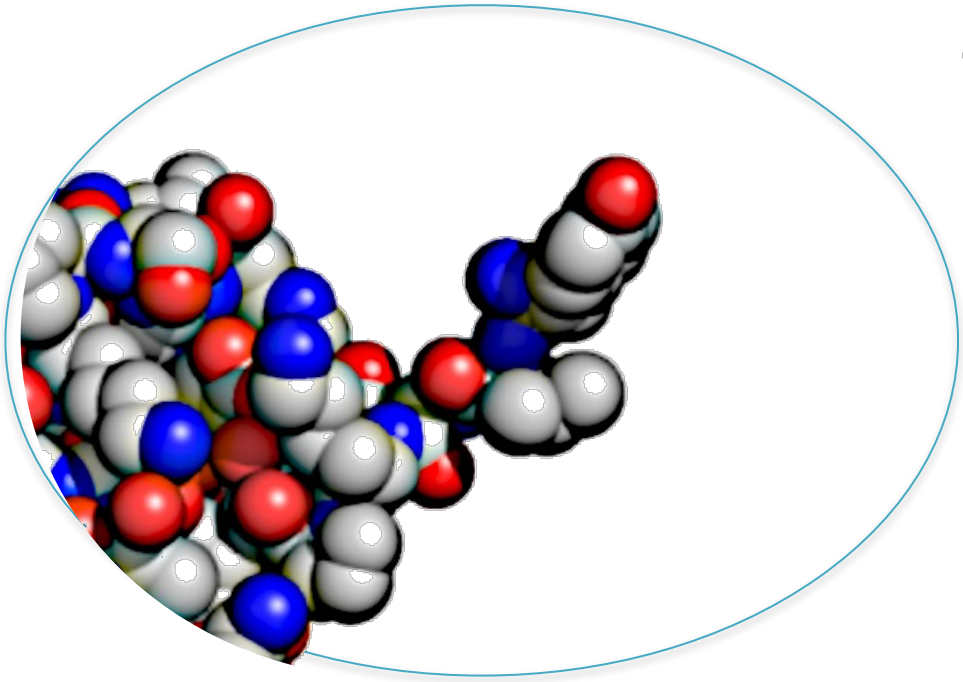
The atoms are at the extremities or intersection of the sticks.



How do we represent protein structures?



spheres



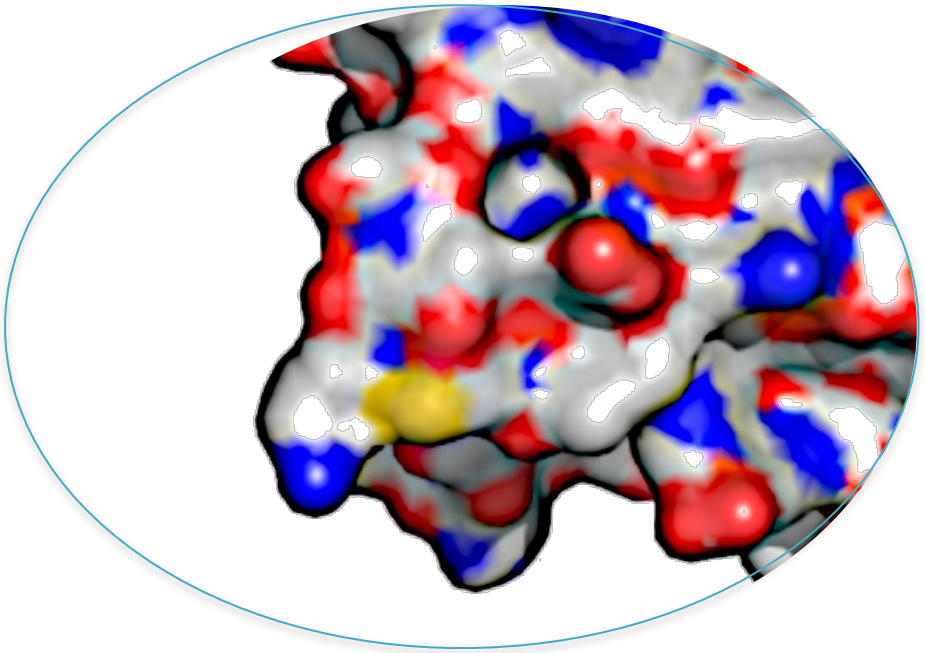
The sphere represents the volume taken by the atom.

The radius of the sphere depends on the type of atom. It is called the van der Waals radius.

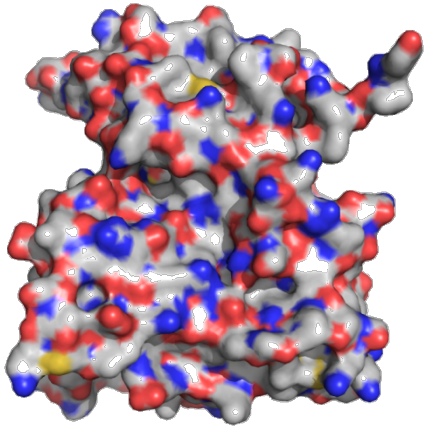
How do we represent protein structures?

The molecular surface delimits the volume that is not penetrated by water molecules in solution.

It is obtained by rolling a probe (sphere of 1.4 Angstroms) on the protein.

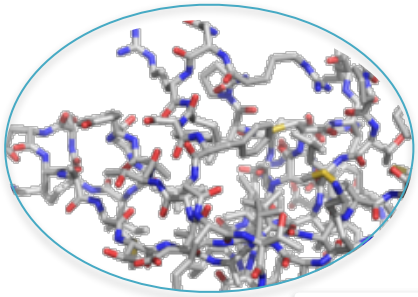


surface

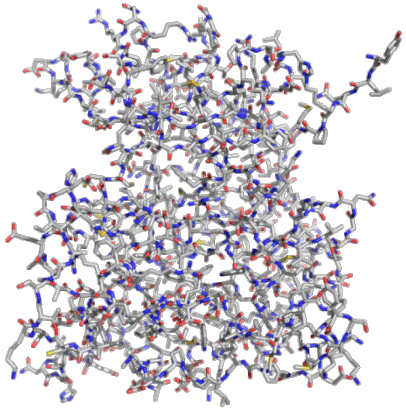


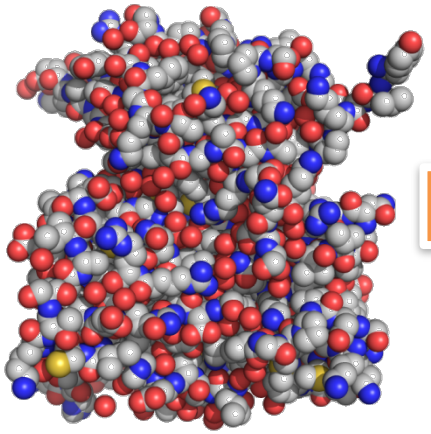
1 Angstrom = 10^{-10} m = 0.1 nm

How do we represent protein structures?

