Principles of protein structure

Biophysical Chemistry 1, Fall 2010

Fundamentals of protein structure

Basics of molecular mechanics and dynamics

Reading assignment: Chaps 1 & 2, Appendices A, B, D & E

Cell biology: not yet biochemistry!

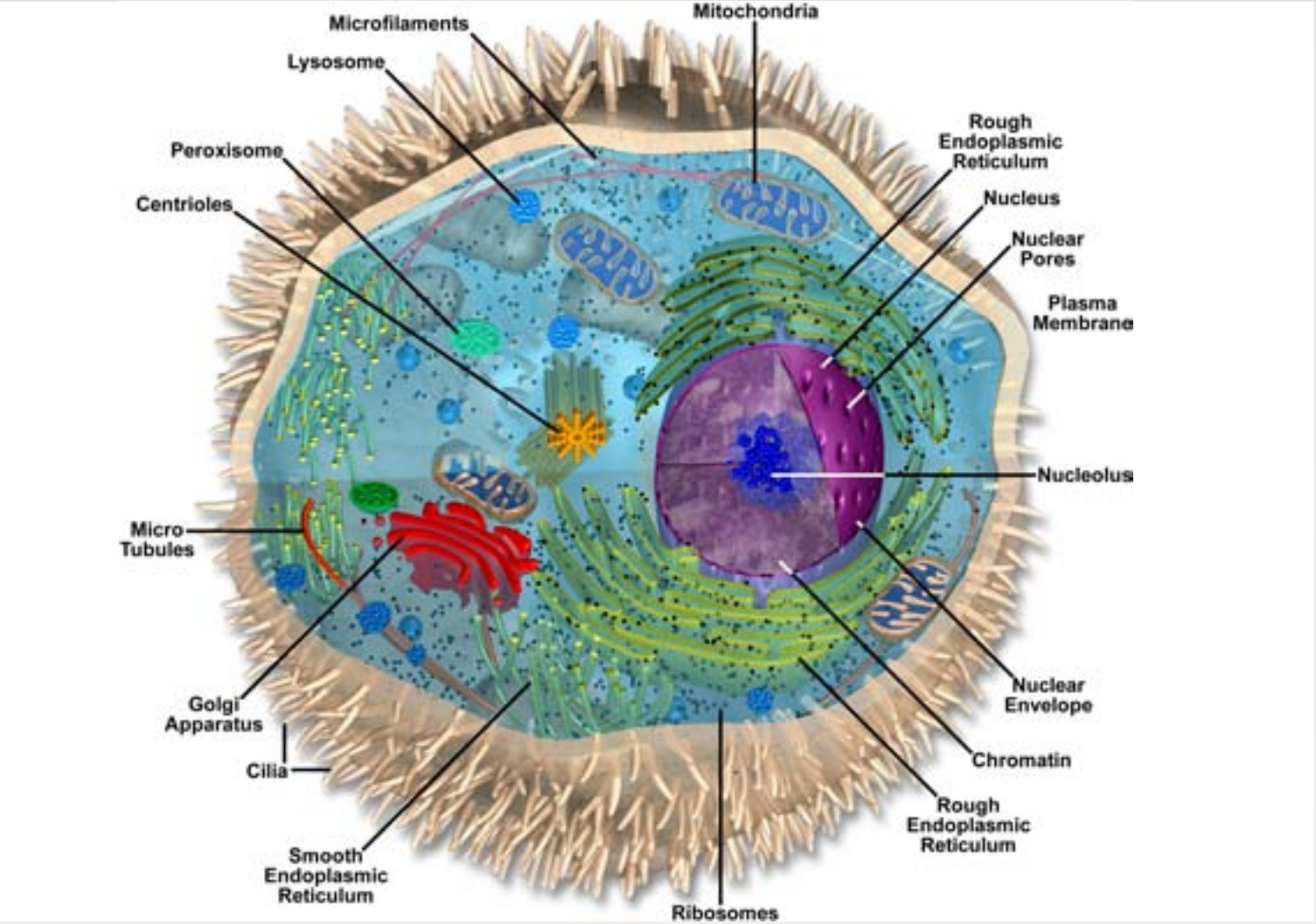


FIGURE 1.3 ■ A schematic picture of an animal cell showing sub-cellular structures, such as nucleus, membrane systems (ER), mitochondrion, etc. (Made by Michael W. Davidson, Florida State University.)

Proteins are polymers of amino acids

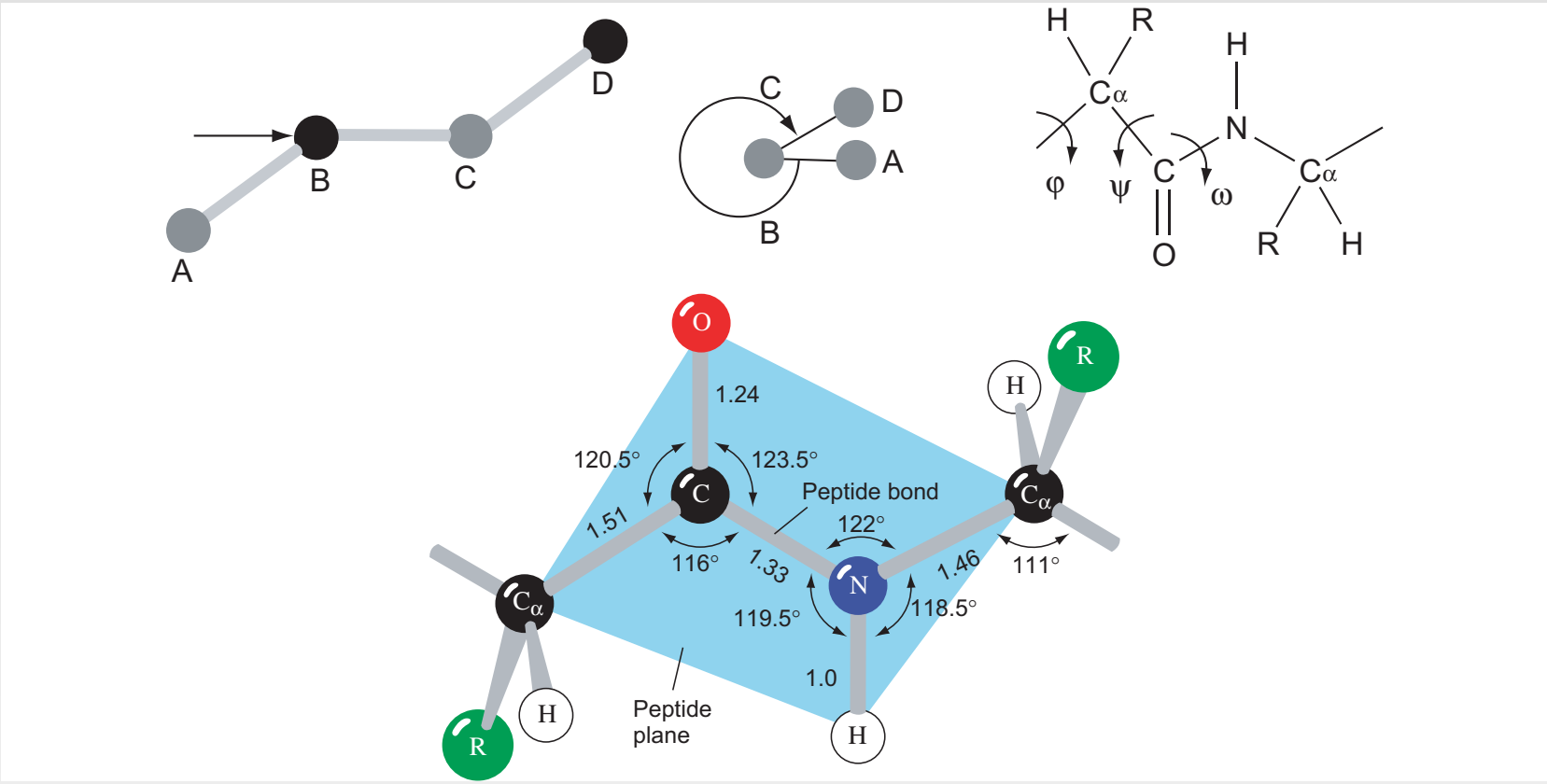


FIGURE 2.3 ■ The peptide bond. The *top left* figure shows the definition of a torsion angle. The *middle* figure, where we look along the bond between atoms B and C, shows how we can determine the angles between the bonds AB and CD. The *top right* drawing shows the names of the torsion angles. In a trans peptide $\omega \approx 180°$, since the atoms of the peptide bond tends to form a plane (the amide plane). *Bottom*: The distances and angles between the atoms of a peptide bond.

There are twenty common side chains

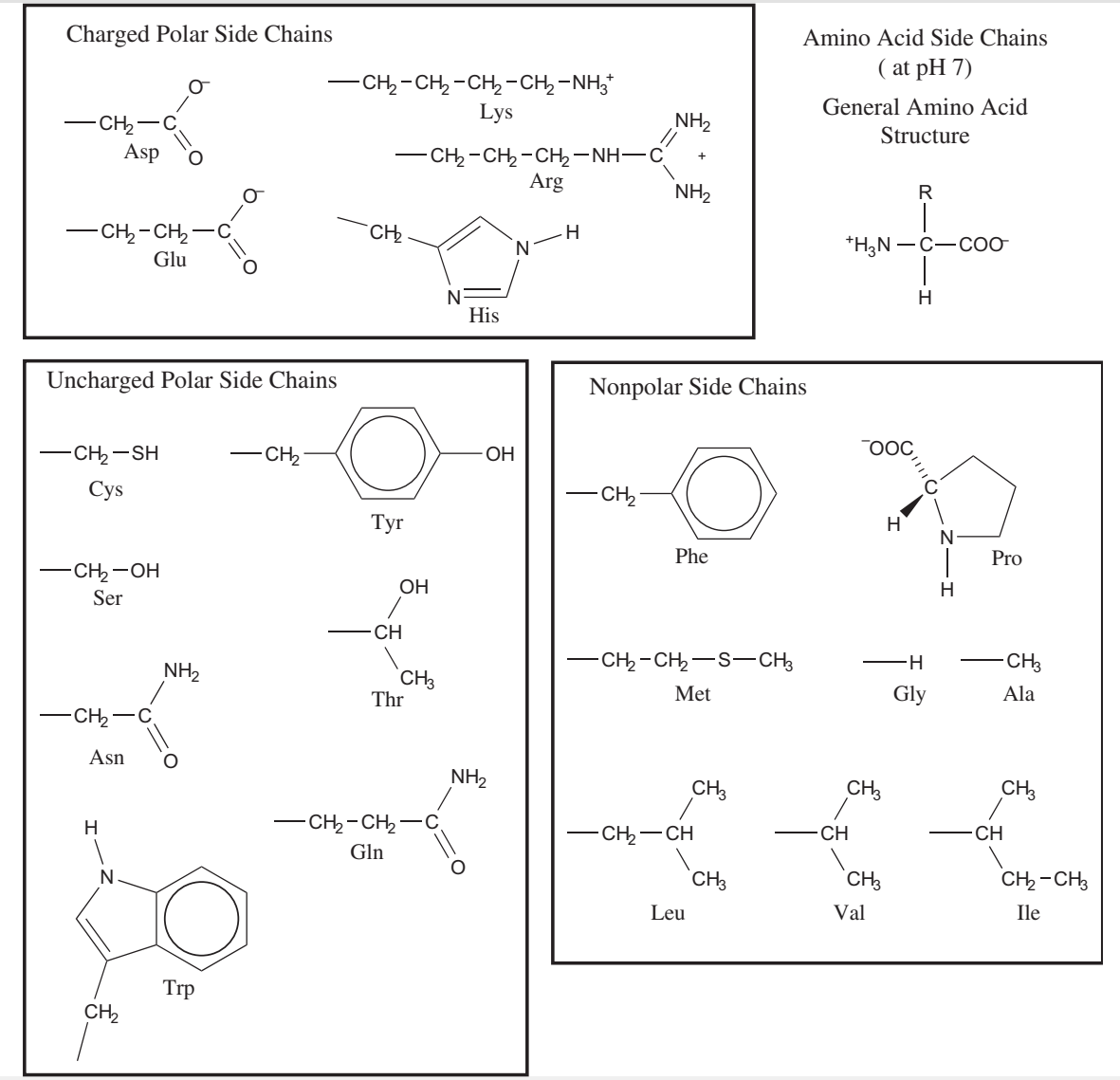


FIGURE 2.2 ■ The 20 different side chains of the amino acids.

The Protein Data Bank (PDB, aka RCSB)

http://www.rcsb.org

Batchelor JD, Doucleff M, Lee CJ, Matsubara K, De Carlo S, Heideker J, Lamers MH, Pelton JG, Wemmer DE.

Structure and regulatory mechanism of *Aquifex aeolicus* NtrC4: variability and evolution in bacterial transcriptional regulation.

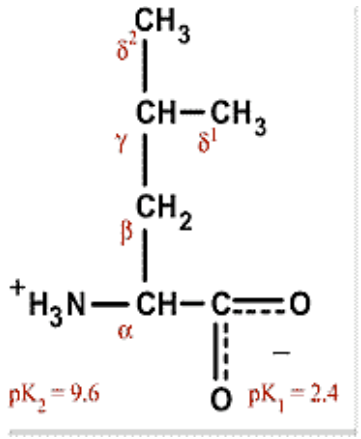
J Mol Biol (2008) **384**(5), 1058-1075.

NtrC4: (nitrogen regulatory protein C4) an activator protein that stimulates gene expression by binding sigma-54
sigma-54 (σ^{54})**:** protein subunit of the RNA polymerase assembly of molecular weight 54 kDa, which is required for the expression of a wide variety of genes
***Aquifex aeolicus*:** thermophilic (heat-loving) bacterium

“ATOM” cards in a PDB file:

```
ATOM     19  N   LEU A  22      13.495  17.685  -7.418  1.00 107.88           N
ATOM     20  CA  LEU A  22      12.855  16.698  -6.555  1.00 100.06           C
ATOM     21  C   LEU A  22      11.626  16.031  -7.172  1.00  97.04           C
ATOM     22  O   LEU A  22      11.508  14.807  -7.158  1.00  94.39           O
ATOM     23  CB  LEU A  22      12.487  17.336  -5.213  1.00  95.81           C
ATOM     24  CG  LEU A  22      13.669  17.846  -4.385  1.00  91.39           C
ATOM     25  CD1 LEU A  22      13.197  18.737  -3.246  1.00  88.32           C
ATOM     26  CD2 LEU A  22      14.492  16.681  -3.861  1.00  88.80           C
```

ATOM ‘card’:
ATOM, atom #, atom name, res name, chain ID, res #, x, y, z, occupancy, B-factor, atom type



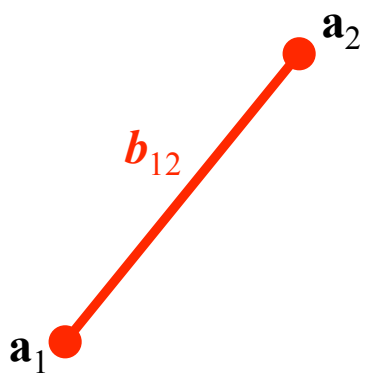
Inter-atomic distances:

Atomic coordinates

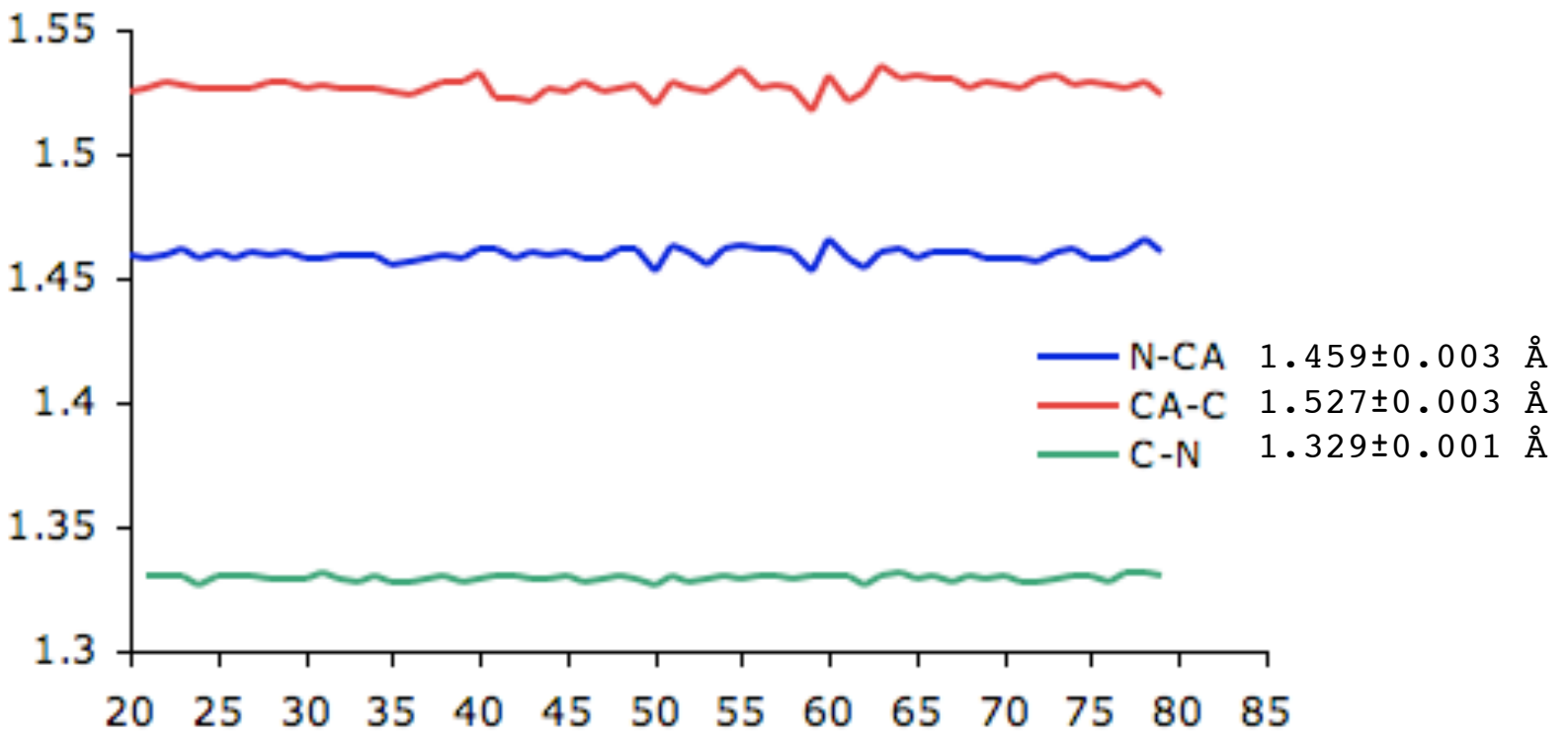
$$\mathbf{a}_1 = (a_{1x},\ a_{1y},\ a_{1z}) \qquad \mathbf{a}_2 = (a_{2x},\ a_{2y},\ a_{2z})$$

