De novo protein design workshop

TA: Jody Mou

Nick Polizzi

Dana-Farber Cancer Institute

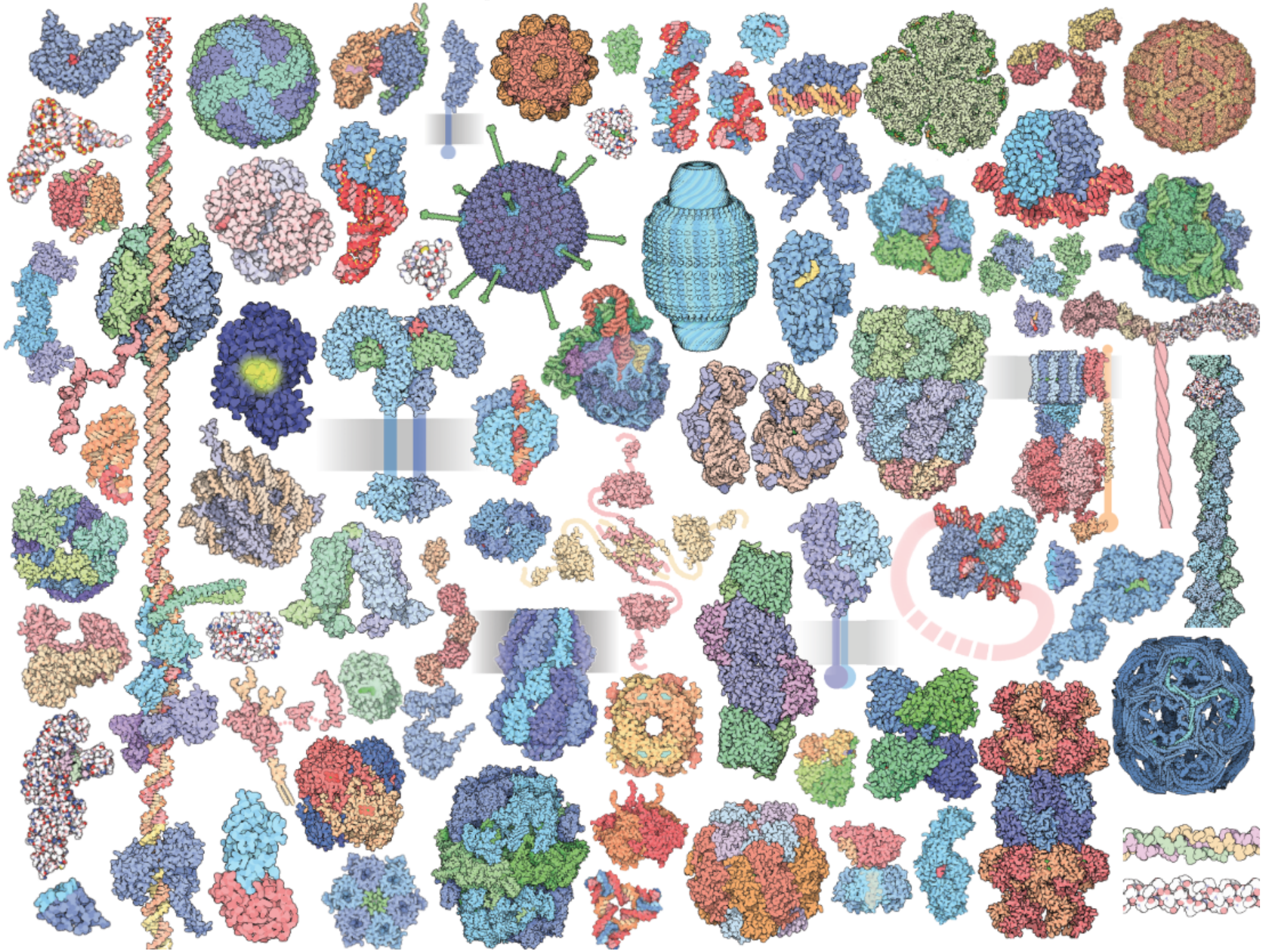
Principal Investigator
Dept of Cancer Biology
Dana-Farber Cancer Institute

HARVARD MEDICAL SCHOOL

Assistant Professor
Dept of Biological Chemistry
and Molecular Pharmacology
Harvard Medical School

https://pdb101.rcsb.org/motm/motm-about

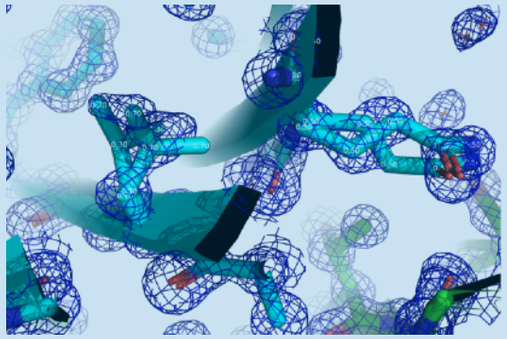
The protein universe is diverse



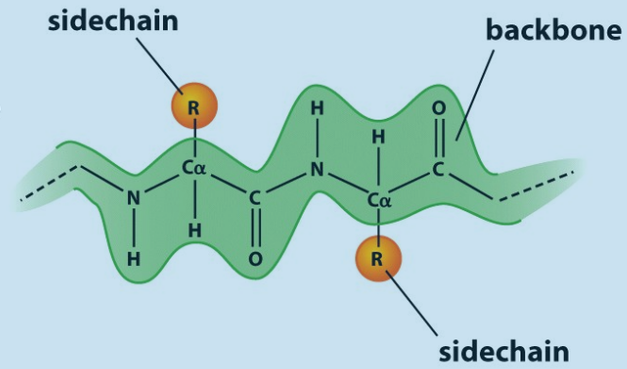
We are increasingly becoming able to design proteins that resemble these

https://pdb101.rcsb.org/motm/motm-about

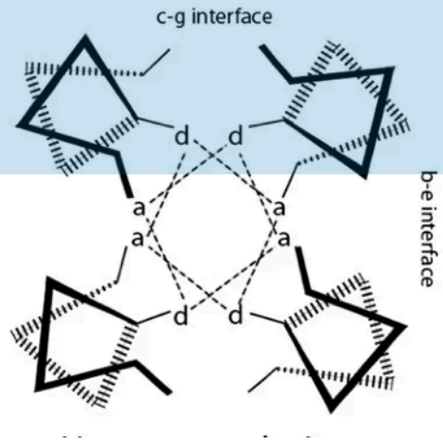
Workshop Learning Objectives



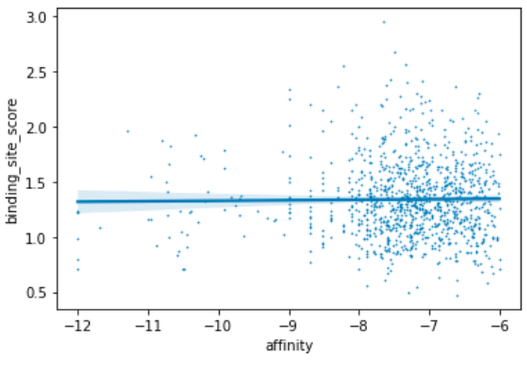
- **Understand basic elements of protein structure**



- **Assess quality of X-ray crystal structures**
  - The data that underpins all modern ML models

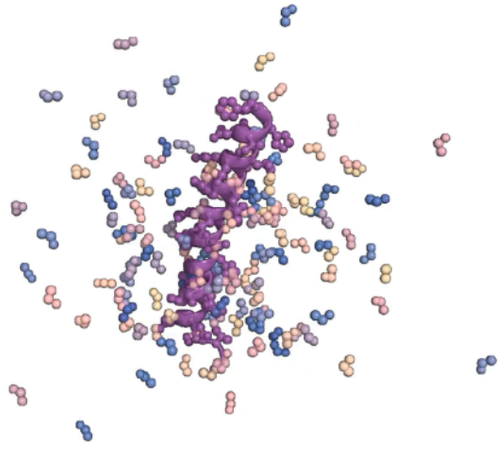


- **Understand sequence <—> structure relationships**



- **Gain familiarity with coding to manipulate proteins**
  - Python packages ProDy, PyTorch, Matplotlib, Plotly, etc

- **Learn to use the latest protein design tools to:**
  - Generate or alter protein structure
  - Design protein sequence, given a structure
  - Assess quality of designed proteins



Workshop Projects

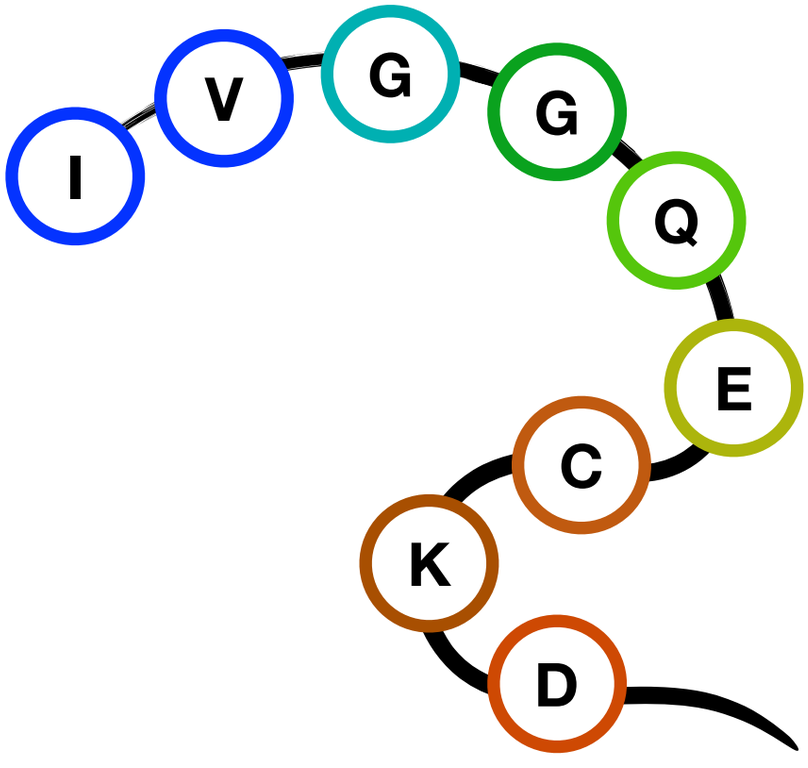
- **Set a design goal and use what you've learned to design a protein(s)**
  - Consult with me and Jody on project ideas

- **Describe your approach and analysis with in a 10 min presentation on Friday**
  - The data that underpins all modern ML models

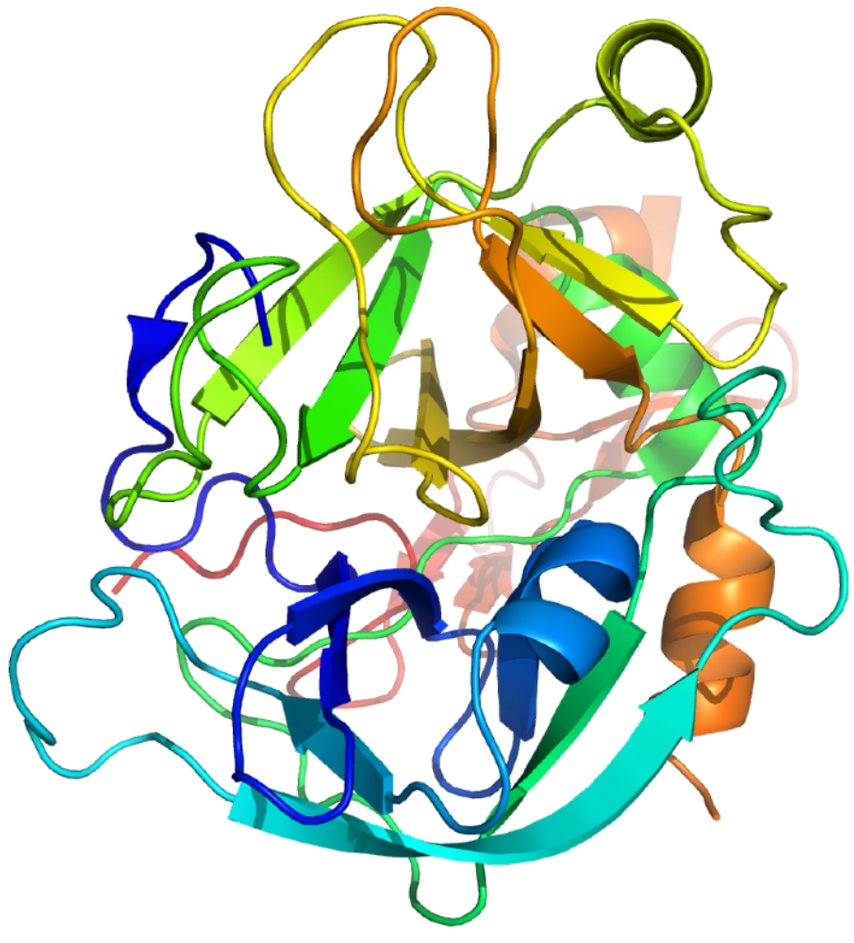
What is protein design?

Anfinsen's hypothesis:

protein sequence



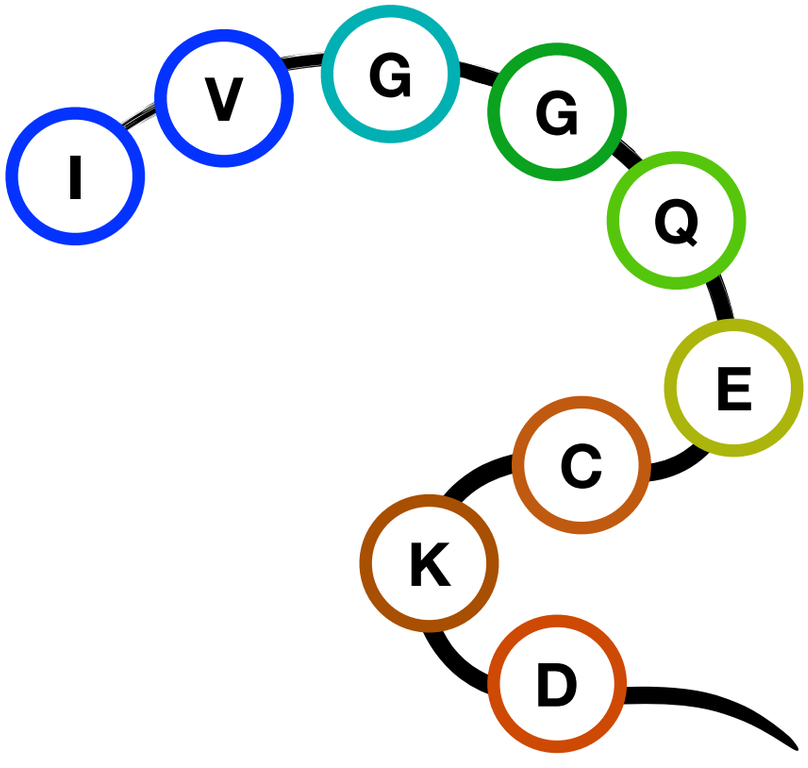
shape and function



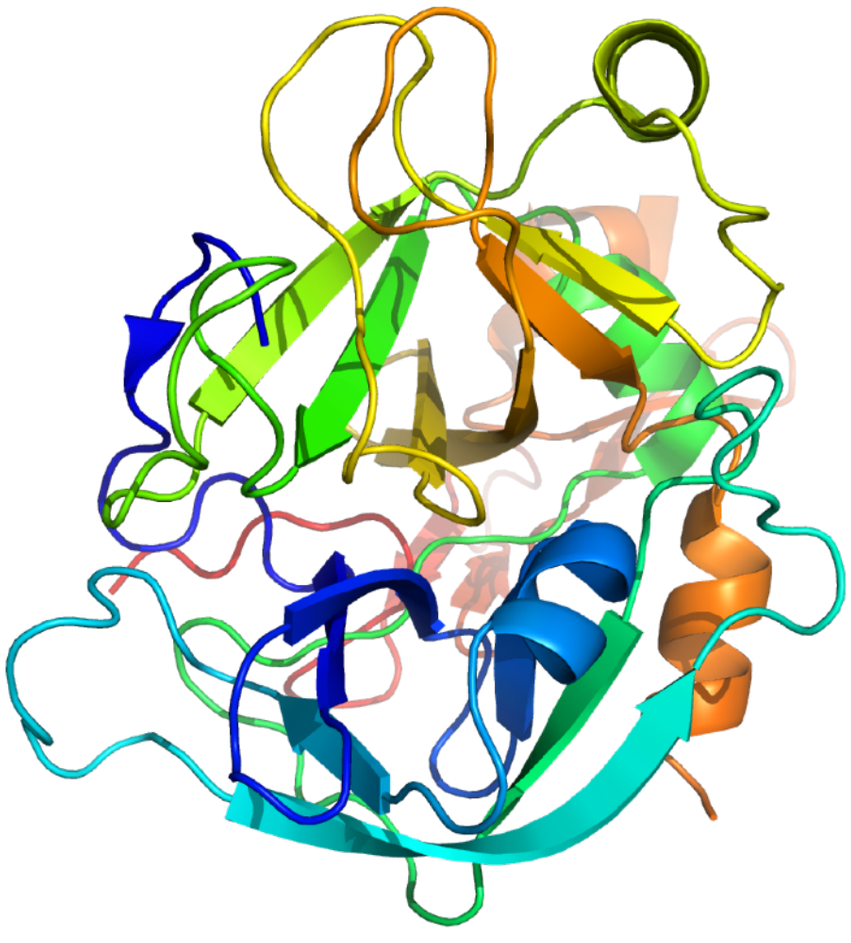
What is protein design?

Invert **Anfinsen's hypothesis:**

protein sequence



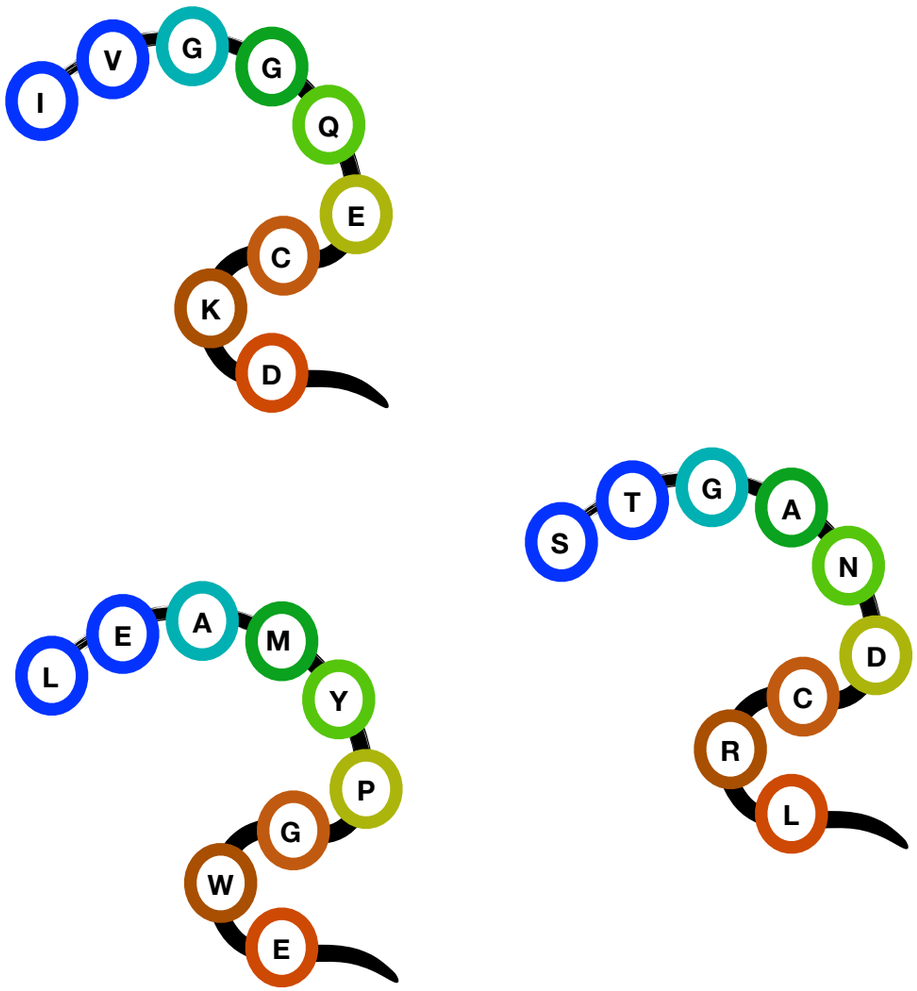
shape and function



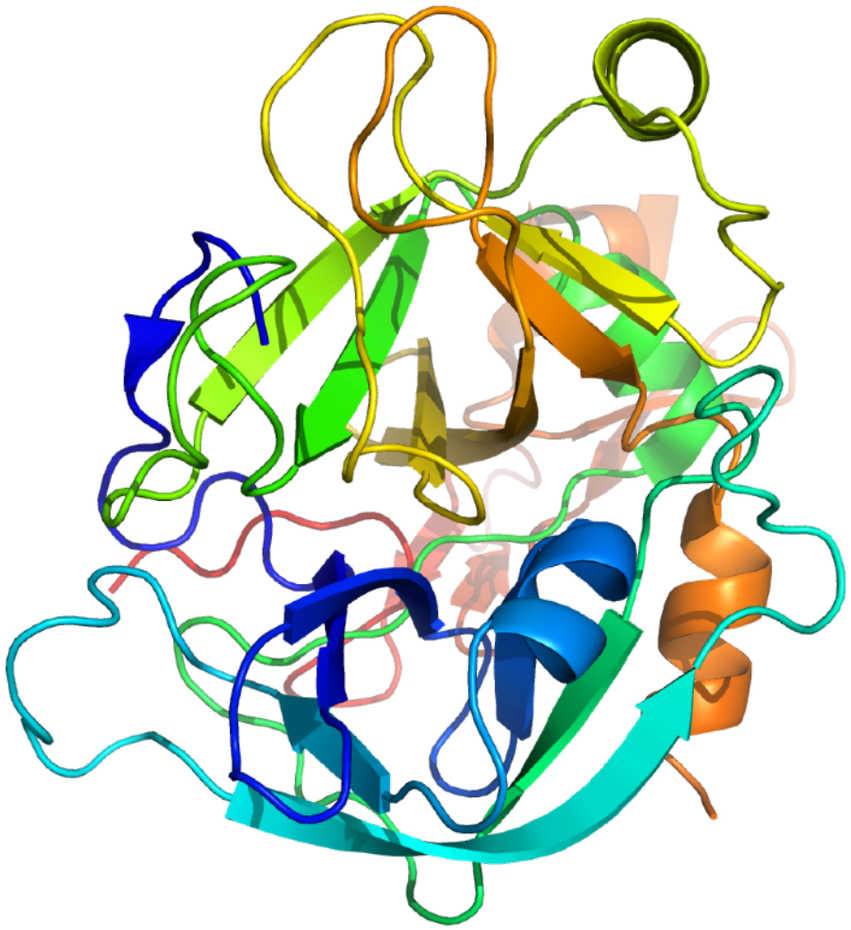
What is protein design?

Invert **Anfinsen's hypothesis:**

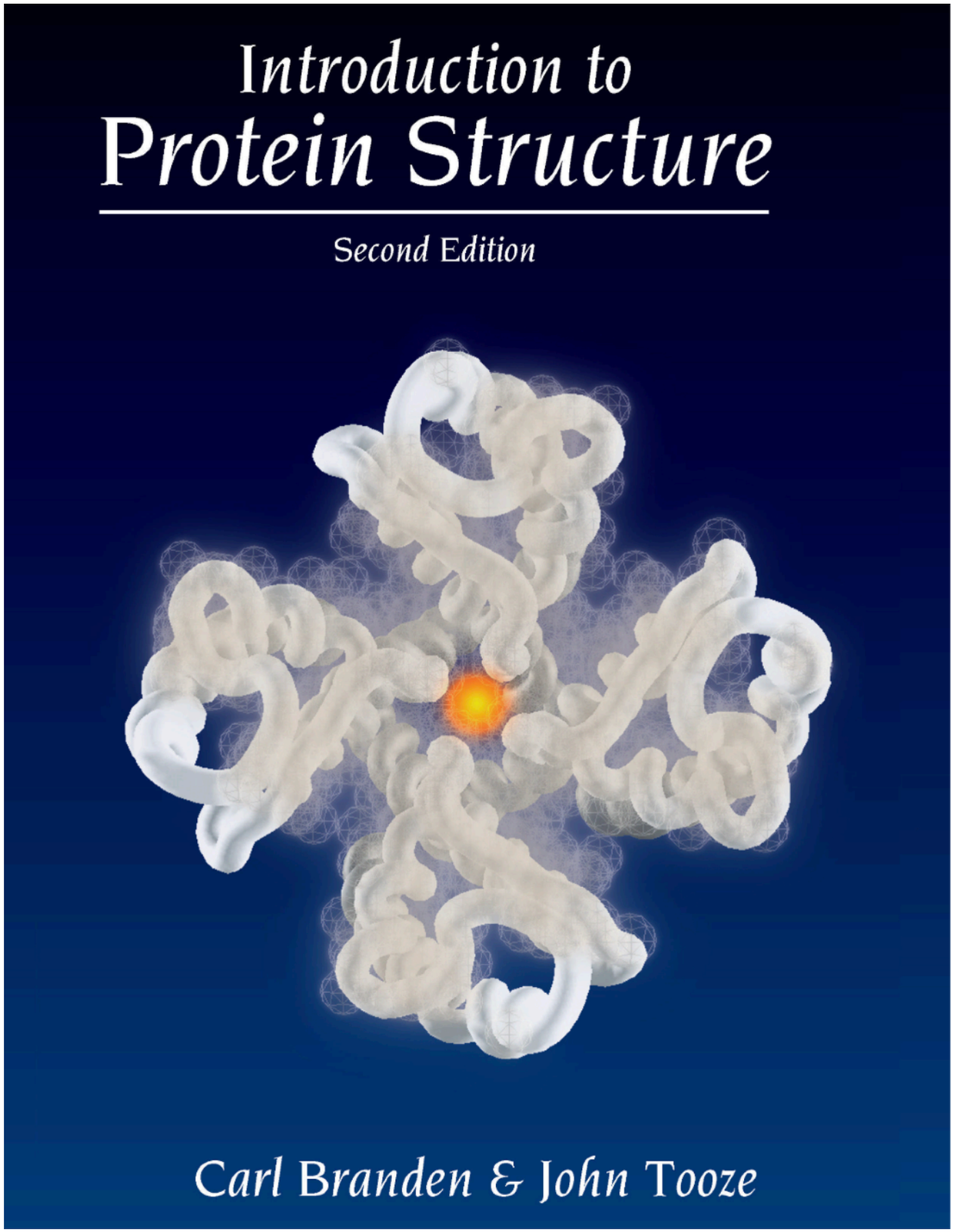
protein sequence(s)

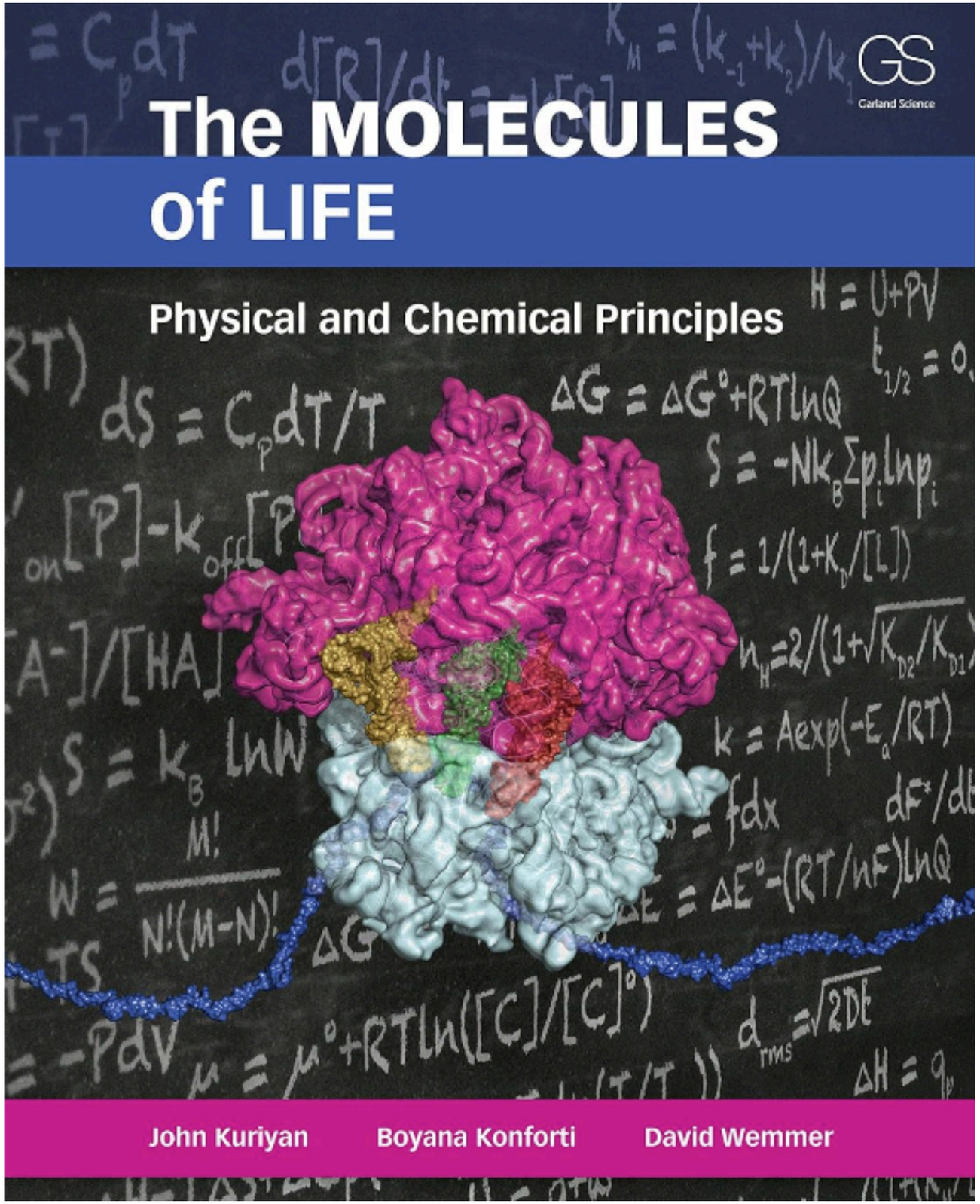


shape and function



Textbook resources

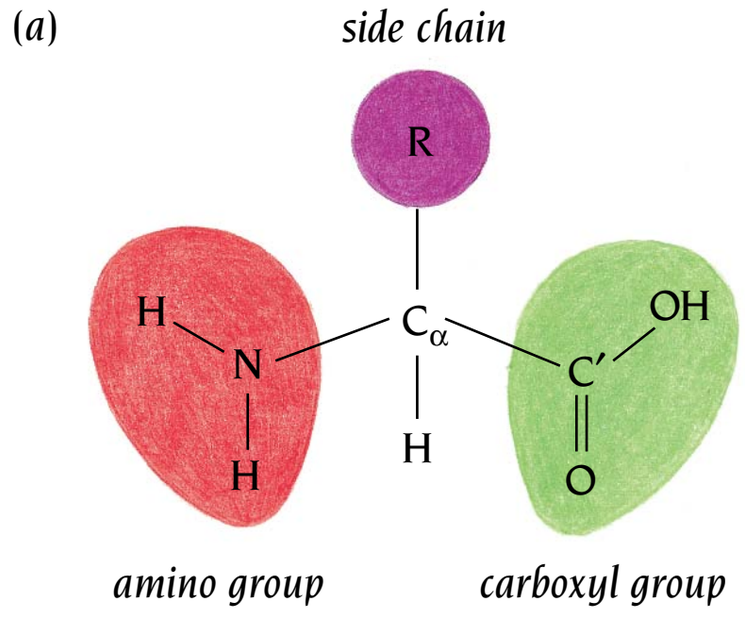


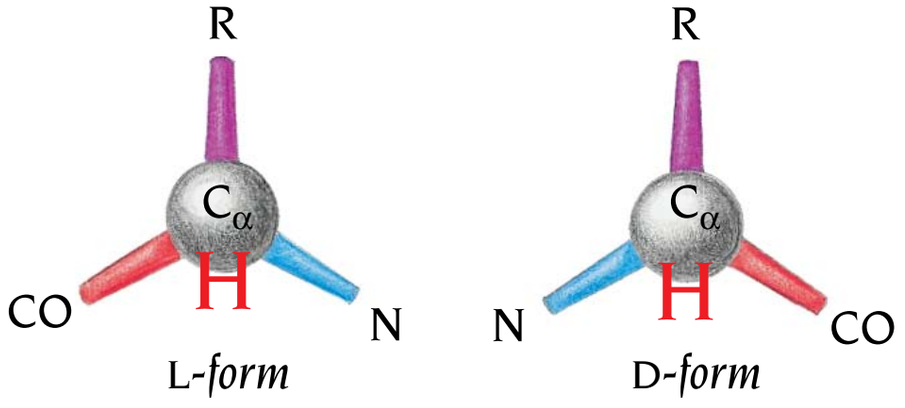


The protein data bank is a protein-structure treasure trove

rcsb.org

The polypeptide “backbone”





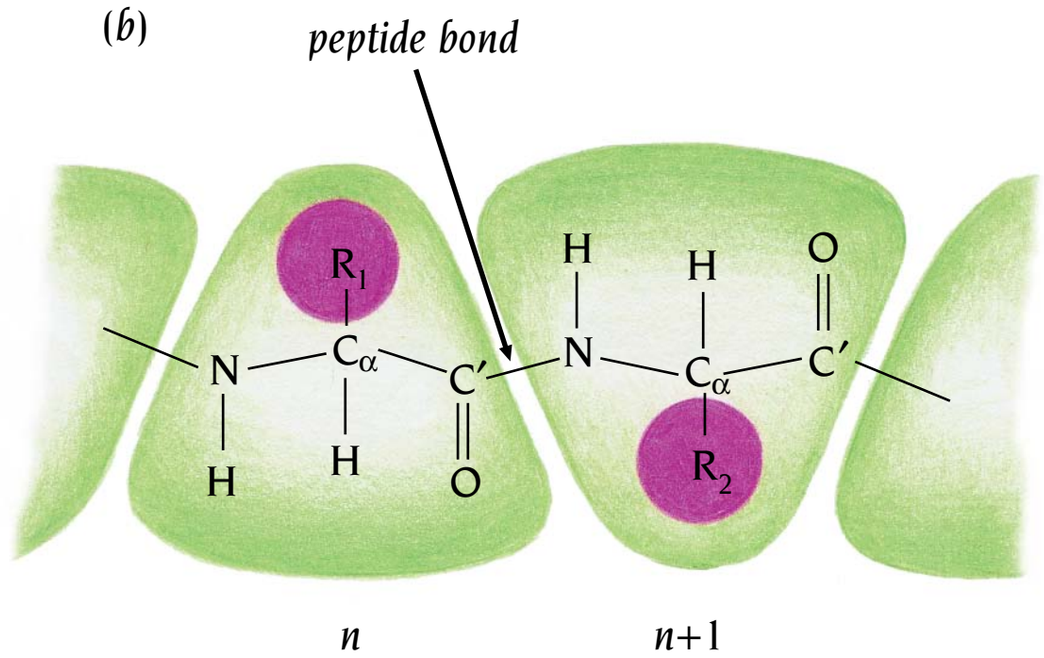
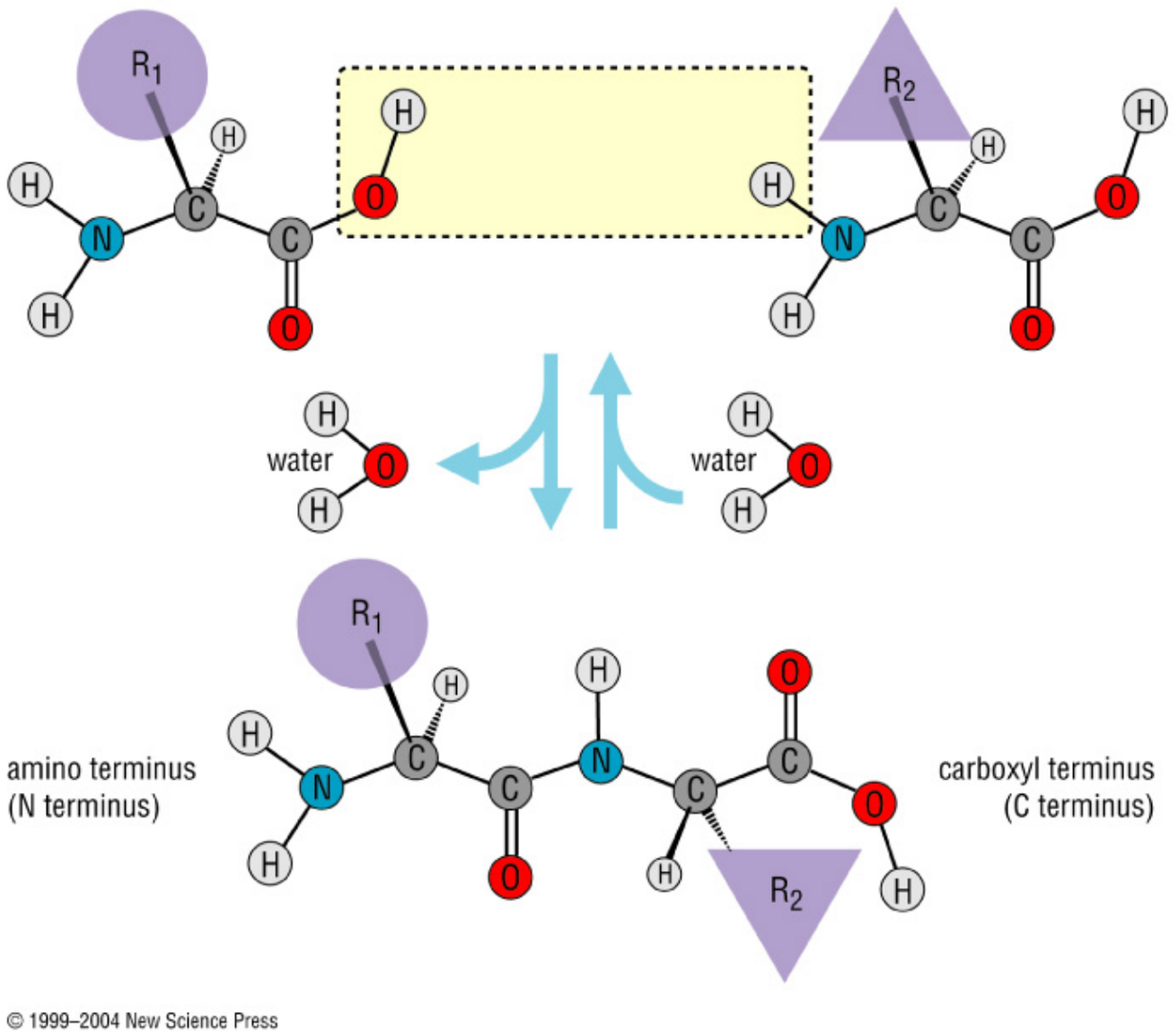


Figure 1.3 The “handedness” of amino acids. Looking down the H–C_α bond from the hydrogen atom, the L-form has CO, R, and N substituents from C_α going in a clockwise direction. There is a mnemonic to remember this; for the L-form the groups read CORN in clockwise direction.