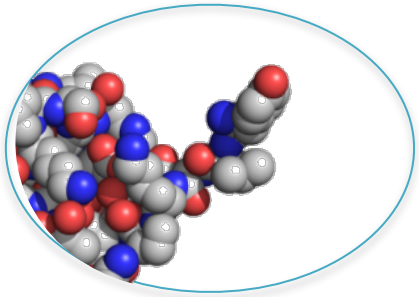


sticks

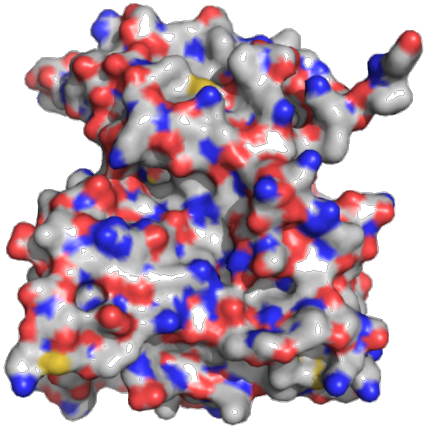


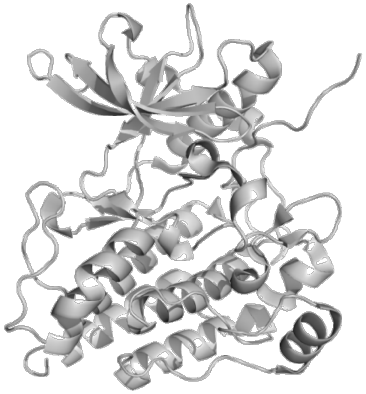


spheres

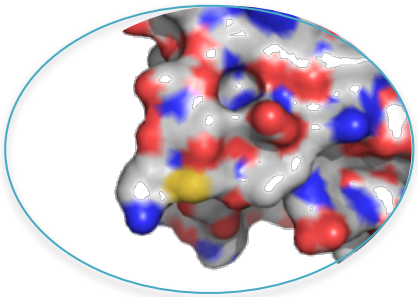


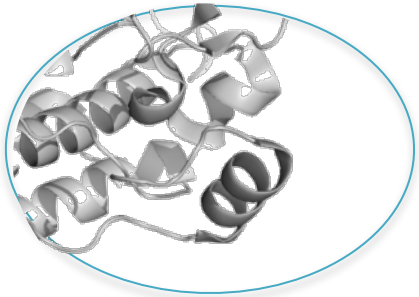
surface



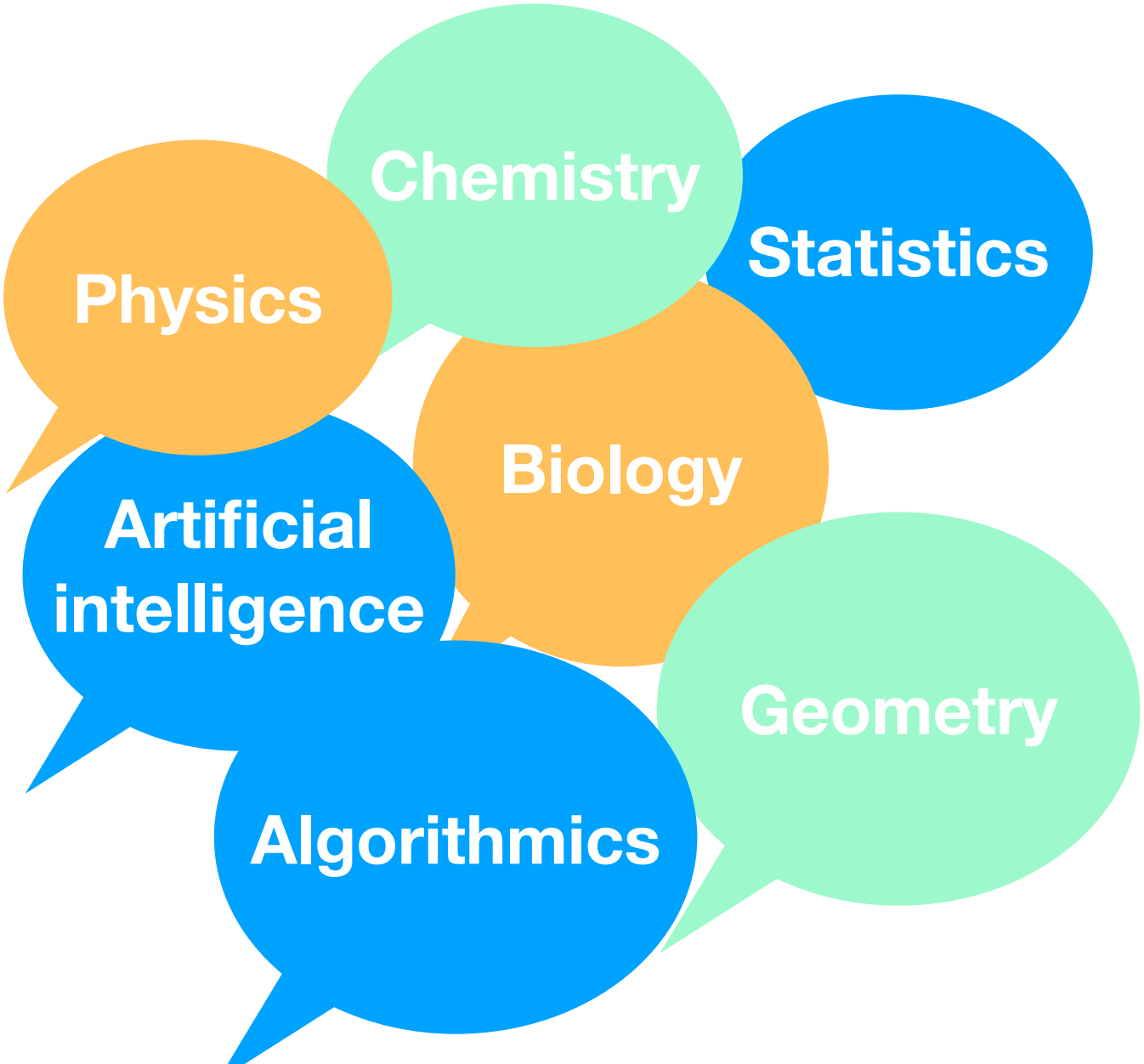


cartoon





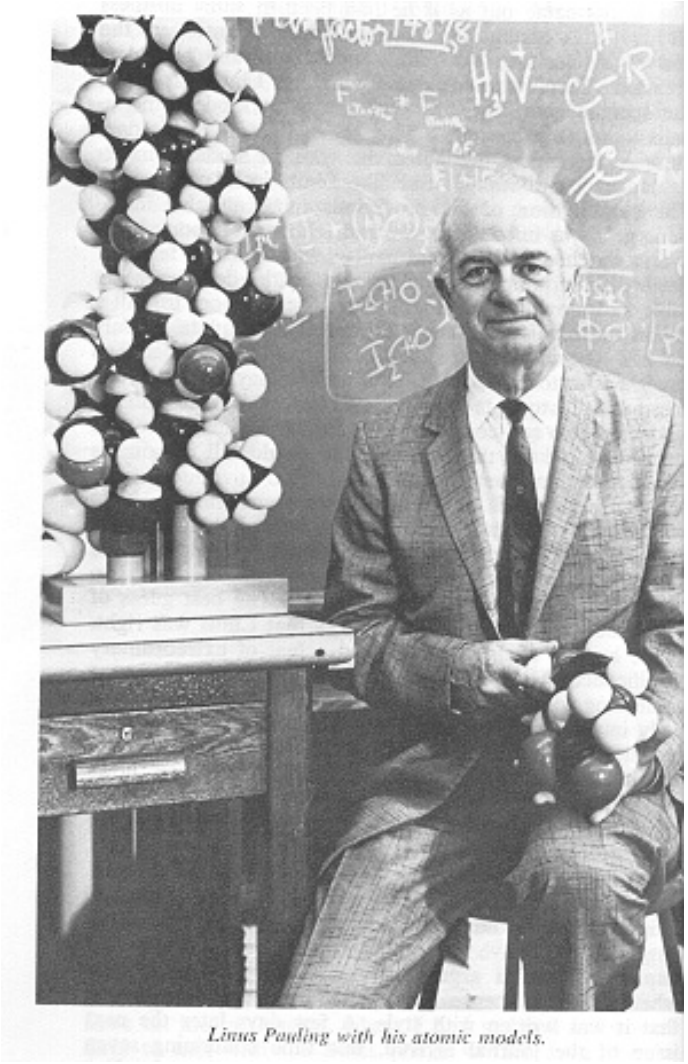
Protein structures are about…



PROTEIN STRUCTURE DETERMINATION

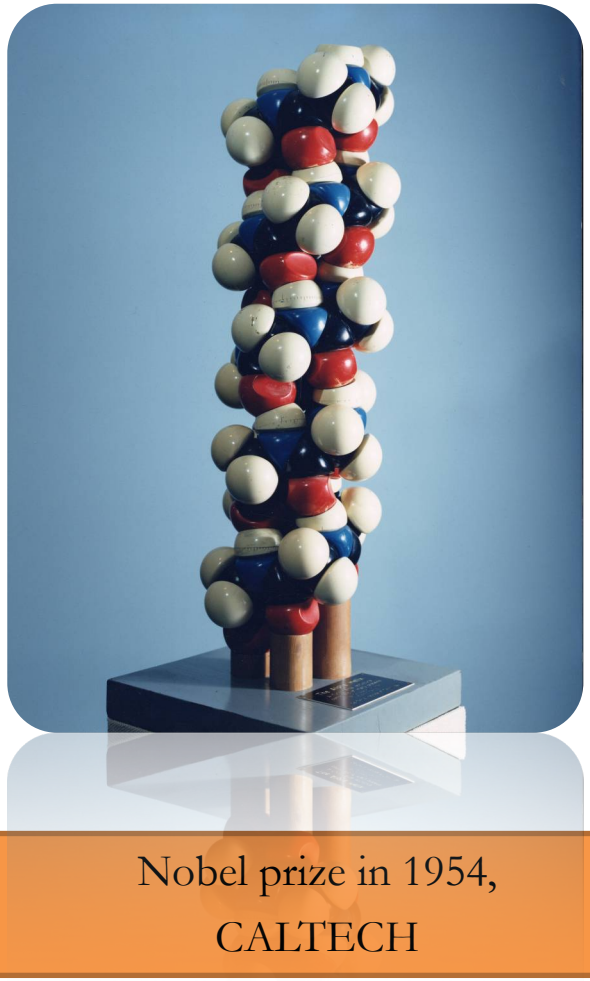
A brief history

Linus Pauling (1901-1994)



Linus Pauling with his atomic models.

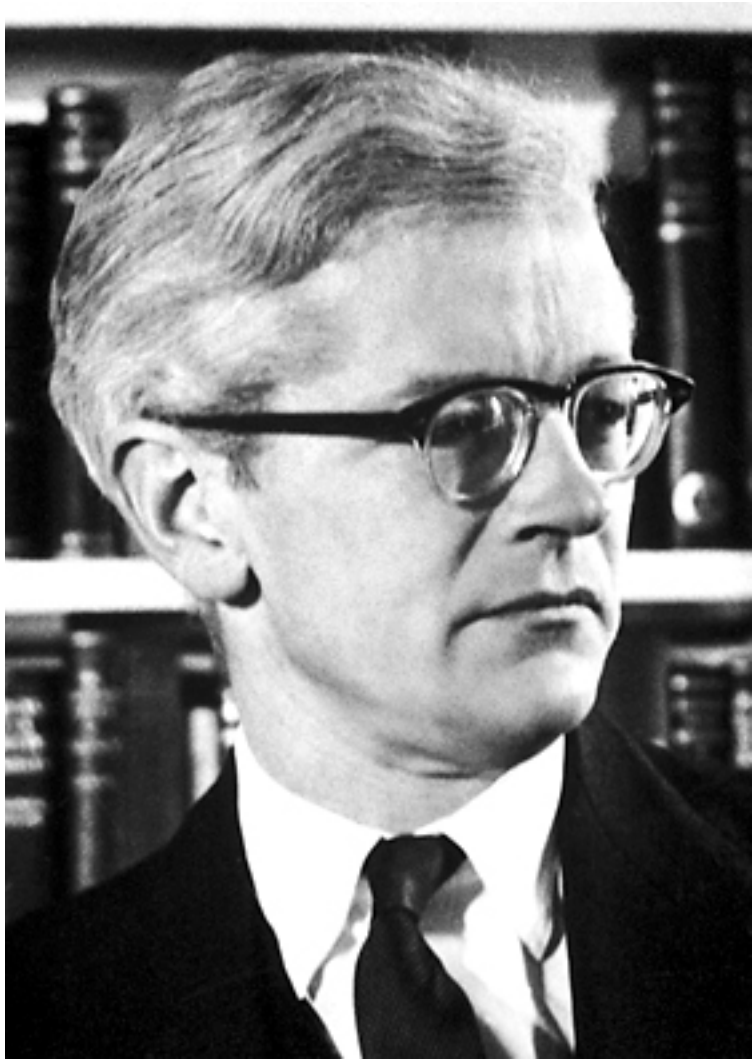
Pauling & Corey (1951) *PNAS*



Nobel prize in 1954,
CALTECH

A brief history

John Kendrew (1917-1997)



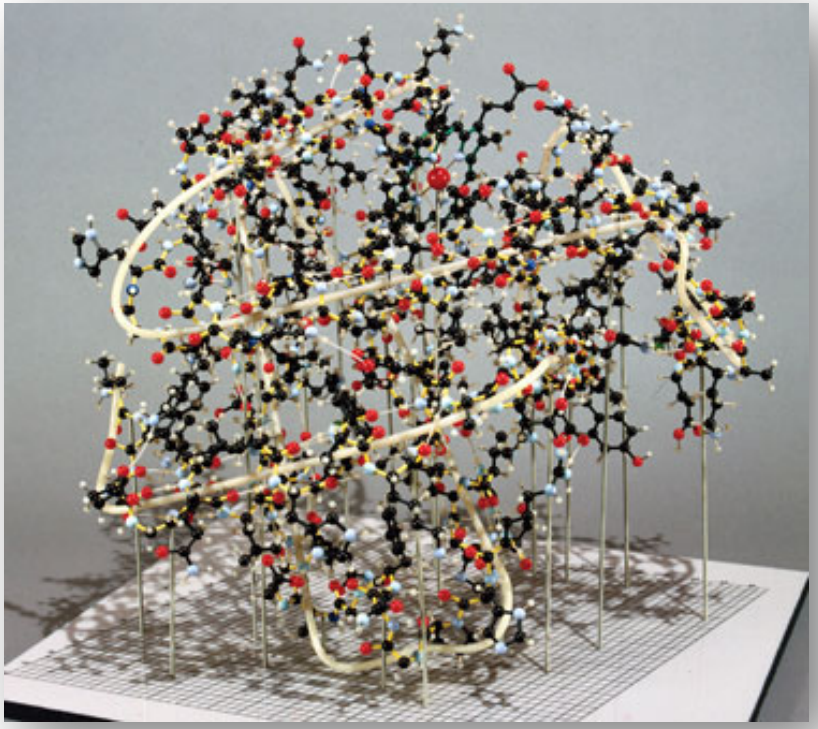
Kendrew *et al.* (1958) *Nature*

Nobel prize in 1962 with Max Perutz, Cavendish Laboratory

Protein structures are puzzling

3-Dimensional structure of myoglobin (1958, 2 Å resolution)

Kendrew *et al.* (1958) *Nature*



Perhaps the most remarkable features of the molecule are its **complexity** and its **lack of symmetry**. The arrangement seems to be almost totally lacking in the kind of regularities which one instinctively anticipates, and it is **more complicated than has been predicted** by any theory of protein structure.