Bond length

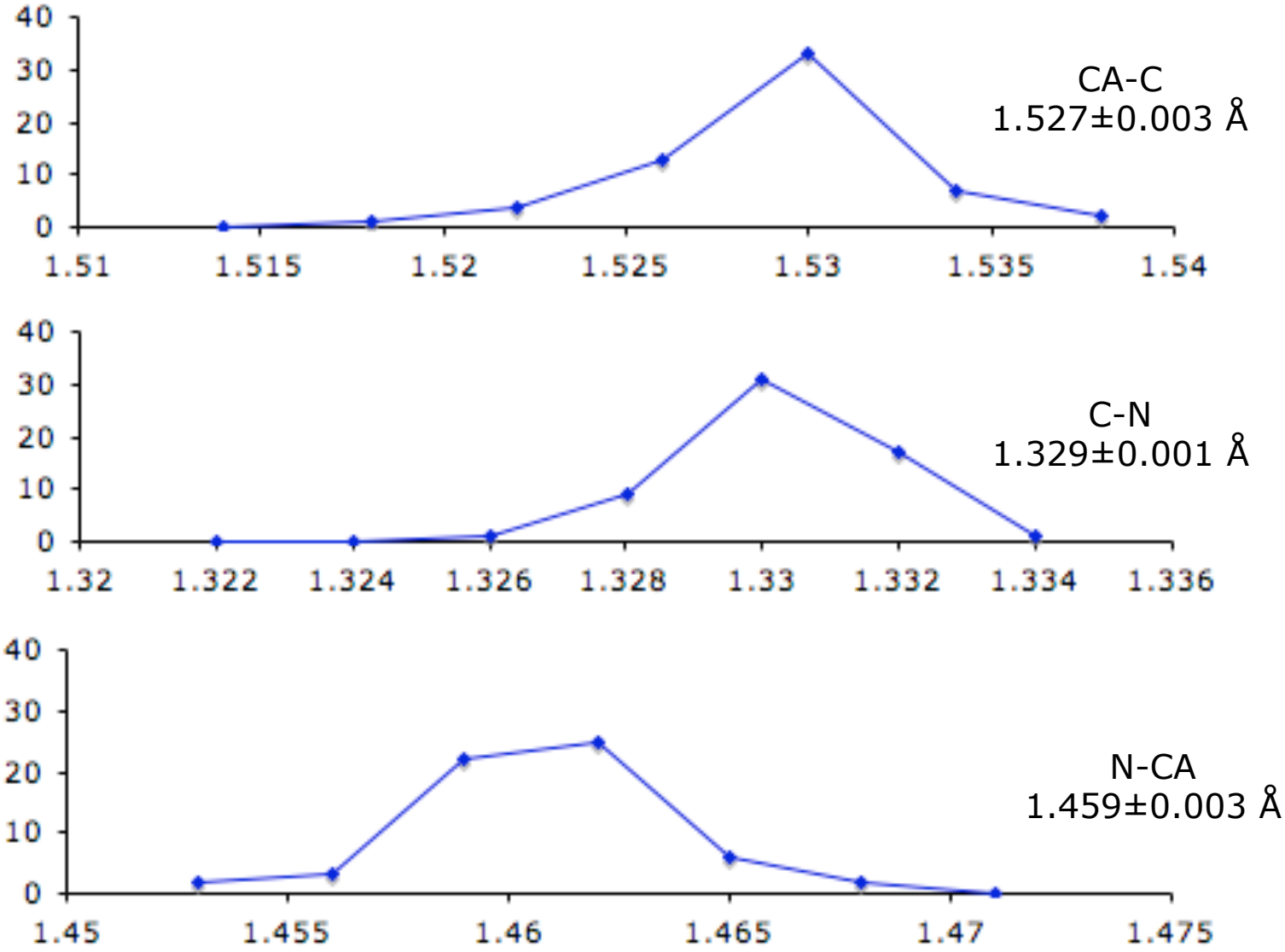
$$b_{12} = ((a_{2x}-a_{1x})^2+(a_{2y}-a_{1y})^2+(a_{2z}-a_{1z})^2)^{1/2}$$



Bonded distances don’t change much:



Histogram of backbone distances:



Bond or “valence” angles:

Atomic coordinates

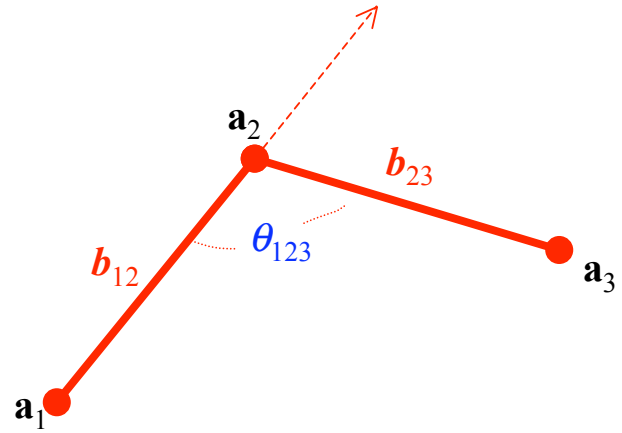
$$\mathbf{a}_1 = (a_{1x},\ a_{1y},\ a_{1z}) \qquad \mathbf{a}_2 = (a_{2x},\ a_{2y},\ a_{2z}) \qquad \mathbf{a}_3 = (a_{3x},\ a_{3y},\ a_{3z})$$

Bond vectors

$$\mathbf{b}_{12} = (a_{2x}-a_{1x},\ a_{2y}-a_{1y},\ a_{2z}-a_{1z}) \qquad \mathbf{b}_{23} = (a_{3x}-a_{2x},\ a_{3y}-a_{2y},\ a_{3z}-a_{2z})$$

Scalar product

$$\mathbf{b}_{12}\cdot\mathbf{b}_{23} = (a_{2x}-a_{1x})(a_{3x}-a_{2x})+(a_{2y}-a_{1y})(a_{3y}-a_{2y})+(a_{2z}-a_{1z})(a_{3z}-a_{2z})$$

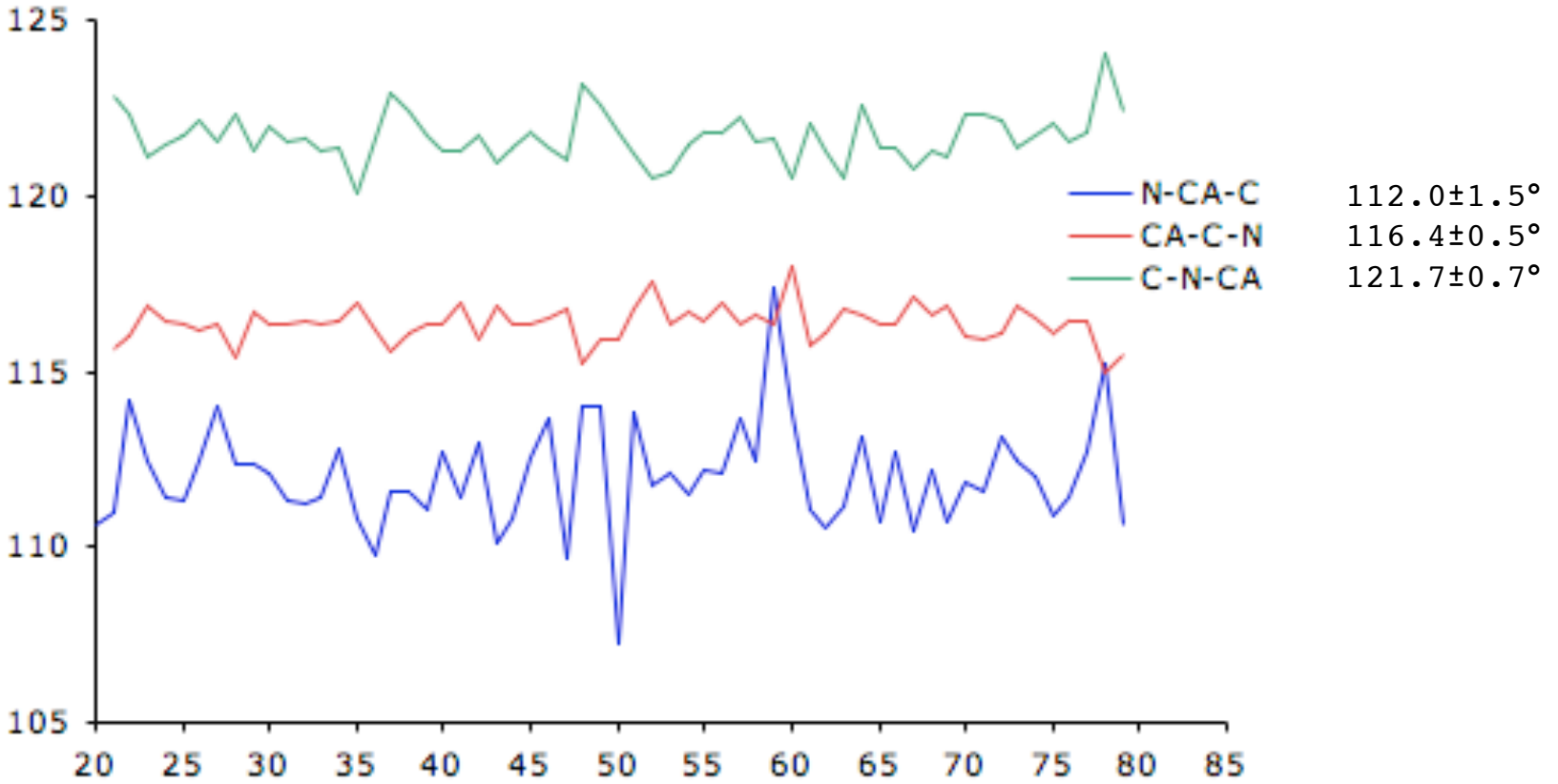


$$\mathbf{b}_{12}\cdot\mathbf{b}_{23} = b_{12}b_{23}\cos(\pi-\theta_{123})$$

$$\cos\theta_{123} = -\mathbf{b}_{12}\cdot\mathbf{b}_{23}/(b_{12}b_{23})$$

$$\cos(\pi-\theta_{123}) = \cos\pi\cos\theta_{123} + \sin\pi\sin\theta_{123} = -\cos\theta_{123}$$

Distribution of backbone angles:



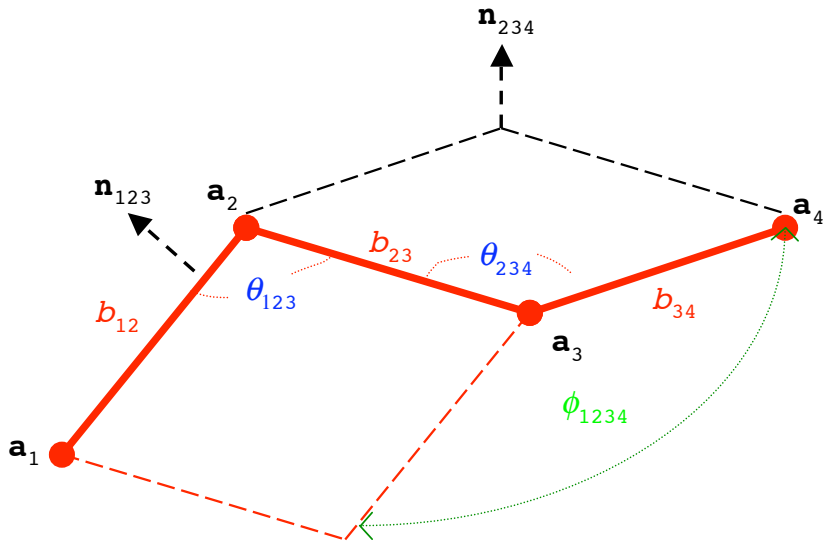
Dihedral or torsion angles:

Atomic coordinates

$\mathbf{a}_1 = (a_{1x}, a_{1y}, a_{1z})$ $\mathbf{a}_2 = (a_{2x}, a_{2y}, a_{2z})$ $\mathbf{a}_3 = (a_{3x}, a_{3y}, a_{3z})$ $\mathbf{a}_3 = (a_{4x}, a_{4y}, a_{4z})$

Bond vectors

$\mathbf{b}_{12}=(a_{2x}-a_{1x}, a_{2y}-a_{1y}, a_{2z}-a_{1z})$ $\mathbf{b}_{23}=(a_{3x}-a_{2x}, a_{3y}-a_{2y}, a_{3z}-a_{2z})$ $\mathbf{b}_{34}=(a_{4x}-a_{3x}, a_{4y}-a_{3y}, a_{4z}-a_{3z})$



$$\mathbf{n}_{123} = \mathbf{b}_{12} \times \mathbf{b}_{23}$$

$$\mathbf{n}_{234} = \mathbf{b}_{23} \times \mathbf{b}_{34}$$

$$\mathbf{n}_{123} \cdot \mathbf{n}_{234} = \cos\phi_{1234}$$

$$(\mathbf{n}_{123} \times \mathbf{n}_{234}) \cdot \mathbf{b}_{23} > 0 \Rightarrow \phi_{1234} > 0$$

$$(\mathbf{n}_{123} \times \mathbf{n}_{234}) \cdot \mathbf{b}_{23} < 0 \Rightarrow \phi_{1234} < 0$$

Vector product or “cross” product:

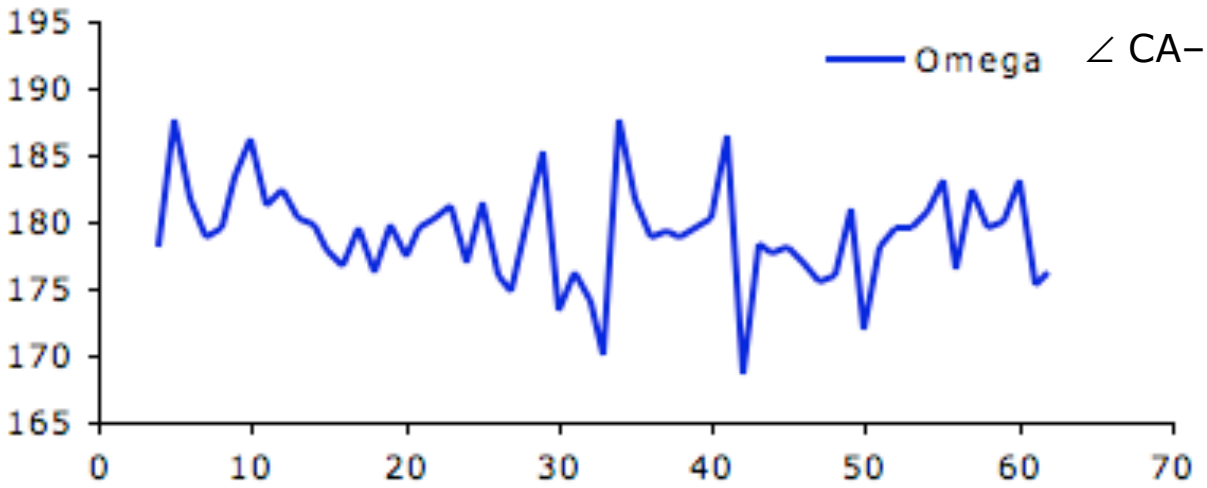
$$\mathbf{V}_1 \times \mathbf{V}_2 = \begin{vmatrix} i & j & k \\ v_{1x} & v_{1y} & v_{1z} \\ v_{2x} & v_{2y} & v_{2z} \end{vmatrix}$$

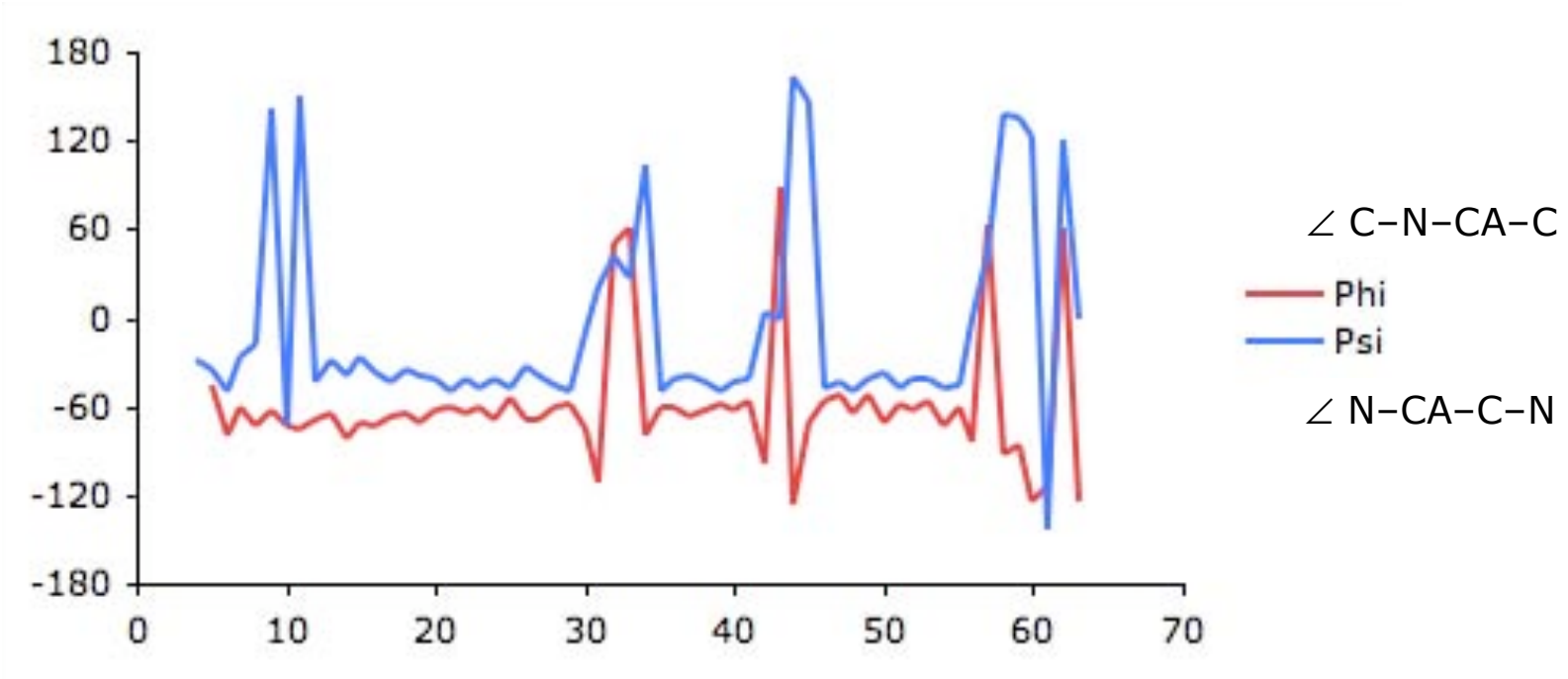
$$= \left(v_{1y}v_{2z} - v_{2y}v_{1z}\right)i + \left(v_{1z}v_{2x} - v_{2z}v_{1x}\right)j + \left(v_{1x}v_{2y} - v_{2x}v_{1y}\right)k$$

$$= \left[\left(v_{1y}v_{2z} - v_{2y}v_{1z}\right), \left(v_{1z}v_{2x} - v_{2z}v_{1x}\right), \left(v_{1x}v_{2y} - v_{2x}v_{1y}\right)\right]$$