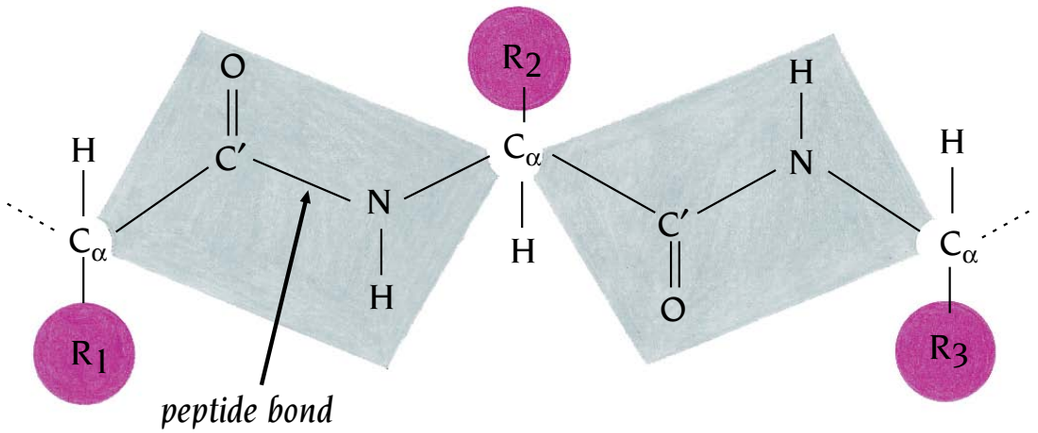
The polypeptide "backbone"

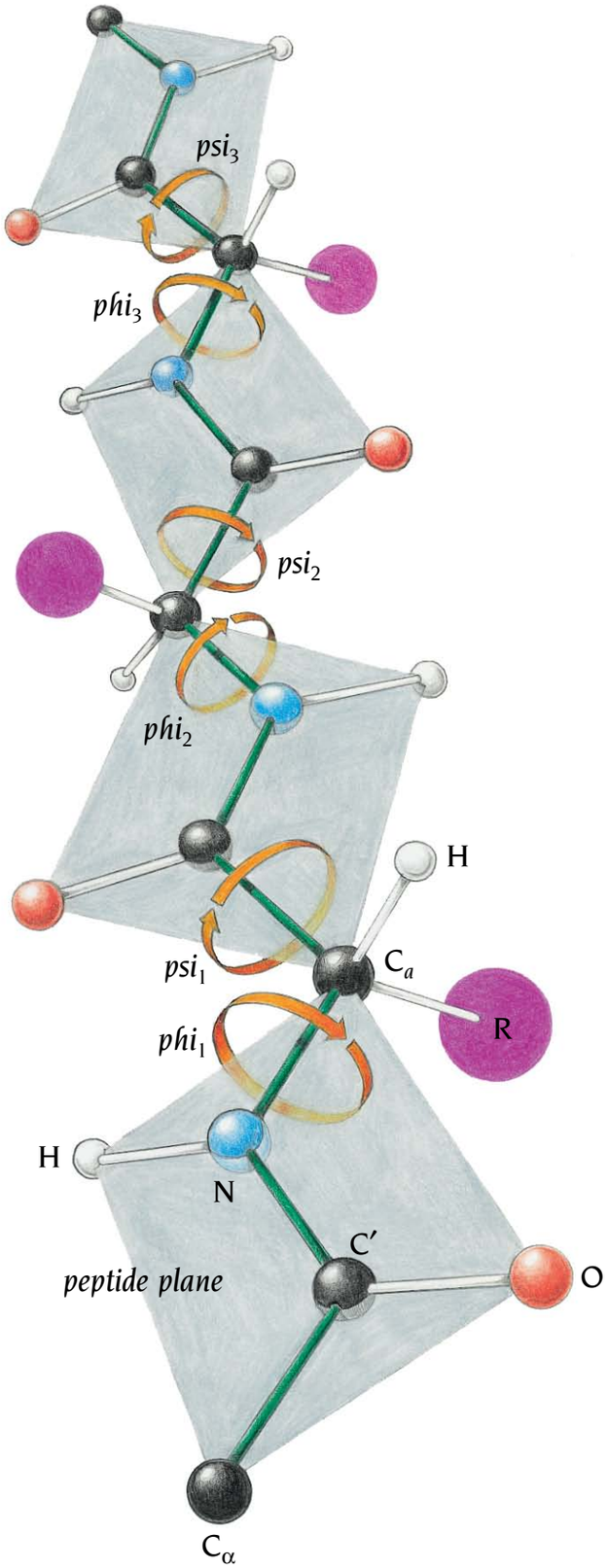
Amino Acids and the Peptide Bond



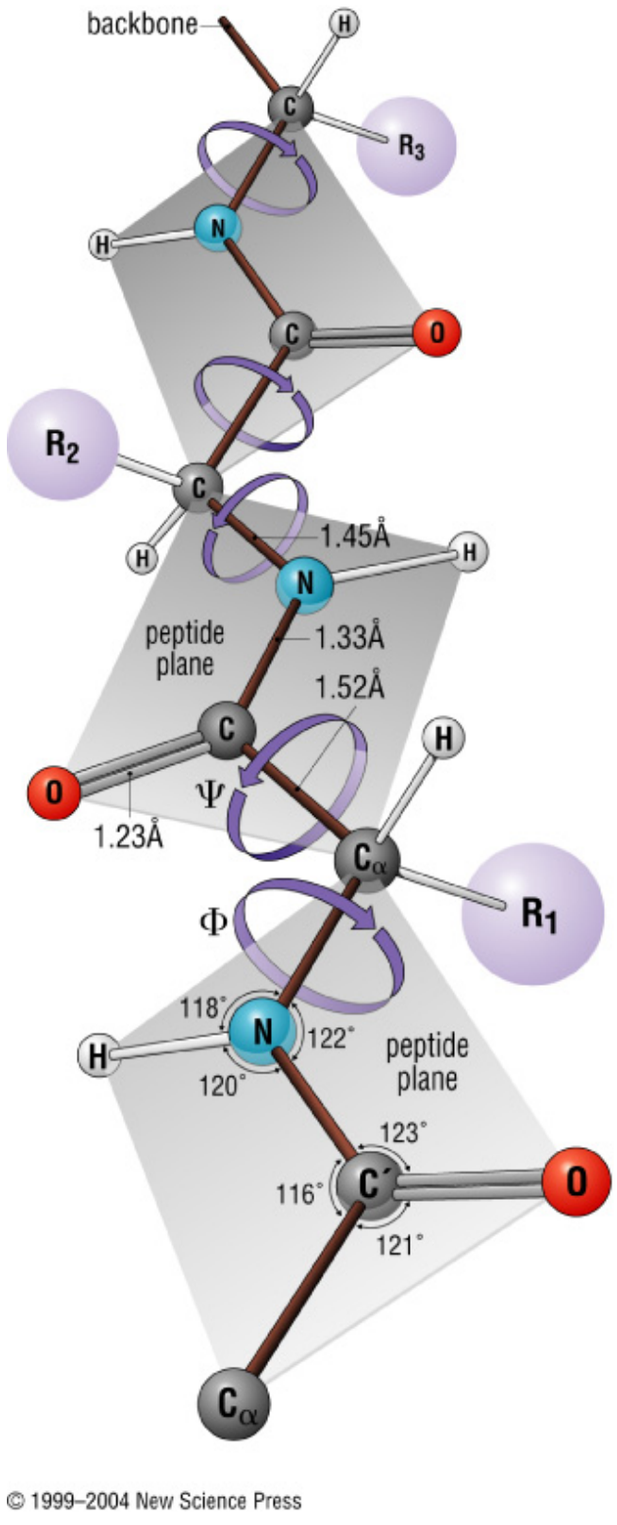
Nomenclature: Cα, amide nitrogen, carbonyl carbon, side chain

The polypeptide “backbone”





The polypeptide “backbone”



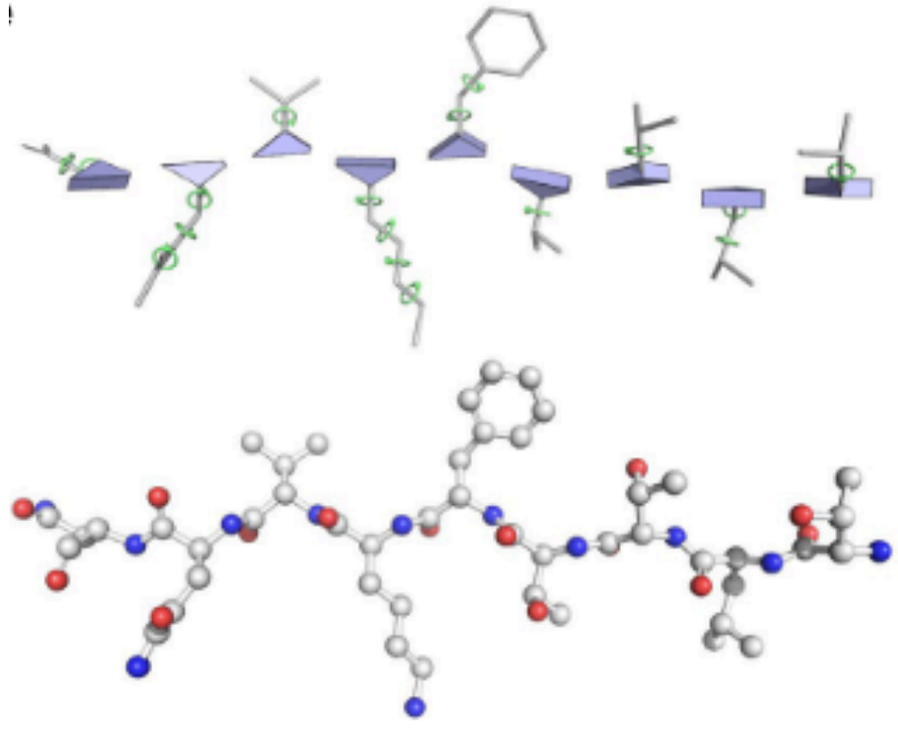
Protein backbone conformation defined by phi and psi angles (rotations between peptide planes)

Side chain torsion angles
chi_1, chi_2, etc.

Bond angles, lengths are essentially fixed (to a first approximation)

C_α- C_α distance is ~3.8 Å

Not quite the same as the “residue gas” from AlphaFold 2



J. Jumper et al, Highly accurate protein structure prediction with AlphaFold. *Nature*. **596**, 583–589 (2021).

The polypeptide “backbone”

Proteins fold into a preferred three-dimensional conformation

- N, C_α and C_O of the coupled amino acid building blocks form the polypeptide backbone
- amino acid sidechains extend from it

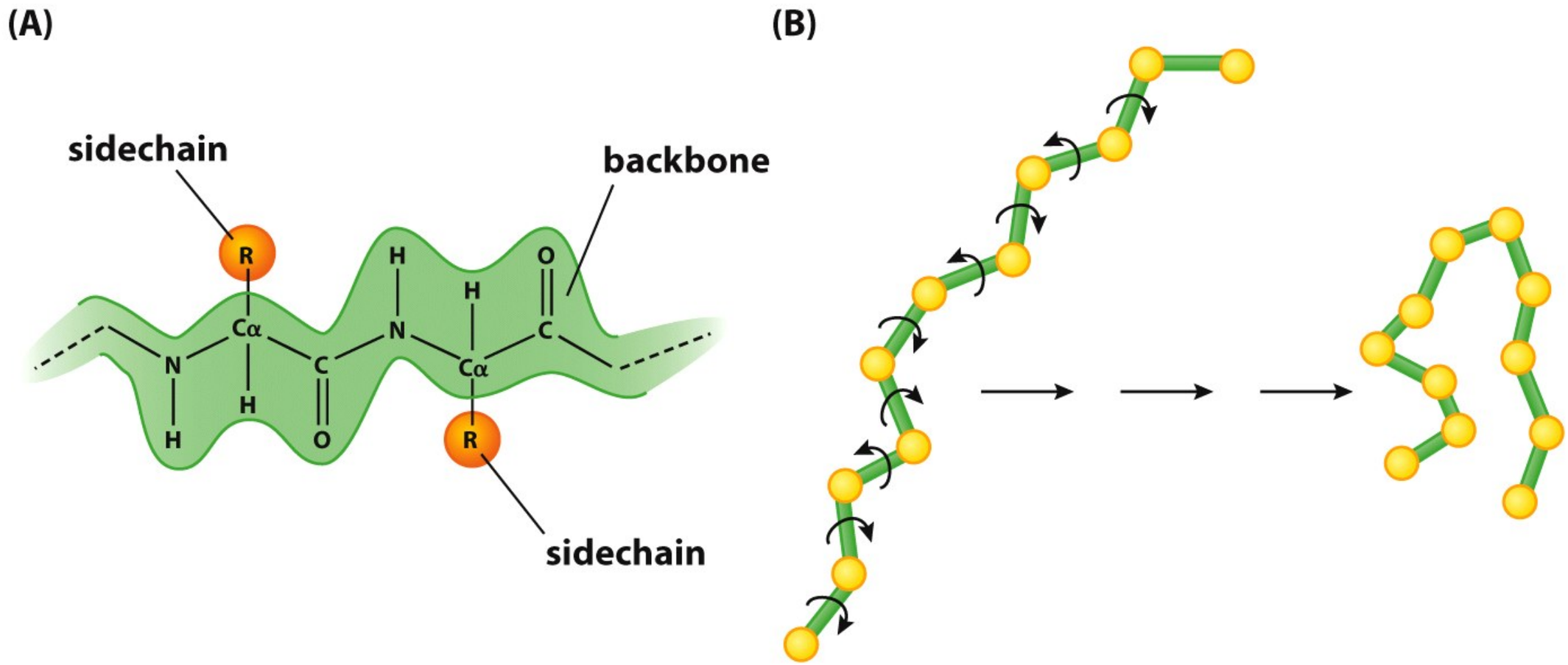
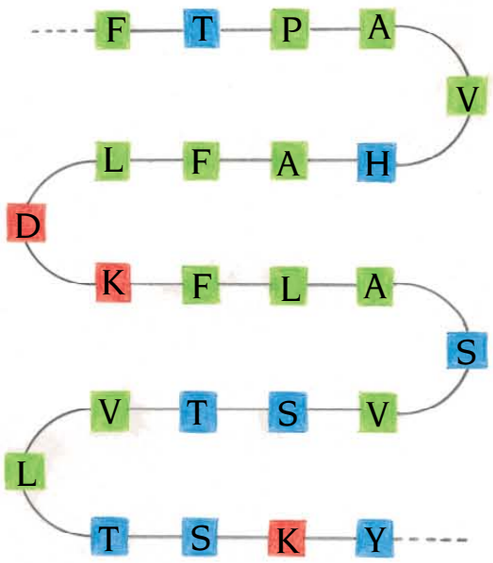


Figure 4.11 The Molecules of Life (© Garland Science 2013)

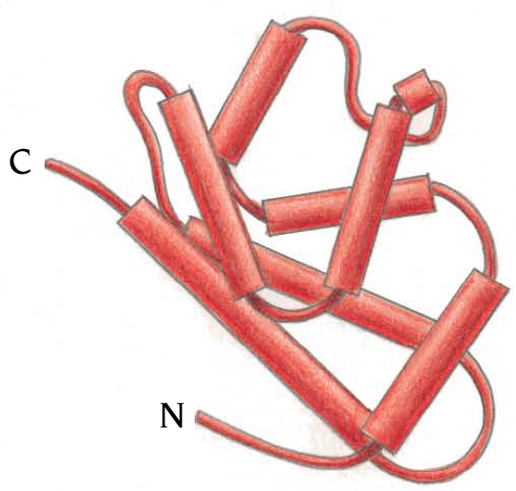
Hierarchy of protein structure

Primary

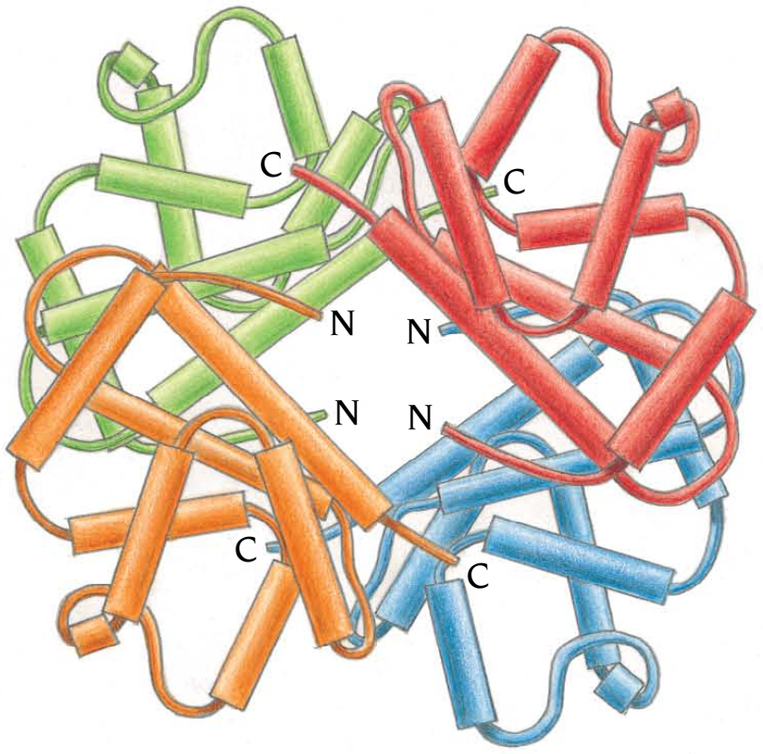


Secondary

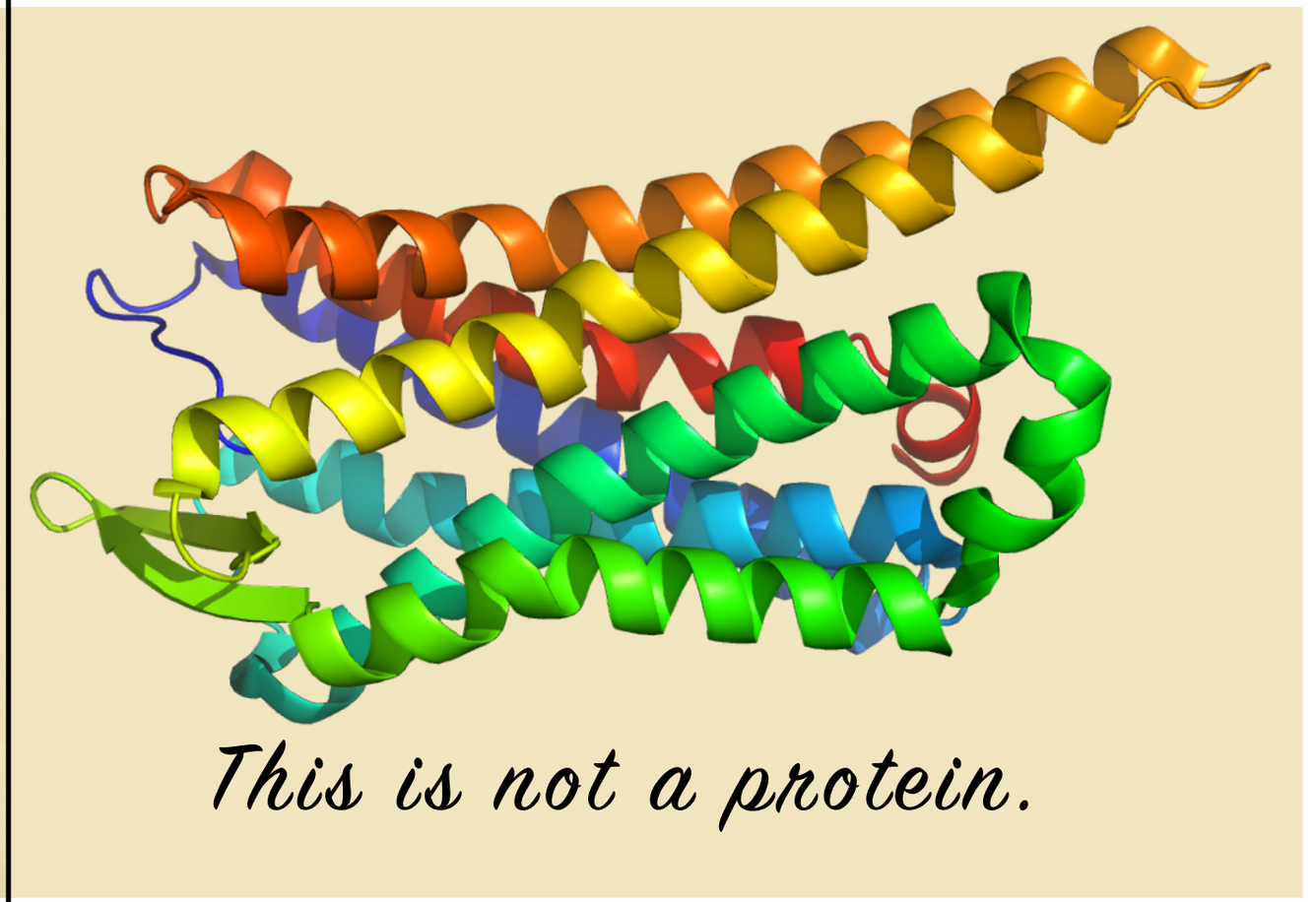
Tertiary



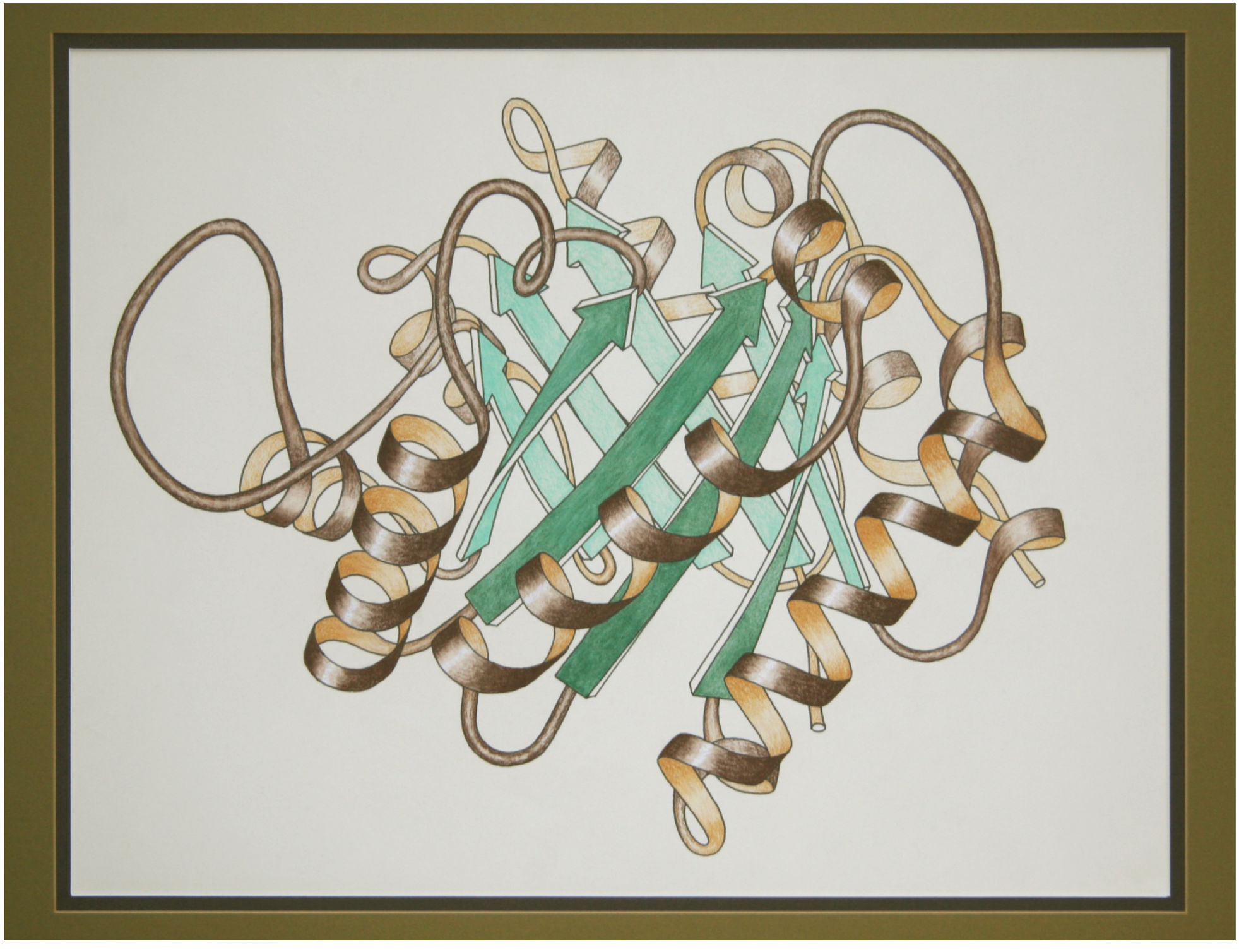
Quarternary



The Treachery of Images
Artist: René Magritte



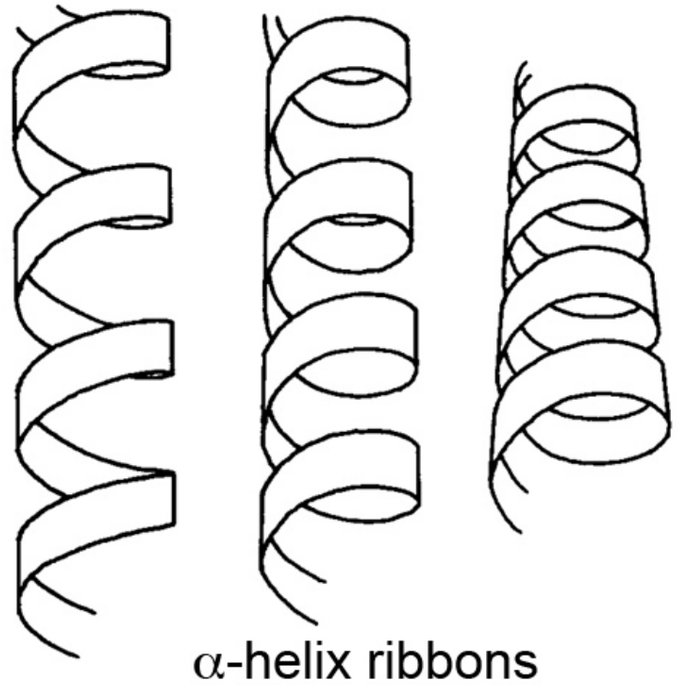
Richardson diagrams of proteins



Ribbon schematic of triose P isomerase monomer (hand-drawn by J. Richardson, 1981) (PDB: 1TIM)

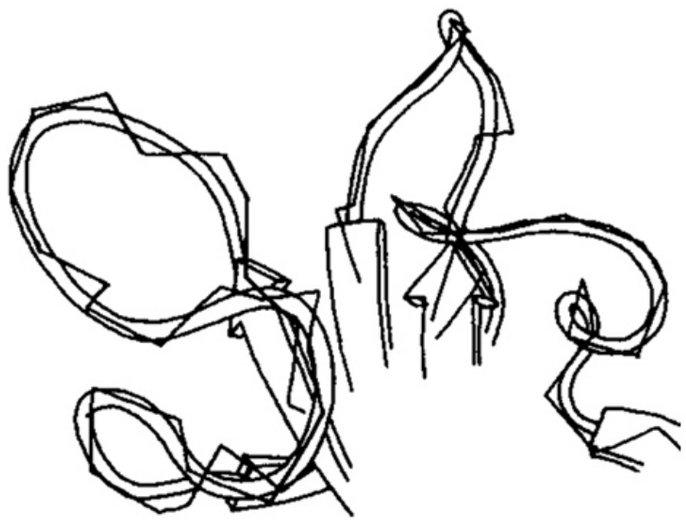
https://en.wikipedia.org/wiki/Ribbon_diagram

Richardson diagrams of proteins



α-helix ribbons

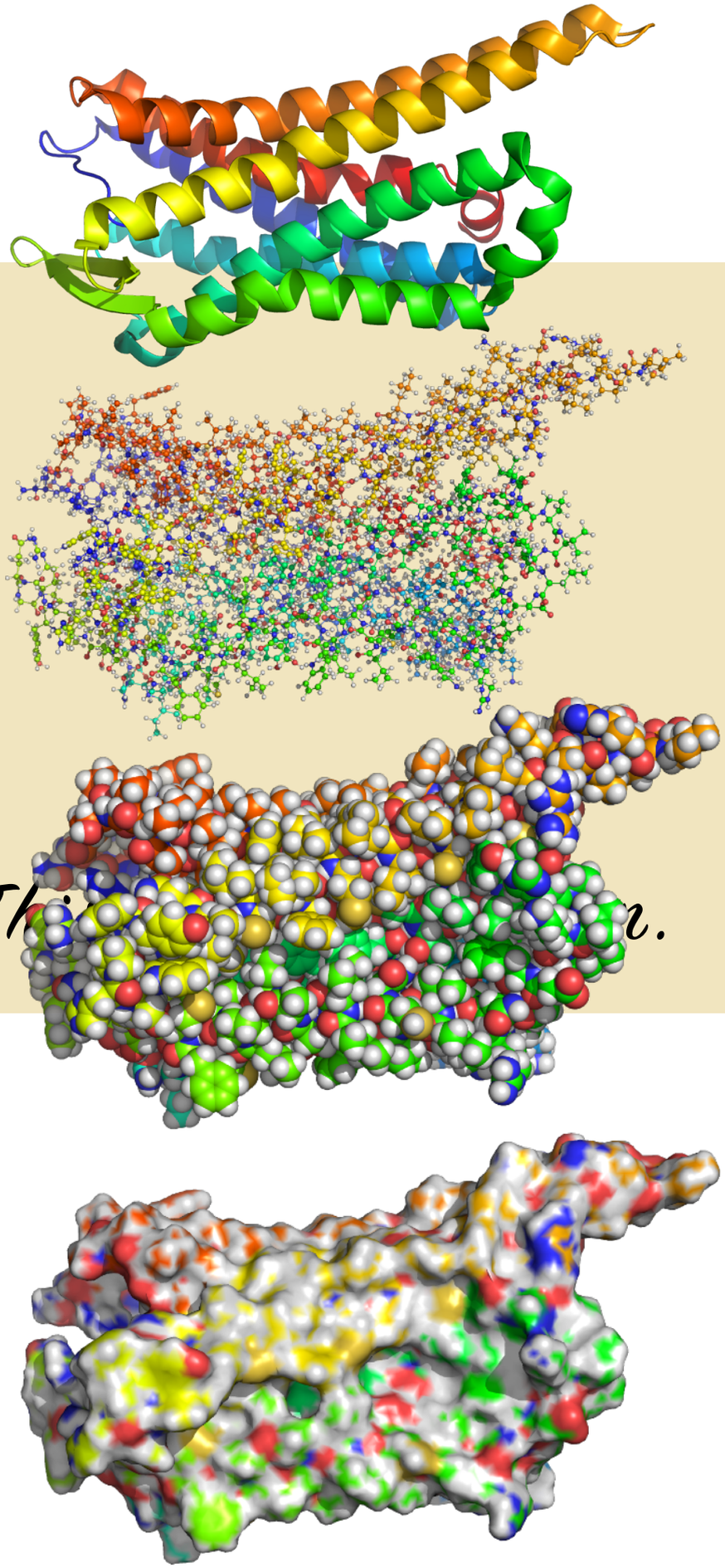
β-strand arrows



smoothed loops

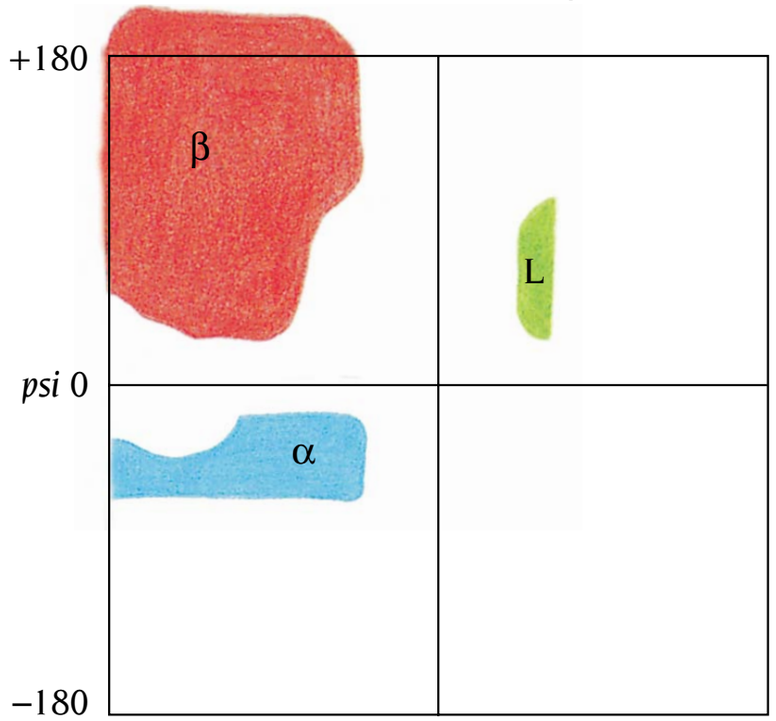
https://en.wikipedia.org/wiki/Ribbon_diagram


Ceci n'est pas une pipe.
Magritte

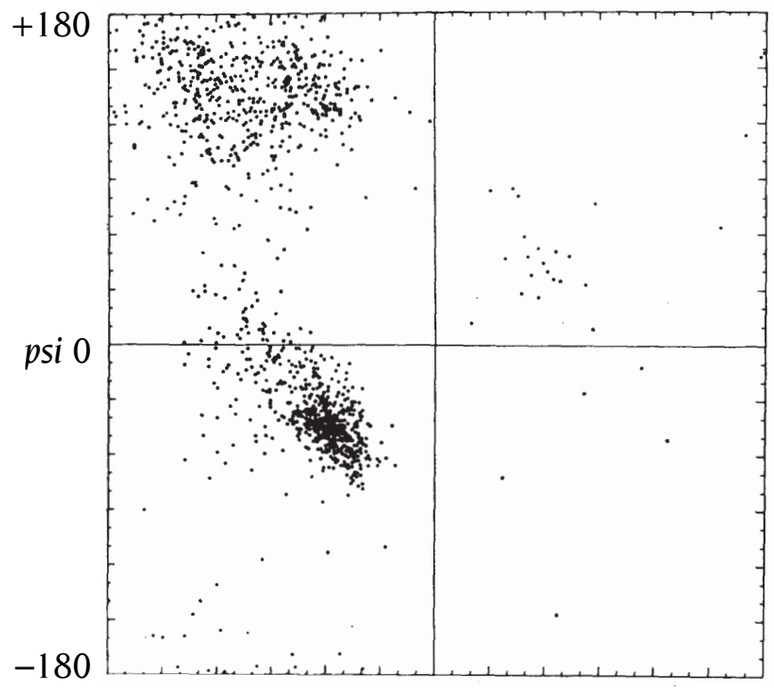

This is not a protein.

Certain backbone dihedral angles are sterically “allowed”

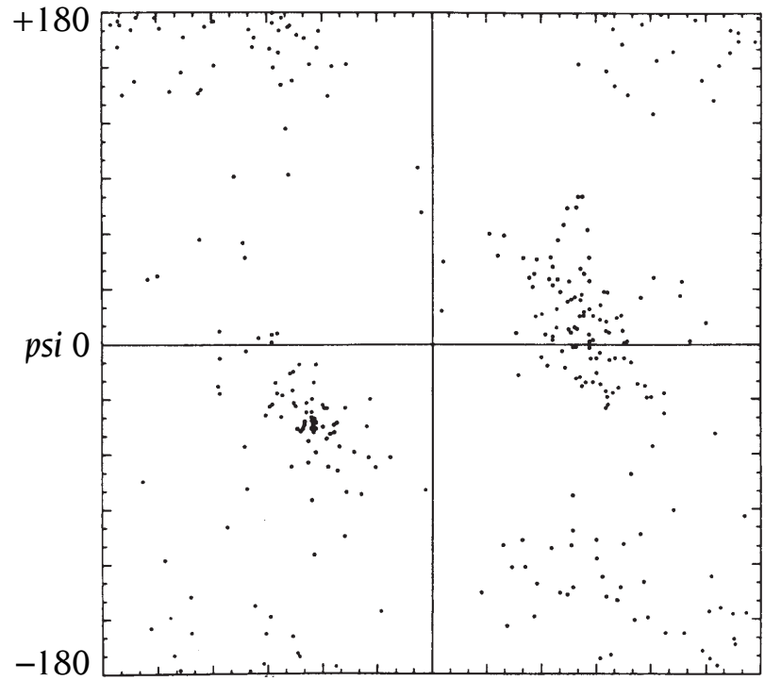
Ramachandran plot

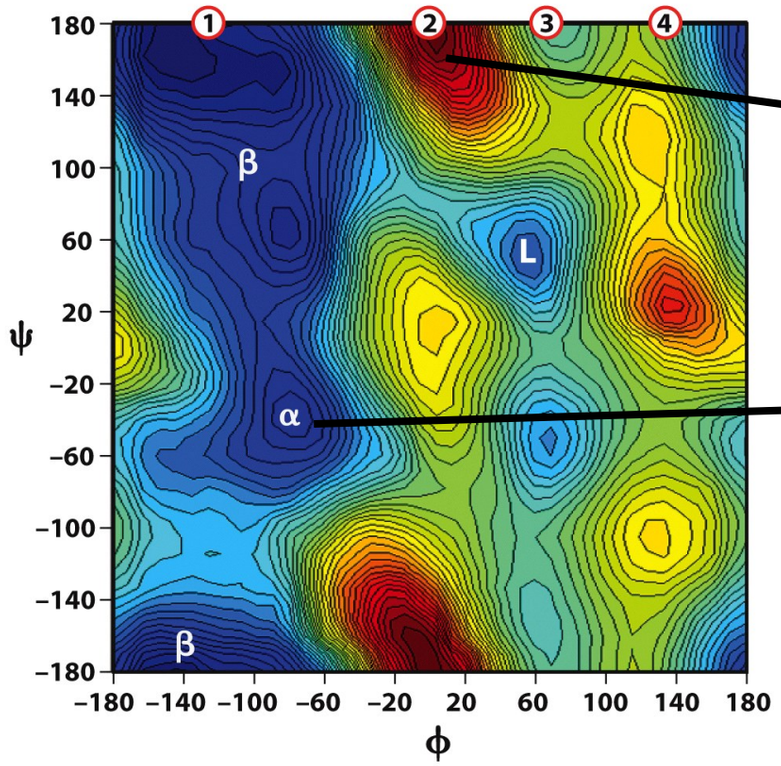


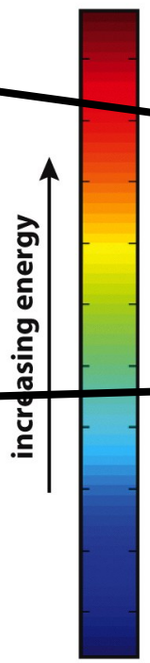
non-GLY residues

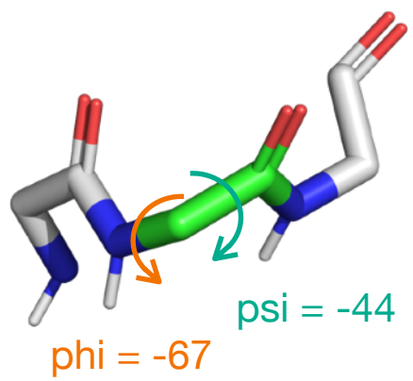


GLY

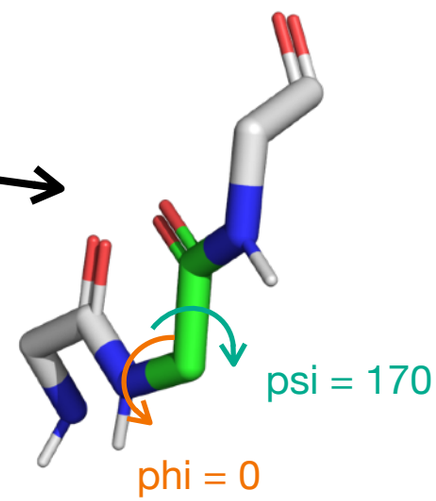








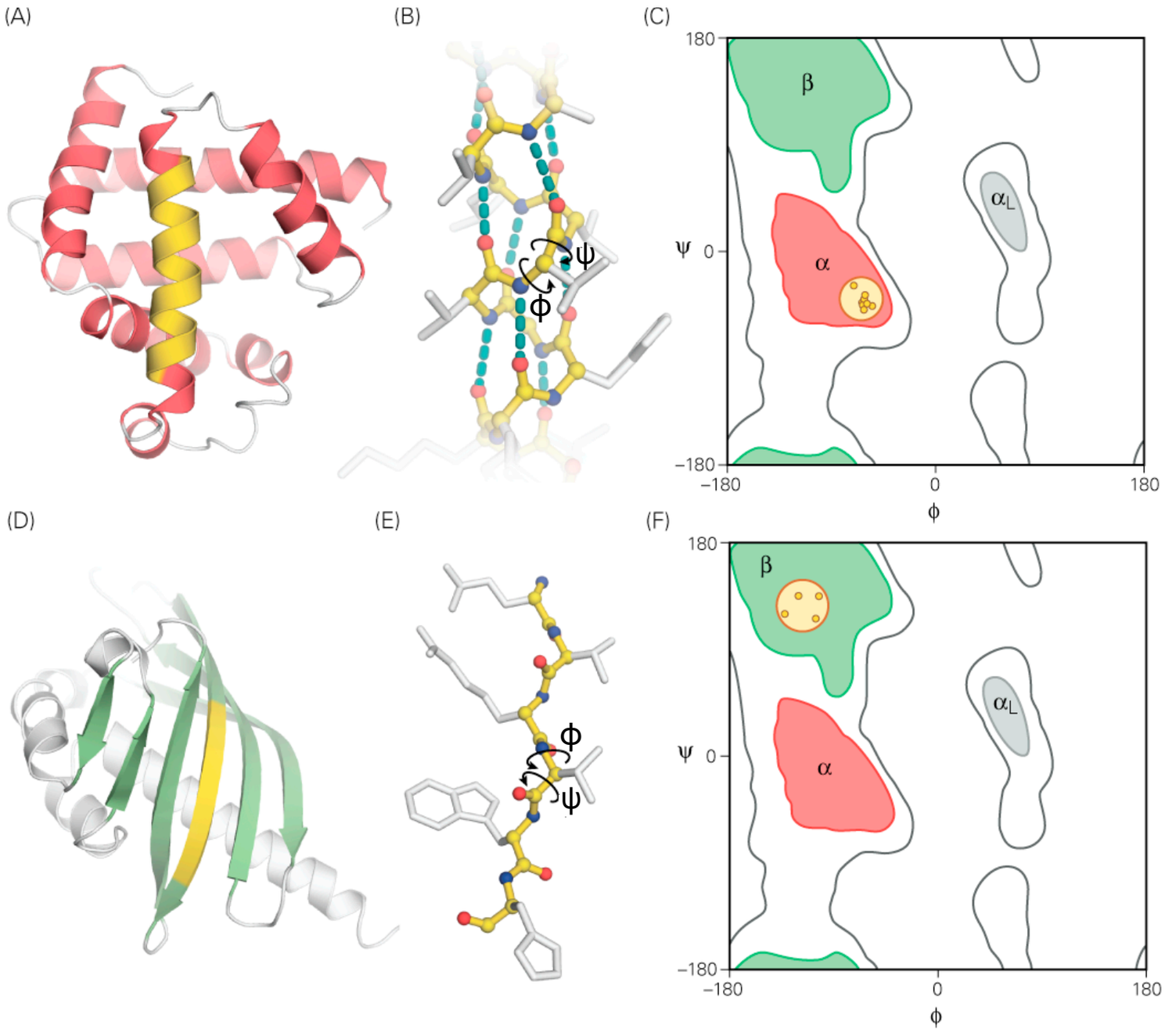
favorable
“allowed”



unfavorable
“disallowed”

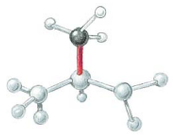
Figure 4.21 The Molecules of Life (© Garland Science 2013)

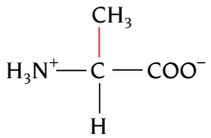
Secondary structure is formed when successive residues lie in a similar region of the Ramachandran plot