Experimental methods

✓ **X-ray crystallography**

➢ average structure

➢ only soluble proteins

➢ (very) high resolution

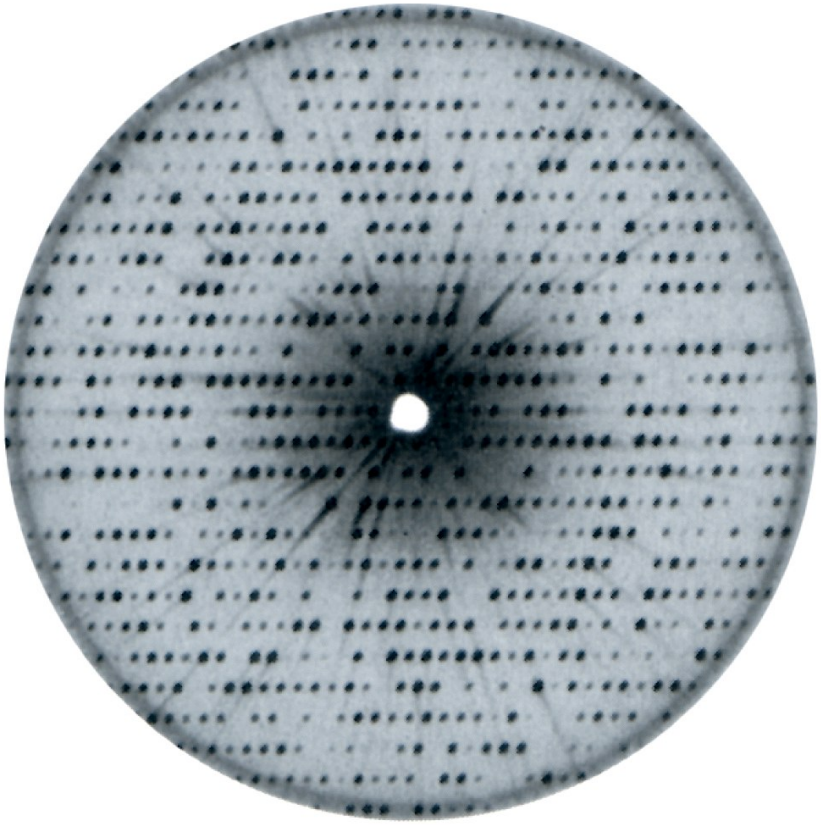


Figure 3-44
Biochemistry, Sixth Edition

diffraction pattern

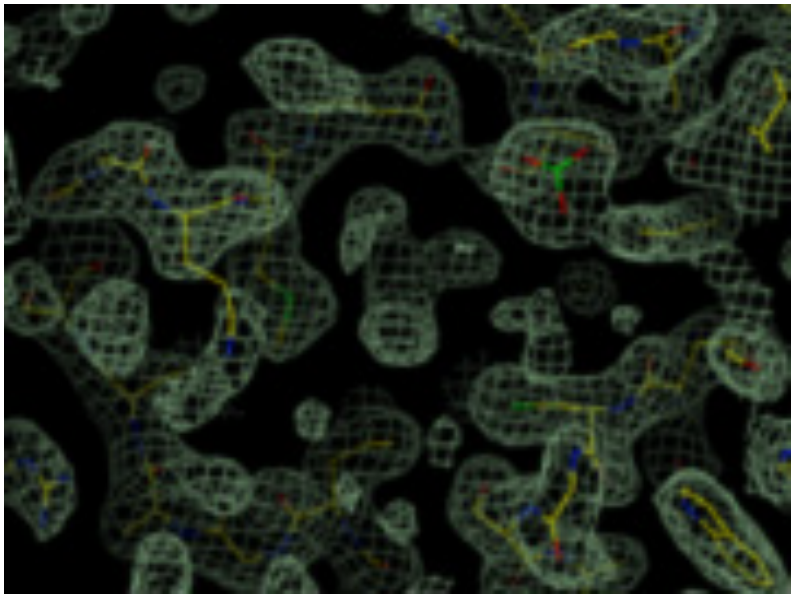
Experimental methods

✓ **X-ray crystallography**

➢ average structure

➢ only soluble proteins

➢ (very) high resolution



electronic density map

Experimental methods

- ✓ X-ray crystallography
- ✓ **NMR spectroscopy**

- ➢ liquid or solid state
- ➢ proteins of limited size
- ➢ Structural ensemble

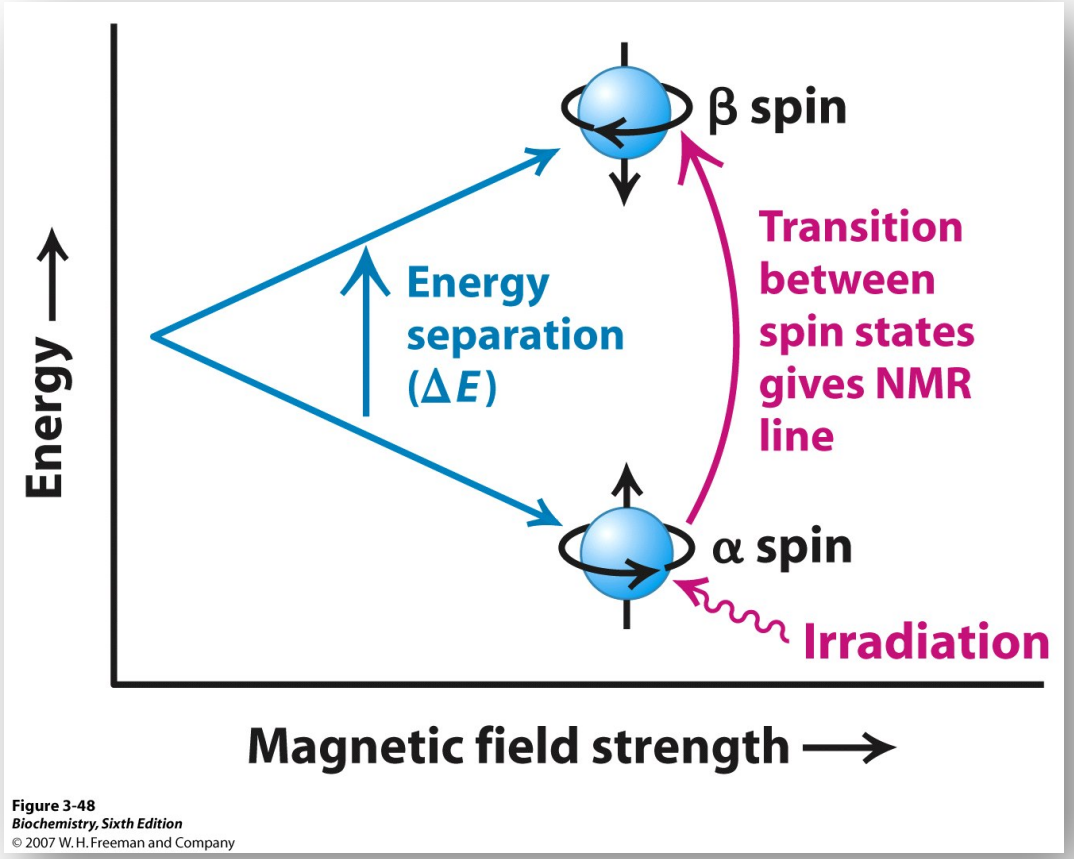
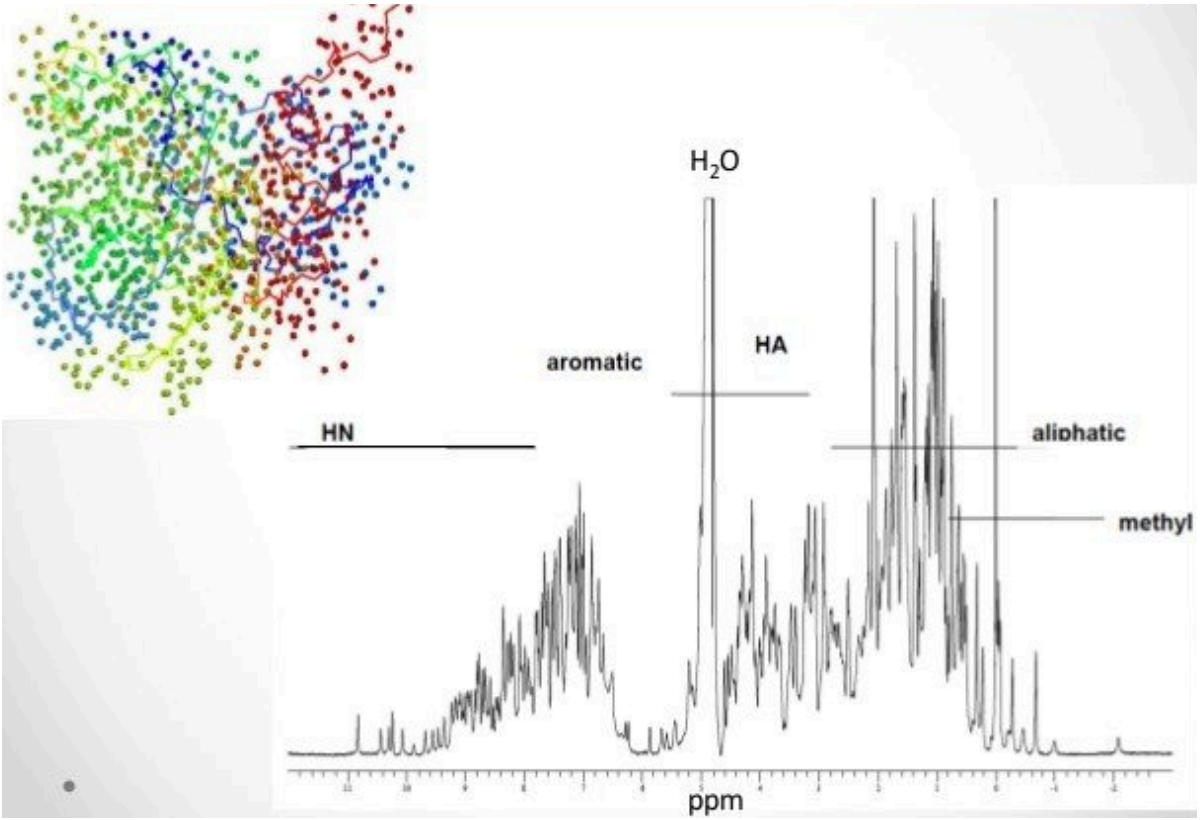


Figure 3-48
Biochemistry, Sixth Edition

Experimental methods

- ✓ X-ray crystallography
- ✓ **NMR spectroscopy**

- ➢ liquid or solid state
- ➢ proteins of limited size
- ➢ Structural ensemble

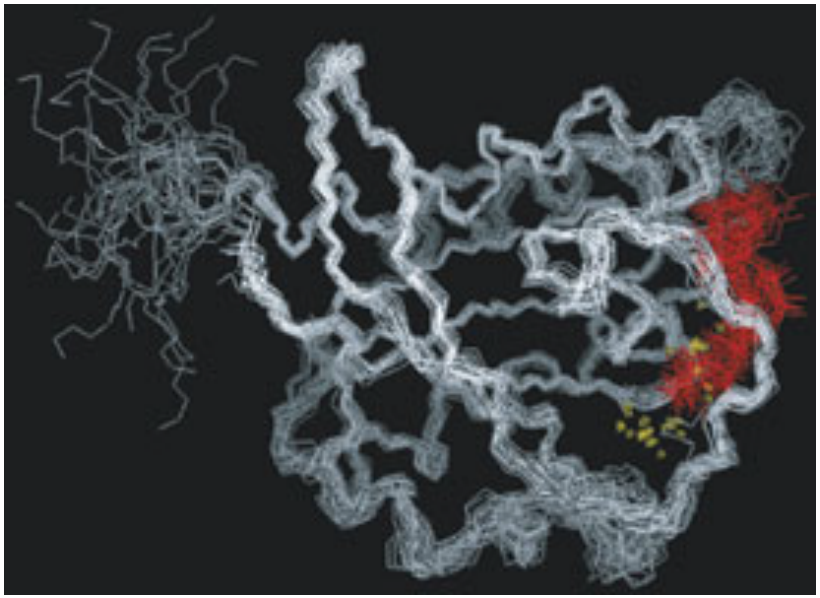


NMR spectrum

Experimental methods

- ✓ X-ray crystallography
- ✓ **NMR spectroscopy**

- ➢ liquid or solid state
- ➢ proteins of limited size
- ➢ Structural ensemble



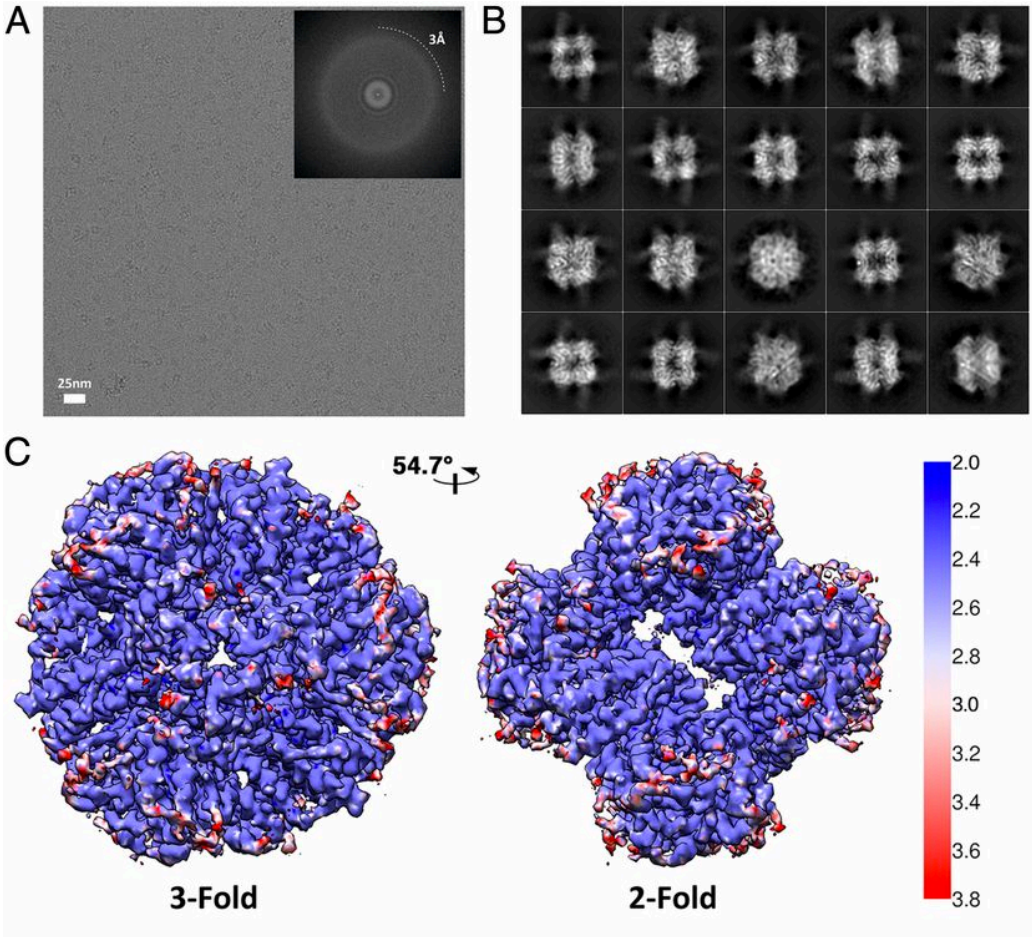
ensemble of predicted models

Experimental methods

- ✓ X-ray crystallography
- ✓ NMR spectroscopy
- ✓ **Cryo-electron microscopy**

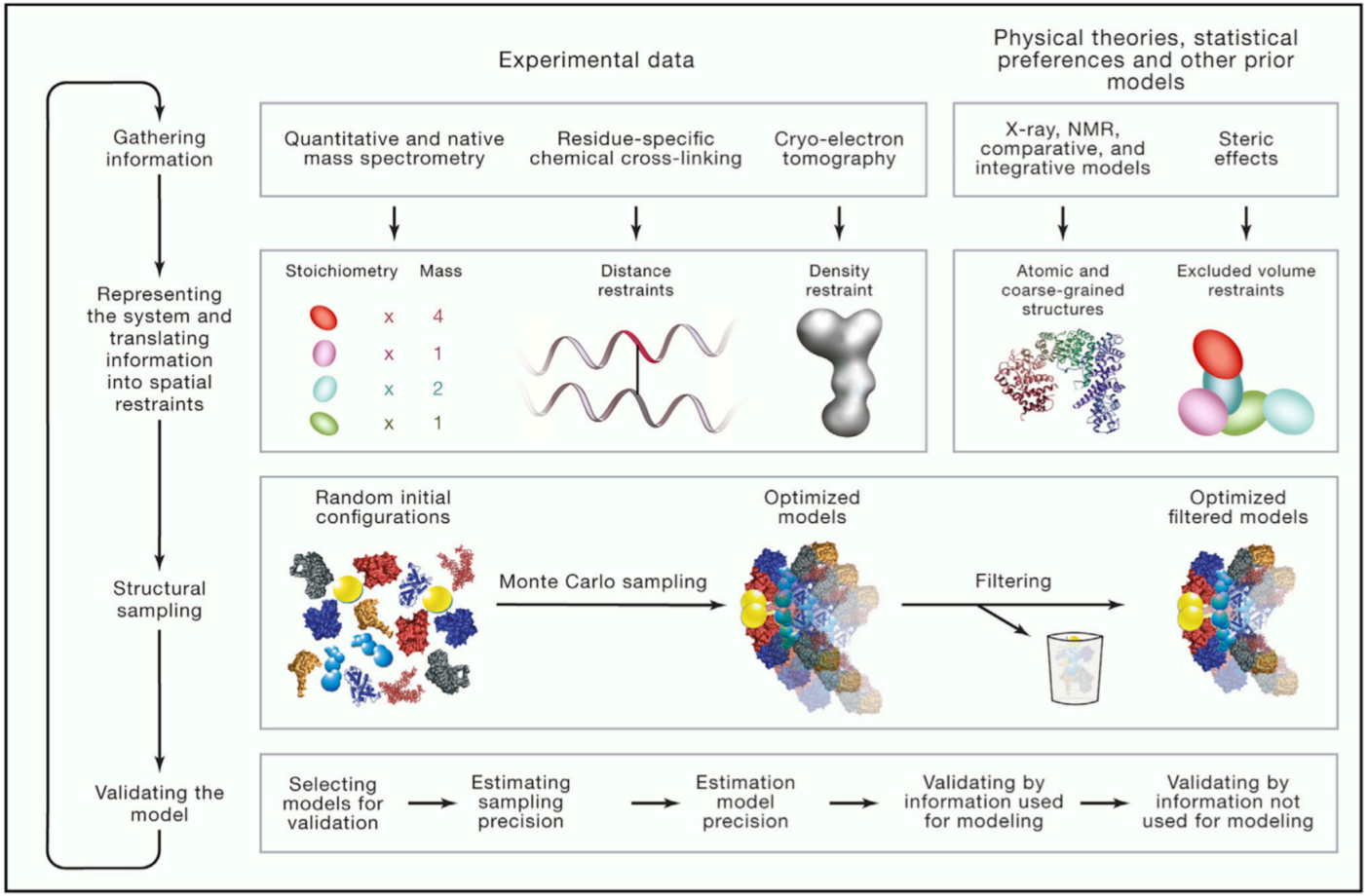
- ➢ physiological conditions
- ➢ Very big systems
- ➢ Low to high resolution
- ➢ A lot of unexploited data!