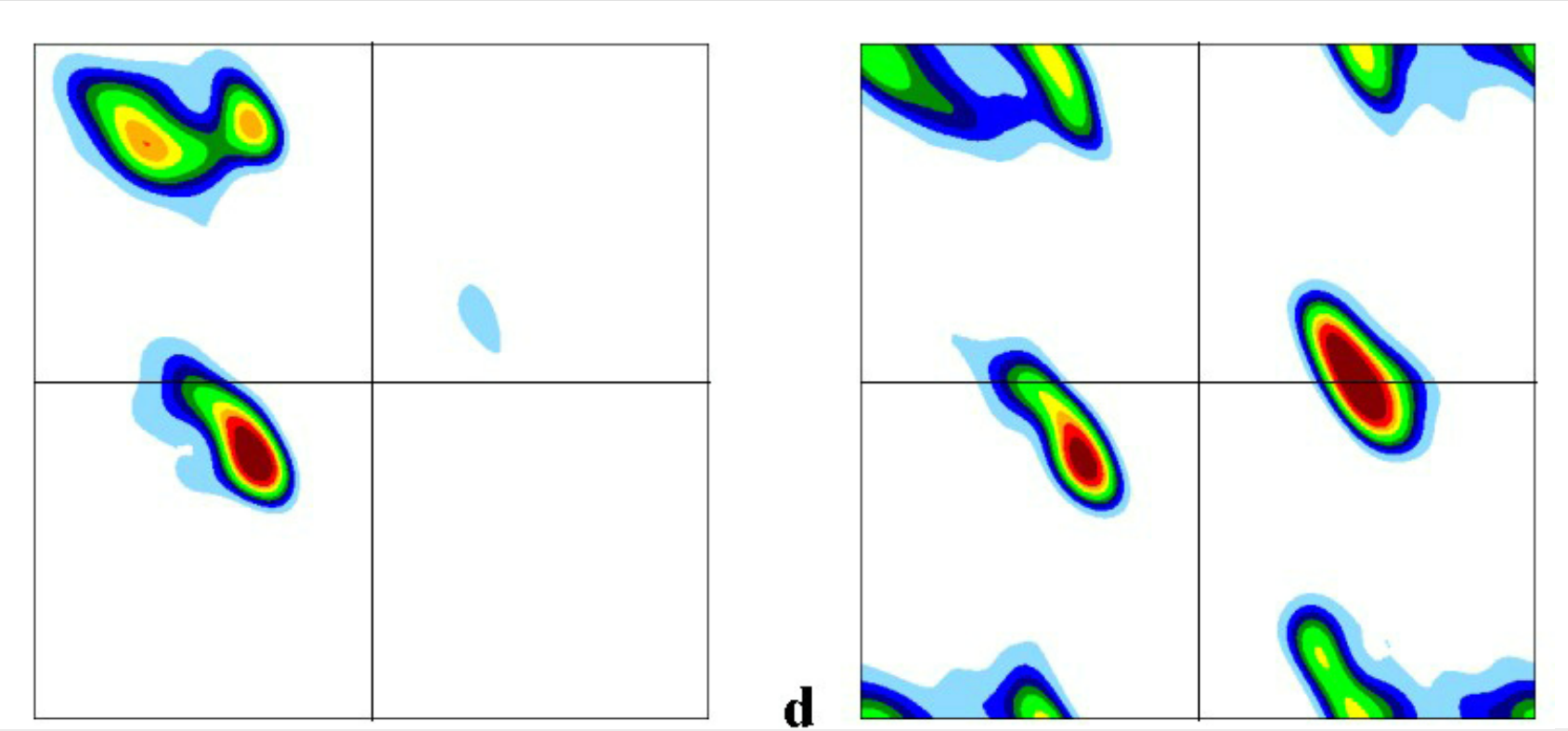
$$= \left(c_1, c_2, c_3\right)$$

Backbone torsion angles:





Ramachandran plots: sidechain torsional potentials



alanine

glycine

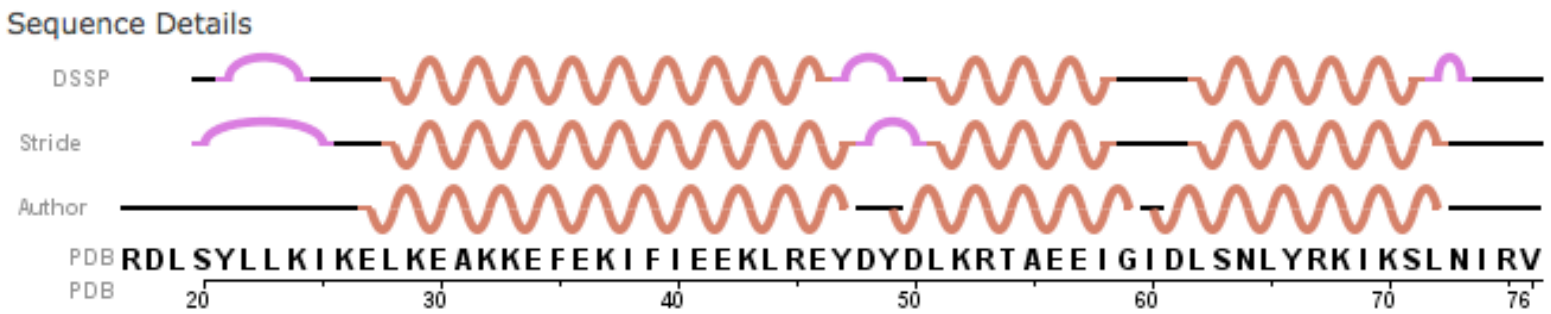
Histogram of backbone distances:

The repetition of local chemical parameters in successive residues generates a regular helical structure.

Three segments of chain A (involving residues 28-49, 54-60, and 62-74) show such repeating patterns:

CA-CA bond lengths ~3.81 Å
CA-CA-CA valence angles ~90°
CA-CA-CA-CA dihedral angles ~50°

The PDB includes the following helical residue assignments for `3e71`:



Secondary structure: the α-helix

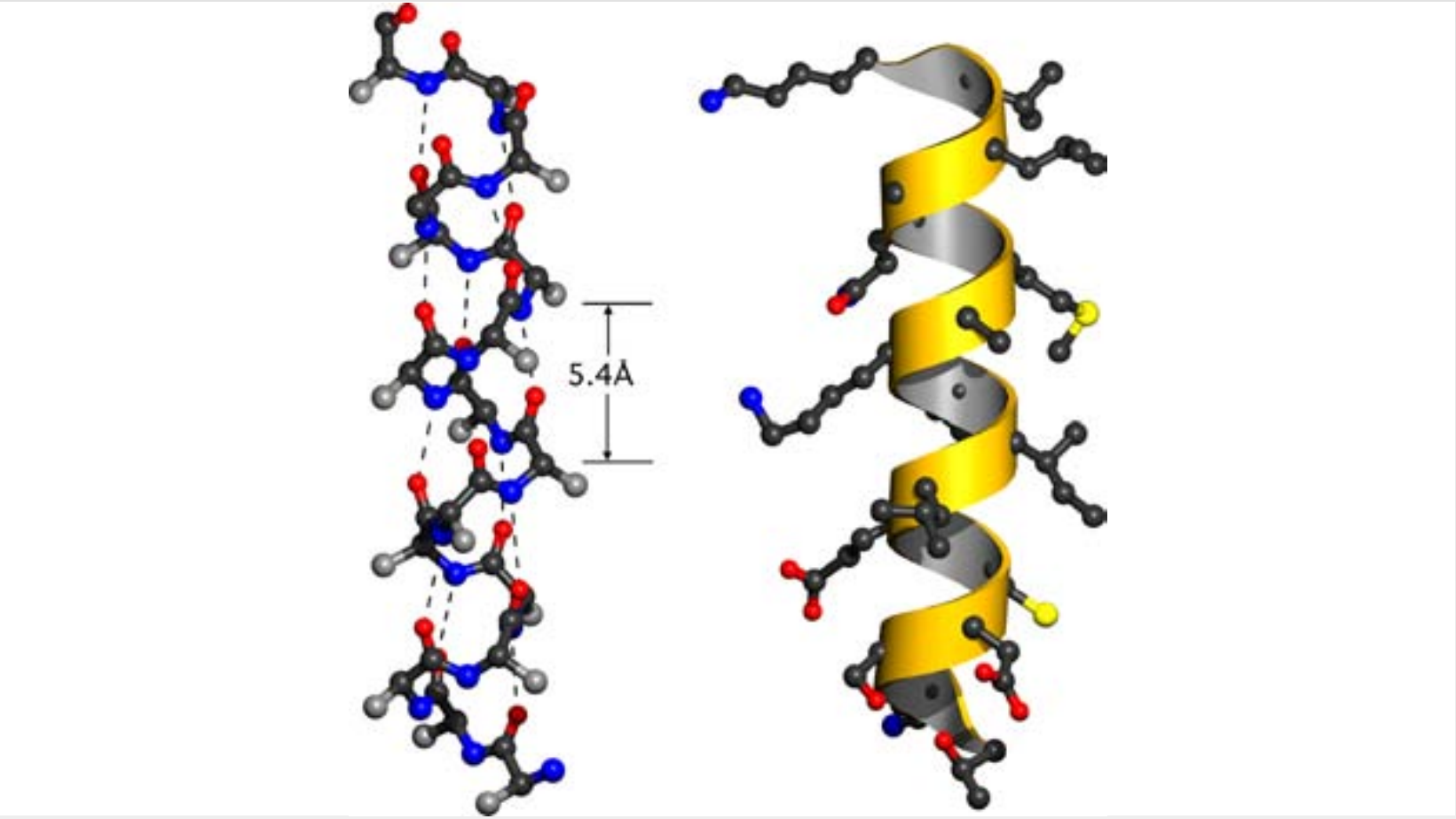


FIGURE 2.7 ■ The α-helix. *Left*: The main chain and Cβ atoms (gray) of an α-helix. The pitch (rise per turn) is 5.4 Å. *Right:* The same α-helix showing the side chains. The backbone is drawn schematically, with the Cβ atoms pointing towards the N-terminus of the helix (down).

Secondary structure: 3_{10} and π helices

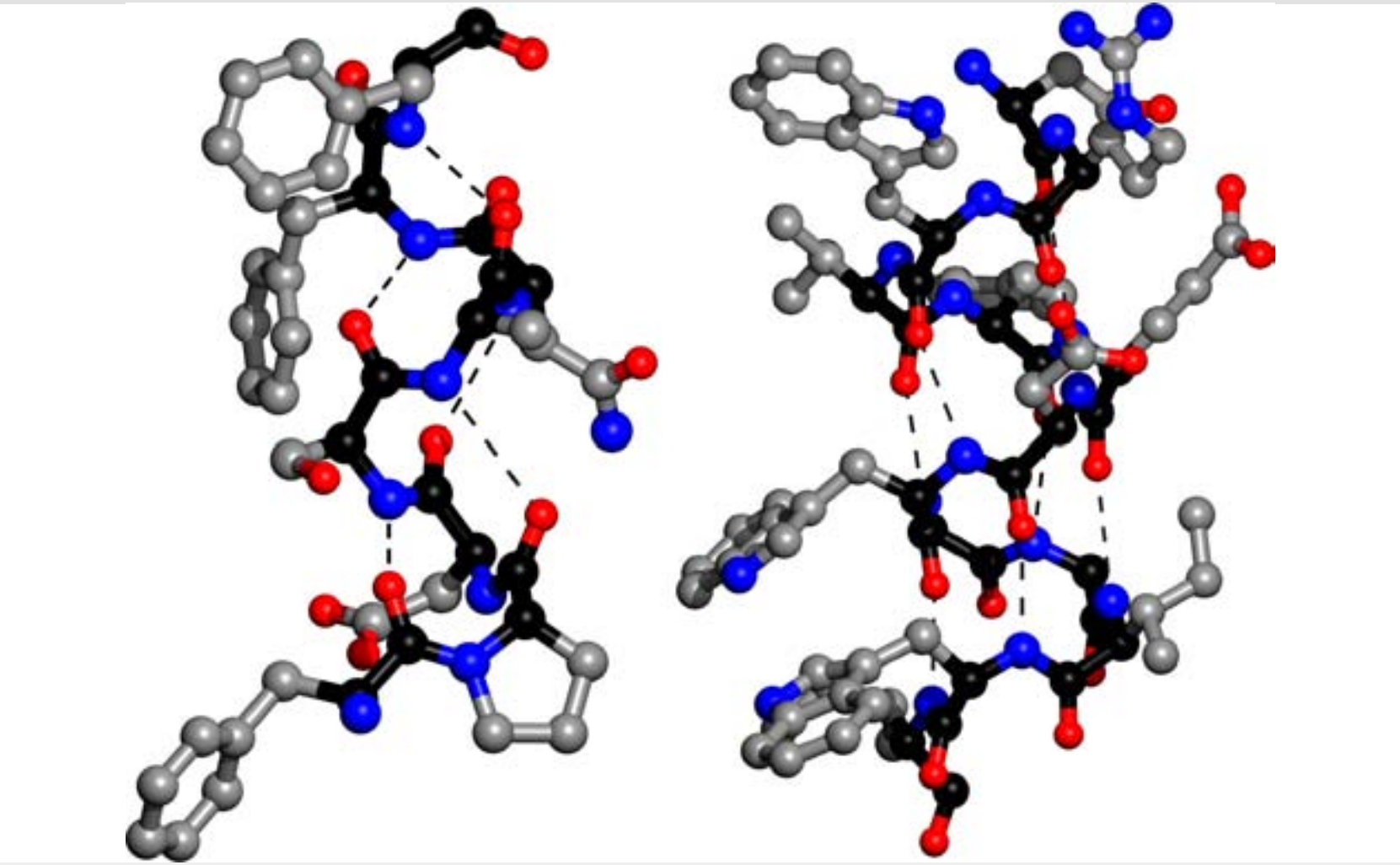


FIGURE 2.8 ■ A 3_{10} helix from *Aplysia limacina* myoglobin (PDB: 1MBA) and a π-helix from methane monooxygenase hydroxylase from *Methylococcus capsulatus* (PDB: 1MTY).

Secondary structure: the β sheets

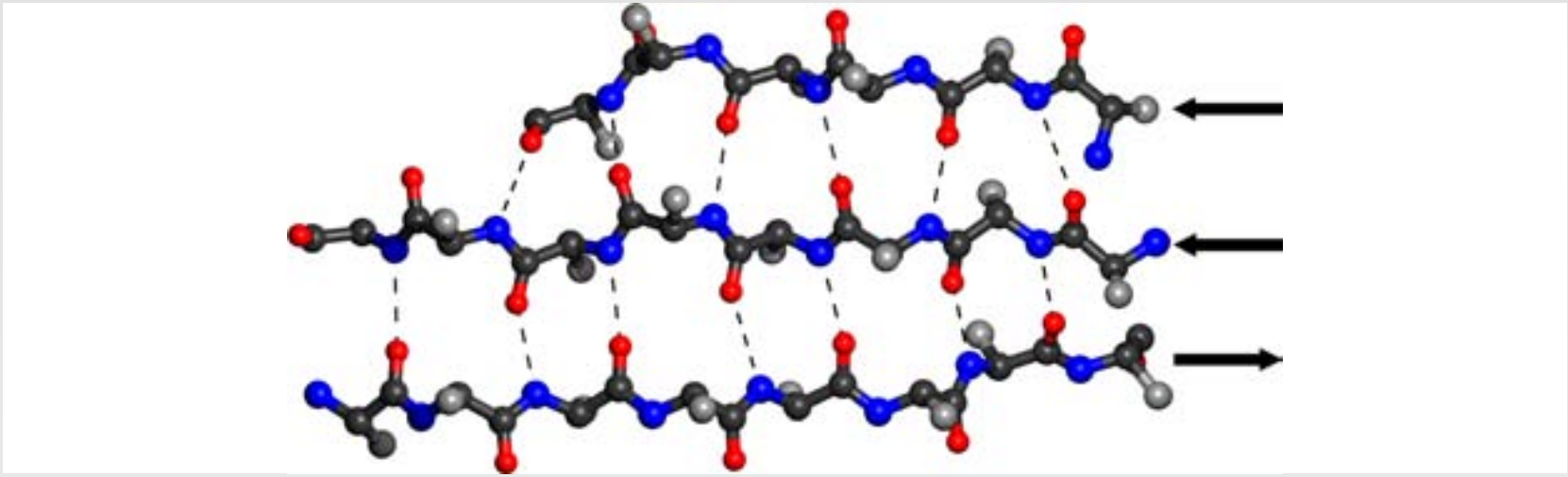
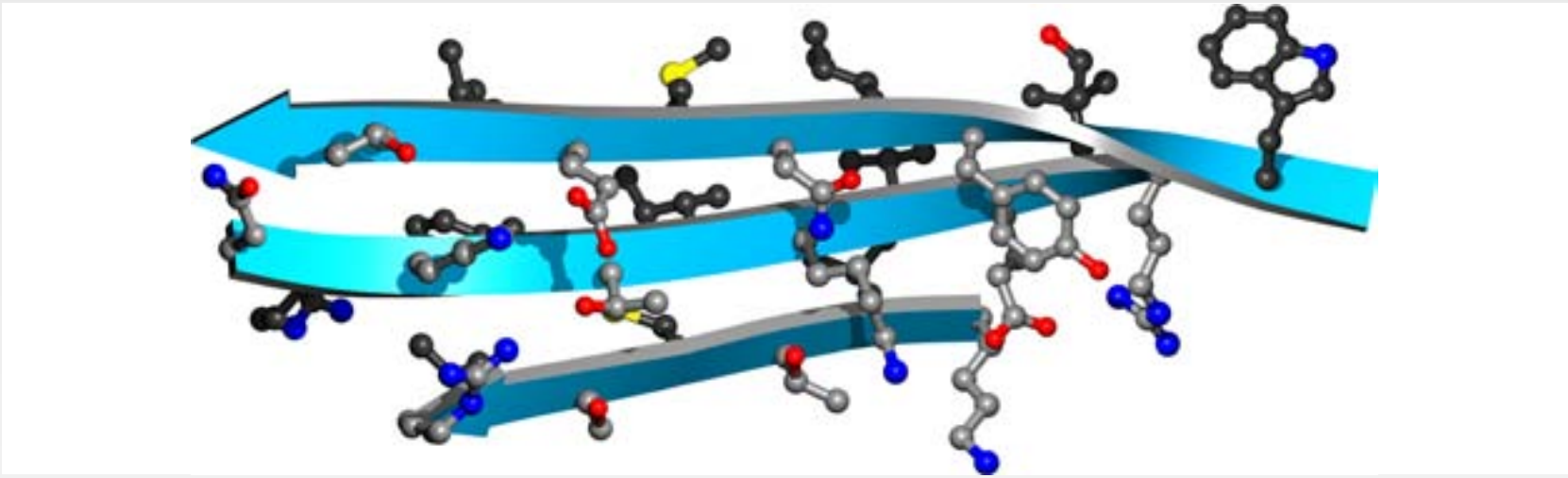


FIGURE 2.9 ■ The hydrogen bonds in a mixed β-sheet. β-sheets are always more or less twisted.