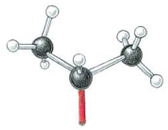
Amino acid sidechains

(a) Hydrophobic amino acids

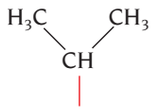


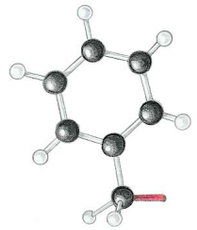


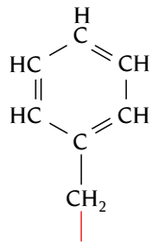
A Ala, Alanine



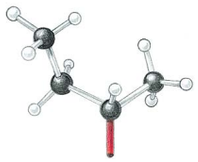
V Val, Valine

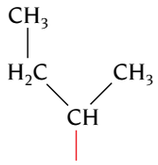






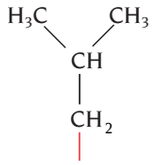
F Phe, Phenylalanine

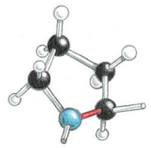


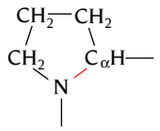


I Ile, Isoleucine

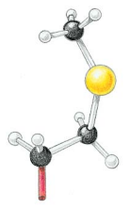
L Leu, Leucine

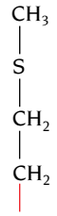






P Pro, Proline



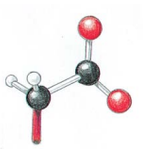


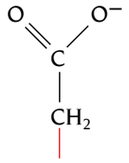
M Met, Methionine

(d) Glycine

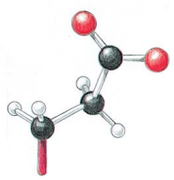
G Gly, Glycine

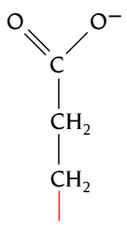
(b) Charged amino acids



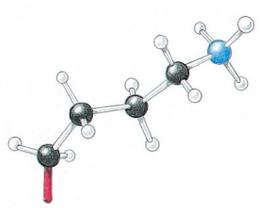


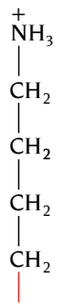
D Asp, Aspartic acid



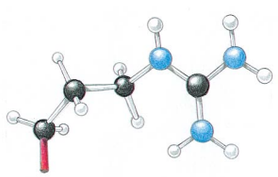


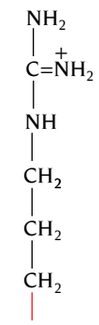
E Glu, Glutamic acid





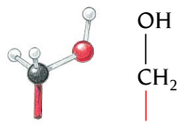
K Lys, Lysine



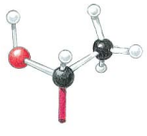


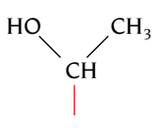
R Arg, Arginine

(c) Polar amino acids

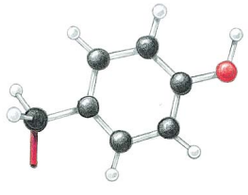


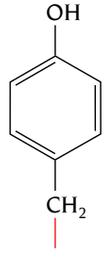
S Ser, Serine



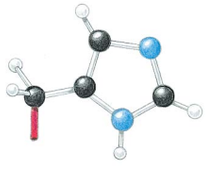


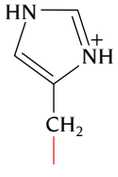
T Thr, Threonine



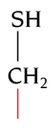


Y Tyr, Tyrosine

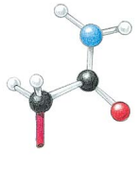


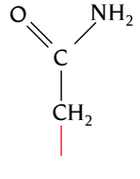


H His, Histidine

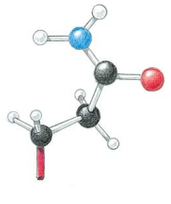


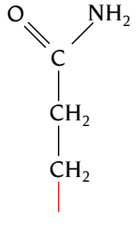
C Cys, Cysteine



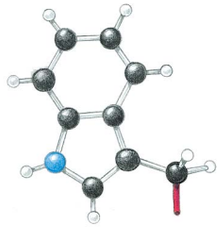


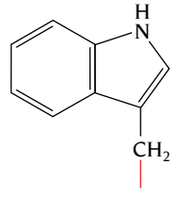
N Asn, Asparagine





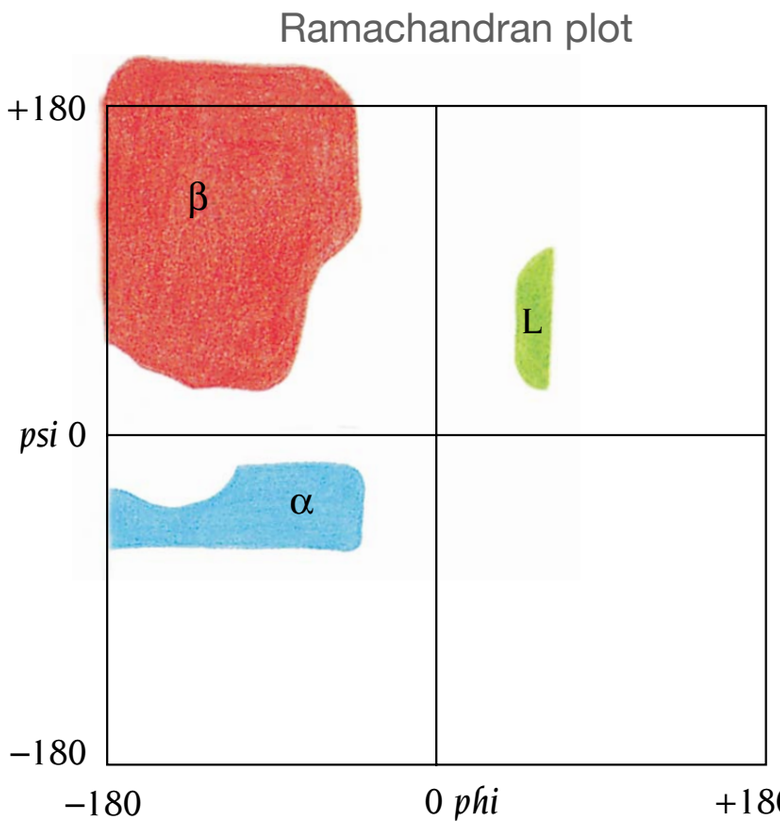
Q Gln, Glutamine

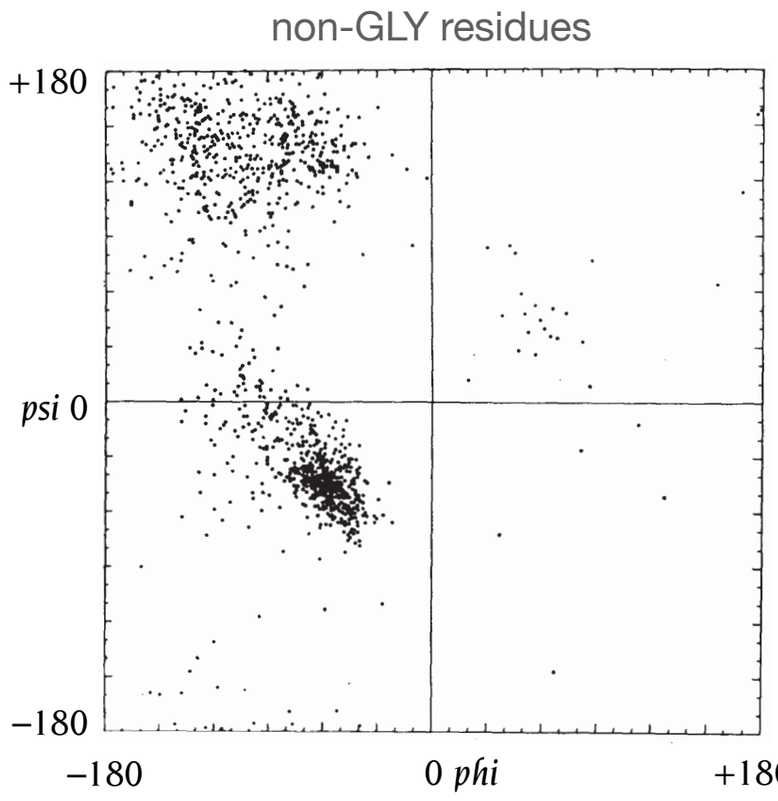


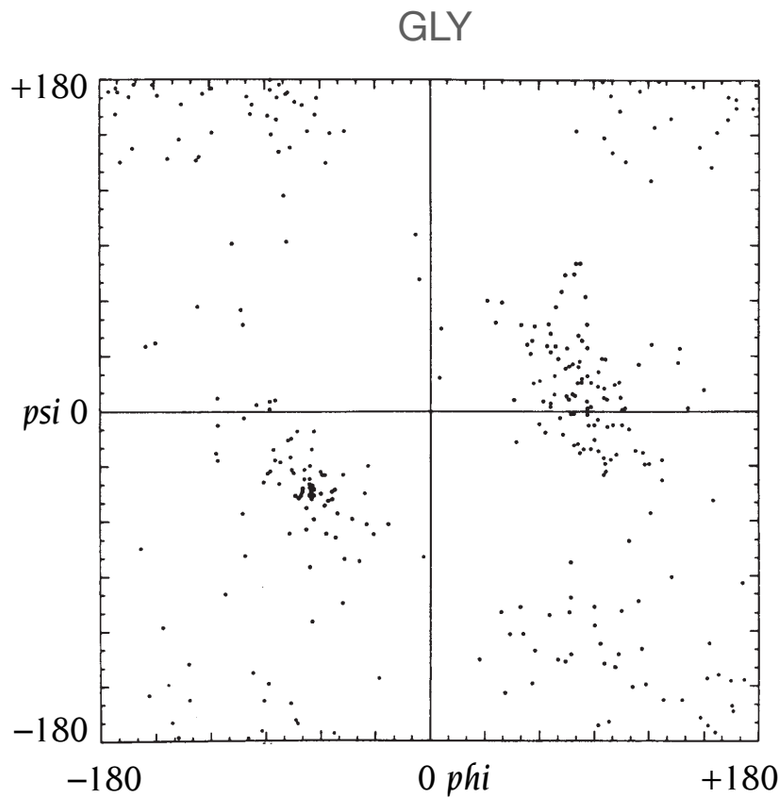


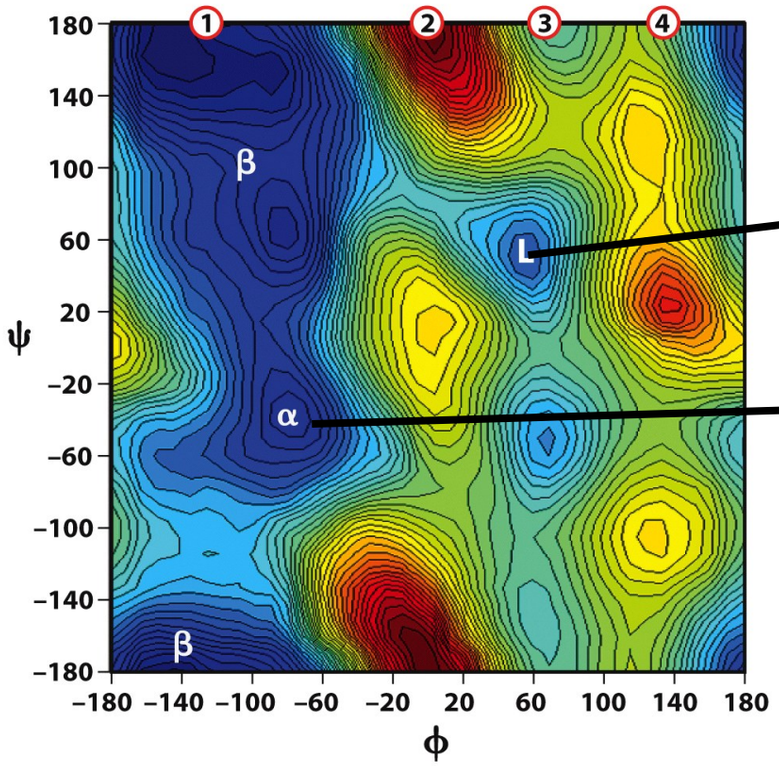
W Trp, Tryptophan

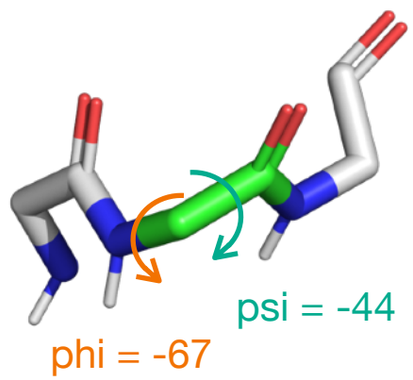
The alpha-left region is most favorable for GLY



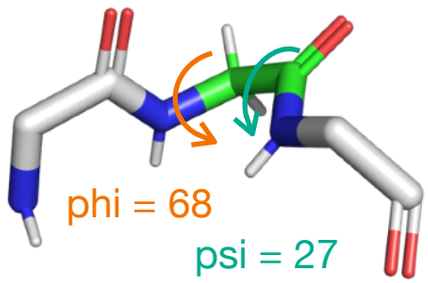




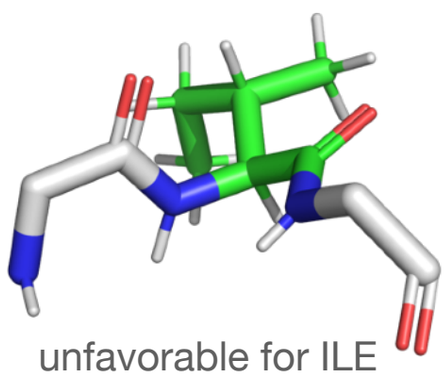




favorable
“allowed”



favorable for GLY



unfavorable for ILE

Figure 4.21 The Molecules of Life (© Garland Science 2013)

Sidechains have preferred low energy conformations

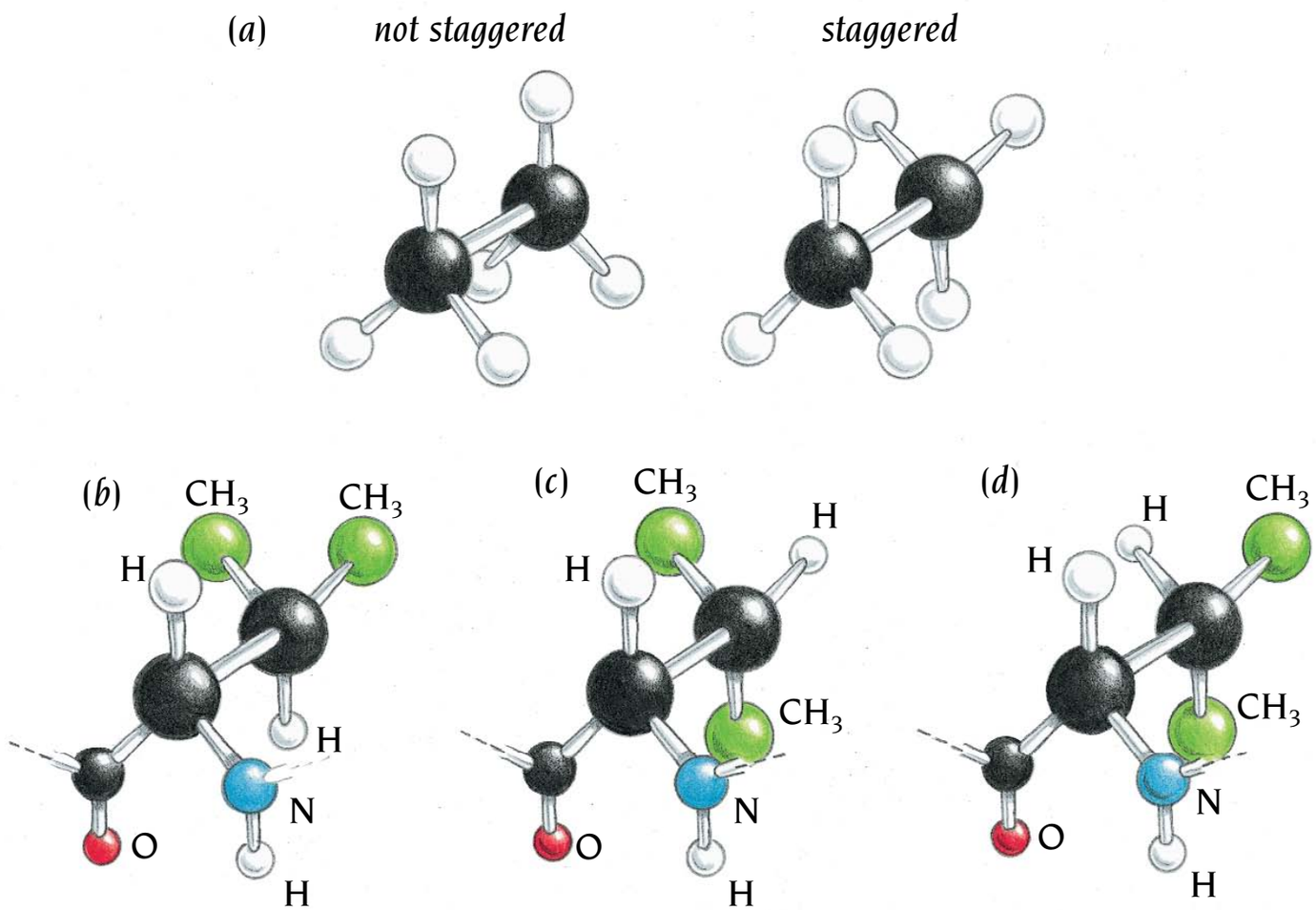
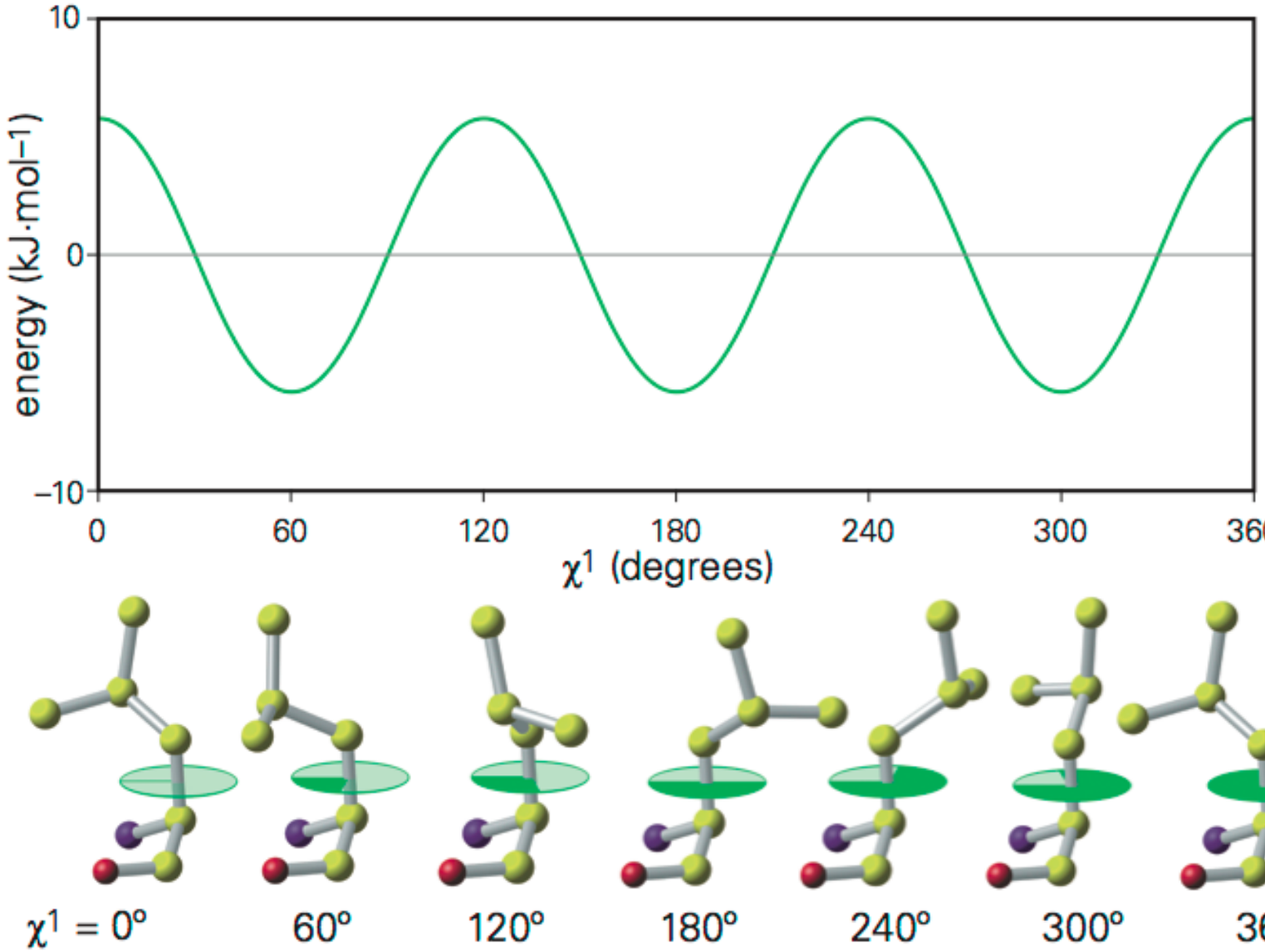
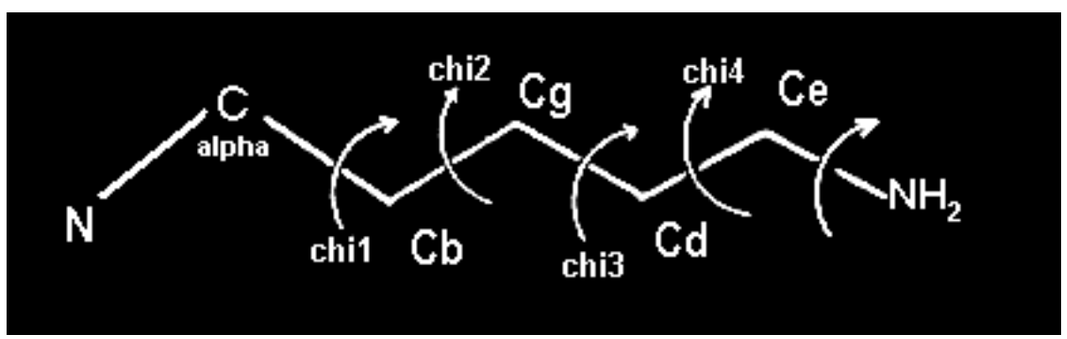


Figure 1.8 The staggered conformations are the most energetically favored conformations of two tetrahedrally coordinated carbon atoms. (a) A view along the C–C bond in ethane (CH_3CH_3) showing how the two carbon atoms can rotate so that their hydrogen atoms are either not staggered (aligned) or staggered.Three indistinguishable staggered conformations are obtained by a rotation of 120° around the C–C bond. (b–d) Similar views as in (a) of valine. The three staggered conformations are different for valine because the three groups attached to C_β are different. The first staggered conformation (b) is less crowded and energetically most favored because the two methyl groups bound to C_β are both close to the small H atom bound to C_α.

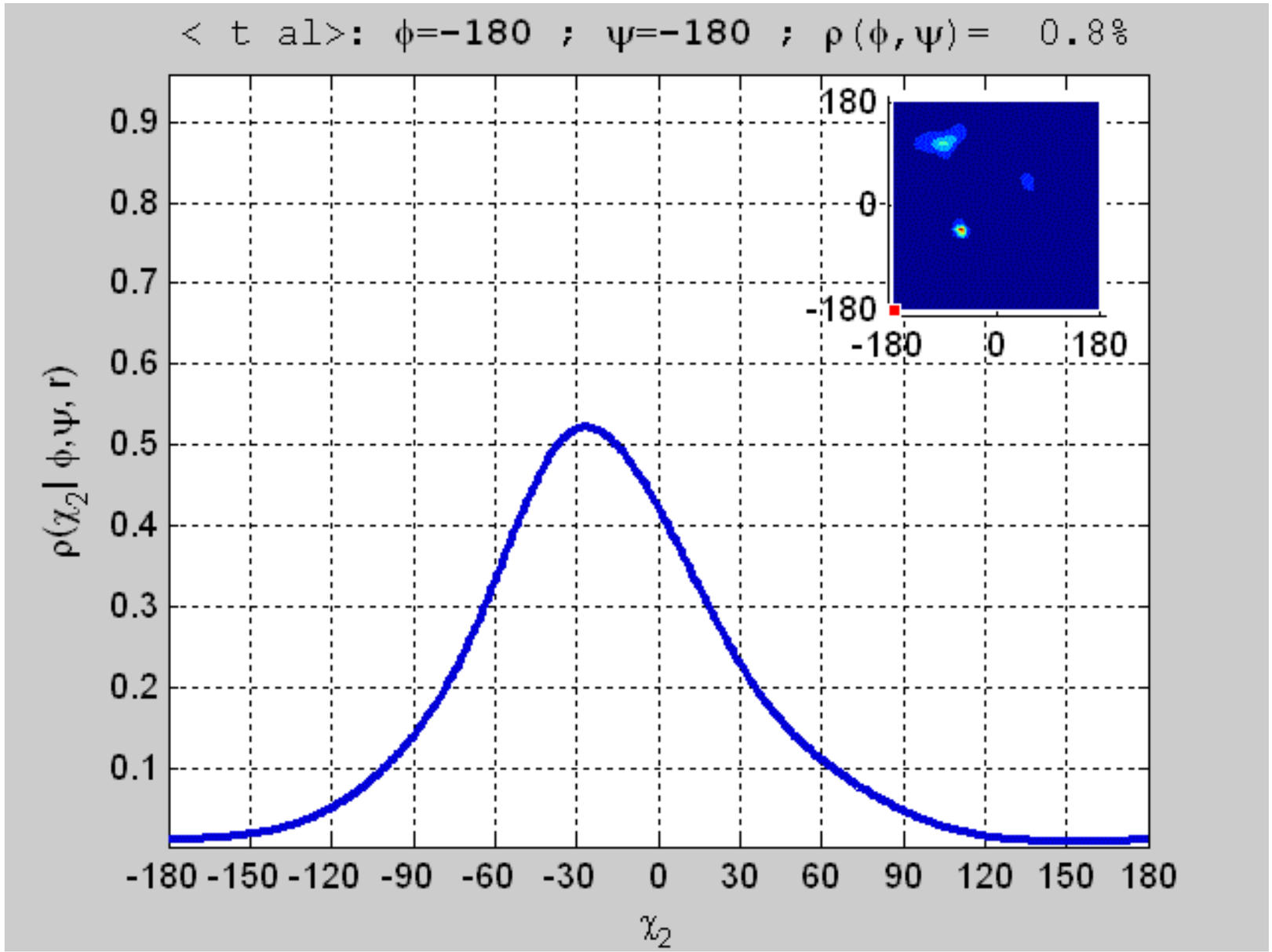
Rotation around side-chain C-C bonds

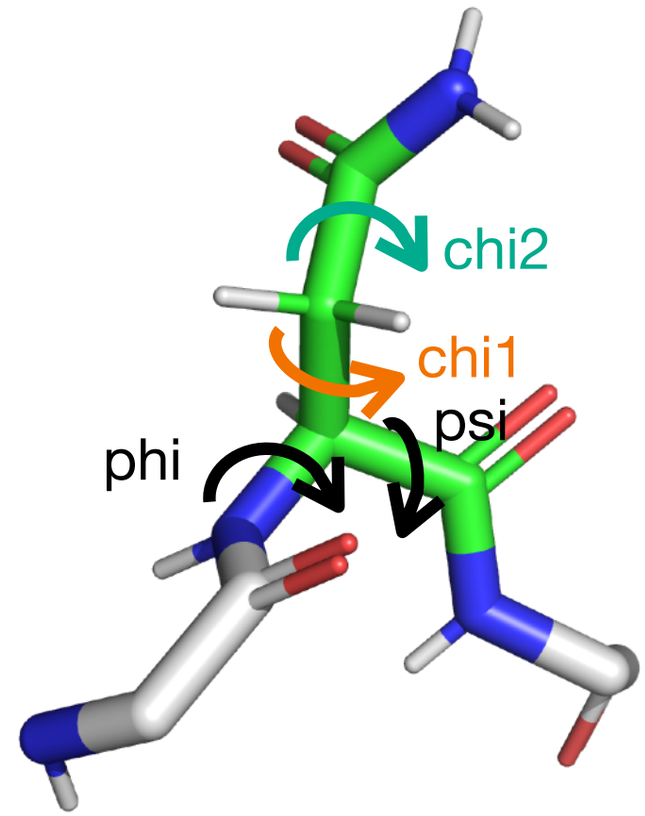




http://dunbrack.fccc.edu/lab/bbdep2010_tutorial

Sidechain conformations can be backbone phi/psi dependent





http://dunbrack.fccc.edu/lab/bbdep2010_tutorial

A new structural unit

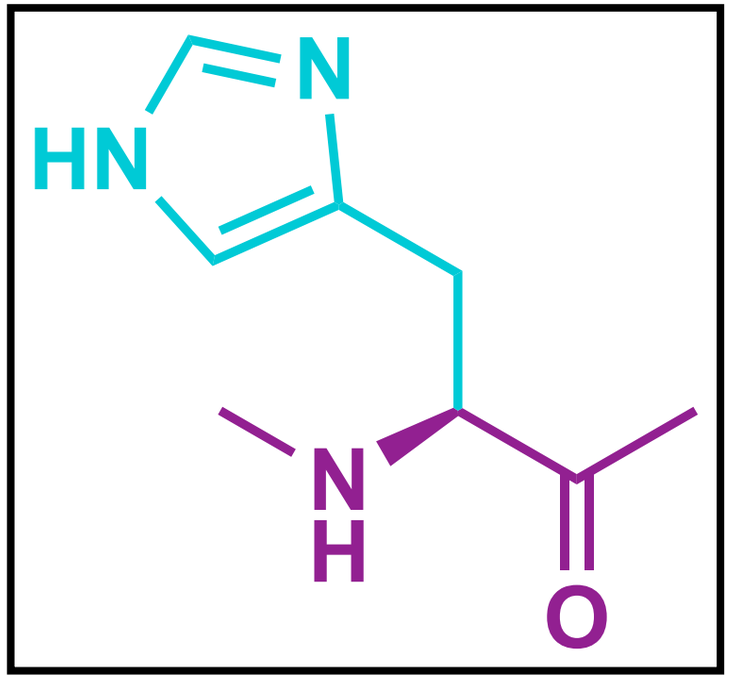
van der Mers map the backbone directly to fragment position

van der Waals contact + rotamer

Basis set for covalent *sidechain* conformations

Rotamer

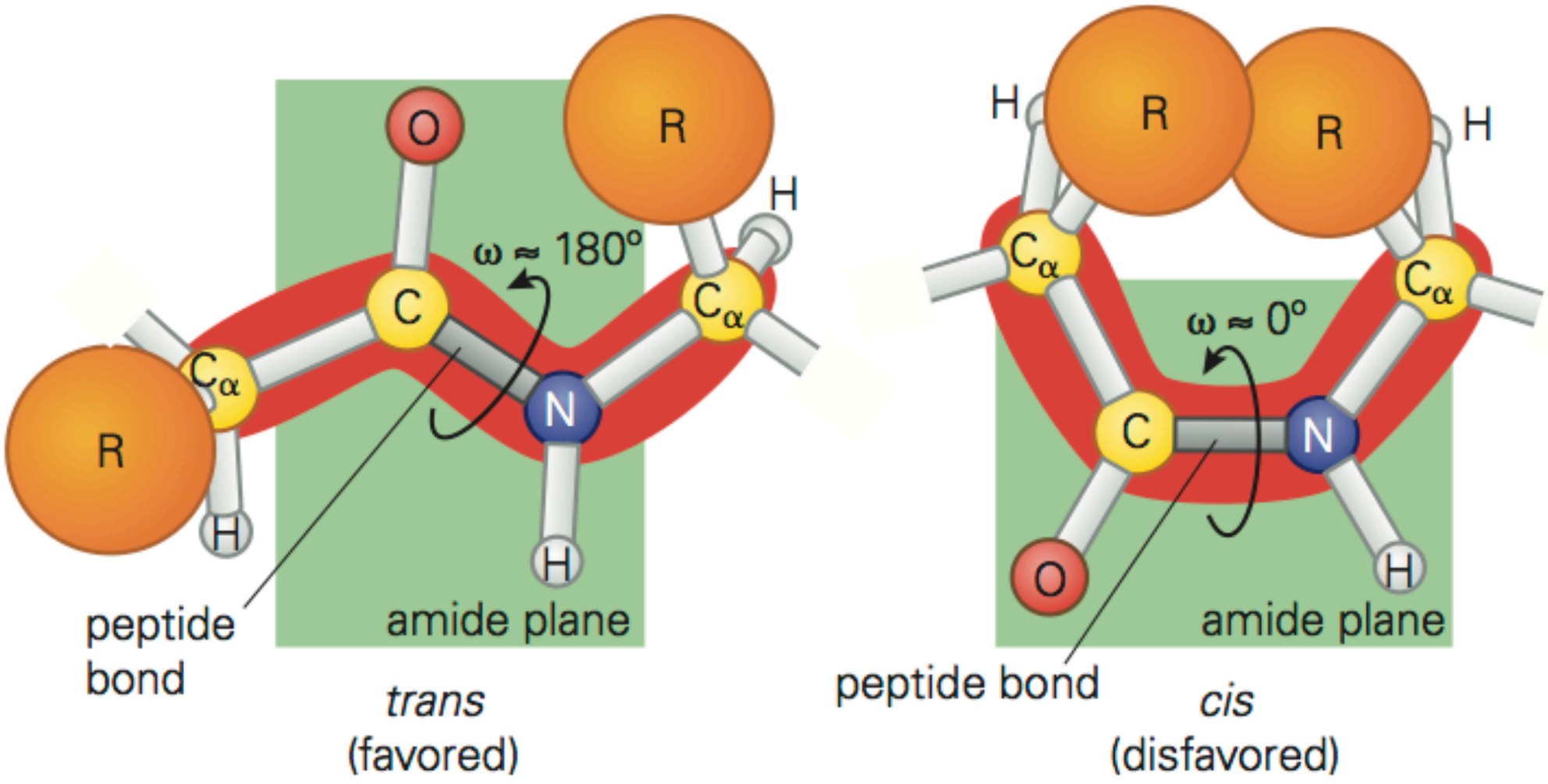
Uses PDB to define placement of **sidechain** relative to **mainchain**



His

Polizzi and DeGrado. *Science.* 2020, 369, 1227-1233

Adjacent amino acids are almost always in the *trans* conformation



Proline can adopt cis conformation

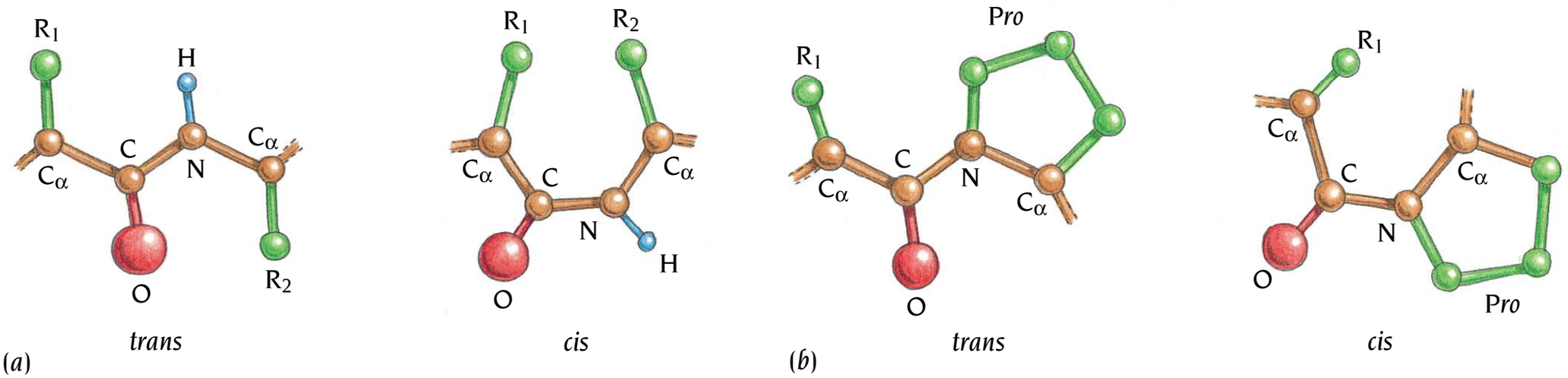
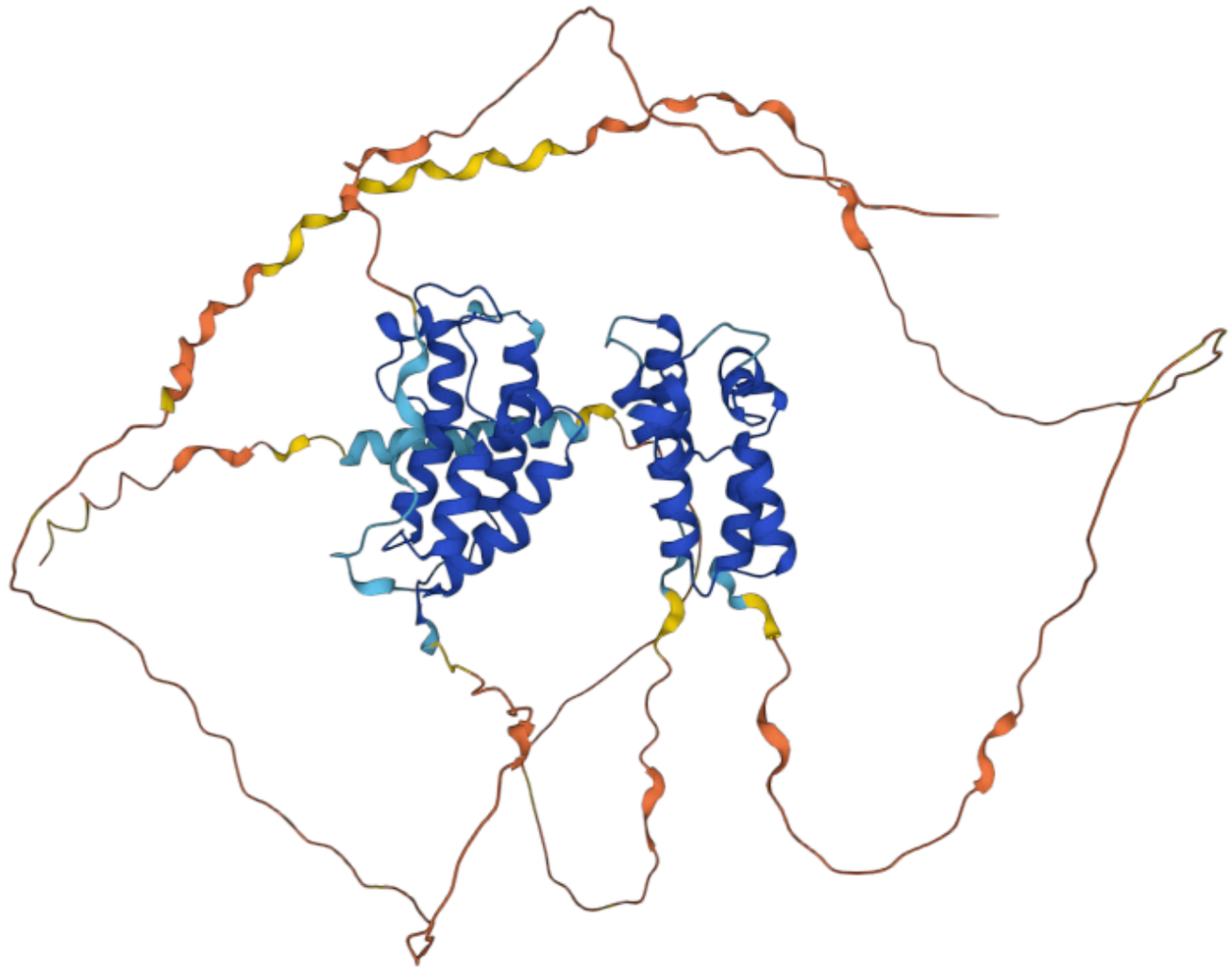


Figure 6.9 (a) Peptide units can adopt two different conformations, *trans* and *cis*. In the *trans*-form the C=O and the N-H groups point in opposite directions whereas in the *cis*-form they point in the same direction. For most peptides the *trans*-form is about 1000 times more stable than the *cis*-form. (b) When the second residue in a peptide is proline the *trans*-form is only about four times more stable than the *cis*-form. *Cis*-proline peptides are found in many proteins.

Why do proteins fold at all?

- The hydrophobic effect
- Hydrogen bonding
- van der Waals interactions
- Ion pairs (salt bridges)
- Longer range electrostatic interactions
- Disulfide bond formation
- Metal coordination, cofactor binding



Predicted structure of BRD4

Can use these principles to design proteins

Natural protein sequences are selected for function (which often means structure)

A few billion natural protein sequences in databases

Number of possible protein sequences for a 300 aa protein

https://www.uniprot.org/

$\cong 10^9$ sequences <<<< $20^{300} \cong 10^{390}$ sequences

vast majority of these likely wouldn't fold to a unique structure

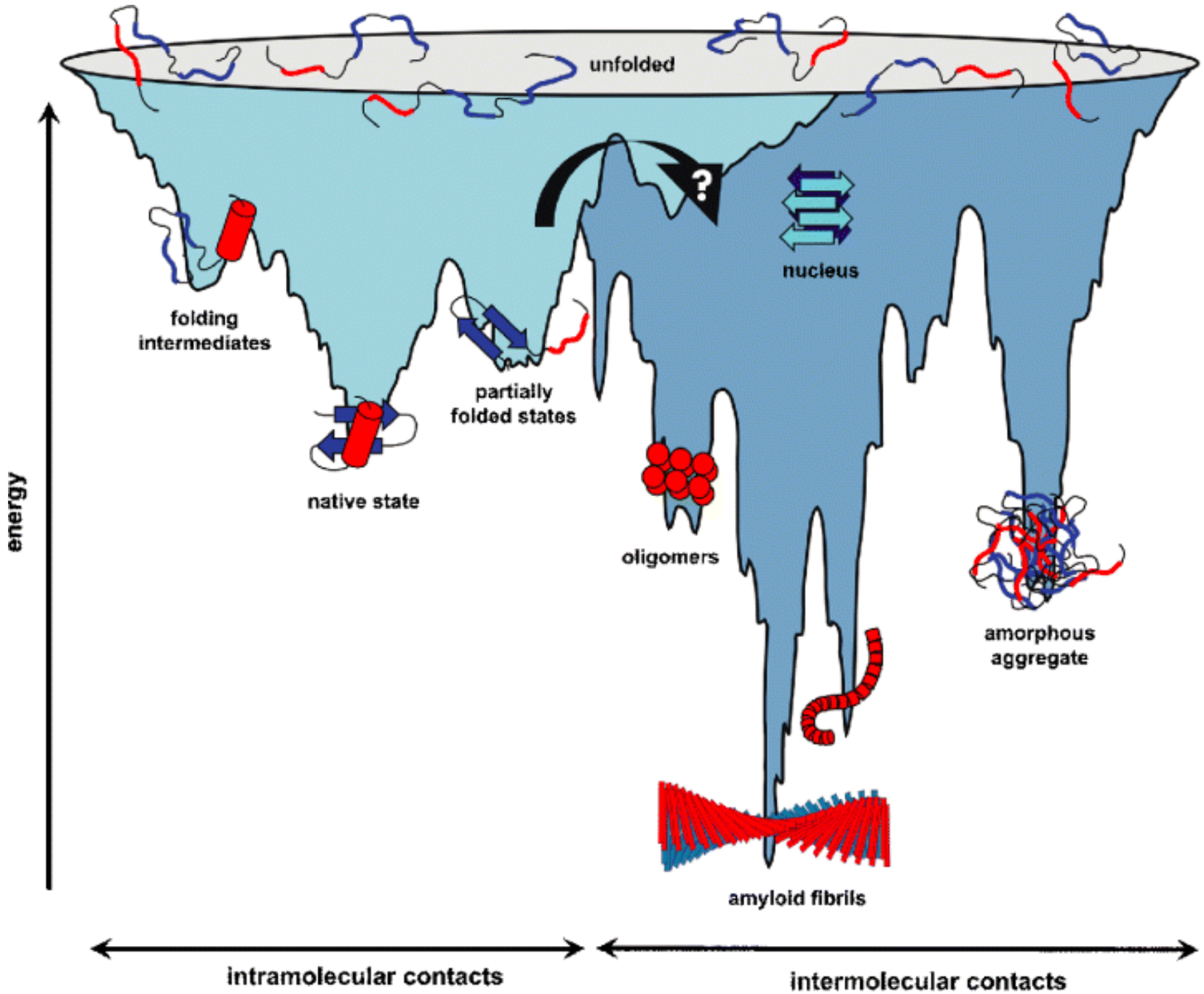
Bioinformatic Methods

BFD was created by clustering 2.5 billion protein sequences from Uniprot/TrEMBL+Swissprot, Metaclust and Soil Reference Catalog Marine Eukaryotic Reference Catalog assembled by Plass.