The resolution revolution
Science 2014



Integrative structural biology

All these experimental methods must be coupled with computational methods, to treat raw data and build 3D models.



Sali *Cell* 2019

Protein degrees of freedom

A protein is composed of tens to thousands of amino acid residues

------ each residue is composed of 10-20 atoms

---------- each atom possesses 3 degrees of freedom (coord X, Y and Z)

Protein degrees of freedom

A protein is composed of tens to thousands of amino acid residues

------ each residue is composed of 10-20 atoms

---------- each atom possesses 3 degrees of freedom (coord X, Y and Z)

But… proteins cannot do « anything »! The interatomic forces (covalent or weak) constrain their motions

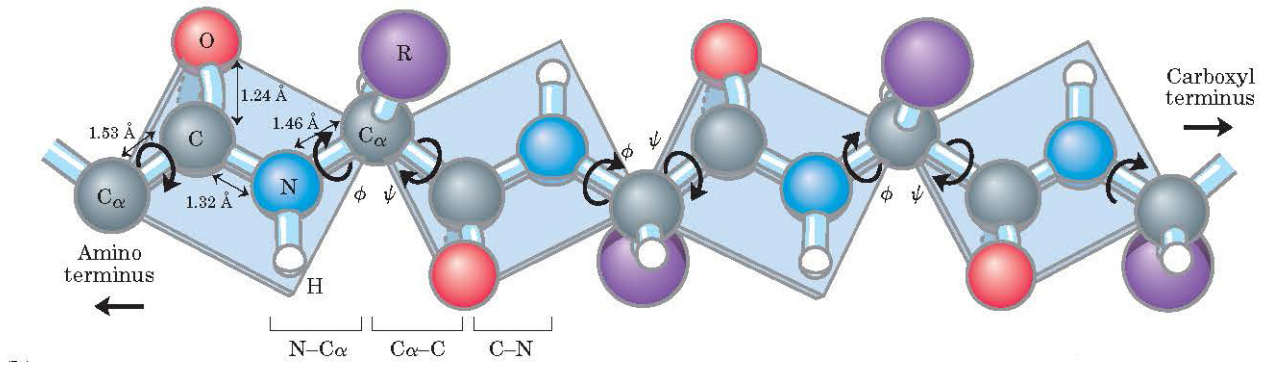
Protein degrees of freedom

A protein is composed of tens to thousands of amino acid residues

------ each residue is composed of 10-20 atoms

---------- each atom possesses 3 degrees of freedom (coord X, Y and Z)

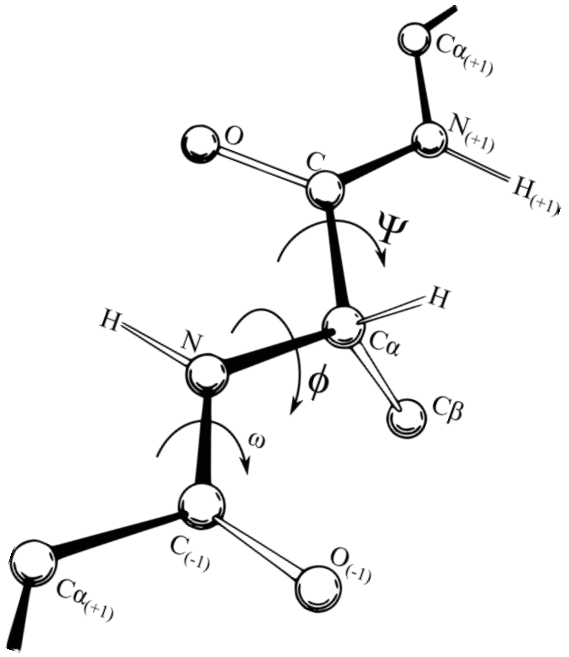
But… proteins cannot do « anything »! The interatomic forces (covalent or weak) constrain their motions



planar peptidic units

Backbone torsion angles

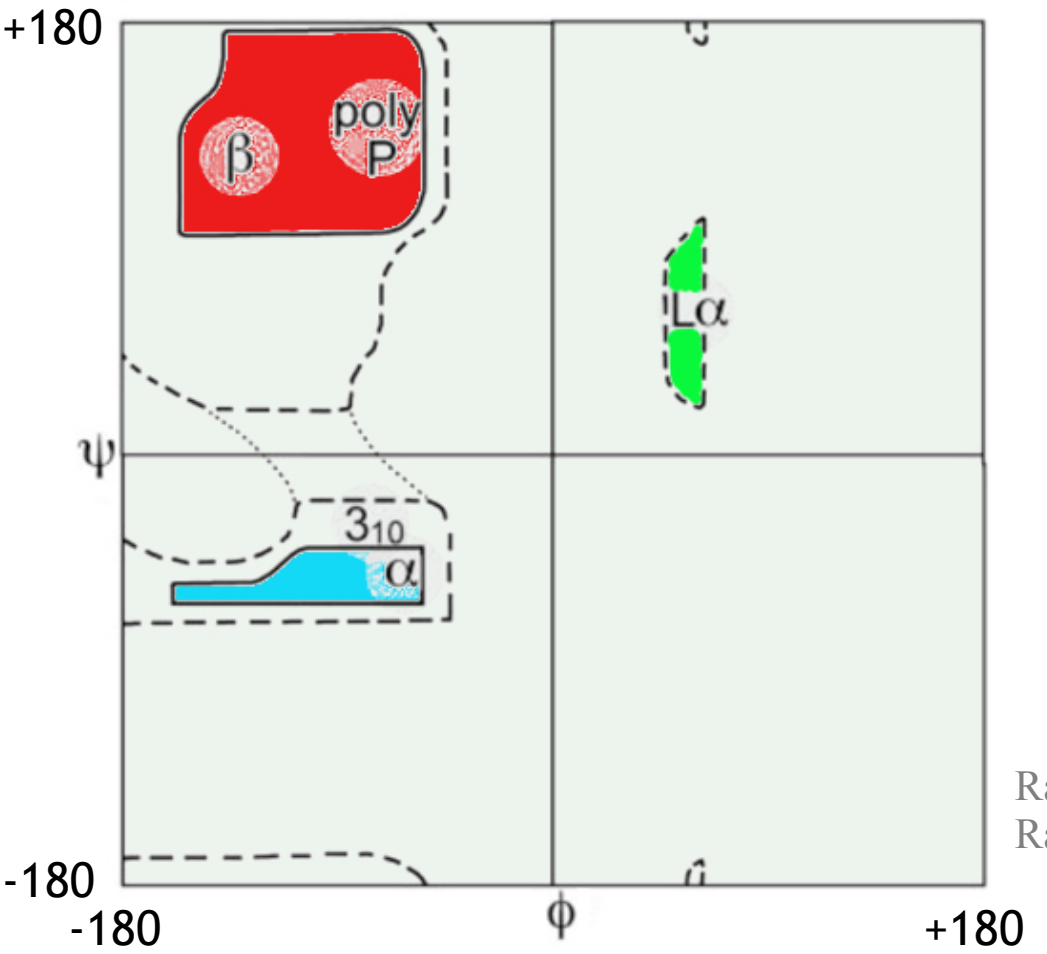
The conformation of the whole backbone is completely determined when the rotation angles **φ (N-Cα)** **etψ = (Cα-C')** are defined with precision.



These torsion angles form a system of internal coordinates in which we can represent the 3D structure of the protein.

Quality assessment

Ramachandran diagram

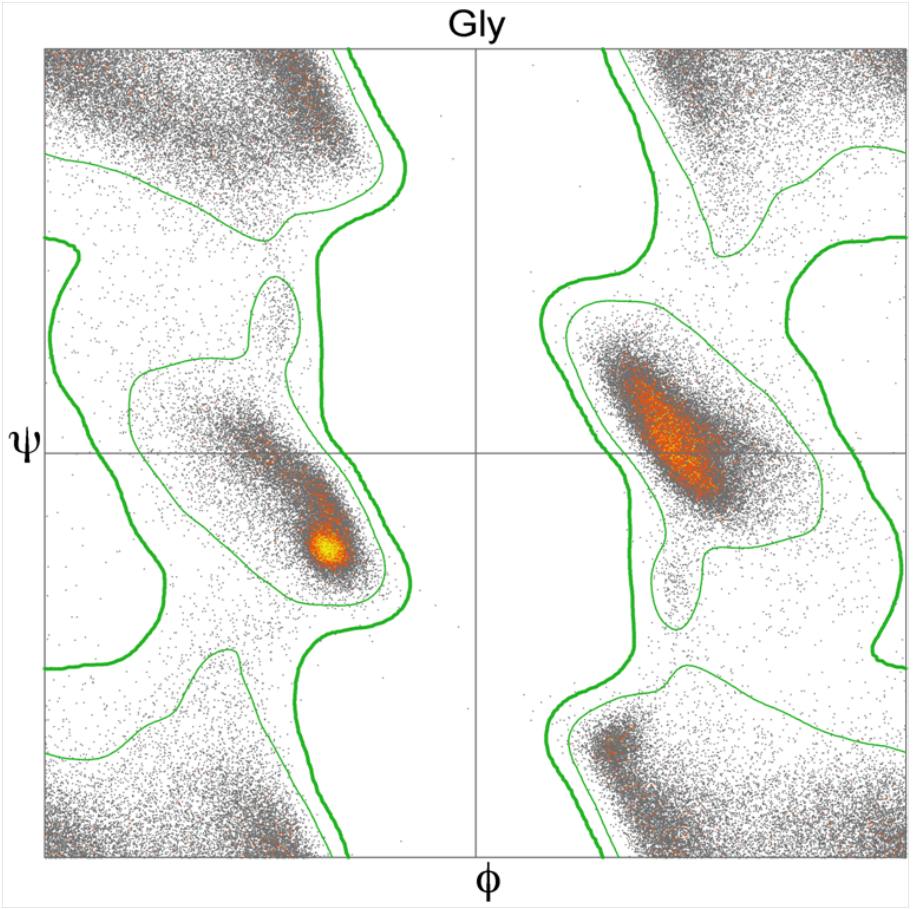


Most of the **φ and ψ** combinations for an amino acid are not allowed because of steroidal clashes between the main and side chains

Ramachandran *et al.* (1963) *J. Mol. Biol.*
Ramachandran *et al.* (1968) *Adv. Protein. Chem.*

Quality assessment

Ramachandran diagram

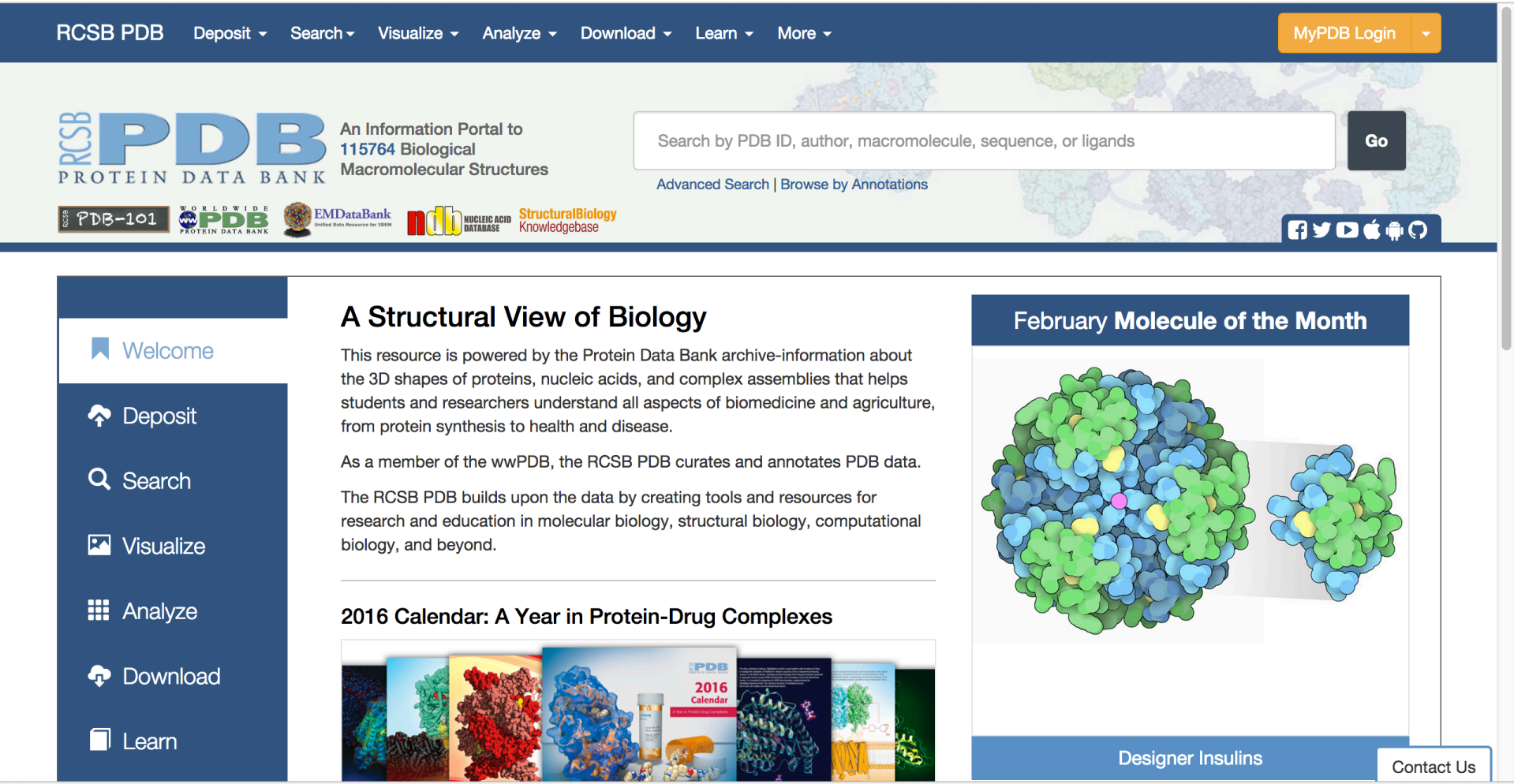


Glycine can adopt many more conformations than the others, and hence play an important structural role