Secondary structure: turns

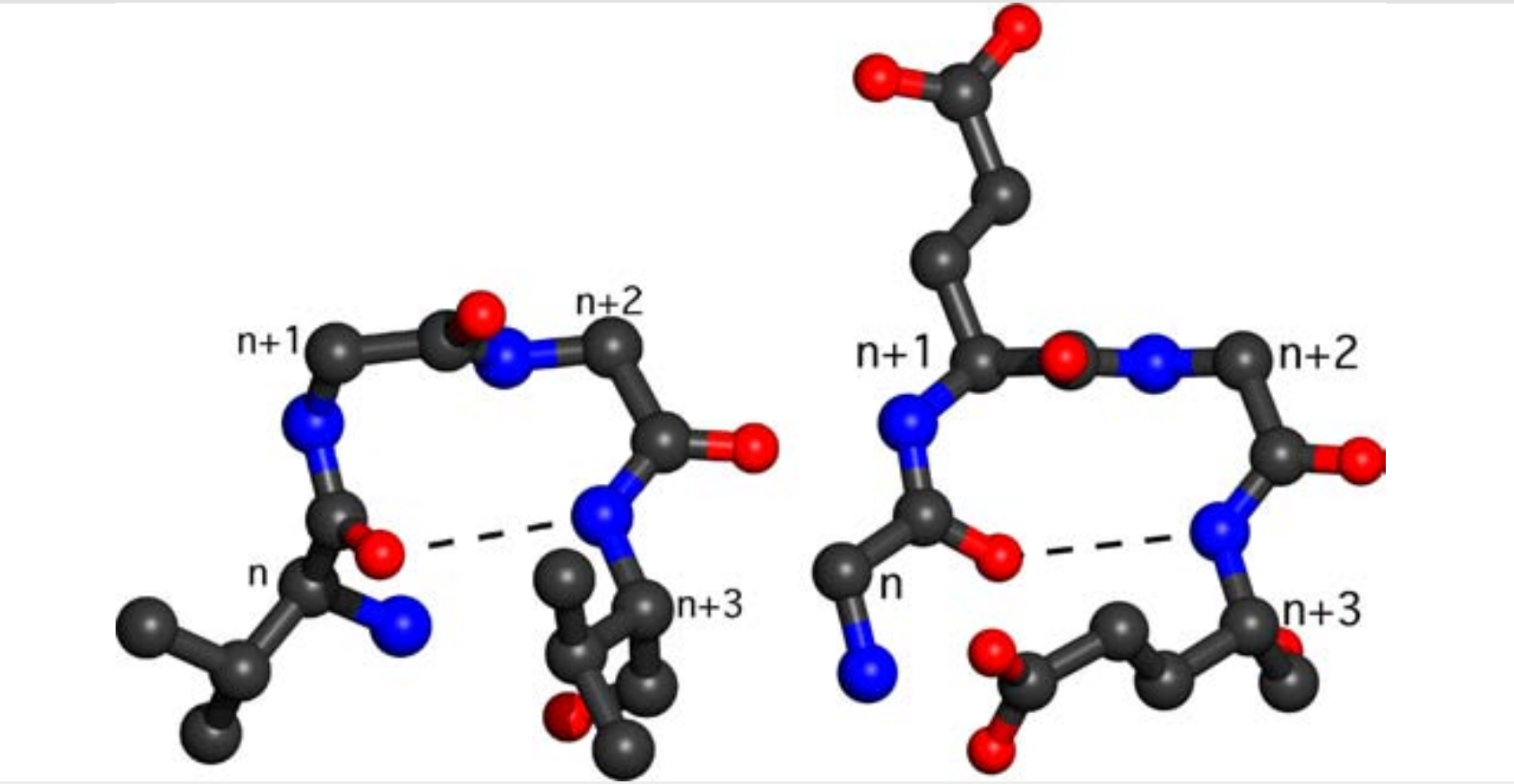
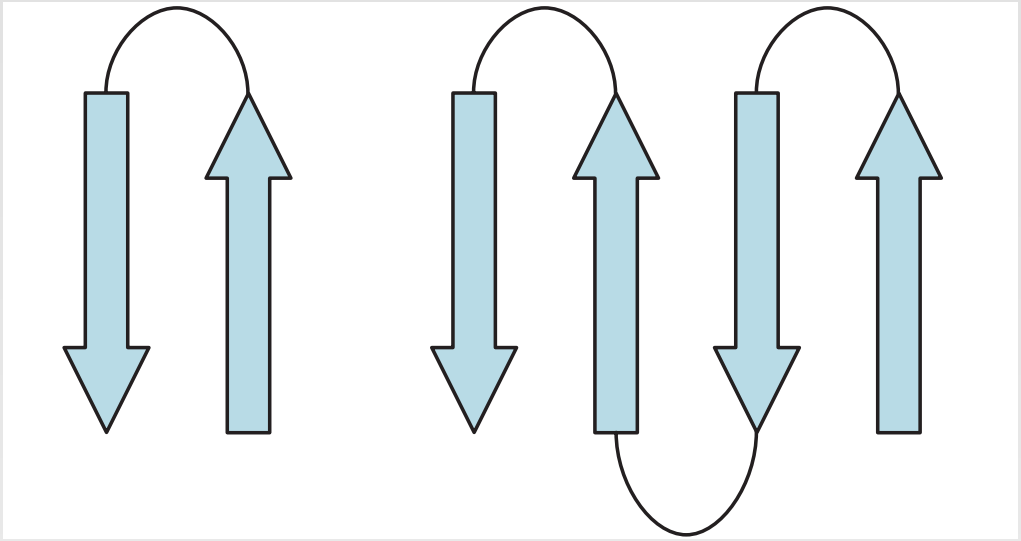


FIGURE 2.13 ■ Two of the most common types of reverse turns, type I′ and II.

Motifs, topologies, folds: antiparallel β sheets



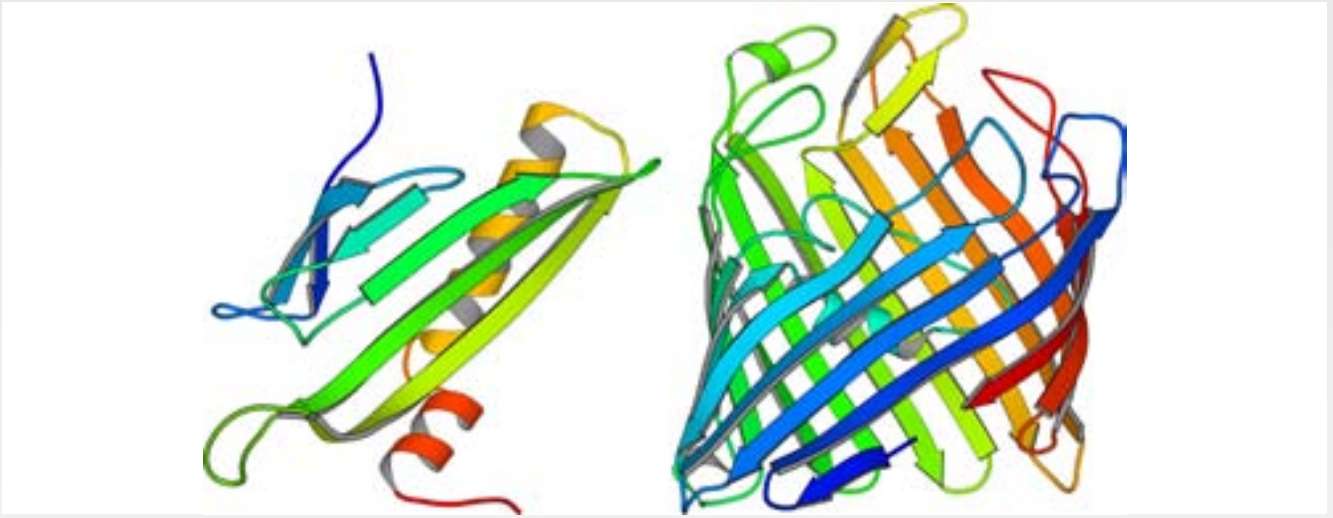


FIGURE 2.15 ■ Two up-and-down sheets, the open sheet in the coat protein subunit of phage MS2 (PDB: 2MS2), and the closed cylinder in a bacterial porin, a protein from the outer membrane of *E. coli* (PDB: 2OMF).

Motifs, topologies, folds: Greek key

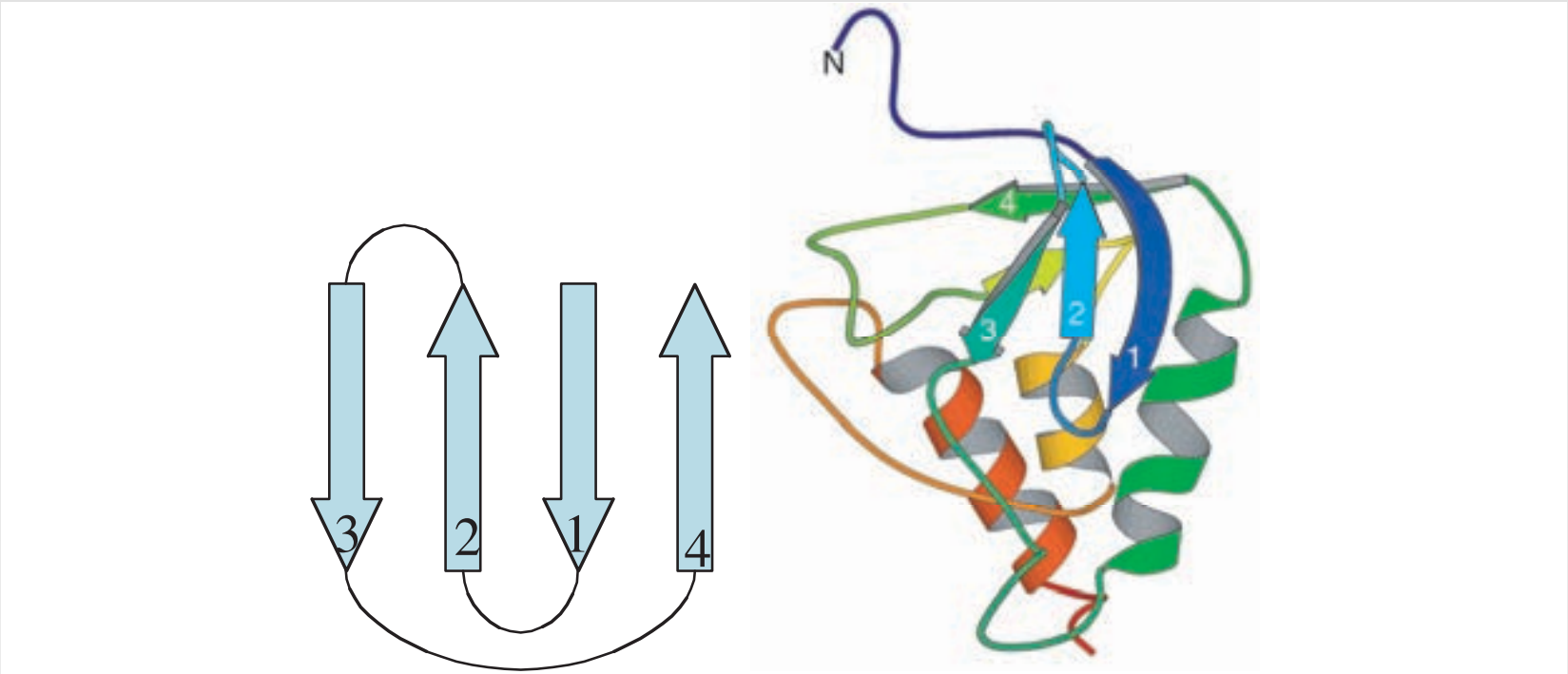


FIGURE 2.17 ■ A schematic drawing of a Greek key motif and the same motif in a protein (Micrococcal nuclease, PDB: 2SNS). The 5-stranded sheet and the helix in the loop connecting strands 3 and 4 is an example of the OB (oligonucleotide/oligosaccharide binding) fold found in many proteins.

Motifs, topologies, folds: jellyroll

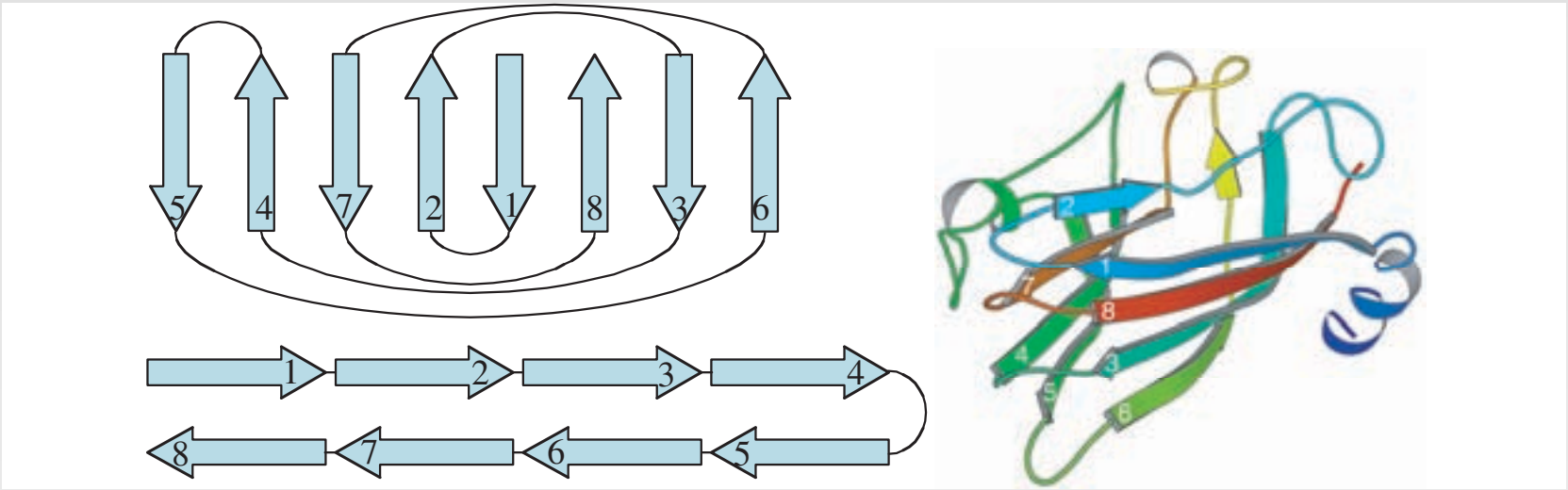
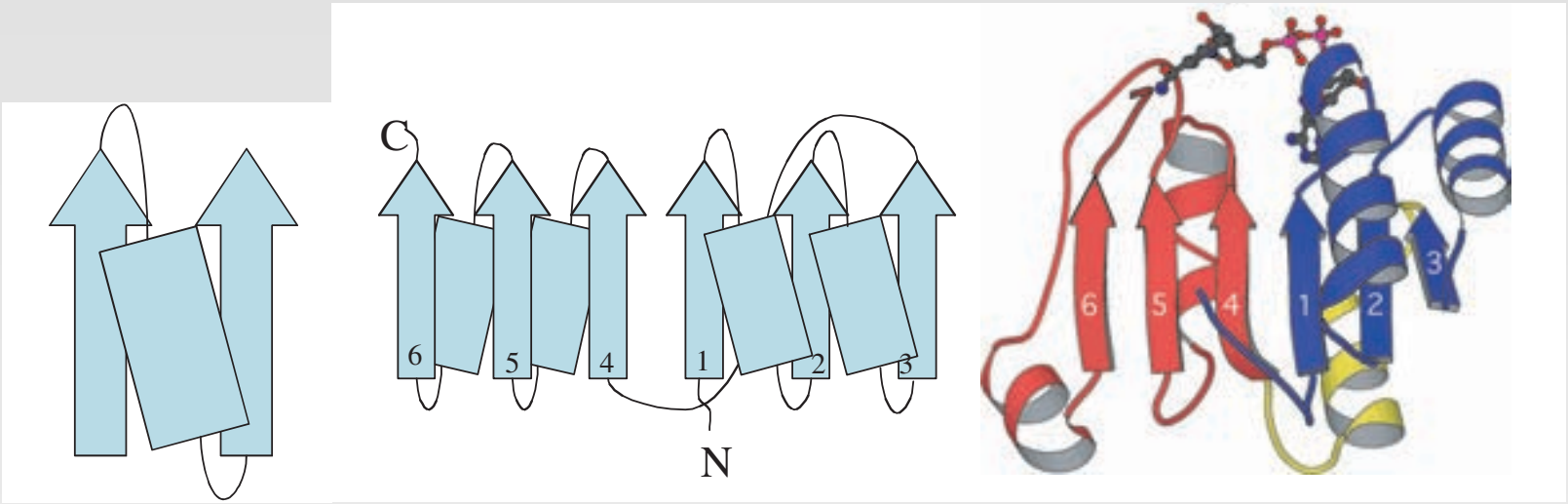


FIGURE 2.18 ■ *Left*: A jellyroll topology (*top*) can be seen as a β-hairpin rolled up. *Right*: The coat protein of the plant virus STNV, a simple jellyroll fold (PDB: 2BUK).