We clustered sequences that could be aligned to a longer sequence with 90% of their residues and a sequence identity of 30% using Linclust/MMseqs2 --cov-mode 1 --min-seq-id 0.3.

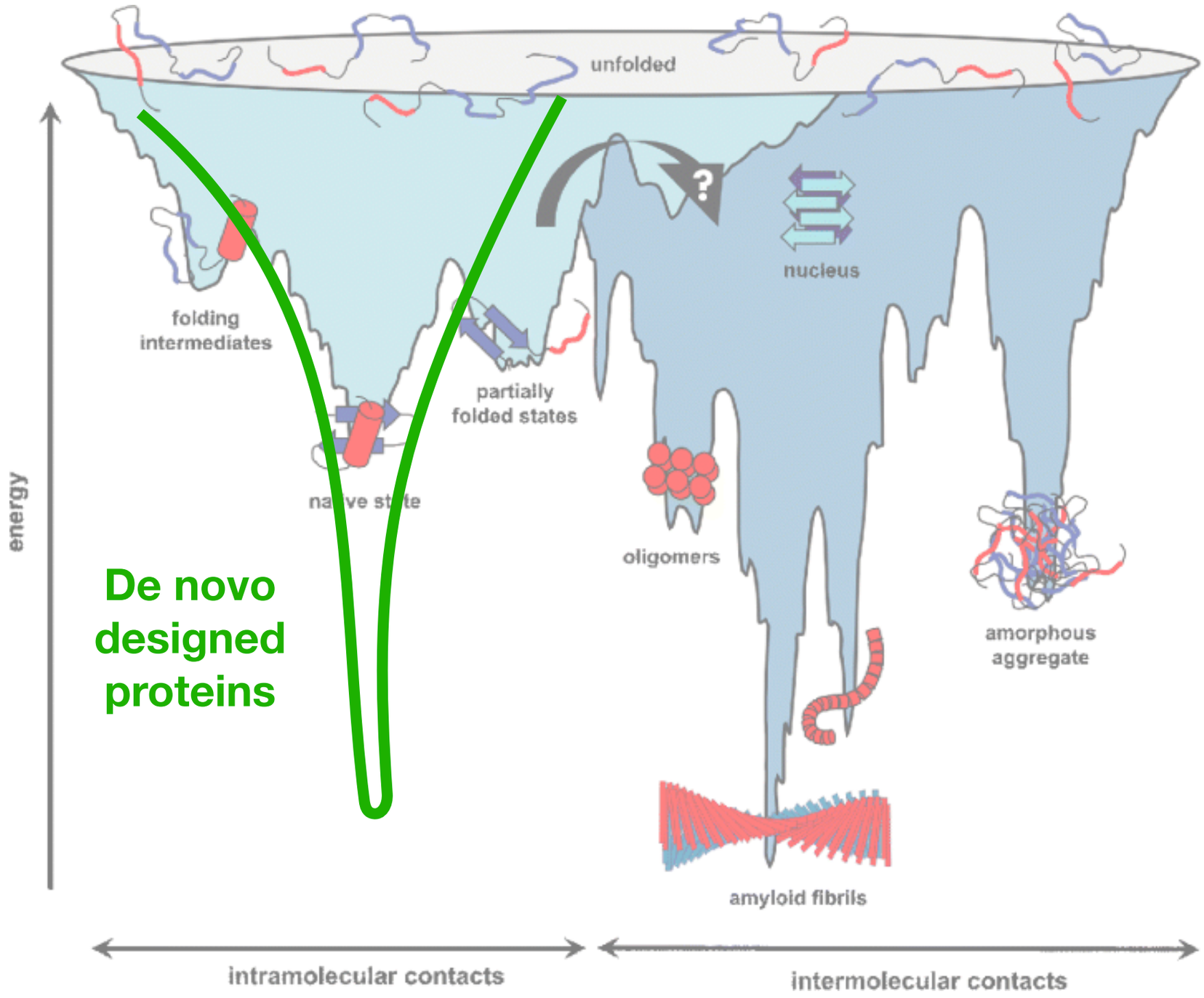
https://bfd.mmseqs.com/

Folding funnel of proteins



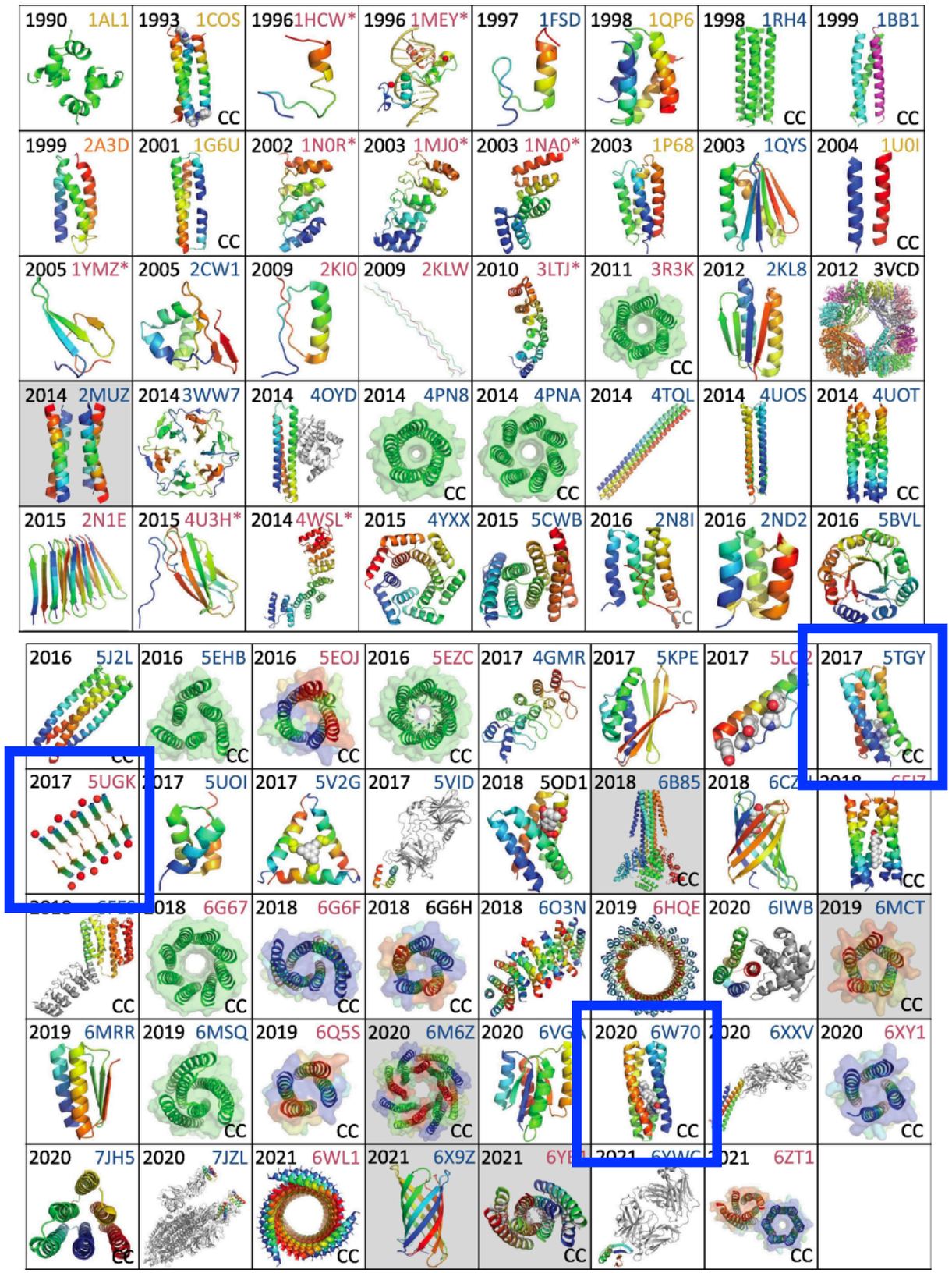
T. R. Jahn, S. E. Radford, The Yin and Yang of protein folding. *FEBS J*. **272**, 5962–5970 (2005).

Designed proteins are often very stable



T. R. Jahn, S. E. Radford, The Yin and Yang of protein folding. *FEBS J*. **272**, 5962–5970 (2005).

A structural history of protein design



Woolfson, *J Mol Biol,* 2021

Earliest

Most recent

Folding can proceed through intermediates without well-defined structure

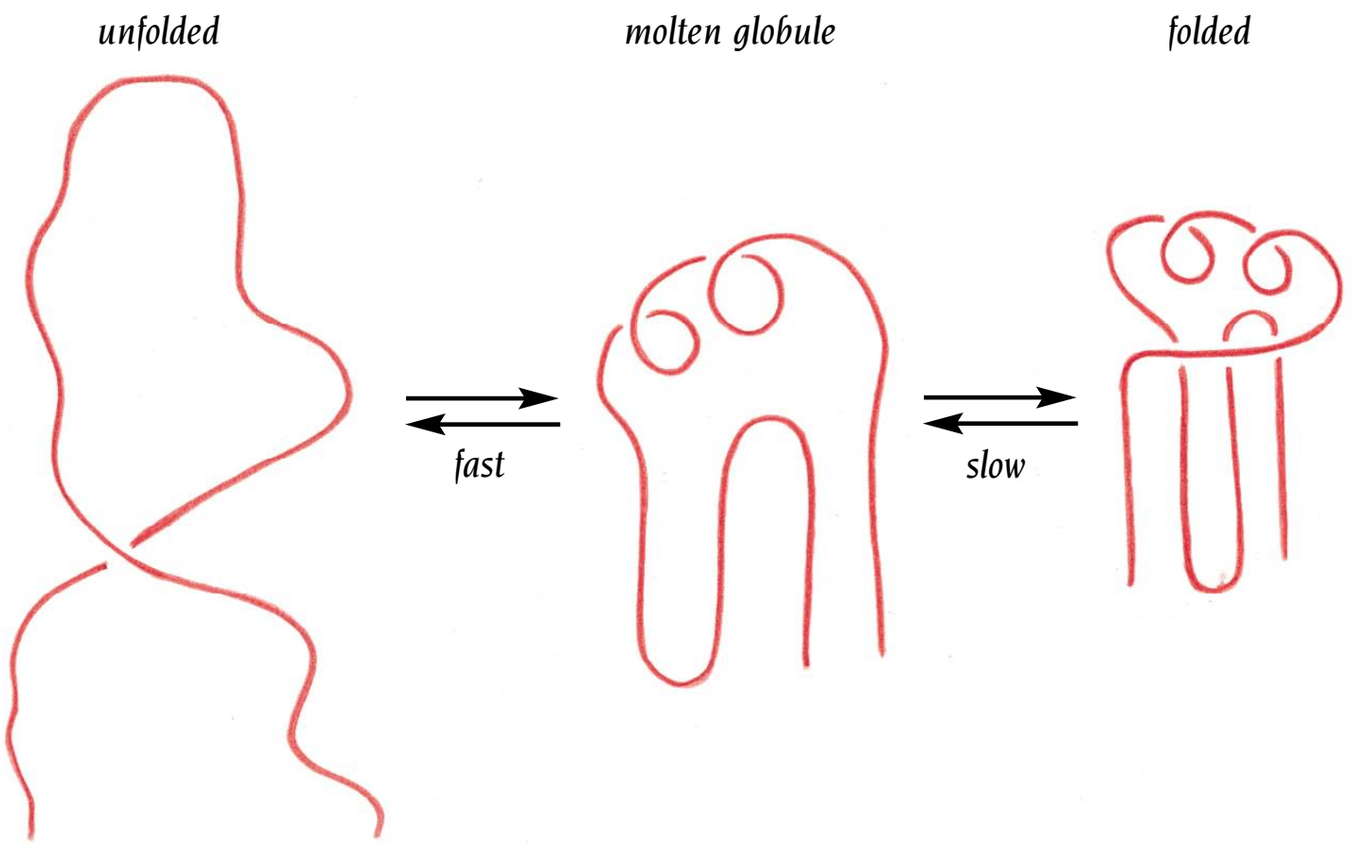


Figure 6.2 The molten globule state is an important intermediate in the folding pathway when a polypeptide chain converts from an unfolded to a folded state. The molten globule has most of the secondary structure of the native state but it is less compact and the proper packing interactions in the interior of the protein have not been formed.

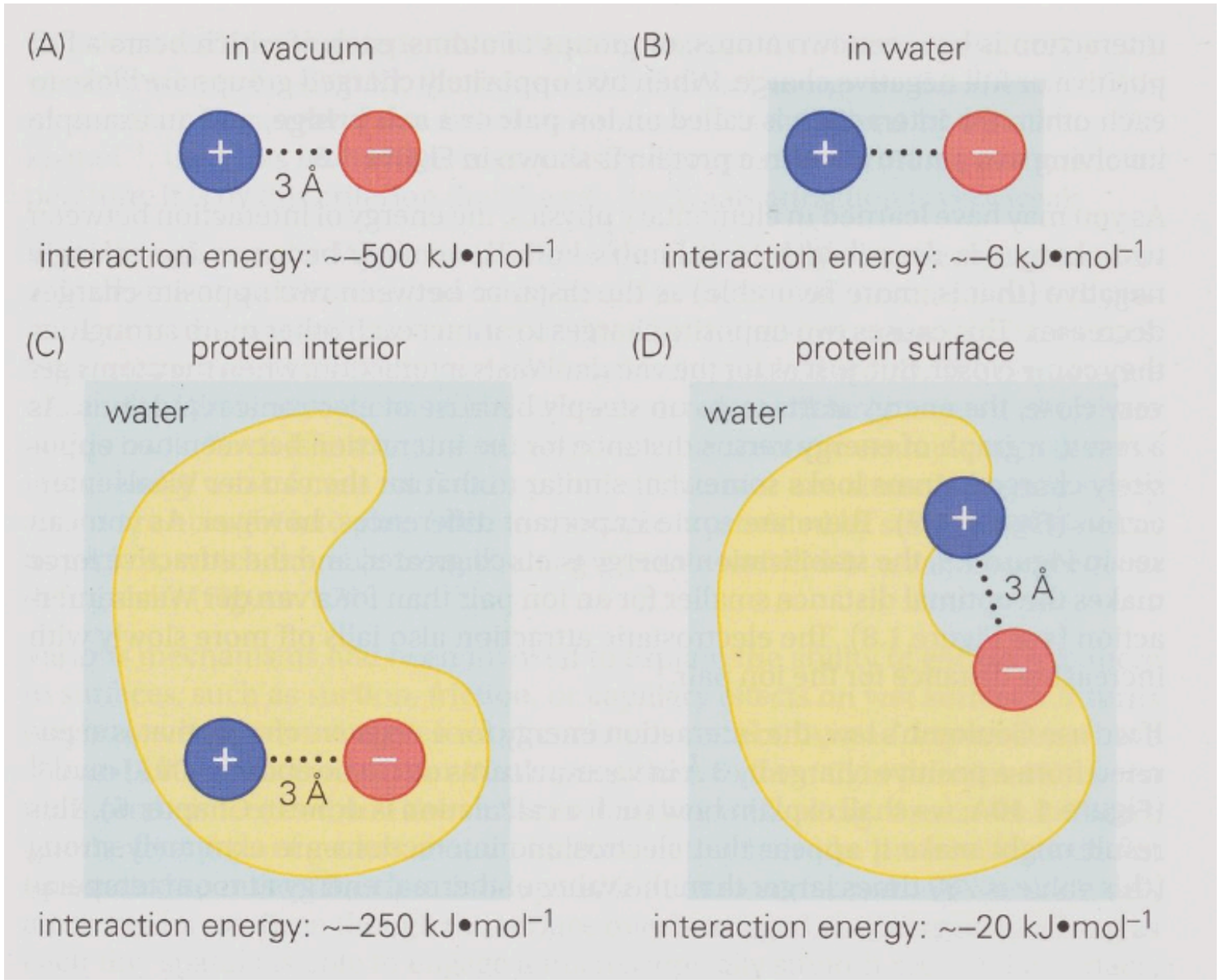
Molecular interactions that govern protein structure

- Electrostatic interactions
  - $E(r) \propto (1/r)$ where r is the distance between the charges

- Dipole-dipole
  - $E(r) \propto (1/r^3)$ where r is the distance between the dipoles
  - H-bonds are a special case of dipole-dipole

- Van der Waals
  - Attractive force due to induced dipoles
  - Lennard-Jones potential $V(r) = 4\varepsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right]$

Electrostatic energy



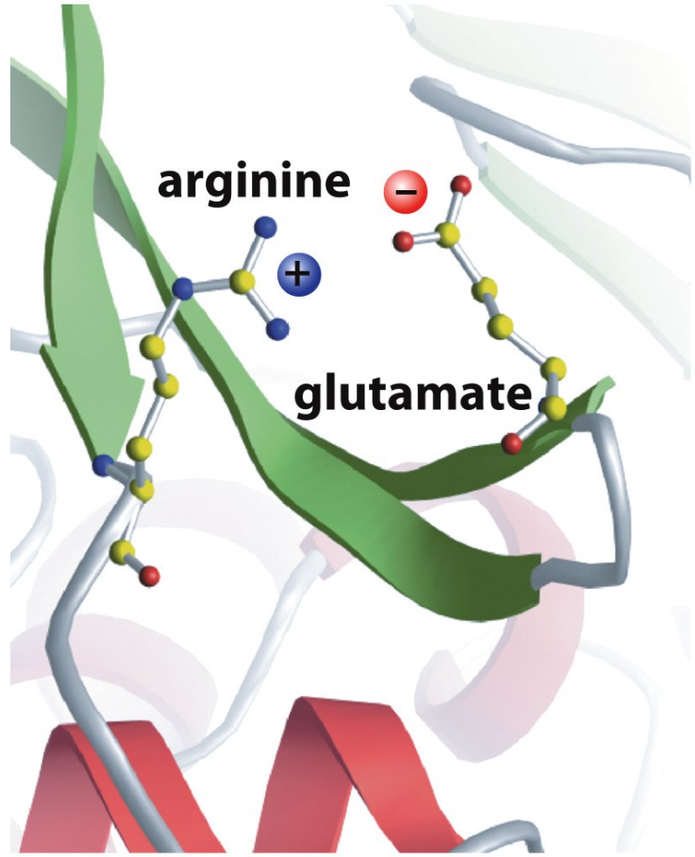
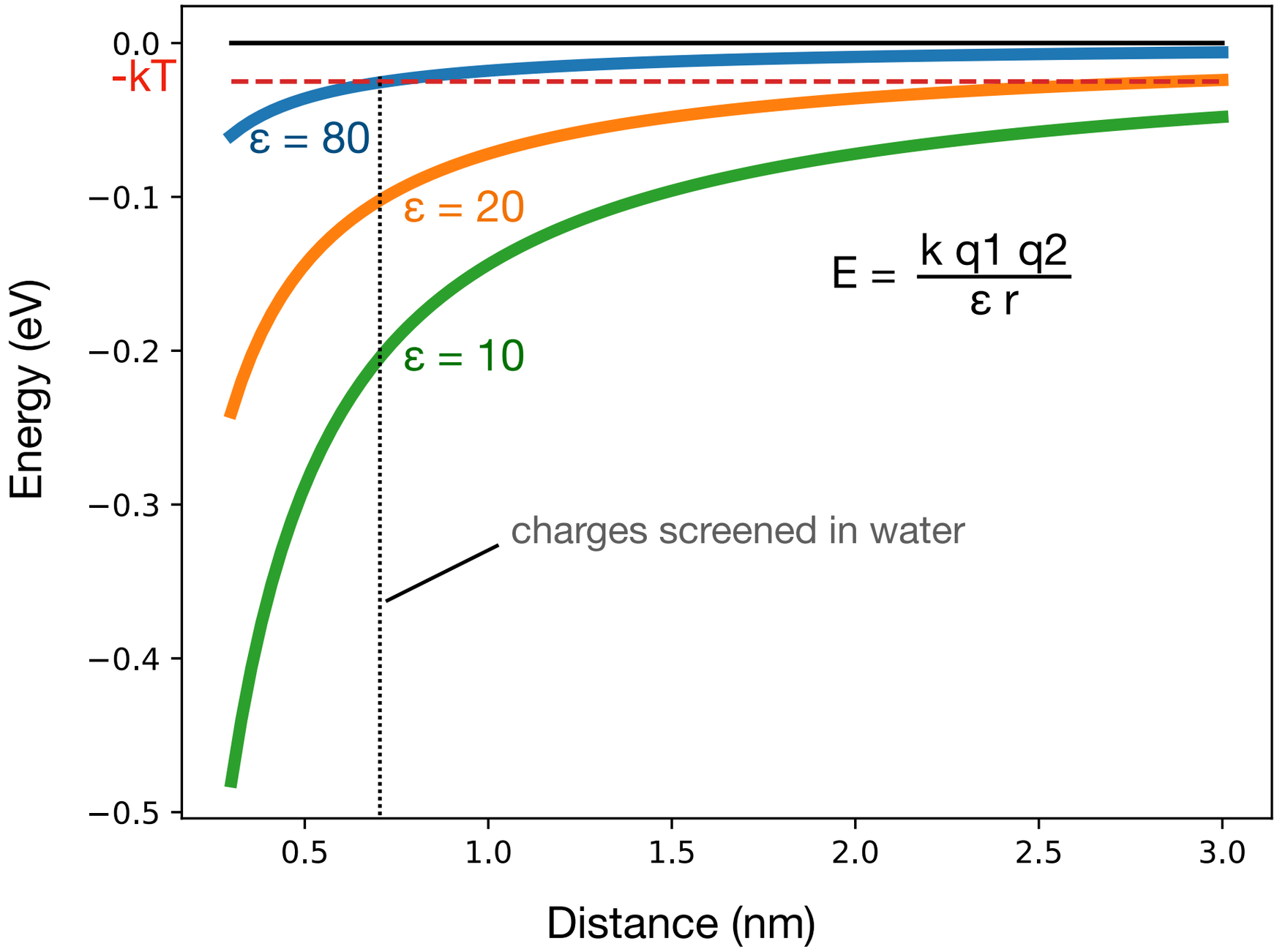


Figure 1.8 The Molecules of Life (© Garland Science 2013)

Electrostatic energy



Hydrogen Bonds: example of dipole-dipole

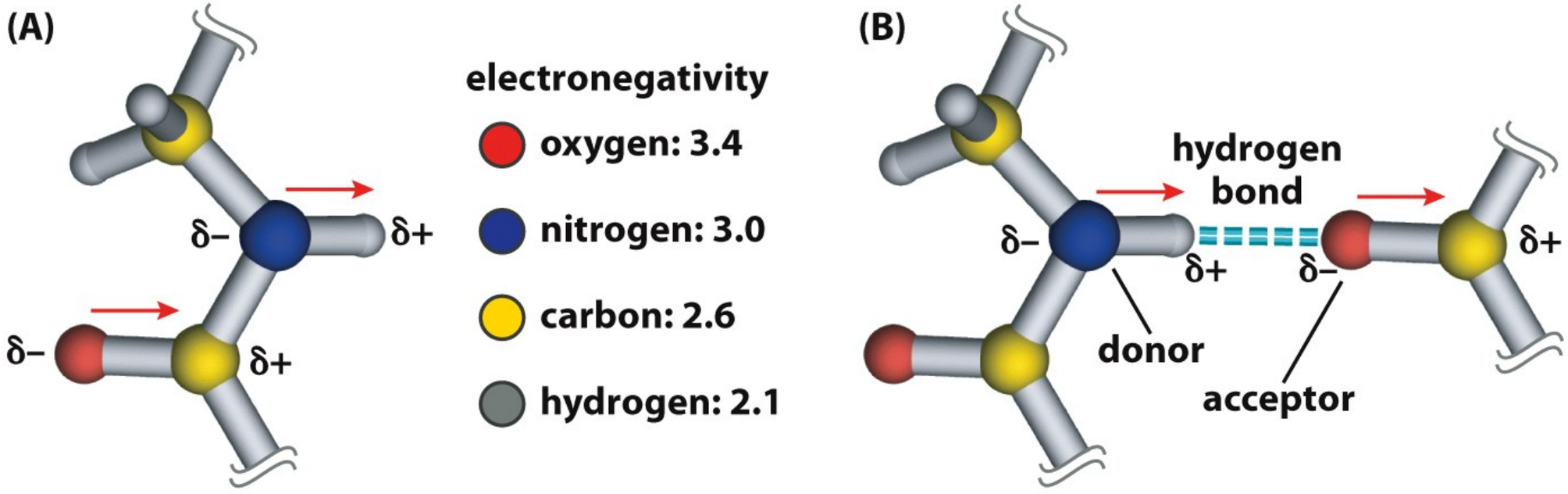


Figure 1.11 The Molecules of Life (© Garland Science 2013)

- ubiquitous in protein structures
- backbone-backbone
- sidechain hydroxyl and NH groups to each other, to backbone amide and carbonyl groups
- Many ordered water molecules coordinated through H-bonds, ”water-mediated” hydrogen bonds…

Hydrogen Bonds: example of dipole-dipole

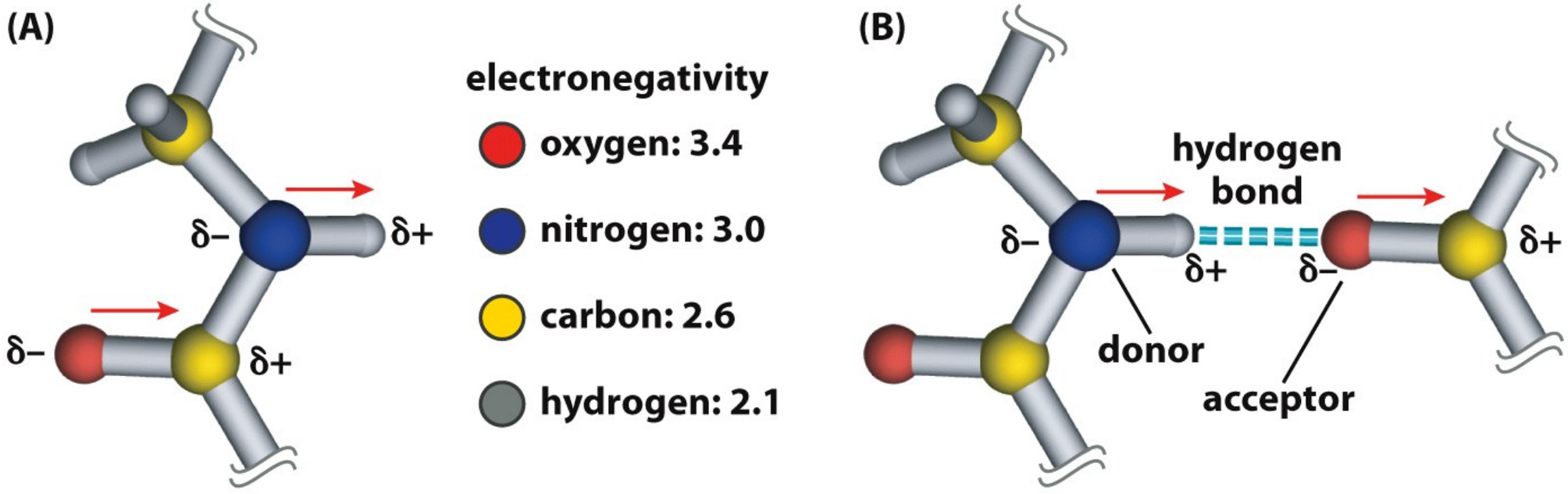


Figure 1.11 The Molecules of Life (© Garland Science 2013)

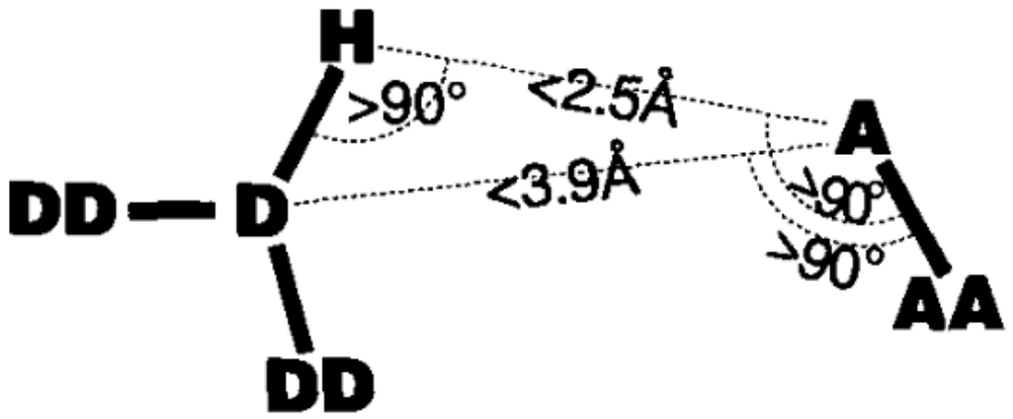


Figure 1. Geometric criteria for hydrogen bonds. D is the donor heavy atom, H the hydrogen, A the acceptor, DD donor antecedent (i.e. an atom 2 covalent bonds away from the hydrogen), AA acceptor antecedent, DDD an atom 3 covalent bonds from the hydrogen, AAA an atom 2 covalent bonds from the acceptor.

I. K. McDonald, J. M. Thornton, Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793 (1994).

London dispersion forces

- arise from induced dipoles in closely adjacent atoms

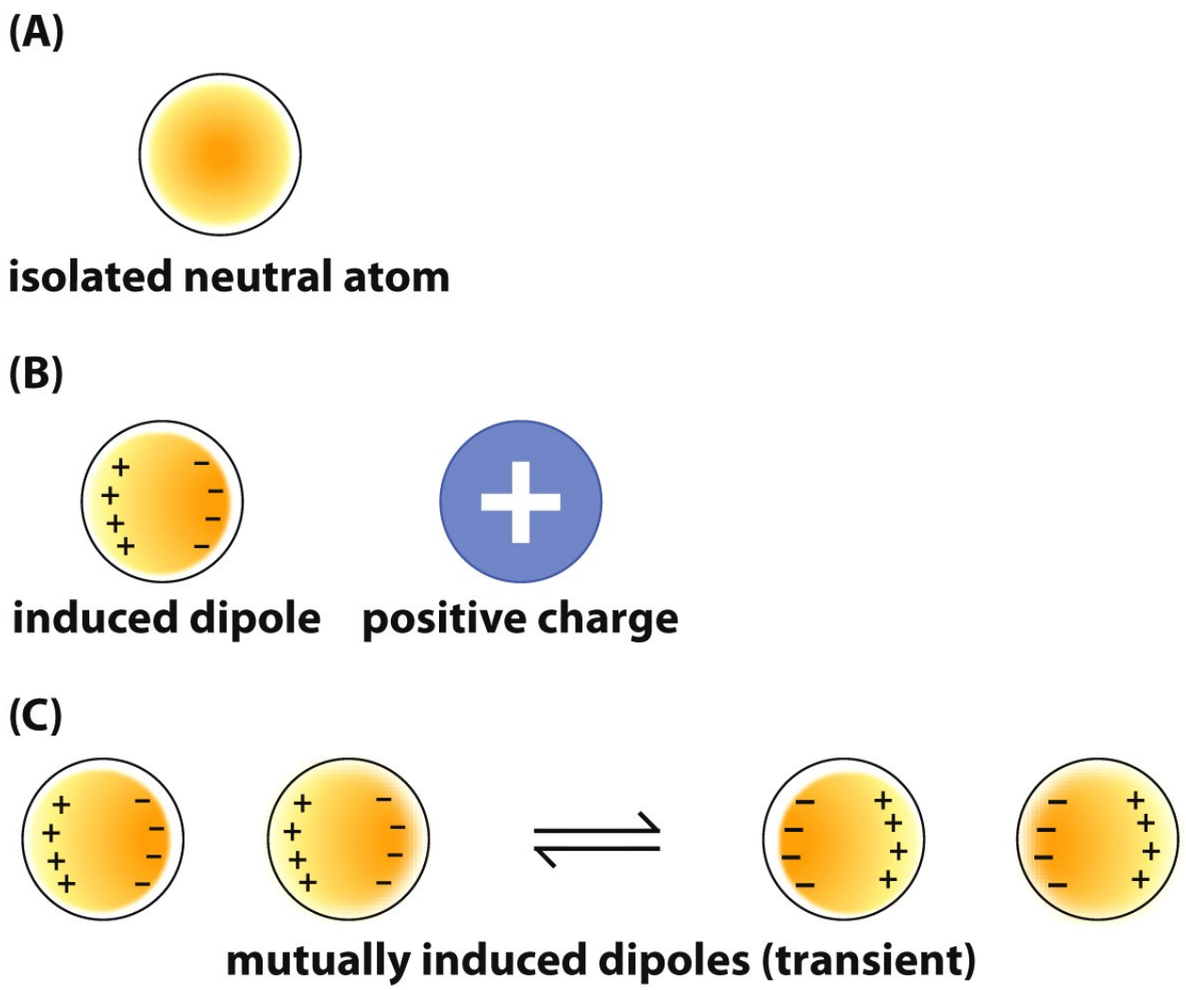


Figure 1.5 The Molecules of Life (© Garland Science 2013)

Table 1.1 Van der Waals radii and the electronegativities for atoms commonly found in biological molecules.

Atom	van der Waals radius (Å)	Electro-negativity (Pauling scale)
O	1.5	3.4
Cl	1.9	3.2
N	1.6	3.0
S	1.8	2.6
C	1.7	2.6
P	1.8	2.2
H	1.2	2.1

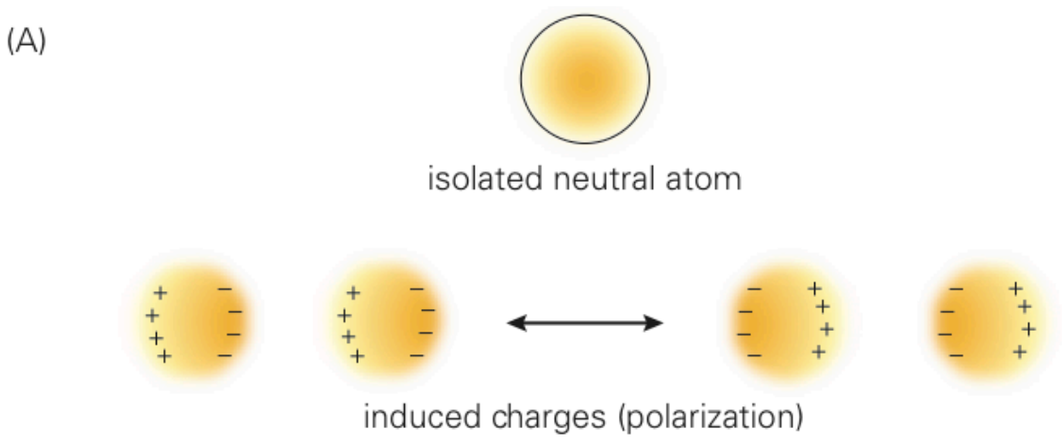
The atoms are listed from largest electronegativity (electron withdrawing ability) to smallest, as determined by Linus Pauling.