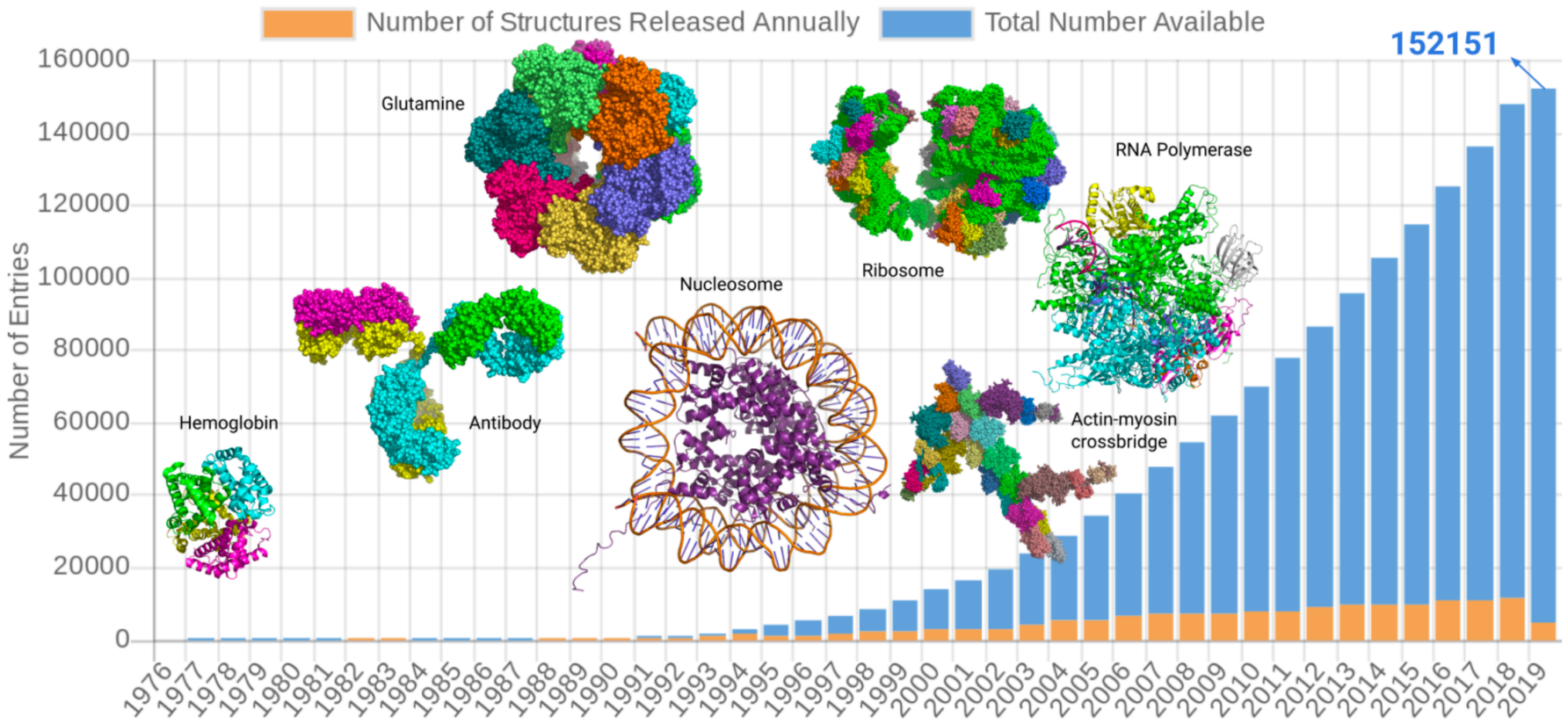
Ramachandran *et al.* (1963) *J. Mol. Biol.*
Ramachandran *et al.* (1968) *Adv. Protein. Chem.*

Database of protein structures

http://www.rcsb.org/ (voir aussi https://www.ebi.ac.uk/services/structures)



Database of protein structures

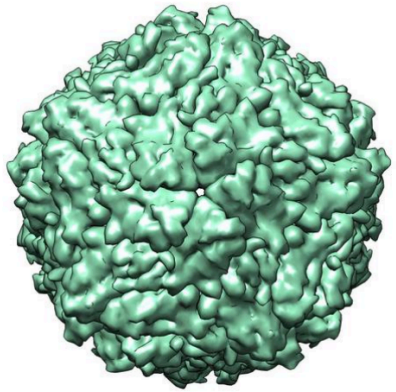


Database of protein structures

Structure solved by cryo-electron microscopy

Structure Summary | 3D View | Annotations | Sequence | Sequence Similarity | Structure Similarity | Experiment

Biological Assembly 1



View in 3D: Jmol (in Browser)

Standalone Viewers Large

Simple Viewer Protein Workshop
Ligand Explorer

Protein Symmetry: Icosahedral (View in 3D)

Protein Stoichiometry: Homo 60-mer - A60

Display Files | Download Files

3JCI

2.9 Angstrom Resolution Cryo-EM 3-D Reconstruction of Close-packed PCV2 Virus-like Particles

DOI: 10.2210/pdb3jci/pdb **EMDataBank**: EMD-6555 Download EM Map

Classification: VIRUS LIKE PARTICLE
Deposited: 2015-12-13 **Released**: 2016-02-03
Deposition author(s): Liu, Z., Guo, F., Wang, F., Li, T.C., Jiang, W.
Organism: Porcine circovirus-2

Structural Biology Knowledgebase: 3JCI SBKB.org

Experimental Data Snapshot

Method: ELECTRON MICROSCOPY
Resolution: 2.9 Å
Aggregation State: Particle
Specimen Type: Vitreous Ice (cryo Em)
Reconstruction Method: Single Particle

wwPDB Validation Full Report

Validation Report Pending

Literature Download Primary Citation

2.9 angstrom Resolution Cryo-EM 3D Reconstruction of Close-Packed Virus Particles

Contact Us

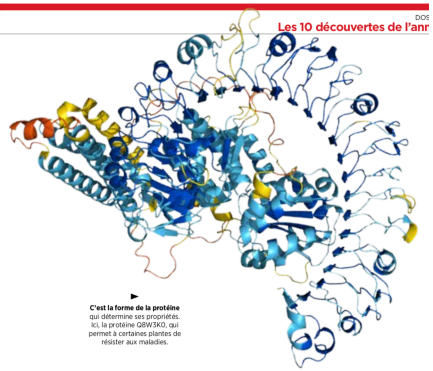
4

Les 10 découvertes de l'anr

Le repliement des protéines résolu par une IA

L'intelligence artificielle, nouveau graal de la biologie? C'est ce que suggère le résultat spectaculaire d'Alphafold, l'algorithme de DeepMind qui prédit la forme des protéines. De quoi révolutionner la conception de nouveaux médicaments.

Le visage de la biologie computationnelle n'est plus le même depuis l'avènement du programme AlphaFold2. Publié en juillet 2021 dans la revue *Nature*, l'article décrivant cette avancée spectaculaire des chercheurs de Deep Mind, la filiale de Google dédiée à l'intelligence artificielle, a beaucoup été commenté. Le problème posé était le suivant: soit une suite d'acides aminés donnée, cules de la protéine. Si la nature sait s'y prendre, les biologistes ont bien du mal à prédire cette forme stable, qui confère ses propriétés à la protéine. En cause, la combinatoire qui est énorme: une protéine comprend plusieurs dizaines de milliers d'atomes en interaction. « *Typiquement, une protéine contient entre 100 et 300 acides aminés* », explique Krzysztof Fidelis, directeur du Protein *est unique et dépend essentiellement de sa séquence d'acides aminés* », raconte Krzysztof Fidelis. Mais la résolution de ce problème d'énoncé simple allait se révéler redoutable, car les interactions entre atomes sont très nombreuses, de sorte que les simulations numériques étaient encore, après plusieurs décennies de travaux, loin du compte.



C'est la forme de la protéine qui détermine ses propriétés. Ici, la protéine G8W3K0, qui permet à certaines plantes de résister aux maladies.

NEWS | BIOLOGY

'The game has changed.' AI triumphs at solving protein structures

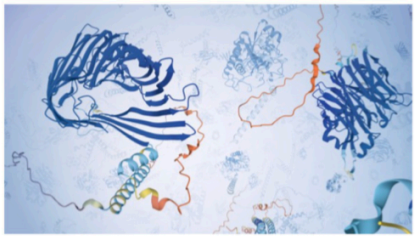
Science

In milestone, software predictions finally match structures calculated from experimental data

FOCUS | 11 JANUARY 2022

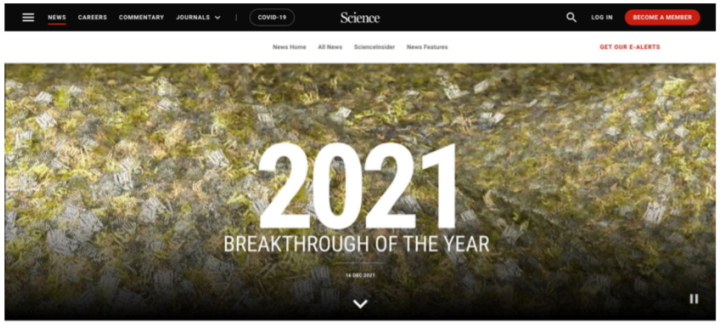
Method of the Year 2021: Protein structure prediction

Protein structure prediction is our Method of the Year 2021, for the remarkable levels of accuracy achieved by deep learning-based methods in predicting the 3D structures of proteins and protein complexes, essentially solving this long-standing challenge.



Boris Johnson
@BorisJohnson
United Kingdom government official

.@DeepMind's announcement today with @emblebi demonstrates the very best in UK science – powered by AI, they are opening up biological data and research globally that will help accelerate transformative scientific breakthroughs for people around the world.



One of biology's biggest mysteries 'largely solved' by AI

BBC

By Helen Briggs
BBC science correspondent

NEWS | 30 November 2020

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

Nature

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

A.I. Predicts the Shapes of Molecules to Come

NY Times

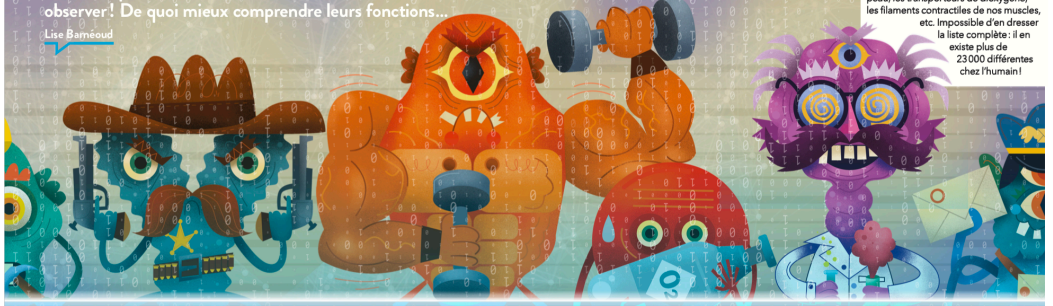
DeepMind has given 3-D structure to 350,000 proteins, including every one made by humans, promising a boon for medicine and drug design.

QUOI DE NEUF?

L'IA QUI A DÉVOILÉ LE VRAI VISAGE DES PROTÉINES

On savait qu'elles avaient toutes leur look, mais on n'arrivait pas à leur tirer le portrait. Et voici qu'une machine intelligente a réussi à prédire la structure de ces molécules, sans même les observer! De quoi mieux comprendre leurs fonctions...

Lise Barnéoud



Sans elles, vous ne seriez pas là, à lire votre journal préféré. Vous ne pourriez faire le moindre mouvement, voir, sentir ou respirer. Elles, ce sont les protéines! Ces molécules fabriquées par nos cellules sont les ouvrières de la vie. Des travailleuses ultra-spécialisées qui bossent jour et nuit pour faire tourner l'immense machine qu'est notre organisme. Plus concrètement, les protéines sont les enzymes qui digèrent nos repas, les anticorps qui nous protègent des microbes, les fibres qui donnent sa structure à notre peau, les transporteurs de dioxygène, les filaments contractiles de nos muscles, etc. Impossible d'en dresser la liste complète: il en existe plus de 23 000 différentes chez l'humain!

PLUS DE 23 000 DIFFÉRENTES CHEZ L'HUMAIN!

À quoi ressemblent-elles? Jusqu'à présent, nous connaissions l'apparence d'à peine 20 % d'entre elles. Mais une équipe de geeks s'en est mêlée. Et a mis au point un nouveau logiciel d'intelligence artificielle, capable de prédire le look de chacune... Un incroyable succès qui devrait permettre de mieux comprendre comment agissent ces travailleuses de l'ombre. Pour saisir l'importance de ces travaux, revenons aux bases. Imaginez les protéines comme des constructions en Lego.

Depuis plus d'un demi-siècle, nous connaissons toutes les pièces avec lesquelles elles sont fabriquées: il y en a 20 différentes, appelées «acides aminés». Nous avons même accès au plan en 2D de chaque protéine, car il est inscrit dans nos gènes.