Motifs, topologies, folds: $\beta\alpha\beta$ structures



Motifs, topologies, folds: TIM barrel

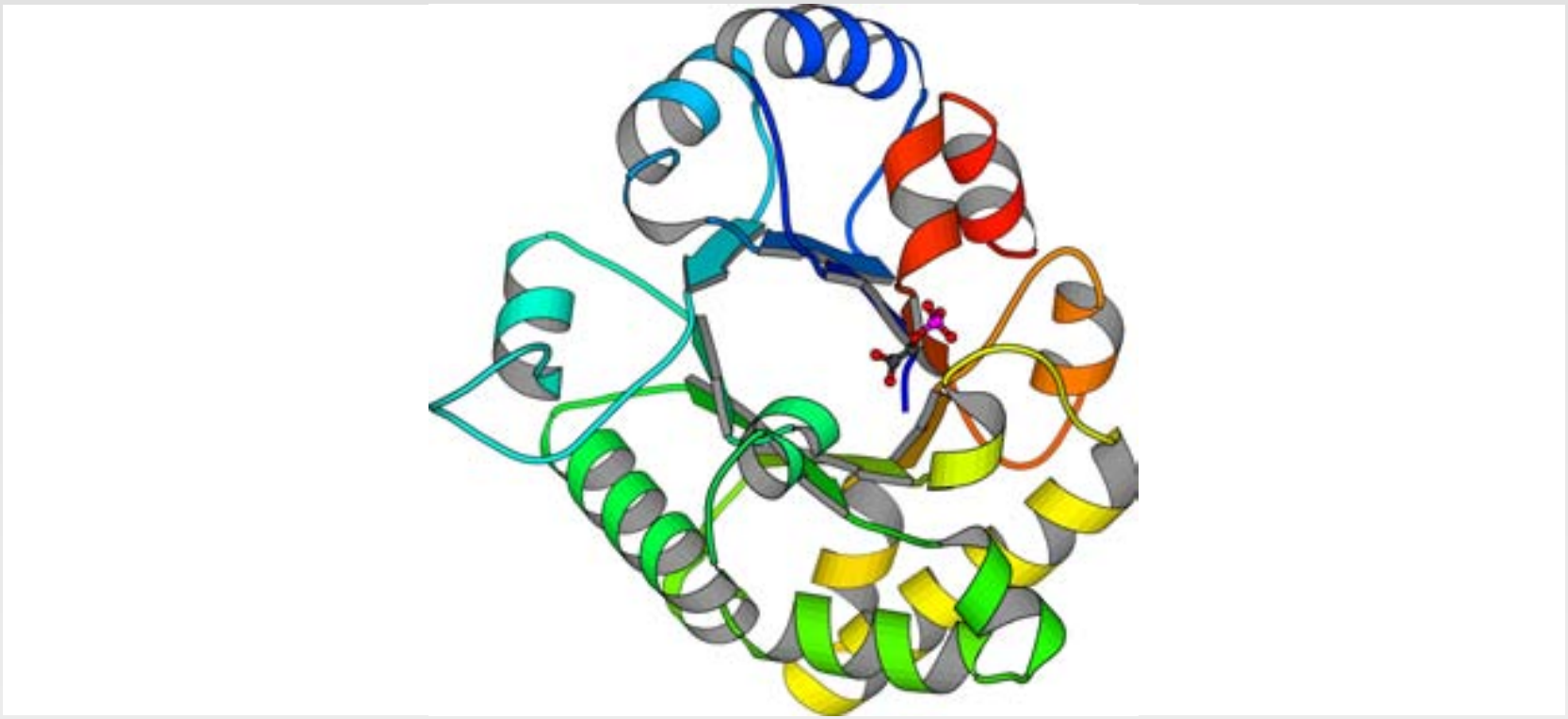


FIGURE 2.22 ■ The TIM barrel of triose phosphate isomerase from *Plasmodium falciparum*. The order of the strands is 12345678. The active site is occupied by a transition-state analog, phosphoglycolate. In all TIM barrel structures, the active site is found at the C-terminal end of the β-strands (PDB: 1LYX).

Motifs, topologies, folds: αhelix packing

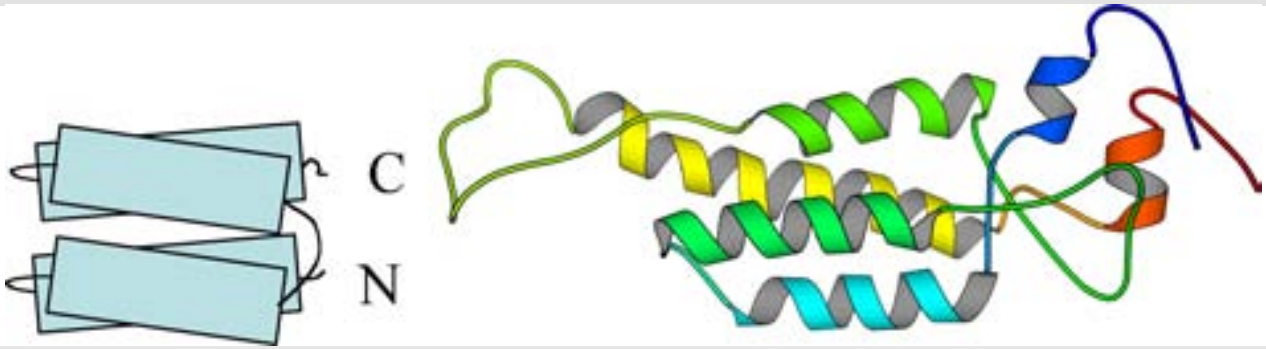


FIGURE 2.23 ■ Schematic drawing of an up-and-down four-helix bundle and the coat protein of tobacco mosaic virus (PDB: 2TMV).

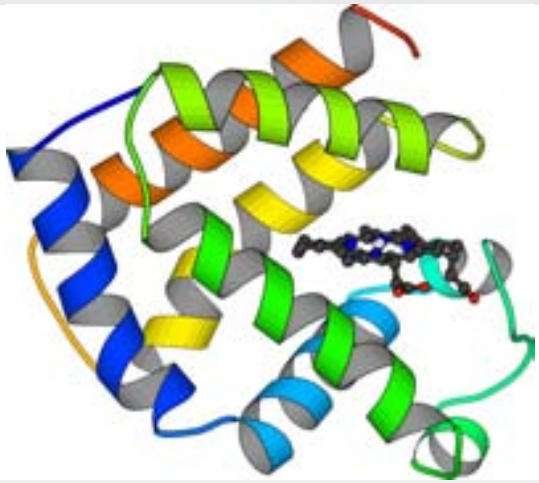


FIGURE 2.24 ■ The myoglobin structure where most of the helices pack with an approximately 50° angle between helix axes. The protein binds a heme group (PDB: 1MBA).

Proteins often consist of modular domains

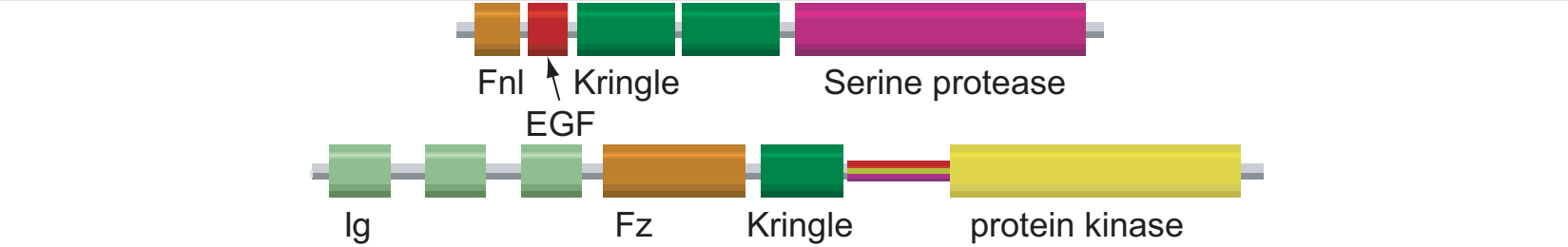


FIGURE 2.27 ■ Domain organization of two typical multi-domain proteins, tissue plasminogen activator (*top*) and a receptor tyrosine kinase (*bottom*) as presented in the Pfam database (http://pfam.sanger.ac.uk), where the cylinders are links to the corresponding family.

FIGURE 2.25 ■ A two-domain protein, transducin α, with one domain (red) as an insertion in the other domain, a G-domain (blue). The amino acid sequence is shown in blue and red for the main and inserted domain, respectively.

Common protein folds

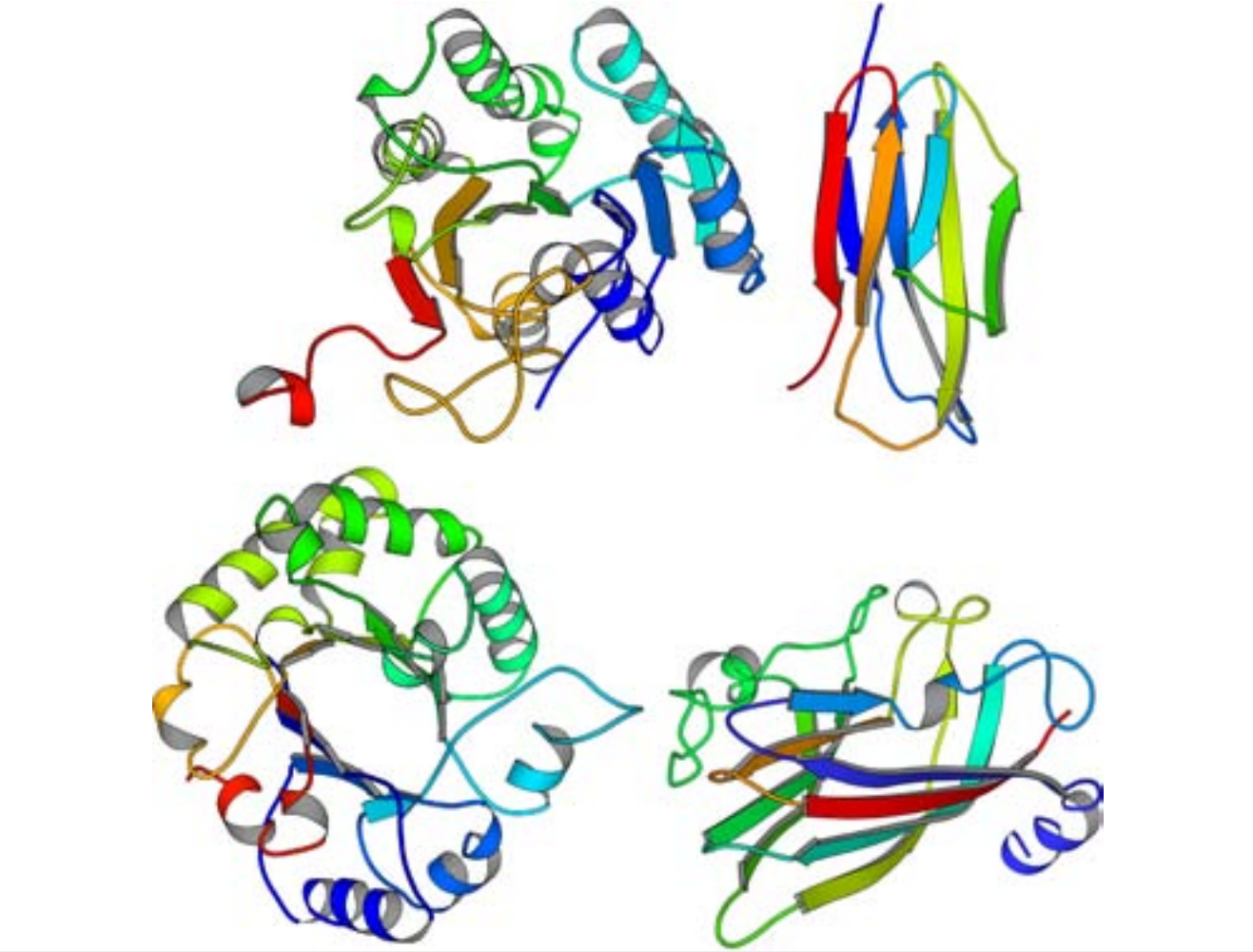


FIGURE 2.31 ■ Schematic drawings of a number of common protein folds. *Top left*: α/β domain with Rossmann fold (3α, 20β-hydroxysteroid dehydrogenase, PDB: 1HDC). *Top right*: Immunoglobulin constant domain (PDB: 1AQK). *Bottom left*: TIM barrel (triose phosphate isomerase, PDB: 1YPI). *Bottom right*: Jellyroll (satellite tobacco necrosis virus coat protein, PDB: 2BUK).

Post-translational modifications: glycosylation

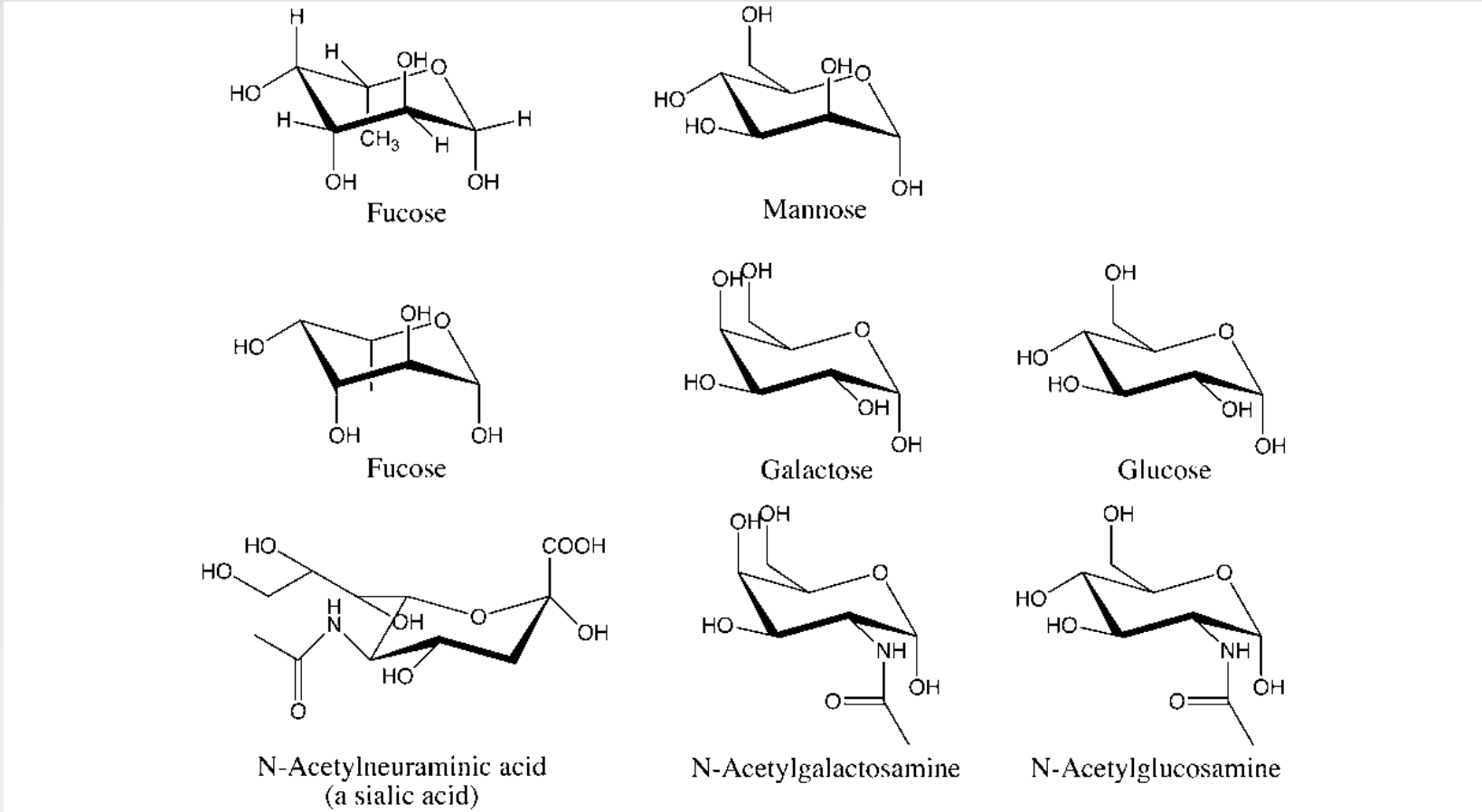


FIGURE E.7 ■ Some types of sugar residues that can be attached to proteins. The top *left* shows the more extensive formula for fucose where all hydrogens are included. *Below* is the same fucose without showing the hydrogens directly attached to the ring carbons.

Post-translational modifications: glycosylation

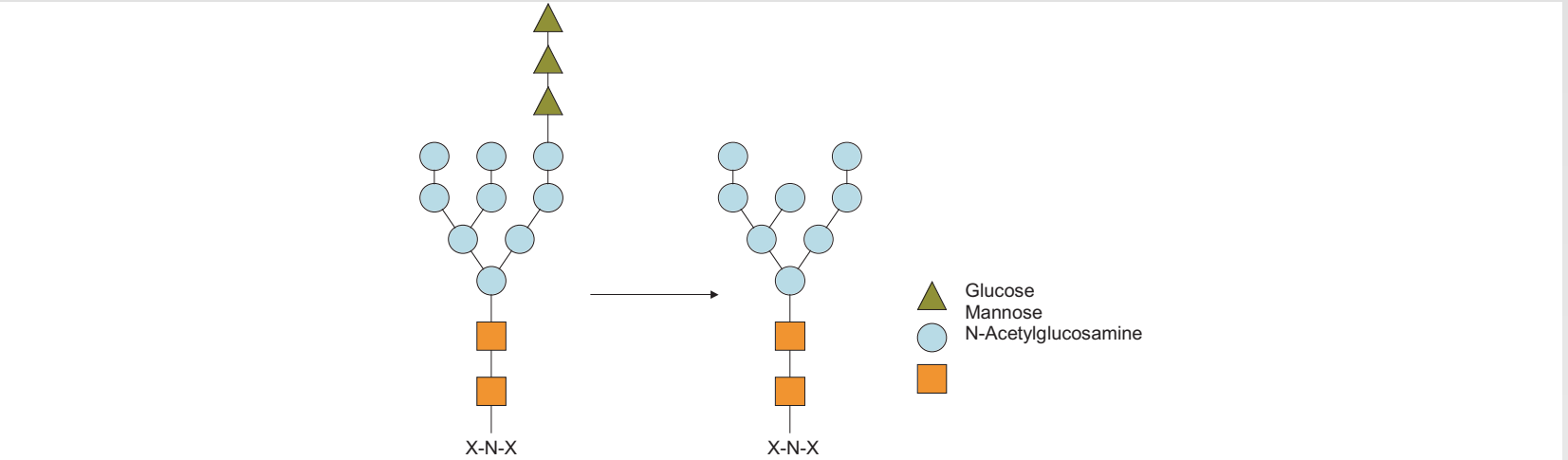


FIGURE E.8 ■ An example of how an asparagine side chain becomes attached by N-linked oligosaccharides in the form of a 14-mer of three different types of sugar moieties. The oligosaccharide is reduced to a 10-mer in several steps before further specific modifications are made.