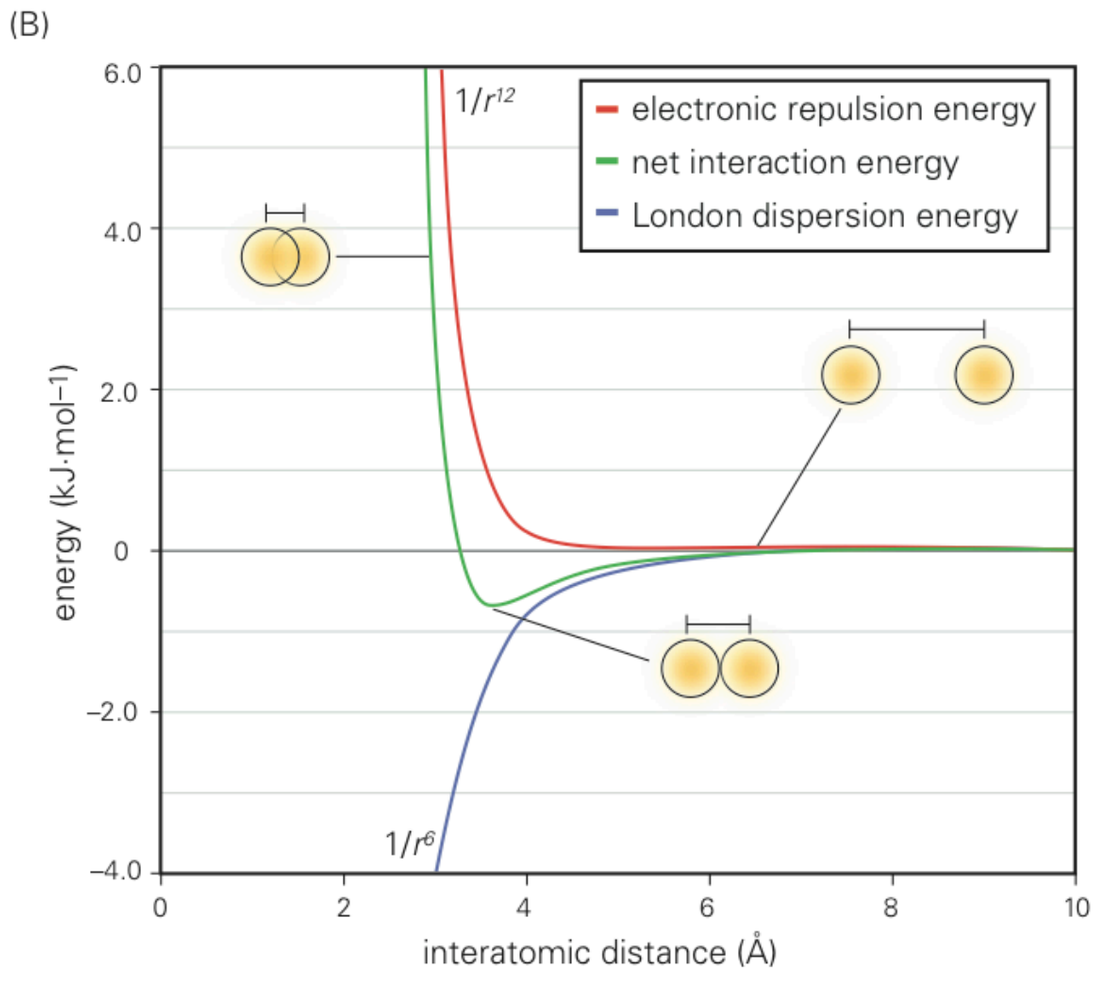
Table 1.1 The Molecules of Life (© Garland Science 2013)

London dispersion forces

Lennard-Jones Potential $$V(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

- model for the close range attractive and repulsive forces between neutral atoms
- ε is the well depth
- σ the "hard-sphere" radius





Relative energy of van der Waals and hydrogen bond interactions as a function of interatomic distance

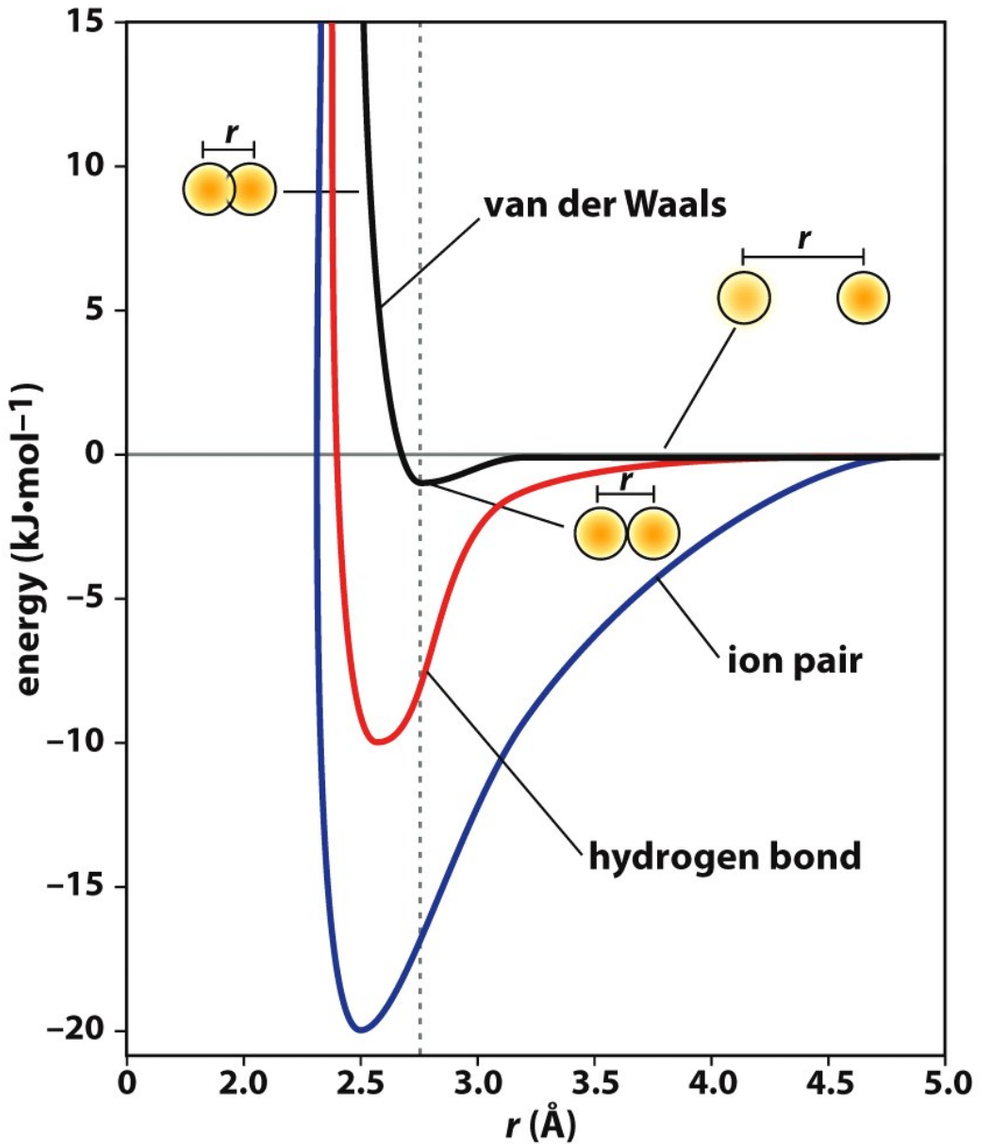


Figure 1.12 The Molecules of Life (© Garland Science 2013)

Disulfide formation

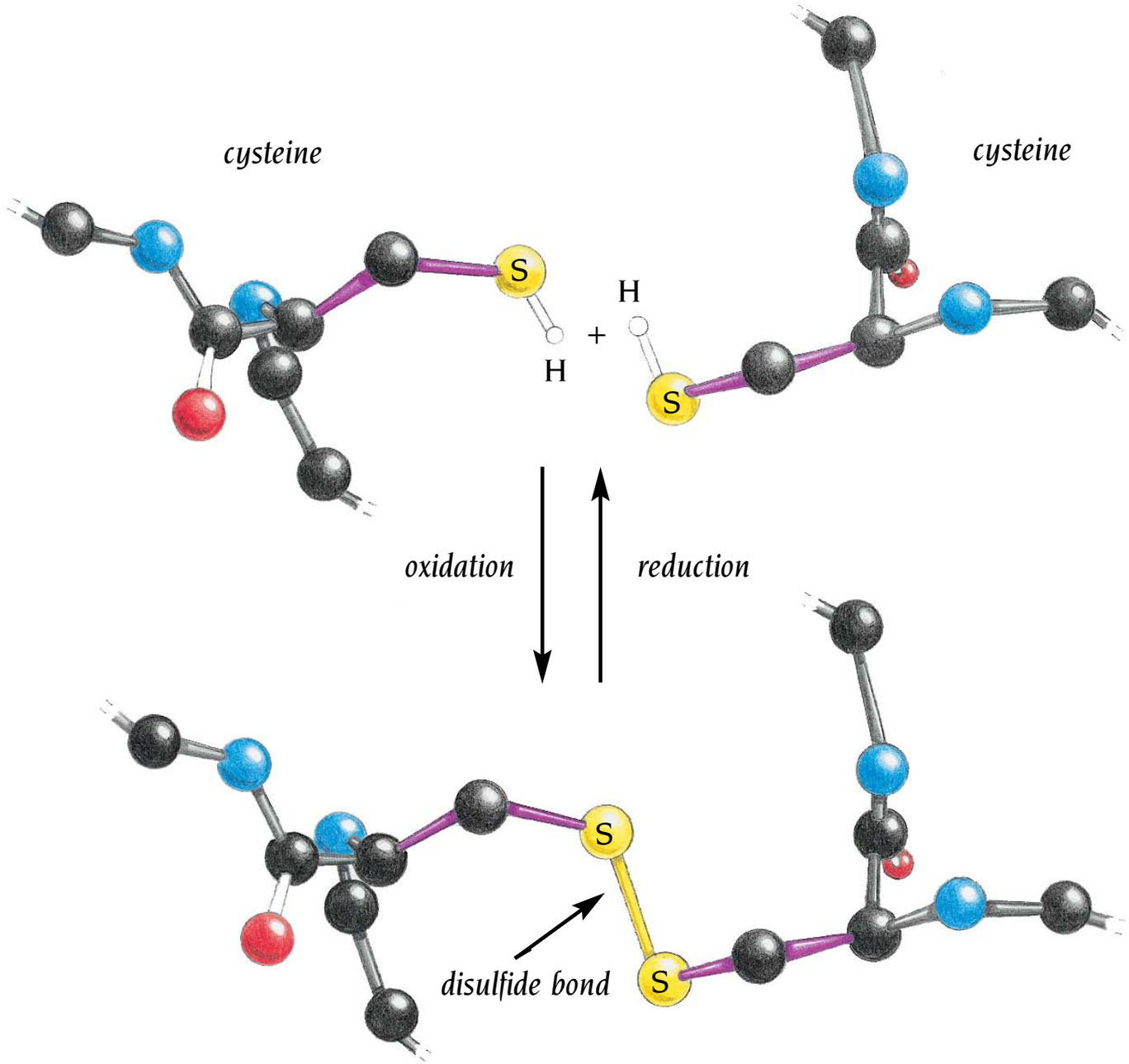
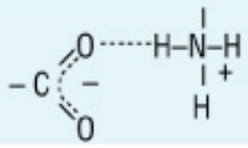
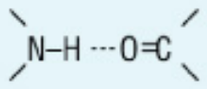
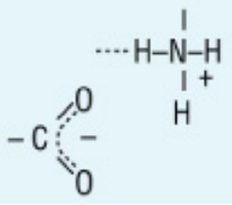
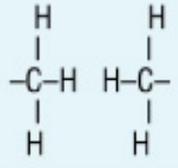


Figure 1.4 The disulfide is usually the end product of air oxidation according to the following schematic reaction scheme:

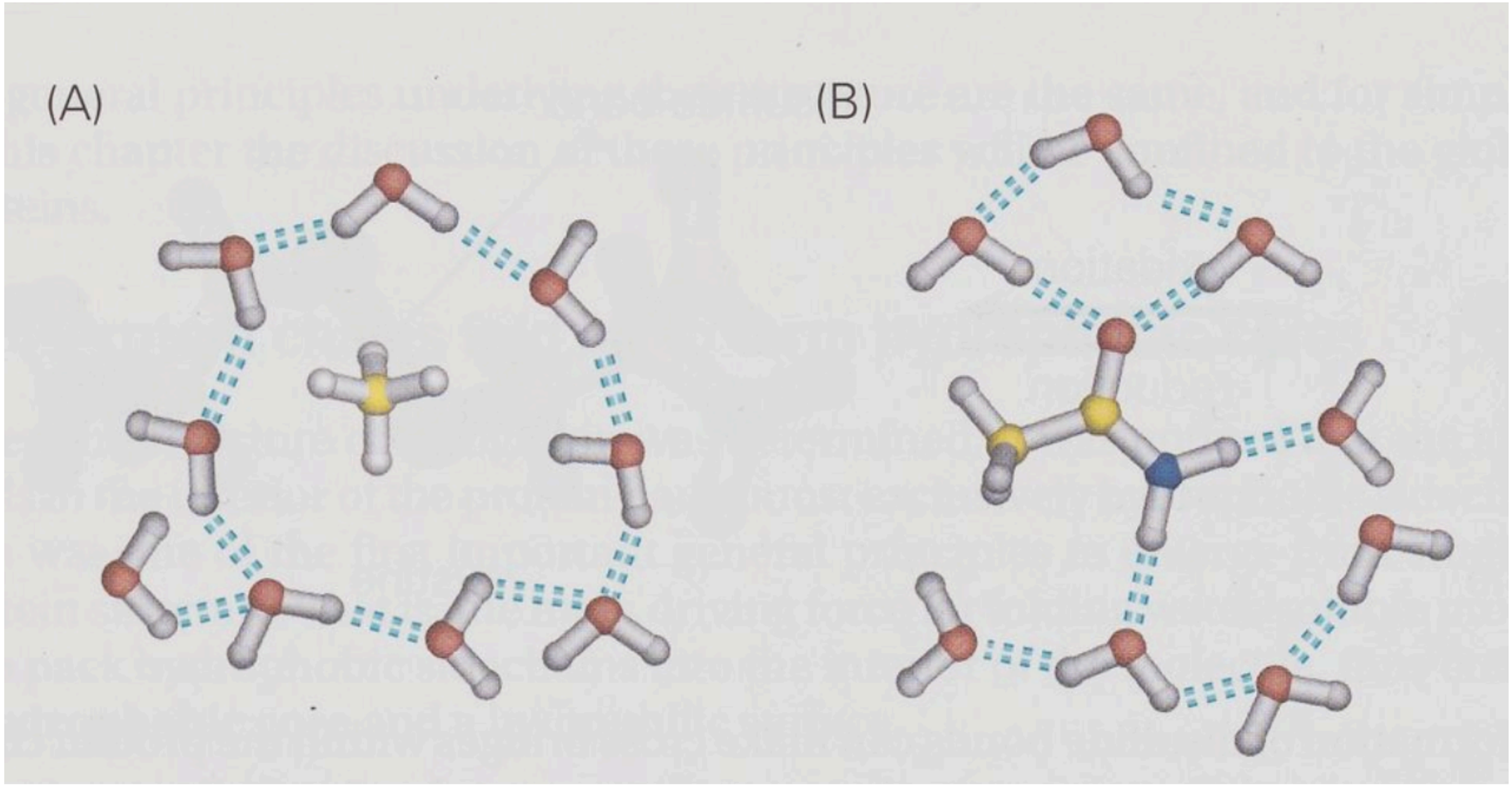
$$2\ \text{–}CH_2SH + \tfrac{1}{2}\ O_2 \rightleftharpoons \text{–}CH_2\text{–}S\text{–}S\text{–}CH_2 + H_2O$$

Disulfide bonds form between the side chains of two cysteine residues. Two SH groups from cysteine residues, which may be in different parts of the amino acid sequence but adjacent in the three-dimensional structure, are oxidized to form one S–S (disulfide) group.

Chemical Interactions that Stabilize Polypeptides

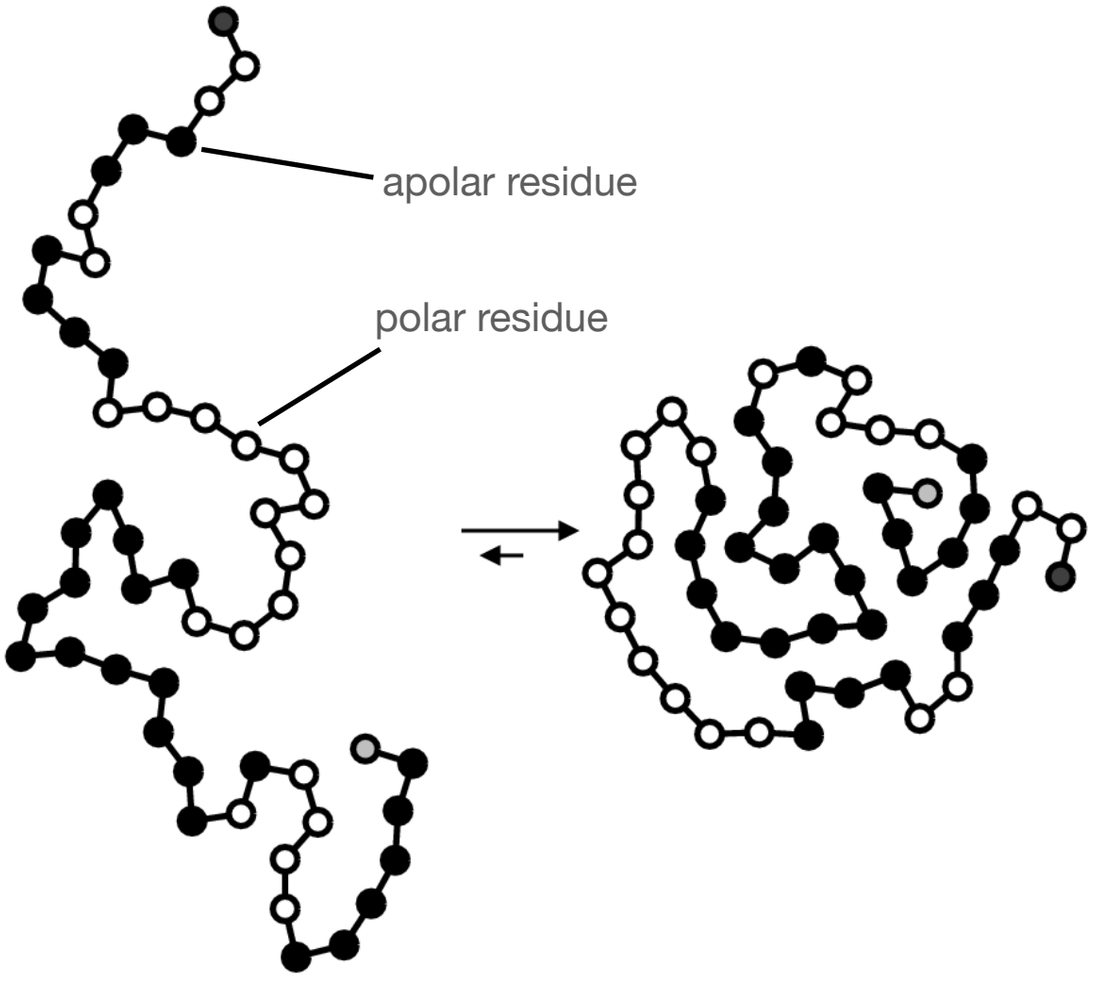
Interaction	Example	Distance dependence	Typical distance	Free energy (bond dissociation enthalpies for the covalent bonds)
Covalent bond	$-C_\alpha-C-$	-	1.5 Å	356 kJ/mole (610 kJ/mole for a C=C bond)
Disulfide bond	–Cys–S–S–Cys–	-	2.2 Å	167 kJ/mole
Salt bridge		Donor (here N), and acceptor (here O) atoms <3.5 Å	2.8 Å	12.5–17 kJ/mole; may be as high as 30 kJ/mole for fully or partially buried salt bridges (see text), less if the salt bridge is external
Hydrogen bond		Donor (here N), and acceptor (here O) atoms <3.5 Å	3.0 Å	2–6 kJ/mole in water; 12.5–21 kJ/mole if either donor or acceptor is charged
Long-range electrostatic interaction		Depends on dielectric constant of medium. Screened by water. 1/r dependence	Variable	Depends on distance and environment. Can be very strong in nonpolar region but very weak in water
Van der Waals interaction		Short range. Falls off rapidly beyond 4 Å separation. $1/r^6$ dependence	3.5 Å	4 kJ/mole (4–17 in protein interior) depending on the size of the group (for comparison, the average thermal energy of molecules at room temperature is 2.5 kJ/mole)

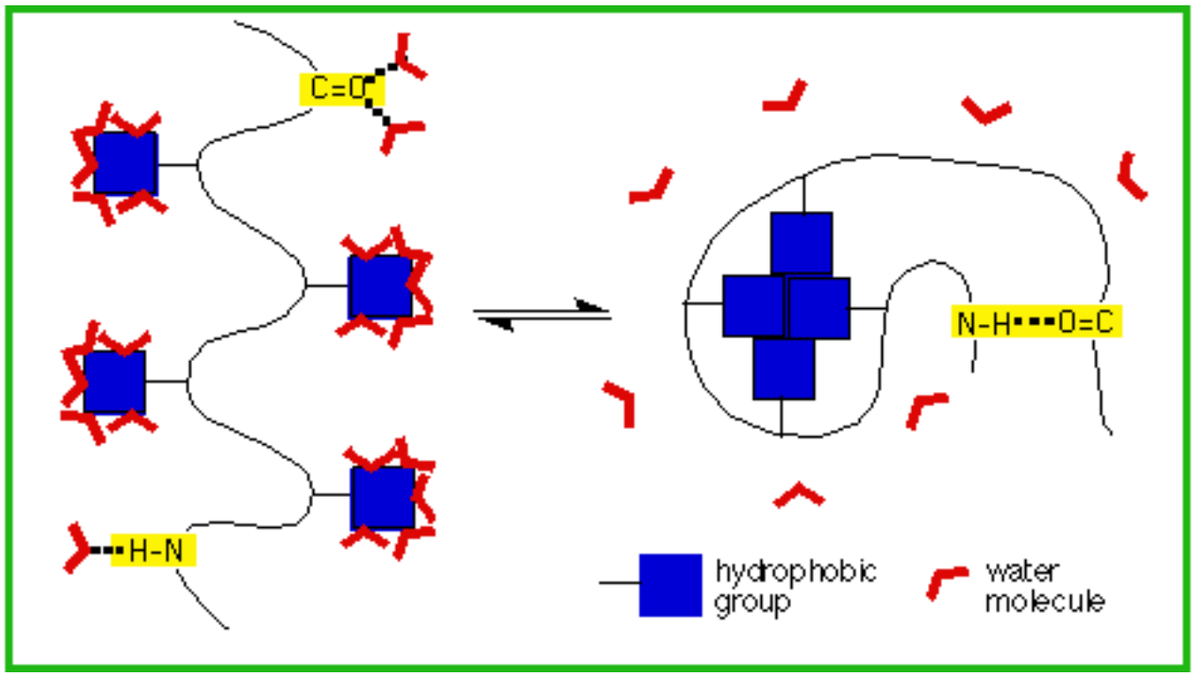
The hydrophobic effect is extremely important for driving folding



The hydrophobic effect is extremely important for driving folding

hydrophobic collapse



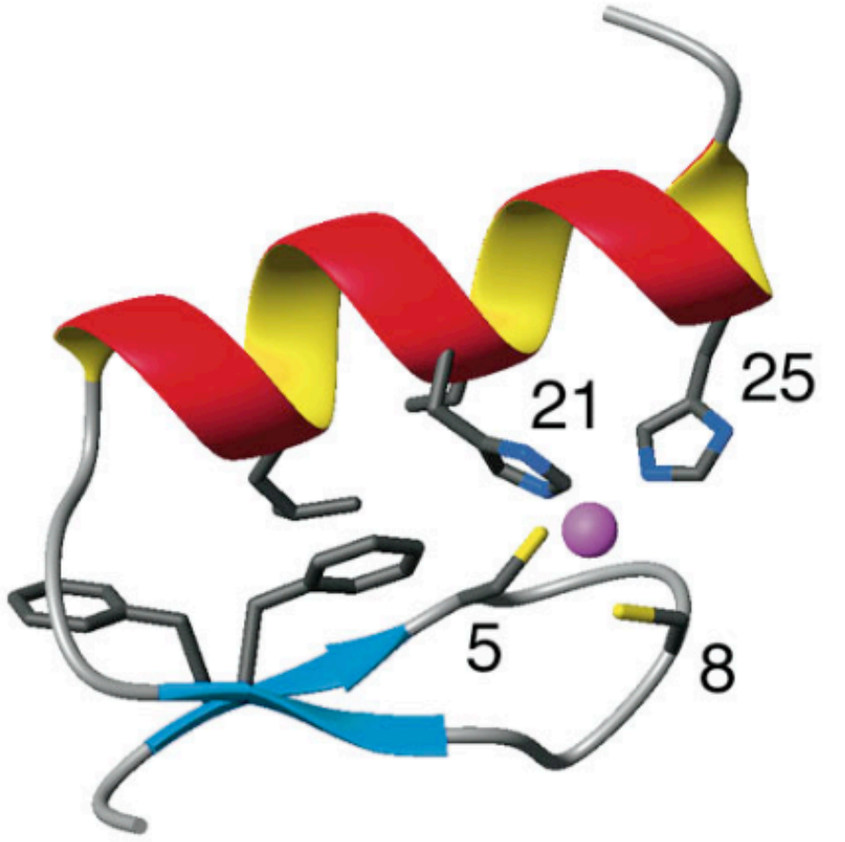


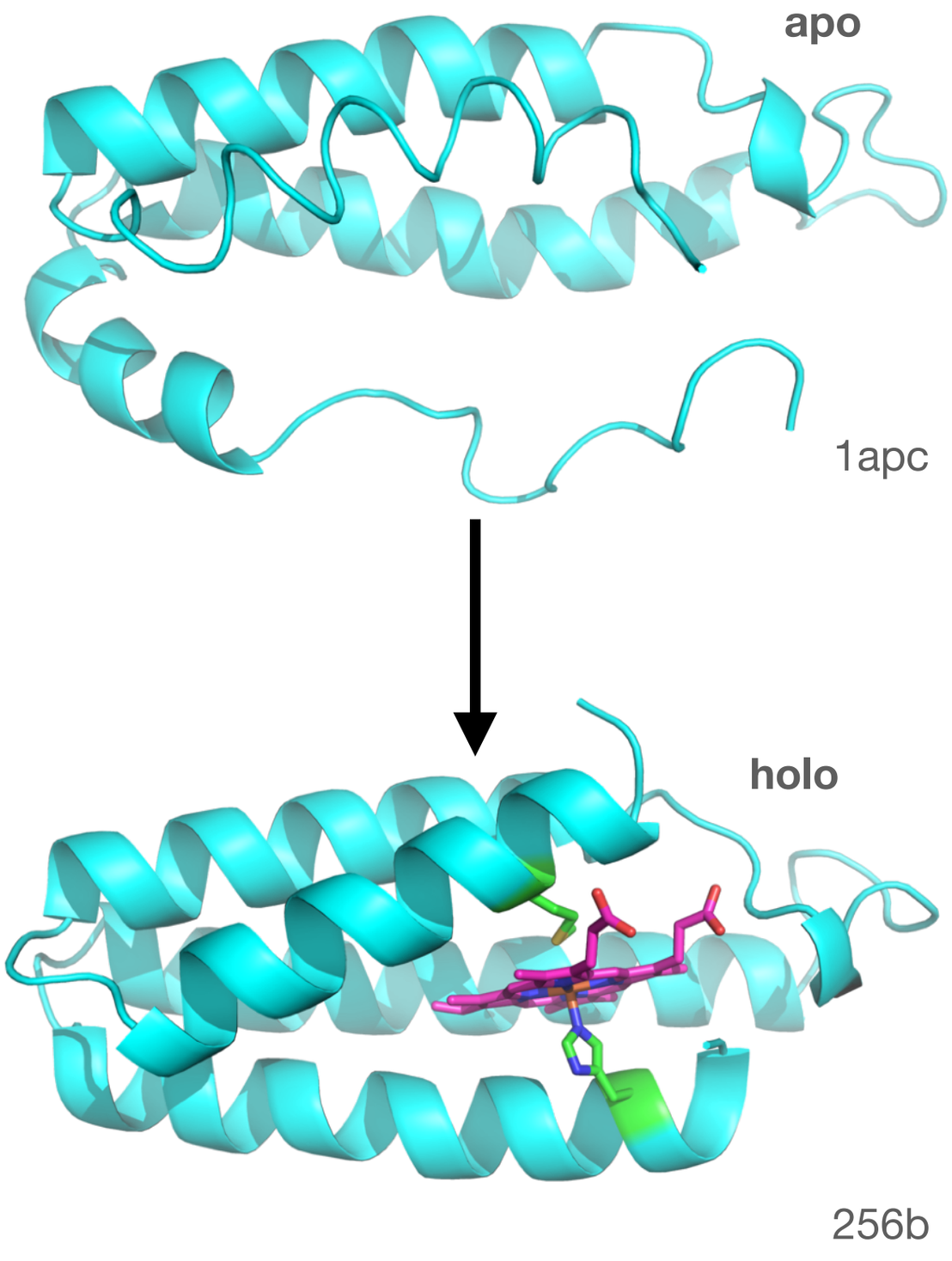
https://www.cryst.bbk.ac.uk/PPS2/projects/day/TDayDiss/Major.html

256b.pse

Proteins bind ligands to fold and function

Metal/cofactor binding can drive folding

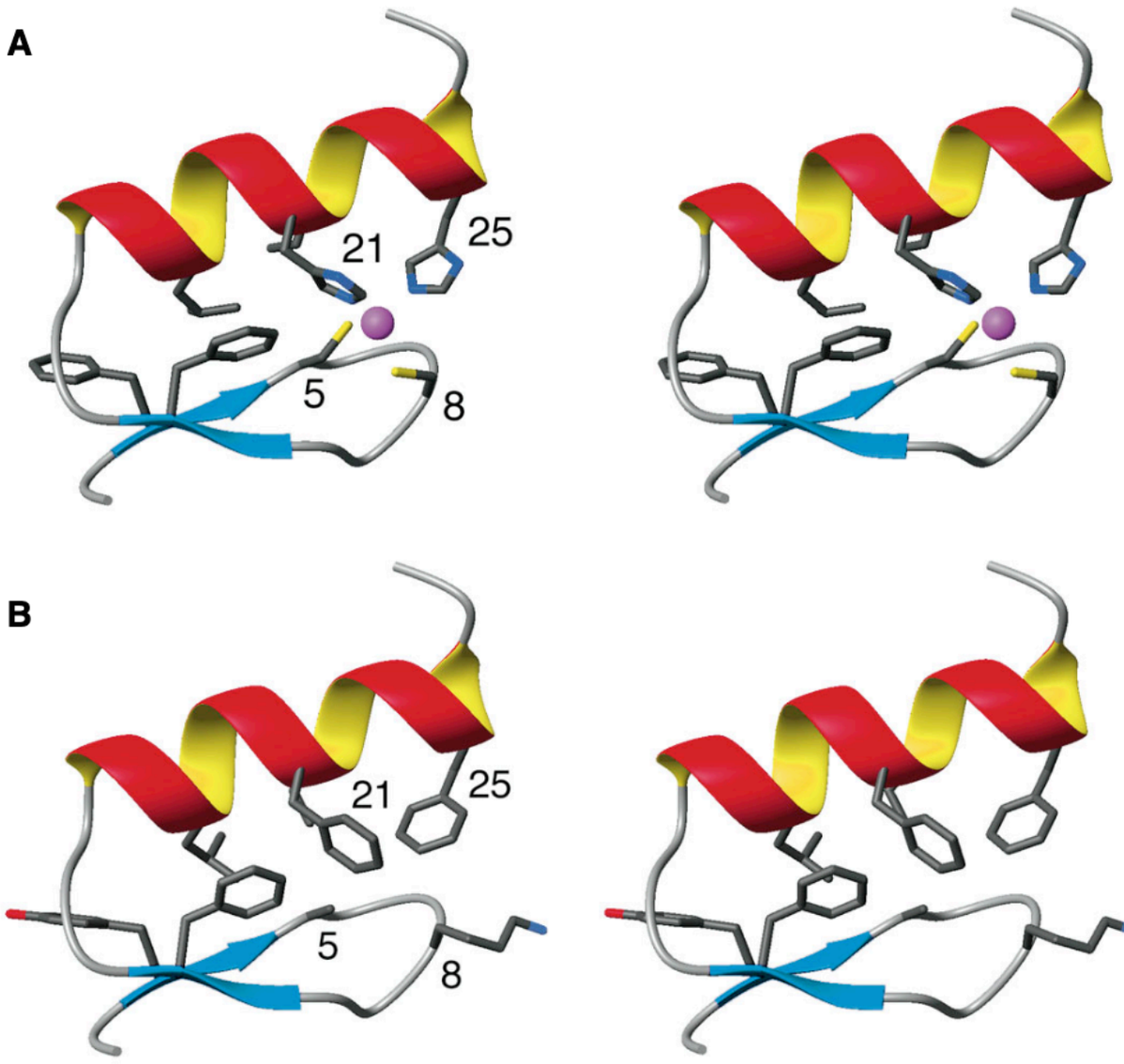




De Novo Protein Design: Fully Automated Sequence Selection

Bassil I. Dahiyat† and Stephen L. Mayo*

SCIENCE • VOL. 278 • 3 OCTOBER 1997 • www.sciencemag.org



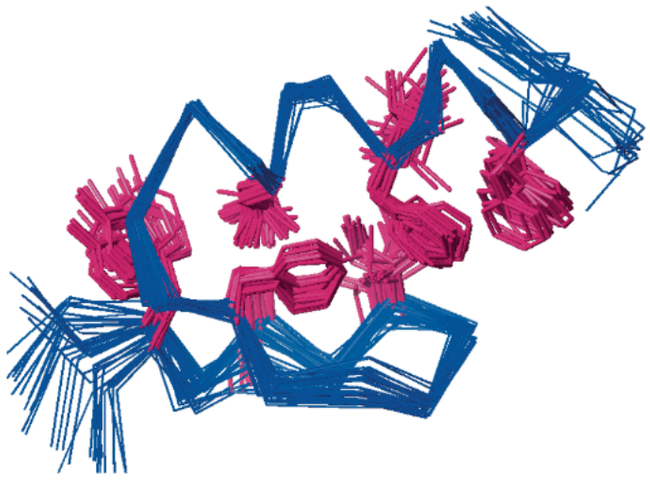


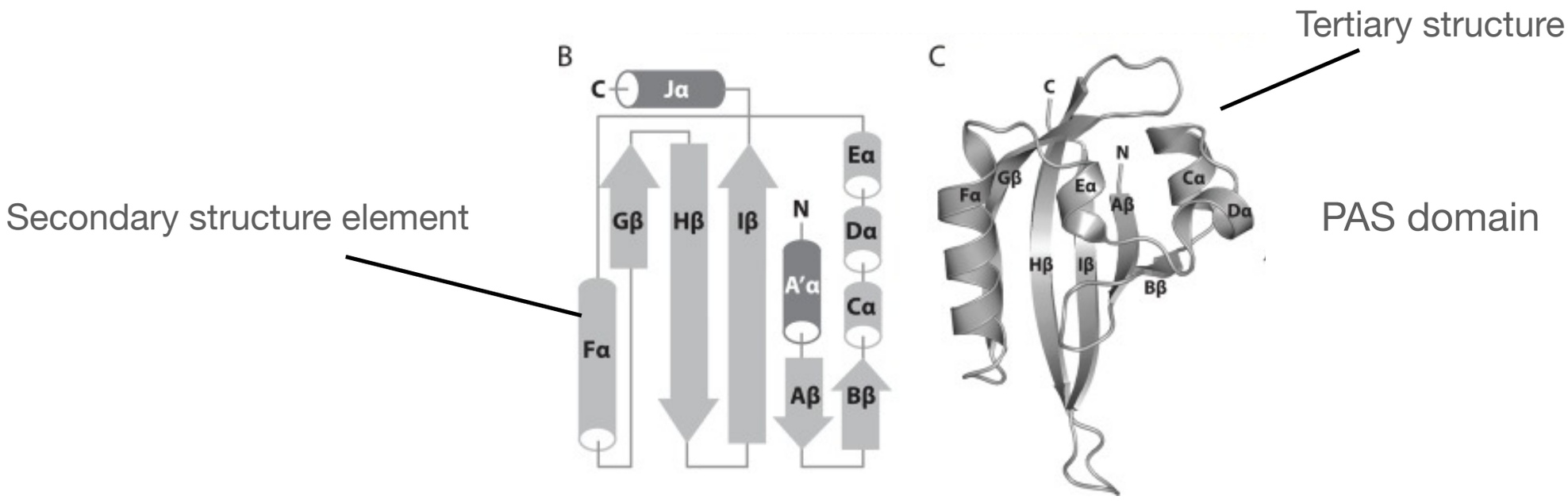
Fig. 2. Comparison of Zif268 (*9*) and computed FSD-1 structures. (**A**) Stereoview of the second zinc finger module of Zif268 showing its buried residues and zinc binding site. (**B**) Stereoview of the computed orientations of buried side chains in FSD-1. For clarity, only side chains from residues 3, 5, 8, 12, 18, 21, 22, and 25 are shown. Color figures were created with MOLMOL (*38*).

Definition of protein secondary structure

Consecutive stretch of amino acid residues exhibiting consistent ϕ,ψ angles and regular, repeating H-bonding interactions

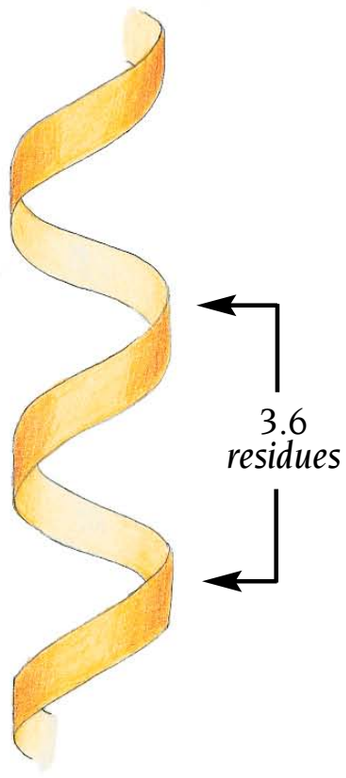
- α helices and β sheets are the major secondary structure elements

Elegant and comprehensive solution to satisfying the H-bonding requirements of the polypeptide backbone

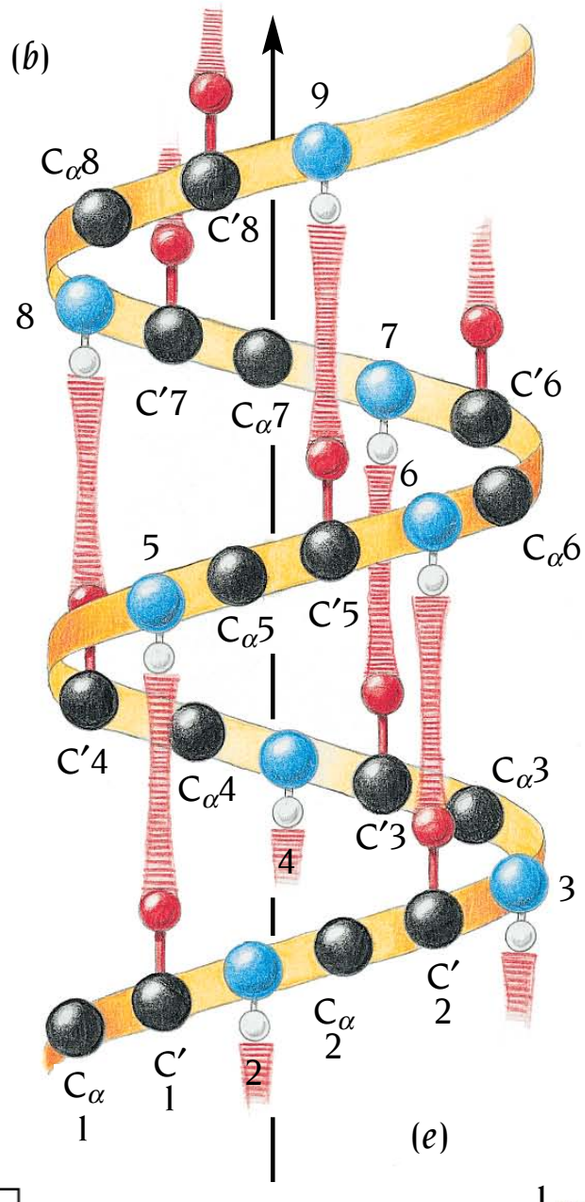


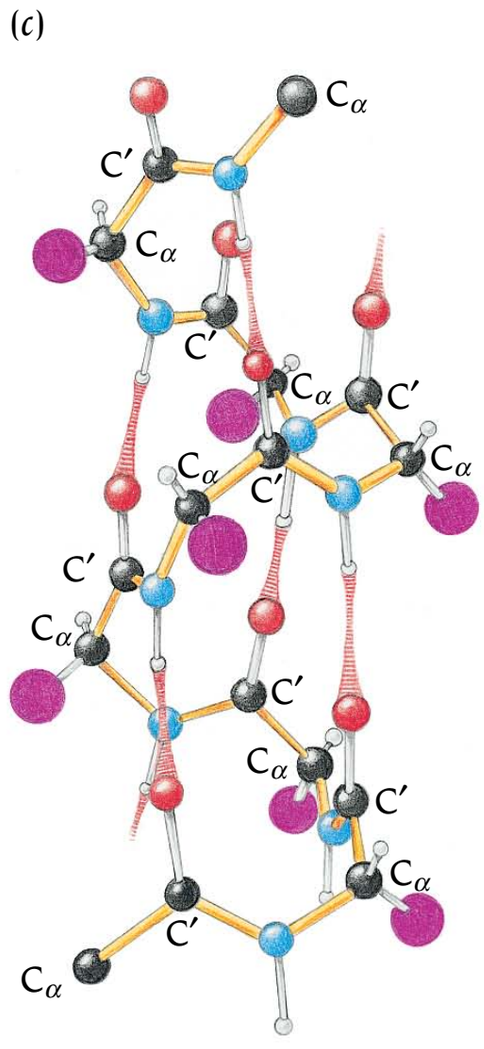
Biochemistry. 2015 Jan 20; 54(2): 134–149.