Le grand Lego de la vie

Un gène ressemble un peu à une fiche sur laquelle seraient dessinées les pièces de Lego nécessaires à l'assemblage d'une protéine donnée. Ces pièces sont alignées les unes à côté des autres, dans un ordre bien précis, comme dans un immense collier. Sauf que, une fois fabriqué dans nos cellules *(voir schéma pages 40-41)*, ce collier de Lego se contorsionne, se replie, si bien que certaines pièces, parfois très éloignées les unes des autres dans la chaîne, finissent par s'emboîter pour former une structure en 3D. Et ce sont justement les images de cet assemblage final, en volume, qui nous manquaient. Plutôt frustrant! C'est comme si l'on disposait de la liste des Lego, mais sans le plan de montage: comment savoir si l'on va obtenir le vaisseau de *Star Wars* ou le

Database of **predicted** protein structures

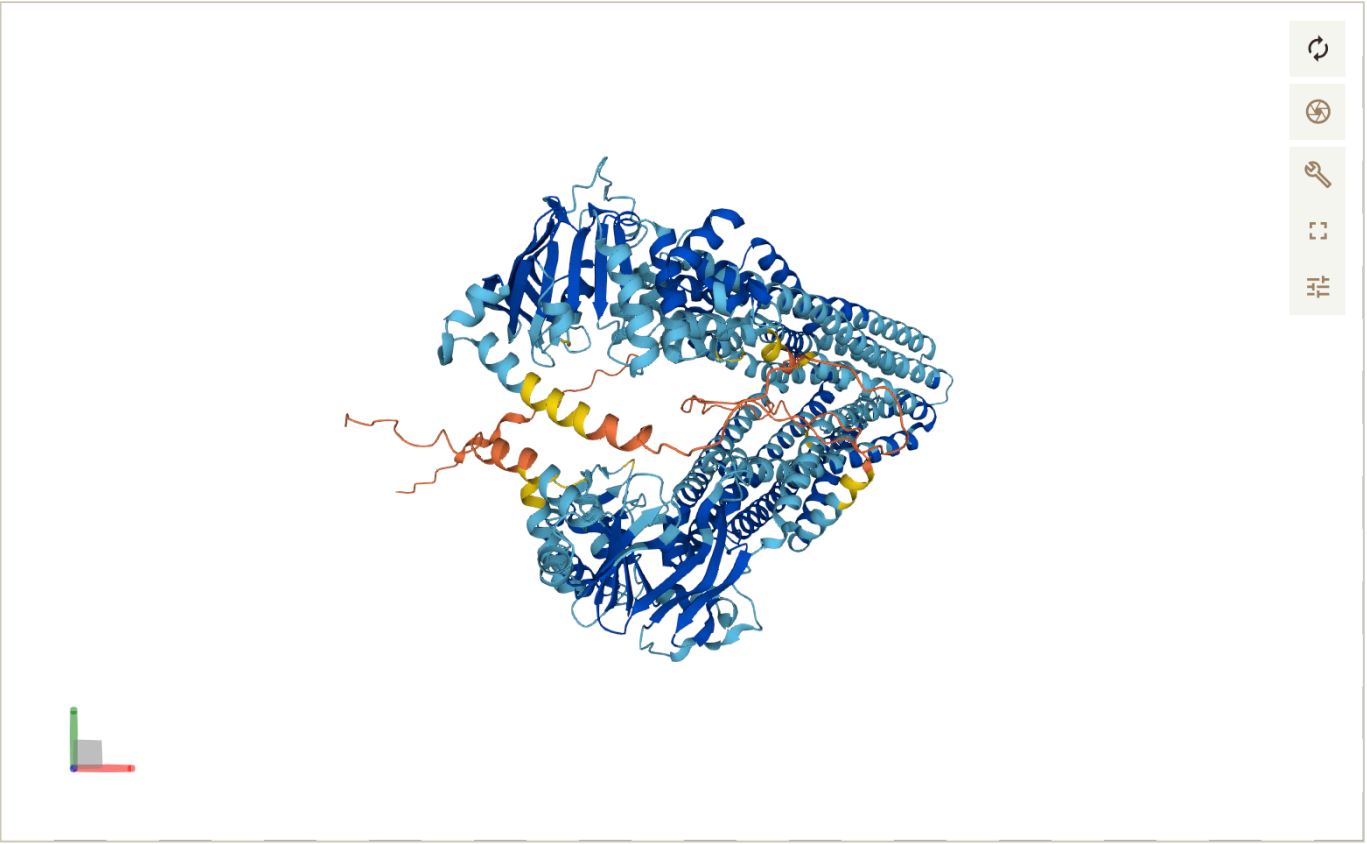
https://alphafold.ebi.ac.uk/ (available from Uniprot and the PDB)

Structure[i]

Model Confidence:

- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions with low pLDDT may be unstructured in isolation.

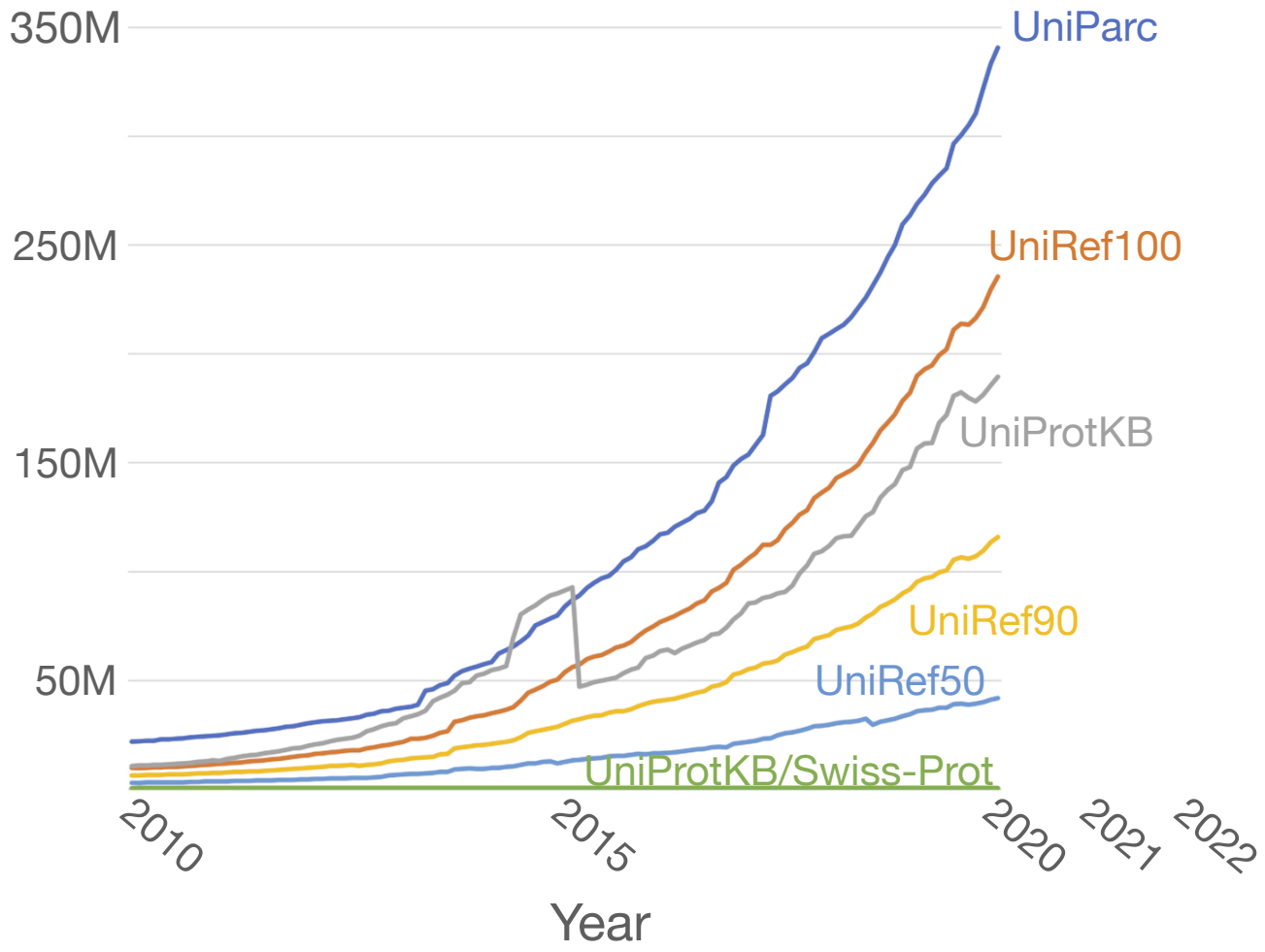


Source	Identifier	Method	Resolution	Chain	Positions	Links
AlphaFold	AF-Q9ZR72-F1	Predicted			1-1286	AlphaFold

ALGORITHMS IN STRUCTURAL BIOINFORMATICS

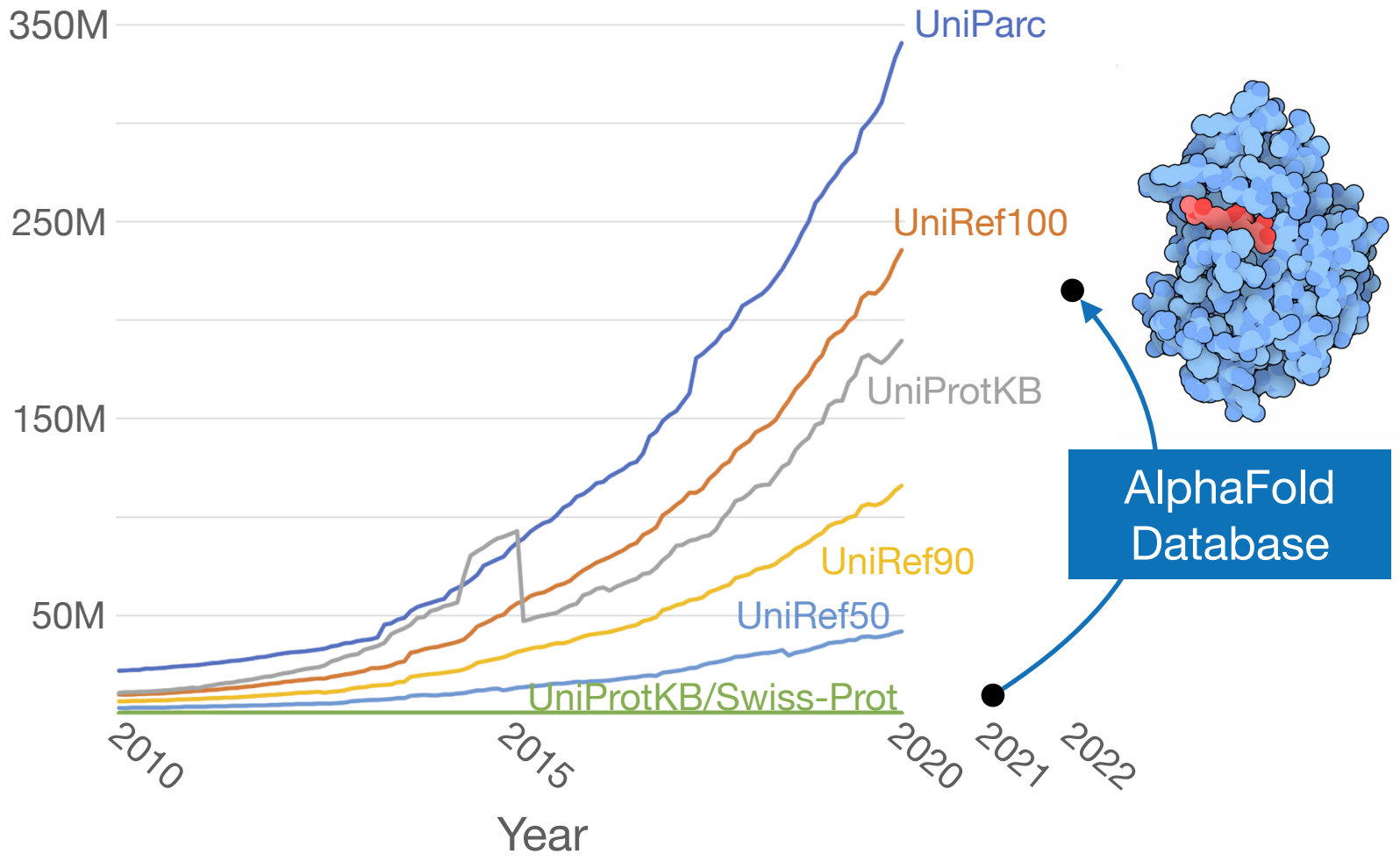
Algorithms, what for?

- **To predict protein structures**
  - experimental data analysis and 3-dimensional model building
  - secondary or tertiary structure prediction based on the sequence



Algorithms, what for?