Post-translational modifications: phosphorylation

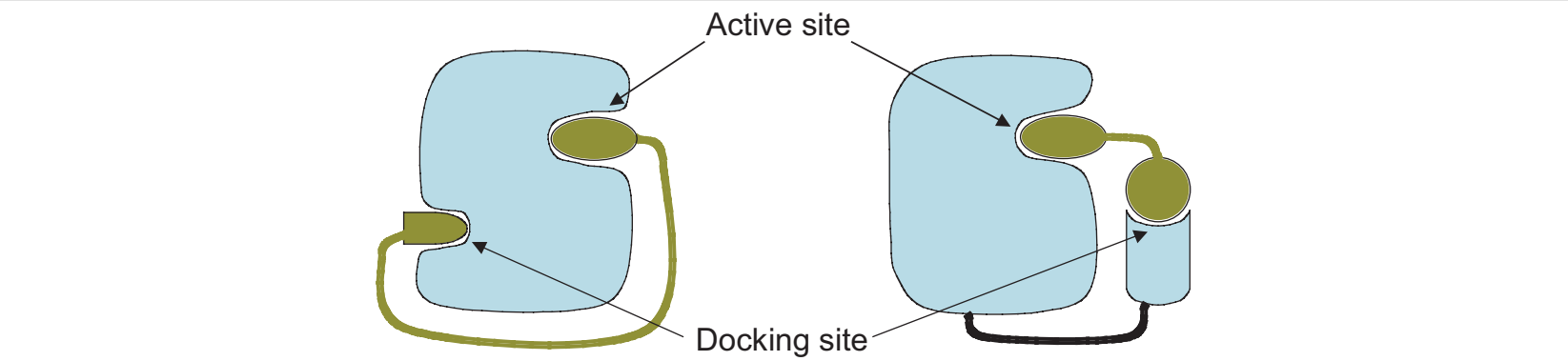


FIGURE E.4 ■ The regulatory role of kinases and phosphatases requires extensive control of the protein interactions. The enzymes (blue) partly need to be activated but also need to identify structures other than the part that will be phosphorylated or dephosphorylated. This is due to substrate (green) interactions with a docking site that can be part of the kinase domain or belong to different parts of the enzyme.

Post-translational modifications: acetylation

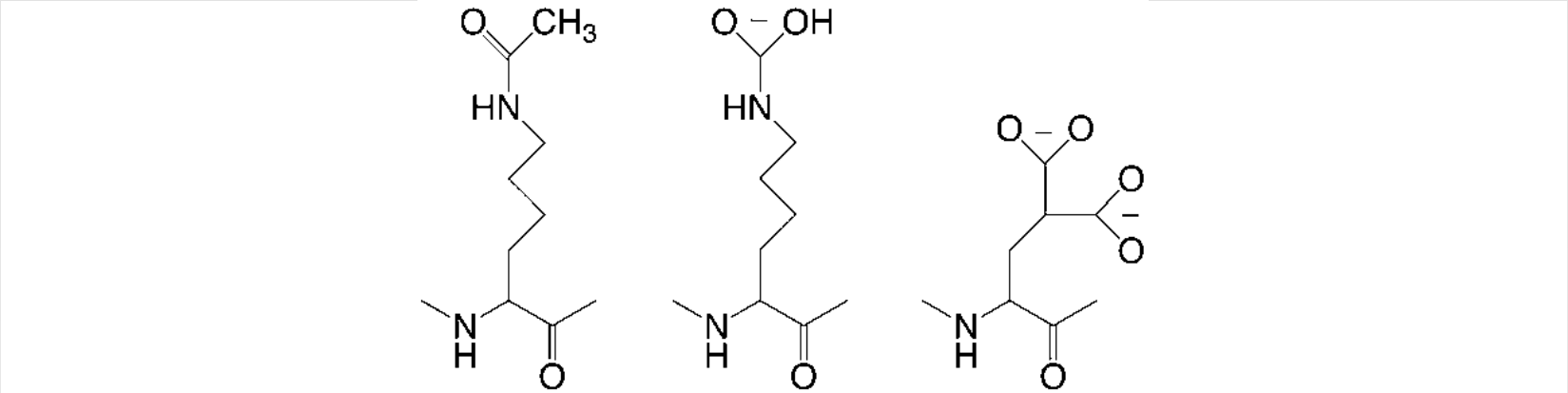


FIGURE E.6 ■ *Left*: An acetylated lysine. This modification is important in the modifications of histones. *Middle*: A carbamylated lysine, which has been found in several enzyme active sites with two metals. The elongated residue can bridge between the two metals. *Right*: A γ-carboxyglutamic acid (Gla) residue. This modification is abundant in the blood clotting system.

Post-translational modifications: disulfide bonds

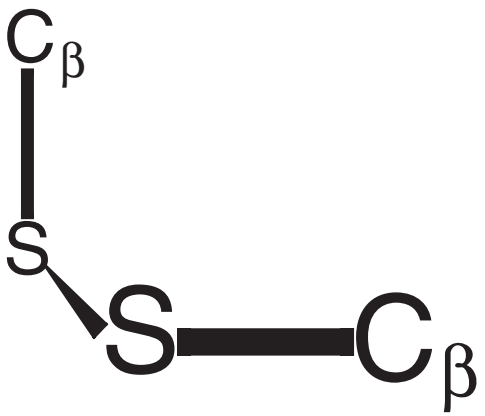


FIGURE A.1 ■ The preferred conformation of a disulfide bond with a 90° angle between the S–C_β bonds viewed down the S–S bond.

Metals in proteins: “blue” copper

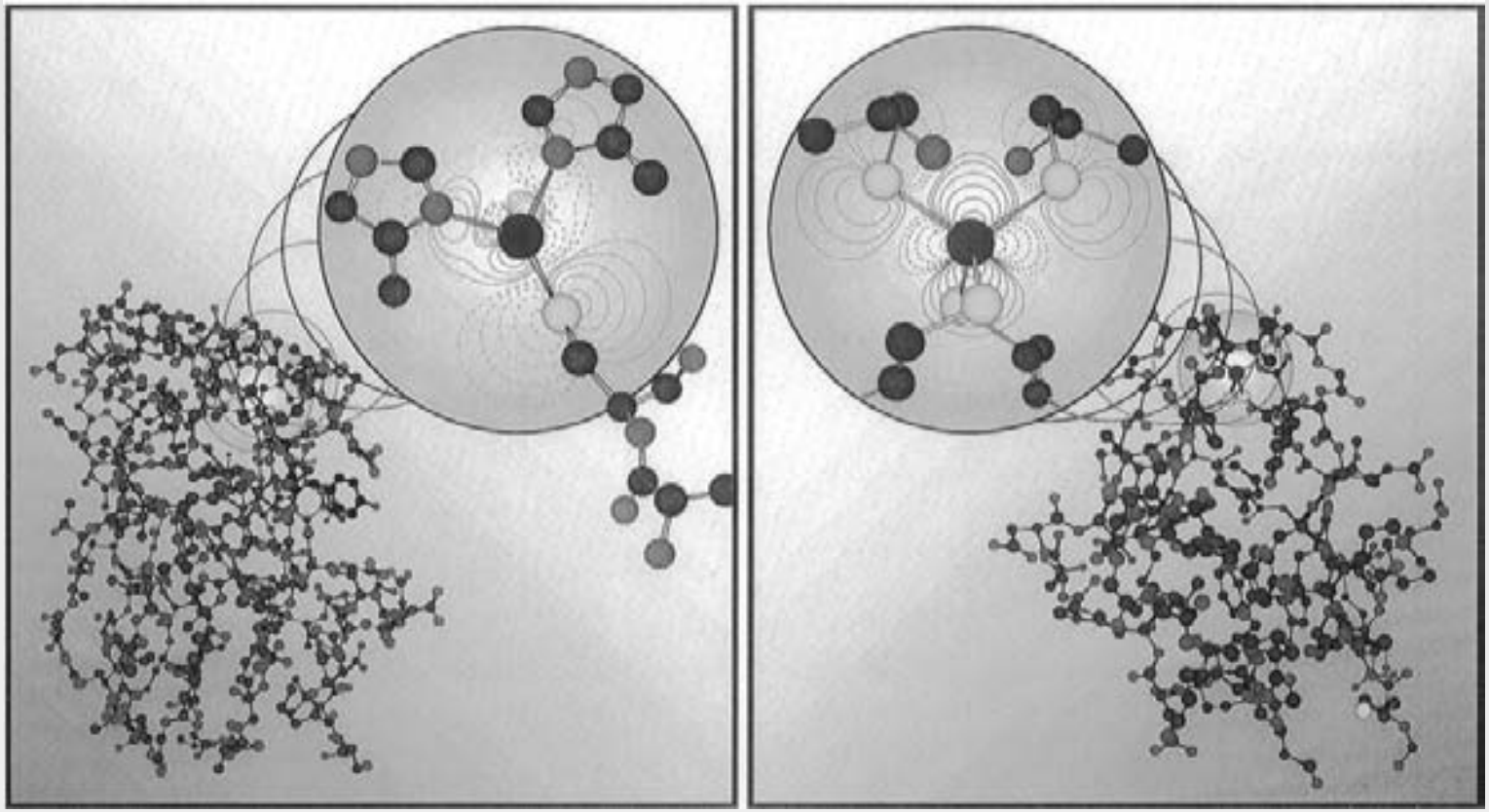
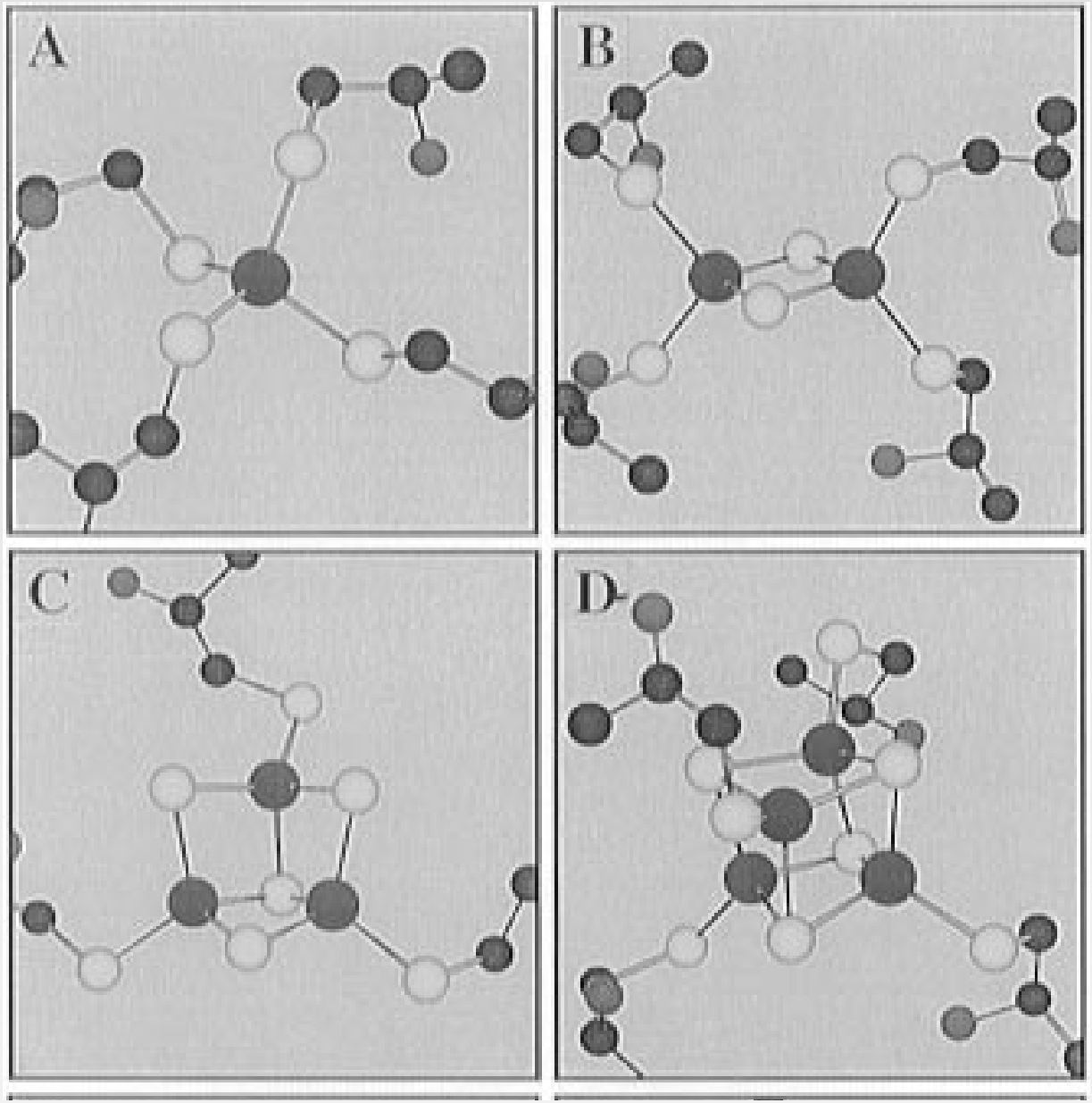


Figure 1. Expanded views of geometric and electronic structure of the active sites of plastocyanin (left) and rubredoxin (right). The expanded plastocyanin site is rotated such that the Met·S−Cu bond is out of the plane of the page. Contour values are set to ±0.16, ±0.08, ±0.04, ±0.02, and ±0.01 (e/μ_B3), with positive contours in solid red and negative values in dashed blue.

Metals in proteins: Iron-sulfur centers



Protein families: the SCOP classification

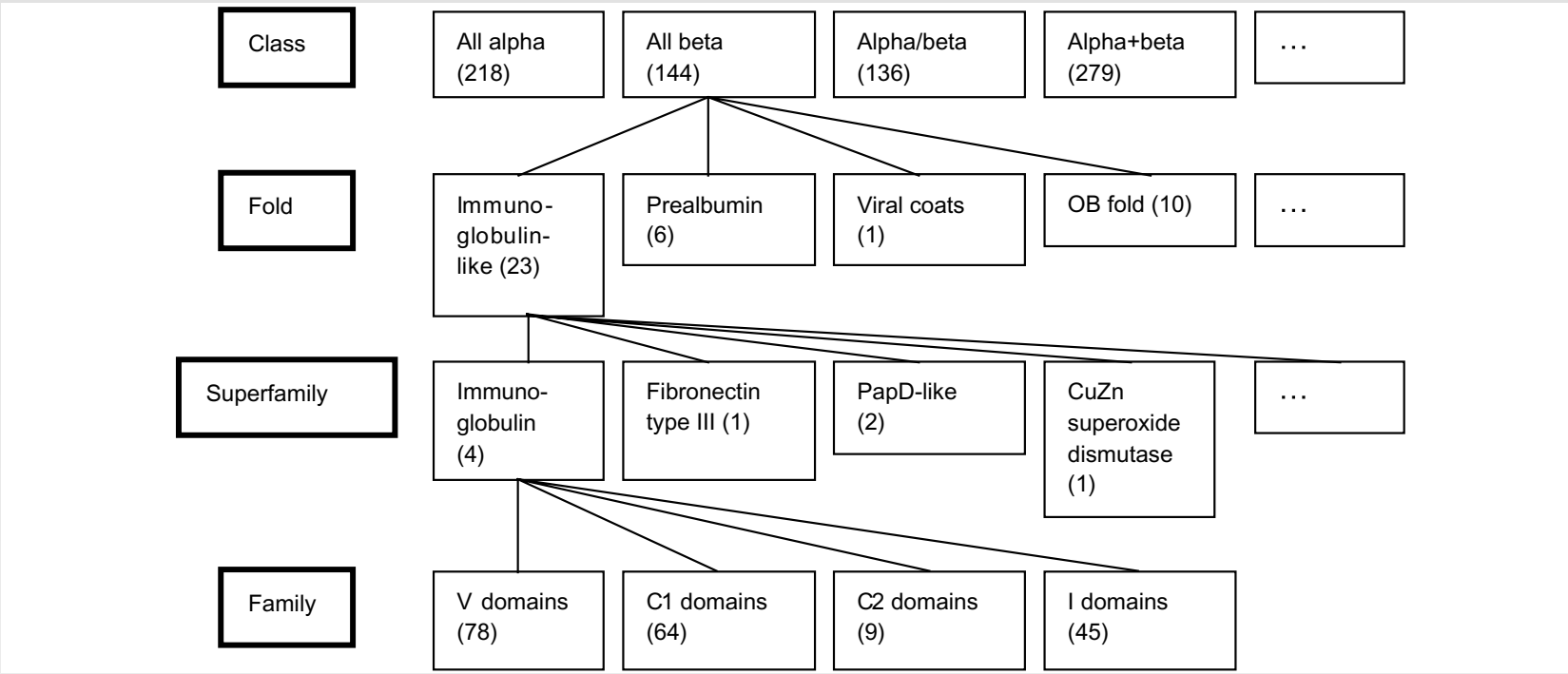


FIGURE B.5 ■ SCOP classification. The example shows the hierarchy for immunoglobulin domains. The number of different subdivisions of a group is shown in parentheses.