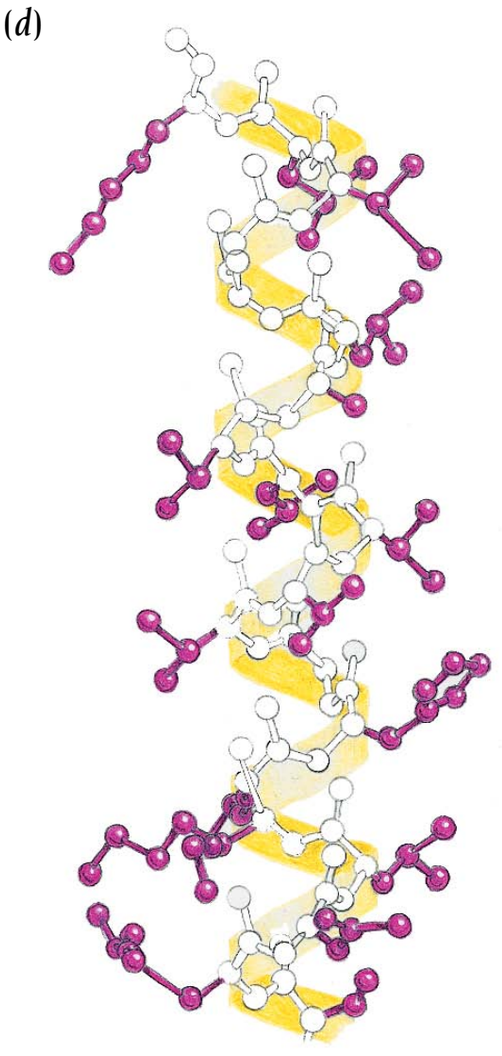
The alpha helix

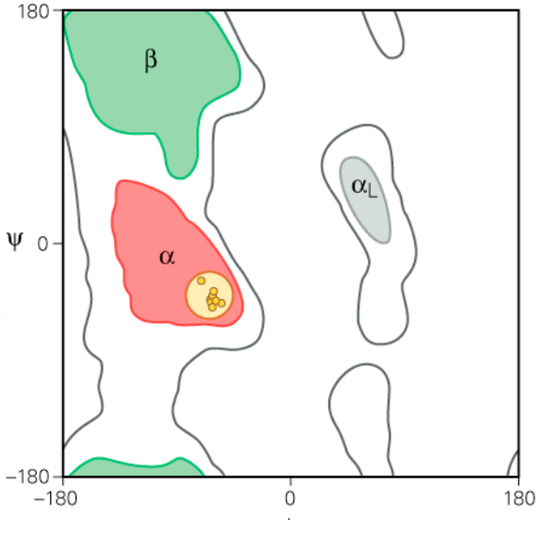


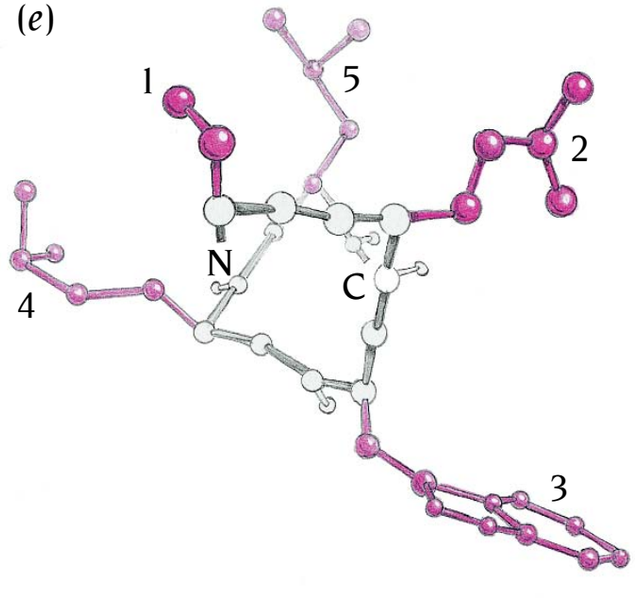
3.6 residues per turn
5.4 Å (1.5 Å per res)





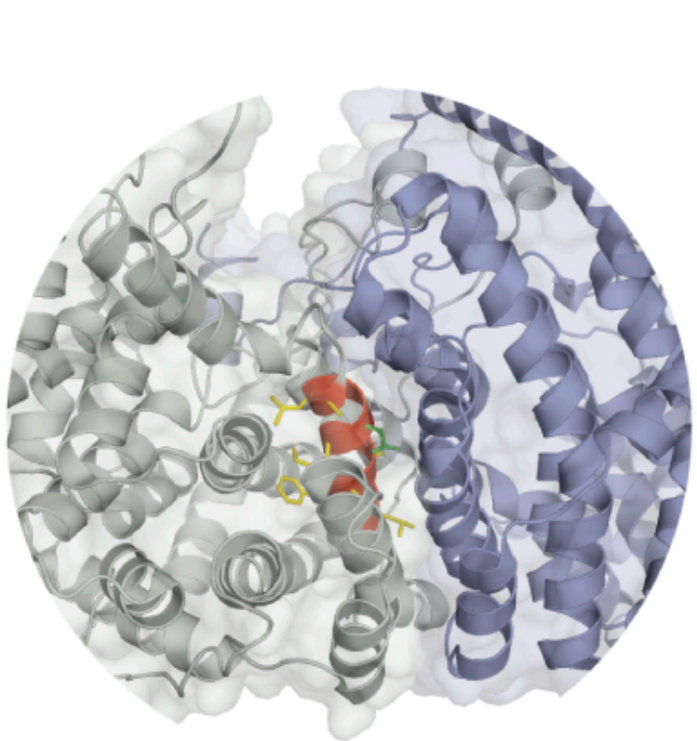




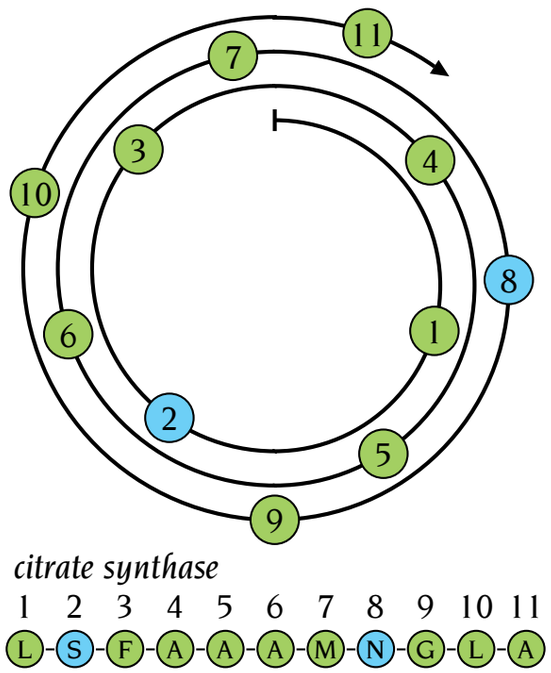


Helical handedness in Florence architecture

Sequence profiles of a helix suggest its native environment

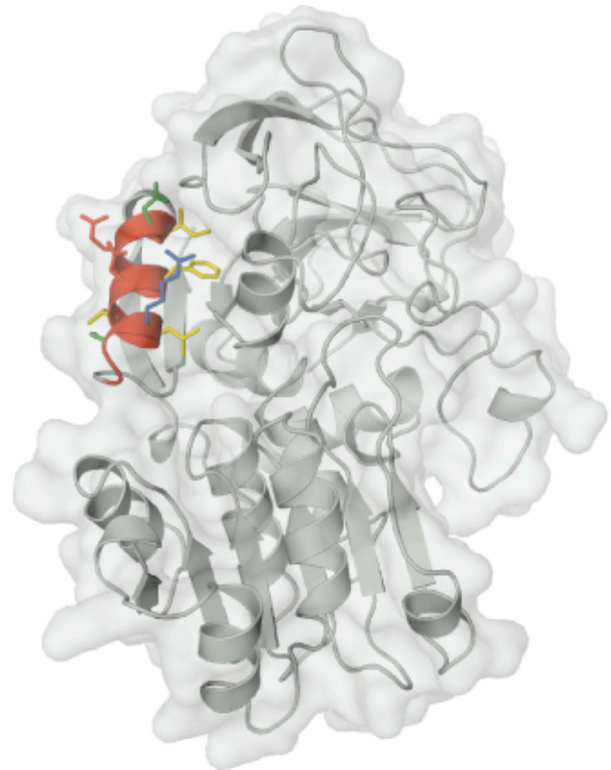


Buried helix in citrate synthase

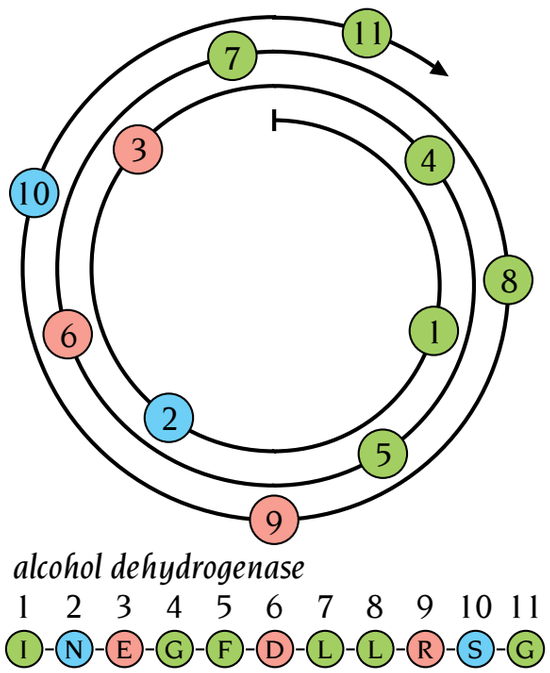


citrate synthase

1	2	3	4	5	6	7	8	9	10	11
L	S	F	A	A	A	M	N	G	L	A



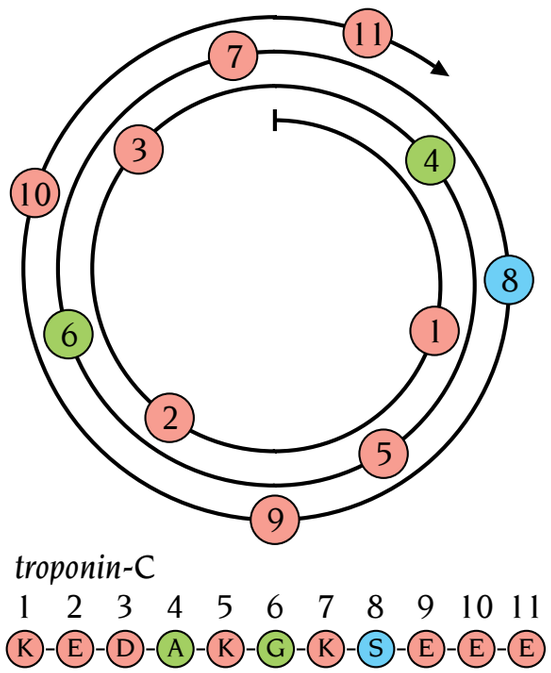
Partially exposed helix from alcohol dehydrogenase



alcohol dehydrogenase

1	2	3	4	5	6	7	8	9	10	11
I	N	E	G	F	D	L	L	R	S	G

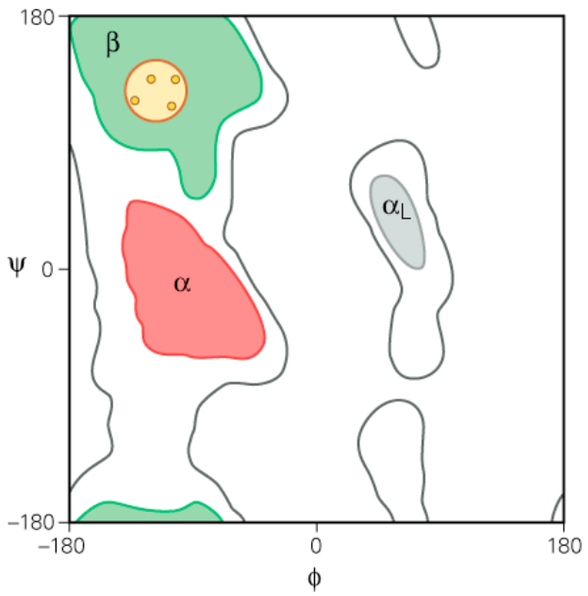
Fully exposed helix from troponin C

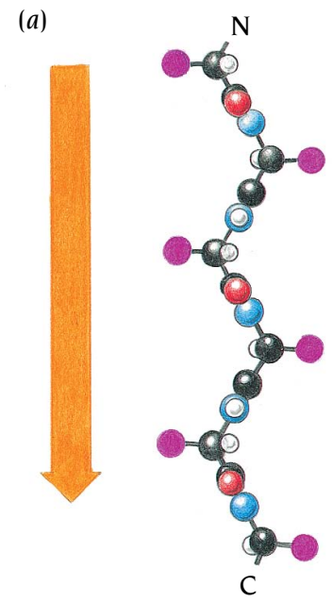


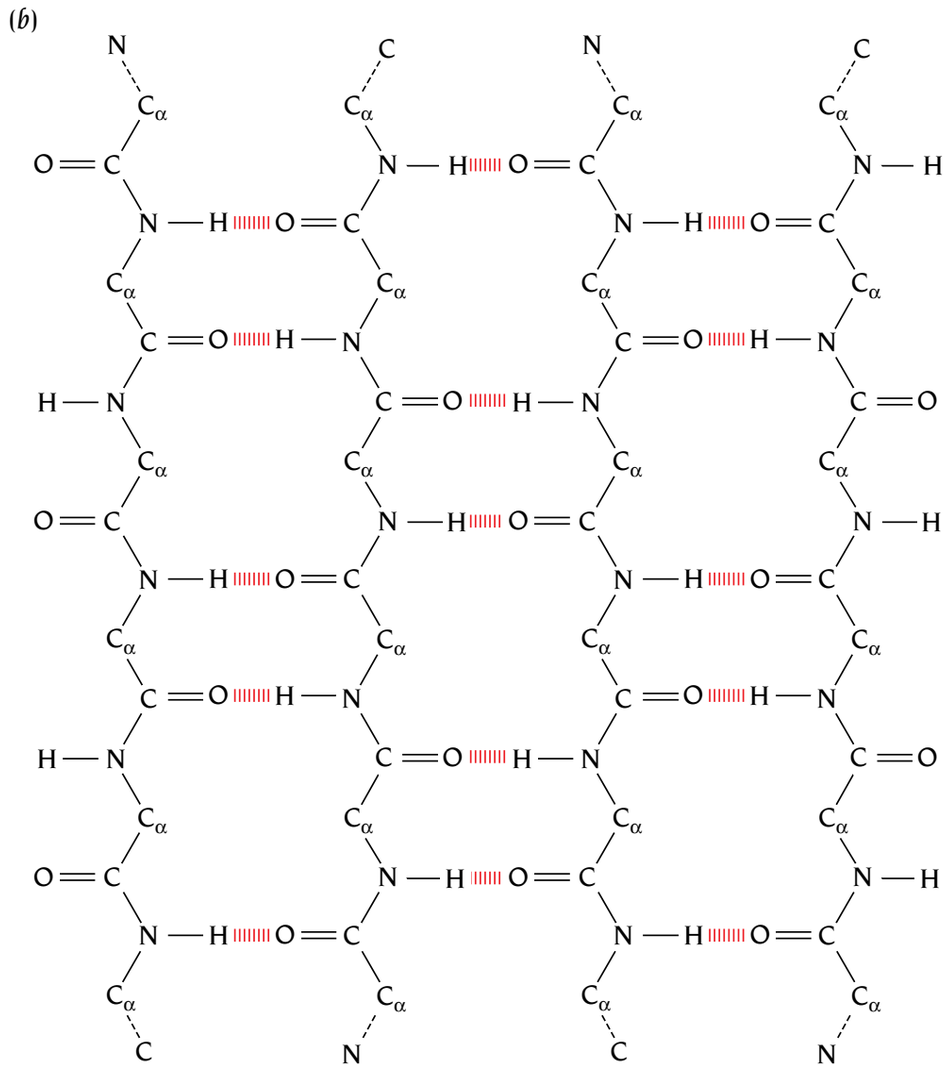
troponin-C

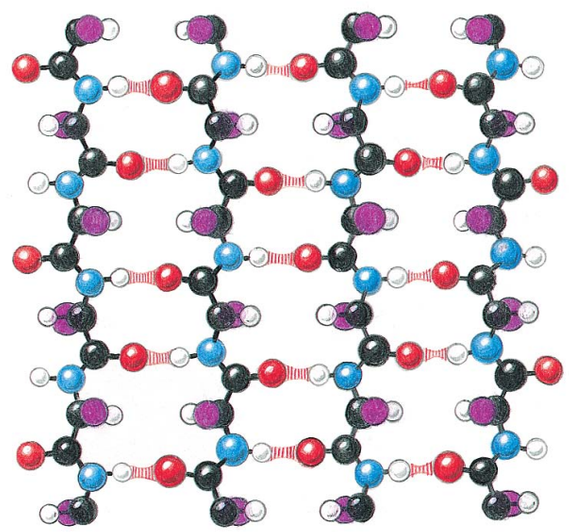
1	2	3	4	5	6	7	8	9	10	11
K	E	D	A	K	G	K	S	E	E	E

Anti-parallel beta sheets

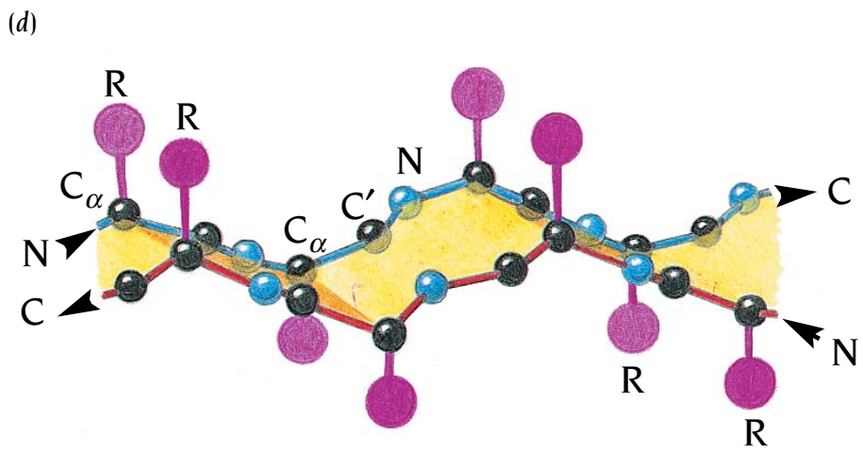






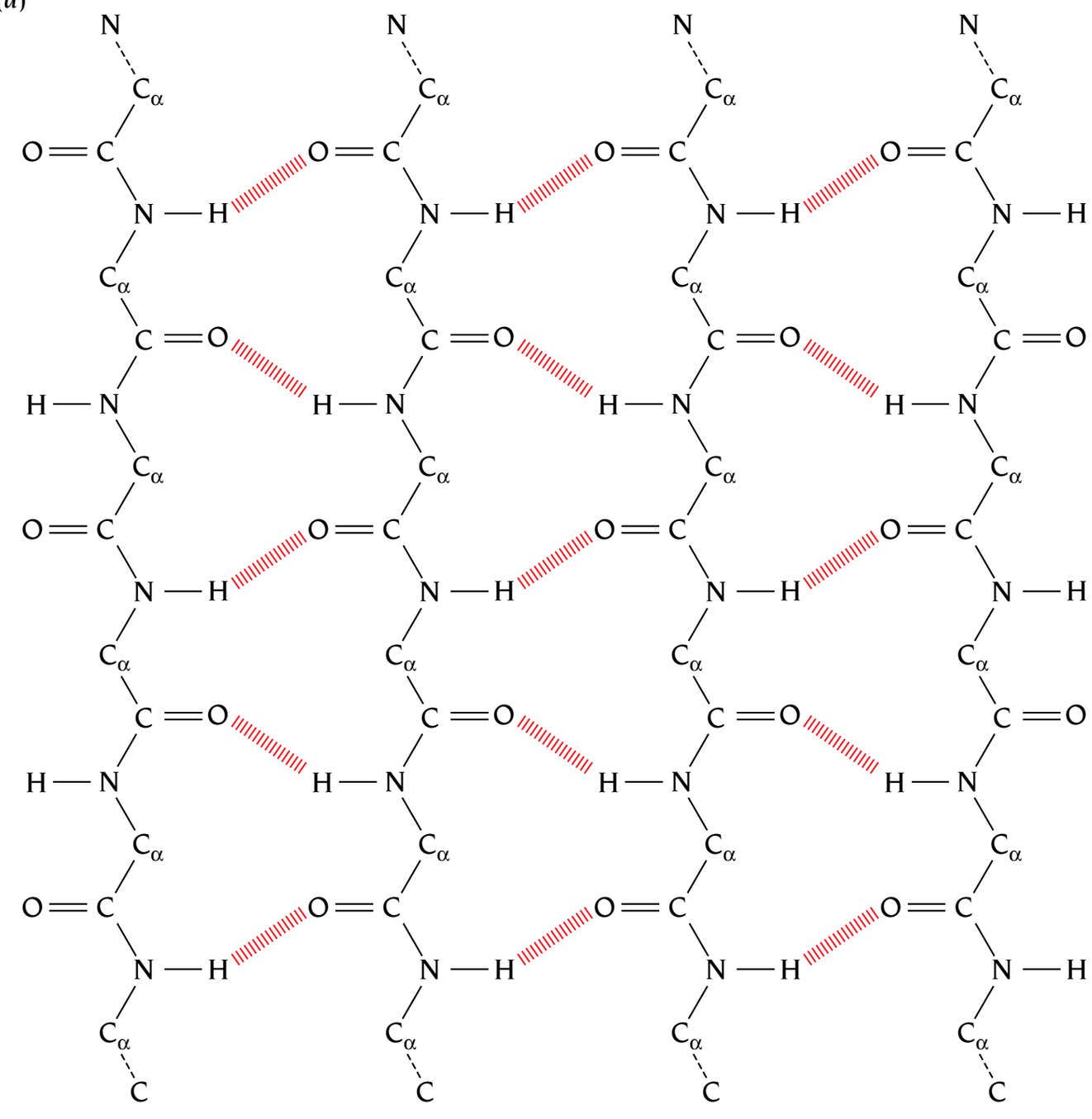


More difficult to design due to non-local nature of backbone H-bonding

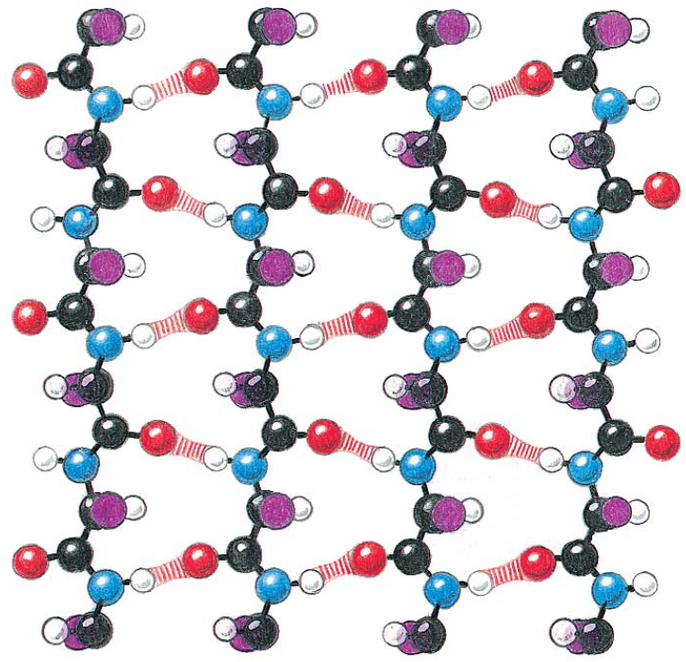


Parallel beta sheets

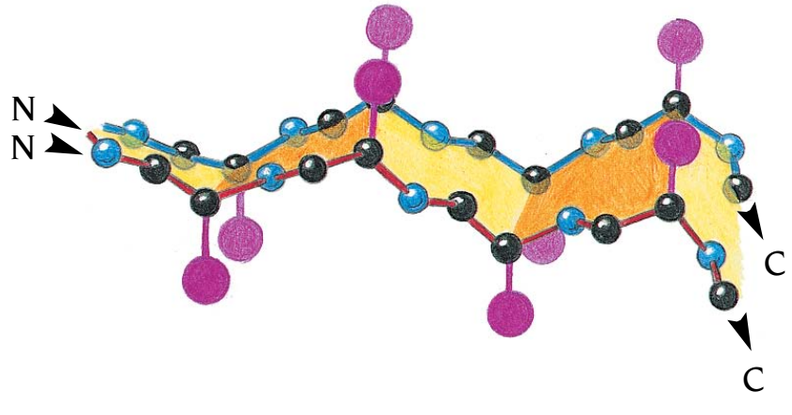
(*a*)



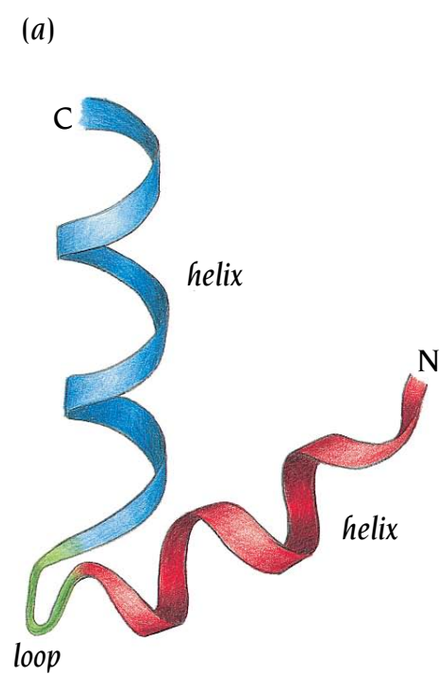
(*b*)



(*c*)



Secondary structure elements are connected to form simple motifs



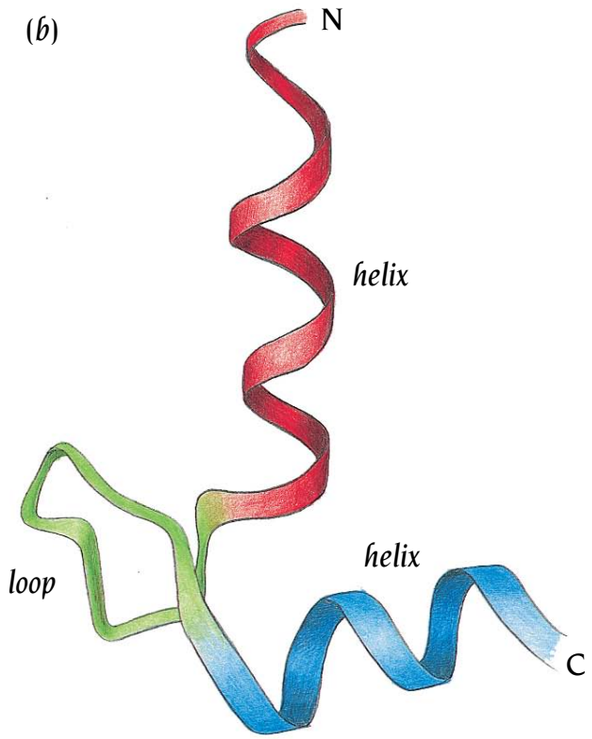
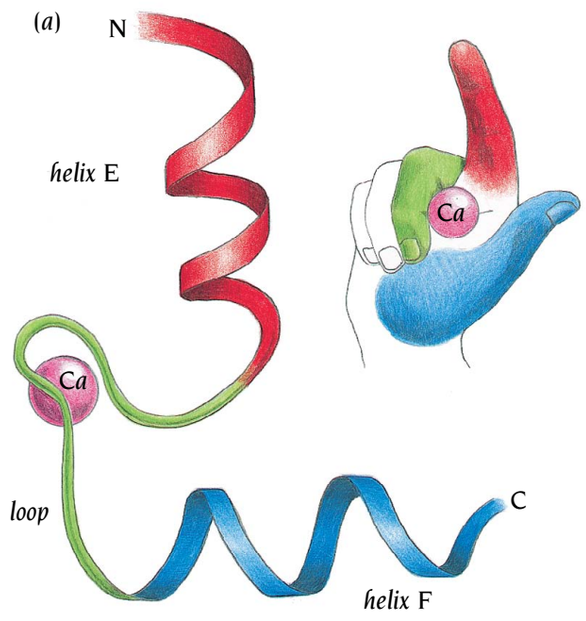
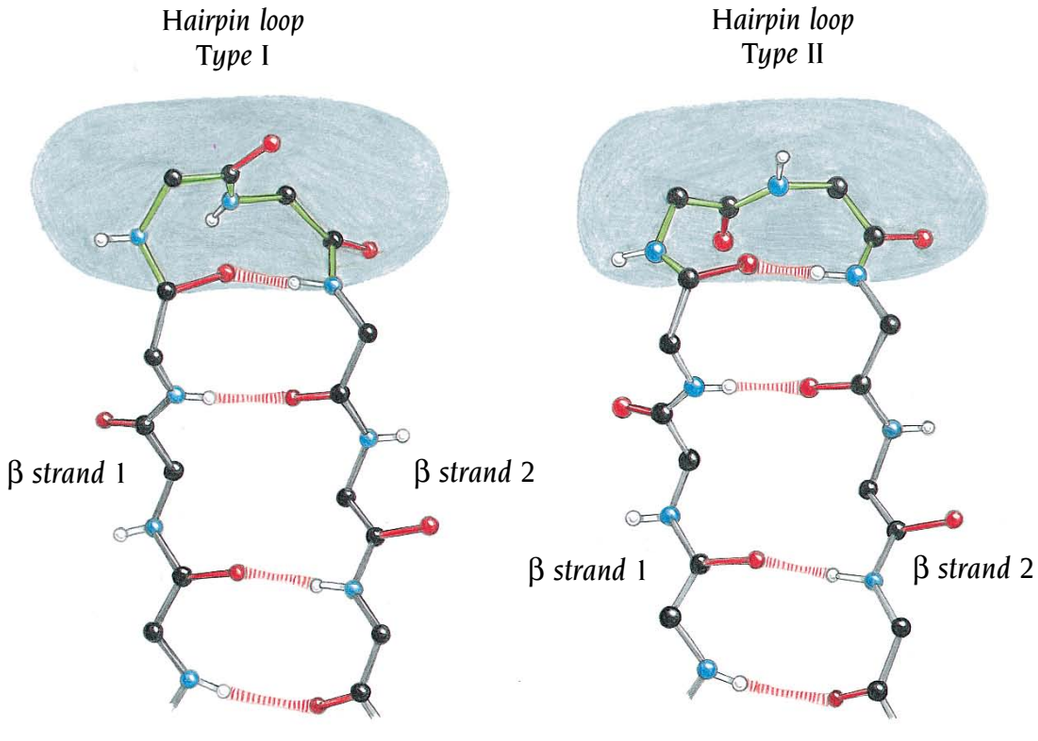


Figure 2.12 Two α helices that are connected by a short loop region in a specific geometric arrangement constitute a helix-turn-helix motif. Two such motifs are shown: the DNA-binding motif (a), which is further discussed in Chapter 8, and the calcium-binding motif (b), which is present in many proteins whose function is regulated by calcium.

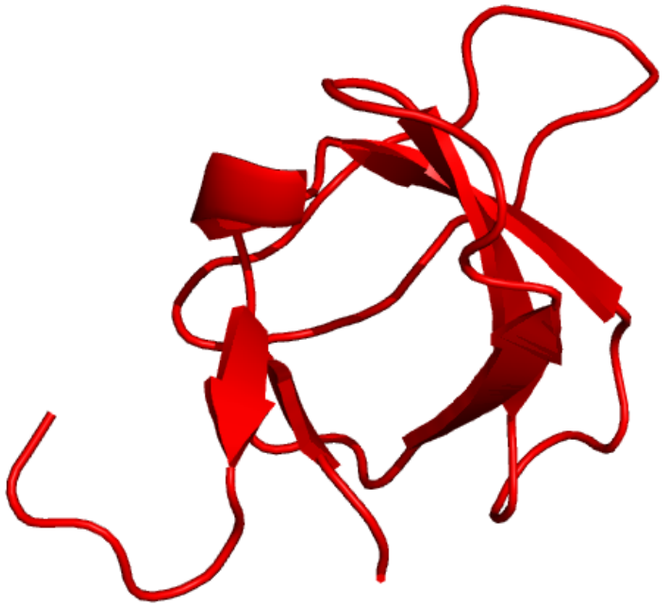




Domains are self-contained “autonomous folding units”

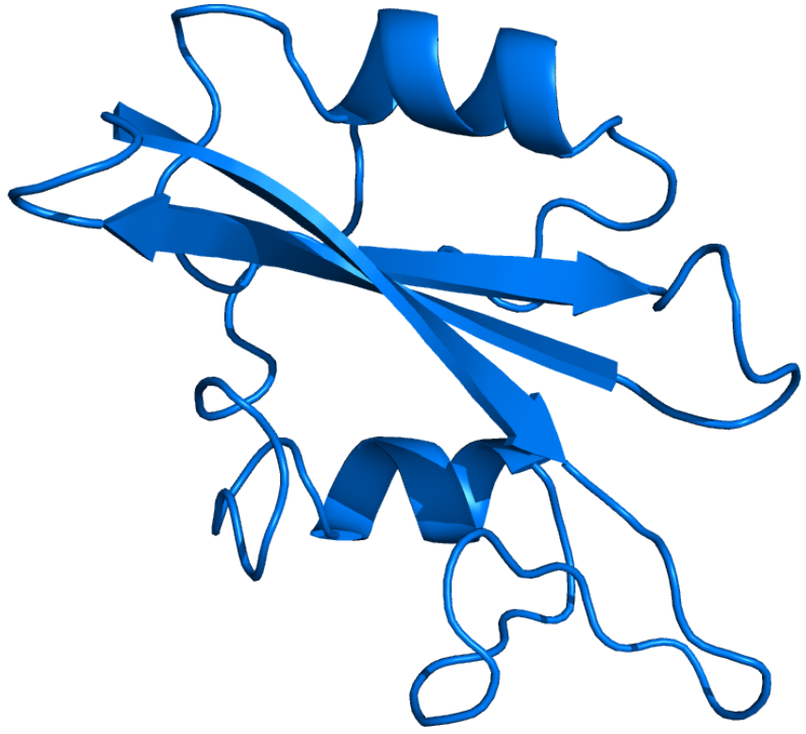
- Size ranges from ~50 – 250 amino acids
- Can comprise a whole protein, or an individual functional module within a larger protein
- Examples: Src-homology (SH) modules

SH3 domain



~60 amino acids
Mostly beta structure

SH2 domain



~100 amino acids
Mixed alpha/beta structure

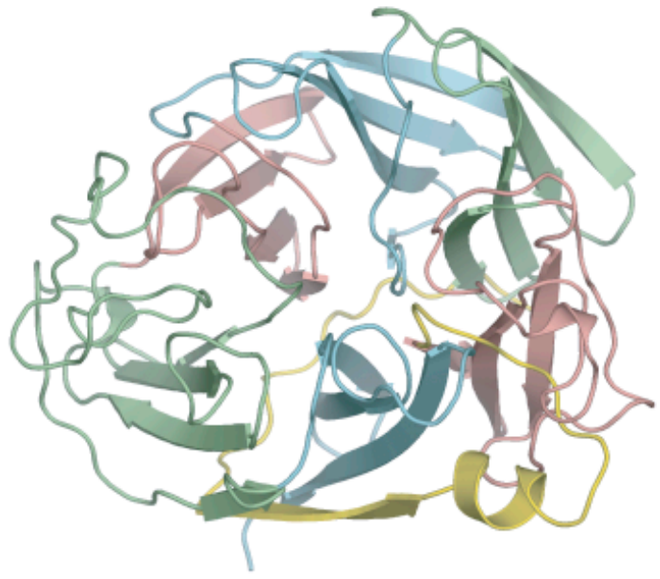
Domains are self-contained “autonomous folding units”

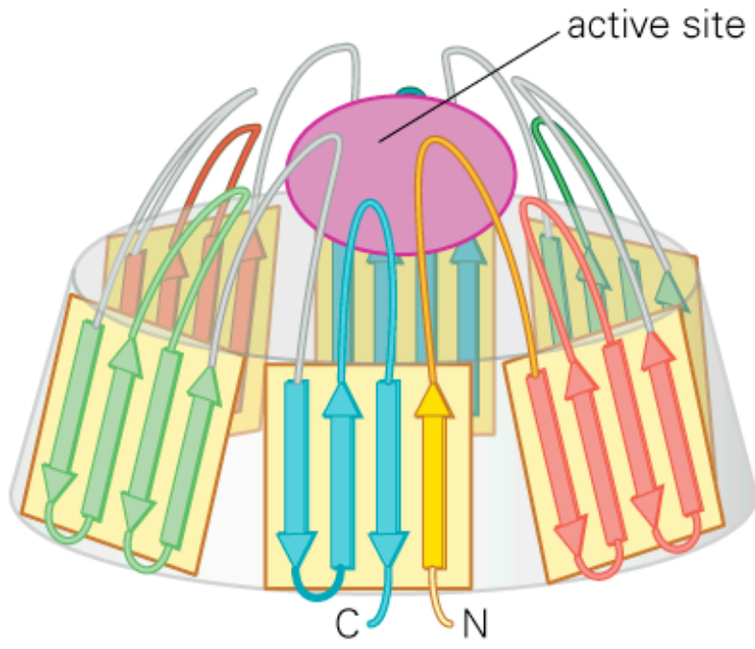
Beta propeller domain (~250 amino acids)

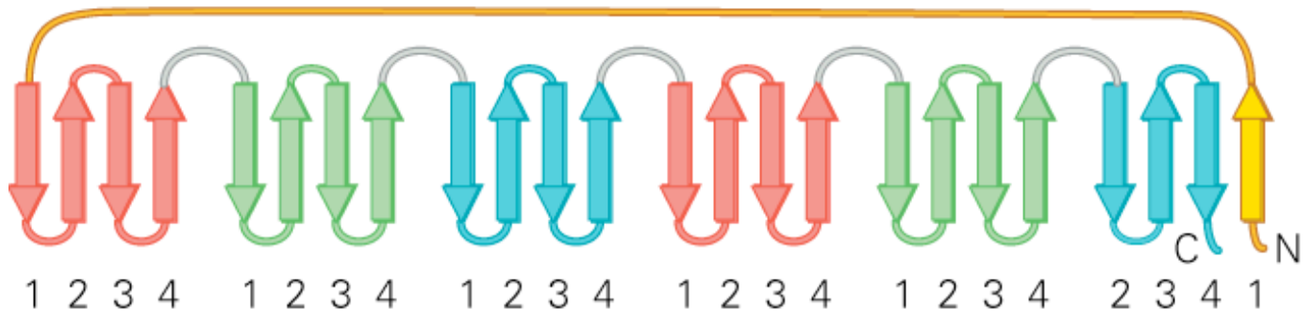
large all beta domain found in receptors (LDL receptor and related proteins), enzymes (influenza neuraminidase, depicted here)

E3 ligase subunits

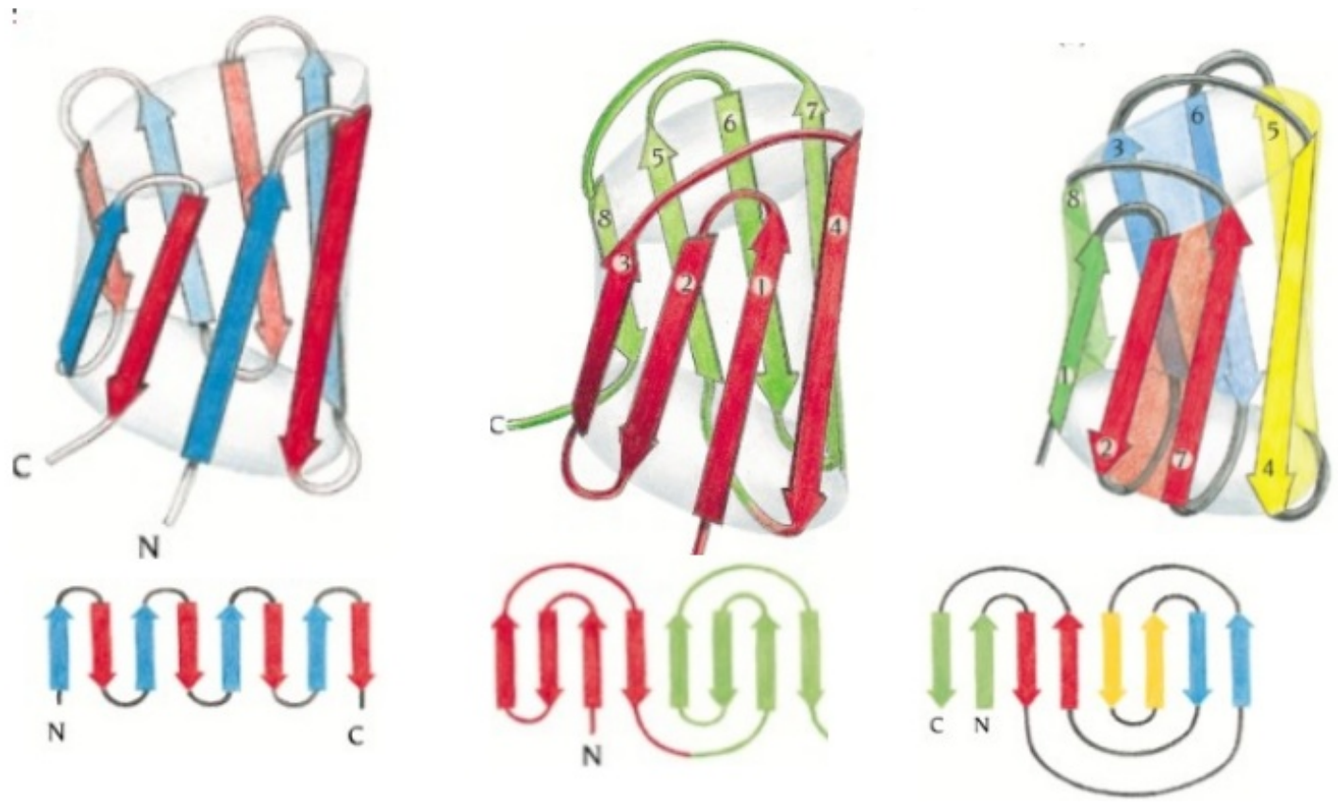
G-protein coupled receptor beta subunits





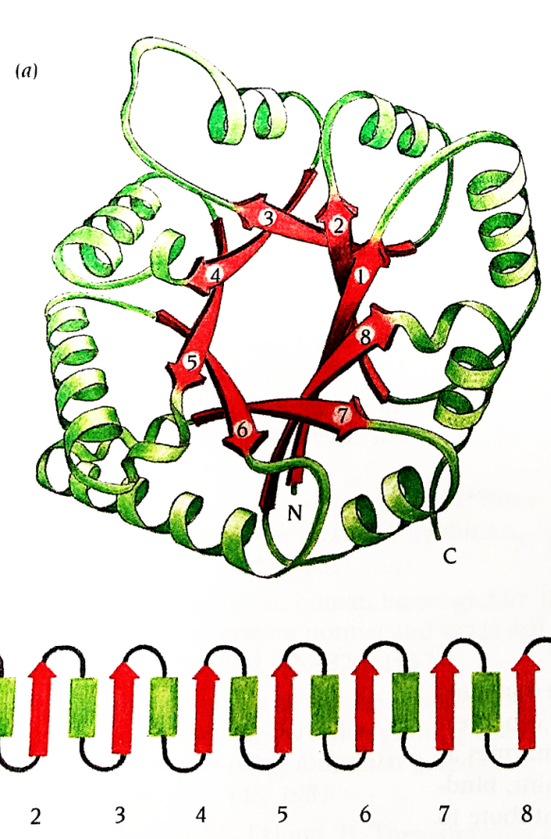


Examples of beta-barrel domains

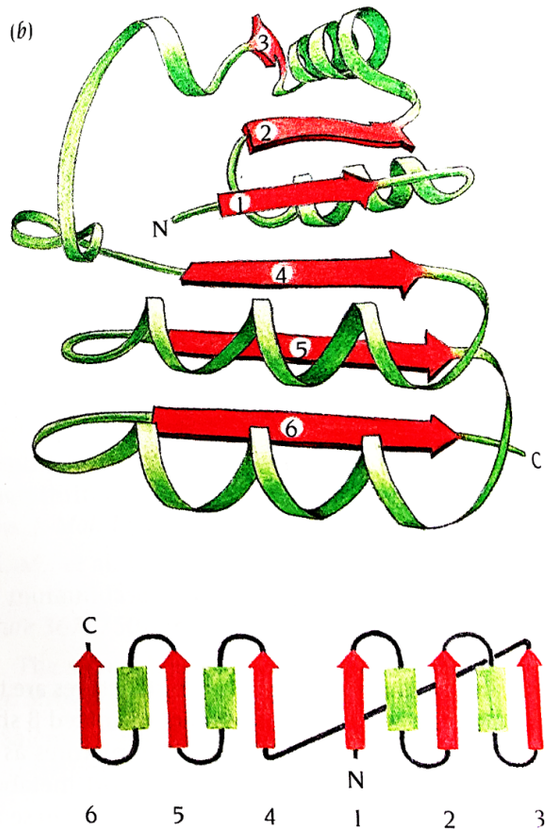


Linear (up and down) Beta-crystallin “jelly-roll”

Topology diagrams of α/β domains



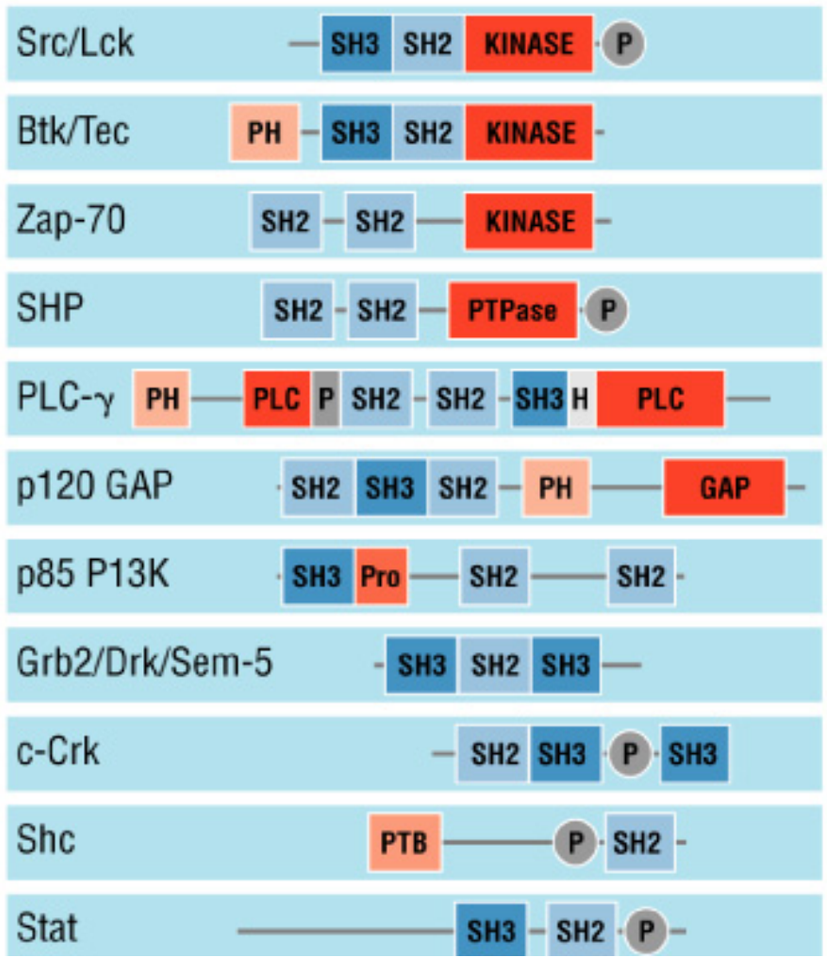
α/β "TIM barrel"