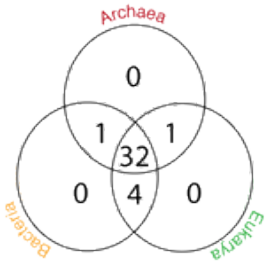
- **To predict protein structures**
  - experimental data analysis and 3-dimensional model building
  - secondary or tertiary structure prediction based on the sequence

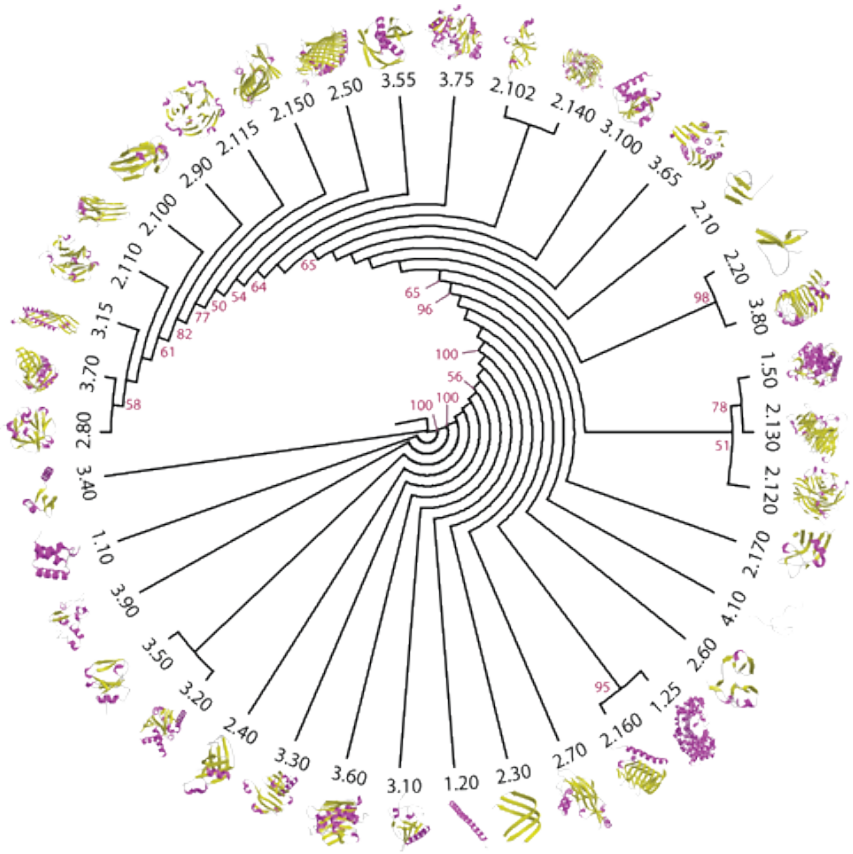


Algorithms, what for?

- **To predict protein structures**

- **To compare protein structures**
  - classification of proteins (divergent/convergent evolution)
  - identification of active sites, functional motifs or binding sites

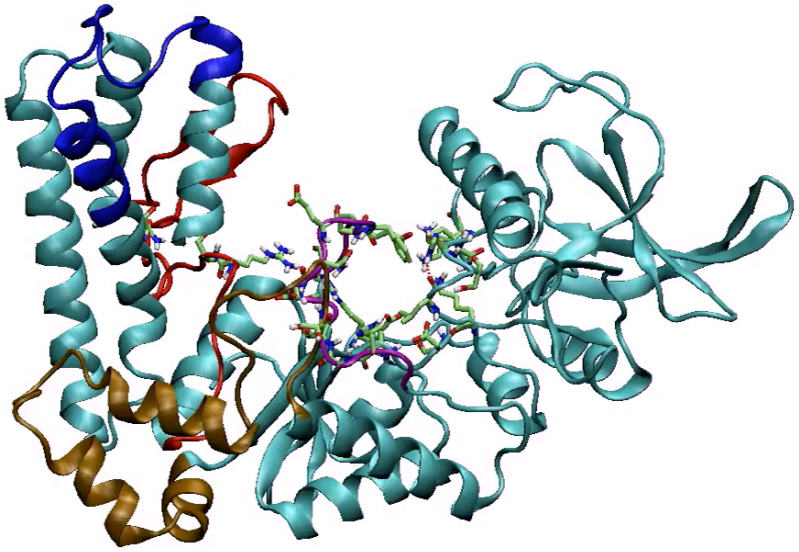




Phylogenetic tree of 38 CATH Architecture domain structures

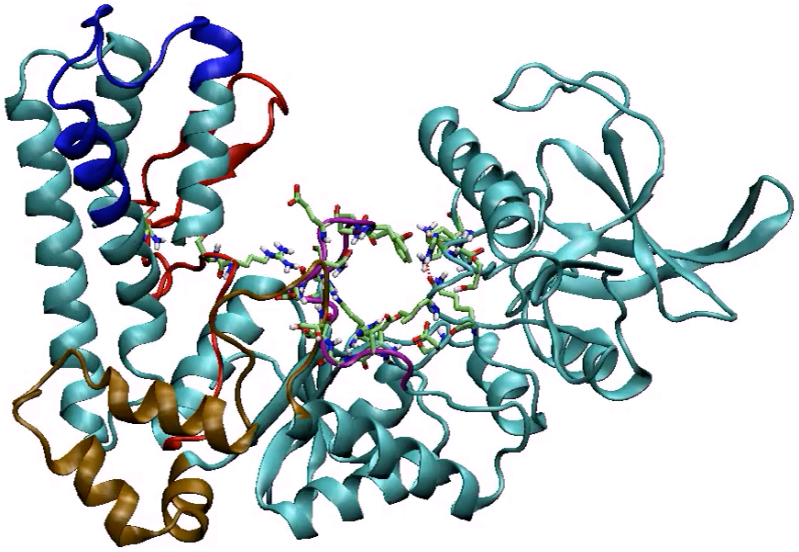
Algorithms, what for?

- **To predict protein structures**
- **To compare protein structures**
- **To simulate protein motions**
  - atomic-level description of the mechanisms underlying protein activity
  - characterization of intermediate conformations that can be targeted by drugs



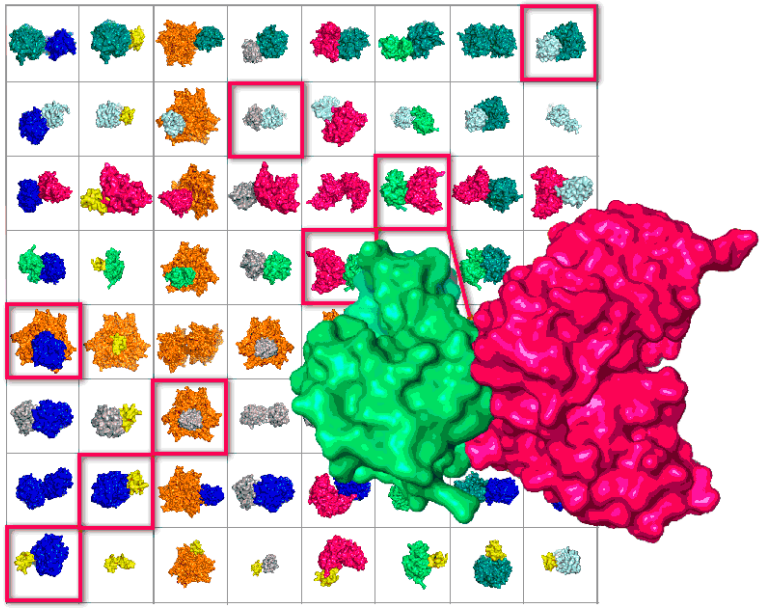
Algorithms, what for?

- **To predict protein structures**
- **To compare protein structures**
- **To simulate protein motions**
  - atomic-level description of the mechanisms underlying protein activity
  - characterization of intermediate conformations that can be targeted by drugs



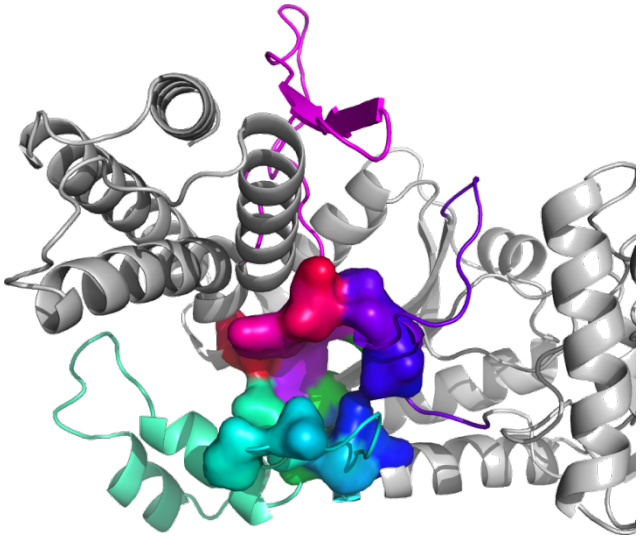
Algorithms, what for?

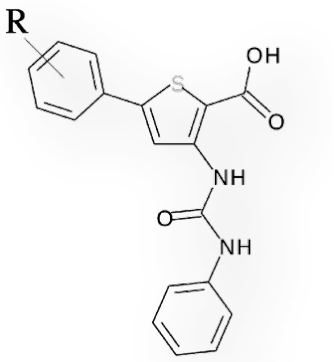
- **To predict protein structures**
- **To compare protein structures**
- **To simulate protein motions**
- **To characterize protein interactions**
  - protein interaction sites identification and complex structures prediction
  - discrimination between true partners in the cell and non-interactors



Algorithms, what for?

- **To predict protein structures**
- **To compare protein structures**
- **To simulate protein motions**
- **To characterize protein interactions**
- **To discover and design drugs**
  - putative druggable pockets identification
  - binding mode and relative affinity prediction





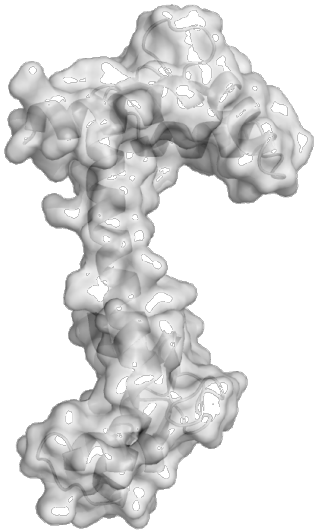
Step 1
Looking at proteins

UniProt

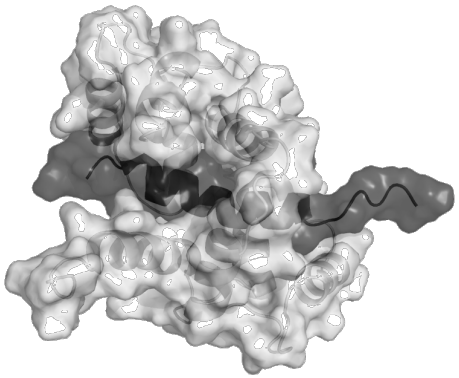
Human Calmodulin

https://www.uniprot.org/uniprotkb/P0DP23/entry

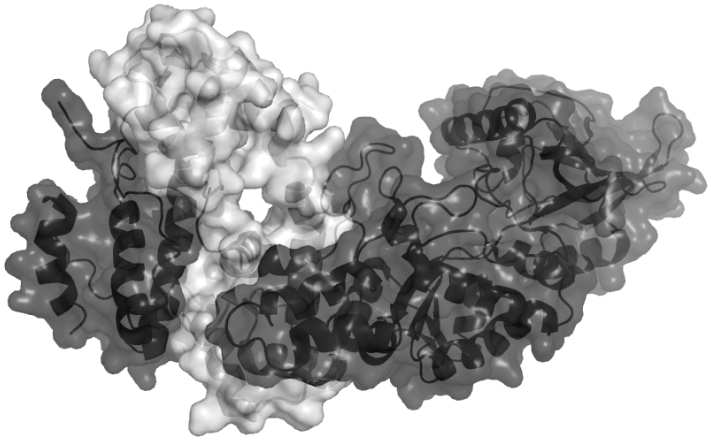
1CLL



2BBM



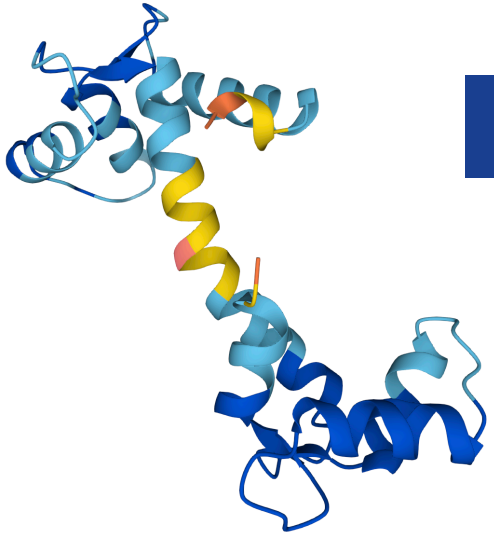
1K93



Model Confidence:

- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions with low pLDDT may be unstructured in isolation.

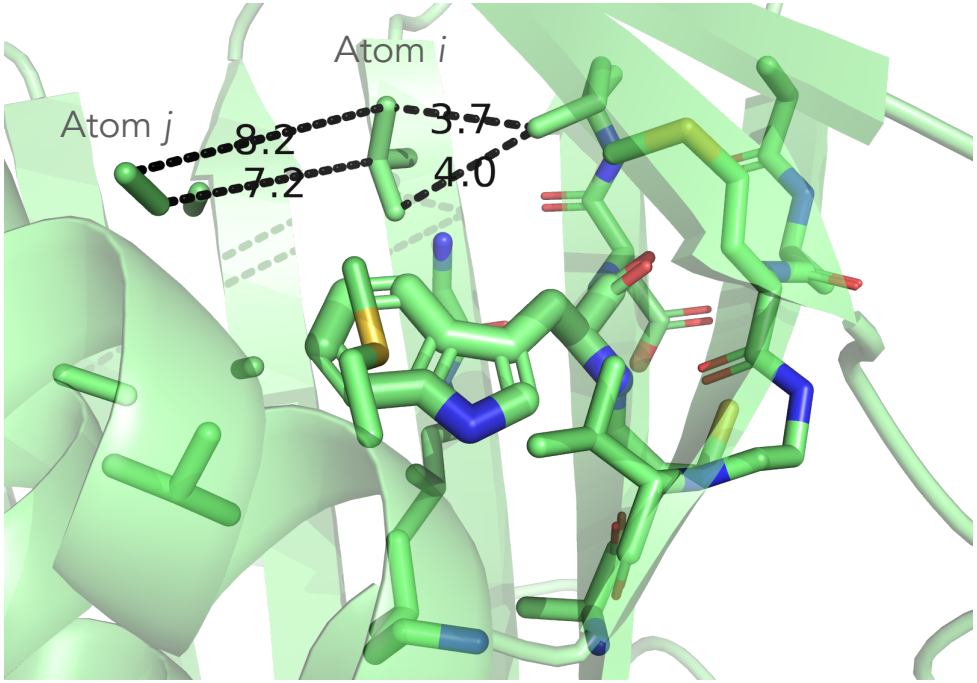


AlphaFold Protein Structure Database

https://alphafold.ebi.ac.uk

Tunyasuvunakool *et al.* 2021

Varadi *et al.* 2021



Predicted lDDT per position

To compute lDDT, we look at the neighbourhood of each residue, within a certain inclusion radius.