Protein families: the CATH classification

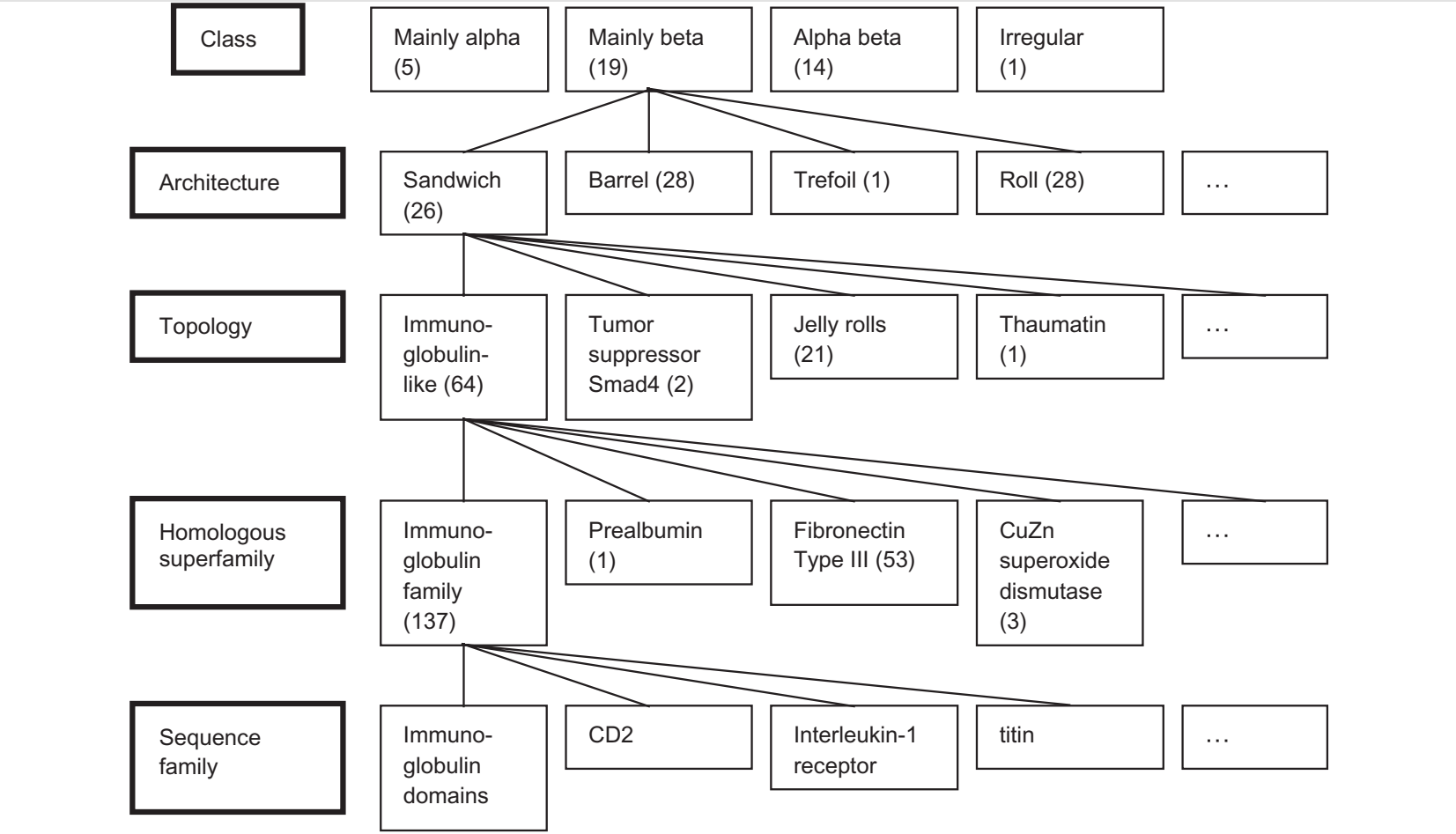
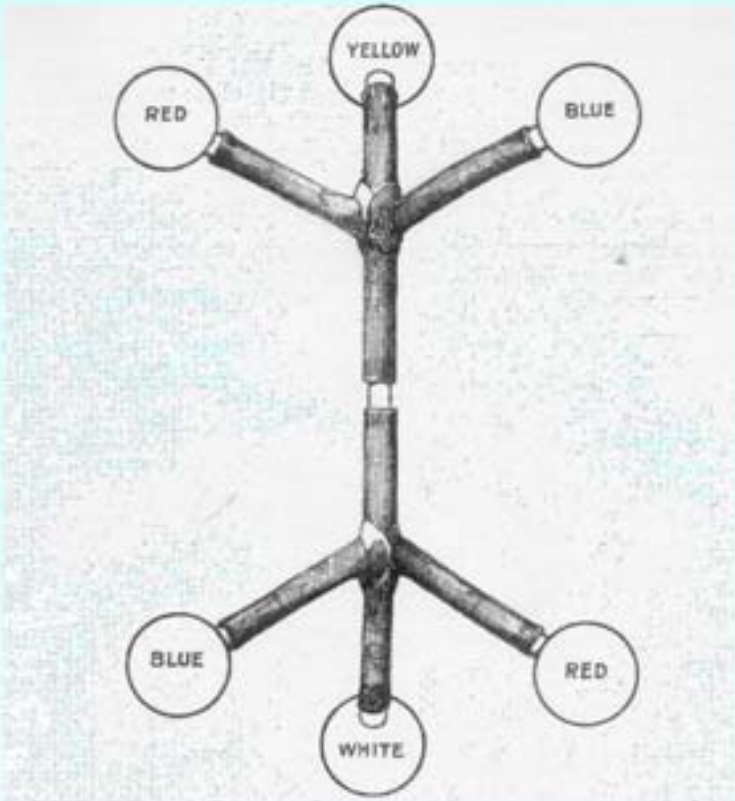


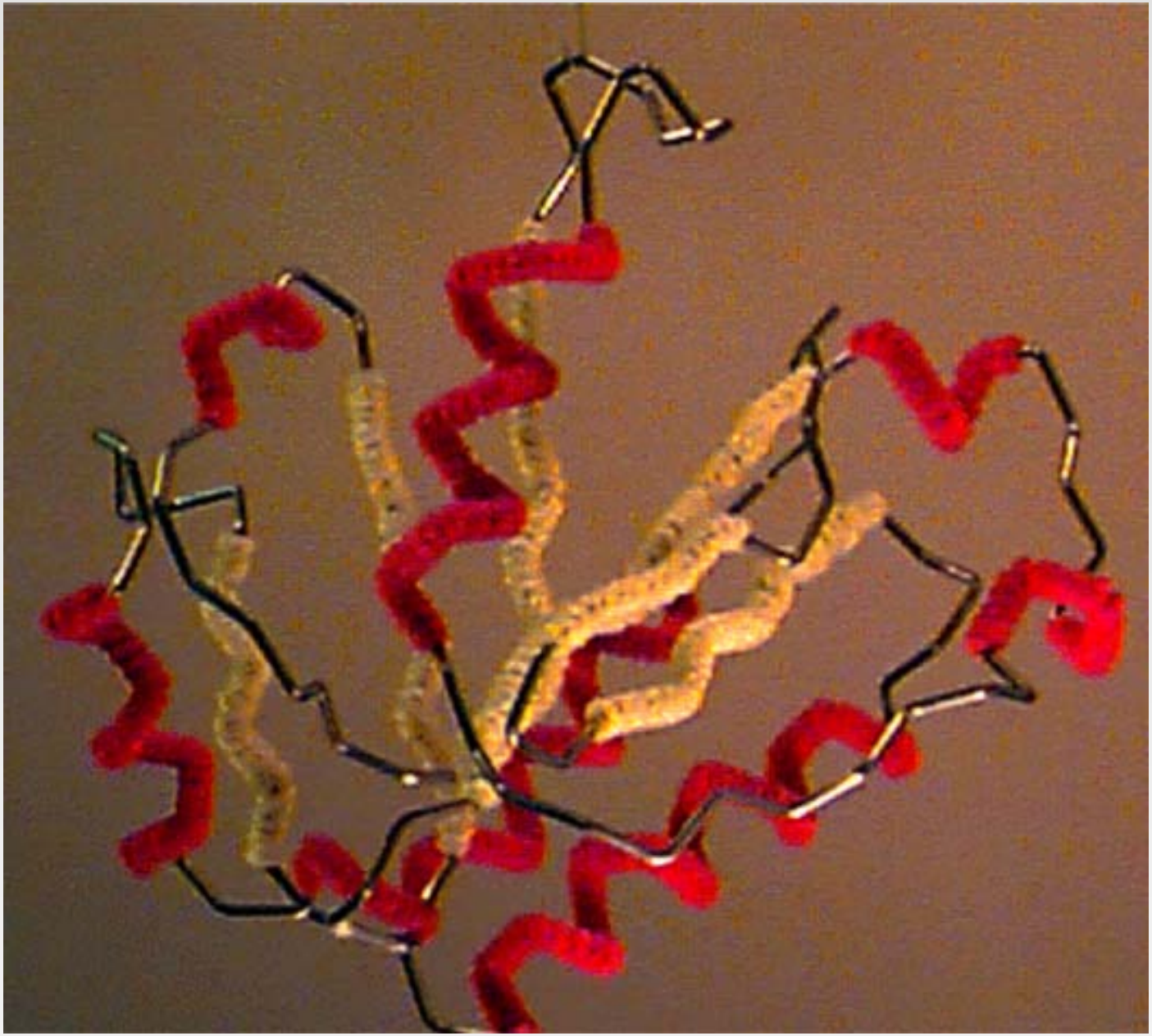
FIGURE B.6 ■ CATH classification. The immunoglobulins are used as an example again. Note that prealbumin (transthyretin) is classified as part of the immunoglobulin-like topology in CATH, but is a separate fold in SCOP (Fig. B.5).

1901 (and earlier?) ball and stick models

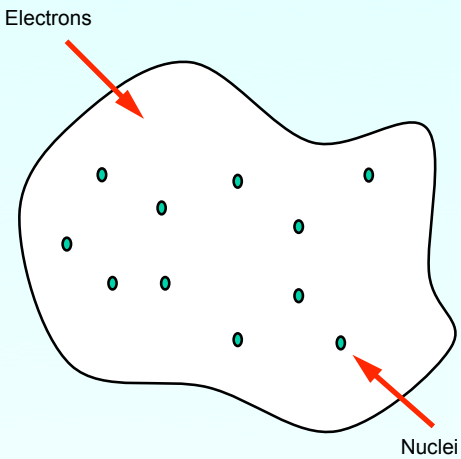


Baird & Tatlock 1901

1950s: wire models of proteins

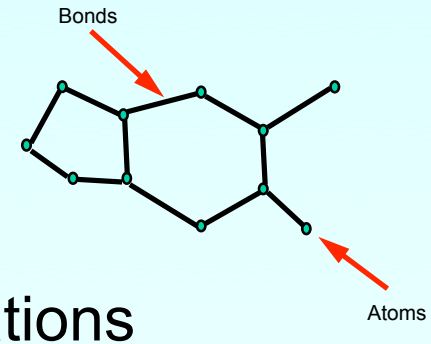


- separate nuclei and electrons
- polarisation, electron transfer and correlation
- can specify electronic state
- can calculate formation energies
- can do chemistry (bond breaking and making)
- variationally bound
- computationally expensive
- typically ~10-100 atoms
- dynamics ~1 ps



QM MOLECULE

- no explicit electrons, net atomic charges
- no polarisation, electron transfer or correlation
- conformational energies for ground state
- no chemistry
- semi-empirical force fields
- not variationally bound
- solvent and counterion representations
- typically ~1000-100000 atoms
- dynamics up to ~100 ns



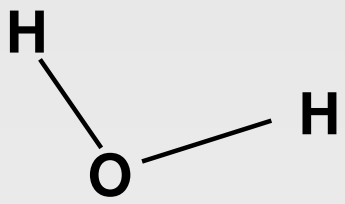
MM MOLECULE

Some force field assumptions

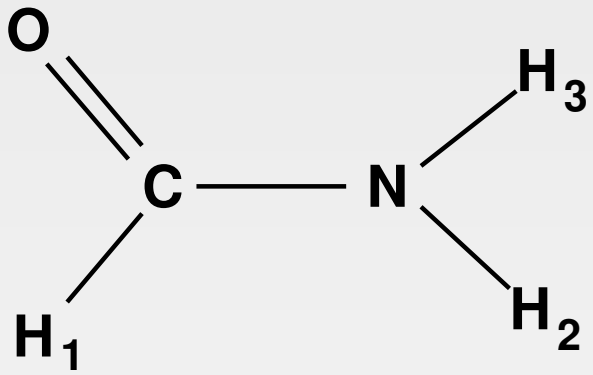
1. Born-Oppenheimer approximation (separate nuclear and electronic motion)
2. Additivity (separable energy terms)
3. Transferability (look at different conformations, different molecules)
4. Empirical (choose functional forms and parameters based on experiment)

What does a force field look like?

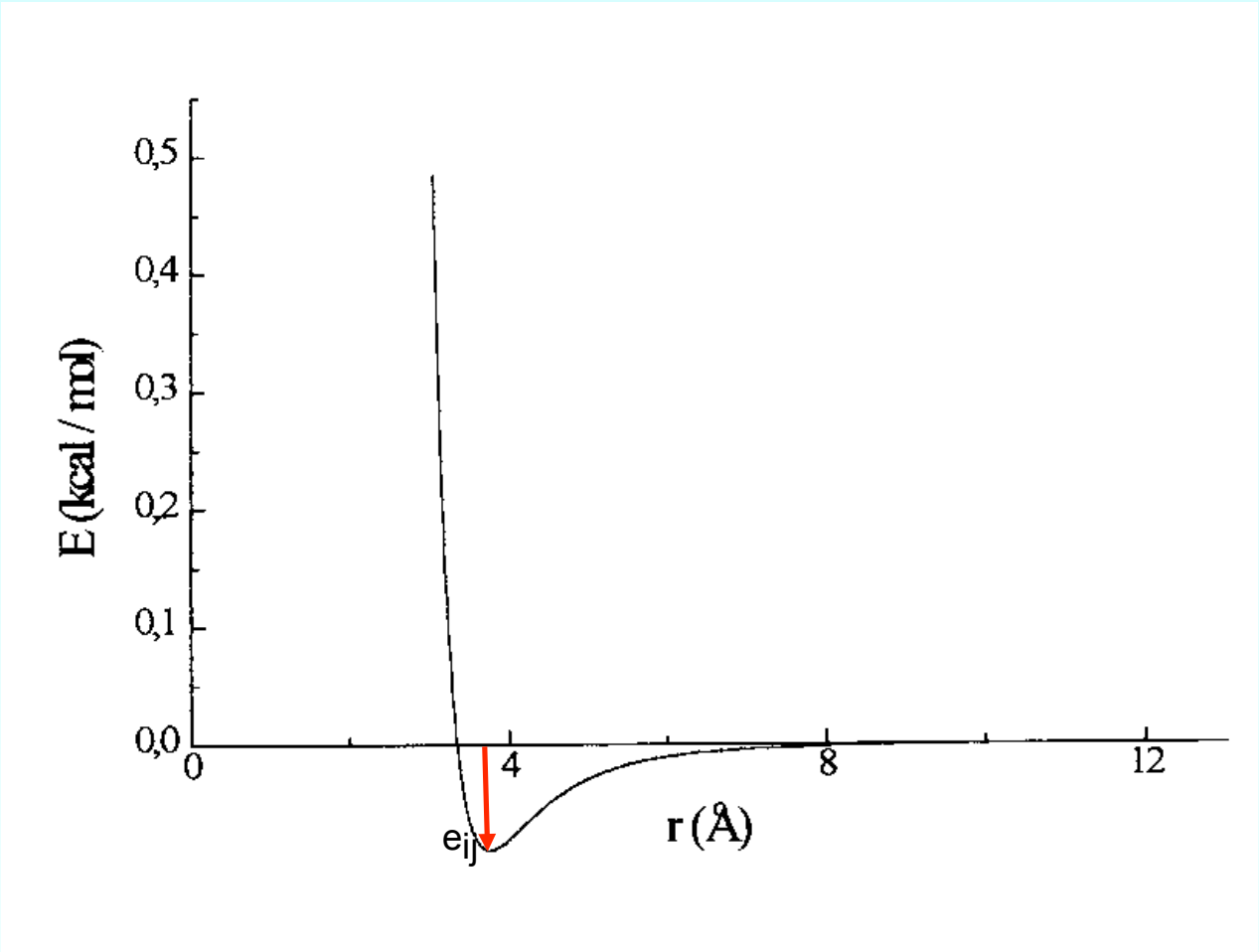
$$U = \sum_{\text{bonds}} K_b(b - b_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 + \sum_{impropers} K_w w^2$$
$$+ \sum_{torsions} K_\phi \cos(n\phi) + \sum_{nonbonded\ pairs} \left\{ 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] + \frac{q_i q_j}{r} \right\}$$



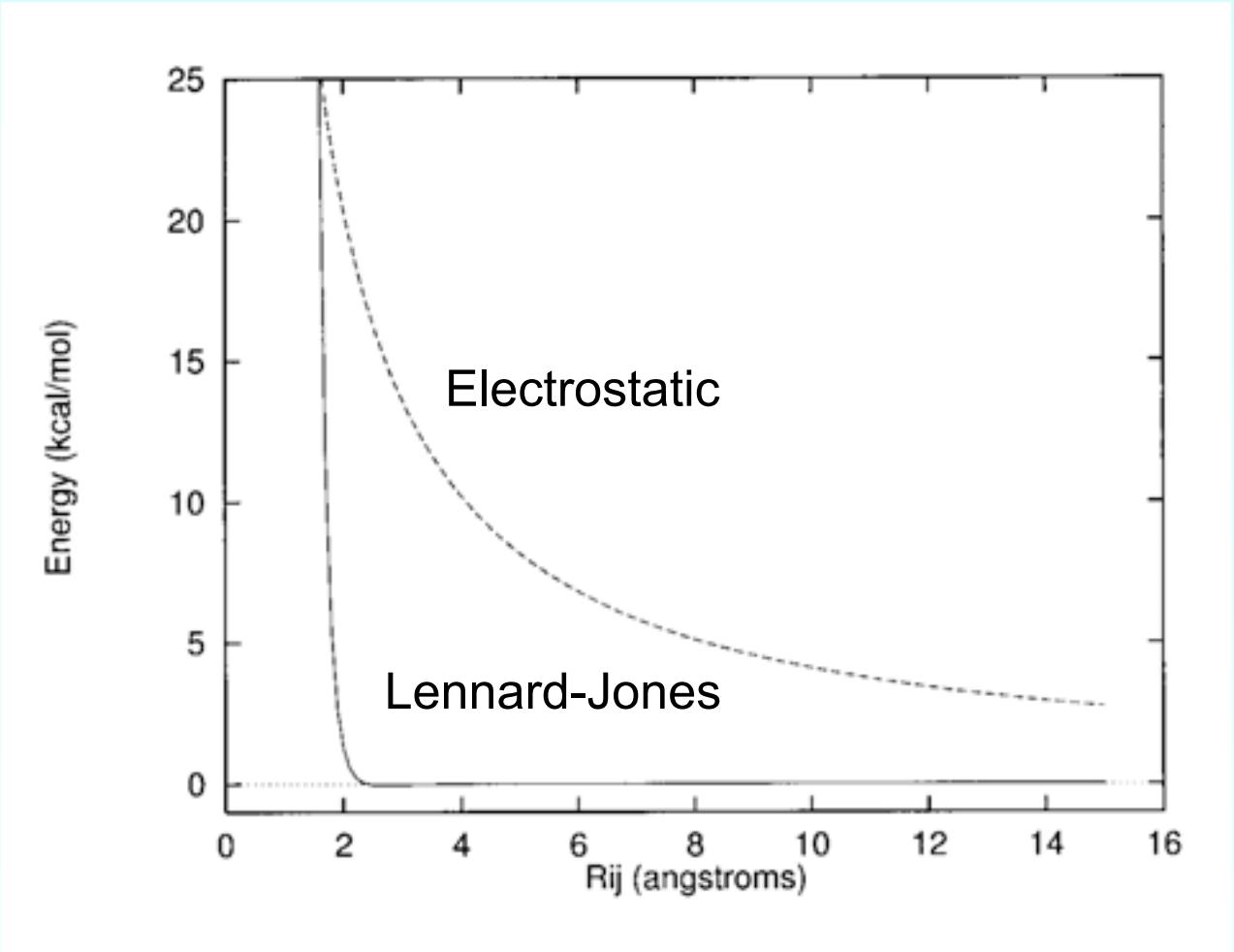
water



formamide



Lennard-Jones energy curve



Distance dependence

H	H bonded to nitrogen atoms
HC	H aliph. bond. to C without electrwd.group
H1	H aliph. bond. to C with 1 electrwd. group
H2	H aliph. bond. to C with 2 electrwd.groups
H3	H aliph. bond. to C with 3 eletrwd.groups
HA	H arom. bond. to C without elctrwd. groups
H4	H arom. bond. to C with 1 electrwd. group
H5	H arom. bond. to C with 2 electrwd. groups
HO	hydroxyl group
HS	hydrogen bonded to sulphur
HW	H in TIP3P water
HP	H bonded to C next to positively charged gr