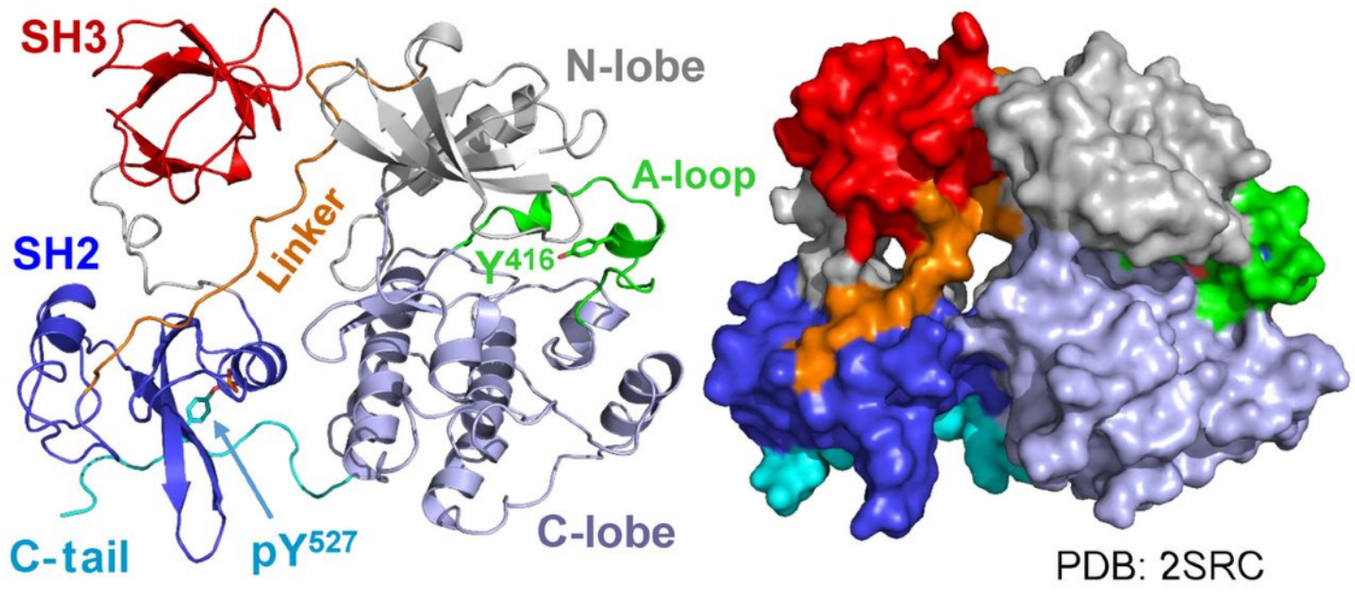
Doubly wound (dehydrogenase enzymes)

Large proteins usually constructed from multiple domains

-consecutive in sequence
-regulatory interdomain interactions frequently occur
-one domain may also interrupt another
-domains can even be “co-folded” from separate chains



Src family kinase domain organization



SCIENCE SIGNALING • 23 Oct 2018 • Vol 11, Issue 553 • DOI: 10.1126/scisignal.aat5916

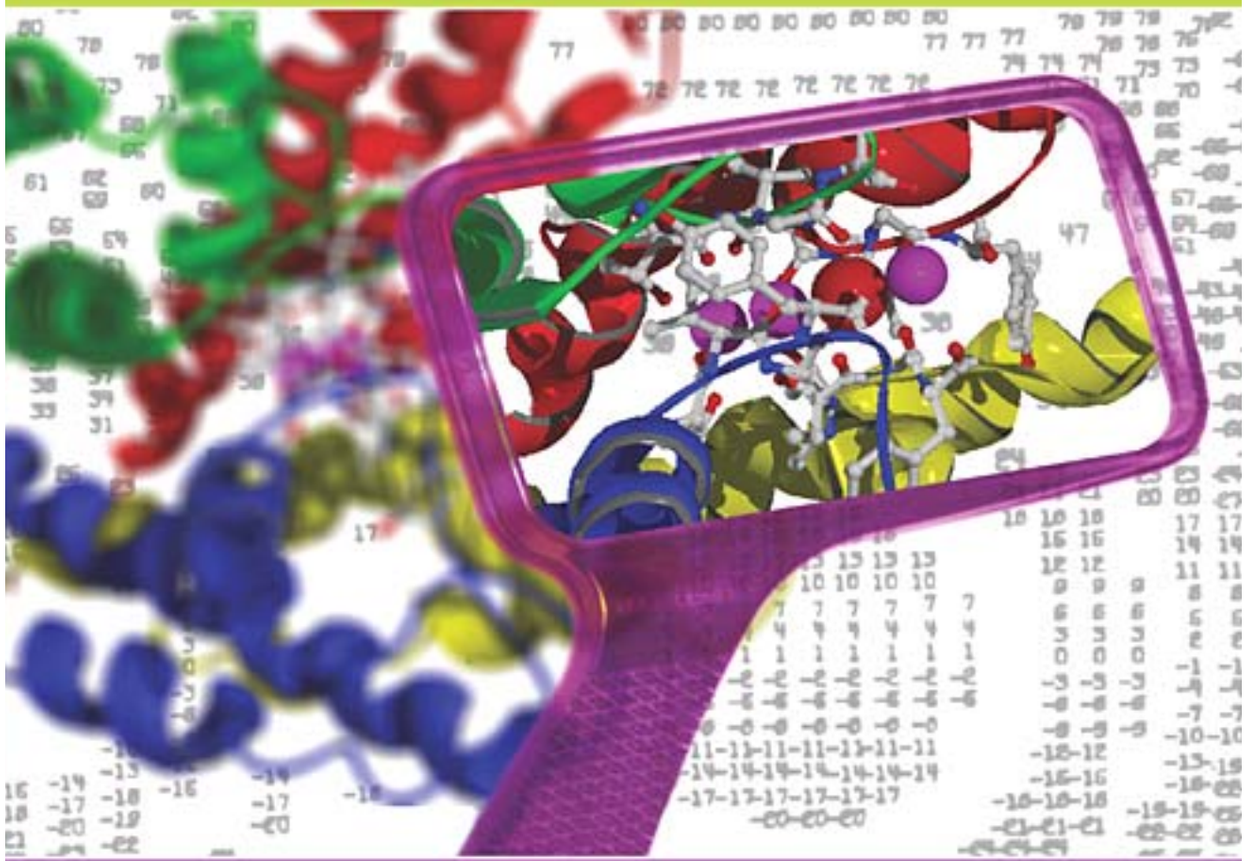
Break

COMPLEMENTARY SCIENCE SERIES

Crystallography Made Crystal Clear

THIRD EDITION

A Guide for Users of Macromolecular Models

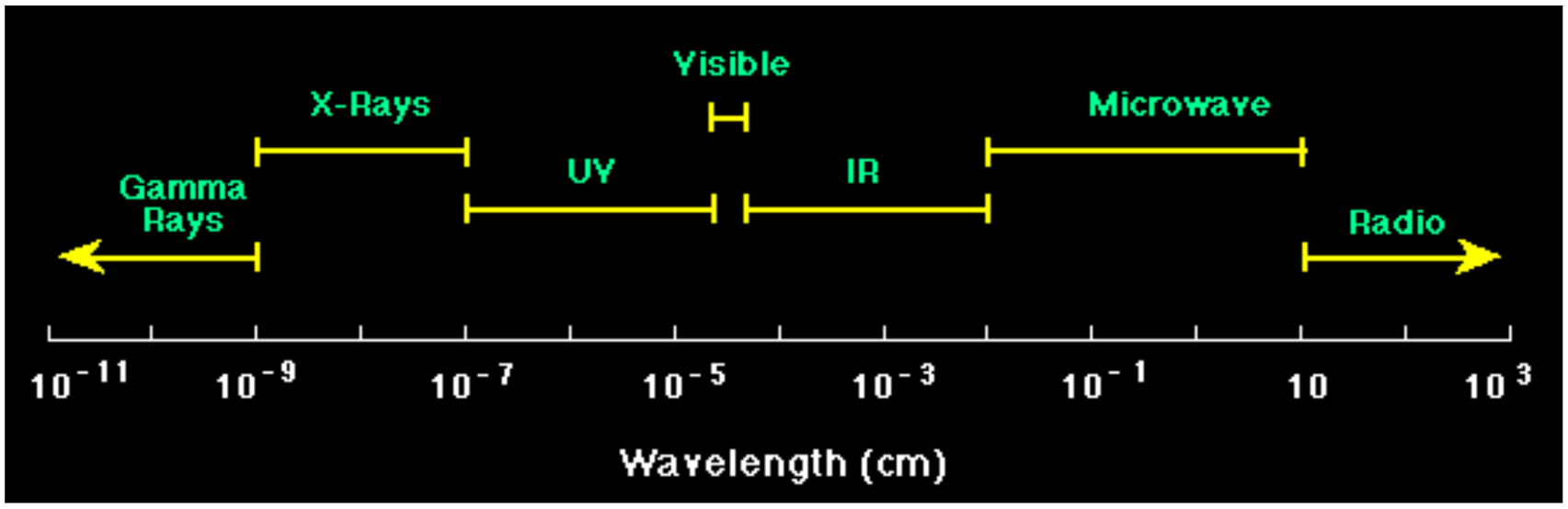


Gale Rhodes

AP

Why ***X-ray*** crystallography?

- Need light of a wavelength that will resolve objects at an atomic scale
- Need signal – crystallization amplifies scattering, even a tiny crystal might contain 10^{12} copies.



Wavelength of X-rays used in crystallography:

~1 Å (10^{-8} cm), typically from synchrotron radiation

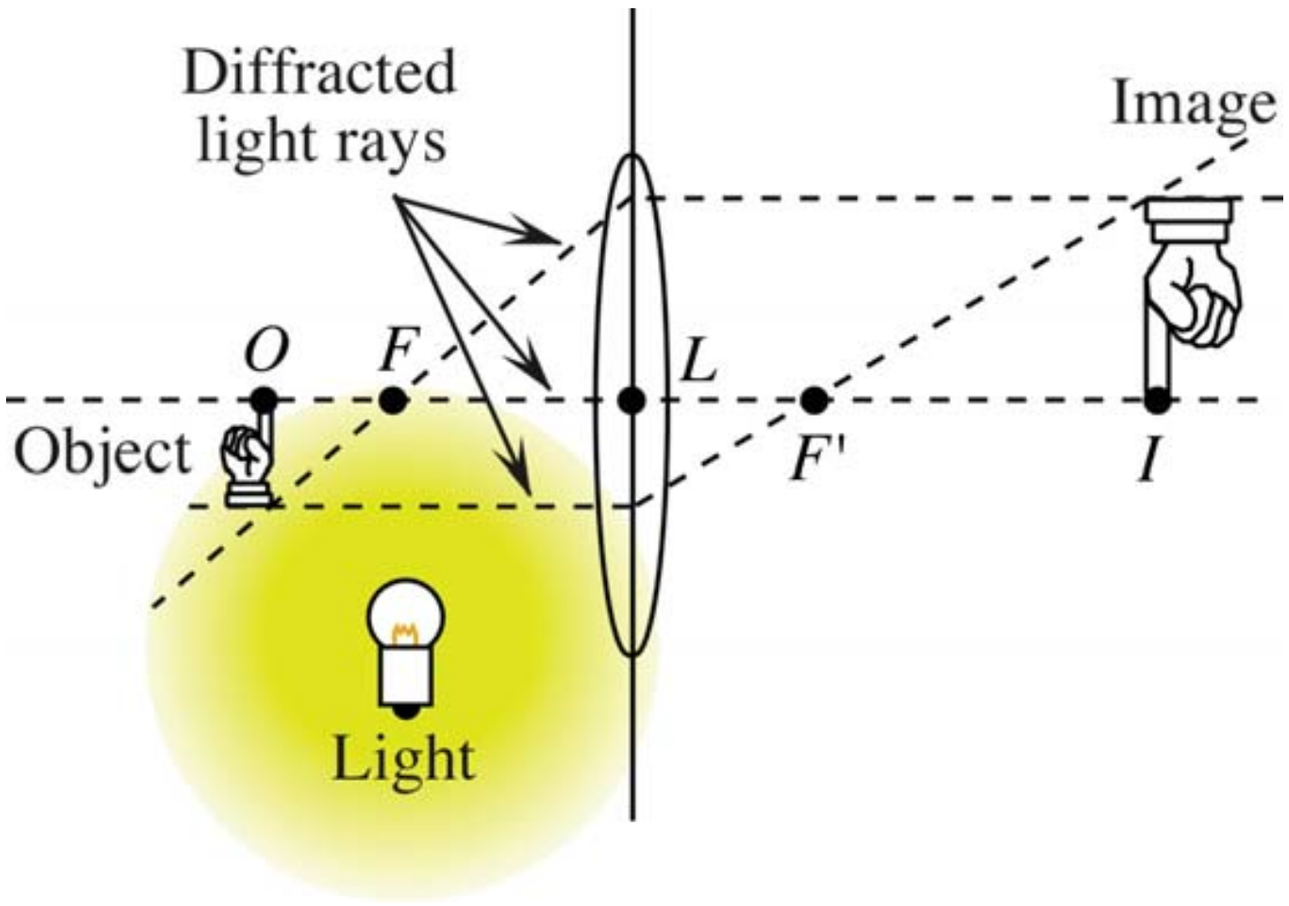


Figure 2.1 ► Action of a simple lens. Rays parallel to the lens axis strike the lens and are refracted into paths passing through a focus (F or F'). Rays passing through a focus strike the lens and are refracted into paths parallel to the lens axis. As a result, the lens produces an image at I of an object at O such that $(OF)(IF') = (FL)(F'L)$.

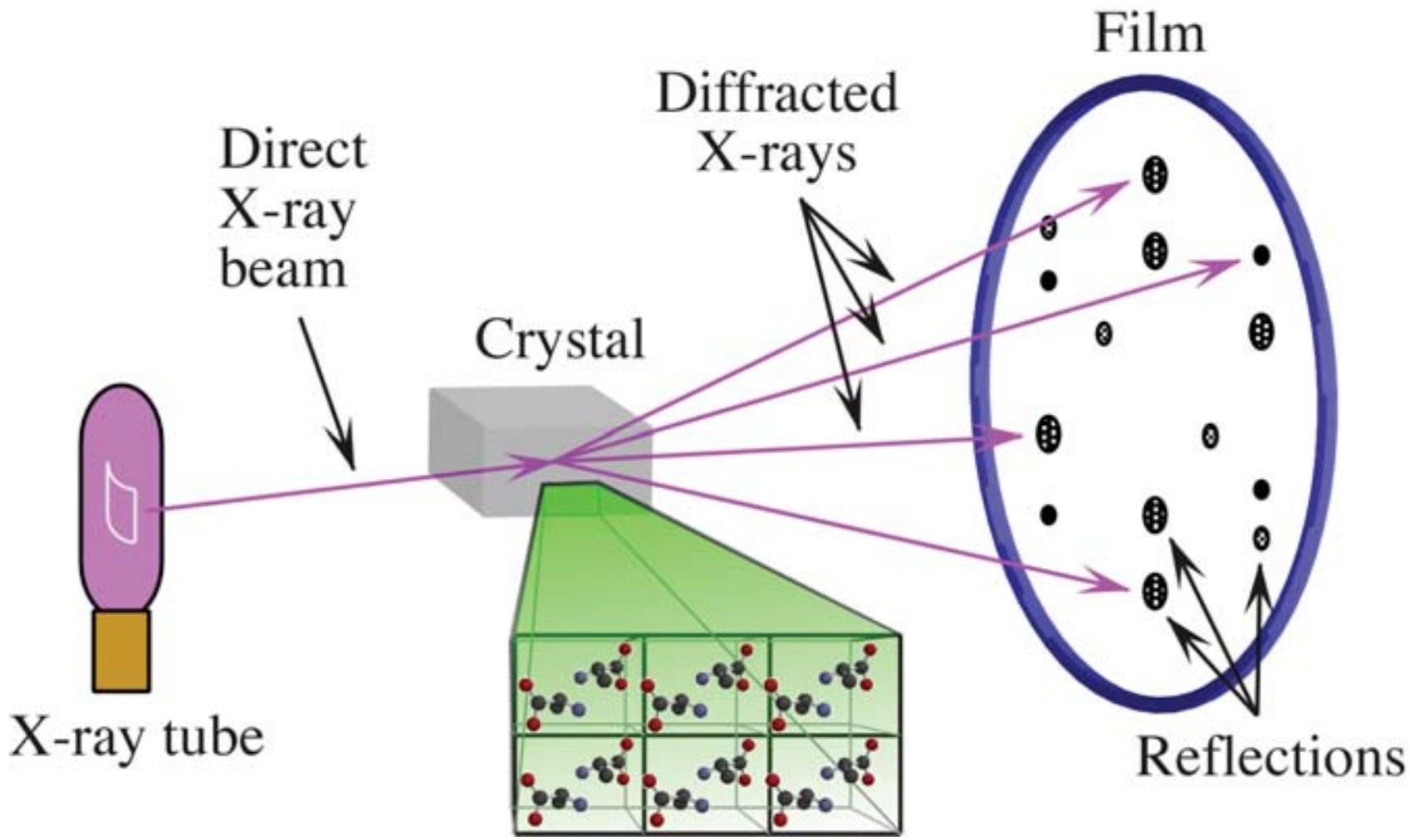


Figure 2.6 ► Crystallographic data collection. The crystal diffracts the source beam into many discrete beams, each of which produces a distinct spot (reflection) on the film. The positions and intensities of these reflections contain the information needed to determine molecular structures.

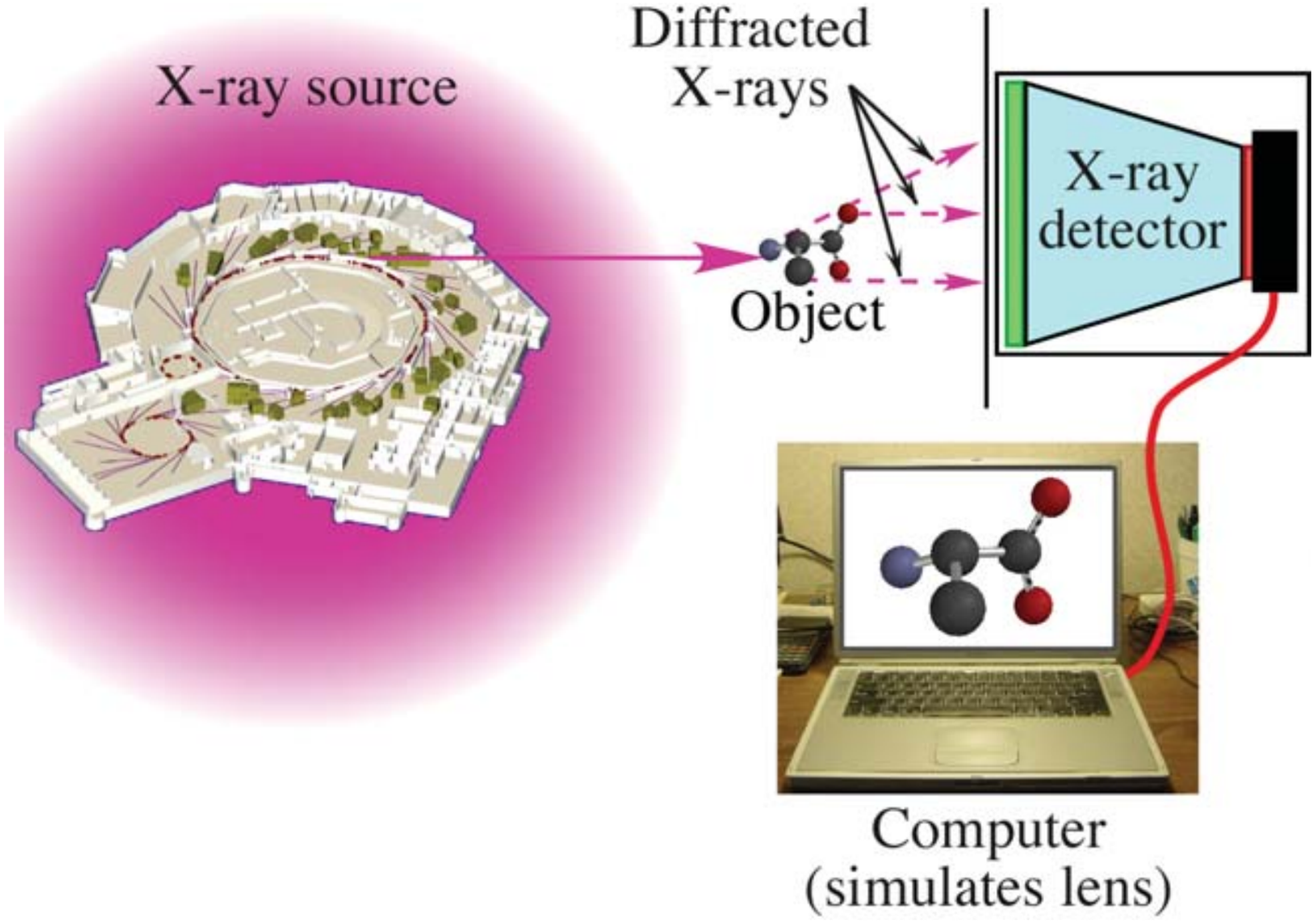
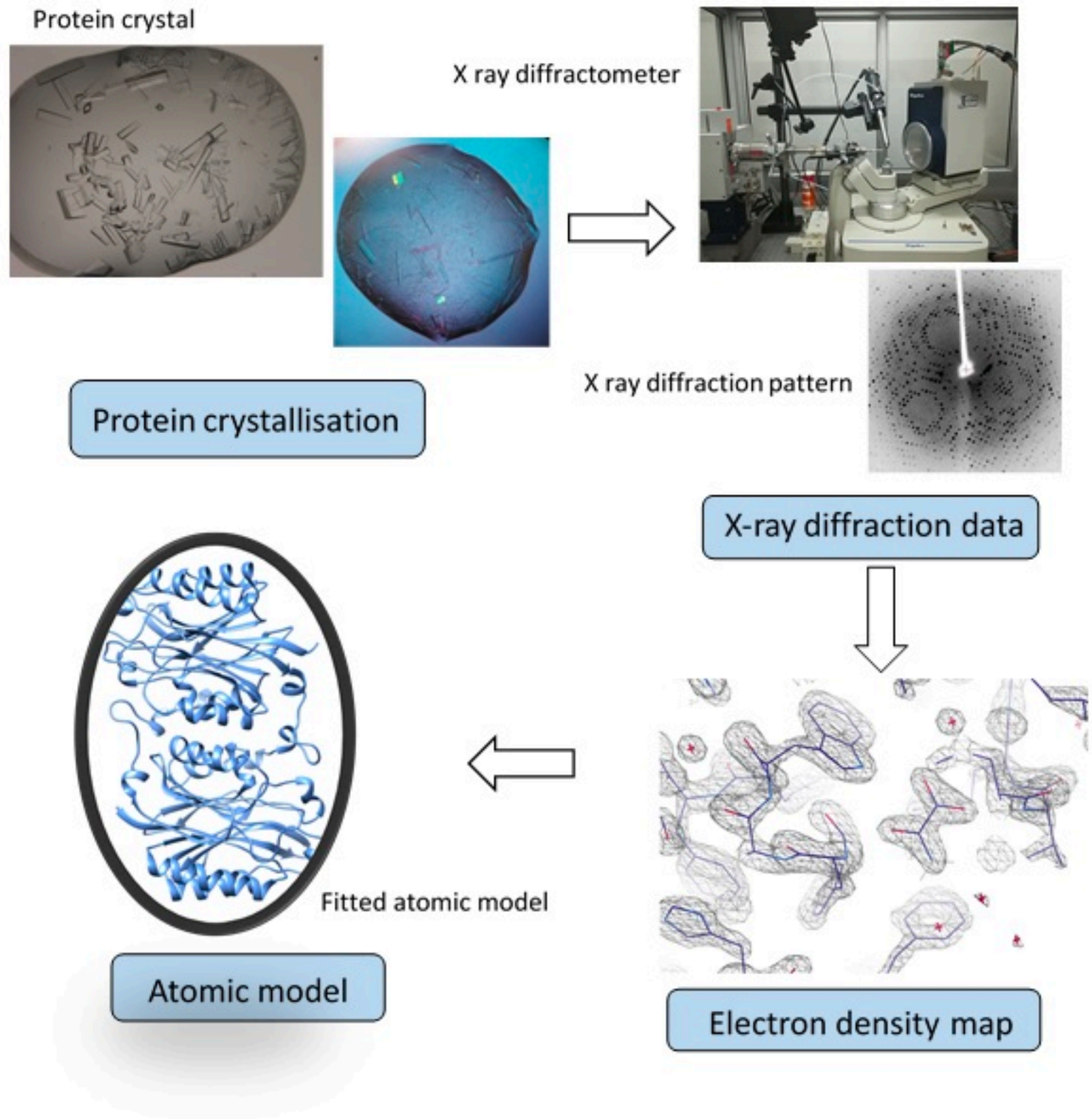


Figure 2.2 ► Crystallographic analogy of lens action. X-rays diffracted from the object are received and measured by a detector. The measurements are fed to a computer, which simulates the action of a lens to produce a graphics image of the object. Compare Fig. 2.2 with Fig. 2.1 and you will see that to magnify molecules, you merely have to replace the light bulb with a synchrotron X-ray source (175 feet in diameter), replace the glass lens with the equivalent of a 5- to 10-megapixel camera, and connect the camera output to a computer running some of the world's most complex and sophisticated software. Oh, yes, and you will need to spend somewhere between a few days and the rest of your life getting your favorite protein to form satisfactory crystals. No, it's not quite as simple as microscopy.

Crystallography workflow



https://ebe.uct.ac.za/emu/instruments-services/x-ray-crystallography

Interpreting the data

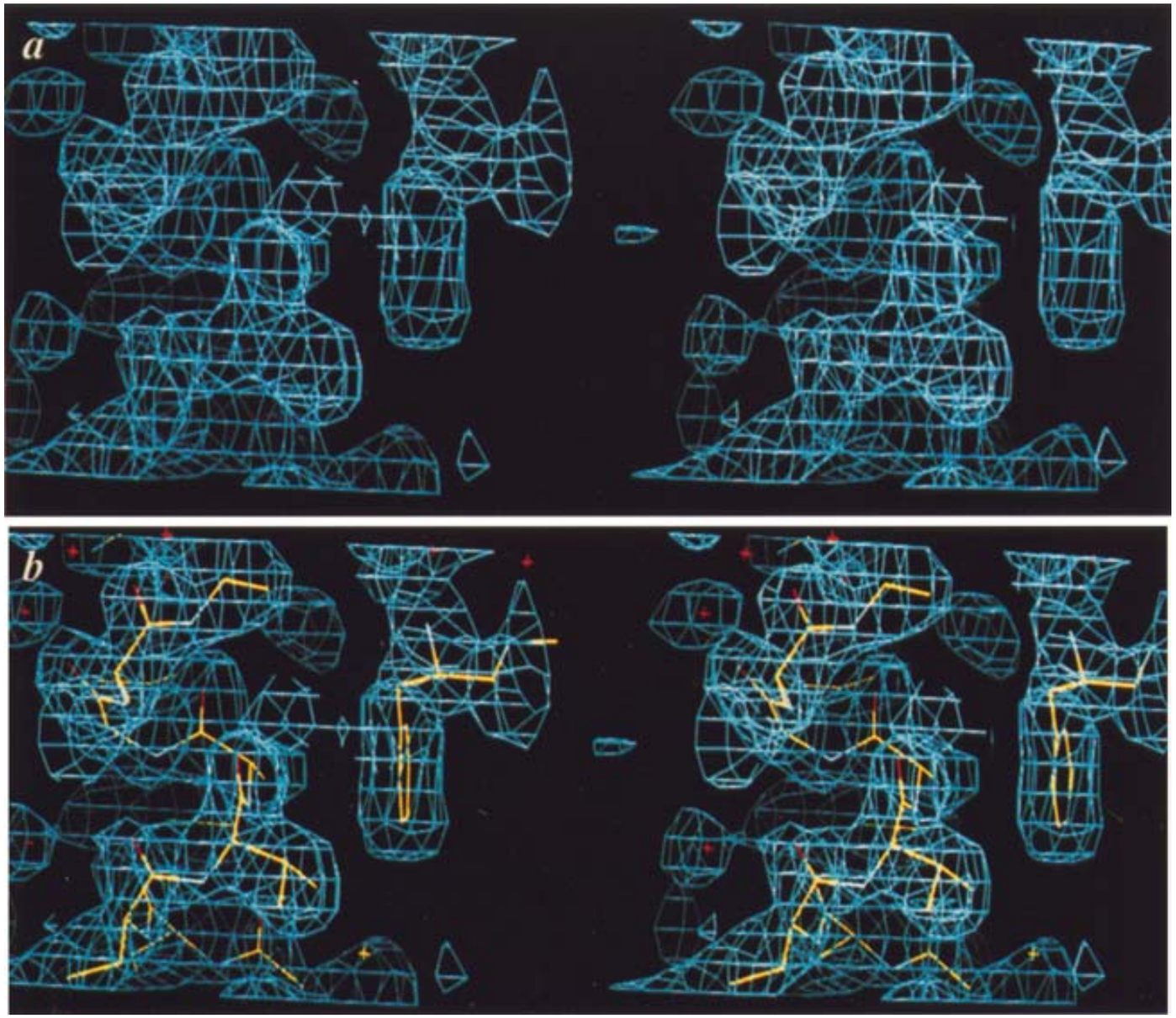
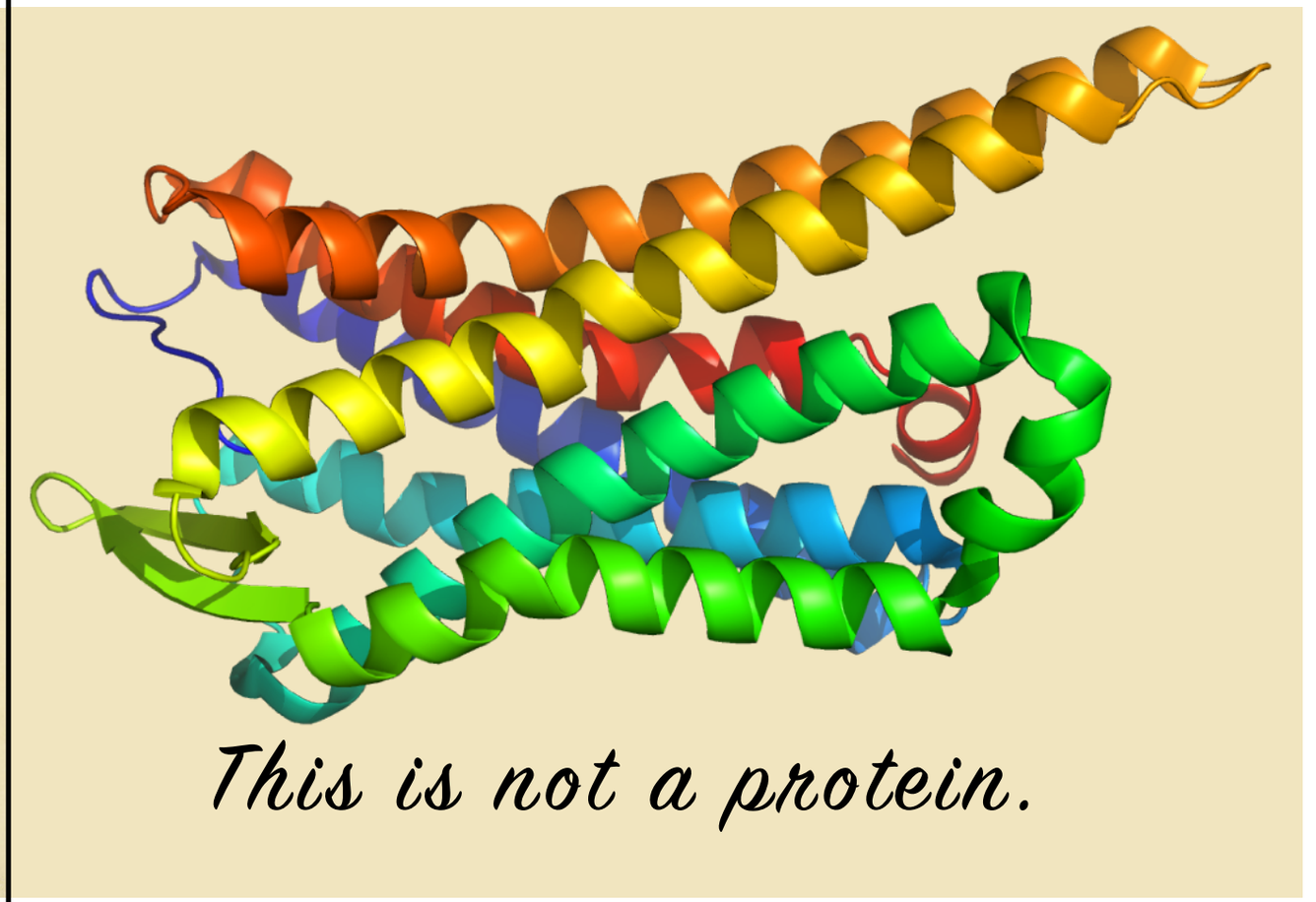
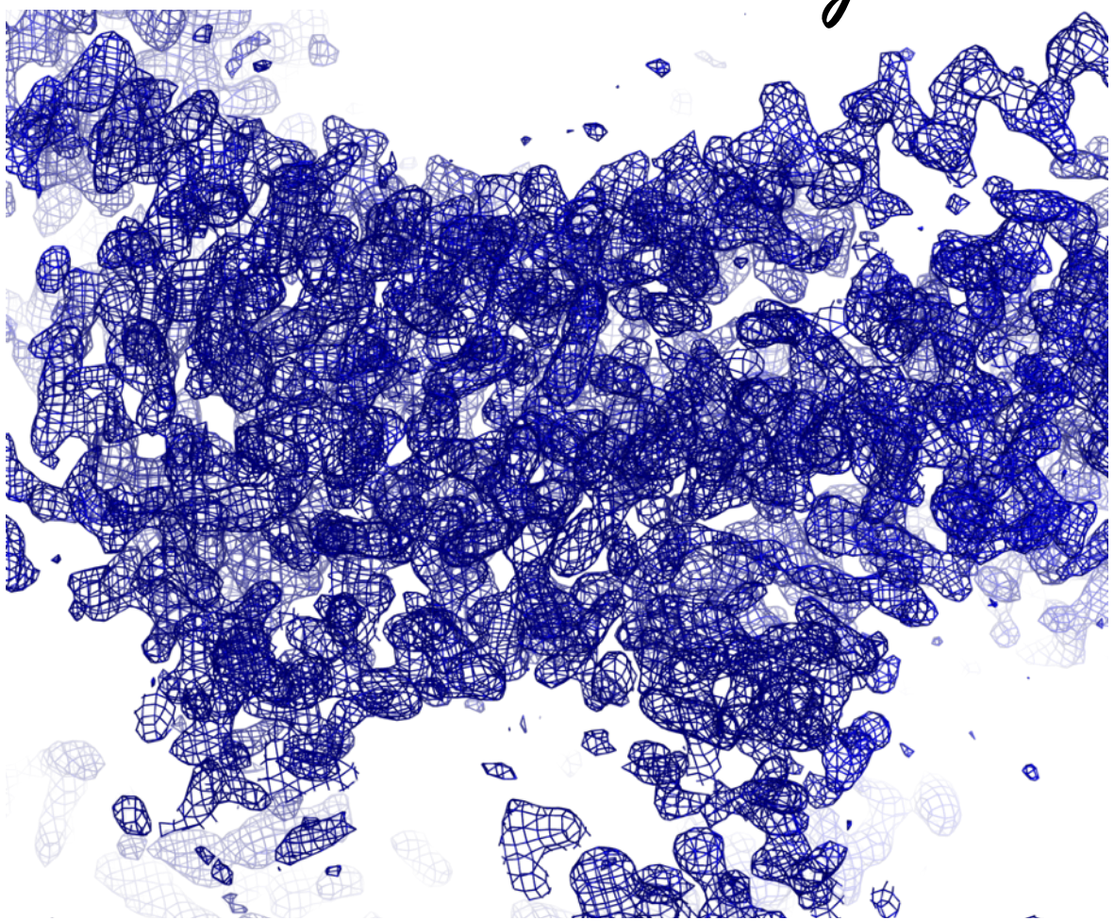


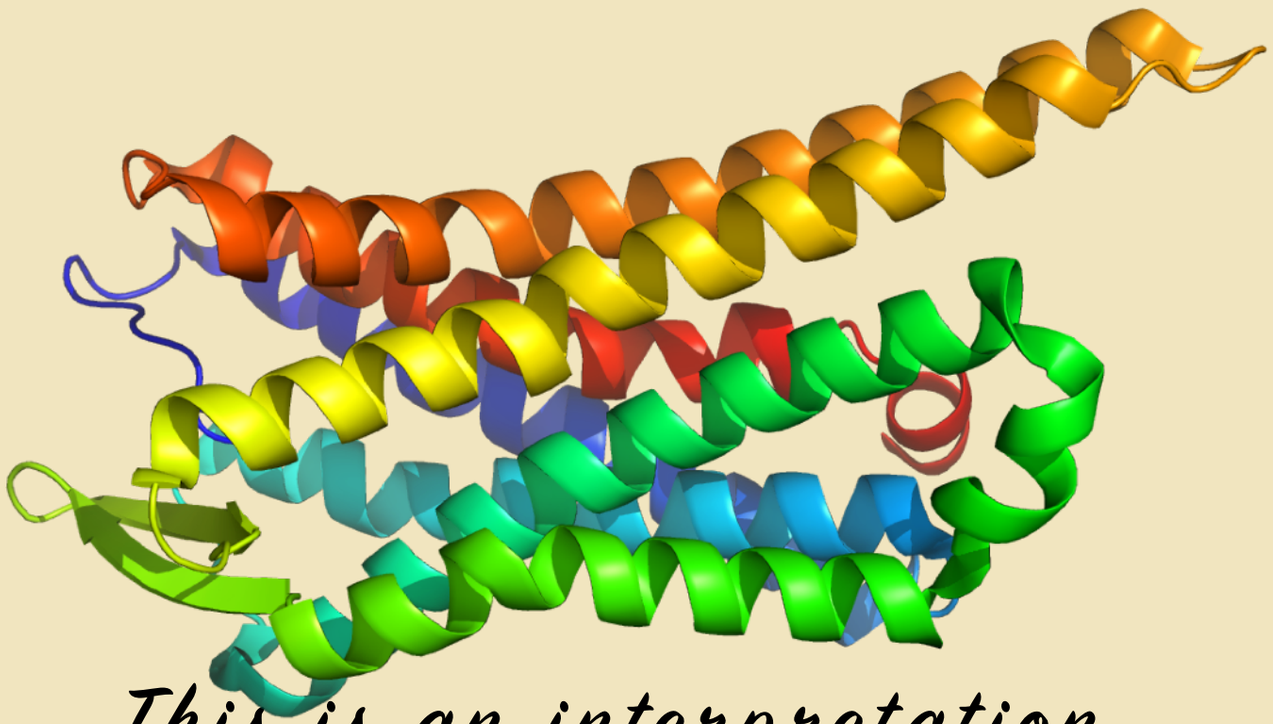
Figure 2.3 ► (*a*) Small section of molecular image displayed on a computer. (*b*) Image (*a*) is *interpreted* by building a molecular model to fit within the image. Computer graphics programs allow the crystallographer to add parts to the model and adjust their positions and conformations to fit the image. The protein shown here is adipocyte lipid binding protein (ALBP, PDB 1alb).

The Treachery of Images
Artist: René Magritte



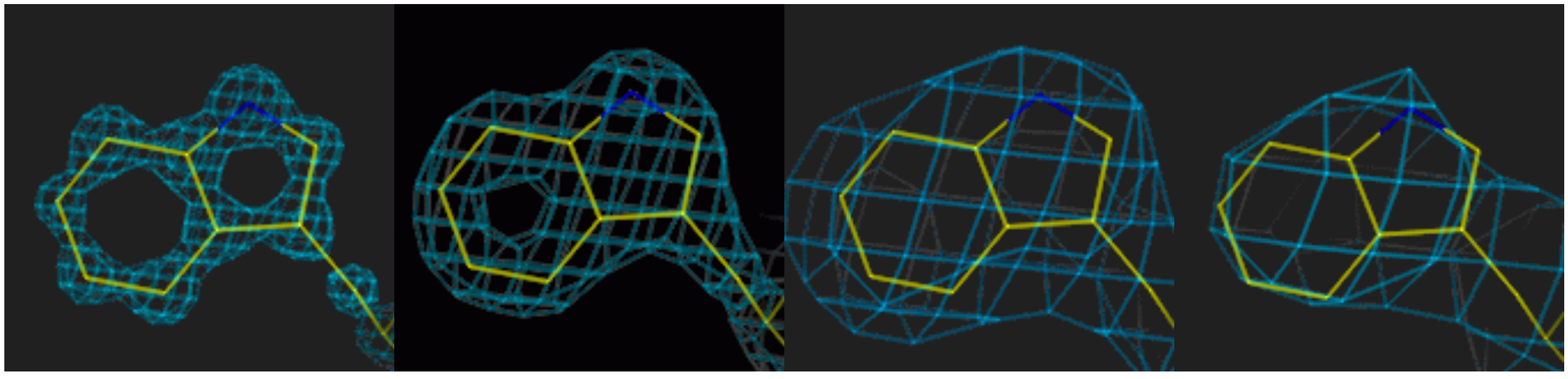
electron density





This is an interpretation of the density

Concept of resolution



1 Angstrom 2.5 Angstrom 3 Angstrom 4 Angstrom

http://www-structmed.cimr.cam.ac.uk/Course/Fitting/fittingtalk.html#resolution

Low (or poor) resolution can give overall fold or shape

In microscopy, the phrase "resolution of 2 Å," implies that we can resolve objects that are 2 Å apart. If this phrase had the same meaning for a crystallographic model of a protein, in which bond distances average about 1.5 Å, we would be unable to distinguish or resolve adjacent atoms in a 2-Å map. Actually, for a protein refined at 2-Å resolution to an R-factor near 0.2, the situation is much better than the resolution statement seems to imply…

Although the final 2-Å map, viewed as an empty contour surface, may indeed not allow us to discern adjacent atoms, prior knowledge in the form of structural constraints on the model greatly increase the precision of atom positions. The main constraint is that we know we can fit the map with *groups* of atoms— amino-acid residues—having known connectivities, bond lengths, bond angles, and stereochemistry.