- For each pair *(i,j)* of atoms in the neighbourhood, compute the distance *dij*.
- For each threshold value *tcut = 0.5, 1, 2, 4 A*, compute the fraction of preserved distances:

$$\frac{\#(|d_{ij} - d_{ij}^{TRUE}| < t_{cut})}{N_{pairs}}$$

The final lDDT value is averaged over the *tcut* values. Mariani *et al.* 2021

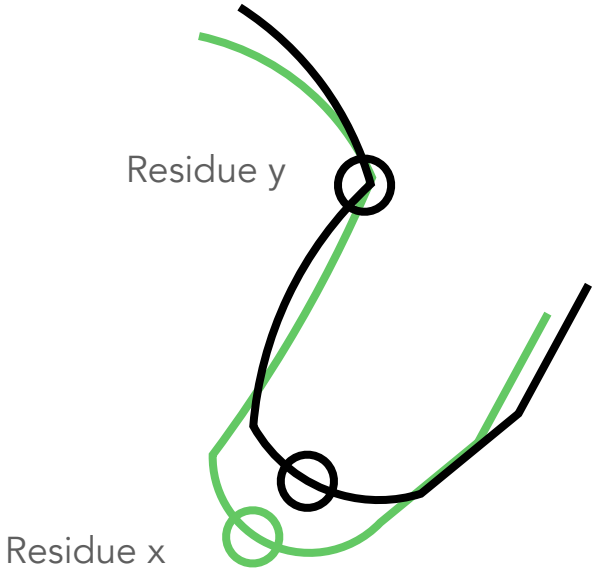
Predicted Aligned Error

Independent of the 3D structure, AlphaFold produces an output called “Predicted Aligned Error”. This is shown at the bottom of structure pages as an interactive 2D plot.

- The colour at (x, y) indicates AlphaFold’s expected position error at residue x if the predicted and true structures were aligned on residue y.
- If the predicted aligned error is generally low for residue pairs x, y from two different domains, it indicates that AlphaFold predicts well-defined relative positions for them.
- If the predicted aligned error is generally high for residue pairs x, y from two different domains, then the relative positions of these domains in the 3D structure is uncertain and should not be interpreted.

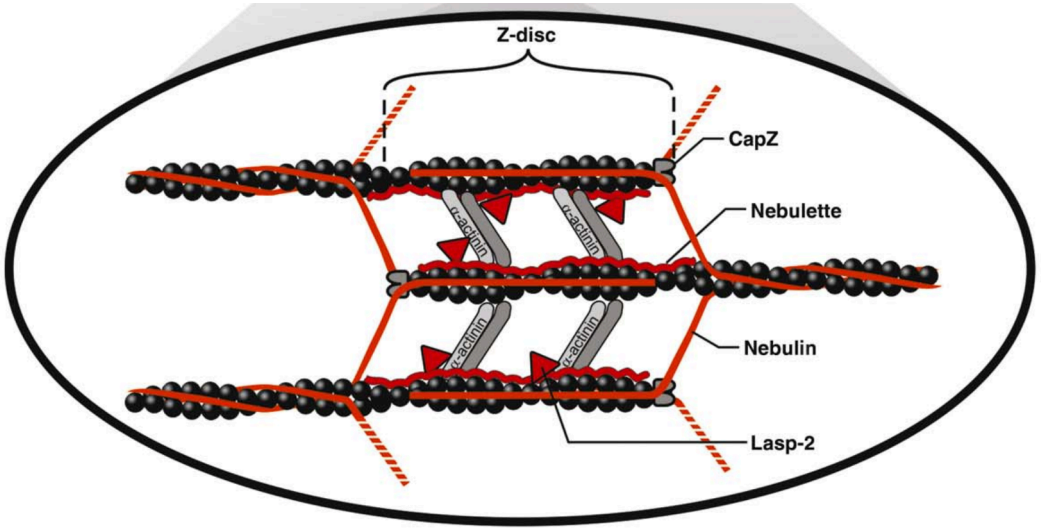
AlphaFold produces a useful inter-domain prediction in some cases. However, in CASP14 intra-domain prediction accuracy was more extensively validated and is therefore expected to be more reliable.

https://alphafold.ebi.ac.uk/faq#faq-6



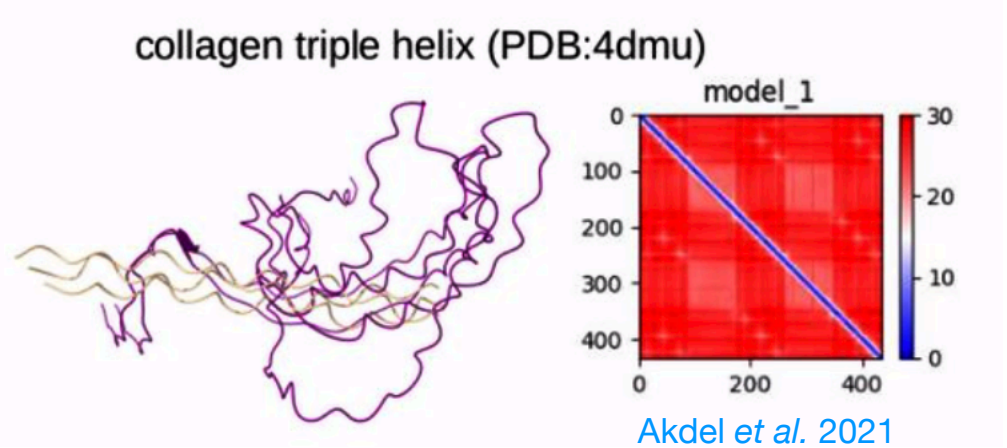
What about structural proteins?

Pappas *et al.* 2010



Nebulette: O76041

Collagen: P02452



Akdel *et al.* 2021

Step 2
Making proteins!

Repository: https://github.com/elolaine/AI4Biologists

Notebook: *AlphaFold Single*

Harvard FAS Center for Systems Biology

What does it do?

Given an input protein sequence, it runs AlphaFold to predict its 3D structure. **S. Ovchinikov** modified AF code to make it faster (compilation tricks). The output is a 3D model.

- Recycles: By default, the prediction is performed by a single pass through the network. Alternatively you can execute the network several times via the « recycling » procedure. The outputs of each pass is recorded and used as additional inputs for the next pass.
- Animate: the Animate cell allows you to visualise the evolution of the 3D predicted model along the recycles.

Task: design a protein that fold into… a helix, 2 helices, a beta-strand…

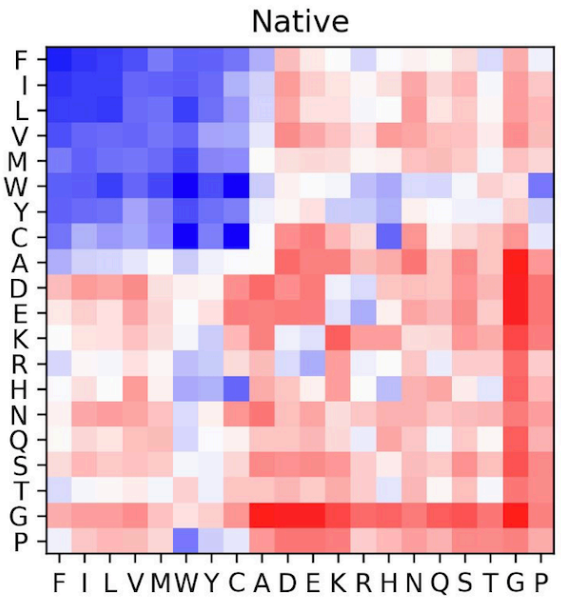
Secondary structure propensities

	C	H	E
A	-0.41	0.46	-0.40
E	-0.27	0.42	-0.53
Q	-0.24	0.36	-0.39
K	-0.05	0.21	-0.32
R	-0.22	0.24	-0.14
M	-0.37	0.26	0.03
L	-0.64	0.35	0.13
W	-0.37	0.06	0.36
Y	-0.36	-0.03	0.46
F	-0.39	-0.03	0.48
I	-0.77	-0.03	0.74
V	-0.69	-0.25	0.89
C	-0.06	-0.25	0.41
T	0.14	-0.39	0.30
H	0.14	-0.15	0.01
S	0.32	-0.25	-0.17
N	0.52	-0.30	-0.65
D	0.50	-0.21	-0.78
G	0.82	-1.05	-0.56
P	0.90	-0.94	-1.06

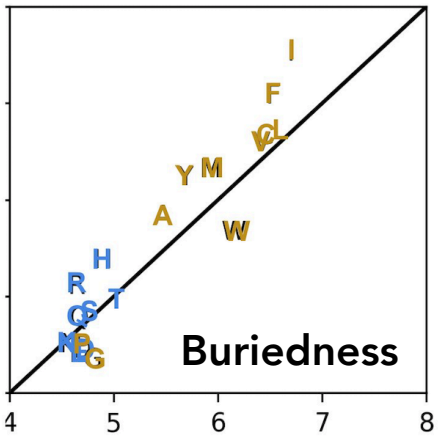
log(P(AA|SS) / P(AA))

H: helix, E: sheets

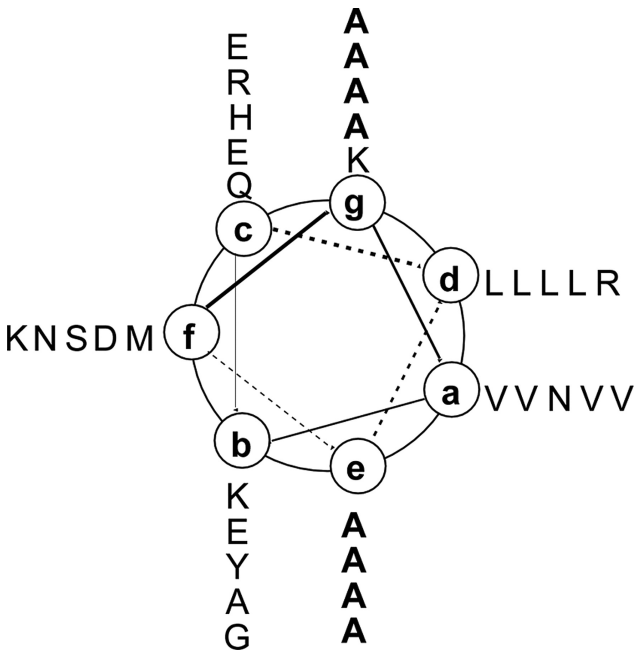
Interaction potential



log(P(A,B|Contact) / P(A)*P(B))



An example of a coiled coil protein



https://www.pnas.org/doi/10.1073/pnas.0604871103

Liu *et al.* 2006

Try and design a coiled-coil !

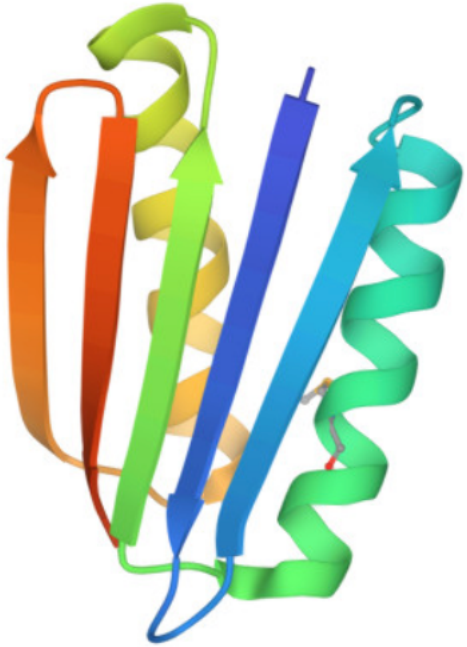
Take a moment to listen: lecture on secondary structure prediction

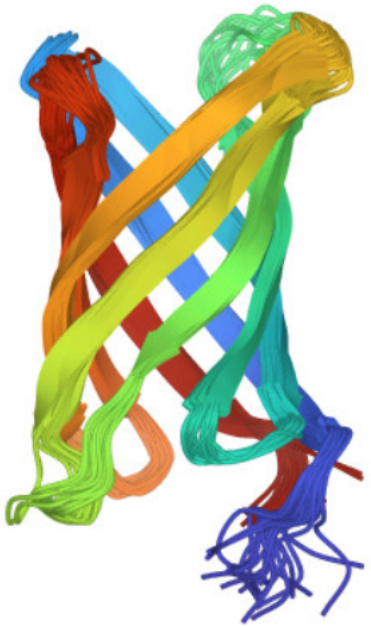
https://www.rcsb.org/structure/1QYS

Maybe the most popular designed protein

TOP7

(solved in 2003)





https://www.rcsb.org/structure/6X1K

What about a more recent one?

(solved in 2020, NOT in AF2 training set)

What happens if you try and predict a natural protein from a single sequence?

For instance **human myoglobin** (3RGK)

Fast and accessible full version of AF2

https://github.com/sokrypton/ColabFold

Mirdita *et al.* 2022

AF2 powered by mmseqs2!

Task: try out your favorite protein! For instance, human myoglobin again, but toggling on the MSA generation this time :) Are the results better?

You can first check what to expect…

- Go to the BLAST website: https://blast.ncbi.nlm.nih.gov/Blast.cgi
- Query the sequence against the PDB. Do we know the structure of a close homolog?

Step 3
Comparing proteins!

Pymol for visualising & manipulating protein structures

https://pymolwiki.org/index.php/Main_Page

Delano 2001

http://legacy.ccp4.ac.uk/newsletters/newsletter40/11_pymol.pdf

Task: Visualise the 2 models predicted by AlphaFold2, with and without MSA, and superimpose them onto the experimental structure 3RGK.

You can color the residues by the values reported in the 12th column.

- Predicted lDDT for the AF2 models
- Temperature factors for the X-ray experimental structure