AMBER parm94 H atom types

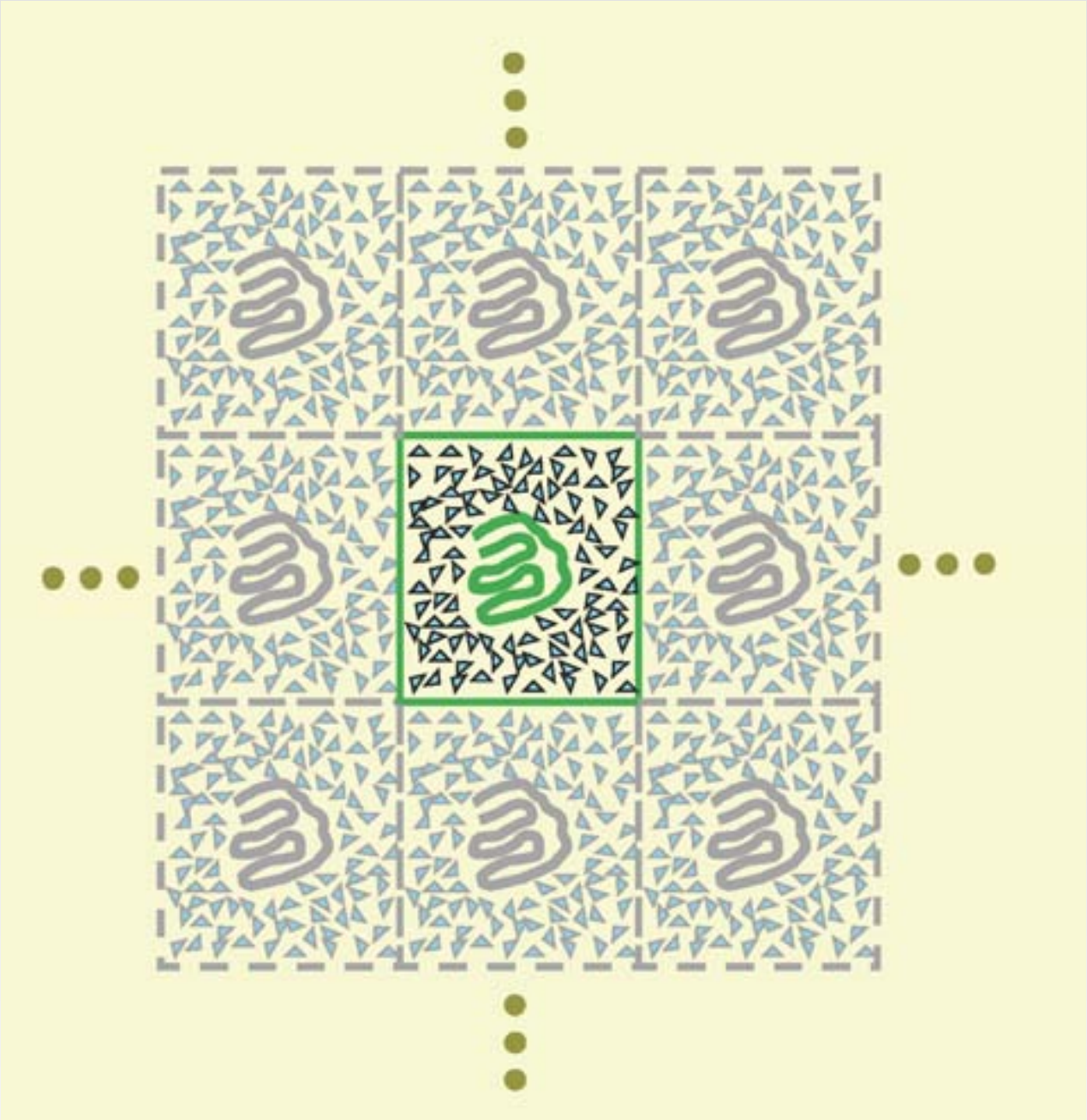
C	sp2 C carbonyl group
CA	sp2 C pure aromatic (benzene)
CB	sp2 aromatic C, 5&6 membered ring junction
CC	sp2 aromatic C, 5 memb. ring HIS
CK	sp2 C 5 memb.ring in purines
CM	sp2 C pyrimidines in pos. 5 & 6
CN	sp2 C aromatic 5&6 memb.ring junct.(TRP)
CQ	sp2 C in 5 mem.ring of purines between 2 N
CR	sp2 arom as CQ but in HIS
CT	sp3 aliphatic C
CV	sp2 arom. 5 memb.ring w/1 N and 1 H (HIS)
CW	sp2 arom. 5 memb.ring w/1 N-H and 1 H (HIS)
C*	sp2 arom. 5 memb.ring w/1 subst. (TRP)

AMBER parm94 C atom types

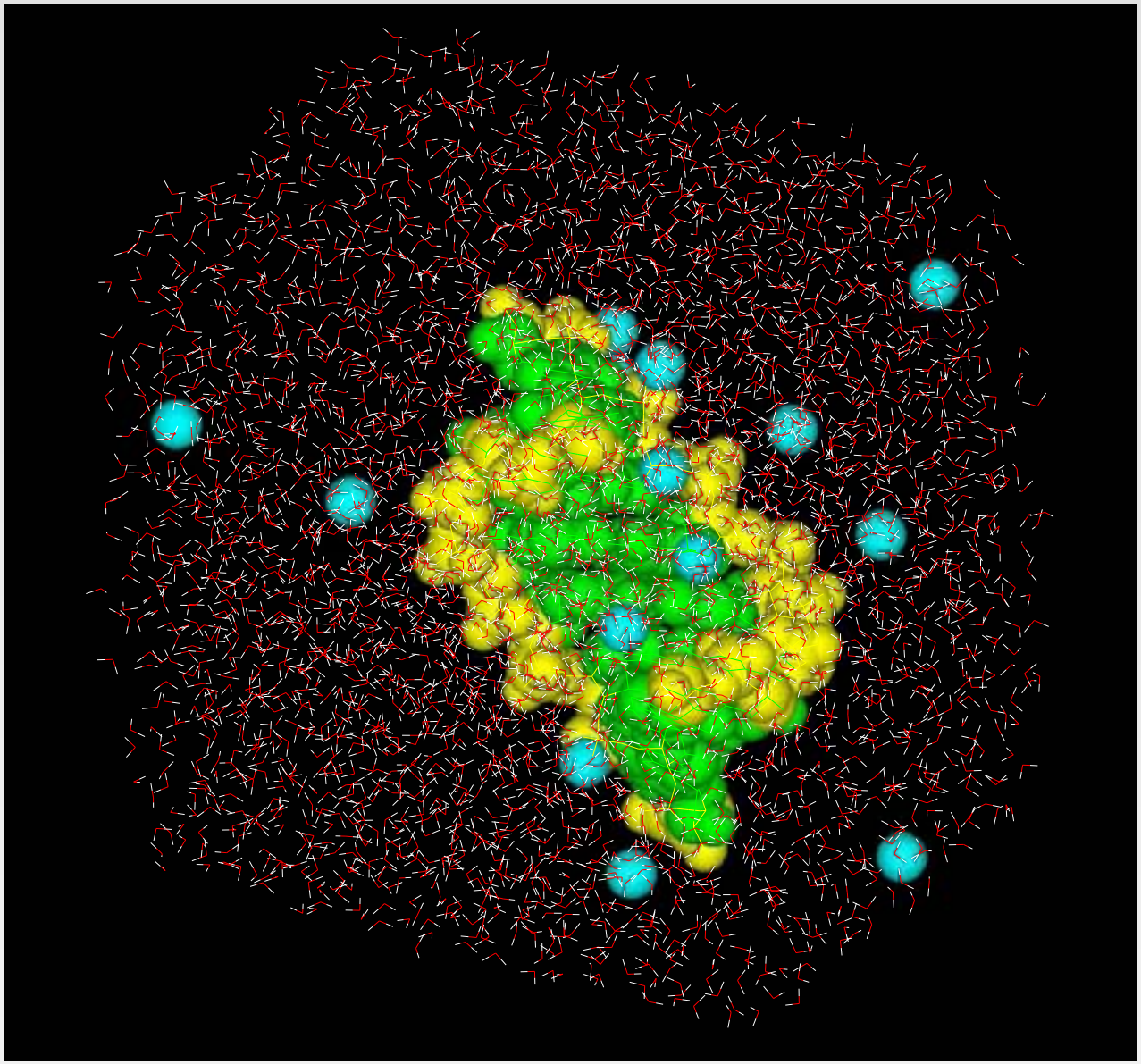
Force fields in Amber

- ff94: widely used ("Cornell et al.), pretty good nucleic acid, too much α-helix for proteins
- ff99: major recalibration by Junmei Wang and others; basis for most current Amber ff's
- ff99SB: recalibration of backbone potentials for proteins by Carlos Simmerling ("SB")
- ff02r1: polarizable extension for ff99
- ff03: new charge model (Yong Duan) + backbone torsions for proteins
- ff03ua: united atom extension

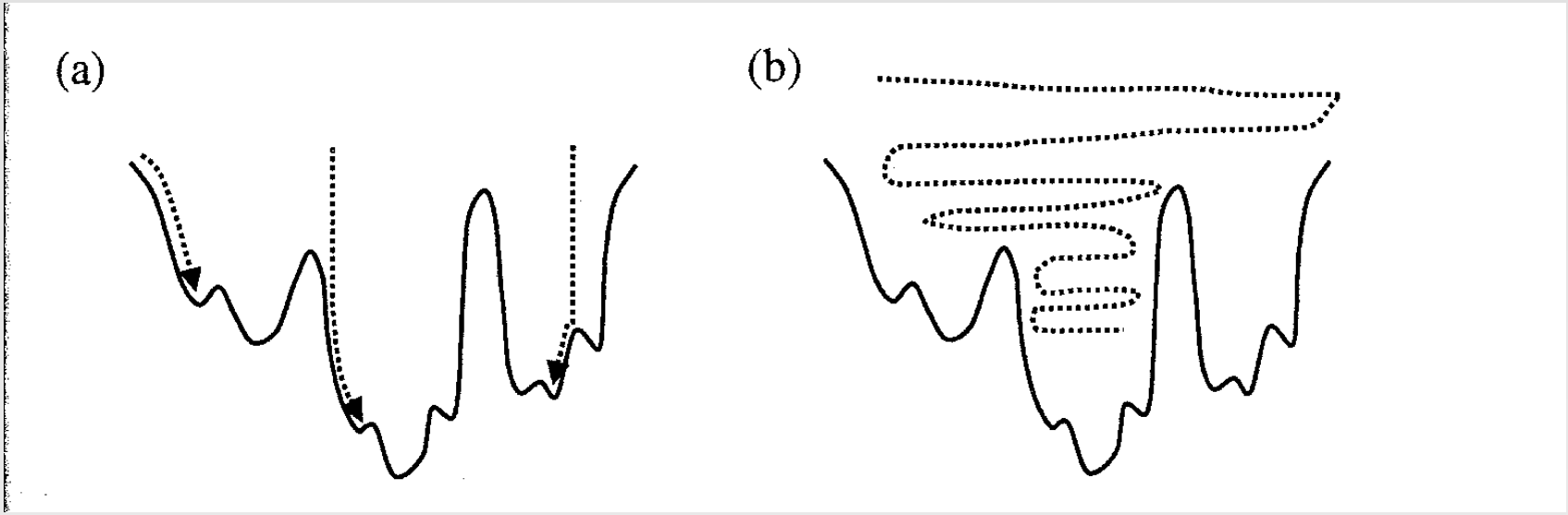
Periodic boundary conditions



Example of explicit solvation setup



Minimization and simulated annealing



Molecular dynamics algorithms

$$
\begin{aligned}
x(t+h) &= x(t) + v(t)h + \frac{1}{2}a(t)h^2 + \frac{1}{6}\frac{d^3x}{dt^3}h^3 + O(h^4) \\
x(t-h) &= x(t) + -v(t)h + \frac{1}{2}a(t)h^2 - \frac{1}{6}\frac{d^3x}{dt^3}h^3 + O(h^4) \\
x(t+h) &= 2x(t) - x(t-h) + a(t)h^2 + O(h^4) \qquad (1) \\
x(t+h) - x(t) &= x(t) - x(t-h) + a(t)h^2 + O(h^4) \\
v(t+\frac{1}{2}h) &= v(t-\frac{1}{2}h) + a(t)h + O(h^3) \qquad (2) \\
x(t+h) &= x(t) + v(t+\frac{1}{2}h)h + O(h^4) \qquad (3)
\end{aligned}
$$

Eq. (1) is the original Verlet propagation algorithm; Eqs. 2 and 3 are the "leap-frog" version of that. Remember that $a = d^2x/dt^2 = F/m = (\partial V/\partial x)/m$. See pp. 42-47 in Becker & Watanabe.

Regulating temperature

"Temperature" is a measure of the mean kinetic energy. The instantaneous KE is

$$T(t) = \frac{1}{k_B N_{dof}} \sum_i^{N_{dof}} m_i v_i^2$$

(cf. classical rule of thumb: "$k_B T/2$ of energy for every squared degree of freedom in the Hamiltonian")
Suppose the temperature is not what you want. At each step, you could scale the velocities by:

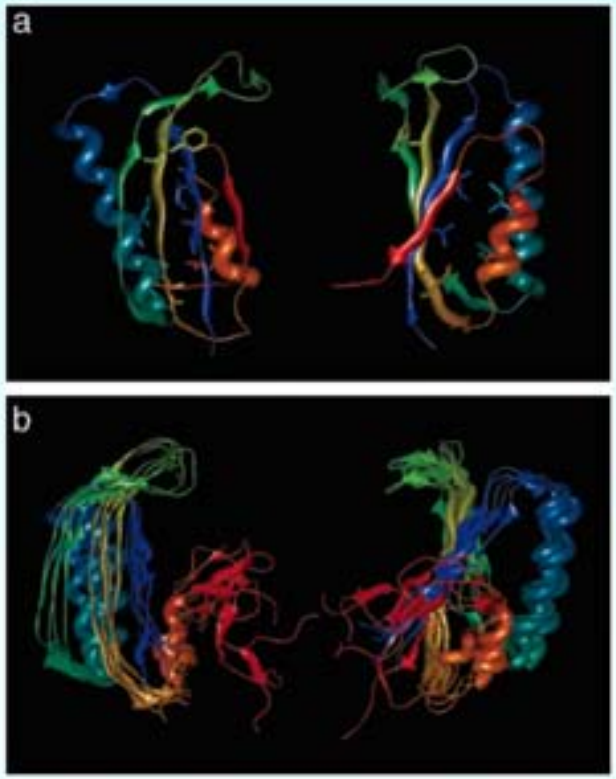
$$\lambda = \left[1 + \frac{h}{2\tau}\left(\frac{T_0}{T(t)} - 1\right)\right]^{1/2}$$

This is the "Berendsen" or "weak-coupling" formula, that has a minimal disruption on Newton's equations of motion. But it does not guarantee a canonical distribution of positions and velocities. See Morishita, J. Chem. Phys. 113:2976, 2000; and Mudi and Chakravarty, Mol. Phys. 102:681, 2004.

Go model for protein folding

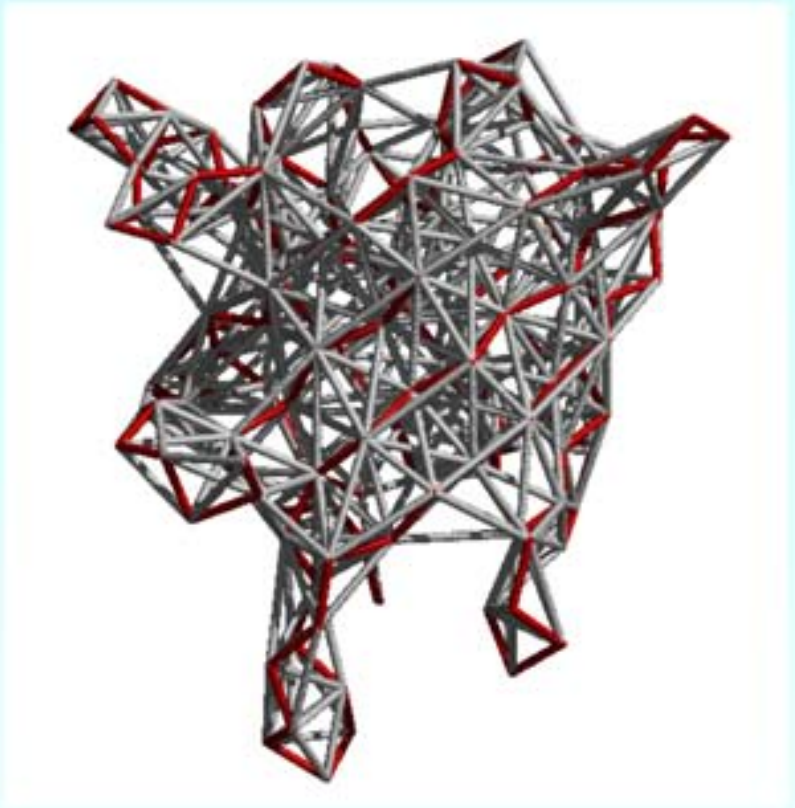
Native

TSE



- square-well potential
- native contacts "+1"
- non-native contacts "-1"
- cannot represent frustration during folding

Gaussian network model

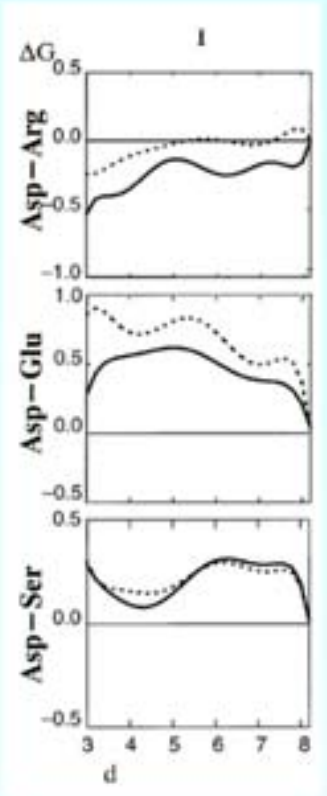


$$r_{ij}^{n} < R_{cut}$$

$$E_{ij} = k(r_{ij} - r_{ij}^{n})^{2}$$

Knowledge-based potentials

- Start from set of known protein structures
- Assume energy can be decomposed into residue pair interactions
- Assume that frequency of interactions within the ensemble ≡ frequency of interactions within the equilibrium ensemble of a single protein
- Derive potential of mean force for residue pairs from observed occurrence probabilities
- Knowledge-based potentials are used in both threading and folding



- - - small proteins

—— large proteins

Rooman & Gilis Eur. J. Biochem. 254, 1998, 135