In crystallography, unlike microscopy, the term *resolution* simply refers to the amount of data ultimately phased and used in the structure determination. In contrast, the precision of atom positions depends in part upon the resolution limits of the data, but also depends critically upon the quality of the data, as reflected by such parameters as R-factors. **Good data can yield atom positions that are precise to within one-fifth to one-tenth of the stated resolution**.

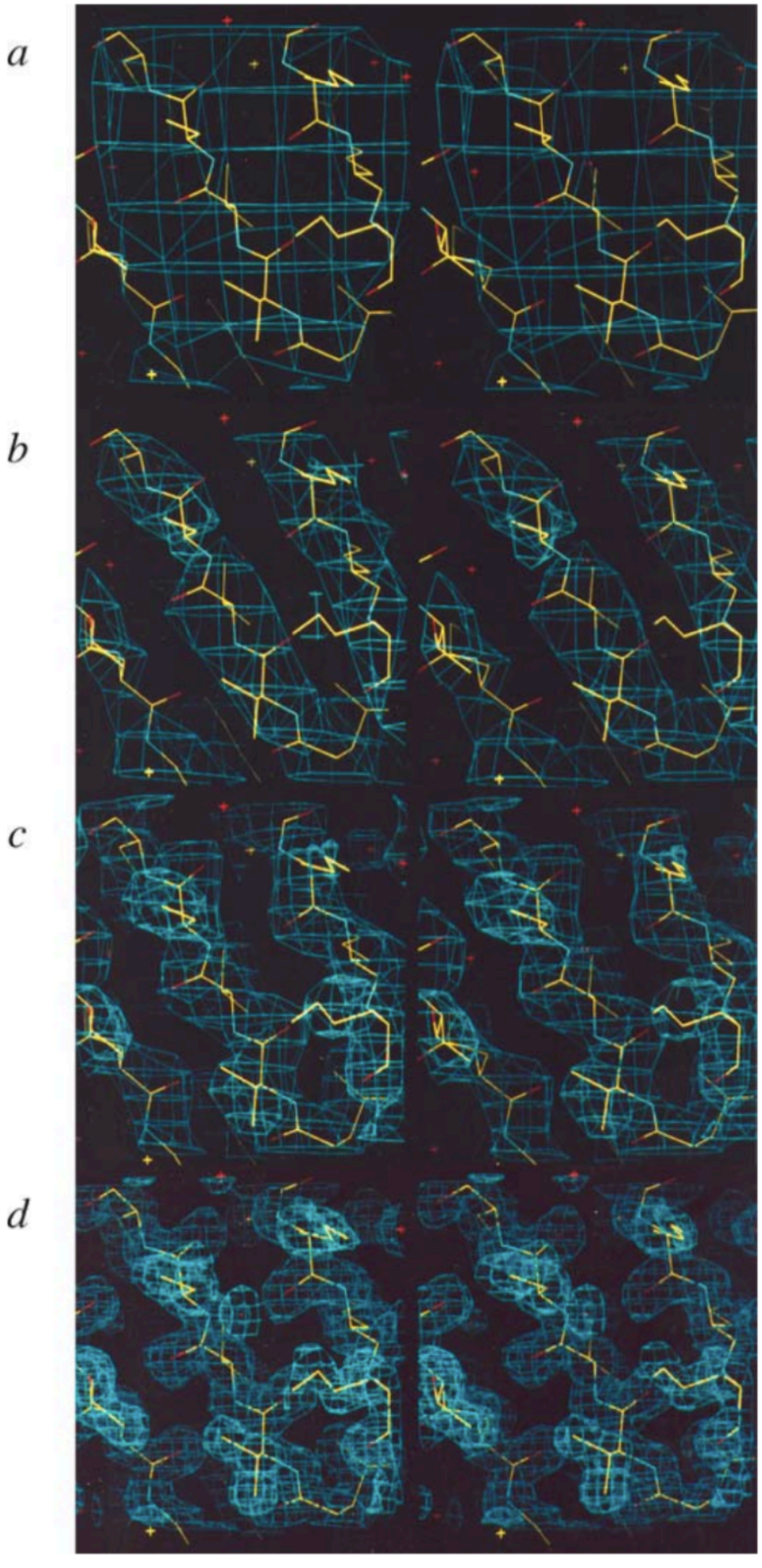
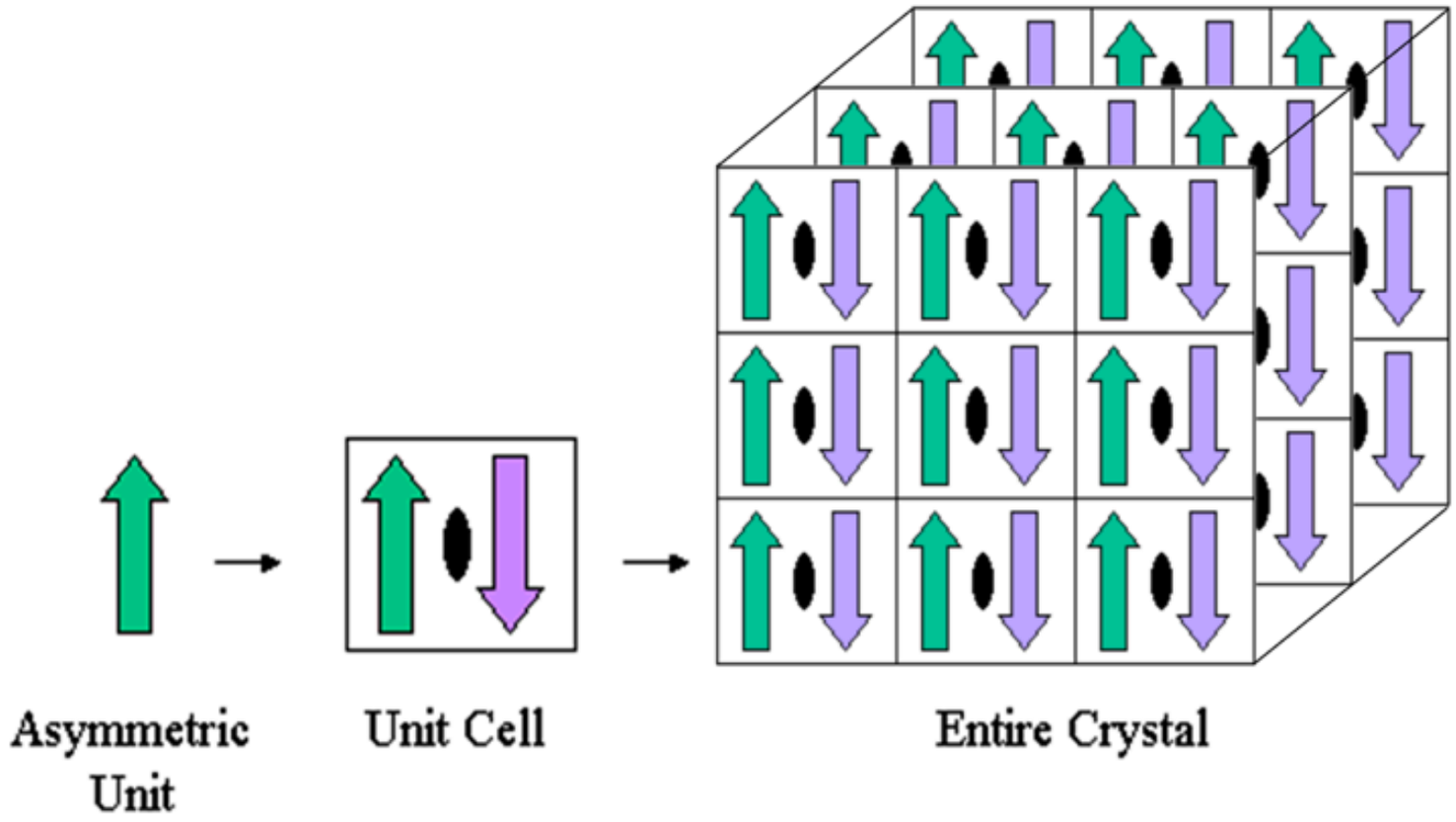


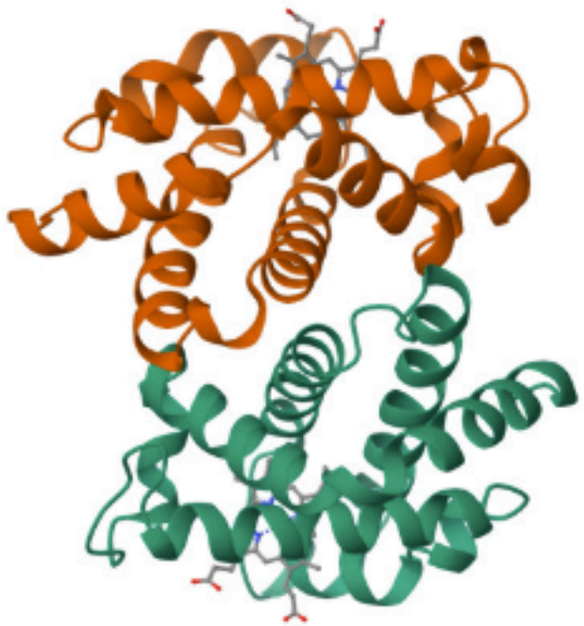
Figure 7.5 ► Electron-density maps at increasing resolution (stereo). Maps were calculated using final phases, and Fourier sums were truncated at the the following resolution limits: (*a*) 6.0 A; (*b*) 4.5 A; (*c*) 3.0 A; (*d*) 1.6 A.

Describing the crystal with smaller repeating units



https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/biological-assemblies

Asymmetric Unit



Explore in 3D: Structure | Sequence Annotations | Validation Report | Ligand Interaction (HEM)

1OUT

TROUT HEMOGLOBIN I

PDB DOI: https://doi.org/10.2210/pdb1OUT/pdb

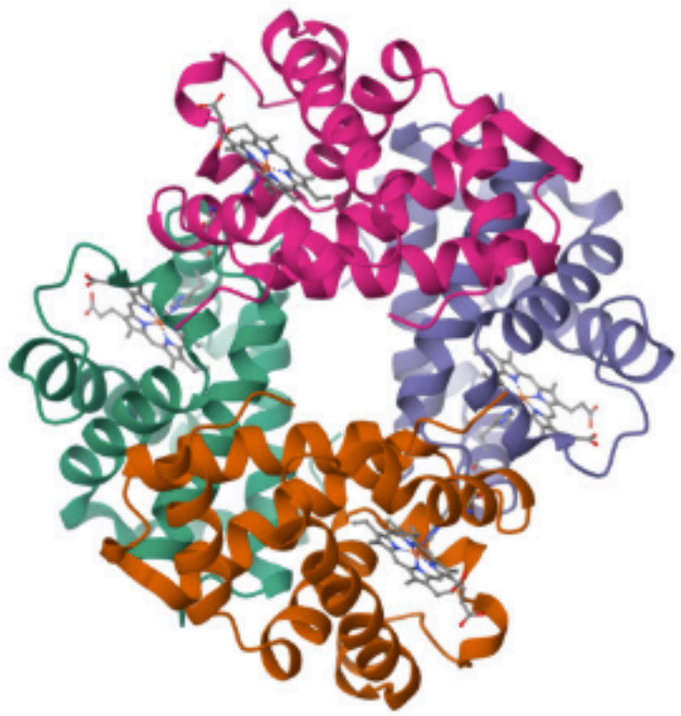
Classification: OXYGEN TRANSPORT
Organism(s): Oncorhynchus mykiss
Mutation(s): No

Deposited: 1996-06-21 **Released:** 1997-01-11
Deposition Author(s): Tame, J., Wilson, J.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.30 Å
R-Value Free: 0.247
R-Value Work: 0.162

Biological Assembly 1



Explore in 3D: Structure | Sequence Annotations | Validation Report | Ligand Interaction (HEM)

Global Symmetry: Cyclic - C2 (Explore in 3D)
Global Stoichiometry: Hetero 4-mer - A2B2

Pseudo Symmetry: Dihedral - D2 (Explore in 3D)
Pseudo Stoichiometry: Homo 4-mer - A4

Find Similar Assemblies

Biological assembly 1 assigned by authors.

1OUT

TROUT HEMOGLOBIN I

PDB DOI: https://doi.org/10.2210/pdb1OUT/pdb

Classification: OXYGEN TRANSPORT
Organism(s): Oncorhynchus mykiss
Mutation(s): No

Deposited: 1996-06-21 **Released:** 1997-01-11
Deposition Author(s): Tame, J., Wilson, J.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.30 Å
R-Value Free: 0.247
R-Value Work: 0.162

This is version 1.2 of the entry. See complete histo

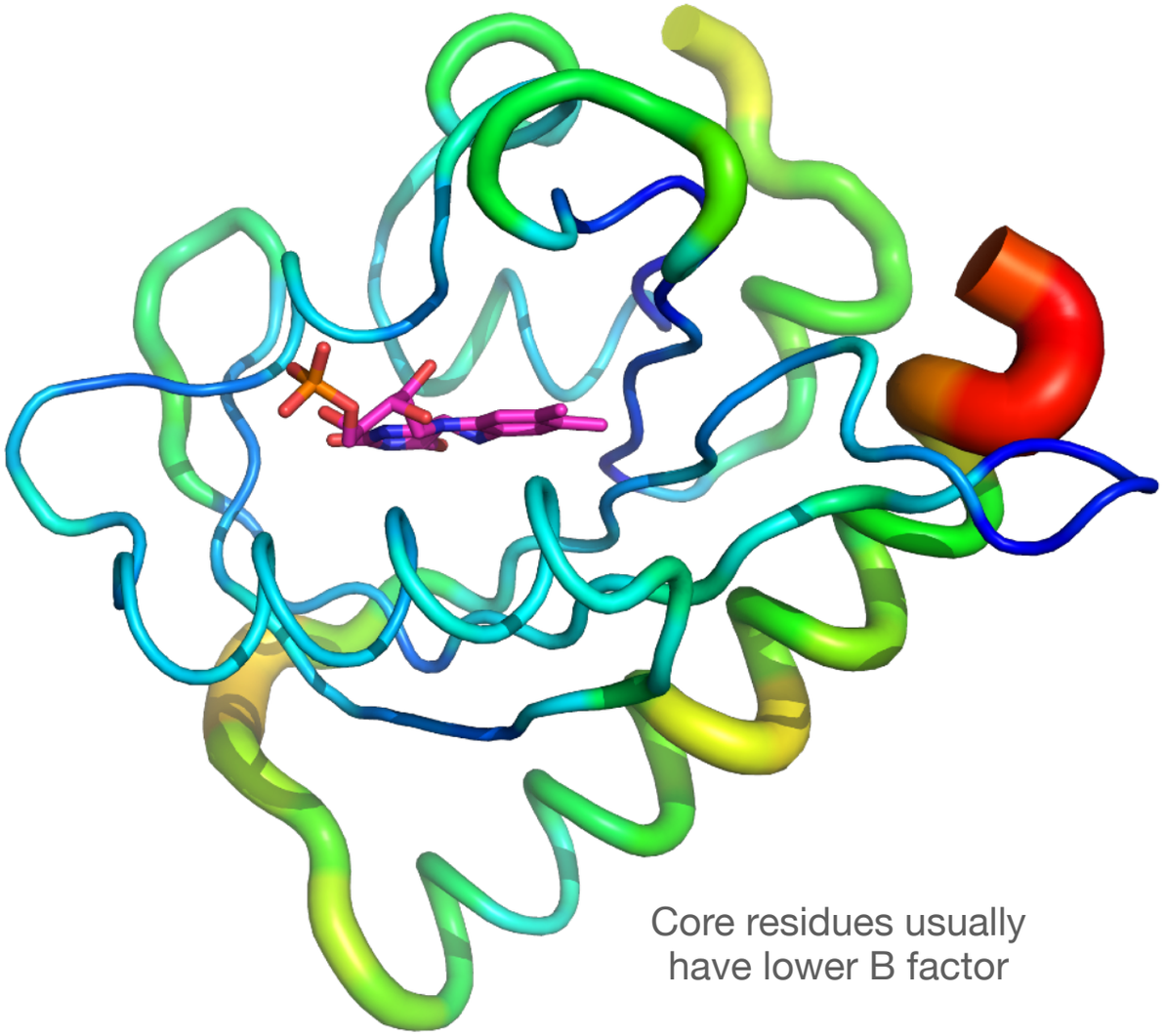
Literature

The crystal structures of trout Hb I in

B factor and occupancy

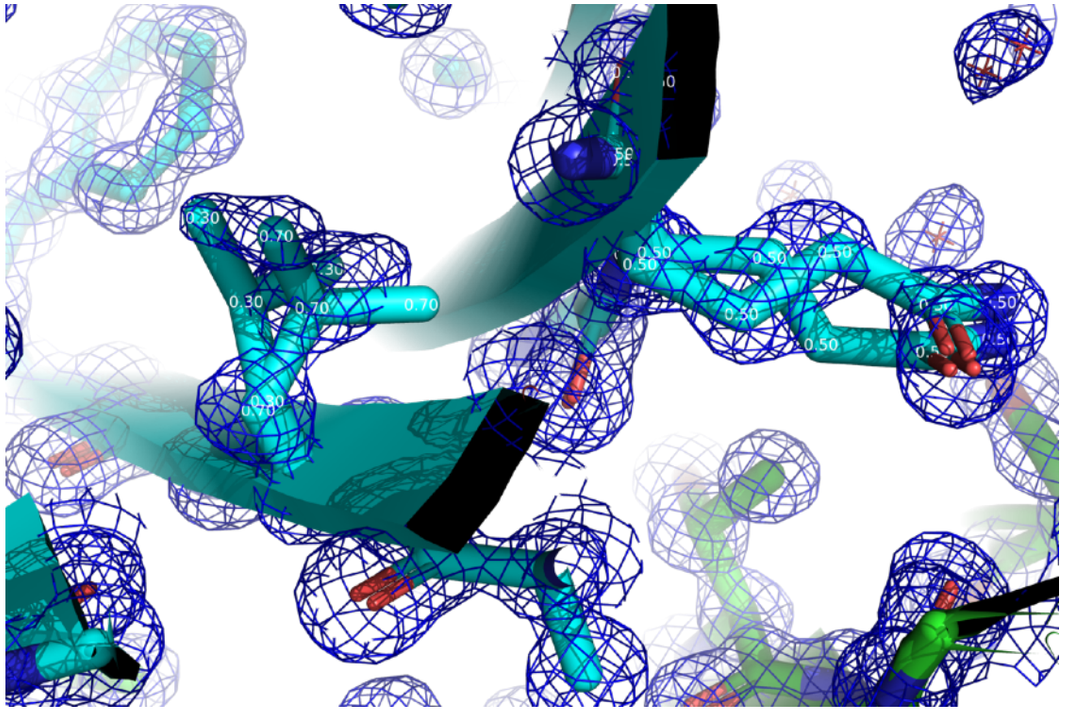
Thermal motion and crystal heterogeneity

LOV2 domain (pdb: 2v1b)



Core residues usually have lower B factor

Streptavidin (pdb: 3ry2, residues 73 and 95, chain B)



Atoms (esp in sidechains and ligands) could have occupancy < 1.0, especially if B factor is higher than surrounding atoms

Thermal motion in the crystal leads to uncertainties in position

If B_j were purely a measure of thermal motion at atom j (and assuming that occupancies are correct), then in the simplest case of purely harmonic thermal motion of equal magnitude in all directions (called *isotropic* vibration), B_j is related to the magnitude of vibration as follows:

$$B_j = 8\pi^2\{u_j^2\} = 79\{u_j^2\}, \tag{8.1}$$

where $\{u_j^2\}$ is the mean-square displacement of the atom from its rest position. Thus if the measured B_j is 79 Å^2, the total mean-square displacement of atom j due to vibration is 1.0 Å^2, and the rms displacement is the square root of $\{u_j^2\}$, or 1.0 Å. The B values of 20 and 5 Å^2 correspond to rms displacements of 0.5 and 0.25 Å. But the B values obtained for most proteins are too large to be seen as reflecting purely thermal motion and must certainly reflect disorder as well.

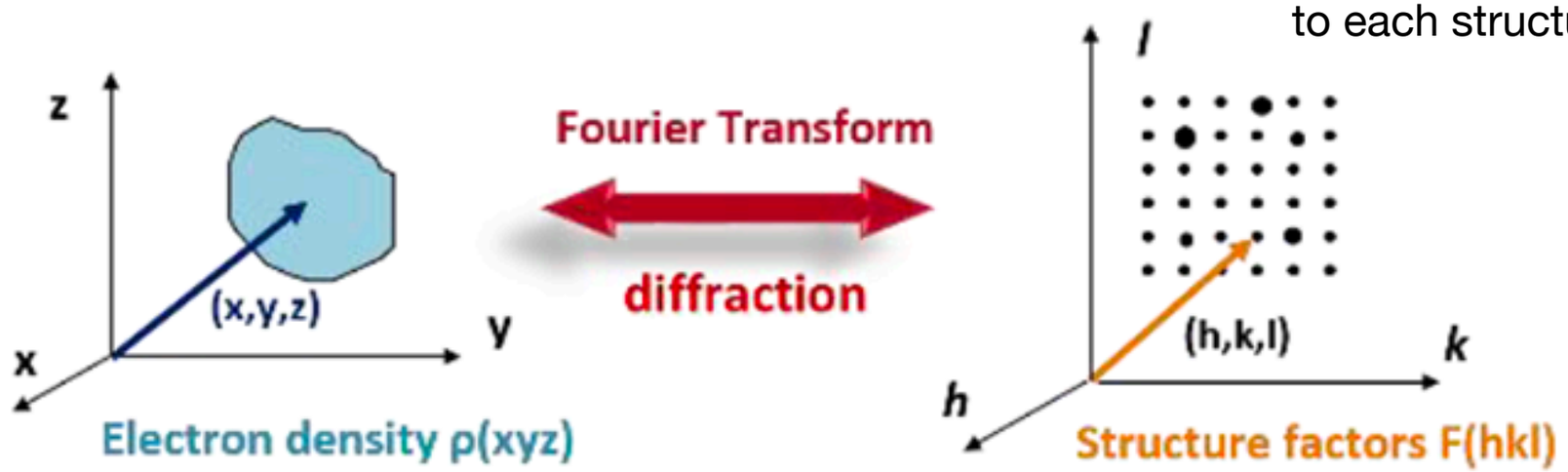
"…if occupancies are constrained to values of 1.00, variation in actual occupancies will show up as increased temperature factors. This is true for other types of constraints as well, and so high temperature factors can mask other kinds of errors. **For this reason, the temperature factor has been uncharitably referred to as a garbage can for model errors**."

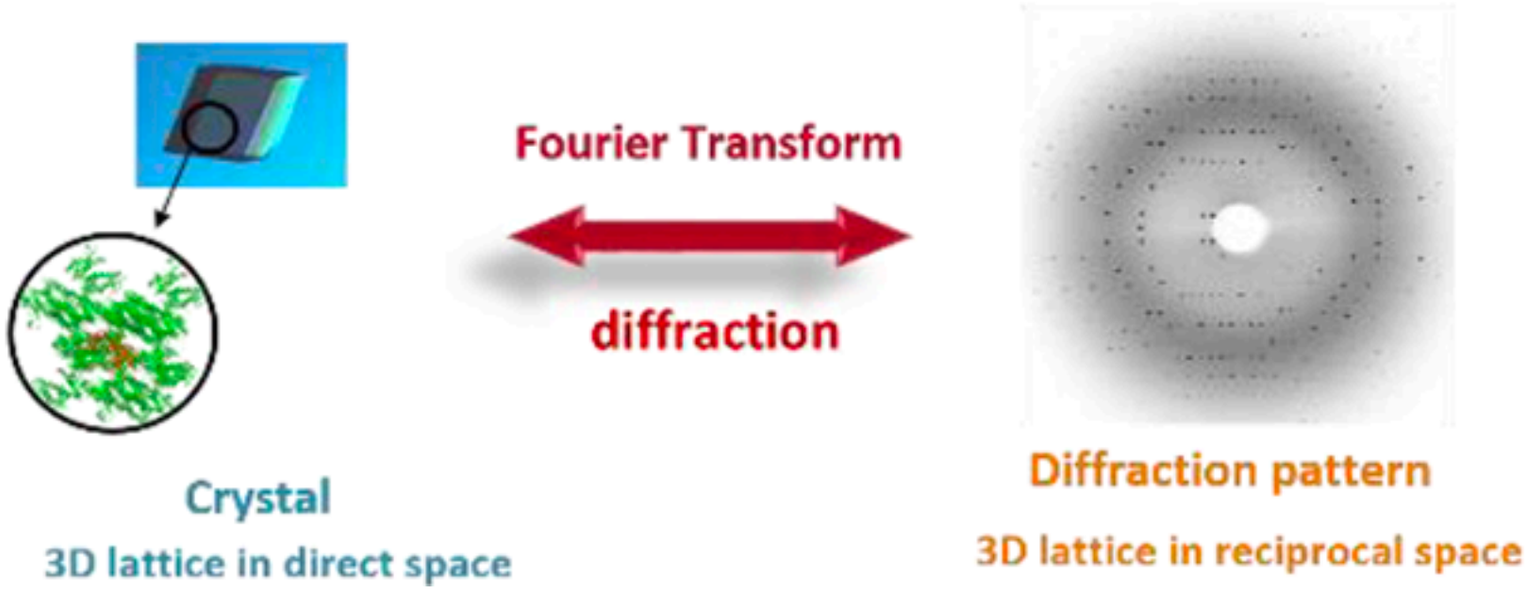
How to compute the density

$$\rho(xyz) = \frac{1}{V} \sum_{hkl}^{+\infty} |F(hkl)| \, . \, e^{-2\pi i[hx+ky+lz-\phi(hkl)]}$$

$$F(hkl) = \sum_j f_j e^{2\pi i(hx_j+ky_j+lz_j)}$$

every atom in the crystal contributes to each structure factor





Molecular replacement or heavy atom soaks can help determine phases

https://www.intechopen.com/chapters/52125

Global error analysis

Least squares fitting during refinement

$$\Phi = \sum_{hkl} w_{hkl} \left(|\mathbf{F}_\mathrm{o}| - |\mathbf{F}_\mathrm{c}|\right)^2_{hkl}$$

$$\mathbf{F}_\mathrm{c} = G \cdot \sum_j n_j f_j e^{2\pi i (hx_j + ky_j + lz_j)} \cdot e^{-B_j [(\sin\theta)/\lambda]^2}$$

2 Fo - Fc	density map
Fo - Fc	difference density map

R factor How well does the model describe the data used in least squares fit?

$$R = \frac{\sum ||\mathbf{F}_\mathrm{obs}| - |\mathbf{F}_\mathrm{calc}||}{\sum |\mathbf{F}_\mathrm{obs}|}$$

Random set of atoms will give R ~ 0.6

R factors of ~0.1 x the resolution are expected, to a point

R free How well does the model describe held-out data (reflections)?

Less biased assessment of model quality.

Usually slightly higher than R factor, but ideally not much higher

Many more kinds of assessments, both local and global

Other considerations when interpreting crystal structure

- Crystal contacts
- Disorder, missing residues
- Steric clashing
- Ramachandran outliers
- Is there density for that?
- Real space residuals (RSR)

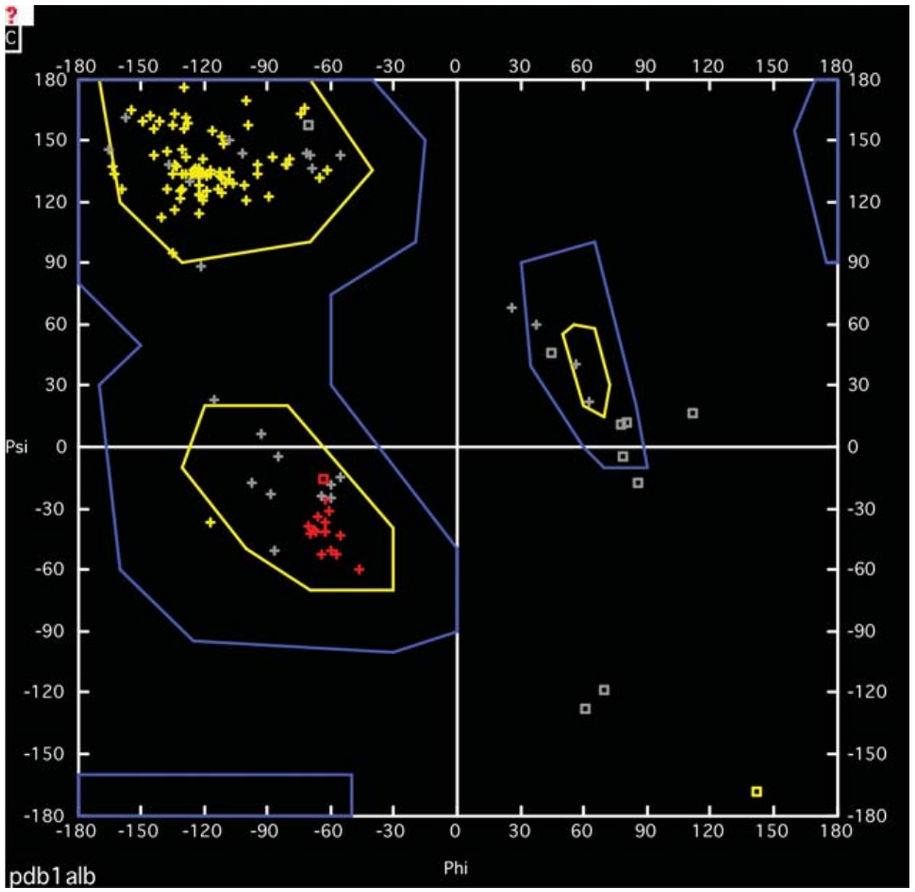


Figure 8.7 ► Ramachandran plot of the crystallographic model of ALBP, generated with DeepView (see Chapter 11). The main-chain torsional angle Φ (N-C_α bond) is plotted versus Ψ (C-C_α bond). The following symbols are used: (+) nonglycine residues; (⊡) glycine residues. The enclosed areas of the plot show sterically allowed angles for nonglycine residues. The symbols are colored according to their inclusion in secondary structural elements: red, alpha helix; yellow, beta sheet; gray, coil.

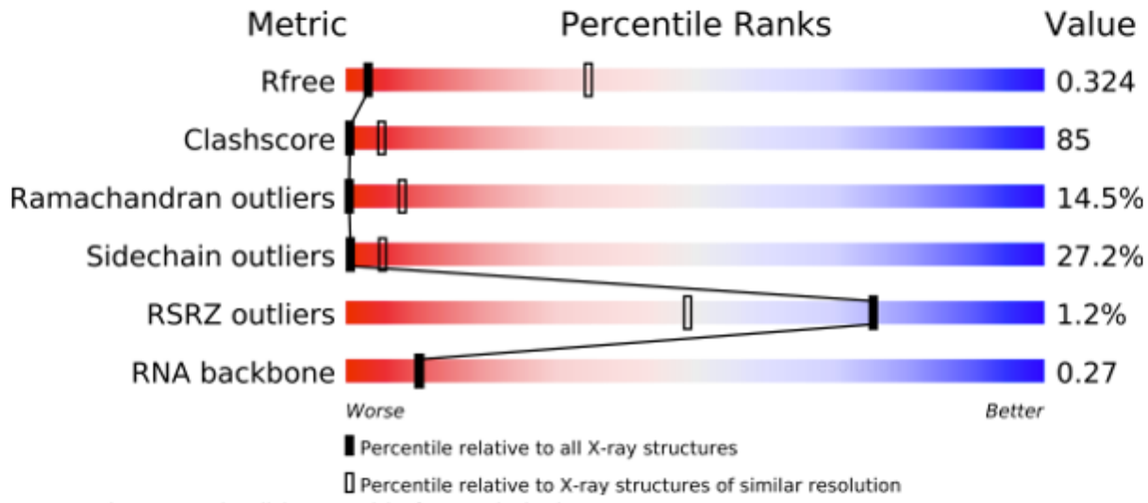
Validation reports are your friend

https://www.wwpdb.org/validation/XrayValidationReportHelp

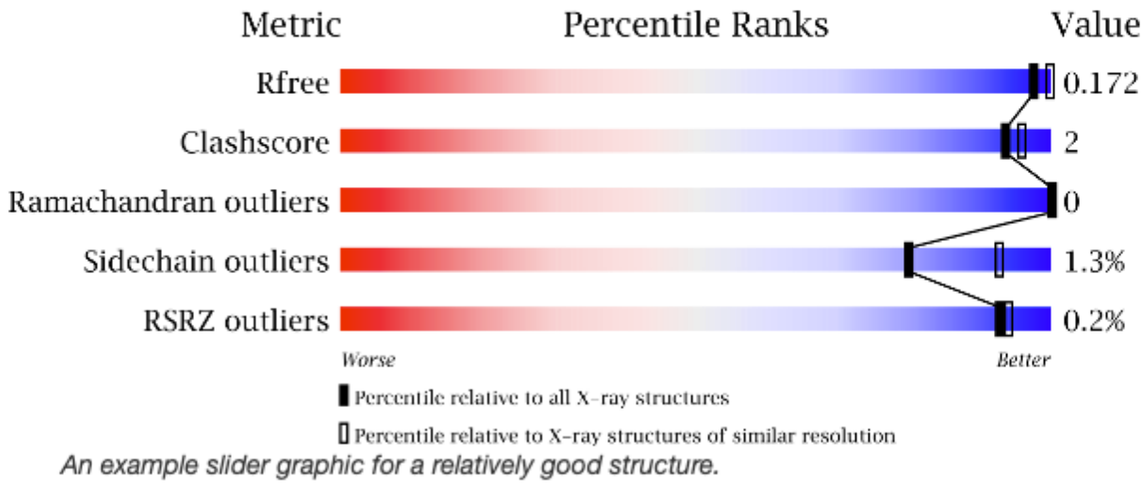
1. Overall quality at a glance

This section provides a succinct "executive" summary of key quality indicators. If there should be serious issues with a structure, this would usually be evident from this summary.

The metrics shown in the "slider" graphic (see examples below) compare several important global quality indicators for this structure with those of previously deposited PDB entries. The comparison is carried out by calculation of the percentile rank, i.e. the percentage of entries that are equal or poorer than this structure in terms of a quality indicator. The global percentile ranks (black vertical boxes) are calculated with respect to all X-ray structures available in the PDB archive up to 27 December 2017. The resolution-specific percentile ranks (white vertical boxes) are calculated with respect to a subset of X-ray entries in the same subset of the PDB archive, but only considering entries with comparable resolution to this entry. In general, one would of course like all sliders to lie far to the right in the blue areas (especially for recently determined structures, and especially the resolution-specific sliders).



An example slider graphic for a relatively poor structure.



An example slider graphic for a relatively good structure.

Note that if you are not an expert you neither need to know what the various quality criteria measure nor whether the values for an entry are unusual or not. However, for increased understanding, below is a brief description of these key global quality indicators:

streptavidin.pse and RCSB

sortase.pse and RCSB

Make your own validation report

Not Secure molprobity.biochem.duke.edu

MOLPROBITY

Main page

Main page
About hydrogens
Evaluate X-ray
Evaluate NMR
Fix up structure
Work with kins

View & download files
Lab notebook
Feedback & bugs
Site map

Save session
Log out

You are using 0% of your 200 Mb of disk space.

Looking at deposited SARS-CoV-2 related structures? Check PDB for updated versions as well as new structures.
(Our Fetch > always returns the latest version.)
Solving or improving them? Look at MolProbity's CaBLAM outliers, and at sparse H-bonds.

FILE UPLOAD/RETRIEVAL (MORE OPTIONS)

PDB/NDB code: type: PDB coords

Choose File No file chosen type: PDB coords

Usage Guidelines:
These web services are provided for analysis of individual structures.
For batch runs, please download and install your own copy of MolProbity.

Walkthroughs, tutorials, and usage FAQs:

Evaluate X-ray structure: Typical steps for a published X-ray crystal structure or one still undergoing refinement.

Evaluate NMR structure: Typical steps for a published NMR ensemble or one still undergoing refinement.

Fix up structure: Rebuild the model to remove outliers as part of the refinement cycle.

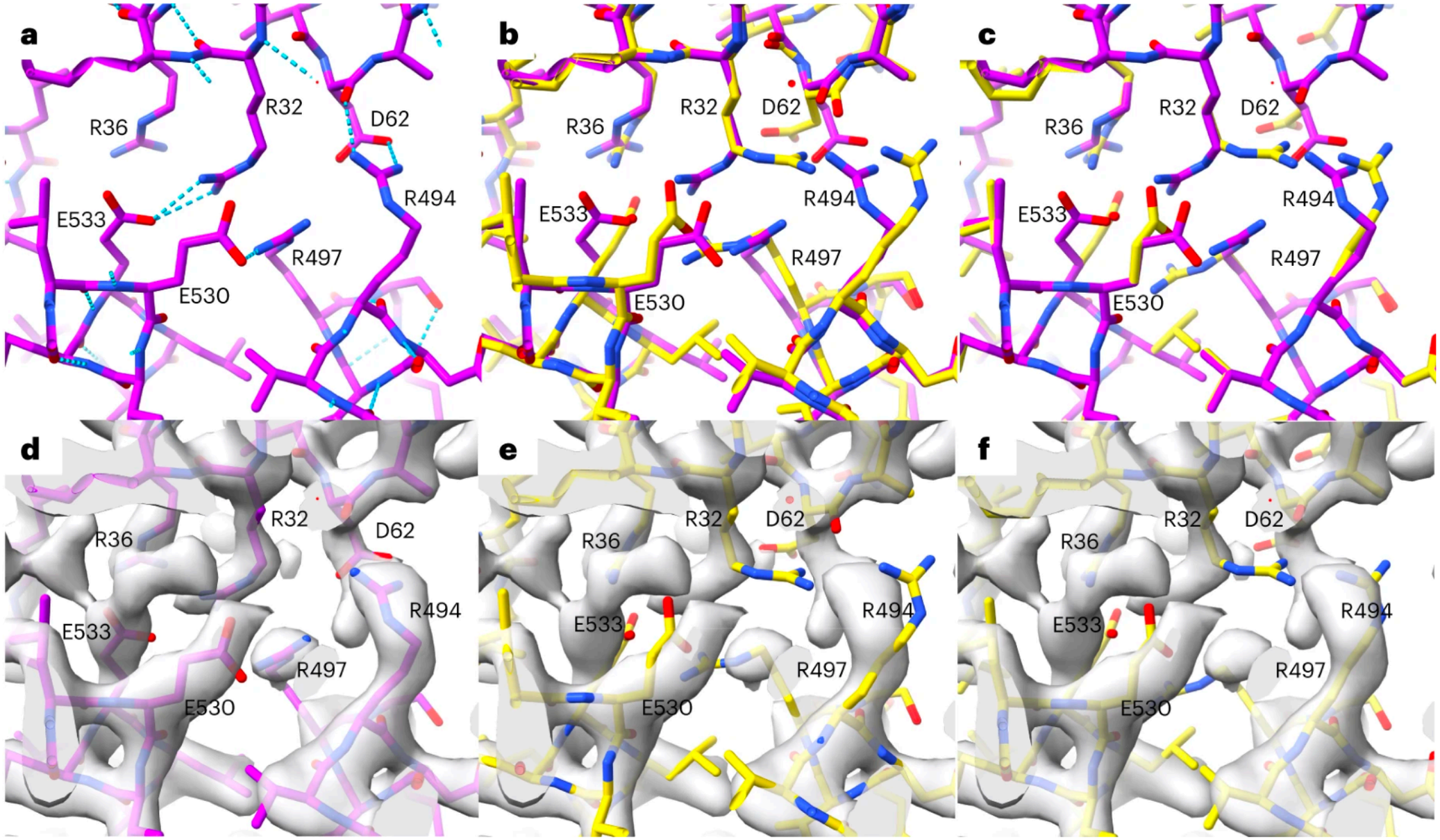
Work with kinemages: Create and view interactive 3-D graphics from your web browser.

Guide to Reduce options: Learn about adding hydrogens to a structure for all-atom contact analysis.

Guide to summary statistics: Interpret structure-level validation statistics.

Guide to validation options: Choose validations appropriate to a structure.

Comparison of AlphaFold side-chain predictions with density map for PDB entry 7vgm



a, PDB entry 7vgm showing hydrogen bonding network. **b**, AlphaFold prediction (yellow) superimposed on deposited model for PDB entry 7vgm (magenta). **c**, As in **b**, except the AlphaFold side chains (yellow) are grafted onto the backbone for PDB entry 7vgm (main-chain atoms for each model are used to superimpose the side chains). **d**, Deposited model as in **a** superimposed on experimental density map (2.3-Å resolution). **e**, AlphaFold prediction as in **b** superimposed on density map. **f**, Grafted AlphaFold model superimposed on density map.

T. C. Terwilliger, et al, AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nat. Methods*, 1–7 (2023).

PDBs with troubling density

7cii
7cig
7cij
5a8d
1xx1
3qqk
3qqf
4mys
5qee
5qel
5qew
3jti
5ryn
3noa
3rm7
3rm6
1pxo
5i8k
2g6o
3r7e
3r73
4wk0
6spw
3tiy
3b9z

3hr8
2fv1
2j7u
2c4e
3fhv
2fhb
3kz7
1i5r
6gue
1um0
5umn
5b72
6dw4
7f20
7f22
2qvu
2dyu
3t00
2qrt
3qrt
5qrv
3qxp
3qx4
1w27
3rni

Project ideas:
Redesign a membrane protein -
Design a transcription factor - make a bundle that has a DNA recognition sequence like a HTH motif.
Design a 2,3,4,5,6,7,8, etc helical bundle.
Design a sequence for two states.
Redesign an nanobody using a nanobody-target structure
If ligandmpnn, check dependence of sequence on ligand for a natural protein.
Change atoms of ligand to see what happens
-use diffdockL to check if it docks?